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»

Выполнил:
Студент 3 курса
Группы АС-65
Касьяник К. 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

data = pd.read_csv(
    "auto-mpg.csv",
    na_values="?"
)

print("Размер датасета:", data.shape)
print(data.head(), "\n")
```

```
Размер датасета: (398, 9)
    mpg  cylinders  displacement  horsepower  weight  acceleration  model year  origin                   car name
0  18.0          8         307.0       130.0    3504          12.0          70       1  chevrolet chevelle malibu
1  15.0          8         350.0       165.0    3693          11.5          70       1          buick skylark 320
2  18.0          8         318.0       150.0    3436          11.0          70       1         plymouth satellite
3  16.0          8         304.0       150.0    3433          12.0          70       1              amc rebel sst
4  17.0          8         302.0       140.0    3449          10.5          70       1                ford torino
```

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

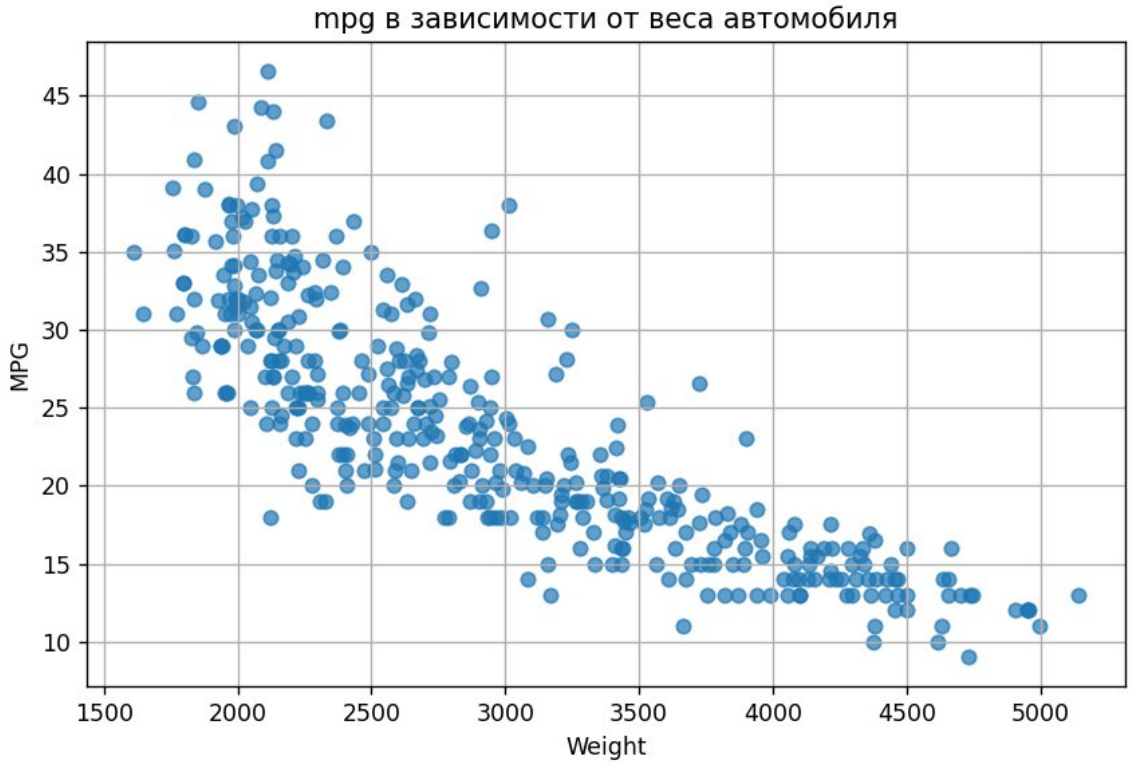
```
data["horsepower"] = data["horsepower"].astype(float)

hp_avg = data["horsepower"].mean()
data["horsepower"].fillna(hp_avg, inplace=True)

print(f"Заполненные пропуски horsepower средним: {hp_avg:.2f}\n")
```

```
Заполненные пропуски horsepower средним: 104.47
```

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

```
plt.figure(figsize=(8, 5))
plt.scatter(
    data["weight"],
    data["mpg"],
    alpha=0.7
)
```

```
plt.title("mpg в зависимости от веса автомобиля")
plt.xlabel("Weight")
plt.ylabel("MPG")
plt.grid(True)
plt.show()
```



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

```
origin_labels = {
    1: 0,
    2: 1,
    3: 2
}

data["origin"] = data["origin"].replace(origin_labels)
print("Уникальные значения origin:", data["origin"].unique(), "\n")
```

```
Уникальные значения origin: [0 2 1]
```

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

```
collection_year = 1983
data["age"] = collection_year - (data["model year"] + 1900)

print(data[["model year", "age"]].head(), "\n")
```

	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.

```
cyl_counts = data["cylinders"].value_counts().sort_index()

plt.figure(figsize=(6, 4))
plt.bar(
    cyl_counts.index.astype(str),
    cyl_counts.values
)
plt.title("Распределение автомобилей по количеству цилиндров")
plt.xlabel("Cylinders")
plt.ylabel("Количество")
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.