Restaurant Location Analysis

Guy Katriel
July 17th, 2020

A. Introduction - The Business Problem

According to CofaceBdi [1] Over 2017, Israelis have spent over 20 Billion shekels on the restaurant sector, a 3.5% increase from 2016. At the same time, the level of risk for the restaurant industry was measured at 6.86 (on a 1-10 scale), over 2018, 835 Israeli restaurants went out of business, while 2076 new venues were opened in this sector [2].

Although a relatively lucrative sector, restaurant entrepreneurship is defined by a sizable measure of risk, any potential business owner in this industry would benefit from maximizing his chances of success.

One of the most important factors leading to the success of a restaurant is its location [3][4]. In this project, I will attempt to recommend, based on the available data and analysis of it, an optimal location for a new restaurant, and it's preferable menu style, on the basis of existing venues in the different cities of Israel.

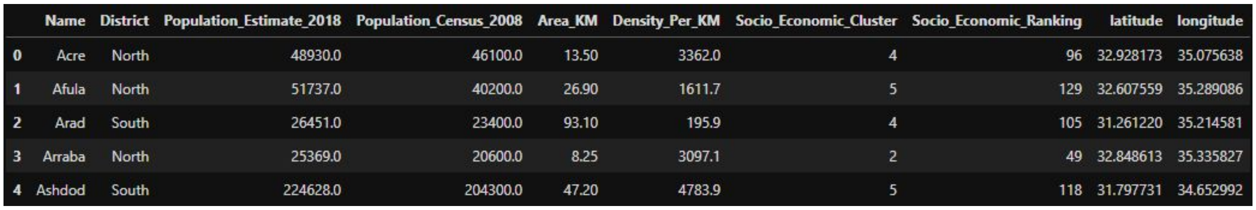
We will assume the business stakeholder in this situation is an entrepreneur interested in opening a new restaurant around Israel & in an urban environment, we will work under the assumption that the prevalence of specific menu types per area is indicative of the demand for that restaurant-style for a specific area.
We will also use socio-economic index data in order to segment Israeli cities, to try and differentiate between preferred venues on the assumption that Socio-Economic variables affect dining preferences [5][6], and eventually, make various recommendations based on this categorization.

B. Data

Data Structure

Structure for basic dataset:

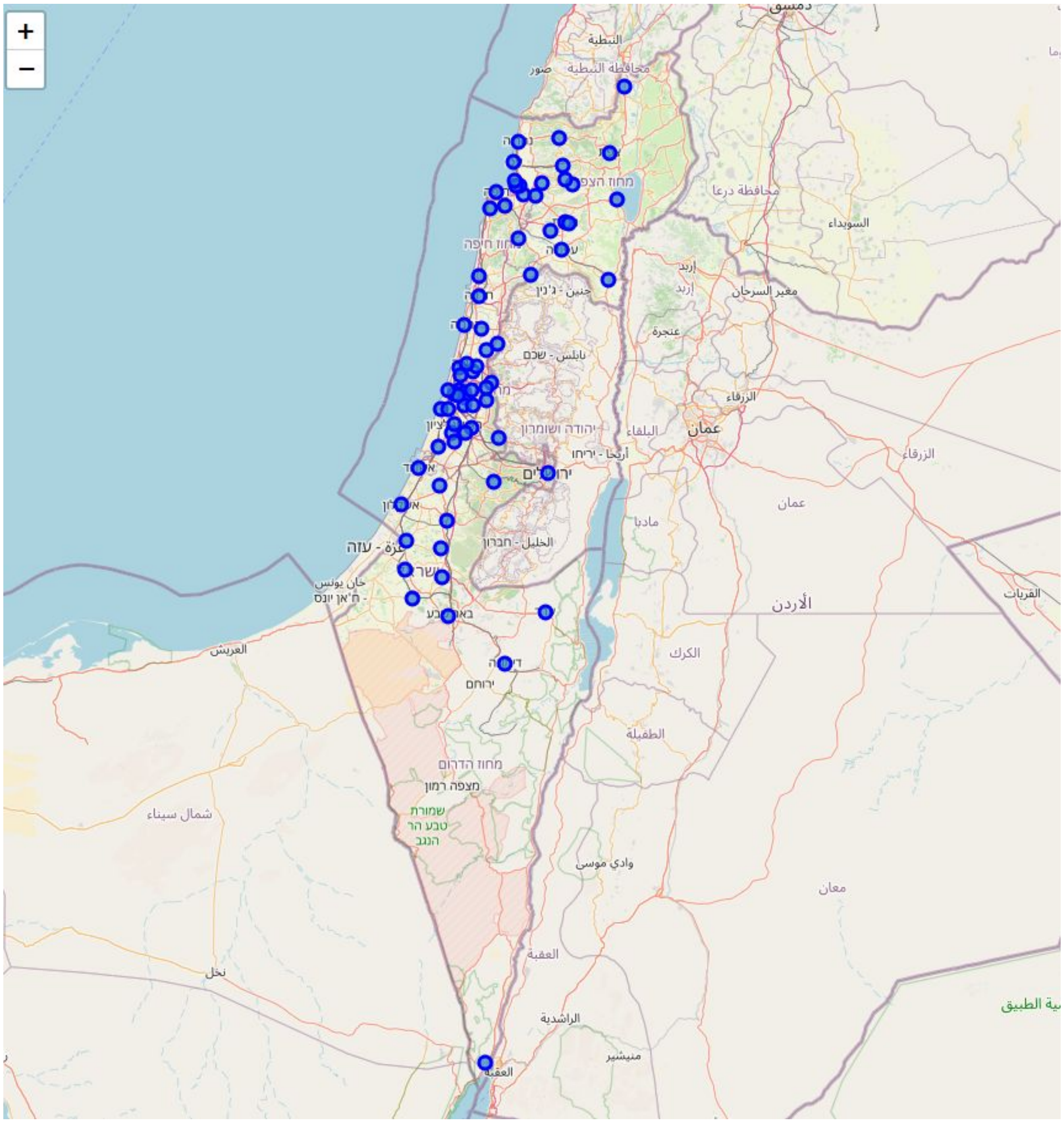
City	Name of city in Israel
District	District attributed to each city
Population estimate 2018	Population estimate for 2018, based on Israel Central Bureau of Statistics[7]
Population Census 2008	Population Census from 2008, based on Israel Central Bureau of Statistics[8]
Area_KM	City area in KM, based on Israel Central Bureau of Statistics[9]
Density per KM	Population density per KM for each city, based on Israel Central Bureau of Statistics[10]
Socio Economic Cluster	Percentile groups, 1-low, 10-high, based on Israel Central Bureau of Statistics [11]
Socio Economic Ranking	Rank among all Iseaeli Settlements, goes from 1- low, to 256 high, based on Israel Central Bureau of Statistics [11]
Latitude	Latitude for each city, based on opencagedata API [12]
Longitude	Longitude for each city, based on opencagedata API [12]



