МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

|  |  |
| --- | --- |
| **Домашнее задание** | |
|  | |
| по дисциплине | Практикум по программированию |

|  |  |  |  |
| --- | --- | --- | --- |
| Студента | Шелепова Дениса Дмитриевича | | |
|  | фамилия, имя, отчество полностью | | |
| Курс | 2 | Группа | ФИТ-231 |
| Направление | 02.03.02. Фундаментальная информатика и информационные технологии ㅤ | | |
|  | код, наименование | | |
| Руководитель | старший преподаватель | | |
|  | должность, ученая степень, звание | | |
| Саматов А. П. | | |
|  | фамилия, инициалы, дата, подпись | | |
| Выполнил | 01.12.2024 ㅤ | | |
|  | дата, подпись студента(ки) | | |

|  |  |
| --- | --- |
| Итоговый рейтинг |  |

Омск 2024

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc10665396)

[1 Поиск, загрузка и подготовка данных 4](#_Toc10665397)

[2 Разведывательный анализ данных 8](#_Toc10665400)

[3 Предварительная обработка данных 13](#_Toc10665403)

[ЗАКЛЮЧЕНИЕ 16](#_Toc10665411)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 17](#_Toc10665412)

ВВЕДЕНИЕ

Анализ данных играет ключевую роль в современном обществе, активно влияя на принятие решений в бизнесе, науке и многих других сферах. Прежде всего он позволяет извлечь ценную информацию из массивов данных, которые в противном случае оставались бы без внимания. Эти данные, будь то числовые показатели продаж, медицинские исследования или результаты социологических опросов, предоставляют необходимую основу для глубинного понимания процессов и тенденций.

Используя статистические методы и современные технологии, организации могут выявлять закономерности, прогнозировать будущее и находить новые возможности для роста. Например, компании могут оптимизировать свою деятельность, анализируя предпочтения клиентов и разрабатывая персонализированные предложения, что существенно увеличивает шансы на успех.

Еще анализ данных помогает в повышении эффективности и снижении рисков. В условиях постоянно меняющейся рыночной среды способность быстро адаптироваться к новым условиям становится конкурентным преимуществом. Таким образом, данные становятся не просто цифрами, а мощным инструментом, открывающим перспективы для инноваций и стратегического развития.

Анализировать данные помогают следующие библиотеки, написанные для работы на языке программирования *Python*: *Pandas*, *Numpy*, *Scipy*, *Seaborn* и *Matplotlib*. Первая из них, *Pandas*, позволяет хранить и обрабатывать датафреймы. *Numpy* и *Scipy* предоставляют широкий спектр математических процедур над различными объектами, например поиск моды, медианы, среднего значения. *Seaborn* и *Matplotlib* содержат набор инструментов, позволяющих строить разнообразные графики и диаграммы.

1 Поиск, загрузка и подготовка данных

Для выполнения задания на платформе *Kaggle* был выбран датасет «*World Cities Dataset*», который отражает информацию о всех городах мира. В нем содержатся такие поля, как название города, его географические координаты (широта и долгота), страна, в которой он находится, её сокращенное название, население города и его рейтинг. Чтобы описать датасет, был создан файл *describe.md*, представленный на рисунке 1.

