IBM Data Science
Professional Certificate Specialization.

Course 9 of 9:
Applied Data Science Capstone.

Final Project.
Full Report.

2019/03

01. Introduction

Moscow is the capital of the Russian Federation, a country of 144.5 million people.

Moscow is not only the largest city in Russia (its total area is 2,511km^2), it is also the most populous city with 13.2 million residents within the city limits, 17 million within the urban area and 20 million within the metropolitan area, according to official data.

Being situated between Europe and Asia, Russia has always been a multinational and multicultural country. So, it's no wonder, that Moscow attracts people from both worlds, and absorbs a lot of their views, traditions and customs.

Moscow. Photo by Deensel (https://www.flickr.com/people/158619309@N03)

A lot of these things can be observed on Moscow streets. Especially, when it comes to food.

There are a lot of stereotypes about the Russians.

Arnold Schwarzenegger as his iconic character Ivan Danko in Red Heat, 1988. One of the funniest movies ever.

Many people think that Russian people are grumpy, rude, unemotional, even blunt, their food is bland and their bears are scary.

And, of course, Russians consume as much vodka as they can.

Well, it's not quite true.

In the 2018 Russia hosted FIFA World Cup.

Official FIFA World Cup 2018 logo

Nearly 7.7 million football (or soccer, if you are from the United States) fans came to Russia to see their favorite teams and, to many people's surprise, enjoy Russian hospitality. And Russian food.

During the World Cup, Moscow, among other Russian cities hosting the tournament, offered its guests the best of entertainment, food and shows.

There were a lot of new venues, opened specially for the event, and most of them still keep their doors open.

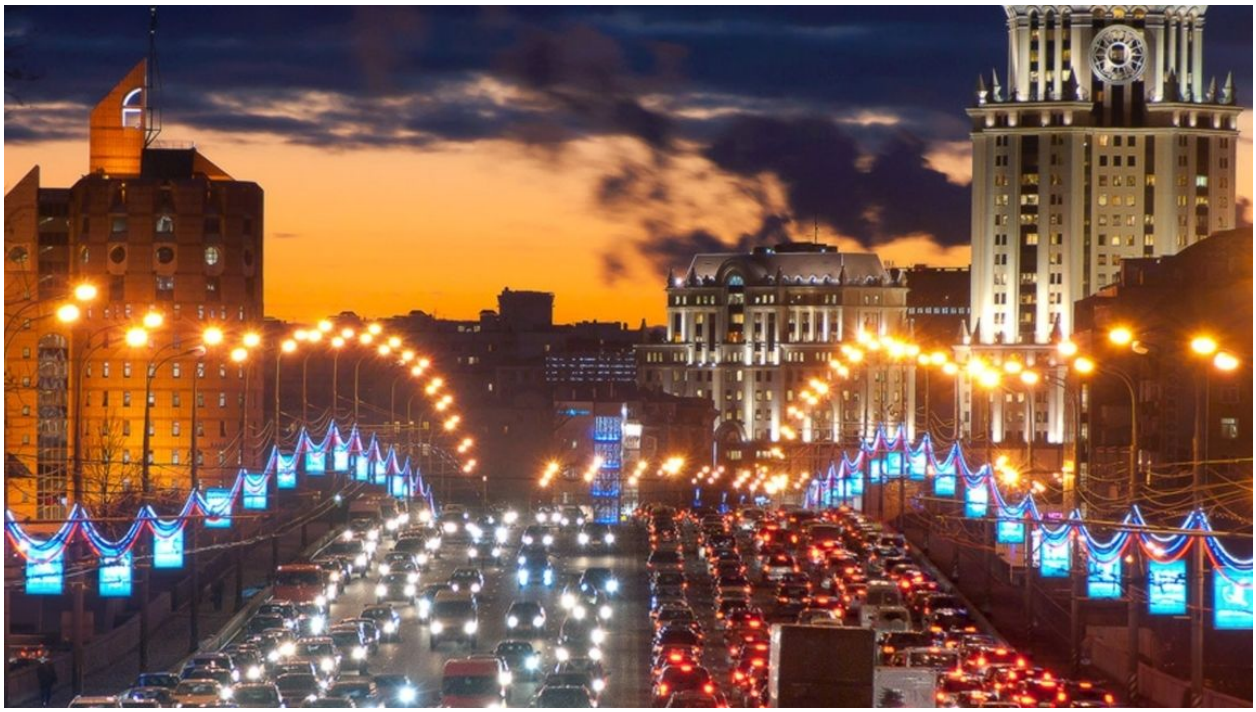
In this project I am going to study Moscow venues from the perspective of food. I am going to inspect three major categories: bars, restaurants and fastfood venues (including coffee shops, diners and cafes; all the places, where you can grab a quick bite).

Russian blinis (crepes) with red caviar. Image source is unknown.

I want to see if Russian people still prefer fancy restaurants, as they used to in the 90's, when so-called 'nouveaux riches' were the target auditory for newly opening places, or if they drink as much as people say, and the most popular category is any kind of bar, pub or any other drinking facility. Or, maybe Moscow is full of fast food and street food places, just like any other megapolis, because people prefer grabbing a cup of coffee or a slice of pizza and run their errands as fast as possible.

I am also going to study the density of venues, because Moscow has several unique features.

First of all, while the majority of citizens live on the outskirts, their everyday routes lay through the downtown, as it's the shortest and the most convenient way from one part of the city to another, and almost all the entertainment events happen in the center of the city. So I want to see if the density of venues is affected by the distance from the center.

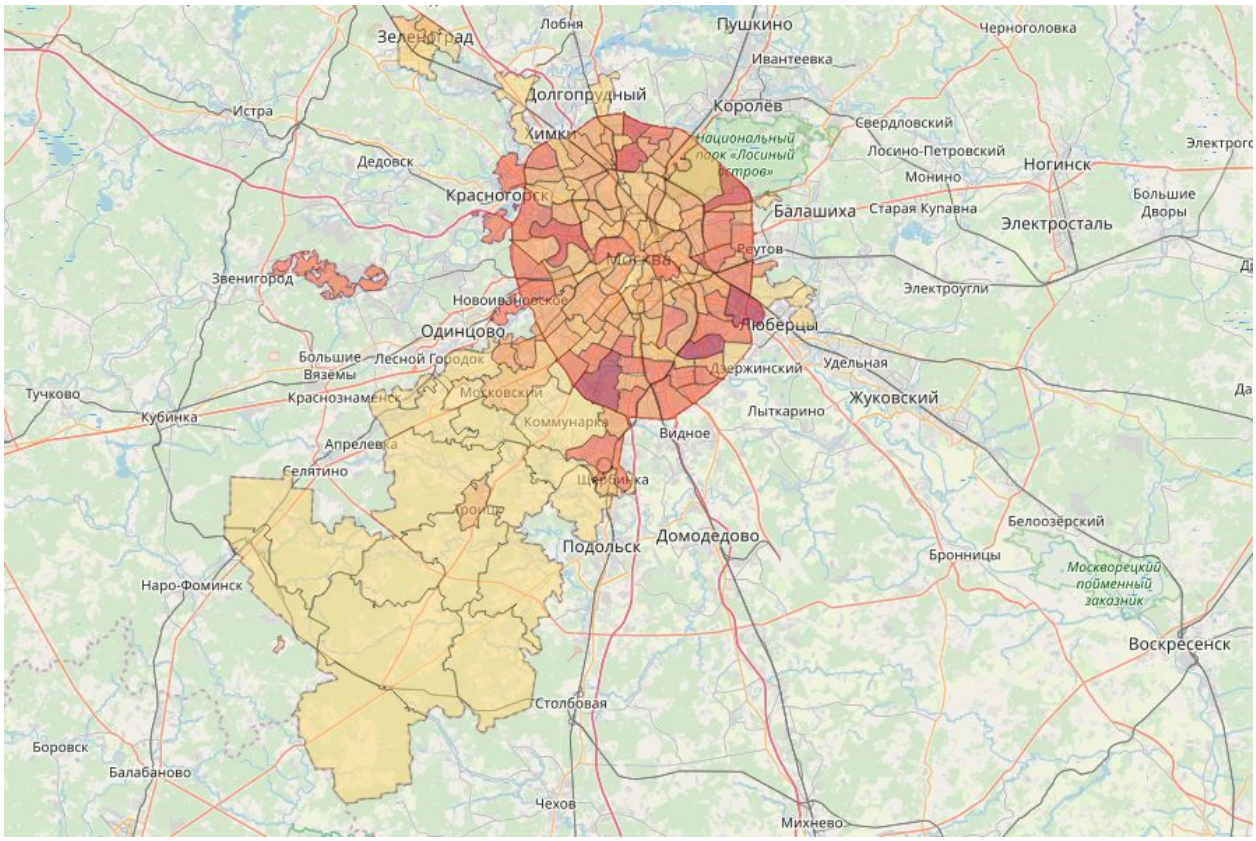


Moscow evening traffic. Photo source: http://gazeta.ru

The other unique Moscow feature is the sparseness. In the year 2012 happened what was called the most ambitious project of expanding the territory in the entire history of Moscow.

The official expansion of the administrative borders of Moscow at the expense of the territory of the Moscow Region was carried out on July 1, 2012, with the result that the area of the city increased by about 2.4 times, due to this, Moscow rose from 11th to 6th place in the ranking of cities of the world by area and became the largest city by area in Europe.

But what was interesting about thin joinment, was the population number. It was only 250,000 of people, which seems to be like a very low number in the view of the size of territory. So, there are districts of Moscow where hundreds of thousands people live, and there are districts, where there are only several hundreds of citizens. It must affect the density of venues, isn't it?



Choropleth map of Moscow population. Map was created for this project with Folium Maps

In terms of application, I believe this research is going to be interesting for two groups of people.

First, tourists who plan to visit Moscow. They will be able to see how diverse this city is, where to look for the most popular venues, where to find their favorite food and where to look for interesting places.

Second, entrepreneurs who are looking for a place to open a new venue. This research is going to highlight overcrowded places, places, where new venues are needed, and what's missing on the plate.

I sincerely hope my research is going to help people to find a spot they've been looking for.

St. Basil's Cathedral. One of the most recognizable images of Moscow. Source is unknown.

Table of Contents

01. Introduction 2

Table of Contents 9

02. Data Description 10

2.1. Moscow borders and districts shapes 10

2.2. Moscow Population 11

2.3. Moscow Venues 11

2.4. Illustrative materials 12

03. Methodology 12

3.1. About This Report 12

3.2. Data Preparation 13

3.3. Data Visualization 19

3.4. Battle of Neighborhoods 38

K-Means Clustering 38

Clusters Examination 42

04. Results 50

05. Discussion 52

06. Conclusion 53

02. Data Description

2.1. Moscow borders and districts shapes

Moscow borders and districts shapes come from the GIS-Lab (http://gis-lab.info) — informal community of russian-speaking GIS/RS specialists.

The data comes in form of GeoJSON file, which is a Feature Collection, describing Moscow districts in the form of sets of coordinates, which can be used to shape polygons and multipolygons on Folium Maps, including the Choropleth Map of Population.

The second source of geospatial data is OpenStreetMap's search engine Nominatim (https://nominatim.openstreetmap.org/), which is a tool for searching OSM data by name and address (geocoding) and to generate synthetic addresses of OSM points.

This data is used mostly for precision.

2.2. Moscow Population

Moscow Population data was manually collected from Wikipedia (http://wikipedia.org). It was put together in a csv file, containing the District Name, Borough Name and Number of People, living in each area.

This data is used to determine density and dependency of number of venues to said density.

2.3. Moscow Venues

All the information about Moscow venues was received from the Foursquare (http://foursquare.com), a service which lets users search for restaurants, nightlife spots, shops and other places of interest.

Foursquare provides data which is the main point of interest of this research, such as venue category and location. Due to specific restrictions of the API, this information may not be considered full and complete, but it's quite enough to observe common trends and special regional features.

2.4. Illustrative materials

Illustrations for the report come from three sources:

1. Photos and images found online (in that case the source is mentioned in the image caption);

2. Screenshots made from the corresponding Jupyter Notebook, such as tables, graphs and dataframe snippets;

3. Some graphs were made with an external tool for sake of readability and more attractive look, though the data they rely on is still prepared in the corresponding Jupyter Notebook. No external tool was used to prepare or analyze the data.

03. Methodology

3.1. About This Report

This Report follows the strict convention

Fonts:

- Normal font — the main text paragraph
- ***Emphasis font*** — important text or remark
- `Console font` — example of code
- `# Commentary` — commented part of code (often for explanation purpose)
- Small font — explanatory illustration caption
- `[dataframe]` — object, such as variable, list or dataframe name
- http://ibm.com — weblink

Color codes for venue categories (used on graphs and maps):

● bar, ● fast food venue, ● restaurant.

3.2. Data Preparation

Environment Preparation

All the work for this project was done in the Jupyter Notebook at http://labs.cognitiveclass.ai.

To get all the necessary tools in place, the following packages were installed into the environment:

```
# install required packages
!conda install -c conda-forge folium=0.5.0 --yes # package to create maps
!conda install -c conda-forge geopy --yes        # client for geocoding web services
# import required libraries
import json                                # JSON encoder and decoder module
import itertools                           # functions creating iterators
import pandas as pd                        # library for data manipulation and analysis
import requests                            # HTTP library
import folium                              # mapping strength of leaflet.js library
import numpy as np                         # multidimensional arrays and math functions
import matplotlib as mpl                   # plotting library
import matplotlib.pyplot as plt            # shortcut for the important function
import seaborn as sns                      # statistical data visualization
import matplotlib.cm as cm                 # matplotlib colormap
import matplotlib.colors as colors         # shortcuts for colors in matplotlib
import collections                         # container datatypes
from scipy import stats                    # statistical functions
from scipy.stats import linregress         # linear regression function
from geopy.geocoders import Nominatim      # tool to search OSM data by geocoding
from sklearn.cluster import KMeans         # K-means Clustering tool
```

Moscow Regions and Population

The first raw data source was a 973,8 KiB GeoJSON file from the the GIS-Lab, containing the coordinates of Moscow regions.

```
{
"type": "FeatureCollection",
"crs": { "type": "name", "properties": { "name": "urn:ogc:def:crs:OGC:1.3:CRS84" } },

"features": [
{ "type": "Feature", "properties": { "NAME": "Kiyevsky", "OKATO": "45298555", "OKTMO": "45945000",
"NAME_AO": "Троицкий", "OKATO_AO": "45298000", "ABBREV_AO": "Троицкий", "TYPE_MO": "Поселение" },
"geometry": { "type": "MultiPolygon", "coordinates": [ [ [ [ 36.8031, 55.44083 ], [ 36.80319, 55.4416 ], [
36.80357, 55.45162 ], [ 36.81253, 55.4514 ], [ 36.82745, 55.45134 ], [ 36.83337, 55.45138 ], [ 36.8338,
55.45164 ], [ 36.83458, 55.45126 ], [ 36.83486, 55.45142 ], [ 36.83499, 55.45149 ], [ 36.8358, 55.45112 ],
```

Cut from mo.geojson file

As you can see, there are sets of coordinates for every district as well as every borough inside those districts. These coordinates form polygons and, in some cases, multipolygons.

