Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по лабораторной работе №1

«Разведочный анализ данных. Исследование и визуализация данных»

*Вариант 12*

Студент:

Крюков Г. М.

Группа ИУ5-61Б

Преподаватель:

Гапанюк Ю. Е.

Москва, 2020 г.

**Цель лабораторной работы:**

Изучение различных методов визуализации данных.

Краткое описание: Построение основных графиков, входящих в этап разведочного анализа данных.

**Задание:**

* Выбрать набор данных (датасет).

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

* Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

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

**Выполнение работы:**

1. Текстовое описание выбранного набора данных

В качестве набора данных мы будем использовать набор данных, задачей которого является попытка предсказывания качества винной продукции.

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

Решение этой задачи может быть актуально для любителей и эстетов в сфере винопотребления.

Датасет состоит из 1 файла: winequality-red.csv

Каждый файл содержит следующие колонки:

"fixed acidity" - фиксированная кислотность

"volatile acidity" – летучая кислотность

"citric acid" – лимонная кислота

"residual sugar" – остаточный сахар

"chlorides" – хлориды

"free sulfur dioxide" – доля свободного диоксида серы

"total sulfur dioxide" – общая доля диоксида серы

"density" – плотность

"pH" – индекс кислотности

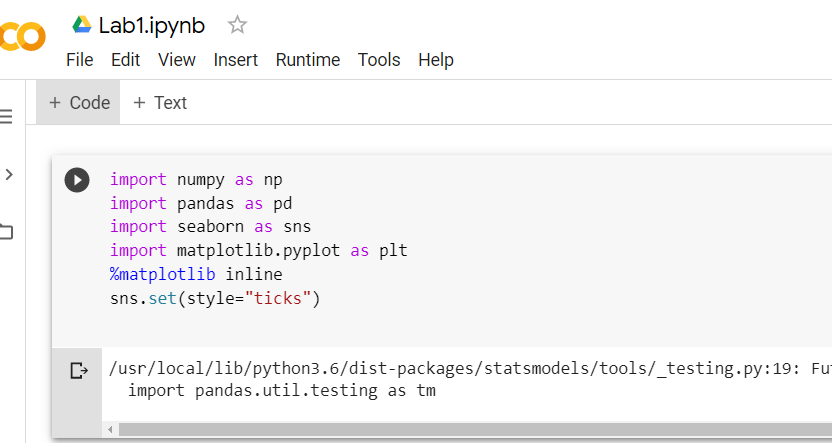
"sulphates" – сульфаты

"alcohol" – доля алкоголя

“quality” – общее качество напитка (от 0 до 10)

