К. Мурашев

Практическое руководство по оценке совокупного износа транспортных средств, относящихся к категориям M1, M2, M3, N1, N2, N3, оснащённых двигателями внутреннего сгорания, не подвергавшихся капитальному ремонту, с помощью современных технологий методами Data Mining

Санкт-Петербург

2019

Сведения о правах на данное произведение

Оглавление

Список таблиц

Список иллюстраций

Список формул

Библиография

1: https://en.wikipedia.org/wiki/KISS_principle
2: https://en.wikipedia.org/wiki/High-level_programming_language
3: https://en.wikipedia.org/wiki/Scripting_language
4: https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai/21669
5: http://adv-r.had.co.nz/Style.html
6: https://style.tidyverse.org
7: https://google.github.io/styleguide/Rguide.xml
8: https://www.bioconductor.org/developers/how-to/coding-style/
9: А.П. Ковалев, д.э.н., профессор; А.А. Кушель, к.т.н., доцент; В.С. Хомяков, д.т.н., профессор; Ю.В. Андрианов, к.т.н.; Б.Е. Лужанский, д.э.н., профессор; И.В. Королев; С.М. Чемерикин., Оценка стоимости машин, оборудования и траснпортных средств, Интерреклама, Москва, 2003, , 488, 5-8137-0101-X
10: В.П. Антонов, Е.В. Антонова, С.К. Шамышев, Р.Г. Шаулова, Оценка стоимости машин и оборудования, Русская оценка, Москва, 2005, , 254,
11: А.П. Ковалёв, А.А. Кушель, И.В. Королёв, П.В. Фадеев, Основы оценки стоимости машин и оборудования, Финансы и статистика, Москва, 2006, , 288, 5-279-03160-7
12: М.А. Федотова, В.Ю. Рослов, О.Н. Щербакова, А.И. Мышанов, Оценка для целей залога: теория, практика, рекомендации, Финансы и статистика, Москва, 2008, , 384, 978-5-279-03287-7
13: Ю.В. Андрианов, Оценка стоимости подвижного состава автомобильного транспорта, Международная академия оценки и консалтинга, Москва, 2003, , 257,
14: Смоляк С.А., Проблемы и парадоксы оценки машин и оборудования: сюита для оценщиков машин и оборудования, Международная академия оценки и консалтинга, Москва, 2009, , 305,
15: https://en.wikipedia.org/wiki/Operations_research
16: https://en.wikipedia.org/wiki/Applied_mathematics
17: Банк России, Положение о единой методике определения размера расходов на восстановительный ремонт в отношении поврежденного транспортного средства, 2014
18: Андрианов Ю.В., Учебно-методическое пособиепо дисциплине«Оценка стоимости транспортных средств», Московская финансово-промышленная академия, Москва, 2010, , 114,
19: Г.И. Микерин, В.Г. Гребенников, Е.И. Нейман, И.Л. Артёменков, А.В. Верхозина, Н.В. Павлов, С.А. Смоляк, Методологические основы оценки стоимости имущества, Интерреклама, Москва, 2003, , 688, 5-8137-0097-8
20: Махнин Е.Л., Новоселецкий И.Н., Федотов С.В., Галевский С.О., Калинин М.А., Кошелев Д.М., Суслов С.Б., Алексеев И.В., Калакутин А.В., Методические рекомендации по проведению судебных автотехнических экспертиз и исследований колёсных транспортных средств в целях определения размера ущерба, стоимости восстановительного ремонта и оценки, 2018
21: М.П. Улицкий, Ю.В. Андрианов, Б.Е. Лужанский, С.М. Чемерикин, Оценка стоимости транспортных средств, Финансы и статистика, Москва, 2005, , 304, 5-279-02879-7
22: Табакова С.А., Андрианов Ю.В., Мухин Е.М., Палочкин Р.Е., Прицкалов М.Е., Подшиваленко Д.В., Ковалёв А.П., Школьников Ю.В., Дарсания С.А., Владимиров П.В., Задорожный А.Н., Кузнецова Е.П., Методические указания Оценка стоимости в отношении транспортных средств, 2010

23: Н.В. Вейг, Оценка стоимости машин и оборудования, Издательство СПбГУЭФ, Санкт-Петербург, 2009, , ,
24: А.И. Попеско, А.В. Ступин, С.А. Чесноков, Износ технологических машин и оборудования при оценке их рыночной стоимости, Российское общество оценщиков, Москва, 2002, , 241, 5-93027-010-4
25: Мышанов А.И., к. т.н., Рослов В.Ю., к.т.н. , Модифицированный метод сроков жизни для расчёта износа оборудования, , , ,
26: Лейфер Л.А, Кашникова П.М., Определение остаточного срока службы машин и оборудования на основе вероятностных моделей, ЗАО "Приволжский центр финансового консалтинга и оценки", , , 2007
27: С.А. Смоляк, Эргодические модели износа машин и оборудования, , , , 2009
28: А. Ковалёв, О. Шинкевич, Определение износа при оценке машин и оборудования, , , , 2007
29: А.П. Ковалёв, Определение срока службы машин и оборудования при их стоимостной оценке, Имущественные отношения в РФ, 10 (181), 10, 2016
30: А.П. Ковалёв, Д. Игонин, Функциональное устаревание машин и оборудования: как учесть его при оценке, Вестник МГТУ Станкин, 4, , 2011
31: Дик Свааб, Мы - это наш мозг: от матки до Альцгеймера., Изд-во Ивана Лимбаха, Санкт-Петербург, 2014, , 544, 978-5-89059-202-6
32: Новицкий П.В., Зограф И.А., Оценка погрешностей результатов измерений., Энергоатомиздат, Ленинград, 1991, , 303,
33: M. P. Wand, Data-Based Choice of Histogram Bin Width, American Statistical Association, 51(1), Feb, 1997
34: Сара Бослаф, Статистика для всех, ДМК Пресс, 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2013, , 586, 978-5-94074-969-1
35: Cochran W.G., Some Methods of Strengthening the Common chi-square Tests, Biometrics, , 1954, V. 10, 417,
36: Mann H.B., Wald A., On the choice of the number of class intervals in the application of the chi square test, , , , 1942
37: Mann H.B., Wald A., On the choice of the number of intervals in the application of the chi-square test, , , , 1942
38: Sturgess H.A., The choice of classic intervals, Am. Statist. Assoc., , Jan, 1926
39: Шторм Р. , Теория вероятностей. Математическая статистика. Статистический контроль качества. , Мир, Москва, 1970, , 368,
40: Heinhold I., Gaede K.W., Ingeniur statistic., Springler Verlag, München; Wien, 1964, , 352,
41: Mann H.B., Wald A., On the choice of the number of intervals in the application of the chi-square test, , V. 13., , 1942
42: Таушанов З., Тонева Е., Пенова Р., Вычисление энтропийного коэффициента при малых выборках, 1973
43: Тонева Е., Аппроксимация распределений погрешности средств измерений, , № 6, , 1981
44: Алексеева И.У., Теоретическое и экспериментальное исследование законов распределения погрешностей, их классификация и методы оценки их параметров, 1975
45: http://www.machinelearning.ru/wiki/index.php?title=Моменты_случайной_величины
46: Бурдун Г.Д., Марков Б.Н., Основы метрологии., Изд-во стандартов, Москва, 1985, , 120,
47: Ченцов Н.Н., Статистические решающие правила и оптимальные выводы., Наука, Москва, 1972, , 520,

48: https://ami.nstu.ru/~headrd/seminar/xi_square/bibliographia.htm#i36
49:https://stackoverflow.com
https://stackoverflow.comhttps://stackoverflow.com/questions/7568356/3d-plot-in-r-patc
50:https://www.r-bloggers.com https://www.r-bloggers.comhttps://www.r-bloggers.com/waterfall-and-3d-plotting-exploration/
51: В. Савельев, Статистика и котики, ООО «Издательство АСТ», Санкт-Петербург, 2018, , 122, 978-5-17-106143-2
52:А. Дьяконов https://dyakonov.org/2018/07/30/байесовский-подход/
53:М. Иришкин https://habr.com/ru/post/170545/
54: http://www.machinelearning.ru/wiki/index.php?title=Несмещённая_оценка
55: http://www.machinelearning.ru/wiki/index.php?title=Квантиль
56: http://www.machinelearning.ru/wiki/index.php?title=Математическое_ожидание
57: http://www.machinelearning.ru/wiki/index.php?title=Коэффициент_асимметрии
58: http://www.machinelearning.ru/wiki/index.php?title=Коэффициент_эксцесса
59: https://r-analytics.blogspot.com/2011/11/r_08.html
60: https://ru.wikipedia.org/wiki/Ящик_с_усами#/media/Файл:Densityvsbox.png
61:Автор: Jhguch at en.wikipedia, CC BY-SA 2.5, Ссылка
https://ru.wikipedia.org/wiki/Ящик_с_усами#/media/Файл:Boxplot_vs_PDF.svg
62: http://www.machinelearning.ru/wiki/index.php?title=Доверительный_интервал

Введение

Целью данной работы является попытка объединения наработок в областях оценочной деятельности и Data mining. Автор доказывает возможность применения современных технологий в сфере оценки имущества, их эффективность и ряд преимуществ относительно иных методов определения износа. В условиях заданного руководством России курса на цифровизацию экономики и особенный фокус на развитие технологий искусственного интеллекта внедрение методов Data mining в повседневную практику оценщиков выглядит логичным и необходимым.

Актуальность данного исследования заключается во-первых в том, что оно даёт практический инструментарий, позволяющий делать обоснованные, поддающие верификации и доказательные выводы на основе одной только информации о предложениях на открытом рынке без использования каких-либо иных источников информации. Во-вторых, предложенные и рассмотренные методы обладают весьма широким функционалом, позволяющим использовать их в круге задач, намного более широком чем определение износа. Важность обеих причин автор видит в том, что по состоянию на 2019 год в оценочной деятельности в России сложилась ситуация, которую можно охарактеризовать двумя состояниями:

1) состояние неопределённости будущего отрасли;

2) состояние интеллектуального тупика.

Первая проблема заключается в неопределённости как правового регулирования отрасли, так и её экономики. Введённая около года назад система квалификационных аттестатов оценщиков, на которую регулятор, заказчики и, возможно, часть самих оценщиков возлагали надежду как на фильтр, позволяющий оставить в отрасли только квалифицированных специалистов, сократить предложение оценочных услуг и, следовательно, способствовать росту вознаграждений за проведение оценки, не оправдала ожиданий. Несмотря на существенное сокращение числа оценщиков, имеющих право подписывать отчёты об оценке, не произошло никаких значимых изменений ни в объёме предложения, ни в уровне цен на эти услуги. Фактически произошло лишь развитие уже существовавшего ранее института подписантов отчётов — оценщиков, имеющих квалификационные аттестаты, и выпускающих от своего имени отчёты, в т.ч. и те, которые они не только не готовили сами, но нередко и не читали, а порой и не видели в силу присутствия в другом регионе. При этом, как ни странно, доход таких «специалистов» также не вырос существенным образом. Всё это очевидным образом приводит к недовольству регуляторов в адрес оценочного сообщества. Следует ожидать дальнейших ужесточений и усугубления положения добросовестных оценщиков и их работодателей.

При этом было бы ошибочным считать, что виной всему только сами оценщики и их работодатели. Во многом проблемы квалификации и качества работы вызваны не нежеланием добросовестно выполнять свою работу, а отсутствием интереса заказчиков к серьёзной

качественной оценке. Не секрет, что во многих случаях оценка является навязанной требованиями закона либо кредитора услугой, не нужной самому заказчику, которого волнует не качество отчёта об оценке, а соответствие определённой в нём стоимости ожиданиям и потребностям заказчика, его договорённостям с контрагентами. В таких условиях, с одной стороны экономика не создаёт спрос на качественную оценку, а с другой – сами оценщики не предлагают экономике интересных решений и новых ценностей, которые могли бы принести финансовые потоки в отрасль.

Вторая проблема тесно связана с первой и выражается в т.ч. в наблюдамом падении качества отчётов об оценке и общей примитивизации работы оценщика на протяжении последних практически 10 лет. Коротко проблему можно сформулировать одной фразой: «раньше молодые оценщики спрашивали «как проанализировать данные и построить модель для оценки», сейчас остался один вопрос «где взять корректировку на «Х?». Установление метода корректировок в качестве доминирующего во всех случаях без анализа применимости других методов стало логичным итогом процесса деградации качества отчётов об оценке. Как и в первом случае винить только самих оценщиков было бы неправильным. В условиях работы в зачастую очень жёстких временных рамках, оценщик часто лишён возможности провести самостоятельный анализ тех или иных свойств открытого рынка и вынужден использовать внешнюю информацию в т.ч. и непроверенного качества. Со временем это становится привычкой.

Данное руководство призвано дать в руки оценщика инструменты, позволяющие ему легко и быстро извлекать информацию с открытых рынков, строить на её основе гипотезы, выбирать из них наиболее перспективные, и в итоге получать готовые модели предсказания различных свойств объекта оценки. В данном случае анализируется износ транспортных средств. Есть некоторая надежда, что применение технологий Data mining позволит без увеличения трудоёмкости, а скорее, напротив, снижая её, повысить качество отчётов, усилит их доказательность и как итог создаст новые ценности, предлагаемые оценщиками экономике, государству и потребителям.

Особенностью данной работы является её практическая направленность: в тексте содержатся все необходимые инструкции, формулы, описания и фрагменты программного кода, необходиые и достаточные для воспроизведения всех описанных процедур.

Важным принципом Data Mining, положенным в основу данной работы является принцип максимального извлечения информации из того набора данных, который есть в наличии.

В настоящей работе будут рассмотрены 6 (шесть) способов определения совокупного износа:

1) метод регрессионного анализа;

2) метод экспоненциальной скользящей средней;

3) метод «ленивого обучения» (lazy learning);

4) метод деревьев классификации (в варианте метода деревьев регрессии);

5) метод случайных лесов (bagging bootstrap aggregation);

6) gradient boosting machine.

Также рассматриваются вопросы факторного анализа, кластеризации, классификации и предсказательного распознавания классов.

Данное руководство основано на двух основополагающих принципах:

1) **Принцип KISS** [1] (keep it simple stupid, вариации: keep it short and simple, keep it simple and straightforward и т.д.), предложенный американским авиаинженером Келли Джонсоном, и ставший официальным принципом проектирования и конструирования ВМС США с 1960 г. Данный принцип заключается в том, что при разработке той или иной системы следует использовать самое простое решение из возможных. Применительно к тематике данной работы это означает, что в тех случаях, когда автор сталкивался с проблемой выбора способа решения задачи в условиях неопределённости преимуществ и недостатков возможных вариантов её решения, он всегда выбирал самый простой способ. Например в задаче кластеризации, выбирая между видами расстояний (евклидово, манхэттенское и др.), автор выбирает евклидово расстояние, выбирая между вариантом использования данных, имеющих распределение отличное от нормального (распределение Гаусса), и преобразования данных к нормальному распределению, происходит выбор в пользу использования исходных данных без дополнительного преобразования, если отличие распределения от нормального является возможным и допустимым для целей дальнейшего анализа.

2) **Принцип «Не дай алгоритму уничтожить здравый смысл»**. Данный принцип означает необходимость осмысления аналитиком всех результатов выполнения процедур, в т.ч. и промежуточных. Возможны ситуации, когда полученные результаты могут противоречить здравому смыслу и априорным знаниям о предметной области, которым обладает аналитик либо пользователь его работы. Следует избегать безоговорочного доверия к результатам, выдаваемым алгоритмами. Если построенная модель противоречит априорным знаниям об окружающей действительности, то следует понимать, что другой действительности у нас нет, тогда как модель можно заменить на другую.

Все описанные этапы действий описаны таким образом, что позволяют сразу же без каких-либо дополнительных исследований воспроизвести всё то, что было реализовано при подготовке руководства. От пользователей потребуется только установить необходимые программные средства, создать свой набор данных для анализа и загрузить его в пакет. Все действия по установке и настройке описаны внутри Руководства.

От пользователей данного руководства не требуется специальных познаний в области разработки ПО, software engineering и иных аспектов программирования. Некоторые понятия

вроде «класс», «метод», «функция», «оператор» и т. п. термины из области программирования могут встречаться в тексте Руководства, однако их понимание либо непонимание пользователем Руководства не оказывает сколько-нибудь существенного влияния на восприятие в целом. В отдельных случаях, когда понимание термина является существенным, как например в случае с термином «переменная», в тексте Руководства приводится подробное объяснение смысла такого термина, доступное для понимания неспециалиста.

Также от пользователей руководства не требуется (хотя и является желательным) глубокого понимания математической статистики, дифференциальных вычислений, линейной алгебры, методов исследования операций, методов оптимизации, хотя и предполагается наличие таких познаний на уровне материала, включённого в школьную программу и программу 1 курса технических и экономических специальностей вузов России. В тексте Руководства приводится описание смысла и техники всех применённых статистических методов, математических операций и вычислений в объёме, достаточном для обеспечения доказательности процедур при использовании методов из Руководства. Автор всегда приводит ссылки на материалы, подтверждающие приведённые им описания. Особое внимание автор уделяет соблюдению требований к информации, имеющей существенное значение для определения стоимости объекта оценки, установленных Федеральным законом «Об оценочной деятельности в Российской Федерации», а также Федеральными стандартами оценки №1,3,10. Информация, приведённая в Руководстве является, по мнению автора, достаточной для обеспечения выполнения вышеуказанных требований к информации, содержащейся в Отчёте об оценке. Таким образом, использование описаний процедур, выполненных в настоящем Руководстве, должно быть достаточным при использовании изложенных в нём методик в целях осуществления оценочной деятельности и составлении Отчёта об оценке. Однако, автор рекомендует уточнять требования, предъявляемые к Отчёту об оценке со стороны СРО, в котором состоит Оценщик, а также требования Заказчиков и регуляторов.

В силу свободной лицензии, на условиях которой распространяется данная работа, любая её часть или она целиком может быть скопирована, воспроизведена, переработана или использована любым другим способом любым лицом в т.ч. в коммерческих целях. Таким образом, автор рекомендует использовать тексты, приведённые в Руководстве для описания сделанных оценщиком процедур.

Данное руководство может быть полезно оценщикам, сотрудникам залоговых служб кредитных организаций, страховым и лизинговым компаниям. С целью приближения руководства к целям кредитных и лизинговых организаций в качестве объекта исследования выбран коммерческий транспорт, а исследуемый срок эксплуатации ограничен 6 (Шестью) годами, что соответствует распространённому сроку кредитования либо лизингового финансирования в 5 (Пять) лет, увеличенному на 1 (Один) год, что на практике может пригодиться в случае необходимости реализации залогового имущества, которое по каким-то

причинам было изъято в пользу Кредитора в последний год кредитования/финасирования и требует оценки износа для последующей реализации.

В данном руководстве не содержится общих выводов касательно параметров износа транспортных средств как таковых вообще, не выводятся общие формулы, применимые для всех марок и моделей. Вместо этого в распоряжение пользователей предоставляется набор инструментов, достаточный для определения износа транспортного средства, конкретной марки и модели по актуальным данным открытого рынка. В случае необходимости пользователь, применяя рассмотренные методы, может самостоятельно разработать предсказательную модель износа для групп транспортных средств.

Забегая вперёд, можно сказать, что при решении конкретной практической задачи применении всех 6 (шести) методов не является обязательным. В тексте Руководства содержаться рекомендации по выбору методов, исходя из свойств данных, рассматриваются сильные и слабые стороны каждого из них.

Несмотря на изначально кажущуюся сложность и громоздкость методов, при более детальном знакомстве и погружении в проблематику становится ясно, что применение предложенных реализаций методов существенно сокращает время, необходимое для выполнения расчёта относительно других методов сопоставимого качества, а сама процедура сводится к написанию и сохранению нескольких строк кода при первом применении и их вторичному многократному использованию для новых наборов данных при будущих исследованиях.

Все расчёты, кроме первичного ручного анализа выгруженных из ИТС Интернет данных выполнены на языке R с использованием IDE[1] R-Studio.

В дальнейших версиях данного руководства автор планирует добавить описание выполнения аналогичного анализа средствами языка Python, а также увеличить число применяемых методик, в частности использовать нейронные сети.

Автор надеется, что для некоторых пользователей данное Руководство станет первым шагом на пути изучения языка R, а также погружения в мир Data mining.

Автор выражает благодарность...

1 IDE – integrated development environment (пер. на русский «интегрированная среда разработки» либо просто «среда разработки»)

Глава 1. Технологическая основа

1.1. Почему R?

1.1.1. А почему вообще не Excel?

На сегодняшний день очевидным является факт того, что программным продуктом доминирующим в среде русских оценщиков в качестве средства для выполнения расчётов, является приложение MS Excel. Следом за ним идут его свободные и бесплатные аналоги LibreOffice Calc и OpenOffice Calc. Не оспаривая достоинств этих продуктов, нельзя не сказать о том, что они являются универсальными средствами обработки данных общего назначения и как любые универсальные средства, стремящиеся объять необъятное, не преуспевают в этом. В них в виде готовых функций реализованы многие основные математические и статистические процедуры. Также само собой присутствует возможность выполнения расчётов в виде формул, собираемых вручную из простейших операторов. Однако возможностей этих продуктов для профессионального анализа данных всё же недостаточно. В частности существуют ограничения на количество строк и столбцов, отсутствуют средства реализации многих современных методов анализа данных. Например, ни одно из вышеперечисленных приложений не позволяет решить задачи метода GBM.

Таким образом, следует признать, что оставаясь высококачественными универсальными средствами для базовых расчётов, вышеперечисленные приложения не могут быть использованы для профессионального анализа данных на современном уровне.

При этом использование их либо каких-то иных аналогичных приложений необходимо на первоначальном этапе. Исходные данные, предоставляемые аналитику для обработки, нередко содержатся в электронных таблицах. Такие таблицы помимо полезной информации могут содержать посторонние данные, тексты, графики и изображения. В практике автора был случай предоставления для анализа данных в виде электронной таблицы формата .xlsx, имеющей размер около 143 МБ, содержащей помимо подлежащей анализу числовой информации о товарах, их рекламные описания в текстовом виде и фотографии, составляющие свыше 95% объёма файла.

Просмотр исходной информации средствами табличных процессоров и создание нового файла, содержащего только необходимые для анализа данные, является подготовительным этапом процесса анализа. В последующих разделах будут данные практические рекомендации касательно реализации данного этапа. В качестве ремарки следует сказать, что по состоянию на июнь 2019 лучшим табличным процессором представляется LibreOffice Calc, превосходящий по ряду характеристик свой прототип Microsoft Excel. Можно добавить, что весь пакет LibreOffice представляется предпочтительным относительно Microsoft Office. Например данное Руководство было набрано именно в LibreOffice Writer, являющимся аналогом Microsoft Word.

1.1.2. R vs Python

По состоянию на первое полугодие 2019 года в области Data mining[2] доминирующими техническими средствами являются языки программирования R и Python. Оба они являются высокоуровневыми скриптовыми языками программирования. Высокоуровневым называется такой язык программирования, в основу которого заложена сильная «абстракция», т. е. свойство описывать данные и операции над ними таким образом, при котором разработчику не требуется глубокое понимание того, как именно машина их обрабатывает [2]. Скриптовым (сценарным) языком называется такой язык программирования, работа которого основана на использовании сценариев, т. е. программ, использующих уже готовые компоненты (например библиотеки) [3].

Оба языка распространяются на условиях свободных лицензий с незначительными отличиями. R распространяется на условиях лицензии GNU GPL, Python – на условиях лицензии Python Software Foundation Licence, являющейся совместимой с GNU GPL. Отличия между ними не имеют никакого практического значения для целей настоящего Руководства и применения любого из этих языков в оценочной деятельности в целом. Следует знать основной факт: использование этих языков является легальным и бесплатным в том числе и для коммерческих целей.

Отличие этих языков заключается в частности в том, что Python – язык общего назначения, применяемый в различных областях, тогда как R – специализированный язык для статистического анализа и Data mining. В целом можно сказать, что задачи Data mining могут одинаково успешно решаться средствами обоих языков.

К преимуществам R можно отнести, тот факт, что он изначально был разработан двумя профессиональными статистиками: Ross Ihaka, Robert Gentleman, по первым буквам имён которых и был назван. Дальнейшее развитие языка также осуществляется прежде всего силами профессиональных математиков и статистиков, что даёт преимущества в виде того, что для R реализовано значительное количество библиотек, выполняющих практически все доступные на сегодняшнем уровне развитии науки и техники статистические процедуры. Кроме того, можно быть уверенным в абсолютной корректности всех алгоритмов, реализованных в библиотеках. К тому же этот язык особенно популярен в академической среде, что означает факт того, что в случае, например выхода какой-то статьи, описывающей новый статистический метод, можно быть уверенным, что соответствующая библиотека, реализующая этот метод выйдет в ближайшее время либо уже вышла. Кроме того, важным преимуществом R являются очень хорошо проработанные средства вывода графической интерпретации результатов анализа.

2 Здесь и далее основным термином, описывающим весь процесс анализа данных будет термин Data mining. В Руководстве в разделе 1.2. Что было раньше – курица или яйцо? Соотношение понятий machine learning, data mining и artificial intelligencена стр. 20 приводится краткое описание соотношения понятий «Анализ данных», «Machine learning», «Data mining», «Artificial intelligence». Не претендуя на истинность и выбор наилучшего термина, автор всё же вводит термин Data mining в качестве основного в целях единообразия в тексте.

Недостатки R, как это часто бывает, следуют из его достоинств. Язык и его библиотеки поддерживаются в первую очередь силами математиков-статистиков, а не программистов, что приводит к тому, что язык относительно плохо оптимизирован с точки зрения software engineering, многие решения выглядят неочевидными и неоптимальными с точки зрения их обращений к памяти, интерпретации в машинные команды, исполнения на процессоре. Это приводит в высокому потреблению ресурсов машины, в первую очередь оперативной памяти, медленном исполнении команд. Хотя конечно, говоря о медленном исполнении, следует понимать относительность этой медлительности. Выполнение команды за 35 мс вместо 7 мс не замечается человеком и обычно не имеет сколько-нибудь определяющего значения. Проблемы с производительностью становятся заметны только при работе с данными большой размерности: миллионы наблюдений, тысячи переменных. В практических задачах, с которыми сталкиваются оценщики, подобная размерность данных выглядит неправдоподобной, вследствие чего можно говорить об отсутствии существенных недостатков языка R для целей применения в оценочной деятельности в целом и задачи, решаемой в данном Руководстве, в частности.

Преимуществом Python является его большая распространённость и универсальность. Освоение основ данного языка для целей одной предметной области может быть полезным в дальнейшем, если по каким-то причинам оценщик захочет решать с его помощью задачи иного класса. Данный язык разработан и поддерживается профессиональными программистами, что означает его относительно высокую оптимизацию.

К недостаткам можно отнести меньшее число библиотек, содержащих статистические процедуры. Кроме того, нет такой же уверенности в безупречности их алгоритмов. Данные библиотеки написаны не математиками-статистиками, а разработчиками ПО, что не одно и тоже. При этом следует отметить, что подобные риски присутствуют лишь в новых библиотеках, реализующих экспериментальные либо экзотические статистические процедуры. Для целей оценки как правило вполне достаточно уже относительно отработанных и проверенных библиотек.

Подводя итог, можно сказать, что нет существенной разницы, какой из упомянутых языков является предпочтительным для целей Data mining в оценке. R развивается, оптимизируется и всё больше избавляется от «детских болезней» неоптимизированности, для Python создаются новые мощные библиотеки статистического анализа. Нельзя говорить о явной предпочтительности того или иного языка. В итоге выбор для первого издания данного руководства был сделан в пользу R без какой-то явной причины. В следующей версии Руководства, планируемой к выходу в декабре 2019 года все решения будут реализованы на обоих языках.

Следует кратко упомянуть о том, что помимо R и Python для Data mining используются программные продукты такие как SAS, SPSS, Statistica, Minitab, Stata, Eviews и ряд других. Однако все они являются платными, при этом стоимость лицензии на самый мощный из них — SAS начинается, как правило от нескольких сотен тысяч долларов. В остальном кроме

привычного для большинства людей графического интерфейса они не имеют никаких преимуществ перед R и Python, предоставляя при этом даже меньше возможностей.

1.2. Что было раньше – курица или яйцо? Соотношение понятий machine learning, data mining и artificial intelligence

Данный [4]

1.3. Установка и настройка

В первую очередь необходимо установить сам пакет R. Для этого следует зайти на страницу, расположенную в ИТС Интернет по адресу: https://cran.r-project.org. Далее следует выбрать дистрибутив, соответствующий используемой операционной системе. Установка пакета R не имеет каких-то специфических особенностей и выполняется так же как установка большинства приложений. На момент написания данного Руководства актуальной версией R является версия 3.6.0 «Planting of a Tree» (релиз от 2019-04-26).

Далее рекомендуется установить IDE. Существует несколько вариантов: RStudio, JGR, RK Ward, SciViews-R, Statistical Lab, R Commander, Rattle и т. д. Автор рекомендует установку RStudio. На момент написания Руководства актуальной версией является 1.2.1135 (релиз от 2019-04-08). Установка также не требует никаких специфических действий.

Как R, так и RStudio имеют дистрибутивы для различных операционных систем: MS Windows, OS X (Mac), Linux: имеются готовые реализации для Debian, Ubuntu, OpenSuse, Redhat/Fedora. Как R, так и RStudio (в некоммерческой версии) представляют собой OpenSource продукты.

Рекомендуется сначала установить сам пакет R, а затем IDE. В этом случае произойдёт автоматическое подключение среды R при первом запуске RStudio.

1.4. Каллиграфия Data mining. Краткое стилевое руководство по R

Как известно, код читают чаще, чем пишут. Чтобы код был читаемым следует придерживаться рекомендаций т. н. стилевых гидов. В данном разделе приводится список основных рекомендаций по оформлению кода на R. Более подробные рекомендации можно найти в соответствующих руководствах: [5], [6], [7], [8]. Ниже приводятся наиболее важные по мнению автора.

1) **Оператор присваивания**. В качестве оператора присваивания следует использовать `<-` , а не привычный =.

2) **Пробелы**. Пробелы следует ставить вокруг операторов (кроме квадратных скобок), а также перед открывающейся скобкой. Пробелы всегда следует ставить после запятой.

✔ `average <- mean (feet / 12 + inches, na.rm = TRUE)` – правильно;
✗ `average<-mean(feet / 12 + inches,na.rm = TRUE)`- неправильно.

3) **Имена функций и переменных**. Имена функция и переменных следует записывать только строчныи буквами с разделением точками внутри. Например: period.separation.

4) **Фигурные скобки**. После открывающей фигурной скобки должна следовать новая строка, а закрывающая находиться на отдельной строке, если только за ней не следует else. Пример:

```
if (x >= 0) {
log(x)
} else {
message("Not applicable!")
}
```

5) **Длина строки.** Рекомендуемая длина — 80 символов. Перенос следует осуществлять сочетанием Shift+Enter.

6) **Отступ строки**. Осуществляется 2 пробелами, не табуляцией.

7) **Имена файлов**. Имена файлов скриптов и путь к ним должны содержать только латинские буквы, цифры и специальные символы. Использование пробелов не допускается. Использование кириллических символов не допускается. Имена файлов должны быть информативными и иметь расширение .R.

✔ O:\Methodics\My\Publications\Iznos_TS\R\Gazelle_iznos.R – правильно.

✗ O:\Методики\My\Статьи\Iznos_TS\R\Газель.R – неправильно.

✗ O:\Methodics\My\Publications\Iznos TS\R\Gazelle iznos.R – неправильно.

✗ O:\AAA\BBB\bla-bla-bla\Publications\Iznos_TS\R\111111.R – неправильно.

8) **Комментарии к коду.** Перед комментариями следует ставить знак #.

9) **Дополнительные рекомендации Автора**. С целью улучшения эргономических показателей при разработке кода автор рекомендует использовать следующие настройки IDE: шрифт – Source Code Pro, размер шрифта — не меньше 11 либо 12 в зависимости от диагонали монитора, тема – «Cobalt» либо любая иная с инвертированной цветовой схемой и пониженной контрастностью. Данные настройки выполняются путём выполнения действий из графического меню Rstudio: Tools – Global Options – Appearance. Далее следует изменить значение параметров Editor Font, Editor fonts size, Editor theme соответственно. Данные рекомендации не являются обязательными, но основаны на личном опыте Автора и рекомендациях многих разработчиков кода и специалистов в области Data mining.

1.5. Ещё что-то про R

Глава 2. Формулировка задачи и исходная информация

2.1. Цель и предмет исследования

Целью данной работы является демонстрация возможностей методов Data Mining в оценочной деятельности на примере решения задачи определения совокупного износа траснпортного средсттва. Предметом исследования является износ транспортного средства конкретной марки и модели. В данной работе не рассматривается подробно суть такого явления как износ. Не описывается соотношение понятий физический, функциональный, внешний и совокупный износ. Скажем только, что для целей настоящей работы под совокупным износом понимается количественная мера потери полезности, нарастающая по мере осуществления процесса эксплуатации объекта вследствие влияния на него всех факторов, существующих на открытом рынке. На практике, это означает, что мерой износа является количественная мера потери части стоимости относительно стоимости аналогичного объекта в новом состоянии.

Поскольку данное Руководство предназначено в первую очередь для оценщиков и сотрудников кредитных, страховых и лизинговых организаций, предполагается наличие определённых знаний сути явления износа, его видов и их соотношений. Автор рекомендует использовать следующие материалы, позволяющие изучить данный вопрос тем, кто с ним пока не знаком, и улучшить своё понимание проблематики тем, кто уже обладает некоторыми знаниями в сфере оценки машин, оборудования и транспортных средств. Курсивом выделен комментарий Автора к каждому из перечисленных материалов.

