ANLY 500 FINAL PROJECT REPORT

By: Huili Xing, Rumana Mohammed Myageri, Xunfan Cai

Introduction

There are numerous factors that affect what a property is worth, from basic factors, like numbers of bedrooms and bathrooms, square footage of the living area, and total floors of the house, to social factors, such as location, crime rate, and population density.

The purpose of this project is to analyze the impact of each character on house sales prices. We chose a dataset of house sales for X County between May 2014 and May 2015 to determine the most significant parameters and study how these variables are related to each other. In order to build a successful model we have to first clean the data, conduct an exploratory data analysis to understand how the variables affect each other, build a model, make multivariate regression, and use ANOVA to test the regression.

Dataset

The dataset consists of house sales for X County between May 2014 and May 2015. It includes sales price and characteristics for over 21,000 houses sold. All the variables/ characteristics are shown below:

Variable Name & Variable Description

Id —— Sales ID

Price —— Sales Price

Bedrooms —— Number of Bedrooms

Bathrooms —— Number of bathrooms

Sqft_living —— Square footage of the living area

Sqft_lot —— Square footage of the plot

Floors —— Total floors (levels) of the house

Waterfront —— Does it have a waterfront view

Condition —— How good the condition of the house is

Grade —— Overall grade given to the housing unit, based on King County grading system

Sqft_above —— Square footage of the house apart from the basement

Sqft_basement —— Square footage of the basement

Yr_built —— Year the house was built

Yr_renovated ——Year the house was renovated

Age —— Age of the house in years at the time of sale

- **Data Screening**

Accuracy: the summary shows that this data is accurate.

```
> summary(df)
       id                 date               price            bedrooms        bathrooms      sqft_living
 Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290
 1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427
 Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000   Median :2.250   Median : 1910
 Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080
 3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
 Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540
    sqft_lot           floors        waterfront            view          condition         grade          sqft_above
 Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   : 290
 1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190
 Median :   7618   Median :1.500   Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000   Median :1560
 Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657   Mean   :1788
 3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210
 Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000   Max.   :9410
 sqft_basement       yr_built     yr_renovated       zipcode           lat             long         sqft_living15
 Min.   :   0.0   Min.   :1900   Min.   :   0.0   Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
 1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
 Median :   0.0   Median :1975   Median :   0.0   Median :98065   Median :47.57   Median :-122.2   Median :1840
 Mean   : 291.5   Mean   :1971   Mean   :  84.4   Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
 3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
 Max.   :4820.0   Max.   :2015   Max.   :2015.0   Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
   sqft_lot15
 Min.   :   651
 1st Qu.:  5100
 Median :  7620
 Mean   : 12768
 3rd Qu.: 10083
 Max.   :871200
```

Missing data: We will only check the first 17 columns since the last 4 columns of data are not being used.

```
> df = df[,0:17]
> df = na.omit(df)
> summary(df)
       id                 date               price            bedrooms        bathrooms      sqft_living
 Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290
 1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427
 Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000   Median :2.250   Median : 1910
 Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080
 3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
 Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540
    sqft_lot           floors        waterfront           view          condition         grade          sqft_above
 Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   : 290
 1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190
 Median :   7618   Median :1.500   Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000   Median :1560
 Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657   Mean   :1788
 3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210
 Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000   Max.   :9410
 sqft_basement       yr_built     yr_renovated       zipcode
 Min.   :   0.0   Min.   :1900   Min.   :   0.0   Min.   :98001
 1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033
 Median :   0.0   Median :1975   Median :   0.0   Median :98065
 Mean   : 291.5   Mean   :1971   Mean   :  84.4   Mean   :98078
 3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118
 Max.   :4820.0   Max.   :2015   Max.   :2015.0   Max.   :98199
```

Outliers: We use mahal scores to find outliers in some of the columns that may have the same data distribution (normal distribution) and drop them.