	Name	District	Population_Estimate_2018	Population_Census_2008	Area_KM	Density_Per_KM	Socio_Economic_Cluster	Socio_Economic_Ranking	latitude	longitude
0	Acre	North	48930.0	46100.0	13.50	3362.0	4	96	32.928173	35.075638
1	Afula	North	51737.0	40200.0	26.90	1611.7	5	129	32.607559	35.289086
2	Arad	South	26451.0	23400.0	93.10	195.9	4	105	31.261220	35.214581
3	Arraba	North	25369.0	20600.0	8.25	3097.1	2	49	32.848613	35.335827
4	Ashdod	South	224628.0	204300.0	47.20	4783.9	5	118	31.797731	34.652992

Data Sources

- Basic data was acquired by scraping the wikipedia website entry for “List of cities in Israel” [13], the table included in the entry was used for basic city information.
- A Geolocating API, OpenCage Geocoder[12] was used to acquire latitude and longitude values for each city in the database.
- A record of venues for cities in the database was acquired through the Foursquare API [14], a limit of 100 venues per city in a radius of 4 Km was set. Map plot of the cities included in the dataset:



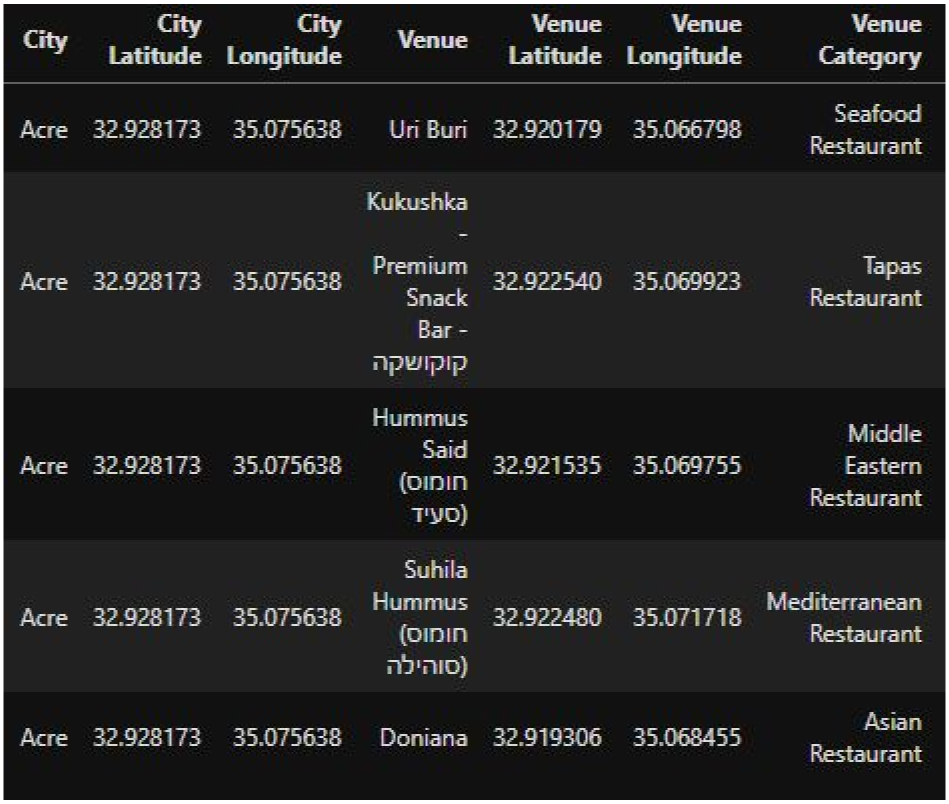
- Data Cleaning

After scraping the base date from Wikipedia, some columns were removed: The columns “First Settlement”, and “Changes 2008-2018” were dropped due to being irrelevant to our purpose. The column “Socio-Economic Index” was dropped due to technical reasons (inappropriate date type) and preference for a non-normalized index. One city was removed from the dataset due to missing values for certain columns, the city dropped was Baqa al-Gharbiyyem.
All HTML leftover code was removed (“\n”) with other irrelevant signs (“,”) in order to adequately identify numeric variables - those relevant columns were then changed to floats and integers accordingly.

Due to the Socio-economic variable included in the original table being unusable, relevant Socio-economic city index was added via a separate source [11], using an accumulated excel tabular data file.

Venue information was collected via the Foursquare API [14], general venue information per city was collected with a limit of 100 per city and 4 KM radius. The collected venues included non-relevant categories beside restaurants - and were therefore cleaned from the data, leaving it with 62 unique types of restaurant categories.

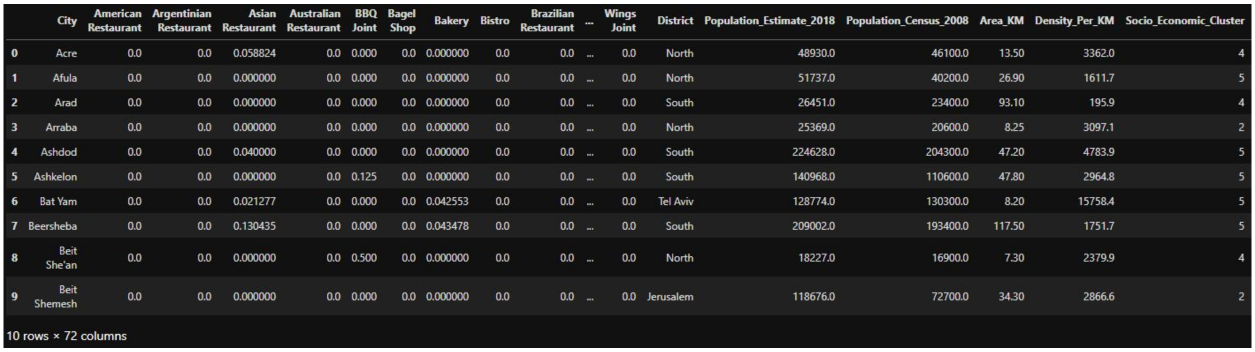
Example of venue information:



City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Acre	32.928173	35.075638	Uri Buri	32.920179	35.066798	Seafood Restaurant
Acre	32.928173	35.075638	Kukushka - Premium Snack Bar - קוקושקה	32.922540	35.069923	Tapas Restaurant
Acre	32.928173	35.075638	Hummus Said (חומוס סעיד)	32.921535	35.069755	Middle Eastern Restaurant
Acre	32.928173	35.075638	Suhila Hummus (חומוס סוהילה)	32.922480	35.071718	Mediterranean Restaurant
Acre	32.928173	35.075638	Doniana	32.919306	35.068455	Asian Restaurant

After collecting the venue information into a venue-level dataset, in order to merge it to the city-level dataset a “one-hot encoding” method was used - dichotomous dummy variables were created for every venue on the basis of the venue category (1 - yes, 0 - no). The rows were then grouped by city and a set of variables containing the mean of the frequency of occurrence of each category was created and merged with the city dataset.

Example of merged dataset:



	City	American Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Brazilian Restaurant	...	Wings Joint	District	Population_Estimate_2018	Population_Census_2008	Area_KM	Density_Per_KM	Socio_Economic_Cluster
0	Acre	0.0	0.0	0.058824	0.0	0.000	0.0	0.000000	0.0	0.0	...	0.0	North	48930.0	46100.0	13.50	3362.0	4
1	Afula	0.0	0.0	0.000000	0.0	0.000	0.0	0.000000	0.0	0.0	...	0.0	North	51737.0	40200.0	26.90	1611.7	5
2	Arad	0.0	0.0	0.000000	0.0	0.000	0.0	0.000000	0.0	0.0	...	0.0	South	26451.0	23400.0	93.10	195.9	4
3	Arraba	0.0	0.0	0.000000	0.0	0.000	0.0	0.000000	0.0	0.0	...	0.0	North	25369.0	20600.0	8.25	3097.1	2
4	Ashdod	0.0	0.0	0.040000	0.0	0.000	0.0	0.000000	0.0	0.0	...	0.0	South	224628.0	204300.0	47.20	4783.9	5
5	Ashkelon	0.0	0.0	0.000000	0.0	0.125	0.0	0.000000	0.0	0.0	...	0.0	South	140968.0	110600.0	47.80	2964.8	5
6	Bat Yam	0.0	0.0	0.021277	0.0	0.000	0.0	0.042553	0.0	0.0	...	0.0	Tel Aviv	128774.0	130300.0	8.20	15758.4	5
7	Beersheba	0.0	0.0	0.130435	0.0	0.000	0.0	0.043478	0.0	0.0	...	0.0	South	209002.0	193400.0	117.50	1751.7	5
8	Beit She'an	0.0	0.0	0.000000	0.0	0.500	0.0	0.000000	0.0	0.0	...	0.0	North	18227.0	16900.0	7.30	2379.9	4
9	Beit Shemesh	0.0	0.0	0.000000	0.0	0.000	0.0	0.000000	0.0	0.0	...	0.0	Jerusalem	118676.0	72700.0	34.30	2866.6	2

10 rows × 72 columns

C. Methodology

In this project I will attempt to identify differing restaurant menu-style preferences between urban areas within Israel, our analysis will be limited to Israeli cities and restaurant-style venues only. We will also attempt to segment these results on the basis of socio-economic index per city, to ascertain whether preferences are impacted by socio-economic city status.

As a first step, I have collected and manipulated data sources in order to acquire a list of Israeli cities and some relevant background information (including Socio-Economic city status), Longitude & Latitude date per city, and a list of relevant venues per city.

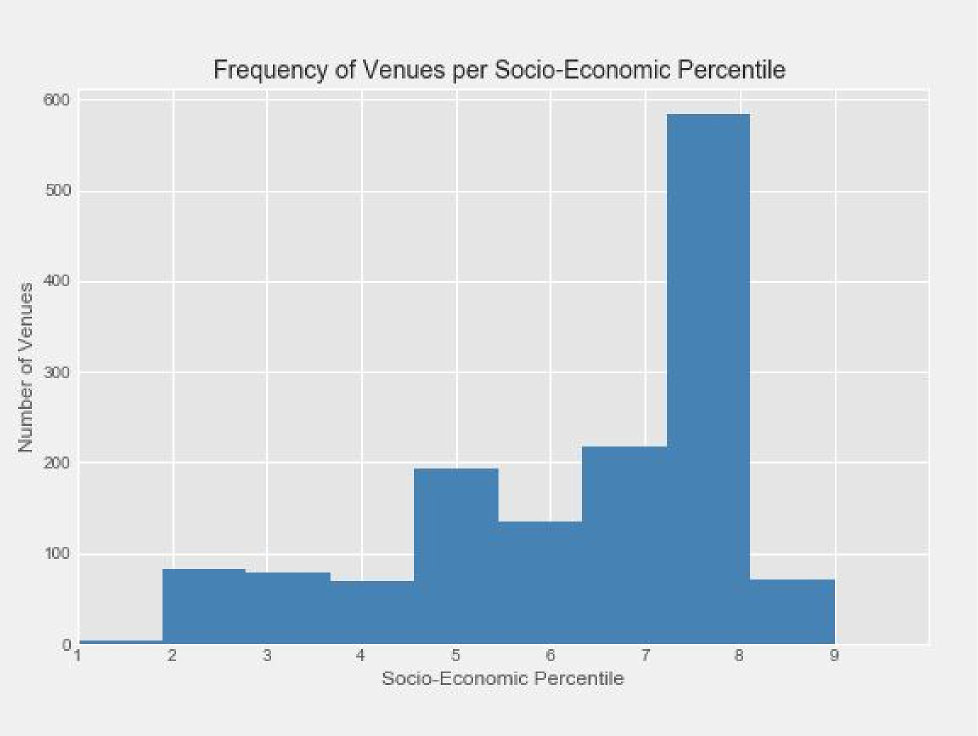
For the second step, we will analyze and assess the properties of the collected data, in order to both elucidate relevant aspects to our cause and to discover possible bias in the data.

