**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный университет**

**Факультет прикладной математики-процессов управления**

**Кафедра “фундаментальная информатика и информационные технологии”**

**отчет**

**по лабораторной работе №1**

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

**на тему «генерация синтетических данных»**

**Вариант – 5**

|  |  |  |
| --- | --- | --- |
| Студент гр. 22Б15-пу |  | Колосков В.Ю. |
| Преподаватель |  | Дик А.Г. |

**Санкт-Петербург**

**2023 г**

**Оглавление**

[Цель работы 3](#_Toc148294014)

[Задача 3](#_Toc148294015)

[Теоретическая часть 3](#_Toc148294016)

[Описание алгоритма 4](#_Toc148294017)

[Блок-схема программы 6](#_Toc148294018)

[Описание программы 12](#_Toc148294019)

[Рекомендации пользователя 15](#_Toc148294020)

[Рекомендации для программиста 16](#_Toc148294021)

[Контрольный пример 16](#_Toc148294022)

[Вывод 17](#_Toc148294023)

[Список использованной литературы 17](#_Toc148294024)

# **Цель работы**

Создать генератор синтетических данных о покупке ЖД билетов

# **Задача**

Изучить принципы генерации синтетических данных. Разработать алгоритм, который генерирует датасет с информацией о билетах на поезд и написать программу, которая при, вводе определенных параметров генерирует на их основе данный датасет.

# **Теоретическая часть**

Существует несколько подходов к созданию датасета покупки ЖД билетов. Один из подходов - использование данных из системы продаж билетов железнодорожных компаний. Для получения датасета эту информацию необходимо собрать в базу данных

Другой подход - создание синтетических данных о покупке ЖД билетов на основе существующих. Этот подход может полезен в случае отсутствия доступа к данным системы продаж билетов. Синтетические данные создаются на основе моделирования поведения покупателей и генерации случайных данных, которые соответствуют этой модели.

При использовании второго типа необходимо учитывать многие факторы при генерации железнодорожных билетов, а именно: направление движения поезда, тип поезда, тип вагонов, время года, расстояние между городами и другие.

Направление движения поезда определяет его номер: поезда с четными номерами двигаются на север и восток. Поезда с нечетными номерами едут на юг и запад[1].

Тип поезда и расстояние между городами определяют среднюю скорость, а также время, которое он находиться в пути, благодаря этому можно составить более точное расписание.

Для определения расстояния между двумя точками на поверхности земли используется геодезическая формула Винсента.

Геодезическая формула Винсента — это математическое выражение, которое используется для определения расстояния между двумя точками на сферической поверхности. Формула была разработана в 19 веке французским геодезистом Фердинандом Винсентом и имеет вид [2]:

, где:

d – расстояние между точками в километрах

lt1 – широта первой точки в радианах

lt2 – широта второй точки в радианах

ln1 – долгота первой точки в радианах

ln2 – долгота второй точки в радианах

R – радиус земли в километрах

Тип вагона определяет более точную стоимость билета, а использование времени года необходимо для генерации поездов, которые ходят только в определенный сезон или всего несколько раз в год

# **Описание алгоритма**

* 1. Задаются параметры, по которым будет генерироваться датасет: количество строк в датасете, значения вероятностей генерации карт с определенными платежными системами, значения вероятностей генерации банков, выдавших карты.
  2. Генерация серий и номеров паспортов в количестве, равному количеству строк в датасете
  3. Генерация номеров банковских карт в количестве, равному количеству строк в датасете
  4. Генерация фамилии имени и отчества человека, который покупал билет в количестве, равному количеству строк в датасете
  5. Генерация билетов:
     1. Генерация расписания поездов
     2. Рандомизированный выбор одного из поездов
     3. Генерация для каждого места в поезде его цены
     4. Повторение шагов b-c до тех пор, пока мест с ценами не будет в количестве больше или равному количеству строк в датасете, при этом в каждом выбранном на шаге b поезде все места должны иметь цены.
  6. Объединение всех данных, полученных на шаге 2 - 5 и создание необходимого количества строк
  7. Запись строк в excel файл

