Report 2b:

Unsupervised Learning: Clustering

Team 5: Yunhao Bai, Zhiyi Chen, Michelle Guan, Huiqiong Wu

BUS212a - Analyzing Big Data II

Prof. Arnold Kamis

Spring 2022

1. Introduction

As housing is a basic human necessity, many people are concerned about apartment rental prices. Those that seek housing are interested in what they can afford with their budgets. Those that can provide housing are interested in what the market is willing to offer. Housing listings can be found all over the internet, and appropriate pricing of rental properties can be unclear without a basis of comparison. To better understand the condition of the rental housing market, this report aims to explore the relationship between pricing and features of a property, specifically apartment units. The variable of interest in this venture is apartment rental prices across the United States. This study will pay more attention to analyzing the practical aspects that lead to the rental pricing among different types of apartments, using classification and regression methods, as well as integrating cluster analysis. By performing this study, suggestions can be produced for people with different expectations based on an apartment unit's location, size, offered amenities, etc.

2. Dataset

The dataset used in this study is an original dataset on Kaggle that was collected from rental housing listings on Craigslist in 2020. As a popular website that caters to communities across the United States, the data from Craigslist can be useful in understanding the rental housing market nationwide. Additionally, the data is collected from early 2020, which avoids the impact of the Covid-19 pandemic.

Since the raw data contains 18 variables with over 384,000 observations, data wrangling and down-sizing is necessary. Outliers and missing values are dropped to avoid misleading information, and duplicates are also removed to ensure randomization during the data preparation process. The focus of the housing type has been limited to only "apartments".

The cleaned new dataset is random sampled and down-sized to 20,000 observations for the representativeness of the data. Moreover, diagnostics and data splits will be performed to check the validity of the data, and classification and regression will then be used to find the best model. To start, the cleaned data should be explored.

3. Exploratory Analysis

3.1 Data Type

```
Descriptive analysis
Target variable: price_range (categorical)
    [Low; High]
    integer replacement: [0, 1]
Target variable: price (numeric)
Other variables (predictor variables):
 1. sqfeet: square feet (numeric)
 2. beds: number of beds (numeric)
 3. baths: number of baths (numeric)
 4. cats_allowed: (binary)
 5. dogs_allowed: (binary)
 6. smoking_allowed: (binary)
 7. wheelchair_access: (binary)
 8. electric_vehicle_charge: (binary)
 9. comes_furnished: (binary)
 10. laundry_options: (categorical)
        [no laundry on site, laundry in bldg, laundry on site,  w/d hookups, w/d in unit]
        integer replacement: [1:4]
 11. parking_options: (categorical)
        [no parking, carpot, attached garage,  off-street parking, street parking, valet parking, detached garage]
        integer replacement: [1:6]
 12. state: (categorical) {for linear model}
```

Figure 3.1

The purpose of this research is to find out how rental housing prices are affected by different variables. In order to perform the analysis, 12 predictor variables are selected based on experience and common knowledge to examine how these variables affect the target variable which is price. The numeric price variable is segmented into a binary variable called “price_range" to further analyze the sensitivity of the target variable when interacting with predictor variables. Price range is defined by using the median of the price variable, which is $1,125. All prices below $1,125 are described as low price range (0), and all prices above $1,125 are described as high price range (1). Using the median of price also helps to eliminate imbalance problems in classification.

3.2 Statistical Summary

	Min	1st Qu.	Median	Mean	Mode	Sd	Variance	Range	Skew	3rd Qu.	Max
Price	215	850	1125	1289	1200	661.86	438058.7	7780	2.22	1535	7995
Sqfeet	501	748	900	949.3	1000	290.1	84158.01	2499	1.52	1100	3000
Beds	0	1	2	1.79	2	0.76	0.5776	8	0.58	2	8
Baths	0	1	1	1.356	1	0.53	0.2809	6	1.07	2	6

Figure 3.2

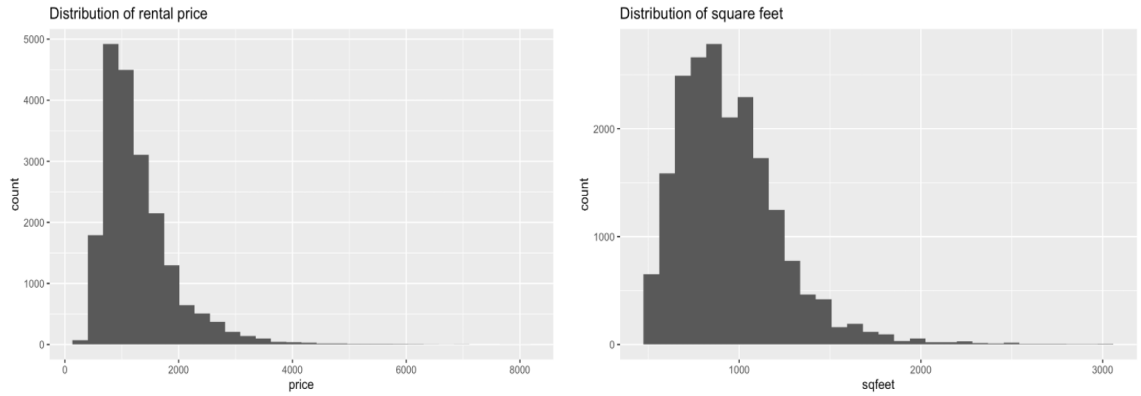


Figure 3.3

The dataset contains 20,000 randomly selected observations about rental apartments across the country. There are four numerical variables that were analyzed: *price*, *sqfeet* (square feet), *beds* (number of bedrooms), and *baths* (number of bathrooms). Based on the statistical analysis, all four variables have a large range between the minimum and maximum values and are highly skewed. In particular, *price* and *sqfeet* have high dispersion with high standard deviations, which means price and square feet have high variability. Additionally, as seen in Figure 3.3, there is skewness in the distributions of both rental price and square footage. The existence of outliers will skew the distribution of the data, and cause bias in future model training.

In order to eliminate potential bias, an IQR method is used to remove outliers. The formula returns the first quartile (25%), Q1, the third quartile (75%), Q3, and the IQR range, 0.0839. Then, the data range is defined by using a lower limit of Q1-1.5*IQR and an upper limit of Q3-1.5*IQR to help subset the cleaned data. Using IQR to eliminate outliers is a reasonable practice here because it uses data points in the middle 50% which provides an unbiased starter to find the range of outlier numbers. By performing IQR, outliers of the dataset are removed, representing 1.075 percent of the total data. This only eliminates a small portion of the original data, causing no integrity issues in the dataset. Moreover, to make both dependent and independent variables more Gaussian, a data transformation is conducted on the skewed variables by introducing logged terms.

3.2.1 Box Plot for Numerical Variables

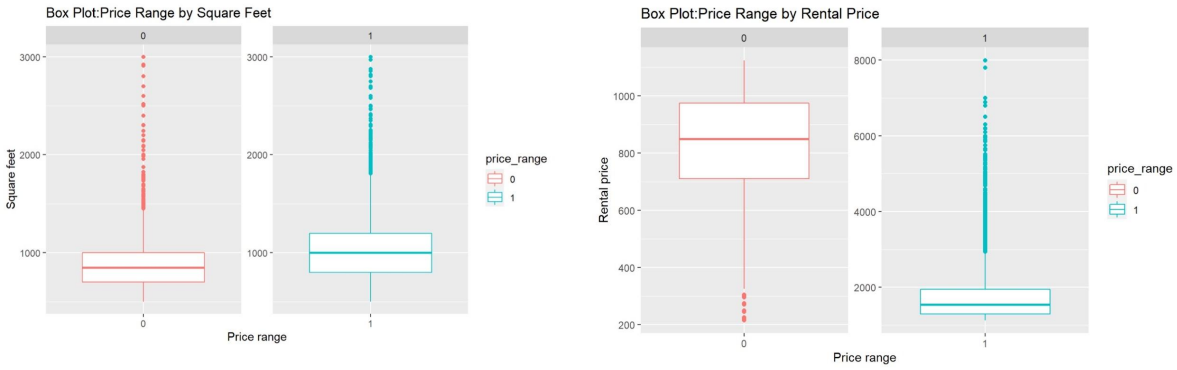


Figure 3.4

Based on the box plot on the left side of Figure 3.4, an apartment in the low price range (0) is around 800 square feet on average and an apartment in the high price range (1) is around 1,000 square feet on average. Both price ranges have no low-value outliers, but have high-value outliers. In both ranges, most apartments are around 1,000 square feet or less, which might indicate square footage is not heavily impactful in determining price range after a certain value of square footage is reached. Rental prices in both price ranges are shown on the right side in Figure 3.4. For the low price range, the mean rental price is around $800 with very few extreme low price values. For the high price range, the mean rental price is around $1,500 with the majority of prices under $2,000 and some extreme high price values.

