House Sales Analysis Project - MATH/DA 220-01

Yen Nguyen, Hoa Nguyen, Minh Le

12/17/2023

I. Describing Data

1. Introduction

King County is the most populous county in Washington, known for its innovation and rich culture. It is a key component of the Seattle-Tacoma-Bellevue metropolitan area, a hub of both economic activity and cultural diversity. In this project, we will use the data set that shape the landscape of "house sale prices" in King County between May 2014 and May 2015 from Kaggle to understand how various factors impose impacts on property values in King County. Through statistical analyses and data visualization, we seek to gain insights that inform both the past and future of real estate in this county.

2. Ethical Consideration

When working with this dataset, it is essential to prioritize privacy and confidentiality. Since it contains various details about homes, anonymizing data that could reveal individual identities is essential to protect privacy rights. Furthermore, it is crucial to ensure that data collection adheres to informed consent principles and legal regulations, respecting individuals' rights. To ensure fairness and objectivity, address biases in pricing and practices, and prioritize transparency through thorough documentation, all of which enhance ethical standards and contribute to the accuracy and significance of house price analysis.

3. Data Exploration

The data set includes homes sold between May 2014 and May 2015 in King County. The original dataset has 21 features of houses; however, we only use the following features about housing in this project:

- `price`: Price of each home sold
- `bedrooms`, `bathrooms`, `floors`: Number of bedrooms, bathrooms, floors, respectively.
- `sqft_living`, `sqft_lot`, `sqft_above`, `sqft_basement`: Square footage of interior living space, land space, floors, interior housing space that is above and below ground level, respectively.
- `waterfront`: A variable for whether the apartment was overlooking the waterfront or not.
- `view`: An index from 0 to 4 of how good the view of the property was.
- `condition`: An index from 1 to 5 on the condition of the apartment.
- `grade`: An index from 1 to 13 on the quality of construction and design.
- `zipcode`: Zip code area of the house.

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
7129300520	2014-10-13	221900	3	1.00	1180	5650	1	0	0
6414100192	2014-12-09	538000	3	2.25	2570	7242	2	0	0
5631500400	2015-02-25	180000	2	1.00	770	10000	1	0	0

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
2487200875	2014-12-09	604000	4	3.00	1960	5000	1	0	0
1954400510	2015-02-18	510000	3	2.00	1680	8080	1	0	0
7237550310	2014-05-12	1225000	4	4.50	5420	101930	1	0	0

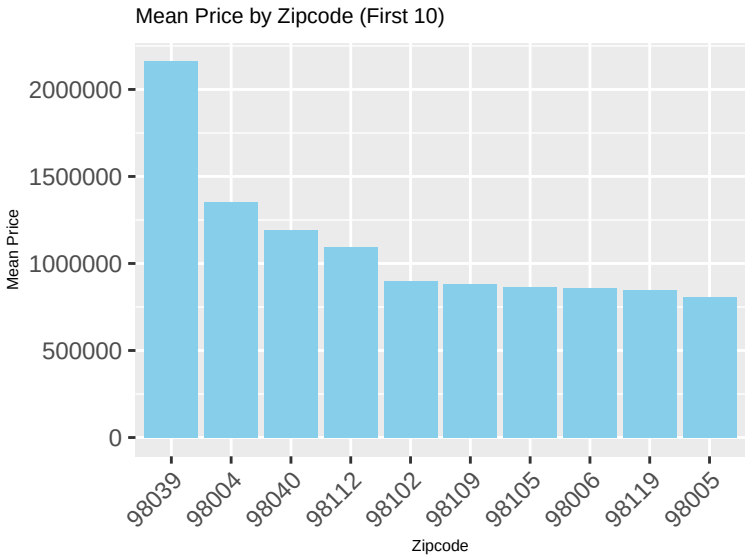
First, we can look into the summary statistics of the variables that we intend to use in the dataset:

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
Min. : 75000	Min. : 0.000	Min. :0.000	Min. : 290	Min. : 520	Min. :1.000
1st Qu.: 321950	1st Qu.: 3.000	1st Qu.:1.750	1st Qu.: 1427	1st Qu.: 5040	1st Qu.:1.000
Median : 450000	Median : 3.000	Median :2.250	Median : 1910	Median : 7618	Median :1.500
Mean : 540088	Mean : 3.371	Mean :2.115	Mean : 2080	Mean : 15107	Mean :1.494
3rd Qu.: 645000	3rd Qu.: 4.000	3rd Qu.:2.500	3rd Qu.: 2550	3rd Qu.: 10688	3rd Qu.:2.000
Max. :7700000	Max. :33.000	Max. :8.000	Max. :13540	Max. :1651359	Max. :3.500

view	condition	grade	sqft_above	sqft_basement
Min. :0.0000	Min. :1.000	Min. : 1.000	Min. : 290	Min. : 0.0
1st Qu.:0.0000	1st Qu.:3.000	1st Qu.: 7.000	1st Qu.:1190	1st Qu.: 0.0
Median :0.0000	Median :3.000	Median : 7.000	Median :1560	Median : 0.0
Mean :0.2343	Mean :3.409	Mean : 7.657	Mean :1788	Mean : 291.5
3rd Qu.:0.0000	3rd Qu.:4.000	3rd Qu.: 8.000	3rd Qu.:2210	3rd Qu.: 560.0
Max. :4.0000	Max. :5.000	Max. :13.000	Max. :9410	Max. :4820.0

The summary table provides a concise overview of key statistics of variables in the dataset, offering insights into its central tendencies and variability. The mean (average) house price of approximately 540,088 dollars, while the median value of price is 450000 dollars, suggesting that the data is somewhat right-skewed. Additionally, the table highlights the summary statistics for other features of the house. These statistics provide valuable context for understanding the distribution of house prices and their features, which are essential for subsequent data analysis and modeling.

One of the feature that we want to look into first is the area where the apartment is sold:



Area with the zip code 98039 has the highest average housing price in King County of 2160606 dollars, followed by areas 98004, 98040, and 98112 with the price range from 1000000 dollars to 1350000 dollars. This implies that locations of the house can be a good determinant for the price of a house.

