**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

**“Большие данные и распределенная цифровая платформа”**

**ОТЧЕТ**

**по лабораторной работе №2**

**по дисциплине «Алгоритмы и структуры данных»**

**на тему Обезличивание датасета»**

**Студент гр. 22Б16-пу \_\_\_\_\_\_\_\_\_\_\_\_\_ Шарабарин М.С.**

**Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_ Дик А.Г.**

**Санкт-Петербург**

**2024 г.**

# Оглавление

[Оглавление 3](#_Toc2114525749)

[Цель работы 4](#_Toc762717844)

[Описание задачи (формализация задачи) 4](#_Toc1734968799)

[Теоретическая часть 4](#_Toc275273379)

[Основные шаги программы 5](#_Toc119929492)

[Ограничения: 5](#_Toc714914723)

[Блок схема программы 6](#_Toc1020108421)

[Описание программы 8](#_Toc821754095)

[Рекомендации для пользователя 8](#_Toc420551492)

[Рекомендации для программиста 9](#_Toc1514026432)

[Исходный код программы 9](#_Toc1188901813)

[Контрольный пример 9](#_Toc1837342827)

[Вывод 10](#_Toc701907079)

[Источники 10](#_Toc39607044)

# Цель работы

Целью лабораторной работы является разработка скрипта для анонимизации данных и обезличивания выходного датасета, содержащего личные данные пассажиров, информацию о поездах и стоимости билетов. Основное внимание уделяется обеспечению конфиденциальности данных.

# Описание задачи (формализация задачи)

Задача состоит в создании анонимизированного датасета, который включает:

* **Обезличенные фиксированные имена пассажиров.**
* **Замаскированные номера паспортов и кредитных карт.**
* **Неопределенные номера поездов и стоимость билетов, влияющие на защиту персональных данных.**
* **Однородное представление временных меток для поездок.**

# Теоретическая часть

* **Pandas**: для работы с данными в формате таблиц.
* **NumPy**: для генерации случайных чисел и математических операций.
* **Datetime**: для обработки временных меток.
* **Lambda-функции**: для применения анонимизации данных в столбцах.

## 

# Основные шаги программы

1. **Анонимизация данных**: Замена реальных значений на анонимизированные, включая ФИО, номера паспортов и карт.
2. **Обобществление номеров и цен**: Применение маскирования и обобщения.
3. **Обработка временных меток**: Добавление случайных значений для анонимизации временных данных.
4. **Выделение уникальных строк и вычисление K-анонимити**: Определение групп записей и их количества для защиты данных.
5. **Представление результатов**: Вывод результатов анализа, таких как размер исходного и анонимизированного наборов данных.