1. Основные характеристики датасета

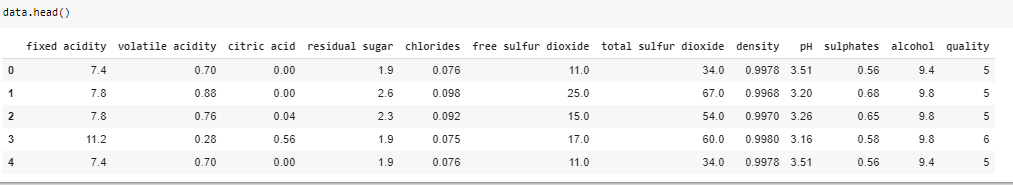
Импорт библиотек



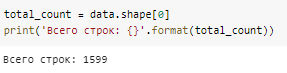
Загрузка данных

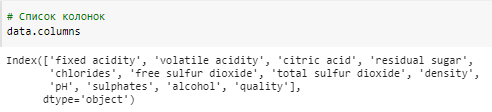


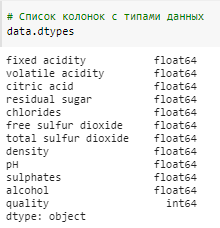
Основные характеристики датасета

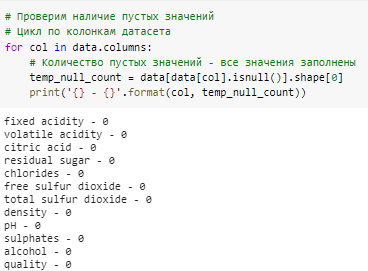


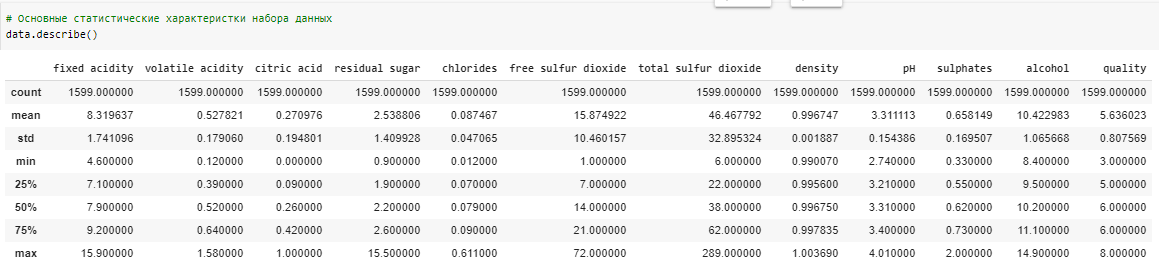


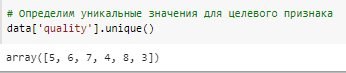










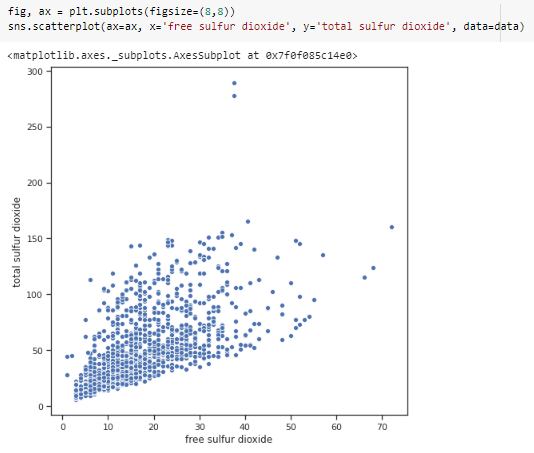


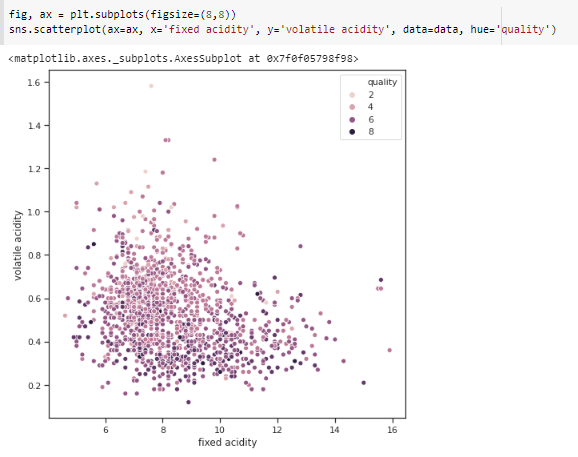
1. Визуальное исследование датасета

Для визуального исследования могут быть использованы различные виды диаграмм, мы построим только некоторые варианты диаграмм, которые используются достаточно часто.

Диаграмма рассеяния

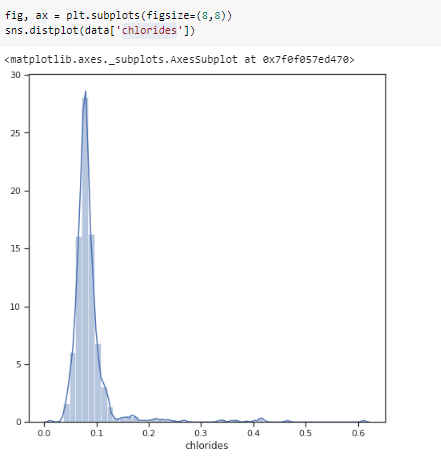
Позволяет построить распределение двух колонок данных и визуально обнаружить наличие зависимости. Не предполагается, что значения упорядочены (например, по времени).