1. «Оценка стоимости машин оборудования и транспортных средств». Под ред. А.Г. Грязновой, Д.С. Львова, М.А. Федотовой. [9]. *Классический учебник, ставший одним из первых в своём роде. Написан и отредактирован большим коллективом авторитетных авторов из академической и преподавательской среды. В нём обобщён существовавший на момент написания в первую очередь международный опыт оценки активов класса «машины, оборудование и ТС». Рассмотрены общие вопросы износа, предложены несколько методов его определения. Несмотря на некоторое устаревание, вызванное неоднократным изменением законодательства, как в сфере оценочной деятельности, так и в сфере гражданского права вообще, благодаря исключительно высокой квалификации авторов, и сейчас, спустя 16 лет, остаётся хорошим пособием для тех, кто начинает изучать вопросы оценки.*

2. «Оценка стоимости машин и оборудования». Под ред. В.П. Антонова. [10]. *Ещё один ставший классикой учебник. Содержит интересные примеры из различных эпох и отраслей, а также некоторые рекомендуемые числовые значения. Как и в случае с первым, однозначно рекомендуется к изучению.*

3. «Основы оценки стоимости машин и оборудования». Под ред. М.А. Федотовой. [11]. *Данное учебное пособие развивает идеи авторов учебника, упомянутого в п. 1. Составлено примерно тем же авторским коллективом. Имеет б**о**льшую практическую направленность.*

4. «Оценка для целей залога: теория, практика, рекомендации». М.А. Федотова. [12]. *Также является более практическим пособием, чем классические учебники. Особенно полезен для сотрудников кредитных и лизинговых организаций, содержит некоторые авторские методики оценки.*

5. «Оценка стоимости подвижного состава автомобильного транспорта». Ю.В. Андрианов. [13]. *Классический учебник по оценке именно транспортных средств. Содержит идею о возможности экспоненциального характера износа, являющуюся важной предпосылкой данного Руководства. Несмотря на устаревание численных значений показателей износа, остаётся одним из лучших пособий для введения в курс специфики оценки транспортных средств.*

6. «Проблемы и парадоксы оценки машин и оборудования». С.А. Смоляк. [14]. *По мнению автора данного Руководства данная книга является одним из самых интересных исследований в оценочной деятельности России за всю историю её существования. Автор, будучи профессиональным математиком и сотрудников ЦЭМИ РАН, использует серьёзный математический аппарат, позволяющй получить очень интересные и ранее не описанные свойства стоимости. В частности исследуется связь между доходностью, полезностью, стоимостью, износом. Принимая метод дисконтированных денежных поток за основу определения базы полезности, автор решает очень интересные задачи, в частности доказывает связь межу ставкой налога на прибыль и мерой приращения износом. Для целей настоящего Руководства особенно полезным является строгое и убедительное доказательство уже упомянутого ранее экспоненциального характера износа транспортных средств. Данная книга достаточно сложна для восприятия, требует знания математического аппарата, развитого абстрактного мышления и понимания сути ряда операций в экономике. Рекомендуется тем, кто уже достаточно владеет категорийным аппаратом, имеет практический опыт в оценке и хочет серьёзно углубить свои знания в данной области, а также области исследования операций [15], [16] как такового. Возможно нам ещё только предстоит в будущем осознать идеи С.А. Смоляка в полной мере.*

7. «Положение о единой методике определения размера расходов на восстановительный ремонт в отношении повреждённого транспортного средства». ЦБ РФ. [17]. *Является нормативным документов, описывает методы оценки и в т.ч. определение износа для специфических целей института ОСАГО. В данном документе также признаётся в качестве основной предпосылки характера совокупного износа — экспоненциальный характер.*

8. «Оценка стоимости транспортных средств». Андрианов Ю.В. [18]. *Представляет собой учебно-методическое пособие, дающее широкий набор практически приёмом оценки транспортных средств. Также содержит сведения, указывающие на экспоненциальный характер процесса износа, как его наиболее вероятную аппроксимацию.*

9. «Методологические основы оценки стоимости имущества». Под ред. М.А. Федотовой. [19]. *Данное пособие не имеет непосредственного отношения к проблематике Руководства, является устаревшим с точки зрения описываемых стандартов. Однако автор всё же рекомендует его в качестве удачного примера самостоятельного осмысления авторами всей существующей международной системы оценки стоимости активов.*

10. «Методические рекомендации по проведению судебных автотехнических экспертиз и исследований колёсных транспортных средств в целях определения размера ущерба, стоимости восстановительного ремонта и оценки». [20]. *Данный материал интересен и полезен тем, что во-первых, является современным и основан на последних обобщениях достижений в области судебной экспертизы в России, во-вторых, он содержит практические рекомендации и требования, предъявляемые пользователями судебных стоимостных экспертиз — судами, органами предварительного расследования и т. д. Не секрет, что помимо Отчётов об оценке оценщики часто готовят заключения судебных экспертиз. Изучение данного пособия даст ответы на то, какая информация и какая её подача обеспечивают доказательность заключения.*

11. «Оценка стоимости транспортных средств». Под ред. М.П. Улицкого. [21]. *Данное издание представляет собой специализированное пособие по оценке различных видов транспортных средств: наземных, воздушных, водных. Его полезность и значение заключаются в частности в том, что вопросы оценки двух последних категорий описаны и проработаны достаточно слабо, в отношении них, в отличие от наземного транспорта, почти нет методик и описаний способ оценки. Данное пособие частично решает проблему информационного вакуума.*

12. Методические указания «Оценка стоимости в отношении транспортных средств» . Комитет по стандартам и методологии в оценочной деятельности НСОД. Рук. раб. группы Табакова С.А. [22]. *Данный документ был разработан практикующими оценщиками. В нём приводится обобщение большей части знаний оценщиков об оценке транспортных средства, в т.ч. о природе процесса их износа.*

13. «Оценка стоимости машин и оборудования». Н.В. Вейг. [23]. *Ещё один учебник общего назначения. Является самым кратким из перечисленных, содержит сжатую информацию. Кроме того автор пособия, являясь жителем Санкт-Петербурга, не мог не привести хотя бы одну книгу, изданную в Санкт-Петербурге, на фоне абсолютного преобладания московских изданий.*

14. «Износ технологических машин и оборудования при оценке их рыночной стоимости». А.И. Попеско. [24]. *Данное издание содержит серьёзное описание сути понятия износ, его видов и их влияния на стоимость.*

15. Статья «Модифицированный метод сроков жизни для расчёта износа оборудования». Мышанов А.И., Рослов В.Ю. [25]. *Авторы приводят интересные доводы в пользу связи износа с доходностью, приводят соответствующие расчёты. Несмотря на некоторое упрощение в части построения общей регрессионной модели износа на основе очень разнородных данных, не соотносимых между собой, данная работа является очень ценным примером применения методов исследования операций [15], [16] к определению износа. В статье сделан вывод об экспоненциальном характере износа машин, оборудования и транспортных средств.*

16. Статья «Определение остаточного срока службы машин и оборудования на основе вероятностных моделей». Лейфер Л.А., Кашникова П.М. [26]. *Ценность данной статьи заключается в первую очередь в том, что автор на основе теоретической информации делает практические выводы о связи нормативного и фактического предельного срока эксплуатации. Рекомендуется к прочтению при наличии соответствующей подготовки.*

17. Статья «Эргодические модели износа машин и оборудования». С.А. Смоляк. [27]. *В краткой форме автор излагает основные понятия связи между доходностью, полезностью и износом. Рекомендуется к прочтению при наличии соответствующей подготовки.*

18. Статья «Определение износа при оценке машин и оборудования». А.П. Ковалёв. [28]. *В статье авторы приходят к выводу о логистическом характере модели износа. Данная гипотеза будет рассмотрена в настоящем Руководстве.*

19. Статья «Определение срока службы машин и оборудования при их стоимостной оценке». А.П. Ковалёв. [29]. *В статье рассматриваются вопросы износа с точки зрения срока службы объекта.*

20. Статья «Функциональное устаревание машин и оборудования: как учесть его при оценке». А.П. Ковалёв, Д. Игонин. [30].

Далее Автор вводит несколько предпосылок, на которых будет основано исследование, без подробного описания в тексте Руководства. Выше приведён исчерпывающий список методических материалов, в которых можно найти описание идей автора:

- наиболее вероятно, что характер прироста износа имеет экспоненциальный характер;
- совокупный износ возникает вследствие мультипликативного сочетания физического, функционального и внешнего износов;
- информация о ценах на отрытом рынке уже учитывает влияние всех факторов.

2.2. Исходная информация

При выборе модели автомобиля, в отношении которой будет проводиться дальнейший анализ выбор пал на модель ГАЗ-А23R32 (Газель Next кузов промтоварный фургон, цельнометаллический фургон). Выбор обоснован тем, что для ряда потенциальных пользователей Руководства, например для сотрудников лизинговых компаний, наибольший интерес представляет именно рынок коммерческого транспорта. Конкретная модель выбрана по признаку того, что с начала её выпуска прошло 6 лет, а их производство продолжается, что означает, что возраст таких машин находится в диапазон от 0 до 6 лет, что соответствует типичному максимальному сроку кредитования/финансирования, увеличенному на 1 год.

За основу была взята информация с сайта в ИТС Интернет, расположенного по адресу: cars.auto.ru. Было найдено 335 предложения о продаже автомобилей с пробегом и 119 — о продаже новых. Ниже приводятся исходные данные.

Первая строка таблиц 1 и 2 содержит описание данных, содержащихся в столбце, вторая — название переменной, используемой при дальнейшем анализе данных в среде R.

Таблица 1: Сведения о предлагаемых к продаже автомобилях ГАЗ-А23R32 (Газель Next кузов промтоварный фургон) с пробегом

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
1	870.0000	2016	3.0000	236.0000	78.6667	Ростов-на-Дону
2	750.0000	2016	3.0000	110.0000	36.6667	Москва
3	1200.0000	2018	1.0000	34.0000	34.0000	Ижевск
4	749.0000	2017	2.0000	78.0000	39.0000	Москва
5	649.0000	2016	3.0000	132.0000	44.0000	Москва
6	1150.0000	2017	2.0000	65.0000	32.5000	Пермь
7	950.0000	2017	2.0000	51.0000	25.5000	Пермь
8	550.0000	2013	6.0000	85.0000	14.1667	Москва
9	790.0000	2015	4.0000	160.0000	40.0000	Москва
10	580.0000	2015	4.0000	123.0000	30.7500	Москва
11	960.0000	2017	2.0000	78.0000	39.0000	Москва
12	740.0000	2016	3.0000	93.0000	31.0000	Москва
13	950.0000	2017	2.0000	150.0000	75.0000	Брянск
14	1200.0000	2017	2.0000	100.0000	50.0000	Москва
15	860.0000	2016	3.0000	80.3030	26.7677	Москва
16	900.0000	2016	3.0000	27.0000	9.0000	Москва

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
17	600.0000	2016	3.0000	148.9390	49.6463	Москва
18	900.0000	2014	5.0000	140.0000	28.0000	Екатеринбург
19	1150.0000	2017	2.0000	34.0000	17.0000	Москва
20	625.0000	2014	5.0000	100.0000	20.0000	Нижний Новгород
21	855.0000	2016	3.0000	65.0000	21.6667	Нижний Новгород
22	625.0000	2014	5.0000	138.0000	27.6000	Москва
23	800.0000	2016	3.0000	88.0000	29.3333	Москва
24	680.0000	2014	5.0000	74.0000	14.8000	Волгоград
25	1200.0000	2017	2.0000	63.4000	31.7000	Москва
26	750.0000	2015	4.0000	139.0000	34.7500	Краснодар
27	735.0000	2014	5.0000	60.4850	12.0970	Москва
28	560.0000	2013	6.0000	112.0000	18.6667	Москва
29	890.0000	2015	4.0000	52.6520	13.1630	Москва
30	960.0000	2016	3.0000	53.6640	17.8880	Москва
31	888.0000	2017	2.0000	31.5000	15.7500	Москва
32	650.0000	2013	6.0000	89.9000	14.9833	Москва
33	525.0000	2013	6.0000	117.0000	19.5000	Пермь
34	870.0000	2017	2.0000	42.0000	21.0000	Москва
35	650.0000	2014	5.0000	85.0000	17.0000	Москва
36	635.0000	2013	6.0000	380.0000	63.3333	Ярославль
37	1100.0000	2013	6.0000	4.7000	0.7833	Ярославль
38	1200.0000	2016	3.0000	47.3000	15.7667	Краснодар
39	1400.0000	2017	2.0000	220.0000	110.0000	Уфа
40	875.0000	2016	3.0000	67.0000	22.3333	Москва
41	685.0000	2014	5.0000	33.0000	6.6000	Казань
42	645.0080	2014	5.0000	139.0000	27.8000	Москва
43	680.0000	2013	6.0000	125.0000	20.8333	Москва
44	1290.0000	2018	1.0000	0.5160	0.5160	Нижний Новгород
45	559.0000	2013	6.0000	132.0000	22.0000	Москва
46	550.0000	2014	5.0000	213.3970	42.6794	Ростов-на-Дону
47	480.0000	2013	6.0000	90.0000	15.0000	Москва
48	665.0000	2014	5.0000	149.0000	29.8000	Москва
49	950.0000	2017	2.0000	22.0000	11.0000	Санкт-Петербург
50	890.0000	2016	3.0000	13.7710	4.5903	Москва

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
51	650.0000	2014	5.0000	130.0000	26.0000	Кострома
52	620.0000	2015	4.0000	240.0000	60.0000	Москва
53	720.0000	2015	4.0000	130.0000	32.5000	Тула
54	920.0000	2016	3.0000	76.6810	25.5603	Москва
55	685.0000	2016	3.0000	87.0000	29.0000	Москва
56	650.0000	2016	3.0000	129.6000	43.2000	Москва
57	677.0000	2013	6.0000	190.0000	31.6667	Тула
58	760.0000	2017	2.0000	55.0000	27.5000	Москва
59	1300.0000	2016	3.0000	10.5000	3.5000	Москва
60	650.0000	2013	6.0000	100.0000	16.6667	Москва
61	1220.0000	2017	2.0000	50.0000	25.0000	Москва
62	759.0000	2016	3.0000	75.0000	25.0000	Новгород
63	670.0000	2013	6.0000	180.0000	30.0000	Москва
64	695.0000	2016	3.0000	153.0000	51.0000	Тверь
65	650.0000	2015	4.0000	151.0000	37.7500	Тверь
66	1100.0000	2016	3.0000	20.0000	6.6667	Екатеринбург
67	1170.0000	2018	1.0000	18.0000	18.0000	Москва
68	999.9990	2017	2.0000	49.9990	24.9995	Москва
69	999.9990	2017	2.0000	51.1500	25.5750	Санкт-Петербург
70	980.0000	2016	3.0000	105.0000	35.0000	Саранск
71	1299.0000	2018	1.0000	8.0000	8.0000	Челябинск
72	655.0000	2014	5.0000	128.0000	25.6000	Москва
73	765.0000	2015	4.0000	250.0000	62.5000	Курск
74	480.0000	2013	6.0000	234.0000	39.0000	Ставрополь
75	645.0000	2014	5.0000	148.0000	29.6000	Москва
76	750.0000	2016	3.0000	120.0000	40.0000	Санкт-Петербург
77	1650.0000	2018	1.0000	70.0000	70.0000	Самара
78	999.9990	2018	1.0000	95.0000	95.0000	Москва
79	850.0000	2016	3.0000	32.0000	10.6667	Ярославль
80	1175.0000	2017	2.0000	29.7820	14.8910	Ростов-на-Дону
81	650.0000	2014	5.0000	212.0000	42.4000	Орёл
82	518.0000	2014	5.0000	33.9080	6.7816	Ростов-на-Дону
83	599.0000	2013	6.0000	2.5120	0.4187	Ростов-на-Дону
84	575.0000	2014	5.0000	38.4250	7.6850	Ростов-на-Дону

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
85	870.0000	2016	3.0000	236.0000	78.6667	Ростов-на-Дону
86	750.0000	2013	6.0000	45.7170	7.6195	Москва
87	1245.0000	2018	1.0000	12.0000	12.0000	Нижний Новгород
88	529.0000	2014	5.0000	37.2430	7.4486	Ростов-на-Дону
89	526.0000	2013	6.0000	19.8890	3.3148	Ростов-на-Дону
90	525.0000	2013	6.0000	2.9390	0.4898	Ростов-на-Дону
91	550.0000	2014	5.0000	54.3500	10.8700	Ростов-на-Дону
92	524.0000	2013	6.0000	51.1940	8.5323	Ростов-на-Дону
93	525.0000	2013	6.0000	37.9140	6.3190	Ростов-на-Дону
94	625.0000	2015	4.0000	145.0000	36.2500	Москва
95	850.0000	2014	5.0000	70.0000	14.0000	Екатеринбург
96	1294.4440	2018	1.0000	38.0000	38.0000	Тюмень
97	1000.0500	2017	2.0000	79.6080	39.8040	Москва
98	1050.0000	2018	1.0000	20.7780	20.7780	Краснодар
99	1030.0000	2018	1.0000	17.6570	17.6570	Краснодар
100	895.0000	2015	4.0000	120.0000	30.0000	Москва
101	1000.0000	2016	3.0000	88.0000	29.3333	Москва
102	998.0000	2017	2.0000	95.0000	47.5000	Нижний Новгород
103	870.0000	2017	2.0000	60.0000	30.0000	Москва
104	1100.0000	2018	1.0000	36.0000	36.0000	Уфа
105	950.0000	2017	2.0000	22.0000	11.0000	Казань
106	815.0000	2015	4.0000	106.0000	26.5000	Москва
107	800.0000	2017	2.0000	37.0000	18.5000	Москва
108	1200.0000	2016	3.0000	182.0000	60.6667	Новосибирск
109	700.0000	2017	2.0000	130.0000	65.0000	Москва
110	950.0000	2017	2.0000	90.0000	45.0000	Москва
111	1100.0000	2018	1.0000	20.0000	20.0000	Самара
112	1000.0000	2017	2.0000	19.7000	9.8500	Москва
113	1206.0000	2018	1.0000	2.1820	2.1820	Санкт-Петербург
114	1020.0000	2018	1.0000	24.0000	24.0000	Москва
115	950.0000	2017	2.0000	71.0000	35.5000	Ярославль
116	1000.0000	2017	2.0000	115.0000	57.5000	Москва
117	999.0000	2016	3.0000	150.8760	50.2920	Краснодар
118	1200.0000	2016	3.0000	15.0000	5.0000	Белгород

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
119	850.0000	2016	3.0000	142.2000	47.4000	Владимир
120	630.0000	2014	5.0000	116.0000	23.2000	Ижевск
121	680.0000	2014	5.0000	130.0000	26.0000	Красноярск
122	900.0000	2017	2.0000	36.0000	18.0000	Москва
123	1070.0000	2017	2.0000	67.0000	33.5000	Санкт-Петербург
124	1750.0000	2015	4.0000	9.0000	2.2500	Ставрополь
125	1200.0000	2017	2.0000	31.0000	15.5000	Псков
126	515.0000	2013	6.0000	194.0000	32.3333	Москва
127	630.0000	2014	5.0000	123.0000	24.6000	Москва
128	1100.0000	2017	2.0000	35.0000	17.5000	Москва
129	720.0000	2016	3.0000	130.0000	43.3333	Москва
130	708.6000	2017	2.0000	153.8430	76.9215	Санкт-Петербург
131	1150.0000	2018	1.0000	17.6500	17.6500	Москва
132	1090.0000	2016	3.0000	14.7000	4.9000	Казань
133	1050.0000	2016	3.0000	12.0000	4.0000	Вологда
134	1090.0000	2017	2.0000	65.0000	32.5000	Краснодар
135	630.0000	2013	6.0000	60.0000	10.0000	Ростов-на-Дону
136	1050.0000	2017	2.0000	38.0000	19.0000	Санкт-Петербург
137	850.0000	2018	1.0000	125.0000	125.0000	Москва
138	660.0000	2014	5.0000	155.0000	31.0000	Пермь
139	620.0000	2014	5.0000	92.0000	18.4000	Москва
140	1200.0000	2018	1.0000	27.0000	27.0000	Москва
141	777.4410	2017	2.0000	53.3740	26.6870	Москва
142	730.0000	2013	6.0000	150.0000	25.0000	Ростов-на-Дону
143	1050.0000	2017	2.0000	39.7000	19.8500	Москва
144	850.0000	2016	3.0000	99.2000	33.0667	Санкт-Петербург
145	710.0000	2016	3.0000	55.1290	18.3763	Москва
146	970.0000	2017	2.0000	105.8000	52.9000	Москва
147	1150.0000	2019	0.0000	2.7070	16.2420	Москва
148	579.0000	2015	4.0000	106.4580	26.6145	Москва
149	950.0000	2014	5.0000	85.0000	17.0000	Воронеж
150	450.0000	2013	6.0000	200.0000	33.3333	Вологда
151	1530.0000	2018	1.0000	12.0000	12.0000	Екатеринбург
152	600.0000	2013	6.0000	39.0000	6.5000	Санкт-Петербург

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
153	750.0000	2014	5.0000	49.0000	9.8000	Тула
154	500.0000	2013	6.0000	60.0000	10.0000	Ярославль
155	980.0000	2016	3.0000	315.0000	105.0000	Чебоксары
156	965.0000	2017	2.0000	41.0000	20.5000	Москва
157	1200.0000	2016	3.0000	41.0000	13.6667	Ставрополь
158	670.0000	2013	6.0000	75.0000	12.5000	Москва
159	550.0000	2014	5.0000	75.0000	15.0000	Санкт-Петербург
160	1060.0000	2017	2.0000	123.0000	61.5000	Челябинск
161	850.0000	2016	3.0000	92.0000	30.6667	Рязань
162	720.0000	2013	6.0000	180.0000	30.0000	Екатеринбург
163	970.0000	2015	4.0000	70.0000	17.5000	Калуга
164	1140.0000	2018	1.0000	31.9110	31.9110	Нижний Новгород
165	1180.0000	2017	2.0000	82.0000	41.0000	Орёл
166	1340.0000	2017	2.0000	12.0000	6.0000	Санкт-Петербург
167	830.0000	2015	4.0000	62.0000	15.5000	Киров
168	1280.0000	2018	1.0000	22.0000	22.0000	Уфа
169	880.0000	2017	2.0000	75.0000	37.5000	Тюмень
170	1200.0000	2017	2.0000	35.0000	17.5000	Нижний Новгород
171	1050.0000	2018	1.0000	59.4730	59.4730	Москва
172	780.0000	2016	3.0000	63.7450	21.2483	Москва
173	525.0000	2013	6.0000	77.0000	12.8333	Москва
174	720.0000	2013	6.0000	200.0000	33.3333	Москва
175	710.0000	2014	5.0000	49.0000	9.8000	Москва
176	635.0000	2015	4.0000	160.0000	40.0000	Брянск
177	440.0000	2013	6.0000	350.0000	58.3333	Краснодар
178	660.0000	2014	5.0000	124.0000	24.8000	Санкт-Петербург
179	1080.0000	2018	1.0000	35.0000	35.0000	Казань
180	1000.0000	2016	3.0000	85.0000	28.3333	Курск
181	1250.0000	2018	1.0000	24.0000	24.0000	Москва
182	650.0000	2015	4.0000	150.0000	37.5000	Ростов-на-Дону
183	665.0000	2013	6.0000	220.0000	36.6667	Екатеринбург
184	799.9900	2017	2.0000	82.3000	41.1500	Санкт-Петербург
185	530.0000	2013	6.0000	115.0000	19.1667	Москва
186	1140.0000	2018	1.0000	20.0000	20.0000	Москва

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
187	900.0000	2017	2.0000	84.1000	42.0500	Санкт-Петербург
188	730.0000	2015	4.0000	120.0000	30.0000	Москва
189	940.0000	2017	2.0000	31.0000	15.5000	Москва
190	1320.0000	2015	4.0000	2.5220	0.6305	Москва
191	740.0000	2015	4.0000	178.0000	44.5000	Курск
192	1020.0000	2017	2.0000	55.0000	27.5000	Санкт-Петербург
193	750.0000	2013	6.0000	64.0000	10.6667	Казань
194	620.0000	2014	5.0000	73.0000	14.6000	Москва
195	620.0000	2015	4.0000	170.0000	42.5000	Ставрополь
196	910.0000	2018	1.0000	25.0000	25.0000	Москва
197	820.0000	2017	2.0000	100.0000	50.0000	Уфа
198	1300.0000	2017	2.0000	15.0000	7.5000	Ярославль
199	710.0000	2015	4.0000	102.0000	25.5000	Москва
200	990.0000	2017	2.0000	120.0000	60.0000	Самара
201	540.0000	2013	6.0000	238.0000	39.6667	Красноярск
202	1050.0000	2013	6.0000	132.0000	22.0000	Краснодар
203	760.0000	2016	3.0000	114.0000	38.0000	Москва
204	670.0000	2013	6.0000	190.0000	31.6667	Екатеринбург
205	1200.0000	2017	2.0000	35.0000	17.5000	Санкт-Петербург
206	1100.0000	2016	3.0000	92.0000	30.6667	Москва
207	800.0000	2015	4.0000	274.5720	68.6430	Иваново
208	1200.0000	2018	1.0000	15.6000	15.6000	Екатеринбург
209	670.0000	2014	5.0000	50.0000	10.0000	Волгоград
210	920.0000	2017	2.0000	159.9000	79.9500	Москва
211	750.0000	2014	5.0000	200.0000	40.0000	Пенза
212	970.0000	2016	3.0000	85.0000	28.3333	Пермь
213	640.0000	2013	6.0000	80.0000	13.3333	Москва
214	1200.0000	2018	1.0000	40.0000	40.0000	Москва
215	650.0000	2013	6.0000	85.0000	14.1667	Москва
216	500.0000	2014	5.0000	138.0000	27.6000	Краснодар
217	820.0000	2016	3.0000	70.0000	23.3333	Челябинск
218	670.0000	2014	5.0000	160.0000	32.0000	Санкт-Петербург
219	1100.0000	2018	1.0000	150.0000	150.0000	Ростов-на-Дону
220	870.0000	2016	3.0000	73.0000	24.3333	Чебоксары

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
221	750.0000	2013	6.0000	120.0000	20.0000	Москва
222	590.0000	2014	5.0000	120.0000	24.0000	Владимир
223	750.0000	2016	3.0000	110.0000	36.6667	Чебоксары
224	750.0000	2016	3.0000	80.0000	26.6667	Чебоксары
225	650.0000	2014	5.0000	136.0000	27.2000	Москва
226	1331.5830	2016	3.0000	0.1800	0.0600	Москва
227	999.0000	2017	2.0000	55.0000	27.5000	Нижний Новгород
228	600.0000	2014	5.0000	200.0000	40.0000	Нижний Новгород
229	1273.9880	2016	3.0000	0.0950	0.0317	Москва
230	1150.0000	2018	1.0000	13.0000	13.0000	Москва
231	489.0000	2013	6.0000	167.0000	27.8333	Санкт-Петербург
232	1100.0000	2017	2.0000	0.1070	0.0535	Тверь
233	575.0000	2014	5.0000	93.0000	18.6000	Нижний Новгород
234	1070.0000	2016	3.0000	136.0000	45.3333	Курск
235	850.0000	2017	2.0000	75.0000	37.5000	Волгоград
236	1200.0000	2018	1.0000	30.0000	30.0000	Пермь
237	890.0000	2013	6.0000	23.0000	3.8333	Томск
238	810.0000	2017	2.0000	120.0000	60.0000	Челябинск
239	995.0000	2014	5.0000	14.4440	2.8888	Ростов-на-Дону
240	580.0000	2013	6.0000	220.0000	36.6667	Москва
241	1600.0000	2017	2.0000	9.9000	4.9500	Хабаровск
242	620.0000	2015	4.0000	125.0000	31.2500	Москва
243	1031.8930	2015	4.0000	0.1200	0.0300	Москва
244	815.0000	2016	3.0000	104.0000	34.6667	Москва
245	579.0000	2013	6.0000	200.0000	33.3333	Москва
246	900.0000	2016	3.0000	130.0000	43.3333	Москва
247	629.0000	2015	4.0000	95.0080	23.7520	Нижний Новгород
248	850.0000	2017	2.0000	50.0030	25.0015	Воронеж
249	950.0000	2017	2.0000	150.0000	75.0000	Краснодар
250	1000.0000	2017	2.0000	13.0000	6.5000	Санкт-Петербург
251	1100.0000	2016	3.0000	37.2000	12.4000	Нижний Новгород
252	775.0000	2014	5.0000	123.0000	24.6000	Нижний Новгород
253	1450.0000	2017	2.0000	37.0000	18.5000	Салехард
254	975.0000	2017	2.0000	33.0000	16.5000	Санкт-Петербург

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
255	1190.0000	2017	2.0000	11.5000	5.7500	Москва
256	1510.0000	2018	1.0000	6.0000	6.0000	Москва
257	1200.0000	2018	1.0000	10.2000	10.2000	Владимир
258	900.0000	2017	2.0000	200.0000	100.0000	Краснодар
259	860.0000	2017	2.0000	67.0000	33.5000	Москва
260	750.0000	2016	3.0000	100.0000	33.3333	Москва
261	1000.0000	2018	1.0000	48.0000	48.0000	Москва
262	1100.0000	2017	2.0000	83.0000	41.5000	Москва
263	1100.0000	2017	2.0000	23.0000	11.5000	Салехард
264	750.0000	2016	3.0000	220.0000	73.3333	Воронеж
265	770.0000	2014	5.0000	160.7070	32.1414	Нижний Новгород
266	980.0000	2016	3.0000	70.0000	23.3333	Липецк
267	1350.0000	2018	1.0000	4.0000	4.0000	Уфа
268	750.0000	2015	4.0000	64.0000	16.0000	Краснодар
269	1050.0000	2018	1.0000	61.0000	61.0000	Вологда
270	950.0000	2016	3.0000	40.0000	13.3333	Нижний Новгород
271	965.0000	2017	2.0000	67.0000	33.5000	Москва
272	1100.0000	2017	2.0000	58.8580	29.4290	Владимир
273	800.0000	2016	3.0000	112.6000	37.5333	Ставрополь
274	550.0000	2014	5.0000	145.0000	29.0000	Москва
275	695.0000	2014	5.0000	200.0000	40.0000	Москва
276	1050.0000	2017	2.0000	60.0000	30.0000	Липецк
277	510.0000	2013	6.0000	300.0000	50.0000	Смоленск
278	500.0000	2013	6.0000	103.0000	17.1667	Мурманск
279	970.0000	2017	2.0000	42.3330	21.1665	Москва
280	950.0000	2014	5.0000	161.0000	32.2000	Краснодар
281	575.0000	2013	6.0000	130.0000	21.6667	Москва
282	1150.0000	2018	1.0000	31.0000	31.0000	Брянск
283	1050.0000	2018	1.0000	31.0000	31.0000	Санкт-Петербург
284	1150.0000	2016	3.0000	65.0000	21.6667	Астрахань
285	850.0000	2016	3.0000	200.0000	66.6667	Пермь
286	555.5550	2014	5.0000	170.0000	34.0000	Москва
287	920.0000	2016	3.0000	50.0000	16.6667	Волгоград
288	540.0000	2013	6.0000	140.0000	23.3333	Салехард

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
289	600.0000	2016	3.0000	350.0000	116.6667	Екатеринбург
290	795.0000	2017	2.0000	200.0000	100.0000	Москва
291	1100.0000	2017	2.0000	4.0000	2.0000	Москва
292	650.0000	2015	4.0000	146.5000	36.6250	Москва
293	720.0000	2014	5.0000	80.0000	16.0000	Сыктывкар
294	950.0000	2015	4.0000	125.0000	31.2500	Ростов-на-Дону
295	950.0000	2016	3.0000	65.0000	21.6667	Томск
296	750.0000	2016	3.0000	110.0000	36.6667	Москва
297	1200.0000	2018	1.0000	30.0000	30.0000	Уфа
298	950.0000	2017	2.0000	69.0000	34.5000	Брянск
299	770.0000	2013	6.0000	130.0000	21.6667	Салехард
300	1090.0000	2016	3.0000	85.0000	28.3333	Новосибирск
301	880.0000	2017	2.0000	50.5000	25.2500	Нальчик
302	750.0000	2015	4.0000	134.0000	33.5000	Смоленск
303	1135.0000	2017	2.0000	44.0000	22.0000	Калуга
304	780.0000	2014	5.0000	170.0000	34.0000	Москва
305	910.0000	2018	1.0000	28.0000	28.0000	Москва
306	890.0000	2016	3.0000	131.4000	43.8000	Тверь
307	650.0000	2014	5.0000	155.0000	31.0000	Москва
308	570.0000	2013	6.0000	230.0000	38.3333	Нижний Новгород
309	950.0000	2017	2.0000	20.0000	10.0000	Краснодар
310	1080.0000	2016	3.0000	20.0000	6.6667	Волгоград
311	1100.0000	2018	1.0000	20.0000	20.0000	Москва
312	640.0000	2013	6.0000	158.0000	26.3333	Москва
313	1015.0000	2017	2.0000	82.1350	41.0675	Ростов-на-Дону
314	565.0000	2014	5.0000	150.0000	30.0000	Владимир
315	890.0000	2016	3.0000	131.4000	43.8000	Тверь
316	1030.0000	2018	1.0000	13.0000	13.0000	Архангельск
317	770.0000	2016	3.0000	71.7500	23.9167	Москва
318	1200.0000	2016	3.0000	12.0000	4.0000	Екатеринбург
319	800.0000	2016	3.0000	109.5120	36.5040	Ярославль
320	900.0000	2016	3.0000	110.0000	36.6667	Липецк
321	1000.0000	2016	3.0000	64.0000	21.3333	Москва
322	669.0000	2013	6.0000	200.0000	33.3333	Красноярск

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Регион предложения
	Price	Year	Age	Mileage	MpY	Region
1	2	3	4	5	6	7
323	780.0000	2014	5.0000	350.0000	70.0000	Ростов-на-Дону
324	800.0000	2016	3.0000	80.0000	26.6667	Самара
325	700.0000	2017	2.0000	68.0000	34.0000	Белгород
326	1150.0000	2018	1.0000	30.0000	30.0000	Красноярск
327	1050.0000	2014	5.0000	13.0000	2.6000	Тула
328	800.0000	2015	4.0000	33.0000	8.2500	Воронеж
329	1120.0000	2017	2.0000	28.0000	14.0000	Севастополь
330	600.0000	2013	6.0000	135.0000	22.5000	Брянск
331	570.0000	2014	5.0000	160.0000	32.0000	Воронеж
332	1500.0000	2016	3.0000	2.5000	0.8333	Новосибирск
333	600.0000	2013	6.0000	130.0000	21.6667	Санкт-Петербург
334	1150.0000	2017	2.0000	60.0000	30.0000	Екатеринбург
335	1040.0000	2015	4.0000	0.3500	0.0875	Москва

Таблица 2: Сведения о предлагаемых к продаже новых автомобилях ГАЗ-А23R32 (Газель Next кузов промтоварный фургон)

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.
	Price.new				
1	2	3	4	5	6
1	1415.0000	Эксплуатация не начата	0	0	0
2	1440.0000	Эксплуатация не начата	0	0	0
3	1425.0000	Эксплуатация не начата	0	0	0
4	1590.0000	Эксплуатация не начата	0	0	0
5	1623.1000	Эксплуатация не начата	0	0	0
6	1621.0000	Эксплуатация не начата	0	0	0
7	1534.0000	Эксплуатация не начата	0	0	0
8	1300.0000	Эксплуатация не начата	0	0	0
9	1495.0000	Эксплуатация не начата	0	0	0
10	1323.0000	Эксплуатация не начата	0	0	0
11	1250.0000	Эксплуатация не начата	0	0	0
12	1250.0000	Эксплуатация не начата	0	0	0
13	1360.0000	Эксплуатация не начата	0	0	0
14	1353.6000	Эксплуатация не начата	0	0	0
15	1204.0000	Эксплуатация не начата	0	0	0

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.
	Price.new				
1	2	3	4	5	6
16	1400.0000	Эксплуатация не начата	0	0	0
17	1575.0000	Эксплуатация не начата	0	0	0
18	1446.0000	Эксплуатация не начата	0	0	0
19	1270.0000	Эксплуатация не начата	0	0	0
20	1755.0000	Эксплуатация не начата	0	0	0
21	1720.0000	Эксплуатация не начата	0	0	0
22	2615.0000	Эксплуатация не начата	0	0	0
23	1245.5000	Эксплуатация не начата	0	0	0
24	1409.0000	Эксплуатация не начата	0	0	0
25	1443.7000	Эксплуатация не начата	0	0	0
26	1359.0000	Эксплуатация не начата	0	0	0
27	1605.0000	Эксплуатация не начата	0	0	0
28	1295.0000	Эксплуатация не начата	0	0	0
29	1377.0000	Эксплуатация не начата	0	0	0
30	1300.0000	Эксплуатация не начата	0	0	0
31	1315.0000	Эксплуатация не начата	0	0	0
32	1312.5000	Эксплуатация не начата	0	0	0
33	1256.0000	Эксплуатация не начата	0	0	0
34	1642.1000	Эксплуатация не начата	0	0	0
35	1676.7000	Эксплуатация не начата	0	0	0
36	1360.5000	Эксплуатация не начата	0	0	0
37	1380.7000	Эксплуатация не начата	0	0	0
38	1521.0000	Эксплуатация не начата	0	0	0
39	1496.0000	Эксплуатация не начата	0	0	0
40	1600.0000	Эксплуатация не начата	0	0	0
41	1410.5000	Эксплуатация не начата	0	0	0
42	1565.5000	Эксплуатация не начата	0	0	0
43	1374.7500	Эксплуатация не начата	0	0	0
44	1585.5000	Эксплуатация не начата	0	0	0
45	1530.0000	Эксплуатация не начата	0	0	0
46	1320.0000	Эксплуатация не начата	0	0	0
47	1250.0000	Эксплуатация не начата	0	0	0
48	1535.0000	Эксплуатация не начата	0	0	0
49	1300.0000	Эксплуатация не начата	0	0	0
50	1400.0000	Эксплуатация не начата	0	0	0

