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

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Создание "истории о данных" (Data Storytelling)»

Выполнил:
студент группы ИУ5И-22М
Лу Жуньда

Москва — 2024 г.

1. Цель лабораторной работы

Изучение различных методов визуализация данных и создание истории на основе данных.

2. Задание

• Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

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.

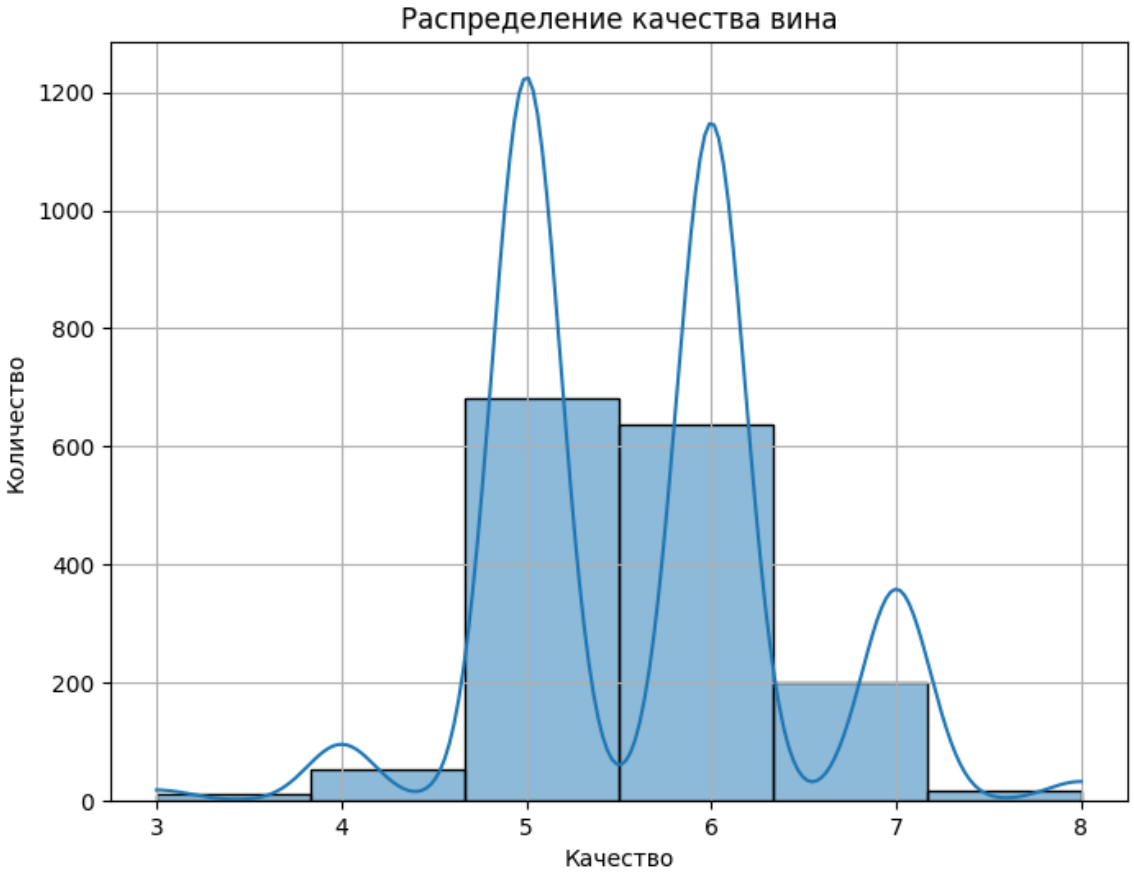
- Сформировать отчет и разместить его в своем репозитории на github.

