Aaron Hung

Real Estate Valuation

Introduction

In this project, we will be using a historical market real estate valuation data set from Sindian Dist., New Taipei City, Taiwan, to find which variables are the most impactful in predicting the price of real estate. Using regression analysis, principal component analysis, and cluster analysis, we will analyze the data set to find the most important predictor variables and identify underlying patterns in the data.

Dataset Description

Link to dataset:
https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set

Variables Table

Variable Name	Role	Type	Description	Units	Missing Values
No	ID	Integer			no
X1 transaction date	Feature	Continuous	for example, 2013.250=2013 March, 2013.500=2013 June, etc.		no
X2 house age	Feature	Continuous		year	no
X3 distance to the nearest MRT station	Feature	Continuous		meter	no
X4 number of convenience stores	Feature	Integer	number of convenience stores in the living circle on foot	integer	no
X5 latitude	Feature	Continuous	geographic coordinate, latitude	degree	no
X6 longitude	Feature	Continuous	geographic coordinate, longitude	degree	no
Y house price of unit area	Target	Continuous	10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared	10000 New Taiwan Dollar/Ping	no

Rows per page 25 0 to 8 of 8

The dataset includes 6 feature variables and 1 target variable with 414 instances. The feature variables include transaction date, house age, distance to the nearest MRT station, number of convenience stores, latitude, and longitude. These variables will be labeled X1 to X6 respectively. The target variable is the house price of unit area. All of the variables are continuous, except the number of convenience stores which is discrete (integral).

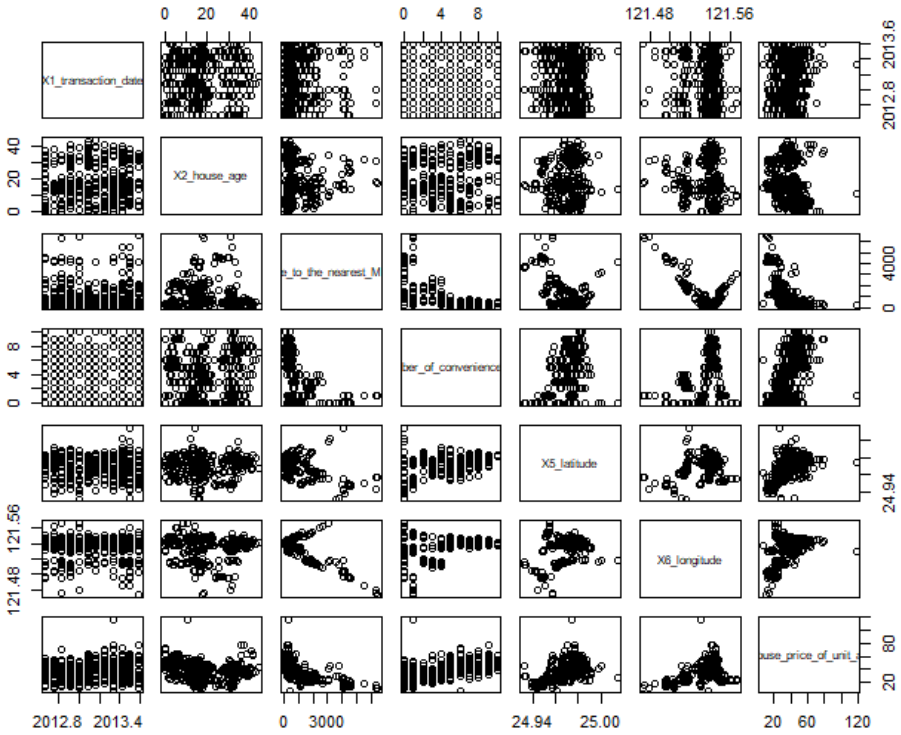
Univariate Analysis

```
X1_transaction_date  X2_house_age    X3_distance_to_the_nearest_MRT_station X4_number_of_convenience_stores  X5_latitude     X6_longitude
Min.   :2013        Min.   : 0.000   Min.   :  23.38                        Min.   : 0.000                  Min.   :24.93   Min.   :121.5
1st Qu.:2013        1st Qu.: 9.025   1st Qu.: 289.32                        1st Qu.: 1.000                  1st Qu.:24.96   1st Qu.:121.5
Median :2013        Median :16.100   Median : 492.23                        Median : 4.000                  Median :24.97   Median :121.5
Mean   :2013        Mean   :17.713   Mean   :1083.89                        Mean   : 4.094                  Mean   :24.97   Mean   :121.5
3rd Qu.:2013        3rd Qu.:28.150   3rd Qu.:1454.28                        3rd Qu.: 6.000                  3rd Qu.:24.98   3rd Qu.:121.5
Max.   :2014        Max.   :43.800   Max.   :6488.02                        Max.   :10.000                  Max.   :25.01   Max.   :121.6
Y_house_price_of_unit_area
Min.   :  7.60
1st Qu.: 27.70
Median : 38.45
Mean   : 37.98
3rd Qu.: 46.60
Max.   :117.50
```

```
> round(apply((rev_data_reduced), MARGIN = 2, FUN = sd),4)
                  X1_transaction_date                         X2_house_age X3_distance_to_the_nearest_MRT_station         X4_number_of_convenience_stores
                               0.2820                              11.3925                              1262.1096                                  2.9456
                         X5_latitude                        X6_longitude            Y_house_price_of_unit_area
                               0.0124                               0.0153                               13.6065
> round(apply((rev_data_reduced), MARGIN = 2, FUN = var),4)
                  X1_transaction_date                         X2_house_age X3_distance_to_the_nearest_MRT_station         X4_number_of_convenience_stores
                               0.0795                             129.7887                           1592920.6308                                  8.6763
                         X5_latitude                        X6_longitude            Y_house_price_of_unit_area
                               0.0002                               0.0002                               185.1365
```

Since the variables have different units, the variances are drastically different in values. It is a good idea to scale the data, which we do as necessary.

