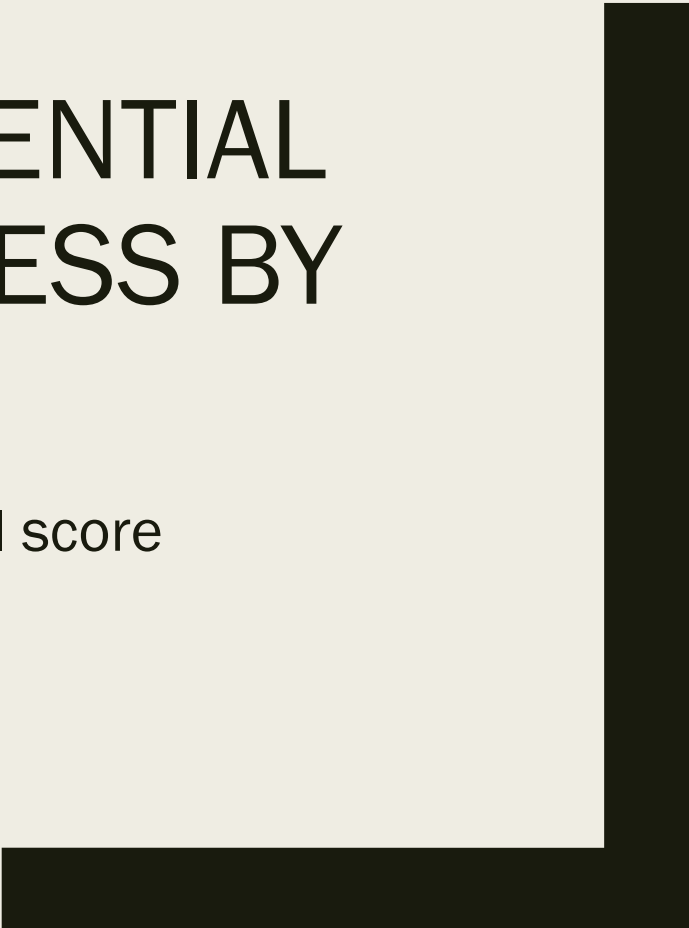




# DETERMINING THE POTENTIAL FOR RESTAURANT SUCCESS BY CUISINE

Regression, clustering, and a novel total score



# Background – should you or should you not?

- Starting a new restaurant business is extremely risky, anecdotally, failure rates are really high
  - *Contrary to popular belief, large chain restaurants have very low mortality*
- It is the small startup restaurant that seems to fail at an alarming rate
- And it is these small startups that need the most amount of help

# Location location location ... and cuisine?

- Location is the begin all of determining if a restaurant will survive
- But does the type of cuisine served also have an impact?
- We develop a model that helps a new startup restaurant determine which location or group of locations are most receptive to their particular cuisine
- We deliver three ways to do this:
  - *K-means clustering to classify similar locations together*
  - *Linear regression, allowing the model to predict the number of restaurants that **should** be in the location*
  - *A total score or metric, indicating the suitability of the location*

# Raw data

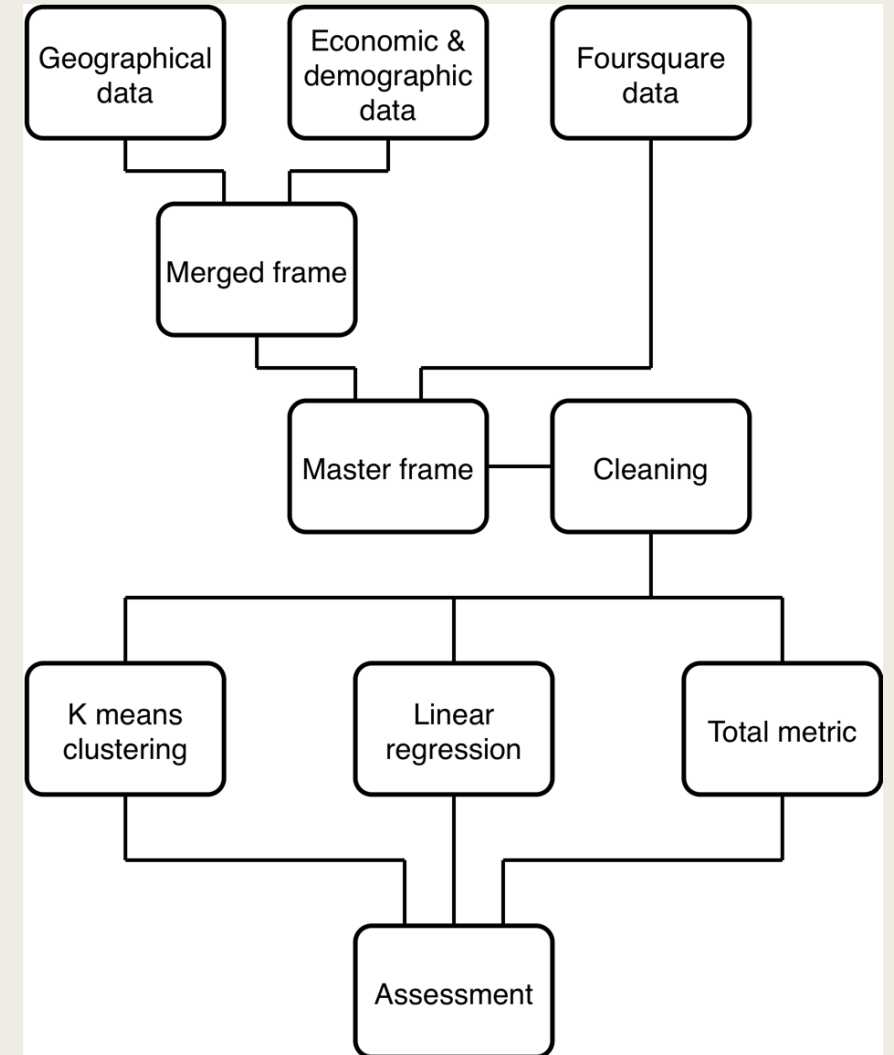
- Data on postal codes, neighborhoods, and boroughs in Toronto
- Demographic data on population by postal code
- Economic data on total taxable income in every postal code
- Foursquare data:
  - *Total number of restaurants from a pool of world cuisines in every postal code*
  - *The cuisines are: Italian, Indian, Chinese, Japanese, Korean, French, Greek, Brazilian, Mexican, Spanish, German, Moroccan, Turkish, Persian, and Thai*

