|  |  |
| --- | --- |
|  | Министерство науки и высшего образования Российской Федерации  Федеральное государственное бюджетное образовательное учреждение  высшего образования  «Московский государственный технический университет  имени Н.Э. Баумана  (национальный исследовательский университет)»  (МГТУ им. Н.Э. Баумана) |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме \_\_\_\_\_\_\_ «Used Cars Dataset» – «Продажа поддержанных автомобилей»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5Ц-74Б \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А. В. Папин

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

Консультант \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

2023 г.

**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 ходе работы будут использоваться следующие библиотеки: Pandas, MatPlotLib, Seaborn, Numpy, Scipy. Анализ будет проводиться через программу Jupyter Python.

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](#_Toc152337654)

[1. Определение данных для анализа 4](#_Toc152337655)

[2. Описание данных 5](#_Toc152337656)

[3. Формулирование гипотез 7](#_Toc152337657)

[4. Изучение общей информации 7](#_Toc152337658)

[5. Подготовка данных для работы 12](#_Toc152337659)

[6. Агрегирование данных 20](#_Toc152337660)

[7. Исследовательский анализ данных 25](#_Toc152337661)

[8. Общий вывод 48](#_Toc152337662)

[ЗАКЛЮЧЕНИЕ 49](#_Toc152337663)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 50](#_Toc152337664)

# ВВЕДЕНИЕ

Для выполнения работы необходимо исследовать базу данных и выявить закономерности.

Поставим цели, которых мы хотим добиться, выполнив научно-исследовательскую работу студента.

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. **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.

1. **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, когда был выпущен автомобиль
4. mileage – миль, пройденных автомобилем с момента выпуска
5. engine – автомобильный двигатель
6. transmission – тип трансмиссии автомобиля
7. drivetrain – тип трансмиссии автомобиля
8. fuel\_type – тип топлива, которое потребляет автомобиль
9. mpg – 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 АВАРИИ
13. one\_owner – принадлежал ли автомобиль одному лицу
14. personal\_use\_only – использовался ли автомобиль только в личных целях
15. seller\_name – имя продавца
16. seller\_rating – рейтинг продавца
17. driver\_rating – рейтинг автомобиля, данный водителями
18. driver\_reviews\_num – количество отзывов об автомобилях, оставленных водителями
19. price\_drop – снижение цены по сравнению с начальной ценой
20. price – цена автомобиля

# 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 топливопотребления.

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

Гипотеза 6: Автомобили с определенными цветами кузова стоят дороже.

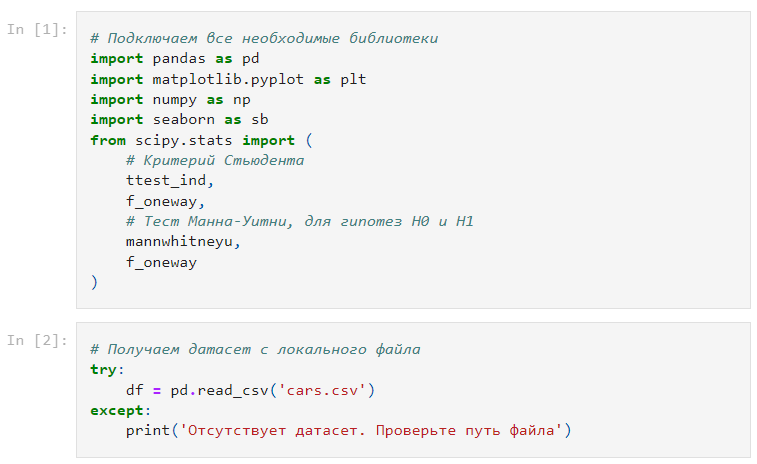
Гипотеза 7: Автомобили с определенными цветами салона стоят дороже.

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

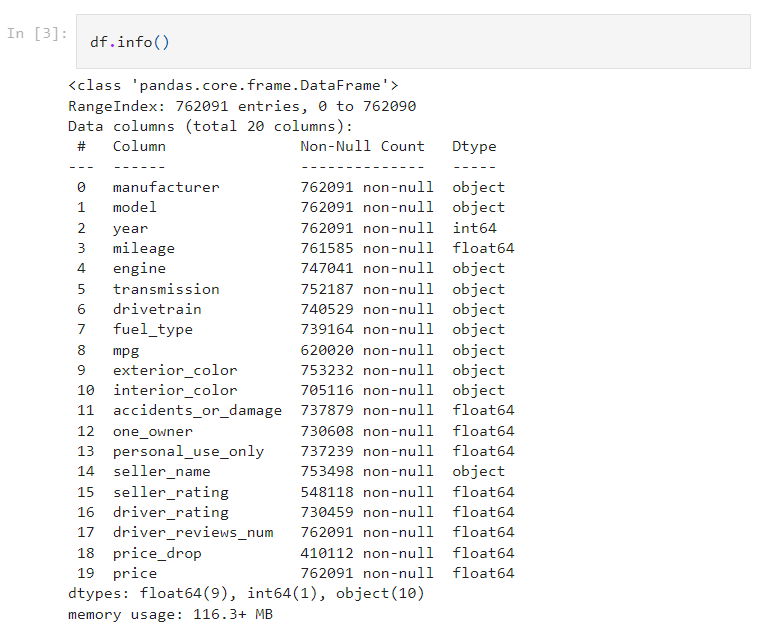
Гипотеза 9: Средняя цена на автомобили меняется с течением времени.

1. **Изучение общей информации**

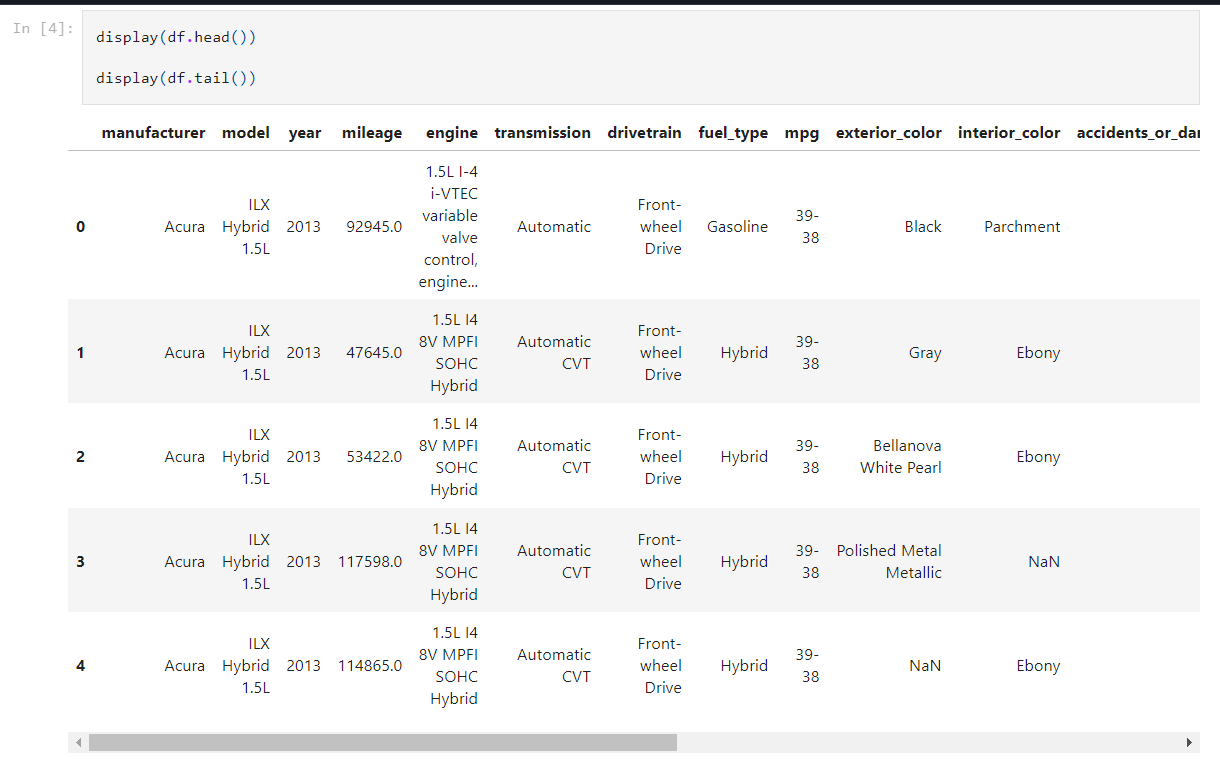
Загружаем датасет и подключаем необходимые библиотеки:



Получим информацию о датасете:



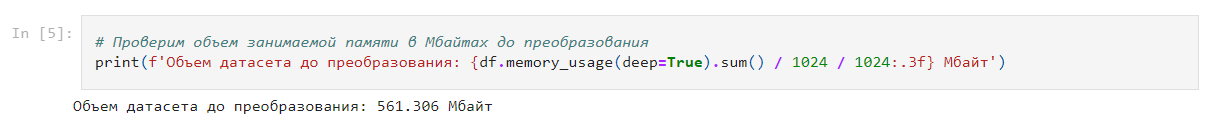
Здесь можно заметить, что в датасете содержатся единиц 762091 строк. А также имеют 3 различные типы: object, int64 и float64. В целях экономии памяти можно преобразовать в другие типы. Далее расмотрим содержимое в датасете, что там описываются.



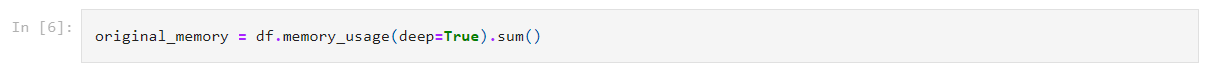


Здесь ничего особенного нет, пока не рассмотрим следующие этапы.

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»:



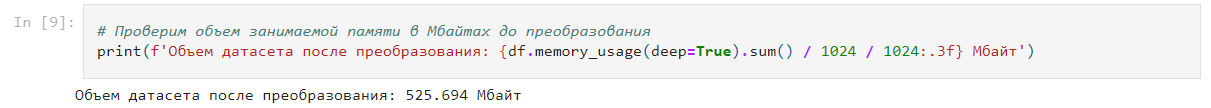
Сохраним объем датасета изначального размера для сравнения:



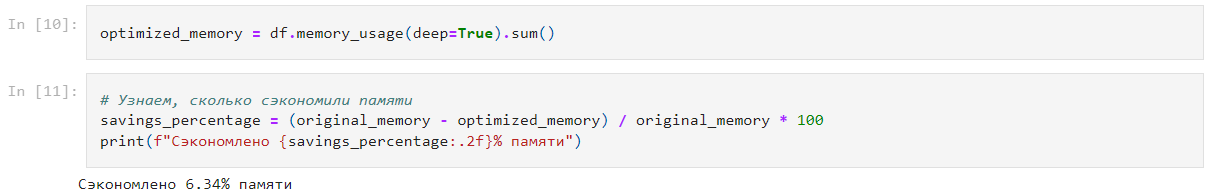
Создадим функцию, которая будет автоматизировать эту работу:



Рассмотрим полученный объем датасаета после преобразования:

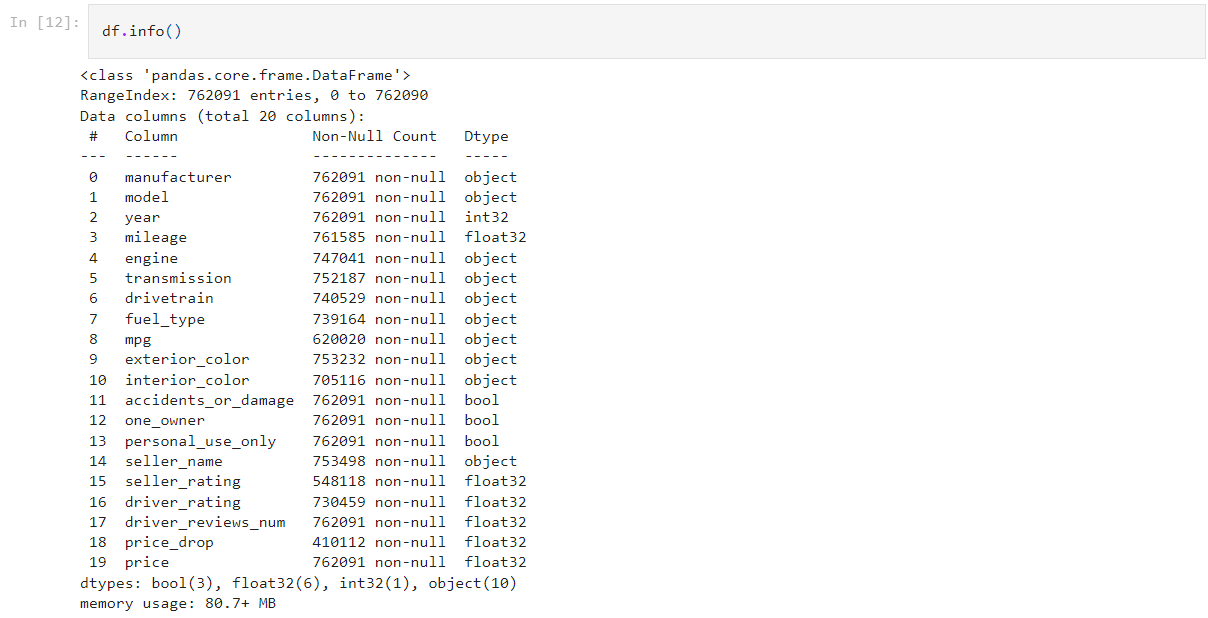


Проверим, на сколько сэкономили места:

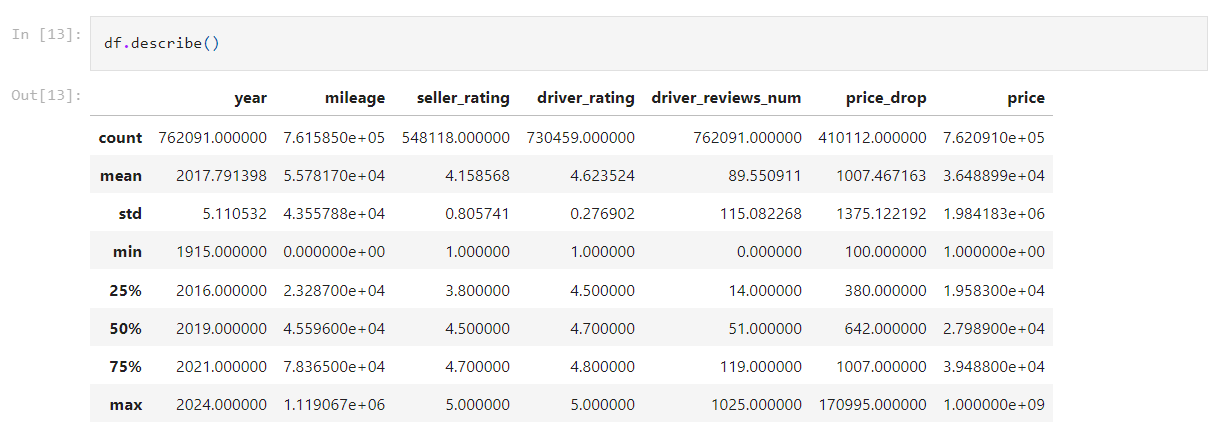


Сэкономили на 6,34% объема памяти.

