Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра автоматизации обработки информации (АОИ)

Вариант 2

Отчёт к Лабораторной работе №1

по дисциплине «Нейронные сети и их применение»

Выполнил:

Студент группы 422–M1

\_\_\_\_Белоус Г.В.

Принял:

К.т.н., Доцент кафедры АОИ

\_\_\_\_Аксёнов С.В.

Оглавление

[1 Введение 3](#_Toc130308366)

[1.1 Цель лабораторной работы 3](#_Toc130308367)

[1.2 Задание на лабораторную работу 3](#_Toc130308368)

[2 Ход выполнения работы 5](#_Toc130308369)

[2.1 Получение описания набора данных и список атрибутов, получение числа пропущенных значений в для каждого атрибута и их обработка 5](#_Toc130308370)

[2.2 Построение pairplot для набора данных 7](#_Toc130308371)

[2.3 Отображение 2-х признаков с неким распределением с помощью гистограмм hist и kdeplot, jointplot 8](#_Toc130308372)

[2.4 Выбор части признаков датасета и отображение корреляцию между ними в виде тепловой карты (heatmap) 10](#_Toc130308373)

[2.5 Выбор 3 признаков, целевую переменную и отобразить для них графики размаха 11](#_Toc130308374)

[2.6 Визуализиция некоторой статистики, для разных атрибутов с использованием violinplot, stripplot, swarmplot, pie. 12](#_Toc130308375)

[3 Вывод 15](#_Toc130308376)

# Введение

## 1.1 Цель лабораторной работы

Исследовательский анализ данных и визуализация c помощью библиотек matplotlib, sns и plotly.

## **1.2 Задание на лабораторную работу**

Написать программу на Python, которая загружает набор данных по вашему варианту и производит исследовательский анализ этих данных и визуализирует ряд зависимостей между признаками в нижеперечисленных вариациях с помощью библиотек matplotlib, sns и plotly.

1. Получить описание набора данных и список атрибутов, получить число пропущенных значений в для каждого атрибута.

2. Обработать пропущенные значения (удалить строки/ удалить атрибуты/ выполнить импутацию значений)

3. Построить pairplot для набора данных

4. Выбрать не менее 2-х признаков с неким распределением значений, и отобразить это распределение с помощью гистограмм hist и kdeplot, jointplot

5. Выбрать часть признаков датасета и отобразить корреляцию между ними в виде тепловой карты (heatmap).

6. Выбрать 3 признака (имеющие некое распределение значений), целевую переменную (категориальную) и отобразить для них графики размаха (ящик с усами, box plot).

8. Визуализировать некоторые статистики, для разных атрибутов с использованием следующих инструментов: violinplot, stripplot, swarmplot, pie.

Напишите короткое заключение о наиболее интересных зависимостях, которые Вы обнаружили в данных.

Содержание отчета.

1. Титульный лист
2. Задача по полученному варианту (выбор варианта осуществляется по последнему номеру в зачетной книжке, если он равен 0 или 5 – то берем вариант 1, если 2 или 6 – вариант 2, если 3 или 7 то вариант 3, если 4 или 8 - вариант 4, если 5 – вариант 5, если 9 – вариант 6)
3. Описание датасета
4. Графики, полученные в ходе анализа набора данных
5. Вывод

# Ход выполнения работы

## Получение описания набора данных и список атрибутов, получение числа пропущенных значений в для каждого атрибута и их обработка

Набор данных о вине был загружен с помощью библиотеки pandas и описание датасета и атрибутов были получены с помощью методов describe() и info(), результаты показаны на рисунке 2.1 и 2.2:

