ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ   
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

###### Московский институт электроники и математики

им. А.Н. Тихонова

Игуменова Марта Александровна – автора МКР, группа БИВ173

**КУРСОВАЯ РАБОТА ПО ДИСЦИПЛИНЕ «ИНФОРМАТИКА И ПРОГРАММИРОВАНИЕ»**

Студент М.А. Игуменова

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

|  |  |
| --- | --- |
|  | Руководитель Топоркова А.С.  Доцент  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  А.С. Топоркова |
| Москва 2018г. | |

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

###### МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОНИКИ И МАТЕМАТИКИ

им. А.Н. Тихонова

**ЗАДАНИЕ**

**на курсовую работу по дисциплине**

**“Информатика и программирование”**

студенту группы БИВ173 Игуменовой Марте Александровой

1. Тема работы

Программа обработки экспериментальных данных, количество которых больше 10000.

1. Требования к работе

Требования у функциональным характеристикам: файл с экспериментальными данными будет открываться в программе, данные из таблицы будут анализироваться и распределяться по какой-либо зависимости. Для отображения используется простой графический интерфейс языка и библиотека **tkinter**. Для реализации меню возможностей будет использоваться меню с кнопками и ввод с клавиатуры.

Требования к составу технических средств: компьютер с монитором, мышь и клавиатура, процессор с тактовой чистотой 1 ГГц 1 Гб или выше, 1 Гб ОЗУ, операционная система- Windows 8/10.

Требования к программной части: среда разработки- Python 3 или версия выше, JetBrains PyCharm Community Edition 2017.2.4 x64, библиотеки Python - **tkinter**, Windows 8/10

1. Содержание работы

Титульный лист, аннотация, оглавление, описание модулей программы, введение, примеры работы программы, база данных с результатами эксперимента, выводы, список использованных источников, описание метода выборки и анализа.

4.Сроки выполнения этапов работы

Первый вариант МКР предоставляется студентом в срок до «15» марта 2018г.

Итоговый вариант МКР предоставляется студентом в срок до «30» мая 2018г.

|  |  |  |
| --- | --- | --- |
| Задание выдано | «10» января 2018г. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.С. Топоркова  подпись руководителя |
| Задание принято к исполнению | «10» января 2018г. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.С. Игуменова подпись студента |

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

###### МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОНИКИ И МАТЕМАТИКИ

им. А.Н. Тихонова

**График выполнения**

**курсовой работы по дисциплине**

**“Информатика и программирование”**

студента группы БИВ173 Игуменовой Марты Александровны

Тема работы

Программа обработки экспериментальных данных, количество которых больше 10000.

|  |  |  |
| --- | --- | --- |
| Дата согласования первого варианта МКР | «15» марта 2018 г. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.С. Топоркова подпись руководителя |
| Дата согласования итогового варианта МКР | «30» мая 2018 г. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.С. Топоркова подпись руководителя |

# 

# Аннотация

Объектом разработки является программа обработки экспериментальных данных в размере больше 10 000 единиц. Целью программы являлась визуализация анализа и обработки данных мэрии Москвы. Благодаря моей программе пользователь получает информацию о подаче заявлений, ответах на заявления, выполнения поставленных вопросов в удобной графической форме, улучшающей восприятие информации. Полученные результаты следует использовать для улучшения взаимодействия между государственными структурами и гражданами России. При дальнейшей разработке программы следует увеличить CSV базу и улучшить пользовательский интерфейс. Объем работы: 25 страниц.

The object of development is the program for processing experimental data in the amount of more than 10,000 units. The purpose of the program was to visualize the analysis and data processing of the Moscow mayor's office. Thanks to my program, the user receives information about submitting applications, responding to applications, performing the questions posed in a convenient graphical form that improves the perception of information. The results obtained should be used to improve the interaction between state structures and Russian citizens. When further developing the program, you should increase the CSV database and improve the user interface. Scope of work: 25 pages.

Оглавление

[Аннотация 4](#_Toc513797881)

[Оглавление 5](#_Toc513797882)

[Введение 6](#_Toc513797883)

[Постановка задачи 8](#_Toc513797884)

[Анализ существующего технического решения задачи 10](#_Toc513797885)

[Обоснование выбора метода решения поставленной задачи 11](#_Toc513797886)

[Обоснование выбора языка программирования и программных средств 12](#_Toc513797887)

[Результаты обработки данных мэрии Москвы 13](#_Toc513797888)

[Внешняя спецификация программы 15](#_Toc513797889)

[Пример программы 17](#_Toc513797890)

[Заключение 26](#_Toc513797891)

[Список использованных источников 27](#_Toc513797892)

# 

# Введение

В результате исследования методов решения задачи, я выявила два способа визуализации программного продукта: реализация базы данных средствами языка sql или использование базы данных CSV и обработка ее данных с помощью библиотек python 3.6, таких как matplotlib, numpy, seaborn, filedialog, pandas. По итогам изучения литературы, я выбрала второй вариант реализации программы.

При сборе экспериментальных данных появляется необходимость обрабатывать большой объем информации, для получения более точных результатов, графического представления данных и основных тенденций в экспериментальных значениях. Следовательно, для оптимизации работы с базами данных, нужна программа, наглядно показывающая пользователю результаты обработки базы, для дальнейшей статистической обработки.

Обработка данных с сайта мэрии Москвы, поможет более точно выражать зависимости между собранными данными и представлять их в более удобном для пользователя интерфейсе.

Целью работы является обработка большого объема данных мэрии Москвы и дальнейшее их представление в графической форме. Задачей работы является изучение библиотек обработки и представления данных языка python 3.6, взаимосвязей между наборами данных, обработка большого объема данных, реализация графического интерфейса средствами библиотеки python 3.6 tkinter.

Для написания скрипта я использовала язык программирования python 3.6, в том числе встроенные библиотеки tkinter, filedialog, seaborn, numpy, matplotlib, requests, bs4, os интерактивная среда разработки JetBrains PyCharm Community Edition 2017.2.4 x64, Jupyter Notebook, операционная система Windows 10, стандартные программы Windows 8-10.

