11/29/24

ISL FA24 Project – Task #3 Report

1) Data description: how many samples? How many features? What type of features?

- There are 99492 samples and 11 features. The features have 6 quantitative and 5 qualitative (including the ID variable)

2) Data preprocessing: are there any null values or outliers? How did you deal with them? How did you handle scaling?

- There were 60,982 null values from various features and 84 duplicate rows. Removing them would remove a large portion of the data, so they had to be replaced with mean, mode, or calculated values.

1. I began by removing the **id** feature since it was not an important feature to have. Then checked for uniqueness among all qualitative and some quantitative features for cleaning purposes.

```
fee
['No' nan 'Yes']

pets_allowed
['Cats' 'Cats,Dogs' nan 'Dogs' 'Cats,Dogs,None']

price_type
['Monthly' nan 'Weekly' 'Monthly|Weekly']

state
['CA' 'VA' 'NC' 'NM' 'CO' 'WV' 'GA' 'MA' 'DC' 'AZ' 'IA' 'WA' 'TX' 'IL'
 'MS' 'OR' 'FL' 'MO' 'PA' 'WI' 'OK' 'UT' 'RI' 'NJ' 'IN' 'MD' 'OH' 'TN'
 'ND' 'NE' 'AR' 'MI' 'MN' 'HI' 'ID' 'SC' nan 'KS' 'AL' 'SD' 'NY' 'KY' 'LA'
 'AK' 'CT' 'NV' 'WY' 'VT' 'NH' 'MT' 'DE' 'ME']

bathrooms
[  1.    1.5   2.    2.5   3.    3.5   4.    nan   7.    4.5 111.    5.
   8.    8.5   6.    5.5   9.    7.5]

bedrooms
[ 1.  3.  2.  4.  0.  5. nan  7.  8. 22.  6.  9.]
```

2. I deleted all rows where both **longitude** and **latitude** had null values. Then found the indices where the **state** feature was null. To fill these with a state, I used the **longitude** and **latitude** to find the correct state.

3. For the **pets_allowed** feature:
   - if the category was 'Cats,Dogs' or 'Cats,Dogs,None' then it's changed to 'Both'
   - If 'nan' then it was changed to 'No'.
   - If **pets_allowed** has a class of 'No' the **fee** is 'No' - otherwise the **fee** is 'Yes'

4. Dropped all rows where **price_type** is not classified as 'monthly' since that is normal.

5. The nulls in **bedrooms** & **bathrooms** are replaced with the mode, since I want integer values than floats

6. The nulls in **price** were replaced with the average of all observations.

7. **Bedrooms** with '0' were switched to '1' room. All observations had small square footage, which makes it acceptable to do.

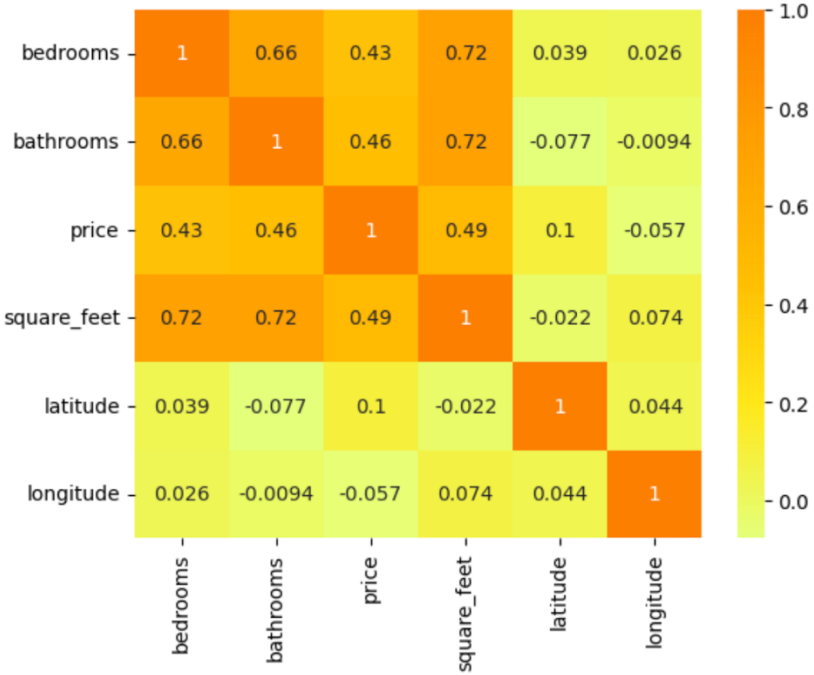
8. To manage high values/outliers in **prices**, all rows where price/bedrooms > $1700 was removed. This removed 10,227 rows. I choose $1700 since it is currently a normal apartment price.

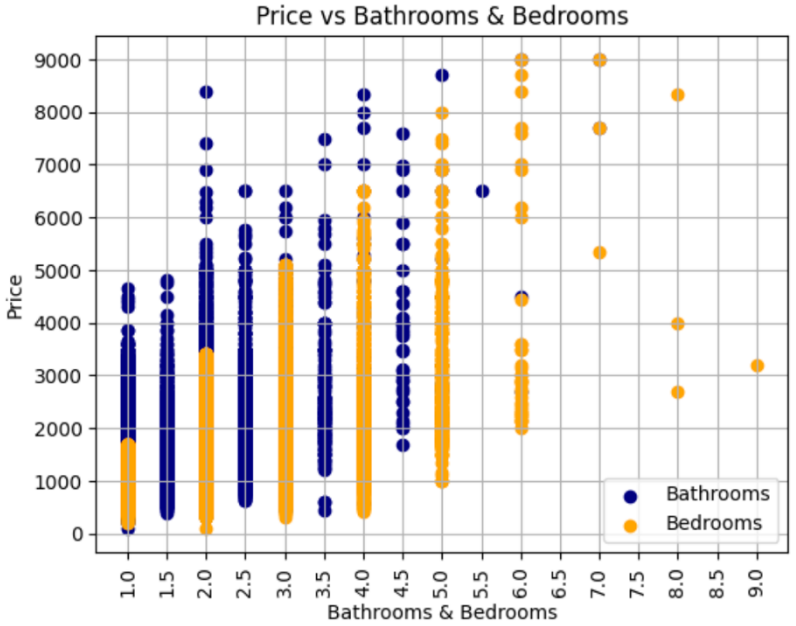
9. Then remove the remaining outliers in square feet, bathrooms, bedrooms.
   - Square Feet > 7000 sqft
   - Bathrooms > 10
   - Bedrooms > 10

