**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**Факультет прикладной математики – процессов управления**

**отчет**

**по лабораторной работе**

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

**на тему «Обезличивание данных»**

**1 вариант**

|  |  |  |
| --- | --- | --- |
| Студент гр. 22.Б15 |  | Агишев А.Б. |
| Преподаватель |  | Дик А.Г. |

**Санкт-Петербург**

**2023 г.**

**Содержание**

[1. Цель работы 2](#_Toc149058268)

[2. Задача 2](#_Toc149058269)

[3. Теоретическая часть 2](#_Toc149058270)

[4. Алгоритм метода 3](#_Toc149058271)

[5. Описание программы 5](#_Toc149058272)

[5.1 Описание функций 5](#_Toc149058273)

[6. Рекомендации пользователю 8](#_Toc149058274)

[7. Рекомендации программисту 8](#_Toc149058275)

[8. Контрольный пример 9](#_Toc149058276)

[9. Заключение 11](#_Toc149058277)

# **Цель работы**

Разработать алгоритм и программу, которая обезличивает набор синтетических данных.

# **Задача**

Приобрести и закрепить знания в области обезличивания данных. Обезличить набор синтетических данных, которые обладают следующими наборами свойств: ФИО, номер телефона, адрес работы, должность, З/П. Рассчитать k-anonymity для обезличенного набора. Достичь приемлемого значения k-anonymity.

# **Теоретическая часть**

*Обезличивание данных* — это процесс обработки персональных данных с целью удаления или замены всех прямых или косвенных идентификационных характеристик, которые могут связать данные с конкретным человеком.

*Методы обезличивания* — техники, применяемый к персональным данным с целью снижения риска идентификации субъекта персональных данных. В результате применения соответствующего метода из набора данных удаляется идентифицирующая личность человека информация. Оценка риска обезличенных данных производится для каждого набора данных и каждого метода. Все методы являются параметризуемыми, то есть содержат настраиваемые характеристики (такие как набор квази-идентификаторов), определяющие конкретный алгоритм в отношении конкретного набора данных в заданных условиях (сценариях и контексте).

*K-анонимность (k-anonymity)* — концепция в области обезличивания данных, представляющая собой методологию, при которой каждая запись в наборе данных неотличима от по крайней мере k-1 других записей по определенному набору атрибутов. Этот подход направлен на устранение уникальных идентификационных признаков и сведение к минимуму риска нежелательного раскрытия личной информации при анализе данных.

# **Описание алгоритма**

1. Считывание файла для обработки и квази-идентификаторов в k-anonymity из поля ввода интерфейса.
2. При нажатии кнопки *«Обезличивание данных»*:
   1. Проверка на корректность выбора файла. Если файл некорректен или отсутствует, то появляется текст ошибки.
   2. Считывание исходного файла, а также файла с соответствием улиц Санкт-Петербурга к районам *«saint\_petersburg\_districts»*.
   3. Применение метода “удаление атрибутов” для ФИО, метода “маскеризация” для номеров телефонов, метода “локальное обобщение” для адресов работ и З/П.

