Group 189: Explaining House Price Variability

First Name	Last Name	Monday or Tuesday class	Share project with ITMD 525? (Y or N)
Yeshwanthi	Jayaraman Durairaj	Tuesday	N
Priyanka	Agrawal	Tuesday	N

Table of Contents

1. Introduction

The aim of the project is to predict housing price of King County located in Washington state. This involves predicting the price of the house (dependent variable) based on various factors which have significant effect on the house price. This will help to identify the factors that contribute in explaining variation in housing price. This will in turn help investors with decision making process of buying home in King County.

There is popular belief that growth of Amazon has contributed in inflating price of the houses located in King county. We are also planning to test this hypothesis by introducing a variable which will calculate the average distance of the house from Amazon HQ.

There is also one more popular belief that crime rate of a region influences the prices of the houses located in that specific region. We are also planning to test this hypothesis by introducing a variable which will calculate the crime rate of the zip codes located in King county.

2. Data

Kaggle Data Set

The Data set for this project belongs to Real estate domain and the data source is available in Kaggle website. The data set comprises of 21613 observations and 17 variables.

Crime Rate & distance from Amazon HQ Data Set

Apart from the housing features from Kaggle dataset, we have introduced two new variables (crime rate and average distance between the house and Amazon HQ) that may explain the variation in prices of the houses in King county.

We identified the independent variables that we are going to use in this project from the data set and derived few more variables from independent variables. For example, renovation work happened in 2015 and after 2015, Square feet of the living area and lot are changed. We derived the new square feet based on year of renovation. Similarly, we calculated crime rate using crime count and population for better standardization.

Independent Variables:

Date
Id
Number of Bedrooms
Number of Bathrooms
Square feet of the living area in 2014
Square feet of the living area in 2015
Square feet of the lot in 2014
Square feet of the lot in 2015
Total floors in the house
Is the house having water front view or not
View
Condition of the house
Grade of the house based on grading system in King county
Crime Count in Zip Code & Year

Population
Distance from Amazon HQ.

Derived Variables

Per Square-Feet - (Price/Living SQFT)
New_SQFT_LIVING - (2014 & 2015 changes due to renovation)
New_SQFT_LOT - (2014 & 2015 changes due to renovation)
Crime rate - (Crime count/Population)

Data set Source:

https://www.kaggle.com/harlfoxem/housesalesprediction.

https://moto.data.socrata.com/dataset/King-County-Sheriff-s-Office/4h35-4mtu/data

3. Problems to be Solved

We are planning to answer below research problems,

1. What are the most important features of a home that will explain variation in housing price in King County.
2. Effect of Crime on Housing price in King county.
3. Effect of distance from Amazon HQ on housing price in King County.

4. Data Processing

Data Cleaning

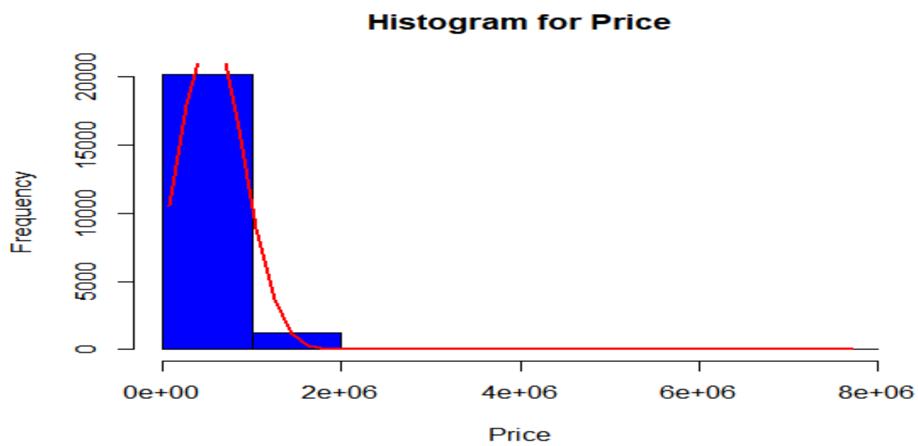
We made cleaning for few variables and resolved few data entry issues in our data set. For example, we encountered a data records having33 bedrooms, but the living area square feet very less. This is could not happen in real world and this is a serious data entry issue. We cleaned all this types of erroneous data from our data set. We also stripped of few commas and $ symbols from numeric data for computation.

