Лабораторная работа 3. Обработка данных на языке Python (100 баллов).

Анализ набора данных "German Credit Data"

Цель работы: закрепить навыки работы с наборами данных, освоить инструменты описательной статистики, визуализации, обработки данных и взаимодействия с базой данных.

Датасет содержит информацию о клиентах и их кредитной истории. Он включает признаки и целевую переменную, которая указывает, является ли клиент добросовестным заемщиком. Данные доступны по ссылке: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

Задачи:

1. Загрузка и подготовка исходных данных
2. Анализ переменных и их распределений
3. Кодирование категориальных признаков
4. Построение графиков и диаграмм
5. Формирование SQL-запросов для работы с базой данных
6. Оформление отчёта.

Порядок выполнения работы:

1. Загрузка и подготовка данных
   Скачать и загрузить набор данных "German Credit Data".
   Назначить именованные столбцы согласно предоставленной спецификации.
   Обработать возможные пропущенные значения.

2. Анализ данных
   Выполнить описание числовых признаков (статистики, распределения).
   Проанализировать категориальные признаки (частоты и распределения).
   Провести кодирование категориальных признаков в числовой формат для дальнейшего анализа.

3. Обработка и исследование взаимосвязей
   Построить корреляционную матрицу числовых признаков.
   Провести группировки и агрегацию по ключевым признакам (например, по статусу кредитования, по целям кредита).

4. Визуализация данных
   Построить гистограммы, boxplot-ы и другие графики для визуального анализа.
   Обосновать выбор типов графиков по конкретным признакам.

5. Работа с базой данных
   Создать базу данных и таблицу для хранения данных.
   Вставить обработанные данные в таблицу.
   Выполнить выборки по ключевым признакам, показать примеры.
   Выполнить агрегатные запросы (например, средний кредит по целям).

6. Документирование
   Оформить код, результаты анализа и объяснения по каждому этапу.

Подготовить краткий отчет по выполненной работе.

Состав сдаваемых материалов:

1. скрипт Python (.py) — полный рабочий код с комментариями;
2. отчет в PDF или Word (.pdf/.docx), включающего все разделы;
3. файлы с графиками (можно вставить в отчет или приложить отдельно).

Дополнительные требования:

1. Весь код должен быть хорошо прокомментирован, понятен и структурирован.
2. Использовать только открытые библиотеки Python (pandas, matplotlib, seaborn, scikit-learn, sqlite3).
3. В процессе анализа применять как минимум 3 различных визуальных метода (гистограммы, boxplot, тепловая карта и др.).
4. SQL-запросы должны быть разнообразными: выборки, группировки, агрегаты.
5. Не допускается использование устаревших методов
6. Итоговую работу оформить аккуратно, без ошибок.

Оценивание:

Критерий	Максимальный балл	Описание оценки
Загрузка и подготовка данных	15	Правильность загрузки, структура данных, обработка пропущенных значений, кодирование категориальных признаков.
Анализ данных	20	Описание статистики, анализ распределений, выявление ключевых характеристик и взаимосвязей.
Визуализация данных	15	Построение и качество графиков: гистограммы, boxplot, тепловая карта, их интерпретация.
Работа с базой данных SQLite	15	Создание базы, таблицы, вставка данных, выполнение разнообразных SQL-запросов и интерпретация их результатов.
Качество кода и оформление	10	Хорошая структура кода, комментарии, читаемость, отсутствие ошибок, соответствие стандартам оформления.
Отчёт и оформление проекта	15	Полнота, логическая структура, ясность изложения, оформление по шаблону, наличие всех разделов.
Дополнительные критерии (выбор методов, оригинальность)	10	Использование разнообразных методов анализа, оригинальные подходы, выводы и обоснования.