**Построение модели определения стоимости автомобиля с пробегом**.

Программа: Искусственный интеллект

Специализация: Инженер ИИ

Студент: Князьков Павел Олегович

**Аннотация**

Этот набор данных содержит данные о 762 091 подержанном автомобиле, собранном из cars.com. Данные были собраны в апреле 2023 года. В данных содержат следующие информации, которые позволяют нам анализировать текущий рынок поддержанных автомобилей, чтобы ориентировать спрос и предложений для увеличения успеха продажи. В задании необходимо сделать следующие задания:

* Проанализировать базу данных «Used Cars Dataset»;
* Устранить ненужные данные: шумы, выбросы, искаженные данные;
* Выдвинуть гипотезы и проанализировать их по выдвинутыми гипотезами построить графики.

В работе будут приложены аналитические информации: корреляционный анализ, агрегирование данных, оптимизация память, удаление дубликатов, очистка данных. Будет производиться анализ гипотез по известным данным и построены соответствующие графики и схемы.

**СОДЕРЖАНИЕ**

[**ВВЕДЕНИЕ** 4](#_Toc178179237)

[Глава 1 Теоретическая часть 5](#_Toc178179238)

[1.1 Выбор инструментов 5](#_Toc178179239)

[1.2 Поиск и загрузка датасетов 5](#_Toc178179240)

[1.3 Концепция ноутбука 5](#_Toc178179241)

[Глава 2 Практическая часть 7](#_Toc178179242)

[**2.1 Определение данных для анализа** 7](#_Toc178179243)

[**2.2 Описание данных** 7](#_Toc178179244)

[**2.3 Формулирование гипотез** 8](#_Toc178179245)

[**2.4 Изучение общей информации** 9](#_Toc178179246)

[**2.5 Подготовка данных для работы** 14](#_Toc178179247)

[**2.6 Агрегирование данных** 20](#_Toc178179248)

[**2.7 Исследовательский анализ данных** 25](#_Toc178179249)

[Гипотеза 1: Автомобили с более низким пробегом имеют более высокую цену. 25](#_Toc178179250)

[Гипотеза 2: Автомобили с определенным типом топлива имеют более высокую эффективность топливо потребления. Рассмотрим эту гипотезу и проверим. 27](#_Toc178179251)

[Гипотеза 3: Рейтинг продавца и рейтинг водителей связаны с ценой автомобиля. Рассмотрим эту гипотезу и проверим. 30](#_Toc178179252)

[Гипотеза 4: Автомобили с определенными цветами салона стоят дороже. 32](#_Toc178179253)

[Гипотеза 5: Автомобили с определенными цветами кузова стоят дороже. 35](#_Toc178179254)

[Гипотеза 6: Рейтинг продавца коррелирует с рейтингом автомобиля, данным водителями. Рассмотрим эту гипотезу и проверим 38](#_Toc178179255)

[Гипотеза 7: Средняя цена на автомобили меняется с течением времени. 40](#_Toc178179256)

[Гипотеза 8: Автомобили с определенным типом топлива имеют более высокую эффективность топливо потребления. 42](#_Toc178179257)

[Гипотеза 9: Автомобили с определенным типом трансмиссии имеют более высокую эффективность топливо потребления. 44](#_Toc178179258)

[**2.8 Общий вывод** 48](#_Toc178179259)

[ЗАКЛЮЧЕНИЕ 49](#_Toc178179260)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ** 50](#_Toc178179261)

# 

# **ВВЕДЕНИЕ**

Для выполнения данной работы мною было выбрано направление «Автомобили с пробегом», где была исследована база данных с целью выявления определенных закономерностей.

Поставим цели, которых мы хотим добиться, выполнив данную работу.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

1. Теоретическая часть
   1. Выбор инструментов

В качестве инструментов для написания кода будем использовать бесплатные программы. JupiterLab который входит в состав программы Anaconda. JupiterLab это практически то же самое, что и Jupiter Notebook, только с расширенными возможностями. JupiterLab как и Jupiter Notebook позволяют выполнять каждую ячейку кода отдельно, при этом результат выполнения и все переменные сохраняются в ноутбуке. Это позволяет запускать ресурсоёмкие части кода один раз (например обучение модели), а дальше экспериментировать с результатами, без надобности запускать весь код с начала. Это очень удобно, по сколько, например процессы кросс-валидации могут занимать более часа.

Для оформления текстового описания работы будем использовать программу Microsoft 365 Word, которая позволяет применять широкие возможности для редактирования документов.

В бесплатной программе Visual Studio Code, можно просматривать ноутбук и копировать ячейки кода с сохранением полного форматирования, для последующего переноса этого текста в документ Word.

В ходе работы будут использоваться следующие библиотеки: Pandas, MatPlotLib, Seaborn, Numpy, Scipy.

* 1. Поиск и загрузка датасетов

Датасет должен содержать информацию о данных по автомобилям с пробегом, например годах выпуска, продажи, количества собственников, типах ДВС, трансмиссий, цветов кузова и салона!

Желательно так же, чтобы датасеты были бесплатными и читаемыми в Excel.

* 1. Концепция ноутбука

Чтобы избавиться от ручного монотонного труда при загрузке, обработке датасетов, отрисовке графиков, нужно продумать программу так, чтобы максимально автоматизировать весь процесс.

Например, датасеты должны автоматически загружаться с какой ни будь директории, в соответствии с правилами загрузки.

Распределение на фичи и таргеты также должно происходить в автоматическом режиме. Например файлы, содержащие символ ‘\_’ автоматически будут инициализированы как таргет, все остальные как фичи.

На всех графиках должны в автоматическом режиме добавляться/удалятся фичи и таргеты. Если есть русская расшифровка таргета/фич, она также должна в автоматическом режиме вставляться в график. Возможна некая корректировка размера полотна графика в коде программы для улучшения вида графиков.

Процесс добавления/удаления фич и таргетов должен сводиться к добавлению/удалению датасета в директорию откуда производиться чтение, без вмешательства в код программы.

Глава 2 Практическая часть

**2.1 Определение данных для анализа**

В качестве данных для анализа был выбран датасет «Used Cars Dataset». В датасете представлены все автомобили, которые были собраны и получены в 2023 году.