Multipolygons appeared hard to process due to the difference of their shapes, so to make it more comfortable to work with, the data was transformed into a flat set of coordinates, and the centers of each borough were calculated:

```
Polygon example:
[[37.84042, 55.73049], [37.84063, 55.73196], [37.84165, 55.74049], [37.84181, 55.74303], [37.84201,
55.74722], [37.84254, 55.74718], [37.84265, 55.74673], [37.84315, 55.74633], [37.84368, 55.7459],
[37.8448, 55.74531], [37.84571, 55.74487], [37.8468, 55.74459], [37.84792, 55.74454], [37.848, 55.7
445], [37.84971, 55.74462], [37.85142, 55.74476], [37.85297, 55.74488], [37.85392, 55.74495], [37.8
5662, 55.74507], [37.85919, 55.74515], [37.86054, 55.74524], [37.86323, 55.74557], [37.86422, 55.74
572], [37.86474, 55.74585], [37.87285, 55.74741], [37.87952, 55.74873], [37.88254, 55.74934], [37.8
8735, 55.74846], [37.88771, 55.74858], [37.88825, 55.74846], [37.88871, 55.74714], [37.88933, 55.74
547], [37.89003, 55.74363], [37.89095, 55.74219], [37.8904, 55.74208], [37.89008, 55.74187], [37.89
009, 55.7416], [37.88583, 55.74261], [37.8836, 55.74269], [37.88175, 55.74273], [37.88075, 55.742
5], [37.88011, 55.74225], [37.88003, 55.74175], [37.87801, 55.73787], [37.87316, 55.73816], [37.868
49, 55.73729], [37.86654, 55.7363], [37.86671, 55.73618], [37.86424, 55.73472], [37.86547, 55.7331
6], [37.86889, 55.73297], [37.86954, 55.73239], [37.86219, 55.73256], [37.8613, 55.73052], [37.854
9, 55.73081], [37.85461, 55.73082], [37.85406, 55.73086], [37.84989, 55.7311], [37.84767, 55.7311
8], [37.84695, 55.73118], [37.84558, 55.7311], [37.84425, 55.73107], [37.84342, 55.73103], [37.8419
5, 55.73088], [37.84119, 55.73073], [37.84042, 55.73049]]
```

Transformed set of coordinates of Novokosino

To improve the accuracy of the geospatial data, precise coordinates were acquired from the OpenStreetMap's search engine Nominatim.

To make sure Nominatim does not confuse coordinates for certain places (there are a lot of places of the same name in different regions), every borough was assigned to its district by joining them in a single string:

	District	Borough
0	Central Administrative Okrug	Arbat
1	Central Administrative Okrug	Basmanny

→

	FullPlaceName
0	Central Administrative Okrug, Arbat
1	Central Administrative Okrug, Basmanny

Joined district and borough names in a single cell

Afterwards, the correct coordinates were acquired.

OpenStreetMap was unable to collect the data for all regions, so coordinates from previously built set were assigned. Inaccurate data was removed.

	District	Borough	Population	AvgLatitude	AvgLongitude	RealLatitude	RealLongitude
0	Central Administrative Okrug	Arbat	25699	55.751192	37.595551	55.751199	37.589872
1	Central Administrative Okrug	Basmanny	100899	55.765157	37.666924	55.767281	37.669773
2	Central Administrative Okrug	Khamovniki	97110	55.740181	37.588317	55.740047	37.573958
3	Central Administrative Okrug	Krasnoselsky	45229	55.773022	37.647721	55.777447	37.654160
4	Central Administrative Okrug	Meshchansky	56077	55.776221	37.627220	55.779169	37.627755

Corrected coordinates of boroughs

Moscow population data was manually collected from open data sources, including Wikipedia.

Next step was to combine all the acquired data together in a dataframe `[moscow_data]` and get rid of rows containing insufficient data:

```
# if no data exists, replace it with data from the geojson
moscow_data.Latitude.fillna(moscow_data.AvgLatitude, inplace=True)
moscow_data.Longitude.fillna(moscow_data.AvgLongitude, inplace=True)
```

	District	Borough	FullPlaceName	Latitude	Longitude	Population
0	Central Administrative Okrug	Arbat	Central Administrative Okrug, Arbat	55.751199	37.589872	25699
1	Central Administrative Okrug	Basmanny	Central Administrative Okrug, Basmanny	55.767281	37.669773	100899
2	Central Administrative Okrug	Khamovniki	Central Administrative Okrug, Khamovniki	55.740047	37.573958	97110
3	Central Administrative Okrug	Krasnoselsky	Central Administrative Okrug, Krasnoselsky	55.777447	37.654160	45229
4	Central Administrative Okrug	Meshchansky	Central Administrative Okrug, Meshchansky	55.779169	37.627755	56077

First rows of `[moscow_data]` dataframe

Moscow Venues from Foursquare

Data, collected from the Foursquare, included the name of the venue, its category and coordinates. A dataframe called `[moscow_venues]` was put together to keep the data together:

	FullPlaceName	Latitude	Longitude	Venue	VenueLatitude	VenueLongitude	Category
0	Central Administrative Okrug, Arbat	55.751199	37.589872	Corner Café & Kitchen	55.751496	37.586757	Japanese Restaurant
1	Central Administrative Okrug, Arbat	55.751199	37.589872	Театр им. Вахтангова	55.749569	37.591638	Theater
2	Central Administrative Okrug, Arbat	55.751199	37.589872	Obedbufet (Обедбуфет)	55.752268	37.592275	Buffet
3	Central Administrative Okrug, Arbat	55.751199	37.589872	Buffalo's	55.751840	37.587376	Wings Joint
4	Central Administrative Okrug, Arbat	55.751199	37.589872	Кофемания	55.752094	37.588102	Coffee Shop

First rows of `[moscow_venues]` dataframe

There were more than 2580 venues total. As the goal of the study was to learn about food venues, the list of venues was restricted to these categories.

Unfortunately, Foursquare does not provide information about general category of a venue, so it has to be done manually.

All the venues were divided by three general categories: bar, restaurant and fast food.

Here's the list of all subcategories, defined by the Foursquare API, that were included in each general category:

Restaurants:

'American Restaurant', 'Argentinian Restaurant', 'Asian Restaurant', 'Belgian Restaurant', 'Caucasian Restaurant', 'Chinese Restaurant', 'Comfort Food Restaurant', 'Czech Restaurant', 'Eastern European Restaurant', 'English Restaurant', 'French Restaurant', 'Gastropub', 'German Restaurant', 'Gourmet Shop', 'Greek Restaurant', 'Halal Restaurant', 'Hawaiian Restaurant', 'Indian Restaurant', 'Israeli Restaurant', 'Italian Restaurant', 'Japanese Restaurant', 'Jewish Restaurant', 'Korean Restaurant', 'Mediterranean Restaurant', 'Mexican Restaurant', 'Middle Eastern Restaurant', 'Modern European Restaurant', 'Moroccan Restaurant', 'Peruvian Restaurant', 'Restaurant', 'Russian Restaurant', 'Scandinavian Restaurant', 'Seafood Restaurant', 'Spanish Restaurant', 'Sushi Restaurant', 'Thai Restaurant', 'Theme Restaurant', 'Turkish Restaurant', 'Ukrainian Restaurant', 'Vegetarian / Vegan Restaurant', 'Vietnamese Restaurant', 'Steakhouse'.

Fast food Venues:

'BBQ Joint', 'Bistro', 'Blini House', 'Creperie', 'Donut Shop', 'Burger Joint', 'Breakfast Spot', 'Dumpling Restaurant', 'Falafel Restaurant', 'Fast Food Restaurant', 'Food Court', 'Fried Chicken Joint', 'Hot Dog Joint', 'Kebab Restaurant', 'Noodle House', 'Pelmeni House', 'Pizza Place', 'Shawarma Place', 'Udon Restaurant', 'Wings Joint', 'Fish & Chips Shop', 'Pastry Shop', 'Pie Shop', 'Salad Place', 'Sandwich Place', 'Snack Place', 'Soup Place', 'Buffet', 'Diner', 'Bakery', 'Bagel Shop', 'Café', 'Cafeteria', 'Coffee Shop', 'Cupcake Shop', 'Dessert Shop', 'Frozen Yogurt Shop', 'Ice Cream Shop'.

Bars:
`'Bar', 'Beer Bar', 'Beer Garden', 'Brewery', 'Cocktail Bar', 'Dive Bar', 'Hotel Bar', 'Irish Pub', 'Karaoke Bar', 'Lounge', 'Pub', 'Sports Bar', 'Whisky Bar', 'Wine Bar', 'Wine Shop'.`

As you can see, venues such as 'Creperie', 'Udon restaurant' or 'Diner' were also included in the Fastfood general category, as it was stated in the Introduction section, for the purpose of simplicity and data readability.

There were total of 322 restaurants, 431 fast food venues and 98 bars found in Moscow.

Next step was to set up a dataset, combining all the data acquired.

First part of the dataset contain the data about a place where the venue is situated:

`['FullPlaceName'] ['District'] ['Borough'] ['Latitude'] ['Longitude'] ['Population']`

Second part of the dataset contain the data about the venue:

`['Venue'] ['VenueLatitude'] ['VenueLongitude'] ['CommonCategory'], ['Category']`

As the dataset contains all the information about each venue, it is very comfortable to work with:

	FullPlaceName	District	Borough	Latitude	Longitude	Population	Venue	VenueLatitude	VenueLongitude	CommonCategory	Category
0	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Corner Café & Kitchen	55.751496	37.586757	restaurant	Japanese Restaurant
1	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Граци Рагацци	55.752137	37.591267	restaurant	Italian Restaurant
2	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Taksim	55.752240	37.591971	restaurant	Turkish Restaurant
3	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Барашка	55.752055	37.586117	restaurant	Middle Eastern Restaurant
4	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Чайхона № 1	55.752383	37.585924	restaurant	Middle Eastern Restaurant
5	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Хинкальная	55.748943	37.588480	restaurant	Caucasian Restaurant
6	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Hard Rock Cafe	55.747995	37.586754	restaurant	American Restaurant
7	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Лодка	55.751183	37.583491	restaurant	Asian Restaurant
8	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Босфор / Bosfor	55.747967	37.587879	restaurant	Turkish Restaurant
9	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Ресторан Дома Актера	55.749118	37.592156	restaurant	Restaurant

First rows of [food_data] dataframe

3.3. Data Visualization

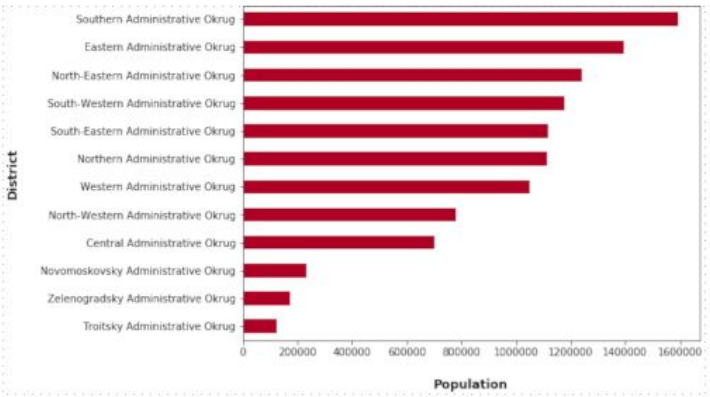
To get fully acquainted with the collected data, it was plotted in several different ways, such as bar plots and Folium maps.

Population

Population by district:

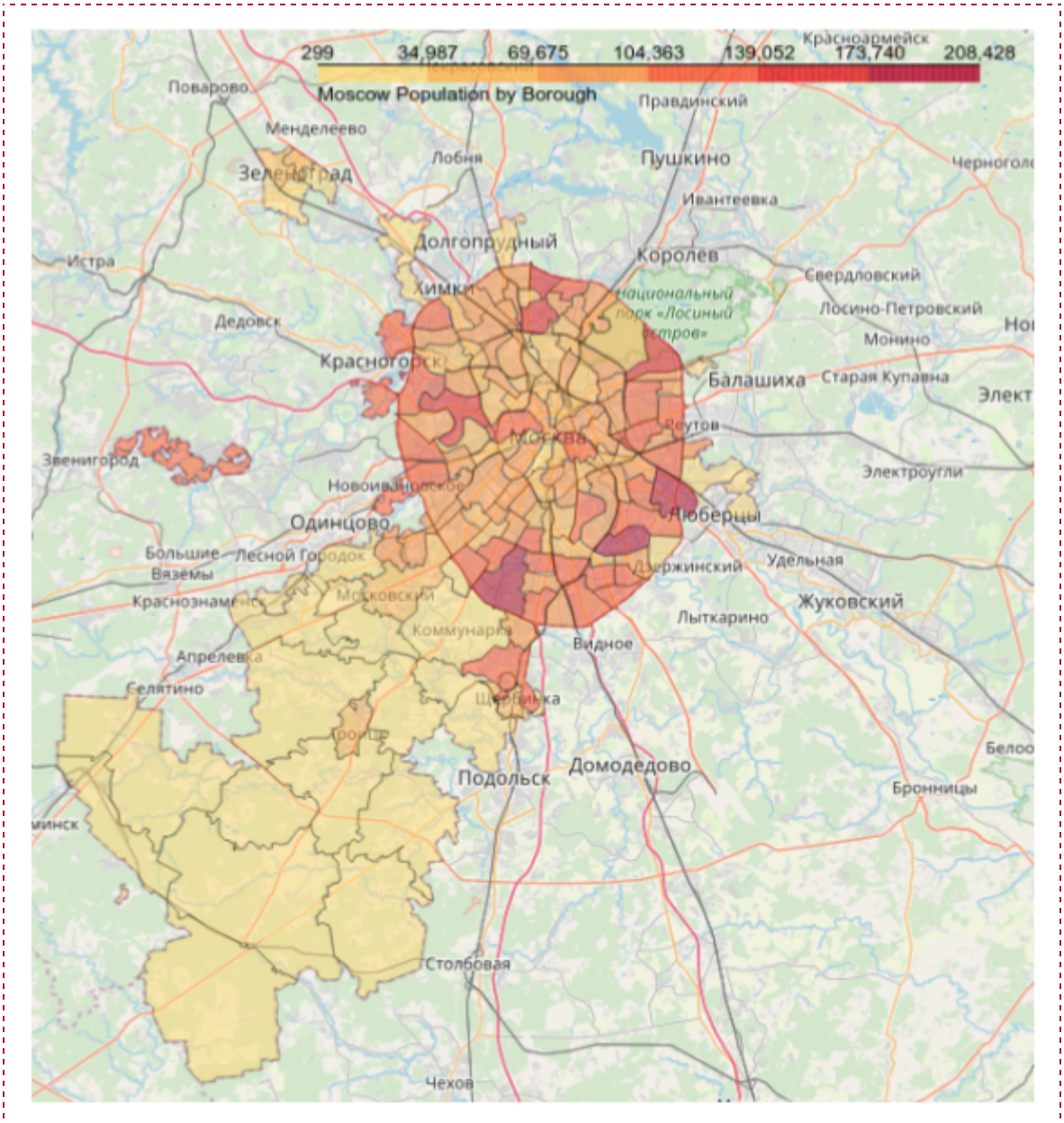
```
# count sum of people living in each borough of a district
population_dist =
      moscow_data.groupby(['District'])['Population'].sum(level="Population")
print(population_dist.sort_values(ascending=False))

# plot the data
ax = population_dist.sort_values(ascending=True).plot(
      kind='barh',color='#B10026',figsize=(8,6)
)
ax.set_xlabel("Population", labelpad=20, weight='bold', size=12)
ax.set_ylabel("District", labelpad=20, weight='bold', size=12)
```

```
District
Southern Administrative Okrug          1593095
Eastern Administrative Okrug           1394497
North-Eastern Administrative Okrug     1240062
South-Western Administrative Okrug     1179211
South-Eastern Administrative Okrug     1116924
Northern Administrative Okrug          1112847
Western Administrative Okrug           1049104
North-Western Administrative Okrug      779965
Central Administrative Okrug            701353
Novomoskovsky Administrative Okrug      232982
Zelenogradsky Administrative Okrug      174805
Troitsky Administrative Okrug           123941
```



