Capstone project
Battle of the neighborhoods
Scoring restaurant startup location by world cuisine

Karim Abbas
Karim.central@gmail.com

June, 2020

1. Executive summary

Toronto is a world alpha city. It is large, sprawling, and diverse. This means two contradictory things for people wanting to open a restaurant business: First, the city is a prime target due to its population, wealth, and connectiveness; and second, there is a great degree of competition and saturation. There is conflicting anecdotal and statistical data on how likely it is for a restaurant business to succeed at conception, but one thing is for sure: failure rates are extremely high for small startups. There are many factors that affect the chance of success of a startup restaurant, but many of these factors collapse into one word: location. Here, we present a tool that advises new restaurants about the types of neighborhoods most suitable for opening the business. We cluster based on a number of parameters, some of them are obtained purely from data sources, and some are derived. The main contribution in this work is that we consider the type of world cuisine the restaurant intends to serve, and then derive parameters based on saturation, potential, as well as amenability to diversity. These are then used to derive three indicators for likelihood to succeed: a total metric, a linear regression score, and a cluster with similar neighborhoods.

2. Detailed discussion of the business problem

Restaurants are among the least likely businesses to thrive. On conception, three out of five restaurants can expect to not make it past the first year. The ratio rises to four out of five at four years [1]. One major problem with restaurants is that they are a low profit margin industry. They are also extremely sensitive to economic and political upheaval as well as natural disasters. For example, the financial crisis of 2008 squeezed restaurant businesses so tight that only the ones that already had very high margins could remain in business [2]. The COVID-19 pandemic is also decimating restaurant businesses around the world, and the impact is still ongoing.

According to CNBC, the number one factor that affects the success of a restaurant is location [3]. This is also known by common wisdom. But we have to ask ourselves, what makes a location more suitable for opening a restaurant? The following are possible factors:

- The location has a high population. Population can translate into clientele, but it does not have to ..
- Wealth of the location. This is probably more important than population. In fact, this could be the single most important factor in determining foot traffic for the restaurant
- Density of restaurants already in the location
- Density of restaurants similar to the one we intend to open

New ventures in the restaurant business desperately need pointers on where a “good” place to open the business would be. There is a need for this advisory especially to small startups with a limited budget. It is important to understand that most of the anecdotal evidence on failing restaurants is actually based on impressions from smaller startups. Larger restaurants and instant multi-branch networks can expect a much higher success rate [4].

In this work, we develop a tool to help new restaurant owners decide the chances that their business will succeed in certain neighborhoods. We introduce a number of metrics that together characterize a neighborhood’s ability to support the business. Some of the metrics are raw and obtained directly from Foursquare or other sources listed below, some are derived. The metrics will characterize the saturation of restaurants in the neighborhood, but will also have indicators of the diversity, willingness, and wealth of the clientele. To do this, we will fine tune our metrics to include information about restaurants by type of cuisine. The metrics are discussed in detail in the methodology section, but are listed below for reference:

- Population of the postcode
- Income per person in the postcode
- Proliferation of restaurant business in the location
- Proliferation of restaurants of particular cuisine in the location
- Diversity of restaurants in the location
- A metric that indicates spending power per person per restaurant

Target audience: The target are prospective owners of new restaurants, particularly small restaurants with a staff count of less than a couple dozen. This is a high risk, low margin, high closure business model that is extremely sensitive to location.

Summary of the problem: Given I will be opening a business of a specific cuisine type, what is the best group of locations that could support my business. The location has to have high potential to generate traffic, but should also show signs of low saturation. Once I have chosen a cluster of locations, what can you tell me about the relative advantage of each location within the cluster?

3. Data needed and data sources

Our master data frame will contain information about a group of postal codes in downtown Toronto. For identification purposes we need the postal code, name of neighborhoods, and borough. We also need latitude and longitude data to query Foursquare. We have to somehow obtain data on income and population to characterize the potential of the location. Finally, we have to obtain restaurant count by world cuisine, for which we will use Foursquare. Datasets and their sources are discussed in detail below.

Postal code, neighborhood, and borough

This has been obtained in week 3. Code for obtaining it is included in the notebook. Sources are listed below. In week 3, we also introduced a reduced frame containing only boroughs with the string “Toronto” in the name. We will also need a reduced frame here for reasons that will be explained below.

Source for neighborhood information:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Latitude and longitude data by postal code
This has also been done in week 3. Data can be obtained by interrogating geoawesomeness or by pulling from the file provided in week 3 (URL attached below). We choose the latter due to the lack of reliability of the tool.
https://cocl.us/Geospatial_data

Number of restaurants around center of postal code
This is the total number of restaurants drawn from a pool of world cuisines around the center of each of the postal codes in our main data frame. We will actually query Foursquare for the number of restaurants for each of the cuisines and will be storing them separately. This allows us to query Foursquare only once, then decide on the target restaurant cuisine offline. The list of target cuisines is ['Italian', 'Indian', 'Chinese', 'Japanese', 'Korean', 'French', 'Greek', 'Peruvian', 'Brazilian', 'Mexican', 'Spanish', 'German', 'Moroccan', 'Egyptian', 'Turkish', 'Persian', 'Thai']. This list can be changed, reduced, or expanded by the user. Once we obtain this information, we save it to a CSV file, allowing the rest of the algorithm to rerun without needing to query Foursquare. This allows us to rerun as many times as we need without being limited by Foursquare quotas.

Density of restaurants of particular cuisine around postal code center
This has to be distinguished from other cuisines when calculating one of the derived parameters. It will also be the label of our regression problem. This can be queried no different from other cuisines using Foursquare. In fact, as discussed above, we obtain and store raw data for all cuisines and store it in our main dataframe, allowing us to make and change the decision on the target cuisine without needing to contact Foursquare.

