Report 2a:

Supervised Learning: Classification and Regression

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. 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

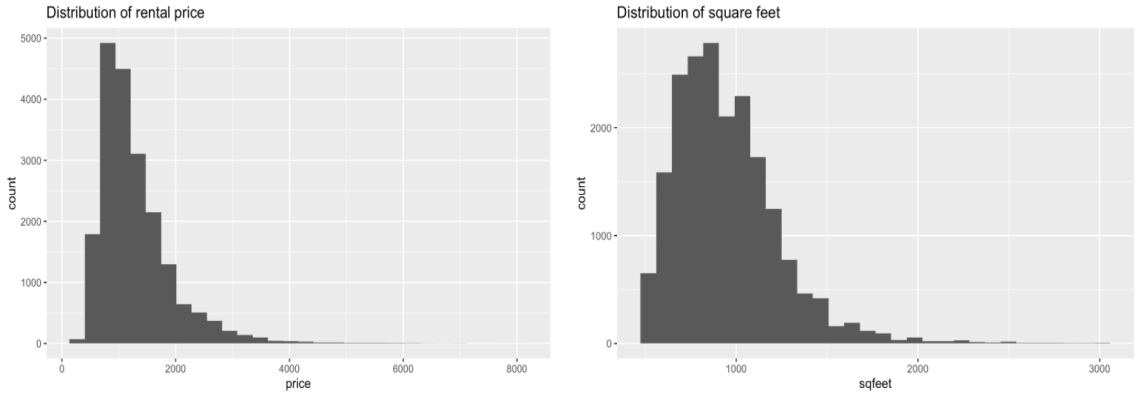

Figure 3.3

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. 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

