**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

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

**ОТЧЁТ**

**По лабораторной работе № 1**

**По дисциплине “Алгоритмы и структуры данных”**

**На тему “Обезличивание датасета и вычисление k-anonymity ”**

**Студент гр. 24Б16-пу**

**Скородумов В.В.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

Оглавление

[**1.** **Цель работы** 3](#_Toc209554251)

[**2.** **Описание задачи** 3](#_Toc209554252)

[**3.** **Теоретическая часть** 3](#_Toc209554253)

[**4.** **Основные шаги программы** 3](#_Toc209554255)

[**5.** **Блок схема программы** 3](#_Toc209554256)

[**6.** **Описание программы** 4](#_Toc209554257)

[**7.** **Рекомендации пользователя** 5](#_Toc209554258)

[**8.** **Рекомендации программиста** 5](#_Toc209554259)

[**9.** **Контрольный пример** 6](#_Toc209554260)

[**10.** **Исследование** 6](#_Toc209554261)

[**Вывод** 6](#_Toc209554262)

1. **Цель работы**

Цель работы — описать и продемонстрировать программу обезличивания синтетического датасета кассовых чеков, имитирующего реальные данные, и вычисления значения K анонимности. Программа формирует обезличенные данные о покупках и пользователях, загружает подготовленные словари и проверяет корректность входных данных.

1. **Описание задачи**

1) Выбрать методы анонимизации для каждого параметра.

2) Разработать алгоритм анонимизации (Создать правила и зависимости между

признаками и целевой переменной, чтобы сохранять приемлимое К анонимити).

3) Реализовать программу обезличивания.

4) Протестировать результат (в результате должен быть сгенерирован обезличенный датасет, посчитано значение К анонимити, выведены 5 худших значений К и посчитано количество уникальных строк по заданным квази-идентификаторам.

5) Написать отчет (файл расширения .doc или .docx)

1. **Теоретическая часть**

Обезличивание основано на таких методах, как:… Програмная реализация выполнена на языке python. Используются библиотеки pandas – для работы с табличными данными, datetime – стандартная библиотека для работы с форматом даты и времени, Path – стандартная библиотека для работы с файловой системой. Полученная в итоге программа может обезличивать датасет входных данных, посчитать значение К анонимити, по заданным пользователем параметрам, вывести количество уникальных строк по заданным квази-идентификаторов.

1. **Основные шаги программы**

1. Загрузка таблицы входных данных.  
2. Валидация типов параметров входной таблицы.  
3. Обезличивание каждого столбца, по заданному для него способу.  
4. Загрузка пользовательского интерфейса для указания квази-идентификаторов.  
5. Вычисление К анонимити и количества уникальных строк.  
6. Запись обезличенного датасета в файл .xlsx.

1. **Блок схема программы**

Рис. 1 Блок-схема основной программы

1. **Описание программы**

Программная реализация состоит из модулей: dictionaries.py (хранение словарей) и anonimization.py (Обезличивание данных и вычисление К анонимити). Ниже приведена таблица всех функций.

