AN INVESTIGATION OF THE FACTORS THAT INFLUENCE HOUSE PRICES IN KING COUNTY, USA.

[redacted]

[redacted]

An investigation of the factors that influence house prices in King County, USA.

[redacted]

[redacted]

1.Introduction

Owning a house is the main aim of individuals. Not only can it enhance the quality of life, studies have also shown that homeownership creates benefits not only for the individual and their families, but also for the surrounding areas and communities. (Kamal, Hassan & Osmadi, 2016). However, affordability has become a critical issue for prospective buyers. In Ireland alone, residential property prices rose 12.2% in the year to August 2017, this being the largest increase in housing prices since June 2015. Further afield in King County, USA, house prices increased 18.2% in the year to August 2017. Homes in King County, USA will for the basis of the analysis presented in this report.

Studies have found that two main factors contribute towards the main concerns of individuals over the affordability of housing: 1) housing is the single largest investment of most individuals and 2) many metropolitan areas have experienced an increase in house prices, (Quigley and Raphael, 2004).

Many factors influence house prices. MacDonald (2011) grouped the factors into five categories:

1. Supply side factors: A decrease in land supply but increase in housing demand
2. Economic factors
3. Industry factors: increased cost of construction materials
4. Physical connectivity and internationalization: the improvement of physical connectivity
5. Demographic factors: the increase in population growth

There are numerous studies which identify the individual factors influencing house prices, and these have been widely discussed and modelled through the literature. Research has shown that that house prices may be influenced by the proximity of schools (Des Rosiers et al., 2001), construction quality (Ooi et al., 2014), proximity to the Central Business District, house and lot size (Abelson et al., 2013) and location (Stacy et al., 2006), amongst other factors. Turnbull, Dombrow and Sirmans (2006) explored the influence of property size on housing price, suggesting that larger houses will sell for higher rates relative to an otherwise identical house in a similar neighborhood.

Using the King County House Sales Data from 2014/15, this analysis will explore if square footage can be used to predict the house price of homes in the King County Sales dataset, with null (H_0) and alternative (H_A) hypotheses stated below.

> H_0: In the presence of other variables, there will be no significant prediction of house prices by the size of the interior living space
>
> H_A: In the presence of other variables, there will be a significant prediction of house prices by the interior living space of the house

Furthermore, this report will build on the work of Ooi et al. (2014) and explore the relationship between construction quality and house price for this dataset. This report will also contribute to the literature and explore if the presence of additional features, namely a basement, can be used as a predictor for house prices, when considering the King County dataset.

2.Methodology

The dataset used in this analysis contains information about homes sold between May 2014 and May 2015 along with the price in US dollars in King County, Washington, USA. The dataset includes 21 variables for 21,613 house sale records. For this analysis the dataset was cut back to include only houses within the price range $0 - $2,000,000. This reduces the dataset to 21,415 records. The data is of good quality, with no missing values present against any variable.

The variables of interest for this study are “price”, “sqft_living”, “sqft_living15”, and are described in detail in section 2.1 Descriptive Statistics. Two extra variables, “basement” and “grade_grp” were derived and added to the dataset for this analysis. The following variables were removed as they were of no relevance to the analysis:

- Date
- ID
- Latitude
- Longitude

Three models were created for this analysis. The baseline regression model is presented first. Model 2 investigates the differential effect between a house having a basement or no, and finally Model 3 investigates the differential effect between the building grade group the house belongs to.

For this analysis a significance level of 0.05 was chosen.

2.1 Descriptive Statistics

2.1.1 Price

For this analysis, the variable “Price” is the response or dependent variable. It is a scale variable and as Figure 2.1 shows, the data is skewed to the right. Furthermore, the standardized scores for skewness (106.11) and kurtosis (129.04) fall well outside the scores of -2 and 2. As the data is skewed, the median is used as a method of centre measurement. In this case, the median is $450,000 with a minimum price of $75,000 and a maximum price of $2,000,000.

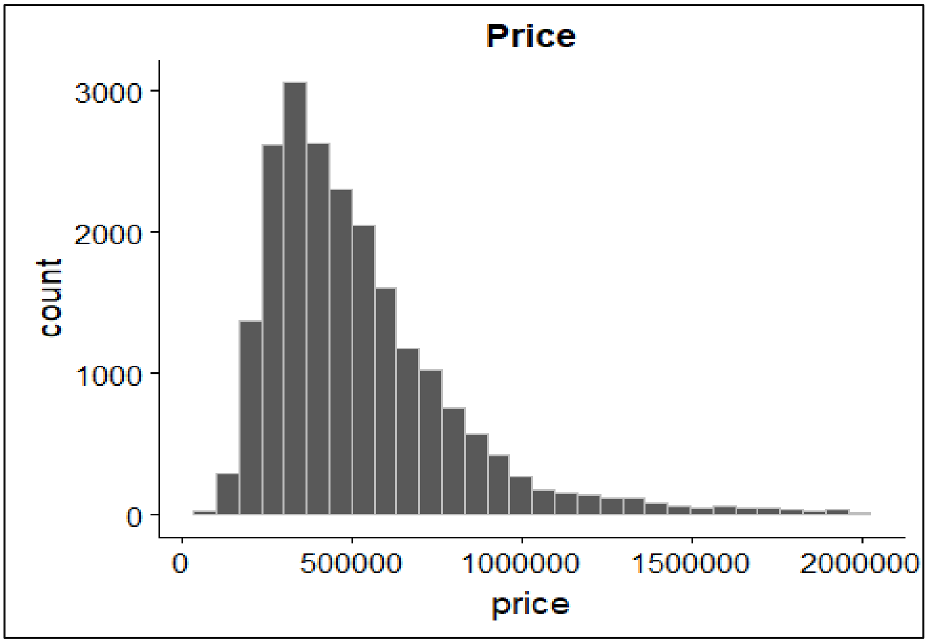


Figure 2.1: Histogram for Price, showing skewness to the right

2.1.2 Sqft_living and Sqft_living15

The variables “sqft_living” and “sqft_living15” are also scale variables, the former describing the square footage of the interior living space of the house and the latter representing the average square footage of the interior living space of the nearest 15 houses. Figure 2.2 highlights that both of these variables are also skewed to the right with standardized skewness and kurtosis values of 64.85 and 57.75 respectively for the variable “sqft_living”, and a skewness value of 64.34 and kurtosis value of 44.85 for “sqft_living15”. As previous, these standardized scores fall outside the scores of -2 and 2. As the data is skewed, the median is used as a method of centre measurement. The variable “sqft_living” has a median of 1900sqft, while “Sqft_living15” has a median of 1830sqft.

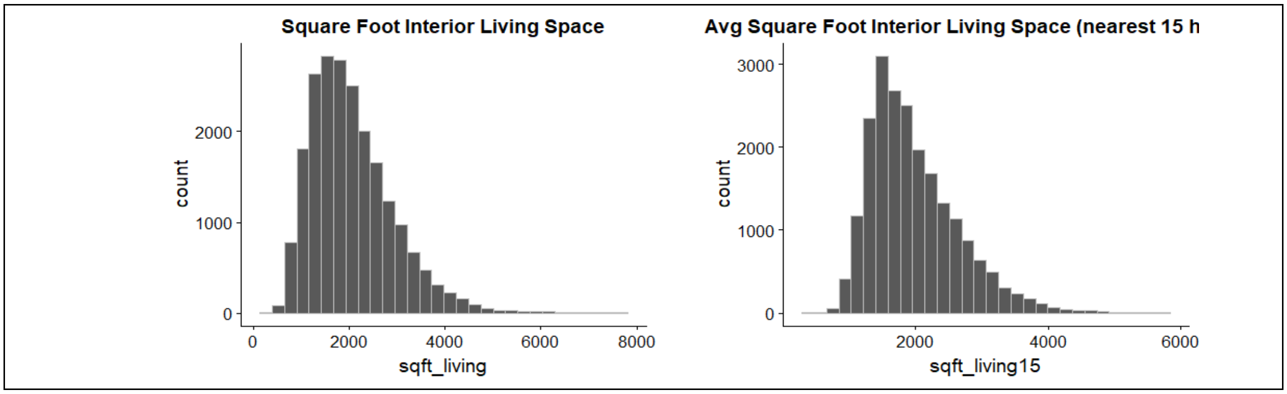
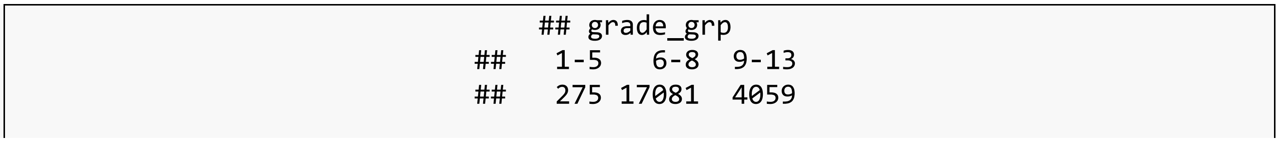