# **Блок-схема программы**

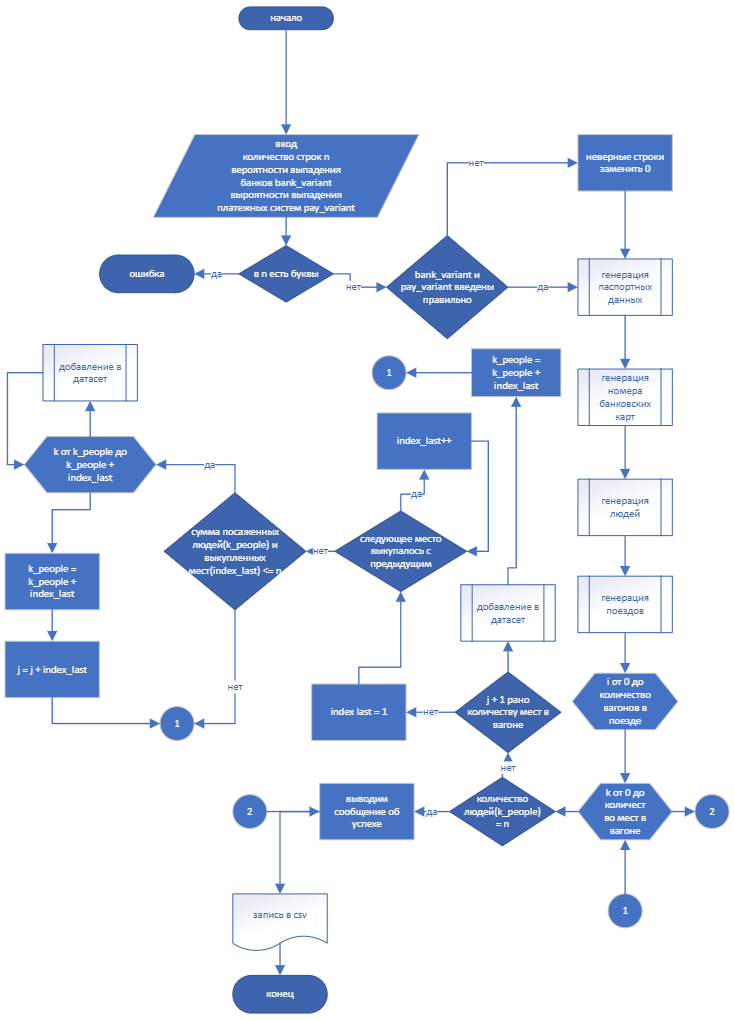
****

Рисунок 1. Основное тело программы

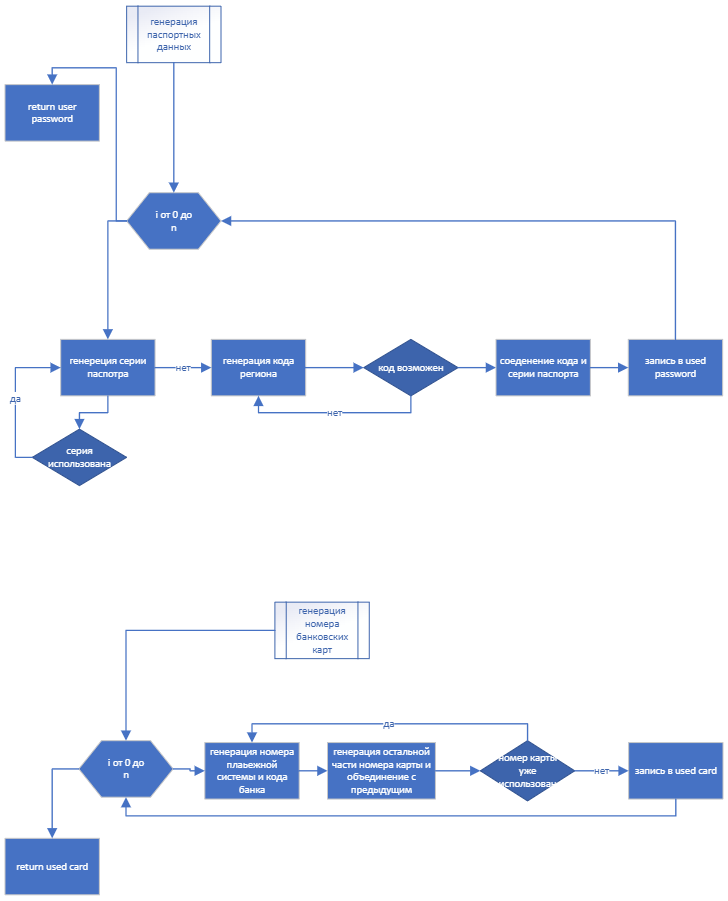
****

Рисунок 2. Функции генерации