МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ им. Р.Е.АЛЕКСЕЕВА

Институт радиоэлектроники и информационных технологий

Кафедра «Вычислительные системы и технологии»

Методы машинного обучения и искусственного интеллекта

Отчет

по лабораторной работе №1

«Предварительная обработка данных»

ПРОВЕРИЛ:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Новиков И.А.

СТУДЕНТ:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Елисеев С. М.

23-ИВТ-5

Нижний Новгород

2025 г.

**Цель лабораторной работы** – получение навыков работы с библиотеками анализа данных, предварительной обработки данных, средствами визуализации.

https://github.com/badend11ng/ML/blob/master/LR\_1\_1.ipynb

Вариант№3

**3. Вторичный рынок машин**

1. Заполнить NaN значения в строковых полях (нейтральными значениями, например «неизвестный»).
2. Оцените диапазоны значений для числовых колонок (например, год регистрации, цена, мощность двигателя) используйте violinplot или boxplot.
3. Удалите выбросы. Выведите размер набора данных до и после удаления.
4. Для числовых признаков и цены машины постройте матрицу корреляций (heatplot).
5. Удалите столбцы, ценность которых для оценки стоимости машины низка исходя из данных корреляций и из логики (опишите почему удалили тот или иной столбец).
6. Постройте histplot-ы по маркам автомобилей (топ 15), типам кузова и используемому топливу.
7. Добавьте в данные новый признак, который представляет собой длину названия автомобиля.
8. Коэффициент корреляции Пирсона и Спирмена между стоимостью автомобиля и мощностью двигателя. Оцените значение p-value. Постройте графики плотностей распределения признаков (kdeplot или displot).

**Общие задания:**

1. Результат выполнения следующих функций библиотеки Pandas: head, tail, info, describe, shape. Показать результат их применения к данным.
2. Пустые значения.
   1. Вывести количество пустых значений по колонкам. (отсортировать по убыванию)
   2. Вывести процент пустых значений по колонкам. (отсортировать по убыванию)
   3. Показать выводы функции info до и после очистки.
3. Удаление дубликатов. (drop\_duplicates) Вывести количество строк до удаления и после. Так же вывести количество найденных дубликатов.

**Описание данных**

Более 370000 подержанных автомобилей были разобраны с помощью Scrapy на досках объявлений eBay. Содержание данных на немецком языке. Эти поля включены: autos.csv:

dateCrawled: когда это объявление было впервые просканировано, все значения полей берутся из этой даты

name: "имя" автомобиля

seller: частное лицо или дилер

offerType: тип предложения

price: цена в объявлении о продаже автомобиля

abtest: тест (проверка)

vehicleType: тип транспортного средства

yearOfRegistration: в каком году автомобиль был впервые зарегистрирован

gearbox: коробка передач

powerPS: мощность автомобиля в лошадиных силах

model: модель

kilometer: сколько километров проехал автомобиль

monthOfRegistration: в каком месяце автомобиль был впервые зарегистрирован

fuelType: тип топлива

brand: бренд

notRepairedDamage: если у автомобиля есть повреждение, которое еще не отремонтировано

dateCreated: дата создания объявления на eBay

nrOfPictures: количество фотографий в объявлении (к сожалению, это поле везде содержит 0 и поэтому бесполезно (ошибка в сканере!))

