# Analysis of Different Cities for Deployment of Successful Experimentation

Varun Rao

April 27th 2020

## 1. Introduction/Business Problem:

When companies deploy new projects, it is done through experimentation. An experiment is run in a controlled environment OR in many cases in a suitable pilot location, and if successful, it is expanded to other locations. While deciding what other locations one should deploy such projects, it is always beneficial to have a good comparison of target locations with the location already experimented on, based on appropriate parameters. This can be achieved by realizing how similar or dis-similar the locations are.

In this project we will aim to solve three separate scenarios (having the same problem above) through Clustering:

> 1. A company like Amazon/Walmart were thinking long-term and wanted to have their own logistics fleet for better control and reliability. They ran an experiment in New York City to determine where to establish new package drop-off locations, to make it more convenient for their customers to return something. The experiment was a success and they have recognized that best target cities are Toronto and San Francisco. We solve the problem of what areas in Toronto and San Francisco should they target to set up these drop off locations.
>
> 2. A company like Trip Advisor wants to recommend me locations in San Francisco based on what I did in New York City. We solve the problem of recommending locations in a new city based on similarities in the previous one.
>
> 3. Clusters in different cities can also help companies determine where, in different cities, they can deploy self-driving cars to pilot test (based on successful experimentation for pilot tests in one city). Let's consider Waymo, Uber or Tesla are at that stage where level 5 cars can be tested on roads with real customers. They would like to know where they can deploy their cars, after a successful test (again) in New York City. They have identified that San Francisco and Toronto are two metros that are open to this. OR This can even help Ride Share companies like Uber and Lyft to deploy more cars around a particular area specifically for a certain **type** of service (like Ride Share). So people would much rather take a shared Uber/Lyft rather than taking a bus for example. Here we solve the problem of Identifying where they can deploy specific type of service in San Francisco based on experimentation in NYC.

**Note there are a lot of assumptions made to create these scenarios and the focus of "this" work is clustering Neighborhoods based on top 10 venues (in each neighborhood) in these three cities to see what are similar areas between the three and how they can be identified after the target locations (San Francisco and Toronto) have been identified.*

**This study does not go into too much detail, and does not take into consideration a lot of real world factors and other data. This is meant to be a presentation of Data Science and Machine Learning abilities to showcase academic comprehension of the subject matter and tools.*

## 2. Data:

In order to cluster different areas of the three cities mentioned above, we will consider each area within a city as a "neighborhood". The data set for all three cities was cleaned from sources (below), to represent the 'Neighborhood' with corresponding latitudes and longitudes of each neighborhood's approximate center and is stored in "**combined_df**" data frame. Details on three Individual data sources as follows:

1. Toronto Dataset: This data frame in csv was combined using two different datasets:
   - Postal Codes List in Toronto (which was exported and saves as an xslx file). This file contained the postal codes and corresponding Neighborhood names. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   - Geospatial File that has the latitude and longitudes and has been provided by Coursera http://cocl.us/Geospatial_data
2. New York Dataset: This data set was extracted from https://geo.nyu.edu/catalog/nyu_2451_34572
3. San Francisco Dataset: This data was extracted from https://geodata.lib.berkeley.edu/catalog/ark28722-s75c8t

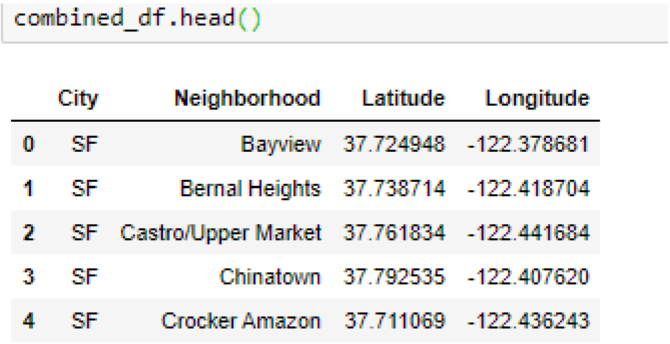
We will use the three datasets to create one merged dataset that has all neighborhoods of all three cities. We will extract Venues for each neighborhood from Foursquare and cluster the neighborhoods to determine similarities in inter-city neighborhoods and intra-city neighborhoods. We will focus more in "inter-city" relationships as it will be used to solve all three problems stated in the first section.

