**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 данных»**

**Студент гр. 23Б15-пу**

**Бек В.А.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2024 г.**

Оглавление

1. [Цель работы 3](#_Toc179124454)
2. [Описание задачи (формализация задачи) 4](#_Toc179124455)
3. [Теоретическая часть 5](#_Toc179124456)

[**Основные компоненты программы** 5](#_Toc179124457)

1. [Основные шаги программы 7](#_Toc179124458)
2. [Блок схема программы 10](#_Toc179124459)
3. [Описание программы 12](#_Toc179124460)
4. [Рекомендации пользователя 14](#_Toc179124461)
5. [Рекомендации программиста 15](#_Toc179124462)
6. [Исходный код программы 15](#_Toc179124463)
7. [Контрольный пример 16](#_Toc179124464)
8. [Вывод 21](#_Toc179124465)
9. [Источники 22](#_Toc179124466)

# Цель работы

**Цель работы** заключается в разработке программы, способной эффективно обезличивать данные о покупках в магазинах и оценивать степень анонимности с помощью метода К-анонимности, для указания на уязвимости обезличенных данных. Также программа должна позволять пользователю выбирать квази-идентификаторы и обеспечивать гибкость при применении различных методов обезличивания.

# Описание задачи (формализация задачи)

* Программа должна считывать входной файл, который является результатом первой лабораторной работы — сгенерированный датасет с имитацией покупок в магазинах. Основная задача программы — обезличивание данных этого датасета.
* Программа должна быть разделена на модули, где каждый отвечает за отдельный функционал:
  + Обезличивание входного набора данных с использованием различных методов (локальное обобщение, маскеризация, и т.д.).
  + Вычисление степени К-анонимности для входного набора данных.
* Пользователь может указывать квази-идентификаторы, на основе которых будет проводиться обезличивание и оценка К-анонимности.
* После обезличивания программа должна рассчитать показатель К-анонимности для каждого из наборов данных и вывести "плохие" значения (с низким К).
* Программа должна анализировать уникальные строки в наборе данных и выводить их в случае, если К = 1.

# Теоретическая часть

В данной работе реализована программа для обработки и обезличиванию датасета с использованием различных методов защиты данных. К-анонимность — это метод анонимизации, который обеспечивает, что каждая запись в анонимизированном наборе данных не может быть однозначно определена среди других K записей. Это уменьшает вероятность утечки конфиденциальной информации.

### **Основные компоненты программы**

* **Загрузка и обработка данных:** Программа считывает входные файлы в формате xlsx, содержащие информацию о покупках, такие как название магазина, дата и время транзакции, координаты, категория товара, бренд, номер карточки, количество товаров и стоимость.
* **Обезличивание данных:** Для защиты конфиденциальности применяется ряд методов анонимизации. Пользователь может выбрать набор квази-идентификаторов — атрибутов, которые могут быть использованы для идентификации личности, но не являются уникальными. Основные методы, реализованные в программе:
  + **Локальное обобщение:** Детализированные данные заменяются более общими значениями. Например, дата транзакции может быть преобразована в сезон (зима, весна, лето, осень), стоимость товара и количество — в диапазон, а координаты заменяются в район (Центральный район, Северный район, Южный район).
  + **Маскеризация:** Чувствительные данные, такие как бренды и или номера карточек, скрываются частично или полностью. Например, номера карточек заменяются на название банка, а бренды на \*\*\*\*\*\*\*.
  + **Локальное подавление:** Удаление записей, которые остаются уникальными или имеют низкий показатель K, для соблюдения К-анонимности.
* **Расчет К-анонимности:** К-анонимность является показателем степени обезличивания набора данных. В процессе работы программа объединяет записи по выбранным квази-идентификаторам и подсчитывает количество записей в каждой группе. Минимальное количество записей в группе устанавливает значение K. Кроме того, программа выводит топ 5 "плохих" значений K, чтобы наглядно отобразить уровень К-анонимности после проведения анонимизации.
* **Полезность данных:** Хоть некоторые данные и маскеризуются полностью, полученные данные всё равно остаются не слишком бесполезными, так как сохраняется их общая структура и обобщённая информация. Это позволяет проводить анализ тенденций и закономерностей, не раскрывая при этом личные данные. Например, собранные данные о покупках могут помочь в выявлении популярных категорий товаров или сезонных колебаний спроса. Важно, что такие данные могут использоваться для принятия обоснованных решений, не нарушая конфиденциальность клиентов.