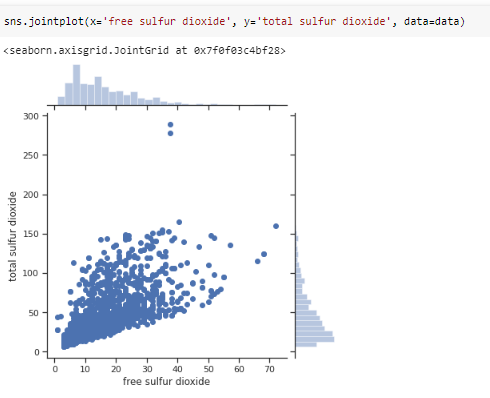
Гистограмма

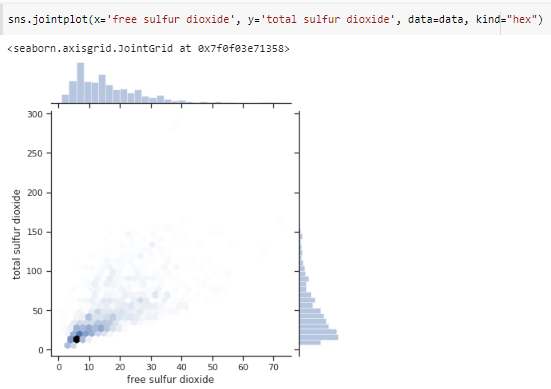
Позволяет оценить плотность вероятности распределения данных.



Jointplot

Комбинация гистограмм и диаграмм рассеивания.