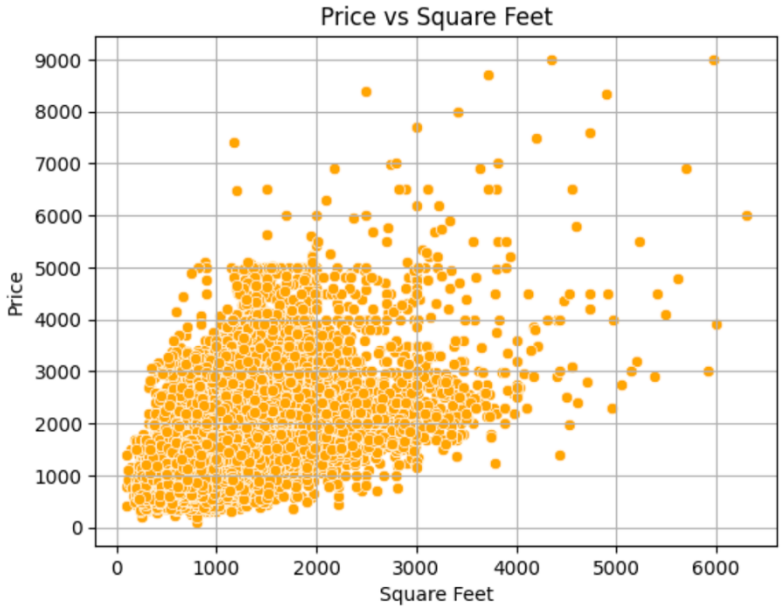
10. After cleaning there was an accumulated 11,567 duplicate rows that was removed and no null values. The uniqueness of the features are much better and majority of the data set was kept. The new dimensions are 77658 x 10

```
fee
['Yes' 'No']

pets_allowed
['Both' 'No' 'Dogs' 'Cats']

price_type
['Monthly']

state
['VA' 'NC' 'CA' 'NM' 'CO' 'WV' 'GA' 'MA' 'DC' 'IA' 'WA' 'TX' 'IL' 'MS'
 'OR' 'FL' 'MO' 'PA' 'WI' 'OK' 'UT' 'RI' 'NJ' 'IN' 'MD' 'OH' 'ND' 'NE'
 'AR' 'MI' 'AZ' 'MN' 'ID' 'SC' 'KS' 'TN' 'AL' 'SD' 'NY' 'KY' 'LA' 'AK'
 'CT' 'NV' 'HI' 'WY' 'VT' 'NH' 'MT' 'DE' 'ME']

bathrooms
[1.5 2.  1.  2.5 3.  4.  3.5 7.  4.5 5.  5.5 6. ]

bedrooms
[3. 2. 1. 5. 4. 7. 8. 6. 9.]
```

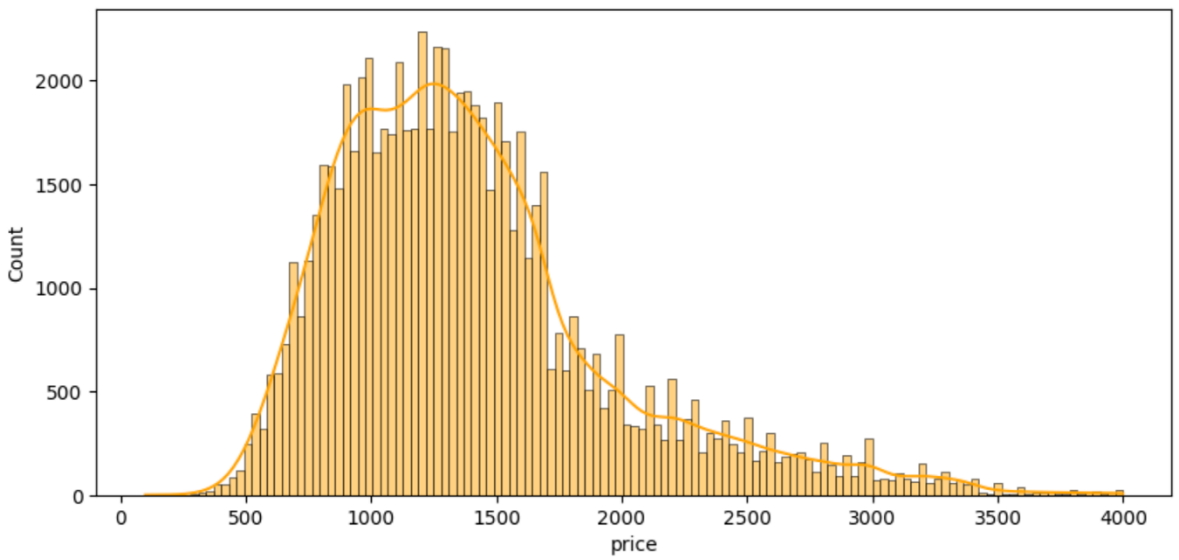
3) Exploratory data analysis: visualize the data with graphs and describe your findings. Did you find any patterns?

