Final Project - The Battle of the Neighborhoods!

# Introduction

## Background

The Israeli socio demographic is a complex issue that’s being discussed throughout the world’s main news channels for the past century.

The main dispute is the everlasting city of Jerusalem. The unique texture of population, which is a combination of (mostly) religious Jews, non-religious Jews and Arabs - makes it a very special place.

## Problem

In this challenge, I'm going to tackle the social texture of the city of Jerusalem in the context of business by neighborhoods analysis. I'll conduct a thorough analysis which will conclude with a clustering for the various venues throughout the different neighborhoods. as a city with a unique socio-demographic texture, this analysis should be super interesting. this problem should appeal to anyone who cares about the socio demographic texture of the city.

# Data acquisition, cleaning and EDA

## Data sources

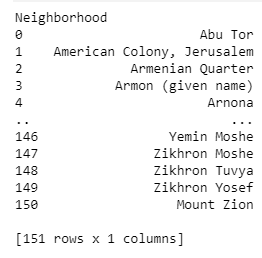
Wikipedia pages which include the list of neighborhoods for the city of Jerusalem.

Geo-code location specifics (geocode + google).

Foursquare venue data.

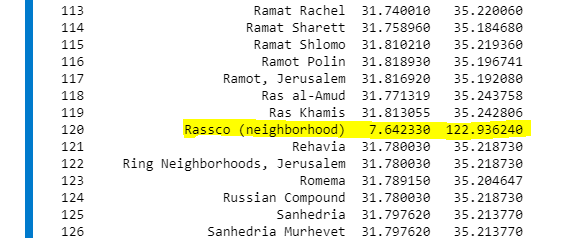
## Data cleaning

Scraping the data from non-tabular web page. Somewhat tedious. Ensuring naming is correct, no duplicates, no missing values or empty fields.

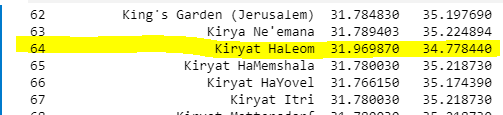


Mapping each of the neighborhoods to the relevant set of coordinates and validating them

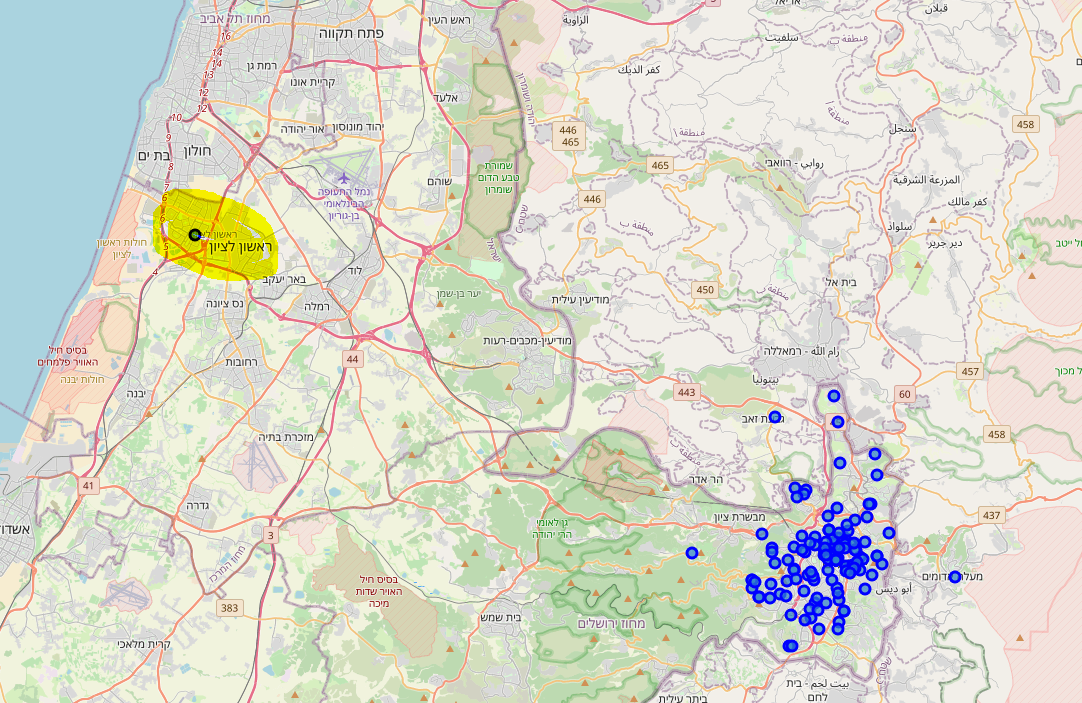
Validating geo location, step one – simple “eyeball” validation, relatively straightforward, a good example of a mismatch is the following highlighted observasion:



Harder to track relatively smaller mismatches, we can use the visualization tool and deep dive into the map to explore any further anomalies, for example, the following row might seem “normal” when just exploring the table:



However, once we visualize it on the map, we can clearly see that it’s an outlier:



# Further data acquisition, EDA and parameters for extraction:

## Radius

Grabbing the relevant venues for the different neighborhoods faces us with another challenge, what should we choose as the radius from the neighborhood’s geo-code?