Population by postal code
As discussed above, population is a critical parameter in determining foot traffic. The tables below allow us to obtain this information by postal code. Since postal code (also known as FSA) will be the key in all our data frames, this will be very helpful.
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0
Data in CSV format:
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=9999&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV

GDP of postal code
https://open.canada.ca/data/en/dataset/1e8a8c6e-a7cc-4f08-a3cc-79c67427a102

https://www.canada.ca/content/dam/cra-arc/prog-policy/stats/individual-tax-stats-fsa/2015-tax-year/tbl1a-en.pdf

It is very hard to obtain income data by postal code in Canada. Census authorities do include very broad data about income in provinces. They also include data about income by gender, group, and other parameters. In terms of geographic distribution, the census office also has income data by census unit, this sometimes maps to FSA, but in other cases the mapping is not so easy. On the other hand CRA has tables on taxable income by FSA. This is as good as the data we need, if not more so. Reported (taxable) income is a very good indicator of the amount of spending power the population has in the location. There are two issues with this data:

- This is total income rather than per capita. This is not a big deal. A simple division can give us a per capita metric. Notice that total income by itself may be a valuable indicator, but it is not as good as per capital income due to lack of normalization by population
- The bigger problem is that data is in PDF format. This means we have to manually intervene to extract the data in a way that allows it to be included in the master frame
- This table provides data on population, however, the population data is misleading since it only indicates taxable population, we will use the population data from statcan instead

In the methodology section we will discuss how this data is preprocessed so that the metrics from the business problem section can be combined into a meaningful assessment of the location. Our aim will be to create “positive” metrics that are best when they increase and “negative” metrics that are the opposite.

Positive metrics will in general indicate potential for spending and traffic, negative metrics are indicators of competition or saturation. We will create six metrics, one of which will play the role of a parameter and a label under different scenarios. We will do three things with these metrics:

- Use K means clustering to cluster similar postcodes together. This will indicate which groups of postal codes are more likely to support our target cuisine. Using map visualization, we may be able to obtain some insight about whether the geographic distribution reflects a particular pattern
- We will calculate a “total metric” using the six parameters. This metric will range from 0 to 6. Zero will represent absolutely no potential for growth in the particular cuisine. A score of 6 represents the highest potential for success
- Linear regression will be used with the number of restaurants from the target cuisine being the label and all other metrics being features. We will use multiple linear regression and then calculate the error for every postal code. This error will then represent how far off the postal code is from what the model suggests. Positive errors could indicate oversaturation, while negative errors can be a sign of unused potential

4. Methodology

The main data frame contains a number of factors that can affect the chances of the target cuisine to success. Some of these factors are positive and some are negative. For positive factors, we will allow them to affect clustering in a direct manner. For those that are considered negative, we must define an inverse relation.

Population will in general be a net positive. More people living in the postal code means more traffic, which means more customers. How highly to weigh this could be controversial because transport, pickup, and home delivery models can also encourage visits from out of neighborhood. This would be particularly true of tightly clustered postal codes.

Wealth is a definite positive, there is also no question that it should be heavily weighed. There are a few caveats though. As discussed above, we should only use per capita income rather than total income. During preliminary data analysis, look for outliers. Postal codes with extremely high or extremely low incomes often indicate special locations without a large residential or commercial presence.

Having more restaurants from the specific kind of cuisine we intend to deliver can be a negative or a positive. On the one hand having a lot of restaurants from the same cuisine could indicate that the population is receptive to the specific kind of food. However, it is also an indication that competition is dire in the neighborhood.

To solve this problem, we will consider a high density of restaurants from the specific cuisine a net negative. However, we will also include two other metrics. The first is the total number of restaurants drawn from a pool of different world, the other is a metric representing the diversity of this pool, including diversity provided by our own genre. Both these metrics are considered net positives.

We will also add another metric to balance out the net positivity of high restaurant presence. This will account for saturation. This metric is the number of dollars available per restaurant in the postal code. This is calculated as the total gdp of the postal code divided by the number of restaurants. This is a net positive metric, but it adds a degree of negativity to restaurant count.

Table 4.1. Metrics used to cluster and fit neighborhoods. The tcuisine metric is the label for linear regression. All others are always features.

Metric name	Description	Positive/Negative
Population	Total population of neighborhood	P
Per_capita	Per capita income	P
Df_merged[tcuisine]	Number of restaurants from our cuisine	N
Total_minus	Number of restaurants of all other cuisines	P
Variance	Variance of restaurant cuisines	P
Rest_pot	Income per capita per restaurant	P

Our metrics exist on completely different scales. For example, population is measured in tens of thousands, while restaurants will be counted in ones. For every positive metric, we will perform normalization as follows:

- Find the maximum of the metric
- Divide all other entries by this maximum

This ensures that positive metrics are all normalized to lie between 0 and 1. Note that tuning these metrics by managing their maxima allows you to manage the relative importance of each. But for starters, let us consider them all to be equally weighed.

For negative metrics (of which there is only one) we will normalize them as follows:

- Find the maximum of the metric
- All other entries are one minus the metric divided by the above maximum

Again, the results will be between 0 and 1, but this time as the metric rises, the normalized metric tends to zero. This is better than calculating the reciprocal of the metric because it avoids distorting the dynamic range of the data. Division tends to magnify small numbers and has no upper limit.

