Zachary Hyman

U13384191

CS506

HW01: AirBnB Data Analytics Report

1. Clustering

The data presented from the AirBnb data set contained many fields regarding all attributes of each listing in New York. The parameters latitude, longitude, and average price listing were the parameters used in calculating the data clusters listed below. In order to properly plot and cluster the data, the value of each listing had to be scaled to a similar scaling as the scope of the latitude and longitude of the listings. Each listing in New York had an approximate coordinate location between 40 to 41 longitude and -73 to -74 latitude, so the scale for location was roughly 1 unit across the two location axis of my model. Thus when scaling the values of each property I found that using mean normalization of the data yielded a spectrum of values which also had a range of approximately 1. This allowed for each axis of the clustering plot to have roughly the same axis and thus allow both the location and value to be reflected equally in the distance of each listing to another in 3d space.

a) Kmeans++

Found below are two models displaying the kmeans++ clustering algorithm using the transformed and normalized AirBnb dataset (Fig. b, c). The plots below depict the algorithm running with a k equal to 4, or four distinct clusters being formed from the data. The parameter for number of clusters was determined by viewing the loss curve for the Kmeans algorithm. The parameter of k=4 is at a location in the loss curve where intra and inter cluster distance is

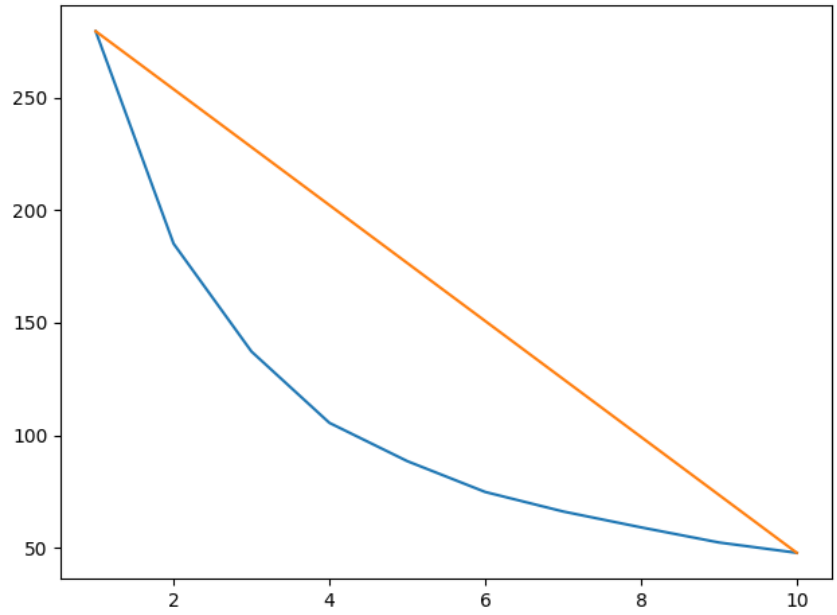


Fig a.) Kmeans++ Loss Curve

minimized, yet is a number of clusters small enough where as an increase would see a diminishing return. If a line is drawn perpendicular to the orange line indicated on the plot (Fig. a), it would indicate approximately 4 clusters to be ideal for this data set. The clustering algorithm preformed also used all 48,000 entries in the .csv file provided.

