STAT 425 Project - Statistical Analysis of King County Housing Data

Sreekanth Krishnaiah

1. Introduction

In this project, we study the King County housing market from the King County housing dataset. The objective is to determine the relationship between the house price in King County and the house features through statistical inference. The content of the report is organised as follows. In section 2, we work on the King County housing dataset. We explore and explain the dataset. Also, we preprocess and tranform the data for statistical analysis. In section 3, we present our first result. We construct a linear regression model which explain the house price in King County. From the model, we identify the significant features which determine the house price. In section 4, we present our second result. We study the pairwise difference in the mean house price from different groups, from which we gives recommendations in boosting the house price. In section 5, we present our third result. We suggest the factors that can increase the selling price of a house in affluent and poor areas. Besides, a data visualization dashboard has also been constructed to better visualize the data.

2. Data Exploration and Preprocessing

2.1 Dataset

In this project, we work on the King County housing dataset. King County located in Seattle of Washington. It is the most populous county in Washington, where the population is about 2100000. Here, we focus on the King County housing data available in https://www.kaggle.com/harlfoxem/housesalesprediction. In the King County housing dataset, there are n=21613 samples of residential houses. Each sample is labelled by its house price together with 20 features:

```
## All features
```

```
##  [1] "id"            "date"          "bedrooms"      "bathrooms"
##  [5] "sqft_living"   "sqft_lot"      "floors"        "waterfront"
##  [9] "view"          "condition"     "grade"         "sqft_above"
## [13] "sqft_basement" "yr_built"      "yr_renovated"  "zipcode"
## [17] "lat"           "long"          "sqft_living15" "sqft_lot15"
```

Below is a brief description of the features in the King County dataset:

1. id - Each house in the county is given a unique ID.
2. date - Date at which the house was sold.
3. bedrooms - No. of bedrooms in the house.
4. bathrooms - No. of bathrooms per bedroom.
5. sqft_living - The total square footage of the house
6. sqft_lot - Lot size of the house.
7. floors - No. of floors in the house.
8. waterfront - Indication of whether the house has a view to waterfront.

9. view - Indication of whether the house has been viewed.
10. condition - Rating of the overall condition.
11. grade - Rating of the oerall grade.
12. sqft_above - Square footage of the house apart from basement
13. sqft_basement - Size of the basement
14. yr_built - The year in which it was built.
15. yr_renovated - The year in which it was renovated.
16. zipcode - Postal code of the house.
17. lat - The latitude location of the house.
18. long - The longitude location of the house.
19. sqft_living15 - The house square footage in 2015.
20. sqft_lot15 - The lot square footage in 2015.

2.2 Data Preprocess

Here, we are going to preprocess the King county housing dataset. As part of the preprocessing of the data, the following tasks have been performed:

1. We eliminate the "id" which does not have much meaning.
2. We code "date" to an ordinal numerical feature "time" in unit of year.
3. We replace the value of "bedrooms" of a sample with an extodinary value 33 by the median.
4. We eliminate "sqft_living" due to the exact collinearity "sqft_living"="sqft_above"+"sqft_basement".
5. We code "floors" as a categorical feature with integral levels 1,2,3.
6. We add a categorical feature "addhalffloor" to indicate if 0.5 floor is added on the house.
7. We add a categorical feature "basement" to indicate if a basment is built.
8. We code "yr_renovated" as a categorical feature "renovated" to indicate if the house is renovated.

After the data preprocessing, there are totally 20 features, in which there are 11 numerical features:

```
## Numerical features
```

```
##  [1] "bedrooms"      "bathrooms"     "sqft_lot"      "sqft_above"
##  [5] "sqft_basement" "yr_built"      "lat"           "long"
##  [9] "sqft_living15" "sqft_lot15"    "time"
```

and 9 categorical features:

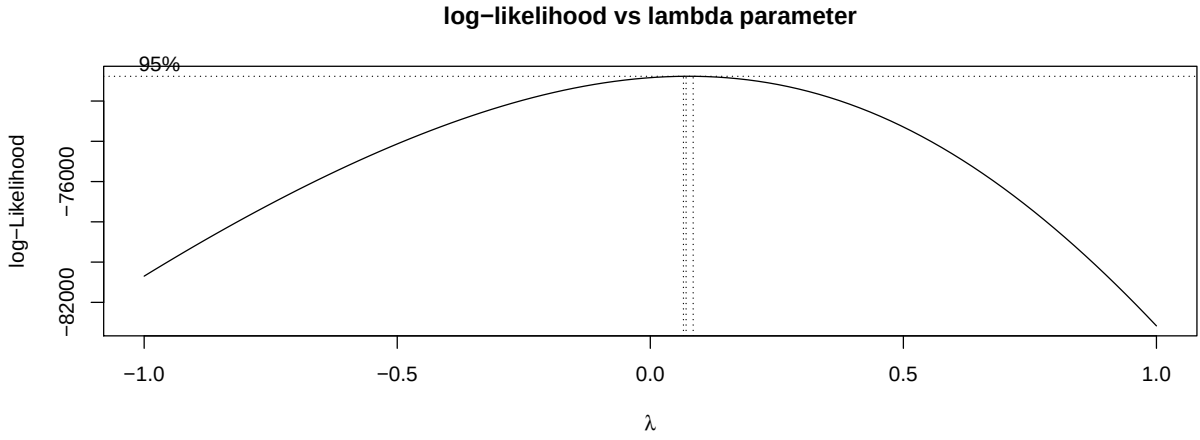
```
## Categorical features
```

```
## [1] "floors"       "waterfront"   "view"         "condition"
## [5] "grade"        "zipcode"      "addhalffloor" "basement"
## [9] "renovated"
```