Для анализа будем чаще всего использовать следующие данные из датасета: название производителя автомобиля, название модели автомобиля, год, когда был выпущен автомобиль, миль, пройденный автомобилем с момента выпуска, автомобильный двигатель, тип трансмиссии автомобиля, тип трансмиссии автомобиля, тип топлива, которое потребляет автомобиль, количество миль, которое автомобиль может проехать, используя один галлон топлива (мили на галлон), цвет кузова автомобиля, цвет салона автомобиля, попадал ли автомобиль в аварии, принадлежал ли автомобиль одному лицу, использовался ли автомобиль только в личных целях, имя продавца, рейтинг продавца, рейтинг автомобиля, данный водителями, количество отзывов об автомобилях, оставленных водителями, снижение цены по сравнению с начальной ценой, цена автомобиля.

Также для анализа изменим некоторые столбцы с типа «object» на тип «float», либо на тип «int», либо на тип «bool» в целях уменьшения занимаемой памяти датасета, делая его более производительным.

**2.2 Описание данных**

Для анализа были собраны данные обо всех проектах, включённых в список. В наборе данных содержатся:

1. manufacturer – название производителя автомобиля
2. model – название модели автомобиля
3. year – год выпуска/производства автомобиля
4. mileage – пробег - количество миль, пройденных автомобилем с момента выпуска
5. engine – тип ДВС - автомобильный двигатель (бенз, газ, дизель, гибрид, электричка)
6. transmission – АКПП/КПП - тип коробки переключения передач у автомобиля
7. drivetrain – тип привода у автомобиля (передний, задний, полный)
8. fuel\_type – тип топлива, которое потребляет автомобиль
9. mpg – запас хода - количество миль, которое автомобиль может проехать, используя один галлон топлива (мили на галлон)
10. exterior\_color – цвет кузова автомобиля
11. interior\_color – цвет салона автомобиля
12. accidents\_or\_damage – попадал ли автомобиль в ДТП
13. one\_owner – принадлежал ли автомобиль одному собственнику
14. personal\_use\_only – использовался ли автомобиль только в личных целях
15. seller\_name – имя продавца
16. seller\_rating – рейтинг у продавца
17. driver\_rating – рейтинг у автомобиля, данный ему водителями
18. driver\_reviews\_num – количество отзывов об автомобиле, оставленных водителями
19. price\_drop – уровень снижение цены по сравнению с начальной ценой
20. price – цена автомобиля

# **2.3 Формулирование гипотез**

Гипотеза 1: Автомобили с более низким пробегом имеют более высокую цену.

Гипотеза 2: Автомобили с более новыми годами выпуска стоят дороже

Гипотеза 3: Рейтинг продавца и рейтинг водителей связаны с ценой автомобиля.

Гипотеза 4: Автомобили с определенным типом топлива имеют более высокую эффективность топливо потребления.

Гипотеза 5: Автомобили с определенным типом трансмиссии имеют более высокую эффективность топливопотребления.

Гипотеза 6: Автомобили с определенными цветами кузова стоят дороже.

Гипотеза 7: Автомобили с определенными цветами салона стоят дороже.

Гипотеза 8: Рейтинг продавца коррелирует с рейтингом автомобиля, данным водителями.

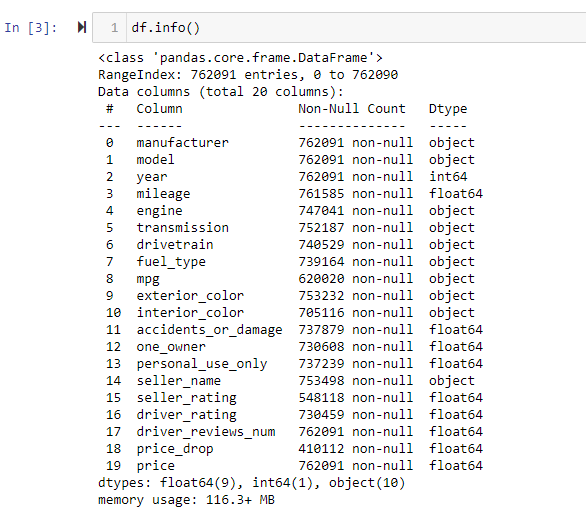
Гипотеза 9: Средняя цена на автомобили меняется с течением времени.

**2.4 Изучение общей информации**

Загружаем датасет и подключаем необходимые библиотеки:



Получим информацию о датасете:



Здесь можно заметить, что в датасете содержатся единиц 762091 строк. А также имеют 3 различные типы: object, int64 и float64. В целях экономии памяти можно преобразовать в другие типы. Далее рассмотрим содержимое в датасете, что там описываются.

Изображение выглядит как текст, снимок экрана, число, Шрифт

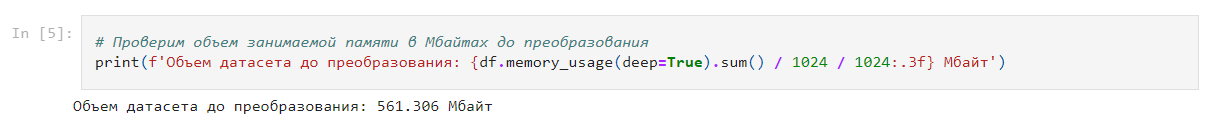
Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, Шрифт

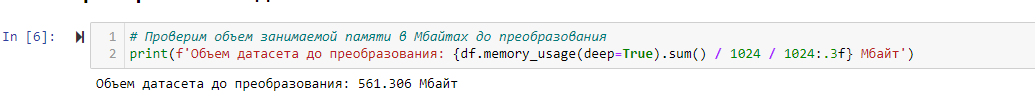
Автоматически созданное описание

Ничего особенного нет, рассмотрим следующие этапы.