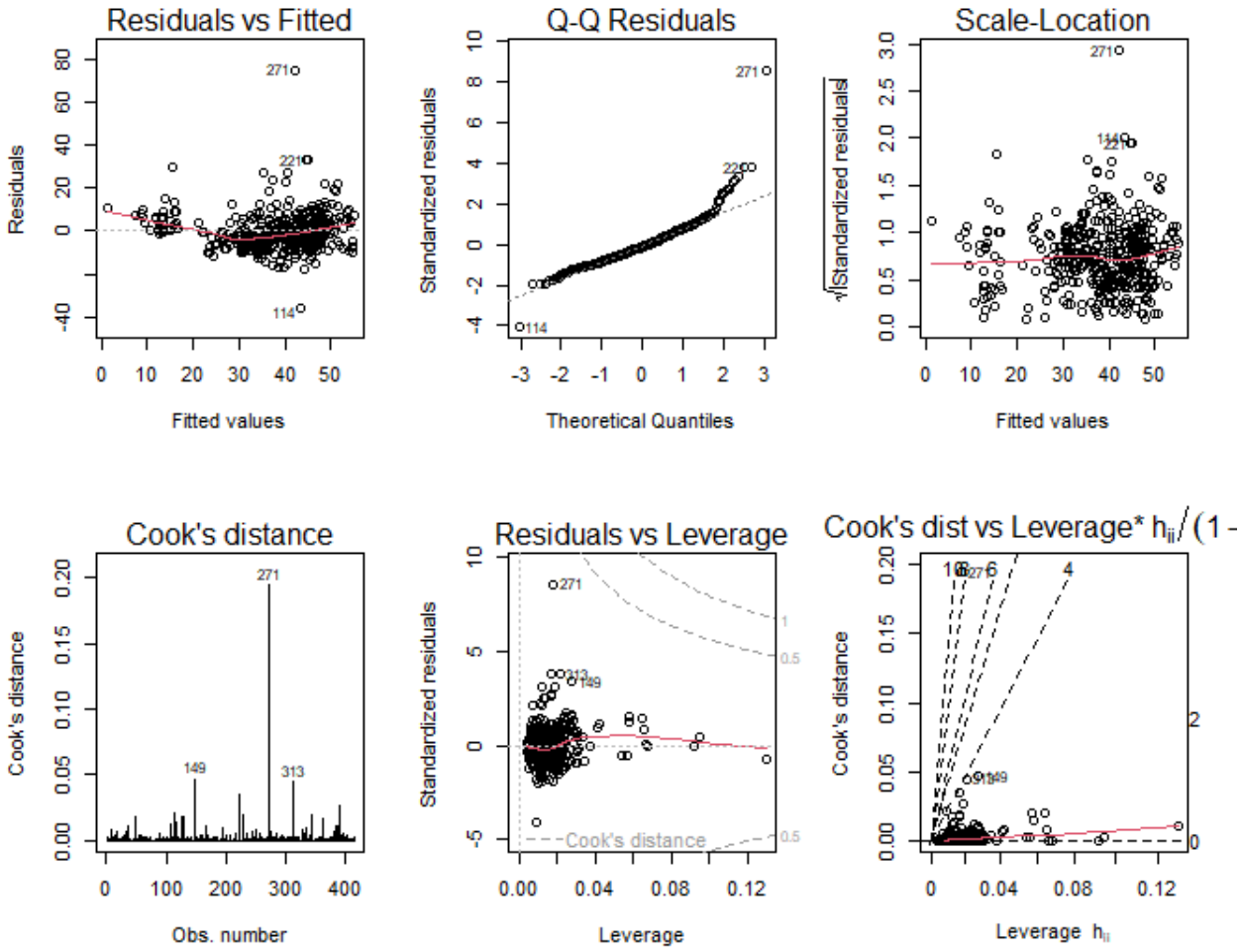
Multiple Linear Regression



The matrix of scatterplots above summarizes the relationships between the response and the predictors.

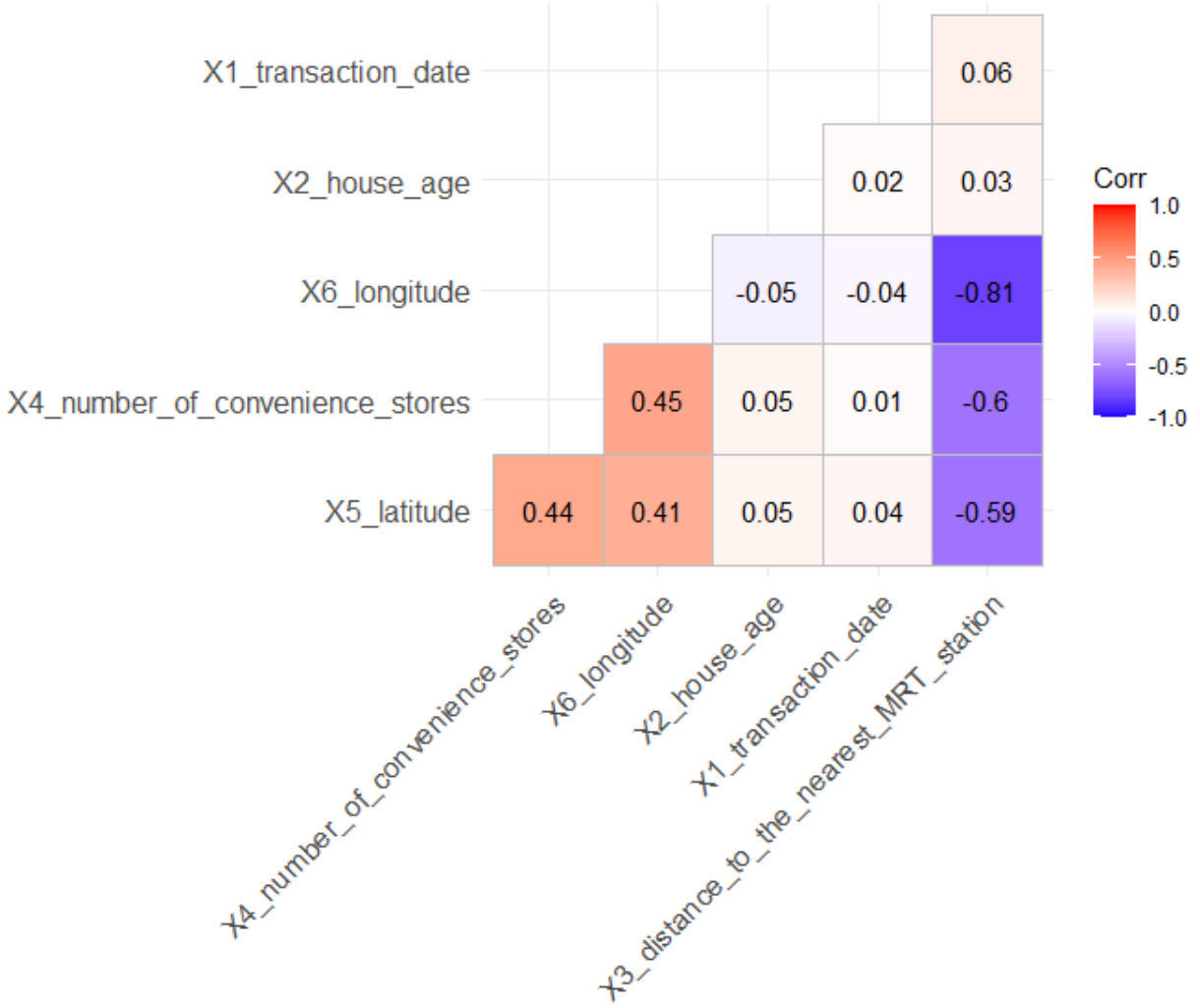
```
Residuals:
    Min      1Q  Median      3Q     Max
-35.667  -5.412  -0.967   4.217  75.190

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -1.444e+04  6.775e+03  -2.132  0.03364 *
X1_transaction_date                     5.149e+00  1.557e+00   3.307  0.00103 **
X2_house_age                           -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
X3_distance_to_the_nearest_MRT_station -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
X4_number_of_convenience_stores         1.133e+00  1.882e-01   6.023 3.83e-09 ***
X5_latitude                             2.255e+02  4.457e+01   5.059 6.38e-07 ***
X6_longitude                           -1.243e+01  4.858e+01  -0.256  0.79820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.858 on 407 degrees of freedom
Multiple R-squared:  0.5824,    Adjusted R-squared:  0.5762
F-statistic:  94.6 on 6 and 407 DF,  p-value: < 2.2e-16
```

The output above is the summary of the linear model. We test the null hypothesis that all regression coefficients are equal to zero against the alternative hypothesis that at least one regression coefficient is not equal to zero. From the output, it is shown that the first five predictors have p-values of less than 0.05 which are significant under the significance level of 0.05. In addition, the F-statistic value is 94.6 which is much greater than 1, with a corresponding p-value of 2.2e-16 which is significant under all the listed significance levels. From this, there is an indication that there is a relationship between at least one of the predictors and the response and we reject the null hypothesis in favor of the alternative hypothesis. Also, we can use the regression coefficient estimates to make predictions for the response variable Y, which is the house price of unit area. The R^2 value is 0.5824 and the R^2adj value is 0.5762. This means that about 58% of the variability observed in the house price of unit area is explained by the model.



