# Assignment3

## Question2

library(tidyverse)

set.seed(17053777)

kc\_house\_data <- read\_csv("kc\_house\_data.csv")

my\_houses <- kc\_house\_data %>% sample\_n(20000)

my\_houses <- my\_houses %>%

select(-id, -date, -lat, -long, -zipcode, sqft\_above,

-sqft\_living15, -sqft\_lot15)

## Question3

library(GGally)

my\_houses %>% cor()

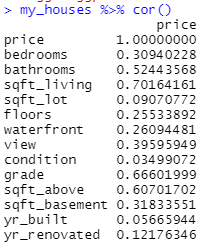
ggpairs(my\_houses, columns = c("price","sqft\_living", "sqft\_above", "grade"))

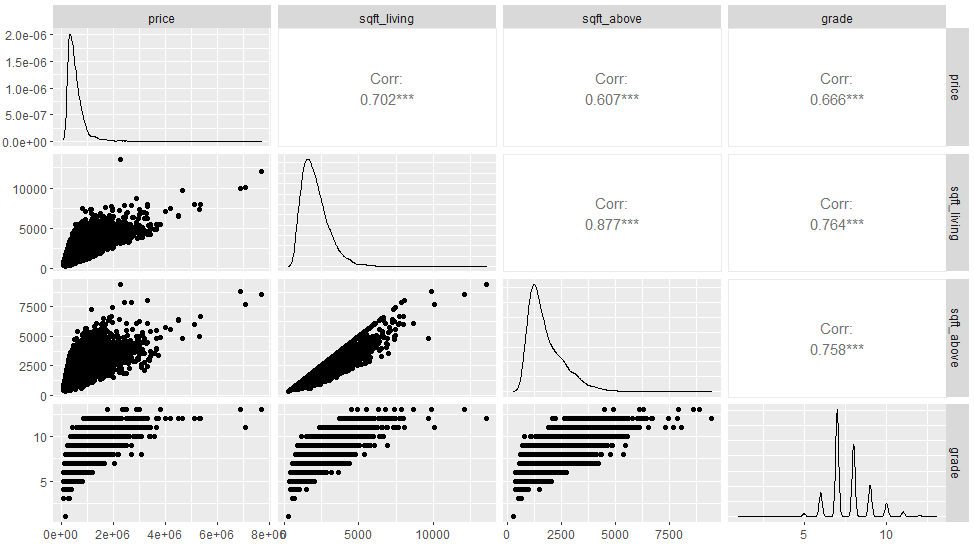
# Based on the correlations, I choose sqft\_living, grade and sqtl\_above as predictors.

# We observe that there is a positive linear relationship between price and sqft\_living with 0.702 strong correlation.

#There is a positive linear relationship between price and sqft\_above and the correlation is 0.607 which is a moderate correlation.

#We can observe there is a positive linear relationship between price and grade but the trend is curved, so they have a quadratic relationship but still linear with 0.666 moderate correlation.

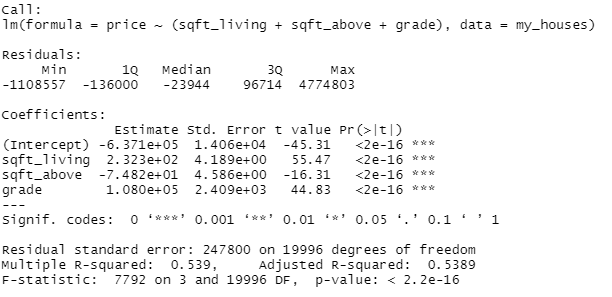




## Question3

m1 <- lm(price ~ (sqft\_living + sqft\_above + grade), data = my\_houses)

summary(m1)



Coefficients:

Price = 237.1 \* sqft\_living – 74.82 \* sqft\_above + 10800 \* grade

Adjusted R^2 = 0.5389

Residual standard error = 247800 on 19996 degrees od freedom

## Question5

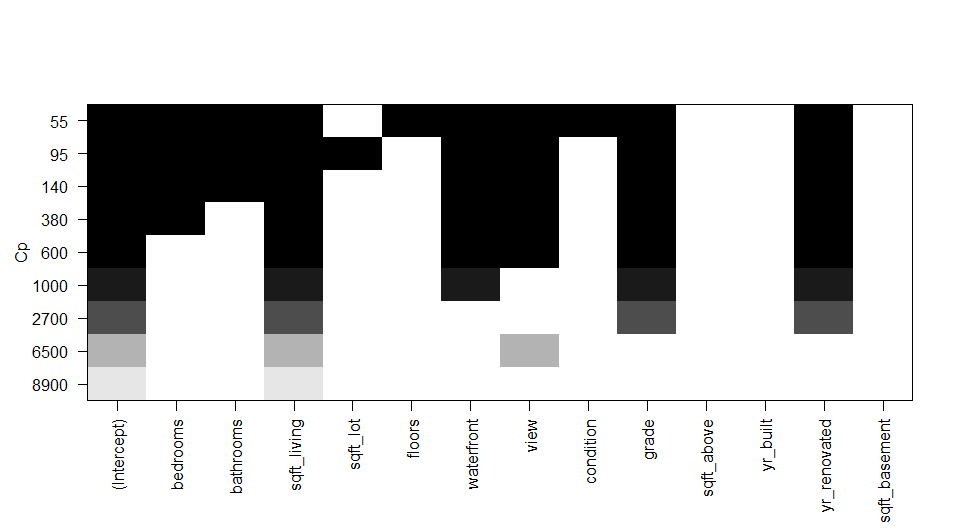
library(leaps)

par(mfrow = c(1, 1))

all\_mods <- regsubsets(price ~ ., data = my\_houses)

plot(all\_mods, scale = 'Cp')

