****Московский государственный технический университет им. Н.Э. Баумана  
Кафедра «Системы обработки информации и управления»

Лабораторная работа №2  
по дисциплине  
«Методы машинного обучения»  
на тему

**«Разведочный анализ данных. Исследование и визуализация данных»**

Выполнил:  
студент группы ИУ5И-21М  
Ван Чжэн

Москва — 2024 г.

**1.Цель лабораторной работы:**

Изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

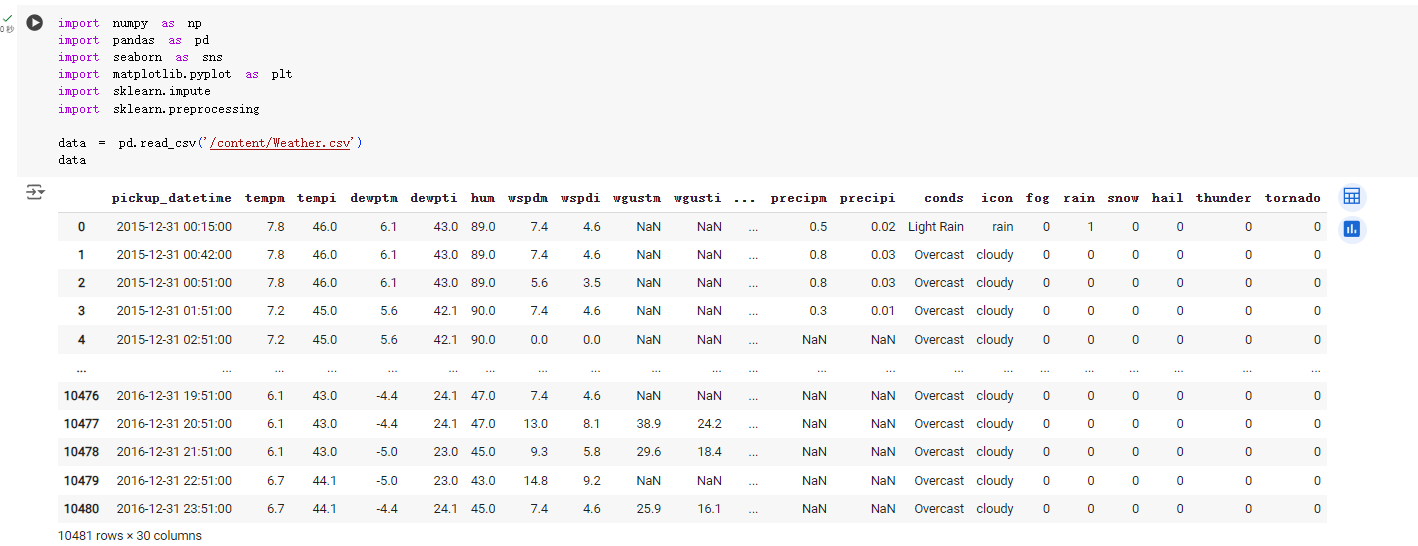
**2.Задание:**

1. Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   * устранение пропусков в данных;
   * кодирование категориальных признаков;
   * нормализацию числовых признаков.

**3.Ход выполнения работы**

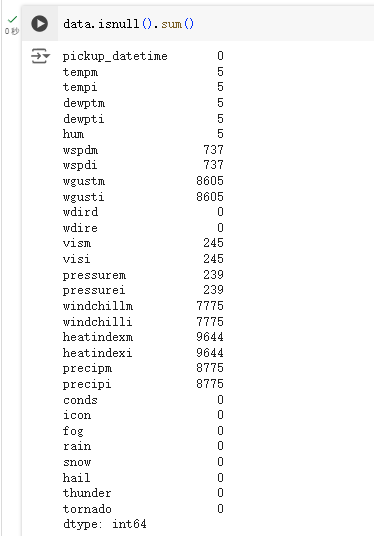
Набор данных-«Нью-Йорк - Почасовые данные о погоде».