The six graphs above summarize part of the checking of potential problems for the linear model. The top left shows the residual plot. If there is a pattern, there may be a problem with the model in terms of non-linearity. However, there is no clear pattern as it seems that the points are pretty spread out around the residual line at 0. This means that there should be no issues with linearity. Transformations such as log and square root can be applied to the predictors for non-linearity. In addition, if the error terms are uncorrelated, then there should be no discernible pattern. This is the case for the data as there is no clear pattern. The top middle shows the Normal Q-Q plot. The majority of the points lie on the line so we can assume the residuals are approximately normally distributed. In the top left and right graph, the red line is approximately linear and there is no clear pattern of a funnel shape so we can assume constant variance of error terms (homoscedasticity). If there is a non-constant variance of error terms (heteroscedasticity) we can use transformations on the response. The bottom three graphs are related to atypical observations. The bottom middle graph shows the standardized or studentized residuals. A rule of thumb is that observations whose standardized residuals are greater than 3 in absolute value are possible outliers. Observation 271 is much greater than 3 in absolute value and as a result is indicated as a possible outlier. There are also a few other observations greater than 3 in absolute

value such as 313 and 149 that are potential outliers. The bottom left graph shows Cook's distance. Cook's distance measures the influence of a data point and how the model changes when the observation is removed. A Cook's distance value larger than 1 usually indicates an influential point. None of the observations are outside of the dashed lines of 0.5 or 1. In the bottom middle and right graphs, we can see that a few of the points have high leverage compared to the rest.



The figure above is the correlation matrix between the predictors. There is one value that stands out in large absolute value which is the one highlighted in dark purple with a value of -0.81. This indicates a strong negative linear relationship between longitude and distance to the nearest MRT station. An explanation for this could be that longitude is location-based, similar to latitude, the number of convenience stores, and the distance to the nearest MRT station. That is why correlation values between these predictors are moderate shown in orange and light purple.

```
            X1_transaction_date                     X2_house_age X3_distance_to_the_nearest_MRT_station
                       1.014655                         1.014287                                4.322984
X4_number_of_convenience_stores                      X5_latitude                             X6_longitude
                       1.617021                         1.610225                                2.926305
```

The output above is the VIF values for the predictors which reveals if multicollinearity exists. A VIF value of 1 indicates the absence of multicollinearity and a value that exceeds 5 or 10 indicates a problematic amount of multicollinearity. None of the values computed exceed 5 or 10, however distance to the nearest MRT station is ~4.323 which is somewhat close compared to the next largest value of ~2.93.

```
  Step Df Deviance Resid. Df Resid. Dev      AIC
1                        407   31931.41 1813.029
```

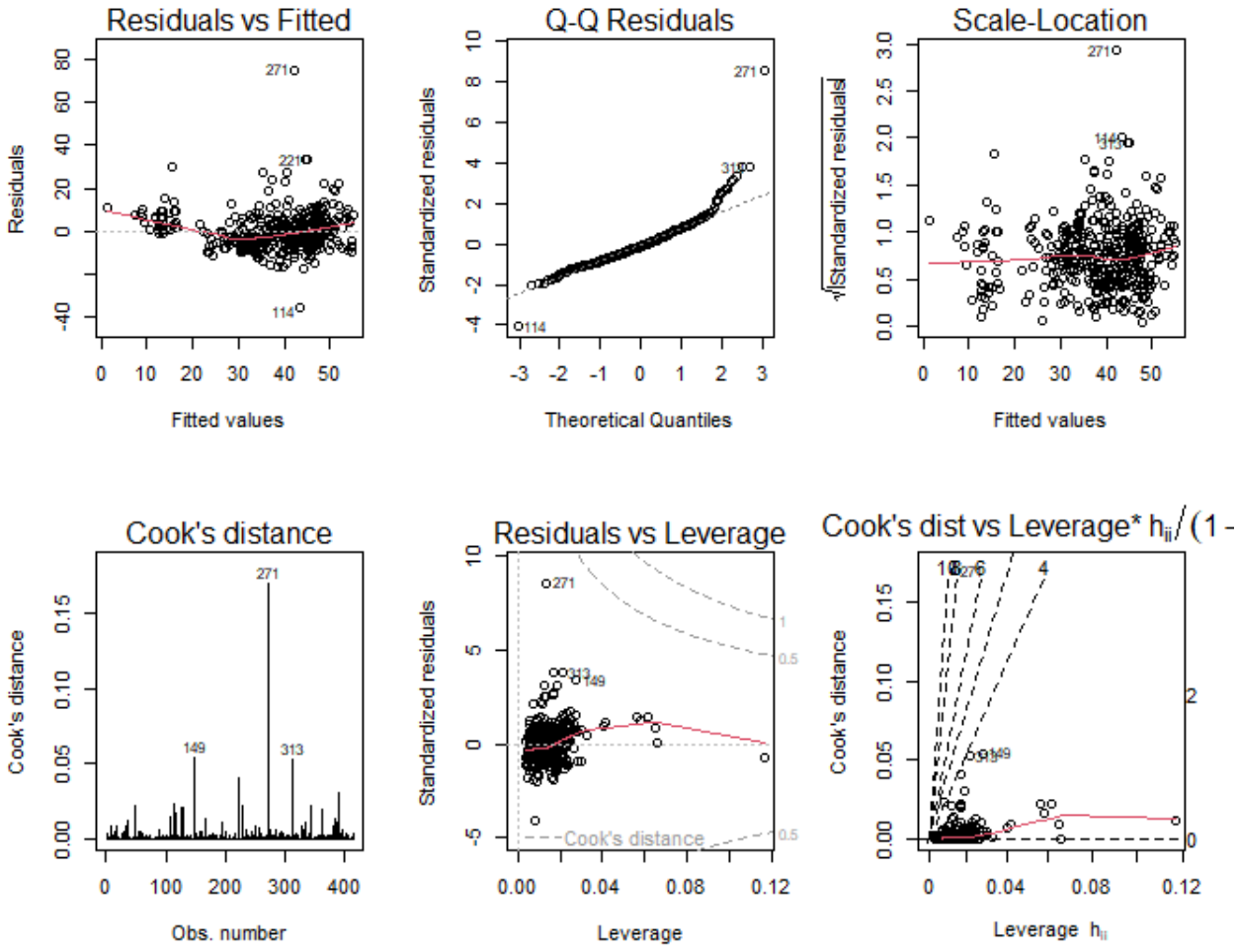
```
            Step Df Deviance Resid. Df Resid. Dev      AIC
1                                  407   31931.41 1813.029
2 - X6_longitude  1 5.135284       408   31936.55 1811.095
```

The two outputs above are the results of forward (top) and backward (bottom) selection using AIC (Akaike information criterion), to find the best selection of predictors. Running forward selection did not change the selection of predictors. However, running backward selection resulted in finding that removing the longitude predictor would result in the best model. We removed the longitude predictor and reran the process to see how the results would change.