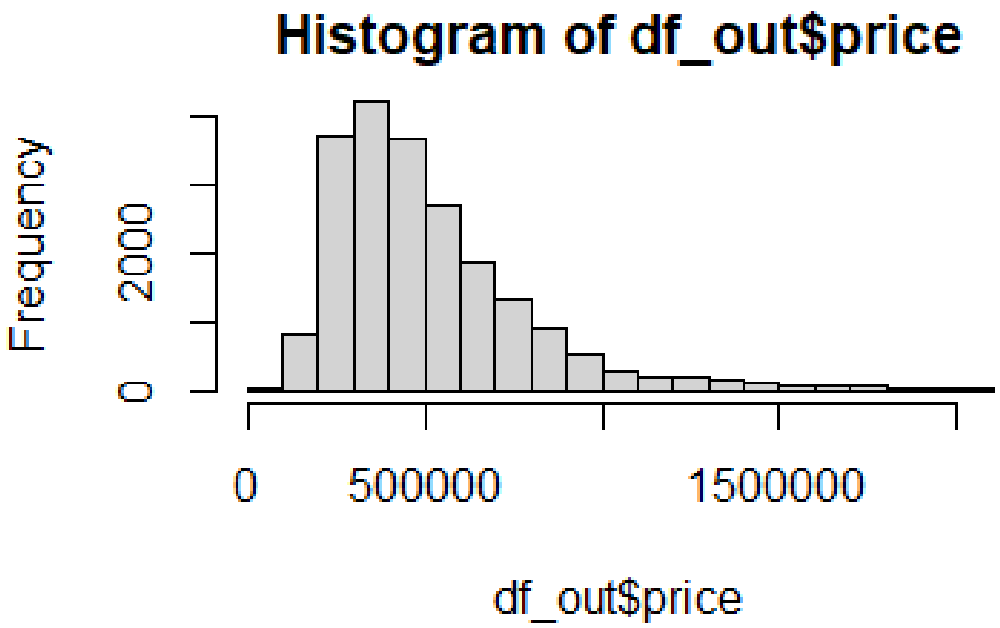
```
> mahal <- mahalanobis(df[,3:7],
+                      colMeans(df[,3:7]),
+                      cov(df[,3:7]))
> cutoff <- qchisq(1-.001, ncol(df[,3:7]))
> cutoff
[1] 20.51501
> summary(mahal < cutoff)
   Mode   FALSE    TRUE
logical     635   20978
> df_out = subset(df, mahal < cutoff)
```

- **Exploratory Data Analysis**

EDA can help us identify issues with data and inform model construction.

Exploratory Data Analysis to explore the distribution of price:

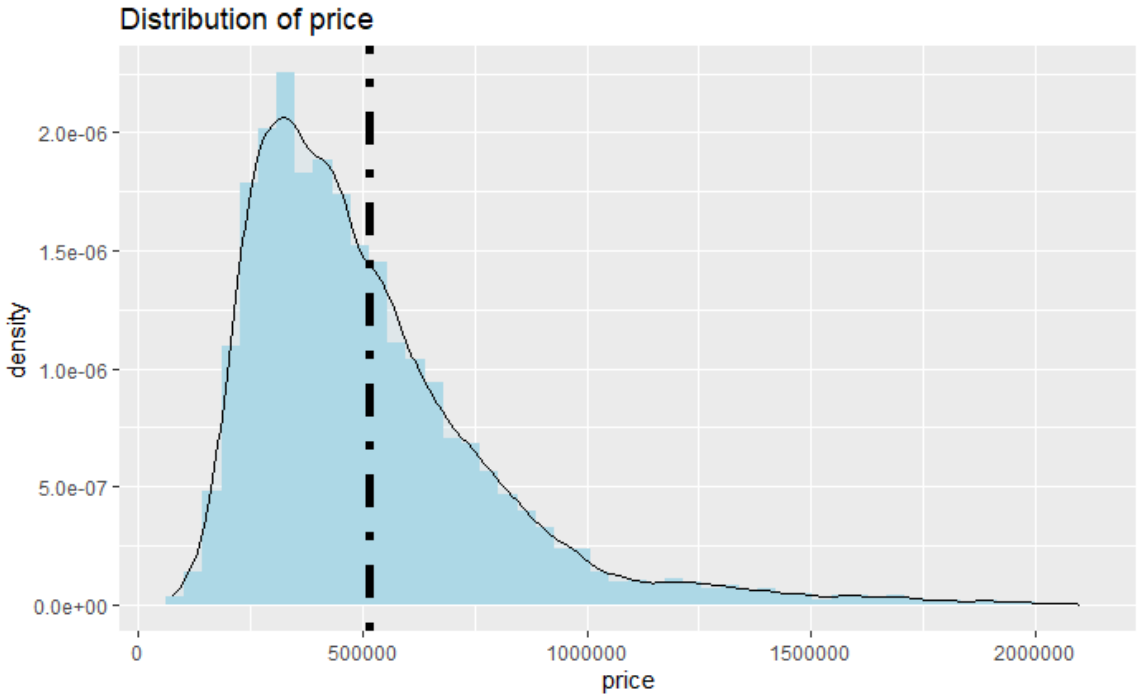
```
>hist(df_out$price)
```



Further exploring the distribution of price using ggplot:

```
>ggplot(df_out, aes(price)) +
  geom_histogram(bins = 50, aes(y = ..density..), fill = "light blue") +
  geom_density(alpha = 0.2, fill = "light blue") +
  ggtitle("Distribution of price")+
```

```
theme(axis.title = element_text(), axis.title.x = element_text()) +
geom_vline(xintercept = round(mean(df_out$price), 2), size = 2, linetype = 4)
```

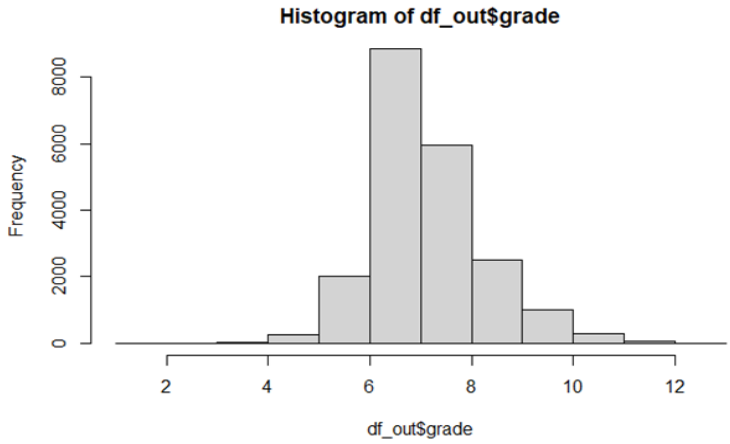


The histogram for price shows that a majority of houses are less than one million dollars. Observe also that the x-axis stretches out to two million dollars, even though there does not appear to be any houses close to that price. This is because there are a very small number of houses with prices closer to two million. The variable price is right-skewed as exhibited by the long right tail.

Exploring the distribution of other characteristics in the dataset:

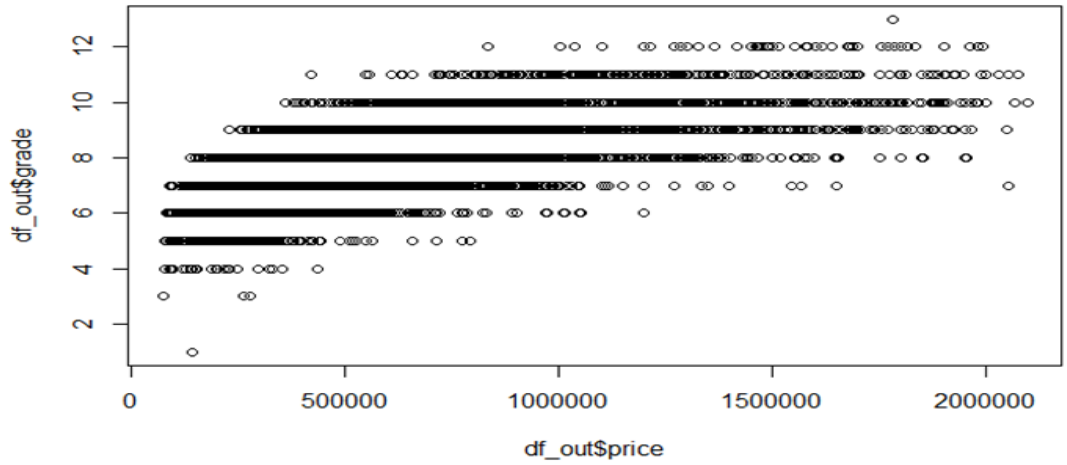
Grade

```
>hist(df_out$grade, breaks = 25)
```