The third and final step will include a k-Means Clustering analysis aiming to segment different areas according to the prevalence of restaurant menu-style in them, this analysis will be performed on all venues in the sample, and also separately for specific socio-economic categories.

D. Analysis

- Exploratory Data Analysis

First, in order to determine an ideal categorization of Socio-economic groups , a histogram of venues per Socio-economic city index was created.
The categorization should take into account the amount of venues per Social economic percentile so that each category contains a sufficient N.

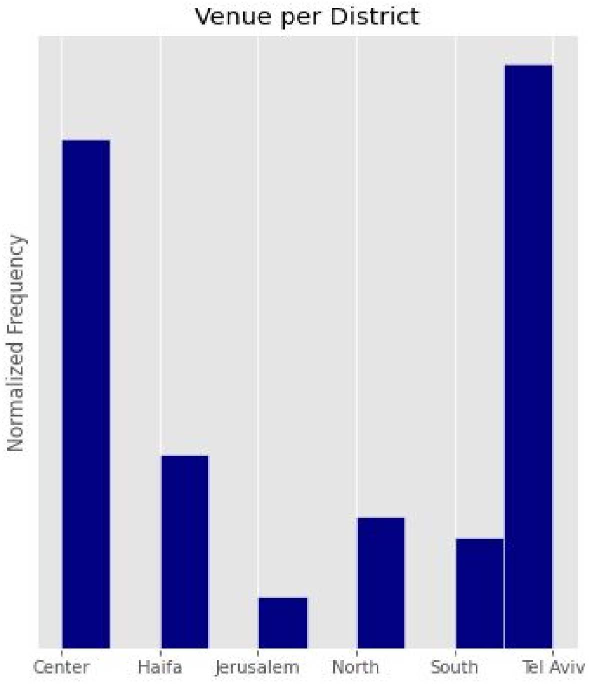


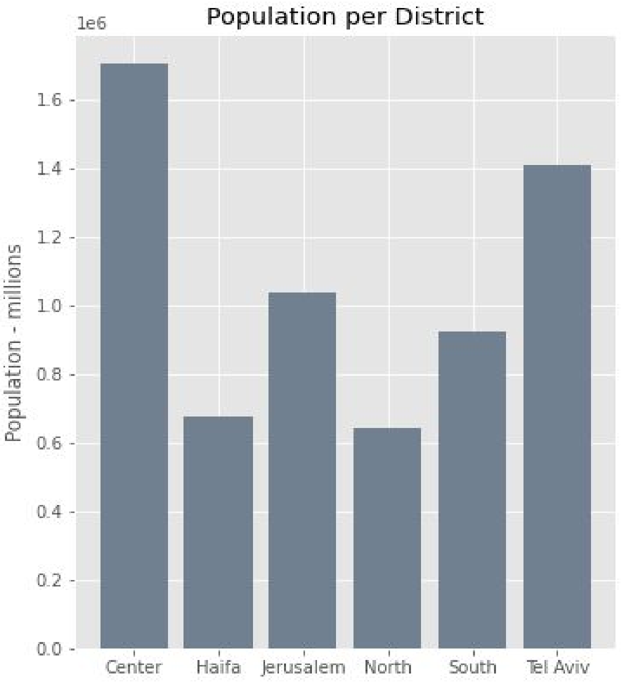
Based on the plot, it is possible the dataset might contain a bias, restaurant location is much more frequent in the higher half of socio-economic percentiles. It can't fully be ascertained if this is due to data bias, or a realistic phenomena represented in reality.

Regardless, this would be considered when constructing socio-economic categories.
As a result, socio-economic categories were constructed from the percentile data as such:

1-4 : “Low”
5-7: “Mid”
8 9: “High”
No cities in the 10th percentile were included in the dataset.

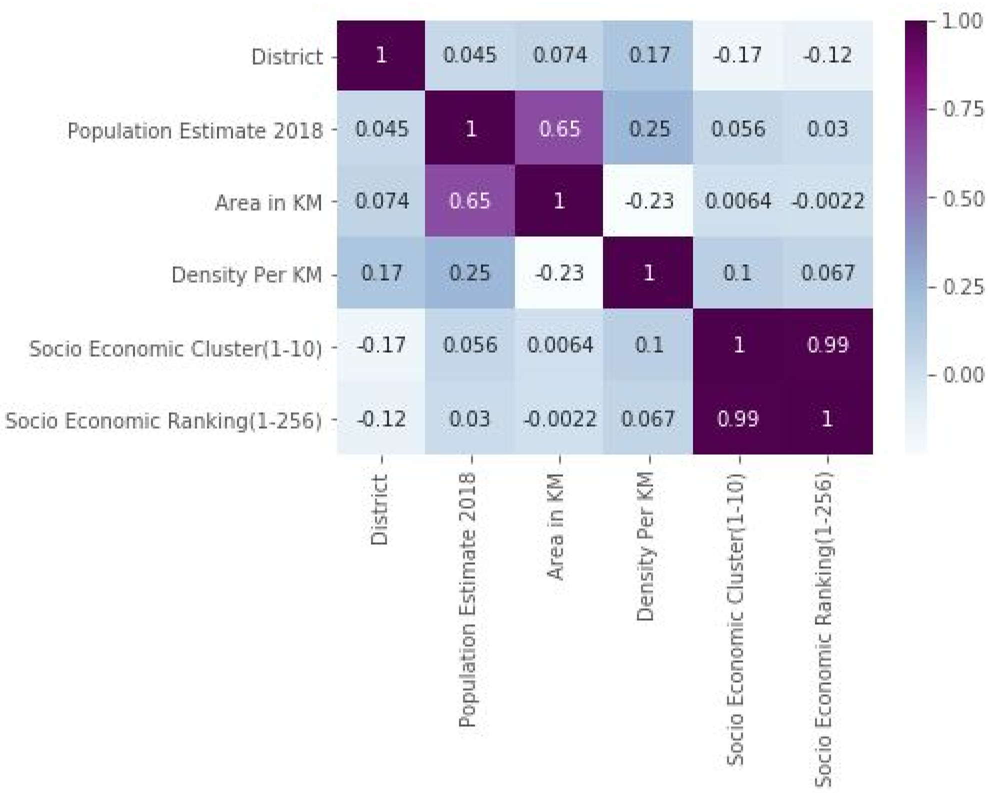
Afterwards, I examined any possible disparities between the distributions of both venue per district and population per district.
Venue count were normalized to better represent the distribution.





