Санкт-Петербургский государственный университет

Факультет прикладной математики – процессов управления

Лабораторная работа №2

**Отчет**

по дисциплине «Алгоритмы и структуры данных»

**Создании программы для обезличивания датасета**

Автор работы: Дацык Р.В.

Группа: 22.Б15-пу

Преподаватель: Дик А.Г.

Санкт-Петербург, 2023

**Оглавление**

[**1.** **Цель работы** 3](#_Toc152061026)

[**2.** **Задачи** 3](#_Toc152061027)

[**3.** **Введение** 3](#_Toc152061028)

[**4.** **Теоретическая часть** 3](#_Toc152061029)

[**5.** **Алгоритм метода** 3](#_Toc152061030)

[**6.** **Описание программы** 5](#_Toc152061031)

[**7.** **Рекомендации пользователю** 7](#_Toc152061032)

[**8.** **Рекомендации программисту** 7](#_Toc152061033)

[**9.** **Контрольный пример** 8](#_Toc152061034)

[**10.** **Вывод** 9](#_Toc152061035)

[**11.** **Список использованной литературы** 9](#_Toc152061036)

# **Цель работы**

Разработка и применение методов анонимизации для защиты личных данных пользователей, а также оценка K-анонимности полученных данных.

# **Задачи**

1. Изучить методы анонимизации данных.
2. Проанализировать их достоинства и недостатки.
3. Реализовать анонимизацию данных.
4. Рассчитать К-анонимность.
5. Проанализировать исходные и обезличенные данные.
6. Сохранение обезличенных данных.

# **Введение**

В современном обществе, где активно развиваются цифровые технологии и сетевые коммуникации, сбор и анализ данных стали неотъемлемой частью многих сфер деятельности. Тем не менее, обеспечение конфиденциальности и безопасности личных данных пользователей остается первостепенной задачей. В этом контексте ключевым является проведение процесса анонимизации данных, чтобы исключить возможность идентификации конкретных личностей.

# **Теоретическая часть**

Генерация персональных данных также имеет важные особенности. Во-первых, в зависимости от определенного сезона, должна показываться ориентированная реклама, значит для каждого сезона должен быть свой словарь с видами рекламы. Во-вторых, время просмотра высчитывается по определенной формуле, которую предоставляет заказчик.

# **Алгоритм метода**

1. Импорт библиотек и данных
2. Применение различных методов анонимизации данных
3. Расчет К-анонимности данных
4. Сравнение уникальных значений исходных и анонимизированных данных
5. Отображение результатов К-анонимности, статистики данных и уникальных сток с K=1
6. Сохранение анонимизированных данных

На рисунке 5.1 представлена блок-схема алгоритма.

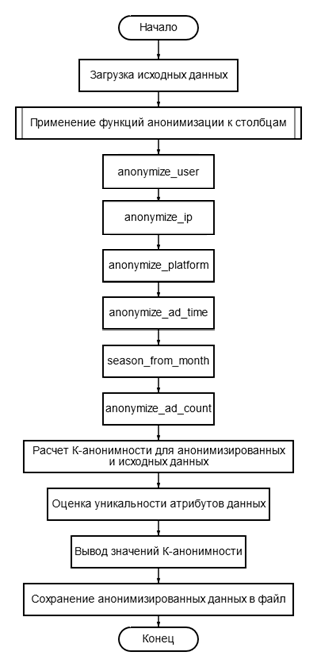


Рисунок 5*.*1 Блок-схема программы.

# **Описание программы**

Алгоритм реализован на языке python 3.10 с использованием следующих пакетов: pandas, random, datetime.

В программе используются 10 функций: 9 связанных с обработкой данных и одна для обработки интерфейса. В таблице 6.1 представлено описание всех функций.