## 3. Methodology:

This section explains the main body of the work conducted in the project. It is divided into 6 sub-sections After these sub-sections, the next main section "RESULTS", will be portrayed.

Before diving into details, to summarize the Methodology, Neighborhood and Location data for three cities (NYC, SF and Toronto) was taken and combined into a single data frame. Venues were extracted for these Neighborhoods from Foursquare and each Neighborhood was explored based on top venues. Then, partitioning clustering (K-means) was used to cluster the neighborhoods into 'four' clusters to see similarity of clusters between different neighborhoods in different cities. The result of this clustering will enable one to make judgements about what location - in new cities (SF, Toronto) can be used to deploy an experiment that succeeded in a pilot-city (NYC).

1. Importing and preparing all three data sets (NY, SF, Toronto): As mentioned in the data section, all three datasets were publically available and were formatted and cleaned to get a combined data frame "combined_df" which consisted of four columns, 'City', 'Neighborhood', 'Latitude', and 'Longitude'.



```
combined_df.head()
```

| | City | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| 0 | SF | Bayview | 37.724948 | -122.378681 |
| 1 | SF | Bernal Heights | 37.738714 | -122.418704 |
| 2 | SF | Castro/Upper Market | 37.761834 | -122.441684 |
| 3 | SF | Chinatown | 37.792535 | -122.407620 |
| 4 | SF | Crocker Amazon | 37.711069 | -122.436243 |

Fig 1. combined_df

2. <u>Creating Maps and Pinouts to visualize the data:</u> A Map of North America was created to visualize the Neighborhood in three cities using Folium. This ensured all the neighborhoods were imported properly and the location data was correct.

