Министерство образования Республики Беларусь
Учреждение образования
«Брестский Государственный технический университет»
Кафедра ИИТ

Лабораторная работа №1
По дисциплине «Основы машинного обучения»

Выполнил:
Студент 3 курса
Группы АС-66
Занько Я.С.
Проверил:
Крощенко А. А.

Брест 2025

Цель: получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Вариант 4

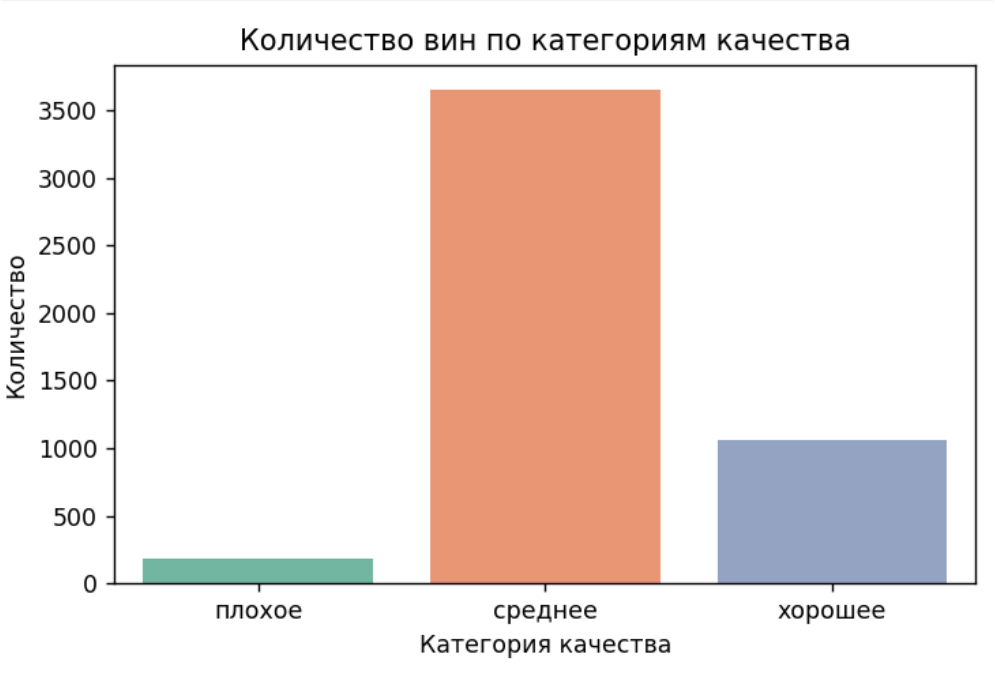
Задачи:

1. Загрузите данные и выведите информацию о типах столбцов.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
df = pd.read_csv('winequality-white.csv', sep=';')
print("Информация о данных:")
print(df.info())
```

2. Преобразуйте целевую переменную quality в категориальную: "плохое" (<=4), "среднее" (5-6), "хорошее" (>=7).

```
def quality_category(q):
    if q <= 4:
        return 'плохое'
    elif q <= 6:
        return 'среднее'
    else:
        return 'хорошее'

df['quality_label'] = df['quality'].apply(quality_category)
df['quality_label'] = pd.Categorical(df['quality_label'], categories=['плохое', 'среднее', 'хорошее'])
```

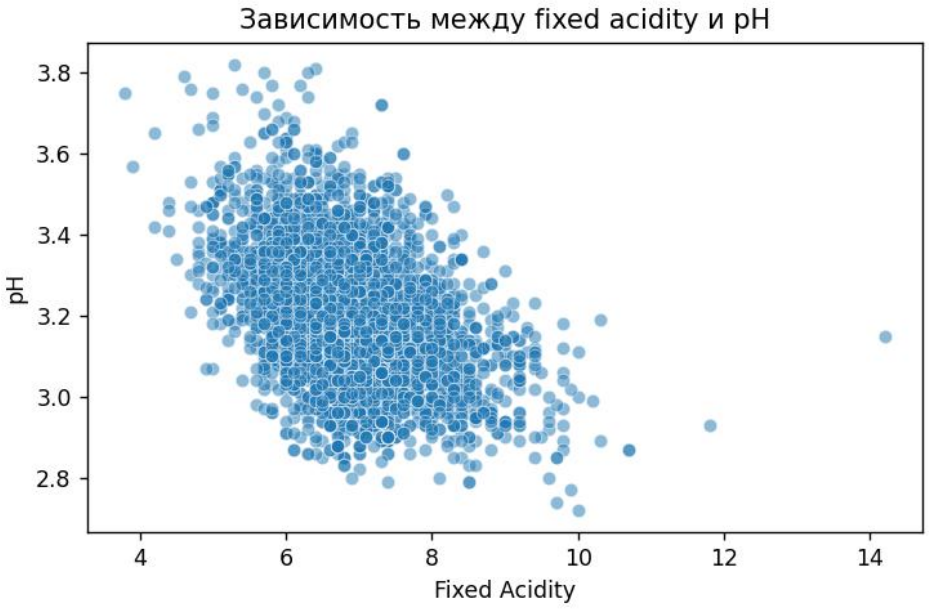


3. Постройте столбчатую диаграмму, показывающую количество вин каждой новой категории качества.