паспортных данных и номера банковских карт

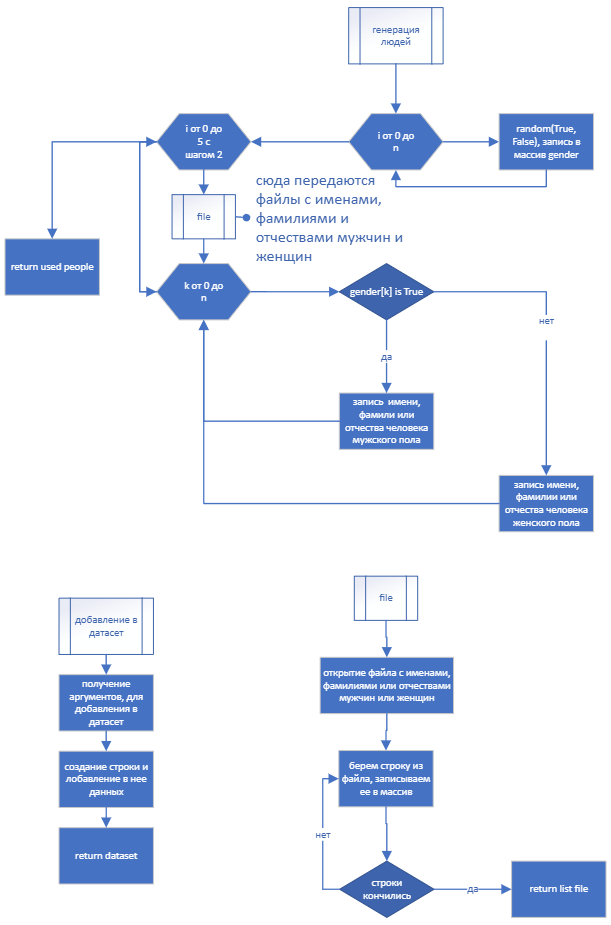


Рисунок 3. Функции генерации имен людей. Добавления в датасет и получения информации из файла

