Applied Data Science Capstone

Final Assignment

Lionel K

December 2019

Table of Contents

1.0 Introduction and Business Problem

1.1 Situation

Our client is a prominent "Buyer's Agent" whose business is based on finding and transacting on property sales for the buyer(s), rather than for the vendor (seller). They take requirements and instructions from Buyers, then seek out properties for those Buyers to consider purchasing. Once a property is chosen, our client the Buyer's Agent will also undertake actions to acquire the chosen property for the Buyer. Typical customers of Buyer’s Agents, Buyers are often time poor, migrants located away from the city, or lack confidence in the purchasing process.

1.2 Complication

Most Buyer's Agencies operate in the market in an 'intuitive' way, based on their market knowledge and experience. For example, if a Buyer says they are looking for a country lifestyle with access to schools, the areas recommended to them are often based on that particular Buyer's Agent's own knowledge and experience from previous sales.

Increasing competition and a slowing market has lead our client to look for a way to differentiate from other Buyer’s Agents. Our client wishes to use data, analytics and data science to build an online tool to guide their Buyer's choices, therefore attracting new customers.

2.0 Problem Statement and Use Case

2.1 Problem Statement

How can we use data, analytics and data science to make informed recommendations on locations for Buyer's, based on their individual requirements?

2.2 Use Case

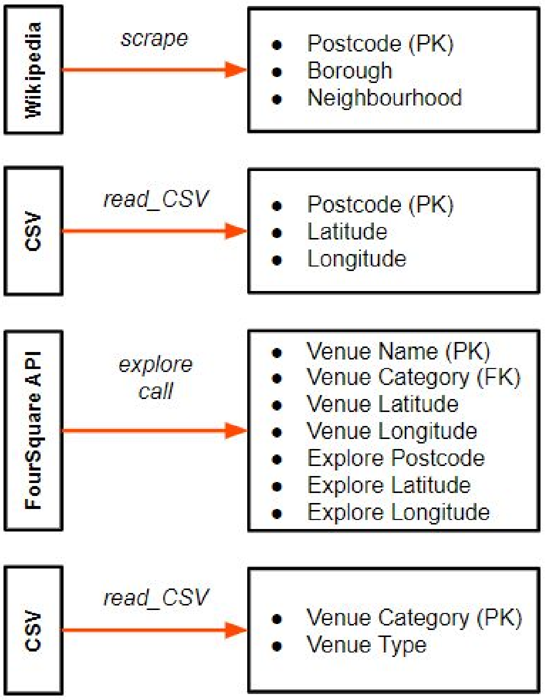
As a pilot project, our client has provided us with a Use Case for a proof of concept. The 'Buyer' in question is a successful middle aged couple, without children. They are relocating to Toronto from Melbourne, Australia and wish to use the move for a change of lifestyle. They have no intention of starting a family, but have grown tired of living in a busy city area. They both work for home / remotely and do not have to travel to an office, but are otherwise active in the community.

The Buyers explicitly provided these instructions to the Buyer's Agent:

"When not working, we spend about 30% of our time dining out, and probably the same again going out to bars and other nightlife, and public amenities like parks and museums. We spend about 20% of our time shopping, and divide the remainder equally between fitness and entertainment. We want a property about 30 mins away from the Toronto CBD by car, but in a location with similar conveniences and facilities that you would find in a downtown area"