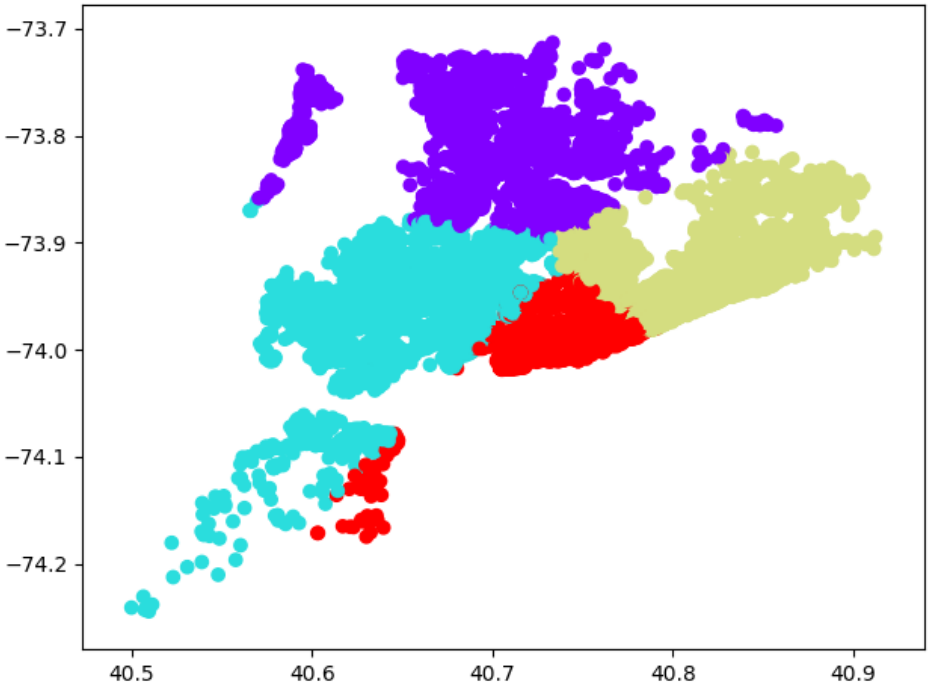


Fig b.) Kmeans++ Algorithm 2D Plot

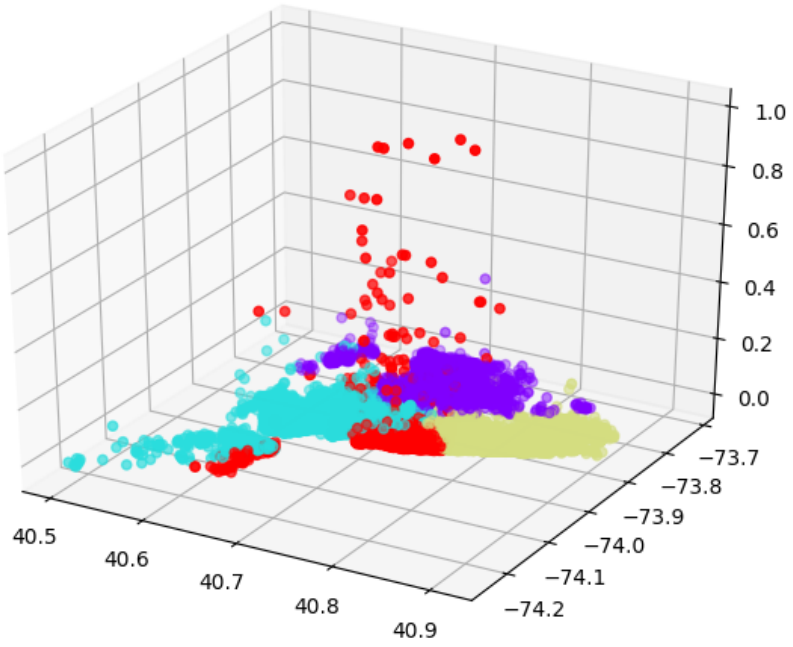


Fig c.) Kmeans++ Algorithm 3D Plot

b.) Hierarchical Clustering

Found below (Fig e, f) are two plots depicting the results of clustering using hierarchical clustering on the provided AirBnb dataset. For this set of clustering a sample size of 8,000 points were used as with the nature of hierarchical clustering, preforming on a data set of almost 50,000 data points would be a task unable for my equipment to complete in reasonable time. To determine the amount of clusters used, I utilized a method from the clustering library known as silhouette scoring to determine the most effective number of clustered to be used. As shown by the graph below, a clustering number of 5 provided a high silhouette score, meaning intra cluster distance was minimized, yet provided a reasonable amount of clusters to observe patterns and shapes on the plots of data. Below is the plot of those scores with many different numbers of

clusters as well as plots of the data utilizing a clustering number of 5, in both 2D and 3D plots (Fig d).

