Asian Restaurant Locations in Toronto

Rakshith Raghu

August 18th, 2019

Introduction

Background

Toronto’s food and beverage sector is a large and important part of the Ontario economy. It produces annual sales of over $17.75 billion. There are over 2,000 such establishments and they represent more than half of all food and beverage companies in the state of Ontario. The industry in Ontario grows at an average of 4% to 5%. This is above average growth to the rest of the economy. It employs more than 64,000 workers, of which over half are employed in the city itself.

Finally, there is a growing level of product development driven by desires for specialty foods (ethnically diverse themes). Specialty foods have grown twice as fast as the average growth in the sector, with an expected growth rate of 12% over the next five years. There is clearly a new and growing demand for ethnically diverse foods in the greater Toronto region.

Problem Statement

Given the new and growing demand for ethnically diverse food in the greater Toronto region, what would be the optimal locations to set up a new specialty food with an Asian (Chinese) theme? The assumption is that an optimal location would be one that is highly trafficked, that contains potential for unmet demand, and that has limited competition from similar themed restaurants. A final assumption is that competition will include other mid and high tier restaurants (non-fast food).

Stakeholders

There are three potential stakeholders would be interested in the analysis this paper presents. The first would be prospective entrepreneurs interested in opening a new themed restaurant. This group could benefit from optimal location with higher sales from better demand and limited competition.

Their potential backers and funders would represent the second group. This group, the financial backers or shareholders, would benefit from a more rational basis for restaurant location. This could help them in rationalizing the risk and level of investment they intend to make in a business.

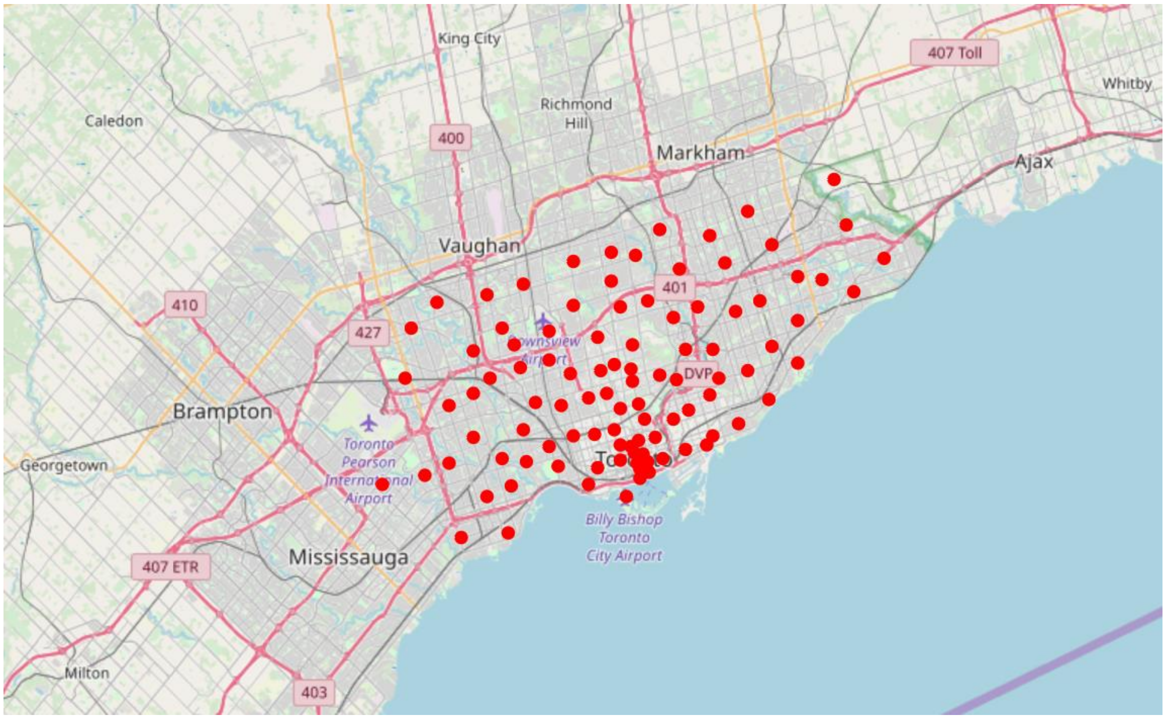
The last group would be potential policy makers. A rational basis for restaurant setup would potentially aid them in consideration of land development in the city. Since their political goal will be to aid the citizenry, this analysis could provide them a rational basis for choosing which locations to set up and prepare for commercial development.

