

**HOUSING PRICE ANALYSIS**

MSBA 305: Business Intelligence and Decision Support

Instructor: Mayank Johri

**By**

**Veena Priyadharshini Chintala**

Golden Gate University

536, Mission Street

San Francisco, CA- 94105

**TABLE OF CONTENT**

1. Introduction……………………………………………………………………………… 3

2. Data Collection ……………………………………………..…………………………… 3

3. Descriptive Analysis ………………………………………………………......................4

3. Explanatory Data Analysis ………………………………………………………............ 7

4. Multiple linear regression and Diagnostics ………………………………………...........12

5. Two sample t-test…………………………………………………………………….……13

6. ANOVA …………………………………………………………………..………………13

7. Time series Analysis…………………………………………………………………...… 14

6. Forecast Analysis…………..…………………………………………………………….. 14

8. Conclusions ……………………………………………………………………………… 15

9. Prescriptive Analytics ………………………………………………………….…………15

10. References …… ………………………………………………………….…………….. 16

11. Appendix …… …………………………………………………………………………..17

**INTRODUCTION**

Residential properties are valued for their physical, locational, neighbourhood and environmental attributes. A scenic view is an environmental amenity that affects the value of a residential property. Evidence from previous studies suggests that a view can add significantly to the value of residential properties. However, in most of the early studies view has been treated generically even though views vary by type (e.g. Ocean, Lake, Mountain, and Forest) and by quality. Failure to treat view in a more elaborate manner was due to the difficulty of obtaining data regarding view variables, which was conquered more recently by the introduction of geographical information system (GIS) data.

This study hypothesises that a view adds significantly to the value of a residential property, where a water view has the highest positive impact. Moreover, it is hypothesised that the impact of a view varies significantly with the quality of the house.

**DATA COLLECTION**

I have acquired this dataset from Kaggle. The dataset consists of 21,613 rows of House sale prices in Seattle. It includes houses sold in the year 2014 and 2015.

The different fields in the dataset include:

id - a notation for a house

date - The Date when house was sold

price –Price of the house

bedrooms - Number of Bedrooms/House

bathrooms - Number of bathrooms/bedrooms

sqft\_living - square footage of the home

floors - Total floors (levels) in house

waterfront - House which has a view to a waterfront

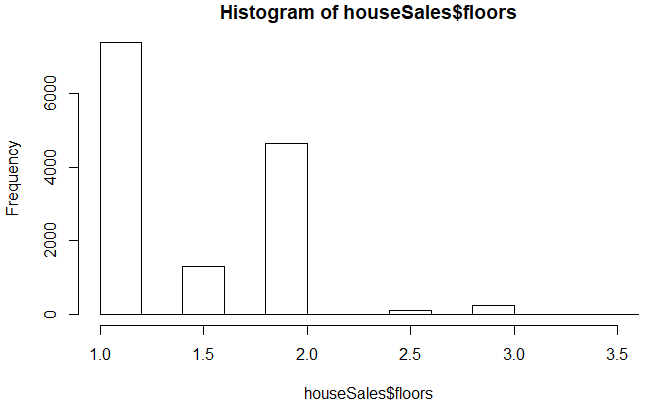
condition - How good the condition is ( Overall )

renovated - house was renovated or not.

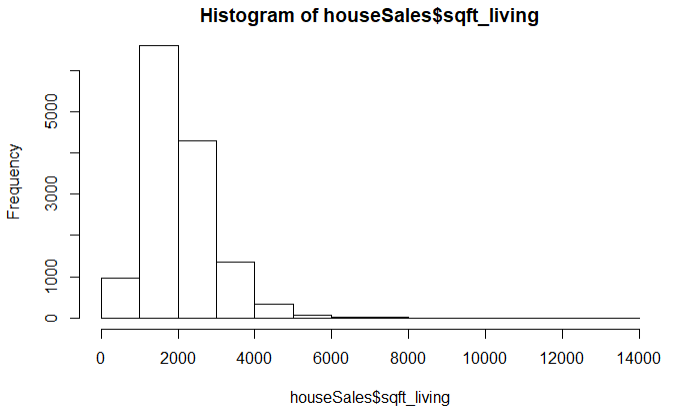
**DESCRIPTIVE ANALYSIS**

The Descriptive analysis helps us to understand the data and discover patterns, and make assumptions for further analysis. It also helps us to understand the relationship between explanatory and independent variables with the dependent variables.