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.
	Price.new				
1	2	3	4	5	6
51	1300.0000	Эксплуатация не начата	0	0	0
52	1414.0000	Эксплуатация не начата	0	0	0
53	1365.0000	Эксплуатация не начата	0	0	0
54	1330.0000	Эксплуатация не начата	0	0	0
55	1320.0000	Эксплуатация не начата	0	0	0
56	1401.5000	Эксплуатация не начата	0	0	0
57	1527.7500	Эксплуатация не начата	0	0	0
58	1580.2500	Эксплуатация не начата	0	0	0
59	1415.0000	Эксплуатация не начата	0	0	0
60	1401.7500	Эксплуатация не начата	0	0	0
61	1575.0000	Эксплуатация не начата	0	0	0
62	1370.0000	Эксплуатация не начата	0	0	0
63	1525.5000	Эксплуатация не начата	0	0	0
64	1500.0000	Эксплуатация не начата	0	0	0
65	1353.0000	Эксплуатация не начата	0	0	0
66	1425.0000	Эксплуатация не начата	0	0	0
67	1528.0000	Эксплуатация не начата	0	0	0
68	1395.0000	Эксплуатация не начата	0	0	0
69	1460.0000	Эксплуатация не начата	0	0	0
70	1333.3330	Эксплуатация не начата	0	0	0
71	1425.0000	Эксплуатация не начата	0	0	0
72	1540.0000	Эксплуатация не начата	0	0	0
73	1331.5000	Эксплуатация не начата	0	0	0
74	1479.5000	Эксплуатация не начата	0	0	0
75	1300.0000	Эксплуатация не начата	0	0	0
76	1425.0000	Эксплуатация не начата	0	0	0
77	1800.0000	Эксплуатация не начата	0	0	0
78	1470.0000	Эксплуатация не начата	0	0	0
79	1687.0000	Эксплуатация не начата	0	0	0
80	1500.0000	Эксплуатация не начата	0	0	0
81	1580.0000	Эксплуатация не начата	0	0	0
82	1635.0000	Эксплуатация не начата	0	0	0
83	1700.0000	Эксплуатация не начата	0	0	0
84	1600.0000	Эксплуатация не начата	0	0	0
85	1330.0000	Эксплуатация не начата	0	0	0

№ пп	Цена предложения, тыс. руб.	Год начала эксплуатации	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.
	Price.new				
1	2	3	4	5	6
86	1455.5000	Эксплуатация не начата	0	0	0
87	1353.2500	Эксплуатация не начата	0	0	0
88	1365.0000	Эксплуатация не начата	0	0	0
89	1330.0000	Эксплуатация не начата	0	0	0
90	1540.0000	Эксплуатация не начата	0	0	0
91	1195.0000	Эксплуатация не начата	0	0	0
92	1540.0000	Эксплуатация не начата	0	0	0
93	1300.0000	Эксплуатация не начата	0	0	0
94	1425.0000	Эксплуатация не начата	0	0	0
95	1505.5930	Эксплуатация не начата	0	0	0
96	1345.0000	Эксплуатация не начата	0	0	0
97	1425.0000	Эксплуатация не начата	0	0	0
98	1340.0000	Эксплуатация не начата	0	0	0
99	1425.0000	Эксплуатация не начата	0	0	0
100	1330.0000	Эксплуатация не начата	0	0	0
101	1380.0000	Эксплуатация не начата	0	0	0
102	1525.0000	Эксплуатация не начата	0	0	0
103	1332.5000	Эксплуатация не начата	0	0	0
104	1332.5000	Эксплуатация не начата	0	0	0
105	1250.0000	Эксплуатация не начата	0	0	0
106	1345.0000	Эксплуатация не начата	0	0	0
107	1425.0000	Эксплуатация не начата	0	0	0
108	1535.0000	Эксплуатация не начата	0	0	0
109	1442.0000	Эксплуатация не начата	0	0	0
110	1600.0000	Эксплуатация не начата	0	0	0
111	1617.0000	Эксплуатация не начата	0	0	0
112	1559.0000	Эксплуатация не начата	0	0	0
113	1496.0000	Эксплуатация не начата	0	0	0
114	1525.0000	Эксплуатация не начата	0	0	0
115	1300.0000	Эксплуатация не начата	0	0	0
116	1210.0000	Эксплуатация не начата	0	0	0
117	1315.0000	Эксплуатация не начата	0	0	0
118	1374.5000	Эксплуатация не начата	0	0	0
119	1320.0000	Эксплуатация не начата	0	0	0

Глава 3. Начало работы

3.1. Импорт исходных данных в R, проверка корректности

Для дальнейшего анализа будем использовать среду разработки R. На основе имеющихся таблиц со сведениями о новых и с пробегом автобилей создадим два файла в формате CSV. Следует сделать небольшую ремарку о том, что под форматом CSV как правило понимают более общий формат DSV, допускающий в отличие от CSV использование разделителей отличных от запятой. В нашем случае таким разделителем будет запятая. Десятичные знаки отделены точкой. Файл со сведениями о новых автомобилях назовём «gazelle_new.csv» об автомобилях с пробегом «gazelle_old.csv». При сохранении электронной таблицы в формате .csv табличный процессор запросит указать кодировку, в которой следует сохранить данные. Рекомендуется указать кодировку Windows-1251. Оба файла сохраняем в одну директорию, которая в дальнейшем станет рабочим каталогом для всего проекта. Ранее в Разделе 1.4. Каллиграфия Data mining. Краткое стилевое руководство по R на стр. 20 уже было сказано о том, что путь к рабочему каталогу должен содержать только латинские буквы, цифры и специальный символы, и не должен содержать пробелы и кириллические символы. Следует обеспечивать осмысленность названий всех файлов, используемых в работе.

Запускаем RStudio. Если ранее всё было сделано правильно, то пользователь увидит следующее:

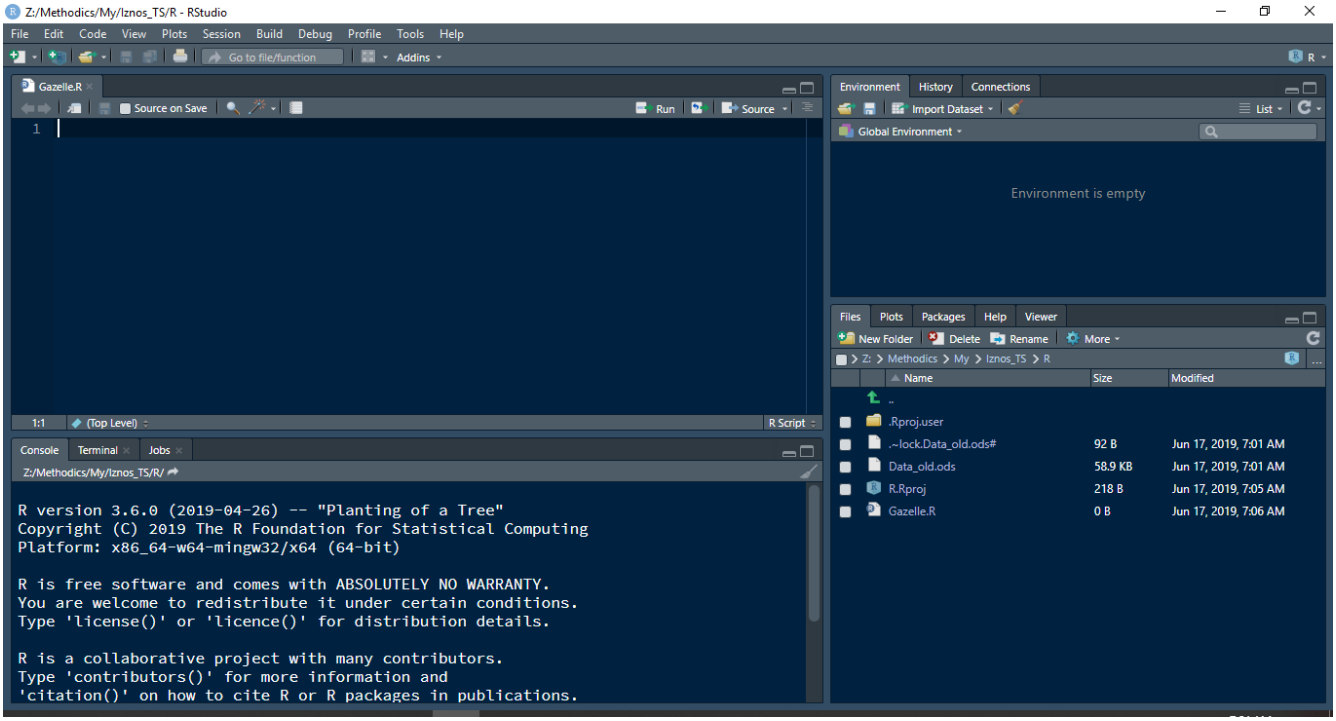

Рисунок 1: Начальный экран при работе с RStudio

Рабочая область разделена на 4 части.

1) В левой нижней части расположена т. н. «консоль» – основной рабочий инструмент R. В неё можно вводить текст кода, там же отображаются сообщения R о результатах выполнения элементов кода. Консоль является единственным рабочим инструментов при работе с R без использования IDE. RStudio же предоставляет и другие средства.

2) В левой верхней части расположена область редактора кода. В ней в отличие от консоли предоставлены расширенные возможности: подсчёт количества символов в строке, возможность настроек шрифта, подсветка элементов кода, исходя из их принадлежности к тому или иному типу его элементов и т. д.

3) В правой верхней части отображается информация о данных, загруженных R в оперативную память машины и базовые средства визуализации этих данных.

4) В правой нижней части расположена многофункциональная зона отображения графических объектов, созданных R, путей к файлам и каталогам, зона доступа к средству управления библиотеками, средству доступа к справке по возможностям и средствам R, средству иной визуализации данных — в зависимости от выбора пользователя.

При работе с R следует придерживаться правила держать все файлы одного проекта в одном каталоге, о принципах именования которого уже было сказано ранее.

Создадим рабочий каталог. Для этого нам потребуется создать первую команду. Этой командой станет setwd – «set working directory». По мере набора первых букв команды IDE начинает помогать пользователю, предлагаю подходящие варианты, а после ввода команды заботливо ставит две скобки. В качестве аргумента команды следует указывать полный путь от названия диска до самого каталога включительно. Аргумент заключается в кавычки с обеих сторон от пути. В случае автора строка выглядит вот так:

```
setwd("Z:\\Methodics\\My\\Iznos_TS\\R") #создаём рабочий каталог
```

Естественно, у каждого будет свой путь.

Запускаем команду путём клика на кнопку “Run” над областью редактора кода.

В консоли отобразится такой же код как в редакторе. Если нет ошибок, то отобразится только он.

В случае ошибки в консоли появится сообщение «Error:...», содержащее описание ошибки.

Иногда вместо ошибки консоль выдаёт предупреждение «Warning:...», указывающее на указание со стороны пользователя выполнить технически корректное, но нетипичное действие. Данное сообщение не требует немедленного исправления кода, но является сигналом для проверки кода и анализа причины.

Автор вводит некоторое упрощение и предупреждает о том, что по тексту Руководства будет использовать термины «процедура» и «функция» как синонимы. Также в качестве синонимов будут использоваться понятия «наблюдение» и «строка»; понятия «переменная», «столбец» и «характеристика».

R является молчуном, и поэтому порой невозможно сразу понять, выполнил ли он то или иное действие. Проверим корректность установки рабочего каталога. Используем для этого команду getwd – «get working directory». Данная команда не требует аргументов.

```
getwd() #проверяем путь к рабочему каталогу
```

Если консоль возвращает значение

```
[1] "Z:/Methodics/My/Iznos_TS/R"
```

значит всё в порядке. Установка рабочего каталога выполнена успешно.

Создадим два объекта – набора данных. Один для обработки первичной информации о новых автомобилях, второй — об автомобилях с пробегом.

Назовём первый из них «gazelle.new.01», второй «gazelle.old.01».

Для создания набора данных из ранее заготовленных .csv файлов используем процедуру read.table. Следует напомнить, что оператором присвоения в R является <-, а не =.

Процедура read.table имеет следующие аргументы, указываемые через запятую, либо опускаемые в случае необязательности какого-либо из них:

Таблица 3: Аргументы процедуры read.table

№ пп	Аргумент	Описание	Правило написания аргумента	Возможные значения	Значение по умолчанию	Обязательность присвоения значения вручную
1	2	3		4	5	6
1	file	Указывает путь к файлу, из которого импортируются данные. Путь можно указать как в абсолютном виде, например: file = "Z:/Methodics/My/Iznos_TS/R/gazelle_new.csv", так и только в виде имени файла: file = “gazelle_new.csv”, при условии, что тот находится в рабочем каталоге. Также возможно указание полной ссылки на файл, расположенный в ИТС Интернет, например: file = https://sovconsult.ru/files/gazelle_new.csv	Заключается в кавычки	Любые	Нет	Да
2	header	Сообщает R о том, содержит ли первая строка заголовки столбцов	Без кавычек	True False	False	Нет
3	row.names	Сообщает R о номере столбца, содержащего наименования строк	Без кавычек	Числовые	Не имеет	Нет
4	sep	Separator – разделитель. Указывет R, какой разделитель используется для разделения столбцов.	Заключается в кавычки	Символы, проблемы, табуляция	Пробел либо табуляция	Нет
5	dec	Указывает R, какой знак в файле отделяет целую часть от дробной	Заключается в кавычки	Как правило . либо ,	.	Нет
6	nrows	Указывает R количеств первых строк, которое должно быть импортировано из файла	Без кавычек	Целые числа >0	Не имеет	Нет
7	skip	Указывает R количеств первых строк, которое должно быть пропущено при импорте данных из файла	Без кавычек	Целые числа >0	Не имеет	Нет

Выполним чтение данных из файлов.

```
gazelle.new.01 <- read.table(file = "gazelle_new.csv", header = TRUE, sep = ",")
gazelle.old.01 <- read.table(file = "gazelle_old.csv", header = TRUE, sep = ",")
```

Если всё правильно, консоль выдаст только код, а в правой верхней части появятся два объекта с названиями gazelle.new.01 и gazelle.old.01 и их основные характеристики.

Следует добавит, что в R, как, пожалуй, и в большинстве языков разработки аргументы процедур/функций можно разделить в том числе на: positional и named, что можно перевести как позиционные и именованные соответственно. Принадлежность аргумента к типу «позицонный» означает, что имеет значение место, где он расположен среди других аргументов процедуры, принадлежность аргумента к типу «именованный» означает, что его следует записывать в виде указания имени аргумента и затем его значения. При указании значения именованной опции следует использовать знак =.

На примере процедуры read.table можно сказать, что аргумент file является позиционным и должен идти первым в строке аргументов, при этом указание имени аргумента не является обязательным. Остальные аргументы являются именованными.

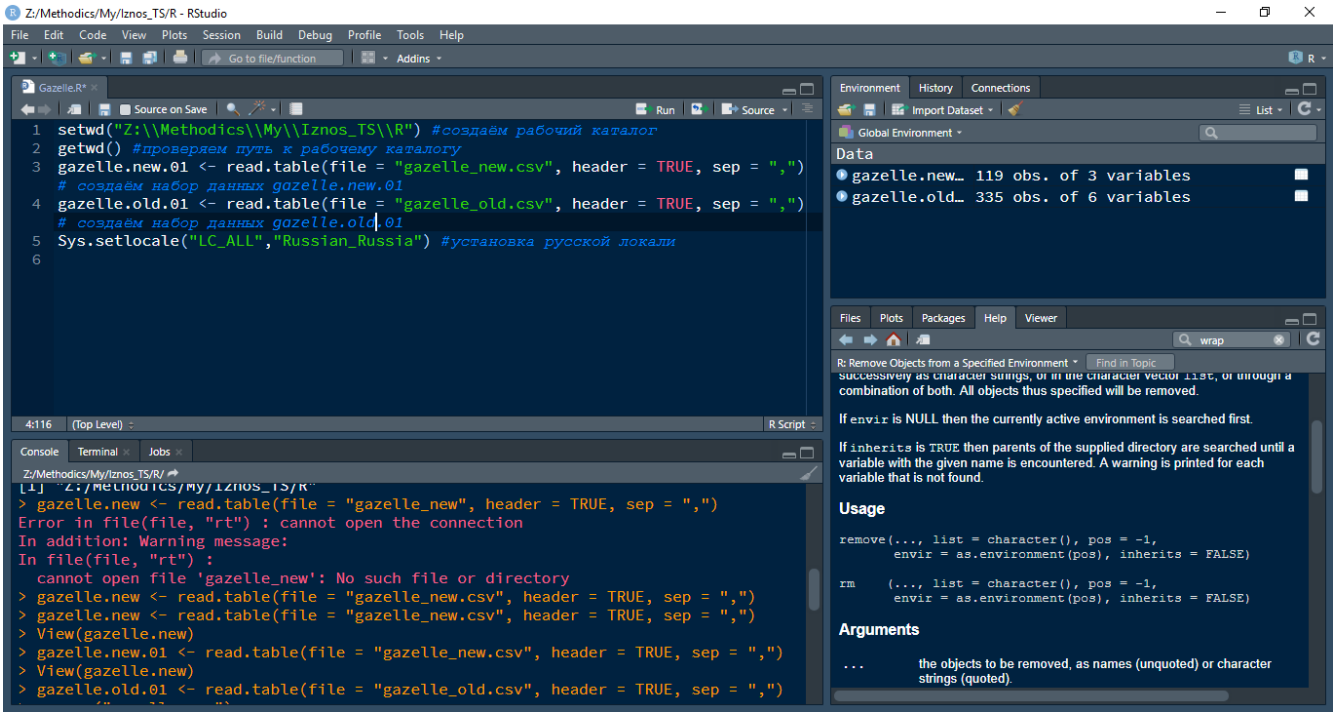

Рисунок 2: Экран после создание объектов – наборов данных

Как видим, помимо названий отображаются показатели obs. и variables. Obs означает количество наблюдений (observations), variables – количество переменных. Количество наблюдений должно совпадать с числом строк исходных файлов минус единица (первая строка с заголовками не содержит сведений о наблюдении), количество переменных — с количество столбцов. В нашем случае всё корректно. 119 наблюдений о ценах новых автомобилей 335 — о ценах автомобилей с пробегом. 3 переменные (Price, Age, Milage) в наборе данных о новых автомобилях, 6 переменных (Price, Year, Age, Mileage, PrY, Region) в наборе данных об автомобилях с пробегом.

Поскольку сведения о регионе приводятся в кириллической кодировке, при начале работы с R желательно проверить корректность их отображения. Для этого используем функцию View. R

различает регистры в названиях процедур. Таким образом вызов функции view ничего не даст. Следует вызывать функцию View. Её единственный значимый аргумент — наименование объекта, подлежащего просмотру. Аргумент не заключается в кавычки.

```
View(gazelle.old.01)
```

В случае успеха R покажет в области редактора кода следующую таблицу:

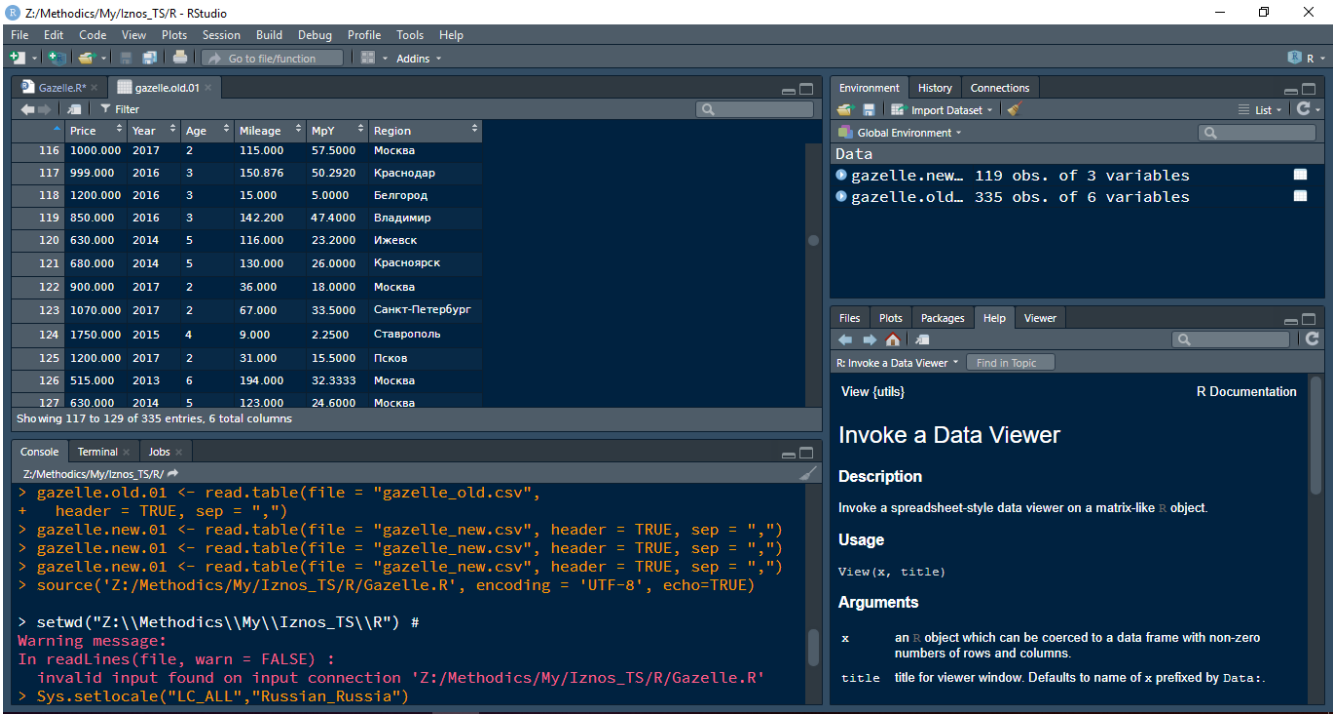

Рисунок 3: Отображение результатов импорта данных

Возможна ситуация, при которой вместо кириллического текста будут отображаться знаки вопроса либо иные символы, свидетельствующие об ошибке определения кириллических символов. В это случае необходимо указать R на необходимость установить русскую локаль. Выполняется при помощи команды:

```
Sys.setlocale("LC_ALL","Russian_Russia") #включение русской локали
```

Консоль выдаст сообщение:

```
[1]
"LC_COLLATE=Russian_Russia.1251;LC_CTYPE=Russian_Russia.1251;LC_MONETARY=Russian_Ru
ssia.1251;LC_NUMERIC=C;LC_TIME=Russian_Russia.1251"
```

После чего повторяем процедуру просмотра. Убедившись, в корректности отображения данных, переходим к дальнейшим действиям.

3.2. Немного теории.

В R существует большое количество типов объектов. В данном Руководстве будут перечислены лишь те, с которыми пользователь столкнётся при решении описываемой задачи либо аналогичных ей.

Прежде всего следует сказать, что объект может обладать такими признаками как type, class, mode, которые можно перевести как «тип», «класс», «режим» соответственно. В рамках данного Руководства, не будет проводиться анализ различий между этими понятиями. В списке литературы приводится значительная подборка книг, детально описывающих язык R. Для целей настоящего Руководства, предназначенного, в первую очередь для практикующих оценщиков, сотрудников залоговых служб банков, андеррайтеров страховых компаний, данные различия не являются существенными, поэтому автор вводит допущение:

type == class == mode

и далее использует термин «тип» во всех случаях.

Типы объектов в R:

1) **Скаляры**. Скаляров в R нет. Скаляр — это вектор длиной 1.
2) **Векторы**. Делятся на:

- numeric (числовые);
- character (текстовые);
- logical (логистические);
- factors (факторы) — им будет посвящена отдельная глава;
- ts (time series) (временные ряды)

3) **Матрицы (matrix).**
4) **Таблицы (data.frame вытесняется data.table).**
5) **Списки (list).**

Матрица — это таблица чисел. За счёт того, что она может содержать только числа — это позволяет экономить память и, тем самым, ускоряет работу. Вследствие этого, следует стремиться кодировать текстовую информацию числовыми кодами. О технологии данного действия будет рассказано в дальнейшем.

Таблица — основной рабочий инструмент. Во многом напоминает привычную электронную таблицу. Содержит важное ограничение: в пределах одного столбца могут быть значения только одного класса. Если числа, то во всём столбце только числа, если текст, то только текст.

Глава 4. Первичная интерпретация информации открытых рынков

Важным первичным этапом анализа рыночной информации является её непосредственная интерпретация аналитиком. Распределение данных, наличие выбросов, асимметрия, наличие выраженных центров плотности — всё это является важными сведениями, позволяющими опытному аналитику сразу же сделать выводы необходимые для быстрой оценки свойств изучаемого явления либо сегмента открытого рынка. Эволюционно мозг человека устроен таким образом, что от 60 до 80 процентов информации поступает в него по визуальному каналу и лишь менее 10 процентов по каналу, который можно назвать смысловым. Подробнее этими и другими выводами можно ознакомиться, например, в работе бывшего руководителя Нидерладского института головного мозга Дика Свааба «Мы — это наш мозг: от матки до Альцгеймера». [31]. Для целей данного Руководства это означает, что не следует пренебрегать визуализацией анализируемой информации. Менее наглядными, но не менее важными являются составление описательных статистик и проверка гипотезы нормальности распределения данных. Эти три операции лежат в основе любого первичного анализа данных открытых рынков.

4.1. Построение гистограмм

Поскольку, как было сказано ранее, предполагается наличие некоторых знаний в сфере исследования операций, статистики и экономики, в Руководстве не будет подробно освещаться суть гистограммы и математическая основа её построения. Достаточно общих сведений о том, что гистограмма — это способ наглядного представления функции плотности вероятности некоторой выборочной случайной величины. Из этого следует, что гистограмма является способом графического отображения распределения данных (значений случайной величины). Частотность событий откладывается по вертикальной оси, группы данных — по горизонтальной. Полосы столбцов имеют одинаковую ширину. На всякий случай следует сделать ремарку: привычная пользователям Excel столбчатая диаграмма не является гистограммой.

Для понимания сути гистограммы кратко рассмотрим алгоритм её построения.

Предположим, что мы имеем n наблюдений, содержащих числовые значения от x_1 до x_n. Возьмём интервал [a,b], содержащий все эти числа. Разбиваем его на k частей. Вопрос нахождения оптимального значения k до сих пор является научной проблемой. Некоторые её аспекты и предлагаемые пути решения приведены в разделе 4.1.1. Выбор числа интервалов гистограмм на стр. 50. Второй проблемой является проблема выбора: должны ли эти части быть одинаковыми по количеству попавших в них наблюдений. На основании эмпирических данных и собственно опыта автор пришёл к выводу о том, что на стадии предварительного анализа данных и их визуализации одинаковость не является проблемой, вследствие чего нет необходимости применять процедуры, позволяющие её избежать.

Созданные части обязательно должны быть непересекающимися. Обозначим их как $\Delta_1 \dots \Delta_k$.

Возникает числовая ось, имеющая границы [a,b], состоящая из k непересекающихся частей.

Затем путём простого деления

$$\frac{n}{k} \quad , \tag{1}$$

где n – количество наблюдений;

k – количество частей, на которые разбиты наблюдения,

определяется количество наблюдений, попавших в каждую часть.

Далее на прямой от a до b строят прямоугольники, высота h которых может быть пропорциональна количеству наблюдений, попавших в отрезок.

$$h_i \propto n_i \quad , \tag{2}$$

где n – количество наблюдений;

i – номер части,

Данный подход является несколько наивным.

Более научный вероятностный подход гласит, что

$$h_i = \frac{n_i}{n|\Delta_i|} \quad , \tag{3}$$

где n_i – количество наблюдений в i-той части;

i – номер части,

n – общее число наблюдений

Δ_i – шаг разбиения

Использование данной формулы обеспечивает условие того, что сумма площадей всех прямоугольников будет всегда равна единице. Выполнение данного условия является необходимым и полезным с точки зрения возможности сравнивать гистограммы, построенные на выборках с разным количеством наблюдений, тогда как в случае определения высоты столбцов по формуле (2) она будет зависеть от свойств конкретного единичного набора данных. Второй довод в пользу использования формулы (3) заключается в следующем. В силу всё того е равенства сумме площадей прямоугольников единице, при таком способе определения высоты мы получаем отображение плотности распределения вероятности значений наблюдений. В этом случае мы можем, хоть и не в явном виде, говорить о возможности определения вероятности того, что измеряемые нами значения наблюдений попадут в интервал [a,b]. В этом случае получается, что данная вероятность может быть рассчитана как интеграл от плотности вероятности.

$$P(x \in [a,b]) = \int_a^b f(t)\,dt \quad , \tag{4}$$

По мере увеличения числа наблюдений и в случае правильного выбора k гистограмма будет всё меньше отличаться от функции плотности.

Ключевым моментом при восприятии гистограммы должно быть то, что при её анализе мы смотрим на площади. Если мы сосчитаем сумму площадей прямоугольников, расположенных на отрезке оси абсцисс от α до β, то это будет означать, что мы определили вероятность того, что значение случайной величины будет находиться в диапазоне от α до β. Разумеется данное утверждение является справедливым только в случае использование формулы (3) для определения высоты столбцов.

При этом следует отметить, что использование формулы (2) может быть оправдано в случаях подготовки гистограмм в презентационных и маркетинговых целях. Но для целей анализа данных осмысленно использование формулы (3).

4.1.1. Выбор числа интервалов гистограмм

Число интервалов группирования, используемое при вычислении оценок параметров, построении гистограмм колеблется в широких пределах. Большинство рекомендуемых формул для оценки числа интервалов k носит эмпирический характер и часто даёт завышенные значения.

В общем случае можно говорить о том, что количество интервалов k связано с количеством наблюдений переменной. Значительное количество рекомендаций по выбору числа интервалов k из различных источников содержится в [32]. Серьёзные теоретические рассуждения приведены в [33]

При выборе интервалов равной длины определяющим является требование, чтобы число наблюдений, попавших в интервалы, было не слишком малым и сравнимым. Ряд авторов [34] утверждает, что количество интервалов должно быть не менее 6. В то же время другие авторы [35] отмечают, что допустимо использование 5 и менее интвервалов.

В работах [36], [37] посвящённых мощности критерия хи-квадрат Пирсона в случае унимодального распределения допускается уменьшение ожидаемых частот попадания наблюдений для одного или двух интервалов до 1 и даже ниже.

В работе Стёрджесса [38] предлагаются следующие формулы:

$$k = \log_2 N + 1 \quad ; \tag{5}$$

$$k = 3.3 lgN + 1 \quad , \tag{6}$$

где N – количество наблюдений переменной (объём выборки).

Следует отметить, что значения полученные путём применения формул (5) и (6) всегда приблизительно равны, вследствие чего можно говорить о двух реализациях одного метода.

В [39] для определения рационального числа интервалов k рекомендуют формулу Брукса и Каррузера:

$$k = 5 lgN \quad . \tag{7}$$

В [40] рекомендуют соотношение:

$$k = \sqrt{N} \quad . \tag{8}$$

В [41] для равновероятных интервалов их число устанавливают порядка:

$$k \approx 4 \sqrt[5]{2 * \left(\frac{N}{t}\right)^{0.4}} \quad , \tag{9}$$

где t – квантиль стандартного нормального распределения для заданного уровня значимости.

В ряде работ приводят модификации данной формулы.

Автор обращает внимание на то обстоятельство, что в ряде публикаций приводится ошибочный вариант данной формулы:

$$k \approx 4 \sqrt[5]{2} \left(\frac{N}{t}\right)^{0.4}$$

Очевидно, что операция извлечения корня из числа 2 и умножение результата на 4 могло быть выполнено авторами самостоятельно и вместо такой неочевидной записи числа они бы предложили уже готовое его значение 4.5948.

В [42] предлагают значение:

$$k = 4 lg(N) \quad . \tag{10}$$

А в [43] дальнейшее развитие этой идеи:

$$k = 5 lg(N) - 5 \quad . \tag{11}$$

В исследовании [44] получено соотношение:

$$k = \frac{4}{\xi} lg \frac{N}{10} \quad , \tag{12}$$

Где ξ – значение контрэксцесса:

$$\zeta = \frac{1}{\sqrt{\frac{\mu_4}{\sigma^4}}} \quad , \tag{13}$$

где μ — центральный момент 4-го порядка [45];

σ — дисперсия.

При больших объёмах выборок N разброс значений k, задаваемых различными формулами, достаточно велик. Поэтому на практике при выборе числа интервалов больше руководствуются тем, чтобы в интервалы попадало число наблюдений не менее 5-10. Так, например, в рекомендациях ВНИИ Метрологии [46] в зависимости от N предлагают следующие значения *k*:

Таблица 4: Рекомендуемое число интервалов k в зависимости от N

№ пп	Количество наблюдений переменной (объём выборки) – N	Количество равных интервалов – *k*
1	2	3
1	40-100	7-9
2	100-500	8-12
3	500-1000	10-16
4	1000-10 000	12-22

Все вышеперечисленные рекомендации опирались на предположение, что k следует выбирать таким образом, чтобы вид гистограммы был как можно ближе к плавной кривой плотности распределения генеральной совокупности.

В [47] показано, что уклонение гистограммы от плотности распределения в лучшем случае имеет порядок:

$$.= \frac{1}{\sqrt[3]{N}} \quad , \tag{14}$$

достигаемый при числе интервалов k порядка

$$.= \sqrt[3]{N} \quad . \tag{15}$$

Очевидно, что рациональное значение k зависит не только от объёма выборки, но и от вида закона распределения и от способа группирования.

При асимптотически оптимальном группировании относительно скалярного параметра при 10-11 интервалах в группированной выборке сохраняется около 98% информации, при оптимальном группировании относительно вектора параметров (два параметра) для 15 интервалов – около 95%. Дальнейшее увеличение числа интервалов существенного значения не имеет.

Конкретное число интервалов при асимптотически оптимальном группировании выбирают, исходя из следующих соображений. При оптимальном группировании вероятности попадания в интервалы в общем случае не равны. Обычно минимальны вероятности попадания в крайние интервалы. Поэтому k желательно выбирать из условия Npi(Ɵ) ≥5-10 для любого интервала при рациональном группировании. По крайней мере, минимальная ожидаемая частота должна быть больше 1. В случае использования равновероятного группирования порядок k должен быть примерно таким же, как и при асимптотически оптимальном группировании.

Все наиболее разумные рекомендации по выбору числа интервалов, в том числе по выбору числа интервалов в случае асимптотически оптимального группирования, исходят из того, чтобы при данном N приблизить плотность распределения её непараметрической оценкой (гистограммой) как можно лучше.

Поскольку определение количества интервалов гистограмм не является особенно существенным для целей настоящей работы, автор не приводит дальнейшие рассуждения об их выборе и переходит к расчётам. Интересующимся данной проблематикой и желающим развить тему рекомендаций в вопросе выбора числе интервалов для построения гистограмм рекомендуется ознакомиться со следующим списком материалов по теме: [48].

Создадим объекты – единичные векторы, содержащие значения о количестве наблюдений по новым и бывшим в употреблении автомобилям. А заодно сразу проверим корректность выполнения команд.

```
  n.rows.new <- nrow(gazelle.new.01) #вычисляем количество наблюдений новых машин

  n.rows.new #проверяем выполнение

  n.rows.old <- nrow(gazelle.old.01) #вычисляем количество наблюдений б/у машин

  n.rows.old #проверяем выполнение
```

Консоль выдаёт следующие сообщения:

```
> n.rows.new <- nrow(gazelle.new.01) #вычисляем количество наблюдений новых машин
> n.rows.new
[1] 119
> n.rows.old <- nrow(gazelle.old.01) #вычисляем количество наблюдений б/у машин
> n.rows.old #проверяем выполнение

[1] 335
```

Всё корректно. Далее будем осуществлять расчёты. Почему мы создали объекты, содержащие сведения о количестве наблюдений, вместо простого использования чисел в формулах? Это азы работы с данными. Количество наблюдений может изменяться. А можно сделать опечатку.

