~~~
---
title: "House Dataset Analysis"
author: "Gaurav Kumar"
date:
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

Exploratory Data Analysis:
Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.
```{r}
Houses<-read.csv("E:\\Predictive Analytics\\Housing.csv",header=TRUE)
summary(Houses$Price)
boxplot(Houses$Price)
hist(Houses$Price, freq=TRUE, xlab="Houses' Price", breaks="FD", main="Histogram of Houses'price")
```
#Considering Garage,Bed,Bath and School as factors because these variables have less number of levels.
#From the above plots and summary statistics :
+The maximum price of the house is 450.
+The minimum price of the house is 155.5.
+25% of House prices are between 155.5 and 242.8.(1st Quartile).
+75% of HOuse prices are between 155.5 and 336.8.(Third Quartile).#
#

#-----Convert categorical variables to factors
factor(Houses$Bath)
```
```

```
factor(Houses$Garage)
factor(Houses$School)
factor(Houses$Bed)

#------Using the summary and a boxplot
#describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.

boxplot(Houses$Price~Houses$School)
boxplot(Houses$Price~Houses$Bath)
boxplot(Houses$Price~Houses$Garage)
boxplot(Houses$Price~Houses$Bed)

by(Houses$Price,Houses$Bed,summary)
by(Houses$Price,Houses$Garage,summary)
by(Houses$Price,Houses$School,summary)
by(Houses$Price,Houses$Bath,summary)

#-----Using the summary, correlation and the pairs plots discuss
#the relationship between the response sales price and each of the numeric predictor variables.
data<-data.frame(Houses$Price,Houses$Size,Houses$Bed,Houses$Garage,Houses$Bath,Houses$Lot)
pairs(data)
cor(data,use="all.obs")
summary(data)

names(Houses)

#------Considering Garage,bed ,bath and school as categorical variable.

#----Regression model
```

#Fit a model using size, lot, bath, bed, year, garage and school as the predictor variables.

#Equation

#Price<-b_0 + b1*Size + b2*Bath + b3*Bed + b4*year + b5*Garage + b6*school

model<-lm(Price ~ Size + Lot + factor(Bath) + factor(Bed) + Year + factor(Garage) + School,data=Houses)

summary(model)

#---Estimate for the intercept term b0.

#The value of b0 is -884.3531

#----Estimate of Î²size the parameter associated with ï¬oor size (Size).

59.4503

#---Estimate of Î²Bath1.1 the parameter associated with one and a half bathrooms.

135.8983

#Discuss and interpret the eï¬ect the predictor variable bed on the expected value of the house prices.

#The values are significant at 1% level of significance and

model2<-lm()

#with increase in value of bed the price of house will decrease.

#List the predictor variables that are signiï¬cantly contributing to the expected value of the house prices

#Size , Lot ,Bath1.1 and Bed

#Since there P value is less than the level of significance.

#For each predictor variable what is the value that will lead to the

#largest expected value of the house prices.

#This is not a good model of the expected value of the house price as the predicted value differs a lot from the

#actual value.

#Adjusted R squared value -

#The Adjusted R squared value is 0.51 which says that 51 % of the variation in Y can be expalined by this model.

# 11.Interpret the F-statistic in the output in the summary of the regression model. Hint: State the

#hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

#Hypothesis being tested -

# All the coefficients are Zero

#F-statistic: 4.942 on 20 and 55 DF, p-value: 1.265e-06

#This says that we can reject the null hypothesis , as the p value is less than the level of significance .

#This means at least one of it is not equal to 0 .

#Hypothesis says that :

#HO: b_0=b1=b2=b3=b4=b5=b6=0

#HA:

#and We can say that there is 4.94% probability that explaination in Y is expalined by atleat o

#one of the expalinatory variable.

#Test Statistics used as -

#MSE ERROR/MSE RESIDUAL 42.13,Reject HO . as the probability value is less.

```
#ANOVA with the sequential sum of squares (or ANOVA type 1 tests)

anova(model)

#11078 of the variation in Y is

#explained by the variable Size given that no other variables are in the model


#65041 of the variation in Y is explained by the variable Lot

#given that size is in the model.


##36824 of the variation in Y is explained by the variable Bath

#given that size and Lot is in the model.


##25502 of the variation in Y is explained by the variable Bed

#given that size , Lot and Bath are in the model.


##16101 of the variation in Y is explained by the variable Garage

#given that size , Lot , Bath and Bed are in the model.


#70112 of the variation in Y is explained by the variable School

#given that size , Lot , Bath,Bed  are in the model.

#Year value doesn't make any significance as the p value is greater than the level of significance 0.05.

#97599 of the variation of Y is not explained by  Size,Lot,Bath,Bed,Garage and School.


#Compute a type 2 anova table

library(car)

Anova(model)

# This also says that Year variable is insignificant and can be removed.
```

```
#Diagnostics------------------

#Added variable plot -----------------------


library(car)


avPlots(red)





crPlots(model)


# The relationship between Price and Year seems to be non linear as the pink line and the dashed line diifer

# dramatically.We will remove the predictor variable that shows that shows the non linearity to make the model

# linear.


dwt(model)

#The Durbin Watson test statistic is 1.614157

#and the p-value is 0.052 so the hypothesis of no autocorrelation is not rejected and the observations can be classed as independent




#To check the multicollinearity and Variation inflation factor.

install.packages("corrplot")

library(corrplot)

M <- cor(Houses[,2:5])

corrplot.mixed(M)
```

```
#The correlation between size and other numerical variable is mild.Only Year and Garage show correlation

#greater than 0.50

#Indicating that it is unlikely we will have a multicollinearity problem with a regression including these two predictor variables.

#will remove the variable that shows the collinearity , here we can remove Year.


vif(model)

#Variation inflation factor_check this.


#4. Check the zero conditional mean and homoscedasticity assumption by interpreting the

#studentized residuals vrs ï¬tted values plots and the studentized residuals vrs predictor variable plots.


plot(fitted(model),rstudent(model))

abline(h=0)

plot(Houses$Size,rstudent(model))

abline(h=0)

plot(Houses$Lot,rstudent(model))

abline(h=0)


#This is to check what the residuals say about the normality distribution .

r = rstudent(model)

r

par(mfrow=c(2,1))

boxplot(r)

hist(r,freq=FALSE)
```

```
lines(density(r, lwd=2, col="blue"))

qqnorm(r)

qqline(r)




#Leverage ,Influence and Outliers

#What is a leverage point? What eï¬ect would a leverage point have on the regression model? Use the leverage values and the leverage plots to see if there is any leverage points.

lev = hat(model.matrix(model))

plot(lev)

Houses[lev >0.9,]

leveragePlots(model)

outlierTest(model)

Houses[44,]

#The observation 44 should be removed from the dataset.


ols_plot_cooksd_bar(model)

ols_plot_dfbetas(model)


influencePlot(model)


#--------------Expected Value,CI and PI---------------------#

library(ggplot2)

plot(model)

New_Price<-Houses[,-1]

New_Price
```

~~~
predict(model,New_Price)
pred.int<-predict(model,interval="prediction")
mydata<-cbind(Houses,pred.int)
predict(model)
p<-ggplot(mydata, aes(Size,Price)) +
  geom_point() +
  stat_smooth(method = lm)+geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y = upr), color = "red", linetype = "dashed")
p



```
~~~