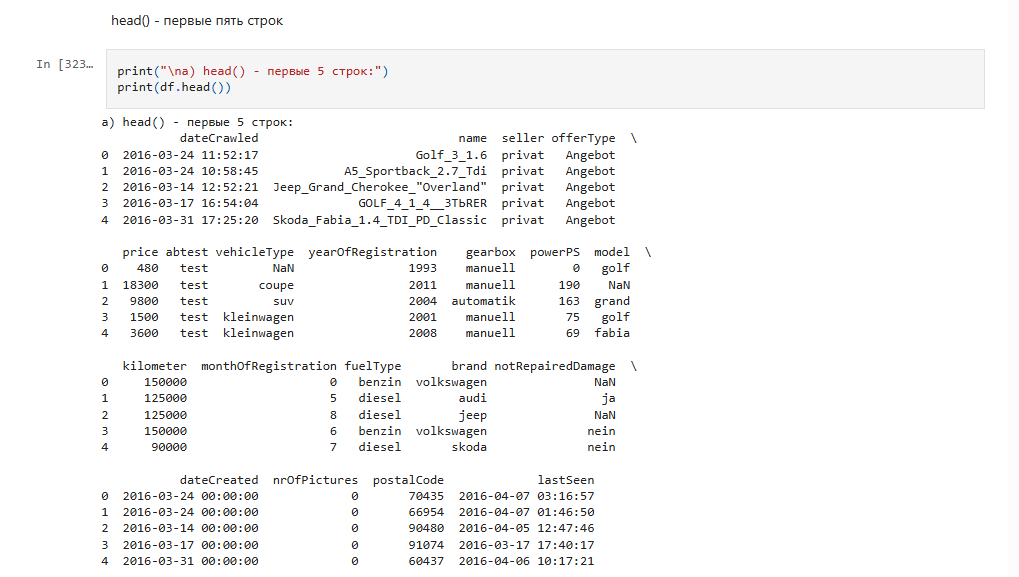
postalCode: почтовый индекс

lastSeenOnline: когда поисковый робот последний раз видел это объявление онлайн

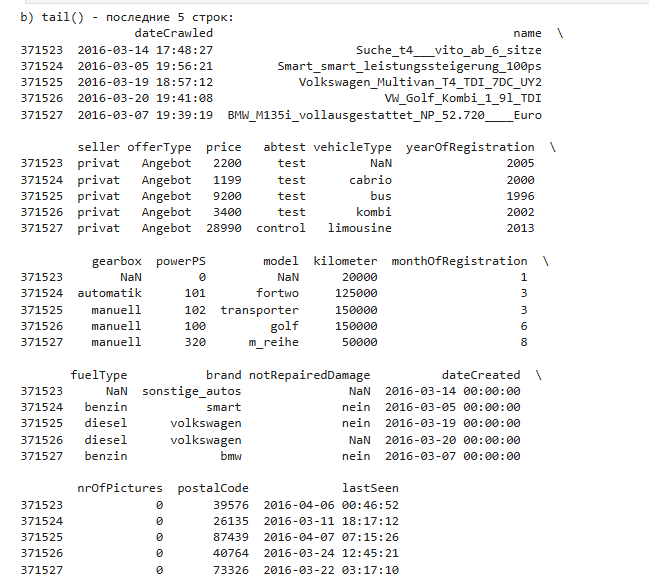
Поля lastSeen и dateCreated можно использовать для оценки того, как долго автомобиль будет находиться в эксплуатации, прежде чем будет продан.

**Задание 1. Выполнение некоторых функций библиотеки Pandas**

df.head() - Отображает первые n строк DataFrame. По умолчанию n = 5.

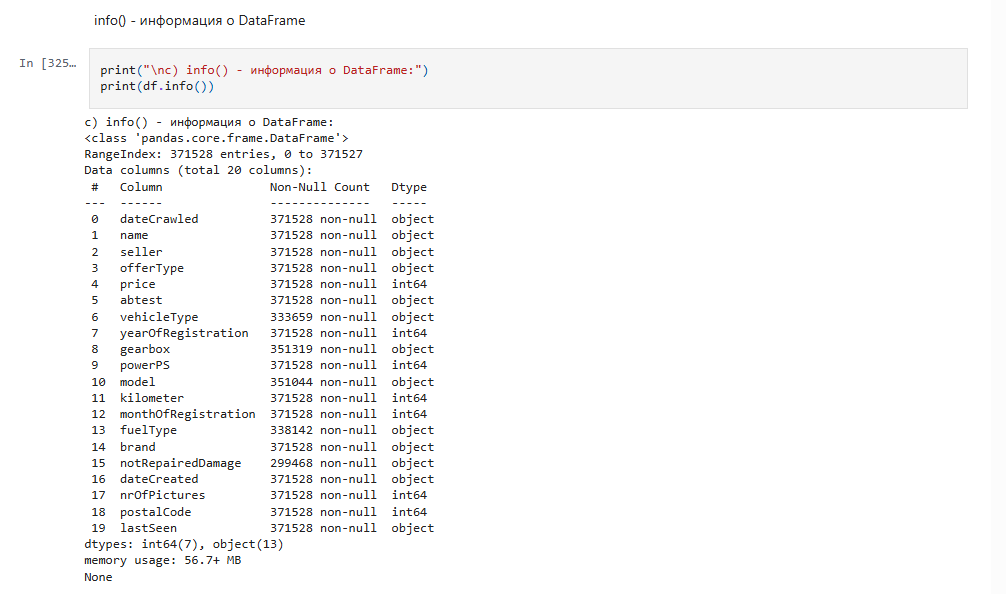


df.tail() - Отображает последние n строк DataFrame. По умолчанию n = 5.



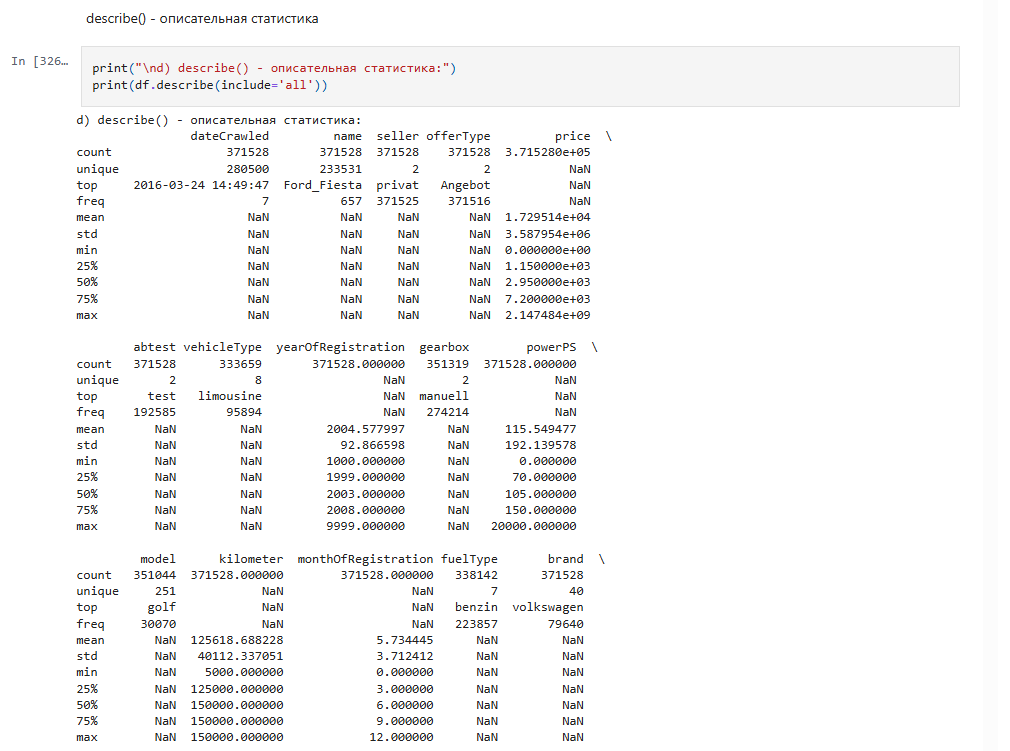
df.info() - Выводит краткую информацию о DataFrame:

* + Количество строк и столбцов.
  + Типы данных в каждом столбце.
  + Количество ненулевых значений.



df.describe() - Предоставляет статистическую сводку по числовым столбцам:

* + Количество, среднее, стандартное отклонение, минимум, максимум и квартили.

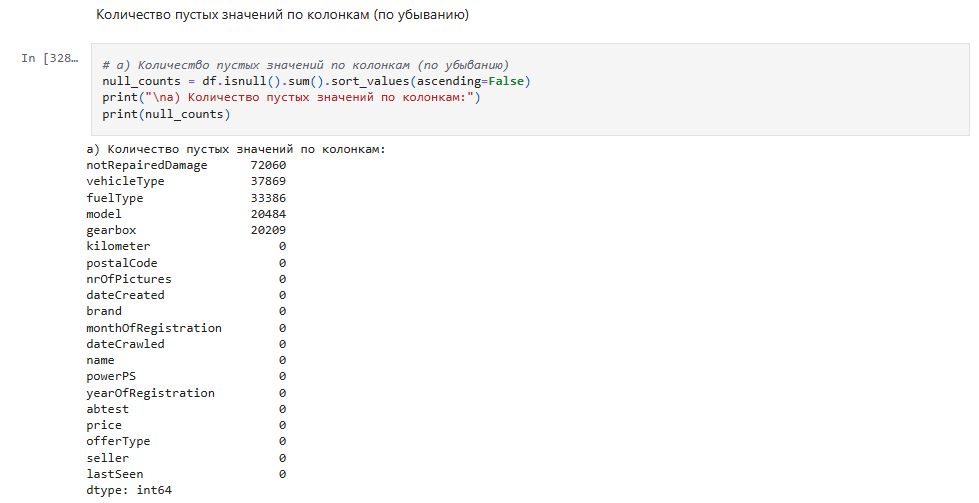


df.shape - Возвращает кортеж с количеством строк и столбцов.

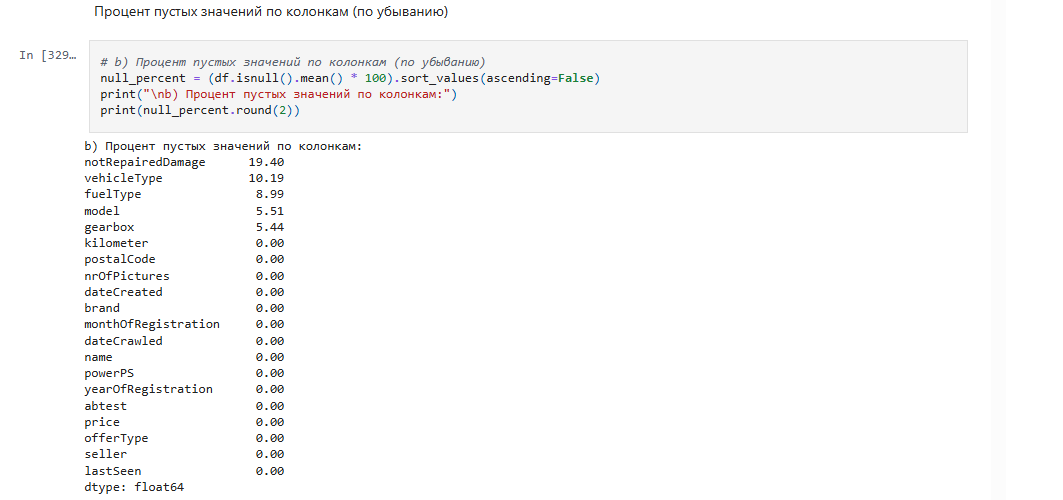


2. Пустые значения.