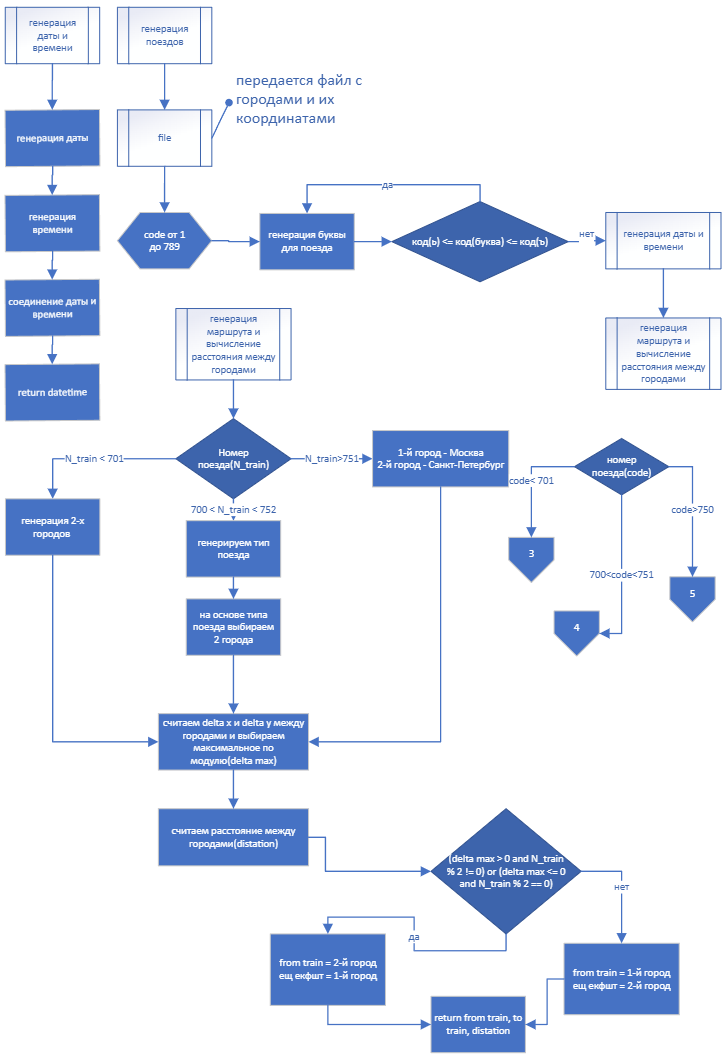


Рисунок 4. Функции генерации даты и времени, маршрута и поездов

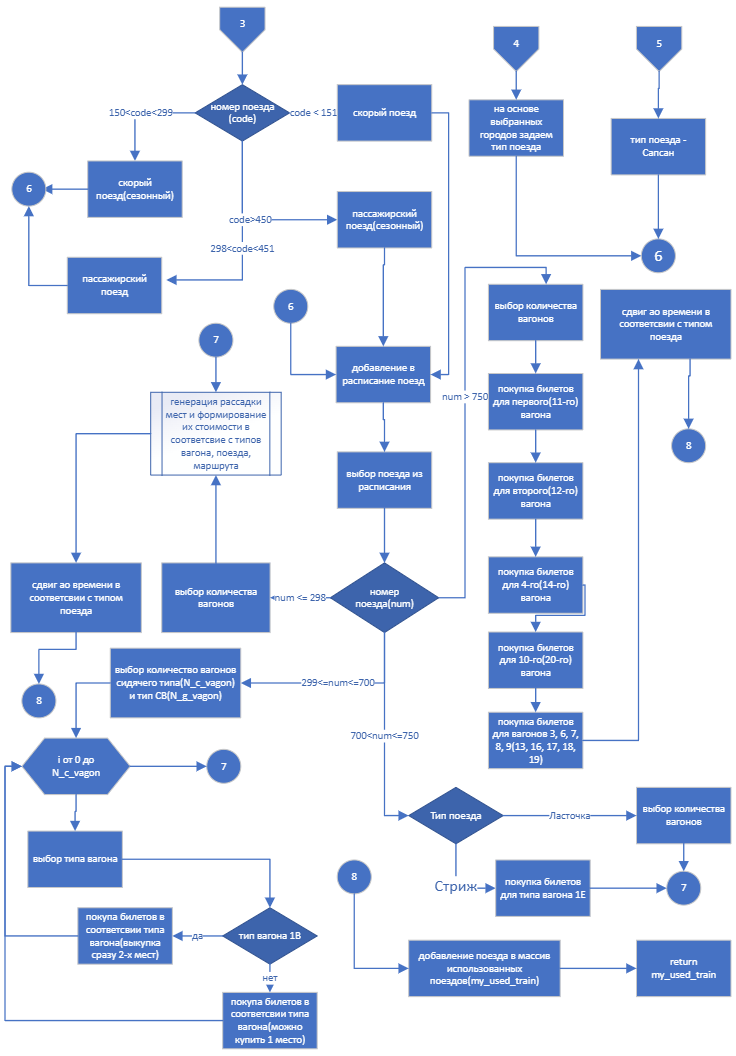
.

