****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.](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:

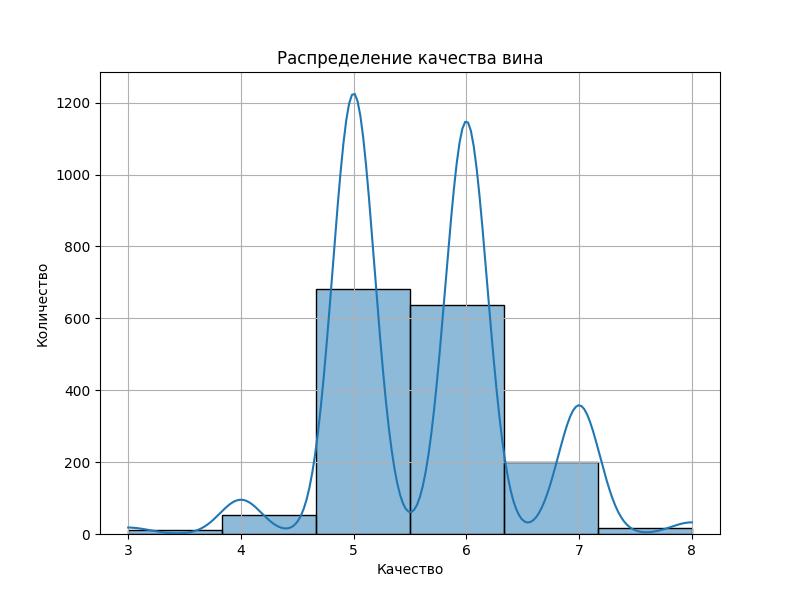
1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется 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.