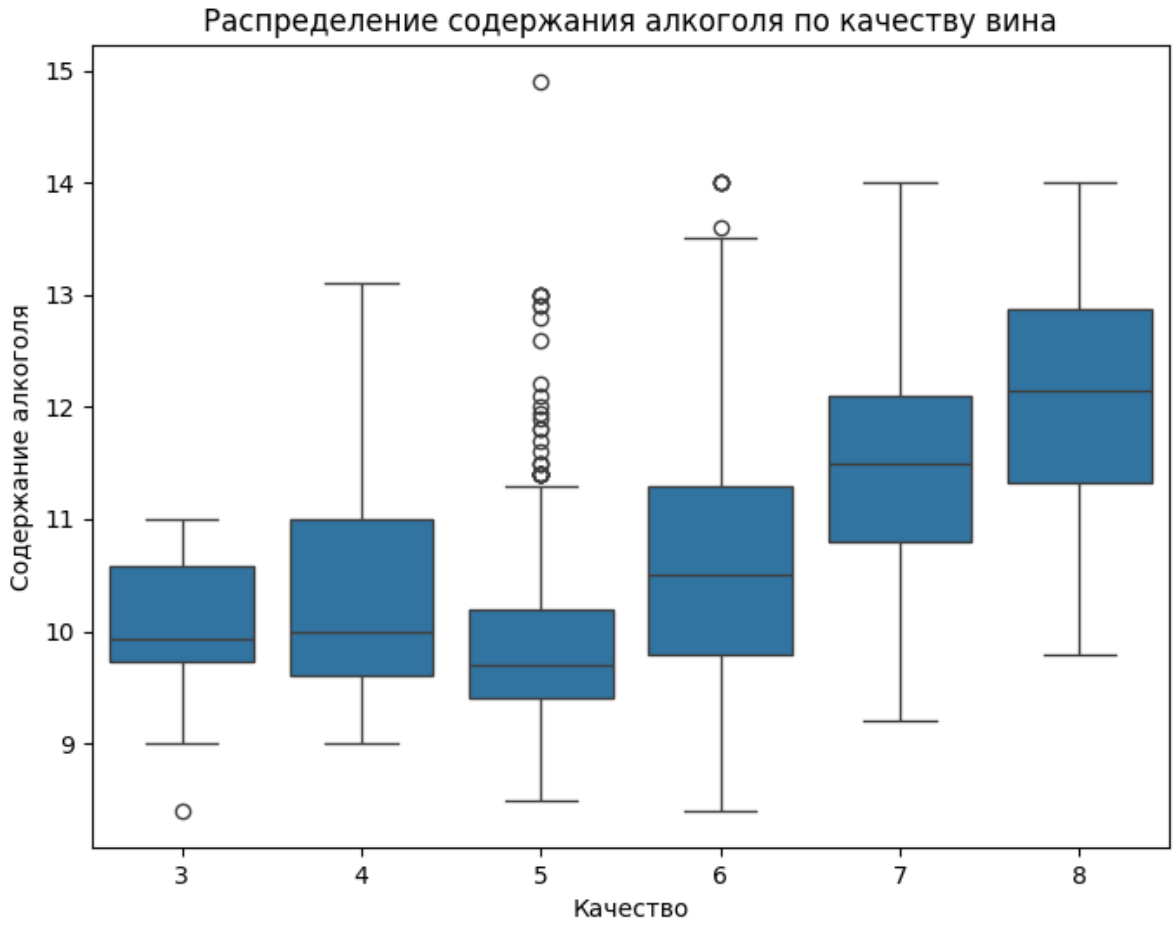
3. Текст программы

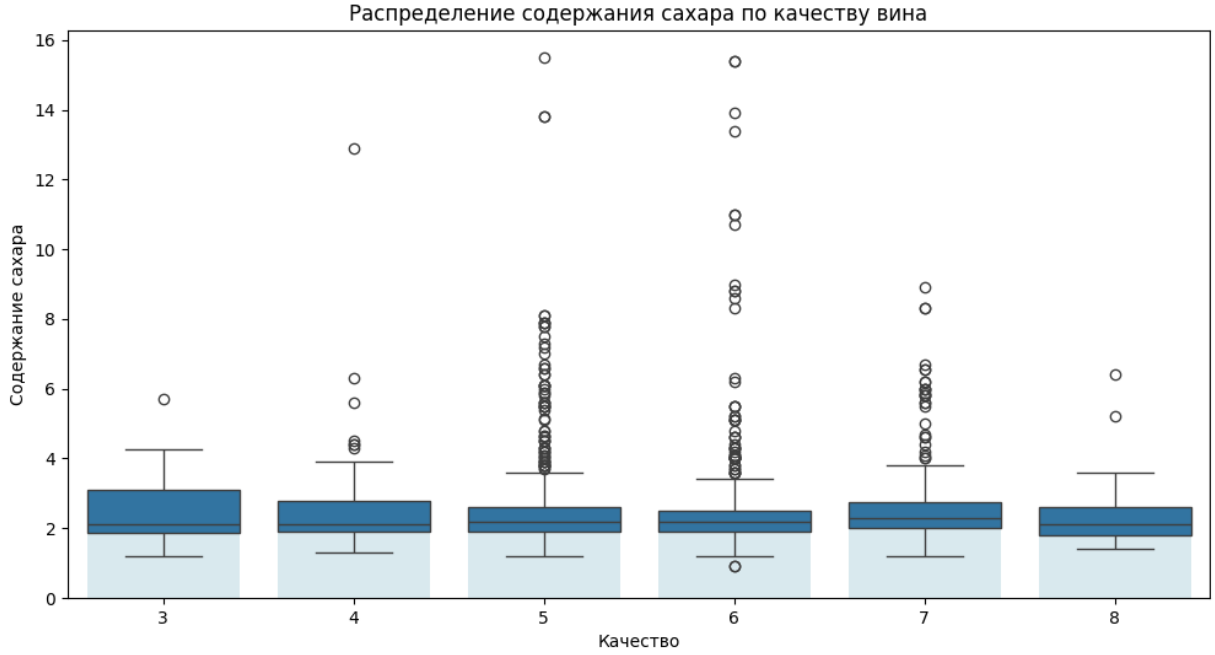
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Загрузим данные
data = pd.read_csv("winequality-red-folds.csv")