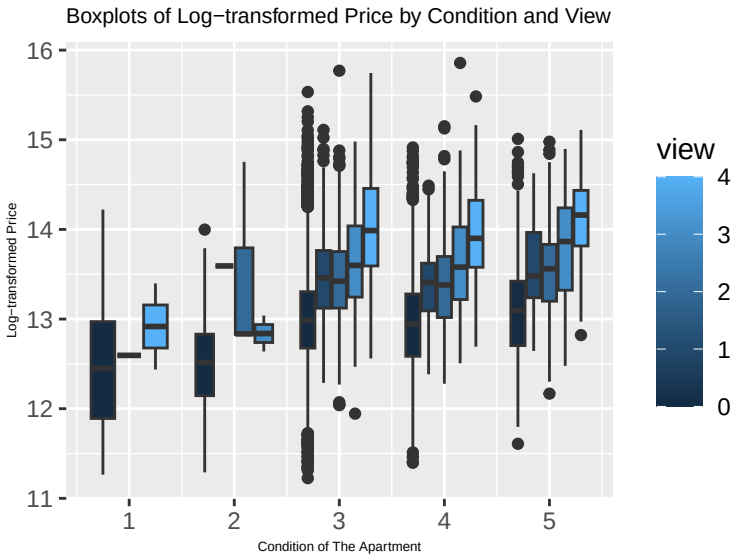
How about the other feature such as whether a house having a waterfront or not? We will look into the average price of the two groups and compare the difference between them.

```
## kc_house_data$waterfront: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   75000  320000  450000  531564  639897 7700000
## ------------------------------------------------------------
## kc_house_data$waterfront: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  285000  760000 1400000 1661876 2215000 7062500
```

Based on the above results, the average price for houses overlooking the waterfront is nearly triple the price of non-waterfront houses. This suggests that the presence of a waterfront view somewhat has an impact on the prices. However, the maximum price of houses without waterfront is higher than that of houses overlooking waterfront, at 7700000 and 7062500, respectively. This means that there are some properties without waterfront views have sold for a higher maximum price compared to those with the waterfront views. This observation can be attributed to outliers within the non-waterfront group for various reasons such as recent renovations, or unique features, have sold at exceptionally high prices.

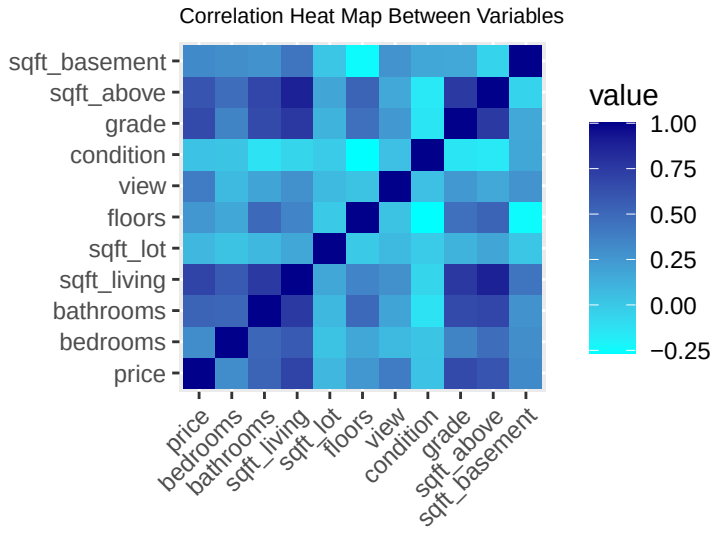
4. Statistical Analysis and Interpretation

To further explore the house prices' relationships with other variables, we then generate box plots illustrating house prices in relation to their condition and view rates.