Figure 2.2: Histograms for sqft_living and sqft_living15

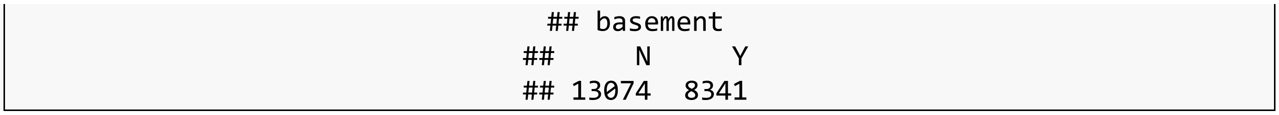
2.1.3 Basement and Grade_Grp

The variable “basement” is a binary variable. It was created using the variable “sqft_basement”, which provides the basement floor area of each house sold. Where “sqft_basement” > 0, “basement” was flagged as ‘Y’ (1), otherwise it was flagged as ‘N’ (0).

“Grade_grp” is an ordinal variable, created using the already existing variable “grade”, an index from 1-13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design. The grades were grouped into three groups for this analysis (1-5, 6-8, 9-13) to form new variable “Grade_grp”. Figure 2.3 shows the sample characteristics for both variables.



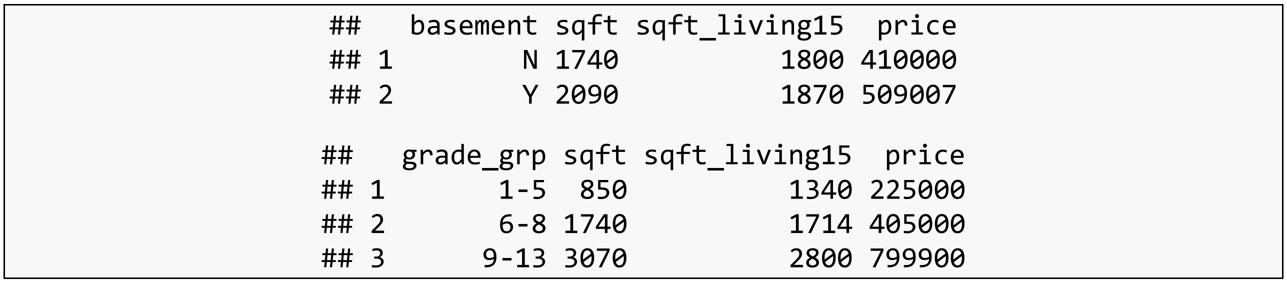
```
         ## grade_grp
## 	1-5   6-8  9-13
## 	275 17081  4059
```



```
## basement
##     N     Y
## 13074  8341
```

Figure 2.3: Characteristics for “grade_grp” and basement

It is also interesting to note. for later comparisons, the medians of the scale variables of interest in relation to the ordinal variables. Figure 2.4 shows the median house price and interior living space broken down by grade group and by whether the house has a basement or not.



```
##   basement sqft sqft_living15  price
## 1        N 1740          1800 410000
## 2        Y 2090          1870 509007

##   grade_grp sqft sqft_living15  price
## 1       1-5  850          1340 225000
## 2       6-8 1740          1714 405000
## 3      9-13 3070          2800 799900
```

Figure 2.4: Median values by basement and grade_grp

The code related to 2.1 Descriptive Statistics is presented in Appendix A for reference.

2.2 Multiple Linear Regression

The method of analysis used is Multiple Linear Regression (MLR). MLR is a method of data analysis that may be appropriate whenever a quantitative variable (the response variable) is to be examined or predicted in relation to any other variables (the predictor variables). Relationships may be non-linear, and the predictors may be quantitative or qualitative. A MLR equation for predicting “y” can be expressed as follows:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots. + b_nx_n + e_i$$

where b_0 is the intercept value and $b_{1\ldots}b_n$ are the regression coefficient for the variables 1 to n. In this analysis, “price” was the quantitative response variable (y) and the predictor variables (b_n) were “sqft_living”, “sqft_living15”, “basement” and “grade_grp”.

3.Results

3.1 Correlations

The relationship between price and the square foot of the interior living space was investigated using a Spearman correlation. A strong positive correlation was found (rho=0.634, n=21415, p<=.001). Using the same test, there was also a strong positive correlation found between price and the average square foot of the interior living space of the nearest 15 houses (rho=0.562, n=21415, p<=.001).

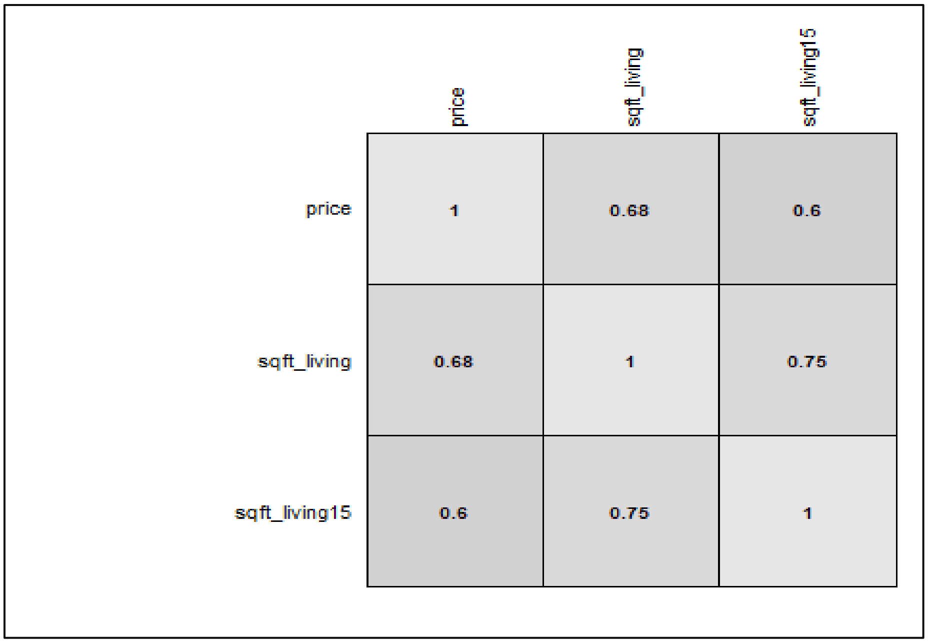
Using an independent t-test it was established that the difference between house prices for houses with a basement and without a basement were different to a statistically significant level (t (21413) = -26.493, p<0.01). The mean difference was substantial ($105,148) with houses with a basement having a higher mean price than houses without a basement.

A one-way ANOVA test was conducted to explore the impact of grade level on the house price. Houses were divided into three groups according to their grade (1-5, 6-8, 9-13). There was a statistically significant difference at the p<0.05 in house prices for the three grade groups (f (21415) = 6018.44, p<0.001). There is a difference of $195,638.49 between the mean price of houses in the groups 1-5 and 6-8, a difference of $629,456.83 between groups 1-5 and 9-13 and a difference of $433,818.34 between groups 6-8 and 9-13.

Full details of all tests conducted are included in Appendix B.

3.2 Baseline Model

The baseline model for this analysis looks to see if the response variable “price” can be predicted by the variable in question, “sqft_living”. The variable “sqft_living15” was also included in the baseline model to give an insight into what kind of effect the surrounding area might have on the house price. Figure 3.1 shows the correlation between these variables. The correlation between “sqft_living” and “sqft_living15” is 0.75. Since this correlation is below 0.8, it is acceptable to proceed with the regression model. The VIF value, 2.323414 is also acceptable.



	price	sqft_living	sqft_living15
price	1	0.68	0.6
sqft_living	0.68	1	0.75
sqft_living15	0.6	0.75	1

Figure 3.1: Correlation plot

As noted in Section 2.1, the response variable “price” was skewed to the right. In order to create the regression model, the variable must be transformed to achieve normality. For this analysis the square root function was used as the transformation method. Figure 3.2 shows the histogram of the transformed variable price. This follows a more normal pattern than before (Figure 2.1).