Таблица 6.1. Описание функций

|  |  |  |  |
| --- | --- | --- | --- |
| Имя функции | Аргумент | Возвращаемое значение | Условия формирования данных |
| anonymize\_ip(ip\_address) | От 8 до 12 цифр, составляющих ip-адрес | От 8 до 12 цифр, составляющих ip-адрес | Функция для маскировки IP-адресов, заменяющая последние два октета на 'X'. |
| anonymize\_user(user) | Имя пользователя | Анонимное имя пользователя | Функция для анонимизации имен пользователей, заменяющая имя на "user\_" с первой цифрой из исходного имени. |
| convert\_time\_to\_float(time\_str) | Время просмотра в формате строк | Время просмотра в минутах | Функция для конвертации времени в минутах в числовой формат с плавающей точкой. |
| anonymize\_ad\_time(ad\_time) | Время просмотра в минутах | Строка  'до 600' или 'больше 600' минут. | Функция для анонимизации времени просмотра рекламы, разделяющая его на две категории: 'до 600' и 'больше 600' минут. |
| anonymize\_platform(platform) | Платформа просмотра | Домен платформы | Функция для анонимизации информации о платформе, оставляющая только домен верхнего уровня. |
| season\_from\_month(month) | Порядковый  номер месяца | Время года | Функция для определения сезона на основе номера месяца, возвращающая 'весна', 'лето', 'осень' или 'зима'. |
| anonymize\_ad\_count(ad\_count) | Количество просмотренной рекламы | Строка  'до 100' и 'больше 100'  реклам | Функция для анонимизации количества рекламы, разделяющая его на две категории: 'до 100' и 'больше 100'. |
| calculate\_k\_anonymity(data, quasi\_identifiers) | Данные из таблицы, квази-идентификаторы | Значение К-анонимности | Функция для расчета K-анонимности на основе заданных квази-идентификаторов. |
| data\_usefulness(original\_data, anonymized\_data) | Данные изначальной таблицы,  Данные анонимизированной таблицы | Уникальные значения атрибутов | Функция для оценки полезности данных путем сравнения уникальных значений различных атрибутов между исходными и анонимизированными данными. |
| main() | None | Таблица Excell | Основная функция, выполняющая все шаги программы: чтение данных, анонимизацию, расчет K-анонимности, оценку полезности данных и сохранение результатов в новый файл. |

# 

# **Рекомендации пользователю**

Для успешного запуска программы необходимо устройство с операционной системой Linux, macOS или Windows, а также среда разработки, поддерживающей запуск python 3.10

Для запуска генерации достаточно запустить файл с кодом. Как только таблица будет создана, появится уведомление об успешном завершении. Файл с данными будет сохранен в директорию, где находится код программы.

