Comparing Popular USA Cities

Simran Kaur Soin

1. Introduction/Business Problem

Which of the four cities in NYC most similar to: Boston, San Francisco, Chicago, or Philadelphia?

In the past, NYC has been a popular place to start companies due to its densely packed population, much of whom are relatively young and often the target audiences of these up and coming industries; furthermore, it attracts more employees looking for an opportunity to move to New York. However, in recent years, real estate prices in NYC have been increasing exponentially, which can be especially detrimental to start-ups that have a limited budget. Comparing NYC with other cities in the USA would be beneficial for start-ups looking to find a permanent location because it could become possible to get more reasonably priced space while still providing a similar experience for employees and similar target populations.

2. Data

Four different data sources will be used in order to address the various aspects of this problem.

Firstly, the Foursquare API can provide in-depth information regarding the types of venues in each major city and the number of each type. This information would be relevant depending on what kind of company was looking to find a location, because it would provide information on how much competition each city would have as well as how well each type of venue generally does in that city.

Secondly, the Governing List of city populations will be used to compare cities for obvious reasons, such as potential customer base and exposure to new audiences.

Thirdly, the Kiplinger list of median home prices in US cities will be used to gain a general idea of how expensive real estate can be in each city

Lastly, the Wikipedia list of largest school districts in the US (by enrollment) will allow for better analysis of types of populations. The school districts in particular, relate to the population of parents and children, which could be relevant, depending

on the type of company looking to find a location. For example, if a nightclub was looking to open in a certain city, it might be helpful to know if there was an exceptionally large school district — as this might be an indicator of a bad location choice. In comparison, if a toy store was looking to open in a certain city, the school district size might be a good indicator for where they should look to open.

3. Methodology

First, the population density data, school district data, median home price data, and venue location data were accessed through scraping the different websites and accessing the Foursquare API. This data was then cleaned by refining it to only include the relevant data from each table — in other words, the data that pertained to the five cities under investigation. Furthermore, the Foursquare venue data was cleaned by accessing the 30 most popular venues within each city, getting their categories, and calculating the 5 most common categories of venues within each city.

Once the data had been cleaned, it was compiled into a singular pandas DataFrame called overall_df and the median home price column was temporarily dropped (since the goal was to search for a similar city with *less expensive* real estate prices, not identical prices). This was then one-hot encoded to make comparisons easier between cities. And the encoded data frame was clustered using k-means clustering with k-values 1-4. This allowed me to see how strong the correlations were between different cities. With k=2, Boston and NYC were in a cluster of their own, while the remaining 3 cities were in a separate cluster. With k=3, Philadelphia was also clustered with Boston and NYC. This showed that Boston had a stronger correlation with NYC than Philadelphia, but both Boston and Philadelphia were more similar to NYC than any of the other locations.

Finally, the median home price column was added again. This showed that median real estate price is actually slightly higher in Boston than NYC.

4. Conclusion

Although Boston and NYC were clearly shown to have the strongest correlation in population density, school district size, and venue category popularity, Philadelphia was closely behind Boston and had a median home price that was almost *half* that of either Boston or NYC. Therefore, the overall conclusion is that Philadelphia would be the best location for start-ups looking to have a similar target audience to that of NYC.