МИНОБРНАУКИ РФ

ФГБОУ Тверской государственный технический университет

Факультет информационных технологий

Кафедра «Программное обеспечение»

Лабораторная работа №1

Дисциплина: «Анализ больших данных»

Вариант №7

Работу выполнил: студент группы

ПИН. РИС.21.06

Койич Марко.

Тверь, 2025

# 1. Формулировка задания, описание условий

1. Создать базу из файлов или скриптов согласно варианту. Сложность программы выбрать из представленных ниже.
2. Подключиться к базе данных из python.
3. Сделать описание данных. Из каких таблиц и полей состоят данные таблиц? Какие из данных являются признаками? К какому типу данных и к какой шкале относятся признаки?
4. Одномерный анализ. Построить гистограммы распределения количественных признаков, которые важны для задачи. Сделать вывод после построения. Какое распределение для каждого из признаков? Почему, по вашему мнению, признаки важны для задачи?
5. Многомерный анализ. Построить графики из 3-4 признаков. Выбрать категориальные (номинальные, порядковые или бинарные) признаки и количественные. Что получилось на каждом графике? Почему, по вашему мнению, признаки важны для задачи?

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

Вариант: №7

Сложность: Rare

• Реализовать Jupiter Notebook или консольное приложение для выполнения задания.

• В качестве базы данных выбрать sqlite3. Создать базу из файлов или скриптов согласно варианту.

• Подключиться к базе данных из python.

• Сделать описание данных и выводы по заданию.

• Соединить признаки в 1 таблицу pandas для анализа

• Одномерный анализ. Построить 2 гистограммы распределения количественных признаков, которые важны для задачи и сделать их описание по заданию.

• Многомерный анализ. Построить хотя бы 1 график из 3-4 признаков и сделать его описание по заданию.

# 3. Формулировка задания, описание условий

В данной базе данных о гонках Формулы 1 содержится несколько таблиц, каждая из которых содержит различные аспекты данных о гонках, гонщиках, командах и результатах. Ниже приведено описание основных таблиц и их полей, а также информация о признаках, их типах данных и шкалах.

1. Таблицы и их поля

1.1. Таблица Races

* raceId: уникальный идентификатор гонки (целое число).
* year: год гонки (целое число).
* round: номер гонки в сезоне (целое число).
* circuitId: идентификатор трассы (целое число).
* date: дата гонки (дата).
* name: название гонки (строка).

1.2. Таблица Drivers

* driverId: уникальный идентификатор гонщика (целое число).
* driverRef: ссылка на гонщика (строка).
* number: номер гонщика (целое число).
* code: код гонщика (строка).
* name: имя гонщика (строка).
* nationality: национальность гонщика (строка).

1.3. Таблица Constructors

* constructorId: уникальный идентификатор команды (целое число).
* name: название команды (строка).
* nationality: национальность команды (строка).

1.4. Таблица Results

* resultId: уникальный идентификатор результата (целое число).
* raceId: идентификатор гонки (целое число).
* driverId: идентификатор гонщика (целое число).
* constructorId: идентификатор команды (целое число).
* grid: стартовая позиция (целое число).
* position: позиция в гонке (целое число).
* points: полученные очки (число с плавающей точкой).

1.5. Таблица Circuits

* circuitId: уникальный идентификатор трассы (целое число).
* name: название трассы (строка).
* location: местоположение трассы (строка).
* country: страна (строка).

2. Признаки и их типы данных

2.1. Признаки

Признаки — это поля, которые могут быть использованы для анализа и построения моделей. В данной базе данных можно выделить следующие признаки:

* Таблица Races:
  + year (количественный)
  + round (количественный)
  + date (категориальный, временной)
  + name (категориальный)
* Таблица Drivers:
  + number (количественный)
  + name (категориальный)
  + nationality (категориальный)
* Таблица Constructors:
  + name (категориальный)
  + nationality (категориальный)
* Таблица Results:
  + grid (количественный)
  + position (количественный)
  + points (количественный)
* Таблица Circuits:
  + name (категориальный)
  + location (категориальный)
  + country (категориальный)

2.2. Типы данных и шкалы

* Типы данных:
  + Целые числа (int): raceId, year, round, circuitId, driverId, constructorId, grid, position, number.
  + Числа с плавающей точкой (float): points.
  + Строки (str): name, nationality, location, country, driverRef, code.
  + Дата (date): date.
* Шкала:
  + Количественные признаки (например, year, grid, position, points) имеют интервальную шкалу, так как они представляют собой числовые значения, которые можно сравнивать и выполнять арифметические операции.
  + Категориальные признаки (например, name, nationality, location, country) имеют номинальную шкалу, так как они представляют собой категории без естественного порядка.