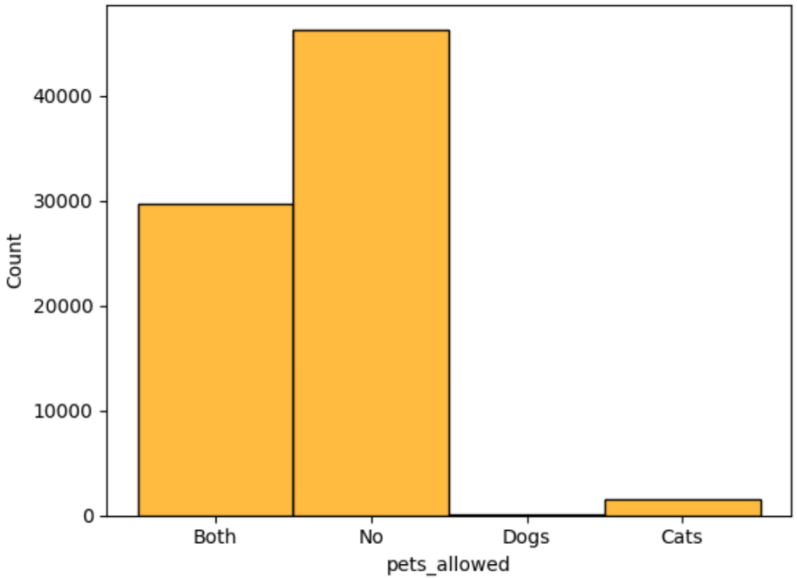


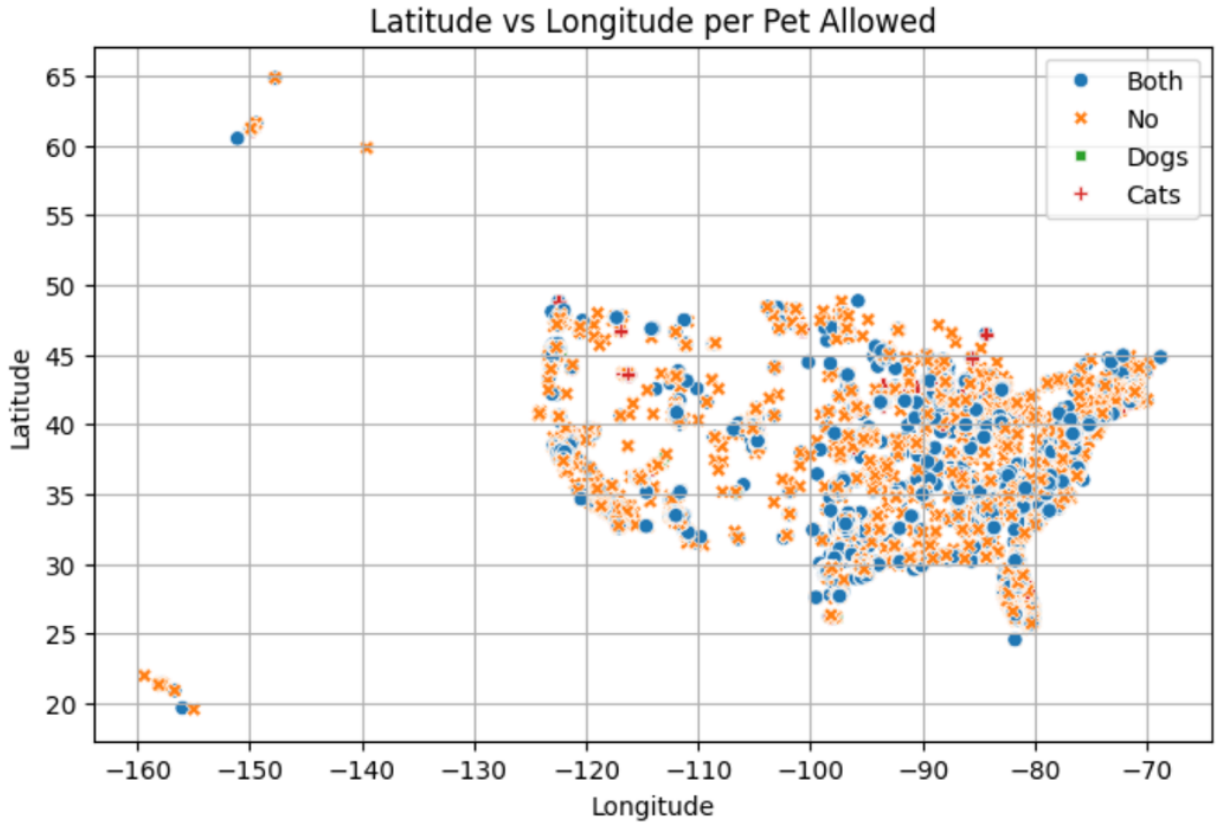




Looking at the correlation matrix the number of bedrooms and bathrooms are positively correlated, and the scatter plot above shows that relationship. As for the influence on price, the number of bedrooms and bathrooms appears to increase price at similar rates. Square footage also has a positive correlation with price. These four features could be a cluster or influence one based on their relationships with price.







The distribution of price has an average of $1200 per month but there are some rent prices that appear to be outliers. If we look at the scatter plots above, the number of bedrooms and square footage can justify the monthly cost.

A little less that 50,000 observations show apartments that do not allow pets. However, 30,000 observations allow for both cats and dogs. The scatter plot to the left reveals where pets are allowed.

On the west coast, most apartments don't allow pets. However, a majority of the eastern half has an even split on allowing or disallowing pets. There could be some clustering for different regions where no pets are allowed and the regions where only cats are allowed.

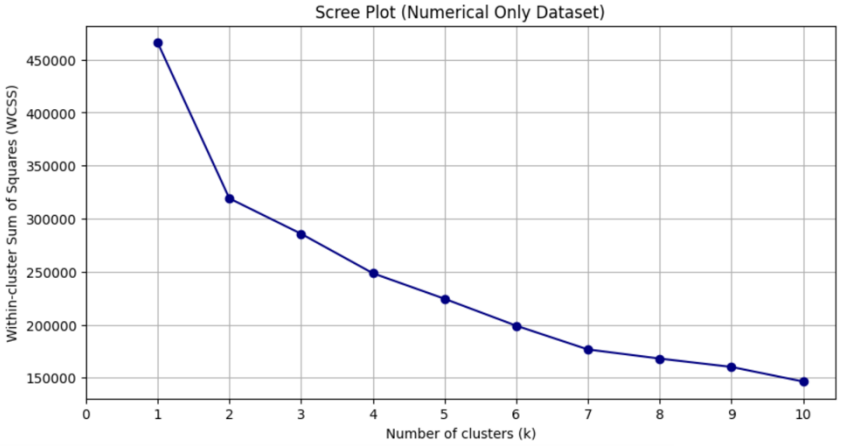
4) Model development: state the hyperparameters selected for the models and how/why you selected those hyperparameters

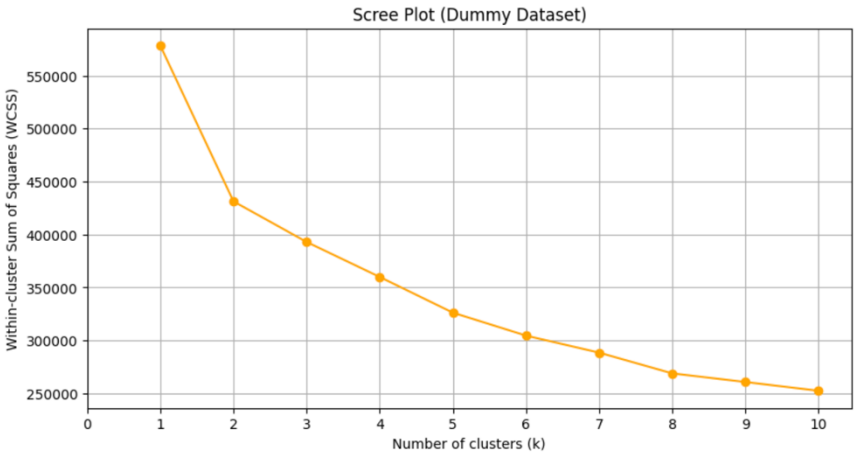
- We tested whether having only scaled quantitative predictors or having scaled quantitative predictors with dummy variables was better for clustering. Thus, we have two different datasets.