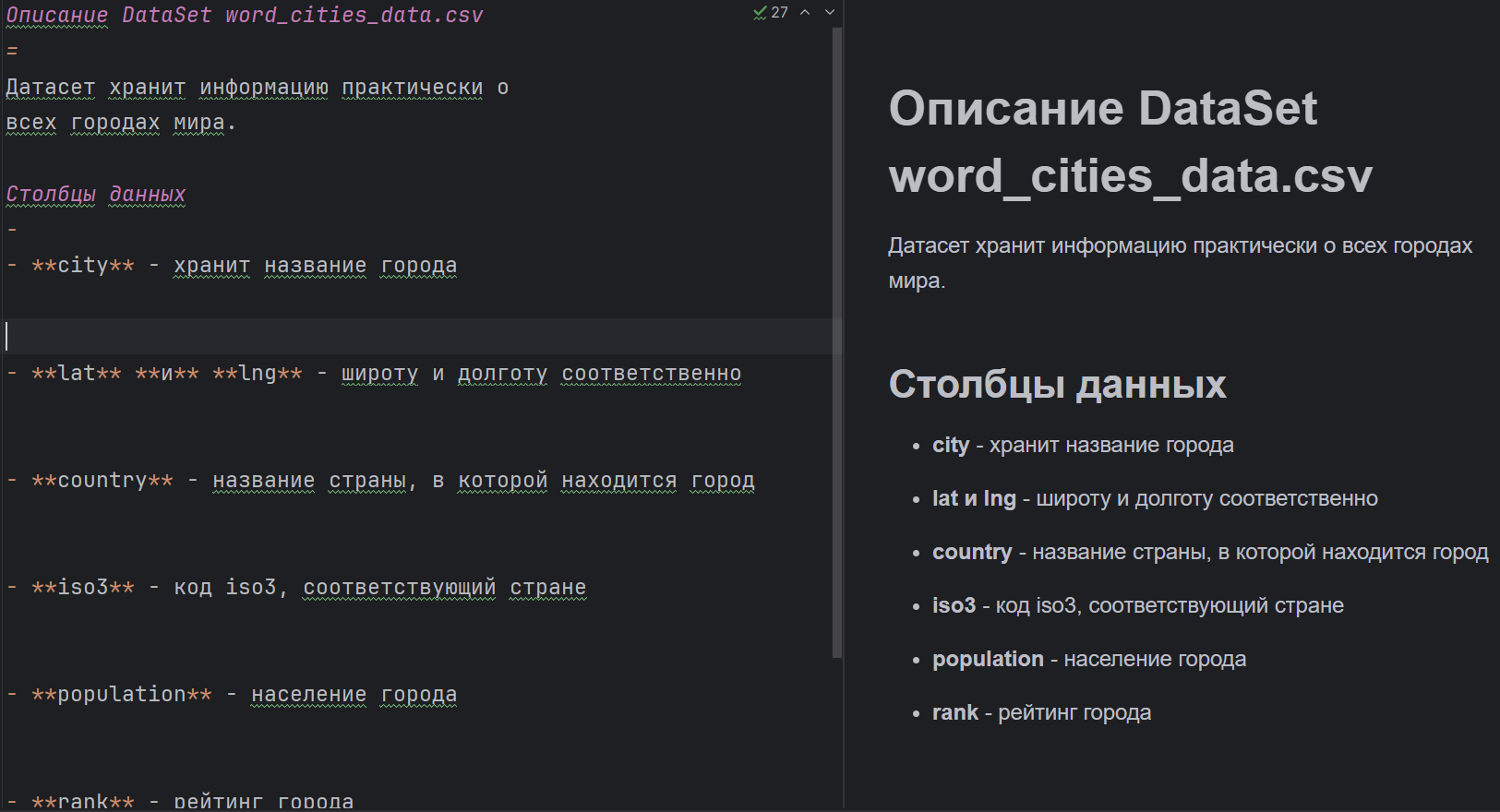


Рисунок 1 – Файл *describe.md*

Датасет представляет собой *csv*-файл, показанный на рисунке 2.

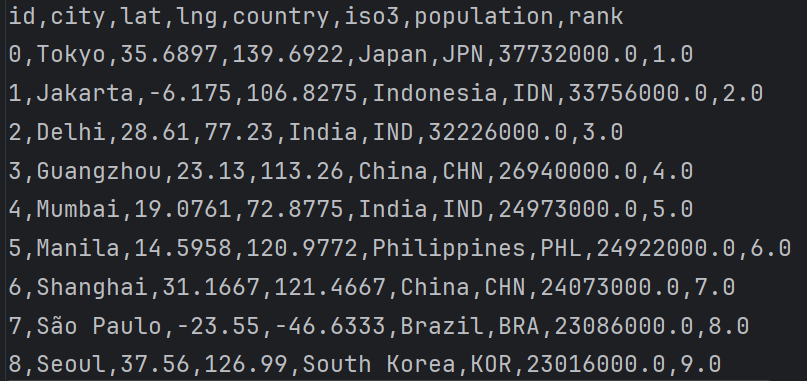


Рисунок 2 – Исходный датасет

Датасет был загружен в программу с помощью функции *read\_csv* библиотеки *Pandas* (рисунок 3).

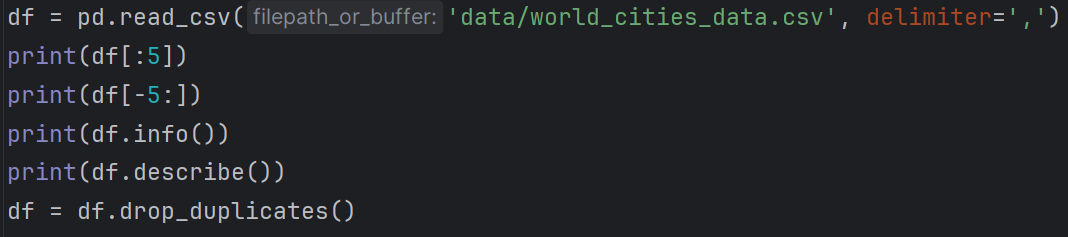


Рисунок 3 – Загрузка данных датасета в программу

На рисунках 4-5 представлены первые пять и последние пять строчек датасета.

