МИНОБРНАУКИ РФ

ФГБОУ Тверской государственный технический университет

Факультет информационных технологий

Кафедра «Программное обеспечение»

Практическая работа №1

Дисциплина: «Анализ больших данных»

Вариант №7

Работу выполнил: студент группы

ПИН. РИС.21.06

Олимов Авазбек.

Тверь, 2025

# Формулировка задания, описание условий

1. Ознакомьтесь с набором данных mpg из библиотеки Seaborn.

(загрузка через df = sns.load\_dataset(’mpg’))

2. Посчитайте количество строк и столбцов.

3. Проведите разведочный анализ, то есть:

(a) для каждой числовой переменной вычислите:

• Долю пропусков

• Максимальное и минимальное значение

• Среднее значение

• Медиану

• Дисперсию

• Квантиль 0.1 и 0.9

• Квартиль 1 и 3

(b) для каждой категориальной переменной вычислите:

• Долю пропусков

• Количество уникальных значений

• Моду

4. Сформулируйте и проверьте минимум 2 статистические гипотезы. Выбор критериев для проверки гипотез требуется обосновать. Сделать выводы в терминах предметной области.

5. Закодируйте категориальные переменные, необходимые для анализа, если требуется. Методом OneHotEncoding или LabelEncoding.

6. Постройте таблицу корреляции признаков и целевого столбца. Обоснуйте, какой столбец является целевым, а какие признаками.

7. Реализуйте стохастический и обычный градиентный спуск вручную, можно использовать ноутбук с лекции ссылка. Для этих данных: y = ’mpg’ и x = ’horsepower’ или ’weight’.

Задание для самостоятельного выполнения

Реализовать задание для данных вашего варианта

1. Преобразовать категориальные переменные в числовые, если это необходимо. Добавить вычисляемые столбцы.

2. Посчитайте количество строк и столбцов.

3. Проведите разведочный анализ, то есть:

(a) для каждой числовой переменной вычислите:

• Долю пропусков

• Максимальное и минимальное значение

• Среднее значение

• Медиану

• Дисперсию

• Квантиль 0.1 и 0.9

• Квартиль 1 и 3

(b) для каждой категориальной переменной вычислите:

• Долю пропусков

• Количество уникальных значений

• Моду

4. Сформулируйте и проверьте минимум 2 статистические гипотезы. Выбор критериев для проверки гипотез требуется обосновать. Сделать выводы в терминах предметной области.

5. Постройте таблицу корреляции признаков и целевого столбца. Обоснуйте, какой столбец является целевым, а какие признаками.

