City Comparer

A CAPSTONE Project for Data Science Professional Certification - compare cities to help users envision what to expect when they move between from one city to another. Provide this comparison using the publicly available data.

Sandeep Wadhwa

Table of Contents

Introduction

The audience of this analysis are people who are contemplating a move across 2 cities and would like to compare their current city and prospective city on various dimensions.
People keep moving to different cities/state due to job or studies requirements. Many of these people, face challenge in terms of how the new city will be and what to expect after they move. If the city is in the same state, they may have an idea about what to expect after they move, but if the move is to a different state, they have not been to the city before and hence do not know what to expect after the move.

One-way people try to solve this is by researching online, the problem is that they cannot compare so many data points side by side unless they create a comparison table and visit many different resources. For example:

- check Wikipedia to get information on population and ethnicity of population,
- check Google Maps to know types of venues (restaurant etc) in the city
- they may want to know whether restaurants for a particular cuisine exists
- check Google/other sites to know what are the points of interest (lake, mountains, downtown etc) around the place they plan to move
- research whether the city supports specific type of activity that they love to do, such as cycling, hiking etc.

The analysis will bring the data from different sources, Wikipedia, Foursquare, Weather.com etc to compare 2 cities side-by-side and provide insights to the person, so that they can envision what to expect after the move and decide whether the move is worth or not.

Data

I plan to use the following data for this analysis:

Data Source	Data to be extracted
Wikipedia/other resources	Population (Ethnicity, if available)
Four Square	Venues in the city and comparison. Cluster of venues and their comparison
Weather.com	Weather data for comparison
Lonely Planet/Expedia/Wikipedia/Google Maps	Point of interest around the city

Zillow	Range of rent in the city and average rent.

While the above data sources were planned, over the course of this project it was discovered that some of the data source couldn't be used due to complicated and inconsistent data clean-up required.

Below are challenges faced while extracting data from different sources:

Data Source	Challenge
Demographic Information	Available on Wikipedia, but is inconsistently provided that requires complicated data clean-up tasks.
Weather.com	Time Constraint
Wikipedia/Google Maps	Time Constraint
Zillow	No Free Version Available.

Methodology

Based on authors interest, two cities 1) Fremont (California, USA), and 2) Saskatoon (Saskatchewan, Canada) were chosen across US and Canada for the comparison. This also provided opportunity to compare the cities across two different countries and providing a glimpse of issues that would exist in dataset when comparing cities across countries.

There are 2 types of data identified that can be compared in following ways to help users get picture of the cities.

1. **Population Metrics:** this includes population density and total population that can be directly compared against each other. This data will extracted from Wikipedia, cleaned and used as is for comparison.

2. **Venue Data:** there are 3 types of comparison that is possible, described as following. For the purpose of this analysis, list of neighborhood will be extracted from Wikipedia and coordinates will be extracted using GeoCode API, and lastly FourSquare API will be used to get venues for each of these neighborhoods.

    a. Compare the total number of venues – this can be compared directly.
    b. Compare the Common venues across the cities – this can be compared directly by looking at the venues and their frequency

c. Compare the Top venues across the cities – this can be compared directly by looking at the venues and their frequency

d. Based on the venues in the neighborhood of cities, segment the neighborhoods – this can be done based on the clustering of venue data

e. Compare the clusters of venues using Cosine Similarity – this can be achieved using the cosine similarity in the venues data for two cities.

“Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. It is defined to equal the cosine of the angle between them, which is also the same as the inner product of the same vectors normalized to both have length 1. The cosine of 0° is 1, and it is less than 1 for any angle in the interval (0, π] radians. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.” [1]

[1] Sourced from Wikipedia

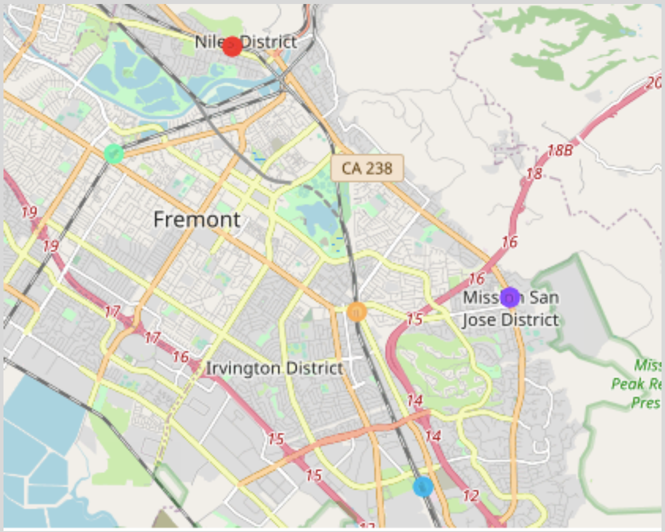
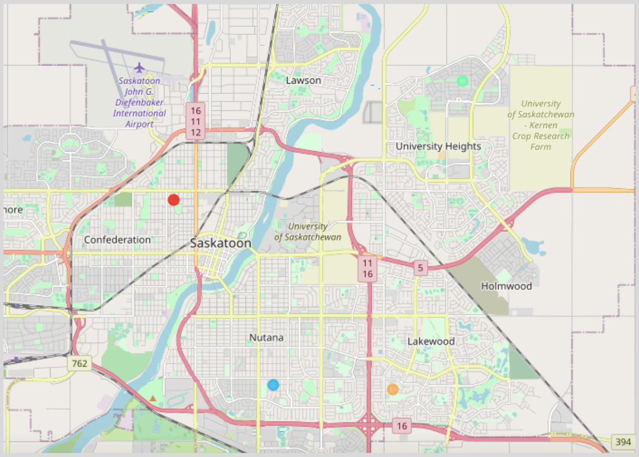
Results

Based on the venue data extracted from FourSquare for the 2 cities, the cities are quite similar.

SIMILARITY INDEX = 57% (using Method 1)[2]

SIMILARITY INDEX = 34% (using Method 2)[3]

Below are detailed results from this analysis:

	City 1 (Fremont)	City 2 (Saskatoon)
MAP	 Niles District CA 238 18B 18 16 Fremont 19 17 Mission San Jose District 15 Irvington District 14 12	 Saskatoon John G. Diefenbaker International Airport Lawson 16 11 12 University Heights University of Saskatchewan - Kernen Crop Research Farm Confederation Saskatoon University of Saskatchewan 11 16 5 Holmwood Nutana Lakewood 762 16 394
POPULATION	246,376	214,089
POPULATION DENSITY	2,797/sq mi	2,400/sq mi

[2] Method 1 – using the cosine_similarity in sklearn package
[3] Method 2 – using the cosine in scipy.spatial package

The different neighborhoods in these cities can be allocated to each of the cluster identified, except cluster 1, which appears only in Fremont City.