Population by District

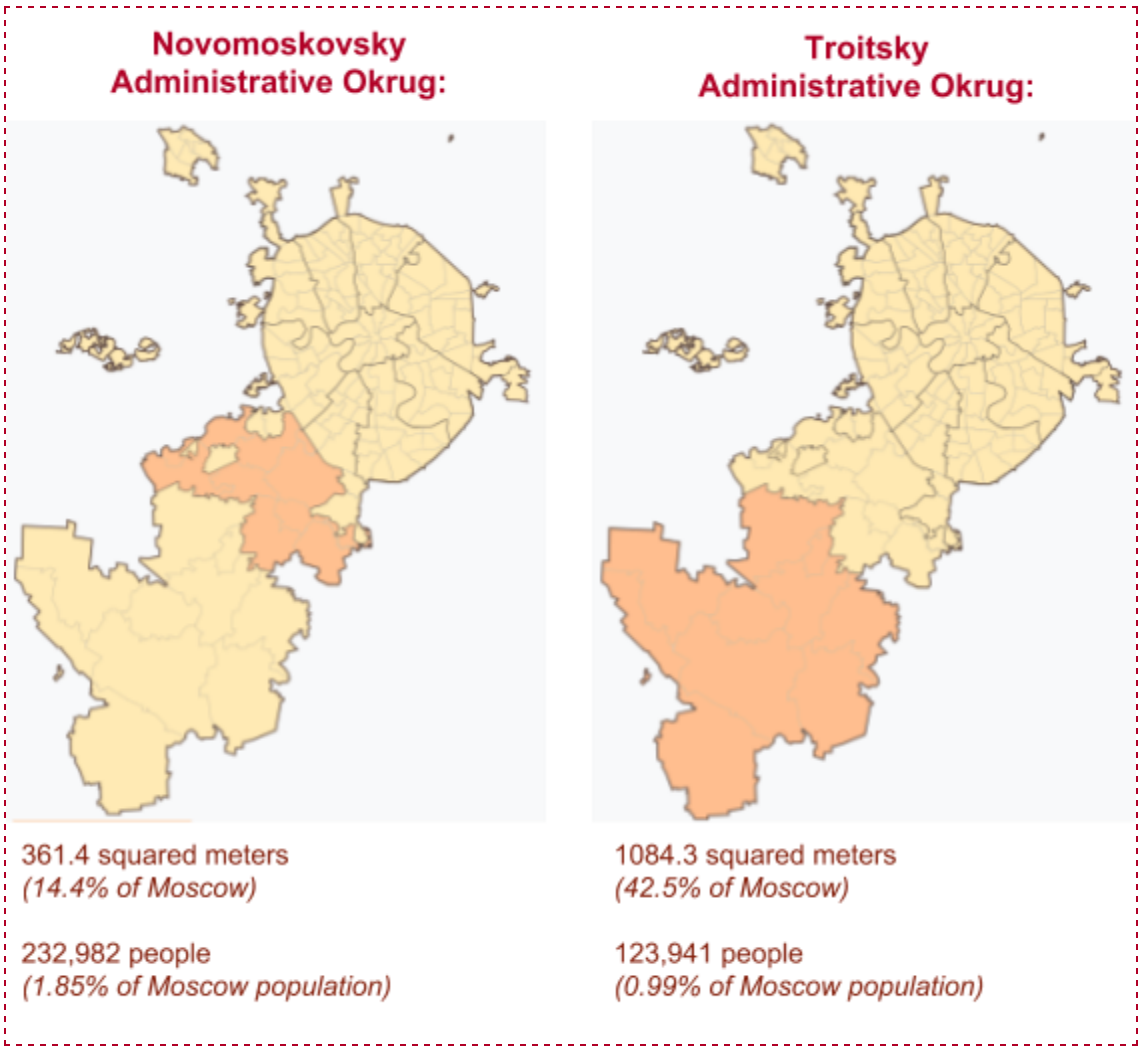
To see this data in details, a choropleth map was created using the Folium package:



Choropleth map created with Folium Maps: Moscow Population

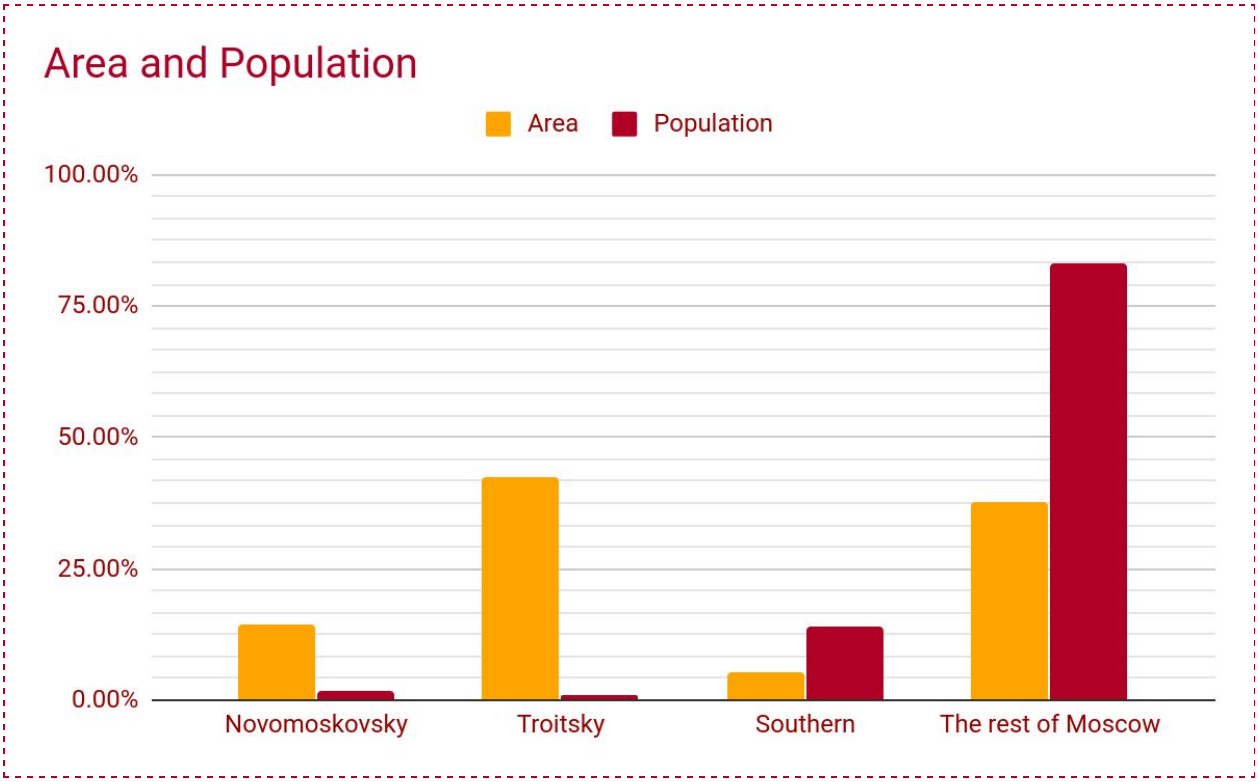
As your can see, Moscow population is extremely uneven. The whole part of Moscow, which was joined in 2012 (bright-yellow color on the map) is way less populated than the old part of Moscow.

In Novomoskovsky Administrative Okrug (361.4 m^2, 14.4% of Moscow area) live only 232,982 people (1.85% of Moscow population). In Troitsky Administrative Okrug (1,084.3 m^2, 42.5% of Moscow area) live only 123,941 people (0.99% of Moscow population).



Comparison between Novomoskovsky and Troitsky districts

For comparison: in Southern Administrative Okrug (131,773 m^2, 5.3% of Moscow area) live 1,593,095 people (14.14% of Moscow population):



Comparison between Novomoskovsky, Troitsky and Southern Districts and the rest of Moscow: area vs population

Such unevenness makes data analysis complicated and increases the possibility of errors and mistakes.

There are two approaches to solving this problem:

1. Discard newly joined regions with the lowest population numbers as insufficient;
2. Keep the data and proceed with the analysis.

I have decided to take the second approach, as it might bring interesting results.

Venue Categories

To see what each general (or common) category contains, the entries were grouped and counted:

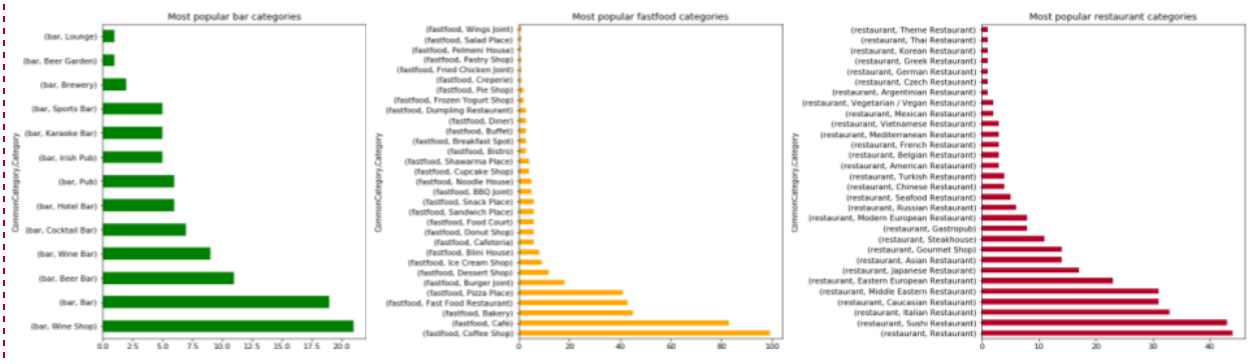
```
# see how popular are categories among common categories
food_data.groupby(['CommonCategory'])['Category'].value_counts()
```

Top 5 of the most popular venue categories are:

Bar		Fast food		Restaurant	
Wine Shop	21	Coffee Shop	99	Restaurant	44
Bar	19	Cafe	83	Sushi Restaurant	43
Beer Bar	11	Bakery	45	Italian Restaurant	33
Wine Bar	9	Fast Food Restaurant	43	Caucasian Restaurant	31
Cocktail Bar	7	Pizza Place	41	Middle East Restaurant	31

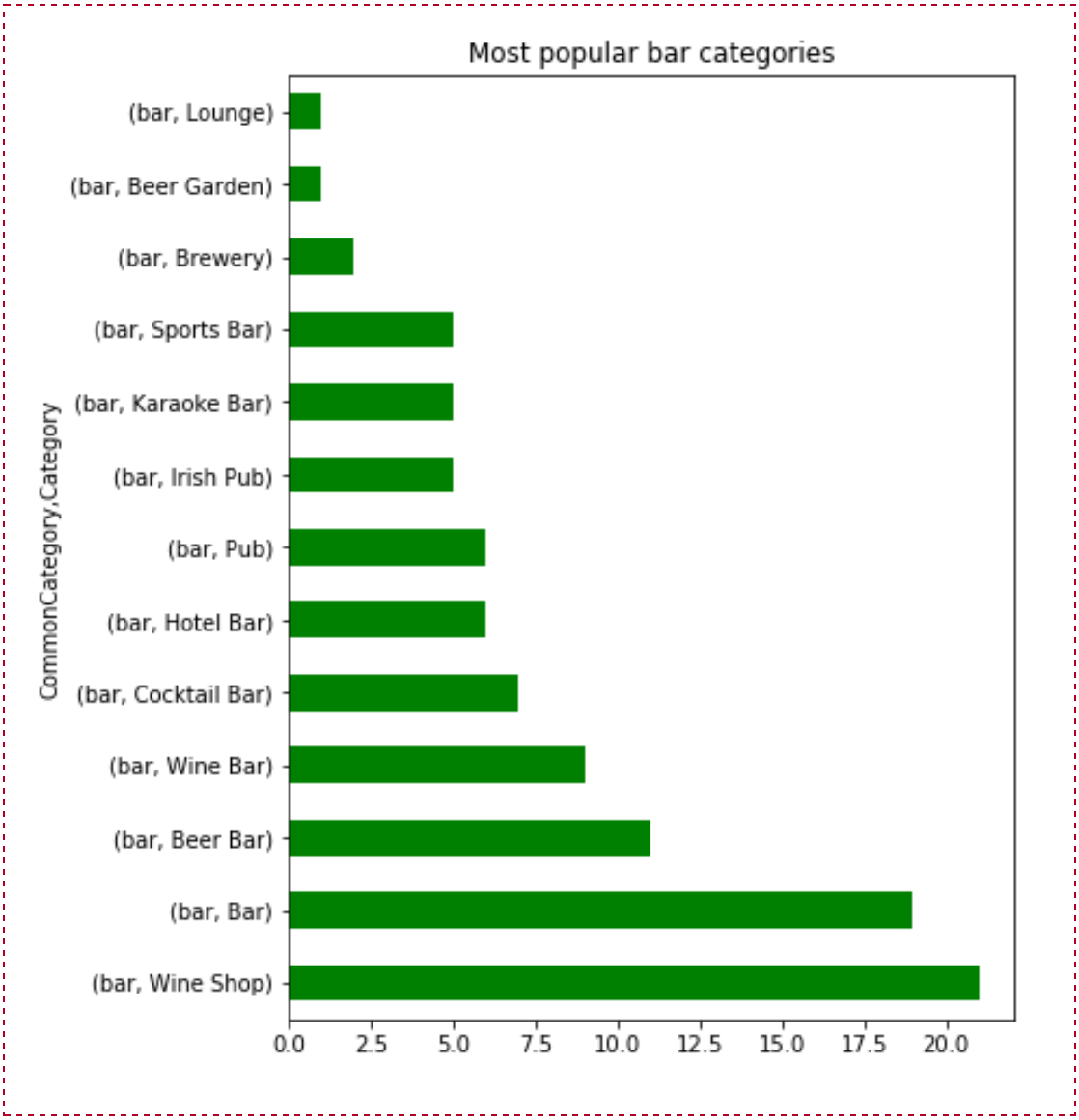
Most popular venue categories in Moscow

This is how it looks on the plot:



Most popular subcategories in general categories

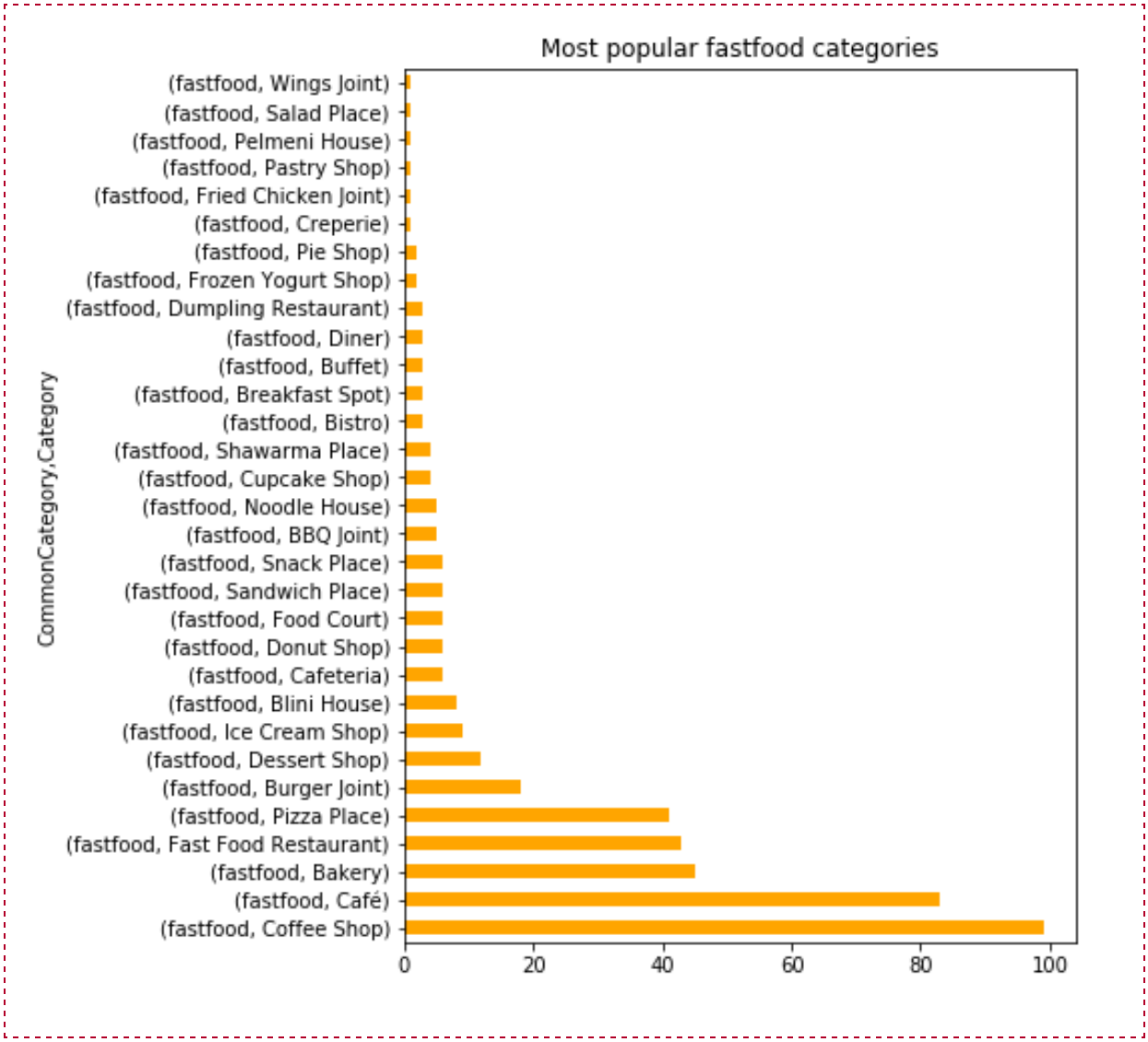
Most popular bar categories:



Most popular categories in general Bar category

So, the most popular bar venues are wine shops, general bars, and beer bars, and the least popular are breweries, beer gardens and lounges.

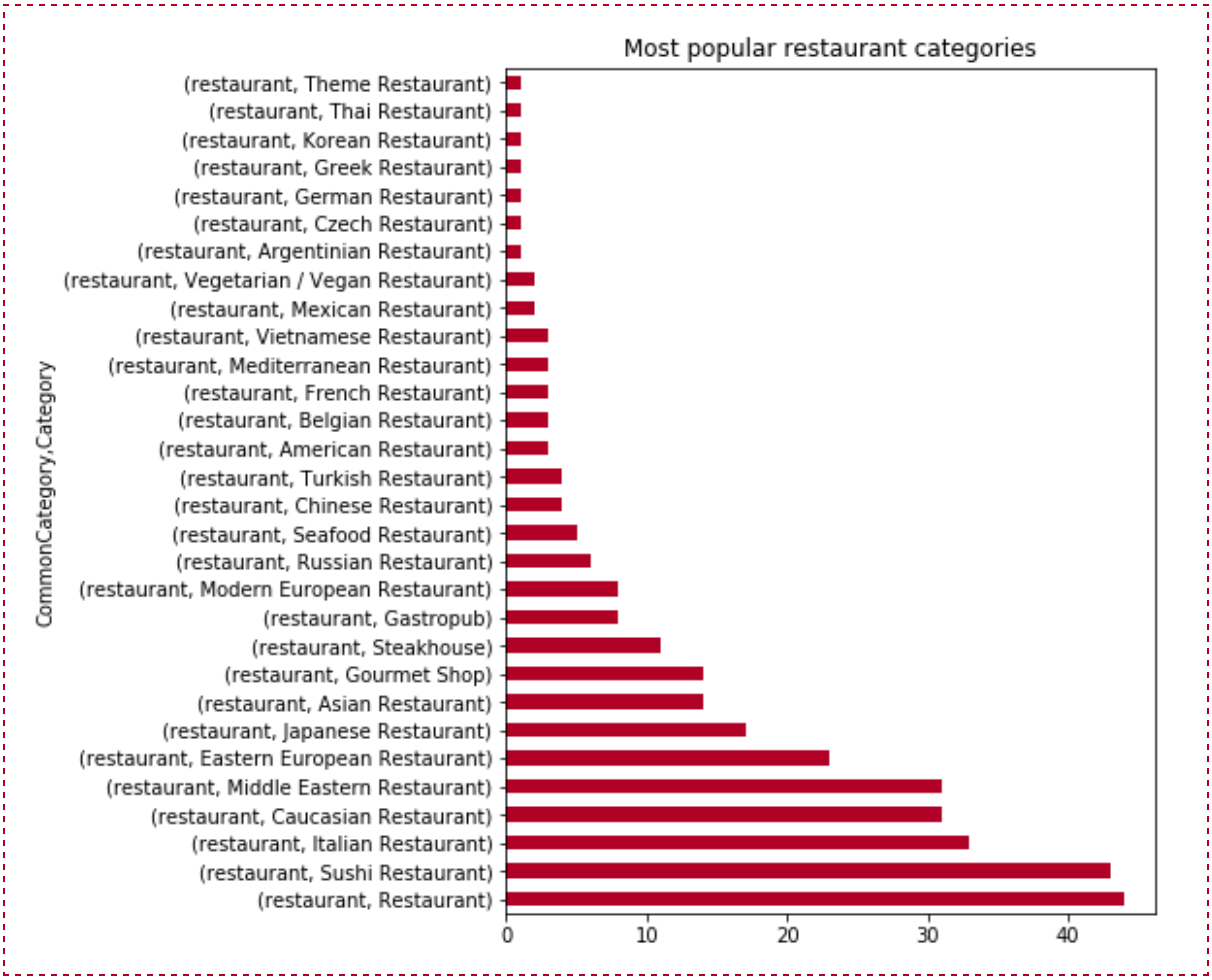
Most popular fastfood categories:



Most popular categories in general Fastfood category

So, the most popular venues are coffee shops, cafes and bakeries, and the least popular venues are pelmeni house, salad places and wing joints.

Most popular restaurant categories:



Most popular categories in general Restaurant category

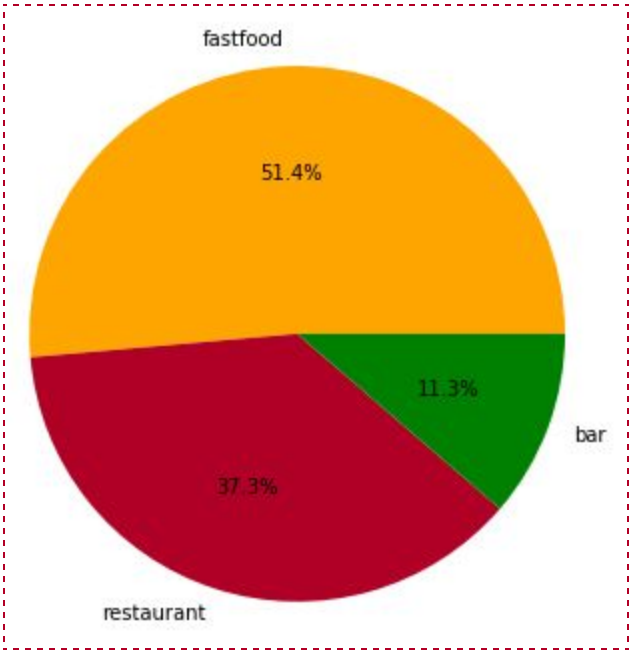
So, the most popular restaurant venues are general restaurants, sushi restaurants and Italian restaurants, and the least popular venues are Korean restaurants, Thai restaurants and Theme restaurants.

Common categories ratio

```
# count how many of each category we have
food_data['CommonCategory'].value_counts()
```

Common Category	Number of Venues
Bar	97
Restaurant	321
Fastfood	440

The data on a pie chart:



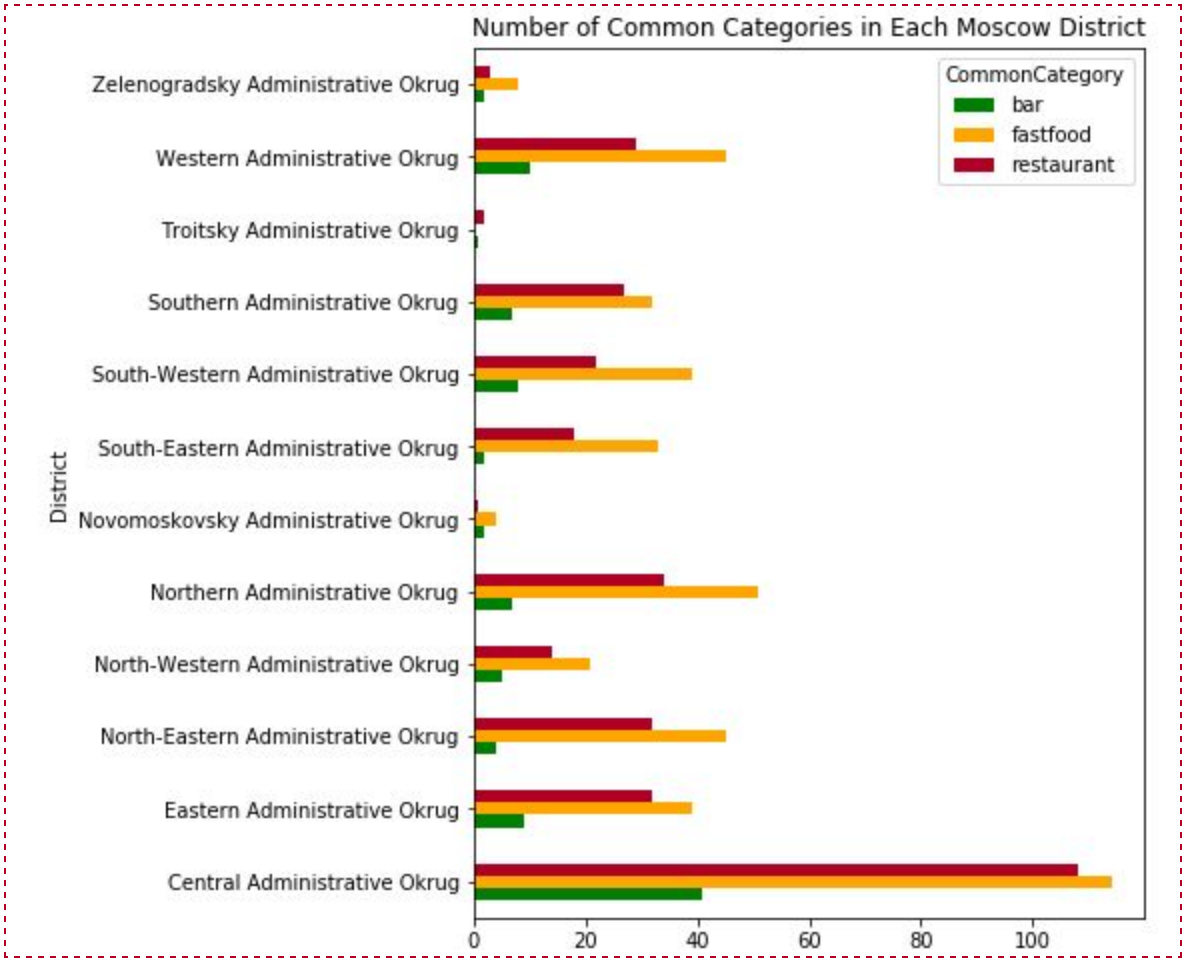
Bar / Fastfood / Restaurant ratio

Common Categories Ratio Among Districts

Now, let's see how popular categories spread across districts.