It can be seen that venue location does not align with population count,
Venues are highly concentrated in the Center & Tel-Aviv districts of Israel, While population is distributed somewhat more equally.
A particularly large miss-alignment is in the Jerusalem District.

A correlation matrix was run to further understand the relationships between relevant numeric & categorical variables in the data:



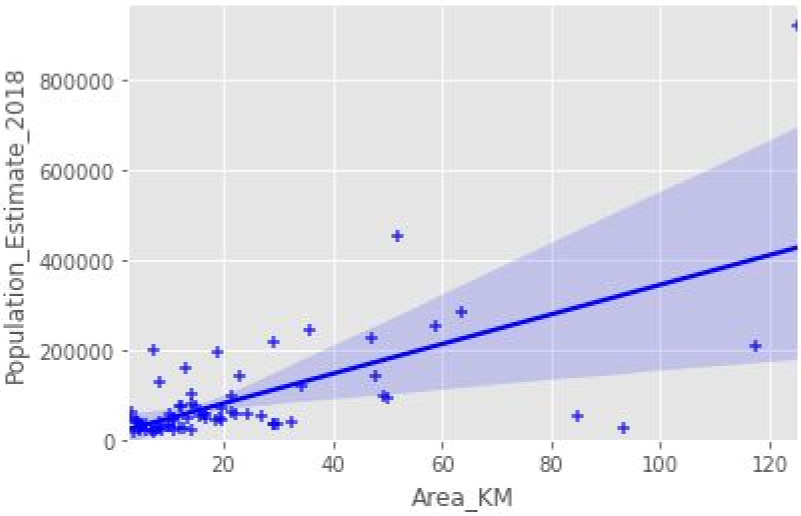
No strong relationships were found between either District or Socio-Economic index to other variables in the dataset (Both variables used to measure Socio-Economic city status were congruent)

A correlation coefficient of 0.25 was found between Population per city and Density per KM.

A correlation coefficient of -0.23 was found between Density in KM and Area in KM

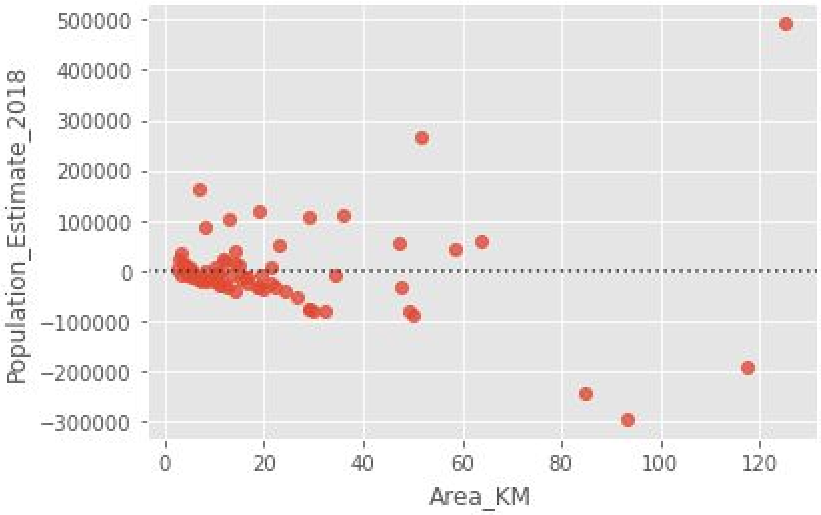
A correlation coefficient of 0.65 was found between Population per city and city area in KM, this is the strongest relationship found in the dataset, and therefore it will be explored further.

A linear regression model was built with Area in KM as a predictor and Population size as the predicted variable.



The intercept for this model is: 14545.990091
The coefficient for this model is: 3295.07170445
The R^2 score for this model is: 0.4228627567446652
The RMSE for this model is 95741.34377894115

A residuals plot was created to approve that the linear model was adequate.



Density was checked to examine if it might be an intervening variable in the model, on its own, it was found to be a poor predictor (R^2 = 0.06398363130012996, RMSE =

121927.50591327358). When a Multiple linear model was constructed using both variables as predictors, an enhanced model was found:

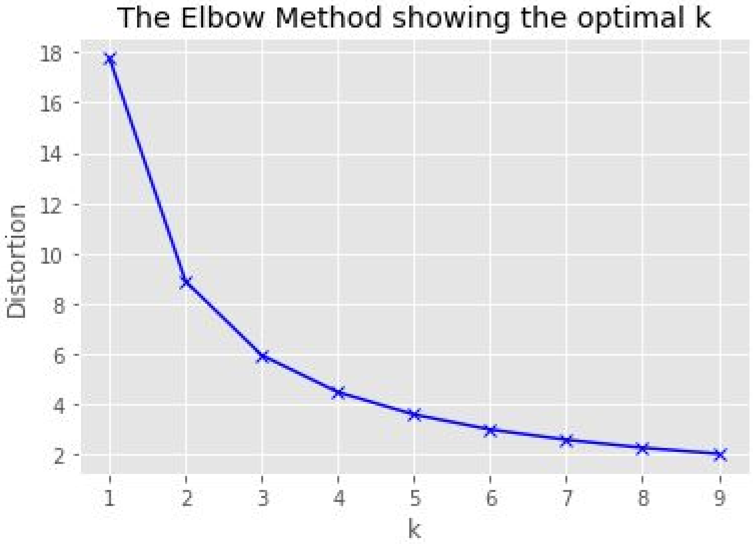
The intercept for this model is: -56393.05183989
The coefficients for this model are:3784.67224659, 12.64953486
The R^2 score for this model is: 0.5927924953023633
The RMSE for this model is: 80420.70112333835

Further model evaluation was not performed due to the relatively small sample size.