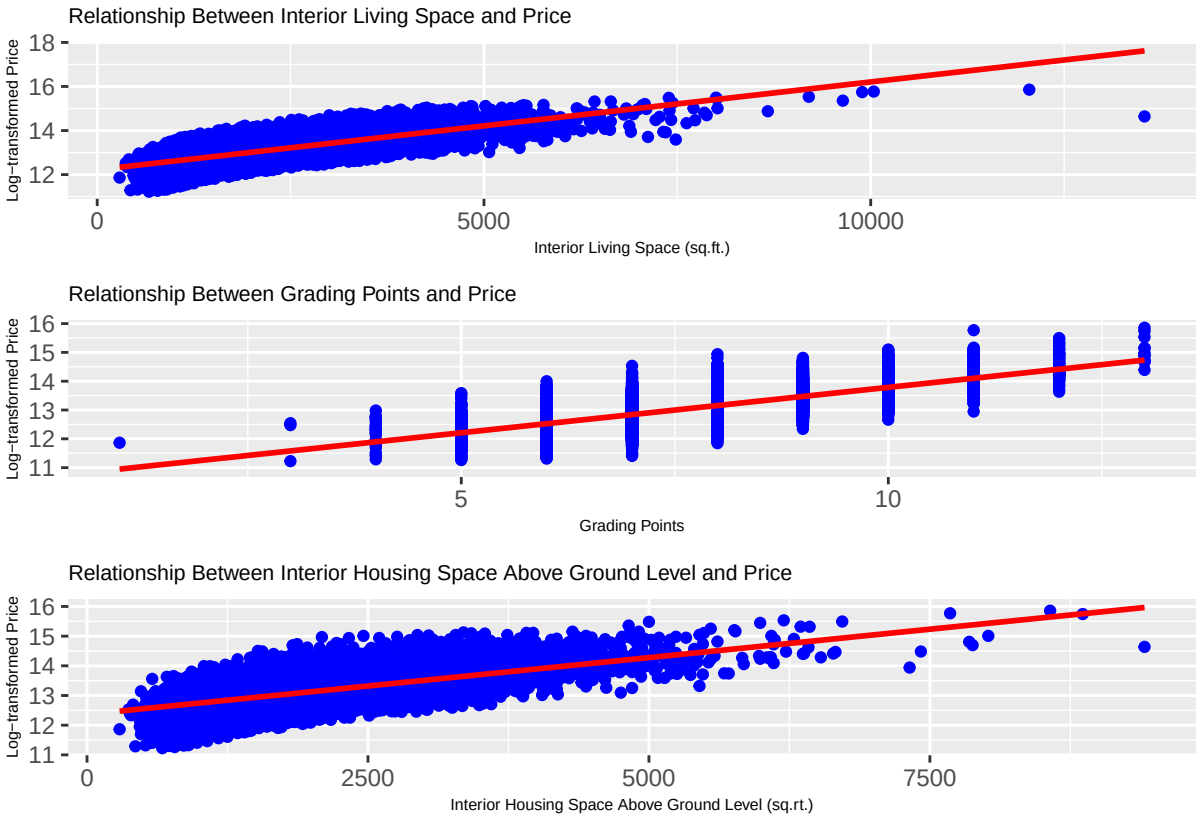


It can clearly be seen that the majority of apartments have condition ratings of at least 3. Moreover, for houses with condition ratings above 3, there is a noticeable trend: as the view rate increases, the house prices tend to rise, which is evident from the increasing price range in the boxplots. This suggests a positive correlation between house prices and their view. Meanwhile, there is no clear association between house prices and their condition as the houses whose conditions are rated 3, 4, and 5 exhibit minimal variation in pricing.

As `price` is our dependent variables and other variables in the dataset that we chose are all independent variables. Therefore, we will create a heat map to see the correlation between every variables.



From the heat map above, we can clearly see that there is a correlation between `price` and variables such as `sqft_living`, `grade`, and `sqft_above`. Hence, we will delve deeper into the connection between these variables by generating a scatter plot.



The scatter plots above depict a relationship between log-transformed housing price and their corresponding features. Regarding the interior living space, we can see that the larger the area inside the house, the higher the price of the house. Even though the data points are not close to the regression line, it is reasonable to say that there is a positive correlation between these two variables. This is also the case for grading points and interior housing above ground line variables. To be specific, the higher the grade for the level of construction of the house or the larger the interior housing space that is above the ground level is, the higher the price of the apartment will be.

II. Inference

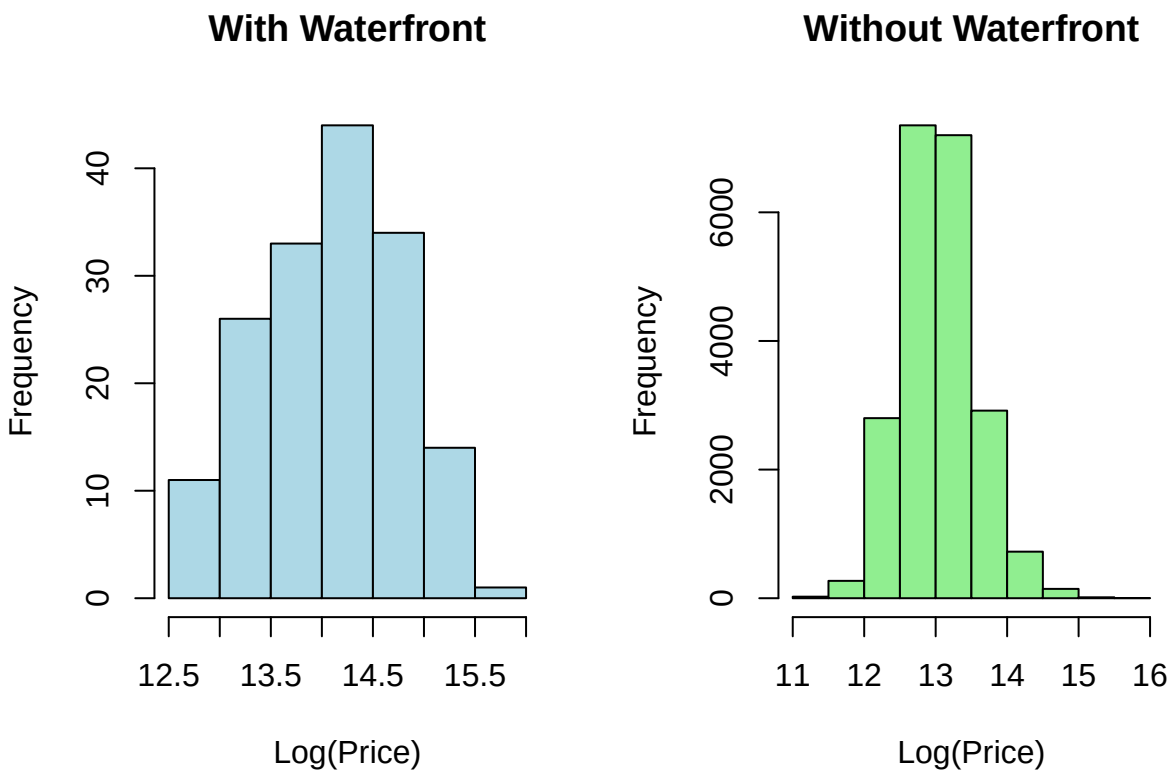
1. Hypothesis Test and Confidence Interval for Difference in Means Between Properties With and Without Waterfront Views

For this part, we want to assess the difference in average housing prices between properties with waterfront views and those without. By constructing this test statistic, we aim to evaluate whether there is a statistically significant difference in average housing prices between the two groups. Note that sample sizes are both large (163 for houses with waterfront and 21450 for ones without waterfront), so it is reasonable to apply the central limit theorem.

Let μ_w denote the average price of a house with a waterfront view (`waterfront = 1`), and μ_o denote the average price of houses that do not have a waterfront view (`waterfront = 0`). Our hypotheses are:

$$H_0 : \mu_w = \mu_o$$
$$H_A : \mu_w \neq \mu_o$$

First, let's look at the distributions of data in both groups:



From the histogram above, the variances of these two groups are not identical (where the range of values for houses without waterfront are more spread out than ones with a waterfront view), so we conduct a t-test under the assumption that the variances are unequal to test whether or not the average prices of houses with waterfront view and those without waterfront view are equal:

```
##
##  Welch Two Sample t-test
##
## data:  with_waterfront$price and without_waterfront$price
## t = 12.876, df = 162.23, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##   956963.3 1303661.6
## sample estimates:
## mean of x mean of y
## 1661876.0  531563.6
```

In our comparison of prices between properties with a waterfront and without a waterfront view, we found our test statistic is $t^* = 12.876$, a measure of how many standard errors the sample mean difference is from the hypothesized population mean difference if the null hypothesis is true, which is considered to be statistically significant and the difference here is unlikely to have occurred by chance. Thus, our p-value is given by $\mathbb{P}(|t_{n-1}| \geq 12.876)$

Our p-value is less than 2.2e-16, so we have sufficiently strong evidence to reject the null hypothesis at the $\alpha = 0.05$ significance level. Namely, we have sufficiently strong evidence to conclude that there is a difference between the average price of properties with a waterfront view and those that do not.