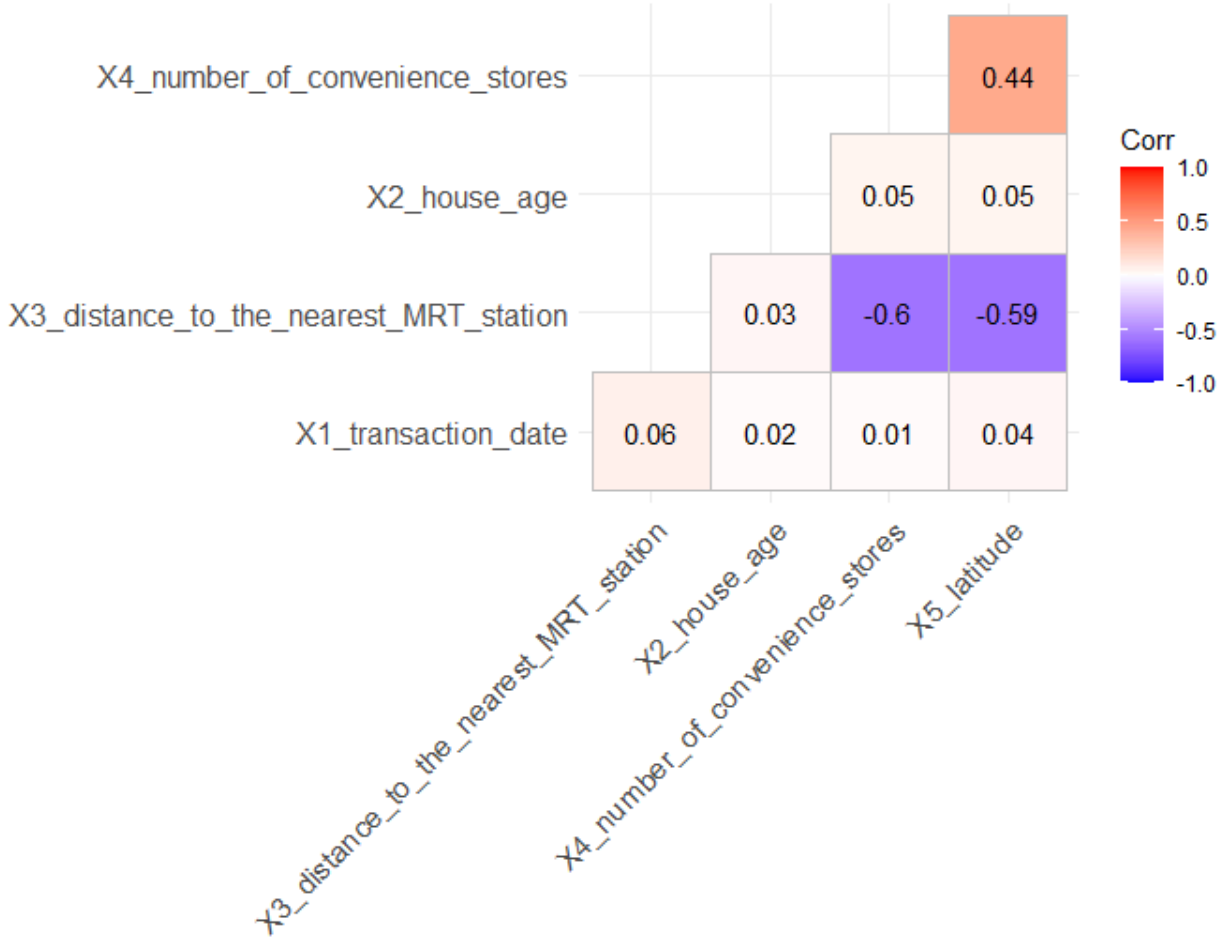
```
Residuals:
    Min      1Q  Median      3Q     Max
-35.625  -5.373  -1.020   4.243  75.343

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -1.596e+04  3.233e+03  -4.938 1.15e-06 ***
X1_transaction_date                    5.138e+00  1.554e+00   3.305  0.00103 **
X2_house_age                          -2.694e-01  3.847e-02  -7.003 1.04e-11 ***
X3_distance_to_the_nearest_MRT_station -4.353e-03  4.899e-04  -8.887  < 2e-16 ***
X4_number_of_convenience_stores        1.136e+00  1.876e-01   6.056 3.17e-09 ***
X5_latitude                            2.269e+02  4.417e+01   5.136 4.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.847 on 408 degrees of freedom
Multiple R-squared:  0.5823,    Adjusted R-squared:  0.5772
F-statistic: 113.8 on 5 and 408 DF,  p-value: < 2.2e-16
```

The output above is the summary of the new linear model with the longitude predictor removed. We test again the null hypothesis that all regression coefficients are equal to zero against the alternative hypothesis that at least one regression coefficient is not equal to zero. From the output, it is shown that all five predictors have p-values of less than 0.01 which are significant under the significance level of 0.01. In addition, the F-statistic value is 113.8 which is much greater than 1, with a corresponding p-value of 2.2e-16 which is significant under all listed significance levels. From this, there is an indication that there is a relationship between at least one of the predictors and the response and we reject the null hypothesis in favor of the alternative hypothesis. The R^2 value is 0.5823 and the R^2adj value is 0.5772. Compared to the model

with longitude included with an R^2 value of 0.5824 and the R^2adj value of 0.5762, both values remained pretty much the same. R^2 value increased by 0.0001 and R^2adj value decreased by 0.001. This means that about 58% of the variability observed in the house price of unit area is still explained by the model with the best selection of predictors. We can again use the new regression coefficient estimates to make predictions for the house price of unit area.



The six graphs above are produced again for the new model. There are subtle differences between the new and old models. However, the results for the potential problems remain unchanged.



The figure above is the new correlation matrix between the best selection of predictors. With longitude removed, there are no more values that are very high in absolute value.

```
          X1_transaction_date                 X2_house_age X3_distance_to_the_nearest_MRT_station
                     1.013815                     1.013243                                2.016820
X4_number_of_convenience_stores                X5_latitude
                     1.611282                     1.585625
```

The output above is the new VIF scores for the predictors. After removing longitude, all values have decreased which is good as it becomes closer to 1 which indicates the absence of multicollinearity.

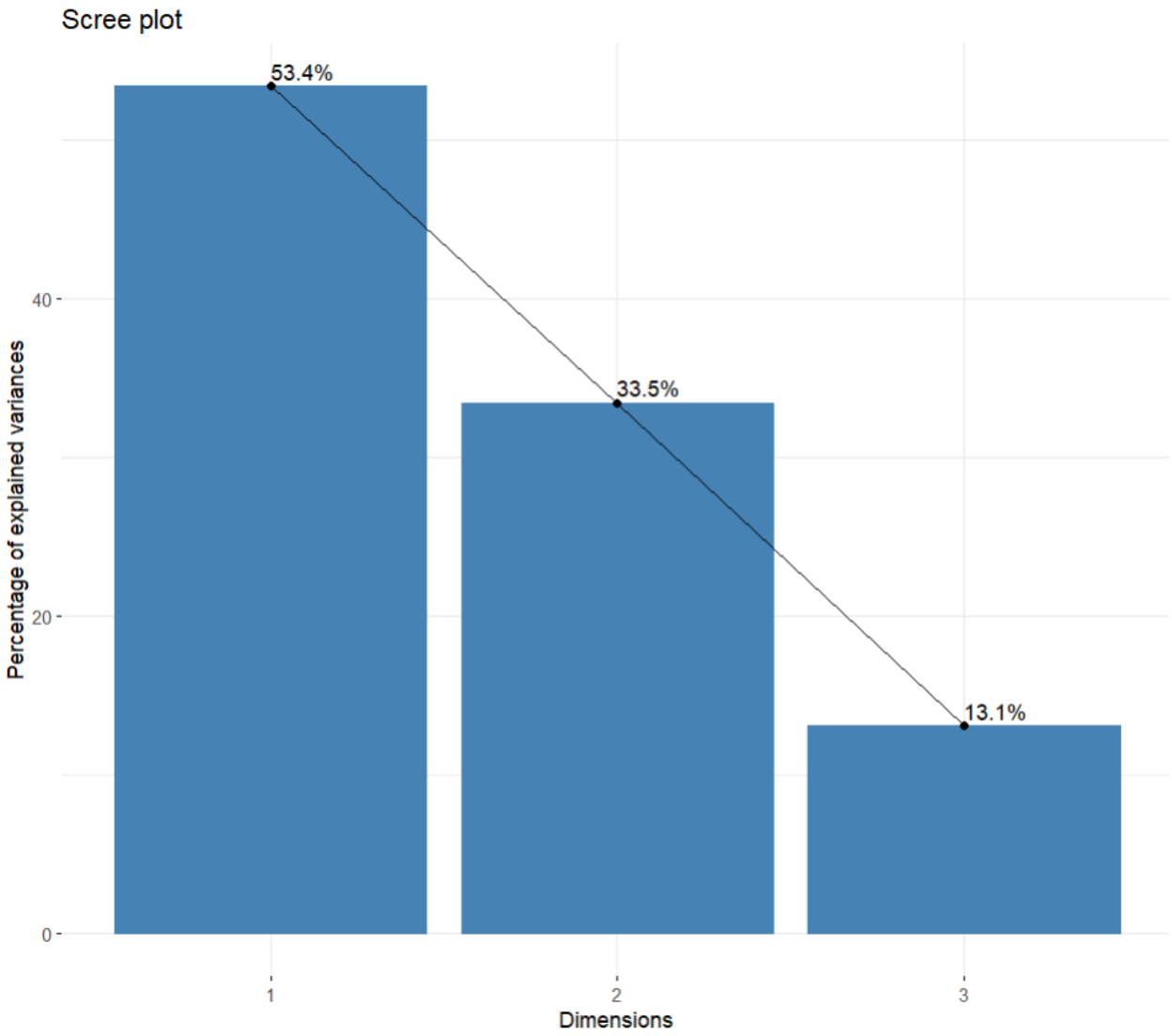
Principal Component Analysis

```
Importance of components:
                          PC1    PC2    PC3
Standard deviation     1.2661 1.0021 0.6268
Proportion of Variance 0.5343 0.3347 0.1309
Cumulative Proportion  0.5343 0.8691 1.0000
```