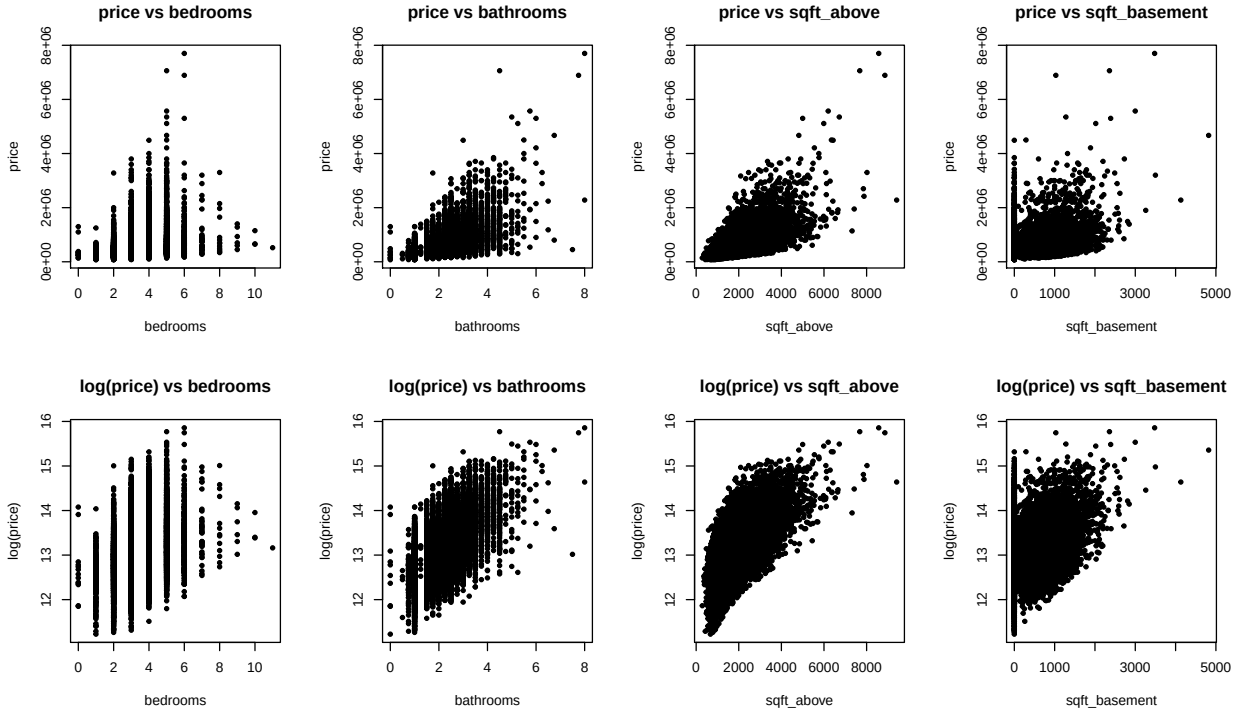
2.3 Data Transformation

Here, we perform appropriate transformation on the housing data. On looking at the distribution of the price, it is quite evident that it is non-linear and is of non-constant varaince. In order to achieve the linearity and constant variance, we perform a box-cox transformation on the house price.

The MLE for the Box-Cox parameter is λ=0.07. So we simply take the closest integer value λ=0, which is to



take log on the price.
Below, we show the changes of house price against several variables before and after the logarithmic transformation. Better linearity and non constant varaince are achieved after the Box-Cox transformation. We will stick to this logorithmic price in the remining of the discussion.



3. Result I: The Model for House Price

Here, we are going to get at a linear regression model which explain the house price. Once we arrive at the linear regression model, we will explain how the features in the model affect the house price. We carry out the following procedure to get at our model:

1. We eliminate highly correlted variables from the VIF. Below shows the VIF of the features:

```
##                       GVIF Df GVIF^(1/(2*Df))
## bedrooms          1.868745  1        1.367021
## bathrooms         3.541036  1        1.881764
## sqft_lot          2.127810  1        1.458702
## floors            3.642775  2        1.381523
## waterfront        1.625406  1        1.274914
## view              2.100294  4        1.097198
## condition         1.585549  4        1.059309
## grade             6.509182 11        1.088876
## sqft_above        6.075212  1        2.464794
## sqft_basement     4.157197  1        2.038921
## yr_built          3.631558  1        1.905665
## zipcode       10267.107051 69        1.069223
## lat              64.698393  1        8.043531
## long             34.642550  1        5.885792
## sqft_living15     3.358056  1        1.832500
## sqft_lot15        2.285674  1        1.511845
## time              1.012772  1        1.006366
## addhalffloor      1.325666  1        1.151376
## basement          3.694800  1        1.922186
## renovated         1.197860  1        1.094468
```

Highly correlated variables generally exhibit VIF >> 1. The VIF above shows that "zipcode", "long" and "lat" are highly correlated, which is as expected since "zipcode" encodes the location information of "long" and "lat". Since "zipcode" is more informative than "long" and "lat", we choose to eliminate the features "long" and "lat" in the later analysis.

