The Battle of the Neighborhoods

Presents

The Battle of Alberta – The Neighborhood Edition

Calgary versus Edmonton

By Pam Pritchett

*This report is a Battle of the Neighborhoods capstone requirement for the IBM Data Science Professional Certificate offered by Coursera.

Introduction

The Battle of Alberta is a term well known to the residents of Alberta that refers to the playful and sometimes intense rivalry between the cities of Calgary and Edmonton. Calgary is Alberta's largest city with the second-highest corporate head offices in Canada. In 1988, it was also the first Canadian city to host the Winter Olympics Games. Edmonton is Alberta's capital city, the northern most major city in North America known as the Gateway to the North, and is also home to North America's largest mall.

This rivalry has its roots dating back to the early 1900s before Alberta was a province, when the cities competed to become its capital city. The rivalry later manifested in sports, most predominately the historic rivalry between their respective NHL sports teams, the Edmonton Oilers and the Calgary Flames. It fuels the passion of their fans, each city boosting that their city is better than the other. But which city is correct? Do the neighborhoods of each city offer the same access to amenities, or do the neighborhoods of one city outshine the other?

This project endeavors to answer the question: Calgary or Edmonton – Which city is better to live in?

This analysis compares the two cities by exploring venues and amenities based on FourSquare data. This is a simplistic approach that does not consider other complex socio-economic factors that impact a neighborhood's livability. It is intended instead, to be a lighthearted approach using location analysis to compare the merits of the neighborhoods of each city.

Target Audience

The target audience is the people of Alberta who have ever questioned the basis of the Calgary / Edmonton rivalry and wondered are the neighborhoods of each city really that different from each other and is so, then how?

Dataset

Data Sources

The data acquired for this analysis are derived from the following sources:

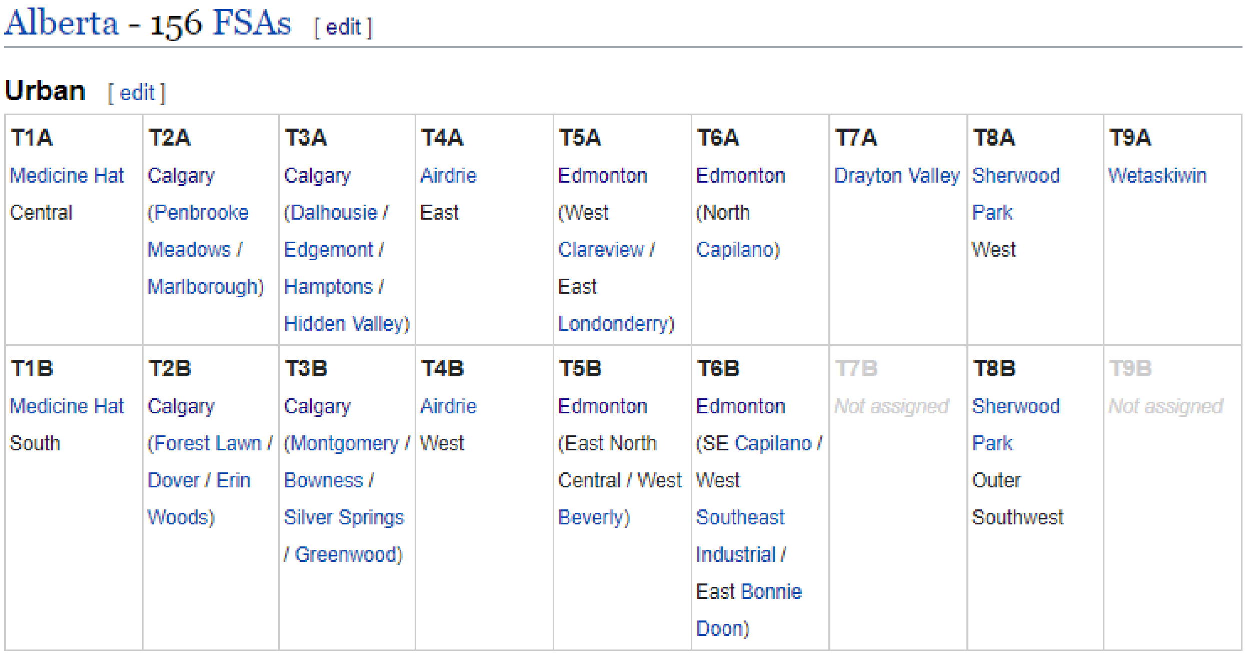
- Postal Codes are scraped from Wikipedia for each city to enable neighborhood groupings will be extracted from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T
- ArcGis geocoding web services to retrieve geological coordinates to create a new dataframe to be passed to FourSquare
- FourSquare API used to explore and retrieve neighborhood venues

Data Challenges

The Wikipedia data structure for the postal code and neighborhood data must be transposed for analysis.

- The required data is combined into a single cell. Analysis requires the data to be scraped and extracted into a dataframes containing ‘Neighborhood’ and ‘Postal Code’ columns for each city.
- The Wikipedia data structure has 9 columns: columns 2 and 3 contain Calgary data and columns 5 and 6 contain Edmonton data. For analysis, these columns need to be stacked into a single column.