This is a unique challenge as Jerusalem’s neighborhoods’ size has high variance. While some neighborhoods are small (both in terms of residence and actual size), other are huge.

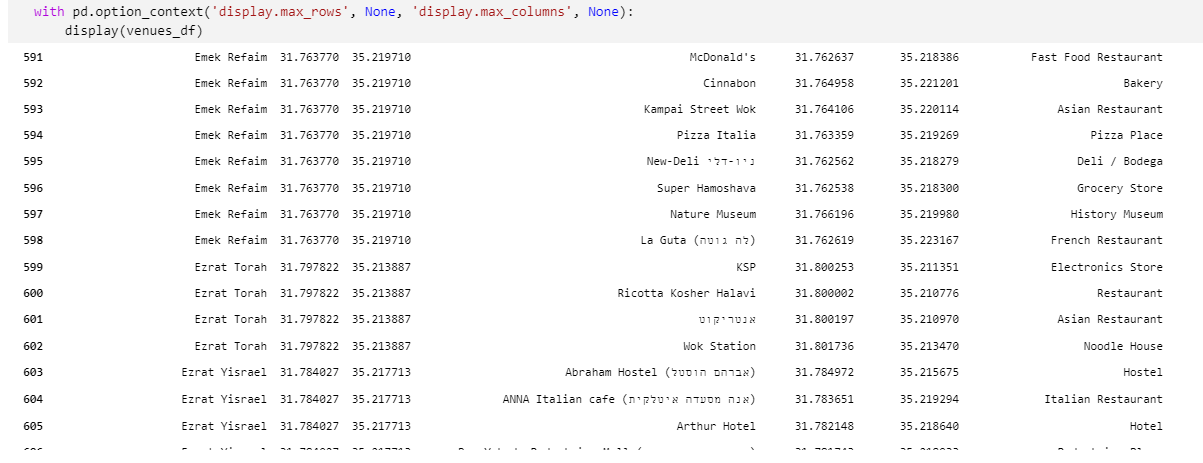
I’ve tried various set of ranges, and it turned out 500M seems representative for most neighborhoods – I work under the assumption that I might be losing a LOT for the larger neighborhoods, however, the need for no-overlap strikes me as more important.

## Venues per Neighborhood

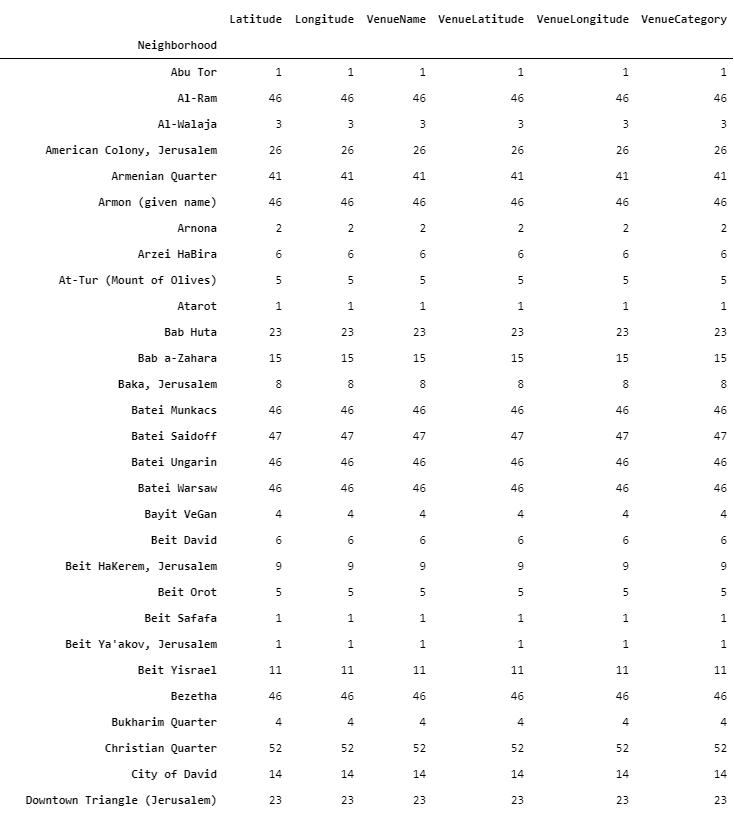
As I fetched the data, I set the venue per neighborhood parameter to 100, so for approx. 140 neighborhoods, I’ve expected a around 13k-14k results.