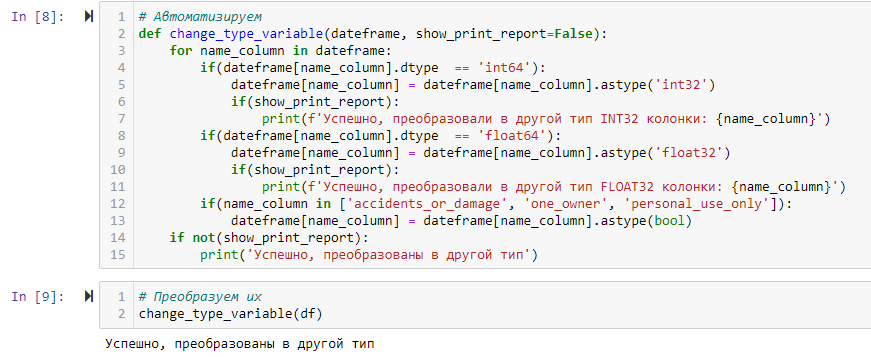
Изменим некоторые столбцы с типа «object» на тип «float», либо на тип «int», либо на тип «bool» в целях уменьшения занимаемой памяти датасета, делая его более производительным»:



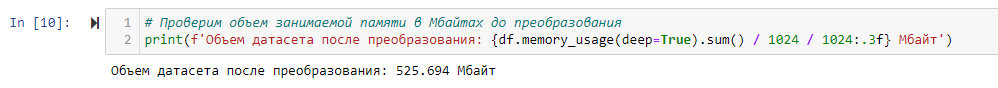
Сохраним объем датасета изначального размера для сравнения:



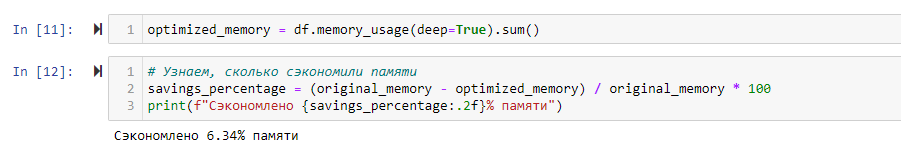
Создадим функцию, которая будет автоматизировать эту работу:



Рассмотрим полученный объем датасета после преобразования:



Проверим, на сколько сэкономили места:



Сэкономили на 6,34% объема памяти.

Рассмотрим полученную информацию о датасете:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Как и видим, что некоторые колонки датасета успешны преобразованы в другой тип. Рассмотрим описательную статистику.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Здесь стоит обратить внимание на следующие колонки:

* **Год**. Здесь мы видим что чаще всего продаются автомобили, примерно `2017 года` выпуска (данное среднее значение преобладает среди остальных). В объявлениях так же можем увидеть автомобиль даже `1915 года`.
* **Пробег.** В объявлениях чаще всего фигурируют автомобили со средним пробегом в `5.57^4 миль`. Так же мы видим, что в продажу выставляют авто с 0 пробегом. Кроме того есть авто, пробег у которого `1,11^6 миль`!.
* **Попадал ли автомобиль в ДТП.** Статистика говорит, что в объявлениях редко указывают, что авто попадал в ДТП.
* **Принадлежал ли автомобиль одному собственнику.** Этот столбец говорит о том, сколько у автомобиля было собственников. Если 0, то одно лицо, в противном случае было несколько собственников у авто. Как показывает статистика, в среднем мы сталкиваемся с авто, у которых было несколько собственников.
* **Использование в личных целях.** Статистика показывает, что чаще всего пользуются автомобилем в личных целях, около `65%`.
* **Рейтинг продавца.** Продавец в среднем чаще всего выставляют автомобиль с рейтингом `4.15`, а самой минимальной - `1.00`.
* **Рейтинг водителя.** Водитель в среднем чаще всего выставляют автомобиль с рейтингом `4.62`, а самой минимальной - `1.00`.
* **Количество отзывов об автомобилях, оставленных водителями.** Водитель в среднем чаще всего выставляют автомобиль с рейтингом `4.62`, а самой минимальной - `1.00`.
* **Снижение цены по сравнению с начальной ценой.** В среднем в объявлениях видно, что начинают падать на сумму в `1007 долларов`, самое большое понижение - `170995 долларов`, минимальный шаг - `100 долларов`.
* **Цена автомобиля.** В среднем мы можем увидеть в объявлениях, что продают автомобилей с `3.64^4 долларов`, а самой максимальной - `10 00 000 000 долларов`, минимальной - `1 долларов`. Наверное совсем хлам...)

***Основываясь на эти данные, можем сделать промежуточный вывод:***

В датасете содержатся широкий диапазон промежутков годов выпуска автомобилей, начиная с 1915 по 2023 года. Да в объявлениях выставляются в продажу также и раритетные автомобили, что видимо и обусловлено высокой их стоимостью. Также в объявлениях вытавляют автомобилей с большими пробегами, которые скорее всего нуждается в востановительных/технических ремонтах, не говоря уж о нескольких собственниках у таких авто. Скорее всего в объвлениях так же выставляют служебные автомобили: фургоны, пикапы, т.к. процент пользования в личных целях невысок (около 65%). Самое интересное, что продавцы оставляют отзыв по автомобилям с оценкой ниже по сравнению с отзывами водителей. Отсюда следует следющая причинно-следсвенная связь - понижение цены по сравнению с начальной стоимостью автомобилей.

**2.5 Подготовка данных для работы**

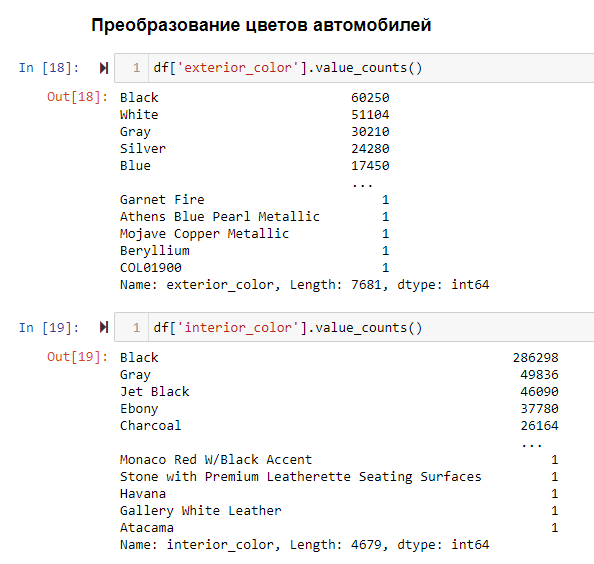
Преобразование в другой тип:

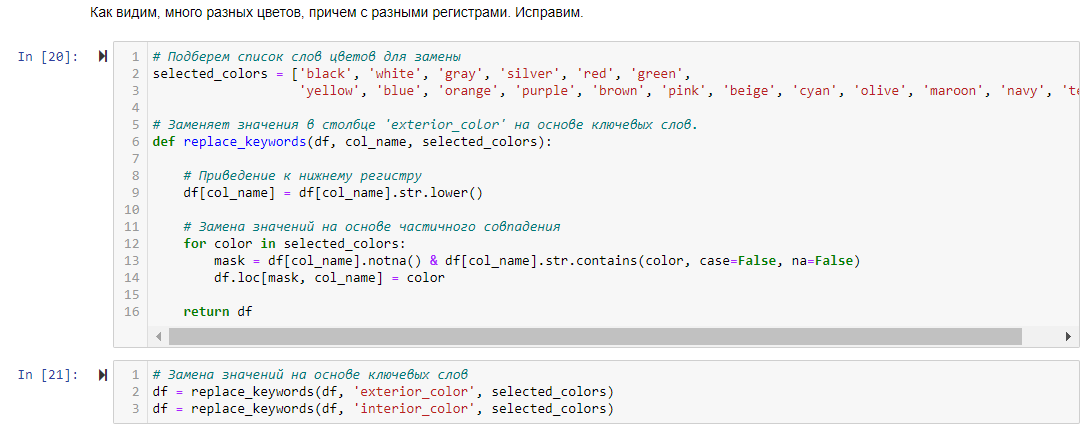
У нас в датасете есть колонка «mpg» не численного типа, которого нужно преобразовать в другой тип для проверки дальнейшей гипотезы.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Также в датасете у нас есть разные названия цветов автомобиля, поэтому необходимо категорировать их.





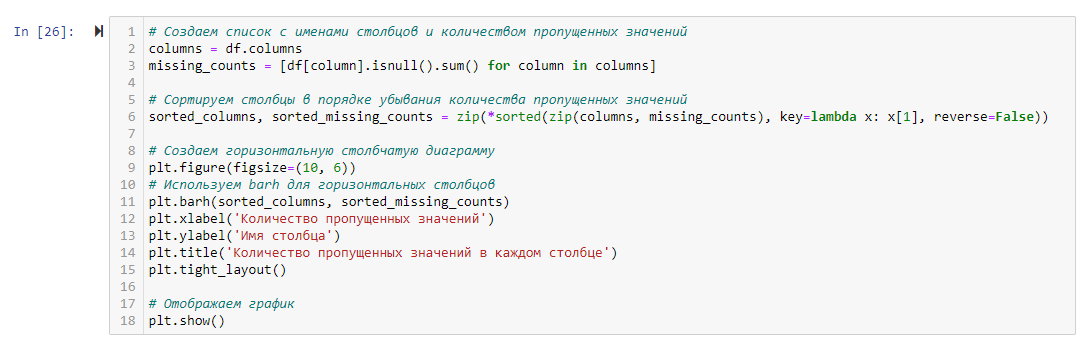
Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рассмотрим пропущенные значения:



Изображение выглядит как текст, снимок экрана, дисплей, число

Автоматически созданное описание

Как видно присутствуют огромные пропуски в столбцах: снижение стоимости, рейтинг продавца и кол-во миль. Заполним пропуски медианными значениями только для численных типов, а у остальных - устраним.

Медианными значениями заполняем, потому что они менее чувствительны к выбросам.

Изображение выглядит как текст, программное обеспечение, веб-страница, Значок на компьютере

Автоматически созданное описание

И снова создадим функцию, которая будет автоматизировать эту работу:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Снова проверим пропущенные значения:

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Посчитаем количество дубликатов:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Удаление выбросов:

Перед удалением нужно заново рассмотреть описательную статистку, чтобы выявить наличие выбросов и устранить их.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Стоимость автомобиля:

По описательной статистке видно, что есть выброс в стоимости автомобиля. Проверим на графике.

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Как видим - есть выброс, устраним его

Изображение выглядит как текст, линия, График, Шрифт

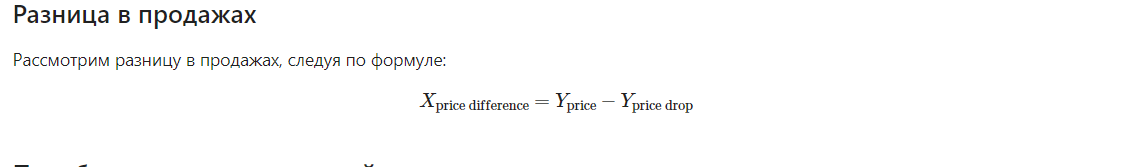
Автоматически созданное описание

График получился более читаемым.

**2.6 Агрегирование данных**

Из нашего датасета рассмотрим разницу в продажах, агрегируя нужные колонки.

Рассмотрим разницу в продажах, следуя по нижеприведенной формуле:



Добавим новый столбец с названием «Price difference»: разница начальной и цены после понижения.

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

Можно сказать, что при "понижение цены" стоимости автомобиля продавцом по какой то причине произошло ее увеличение, рассмотрим ниже сортировку по возрастанию разницы стоимости.

Изображение выглядит как текст, снимок экрана, число, программное обеспечение

Автоматически созданное описание

Рассмотрим график разницы стоимости автомобиля

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Изображение выглядит как диаграмма, График, снимок экрана, линия

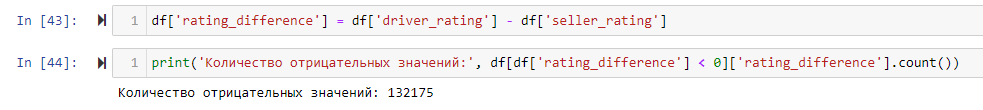
Автоматически созданное описание

На этом графике можно заметить (желтые вертикальная линия), что в некоторых местах изначальная стоимость автомобиля была высокой, а затем значительно изменилась в меньшую сторону.

