IBM Applied Data Science Capstone Project

Restaurant Business Analysis in Toronto

Introduction

According to restaurantscanada.org, 2020 would witness commercial foodservice sales improvement by 4%. Alberta and Ontario will lead the way with 4.4% and 4.2% growth, respectively. By 2021, foodservice sales are forecasted to go over 100-billion dollar.

The restaurant industry represents 4% of Canada GDP, with around 85 Billon dollars in annual sales that are generated by the restaurant industry, it employs 1.2 Million people not less than 7% of the country total workforce. These figures could seem a bit staggering but when we put them in the context of Canadians making, daily, 22 Million visits to restaurants, it may completely change our perception of the restaurant business outlook.

Toronto, being the capital of Ontario, and having by far the biggest population and strongest economy within the province. Being a cosmopolitan metropole, it fosters diverse culinary backgrounds when it comes to food businesses, catering and, restaurants and attracting a huge number of immigrants makes its restaurant business a blooming one.

Business Problem

This project objective is to come up with an analysis of the restaurant business in Toronto based on several features or characteristics such as population levels, income ranges, restaurant business categories, geolocation within the city and so on. In other words, data science methodology will allow to answer the following question: Within the city of Toronto, where would be the best location to invest in a restaurant business and also for a given area within the city, what would best category of restaurant to invest on.

Target Audience

The target audience of this project are potential restaurant business investors or restaurant managers who need to conduct a market research in order to get specific insights and be able to determine the business arguments to invest in a given category of food establishment within a specific area with a unique set of characteristics. This project being specific to the city of Toronto aims to provide a more targeted view than reports or studies related to Canada and would help anyone interested in investing in this sector within Toronto.

Data acquisition

1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M: this Wikipedia page provides the exhaustive list of Toronto postal codes or FSA, with their corresponding boroughs and neighborhoods.

1. “https://cocl.us/Geospatial_data” this CSV file provides the longitude and latitudes of the Toronto FSA or postal codes.

2. https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm - This repository from Statistics Canada, provides a boundary file representing the FSAs of Canada, this file is in Shape format. A conversion is to be conducted to transform this file to GeoJSON format using QGIS software - please note that a reduction of the file has been conducted to keep only Toronto FSA with the GeoJSON file.

3. https://www.canada.ca/en/revenue-agency/programs/about-canada-revenue-agency-cra/income-statistics-gst-hst-statistics/individual-tax-statistics-fsa/individual-tax-statistics-fsa-2017-edition-2015-tax-year.html#t this page of the Government of Canada, provides the Individual Tax Statistics by Forward Sortation Area (FSA) − 2017 Edition, this CSV dataset will provide us with income inputs, this will be one of the features of our analysis and machine learning algorithm.

4. https://www12.statcan.gc.ca/ Statistics Canada provides a CSV file of the 2016 population census; the population level feature will be assessed to see whether it has an incidence on the restaurant business in Toronto.

5. https://developer.foursquare.com/docs/api/endpoints: Restaurant businesses category related data is retrieved via the Foursquare API; this category is identified within Foursquare database with the following ID 4d4b7105d754a06374d81259 (https://developer.foursquare.com/docs). An HTTP request will be sent to the API as follows: GET https://api.foursquare.com/v2/venues/search, it will return a list of venues, matching our category ID, near the specified location (providing its latitude and longitude).

Methodology

As we are going to analyze Toronto and assess areas where restaurant business could thrive, we would need first to get a complete list of the different neighborhoods of Toronto; we will perform a web scraping using Beautifulsoup Library of the Wikipedia page listing the postal codes of Canada, after some data cleanup, we will obtain a data frame containing the costal code, borough name and its corresponding neighborhoods.

Second step would be to add the geo tagging of these neighborhoods, we could get this information from various Map content providers API, but we will retrieve the information from the CSV file at the URL location : https://cocl.us/Geospatial_data/Geospatial_Coordinates.csv; The geo tagging is obtained via an inner merge between the Neighborhoods dataframe and the Geo dataframe, resulting in a combined Toronto dataframe .

Then we will enrich the data with the 2016 Census population data, for that purpose we will parse a CSV file obtained from the Canada Statistics website and saved locally. We will drop unneeded information for our study, such as the 'Incompletely enumerated Indian reserves and Indian settlements, 2016' and 'the private dwellings' related information. This data covers the whole country of Canada and we would need to select only the Toronto items, but we would opt for a simple left join between the combined Toronto dataframe and the population dataframe (2 birds, 1 stone), we obviously would need to perform a sanity check on the resulting dataset.

Using Folium, we can then perform a data visualization to check the consistency of the progress made so far, using the constructed Toronto GeoJSON file and our Toronto Combined Dataframe and create a map depicting the Toronto Population Levels by FSA (Forward Sortation Areas)

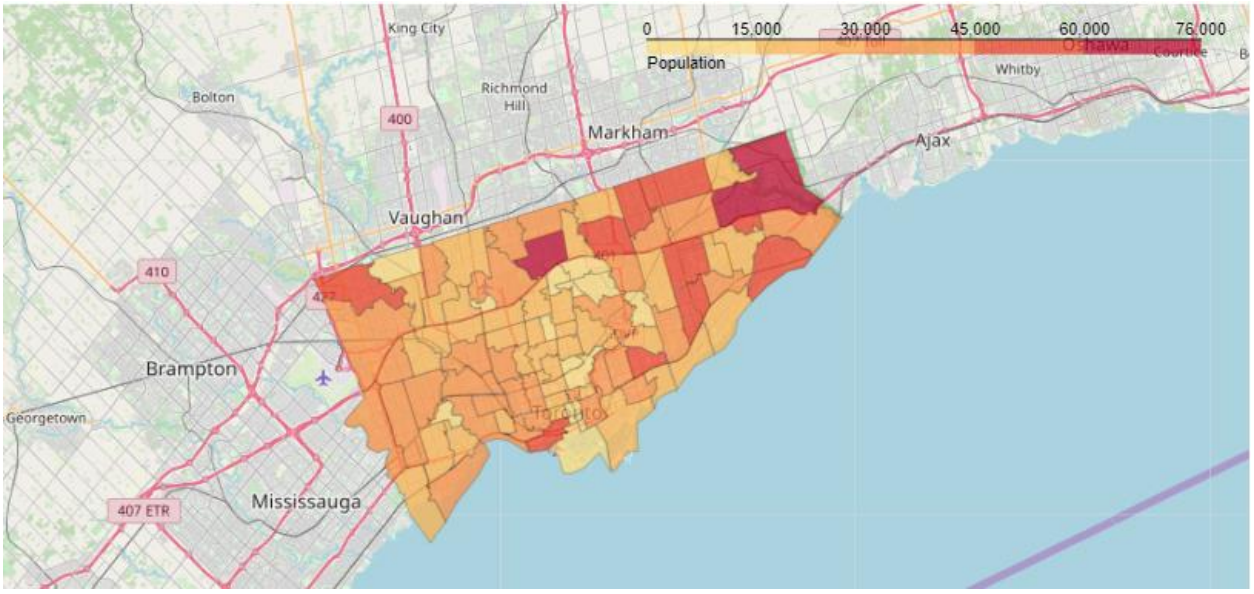

Figure: Toronto Population levels

Our next step would be to enrich our dataset with annual income information from Canada Statistics, using a CSV file; we then perform a data cleanup by dropping again unnecessary data and also by reordering the data columns to have a consistent data that we will merge with the Toronto Combined dataframe via a left join operation. After that we can perform another round of data check and cleanup. Lastly, we will visualize our income data by plotting the annual income per FSA to check for any potential inconsistency (we could find an improper float conversion through this step)

