IBM Data Science Capstone: Final Exam – Week 2 – Report

Using Clustering to Target Best Choices for a New Store-Front

Introduction – Business Problem

Companies opening up new store fronts, have to deal with a myriad of issues: business licenses, regulatory requirements, sales and business taxes, competitors, and of course the biggest item: location, location, location. Positioning a store front in an advantageous location is one of, if not the, most important, key factor in a business's ability to survive and thrive. The business problem this analysis is focused on, is to assist new business owners with finding the best set of potential locations to maximize their chances of success when opening a new store front. The specific "client" that this study will address, is a business owner that wants to open a new upscale Steakhouse restaurant in the Toronto, Canada area.

The overall strategy will be to find the best potential upscale neighborhoods, in combination with the lowest number of existing competitors. As part of the analysis, the total number of other nearby venues are factored in, which can act as additional "magnets" to entice more customers and foot traffic into the area that the new business will operate in.

Both a text summary of the best potential neighborhoods, as well as a graphic Folium map showing the recommended potential neighborhoods are produced by the project.

Who Would be Interested

The typical client for this kind of service is any business that wants to open a new store-front in any of Toronto's neighborhoods. The customer is able to get a list of the top 10 neighborhoods that would have the best desired "draw" characteristics, as well as good potential (under-served areas) for the target business.

While this specific project focuses on selecting Steakhouse restaurant locations, any type of store front business (not just restaurants) can be evaluated, by changing the target venue of the new business being analyzed.

Data Used in this Study

This report focuses on the city of Toronto Canada. The datasets used in this study were downloaded from the various "Open Data" Portals that the city of Toronto provides to online researchers, as well as data generated and collected from Nominatim (latitude and longitude), and Foursquare Labs (venues).

The following specific datasets were used:

City of Toronto:

A large (multi-megabyte) XLSX list of neighborhoods and associated housing data from Toronto's "Wellbeing Toronto - Housing" website at their Open Data Portal at URL:
https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#0ee5007f-7c8b-5107-7fa8-24de3ae06f22
was used as a primary source.

That file contains all of the Toronto neighborhood names, housing demographics, and population demographics broken down neighborhood by neighborhood.

That XLSX dataset was downloaded and converted to CSV format into a local directory. The resulting local CSV file was later loaded into a Pandas "Neighborhoods" DataFrame, that acted as one of the primary datasets in the study.

Nominatim:

The Nominatim geocode() network API was used to obtain the latitude and longitude codes for each Toronto neighborhood, using the Neighborhood names supplied from the Toronto Neighborhood data above.

Nominatim is an online geocoding search engine for OpenStreertMap data. As part of this service, it can return machine-readable latitude and longitude information for various cities, and neighborhood districts within large cities, including Toronto

Nominatim was selected because of Google's recently increased pricing for obtaining geo-coded data. Nominatim has a policy of no more than 1 request per second (otherwise will get throttled or rejected). A 2 second timer was used between API invocations, with one neighborhood looked up per API call. The results were saved in a CSV file, that was later merged into a Python Pandas "Neighborhoods" DataFrame.

Foursquare Labs:

The Foursquare Labs venues search network API was used to obtain lists of venues for each of the Toronto neighborhoods. The neighborhood latitudes and longitudes obtained from Nominatim above, were used to identify the starting location for venues searches.

Foursquare Labs is a search and discovery service, that records and reports about all kinds of venues and recommendations for various cities, including Toronto.

The Foursqaure returned venues results were then processed to generate;

- Detailed counts of counts of current general venues in a each neighborhood

- Counts of current Steakhouse and restaurant venues, including number of direct competitors, in each of the Toronto neighborhoods.
- Aggregated counts of general venues in a given neighborhood to ascertain additional "magnets" and points of interest that would help increase foot traffic and potential "draw in" business in the area.