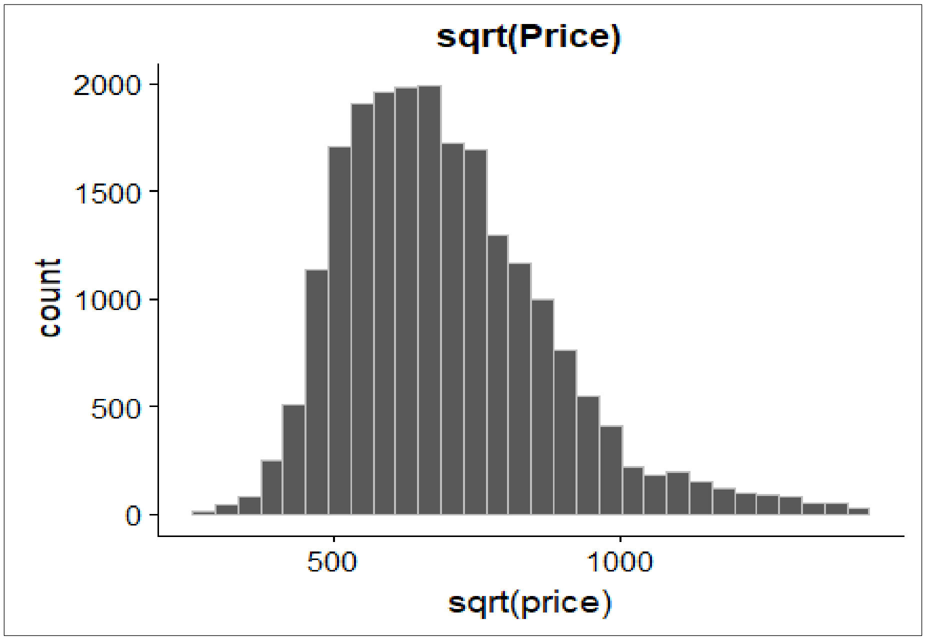


Figure 3.2: Histogram for sqrt(price)

A multiple linear regression was derived to predict house prices based on the square foot of the interior living space of the house and the average square footage of the interior living space of the nearest fifteen houses. A significant regression equation was found

(F (2,21412) = 1.024e+4, p<.001),

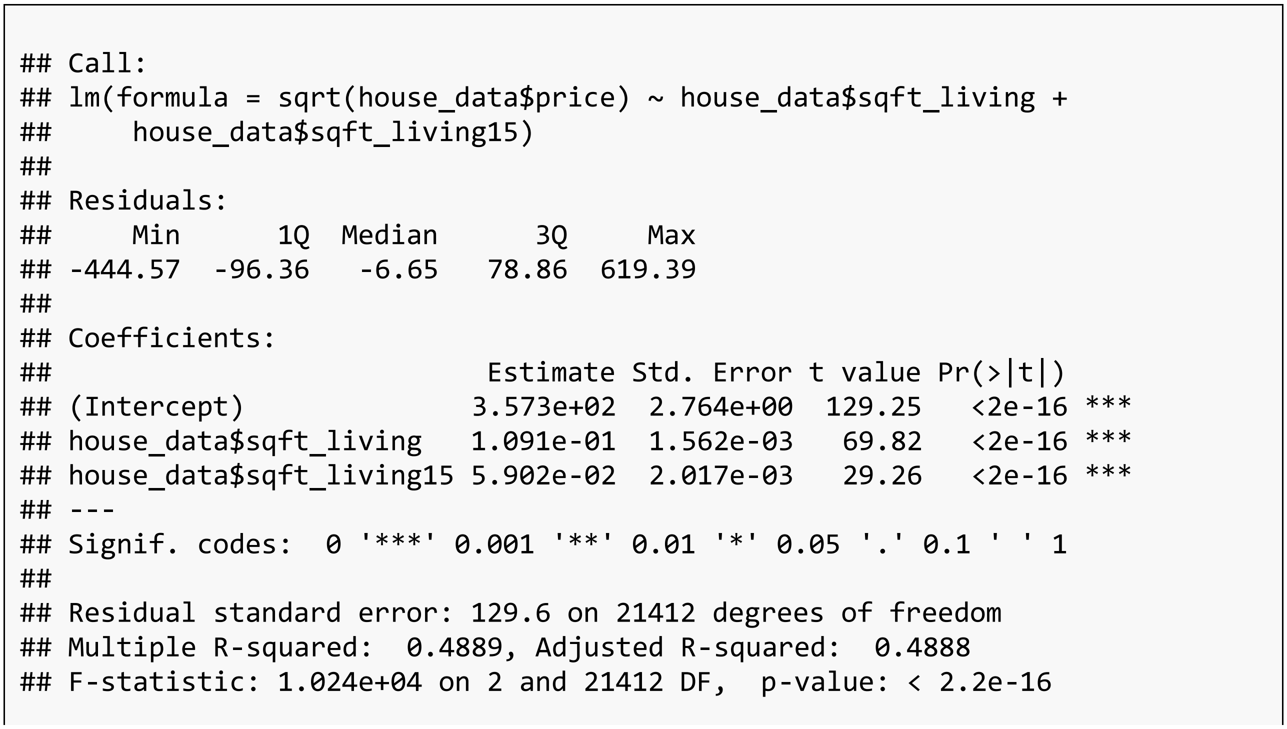
with an R^2 of 0.4888. The house’s predicted price can be calculated using the following regression equation:

(357.268087 + 0.109074(sqft_living) + 0.059023(sqft_living15))2

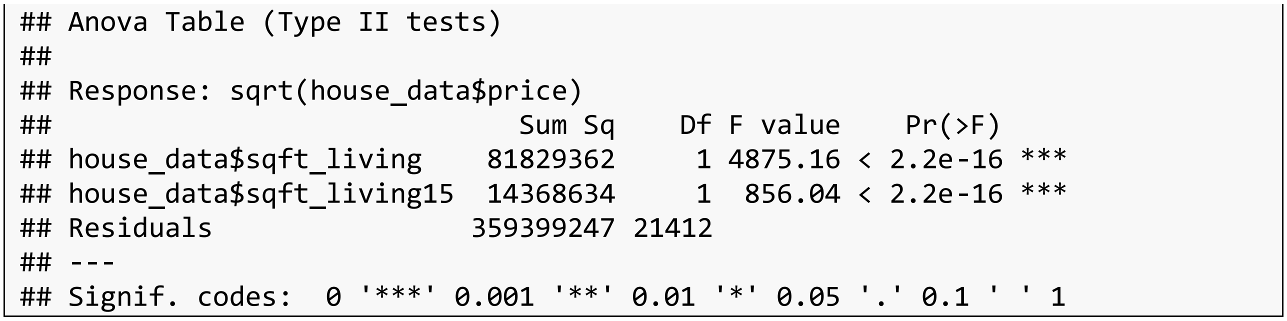
where both variables are measured in square foot. According to the analysis the following results can be deduced:

1. If the interior living space of the house increases, the price of the house increases
2. If the average interior living space of the nearest fifteen houses increases, the price of the house increases

