**Федеральное государственное бюджетное образовательное учреждение профессионального образования**

**Нижегородский государственный**

**технический университет имени Р.Е. Алексеева**

**Лабораторная работа**

“ **Предварительная обработка данных**”

Работу выполнил Борисов Д.Я

Работу принял Гай В.Е.

г. Нижний Новгород

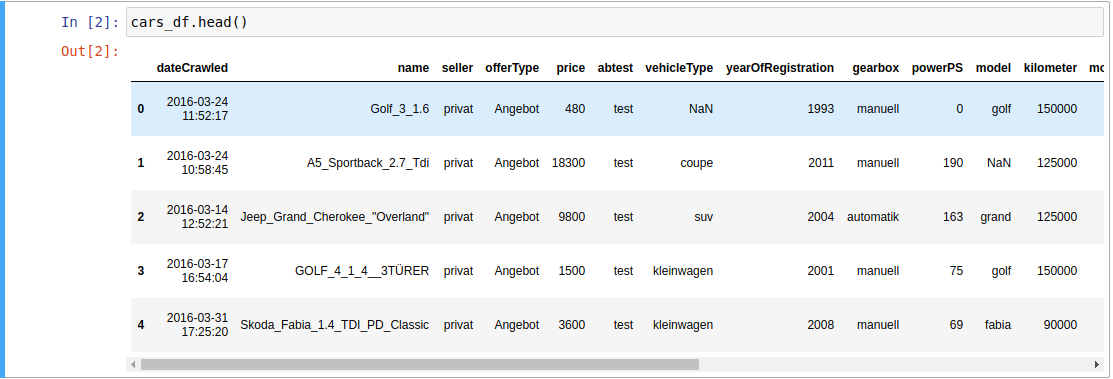
2020 г.

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

**Вариант 3. Вторичный рынок машин**

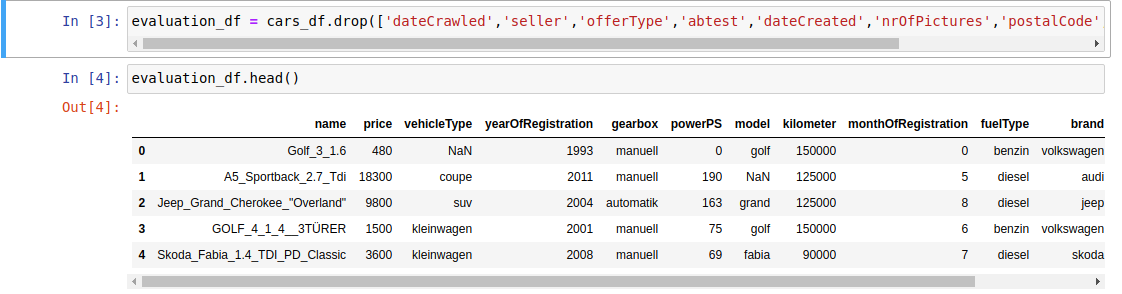
Для начало импортируем необходимые библиотеки и создаем переменную cars\_df в которую поместим значения данных с которыми будем работать.

Команда %matplotlib inline указывает, что график необходимо построить все в той же оболочке Jupyter, но теперь он выводится как обычная картинка.



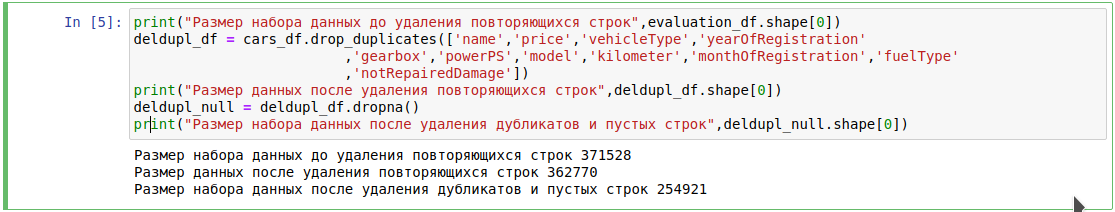
С помощью метода head() выводим заданное количество строк на экран.