Rooms_Issue

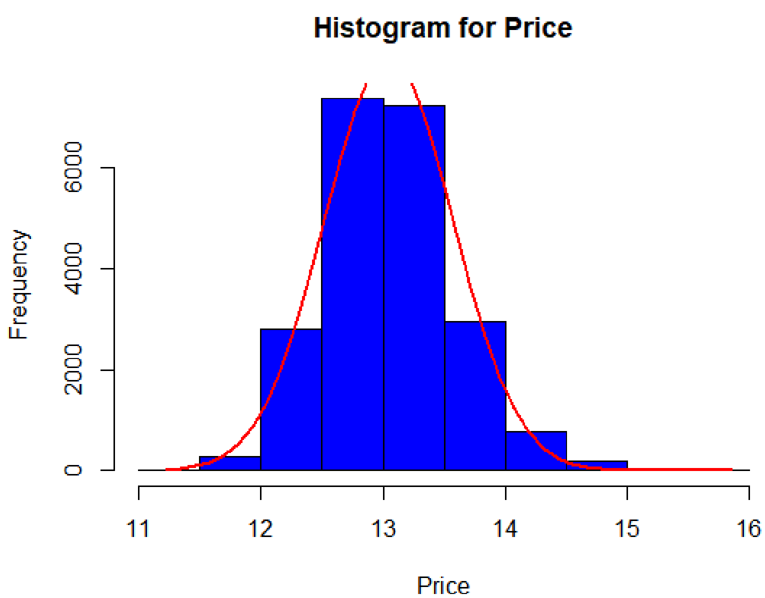
0	1	2	3	4	5	6	7	8	9	10	11	33
13	199	2760	9824	6882	1601	272	38	13	6	3	1	1

We analyzed our x and y variables and discovered that not all data are normalized and in standard form. We applied transformations on our dependent variable (price) for better normalization.

Explorative Analysis of Home Price – Dependent variable



Home Price- After Log Transformation



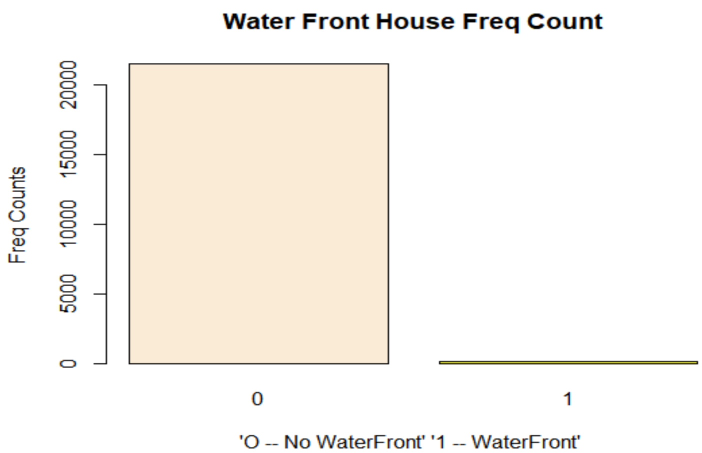
Frequency Analysis of X Variables

We made frequency analysis of few categorical and ordinal variables and figured out that the frequency distribution of some variables are not distributed normally and did not include those variables in our model.

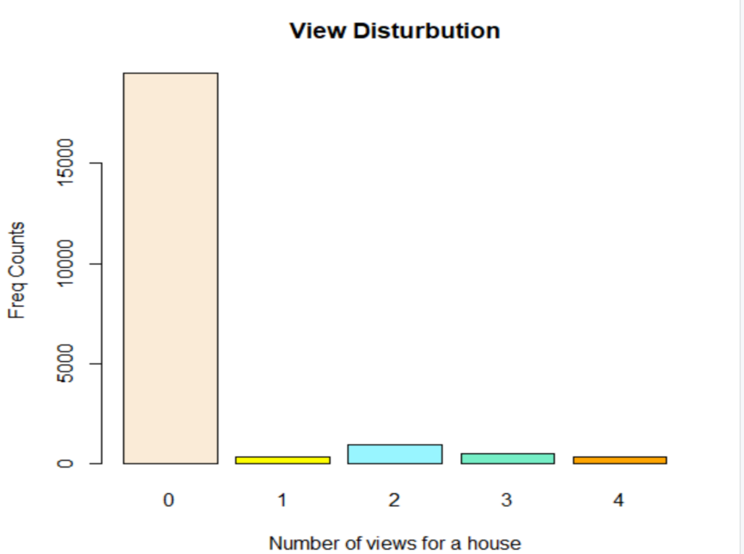
Water front variables denotes whether the house has a water front view or not and more than 99 % of the houses did not have a water front view.

View variable denotes number of time the house has been viewed and 99 % of the houses has not been view and the value is 0.