- After testing the two datasets with K-means & Hierarchical Clustering Methods, having only quantitative predictors resulted in lower within cluster sum or squares and higher silhouette scores for k-means. The silhouette scores for hierarchical decreased slower in an only quantitative predictor dataset

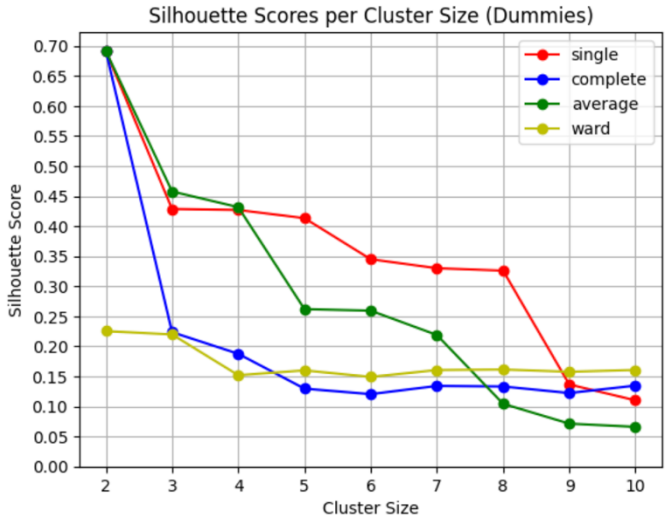
- To find labels in hierarchical, I used AgglomerativeClustering with cluster sizes from 2-10 for four different linkage methods: average, ward, complete, single. Then found the silhouette scores with those predictions & the test data. The dendrograms where truncated to show only 5 levels of clustering within each distinct cluster for easier viewing.

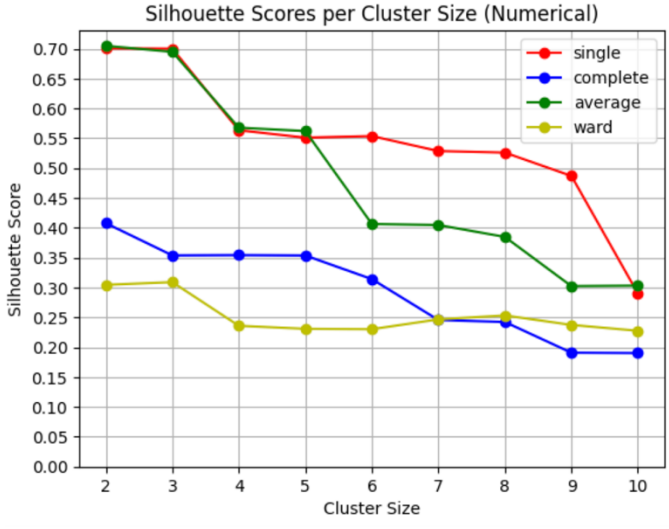
```
Cluster Size 2
Train Silhouette Score: 0.309
Test Silhouette Score: 0.307

Cluster Size 3
Train Silhouette Score: 0.215
Test Silhouette Score: 0.295

Cluster Size 4
Train Silhouette Score: 0.229
Test Silhouette Score: 0.228

Cluster Size 5
Train Silhouette Score: 0.246
Test Silhouette Score: 0.246

Cluster Size 6
Train Silhouette Score: 0.237
Test Silhouette Score: 0.239
```



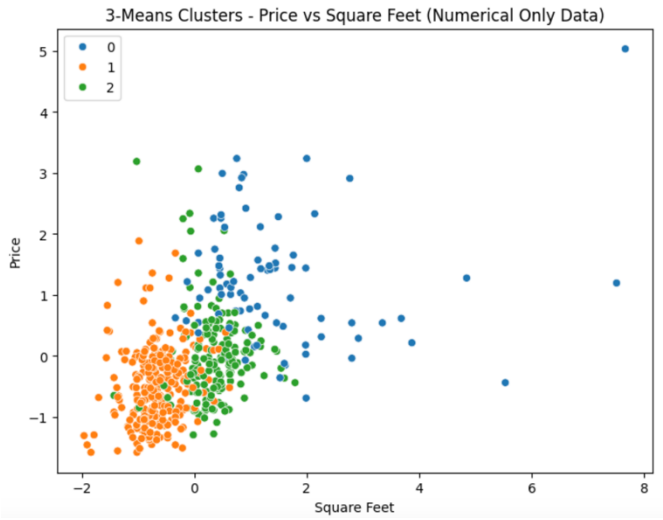


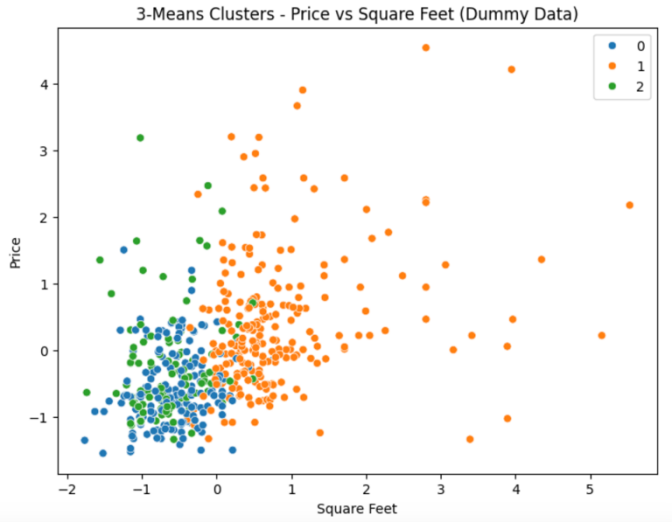

```
Cluster Size 2
Train Silhouette Score: 0.238
Test Silhouette Score: 0.237

Cluster Size 3
Train Silhouette Score: 0.196
Test Silhouette Score: 0.225

Cluster Size 4
Train Silhouette Score: 0.160
Test Silhouette Score: 0.186

Cluster Size 5
Train Silhouette Score: 0.179
Test Silhouette Score: 0.176

Cluster Size 6
Train Silhouette Score: 0.173
Test Silhouette Score: 0.170
```