Data

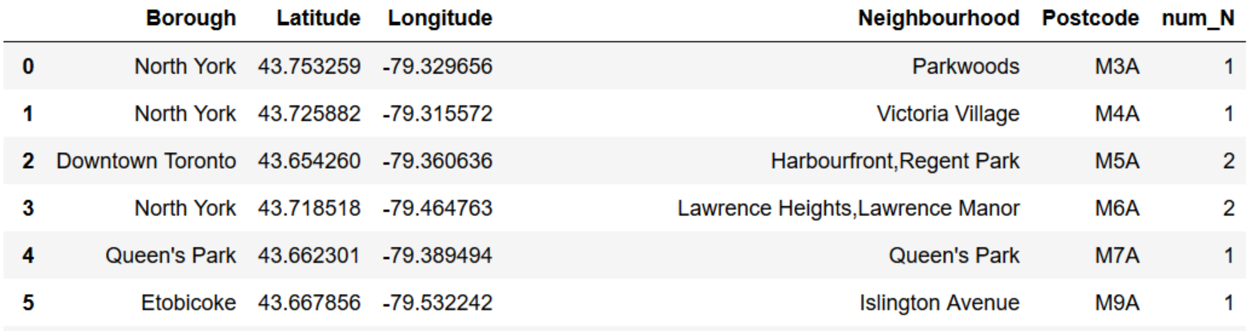
Data of Postcodes

The data was initially obtained from two locations. The first was a Wikipedia table of postcodes in Toronto itself. The table was split into three columns the Postcode, Borough, and Neighborhood. The second source of data was a data set provided by IBM. This data set contained the Postcode, Latitude, and Longitude for columns. The postcode appeared to serve as a primary key.

A level of cleaning was applied to the Wikipedia table dataset. First, rows that were empty for the Borough column were dropped from the data set. Secondly, where Neighborhood was empty, the data in the Borough column was copied over. Finally, postcode was made to serve as a primary key, so multiple rows were combined with the neighborhood column being used to store a list of individual neighborhoods for each index. This dataset was then merged with the IBM dataset to provide one table with columns Postcode, Borough, Neighborhood, Latitude, and Longitude. A new column num_N, representing number of strings in the Neighborhood column was also added. This data was visualized:

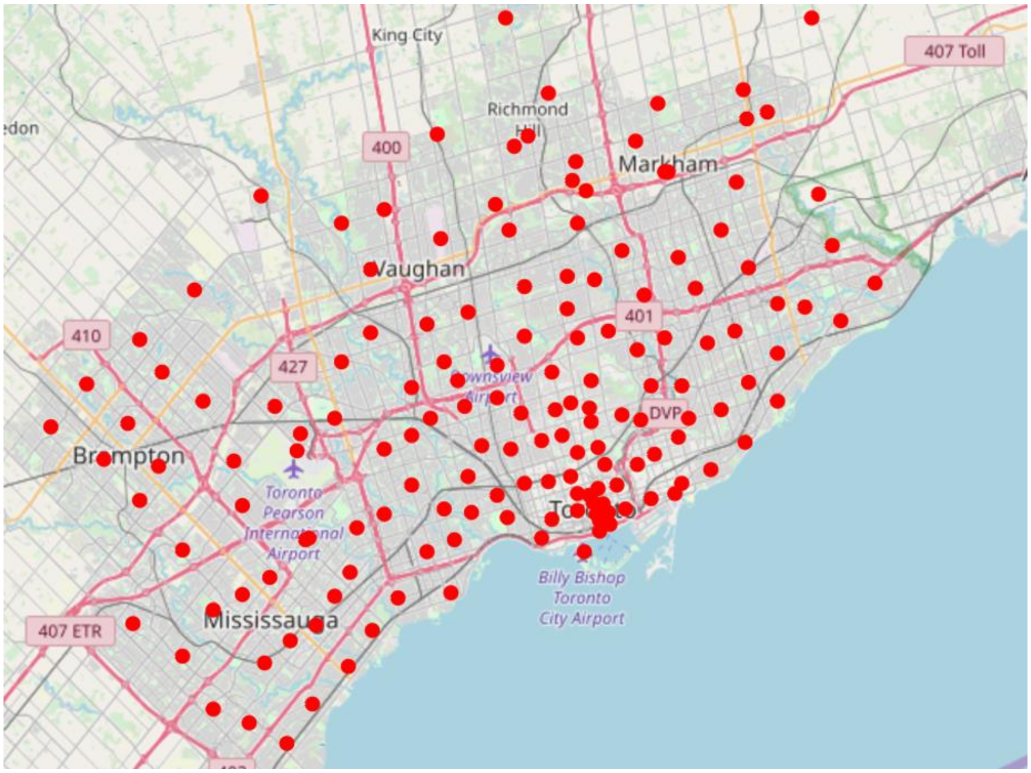


As the analysis expanded, it was quickly understood that the major suburbs connected to Toronto should also be included. This included Brampton, Mississauga, Vaughan, Richmond Hill, and Markham. A google fusion table was found that contained the relevant columns for FSA (postcodes) across all of Canada. This dataset was restricted to the relevant suburbs using the column “Place Name.” Data cleaning then occurred, with the columns being restricted and renamed to Postcode, Borough, Latitude, and Longitude. A column for Neighborhood was added, with values copied over from Borough. A column for num_N was added with value 1 for every row. Finally, this dataset was concatenated with the original data set to get the final table of postcodes of which the first five rows are visualized below.