В результате выполнения работы, была получена программа с представлением экспериментальных данных, взятых с сайта мэрии Москвы и графической оболочкой, написанной с помощью библиотеки tkinter.

# 

# Постановка задачи

Программа будет выполнена в двух частях: обработка экспериментальных данных и программа взаимодействия пользователя с результатами обработки.

Код обработки данных будет выполнен средствами Jyputer Notebook. Для анализа использую различные методы, представленные в возможностях языка python и средствах Notebook:

Корреляция – статистическое взаимодействие двух или более величин. Этот метод нужен для выяснения зависимости между несколькими наборами данных. Результатом выполнения метода будет коэффициент корреляции.

Для каждого набора данных выведем основные статистические характеристики:

1. Число непропущенных значений -наполненность ячеек датафрейма.
2. Среднее и стандартное отклонение значений в каждой ячейке набора данных.
3. Диапазон значений - вывод минимального и максимального значения конкретного набора.
4. Медиану данных - среднее значение в каждой ячейке.
5. Квартиль 25% - значение ниже среднего в наборе.
6. Квартиль 75% - значение выше среднего в наборе.

Взаимодействие пользователя с данными будет выполнено средствами языка python 3.6, а именно возможностями графической библиотеки tkinter. В программе будут представлены результаты обработки экспериментальных данных в виде картинок, код обработки данных в виде файла в формате ipynb с возможностью открытия через браузер. Так же будет реализована возможность просмотра базы данных в формате xls. Открыть датафрейм можно будет из программы, используя средства Windows.

Так же программа будет показывать другие результаты обработки данных, связанные с взаимодействием нескольких ячеек набора данных, не требующих использования метода для анализа.

Так же в программе реализована возможность делать заметки. В интерфейс встроен простой блокнот, с возможностью записи текстовых файлов и сохранения их на компьютере пользователя. Это средство нужно для улучшения взаимодействия пользователя с данными.

Графический интерфейс программы будет содержать кнопки для взаимодействия с результатами данных, вызов средств блокнота, кнопка просмотра базы данных средствами Windows, кнопка просмотра кода обработки данных средствами браузера.

# Анализ существующего технического решения задачи

Обработка экспериментальных данных предполагает реализацию математических и графических решений зависимостей между отдельными частями массива данных. Следовательно, для понимания метода решения задачи, нужно знать основную информацию о том, что такое данные, какие они бывают, как с ними работать, какие методы обработки существуют.

Любые данные имеют свои признаки, которые делятся на несколько категорий: количественные, категориальные, бинарные. К количественным признакам относят данные типа int64, к категориальным типа string, распределяющим основные массивы данных по категориям, количество которых больше двух, бинарным типа sting и float64, если количество категориальных признаков равно двум.

Признаки являются основным объектом изучения и обработки. Обычно методы изучения признаков делятся на две группы: количественные признаки изучаются с помощью математических методов анализа, построения графиков и гистограмм, категориальные и бинарные признаки объединяются в одну группу и являются вспомогательными признаками. Изучая взаимосвязи признаков, используются методы: количественные признаки с количественными, категориальные с категориальными, количественные с категориальными.

Для нахождения зависимостей между данными используются математические и аналитические методы: построение графиков корреляции объектов data frame, диаграмм, bar chart, pie chart, нахождение медиан, непропущенных значений, квартилей, отклонений и диапазона значений.

Следовательно, для написания программы, я буду использовать математический подход к обработке экспериментальных данных.

# Обоснование выбора метода решения поставленной задачи

На данный момент, метод интеллектуального анализа данных является наиболее распространённым способом парсинга и преобразования большого объема данных. Одно из важнейших качеств метода – наглядное представление информации. Следовательно, человек с незначительным уровнем знания математики, не владеющий методами анализа, сможет понять представленные данные и оценить зависимости между их наборами. Так как одной из целей моей работы является графическая визуализация данных, то метод интеллектуального анализа подходит для решения моей задачи.

# Обоснование выбора языка программирования и программных средств

Для написания программы я выбрала язык программирования Python 3.6. Python простой язык программирования, производительный, с большим набором методов и программных средств. Язык имеет достаточно библиотек для реализации графического интерфейса, доступной обработки данных, визуализации, доступа к базам данных, связки программы с программным обеспечением компьютера. Библиотека языка Numpy подходит для научных вычислений, Pandas позволит упростить обработку большого количества данных, Matplotlib используется для построения графиков, Seaborn сделает графики доступными и понятными для пользователя программы.

Я выбрала среду разработки JetBrains PyCharm Community Edition 2017.2.4 x64, потому что это самая удобная и интеллектуальная IDE для работы с языком Python. Среда имеет простую навигацию в коде, подсказки, коллекцию плагинов, встроенные библиотеки.

Для обработки данных я использовала Jupyter Notebook – веб-приложение с открытым исходным кодом, позволяет визуализировать данные и писать комментарии прямо в файле тетради. Notеbook поддерживает основные языки программирования.

# Результаты обработки данных мэрии Москвы

В качестве базы были взяты открытые данные мэрии Москвы за 2016 и часть 2017 года.

Загруженная таблица данных имеет большое количество столбцов, в том числе девять основных:

1. year – год отчета по письмам
2. month – месяц отчета
3. total\_appeals – общее количество обращений за месяц
4. appeals\_to\_mayor – обращения за месяц в адрес Мэра
5. res\_positive – положительные решения за месяц
6. res\_negative – отклоненные письма
7. res\_explained – количество писем ,на которые был дан ответ
8. El\_from\_to\_mayor – электронные письма Мэру
9. Pap\_from\_to\_mayor – обращения в письменной форме

Таким образом, всего в таблице столбцов 32, строк 20. Моя программа работает с файлом в формате CSV, где запятая/точка/двоеточие/точка с запятой – знак табуляции.