# Посмотрим на первые несколько строк датасета
print(data.head())

# Посмотрим на информацию о датасете
print(data.info())

# Посмотрим на статистику датасета
print(data.describe())

# Построим гистограмму распределения качества вина
plt.figure(figsize=(8, 6))
sns.histplot(data['quality'], bins=6, kde=True)
plt.title('Распределение качества вина')
plt.xlabel('Качество')
plt.ylabel('Количество')
plt.grid(True)
plt.show()

# Построим тепловую карту корреляции
plt.figure(figsize=(10, 8))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Тепловая карта корреляции признаков')
plt.show()

# Построим ящик с усами для распределения алкоголя
plt.figure(figsize=(8, 6))
sns.boxplot(x='quality', y='alcohol', data=data)
```

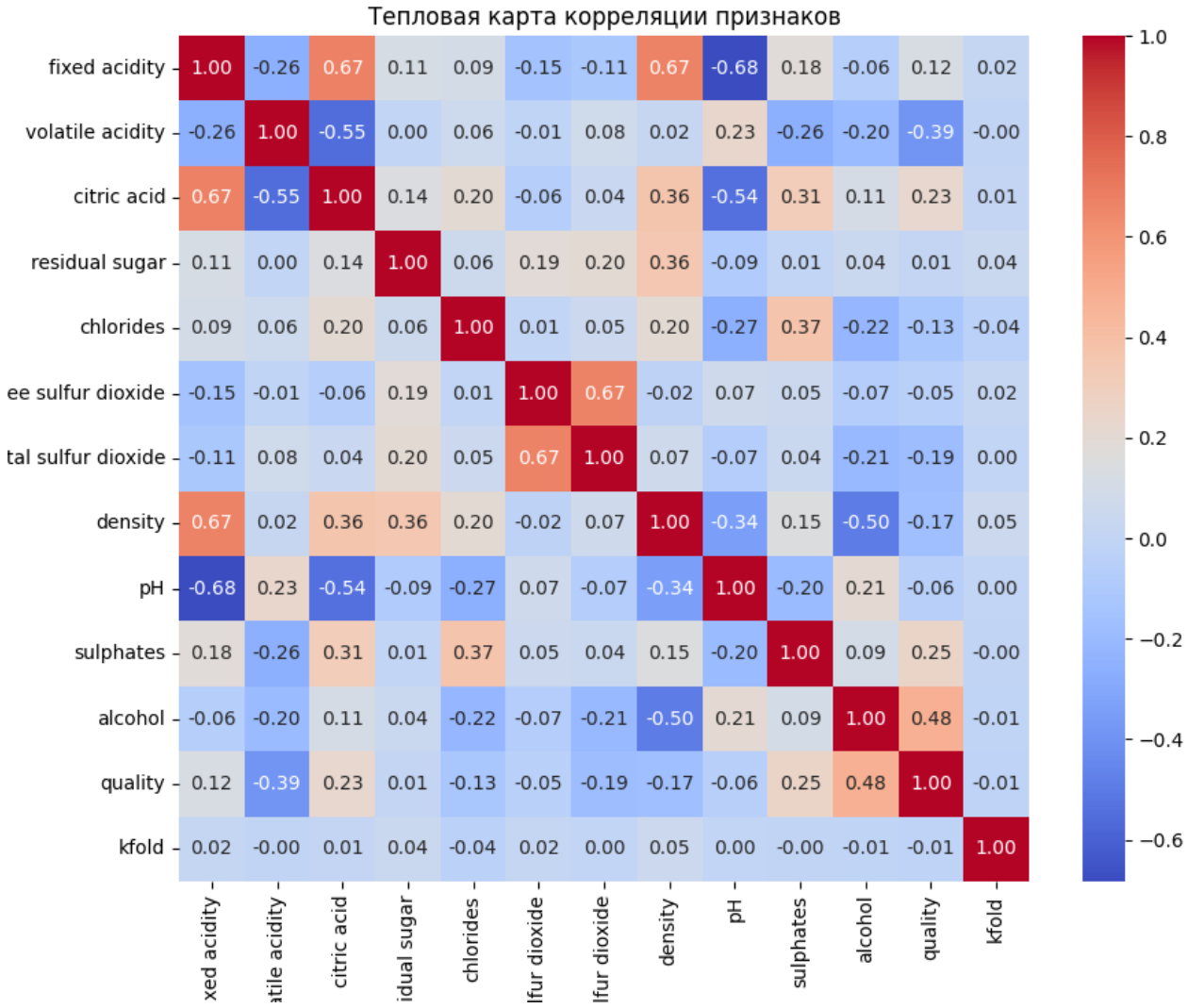
```
plt.title('Распределение содержания алкоголя по качеству вина')
plt.xlabel('Качество')
plt.ylabel('Содержание алкоголя')
plt.show()

