Predicting House Prices

Zahraa Alshalal

2023-03-17

## libraries

# packages  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(plotly)

## Loading required package: ggplot2

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.1.8  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0  
## ✔ readr 2.1.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ plotly::filter() masks dplyr::filter(), stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:plotly':  
##   
## select  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(DataExplorer)  
library(Hmisc)

## Loading required package: lattice  
## Loading required package: survival  
## Loading required package: Formula  
##   
## Attaching package: 'Hmisc'  
##   
## The following object is masked from 'package:plotly':  
##   
## subplot  
##   
## The following objects are masked from 'package:dplyr':  
##   
## src, summarize  
##   
## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(polycor)  
library(corrplot)

## corrplot 0.92 loaded

library(htmlwidgets)  
library('IRdisplay')

## end of library (ignore top)

#import data  
housedata = read.csv("~/Desktop/spring23/math448/project/house\_data.csv")  
# look at the data  
glimpse(housedata)

## Rows: 21,613  
## Columns: 21  
## $ id <dbl> 7129300520, 6414100192, 5631500400, 2487200875, 19544005…  
## $ date <chr> "20141013T000000", "20141209T000000", "20150225T000000",…  
## $ price <dbl> 221900, 538000, 180000, 604000, 510000, 1225000, 257500,…  
## $ bedrooms <int> 3, 3, 2, 4, 3, 4, 3, 3, 3, 3, 3, 2, 3, 3, 5, 4, 3, 4, 2,…  
## $ bathrooms <dbl> 1.00, 2.25, 1.00, 3.00, 2.00, 4.50, 2.25, 1.50, 1.00, 2.…  
## $ sqft\_living <int> 1180, 2570, 770, 1960, 1680, 5420, 1715, 1060, 1780, 189…  
## $ sqft\_lot <int> 5650, 7242, 10000, 5000, 8080, 101930, 6819, 9711, 7470,…  
## $ floors <dbl> 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1…  
## $ waterfront <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ view <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0,…  
## $ condition <int> 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 4, 4,…  
## $ grade <int> 7, 7, 6, 7, 8, 11, 7, 7, 7, 7, 8, 7, 7, 7, 7, 9, 7, 7, 7…  
## $ sqft\_above <int> 1180, 2170, 770, 1050, 1680, 3890, 1715, 1060, 1050, 189…  
## $ sqft\_basement <int> 0, 400, 0, 910, 0, 1530, 0, 0, 730, 0, 1700, 300, 0, 0, …  
## $ yr\_built <int> 1955, 1951, 1933, 1965, 1987, 2001, 1995, 1963, 1960, 20…  
## $ yr\_renovated <int> 0, 1991, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ zipcode <int> 98178, 98125, 98028, 98136, 98074, 98053, 98003, 98198, …  
## $ lat <dbl> 47.5112, 47.7210, 47.7379, 47.5208, 47.6168, 47.6561, 47…  
## $ long <dbl> -122.257, -122.319, -122.233, -122.393, -122.045, -122.0…  
## $ sqft\_living15 <int> 1340, 1690, 2720, 1360, 1800, 4760, 2238, 1650, 1780, 23…  
## $ sqft\_lot15 <int> 5650, 7639, 8062, 5000, 7503, 101930, 6819, 9711, 8113, …

#Summary of Data  
summary(housedata)