**1. Удалите столбцы, ценность которых для оценки стоимости машины низка.**



В переменную evalution\_df помещаем таблицу с удаленными столбцами с помощью метода drop.

### **2. Удалите повторяющиеся строки, строки содержащие пропуски в данных. Выведите размер набора данных до и после удаления.**

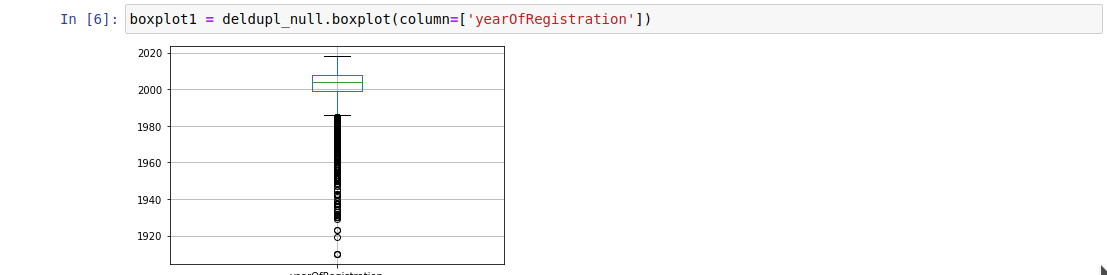


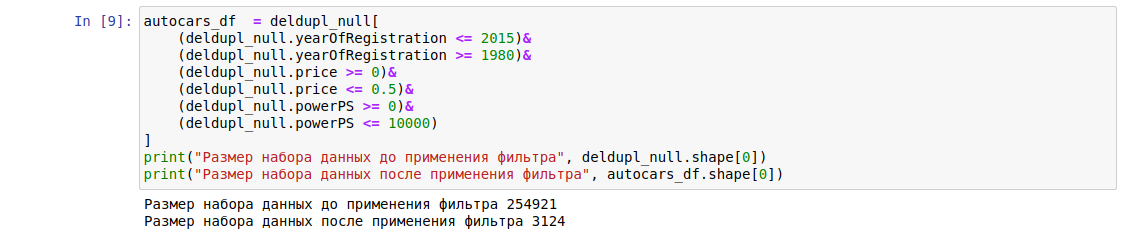
С помощью метода .shape вычисляем количество строк в таблице.

С помощью метода .drop\_duplicates удаляем повторяющиеся строки.

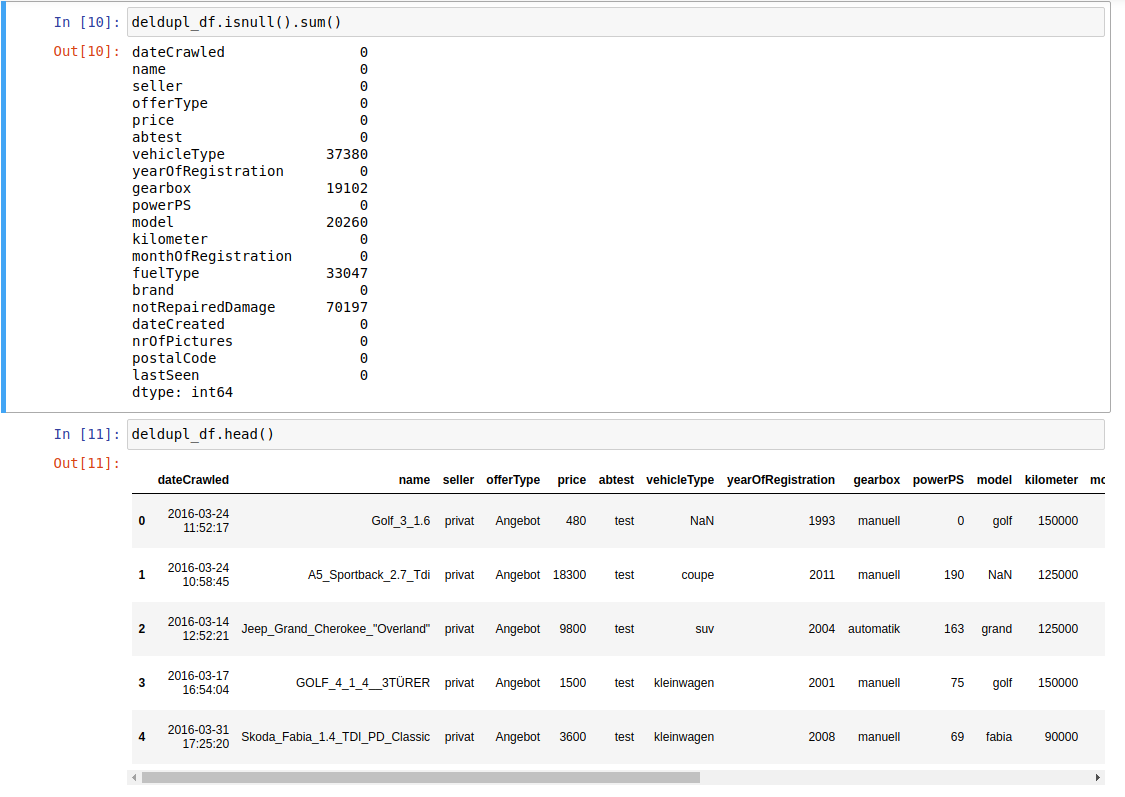
С помощью метода .dropna удаляем пустые строки.