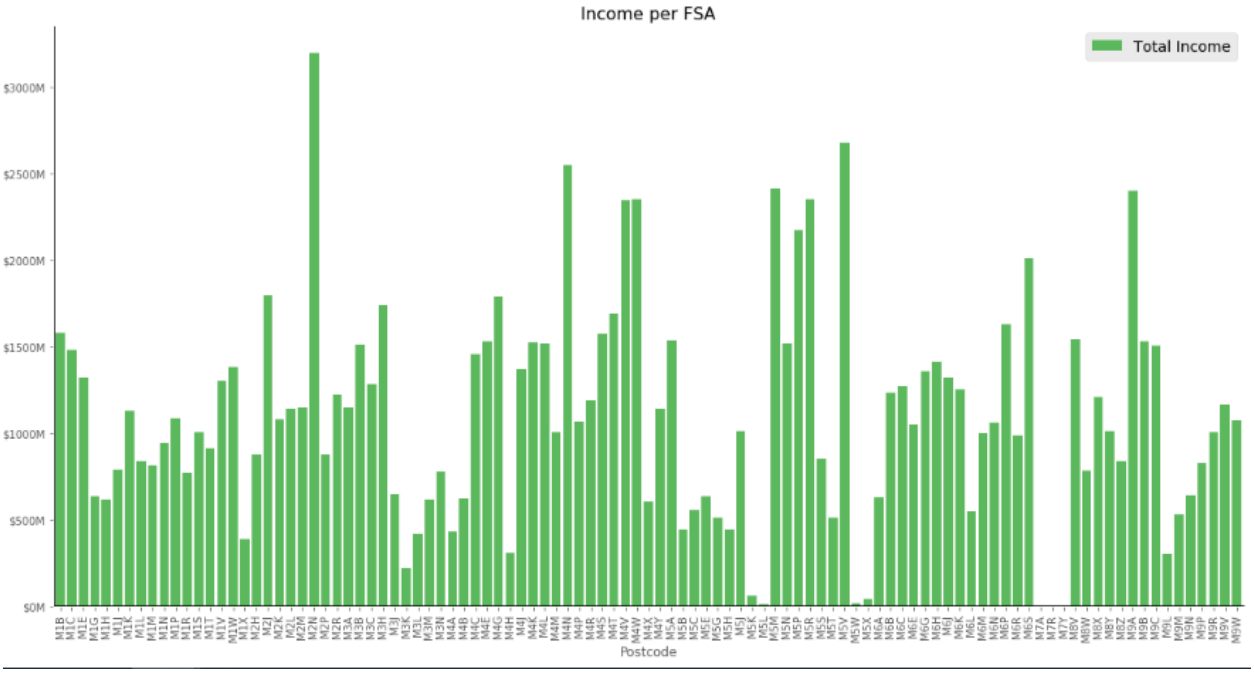

Figure: Income per FSA

We then proceed with retrieving all the restaurant business occurrences within each neighborhood of Toronto, for that, we will use Foursquare API, via an HTTP GET search request provided with the latitude and longitude of the neighborhood and the food category ID: 4d4b7105d754a06374d81259. We will iterate through the neighborhoods within the Toronto Combined dataframe to get all the restaurant businesses of the city. Foursquare is returning a JSON file, that needs to be parsed through an iterative process which outcome will be a dataframe containing each business's name, its latitude & longitude and its category (Coffee shop, fast food, restaurant, Italian restaurant…)

	Postcode	Borough	Neighbourhood	Catering Name	Latitude	Longitude	Category
0	M1B	Scarborough	Malvern / Rouge	Meena's Fine Foods	43.804476	-79.199753	Indian Restaurant
1	M1B	Scarborough	Malvern / Rouge	Wendy's	43.807448	-79.199056	Fast Food Restaurant
2	M1B	Scarborough	Malvern / Rouge	Second Cup	43.802165	-79.196114	Coffee Shop
3	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	Shamrock Burgers	43.783823	-79.168406	Burger Joint
4	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	Amigo's	43.783749	-79.168691	Breakfast Spot

Figure: Dataframe resulting after Foursquare JSON parsing

We will then enrich each business data with its geographical distance from the city center, we will consider Toronto City Hall for that exercise, the aim of this enrichment is to assess whether closeness to the city center have a positive or negative correlation on the number of restaurant businesses. Let us use the Haversine method to compute the geographical distance between two points on earth (each restaurant and Toronto City Hall), this method considers earth as spherical but with some flattening around the poles, in our specific context, the distances between our different points are not so large so we expect to obtain a fair accuracy for the distances calculation. (https://en.wikipedia.org/wiki/Geographical_distance).

```
def haversine_distance(lat1,lon1,lat2,lon2):

    radius = 6371 # earth's radius in kms

    dlat = math.radians(lat2-lat1)
    dlon = math.radians(lon2-lon1)
    a = math.sin(dlat/2) * math.sin(dlat/2) + math.cos(math.radians(lat1)) \
        * math.cos(math.radians(lat2)) * math.sin(dlon/2) * math.sin(dlon/2)
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1-a))
    d = radius * c

    return(d)
```

Figure: The Haversine Function

The coordinates of Toronto City Hall are based on www.latlong.net website. After this operation completion we can insert the geographical distance data into our catering dataframe.

As we have seen the category of restaurant business is a categorical information, and that would not be a valuable information for machine learning algorithms, we would need to convert this data from categorical to numerical via the hot one encoding technique, using Pandas getdummies () method. We then perform a Pandas groupby operation on Neighborhood to sum the business category types.

	Neighbourhood	Afghan Restaurant	African Restaurant	American Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	...	Tapas Restaurant	Tea Room	Thai Restaurant
0	Agincourt	0	0	0	0	0	0	0	0	0	...	0	0	0
1	Alderwood / Long Branch	0	0	0	0	2	0	2	0	0	...	0	0	1
2	Bathurst Manor / Wilson Heights / Downsview North	0	0	0	0	0	0	0	0	0	...	0	0	0
3	Bayview Village	0	0	0	0	2	0	0	0	0	...	0	0	0
4	Bedford Park / Lawrence Manor East	0	0	2	0	1	0	0	0	1	...	0	0	1

Figure: One-Encoding and Groupby resulting dataframe

Last step in the data preparation would be to merge this newly obtained data with the Toronto Combined dataframe and perform a data cleanup

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Population, 2016	Total Income	Total	Under $5,000	000to 9,999	...	Tapas Restaurant	Tea Room	Thai Restaurant	The Restau
0	M1B	Scarborough	Malvern / Rouge	43.806686	-79.194353	66108.0	1.577233e+09	51410	8140	4340	...	0.0	0.0	0.0	
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	43.784535	-79.160497	35626.0	1.483624e+09	29080	3340	1780	...	0.0	0.0	0.0	
2	M1E	Scarborough	Guildwood / Morningside / West Hill	43.763573	-79.188711	46943.0	1.320927e+09	36220	4810	2970	...	0.0	0.0	0.0	

Figure: Toronto Combined Dataframe Snaphot

Our next step would be to follow a top down approach in analyzing our dataset and determine if there are any correlation between the different features (income, population, category types…) within our dataset. We will draw a global heatmap using the Seaborn library and start form there.