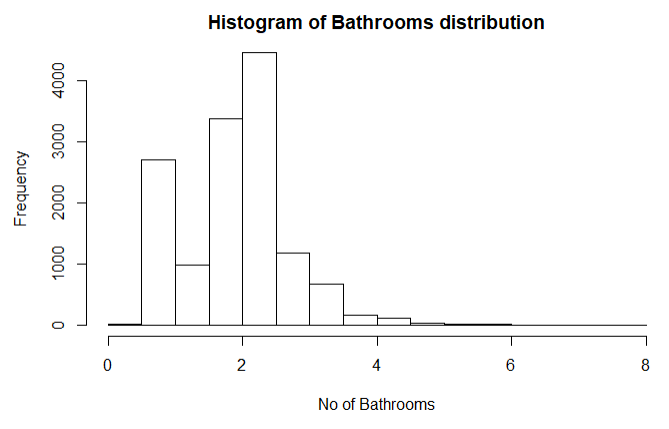
For this analysis, I am including Boxplot, Histogram, and Scattered plot to understand the data set.

****

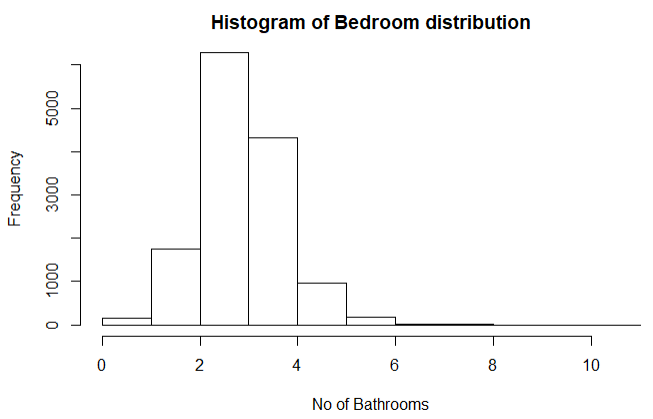
From the above graph we can see that more number of houses are built with one floor and least with 2.5 floors.

****

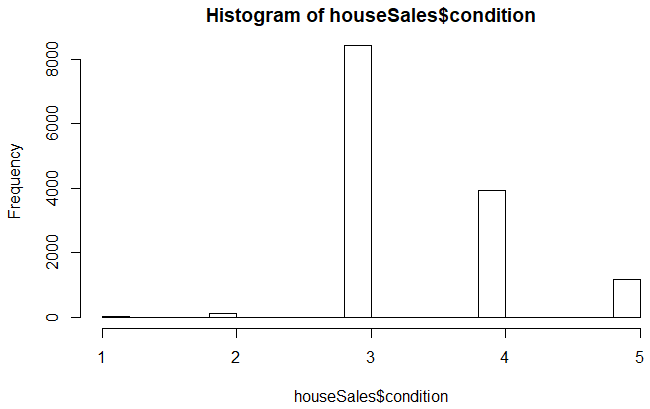
From the above graph we can see that more number of houses are built with 2000 sqft living space.

****

From the above graph we can see that more number of houses are built with 2 and 2 1/2 bathrooms.

****

From the above graph we can see that more number of houses are built with 2 to 3 bedrooms.

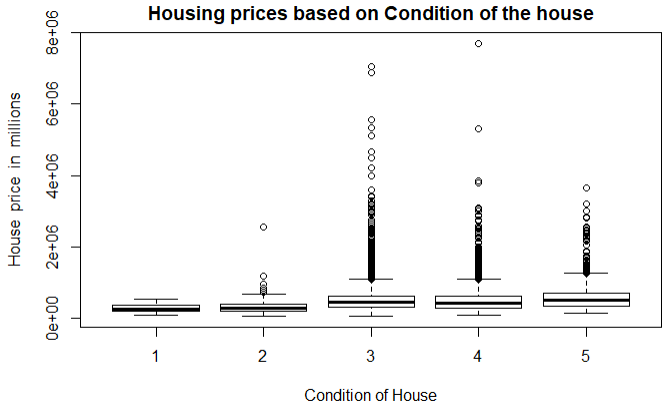
****

From the above graph we can see that more number of houses are in better condition than the best.