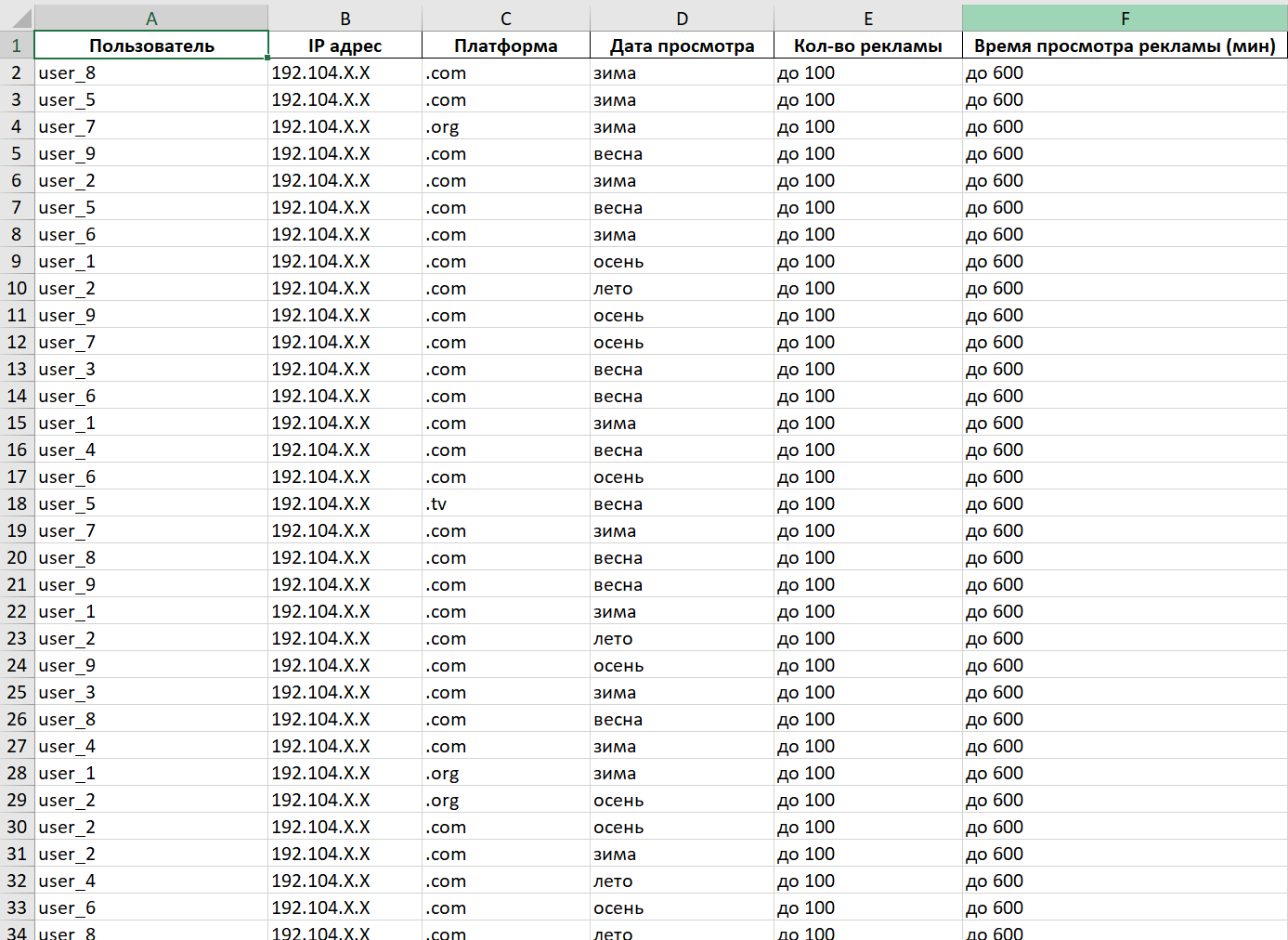
# **Рекомендации программисту**

Для запуска программы необходима 64-битная операционная система Windows, Linux или macOS. Для работы с кодом необходима среда разработки, совместимая с python 3.10 и библиотеки pandas, random, datetime.

Исходный код программы и необходимые текстовые файлы доступны по ссылке:

[https://github.com/CapTopGrade/Algorithms/blob/main/2%20lab.py](https://github.com/CapTopGrade/Algorithms/blob/main/Alg1.py)

# **Контрольный пример**

Пример полученной таблицы, а также вывода программы изображены на следующих рисунках. Рисунок 9.1 Пример полученной таблицы