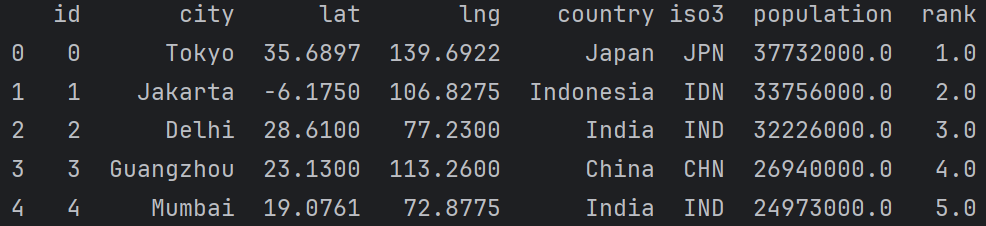


Рисунок 4 – Первые 5 строчек датасета

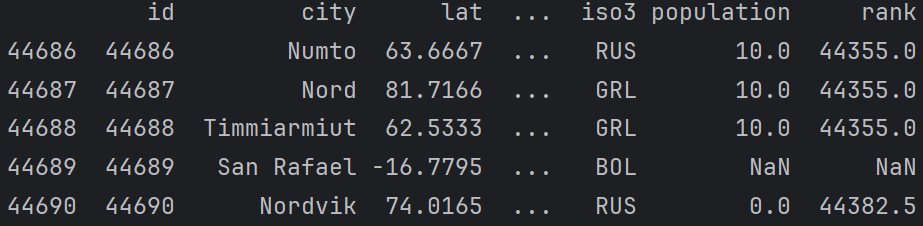


Рисунок 5 – Последние 5 строчек датасета

На рисунке 6 находится статистическая информация о датасете, а на рисунке 7 – информация о типах данных в наборе.

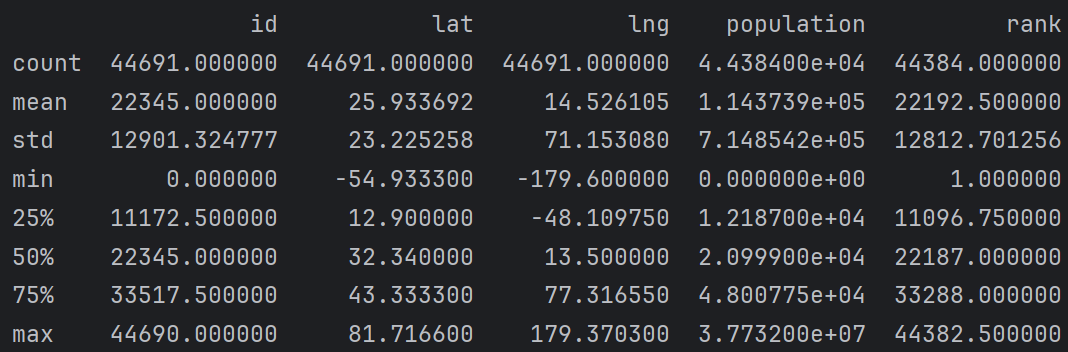


Рисунок 6 - Статистическая информация о датасете

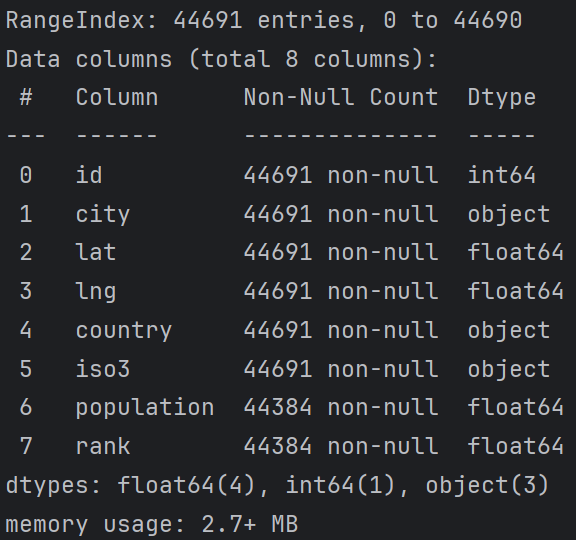


Рисунок 7 – Информация о типах данных в датасете

Были удалены дубликаты строк (рисунок 8) и поле *ID* (рисунок 9). Также поле «*rank*» переименовано в «*rating*» (рисунок 10).