The histogram of grade shows the normal distribution. So we can further investigate the relationship between grade and price.

```
>plot(df_out$price, df_out$grade)
```



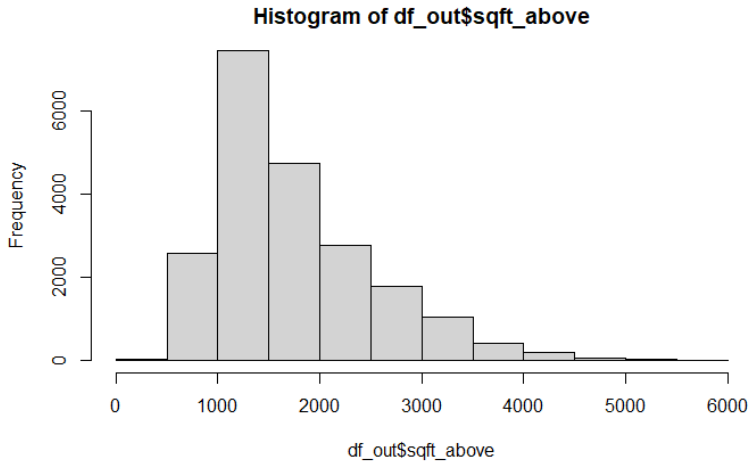
There is a positive relationship between price & grade.

Sqft_living

```
>hist(df_out$sqft_living)
```

This distribution of sqft_living is close to normal distribution. So we will further explore the relationship between price and sqft of living area using ggplot.

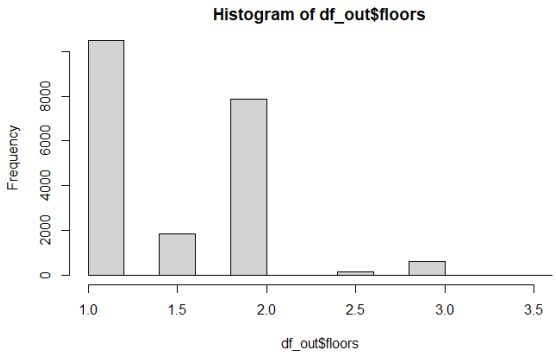
```
>hist(df_out$sqft_above)
```



The variable sqft_above is right-skewed as exhibited by the long right tail. Most of the square footage of the house apart from the basement is less than 2000.

Floors

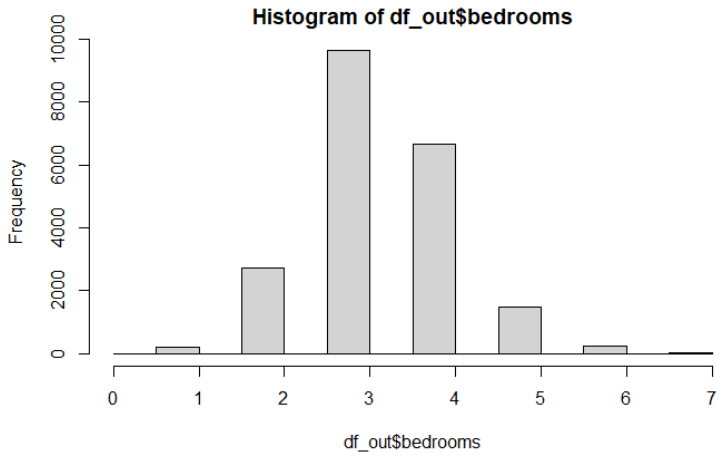
```
>hist(df_out$floors)
```



One- and two-floor make up the majority of the floors variable.