### **3. Удалите данные в строках, выходящие за некоторые пределы (год регистрации, цена, мощность двигателя). Выведите размер набора данных до и после удаления. Для оценки диапазонов значений признаков используйте BoxPlot.[¶](http://localhost:8888/notebooks/my_project_env/Borisov_lab1_var3.ipynb" \l "3.-Удалите-данные-в-строках,-выходящие-за-некоторые-пределы-(год-регистрации,-цена,-мощность-двигателя).-Выведите-размер-набора-данных-до-и-после-удаления.-Для-оценки-диапазонов-значений-признаков-используйте-BoxPlot.)**

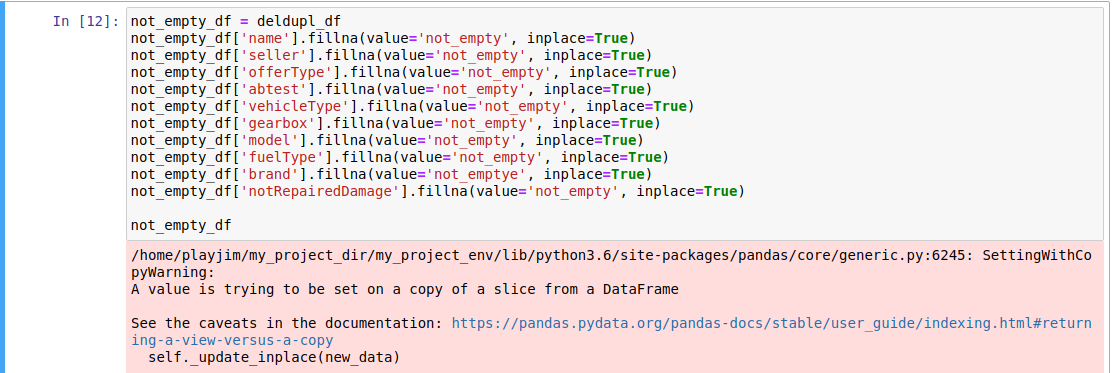
****

В таблице autocars\_df удаляем строки которые не входят в заданные значения.