# Derived data and main data frame

- Six metrics are derived to characterize every postal code for the target cuisine:
  - *Total population of the postal code*
  - *Per capita taxable income, representing purchasing power*
  - *Taxable income per restaurant per capita*
  - *Variance of restaurants among different world cuisines*
  - *Number of restaurants not including our target cuisine*
  - *Number of restaurants from our target cuisine*

# Approaching the problem

- After the data is collected and cleaned it can be used in all three analysis methods
- Information from all three can then be used to make an assessment



# Normalization

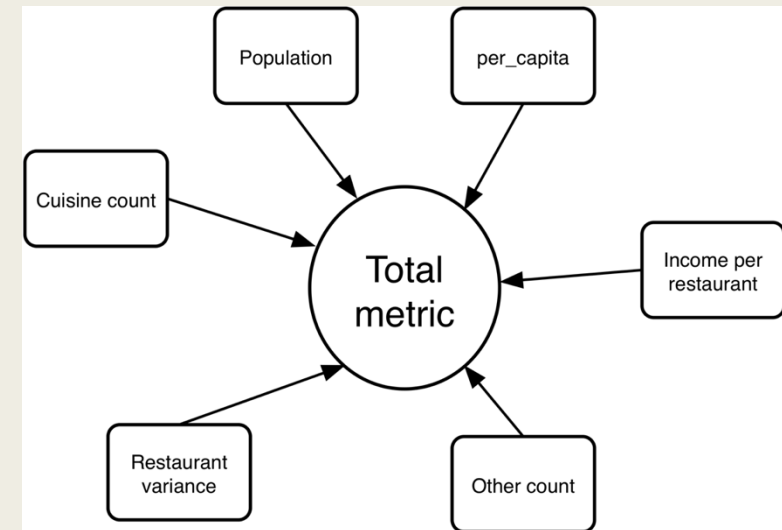
Out [142]:

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

- The six features are normalized to a range between 0 and 1
- The metric for the number of restaurants from target cuisine is normalized to drop as the number of restaurants rises

# The total metric

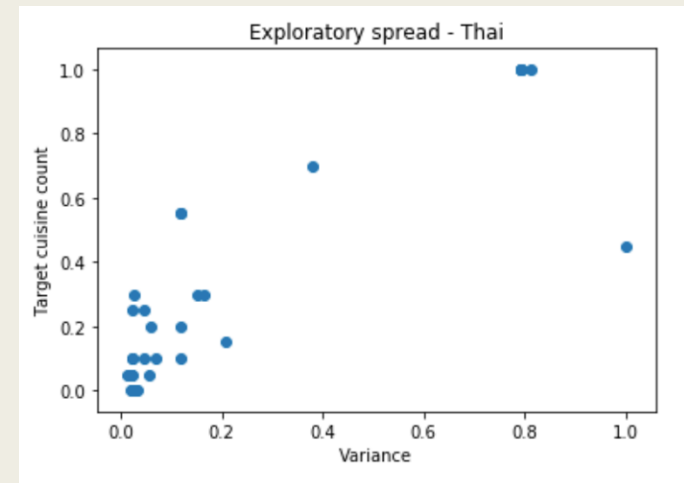
- This is the sum of normalized metrics
  - *0 for worst fit, 6 for best*
- We can confirm the metric changes with target cuisine
- For Thai restaurants, the metric points to very wealthy high finance districts
- For German restaurants (less common) it points to locations with high diversity