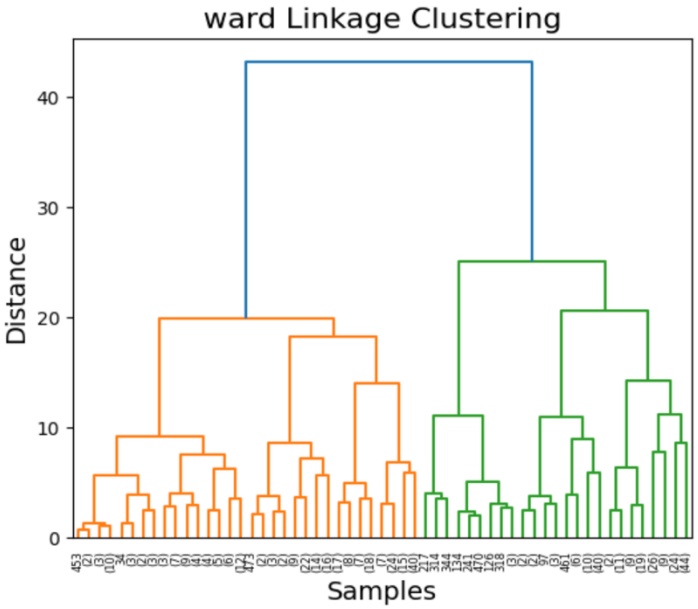


For k-means, I compared hyperparameter n_clusters = 3 for both datasets, and you can see how the numerical only dataset clusters better than a dummy variables dataset. I chose three clusters since both datasets could be better compared and their silhouette scores would be similar.



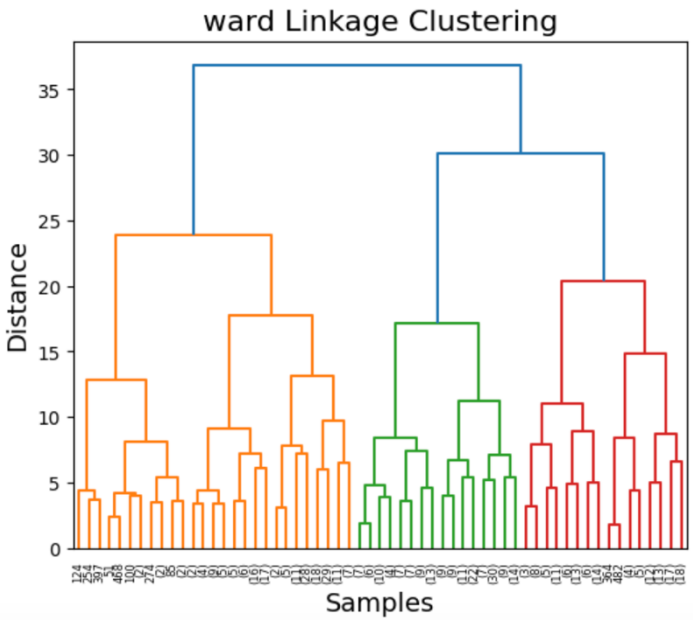


For hierarchical clustering, the ward linkage method produced clusters without having a single observation as its own cluster in both datasets. The line graphs of silhouette scores above indicate that a numerical only dataset performs better in all four linkage methods. However, in both line graphs, the silhouette scores decrease quickly as the cluster size increases.



The dendrogram on the left uses numerical variables only. The ward linkage model prefers two distinct clusters but can easily distinguish as high as 6 to 7 clusters as well. This conclusion is supported with the silhouette score around 0.25 for 7 clusters.

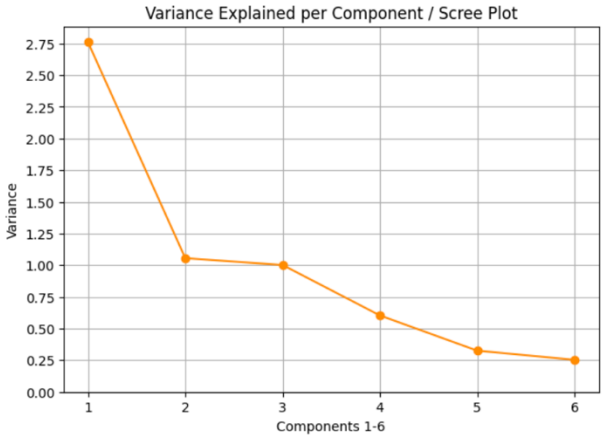
The dendrogram on the right is from the dummy variables dataset and utilizes the ward linkage method. This model prefers 3 clusters, and its silhouette graph shows that clusters greater than 3 will have a silhouette score near 0.16. This could indicate overlapping clusters as seen in k-means.