2. We check for the relevant numerical features by the T-test. We perform a temporary linear regression on the housing dataset. The p-value for the T-statistics of "bedrooms" and "sqft_lot15" are given below:

```
## p-values
```

```
##   bedrooms sqft_lot15
## 0.07593311 0.42351639
```

Note that both of the p-values > 0.05, meaning that they do not have significant linear relation witht the logged price. Hence, we eliminate the features "bedrooms" and "sqft_lot15". While it can be checked that marginally the logged price of a house increases with the number of bedrooms, the feature does not shows to be significant in the multiple regression model because "bedrooms" is correlated with other variables like "bathrooms" and "sqft_living".

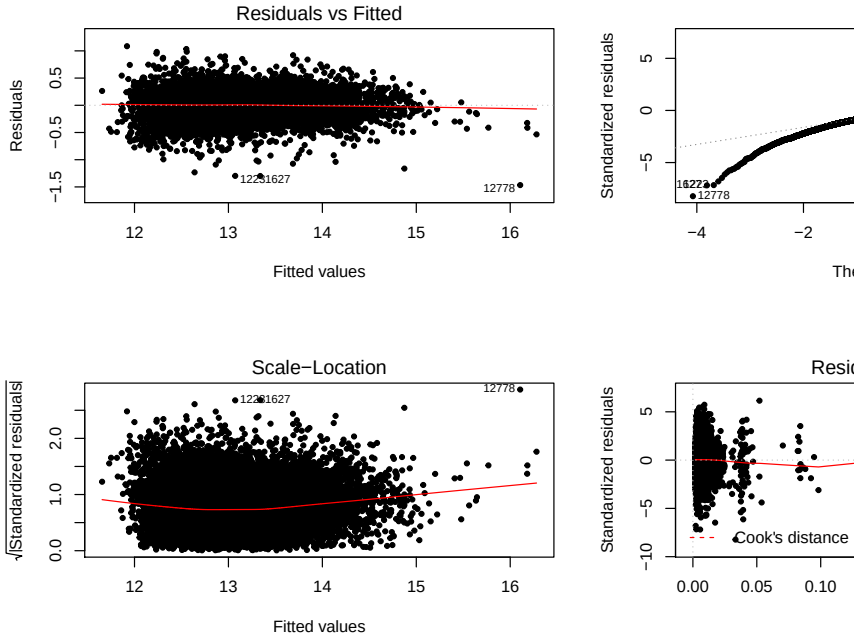
3. We check for the relevant categorical features by the F-test. We again perform a temporary linear regression. Then we perform the anova on the temporary model. The anova table is given as below:

```
## Analysis of Variance Table
##
## Response: log(price)
##                Df  Sum Sq Mean Sq    F value    Pr(>F)
## bathrooms       1 1819.87 1819.87 55398.3982 < 2.2e-16 ***
```

```
## sqft_lot          1   15.91   15.91   484.1874 < 2.2e-16 ***
## floors            2    0.28    0.14     4.3027   0.01354 *
## waterfront        1  116.12  116.12  3534.8611 < 2.2e-16 ***
## view              4  278.83   69.71  2121.9603 < 2.2e-16 ***
## condition         4   66.81   16.70   508.4495 < 2.2e-16 ***
## grade            11 1085.28   98.66  3003.3529 < 2.2e-16 ***
## sqft_above        1   60.90   60.90  1853.7835 < 2.2e-16 ***
## sqft_basement     1  144.36  144.36  4394.4193 < 2.2e-16 ***
## yr_built          1  317.68  317.68  9670.3860 < 2.2e-16 ***
## zipcode          69 1347.04   19.52   594.2762 < 2.2e-16 ***
## sqft_living15     1   19.96   19.96   607.7002 < 2.2e-16 ***
## time              1   12.20   12.20   371.5306 < 2.2e-16 ***
## addhalffloor      1    0.17    0.17     5.1413   0.02337 *
## basement          1    1.94    1.94    59.0388 1.612e-14 ***
## renovated         1    4.23    4.23   128.6623 < 2.2e-16 ***
## Residuals     21511  706.65    0.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistics of all features has value < 0.05. Hence, we do not do further elimination on the categorical features.

4. We perform the stepwise feature selection on top the previous feature selection based on the hypothesis testing. More precisely, we perform the stepwise model selection with AIC and BIC. The feature selection with AIC do not further eliminate any feature whereas the BIC eliminate the feature "addhalffloor" in the selection. Note that the feature "addhalffloor" also does not show to be very significant in the previous F-test. Nevertheless, intuitively "addhalffloor" should contribute to the house price. So we adopt the selection by AIC where no feature is eliminated.



5. We perform diagnostic checking on our model.