|  |  |  |
| --- | --- | --- |
| Метод / Функция | Модуль | Назначение |
| Load\_table(path) | anonimization.py | Загрузка таблицы из Excel-файла по указанному пути. |
| table\_validate(table) | anonimization.py | Приведение типов данных, проверка и подготовка таблицы к обработке. |
| correct\_output(table) | anonimization.py | Переименование столбцов в итоговой таблице в стандартные имена. |
| export\_output(table, path) | anonimization.py | Экспорт обработанной таблицы в Excel-файл. |
| anonymize\_card\_number(card) | anonimization.py | Маскирование номера банковской карты (оставляет только первые 4 цифры). |
| anonymize\_date\_time(date\_time) | anonimization.py | Преобразование даты/времени в формат YYYY-MM для сокрытия точных временных данных. |
| anonymize\_store(store) | anonimization.py | Замена полного адреса магазина на его обобщённое название (по словарю anonymized\_stores из dictionaries.py). |
| anonymize\_coords(coords) | anonimization.py | Замена координат на соответствующий административный район (по словарю districts). |
| anonymize\_total\_cost(cost) | anonimization.py | Категоризация общей стоимости покупки по диапазонам. |
| anonymize\_num\_products(num) | anonimization.py | Преобразование количества товаров в интервальные категории. |
| anonymize\_price(price) | anonimization.py | Преобразование цены товара в интервальные диапазоны. |
| anonymize\_categories(cat) | anonimization.py | Замена конкретной категории товара на обобщённую категорию (по словарю categories). |
| anonymize\_brand(brand) | anonimization.py | Замена бренда на его ценовой сегмент (по словарю brands). |
| anonymize\_column(table, column) | anonimization.py | Применение соответствующей функции обезличивания к одному столбцу таблицы. |
| anonymize\_direct\_identifiers(table) | anonimization.py | Удаление прямых идентификаторов (например, receipt\_id) и маскирование номера карты. |
| get\_good\_k(table) | anonimization.py | Определение минимального допустимого значения k для оценки K-анонимности в зависимости от размера таблицы. |
| get\_k\_anonymity(table, quasi\_ids) | anonimization.py | Расчёт степени анонимности: процент строк, удовлетворяющих условию k-анонимности, и распределение по «плохим» k-группам. |
| full\_anonymization(table) | anonimization.py | Полная последовательная анонимизация всех столбцов датасета. |
| print\_result(fraction, k\_anonymity, bad\_k) | anonimization.py | Вывод результатов анализа K-анонимности в консоль. |
| get\_quasis(keys) | anonimization.py | Преобразование введённых пользователем номеров квази-идентификаторов в список названий столбцов. |
| user\_interface(table) | anonimization.py | Интерактивный интерфейс для выбора квази-идентификаторов и отображения метрик анонимности. |
| get\_anonymized\_columns(table, quasi\_ids) | anonimization.py | Получение уникальных комбинаций квази-идентификаторов и их количества. |
| main (блок if name == "main") | anonimization.py | Последовательный запуск всех этапов: загрузка данных, валидация, анонимизация, экспорт и анализ. |

# **Рекомендации пользователя**

Для запуска:  
1) Убедитесь, что в папке data репозитория находится файл table.xlsx с такими же столбцами как в примере.  
2) Запустите скрипт run\_anonymization.bat — он установит зависимости из requirements.txt, выполнит обезличивание и посчитает К анонимити для квази идентификаторов, которые вам нужно будет указать.  
3) Итоговый файл output.xlsx появится в папке output исходного проекта.

# **Рекомендации программиста**

Для корректного функционирования программы, убедитесь, что выполнены следующие пункты:

• Python 3.10+.  
•Установлены библиотеки pandas, openpyxl.  
• Формат таблиц входных данных должен соответствовать ожидаемым столбцам из примера в исходном репозитории.  
• Корректно указаны пути к файлам, директории уже существуют.  
• Корректный ввод пользователя в интерфейсе, рекомендуется добавить валидацию пользовательского ввода.

# **Контрольный пример**

1) Запуск run\_anonymization.bat.

2) Проверка наличия виртуального окружения в корне проекта и необходимых, создание окружения и загрузка библиотек в случае отсутствия.

3) Загрузка входных данных.

4) Валидация входных данных.

5) Обезличивание столбцов исходной таблицы по выбранным методам.

6) Запуск пользовательского интерфейса для указания квази-идентификаторов.

7) Вычисление К анонимити для указанных квази-идентификаторов и подсчет уникальных строк.  
8) В директории вывода появляется output.xlsx с обезличенными данными.

# **Вывод**

В рамках данной работы были изучены библиотеки pandas, datetime и реализованы функции обезличивания датасета.

В итоге было написано полностью работоспособное, настраиваемое програмное обеспечение для обезличивания датасета и вычисления К анонимити по заданным квази-идентификаторам.