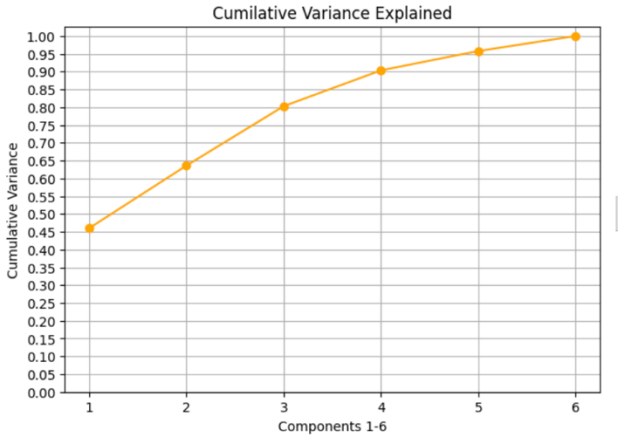


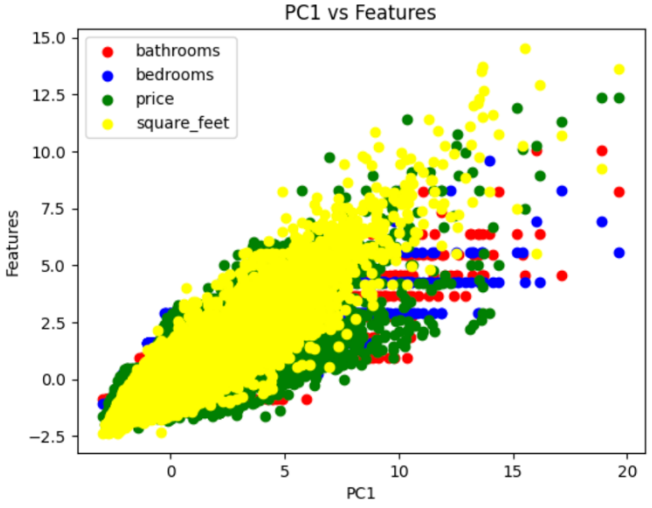
PCA

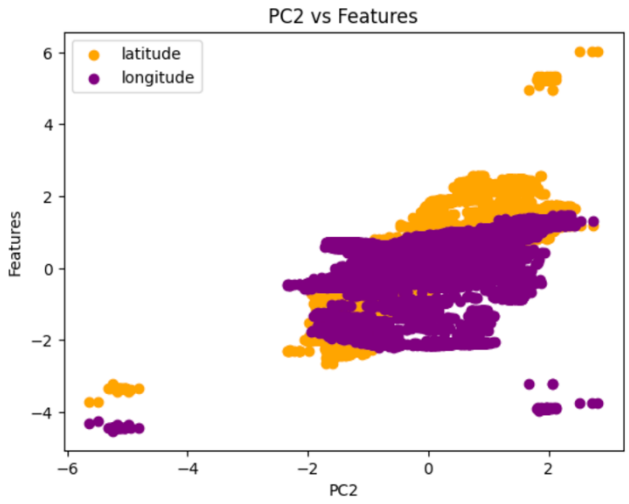
After testing which dataset performs best, I conducted **PCA** on the numerical only data. Another, way to show that the numerical only data is better for clustering is to look at the heatmaps for both datasets. The heatmap of the dummy variables shows a large feature space having no collinearity while, the numerical only data has a smaller feature space where all features have some collinearity. (View in the code)

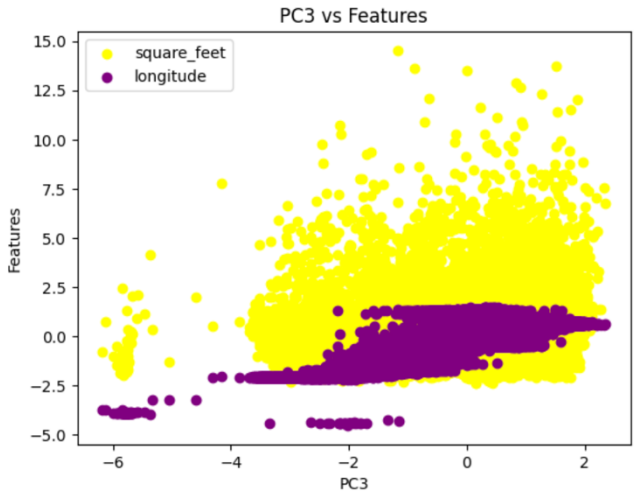
The hyperparameter is 6 components for the six features in the numerical only dataset. We see that 4 components explain 90% of the data. In the scatter plots below, we find that PC1 has a positive correlation with 4 variables. Additionally, PC2 is correlated with latitude and longitude.