Both “sqft_living” and “sqft_living15” were significant predictors of house prices, explaining 49% of the variance in house price. The R outputs for the baseline model can be seen in Figure 3.3 below.



```
## Call:
## lm(formula = sqrt(house_data$price) ~ house_data$sqft_living +
##     house_data$sqft_living15)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.57  -96.36   -6.65   78.86  619.39
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.573e+02  2.764e+00  129.25   <2e-16 ***
## house_data$sqft_living   1.091e-01  1.562e-03   69.82   <2e-16 ***
## house_data$sqft_living15 5.902e-02  2.017e-03   29.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129.6 on 21412 degrees of freedom
## Multiple R-squared:  0.4889, Adjusted R-squared:  0.4888
## F-statistic: 1.024e+04 on 2 and 21412 DF,  p-value: < 2.2e-16
```



```
## Anova Table (Type II tests)
##
## Response: sqrt(house_data$price)
##                              Sum Sq    Df F value    Pr(>F)
## house_data$sqft_living     81829362     1 4875.16 < 2.2e-16 ***
## house_data$sqft_living15  14368634     1  856.04 < 2.2e-16 ***
## Residuals                359399247 21412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.3: Baseline Model Output

Residual v Fitted plots and Normal Q-Q plots were used to test for normal distributions. The residual plot in Figure 3.4 shows roughly an even number of cases spread around the horizontal at 0, indicating a normal distribution and a linear relationship. In addition, the normal Q-Q plot in Figure 3.5 also shows that the residuals are normally distributed.

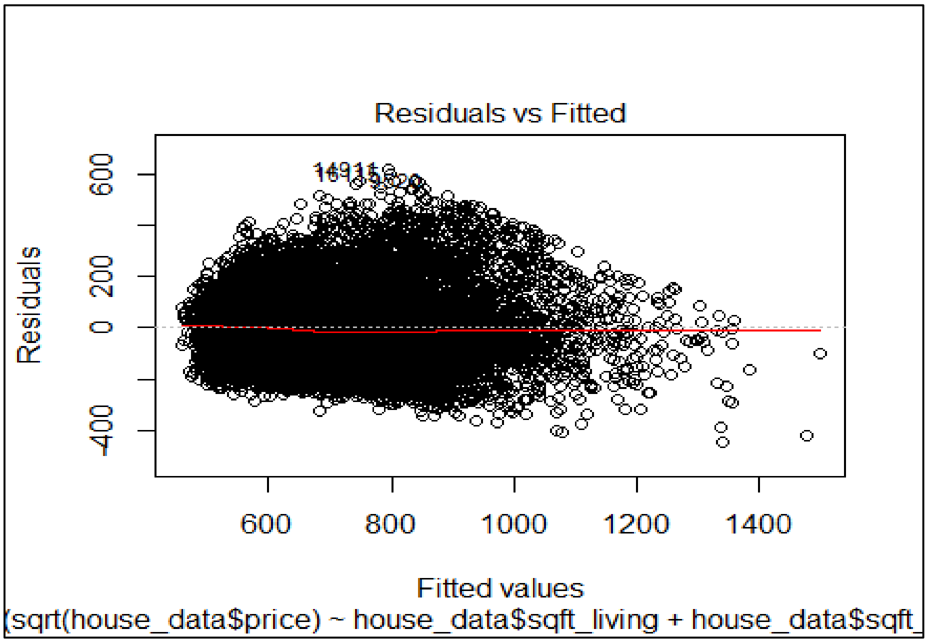


Figure 3.4: Residual V Fitted for Model 1

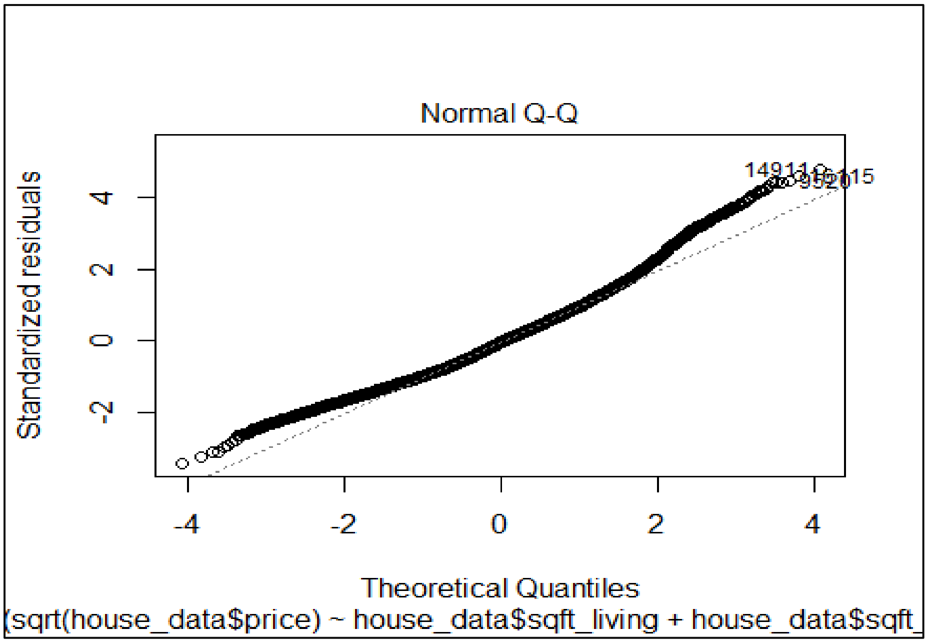


Figure 3.5: Normal Q-Q Plot for Model 1

The Cook's Distance graph in Figure 3.6 highlights three cases that may be influential towards the regression results. Excluding these cases may improve the results but due to the low number of outliers (3) relative to the number of cases in this analysis (21,415), it is not expected to have a significant impact.

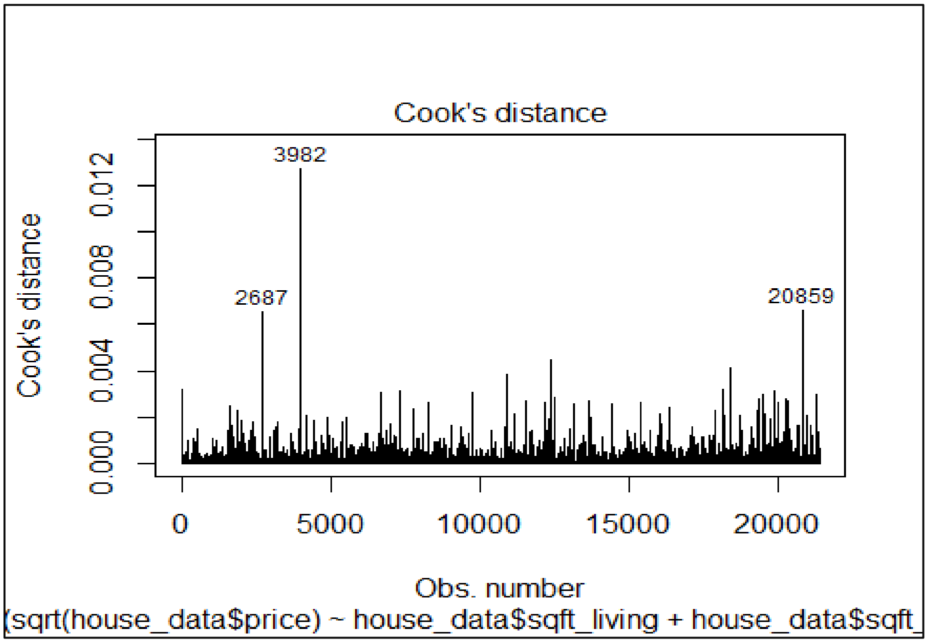


Figure 3.6: Cooks Distance Graph for Model 1

Using the median values for “sqft_living” and “sqft_living15”, it is possible to illustrate the above findings:

$$(357.268087 + 0.109074(1900) + 0.059023(1830))^2$$

$$(357.268087 + 207.1 + 108.01209)^2 = (672.48009)^2 = \$452{,}229.47$$

Based on the above, a house with an interior living space of 1900sqft and the interior living space of the nearest fifteen houses, on average 1830sqft, would cost $452,229.47. Comparing to this to the median house price of $450,000 presented earlier, provides a good indication of the accuracy of this model.

3.3 Model 2

The second model in this analysis investigated if there was a differential effect for having a basement in the house or not. Before creating this model, the variable basement was transformed into a dummy variable. In this case, having a basement (Y) was recoded to 1 (the category of interest) and not having a basement (N) was recoded to 0 (the reference category).

The outputs highlighted in Figure 3.7 show that a significant regression equation was found

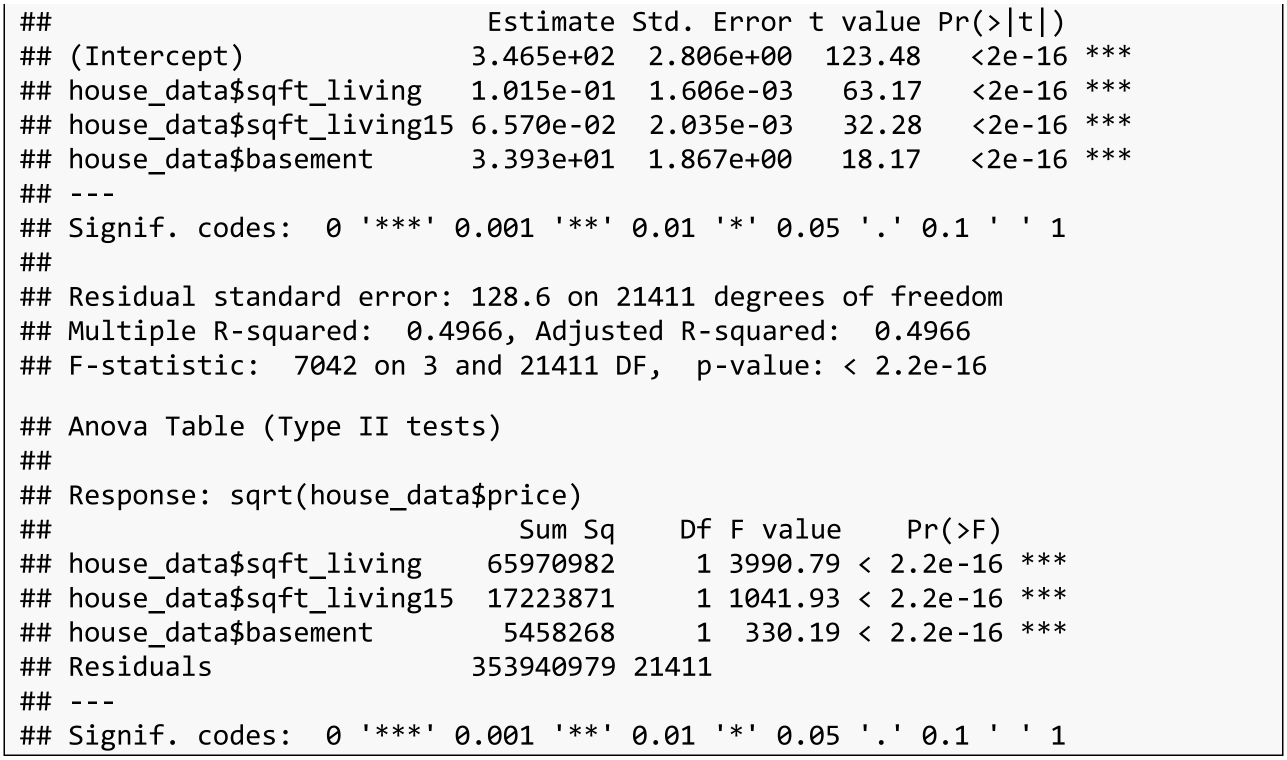
$(F(3,21411)=7042, p<0.001)$,

with an R^2 of 0.4966. The house price can be predicted using the following regression equation:

$$(346.515 + 0.101455(\text{sqft\_living}) + 0.065702(\text{sqft\_living15}) + 33.931363(\text{basement}))^2$$

where basement is coded as 0 = has no basement, 1 = has basement, and interior living space is measured in square foot.

