

Applied Data Science Capstone

Part 1 and 2

Coursera

Lionel K

December 2019

Introduction and Business Problem

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, the Buyers, are often time poor, migrants located away from the city, or lack confidence in the purchasing process.

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.

Problem Statement and Use Case

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?

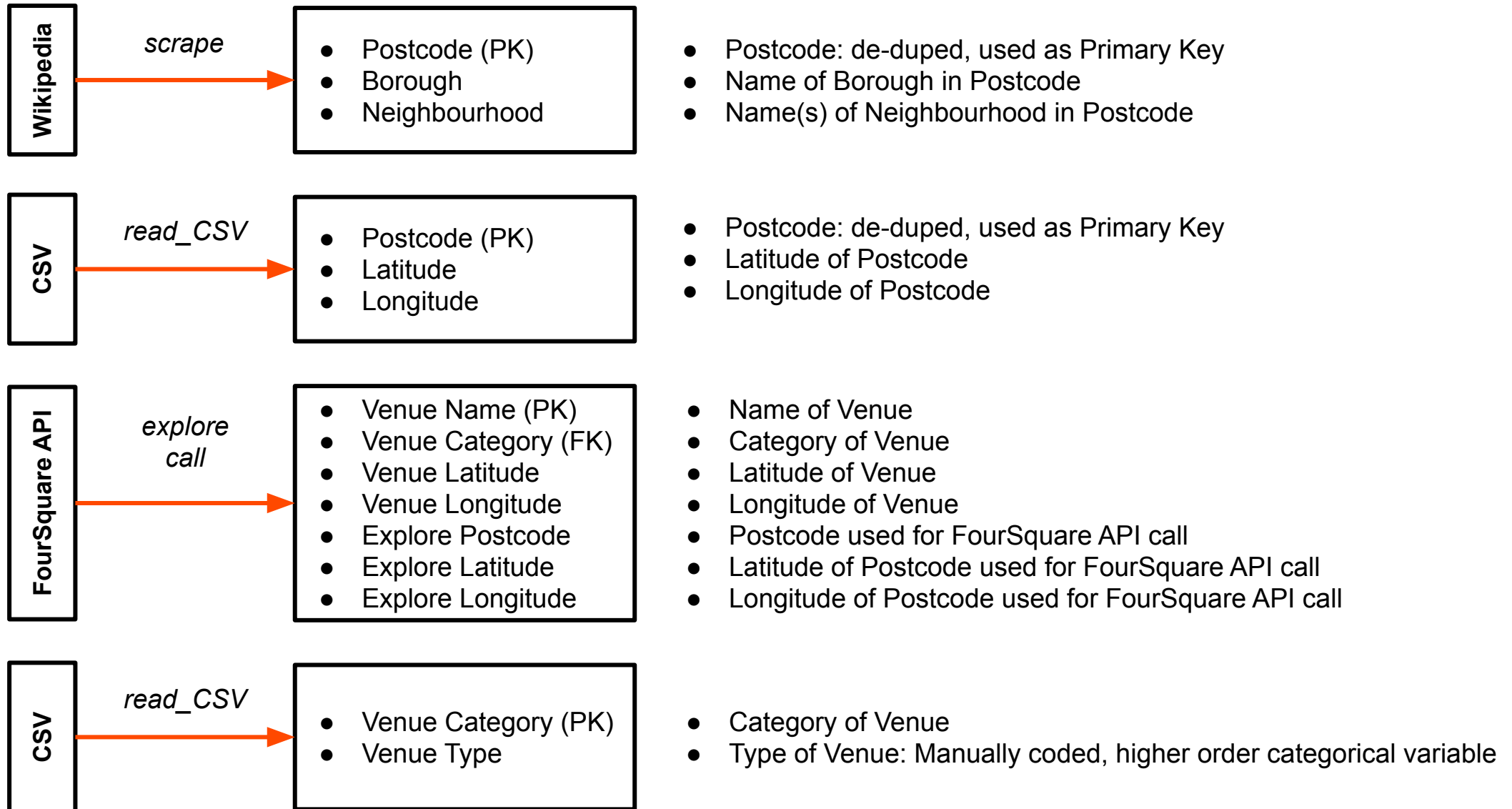
Use Case:

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.

