The Battle of Neighborhood

...

Opening Restaurant

Introduction

Nowadays, it is difficult to imagine a city without a restaurant or a venue for food where people can have a meal or drink.

The city of my choice is Taganrog that has following industry and businesses:

- Aerospace
- Steel
- Food
- Farming
- See port
- etc.

That means there are a lot of business opportunities for restaurant business what leads to high competition.

Problem

There are a lot of business opportunities for restaurant business what leads to high competition.

To survive in such competitive market it is very important to find right place and take into account many other important factors such as:

- City population
- Sport and Entertainment zones
- Food markets with products of local farmers
- Local competitors and their ratings
- etc.

Data Description

For further analysis will be used the following data sources:

- Wikipedia page for city population (https://en.wikipedia.org/wiki/Taganrog)
- Nominatim search engine for OpenStreetMap data to get the bounding box of the city (https://nominatim.openstreetmap.org/search?format=json&q=Taganrog&polygon_geojson=1)
- Data from Foursquare API
    - search endpoint
    - categories (list of all categories)
    - details (full details about a venue)

Data Example

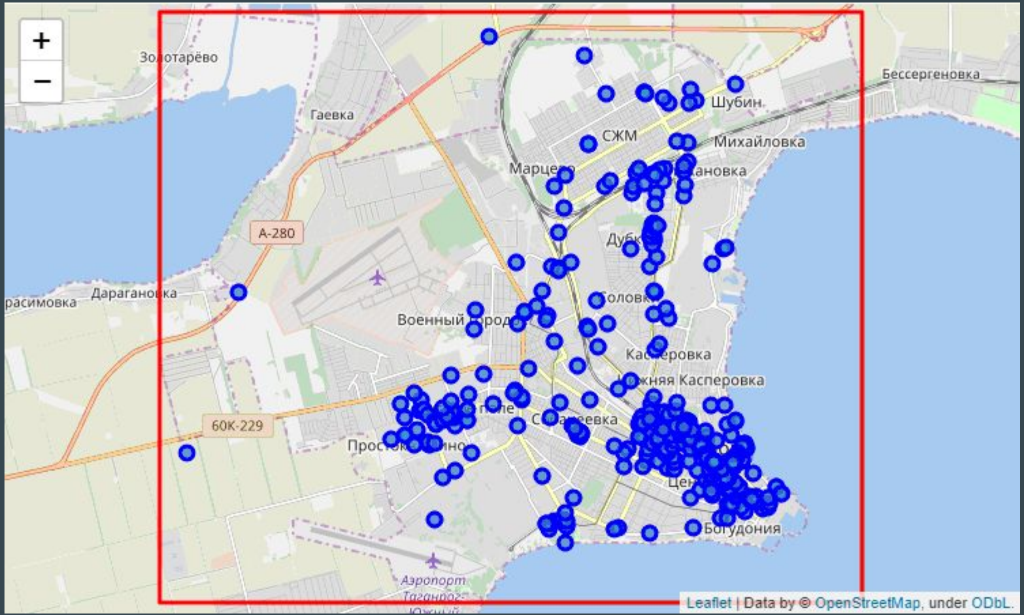
- Taganrog city has no neighborhood division, that why a coordinate grid covering the entire city by cells will build
- The first five venues within such cell (about 700 meters) from foursquare data

	Venue	Latitude	Longitude	Category	Id
0	Площадь перед администрацией города	47.215733	38.928230	Plaza	5368f4ad498ea0cb80cef632
1	Культ вина	47.215510	38.929310	Wine Bar	5c74142e60255e002c1aefbc
2	Театр имени А. П. Чехова	47.216325	38.928217	Theater	4dcbe98a1f6ea1401d49d12a
3	Администрация Таганрога	47.215517	38.928420	City Hall	4da693d90cb66f658708dafc
4	Л'Этуаль	47.215416	38.929266	Cosmetics Shop	4f83002ee4b0b2237e8a6cb1

Data Collection

This section covers data acquisition steps:

- Display venues on map with the bounding box
- Get detailed information about each venue and put it in a new dataset



Methodology

Understanding and preparation

Data needs to be cleaned and prepared

Fields renamed and NaN's fixed

Exploratory analysis

- Check distributions,
- Detect any outliers,
- Exploring correlation between fields
- Discarding erroneous or not needed data