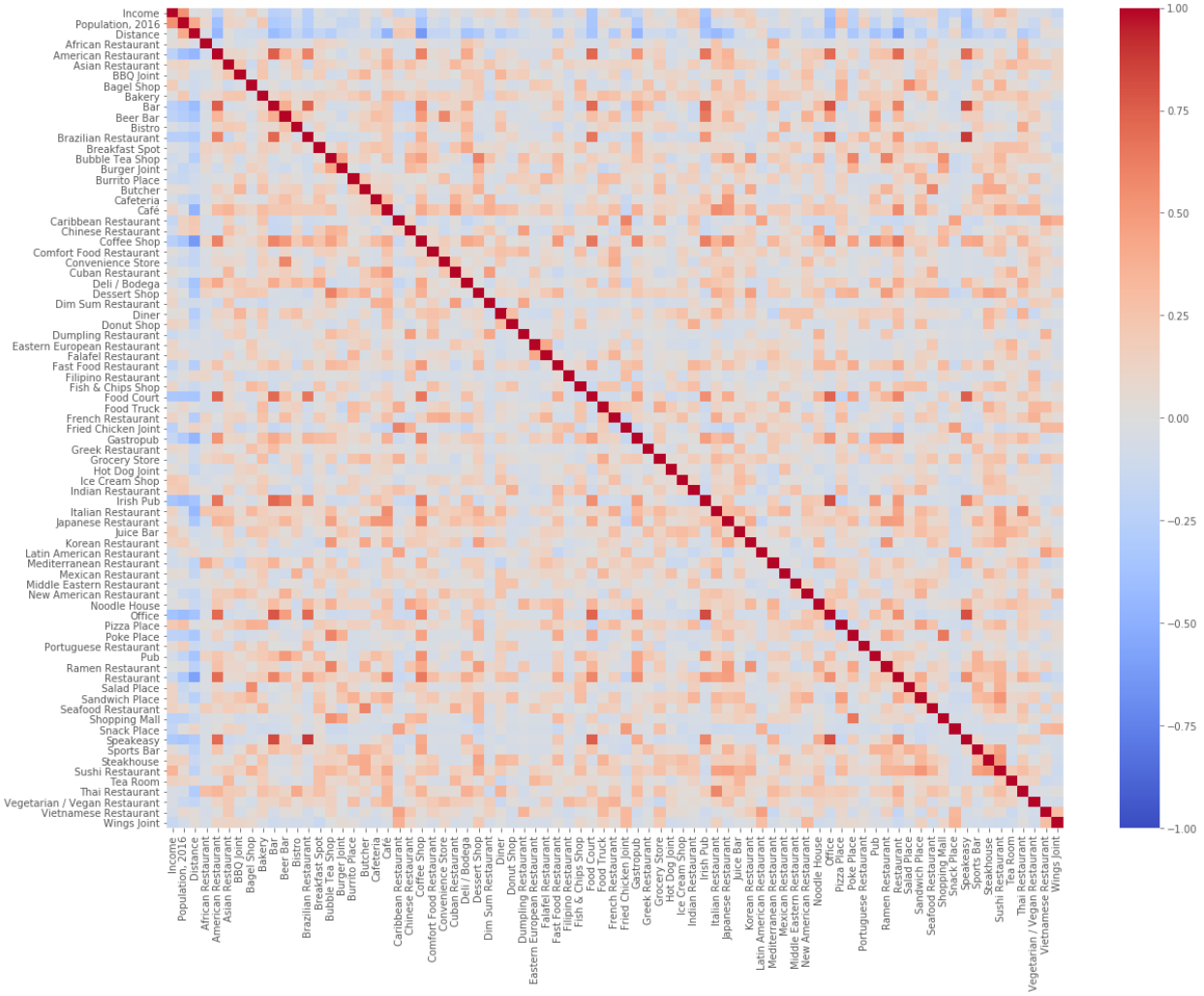

Figure: Global Data Heatmap

Clearly, through the red spots can we say that there are some correlations between some kind of catering businesses, also the blue left and top edges show vague some signs of negative correlation between either income or population levels or even distance from Toronto city center. let's explore further down in details these correlation relationships to get further insights.

Let’s analyze then as a first step the relationship between restaurant counts with distance from Toronto city hall, income and population, we will draw a heatmap map for these features alone.

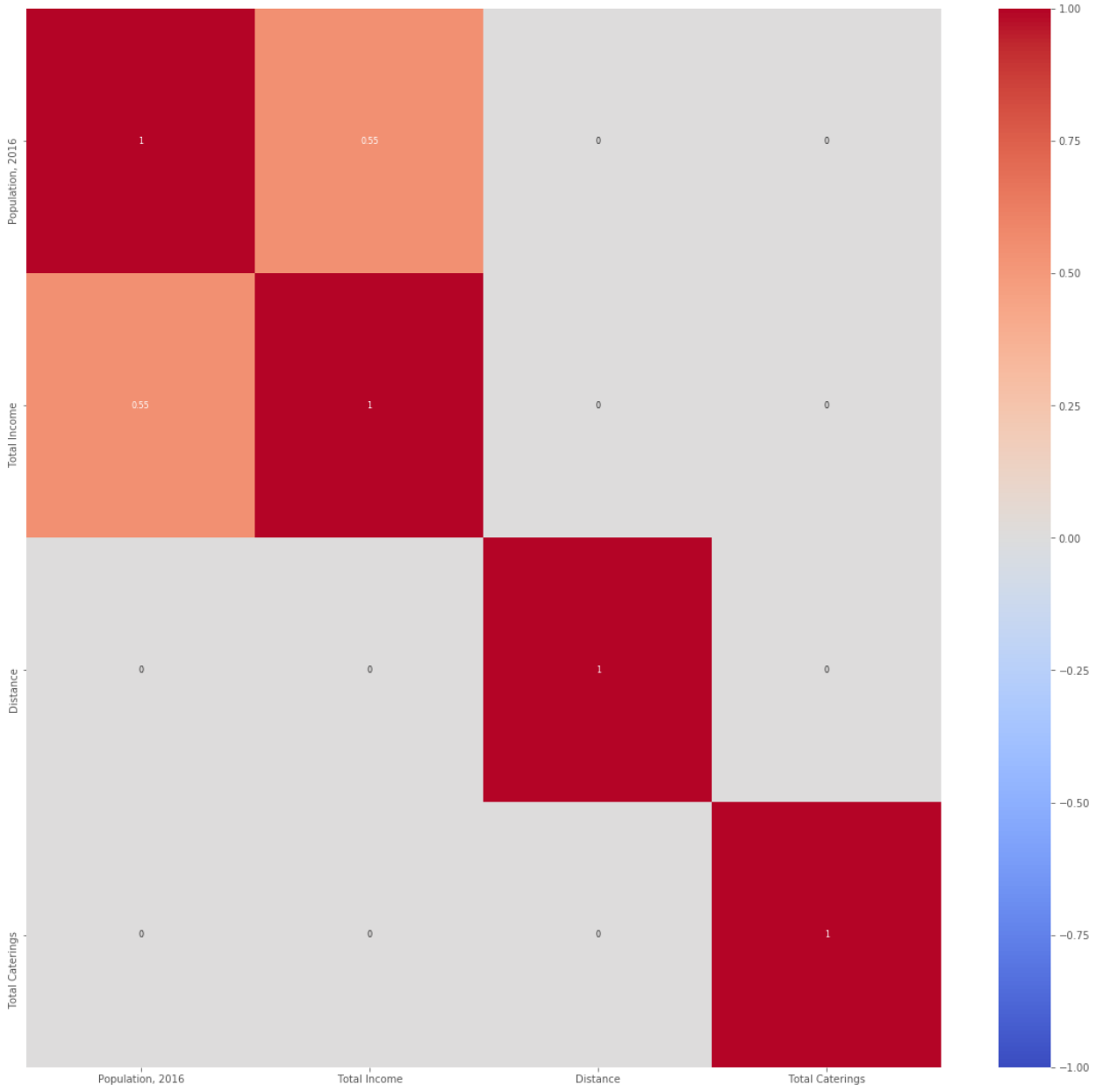

Figure: Heatmap of business count, distance, income, population

There is no palpable correlation between the number of restaurant business and the population and income levels, yet we can see that there is slight positive correlation between population and income levels with a Pearson correlation factor of 0.55. An important note to consider is that correlation does not imply causation meaning that we could have cases of FSA with low population figures characterized with high level income.