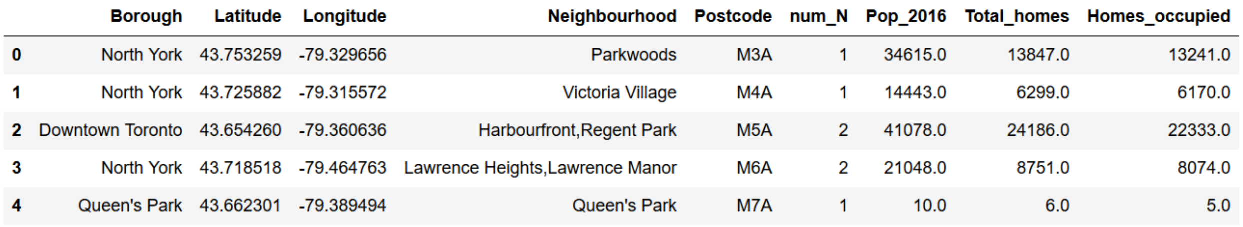


	Borough	Latitude	Longitude	Neighbourhood	Postcode	num_N
0	North York	43.753259	-79.329656	Parkwoods	M3A	1
1	North York	43.725882	-79.315572	Victoria Village	M4A	1
2	Downtown Toronto	43.654260	-79.360636	Harbourfront,Regent Park	M5A	2
3	North York	43.718518	-79.464763	Lawrence Heights,Lawrence Manor	M6A	2
4	Queen's Park	43.662301	-79.389494	Queen's Park	M7A	1
5	Etobicoke	43.667856	-79.532242	Islington Avenue	M9A	1

As a final test, the data was visualized by latitude and longitude. It was deemed, based on that visualization, that the data effectively captured all relevant postcodes in greater Toronto.



A final element was added to this table. This was population and housing statistics for each postal code. A dataset from Statistics Canada with columns such as Geographic code (postcode), Population 2016, Total private dwellings 2016, Private Dwellings occupied by usual residents 2016. Data cleaning was done, dropping unimportant columns and renaming the formerly named columns Postcode, Pop_2016, Total_homes, and Homes_occupied respectively. The data was then merged with the main postcode table by an inner join on postcode. On the next page is the first five rows of this table.



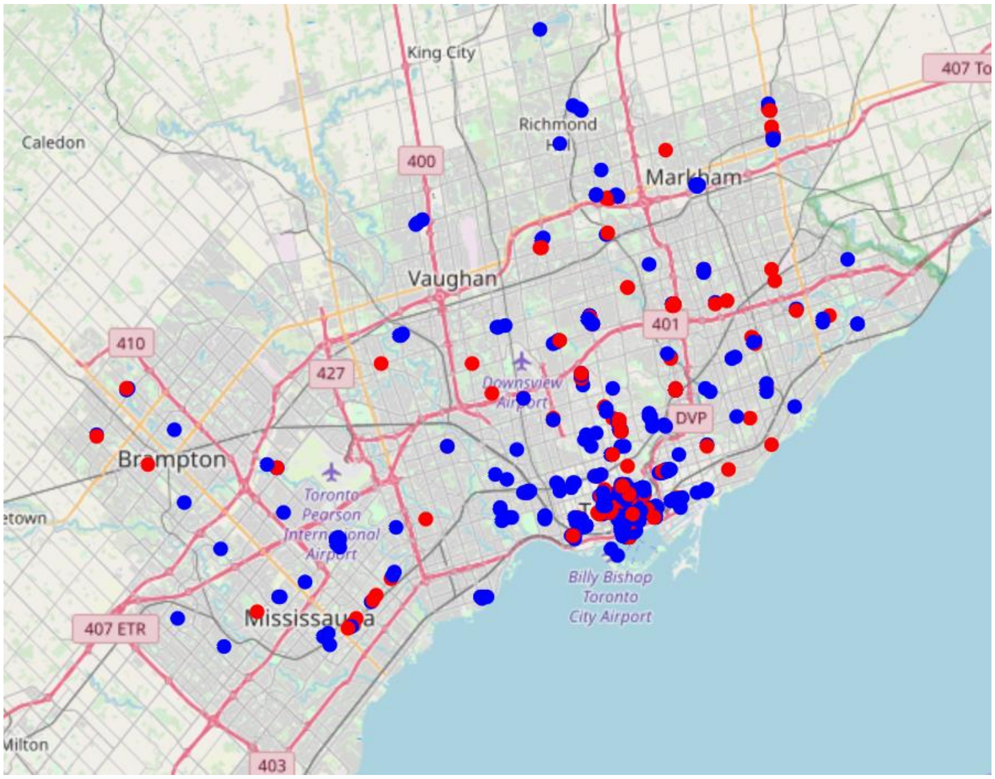
	Borough	Latitude	Longitude	Neighbourhood	Postcode	num_N	Pop_2016	Total_homes	Homes_occupied
0	North York	43.753259	-79.329656	Parkwoods	M3A	1	34615.0	13847.0	13241.0
1	North York	43.725882	-79.315572	Victoria Village	M4A	1	14443.0	6299.0	6170.0
2	Downtown Toronto	43.654260	-79.360636	Harbourfront,Regent Park	M5A	2	41078.0	24186.0	22333.0
3	North York	43.718518	-79.464763	Lawrence Heights,Lawrence Manor	M6A	2	21048.0	8751.0	8074.0
4	Queen's Park	43.662301	-79.389494	Queen's Park	M7A	1	10.0	6.0	5.0

Data of Restaurants

Data was found by using Foursquare api. A query was done for restaurants within 500 meters (0.5 km) of each postcode. The restaurants were considered of category restaurant, diner, tanerna, steakhouse. A final element was added to each restaurant, whether it was of the category Asian, which was defined by Foursquare. This element was a Boolean True/False.