Bedrooms

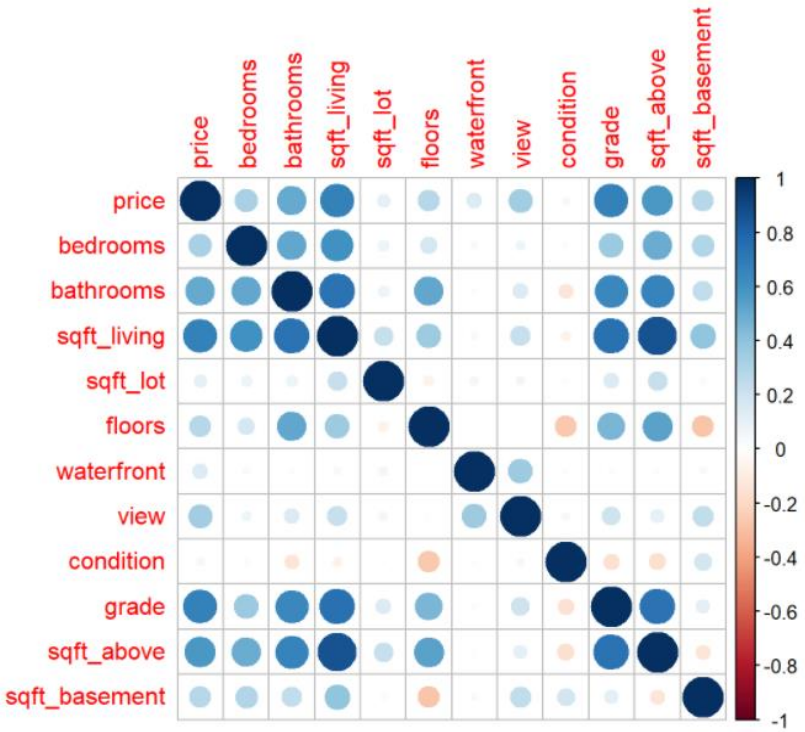
```
>hist(df_out$bedrooms, breaks = 25)
```



Three- and four-bedroom make up the majority of the bedrooms variable.

Modeling

correlation: The graph shows that sqft_living and grade have a higher correlation with price.



- **Simple Linear Regression**

From the correlation table above, we choose sqft_living as the independent variable to run a simple linear regression because it has a higher correlation with price.

>screen1 = lm(price ~ sqft_living, df_out)

>summary(screen1)

```
Call:
lm(formula = price ~ sqft_living, data = df_out)

Residuals:
    Min      1Q  Median      3Q     Max
-593021 -136222  -21476  100541 1199515

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47745.46    3804.30   12.55   <2e-16 ***
sqft_living   229.94       1.74  132.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 207100 on 20976 degrees of freedom
Multiple R-squared:  0.4543,    Adjusted R-squared:  0.4542
F-statistic: 1.746e+04 on 1 and 20976 DF,  p-value: < 2.2e-16
```

From the above results, we could know that the variance of the model is 0.4543, and sqft_living is an important feature since the P-value is pretty small and significant.