Все данные - 10481\*30.

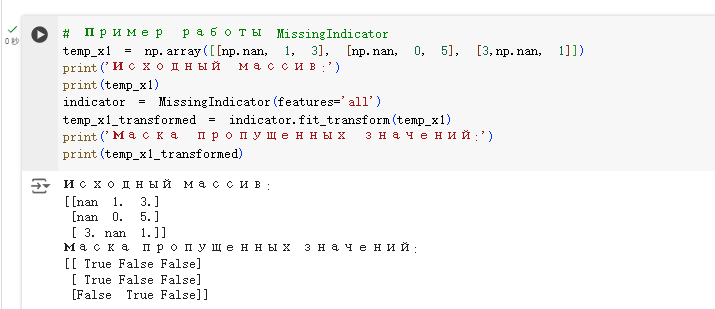


* 1. **Обработка пропусков в данных**

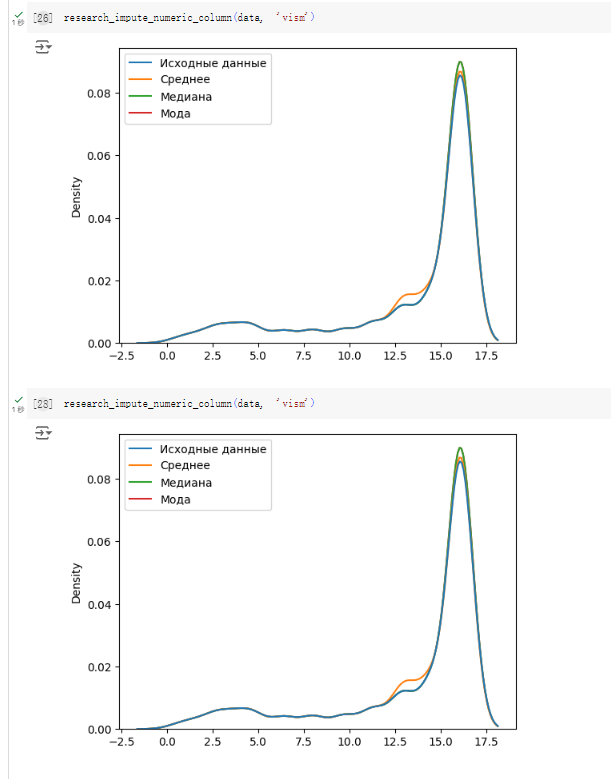
Найдем все пропуски в данных:



Очевидно, что мы можем выбрать столбец «vism»( Vivibility in Km/Видимость в км). Давайте попробуем четыре разные стратегии заполнения на основе: «среднее», «медиана», «наиболее часто встречающаяся» и «постоянная»



