|  |  |
| --- | --- |
|  | **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»**

Отчет по лабораторной работе №3

**«Подготовка обучающей и тестовой выборки, кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б   
Папин А.В.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы: 3](#_Toc1)

[2. Описание задание 3](#_Toc2)

[3. Основные характеристики датасета 3](#_Toc3)

[4. Листинг 5](#_Toc4)

[4.1. Изучение данных 5](#_Toc5)

[4.2. Преобразование данных 6](#_Toc6)

[4.3. Описательная статистика 7](#_Toc7)

[4.4. Предобработка данных 9](#_Toc8)

[4.4.1. Пропущенные значения 9](#_Toc9)

[4.4.2. Дубликаты 10](#_Toc10)

[4.4.3. Удаление выбросов 11](#_Toc11)

[4.4.4. Преобразование в численный тип 12](#_Toc12)

[4.4.5. Преобразование цветов автомобилей по ключевому названию 12](#_Toc13)

[4.4.6. Добавление новых фич для машинного обучения 14](#_Toc14)

[4.4.6.1. Разница в продажах 14](#_Toc15)

[4.4.6.2. Разница в рейтингах, оставленных водителями и продавцами 15](#_Toc16)

[4.5. Отсев до определенного кол-во уникальных значений 16](#_Toc17)

[4.6. Машинное обучение 18](#_Toc18)

[4.6.1. Деление на обучающей и валидационной выборки 19](#_Toc19)

[4.7. Кодирование признаков - прямое кодирование (One-Hot Encoding) 19](#_Toc20)

[4.8. Обучение модели 20](#_Toc21)

[4.8.1. KNeighborsRegressor 20](#_Toc22)

[4.9. Итог 21](#_Toc23)

[4.9.1. Анализ моделей 21](#_Toc24)

# **Цель лабораторной работы:**

Изучение способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

# **Описание задание**

1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите модель ближайших соседей для произвольно заданного гиперпараметра K. Оцените качество модели с помощью подходящих для задачи метрик.
5. Произведите подбор гиперпараметра K с использованием GridSearchCV и RandomizedSearchCV и кросс-валидации, оцените качество оптимальной модели. Используйте не менее двух стратегий кросс-валидации.
6. Сравните метрики качества исходной и оптимальной моделей.

# **Основные характеристики датасета**

Название датасета: Used Cars Dataset (Датасет поддержанных (б\у) автомобилей)

Ссылка: https://www.kaggle.com/datasets/andreinovikov/used-cars-dataset

**О датасетах**

Этот набор данных содержит данные о 762 091 подержанном автомобиле, собранном из cars.com . Данные были собраны в апреле 2023 года.

Датасет состоит из 20 столбцов и 762 091 строк, где каждая строка представляет:

manufacturer - название производителя автомобиля

model - название модели автомобиля

year - год, когда был выпущен автомобиль

mileage - миль, пройденных автомобилем с момента выпуска

engine - автомобильный двигатель

transmission - тип трансмиссии автомобиля

drivetrain - тип трансмиссии автомобиля

fuel\_type - тип топлива, которое потребляет автомобиль

mpg - количество миль, которое автомобиль может проехать, используя один галлон топлива (мили на галлон)

exterior\_color - цвет кузова автомобиля

interior\_color - цвет салона автомобиля

accidents\_or\_damage - попадал ли автомобиль в АВАРИИ

one\_owner - принадлежал ли автомобиль одному лицу

personal\_use\_only - использовался ли автомобиль только в личных целях

seller\_name - имя продавца

seller\_rating - рейтинг продавца

driver\_rating - рейтинг автомобиля, данный водителями

driver\_reviews\_num - количество отзывов об автомобилях, оставленных водителями

price\_drop - снижение цены по сравнению с начальной ценой

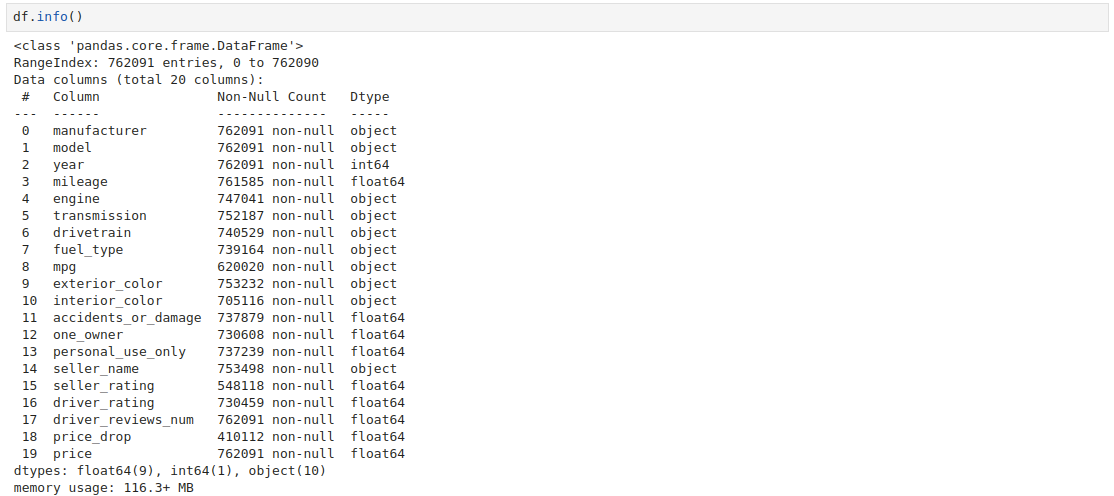
price - цена автомобиля

**Выбор признаков для машинного обучения**

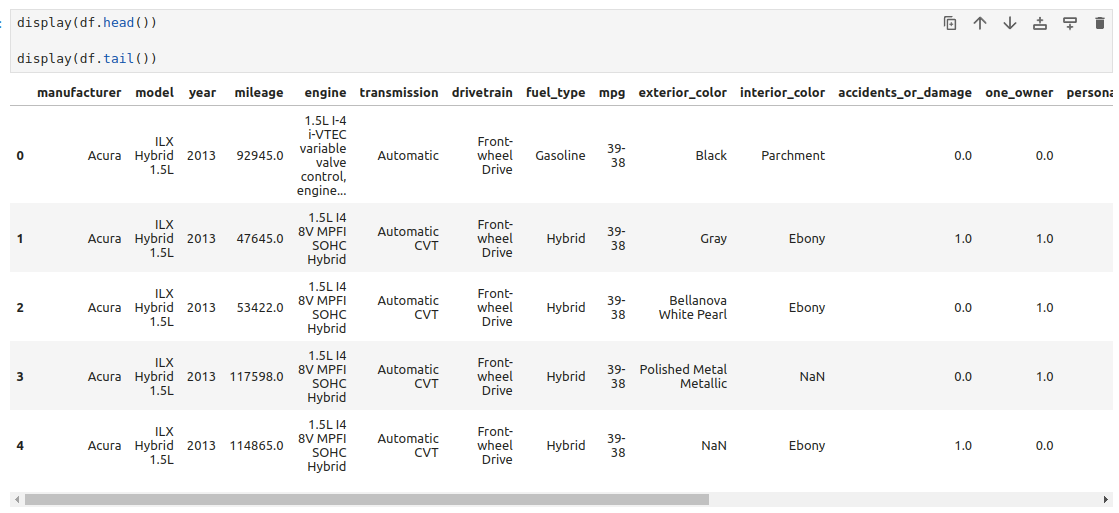
Для машинного обучения выберем целевой признак - стоимость автомобиля. Сопоставим с остальными признаками, а именно, характеристики и конфигурации автомобиля выявяляем примерную стоимость автомобиля.