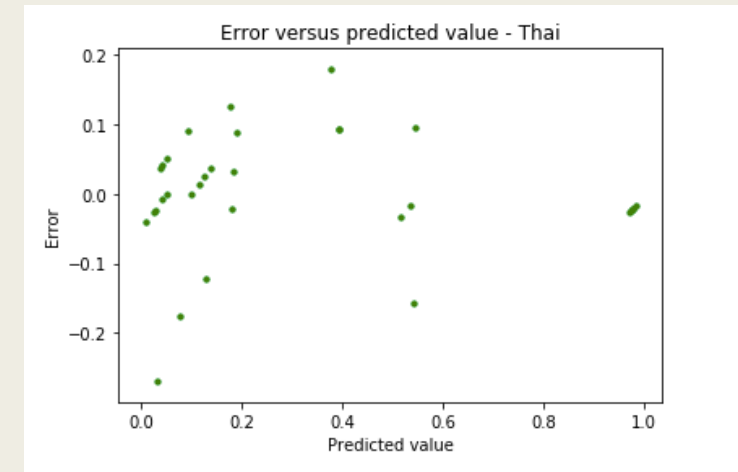
# Linear regression and excess error

- Consider target restaurant count to be the label
- The remaining five variables are features
- Could we fit a model to predict restaurant count?
- Plotting restaurant count against most of the features shows good correlation
  - *There is potential*



# Results of regression

- Linear regression with five features
- Error is zero mean regardless of target cuisine
- Model varies greatly depending on target



# Interpreting regression results

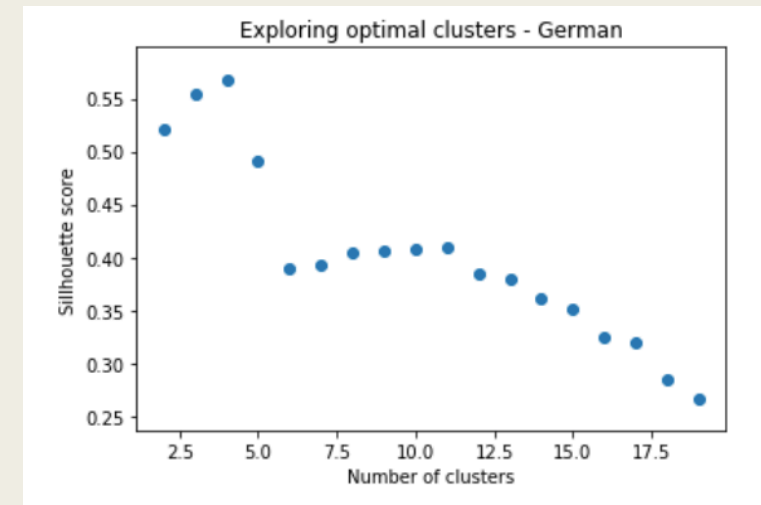
- We find the “error” for each postal code based on the model
  - *A positive error indicates the model predicted more target restaurants*
  - *A negative error indicates the model predicted less target restaurants*
- In the report we look at the extremes of errors, where a positive error could insinuate missed opportunity and a negative error could indicate saturation
- We also show that the distribution of errors is highly dependent on the target cuisine

# Choosing optimal clusters

- We use the silhouette score to find the optimal number of clusters
- The silhouette score measures how close an element is to its cluster versus how far it is from other clusters
- The score varies between -1 and 1, with 1 being the best
- It is a good idea to choose number of clusters so that the silhouette score is maximized