Frequency of Water Front Houses



Frequency of the Houses which has been viewed or not

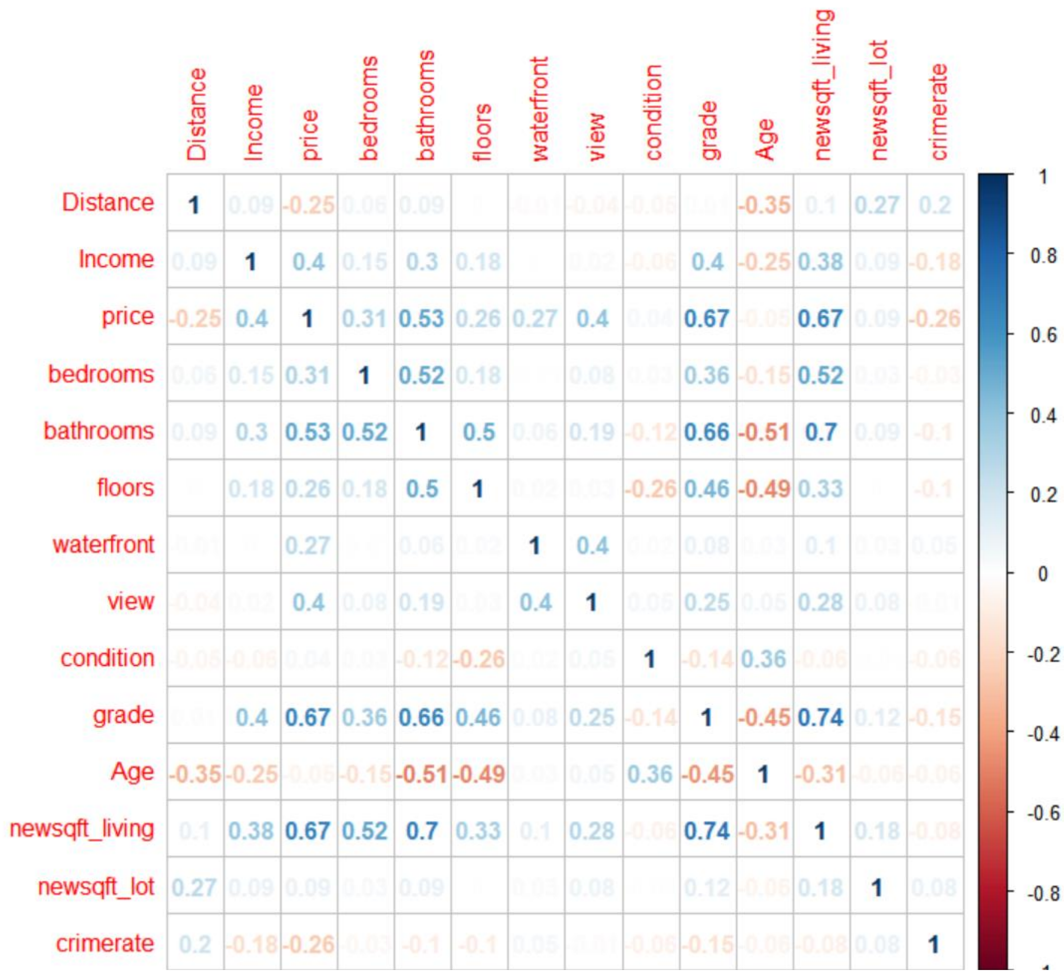


4.Methods and Process

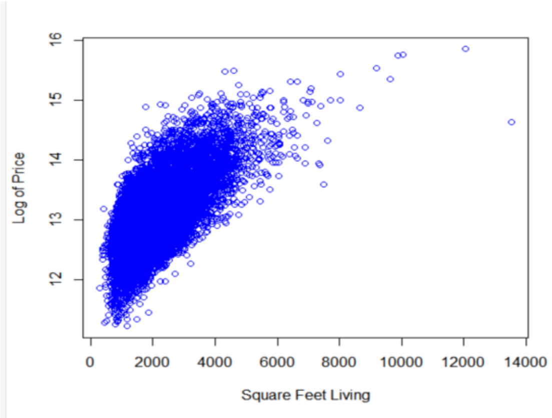
A multiple linear regression was performed to assess if the independent variables (housing features) explain the variations in prices of house. This in turn identified the variables having significant effect on the prices of house. For significant independent variable, every unit increase or decrease in the independent variable the price of the house can be determined using coefficients of beta.

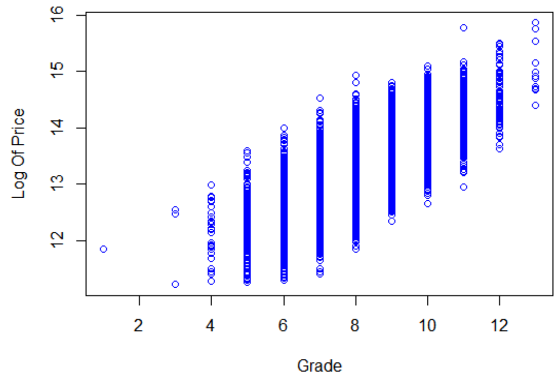
Pearson Correlation Matrix:

We used Pearson Correlation matrix to determine the correlation of each pair of variables in our data set. From the correlation matrix we observed that condition, newsqft_lot has very meagre correlation with price and it is not linear even after applying transformations. We dropped these variables and did not include in our model.



Linearity Assumptions: We checked for linear assumptions of X variable with y variable using scatter plot and correlation values.