In terms of confidence interval, we are 95% confident that the average price for houses with a waterfront view is between 956963.3 more and 1303661.6 more than the average price for houses that do not have a waterfront view. By 95% confidence level, we mean that if we repeated the study many times, and then constructed many 95% confidence intervals for the difference in average prices between houses with and without waterfront view, then around 95% of those confidence intervals would contain the true difference.

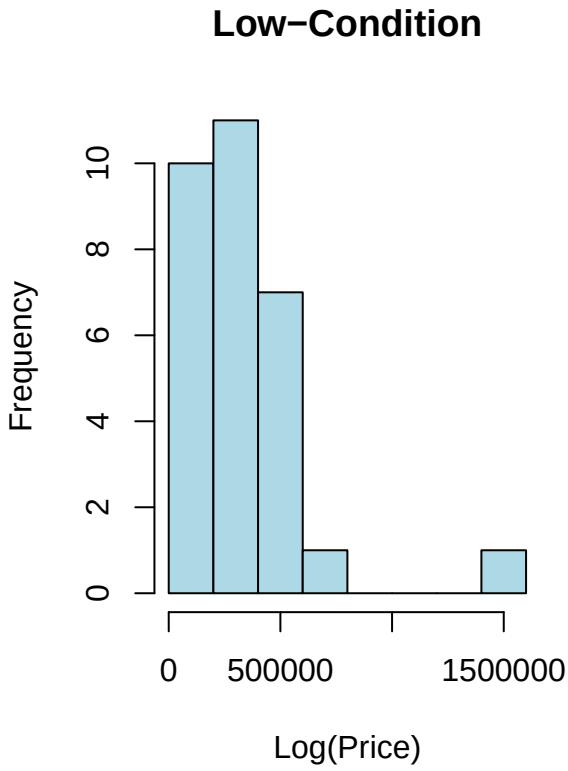
2. Hypothesis Testing and Confidence Interval for Difference in Means Between Low and High Condition Properties

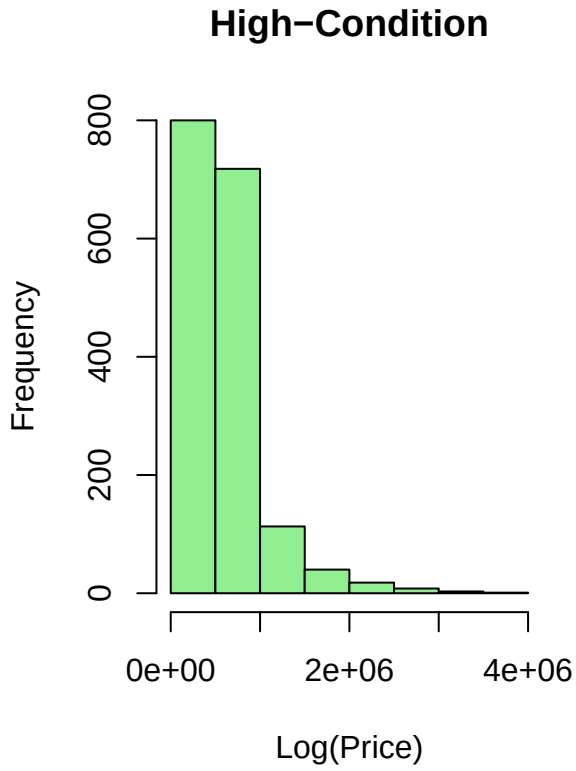
The condition of the house is one of the factor that people need to consider when buying a new house. Hence, we want to test whether the average price of low-condition house (`condition = 1`) is different from the average price of high-condition house (`condition = 5`). Note that the sample size for both our groups are large enough (30 for low condition houses and 1701 for high condition houses), it is reasonable to apply the central limit theorem.

Let μ_S denote the average price of low-condition house, and μ_N denote the average price of high-condition house. Our hypotheses are:

$$H_0 : \mu_S = \mu_N$$
$$H_A : \mu_S \neq \mu_N$$

First, let's look at the distributions of data in both groups:





From histograms above, we know that the variances of these two groups are not identical, so we conduct a t-test under the assumption that the variances are unequal.

```
##
##  Welch Two Sample t-test
##
## data:  low_condition and high_condition
## t = -5.5045, df = 31.396, p-value = 4.857e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -380933.2 -175039.6
## sample estimates:
## mean of x mean of y
##  334431.7  612418.1
```

From the t-test, our test stastic is $t^* = -5.5045$, which measures of how many standard errors the sample mean difference is from the hypothesized population mean difference under the null hypothesis. Thus, our p-value is given by $\mathbb{P}(|t_{n-1}| \geq 5.5045)$.

Our p-value is 4.857e-06, so we can reject the null hypothesis at the $\alpha = 0.05$ significance level. Indeed, we have strong enough evidence to conclude that there is a difference in the average price of low and high-condition houses.

The confidence interval for the difference between the average price between the two groups is (-380933.2, -175039.6). We are 95% confident that the true difference in average price between low-condition houses and high-condition houses in King County is between - $380933.2 and - $175039.6. By 95% confidence, we mean that if we repeated the sample many times and constructed many different confidence intervals, then around 95% of them would contain the true difference in average price between the two groups. Since the interval represents a range of "plausible" values, and all values in this range are smaller than 0, then we know that there is the average price for high-condition houses is higher than the average price for low-condition houses.