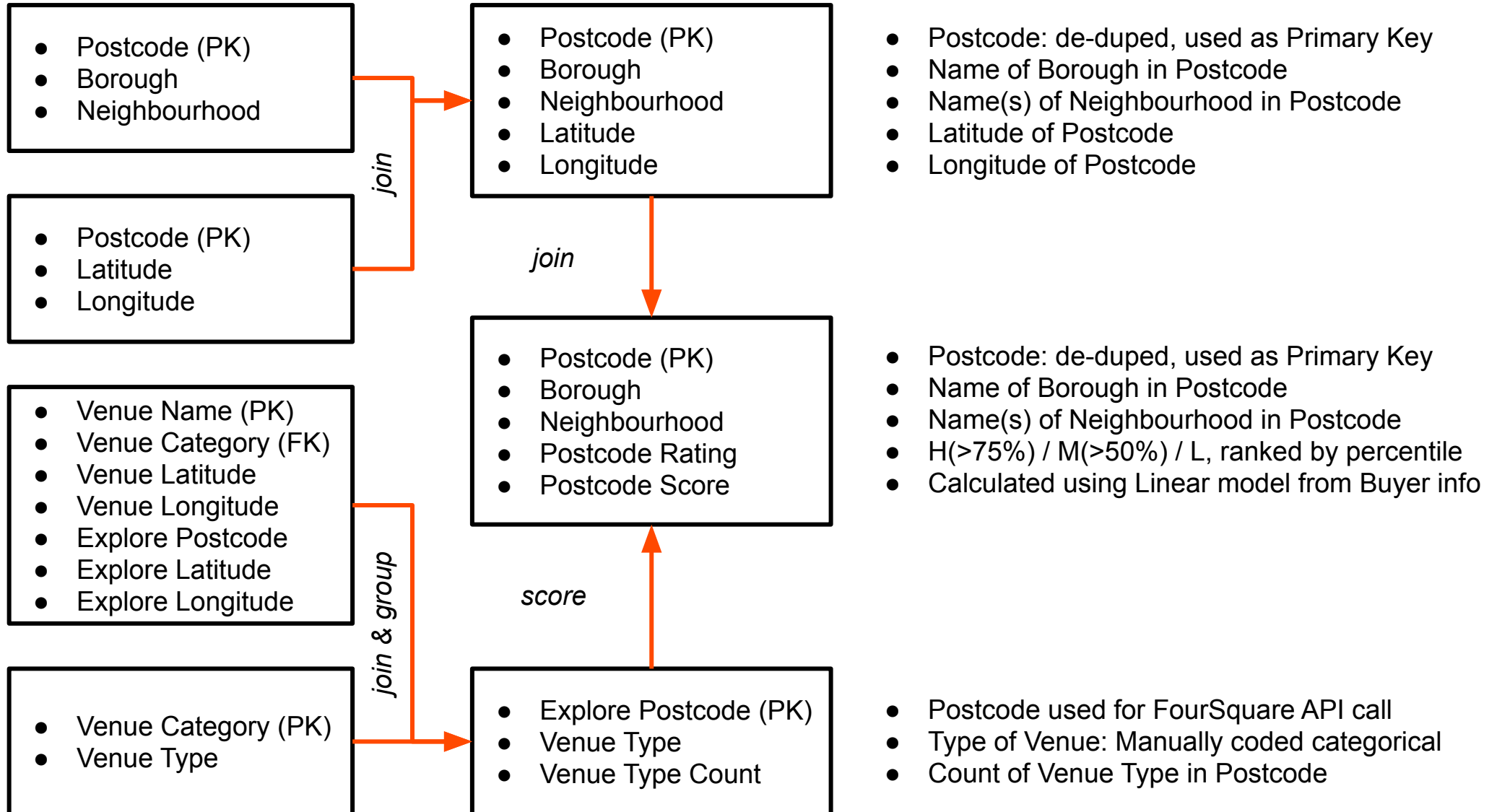
The Buyers explicitly provided these instructions to the Buyer's Agent (our client):

"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"

Data Sources and Definitions



Data Transformations



Methodology, Rationale

Upon reviewing the Client and Buyer's instructions and scope we can determine the following...