## id date price bedrooms   
## Min. :1.000e+06 Length:21613 Min. : 75000 Min. : 0.000   
## 1st Qu.:2.123e+09 Class :character 1st Qu.: 321950 1st Qu.: 3.000   
## Median :3.905e+09 Mode :character Median : 450000 Median : 3.000   
## Mean :4.580e+09 Mean : 540088 Mean : 3.371   
## 3rd Qu.:7.309e+09 3rd Qu.: 645000 3rd Qu.: 4.000   
## Max. :9.900e+09 Max. :7700000 Max. :33.000   
## bathrooms sqft\_living sqft\_lot floors   
## Min. :0.000 Min. : 290 Min. : 520 Min. :1.000   
## 1st Qu.:1.750 1st Qu.: 1427 1st Qu.: 5040 1st Qu.:1.000   
## Median :2.250 Median : 1910 Median : 7618 Median :1.500   
## Mean :2.115 Mean : 2080 Mean : 15107 Mean :1.494   
## 3rd Qu.:2.500 3rd Qu.: 2550 3rd Qu.: 10688 3rd Qu.:2.000   
## Max. :8.000 Max. :13540 Max. :1651359 Max. :3.500   
## waterfront view condition grade   
## Min. :0.000000 Min. :0.0000 Min. :1.000 Min. : 1.000   
## 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.: 7.000   
## Median :0.000000 Median :0.0000 Median :3.000 Median : 7.000   
## Mean :0.007542 Mean :0.2343 Mean :3.409 Mean : 7.657   
## 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:4.000 3rd Qu.: 8.000   
## Max. :1.000000 Max. :4.0000 Max. :5.000 Max. :13.000   
## sqft\_above sqft\_basement yr\_built yr\_renovated   
## Min. : 290 Min. : 0.0 Min. :1900 Min. : 0.0   
## 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951 1st Qu.: 0.0   
## Median :1560 Median : 0.0 Median :1975 Median : 0.0   
## Mean :1788 Mean : 291.5 Mean :1971 Mean : 84.4   
## 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997 3rd Qu.: 0.0   
## Max. :9410 Max. :4820.0 Max. :2015 Max. :2015.0   
## zipcode lat long sqft\_living15   
## Min. :98001 Min. :47.16 Min. :-122.5 Min. : 399   
## 1st Qu.:98033 1st Qu.:47.47 1st Qu.:-122.3 1st Qu.:1490   
## Median :98065 Median :47.57 Median :-122.2 Median :1840   
## Mean :98078 Mean :47.56 Mean :-122.2 Mean :1987   
## 3rd Qu.:98118 3rd Qu.:47.68 3rd Qu.:-122.1 3rd Qu.:2360   
## 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

#unique(housedata$yr\_renovated)

#Structure of Data   
str(housedata)

## 'data.frame': 21613 obs. of 21 variables:  
## $ id : num 7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...  
## $ date : chr "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...  
## $ price : num 221900 538000 180000 604000 510000 ...  
## $ bedrooms : int 3 3 2 4 3 4 3 3 3 3 ...  
## $ bathrooms : num 1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...  
## $ sqft\_living : int 1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...  
## $ sqft\_lot : int 5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...  
## $ floors : num 1 2 1 1 1 1 2 1 1 2 ...  
## $ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ view : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ condition : int 3 3 3 5 3 3 3 3 3 3 ...  
## $ grade : int 7 7 6 7 8 11 7 7 7 7 ...  
## $ sqft\_above : int 1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...  
## $ sqft\_basement: int 0 400 0 910 0 1530 0 0 730 0 ...  
## $ yr\_built : int 1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...  
## $ yr\_renovated : int 0 1991 0 0 0 0 0 0 0 0 ...  
## $ zipcode : int 98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...  
## $ lat : num 47.5 47.7 47.7 47.5 47.6 ...  
## $ long : num -122 -122 -122 -122 -122 ...  
## $ sqft\_living15: int 1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...  
## $ sqft\_lot15 : int 5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...

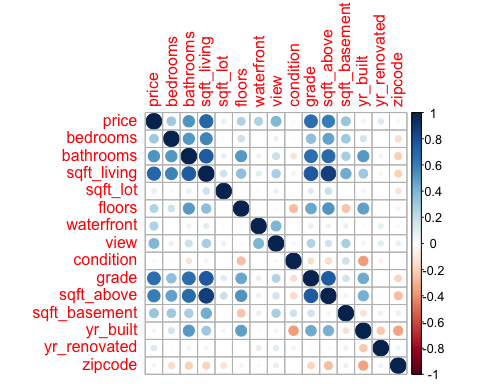
#Checking Null Value  
sum(is.na(housedata))

