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

Лабораторная работа №1  
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**

Изучить различные методы визуализации данных [1].

**2. Задание**

* Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов [здесь.](https://github.com/ugapanyuk/courses_current/wiki/DSLIST)

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

* Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

* Сформировать отчет и разместить его в своем репозитории на github.

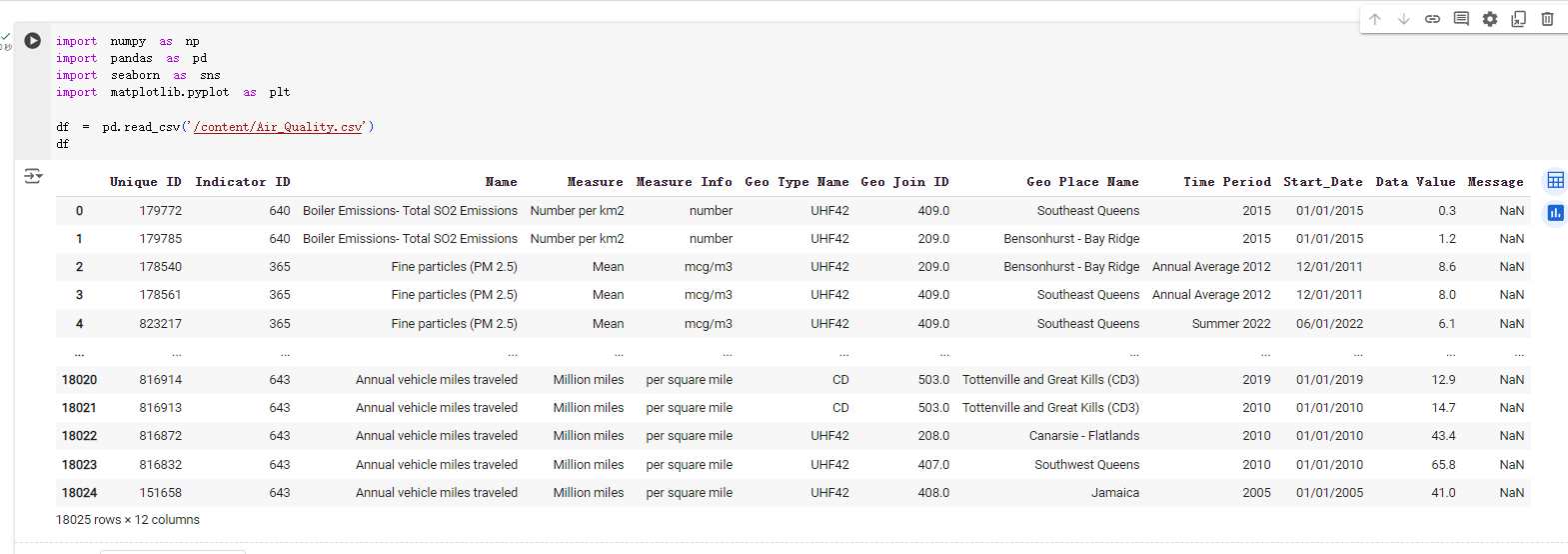
**3. Ход выполнения работы****3.1 Масштабирование признаков**

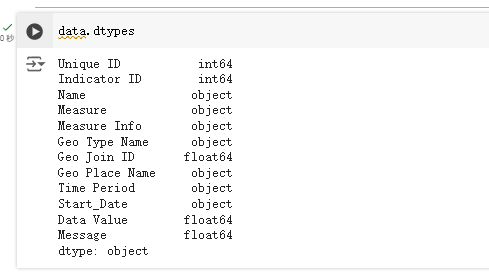
В этой тетради я буду использовать графики для визуализации взаимосвязи между переменными в наборе данных "Качество воздуха из Нью-Йорка".

О наборе данных:

Набор данных содержит информацию о данных наблюдения за качеством воздуха в Нью-Йорке.

Загрязнение воздуха является одной из наиболее важных экологических угроз для городского населения, и, хотя воздействию подвергаются все люди, выбросы загрязняющих веществ, уровни воздействия и уязвимость населения различаются в зависимости от района.