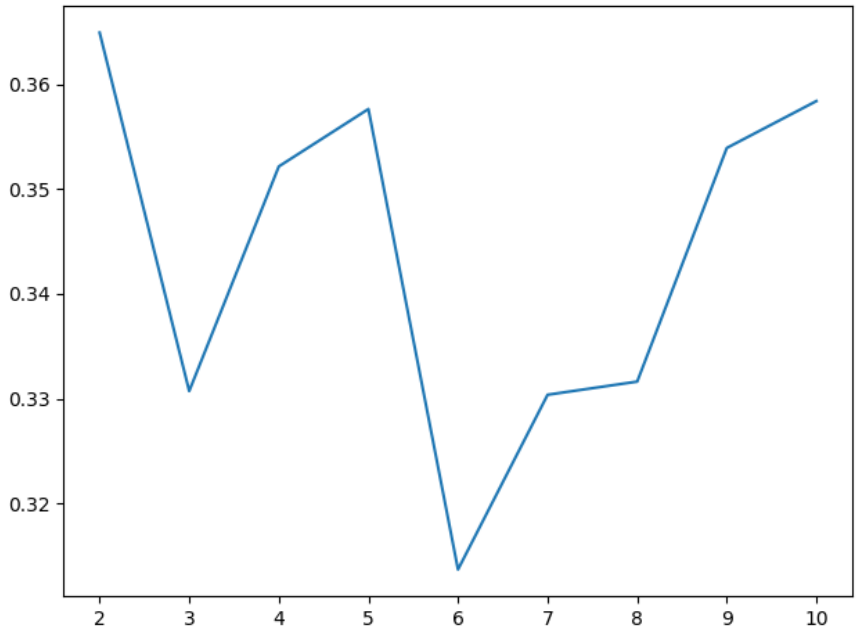


Fig d.) Silhouette Scores

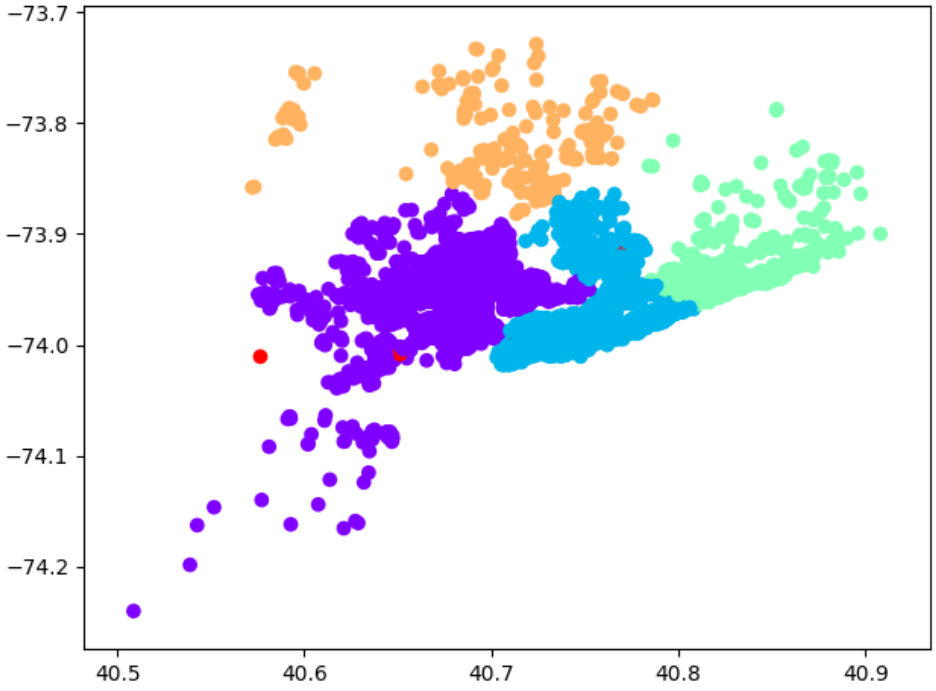