# As we know, the lower the Cp, the better the model. We can observe that sqft\_lot, sqft\_above, yr\_build and sqft\_basement are the variables that need to be dropped. So the predictors which need to be contained in the best model is bedrooms, bathrooms, sqft\_living, floors, waterfront, view, condition, grade and yr\_renovated.



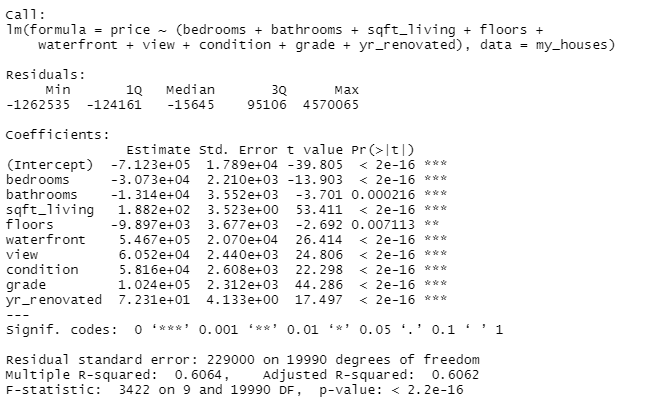
## Question6

m2 <- lm(price ~ (bedrooms + bathrooms + sqft\_living + floors + waterfront + view + condition + grade + yr\_renovated), data = my\_houses)

summary(m2)

# adjusted R^2 = 0.6062, residual standard error = 229000 on 19990 degrees of freedom

# From m1 to m2, we hae more variables as predictors. We observe that the adjusted R^2 increased from 0.5389 to 0.6062 and the residual standard error is smaller. As the adjusted R^2 increases, we can say the model is improved. As residual standard error decreases, we can say the predictions are better and m2 is better than m1.



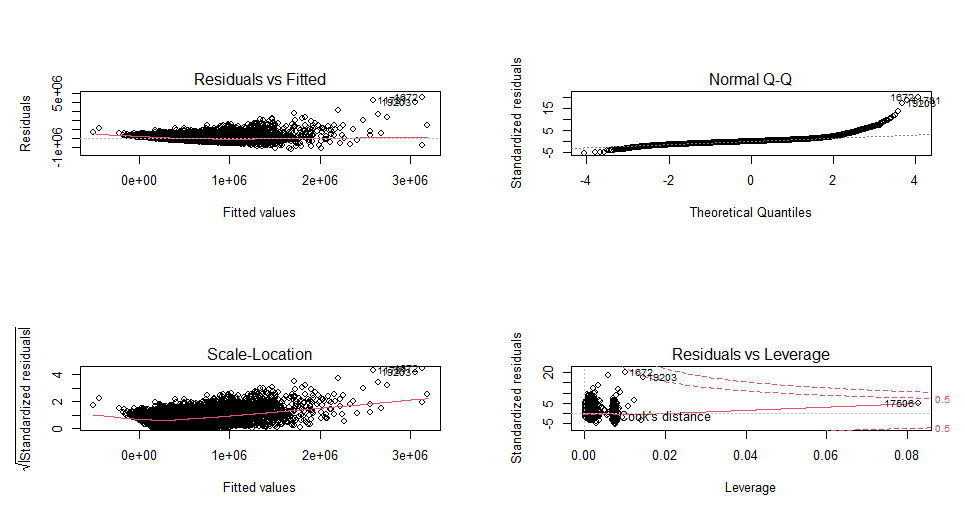
## Question7

par(mfrow = c(2, 2))

plot(m2)

# From Normal Q-Q, we observe that after theoretical qualities is greater than 2, the observation points are further way from predicted line. The residuals increase as the qualities increase.

# From Residuals vs Leverage, we observe that there are some points which are close to the red dash line 0.5, which could be considered as outliers.



## Question8

new\_houses <- tibble(bedrooms = c(3, 4), bathrooms = c(1.5, 2.5),

sqft\_living = c(1200, 1920), sqft\_lot = c(15606, 8562),

floors = c(1, 2), waterfront = c(0, 0), view = c(0, 0),

condition = c(3, 4), grade = c(7, 7),

sqft\_basement = c(0, 0), yr\_built = c(1985, 1994),

yr\_renovated = c(0, 0))

predict(m2, new\_houses, interval = 'confidence')

predict(m2, new\_houses, interval = 'prediction')

# the prediction interval is for an individual outcome drawn from the distribution of the response.

# he confidence interval is for an estimate of the mean of the distribution of the response.