District	Fastfood	Restaurant	Bar
Central Administrative Okrug	114	109	41
Eastern Administrative Okrug	39	32	9
North-Eastern Administrative Okrug	45	32	4
North-Western Administrative	21	14	5
Northern Administrative Okrug	51	34	7
Novomoskovsky Administrative Okrug	4	2	1
South-Eastern Administrative Okrug	33	18	2
South-Western Administrative Okrug	39	22	8
Southern Administrative Okrug	32	27	7
Troitsky Administrative Okrug	7	2	1
Western Administrative Okrug	45	29	10
Zelenogradsky Administrative Okrug	8	3	2

Number of venues of category in each district



Common categories ratio in each district

Bars ratio in each district:

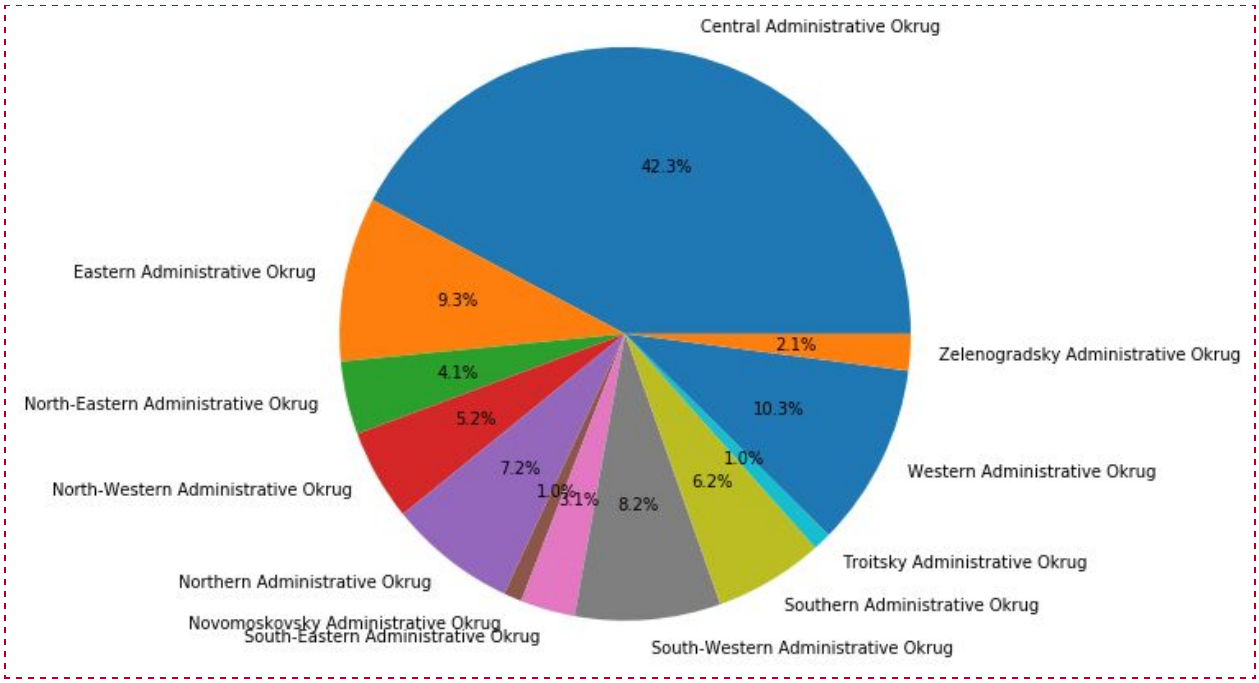


Illustration of Bar popularity in each district

Restaurants ratio in each district:

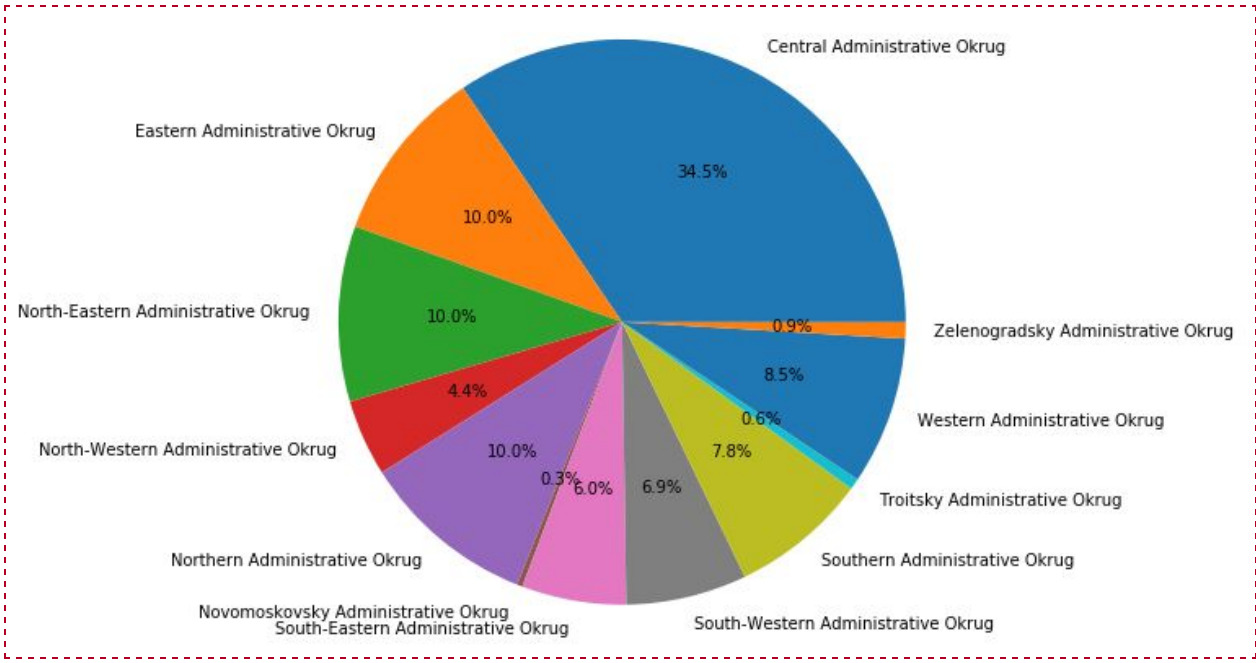


Illustration of Restaurant popularity in each district

Fast food Bars ratio in each district:

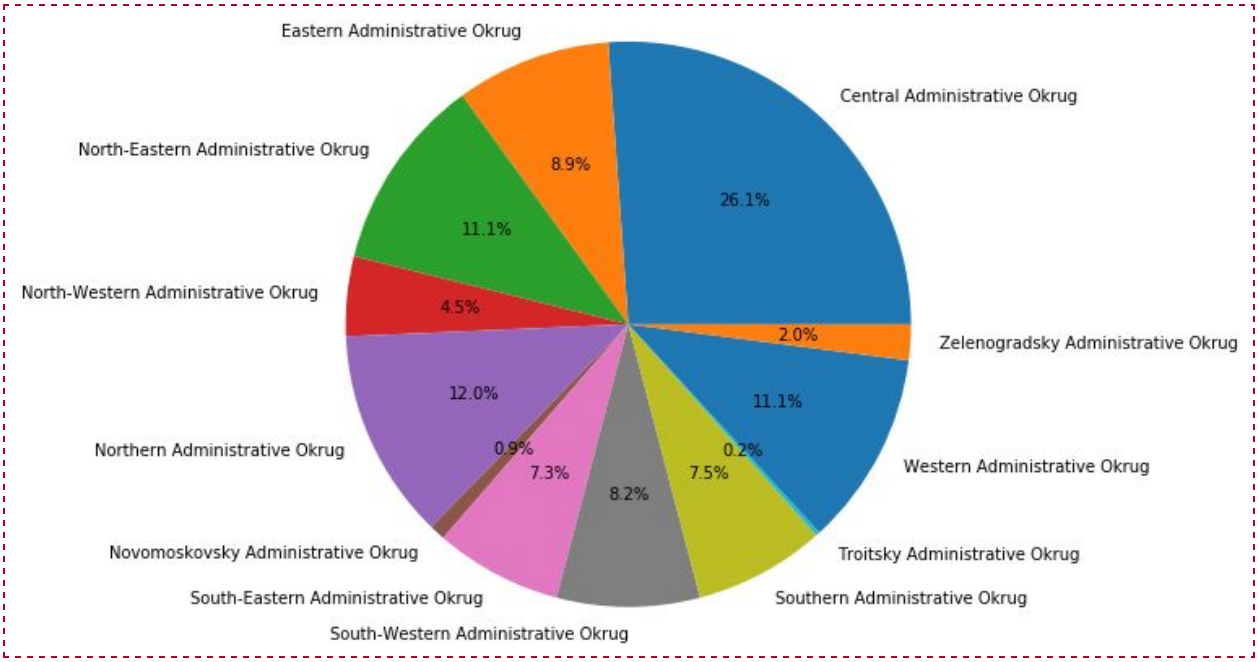


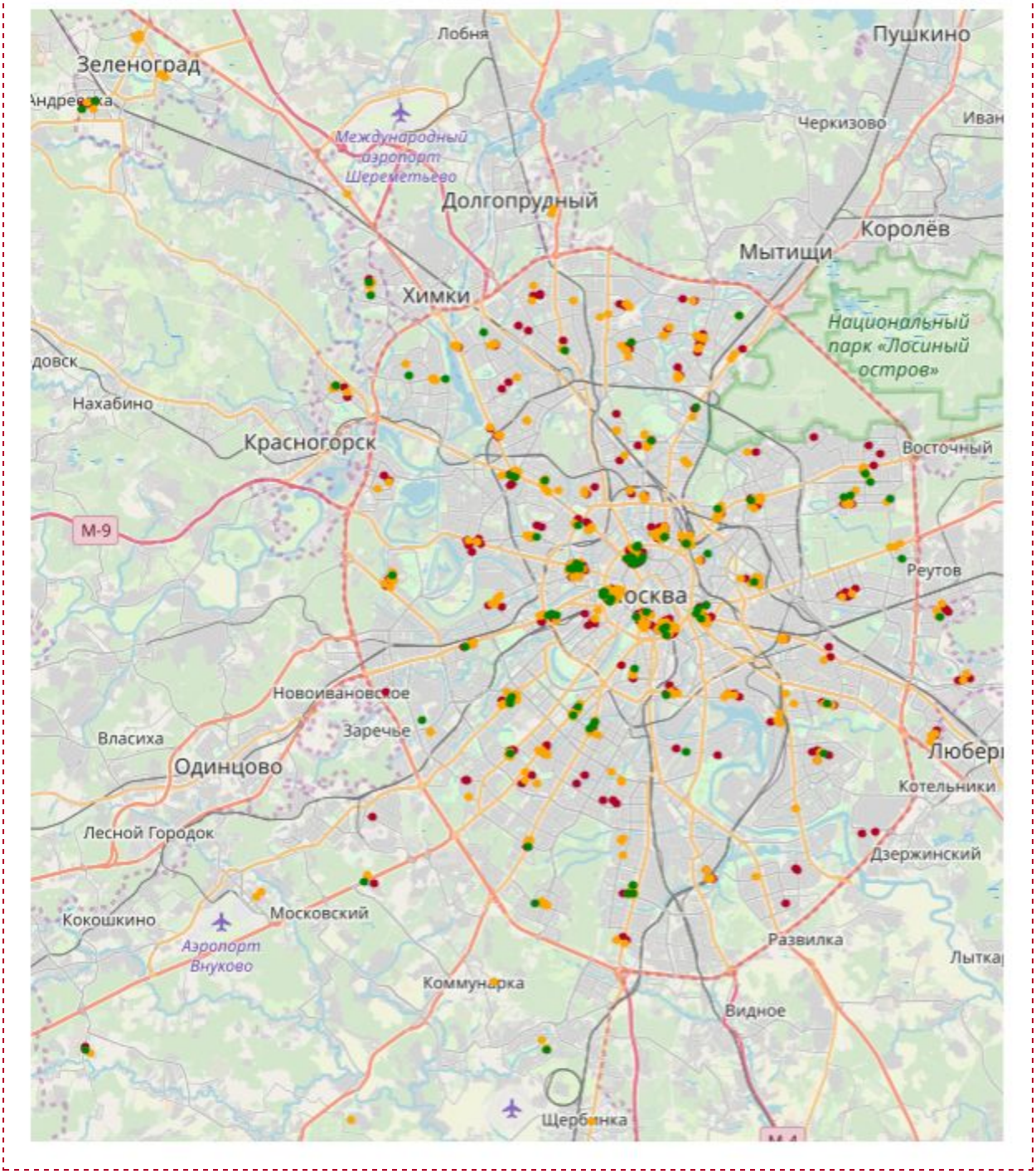
Illustration of Fastfood popularity in each district

So, it's quite obvious, that most of the venues are situated in the Central Administrative Okrug.

It confirms the statement, that all activities in Moscow are concentrated around the center, while the rest of the city is more populated, but is mostly purposed as dormitory areas, or industrial zones.

Such districts as Northern, Western or Eastern Administrative Districts, do have their own part of venues, though it's much smaller than the Central district, and later on in this research the correlation between population and number of venues is going to be studied.