Model selection process:

(i) Model built using Backward elimination process

We built the multi linear regression model by manual backward elimination process. This involves in building the model with all variables and by manually dropping variable which are not significant on price (p value is less than 0.05)

Model1:

```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + grade + floors +
    newsqft_living + crimerate + Distance, data = train1.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90442 -0.18153  0.00387  0.17504  1.46966

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.156e+01  1.915e-02 603.466  < 2e-16 ***
bedrooms       -1.913e-03  2.890e-03  -0.662    0.508
bathrooms       4.634e-02  4.693e-03   9.876  < 2e-16 ***
grade           1.705e-01  3.072e-03  55.523  < 2e-16 ***
floors         -3.917e-02  4.899e-03  -7.996 1.37e-15 ***
newsqft_living  2.277e-04  4.465e-06  50.999  < 2e-16 ***
crimerate      -4.924e+00  1.195e-01 -41.189  < 2e-16 ***
Distance       -1.417e-02  1.952e-04 -72.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2909 on 17282 degrees of freedom
Multiple R-squared:  0.6932,    Adjusted R-squared:  0.6931
F-statistic:  5579 on 7 and 17282 DF,  p-value: < 2.2e-16
```

Model 2 – Excluding bedroom from model1

We built model 2 by removing bedroom from model 1 as bedroom does not have any significant effect on price.

```
Call:
lm(formula = log(price) ~ bathrooms + grade + floors + newsqft_living +
    crimerate + Distance, data = train1.data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8980 -0.1813  0.0040  0.1755  1.4709

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.155e+01  1.753e-02 659.129  < 2e-16 ***
bathrooms       4.547e-02  4.505e-03  10.095  < 2e-16 ***
grade           1.708e-01  3.050e-03  55.992  < 2e-16 ***
floors         -3.893e-02  4.885e-03  -7.969  1.7e-15 ***
newsqft_living  2.269e-04  4.281e-06  52.999  < 2e-16 ***
crimerate      -4.924e+00  1.195e-01 -41.192  < 2e-16 ***
Distance       -1.417e-02  1.952e-04 -72.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2909 on 17283 degrees of freedom
Multiple R-squared:  0.6932,    Adjusted R-squared:  0.6931
F-statistic:  6509 on 6 and 17283 DF,  p-value: < 2.2e-16
```

After removing bedroom, all variables have significant effect on price. That is p value of all variables are less than 0.05.

(ii) Model built using Step wise Regression

We used step wise regression approach to build new models. Step wise regression suggest variables to be included in our model .

Forward:

Forward stepwise regression suggested to include all features in model 1.

```
> #Forward stepwise regression
> step(m1,direction="forward",trace=T)
Start:  AIC=-42596.26
log(price) ~ bathrooms + grade + bedrooms + floors + newsqft_living +
    crimerate + Distance


Call:
lm(formula = log(price) ~ bathrooms + grade + bedrooms + floors +
    newsqft_living + crimerate + Distance, data = train1.data)

Coefficients:
   (Intercept)       bathrooms           grade        bedrooms          floors
    11.5554121       0.0462107       0.1661926      -0.0006836      -0.0308359
newsqft_living       crimerate        Distance
     0.0002340      -4.6785912      -0.0141153
```

Backward:

Backward step wise regression suggested to remove bedrooms which is same like model 2.

```
> step(m1,direction="backward",trace=T)
Start:  AIC=-42596.26
log(price) ~ bathrooms + grade + bedrooms + floors + newsqft_living +
    crimerate + Distance

                 Df Sum of Sq    RSS    AIC
- bedrooms        1      0.00 1470.2 -42598
<none>                        1470.2 -42596
- floors          1      3.35 1473.5 -42559
- bathrooms       1      8.25 1478.4 -42502
- crimerate       1    134.18 1604.3 -41088
- newsqft_living  1    233.63 1703.8 -40048
- grade           1    245.95 1716.1 -39924
- Distance        1    449.48 1919.6 -37986

Step:  AIC=-42598.21
log(price) ~ bathrooms + grade + floors + newsqft_living + crimerate +
    Distance

                 Df Sum of Sq    RSS    AIC
<none>                        1470.2 -42598
- floors          1      3.35 1473.5 -42561
- bathrooms       1      8.93 1479.1 -42495
- crimerate       1    134.22 1604.4 -41090
- grade           1    249.59 1719.7 -39889
- newsqft_living  1    253.96 1724.1 -39845
- Distance        1    449.52 1919.7 -37988

Call:
lm(formula = log(price) ~ bathrooms + grade + floors + newsqft_living +
    crimerate + Distance, data = train1.data)
```

Both:

Both step wise regression suggested to remove bedrooms which is like model 2.

```
Step:  AIC=-42598.21
log(price) ~ bathrooms + grade + floors + newsqft_living + crimerate +
    Distance

                 Df Sum of Sq    RSS    AIC
<none>                        1470.2 -42598
+ bedrooms        1      0.00 1470.2 -42596
- floors          1      3.35 1473.5 -42561
- bathrooms       1      8.93 1479.1 -42495
- crimerate       1    134.22 1604.4 -41090
- grade           1    249.59 1719.7 -39889
- newsqft_living  1    253.96 1724.1 -39845
- Distance        1    449.52 1919.7 -37988

Call:
lm(formula = log(price) ~ bathrooms + grade + floors + newsqft_living +
    crimerate + Distance, data = train1.data)

Coefficients:
   (Intercept)       bathrooms           grade          floors  newsqft_living
    11.5536187       0.0458941       0.1662740      -0.0307505       0.0002337
     crimerate        Distance
    -4.6789240      -0.0141148

>
```

(iii) Models built using Best Subset Regression

We built models using best subset regression approach. It finds the best subset of x variabes using cp, R2 or Adj R2

Subset Selection by Cp

Subset selection regression using cp as metric suggested features used in model 1.

```
$size
 [1] 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6
[46] 6 6 7 7 7 7 7 7 7 8

$Cp
 [1]  7760.0908  7769.8755 15445.4308 23568.6379 24675.0668 24811.9678 25059.7835  3371.7715
 [9]  4741.1099  4813.4615  5818.6001  6615.4250  7241.1756  7476.3602  7503.5800  7620.7734
[17]  7669.0860   909.4935  2413.9683  3006.1672  3281.7667  3340.9401  3395.8142  3987.9790
[25]  4276.0060  4447.8743  4614.4222   296.1984   778.9309   863.5045   896.8064  2120.4983
[33]  2326.5042  2403.3453  2793.9830  3003.6146  3243.2599   148.0933   249.2962   288.1154
[41]   715.3316   736.2879   830.2698  1925.8581  2109.6429  2315.7100  2782.2558    91.5139
[49]   104.7717   223.5023   630.3824  1901.9783  2934.2247  3160.6373     8.0000

>
```

Subset Selection by Adj RSquared

Subset selection regression using adjusted R squared as metric suggests newsqft_living, Distance and grade features

```
+        )],method="adjr2")
$which
  Distance bedrooms bathrooms floors grade newsqft_living crimerate
1    FALSE    FALSE     FALSE  FALSE FALSE           TRUE     FALSE
1    FALSE    FALSE     FALSE  FALSE  TRUE          FALSE     FALSE
1    FALSE    FALSE      TRUE  FALSE FALSE          FALSE     FALSE
1    FALSE     TRUE     FALSE  FALSE FALSE          FALSE     FALSE
1    FALSE    FALSE     FALSE   TRUE FALSE          FALSE     FALSE
1    FALSE    FALSE     FALSE  FALSE FALSE          FALSE      TRUE
1     TRUE    FALSE     FALSE  FALSE FALSE          FALSE     FALSE
2     TRUE    FALSE     FALSE  FALSE FALSE           TRUE     FALSE
2    FALSE    FALSE     FALSE  FALSE  TRUE           TRUE     FALSE
2     TRUE    FALSE     FALSE  FALSE  TRUE          FALSE     FALSE
2    FALSE    FALSE     FALSE  FALSE FALSE           TRUE      TRUE
2    FALSE    FALSE     FALSE  FALSE  TRUE          FALSE      TRUE
2    FALSE    FALSE      TRUE  FALSE FALSE           TRUE     FALSE
2    FALSE     TRUE     FALSE  FALSE FALSE           TRUE     FALSE
2    FALSE    FALSE     FALSE   TRUE FALSE           TRUE     FALSE
2    FALSE    FALSE      TRUE  FALSE  TRUE          FALSE     FALSE
2    FALSE     TRUE     FALSE  FALSE  TRUE          FALSE     FALSE
3     TRUE    FALSE     FALSE  FALSE  TRUE           TRUE     FALSE
3     TRUE    FALSE     FALSE  FALSE FALSE           TRUE      TRUE
3     TRUE    FALSE      TRUE  FALSE FALSE           TRUE     FALSE
3     TRUE     TRUE     FALSE  FALSE FALSE           TRUE     FALSE
```

```
$label
[1] "(Intercept)"    "Distance"       "bedrooms"       "bathrooms"
[5] "floors"         "grade"          "newsqft_living" "crimerate"

$size
 [1] 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 6
[39] 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 8

$adjr2
 [1] 0.46015157 0.44496535 0.27753200 0.10138200 0.06905934 0.06441928
 [7] 0.06132630 0.55724581 0.52021049 0.50820007 0.50255381 0.46969272
[13] 0.46566586 0.46280380 0.46174241 0.45713131 0.45107047 0.60427157
[19] 0.57773604 0.56375101 0.55951895 0.55799056 0.55017674 0.52679611
[25] 0.52223128 0.52179086 0.52018573 0.61797473 0.60654541 0.60567742
[31] 0.60442863 0.58298700 0.57982413 0.57801703 0.56883597 0.56375202
[37] 0.56028789 0.62063052 0.61931059 0.61807500 0.60785879 0.60748558
[43] 0.60630440 0.58749779 0.58312037 0.58012150 0.56897847 0.62186811
[49] 0.62149974 0.61980636 0.60976812 0.58786837 0.55479824 0.54936889
[55] 0.62364338
```