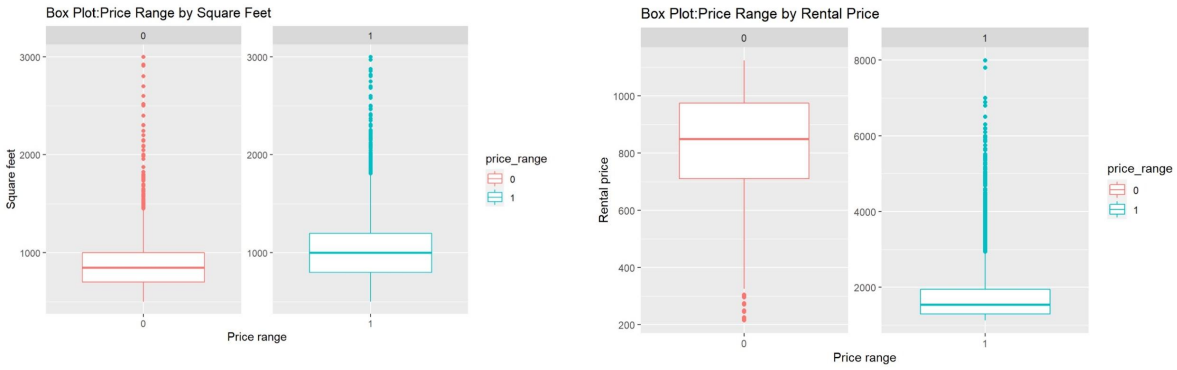

Figure 3.4

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

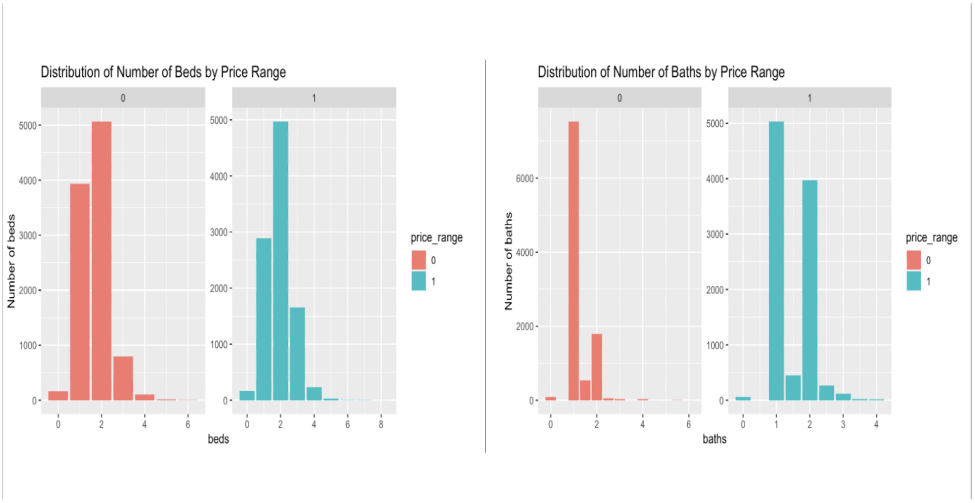

Figure 3.5 Bar Plot for Numerical Discrete Variables

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

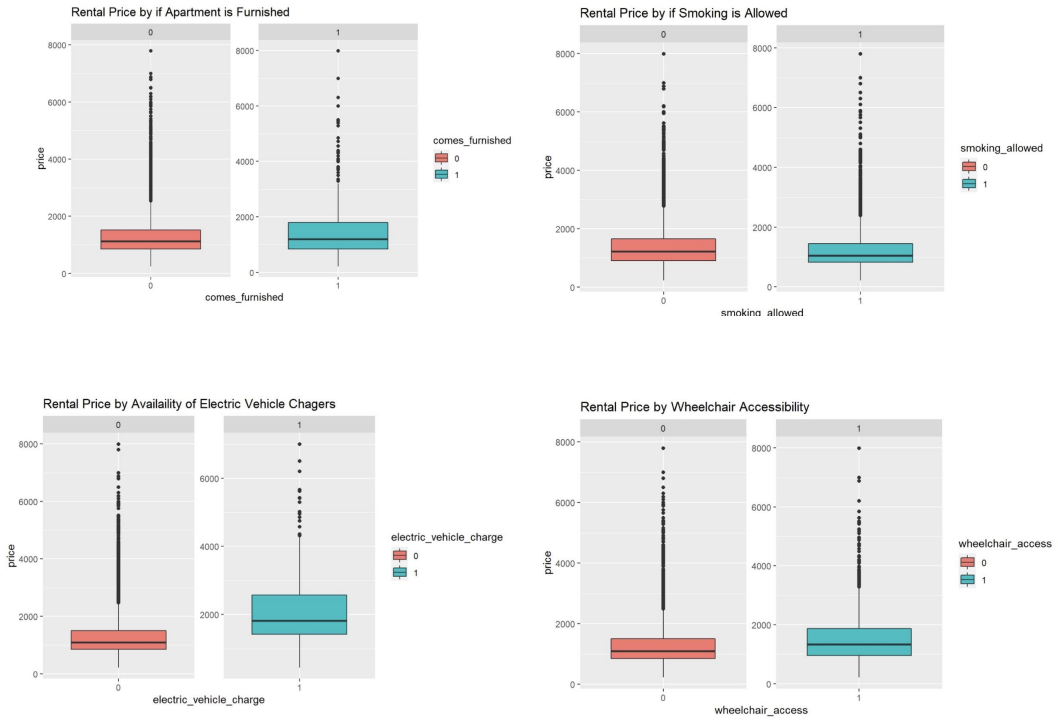

Figure 3.6 Box Plots for Furnishing, Smoking, Electric Vehicle Charger Availability, Wheelchair Accessibility