```
## Number of high leverage points =  1188
```

```
## Number of outliers =  35
```

```
## Number of influential points =  1
```

It can be seen that the constant variance and the normality is approximately valid. Besides, from the analysis of leverages, we see that there are 1188 high leverage points. From the T-test of the studentized residuals, we see that there are 35 outliers. From the cook's distance, we see that there is 1 influential point.

The result here is that we get a linear regression model that explain the King County house price with 16 significant features. Within the 16 significant features in the linear regression model, 7 features are numerical features:

```
## Numerical features
```

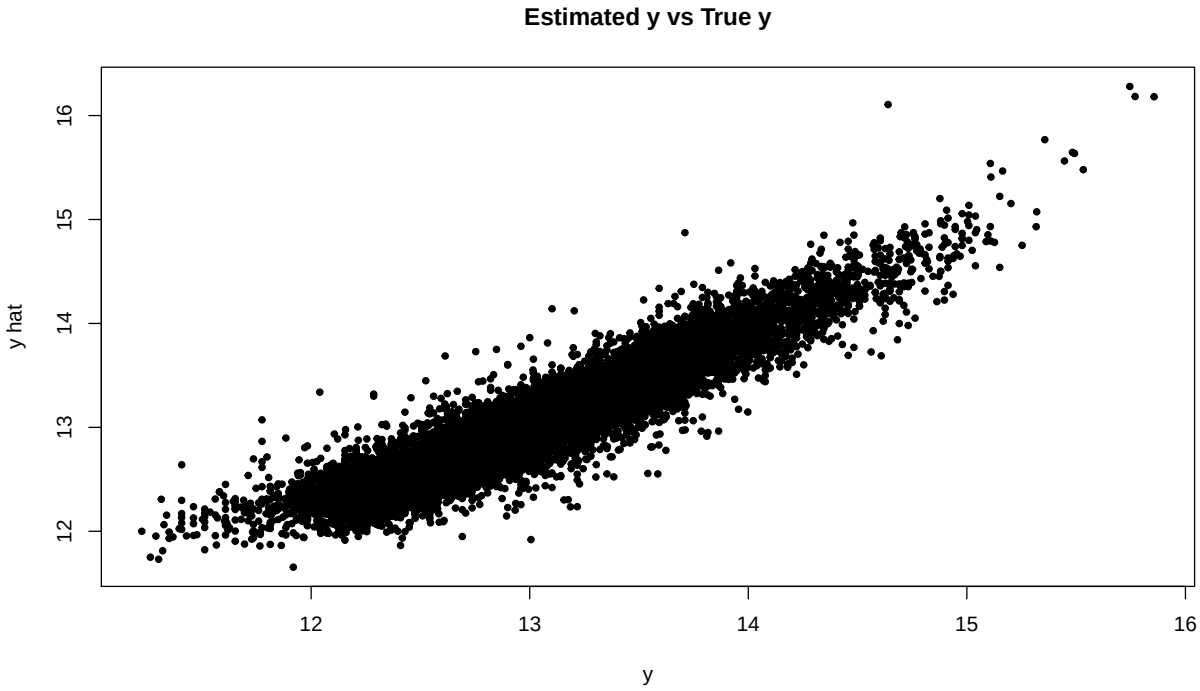
```
## [1] "bathrooms"     "sqft_lot"      "sqft_above"    "sqft_basement"
## [5] "yr_built"      "sqft_living15" "time"
```

whereas the remaining 9 features are categorical features:

```
## Categorical features
```

```
## [1] "floors"       "waterfront"   "view"         "condition"
## [5] "grade"        "zipcode"      "addhalffloor" "basement"
## [9] "renovated"
```

where "waterfront", "addhalffloor", "basement" and "renovated" takes only on values 0 or 1. The linear regression model explains a large fraction of variance of the logged price with $R^2_{adj} = 0.881$. The large R^2_{adj} for the linear regression model is visualized by the high correlation between the logged price y and the estimated logged



price $\hat{y}$ shown below

Now, we interpret the coefficients in the linear regression model. For the numerical features, the coefficients are given by

```
## Coefficients of numerical features
```

```
##     bathrooms      sqft_lot    sqft_above sqft_basement      yr_built
##      3.54e-02      6.55e-07      2.10e-04      9.72e-05     -4.06e-04
## sqft_living15          time
##      8.31e-05      7.81e-02
```

The coefficients of "bathrooms", "sqft_lot", "sqft_above", "sqft_basement" and "sqft_living15" are positive because they measure the house quality. The coefficients of "time" is positive because the house price increase with time. The negative coefficient for "yr_built" is counter intuitive since the house price should decrease as its age increases. Such negativity can be explained by the "yr_built" and other features. Besides, the comparsion of magnitudes of coefficients is meanningful for "sqft_above","sqft_living15" and "sqft_lot" since they are of the same unit. The order of magnitudes of coefficients is given by "sqft_above" > "sqft_living15" > "sqft_lot". Such order indicates their important in determining the house price. For the categorical features which takes on values 0 and 1, the coefficients are given by