# **Листинг**

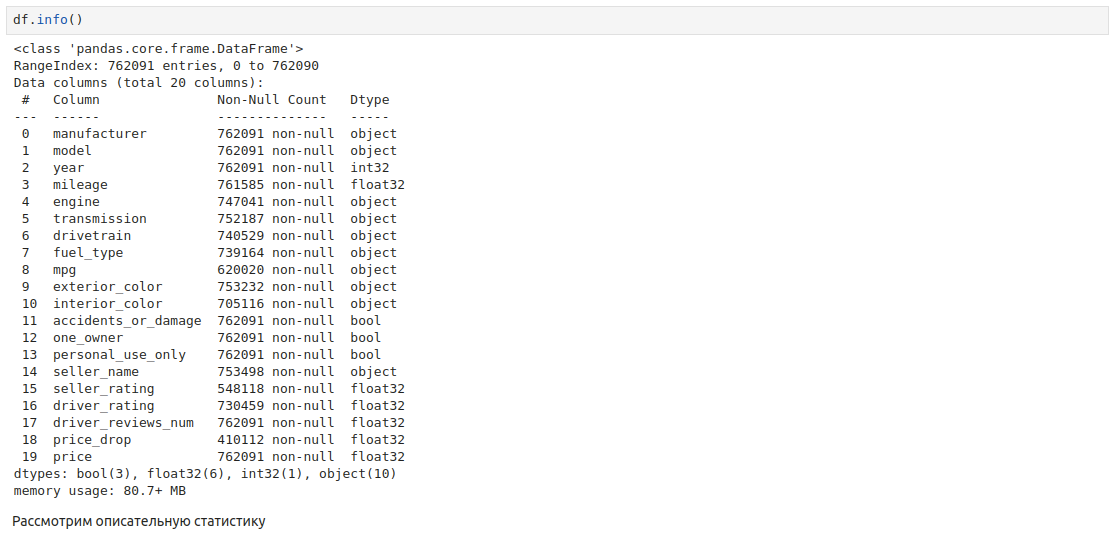
# **Изучение данных**



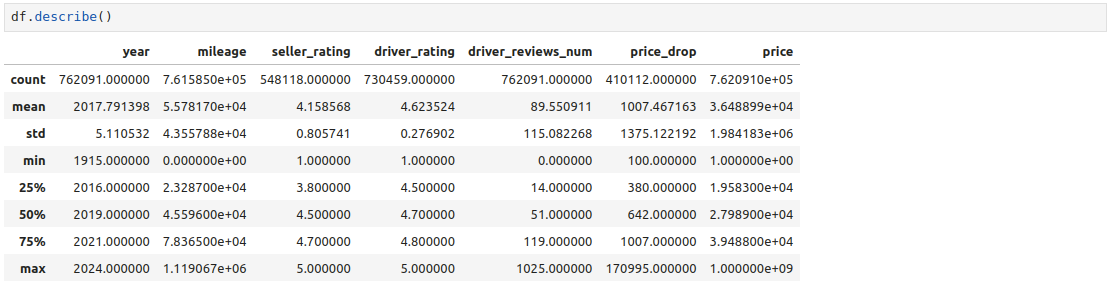
Здесь можно заметить, что в датасете содержатся единиц 762091 строк. А также имеют 3 различные типы: object, int64 и float64. В целях экономии памяти можно преобразовать в другие типы.



# **Преобразование данных**



# **Описательная статистика**



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. Нелья 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.

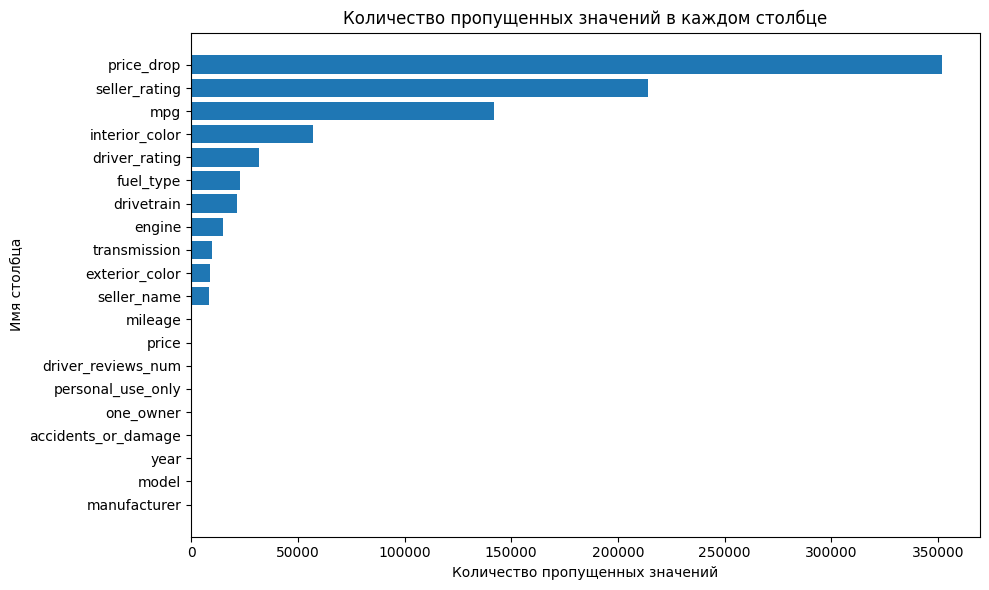
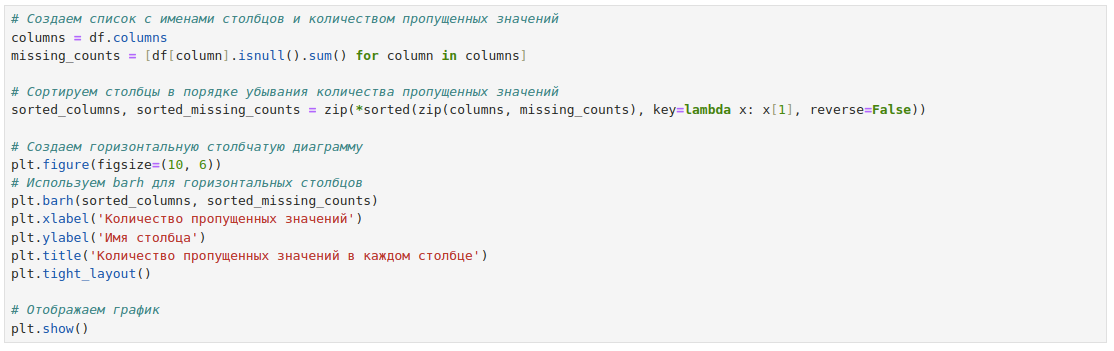
**Снижение цены по сравнению с начальной ценой**. В среднем мы можем увидеть в объявлениях, что продают автомобилей с 1007 долларов, а самой максимальной - 170995 долларов, минимальной - 100 долларов.

**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.

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.

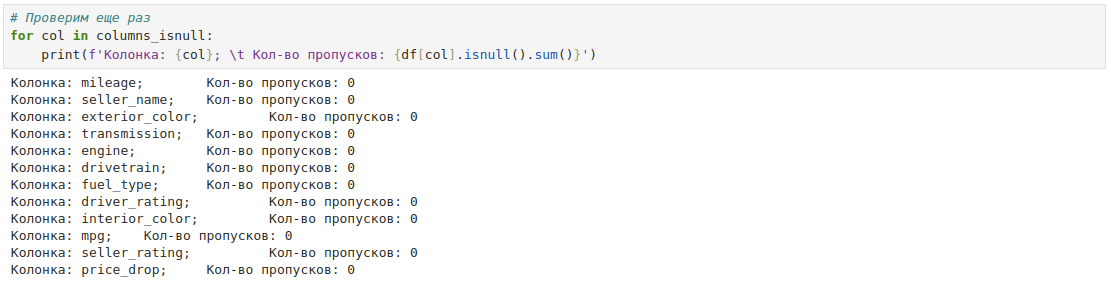
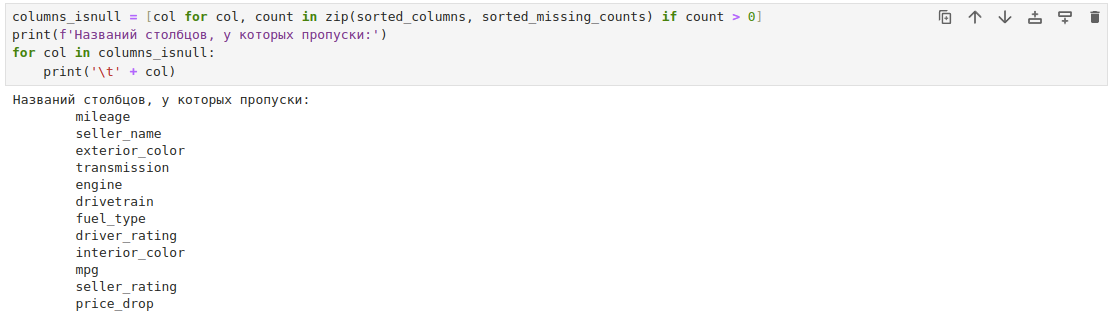
# **Предобработка данных**

# **Пропущенные значения**

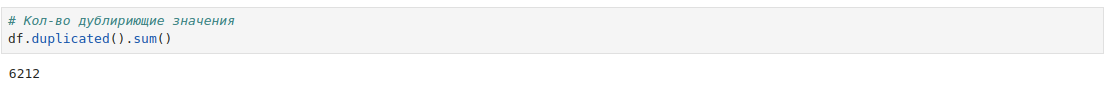


Как видим, что присутствуют огромные пропуски в столбцах: сниженная стоимость, рейтинг продаца и кол-во миль. Заполним пропуски медианными значениями только для численных типов, а остальных - устраним.