3.0 Data

3.1 Data Sources and Definitions



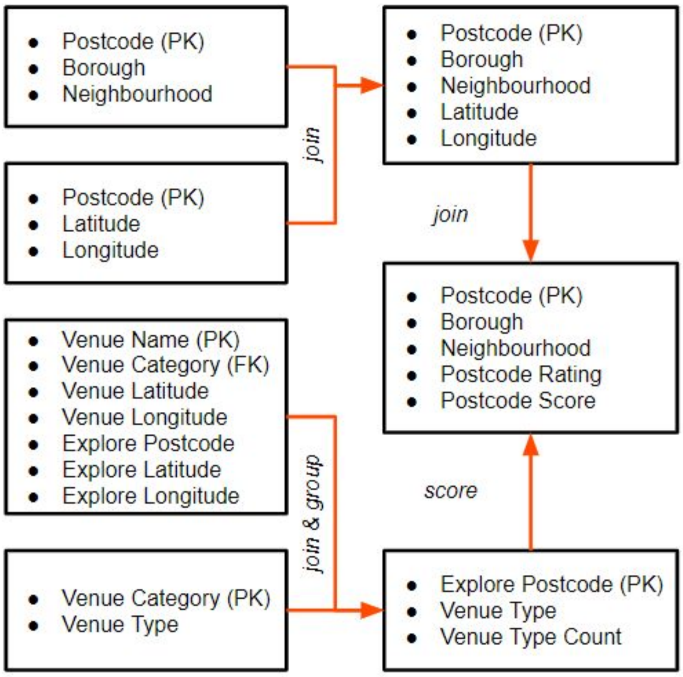
- Postcode: de-duped, used as Primary Key
- Name of Borough in Postcode
- Name(s) of Neighbourhood in Postcode

- Postcode: de-duped, used as Primary Key
- Latitude of Postcode
- Longitude of Postcode

- Name of Venue
- Category of Venue
- Latitude of Venue
- Longitude of Venue
- Postcode used for FourSquare API call
- Latitude of Postcode used for FourSquare API call
- Longitude of Postcode used for FourSquare API call

- Category of Venue
- Type of Venue: Manually coded, higher order categorical variable

3.1 Data Transformations



- Postcode: de-duped, used as Primary Key
- Name of Borough in Postcode
- Name(s) of Neighbourhood in Postcode
- Latitude of Postcode
- Longitude of Postcode

- Postcode: de-duped, used as Primary Key
- Name of Borough in Postcode
- Name(s) of Neighbourhood in Postcode
- H(>75%) / M(>50%) / L, ranked by percentile
- Calculated using Linear model from Buyer info

- Postcode used for FourSquare API call
- Type of Venue: Manually coded categorical
- Count of Venue Type in Postcode

4.0 Methodology

4.1 Objectives

1. Buyer is seeking a location similar to downtown, but about 30 mins away
2. Buyer spends their time on dining, nightlife, public amenities and shopping
3. Client wishes this Use Case to be a POC; approach must be scalable

4.2 Approach

1. We want to use FourSquare data to:
   a. identify the Types of venues in each postcode, as described by Buyer's instructions (Dining, Shopping, Nightlife, ect)
   b. use the count of these types to rate the availability of venues between every postcode in Toronto
   c. identify clusters of postcodes that have similar characteristics

2. Thus, we choose these techniques and visualisations to solve for above:
   a. sums / counts for every Venue and Venue Type, for each postcode
   b. descriptive statistics for each distribution of Venue type per postcode
   c. bar charts and box plots, to visualise the above
   d. derive a Rating variable based on the distribution of Venue Type
   e. derive a Score for each postcode, based on Buyer instructions, in order to rank postcodes

4.3 Data Preparation and Wrangling

- Scrape Postcode data from Wikipedia using Beautiful Soup library
- Import Geospatial data from CSV using .read_csv method in Pandas
- Wrangling for missing wikipedia data, dedupe Postcode, group Neighbourhoods
- Create For loop to iterate over Postcode to extract 100 obs per postcode
- Extract FourSquare data using Explore Call data via API
- Transform FourSquare Explore Call data, keeping required features
- Create list of unique Venue Categories found in Toronto
- Manually code Venue Types per Buyer instructions
- Import Venue Types data using .read_csv method, join to Foursquare Data

4.4 Analysis & Modelling

- Count of Venue Types per Postcode
    - produce ranked ordered bar chart visualisation

- Calculate descriptive statistics on Count of Venue Type
    - produce Boxplot by Venue Type

- Assign Rating variable using =
    - rank of Count of Venue Type in percentile by Venue Type

- Rating =
    - if CountVenueType > 75%tile = HIGH
    - elif CountVenueType > 75%tile = MED
    - else LOW)