Secondly, let us explore the potential correlation relationships between the different restaurant business types; as we have seen from the overall heatmap that the correlation relationships are not so “talkative”, as it looks from the figure that there are several restaurant types with strong intra correlations, still are these business relevant quantitatively to be meaningful? We will follow this intuition and perform some filtering and keep only the catering businesses with relevant number of occurencies across the Toronto FSAs. Let us plot the data to see if the intuition makes sense

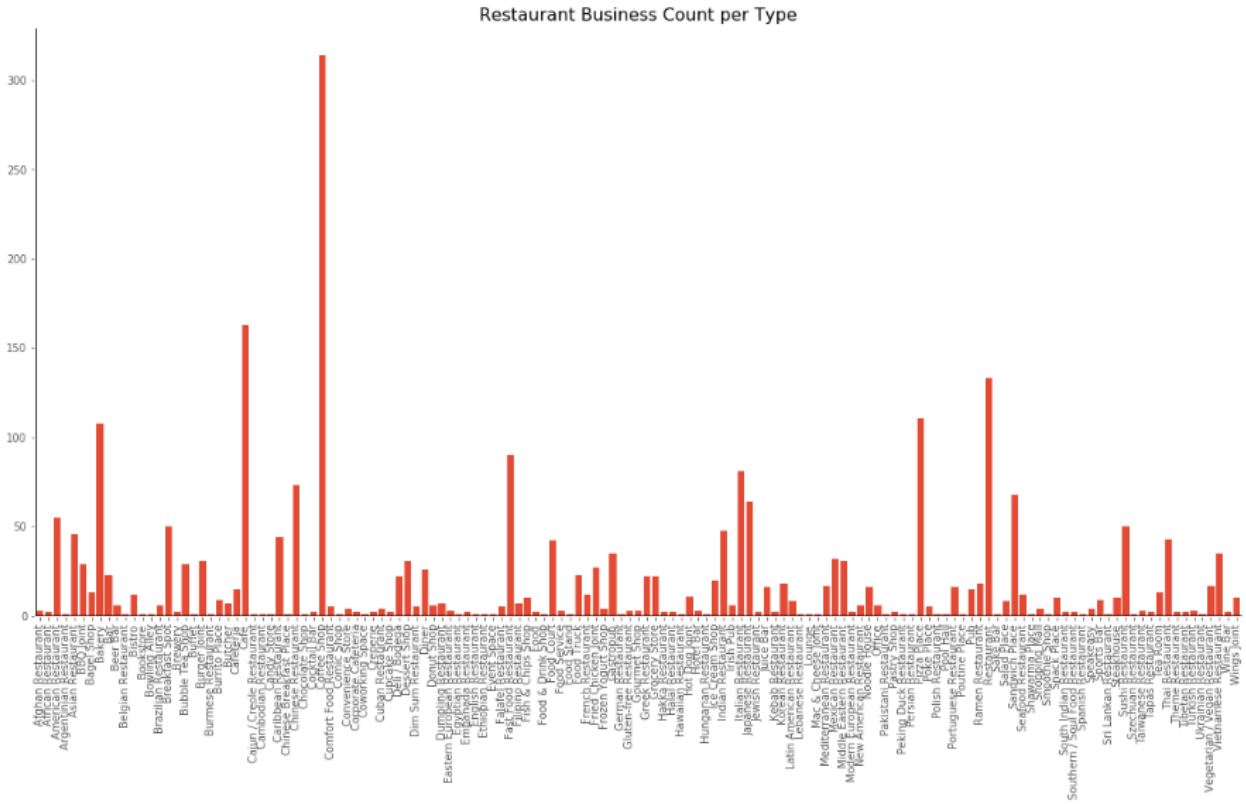

Figure: Restaurant Business Levels per Category Type

From the figure, we notice that there are many category types that are not so well represented quantitively within the dataset we obtained from the Foursquare API, let us analyze further this distribution by running the Pandas describe() method.

```
plot_data.describe().loc["75%"].max()
```

```
4.0
```

Figure: Describe method output snapshot

From the describe() method figure, we see that the 75th percentile of the restaurant have a count of 4; Our filtering condition is then set, let's keep only the catering business types with at least a count of 4. Our restaurant business dataset is filtered accordingly, and we will plot another heatmap of the result.