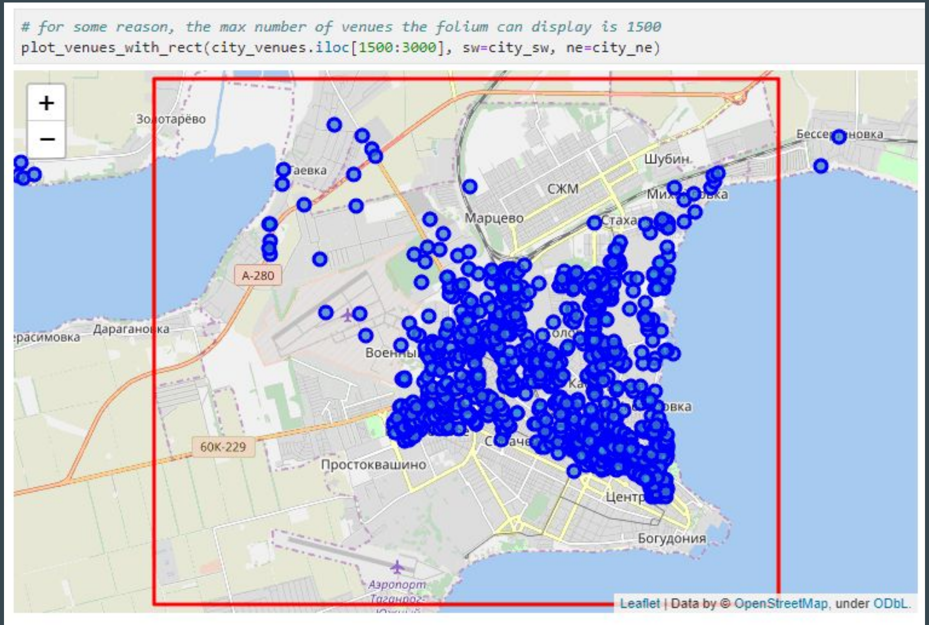
Clustering

- Determine the optimum number of clusters
- Finding suitable cluster

Understanding and Preparation

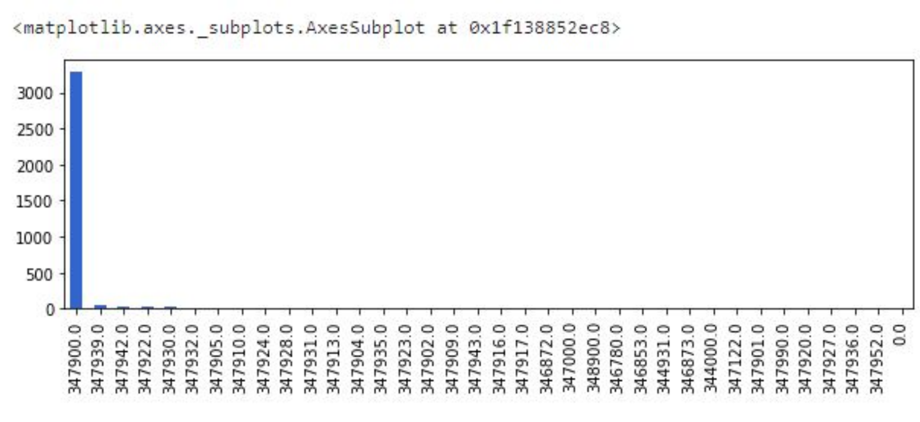
There are a number of venues out of city bounding box.

Remove them.

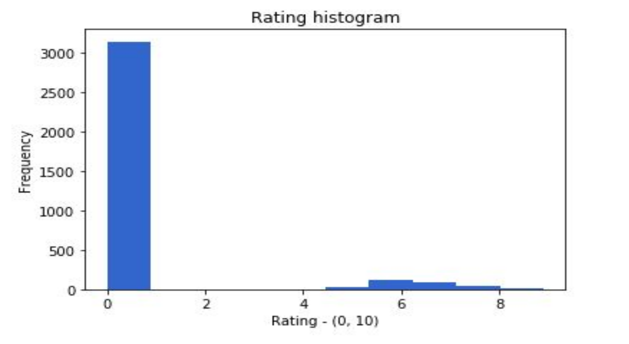


Exploratory Analysis

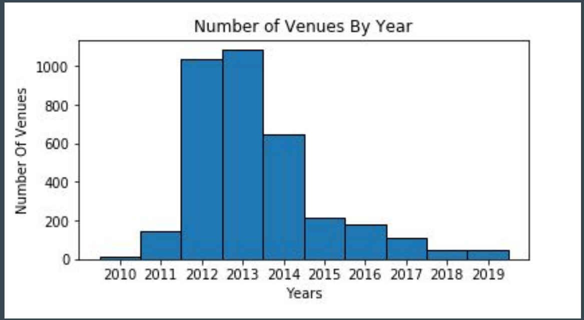
Explore the PostalCode field and plot its values.



Next explore the Rating field and plot its histogram.