However, the number of fetched venues was about 3k. Initially, I figured it had to be a coding error / API issue. After further exploring the data and, code, API and responses per iteration, I decided to deep dive and see the breakdown by neighborhood.

Sorted by neighborhood, we see the below table:



If we’d like to group, it per neighborhood to see the aggregated results we’ll get the below summary:



As we can see, the variance in the number of extracted venues per neighborhood is very high.

After exploring it further, it appears to be related to the NATURE of the population that resides in the neighborhood.

As discussed in the beginning, Jerusalem has three main types of population:

1. Religious Jews
2. Non-religious Jews
3. Arabs

There are some other minority groups, however, they’re not crucial or relevant for this analysis.

the distribution between the 3 groups are 1/3 in each (approximately). In addition to have a very different religious views, they have very different social/technological approach. In general Religious Jews don’t interact with social media / go online (some of them have “Kosher” phones- which isn’t a smart phone). They don’t use the web, etc. The Arab population, while they don’t share the restrictions, they suffer from low income, and educational level, which correlates with social media / online interactions etc.

At this point I was facing a decision. There are two types of analysis I can perform,

1. Between analysis – try to estimate the difference between the various groups. Hopefully, there will be some interesting insights and the data will align with the above assumptions and priors, or perhaps, we’ll discover a different “truth” which we’ll need to explore further. All of that – of course when taking into account the fact of sufficient reliable and robust data to support the assumptions (discussed later)
2. Within analysis – pick and choose a population and explore it further. A good approach would be to sample the non-religious Jewish population (again, will be explained later).

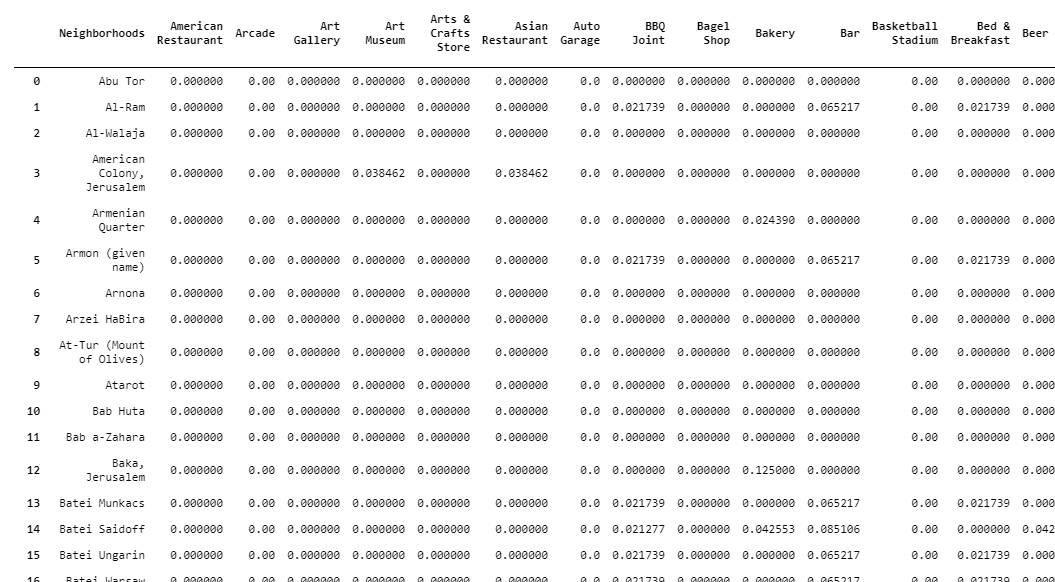
There are two relevant issues to consider,

1. Data robustness per analysis. In the end of the day, we need good and reliable data (and “enough” of it) in order to conduct any type of analysis and check out hypothesis/ find trends etc. In our case, there’s a big issue of lacking data for 2 out of the 3 sectors. Effectively we have sufficient amount of data to conduct a reliable analysis only for the non-religious Jewish sector.
2. Interest of study – while data robustness is important, this isn’t a statistics class :D I think discussing between population difference is way more interesting.