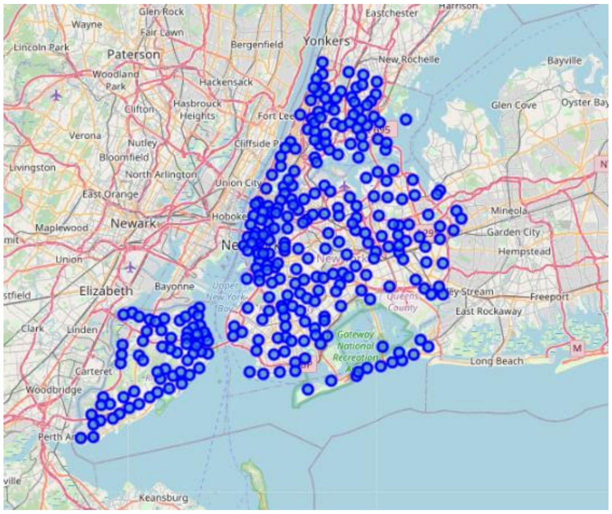


Fig 2. NYC Neighborhoods

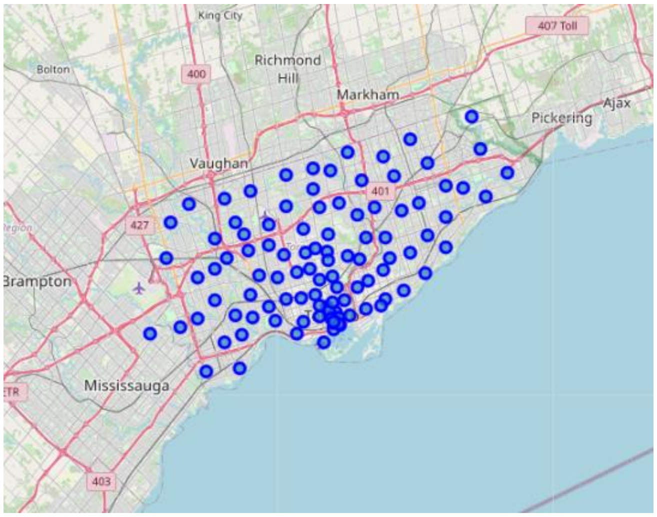


Fig 3. Toronto Neighborhoods

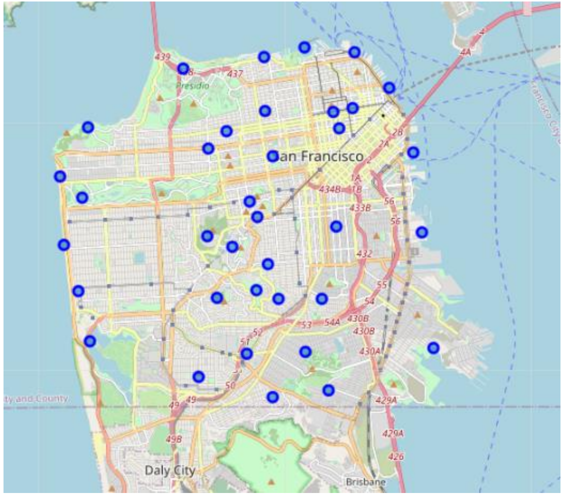
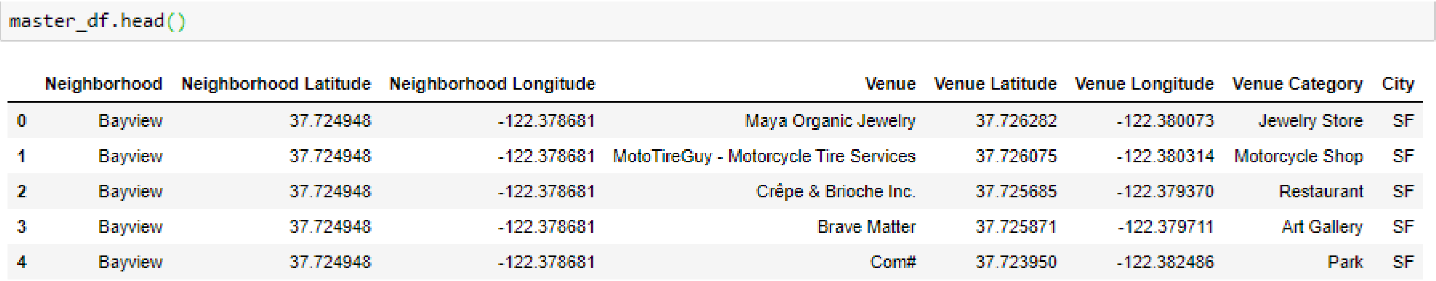


Fig 4. San Francisco Neighborhoods

3. Using FourSquare to get the Venue information for combined_df: FourSquare is an application that has data for Venues and Users and Experiences. Using this app, data for Venues (limited to 100 venues per neighborhood) was extracted using API calls (more on APIs here https://developer.foursquare.com/docs/). The main interest of this project was to segment the Neighborhoods based on different types of Venues in that Neighborhood. Hence 'Venue_Categories' were exported and cleaned up to create a master_df containing all Cities, Neighborhoods, Venues and Venue Categories, along with their location information.



```
master_df.head()
```

|  | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category | City |
|---|---|---|---|---|---|---|---|---|
| 0 | Bayview | 37.724948 | -122.378681 | Maya Organic Jewelry | 37.726282 | -122.380073 | Jewelry Store | SF |
| 1 | Bayview | 37.724948 | -122.378681 | MotoTireGuy - Motorcycle Tire Services | 37.726075 | -122.380314 | Motorcycle Shop | SF |
| 2 | Bayview | 37.724948 | -122.378681 | Crêpe & Brioche Inc. | 37.725685 | -122.379370 | Restaurant | SF |
| 3 | Bayview | 37.724948 | -122.378681 | Brave Matter | 37.725871 | -122.379711 | Art Gallery | SF |
| 4 | Bayview | 37.724948 | -122.378681 | Com# | 37.723950 | -122.382486 | Park | SF |

Fig 5. master_df

4. Create data frame for Clustering: The goal of this sub-section, was to have a data frame, that could be used for Clustering. This data frame is "master_grouped" and was created by grouping "one_hot encoded venues table", by Neighborhood and City to maintain uniqueness of each Neighborhood. Also another data frame was created, that would show the top 10 Venues by Neighborhood, so we can add our Cluster labels to this data frame to analyze what each cluster contains (after clustering). This data frame is "neighborhoods_venues_sorted".