## [1] 0

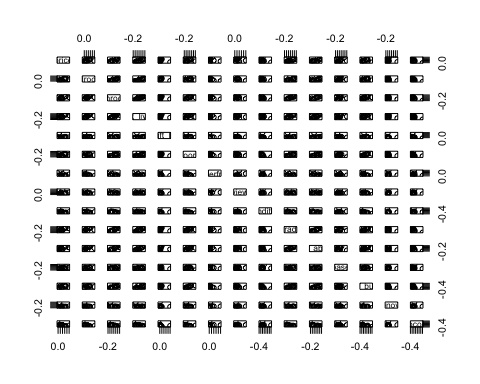
#Correlation  
data\_cor = cor(housedata[,c(3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)])  
data\_cor

## price bedrooms bathrooms sqft\_living sqft\_lot  
## price 1.00000000 0.308349598 0.52513751 0.70203505 0.089660861  
## bedrooms 0.30834960 1.000000000 0.51588364 0.57667069 0.031703243  
## bathrooms 0.52513751 0.515883638 1.00000000 0.75466528 0.087739662  
## sqft\_living 0.70203505 0.576670693 0.75466528 1.00000000 0.172825661  
## sqft\_lot 0.08966086 0.031703243 0.08773966 0.17282566 1.000000000  
## floors 0.25679389 0.175428935 0.50065317 0.35394929 -0.005200991  
## waterfront 0.26636943 -0.006582479 0.06374363 0.10381782 0.021603683  
## view 0.39729349 0.079531852 0.18773702 0.28461119 0.074710106  
## condition 0.03636179 0.028472104 -0.12498193 -0.05875259 -0.008958250  
## grade 0.66743426 0.356966725 0.66498253 0.76270448 0.113621124  
## sqft\_above 0.60556730 0.477600161 0.68534248 0.87659660 0.183512281  
## sqft\_basement 0.32381602 0.303093375 0.28377003 0.43504297 0.015286202  
## yr\_built 0.05401153 0.154178069 0.50601944 0.31804877 0.053080367  
## yr\_renovated 0.12643379 0.018840823 0.05073898 0.05536293 0.007643505  
## zipcode -0.05320285 -0.152668487 -0.20386627 -0.19943004 -0.129574486  
## floors waterfront view condition grade  
## price 0.256793888 0.266369434 0.39729349 0.036361789 0.66743426  
## bedrooms 0.175428935 -0.006582479 0.07953185 0.028472104 0.35696673  
## bathrooms 0.500653173 0.063743629 0.18773702 -0.124981933 0.66498253  
## sqft\_living 0.353949290 0.103817818 0.28461119 -0.058752587 0.76270448  
## sqft\_lot -0.005200991 0.021603683 0.07471011 -0.008958250 0.11362112  
## floors 1.000000000 0.023698320 0.02944382 -0.263767946 0.45818251  
## waterfront 0.023698320 1.000000000 0.40185735 0.016653157 0.08277491  
## view 0.029443820 0.401857351 1.00000000 0.045989737 0.25132058  
## condition -0.263767946 0.016653157 0.04598974 1.000000000 -0.14467367  
## grade 0.458182514 0.082774914 0.25132058 -0.144673671 1.00000000  
## sqft\_above 0.523884710 0.072074592 0.16764934 -0.158213616 0.75592294  
## sqft\_basement -0.245704542 0.080587939 0.27694658 0.174104914 0.16839182  
## yr\_built 0.489319425 -0.026161086 -0.05343985 -0.361416562 0.44696320  
## yr\_renovated 0.006338401 0.092884837 0.10391729 -0.060617787 0.01441428  
## zipcode -0.059120642 0.030284728 0.08482692 0.003025524 -0.18486209  
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode  
## price 0.60556730 0.32381602 0.05401153 0.126433793 -0.053202854  
## bedrooms 0.47760016 0.30309338 0.15417807 0.018840823 -0.152668487  
## bathrooms 0.68534248 0.28377003 0.50601944 0.050738978 -0.203866274  
## sqft\_living 0.87659660 0.43504297 0.31804877 0.055362927 -0.199430043  
## sqft\_lot 0.18351228 0.01528620 0.05308037 0.007643505 -0.129574486  
## floors 0.52388471 -0.24570454 0.48931942 0.006338401 -0.059120642  
## waterfront 0.07207459 0.08058794 -0.02616109 0.092884837 0.030284728  
## view 0.16764934 0.27694658 -0.05343985 0.103917288 0.084826917  
## condition -0.15821362 0.17410491 -0.36141656 -0.060617787 0.003025524  
## grade 0.75592294 0.16839182 0.44696320 0.014414281 -0.184862093  
## sqft\_above 1.00000000 -0.05194331 0.42389835 0.023284688 -0.261189977  
## sqft\_basement -0.05194331 1.00000000 -0.13312410 0.071322902 0.074844608  
## yr\_built 0.42389835 -0.13312410 1.00000000 -0.224873518 -0.346869178  
## yr\_renovated 0.02328469 0.07132290 -0.22487352 1.000000000 0.064357057  
## zipcode -0.26118998 0.07484461 -0.34686918 0.064357057 1.000000000