# 4. Ссылка на репозиторий с программной реализацией

<https://github.com/Koyara/bigdata>

# 5. Описание проделанной работы

# В рамках выполнения лабораторной работы была реализована программа на языке Python с использованием библиотек sqlite3, pandas, matplotlib и seaborn для анализа данных о гонках Формулы 1. Ниже приведены основные этапы работы:

# Подключение к БД:

# Подключение к базе данных осуществлено с помощью библиотеки sqlite3 в Python.

# Описание данных:

# Проведен анализ структуры данных, описание таблиц и их полей. Определены типы данных и шкалы для каждого признака.

# Выделены ключевые признаки для анализа, такие как year, round, grid, position, points, name, nationality и другие.

# Одномерный анализ:

# Построены гистограммы для количественных признаков, таких как points (очки, полученные гонщиком) и grid (стартовая позиция). Гистограммы позволили визуализировать распределение данных и сделать выводы о том, как часто гонщики набирают определенное количество очков или стартуют с определенной позиции.

# Выявлено, что распределение очков имеет правосторонний перекос, что указывает на то, что большинство гонщиков набирают небольшое количество очков, а лишь немногие получают высокие баллы. Распределение стартовых позиций более равномерное, но с небольшим смещением в сторону более высоких позиций.

# Многомерный анализ:

# Построены графики, объединяющие несколько признаков. Например, был создан scatter plot (точечный график), который отображает зависимость между стартовой позицией (grid) и финишной позицией (position), а также количеством набранных очков (points).

# Также был построен box plot (ящик с усами) для анализа распределения очков в зависимости от национальности гонщика (nationality). Это позволило выявить, что гонщики из некоторых стран в среднем набирают больше очков, чем другие.

# На графиках видно, что стартовая позиция имеет значительное влияние на финишную позицию и количество набранных очков, что подтверждает важность квалификации в гонках Формулы 1.

# Объединение данных в одну таблицу:

# Для удобства анализа данные из нескольких таблиц были объединены в одну таблицу с использованием библиотеки pandas. Это позволило проводить анализ на основе всех доступных признаков.

# Реализация программы:

# Программа реализована в виде Jupyter Notebook, что позволяет интерактивно выполнять код и визуализировать результаты. Все этапы работы, включая загрузку данных, анализ и построение графиков, описаны в ноутбуке.

# 6. Краткий вывод по работе. Описание реализованной программы и её функций.

В ходе выполнения лабораторной работы была разработана программа для анализа данных о гонках Формулы 1. Программа позволяет:

* Подключаться к базе данных SQLite и загружать данные из нескольких таблиц.
* Проводить одномерный анализ данных, включая построение гистограмм для количественных признаков.
* Проводить многомерный анализ данных, включая построение scatter plots и box plots для выявления взаимосвязей между признаками.
* Объединять данные из разных таблиц в одну для удобства анализа.
* Визуализировать результаты анализа с использованием библиотек matplotlib и seaborn.

Выводы:

* Анализ данных показал, что стартовая позиция (grid) имеет значительное влияние на финишную позицию (position) и количество набранных очков (points). Это подтверждает важность квалификации в гонках Формулы 1.
* Распределение очков среди гонщиков имеет правосторонний перекос, что указывает на то, что лишь немногие гонщики набирают большое количество очков.
* Национальность гонщика также может влиять на его результаты, что подтверждается анализом распределения очков в зависимости от национальности.

Программа успешно справляется с поставленными задачами и может быть использована для дальнейшего анализа данных о гонках Формулы 1.