5. Clustering: Once the data frames above were available, partitioning Clustering (K-means) was used to create 6 clusters of Neighborhoods. Cluster labels were inserted to the "neighborhoods_venues_sorted" and this data frame was merged with "combined_df" which contained latitudes and longitudes (location) information and a final data frame "master_merged" was created that could be analyzed for results. Scikitlearn package was used for clustering algorithm.

| Cluster Label | Color on Map | # of Neighborhoods across all three cities |
|---|---|---|
| 0 | Red | 23 |
| 1 | Purple | 1 |
| 2 | Blue | 20 |
| 3 | Light Blue(Turquoise) | 2 |
| 4 | Light Green | 5 |
| 5 | Orange | 372 |

Fig 6. Clustering Summary

6. Created a MAP for visualization: Finally, another Map was created using folium to visualize the clusters and how they were spread in different cities.

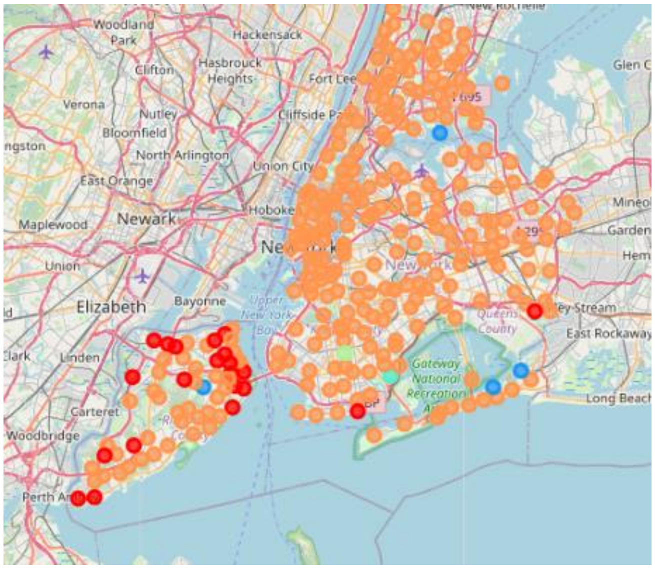


Fig 7. NYC Clusters

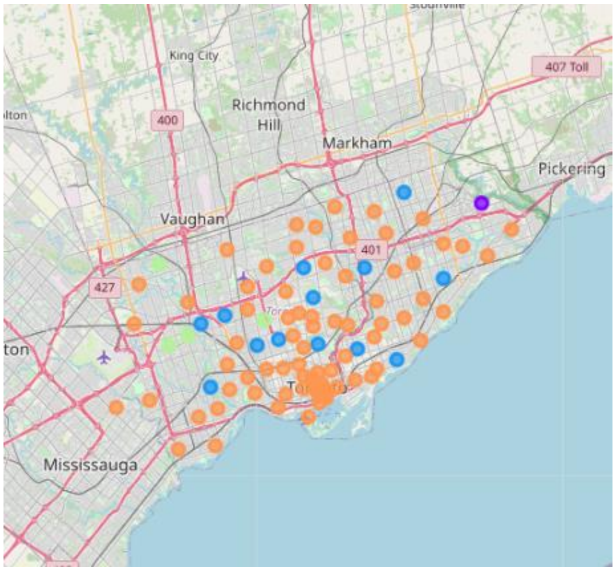


Fig 8. Toronto Clusters

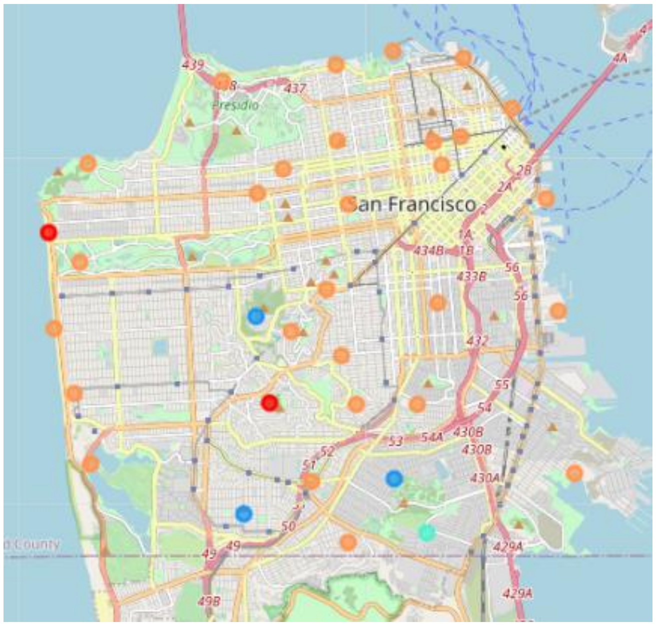


Fig 9. San Francisco Clusters

## 4. Results & Discussion:

The data successfully divided the Neighborhoods in all three cities into clusters. Clusters 0 to 5. Of which three clusters will be considered for analysis (0, 2 & 5)

1. Cluster 0 (RED Colored Clusters): Theses were areas where the most popular venue was a Bus Stops. There were two main locations in SF that fit this category. This is a bit surprising but these come up as a lot of Students/Tourists who want to travel from the City to these two destinations, end up getting here taking multiple modes of transport and prefer Buses (single mode of transport, even though its slower) to get back to their residence as they are tired after a day in the beach or walking/hiking around twin peaks.

2. Cluster 1 (PURPLE Colored Clusters): Just had one location. So we will not consider this.

3. Cluster 2 (DARK BLUE Colored Clusters): These areas had trails, parks, playgrounds and Zoos. These are places where people visit to for outdoor activities or as weekend trips. Also a lot of tourists visit these places to see local nature and points of interest. This cluster is a good recommendation that companies like Trip Advisor will give to a person who enjoys outdoor activities and is visiting San Francisco or Toronto.

4. Cluster 3 (LIGHT BLUE (Turquoise) Colored Clusters): Just had 2 locations. So we will not consider this.