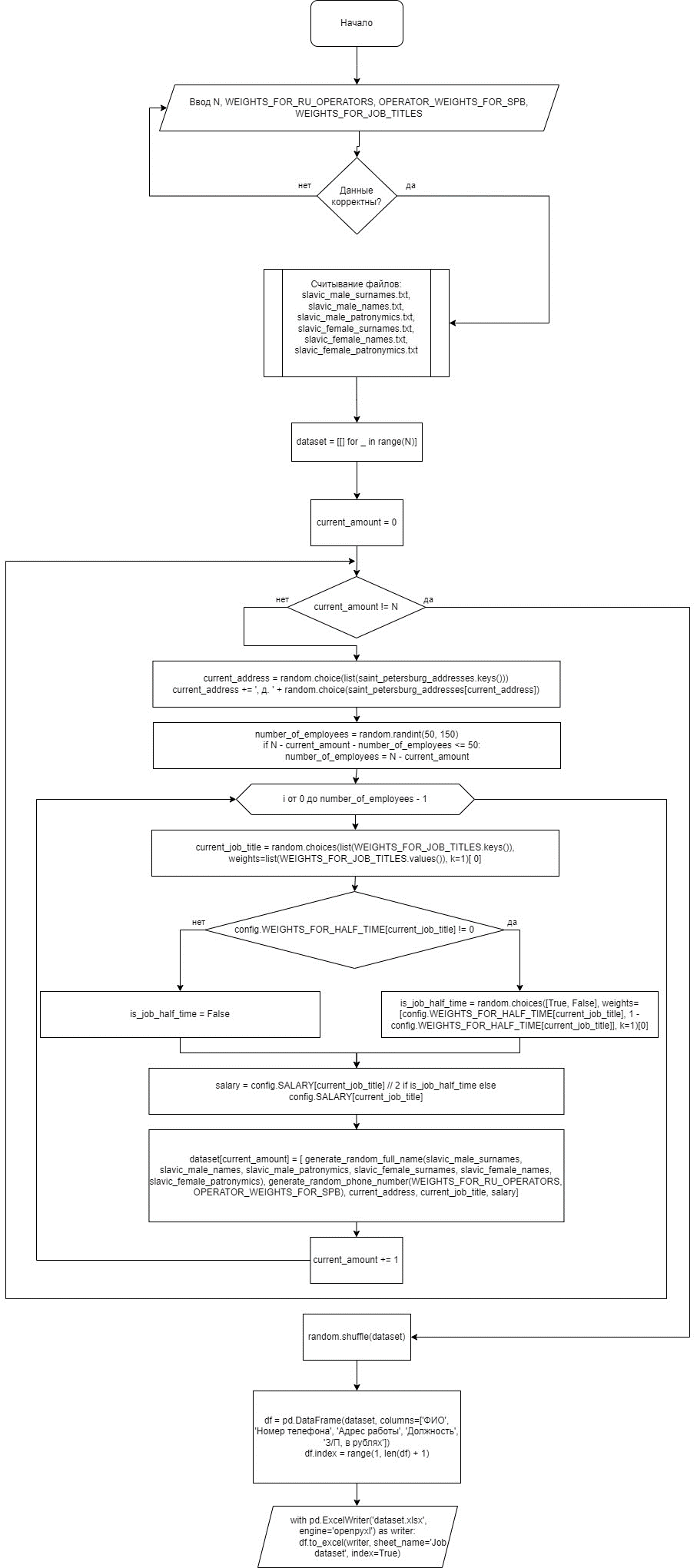


Рисунок 4.1. Блок-схема алгоритма генерации датасета.

# **Описание программы**

## **Описание функций**

В программе используется 6 функций: 2 связанных с интерфейсом и 3 – с генерацией данных. В таблице 5.1 представлено описание функций.

*Таблица 5.1. Описание функций*

|  |  |  |
| --- | --- | --- |
| Имя функции | Входные данные | Описание функции |
| Main | — | Создание интерфейса, вызов функции *«threaded\_run\_generator»* |
| threaded\_run\_generator | — | Запуск функции *«run\_generator»* в отдельном потоке |
| run\_generator | — | Проверка на корректность введенных данных, вызов функции *«generate\_dataset»* |
| generate\_dataset | Размерность, статус интерфейса, вероятностные коэффициенты выбора оператора связи, вероятностные коэффициенты выбора номеров операторов с домашним регионом Санкт-Петербург, вероятностные коэффициенты выбора работы на полставки | Генерация базы данных |
| generate\_random\_full\_name | Славянские мужские и женские ФИО | Генерация случайного ФИО |
| generate\_random\_phone\_number | Вероятностные коэффициенты выбора оператора связи, вероятностные коэффициенты выбора номеров операторов с домашним регионом Санкт-Петербург | Генерация случайного номера телефона |

Описание переменных функции *«generate\_dataset»* представлено в таблице 5.2.

*Таблица 5.2. Описание переменных функции «generate\_dataset»*

|  |  |  |
| --- | --- | --- |
| Тип | Имя | Описание |
| str[] | slavic\_male\_surnames | Список славянских мужских фамилий |
| str[] | slavic\_male\_names | Список славянских мужских имён |
| str[] | slavic\_male\_patronymics | Список славянских мужских отчеств |
| str[] | slavic\_female\_surnames | Список славянских женских фамилий |
| str[] | slavic\_female\_names | Список славянских женских имён |
| str[] | slavic\_female\_patronymics | Список славянских женских отчеств |
| str[] | saint\_petersburg\_addresses\_list | Содержимое файла *«saint\_petersburg\_addresses.txt»* |
| dict | saint\_petersburg\_addresses | Словарь всех улиц со всеми домами |
| str[][] | Dataset | Итоговая база данных |
| int | current\_amount | Текущее количество строк в базе |
| str | current\_address | Текущий адрес |
| int | number\_of\_employees | Количество работников, работающих на одном адресе |
| str | current\_job\_title | Текущая профессия |
| bool | is\_job\_half\_time | Работа на полставки |
| int | Salary | Зарплата |
| pd.DataFrame | Df | Итоговая таблица сгенерированных данных |
| dict | column\_widths | Название колонок в итоговой таблице и их ширина |
| worksheet | Worksheet | Название листа в Excel |

