Лабораторная работа №6. Исследовательский анализ данных (EDA)

Изучить методы статистического анализа и визуализации данных.

Датасет для работы

Используется набор данных Students Performance Dataset.

Источник:
Students Performance Dataset

Теоретические сведения

Исследовательский анализ данных (EDA) позволяет выявить:

- закономерности;
- выбросы;
- зависимости между признаками;
- распределение данных.

Для визуализации используются библиотеки Matplotlib и Seaborn.

Задание

1. Загрузить подготовленный набор данных.
2. Рассчитать основные статистические показатели.
3. Построить гистограммы распределения.
4. Построить диаграмму корреляции.
5. Выполнить анализ зависимостей между признаками.

Пример выполнения import

pandas as pd

import matplotlib.pyplot as plt import
seaborn as sns

```
# Загрузка данных
data = pd.read_csv("clean_data.csv")
```

```
# Статистика
print(data.describe()) #
Гистограмма
data.hist(figsize=(10, 8)) plt.show()

# Корреляционная матрица corr =
data.corr(numeric_only=True)

# Тепловая карта
sns.heatmap(corr, annot=True, cmap="coolwarm") plt.show()
```

Контрольные вопросы

1. Что такое EDA?
2. Для чего используется корреляционная матрица?
3. Что показывает гистограмма?
4. Какие библиотеки используются для визуализации?
   Что такое выбросы в данных?