5. Cluster 4 (LIGHT GREEN Colored Clusters): Pizza Places. Not of much use for our applications. So we will not consider this.

6. Cluster 5 (ORANGE Colored Clusters): Restaurants, Coffee Shops, Cafe's, Convenience Stores, Pubs & Bars, etc. This is the most prominent cluster as all three locations are densely populated cities. These are the places where lot of movement happens. People are constantly going to Cafe's and Coffee shops on a regular basis. Neighborhood specific Cafe's, Coffee Shops, Bakeries and Restaurants serve as a great place for people to regularly go with friends and family. These locations can be used by Amazon/Walmart to establish package drop-off locations.

** For detailed results, please refer to Notebook link RESULTS section:*
*https://github.com/varunrao1989/Coursera_Capstone/blob/master/Analysis%20of%20Different%20Cities%20for%20Deployment%20of%20Successful%20Experimentation.ipynb*

## 5. Conclusion:

This study successfully segmented Neighborhoods, based on similarities in most popular Venues (between cities) which helped companies make decisions on what area to deploy a project in a new city based on successful experimentation in a pilot - city. This was extended to three cases introduced in the first section.

1. Companies like Amazon/Walmart were able to deploy "package drop-off locations" around Neighborhood venues that belong to cluster 5 (Orange Colored Clusters) in Toronto and San Francisco.

2. Trip Advisor (or similar) was able to give good relevant recommends - cluster 2 (Dark Blue Colored Clusters) to to its users who were into outdoor activities and were either tourists or just exploring the city of Toronto and San Francisco.

3. Companies like Uber/Lyft or Super Shuttle were able to deploy new Ride Share service routes in San Francisco - cluster 0 (Red Colored Clusters) where majority of the population would wait around a bus stop and get a single mode of transportation, rather than use other modes of transport where they would have to change stations/modes multiple times.

** Again, this study does not go into too much detail, and does not take into consideration a lot of real-world factors and other data. This is meant to be a presentation of Data Science and Machine Learning abilities to showcase academic comprehension of the subject and tools.*

*****End of Report*****