The collected data was be used to determine which neighborhoods exist in the Toronto area, what are the economic characteristics of each neighborhood, what competitors exist in each neighborhood, which neighborhoods might be being "under-served" (very few offerings or competitors in the area), and what "magnets" does the respective neighborhood have to draw people into the area. That information was then used to cluster rank the "Top 10" candidate neighborhoods

Methodology

The overall methodology process encompassed the following major elements:

- the Toronto city neighborhood and real-estate datasets were downloaded from the data sources listed in the previous section, that data was converted (data wrangled) into "cleaned up" form (removing any duplicates and nulls),
- the neighborhood names and city were sent to Nominatim's online API to retrieve and save the longitude and latitude for each of the 140 Toronto neighborhoods,
- the neighborhood longitude and latitude data was sent to Foursquare Labs online APIs to retrieve and save the venues associated with each neighborhood,
- a master Pandas dataframe was created using the above data,
- various exploratory analysis (histograms, scatter plots, "one hot data" encoding) were used to tease out key elements of the data,
- aggregation of key indicators (total number of venues, total restaurants, and total Steakhouse competitors) were generated,
- Kmeans clustering techniques were used against the collected data to select out the high-end neighborhoods with the largest number of venues,
- The results were then ranked, and a list of top 10 best candidates were selected, and
- a city map showing the recommended locations was generated.

Python-based Data Science tools were an integral part of analyzing the data. The associated code (in Python Notebook form) is located at the Github URL associated with this project.

Exploratory Data Analysis

The initial Toronto demographic data was loaded from the project's CSV file, the latitude and longitude information retrieved for each neighborhood, and the venues information retrieved from FourSquare Labs. A set of histograms and scatter plots were then

constructed to explore the data. Each of the key plots used in the Data Analysis phases are discussed below.

Figure 1 shows a Histogram of Real Estate pricing for all of Toronto's 140 neighborhoods. Not surprisingly, it is left skewed, with the bulk of the real estate in the 200K – 400K range, which would be the middle and upper middle class income ranges. Since we are targeting a "high-end" Steakhouse, our goals will focus on the right part of the graph, in the neighborhoods at or above the 548K mean value.