```
## Coefficients of categorical features
```

```
##    waterfront addhalffloor1     basement1    renovated1
##    0.44327175    0.01333445    0.03820276    0.07594021
```

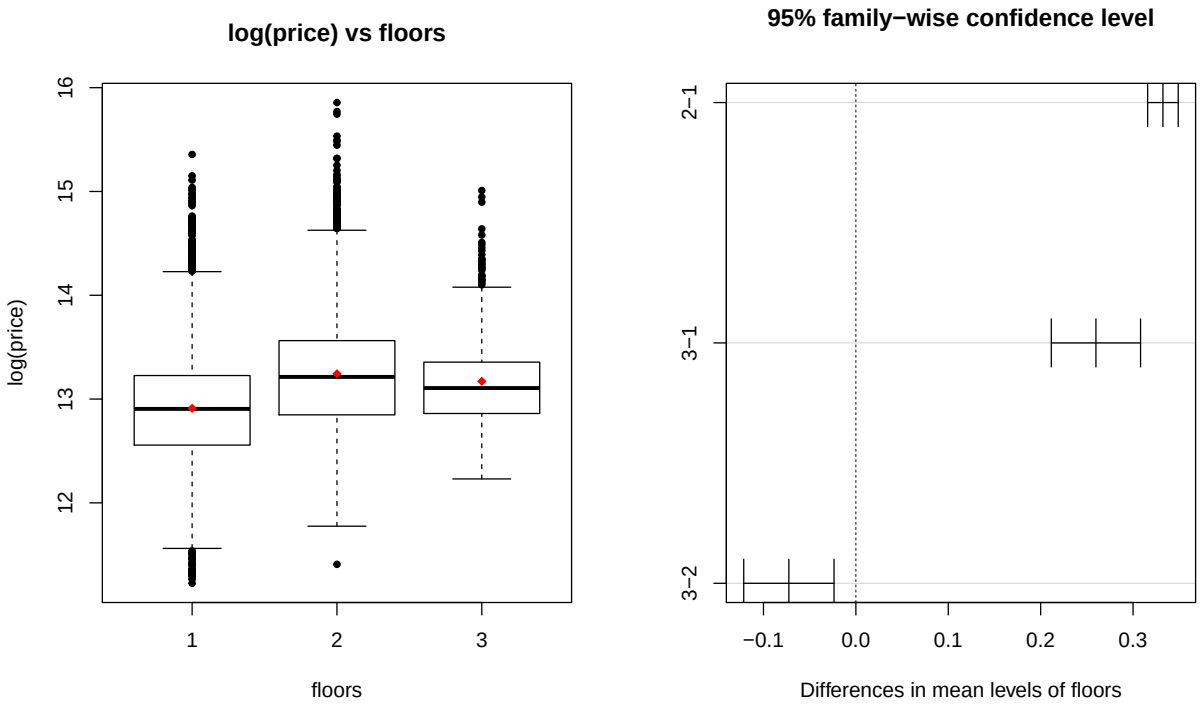
The coefficients of "waterfront", "addhalffloor", "basement" and "renovated" is positive because they are all improvement of the house. The comparsion of magnitudes of coefficients is meanningful for the categorical features which takes on values 0 and 1. we have the order of magnitudes of coefficients given by "waterfront" > "renovated" > "basement" > "addhalffloor" which indicates their important in determining the house price.

4. Result II: Boosting the House Price

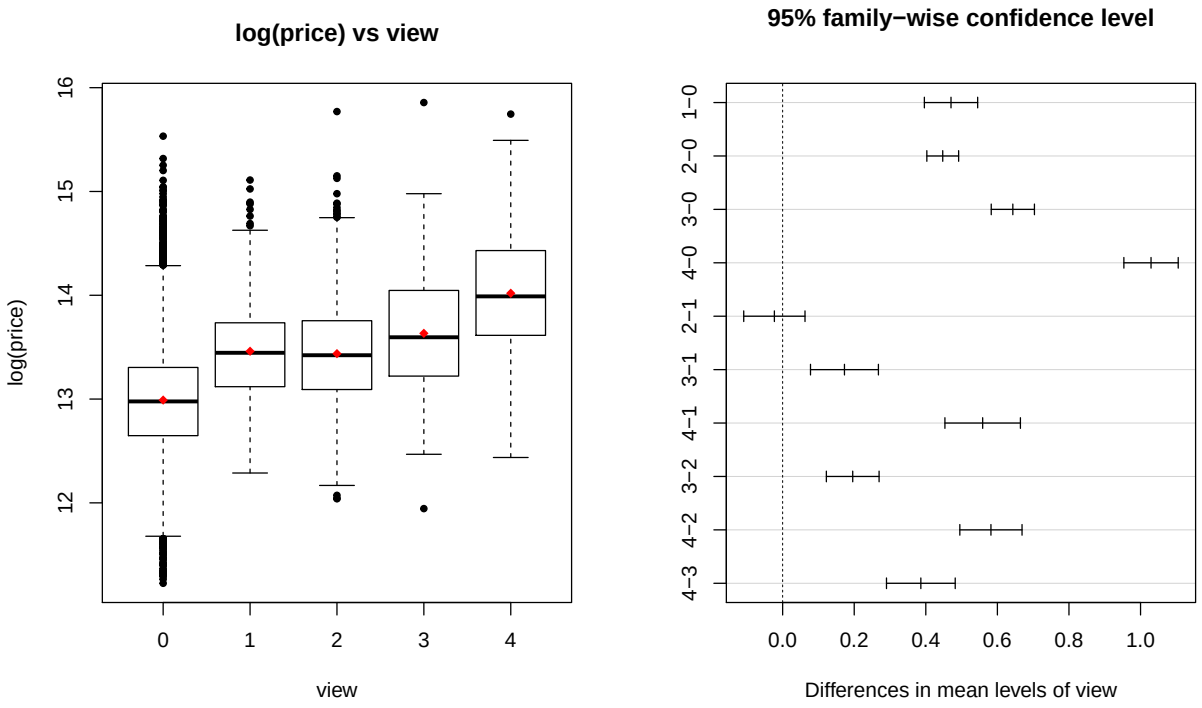
Here, we are going to suggest methods to boost the house price. We explore if there is any difference in the mean of logged prices in different groups for the categorical features "floors", "view", "condition" and "grade". A pairwise comparision is applied to determine this. From the pairwise comparision, we are going to draw conclusion and recommendation to increase the house price.

In the following, we give the box plot of the logged price against the feature of interest, in which the red dot indicates the mean logged price for each group. Each box plot is accompanied with a visualization of the confidence intervals for the Tukey's honest significant difference in group mean. Conclusion and recommendation are draw from the statistical analysis.

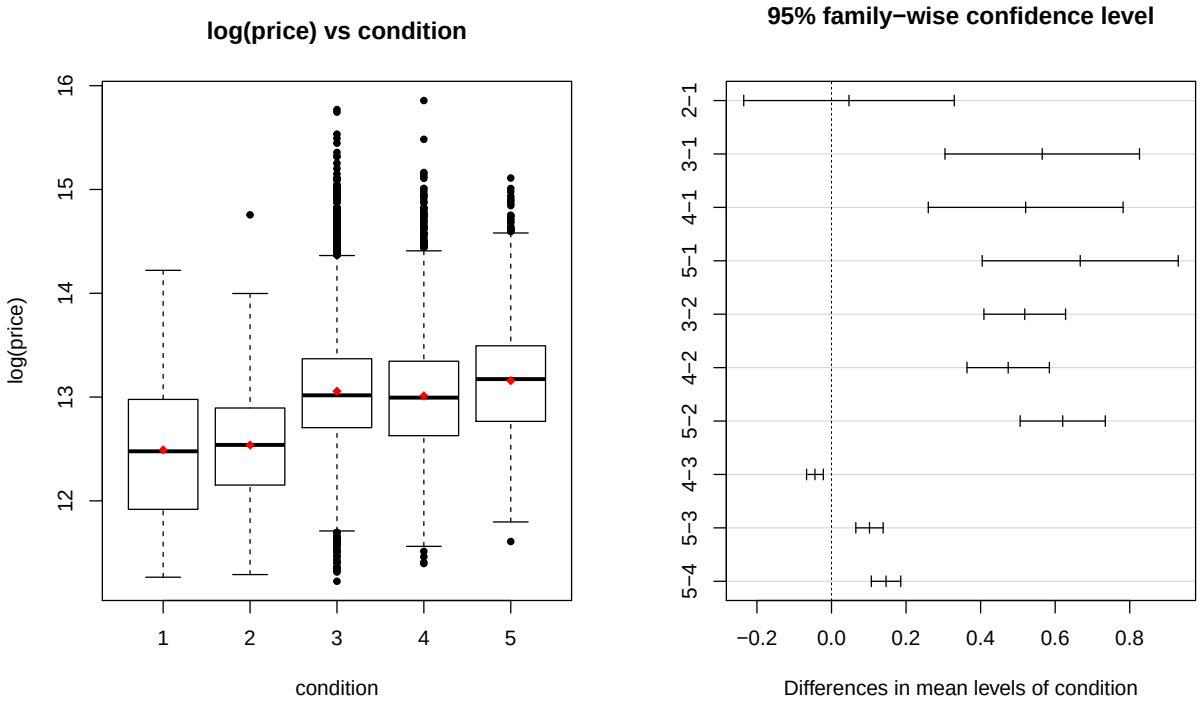
1. "floors": From the boxplot, we can see that the group mean goes up from "floors"=1 to "floors"=2 but goes down from "floors"=2 to "floors"=3. Such trends of group mean is further confirmed by the plot of confidence intervals. The data suggests that increasing the number of floors does not generally increase the house price. While increasing from "floors"=1 to "floors"=2 boost the house price, increasing from "floors"=2 to "floors"=3 does not.