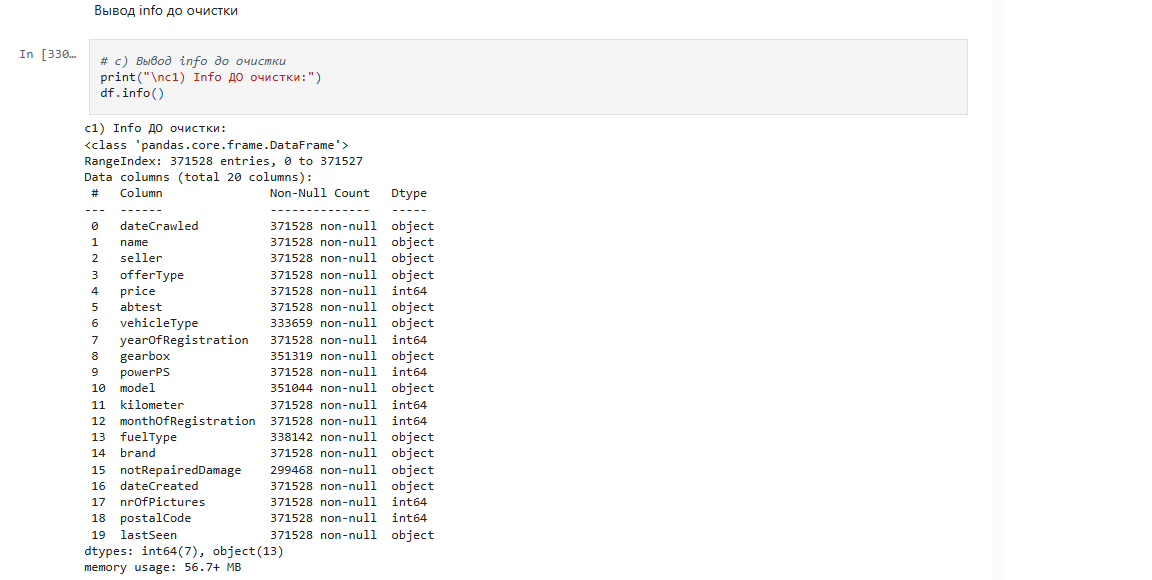
* 1. Вывести количество пустых значений по колонкам. (отсортировать по убыванию)

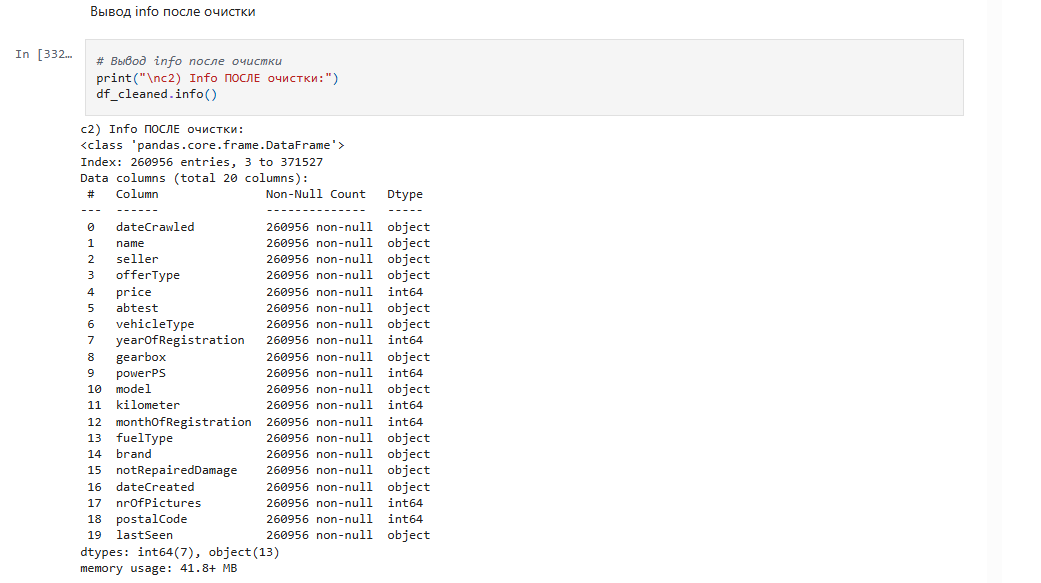


* 1. Вывести процент пустых значений по колонкам. (отсортировать по убыванию)



* 1. Показать выводы функции info до и после очистки.





3. Удаление дубликатов. Вывести количество строк до удаления и после. Так же вывести количество найденных дубликатов.

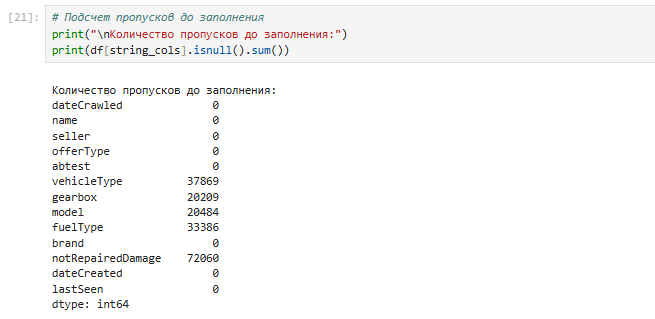




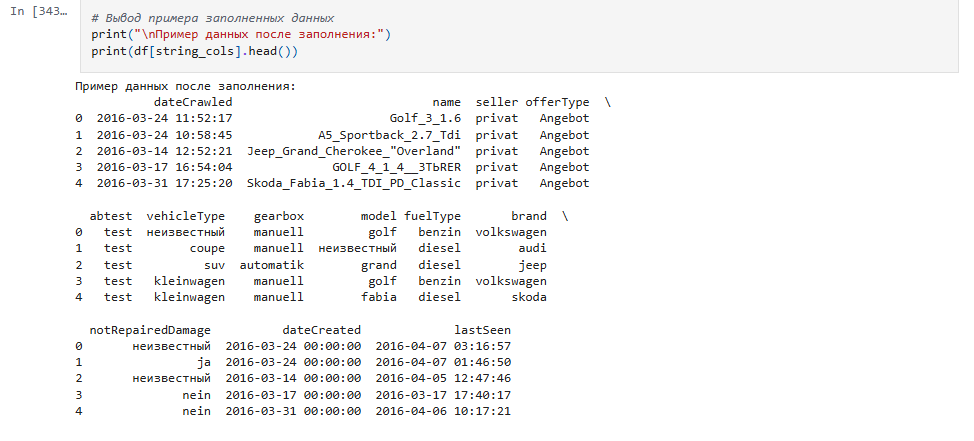
**Задание по варианту**

1.Заполнить NaN значения в строковых полях (нейтральными значениями, например «неизвестный»).