# 2. Вариант и условия задач. Указать какой сложности выполняется задание.

Вариант: №7  
Сложность: Rare

Условия**:**

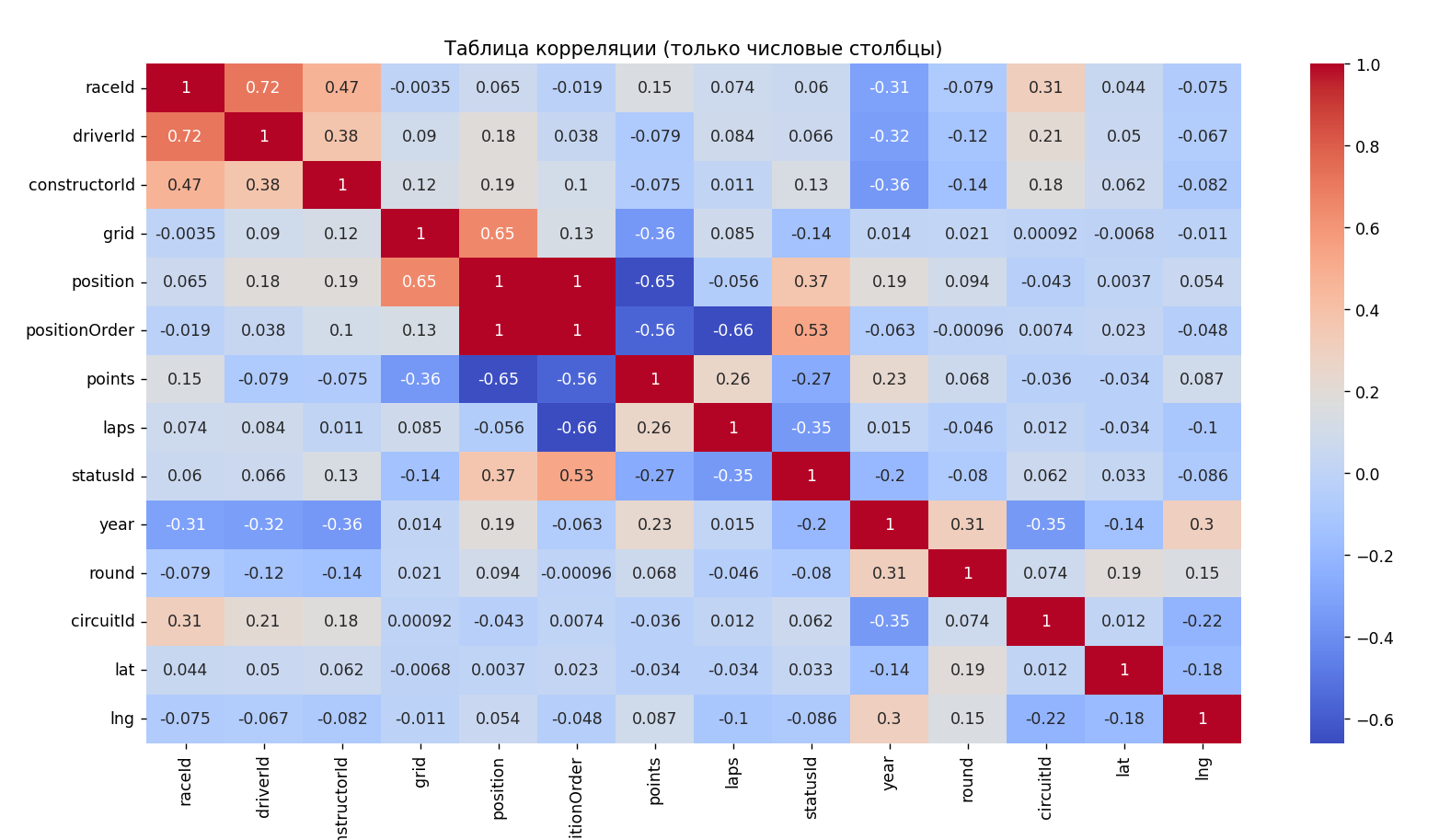
Использовать базу данных SQLite.

Реализовать Jupiter Notebook или консольное приложение для выполнения задания.

Провести анализ данных о гонках Формулы 1.

# Ссылка на репозиторий с программной реализацией

<https://github.com/Avazbek22/BigData>

1. Описание проделанной работы
2. Загрузка данных:
   * Данные загружены из базы данных SQLite, содержащей информацию о гонках Формулы 1.
   * Таблицы объединены в один DataFrame для удобства анализа.
3. Разведочный анализ:
   * Для числовых переменных вычислены: доля пропусков, минимальное и максимальное значения, среднее, медиана, дисперсия, квантили и квартили.
   * Для категориальных переменных вычислены: доля пропусков, количество уникальных значений и мода.
4. Проверка гипотез:
   * Гипотеза 1: Средний результат (points) для гонщиков из разных стран отличается.
   * Гипотеза 2: Корреляция между стартовой позицией (grid) и финишной позицией (position).
5. Кодирование категориальных переменных:
   * Категориальные переменные закодированы с использованием OneHotEncoding и LabelEncoding.
6. Таблица корреляции:
   * Построена тепловая карта корреляции для числовых столбцов.
   * 
7. Градиентный спуск:
   * Реализованы обычный и стохастический градиентный спуск для предсказания points на основе grid.

# Веб версия с использованием стримлит

Изображение выглядит как текст, снимок экрана, число

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, дисплей, График

Контент, сгенерированный ИИ, может содержать ошибки.

# Краткий вывод по работе

В ходе выполнения лабораторной работы были выполнены следующие задачи:

* Проведён разведочный анализ данных о гонках Формулы 1.
* Проверены две статистические гипотезы, подтверждена значимость различий в результатах гонщиков из разных стран и выявлена корреляция между стартовой и финишной позициями.
* Категориальные переменные успешно закодированы.
* Построена таблица корреляции, визуализированы взаимосвязи между признаками.
* Реализованы алгоритмы градиентного спуска для задачи регрессии.

Программа успешно справляется с поставленными задачами и может быть использована для дальнейшего анализа данных.

# Ссылки на используемые материалы

Документация Pandas: https://pandas.pydata.org/docs/

Документация Seaborn: https://seaborn.pydata.org/

Документация Scipy: https://docs.scipy.org/doc/scipy/

Документация Scikit-learn: https://scikit-learn.org/stable/