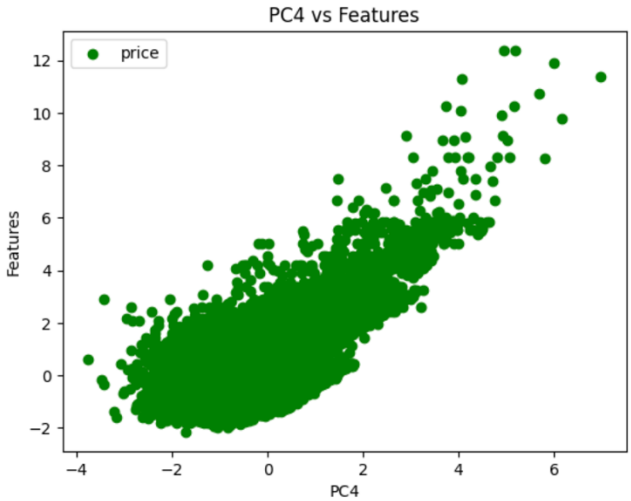






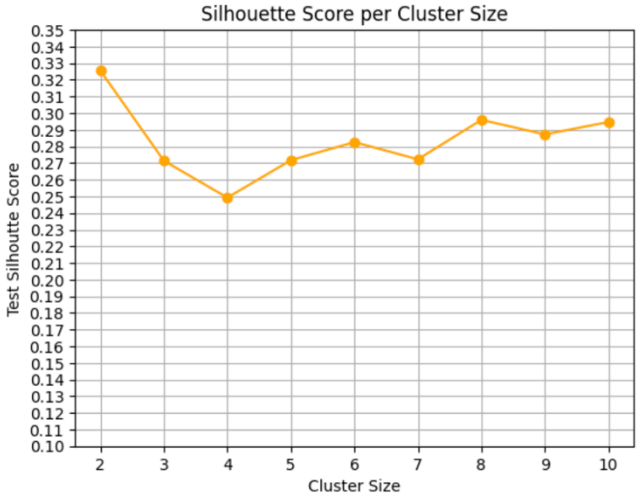


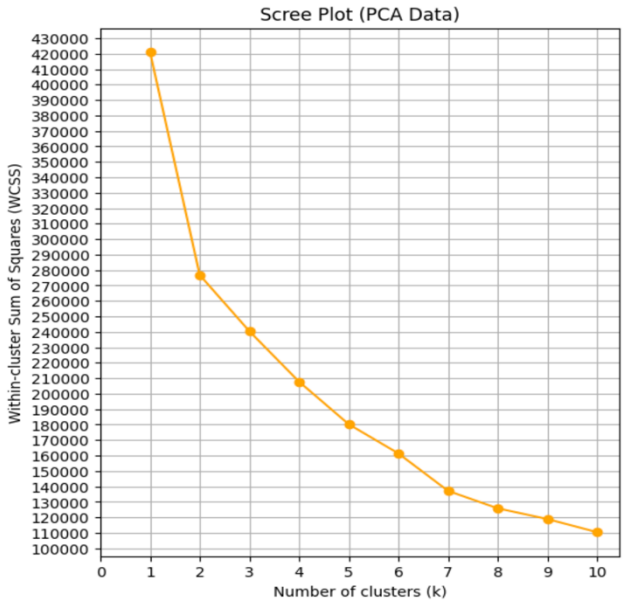




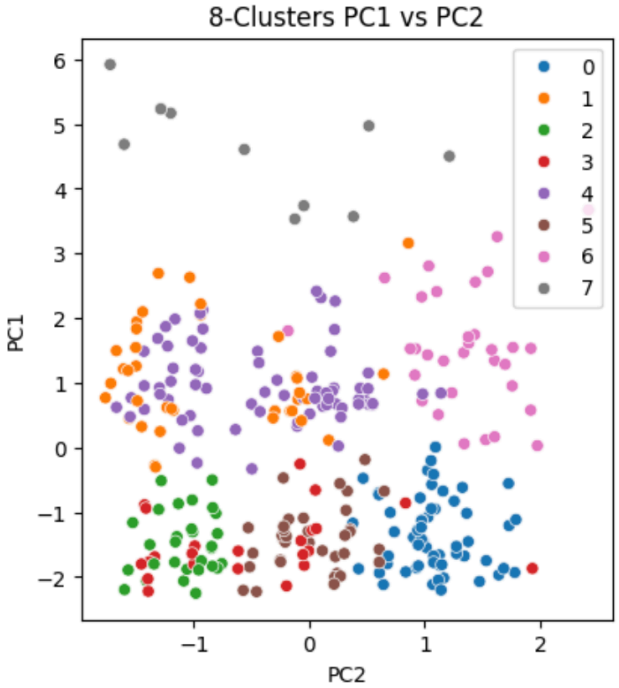
5) Performance evaluation: state the model results - accuracy, loss, precision, recall, f1-score, confusion matrix, etc.
After PCA, K-means and Hierarchical was conducted again. The results of the models are shown below.

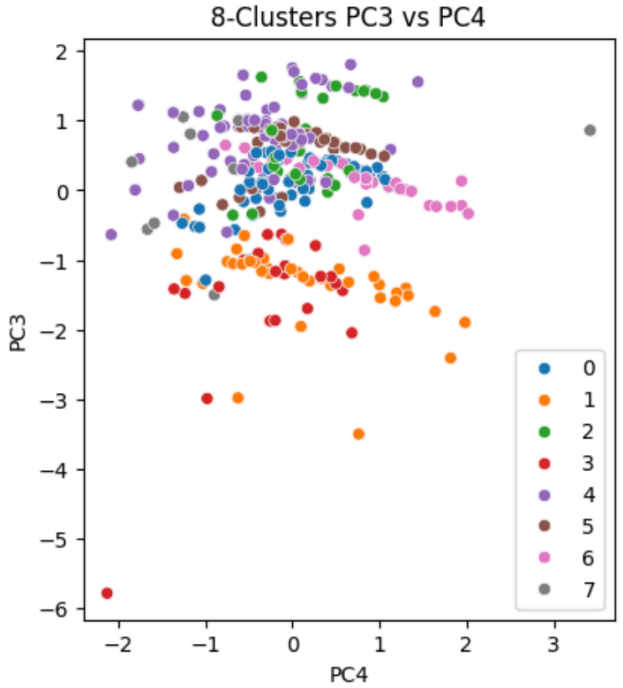
K-Means.





The cluster size chosen was 8 because of its silhouette score and how much within cluster variance is reduced according to the scree plot. The WCSS has lowered, and silhouette scores has increased after PCA.





<u>On the left scatter plot</u>
PC1 is split into 3 regions:

- PC1<0 (Grn, Rd, Br, Bl)
- 0<PC1<3 (Or, Pp, Pk)
- 3<PC1 (Gry)

PC2 is split into 3 regions:

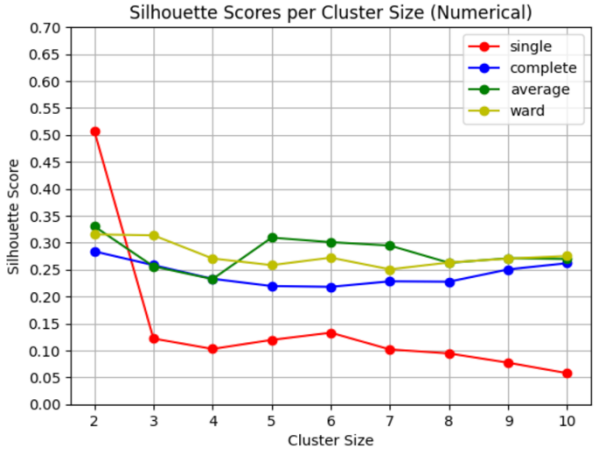
- PC2 < -0.5 (Grn, Rd, Pp, Or)
- -0.5 < PC2 < 0.5 (Br, Rd, Pp, Or)
- 0.5 < PC2 (Pk, Bl)

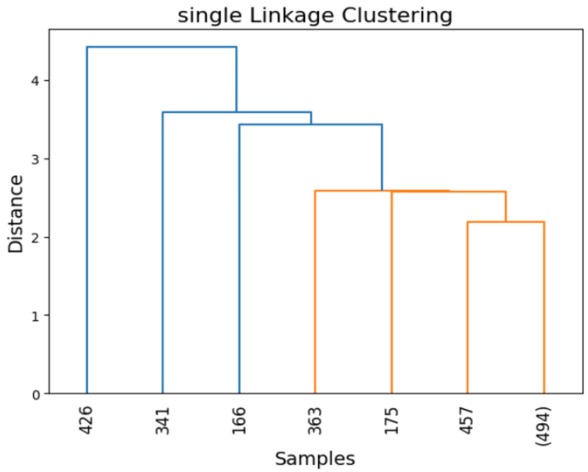
<u>On the right scatter plot</u>
PC3 is split in half:

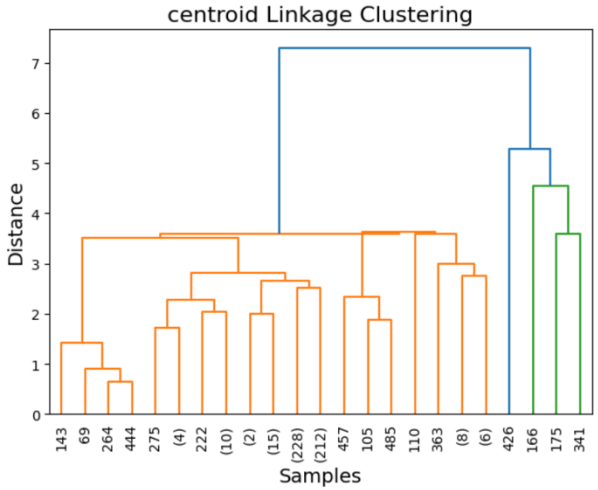
- PC3 < -1 (Rd, Or – below avg sqft)
- -1 < PC3 < 0.5 (Pk, Bl, Grn – avg sqft)
- 0.5 < PC3 (Br, Pp, Grn, Gry – above avg sqft)

Pk, Pp, and Or are a little above average in price, sqft, bedrooms, and bathrooms. Pp & Or tend to be Midwest to western states. Bl, Br, Rd, and Grn are a little cheaper than average in price but smaller bedrooms & bathrooms. Grn is around southwestern region, Br is around central, and Bl is northeast. Gry is the high priced, multi bedroom homes. Or & Rd have different prices but PC3 indicates they might have different square footage and regions

Hierarchical Clustering







- The silhouette scores have become more stable than the previous hierarchical clustering metrics.