A sample of the data format from Wikipedia is displayed below:



Alberta - 156 FSAs [edit]

Urban [edit]

T1A Medicine Hat Central	T2A Calgary (Penbrooke Meadows / Marlborough)	T3A Calgary (Dalhousie / Edgemont / Hamptons / Hidden Valley)	T4A Airdrie East	T5A Edmonton (West Clareview / East Londonderry)	T6A Edmonton (North Capilano)	T7A Drayton Valley	T8A Sherwood Park West	T9A Wetaskiwin
T1B Medicine Hat South	T2B Calgary (Forest Lawn / Dover / Erin Woods)	T3B Calgary (Montgomery / Bowness / Silver Springs / Greenwood)	T4B Airdrie West	T5B Edmonton (East North Central / West Beverly)	T6B Edmonton (SE Capilano / West Southeast Industrial / East Bonnie Doon)	T7B *Not assigned*	T8B Sherwood Park Outer Southwest	T9B *Not assigned*

Table 1 - Wikipedia Table of Alberta Postal Codes

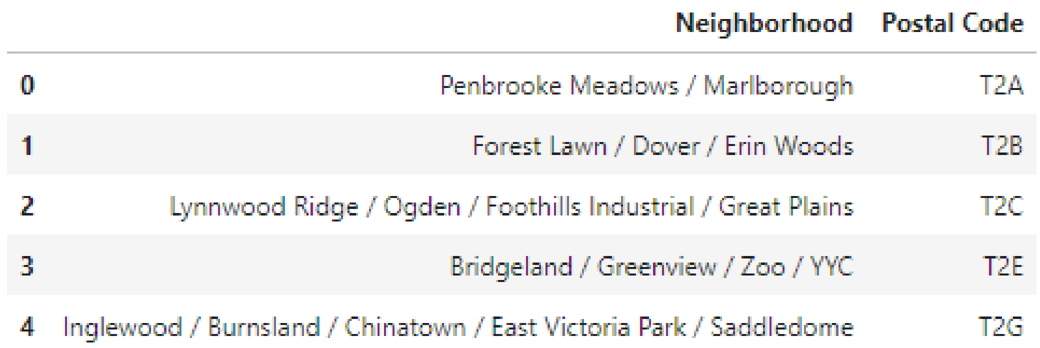
Data Preparation Methods

1. Scrape postal codes and associated neighborhoods from Wikipedia
2. Use Beautiful Soup library to parse the HTML table data
3. Extract parsed data into separate Calgary and Edmonton pandas dataframes
4. Stack the ‘Neighborhood’ columns into a single column
5. Create a ‘Postal Code’ column in both dataframes and extract from the existing ‘Neighborhood’ data and perform a cleanup on the ‘Neighborhood’ names.

Explore Data

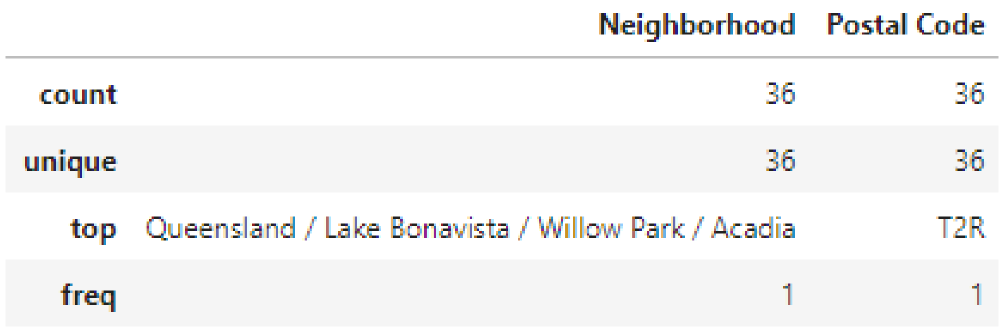
Calgary

Calgary dataframe containing scraped ‘Neighborhood’ and ‘Postal Code’ data



	Neighborhood	Postal Code
0	Penbrooke Meadows / Marlborough	T2A
1	Forest Lawn / Dover / Erin Woods	T2B
2	Lynnwood Ridge / Ogden / Foothills Industrial / Great Plains	T2C
3	Bridgeland / Greenview / Zoo / YYC	T2E
4	Inglewood / Burnsland / Chinatown / East Victoria Park / Saddledome	T2G

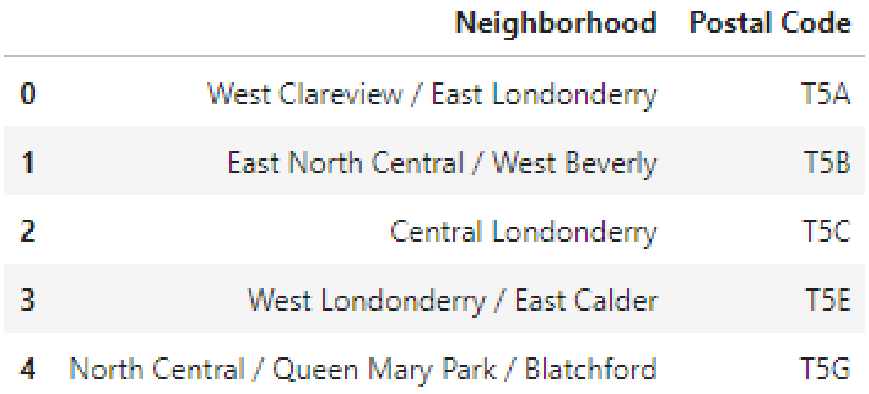
Calgary summary statistics identifies 36 distinct neighborhoods.



	Neighborhood	Postal Code
count	36	36
unique	36	36
top	Queensland / Lake Bonavista / Willow Park / Acadia	T2R
freq	1	1

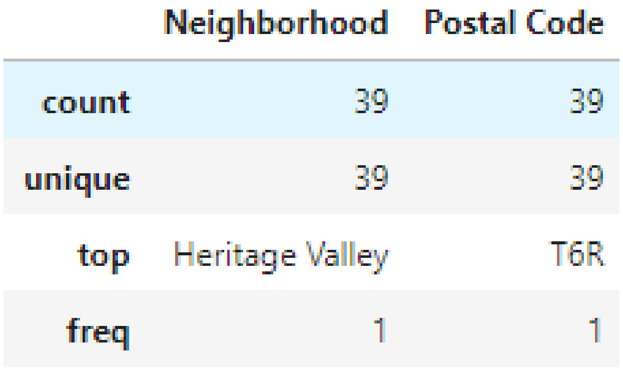
Edmonton

Edmonton dataframe containing scraped 'Neighborhood' and 'Postal Code' data



	Neighborhood	Postal Code
0	West Clareview / East Londonderry	T5A
1	East North Central / West Beverly	T5B
2	Central Londonderry	T5C
3	West Londonderry / East Calder	T5E
4	North Central / Queen Mary Park / Blatchford	T5G