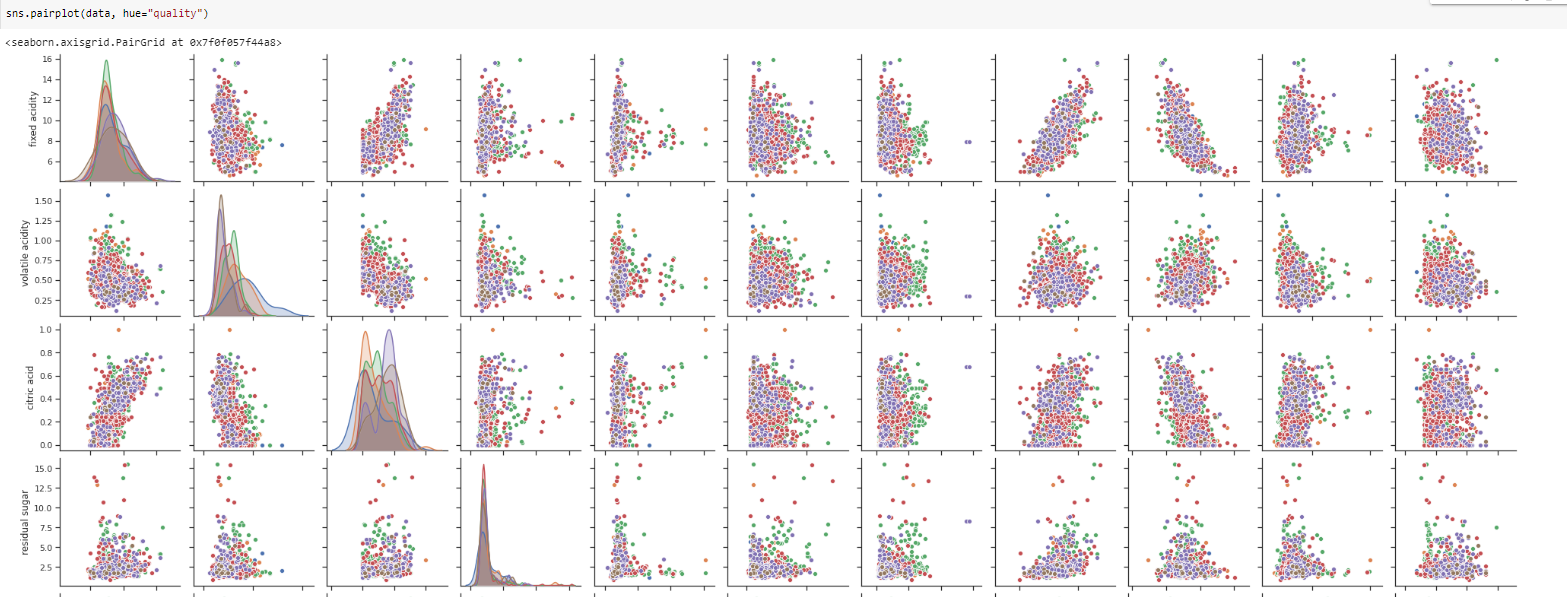


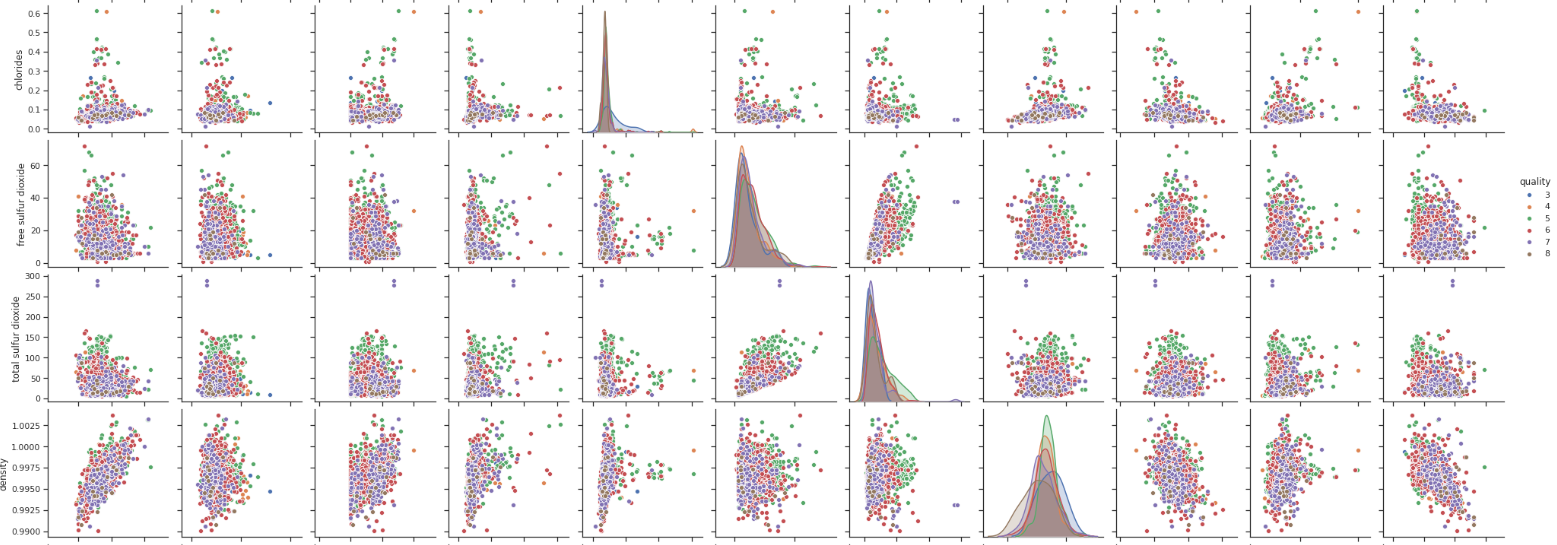


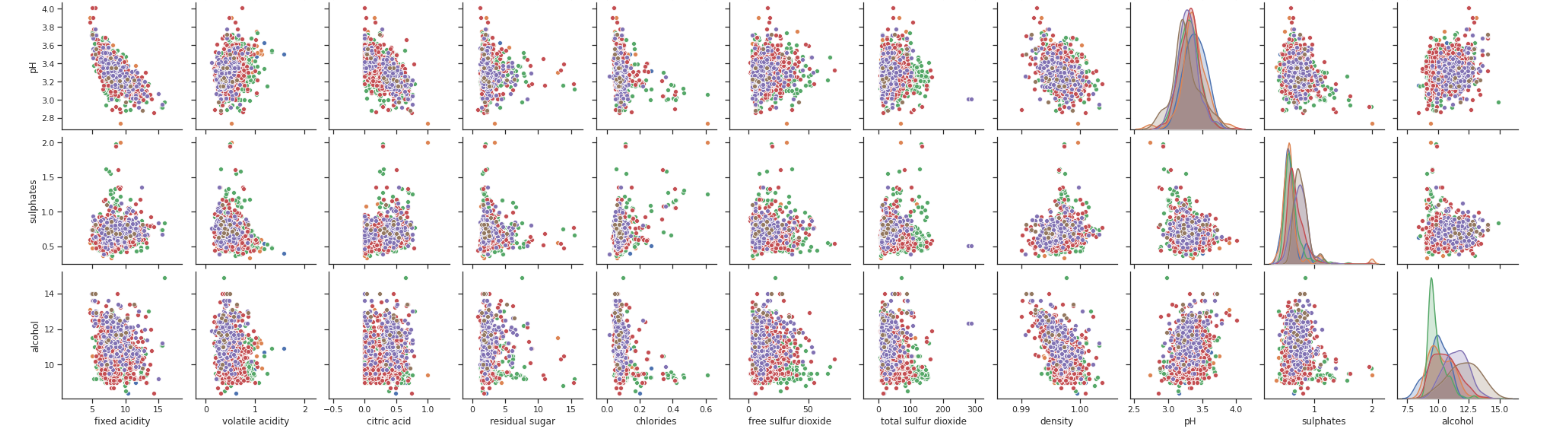
"Парные диаграммы"

Комбинация гистограмм и диаграмм рассеивания для всего набора данных.

Выводится матрица графиков. На пересечении строки и столбца, которые соответствуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.

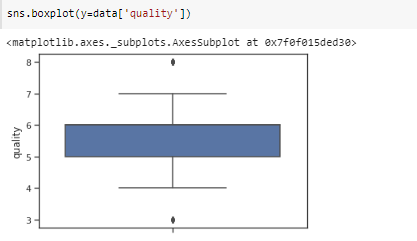


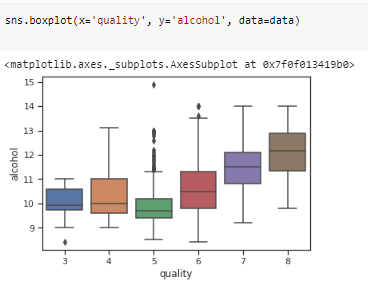




Ящик с усами

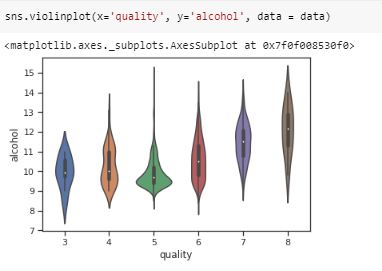
Отображает одномерное распределение вероятности.





Violin plot

Похоже на предыдущую диаграмму, но по краям отображаются распределения плотности.