The data was stored in three variables. The first was a dictionary of all restaurants queried. The second was a list, by postcodes index from the main table, of lists. The inner list was every restaurant in 500 meters of that postcode center. The third was a dictionary of all Asian restaurants within 500 meters of some postcode center.

With this method, 746 restaurants were captured with 193 being specifically Asian themed. This is less than half of the total food and beverage locations potentially in Toronto. This does represent the mid to high brow food locations in the city. These restaurants are visualized geographically below with blue being non-Asian and red being Asian themed.



Three points of interest were noted from the visualization. First, most of the restaurants are located along a section of southern Toronto, likely being the downtown. Second, there is a decent overlap between the locations of Asian and non-Asian themed restaurants. Finally, in the more suburban areas, the majority of the restaurants are next to major roadways.

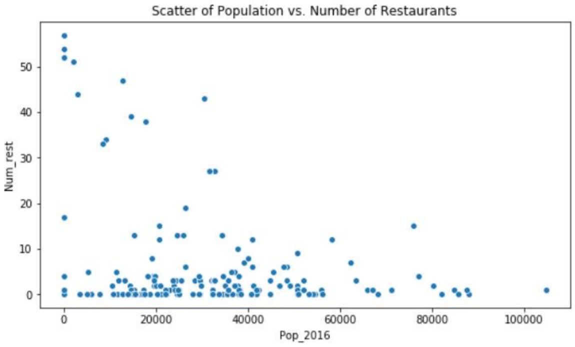
With the list and dictionaries, two more columns were added to the main table. The first is the count of the number of restaurants generally in each postcode. The second was the number of Asian restaurants in each postcode.

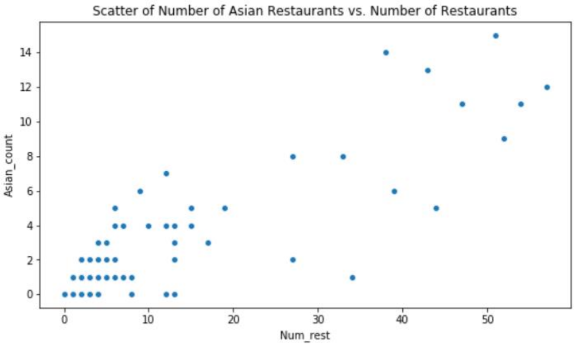
Methodology and Results

Exploratory Data Analysis

Exploratory data analysis was done through two methods. The first was through data visualizations using charts and heatmaps. The second was through looking at correlations of the features with each other.