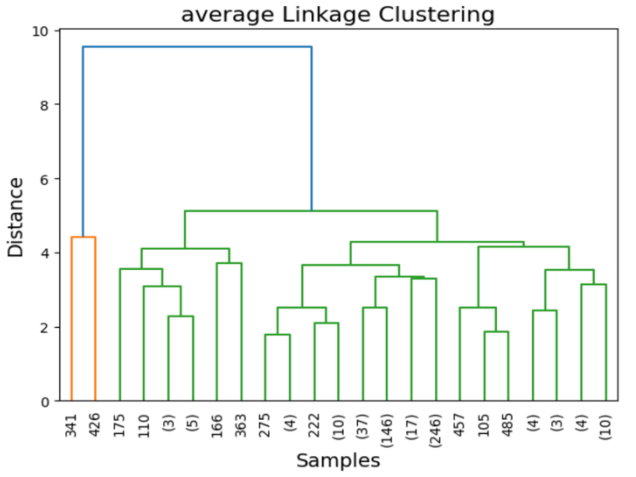
- Complete, average, and ward are performing similarly. Looking at their dendrograms they all have two clusters that are well separated.

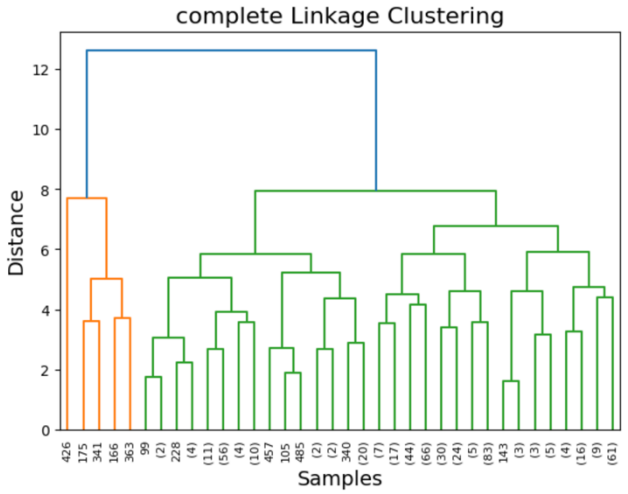
- Centroid and single linkage struggled since they have single observation clusters. Additionally, they are not as well separated. The silhouette score for single linkage explains its dendrogram distances.

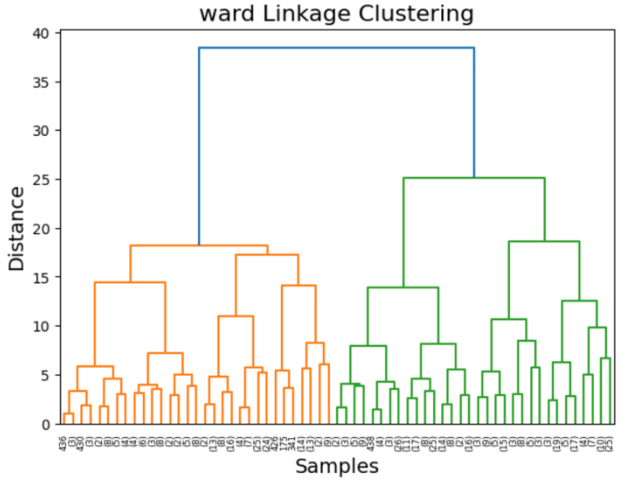
- 7 to 8 clusters are a decent number of clusters that can explain our feature space without too much overlapping in k-means. We would cut at these distances: For Ward linkage, cut at 15. For complete cut at 5.5. For average cut at 4.

- The model tends to cluster for 2 clusters and appears to have one cluster that has a majority of the observations.

- Average has a similar shape to centroid which may indicate smaller distribution of values/removed outliers. Complete and Ward are better at separating the clusters which could be from how outliers were handled. However, Ward is much better at creating balance & reducing within cluster variance.







6) Interpretation: state why a model is performing better/worse. What are the most significant features? Why is the model classifying or clustering to a specific cluster? How do the model results relate to the dataset and the problem?

The models are performing better after PCA. For K-Means, the clusters where more pronounced than before. Additionally, the silhouette scores from the first trial decreased as cluster size increased. However, with PCA, silhouette scores initially decreased and then increased as cluster sizes increased.

The hierarchical clusters group into 2 clusters. From k-means scatter plots you can see two distinct groups at PC1 = 0 & PC3 = -1, but one group is always bigger than the other. This could explain why the green cluster is much larger than the orange in some hierarchical linkage methods.

The most significant features appear to be square feet, price, bedrooms, and bathrooms. PC1 is correlated with the most significant features, thus splitting at PC1=0 shows cheap vs expensive apartments. PC3 is associated with square footage and splitting at PC3 = -1 shows larger vs smaller apartments. This means expensive apartments tends to be bigger and have more bedrooms/bathrooms. The model may classify to the large cluster more, since there are more observations for a certain cluster. Furthermore, it may also classify to the larger cluster due to how the similar most of the observations are.

The model results show that the data has two main groups and the relationship between the groups depend on the significant features, especially price and square feet.

7) Conclusions: summarize the project.

After cleaning and preprocessing the dataset, I tested two datasets – one with dummy variables and one with only numerical features. The numerical only dataset performed better than the other dataset. Then, I performed PCA on the numerical only dataset to find improvements to the clustering methods. The improvements involved more distinguished clusters/grouping and more stable silhouette scores as cluster size increased. The feature space is now understood that price, square footage, and number of bedrooms & bathrooms are the most important features which means the observations will be grouped based on those features first. For future improvements, I would like to perform DBSCAN.

8) References

James, G. *et al.* (2023) *An Introduction to Statistical Learning with Applications in Python*.

pcp21599 (2023) *Hierarchical clustering in data mining*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/ (Accessed: 01 December 2024).

sklearn developers (no date) *Kmeans*, *scikit*. Available at: https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html (Accessed: 01 December 2024).