Рассмотрим полученную информацию о датасете:



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



Здесь стоит обратить внимание на следующие колонки:

* **Год**. Мы чаще всего можем увидеть автомобиль, которая продается около 2017 года (среднее значение преобладает остальных). В объявлениях можем увидеть автомобиль с 1915 года.
* **Пройденный миль**. В объявлениях чаще всего выставляют автомобиль с 5.58^4 милях. Нельзя не отрицать, что в продажах выставляют автомобиль, которая ни разу не проехала. Существует автомобиль, которая проехала 1,11^6 миль.
* **Наличие авария автомобили**. Статистика говорит, что в объявлениях редко указывают, что автомобиль попадает в аварию. Мы можем сталкиваться с автомобилей, у которой была авария, с вероятностью около 22%.
* **Одно лицо у автомобилей**. Эта колонка говорит о том, что у этой автомобилей было только одно лицо - водитель. Если да, то одно лицо, в противном случае несколько лиц было у этой автомобили. Статистика говорит, что в среднем мы сталкиваемся с автомобилей, у которой было несколько лиц.
* **Пользование в личных целях**. Статистика говорит, что чаще всего пользуются автомобилей 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 мы можем увидеть в объявлениях, что продают автомобилей с 1007 долларов, а самой максимальной - 170995 долларов, минимальной - 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. Интересно узнать, какие же автомобили же.

Основываясь на это, можем сделать промежуточный вывод:

В датасете содержатся широкий диапазон промежутков года автомобилей, начиная с 1915 по 2024 года. Нельзя не отрицать, что в объявлениях выставляют продажи раритетных автомобилей, что было обусловлено высокой стоимостью. Также в объявлениях вытавляют автомобилей с большими пробегами, которая нуждается в технических ремонтах, не говоря уж о несколько лиц у этой автомобилей. Скорее всего в объвлениях выставляют служебные автомобили: фургоны, пикапы, т.к. процент пользования в личных целях невысок (около 65%). Самое удивительное, что продавцы оставили отзыв автомобиля ниже по сравнению с отзывом водителей. Поэтому отсюда следует причина - сильное понижение цены по сравнению с начальной стоимостью автомобилей.

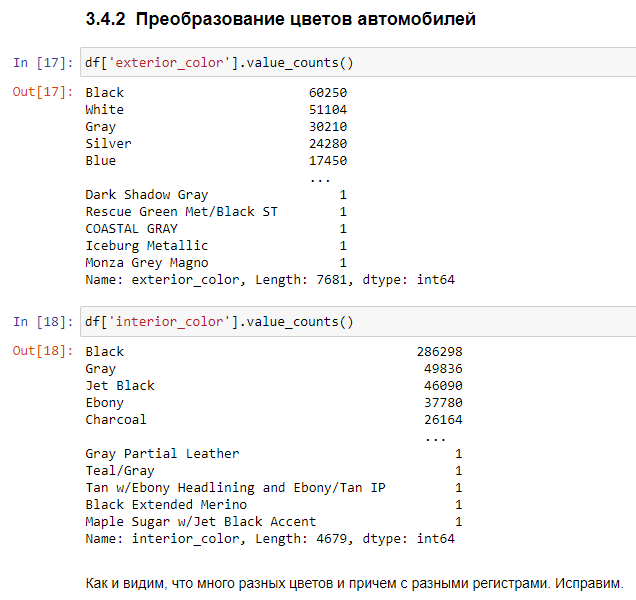
1. **Подготовка данных для работы**

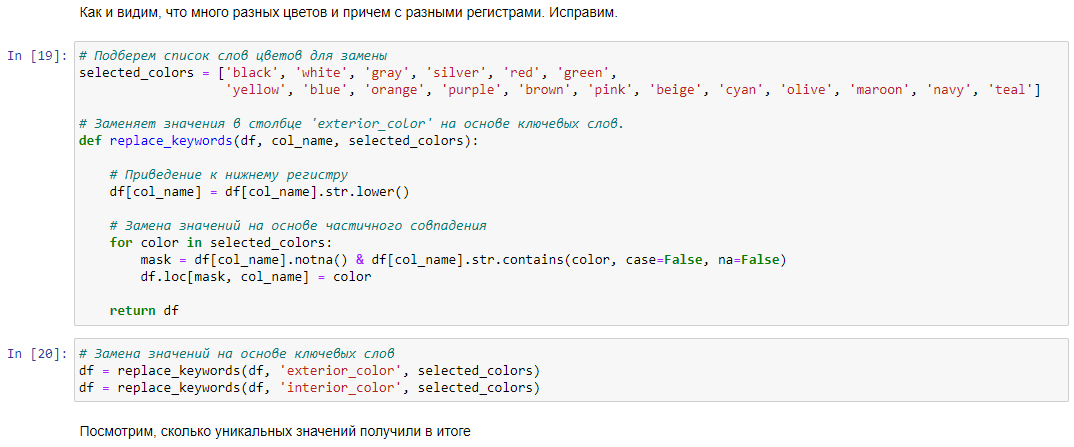
Преобразование в другой тип:

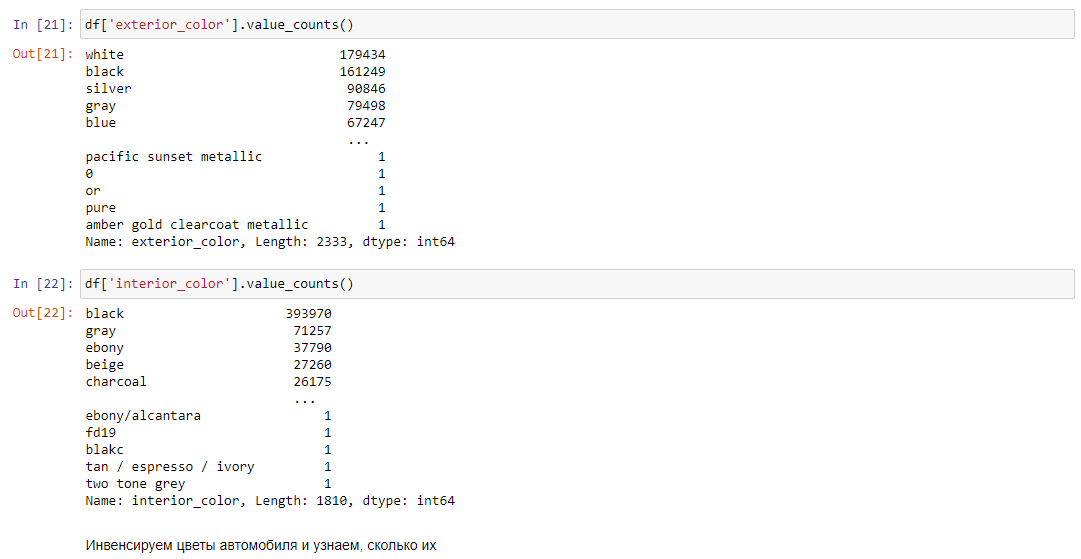
У нас в датасете есть колонка «mpg» не численного типа, которого нужно преобразовать в другой тип для проверки дальнейшей гипотезы.

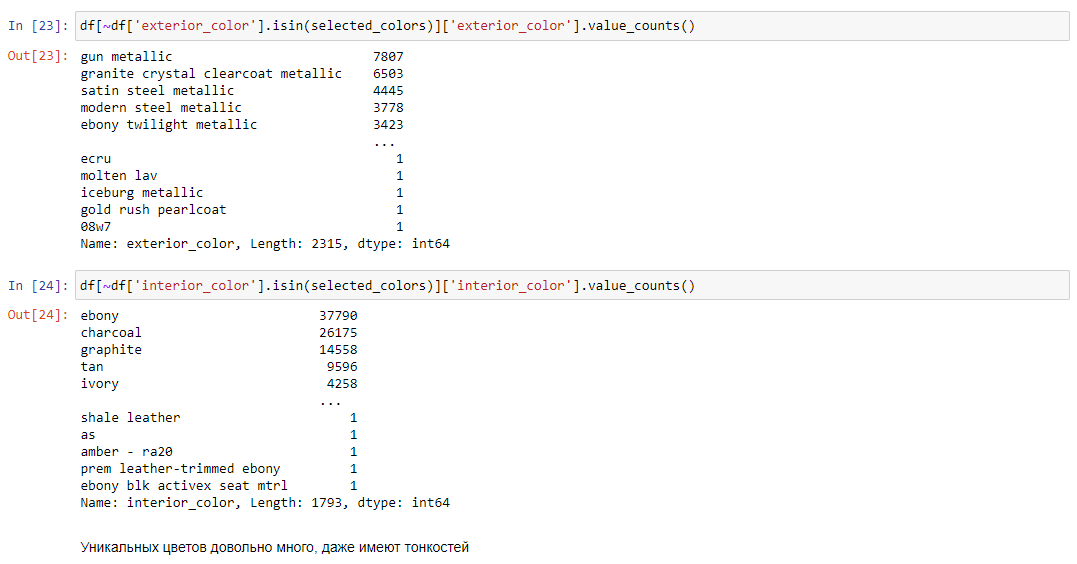


Также в датасете у нас есть разные названия цветов автомобиля, поэтому необходимо категорировать их.

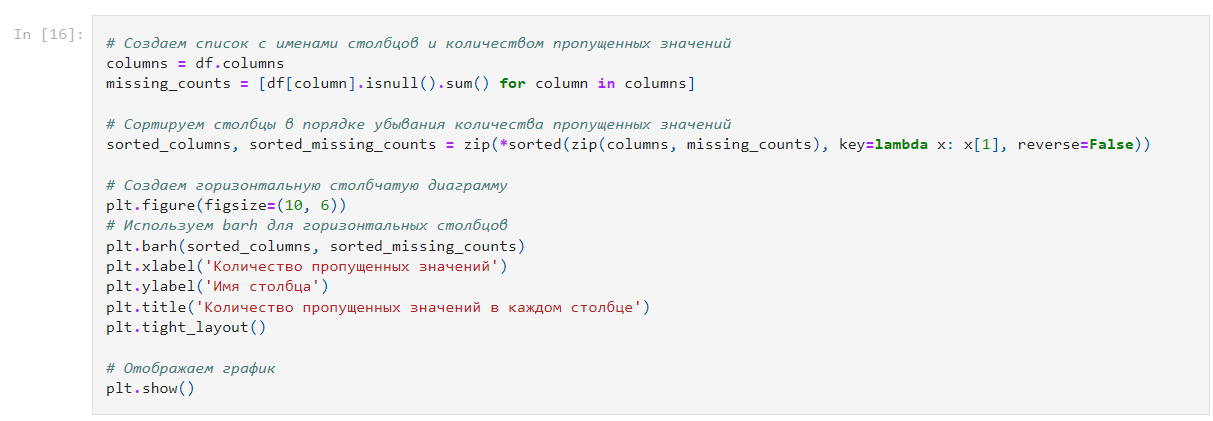


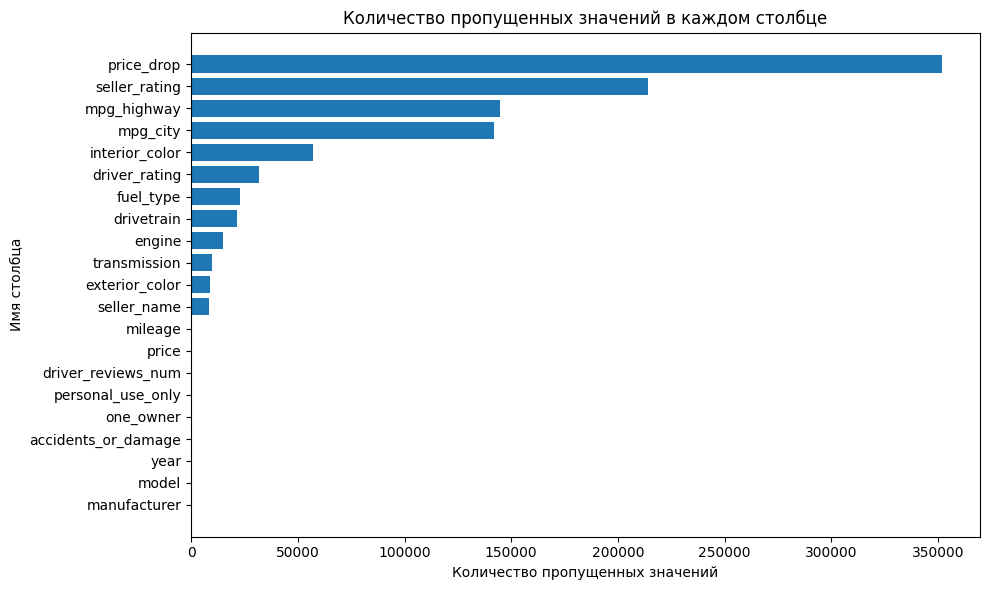






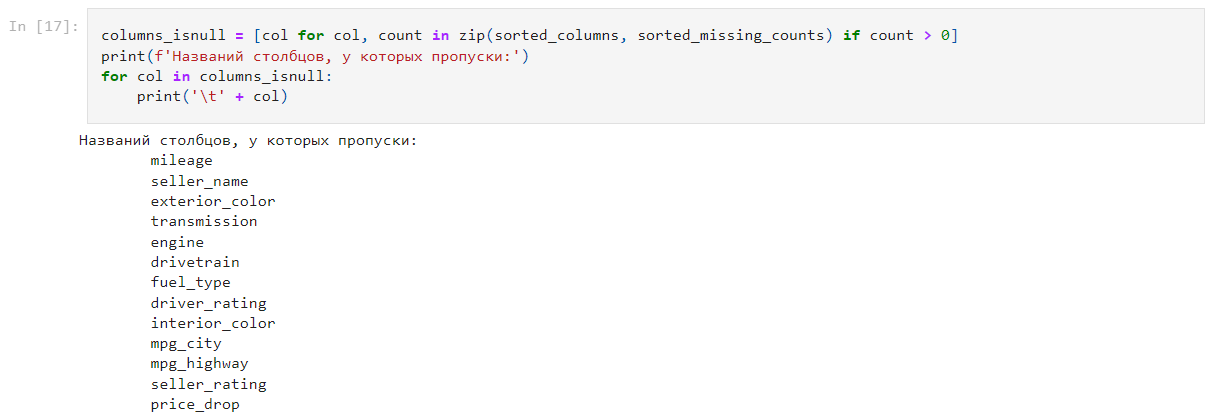
Рассмотрим пропущенные значения:





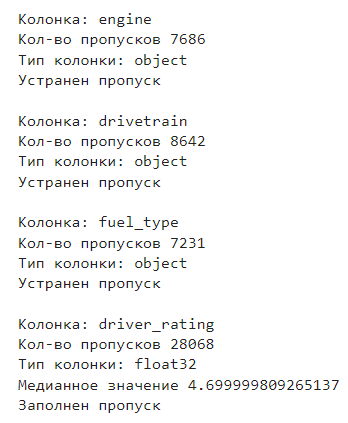
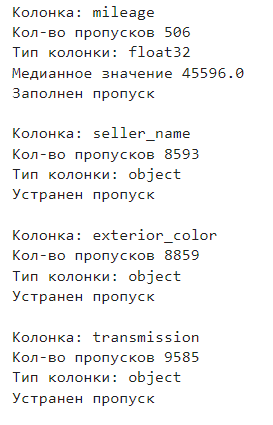
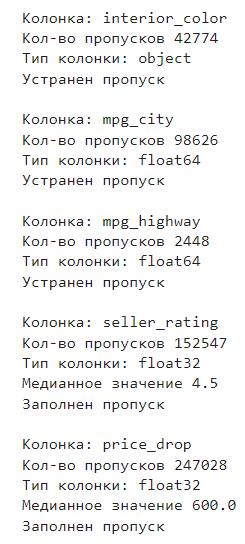
Как видим, что присутствуют огромные пропуски в столбцах: сниженная стоимость, рейтинг продавца и кол-во миль. Заполним пропуски медианными значениями только для численных типов, а остальных - устраним.