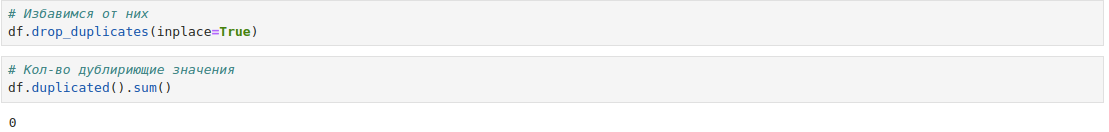
Медианными значениями заполняем, потому что они менее чувствительны к выбросам.



# **Дубликаты**

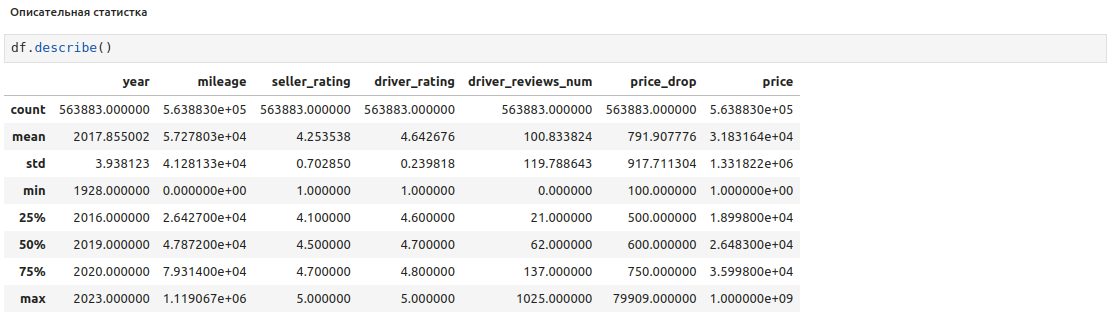


Как видим, что присутствуют очень много дубликатов. Устраним их.

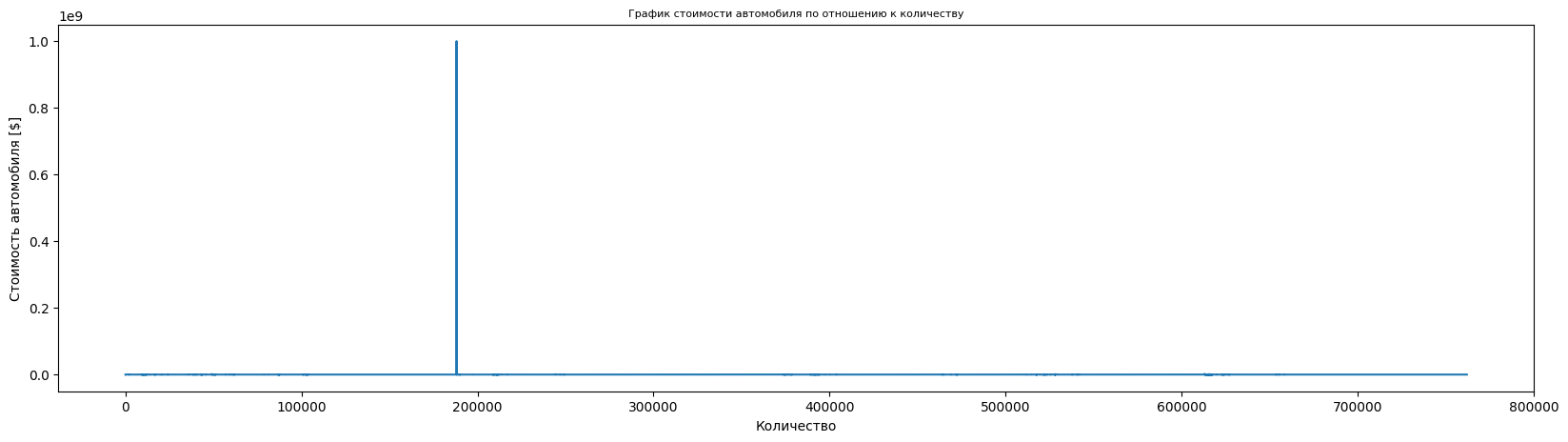


# **Удаление выбросов**

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



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



Как и видим, устраним их.

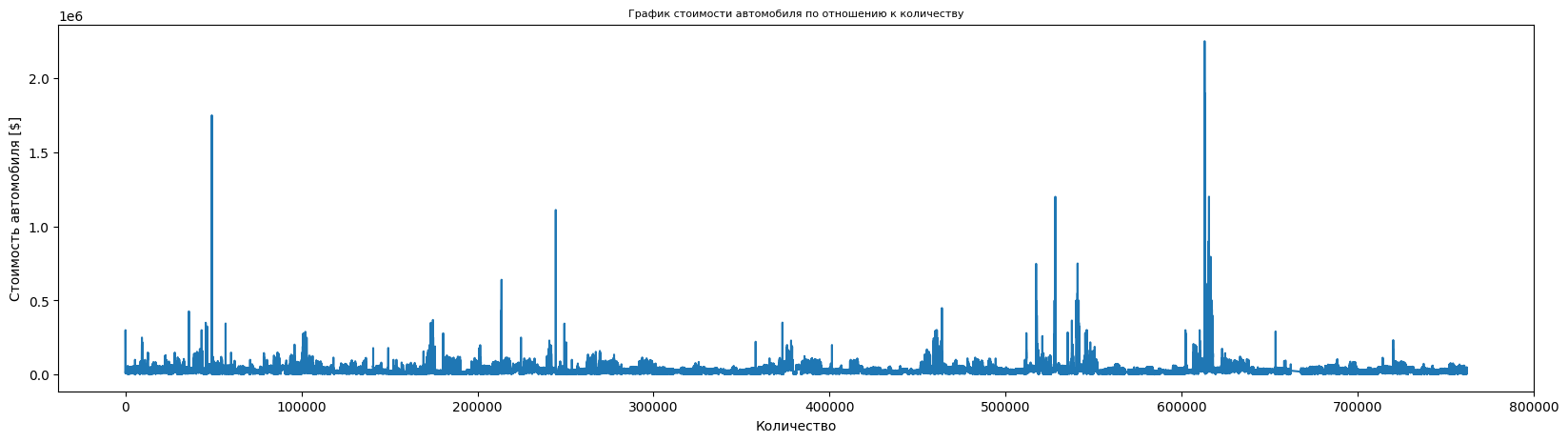
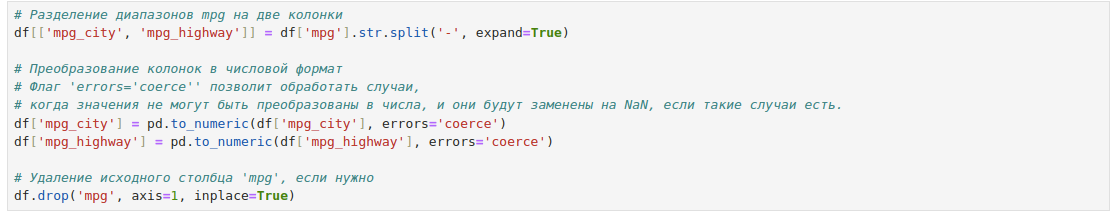
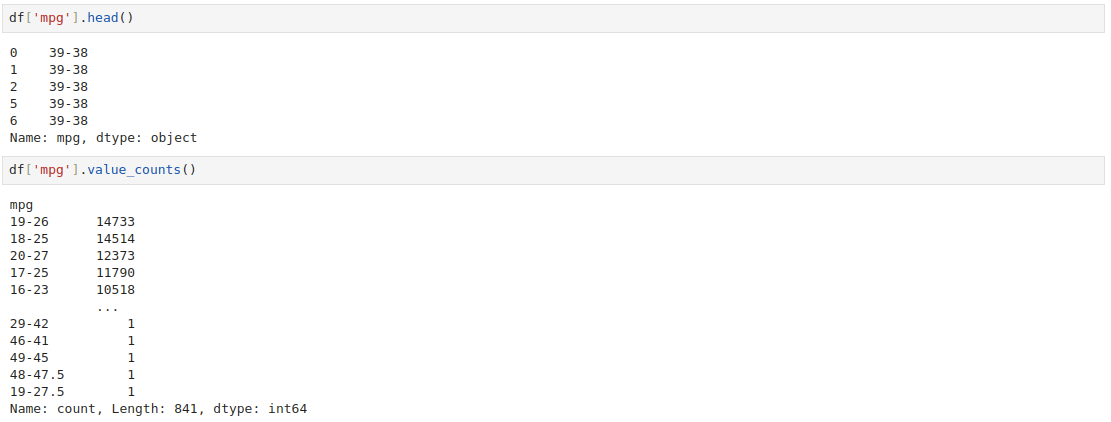


График получился более менее адекватным.