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

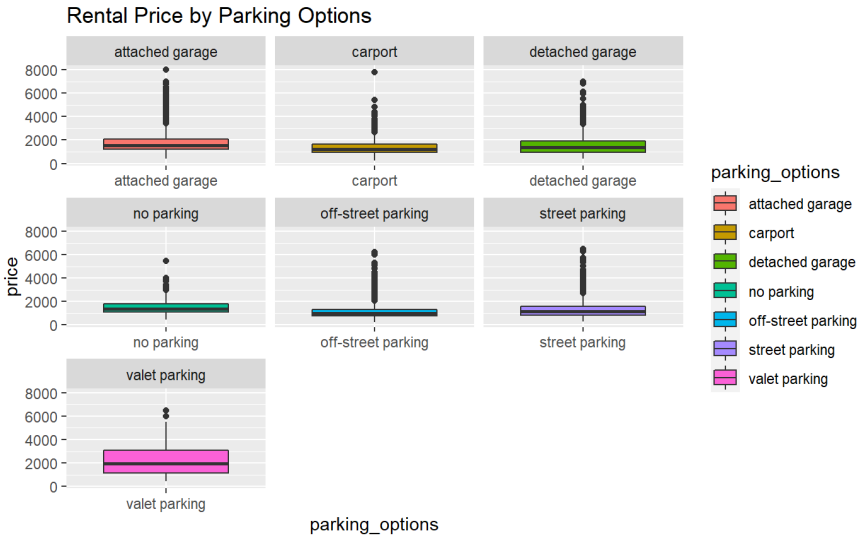

Figure 3.8 Box Plot for Parking

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

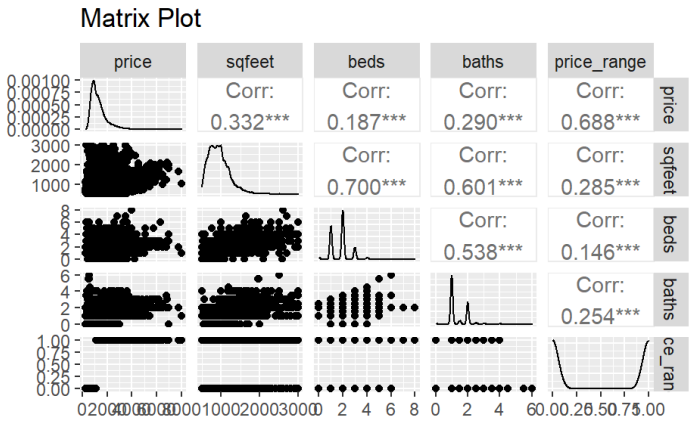

Figure 3.9 Matrix Plot

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

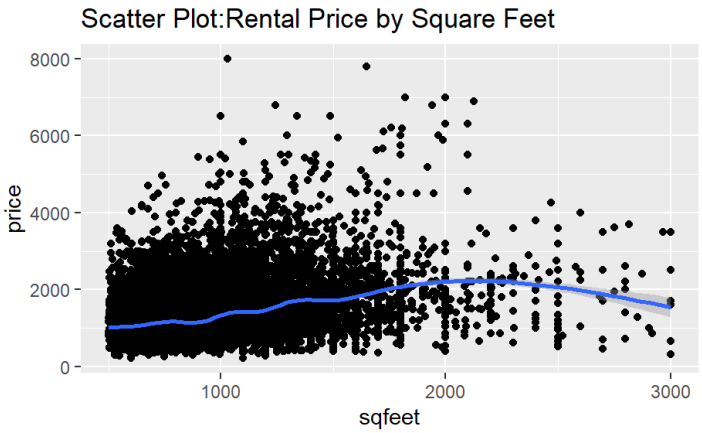

Figure 3.10 Scatterplot

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. Multiple Regression

4.1 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*, has been transformed into logged terms