3. Hypothesis Testing and Confidence Interval for Difference in Two Proportions

The location of the house can be a good determinant for the price of the house. Therefore, we curious about whether houses with above average quality are more likely to locate in a place with high average housing price. Note that the sample size

for both our groups are large enough, 199 and 590, respectively, it is reasonable to apply the central limit theorem.

If p_c denotes the proportion of above average quality houses locate in 98038 (location with highest average housing price), and p_n denotes the proportion of above average quality houses locate in 98002 (location with lowest average housing price), then we are interested in conducting the following hypothesis test:

$$H_0 : p_c = p_n$$
$$H_A : p_c > p_n.$$

We begin by testing our test statistic,

$$z^* = \frac{p_c - p_n}{SE(pc - pn)}.$$

where $SE(pc - pn) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{m}}$ and $\hat{p} = \frac{np_c + mp_n}{n+m}$

We calculated our test stastic to be $z^* = 9.909209$. Thus, our p-value is given by $\mathbb{P}(z \geq 9.909209)$, which is

```
## [1] 1.898201e-23
```

Our p-value is 1.898201e-23, so we can reject the null hypothesis at the $\alpha = 0.05$ significance level. Indeed, we have strong enough evidence to conclude that houses with above average quality are more likely to locate in a place with high average housing price.

For further understanding, we will now construct a confidence interval for difference in proportions of these groups, which is given by the following formula:

$$(p_c - p_n - z_{\alpha/2}SE(p_c - p_n), p_c - p_n + z_{\alpha/2}SE(p_c - p_n))$$

Here, $SE(p_c - p_n) = \sqrt{\frac{p_c(1-p_c)}{c} + \frac{p_n(1-p_n)}{n}}$.

We then got this interval:

```
## [1] 0.1939668
```

```
## [1] 0.3274538
```

We are 95% confidence that the true difference in the proportion of above average quality houses located in the highest and lowest average housing price is between (0.1939668, 0.3274538). By 95% confident, we mean that if we repeated the sample many times and constructed many different confidence intervals, then around 95% of them would contain the true difference in proportions between the two groups. Since the interval represent a range of "plausible" values, and all values in this range are bigger than 0, then we know that houses with above average quality are more likely to locate in a place with high average housing price. We now are more confidence to conclude that location has a great impact on the price of the properties.

4. Bootstrapped Hypothesis Test for The Variances of Prices of Houses with Low-quality and High-quality Construction and Design

We want to see whether the variance of house prices differs between houses with low-quality construction and design (`grade < 7`) and houses with good construction and design (`grade >= 7`). It's worth noting that houses with a grade of 7 represent an average level of construction and design.

Let σ_x denote the variance of prices of low condition house, and σ_y denote the variance of price of high condition house. Our hypotheses are:

$$H_0 : \sigma_x = \sigma_y$$
$$H_A : \sigma_x \neq \sigma_y$$

To assess the hypothesis, we conducted a bootstrapped hypothesis test by simulate repeating the study R = 1,000 times. We randomly sampled prices for houses with low-quality construction and design (`grade < 7`) and those with good construction and design (`grade >= 7`). The observed difference in sample variances was calculated, and by comparing this difference to the variances obtained through resampling, we determined the following p-value:

```
## [1] 0.48
```

Since our p-value of 0.48 was greater than $\alpha = 0.05$, we fail to reject the null hypothesis. This means that based on the data we have and the results of our bootstrap hypothesis test, there is no significant difference in the population variances of house prices between the two groups with different grades of construction and design. Our analysis suggests that the variances in house prices for properties with a grade of less than 7 and properties with a grade of 7 or higher are similar within the sample we've analyzed.

5. Bootstrapped Confidence Interval for the Variance of the Population

In overall, we interested in looking into the variance of the housing price for King County, so we constructed a 95% confidence interval using 1,000 bootstrap resamples. This interval provides a plausible range for the population variance of house prices:

```
##         2.5%
## 124975814733
```

```
##      97.5%
## 1.4557e+11
```

The 95% confidence interval for the variance of house prices, based on the bootstrapped samples, ranges from approximately $124.98 billion to $145.58 billion.

This confidence interval gives us a range of plausible values for the population variance of house prices. We are 95% confident that the true population variance is likely to fall within this interval. In other words, if I repeatedly take samples of the same size from the same population and compute a 95% CI for the variance using the same resampling method, approximately 95% of those intervals would contain the true population variance. This interval suggests that the housing market may exhibit diverse and fluctuating price dynamics. Stakeholders in the real estate sector should be aware of this broad range, emphasizing the importance of adaptive strategies to navigate potential shifts in house prices and market conditions.

6. Correlation Test

For the last part of our project, we will look into the number of functioning rooms in the house. We will do it by making a correlation test to see the connection between the price of the house and its feature.

```
##
##  Pearson's product-moment correlation
##
## data:  kc_house_data$price and kc_house_data$bedrooms
## t = 47.651, df = 21611, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2962354 0.3203646
## sample estimates:
##       cor
## 0.3083496
```

```
##
##  Pearson's product-moment correlation
##
## data:  kc_house_data$price and kc_house_data$bathrooms
## t = 90.714, df = 21611, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5154140 0.5347258
## sample estimates:
##       cor
## 0.5251375
```

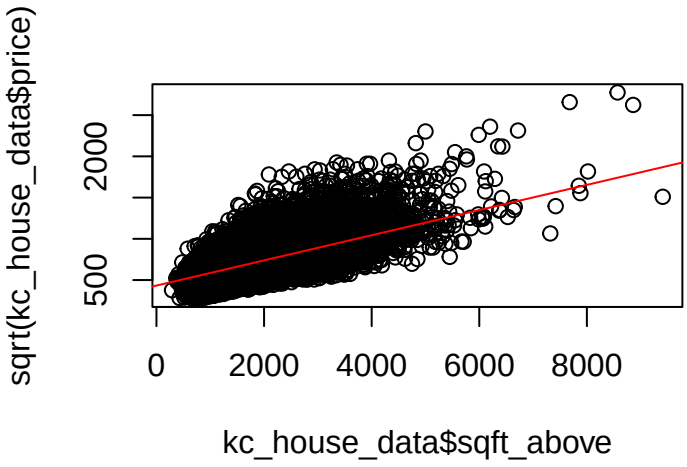
```
##
##  Pearson's product-moment correlation
##
## data:  kc_house_data$price and kc_house_data$floors
## t = 39.06, df = 21611, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2442983 0.2692042
## sample estimates:
##       cor
## 0.2567939
```