Neighborhood	Evergreen	Adelaide	Mayfair	Lakeview	Centerville	Niles	Irvington	Warm Springs	Mission San Jose
City	Saskatoon	Saskatoon	Saskatoon	Saskatoon	Fremont	Fremont	Fremont	Fremont	Fremont
Cluster Labels	3	2	0	4	3	0	4	2	1
1st Most Common Venue	Coffee Shop	Pub	Hotel	Pub	Coffee Shop	Park	Coffee Shop	Coffee Shop	Coffee Shop
2nd Most Common Venue	Hotel	Coffee Shop	Restaurant	Coffee Shop	Grocery Store	Coffee Shop	Grocery Store	Fast Food Restaurant	Ice Cream Shop
3rd Most Common Venue	Restaurant	Café	Coffee Shop	Café	Sushi Restaurant	Grocery Store	Ice Cream Shop	Gym	Trail
4th Most Common Venue	Bakery	Grocery Store	Pub	Pizza Place	Bakery	Breakfast Spot	Fast Food Restaurant	Grocery Store	Bakery
5th Most Common Venue	Pub	Pizza Place	Café	Mexican Restaurant	Ice Cream Shop	Sushi Restaurant	Gym	Sushi Restaurant	Fast Food Restaurant
6th Most Common Venue	Asian Restaurant	Restaurant	Bakery	Restaurant	Park	Bakery	Pizza Place	Juice Bar	Grocery Store
7th Most Common Venue	Café	Mexican Restaurant	Steakhouse	American Restaurant	Thai Restaurant	Fast Food Restaurant	Trail	Trail	Gym
8th Most Common Venue	Pizza Place	Sandwich Place	Pizza Place	Bakery	Breakfast Spot	Chinese Restaurant	Bakery	Thai Restaurant	Mexican Restaurant
9th Most Common Venue	American Restaurant	Breakfast Spot	Asian Restaurant	Hotel	Trail	Mexican Restaurant	Falafel Restaurant	Bakery	Pizza Place
10th Most Common Venue	Grocery Store	Bookstore	American Restaurant	Breakfast Spot	Mexican Restaurant	Thai Restaurant	Park	Mexican Restaurant	Falafel Restaurant

Discussion

I used two methods to get the similarity between the venues across these cities.

Method 1 – using the cosine_similarity in sklearn package, which works on one-hot encoded data – we are able to identify that in terms of venue types, the cities are similar and there is 57% similarity in terms of the types of venues across these cities. However, this doesn't account for the frequency of the venues. Therefore, we use another method that considers the frequency of venues.

Method 2 – using the cosine in scipy.spatial package, we are able to measure the similarity based on the frequency of the venues.

To investigate this further, additional analysis was done, which is as following.

Top 10 Venues

Based on the top 10 venues across the cities, it is clear the Coffee Shop has the highest frequency across the cities and it is the top venue for both the cities. However, that is the only similar point, for rest of the venues, both the cities have lot of differences. For example, Fremont has Park, different types of restaurant and no hotels; Saskatchewan has hotel, Pizza Place and Pubs.

City	Fremont, California, USA
Coffee Shop	34
Grocery Store	24
Park	19
Ice Cream Shop	19
Bakery	18
Fast Food Restaurant	17
Trail	15
Sushi Restaurant	15
Thai Restaurant	14
Mexican Restaurant	13

City	**Saskatoon, Saskatchewan, Canada**
Coffee Shop	18
Pub	16
Hotel	14
Café	14
Restaurant	14
Bakery	12
Pizza Place	12
American Restaurant	10
Grocery Store	9
Asian Restaurant	9

Exclusive Venues

Next, we find the venues that are exclusive to a city and do not exist across the cities.

There are many venues that SASKATOON city lacks compared to Fremont. Most prominently – Parks, Sushi Restaurant, Falafal Restaurant, Pet Stores, Lake and Spa. Similarly, there are many venues that FREMONT city lacks compared to Saskatoon. Most prominently – Steakhouse, Gluten-Free Restaurant, Bar, and Brazilian, Cajun, and Caribbean Restaurants.

City	Fremont, California, USA	Saskatoon, Saskatchewan, Canada
Park	19	0
Sushi Restaurant	15	0
Falafel Restaurant	12	0
Pet Store	10	0
Lake	7	0
Spa	7	0
Dessert Shop	6	0
Donut Shop	6	0
Japanese Restaurant	6	0
Juice Bar	6	0
Shanghai Restaurant	6	0
Burmese Restaurant	5	0
Dog Run	5	0
Library	5	0

City	Fremont, California, USA	Saskatoon, Saskatchewan, Canada
Steakhouse	0	8
Gluten-free Restaurant	0	5
Bar	0	4
Brazilian Restaurant	0	4
Cajun / Creole Restaurant	0	4
Caribbean Restaurant	0	4
Cheese Shop	0	4
Gift Shop	0	4
Indie Movie Theater	0	4
Italian Restaurant	0	4
New American Restaurant	0	4
Zoo	0	4
Gastropub	0	3
Golf Course	0	3

Based on this analysis, I found that we can use two different methods to measure the similarity across the cities based on the types of venues – 1) based on the type of venues without frequency and 2) based on the type of venues and frequency of venues. I am sure this will be very helpful in future version of this project.