После поиска коэффициента корреляции первых семи столбцов, станет понятно, что общее количество разъясненных обращений, превышает количество других ячеек базы данных, следовательно, делаем вывод о работе мэрии – государственный орган старается по возможности ответить на все письма. Поэтому, чем больше обращений в мэрию, тем больше разъяснений пишется. Отсюда ясно значение коэффициента корреляции, который равен 0,83.

Рассмотрим другие 22 столбца таблицы, которые содержат данные об обращения по округам Москвы. Построим столбчатую диаграмму обращений по округам, которая находится в программе. Заметим, что обращений в центральном округе больше всего, можно сделать вывод, что в этом округе чаще всего проводятся какие-либо работы по благоустройству.

Другие функции программы позволяют увидеть аналогичные зависимости между наборами данных. Пользователь может увидеть визуализацию коэффициента корреляции, столбчатые диаграммы по наборам данных, основные статистические признаки наборов. Пользователь сможет сделать свои выводы по представленным графическим моделям.

# Внешняя спецификация программы

Пользовательский интерфейс программы выполнен средствами языка Python 3.6 и библиотеки tkinter.

Пользователь взаимодействует с программой мышкой. Графики будут отображаться после нажатия соответствующих кнопок в окне программы рис. 1. Так же пользователь может посмотреть используемую базу данных в формате .xls, открытый код тетради Jupyter Notebook. Для удобства анализа в программу был добавлен простой текстовый редактор рис. 2, для написания заметок. Его можно вызвать, нажав кнопку «Блокнот для заметок». Текстовый редактор может сохранить или открыть файл .txt, позволяя его редактировать.