```
##
## Call:
## lm(formula = sqrt(house_data$price) ~ house_data$sqft_living +
##     house_data$sqft_living15 + house_data$basement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -429.53  -94.32   -6.58   77.72  632.73
##
## Coefficients:
```



```
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.465e+02  2.806e+00  123.48   <2e-16 ***
## house_data$sqft_living    1.015e-01  1.606e-03   63.17   <2e-16 ***
## house_data$sqft_living15 6.570e-02  2.035e-03   32.28   <2e-16 ***
## house_data$basement       3.393e+01  1.867e+00   18.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128.6 on 21411 degrees of freedom
## Multiple R-squared:  0.4966, Adjusted R-squared:  0.4966
## F-statistic:  7042 on 3 and 21411 DF,  p-value: < 2.2e-16

## Anova Table (Type II tests)
##
## Response: sqrt(house_data$price)
##                              Sum Sq    Df F value    Pr(>F)
## house_data$sqft_living     65970982     1 3990.79 < 2.2e-16 ***
## house_data$sqft_living15   17223871     1 1041.93 < 2.2e-16 ***
## house_data$basement         5458268     1  330.19 < 2.2e-16 ***
## Residuals                 353940979 21411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.7: Model 2 Outputs

According to the analysis the following results can be deduced:

1. If the interior living space of the house increases, the price of the house increases
2. If the average interior living space of the nearest fifteen houses increases, the price of the house increases
3. If the house has a basement, the price of the house increases.

All variables, “sqft_living”, “sqft_living15” and “basement” were significant predictors.

As per the Baseline Model, the residual plot shows evidence of linear regression, with the Normal Q-Q plot showing normally distributed residuals. The same three cases are highlighted as potential issues in the Cooks Distance graph, albeit this time the distance has reduced.