**EXPLANATORY DATA ANALYSIS**

The following box plot provides the relation between explanatory variables and the dependent variable. Boxplots provide a five number summary: minimum, first quartile, median, third quartile, and maximum.

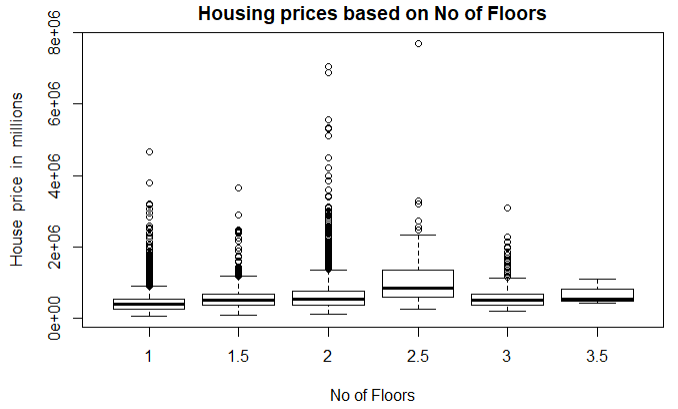
In this analysis house sale prices is examined with bedrooms, bathrooms, no of floors, renovation, waterfront, and condition of house.



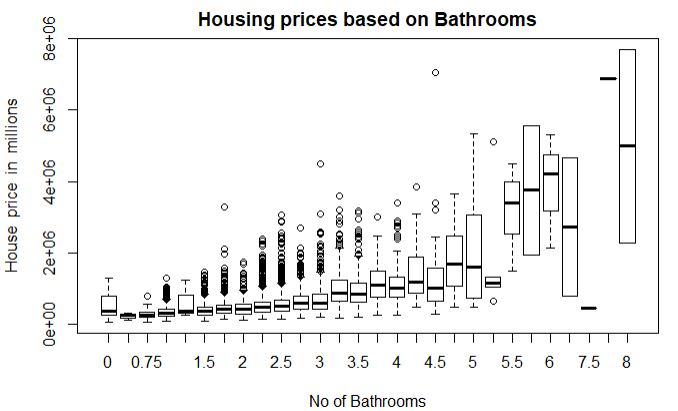
From the above graph we can see that houses that are well maintained and have good condition (i.e number 5) are costlier than others.



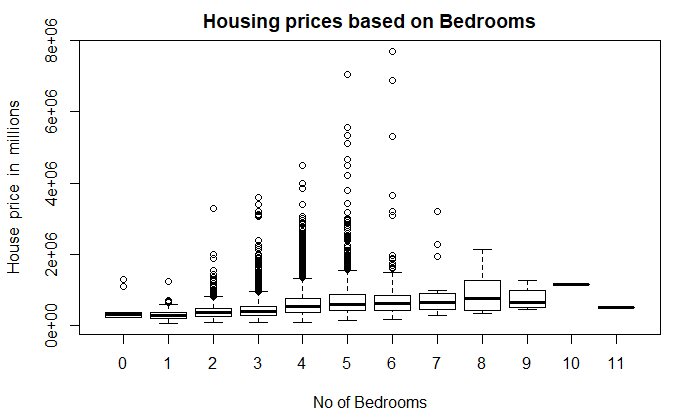
From the above graph we can see that houses that have a waterfront view are more valuable than others.



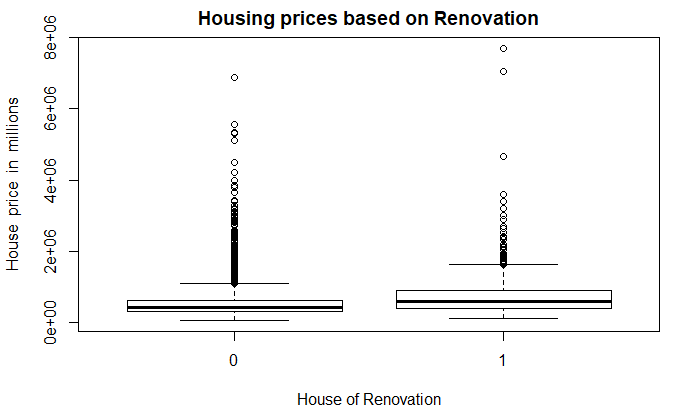
From the above graph we can see that houses with 2.5 floor are with higher price than other.