Определение параметра k с использованием формул Стёрджесса.

Для данных по рынку новых автомобилей значение k, определённое по формуле (5), составляет 7.8948, по формуле (6) – 7.8493.

Для данных по рынку автомобилей с пробегом значение k, определённое по формуле (5), составляет 9.3880, по формуле (6) – 9.3326.

Определение параметра k с использованием формулы Брукса-Каррузера

Для данных по рынку новых автомобилей значение k, определённое по формуле (7), составляет 10.3777.

Для данных по рынку автомобилей с пробегом значение k, определённое по формуле (7), составляет 12.6252.

Расчёт по формуле Heinhold I., Gaede K.W.

Для данных по рынку новых автомобилей значение k, определённое по формуле (8), составляет 10.9087.

Для данных по рынку автомобилей с пробегом значение k, определённое по формуле (8), составляет 18.3030.

Расчёт по формуле Mann H.B., Wald A.

Данная формула требует выбора значения ранее не упоминавшегося показателя уровня значимости.

Уровень значимости статистического теста — это допустимая для решаемой задачи вероятность ошибки первого рода (ложноположительного решения, false positive), то есть вероятность отклонить нулевую гипотезу, когда на самом деле она верна.

Данное определение, хотя и является каноническим определением уровня значимости, всё же выглядит не вполне соотносящимся с текущей задачей выбора значения k, поэтому дополнительно сформулируем иначе.

Уровень значимости — это такое (достаточно малое) значение вероятности события, при котором событие уже не является случайным.

Уровень значимости обычно обозначают греческой буквой α (альфа).

В дальнейшем в Руководстве будет приведён расширенный анализ данных понятий. Пока только введём допущение о то, что **здесь и далее по всему тексту Руководства будет использоваться уровень значимости α = 0.05, если только в тексте не указано иное. Значение дисперсии при этом принимается равным 1, если только также прямо не указано иное.**

Для данных по рынку новых автомобилей значение k, определённое по формуле (9), составляет 6.4717.

Для данных по рынку автомобилей с пробегом значение k, определённое по формуле (9), составляет 7.0304.

Расчёт по формуле Таушанова-Тоневой-Пеновой

Для данных по рынку новых автомобилей значение k, определённое по формуле (10), составляет 8.3022.

Для данных по рынку автомобилей с пробегом значение k, определённое по формуле (10), составляет 10.1002.

Расчёт по формуле Е. Тоневой

Для данных по рынку новых автомобилей значение k, определённое по формуле (11), составляет 5.3777.

Для данных по рынку автомобилей с пробегом значение k, определённое по формуле (11), составляет 7.6252.

Расчёт по формуле И.У. Алексеевой

Для расчёта по формуле (12) нам потребуется рассчитать значение такого показателя как контрэксцесс. В стандартных средствах R нет готовых функций для этого. Нам потребуется установить библиотеку «moments». Сделаем это при помощи команды install.packages, аргументом которой является название библиотеки, заключённое в кавычки.

После установки библиотеки она не запускается автоматически. Её нужно включить командой library, аргументом которой также является название библиотеки, заключённое в кавычки.

Код для выполнения вышеуказанных действий.

```
install.packages("moments") #устанавливаем библиотеку

library("moments") #включаем библиотеку
```

Следует отметить, что включённые библиотеки остаются таковыми только до конца сессии R. После его перезапуска для экономии памяти включаются только базовые пакеты. Процедура включения нужных библиотек уже описана выше.

Для переменной Price из набора данных по рынку новых автомобилей значение k, определённое по формуле (12), составляет 7.0047.

Для переменной Price из набора данных по рынку автомобилей с пробегом значение k, определённое по формуле (12), составляет 7.3624.

Для переменной Age из набора данных по рынку автомобилей с пробегом значение k, определённое по формуле (12), составляет 5.7839.

Для переменной Mileage из набора данных по рынку автомобилей с пробегом значение k, определённое по формуле (12), составляет 9.3750.

Для переменной MpY из набора данных по рынку автомобилей с пробегом значение k, определённое по формуле (12), составляет 12.9065.

Данная формула интересна тем, что только она учитывает не только количество данных, но их их распределение. Она безошибочно определила, что у переменной Age всего 6 значений и никакого смысла вводить количество столбцов больше 6 нет.

```
log(n.rows.new, 2)+1 #определяем k для нов. машин по 1 формуле Стёрджесса

3.3*log(n.rows.new, 10)+1 #определяем k для нов. машин по 2 формуле Стёрджесса

log(n.rows.old, 2)+1 #определяем k для б/у машин по 1 формуле Стёрджесса

3.3*log(n.rows.old, 10)+1 #определяем k для б/у машин по 2 формуле Стёрджесса

5*log(n.rows.new, 10) #определяем k для нов. машин по ф. Брукса-Каррузера

5*log(n.rows.old, 10) #определяем k для б/у машин по ф. Брукса-Каррузера

sqrt(n.rows.new) #определяем k для новых машин по ф. Heinhold I., Gaede K.W.

sqrt(n.rows.old) #определяем k для б/у машин по ф. Heinhold I., Gaede K.W.

4*((((n.rows.new/qnorm(0.95))^0.4)*2)^(1/5)) #определяем k для новых машин по
формуле Mann H.B., Wald A.

4*((((n.rows.old/qnorm(0.95))^0.4)*2)^(1/5)) #определяем k для б/у машин по
формуле Mann H.B., Wald A.

4*log(n.rows.new, 10) #определяем k для новых машин по ф. Таушанова-Тоневой-
Пеновой

4*log(n.rows.old, 10) #определяем k для б/у машин по ф. Таушанова-Тоневой-
Пеновой

5*log(n.rows.new, 10)-5 #определяем k для новых машин по ф. Тоневой

5*log(n.rows.old, 10)-5 #определяем k для б/у машин по ф. Тоневой

install.packages("moments") #устанавливаем библиотеку

library("moments") #включаем библиотеку

(4/(1/(sqrt(kurtosis(gazelle.new.01$Price)))))*(log((n.rows.new/10), 10))
#определяем k для новых машин по формуле Алексеевой

(4/(1/(sqrt(kurtosis(gazelle.old.01$Price)))))*(log((n.rows.new/10), 10))
#определяем k для переменной Price для б/у машин по формуле Алексеевой
```

```
  (4/(1/(sqrt(kurtosis(gazelle.old.01$Age)))))*(log((n.rows.new/10), 10))
#определяем k для переменной Age для б/у машин по формуле Алексеевой

  (4/(1/(sqrt(kurtosis(gazelle.old.01$Mileage)))))*(log((n.rows.new/10), 10))
#определяем k для переменной Mileage для б/у машин по формуле Алексеевой

  (4/(1/(sqrt(kurtosis(gazelle.old.01$MpY)))))*(log((n.rows.new/10), 10))
#определяем k для переменной MpY для б/у машин по формуле Алексеевой
```

Внимательный читатель может обратить внимание на порой избыточное количество скобок. Данное обстоятельство не является прихотью автора, а основано на практическом опыте использования R. R – это такой язык, в котором всегда лучше поставить лишнюю скобку, чем не поставить необходимую. В дальнейшем мы ещё столкнёмся с тем, что отсутствие скобки в неочевидном месте приводит к ошибкам кода. Таким образом, по мере работы с R возникает привычка перестраховываться и ставить скобки везде, где есть хотя бы теоретическая возможность потребности в них.

Обобщим полученные результаты:

Таблица 5: Сведения о рациональном количестве интервалов значений переменных, полученном различными методами

		Расчётное рациональное число интервалов				
№ пп	Наименование метода	Цена предложения автомобилей с пробегом, тыс. руб.	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Цена предложения новых автомобилей, тыс. руб.
		Price	Age	Mileage	MpY	Price
1		2	3	4	5	6
1	Метод Стёрджеса	9	9	9	9	8
2	Метод Брукса-Каррузера	13	13	13	13	10
3	Метод Хайнхольда-Гайде	18	18	18	18	11
4	Метод Манна-Вальда	7	7	7	77	6
5	Метод Таушанова-Тоневой-Пеновой	10	10	10	10	8
6	Метод Тоневой	8	8	8	8	5
7	Метод Алексеевой	7	6	9	13	7

Поскольку, как было сказано ранее, выбор числа интервалов гистограмм не является принципиальным вопросом для данной работы, дальнейший анализ данного вопрос не проводится. Автор принимает значение количества интервалов гистограмм равным 7 для данных по ценам новых автомобилей, равным 10 для переменной Price для автомобилей с пробегом, равным 6 для Age, равным 10 для Mileage, равным 13 для MpY. **При недостатке времени на анализ значений всех формул, автор рекомендует в первую очередь использовать формулу Алексеевой, во вторую — Стёрджесса.**

4.1.2. Построение гистограмм

Построим гистограммы, совмещённые с кривой плотности вероятности для всех исследуемых переменных. Хочется предостеречь тех, кто только знакомится с R от соблазна использовать его поистине фантастические возможности в области визуализации данных на полную мощность. Безусловно для целей представления данных руководству и заказчикам, подготовки итоговых материалов, в т.ч. и для отчётов об оценке возможности R в области визуализации являются одним из аргументов в пользу его внедрения в повседневную практику. Однако, собственно для анализа данных предпочтительным является минимализм, использование чёрно-белых цветов, оттенков серого. В случае использования цветных объектов, предпочтение следует отдавать синему цвету: среди нас есть люди с пониженным цветовосприятием. Синий же цвет воспринимается ими относительно нормально. Ниже приводится два примера визуализации данных, созданных в R, после чего перейдём к построению гистограмм для наших данных.

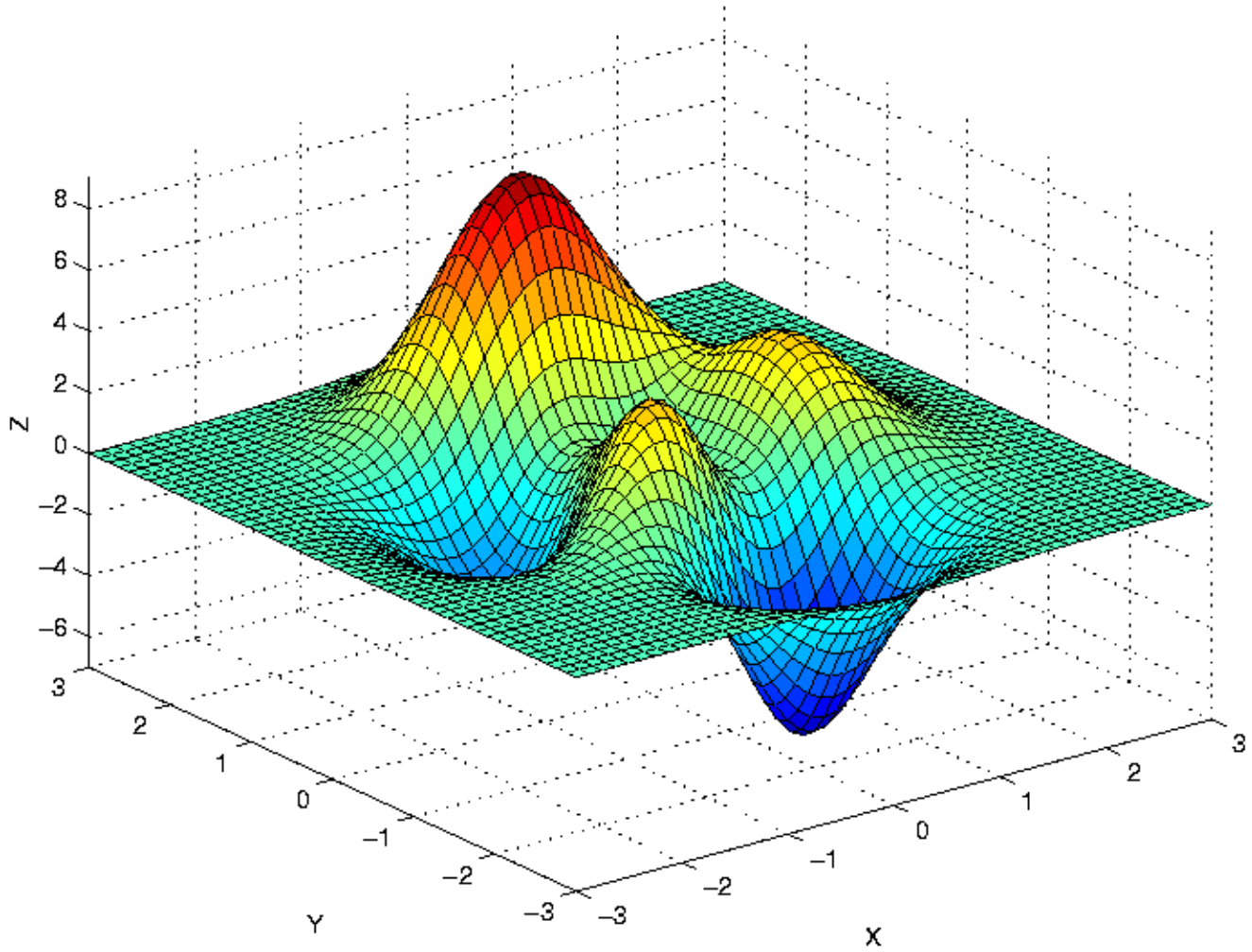

Рисунок 4: Пример визуализации, созданной средствами R
Источник [49]

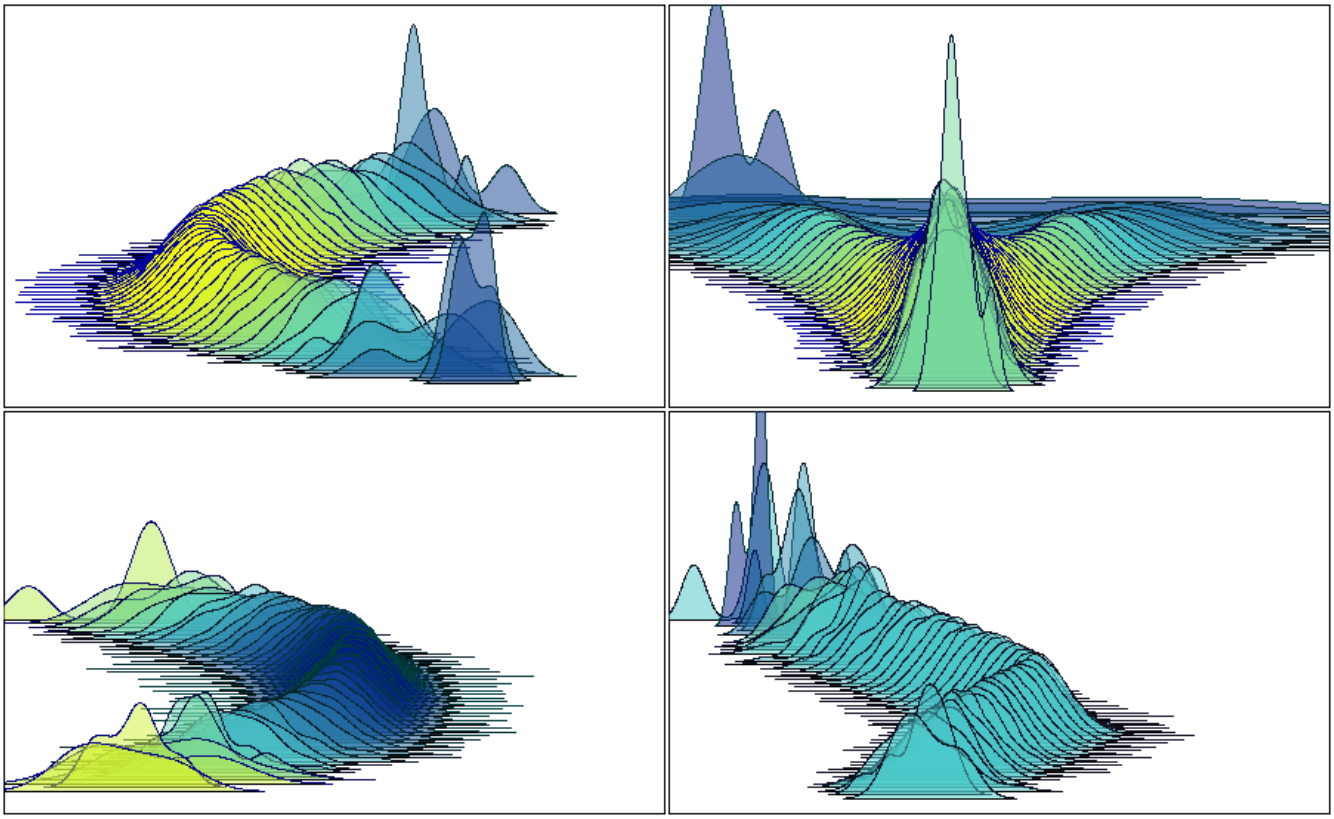

Рисунок 5: Ещё один пример визуализации данных, созданной средствами R
Источник [50]

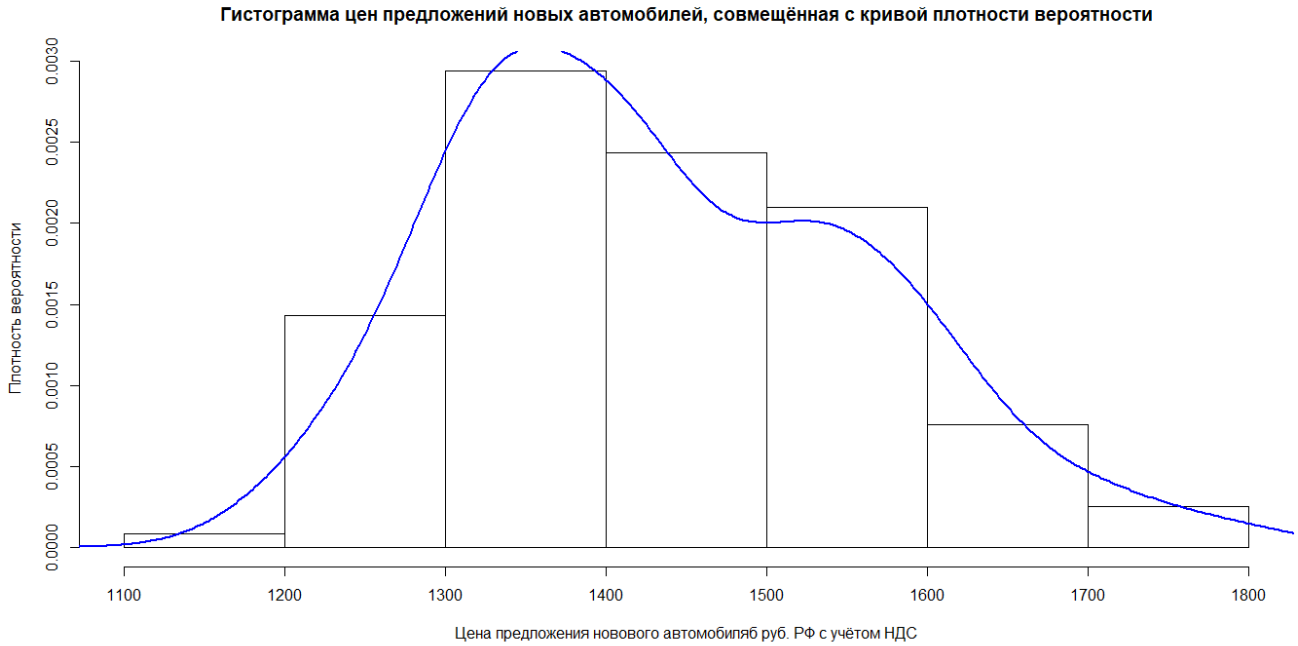

Рисунок 6: Гистограмма цен предложений новых автомобилей, совмещённая с кривой плотности вероятности

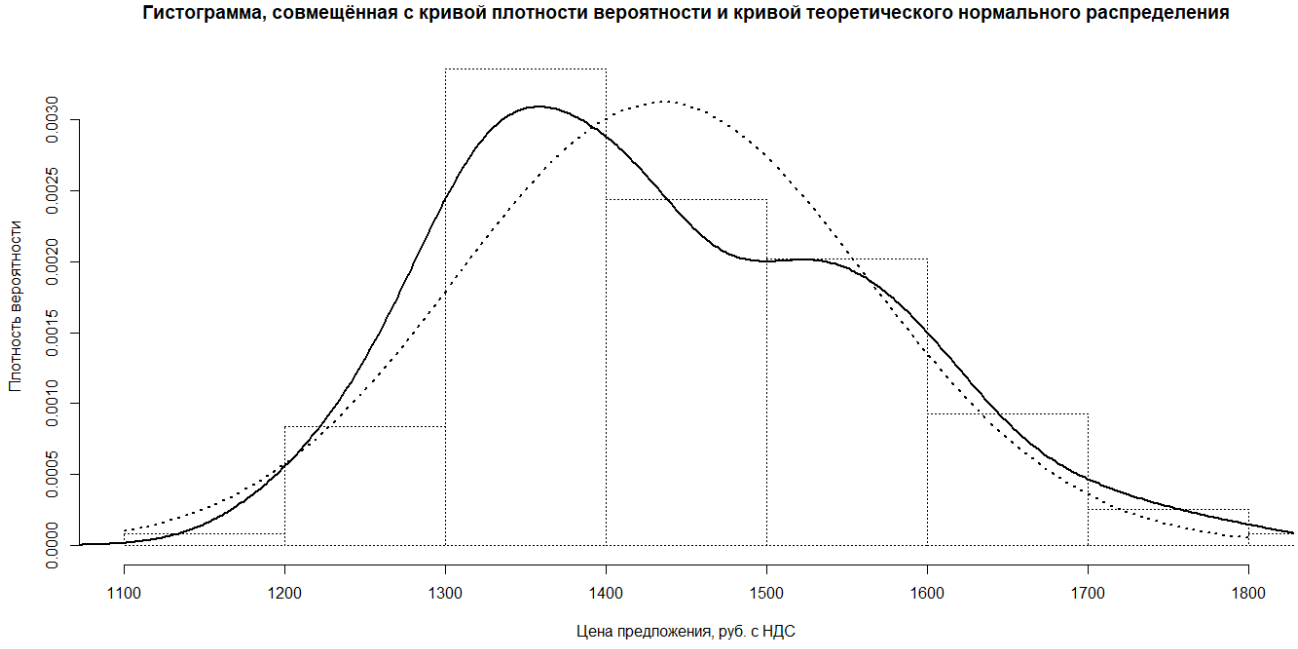

Рисунок 7: Гистограмма цен предложений новых автомобилей, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

Рисунок 8: Гистограмма цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности

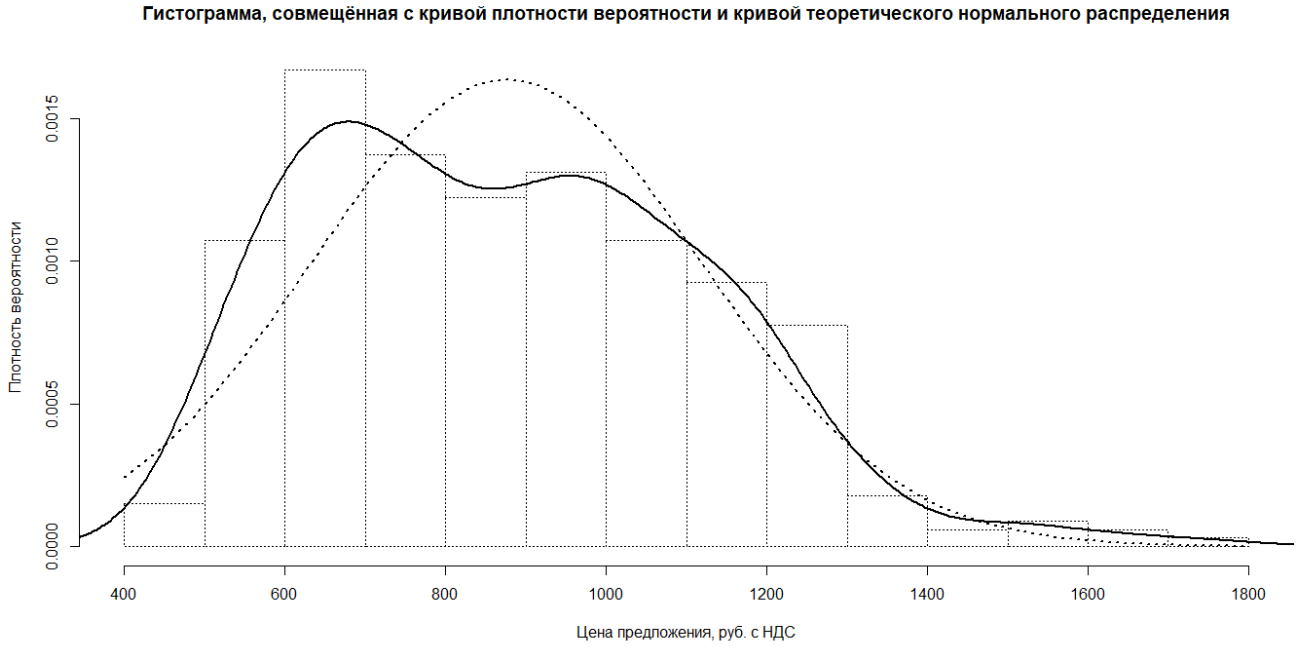

Рисунок 9: Гистограмма цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

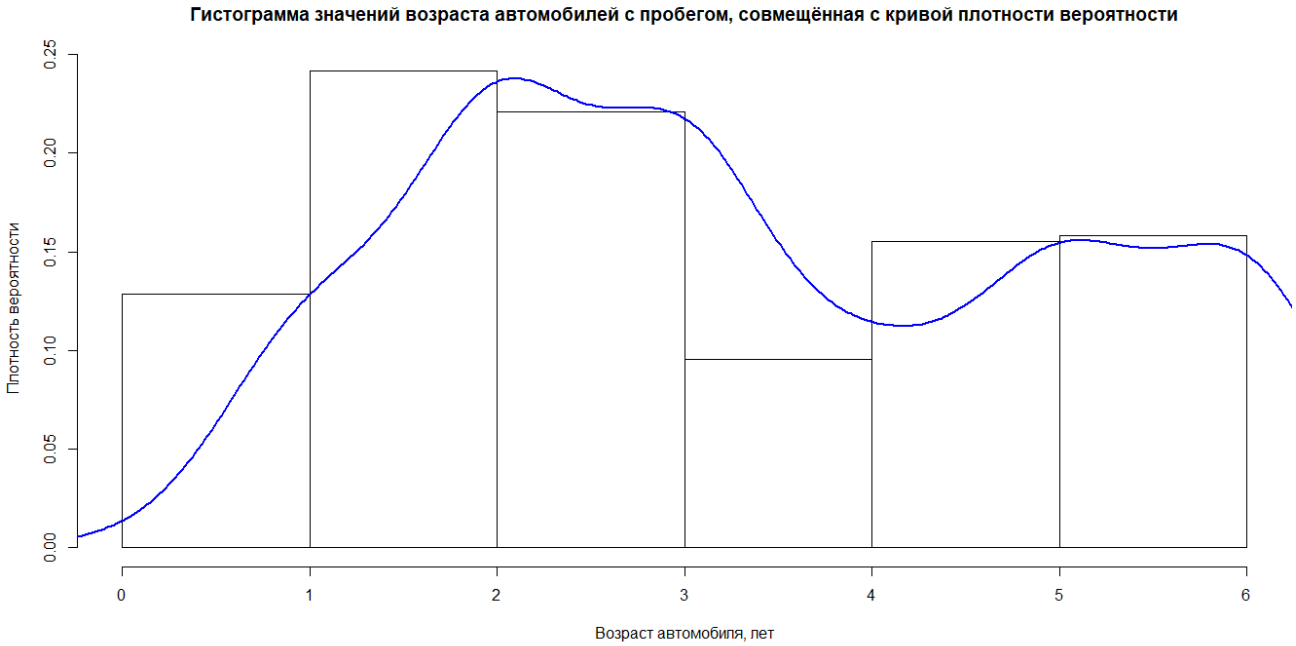

Рисунок 10: Гистограмма значений возраста автомобилей с пробегом, совмещённая с кривой плотности вероятности

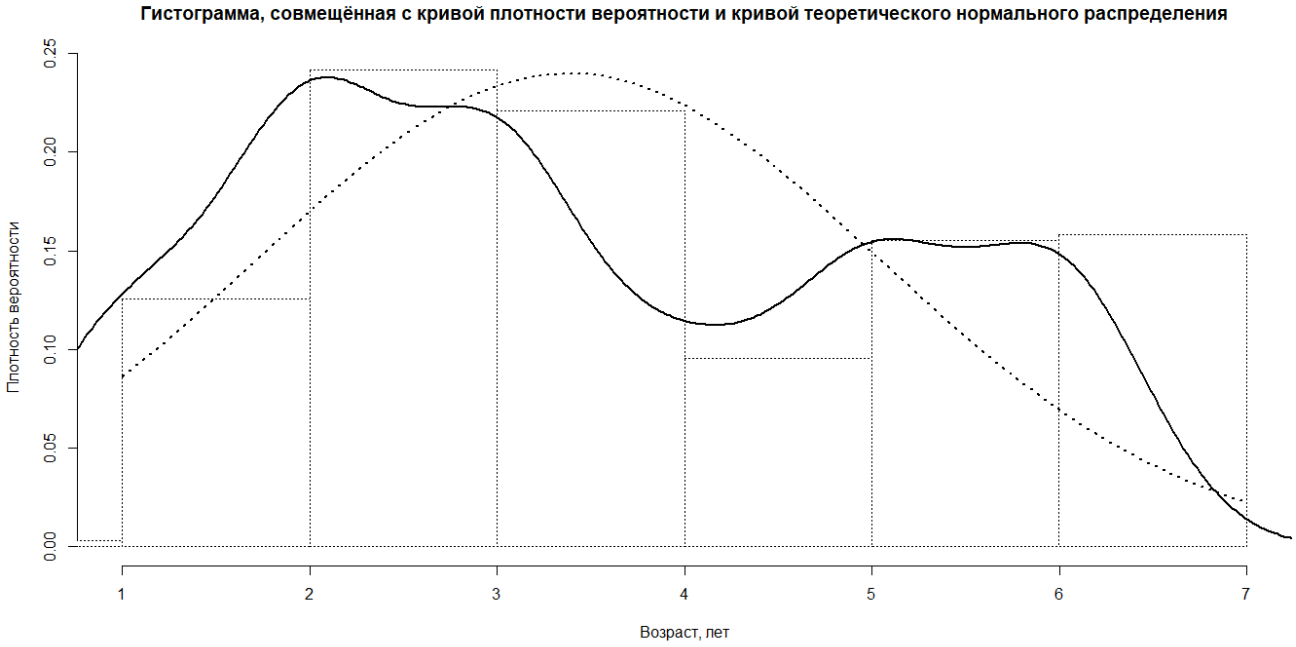

Рисунок 11: Гистограмма значений возраста автомобилей с пробегом, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

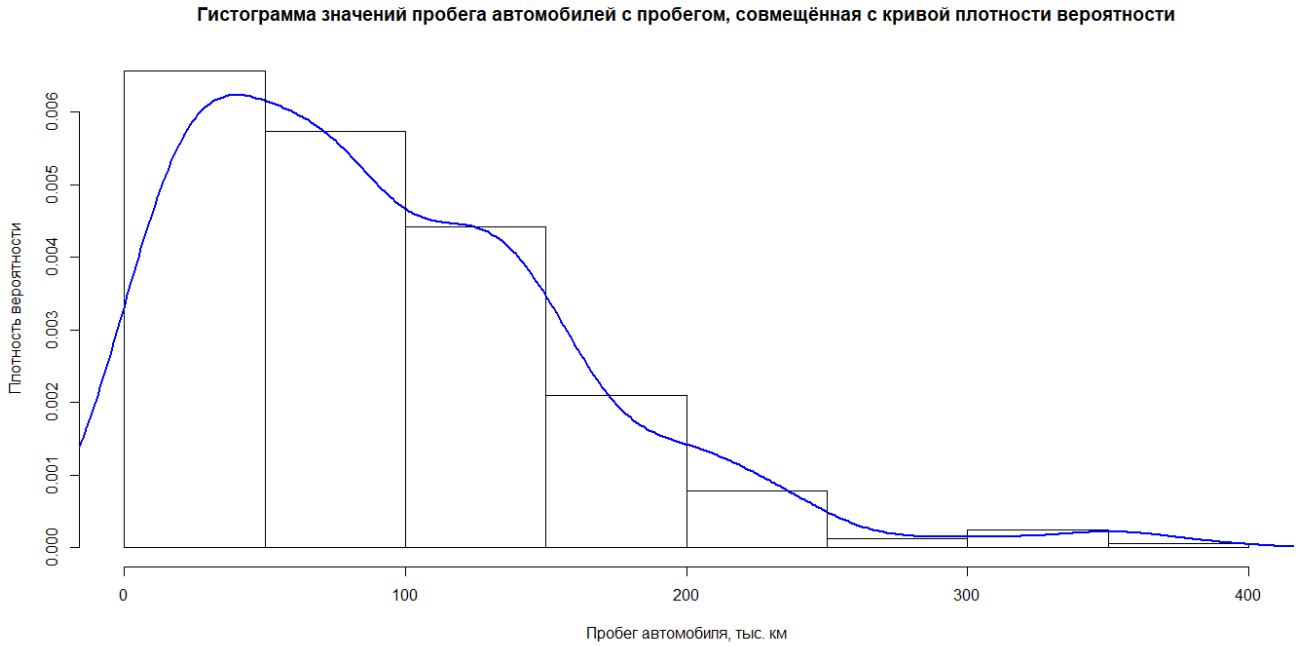

Рисунок 12: Гистограмма значений пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности

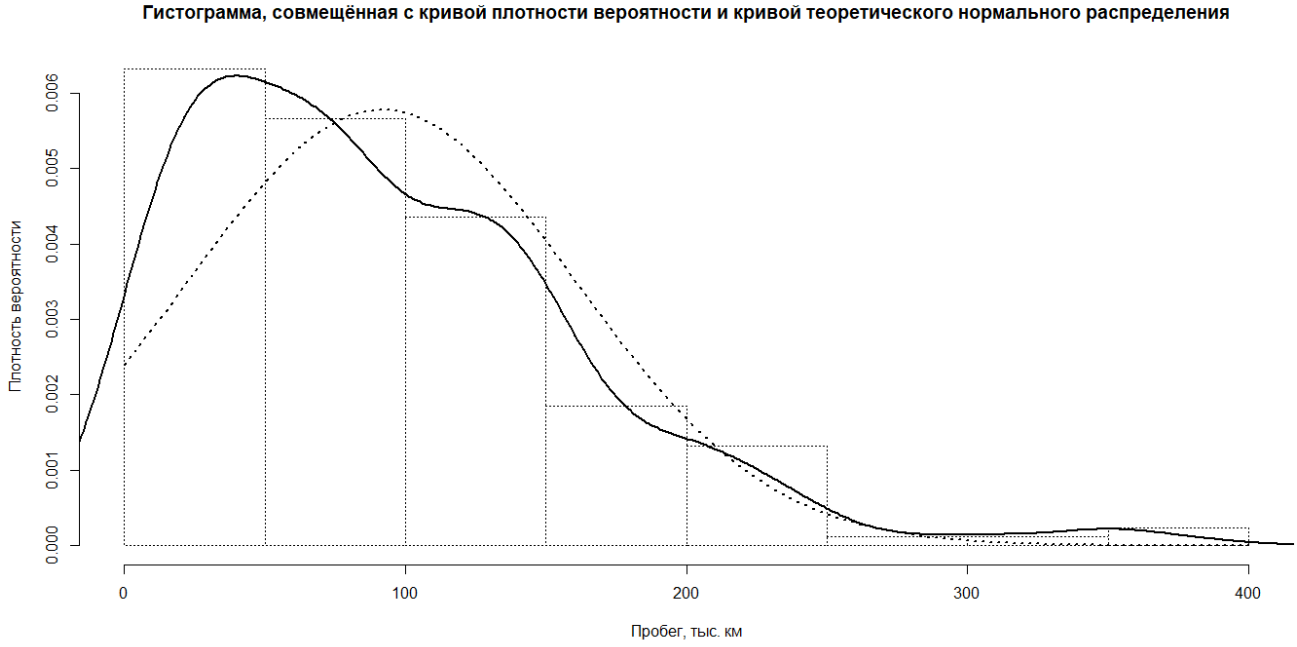

Рисунок 13: Гистограмма значений пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