corrplot(data\_cor)



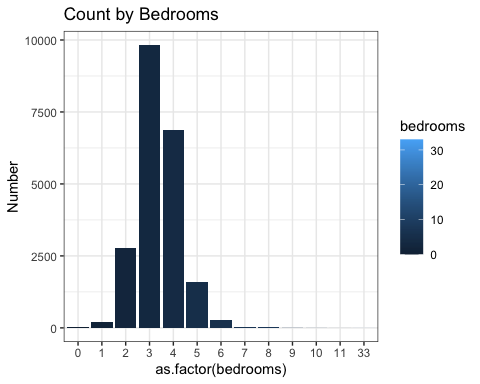
pairs(data\_cor)



# the number of bedroom's in houses  
bedrooms\_count = housedata%>%group\_by(bedrooms)%>%summarise(Number=length(bedrooms))  
head(bedrooms\_count)

## # A tibble: 6 × 2  
## bedrooms Number  
## <int> <int>  
## 1 0 13  
## 2 1 199  
## 3 2 2760  
## 4 3 9824  
## 5 4 6882  
## 6 5 1601

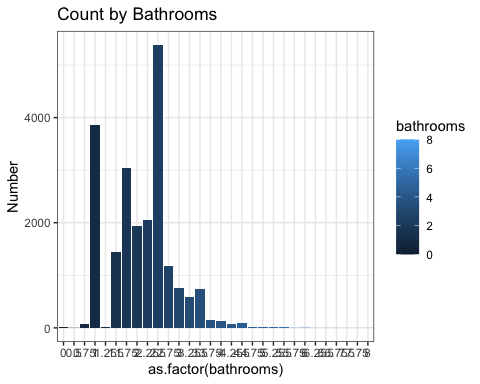
ggplot(bedrooms\_count, aes(as.factor(bedrooms), Number, fill=bedrooms)) + geom\_bar(stat='Identity') + labs(title='Count by Bedrooms') + theme\_bw()



# the number of bathrooms  
bathrooms\_count = housedata%>%group\_by(bathrooms)%>%summarise(Number=length(bathrooms))  
head(bathrooms\_count)

## # A tibble: 6 × 2  
## bathrooms Number  
## <dbl> <int>  
## 1 0 10  
## 2 0.5 4  
## 3 0.75 72  
## 4 1 3852  
## 5 1.25 9  
## 6 1.5 1446

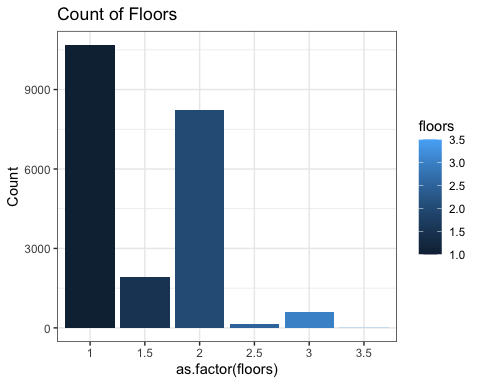
ggplot(bathrooms\_count, aes(as.factor(bathrooms), Number, fill=bathrooms)) + geom\_bar(stat='Identity') + labs(title='Count by Bathrooms') + theme\_bw()



#the number of floors are there  
housefloor = housedata%>%group\_by(floors)%>%summarise(housefloor=length(floors))  
housefloor