3.2.2 Bar Plot for Numerical Discrete Variables

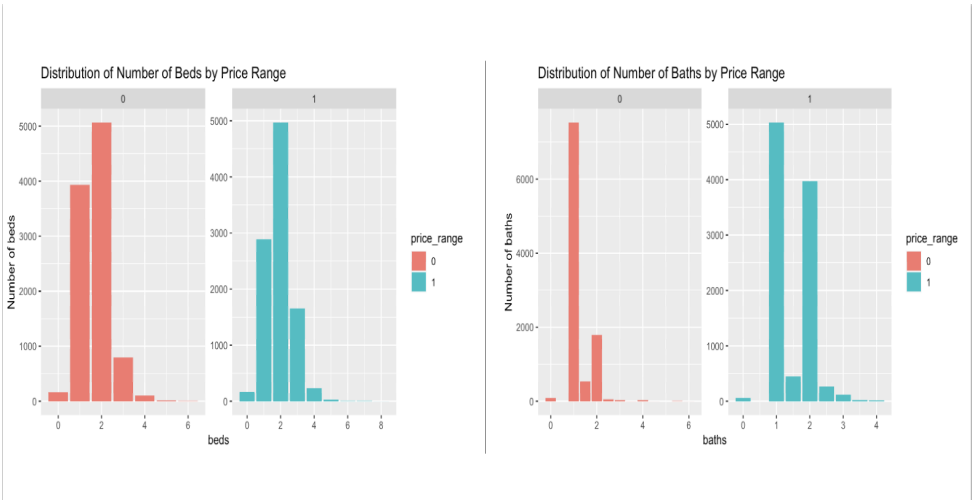


Figure 3.5 Bar Plot for Numerical Discrete Variables

Bar charts are very useful to show the distributions of the number of beds and

number of baths in each of the two price ranges. On the left side of Figure 3.5, the majority of the number of beds is within the range of 1 to 3 for both low and high price range. There are more 4-beds data in the high price range than in the low price range and more 1-bed data in the low price range than in the high price range. The overall distributions of the two price ranges by number of beds are very similar, so the number of beds might not be very influential on the target variable. From the right side of Figure 3.5, there is a notable difference in the number of baths between price ranges. The high price range has more baths than the low price range. Most of the baths are in the range 1 to 2. Therefore, the number of baths might be impactful on the target variable.

3.2.3 Box Plot for Categorical Variables

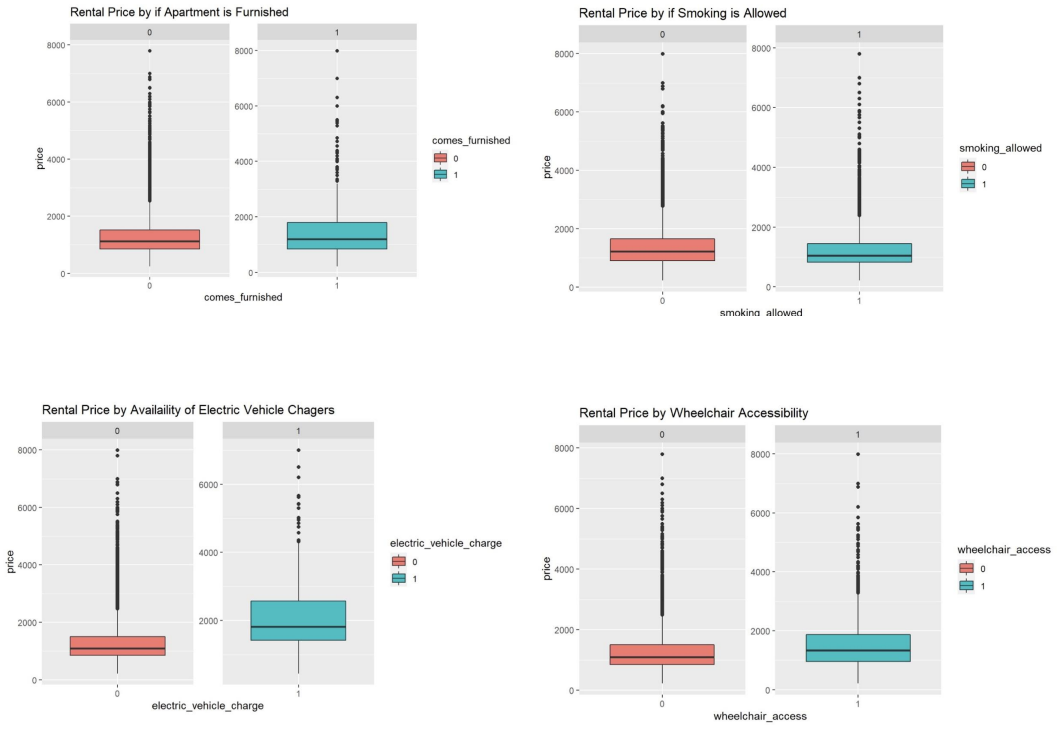


Figure 3.6 Box Plots for Furnishing, Smoking, Electric Vehicle Charger Availability, Wheelchair Accessibility

In order to have a better understanding of the predictor variables, box plots on four variables describing furnishing, smoking, wheelchair access and electric vehicle chargers availability were created. Based on the plots, prices for apartments that are furnished are not significantly different from those that are not furnished. Similarly, smoking does not seem to have a significant impact on apartment rental prices. These two variables are likely not priorities for renters when searching for an apartment, so it is reasonable for them to not have a strong impact on prices. The plot shows no

significant difference on price with or without wheelchair accessibility. Electric cars are a more common option for drivers than in the past, so many apartments are also equipped with chargers. Based on the plot, an apartment with chargers has a higher mean price than those without chargers. This is an interesting result considering that electric cars have not been in the market until relatively recently, but their influence can already be seen in the housing market, which might owe to newly built luxury apartments being equipped with chargers to attract higher-paying residents.

Figure 3.7 Box Plot for Laundry

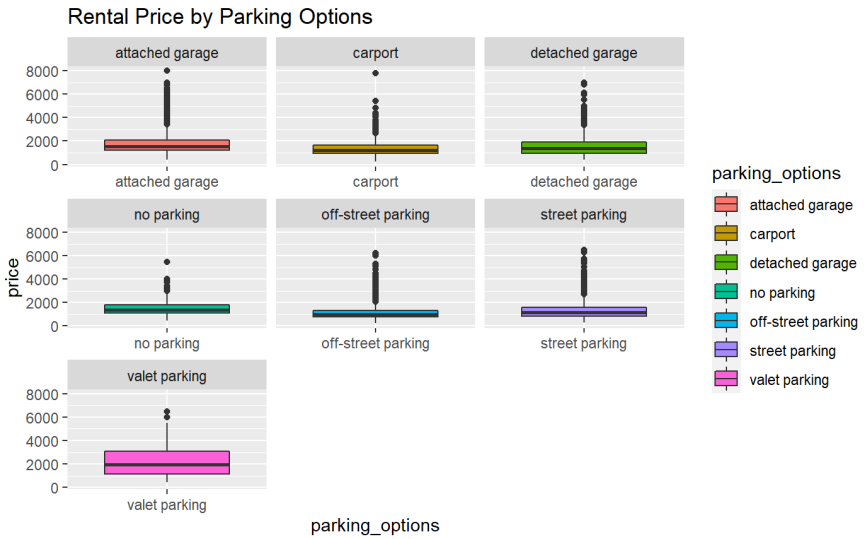


Figure 3.8 Box Plot for Parking

For many apartments, having parking and laundry facilities are expensive amenities which could impact prices. Apartments with laundry facilities inside the unit have higher mean prices compared to those with other laundry options. Overall, apartments that have laundry facilities, whether they are shared or private, can be seen

to have an impact on apartment prices. However, apartments that have washer and dryer hookups are not significantly different from apartments with no laundry facilities. This may be because those with washer and dryer hookups may not be using them. In terms of parking, we can see that, other than valet parking, there is no significant difference between parking options and apartment rental price.

3.2.4 Matrix Plot

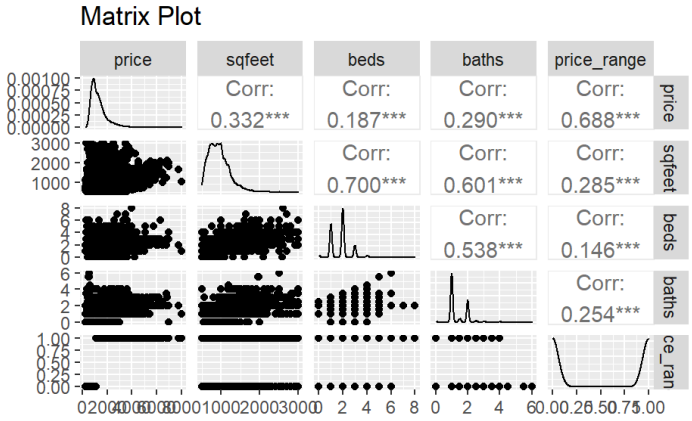


Figure 3.9 Matrix Plot

A matrix plot is useful to provide better understanding of the relationship between existing variables. For the target variable *price*, the scatter plots show positive relationships with *sqfeet*, *beds*, and *baths*, which is reasonable because larger and better equipped apartments typically are associated with a higher price point. Based on the matrix plot, *sqfeet*, *beds*, and *baths* have strong correlations (over 0.5) with one another. Their correlations indicate a potential collinearity problem for further regression analysis.

3.2.5 Scatterplot

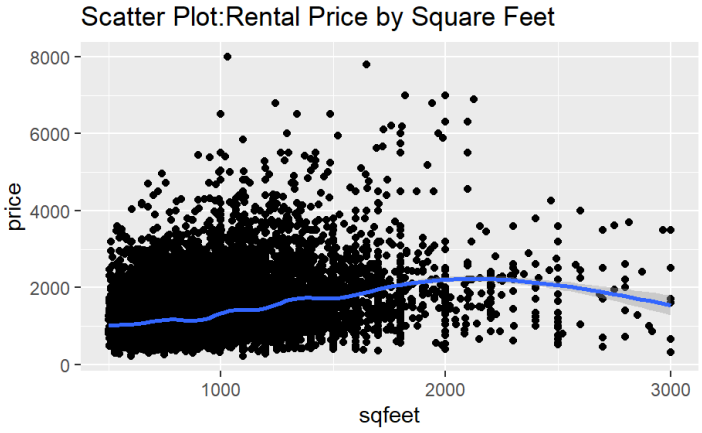


Figure 3.10 Scatterplot

The scatterplot between price and square footage seen in Figure 3.10 demonstrates the relationship between these two variables. As the plot shows, there is a curve in the line across the dots to create a somewhat arched regression line, signaling a curvilinear relationship. This may be reasonable, since after an apartment reaches a certain size, there may be other factors that influence the price that renters are willing to pay.

