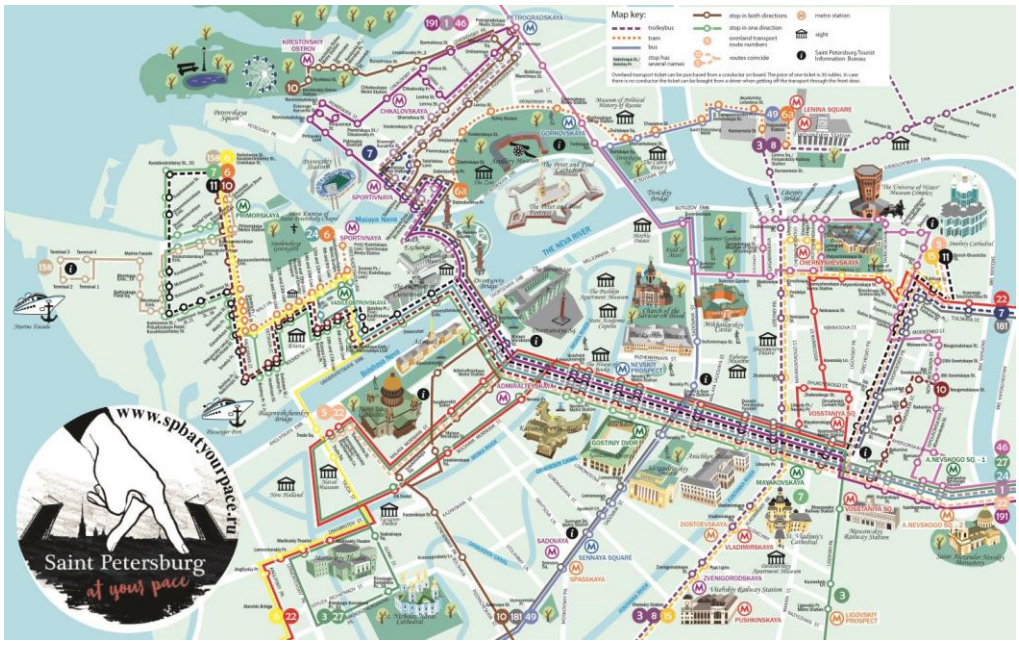
Choosing the district, where to open an Italian Restaurant in Saint-Petersburg, Russia

Aleksandr Kan

July 2021



1. Introduction

1.1.Background

Starting your own business is always a challenge and big risks, something can always go wrong. You may not have enough money, your product may not find a response from buyers, you may be wrong with the chosen marketing strategy, or choose the wrong place to open your business. Among other things, opening your own restaurant is one of the most difficult beginnings in my opinion. Choosing the right area to open is a difficult choice and a lot depends on it. Will there be a sufficient number of people around, how strong is the competition and if there is none at all, then perhaps for a reason.

1.2.Problem

In this study, we will consider one of the factors for the successful opening of a restaurant - the population density in a particular area of St. Petersburg and the number of Italian restaurants open around. We will be based on the fact that if there are no Italian restaurants in a certain radius in the area, then this area is not suitable for us, since most likely other restaurateurs conducted research and the absence of an Italian restaurant in it for a reason.

1.3.Interest

This research will be of interest to those who want to open their own restaurant and choose a suitable area in St. Petersburg. Also, the study will be interesting to anyone who wants to open a business, to see the competition around, since it will be enough just to change a few parameters. For example, by entering the coordinates of an existing grocery store, changing the search radius and the name of the place found in Foursquare, you will see the number of competitors around.

2. Data acquisition and cleaning

2.1.Data description

We will use different types of data in our research. First of all, we will need to determine the coordinates of the city of St. Petersburg, as well as the coordinates of the city districts. We need the coordinates so that using the application programming interface (API) we can see what is around a given point and determine the number of Italian restaurants. We will also need the population size in the city districts, as well as their area to determine the population density. The data sources for our research will be the site, where the coordinates of the city will be indicated. Wikipedia, from there we can take data on the population in each district, as well as the coordinates of the districts and the area of each district. And of course, data from Foursquare, from where we will take the data we need on the location of Italian restaurants.

