Saint-Petersburg District Clustering Project Report

by Artem F., 2020

Introduction/Business Problem

Prior setting for this project is a hypothetical restaurateur's request for market analysis. The well-known restaurateur decided to open a new venue in Saint Petersburg, Russia. It would be an obvious suggestion that the most common place for it is the center of the city. But there are several districts in the central area. Which one is the best candidate? Area is one of the most major aspects of the site selection as Russian cities are widely spaced along the land. Also, there are too many competitive restaurants may be placed there, so it is important to find the best possible place to get more profit for effective pay back.

The problem of this project is to define better district for opening a new high cuisine restaurant based on points:

1. Focus on customers with high-income
2. Total population of districts
3. Less possible competitive venues along the area
4. Venue's cuisine selection for district

Data

For this project we will use data:

- Foursquare data about venues (via Foursquare API)
- Administrative divisions of Saint Petersburg (parsed directly from Wikipedia with BeautifulSoup4 library)
- Statistics data of average income from Official Statistics Department of Russia in Saint Petersburg from jan to sept 2018
- Price of land from Rusland SP
- Price of land from “Restate” agency

Data description:

Foursquare API will return data about venues in format listed below. We use following tags for our dataset from the Object:

- name
- location
- categories
- price

Object example


```
{'id': '4f3232e219836c91c7bfde94',
'name': 'Conca Cucina Italian Restaurant',
'location': {'address': '63 W Broadway',
        'lat': 40.714484000000006,
        'lng': -74.00980600000001,
        'labeledLatLngs': [{'label': 'display',
        'lat': 40.714484000000006,
        'lng': -74.00980600000001}],
'distance': 469,
'postalCode': '10007',
'cc': 'US',
'city': 'New York',
'state': 'NY',
'country': 'United States',
'formattedAddress': ['63 W Broadway',
        'New York, NY 10007',
        'United States']},
'categories': [{'id': '4d4b7105d754a06374d81259',
```

'name': 'Food',
'pluralName': 'Food',
'shortName': 'Food',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/default_',
'suffix': '.png'},
'primary': True}],
'referralId': 'v-1588466252',
'hasPerk': False}

Yandex.Maps API will return data about location position:

- position (str)

Administrative divisions of Saint Petersburg is a table with the shape 5 x 19 shape with columns:

- index (int)
- district name (str)
- population (int)
- district size (float)
- density of population (float)

Statistics data of average income from Official Statistics Department of Russia in Saint Petersburg from jan to sept 2018 is a xmlx table with the shape 2 x 19 with columns:

- district name (str)
- average salary (int)

Price of land from Rusland SP is a table with the shape 3 x 17 (some districts are missing) with columns:

- index (int)
- district name (str)
- average m2 land cost (float)

Price of land from “Restate” agency is a table set (19 units) with different shapes. Shape of the specific table depends on district's non-living building cost per m^2 with columns:

- date (date)
- type of building average m2 cost (float)
- type of building average m2 cost dynamic (float)

Methodology

1. Finding best Radius param for Foursquare API

    - Building Voronoi diagram and closest vertice calculation
    - R-function (geometric)
    - R-function (with square)
    - R-mean (limited to 12k)

2. Venues assignment to nearest district

    - Collecting duplicates
    - Searching the least dist value

3. Venues data wrangling

    - Collecting venues data for all Districts
    - Duplicates removing
    - Generating top 10 district venues data with one hot encoding method
    - Generating price policy labels and stats for venues with one hot encoding method

4. Finding the best Features for a number of high cuisine restaurants. We assume that all restaurants with top price are the best match for our target venue:

    - Correlation matrix with main features with the Pearson evaluation method
    - Selecting values with corr score > 0.4

5. Finding the best K for KMean clustering algorithm with metric scores:

    - Preprocess Features data for a model
    - Davies Bouldin score
    - Calinski Harabasz score
    - Silhouette score

6. Building the KMean model with the best cluster-number for Districts clustering

7. City statistics comparison with resulting district data

- Creating top 30 and top 10 city venues lists
- District and city tops comparison

Results

There are 2 type of districts (from best correlation score) in the city:

- High Income - 4 Districts
- Medium / Low Income - 14 Districts

The most appropriate District for a new venue is Central (Centralniy) District as:

- Avg income rate is very high
- No competitive venues along the area
- There are over 200k population

The best match for a new venue (according to city top):

- Pizza Place
- Sushi Restaurant
- Middle Eastern Restaurant
- Fast Food Restaurant

Discussion

There might be some drawbacks in methodology:

- Incomplete or wrong venue data in Foursquare
- Minor Radius inaccuracies
- Venue assignment to the District may be inaccurate due to direct distance calculation and the results may be improved by smart distance matrix
- Venue type selection uses data from all price type venues and may be too noisy as high cuisine restaurants share is low in comparison with all dataset
- We don’t have enough data to evaluate venue type as there is no targeted customers data

As far as we can guess Easter and Italian cuisine are the most popular in the city. But this prediction is pretty blant.

Conclusion

This project has approved our main point and made predictions for a central location in Saint-Petersburg. Also it described tastes of SPb citizens which can be used for future investment.