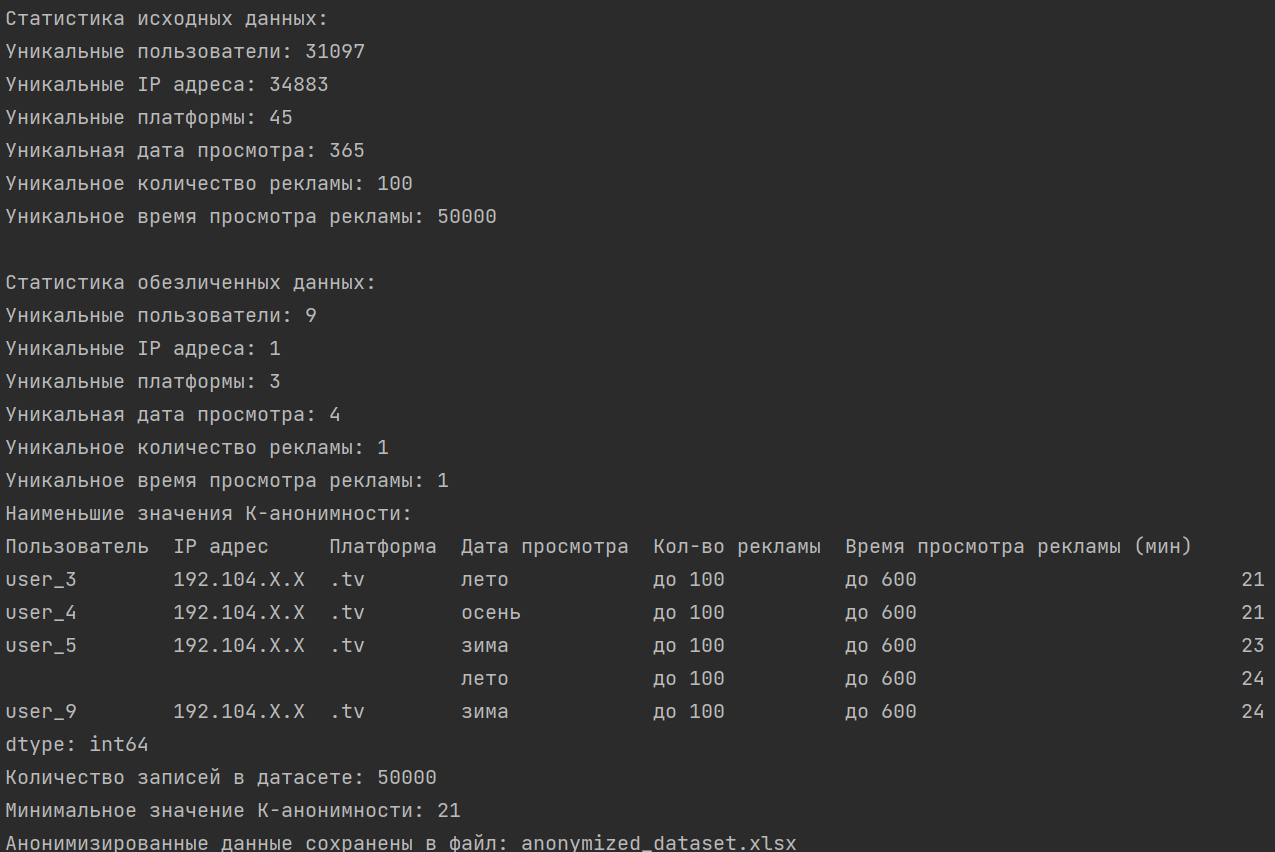


Рисунок 9.2 Пример вывода программы

# **Вывод**

Была реализована программа для обезличивания данных, обеспечивая безопасность и конфиденциальность информации. Стало возможным анализировать на каких платформах обычно смотрят рекламу и в каком количестве. В ее основе лежат несколько алгоритмов анонимизации: маскеризация, локальное обобщение, удаление атрибутов и создание псевдонимов. Квази-идентификаторами являются: 'Пользователь', 'IP адрес', 'Платформа', 'Дата просмотра', 'Кол-во рекламы', 'Время просмотра рекламы (мин)'.

Важным аспектом в рамках программы является оценка уровня K-анонимности, которая измеряет степень анонимизации как для обезличенных, так и для исходных данных. Этот процесс позволяет оценить, насколько успешно обезличенные данные убирают личную информацию, обеспечивая при этом безопасность. Результаты процедуры анонимизации сохраняются в новом файле формата Excel, что упрощает последующий анализ и использование данных.

# **Список использованной литературы**

[1] Документация библиотеки Pandas: https://pandas.pydata.org/docs/

[2] Документация по работе с датами и временем в Python: <https://docs.python.org/3/library/datetime.html>