4. Clustering

Within the rental housing dataset, while shared features can be seen across observations, there are no clear and distinct classes of similar data. To better understand any latent structures within the data, a cluster analysis is performed, which groups the data based on similarities in features. Sorting through all of the data to manually form groups with like characteristics is not practical with over 19,000 observations, so clustering can be helpful in structuring the data. The analysis is done using hierarchical clustering and k-means clustering. The dataset used is normalized to reduce the influence of outliers, as clusters are sensitive to changes in differences of distance. Both methods evaluate the numerical variables available (*price*, *sqfeet*, *beds*, *baths*). Categorical variables were not utilized in the clustering methods discussed in this section.

4.1 Hierarchical Clustering

With hierarchical clustering, a dendrogram showing the ways that the data splits can be cut at various points to form clusters. The tree can be cut at a given height or by specifying a desired number of clusters. In investigating an appropriate number of clusters, dendrograms were created using five different distance methods: “canberra”, “minkowski”, “euclidean”, “manhattan”, and “maximum”. With an exploratory approach of cutting the dendrograms by height, the most balanced clusters were produced using the “canberra” distance method, cut at h=5000. In this case, there are three clusters (k=3).

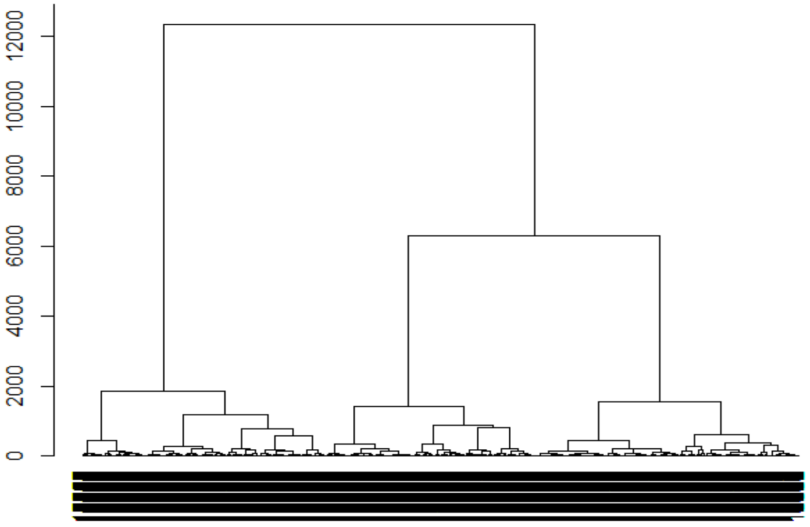


Figure 4.1 Dendrogram using "canberra" distance

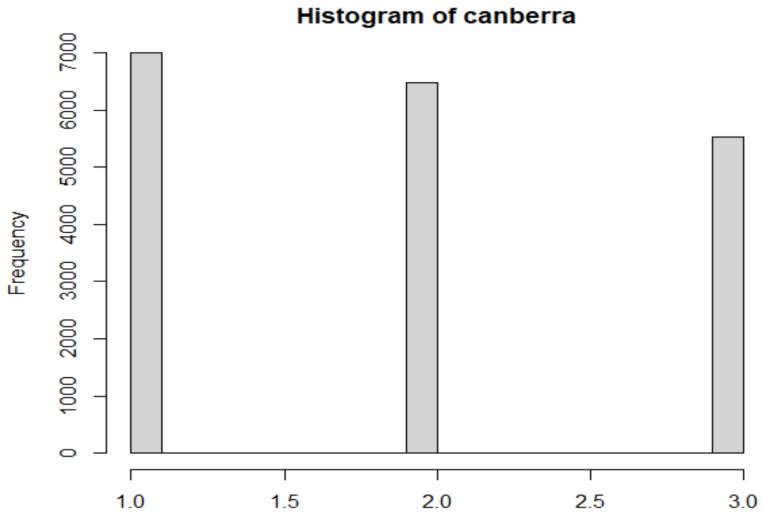


Figure 4.2 Histogram of cluster membership at h=5000

4.2 K-means Clustering

For k-means clustering, reasonable values of k can be determined using an elbow plot. By plotting possible k values against the within-cluster sum of squares, an estimation of k is found. Figure 4.3 shows that as k increases, the within-cluster sum of squares decreases. This sum of squares comes from the Euclidean distance between each data point in a cluster and the cluster centroid. The "bend" of the elbow indicates where an optimal k might lie. It can be seen that the bend is at k=5.

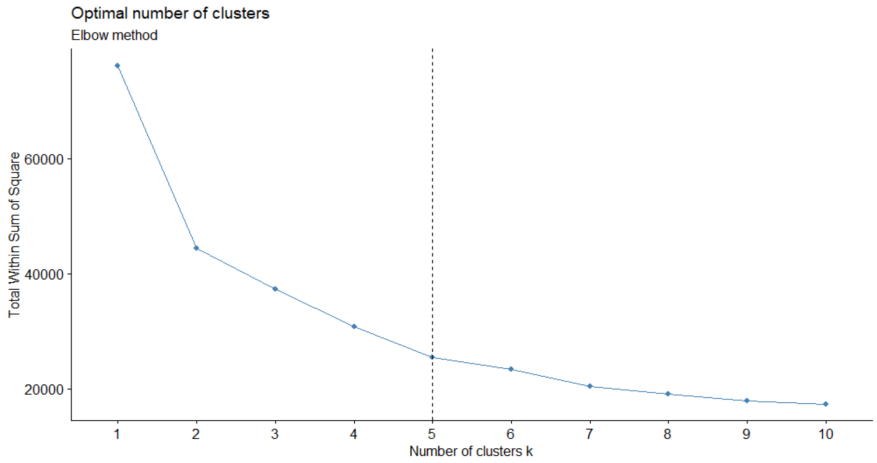


Figure 4.3 Elbow Plot for K-Means

4.3 Cluster Interpretation

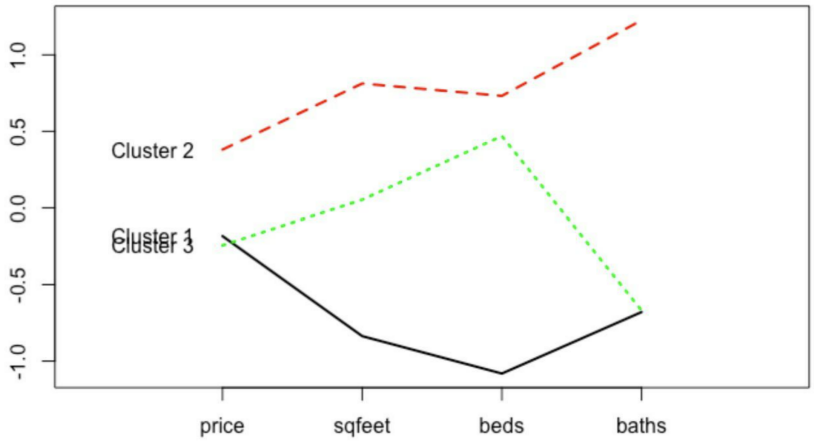


Figure 4.4 Profile Plot of Cluster Centroids using Hierarchical Clustering

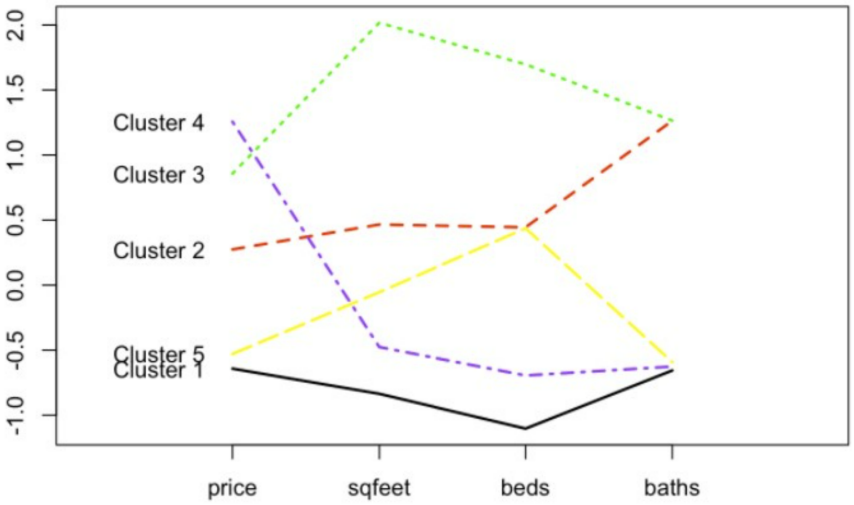


Figure 4.5 Profile Plot of Cluster Centroids using K-Means Clustering

By plotting the centroids against the features used to determine the clusters, the defining characteristics of each cluster becomes clearer.

In the three clusters produced by the hierarchical method seen in Figure 4.4, Cluster 1 appears to contain apartments that are smaller (fewer bedrooms and less square footage) at a similar price point as Cluster 3, which has comparatively more bedrooms and space. Cluster 2 is representative of apartments at a higher price level with more bedrooms, square footage, and bathrooms.