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

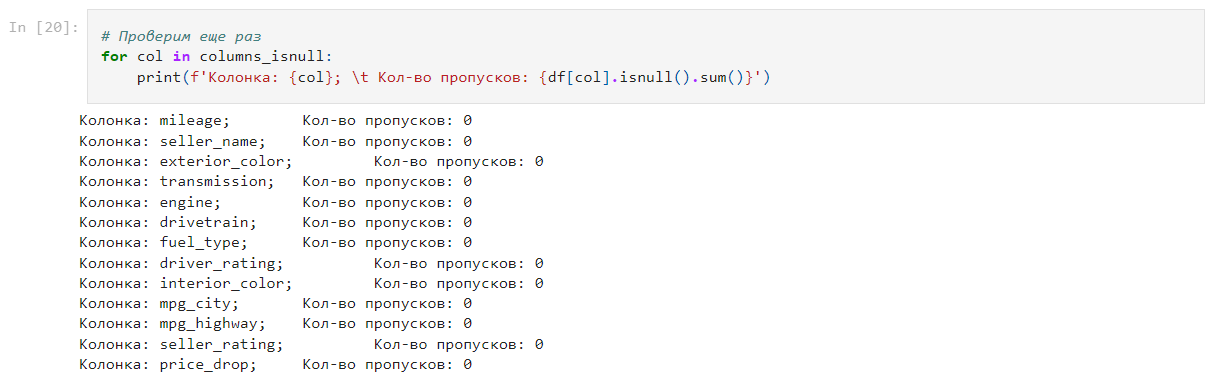


И снова создадим функцию, которая будет автоматизировать эту работу:

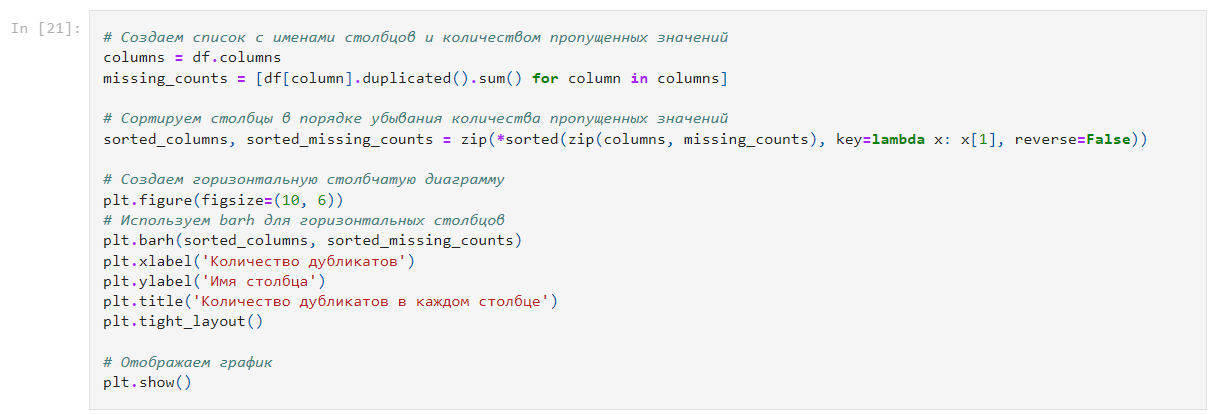


Снова проверим пропущенные значения:



Посчитаем количество дубликатов:



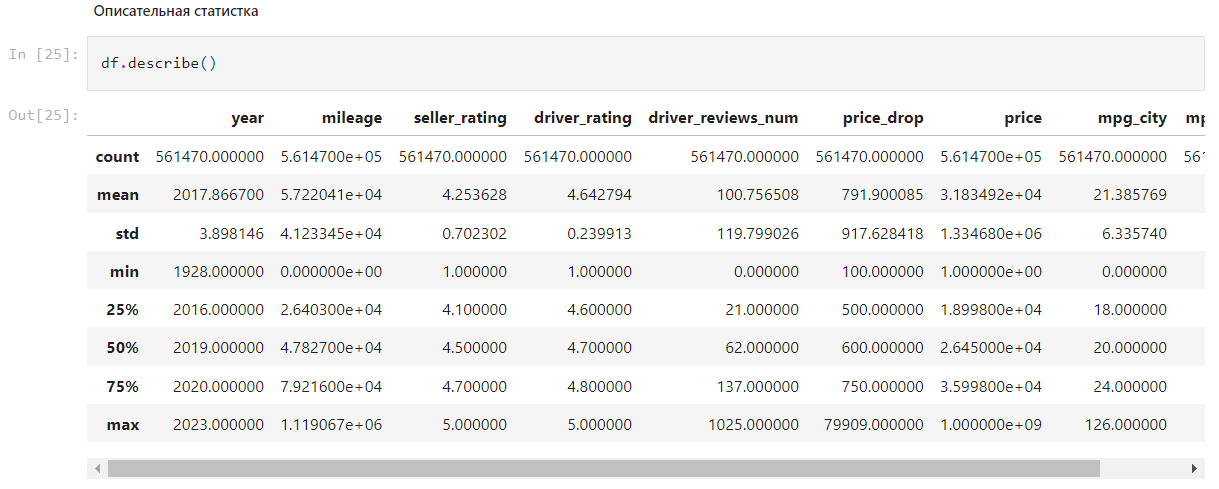


Как видим, что присутствуют очень много дубликатов в каждом столбике. Устраним их.



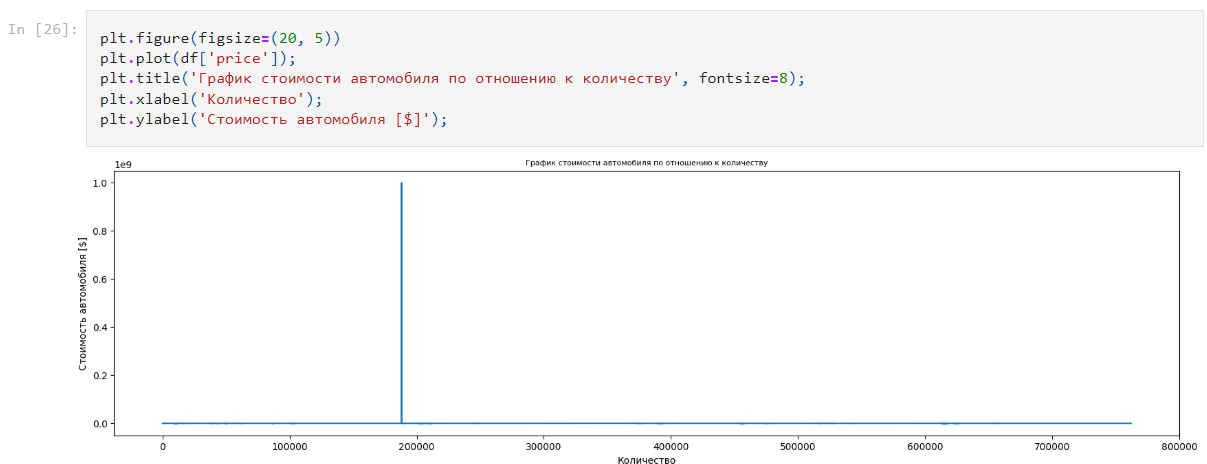
Удаление выбросов:

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



Стоимость автомобиля:

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



Как и видим, устраним их

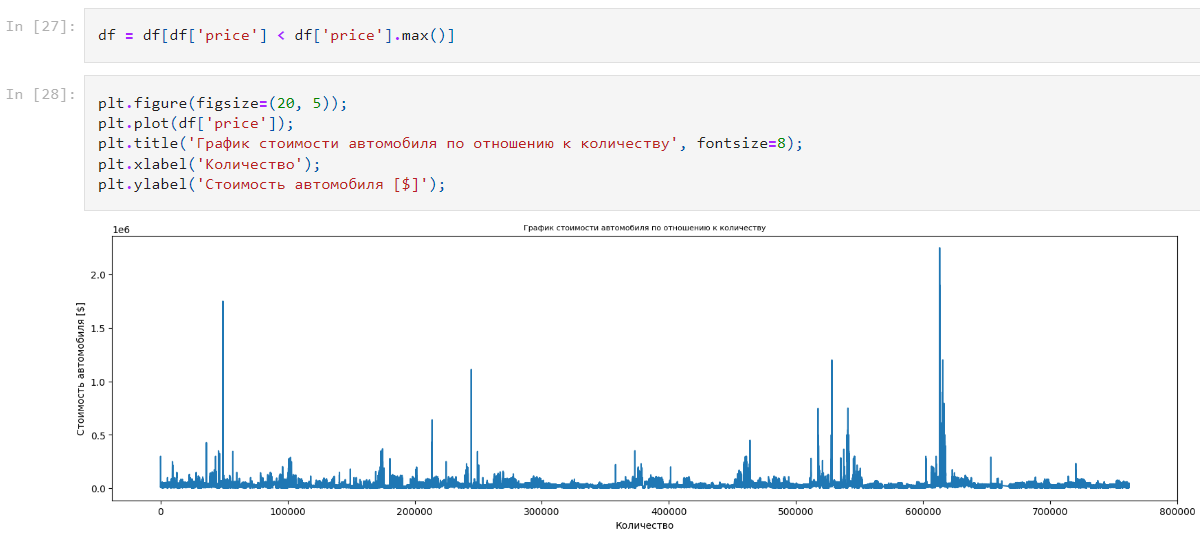
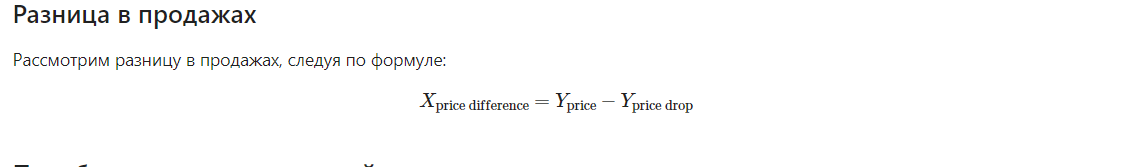


График получился более-менее адекватным.

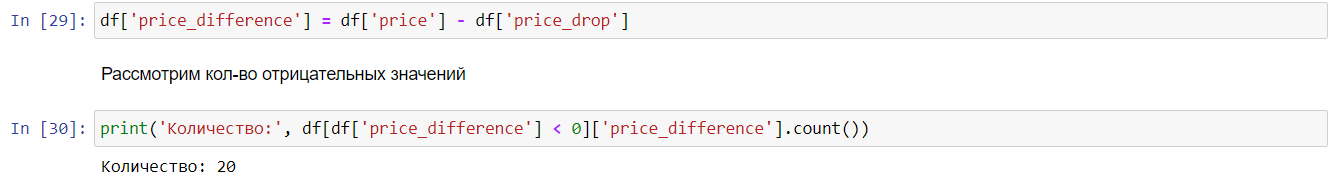
1. **Агрегирование данных**

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

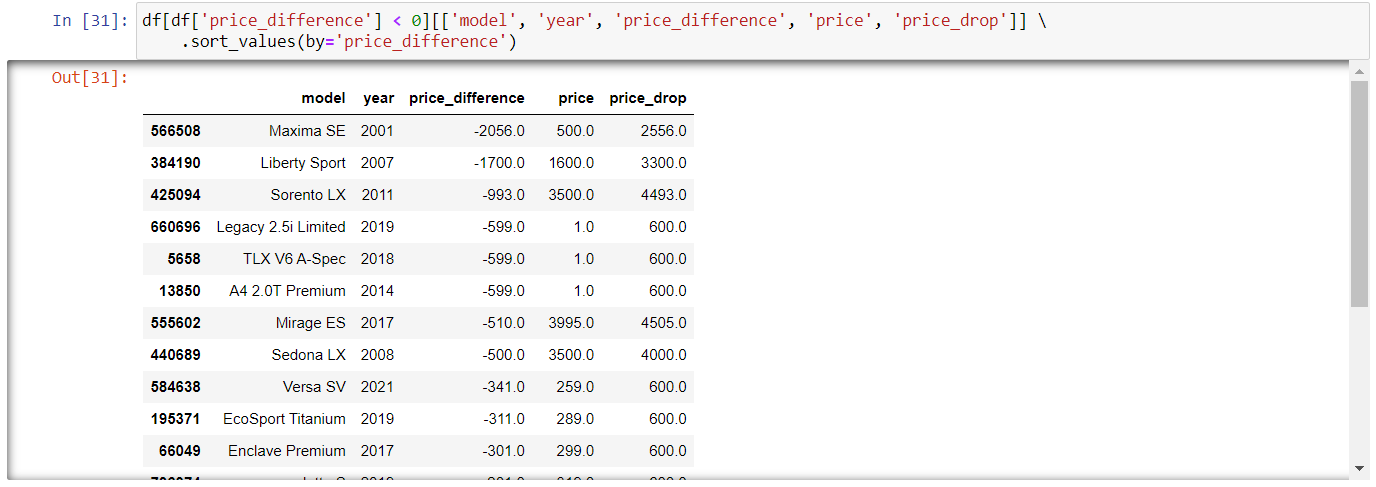
Рассмотрим разницу в продажах, следуя по нижеприведенной формуле:



Добавим новый столбец с названием «Price difference»: разница в стоимостях выставленного и продажного автомобиля.