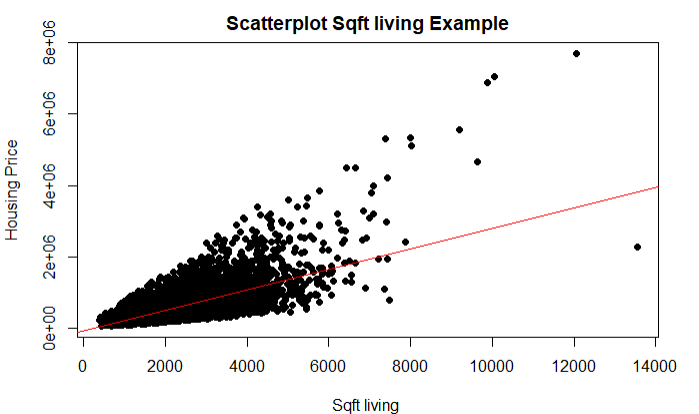
From the above graph we can see that houses with 8 bathrooms are expensive than others.



From the above graph we can see that houses with 8 bedrooms are expensive than others.



From the above graph we can see that renovated houses cost more than the non-renovated once.



The scattered plot above shows that, most of the houses with 2000 to 4000 sqft living space are priced higher than others.

**MULTIPLE LINEAR REGRESSION AND DIAGNOSTICS**

According to Crawley (2015), multiple linear regression (MLR) is a statistical method used to predict the value of a dependent variable based on the values of multiple explanatory variables.

The MLR is used to identify the explanatory variable that have a significant effect on the response variable and hence provide an equation to calculate the response variable from the independent variable.

In this dataset, house price is the response variable and bedrooms, bathrooms, sqft\_living, floors, waterfront, and condition are independent variables. Multiple linear regression analysis will help us to explain how significant the relation between response and independent variables is. The results show that only bedrooms, sqft\_living, floors, waterfront, condition has significant effect on house prices.

Based on the MLR results only 5 variables are significant for the house prices. Hence, I created the first model with dependent variable house price and the independent variables individually. In the second model, I only included the dependent variable and the independent variables which were shown significant from the MLR analysis. I created AIC (Akaike’s Information Criterion) to choose the best model. The model with the lowest AIC score is the best fit for further analysis. The result showed that the second model is the best regression model in all the four cases. Thus, I used the second model to conduct my research.

**Two regression models for house prices**

|  |  |  |
| --- | --- | --- |
|  | Dependent Variable | Independent Variable |
| Model1 | price | bedrooms, bathrooms, sqft\_living, floors, waterfront, and condition |
| Model2 | price | bedrooms, sqft\_living, floors, waterfront, condition |

Finally, I used a regression diagnostic to see if the model works well with the research.

Linear regression model has 4 assumptions and 4 diagnostic plots. I created the diagnostic plots for all the four sales groups

The four assumptions are:

1. Linearity: The relationship between the dependent variable and the mean of the independent variable is linear.

2. Homoscedasticity: The variance of the residual is the same for any value of the dependent variable.

3. Independence: Observations are independent of each other.

4. Normality: For any fixed value of the dependent variable, the independent variable is normally distributed.

There are 4 diagnostic plots showing residuals:

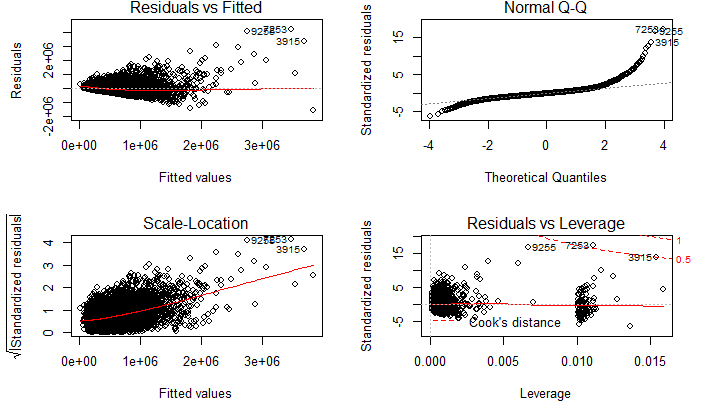
1. Normal Q-Q: This plot shows if residuals are normally distributed.