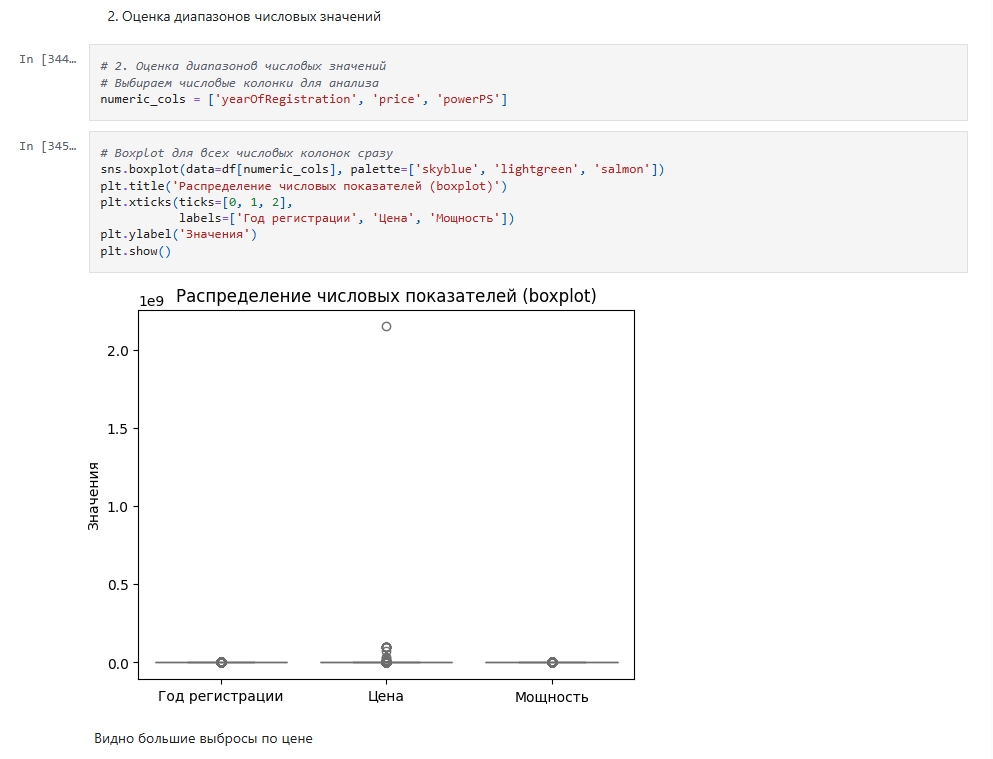






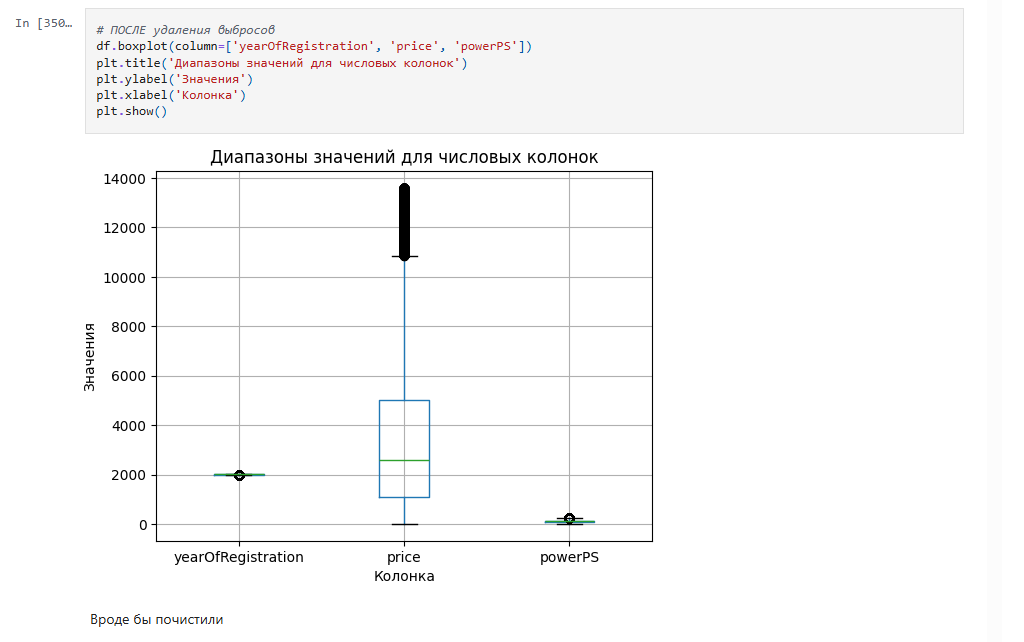


2.Оцените диапазоны значений для числовых колонок (например, год регистрации, цена, мощность двигателя) используйте violinplot или boxplot.