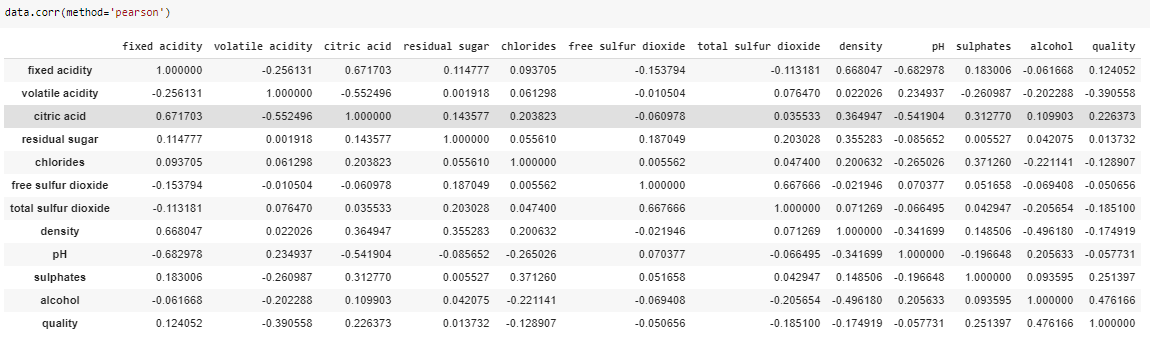


1. Информация о корреляции признаков

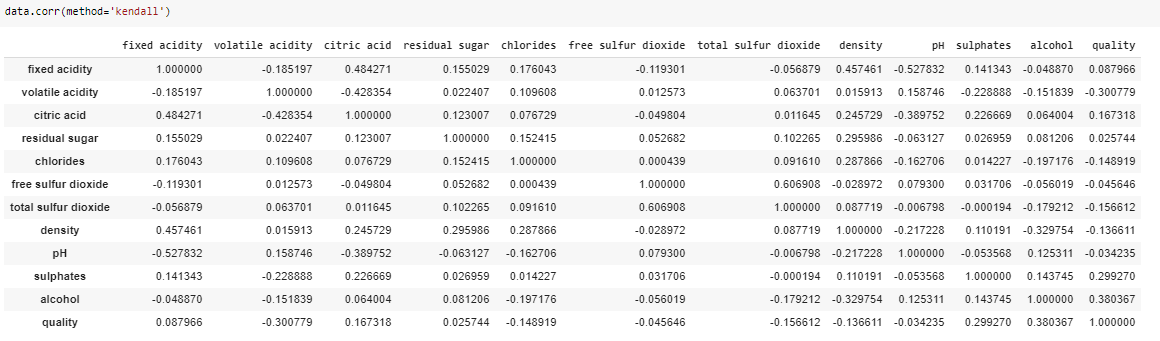
Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (в нашем примере примере нет целевого признака) Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

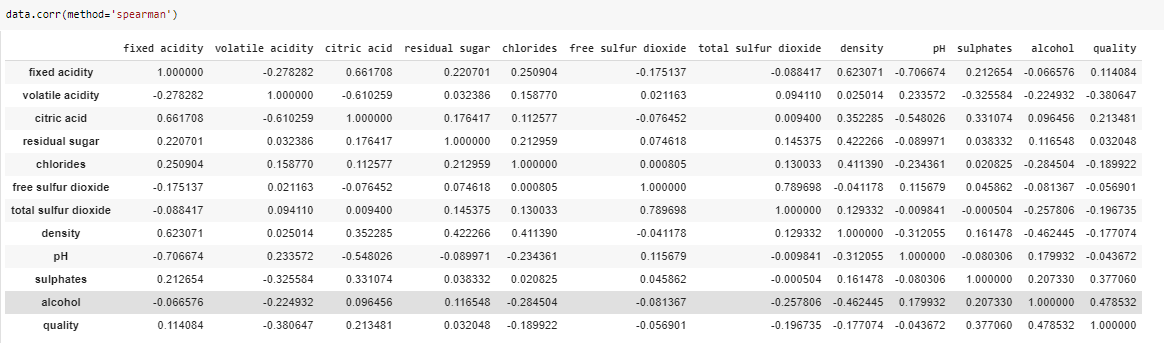
Корреляционная матрица Пирсона:



Корреляционная матрица Кендалла:



Корреляционная матрица Спирмана:



Можем видеть, что значения во всех трёх матрицах примерно равны в соответствующих ячейках.

Для визуализации корреляционной матрицы будем использовать "тепловую карту" heatmap, которая показывает степень корреляции различными цветами.

