Analysis of Different Cities (for deployment of Successful Experimentation) using Clustering

VARUN RAO

04/27/2020

Introduction

This study provides a solution for projects & applications that are deployed in different cities and locations world wide, and serve a twofold purpose:

Real World projects are expensive. They need a lot of planning, co-ordination, time and resources to get things going.

- Although experimentation by itself is expensive, deploying the project after successful experimentation is more expensive and cost of failure is the most expensive.
- Hence, assuming the experimentation was a success, it would be beneficial to simulate and plan the deployment of a project, based on factual data driven approaches, which not only will significantly lower the probability of failure of a project but will also serve as the starting point for deployment.
- Understanding the similarities between the 'experimentation location' and 'target market' location will benefit in such cases.

The second application of this study is for location based recommendation systems, which suggests locations in a new place, similar to what one liked in another place.

Business Problems

In this project we will aim to solve three separate scenarios through Clustering:

- 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.*

Path to Solution: Data and Methodology

In order to achieve these goals, we can break down the locations that we want to compare, into a viable resolution, and compare those locations based on some parameters that are related to the project. This will enable us to pin point areas in 'target-markets' that will be similar to the 'experimentation location' where probability of success will increase.

Data:

- This study takes into consideration the 'experimentation location' as New York City and 'target-markets' as San Francisco and Toronto. The resolution in these three cities goes down to the 'neighborhood' level. A full in-depth analysis of the data can be found in this link below. [Report.https://github.com/varunrao1989/Coursera_Capstone/blob/master/Analysis%20of%20Different%20Cities%20for%20Deployment%20of%20Successful%20Experimentation.pdf](https://github.com/varunrao1989/Coursera_Capstone/blob/master/Analysis%20of%20Different%20Cities%20for%20Deployment%20of%20Successful%20Experimentation.pdf)

Methodology:

- Venues were extracted for each neighborhood from Four Square application and was categorized accordingly. Then "K-means" clustering was used to group the data by venue, in three cities, so that comparisons can be made, to find locations in San Francisco and Toronto, that are similar to successful locations in New York City. Details can be found in the Methodology section of this link below. [Report.https://github.com/varunrao1989/Coursera_Capstone/blob/master/Analysis%20of%20Different%20Cities%20for%20Deployment%20of%20Successful%20Experimentation.pdf](https://github.com/varunrao1989/Coursera_Capstone/blob/master/Analysis%20of%20Different%20Cities%20for%20Deployment%20of%20Successful%20Experimentation.pdf)

Results - 1

Data was successfully grouped into 6 Clusters. Dots of the same color in different cities represent the neighborhoods that belong to the same cluster.

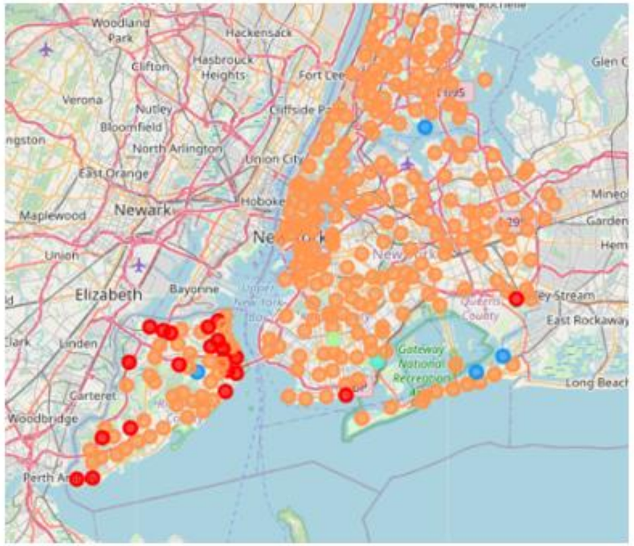

Fig 7. NYC Clusters

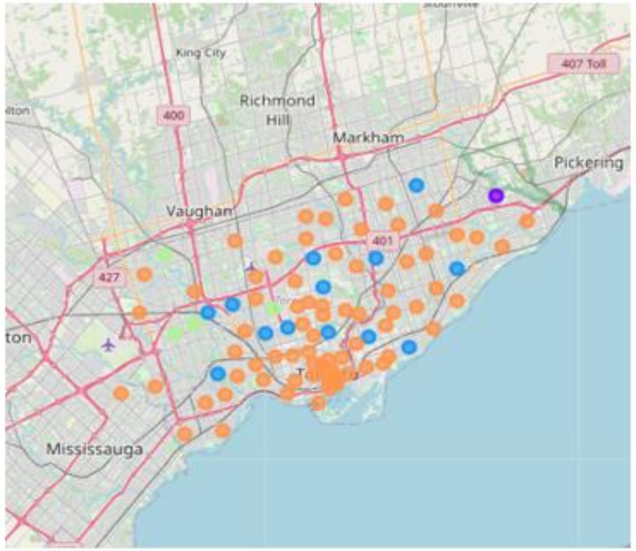

Fig 8. Toronto Clusters

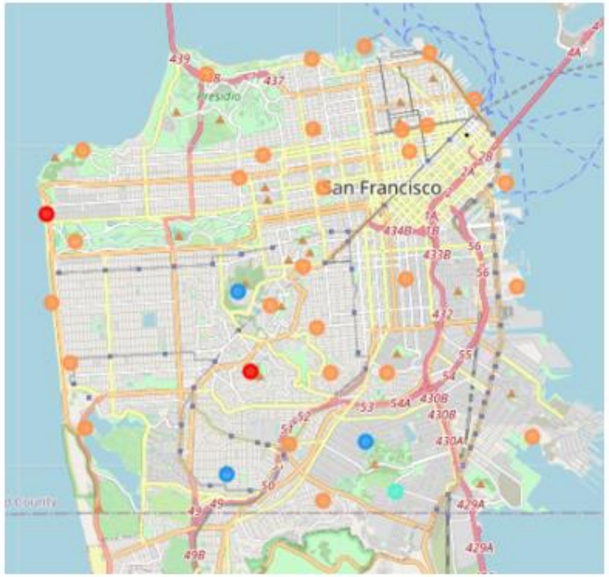

Fig 9. San Francisco Clusters

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

Results - 2

- 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

Conclusion

This study successfully segmented Neighborhoods, based on similarities in most popular Venue categories (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 can solve the following case applications:

- 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.

** 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.*