Рисунок 5. Генерация количества мест в поезде в зависимости от их типа

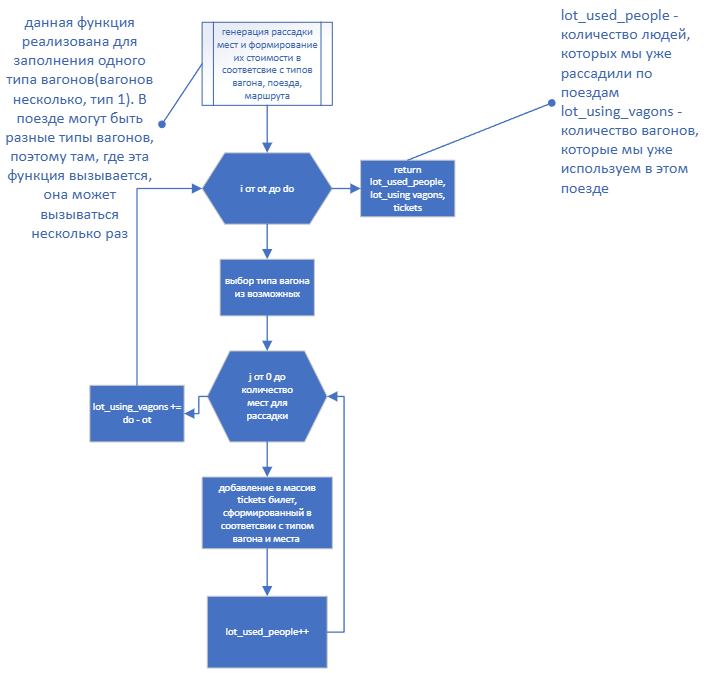


Рисунок 6. Генерация рассадки пассажиров

# **Описание программы**

Программа реализована на языке python 3.11.4, с использованием библиотек: os, random, geopy, datetime, pandas, tkinter. В программе использовались 10 функций и 20 структур данных. В таблице 1 представлено описание функций

Таблица 1. Описание функций

|  |  |  |
| --- | --- | --- |
| Имя функции | Тип возвращаемого значения | Описание функции |
| Creating\_generation | None | Проверяет правильность введенных данных, запускает функции генерации данных и добавления в датасет, сохраняет датасет в файле с расширением xlsx |
| Password\_generation | list | Генерирует серии и номеров паспортов людей |
| Card\_generation | list | Генерирует номера банковских карт, с которых человек может произвести платеж |
| People\_generation | list | Генерирует имена, фамилии и отчества людей, которые могут купить билеты |
| Train\_generation | list | Генерирует расписание поездов, а также билеты на эти поезда |
| Open\_file | list | Получает необходимые для генерации данные из файла и записывает их в список |
| Random\_datetime | datetime | Генерирует дату и время отправления поезда |
| Route\_generation | tuple | Генерирует пункт отправки и пункт назначения, рассчитывает расстояние между городами |
| Completion\_vagons | tuple | Сопоставляет место с его ценой, возвращает количество место, которым сопоставлена цена во всех поездах и количество вагонов, используемых в данном поезде |
| Add\_in\_dataset | dict | Добавляет строку в датасет, возвращает словарь строк датасета |

В таблице 2 представлено описание структур данных, используемых в программе.

Таблица 2. Описание структур

|  |  |  |  |
| --- | --- | --- | --- |
| Имя структуры | Тип структуры | Тип хранимых данных | Описание |
| Pay\_name | list | string | Хранит имена возможных платежных систем |
| Bank\_name | list | string | Хранит имена возможных банков |
| Pay\_insert | list | string | Хранит поля, в которые вводятся вероятности выпадения определенных платежных систем |
| Bank\_insert | list | string | Хранит поля, в которые вводятся вероятности выпадения определенных банков |
| Used\_series | set | string | Хранит использованные серии паспортов |
| Used\_password | set | string | Хранит использованные номера паспортов |
| Not\_code\_base | set | int | Хранит несуществующие номера регионов |
| Used\_card | set | string | Хранит использованные номера банковских карт |
| Pay\_variant | array | int | Хранит коды возможных платежных систем |
| List\_file | array | string | Хранит данные, полученные из файла |
| Name\_file | array | string | Хранит имена файлов, в которых находятся необходимые данные |
| Used\_people | array | string | Хранит полные имена людей |
| Gender\_people | array | bool | Хранит пол людей. True – мужчина. False - женщина |
| Branded\_trains | dict | array | Ключ – особые поезда, значение – массив из названий городов, в которые они могут ехать |
| Type\_vagon | dict | dict | Ключ – класс вагона. – значение словарь его подклассов, значение второго словаря – коэффициенты, на которые умножается стоимость билета в зависимости от вагона |
| scredule | array | tuple | Хранит информацию о поезде |
| My\_used\_train | array | tuple | Хранит информацию о заполненном поезде и билеты на места в его вагонах |
| dataset | dict | array | Хранит строки датасета, которые были сгенерированы |
| tickets | array | tuple | Хранит билеты на данный поезд |
| Using\_train | array | tuple | Список поездов, которые можно использовать |