# Построим точечную диаграмму для сравнения pH и качества вина
plt.figure(figsize=(8, 6))
sns.scatterplot(x='pH', y='quality', data=data)
plt.title('Сравнение pH и качества вина')
plt.xlabel('pH')
plt.ylabel('Качество')
plt.show()

# Построим полосчатую диаграмму для сравнения содержания алкоголя по качеству
вина
plt.figure(figsize=(8, 6))
sns.barplot(x='quality', y='alcohol', data=data, ci=None)
plt.title('Среднее содержание алкоголя по качеству вина')
plt.xlabel('Качество')
plt.ylabel('Среднее содержание алкоголя')
plt.show()
```

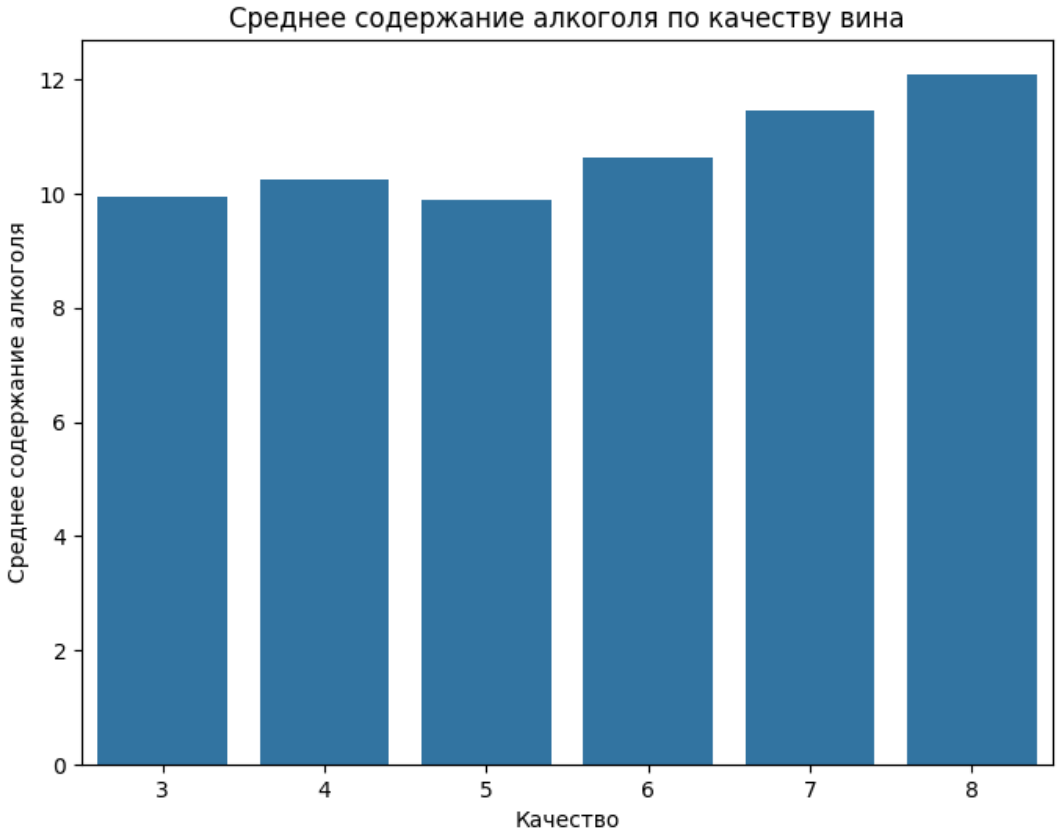
4. Экранные формы с примерами выполнения программы




Распределение содержания алкоголя по качеству вина
Содержание алкоголя
Качество
15
14
13
12
11
10
9
3
4
5
6
7
8


Распределение содержания сахара по качеству вина
Содержание сахара
Качество
16
14
12
10
8
6
4
2
0
3
4
5
6
7
8


Тепловая карта корреляции признаков
fixed acidity 1.00 -0.26 0.67 0.11 0.09 -0.15 -0.11 0.67 -0.68 0.18 -0.06 0.12 0.02
volatile acidity -0.26 1.00 -0.55 0.00 0.06 -0.01 0.08 0.02 0.23 -0.26 -0.20 -0.39 -0.00
citric acid 0.67 -0.55 1.00 0.14 0.20 -0.06 0.04 0.36 -0.54 0.31 0.11 0.23 0.01
residual sugar 0.11 0.00 0.14 1.00 0.06 0.19 0.20 0.36 -0.09 0.01 0.04 0.01 0.04
chlorides 0.09 0.06 0.20 0.06 1.00 0.01 0.05 0.20 -0.27 0.37 -0.22 -0.13 -0.04
