Segmentation and Clustering Districts in Penza

The main points of the project:

- Introduction
  - Description of the problem
  - Description of the data
- Data
  - Import Libraries
  - Prepping Data
- Exploring and clustering
  - Methodology
  - Explore data
  - Utilizing the Foursquare API
  - Cluster Districts
  - Analyze
- Results
- Conclusion

Introduction

This project was created as the final assignment for the Applied Data Science Capstone course on Coursera. Here we will describe a problem that can be solved using Foursquare location data.

1. Description of the problem

Have you ever rented a house? Today, a large number of people are renting apartments. There can be many reasons: moving to another city for study, not enough money to buy housing, a business trip, etc. However, **the problem of choosing the place** where it is necessary to stop **is now more urgent than ever**.

Thousands of sites with various information are ready to offer you a huge variety of apartment options for every taste.

- Here are some examples in Russia:
  - cian
  - avito
  - realty.yandex
  - domofond

Cheap and expensive, small and large, new and old, etc. But only a few of them give a rough understanding of the area in which the apartment is located. After all, an apartment may be wonderful, but the area will spoil the impression of it. Especially true in provincial cities. Often there is **little information available**.

In this project, we will conduct an up-to-date analysis and segment the districts of the city of Penza where apartments are rented.

The city of Penza was chosen, but the analysis is applicable to any other.

Some information about the city of Penza:

- Country - Russia
- Subject of the Federation - Penza Region
- Center of the European part of Russia

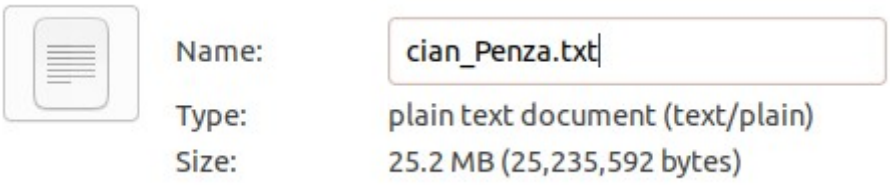
- Area 290.377 km²
- Founded in 1663
- Population 520,300 (2020)
- Penza is divided into four urban areas::
  - Железнодорожный (Zheleznodorozhnyy) — 114,408 people
  - Ленинский (Leninskiy) — 90,479 people
  - Октябрьский (Oktyabr'skiy) — 182,336 people
  - Первомайский (Pervomayskiy) — 133,077 people

The information was taken from [Wikipedia]

2. Description of the data

For the project I am using the following data:

1. Open Data from the cian website with current apartment rental offers on 03/02/2021
    - The raw data is a .txt file with HTML code obtained from the site.



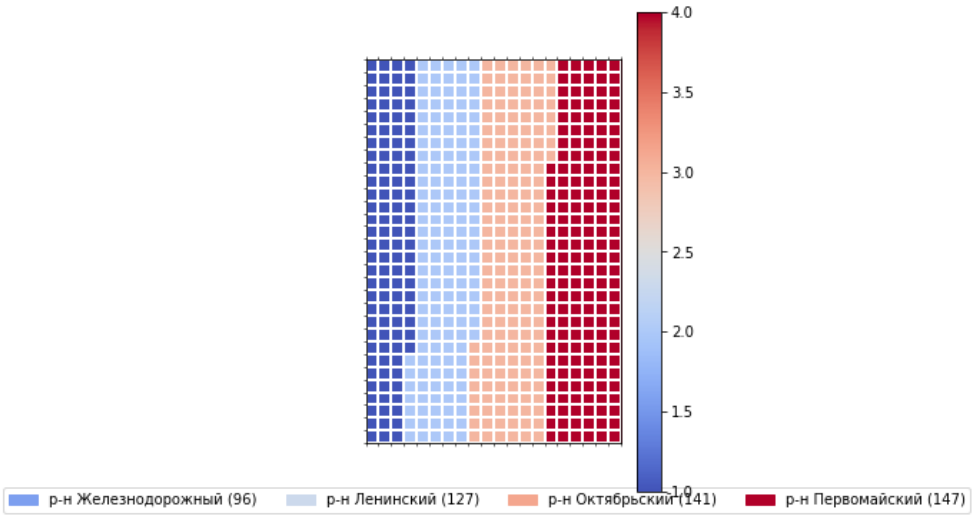
---

1. Location data retrieved from GeoPy
    - With this tool, we get **Latitude** and **Longitude** of Locations
1. The Foursquare location data