2. Residuals vs Fitted: This plot shows if residuals have non-linear patterns.

3. Scale-Location: This is to check the assumption of equal variance.

4. Residuals vs Leverage: This plot helps us to find influential cases.

The results for house price show that the points almost lay on the line hence we can say that, the independent variables are normally distributed.



**TWO SAMPLE T-TEST**

Independent 2 sample t-test are statistical methods for comparing the difference in means for 2 populations. The two sample t-test is conducted with a 95% confidence interval on house price on the explanatory variable renovated.

**Null Hypothesis: There is no relationship between house price and renovated.**

The goal of this test is to check if renovated home have any effect on the house price. Here the dependent variable is house price and the independent variable is a renovated which has 2 levels (0 and 1(renovated house).

**The result of the two sample t-test is:**

**Null Hypothesis: There is no relationship between house price and renovated.**

|  |  |  |
| --- | --- | --- |
| **Mean not renovated** | **Mean renovated house** | **P – value** |
| 521267.6 | 765771.4 | 2.2e-16 |

From the above chart, we can see that house price have a p-value less than 0.05 hence we reject the null hypothesis and accept that there is a significant difference between the means. Hence we have evidence to believe that renovation of a house increases the price of the house.

**ANOVA TEST**

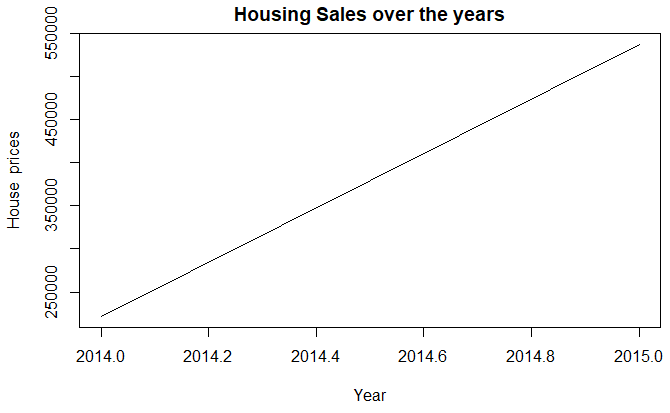
Anova is a statistical method used for comparing the means of two or more independent population. In this project, one-way ANOVA is conducted to check if there is any significant difference in the house sales due to the independent variables waterfront.

**Null Hypothesis: Waterfront has no effect on House Prices.**

The results show that the p-value is less than 0.05 and has a high F- value, hence we reject the null hypothesis and have evidence to believe that Waterfront has significant effect on raising house prices.

**TIME SERIES AND FORECAST ANALYSIS**

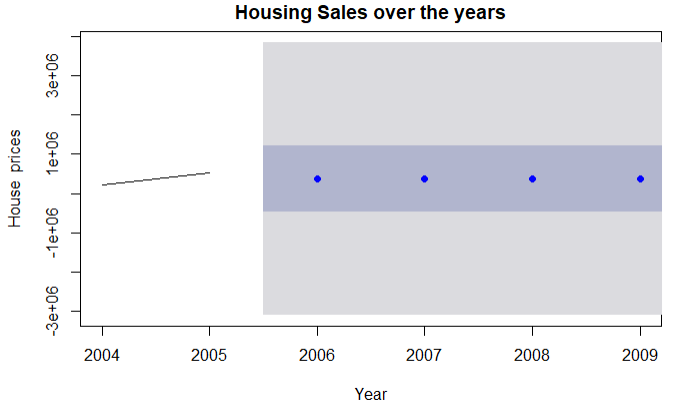
Any data analyses over regular intervals of time are described as time series analysis. Time series analysis is very important for the business to know the trend and also to perform the forecast analysis. In this dataset, I have analyzed the house prices from the year 2014 to 2015.



From the above graph, it shows that House prices are constantly increasing from the year 2014 to 2015.

**FORECAST ANALYSIS**

Forecasting is the process of using statistical methods to predict the future values based on the historical trend in the time series. In this project forecast analysis helps us to predict the housing prices for the next 5 years. It is required to install forecast package to do forecast analysis.



From the above graph, we can see that house prices are constant over the next 5 years. The area around the darkly shaded region is the place where the maximum housing prices will be.