2.2.Data sources

- Foursquare API
- Coordinates of Saint-Petersburg and all districts (Wikipedia)
- Square area and population of districts (Wikipedia)
- Python libraries:
- Pandas: to create and manipulate Data Frames

Folium: to show the map of the city

Geopy: to get the coordinates

BeautifulSoup: to parse necessary data from websites

Matplotlib: to plot regression

Scikit – learn: to get regression model

2.3.Data shaping and cleansing

First of all, I needed the coordinates of St. Petersburg. In order to find out the coordinates, I used two different methods. The first was to use the Beautifulsoup python library, which helps parse data from websites. I used the site:

https://www.gps-latitude-longitude.com/gps-coordinates-of-st-petersburg-russia

where the latitude and longitude of the city were presented in a table. Parsing without conversion allowed me to get data like this:

```
Latitude of st petersburg russia
59.934280
Longitude of st petersburg russia
30.335099
```

I converted this data into a tabular form using pandas:

	Place	Latitude	Longitude
0	Saint-Petersburg	59.934280	30.335099

I also decided to get the coordinates in a different way, using the Geopy library.

Geopy is a Python client for several popular geocoding web services, makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

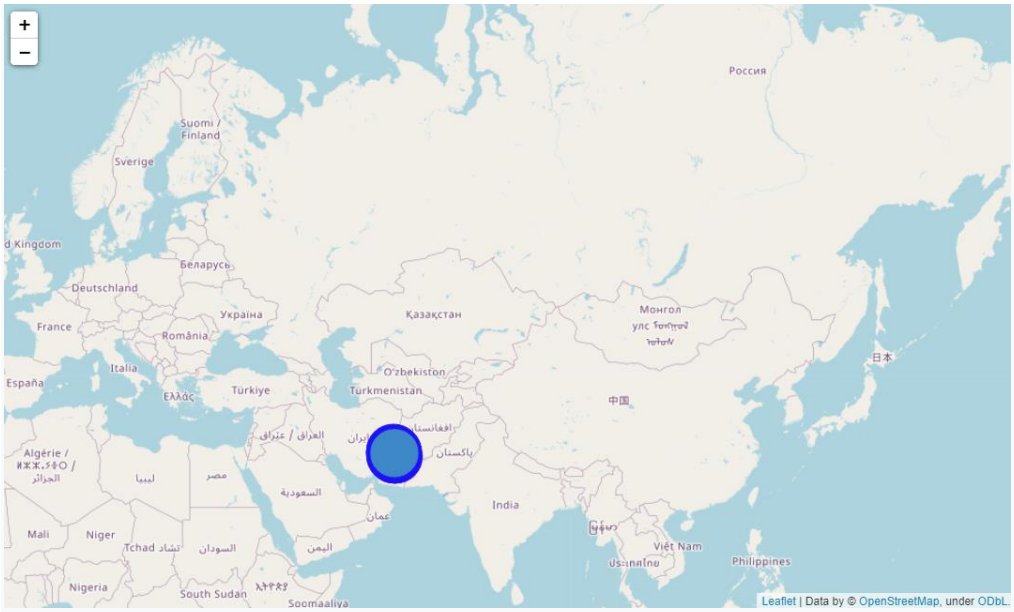
```
The geograpical coordinate of Saint-Petersburg City are 59.938732, 30.316229.
```

As you can see, both methods are working, difference in coordinates is miserable and both point the same city.

Next step is getting districts, it's square area and population. Instead of parsing necessary data step by step from different pages, I decided to copy the data to excel file and use panda to work with it. And here is the table I imported using pandas to my notebook.