**3.2. Основные характеристики набора данных**  




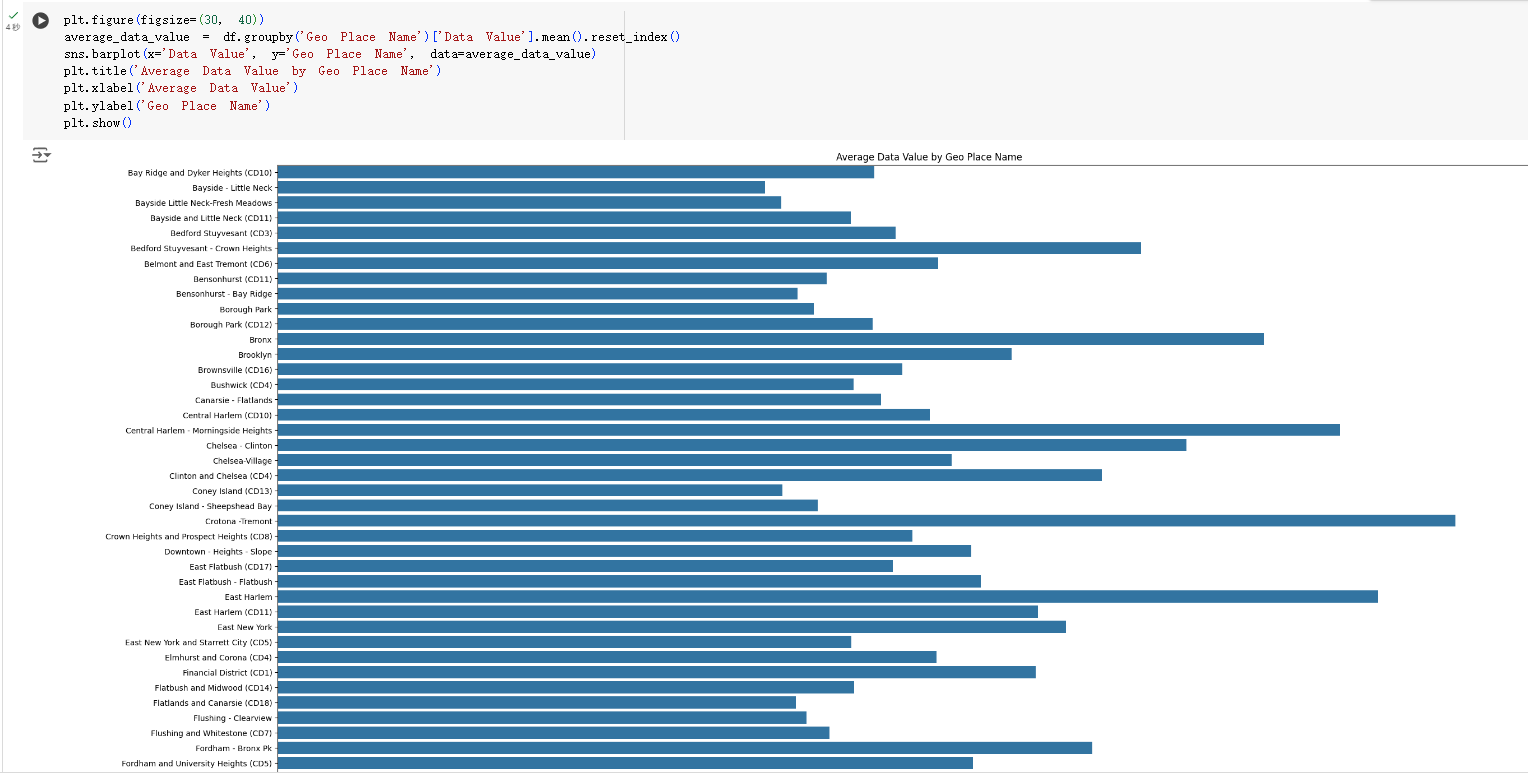


Рис 1. -Среднее значение данных по географическим названиям

