R CODE

#load in the libraries

library(readxl)

library(caret)

library(dplyr)

library(tidyverse)

#Read in the Data

ames <- read\_excel("ames.xlsx")

#converting into factor

str(ames)

ames1 <- ames

ames1 %>% mutate\_if(is.character, as.factor) -> ames2#Assigned a new table as ames2 where all character is factor for analysis.

str(ames2)

#sorting Ascending

sort(colnames(ames2))

#solving Data Quality issues

#Year built

boxplot(ames2$Year.Built)

ames2 %>% mutate(Year.Built=replace(Year.Built, Year.Built<1950, NA),

Year.Built=replace(Year.Built, Year.Built>2010,NA))-> ames2\_clean #used mutate to solve DQ issues of year bulit and created new table with ames2\_clean

boxplot(ames2\_clean$Year.Built) # to check the accuracy

#House style

sum(is.na(ames2\_clean$House.Style)) #we have zero NA values

table(ames2\_clean$House.Style)

plot(ames2\_clean$House.Style) # can see values throught scatter plot, round off is not required

#overall condition

sum(is.na(ames2\_clean$Overall.Cond)) #no NA found

plot(ames2\_clean$Sale.Price, ames2\_clean$Overall.Cond) # see a attached plot

ames2\_clean %>% mutate(Overall.Cond=replace(Overall.Cond, Overall.Cond > 9, NA),

Overall.Cond=replace(Overall.Cond, Overall.Cond < 1, NA)) -> ames3\_clean #anything above 9 or less than 1 to replace it by NA and in a new table ames3\_clean

plot(ames3\_clean$Overall.Cond, ames3\_clean$Sale.Price)

#Garage Area

sum(is.na(ames3\_clean$Garage.Area)) #found one NA/missing values

ames3\_clean%>%filter(!is.na(Garage.Area)) ->ames4\_clean #removed missing values and created new table ames3\_clean

sum(is.na(ames4\_clean$Garage.Area)) #

#Greater Living Area

sum(is.na(ames3\_clean$Gr.Liv.Area))

ames4\_clean%>%filter(!is.na(Gr.Liv.Area)) ->ames5\_clean #created new table with removed NA values and stored it in ames5\_clean

sum(is.na(ames5\_clean$Gr.Liv.Area)) # found zero mssing vlaues

#full bathroom

sum(is.na(ames5\_clean$Full.Bath)) #found zero missing values

#Sale price

plot(ames5\_clean$Sale.Price)

filter(ames5\_clean, Sale.Price>750000) #2 values more than 750000

sum(is.na(ames5\_clean$Sale.Price))

ames5\_clean %>% mutate(Sale.Price=replace(Sale.Price, Sale.Price > 750000, NA),

Sale.Price=replace(Sale.Price, Sale.Price < 1, NA)) -> ames6\_clean #removed outliers and created a new table as ames6\_clean

boxplot(ames6\_clean$Sale.Price)

#some misllenacious sorting

ID=na.omit(ames6\_clean$ID)

Sale.Price = na.omit(ames6\_clean$Sale.Price)

sum(is.na(ames5\_clean$Sale.Price))

#Testing coorelation/hypothesis b/w different variables

cor(ames2$Year.Built,ames2$Sale.Price) #significant but not as expected

cor(ames5\_clean$House.Style,ames5\_clean$Sale.Price)

cor(ames2\_clean$Overall.Cond,ames2\_clean$Sale.Price) #its in minus which is unexpected

cor(ames5\_clean$Gr.Liv.Area,ames5\_clean$Sale.Price) #a good correlation

cor(ames5\_clean$Full.Bath,ames5\_clean$Sale.Price) #average relation

#visualizations

library(ggplot2)

install.packages("ggnewscale")

library("ggnewscale")

#sale price and Year Built

ggplot() +

geom\_point(data = ames6\_clean, aes(x = Year.Built, y = Sale.Price , color = Year.Built)) +

scale\_color\_gradient(low = "green", high = "blue") +

new\_scale\_color() +

scale\_color\_gradient(low = "yellow", high = "red") +

labs(title = "Distribution of house prices", x = "Year Built", y = "Sale Price") +

theme\_minimal() +

geom\_smooth() + geom\_smooth(method = "lm")

#sale price and house style

ggplot() +

geom\_point(data = ames6\_clean, aes(x = House.Style, y = Sale.Price , color = House.Style)) +

scale\_color\_gradient(low = "yellow", high = "blue") +

new\_scale\_color() +

scale\_color\_gradient(low = "yellow", high = "red") +

labs(title = "Distribution of house prices as per house style", x = "House Style", y = "Sale Price") +

theme\_minimal() +

geom\_smooth() + geom\_smooth(method = "lm")