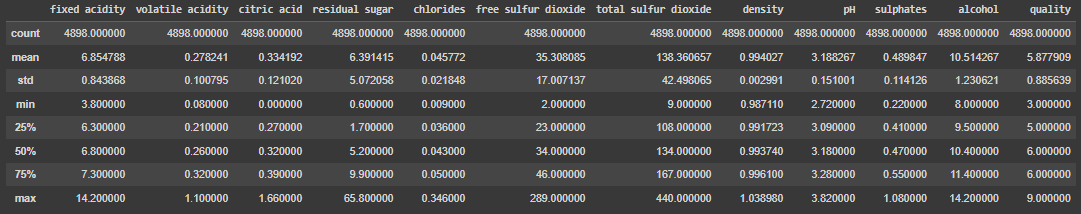


Рисунок 2.1 – Описание набора данных

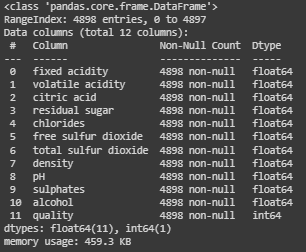


Рисунок 2.2 – Список атрибутов

Запрос о числе пропущенных значений осуществлялся с помощью метода snull().sum(), результат выполнения показан на рисунке 2.3:

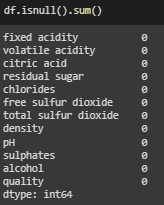


Рисунок 2.3 – Число пропущенных значений в датасете

Так как в наборе данных не было обнаружено пропущенные значений, пункт с обработкой пропущенных значений и их импутацией был пропущен.

## Построение pairplot для набора данных

На рисунке 2.5 показан pairplotпостроенный для всех атрибутов датасета:

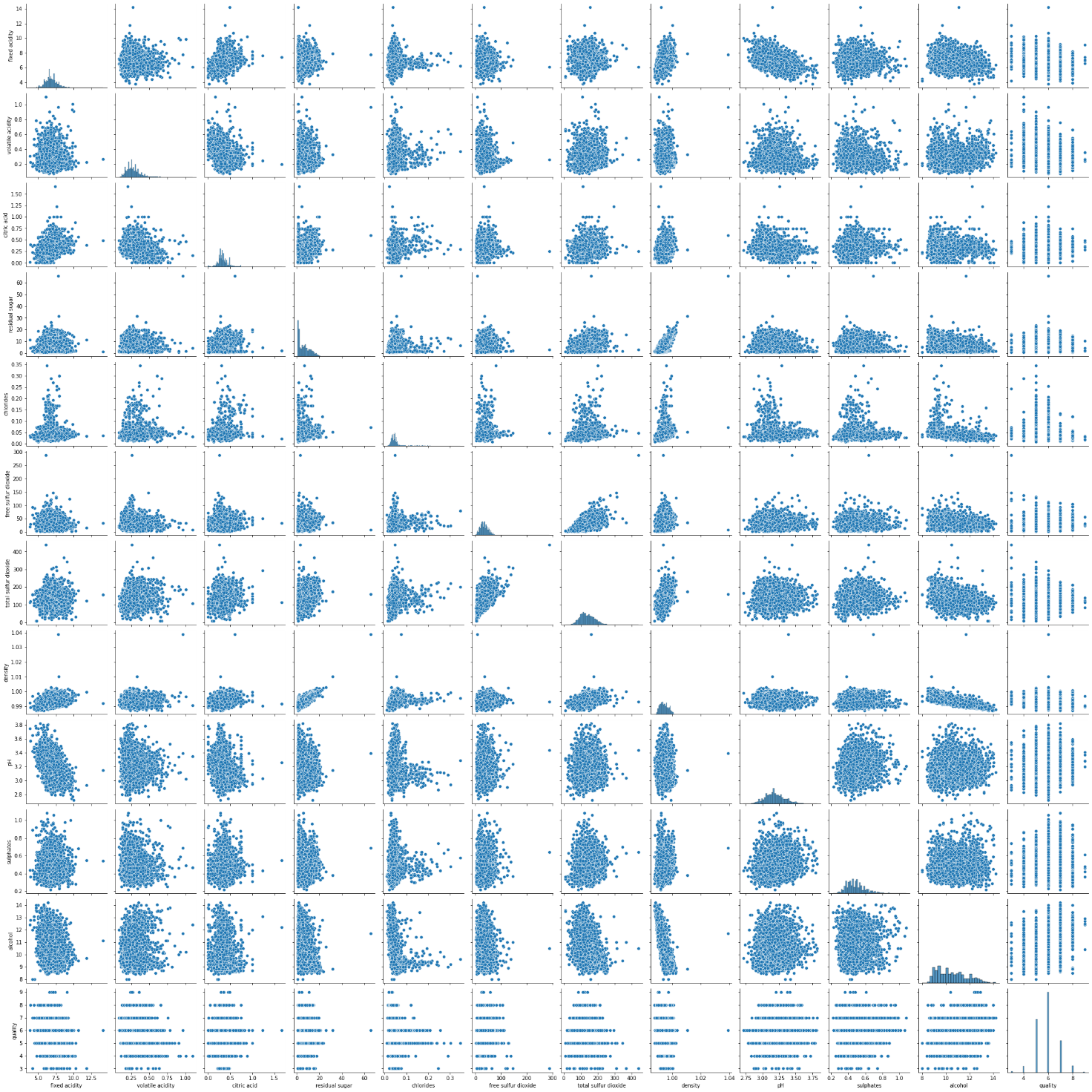


Рисунок 2.4 – Построенный pairplot для всех атрибутов

## Отображение 2-х признаков с неким распределением с помощью гистограмм hist и kdeplot, jointplot

На рисунке 2.5 продемонстрирована гистограмма влияния летучей кислотности на оценку вина:

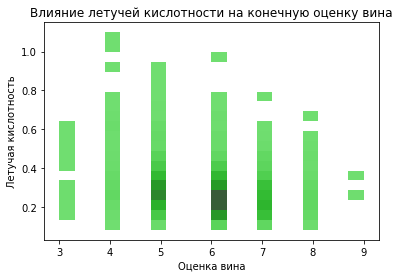


Рисунок 2.5 – Отображение признаков volatile acidity и quality с помощью hist

Так как отображение признаков volatile acidity и quality с помощью kdeplot не несёт информативности, из-за того, что quality является целочисленным признаком, для последующих гистограмм вместо него был взят признак alcohol, полученные гистограммы показаны на рисунках 2.6, 2.7:

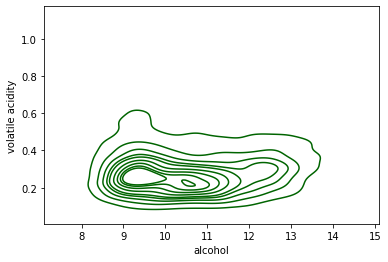


Рисунок 2.6 – Отображение признаков volatile acidity и alcohol с помощью kdeplot

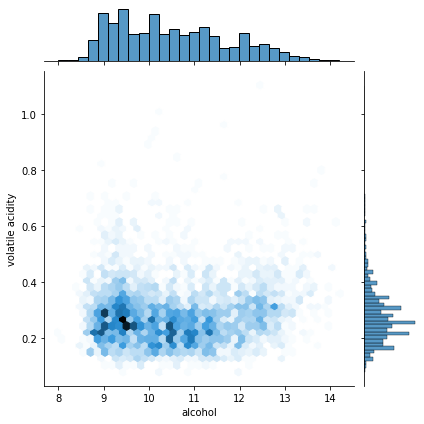


Рисунок 2.7 - Отображение признаков volatile acidity и alcohol с помощью jointplot

## Выбор части признаков датасета и отображение корреляцию между ними в виде тепловой карты (heatmap)

Для построения тепловой карты были выбраны все признаки из набора данных, результат построения карты показан на рисунке 2.8, как видно самое низкое значение у пары alcohol и density, а самое высокое у density и residual sugar:

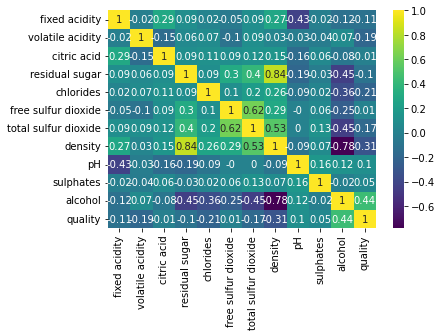


Рисунок 2.8 – Корреляция между всеми признаками датасета с помощью hitmap

## Выбор 3 признаков, целевую переменную и отобразить для них графики размаха

Для построения графика размаха были выбраны три признака (fixed acidity, alcohol и total sulfur dioxide) от целевой переменной quality, которая не отображается на рисунке 2.9, так как она категориальная:

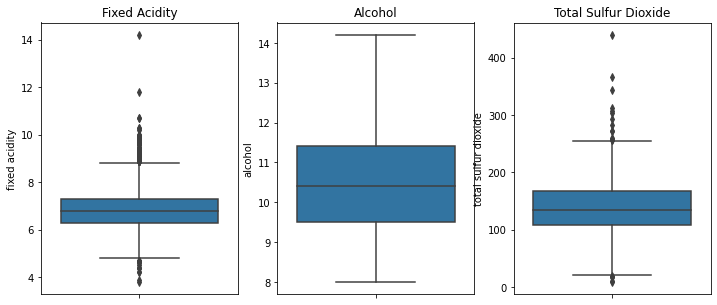


Рисунок 2.9 – Графики размаха для признаков fixed acidity, alcohol и total sulfur dioxide от признака quality

## Визуализиция некоторой статистики, для разных атрибутов с использованием violinplot, stripplot, swarmplot, pie.

**Violinplot** - это тип графика, который отображает распределение данных по заданному атрибуту. Каждый violinplot состоит из двух половинок, где ширина каждой половинки отражает частоту появления значений в данных. На рисунке 2.10 показан график для признака alcohol:

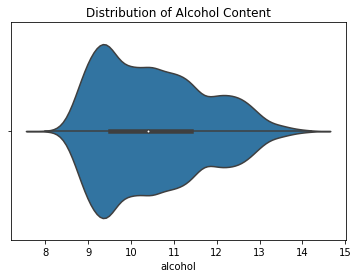


Рисунок 2.10 – График violinplot для признака alcohol

**Stripplot** - это тип графика, который отображает каждую точку данных на графике. Это особенно полезно, когда у вас есть небольшой набор данных. На рисунке 2.11 показан график для признака total sulfur dioxide:



Рисунок 2.11 – График stripplot для признака total sulfur dioxide

**Swarmplot** - это тип графика, который отображает каждую точку данных на графике, разделяя их таким образом, чтобы они не перекрывали друг друга. Так как данный тип графика отображается каждую точку, в результате выполнения кода по построению данного графика было получено предупреждение о том, что 2/3 точек не могут быть размещены на графике, а полученный рисунок 2.12 мало информативен:

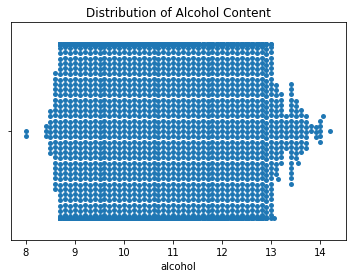


Рисунок 2.12 – График swarmplot для признака alcohol

**Pie chart** - это круговая диаграмма, которая показывает, как данные в наборе распределяются по категориям. Пример графика на категориальном признаке quality продемонстрирован на рисунке 2.13:

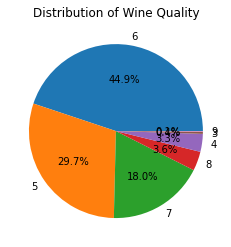


Рисунок 2.13 – График pie chart для признака quality

# Вывод

В ходе выполнения лабораторной работы был сделан исследовательский анализ данных и визуализация c помощью библиотек matplotlib, sns и plotly. В отчёте представлены все полученные графики.

Интересная зависимость представлена на рисунке 2.5. Видно, как летучая кислотность влияет на конечную оценку вина, чем меньше кислотности, тем лучше оценки вина.

Так же при построений swarmplot было получено предупреждение о невозможности разместить все точки на графике – это значит, что данный тип графика наиболее подходит для малого набора данных.