Техническое задание\_v2.0

**Программа для первичного анализа набора данных**

‌Программа **позволяет** пользователю **произвести первичный анализ набора данных** (см. далее в П. 4), получаемыми из файлов данных различных табличных типов (csv, xlsx, SQL). Используя данное приложение, пользователь сможет быстро получить большой объем полезной информации о дата сете без установки сторонних библиотек.

**Цель**: разработать программу для первичного анализа набора данных с графическим интерфейсом, выполняющую получение основных предварительных метрик из дата сета языке программирования Python c использованием библиотеки PyQT5.

**Задачи**:

Изучить основы работы с библиотекой PyQT5.

Изучить основы работы с программой создания графического дизайна и подключения его к исполняемому файлу.

Изучить работу с библиотеками для анализа данных

Изучить основные принципы в статистике и анализе данных

Изучить работу с базами данных

**Основные функциональные возможности приложения**

· Программа имеет интуитивно понятный графический интерфейс

· Окно программы имеет стандартные элементы окна Windows, экранные кнопки закрыть, развернуть и свернуть в верхнем правом углу, строка заголовка содержит название приложения и иконку.

· В следующем пункте будет описан весь функционал и разметка

**Весь функционал**

**Обзор всего дата сета**

1.1. Статистика по дата сету

1.1.1. Количество столбцов

1.1.2. Количество строк

1.1.3. Количество пропущенных клеток

1.1.4. Процент пропущенных клеток

1.1.5. Количество дублированных строк

1.1.6. Процент дублированных строк

1.2. Типы столбцов (тип данных – количество столбцов)

**Обзор каждой переменной в отдельности**

2.1. Название

2.3. Количество уникальных

2.4. Количество пропущенных

2.5. Процент пропущенных

Если переменная типа int или float

2.6. Среднее

2.7. 5 перцентиль

2.8. Минимальное

2.9. Максимальное

2.10. Q1

2.11. Медиана

2.12. Q2

2.13. 95 перцентиль

2.14. Количество нулей

2.15. Процент нулей

2.16. Количество отрицательных

2.17. Процент отрицательных

2.18. Стандартное отклонение

2.19. Сумма

2.20. Гистограмма распределение

2.21. Box plot

2.22. Наиболее повторяющиеся значения

Корреляции одного столбца с другим

3.1. График корреляции

4.1 Вывод 5 случайных строк дата фрейма

5.1 Очистка дата фрейма по дублям и от выбросов

6.1 Сохранение датафрейма в выбранный формат (csv, xlsx, SQL)