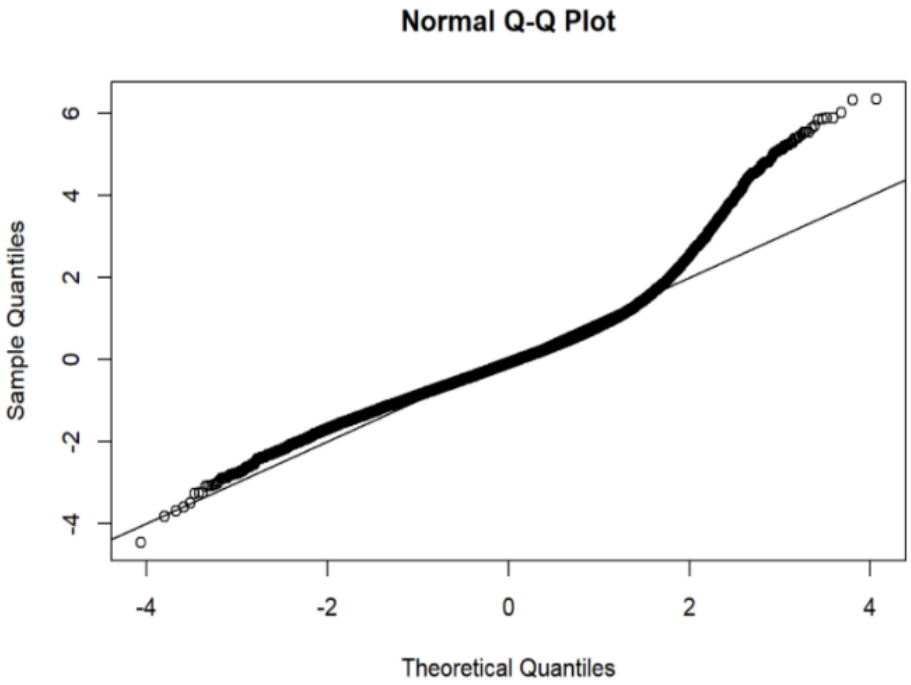
- **Assumption test**

we will do some assumption tests here to check our LR model results

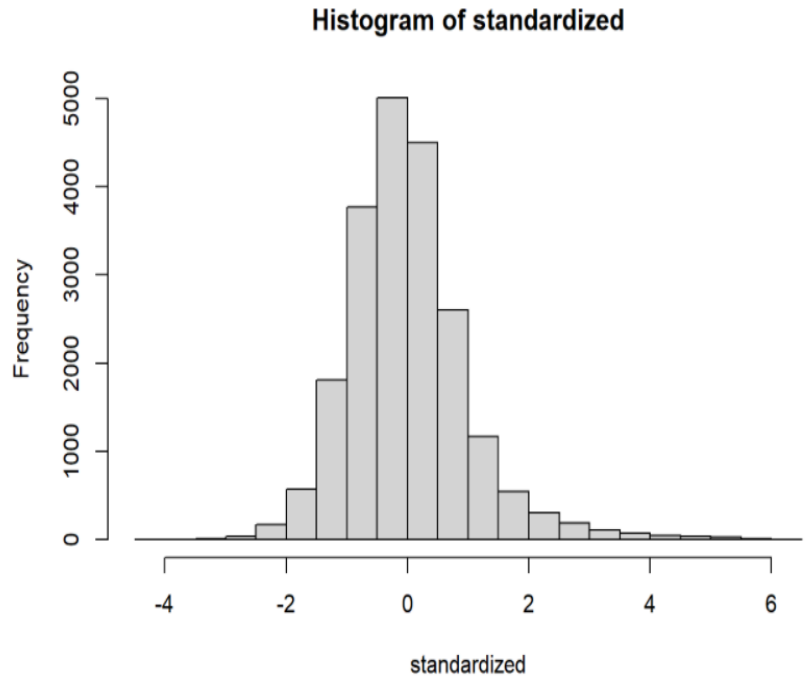
Linearity & Normality: We plot the standardized residual of the linear regression model and see the error term ϵ do have normally distributed. And the standardized plot also indicates we have achieved normality.