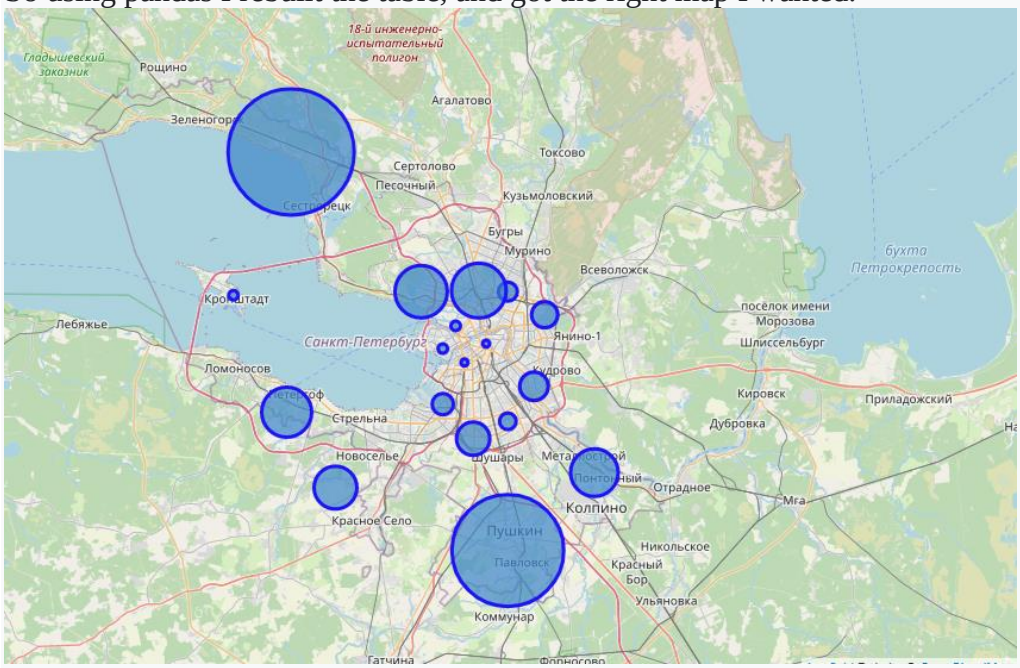
	District	Population	Square area	Longitude	Latitude
0	Admiraltejskij	156958	13,82	59.91683	30.30000
1	Central'nyj	210899	17,77	59.93860	30.35000
2	Frunzenskij	384385	37,52	59.84870	30.40000
3	Kalininskij	521875	40,18	59.99860	30.40000
4	Kirovskij	331550	47,46	59.86860	30.25000
5	Kolpinskij	194934	102,25	59.78920	30.59970
6	Krasnogvardejskij	355621	56,35	59.97200	30.48500
7	Krasnosel'skij	412886	90,49	59.77180	30.00230
8	Kronštadtskij	44353	19,53	59.99450	29.76680
9	Kurortnyj	78910	268,19	60.15930	29.90000
10	Moskovskij	347022	73,07	59.82860	30.32000
11	Nevskij	538323	60,66	59.88930	30.46030
12	Petrodvorcovyj	142655	107,08	59.85930	29.88970
13	Petrogradskij	125731	19,54	59.95910	30.27970
14	Primorskij	580100	109,90	59.99860	30.20000
15	Puškinskij	229403	240,09	59.69860	30.40000
16	Vasileostrovskij	205240	21,47	59.93283	30.25000
17	Vyborgskij	523497	115,52	59.99983	30.33301

After getting coordinates we can build a map using Folium library, so we could watch where districts locate in the city. But after building it I found that my districts are displayed somewhere between Iran and Afghanistan.



Overwatched coordinates in the table, mentioned that longitude and latitude should be reversed.

So using pandas I rebuilt the table, and got the right map I wanted.



Now we are ready to connect to Foursquare API to get venues. I took 2 km radius for each district and got 1 344 venues.