**CONCLUSION**

The test shows that, price of the houses which are in very good condition are expensive than others. The analysis also shows that the view has positive influence in increasing the prices of the houses. The houses with 8 bedrooms and 8 bathrooms on 2.5 floors have house prices more than others. The test also shows that, renovation of houses has positive influence on raising the housing prices.

Finally, the time series plots show that the house prices increases constantly from year 2014 to 2015. The forecast shows that the house prices remains a constant for the next 5 years.

**PRESCRIPTIVE ANALYTICS**

* If houses are built with 2000 to 4000 sqfl living space with 2.5 floors and 8 bedrooms and 8 bathrooms which has a waterfront view can increase the housing prices.
* Renovating the house can raise the housing price.

**REFERENCE**

Kabacoff, R. (August, 2011). R in Action (2nd ed.). Manning Publications Co. : Publisher.

Bourassa, S.C., M. Hoesli, and J. Sun. 2004. What’s in a View? *Environment and Planning A.* 36(8): pp 1427-50.

Darling, A.H. 1973. Measuring Benefits Generated by Urban Water Parks. *Land Economics.* 49: pp 22-34.

LeSage, J.P. 1999. *Spatial Econometrics*. Unpublished manuscript available online from <http://www.econ.utoledo.edu/faculty/lesage/lesage.html>.

Paterson, R., and K. Boyle. 2002. Out of Sight, Out of Mind? Using GIS to Incorporate Visibility in Hedonic Property Value Models. *Land Economics.* 78: pp 417-25.

**APPENDIX**

# R Notebook

This is an [R Markdown](http://rmarkdown.rstudio.com/) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the Run button within the chunk or by placing your cursor inside it and pressing Ctrl+Shift+Enter.

house<- read.csv("C:\\Users\\Veena Chintala\\Desktop\\305\\Final project\\kc\_house\_data.csv")

head(house)

|  |
| --- |
|  |

|  | **id**  <dbl> | **date**  <int> | **numericdate**  <fctr> | **price**  <dbl> | **bedrooms**  <int> | **bathrooms**  <dbl> | **sqft\_living**  <int> | **sqft\_lot**  <int> | **floors**  <dbl> |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 7129300520 | 2014 | 20141013T000000 | 221900 | 3 | 1.00 | 1180 | 5650 | 1 |  |
| 2 | 6414100192 | 2014 | 20141209T000000 | 538000 | 3 | 2.25 | 2570 | 7242 | 2 |  |
| 3 | 5631500400 | 2015 | 20150225T000000 | 180000 | 2 | 1.00 | 770 | 10000 | 1 |  |
| 4 | 2487200875 | 2014 | 20141209T000000 | 604000 | 4 | 3.00 | 1960 | 5000 | 1 |  |
| 5 | 1954400510 | 2015 | 20150218T000000 | 510000 | 3 | 2.00 | 1680 | 8080 | 1 |  |
| 6 | 7237550310 | 2014 | 20140512T000000 | 1230000 | 4 | 4.50 | 5420 | 101930 | 1 |  |

6 rows | 1-10 of 17 columns

houseSales<- na.omit(house)

boxplot( houseSales$price~houseSales$bedrooms, main="Housing prices based on Bedrooms", xlab="No of Bedrooms", ylab="House price in millions")

boxplot(houseSales$price~houseSales$bathrooms, main="Housing prices based on Bathrooms", xlab="No of Bathrooms", ylab="House price in millions")

hist(houseSales$bedrooms, main="Histogram of Bedroom distribution", xlab="No of Bedrooms")

hist(houseSales$bathrooms, main="Histogram of Bathrooms distribution", xlab="No of Bathrooms")

plot(houseSales$sqft\_living, houseSales$price, main="Scatterplot Sqft living Example", xlab="Sqft living ", ylab=" Housing Price ", pch=19)

abline(lm(houseSales$price ~ houseSales$sqft\_living), col="red") *# regression line (y~x)*

hist(houseSales$sqft\_living)

boxplot(houseSales$price~ houseSales$floors, main="Housing prices based on No of Floors", xlab="No of Floors", ylab="House price in millions")

hist(houseSales$floors)

boxplot(houseSales$price~houseSales$waterfront, main="Housing prices based on Waterfromt", xlab="Waterfront(1)", ylab="House price in millions")

boxplot(houseSales$price~ houseSales$condition, main="Housing prices based on Condition of the house", xlab="Condition of House", ylab="House price in millions")

hist(houseSales$condition)

boxplot(houseSales$price~houseSales$yr\_renovated, main="Housing prices based on Renovation", xlab="House of Renovation", ylab="House price in millions")

model1<- lm(houseSales$price ~ houseSales$bedrooms+houseSales$bathrooms+houseSales$sqft\_living+houseSales$floors+houseSales$waterfront+houseSales$condition)

summary(model1)