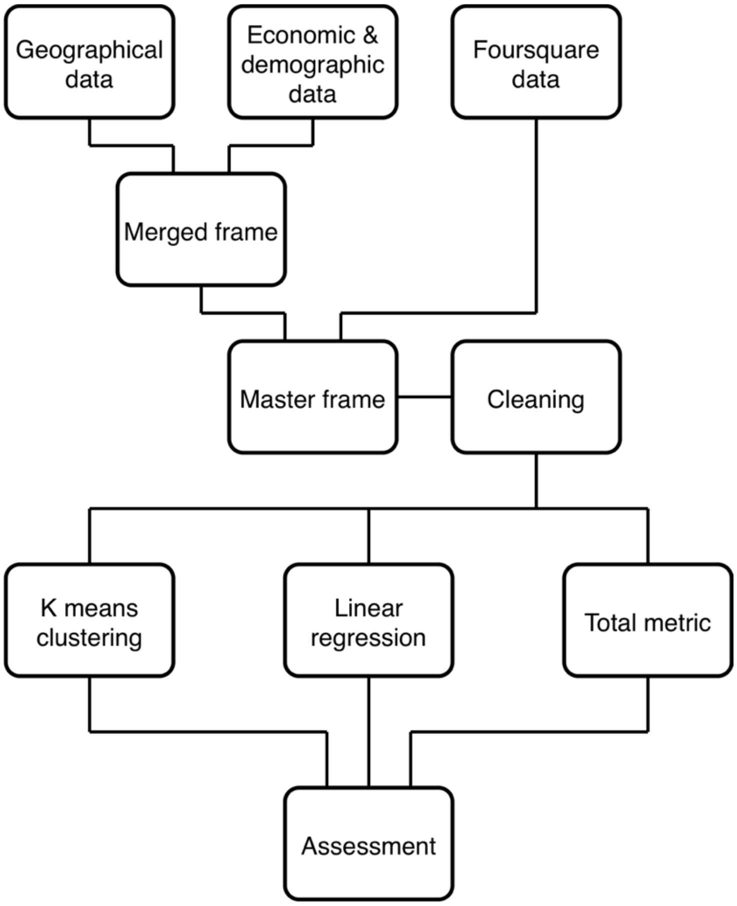


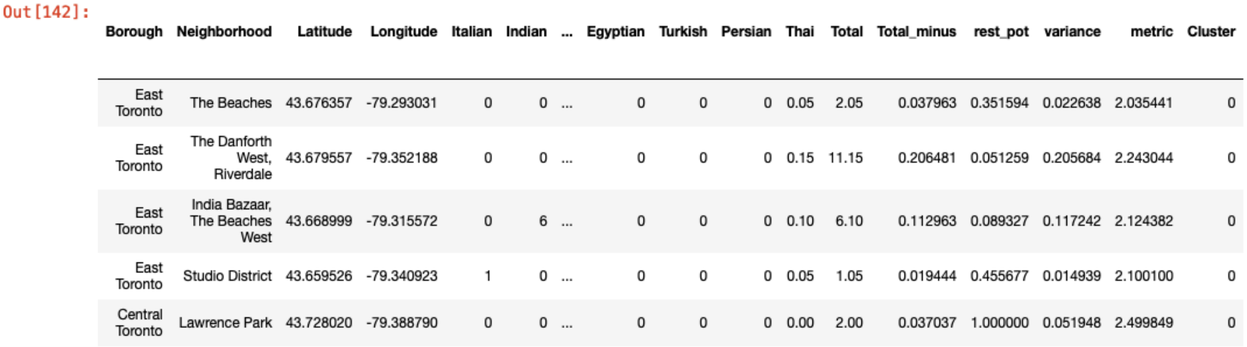
Figure 4.1. Flowchart for data acquisition and processing.

To summarize the steps so far, we have (figure 1):

- Obtained naming data on postal codes
- Merged with latitude and longitude data
- Used lat and long data to query Foursquare on the number of restaurants from a list of world cuisines
- Merged with the results of all Foursquare queries
- Calculated a new metric called rest_pot=per capita income/total restaurants in postal code
- Calculated and normalized the variance of restaurants by genre. Although variance is itself normalized by the number of entries, we normalize so that the maximum of the

variance variable is itself 1. This will measure the spread of restaurants across world cuisines

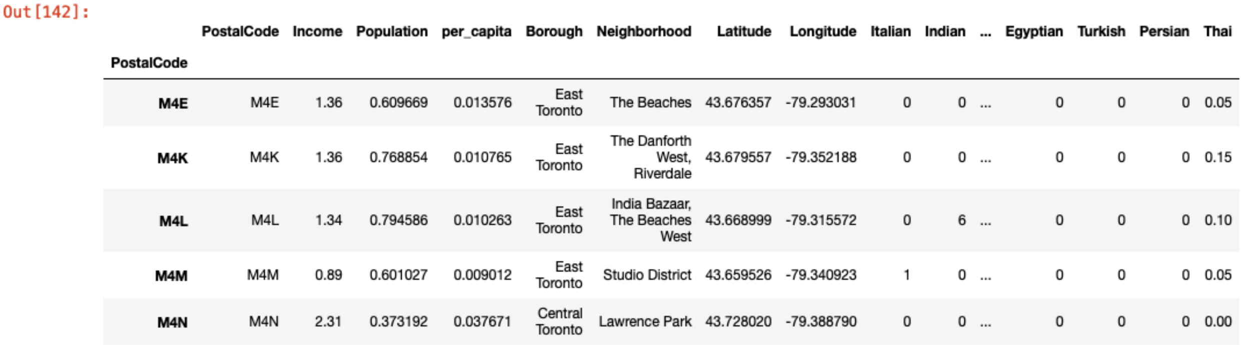
- We normalize all metrics



Out[142]:

Borough	Neighborhood	Latitude	Longitude	Italian	Indian	...	Egyptian	Turkish	Persian	Thai	Total	Total_minus	rest_pot	variance	metric	Cluster
East Toronto	The Beaches	43.676357	-79.293031	0	0	...	0	0	0	0.05	2.05	0.037963	0.351594	0.022638	2.035441	0
East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	0	0	...	0	0	0	0.15	11.15	0.206481	0.051259	0.205684	2.243044	0
East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572	0	6	...	0	0	0	0.10	6.10	0.112963	0.089327	0.117242	2.124382	0
East Toronto	Studio District	43.659526	-79.340923	1	0	...	0	0	0	0.05	1.05	0.019444	0.455677	0.014939	2.100100	0
Central Toronto	Lawrence Park	43.728020	-79.388790	0	0	...	0	0	0	0.00	2.00	0.037037	1.000000	0.051948	2.499849	0

Figure 4.2. Main data frame part 1.