	0	1	2	3	4	5	6
0	Admiraltejskij	59.91683	30.30000	Усачевские бани. Купеческий класс	59.918335	30.299474	Historic Site
1	Admiraltejskij	59.91683	30.30000	Alexander House	59.919205	30.300146	Hotel
2	Admiraltejskij	59.91683	30.30000	The Seven Bridges Point (Семимостье)	59.920811	30.298543	Historic Site
3	Admiraltejskij	59.91683	30.30000	Палантин	59.913655	30.296419	Hotel
4	Admiraltejskij	59.91683	30.30000	Kryukov Canal (Крюков канал)	59.920978	30.299178	Canal
...	...	...	...	...	...	...	...
1340	Vyborgskij	59.99983	30.33301	Pandora	60.004540	30.299314	Jewelry Store
1341	Vyborgskij	59.99983	30.33301	Буквоед	60.002338	30.299127	Bookstore
1342	Vyborgskij	59.99983	30.33301	Black Cat Pub	60.013036	30.313266	English Restaurant
1343	Vyborgskij	59.99983	30.33301	Автосервис для Land Rover	60.000118	30.324379	Auto Workshop
1344	Vyborgskij	59.99983	30.33301	Ярославские бани	60.013204	30.321013	Bath House

Now it seems that we have all data needed, but I decided to check if I can check all venues only in one district, and got nothing:

District	District_Latitude	District Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category

So, I decided to check what is the problem and found that space symbol after the name of district looks like:

```
'Admiraltejskij\xa0'
```

So, I transformed the data and after the check everything seemed fine and groupby function worked as it should.

	District	District_Latitude	District Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Admiraltejskij	59.91683	30.3	Усачевские бани. Купеческий класс	59.918335	30.299474	Historic Site
1	Admiraltejskij	59.91683	30.3	Alexander House	59.919205	30.300146	Hotel
2	Admiraltejskij	59.91683	30.3	The Seven Bridges Point (Семимостье)	59.920811	30.298543	Historic Site
3	Admiraltejskij	59.91683	30.3	Палантин	59.913655	30.296419	Hotel
4	Admiraltejskij	59.91683	30.3	Kryukov Canal (Крюков канал)	59.920978	30.299178	Canal
...	...	...	...	...	...	...	...
95	Admiraltejskij	59.91683	30.3	WöD	59.932163	30.305139	Café
96	Admiraltejskij	59.91683	30.3	Lotte Hotel	59.931383	30.310401	Hotel
97	Admiraltejskij	59.91683	30.3	Додо Пицца	59.920642	30.317957	Pizza Place
98	Admiraltejskij	59.91683	30.3	Северянин	59.926393	30.312993	Russian Restaurant
99	Admiraltejskij	59.91683	30.3	Gutenberg Hotel	59.928221	30.312705	Hotel

3. Methodology

After how the data was collected and cleaned through various transformations, we can carry out some of the actions that we need for analysis. In order for a restaurant to be successful, it is necessary that the population density in the selected area is the highest. Therefore, in order to get the population density per square kilometer, we will add a population density column, which will be equal to the population divided by the area.