The purpose of choosing the number of PCs is dimensionality reduction. We want to reduce the number of variables from d to q while keeping as much variability as we can. One of the three ways we learned in class is to use the cumulative proportion of variance explained. This method extracts the number of principal components that are able to explain a majority of the variance, typically 80% or greater. From the output above, we can see that the first principal component explains 53.43% of the variance. The second principal component explains 33.47% of the variance. And the third principal component explains 13.09% of the variance. The first two principal components together explain 86.91% of the variance which is over 80%. Therefore, we select only the first two principal components.

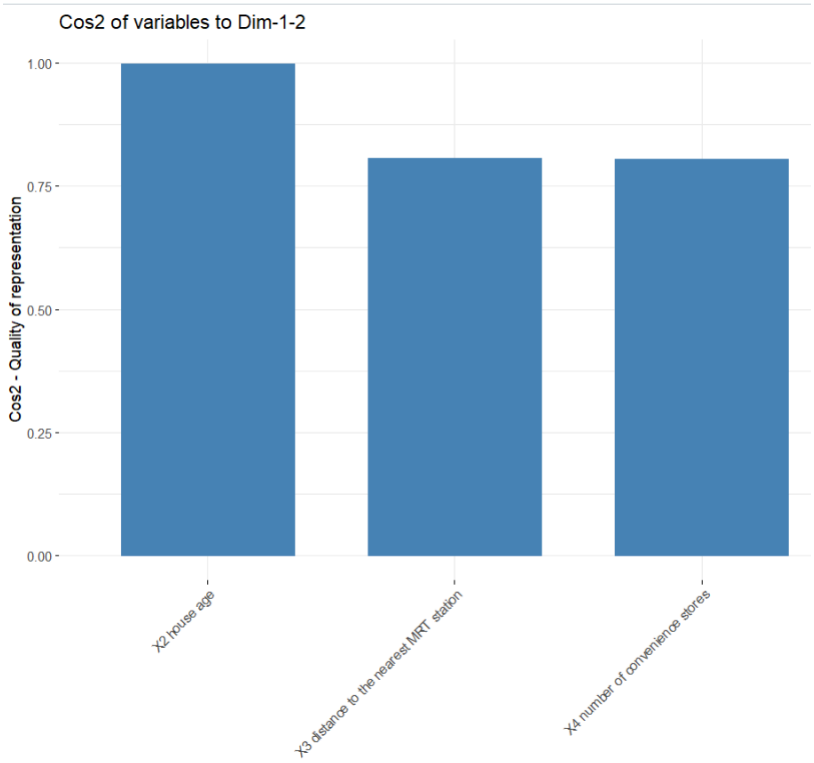
```
      eigenvalue variance.percent cumulative.variance.percent
Dim.1  1.6029974         53.43325                    53.43325
Dim.2  1.0041821         33.47274                    86.90599
Dim.3  0.3928204         13.09401                   100.00000
```

The second way we were taught is to use Kaiser's rule. The rule is to omit the principal components that contain less information than the average information per principal component. Since our data has been standardized, we omit the principal components with eigenvalues less than 1 and keep the principal components with eigenvalues larger than 1. From the output above, the third principal component eigenvalue is ~0.393 which is less than 1. As a result, we omit the third principal component. The first and second principal component eigenvalues are ~1.603 and ~1.004 respectively. Since both eigenvalues are larger than 1, we keep only the first two principal components in our selection which matches the result from the cumulative proportion of variance explained.

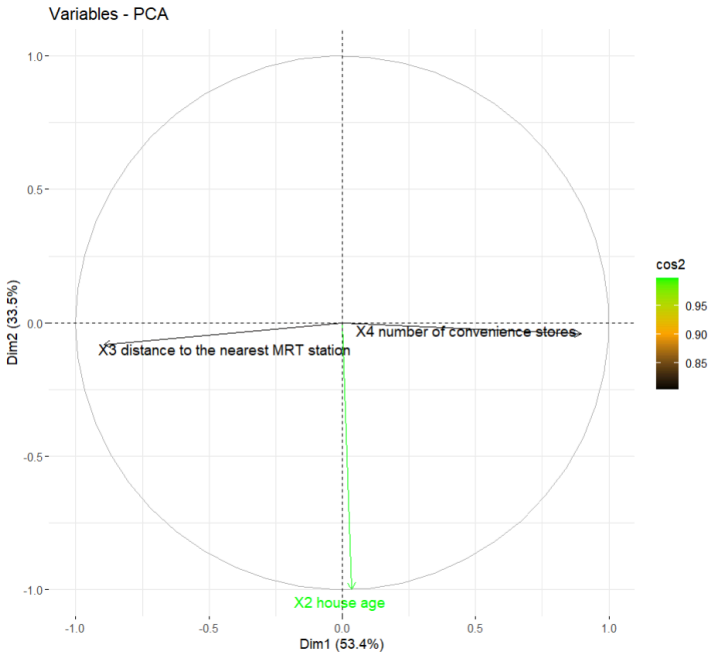


The last way we learned in class is to use a scree plot. To conclude the number of variables from a scree plot, all we have to do is retain as many components as possible until a significant jump on the scree plot appears. In other words, to keep the components to the left of where an elbow shape is present. And if the elbow shape is not discernible, we refer to the cumulative PVE rule. In this case, the elbow shape is indeed indiscernible. The first principal component accounts for 53.4% of the variability and the second principal component accounts for 33.5% of the variability and together accounts for 86.9% of the variability. This is a good compromise and we select only these two principal components.

The bar plot to the right shows how well each feature represents its respective principal component. As you can see, X2 house age is 1 because it is the only feature on that principal component. On the other hand, features X3 and X4 both hover around the .80 mark which suggests that they both equally contribute to their principal component.



The picture to the right shows the biplot for the 3 variables. From what we discussed earlier, X2 house age representation value is 1 which is demonstrated by the length of the arrow X2 in the biplot and the length of X3 and X4 is the same value of around 0.8 also follows the previous graph. We also note that all 3 arrows form very small angles with their respective axis which suggests that there is a high correlation between the feature and its principal component. Also note that the arrows representing features X3 and X4 point in opposite directions which means there is a negative correlation between the 2 features.



Cluster Analysis

To perform cluster analysis on the data, we first need to determine what portion of the data set to use. We start by removing the target variable "Y house price of unit area" and the feature variable "No" which is the ID assigned to each observation, "X1 transaction date", "X5 latitude" and "X6 longitude". We decided to remove variables X5 and X6 because these variables take into consideration the location of the property, an aspect of the data that is already taken into account by other variables such as feature X4, the number of convenience stores in the area, and feature X3, the distance to the nearest MRT station.

Next, we must scale the data. This is an important step because the features in the data set all have drastically different values. For example, the values in feature X2, house age, range from 0 to 43.6, while feature X3, distance to the nearest MRT station, ranges from 23 to 6488. This significant difference in variances will cause the cluster analysis to take into account one feature more than the other. But if we scale the features, the cluster analysis will equally consider all features.

After cleaning the data we can begin our clustering analysis. There are multiple ways to perform cluster analysis. We will start with Hierarchical clustering. In hierarchical clustering, we will use a method called agglomerative hierarchical clustering.