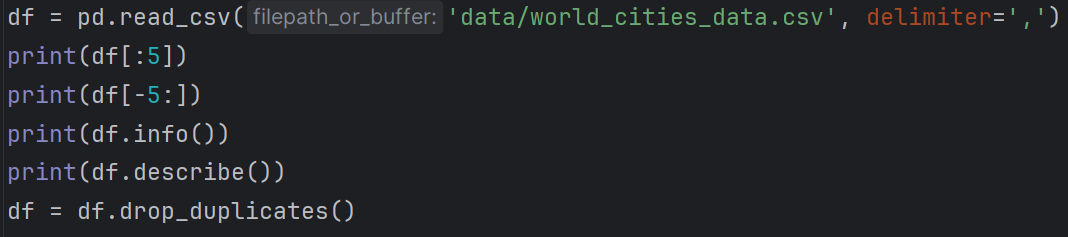


Рисунок 8 – Удаление дубликатов строк

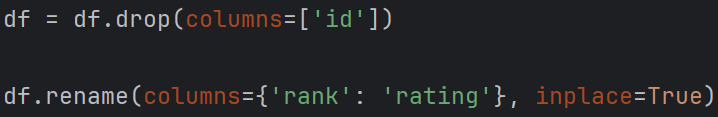


Рисунок 9 – Удаление поля *ID*

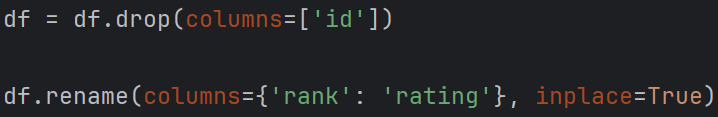


Рисунок 10 – Переименование поля «*rank*»

Таким образом, датасет был загружен в программу и подготовлен для дальнейшей работы с ним.

2 Разведывательный анализ данных

Был проведён анализ данных датасета. Инструменты библиотеки *Matplotlib* и *SeaBorn* позволили визуализировать данные. Была построена гистограмма распределения городов по широтам планеты (рисунок 11).

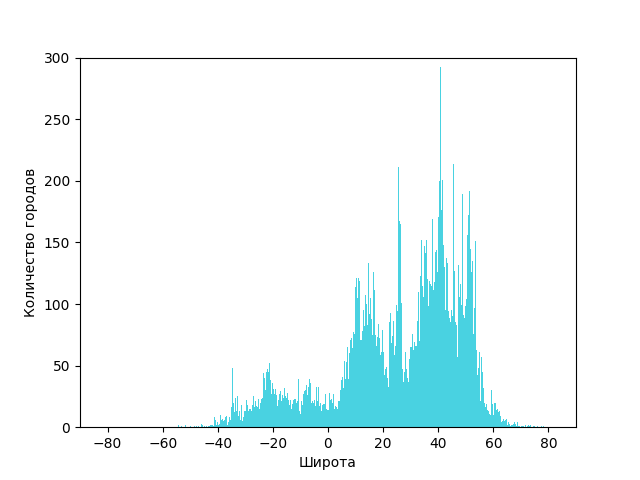


Рисунок 11 – Распределение городов по широтам

По гистограмме видно, что больше всего городов располагаются примерно на широте 42°.

Была построена диаграмма «ящик с усами» для долготы городов (рисунок 12).

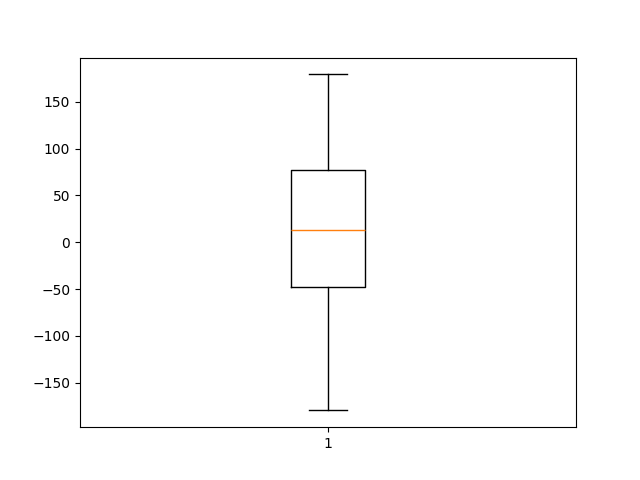


Рисунок 12 – Диаграмма «ящик с усами»

Эта диаграмма показывает, что медиана значений долготы приблизительно равна 10°. Её межквартильный размах идёт от -50° до 75°.

Далее была построена круговая диаграмма, в которой представлены страны с наибольшим количеством городов. Поскольку стран очень много, для удобства на диаграмме показано только 15 наиболее крупных стран, оставшаяся доля представлена в части «*Other*» (рисунок 13).