in the multiple regression methods in the following 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. Lastly, the adjusted R-squared and root mean square error (RMSE) in each model will be compared to find the best model.

4.2 Diagnostics

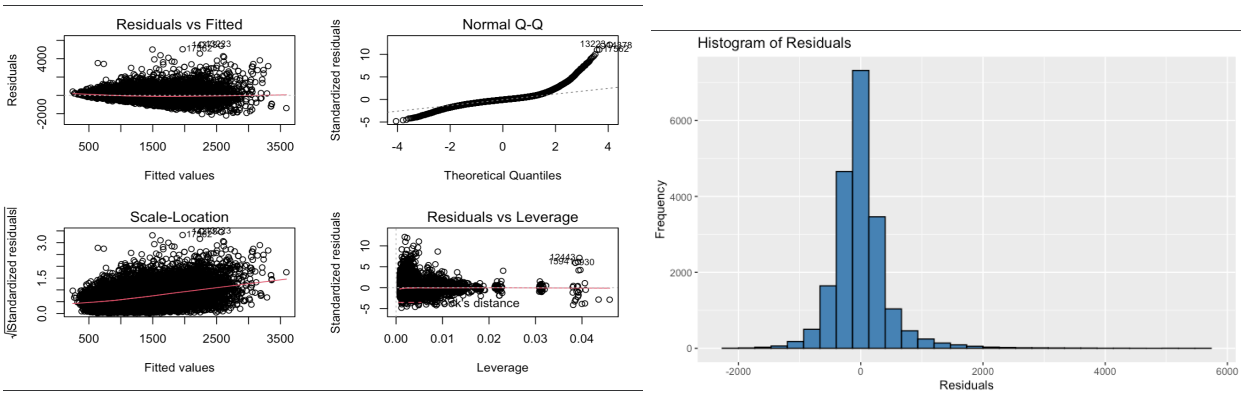

Figure 4.1 Diagnostic Plots before the change

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.

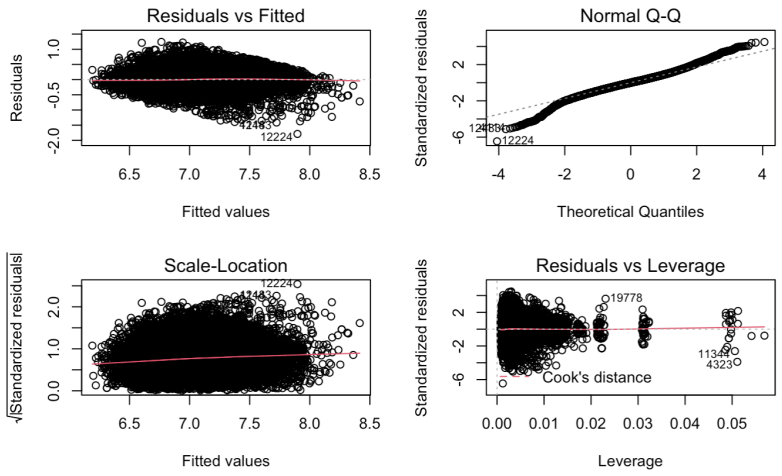

Figure 4.2 Diagnostic Plots after the change

Figure 4.2 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.