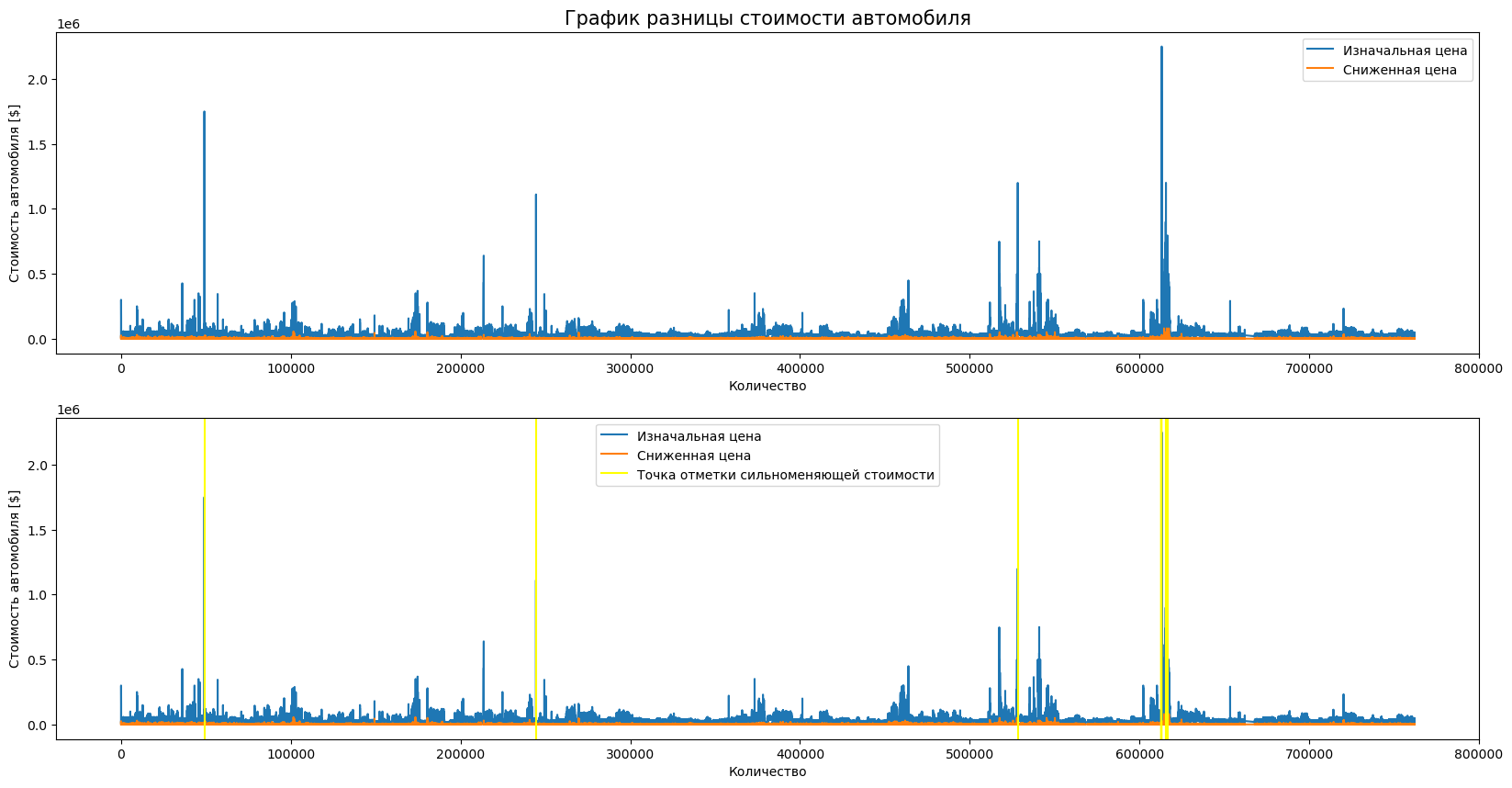


Можно сказать, что "сниженная стоимость" продавец по какой-то причине стал увеличить стоимость автомобиля, рассмотрим ниже в сортировке по возрастанию разницы стоимости.



Рассмотрим график разницы стоимости автомобиля

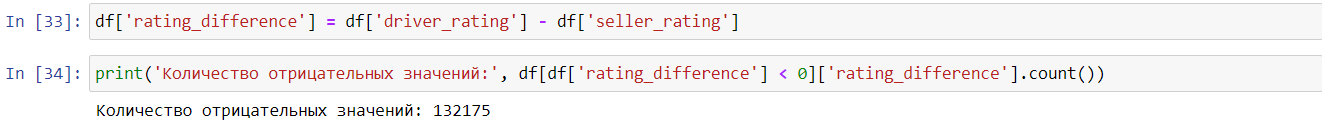




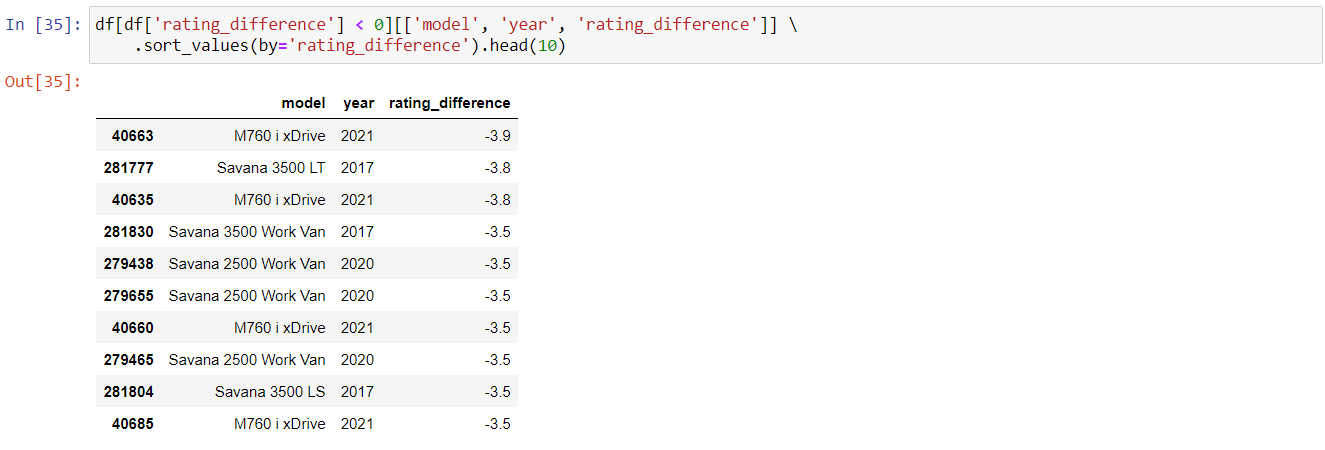
По этой графике можно заметить (желтая вертикальная линия), что в некоторых местах изначальная стоимость автомобиля была высокой, а потом сильно опустила вниз.

Агрегируем колонку – разницы в рейтингах, оставленных водителями и продавцами.

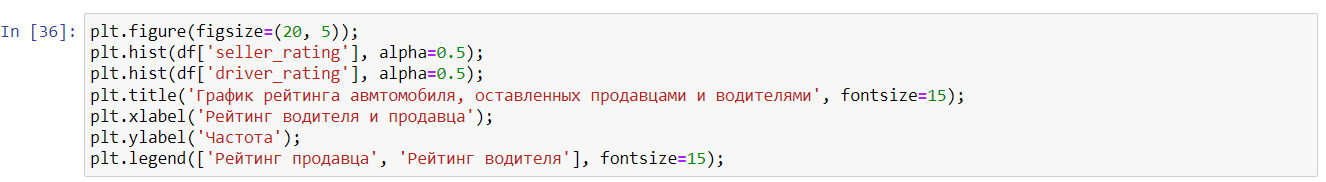


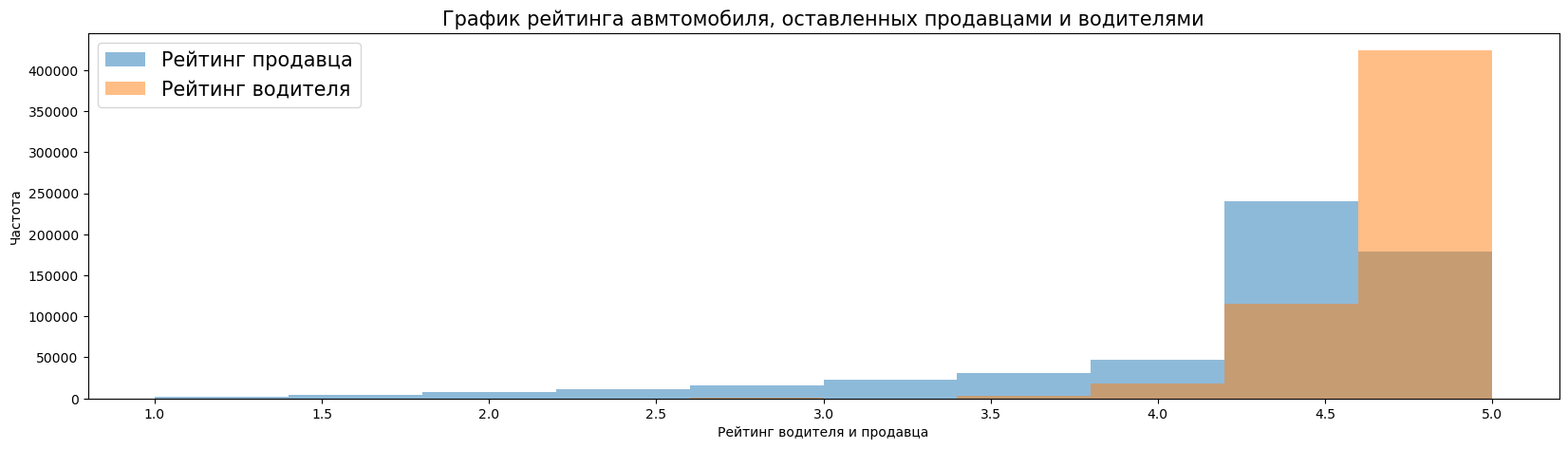


Рассмотрим только ТОП-10 данных с отрицательными значениями



Как и здесь видим, что в какой-то причине рейтинг получилось сильно разным. Давайте рассмотрим график.

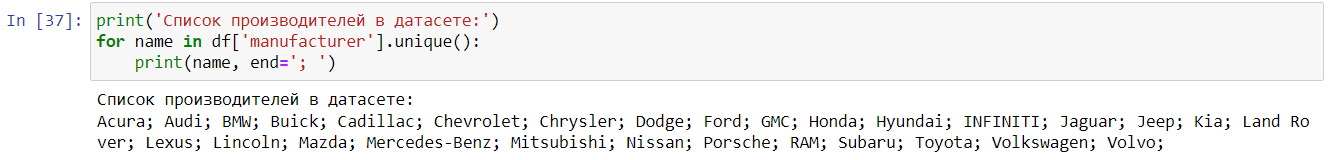


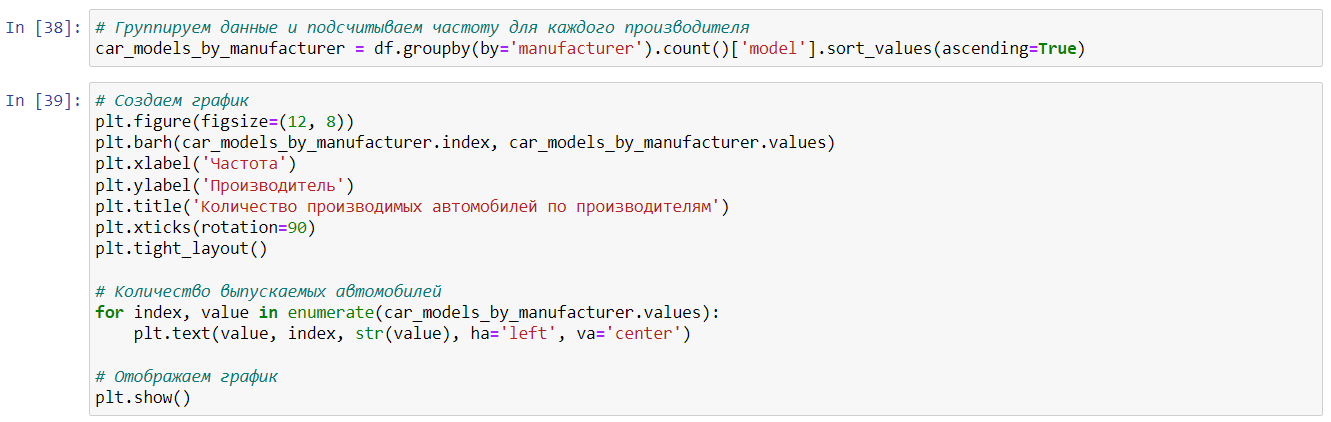


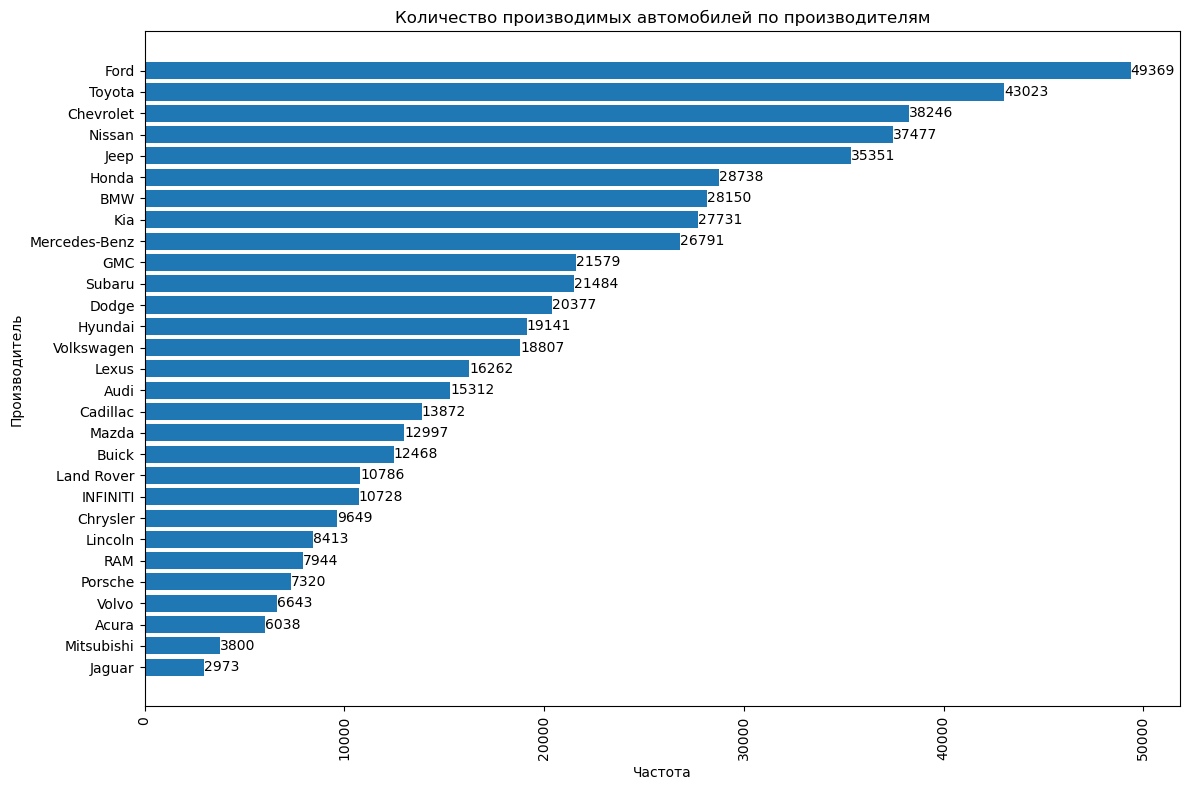
По графике можно сказать, что видны существенные разницы рейтинг между водителями и продавцами. Однако, стоит обратить внимание, что водитель чаще всего оставляют высокий рейтинг нежели продавца.