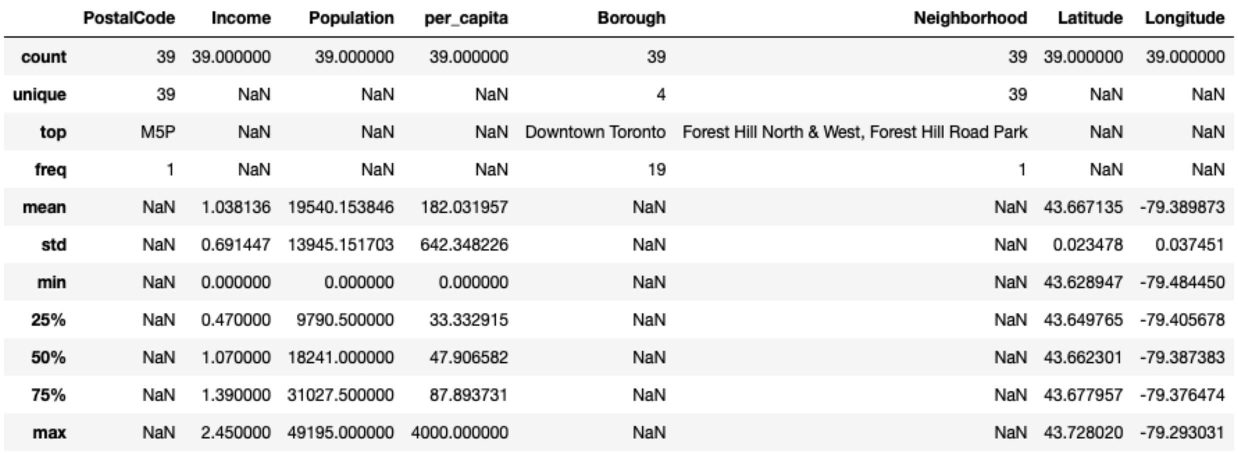
Out[142]:

	PostalCode	Income	Population	per_capita	Borough	Neighborhood	Latitude	Longitude	Italian	Indian	...	Egyptian	Turkish	Persian	Thai
PostalCode															
M4E	M4E	1.36	0.609669	0.013576	East Toronto	The Beaches	43.676357	-79.293031	0	0	...	0	0	0	0.05
M4K	M4K	1.36	0.768854	0.010765	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	0	0	...	0	0	0	0.15
M4L	M4L	1.34	0.794586	0.010263	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572	0	6	...	0	0	0	0.10
M4M	M4M	0.89	0.601027	0.009012	East Toronto	Studio District	43.659526	-79.340923	1	0	...	0	0	0	0.05
M4N	M4N	2.31	0.373192	0.037671	Central Toronto	Lawrence Park	43.728020	-79.388790	0	0	...	0	0	0	0.00

Figure 4.3. Main data frame part 2.

The header of this final normalized data frame is shown above. Notice there is an additional column called "Cluster" that will make sense when we discuss K-means clustering below. The rest of the columns are self-explanatory. Columns that will be used as parameters are normalized. Other columns, particularly restaurant counts and coordinates are left as is.

The table below summarizes some of the main statistics of numerical columns in the main dataframe. This does not include restaurant counts. Some things to notice is that both population and per capita income have a huge spread. This leads to misleading means. For example the mean per capita income is about $182,000, which is highly unrealistic. Notice also that the population shows a huge spread, with some postal codes having no resident population at all. Means would have been meaningful if postal codes had equal populations, or even close. A much better indicator is thus median. The median income per capita is about $48,000, which is a lot more reasonable. The median population is about 18,200. These statistics indicate that some postal codes are anomalous.



	PostalCode	Income	Population	per_capita	Borough	Neighborhood	Latitude	Longitude
count	39	39.000000	39.000000	39.000000	39	39	39.000000	39.000000
unique	39	NaN	NaN	NaN	4	39	NaN	NaN
top	M5P	NaN	NaN	NaN	Downtown Toronto	Forest Hill North & West, Forest Hill Road Park	NaN	NaN
freq	1	NaN	NaN	NaN	19	1	NaN	NaN
mean	NaN	1.038136	19540.153846	182.031957	NaN	NaN	43.667135	-79.389873
std	NaN	0.691447	13945.151703	642.348226	NaN	NaN	0.023478	0.037451
min	NaN	0.000000	0.000000	0.000000	NaN	NaN	43.628947	-79.484450
25%	NaN	0.470000	9790.500000	33.332915	NaN	NaN	43.649765	-79.405678
50%	NaN	1.070000	18241.000000	47.906582	NaN	NaN	43.662301	-79.387383
75%	NaN	1.390000	31027.500000	87.893731	NaN	NaN	43.677957	-79.376474
max	NaN	2.450000	49195.000000	4000.000000	NaN	NaN	43.728020	-79.293031

Figure 4.4. Statistics of demographic and economic data.

It turns out that some of the postal codes are special. They represent either government buildings or parks. As such, they do not have a resident population, or have a very small population. Even with a very limited taxable income, these postal codes will have huge per capita incomes. This will lead to incorrect normalization and will incorrectly bias machine learning outcomes. This is especially true given these postal codes cannot be zoned for restaurants. We will thus remove all these postal codes. Foursquare data also indicates there are zero instances of certain world cuisines in all postal codes. We will remove these cuisines both from affecting other restaurant types and from being potentials themselves. The fact that these cuisines do not exist will cause the algorithms to highly recommend opening them everywhere, but their absence can be caused by a particularly high barrier to entry for the specific kind of food. This requires a separate analysis.

We then proceed to do three things with the data: K-means clustering, linear regression, and total metric calculation. We will discuss the details of each analysis below.

K-means clustering: The number of clusters will vary depending on the target cuisine. Cuisines with a more diverse distribution require more clusters to properly represent categories of potential, while highly saturated cuisines can be represented in as few as a couple of clusters. To determine the number of clusters to use, we use the silhouette score method [5]. The silhouette score is often used as a visual representation method, but it can also be used to determine the suitability of the value of K by sweeping K and finding the K that achieves the maximum score.

We also applied built-in normalization from sklearn to normalize the features before they are used. We used all six features in Table 4.1 as inputs to clustering. This includes the number of restaurants from our given cuisine but represented as (1-df_merged[tcuisine]) to indicate its negative influence. The seven features allow us to cluster similar neighborhoods together.

Total metric calculation: From our custom normalization, we have six features each ranging from 0 to 1 in Table 4.1. We calculate their sum as a total metric. This total metric represents the readiness of the location to host restaurants from our particular cuisine. It summarizes positive influences that increase potential to generate customers, interest, and diverse tastes; as well as negative factors mainly related to saturation.