## # A tibble: 6 × 2  
## floors housefloor  
## <dbl> <int>  
## 1 1 10680  
## 2 1.5 1910  
## 3 2 8241  
## 4 2.5 161  
## 5 3 613  
## 6 3.5 8

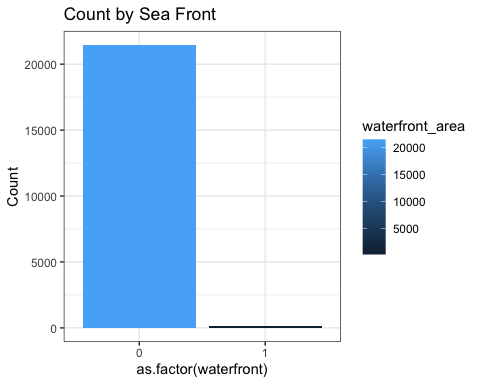
ggplot(housefloor, aes(as.factor(floors), housefloor, fill=floors)) + geom\_bar(stat='Identity') + labs(title='Count of Floors', y='Count') + theme\_bw()



#waterfront   
waterfront = housedata%>%group\_by(waterfront)%>%summarise(waterfront\_area=length(waterfront))  
waterfront

## # A tibble: 2 × 2  
## waterfront waterfront\_area  
## <int> <int>  
## 1 0 21450  
## 2 1 163

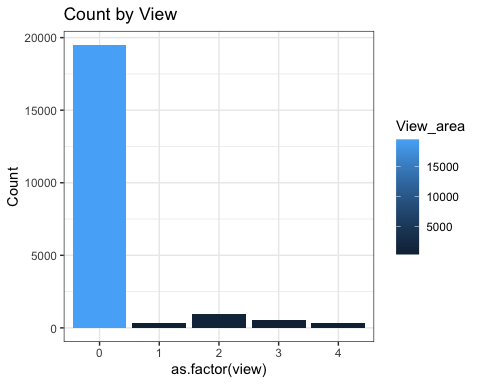
ggplot(waterfront, aes(as.factor(waterfront), waterfront\_area, fill=waterfront\_area)) + geom\_bar(stat='Identity') + labs(title='Count by Sea Front', y='Count') + theme\_bw()



#view  
view = housedata%>%group\_by(view)%>%summarise(View\_area=length(view))  
view

## # A tibble: 5 × 2  
## view View\_area  
## <int> <int>  
## 1 0 19489  
## 2 1 332  
## 3 2 963  
## 4 3 510  
## 5 4 319

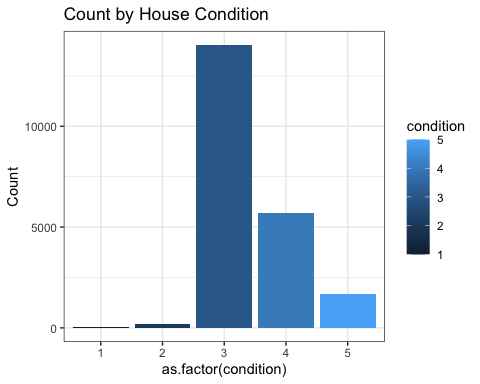
ggplot(view, aes(as.factor(view), View\_area, fill=View\_area)) + geom\_bar(stat='Identity') + labs(title='Count by View', y='Count') + theme\_bw()



#Condition of the house  
condition = housedata%>%group\_by(condition)%>%summarise(house\_condition=length(condition))  
condition

## # A tibble: 5 × 2  
## condition house\_condition  
## <int> <int>  
## 1 1 30  
## 2 2 172  
## 3 3 14031  
## 4 4 5679  
## 5 5 1701

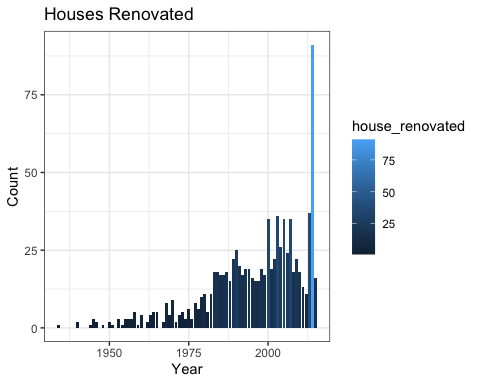
ggplot(condition, aes(as.factor(condition), house\_condition, fill=condition)) + geom\_bar(stat='Identity') + labs(title='Count by House Condition', y='Count') + theme\_bw()