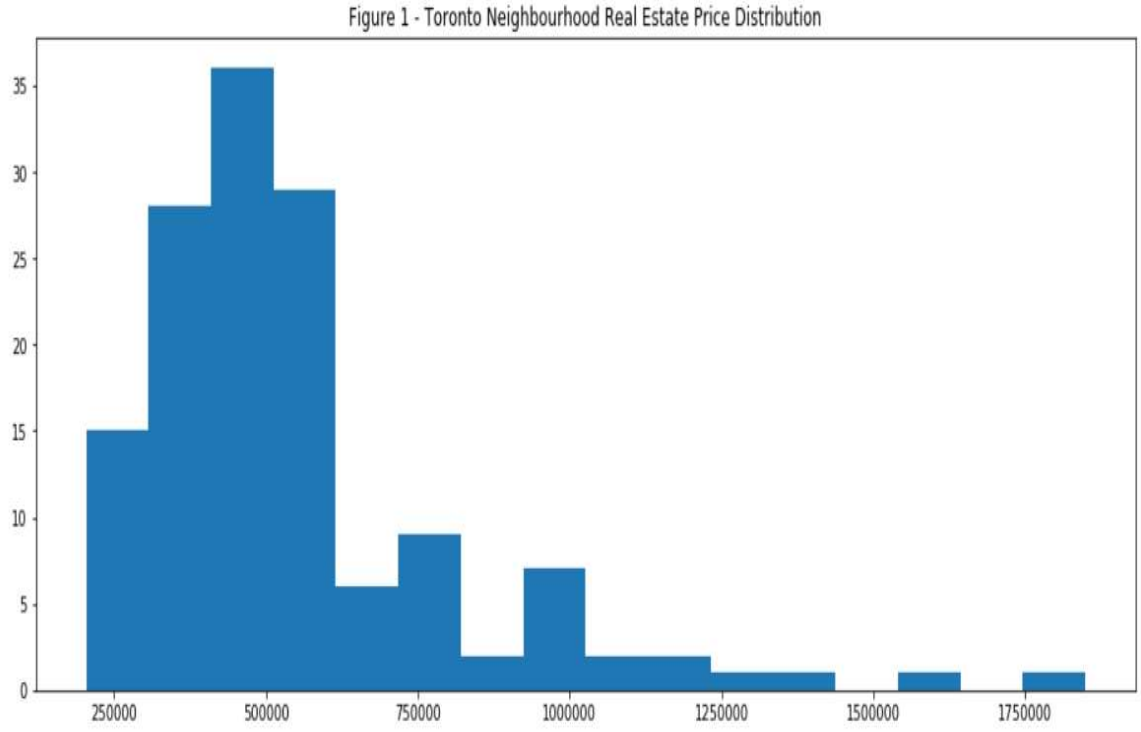


Figure 2 (below) shows a scatter plot of the number of Venues versus the associated Real Estate prices for each of the 140 neighborhoods. Looking at the chart, it clearly is not a linear plot, but rather a small set of fairly obvious (low, middle, top) clusters. A large number of the neighborhoods have 25 or less venues, which can often mean they are primarily residential, and there are zoning restrictions on the number of business allowed in those areas. For our purposes, we want to focus on the top upper band in the chart, where there are a lot of venues, and consequently, a lot of "foot traffic" acting as a magnet to the area.

This was a key chart that oriented the analysis toward using a clustering approach, since the data shows a clear clustering pattern, rather than a linear or regression oriented pattern.

