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

Лабораторная работа №1
По дисциплине «Основы машинного обучения»
Тема: **«Знакомство с анализом данных: предварительная обработка и визуализация»**

Выполнила:
Студентка 3 курса
Группы АС-65
Степанова Д. А.
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 7

Выборка Auto MPG. Содержит технические характеристики различных автомобилей и данные о расходе топлива (миль на галлон).

Задачи:

1. Загрузите данные. Обратите внимание, что пропуски в столбце horsepower могут быть обозначены знаком (?).

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

df = pd.read_csv("auto-mpg.csv")
print("Первые строки:\n", df.head(), "\n")
```

```
Первые строки:
    mpg  cylinders  displacement horsepower  weight  acceleration  model year  origin                   car name
0  18.0          8         307.0        130    3504          12.0          70       1  chevrolet chevelle malibu
1  15.0          8         350.0        165    3693          11.5          70       1          buick skylark 320
2  18.0          8         318.0        150    3436          11.0          70       1         plymouth satellite
3  16.0          8         304.0        150    3433          12.0          70       1              amc rebel sst
4  17.0          8         302.0        140    3449          10.5          70       1                ford torino
```

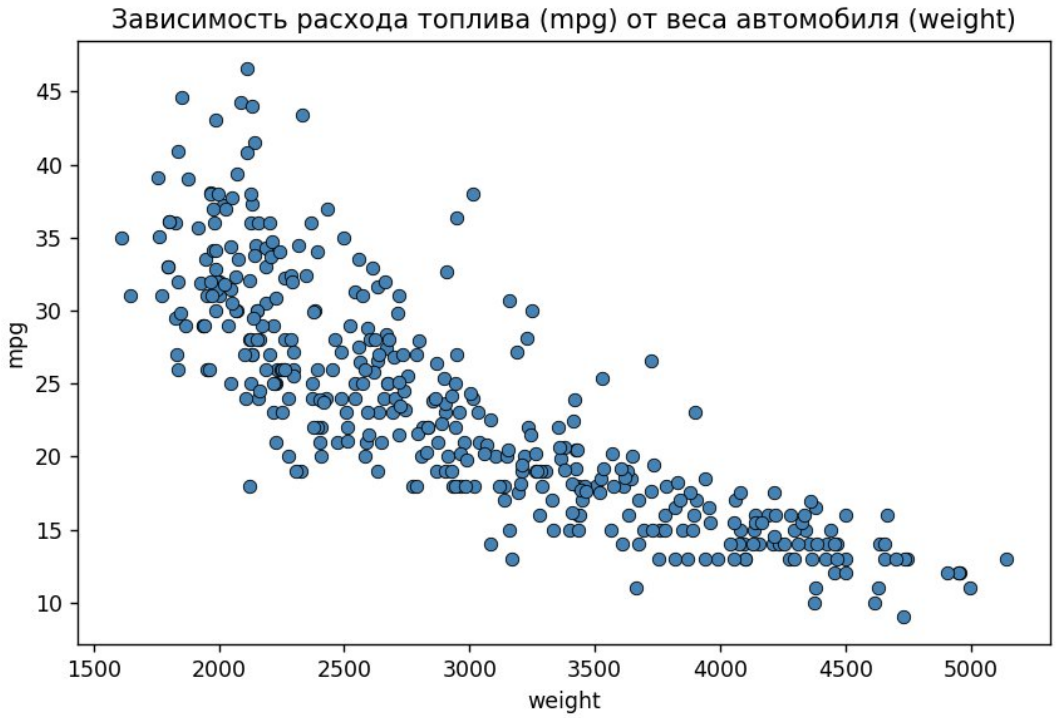
2. Преобразуйте столбец horsepower в числовой формат и заполните пропуски средним значением.

```
df["horsepower"] = pd.to_numeric(df["horsepower"], errors="coerce")
mean_hp = df["horsepower"].mean()
df["horsepower"] = df["horsepower"].fillna(mean_hp)
print(f"Среднее значение horsepower (замена пропусков): {mean_hp:.2f}\n")
```

```
Среднее значение horsepower (замена пропусков): 104.47
```

3. Постройте диаграмму рассеяния, чтобы изучить зависимость расхода топлива (mpg) от веса автомобиля (weight).

```
plt.figure(figsize=(8, 5))
sns.scatterplot(data=df, x="weight", y="mpg", color="steelblue",
edgecolor="black")
plt.title("Зависимость расхода топлива (mpg) от веса автомобиля (weight)")
plt.xlabel("weight")
plt.ylabel("mpg")
plt.show()
```



4. Преобразуйте категориальный признак origin (страна производства) в числовой.

```
# 1-usa 2-eu 3-japan
df["origin"] = df["origin"].map({1: "USA", 2: "Europe", 3: "Japan"})
df = pd.get_dummies(df, columns=["origin"], drop_first=False)
print("Последние колонки после кодирования:\n", df.columns.tolist()[-5:],
"\n")
```

```
Последние колонки после кодирования:
 ['model year', 'car name', 'origin_Europe', 'origin_Japan', 'origin_USA']
```

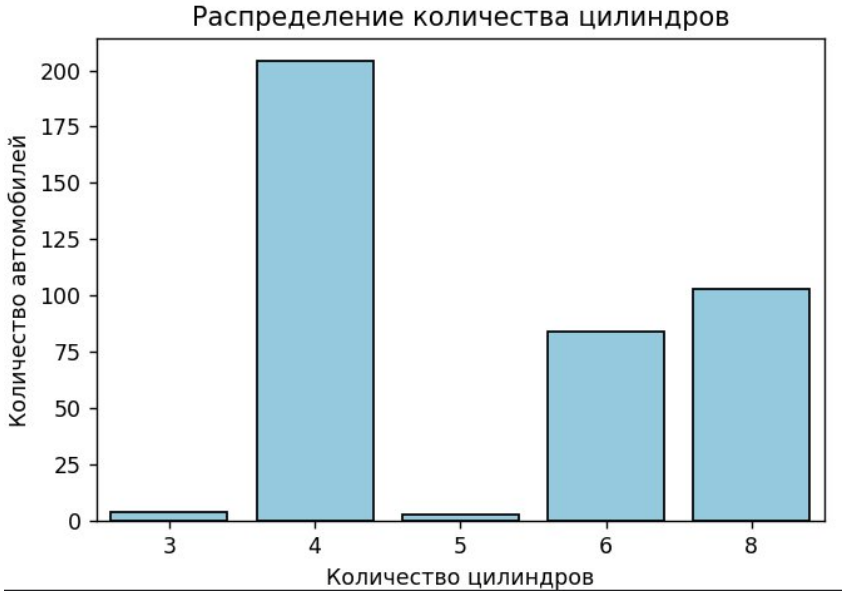
5. Создайте новый признак age, рассчитав возраст автомобиля относительно года, когда были собраны данные (например, 1983 - model year).

```
df["age"] = 1983 - (1900 + df["model year"])
print("Пример возрастов автомобилей:\n", df[["model year", "age"]].head(),
"\n")
```

```
Пример возрастов автомобилей:
    model year  age
0          70   13
1          70   13
2          70   13
3          70   13
4          70   13
```

6. Визуализируйте распределение количества цилиндров (cylinders) с помощью столбчатой диаграммы.

```
plt.figure(figsize=(6, 4))
sns.countplot(data=df, x="cylinders", color="skyblue", edgecolor="black")
plt.title("Распределение количества цилиндров")
plt.xlabel("Количество цилиндров")
plt.ylabel("Количество автомобилей")
plt.show()
```



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.