# **Рекомендации пользователя**

Для запуска программы необходимо наличие устройства с установленной операционной системой Linux, macOS или Windows, а также среды разработки, поддерживающей запуск python версии 3.11.4

Необходимо установить python библиотеки: geopy, pandas, tkinter. Для этого необходимо открыть терминал и прописать следующие команды: pip install geopy, pip install pandas, pip install tkinker

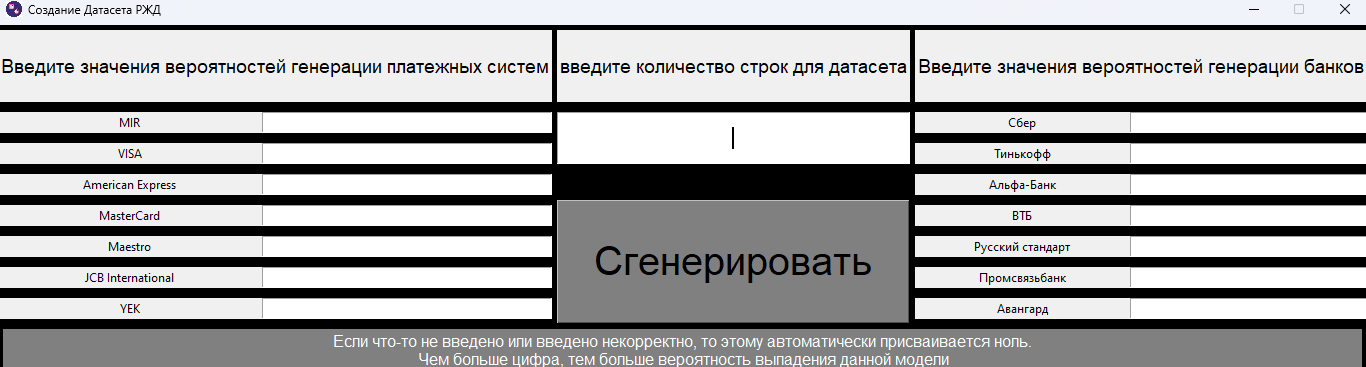
Под строкой «Введите значения вероятности генерации платежных систем» находятся названия платежных систем, а справа от них, соответствующая каждой системе, строка ввода. В эти строки необходимо ввести значения вероятности генерации для каждой платежной системы (рисунок. 7). Если оставить поле пустым или данные, введенные в него, будут некорректными, то вероятность автоматически заменится на 0.

Рисунок 7

Под строкой “Введите значения вероятностей генерации банков” находятся названия банков, а справа от них, соответствующая каждой системе, строка ввода. В эти строки необходимо ввести значения вероятности генерации для каждой платежной системы. (рисунок. 8) Если оставить поле пустым или данные, введенные в него, будут некорректными, то вероятность автоматически заменится на 0.