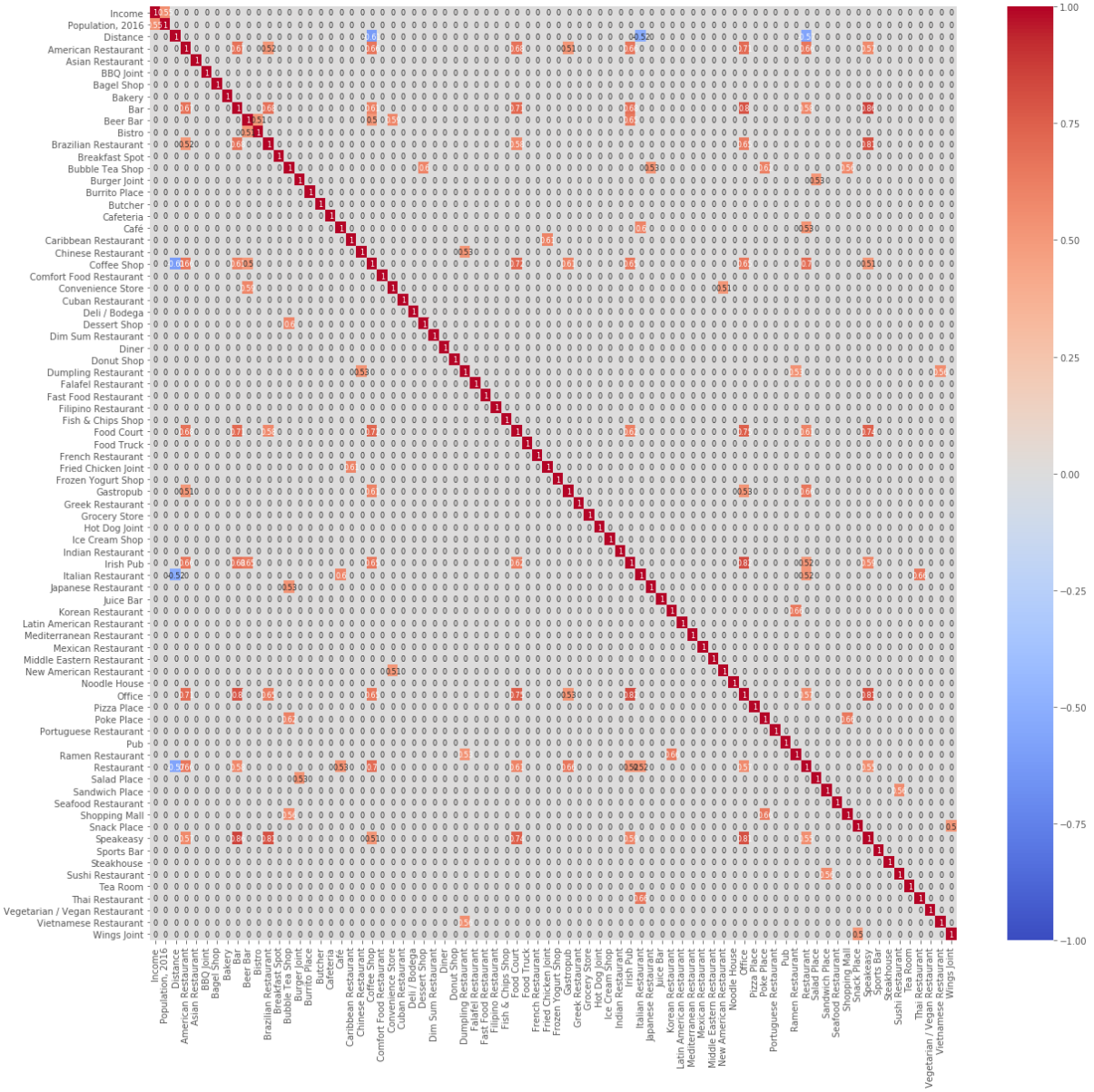

Figure: “Filtered” Dataset Heatmap

From the above figure, we have a clear view , and we get the confirmation that income and population levels have no correlation with any type of catering businesses, however we see that there is a negative correlation between the distance to the city center and few catering business types, namely coffees shops and general type of restaurants and Italian restaurants. so, the closer we are from Toronto City Halls, the more important would be the occurrences of coffees shops and restaurants.

Let us focus now on correlations between categories of restaurant businesses, in order to keep only consistent inter categories correlation relationships we will consider correlation coefficient that are quite significant meaning with a value greater or equal to 0.60 and P-values less or than 0.05 (moderate certainty in the correlation result), If our alpha level is 0.05, there would a 5% chance that we will incorrectly reject the null hypothesis. Or to put it another way, an alpha value of 0.05 indicates that the risk of concluding that a correlation exists, when, actually, no correlation exists is 5%. We will compute then correlation coefficient for each restaurant occurrence and the other restaurant instances within the dataset, and its P-value, then filter out entries with values that are not satisfactory. We come up with below result.

	Restaurant Type 1	Restaurant Type 2	Correlation Coefficient 'r'	p-value
0	Bar	Speakeasy	0.857367	6.783177e-31
1	Irish Pub	Office	0.823024	1.482488e-26
2	Brazilian Restaurant	Speakeasy	0.808207	5.794952e-25
3	Office	Speakeasy	0.808207	5.794952e-25
4	Bar	Office	0.797959	6.101113e-24
5	Food Court	Office	0.750266	7.407324e-20
6	Food Court	Speakeasy	0.744547	1.984420e-19
7	Coffee Shop	Food Court	0.721999	7.592856e-18
8	Bar	Food Court	0.714163	2.480615e-17
9	American Restaurant	Office	0.706459	7.648626e-17
10	Coffee Shop	Restaurant	0.696733	3.012309e-16
11	Bar	Brazilian Restaurant	0.681129	2.432895e-15
12	Bar	Irish Pub	0.681129	2.432895e-15
13	American Restaurant	Food Court	0.675422	5.059713e-15
14	American Restaurant	Bar	0.673231	6.672795e-15
15	American Restaurant	Restaurant	0.661410	2.854766e-14
16	Gastropub	Restaurant	0.660812	3.067260e-14
17	Italian Restaurant	Thai Restaurant	0.659723	3.494459e-14
18	American Restaurant	Irish Pub	0.658713	3.941544e-14
19	American Restaurant	Coffee Shop	0.657045	4.803927e-14
20	Korean Restaurant	Ramen Restaurant	0.656132	5.350387e-14
21	Poke Place	Shopping Mall	0.656073	5.387595e-14
22	Coffee Shop	Irish Pub	0.649657	1.136857e-13
23	Coffee Shop	Office	0.649657	1.136857e-13
24	Beer Bar	Irish Pub	0.646048	1.716993e-13
25	Brazilian Restaurant	Office	0.646048	1.716993e-13
26	Food Court	Irish Pub	0.622040	2.331087e-12
27	Bubble Tea Shop	Poke Place	0.616821	3.991578e-12
28	Food Court	Restaurant	0.609276	8.536802e-12
29	Bar	Coffee Shop	0.607710	9.971349e-12
30	Caribbean Restaurant	Fried Chicken Joint	0.607052	1.064067e-11
31	Coffee Shop	Gastropub	0.606656	1.106449e-11
32	Café	Italian Restaurant	0.601408	1.846833e-11
33	Bubble Tea Shop	Dessert Shop	0.600360	2.043417e-11

Figure: Inter Categories Correlations

The above figure shows many interesting correlations, but it is not yet so talkative, a better visualization would be certainly a heatmap covering all these resulting restaurant business categories. In order to produce such a visualization, we will compute a correlation matrix out of the above results that could be fed as an input to the Seaborn heatmap function. And voila!