Таким образом, программа не только обеспечивает защиту конфиденциальной информации, но и сохраняет полезность данных для дальнейшего анализа и обработки.

# Основные шаги программы

* **Инициализация графического интерфейса**:

Программа начинается с вызова функции setup\_window(), которая создает главное окно с заголовком "K-anonymity". В окне устанавливаются размеры (ширина 480 пикселей и высота 260 пикселей), и оно центрируется на экране, используя вычисления для позиции окна.

* **Создание пользовательских элементов интерфейса**:

Функция create\_widgets() отвечает за создание основных элементов интерфейса:

* + Создается основной фрейм, который упрощает размещение других виджетов.
  + Внутри этого фрейма добавляется LabelFrame, который содержит чекбоксы для выбора квази-идентификаторов. Пользователь может выбрать атрибуты, которые будут использованы для анонимизации и подсчёта k-anonymity:
    - "Название магазина"
    - "Дата и время"
    - "Координаты"
    - "Категория"
    - "Бренд"
    - "Номер карточки"
    - "Количество товаров"
    - "Стоимость"
  + Создается другой LabelFrame, отображающий таблицу с заголовком "Топ плохих k-anonymity". В таблице будут выводиться результаты расчета K-анонимности.
  + Также создается фрейм для размещения кнопок управления:
    - Кнопка "Выбрать файл" для выбора файла Excel с данными.
    - Кнопка "Рассчитать k-anonymity" для начала процесса анализа данных.
    - Кнопка "Обезличить" для анонимизации выбранных данных.
    - Кнопка "Выход" для завершения работы программы.
* **Выбор файла**:

При нажатии на кнопку "Выбрать файл" вызывается функция select\_file(), которая открывает диалоговое окно для выбора файла xlsx. Путь к выбранному файлу сохраняется в переменной file\_path, что позволяет программе впоследствии работать с данными из этого файла.

* **Расчет K-анонимности**:

По нажатию на кнопку "Рассчитать k-anonymity" запускается функция calculate(), которая выполняет следующие шаги:

* + Сначала очищается таблица от предыдущих результатов.
  + Из полученного списка квази-идентификаторов извлекаются выбранные пользователем.
  + Открывается файл xlsx, и данные считываются в виде списка записей.
  + Программа формирует комбинации записей на основе выбранных идентификаторов.
  + Если количество записей в группе меньше 10, такие записи удаляются из исходного файла.
  + Программа также собирает уникальные комбинации, у которых K = 1, и выводит их в таблицу.
  + Результаты расчета K-анонимности (значения K и количество таких записей) отображаются в таблице, информируя пользователя о качестве анонимизации.
* **Обезличивание данных**:

При нажатии на кнопку "Обезличить" запускается функция anonymize(), которая выполняет следующие действия:

* + Открывается файл с исходными данными и создается новый файл для сохранения обезличенных данных.
  + В зависимости от выбранных квази-идентификаторов, программа обрабатывает каждую строку данных:
    - Если выбран "Название магазина", программа заменяет его на категорию, к которой он принадлежит.
    - Для "Дата и время" вычисляется сезон, в который попадает дата.
    - Для "Координаты" преобразуются в названия районов.
    - Бренды заменяются на маску "\*\*\*\*\*\*\*\*\*\*\*\*\*\*".
    - Номера карточек обрабатываются, и выводится название банка на основе цифр со второй по шестую.
    - Для "Количество товаров" и "Стоимость" устанавливаются диапазоны, что помогает сохранить общий вид данных без раскрытия конкретной информации.
  + Все изменения сохраняются в новом файле anonymized\_shopping\_data.xlsx.
* **Сохранение обезличенных данных**:

После завершения анонимизации программа сохраняет изменения в новом файле и выводит сообщение об успешном завершении процесса, информируя пользователя о том, что данные были успешно обезличены.