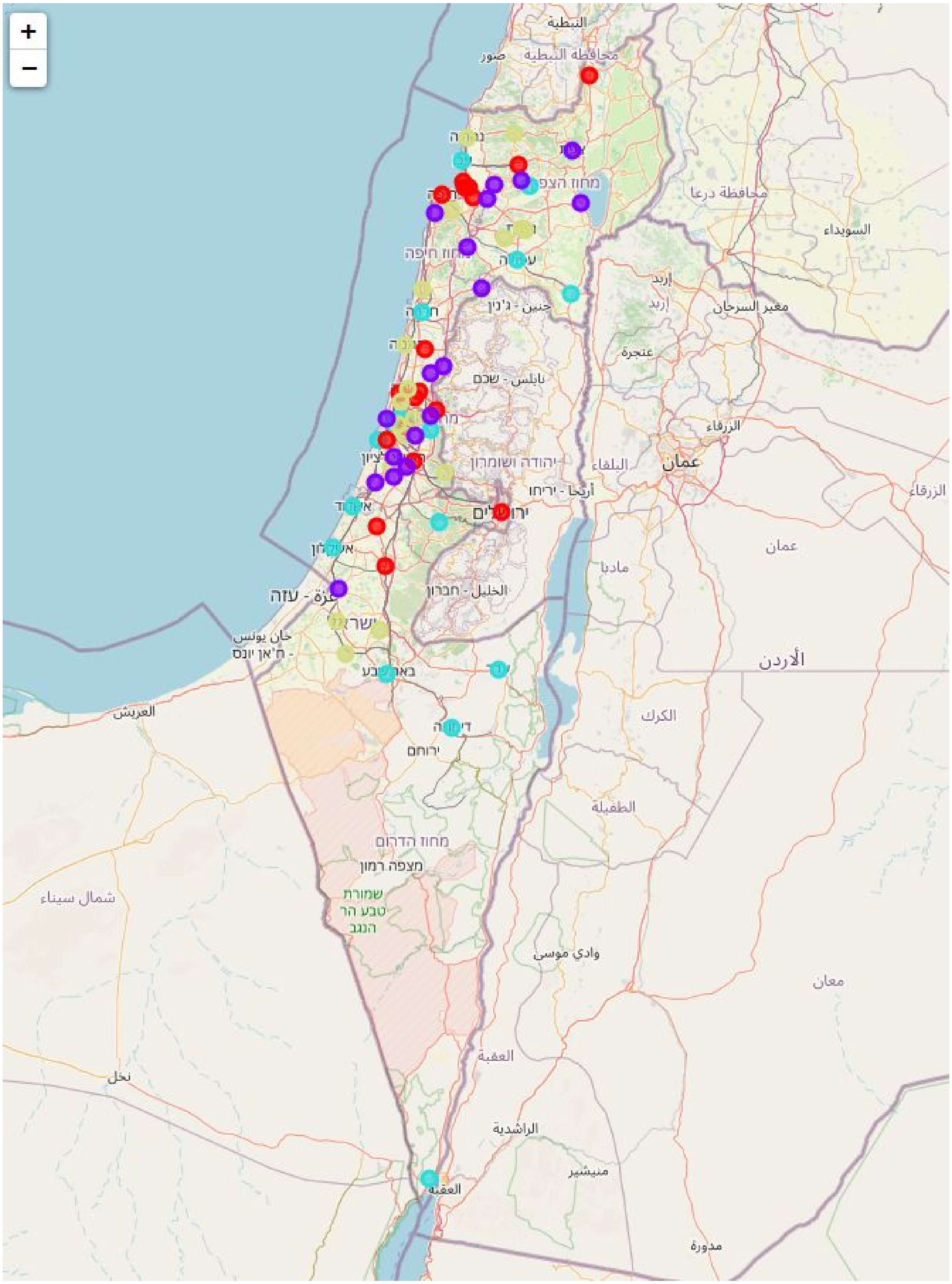
- Clustering

A k-Means Clustering analysis was conducted on the “one-hot encoding” venue category aspects of the dataset, in order to segment the cities into groups with distinct preferable venues.

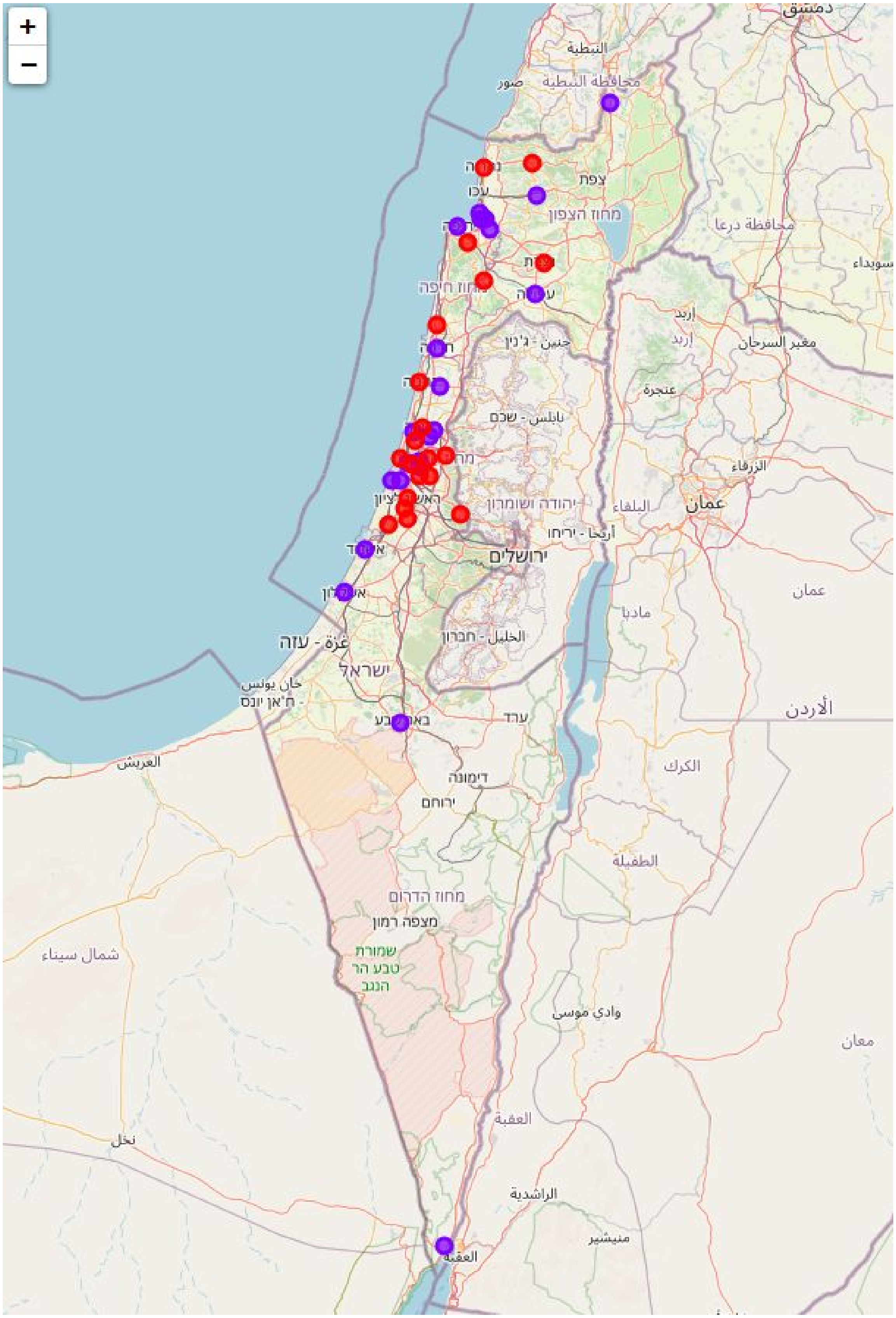
The elbow method to determine optimal k - cluster groups was used:



4 clusters were selected to maintain a reasonable amount of clusters for the amount of cities we have in the data.
The 4 clusters illustrated:


+
−
النبطية
محافظة النبطية
صور
محافظة درعا
السويداء
מחוז חיפה
إربد
مغير السرحان
جنين - ג'נין
عنجرة
نابلس - שכם
الزرقاء
יהודה ושומרון
البلقاء
عمان
أريحا - יריחו
عمان
مادبا
الخليل - חברון
خان يونس
ח'אן יונס -
الأردن
العريش
الكرك
ירוחם
الطفيلة
מחוז הדרום
מצפה רמון
شمال سيناء
שמורת
טבע הר
הנגב
وادي موسى
معان
العقبة
نخل
الراشدية
منيشير
العقبة
مدورة
الزرقاء

A second k-Means Clustering analysis was conducted, this time including only a subset of the data containing the “mid” and “high” socio-economically ranked cities in Israel. For this analysis, 2 clusters were selected due to the relatively low sample size.The 2 clusters illustrated:


+
−
دمشق
النبطية
محافظة النبطية
صور
צפת
עכו
מחוז הצפון
محافظة درعا
سويداء
מחוז חיפה
إربد
إربد
جنين - ג'נין
مغير السرحان
عنجرة
نابلس - שכם
الزرقاء
יהודה ושומרון
البلقاء
عمان
أريحا - יריחו
ירושלים
عمان
مادبا
غزة - עזה
الخليل - חברון
ישראל
خان يونس
ח'אן יונס -
الأردن
ערד
العريش
الكرك
דימונה
ירוחם
الطفيلة
מחוז הדרום
מצפה רמון
شمال سيناء
שמורת
טבע הר
הנגב
وادي موسى
معان
العقبة
الراشدية
نخل
منيشير
العقبة

E. Results and Discussion

When exploring the properties of the dataset, a trend was found towards venues being more densely located in cities that are characterized by a high socio-economic index. it is uncertain whether this trend is represented in reality, or if it may be an inherent bias of the Foursquare API used to obtain the venue information. The high end of the Socioeconomic index cities (the index: 8, for this dataset), included 655 venues out of the total 1437 relevant venues returned from the API, 45.6% of all venues.

After exploring the venue frequency per socio-economic categorization, the distribution of venues by area was examined; Venues were found to be more frequent in the geographical center of Israel, and specifically in the Tel-Aviv area (69% combined for both districts). When compared to population numbers for those districts, it can be seen that venues are over-represented (48.6% of the population for both districts), this might be explained by the Tel-Aviv metropolitan area being a center of Israeli economics and leisure, drawing people from other districts of Israel on a daily basis.

The clearest under-representation of venues can be seen in the Jerusalem district (3.34% of venues compared to 16.2% of the population), this finding is likely a result of the Foursquare API's bias. Based on this disparity, conclusions regarding preferable location according to venue frequency should be made with caution.

A correlation matrix was run to further understand the relationships between relevant numeric & categorical variables in the data.

No strong relationships were found between either District or Socio-Economic index to other variables in the dataset (Both variables used to measure Socio-Economic city status were congruent)

A correlation of 0.25 was found between Population per city and Density per KM.

A correlation of -0.23 was found between Density in KM and Area in KM

A correlation of 0.65 was found between Population per city and city area in KM, this is the strongest relationship found in the dataset, and therefore it will be explored further.

A linear regression model was built with Area in KM as a predictor and Population size as the predicted variable.

The intercept for this model is: 14545.990091
The coefficient for this model is: 3295.07170445
The R^2 score for this model is: 0.4228627567446652
The RMSE for this model is 95741.34377894115

A residuals plot was created to approve that the linear model was adequate.

Density was checked to examine if it might be an intervening variable in the model, on its own, it was found to be a poor predictor (R^2 = 0.06398363130012996, RMSE = 121927.50591327358). When a Multiple linear model was constructed using both variables as predictors, an enhanced model was found:

The intercept for this model is: -56393.05183989
The coefficients for this model are:3784.67224659, 12.64953486
The R^2 score for this model is: 0.5927924953023633
The RMSE for this model is: 80420.70112333835

- Cluster Results

A K-means clustering model was constructed on the basis of a dataset containing the frequency of venue types in every given city.

The ideal number of clusters was assessed using the elbow method, and a 4 cluster model was chosen.