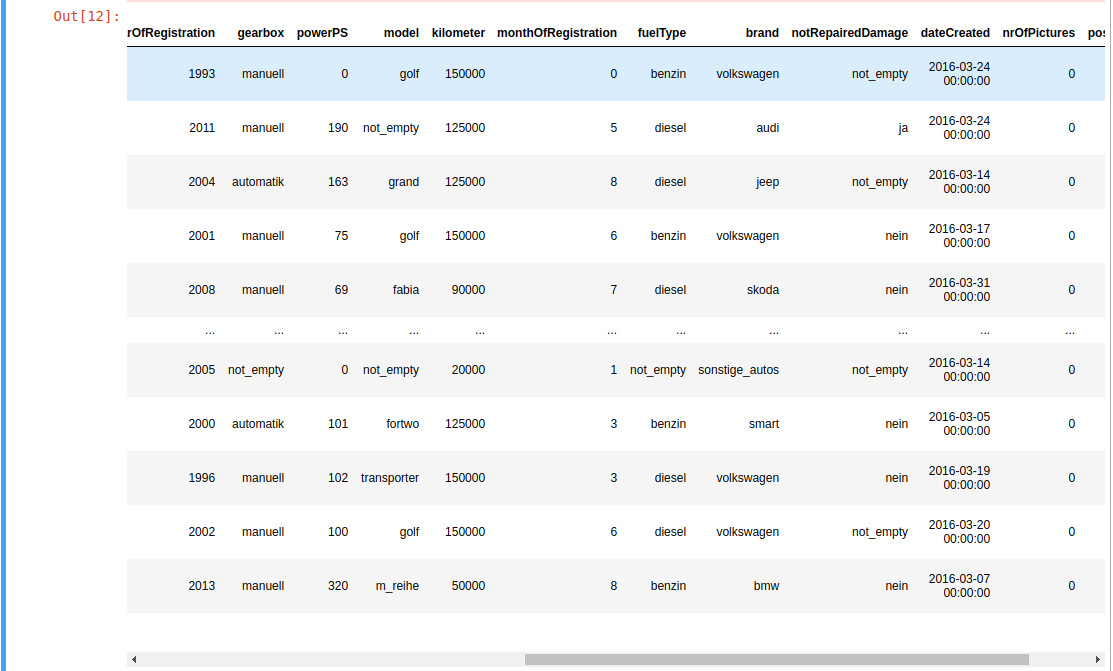
### **4. Заполнить пропущенные данные в строковых полях.**



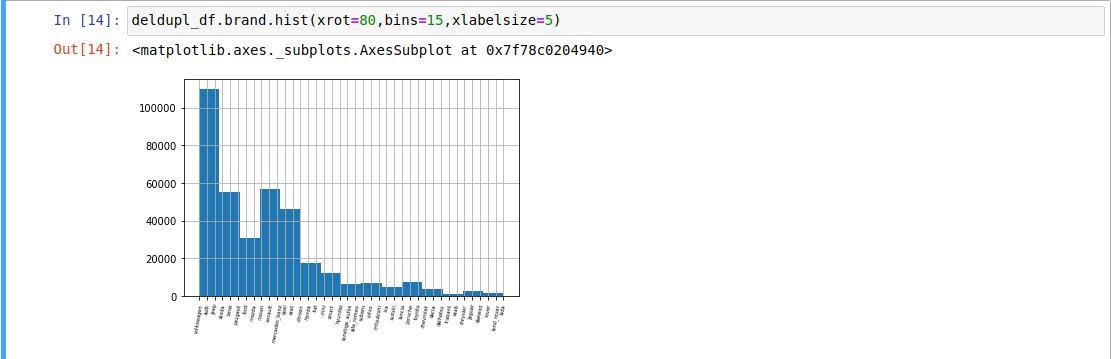
С помощью методов .isnull().sum() считаю количество пустых значений аттрибутов.