Objectives

1. Buyer desires to purchase in a location similar to downtown, but about 30 mins away
2. Buyer spends most of their time on dining, nightlife, public amenities and shopping
3. Client wishes to use this Use Case as a POC, so methodology must be scalable

Approach

1. We want to use FourSquare data to:
 - a. identify the Types of venues in each postcode, as described by Buyer's instructions
 - b. use the count of these types to rate the availability of venues between postcodes
 - c. identify clusters of postcodes that have similar characteristics to each other
2. Thus, we will use the following 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, for each postcode
 - c. bar charts and box plots, to visualise the above
 - d. derive a Rating variable based on the distribution of Venue Type, for each postcode
 - e. derive a Score for each postcode, based on Buyer instructions, in order to rank postcodes

Methodology, Outline

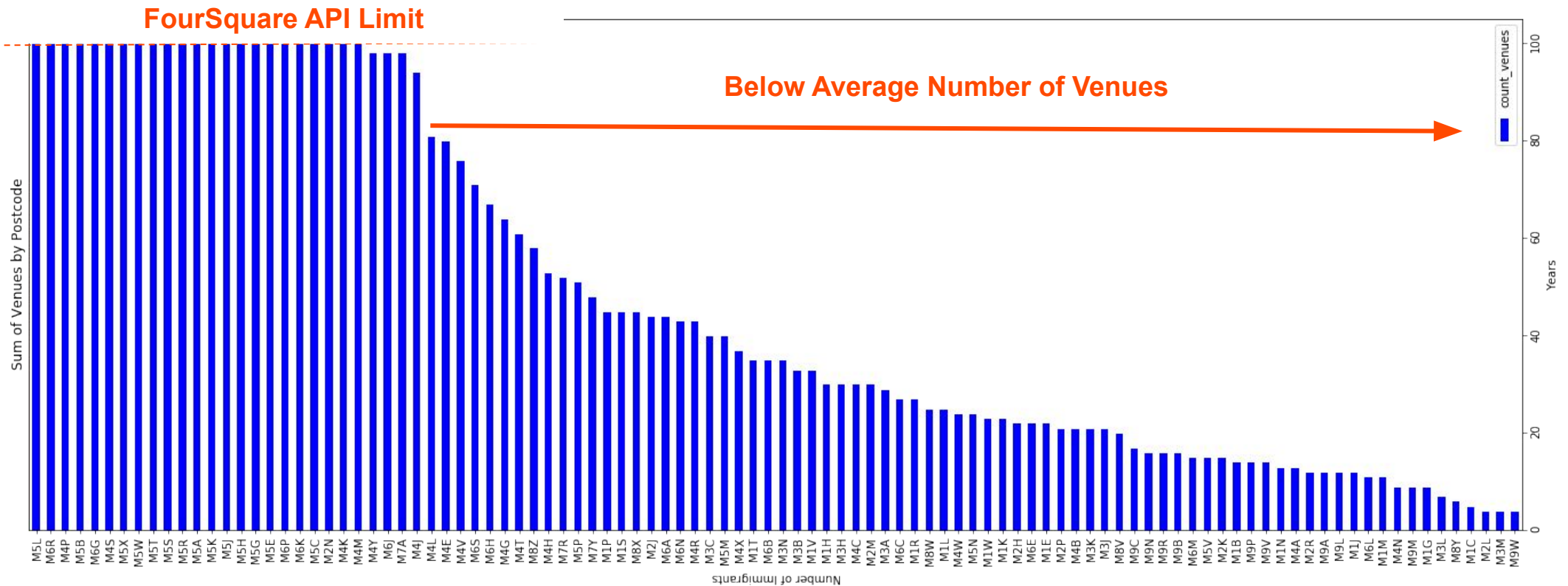
1. Data Preparation and Wrangling

- Scrape Postcode data from Wikipedia using Beautiful Soup library; create Dataframes
- Import Geospatial data from CSV using `.read_csv` method in Pandas
- Wrangling for missing wikipedia data, dedupe by Postcode and group Neighbourhoods
- Create For loop to iterate over Postcode, and extract FourSquare Explore Call data via API
- Transform FourSquare Explore Call data, keep required features and create Dataframe
- Create list of unique Venue Categories, manually code Venue Types per Buyer requirement
- Import Venue Types data using `.read_csv` method, join to Foursquare Data

2. 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 $\text{Count}_{\text{VenueType}} > 75\% \text{tile} = \text{HIGH}$, elif $\text{Count}_{\text{VenueType}} > 75\% \text{tile} = \text{MED}$, else LOW)
- Calculate Score using linear model as an 'Expert Model' from Buyer's instructions
 - Score = $0.3 * \text{Dining} + 0.3 * (\text{Nightlife} + \text{Amenities}) + 0.2 * \text{Shopping} + 0.2 * (\text{Fitness} + \text{Entertain})$
- Assign Postcode Cluster using k-means clustering to identify similar postcodes