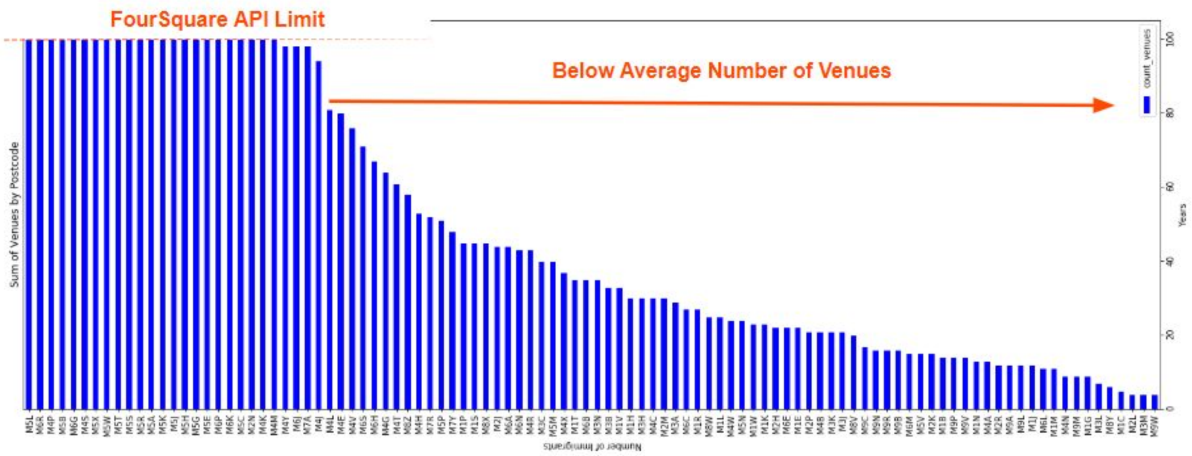
- Calculate Score using linear model as an ‘Expert Model’
    - Equation derived as a SLR based on Buyer’s instructions
    - Score =

*0.3*Dining + 0.3*(Nightlife+Amenities) + 0.2*Shopping + 0.2*(Fitness+Entertainment)*