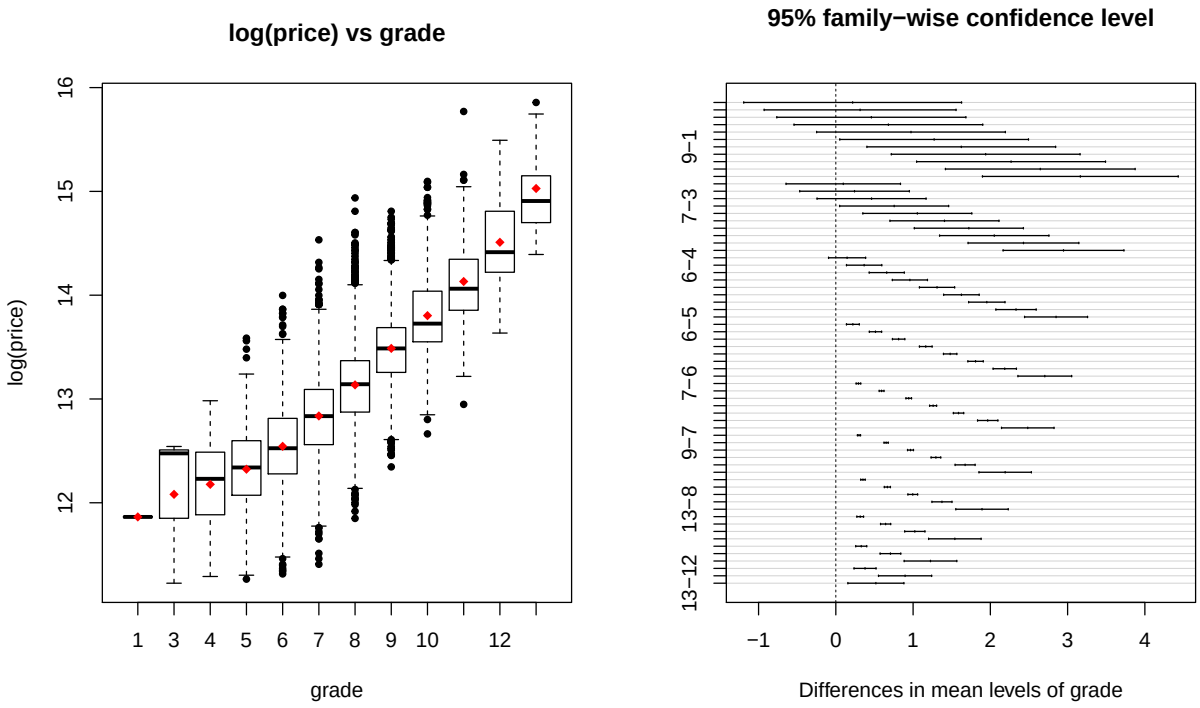
2. "view": From the box plot, we see that the mean of logged price increases generally as the house got more viewed. But the mean of logged price stays almost the same from "view"=1 to "view"=2. The plot of confidence intervals shows that while there is a jump in the price from "view"=0 to "view"=1, there is no significant difference in the mean from "view"=1 to "view"=2. So to boost the house price, the house should at least be viewed once. Viewing the house two times and three times does not leads to a significant change in the house price.



3. "condition": The box plot shows that the mean of logged price increases generally if the house has a better condition. However, the mean price stays almost the same from "condition"=1 to "condition"=2. The confidence intervals on the pairwise differences also suggest that the groups "condition"=1 and "condition"=2 do not lead to a significant different in the mean price, whereas there is a significant improve in the mean price from "condition"=2 and "condition"=3. So to boost the price by increasing the "condition", the "condition" should at least be increased to 3.



4. "grade": From the box plot, we see that the mean of logged price generally increasesas the house got more viewed. If we look further into the confidence interval for the Tukey's honest significant difference, we see that actually the mutual difference in the mean for "grade" = 1 to 5 are not significant. Hence grade 1 to 5 do not make significant change in the house price. To boost the house price, the house should be at least of grade 6.



5. Result III: Specific recommendations for selling house in affluent and poor areas