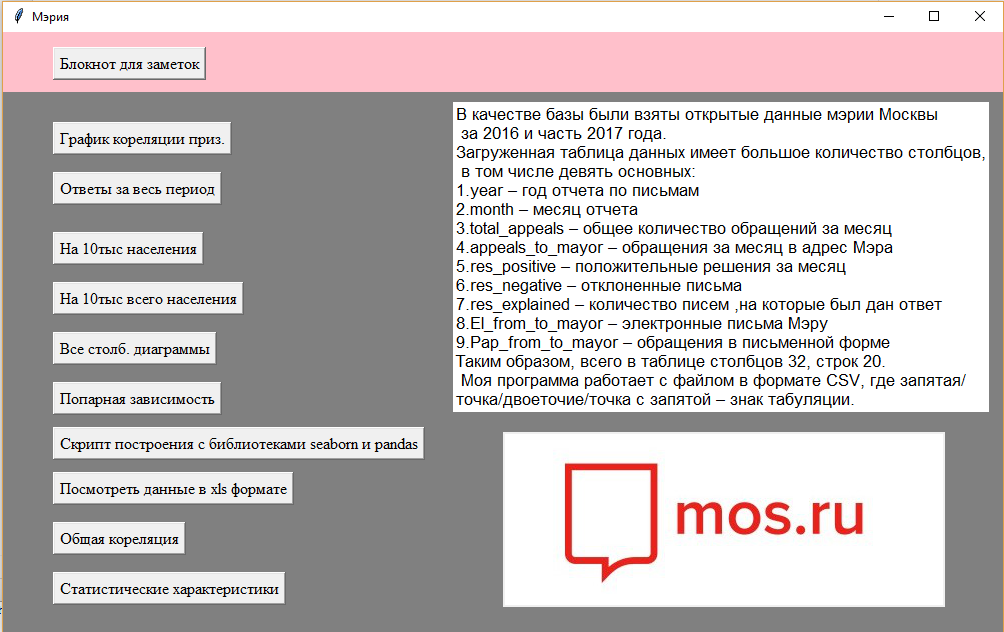


Рис.1 –Пользовательский интерфейс приложения



Рис. 2 – Пользовательский интерфейс блокнота

# 

# Пример программы

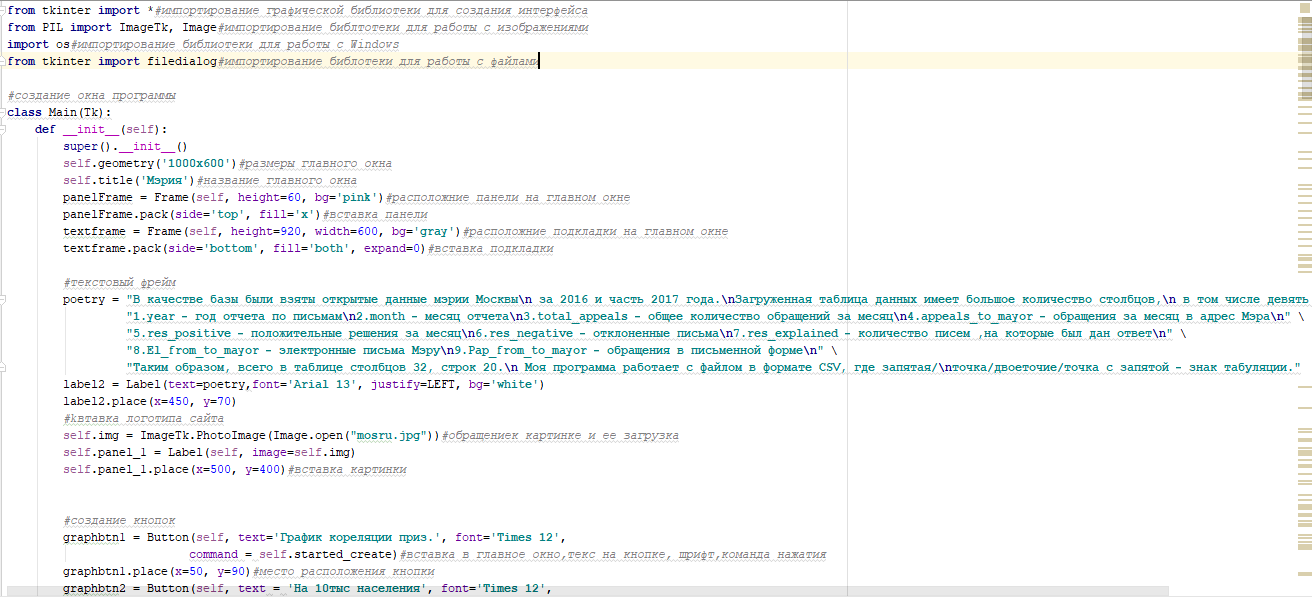


Рис. 3 - Код внешней программы 1



Рис. 4 - Код внешней программы 2

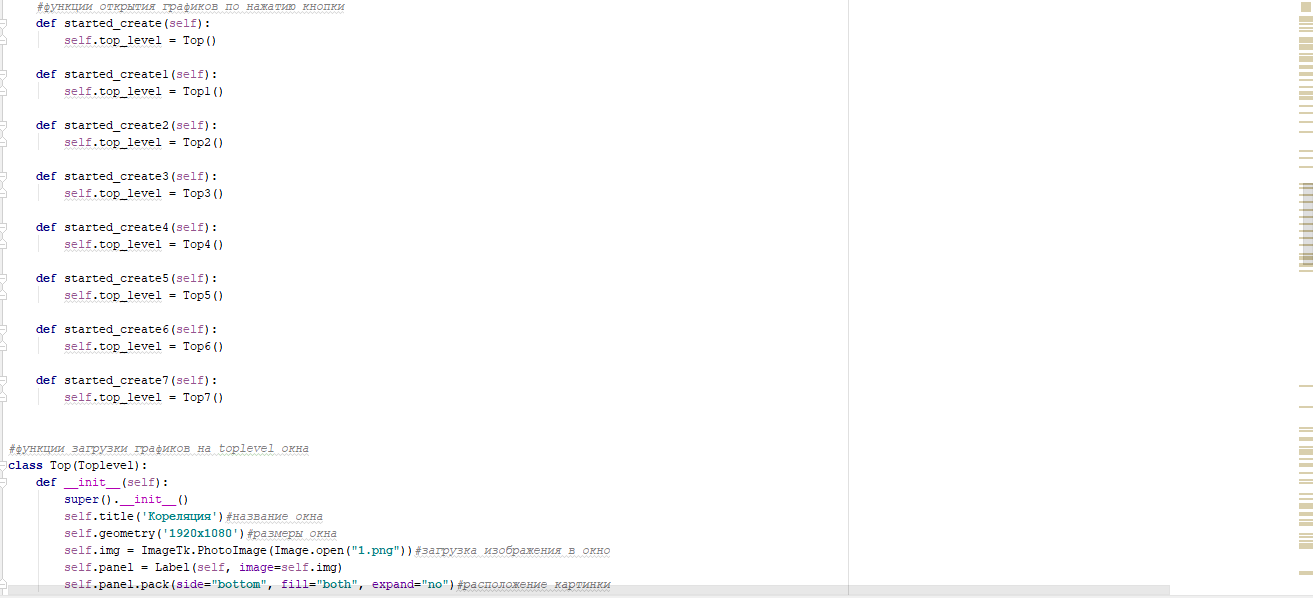


Рис. 5 - Код внешней программы 3

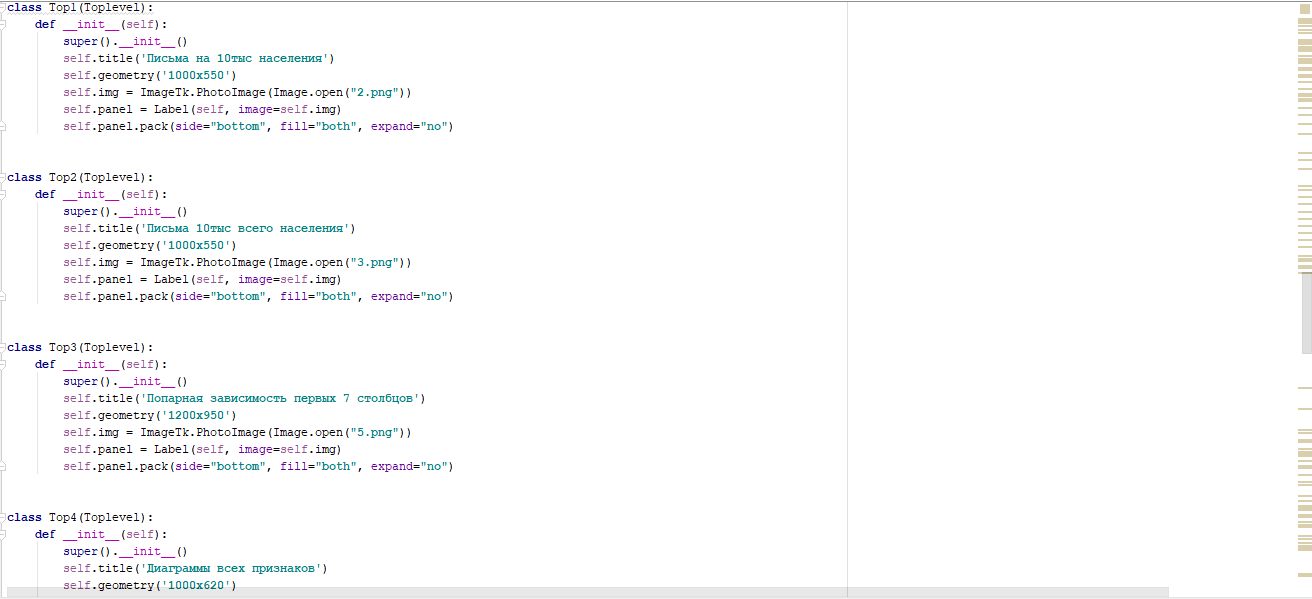


Рис. 6 - Код внешней программы 4

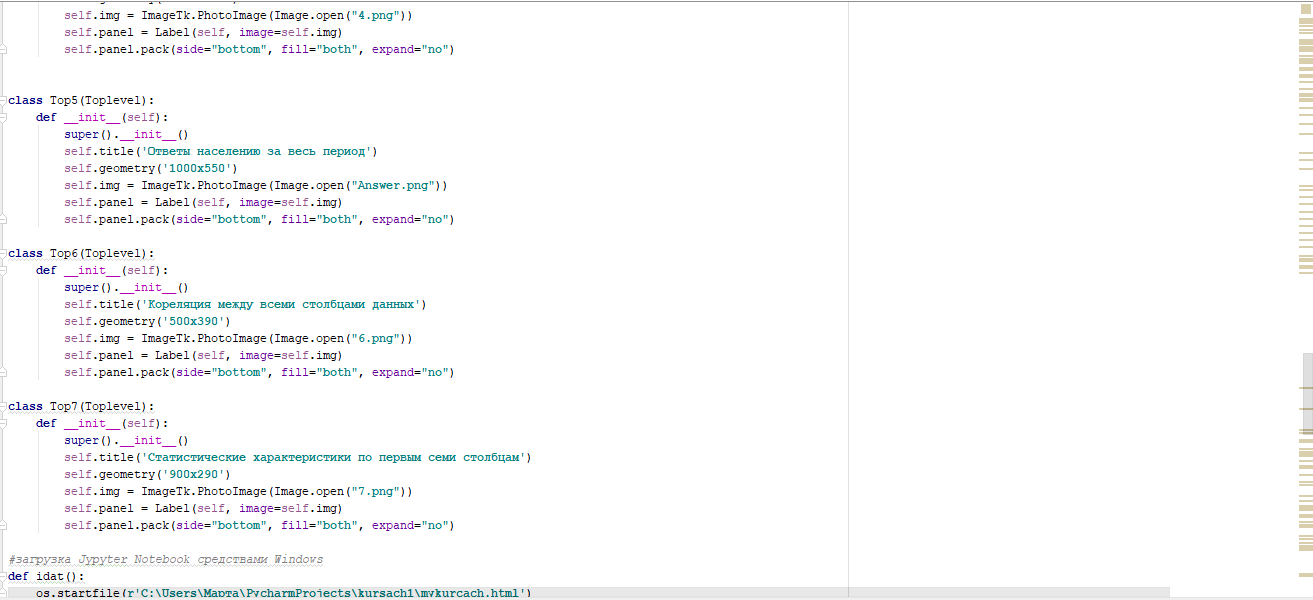


Рис. 7 - Код внешней программы 5

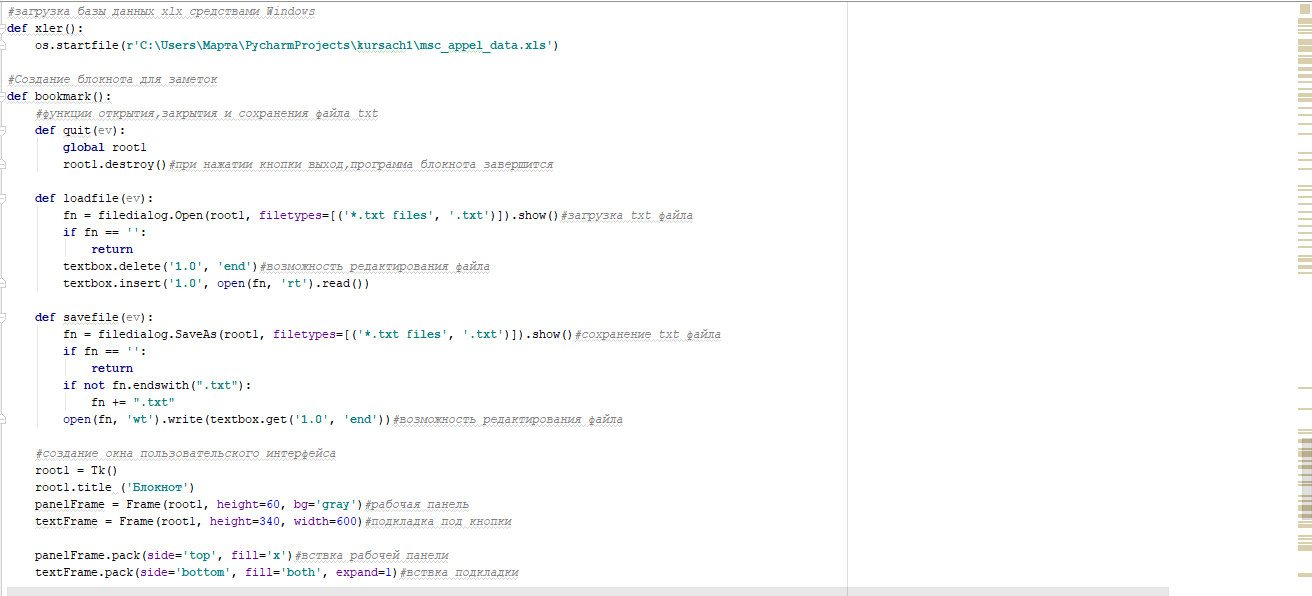


Рис. 8 - Код внешней программы 6

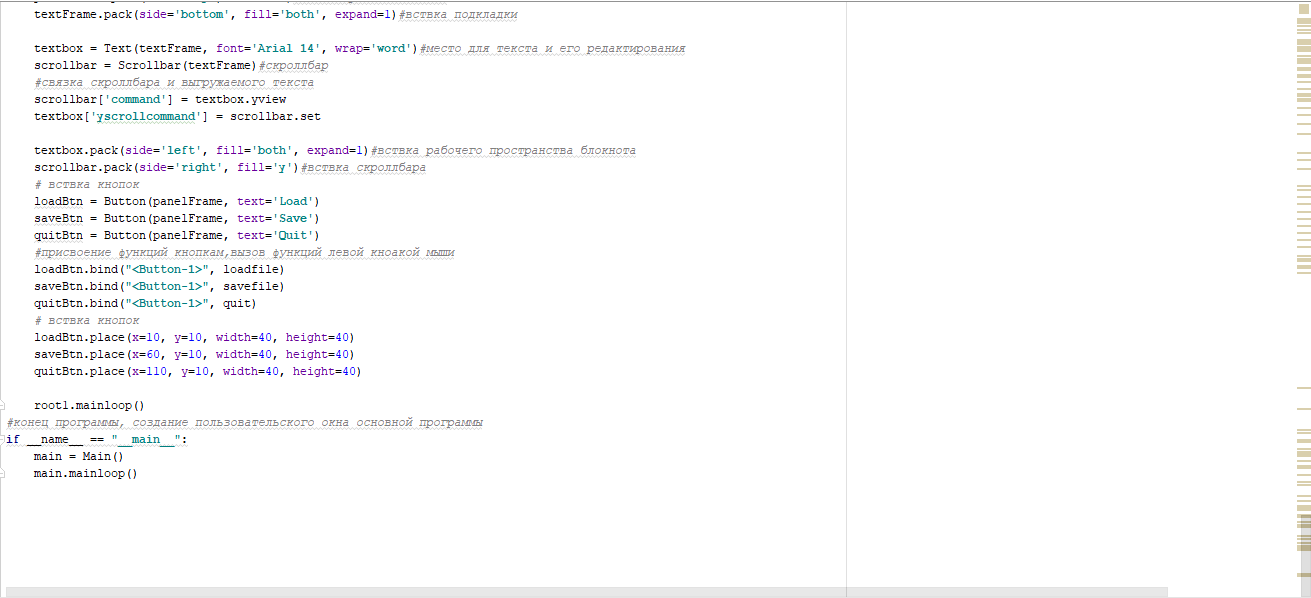


Рис. 9 - Код внешней программы 7

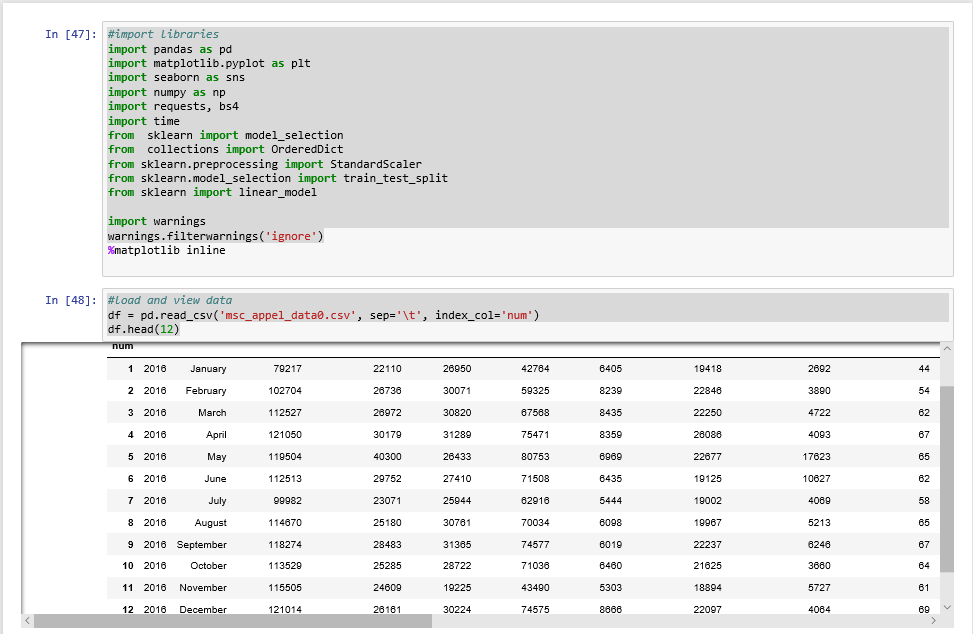


Рис. 10 - Код внутренней программы 1