See venues on the map:

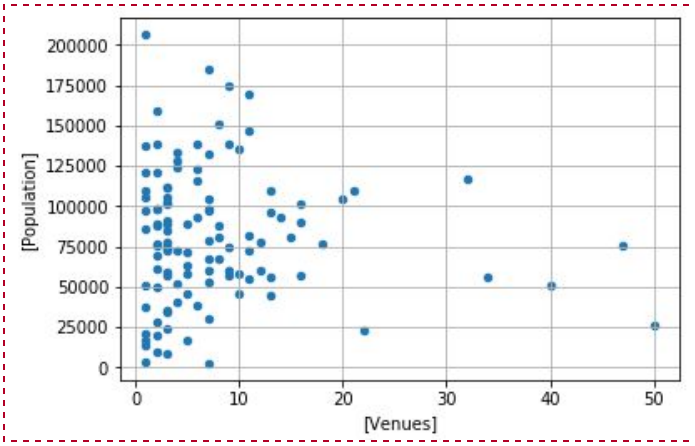


*Folium map of venue categories in each district, where each category **bar**, **fastfood**, **restaurant** is color coded.*

Number of Venues and Population Correlation

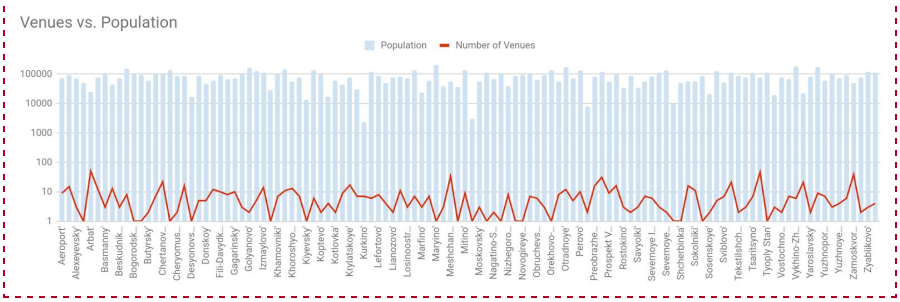
To find out if number of venues depends on population, I use corrcoef function from NumPy and plot the results:

```
keys = set(num_ven.keys() | num_pop.keys()) # number of venues and population
np.corrcoef(
    [num_ven.get(x, 0) for x in keys],
    [num_pop.get(x, 0) for x in keys])[0,1]

Result: -0.04201284243983189
```



Scatter plot showing how number of venues depend on population

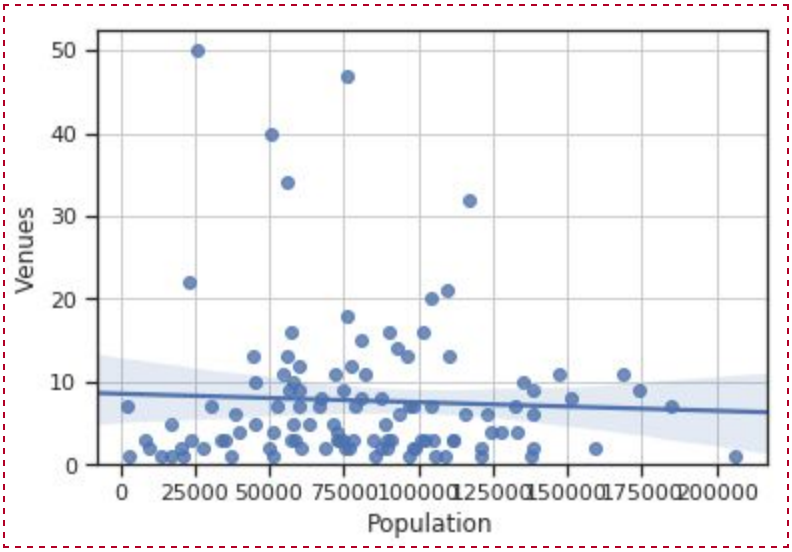
Weirdly enough, there is almost no correlation between population and number of venues. To make sure, I created a combo chart on a raw data:



Combo chart

Finally, I created regplot from Seaborn to confirm the results:

```
sns.regplot(x=pop_ven['Population'], y=pop_ven['Venues'])
plt.ylim(0,)

Result: (0, 53.512462119011644)
```



Regplot to show how number of venues depend on population

The most obvious reason for such unexpected result is that life in Moscow is quite unbalanced. While most people live on the outskirt of Old Moscow (Moscow borders before the join in 2012), most venues are closer to the center, as the major part of all business. This situation has already been observed, so let's prove this statement with numbers.

Number of Venues and Distance from Center Correlation

For that purpose I use distance function from GeoPy, which measures distance between coordinates and returns the value:

```
def distancometer(x, y):
    v = (x,y)
    msk = (55.751244, 37.618423)
    return geopy.distance.distance(msk, v)

food_data['Distance'] = food_data.apply(
    lambda x: distancometer(x['Latitude'],
    x['Longitude']),axis=1
)
```

	FullPlaceName	District	Borough	Latitude	Longitude	Population	Venue	VenueLatitude	VenueLongitude	CommonCategory	Category	Distance
0	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Corner Café & Kitchen	55.751496	37.586757	restaurant	Japanese Restaurant	1.792840486909856 km
1	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Граци Рагацци	55.752137	37.591267	restaurant	Italian Restaurant	1.792840486909856 km
2	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Taksim	55.752240	37.591971	restaurant	Turkish Restaurant	1.792840486909856 km
3	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Барашка	55.752055	37.586117	restaurant	Middle Eastern Restaurant	1.792840486909856 km
4	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Чайхона № 1	55.752383	37.585924	restaurant	Middle Eastern Restaurant	1.792840486909856 km

food_data dataset with an extra column [Distance]

The only issue with this is that geopy.distance returns the distance as a geopy object, not number, so I had to transform the string into a float to make all further work possible:

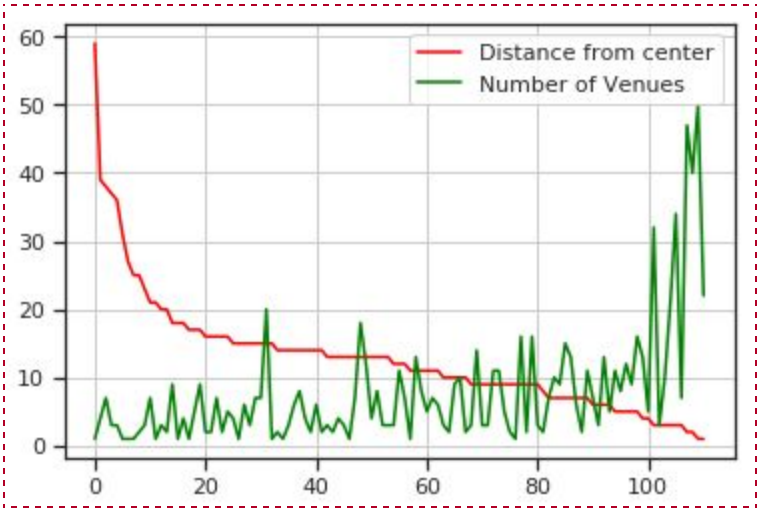
```
food_data['Distance'] = food_data.Distance.astype(str).str[:-3].astype(float)
```

So, into the food_data dataset was added another column called [Distance]:

	FullPlaceName	District	Borough	Latitude	Longitude	Population	Venue	VenueLatitude	VenueLongitude	CommonCategory	Category	Distance
0	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Corner Café & Kitchen	55.751496	37.586757	restaurant	Japanese Restaurant	1.79284
1	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Граци Рагацци	55.752137	37.591267	restaurant	Italian Restaurant	1.79284
2	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Taksim	55.752240	37.591971	restaurant	Turkish Restaurant	1.79284
3	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Барашка	55.752055	37.586117	restaurant	Middle Eastern Restaurant	1.79284
4	Central Administrative Okrug, Arbat	Central Administrative Okrug	Arbat	55.751199	37.589872	25699	Чайхона № 1	55.752383	37.585924	restaurant	Middle Eastern Restaurant	1.79284

food_data dataset with an extra column [Distance] converted to a float type

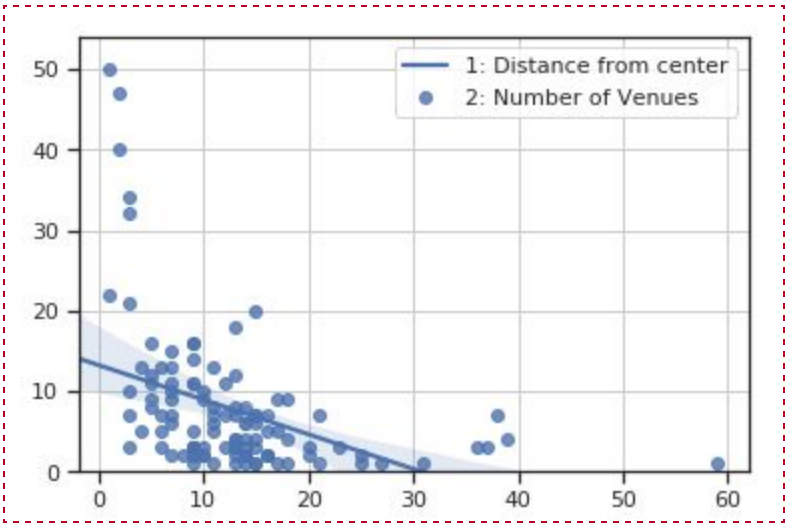
Now, let's look at the data plotted:



Plot to show if number of venues depend on the distance

So, it appears like the number of venues strongly depend on the distance: the closer to the center, the more venues we can see, and vice versa.

Let's build the regplot to confirm:



Regplot to show if number of venues depend on the distance

3.4. Battle of Neighborhoods

Now, let's analyze what do boroughs have in common. Let's look if there are regions where some kinds of venues are more popular than the others.

Before doing so, only the necessary columns of data were chosen to work with:

	District	Borough	Latitude	Longitude	Venue	VenueLatitude	VenueLongitude	Category
0	Central Administrative Okrug	Arbat	55.751199	37.589872	Corner Café & Kitchen	55.751496	37.586757	Japanese Restaurant
1	Central Administrative Okrug	Arbat	55.751199	37.589872	Граци Рарацци	55.752137	37.591267	Italian Restaurant
2	Central Administrative Okrug	Arbat	55.751199	37.589872	Барашка	55.752055	37.586117	Middle Eastern Restaurant
3	Central Administrative Okrug	Arbat	55.751199	37.589872	Taksim	55.752240	37.591971	Turkish Restaurant
4	Central Administrative Okrug	Arbat	55.751199	37.589872	Чайхона № 1	55.752383	37.585924	Middle Eastern Restaurant

[ven_data] dataframe

K-Means Clustering

K-Means is a type of unsupervised learning, which gives pretty accurate results and does not take a lot of time and resources.

It was chosen as a main approach in this study, because it works with unlabeled data (which can be easily obtained with one hot encoding method from the data already present), and suits perfectly the goal to see similarity between different Moscow regions, according to what venue categories are present in each region.

One Hot Encoding

I started the process with one hot encoding venues categories with the built-in pandas tool, which gives binary code to each category:

```
# one hot encoding
m_onehot = pd.get_dummies(complete_data[['Category']], prefix="", prefix_sep="")
```

	Borough	American Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Bar	Belgian Restaurant	Bistro	...
0	Arbat	0	0	0	0	0	0	0	0	0	...
1	Arbat	0	0	0	0	0	0	0	0	0	...
2	Arbat	0	0	0	0	0	0	0	0	0	...
3	Arbat	0	0	0	0	0	0	0	0	0	...
4	Arbat	0	0	0	0	0	0	0	0	0	...

One hot encoding the categories

Frequency of occurrence

Next, rows were grouped by borough, and the mean value was taken to see the frequency of occurrence of each category:

```
m_grouped = m_onehot.groupby('Borough').mean().reset_index()
m_grouped.head()
```

	Borough	American Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Bar	Belgian Restaurant	Bistro	...
0	Aeroport	0.00	0.0	0.00	0.00	0.333333	0.00	0.0	0.0	0.0	...
1	Akademichesky	0.00	0.0	0.00	0.00	0.153846	0.00	0.0	0.0	0.0	...
2	Alexeyevsky	0.00	0.0	0.00	0.00	0.000000	0.00	0.0	0.0	0.0	...
3	Altufyevsky	0.00	0.0	0.00	0.00	0.000000	0.00	0.0	0.0	0.0	...
4	Arbat	0.02	0.0	0.02	0.02	0.100000	0.02	0.0	0.0	0.0	...

Frequency of occurrence of each category

Top 10 venues

Then the top 10 venues for each borough were found:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th
0	Aeroport	Coffee Shop	Bakery	Wine Shop	Burger Joint	Donut Shop	
1	Akademichesky	Sushi Restaurant	Bakery	Pizza Place	Cocktail Bar	Pub	
2	Alexeyevsky	Pizza Place	Diner	German Restaurant	Cocktail Bar	Coffee Shop	
3	Altufyevsky	Pizza Place	Café	Wings Joint	Coffee Shop	Creperie	
4	Arbat	Coffee Shop	Bakery	Burger Joint	Italian Restaurant	Turkish Restaurant	