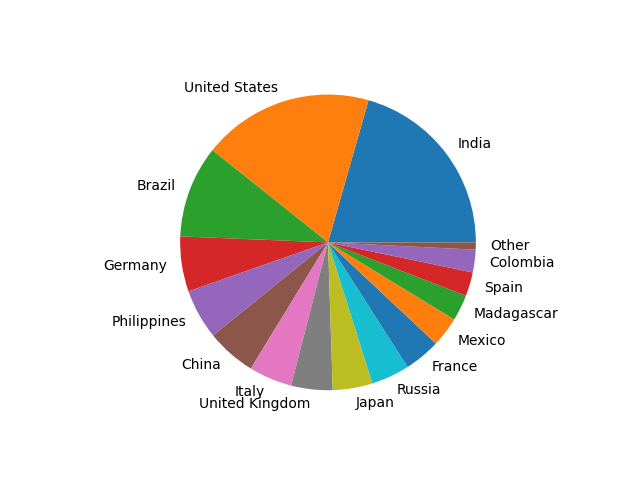


Рисунок 13 – Круговая диаграмма

По диаграмме видно, что на первом месте по наибольшему количеству городов находится Индия, на втором месте – США, на третьем – Бразилия.

Была построена тепловая карта со значениями взаимной корреляции между всеми парами признаков набора данных (рисунок 14).

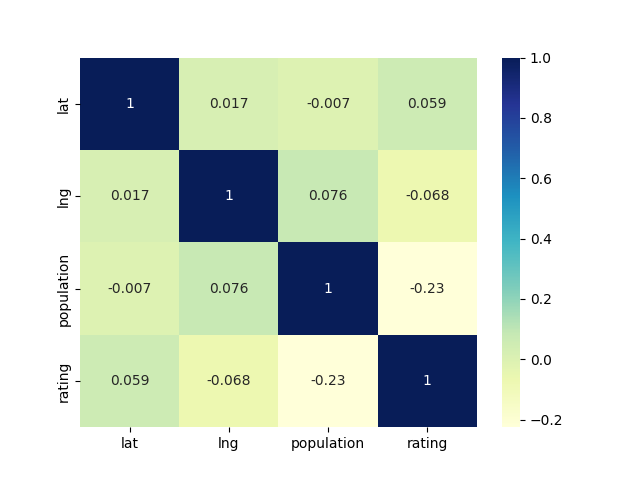


Рисунок 14 – Тепловая карта со значениями взаимной корреляции

Данная тепловая карта показывает отрицательную зависимость рейтинга города от его населения. Остальные пары признаков практически независимы друг от друга.

Также была построена диаграмма, показывающая население некоторых городов (рисунок 15).

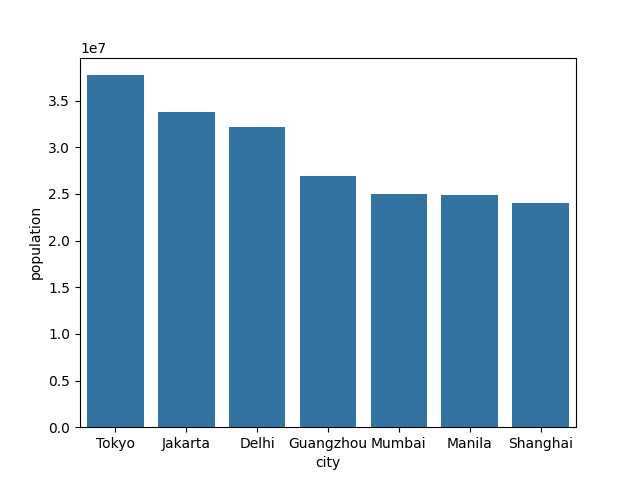


Рисунок 15 – Диаграмма населения городов

Среди выбранных городов самым густонаселённым является Токио.

Также была проанализирована на нормальность выборка из 200 первых значений населения городов (рисунок 16).



Рисунок 16 – Результат проверки выборки на нормальность

Выборка не оказалась нормальной.

На рисунках 17-22 представлены реализации построения диаграмм и проверки выборки на нормальность.

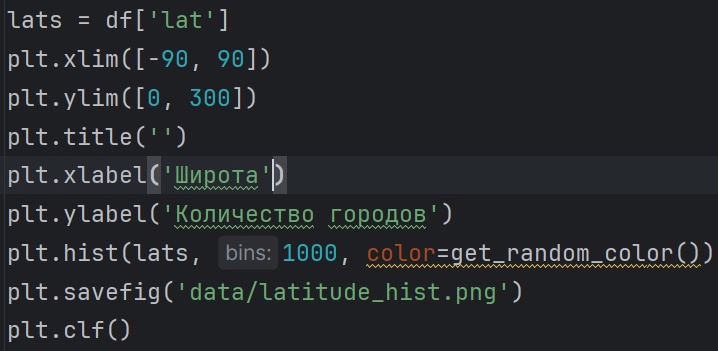


Рисунок 17 – Реализация построения гистограммы

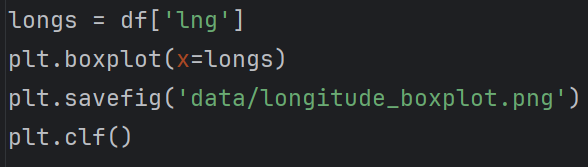


Рисунок 18 – Реализация построения диаграммы «ящик с усами»