# Ограничения:

* Данные должны быть в формате csv

# Блок схема программы

****

Рис 1. Блок-схема основной программы

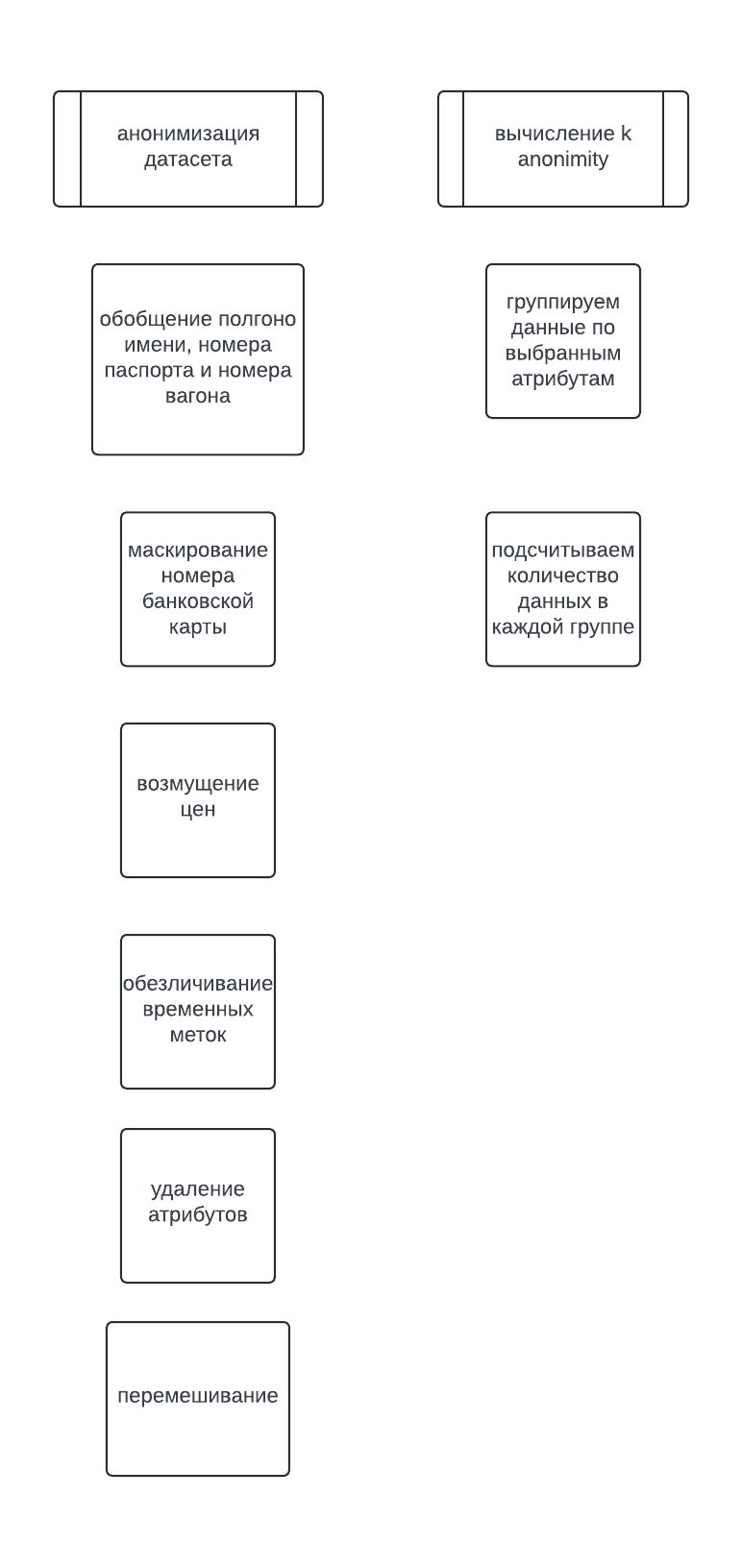


Рис 2. Блок-схема подпрограмм

# Описание программы

Программа написана на Python и состоит из нескольких функций, каждая из которых отвечает за свою задачу. Основная функция main выполняет последовательные шаги анонимизации, валидации и выводит отчеты по анализу.

* **anonymize\_full\_name**: Анонимизирует ФИО, заменяя их на фиктивные.
* **anonymize\_passport\_number**: Замаскирует номера паспортов.
* **anonymize\_card\_number**: Маскирует номера кредитных карт.
* **anonymize\_train\_number**: Обобщает номера поездов.
* **anonymize\_price**: Изменяет цены с учетом случайных колебаний.

# Рекомендации для пользователя

При использовании программы рекомендуем:

* Обеспечить корректный ввод данных.
* Следить за сохранением оригинального датасета для сравнения.
* Внимательно относиться к настройкам анонимизации.

# Рекомендации для программиста

* Документировать код и функции для лучшей читаемости.
* Проводить тестирование каждой функции.
* Обработать возможные исключения для повышения надежности.

# Исходный код программы

**<https://colab.research.google.com/drive/1QD3IdGrSXZHZ9Y2NhbSFuQ1rumTwVJAm>**

# Контрольный пример

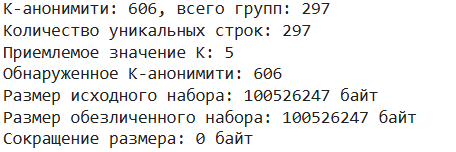
**

Рис 3. пример ввода значений и k-анонимити

# Вывод

В данной лабораторной работе реализован генетический алгоритм для поиска минимального значения целевой функции. Достигнутые результаты подтверждают эффективность генетических алгоритмов в оптимизации сложных задач.

# Источники

1. Pandas библиотека: <https://pandas.pydata.org/>
2. Numpy библиотека: https://numpy.org/
3. CSV модуль: <https://docs.python.org/3/library/csv.html>
4. Concurrent Futures модуль: <https://docs.python.org/3/library/concurrent.futures.html>