**Разница в рейтингах, оставленных водителями и продавцами**

Рассмотрим разницу в рейтингах, следуя по формуле:





Рассмотрим только ТОП-10 данных с отрицательными значениями

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

Как и здесь видим, что по какой то причине данные рейтинга получились сильно разными.

Тогда рассмотрим график.

Изображение выглядит как текст, Шрифт, снимок экрана, линия

Автоматически созданное описание

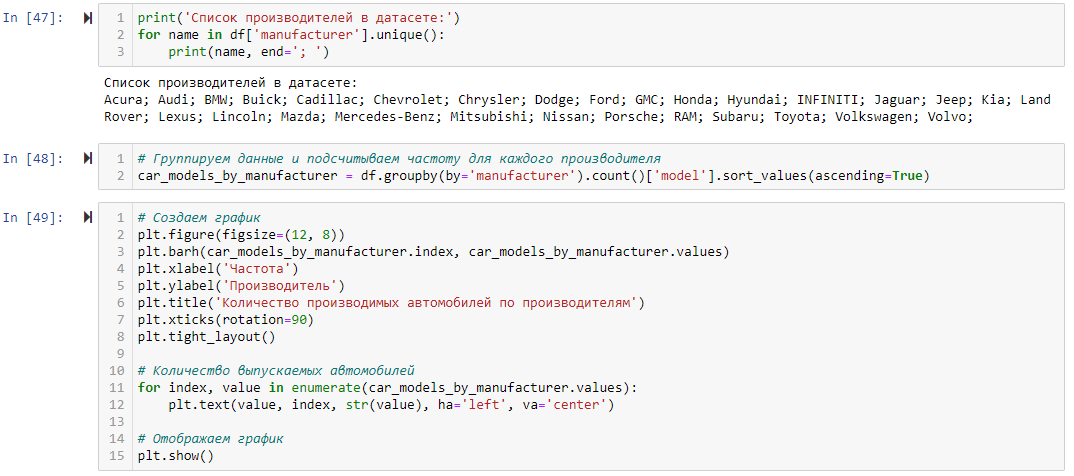
Изображение выглядит как снимок экрана, текст, линия, График

Автоматически созданное описание

По данным на графике видим, что существуют значимая разница в рейтинге между водителями и продавцами. Однако, стоит обратить внимание, что именно водители чаще всего оставляют высокий рейтинг нежели продавцы.

**Количество выпускаемых моделей производителями**

Рассмотрим список разных моделей, которые производители успели выпустить за промежуток в течении года



Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Вывод:

1. Произведена оптимизация памяти и изменение типов данных для дальнейшего анализа;
2. Обработаны пропущенные значения;
3. Обработаны дубликаты.

**2.7 Исследовательский анализ данных**

**Проверка гипотез.**

**Рассмотрим следующие гипотезы:**

* Влияние характеристик на цену автомобиля:
  + Гипотеза: Автомобили с более низким пробегом имеют более высокую цену.
  + Гипотеза: Автомобили с более свежими годами выпуска стоят дороже.
  + Гипотеза: Рейтинг продавца и рейтинг водителей связаны с ценой автомобиля.
* Сравнение цвета кузова и цвета салона с ценой:
  + Гипотеза: Автомобили с определенными цветами кузова и салона стоят дороже.
* Исследование взаимосвязей между рейтингами продавца и водителей:
  + Гипотеза: Рейтинг продавца коррелирует с рейтингом автомобиля, данным водителями.
* Анализ изменения цен по годам:
  + Гипотеза: Средняя цена на автомобили меняется с течением времени.
* Влияние типа топлива и трансмиссии на эффективность топливо потребления (mpg):
  + Гипотеза: Автомобили с определенным типом топлива имеют более высокую эффективность топливо потребления.
  + Гипотеза: Автомобили с определенным типом трансмиссии имеют более высокую эффективность топливо потребления.

**Гипотеза 1: Автомобили с более низким пробегом имеют более высокую цену.**

Сначала рассмотрим корреляцию тепловой карты, чтобы убедиться в них корреляционных значений.

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

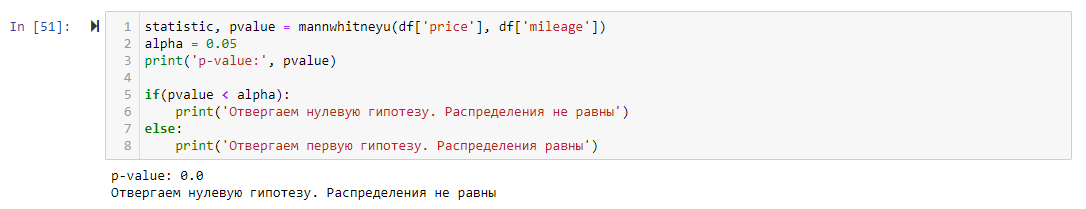
Изображение выглядит как текст, снимок экрана, число, диаграмма

Автоматически созданное описание

В данном случае корреляция говорит, что значение получилось отрицательным, следовательно, с увеличением пробега стоимость автомобиля снижается. И в противном случае, чем дороже автомобиль, тем выше вероятность того, что у автомобиля небольшой пробег.

Попробуем применять математический метод, а именно, равенство распределений выборок тестом Манна-Уитни, чтобы выявлять, распределения равны или нет.

* H0 - распределения равны
* H1 - распределения не равны



p-value получилось слишком маленьким или стремится к нулю. Рассмотрим критерий Стьюдента.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

В данном случае получен очень маленький p-value, что говорит о том, что различия между данными статистически значимы. Это означает, что существует значимая разница между средними значениями двух выборок и эта разница не может быть объяснена случайными факторами.

Также можно рассмотреть значение t-статистики. Если оно больше 2 или меньше -2 (при уровне значимости 0,05), то различия между выборками считаются статистически значимыми.

Подытожим, это означает, что нулевая гипотеза о равенстве распределений выборок отвергается на уровне значимости 0,05. То есть, существует статистически значимая разница между ценами и пробегом автомобилей. И подтвердим гипотезу на основе вывода корреляции тепловой карты.