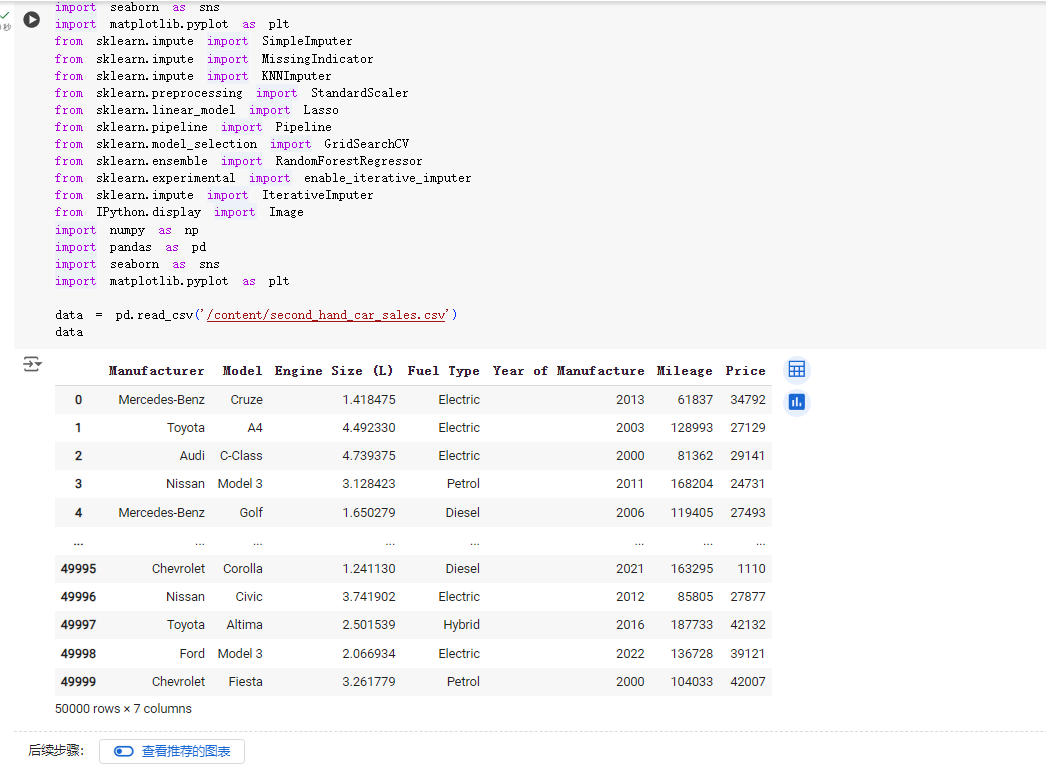


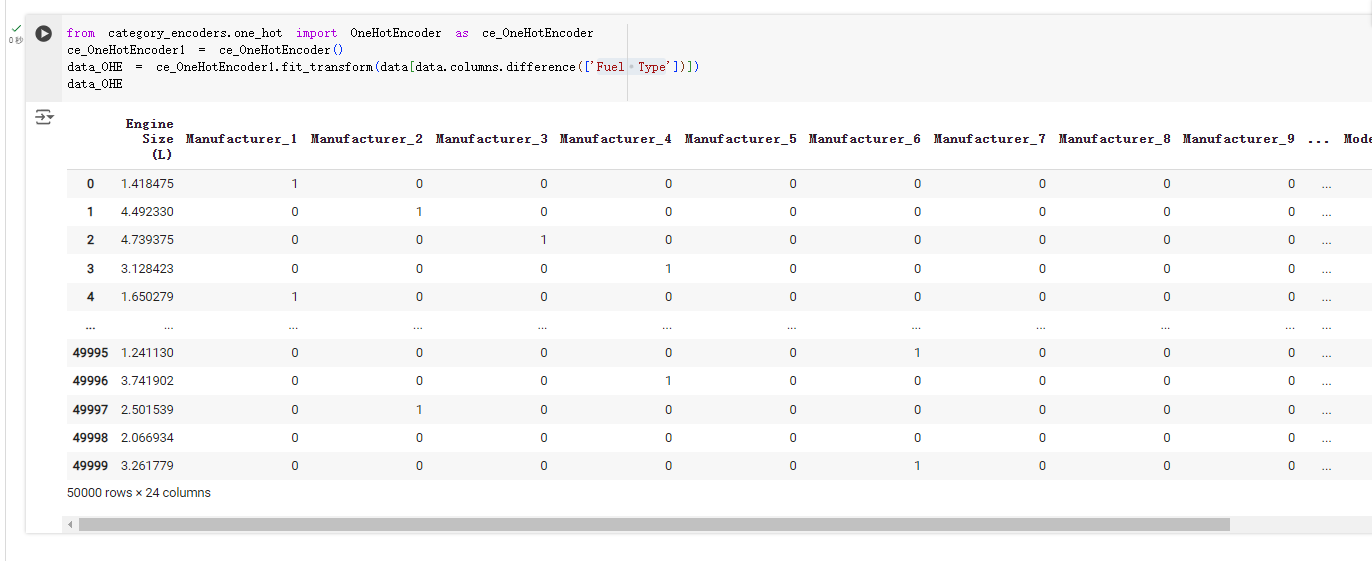


Распределения одномодальные, поэтому можно использовать для импутации моду.