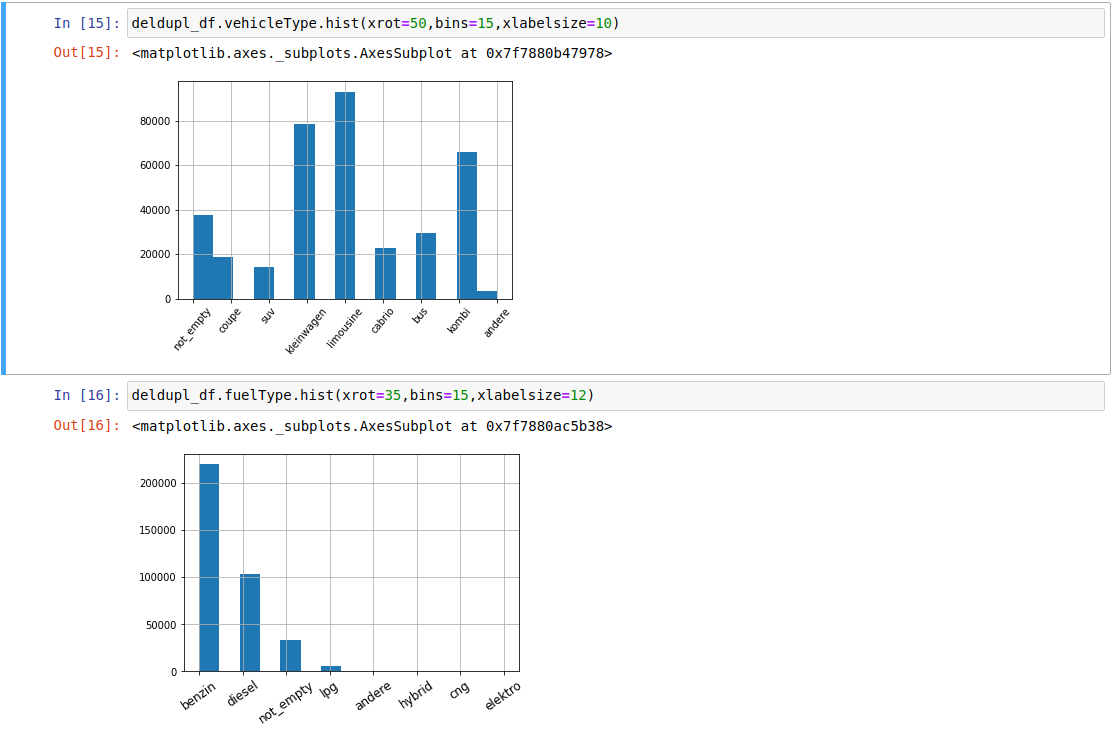


Методом .fillna заполняю пустые аттрибуты значением not\_empty

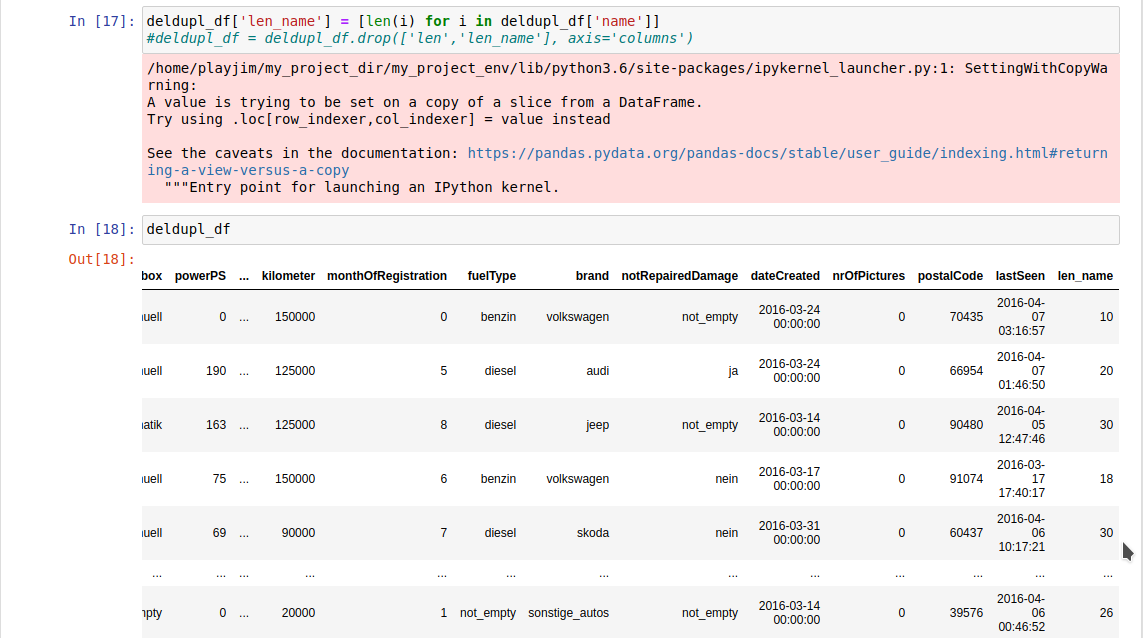


### **5. Постройте гистограмму по маркам