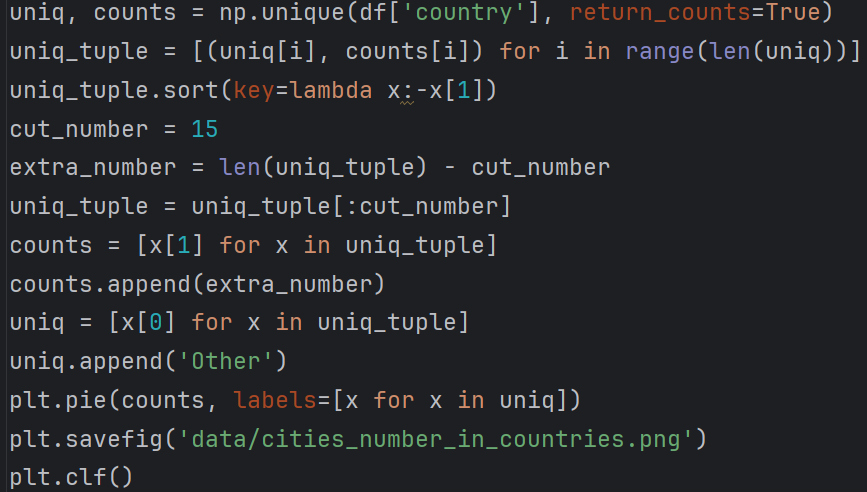


Рисунок 19 – Реализация построения круговой диаграммы

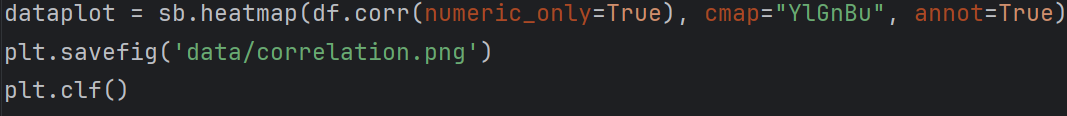


Рисунок 20 – Реализация построения тепловой карты

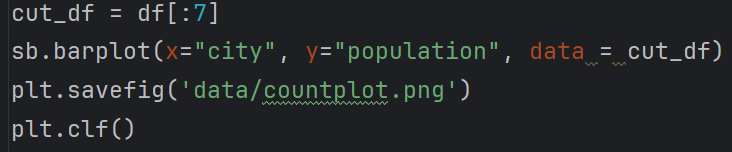


Рисунок 21 – Реализация построения диаграммы населения городов

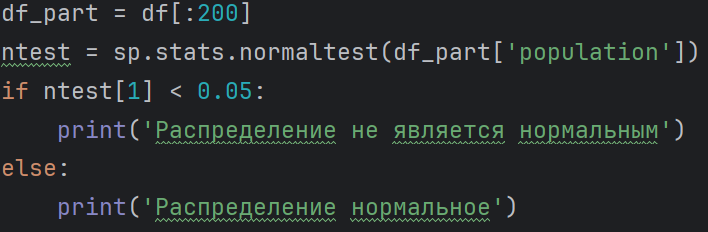


Рисунок 22 – Проверка выборки на нормальность

Таким образом, с помощью различных визуализаций был проанализирован датасет.

3 Предварительная обработка данных

Данные в датасете оказались неполными – нашлись пропуски в значениях. На рисунке 23 показан поиск полей, в которых присутствуют пустые значения, а на рисунке 24 – собственно поля с пропусками.