4.3 Train and Validate

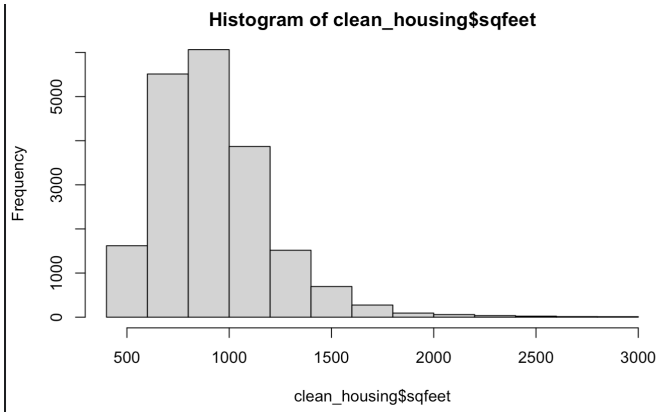

Figure 4.3

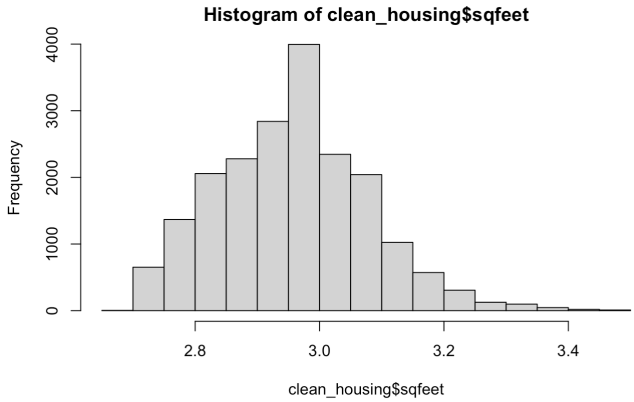

Figure 4.4

The skewness of the explanatory variables are checked before the data is split into training and validating sets. Figure 4.3 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 4.4 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.

4.4 Models

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.

4.4.1 Forward Selection

Forward selection is a method that adds one additional predictor whose addition leads 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 final 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

                   Df Sum of Sq    RSS    AIC
<none>                          584.52 -20483
+ wheelchair_access  1  0.025522 584.49 -20481
```

Figure 4.5 Forward Selection

4.4.2 Backward Elimination

Backward elimination is a method that is opposite to forward selection - it contains all the variables in the model and removes predictors as the removal leads to the most improvement of the criterion. The removed variables are the least statistically significant ones.

The final 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

                         Df Sum of Sq     RSS    AIC
<none>                                  911.03 -30338
- dogs_allowed            1      0.68  911.72 -30331
- comes_furnished         1      0.77  911.80 -30330
- beds                    1      1.56  912.59 -30320
- electric_vehicle_charge 1      2.96  913.99 -30302
- cats_allowed            1      3.81  914.84 -30291
- baths                   1      5.05  916.08 -30274
- smoking_allowed         1      7.18  918.21 -30247
- parking_options         6     55.44  966.47 -29649
- sqfeet                  1     56.18  967.21 -29630
- laundry_options         4     62.15  973.19 -29563
- state                  50    597.17 1508.20 -24454
```

Figure 4.6 Backward Elimination

4.4.3 Bidirectional Search

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

                          Df Sum of Sq     RSS    AIC
<none>                                  911.03 -30338
+ wheelchair_access        1      0.03  911.00 -30337
- dogs_allowed             1      0.68  911.72 -30331
- comes_furnished          1      0.77  911.80 -30330
- beds                     1      1.56  912.59 -30320
- electric_vehicle_charge  1      2.96  913.99 -30302
- cats_allowed             1      3.81  914.84 -30291
- baths                    1      5.05  916.08 -30274
- smoking_allowed          1      7.18  918.21 -30247
- parking_options          6     55.44  966.47 -29649
- sqfeet                   1     56.18  967.21 -29630
- laundry_options          4     62.15  973.19 -29563
- state                   50    597.17 1508.20 -24454
```

Figure 4.7 Bidirectional Search

4.4.4 Polynomial Term

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.

4.5 Best Model and Interpretation

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. 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 4.8 Adjusted R-squared and RMSE

Figure 4.8 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.

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.

The variable, *comes_furnished*, is found to be not statistically significant after the forward selection, so this variable will also 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 term, 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 terms, 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 the apartment contains electric vehicle charge, 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 in price 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.

5. Classification

5.1 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 regression very difficult. It is better to group the states into a smaller categorical variable and run regression on this variable.

5.2 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 5.1 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 5.2 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.

5.3 Model Perfection