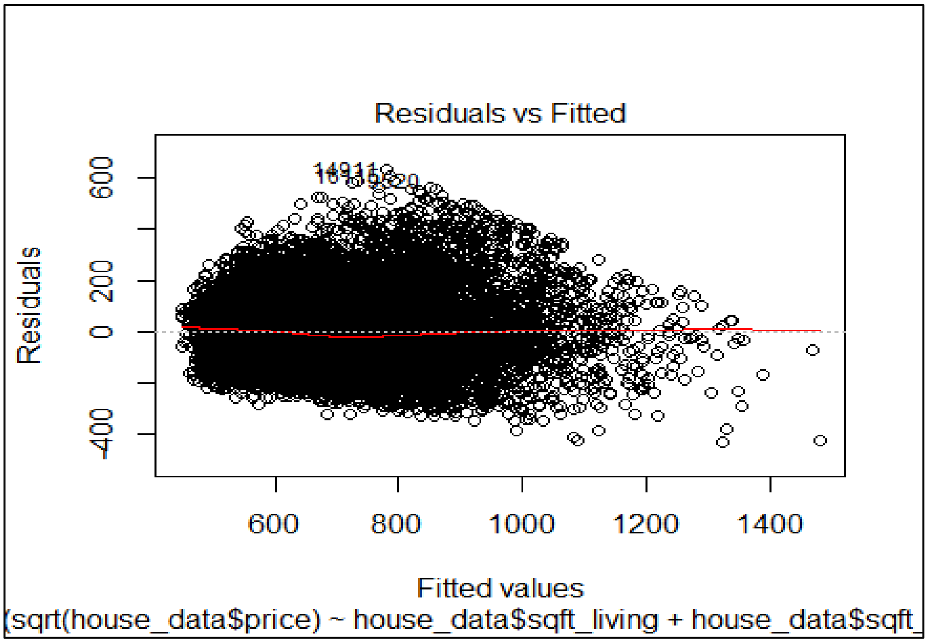


Figure 3.8: Residual V Fitted Plot for Model 2

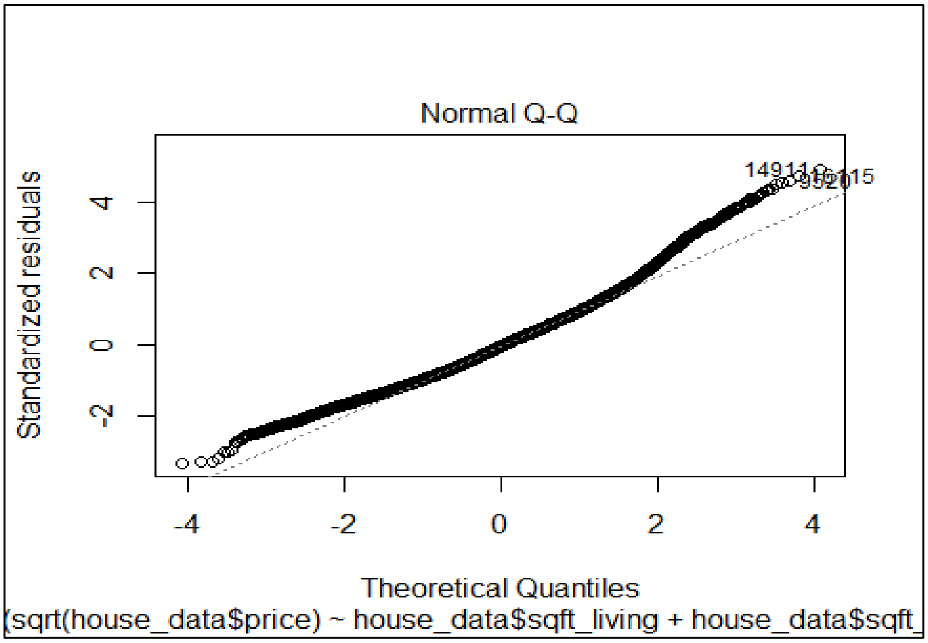


Figure 3.9: Normal Q-Q Plot for Model 2

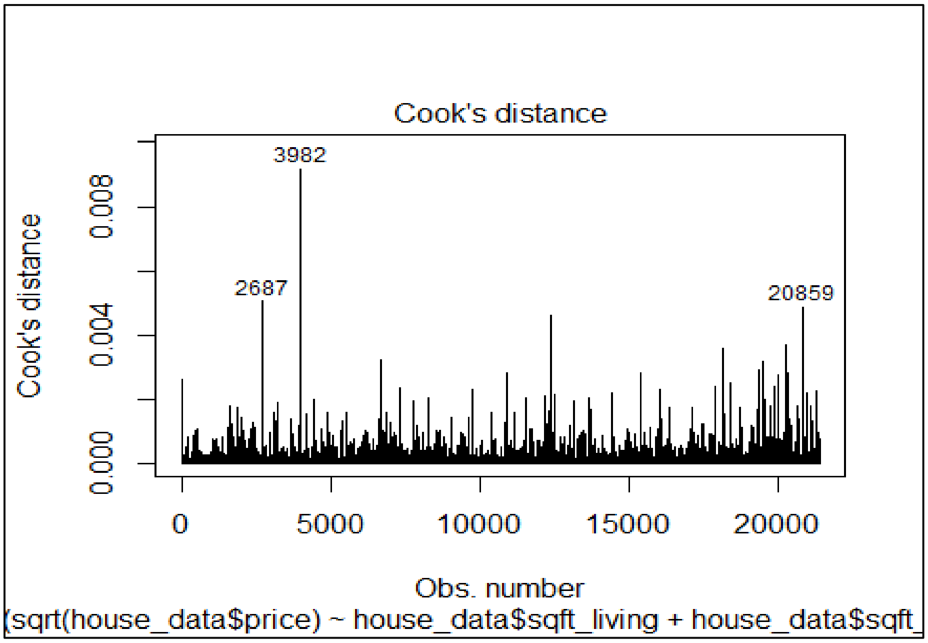


Figure 3.10: Cooks Distance Graph for Model 2

Again, we can illustrate the findings using the median values for “sqft_living” and “sqft_living15” grouped by whether they have a basement or not.

With basement - $(346.515 + 0.101455(2090) + 0.065702(1800) + 33.931363(1))^2$

$(346.515 + 212.041 + 118.26 + 33.93)^2 = (710.751)^2$ =$505,166.86

Without basement - $(346.515 + 0.101455(1740) + 0.065702(1870) + 33.931363(0))^2$

$(346.515 + 176.53 + 122.86)^2 = (645.91)^2$ = $417,199.00

Based on the above calculations, a house with an interior living space of 2090sqft and the interior living space of the nearest fifteen houses, on average 1800sqft and with a basement would be priced at approximately at $505,166.86. A house with an interior living space of 1740sqft and the interior living space of the nearest fifteen houses, on average 1870sqft and no basement would be priced at approximately $417,199. Like the baseline model, these approximates are relatively close to the median price of houses both with ($509,007) and without a basement ($410,000). This result is intuitive, as the median floor area of a house with a basement is bigger than that without a basement.

To get an understanding of the true differential effect for having a basement or not, the same regression equation can be calculated, using the baseline value for the interior living space for both houses with and without a basement.

With basement - $(346.515 + 0.101455(1900) + 0.065702(1830) + 33.931363(1))^2$

$$(346.515 + 192.7645 + 120.2347 + 33.93)^2 = (693.4455)^2 = \$480{,}866.70$$

Without basement - $(346.515 + 0.101455(1900) + 0.065702(1830) + 33.931363(0))^2$

$$(346.515 + 192.7645 + 120.2347)^2 = (659.5142)^2 = \$434{,}958.90$$

3.4 Model 3

The final model in this analysis explored if there was a differential effect depending on the building construction grade given to the house. The variables “sqft_living15” and “basement” were removed such that only the characteristics of each house in the dataset were examined. Based on the literature, it is expected that houses with a higher grade will have a higher price. The variable “grade_grp” was transformed into a dummy variable, 1-5 = 0, 6-8 = 1, 9-8 = 2.

The outputs, seen below in Figure 3.11, show that a significant regression equation was found

(F(3,21411)=7471, $p<0.001$) with an R^2 of 0.5114.

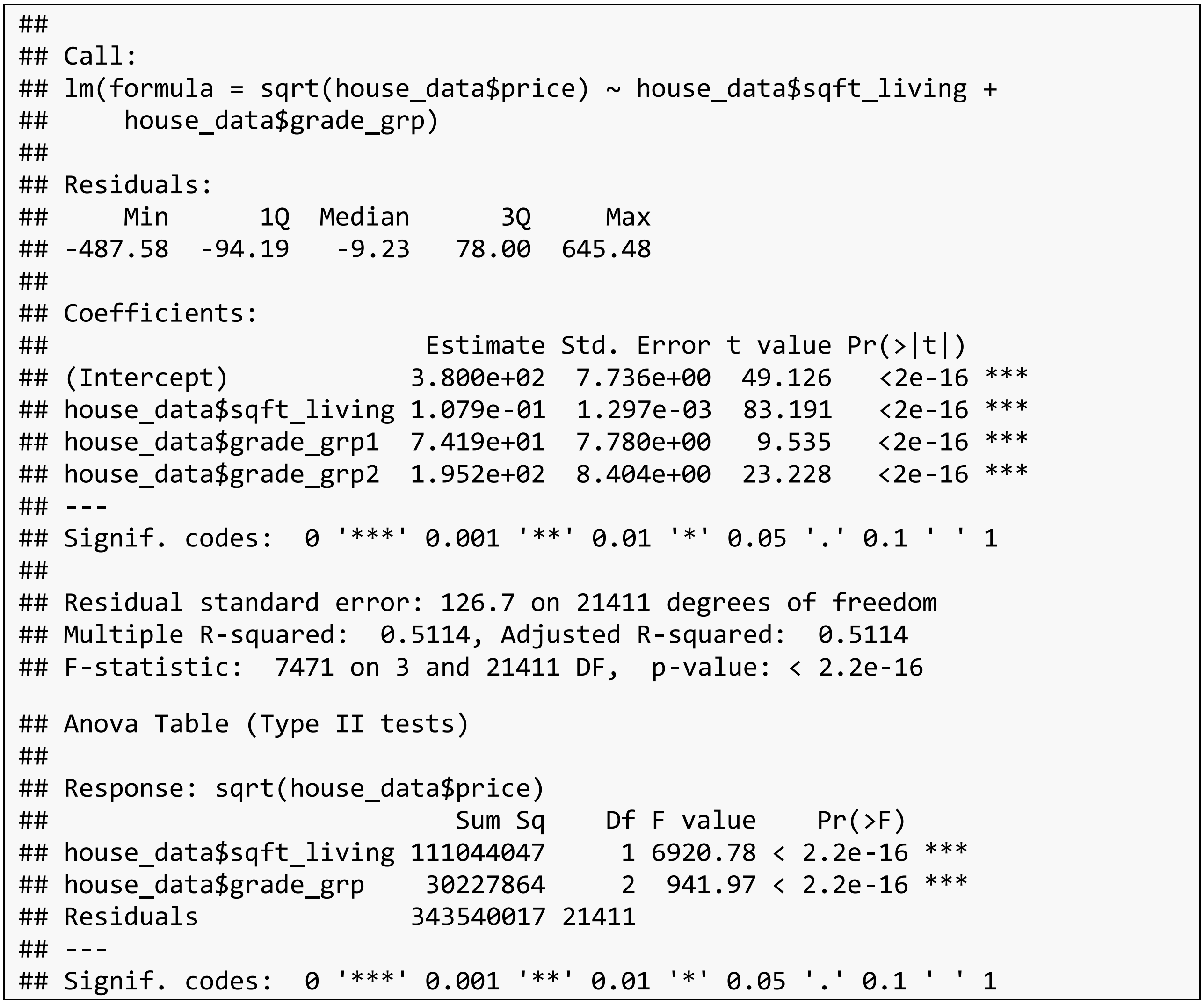
The regression equation is as follows:

$$(380.015 + 0.107859(\text{sqft\_living}) + 74.187(\text{6-8 group}) + 195.203(\text{9-13 group}))^2$$

where “grade_grp” is coded as 0 = 1-5, 1 = 6-8, 2= 9-13, and interior living space is measured in square foot. According to the analysis the following results can be derived:

1. If the interior living space of the house increases, the price of the house increases
2. If the house has a building grade between 6-8 and 9-13, the price of the house increases.