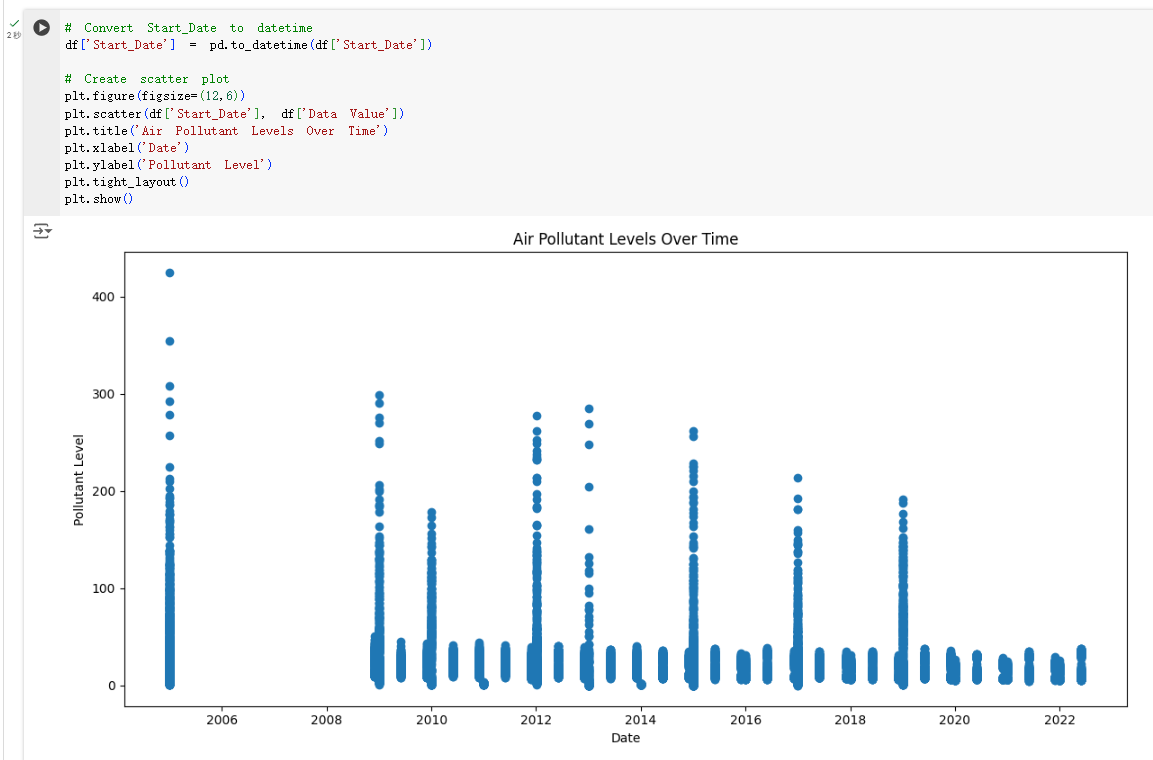


Рис 2. - Уровни загрязнителей воздуха с течением времени

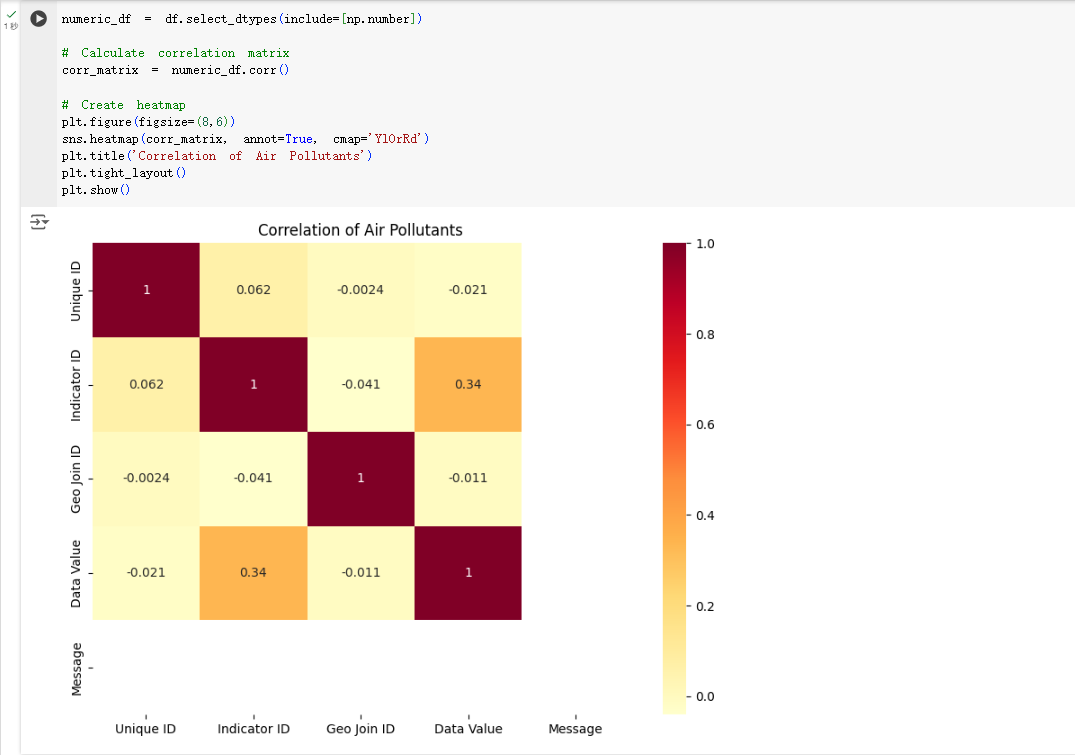
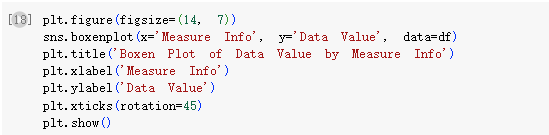


