
**Assignment 2**

Jiaxing Lin

Metropolitan College, Boston University

AD571 A1: Business Analytics Foundations

Prof. Hyunuk Kim

Assignment Due Date: 06/12/2023

In the beginning stage of analysis, I removed data that is not useful including the sale prices and square feet of buildings equal to zero in order to provide more statistically accurate results. Through the process of pulling out the sales year and connecting the corresponding building code IDs, I narrowed down the data to include only residential building types to find specific information.

Part 1

In this part, I focused mainly on my assigned area - HARLEM-CENTRAL neighborhood. After explicitly pinpointing the neighborhood in the data set, the sales data was segmented into separate clusters for each year spanning from 2000 to 2021. By generating a list of calculated average sale prices per square foot for each year by dividing the sale prices by the usable interior space measured in square feet and finding the mean, I can conclude that the average sales price reached a peak in 2019 with approximately $815 per square foot.

Embarking on an exploration of our dataset, I compiled the five-number summary for both the sale price and gross square footage of residential buildings in HARLEM-CENTRAL starting from the year 2009. This essential statistical technique has provided valuable insights into the distribution of the selected group. From 2009 to 2021, the average sales value for residential properties stood at $3,759,871, while the highest recorded sale price reached a staggering $200,850,000. And the range of residential building sales in terms of gross square feet is substantial, from a minimum of 335 sqft to a maximum of 7,156,400 sqft.

Through analysis, I examined the relationship between the sale price and gross square feet of residential properties since 2009, revealing a correlation coefficient of 0.62. This indicates a positive correlation, suggesting that as the gross square feet of a property increase, the sale price tends to increase as well.

Part 2

In order to examine the patterns and gather insights into the characteristics of all neighborhoods, specifically in residential buildings in NYC since 2009, I computed the median sale prices, the total sum of sales units, and the standard deviation of sales units. I then combined these three key performance indicators (KPIs) into a single dataset and performed k-means clustering to identify common groups or clusters within the neighborhoods. At the beginning of the analysis, I excluded neighborhood IDs as they serve as unique identifiers and changed the row name to neighborhood ID in order to better pinpoint the ID. Subsequently, I standardized all three KPIs to ensure they are on the same scale for further analysis. After removing the missing value inside the dataset, I can then find the optimal number of clusters by calculating the within-cluster sum of squares for a maximum of 25 clusters and the visual representation shown below.

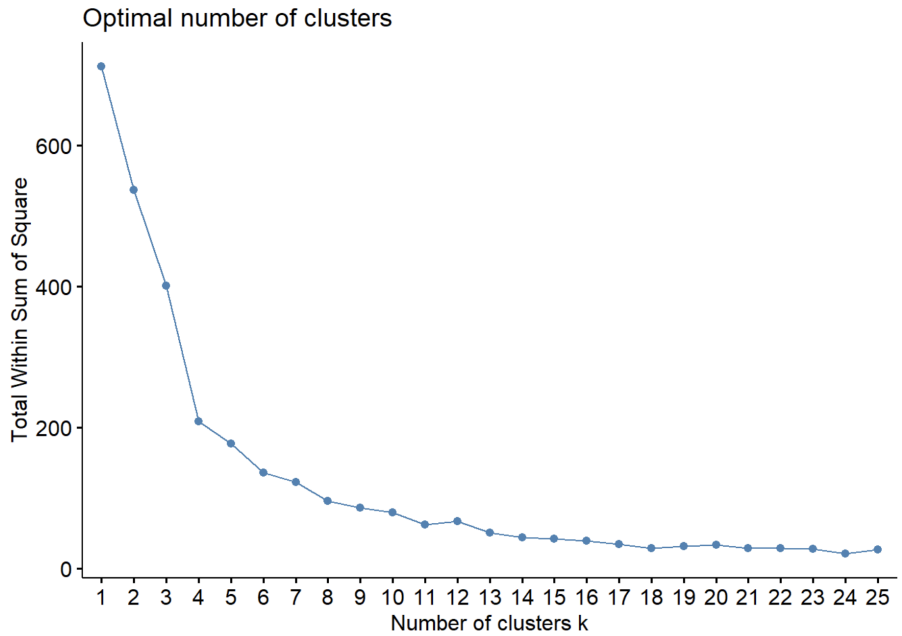


Figure 1: The graph of optimal number of clusters

According to *Figure 1*, I believe 6 is the ideal cluster based on the Elbow method. Beyond this point, introducing additional clusters may not markedly enhance the effectiveness of the clustering. By graphing the k-mean cluster, the cluster plot indicates as

below in *Figure 2* where the assigned neighborhood HARLEM-CENTRAL in ID 118 is in cluster 1. There are 13 neighborhoods in the same cluster that share similar characteristics, indicating that they share similarities in terms of median sale prices, total volumes of sales units, and the variability of their sales units. Additionally, this step may also highlight any outliers, such as in this instance, neighborhood 142, which deviates significantly from the common characteristics observed among the other neighborhoods.