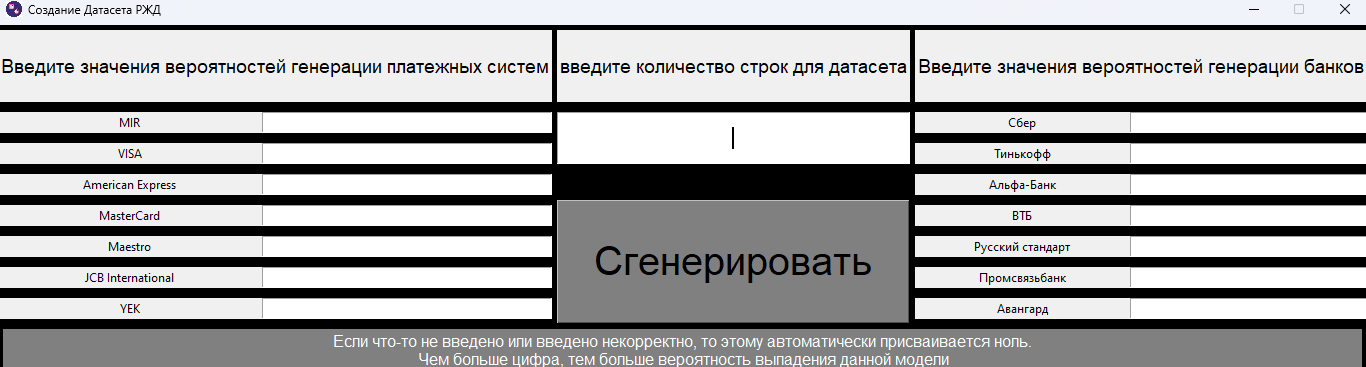


Рисунок 8

В поле расположенное под строкой “Количество строк в датасете” необходимо ввести количество строк, которые нужно сгенерировать для датасета (рисунок. 9). Если оставить поле пустым или данные, введенные в него, будут некорректными, то программа выдаст ошибку.

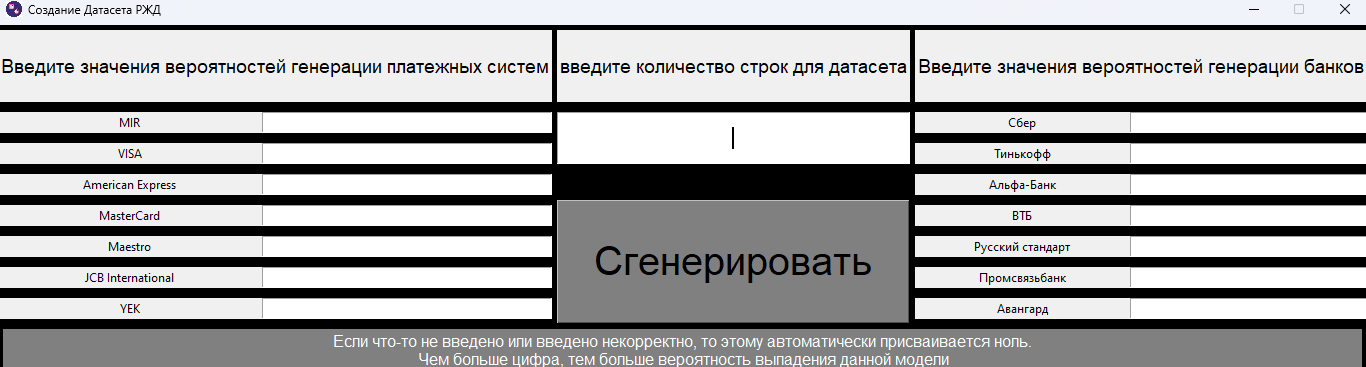
После того, как вы ввели данные, для генерации датасета необходимо нажать на кнопку “Сгенерировать” и дождаться уведомления об успехе. Excel файл будет сохранен в ту директорию, в которой находится программа.

Рисунок 9

# **Рекомендации для программиста**

Для внесения изменений в исходный код программы необходимо наличие устройство с установленное операционной системой Linux, macOS или Windows, а также среды разработки, поддерживающей запуск python версии 3.11.4

Исходный код программы доступен по ссылке ниже:

[study/dataset/main.py at main · ARnoD0rian/study (github.com)](https://github.com/ARnoD0rian/study/blob/main/dataset/main.py)

# **Контрольный пример**

В этом разделе представлен пример работы программы.

На рисунке 10 представлены параметры, установленные при генерации.