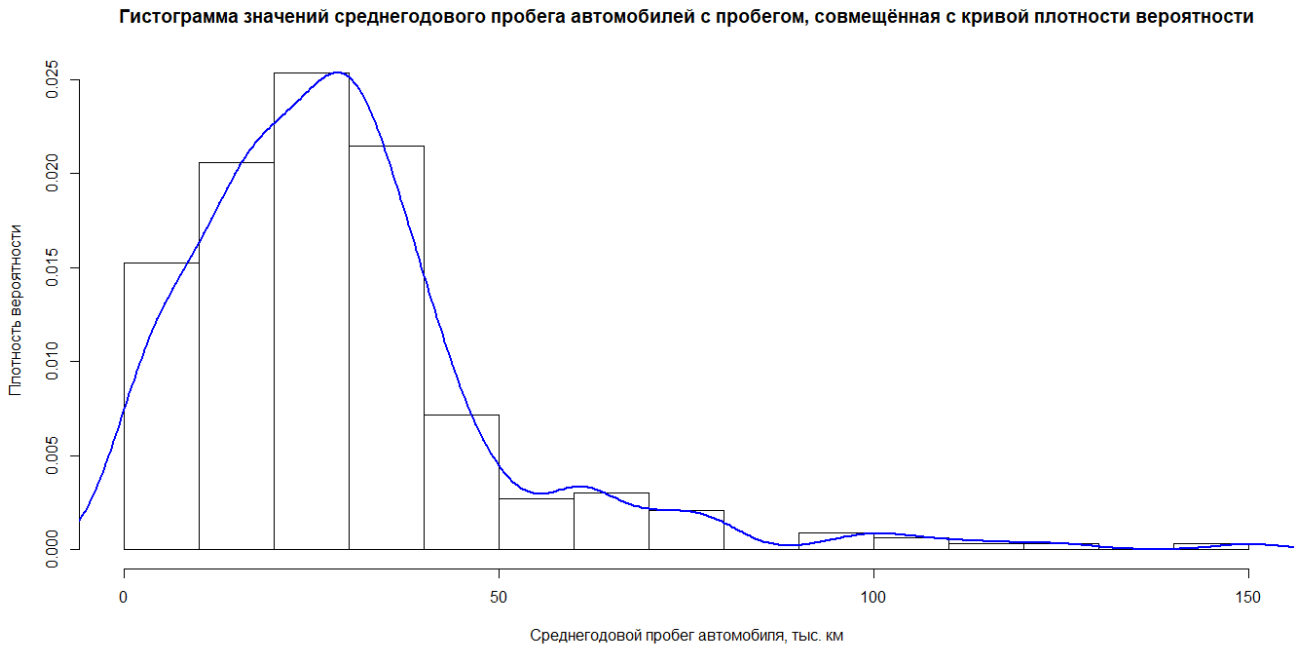

Рисунок 14: Гистограмма значений среднегодового пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности

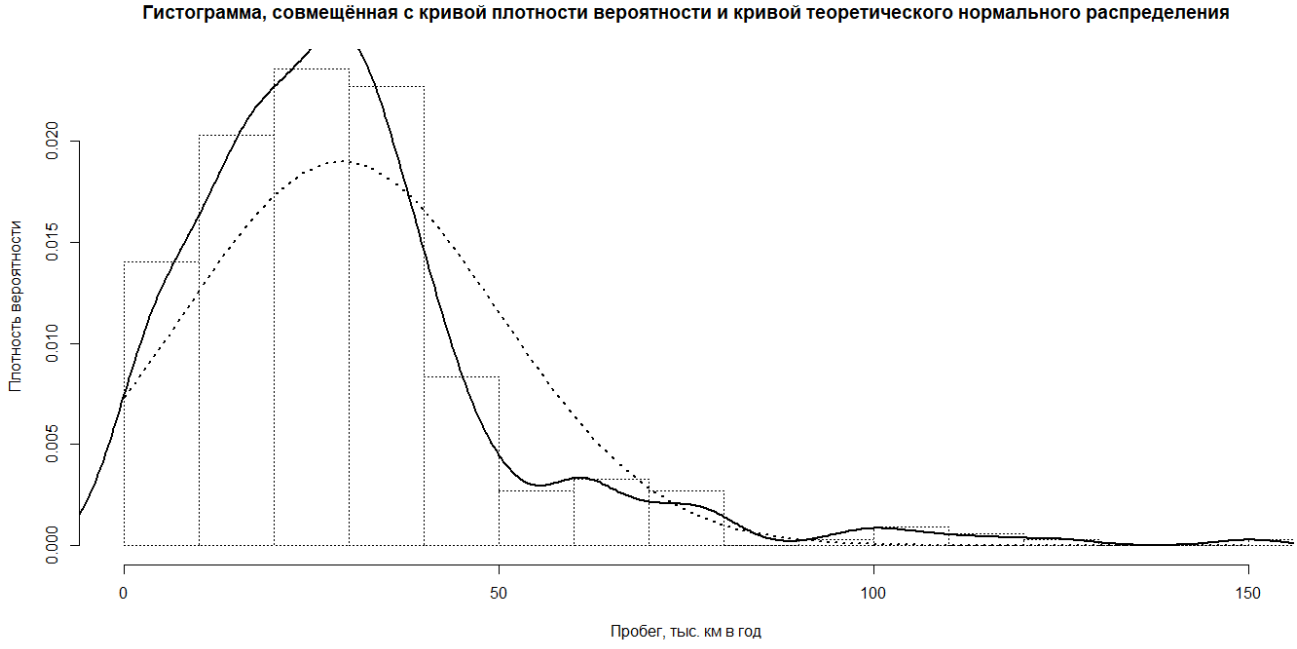

Рисунок 15: Гистограмма значений среднегодового пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

Ниже приводится код, вызывающий построение вышеприведённых гистограмм.

```
  hist(gazelle.new.01$Price, breaks = 7, freq = FALSE, xlab = "Цена предложения нововогo автомобиляб руб. РФ с учётом НДС", ylab = "Плотность вероятности", main = "Гистограмма цен предложений новых автомобилей, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений цен новых автомобилей

  lines(density(gazelle.new.01$Price), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$Price, breaks = 10, freq = FALSE, xlab = "Цена предложения автомобиля с пробегом, руб. РФ с учётом НДС", ylab = "Плотность вероятности", main = "Гистограмма цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений цен б/у автомобилей

  lines(density(gazelle.old.01$Price), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$Age, breaks = 6, freq = FALSE, xlab = "Возраст автомобиля, лет", ylab = "Плотность вероятности", main = "Гистограмма значений возраста автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений возраста

  lines(density(gazelle.old.01$Age), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$Mileage, breaks = 10, freq = FALSE, xlab = "Пробег автомобиля, тыс. км", ylab = "Плотность вероятности", main = "Гистограмма значений пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений пробега

  lines(density(gazelle.old.01$Mileage), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$MpY, breaks = 13, freq = FALSE, xlab = "Среднегодовой пробег автомобиля, тыс. км", ylab = "Плотность вероятности", main = "Гистограмма значений среднегодового пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений среднегодового пробега

  lines(density(gazelle.old.01$MpY), col = "blue", lwd = 2) #добавляем кривую плотности вероятностиhist(gazelle.new.01$Price, breaks = 7, freq = FALSE, xlab = "Цена предложения нововогo автомобиляб руб. РФ с учётом НДС", ylab = "Плотность вероятности", main = "Гистограмма цен предложений новых автомобилей, совмещённая с
```

```
кривой плотности вероятности") #создаём гистограмму для значений цен новых автомобилей

  lines(density(gazelle.new.01$Price), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$Price, breaks = 10, freq = FALSE, xlab = "Цена предложения автомобиля с пробегом, руб. РФ с учётом НДС", ylab = "Плотность вероятности", main = "Гистограмма цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений цен б/у автомобилей

  lines(density(gazelle.old.01$Price), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$Age, breaks = 6, freq = FALSE, xlab = "Возраст автомобиля, лет", ylab = "Плотность вероятности", main = "Гистограмма значений возраста автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений возраста

  lines(density(gazelle.old.01$Age), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$Mileage, breaks = 10, freq = FALSE, xlab = "Пробег автомобиля, тыс. км", ylab = "Плотность вероятности", main = "Гистограмма значений пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений пробега

  lines(density(gazelle.old.01$Mileage), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$MpY, breaks = 13, freq = FALSE, xlab = "Среднегодовой пробег автомобиля, тыс. км", ylab = "Плотность вероятности", main = "Гистограмма значений среднегодового пробега автомобилей с пробегом, совмещённая с кривой плотности вероятности") #создаём гистограмму для значений среднегодового пробега

  lines(density(gazelle.old.01$MpY), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  histDist(gazelle.new.01$Price, density = TRUE, nbins = 7, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Цена предложения, руб. с НДС", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab =
```

```
"black", col.axis = "black", xlim = c(1100, 1800)) #ещё один вариант гистограммы для цен новых машин

  histDist(gazelle.old.01$Price, density = TRUE, nbins = 10, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Цена предложения, руб. с НДС", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab = "black", col.axis = "black", xlim = c(400, 1800)) #ещё один вариант гистограммы для цен б/у машин

  histDist(gazelle.old.01$Age, density = TRUE, nbins = 6, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Возраст, лет", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab = "black", col.axis = "black", xlim = c(1, 7)) #ещё один вариант гистограммы возраста

  histDist(gazelle.old.01$Mileage, density = TRUE, nbins = 10, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Пробег, тыс. км", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab = "black", col.axis = "black", xlim = c(0, 400)) #ещё один вариант гистограммы пробега

  histDist(gazelle.old.01$MpY, density = TRUE, nbins = 13, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Пробег, тыс. км в год", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab = "black", col.axis = "black", xlim = c(0, 150)) #ещё один вариант гистограммы среднегодового пробега
```

4.1.3. Введение дополнительных переменных

Поскольку мы обладаем априорными знаниями об экспоненциальном характере нарастания износа, основанными на многочисленных исследованиях теоретиков и практиков оценки, часть из которых приведена в разделе 2.1. Цель и предмет исследования на стр. 23, мы сразу создадим вектор данных, содержащих сведения о значениях натуральных логарифмов цен. Несложно догадаться, что известная многим оценщикам формула определения значения износа, имеющая вид:

$$И_c = 1 - \exp(-\Omega) \quad , \tag{16}$$

где $И_с$ – величина совокупного износа;

Ω – параметр, характеризующий степень утраты полезности, являющийся функцией от переменных, описывающих некие физические свойства объекта (например возраст, пробег, условия эксплуатации и т. д.),

представляет собой частный случай операции потенцирования, обратной которой является операция логарифмирования. Таким образом, следует необходимость сразу же рассчитать значения натуральных логарифмов цен.

Данный подход является важной частью процесса анализа данных. Априорные знания не следует игнорировать. В основе такого суждения лежит т. е. «Байесовский подход», см. [51], [52], [53].

Заодно в данном разделе будет рассмотрен вопрос технологии присоединения векторов данным к таблицам.

Создадим два вектора данных, содержащих значения натуральный логарифмов цен предложений автомобилей как новых, так и с пробегом. В полученных векторах будет длинное и неочевидное название переменной == заголовка столбца. Изменим его на человекочитаемое интуитивно понятно. Важным дополнением являет то, что сделать это лучше именно на векторе до его присоединения к таблице. В R технология переименования заголовка столбцов, равно как многие технологии редактирования данных реализованы специфическим методом создания копии объекта, внесения изменений в копию и замены ей изначального объекта. Т.е. в некоторый момент времени объём оперативной памяти, занимаемый объектом увеличивается в два раза, что может привести к остановке обработки вследствие недостатка памяти. Поэтому переименование столбца в момент, пока он находится в меньшем объекте — векторе, является технологически обоснованным.

После переименования заголовка столбца == названия переменной объединяем объекты. Ниже приводится код.

```
colnames(ln.price.new) <- c("logPrice") #переименовываем столбец

colnames(ln.price.old) <- c("logPrice") #переименовываем столбец

gazelle.new.01 <- data.frame(gazelle.new.01, ln.price.new) #добавляем столбец с логарифмами == добавляем вектор к таблице

gazelle.old.01 <- data.frame(gazelle.old.01, ln.price.old) #добавляем столбец с логарифмами == добавляем вектор к таблице

remove(ln.price.new) #удаляем уже не нужный вектор

remove(ln.price.old) #удаляем уже не нужный вектор
```

Построим гистограммы для значений логарифмов цен.

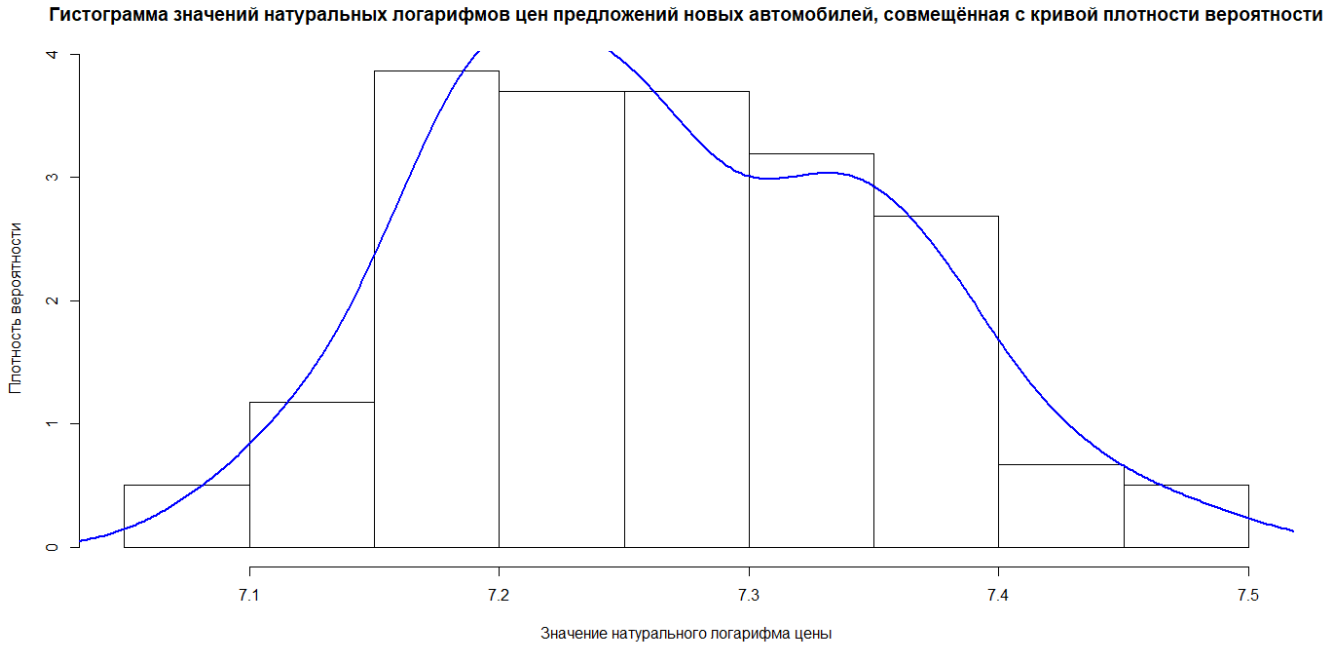

Рисунок 16: Гистограмма значений натуральных логарифмов цен предложений новых автомобилей, совмещённая с кривой плотности вероятности

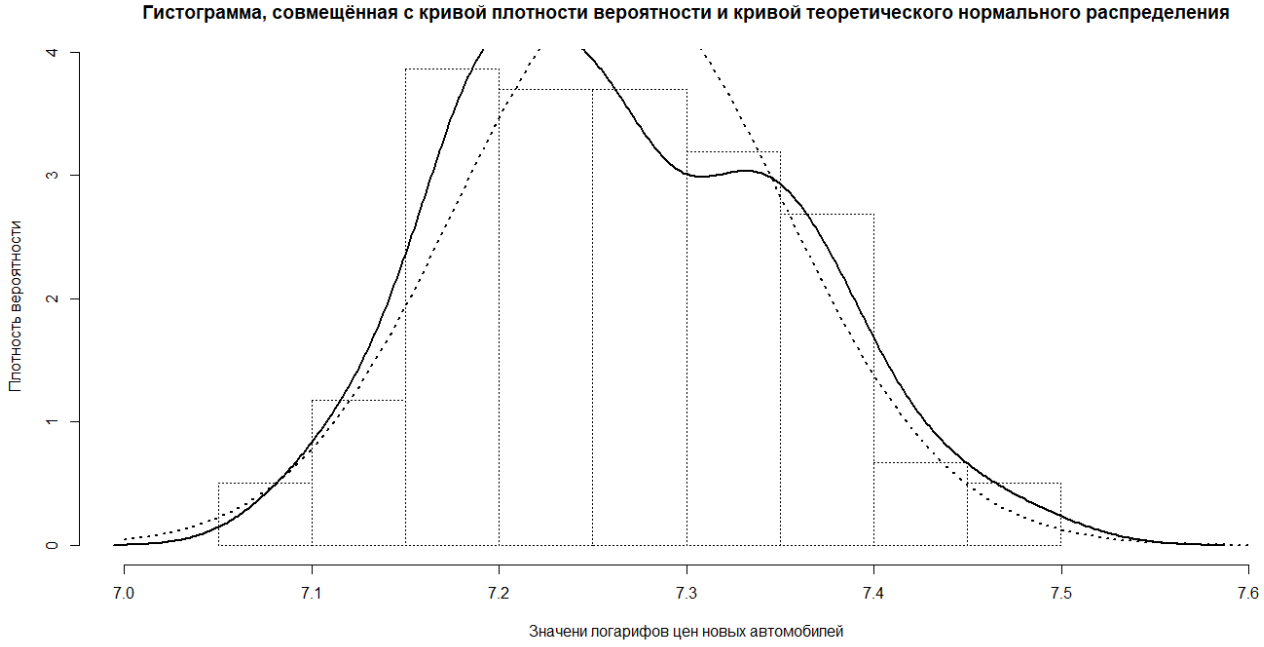

Рисунок 17: Гистограмма значений натуральных логарифмов цен предложений новых автомобилей, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

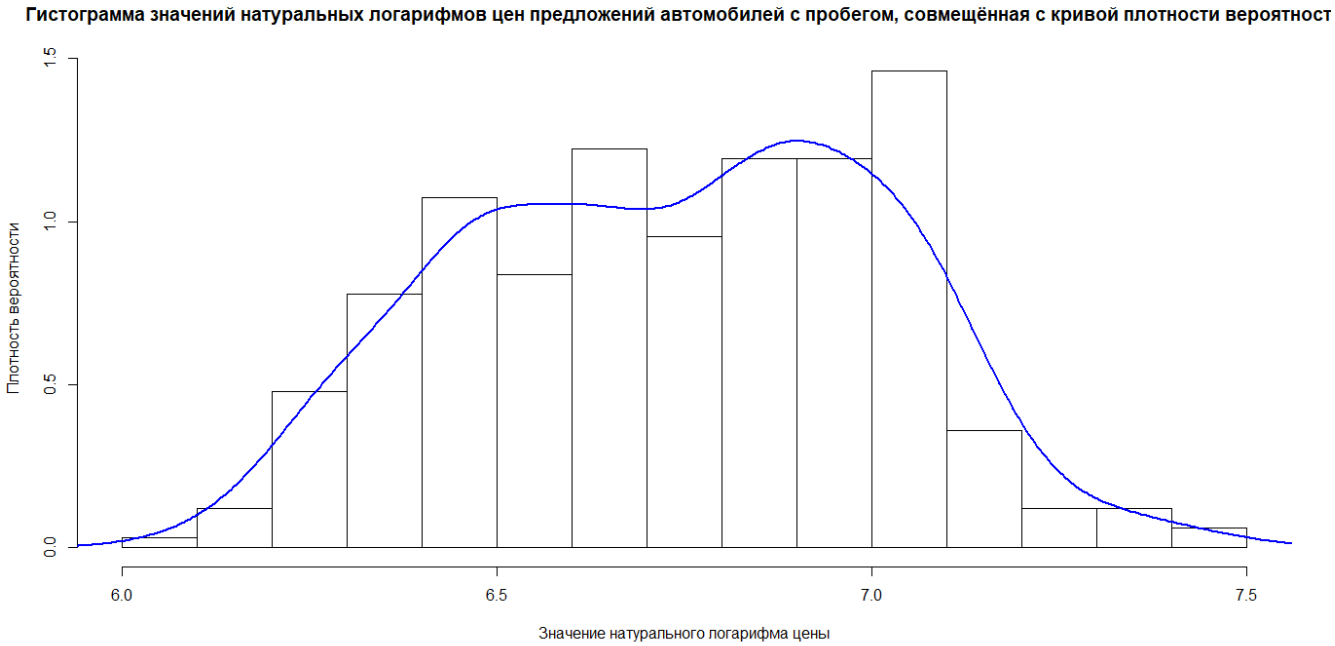

Рисунок 18: Гистограмма значений натуральных логарифмов цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности

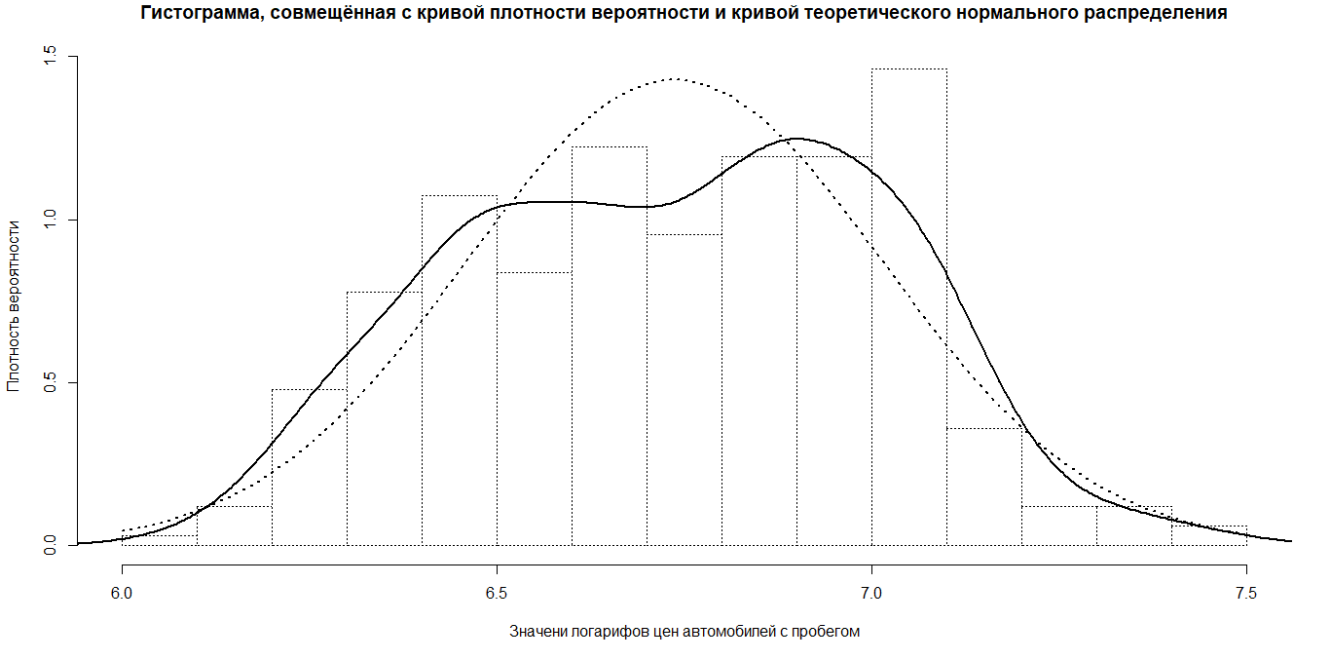

Рисунок 19: Гистограмма значений натуральных логарифмов цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения

```
  hist(gazelle.new.01$logPrice, breaks = 7, freq = FALSE, xlab = "Значение натурального логарифма цены", ylab = "Плотность вероятности", main = "Гистограмма значений натуральных логарифмов цен предложений новых автомобилей, совмещённая с кривой плотности вероятности") #строим гистограмму логарифмов цен новых автомобилей

  lines(density(gazelle.new.01$logPrice), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$logPrice, breaks = 10, freq = FALSE, xlab = "Значение натурального логарифма цены", ylab = "Плотность вероятности", main = "Гистограмма значений натуральных логарифмов цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности") #строим гистограмму логарифмов цен новых автомобилейhist(gazelle.new.01$logPrice, breaks = 7, freq = FALSE, xlab = "Значение натурального логарифма цены", ylab = "Плотность вероятности", main = "Гистограмма значений натуральных логарифмов цен предложений новых автомобилей, совмещённая с кривой плотности вероятности") #строим гистограмму логарифмов цен новых автомобилей

  lines(density(gazelle.new.01$logPrice), col = "blue", lwd = 2) #добавляем кривую плотности вероятности

  hist(gazelle.old.01$logPrice, breaks = 10, freq = FALSE, xlab = "Значение натурального логарифма цены", ylab = "Плотность вероятности", main = "Гистограмма значений натуральных логарифмов цен предложений автомобилей с пробегом, совмещённая с кривой плотности вероятности") #строим гистограмму логарифмов цен новых автомобилей

  histDist(gazelle.new.01$logPrice, density = TRUE, nbins = 7, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Значени логарифов цен новых автомобилей", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab = "black", col.axis = "black", xlim = c(7.0, 7.6)) #ещё один вариант гистограммы логарифмов цен новых автомобилей

  histDist(gazelle.old.01$logPrice, density = TRUE, nbins = 10, main = "Гистограмма, совмещённая с кривой плотности вероятности и кривой теоретического нормального распределения", xlab = "Значени логарифов цен автомобилей с пробегом", ylab = "Плотность вероятности", col.hist = "white", border.hist = "black", fg.hist = "black", line.wd = 2, line.ty = c (3, 1), line.col = c (1, 1), col.main = "black", col.lab = "black", col.axis = "black", xlim = c(6.0, 7.5)) #ещё один вариант гистограммы логарифмов цен новых автомобилей
```

4.2. Описательные статистики

Перед началом процедуры создания описательных статистик проведём проверка целостности и корректности данных.

Команда dim даёт нам возможность проверить размерность объекта. Проверяем количество строк и столбцов. Команда str показывает нам структуру данных по каждой переменной, тип данных, знак, отделяющий целую часть от дробной.

Посмотрим. В объекте gazelle.new.01 4 столбца == переменных и 119 строк == наблюдений. Всё правильно. Типы данных: num (числовой) для Price и logPrice и integer (целочисленный). Умница R определил, что в переменной все числа целые и для экономии памяти присвоил тип, требующий меньше памяти, и меньше нагружающий процессор, чем тип num, предполагающий наличие плавающей точки.

В объекте gazelle.old.01 7 столбцов == переменных и 335 строк == наблюдений. Также всё корректно. Типы данных: num (числовой) для Price, Age, Mileage, MpY, logPrice, integer (целочисленный) для Year, Factor (фактор) для Region. Поскольку в нашем исследовании возраст принимается в целых годах без детализации вследствие отсутствия возможности получения таких данных по наблюдениям с открытого рынка, нет необходимости сохранять возможность использования плавающей точки, а значит можно присвоить данным этой переменной тип integer для оптимизации работы. Это уже инженерный подход программиста. Хотя данное Руководство не преследует цели рассмотрения вопросов software engineering, а объём используемых данных ничтожно мал по сравнению с возможностями вычислительных машин по состоянию на 2019 год, всё же представляется правильным изначально давать некоторые рекомендации по правильному подходу к вопросам разработки кода. Для изменения типа данных используем команду as.integer. Также сейчас мы впервые сталкиваемся с таким типом данных как Factor. Подробное описание этого типа и его значения для Data mining будет приведено в соответствующем разделе. Пока что можно сказать только, что понятие «фактор» в Data mining не соотносится с известным в оценке понятием «ценообразующий фактор». С точки зрения оценщика факторами являются все переменные, влияющие на стоимость, с точки зрения Data mining факторами являются строго определённые переменные. Об это будет подробно рассказано в соответствующем разделе. Ниже приведён код для описанных действий.

```
  dim(gazelle.new.01) #проверям размерность объекта

  dim(gazelle.old.01) #проверям размерность объекта

  str(gazelle.new.01) #проверяем структуру данных на типы данных и ошибки

  str(gazelle.old.01) #проверяем структуру данных на типы данных и ошибки

  gazelle.old.01$Age <- as.integer(gazelle.old.01$Age) #изменение типа данных переменной

  str(gazelle.old.01) #проверяем изменённую структуру данных на типы данных и ошибки
```

Результаты проверок выводятся в консоли.

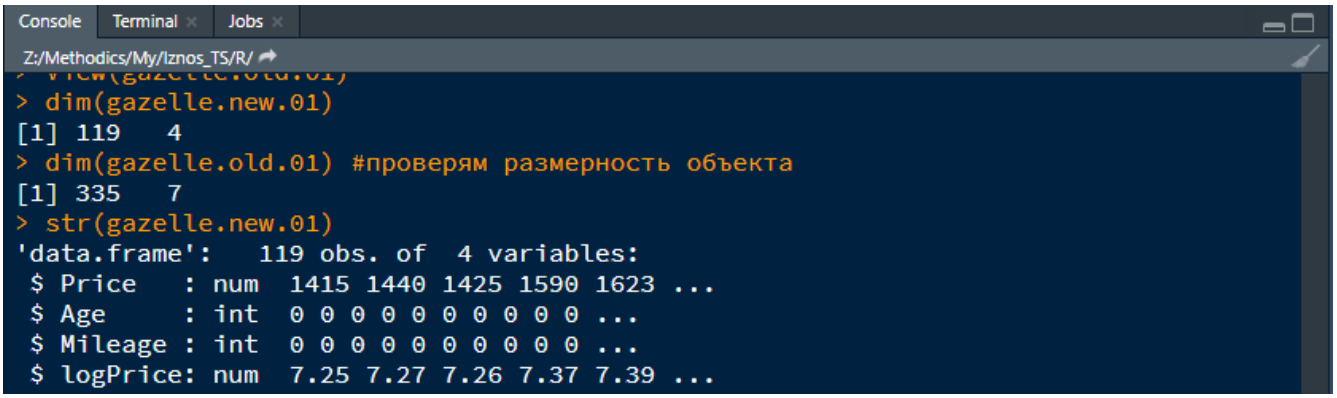

Рисунок 20: Результаты проверки размерности и типа данных

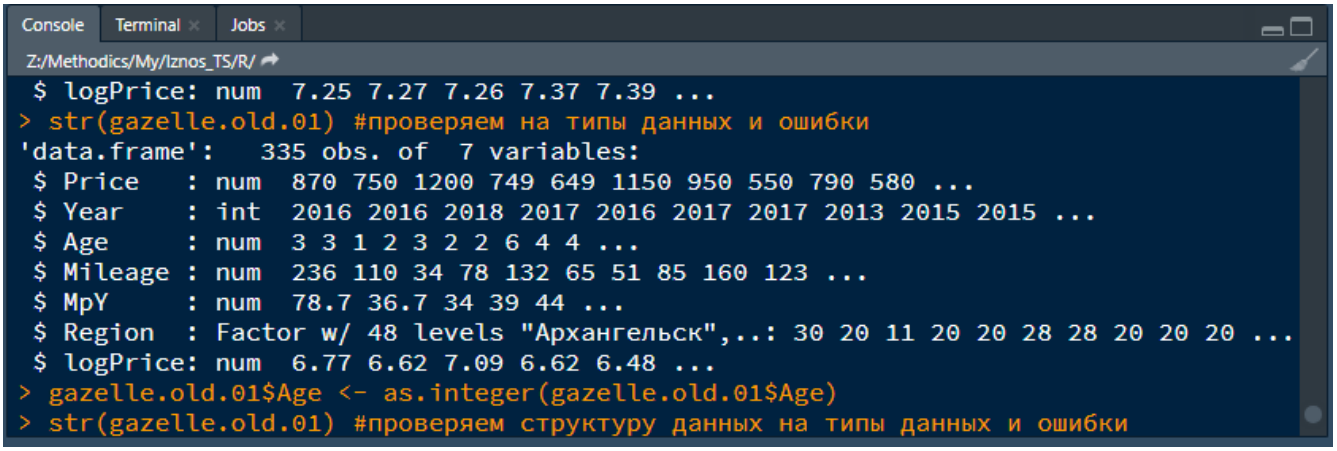

Рисунок 21: Результаты проверки типов данных, изменение типа данных

Следует обратить внимание на запись описания типа данных переменной Region. Помимо указания типа данных переменной Factor приводится показатель levels, возвращающий количество возможных значений фактора. В нашем случае значение 48, это означает, что всего есть 48 различных регионов, в которых есть предложения о продаже рассматриваемой модели автомобилей.

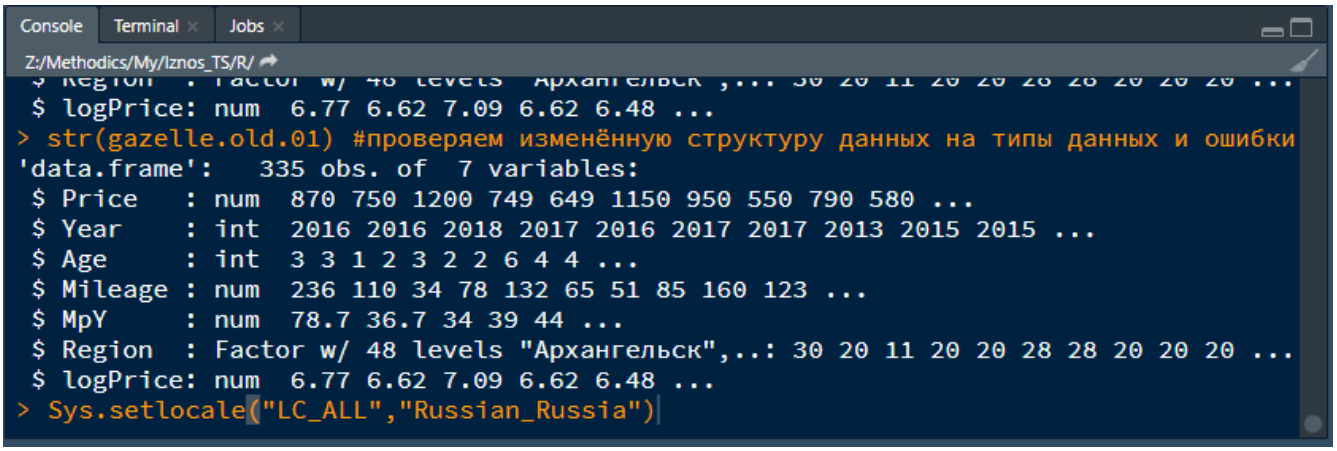

Рисунок 22: Результаты проверки типов данных объекта gazelle.old.01 после изменения типа данных переменной Age с nit на integer

Рассчитаем основные описательный статистики. Для самых базовых используем команду summary. Она выдаст немного статистики.