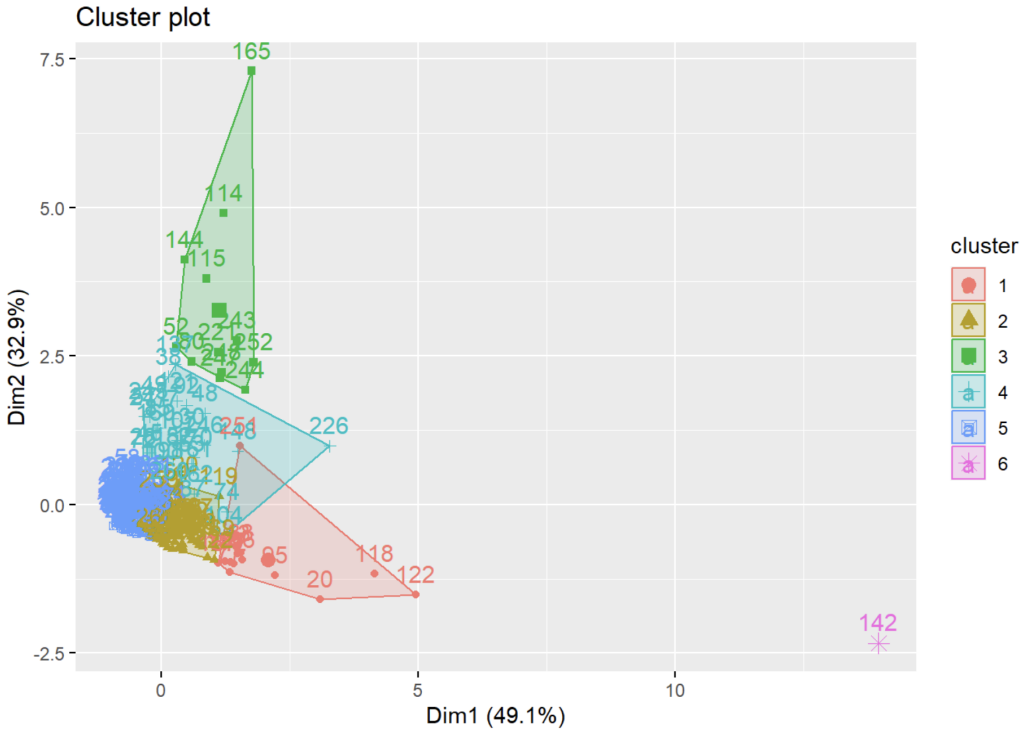


Figure 2: The graph of cluster plot

## Part 3

To compare the average residential building costs since 2009 of the assigned neighborhood HARLEM-CENTRAL with another neighborhood, I randomly selected a neighborhood in ID 238 and filtered the sale price with my assigned neighborhood ID 118. By running an unpaired t-test with the null hypothesis assuming no difference in means and

the alternative hypothesis suggesting the mean difference is higher than zero, we can reject the null hypothesis at a 95% confidence level when the p-value is less than 0.05. This indicates statistical significance, providing evidence to support the existence of a meaningful difference between the two variables. Overall, when comparing the average sale prices between the selected neighborhood and HARLEM-CENTRAL, it is evident that HARLEM-CENTRAL exhibits significantly higher average property costs in the residential building category.

After all the analysis above, I do believe my neighborhood presents an appealing opportunity to open a real estate office due to the following reasons. Firstly, there is an overall increase in sale price per square foot from 2003 to 2021 even with some fluctuations. Secondly, the analysis of residential property sales across all neighborhoods since 2009 reveals that HARLEM-CENTRAL stands as the second highest-selling neighborhood in New York City. And other KPIs we analyzed also indicate my neighborhood is among the top choices. Final but not least, based on the t-test conducted, the average sale price of my neighborhood exceeds nearly tenfold when compared to a randomly selected neighborhood in New York City. Thus, I believe HARLEM-CENTRAL is an attractive neighborhood to open a real estate office.