Fig e.) Hierarchical Clustering 2D Plot

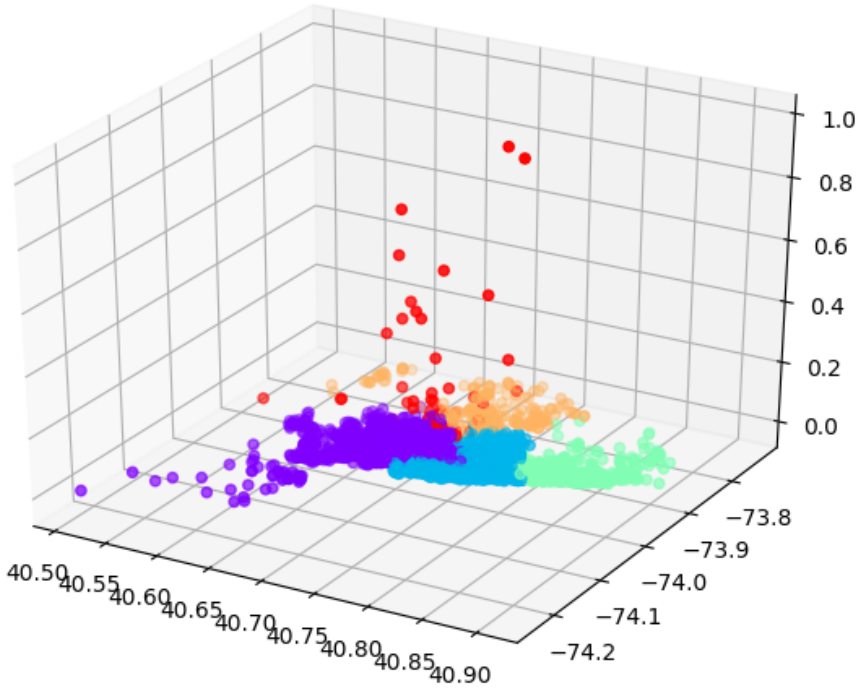