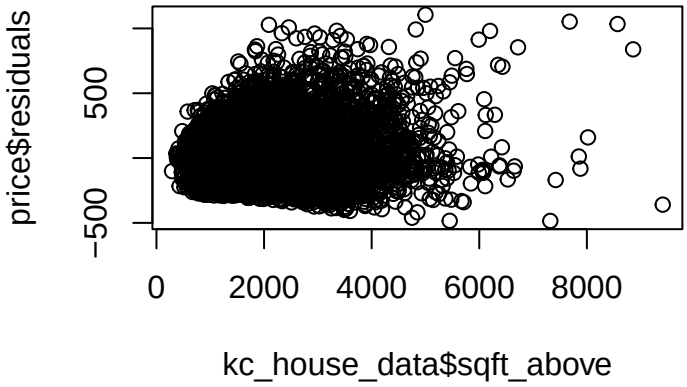
According to the result of the correlation test above, the correlation coefficient (cor) is positive in this case for all three tests, which indicates that as one variable increases, the other variable (price) tends to increase as well. However, the correlation here is pretty moderate, suggesting that there is not an extremely strong relationship between them. Surprisingly, the number of bathrooms seem to have the highest association with the price of the house, with correlation of 0.525. However, the p-value of all the tests are less than the significance level alpha = 0.05. Thus, we can reject the null hypothesis, concluding that there is a statistically significant correlation between `price` and the number of functioning rooms.

III. Regression

1. Simple Linear Regression Models

For this part, we curious about looking into the relationship between `price` and `sqft_above`. We will first check whether the assumptions for linearity appear to be satisfied.





Technical conditions for SLR:

- Linear function: Based on the plot between `price` and `sqft_above`, we can see that our data can roughly fit a line
- Independence of errors: The errors are independent from each other and do not follow any particular trend
- Normally distributed: As we plot the histogram, the errors are normally distributed
- Equal variances: Our errors have equal variances at each value of the predictor.

Therefore, we will now fit the linear regression model for `price` and `sqft_above`:

```
##
## Call:
## lm(formula = kc_house_data$price ~ kc_house_data$sqft_above)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -913132 -165624  -41468  109327 5339232
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               59953.2     4729.8   12.68   <2e-16 ***
```

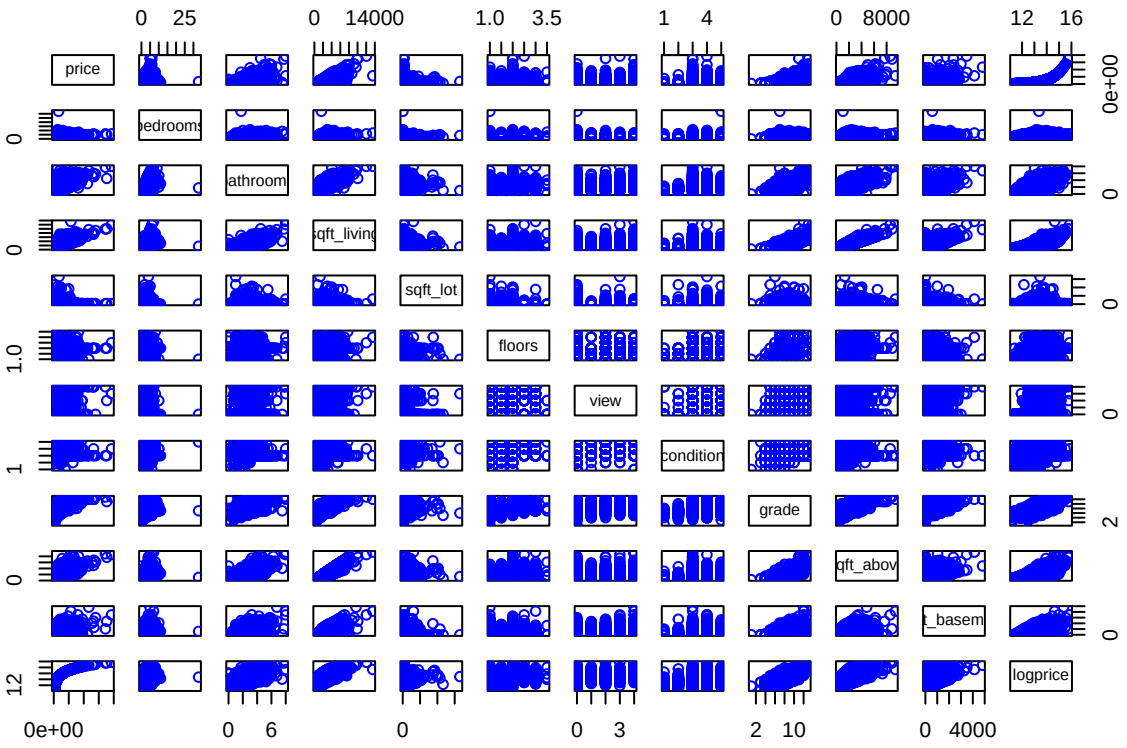
```
## kc_house_data$sqft_above    268.5        2.4   111.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292200 on 21611 degrees of freedom
## Multiple R-squared:  0.3667, Adjusted R-squared:  0.3667
## F-statistic: 1.251e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

According to the result above, the coefficient of `sqft_above` is 268.5, which is pretty large and positive. This means that as the value of `sqft_above` increases, the mean of `price` also tends to increase significantly. If `sqft_above` increases by one square foot, the `price` increases by $268.5 according to the model. Also, the p-value in this case is less than 2e-16, which indicates the variable `sqft_above` has a significant influence on the `price`.

From the results, we can have a regression equation: $price = 59953.2 + 268.5 * sqft\_above \pm 2.4$.

2. Mutiple Regression Models

Before fitting the model, we will create the `pairs` plot to check that the assumptions for linear regression appear to be satisfied.



From the `pairs` plot above, we can see that the relationship between `price` and all the variables is non-linear, which is specifically exponential relationship. However, `log(price)` has a relationship with these variables: number of bathrooms (`bathrooms`), the square footage of interior living space (`sqft_living`), the quality of construction and design (`grade`), interior housing space that is above ground level (`sqft_above`), the square footage of basement (`sqft_basement`).

For our first multiple regression model, we have focused on capturing the essence of house pricing through a set of basic features that we think are important. This model includes the variables such as the square footage of interior living space (`sqft_living`), interior housing space that is above ground level (`sqft_above`), and the quality of construction and design (`grade`). We will include response variable `log(price)` instead of `price`:

```
##
```

```
## Call:
## lm(formula = log(price) ~ sqft_living + sqft_above + grade, data = kc_house_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76362 -0.24675  0.00378  0.22910  1.41672
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.108e+01  1.893e-02  585.45   <2e-16 ***
## sqft_living  3.022e-04  5.643e-06   53.56   <2e-16 ***
## sqft_above  -1.309e-04  6.183e-06  -21.18   <2e-16 ***
## grade        2.049e-01  3.241e-03   63.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3477 on 21609 degrees of freedom
## Multiple R-squared:  0.5643, Adjusted R-squared:  0.5643
## F-statistic:  9330 on 3 and 21609 DF,  p-value: < 2.2e-16
```

The R^2 of my model is 0.5643, indicating that 56.43% of the variability in the logarithm of housing prices is explained by the regression model. Additionally, all coefficients are statistically significant, as the p-value are all smaller than 0.05. This implies that all the variables that I chose above help explain the housing price.

Note: As `sqft_living` and `grade` are closely correlated to each other, we will only include one variable in the model since it will take the key credit, making the remaining variable redundant. In addition, `sqft_living` and `sqft_above` are also correlated. Therefore, including all three variables in the regression model might not necessary.

However, there might be a different variables that are also important in understanding the variability of the logarithm of housing price. Therefore, our second regression models adopts a more comprehensive approach, which is called step wise variable selection methods. To be specific, I will use backward selection to select the variables for the regression model:

```
##
## Call:
## lm(formula = log(price) ~ bathrooms + sqft_living + grade + sqft_above,
##     data = df_mod2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77021 -0.24529  0.00369  0.23006  1.41866
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.108e+01  1.893e-02 585.608  < 2e-16 ***
## bathrooms   -1.651e-02  4.787e-03  -3.448 0.000565 ***
## sqft_living  3.102e-04  6.094e-06  50.894  < 2e-16 ***
## grade        2.072e-01  3.306e-03  62.666  < 2e-16 ***
## sqft_above  -1.306e-04  6.182e-06 -21.118  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3476 on 21608 degrees of freedom
## Multiple R-squared:  0.5646, Adjusted R-squared:  0.5645
## F-statistic:  7004 on 4 and 21608 DF,  p-value: < 2.2e-16
```

After using backward selecting method, `bathrooms, sqft_living, grade, sqft_above` are selected for the multiple linear regression model. All the variables included in the model are significantly important, as all the p-value are smaller than 0.05.

Coefficient:

- `Intercept`: The intercept is 1.108e+01. This is the estimated value of the response variable (`log(price)`) when all predictor variables are zero.

- `bathrooms`: The coefficient for `bathrooms` is -1.651e-02. This suggests that as `bathrooms` increases by one unit, the logarithm of price is expected to decrease by 1.651e-02 units.
- `sqft_living`: The coefficient for `sqft_living` is 3.102e-04. This suggests that as `sqft_living` increases by one unit, the logarithm of price is expected to increase by 3.102e-04 units.
- `sqft_above`: The coefficient for `sqft_above` is -1.306e-04. This suggests that as `sqft_above` increases by one unit, the logarithm of price is expected to decrease by 1.306e-04 units.
- `grade`: The coefficient for `grade` is 2.072e-01. This suggests that as `grade` increases by one unit, the logarithm of price is expected to increase by 2.072e-01 units.

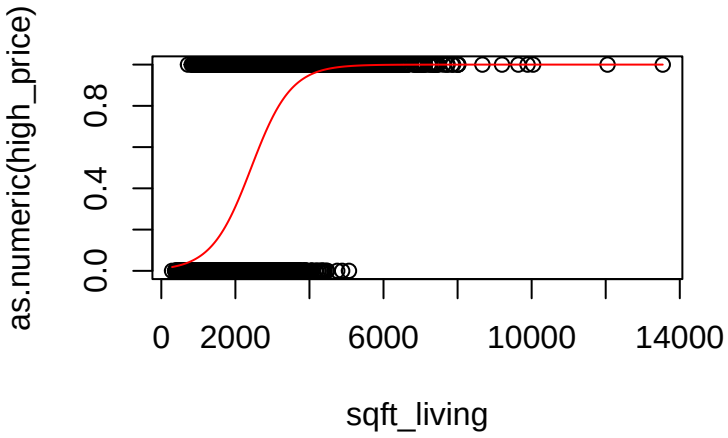
The R-squared value for the model is 0.5646, which indicates that 56.46% of the variability of the logarithm of price is explained by the model. The Adjusted R-squared for the model is 0.5645. This suggests that all the variables included in the model are important for expalining the variability of the logarithm of housing price.

Note: As `sqft_living`, `sqft_above`, `bathrooms`, and `grade` are closely correlated to each other, we will only include one variable in the model since it will take the key credit, making the remaining variables redundant. Therefore, including all these variables in the regression model might not necessary since one of the three variables will take the key credit, making the others redundant.

Since this dataset has a few variables that have a linear relationship with housing price and do not correlate with other variables, building a linear regression model is not the best way to predict the housing price in this case.

3. Logistic Regression Models

In terms of logistic regression, we decided to use `sqft_living` to predict whether the property has a `high_price` or not. Also, we used `mutate` to create a new column to determine if the price of a particular property is above 540088 which is an average price in our data set, then it is considered to have a `high_price`; otherwise, it will be assigned 0.