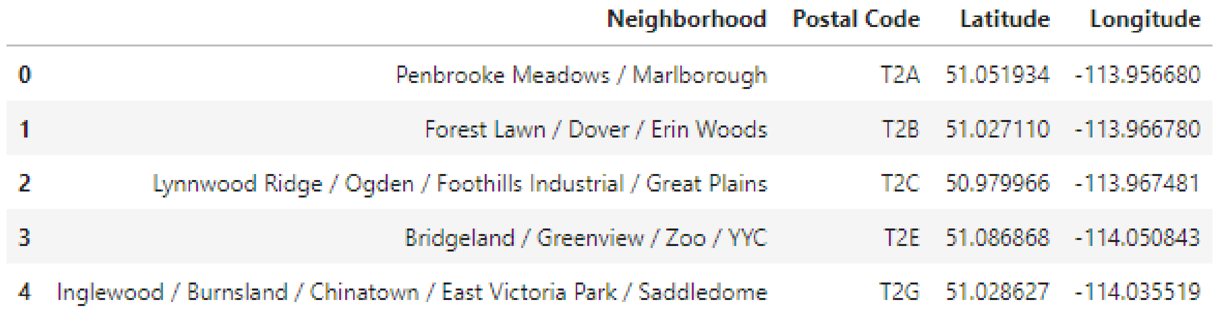
Edmonton summary statistics identifies 39 distinct neighborhoods.



	Neighborhood	Postal Code
count	39	39
unique	39	39
top	Heritage Valley	T6R
freq	1	1

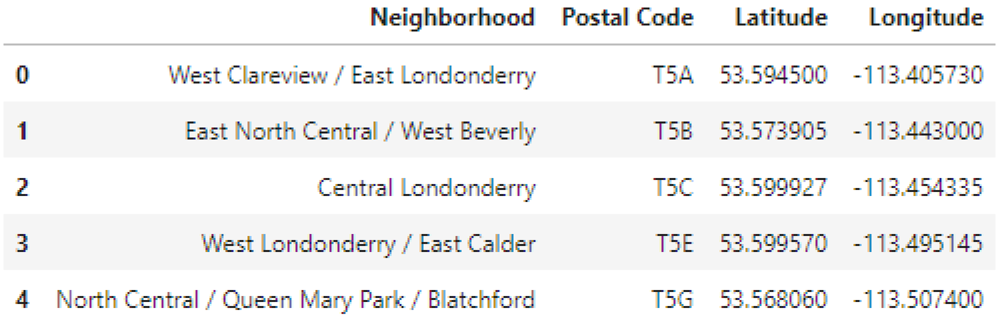
ArcGIS geocoder services and Nominatim APi use to locate latitude and longitude geographic coordinates. Dataframes merged with the coordinates for analysis using the FourSquare API

Calgary



	Neighborhood	Postal Code	Latitude	Longitude
0	Penbrooke Meadows / Marlborough	T2A	51.051934	-113.956680
1	Forest Lawn / Dover / Erin Woods	T2B	51.027110	-113.966780
2	Lynnwood Ridge / Ogden / Foothills Industrial / Great Plains	T2C	50.979966	-113.967481
3	Bridgeland / Greenview / Zoo / YYC	T2E	51.086868	-114.050843
4	Inglewood / Burnsland / Chinatown / East Victoria Park / Saddledome	T2G	51.028627	-114.035519

Edmonton

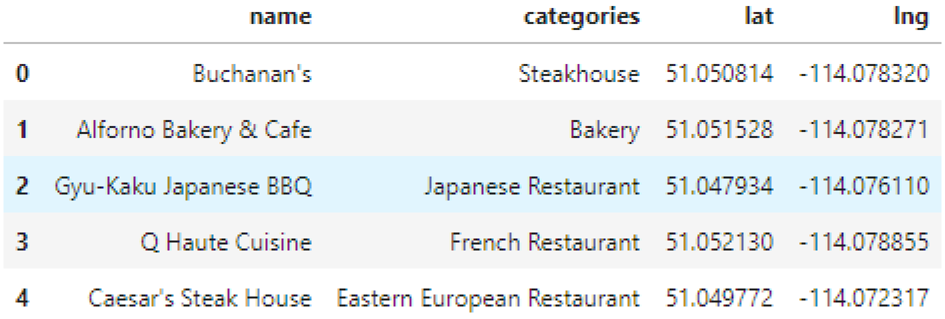


	Neighborhood	Postal Code	Latitude	Longitude
0	West Clareview / East Londonderry	T5A	53.594500	-113.405730
1	East North Central / West Beverly	T5B	53.573905	-113.443000
2	Central Londonderry	T5C	53.599927	-113.454335
3	West Londonderry / East Calder	T5E	53.599570	-113.495145
4	North Central / Queen Mary Park / Blatchford	T5G	53.568060	-113.507400

FourSquare API used to retrieve and explore venues in downtown neighborhoods of the rival cities

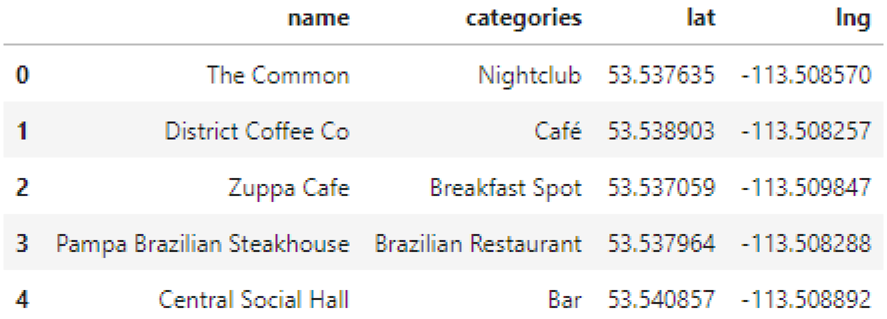
Top 5 venues in downtown areas of each city