#Count of house renovated by Year  
renovated\_Year = housedata%>%group\_by(yr\_renovated)%>%summarise(house\_renovated=length(yr\_renovated))%>%arrange(desc(house\_renovated))  
  
head(renovated\_Year)

## # A tibble: 6 × 2  
## yr\_renovated house\_renovated  
## <int> <int>  
## 1 0 20699  
## 2 2014 91  
## 3 2013 37  
## 4 2003 36  
## 5 2000 35  
## 6 2005 35

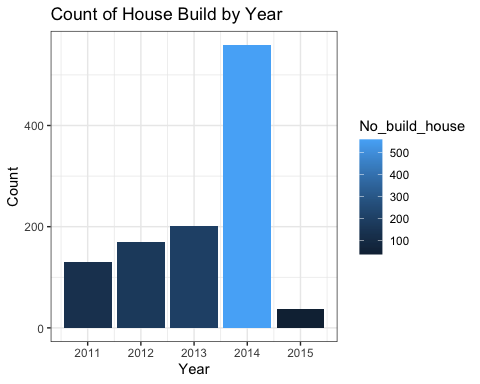
#the number of house renovated after 2010  
  
renovated\_aftr\_2010<-subset(renovated\_Year, renovated\_Year$yr\_renovated>1900)  
  
ggplot(renovated\_aftr\_2010, aes(yr\_renovated ,house\_renovated, fill=house\_renovated)) + geom\_bar(stat='Identity') + labs(title='Houses Renovated', y='Count', x='Year') + theme\_bw()



#Count of house Build by Year   
house\_build\_year = housedata%>%group\_by(yr\_built)%>%summarise(No\_build\_house=length(yr\_built))%>%arrange(desc(No\_build\_house))  
  
head(house\_build\_year)

## # A tibble: 6 × 2  
## yr\_built No\_build\_house  
## <int> <int>  
## 1 2014 559  
## 2 2006 454  
## 3 2005 450  
## 4 2004 433  
## 5 2003 422  
## 6 1977 417

#house built after 2010   
  
house\_build\_aftr\_2010 = filter(house\_build\_year, house\_build\_year$yr\_built>2010)  
  
ggplot(house\_build\_aftr\_2010, aes(yr\_built, No\_build\_house, fill=No\_build\_house)) + geom\_bar(stat='Identity') + labs(title='Count of House Build by Year', x='Year', y='Count') + theme\_bw()



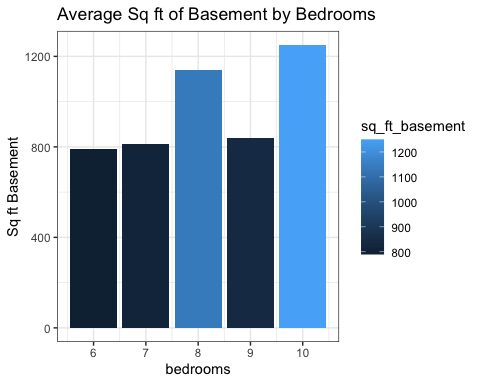
#the average sq ft of basement by bedrooms  
sqft\_basement\_besment = housedata%>%group\_by(bedrooms)%>%summarise(sq\_ft\_basement=mean(sqft\_basement))%>%top\_n(5)

## Selecting by sq\_ft\_basement

head(sqft\_basement\_besment)

## # A tibble: 5 × 2  
## bedrooms sq\_ft\_basement  
## <int> <dbl>  
## 1 6 790.  
## 2 7 814.  
## 3 8 1139.  
## 4 9 840   
## 5 10 1250

ggplot(sqft\_basement\_besment, aes(bedrooms ,sq\_ft\_basement, fill = sq\_ft\_basement)) + geom\_bar(stat='Identity') + theme\_bw() + labs(title='Average Sq ft of Basement by Bedrooms', y='Sq ft Basement')