ee sulfur dioxide -0.15 -0.01 -0.06 0.19 0.01 1.00 0.67 -0.02 0.07 0.05 -0.07 -0.05 0.02
tal sulfur dioxide -0.11 0.08 0.04 0.20 0.05 0.67 1.00 0.07 -0.07 0.04 -0.21 -0.19 0.00
density 0.67 0.02 0.36 0.36 0.20 -0.02 0.07 1.00 -0.34 0.15 -0.50 -0.17 0.05
pH -0.68 0.23 -0.54 -0.09 -0.27 0.07 -0.07 -0.34 1.00 -0.20 0.21 -0.06 0.00
sulphates 0.18 -0.26 0.31 0.01 0.37 0.05 0.04 0.15 -0.20 1.00 0.09 0.25 -0.00
alcohol -0.06 -0.20 0.11 0.04 -0.22 -0.07 -0.21 -0.50 0.21 0.09 1.00 0.48 -0.01
quality 0.12 -0.39 0.23 0.01 -0.13 -0.05 -0.19 -0.17 -0.06 0.25 0.48 1.00 -0.01
kfold 0.02 -0.00 0.01 0.04 -0.04 0.02 0.00 0.05 0.00 -0.00 -0.01 -0.01 1.00
xed acidity
atile acidity
citric acid
idual sugar
chlorides
lfur dioxide
lfur dioxide
density
pH
sulphates
alcohol
quality
kfold
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Среднее содержание алкоголя по качеству вина
Среднее содержание алкоголя
0
2
4
6
8
10
12
3
4
5
6
7
8
Качество

Список литературы

[1] Гапанюк Ю. Е. LAB_MMO__DATA_STORYЛабораторная работа №1Создание "истории о данных" (Data Storytelling)// GitHub. — 2024. — Режим доступа:https://github.com/ugapanyuk/courses_current/wiki/LAB_MMO__DATA_STORY#%D0%BB%D0%B0%D0%B1%D0%BE%D1%80%D0%B0%D1%82%D0%BE%D1%80%D0%BD%D0%B0%D1%8F-%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%B0-1

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