Calgary - City Centre / Calgary Tower



	name	categories	lat	lng
0	Buchanan's	Steakhouse	51.050814	-114.078320
1	Alforno Bakery & Cafe	Bakery	51.051528	-114.078271
2	Gyu-Kaku Japanese BBQ	Japanese Restaurant	51.047934	-114.076110
3	Q Haute Cuisine	French Restaurant	51.052130	-114.078855
4	Caesar's Steak House	Eastern European Restaurant	51.049772	-114.072317

Edmonton - South Downtown / South Downtown Fringe / AB Government



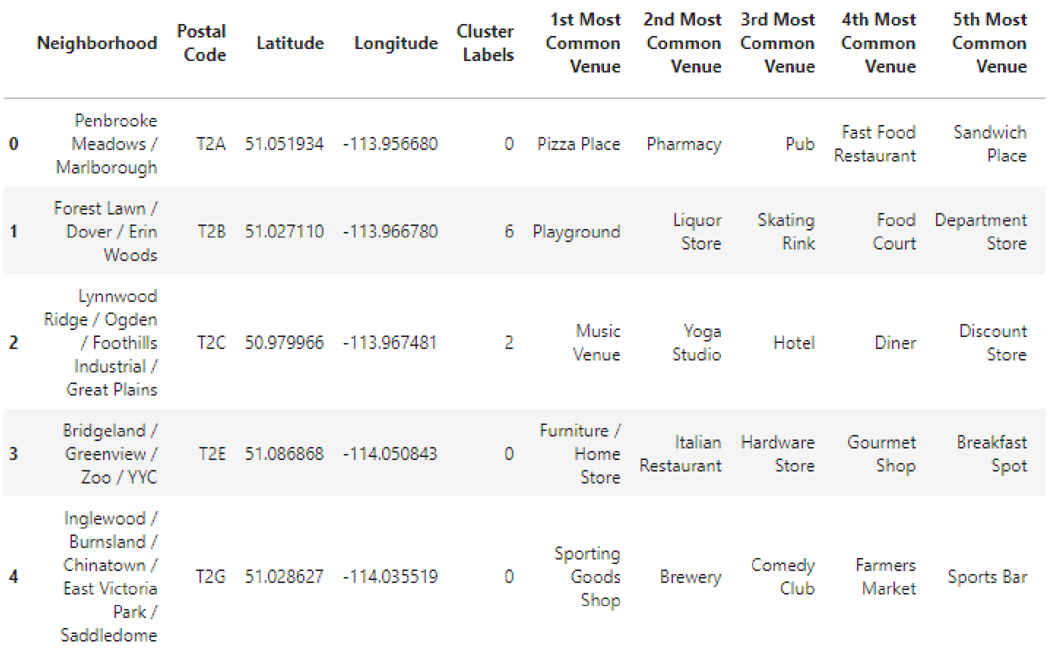
	name	categories	lat	lng
0	The Common	Nightclub	53.537635	-113.508570
1	District Coffee Co	Café	53.538903	-113.508257
2	Zuppa Cafe	Breakfast Spot	53.537059	-113.509847
3	Pampa Brazilian Steakhouse	Brazilian Restaurant	53.537964	-113.508288
4	Central Social Hall	Bar	53.540857	-113.508892

Methodology

Use machine learning model to cluster and compare neighborhood venues

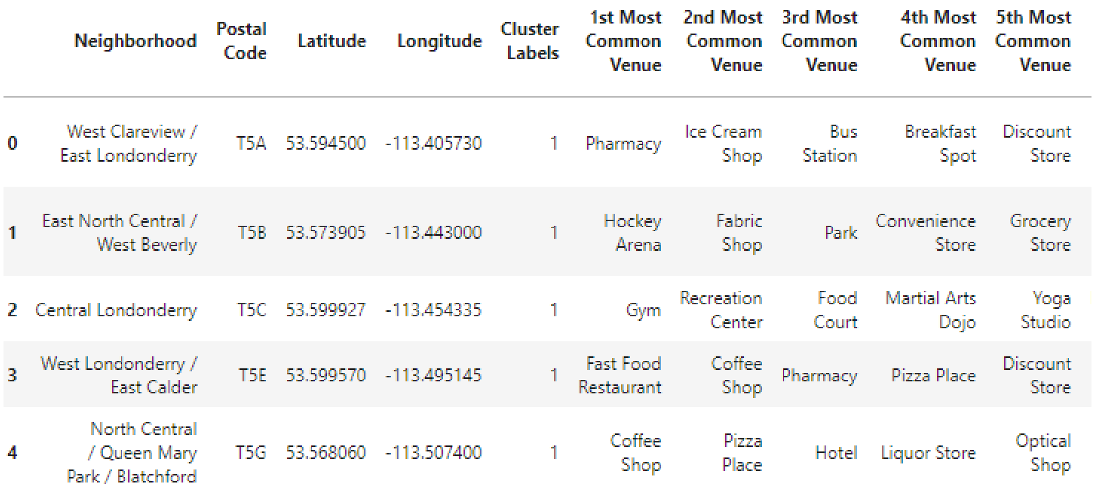
- Use OneHot Encoding to encode venue categories into an array to train data and group by neighborhood
- Identify top 5 most common venues of each neighborhood and create dataframes
- Cluster similar neighborhoods using K-means clustering, an unsupervised machine learning algorithm to identify underlying patterns of the clusters