1. Initial Scatterplots

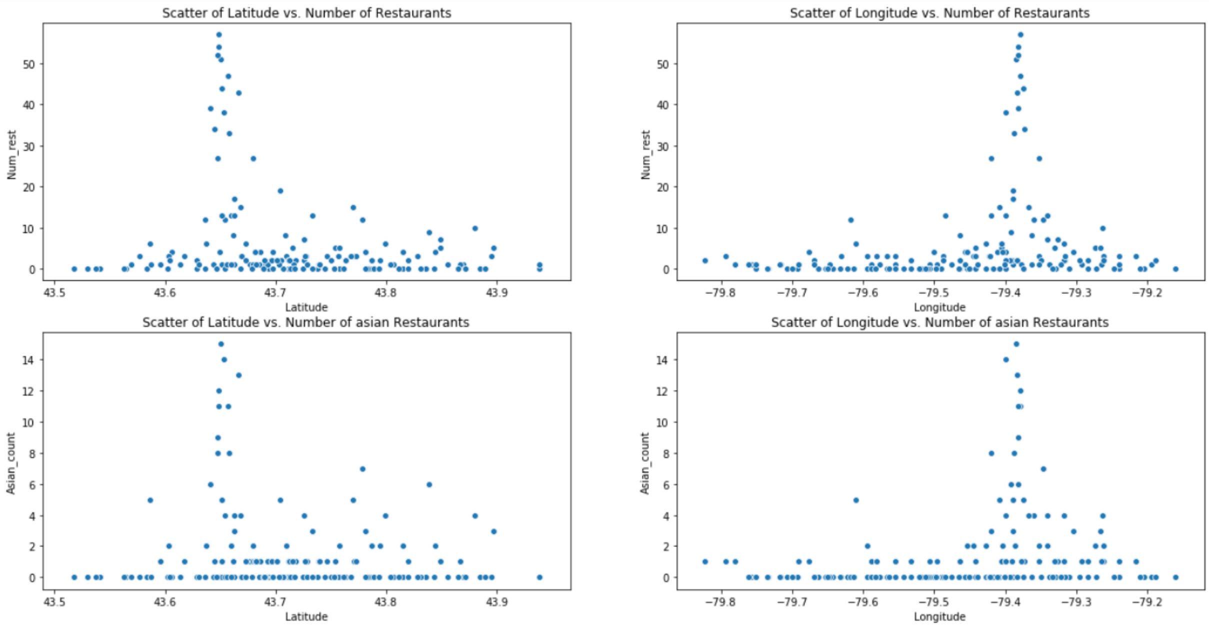




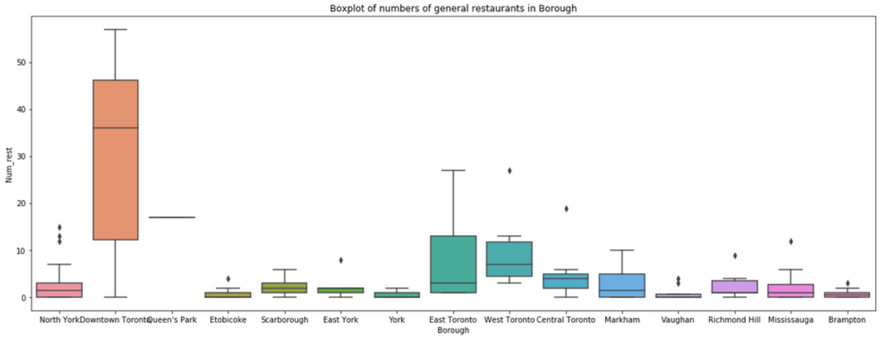
The first scatterplot, the left, compares features Pop_2016 to Num_rest. This was a comparison of population vs. number of restaurants. This implied a deeply negative relationship between population and number of restaurants. What may sit behind this relationship is a policy split between residential and commercial districts. The locations highly trafficked may be locations close to a workplace rather than close to a residential area.

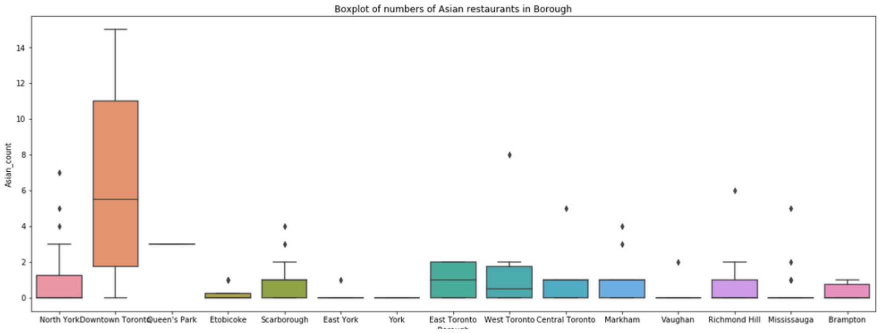
The second chart looks at the relationship between the number of Asian restaurants and the number of restaurants generally. It appears that at low numbers, there is a directly linear relationship between the two features. But as the number of restaurants increases, the number of expected Asian restaurants becomes saturated and approaches a 0 slope.

2. Longitude and Latitude against restaurant number



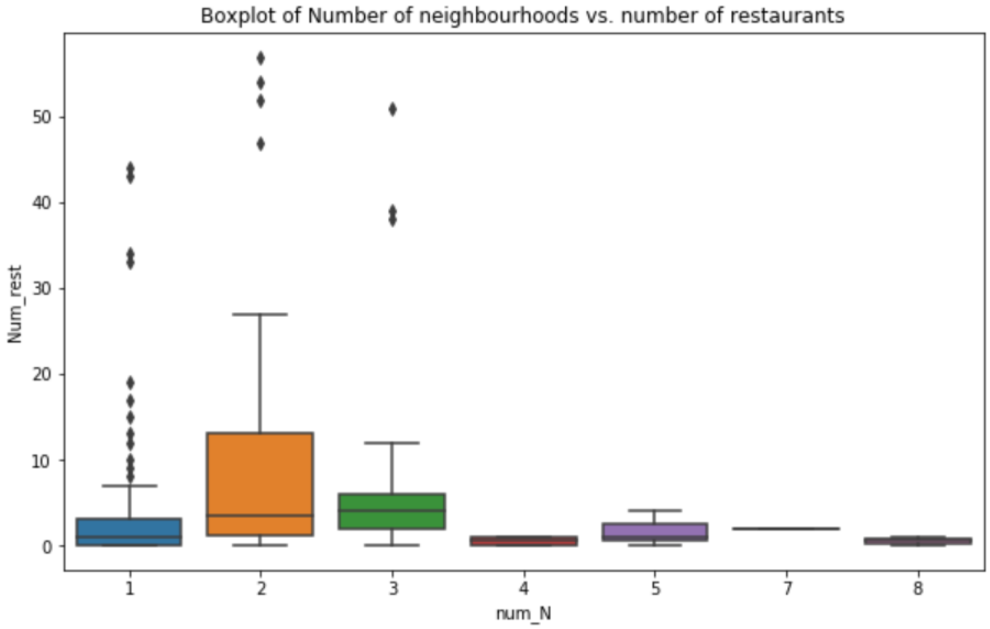
It appears to be little difference between Asian restaurant numbers and general restaurant numbers in the latitude and longitude scatterplots. There appears to be a concentration of the restaurants between 43.6 to 46.7 Latitude and -79.5 to -79.35 Longitude. This is believed to correspond with the difference between a downtown area and non-downtown areas. A boxplot of Boroughs was used to check validate this hypothesis.