- Assign Postcode Cluster using k-means clustering to identify similar postcodes

5.0 Results

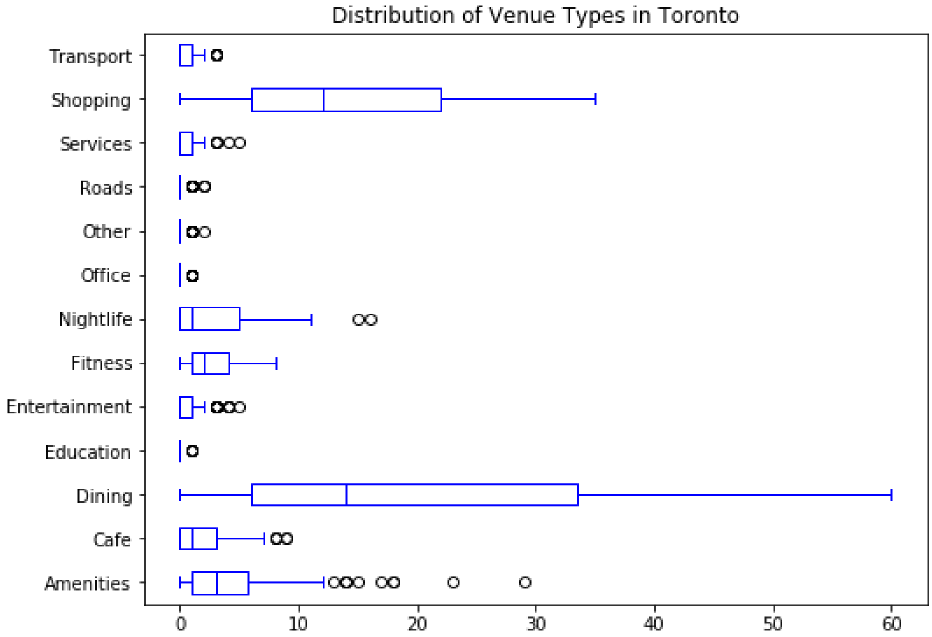
5.1 Number of Venues per Postcode

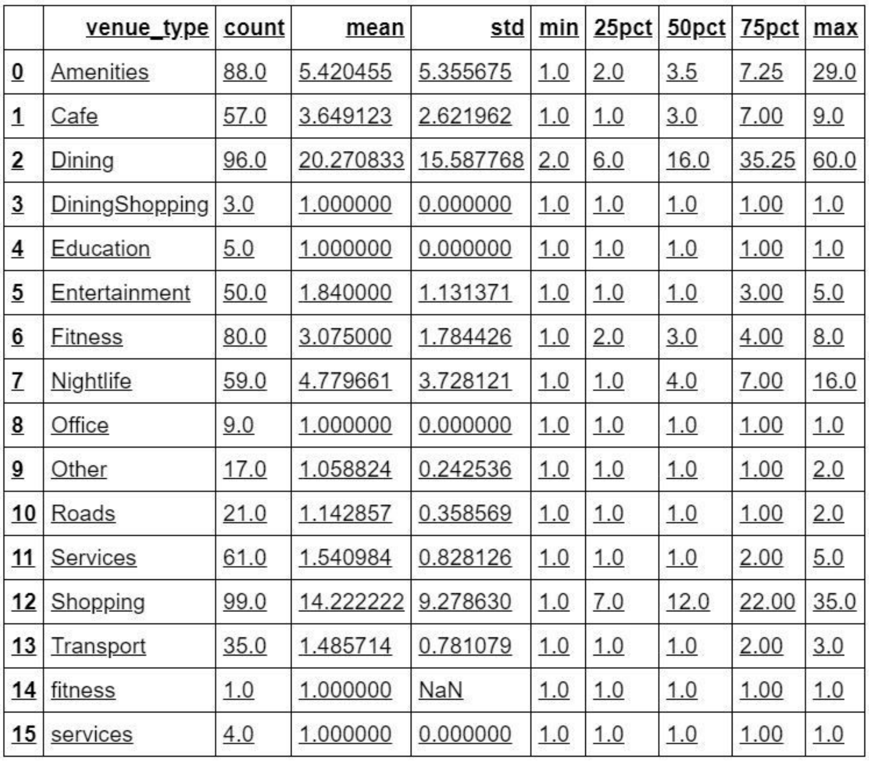


Observations

- Distribution of Count of Venues across Toronto is log-natural
  - ie less postcodes have many venues, most postcode have less than average venues
  - Import Geospatial data from CSV using .read_csv method in Pandas
- Buyers are looking for similar to downtown, postcodes with < ave no. of venues will not do
- However, most postcode with > ave no. of venues are in downtown! Keeping looking…