**Гипотеза 2: Автомобили с более свежими годами выпуска стоят дороже. Рассмотрим эту гипотезу и проверим.**

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

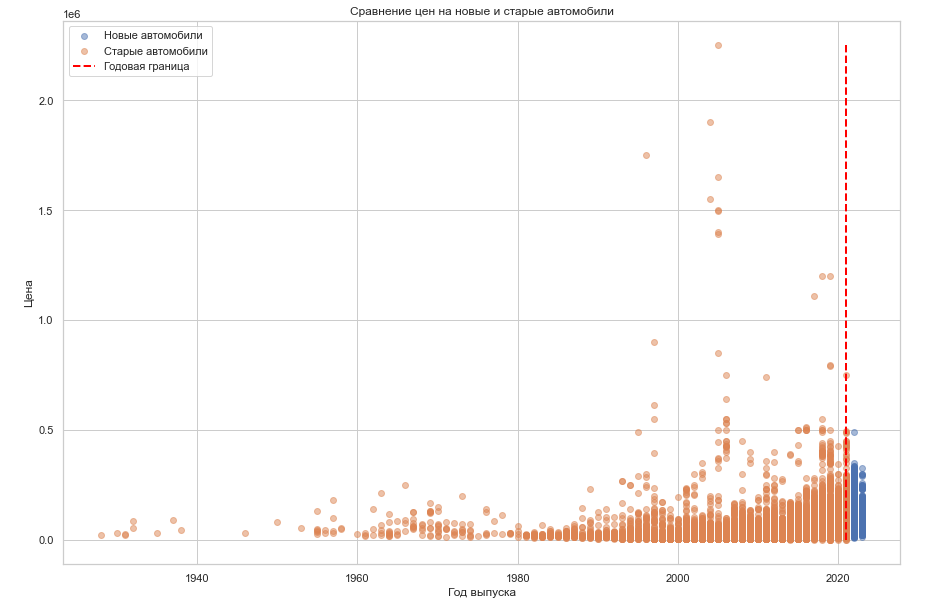
Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Здесь корреляция говорит, что значение получилось слабоположительной, следовательно, с уменьшением года выпуска увеличивается и стоимости автомобиля. И в противном случае, чем старее автомобиль, тем дешевле. Рассмотрим график.

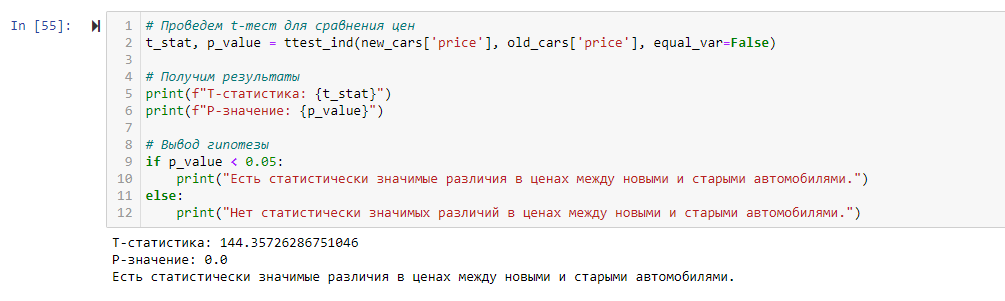
Изображение выглядит как текст, снимок экрана, Шрифт, веб-страница

Автоматически созданное описание



Попробуем применять математический метод, а именно, равенство распределений выборок тестом "t-тест" для сравнения цен

* H0 - Нет статистически значимых различий в ценах между новыми и старыми автомобилями.
* H1 - Есть статистически значимые различия в ценах между новыми и старыми автомобилями



Результат t-теста позволяет отвергнуть нулевую гипотезу о том, что нет статистически значимых различий в ценах между новыми и старыми автомобилями. P-значение равно 0.0, что гораздо меньше уровня значимости 0.05.

Таким образом, на основе проведенного t-теста можно сделать вывод, что существует статистически значимая разница в ценах между автомобилями нового и старого года выпуска. Поскольку T-статистика положительна, это также указывает на то, что средние цены у новых автомобилей выше, чем у старых.

Однако, несмотря на статистическую значимость, важно помнить, что корреляция не означает причинно-следственной связи, и другие факторы также могут влиять на цены автомобилей.

**Гипотеза 3: Рейтинг продавцов и рейтинг водителей связаны с ценой автомобиля.**

. Рассмотрим эту гипотезу и проверим.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Изображение выглядит как снимок экрана, текст, Прямоугольник, прямоугольный

Автоматически созданное описание

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

Изображение выглядит как линия, текст, График, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Изображение выглядит как линия, График, текст, диаграмма

Автоматически созданное описание

Как видим, что корреляция получилась слабоположительной, следовательно, присутствует незначительный рост со стоимостью автомобиля и с рейтингом продавца. Например, если продавцу понравилась эта машина, значит возможно будет продавать его чуть подороже.

**Сравнение цвета кузова и цвета салона с ценой:**

**Гипотеза 4: Автомобили с определенными цветами салона стоят дороже**.

Поскольку нам известно, что слова — это не численный тип, поэтому сделать матрицу тепловой карты невозможно. На эту помощь придет декодирование признаков - get\_dummies, который будет заменять слова на уникальный номер.

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание



У нас получились 2 корреляции:

* **Положительная корреляция**:

Black (черный), Green (зеленый), White (белый), Yellow (желтый) - эти цвета интерьера имеют положительную корреляцию со стоимостью автомобиля, что может означать, что автомобили с этими цветами интерьера могут стоить дороже.

* **Отрицательная корреляция**:

Silver (серебристый), Red (красный), Gray (серый), Blue (синий) - эти цвета интерьера имеют отрицательную корреляцию со стоимостью автомобиля. Возможно, автомобили с этими цветами интерьера могут быть более доступными или иметь более низкую стоимость.

Рассмотрим нулевую и альтернативную гипотезу.

Рассмотрим нулевую и альтернативную гипотезу.