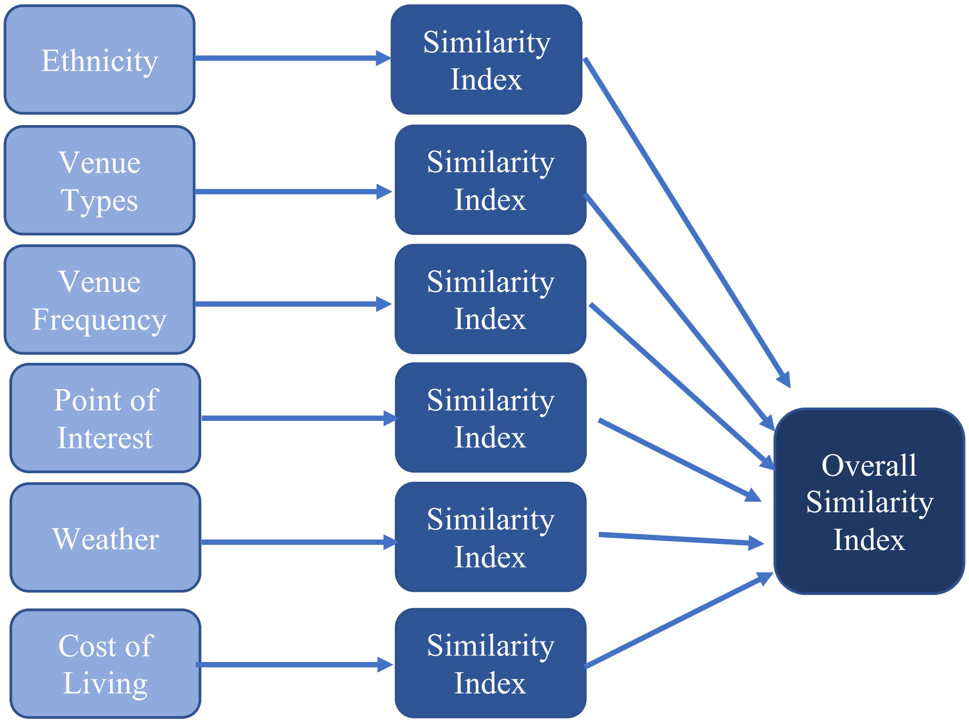
Conclusion

- Using the publicly available data and using analytics on it, the overall comparison across the cities was made easier as there was less noise and more meaningful information.
- More information can be included such as demographics, weather and point of interests (as originally planned) between the cities. However, the data is not available in easy to consume format and more work is required to make this data available.
- Specifically, data across different countries are available on different platforms in different formats – making the data acquisition a time consuming issue.

Future

- The original idea behind this project was to convert this into a tool that can be made available to general public. However, in current state of data, which is not easily consumable, it is not possible to make this available in tool format. For example, on Wikipedia same information is available differently for these 2 cities, which means same code cannot work to extract the information for both the cities. And hence making it difficult to convert this into a tool.

- In future, if this data is made available easily, the comparison using this method can be enhanced and made available in easy to consume format.
    - Furthermore, capability to compare more than 2 cities can also be added when data is easily available.
    - Similarity index can be derived separately for other comparison points such as ethnicity, point of interest etc.
    - A final Similarity Index can be derived based aggregation of similarity indices on each dimension (as shown below).


Ethnicity
Similarity Index
Venue Types
Similarity Index
Venue Frequency
Similarity Index
Point of Interest
Similarity Index
Weather
Similarity Index
Cost of Living
Similarity Index
Overall Similarity Index