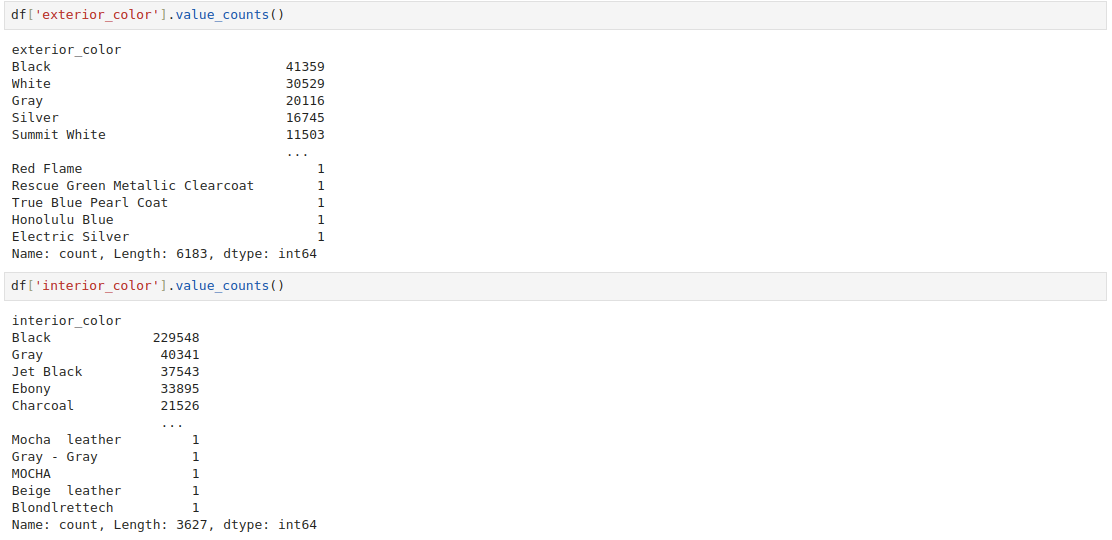
# **Преобразование в численный тип**

У нас есть колонка не численного типа, которого нужно преобразовать в другой тип.



# **Преобразование цветов автомобилей по ключевому названию**

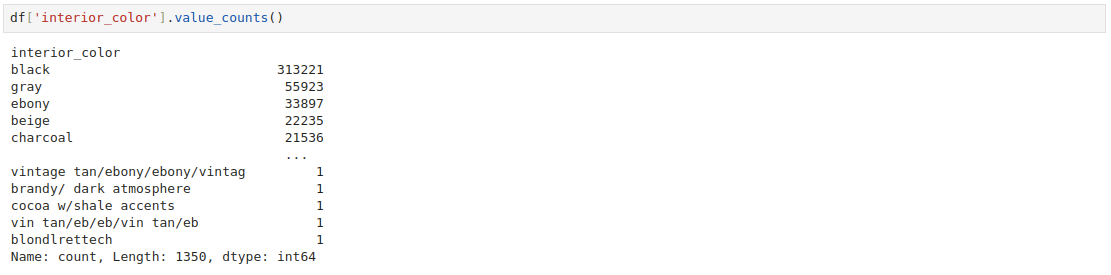
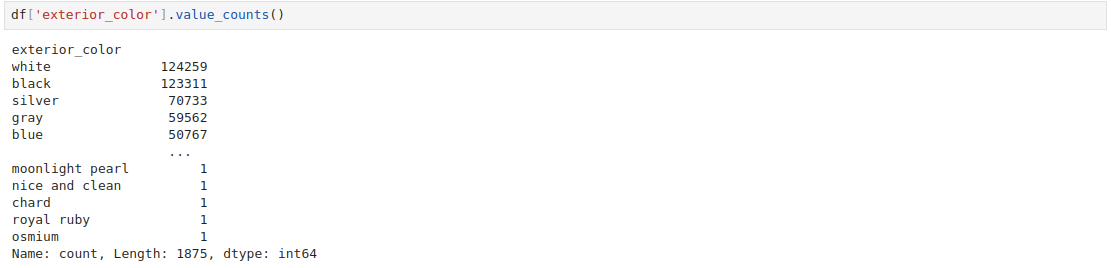
Бесспорно, что существуют много разных типов цветов автомобилей. Можно через TF-IDF пропустить, чтобы составить ключевые цвета со списков разных цветов автомобилей со всего датасета.



Как и видим, что много разных цветов и причем с разными регистрами. Выделим конкретные уникальные цвета.



Посмотрим, сколько уникальных значений получили в итоге.