* **H0 (нулевая гипотеза):** Существует статистически значимая разница в стоимости автомобилей в зависимости от цвета кузова и цвета салона.
* **H1 (альтернативная гипотеза):** Не существует статистически значимой разницы в стоимости автомобилей в зависимости от цвета кузова и цвета салона.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

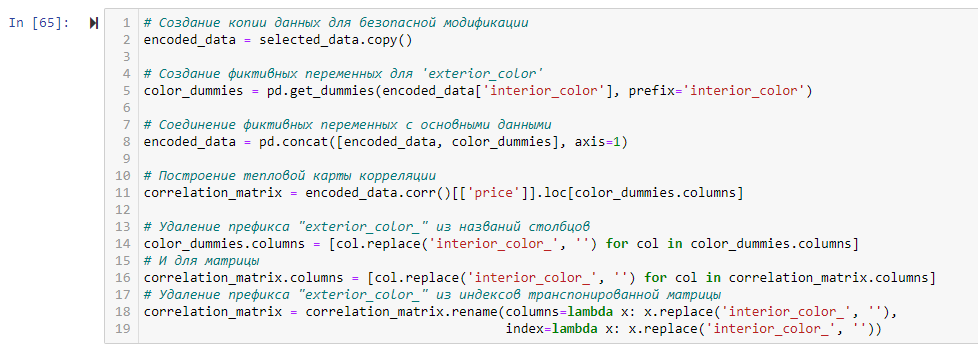
Так как P-значение меньше общепринятого уровня значимости, мы отвергаем нулевую гипотезу (H0). Это говорит о том, что существует статистически значимая разница в стоимости между автомобилями с черным и белым цветами кузова. Таким образом, можно сделать вывод о том, что цвет кузова влияет на цену автомобиля.

**Гипотеза 5: Автомобили с определенными цветами кузова стоят дороже**.

Рассмотрим эту гипотезу и проверим.

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание



Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Изображение выглядит как снимок экрана, текст, линия, число

Автоматически созданное описание

Исходя из результатов анализа, можно сделать следующие выводы:

* **Цвет салона влияет на стоимость автомобиля:** Наблюдаются статистически значимые различия в стоимости автомобилей в зависимости от цвета салона. Например, автомобили с серым цветом салона имеют существенно более низкую стоимость, чем с автомобилями с черным, белым или красным цветами салона.
* **Наибольшее влияние оказывают цвета с черной и красной обивкой:** Автомобили с черным и красным цветами салона имеют более высокие цены по сравнению с другими цветами салона. Это может быть связано с предпочтениями покупателей или особенностями рынка.
* **Стоимость может зависеть от индивидуальных предпочтений покупателей:** Например, автомобили с бежевым, белым и синим цветами салона также могут иметь более высокую цену, чем автомобили с серым или оливковым цветами.

Рассмотрим нулевую и альтернативную гипотезу.

* **H0 (нулевая гипотеза):** Средние стоимости автомобилей для разных цветов кузова равны.
* **H1 (альтернативная гипотеза):** Существует статистически значимая разница в средних стоимостях автомобилей для разных цветов кузова.

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

Так как P-значение меньше общепринятого уровня значимости, мы отвергаем нулевую гипотезу (H0). Странно, но можно предположить, что в данном случае у нас нет достаточных данных для выявления статистически значимых различий в средних стоимостях автомобилей для разных цветов салона. Рассмотрим еще другую гипотезу.

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Так как P-значение меньше общепринятого уровня значимости, мы отвергаем нулевую гипотезу (H0). Это говорит о том, что существует статистически значимая разница в средних стоимостях автомобилей для разных цветов кузова. Таким образом, можно сделать вывод о том, что цвет салона влияет на цену автомобиля.

Исследование взаимосвязей между рейтингами продавцов и водителей:

Гипотеза 6: Рейтинг продавцов коррелирует с рейтингом автомобиля, данным им водителями.

Рассмотрим эту гипотезу и проверим

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Корреляция получилась слабоположительной, рассмотрим далее.

Проверим гипотезу, выбрав тип топлива бензина и гибрид.

* Нулевая гипотеза (H0): Рейтинг продавца не коррелирует с рейтингом водителей
* Альтернативная гипотеза (H1): Рейтинг продавца коррелирует с рейтингом водителей.

Изображение выглядит как текст, снимок экрана, Шрифт, веб-страница

Автоматически созданное описание

Результаты статистического теста свидетельствуют о том, что существует статистически значимая корреляция между рейтингами продавцов и водителей.

**Анализ изменения стоимости по годам:**

**Гипотеза 7: Средняя стоимость на автомобили меняется с течением**

Рассмотрим эту гипотезу и проверим.

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, График, линия, диаграмма

Автоматически созданное описание

По графику видно, что раритетные автомобили более дорогие, а самым дешевым оказались автомобили с промежутком между 2000 по 2010 года. Проверим гипотезу математическим способом.

Проверим гипотезу, выбрав тип топлива бензина и гибрид.

* **Нулевая гипотеза (H0):** Существует статистически значимое изменение средней стоимости на автомобили с течением времени.
* **Альтернативная гипотеза (H1):** Нет статистически значимого изменения средней стоимости на автомобили с течением времени

Изображение выглядит как текст, Шрифт, число, линия

Автоматически созданное описание

Результаты теста показали, что существует статистически значимое изменение средней стоимости на автомобили с течением времени.

**Влияние типа топлива и трансмиссии на эффективность топливопотребления (mpg):**

**Гипотеза 8: Автомобили с определенным типом топлива имеют более высокую эффективность топливо потребления.**

Рассмотрим эту гипотезу и проверим. Но сначала рассмотрим, какие типы топлива популярны в Америке

Изображение выглядит как текст, Шрифт, веб-страница, программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описание

Как и видим, что в Америке популярный тип топлива - **бензин**. Второе место занимает - **гибрид**. А дальше уже смешанный тип топлива и дизель.

Проверим гипотезу, выбрав тип топлива бензина и гибрид.

* **Нулевая гипотеза (H0):** Средние значения mpg в группах (типах топлива) равны, то есть нет статистически значимой разницы между типами топлива в отношении эффективности топливо потребления.
* **Альтернативная гипотеза (H1):** Средние значения mpg в группах (типах топлива) различаются статистически значимо, что подразумевает, что существует влияние типа топлива на эффективность топливо потребления.