# Блок схема программы

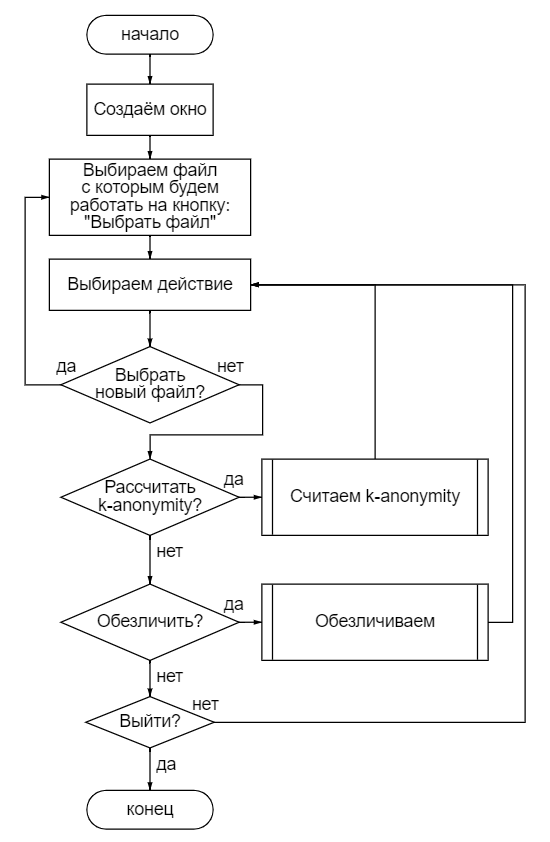


Рис 1. Блок-схема основной программы

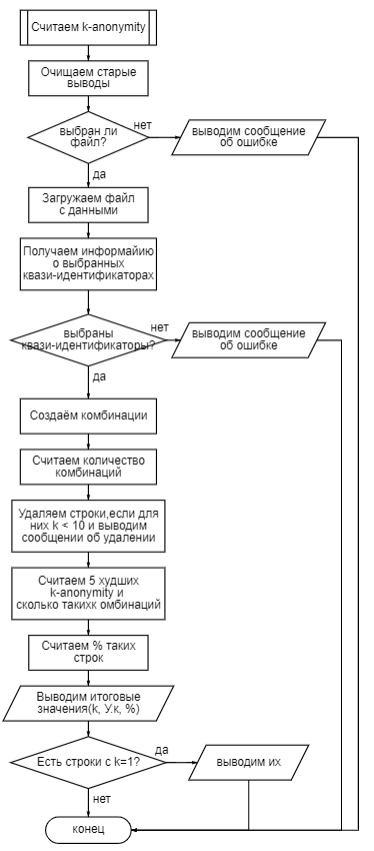


Рис 2. Блок-схема подпрограммы 1

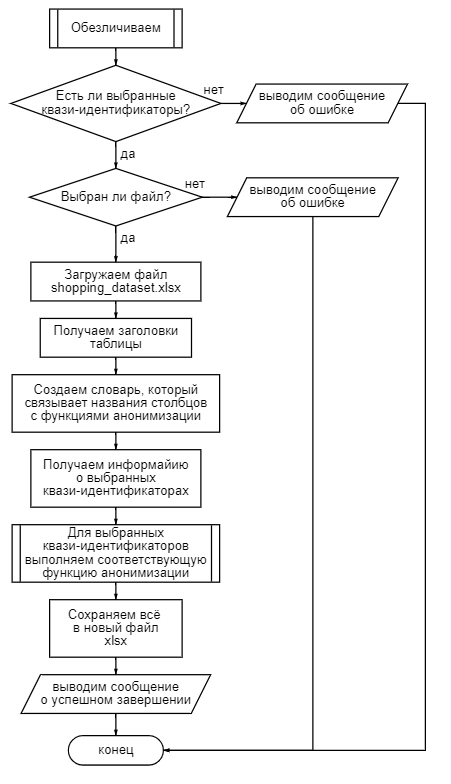


Рис 3. Блок-схема подпрограммы 2

# Описание программы

Программная реализация написана на языке Python 3.12.7 с использованием следующих библиотек: tkinter, openpyxl, collections, datetime и dateutil. Программа организована в едином модуле, который фокусируется на анонимизации данных о покупках и подсчете их K-анонимности. В процессе разработки программы использовалось 13 функций, каждая из которых имеет четко определенное назначение:

Таблица 1. shopping\_dataset\_anonymization\_and\_k\_anonymity\_tool.py

|  |  |  |
| --- | --- | --- |
| Функция | Описание | Возвращаемое значение |
| setup\_window | Настройка главного окна приложения, установка заголовка и размеров. | root |
| create\_widgets | Создание элементов интерфейса: чекбоксов для выбора квази-идентификаторов, кнопок для управления. | None |
| select\_file | Открытие диалогового окна для выбора файла Excel с данными о покупках. | None |
| calculate | Расчет K-анонимности, определение уникальных записей и группировка данных. | None |
| anonymize | Анонимизация данных, замена персональной информации на общие значения. | None |
| anonymize\_store | |  | | --- | |  |  |  | | --- | | Обобщение названия магазина, замена на категорию. | | None |
| |  | | --- | | anonymize\_time |  |  |  | | --- | --- | |  |  | | |  | | --- | |  |  |  | | --- | | Преобразование даты и времени в строку с указанием сезона. | | None |
| anonymize\_coordinates | |  | | --- | |  |  |  | | --- | | Преобразование координат в названия районов. | | None |
| anonymize\_category | Обобщение категории товара, замена на общую категорию. | None |
| anonymize\_brand | Маскирование информации о бренде, замена на "\*\*\*\*\*\*\*\*\*\*\*\*\*\*". | None |
| anonymize\_card | Маскирование номера карточки, замена на название банка на основе BIN-кода. | None |
| anonymize\_quantity | Обобщение количества товаров, установка диапазонов. | None |
| anonymize\_cost | Обобщение стоимости, установка диапазонов. | None |

# Рекомендации пользователя

При работе с программой для анонимизации данных и расчета K-анонимности рекомендуется соблюдать следующие шаги и советы для эффективного использования приложения:

* **Выбор файла**: Для начала работы необходимо выбрать файл формата xlsx, содержащий данные для анализа. Это можно сделать, нажав на кнопку "Выбрать файл". Убедитесь, что данные в файле организованы корректно, с указанием всех нужных атрибутов (например, название магазина, дата и время, стоимость и т.д.).
* **Выбор квази-идентификаторов**: После загрузки файла выберите квази-идентификаторы — атрибуты, которые могут использоваться для анонимизации данных. Например, можно выбрать такие поля, как "Название магазина", "Дата и время", "Координаты", которые помогут сгруппировать данные и обезличить их.
* **Обезличивание данных**: После выбора квази-идентификаторов нажмите кнопку "Обезличить", чтобы программа выполнила анонимизацию данных. Информация будет обобщена или замаскирована в зависимости от выбранных полей. Например, даты преобразуются в сезоны, бренды маскируются, а номера карт заменяются названиями банков.
* **Выбор обезличенного файла**: Теперь вы можете выбрать новый, обезличенный файл: anonymized\_shopping\_data.xlsx. Это можно сделать при помощи всё той же кнопки "Выбрать файл".
* **Расчет K-анонимности**: Для оценки текущего уровня анонимности набора данных нажмите кнопку "Рассчитать k-anonymity". Программа проанализирует данные и определит уникальные записи. Записи где K-анонимность ниже порогового значения (K < 10) программа удалит. Удаление таких записей повысит уровень конфиденциальности.

# Рекомендации программиста

**Установка необходимых библиотек:**

* Перед запуском программы убедитесь, что все необходимые библиотеки установлены. В данном случае вам потребуется библиотеки **openpyxl, tkinter** и **dateutil**. Установите их с помощью команд:





**Подготовка файлов данных:**

* Убедитесь, что существует файл shopping\_dataset.xlsx который содержит корректные данные с имитацией покупок в магазинах.