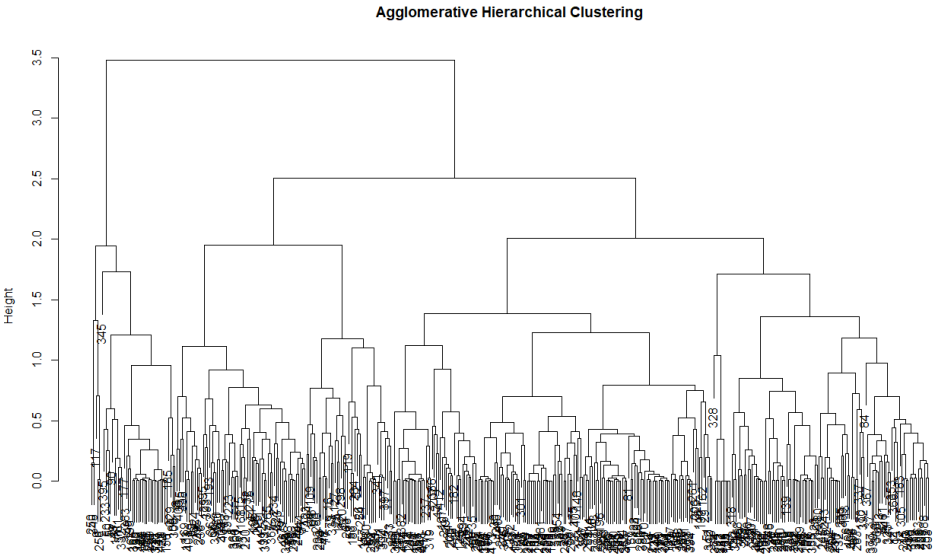
In agglomerative hierarchical clustering, all the individual data entries start as individual clusters. After computing the distance matrix for the data, points that have the lowest distance according to a chosen lineage method are combined. This process repeats until there is one large cluster containing all the points.

For agglomerative hierarchical clustering, we have to determine which linkage method produces the dendrogram that most accurately preserves the similarity and dissimilarity of the data points. To do this we will calculate the cophenetic distances and compare those values to the original distance values. This will result in a correlation coefficient that tells us which clustering method most accurately represents the original data. After testing 8 different clustering methods we can see that average linkage has the highest correlation coefficient with a value of 0.778. So moving forward we will use average linkage.

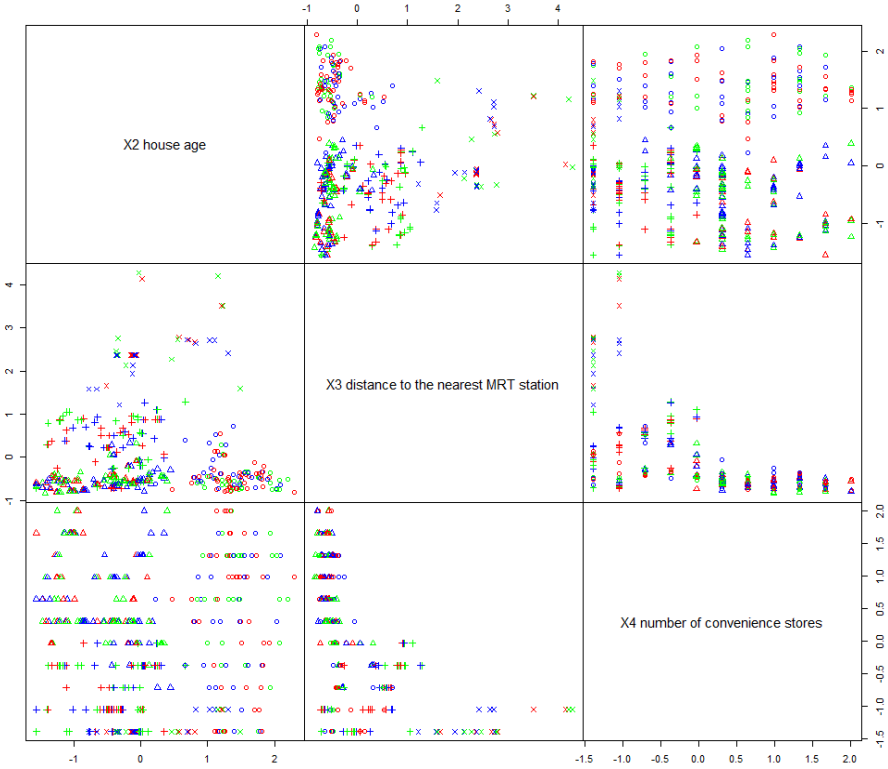
	Method	Correlation
1	ward.D2	0.6384483
2	average	0.7783322
3	ward.D	0.6184286
4	single	0.5748337
5	complete	0.7447914
6	mcquitty	0.6903277
7	median	0.4881614
8	centroid	0.7294392

Using average linkage, we can create a dendrogram to help visualize which points are similar and which are not.

The dendrogram is unable to give us a clear representation of which points are similar and which points are not due to the sheer number of data points. Instead, we will look at the pairs plot, cluster plot, and values within those clusters.



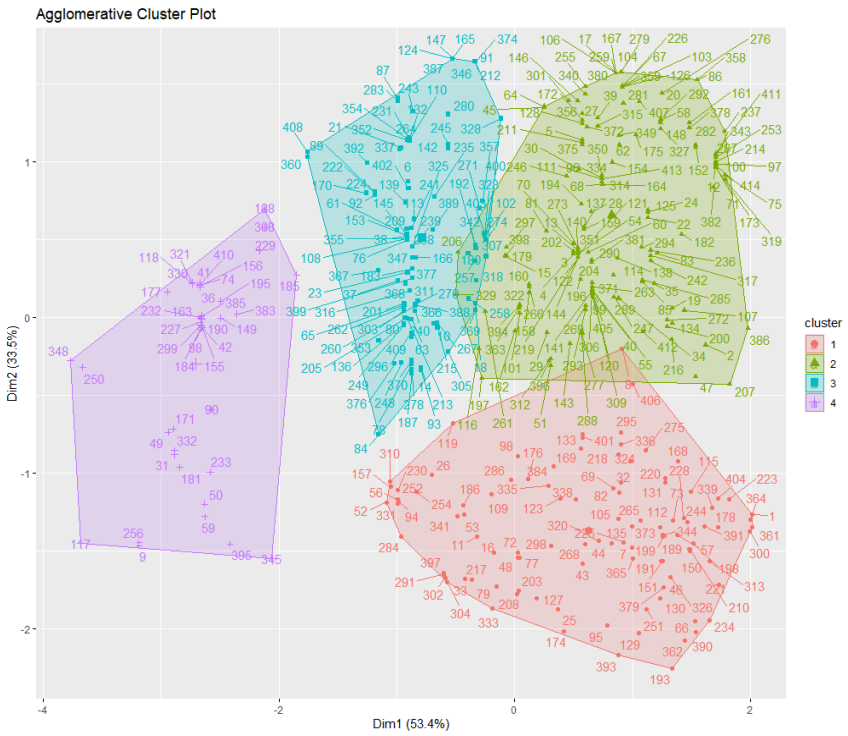
This pairs plot is used to see if there are any pairwise relationships between the features. The lack of separation suggests that the dataset may not exhibit any underlying clustering.



cluster	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores
1	33.57196	533.0879	5.2523364
2	10.40063	390.6678	5.9873418
3	11.98598	1413.1099	1.6635514
4	19.40476	4256.1906	0.2142857

Above is the aggregate data for each variable in 4 different clusters. Looking at the data, we can see that cluster one groups properties that are older, are relatively close to an MRT station, and have a higher number of convenience stores. Contrast that to cluster 4 which groups properties that are average in age, are very far from an MRT station, and have very few convenience stores nearby. The difference in these variables is what creates these clusters. Earlier we saw that variables X3 and X4 have a negative relationship meaning that as one increases the other decreases. We can see that relationship in the clusters. Clusters 1 and 2 both contain points that are closer to the nearest MRT station with a high number of convenience stores while clusters 3 and 4 contain points that are far away from the nearest MRT station and have a low number of convenience stores.