```
##
## Call:
## glm(formula = as.numeric(high_price) ~ sqft_living, family = binomial,
##     data = kc_house_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.513e+00  6.122e-02  -73.71   <2e-16 ***
## sqft_living  1.859e-03  2.743e-05   67.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28400  on 21612  degrees of freedom
## Residual deviance: 20619  on 21611  degrees of freedom
## AIC: 20623
##
## Number of Fisher Scoring iterations: 5
```

Based on the results of the logistic regression model, we can see that the coefficient is 1.859e-03, which is positive. As `sqft_living` increases by one square foot, the log-odds of `high_price` increases by 1.859e-03. Also, both the intercept and sqft_living coefficients are statistically significant ($p < 0.001$), suggesting that they make a remarkable contribution to predicting whether a property has a high price or not. In our case, the logistic regression equation is: $\hat{p}(high\_price = 1|sqft\_living = x) = \frac{e^{-4.513+1.859e-03\times sqft\_living}}{1+e^{-4.513+1.859e-03\times sqft\_living}}$

4. New Technique - Random Forest

Random Forest is a robust machine learning algorithm used for both classification and regression tasks. It constructs multiple decision trees during training, where each tree learns from random subsets of the data, and combines their outputs to make predictions. This ensemble technique helps to reduce overfitting and improve accuracy.

Here, we would like to conduct a Random Forest classification on our dataset:

1. **Data Preprocessing:** Create a new categorical variable (`price_rate`) based on the `price` column's distribution. Specifically, we categorize it into either "low" (<$300,000), "medium" (<$300,000 and ≤$650,000), or "high" (> $650,000).
2. **Data Splitting:** Split the dataset into training and testing subsets (80% for training, 20% for testing) using the `sample` function.
3. **Random Forest Model Building:** Use the `randomForest` function from the `randomForest` library in R to build the model. Specify the formula (`price_rate ~ . - price`) to predict `price_rate` based on other features in the dataset.
4. **Model Evaluation:** Validate the model using the test dataset by making predictions (`predict`) and comparing them with the actual `price_rate` values. Generate a table (`table`) to analyze predicted vs. actual values.
5. **Accuracy Calculation:** Calculate the accuracy of the model's predictions by comparing predicted values with the actual `price_rate` values from the test set.

```
## [1] 17290    14
```

```
## [1] 4323   14
```

```
##
## Call:
##  randomForest(formula = price_rate ~ . - price, data = train,      ntree = 1000, mtry = 5)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 18.72%
## Confusion matrix:
##        high  low medium class.error
## high   3207   15    973   0.2355185
## low       5 2598   1051   0.2889984
## medium  633  560   8248   0.1263637
```

There are 17290 and 4323 observations in the training and test dataset, respectively.

Our forest comprises 1000 trees, and we've designated `mtry` as 5, representing the count of randomly chosen variables assessed for potential splits at each stage. The out-of-bag error rate is around 18.72%, indicating the model's estimated error on unseen data.

Confusion matrix

The confusion matrix displays the model's predictions for different classes (high, low, medium).

- The 'high' and 'low' classes have higher error rates (23.55% and 28.90%) compared to the 'medium' class (12.64%). This indicates that the model struggles relatively more in accurately predicting these classes.
- For 'high' class predictions, the model has a noticeable error rate in misclassifying instances as 'medium'.
- For 'low' class predictions, the model has a considerable error rate in misclassifying instances as 'medium'.

```
##
## prediction high  low medium
##     high    801    1    153
##     low       0  648    143
##     medium  211  267   2099
```

The table shows the results of using our model built from the training data to predict the test data.

Predicted 'high' class:

Out of the instances predicted as 'high' (801 + 1 + 153 = 955):

- 801 instances were correctly predicted as 'high'.
- 1 instance was mistakenly predicted as 'low'.
- 153 instances were mistakenly predicted as 'medium'.

Predicted 'low' class:

Out of the instances predicted as 'low' (0 + 648 + 143 = 791):

- 648 instances were correctly predicted as 'low'.
- No instance was mistakenly predicted as 'high'.
- 143 instances were mistakenly predicted as 'medium'.

Predicted 'medium' class:

Out of the instances predicted as 'medium' (211 + 267 + 2099 = 2577):

- 2099 instances were correctly predicted as 'medium'.
- 211 instances were mistakenly predicted as 'high'.
- 267 instances were mistakenly predicted as 'low'.

This table, similar to the confusion matrix, helps to assess the model's performance by illustrating the types of errors it makes when predicting different classes. It gives a detailed breakdown of how the model's predictions align with the actual classes in the test dataset.

```
## [1] 0.8207263
```

Based on the results from using the model to predict our test data, we can have the accuracy rate of the model as approximately 0.821. This suggests that the model correctly predicts the `price_rate` category for approximately 82.1% of the instances in the test dataset.

IV. Limitations & Conclusions

One notable limitation of this study is the reliance on a single dataset from King County within a specific time frame. While this dataset provides valuable insights into housing prices within the region, it may not capture the full spectrum of housing market dynamics in other geographical areas or over different time periods. Local factors, economic conditions, and market trends unique to King County may limit the generalizability of our findings to broader contexts. Therefore, caution should be exercised when applying the conclusions drawn from this study to other real estate markets with distinct characteristics. Future research could benefit from incorporating multiple datasets or extending the analysis to diverse geographic regions for a more comprehensive understanding of housing price determinants.

In conclusion, our analysis and test statistics of housing data in King County have yielded valuable insights into the determinants of house prices. Location, specifically zip codes, plays a pivotal role, with the 98039 area commanding the highest average housing price, followed closely by areas 98004, 98040, and 98112. Besides that, the presence of a waterfront view as well as the number of functioning rooms also impact prices. Moreover, our investigation revealed a positive correlation between house prices and interior living space, grade, and above-ground living area. These findings contribute to a nuanced understanding of the factors influencing housing prices in King County, providing valuable insights for both buyers and sellers in the markets of King County.

V. Reference

Harlfoxem. "House Sales in King County, USA." Kaggle, August 25, 2016. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data.