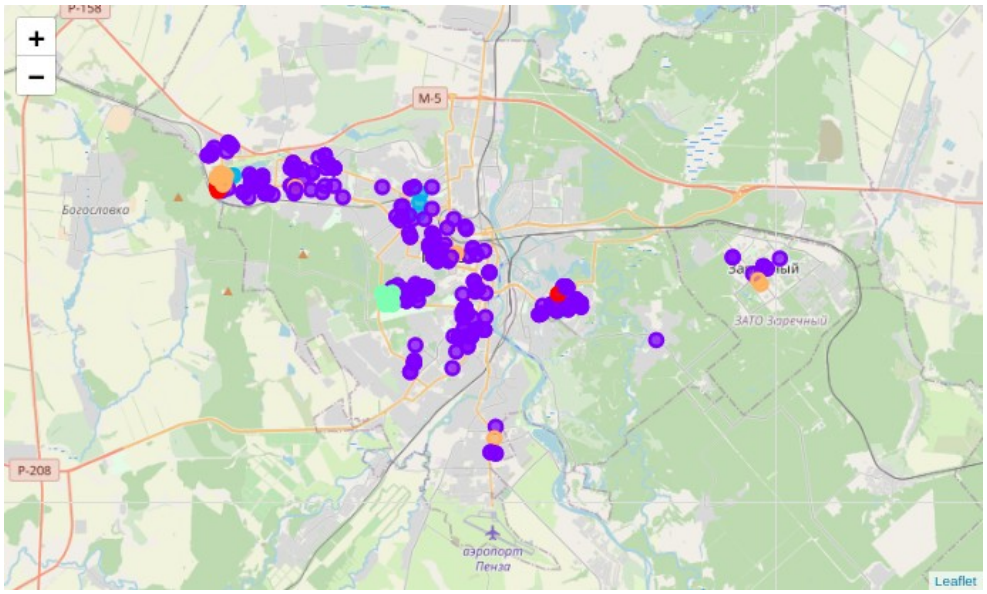
An example of a received ad:

```
data[0]
['Пензенская область',
 'Пенза',
 'р-н Октябрьский',
 'проспект Строителей',
 '152',
 'student city',
  'https://cdn-p.cian.site/images/43/309/101/kvartira-penza-prospekt-stroiteley-1019033481-4.jpg',
 '3-комн. кв., 64 м², 4/10 этаж',
 '11 000 ₽/мес.',
 '+7 960 325-16-...',
 '2 недели назад',
 '12 фев, 11:35',
 'От года, 11 000 ₽ + 2 000 ₽ комм. платежи (без счётчиков), комиссия 50%, без залога',
  'Сдам 3-х ком. квартиру на длительный срок по пр.Строителей 152 квартира чистая, теплая, мебелированная, имеется вся бытовая техника']
```

```
Total number of tiles is 600
р-н Железнодорожный: 113
р-н Ленинский: 149
р-н Октябрьский: 166
р-н Первомайский: 173
<Figure size 432x288 with 0 Axes>
```



Cluster Districts



Results

The analysis performed gave us the segmentation of the Penza city districts. We received 5 groups, each of which has its own average rental price.

Thanks to the **BeautifulSoup** library, we were able to render the HTML page. This gave us the basis of our data and the object of analysis. The data cleansing section showed how important it is to select the right information.

Next, we used the **geopy** library to get location data (latitude and longitude).

After that, we used **Foursquare** - API to obtain information on the coordinates of an object.

Data was visualized using the **folium** library.

For each group received, an analysis was carried out, in which information was disclosed, allowing you to rent a good apartment. The **k-means** algorithm was chosen for classification.

Conclusion

In this project, we carried out an up-to-date analysis and segmented the areas of the city of Penza where apartments are for rent. Based on the data obtained, we can choose an apartment based not only on the monetary parameter, but also on the classification of the area, infrastructure, personal preferences and location.

I did not begin to determine which apartment is better, since this is a purely individual task for everyone. After all, as I said in the introduction, there are a lot of reasons for renting an apartment. However, you now have enough data to make a choice.