The Battle of Neighborhoods with Irish Pubs

Introduction

Background

Irish pubs. From Wikipedia: "Irish pubs are characterised by a unique culture centred around a casual and friendly atmosphere, hearty food and drink, Irish sports, and traditional Irish music. Their widespread appeal has led to the Irish pub theme spreading around the world." I love irish pubs. I love irish food. I love irish beer. But I live in Moscow, Russia. And I'm not irish at all.

Business Problem

There are a lot of Irish pubs in Moscow. And in this project we will analyze them. We will explore and compare neighborhoods of Moscow based on presence of Irish pubs. We will look for areas that have pubs, but with a low rating or no pubs at all. And will look for areas with a high density of pubs and high ratings.

Interest

This report will be directed to customers looking for a good location to open an Irish pub. The report will also be of interest to people who are looking for areas for renting an apartment with good pubs within walking distance. We will use Foursquare location data and our data science powers to come up with solution that best suits our needs.

Data acquisition and cleaning

Data sources

Number of pubs and their type and location, ratings in every neighborhood will be obtained using **Foursquare API.**

Moscow neighbourhoods and their location and borders will be obtained using http://mosopen.ru/regions. Borders will be stored in GeoJSON format.

Moscow neighbourhoods areas and population will be obtained using https://ru.wikipedia.org/wiki/Районы_и_поселения_Москвы

Data cleaning

Moscow neighbourhoods data will be scraped from multiple sources and combined into one table. But there is a problem with datasets. Moscow neighbourhoods identified by their
names. And different data sources may use a different word order in the name of the neighbourhood or may use lowercase letters. It should be manually fix before scrapping data from different sources. Otherwise there will be missing values.

Searching query 'Irish Pub' can cause Foursquare API to return not related to pubs venues, such as offices. This will be cleaned using venue category.

Also searching fro venues in particular neighbourhood will be based on latitude and longitude of the center of neighbourhood. And radius of the search will be based on neighbourhood area. So there is a chance that Foursquare can return venues from nearest neighbourhood beyond the borders of current neighbourhood. This will cause some venues to be duplicated in resulting datasets. Duplicates will be removed based on venue id.

Feature selection

After data cleaning there will be two datasets with general information about neighbourhoods. Such as: name, latitude, longitude, area, population density, borders. And data set with information about pubs. Such as: name, latitude, longitude, neighbourhood, category, rating, likes. Based on this we will compare our neighbourhoods and find most promising ones for opening an Irish pub or living perspective.