Еще агрегируем колонку – количество выпускаемых моделей производителями.

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





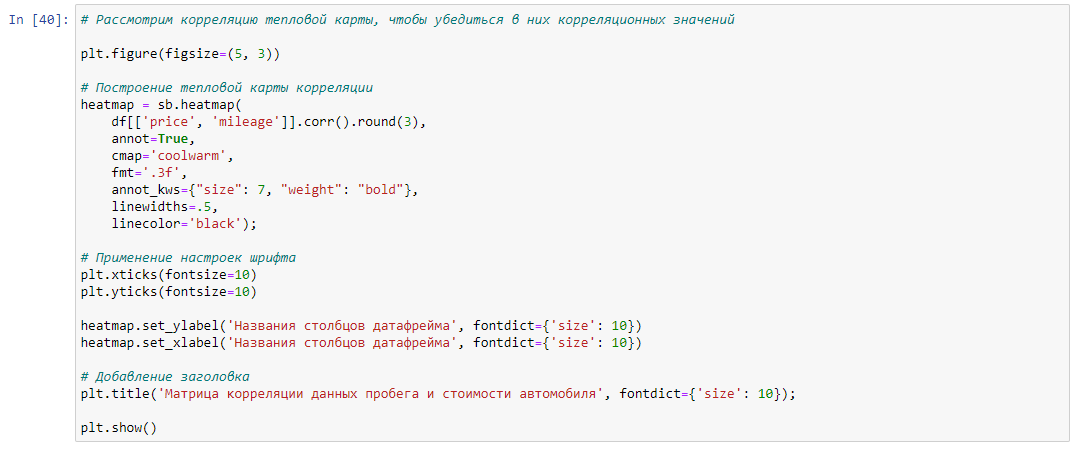


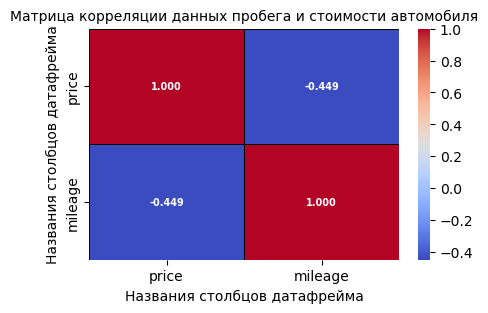
Вывод:

1. была произведена оптимизация памяти и изменение типов данных для дальнейшего анализа;
2. пропущенных значений не имеется;
3. дубликатов не имеется.
4. **Исследовательский анализ данных**

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

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

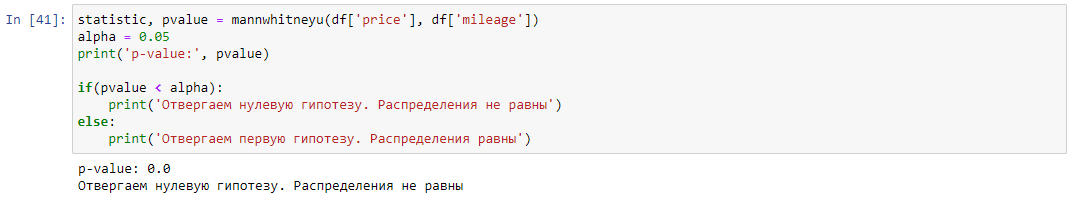




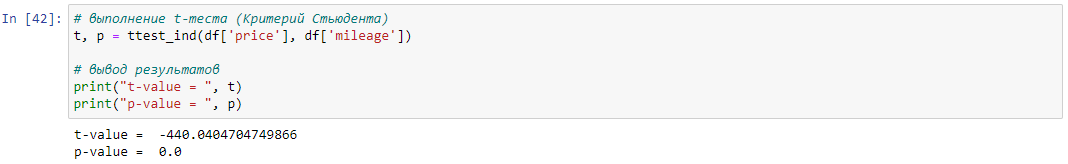
Здесь корреляция говорит, что значение получилось отрицательным, следовательно, с увеличением пробега стоимости автомобиля снижается. И в противном случае, чем дороже автомобиль, тем выше вероятность того, что у автомобиля небольшой пробег.

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.

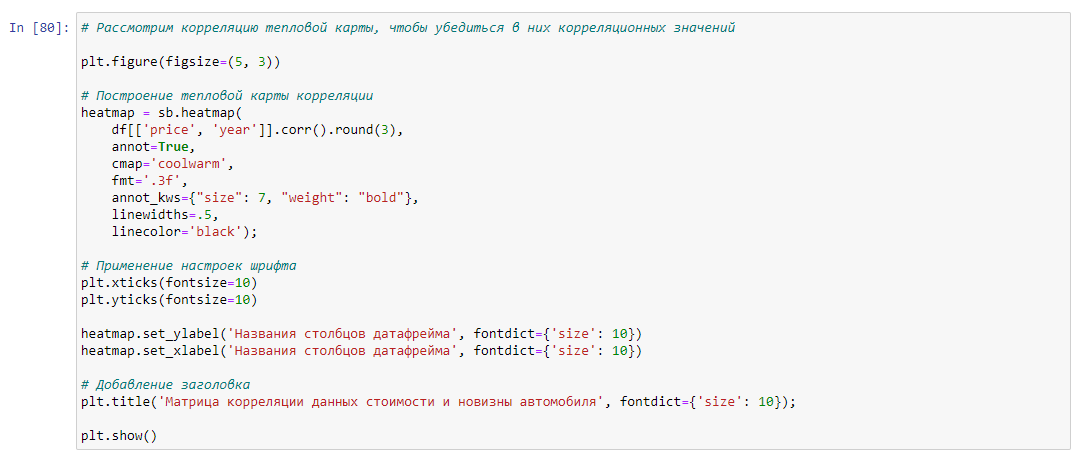


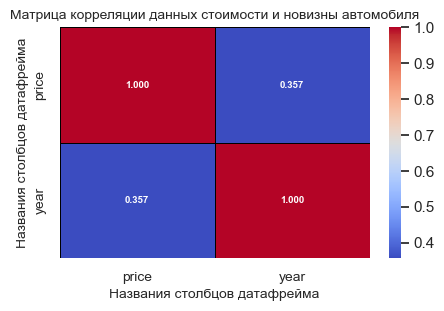
В данном случае получен очень маленький p-value, что говорит о том, что различия между данными статистически значимы. Это означает, что существует значимая разница между средними значениями двух выборок и эта разница не может быть объяснена случайными факторами.

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.

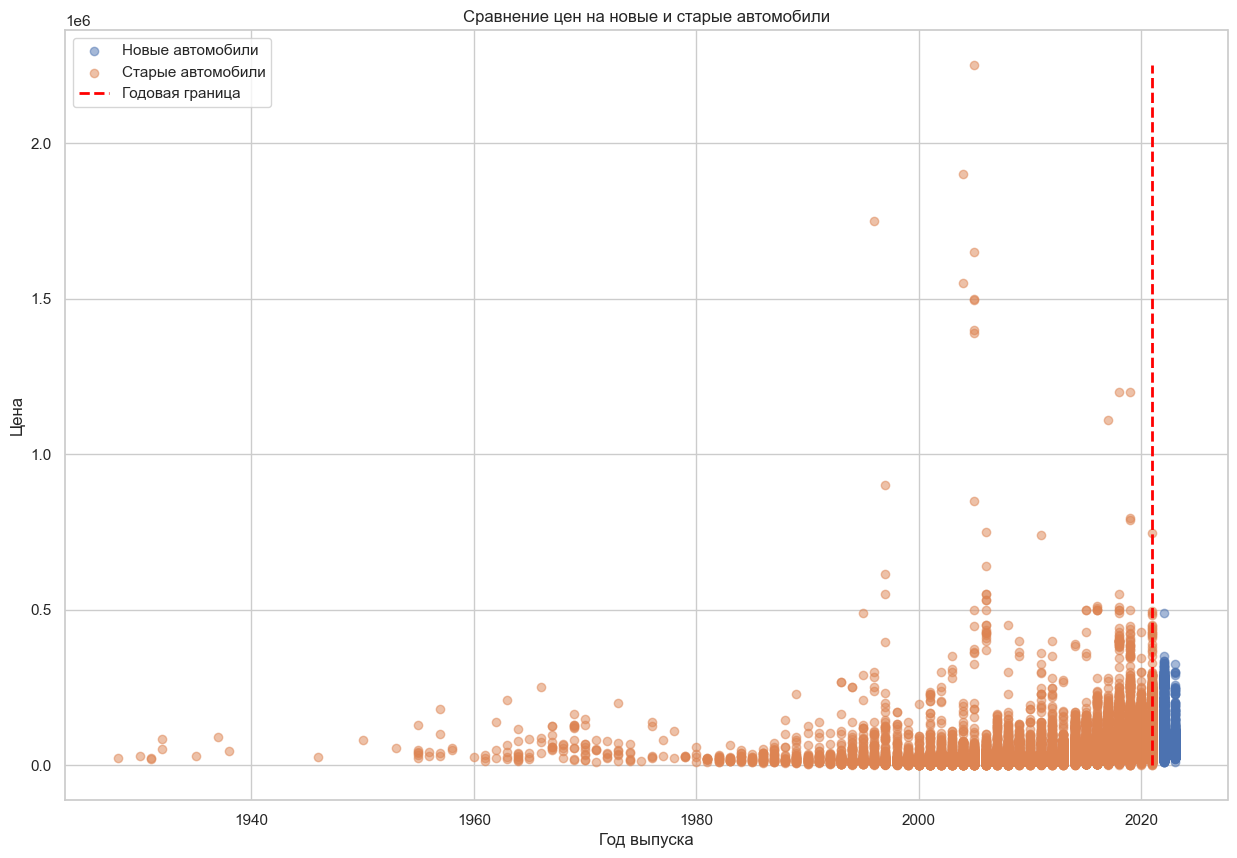
Гипотеза 2: Автомобили с определенным типом топлива имеют более высокую эффективность топливопотребления. Рассмотрим эту гипотезу и проверим.





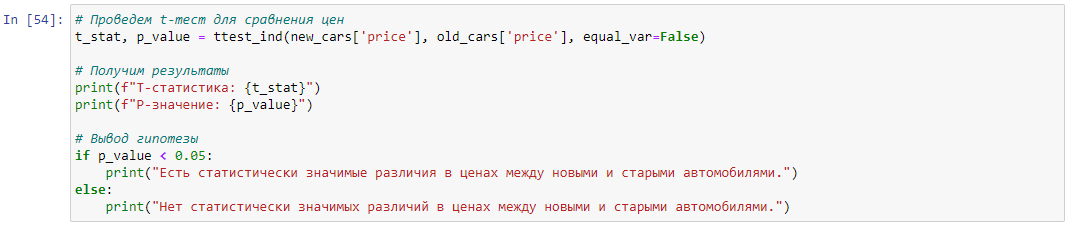
Здесь корреляция говорит, что значение получилось слабоположительной, следовательно, с увеличением года увеличивается стоимости автомобиля. И в противном случае, чем старее автомобиль, тем дешевлее. Рассмотрим график.





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

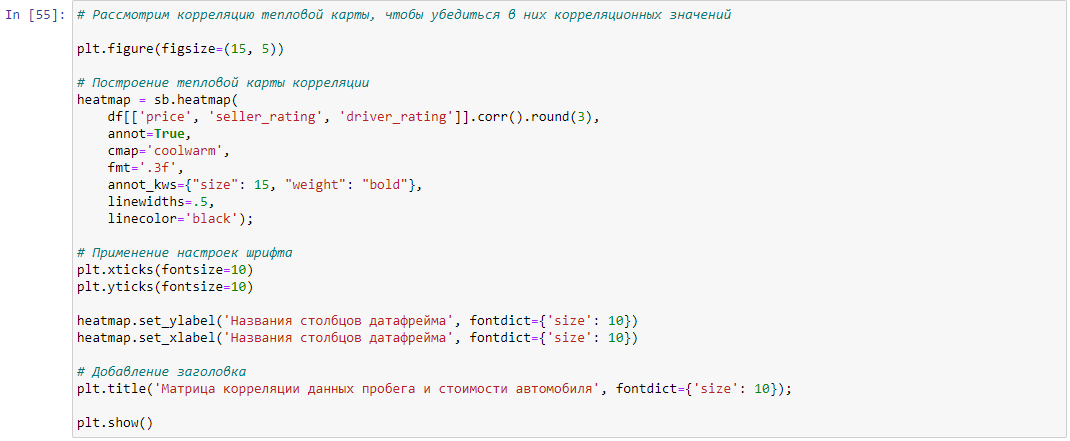


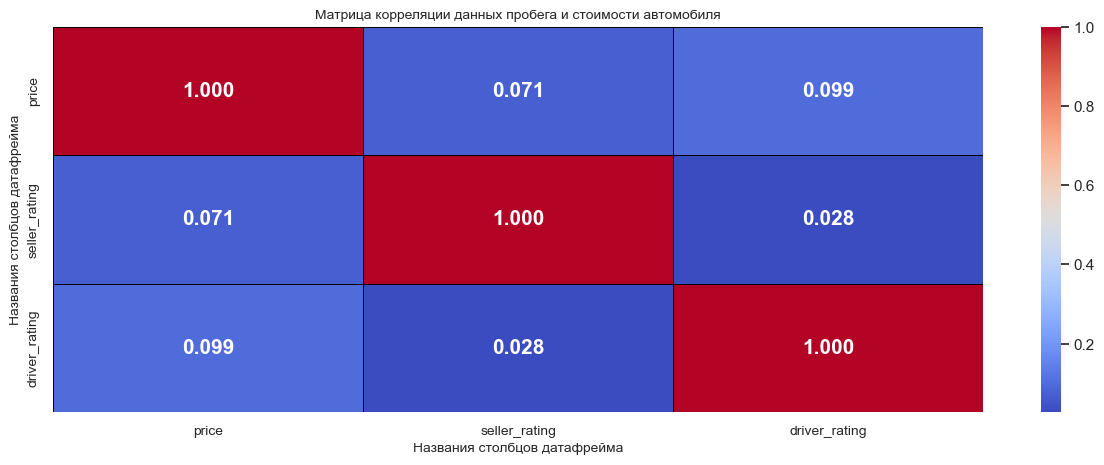
Результат t-теста позволяет отвергнуть нулевую гипотезу о том, что нет статистически значимых различий в ценах между новыми и старыми автомобилями. P-значение равно 0.0, что гораздо меньше уровня значимости 0.05.