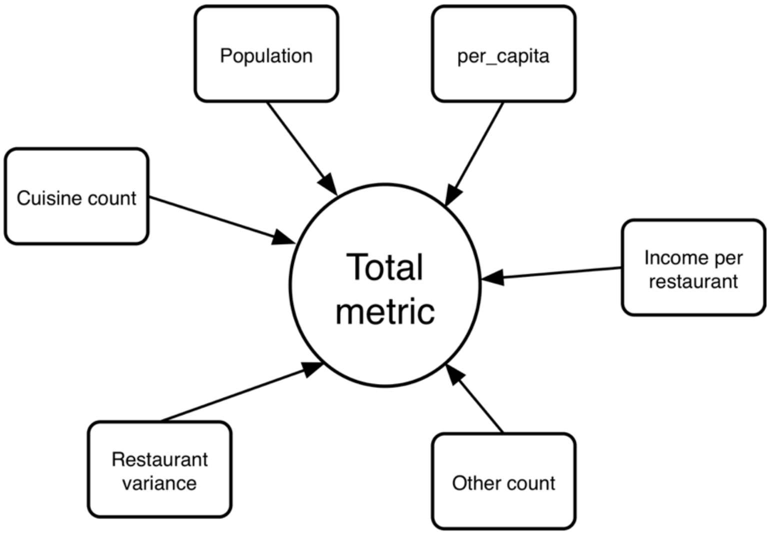


Figure 4.5. Features affecting the total metric.

The total metric (figure 4.5) is a unitless floating number ranging from 0.0 to 6.0. A postal code with a total score of 0.0 has absolutely no potential for success. It is saturated with our cuisine, has no diversity, no generated value, no population to spend such value, and no restaurants culture. A location with a score of 6.0 is at the other extreme. It is particularly ready to welcome our regional cuisine, has a vibrant restaurant culture, culinary diversity, and a large wealthy population.

Notice that the normalization we have here is within-set. This means that a score of 1.0 in any of the metrics does not mean an absolute best value, it means the best value in our dataset. A postal code that scores 6.0 will thus be the best in all categories among all the locations we considered, not among all locations in the world.

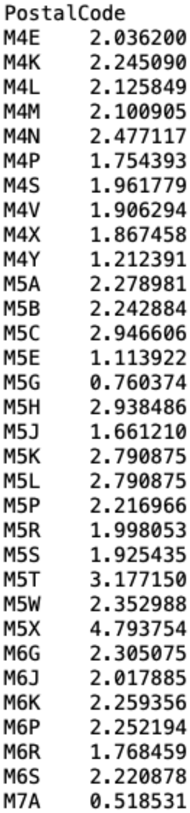
Linear regression and excess opportunity: We will consider a multivariate linear model where all the parameters in Table 4.1 (except tcuisine) are considered equal weight features. The label of the problem is the number of restaurants from the particular cuisine in the neighborhood. We fit the linear regression model with all the available postal codes as a training set. We then use them again as a test set. There are two questions here: Why choose a linear model? And will this not expose us to out of sample error?

To answer the first question, we have to look at some preliminary results in the next section. But even before seeing results, the aim of this model is not just to fit the points. In fact, overfitting would be the worst thing we can do here. What we aim to do is find a linear model that predits the number of restaurants from a particular cuisine that should exist in the current location, then from this we calculate the error, which is the only thing we care about. We will stil have to look at the error to detect any trends that indicate the need for a polynomial or otherwise nonlinear model.

The answer to the second question is related to the first question. We are not interested in predicting new values as much as we are interested in finding the in-sample error. In fact, there is no extraneous test data that will ever be present in this situation. What we want to do is find the absolute error as: the number of tcuisine restaurants in the postal code – number of tcuisine restaurants the model predicts for an identical postal code. A positive error means the postal code has more restaurants than the model would justify. This could be a sign of saturation. A negative error on the other hand, could indicate thirst for this particular cuisine in this particular location.

5. Results

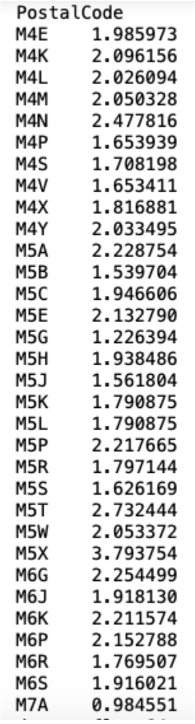
We ran the script for two target cuisines "Thai" and "German". The two were chosen based on the preliminary Foursquare results because they each represent one extreme. Thai is an extremely popular cuisine found in almost every postal code and in large numbers. German cuisine is very rare. Thus, it is interesting to see how the results will vary for each. Notice that results will be very different for each of the targets.



```
PostalCode
M4E    2.036200
M4K    2.245090
M4L    2.125849
M4M    2.100905
M4N    2.477117
M4P    1.754393
M4S    1.961779
M4V    1.906294
M4X    1.867458
M4Y    1.212391
M5A    2.278981
M5B    2.242884
M5C    2.946606
M5E    1.113922
M5G    0.760374
M5H    2.938486
M5J    1.661210
M5K    2.790875
M5L    2.790875
M5P    2.216966
M5R    1.998053
M5S    1.925435
M5T    3.177150
M5W    2.352988
M5X    4.793754
M6G    2.305075
M6J    2.017885
M6K    2.259356
M6P    2.252194
M6R    1.768459
M6S    2.220878
M7A    0.518531
```

Figure 5.1. Total metrics for each postal code with a target of German cuisine.

Figure 5.1 shows the total metric values for each of the postal codes. Most of the postcodes have a moderate metric in the 1-3 range. This is a good indicator that this metric measures something meaningful. Large variances would indicate the metric is somehow biased. Common sense dictates that for districts that get hit on a certain feature, another should carry the total metric so that we end up with a reasonable total metric. The mean score is 2.16, which is again fairly reasonable. This means should inform our judgement of total scores. Districts with lower than average totals should certainly be questioned, although not necessarily dismissed out of hand.



```
PostalCode
M4E    1.985973
M4K    2.096156
M4L    2.026094
M4M    2.050328
M4N    2.477816
M4P    1.653939
M4S    1.708198
M4V    1.653411
M4X    1.816881
M4Y    2.033495
M5A    2.228754
M5B    1.539704
M5C    1.946606
M5E    2.132790
M5G    1.226394
M5H    1.938486
M5J    1.561804
M5K    1.790875
M5L    1.790875
M5P    2.217665
M5R    1.797144
M5S    1.626169
M5T    2.732444
M5W    2.053372
M5X    3.793754
M6G    2.254499
M6J    1.918130
M6K    2.211574
M6P    2.152788
M6R    1.769507
M6S    1.916021
M7A    0.984551
```

Figure 5.2. Total metric with a target of Thai restaurants.