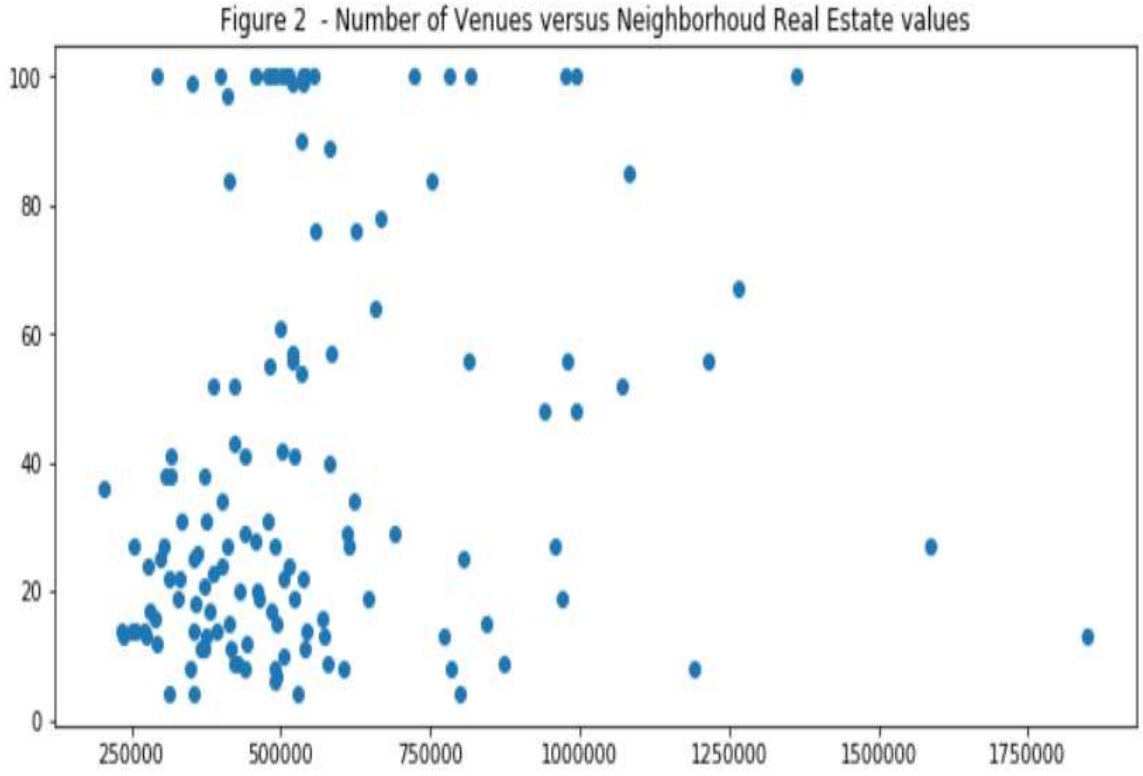
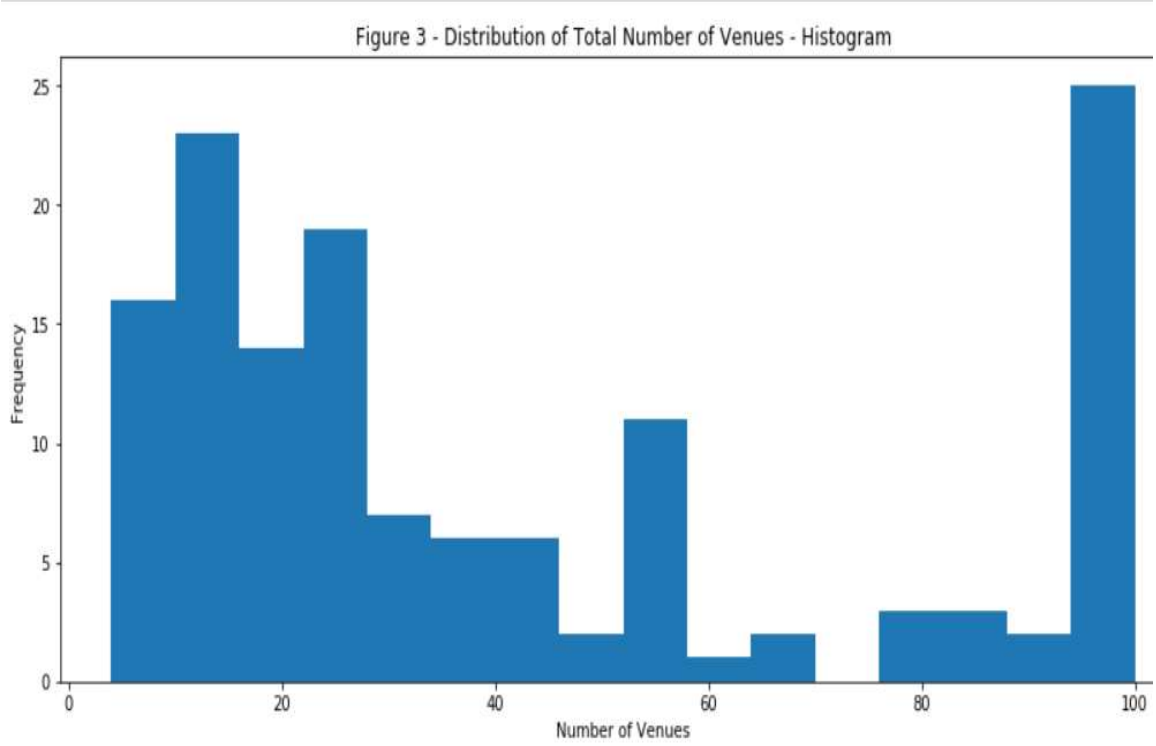


Figure 3 is a histogram showing the distribution number of venues in the various neighborhoods. Again, this is a highly skewed graph, with the bulk of the venues being skewed either to the left end (small number of venues) or to the right end (high number of venues). As mentioned above, this is probably driven by a combination of zoning restrictions in various neighborhoods, and the population and/or business mix density of a given neighborhood. Again, for a "high-end" Steakhouse, we are oriented toward the right end of the chart.



In Figure 4, the plot on the left shows the 3 primary cluster bands that were the result of the Kmeans clustering algorithm. Various clustering numbers were used from 2 to 6, but the best, and clearest results came from a clustering setting of 3 clusters. This resulted in a list of 24 candidate neighborhoods that had the best mix of venues and upper income ranges.