автомобилей, типам кузова и используемому топливу.**



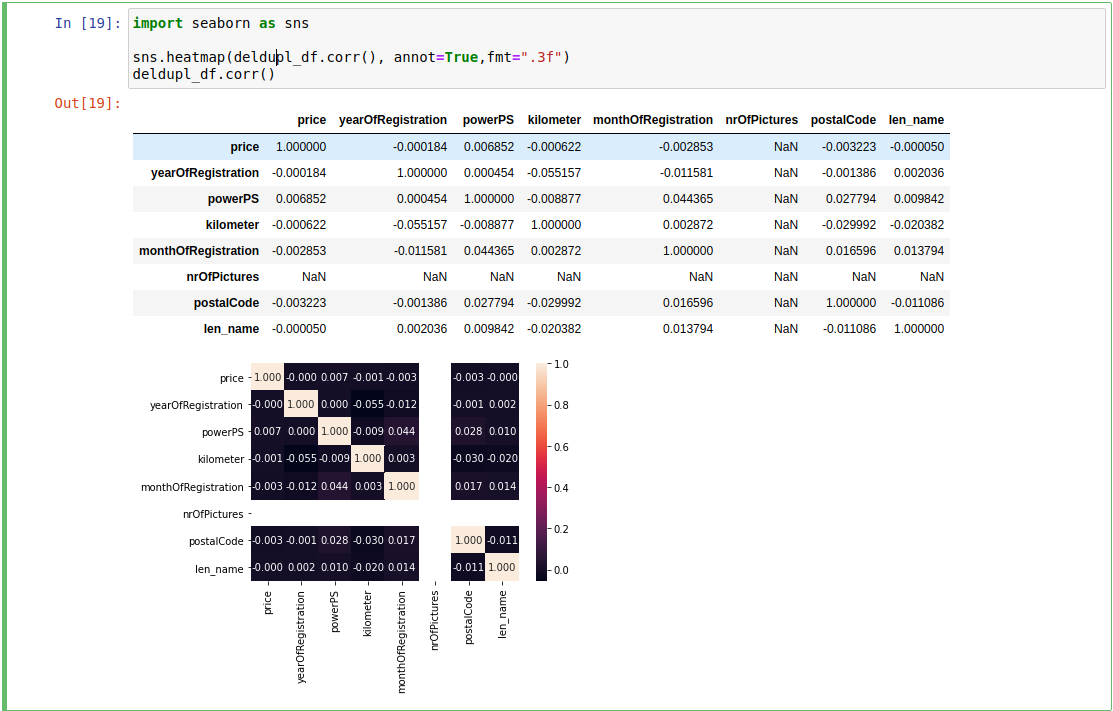
С помощью метода .hist построил гистограммы по маркам автомобилей, типам кузова и используемому топливу.