Top 10 venues for every borough

K-Means Clustering

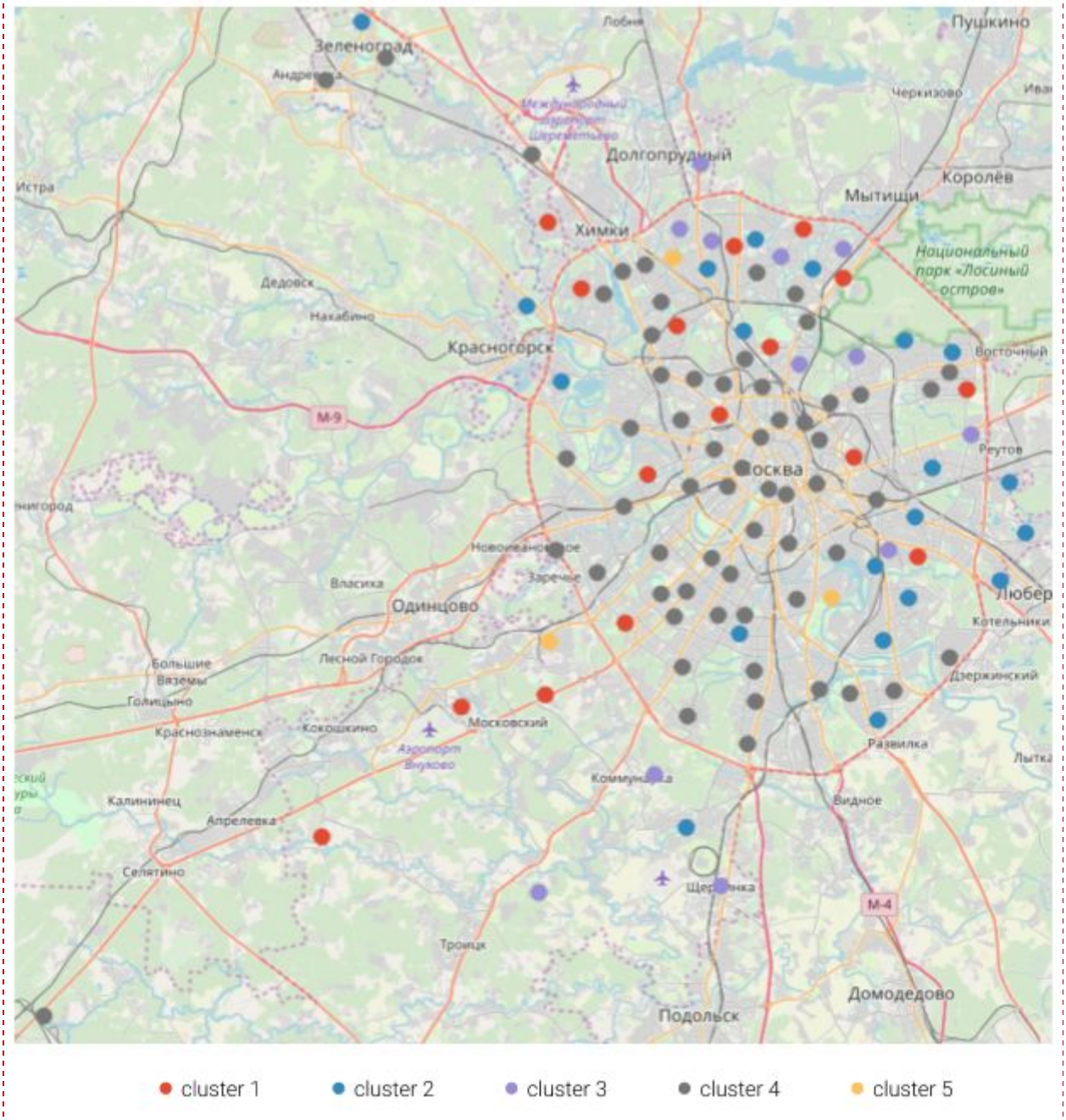
As there are not too many venues found by Foursquare free API, the logical decision was to limit the number of clusters to 5: less number is not informative enough, while larger number of clusters would be too meticulous.

After all clusters were determined, a `[m_merged]` dataframe looks like this:

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Co
0	Central Administrative Okrug	Arbat	55.751199	37.589872	3	Coffee Shop	Bakery	Burger Joint	Italian Restaurant	T Rest
50	Central Administrative Okrug	Basmanny	55.767281	37.669773	3	Coffee Shop	Irish Pub	Mexican Restaurant	Donut Shop	Cl
53	Central Administrative Okrug	Khamovniki	55.740047	37.573958	3	Pizza Place	Pastry Shop	Caucasian Restaurant	Bakery	M Eur Rest
60	Central Administrative Okrug	Krasnoselsky	55.777447	37.654160	3	Hotel Bar	Bakery	Coffee Shop	Asian Restaurant	Middle E Rest

Dataframe with cluster labels

Clusters on the Map



Folium map with clusters on it

Clusters Examination

Categories of Venues in Clusters

Total Number of Categories

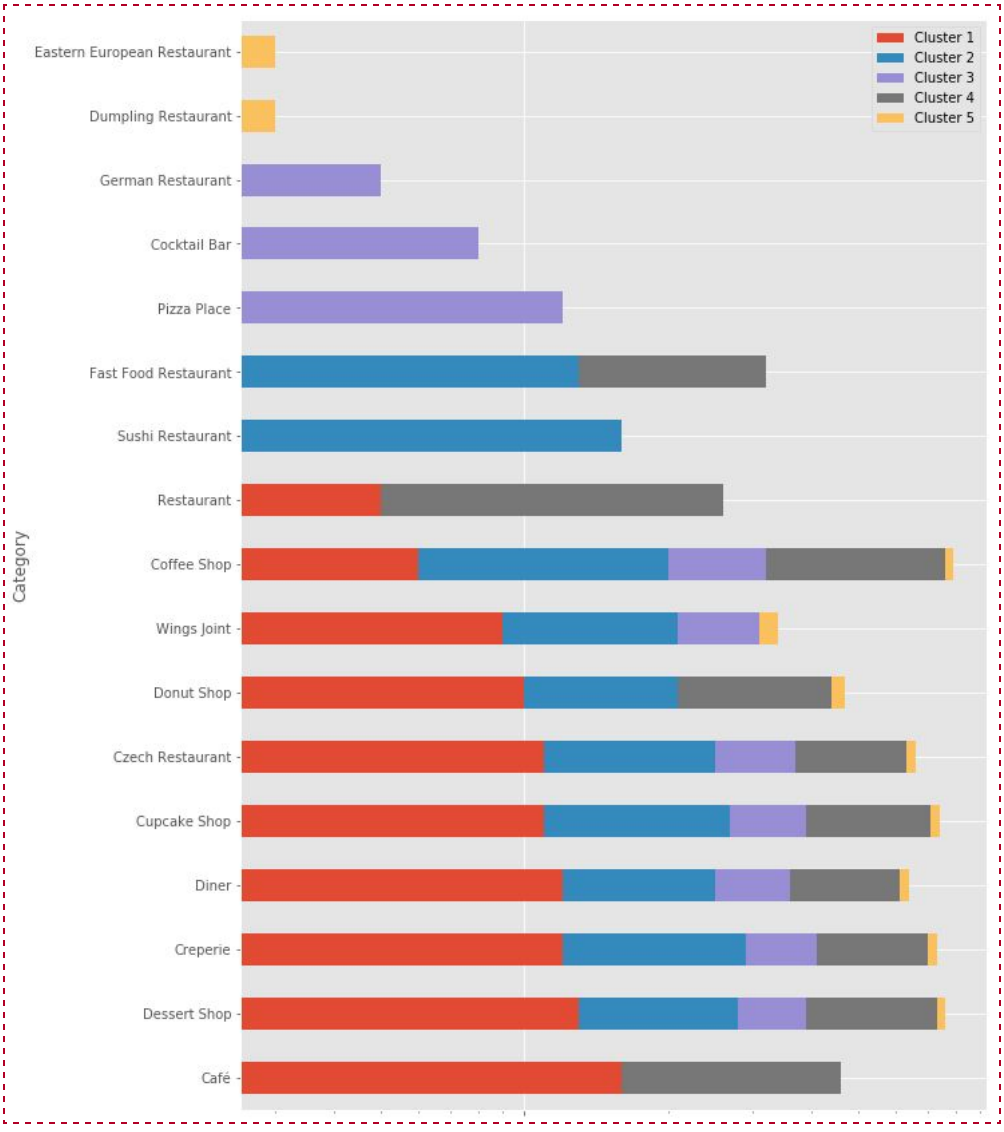
Among top-10 most popular venues, there are 16 categories.

	Category	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
0	Café	16.0	NaN	NaN	30.0	NaN	46.0
1	Dessert Shop	13.0	15.0	11.0	34.0	3.0	76.0
2	Creperie	12.0	17.0	12.0	29.0	3.0	73.0
3	Diner	12.0	13.0	11.0	25.0	3.0	64.0
4	Cupcake Shop	11.0	16.0	12.0	32.0	3.0	74.0
5	Czech Restaurant	11.0	14.0	12.0	26.0	3.0	66.0
6	Donut Shop	10.0	11.0	NaN	23.0	3.0	47.0
7	Wings Joint	9.0	12.0	10.0	NaN	3.0	34.0
8	Coffee Shop	6.0	14.0	12.0	44.0	3.0	79.0
9	Restaurant	5.0	NaN	NaN	21.0	NaN	26.0
10	Sushi Restaurant	NaN	16.0	NaN	NaN	NaN	16.0
11	Fast Food Restaurant	NaN	13.0	NaN	19.0	NaN	32.0
12	Pizza Place	NaN	NaN	12.0	NaN	NaN	12.0
13	Cocktail Bar	NaN	NaN	8.0	NaN	NaN	8.0
14	German Restaurant	NaN	NaN	5.0	NaN	NaN	5.0
15	Dumpling Restaurant	NaN	NaN	NaN	NaN	3.0	3.0
16	Eastern European Restaurant	NaN	NaN	NaN	NaN	3.0	3.0

Categories of venues in clusters

Categories Ratio in Clusters

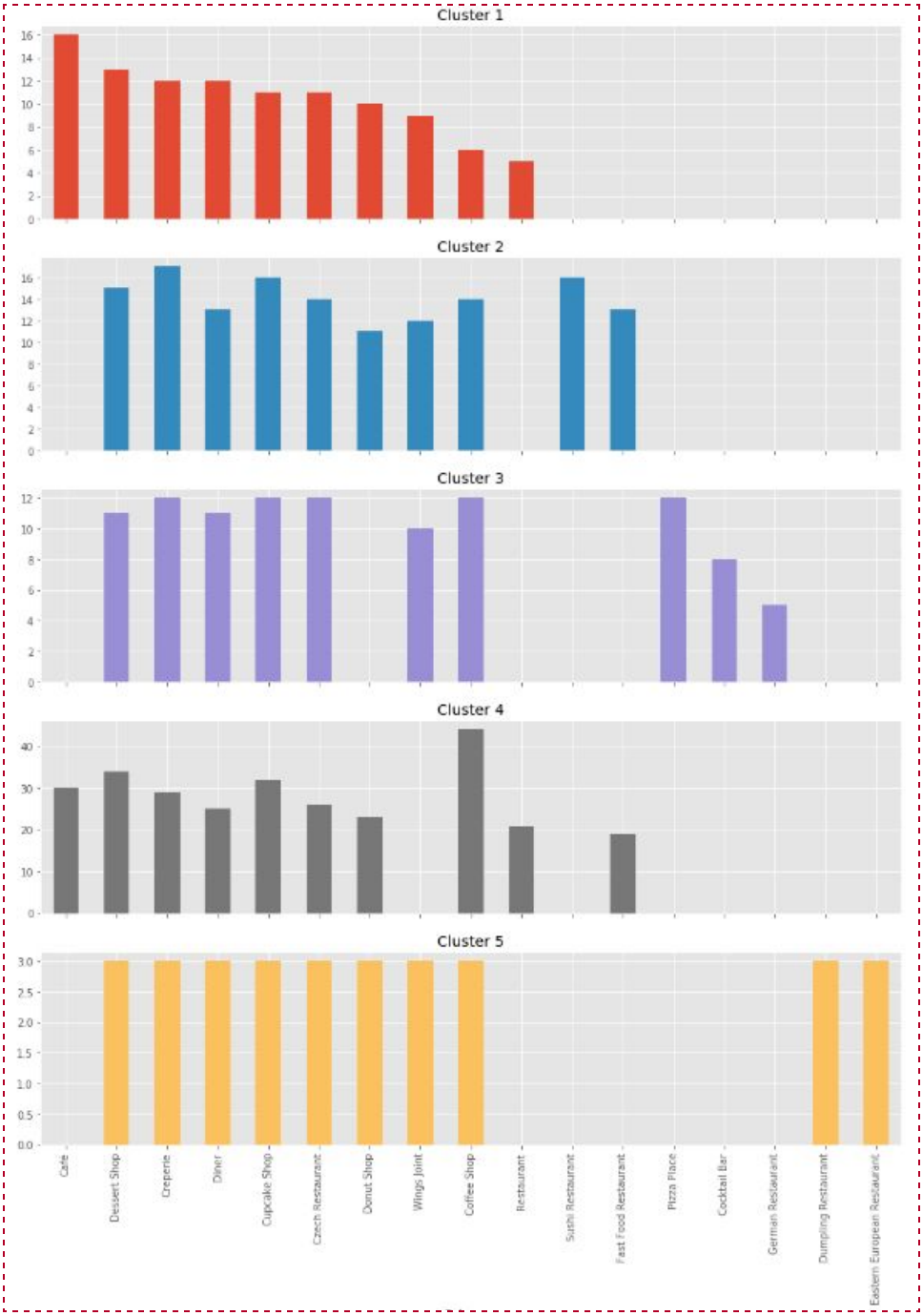
The following graph allows us to see how different categories of venues are spread among clusters. It becomes obvious why cluster 5 was highlighted by the algorithm, as it contains two unique categories.



Ratio of Venue Categories by Cluster

Number of Venues in Each Cluster

The following illustration shows categories of each cluster in numerical terms.

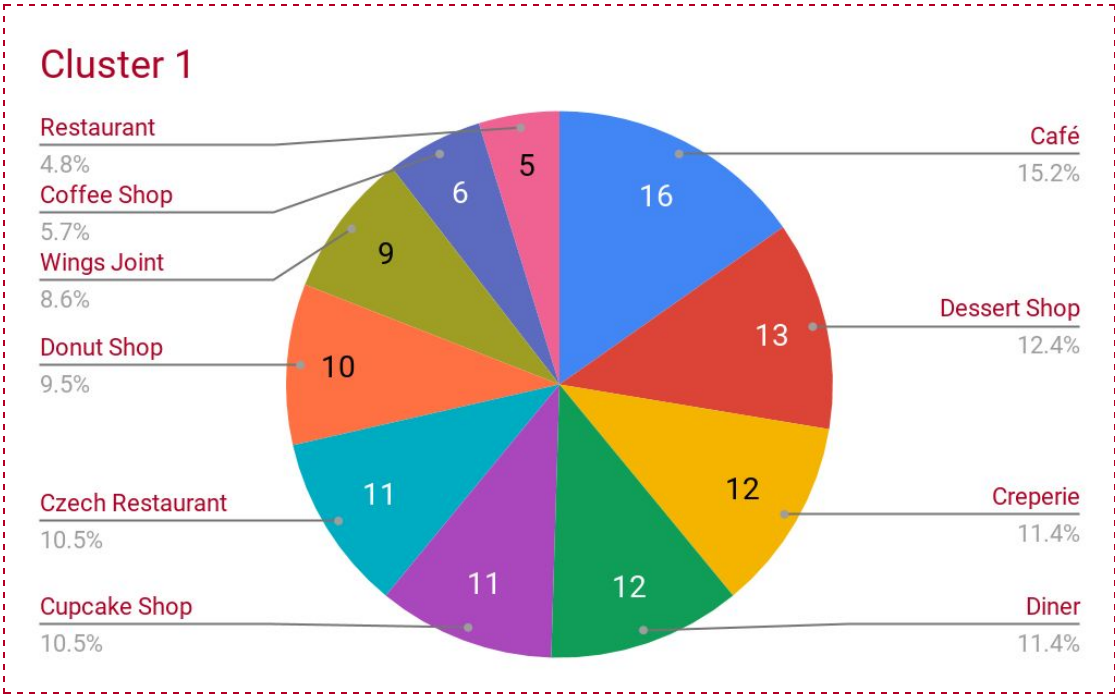


Number of Venue Categories by Cluster

Individual Clusters Examination

Cluster 1

Top venues are: Café (15.2%), Dessert Shop (12.4%) and Creperie (11.4%).