Top 5 venues by neighborhood in Calgary with cluster labels



	Neighborhood	Postal Code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Penbrooke Meadows / Marlborough	T2A	51.051934	-113.956680	0	Pizza Place	Pharmacy	Pub	Fast Food Restaurant	Sandwich Place
1	Forest Lawn / Dover / Erin Woods	T2B	51.027110	-113.966780	6	Playground	Liquor Store	Skating Rink	Food Court	Department Store
2	Lynnwood Ridge / Ogden / Foothills Industrial / Great Plains	T2C	50.979966	-113.967481	2	Music Venue	Yoga Studio	Hotel	Diner	Discount Store
3	Bridgeland / Greenview / Zoo / YYC	T2E	51.086868	-114.050843	0	Furniture / Home Store	Italian Restaurant	Hardware Store	Gourmet Shop	Breakfast Spot
4	Inglewood / Burnsland / Chinatown / East Victoria Park / Saddledome	T2G	51.028627	-114.035519	0	Sporting Goods Shop	Brewery	Comedy Club	Farmers Market	Sports Bar

Top 5 venues by neighborhood in Edmonton with cluster labels

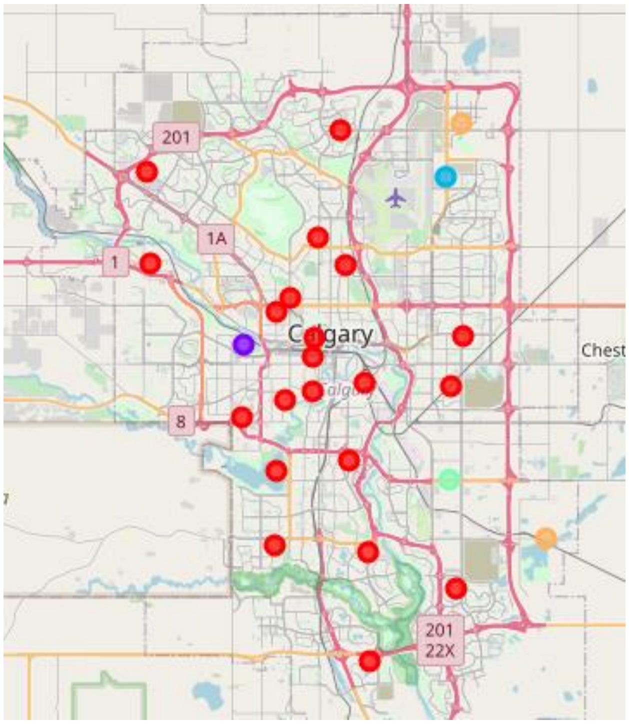


	Neighborhood	Postal Code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	West Clareview / East Londonderry	T5A	53.594500	-113.405730	1	Pharmacy	Ice Cream Shop	Bus Station	Breakfast Spot	Discount Store
1	East North Central / West Beverly	T5B	53.573905	-113.443000	1	Hockey Arena	Fabric Shop	Park	Convenience Store	Grocery Store
2	Central Londonderry	T5C	53.599927	-113.454335	1	Gym	Recreation Center	Food Court	Martial Arts Dojo	Yoga Studio
3	West Londonderry / East Calder	T5E	53.599570	-113.495145	1	Fast Food Restaurant	Coffee Shop	Pharmacy	Pizza Place	Discount Store
4	North Central / Queen Mary Park / Blatchford	T5G	53.568060	-113.507400	1	Coffee Shop	Pizza Place	Hotel	Liquor Store	Optical Shop