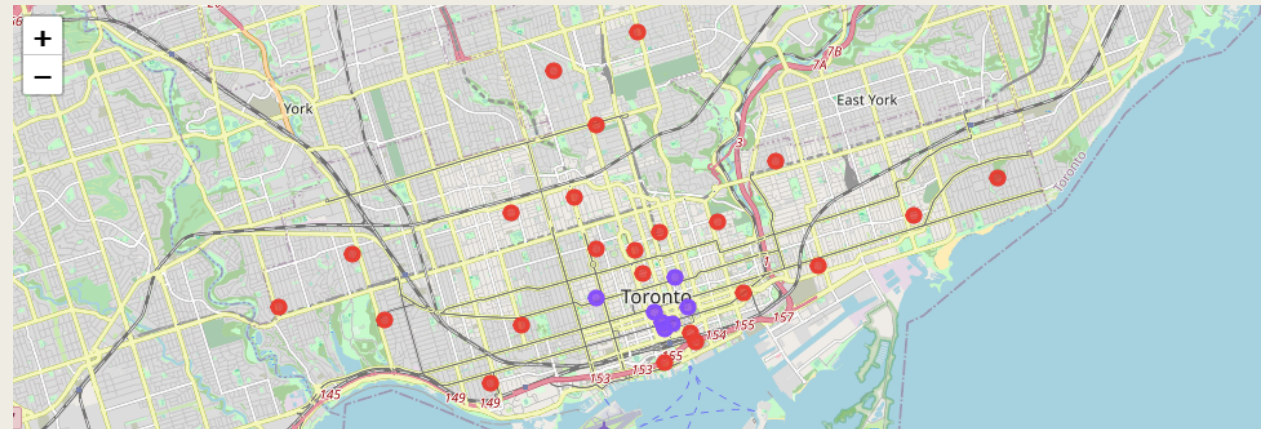
# Results of silhouette simulation

- It is a computationally intensive process because it loops through multiple clustering problems
- We ran it for all target cuisines
- In general, popular cuisines are best clustered into a small number of clusters (2)
- Less common restaurants need to be clustered into more clusters (German restaurants need 4 clusters)



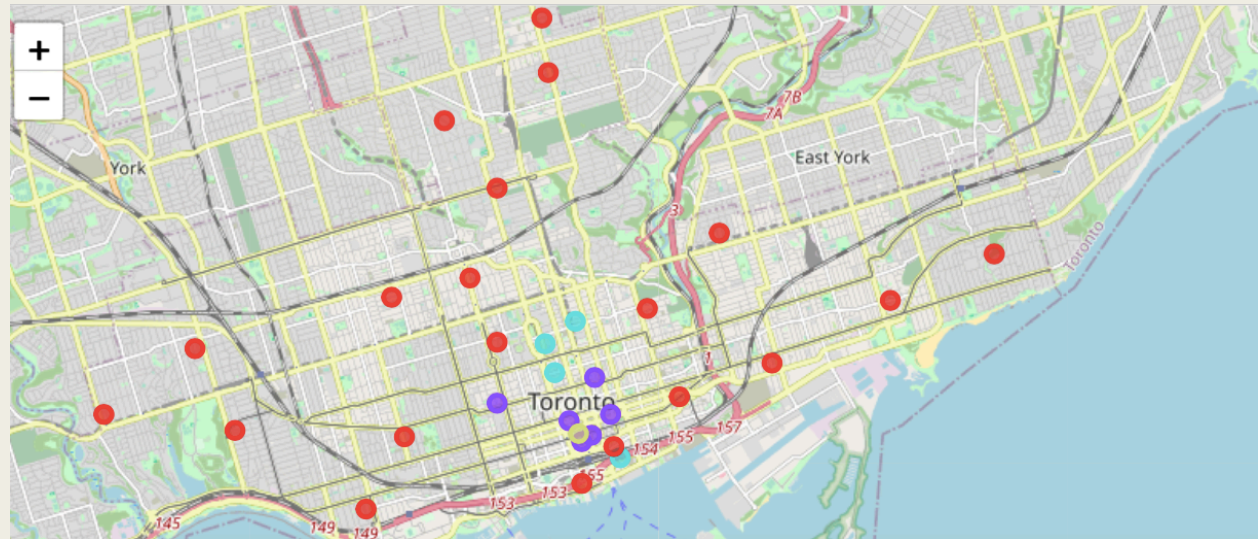
# Where to start a Thai restaurant?

- Toronto is awash with Thai restaurants
- There are only two types of places to start the restaurant
- The blue cluster is high competition high potential



# Where to start a German restaurant?

- This is a fairly uncommon cuisine
- There are four different clusters in close geographic proximity
- This cuisine is more location sensitive



# Conclusions

- Clustering is highly sensitive to target cuisine
- Linear regression varies wildly based on target cuisine
- The total metric score will vary substantially based on target cuisine
- This combined means that the choice of location must be informed by the target cuisine and its interaction with other cuisines
- Our three methods to aid the client give non-conflicting and complementary information
  - *You can use clustering, followed by linear regression, followed by scoring to choose a location*



# Caveats and outliers

- We considered central Toronto for our data set
- There are four postal codes that are particularly troublesome
  - *They have an extremely small resident population*
  - *An artificially high per capita income*
- These locations greatly distort the results
- They are also non-residential and non-commercial
  - *Mostly government and park districts*

# Areas for improvement

- Extend the setup to other geographic locations
  - *Main challenge is data acquisition*
- Consider the interaction of close neighborhoods through commutes and transport
  - *This is particularly true for locations with low resident population, but high dining culture*
- Normalize the total metric so that it corresponds to chances of success
- Study the interaction of related cuisines and their impact on each other's potential for success