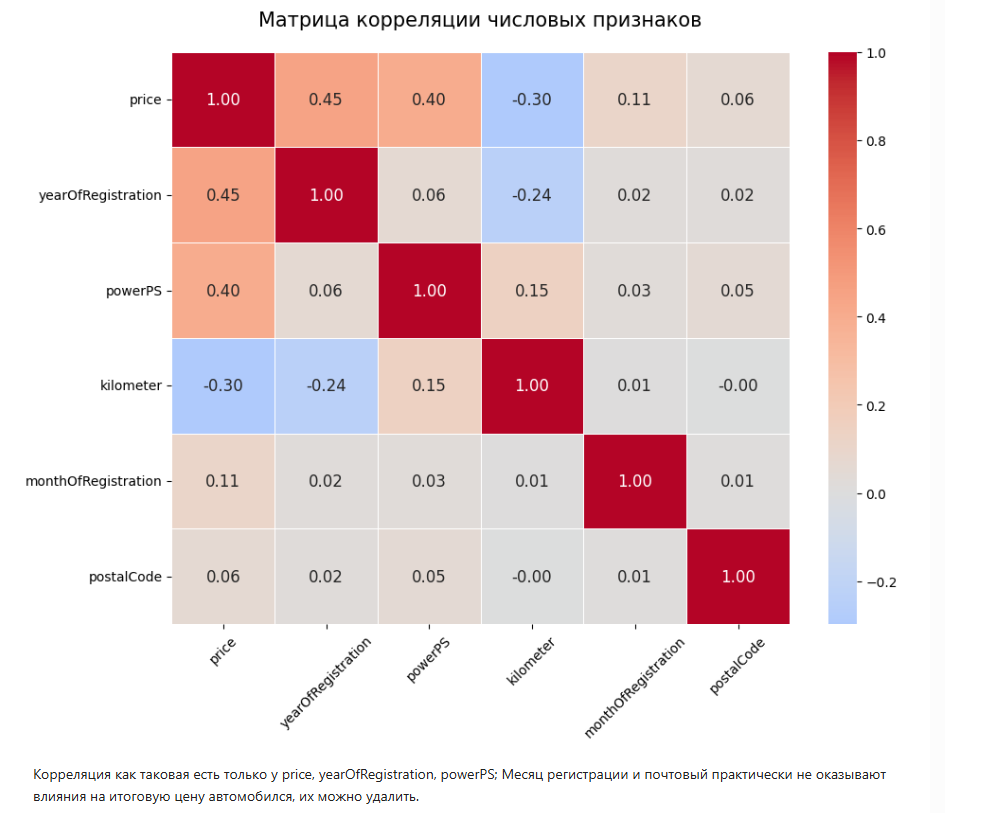
3.Удалите выбросы. Выведите размер набора данных до и после удаления.





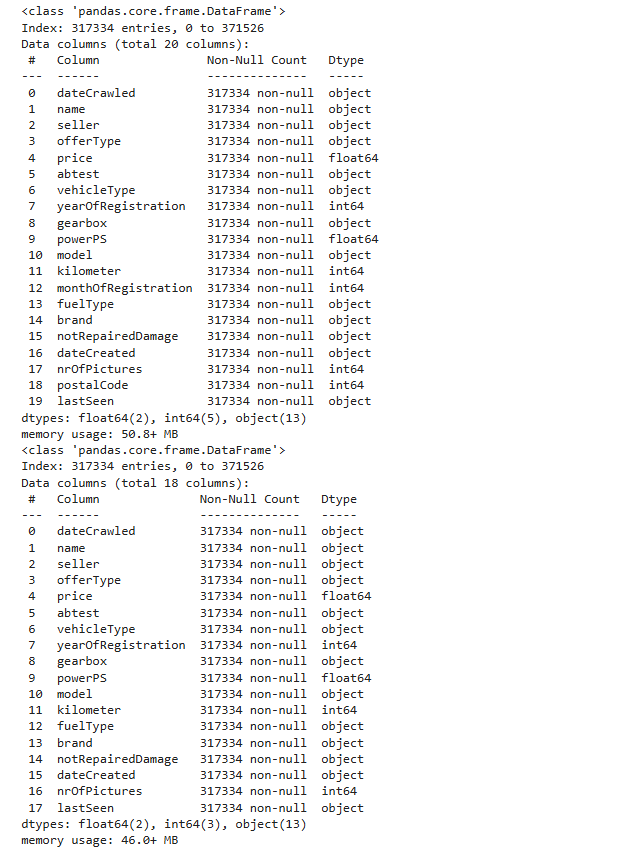
4.Для числовых признаков и цены машины постройте матрицу корреляций (heatplot).