Convert timestamp in seconds to years and find a range of years



Correlation between fields

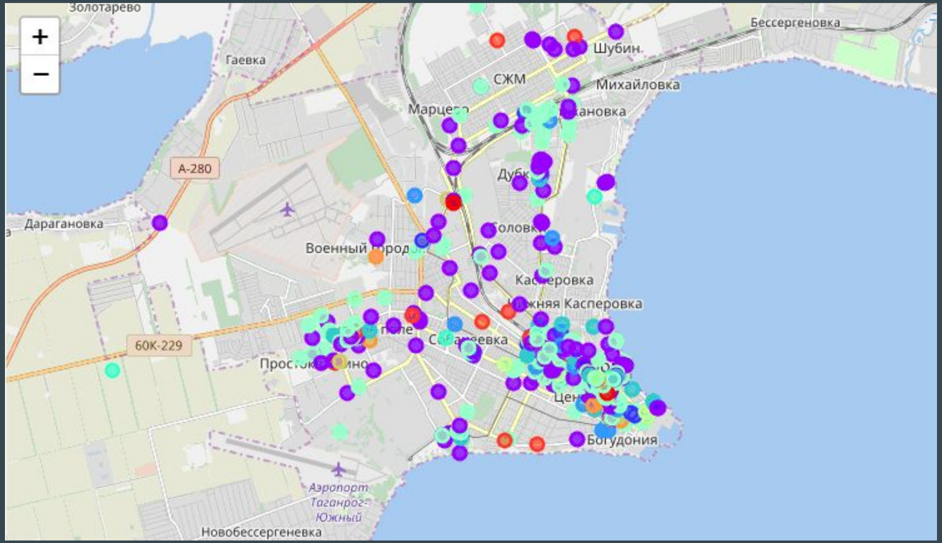
	Latitude	Longitude	Rating	Likes	Tips	Price_tier	Food_venue	Created_year	Venue_age
Latitude	1.000000	-0.115304	-0.045754	-0.061645	-0.062015	-0.054484	-0.076316	-0.026311	0.026311
Longitude	-0.115304	1.000000	0.055542	0.068591	0.080413	0.087981	0.088966	0.025289	-0.025289
Rating	-0.045754	0.055542	1.000000	0.495428	0.435916	0.153763	0.142334	-0.151191	0.151191
Likes	-0.061645	0.068591	0.495428	1.000000	0.871930	0.144708	0.128512	-0.128767	0.128767
Tips	-0.062015	0.080413	0.435916	0.871930	1.000000	0.203122	0.182537	-0.117383	0.117383
Price_tier	-0.054484	0.087981	0.153763	0.144708	0.203122	1.000000	0.673304	0.046388	-0.046388
Food_venue	-0.076316	0.088966	0.142334	0.128512	0.182537	0.673304	1.000000	0.095791	-0.095791
Created_year	-0.026311	0.025289	-0.151191	-0.128767	-0.117383	0.046388	0.095791	1.000000	-1.000000
Venue_age	0.026311	-0.025289	0.151191	0.128767	0.117383	-0.046388	-0.095791	-1.000000	1.000000

Summary.
Almost all variables demonstrate very weak positive and negative correlations except those that obviously related to each other. For example **Tips**, **Likes** and **Rating** correlate between each other as well as **Food_venue** and **Food_tier**.

Clustering

The core of the analytic method is clustering of all venues in Taganrog city by its fields, determining the optimum number of clusters and find suitable cluster for a new restaurant.

- Transform categorical field **Category** to dummies
- Normalize the dataset
- Run k-means to cluster all venues into 11 clusters
- Assign labels to venues
- Display our clustered venues on map



Results

To help with interpreting the data, let's sort the clusters by a meaningful attribute. A good candidate for such attribute is total number of venues in each cluster.

1. Clusters 1 and 6 show that the number of restaurants is high and it is risky to open a restaurant here
2. Cluster 4 contains of cuisines of different countries: Eastern European, Italian, Japanese
3. Assign labels to venues
4. The rest clusters have a small number of venues and introduced by a single category

	Label	1st	2nd	3rd	4th	5th	Total
1	1	Restaurant	Sushi Restaurant	Pizza Place	Snack Place	BBQ Joint	162
6	6	Café	Cafeteria	Bistro	Caucasian Restaurant	Pastry Shop	146
4	4	Eastern European Restaurant	Italian Restaurant	Restaurant	Japanese Restaurant	Café	27
7	7	Coffee Shop	None	None	None	None	23
2	2	Fast Food Restaurant	None	None	None	None	18
10	10	Bakery	None	None	None	None	15
3	3	Diner	None	None	None	None	12
9	9	Burger Joint	None	None	None	None	10
8	8	Sandwich Place	None	None	None	None	9
0	0	Dessert Shop	None	None	None	None	7
5	5	Modern European Restaurant	None	None	None	None	5

Discussion:

One interesting observation is that small restaurants (such as Cafe, Cafeteria, Sushi, Pizza, etc.) prevail over big restaurants.

There is a small number of restaurants of different cuisines and a recommendation might be to open a restaurant of national cuisine (for instance Indian Restaurant).

Conclusion

- The project analyzes venue similarity based on data obtained from Foursquare.
- The data needed cleaning as it contained nonexistent venues and missing data. We can observe a lack of restaurants of national cuisines and suitable cluster for opening a restaurant.
- However, we wish Foursquare had returned more venues as the data was very sparse and difficult to use for clustering. In future work, it may also be useful to include data pertaining to schools, universities, and such.