Лабораторная работа 1.

Тема: Первичный анализ и предобработка данных

Цель работы: Освоить на практике основные методы первичного анализа данных (EDA), их визуализации и предобработки с использованием библиотек Python.

Задание

1. Выберите набор данных на платформе Kaggle. Рекомендуется выбрать набор с различными типами признаков (числовые, категориальные) и наличием пропусков (например, House Prices, Wine Quality, Spotify Tracks).
2. Загрузите данные в среду разработки (Jupyter Notebook / Google Colab).
3. Проведите все этапы разведочного анализа (EDA), описанные ниже.
4. Оформите отчет в электронном виде, приложив ссылку на Jupyter Notebook/ Google Colab, где код сопровождается краткими выводами по каждому шагу, электронный вид отчёта в формате pdf в ЭИОС.

Подробная информация о задании:

1. Загрузка и первичный осмотр:
   - Загрузите данные в DataFrame.
   - Выведите первые 5-10 строк.
   - Используйте методы .info(), .describe(), .shape для получения общей информации.
   - Вывод: Опишите общий размер набора данных, типы признаков и наличие пропущенных значений.
2. Анализ пропусков:
   - Посчитайте количество и долю пропусков в каждом столбце.
   - Визуализируйте матрицу пропусков с помощью sns.heatmap().
   - Вывод: Определите столбцы с наибольшим процентом пропусков. Предложите стратегию их обработки (удаление, заполнение медианой/модой).
3. Анализ числовых признаков:
   - Для всех числовых столбцов постройте гистограммы и boxplot'ы.
   - Рассчитайте стандартные статистики (среднее, медиана, стандартное отклонение, асимметрия) для ключевых числовых признаков.
   - Вывод: Охарактеризуйте распределения, наличие выбросов.

4. Анализ категориальных признаков:
   - Для категориальных столбцов постройте столбчатые диаграммы (sns.countplot()).
   - Посчитайте количество уникальных категорий в каждом признаке.
   - Вывод: Определите категориальные признаки с большим количеством уникальных значений (высокая кардинальность).
5. Анализ взаимосвязей:
   - Постройте матрицу корреляций для числовых признаков и визуализируйте ее тепловой картой (sns.heatmap()).
   - Для пар ключевых признаков постройте диаграммы рассеяния (sns.scatterplot()).
   - Исследуйте взаимосвязь категориальных и числовых признаков с помощью boxplot'ов (например, sns.boxplot(x='категория', y='число')).
   - Вывод: Назовите наиболее коррелирующие пары признаков. Есть ли видимая зависимость между целевой переменной и другими признаками?
6. Базовая предобработка:
   - Приведите названия столбцов к удобному формату (например, нижний регистр).
   - Обработайте пропуски в соответствии с выводами из п.2.
   - Преобразуйте категориальные признаки в числовой формат выбранным методом.
7. Обработка выбросов (*):
   - Выберите один числовой признак с сильными выбросами.
   - Примените к нему один из методов обработки выбросов (например, логарифмирование или "обрезку" на основе IQR).
   - Постройте boxplot до и после обработки и прокомментируйте результат.

Требования к отчету

Отчет должен содержать:

1. Титульный лист: Название работы, ФИО, группа.
2. Введение: Краткое описание выбранного набора данных и постановка задачи.
3. Основная часть: Последовательное выполнение всех пунктов задания с кодом и текстовыми выводами после каждого шага.
4. Заключение: Общие выводы по проведенному анализу:
   - Какие основные проблемы данных были выявлены?
   - Какие методы предобработки были применены и почему?
   - Как проведенная работа повлияет на дальнейшее построение моделей?
5. Ссылка на выбранный датасет на Kaggle.