**Introduction**

#### Boston is one of the biggest and the 21st most populous city in the US. The city is also placed among the top 30 most economically powerful cities in the world. Consider all these factors, Boston always attracts real estate investor’s attention when it comes down to real estate investment. However, Boston consists of several neighborhoods, and each neighborhood has its own characteristic. Therefore, it’s important to analyze the relationship between the neighborhood’s venue data and its housing price, and see what’s the most common venues in high housing price neighborhood.

## Data Description

#### The Data I am going to use in this project are listed as below:1. The list of the neighborhood in Boston from Wikipedia 2. Foursquare API to get the most common venues of given Neighborhood in Boston3. The average housing price of each neighborhood from redfin. Since it is hard to find the aggregated data online, I manually record the data from the redfin website and put them into a CSV file.

## Methodology

## *As a database, I used GitHub repository in my study. My master data include neighborhood, Housing price per Sq. Ft, each neighborhood’s Latitude, and Longitude.*

## 

## *I used folium to visualize geographic details of Boston and its neighborhood, and I create a map of Boston with neighborhoods superimposed on top. I used latitude and longitude values to get the visual as below.*

## 

## *I also utilized the Foursquare API to explore the boroughs and segment them. I designed the limits as 100 venue and the radius 500 meter for each neighborhood from their given latitude and longitude information. Below is a head of the list Venues name, category, latitude and longitude information from Foursquare API.*

## 

## *In summary of this data, 844 venues were returned by Foursquare.*

## *In order to analyze the venue data, I used unsupervised learning K-means algorithm to cluster the neighborhood. First, I use get\_dummies function to turn the categorical value into numeric value. Then I import standardScaler from Sklearn to standardlize the data. This process make sure the housing price value won’t have much more influence on clustering than any other numeric value. Then I set the number of cluster to 3 and label the each neighborhood. The result shows below.*

## 

## Result

## *By calculating the average housing price of each cluster, the result shows that three cluster’s housing price has significant differences.*

## *Therefore, I named the cluster 0 as high price area, cluster 2 as middle price area, and cluster 3 as low price area.*

## *In high price area venue data set, we can see that the most common venue for each neighborhood usually Korean restaurant, Italian restaurant, Chinese restaurant, and hotel.*

## 

## *In middle price area venue, the most common venue are usually pizza place and coffee shop, bakery and pharmacy.*

## 

## *When we look at the low price area, the most common venue are usually metro station, sandwich place, park, and home service.*

## 

## Discussion

## *By looking at the result, we can see that the neighborhood’s housing price are certainly tied to the category of venue. The neighborhood with more Italian restaurant, Korean restaurant, and Chinese restaurant have higher housing price compare with other neighborhood, where doesn’t have many these restaurant.*

## *The housing price is affected by many other factors other than category of venue. However, by analyzing the category of venue in certain neighborhood, we can definitely get some insights regarding to its housing price.*

## Conclusion

## *When a real estate investor want to invest in certain neighborhood, he/she can start analyzing the neighborhood ‘s venue category to gain some insights regard this neighbor. However, the housing price analysis can be completed by adding more other factors such as education recourse, population density to get more accurate result.*