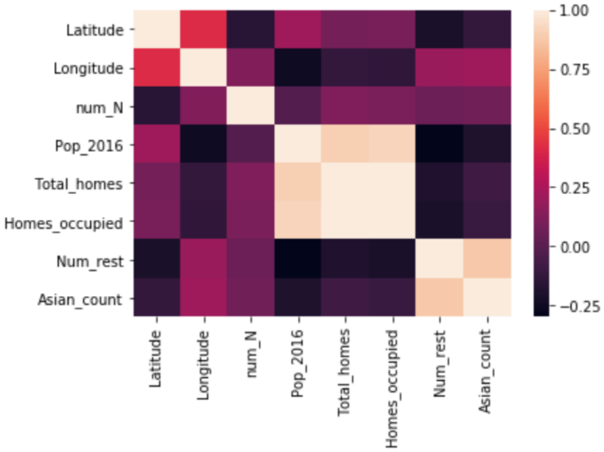
The boxplots of the boroughs seem show this relationship. 75% of Downtown Toronto counts of Asian restaurants and restaurants generally greater than 75% of the counts from all other Boroughs. A new binary feature, Downtown, was added to represent the difference between downtown Toronto and non-downtown Toronto.

3. Number of Neighborhoods vs. number of restaurants



This supports the hypothesis from the initial scatter plots that postcodes with a smaller number of neighborhoods, and thus a smaller number of residential spaces, would contain a higher number of restaurant counts. Given that the data seems to reach a highest 50% percentile at 2 and 3, this appears to be an inverted X^2 curve. Based on this plot and the initial scatter, a 2nd order parameter, num_N^2 will be added. This feature and the added one will be treated as continuous variables even though they may not necessarily be so.

4. Correlation Heatmap



This heatmap looks at the correlation between each feature. One thing that jumps is that there is a high correlation between the count of restaurants generally and the count of Asian restaurants. These two features will not be used in predicting the other.

Features Total_homes and Homes_occupied appear to have the same correlation with every other feature. It will be assumed that these features have a nearly exact linear relationship and thus both should not be used in the same model (multicollinearity). Homes_occupied was chosen as it is closer to representing residential demand.

5. Final feature selection

Given the analysis that has been done, a final set of features was created/down-selected. These features are Latitude, Longitude, Pop_2016, Homes_occupied, num_N, num_N_2, and Downtown.

Building Models

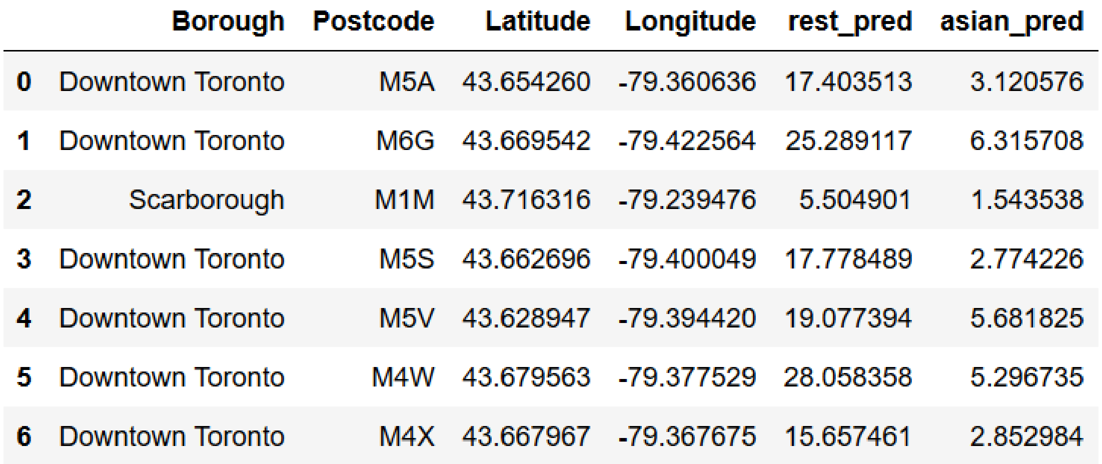
The strategy for the use of models is two-fold. In the first stage, two simple linear regression models will be built using the features to predict the number of restaurants and the number of Asian restaurants for each postcode. The predictors in these models would be scaled by subtracting the mean and dividing by standard deviation. Once built, and verified as informative based on metrics such as Mean squared error and R^2 score, the models were then run on the data to predict the number of restaurants and Asian restaurants.