Рис 3. - Корреляция загрязнителей воздуха



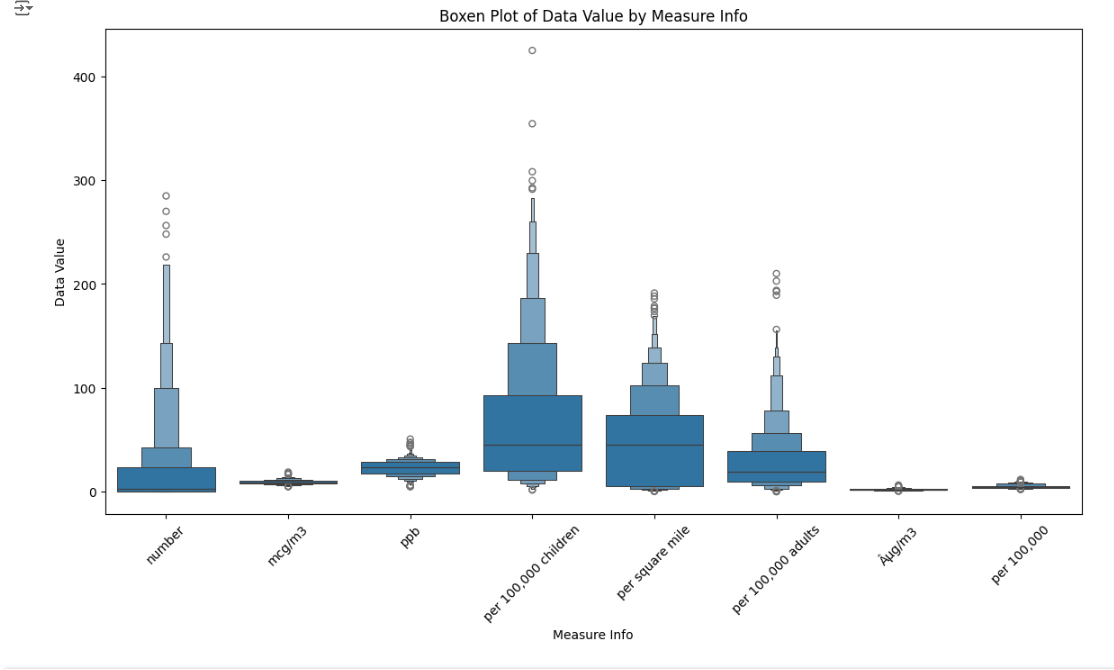


Рис 4. - График Боксена значения данных по мере информации

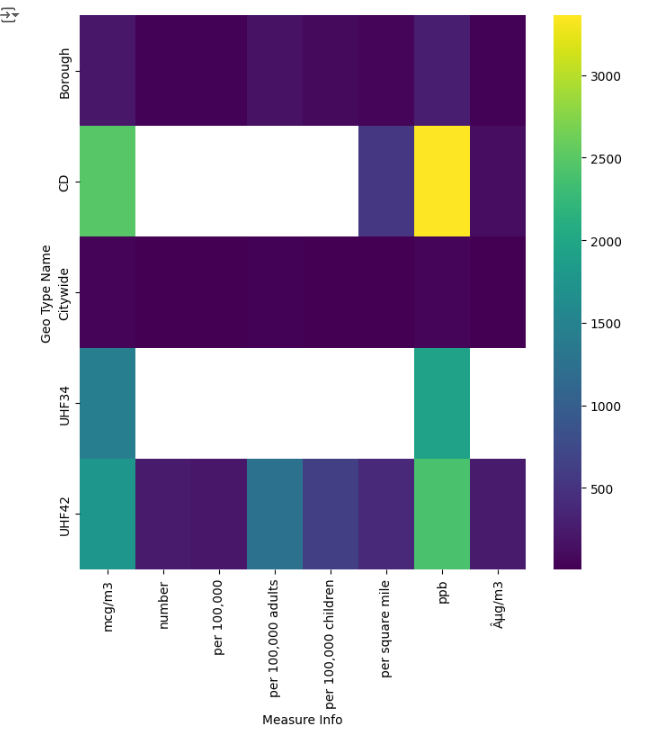
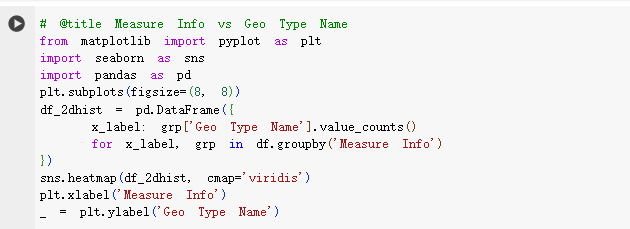


Рис 5. - Измерение информации по названию типа GEO

**Список литературы**

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование  
и визуализация данных» [Электронный ресурс] // GitHub. –– 2019. –– Режим доступа: https://github.com/ugapanyuk/ml\_course/wiki/LAB\_EDA\_VISUALIZATION (дата  
обращения: 13.02.2019)

[2] <https://www.kaggle.com/datasets>