>qqnorm(standardized)

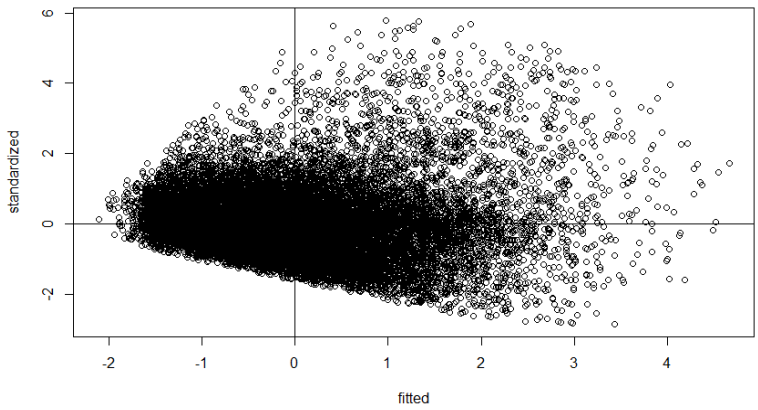
abline(0,1)



hist(standardized)

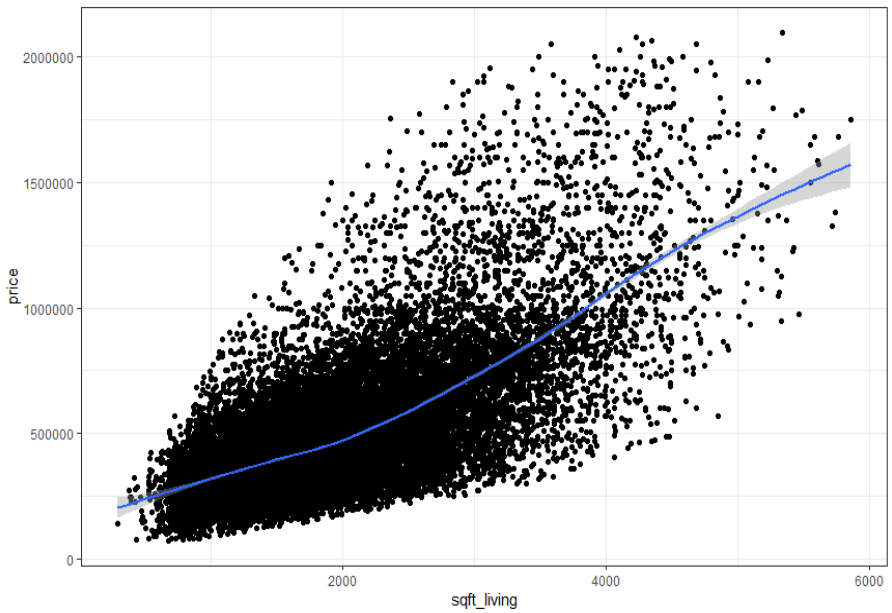


Homogeneity and Homoscedasticity: the plot shows we have met the assumption for homogeneity & homoscedasticity



scatter plot: plot the relationship between price and sqft_living

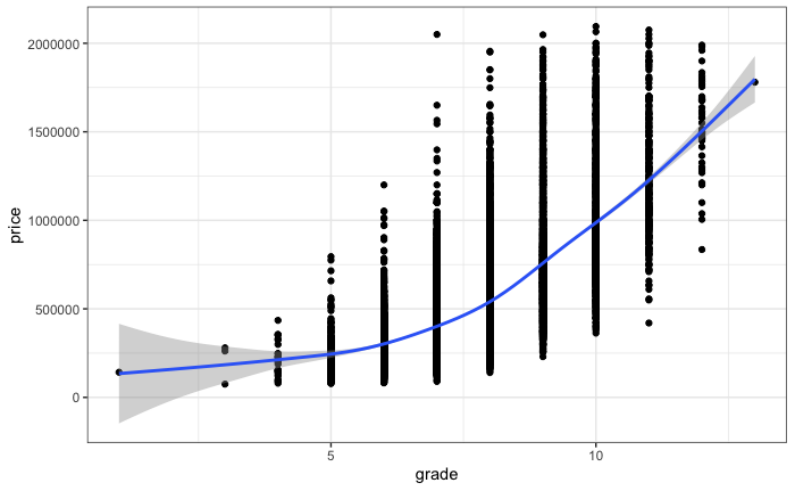
```
scat1 <- ggplot(df_out, aes(sqft_living, price)) +
    geom_point() +
    geom_smooth( ) +
    xlab("sqft_living") +
    ylab("price") +
    theme_bw()
scat1
```



There is a positive relationship between sqft_living & price.