**Код программы:**

**<https://github.com/Kliooo/Algorithms-and-data-structures>**

# Контрольный пример

* Запуск программы: Для запуска программы используйте файл **shopping\_dataset\_anonymization\_and\_k\_anonymity\_tool.py.** Программа запустит графический интерфейс (Рис. 4), в котором можно будет выбрать “Выбрать файл” или “Рассчитать k-anonymity”, “Обезличить” и “Выход”, предварительно выбрав квази-идентификаторы в столбике слева.

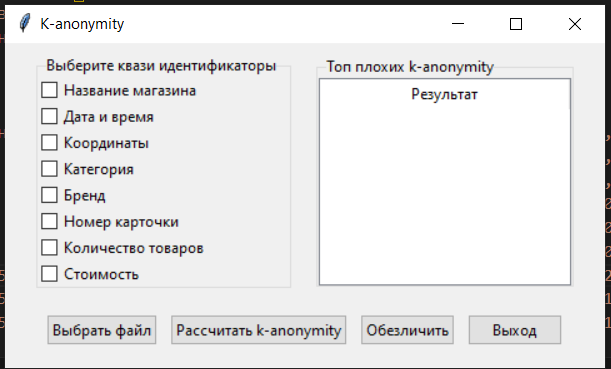


Рис 4. Графический интерфейс программы

* Выбор файла: Первым делом нам надо выбрать файл для работы. При нажатии на копку: “Выбрать файл” программа откроет проводник, где пользователю предложено выбрать xlsx файл (Рис.5).

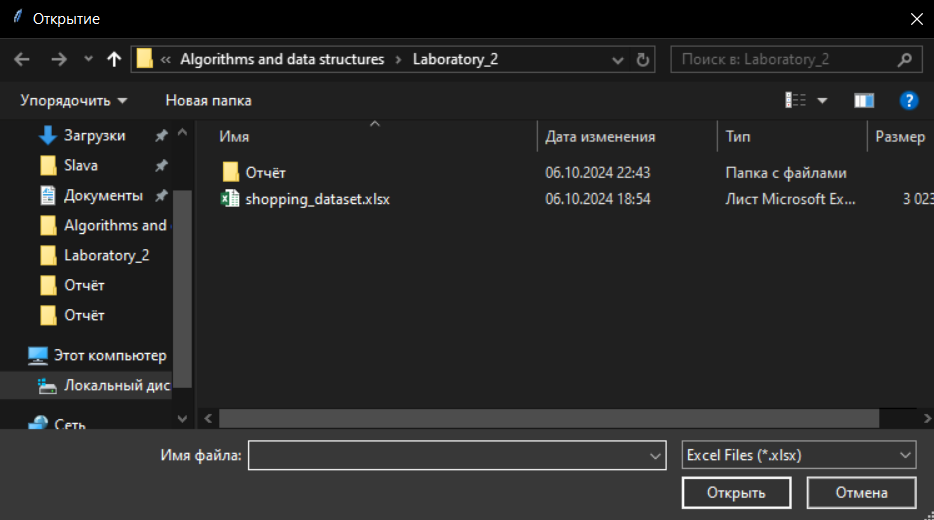


Рис 5. Выбор файла

* Обезличить датасет: При выборе квази-идентификаторов и нажатии кнопки “Обезличить датасет” программа начнёт обезличивание файла, который мы выбрали нажатием на кнопку: “Выбрать файл”. И по результату создаст: anonymized\_shopping\_data.xlsx.
* Подсчёт k-anonymity: Первым делом нам надо снова выбрать файл, с которым будем работать на кнопку “Выбрать файл”. Нам нужен: anonymized\_shopping\_data.xlsx (Рис.6). Теперь, после выбора квази-идентификаторов и нажатия на кнопку: “Рассчитать k-anonymity” начнётся процесс и через несколько секунд в таблице справа появятся топ 5 худших k-anonymity, количество групп с таким k-anonymity и их процент от общего числа.(Рис. 7).