There are five clusters suggested by the k-means method, shown in Figure 4.5. Cluster 1 seems to be the lowest priced, with the fewest bedrooms and least square footage, which may be indicative of less desirable housing within cities (accounting for the lack of space). Cluster 2 is moderately priced, and likely captures the average number of bedrooms and square footage offered, with more bathrooms. Apartments in Cluster 3 have a similar number of bathrooms as those in Cluster 2, though the square footage and number of bedrooms is notably higher, as well as the price level. These may be large, higher-end units. Cluster 4 represents the highest priced apartments with little space and fewer bedrooms, possibly reflecting housing in desirable areas of cities. Cluster 5 is low-priced, and contains apartments that are smaller and with fewer bathrooms than those in Cluster 2, though with a similar number of bedrooms.

The cluster membership assigned by hierarchical clustering (hclust_clusters) and k-means clustering (km_clusters) is added to the dataset to be included in regression and classification as predictor variables.

5. Multiple Regression

5.1 Target Variable Description

Multiple regression is performed to find relevant factors that contribute to apartment rental price, with *price* as the target variable. This numerical variable is chosen because price is an important consideration and reference point when conducting research in the housing rental market. Businesses and individuals alike can find value in the relationship between a rental property's price and its features. In order to keep the variance constant, the target variable, *price*, will be transformed into logged terms in the multiple regression methods in this section.

Firstly, diagnostics will be performed to ensure the model follows the assumptions of the best linear model. Secondly, after the model is checked, three different selection methods will be used to eliminate insignificant factors. Then, the goodness of fit will be measured by the two parameters: adjusted R-squared and root mean square error (RMSE) in each model and will be compared to find the best model. Lastly, higher-order terms and clustering will be applied to the model in the hopes of improving prediction accuracy.

5.2 Diagnostics

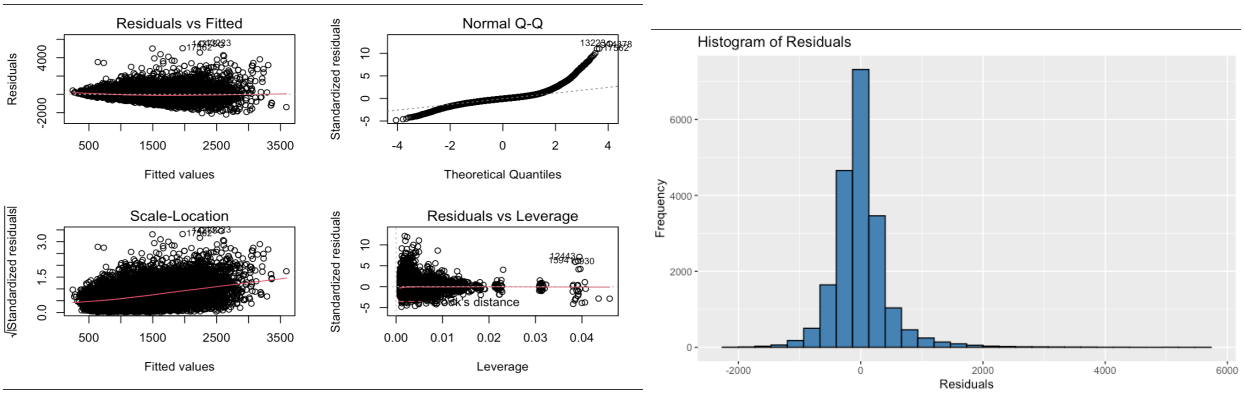


Figure 5.1 Diagnostic Plots before the change

First of all, the residuals vs fitted plot will check the linearity of the model. The red dotted line is approximately horizontal at zero, and there is no clear pattern of spread in the residual plots, except for a few outliers that may cause skewness. Therefore, the linearity assumption is not violated, indicating that the target variable and the explanatory variables are largely in a linear relationship. Next, the normal probability plot shows a mirrored S-shape curve, so the data is heavily tailed, and the histogram of residuals are also right skewed, indicating the normality is violated by some outliers. Thus, the interquartile range method is used to remove these outliers. Furthermore, the scale-location plot checks the assumption of homoskedasticity, with

fitted values of the regression model on the x-axis and the square root of standardized residuals on the y-axis. It can be seen that the variance of the residual points increases gently with the value of the fitted values, thus may present heteroskedasticity problems in the data. In order to reduce heteroskedasticity, the target variable *price* is transformed to logged terms. Lastly, the residuals vs leverage plot identifies the extreme cases that might influence the regression, and from the diagnostic analysis plot, there are some extreme points being labeled with standard residuals above 5. Thus, these extreme values will be removed along with the outliers using the interquartile method introduced earlier.

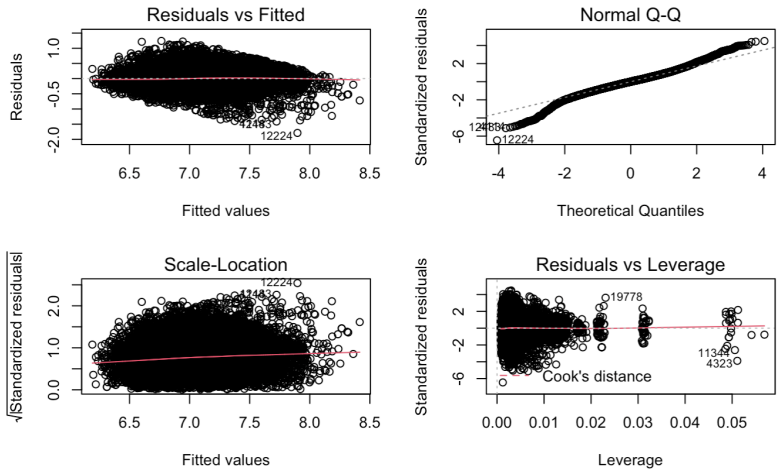


Figure 5.2 Diagnostic Plots after the change

Figure 5.2 above shows the diagnostic plots after removing outliers, along with transforming the target variable to a logged term. 1.075% of the entire dataset are outliers and have been removed. This is reasonable because some of the very high price housing data may not be representative of the local rental market, or possibly be fraudulent listings, that can cause bias in the entire dataset. The removal of the outliers is acceptable because it is a very small percentage of the total dataset, so there is very little impact on the integrity of the data with this omission. The residuals vs fitted plot still follows the rule of linearity assumption, and the normal probability plot of residuals approximately follows a straight line even though there is a small tail when the standardized residuals are low. The scale-location plot shows a straight horizontal line with equally spread points, which indicates homoskedasticity. Lastly, the residuals vs leverage plot still shows several influential values, but the standard deviation is limited to 2, which is better than seen before.

	GVIF	Df	GVIF^(1/(2*Df))
sqfeet	2.541558	1	1.594226
beds	2.249623	1	1.499874
baths	1.989537	1	1.410510
cats_allowed	2.813169	1	1.677250
dogs_allowed	2.961010	1	1.720759
smoking_allowed	1.170596	1	1.081941
wheelchair_access	1.150963	1	1.072829
electric_vehicle_charge	1.096826	1	1.047295
comes_furnished	1.064396	1	1.031695
laundry_options	1.894517	4	1.083147
parking_options	2.083723	6	1.063090
state	3.316391	50	1.012061

Figure 5.3 VIF of variables

Figure 5.3 shows the Variance Inflation Factor (VIF) of the explanatory variables used in the regression model, and the figure suggests that all the variables have a VIF value smaller than 5, which means the variables have low correlation among variables, and the regression results are reliable.

5.3 Train and Validate

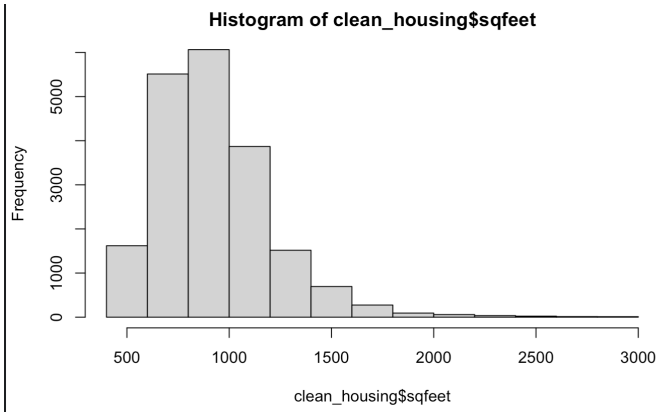


Figure 5.4 sqfeet Before Logging

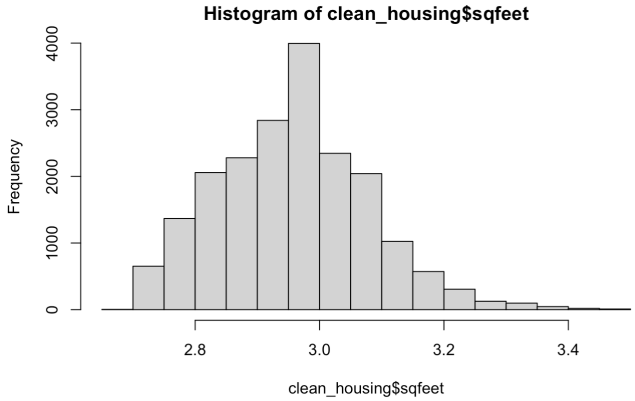


Figure 5.5 sqfeet After Logging

The skewness of the explanatory variables are checked before the data is split into

training and validating sets. Figure 5.4 shows the original distribution of the explanatory variable *sqfeet*, and as it is clearly right skewed, the variable is transformed to logged terms to follow the normal distribution (as Figure 5.5 shows). Then, because the variables *beds* and *baths* are small numeric numbers with ranges from 0 to 6, transformation is not performed.