Figure 5.2 shows the total metric calculation with a target cuisine of Thai. Notice that we need to recalculate because the balance of target and other restaurants is changed. Variance is lower for this target and most locations score close to the mean.

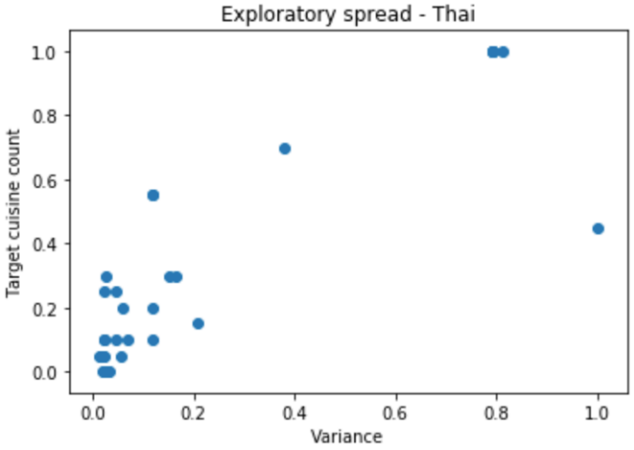


Figure 5.3. Spread of number of Thai restaurants versus variance of cuisines in the postal codes.

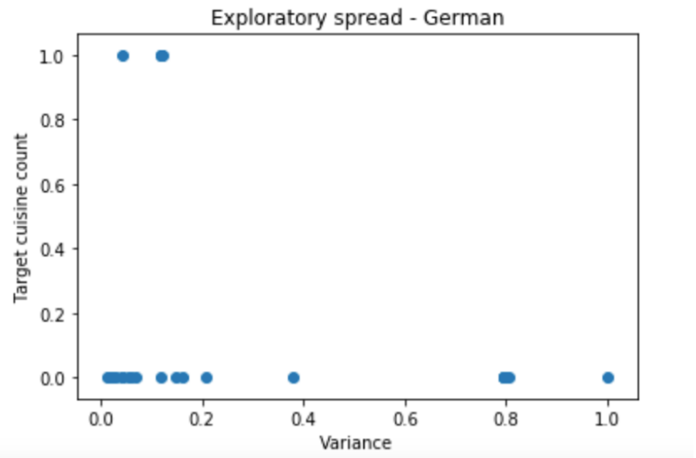


Figure 5.4. Spread of number of German restaurants versus variance of cuisines in the postal codes.

Figures 5.3 and 5.4 show exploratory plots for correlation of number of target cuisine restaurants versus the remaining five features from table 4.1. This was repeated for a number of features and a number of targets.

Our third set of results have to do with K-means clustering. We apply the algorithm to cluster postal codes according to their suitability for starting the restaurant. All six features in table 4.1 are used. We also apply the built-in normalization function from sklearn to prepare the data.

One major step to prepare for clustering is determining the number of clusters to use. As discussed in section 4, we use the silhouette score for this.

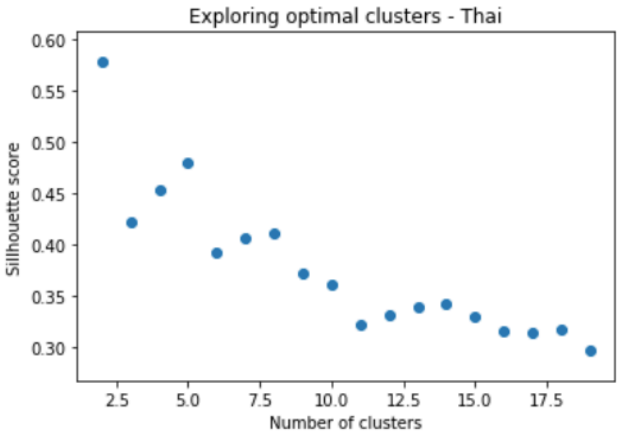


Figure 5.5. Silhouette score vs. number of clusters for Thai restaurants.

Figures 5.5 and 5.6 show the results of this sweep. For Thai restaurants, the optimal number of clusters is 2, for German restaurants, it is 4. The main takeaway here is that there **is** a difference in the optimal number of clusters based on target cuisine, and that the difference will not be minor by any measure.

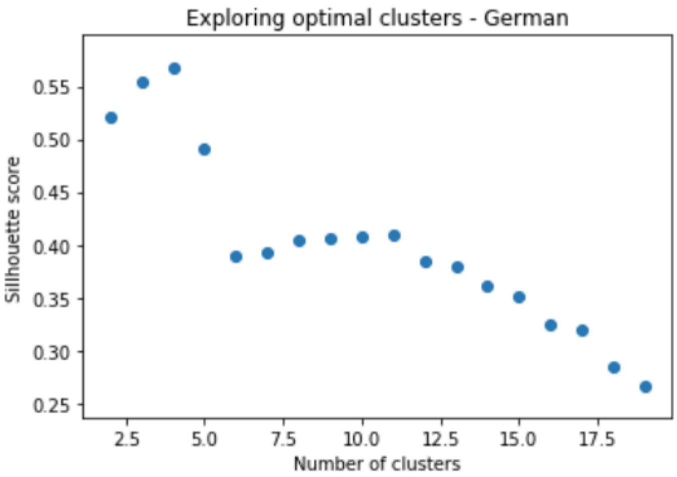


Figure 5.6. Silhouette score vs. number of clusters for German restaurants.

-Kmeans label results
-K means visualization on map

6. Discussion

The total metric is an interesting idea. It assigns a total numerical value to the different aspects that affect the chances of a restaurant by postal code. However, we need to be a little careful interpreting the results. The lowest total score from figure 5.1 is for postal code M7A at 0.52. M7A is discovery district. This is a relatively high income postal code, however, it scores so low mainly based on the lack of a resident population and taxable income reported from district. This is a good indication of lack of fine dining activity in the area, but it also misses the fact that a lot of non-residents may visit the location. Consult section 7 for ways future improvements can be made to account for this.