	District	Population	Square area	Latitude	Longitude	Population density
3	Kalininskij	521875	40.18	59.99860	30.40000	12988.427078
1	Central'nyj	210899	17.77	59.93860	30.35000	11868.261114
0	Admiraltejskij	156958	13.82	59.91683	30.30000	11357.308249
2	Frunzenskij	384385	37.52	59.84870	30.40000	10244.802772
16	Vasileostrovskij	205240	21.47	59.93283	30.25000	9559.385189
11	Nevskij	538323	60.66	59.88930	30.46030	8874.431256
4	Kirovskij	331550	47.46	59.86860	30.25000	6985.882849
13	Petrogradskij	125731	19.54	59.95910	30.27970	6434.544524
6	Krasnogvardejskij	355621	56.35	59.97200	30.48500	6310.931677
14	Primorskij	580100	109.90	59.99860	30.20000	5278.434941
10	Moskovskij	347022	73.07	59.82860	30.32000	4749.172027
7	Krasnosel'skij	412886	90.49	59.77180	30.00230	4562.780418
17	Vyborgskij	523497	115.52	59.99983	30.33301	4531.656856
8	Kronštadtskij	44353	19.53	59.99450	29.76680	2271.018945
5	Kolpinskij	194934	102.25	59.78920	30.59970	1906.444988
12	Petrodvorcovyj	142655	107.08	59.85930	29.88970	1332.228241
15	Puškinskij	229403	240.09	59.69860	30.40000	955.487526
9	Kurortnyj	78910	268.19	60.15930	29.90000	294.231701

Also, we need to know, how many of Italian restaurants are in the district (we took 2km radius, which doesn't cover all district, but enough for this research):

	District	Number_of_IT_Restar
0	Admiraltejskij	4
1	Central'nyj	3
2	Frunzenskij	2
3	Kalininskij	1
4	Kirovskij	3
5	Petrogradskij	3
6	Vasileostrovskij	1
7	Vyborgskij	1

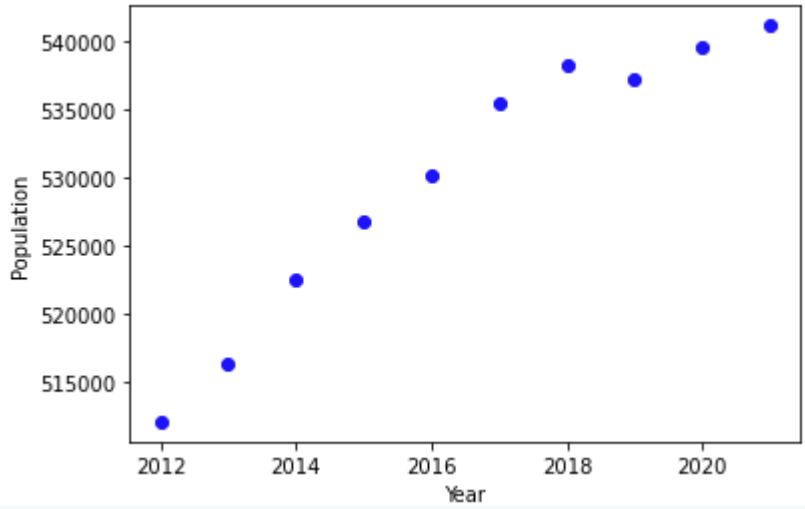
After that we need to merge the number of Italian restaurants with our table with density population. To understand the density for a restaurant, we need to divide the density population by number of restaurants and multiply by 2(because density is for 1 km and we took 2 km radius for Foursquare venues):

	District	Population	Square area	Latitude	Longitude	Population density	Number_of_IT_Restar	Pop_density_per_IT_Restr
3	Kalininskij	521875	40.18	59.99860	30.40000	12988.427078	1.0	25976.854156
16	Vasileostrovskij	205240	21.47	59.93283	30.25000	9559.385189	1.0	19118.770377
2	Frunzenskij	384385	37.52	59.84870	30.40000	10244.802772	2.0	10244.802772
17	Vyborgskij	523497	115.52	59.99983	30.33301	4531.656856	1.0	9063.313712
1	Central'nyj	210899	17.77	59.93860	30.35000	11868.261114	3.0	7912.174076
0	Admiraltejskij	156958	13.82	59.91683	30.30000	11357.308249	4.0	5678.654124
4	Kirovskij	331550	47.46	59.86860	30.25000	6985.882849	3.0	4657.255232
13	Petrogradskij	125731	19.54	59.95910	30.27970	6434.544524	3.0	4289.696349