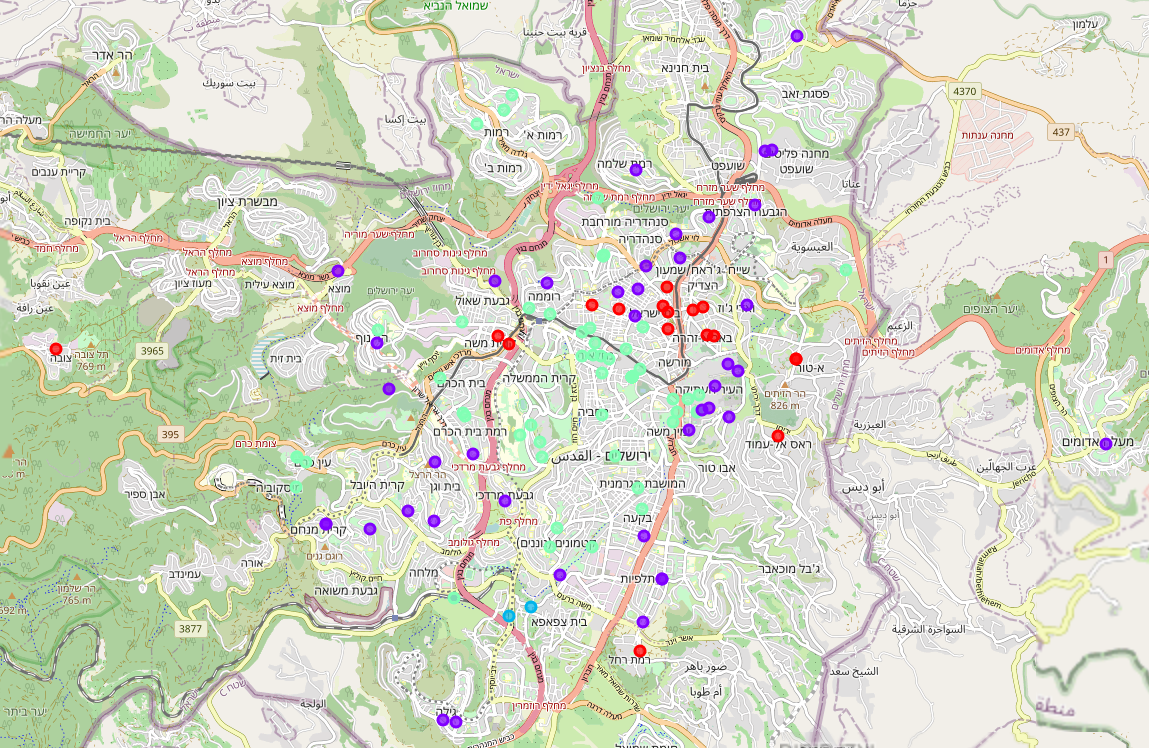
In the end of the day, I’ve decided to go with my heart and choose the between analysis, with the hope that someday, I’ll get enough reliable data, and will be able to continue the study.

# Clustering method

Next, I’ve prepared the data for clustering, everything’s in the code, but essentially, I’ve utilized foursquare data and extracted the unique category of venues per neighborhood. But a OHE matrix, then the frequency matrix for the entire set of neighborhoods, and ran the kmeans algo on it.



I’ve then ran several specifications of the kmeans algorithm and received the following allocation to clusters: Of course we should visualize it in order to evaluate and explore:



# Results and Summary

I've started with grabbing the data from the Jerusalem neighborhoods wiki page. I've then grabbed the relevant longitudes and latitudes per neighborhood. validated the data, cleaned for outliers. displayed over the map of the Jerusalem Area (including suburbs). The distribution makes sense and fits the real state of the world.

I've then created the relevant datasets utilizing the google geocode and foursquare APIs to grab relevant venues for each one of the neighborhoods (around 135 neighborhoods in total). an initial observation was that we've received a (relatively) low number of venues. As i was curious about the reason for that, i had a breakdown by the neighborhood. where it shows a drastic variance of venues for neighborhoods.

An important aspect of Israel, and Jerusalem in particular is the social texture. the city is somewhat split - 1/3 are religious Jews, 1/3 are non-religious Jews and 1/3 are Arabs (roughly). It appears that we have less information as a whole for the 2/3 of the Arab and the religious Jewish neighborhoods - which aligns with our expectations (these fractions of the population are somewhat less vanguard when it comes to technology). I've decided to move forward and analyze the entire set of neighborhoods as I figured a breakdown between populations is more interesting than a within population analysis (however, perhaps for future tasks...)

the clustering results make some sense as it appears to capture the difference between relatively mid-income non-religious Jewish neighborhoods (light green), and low income neighborhoods (purple), the Arab neighborhoods either fall in the purple (low income) neighborhoods or in red (east Jerusalem) . when conducting this analysis, we need to consider the low number of observations for some of the neighborhoods which affect the results drastically.