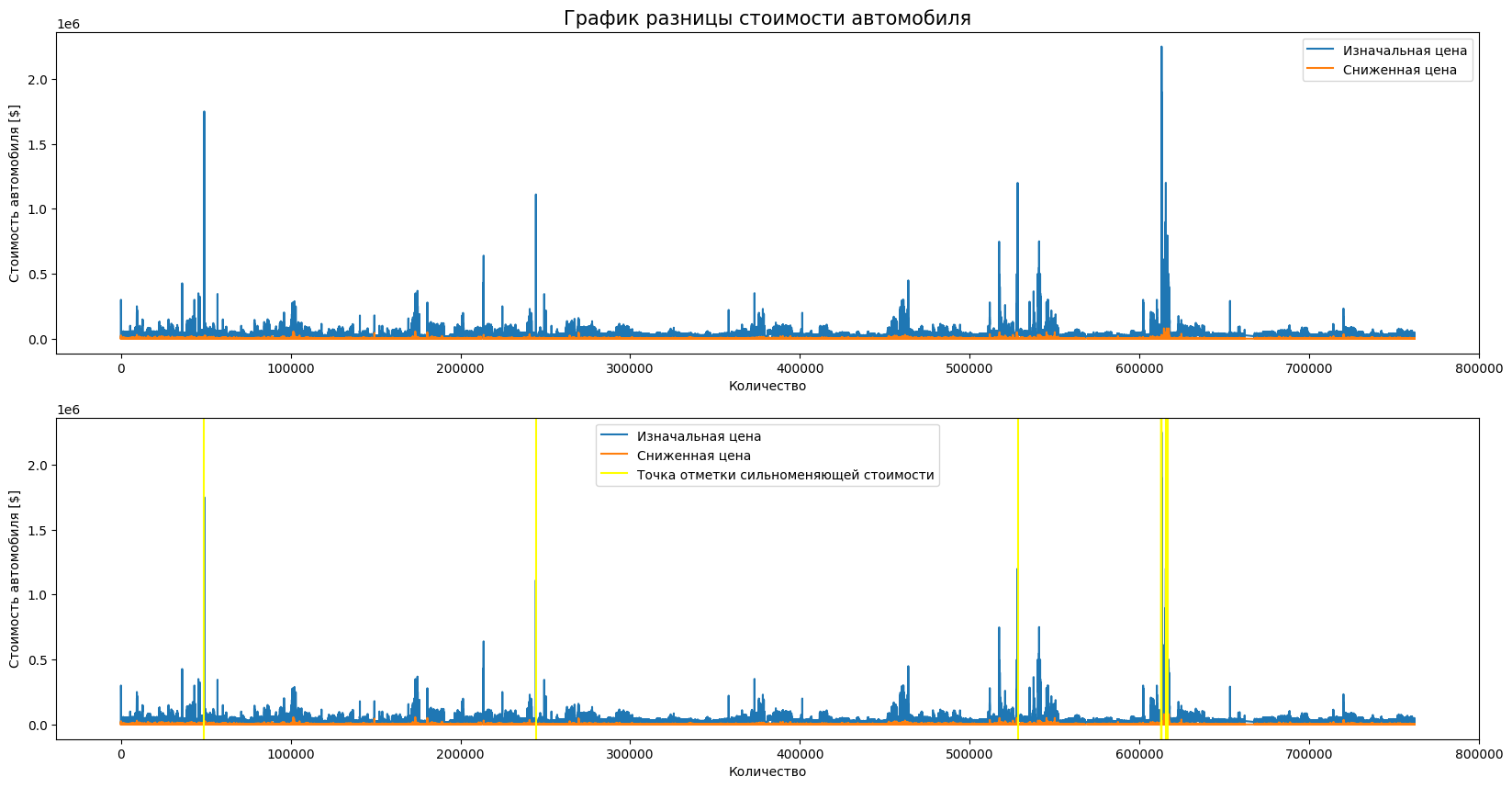
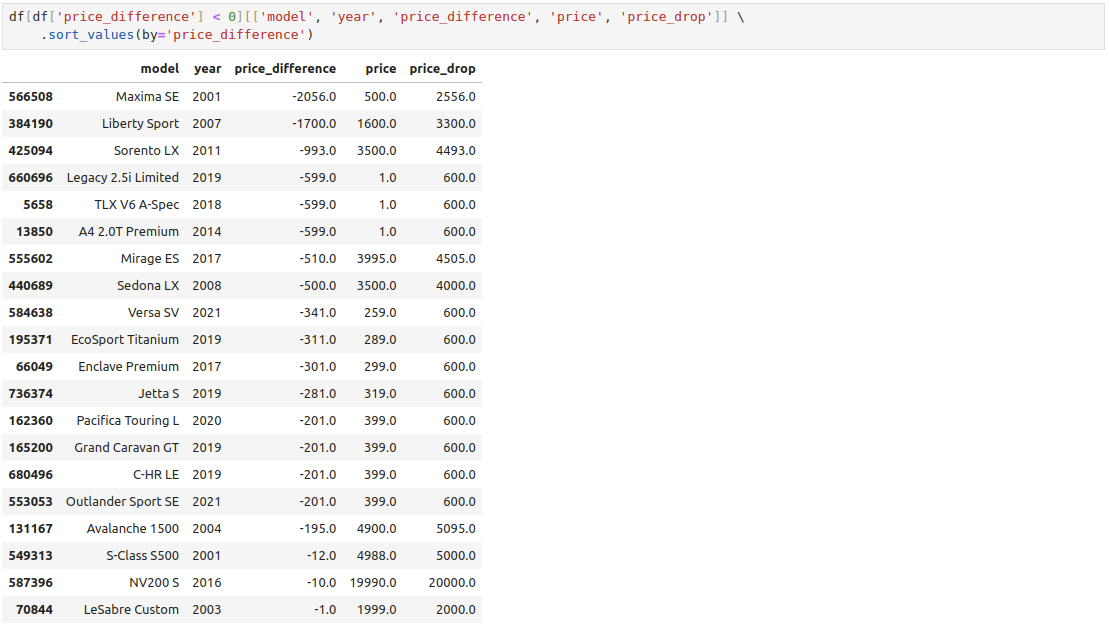
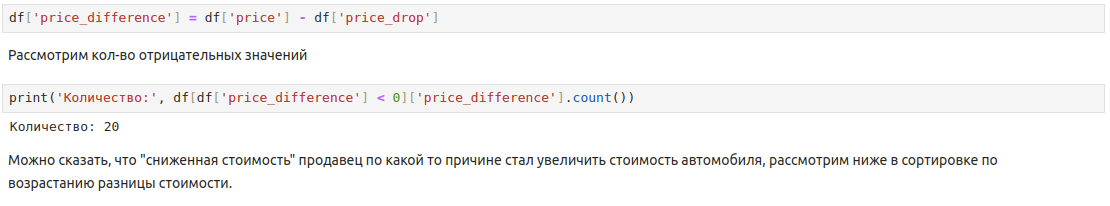


Уникальных цветов довольно много, можно отсеивать их для машинное обучение.

# **Добавление новых фич для машинного обучения**

# **Разница в продажах**

Рассмотрим разницу в продажах, следуя по формуле:



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

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

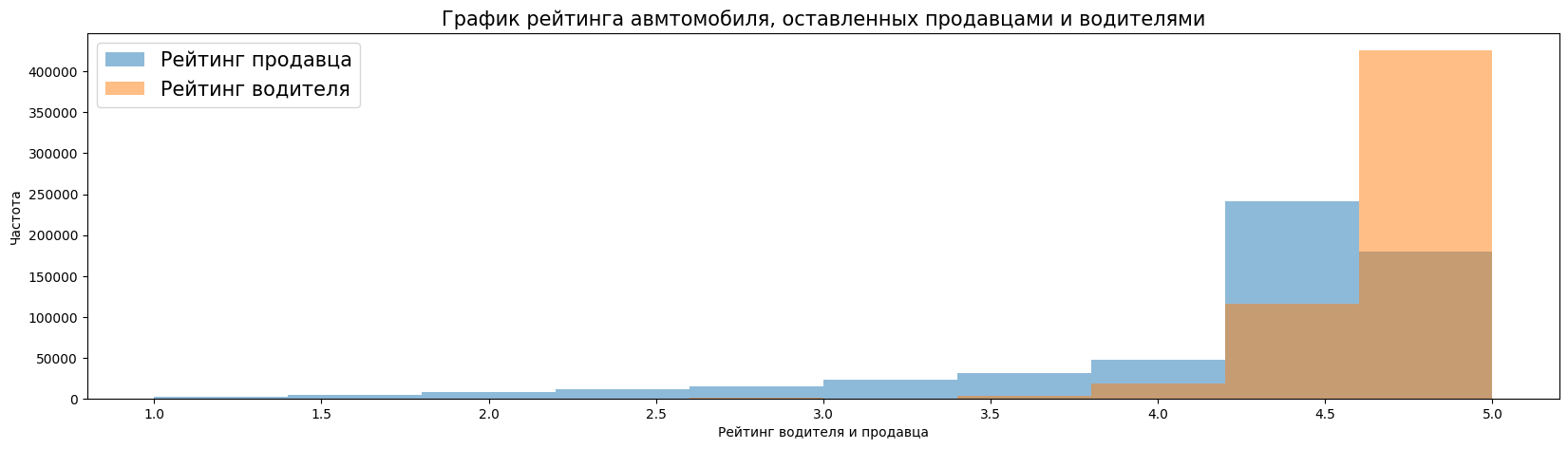
Рассмотрим разницу в рейтингах, следуя по формуле:



Рассмотрим только ТОП-10 данных с отрицательными значениями



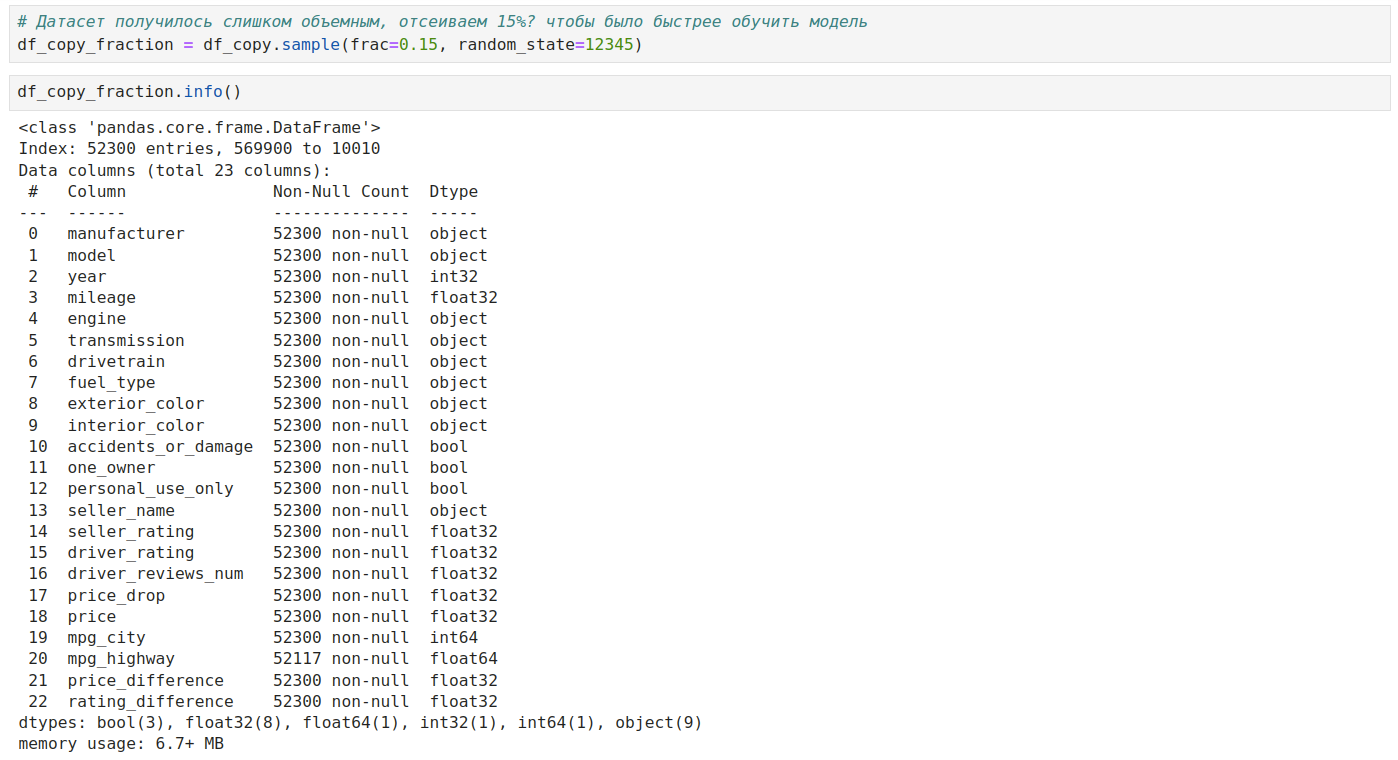
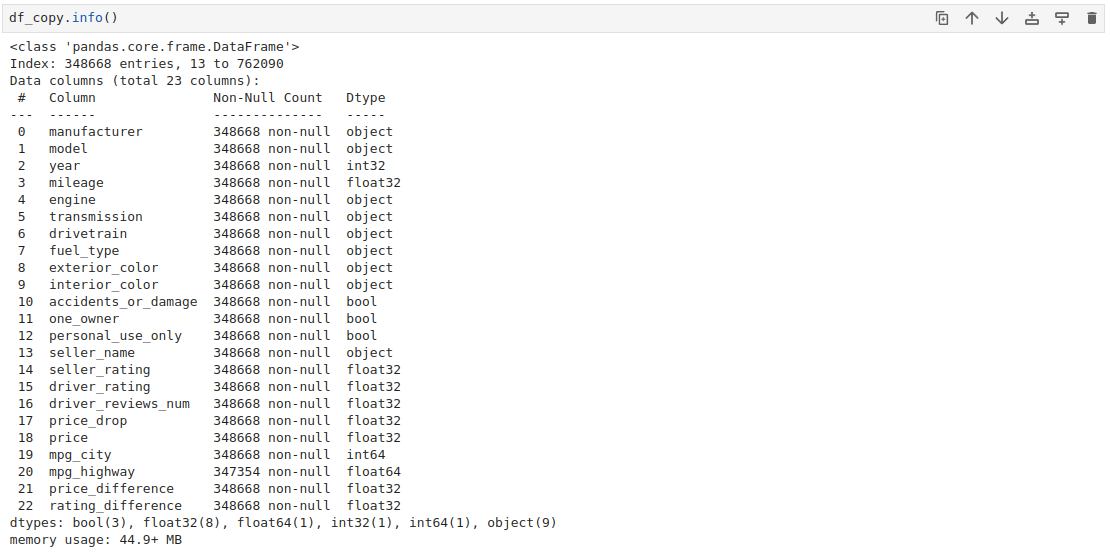
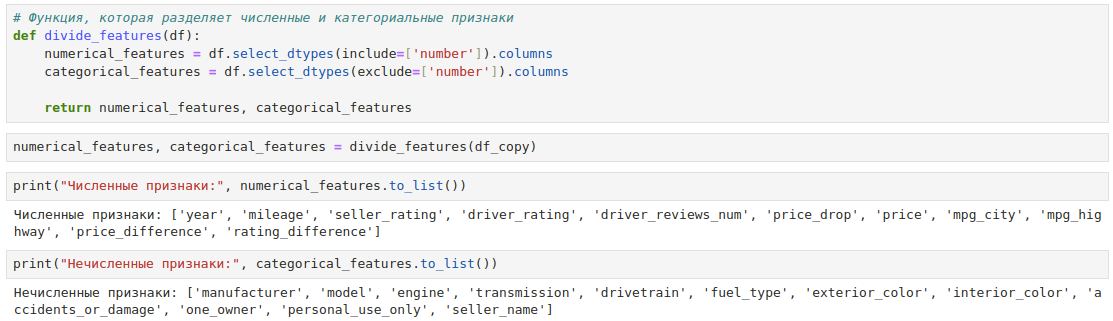
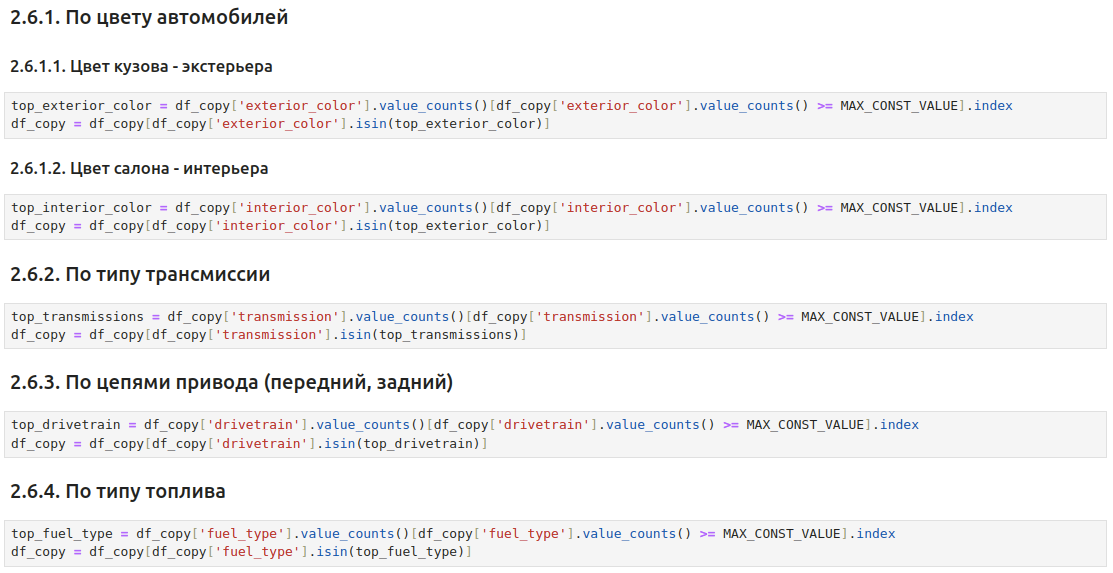
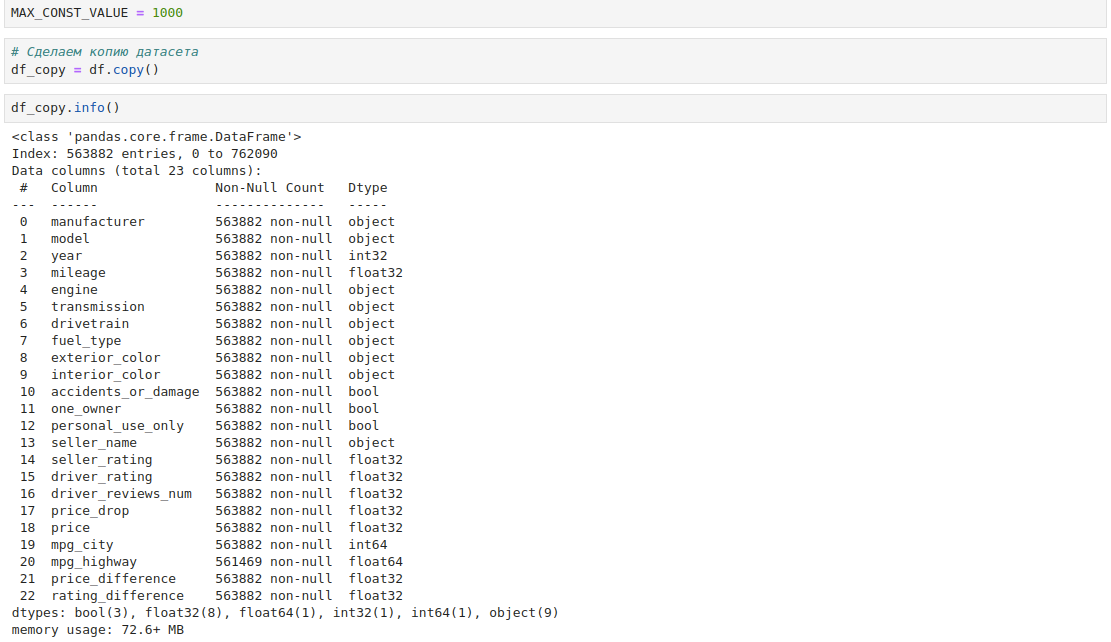
Как и здесь видим, что в какой-то причине рейтинг получилось сильно разным. Давайте рассмотрим график.