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



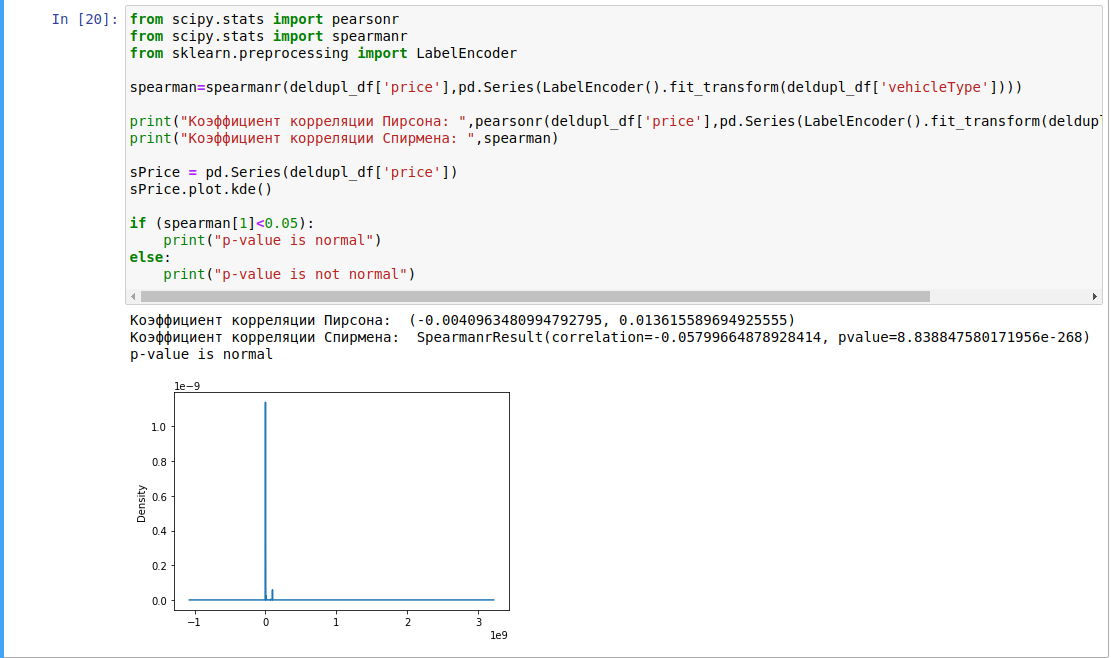
Создаю столбец len\_name и передаю в него значение длины названия автомобиля с помощью цикла for.

### **7. Постройте карту корреляций между признаками. Выведите также числовые значения признаков.**



Импортирую библеотеку seaborn и создаю карту корреляции таблицы deldupl\_df.

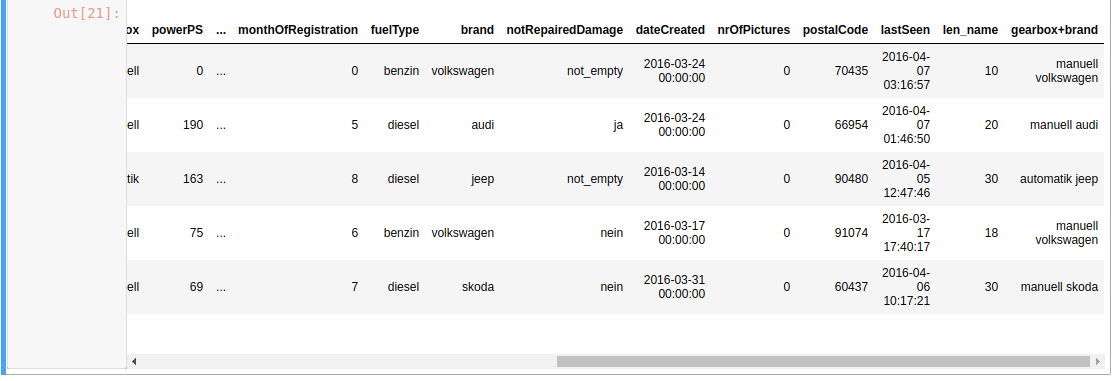
### **8. Определите коэффициент корреляции Пирсона и Спирмена между стоимостью автомобиля и типом кузова. Оцените значение p-value. Постройте плотность распределения признаков.**



С помощью метода pearsonr вычисляем значение корреляции Пирсона, для Спирмена метод spearmanr. Строим гистограмму с помощью plot.kde().

### **9. Добавьте в набор данных признак, являющийся суммой двух других признаков.[¶](http://localhost:8888/notebooks/my_project_env/Borisov_lab1_var3.ipynb" \l "9.-Добавьте-в-набор-данных-признак,-являющийся-суммой-двух-других-признаков.)**



Добавил в таблицу столбец gearbox+brand и заполнил его значением gearbox и brand(Коробка передачи и название бренда).