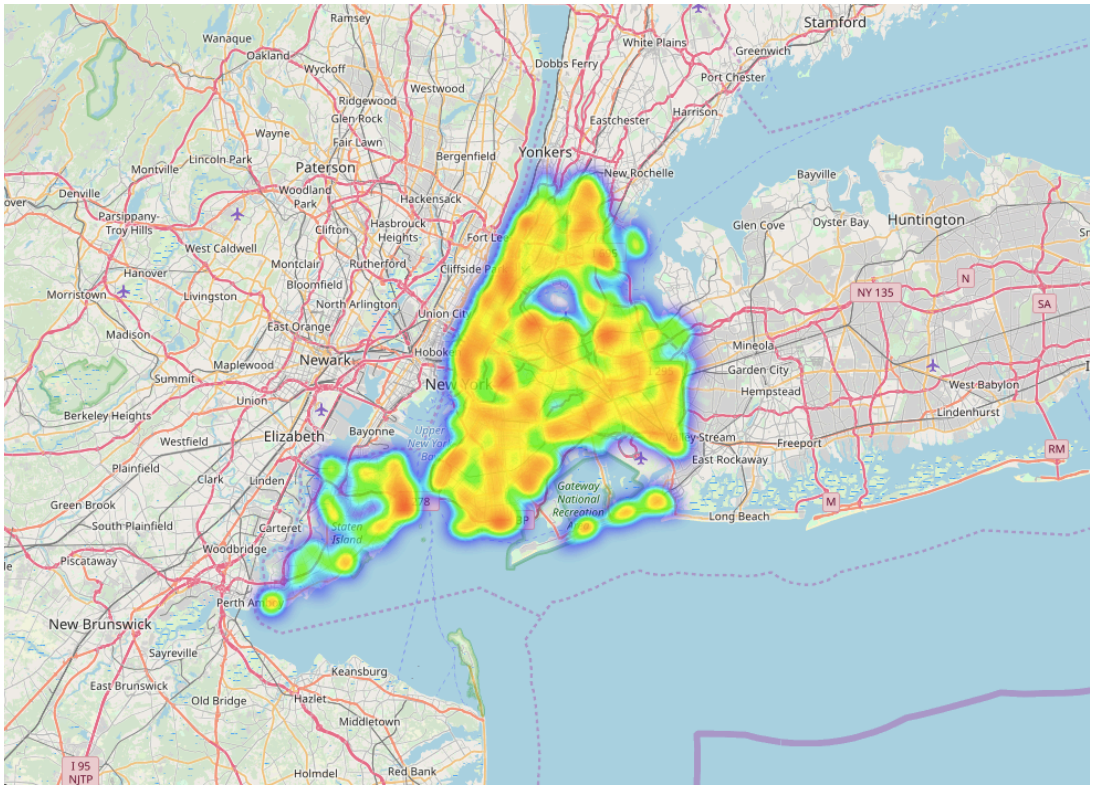
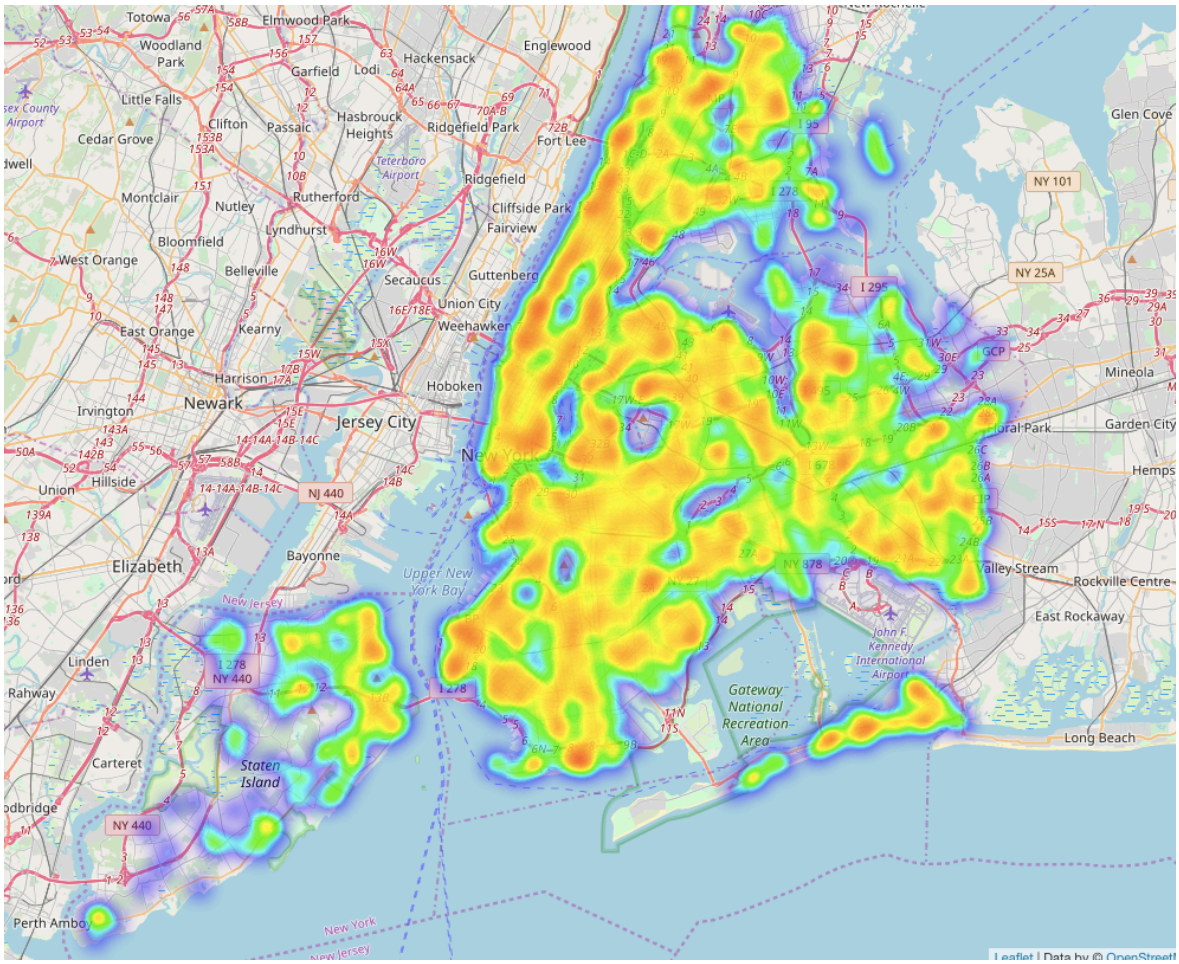