The training data is 60 percent of the total data, and the validating data is 40 percent of the total. The data is randomly split into two groups, so the representativeness of the data is preserved.

5.4 Model Selection

The multiple regression model used is $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e$, where n is the number of explanatory variables included in the model, β is the coefficient of the corresponding variable, and *e* is the error term.

To create the best model to predict the target variable of *price*, four methods were tried: forward selection, backward elimination, bidirectional search, and the addition of a polynomial term.

Forward selection incorporates one predictor at a time, with each addition leading to the most improvement of the criterion. The added variable will be the most statistically significant one, and further additions will stop once there are no more significant variables. The model using forward selection excludes the variable *wheelchair_access* and includes the rest of the explanatory variables.

```
Step:  AIC=-20482.74
price ~ state + sqfeet + laundry_options + parking_options +
    cats_allowed + smoking_allowed + baths + electric_vehicle_charge +
    dogs_allowed + beds + comes_furnished
```

Figure 5.6 Forward Selection

Backward elimination is essentially opposite to forward selection - it starts with all the variables in the model and removes predictors for the most improvement of the criterion. The removed variables are the least statistically significant ones. The model using backward elimination excludes the variable *wheelchair_access*.

```
Step:  AIC=-30338.15
price ~ sqfeet + beds + baths + cats_allowed + dogs_allowed +
    smoking_allowed + electric_vehicle_charge + comes_furnished +
    laundry_options + parking_options + state
```

Figure 5.7 Backward Elimination

Bidirectional search is a method that chooses between forward and backward that leads to the most improvement of the criterion. The result of bidirectional search includes all the explanatory variables except for *wheelchair_access*.

```
Step:  AIC=-30338.15
price ~ sqfeet + beds + baths + cats_allowed + dogs_allowed +
    smoking_allowed + electric_vehicle_charge + comes_furnished +
    laundry_options + parking_options + state
```

Figure 5.8 Bidirectional Search

In the exploratory analysis, a curve can be seen in the relationship between price and square feet (Figure 3.10). In order to check if the model will perform better when this perceived nonlinearity is accounted for, a polynomial term of the explanatory variable, *sqfeet*, is added to the model. The variable is squared in an attempt to capture the prominent curve in the regression line. However, because of the smaller bends seen in the line, the squared *sqfeet* term may not sufficiently explain all of the nonlinearity.

5.5 Best Model

The best model is determined by two features: adjusted R-squared and root mean square error (RMSE). Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The R-squared is a measure of goodness of fit, how close the estimated values are to the actual values. Therefore, the adjusted R-squared increases when the predictors improve the model more than expected. The higher the adjusted R-squared, the better the fit of the model to the data. RMSE is the standard deviation of the residuals, meaning how spread out the residuals are from the fitted values. A lower RMSE indicates a better fit of the model.

	AdjR2 <dbl>	RMSE <dbl>
Forward Selection	0.5773326	0.2717694
Backward Elimination	0.5677354	0.2739760
Bidirectional Search	0.5677354	0.2739760
Polynomial Regression	0.5677511	0.2739954

Figure 5.9 Adjusted R-squared and RMSE

Figure 5.9 shows the adjusted R-squared values and RMSE values of four models, and the result indicates the model selected by forward selection is the best model, which has the highest adjusted R-squared value and the lowest RMSE among all other models.

5.5.1 Best Model with Clusters

	GVIF	Df	GVIF^(1/(2*Df))
sqfeet	2.728492	1	1.651815
cats_allowed	2.799024	1	1.673028
dogs_allowed	2.947930	1	1.716954
smoking_allowed	1.150077	1	1.072417
electric_vehicle_charge	1.064308	1	1.031653
comes_furnished	1.057215	1	1.028210
laundry_options	1.982141	4	1.089286
parking_options	2.081506	6	1.062996
state	4.125763	50	1.014273
km_clusters	4.425905	4	1.204343

Figure 5.10 VIF of variables with K-Means Clusters

	GVIF	Df	GVIF^(1/(2*Df))
sqfeet	2.220790	1	1.490232
cats_allowed	2.795855	1	1.672081
dogs_allowed	2.945863	1	1.716352
smoking_allowed	1.147402	1	1.071169
electric_vehicle_charge	1.060743	1	1.029924
comes_furnished	1.057254	1	1.028229
laundry_options	1.905764	4	1.083948
parking_options	2.025161	6	1.060567
state	3.296491	50	1.012000
hclust_clusters	2.489458	2	1.256106

Figure 5.11 VIF of variables with Hierarchical Clusters

As previously mentioned in section 4, variables concerning cluster membership, derived from hierarchical clustering and k-means clustering, are added to the best model of multiple regression (forward selection). However, as the VIF value is above 10, numeric variable *baths* and *beds* have been dropped to avoid multicollinearity problems, and as the two clusters together will cause multicollinearity, the cluster variables will be added once at a time, and the accuracy will be compared. The adjusted R-squared value and RMSE are measured to detect if the addition of clusters will improve the best model.

	AdjR2 <dbl>	RMSE <dbl>
Best Model	0.5773326	0.2717694
Best Model with Kmeans Cluster	0.6898821	0.2308018
Best Model with Hierarchical Clusters	0.5700037	0.2728948

Figure 5.12 Adjusted R-squared and RMSE for cluster

Figure 5.12 shows the adjusted R-squared values and RMSE values for the selected best model and the best model with the addition of clusters. The figure suggests that the adjusted R squared value increases from 0.5773 to 0.6899, approximately 19.5%, and the RMSE value decreases by 15%, from 0.2718 to 0.2308 after the addition of k-means clusters. However, the adjusted R squared value decreases by 1.26%, and the RMSE value increases by approximately 0.4% after adding the hierarchical clusters. Therefore, the addition of k-means clusters will improve the accuracy of the prediction,

but, on the other hand, the addition of hierarchical clusters will decrease the accuracy. This suggests that there are other predictors that can explain the target variable better than the hierarchical clusters. Hierarchical clustering tends to create fewer, larger clusters while k-means tends to create more clusters that are smaller. Because of this, it is possible that the k-means clusters have more similarities or fewer differences than the hierarchical clusters.

5.5.2 Interpretation

The best model regresses the dependent variable, *price*, on *sqfeet*, *beds*, *baths*, *cats_allowed*, *dogs_allowed*, *smoking_allowed*, *electric_vehicle_charge*, *comes_furnished*, *parking_options*, *laundry_options*, and *state*. As the variables *cats_allowed*, *dogs_allowed*, *smoking_allowed*, *electric_vehicle_charge*, *comes_furnished* are binary categorical variables, and *parking_options*, *laundry_options*, and *state* are nominal categorical variables, the model will treat these variables as dummy variables, and a benchmark model is defined to avoid multicollinearity. Additionally, since the *km_clusters* variable improves the accuracy of the model in terms of adjusted R-squared and RMSE, it will be added as a predictor in the best model, and the coefficients will be compared.

The variable, *comes_furnished*, was found to be not statistically significant after the forward selection, so this variable will be excluded during the interpretation.

The base model considers the state Alaska as reference, with the apartment having laundry facilities in the building and an attached garage, which does not allow cats, dogs, and smoking, and does not have electric vehicle chargers. The interpretation of the benchmark model is that the apartment rental price will decrease by 1.38%, as the price is in logged terms, with a one unit increase in the number of bedrooms, holding other variables constant. When the area of the apartment in square feet increases by 1%, as variable *sqfeet* is in logged term, the price will approximately increase by 0.95%, holding other variables constant. If smoking is then allowed in the base model, the price will decrease by 5.42%, and if cats and dogs are allowed in the apartment, the price will increase by 6.98% and 3.22%, respectively. Moreover, if one apartment contains electric vehicle chargers, the price will increase by 8.27%. Different laundry options and parking options will vary the price as well, for instance, if there is in-unit laundry, the price will be 10.25% more than a laundry option contained in the building. Off-street parking will be 20.98% less expensive than an apartment with an attached garage. Additionally, apartment prices will be distinctive based on different states. For example, the price will be 46.2% more in California than in Alaska, and the price will be 46.96% less in Missouri than in Alaska.

Two numeric variables, *baths* and *beds*, are removed to ensure the

multicollinearity problem is avoided after the addition of clusters. After adding the k-means cluster variable to the model, the price will increase by only 0.40% with a 1% increase in the square footage, which is 57.9% less compared to the model without the addition of clusters, holding other variables constant. The price will decrease by 3.83% if smoking is allowed, 29.3% less than the model without clusters. If dogs and cats are allowed, the price will increase by 1.05% and 4.84%, respectively, which are 67.4% and 30.7% less than without clusters. A conclusion can be made that the addition of k-means clusters will decrease the coefficients in the best model, and increase the prediction accuracy.

6. Classification

6.1 Target Variable Description

The target variable, *price*, has been segmented into two categories: High Price and Low Price based on the median of *price* to create a balanced dataset. Since *price* will be segmented into a binary variable, there is no need to do normalization. By separating into two categories, the classification model is able to analyze how different price ranges react to predictor variables and show factors that influence price. This binary categorical variable will provide further insight about how an apartment compares to the national median, and will have simpler interpretability for businesses and individuals than the numerical variable, *price*, used in the multiple regression model. The variable, *states*, is eliminated for classification, due to the fact that there are 51 states (including District of Columbia) which makes the interpretation of the logistic regression very difficult. It is better to group the states into a smaller categorical variable and run regression on this variable.