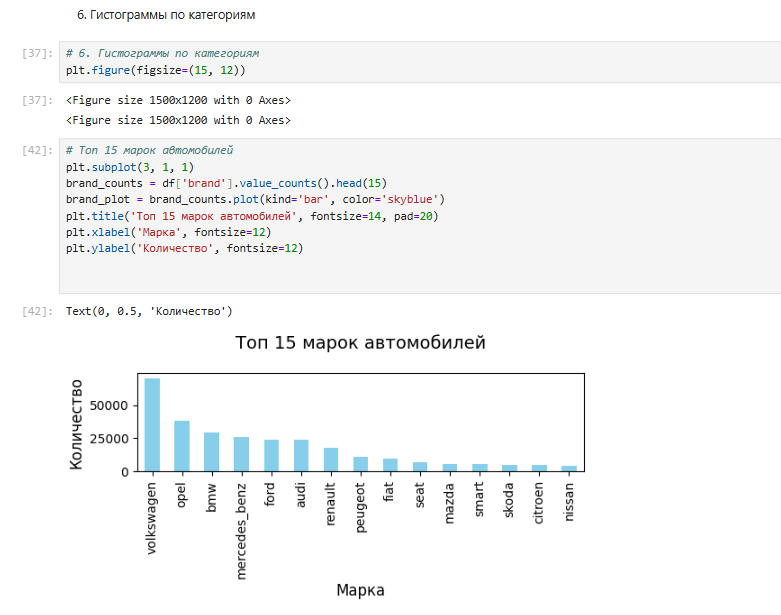


5.Удалите столбцы, ценность которых для оценки стоимости машины низка исходя из данных корреляций и из логики (опишите, почему удалили тот или иной столбец).

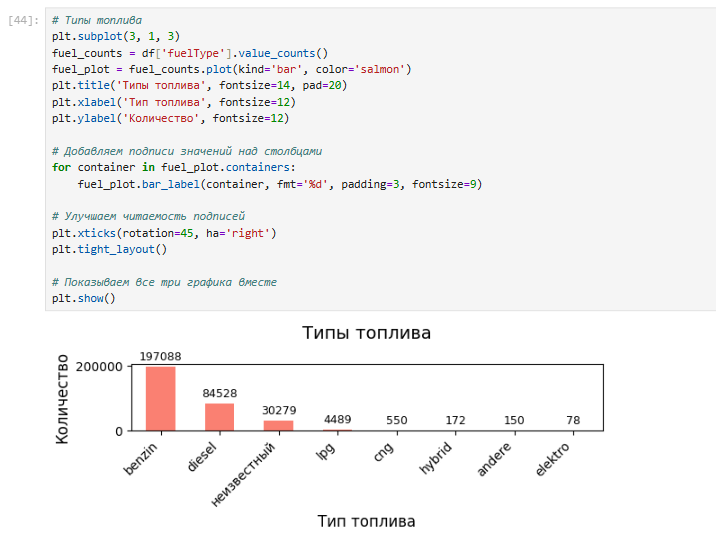




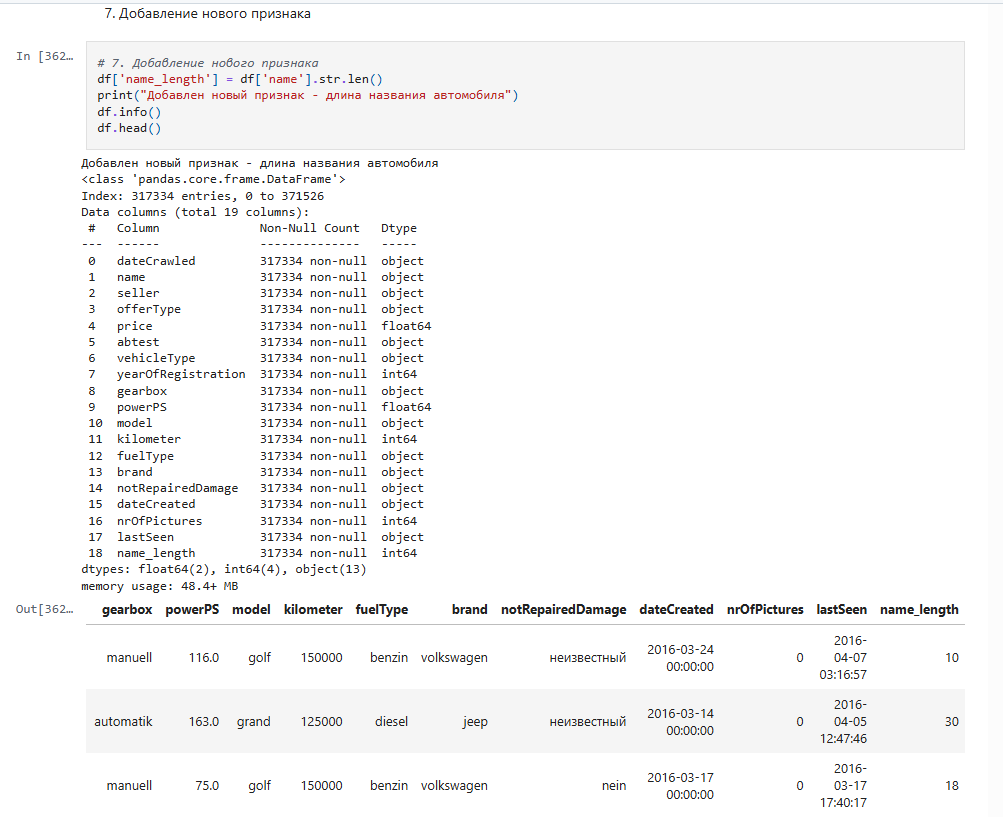
6.Постройте histplot-ы по маркам автомобилей (топ 15), типам кузова и используемому топливу.