Fig f.) Hierarchical Clustering 3D Plot

2. Data Visualization

a.)





I found the Heat maps produced by the algorithm provided to be rather uninformative in drawing any conclusions about the expressiveness of any particular area in New York. This heat map simply shows what areas provide a greater amount of AirBnb location, but in a city such as New York where they are extremely common, all data is lost into the large heat map blob. I found the clustering produced to be a much more informative representation of the data which provided insight into price disparities in different burrows of New York. The lack of distinct features in the heat map model provides a great lacking to draw conclusions on the data.

b.) See above Figures A-F.

c.)

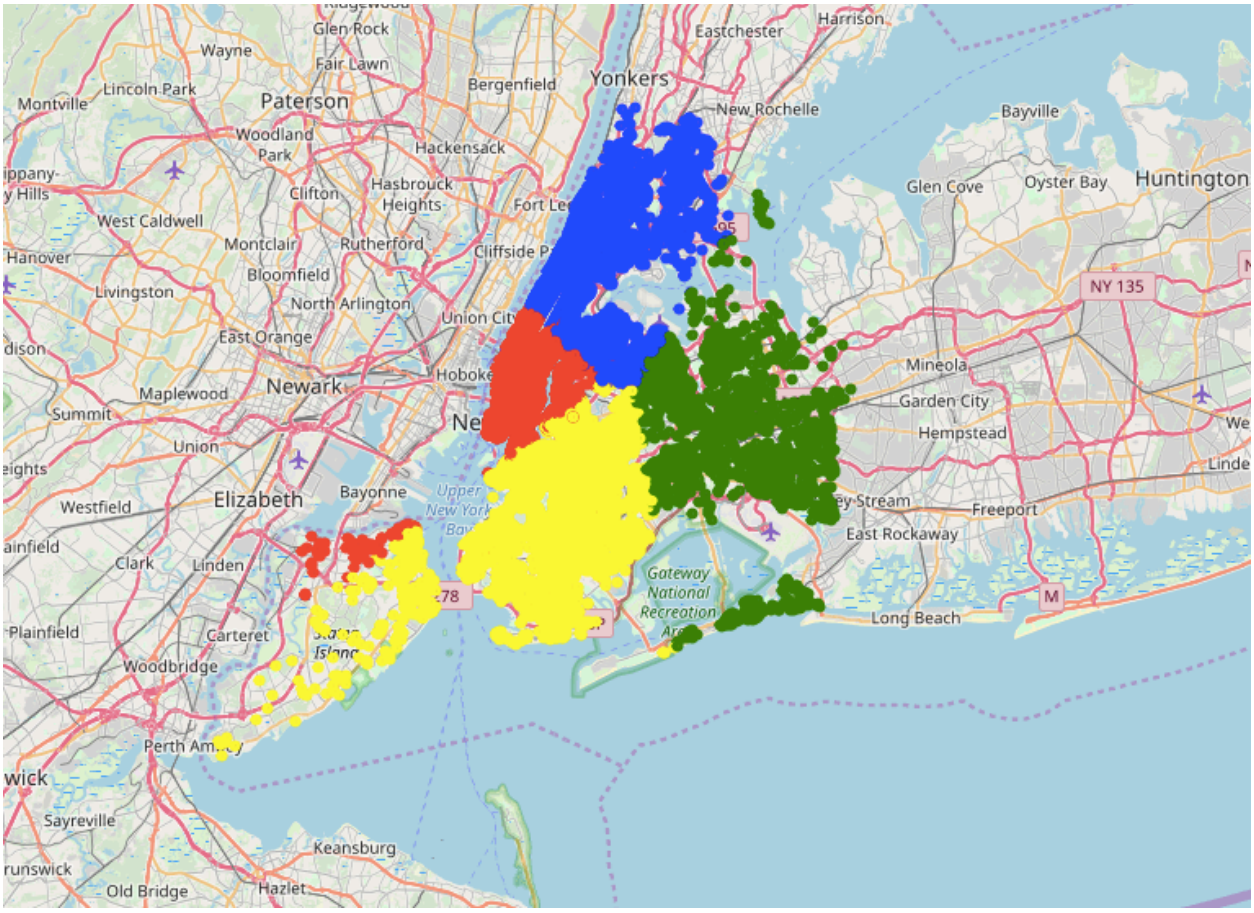
Found below are the average price in dollars listings determined from the Kmeans++ algorithm used in the first part of this report with a setting of four clusters used. Below the prices listed are the centers of each of the clusters displayed in numerical order with the format:

[Latitude, Longitude , Normalized Price].

```
Average Price in Cluster 0: 233.50825412706243
Average Price in Cluster 1: 116.74466403162023
Average Price in Cluster 2: 97.96953010279003
Average Price in Cluster 3: 110.34146991419234
[[ 4.07371927e+01 -7.39849478e+01  8.19660053e-03]
 [ 4.08011657e+01 -7.39367104e+01 -3.46940162e-03]
 [ 4.07083901e+01 -7.38217532e+01 -5.34931175e-03]
 [ 4.06803243e+01 -7.39517233e+01 -4.11238253e-03]]
```

d.)

HTML file can be found in submitted directory.



e.)

I would say that the clusterings provided agree with my somewhat limited understanding of how pricing in the burrows of New York are. I was not surprised to see many of the outlying high prices were found to be located on Long Island, as I have knowledge that there are many mansions and vacation destinations on Long Island which would support a much higher average price than any other burrow. There were also clear divides that could be indicated between areas such as Manhattan and Brooklyn which I would expect as many different areas of New York demonstrate different standards for living as well as architecture and environment. The same goes for the division in this example of uptown and downtown Manhattan. These differences would be reflected in the listed pricing of establishments and the clusters do support this hypothesis.