6.2 Examining Variables

Based on the scatterplot (Figure 3.10), there is a curvilinear relationship between price and square feet. In order to create a model that fits the best for the data, the *sqfeet* variable has been transformed into a polynomial term which is the squared term of *sqfeet*. Moreover, as the descriptive analysis revealed, the continuous variables price and square feet are not normally distributed. Before using them to train the logistic model, *price* and *sqfeet* are each transformed into logged terms so that the data is normally distributed.

```
Call:
glm(formula = price_range ~ . - price - sqfeet + I(sqfeet^2),
    family = binomial(link = "logit"), data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8761  -0.9910  -0.5986   1.0331   2.3213

Coefficients:
                          Estimate Std. Error z value             Pr(>|z|)
(Intercept)             -8.627767   0.360787 -23.914 < 0.0000000000000002 ***
beds                    -0.383703   0.041942  -9.148 < 0.0000000000000002 ***
baths                    0.208860   0.053249   3.922  0.00008770928576293 ***
cats_allowed             0.418167   0.075865   5.512  0.00000003548040279 ***
dogs_allowed            -0.133804   0.072461  -1.847               0.0648 .
smoking_allowed         -0.504084   0.042864 -11.760 < 0.0000000000000002 ***
wheelchair_access       -0.060256   0.074729  -0.806               0.4201
electric_vehicle_charge  1.440114   0.184785   7.793  0.00000000000000652 ***
comes_furnished         -0.058073   0.098739  -0.588               0.5564
laundry_options          0.312011   0.017828  17.501 < 0.0000000000000002 ***
parking_options         -0.122840   0.016682  -7.364  0.00000000000017913 ***
I(sqfeet^2)              0.181687   0.009079  20.011 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15758  on 11403  degrees of freedom
Residual deviance: 13899  on 11392  degrees of freedom
AIC: 13923
```

Figure 6.1 Initial Model Result

The result of the logistic regression shows that most predictor variables are statistically significant with a p-value of less than 0.05, but the variables *dog_allowed*, *wheelchair_access*, and *comes_furnished* are not significant.

```
Number of Fisher Scoring iterations: 4
```

beds	baths	cats_allowed	smoking_allowed
2.192882	1.777135	1.098403	1.037765
electric_vehicle_charge	**laundry_options**	**parking_options**	**I(sqfeet^2)**
1.017620	1.208791	1.016777	2.442576

Figure 6.2 Initial Variables VIF Score

A VIF check was performed on the predictor variables which all returned values less than 5. The very low VIF scores means predictor variables are not highly collinear with the other variables which may further prove the accuracy of the model.

6.3 Model Perfection with Cluster Variables

```
Call:
glm(formula = price_range ~ . - price - dogs_allowed - comes_furnished -
    sqfeet - wheelchair_access + I(sqfeet^2), family = binomial(link = "logit"),
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1064  -0.9398  -0.5410   0.9839   2.4954

Coefficients:
                           Estimate Std. Error z value             Pr(>|z|)
(Intercept)              -8.909245   0.377730 -23.586 < 0.0000000000000002 ***
beds                     -0.236970   0.058381  -4.059      0.0000492857726 ***
baths                     0.646870   0.065267   9.911 < 0.0000000000000002 ***
cats_allowed              0.318308   0.049319   6.454      0.0000000001089 ***
smoking_allowed          -0.476804   0.043331 -11.004 < 0.0000000000000002 ***
electric_vehicle_charge   1.116271   0.189328   5.896      0.0000000037250 ***
laundry_options           0.282510   0.018093  15.614 < 0.0000000000000002 ***
parking_options          -0.112269   0.017034  -6.591      0.0000000000438 ***
hclust_clusters          -0.904002   0.049770 -18.163 < 0.0000000000000002 ***
km_clusters               0.493441   0.022210  22.217 < 0.0000000000000002 ***
I(sqfeet^2)               0.176445   0.009449  18.674 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15758  on 11403  degrees of freedom
Residual deviance: 13328  on 11393  degrees of freedom
AIC: 13350

Number of Fisher Scoring iterations: 4
```

Figure 6.3 Cluster Model with Outliers

After introducing the two cluster variables (km_clusters, hclust_clusters), the best logistic model has been regressed again to observe potential changes on the significance of variables. Based on the figure above, all the variables are statistically significant with p-values under 0.05.

```
                  beds                 baths           cats_allowed
smoking_allowed
              4.019189              2.499492              1.098472
1.040473
electric_vehicle_charge       laundry_options       parking_options
hclust_clusters
              1.026997              1.220684              1.018059
3.548508
           km_clusters          I(sqfeet^2)
              2.626582              2.461126
```

Figure 6.4 VIF Score with Outliers

The VIF scores of the variables in the model are less than 5 for all the variables, which means that after introducing the cluster variables, all the predictor variables are still not collinear with the other variables. When deciding which cluster variables to include, it was found that removing either predictors does not significantly change the VIF score, so both cluster variables were added to the classification models.

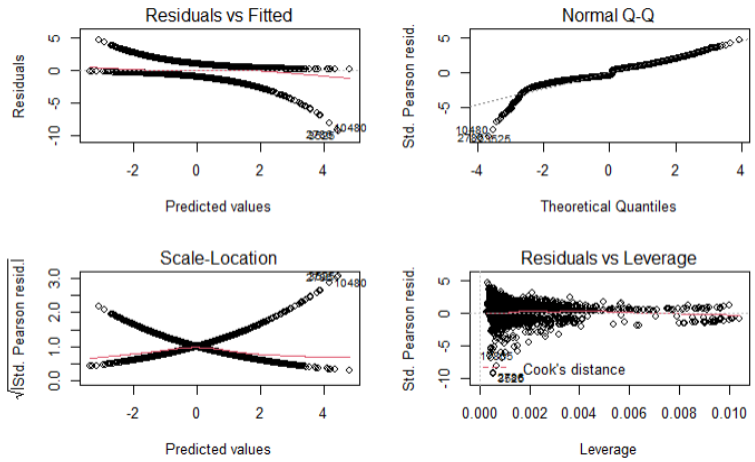


Figure 6.5 Q-Qplot with Outliers

Based on the Q-Q plot above, the overall shape of the plot is close to a line, however, there is a section of the left tail that pulls away from the straight line, which means there are outliers in the dataset that need to be removed.

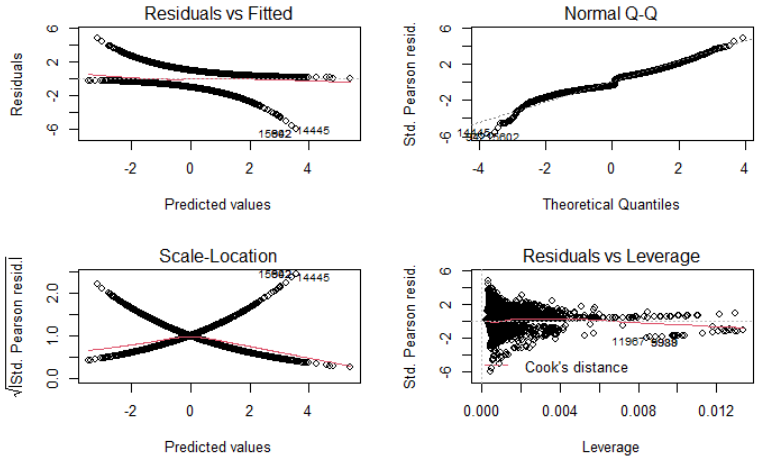


Figure 6.6 Q-Qplot without Outliers

After removing outliers which make up 0.14% of the entire dataset, the Q-Q plot is close enough to be a straight line with only few outliers left. By removing influential outliers, the model aligns with the assumption that no significant outliers causes bias. This amount of outlier omission is acceptable because it is very small when compared to the entire dataset.

```
Call:
glm(formula = price_range ~ . - price - dogs_allowed - comes_furnished -
    sqfeet - wheelchair_access + I(sqfeet^2), family = binomial(link = "logit"),
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6846  -0.9323  -0.5296   0.9745   2.5218

Coefficients:
                          Estimate Std. Error z value             Pr(>|z|)
(Intercept)             -9.247922   0.382645 -24.168 < 0.0000000000000002 ***
beds                    -0.233142   0.058913  -3.957    0.000075777912568 ***
baths                    0.682329   0.065796  10.370 < 0.0000000000000002 ***
cats_allowed             0.313165   0.049636   6.309    0.000000000280467 ***
smoking_allowed         -0.488212   0.043646 -11.186 < 0.0000000000000002 ***
electric_vehicle_charge  1.558481   0.218119   7.145    0.000000000000899 ***
laundry_options          0.280150   0.018199  15.394 < 0.0000000000000002 ***
parking_options         -0.115168   0.017158  -6.712    0.000000000019180 ***
hclust_clusters         -0.947060   0.050332 -18.816 < 0.0000000000000002 ***
km_clusters              0.510813   0.022452  22.752 < 0.0000000000000002 ***
I(sqfeet^2)              0.183922   0.009566  19.226 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15727  on 11379  degrees of freedom
Residual deviance: 13173  on 11369  degrees of freedom
AIC: 13195

Number of Fisher Scoring iterations: 4
```

Figure 6.7 Cluster Model without Outliers

Regression results for the cluster model without outliers shows that all the variables are still statistically significant with p-values less than 0.05. *beds*, *baths*, *smoking_allowed* and *hclust_clusters* are negative in the regression model, with the rest of the variables being positive.