Чем больше значение F-статистики, тем более вероятно, что существует статистически значимая разница между группами.

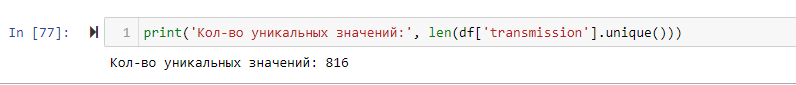
Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Результаты теста показали, что средние значения mpg\_city для бензиновых и гибридных автомобилей различаются статистически значимо, что подтверждает альтернативную гипотезу и опровергает нулевую гипотезу.

**Гипотеза 9: Гипотеза: Автомобили с определенным типом трансмиссии имеют более высокую эффективность топливо потребления.**

Для начала рассмотрим уникальные значения по названию и моделей трансмиссии.



Уникальных значений очень много, поэтому рассмотрим их до ТОП-25 мест.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Рассмотрим нулевую и альтернативную гипотезу для анализа:

* **Нулевая гипотеза (H0):** Средние значения эффективности топливопотребления (mpg) одинаковы для всех типов трансмиссии. То есть, нет статистически значимой разницы в mpg между автомобилями с разными типами трансмиссии.
* **Альтернативная гипотеза (H1):** Средние значения эффективности топливопотребления (mpg) различаются для по меньшей мере двух типов трансмиссии. То есть, существует статистически значимая разница в mpg между какими-то из типов трансмиссии.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, документ

Автоматически созданное описание

* **F-статистика: 341.08650483653787** - это высокое значение F-статистики, которое указывает на статистически значимую разницу в средних значениях mpg\_city между разными типами трансмиссии. В данном случае, высокое значение F-статистики говорит о том, что существует статистически значимая разница в эффективности топливопотребления в зависимости от типа трансмиссии.
* **p-значение: 0.0** - нулевое p-значение означает, что вероятность получить такие или более экстремальные результаты, как в ваших данных, при условии, что нулевая гипотеза (отсутствие разницы в средних значениях) верна, крайне низка. Это говорит о том, что разница в mpg\_city между разными типами трансмиссии статистически значима.

Интерпретация:

Исходя из результатов ANOVA теста, у вас есть статистически значимые доказательства в пользу того, что средние значения эффективности топливо потребления (mpg\_city) различаются для по меньшей мере двух типов трансмиссии. Таким образом, гипотеза о влиянии типа трансмиссии на эффективность топливо потребления подтверждена, и средняя эффективность топливо потребления различается в зависимости от типа трансмиссии.

**Корреляционный анализ**

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Давайте условно назовем сильную корреляцию, когда значение превышает 0.5 / -0.5.

Здесь в большинствах случаях корреляция равна нулю или слабо положительная / отрицательная.

Есть то, что стоит обратить внимание на очевидные вещи:

* **Год выпуска и Пробег:** Присутствует средняя отрицательная корреляция между годом выпуска автомобиля и его пробегом, что логично, так как новые автомобили обычно имеют меньший пробег.
* **Год выпуска и Цена:** Наблюдается положительная корреляция между годом выпуска и ценой. Это также ожидаемо, поскольку новые автомобили, как правило, дороже.
* **Пробег и Цена:** Существует отрицательная корреляция между пробегом и ценой, что подтверждает интуитивное представление: чем меньше пробег, тем выше цена.
* **Оценка продавца и Цена:** Отмечается слабая положительная корреляция между оценкой продавца и ценой. Такие автомобили, вероятно, более дорогие.
* **Оценка водителя и Цена:** Слабая положительная корреляция между оценкой водителя и ценой. Автомобили с более высокими оценками водителей, вероятно, также более дорогие.
* **Число отзывов водителей и Цена:** Отсутствует значительная корреляция между числом отзывов водителей и ценой.
* **Рейтинг продавца и Рейтинг водителя:** Существует слабая положительная корреляция между рейтингами продавца и водителя. Это может свидетельствовать о том, что продавцы и водители оценивают автомобили схожим образом.
* **Разница в рейтинге и Цена:** Отсутствует значительная корреляция между разницей в рейтинге и ценой.
* **Новизна автомобиля и Цена:** Присутствует положительная корреляция между новизной автомобиля и его ценой.

# **2.8 Общий вывод**

Перед анализом было оптимизированно использование памяти путем преобразования данных к более корректным типам. Также были добавлены необходимые столбцы при помощи агрегирования данных, что в дальнейшем нам облегчили провести исследование и подтвердить гипотезу, а также получить ценные знания в предметной области «Автомобили с пробегом».

В ходе анализа было выявлено:

* Самым дешевым автомобилем будет только тот, пробег у которого будет больше всего;
* Существует высокая вероятность того, что автомобиль свежего года выпуска при этом будет стоить дешевле.
* Не исключено, что продавец и покупатель, оставленные высокий рейтинг автомобиля, будут продавать дороже.
* Самый распространенный цвет кузова автомобиля – черный, белый и желтый. Менее популярный – серебристый, красный и бордовый.
* Самый распространенный цвет салона автомобиля – красный, черный и коричневый. Менее популярный – серый, бежевый и оливковый.
* Есть слабая положительная корреляция между оценками водителей и продавцами.
* Средняя цена автомобиля меняется в течение года его выпуска.
* Существует и самый предпочитаемый тип топлива – бензин, энергоэффективность которого превосходит другие типы топлива.
* Автомобиль с автоматической коробкой переключении передачи расходует намного меньше топлива по отношению к автомобилям, у которых механическая коробка переключения передачи.

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «Автомобили с пробегом», в ходе которого были опровергнуты или подтверждены выдвинутые гипотезы бытующие на рынке.

Для выполнения данной работы было использовано программное обеспечение: Pandas, MatPlotLib, Seaborn, Numpy, Scipy на базе языка программирования Python. Также мы пользовались Jupyter Notebook, в нём мы писали код и анализировали диаграммы.

Оптимизация использования памяти, агрегирование данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. <https://pandas.pydata.org/>
2. <https://matplotlib.org/>
3. <https://www.geeksforgeeks.org/difference-between-null-and-alternate-hypothesis/>
4. <https://seaborn.pydata.org/>