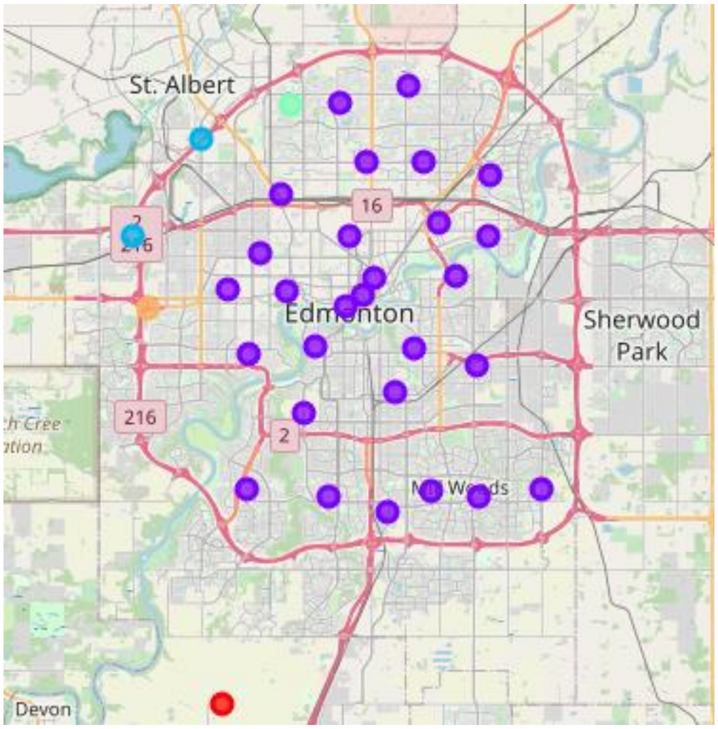
Results

Mapping Results

Cluster Similar Calgary Neighborhoods



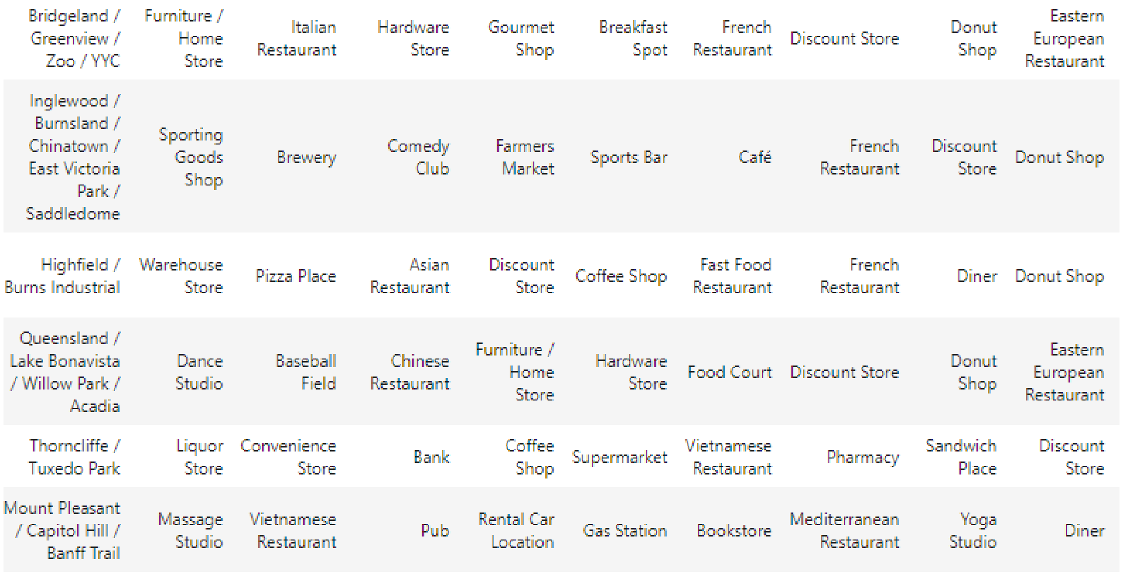
Cluster Similar Edmonton Neighborhoods



Cluster Results

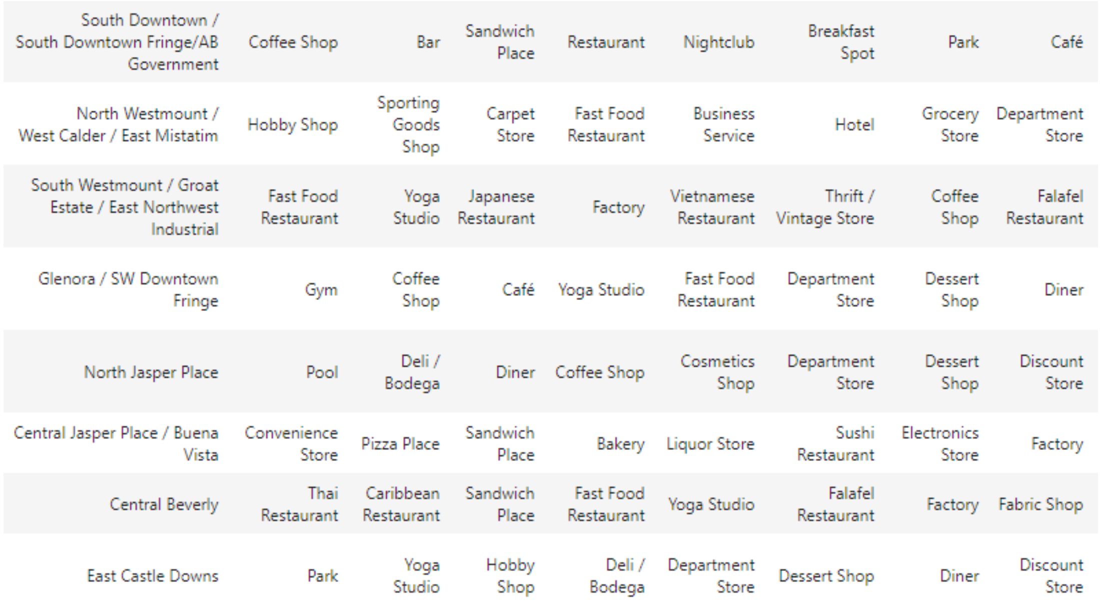
Neighborhoods clustered similarly in Calgary and Edmonton. Most neighborhoods clustered together with comparable venues.

Sample of largest cluster of Calgary neighborhoods based on venues patterns



Bridgeland / Greenview / Zoo / YYC	Furniture / Home Store	Italian Restaurant	Hardware Store	Gourmet Shop	Breakfast Spot	French Restaurant	Discount Store	Donut Shop	Eastern European Restaurant
Inglewood / Burnsland / Chinatown / East Victoria Park / Saddledome	Sporting Goods Shop	Brewery	Comedy Club	Farmers Market	Sports Bar	Café	French Restaurant	Discount Store	Donut Shop
Highfield / Burns Industrial	Warehouse Store	Pizza Place	Asian Restaurant	Discount Store	Coffee Shop	Fast Food Restaurant	French Restaurant	Diner	Donut Shop
Queensland / Lake Bonavista / Willow Park / Acadia	Dance Studio	Baseball Field	Chinese Restaurant	Furniture / Home Store	Hardware Store	Food Court	Discount Store	Donut Shop	Eastern European Restaurant
Thorncliffe / Tuxedo Park	Liquor Store	Convenience Store	Bank	Coffee Shop	Supermarket	Vietnamese Restaurant	Pharmacy	Sandwich Place	Discount Store
Mount Pleasant / Capitol Hill / Banff Trail	Massage Studio	Vietnamese Restaurant	Pub	Rental Car Location	Gas Station	Bookstore	Mediterranean Restaurant	Yoga Studio	Diner