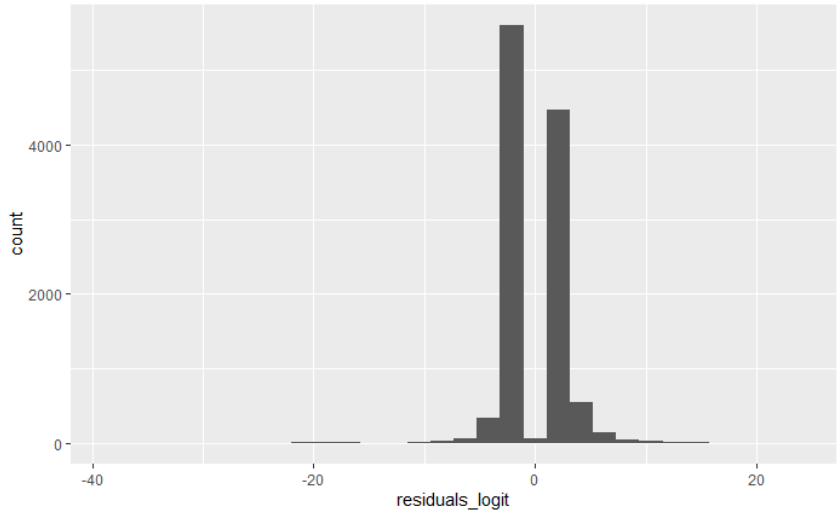


Figure 6.8 Residuals Plot with Cluster Variables

The residuals plot for logistic regression with cluster variables is very similar to the residual plot for the logistic model without the cluster variables. Because it is very possible for the residuals to not contain values of 0, there is no concern. Overall, the residuals plot is very close to normal distribution which means the assumptions for the model during inference are valid and reasonable.

6.4 Interpretation of Odds with Cluster Variables

	odds <dbl>	odds_1 <dbl>
(Intercept)	0.00009631155	-0.9999037
beds	0.79204128416	-0.2079587
baths	1.97848027008	0.9784803
cats_allowed	1.36774666726	0.3677467
smoking_allowed	0.61372276855	-0.3862772
electric_vehicle_charge	4.75160030845	3.7516003
laundry_options	1.32332773103	0.3233277
parking_options	0.89121663220	-0.1087834
hclust_clusters	0.38787965117	-0.6121203
km_clusters	1.66664635736	0.6666464

	odds <dbl>	odds_1 <dbl>
I(sqfeet^2)	1.20192163872	0.2019216

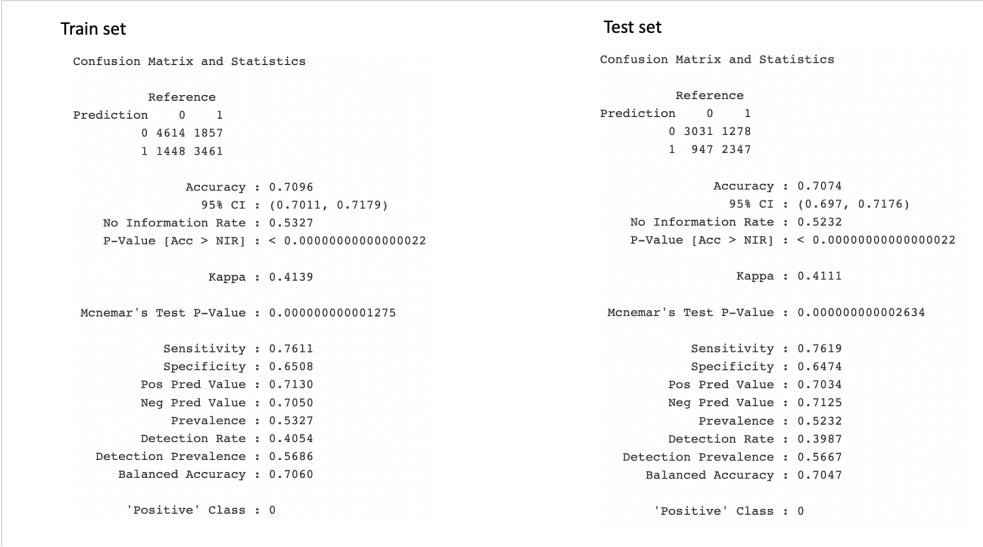
Figure 6.9 Odds with Cluster Variables Results

After calculating the odds for regression, interpretation is needed to analyze how the explanatory variables impact the possibility of getting a high or low price range. The difference (odds - 1) was obtained for each variable to produce the column 'odds_1'. The variables *beds*, *smoking_allowed*, *parking_options, and hclust_clusters* have odds that are less than 1. The variables *baths*, *cats_allowed*, *electric_vehicle_charge*, *laundry_options*, squared *sqfeet, and km_clusters* have odds that are greater than 1. Odds that are greater than 1 indicate a positive effect on the reference case while odds that are less than 1 indicate a negative effect on the reference case. For *beds*, the reference case is an apartment in the high price range with 0 bedrooms (studio apartment). Keeping other variables constant, an increase of 1 unit in *beds* would cause a 20.8% decrease of the probability of the apartment being in the high price range. For *baths*, the reference case is an apartment in the high price range with 0 bathrooms. Keeping other variables constant, an increase of 1 unit in *baths* would cause a 97.8% increase of the probability of a high price apartment. For *cats_allowed*, the reference case is an apartment in the high price range that does not allow cats. Keeping other variables constant, allowing cats in apartments would cause a 36.8% increase of the probability of a high price apartment. For *smoking_allowed*, the reference case is an apartment in the high price range that does not allow smoking. Keeping other variables constant, allowing smoking in apartments would cause a 38.6% decrease of the probability of a high price apartment. For *electric_vehicle_charge*, the reference case is an apartment in the high price range without electric vehicle chargers. Keeping other variables constant, electric vehicle chargers being available would cause a 375% increase of the probability of a high price apartment. It's unusual to have a 375% increase, but there is a plausible explanation for this situation. Electric vehicles were not popular until recently, however, many apartments with electric vehicle chargers may be newly built luxury buildings aiming to attract renters that will pay more. Additionally, because of the cost of electric cars, those with means to own such vehicles may inherently be able

to spend more on housing. For laundry options, the reference case is an apartment in the high price range with no laundry. Keeping other variables constant, any other laundry option would cause a 32.3% increase of the probability of a high price apartment. For parking options, the reference case is an apartment in the high price range with no parking. Keeping other variables constant, any other parking option would cause a 10.9% decrease of the probability of a high price apartment. This might be due to the tendency for high priced apartments to be in large cities, where residents may be less likely to have cars because of easy access to other forms of transportation. For the squared *sqfeet* term, keeping other variables constant, an increase in 1 unit in squared *sqfeet* would cause a 20.19% increase of the probability of a high price apartment.

For the k-means cluster, the reference case is Cluster 1. Keeping other variables constant, any other laundry option would cause a 66.7% increase of the probability of a high price apartment. For *hclust_clusters*, the reference case is Cluster 1. Keeping other variables constant, any other cluster would cause a 61.2% decrease of the probability of a high price apartment.

6.5 Interpretation of Confusion Matrix with Cluster Variables



Train set

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4614 1857
         1 1448 3461

               Accuracy : 0.7096
                 95% CI : (0.7011, 0.7179)
    No Information Rate : 0.5327
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4139

 Mcnemar's Test P-Value : 0.000000000001275

            Sensitivity : 0.7611
            Specificity : 0.6508
         Pos Pred Value : 0.7130
         Neg Pred Value : 0.7050
             Prevalence : 0.5327
         Detection Rate : 0.4054
   Detection Prevalence : 0.5686
      Balanced Accuracy : 0.7060

       'Positive' Class : 0
```

Test set

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3031 1278
         1  947 2347

               Accuracy : 0.7074
                 95% CI : (0.697, 0.7176)
    No Information Rate : 0.5232
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4111

 Mcnemar's Test P-Value : 0.000000000002634

            Sensitivity : 0.7619
            Specificity : 0.6474
         Pos Pred Value : 0.7034
         Neg Pred Value : 0.7125
             Prevalence : 0.5232
         Detection Rate : 0.3987
   Detection Prevalence : 0.5667
      Balanced Accuracy : 0.7047

       'Positive' Class : 0
```

Figure 6.10 Confusion Matrix for Logistic Regression with Cluster Variables

Based on the confusion matrix, the prediction accuracy for the test set is 0.7074 with a sensitivity of 0.7619 and specificity of 0.6474. The sensitivity measures how correctly a classifier can predict members in the data, which means logistic regression can correctly predict 76.19% of the class. The specificity measures how correctly a classifier can rule out the wrong members, which shows that the model can detect 64.74% of the false values. Overall accuracy of the model is 70.74%, which is reasonable, but there is still potential to improve in the future. The accuracy of the test set is very

similar to the training set. Both the precisions and the recalls of the training set and the test set are also very similar. The logistic model reasonably fits both the training set and the test set, indicating the model does not have overfitting problems.

6.6 Comparison/Contrast between Logistic Regression and K-NN With Cluster Variables

```
k-Nearest Neighbors

11380 samples
   12 predictor
    2 classes: 'X0', 'X1'

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 10242, 10242, 10242, 10243, 10241, 10241, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.7643872  0.7517863  0.6496787
   7  0.7663785  0.7510731  0.6464207
   9  0.7687447  0.7526143  0.6408441
  11  0.7723037  0.7517875  0.6430390
  13  0.7718702  0.7537686  0.6383390
  15  0.7715190  0.7530530  0.6392162
  17  0.7723666  0.7564618  0.6409704
  19  0.7718073  0.7529417  0.6386510
  21  0.7723578  0.7551946  0.6410344
  23  0.7728696  0.7542043  0.6429756

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 23.
Confusion Matrix and Statistics
```

Figure 6.11 K-NN Result With Cluster Variables

The K-Nearest Neighbors classification is performed by using 12 predictor variables (state excluded) to target the variable *price_range*. The dataset has been cleaned to omit outliers and could be used to generate unbiased results. Based on accuracy, the best neighbor for the model is K=23 where ROC has the largest value. ROC measures the performance of a model by plotting the pairs of sensitivity and specificity, so the largest value of ROC means that K=23 when sensitivity and specificity are also at the highest values for the model. For this model, when 23 neighbors are used, K-NN can provide the best classification result.