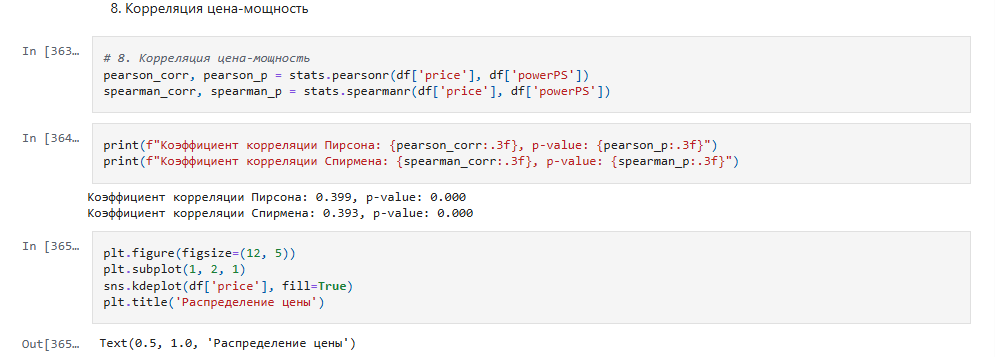


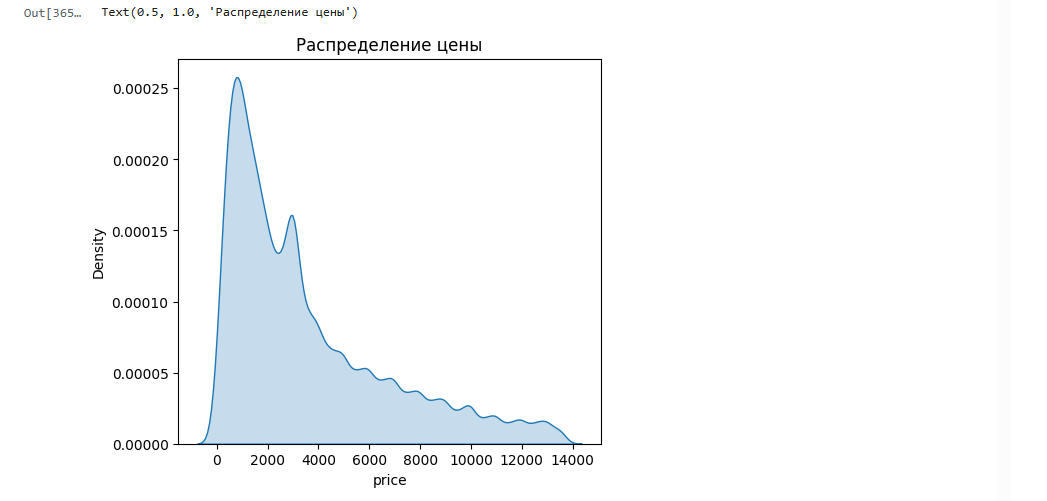


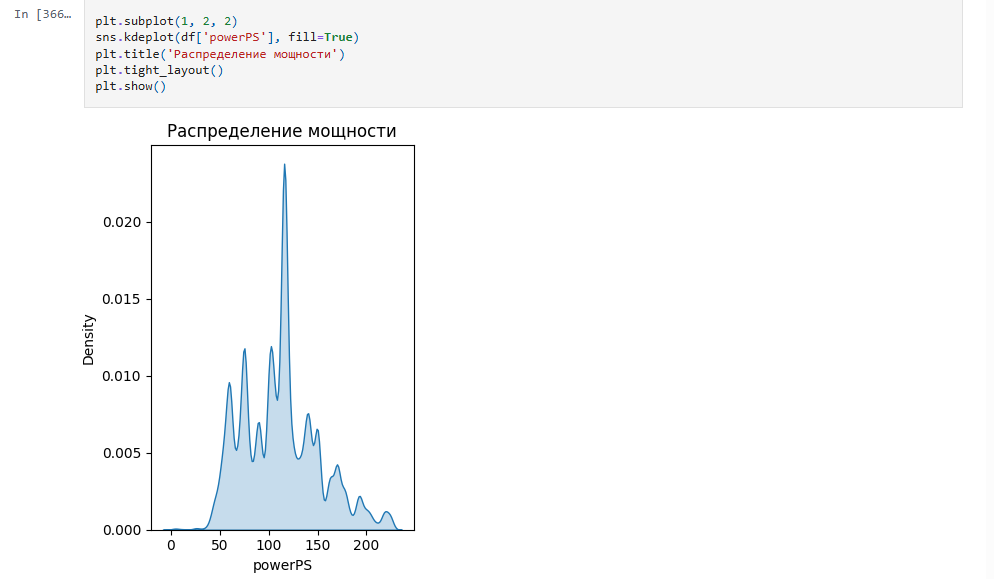
7.Добавьте в данные новый признак, который представляет собой длину названия автомобиля.



8.Коэффициент корреляции Пирсона и Спирмена между стоимостью автомобиля и мощностью двигателя. Оцените значение p-value. Постройте графики плотностей распределения признаков (kdeplot или displot).







**Выводы:** удалось понять работу основных функций библиотеки Pandas, появилось понимание об этапах обработки данных, были освоены базовые инструменты визуализации.