5.2 Distribution of Venue Types

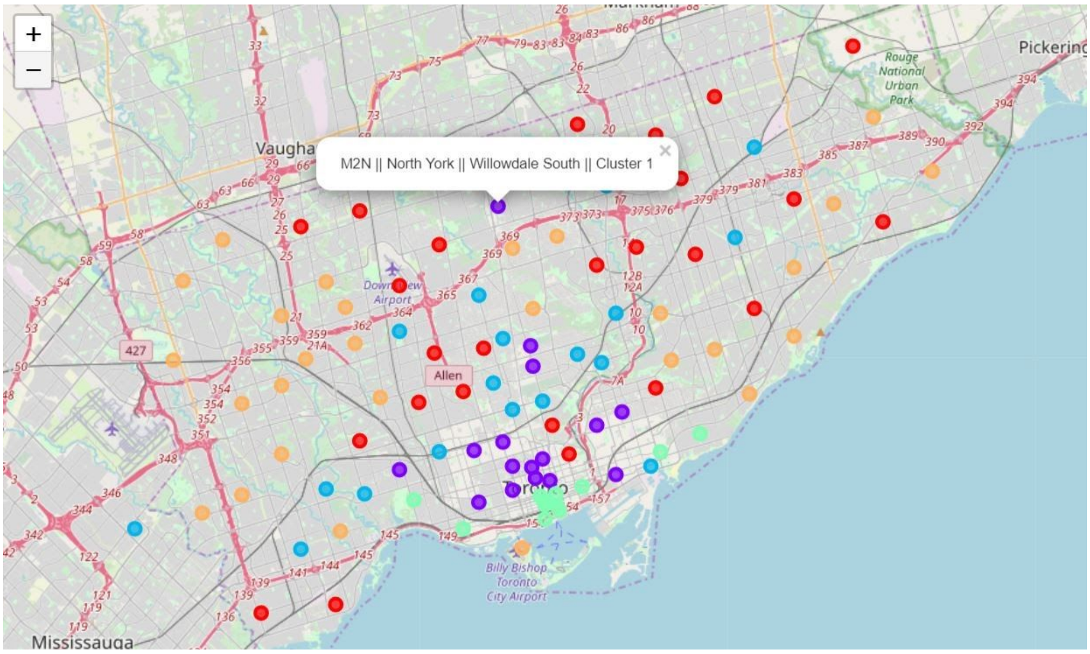




	venue_type	count	mean	std	min	25pct	50pct	75pct	max
0	Amenities	88.0	5.420455	5.355675	1.0	2.0	3.5	7.25	29.0
1	Cafe	57.0	3.649123	2.621962	1.0	1.0	3.0	7.00	9.0
2	Dining	96.0	20.270833	15.587768	2.0	6.0	16.0	35.25	60.0
3	DiningShopping	3.0	1.000000	0.000000	1.0	1.0	1.0	1.00	1.0
4	Education	5.0	1.000000	0.000000	1.0	1.0	1.0	1.00	1.0
5	Entertainment	50.0	1.840000	1.131371	1.0	1.0	1.0	3.00	5.0
6	Fitness	80.0	3.075000	1.784426	1.0	2.0	3.0	4.00	8.0
7	Nightlife	59.0	4.779661	3.728121	1.0	1.0	4.0	7.00	16.0
8	Office	9.0	1.000000	0.000000	1.0	1.0	1.0	1.00	1.0
9	Other	17.0	1.058824	0.242536	1.0	1.0	1.0	1.00	2.0
10	Roads	21.0	1.142857	0.358569	1.0	1.0	1.0	1.00	2.0
11	Services	61.0	1.540984	0.828126	1.0	1.0	1.0	2.00	5.0
12	Shopping	99.0	14.222222	9.278630	1.0	7.0	12.0	22.00	35.0
13	Transport	35.0	1.485714	0.781079	1.0	1.0	1.0	2.00	3.0
14	fitness	1.0	1.000000	NaN	1.0	1.0	1.0	1.00	1.0
15	services	4.0	1.000000	0.000000	1.0	1.0	1.0	1.00	1.0

- Dining venues dominate Toronto, which our Buyers will enjoy
- Shopping venues are the 2nd most common in Toronto
- Our Buyers may be disappointed in Toronto’s nightlife and amenities

5.3 K-means Clustering


M2N || North York || Willowdale South || Cluster 1
Vaugha
Pickering
Rouge National Urban Park
Downsview Airport
Allen
Toronto
Billy Bishop Toronto City Airport
Mississauga
427

- Purple Cluster postcodes are downtown locations, except one!
- Could M2N Willowdale South, 30 mins north of Toronto's CBD be the location our Buyers are looking for?