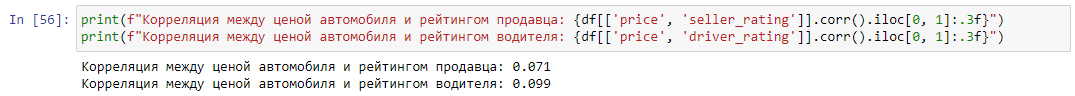
Таким образом, на основе проведенного t-теста можно сделать вывод, что существует статистически значимая разница в ценах между автомобилями нового и старого года выпуска. Поскольку T-статистика положительна, это также указывает на то, что средние цены новых автомобилей выше, чем у старых.

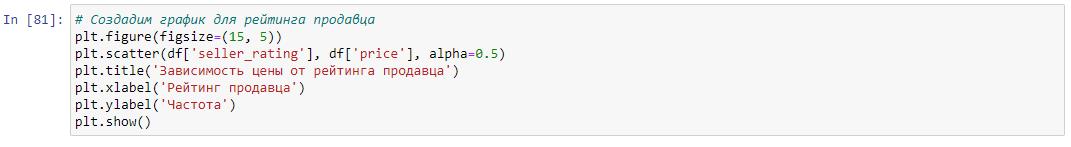
Однако, несмотря на статистическую значимость, важно помнить, что корреляция не означает причинно-следственной связи, и другие факторы также могут влиять на цены автомобилей.

Гипотеза 3: Рейтинг продавца и рейтинг водителей связаны с ценой автомобиля. Рассмотрим эту гипотезу и проверим.

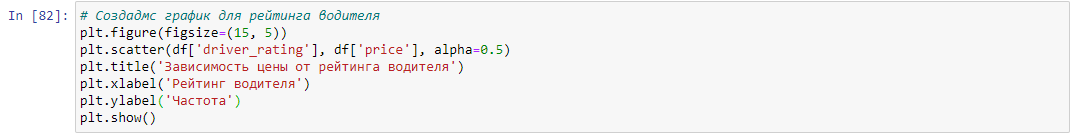


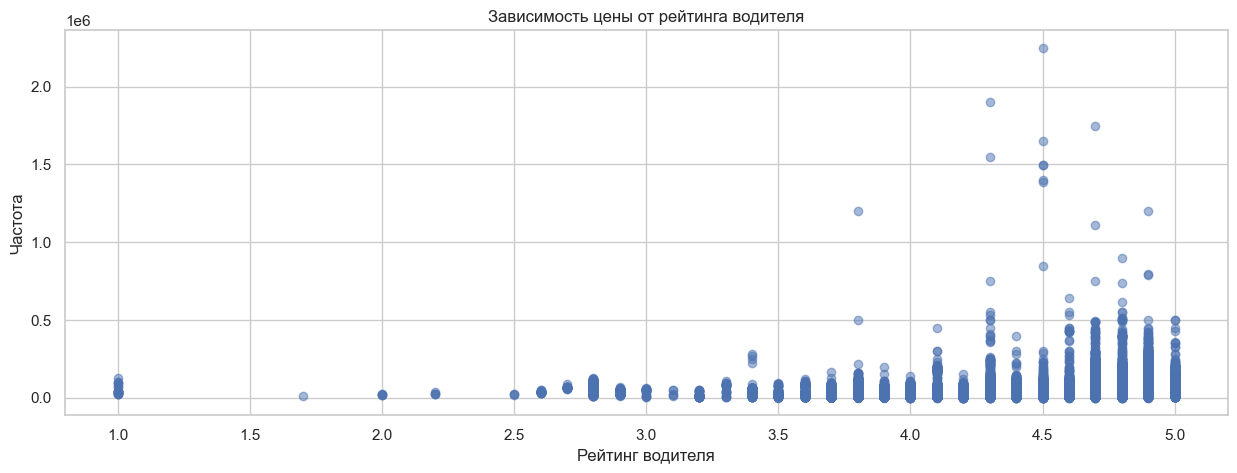








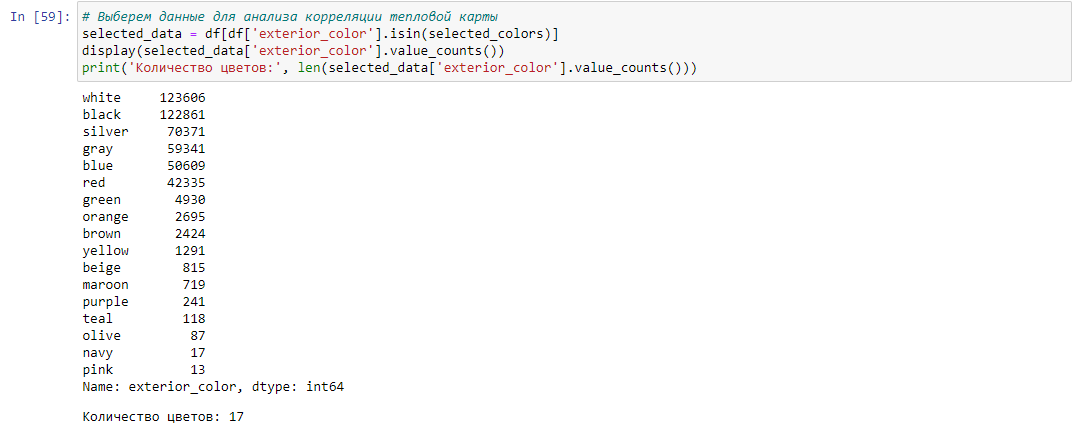


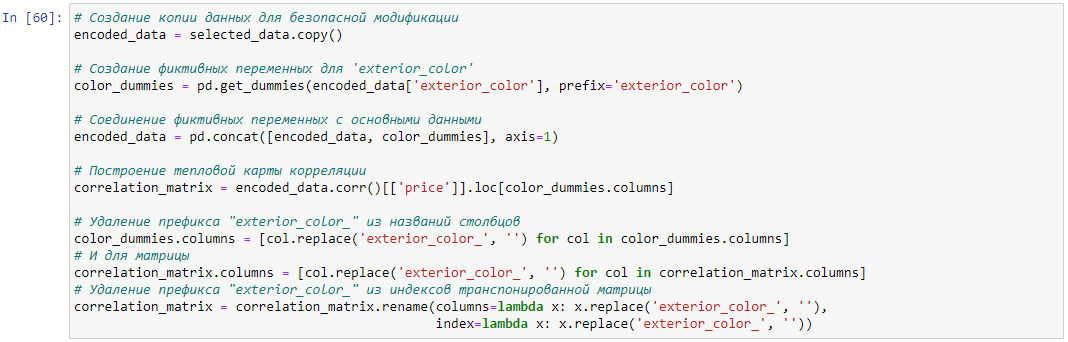


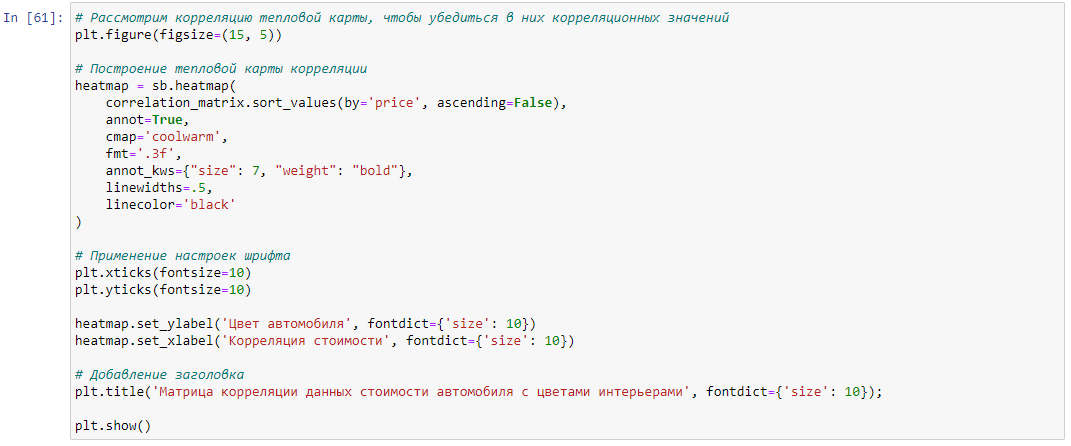
Как и видим, что корреляция получилась слабоположительной, следовательно, присутствует незначительный рост со стоимостью автомобиля и с рейтингом продавца. Например, если продавцу понравилась эта машина, значит возможно будет продавать его чуть подороже.

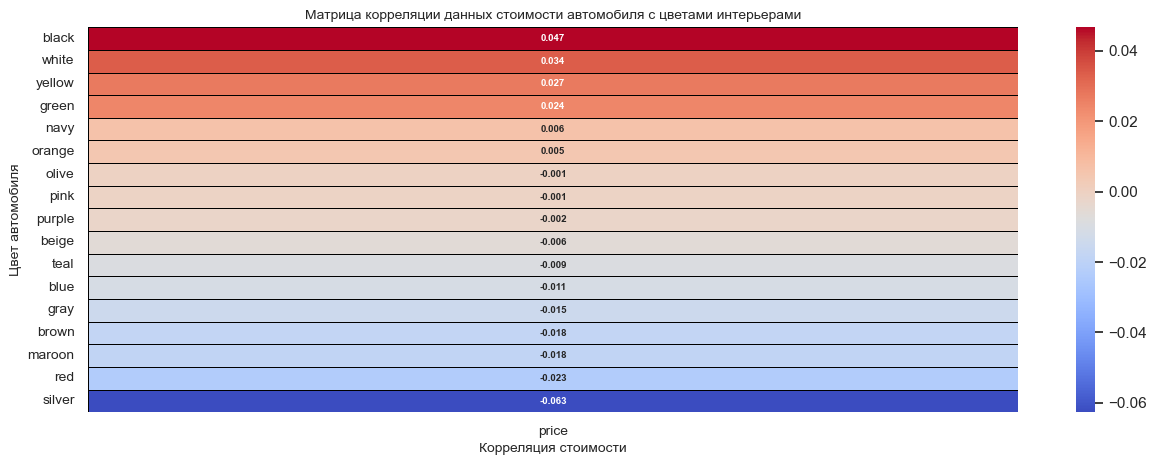
Гипотеза 4: Автомобили с определенными цветами салона стоят дороже.

Рассмотрим эту гипотезу и проверим. 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.







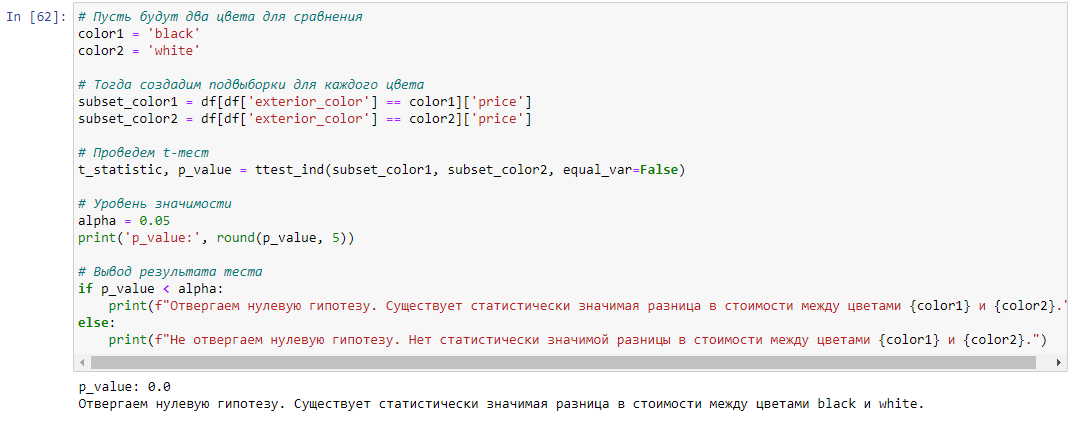


У нас получились 2 корреляции:

* 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.

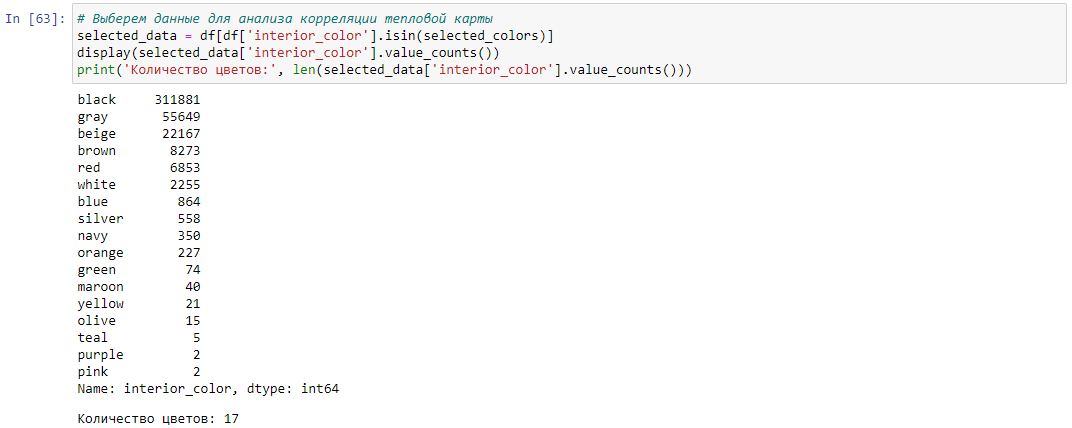
* H0 (нулевая гипотеза): Существует статистически значимая разница в стоимости автомобилей в зависимости от цвета кузова и цвета салона.
* H1 (альтернативная гипотеза): Не существует статистически значимой разницы в стоимости автомобилей в зависимости от цвета кузова и цвета салона.