```
plt.figure(figsize=(6, 4))
sns.countplot(data=df, x='quality_label', palette='Set2')
plt.title('Количество вин по категориям качества')
plt.xlabel('Категория качества')
plt.ylabel('Количество')
plt.tight_layout()
plt.show()
```

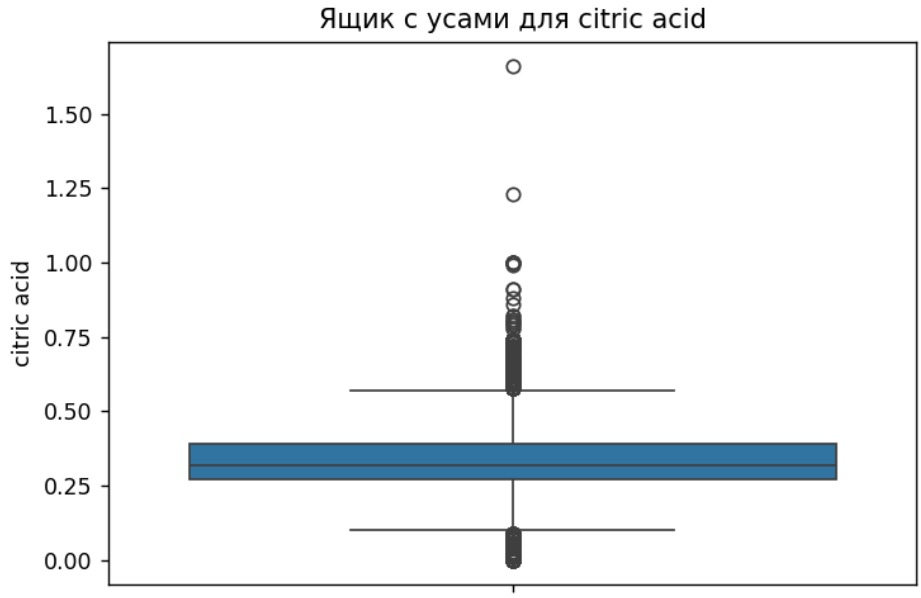
4. Проверьте корреляцию между fixed acidity и pH. Визуализируйте эту зависимость на диаграмме рассеяния.

```
correlation = df['fixed acidity'].corr(df['pH'])
print(f"\nКорреляция между fixed acidity и pH: {correlation:.2f}")

plt.figure(figsize=(6, 4))
sns.scatterplot(data=df, x='fixed acidity', y='pH', alpha=0.5)
plt.title('Зависимость между fixed acidity и pH')
plt.xlabel('Fixed Acidity')
plt.ylabel('pH')
plt.tight_layout()
plt.show()
```



5. Найдите признак с наибольшим количеством выбросов, используя "ящик с усами" (box plot).

```
def count_outliers(series):
    q1 = series.quantile(0.25)
    q3 = series.quantile(0.75)
    iqr = q3 - q1
    lower = q1 - 1.5 * iqr
    upper = q3 + 1.5 * iqr
    return ((series < lower) | (series > upper)).sum()

outlier_counts = df.select_dtypes(include='number').apply(count_outliers)
most_outliers_feature = outlier_counts.idxmax()
print(f"\nПризнак с наибольшим количеством выбросов: {most_outliers_feature}")
```



6. Выполните стандартизацию всех числовых признаков.

```
numeric_cols = df.select_dtypes(include='number').columns
scaler = StandardScaler()
df_scaled = df.copy()
df_scaled[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```

Вывод: получили практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научились выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.