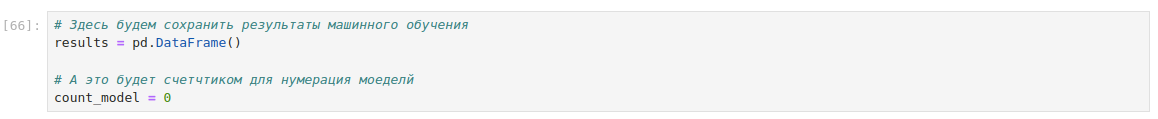
По графике видим, что видны существенные разницы рейтинг между водителями и продавцами. Однако, стоит обратить внимание, что водитель чаще всего оставляют высокий рейтинг нежели продавца.

# **Отсев до определенного кол-во уникальных значений**

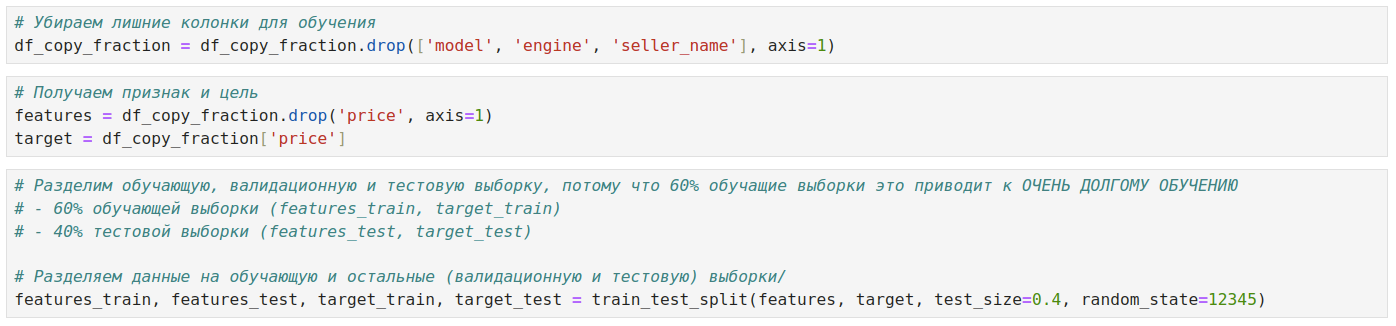
Для кодирования признаков OHE или OH будет черевато, если оставить много уникальных названия, потому что это приведет к созданию много новых признаков. Отсеиваем до небольших количеств, то есть сделаем так, чтобы создали максимум небольших новых закодированных признаков.



# **Машинное обучение**

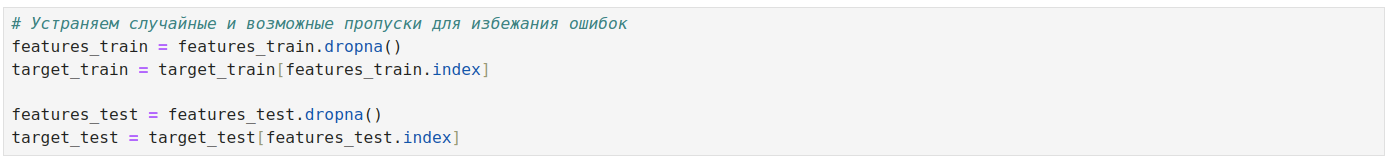
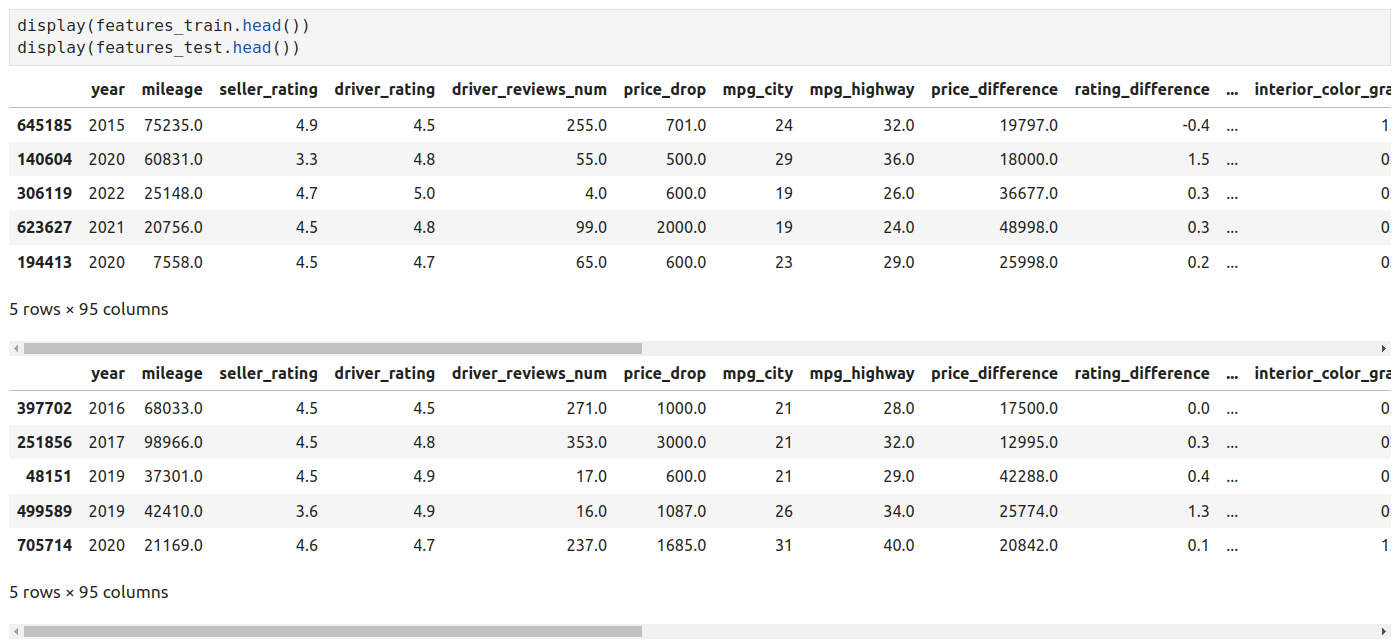
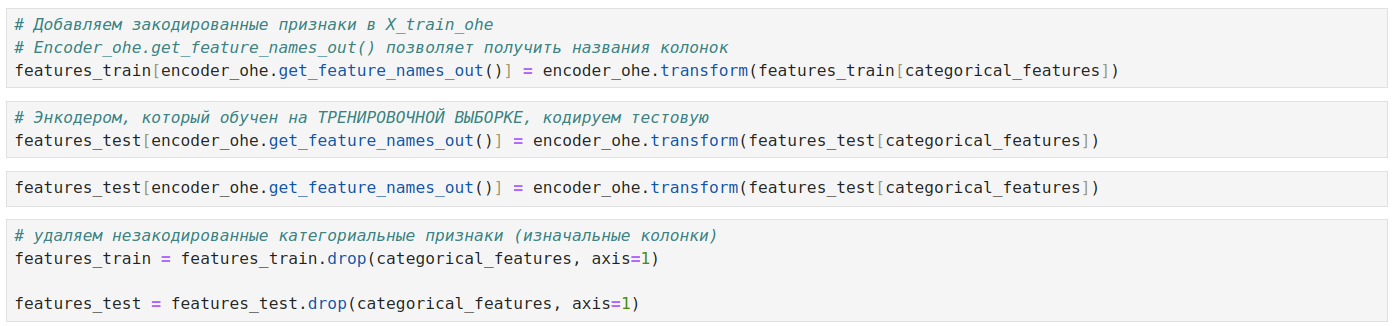
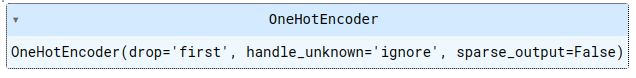
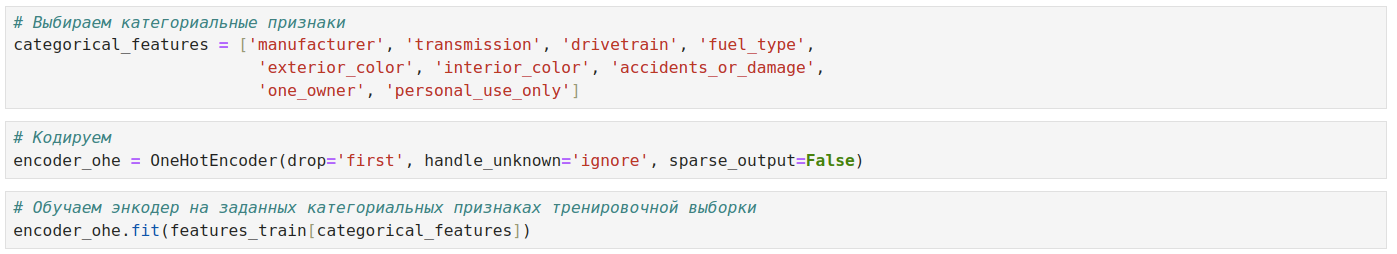