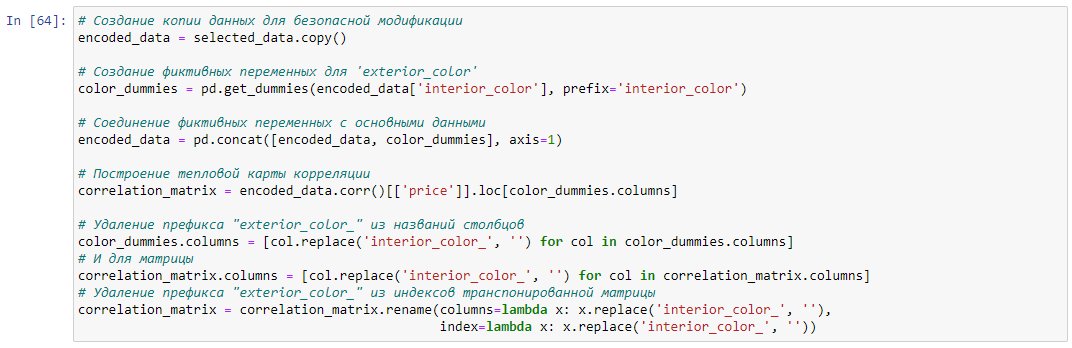


Так как P-значение меньше общепринятого уровня значимости, мы отвергаем нулевую гипотезу (H0). Это говорит о том, что существует статистически значимая разница в стоимости между автомобилями с черным и белым цветами кузова. Таким образом, можно сделать вывод о том, что цвет кузова влияет на цену автомобиля.

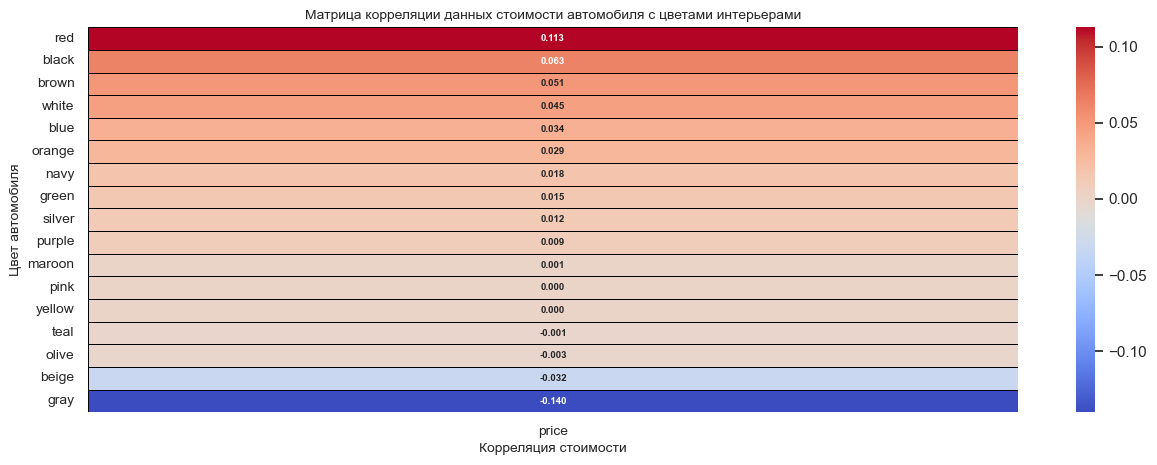
Гипотеза 5: Автомобили с определенными цветами кузова стоят дороже.

Рассмотрим эту гипотезу и проверим.







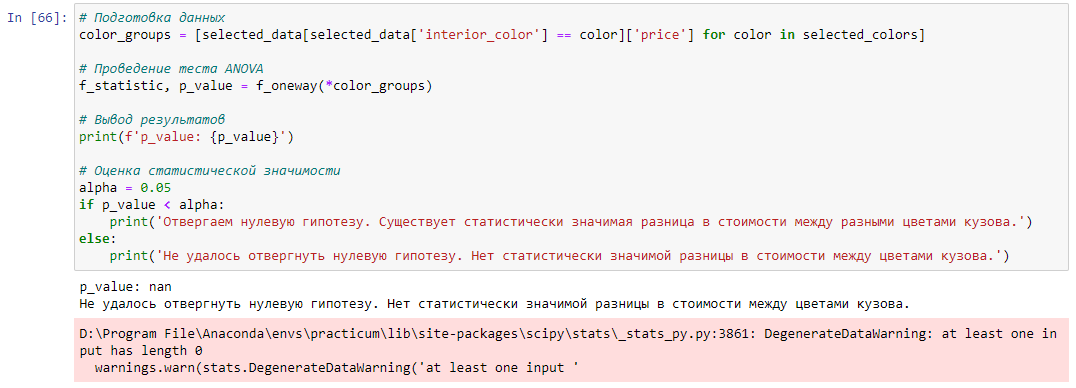


Исходя из результатов анализа, можно сделать следующие выводы:

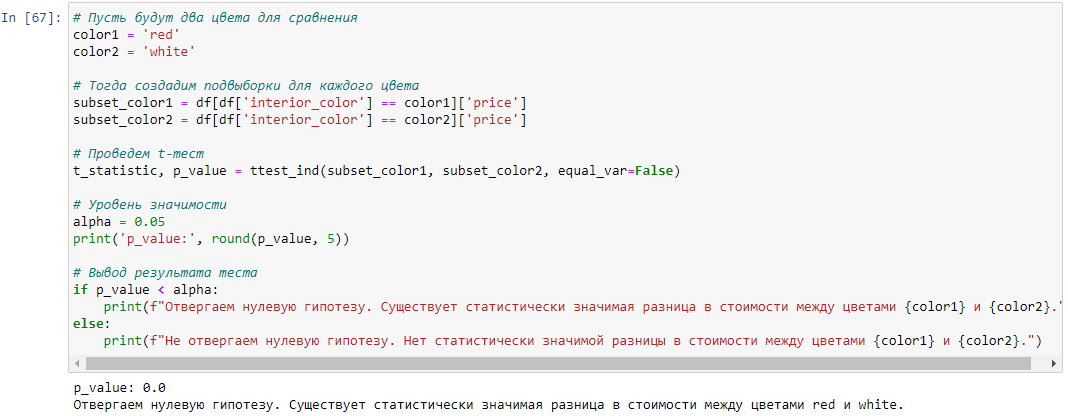
* Цвет салона влияет на стоимость автомобиля: Наблюдаются статистически значимые различия в стоимости автомобилей в зависимости от цвета салона. Например, автомобили с серым цветом салона имеют существенно более низкую стоимость, чем с автомобилями с черным, белым или красным цветами салона.
* Наибольшее 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.



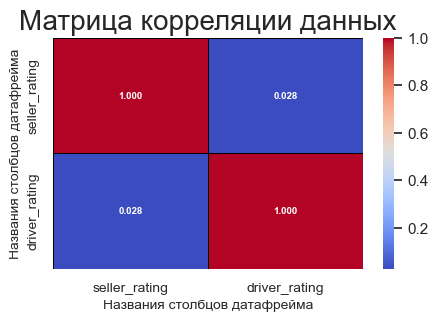
Так как P-значение меньше общепринятого уровня значимости, мы отвергаем нулевую гипотезу (H0). Странно, но можно предположить, что в данном случае у нас нет достаточных данных для выявления статистически значимых различий в средних стоимостях автомобилей для разных цветов салона. Рассмотрим еще другую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



Так как P-значение меньше общепринятого уровня значимости, мы отвергаем нулевую гипотезу (H0). Это говорит о том, что существует 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.

Гипотеза 6: Рейтинг продавца коррелирует с рейтингом автомобиля, данным водителями. Рассмотрим эту гипотезу и проверим

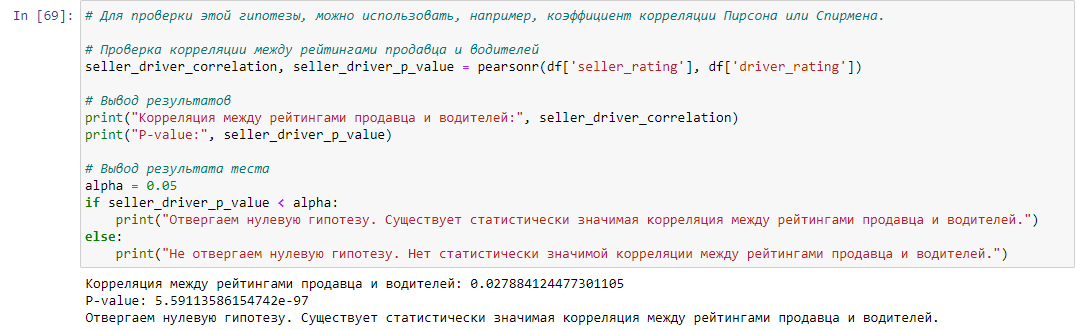




Корреляция получилась слабоположительной, рассмотрим далее.

Проверим гипотезу, выбрав тип топлива бензина и гибрид.

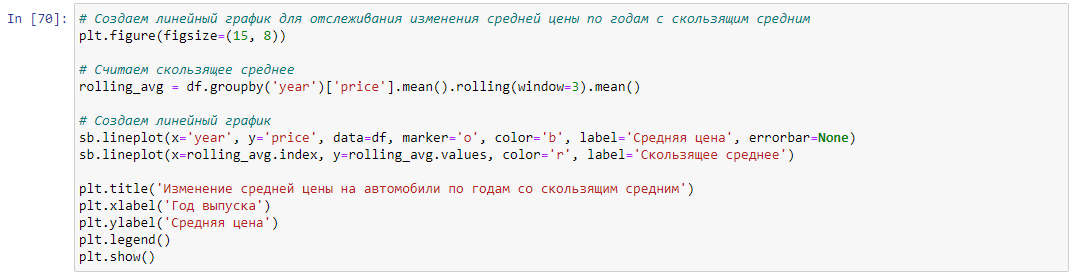
* Нулевая гипотеза (H0): Рейтинг продавца не коррелирует с рейтингом водителей
* Альтернативная гипотеза (H1): Рейтинг продавца коррелирует с рейтингом водителей.

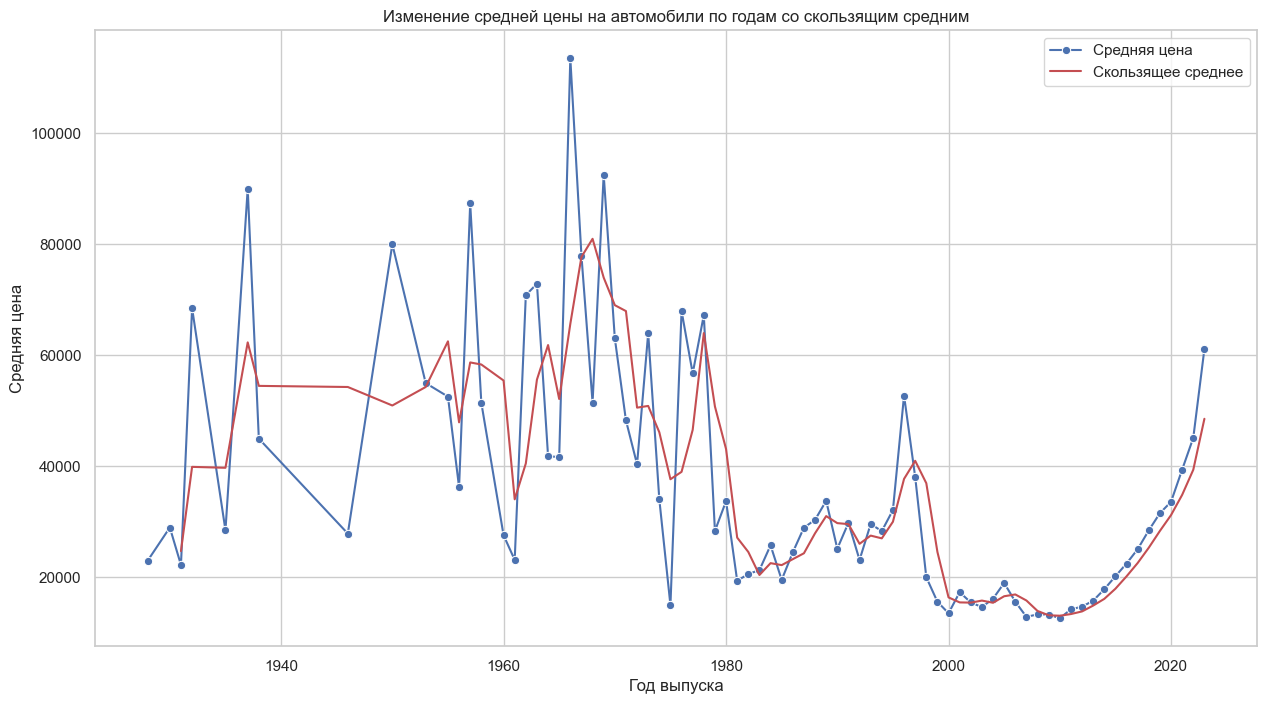


Результаты статистического теста свидетельствуют о том, что существует статистически значимая корреляция между рейтингами продавца и водителей.

Гипотеза 7: Средняя цена на автомобили меняется с течением времени.

Рассмотрим эту гипотезу и проверим.

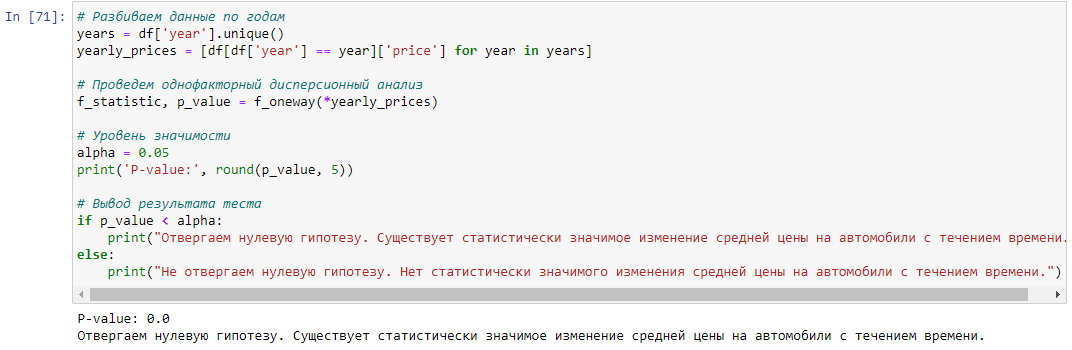




По графику видно, что раритетные автомобили дорогие, а самым дешевым оказались автомобиль с промежутком 2000 по 2010 года. Проверим гипотезу математическим способом.

Проверим гипотезу, выбрав тип топлива бензина и гибрид.

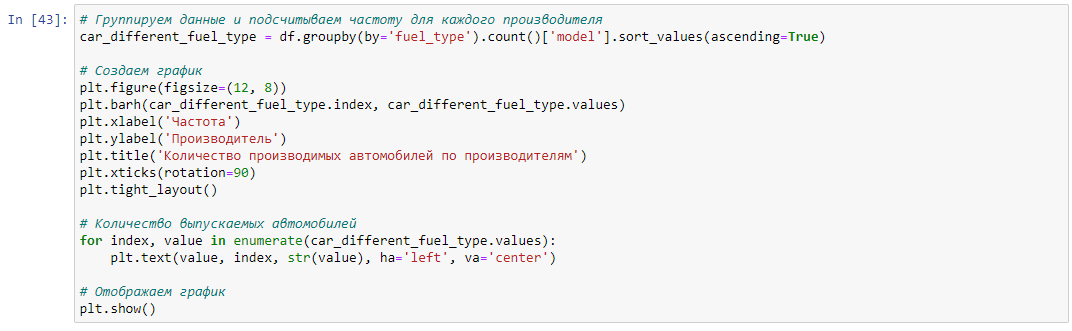
* Нулевая гипотеза (H0): Существует статистически значимое изменение средней цены на автомобили с течением времени.
* Альтернативная гипотеза (H1): Нет статистически значимого изменения средней цены на автомобили с течением времени.