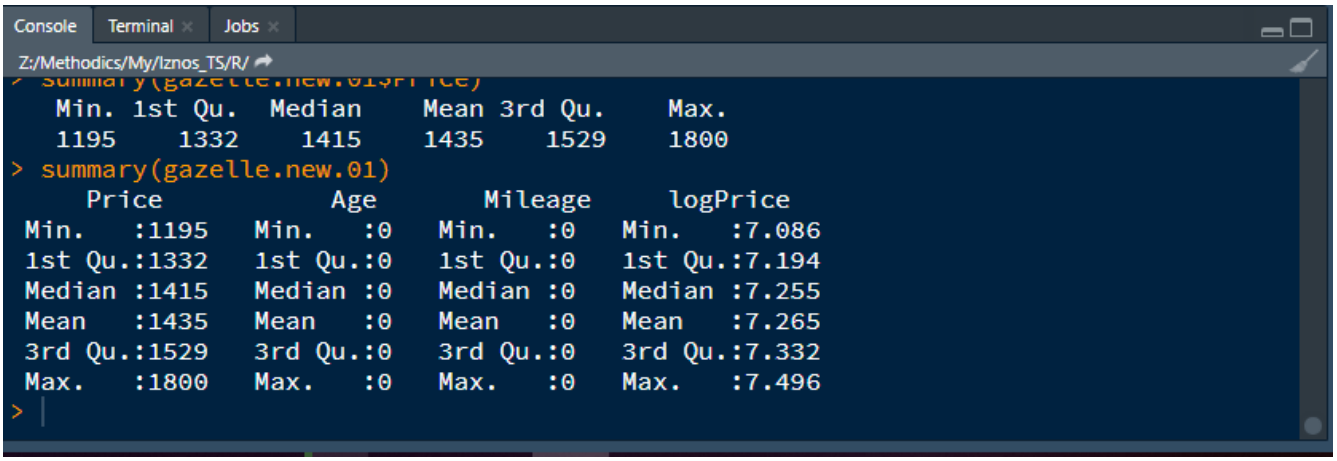

Рисунок 23: Базовые описательные статистики для для объекта gazelle.new.01

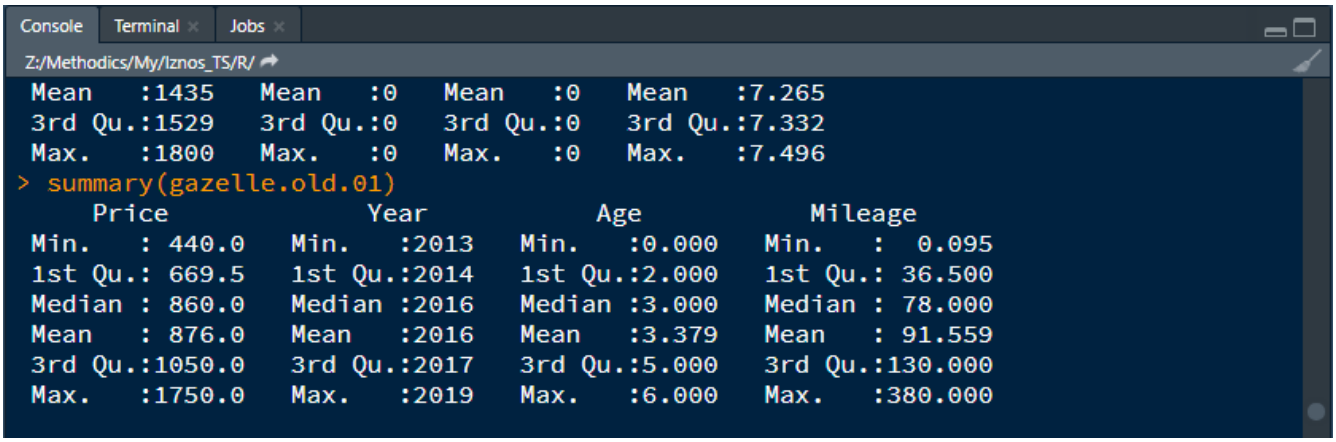

Рисунок 24: Базовые описательные статистики для объекта gazelle.old.01

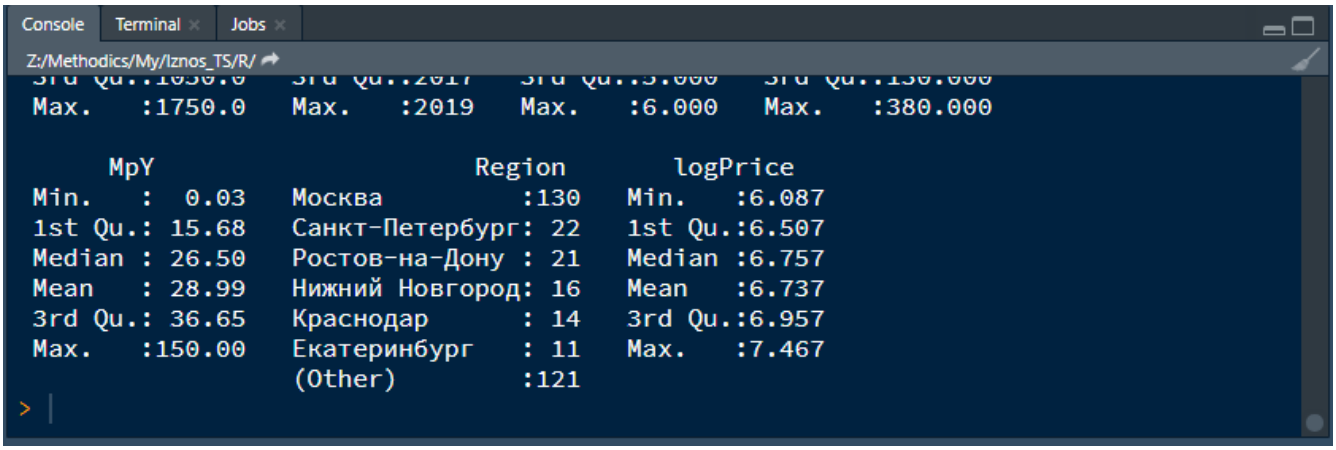

Рисунок 25: Базовые описательные статистики для объекта gazelle.old.01 (продолжение)

Как видим, для числовых данных функция выдаёт значения минимального, максимального, среднего арифметического, а также квартили. Для набора данных типа «фактор» подсчитывает количество встречаемых значений фактора. Как видим Москва далеко впереди по количеству предложений.

Данная функция даёт только самые основные статистики. Если мы хотим получить дополнительные, то придётся сделать это своими руками. Создадим собственную функция summaryplus, которая будет вычислять всё, что нам интересно. Почему необходимо создать функцию? Почему просто не посчитать? Преимущество функции в том, что её можно использовать многократно для различных объектов. Создав один раз, ей можно пользоваться в дальнейшем.

```
summaryplus <- function(x, na.omit = FALSE) { #создаём функцию
  minimum <- min(x) #определяем минимальное значение
  quartil.01 <- quantile(x, 0.25) #определяем значение 1-го квартиля
```

```
    mediana <- quantile(x, 0.5) #определяем значение 2-го квартиля == медиану
    srednee <- mean(x) #определяем среднее арифметическое
    quartil.03 <- quantile(x, 0.75) #определяем значение 3-го квартиля
    maximum <- max(x) #определяем максимальное значение
      trim.95 <- mean(x, trim = 0.025) #определяем усечённое 5% среднее арифметическое
      trim.68 <- mean(x, trim = 0.16) #определяем усечённое 32% среднее арифметическое
    stderror <- std.error(x) #выччисляем стандартную ошибку
    disp.01 <- var(x) #выччисляем выборочную дисперсию
    standotklon <- sd(x) #выччисляем стандартное отклонение
    k.var <- sd(x)/mean(x) #выччисляем коэффициент вариации
    sk <- skewness(x) #выччисляем ассиметрию
    kurt.01 <- kurtosis(x) #выччисляем эксцесс
    kontrkurt <- 1/(sqrt(kurtosis(x))) #выччисляем контрэксцесс
    razmakh <- max(x)-min(x) #выччисляем диапазон
        return(c(минимум=minimum, перв.квартиль=quartil.01, медиана=mediana, среднее=srednee, трет.квартиль=quartil.03, максимум=maximum, усеч.среднее.95=trim.95, усеч.среднее.68=trim.68, станд.ошибка=stderror, выб.дисперсия=disp.01, станд.отклон=standotklon, коэф.вариации=k.var, ассиметрия=sk, эксцесс=kurt.01, контрэксцесс=kontrkurt, диапазон=razmakh)) #вычисляем диапазон указываем отображаемые имена показателей
  } #конец функции
  options(digits = 4) #указываем число знаком после точки (влияет только на отображение, расчёты всегда с максимальной точностью)
  sapply(gazelle.old.01, summaryplus) #применяем функцию к объекту
```

Отметим следующие моменты:

1) Усечённое среднее определяется таким образом, что отсекаются 5% крайних значений, т. е. по 2.5% значений переменных, имеющие наименьшее и наибольшее значение. В этом случае вычисление среднего осуществляется по 95% значений. Затем определяется усечённое среднее по 68% значений, т. е. отсекается 32% крайних значений переменных по 16% наименьших и наибольших. Данный подход Автора основан на свойствах стандартного нормального распределения, одним из которых является то, что 68%

значений наблюдений находятся на расстоянии 1 стандартного отклонения от математического ожидания, а 95% значений наблюдений находятся на расстоянии 2 стандартных отклонений от математического ожидания. Таким образом, предложенный вариант отсечения крайних значений может служить первичным нечётким способом оценки соответствия распределения значений переменной критерию нормальности. В случае приблизительного равенства среднего арифметического и его усечённых 5%, 32% значений можно с некоторой степенью приближения ожидать нормальность распределения данных. Следует отметить, что **данный метод является сугубо разведочным и ни в каких случаях не может заменять собой стандартные тесты проверки нормальности**, которые будут рассмотрены подробно в соответствующем разделе. В повседневной практике аналитик может использовать любые меры отсечения части значений.

2) При определении дисперсии речь идёт о т. н. выборочной дисперсии.

3) Поскольку мы имеем место с выборочной дисперсией, стандартная ошибка определяется по формуле:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \quad , \tag{17}$$

где $SE_{\bar{x}}$= (standard error) – искомое значение стандартной ошибки

s – стандартное отклонение значений переменной x на основе несмещённой [54] оценки её выборочной дисперсии

n – количество наблюдений в выборке

4) При определении стандартного отклонения и стандартной ошибки используется несмещённая [54] оценка первой.

Таблица 6: Основные описательные статистики значений переменных

№ пп	Показатель	Цена предложения, тыс. руб.	Натуральный логарифм цены предложения	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Цена предложения новых, тыс. руб.	Натуральный логарифм цены предложения
	Объект	gazelle.old.01	gazelle.old.01	gazelle.old.01	gazelle.old.01	gazelle.old.01	gazelle.new.01	gazelle.new.01
	Имя переменной	Price	logPrice	Age	Mileage	MpY	Price	logPrice
1		2		3	4	5	6	
1	Минимум						1195.0000	7.0860
2	1-й квартиль						1332.0000	
3	Медиана						1415.0000	
4	Среднее арифметическое						1435.0000	

№ пп	Показатель	Цена предложения, тыс. руб.	Натуральный логарифм цены предложения	Возраст, лет	Пробег, тыс. км	Средний пробег в год, тыс. км.	Цена предложения новых, тыс. руб.	Натуральный логарифм цены предложения
	Объект	gazelle.old.01	gazelle.old.01	gazelle.old.01	gazelle.old.01	gazelle.old.01	gazelle.new.01	gazelle.new.01
	Имя переменной	Price	logPrice	Age	Mileage	MpY	Price	logPrice
5	3-й квартиль						1529.0000	
6	Максимум						1800.0000	
7	Усечённое среднее, 5%						1432.6750	
	Усечённое среднее, 32%						1424.7650	7.26030
	Стандартная ошибка						11.7594	0.0081
	Дисперсия (выборочная)							
	Стандартное отклонение						128.2794	
	Коэффициент вариации						0.0894	0.0121
	Ассиметрия							
	Эксцесс							
	Контрэксцесс							
	Интервал							

Возможно некоторым пользователям Руководства знакомы не все вышеприведённые статистические показатели. Пробежимся ним.

1) **Минимум**. Здесь всё очевидно. Минимальное значение из всех значений переменной.
2) **Квартили**. Здесь немного сложнее. Начнём с более общего понятия «квантиль». Более строгой формой записи является «квантиль уровня p» [55] .

Приведём формульную запись этого понятия.

$$x(\lfloor np+1 \rfloor) \quad , \tag{18}$$

Предположим, что у нас есть вектор, содержащий числовые данные от x_1 до x_n. Упорядочим данные от по возрастанию. Получим новую структуру данных вектора, которую иногда называют «вариационный ряд». В этом случае x_1 имеет минимальное значение, x_n – максимальное. Этому объяснению соответствую круглые скобки. Значение скобок вида ⌋ должно быть понятно в первую очередь тем, кто знаком с программирование. Их значение аналогично команде floor. Т.е. их смысл можно описать как «наибольшее целое меньшее данного». Т.е. ⌊4.

$8\rfloor == 4$, $\lfloor 27.59 \rfloor == 27$, $\lfloor -8.11 \rfloor == -9$. Соответственно квантилем является число, получаемое как:

$$p * 100\% \quad , \tag{19}$$

где p — заданный уровень

Соответственно, медиана — это квантиль уровня 0.5, т. е. то наблюдение, по отношению к которому ½ наблюдений имеет меньшее значение, а ½ — большее. В случае чётного количества наблюдений в качестве медианы принимается среднее арифметическое двух наблюдений, ближайших к уровню p = ½. Первым квартилем является квантиль уровня ¼, вторым — уже упомянутая медиана, третьим — квантиль уровня ¾. Иногда первый квартиль называет нижним квартилем, а третий — верхним.

Следует сказать, что медиана, как и квартили является примером робастной оценки центральной тенденции. В отличие, например, от более распространённого среднего арифметического. О том, что такое робастность будет сказано в дальнейшем в соответствующем разделе.

3) **Среднее арифметическое**. [56]
4) **Максимум.**
5) **Усечённое среднее.**
6) **Выборочная дисперсия.**
7) **Стандартное отклонение.**
8) **Стандартная ошибка.**
9) **Коэффициент вариации.**
10) **Ассиметрия.** [57]
11) **Эксцесс.** [58]
12) **Контрэсцесс.**
13) **Интервал.**

4.3. Проверка гипотезы о нормальности распределения значений переменных

4.3.1. Здесь будет общий текст про распределения.

4.3.2. Здесь будет текст про нормальное распределение.

4.3.3. Здесь будет общий текст про проверку нормальности.

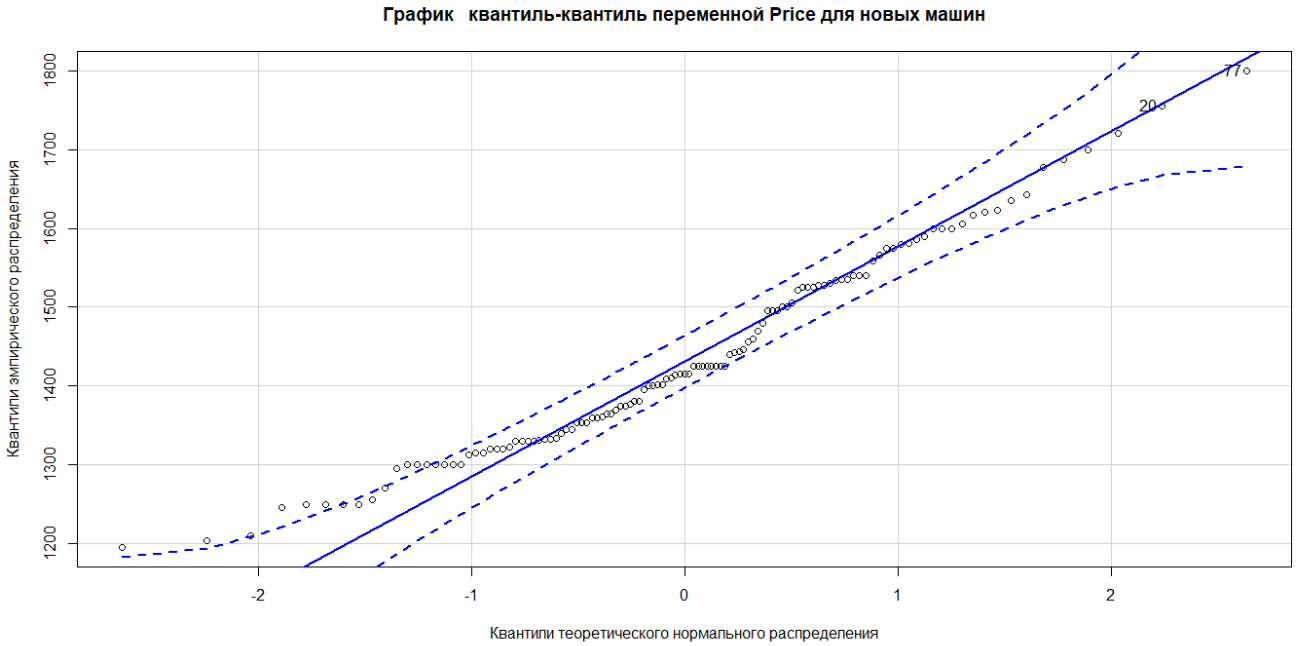

Рисунок 26: График квантиль-квантиль переменной Price для новых машин

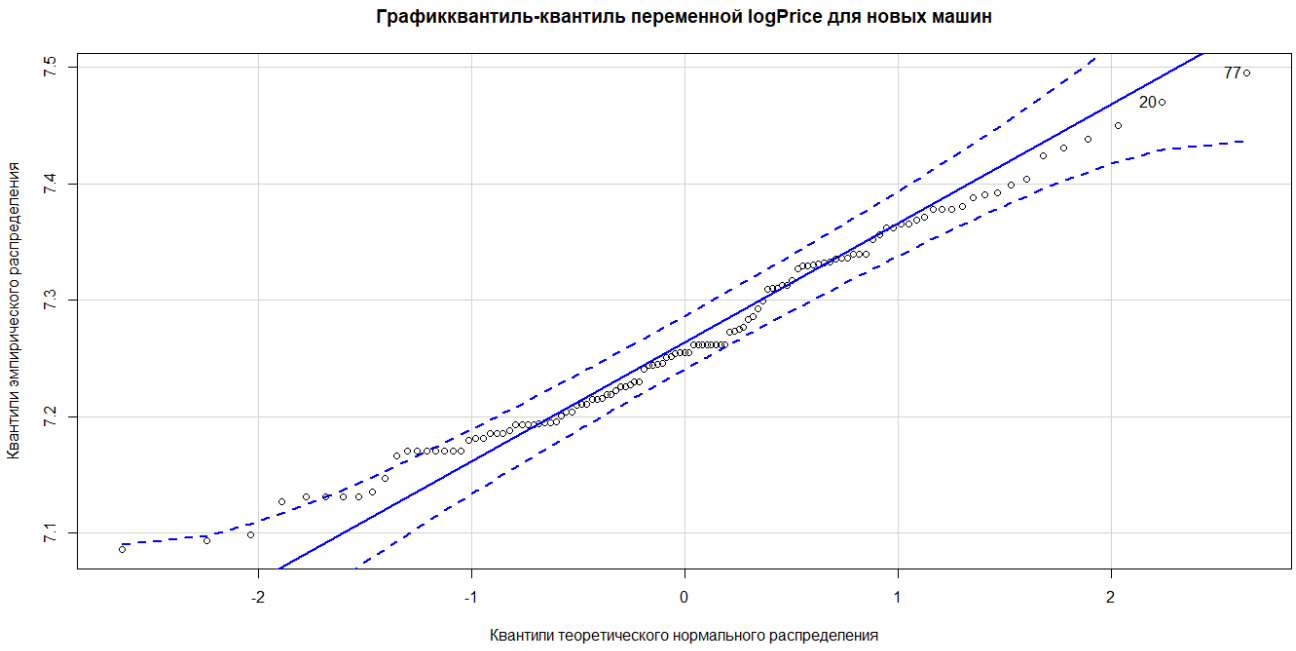

Рисунок 27: График квантиль-квантиль переменной logPrice для новых машин"

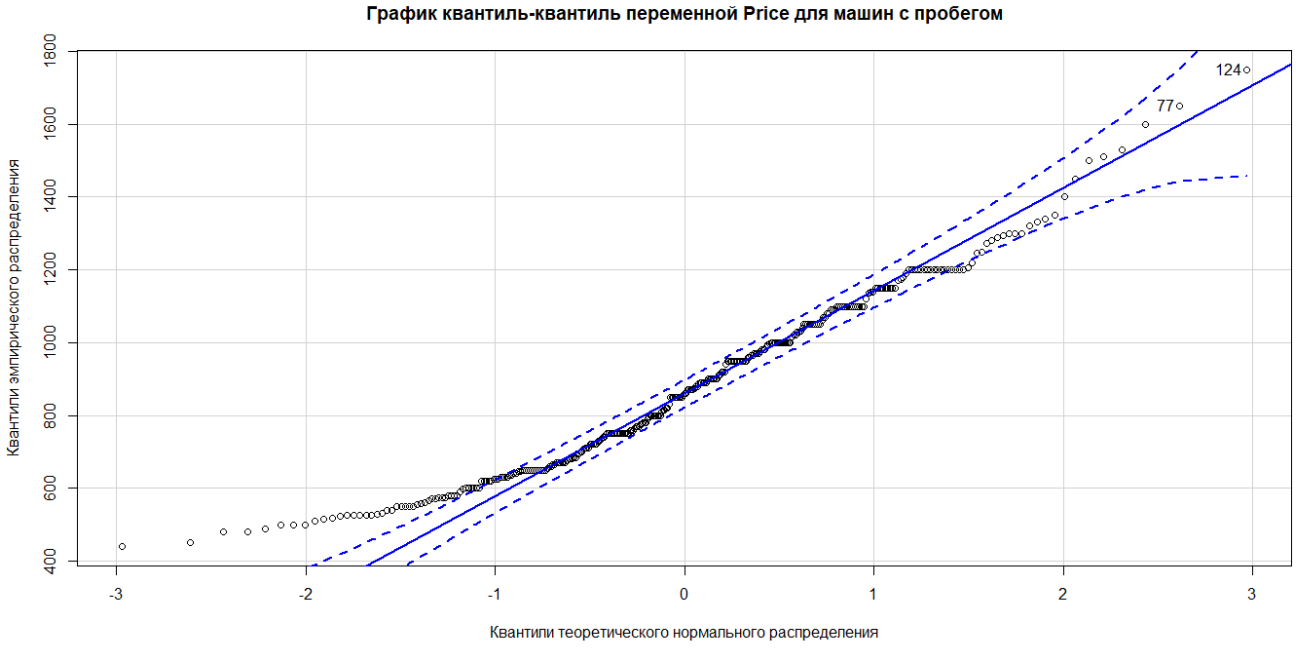

Рисунок 28: График квантиль-квантиль переменной Price для машин с пробегом

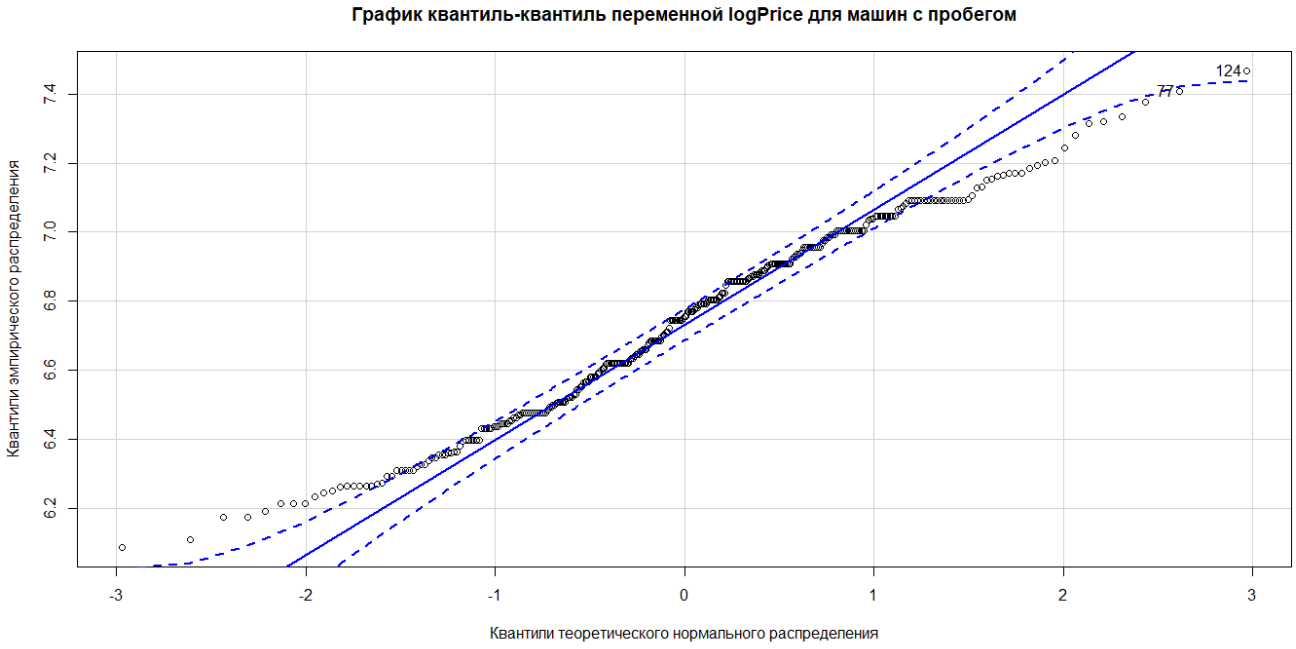

Рисунок 29: График квантиль-квантиль переменной logPrice для машин с пробегом

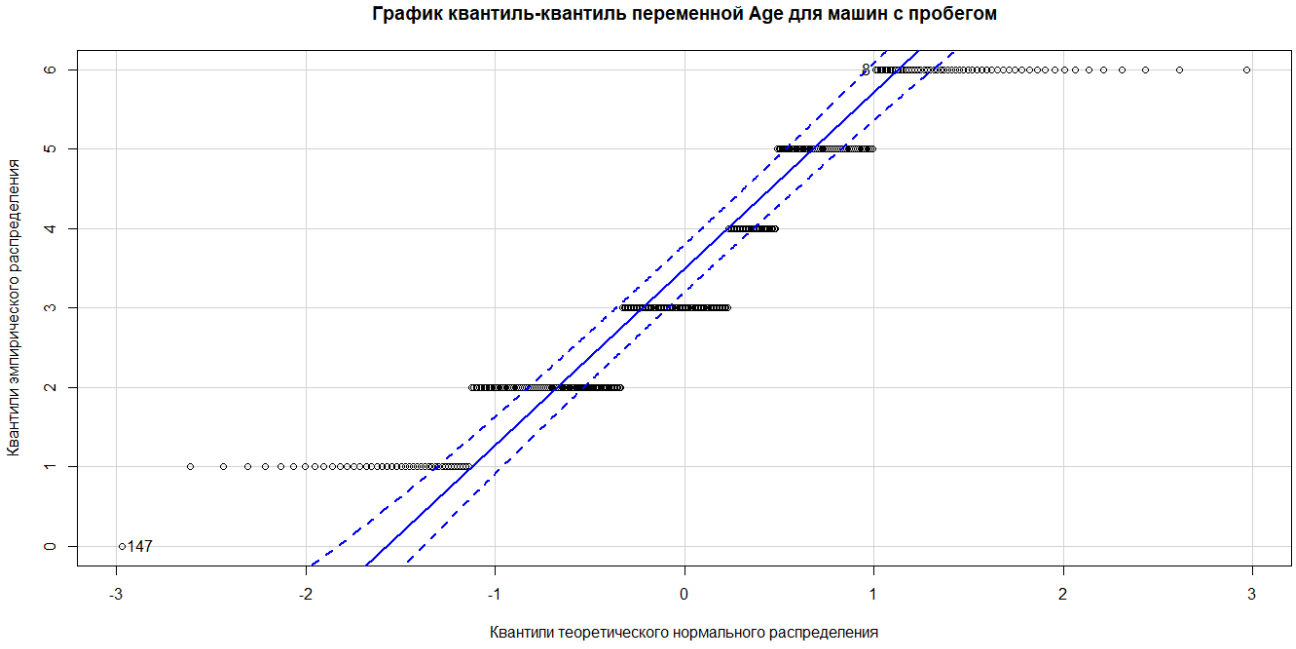

Рисунок 30: График квантиль-квантиль переменной Age для машин с пробегом

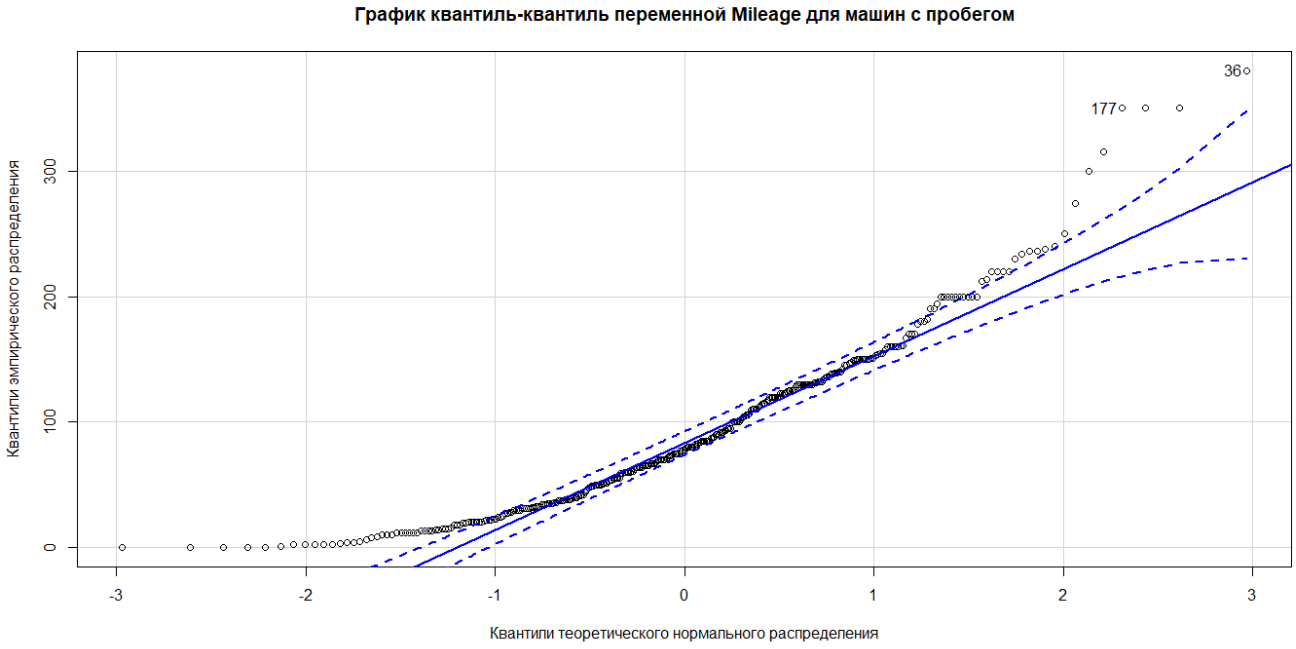

Рисунок 31: График квантиль-квантиль переменной Mileage для машин с пробегом

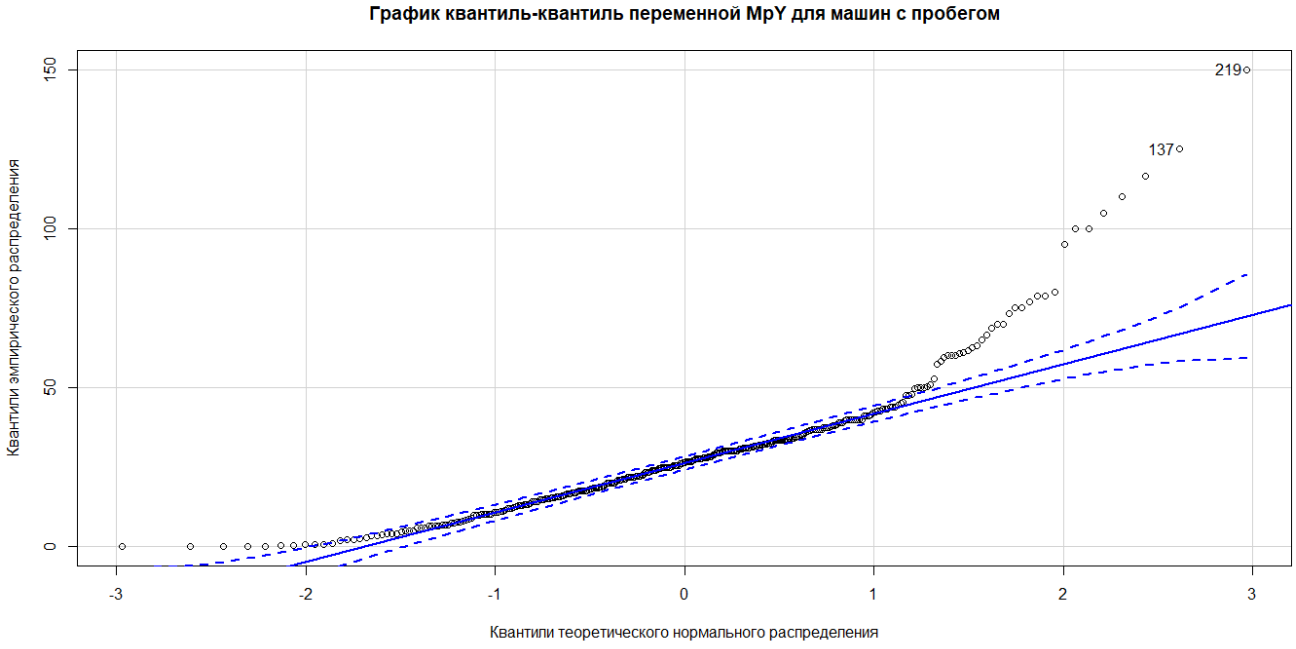

Рисунок 32: График квантиль-квантиль переменной MpY для машин с пробегом

К достоинствам графического метода анализа соответствия эмпирического распределения теоретическом можно отнести наглядность. К недостаткам — то, что они ничего не объясняют и не доказывают. Смотря на график можно разве что сделать предположение, но оно не может служить основанием для выводов. Следует использовать точные критерии, которые будут рассмотрены далее в Разделе Ошибка: источник перекрёстной ссылки не найден на стр. Ошибка: источник перекрёстной ссылки не найден.

4.3.4. Здесь будет общий текст про критерии проверки нормальности.

4.3.4.1. Критерий Шапиро — Уилка

4.3.4.2. Критерий Колмогорова-Смирнова

4.3.4.3. Критерий Андерсона-Дарлинга

4.3.4.4. Криткрий Крамера — фон Мизеса

4.3.4.5. Критерий Колмогорова-Смирнова-Лиллефорса

4.3.4.6. Критерий χ2 Пирсона

4.3.4.7 Критерий Шапиро - Франчия

4.3.4.8 Критерий Д'Агостино

4.3.4.9 Критерий Бонетта – Сайера

4.3.4.10 Критерий Жарка - Бера

4.3.4.11 Робастный критерий

4.3.4.12. Экспериментальный критерий, он не описан в литературе, я просто тренируюсь.