Model 3: Features suggested by Best subset Regression -AdjR2

We built a new model using the subset of x variables suggested by Adj R2 metric

```
Call:
lm(formula = log(price) ~ grade + newsqft_living + Distance,
    data = train1.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.45109 -0.20398 -0.00155  0.19264  1.56364 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.140e+01  1.799e-02  633.55   <2e-16 ***
grade           1.837e-01  2.948e-03   62.32   <2e-16 ***
newsqft_living  2.478e-04  4.057e-06   61.07   <2e-16 ***
Distance       -1.555e-02  1.989e-04  -78.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3056 on 17285 degrees of freedom
Multiple R-squared:  0.6664,    Adjusted R-squared:  0.6663 
F-statistic: 1.151e+04 on 3 and 17285 DF,  p-value: < 2.2e-16
```

Multi Collinearity:

There were no serious multicollinearity issues in our model. That is none of the independent variable has strong correlation with other independent variables used in the model

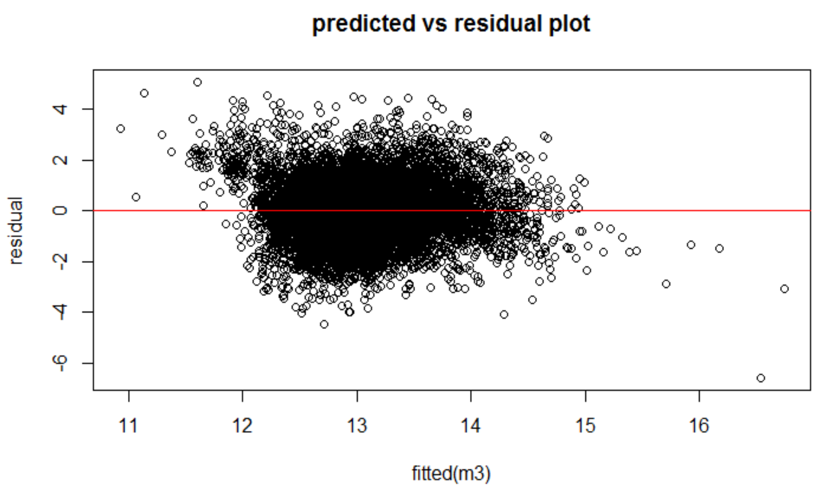
```
> #Testing multicollinearity for model 1
> vif(m1)
     bathrooms          grade       bedrooms         floors newsqft_living 
      2.669216       2.696074       1.540936       1.434268       2.996678 
     crimerate       Distance 
      1.070291       1.064465 
> #Testing multicollinearity for model 2
> vif(m2)
     bathrooms          grade         floors newsqft_living      crimerate 
      2.430971       2.659373       1.425761       2.749930       1.070124 
      Distance 
      1.064310 
> #Testing multicollinearity for model 3
> vif(m3)
         grade newsqft_living       Distance 
      2.233930       2.253016       1.017495 
> 
```

Residual Analysis:

Residual analysis was performed for our model and all residual assumptions are satisfied. From the residual plots we can observe that residuals are plotted with constant variance.

Predicted Versed Residual:

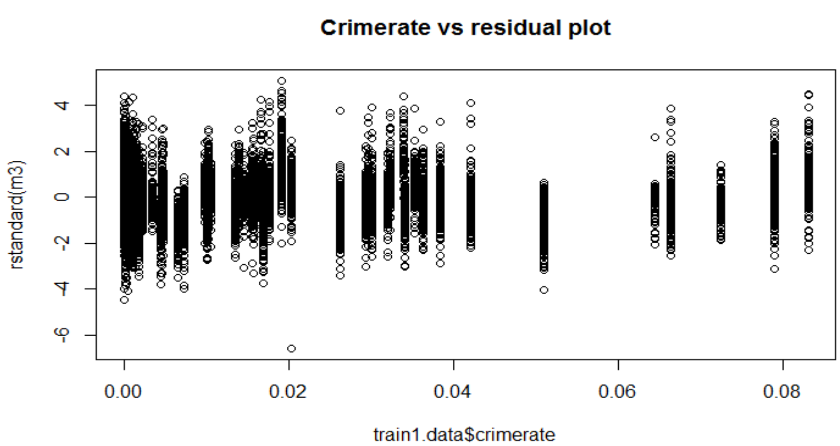
The plot between predicted and residual are with constant variance.