Both “sqft_living” and “grade_grp” were significant predictors.



```
##
## Call:
## lm(formula = sqrt(house_data$price) ~ house_data$sqft_living +
##     house_data$grade_grp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -487.58  -94.19   -9.23   78.00  645.48
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.800e+02  7.736e+00  49.126   <2e-16 ***
## house_data$sqft_living 1.079e-01  1.297e-03  83.191   <2e-16 ***
## house_data$grade_grp1  7.419e+01  7.780e+00   9.535   <2e-16 ***
## house_data$grade_grp2  1.952e+02  8.404e+00  23.228   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126.7 on 21411 degrees of freedom
## Multiple R-squared:  0.5114, Adjusted R-squared:  0.5114
## F-statistic:  7471 on 3 and 21411 DF,  p-value: < 2.2e-16

## Anova Table (Type II tests)
##
## Response: sqrt(house_data$price)
##                           Sum Sq    Df F value    Pr(>F)
## house_data$sqft_living 111044047     1 6920.78 < 2.2e-16 ***
## house_data$grade_grp    30227864     2  941.97 < 2.2e-16 ***
## Residuals              343540017 21411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.11: Outputs for Model 3

Of all three models, the residual plot for model 3, shows the best fit, with the spread of cases a lot closer to the horizontal at 0 (Figure 3.12). Again, there were only three cases of the 21,415 identified as potential issues in the Cooks Distance graph (Figure 3.14).

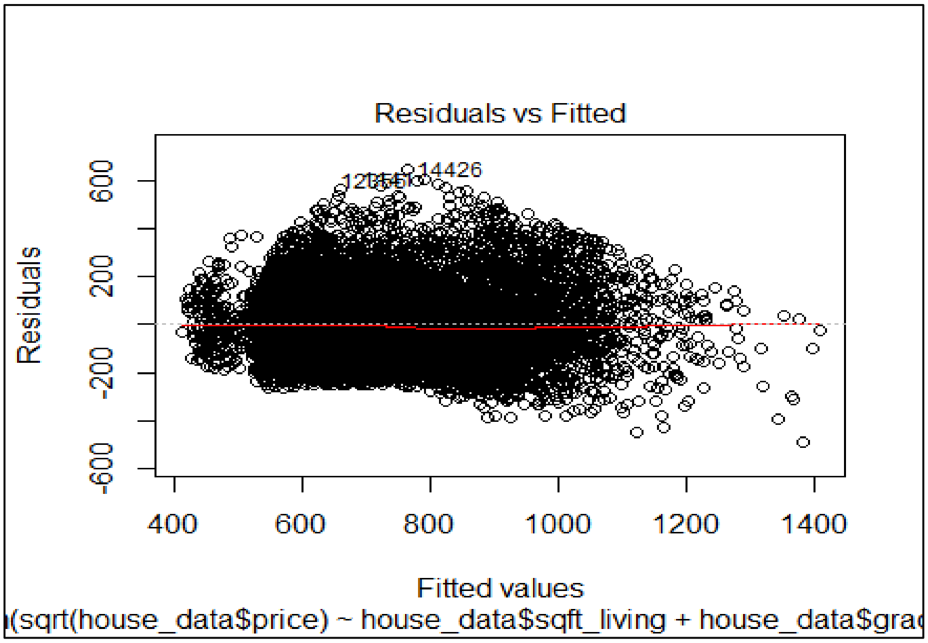


Figure 3.12: Residual V Fitted Plot for Model 3

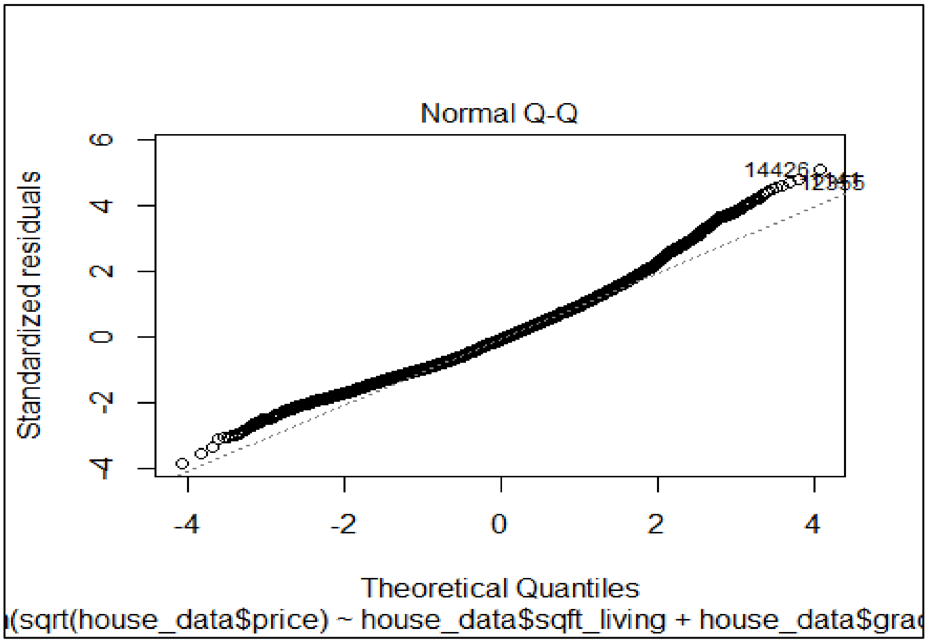


Figure 3.13: Normal Q-Q Plot for Model 3

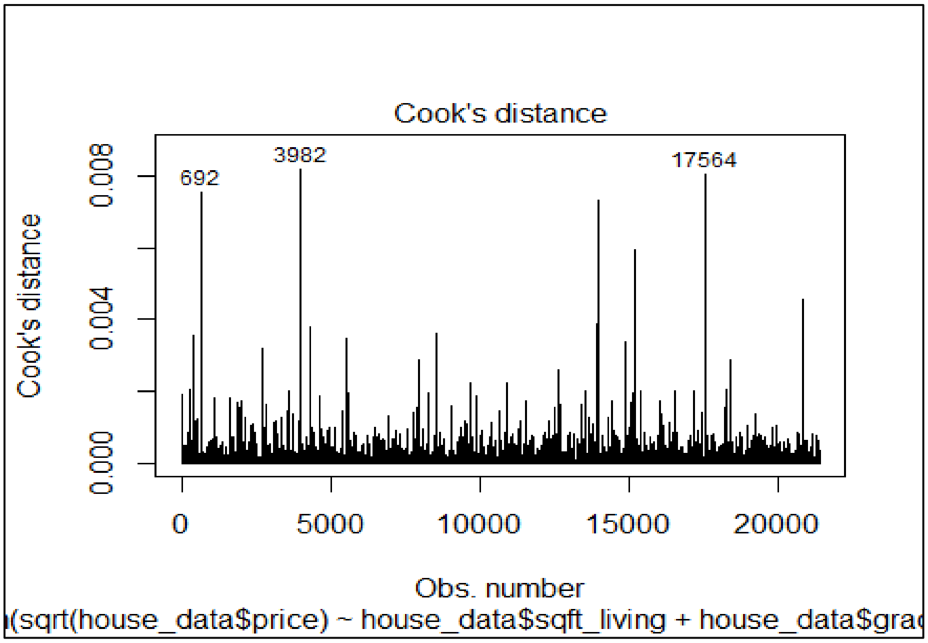


Figure 3.14 Cooks Distance Graph for Model 3

Using the regression equation, we can illustrate the findings for model 3 using the median value for the interior living space for each grade group.

$$1\text{-}5: (380.015 + 0.107859(850) + 74.187(0) + 195.203(0))^2$$

$$(380.015 + 91.68015)^2 = (471.69)^2 = \$222{,}496.30$$

$$6\text{-}8: (380.015 + 0.107859(1740) + 74.187(1) + 195.203(0))^2$$

$$(380.015 + 187.6747 + 74.187)^2 = (641.8767)^2 = \$412{,}005.60$$

$$9\text{-}13: (380.015 + 0.107859(3070) + 74.187(0) + 195.203(1))^2$$

$$(380.015 + 331.1271 + 195.203)^2 = (906.3451)^2 = \$821{,}461.50$$

Comparing these results back to the median values of price by the grade group, again provide a sense of the accuracy of the model with the median price of homes with a grade between 1-5 being $225,000, the median price of homes with a grade between 6-8 being $405,000 and the median price of homes with a grade between 9-13 being $799,900 as presented earlier.

4.Discussion

If the significance level is accepted as 0.05, all variables used in these models have a significant impact on the price of a house. Statistically, all three of these models are quite similar, with the adjusted R^2 value increasing with each addition and/or exclusion to the model.

The baseline model had an adjusted R^2 value of 0.4888, meaning that approximately 48.88% of the variance of the house price was explained by the square foot of the interior living space. This is consistent with the literature which indicates that the while square footage is an important factor in determining house price, there are multiple variables which influence house price not least the proximity of schools, proximity to the Central Business District and location. However, these variables were outside the scope of this analysis.

Adding in the differential effect for basement increased this by a small margin. The interior living space and the basement indicator explained for 49.66% of the variance of the house price. This model includes floor space, and the two models presented show that the main predictor for house price is floor area. The presence of a basement (y/n) in isolation does not have a significant impact on house price and therefore should not be used independently to predict house prices for this dataset.

The final model, which removed the average square footage of the interior living space of the nearest fifteen house and the basement indicator and included the condition grade group, was the best fit model. This model produced an adjusted R^2 value of 0.5114, meaning that the interior living space and the grade group explained 51% of the variance of the house price. There may be some bias towards this model however as it would be expected that house prices with a better building grade would be more expensive. This model provided a close approximation to the expected values. Higher accuracy may be achieved by not grouping building grades, but rather considering them individually.

The aim of this analysis, as highlighted in the Introduction was to explore the factors that influence house prices focusing on the size of the interior living space of the house and whether it could be considered to predict house prices. It has been demonstrated through this analysis that house price is dependent on square footage of living space, and the floor area can be used to predict house prices, hence rejecting the null hypothesis as stated in the Introduction and accepting the alternative hypothesis that in the presence of other variable, there is a significant prediction of house prices by the size of the interior living space. A stronger model could be produced using the additional pertinent variables available in the dataset, as evidenced by the addition of grade group to the analysis.

The analysis presented herein is limited to predicting house prices for the selected dataset. Since additional factors, for example location, impact house price further analysis could be conducted to include location to improve both the accuracy and scope of the model.

5. References

Abelson, P., Joyeux, R., Mahuteau S., *Modelling House Prices across Sydney,* The Austrailian Economic Review, 46(3), 269-85

Des Rosiers, F., Lagana, A., Theriault, M., *Size and proximity effects of primary schools on surrounding house values*, Journal of Property Research, 2001, 18(2) 149 -168

Kamal, E.M, Hassan, H. and Osmadi, A., *Factors Influencing the Housing Price: Developers' Perspective*, 2016

MacDonald, S., *Drivers of house price inflation in Penang, Malaysia: Planning a more sustainable future*, 2011, Penang Institute: Penang.

Ooi, J.T.L., Thao, T.T., LeNai-Jia, L., *The impact of construction quality in house prices,*Journal of Housing Economincs, 2014, 26, 126-138

Quigley, J.M. and S. Raphael, *Is Housing Unaffordable? Why Isn't It More Affordable? Journal of Economic Perspectives*, 2004

Turnbull, G.K, Dombrow, J., Sirmans C.F, *Big House, Little House: Relative Size and Values,* 2006 34(3), 439-456

Sirmams, S., MacDonald, L., Macpherson, D., Norman Zietz, E., *The Value of Housing Charachteristics: A Meta Analysis,* 2006

Yusof, A.M, Ismail, S, *Multiple Regressions in Analysing House Price Variations*, 2012

Appendix A

Descriptive Statistics

Price