Several assumptions were made in doing this. First, there was no split in the data into training, test, and validation sets. Instead, the models were trained on the whole of the data, and then asked to predict using the same data. The objective is not to deploy/use those models generally. The objective is to instead use those models, trained specifically to predict this data, to find postcodes where the number of restaurants and number of Asian restaurants were overpredicted. This is assumed to be a sign of unmet potential demand in that postcode. The models produced the following scores below:

	Mean Squared Error	R^2 score
Model for Asian prediction	4.48	0.4448
Model for general restaurant prediction	59.4797	0.5667

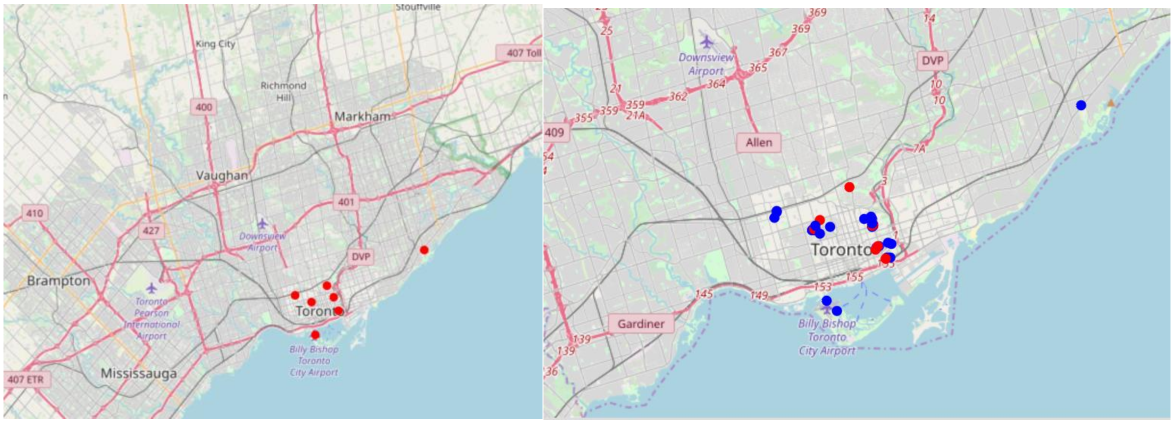
In general, the R^2 scores argue that the models are not very effective at capturing total variability in the data. But, with scores above 0.4, they do capture a significant amount of it. With an average error of 2 for Asian counts, and slightly less than 7 for restaurants generally, the models were assumed to be useful.

After the predictions were made, two questions were asked. First, what postcodes do the linear model for restaurants generally overpredicted. Second, what postcodes do the linear model for Asian restaurants generally overpredicted. These two questions were used to create a query getting only postcodes where both those questions were answered “yes”. An assumption was made here that the overprediction for restaurants generally should be > 5, and that the overpredictions for Asian restaurants should be greater than 1. This was done to hopefully only grab postcodes where significant demand might be unmet. This queried set was produced below:



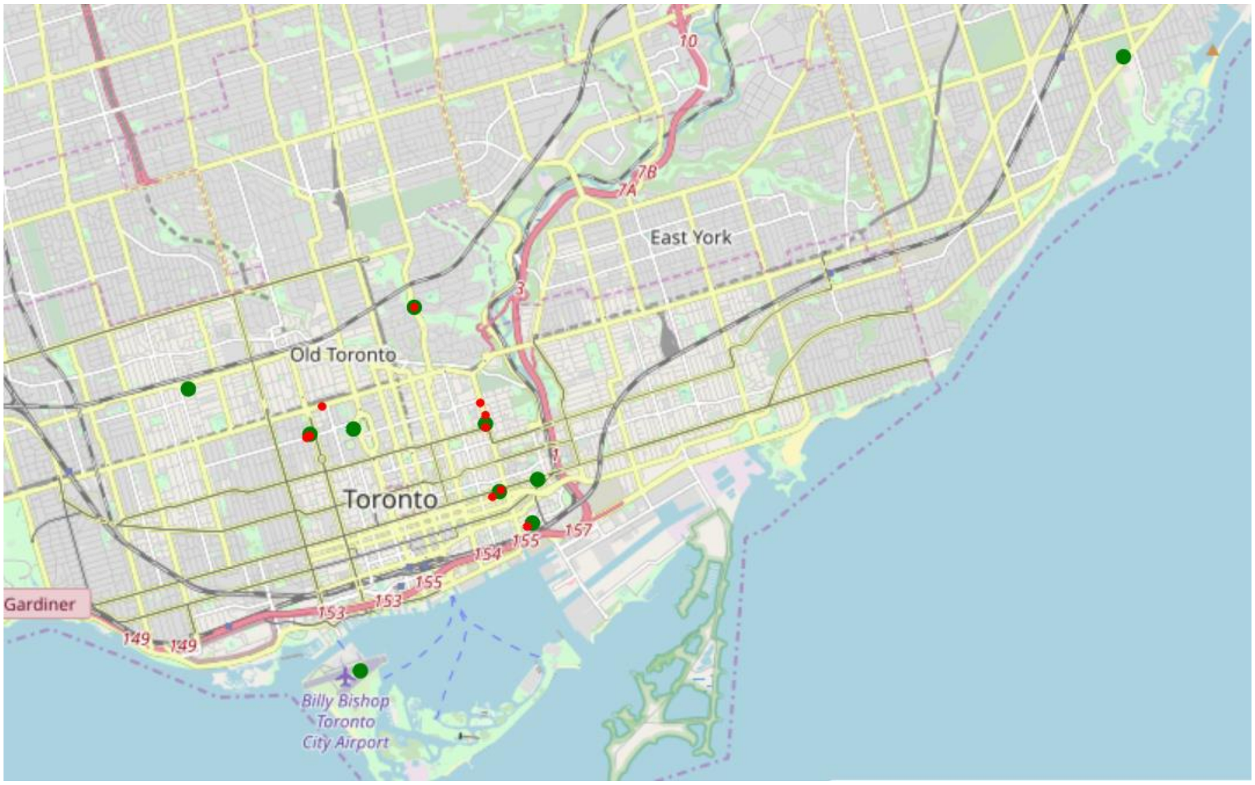
	Borough	Postcode	Latitude	Longitude	rest_pred	asian_pred
0	Downtown Toronto	M5A	43.654260	-79.360636	17.403513	3.120576
1	Downtown Toronto	M6G	43.669542	-79.422564	25.289117	6.315708
2	Scarborough	M1M	43.716316	-79.239476	5.504901	1.543538
3	Downtown Toronto	M5S	43.662696	-79.400049	17.778489	2.774226
4	Downtown Toronto	M5V	43.628947	-79.394420	19.077394	5.681825
5	Downtown Toronto	M4W	43.679563	-79.377529	28.058358	5.296735
6	Downtown Toronto	M4X	43.667967	-79.367675	15.657461	2.852984