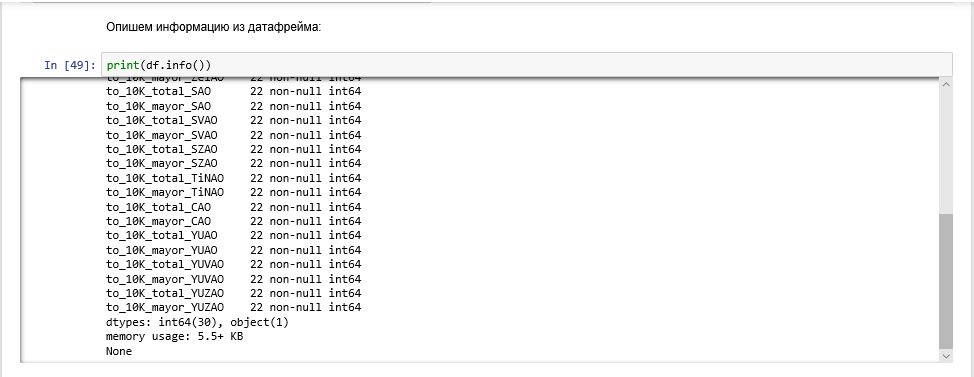


Рис. 11 - Код внутренней программы 2

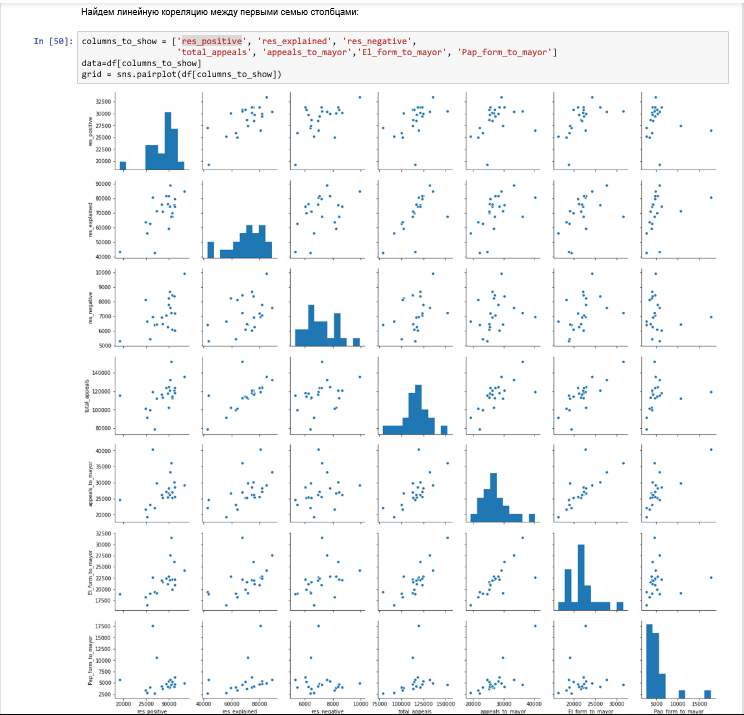


Рис. 12 - Код внутренней программы 3

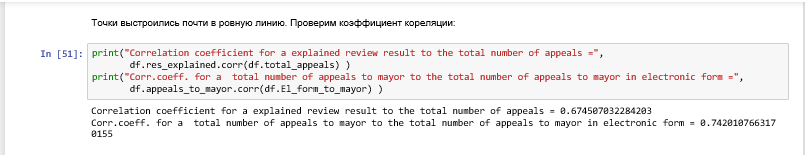


Рис. 13 - Код внутренней программы 4

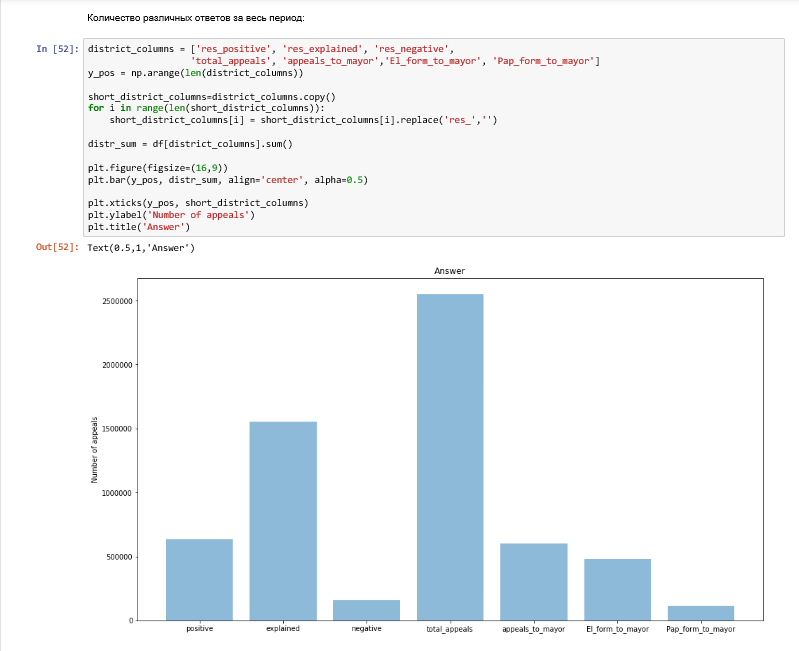


Рис. 14 - Код внутренней программы 5

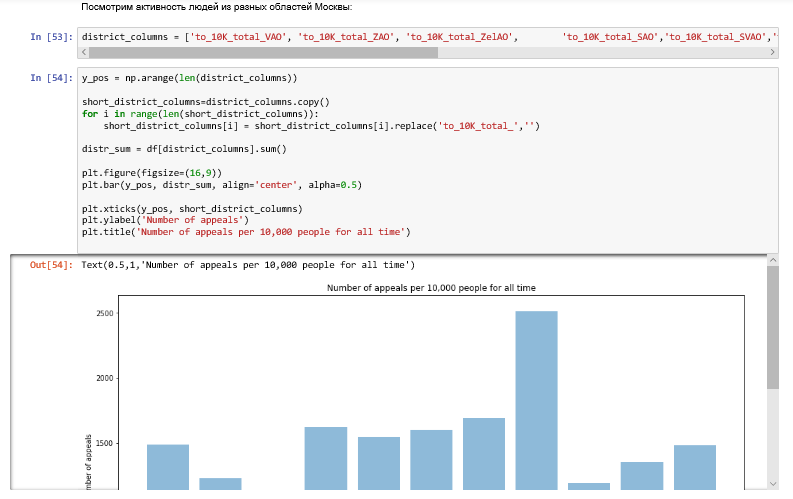


Рис. 15 - Код внутренней программы 6



Рис. 16 - Код внутренней программы 7

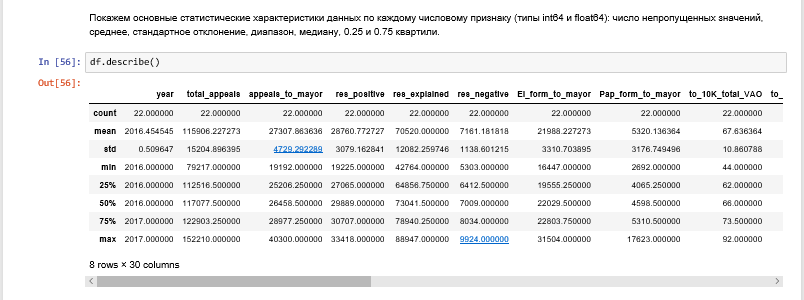


Рис. 17 - Код внутренней программы 8

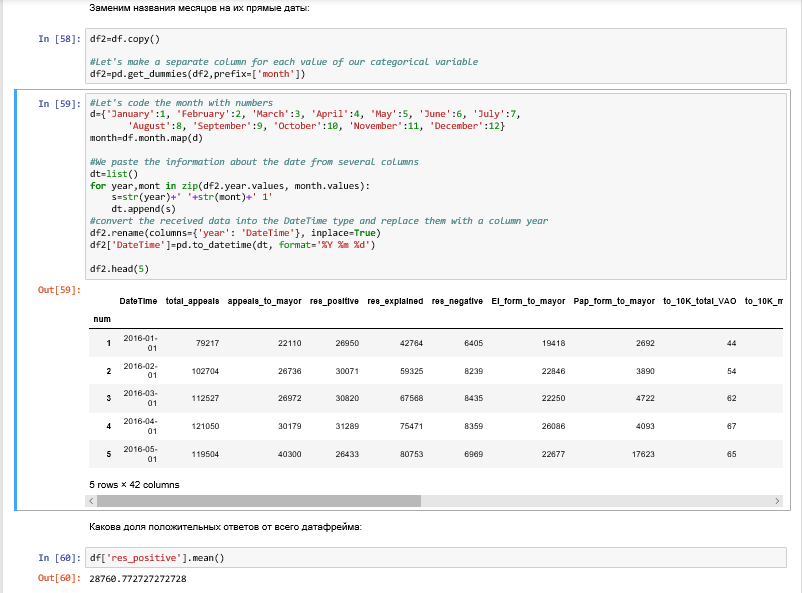


Рис. 18- Код внутренней программы 9

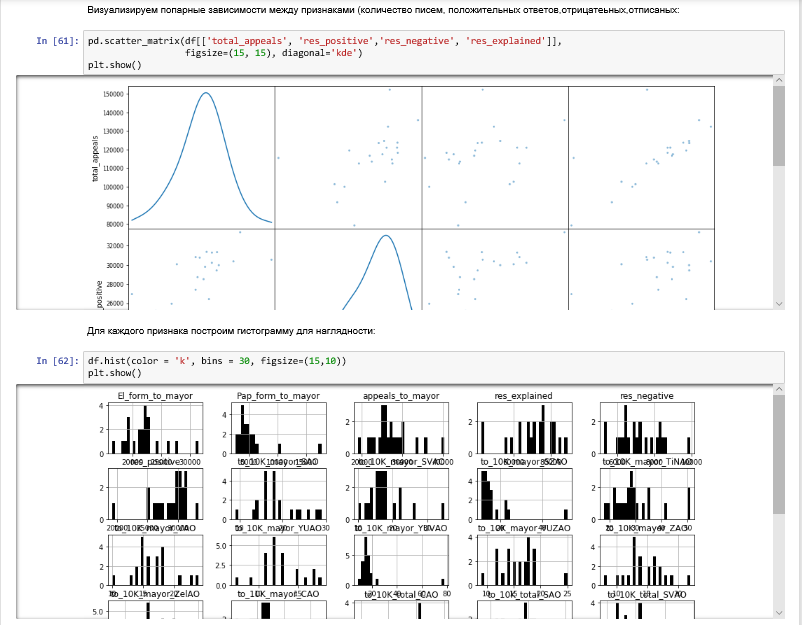


Рис. 19- Код внутренней программы 10

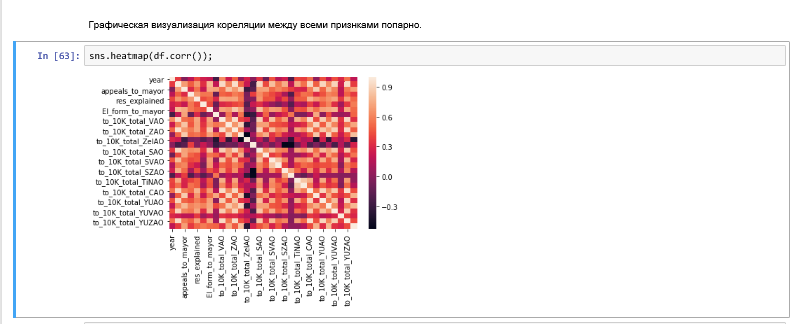