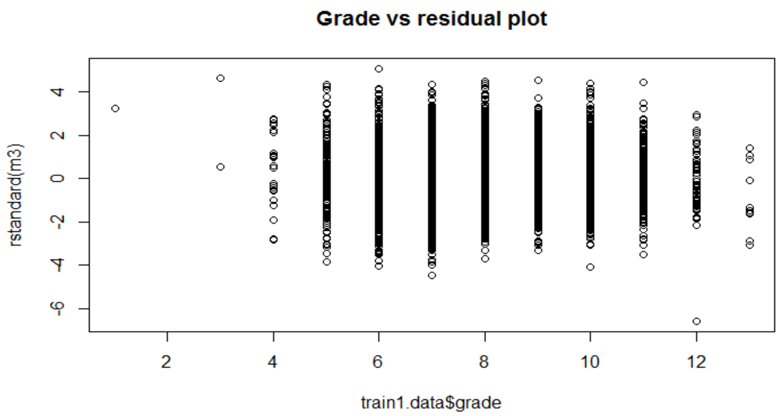
<u>Residual verses Independent variables</u>

The plot between the residual and independent variables are more or less with constant variane except for few outliers.

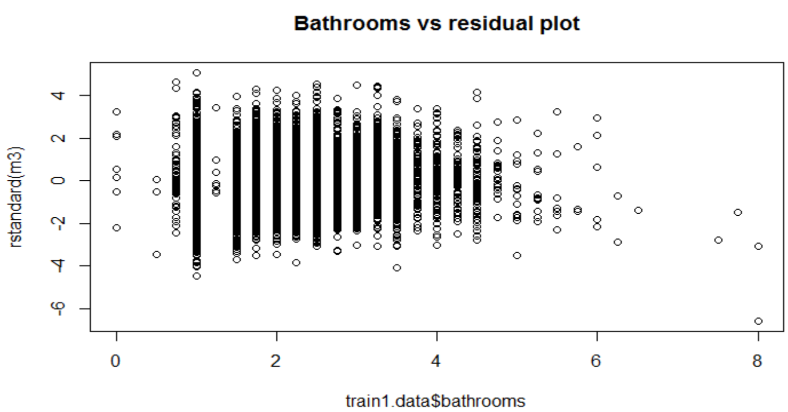
<u>Residual verses Crime rate</u>



Grade verses Residual

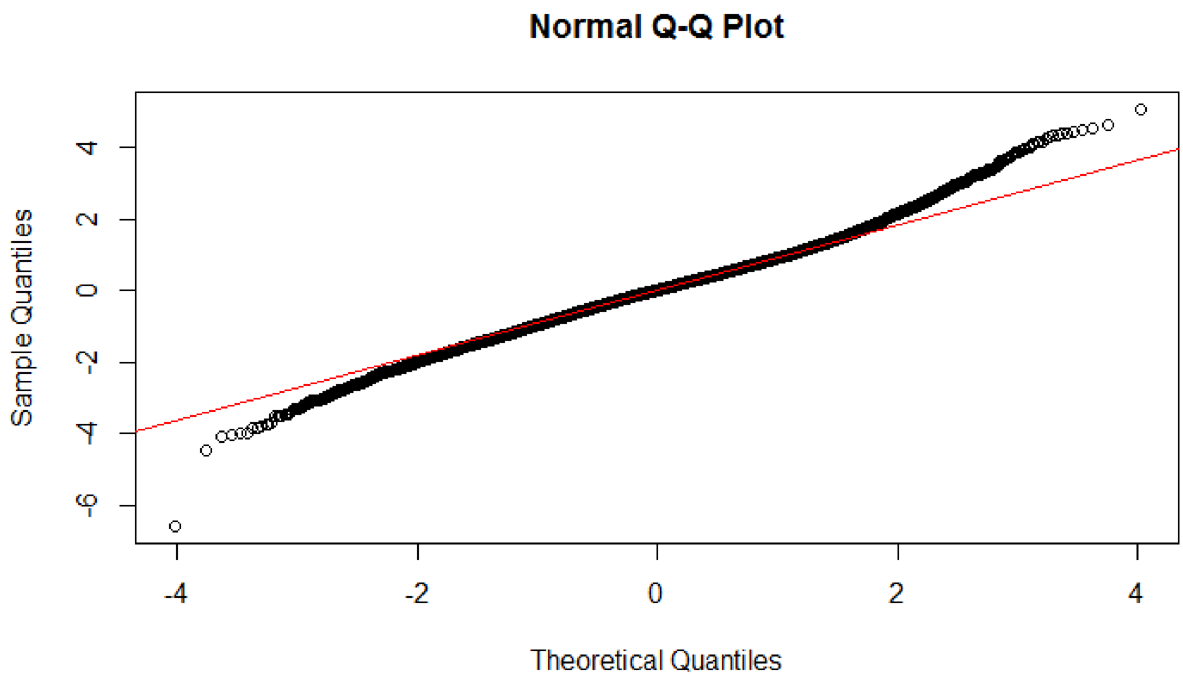


Residual verses Bathrooms



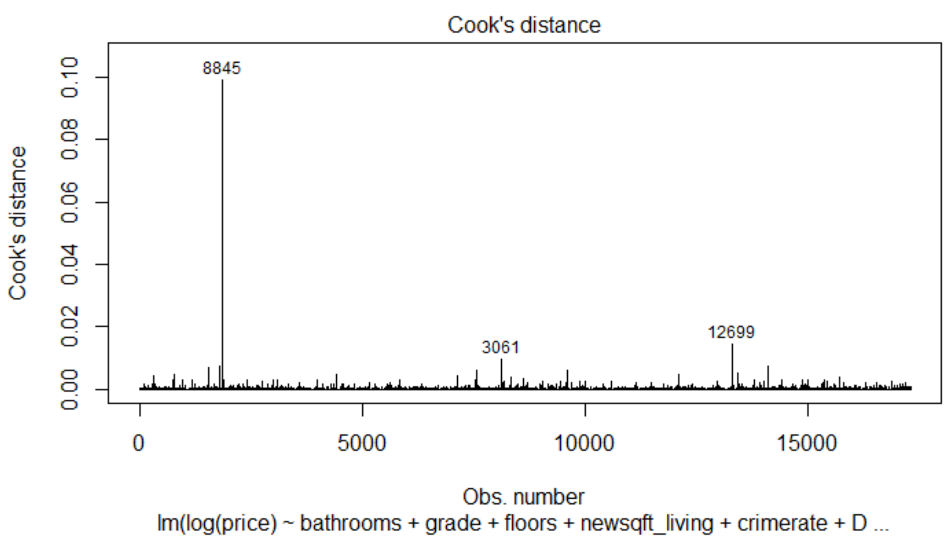
Normality Plot

The normality plot of residual analysis with constant variance except for few outliers.



Influential Points:

There are influential points in our model and removing the influential points will further improve the model.



Hypothesis Testing

Effect of Crime on Price

We used two sample Z-Test to Crime has no effect on average price per square feet

identify if there is statistically significant difference in average housing price between high crime rate vs low crime rate after breaking the price per square feet of the house into two groups based on the median value of the crime rate

Null hypothesis: Crime has no effect on average price per square feet

Alternate hypothesis: Crime is contributing on average price per square feet

Two Sample Z-Test

```
        Two-sample z-Test

data:  x and y
z = 77.541, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 104.1607 109.5629
sample estimates:
mean of x mean of y
 317.3347  210.4729
```

Conclusion: Two Sample Z- Test failed to reject null hypothesis so crime is contributing on average price per square feet.

Effect of Distance on Price

We used two sample Z-Test to identify if there is statistically significant difference in average housing price between two groups split using distance of the house from Amazon HQ.

Null hypothesis: Crime has no effect on average price per square feet

Alternate hypothesis: Crime is contributing on average price per square feet

Two Sample Z-Test

```
        Two-sample z-Test

data:  x and y
z = 65.809, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 89.88797 95.40653
sample estimates:
mean of x mean of y
 314.6669  222.0196
```

Conclusion: Two Sample Z- Test failed to reject null hypothesis so distance is contributing on average price per square feet.

6. Evaluations and Results

6.1. Evaluation Methods