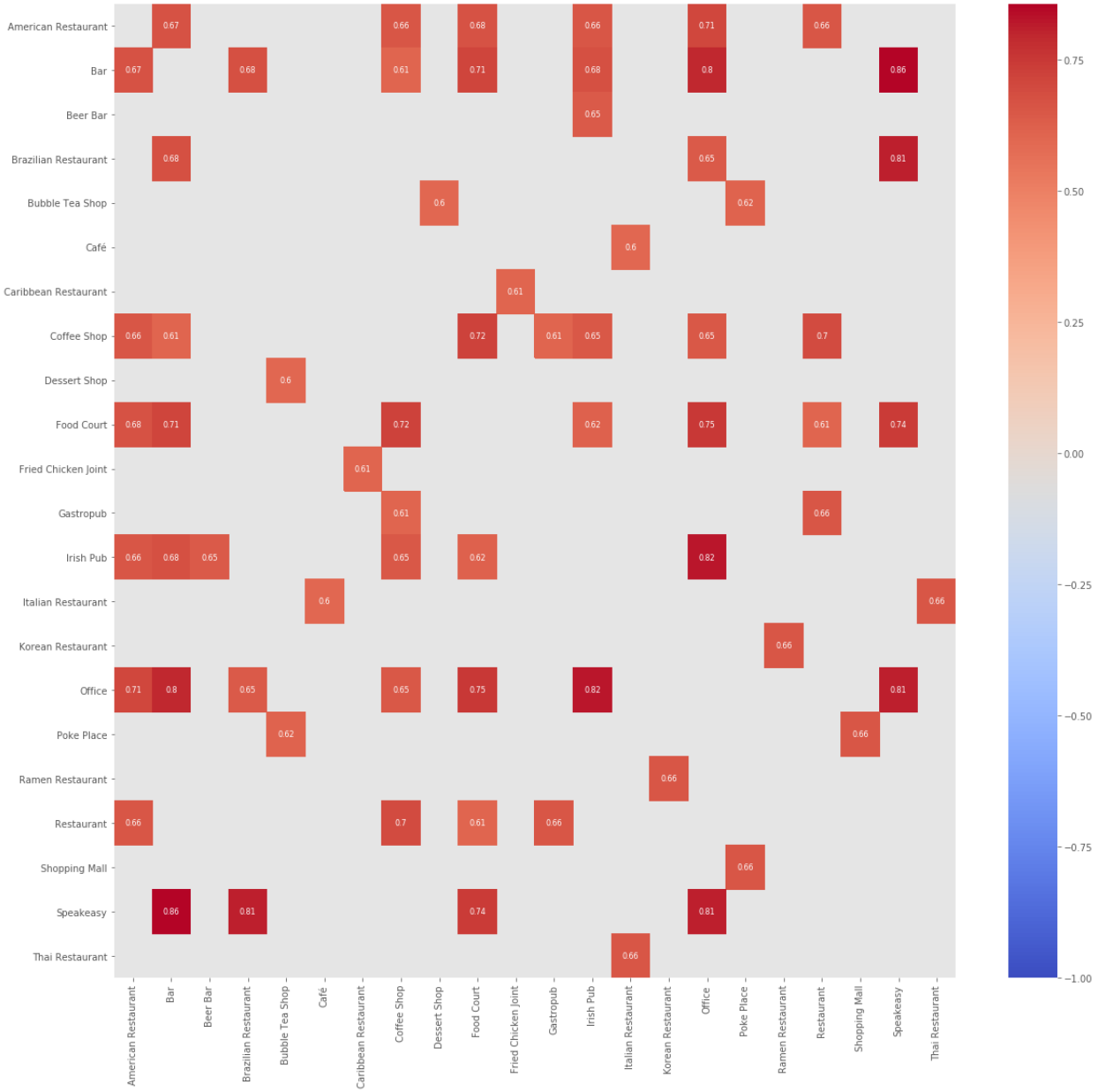

Figure: Strong Inter Categories Correlation Heatmap

Correlation Analysis Results

There are several interesting findings from the above correlation exercise we have made:

- Within Food courts in Toronto, one might say where could I invest, well, the heatmap is clear, good business opportunities are Speakeasy, Restaurants, Office, Coffee Shops, American restaurants.
- Coffee shops (being the number one restaurant business in Toronto) are highly correlated with bars, Gastropubs, Food courts, Irish pubs, office and last but not least restaurants, this last instance could be explained by the fact that consumers may tend to take a coffee after their meal at restaurants, these insights are valuable for an investor willing to invest in an area very well covered by Coffee shops. Also a further insight on this last correlation, the heatmap is also showing a moderate link between the number of Italian restaurants and Café places (not coffee shops, this time), this could be explained by the fact that consumers, after completing their Italian meal, would be to consume maybe an Italian café served in a pleasant and cozy environment, thing that the common coffee shop fails to offer
- Office or catering type of businesses are strongly linked to Bars, Irish Pubs and speakeasy, with correlation coefficients over 0.80, which makes a lot of sense, as catering would be close to offices location, and employees would tend to go to bars and pubs after office hours are completed.

Toronto Neighborhood Clusters Analysis

Let us now analyze the Toronto neighbourhoods and assess if there are similarities, differencies between groups of neighbourhoods based on the features set we have gathered in the PandasToronto combined dataframe.

We will rely on a unsupervised clustering machine learning algorithm, k-means, to accomplish this task; k-means clustering, is a method of vector quantization, originally created from signal processing, its objective is to partition n entities into k clusters in which each entity is related to the cluster with closest mean or basically the centroid of the cluster.

Our first step would be then to determine the optimal k parameter, that would provide us with the most accurate results; we can rely on the elbow method in this regard, which is a heuristic method of determining a consistency within a cluster analysis in order to find the appropriate k parameter, namely the number of clusters we can get out of a given dataset.

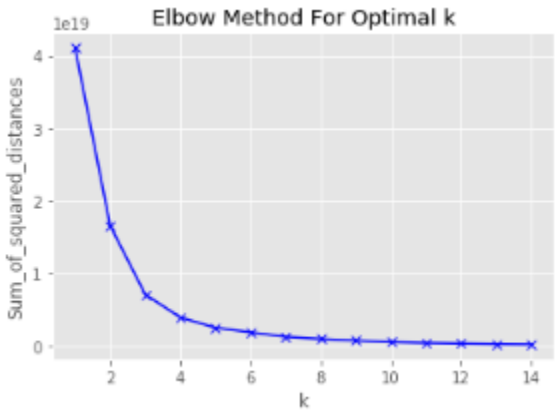

Figure: Elbow Method

It is not straightforward, from the elbow method, to say with confidence that K is equal to 3 or 4, let us complement this method by another approach: The Silhouette method, that is another mean to determine the optimal K parameter or the number of clusters. The silhouette coefficient of a dataset measures how well close a given data is towards its own cluster and how far it is from other clusters. A silhouette coefficient close to 1 represent data points that are labeled under an appropriate cluster, whereas a coefficient close to −1 shows data points that are wrongly labeled implies under clusters.

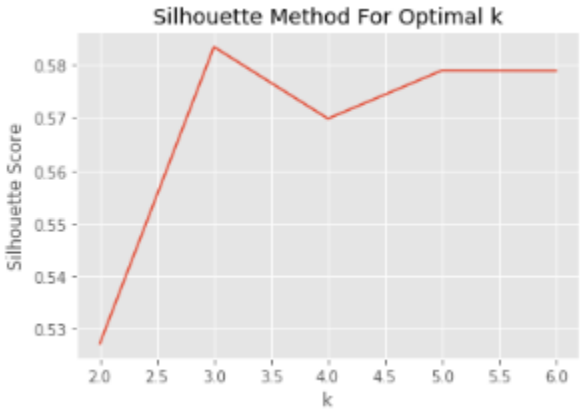

Figure: Silhouette Method

We see from the above figure that with K is equal to 3, as we have a better silhouette score than with K is equal to 4. Please note that the Elbow Method and the Silhouette Method are not really substitutes to each other but more means that can be used altogether to find K especially when finding the elbow is not enough by itself.

We will run K-Means with K equal to 3 on the Toronto combined dataset, having removed all unnecessary categorical data such names of the borough, postcodes etc., all numerical data is kept except the geodata. We obtain a labeling of our dataset, as per our initial definition, labels span from 0 to 2 (K or 3 clusters) and categorize each neighborhood as belonging to a given cluster. Using Folium and the population level, we plot the obtained Toronto neighborhood clusters.

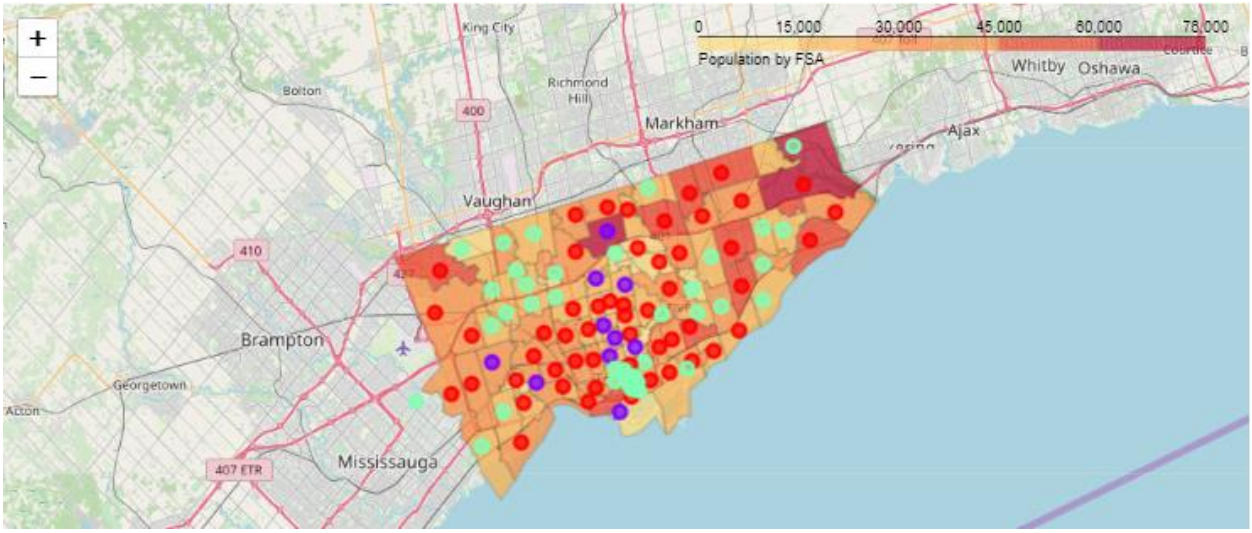

Figure: Toronto Neighborhood Clusters Vs Population level

We see from above figure that there is no relationship between population levels and our clusters definition. Let us try now with the income feature from our dataset and plot it against the neighborhood clusters.