The highest total score is for postal code M5X at 4.79. This is the financial district and it exhibits the highest potential in several areas, including income per capita and number of restaurants. The district benefits from a small resident population in calculating a large per capita income, but is penalized for it in the total population column.

In table 5.2, all metrics are recalculated for a target cuisine of Thai. The range of numbers shrinks noticeably. This makes sense because there is a much larger proliferation of Thai restaurants. This reduces the number of districts with a vacuum of target restaurants and lowers high scores. The highest score is still code M5X at 3.79 which is a great sanity check for the sensibility of the total metric. The lowest score is also still M7A at 0.98. Notice how the range has been reduced from both ends.

Figure 5.3 shows a scatter plot of the number of Thai restaurants versus the variance of restaurants in the different postal codes. Figure 5.4 recreates this for German restaurants. This is part of initial exploration of linear regression as a model. Figure 5.3 shows the kind of dependence we can expect from linear regression. Figure 5.4 is definitely bad, but this is clearly due to the dearth of German restaurants in Toronto. We repeated this for the other features, showing that as the sample size grows linear regression will start to work. The ultimate judgement though has to come from error analysis.

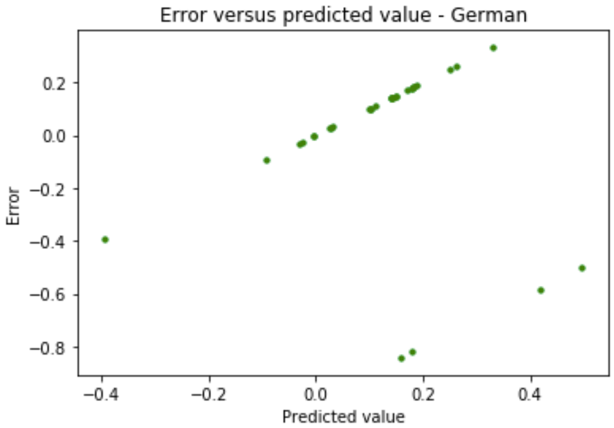


Figure 6.1. Error versus predicted value for German restaurants.

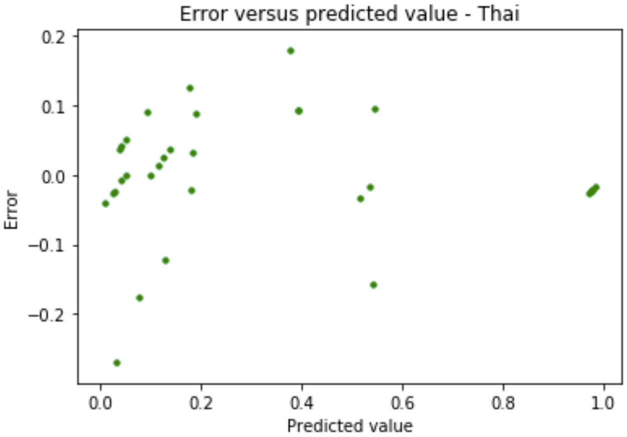


Figure 6.2. Error versus predicted value for Thai restaurants.

Figures 6.1 and 6.2 show error on the y-axis versus the predicted value given by the linear regression model on the x-axis. We are not able to plot versus the feature since this is a multi-feature problem. Figure 6.2 shows a zero mean more or less normally distributed error, suggesting the linear model is sufficient. Again, the rarity of German restaurants makes the pattern seen in figure 6.1 non-indicative.

The average error per postal code for Thai restaurants is 4.8e-17, which is null for all intents and purposes. This proves the error is zero mean. The maximum error is 0.178 for postal code M4Y, the minimum error is -0.268 for M6S. Notice that the label used in fitting is normalized to a maximum value of 1, thus these errors are actually significant. For M4Y, the error indicates the model expects a larger number of Thai restaurants than in reality, indicating potential. M4Y is a diverse, wealthy, and LGBT friendly neighborhood with a good dining culture. M6S is Bloor West Village, which is a highly saturated shopping district, as indicated by the large negative score.

For German restaurants as a target, the average error is -1.38e-17, again showing effective zero mean error despite the small samples of the rare cuisine. The maximum error is 0.332 at M5W, and the minimum is -0.841 for M4Y. This is interesting because it shows the model could record excess potential for one type of cuisine, and a deficit of potential for another cuisine in the same postal code, showing sensitivity to the type of cuisine.

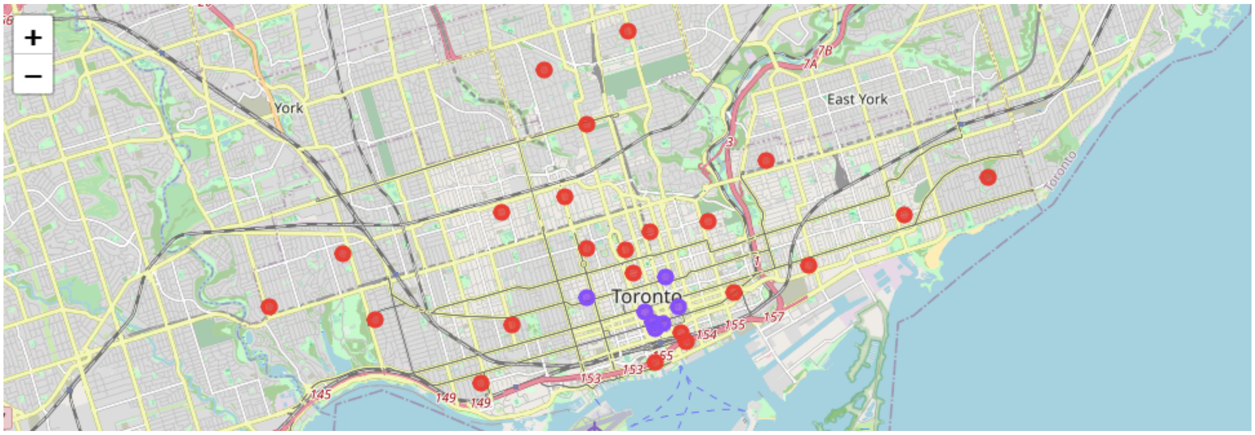


Figure 6.3. Clusters for opening a Thai restaurant. The blue cluster is more favorable.