We got that Kalininskij district has the highest density, but what if population lowers from year to year? This question leads us again to Wikipedia where we get the data of population in the Kalininskij district by year from 2012 to 2021. We copy it to .xlsx file and input to our notebook using pandas:

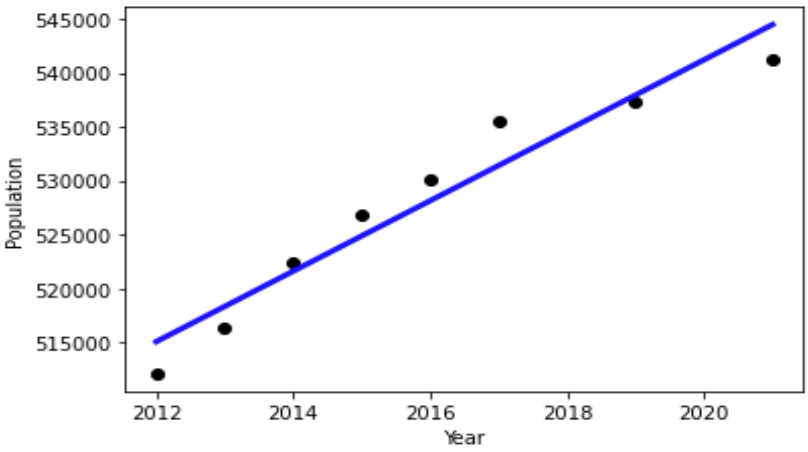
	Year	Population
0	2012	512121
1	2013	516312
2	2014	522493
3	2015	526876
4	2016	530163
5	2017	535428
6	2018	538258
7	2019	537280
8	2020	539600
9	2021	541200

We can see, that it increases, but let’s build a plot to see how it distributes on graph:



As we can see simple linear regression model would fit our data well, so I built it with scikit-learn:

```
Coefficient of determination: 0.93
```



4. Results

After receiving the data, processing and clearing them, we received an answer to our question, in which district of St. Petersburg to open an Italian restaurant and this is the Kalininskij district. Since we took for the analysis only the indicator of population density per 1 Italian restaurant in a particular area, we checked whether the population will grow or, on the contrary, will fall in the selected area. By constructing a simple regression model, we made sure that the population in this area is likely to grow. The determination coefficient was 0.92, which is a good result.

5. Discussion

In the process of research, I was faced with the fact that many of the data that I needed are very fragmented and require efforts to collect them from different sources, and then process them, since they are usually in a form that is not suitable for applying the methods and functions used in data analysis. One of the banal recommendations after this study is that companies set up data collection correctly and better, so that for this they involve a person who will analyze this data in the future. As for the result obtained, as a resident of St. Petersburg, I am not surprised by the result obtained. Subjectively, the Kalininskij district is indeed a good place to open an Italian restaurant. It was also a good decision to check if the population in the selected area is falling, as in many areas of the city the population is falling. Also, the collection of the same data from different sources, as in the example with obtaining coordinates from the site and the Geopy, showed that the data may differ (in our case, it is not critical). Therefore, another recommendation is to check, if possible, the data that you extract from certain sources, if there is such a possibility (which of course does not always happen)

6. Conclusion

To In conclusion, I would like to note that to solve such a problem, I used only 1 factor - population density and the number of restaurants already open, since if you delve deeper into this study, you can go deeper for a couple of months for sure. For example, it is possible to assess the level of well-being of areas using data on the cost of housing, which can be obtained through the IPA of the cyan site(site and app aggregating cost of property), but as far as I understand this will be a paid service. Depending on the cost of housing, we understand the level of income of people living in the area and can guide the future owner of the restaurant what price he can set for his dishes. There are a lot of similar things for research, and if you get carried away, you can go deeper into research for the sake of research, forgetting the purpose of the project. The result obtained is sufficient to determine the choice of the area and opens up new opportunities for a deeper study of the area itself, in order to determine the best place in the Kalininskij district.