The plot on the right, is a “side by each” copy of the previous Figure 2 “Venues vs Neighborhood Prices” scatter plot, to show how the Kmeans clustering results matched up with the scatter plot.

Figure 4

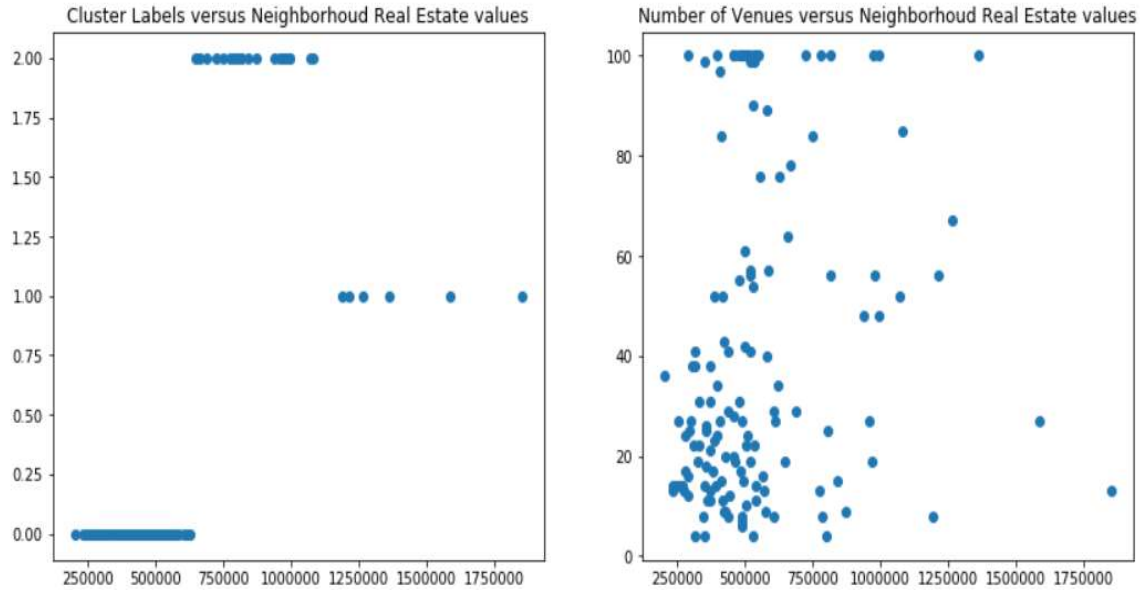


Table 1 shows the results of ranking the initial 24 candidates that were output from the Kmeans clustering step, into a final "Top 10" list of best candidates. To create the "Top 10" list, the Kmeans clustering results were first ranked by number of venues (the more the venues, the higher the ranking, since it indicates more potential foot traffic and "magnets" to pull people in), and the second pass ranking was based on the number of Streakhouse competitors in the area. The goal was to seek out areas that had the fewest number of Steakhouse competitors.

Table 1 - Final "Top 10" Neighborhoods

Neighbourhood	Home Prices	Latitude	Longitude	Venues_Count	Venue_Num_Steakhouses	Venue_Nu
North Riverdale	818592	43.665470	-79.352594	100	0	
Trinity-Bellwoods	723909	43.647627	-79.413879	100	0	
Palmerston-Little Italy	781568	43.655879	-79.410076	100	0	
Annex	993491	43.670338	-79.407117	100	0	
The Beaches	751945	43.671024	-79.296712	84	0	
Runnymede-Bloor West Village	666204	43.651778	-79.475923	78	0	
Wychwood	656868	43.682094	-79.423855	64	0	
Casa Loma	1083381	43.678111	-79.409408	85	1	
Mount Pleasant East	815247	43.708417	-79.390135	56	1	
Yonge-Eglinton	975449	43.706748	-79.398327	100	2	

Another pattern that emerged from the analysis was the relatively "tight" range of real estate pricing for the highest number of venues. Looking at Table 1 the top candidates were in the 723K to 993K range of pricing. Neighborhoods above that range typically had fewer venues, which was probably driven by a combination of zoning restrictions ("exclusive neighborhoods") and much higher rents in higher priced areas, resulting in a squeezing of business margins. Below the 723K range, many of the neighborhoods were below the 50 venues level, which would result in less foot traffic and fewer "magnets" to pull people in.