The cluster plot shows the 4 different clusters found during agglomerative hierarchical clustering. We can see that the clusters are relatively well separated with a slight overlap between clusters 2 and 3. From the aggregate data, we saw that clusters 2 and 3 are very similar in property age but, one cluster is much further away from an MRT station while the other cluster has many more convenience stores in their area.

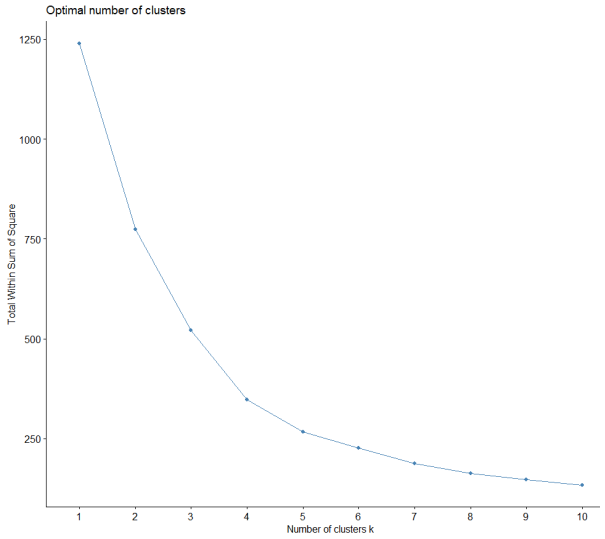


Another clustering method we can employ is K-means clustering. In K-means, we start by selecting K random points to be the centers of the clusters. We then assign each point to the closest center by using the Euclidean distance. Then for each cluster K, compute the new mean value of the cluster. We repeat this process until all the points remain in the same cluster as the previous iteration.

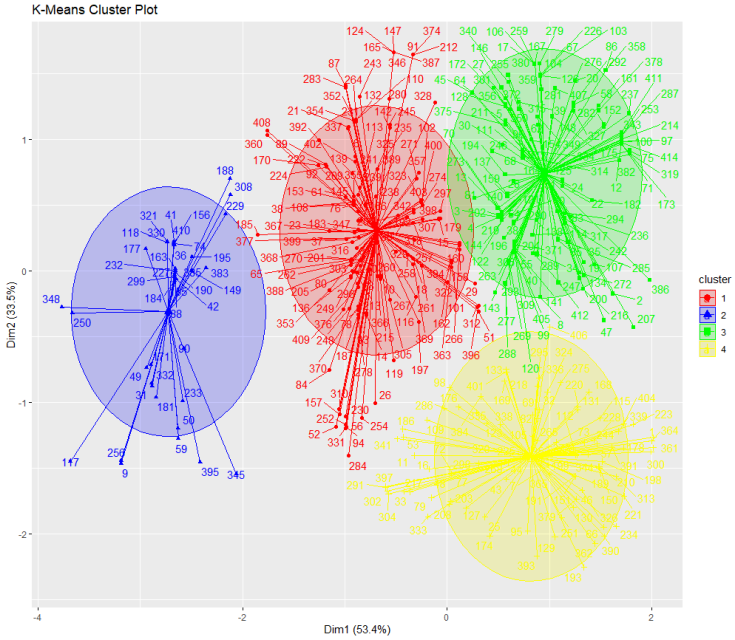
K-means clustering requires that we determine the number of clusters before we apply the algorithm but this means that the data will be split ensuring that the within-cluster dissimilarity is minimized. K-means clustering is also very efficient in dealing with large data sets which may be beneficial to us.

To apply the K-means clustering we must first determine the optimal amount of clusters. To do this we can apply an elbow plot which calculates the total within the sum of squares using different numbers of clusters k to determine the optimal number of clusters. When using this plot we will look for a point where the rate of decrease in total within the sum of squares slows down resulting in an "elbow point" in the graph.

On the graph to the right, it is hard to determine an elbow point but the rate of decrease of the total within the sum of squares drastically decreases after reaching 4 clusters. Therefore, the optimal amount of clusters is 4.



After determining the optimal amount of clusters we can graph the data points using k=4. The graph below shows the 4 different clusters created by K-Means. As you can see the cluster shapes are slightly similar to the clusters we found using Hierarchical clustering.

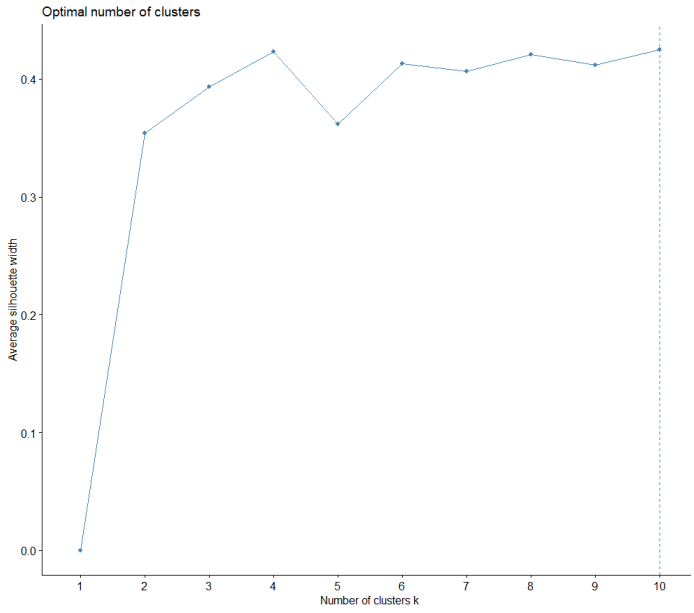


The final clustering method uses K-Medoids.

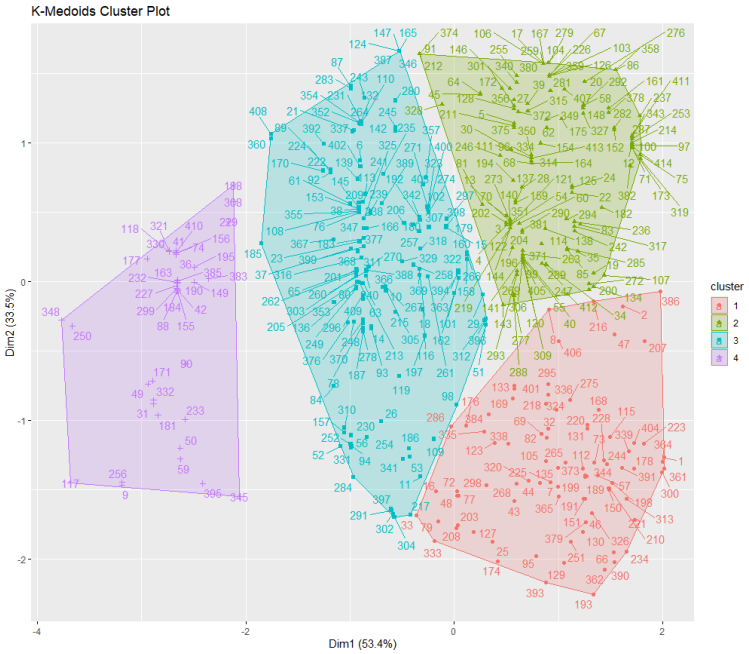
Just like K-Means clustering, in K-Medoids we start by randomly selecting K points to become cluster centers or medoids. Then we assign each point to the closest medoid. Then for each cluster see if any of the points in the cluster decreases the total dissimilarity coefficient. If there is a decrease, the point results in the largest decrease in the total dissimilarity coefficient. We repeat this process until all clusters medoid remain the same.

For K-Medoids clustering, we must also determine the optimal number of clusters. We will use the silhouette method. The silhouette method measures how similar a data point is to its cluster vs another cluster. A higher value indicates that more data points are properly assigned to their clusters making that number of clusters optimal.