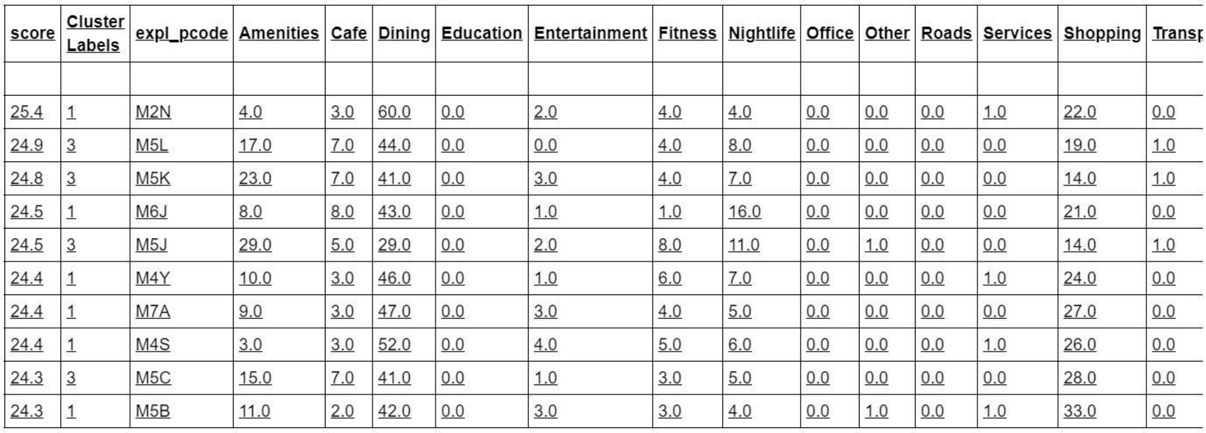
5.4 Scoring

FourSquare no longer allows through API calls, the retrieval of:

- non-Self User data, or
- checkinCount

As such we lack sufficient volume data to statistically calculated any regression model ect. Instead we will approximate a score as an Expert Model, using the following instructions provided by the Buyer.

Equation is described in Section 4.4. The result of scoring is shown below:



score	Cluster Labels	expl_pcode	Amenities	Cafe	Dining	Education	Entertainment	Fitness	Nightlife	Office	Other	Roads	Services	Shopping	Transp
25.4	1	M2N	4.0	3.0	60.0	0.0	2.0	4.0	4.0	0.0	0.0	0.0	1.0	22.0	0.0
24.9	3	M5L	17.0	7.0	44.0	0.0	0.0	4.0	8.0	0.0	0.0	0.0	0.0	19.0	1.0
24.8	3	M5K	23.0	7.0	41.0	0.0	3.0	4.0	7.0	0.0	0.0	0.0	0.0	14.0	1.0
24.5	1	M6J	8.0	8.0	43.0	0.0	1.0	1.0	16.0	0.0	0.0	0.0	0.0	21.0	0.0
24.5	3	M5J	29.0	5.0	29.0	0.0	2.0	8.0	11.0	0.0	1.0	0.0	0.0	14.0	1.0
24.4	1	M4Y	10.0	3.0	46.0	0.0	1.0	6.0	7.0	0.0	0.0	0.0	1.0	24.0	0.0
24.4	1	M7A	9.0	3.0	47.0	0.0	3.0	4.0	5.0	0.0	0.0	0.0	0.0	27.0	0.0
24.4	1	M4S	3.0	3.0	52.0	0.0	4.0	5.0	6.0	0.0	0.0	0.0	1.0	26.0	0.0
24.3	3	M5C	15.0	7.0	41.0	0.0	1.0	3.0	5.0	0.0	0.0	0.0	0.0	28.0	0.0
24.3	1	M5B	11.0	2.0	42.0	0.0	3.0	3.0	4.0	0.0	1.0	0.0	1.0	33.0	0.0

We see that M2N Willowdale South is the most highly scored for this Buyer.

6.0 Discussion

- We used Toronto Geospatial and FourSquare location data to identified that most postcode had too few venues to meet our Buyer’s desire for a downtown-like lifestyle
- However, most postcodes with above average number of venues were located downtown
- We used descriptive statistics and boxplots to uncover that there is abundant Dining and Shopping venues in Toronto, but Nightlife venues may be lacking
- We then built a Rating for each venue type, for each postcode
- K-means clustering helped us to discover that M2N Willowdale South has similar characteristics to downtown postcodes
- We derived a linear model based on the Buyer’s instructions, and scored all postcodes. The results show that M2N Willowdale South is the most highly scored for this Buyer’s desires

7.0 Conclusion

- The Buyer in this Proof of Concept should look for property in ***M2N Willowdale South***
- We have proven to the Client that we can use data science to differentiate their business
- We can scale this Proof of Concept for the Client using a simple Buyer survey