Лабораторная работа 3: Предобработка и визуализация данных

Цель
Цель данной лабораторной работы — изучить методы предобработки данных и их визуализации. Студенты научатся очищать и нормализовать данные, а также визуализировать их с использованием библиотек Matplotlib и Seaborn, чтобы проводить исследование данных и делать выводы.

Описание
В рамках лабораторной работы студенты будут заниматься предобработкой данных, что включает в себя очистку и нормализацию данных, а также их визуализацию. Студенты используют Matplotlib и Seaborn для создания графиков и диаграмм, помогающих в анализе данных. Лабораторная работа охватывает как подготовку данных к анализу, так и визуальное представление результатов.

Общие задачи:
1. **Очистка данных**: Обнаружение и обработка пропущенных значений, удаление или замена выбросов.
2. **Нормализация данных**: Приведение данных к единому масштабу и диапазону для улучшения анализа и моделирования.
3. **Визуализация данных**: Построение различных типов графиков и диаграмм для визуального анализа данных.
4. **Исследование данных**: Проведение анализа данных для выявления трендов, закономерностей и аномалий.

Индивидуальные задания:
Каждому студенту предоставляется уникальная задача в соответствии с его номером в списке группы (смотреть в SSO).
1. **Анализ данных о пассажирах Титаника:**
- **Задача**: Загрузить набор данных о пассажирах Титаника. Очистить данные от пропусков (например, возраст, порт посадки). Визуализировать распределение пассажиров по возрасту и классу с помощью гистограмм и boxplot.
- **Визуализация**: Использовать Seaborn для создания boxplot и Matplotlib для гистограмм.

2. Предобработка и визуализация данных о доходах:
- Задача: Загрузить набор данных с информацией о доходах населения. Нормализовать данные по возрасту, уровню образования и заработной плате. Визуализировать взаимосвязь между этими параметрами с помощью scatterplot и heatmap.
- Визуализация: Использовать Matplotlib для scatterplot и Seaborn для heatmap.

3. Анализ данных о потреблении электроэнергии:
- Задача: Загрузить данные о потреблении электроэнергии за несколько лет. Провести очистку данных от выбросов и заполнение пропусков. Визуализировать временные ряды с использованием линии тренда и распределение потребления по месяцам.
- Визуализация: Использовать Matplotlib для временных рядов и Seaborn для распределения по месяцам.

4. Визуализация корреляций в данных о здоровье:

- Задача: Загрузить набор данных о здоровье населения. Провести корреляционный анализ показателей (например, уровень сахара, артериальное давление). Визуализировать корреляционную матрицу и пары признаков с высокой корреляцией.
- Визуализация: Использовать Seaborn для построения корреляционной матрицы (heatmap) и pairplot.

5. Предобработка и анализ данных о недвижимости:
- Задача: Загрузить данные о продажах недвижимости. Очистить данные от выбросов, заполнить пропуски, нормализовать цены. Визуализировать распределение цен по районам и связь цены с другими характеристиками (площадь, год постройки).
- Визуализация: Использовать Matplotlib для гистограмм и Seaborn для scatterplot.

6. Анализ данных о демографических показателях:
- Задача: Загрузить набор данных с демографической информацией (возраст, пол, образование). Очистить данные и провести нормализацию. Визуализировать распределение возрастных групп по полу и уровню образования с помощью barplot и violinplot.
- Визуализация: Использовать Seaborn для barplot и violinplot.

7. Предобработка данных о погоде и их визуализация:
- Задача: Загрузить данные о погоде (температура, влажность, осадки). Провести предобработку данных: удаление выбросов, нормализация. Визуализировать временные ряды температур и осадков, а также распределение влажности.
- Визуализация: Использовать Matplotlib для временных рядов и Seaborn для распределений.

8. Анализ данных о посещаемости студентов:
- Задача: Загрузить данные о посещаемости студентов. Очистить данные от пропусков и нормализовать их. Визуализировать распределение посещаемости по дням недели и по успеваемости студентов.
- Визуализация: Использовать Seaborn для построения countplot и Matplotlib для визуализации временных рядов.

9. Предобработка и визуализация данных о транспортных средствах:
- Задача: Загрузить данные о характеристиках автомобилей (мощность, расход топлива, тип двигателя). Очистить и нормализовать данные. Визуализировать распределение автомобилей по мощности и расходу топлива, построить парные графики.
- Визуализация: Использовать Seaborn для pairplot и гистограмм.

10. Анализ данных о социально-экономических показателях:
- Задача: Загрузить данные о социально-экономических показателях (уровень безработицы, доход, образование). Провести предобработку данных и нормализацию. Визуализировать связь между показателями с помощью scatterplot и jointplot.
- Визуализация: Использовать Seaborn для scatterplot и jointplot.

11. Предобработка и анализ данных о потребительских предпочтениях:
- Задача: Загрузить данные о предпочтениях потребителей (например, частота покупок, средний чек). Очистить данные от выбросов и нормализовать их.

Визуализировать распределение предпочтений и зависимость между частотой покупок и средним чеком.
- Визуализация: Использовать Matplotlib для распределений и Seaborn для pairplot.

12. Анализ данных о результатах спортивных соревнований:
- Задача: Загрузить данные о результатах спортивных соревнований (например, время прохождения дистанции, количество участников). Провести предобработку данных и нормализацию. Визуализировать временные ряды и распределение результатов по категориям.
- Визуализация: Использовать Matplotlib для временных рядов и Seaborn для распределений.

13. Предобработка данных о преступности и их визуализация:
- Задача: Загрузить данные о преступности в городе (виды преступлений, районы). Очистить данные от выбросов и нормализовать их. Визуализировать распределение преступлений по районам и их типам.
- Визуализация: Использовать Seaborn для heatmap и Matplotlib для barplot.

14. Анализ данных о туристическом потоке:
- Задача: Загрузить данные о туристическом потоке (число туристов, длительность пребывания, расходы). Провести предобработку данных и нормализацию. Визуализировать временные ряды и зависимость между длительностью пребывания и расходами.
- Визуализация: Использовать Matplotlib для временных рядов и Seaborn для scatterplot.

15. Предобработка и анализ данных о здоровье:
- Задача: Загрузить данные о здоровье (вес, рост, индекс массы тела). Провести предобработку данных, заполнить пропуски, нормализовать данные. Визуализировать распределение индекса массы тела по возрастным группам и полу.
- Визуализация: Использовать Seaborn для построения violinplot и barplot.

Вопросы для подготовки

1. Какие методы существуют для обработки пропущенных значений в наборе данных?
2. Как выявить и обработать выбросы в данных?
3. Что такое нормализация данных и какие методы используются для этого?
4. Какие виды визуализаций подходят для анализа распределений данных?
5. Как интерпретировать корреляционную матрицу и тепловую карту?
6. Что такое диаграмма рассеяния и как она помогает в анализе данных?
7. Как анализировать временные ряды и какие графики для этого используются?
8. Каковы преимущества и недостатки различных методов нормализации данных?
9. Какие визуализации можно использовать для анализа категориальных данных?
10. Как интегрировать методы предобработки данных и визуализацию для комплексного анализа?

Критерии Оценки:
- Написание кода для решения индивидуальной задачи: 2 балл

- Объяснение и понимание написанного кода во время защиты: 1 балла
- Ответ на один из теоретических вопросов, выбранный преподавателем: 1 балл

**** Обязательно предоставьте отчет в формате Word, соответствующий всем стандартам.***