Results 1, Number of Venues per Postcode

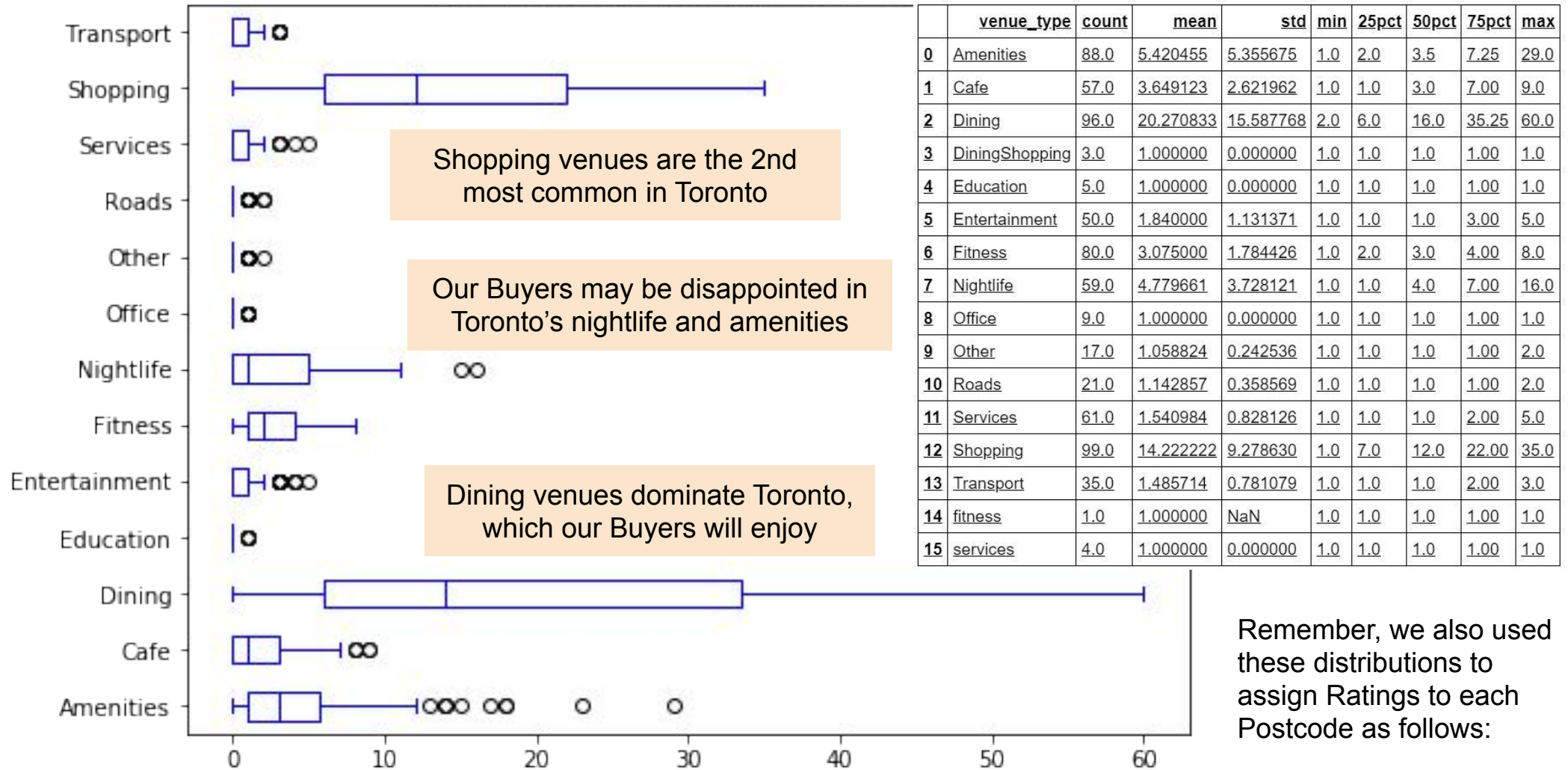


Observations

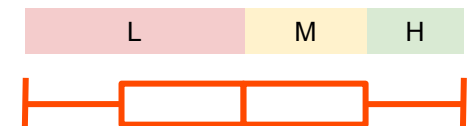
- Distribution of Count of Venues across Toronto is log-natural
 - ie less postcodes have many venues, most postcode have less than average venues
 - Import Geospatial data from CSV using `.read_csv` method in Pandas
- Buyers are looking for similar to downtown, postcodes with < ave no. of venues will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