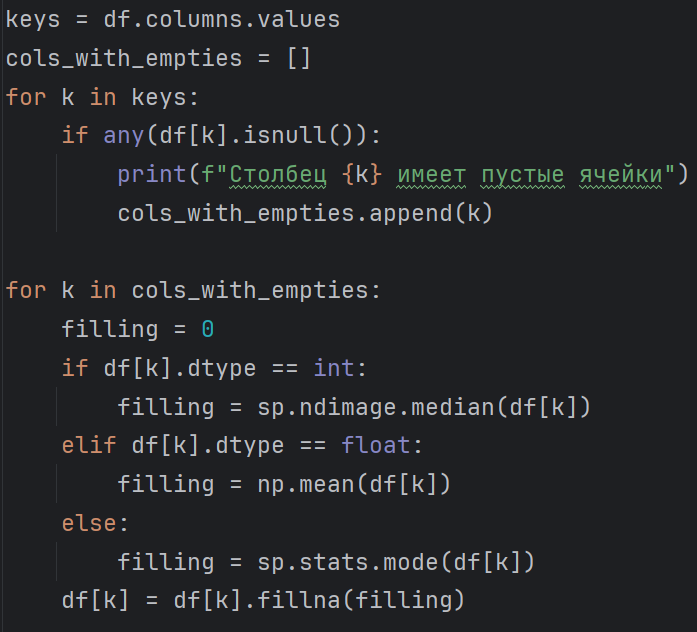


Рисунок 23 – Поиск столбцов с пропусками

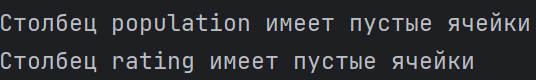


Рисунок 24 – Поля с пропусками

Пропуски были заполнены по следующему правилу: если значением признака является целое число, то пропуск заполняется значением медианы по столбцу, если – действительное число, то в пустое значение становится равным средним значением по столбцу, иначе – пропуск заполняется значением моды. Реализация этого правила представлена на рисунке 25.

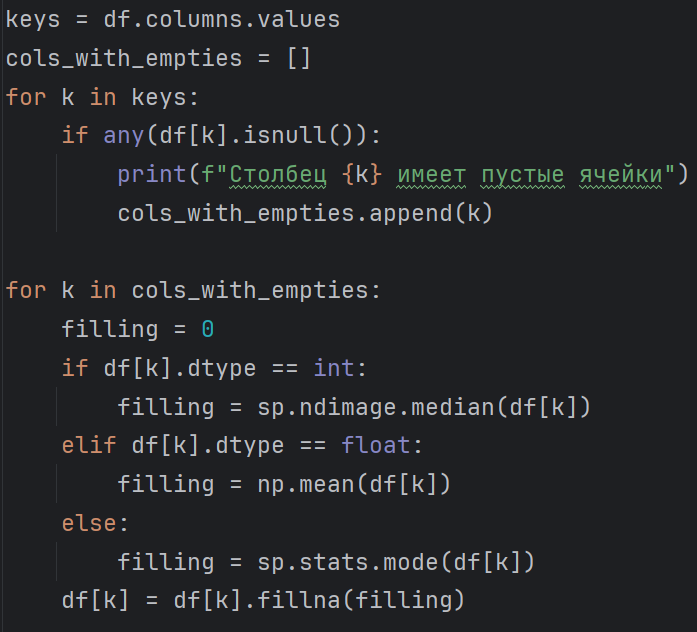


Рисунок 25 – Заполнение пропусков значениями

Результат заполнения пустых значений показан на рисунке 26.

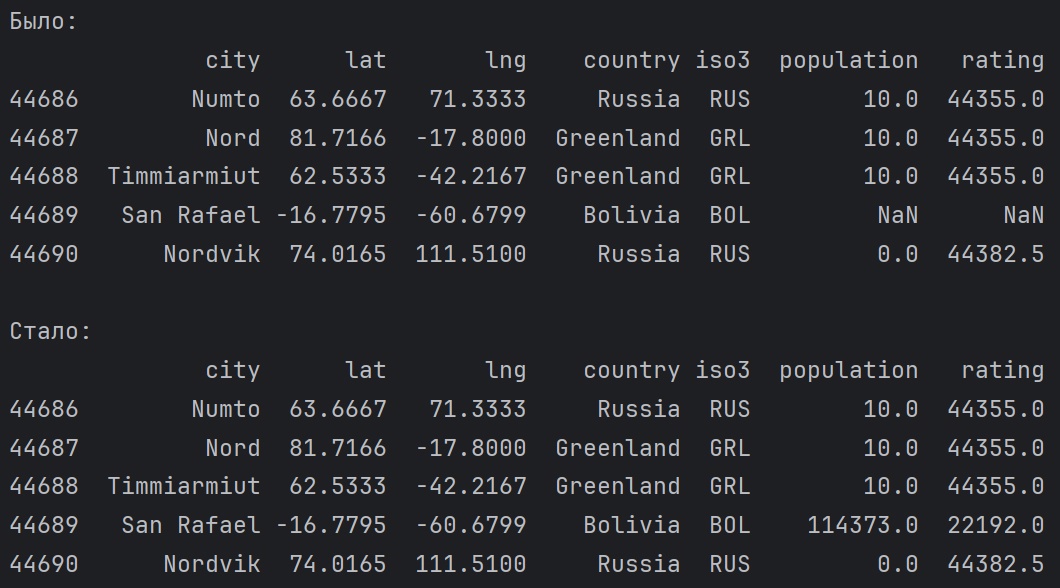


Рисунок 26 – Результат заполнения пропусков значениями

Как видно, значения предпоследней строчки вместо *NaN* стали равны медиане по столбцу.