The clusters can be characterized as such:

Cluster 1

"Free Urbanites" : High population cities, high socio-economic index & relatively less dense. Preferred venus are:

1 Café 0.29

2 Middle Eastern Restaurant 0.09

3 Sushi Restaurant 0.06

4 Italian Restaurant 0.06

5 Bakery 0.05

6 Burger Joint 0.05

7 Fast Food Restaurant 0.05

8 Restaurant 0.05

9 Asian Restaurant 0.04

10 Coffee Shop 0.04

Cluster 2

"Well-To-Do & Satisfied" : Mid-High Socio-economic index, low density & live in small cities with small population size. Preferred venus are:

1 Middle Eastern Restaurant 0.24

2 Café 0.16

3 American Restaurant 0.06

4 Restaurant 0.05

5 Mediterranean Restaurant 0.05

6 Bakery 0.04

7 Ice Cream Shop 0.04

8 Sushi Restaurant 0.03

9 Falafel Restaurant 0.03

10 Coffee Shop 0.03

Cluster 3

"Comfy Middle-Class" : Small to medium population size, small in area but not dense, Socio economically index low to medium. Preferred venus are:

1 Café 0.31

2 Middle Eastern Restaurant 0.11

3 Coffee Shop 0.07

4 Breakfast Spot 0.05

5 BBQ Joint 0.05

6 Restaurant 0.04

7 Burger Joint 0.04

8 Asian Restaurant 0.03

9 Seafood Restaurant 0.03

10 Sushi Restaurant 0.03

Cluster 4

"Urban Warriors" : Very high density in very large cities, Socio-economically low. Preferred venus are:

1 Café 0.33

2 Middle Eastern Restaurant 0.08

3 Italian Restaurant 0.06

4 Pizza Place 0.05

5 Burger Joint 0.05

6 Restaurant 0.04

7 Ice Cream Shop 0.04

8 Falafel Restaurant 0.04

9 Asian Restaurant 0.04

10 Bakery 0.03

Due to the tendency of venues to be located in Socio-economically high cities in greater concentration, a second k-means clustering analysis was performed to attempt a greater focus of this subset of the cities in the data.

For this analysis, 2 clusters were selected due to the relatively low sample size.

The clusters can be characterized as such:

Cluster 1

"Center-Stage" : High population counts with medium density in medium-sized cities, very high socio-economic index. Prefered venus are:

1 Café 0.29

2 Middle Eastern Restaurant 0.10

3 Italian Restaurant 0.06

4 Bakery 0.06

5 Burger Joint 0.04

6 Asian Restaurant 0.04

7 Coffee Shop 0.04

8 Ice Cream Shop 0.04

9 Restaurant 0.04

10 Sushi Restaurant 0.04

Cluster 2

"Closing on the Spotlight": Smaller population counts and higher density cities, high socio-economic index. Preferred venus are:

1 Café 0.28

2 Middle Eastern Restaurant 0.06

3 Italian Restaurant 0.06

4 Burger Joint 0.06

5 Asian Restaurant 0.05

6 Bakery 0.05

7 Coffee Shop 0.05

8 Restaurant 0.05

9 Sushi Restaurant 0.04

10 Fast Food Restaurant 0.04

F. Conclusion & Recommendations

To begin with, our first meaningful finding was that relevant venues tend to more frequently appear at the Center & Tel-Aviv Districts of Israel, and in cities with higher socio-economic index. On the basis of this finding, we can recommend to stakeholders that a greater demand exists in these areas - and the likelihood of financial success might increase if the choice is made to open a restaurant in those areas of Israel.

This recommendation must be limited though, as it might be derived from the data's bias rather than a reliable depiction of venue location. What we can say with a higher likelihood is that the presented findings can be interpreted as more reliable for these areas that have a higher frequency of venues.

The clearest and most distinct finding that can be derived from the clustering analysis is the overwhelming popularity of Cafe venues, throughout almost all clusters - Cafes are the most common venue. We can easily recommend that the stakeholder open a Cafe, regardless of area, under the assumption that frequency is an indication of demand. The high amount might also indicate a saturated market for these specific venues, competition might be particularly harsh and larger, more veteran companies might be harder to compete with.

It should also be mentioned that no clear geographic area preferences for venues were found - as each cluster segment clearly contains cities from across Israel.

- Recommendations per Cluster

Our first cluster analysis can help define further recommendations:

A clear outlier among clusters are the "Well-To-Do & Satisfied" group, this group is the only one not dominated by Cafes, with a higher frequency of Middle Eastern restaurants. If the stakeholders are interested in opening a venue in these locations, a Middle Eastern restaurant clearly seems to have great demand. If the stakeholder is more interested in a niche among these cities, American restaurants seem to have distinct popularity here.

All other clusters, as mentioned, are dominated by Cafes, with Middle Eastern restaurants being consistent runner-ups. For each of these we can still make niche market recommendations:

For the "Free Urbanites" - Sushi restaurants are in higher demand than in any other segment.

For the "Comfy Middle-Class" - Breakfast spots & BBQ joints seem to have high popularity.

For the "Urban Warriors" - Italian restaurants, Pizza places & Burger joints are preferred

Our second clustering analysis, focusing on mid to high index socio-economical cities - also has some distinctions to provide:

The "Center-Stage" cluster has more demand for Bakeries, while the "Closing on the Spotlight" cluster has a higher demand for Burger joints.

- Limitations

All results and recommendations throughout this project are dependent on the collected data's reliability in reflecting reality - as was previously established, the Foursquare API used to gather venues for this project seems to have a bias towards venues in the Center & Tel-Aviv Districts of Israel. Also, the collected information from Wikipedia regarding Israeli cities is not all fully updated - Population counts were estimated only, and for 2018. Other aspects of the data might have also changed somewhat for a portion of the cities (Most likely increases in density and in some cases increases city size)

References:

[1] - "Israelis spend 20 billion a year on Cafes and restaurants"
[2] - "How many restaurants really closed this year? and how many opened?"
[3] - Restaurant Location and Price Fairness as Key Determinants of Brand Equity: A Study on Fast Food Restaurant Industry
[4] - Why Restaurants Fail? Part IV
[5] - Factors Affecting Consumers' Eating-Out Choices in India: Implications for the Restaurant Industry
[6] - Consumers' Restaurant Choice Behavior and the Impact of Socio-Economic and Demographic Factors
[7] - "Population in the Localities 2018"
[8] - "Profiles by Locality"

[9] - "2004 local government profile"
[10] - Population and Density per Sq. Km. in Localities Numbering 5,000 Residents and More on 31.12.2016
[11] - SOCIO-ECONOMIC INDEX 2013 OF LOCAL AUTHORITIES
[12] - OpenCage Geocoder
[13] - List of cities in Israel
[14] - https://foursquare.com