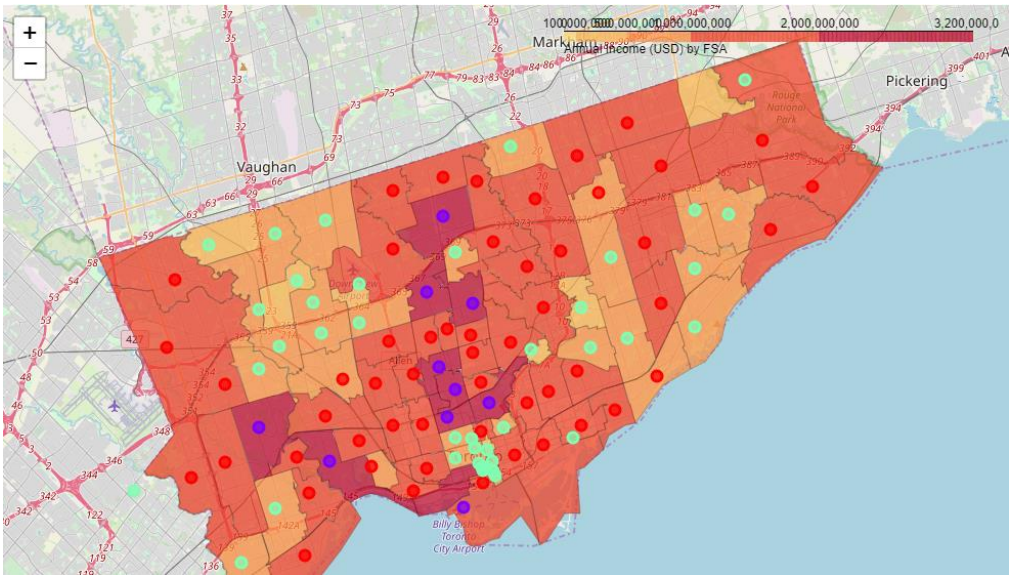

Figure: Toronto Neighborhood Clusters Vs Income Levels

Cluster Analysis Results

From the above figure, we definitely see that clusters follow are certain logic when mapped against the Toronto FSA annual income levels. Let us conduct some further analysis out of these clusters:

```
Cluster : 1

Nb of Neighborhoods : 53
Total Income in BUSD: 68.74
Total population : 1774689.0
Total Food Businesses : 1324.0
Businesses per Neighborhood : 24.98
Nb of Habitants per food business : 1340.0
Income in MUSD per food business : 52.0
Average income per habitant : 38732.0
Percentage of Neighboorhoods with no food business :9.43%

Top 10 food businesses :
 Cluster         Catering type  Totals  Habitants per business
       0           Coffee Shop   123.0                 14428.0
       0                 Café    95.0                 18681.0
       0               Bakery    64.0                 27730.0
       0           Pizza Place    64.0                 27730.0
       0            Restaurant    61.0                 29093.0
       0  Fast Food Restaurant    45.0                 39438.0
       0    Italian Restaurant    44.0                 40334.0
       0    Chinese Restaurant    39.0                 45505.0
       0        Sandwich Place    35.0                 50705.0
       0   Japanese Restaurant    34.0                 52197.0
```

Figure: Cluster 1 (Red) Highlights

```
Cluster : 2

Nb of Neighborhoods : 10
Total Income in BUSD: 24.46
Total population : 315011.0
Total Food Businesses : 142.0
Businesses per Neighborhood : 14.2
Nb of Habitants per food business : 2218.0
Income in MUSD per food business : 172.0
Average income per habitant : 77644.0
Percentage of Neighboorhoods with no food business :10.0%

Top 10 food businesses :
 Cluster        Catering type  Totals  Habitants per business
       1          Coffee Shop    23.0                 13696.0
       1                Café    21.0                 15001.0
       1          Pizza Place    16.0                 19688.0
       1           Restaurant    15.0                 21001.0
       1   Italian Restaurant    12.0                 26251.0
       1     Sushi Restaurant    12.0                 26251.0
       1       Sandwich Place    10.0                 31501.0
       1  American Restaurant     6.0                 52502.0
       1  Japanese Restaurant     6.0                 52502.0
       1      Thai Restaurant     6.0                 52502.0
```

Figure: Cluster 2 (Blue) Highlights

```
Cluster : 3

Nb of Neighborhoods : 40
Total Income in BUSD: 20.45
Total population : 642439.0
Total Food Businesses : 898.0
Businesses per Neighborhood : 22.45
Nb of Habitants per food business : 715.0
Income in MUSD per food business : 23.0
Average income per habitant : 31834.0
Percentage of Neighboorhoods with no food business :10.0%

Top 10 food businesses :
 Cluster         Catering type  Totals  Habitants per business
       2           Coffee Shop   168.0                  3824.0
       2            Restaurant    57.0                 11271.0
       2                 Café    47.0                 13669.0
       2  Fast Food Restaurant    40.0                 16061.0
       2               Bakery    39.0                 16473.0
       2   American Restaurant    33.0                 19468.0
       2            Food Court    32.0                 20076.0
       2           Pizza Place    31.0                 20724.0
       2    Chinese Restaurant    29.0                 22153.0
       2    Italian Restaurant    25.0                 25698.0
```

Figure: Cluster 3 (Green) Highlights

Cluster 1 represents the largest cluster of neighborhoods not only in terms of number of neighborhoods but also when it comes to population (85% more than the two other clusters combined), income and number of food businesses. Geographically spread over the east and west and the center, on the FSAs for which the global income is over 1B USD as depicted in the map, the first business rank is won by far by Coffee shops and Cafes, which comes as no surprise after our correlation analysis with the distance from the city center (Cluster 2 has strong part close to the city center); then we have at fairly the same importance Pizza Restaurants, restaurants and bakery. Obviously going for a Coffee place within this geographical cluster would be a quick win but the bakery business would be very interesting as the number of habitants per bakery is the highest within the most successful food businesses

Cluster 2, the smallest cluster in terms of geographical area and population, but not the least in terms of income, as its FSA have the highest annual income levels in Toronto, the 2016 annual average income per habitant is hovering around 77K USD more than the combined averages of the two other clusters. Another interesting aspect of this cluster is the number of served habitants per food business of 2218, approximately the double of the Toronto average, reflecting a far below average of food industry coverage and translating into a strong investment factor in this area. The top ranked businesses are again Coffee shops/Cafés, still they are not really leaving behind the other business as we see in the other clusters, we see a flatter distribution of business types, with Asian cuisine restaurants (Japanese, sushi, Indian) and Italian restaurants. The high income per habitant in this cluster could explain that we don't have a relative strong presence of fast foods.