Предобработанные данные подверглись *one-hot* кодированию, и затем с помощью метода *to\_csv* были сохранены в файл *cities\_stats.csv* (рисунок 27).

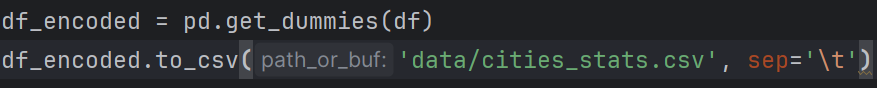


Рисунок 27 – Кодирование и сохранение данных

Как выглядит файл с закодированными данными, представлено на рисунке 28.

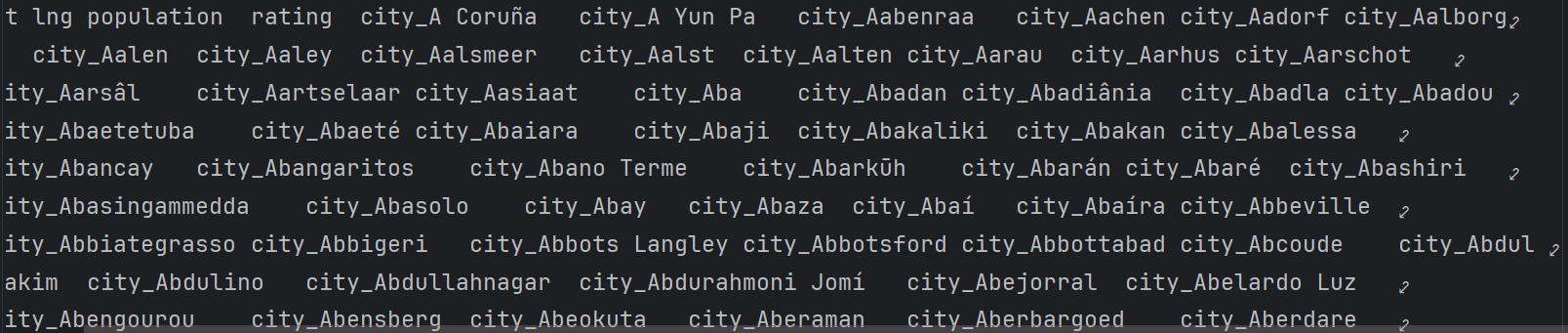


Рисунок 28 – Файл с закодированными данными

Таким образом были обработаны данные датасета: устранены пропуски в значениях. Данные были закодированы и сохранены в *csv*-файл.

ЗАКЛЮЧЕНИЕ

В ходе выполнения задания был:

– найден и загружен в программу датасет,

– написан *md*-файл, описывающий структуру датасета,

– произведен разведывательный анализ данных, в процессе которого были построены:

1) гистограмма распределения городов по широтам,

2) диаграмма «ящик с усами» по долготе городов,

3) круговая диаграмма, показывающая страны с наибольшим числом городов,

4) тепловая карта со значениями взаимной корреляции по всем парам признаков,

5) диаграмма населения городов,

– проверена на нормальность выборка из 200 значений населения городов,

– произведено *one-hot* кодирование всех категориальных признаков,

– данные сохранены в файл *cities\_stats.csv*.

В результате выполнения задания получены начальные навыки анализа данных датасета, работы с датафреймами и изучены виды диаграмм – «ящик с усами» и тепловая карта значений взаимной корреляции.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Skillbox – образовательная платформа. Библиотека Matplotlib для построения графиков, 2023. URL: <https://skillbox.ru/media/code/biblioteka-matplotlib-dlya-postroeniya-grafikov/> (дата обращения 26.11.2024)
2. Skillbox – образовательная платформа. Библиотека Numpy: все, что нужно знать новичку, 2024. URL: <https://skillbox.ru/media/code/biblioteka-numpy-vsye-chto-nuzhno-znat-novichku/> (дата обращения 24.11.2024)
3. ВикипедиЯ. Свободная энциклопедия. Ящик с усами, 2024. URL: <https://ru.wikipedia.org/wiki/Ящик_с_усами> (дата обращения 26.11.2024)
4. Романовский Р. К. Элементы теории вероятностей и случайных процессов: учеб. пособие*.* Омск: Изд-во ОмГТУ, 2015. – 116с.
5. Сайт Pandas. Pandas Documentation, 2024. URL: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html> (дата обращения 24.11.2024)
6. Сайт PythonRu. Руководство по Scipy: что это и как её использовать. URL: <https://pythonru.com/biblioteki/scipy-python> (дата обращения 24.11.2024)
7. Сайт Seaborn. seaborn-heatmap, 2024. URL: <https://seaborn.pydata.org/generated/seaborn.heatmap.html> (дата обращения 26.11.2024)