4.3.4.13. Обобщение данных

Таблица 7: Сравнительная таблица мощности критериев проверки нормальности распределения случайных величин

Наименование критерия	Характер альтернативного распределения					Ранг
	ассиметричное		симметричное		≈ нормальное	
	$\alpha_4<3$	$\alpha_4>3$	$\alpha_4<3$	$\alpha_4>3$	$\alpha_4\approx3$	
1	2		3		4	5
Критерий Шапиро-Уилка	1	1	3	2	2	1
Критерий K^2	7	8	10	6	4	2
Критерий Дарбина-Уотсона	11	7	7	15	1	3
Критерий Д'Агостино	12	9	4	5	12	4
Критерий α_4	14	5	2	4	18	5
Критерий Васичека	2	14	8	10	10	6
Критерий Дэвида-Хэртли-Пирсона	21	2	1	9	1	7
Критерий χ^2	9	20	9	8	3	8
Критерий Андресона-Дарлинга	18	3	5	18	7	9
Критерий Филлибена	3	12	18	1	9	10
Критерий Колмогорова-Смирнова-Лиллиефорса	16	10	6	16	5	11
Критерий Мартинеса-Иглевича	10	16	13	3	15	12
Критерий Лина-Мудхолкэра	4	15	12	12	16	13
Критерий α_3	8	6	21	7	19	14
Критерий Шпигельхальтера	19	13	11	11	8	15
Критерий Саркади	5	18	15	14	13	16
Критерий Смирнова-Крамера-фон Мизеса	17	11	20	17	6	17
Критерий Локка-Спурье	13	4	19	21	17	18
Критерий Оя	20	17	14	13	14	19
Критерий Хегази-Грина	6	19	16	19	21	20
Критерий Муроты-Такеучи	15	21	17	20	20	21

4.3.5 Собственно тесты

Таблица 8: Результаты проверки нормальности значений переменной Price для новых автомобилей

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка	0.01	<	0.05	Распределение нормально	Отклоняется

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерия Колмогорова-Смирнова	$<2*10^{-16}$	<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга	0.004	<	0.05	Распределение нормально	Отклоняется
Критерий Крамера - фон Мизеса	0.003	<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса	0.001	<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона	0.0008	<	0.05	Распределение нормально	Отклоняется
Критерий Шапиро - Франчиа	0.02	<	0.05	Распределение нормально	Отклоняется
Критерий Д'Агостино	0.03	<	0.05	Распределение нормально	Отклоняется
Критерий Бонетта – Сайера	0.1	>	0.05	Распределение нормально	Не отклоняется
Критерий Жарка - Бера	0.075	>	0.05	Распределение нормально	Не отклоняется
Робастный критерий	0.02	<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий	$<2*10^{-16}$	<	0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений цен является **отличным от нормального**.

Таблица 9: Результаты проверки нормальности значений переменной logPrice для новых автомобилей

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка	0.07	>	0.05	Распределение нормально	Не отклоняется
Критерия Колмогорова-Смирнова	$<2*10^{-16}$	<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга	0.02	<	0.05	Распределение нормально	Отклоняется
Критерий Крамера - фон Мизеса	0.02	<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса	0.01	<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона	0.08	>	0.05	Распределение	Не

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
				нормально	отклоняется
Критерий Шапиро - Франчиа	0.1	>	0.05	Распределение нормально	Не отклоняется
Критерий Д'Агостино	0.1539	>	0.05	Распределение нормально	Не отклоняется
Критерий Бонетта – Сайера	0.1	>	0.05	Распределение нормально	Не отклоняется
Критерий Жарка - Бера	0.1966	>	0.05	Распределение нормально	Не отклоняется
Робастный критерий	-1.1	<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий	$<2*10^{-16}$	<	0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений логарифмов цен является **приблизительно нормальным**.

Таблица 10: Результаты проверки нормальности значений переменной Price автомобилей с пробегом

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка	1e-06	<	0.05	Распределение нормально	Отклоняется
Критерия Колмогорова-Смирнова	2e-16	<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга	2e-06	<	0.05	Распределение нормально	Отклоняется
Критерий Крамера - фон Мизеса	5e-05	<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса	1e-05	<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона	3e-07	<	0.05	Распределение нормально	Отклоняется
Критерий Шапиро - Франчиа	5e-06	<	0.05	Распределение нормально	Отклоняется
Критерий Д'Агостино	2e-04	<	0.05	Распределение нормально	Отклоняется
Критерий Бонетта – Сайера	0.004	<	0.05	Распределение нормально	Отклоняется

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Жарка - Бера	6e-04	<	0.05	Распределение нормально	Отклоняется
Робастный критерий	-2.6	<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий	2e-16		0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений цен является **отличным от нормального**.

Таблица 11: Результаты проверки нормальности значений переменной logPrice автомобилей с пробегом

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка		<	0.05	Распределение нормально	Отклоняется
Критерия Колмогорова-Смирнова		<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга		<	0.05	Распределение нормально	Отклоняется
Критерий Крамера - фон Мизеса		<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса		<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона		<	0.05	Распределение нормально	Отклоняется
Критерий Шапиро - Франчиа		<	0.05	Распределение нормально	Отклоняется
Критерий Д'Агостино		>	0.05	Распределение нормально	Не отклоняется
Критерий Бонетта – Сайера		<	0.05	Распределение нормально	Отклоняется
Критерий Жарка - Бера		<	0.05	Распределение нормально	Отклоняется
Робастный критерий		<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий			0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений логарифмов цен является **отличным от нормального**.

Таблица 12: Результаты проверки нормальности значений переменной Age автомобилей с пробегом

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка		<	0.05	Распределение нормально	Отклоняется
Критерия Колмогорова-Смирнова		<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга		<	0.05	Распределение нормально	Отклоняется
Критерий Крамера - фон Мизеса		<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса		<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона		<	0.05	Распределение нормально	Отклоняется
Критерий Шапиро - Франчиа		<	0.05	Распределение нормально	Отклоняется
Критерий Д'Агостино		>	0.05	Распределение нормально	Не отклоняется
Критерий Бонетта – Сайера		<	0.05	Распределение нормально	Отклоняется
Критерий Жарка - Бера		<	0.05	Распределение нормально	Отклоняется
Робастный критерий		<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий			0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений возраста является **отличным от нормального**.

Таблица 13: Результаты проверки нормальности значений переменной Mileage автомобилей с пробегом

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка		<	0.05	Распределение нормально	Отклоняется
Критерия Колмогорова-Смирнова		<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга		<	0.05	Распределение	Отклоняется

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
				нормально	
Критерий Крамера - фон Мизеса		<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса		<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона		<	0.05	Распределение нормально	Отклоняется
Критерий Шапиро - Франчиа		<	0.05	Распределение нормально	Отклоняется
Критерий Д'Агостино		>	0.05	Распределение нормально	Отклоняется
Критерий Бонетта – Сайера		<	0.05	Распределение нормально	Не отклоняется
Критерий Жарка - Бера		<	0.05	Распределение нормально	Отклоняется
Робастный критерий		<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий		<	0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений пробега является **отличным от нормального**.

Таблица 14: Результаты проверки нормальности значений переменной MpY автомобилей с пробегом

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро-Уилка		<	0.05	Распределение нормально	Отклоняется
Критерия Колмогорова-Смирнова		<	0.05	Распределение нормально	Отклоняется
Критерий Андерсона - Дарлинга		<	0.05	Распределение нормально	Отклоняется
Критерий Крамера - фон Мизеса		<	0.05	Распределение нормально	Отклоняется
Критерий Колмогорова - Смирнова-Лиллиефорса		<	0.05	Распределение нормально	Отклоняется
Критерий χ2 Пирсона		<	0.05	Распределение нормально	Отклоняется

Вид теста	Возвращённое значение p	Отношение	Заданный уровень значимости α	Нулевая гипотеза	Результат проверки нулевой гипотезы
1	2	3	4	5	6
Критерий Шапиро - Франчиа		<	0.05	Распределение нормально	Отклоняется
Критерий Д'Агостино		>	0.05	Распределение нормально	Отклоняется
Критерий Бонетта – Сайера		<	0.05	Распределение нормально	Отклоняется
Критерий Жарка - Бера		<	0.05	Распределение нормально	Отклоняется
Робастный критерий		<	0.05	Распределение нормально	Отклоняется
Экспериментальный критерий		<	0.05	Распределение нормально	Отклоняется

Можно сделать вывод, что распределение значений среднегодовых пробегов является **отличным от нормального**.

4.4. Графические методы описания данных

4.4.1. Boxplots: диаграммы размаха, ящики с усами

Визуализируем описательные статистики с помощью т. н. диаграмм размаха, которые также иногда называют диграмма «ящик с усами». Оригинальное название «Boxplot».

Диаграммы размахов, или "ящики с усами" (англ. box-whisker plots), получили своё название за характерный вид: точку или линию, соответствующую медиане или средней арифметической, окружает прямоугольник ("ящик"), длина которого соответствует одному из показателей разброса или точности оценки генерального параметра. Дополнительно от этого прямоугольника отходят "усы", также соответствующие по длине одному из показателей разброса или точности. Графики этого типа очень популярны, поскольку позволяют дать очень полную статистическую характеристику анализируемой совокупности. Кроме того, диаграммы размаха можно использовать для визуальной экспресс-оценки разницы между двумя и более группами (например, между ценами на недвижимость в различных муниципальных округах, ставками по кредитам в различных банках и т.д.).

Принципиальная схема диаграммы размаха выглядит следующим образом:

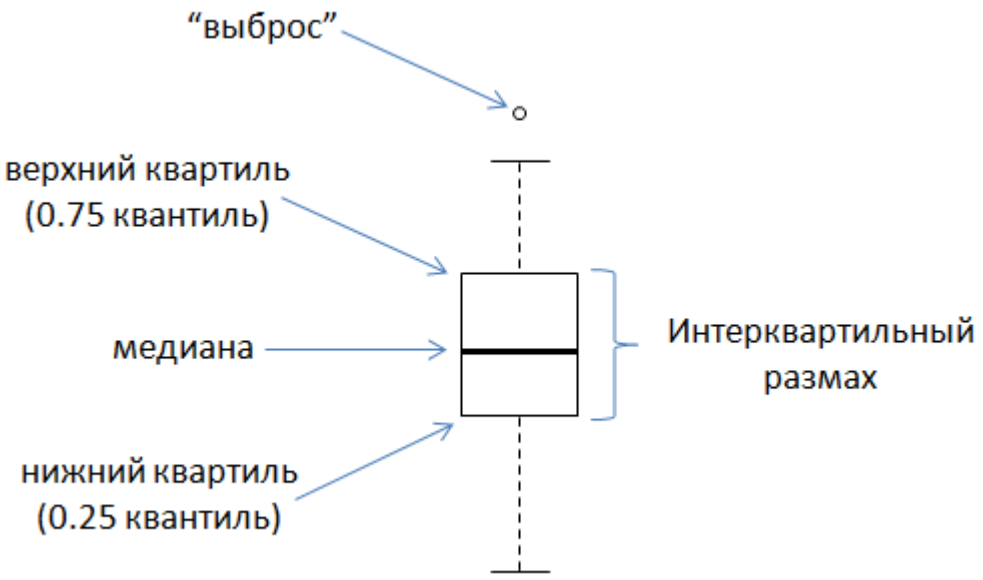

Рисунок 33: Принципиальная схема диаграммы размаха

Таким образом, в в R при построении диаграмм размахов используются устойчивые (робастные) оценки центральной тенденции (медиана) и разброса (интерквартильный размах, ИКР). Верхний "ус" простирается от верхней границы "ящика" до наибольшего выборочного значения, находящегося в пределах расстояния 1.5 (естественно аналитик может изменять данное значение) x ИКР от этой границы. Аналогично, нижний "ус" простирается от нижней границы "ящика" до наименьшего выборочного значения, находящегося в пределах расстояния 1.5 x ИКР от этой границы. Длину данного интервала (т.е. 1.5 x ИКР) можно изменить при помощи аргумента range функции boxplot(). Наблюдения, находящиеся за пределами "усов", потенциально могут быть выбросами. Однако всегда следует внимательно относиться к такого рода нестандартным наблюдениям - они вполне могут оказаться "нормальными" для исследуемой совокупности, и поэтому не должны удаляться из анализа без дополнительного расследования причин их появления [[59]].

График «ящик с усами», или «ящичковая диаграмма», был разработан Джоном Тьюки в 1970-х годах.

Другими словами можно сказать, что диаграмма размаха представляет собой реализацию отображения одномерного распределения плотности вероятности.

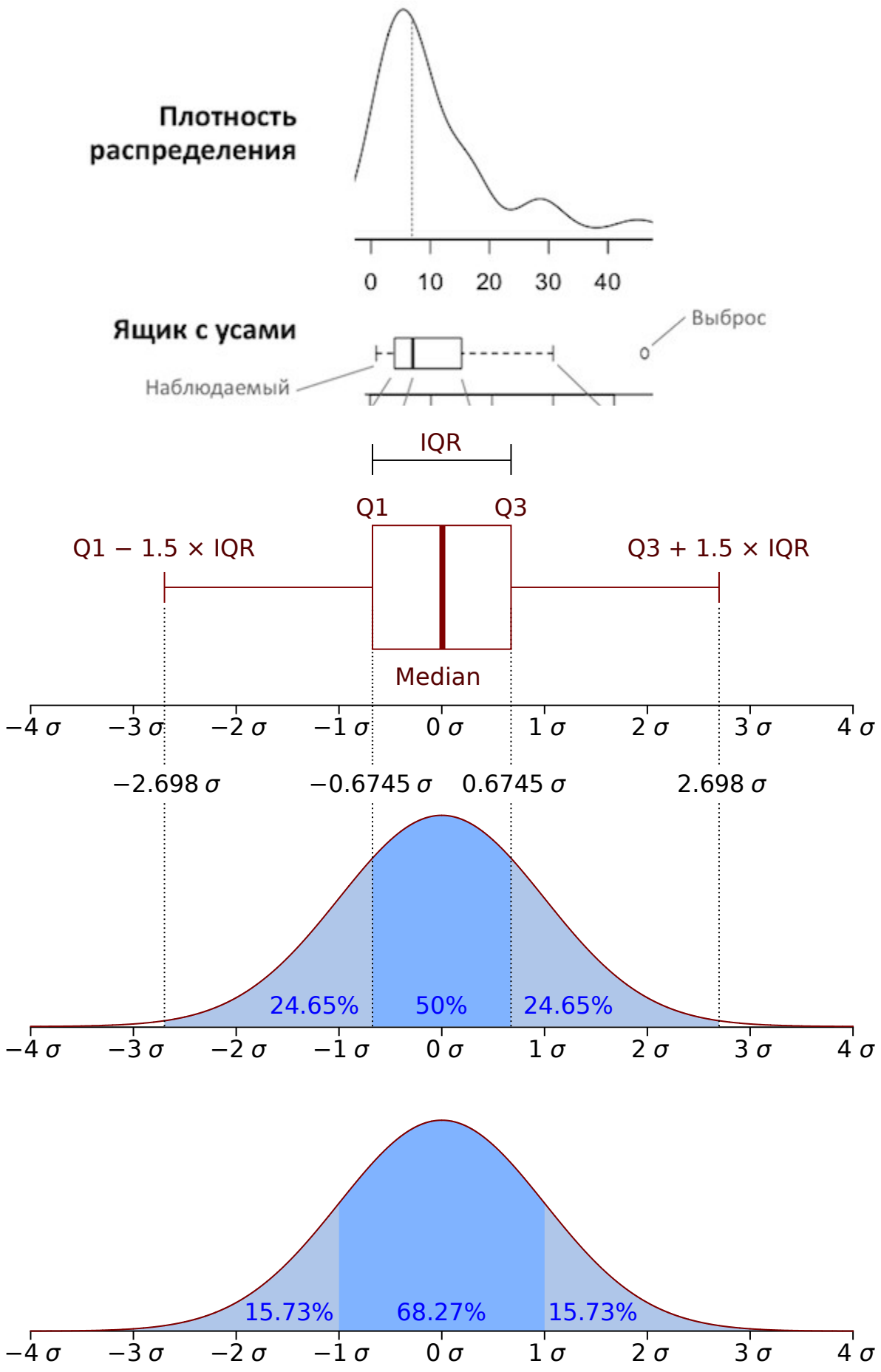

Рисунок 35: Диаграмма размаха и функция вероятности нормального распределения

Источник: [60]

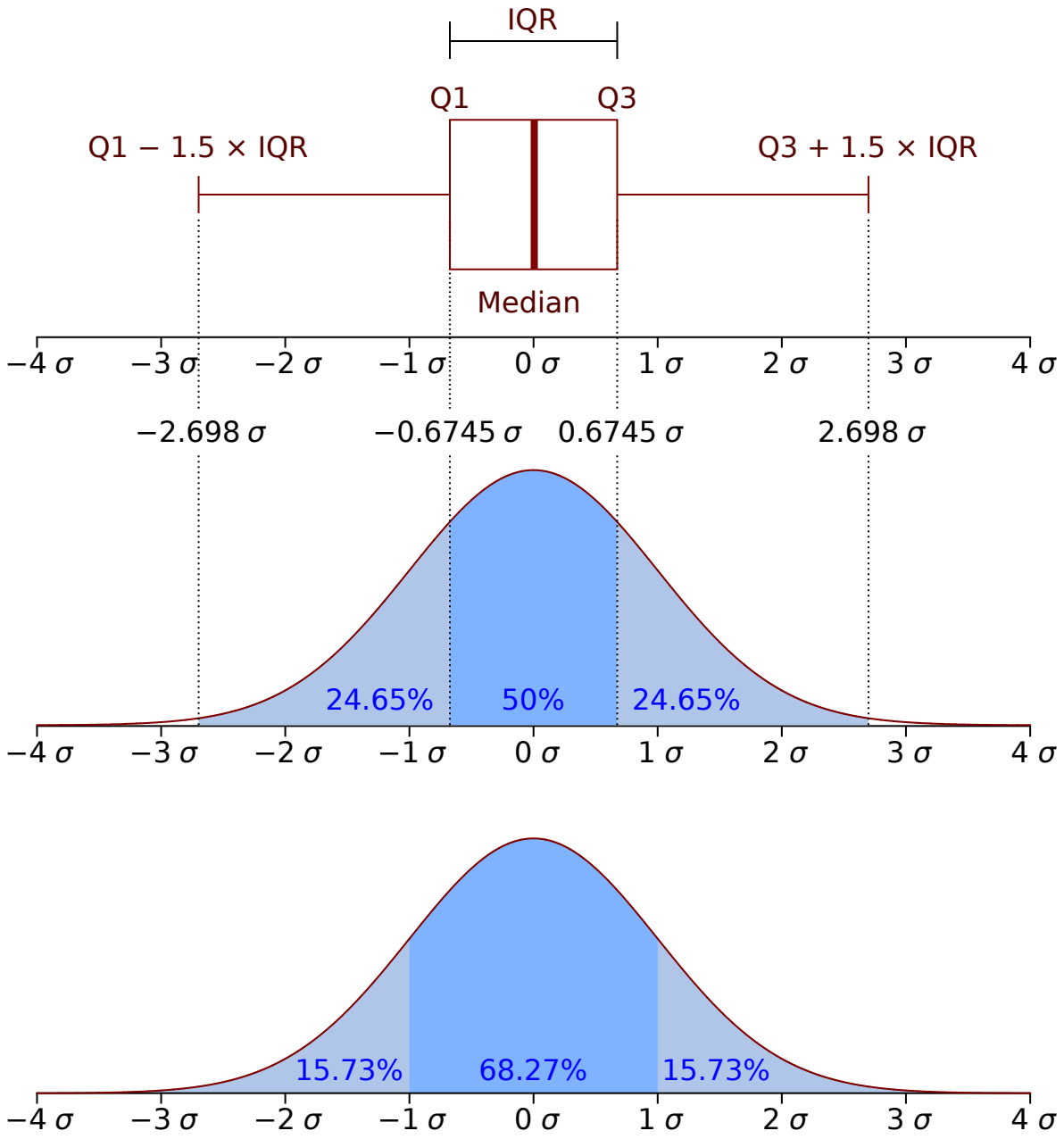

Рисунок 35: Диаграмма размаха и функция вероятности нормального распределения
Источник: [61].

Несмотря на свою простоту и удобство, первоначальная форма ящика с усами обладает и некоторыми недостатками. Один из таких существенных недостатков — отсутствие на графике информации о количестве наблюдений по выборке. Действительно, ящик с усами позволяет сравнить медианы, квартили, минимумы и максимумы по различным выборкам, но если мы

захотим сделать вывод об общей медиане по всей совокупности выборок, то мы не сможем этого сделать, не прибегая к расчётам на исходных данных. В 1978 году первоначальная форма ящика с усами была модифицирована МакГиллом, Ларсеном и Тьюки. Они предложили учитывать размер выборочной совокупности, рисуя ящики разного размера, а также изобразили на графике доверительный интервал [[62]] для медиан в виде расходящихся клиньев. Чем больше ящик по размерам, тем больше количество наблюдений в выборке, по которой строился этот ящик. Что касается доверительного интервала, то он представляет собой выемки на каждом из ящиков; в случае, если получившиеся выемки разных ящиков не пересекаются, их медианы статистически значимо различаются.

Иная модификация получила название «histplot» (сокр. от «histogram plot», с англ. — «график-гистограмма»). Теперь на графике отображаются плотности распределения по трём точкам: медиане, первому и третьему квартилю. Соответственно, вместо прямоугольника, «ящик» теперь представляет собой две равнобедренные трапеции, имеющие смежное основание.

Дальнейшее изменение получило название «vaseplot» (с англ. — «график-ваза») из-за визуального сходства «ящика» с вазой. На данном графике производится отображение всех плотностей вероятностей от первого до третьего квартиля. Затемнённые области представляют собой доверительный интервал медианы

Придерживаясь ранее провозглашённых принципов KISS и графического минимализма ограничимся построение базовой версии диаграмм размаха.

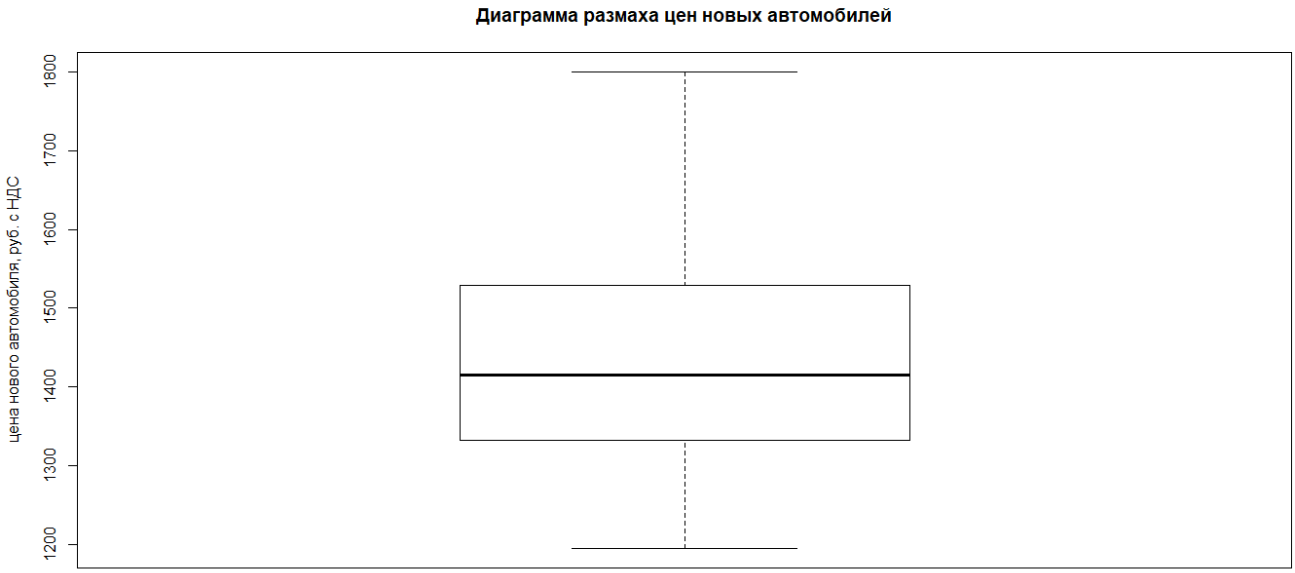

Рисунок 36: Диаграмма размаха значений цен новых автомобилей

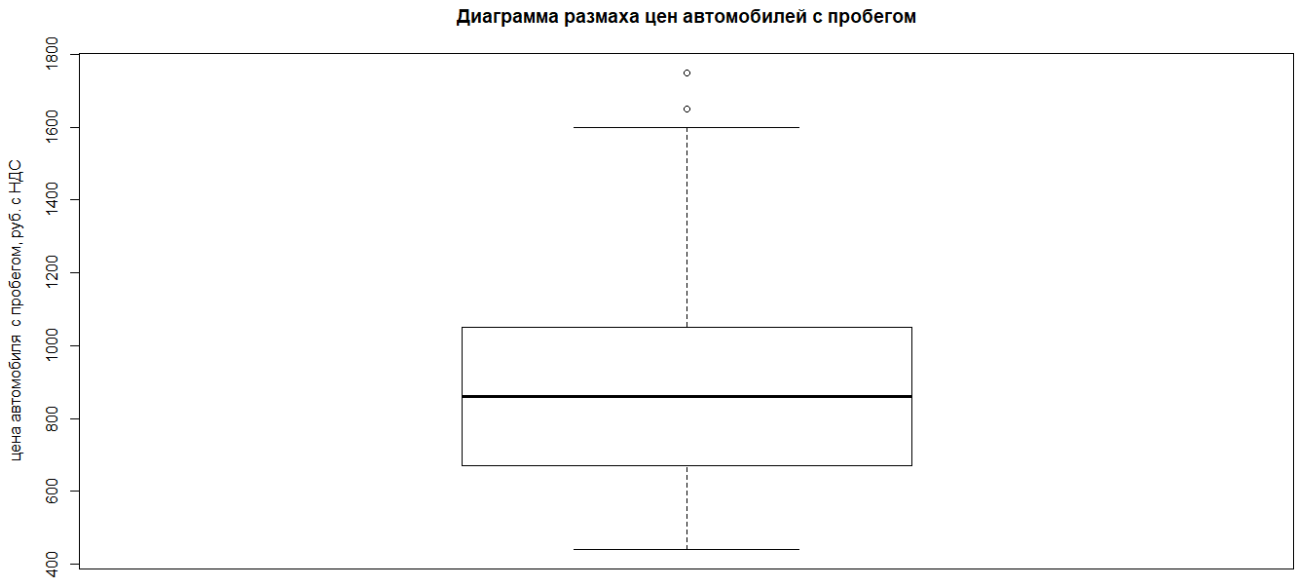

Рисунок 37: Диаграмма размаха значений цены автомобилей с пробегом

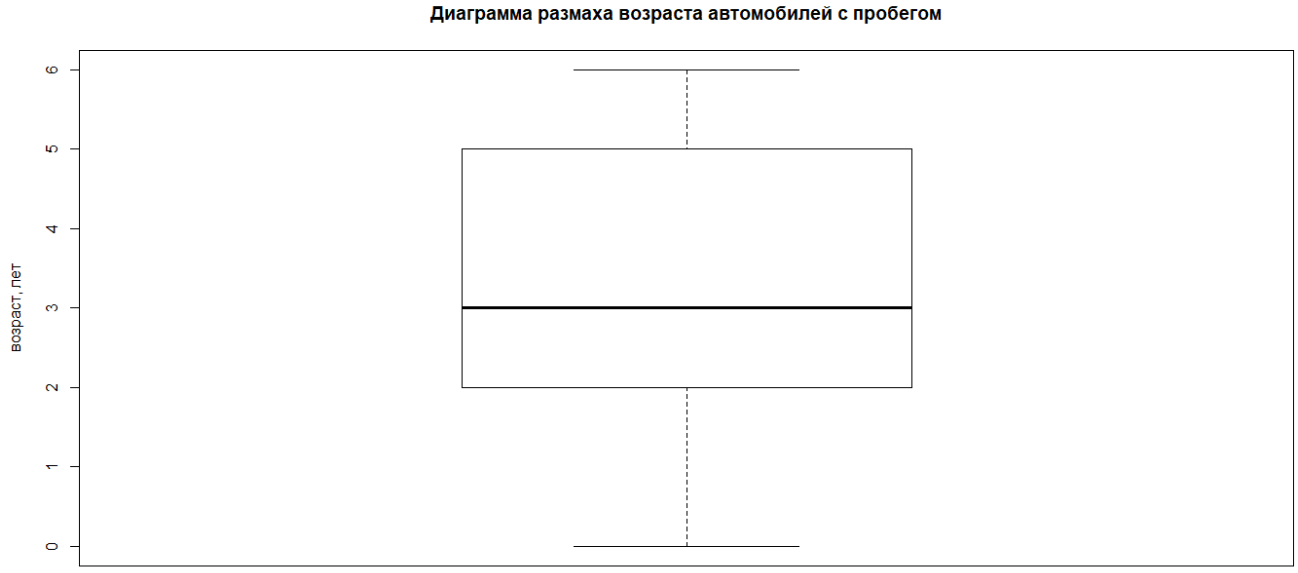

Рисунок 38: Диаграмма размаха значений возраста автомобилей с пробегом

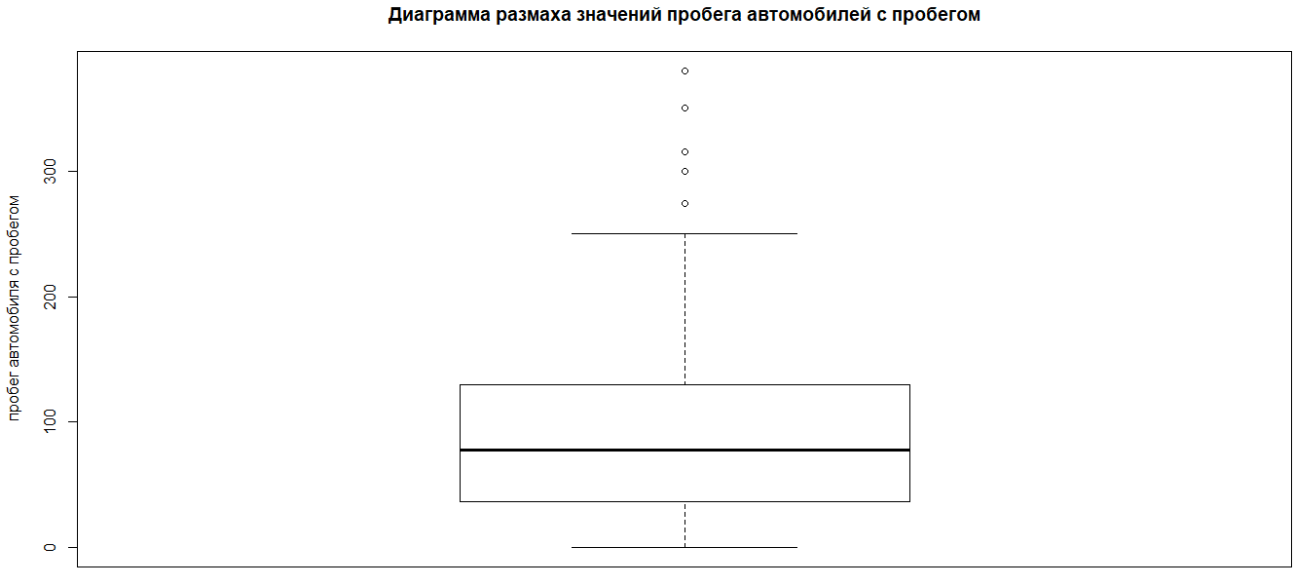

Рисунок 39: Диаграмма размаха значений пробега автомобилей с пробегом

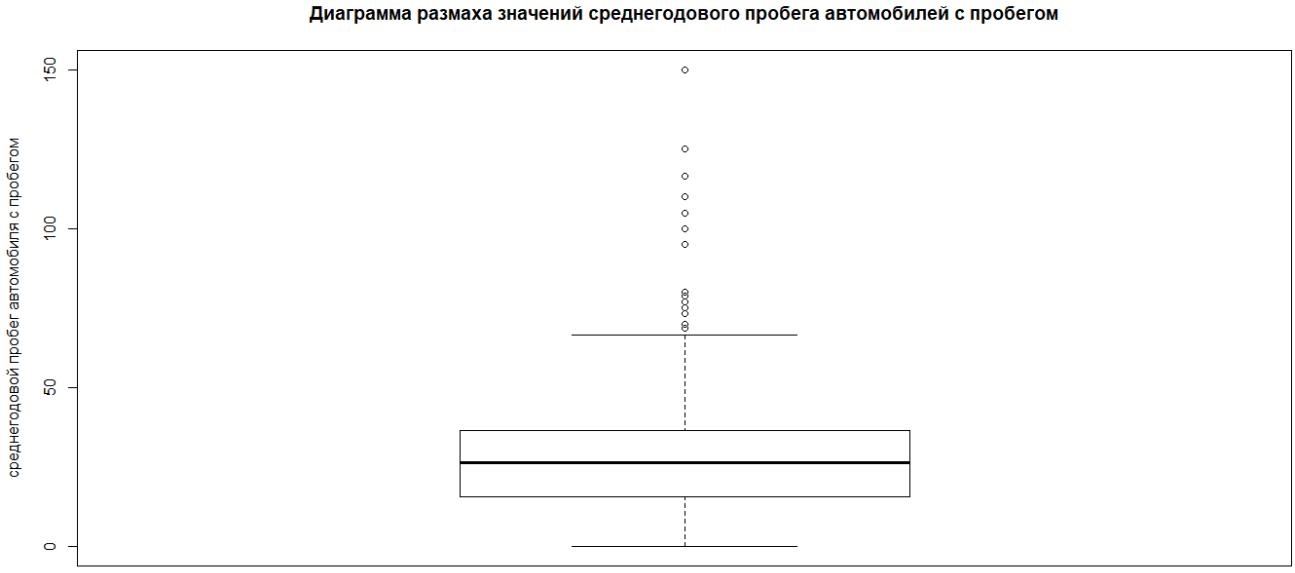

Рисунок 40: Диаграмма размаха значений среднегодового пробега автомобилей с пробегом