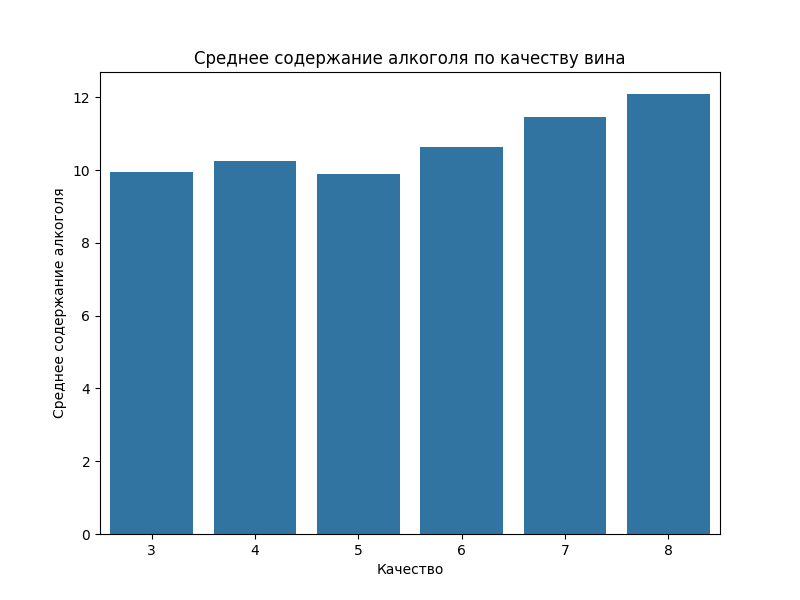
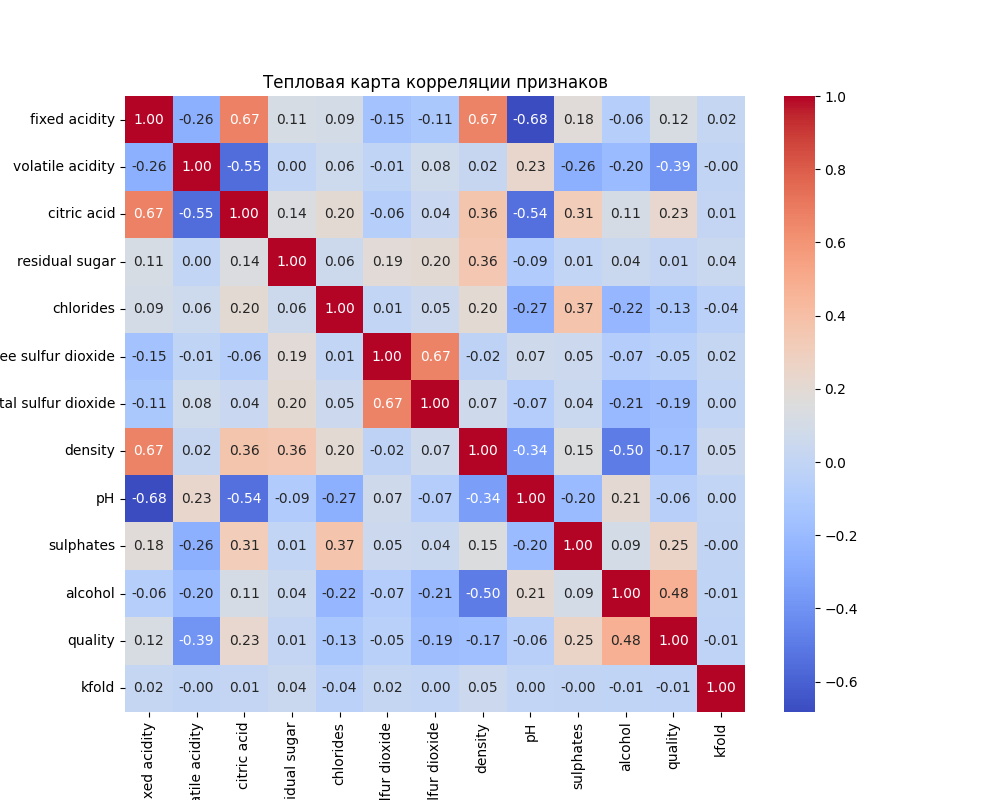
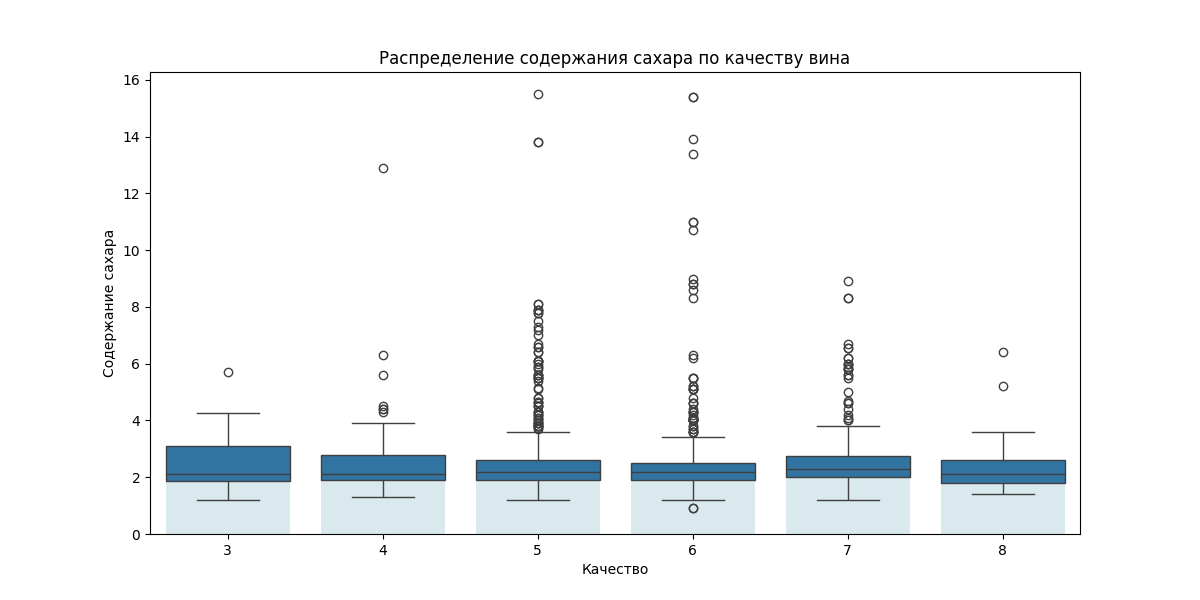
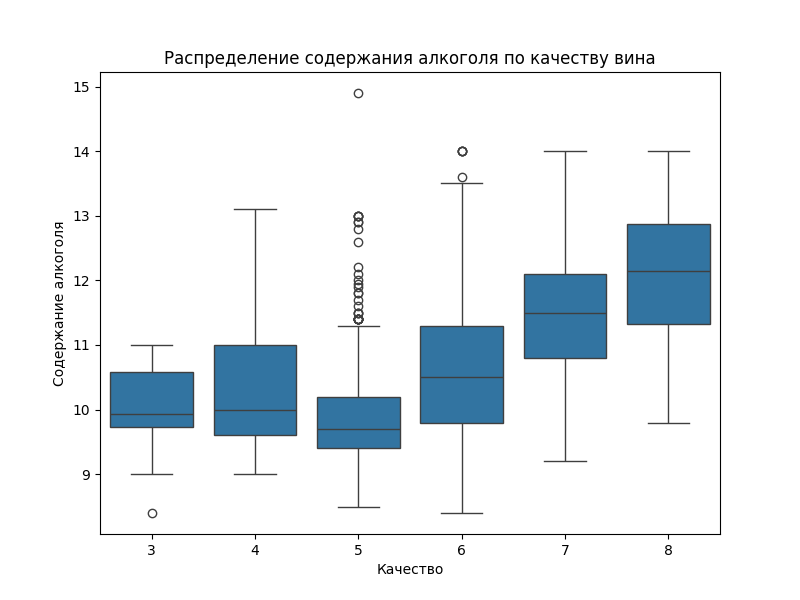
* Сформировать отчет и разместить его в своем репозитории на github.

**3. 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**





**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>