Relationship between price & grade

```
scat2 <- ggplot(df_out, aes(grade, price)) +
    geom_point() +
    geom_smooth() +
    xlab("grade") +
    ylab("price") +
    theme_bw()
scat2
```



As seen in the graph, grade and price are also positively related.

- **More LR & MLR**

Additionally, we also want to take a look at the relationship between price and other features; we first take look at the floor:

> screen3 = lm(price ~ floors, df_out)

> summary(screen3)

```
Call:
lm(formula = price ~ floors, data = df_out)

Residuals:
    Min      1Q  Median      3Q     Max
-524215 -187059  -56637  113363 1605941

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   301481       5457   55.25   <2e-16 ***
floors        142578       3449   41.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 269600 on 20976 degrees of freedom
Multiple R-squared:  0.07534,   Adjusted R-squared:  0.0753
F-statistic:  1709 on 1 and 20976 DF,  p-value: < 2.2e-16
```

we could find that the R-squared value is much smaller than the first one, which means the performance is worse. We also use Anova to compare results

> anova(screen1, screen2)

```
Analysis of Variance Table

Model 1: price ~ sqft_living
Model 2: price ~ floors
  Res.Df        RSS Df   Sum of Sq F Pr(>F)
1  20976 9.0002e+14
2  20976 1.5249e+15  0 -6.2488e+14
```

This also tells us the model is not good.

Finally, we choose sqft_living, grade, sqftt_above, and bathrooms as independent variables to run a multiple linear regression because these variables all have a high correlation with price

> screen2 = lm(price ~ sqft_living + grade + sqft_above + bathrooms, df_out)

> summary(screen2)

```
Call:
lm(formula = price ~ sqft_living + grade + sqft_above + bathrooms,
    data = df_out)

Residuals:
    Min      1Q  Median      3Q     Max
-678968 -121158  -20398   91147 1460041

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.373e+05  1.076e+04  -49.94   <2e-16 ***
sqft_living  1.925e+02  3.617e+00   53.22   <2e-16 ***
grade        1.110e+05  1.890e+03   58.72   <2e-16 ***
sqft_above  -6.924e+01  3.558e+00  -19.46   <2e-16 ***
bathrooms   -3.008e+04  2.764e+03  -10.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 191900 on 20973 degrees of freedom
Multiple R-squared:  0.5316,    Adjusted R-squared:  0.5315
F-statistic:  5950 on 4 and 20973 DF,  p-value: < 2.2e-16
```

Results show that:

The coefficient of sqft_living and grade are positive.

The coefficient of sqft_above and bathrooms are negative.

p-value < 2.2e-16 which is significant

R-squared is **0.5316** which is higher than the R-squared of simple linear regression.

- **ANOVA**

> anova(screen1, screen2)

```
Analysis of Variance Table

Model 1: price ~ sqft_living
Model 2: price ~ sqft_living + grade + sqft_above + bathrooms
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1  20976 9.0002e+14
2  20973 7.7252e+14  3 1.275e+14 1153.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA test shows that there is a significant improvement of multiple linear regression compared to simple linear regression.

- **Conclusion**

Based on the results of previous studies, we can draw the following conclusions.

Firstly, EDA showed us that price of houses and sqft of living room and grade are normally distributed. Secondly, there are high correlations between house prices and characters like sqft_living + grade + sqft_above + bathrooms. Thirdly, the Multiple linear regression model achieved an adjusted R- Squared of 53%. Therefore, we can say that the prices of houses for county X are significantly impacted by the above features.