Our main object of analysis in this section is to suggest factors that can help one increase the selling price of a house in affluent and poor areas. Our assumption is that rich neighbourhoods and poor neighbourhoods might have different factors affecting the selling price of the house. We identify rich and poor neighbourhoods depending on the average price of the zipcodes. The zipcodes with top five average prices were considered as "rich neighbourhoods" and the zipcodes with least five zipcodes were considered as "poor neighbourhoods". We first look into the "grade" of the house. The below is a the brief description on how the house is graded.

1-3 Falls short of minimum building standards. Normally cabin or inferior structure.

4 Generally older, low quality construction. Does not meet code.

5 Low construction costs and workmanship. Small, simple design.

6 Lowest grade currently meeting building code. Low quality materials and simple designs.

7 Average grade of construction and design. Commonly seen in plats and older sub-divisions.

8 Just above average in construction and design. Usually better materials in both the exterior and interior finish work.

9 Better architectural design with extra interior and exterior design and quality.

10 Homes of this quality generally have high quality features. Finish work is better and more design quality is seen in the floor plans. Generally have a larger square footage.

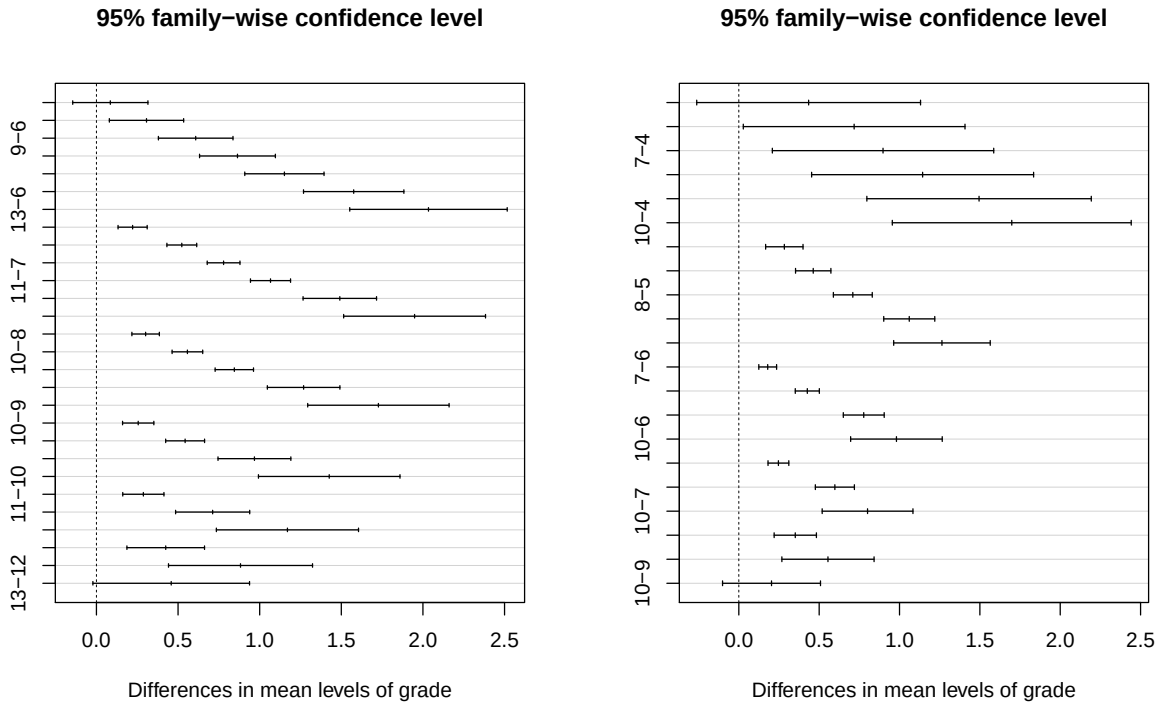
11 Custom design and higher quality finish work with added amenities of solid woods, bathroom fixtures and more luxurious options.

12 Custom design and excellent builders. All materials are of the highest quality and all conveniences are present.

13 Generally custom designed and built. Mansion level. Large amount of highest quality cabinet work, wood trim, marble, entry ways etc.