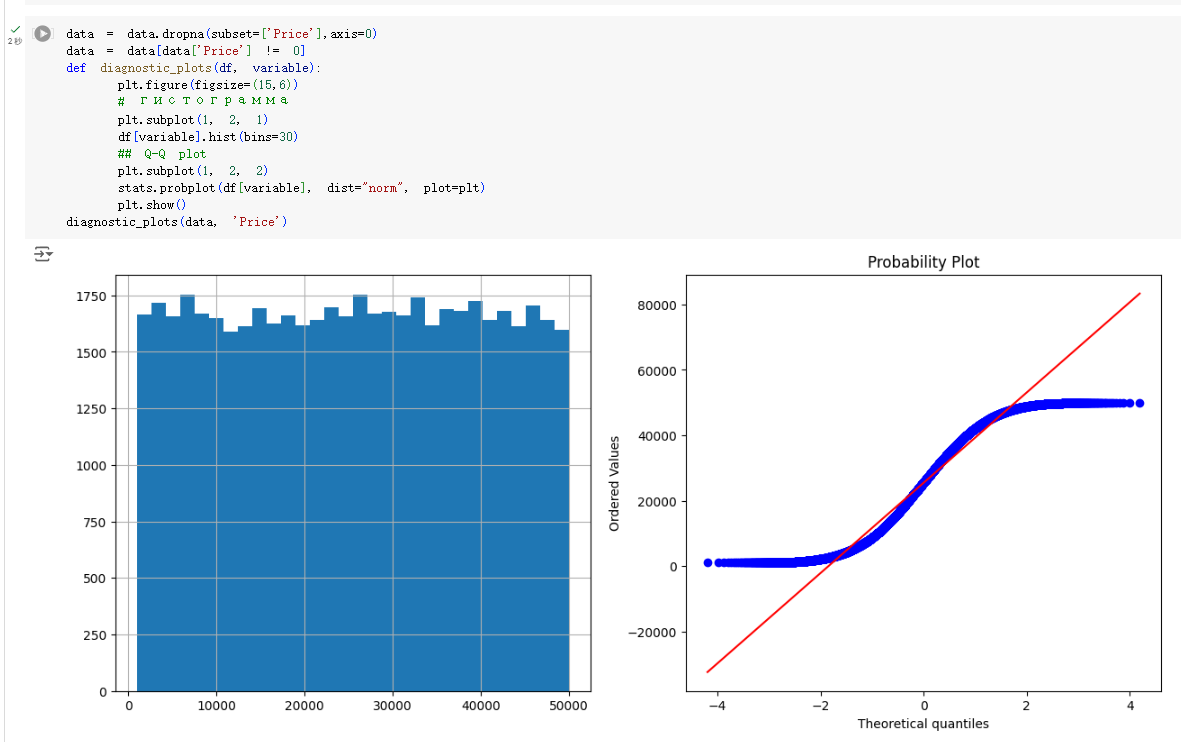
* 1. **Кодирование категориальных признаков**

Я выбрал ещё набор данных-«продажа подержанных автомобилей». Я использовал One-hot encoding для кодирования категориальных признаков.





* 1. **Нормализация числовых признаков**



**Список литературы**

[1] Гапанюк Ю. Е. Лабораторная работа «Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных» [Электронный ресурс] // GitHub. –– 2019. –– Режим доступа: https://github.com/ugapanyuk/ml\_course/wiki/LAB\_MISSING (дата обращения: 05.04.2019).

[2] Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource]//Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/stable/ (online; accessed: 20.02.2019).

[3] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. — Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).

[4] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode: http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).

[5] Gupta L. Google Play Store Apps [Electronic resource] // Kaggle. — 2019. — Access mode: https://www.kaggle.com/lava18/google-play-store-apps (online; accessed:05.04.2019).