*Лабораторная работа №2*

**Исследовательский анализ данных. Постановка гипотез Категориальные данные**

**Цель:** провести исследовательский анализ данных, поставить гипотезы и выявить основные статистики.

**Содержание**

[Общее задание на работу 1](#_Toc188279515)

[Задание для самостоятельного выполнения 2](#_Toc188279516)

[Требования к проекту 2](#_Toc188279517)

[Сложность выполнения 3](#_Toc188279518)

[Варианты 3](#_Toc188279519)

[Отчёт 5](#_Toc188279520)

[Формат выполнения задания 5](#_Toc188279521)

# Общее задание на работу

1. Ознакомьтесь с набором данных mpg из библиотеки Seaborn.

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

1. Посчитайте количество строк и столбцов.
2. 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

1. Сформулируйте и проверьте минимум 2 статистические гипотезы. Выбор критериев для проверки гипотез требуется обосновать. Сделать выводы в терминах предметной области.
2. Закодируйте категориальные переменные, необходимые для анализа, если требуется. Методом OneHotEncoding или LabelEncoding.
3. Постройте таблицу корреляции признаков и целевого столбца. Обоснуйте, какой столбец является целевым, а какие признаками.
4. 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

1. Сформулируйте и проверьте минимум 2 статистические гипотезы. Выбор критериев для проверки гипотез требуется обосновать. Сделать выводы в терминах предметной области.
2. Постройте таблицу корреляции признаков и целевого столбца. Обоснуйте, какой столбец является целевым, а какие признаками.

# Требования к проекту

* Проект реализован локально и выгружен на github или прикреплен в архиве к заданию в elearning
* Проект состоит из файлов программы (py или ipynb), выгрузки (dump) базы данных и файла ReadMe.md
* Проект реализован на python с применением библиотеки pandas.
* README.md должен содержать:
* Название вашего приложения
* Описание
* Инструкции по запуску проекта и базы данных

# Сложность выполнения

**Сложность: Rare**

* Реализовать только общую часть задания

**Сложность: Medium**

* Реализовать общую и самостоятельную часть задания
* Лабораторная работа 1 реализована на **Medium**

**Сложность:** **Well-done**

* Реализовать общую и самостоятельную часть задания
* Лабораторная работа 1 реализована на **Well-done**
* Закодируйте категориальные переменные, необходимые для анализа, если требуется. Методом OneHotEncoding или LabelEncoding.
* Реализуйте стохастический и обычный градиентный спуск вручную для вашего варианта, можно использовать ноутбук с лекции ссылка.

# Варианты

|  |  |  |
| --- | --- | --- |
| № | Вариант | Ссылка на kaggle |
| 1 | SQL Murder Mystery Database | <https://www.kaggle.com/datasets/johnp47/sql-murder-mystery-database> |
| 2 | SQLite Sakila Sample Database | <https://www.kaggle.com/datasets/atanaskanev/sqlite-sakila-sample-database?select=SQLite3+Sakila+Sample+Database+ERD.png> |
| 3 | Bike Store Relational Database | SQL | <https://www.kaggle.com/datasets/dillonmyrick/bike-store-sample-database> |
| 4 | Indian Premier League SQLite Database | <https://www.kaggle.com/datasets/harsha547/ipldatabase> |
| 5 | Housing - SQL Project | <https://www.kaggle.com/datasets/bvanntruong/housing-sql-project> |
| 6 | NBA Database | <https://www.kaggle.com/datasets/wyattowalsh/basketball> |
| 7 | Formula 1 Race Data (SQLite) | <https://www.kaggle.com/datasets/davidcochran/formula-1-race-data-sqlite> |
| 8 | League of Legends Ranked Matches | <https://www.kaggle.com/datasets/paololol/league-of-legends-ranked-matches> |
| 9 | Sample databases for the SQL Server | <https://www.kaggle.com/datasets/emrahaydemr/sample-databases-for-the-sql-server-course> |
| 10 | UEFA Champions League 2016-2022 Data | <https://www.kaggle.com/datasets/cbxkgl/uefa-champions-league-2016-2022-data> |
| 11 | Kensho Derived Wikimedia Dataset | <https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data> |
| 12 | SQL Murder Mystery Database | <https://www.kaggle.com/datasets/johnp47/sql-murder-mystery-database> |
| 13 | SQLite Sakila Sample Database | <https://www.kaggle.com/datasets/atanaskanev/sqlite-sakila-sample-database?select=SQLite3+Sakila+Sample+Database+ERD.png> |
| 14 | Bike Store Relational Database | SQL | <https://www.kaggle.com/datasets/dillonmyrick/bike-store-sample-database> |
| 15 | Indian Premier League SQLite Database | <https://www.kaggle.com/datasets/harsha547/ipldatabase> |
| 16 | Housing - SQL Project | <https://www.kaggle.com/datasets/bvanntruong/housing-sql-project> |
| 17 | NBA Database | <https://www.kaggle.com/datasets/wyattowalsh/basketball> |
| 18 | Formula 1 Race Data (SQLite) | <https://www.kaggle.com/datasets/davidcochran/formula-1-race-data-sqlite> |
| 19 | League of Legends Ranked Matches | <https://www.kaggle.com/datasets/paololol/league-of-legends-ranked-matches> |
| 20 | Sample databases for the SQL Server | <https://www.kaggle.com/datasets/emrahaydemr/sample-databases-for-the-sql-server-course> |
| 21 | UEFA Champions League 2016-2022 Data | <https://www.kaggle.com/datasets/cbxkgl/uefa-champions-league-2016-2022-data> |
| 22 | Kensho Derived Wikimedia Dataset | <https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data> |
| 23 | Housing - SQL Project | <https://www.kaggle.com/datasets/bvanntruong/housing-sql-project> |

# Отчёт

Отчёт должен содержать:

* Титульный лист
* Формулировка задач, описание условий
* Вариант и условия задач. Указать какой сложности выполняется задание.
* Ссылка на репозиторий с программной реализацией
* Описание проделанной работы
* Краткий вывод по работе. Описание реализованной программы и её функций.
* Ссылки на используемые материалы. Документация
* Листинг кода

# Формат выполнения задания

1. Код в репозитории github | gitlab | bitbucket. Для отчета по работе выполнить задание в файле.py или .ipynb.

Желательна реализация в файлах py. Сохранить задачи (сделать коммиты для каждой) в локальном git и опубликовать в удаленном репозитории.

2. Отчет в форматах google docs, docx или tex (+ pdf).

3. База данных в формате db или dump базы данных.