```
                          Estimate Std. Error z value             Pr(>|z|)
(Intercept)             -8.628555   0.360611 -23.928 < 0.0000000000000002 ***
beds                    -0.379128   0.041870  -9.055 < 0.0000000000000002 ***
baths                    0.197748   0.052981   3.732              0.00019 ***
cats_allowed             0.308674   0.048058   6.423   0.0000000001336118 ***
smoking_allowed         -0.500388   0.042193 -11.859 < 0.0000000000000002 ***
electric_vehicle_charge  1.409551   0.182851   7.709   0.0000000000000127 ***
laundry_options          0.305414   0.017517  17.436 < 0.0000000000000002 ***
parking_options         -0.121314   0.016659  -7.282   0.0000000000003287 ***
I(sqfeet^2)              0.181754   0.009078  20.021 < 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: 13904  on 11395  degrees of freedom
AIC: 13922
```

Figure 5.3 Model without Insignificant Variables

```
Number of Fisher Scoring iterations: 4

                 beds                 baths           cats_allowed           dogs_allowed
             2.199323              1.792301              2.713766              2.818956
       smoking_allowed     wheelchair_access electric_vehicle_charge       comes_furnished
             1.069097              1.105720              1.053257              1.029712
       laundry_options       parking_options           I(sqfeet^2)
             1.248253              1.019362              2.443070
```

Figure 5.4 Significant Variables' VIF Score

After dropping insignificant variables, all the remaining variables are statistically significant with p-values of less than 0.05. The predictors in the model also have VIF values of less than 5, signifying no collinearity. Therefore, we meet the assumption of the logistic model, which is the absence of multicollinearity. However, in order to achieve the best model possible, an analysis is run to find if there are outliers that could impact the model.

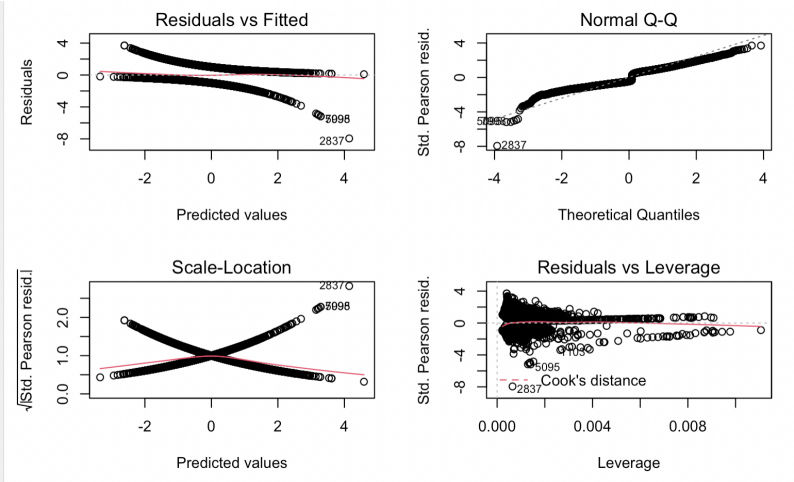

Figure 5.5 Plots with Outliers

Based on the Normal Q-Q plot above, it can be seen that there is an extreme outlier present. As logistic regression assumes that there should be no highly influential outliers, the data at row 2,837 is removed to prevent potential bias.

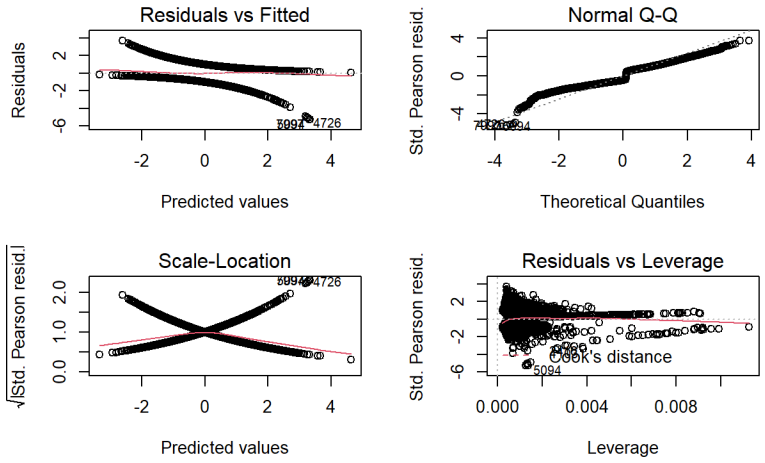

Figure 5.6 Plots without Outliers

After removing the outlier, the Normal Q-Q plot became closer to the straight line, which means the effect of outliers is decreased. The model was rerun without outliers.