```
## [1] "98002" "98168" "98032" "98001" "98148"
```

```
## [1] "98102" "98112" "98040" "98004" "98039"
```



We perform a one way ANOVA on the data from both the regions. Although, the "grade" of a house is significant, we wanted to explore if there was any significant difference in the mean prices of houses for different grades and if this difference was similar for rich and poor neighbourhoods.

Key inferences :

1. Firstly, none of the houses in both the rich and poor neighbourhoods have houses rated between 1 and 3. This makes sense as any house that falls short of minimum building standards isnt being sold in the market irrespective of the neighbourhood.
2. It is interesting to note that although the houses with grade 4 and 5 have been sold in the poor neighbourhoods, there werent any houses which were sold with similar grades in the rich neighbourhood. We can fairly assume that the buildings with this grade which are generally older, low quality construction and simple design and are most likely to be located in the poor neighbourhoods. It makes sense that rich neighbourhoods wouldnt have any houses with this grade.
3. Also, according the results from the Tukey's plot, it doesnt matter if your house grade is 4 or 5 in the poor neighbourhoods. They are likely going to be sold for a similar price probably because they are equally bad. However, if try to get the grade to 5, ther is a significant increase in the price of your house.
4. If you want to make a better deal if your house is in the poor neighbourhood, getting to improve the grade from 9 to 10 alone will not lead to higher selling price. Improving the grade from 9 to 11 or 10 to 11 is more likely to lead to a significant increase in the selling price.
5. In the rich neighbourhood, the scenario is different. The selling price is not going to increase significantly from 6 to 7 i.e from "average grade" to an "above average grade". However, improving the grade from 6 to 8 or 7 to 8 is more likely to lead to a significant increase in the selling price. We can fairly conclude that if people are going to invest in the rich neighbourhood areas, they might as well see more than just an increase from "average" to "above average"

R Shiny UI Dashboard

The following is the link to the interactive dashboard that has been constructed in R Shiny. https://sreekanth29.shinyapps.io/kings_county_analysis/ It contains three tabs - "Graphs", "Data" and "Maps". In the first tab "graphs", we have two graphs - the first one is a trend graph between the average price and different variables and the second shows a box plot between price and different variables. The second tab "Data" contains two sections - the first one shows all the observations based on the selections that have been made and the second shows the average price againt the variable selected. The third tab "map" shows the map of the king's county with the circles represting different houses in the county which are present in the King's County dataset. All the three componenets can be subsetted according the five variables - view, condition, floors, month, sqft_living and zip code.

Conclusion

In this project, we study the King County house price from the King County Housing dataset. By going through the data preprocessing, data transformation and feature selection processes, we build a linear regression model to explain the house price. We end up with 7 significant numerical features "bathrooms", "sqft_lot", "sqft_above", "sqft_basement", "yr_built", "sqft_living15", "time" and 9 significant categorical features :"waterfront", "addhalffloor1", "basement1", "renovated1". We also perform diagnostic checking to justify the validity of our model. In addition, we also made suggestion in boosting the house price by looking into the mean price versus different levels for a given feature. For example, we see that Houses with "grade" 1 to 5 are of the same mean price, so one have to made the house to be of least "grade" 6 so as to increase the house selling price. Beside, we also deepen our analysis by looking into the "rich" area and "poor" area. Comments specific to the "rich" area and "poor" area in increasing the house price are made. Finally, to visualize the data well, we developed an interactive user interface for data visualization.

Along this line of work, one could also use the model we built for predicting house price. One can use the significant variables we selected in the statistical analysis to build a more sophisticated model for house price prediction. For example, nonlinear regression, random forest model and ensemble models could be used instead of the linear regression model. It would be interesting to promote the model built to real market prediction and see its implications.