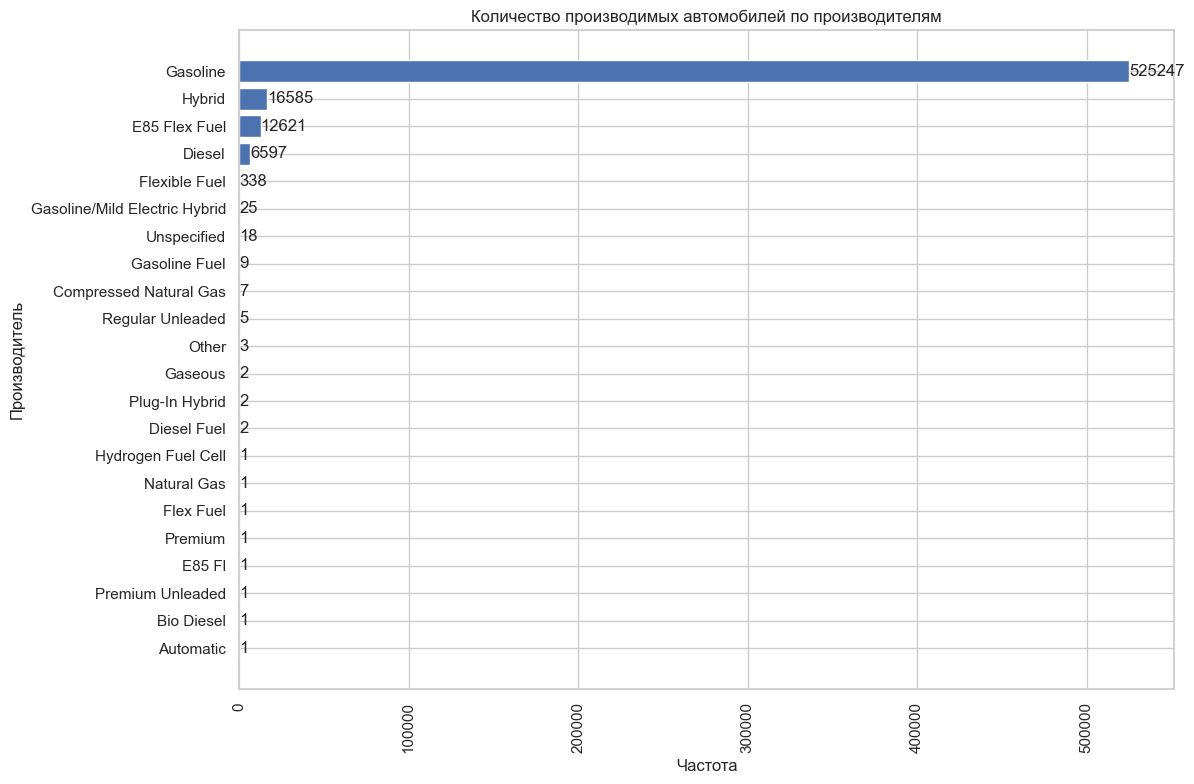


Результаты теста показали, что существует статистически значимое изменение средней цены на автомобили с течением времени.

Гипотеза 8: Автомобили с определенным типом топлива имеют более высокую эффективность топливопотребления.

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





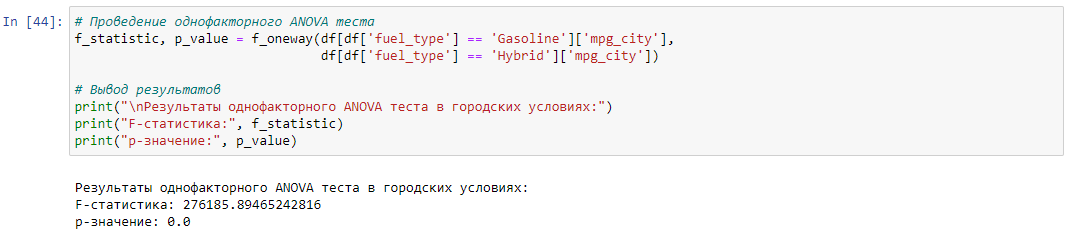
Как и видим, что в Америке популярный тип топлива - **бензин**. Второе место занимает - **гибрид**. А дальше уже смешанный тип топлива и дизель.

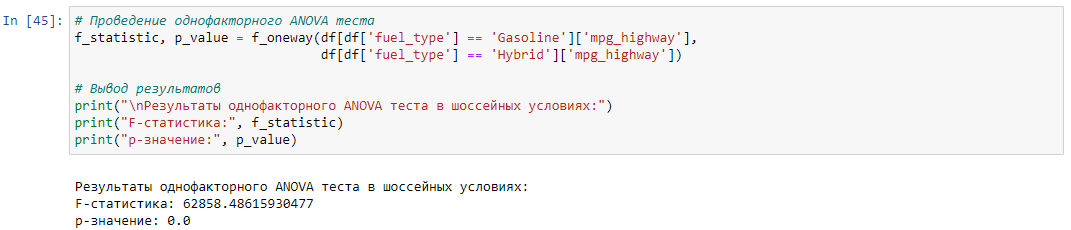
Проверим гипотезу, выбрав тип топлива бензина и гибрид.

Нулевая гипотеза (H0): Средние значения mpg в группах (типах топлива) равны, то есть нет статистически значимой разницы между типами топлива в отношении эффективности топливопотребления.

Альтернативная гипотеза (H1): Средние значения mpg в группах (типах топлива) различаются статистически значимо, что подразумевает, что существует влияние типа топлива на эффективность топливопотребления.

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

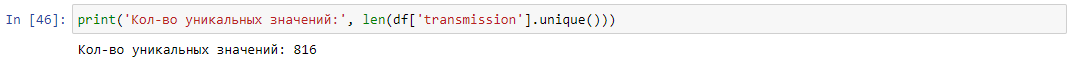




Результаты теста показали, что средние значения mpg\_city для бензиновых и гибридных автомобилей различаются статистически значимо, что подтверждает альтернативную гипотезу и \*опровергает нулевую гипотезу.

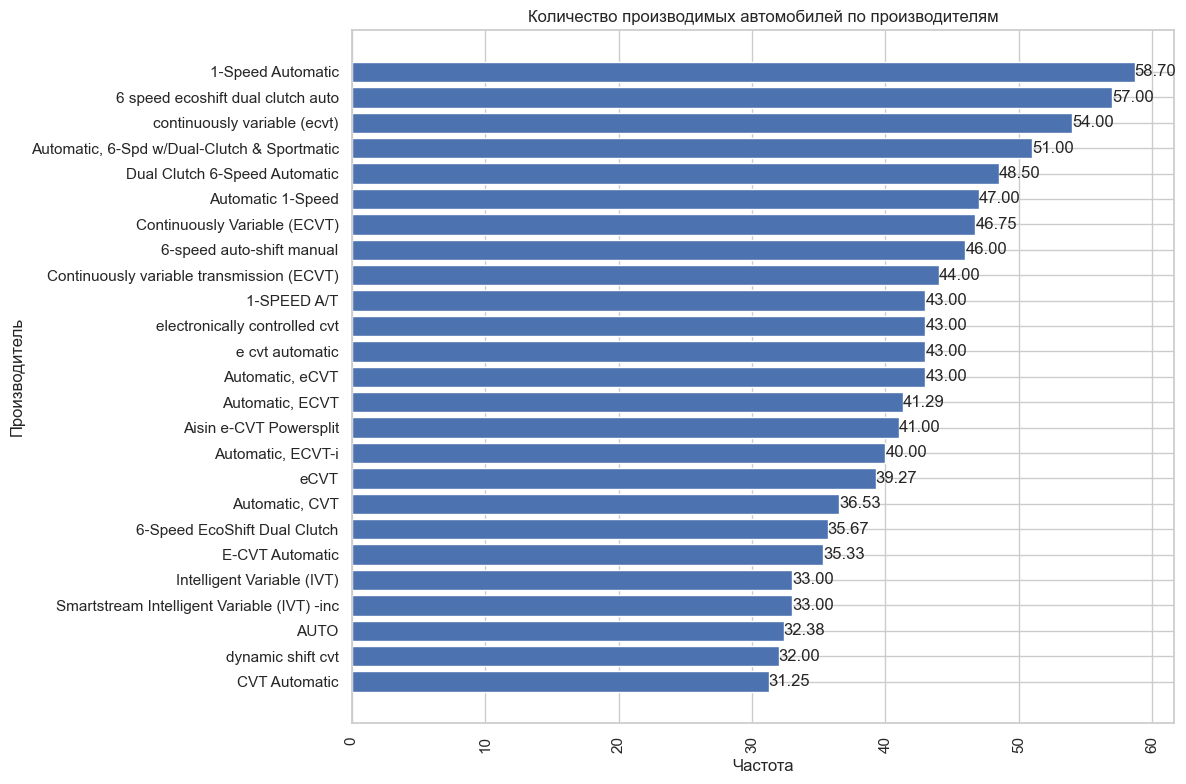
Гипотеза 9: Автомобили с определенным типом трансмиссии имеют более высокую эффективность топливопотребления.

Для начала рассмотрим уникальные значения по названию и моделей трансмиссии.



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



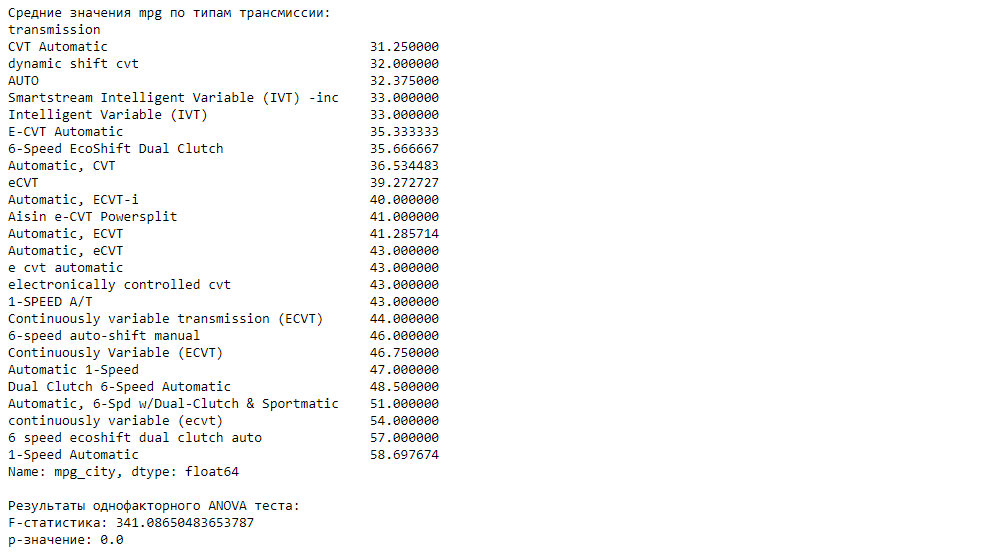


Рассмотрим нулевую и альтернативную гипотезу для анализа:

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.





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.

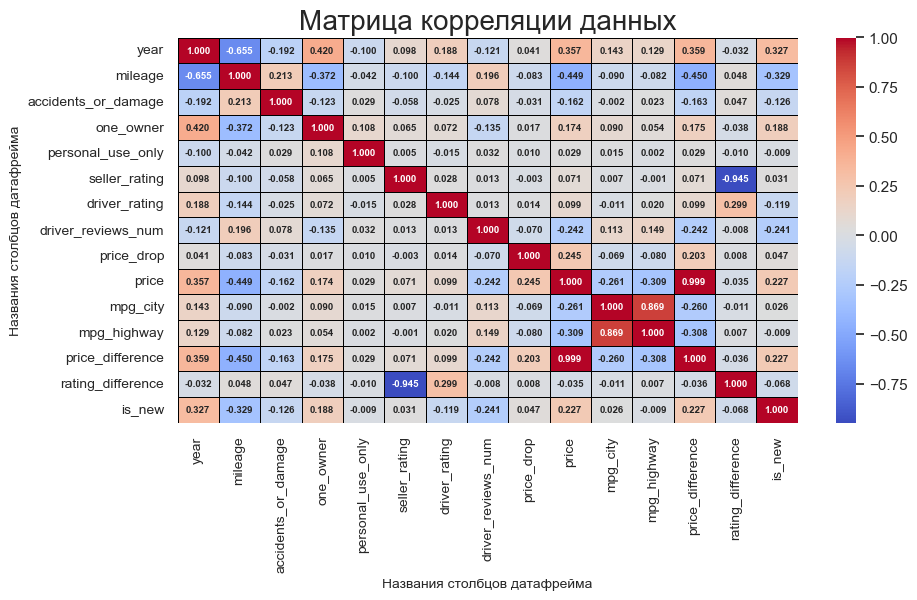
p-значение: 0.0 - нулевое p-значение означает, что вероятность получить такие или более экстремальные результаты, как в ваших данных, при условии, что нулевая гипотеза (отсутствие разницы в средних значениях) верна, крайне низка. Это говорит о том, что разница в mpg\_city между разными типами трансмиссии статистически значима.

Интерпретация:

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

**Корреляционный анализ**





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.

# 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 «Used Cars Dataset».

В ходе анализа было выявлено:

1. Самый дешевый автомобиль будет только тогда, когда у него пробег большой;
2. Есть вероятность того, что автомобиль с новым выпуском будет стоить дешевле.
3. Не исключено, что продавец и покупатель, оставленные высокий рейтинг автомобиля, будут продавать дороже.
4. Самый распространенный цвет кузова автомобиля ­– черный, белый и желтый. Менее популярный – серебристый, красный и бордовый.
5. Самый распространенный цвет салона автомобиля ­– красный, черный и коричневый. Менее популярный – серый, бежевый и оливковый.
6. Есть слабая положительная корреляция между оценками водителей и продавцами.
7. Средняя цена автомобиля меняется в течение года его выпуска.
8. Существует и самый предпочитаемый тип топлива – бензин, энергоэффективность которого превосходит другого.
9. 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 «Used Cars Dataset», в ходе которого были опровергнуты или подтверждены выдвинутые гипотезы.

Для выполнения данной работы было изучено новое программное обеспечение: Pandas, MatPlotLib, Seaborn, Numpy, Scipy на базе языка программирования Python. Также мы пользовались Jupyter Python, в нём мы писали код и анализировали диаграммы. В ходе работы я получил новые навыки в анализе данных и познакомился с вышеперечисленными программами.

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/>