# **Рекомендации пользователю**

Поле *«Введите число N»* устанавливает количество записей в базе данных. По умолчанию начальное значение 50000.

Поля с заданием настроек выбора вероятностей случайных величин модифицируемые. По умолчанию указаны корректные величины. При неправильных настройках поля подсвечиваются красным цветом.

Кнопка *«Генерировать»* отвечает за запуск генерации. Программа выполнит генерацию синтетических данных на основе заданных настроек.

Данные сохраняются в файл *«dataset.xlsx»* в директории, хранящей в себе программу.

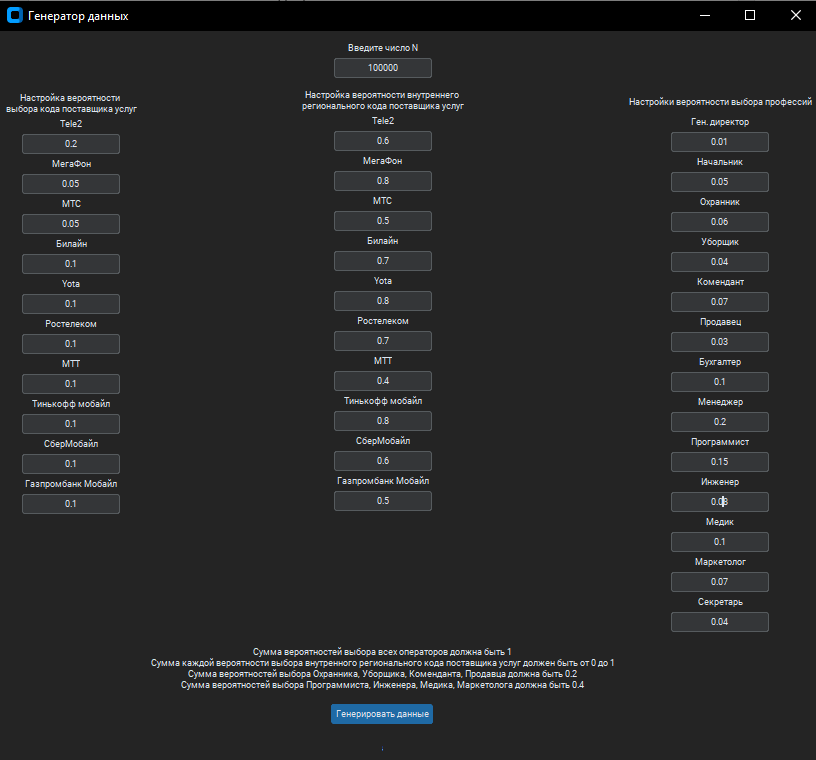
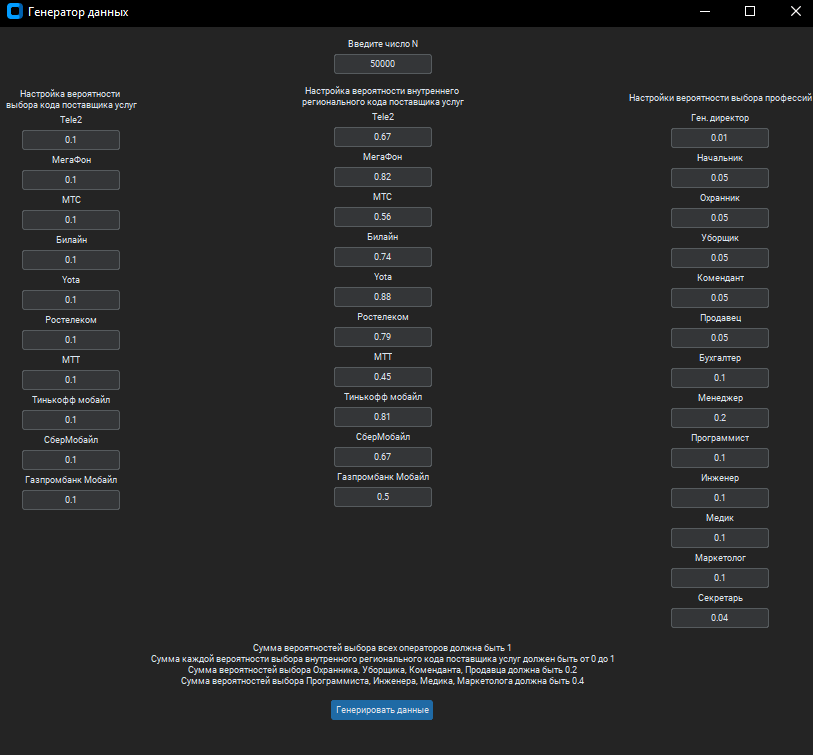
Для завершения работы нажмите на крестик в левом верхнем углу.

