Boston: A Tale of Two Cities

Coursera IBM Data Science Capstone

Sunny Yan

Intro:
My business problem is analyzing the rich and poor neighborhoods of Boston. Want to compare and contrast what kind of venues they have. Ie. do rich neighborhoods have more restaurants, tourist attractions, parks, etc. My hypothesis is that rich neighborhoods will have more of these things that will attract people to come to the area and bring money to the local economy. Whereas poorer neighborhoods will have things such as laundromats, convenience stores, fast food restaurants etc. Things that are more for the citizens of the area and not tourists who can come in and drive money. The audience for my study will be sociologists and also people who are interested in trying to improve the economy of their local community.

Data:
Will be using data that I scraped from wikipedia that has information on all of the zip codes that are in Boston as well as the income data for each of these distinct districts. Also using a dataset that contains the latitude and longitudinal values for each district.

In conjuction, will use a foursquare api that allows me to look at all of the locations/venues that are located in each district.

Methodology:
I started by scraping data from the Boston wikipedia page, the table contained zip code and district information, as well as Income information for each district. I cleaned the data and sorted the table by MedianFamilyIncome descending. With this method I was quickly able to identify which neighborhoods where wealthy and which were comparatively less well off.

After succesfully cleaning the dataframe and creating a completed table. I then merged my dataframe with another dataframe that contained latitude and longitude values for each district. This needed to be done to in order to use the foursquare API. But first, I created a folium map with each district labeled to get a visual sense of where each district was located. This way I could see how far away the poor and rich neighborhoods were in comparison to one another.

Then I began the foursquare API portion of my data research. I began by aggregating all of the nearby venues of the North End district, one of the most prosperous ones from Boston, to see what kind of venues were most seen there.

After that I created a function to get the nearby venues of all the districts to get a more in depth analysis with a larger sample size. Then I created another function to be able to see what the top 10 venues of each district where. I used this function to create a table that would allow me to easily analyze which venues were the most popular in each region. This allowed me to glean a lot of good insights about how the rich and poor neighborhoods differed. The results were quite different, more on that in the next section.

Finally I used machine learning by creating a k-means clustering algorithm that allowed me to separate the neighborhoods into 3 distinct clusters to better categorize the differences.

Results:
I found out that in the rich neighborhoods such as the North End district, there are many expensive restaurants, such as Italian or Seafood Restaurants. Also in these regions you are far more likely to see Parks, Bakeries and cafe's. Also the 8th most common venue in the North End are historic sites. Contrasted to a district like Roxbury, you will see that Roxbury's most popular venues are Convenience Stores, Fast Food Restaurants, Business Services as well as American Restaurants. Both the North End and Roxbury have a solid amount of parks.

Discussion:
So what do these results mean? We see that the North End has a lot more businesses that are attractive to tourists. The restaurants are exotic and expensive Italian and Seafood Restaurants, and there are plenty of bakeries and cafés in the area where tourists can come for quick eat. Whereas in Roxbury we see that the Businesses are more practical and on the cheaper end of things. For example, the convenience stores and the fast food restaurants. This makes sense because it is a lower income area and their citizens cannot afford to go out for expensive dinners. Recommendations for

Roxbury would be to make the area more tourist friendly by opening more bakeries and cafe’s. This move coupled with making their parks beautiful will drive tourist traffic and money into the Roxbury, boosting their local economy.

Conclusion:
This project used Boston zip code and income data to investigate the difference between rich and poor neighborhoods. Scraped data from Wikipedia and used a Foursquare API.

Exploratory Data Analysis revealed that the rich neighborhoods had more tourist attractions driving money into the local economy. Lower income neighborhoods can improve their economies by opening up more businesses that can drive tourism into the region.