#sale price and Overall Condition

ggplot() +

geom\_point(data = ames6\_clean, aes(x = Overall.Cond, y = Sale.Price , color = Overall.Cond)) +

scale\_color\_gradient(low = "red", high = "blue") +

new\_scale\_color() +

scale\_color\_gradient(low = "yellow", high = "red") +

labs(title = "Distribution of house prices As per overall condition", x = "Overall Condition", y = "Sale Price") +

theme\_minimal() +

geom\_smooth() + geom\_smooth(method = "lm")

#sale price and Greater Living area

ggplot() +

geom\_point(data = ames6\_clean, aes(x = Gr.Liv.Area, y = Sale.Price , color = Gr.Liv.Area)) +

scale\_color\_gradient(low = "deeppink", high = "blue") +

new\_scale\_color() +

scale\_color\_gradient(low = "yellow", high = "red") +

labs(title = "Distribution of house prices As per Greater Living Area", x = "Greater Living Area", y = "Sale Price") +

theme\_minimal() +

geom\_smooth() + geom\_smooth(method = "lm")

#sale price and Full bathroom

ggplot() +

geom\_point(data = ames6\_clean, aes(x = Full.Bath, y = Sale.Price , color = Full.Bath)) +

scale\_color\_gradient(low = "red", high = "blue") +

new\_scale\_color() +

scale\_color\_gradient(low = "yellow", high = "red") +

labs(title = "Distribution of house prices As per Full Bathroom", x = "Full Bathroom", y = "Sale Price") +

theme\_minimal() +

geom\_smooth() + geom\_smooth(method = "lm")

#### split the data

library(caret)

#set seed

set.seed(40386815)

#create index

index <- createDataPartition(ames5\_clean$Sale.Price, list = FALSE, p = 0.8, times = 1)

train <- ames5\_clean[index,]

test <- ames5\_clean[-index,]

# Regression MODEL 1, model on train data

formula <- Sale.Price ~ Year.Built + House.Style

model1 <- lm(formula, train) #first regression model with name model1

summary(model1)

#Regression MODEL 2, model on train data

formula <- Sale.Price ~ Year.Built + Full.Bath

model2 <- lm(formula, train)

summary(model2)

#Regression MODEL 3, model on train data

formula <- Sale.Price ~ Gr.Liv.Area + House.Style

model3 <- lm(formula, train)

summary(model3)

#Regression MODEL 4, model on train data

formula <- Sale.Price ~ House.Style + Full.Bath

model4 <- lm(formula, train)

summary(model4)

#Regression MODEL 5, model on train data

formula <- Sale.Price ~ Gr.Liv.Area + Full.Bath

model5 <- lm(formula, train)

summary(model5)

#Final Regression Model

formula <- Sale.Price ~ Gr.Liv.Area + Full.Bath + Year.Built + House.Style + Overall.Cond + Lot.Area + Pool.Area + Condition.1 + Fireplaces + Kitchen.AbvGr + Overall.Qual + Lot.Frontage + Exter.Qual + Bedroom.AbvGr

final\_model <- lm(formula, train)

summary(final\_model)

## check accuracy on the test data (20 % data that we didn't use in building the model)

# Accuracy of Model 1

predictions <- predict(model1, test)

postResample(predictions, test$Sale.Price)

# Accuracy of Model 2

predictions <- predict(model2, test)

postResample(predictions, test$Sale.Price)

# Accuracy of Model 3

predictions <- predict(model3, test)

postResample(predictions, test$Sale.Price)

# Accuracy of Model 4

predictions <- predict(model4, test)

postResample(predictions, test$Sale.Price)

# Accuracy of Model 5

predictions <- predict(model5, test)

postResample(predictions, test$Sale.Price)

# Accuracy of final model

predictions <- predict(final\_model, test)

postResample(predictions, test$Sale.Price)

##with this we have prediction price and actual price side by side

test$prediction <- predictions

##check the assumptions

library(caret)

library(rms)

vif(final\_model)

#To check outliers through cooks distance > 1

cooks <- cooks.distance(final\_model)

sum(cooks >1)

#plotting of final model to check outliers

plot(final\_model)

#To check residuals

train$residuals <- resid(final\_model)

train$predcictions <- fitted(final\_model)

#to Check DW test

install.packages("lmtest")

library(lmtest)

dwtest(final\_model) # A result of 1.73 which lies between 1.5 to 2.5.