# **Деление на обучающей и тестовой выборки**



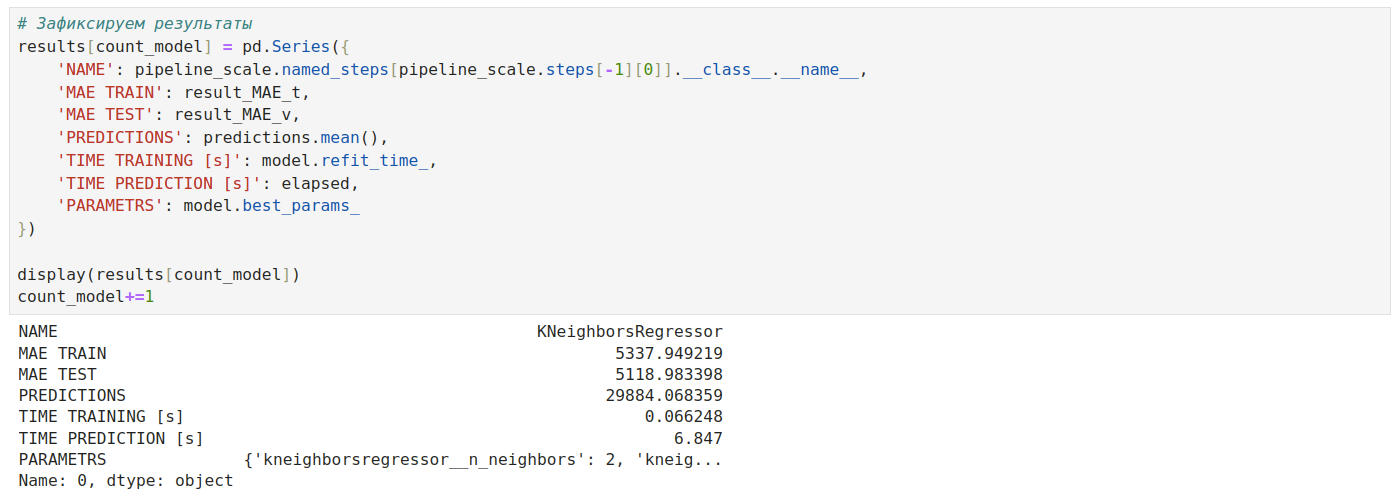
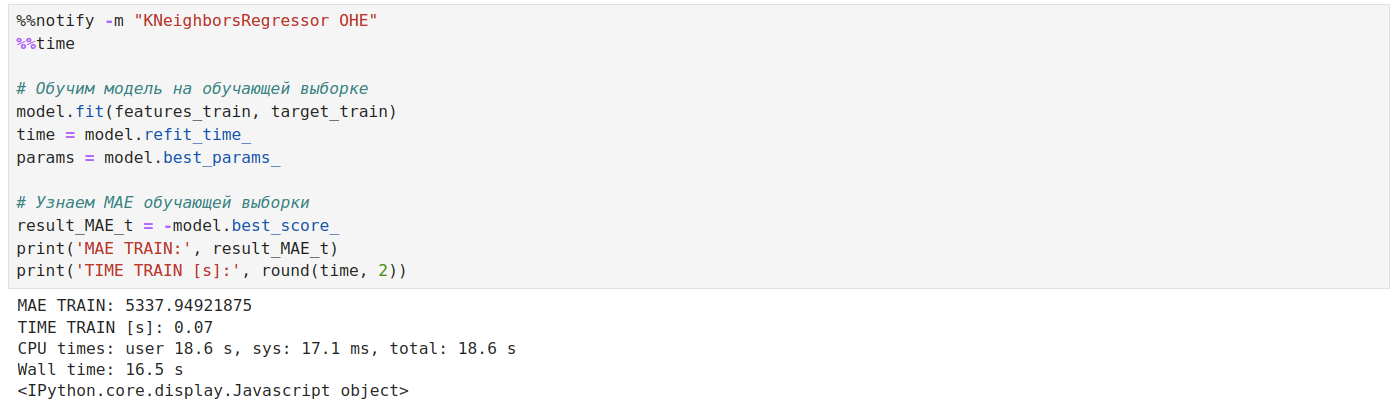
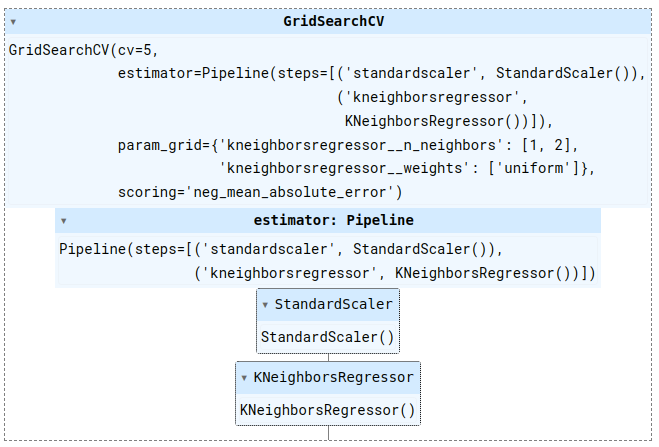
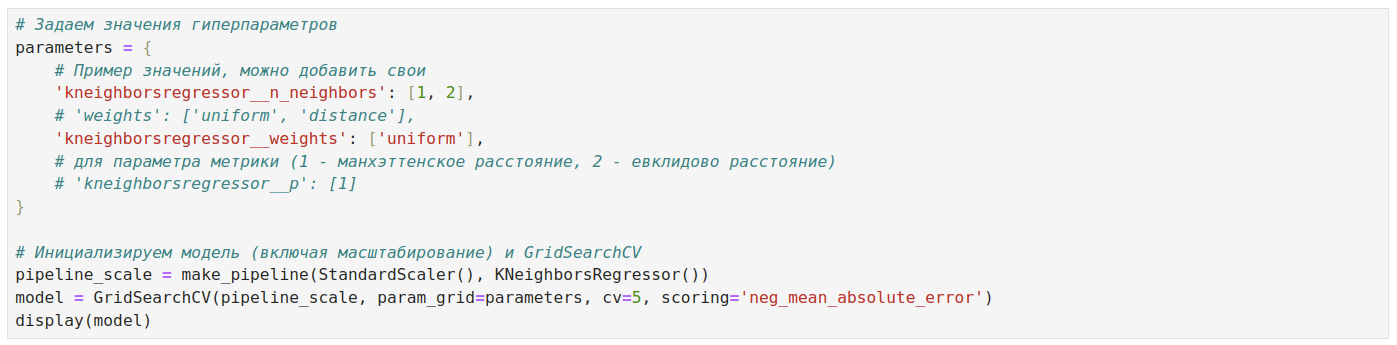
# **Кодирование признаков - прямое кодирование (One-Hot Encoding)**

Закодируем признаки: manufacturer, transmission, drivetrain, fuel\_type, exterior\_color, interior\_color, accidents\_or\_damage,one\_owner, personal\_use\_only



# **Обучение модели**

# **KNeighborsRegressor**



# **Итог**

# **Анализ моделей**