```
                          Estimate Std. Error z value             Pr(>|z|)
(Intercept)             -8.658236   0.360941 -23.988 < 0.0000000000000002 ***
beds                    -0.380055   0.041888  -9.073 < 0.0000000000000002 ***
baths                    0.196762   0.053000   3.712              0.000205 ***
cats_allowed             0.309425   0.048075   6.436  0.00000000012239401 ***
smoking_allowed         -0.501346   0.042210 -11.877 < 0.0000000000000002 ***
electric_vehicle_charge  1.442545   0.184583   7.815  0.00000000000000549 ***
laundry_options          0.305354   0.017522  17.427 < 0.0000000000000002 ***
parking_options         -0.121688   0.016666  -7.302  0.00000000000028454 ***
I(sqfeet^2)              0.182484   0.009086  20.084 < 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: 15757  on 11402  degrees of freedom
Residual deviance: 13896  on 11394  degrees of freedom
AIC: 13914
```

Figure 5.7 Model without Outliers

It can be seen that all variables are statistically significant with p-values of less than 0.05 after removing outliers. The coefficients of *beds*, *smoking_allowed*, and *parking_options* are negative while the coefficients of the *baths*, *cat_allowed*, *electric_vehicle_charge*, *laundry_options*, and squared *sqfeet* are positive. The interpretation of coefficients of the logistic model depends on the odds ratio, which can be seen in section 5.4.

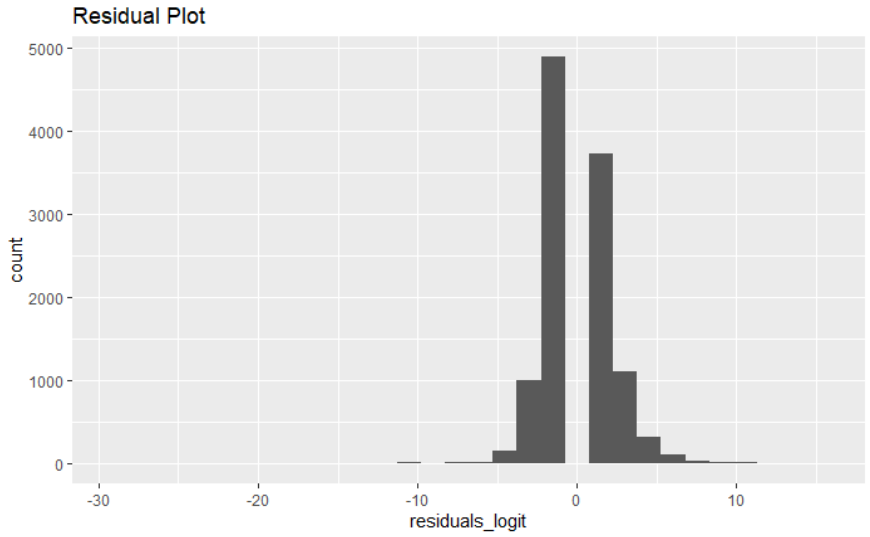

Figure 5.8 Residual Plot

Although the residuals plot for logistic regression does not contain values of 0, it is a possible situation for residuals from logistic regression. Overall, the residual plot is very close to normal distribution which means the assumptions for the model during

inference are valid and reasonable.

5.4 Interpretation of Odds

	odds <dbl>	odds_1 <dbl>
(Intercept)	0.0001736905	-0.9998263
beds	0.6838239277	-0.3161761
baths	1.2174540335	0.2174540
cats_allowed	1.3626413395	0.3626413
smoking_allowed	0.6057148389	-0.3942852
electric_vehicle_charge	4.2314504539	3.2314505
laundry_options	1.3571056852	0.3571057
parking_options	0.8854244584	-0.1145755
I(sqfeet^2)	1.2001952039	0.2001952

Figure 5.9 Odds 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*, and *parking_options* have less than one odds. The variables *baths*, *cats_allowed*, *electric_vehicle_charge*, *laundry_options*, and squared *sqfeet* have odds 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 31.6% 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 21.7% 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 36.3% 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 39.4% 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 323% increase of the probability of the high price apartment. It's unusual to have a

323% 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 35.7% increase of the probability of the 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 11.5% 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% increase of the probability of a high price apartment.

5.5 Interpretation of Confusion Matrix

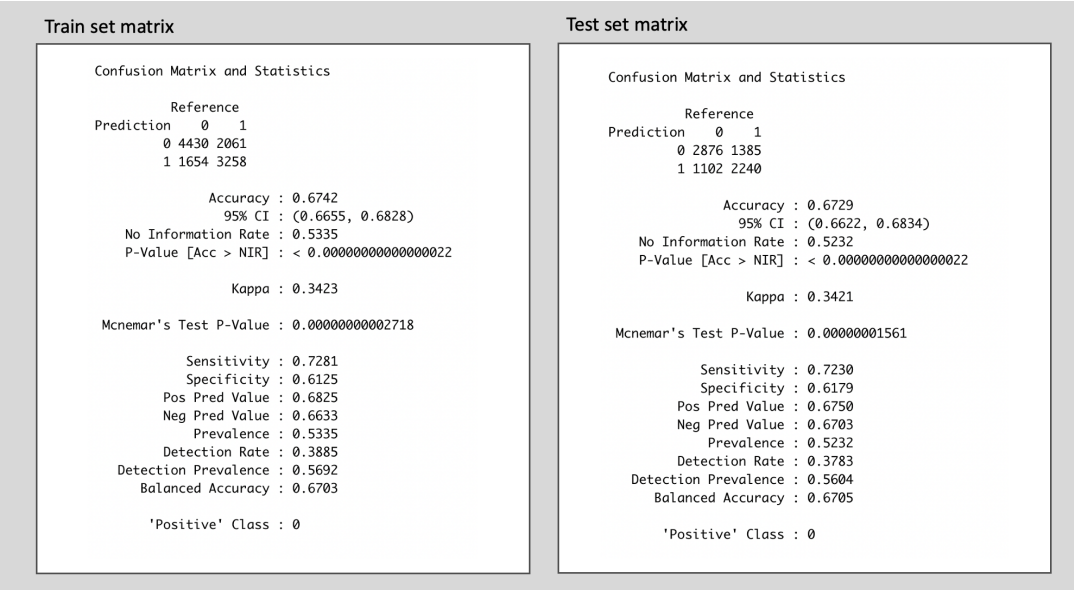

Figure 5.10 Confusion Matrix for Logistic Regression

Based on the confusion matrix, prediction accuracy for the test set is 0.6729 with a sensitivity of 0.7230 and specificity of 0.6179. The sensitivity measures how correctly a classifier can predict members in the data, which means logistic regression can correctly predict 72.3% of the class. The specificity measures how correctly a classifier can rule out the wrong members, which shows that the model can detect 61.8% of the false values. Overall accuracy of the model is 67.29%, 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.

5.6 Comparison/Contrast between Logistic Regression and K-NN.