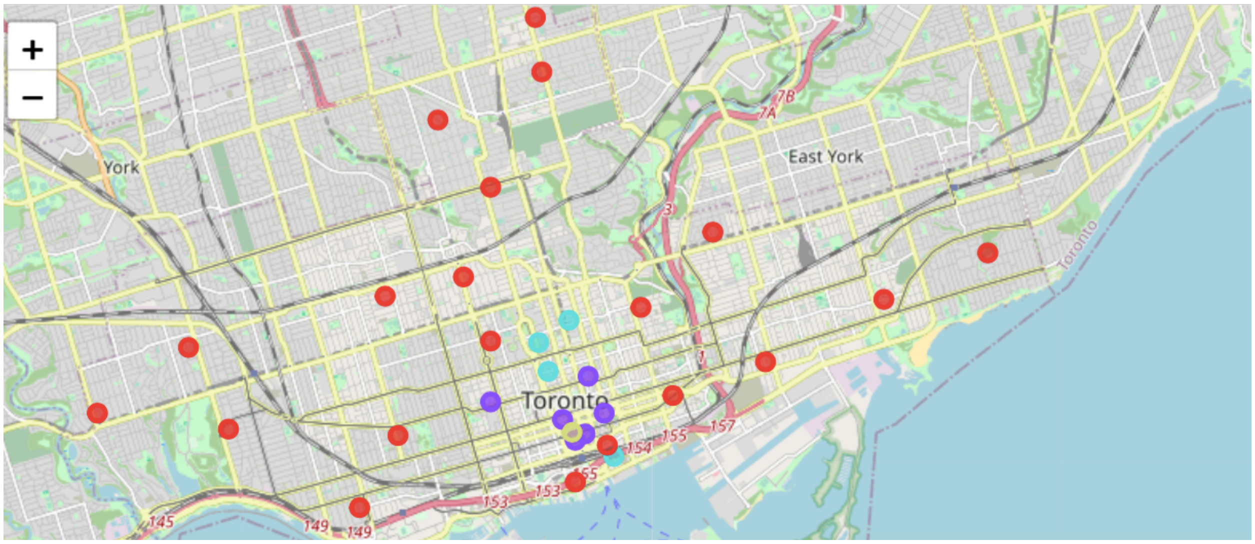


Figure 6.4. Clusters for opening a German restaurant.

Figures 6.3 and 6.4 show the results of clustering for Thai and German restaurants respectively. Recall from section 5 that the optimal number of clusters for Thai restaurants is 2. All postal codes are clustered either in the blue group or the red group. Blue dots represent more dense, urban, and possibly high potential locations, while red dots represent more suburban locations. Notice that these clusters are not simply classifying locations by their similarity, the way the features are designed, the clusters are highly biased towards potential for Thai restaurants.

German restaurants are clustered into four groups. This has the effect of introducing more distinction to the downtown clusters as well as classifying semi-urban postal codes separately from more suburban locations. In general, cuisines with a small presence such as Persian and Brazilian tend to favor a larger number of clusters, while more common cuisines such as Italian and Mexican favor two clusters. This is an indication of the potential of cuisines with a small presence. They have more space to entry and thus, we can distinguish more groups of postal codes.

7. Conclusions

Opening a restaurant is a risky business, but does the type of cuisine we serve have an impact on chances of success? The above discussion certainly shows that it at least should inform our choice of location. In three different ways, we showed that the type of cuisine greatly changes which locations are more favorable, and even which are similar. Through judicious choice of metrics, we can calculate a total score and rank different postal codes by their suitability. Using linear regression, we can get a surprisingly well-behaved multi-feature model. This allows us to predict the number of restaurants from a particular cuisine in a postal code with specific demographic, economic, and venue data. This can then be used to calculate an excess or deficit potential. While clustering, we see that the kind of cuisine not only affects how postal codes cluster, but even the optimal number of clusters. We do notice, though, that suburban locations tend to cluster together regardless of the cuisine.

There are several areas of improvement for this work, some have to do with data collection, but some also have to do with the model and methodology. We summarize them below:

- The highest priority is probably translating the combined metric into something which is easily interpretable. The value of this metric currently is that it compares apples to apples. But this is not enough. Using data fitting (possibly even regression) and diligent normalization, we may be able to translate the total metric into a chance of survival at one, two, three, and four years. The most challenging part of this is obtaining enough data about records of restaurant closures by postal code. This will also require differential weighing of the parameters. Here, we normalized and allowed all factors to impact the cluster equally. As a general rule, different factors will have different levels of impact, and equal weights may not be optimal
- The second most useful addition would be to adapt the setup for other cities, possibly also allow it to perform its normalization on multiple cities. This would be particularly useful in large conurbations such as Southern California. The main challenge again will be obtaining data in a systematic and automated manner. So far, we have to point the software to data ourselves. In countries with publicly available data, the process can be automated
- Third is an improvement to the clustering algorithm. Namely, we have to relate similar cuisines to each other. For example, while Korean, Chinese, and Japanese cuisines are

distinct; it is not unfair to think that the presence of one may provide competition to the other. This issue is a data analysis problem of its own and will require a side project to generate correlation matrices between different cuisines. This is especially true because some cuisines may be similar without even sounding like they should be similar

- The total metric (and clustering) have one factor missing: transport. Some locations will display low buying power, but this is not because they have low potential, it is due to the nature of the location. However, many such locations will attract a lot of visitors from surrounding areas
- Postal code is a great key field. It allows us to merge multiple data frames consistently. However, it does not indicate any meaningful information about area. Some urban postal codes are only a block in size, some suburban postal codes are the size of a town. This can greatly skew our results because our model assumes that each postal code is self-contained. This is not true, close and tightly packed postal codes often interfere with and affect each other. A model that takes into consideration the size of the postal code (in terms of area) can help address this

References

[1] https://www.getorderly.com/blog/high-restaurant-failure-rate

[2] https://smallbusiness.chron.com/successful-profit-margins-restaurant-business-23578.html

[3] https://www.cnbc.com/2016/01/20/heres-the-real-reason-why-most-restaurants-fail.html

[4] https://www.forbes.com/sites/modeledbehavior/2017/01/29/no-most-restaurants-dont-fail-in-the-first-year/#21742fc04fcc

[5] https://en.wikipedia.org/wiki/Silhouette_(clustering)