Cluster 1: most popular venues

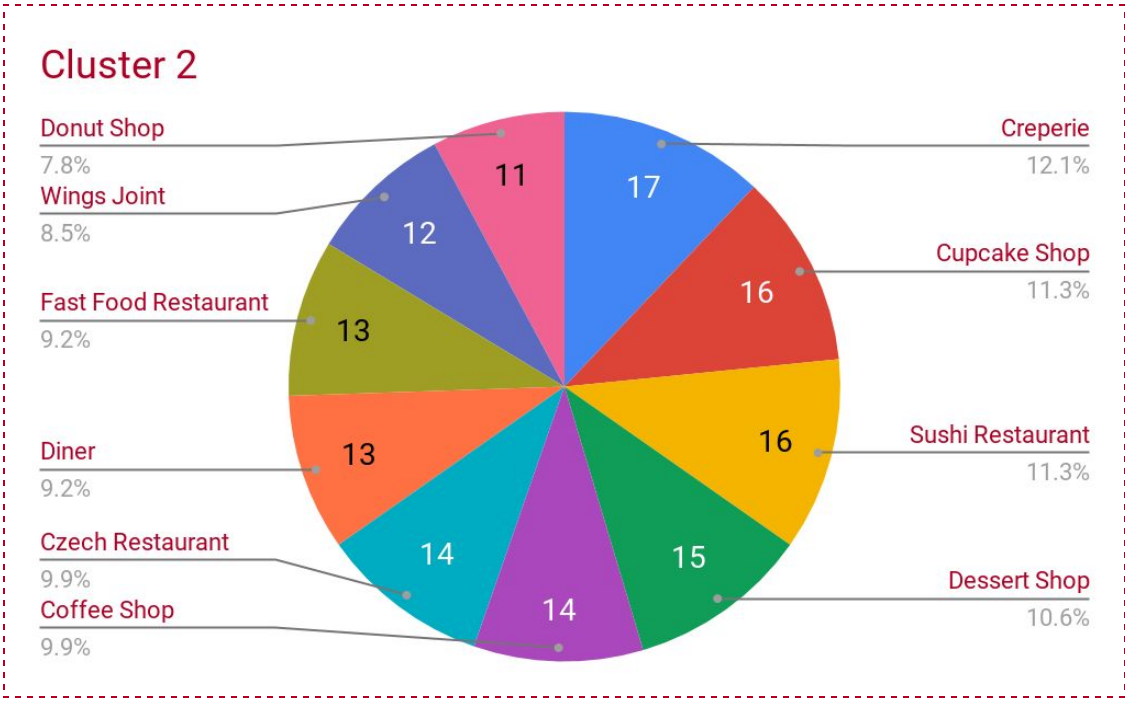
It seems like people in this cluster are seriously into sweet small bites.

There are some restaurants, like Czech restaurants, and Wings Joints, yet most of these venues are places where you can get something sweet: cupcakes, donuts, crepes.

About ⅓ of points is situated in the newly joined territory, which has low population and the majority of venues opened there are networks, like KFC or Dunkin Donuts, so there is a good chance, that this is where the popularity of Wing Joints and Cupcake Shops come from.

Cluster 2

Top venues are: Creperie (12.1%), Cupcake Shop (11.3%), Sushi Restaurant (11.3%).



Cluster 2: most popular venues

Cluster 2 points are mostly situated on the East of Moscow.

The popularity of Creperies is not surprising as well: crepes, or blinis, are considered Russian national meal, so people do love it here.

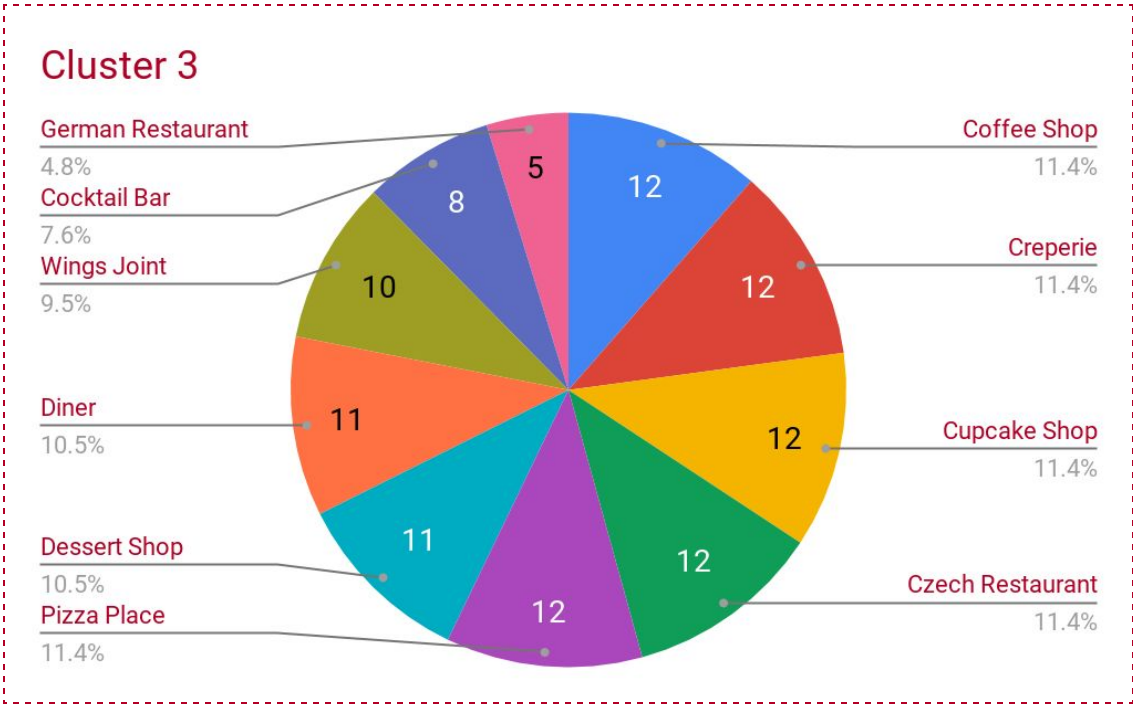
Sushi Restaurants are the third popular type of restaurant in Moscow, somehow Russian people fell in love with Japanese cuisine, and are not going to let it go.

Eastern districts are large and well-developed, so it's no wonder the beloved cuisine is very popular there.

There are a lot of historical parks and places, where you can grab a quick bite while sightseeing.

Cluster 3

Top venues are: Coffee Shop (11.4%), Creperie (11.4%) and Cupcake Shop (11.4%).



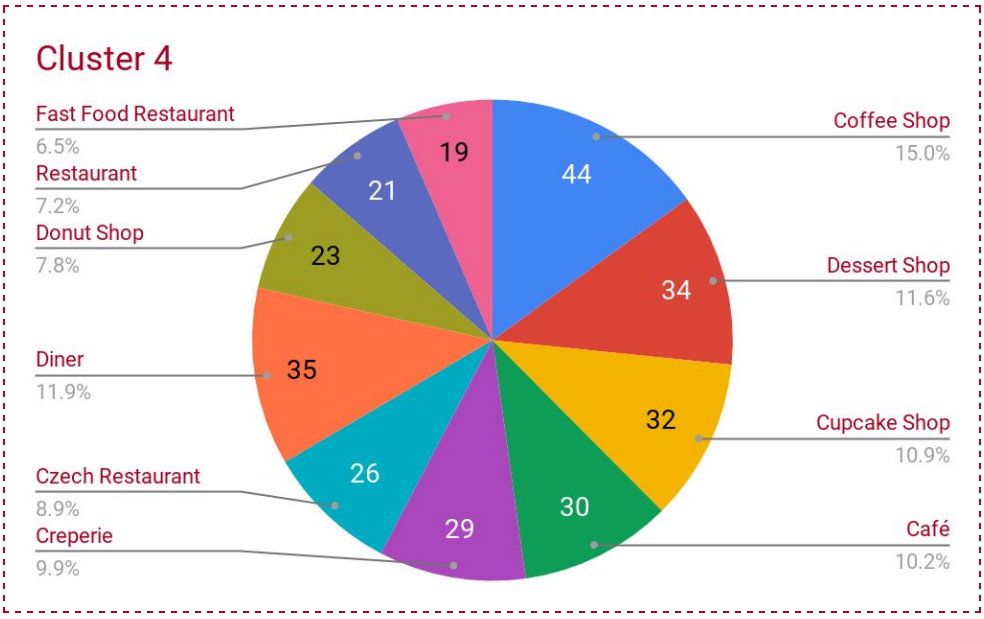
Cluster 3: most popular venues

According to the map, most of points of Cluster 3 are situated on the North of Moscow. It is a part of the city, mostly built in the 1960 - 1970's, and historically artists and aviators used to live there.

Interestingly, the only 3 points on the South of Moscow are nearby the Ostafyevo Airport, which is a class "B" international airport, renovated and opened for civilian flights in 2000 on the grounds of a former military airbase. So it seems like there might be some kind of correlation between aviators and popularity of Cocktail Bars and German cuisine, as this is the only cluster, where German restaurants and Cocktail bars are in the top-10 most popular venues.

Cluster 4

Top venues are: Coffee Shop (15.0%), Dessert Shop (11.6%) and Cupcake Shop (10.9%).



Cluster 4: most popular venues

Cluster 4 points are spread all over Moscow, which seem to be quite logical: Moscow loves its coffee. There are coffee shops literally in every block, so it's no wonder it's the most popular category in the largest cluster.

Russia is famous for its cold winters, and even though winter in Moscow is usually not as bad as in the most of the country, still, it takes a lot of coffee to get through many months of gloomy grey mornings.

Cluster 5

Top venues are: Coffee Shop (10.0%), Creperie (10.0%), Cupcake Shop (10.0%):

Cluster 5: most popular venues

This cluster is very small, it has only three points on the map, so it looks more like a residue. Interesting that it contains two unique categories for other clusters: Dumpling Restaurant and Eastern European Restaurant.

04. Results

There are 12 districts (or *okrugs*), 146 boroughs (or *rayons*) and settlements in Moscow. Each district sometimes seems very different from another, but this study showed how similar these regions might be concerning food.

All Moscow venues, acquired from the Foursquare API, were split into three categories: bars, fast food venues and restaurants. It appeared that most of venues in Moscow are fast food venues (along with cafes, coffee shops, creperies, etc.) (51.4%), and the least venues are bars (11.3%), and this ratio keeps the same among all the districts.

Moscow population is around 11.92 million people. After joining the large territory in 2012, Moscow are grew almost by half, with no significant change in population, so one of goals of this study was to find out if number of venues in each district depends on the population. Surprisingly enough, there was no correlation between these numbers.

Moscow is very disperse in terms of concentration of all kinds of venues, except for the city center. So another goal was to find out how the distance from the center of Moscow affects the number of venues. Initial assumption was confirmed: venue density significantly decreases with distance from the center. The main goal of this study was to split Moscow into clusters and compare them.

There were 5 clusters, chosen by the top-10 most popular venue categories.

There were categories, common for all clusters, such as coffee shops and creperies, but some categories appeared to be unique for some clusters, like German restaurant and Cocktail bar (cluster 5), or Sushi restaurant (cluster 2).

There were clusters, situated mostly in working zones or dormitory areas (such as cluster 2), or spread all over the city (4).

The most popular category in Moscow is Coffee Shop (it was the most popular category in 3 clusters of 5).

The result of the study is quite expected:

Moscow life is focused in the center, and the outskirts are used as dormitory areas, or manufacturing zones, where people prefer to eat on the go.

Russian people prefer cozy inexpensive places where they can have a coffee and desserts, instead of fancy restaurants. Speaking of restaurants, people in Moscow prefer sushi restaurants and Czech restaurants.

There are a lot of places, where there are too few venues of any kind, and it can be a good business strategy to place a fast food venue, if you want to create a very popular place, or local bar, if you prefer filling the niche.

05. Discussion

I've been living in Moscow my whole life, but this research brought several interesting discoveries. For example, I was sure about correlation between population and density of venues, but it appeared to be untrue.

I've always known that Moscow loves coffee, and that there were a lot of coffee shops around, but I had no idea of the number of creperies or donut shops. The other thing that surprised me, was the amount of bars. I thought there were going to be more of those, but it appeared that bars are one of the least popular venues (in the top-10, of course).

In my area there are a lot of all kinds of restaurants, especially, expensive ones, but the study showed that it's not that common in Moscow, and among all the restaurants, more popular are Czech restaurants along with sushi places (which was not surprising, as Japanese cuisine has been very popular in Moscow since the early 2000). I thought that Italian restaurants were going to be very popular, because several years ago there were more Italian restaurants than wing joints, but the study showed that things can change, as the popularity of this kind of venues dropped to insignificant numbers.

As a goal for further research, I would like to study the location of venues of different categories, what affects it, how does it change with the distance from center. And, it also would be interesting to repeat that research in several years, so see how Moscow changes in time, see the trends and make prognosis for future venues, and future businesses.

06. Conclusion

The goal of this research was to determine what and where does Moscow eat and drink. Is it true, that Russian people stick to their traditions and don't accept new streams? Is it true, that people in Russia drink a lot, and prefer bars to any other kind of venue? Is it true, that outside of the Moscow center there's almost nowhere to grab a bite.

Also, there was a task to split Moscow regions in clusters and see what they have in common, and what makes them different.

Well, all the goals were achieved.
Russian people do stick to their traditions, they still prefer creperies and dumpling shops to burgers or udon shops. And among all the cuisines of the world they chose Czech and German cuisine, as something, that's closer to their taste and habits. But they also love trying new things, and if they love something, they love it deeply like coffee shops and sushi restaurants. And no, Russians don't drink as much as Hollywood movies show: there are very few bars in Moscow comparing to other kinds of venues, and the only category that made it to the top-10 was not vodka bar, it was a cocktail bar.

Yes, in the outskirts of Moscow there might be from very few to no restaurants, but there's always a coffee shop or a cafe where you can grab a bite. So you'll never stay hungry, wherever you are in Moscow. So, welcome!

And thank you for reading my report!