```
k-Nearest Neighbors

11400 samples
   11 predictor
    2 classes: 'X0', 'X1'

Pre-processing: centered (11), scaled (11)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 10260, 10260, 10260, 10261, 10260, 10259, .
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.7620306  0.7488215  0.6443596
   7  0.7639644  0.7493697  0.6430418
   9  0.7677684  0.7533145  0.6408488
  11  0.7701740  0.7521096  0.6387192
  13  0.7698342  0.7515080  0.6357720
  15  0.7704118  0.7527668  0.6347695
  17  0.7709010  0.7570975  0.6327022
  19  0.7706349  0.7558910  0.6325748
  21  0.7710657  0.7568215  0.6334506
  23  0.7713874  0.7581373  0.6351424

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 5.11K-NN Result

The K-Nearest Neighbors classification is performed by using 11 predictor variables (state excluded) to target 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 K=23, the K-NN can provide the best classification result.

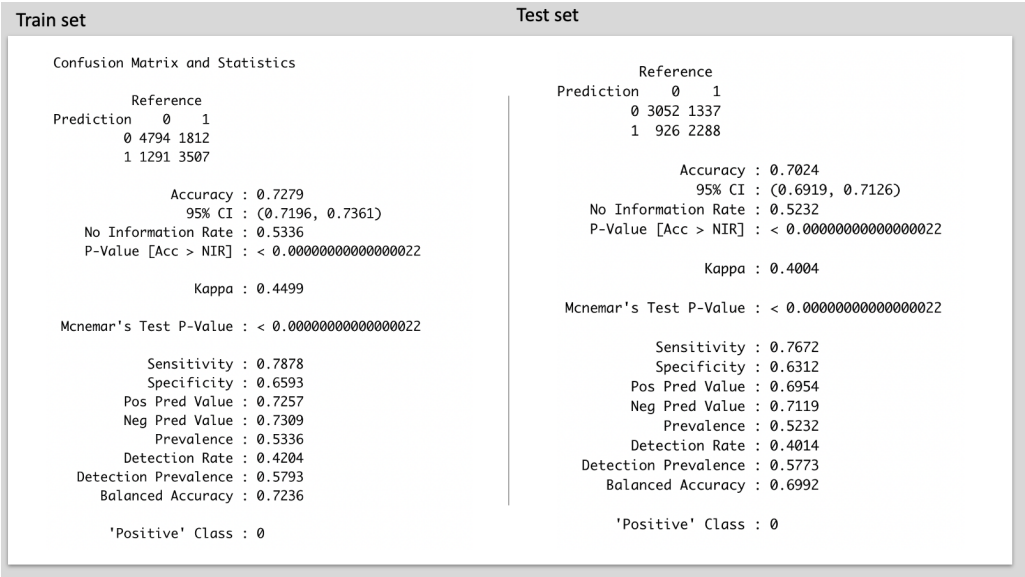

Figure 5.12 Confusion Matrix For K-NN

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 70.24%, which is 2.92% higher than the accuracy of logistic regression. The sensitivity for K-NN is 76.72%, which is 4.4% higher than the result of logistic regression. The specificity for K-NN is 63.12%, which is 1.3% higher than the specificity of logistic regression. The accuracies, the sensitivities, and the specificities for the train set and the test set are very similar. The K-NN model eliminates the overfitting problem.

Aside from better performance values than those of the logistic model, the K-NN method also has more advantages in the training model for the rental dataset. As the descriptive analysis mentioned above, some data of our variables are not normally distributed, which requires us to conduct data transformation. However, compared with the logistic method, the K-NN method does not assume the underlying data distribution pattern, so the K-NN method would have better performance than the logistic model.

F1_logistic	F1_knn
0.6663311	0.6925052

Figure 5.13 F1 Score

Both the logistic model and the K-NN model have different precision and recall, so F1 score should also be considered. F1 score is the weighted average of precision and recall. F1 score reaches its best value at 1 and its worst value at 0. As the figure shown above, the K-NN model has a higher F1 score than the logistic model. Therefore, the overall performance of K-NN is better than the logistic model.

6. Summary

6.1 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 high cost-of-living states such as 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, thus the probability is high. 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 regression model creates dummy variables of each type of option, which can provide a much clearer picture on prices for each feature.

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.

6.2 Reflections & Further Study

Throughout the process, we learned how the cleaning/handling of the data and the reliability of its source affects how we could hope to use the data. Before we could begin building the models, we needed to explore the data. After finding significant outliers and seeing that the distribution was skewed, we performed transformations so that the distributions became more Gaussian. Choosing cutoffs for what we considered outliers was challenging for certain variables. For example, there were significantly more apartments with 0 bedrooms (studio apartments) than apartments with over 4 bedrooms, which would make sense due to the nature of apartments being smaller-sized dwellings (compared to a house). However, omitting larger apartments would bias the results towards smaller apartments and we wouldn't learn anything from the larger units - although we would have lower variance. We certainly experienced the bias/variance tradeoff in many parts of the process.

When building the models, we remained cognizant of overfitting, 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 validated the data in our classification models and made sure performance was good on our testing data for the multiple regression models.

After having models that worked, interpretability of the models was not always straightforward. We saw this particularly in the logistic regression where the algorithm produces a probability rather than coefficients that can be used to directly describe the relationship between a predictor and the target. There was also added complexity in understanding transformed data with polynomial terms. Had we wanted to dive into the multiple regression with an included polynomial, the squared log of *sqfeet* in relation to

price and apartment square footage would have been difficult to explain in absolute terms.

The relationship between price and square footage was interesting, because it was obviously nonlinear. As briefly mentioned in the multiple regression session, we tried to add one polynomial term, to moderate success - it performed about the same as our other models, but not as well as our best one. To that end, it would be worth investigating how to better account for the smaller oscillations in the curve. We would also like to investigate interaction terms further, since it could be that multiple variables affect one another - we observed some correlation between a few of our predictors. This may help us achieve a better model.

For this report, we did classification without the inclusion of the state variable - data points across the country were aggregated. However, this does not necessarily encapsulate the rental pricing conditions in certain locales (regions/states/cities). The goal of examining factors affecting rental pricing may be more effective if locality is controlled for - the rental markets in Florida may be quite different than that of Montana. Still, the methods we used throughout this report can be applied for each individual state, and the granularity of data available within our dataset could support such a venture in future iterations of this project.