Cluster 3, is the second largest cluster of neighborhoods in terms of number of neighborhoods and population, the striking aspect coming out of this cluster is that its revenues are the lowest, and the annual average revenue per habitant is around 30K USD below the Toronto average by more than 20%; the food sector coverage is around 716 habitants per business, in other words, a staggering -38% compared to the Toronto average. By no surprise the thriving business within this sector is also the coffee shops, yet Cluster 3 get the first rank when it comes to the coffee shops figures; which is aligned with our correlation analysis findings stating that coffee shops counts increases when distance from the city center decreases with Cluster 3 having a strong presence in the city center. The

distribution of food business category types is rather aggressive here, as Restaurants, the second top performer is almost three times less frequent, it comes as no surprise to see Restaurant type in this position as well from our previous correlation analysis. Fast foods and food courts are also very strong here, this could be explained by the relatively low average income, and their coverage make them quite interesting investments assets, while the coffee shops category, despite being a top performer, is not so appealing with its very low habitants/business ratio, especially when we have witnessed tremendously higher figures for this category within the other 2 clusters.

Further Results from Income ranges

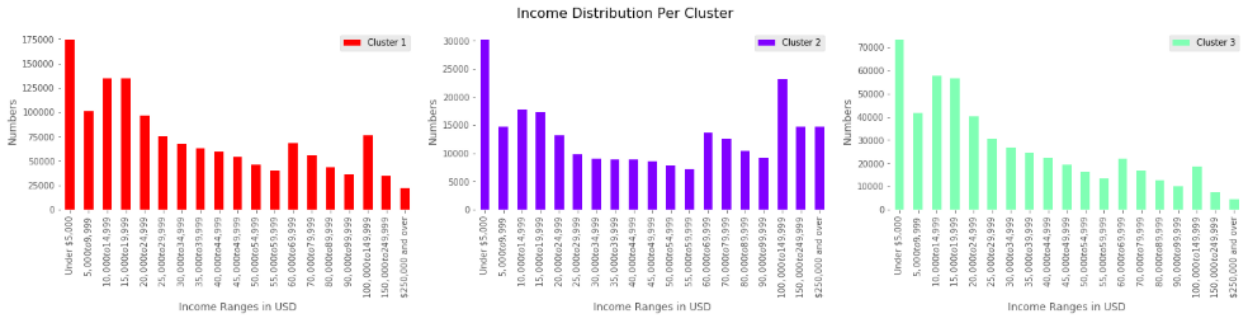

Figure Income Distribution per Toronto Neighborhood sCluster

The cluster 1 income distribution depicts that a majority of the inhabitants have their annual income ranging from below 5,000 to 29,999 USD, We can see that there is also a part of the incomes related to a strong middle class (78373 CAD in 2015,83020 CAD in 2017, source from Wikipedia), the high income population is also well represented and abide to the ratio of 10% as stated by Statistics Canada (https://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-014-x/99-014-x2011003_2-eng.cfm)

Cluster 2, as seen in the cluster analysis had the highest average annual income per habitant, this comes with no surprise as we see from the income distribution that both the middle class and high-income fractions are quite significant. Based on this and on previous analysis, more important investments could be undertaken in this cluster to target a high-end food market.

The income distribution of Cluster 3 on the other hand shows that most of the population within this cluster is ranked as low-income class ranging from under 5K to 30K USD for the annual income. This matches quite well with the cluster analysis and insights we went through in the previous section.

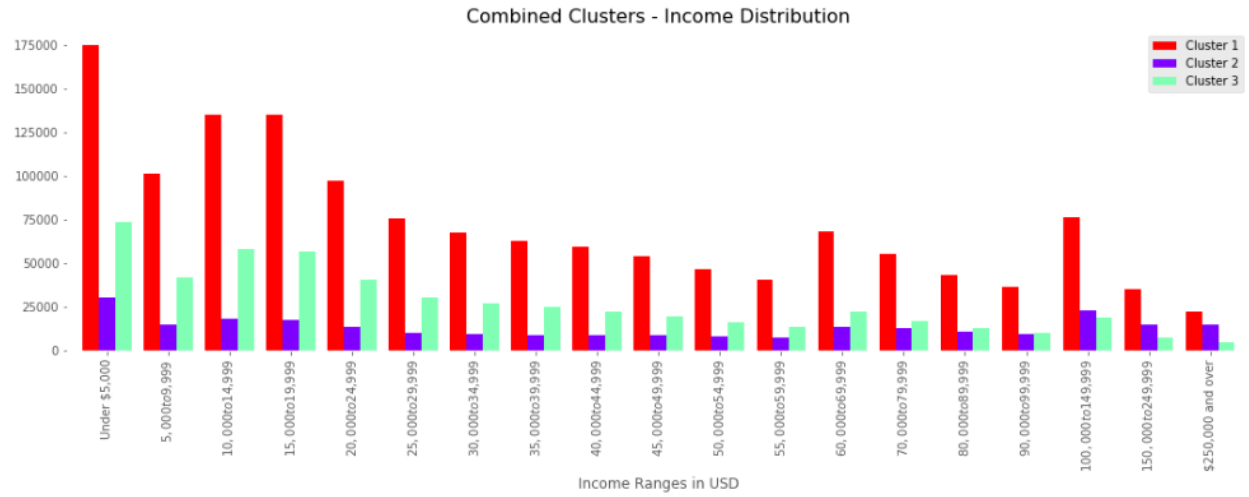

Figure: Neighborhoods Clusters Income Distribution - Comparative View

Discussion

Cluster 1 (red) represents the safest area to invest in a restaurant business, as there is a strong middle class and even 10% of its population would fit within the high income class, the trend would advocate to opt for a coffee shop as we don't see a saturation (14428 habitants per coffee shop) yet in this sector compared to the Toronto average figures we have seen for this type of business, but according our correlation analysis we could even invest in business that are highly correlated to the top trendy in order to avoid risks of competition in the future: Restaurants which are also well ranked in this cluster.

Cluster 2 (blue) is special as it is the only one having a strong middle and high income class, with the highest annual average revenue per habitant in Toronto, it looks like an attractive area to invest as it has the highest habitant per business ratio (double of the Toronto average), meaning that, there is still ample room to invest heavily in this cluster, yet one must bear in mind that its strong high and medium income classes would require heavy investments to undertake business that would be up to their expectation levels.

Cluster 3 (green) is the one with the lowest incomes levels (both global and average) is characterized by the supremacy of coffee shops yet the ratio of habitants per coffee shop is very low compared (factor of 4) compared to the other clusters, depicting an extraordinary coverage which should be taken with caution, as this specific sector may be too competitive to risk it. The recommendation would be to undertake a fast food or restaurant business with a low or aggressive pricing taking into consideration the characteristics of the population of this cluster.

Conclusion

We addressed a business problem with the data science methodology, we have used Python and its various libraries to extract, transform then manipulate data to visualize it (on maps or graphs or heatmaps) and analyze it using a top down correlation assessment approach. We have also relied on an unsupervised Machine Learning clustering algorithm, K-means, to further analyze the specificities of Toronto Neighborhoods, provided with several features such as population levels, annual average incomes, income distribution, restaurant businesses categories counts, mean distance from city center of all restaurant business within each neighborhood. Areas of improvement would be to first to enrich our dataset with further features such as age, ethnicity, housing prices for example, secondly to try out different machine learning algorithms.