Рис. 20- Код внутренней программы 11

# Заключение

При выполнении работы были получены следующие результаты:

1. Написана программа взаимодействия с данными мэрии Москвы.
2. Обработаны данные базы мэрии Москвы
3. Изучены библиотеки обработки данных языка Python 3.6, такие как seaborn, matplotlib, numpy, pandas.
4. Изучена библиотека графического интерфейса программы tkinter.
5. Написан графический интерфейс программы взаимодействия с данными мэрии Москвы.
6. Изучены основные способы обработки статистических и экспериментальных данных.

# Список использованных источников

1. Викиверситет. "Курс по библиотеке Tkinter языка Python". <URL:https://ru.wikiversity.org/wiki/%D0%9A%D1%83%D1%80%D1%81_%D0%BF%D0%BE_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B5_Tkinter_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0_Python> (дата обращения 10.02.2018).
2. Википедия. "Data Mining". <URL:https://ru.wikiversity.org/wiki/%D0%9A%D1%83%D1%80%D1%81_%D0%BF%D0%BE_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B5_Tkinter_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0_Python> (дата обращения 12.01.2018).
3. Юрий Кашницкий. Хабрхабр. "Материалы открытого курса OpenDataScience и Mail.ru Group по машинному обучению". [URL:https://habrahabr.ru/company/ods/blog/344044/](URL:https://habrahabr.ru/company/ods/blog/344044/%20)  (дата обращения 12.01.2018).
4. "Анализ данных при помощи Python. Графики в pandas и matplotlib" [URL: http://malev.ru/%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7-%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85-%D0%BF%D1%80%D0%B8-%D0%BF%D0%BE%D0%BC%D0%BE%D1%89%D0%B8-python-%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B8-%D0%B2-pandas/](URL:https://ru.wikiversity.org/wiki/%D0%9A%D1%83%D1%80%D1%81_%D0%BF%D0%BE_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B5_Tkinter_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0_Python) (дата обращения 12.02.2018).
5. Марк Лутц. Программирование на Python. Т. 1. М.: Символ-Плюс, 2011
6. Доля П.Г. "Введение в научный Python. Часть 2."  [http://docplayer.ru/33176756-Vvedenie-v-nauchnyy-python-oglavlenie.html](URL:https://habrahabr.ru/company/ods/blog/344044/%20)  (дата обращения 02.02.2018).