Рисунок 10. Меню программы

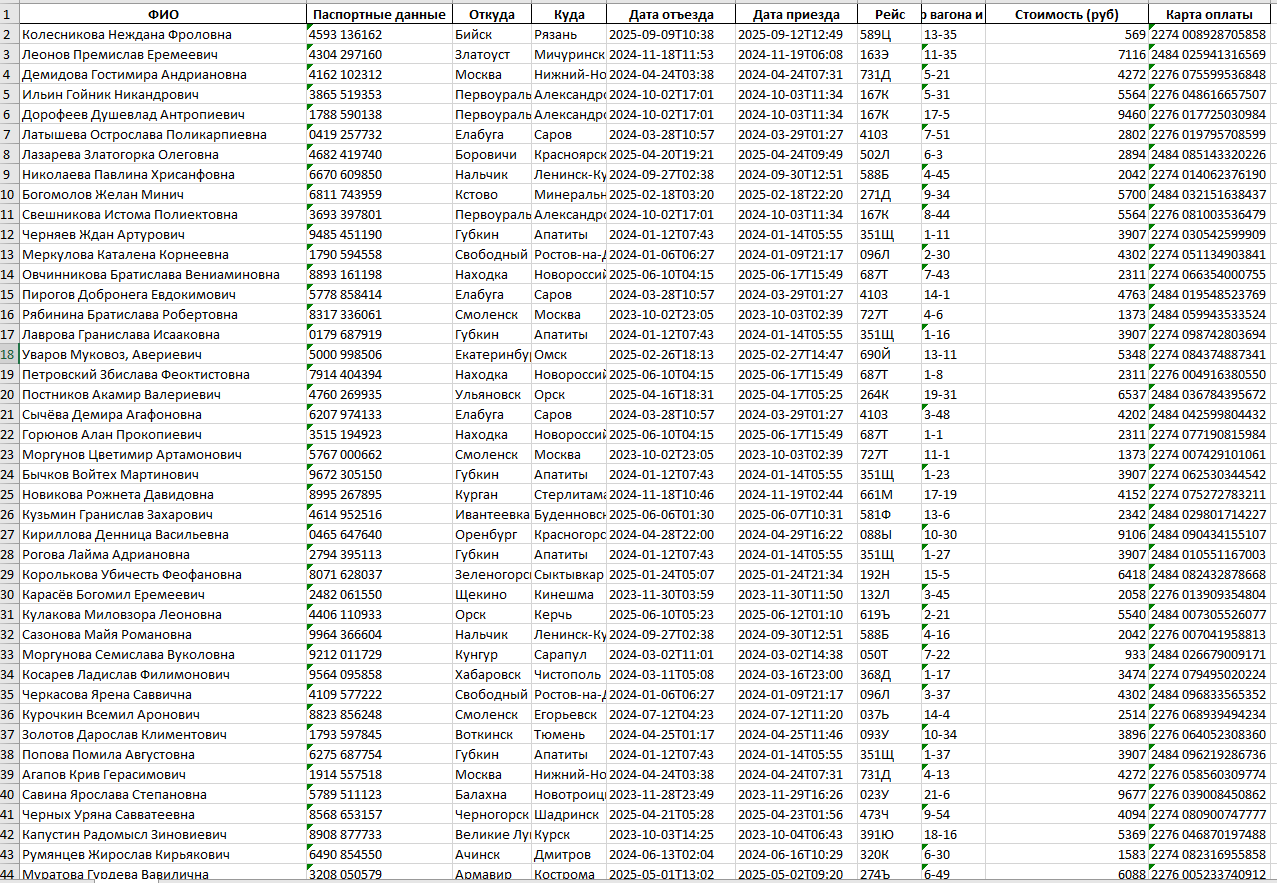
На рисунке 11 представлена часть данных, которые программа сгенерировала.

Рисунок 11. Результат работы программы

# **Вывод**

В ходе выполнения данной работы были изучены принципы генерации синтетических данных о покупке ЖД билетов, разработан алгоритм генерации синтетических данных данного типа и написана программа, которая при вводе определенных параметров генерирует датасет покупки ЖД билетов.

# Список использованной литературы

[1] Определение номера поезда: [(poezd.ru)](https://poezd.ru/text/poleznaya_informaciya/nomer_poezda.php)

[2] Определение расстояние на поверхности земли: [(kobzarev.com)](https://www.kobzarev.com/programming/calculation-of-distances-between-cities-on-their-coordinates/)