How Similar are They?

CHICAGO VS. MANHATTAN

Shaurya Chetal

Coursera Capstone

Problem and Discussion of the Background

While it is widely known that New York is the financial center of the world, people sometimes fail to realize that there is more than one financial center within the United States (US) itself, let alone internationally. I'm interested in assessing how similar (or diverse for that matter) different financial capitals are. For this reason, I'd like to compare the Manhattan, NY to Chicago, IL (another important financial center for the US) to see if the makeup of those areas shows similar trends or do other factors take precedence and explain the dissimilarities. Using the clustering analysis, I will compare the results to make inferences about the areas. I hope to learn more about the areas even though I have visited both places numerous times. My target audience for this would be the residents of Manhattan, NY or Chicago, IL to show them how homogenous/diverse (depending on the outcome) they are.

Data Needed and Solution Outline

Based on what we've learned throughout this program, the data needed for my discussion revolves around collecting geospatial data along with some descriptive information (i.e. venue information) linked to the geospatial data. FourSquare is a great platform to collect such information as it provides coordinates as well which will make mapping easier. Additionally, I'll need to scrape "neighborhood" information from web sources that list the different ZIP codes for both Manhattan, NY and Chicago, IL. Once the data is collected, I'll have to clean it and extract only the data that will be pertinent to my discussion. Once, I complete the data cleaning process, I will move on to the analysis stage where I map and cluster the data using k-Means clustering. Finally, using the information computed by

the k-Means clustering, I will make my conclusions about the aforementioned locations by looking at the most common clusters in both locations.

Methodology

The data acquired for this project comes from the City of Chicago who formally defines each neighborhood that was used. The New York dataset originates from New York University's Spatial Data Repository. In essence, both sources are credible as they are derived from verified and reputed institutions.

KMeans clustering was conducted based on different K values. KMeans clustering is a popular unsupervised machine learning algorithm which groups datapoints that are similar hopefully producing patterns. KMeans has an accuracy measure of the clustering. Further, the values are plotted against the accuracy measure and the best value of k is picked using the "elbow method". Both locations produced different optimal values for k which shows diversity between the locations. Both locations were clustered differently based on their respective optimal k value. Once the array was generated after clustering, both locations had their dataframes append to include the new labeling with each neighborhood being given a label showing which cluster they belong to. The new dataframes were mapped for both locations based on the new cluster labels where each cluster was assigned a different color for ease of identifying on the map itself.

Results and Discussion

For the discussion, observations that stand out will be included as well possible explanations to such observations. The recommendation in this case is really more of an informed determination rather than an actionable plan. Additionally, the determination in this context is to discern whether the locations, Manhattan and Chicago, are similar or diverse solely based on venue information from a location intelligence website (FourSquare).

Starting with the "chicago_venues" and "manhattan_venues" shapes, we can see that Manhattan has more datapoints, even though the Manhattan has fewer neighborhoods (as shown in "manhattan_grouped" and "chicago_grouped"). The risk of a few neighborhoods holding more "venues" is mitigated since a limit was put on how many locations can be returned for a single neighborhood. Furthermore, it is evident that Manhattan has more "venues" in each neighborhood compared to Chicago because most neighborhoods in Manhattan met the 50 "venue" limit when looking at the counts grouped by neighborhoods. From the comparison, one can say that the data shows Manhattan has more to do within each neighborhood when compared to Chicago. However, there is also a possibility that there wasn't enough data captured on Chicago's neighborhoods.

When looking the K-means clustering, we find that Chicago and Manhattan have different optimal clusters. Clusters are datapoints that share similarities as detected by the machine learning algorithm. Using the "elbow method", it was found that Chicago's optimal K value was 4 and Manhattan's optimal K value was 12. In this case, optimal means that value that would provide the most accuracy. We can see here that Manhattan needs more clusters to for higher accuracy compared to Chicago. This shows that Manhattan needs more clusters to distinctly separate the neighborhoods as opposed to Chicago. We can

say here that Manhattan is more diverse than Chicago with the machine being able to find more to separate the neighborhoods.

With regards to the mapping of the clusters, we can see that the central areas for both locations are quite similar in that they fall in the same cluster. For Manhattan, moving towards the ends it can be seen that the clusters start to change. On the other hand, Chicago is pretty similar across the map with a significant majority falling under one cluster. In Chicago, around 82% (61 out of the 74) of neighborhoods fall under the same cluster, whereas around 25% (10 out of 39) of neighborhoods fall in the same cluster for Manhattan. Looking at the most common type of venues in the most prominent cluster in each location, Chicago is more food-oriented (i.e. bars, restaurants, groceries) while Manhattan is more fitness-oriented (i.e. gyms and yoga studios).

Conclusion

Based on the analysis of data, one can determine that Manhattan and Chicago are quite diverse locations even though both of them are in the United States. Both locations play an important in the global financial operations and are the most heavily populated cities in the Unites States. However, the findings suggest that the makeups of the neighborhoods are entirely different as Chicago is more homogeneous across the map while Manhattan is similar in the central region. Adding to that, the optimal clusters for Manhattan compared to Chicago is significantly more which proves that Manhattan is more diverse. Finally, comparing the types of venues that exist in the most common cluster of neighborhoods, we see that Chicago leans towards to dining and Manhattan leans towards wellness.

Having said that, it is important to note that using just a single method of analysis is not enough to make this determination. There are other factors that can impact the similarities or differences in the

above locations. Culture can play a role in how the neighborhoods are structured and demographic information can also cause a change. It is also good practice to scrutinize the data to make sure the source is credible. In this case, some categories counted separately even when they were describing the same venue type (i.e. beer bar and beer garden).

Works Cited

New York Dataset: *https://cocl.us/new_york_dataset*

Chicago Dataset: *https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6*

KMeans Clustering: *https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1*