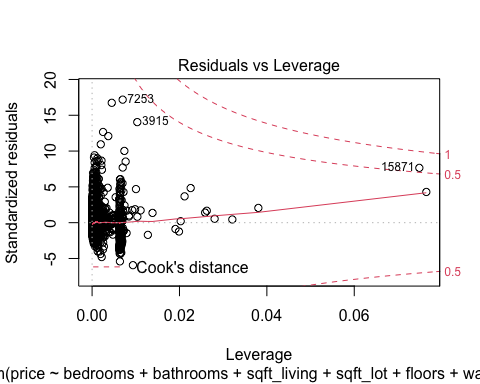
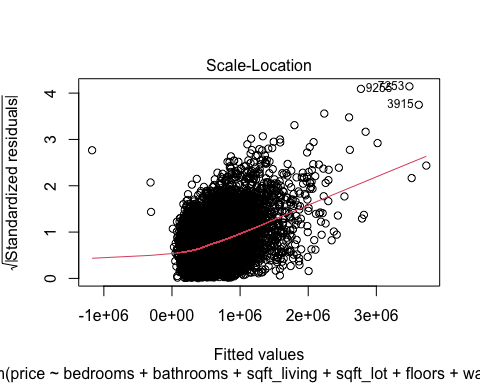
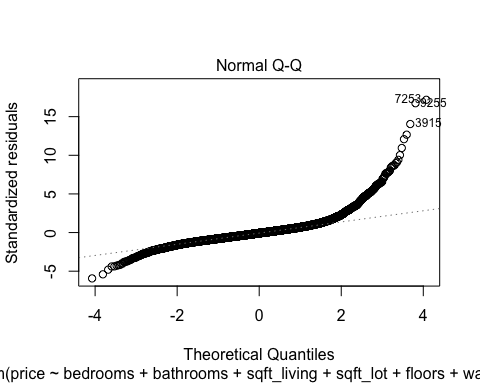
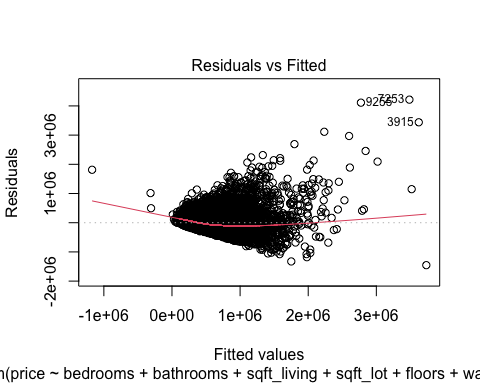
# Training the Model  
model = lm(price~ bedrooms + bathrooms + sqft\_living + sqft\_lot + floors + waterfront + condition, data = housedata)  
  
summary(model)

##   
## Call:  
## lm(formula = price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot +   
## floors + waterfront + condition, data = housedata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1453628 -137375 -18771 102613 4214690   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.175e+05 1.242e+04 -9.459 < 2e-16 \*\*\*  
## bedrooms -5.539e+04 2.263e+03 -24.480 < 2e-16 \*\*\*  
## bathrooms 1.033e+04 3.646e+03 2.832 0.00463 \*\*   
## sqft\_living 3.003e+02 3.013e+00 99.676 < 2e-16 \*\*\*  
## sqft\_lot -3.608e-01 4.133e-02 -8.729 < 2e-16 \*\*\*  
## floors 1.618e+04 3.706e+03 4.365 1.28e-05 \*\*\*  
## waterfront 7.845e+05 1.953e+04 40.161 < 2e-16 \*\*\*  
## condition 5.080e+04 2.679e+03 18.959 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 246200 on 21605 degrees of freedom  
## Multiple R-squared: 0.5505, Adjusted R-squared: 0.5504   
## F-statistic: 3781 on 7 and 21605 DF, p-value: < 2.2e-16

resdata = residuals(model)  
  
resdata = as.data.frame(resdata)  
  
head(resdata)

## resdata  
## 1 -25659.36  
## 2 -155515.45  
## 3 1754.29  
## 4 55098.35  
## 5 102831.07  
## 6 -241939.81

plot(model)



options(repr.plot.width=8, repr.plot.height=4)  
plot\_histogram(housedata)