The models built in this project are evaluated based on hold out method to evaluate and figure out the better model. We separate the training and test data from the data set and build regression models using the training data and evaluate the model using test data. The values predicted using the model are compared with the actual values and the models are evaluated based on RMSE.

```
> #Model 1
> rmsem1
         [,1]
[1,] 3143.924
> #Model 2
> rmsem2
         [,1]
[1,] 3140.601
> #Model 3
> rmsem3
         [,1]
[1,] 3245.949
> 
```

By comparing the RMSE of all the 3 models, Model 2 is better than all other models, as it has lesser RMSE value.

6.2. Results and Findings

The best model is evaluated based on the RMSE and model 3 is evaluated as the best model compared to all other models as it has lesser RMSE value when compared with RMSE value of other models. After successful evaluation of the model, factors which can explain the variations in price of the houses are identified

- Closer to Amazon – More you pay
- Safer the area – less the crime rate – Higher Price tag
- More number of Bathrooms – More Expensive it is
- Higher the Grade – Standardized Seattle Housing Grade – Price Increases

- Obvious More the SQFT Area Greater is the Price

7. Conclusions and Future Work

7.1. Conclusions

Thus, by evaluating the models built and hypothesis testing made in this project, we conclude that the variations in house price in King county are explained by Square feet of the living area, Grade of the house, Number of bathrooms, Crime rate of the zip code, Distance of the house from Amazon HQ, Number of floors.

Crime is contributing on average price per square feet. As crime rate increase, average price per square feet area of the house decreases.

Distance is contributing on average price per square feet. As distance increase, average price per square feet area of the house decreases.

7.2. Limitations

The data is restricted to area of King county and we can not apply this project and hypothesis on Amazon HQ distance to locations.

7.3. Potential Improvements or Future Work

We are planning to segment the zip codes based on the average house price per zip code and classify the zip codes having high graded, medium graded and low graded houses. As a future work, we are planning to predict what type and size of the house built in future will increase its sales. This will benefit the builders to build appropriate type of house which will increase the probability of the sale of the house.