Лабораторная работа № 4
по дисциплине
Статистический анализ и визуализация данных

Тема: Анализ различий между группами и динамики числовых данных во времени с использованием статистических тестов и интерактивной визуализации

Задание:

1. Выбрать набор данных, содержащий как минимум одну числовую переменную, одну категориальную переменную и одну временную переменную, с не менее чем 50 наблюдениями.
2. Загрузить данные в Python или R, проверить их структуру, выявить и обработать пропущенные значения, а также правильно установить типы переменных. В частности, столбцы с датами должны быть преобразованы в формат времени (date/time), а переменные, представляющие группы — в категориальные переменные.
3. Провести первоначальную визуализацию временного ряда, включая линейный график для динамики, гистограмму для распределения и boxplot для сравнения значений между группами.
4. Смоделировать временную динамику числовой переменной с использованием моделей ARIMA и SARIMA для выявления трендов и сезонности. Протестировать несколько конфигураций, а окончательный выбор сделать на основе показателей AIC и ошибок прогнозирования.
5. Проверить нормальность распределения анализируемой величины для каждой группы с помощью теста Шапиро-Уилка и функций qqnorm(), qqline(). Если нормальность не соблюдается, применить тест Крускала-Уоллиса.
6. Проверить гипотезу о наличии значимых различий между средними значениями числовой переменной в зависимости от групп, используя однофакторный дисперсионный анализ (ANOVA), и интерпретировать результат на основе p-значения.
7. Вручную выполнить полный расчет теста ANOVA для трех групп, включая средние по группам, общее среднее, суммы квадратов между группами и внутри групп, средние квадраты и F-значение, которое затем сравнить с результатом, полученным в Python/R.
8. Если обнаружены значимые различия, провести пост-хок анализ, применив тест Тьюки в случае ANOVA или тесты Манна-Уитни/Уилкоксона для парных сравнений групп.
9. Выбрать подмножество данных, содержащее ровно две группы, и применить тест Манна-Уитни как с помощью Python/R, так и вручную, рассчитав объединенные ранги, сумму рангов и значение статистики U.
10. Создать сравнительные визуализации между группами, включая boxplot, наложенные гистограммы и тепловые карты для корреляций.

11. Подготовить полную интерпретацию полученных результатов, включая анализ различий между группами, обоснованность примененных методов, сопоставление с визуализациями и возможные ограничения набора данных.