The graph on the right shows the average silhouette width for several clusters from 1 to 10. As we can see the number of clusters with the highest average silhouette cluster is 4. This matches what we found using the elbow plot in K-Means clustering.



The K-Medoid plot to the right shows similar results to the K_Means and Agglomerative clustering plots. All three clustering methods produce slightly different cluster plots which supports the pair plot's conclusion that there are no clear clusters in the data.



Finally, we can compare the aggregate data for each of the clustering methods to see if they support the same conclusions as the plots.

cluster	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores
1	1.3920934	-0.4364104	0.3931792
2	-0.6418203	-0.5492533	0.6427089
3	-0.5026629	0.2608523	-0.8251911
4	0.1485367	2.5134940	-1.3172079

cluster	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores
1	-0.2941288	0.2101032	-0.7519008
2	0.1598937	2.5452010	-1.3154335
3	-0.7251369	-0.6149635	0.7626310
4	1.4360151	-0.5224770	0.5819969

cluster	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores
1	1.3482002	-0.5549417	0.8377332
2	-0.7889026	-0.6132708	0.6794579
3	-0.1300994	0.1921716	-0.7637281
4	0.1598937	2.5452010	-1.3154335

As we can see, the cluster's aggregate data are all relatively similar. For example, cluster 1 in Picture 1, cluster 4 in Picture 2, and Cluster 1 in Picture 3, all hold the data points with the highest house age, a shorter distance from the nearest MRT station, and a relatively higher number of convenience stores nearby.

We want to check the optimal number of clusters given for K-means, ward.D2, and average clustering methods based on different clustering indices. Ward.D2 suggests that the optimal number of clusters is 5 while K-means and average both suggest 3 to be the optimal number of clusters. Even though the optimal number of clusters varies between 3 and 5, in all 3 tests, 3 and 5 clusters were top candidates regardless of method suggesting that the results are very close and could vary based on the clustering method.

K-Means:

```
*******************************************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 4 proposed 4 as the best number of clusters
* 6 proposed 5 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

ward.D2:

```
*******************************************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 5 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 8 proposed 5 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  5
```

Average:

```
*******************************************************************
* Among all indices:
* 6 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 6 proposed 5 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 3 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

We can also compare different clustering algorithms and determine which clustering solution is the best given several criteria. There are 2 different measures we will be using, the internal measures and the stability measures. There are 3 different types of internal measures, connectivity, Dunn, and Silhouette. Connectivity measures how well the clusters are formed, meaning are points close to each other assigned to the same cluster. Then there's the Dunn index, the Dunn index of how separated the clusters are while also recording how compact the points in individual clusters are. Finally, we have the Silhouette which accomplishes a similar goal using a different approach. Instead of measuring how separated different clusters are and how compact points in each cluster are, the silhouette index determines the quality of the clusters by seeing how well each data point fits in its assigned cluster versus another cluster. Both the Dunn and Silhouette indexes want to be maximized while the Connectivity index wants to be minimized

```
Validation Measures:
                                 2       3       4       5

hierarchical Connectivity   4.2401  9.5687 19.6175 29.9813
             Dunn           0.1303  0.0807  0.0797  0.0837
             Silhouette     0.4461  0.4011  0.4146  0.4301
kmeans       Connectivity  10.8484 23.8996 22.5349 32.9972
             Dunn           0.0753  0.0264  0.0773  0.0593
             Silhouette     0.4335  0.4005  0.4350  0.4469
pam          Connectivity  27.2595 32.5607 46.2671 45.7075
             Dunn           0.0301  0.0270  0.0227  0.0549
             Silhouette     0.3538  0.3931  0.4231  0.3620

Optimal Scores:

             Score  Method       Clusters
Connectivity 4.2401 hierarchical 2
Dunn         0.1303 hierarchical 2
Silhouette   0.4469 kmeans       5
```

In the chart above we can see that in terms of connectivity and Dunn, hierarchical clustering with 2 clusters is the best available method. According to silhouette, K-means with 5 clusters is optimal

Next, we have the stability measures. The 4 stability measures include average proportion of overlap (APN), average distance (AD), average distance between means (ADM), and figure of merit (FOM). All 4 measures are based on the confusion matrix of the original clustering of the full data versus clustering when a column is removed. APN measures the average proportion of observations not placed in the same cluster before and after removing the column. AD measures the average distance between observations placed in the same cluster before and after removing a column. ADM measures the average distance between cluster centers for observations placed in the same cluster before and after removing a column. FOM measures the average intra-cluster variance of the deleted column versus all the columns. APN, ADM, and FOM all range from 0 to 1, while AD ranges from 0 to infinity. In general, lower values suggest high stability within the cluster sample.

The table to the right shows the validation measures for Hierarchical, K-means, and PAM for 2 to 5 clusters using the 4 stability measures. The scores for APN, AD, and ADM are all relatively low and conclude that 2 or 5 clusters using Hierarchical or K-Means is optimal. While FOM scores a high score of 0.8717 which is not preferred suggested 4 clusters with K-means.
While doing K-Means clustering, the scree plot concluded that 4 clusters were optimal but the validation measures suggest that this is not the case.

```
Validation Measures:
                        2      3      4      5

hierarchical APN   0.1432 0.1977 0.1494 0.1951
             AD    1.9543 1.6928 1.5184 1.4574
             ADM   0.3770 0.6111 0.7223 0.8063
             FOM   0.9673 0.9488 0.9152 0.9139
kmeans       APN   0.1437 0.2914 0.2046 0.2594
             AD    1.9540 1.6823 1.4462 1.3743
             ADM   0.3889 0.7344 0.6678 0.7529
             FOM   0.9660 0.9020 0.8717 0.8800
pam          APN   0.2679 0.2645 0.3438 0.3354
             AD    2.0195 1.6796 1.4972 1.3889
             ADM   0.9025 0.7562 0.7456 0.7065
             FOM   0.9925 0.9025 0.8924 0.8814

Optimal Scores:

    Score  Method       Clusters
APN 0.1432 hierarchical 2
AD  1.3743 kmeans       5
ADM 0.3770 hierarchical 2
FOM 0.8717 kmeans       4
```

Conclusion

In multiple linear regression, we first looked at the matrix of scatterplots to briefly view the relationships between the response and the predictors. We then obtained the linear model output which contained the intercept and regression coefficient estimates which we can use to predict the value of house price. In addition, we can use the p-values and the F-statistic to test the null hypothesis against the alternative. Furthermore, the R^2 and the R^2adj tell us how much of the variability in the response is explained by the model. After that, we checked for potential problems: normality of residuals, non-linearity, non-constant variance of error terms, correlation of error terms, collinearity, multicollinearity, and atypical observations. Then, we ran stepwise selection to find the best selection of predictors. Finally, we repeated the process using the best selection of predictors and made comparisons with the previous model.

In principal component analysis, we used the cumulative proportion of variance explained, Kaiser's rule, and Scree Plot to determine the optimal number of principal components which ended up being 2 and together were able to explain 86.9% of the variability. In addition, the number of variables was reduced from 6 to 3. X2 house age was the major contribution to principal component 1 as the angle between the respective arrow and principal component 1 was close to 180 degrees. And for X3 and X4, the angles were both close to 90 with respect to principal component 1. X3 and X4 were roughly equal in contribution to principal component 2

as the angles were close to 180 and 0 respectively. X2's angle was close to 90 with respect to principal component 2.

In cluster analysis, we attempted to create clusters using hierarchical clustering, k-means clustering, and k-medoids clustering all to middling success. To determine if there were real clusters in the dataset and see the validity of the 3 cluster plots, we performed cluster validation. There was no evidence to support our original analysis from the cluster validation tests as they all suggested a different number of optimal clusters. The only test that supported the idea of 4 clusters was the FOM stability measure which also had a very high score of 0.8717 which is not ideal. Overall, we can conclude that this data set has no clear cluster groups as proven by the pairs plot, changes in cluster plots when using different methods, and the lack of cohesion behind all the validation tests.