The rest_pred is the difference between the predicted number of restaurants and the actual number of restaurants. The Asian_pred is the difference between the predicted number of Asian restaurants and the actual number of Asian restaurants. With most of these differences between outside the mean error of the models, it assumed that these are locations where potential demand is being unmet. There were 47 restaurants within 500 meters of these locations. Their locations, and the locations of the restaurants are visualized below:



On the left side is a map with markers for the postcodes. As can be seen, nearly all of these postcodes are located in a part of Downtown Toronto, except for one in the eastern part of the city. On the right side is a map with markers for the 47 restaurants. Here they appear to be clumped in certain areas, likely representing shopping/commercial areas. The blue markers represent non-Asian restaurants and the red Asian. When zoomed in on, it appears that some of these clumped areas contain no Asian restaurants, and thus would likely serve as ideal locations for setting up a new one.

The final element of the analysis was to get the coordinates and thus addresses of these areas. About 11 discernable clumps were recognized and thus a KMeans model, which seeks to cluster data into groups. 11 cluster centers were randomized and the model was run on the 47 restaurants caring only for their latitude and longitude. The clusters were then visualized.



The red represents the location of Asian restaurants while the green markers represent the cluster centers. There are 5 cluster centers that do not have Asian restaurants in their immediate vicinity. One of them, located at the Billy Bishop Toronto airport was ignored while the rest were reversed geotagged. These last 4 locations were concluded to be the optimal locations to start a search for an Asian restaurant location.

Discussion

Four major areas/clusters were noted to be where other restaurants are. These areas have, based on the analysis, unmet demand and specifically have no Asian restaurants. In terms of general features, they tend to be areas in commercial districts with lower housing in the specific postcode. They also tended to be in the downtown area. The four areas are described below:

1. *Burnfield Avenue, Christie Pits, Toronto, M6G 3N1*. It is a road connected to a public recreation area (park) in downtown Toronto. It is also near a subway station. It is a generally multicultural neighborhood that has served as a transitional living space for immigrants. This means a decent number of apartments located there and thus has a demand for lunch and dinner.

2. *Hart House Circle, Bloor Street Culture Corridor, Toronto, M5S 3H3*. This is part of a diverse cultural district with many distinct, world-class, cultural and art organizations. That it has limited number of Asian restaurants means this is a space to open a new cultural theme in a district designed to promote multiculturalism. It is also nearby a university and is downtown which means it will be a highly trafficked place.

3. *Ashby Place, Corktown, Toronto,M51 1G1*.This is an old residential district close to the heart of downtown Toronto. It is currently in the process of being gentrified, with many of its residential buildings being torn down in high value bidding for the land. There are a number of vacant industrial buildings there that have been repurposed for restaurants and other shops. That it is now being gentrified and converted into a commercial location makes it great for staking out an initial location for a market. The downside of this location is that there will be high competition to grab space as it is renovating.

4. *Kingston Road, Cliffside, Toronto, M1N 4A7*. This is an upscale suburban housing area with a number of parks and waterfront areas. Much of the population are of descent from the British Isles, which might provide a desire for exotic/ethnic food. Its distance from the highly trafficked areas of downtown Toronto does mean less demand than the other areas, but it does also have potential competition.

Conclusion

In this analysis, various data sets were combined to get the location, population, number of restaurants, housing, and borough of individual postcodes. The analysis was done on the greater Toronto area which can be defined as the city itself and the suburbs to the north and west side of the city. To apply this analysis to other cities would likely require certain changes to how census style data is collected. More importantly, policy preferences and the economic superstructure will lead to different cities being organized along different lines. This means that certain factors, such as the relationship between the number of neighborhoods and the number of restaurants may not be the same in other cities.

If a generalizable principle were to be made, commercial areas in downtown cities close to public transportation and/or major roads would be good places to begin looking for a restaurant location. The method of analysis, looking for unmet demand could also be generalized for other locations, using factors more correlated with actual spending/demand in a certain geographic area. A more in-depth analysis considering the relationship between types of restaurants and the number of restaurants could also be considered. If the desire is to create deployable models to predict the number of restaurants, geographic features such as latitude and longitude would have to be excluded.