Clustering Russian cities to determines food tastes in different regions

Alexander V. Ivanov

11.08.2020

1. Introduction

1.1 Background

Russia or the Russian Federation is one of the largest country in the world area. Its territory extends from Baltic Sea in the west to the Pacific Ocean in the east. Russia is the most populous nation in Europe. It is a multi-national state with over 185 ethnical groups, the population of those groups very enormously, from millions (Russian and Tatars) to under 10.000(Samis and Eskimos)

1.2 Problem

Because Russia's multi-national and wide-spread territory Russian's people tastes can vary from one region to another. So if your company is big and you what to expand your restaurant busines you need to know what kind of food/kitchen people prefer in desirable region.

2 Data acquisition and cleaning

2.1 Data sources

In order to provide information about Russian people tastes in different regions we need collect data containing Russian cities including theirs coordinates and we need information about food points in these cities such as restaurants, café, etc. And we need to know type of these point, what kind of food people prefer.
For Russian cities we will use information from Wikipedia page:

http://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Russia_by_population

We will read it into pandas data frame. To determine cities coordinates we will use Nominatim API.
All data related to the food points will be obtained via the FourSquare API.
So we build pandas dataframe containing cities, its coordinates and food points in these cities. After we build dataframe, we can cluster and visualize all this information to determine food tastes in different Russian cities or regions.

2.2 Data cleaning

In Wikipedia page we need only City Name. So we need to drop all other columns. And we need to correct some city coordinates, such as Saint Petersburg, because derived from Nominatim coordinates do not pointing to the city center. So our data with cities and coordinates will look like following:

	City	Latitude	Longitude
0	Moscow	55.750446	37.617494
1	Saint Petersburg	59.960674	30.158655
2	Novosibirsk	55.028217	82.923451
3	Yekaterinburg	56.839104	60.608250
4	Kazan	55.782355	49.124227

3.Metodology

We will use following libraries:

1)Pandas for manipulate and data analysis.

2)Requests that is used to receive json files from FourSquare.

3)Matplotlib for plotting.

4)Nominatim API to determine cities coordinates.

5)Folium to visual analyze dataframes and results.

6)sklearn to cluster derived data

First step is to get dataframe with city names and its locations. We get it by parsing Wikipedia page and using Nominatim to get coordinates.

Second step is to create dataframe with cities and corresponding venues. We will use FourSquare API for this and we will limit our requests by maximum of 100 venues per city and search radius 5km:

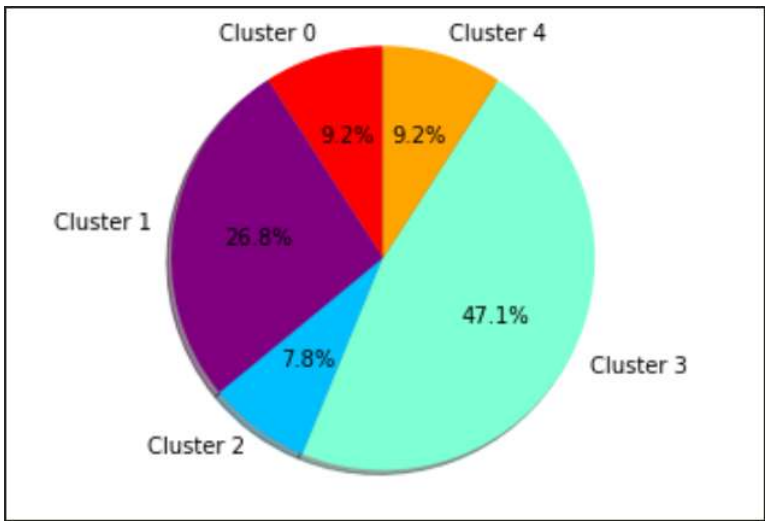
	City	City_Latitude	City_Longitude	Venue	Venue_Category
0	Moscow	55.750446	37.617494	Beluga (Белуга)	Russian Restaurant
1	Moscow	55.750446	37.617494	Траппист	Belgian Restaurant
2	Moscow	55.750446	37.617494	Dr. Zhivago (Dr. Zhivago (Dr. Живаго))	Russian Restaurant
3	Moscow	55.750446	37.617494	5642 высота	Caucasian Restaurant
4	Moscow	55.750446	37.617494	Pinzeria by Bontempi	Pizza Place

Third step is to create dataframe for clustering purpose. We will use columns as food point category for every venue and then we will group it by City.

Final step is to use Kmeans clustering model to define clusters. For our analysis we chose 5 clusters model.

4.1 Results and Discussion

After applying clustering model to collected data, we got next cluster distribution:



We got two big clusters and three small clusters. We can describe this clusters as following:

Cluster 0: In this cluster people prefers Pizzas and sushi.

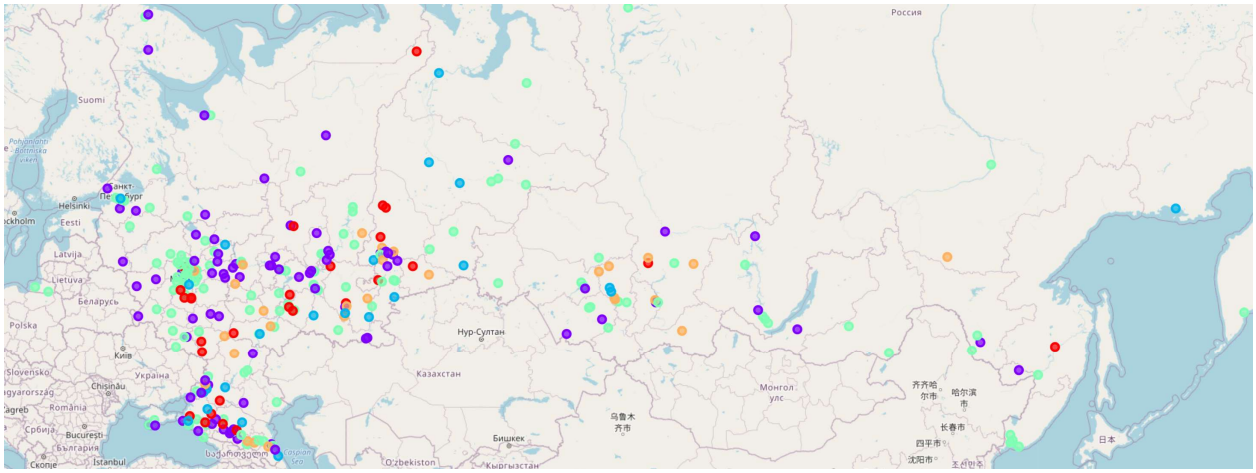
Cluster 1: People most time visit small cafes and eat national cousins.

Cluster 2: People eats variety of foods from European to Asian.

Cluster 3: In this cluster people prefers European foods and love to visit cafes.

Cluster 4: Peoples prefers cafes and Asian kitchen.

We can visualize our clustering results in world map by using folium library to determine what kind food people prefer in different cities or regions.



4.2 Recommendations

Without doubt we can say that café is most popular food point in Russia. European Food is the second choice for opening restaurant. In small towns people prefer pizzas and sushi.

5 Conclusion

The purpose of this project was to identify what kind of food prefer Russian people in different cities or regions. To achieve our goal we used data from Wikipedia and FourSquare API and clustering Kmean model. We created 5 clusters with similar food preferences.

The recommendation is to use our cluster model to decide what kind of restaurant to open in chosen city or region.