Results 2, Distribution of Venue Types

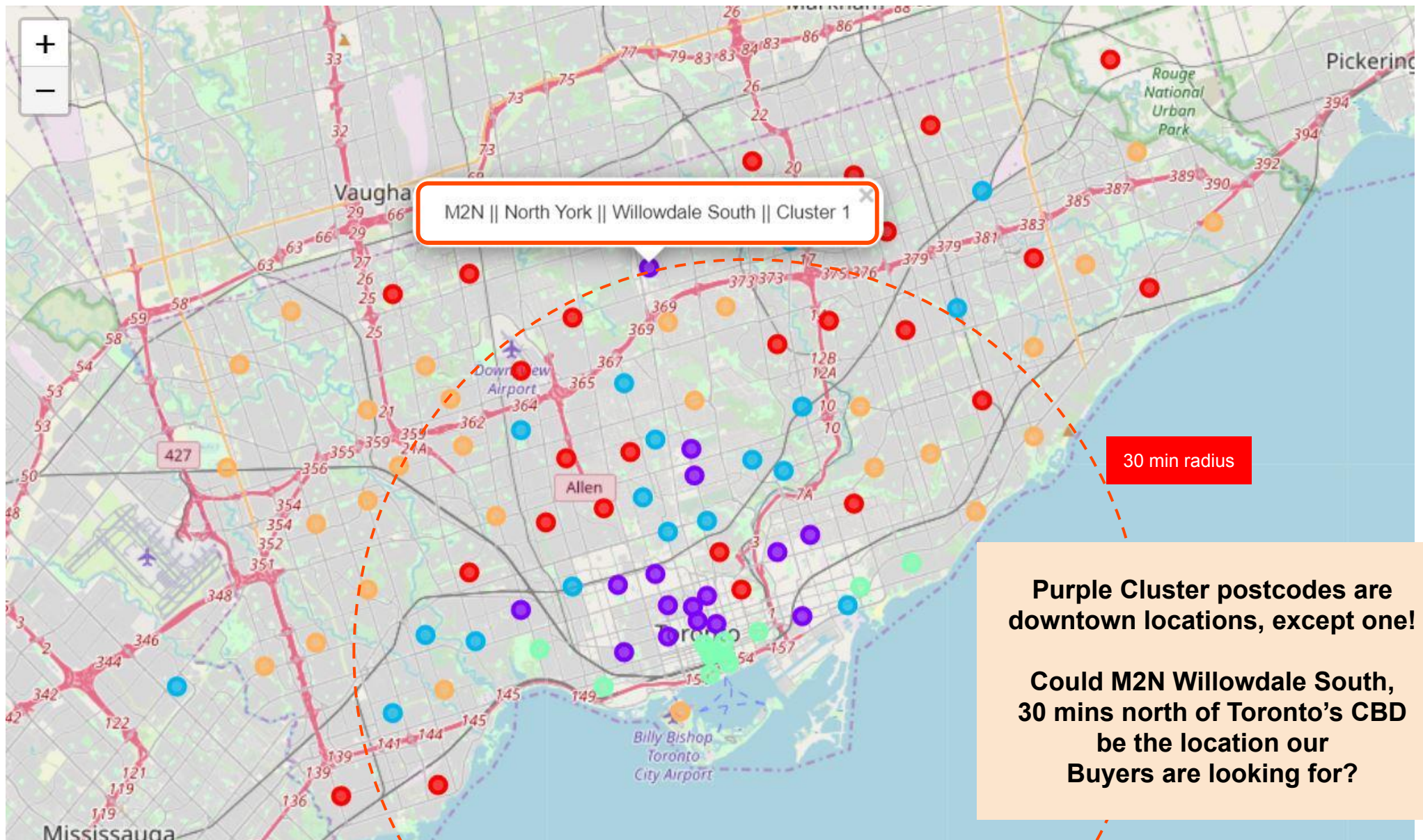
Distribution of Venue Types in Toronto



Remember, we also used these distributions to assign Ratings to each Postcode as follows:



Results 2, K-means Clustering



Results 2, Postcode Score

"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"

- Score = $0.3 \times \text{Dining} + 0.3 \times (\text{Nightlife} + \text{Amenities}) + 0.2 \times \text{Shopping} + 0.2 \times (\text{Fitness} + \text{Entertain})$
- M2N Willowdale South is the most highly scored postcode for our Buyer's preferences!

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

Discussion and Conclusion

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

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

thank you!