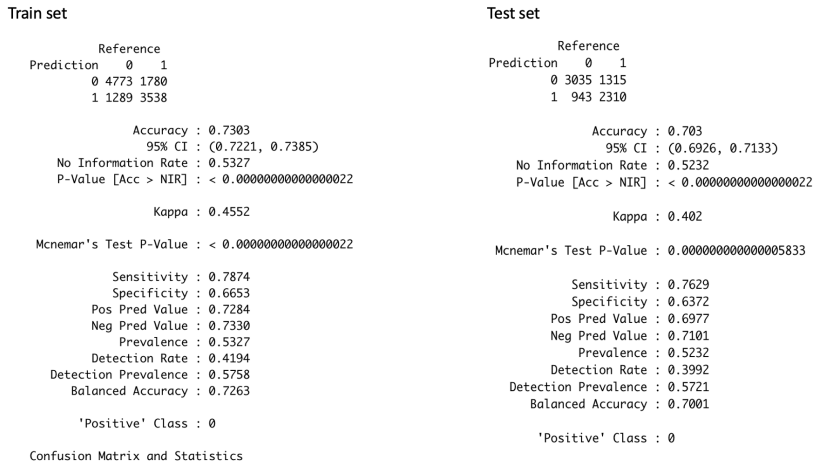


Figure 6.12 Confusion Matrix For K-NN with Cluster Variables

The confusion matrix of K-NN is overall better than the confusion matrix for logistic regression. For K-NN the accuracy of the test set is 70.3%, which is 0.33% lower than the accuracy of logistic regression. The sensitivity for K-NN is 76.3%, which is 0.1% higher than the result of logistic regression. The specificity for K-NN is 63.72%, which is 1.02% lower than the specificity of logistic regression. The accuracies, the sensitivities, and the specificities for the train set and the test set are very similar. Because the value of k is not very small, the K-NN model avoids overfitting problems.

Although the performance of these two kinds of classification methods is really similar, the K-NN method has advantages in the training model for the rental housing dataset. As the descriptive analysis mentioned above, some data are not normally distributed, which requires data transformation. However, compared with the logistic method, the K-NN method does not assume the underlying data distribution pattern. Even though the K-NN method slightly underperforms compared to the logistic regression method, it is still a better fit for the dataset.

F1_Logstic	F1_KNN
0.7000441	0.6944511

Figure 6.13 F1 Score with Cluster Variables

Both the logistic model and the K-NN model have very close precision and recall, so the F1 scores are also very close. The F1 score is the weighted average of precision and recall. F1 reaches its best value at 1 and its worst value at 0. Based on the result above, both models perform as well for the dataset. However, the interesting thing here is that the F1 score of the K-NN model is slightly lower than the F1 score of the logistic model. This is because the K-NN method struggles with larger datasets. The dataset used for analysis has approximately 20,000 data points and 14 features with the inclusion of the

cluster variables, making this a high-dimensional large dataset. In high-dimensional data, it is possible for points that are very similar to have very large distances. Therefore, the K-NN model's performance would not perform better than the logistic model in a large high-dimensional dataset.

7. Summary

7.1 Comparing Multiple Regression and Classification Models

The best multiple regression model is selected based on the forward selection method, and explanatory variables, *wheelchair_access* and *comes_furnished*, are excluded for being statistically insignificant. The target variable, *price*, is logged for the constant variance assumption, and one explanatory variable, *sqfeet*, is also logged to ensure normal distribution. Based on the result of the final model, apartments that allow cats and dogs tend to have higher prices in percentage terms than those that do not, and apartments that allow smoking have a relatively lower price in percentage terms than those that prohibit smoking. More bathrooms and larger square footage indicate higher prices, but increasing the number of bedrooms seems to decrease the price in percentage terms. Different laundry options may influence the price, where the price is approximately the lowest when there is no laundry option on site, and the price is the highest when the laundry option is contained in the unit, and apartments with valet parking options tend to have higher prices than apartments with an attached garage. Various states may have different markets and standards on apartment rental pricing, where the baseline model is referred to Alaska, and apartments in California may have higher price in percentage terms than in Alaska. Since all the predictors are selected and significant, these factors may be considered as important contributors to the price for an apartment.

In the classification model, the target variable is price range, and the level is defined as low if the price is below the median price and high if the price is above the median. The variable, *state*, is not included in the model because there are too many categories for logistic regression to currently handle. Several explanatory variables, *dogs_allowed*, *wheelchair_access*, and *comes_furnished* are not statistically significant, and thus eliminated from the model. Based on the result of the classification model, it suggests that more bedrooms will decrease the probability of an apartment in the high price range, whereas more bathrooms will increase the probability of apartments in the high price range. Allowing cats in the apartment will increase the probability of apartments being in the group of high price range, but allowing smoking will then decrease the probability. The presence of electric vehicle chargers will

significantly increase the probability for apartments to be in the high price range, likely because electric vehicles are relatively new, so housing that has this amenity will tend to be newer, possibly luxury, buildings. A change in the availability of laundry options from no laundry options will increase the probability of apartments being in a high price range, but a change in the parking options will decrease the probability. Lastly, a one unit increase in squared *sqfeet* will also increase the probability to be in the high price range.

The multiple regression model and the classification model both suggest that an increase in *cats_allowed, sqfeet, baths, electric_vehicle_charge* will increase the price and probability of apartments in high price range and will decrease if there are increases in *beds* and *smoking_allowed*. For variables *laundry_options* and *parking_options*, the classification model has transformed these two variables to categorical variables, while the multiple regression model creates dummy variables of each type of option, which can provide a much clearer picture on prices for each feature.

Moreover, this report introduces a method that aims to improve the prediction accuracy of the model by clustering the data using hierarchical and k-means algorithms and adding cluster membership as predictor variables in the best classification and regression models. Unlike the original supervised learning method which focuses on finding the relationship between the predictors and the target variables, clustering benefits from the similarities between the instances to improve the prediction accuracy. Two models are used with the proposed method, and the results show a significant improvement from the point view of adjusted R-squared and RMSE in the regression model, with a 19.5% increase in the adjusted R-square and 15% decrease in RMSE. For classification, clustering helps improve the performance of the logistic model, which increases F1 score from 0.67 to 0.70, by 4.5%. However, the performance of the K-NN model declines as compared to before clustering, because the K-NN model performs better on smaller and less complex datasets. Therefore, dimension reduction techniques would be helpful in the analysis.

The insights based on the results of the regression model can help benefit potential developers looking to build new units that meet the demand of the market, property owners or managers considering adding amenities to attract tenants, and realtors understand pricing on a local level, as the regression model gives specific details in different options. The classification model may help renters who are seeking for new housing and people who are conducting reports on housing markets, as the classification model gives a clearer comparison between the options.

7.2 Reflections & Further Study

In this project, we built off of our work and learnings from the previous report.

The main point of divergence from earlier work was the addition of clusters in our models. During the clustering process, we were able to utilize the hierarchical and k-means methods to create the clusters, and good judgment was necessary to figure out reasonable values for the number of clusters. The rule of thumb used was to create clusters that were as balanced as possible, though there was no way to know if there was a "right way" to choose clusters. In the cluster interpretation (section 4.3), the profile plots of cluster centroids were a good way to show the differences between clusters, but identifying patterns in the features was not very clear or straightforward. We were able to "check" our clusters with the knowledge that hierarchical is better at identifying large clusters, and k-means is better at identifying small clusters, which was supported by the values of k we had.

Additionally, the methods used to divide the data into clusters used the distances from only the numerical variables. This means that the binary/categorical variables were not considered in our cluster analysis. We considered incorporating categorical features that contained multiple options by relabeling each option with numeric values, though ultimately decided against it, since the numbers would be used in finding the distances. There would be no reason why parking option 1 would be further away from parking option 7 than any of the other options. In future cluster analyses, it may be beneficial to experiment with alternative clustering methods, such as k-modes or k-prototype, that would work on the categorical data.

Similar to Report 2a, we remained cognizant of overfitting in our models, since it would not serve us to have a model that only performed well on the data it was trained with. To prevent this issue, we made sure performance was good on the testing data for all our best models.

To see if the performance of our best models would improve with the presence of clusters, we included the variables created by both clustering methods. In the classification models, including both of the cluster variables (*hclust_clusters* and *km_clusters*) seemed to work - the VIF check returned acceptable values of less than 5 for all variables. Removing one of the cluster variables did not change the VIF scores by very much, so we kept them both. For the multiple regression model, including both cluster variables caused some multicollinearity, necessitating the removal of one, as described in section 5.5.1. For further accuracy improvement, it may help to drop the *comes_furnished* variable as it is not statistically significant.

In interpreting the results of our classification models, there was some inconsistency from the inclusion of the cluster variables in the logistic regression. Compared to the results from Report 2a, where the increase of one bathroom in a property increased the probability of an apartment being high-priced by 21.7%, in this report, a one-unit increase in baths resulted in a 97.8% (near-certainty) increase in

probability for an apartment to be high-priced. This might be because of differences in cluster attributes, but other factors may also be contributing to this jump.

Finally, after adding the cluster variables, we found that the F1 for the K-NN model had decreased compared to its performance without the clusters. We suspect that the data may have been too complex for K-NN to run, as mentioned in section 6.6. However, decision trees can handle more complicated data, which we will investigate in the next report.