Sample of largest cluster of Edmonton neighborhoods based on venue patterns



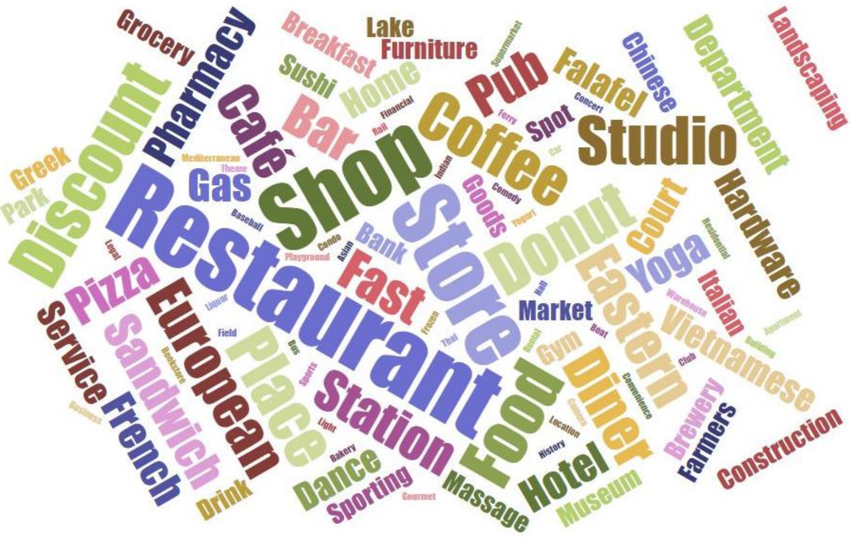
South Downtown / South Downtown Fringe/AB Government	Coffee Shop	Bar	Sandwich Place	Restaurant	Nightclub	Breakfast Spot	Park	Café
North Westmount / West Calder / East Mistatim	Hobby Shop	Sporting Goods Shop	Carpet Store	Fast Food Restaurant	Business Service	Hotel	Grocery Store	Department Store
South Westmount / Groat Estate / East Northwest Industrial	Fast Food Restaurant	Yoga Studio	Japanese Restaurant	Factory	Vietnamese Restaurant	Thrift / Vintage Store	Coffee Shop	Falafel Restaurant
Glenora / SW Downtown Fringe	Gym	Coffee Shop	Café	Yoga Studio	Fast Food Restaurant	Department Store	Dessert Shop	Diner
North Jasper Place	Pool	Deli / Bodega	Diner	Coffee Shop	Cosmetics Shop	Department Store	Dessert Shop	Discount Store
Central Jasper Place / Buena Vista	Convenience Store	Pizza Place	Sandwich Place	Bakery	Liquor Store	Sushi Restaurant	Electronics Store	Factory
Central Beverly	Thai Restaurant	Caribbean Restaurant	Sandwich Place	Fast Food Restaurant	Yoga Studio	Falafel Restaurant	Factory	Fabric Shop
East Castle Downs	Park	Yoga Studio	Hobby Shop	Deli / Bodega	Department Store	Dessert Shop	Diner	Discount Store

Additionally, only 5 neighborhoods in each city formed a cluster outside of the largest cluster, representing mostly industrial areas.

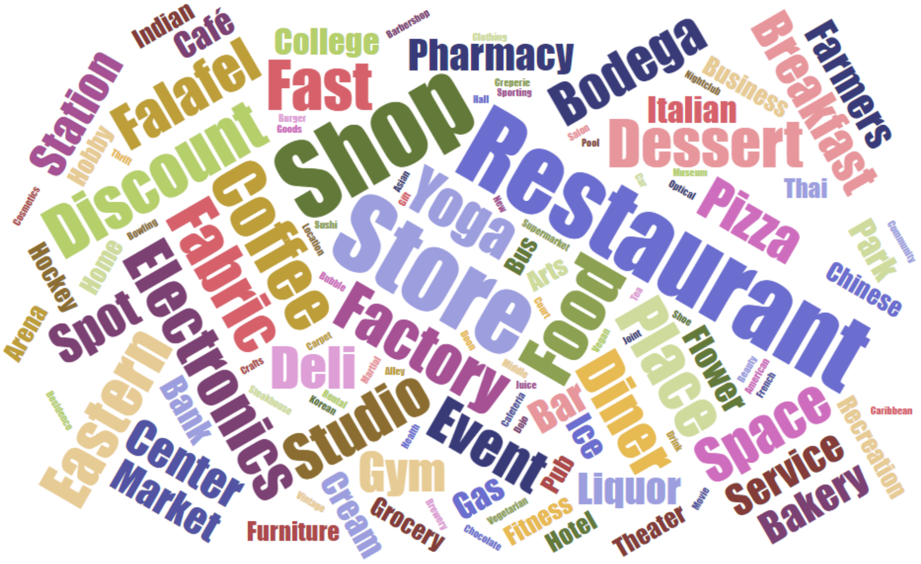
Word Cloud Results

A word cloud derived from the venue categories in the largest cluster for each city show results expected in any urban community and indicate similarities of the neighborhoods in Calgary and Edmonton.