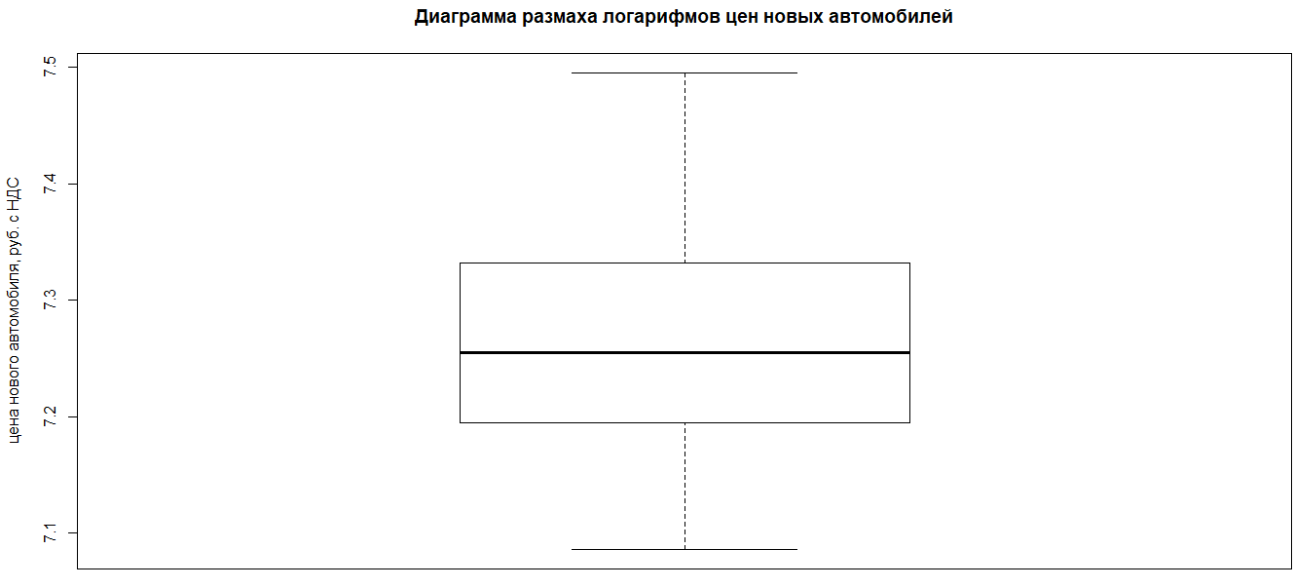

Рисунок 41: Диаграмма размаха значений логарифмов цен новых автомобилей

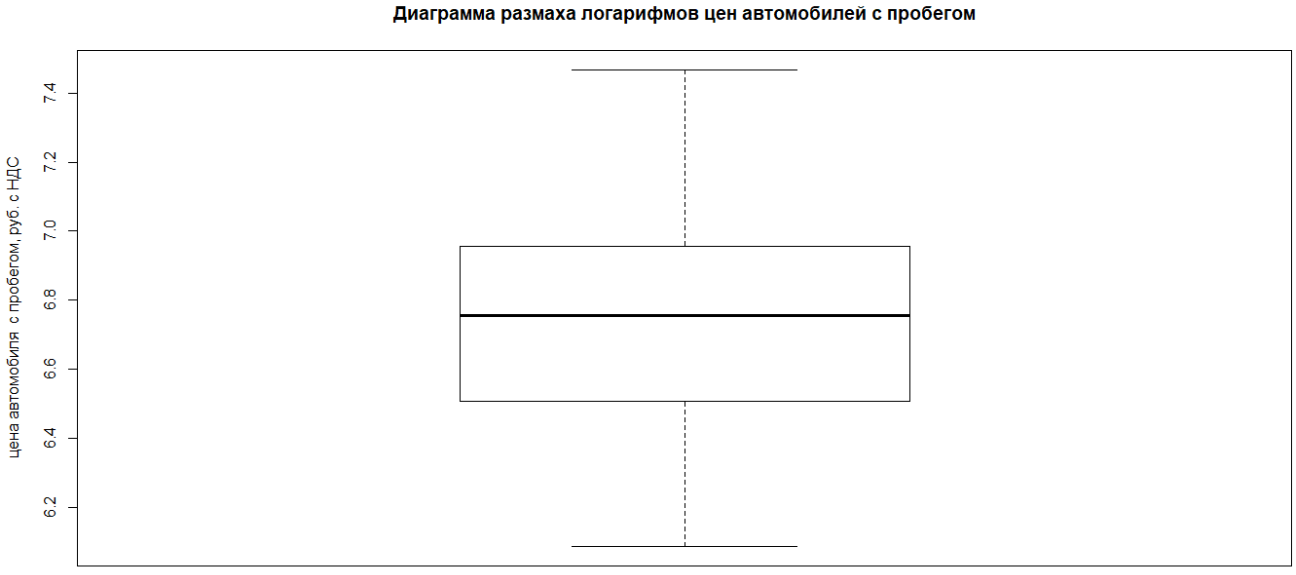

Рисунок 42: Диаграмма размаха значений логарифмов цен автомобилей с пробегом

```
  boxplot(gazelle.new.01$Price, ylab = "цена нового автомобиля, руб. с НДС", main = "Диаграмма размаха цен новых автомобилей") #строим диаграмму размаха цен новых автомбилей

  boxplot(gazelle.old.01$Price, ylab = "цена автомобиля  с пробегом, руб. с НДС", main = "Диаграмма размаха цен автомобилей с пробегом") #строим диаграмму размаха цен б/у автомбилей

  boxplot(gazelle.old.01$logPrice, ylab = "логарифм цены автомобиля с пробегом", main = "Диаграмма размаха значений логарифмов цены автомобилей с пробегом") #строим диаграмму размаха логарифмов цен автомобилей с пробегом

  boxplot(gazelle.old.01$Age, ylab = "возраст, лет", main = "Диаграмма размаха возраста автомобилей с пробегом") #строим диаграмму размаха возраста

  boxplot(gazelle.old.01$Mileage, ylab = "пробег автомобиля с пробегом", main = "Диаграмма размаха значений пробега автомобилей с пробегом") #строим диаграмму размаха логарифмов цен автомобилей с пробегом

  boxplot(gazelle.old.01$MpY, ylab = "среднегодовой пробег автомобиля с пробегом", main = "Диаграмма размаха значений среднегодового пробега автомобилей с пробегом") #строим диаграмму размаха логарифмов цен автомобилей с пробегом

  boxplot(gazelle.new.01$logPrice, ylab = "цена нового автомобиля, руб. с НДС", main = "Диаграмма размаха логарифмов цен новых автомобилей") #строим диаграмму размаха лог цен новых автомбилей

  boxplot(gazelle.old.01$logPrice, ylab = "цена автомобиля  с пробегом, руб. с НДС", main = "Диаграмма размаха логарифмов цен автомобилей с пробегом") #строим диаграмму размаха лог цен б/у автомбилей
```

Здесь будут выводы

4.4.2. Диаграммы рассеяния

Построим диаграммы рассеивания: «Цена-возраст», «Цена-пробег», «Натуральный логарифм цены — возраст», «Натуральный логарифм цены — пробег», «Возраст-пробег». Оценщикам такие диаграммы хорошо знакомы. В табличных процессорах из офисных пакетов: LibreOffice, OpenOffice, Microsoft office и т. п. – они часто используются не только для визуализации данных, но и для построения прямо с них регрессионных моделей. Нет необходимости подробно объяснять, что это такое. Отметим только, что сейчас мы введём два новых термина, которые в дальнейшем будем использовать по всему тексту Руководства. Отныне переменную, определение влияния других на которую мы и хотим выяснить (в нашем случае на данном этапе, это цена, либо логарифм цены, а в дальнейшем это будет остаточная полезность и износ) мы будем называть «**Отклик**», а влияющую на отклик переменную/переменные (в нашем случае это будут «Возраст», «Пробег» и их модифицированные значения) будем называть «**Предиката**». Необходимость использования таких неочевидных терминов часто возникает, когда речь идёт о междисциплинарых исследованиях, в которых одни и те же термины означают разные понятия в тех областях знаний, которые исследователи пытаются совместить. Поскольку термин «**Фактор**» уже занят в таком подходе к Data mining как факторный анализ, придётся смириться с отказом от использования привычного оценщикам термина «ценообразующий фактор».

Таким образом на диаграммах рассеяния на оси абсцисс будут откладываться значения предикат, а по оси ординат — откликов.

Продемонстрируем два варианта диаграмм рассеяния — минималистичный, схожий с теми, которые строят табличные процессоры и расширенный, содержащий полезную в ряде случаев дополнительную информацию.

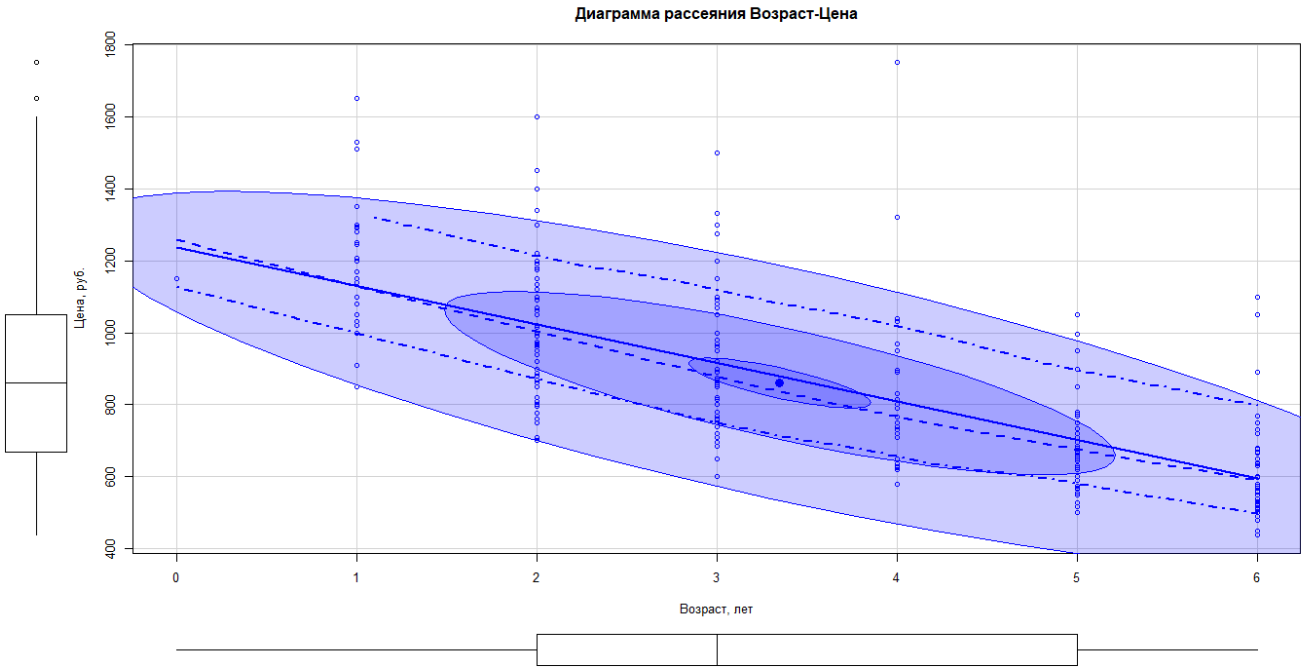

Рисунок 43: Диаграмма рассеяния "возраст-цена" с добавлением диаграмм размаха для шкал, аппроксимирующей линии линейной зависимости, сглаживающих линий (спаном 0.5), эллипсов уровней: 0.05, 0.5, 0.95

Как видим всё привычно. Построим расширенную версию. Как уже было сказано ранее в разделе 4.1.2. Построение гистограмм на стр. 58, с одной стороны R предоставляет большие возможности для визуализации данных, с другой — не следует чрезмерно увлекаться этим. Построим диаграмму рассеяния с дополнительными элементами: уже известными нам диаграммами размаха для шкал, аппроксимирующей линией линейной зависимости, сглаживающими линиями (спан = 0.5), эллипсами, описывающими 0.05, 0.5 и 0.95 всех наблюдений. По мнению автора, данная комбинация показателей может считаться рекомендуемой для подобных диаграмм в целях всего Руководства.

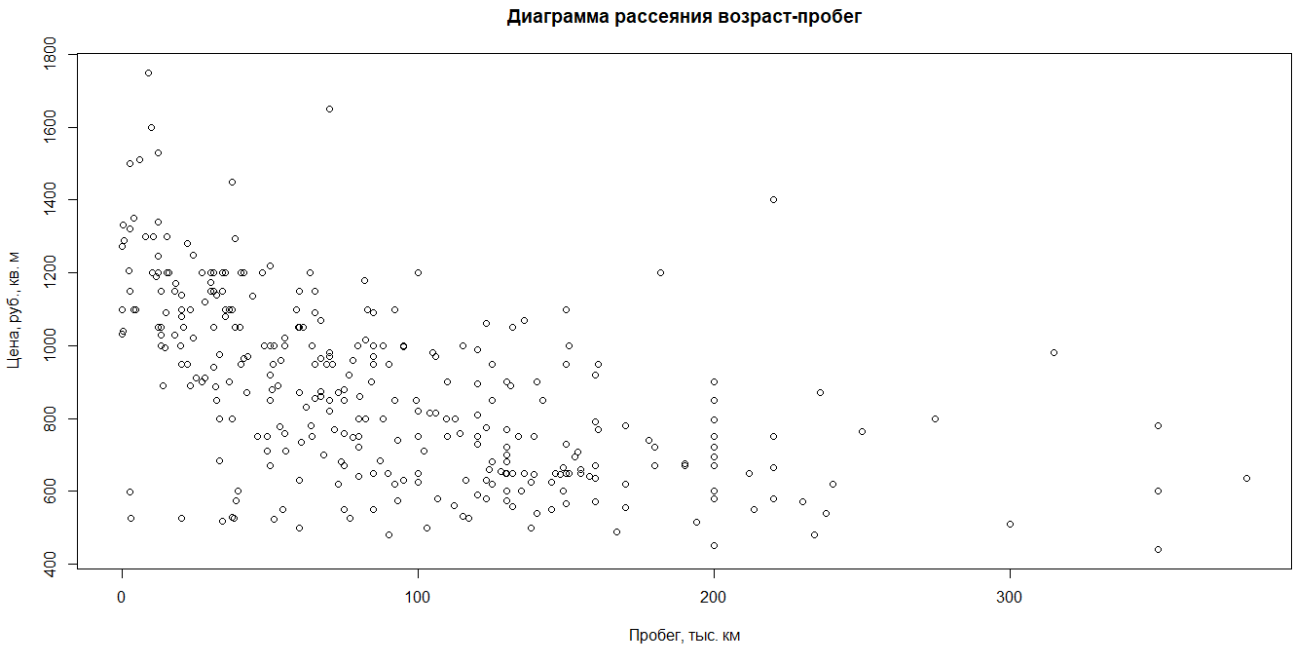

Рисунок 44: Простая диаграмма рассеяния "пробег-цена"

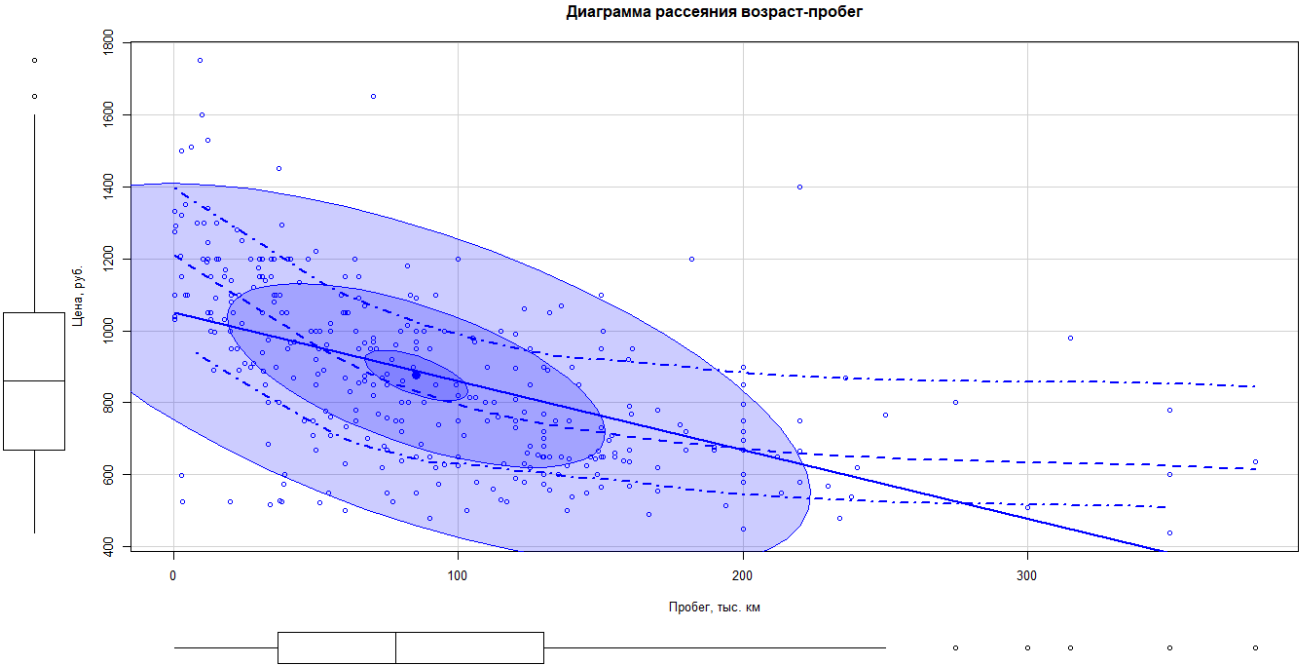

Рисунок 45: Диаграмма рассеяния "пробег-цена" с добавлением диаграмм размаха для шкал, аппроксимирующей линии линейной зависимости, сглаживающих линий (спаном 0.5), эллипсов уровней: 0.05, 0.5, 0.95

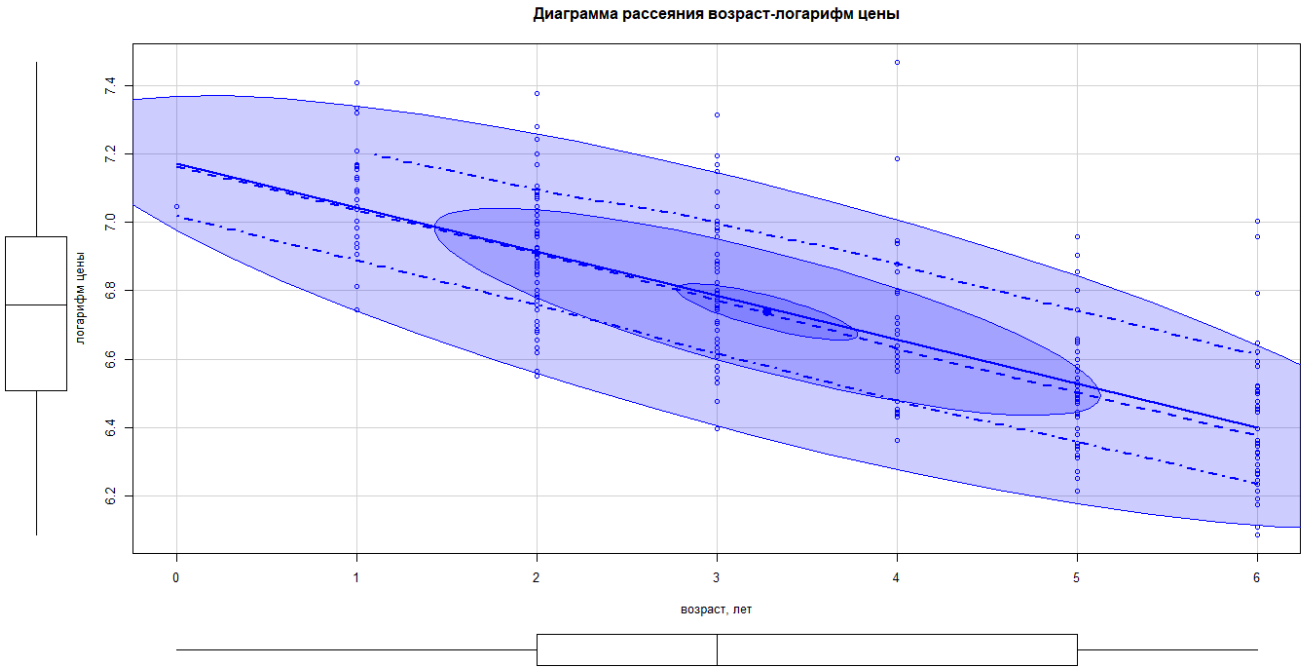

Рисунок 47: Диаграмма рассеяния "возраст-логарифм цены" с добавлением диаграмм размаха для шкал, аппроксимирующей линии линейной зависимости, сглаживающих линий (спаном 0.5), эллипсов уровней: 0.05, 0.5, 0.95

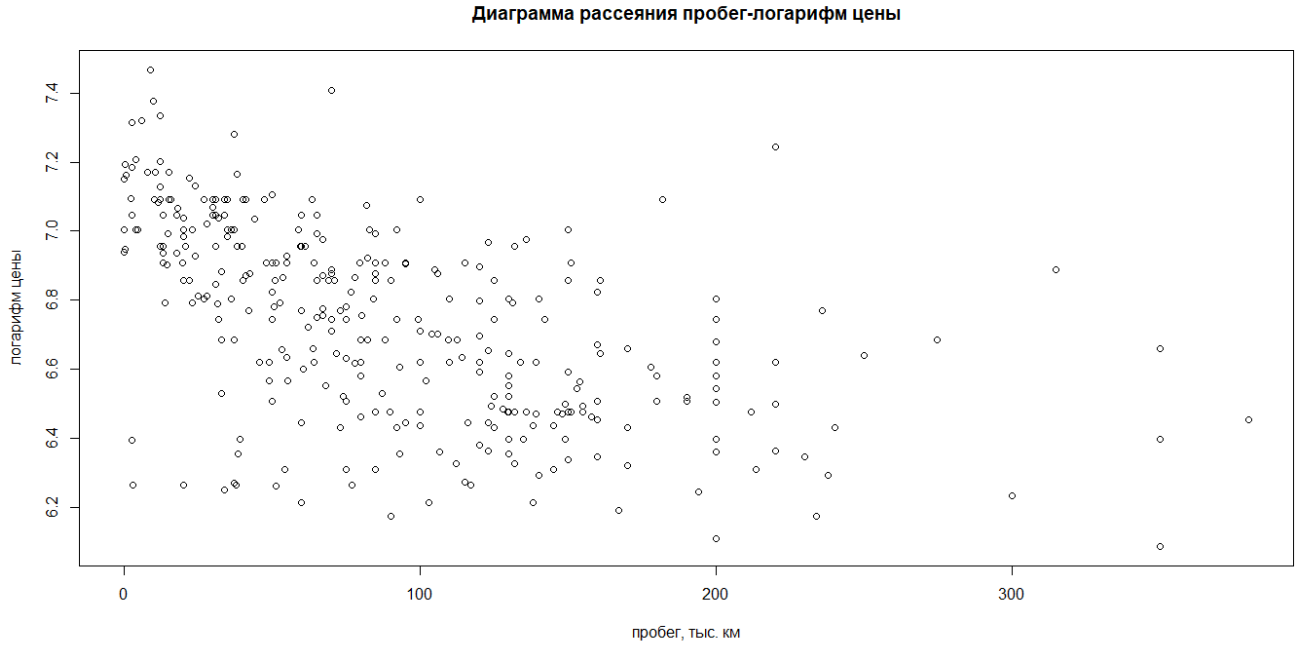

Рисунок 47: Диаграмма рассеяния "пробег-логарифм цены"

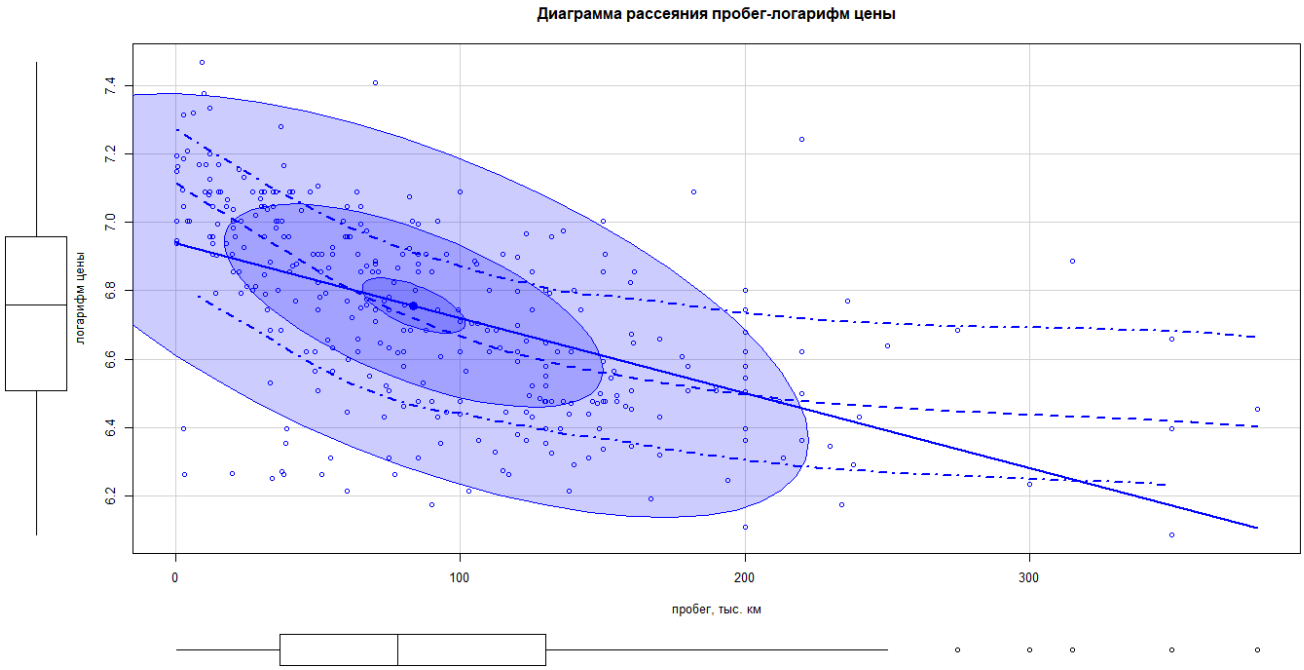

Рисунок 48: Диаграмма рассеяния "пробег-логарифм цены" с добавлением диаграмм размаха для шкал, аппроксимирующей линии линейной зависимости, сглаживающих линий (спаном 0.5), эллипсов уровней: 0.05, 0.5, 0.95

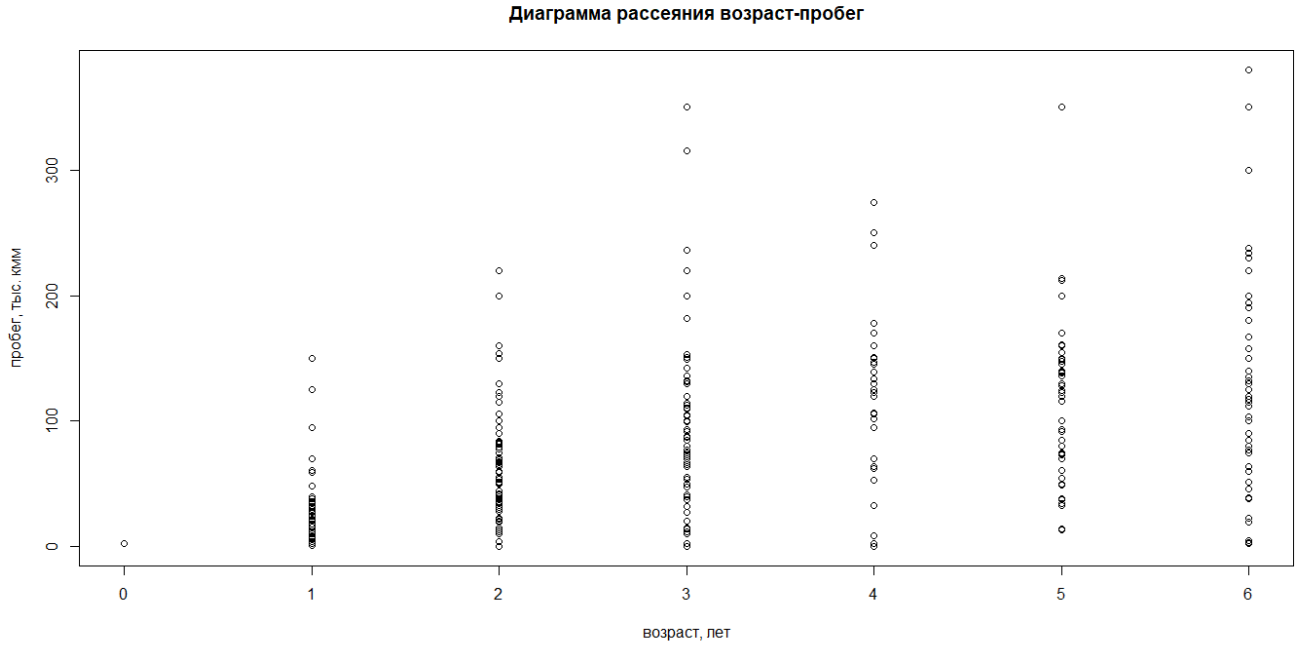

Рисунок 49: Диаграмма рассеяния "возраст-пробег"

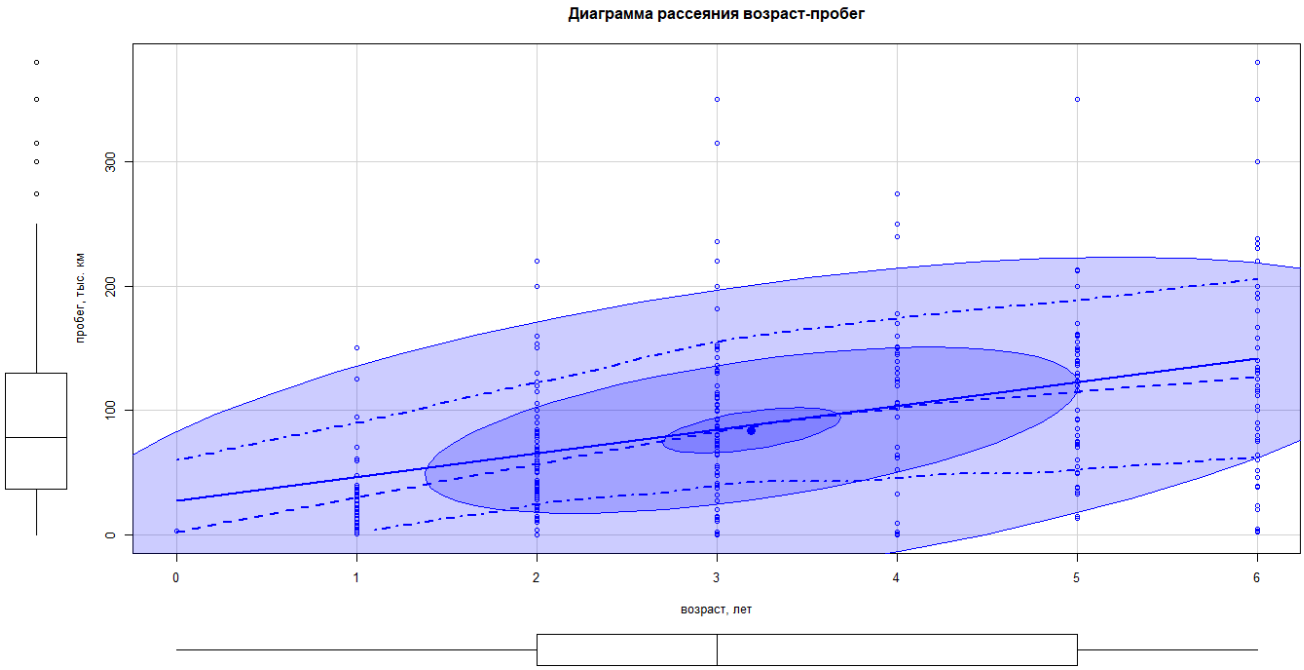

Рисунок 50: Диаграмма рассеяния "возраст-пробег" с добавлением диаграмм размаха для шкал, аппроксимирующей линии линейной зависимости, сглаживающих линий (спаном 0.5), эллипсов уровней: 0.05, 0.5, 0.95

Ниже приводится код, вызывающий построение диаграмм рассеяния.

```
  plot(gazelle.old.01$Age, gazelle.old.01$Price, xlab = "Возраст, лет", ylab = "Цена, руб., кв. м", main = "Диаграмма рассеяния Возраст-Цена") #строим ростую диаграмму рассеяния для "Возраст-Цена"

  scatterplot(gazelle.old.01$Price~gazelle.old.01$Age, ylab = "Цена, руб.", xlab ="Возраст, лет", main = "Диаграмма рассеяния Возраст-Цена", ellipse=list(levels=c(.05, .5, .95))) #строим расширенную диаграмму рассеяния для "Возраст-Цена"

  plot(gazelle.old.01$Mileage, gazelle.old.01$Price, xlab = "Пробег, тыс. км", ylab = "Цена, руб., кв. м", main = "Диаграмма рассеяния возраст-пробег") #строим ростую диаграмму рассеяния для "возраст-вробег"

  scatterplot(gazelle.old.01$Price~gazelle.old.01$Mileage, ylab = "Цена, руб.", xlab ="Пробег, тыс. км", main = "Диаграмма рассеяния возраст-пробег", ellipse=list(levels=c(.05, .5, .95))) #строим расширенную диаграмму рассеяния для "возраст-пробег"

  plot(gazelle.old.01$Age, gazelle.old.01$logPrice, xlab = "возраст, лет", ylab = "логарфим цены", main = "Диаграмма рассеяния возраст-логарифм цены") #строим ростую диаграмму рассеяния для "возраст-логарифм цены"
```

```
  scatterplot(gazelle.old.01$logPrice~gazelle.old.01$Age, ylab = "логарифм цены", xlab ="возраст, лет", main = "Диаграмма рассеяния возраст-логарифм цены", ellipse=list(levels=c(.05, .5, .95))) #строим расширенную диаграмму рассеяния для "возраст-логарифм цены"

  plot(gazelle.old.01$Mileage, gazelle.old.01$logPrice, xlab = "пробег, тыс. км", ylab = "логарифм цены", main = "Диаграмма рассеяния пробег-логарифм цены") #строим ростую диаграмму рассеяния для "возраст-логарифм цены"

  scatterplot(gazelle.old.01$logPrice~gazelle.old.01$Mileage, ylab = "логарифм цены", xlab ="пробег, тыс. км", main = "Диаграмма рассеяния пробег-логарифм цены", ellipse=list(levels=c(.05, .5, .95))) #строим расширенную диаграмму рассеяния для "пробег-логарифм цены"

  plot(gazelle.old.01$Age, gazelle.old.01$Mileage, xlab = "возраст, лет", ylab = "пробег, тыс. кмм", main = "Диаграмма рассеяния возраст-пробег") #строим простую диаграмму рассеяния для "возраст-пробег"

  scatterplot(gazelle.old.01$Mileage~gazelle.old.01$Age, ylab = "пробег, тыс. км", xlab ="возраст, лет", main = "Диаграмма рассеяния возраст-пробег", ellipse=list(levels=c(.05, .5, .95))) #строим расширенную диаграмму рассеяния для "возраст-пробег"

  Здесь будут выводы
```

Глава 5. Кластерный анализ (распознавание образов без учителя)

В общем случае задача кластерного анализа состоит в:

- разбиение наблюдений на группы;
- определение числа групп.

При иерархическом кластерном анализе число групп заведомо неизвестно. Упрощённо можно сказать, что на первоначальном этапе задача классификации кластерного анализа (не путать с задачей классификации в Machine learning (классификация с учителем, распознавание образов)) решается геометрическими методами. Введём следующие предпосылки:

- каждое наблюдение — точка;
- «похожие» наблюдения расположены «близко» друг к другу;
- различающиеся объекты расположены «далеко»;
- скопление точек — есть кластер.

Для графического описания процесса кластеризации можно использовать денндрограммы.

Выбор метода определения расстояния между наблюдениями.

Выбор метода определения расстояния между кластерами.

Выбор алгоритма классификации.

Преобразование данных, т. е. их стандартизация. Можно от 0(-1) до 1, а можно через z-метки, т. е. преобразование данных таким образом, чтобы среднее арифметическое было равно 0, а выборочная дисперсия — 1. Суть метода заключается в том, что для значения наблюдения переменной x_i векторая данных X выполняется преобразование таким образом, что на основе вектора X создаётся вектор Z, в котором:

$$z_i = \frac{x_i - \bar{x}}{\sqrt{\sigma_X^2}} \quad , \tag{20}$$

где x_i – значение i-того значений переменной X;

$\bar{x}$ – среднее арифметическое значений переменной X;

σ^2_x – выборочная дисперсия значений переменной X;

$\sqrt{\sigma^2_x}$ – стандартное отклонение значений переменной X;

либо:

$$z_i = \frac{x_i - min_x}{max_x - min_x}$$, где (21)

где x_i – значение i-того значений переменной x;

max_i – максимальное значение среди всех наблюдений переменной X;

min_i – минимальное x_i – значение i-того значений переменной X.

Следует отметить, что заранее невозможно предсказать, какая из двух формул даст лучший результат. При этом у метода, описанного формулой (20) есть ограничение — применимость только в случае с данными, имеющими нормальное распределение. Это связано с тем, что...