Coursera

IBM Data Science Professional Certificate

Predicting the success of a certain type of restaurant in Russia

Data Section

Applied Data Science Capstone

by

Susanin Dmitry

November, 2019

1. Data acquisition and cleaning

1.1 Data sources

Most data was taken from wiki page https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Russia_by_population, to define the location of each city I used geolocator. I also found and downloaded additional dataset ('ru.csv', from https://simplemaps.com/data/ru-cities), just as a precaution (because sometimes geolocator can't find any information about some cities). I didn't use it as main dataset because it hadn't fresh info (only 2010).

1.2 Data cleaning

I didn't use such columns as Rank, Federal subject names, Federal district names, Population in 2010 and change in it between 2010 and 2017, that's why I dropped them. That was enough to get a clean dataset.

1.3 Feature selection

I am only using 300 cities, because I think that the interest of investors is directly proportional to the population of the city, so I don't need to analyze small cities (towns). That's why I got 300 samples with 5 features. It may seem that this is not enough, but do not forget that the main part of the data I will get from the results of the FourSquare API work.