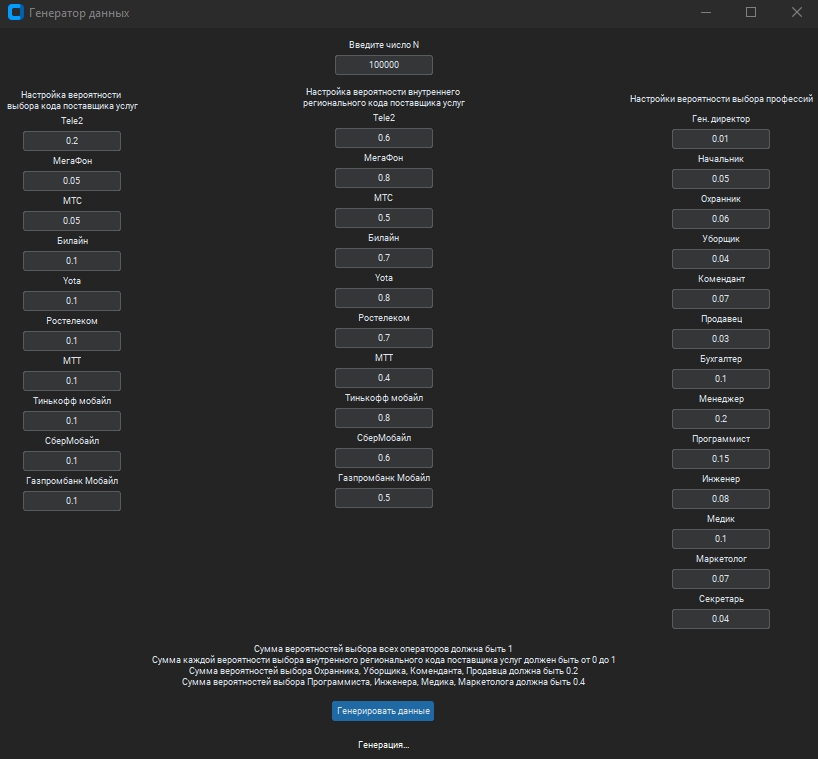
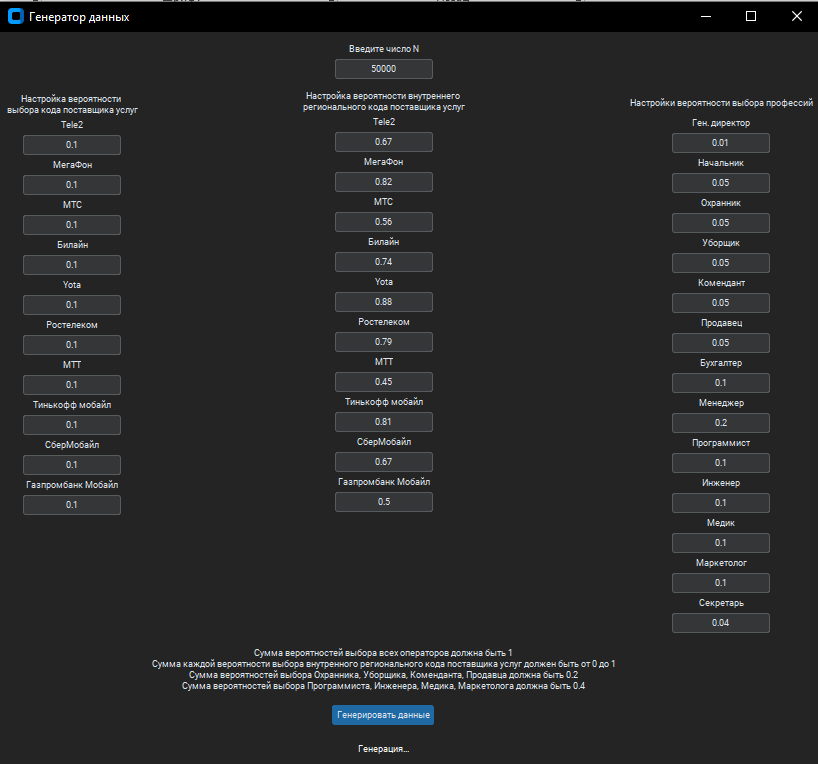
# **Рекомендации программисту**