```
#price
summary(house_data$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   75000  320000  450000  519421  635000 2000000

median(house_data$price)

## [1] 450000

skew(house_data$price)

##   skew (g1)          se           z           p
##   1.7761615   0.0167385 106.1123330   0.0000000

kurtosis(house_data$price)

## Excess Kur (g2)              se               z               p
##        4.310795        0.033477      128.768851        0.000000
```

sqft_living/sqft_living15

```
#sqft_living
summary(house_data$sqft_living)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     290    1420    1900    2052    2520    7730

median(house_data$sqft_living)

## [1] 1900

skew(house_data$sqft_living)

##  skew (g1)         se          z          p
##  1.0826899  0.0167385 64.6826037  0.0000000

kurtosis(house_data$sqft_living)

## Excess Kur (g2)              se               z               p
##        1.929331        0.033477       57.631528        0.000000

#sqft_living15
summary(house_data$sqft_living15)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     399    1480    1830    1973    2340    5790

median(house_data$sqft_living15)
```

```
## [1] 1830
```

```r
skew(house_data$sqft_living15)
```

```
##  skew (g1)         se          z          p
##  1.0745240  0.0167385 64.1947553  0.0000000
```

```r
kurtosis(house_data$sqft_living15)
```

```
## Excess Kur (g2)              se               z               p
##        1.498054        0.033477       44.748752        0.000000
```

Appendix B

Correlations

Price vs Square Foot of Interior Living Space

```
# Spearman correlation
cor.test(house_data$price, house_data$sqft_living, method = "spearman")

## Warning in cor.test.default(house_data$price, house_data$sqft_living,
## method = "spearman"): Cannot compute exact p-value with ties

## 
##  Spearman's rank correlation rho
## 
## data:  house_data$price and house_data$sqft_living
## S = 5.9832e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho 
## 0.6344653
```

Price vs Avg Square Foot of Interior Living Space (nearest 15 houses)

```
# Spearman correlation
cor.test(house_data$price, house_data$sqft_living15, method = "spearman")

## Warning in cor.test.default(house_data$price, house_data$sqft_living15, :
## Cannot compute exact p-value with ties

## 
##  Spearman's rank correlation rho
## 
## data:  house_data$price and house_data$sqft_living15
## S = 7.1712e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho 
## 0.5618865
```

Price Vs Basement

```
#Look at differences in prices for houses with/without a basement
#Conduct Levene's test for homogeneity of variance in library car
library(car)

#undo dummy variables
house_data$basement <- as.factor(recode(house_data$basement,"1"="Y","0"="N"))
leveneTest(price ~ basement, data=house_data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)    
## group     1  106.61 < 2.2e-16 ***
##       21413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Conduct the t-test
#You can use the var.equal = TRUE option to specify equal variances and a poo
led variance estimate
t.test(price~basement,var.equal=FALSE,data=house_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  price by basement
## t = -25.576, df = 15696, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -113206.98  -97089.77
## sample estimates:
## mean in group N mean in group Y
##        478466.1        583614.5
```

Price Vs Grade_grp

```
library(userfriendlyscience)
house_data$grade_grp <- recode(house_data$grade_grp,"0"="1-5","1"="6-8","2"="
9-13")
one.way <- oneway(house_data$grade_grp, y = house_data$price, posthoc = 'Tuke
y')
one.way
```

```
## ### Oneway Anova for y=price and x=grade_grp (groups: 1-5, 6-8, 9-13)
##
## Omega squared: 95% CI = [.35; .37], point estimate = .36
## Eta Squared: 95% CI = [.35; .37], point estimate = .36
##
##                                               SS    Df                MS
## Between groups (error + effect)  638345939264360     2 319172969632180
## Within groups (error only)      1135531244556984 21412   53032469856.01
##                                         F      p
## Between groups (error + effect) 6018.44 <.001
## Within groups (error only)
##
##
## ### Post hoc test: Tukey
##
##           diff      lwr       upr       p adj
## 6-8-1-5   195638.49 162828.58 228448.4  <.001
## 9-13-1-5  629456.83 595823.36 663090.31 <.001
## 9-13-6-8  433818.34 424393.17 443243.52 <.001
```

Appendix C

Model 1

```
vif(model1)

##   house_data$sqft_living house_data$sqft_living15 
##                 2.323414                 2.323414 

sqrt(vif(model1))

##   house_data$sqft_living house_data$sqft_living15 
##                 1.524275                 1.524275 
```

Model 2

```
vif(model2)

##   house_data$sqft_living house_data$sqft_living15      house_data$basement 
##                 2.493371                 2.401730                 1.074129 

sqrt(vif(model1))

##   house_data$sqft_living house_data$sqft_living15 
##                 1.524275                 1.524275 
```

Model 3

```
vif(model3)

##                           GVIF Df GVIF^(1/(2*Df))
## house_data$sqft_living 1.67421  1        1.293913
## house_data$grade_grp   1.67421  2        1.137503

sqrt(vif(model1))

##   house_data$sqft_living house_data$sqft_living15 
##                 1.524275                 1.524275 
```