The last figure provides a map of where the best ranked candidates are located in the city of Toronto. The client can then use this to further research any zoning constraints, store-front rental pricing, and visit the candidate neighborhoods, to get a feel for the verve of the area.

Results

The results showed that data science clustering techniques can be used to help narrow down locations to locate new store fronts. As seen in the various histograms and scatter plots, there is a high degree of skew in large cities between neighborhoods.

Initial exploratory data analysis showed a “natural clustering” of venues and upper scale neighborhoods. Since our “client” was seeking to open an “upscale’ Steakhouse, we were able to use these clustering techniques to help narrow down the best candidate matches, that were summarized in the “Top 10” list in Table 1, and in the associated selected neighborhoods in the Toronto city map.

In this analysis, a “sweet spot” emerged in areas that had real-estate pricing in the 723K to 993K range, in combination with a large number (90 or more venues in a given area).

Discussion

This analysis demonstrated that neighborhoods can be analyzed and "data mined" via online tools, to help tease out key defining characteristics that are needed to target where to locate a new business. It also showed that often clustering (unsupervised learning) techniques can be used to analyze the neighborhoods under consideration.

As seen in this analysis of Toronto (and can be shown in other areas like San Francisco, Seattle, or New York), specific areas of a city act as magnets for businesses (venues), but at the same time constraints such as zoning rules (limiting the number of venues), and real-estate pricing (very high-end neighborhoods hammer margins), can be mined to locate the "sweet spots" where a new business has a better chance of survival.

The recent availability (last 10 years) of online datasets (typically CSVs) and APIs that can be used to extract needed information about a neighborhood, such as location (latitude, longitude), types of venues, real estate pricing and other demographics opens up a whole new layer of analysis and insights for an owner or investor that wants to start a new store-front business.

For now, the lack of freely available, public online detailed demographic (neighborhood by neighborhood) datasets is still an issue for most large and medium cities. This is gradually being filled in by new entrants like Zillow's new online offerings that categorize both home price, rent, and income statistics for given areas/neighborhoods (e.g. its Zillow Research arm at https://www.zillow.com/research/data) are an example of how over time, these data gaps will be filled in.

In an ideal world, the breakdown of overall retails sales or sales tax revenues by each neighborhood, would have been the most valuable piece of data, in order to target the highest revenue generating areas. While the city of Toronto has very good data portals, they currently do not provide this level of detail. Instead, relevant sales revenue and sales taxes are currently only available as aggregate summaries by city, not by neighborhood. So as an approximation, housing prices and rent data (by neighborhood) were used as a proxy to select those neighborhoods with the highest potential.

In the future, adding additional inputs, such as store-front rental prices and crime rates for each of the neighborhoods, would be a useful addition to the selection process. Some form of categorical ranking by zoning restrictions would complete the picture.

Conclusion

This analysis shows that for large cities with a large number of neighborhoods, Data Science can be used to help narrow down the best initial set of candidate areas that would match up with the type of store-front the business owner wants to create. This helps quantify the areas in a more scientific way, rather than just using "by guess and by golly" techniques of scouting out neighborhoods on foot or on generic recommendations.

These techniques are generally applicable for any large cities that have good statistical breakdowns of real-estate pricing, incomes, and venues. These techniques can be a valuable tool for doing initial targeting of neighborhoods that would be a good fit for a potential new store-front business in such cities. They save time and provide a good starting point for the owner or investor of new store-fronts. These techniques are generalizable for other types of business ventures (venues) in urban areas.