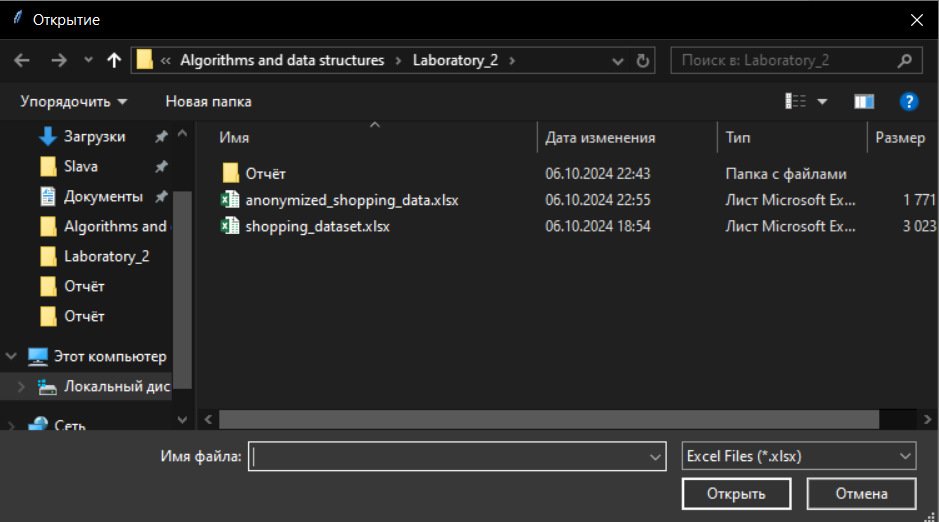


Рис 6. Выбор обезличенного файла

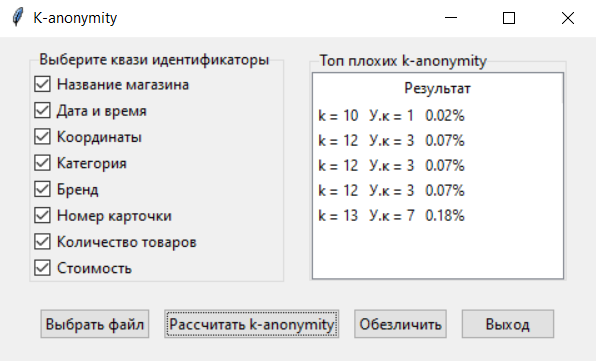


Рис 7. Ввода после подсчета k-anonymity

# Вывод

В ходе выполнения данной работы была разработана программа, обеспечивающая анонимизацию данных, что является важным аспектом в современных условиях защиты личной информации. Основное внимание было уделено внедрению методов k-анонимности, которые позволяют минимизировать риск идентификации индивидуумов при анализе агрегированных данных.

В процессе работы над проектом были проанализированы и реализованы различные подходы к обработке квази-идентификаторов, таких как название магазина, дата и время покупки, координаты и данные банковских карт. Программа позволяет пользователям легко взаимодействовать с интерфейсом и эффективно выполнять необходимые операции по анонимизации.

# Источники

* Редактор блок-схем.

[*https://programforyou.ru/block-diagram-redactor*](https://programforyou.ru/block-diagram-redactor)

*дата обращения: (05.10.2024)*

* dateutil — Упрощение работы с датами и временем. [*https://dateutil.readthedocs.io/en/stable/*](https://dateutil.readthedocs.io/en/stable/)  
  дата обращения: (02.10.2024)
* tkinter — Библиотека для создания графических интерфейсов в Python.

[*https://docs.python.org/3/library/tkinter.html*](https://docs.python.org/3/library/tkinter.html%20)

*дата обращения: (02.10.2024)*

* collections — Контейнерные данные, предоставляемые стандартной библиотекой Python.

[*https://docs.python.org/3/library/collections.html*](https://docs.python.org/3/library/collections.html%20)

*дата обращения: (02.10.2024)*