Для запуска программы необходим Python версии не ниже 3.10.6, а также 64-битная операционная система Windows/Linux/macOS. Предварительно необходимо установить библиотеки: customtkinter версии не ниже 5.2.0, pandas версии не ниже 2.1.0. Для работы с кодом необходим PyCharm версии не ниже 2022.2.1.

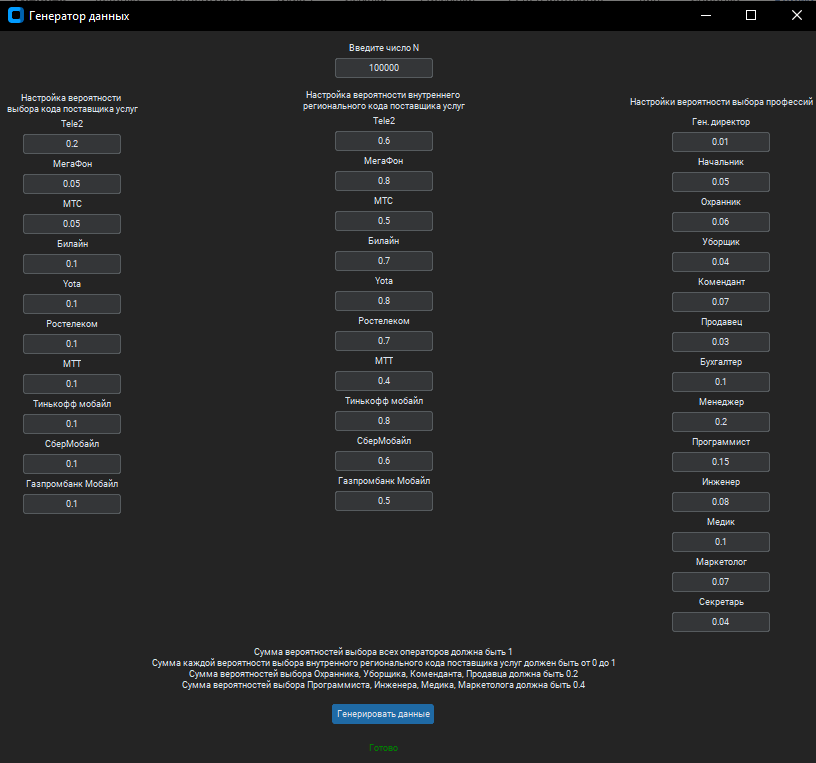
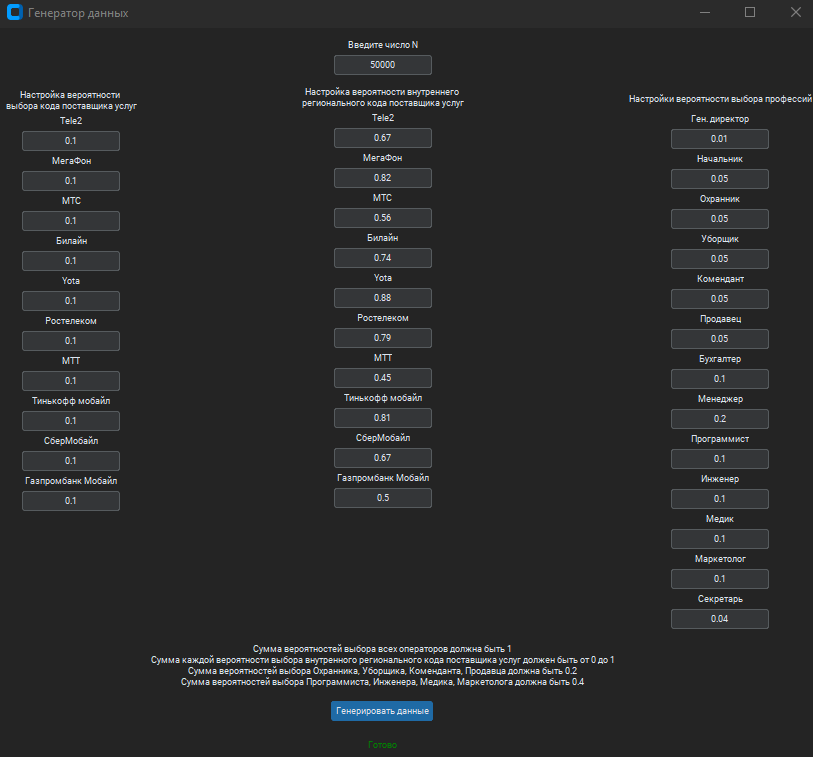
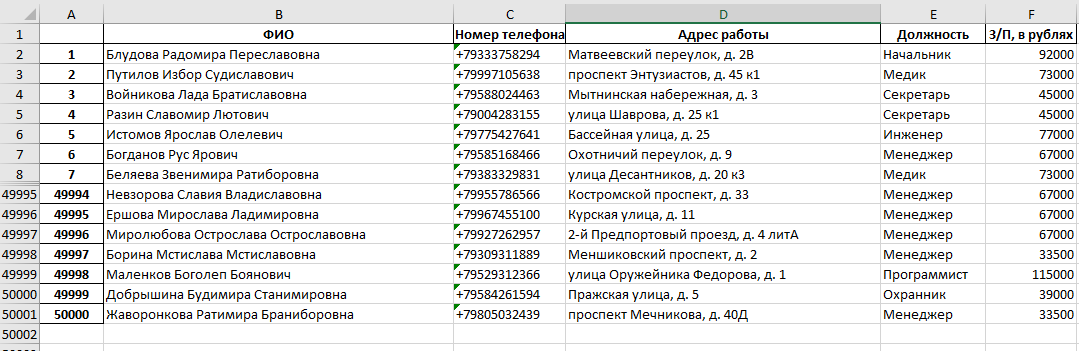
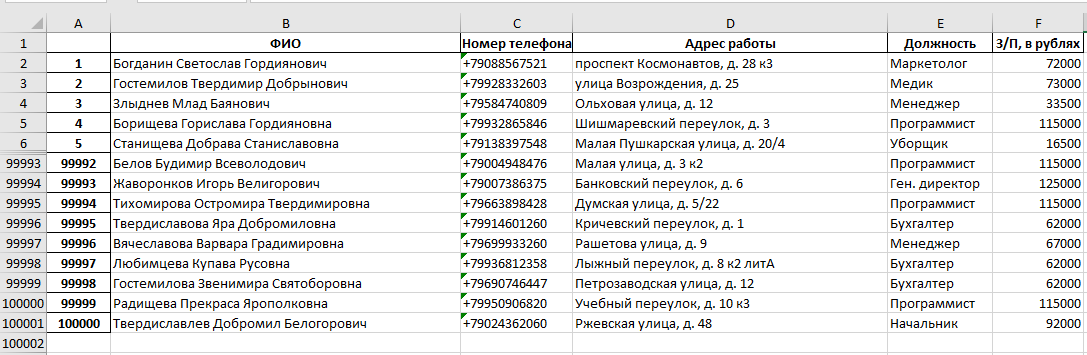
Минимальное необходимое место на диске: 3.5 МБ. Минимальное необходимое количество оперативной памяти: 400 МБ.

# **Контрольный пример**

 В данном разделе представлены два контрольных примера (описываются они параллельно), демонстрирующие способность программы генерировать данные, основываясь на ранее заданных параметрах.

*Рисунок 8.1–8.2. Интерфейс программы*

*Рисунок 8.3–8.4. Окно программы в процессе генерации*

*Рисунок 8.5–8.6. Окно программы после окончания генерации*

*Рисунок 8.7–8.8. Результат работы программы*

# **Заключение**

В рамках данной работы была успешно достигнута поставленная цель - разработка алгоритма и программы для генерации синтетических данных. Данная лабораторная работа позволила не только приобрести, но и закрепить знания в области генерации данных. Был создан набор синтетических данных, отвечающих определенным требованиям и характеристикам, что имеет важное значение для последующего использования в различных учебных целях.