##

## Call:

## lm(formula = houseSales$price ~ houseSales$bedrooms + houseSales$bathrooms +

## houseSales$sqft\_living + houseSales$floors + houseSales$waterfront +

## houseSales$condition)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1560658 -134032 -18654 101277 4216443

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -1.029e+05 1.521e+04 -6.764 1.39e-11 \*\*\*

## houseSales$bedrooms -5.520e+04 2.921e+03 -18.897 < 2e-16 \*\*\*

## houseSales$bathrooms 3.371e+03 4.630e+03 0.728 0.46659

## houseSales$sqft\_living 3.042e+02 3.791e+00 80.254 < 2e-16 \*\*\*

## houseSales$floors 1.348e+04 4.875e+03 2.765 0.00569 \*\*

## houseSales$waterfront 8.288e+05 2.503e+04 33.110 < 2e-16 \*\*\*

## houseSales$condition 4.774e+04 3.254e+03 14.673 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 246900 on 13682 degrees of freedom

## Multiple R-squared: 0.5498, Adjusted R-squared: 0.5496

## F-statistic: 2785 on 6 and 13682 DF, p-value: < 2.2e-16

model2<- lm(houseSales$price ~ houseSales$bedrooms+houseSales$sqft\_living+houseSales$floors+houseSales$waterfront+houseSales$condition)

summary(model2)

##

## Call:

## lm(formula = houseSales$price ~ houseSales$bedrooms + houseSales$sqft\_living +

## houseSales$floors + houseSales$waterfront + houseSales$condition)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1562788 -134357 -18363 101588 4217840

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -1.022e+05 1.518e+04 -6.731 1.75e-11 \*\*\*

## houseSales$bedrooms -5.476e+04 2.859e+03 -19.155 < 2e-16 \*\*\*

## houseSales$sqft\_living 3.059e+02 3.063e+00 99.869 < 2e-16 \*\*\*

## houseSales$floors 1.466e+04 4.598e+03 3.189 0.00143 \*\*

## houseSales$waterfront 8.290e+05 2.503e+04 33.119 < 2e-16 \*\*\*

## houseSales$condition 4.767e+04 3.252e+03 14.658 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 246900 on 13683 degrees of freedom

## Multiple R-squared: 0.5498, Adjusted R-squared: 0.5496

## F-statistic: 3342 on 5 and 13683 DF, p-value: < 2.2e-16

AIC(model1,model2)

|  |
| --- |
|  |

|  | **df**  <dbl> | **AIC**  <dbl> |
| --- | --- | --- |
| model1 | 8 | 378797.2 |
| model2 | 7 | 378795.7 |

par(mfrow=c(2,2))

plot(model2)

abline(model2)

## Warning in abline(model2): only using the first two of 6 regression

## coefficients

t.test(houseSales$price~houseSales$yr\_renovated)

##

## Welch Two Sample t-test

##

## data: houseSales$price by houseSales$yr\_renovated

## t = -9.3187, df = 632.65, p-value < 2.2e-16

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

## -296027.8 -192979.8

## sample estimates:

## mean in group 0 mean in group 1

## 521267.6 765771.4

anovosum<- aov(houseSales$price~houseSales$waterfront)

summary(anovosum)

## Df Sum Sq Mean Sq F value Pr(>F)

## houseSales$waterfront 1 1.444e+14 1.444e+14 1158 <2e-16 \*\*\*

## Residuals 13687 1.708e+15 1.248e+11

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

housesalesTs<- ts(house$TimeSeries\_price, start = c(2014,1), end = c(2015))

plot(housesalesTs, main="Housing Sales over the years", xlab="Year", ylab="House prices")

**library**(forecast)

etsForecst<- ets(housesalesTs)

plot(forecast(housesalesTs, 4), main="Housing prices over the years", xlab="Year", ylab="House prices")