Word cloud derived from the largest cluster of Calgary Neighborhoods

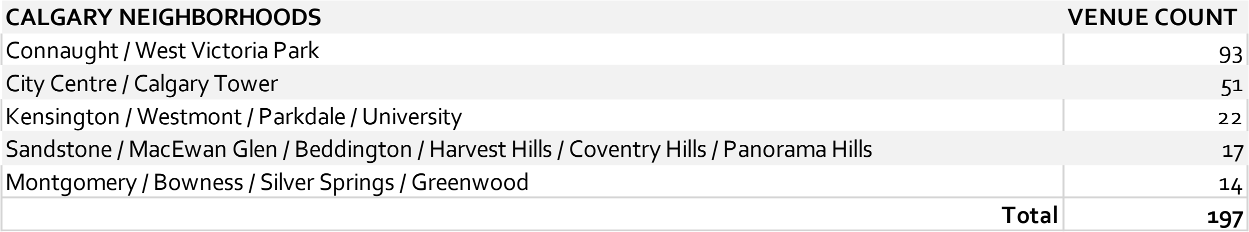


Word cloud derived from the largest cluster of Edmonton Neighborhoods

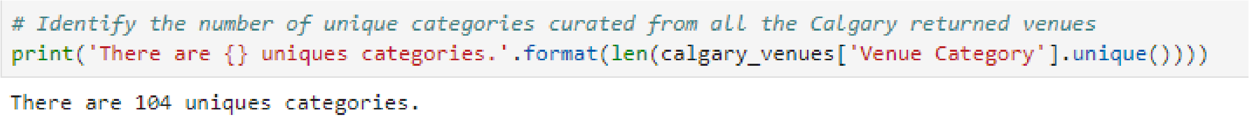


Results of the FourSquare venue count of the top 5 neighborhoods

Venue count of top 5 neighborhoods in Calgary



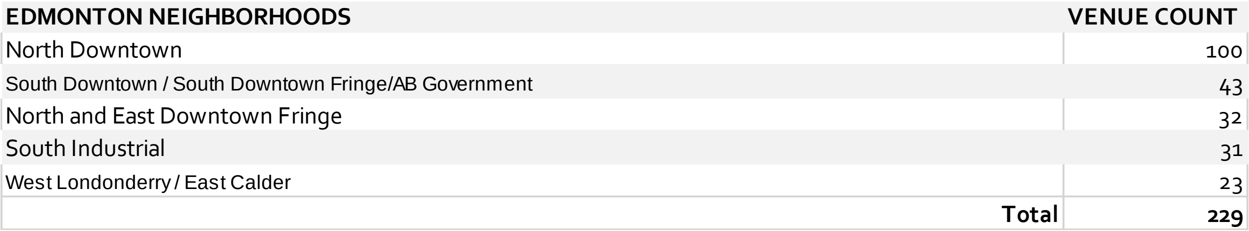
CALGARY NEIGHBORHOODS	VENUE COUNT
Connaught / West Victoria Park	93
City Centre / Calgary Tower	51
Kensington / Westmont / Parkdale / University	22
Sandstone / MacEwan Glen / Beddington / Harvest Hills / Coventry Hills / Panorama Hills	17
Montgomery / Bowness / Silver Springs / Greenwood	14
Total	**197**



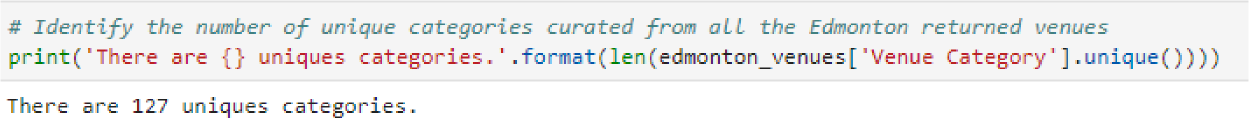
```
# Identify the number of unique categories curated from all the Calgary returned venues
print('There are {} uniques categories.'.format(len(calgary_venues['Venue Category'].unique())))
```

```
There are 104 uniques categories.
```

Venue count of top 5 neighborhoods in Edmonton



EDMONTON NEIGHBORHOODS	VENUE COUNT
North Downtown	100
South Downtown / South Downtown Fringe/AB Government	43
North and East Downtown Fringe	32
South Industrial	31
West Londonderry / East Calder	23
Total	**229**



```
# Identify the number of unique categories curated from all the Edmonton returned venues
print('There are {} uniques categories.'.format(len(edmonton_venues['Venue Category'].unique())))
```

```
There are 127 uniques categories.
```

Result Highlights

- Calgary population is ~6% > Edmonton
- Edmonton has 16% > total venues per 5 top neighborhoods
- Edmonton has 22% more unique venue categories
- Venue categories may not be comparable across the two cities

Discussion

This analysis compares Calgary and Edmonton neighborhoods with data derived from FourSquare to determine similarities/differences in the number and categories of venues in each city. Overall, it presents a process for web scraping data and using geocoding services to preprocess data and preparation for data extraction using the FourSquare API. The data is then clustered into similar neighborhoods using the K-means clustering method, an unsupervised machine learning algorithm used to identify underlying patterns of the clusters.

The results show that despite a population deficit of ~85k, Edmonton took the lead in total number of venues by 16%. Edmonton also leads in the number of unique categories by 22%. This is where the FourSquare venue categories is questionable and deserves further investigation.

Although the word cloud results appear similar, there are 23 venue categories that exist only Edmonton. Does this mean that Edmonton has a more diverse categories of venues or does it imply that the FourSquare categories are applied differently in each city?

A quick observation suggests that at least some similar venues are categorized differently in each city. For instance, Calgary has a ‘Donut’ category with 8 venues assigned to it and in Edmonton this category does not exist. Does this mean that people don not eat donuts in Edmonton or does it imply a different set of criteria is used to assign a category in each city? The presence of Tim Horton shops in Edmonton suggests the later.

In another example, Edmonton has a Deli / Bodega category that contains 8 venues and this category does not exist in the Calgary data. This does not however preclude that Calgary does not have any venues that satisfy this category, rather they are simply assigned to an existing category by an unknown set of criteria.

These examples suggest the FourSquare data used in this analysis lacks consistency and therefore will impact the results. Further analysis is required to determine the criteria used by FourSquare to assign venues to categories in each city.

Conclusion

In conclusion, this analysis was intended to be a simplistic and lighthearted approach to the question: “Calgary or Edmonton – Which city is better to live in?”

A severe limitation to this approach lies both in its simplicity and in data accuracy. A future approach would include other socioeconomic factors such as real estate and rental prices, commute time, weather, walk ability of neighborhoods, etc. to provide a more holistic approach. Also, data accuracy is dependent on data retrieved from FourSquare, concerns of inconsistency previously discussed would require further investigation. Therefore, the results of this analysis are inconclusive.

The question remains: “Calgary or Edmonton – Which city is better to live in?”

This analysis was designed to meet the Battle of the Neighborhoods capstone requirement for the IBM Data Science Professional Certificate offered by Coursera. These requirements were meet through the following steps:

- Data acquisition and Preparation: webscraping, utilize various python libraries
- Process data: geocode data to pass to FourSquare API for venue data retrieval
- Data Exploration: using Folium to visualize data
- Analysis: Machine learning techniques one hot encoding and k-means clustering to uncover patterns
- Communicate results

Data science can be used to answer a myriad of questions and there is plenty of free data available to satisfy the most curious minds. Go ahead and start exploring, you will find insights into your most pressing questions.