getwd()

setwd("C:/Users/MyThinkpad/Desktop/data science NCI/advanced statistics")

house <- read.csv("house.csv", sep=",", header=T)

#shows house price sales based on various different factors.

#I would like to find out if sale prices of homes increase due to these factors

print(house)

attach(house)

#lets do a quick anova test to see if the various different factors have an impact on house prices

house$Bedrooms = as.factor(house$Bedrooms)

house$Baths = as.factor(house$Baths)

house$SquareFeet = as.factor(house$SquareFeet)

house$Agehouse = as.factor(house$Agehouse)

typeof(house$SalePrice)

#lets now first change all values to numeric

house$Agehouse <- as.numeric(house$Agehouse)

house$SalePrice <- as.numeric(house$SalePrice)

house$SquareFeet <- as.numeric(house$SquareFeet)

house$Bedrooms <- as.numeric(house$Bedrooms)

house$Baths <- as.numeric(house$Baths)

cor(house$SalePrice, house$Agehouse)

install.packages("dplyr")

library("dplyr")

install.packages("plyr") #alternative installation of the %>%

library(plyr)

#lets create a subset just to make it easier to show which data we will use exactly for the multi-linear regression

sub\_house <- house %>% select(SalePrice, Agehouse, SquareFeet, Bedrooms, Baths)

print(sub\_house)

#lets now find the correlations!

cor(sub\_house) # simple correlation matrix. #we can see that the sale price of a house is closest positively correlated to number of baths in a house!

#lets now see this in action!

install.packages('corrplot')

library('corrplot') #package corrplot # found from this link. https://stackoverflow.com/questions/10680658/how-can-i-create-a-correlation-matrix-in-r

help("corrplot-package")

M <- cor(sub\_house)

corrplot(M, method = "number") #plot matrix

#this is the correlation matrix

#lets now apply the anova test to it!

house\_anova = aov(sub\_house$SalePrice ~ sub\_house$Agehouse + sub\_house$SquareFeet + sub\_house$Bedrooms + sub\_house$Baths)

summary(house\_anova)

#the interaction effect of each factor is huge.

#for the interaction rate of number of bedrooms on the price of house, there is 0.884% chance that we will arrive at these results by random chance.

boxplot(sub\_house$SalePrice ~ sub\_house$Bedrooms)

hist(sub\_house$SalePrice ~ sub\_house$Bedrooms)

# by appluying anova, we can see that all factors have a storng impact on house sale price. And that it quite difficult to predict.

pairs(sub\_house)

Model1 <- lm(sub\_house$SalePrice ~ sub\_house$Agehouse + sub\_house$SquareFeet + sub\_house$Bedrooms + sub\_house$Baths, data=sub\_house)

Model1

summary(Model1)

plot(sub\_house$SalePrice ~ sub\_house$Agehouse + sub\_house$SquareFeet + sub\_house$Bedrooms + sub\_house$Baths, data=sub\_house, names=c("Sale Price of House", "Age of House","Square Feet","No. Bedrooms", "No. Bathrooms"))

abline(Model1)

#lets do a prediction! Lets say the age of the house is 30, 300 sqft, 4 bedrooms, 3 bathrooms.

x = 15

a = 223

b = 5

c = 3

Y = -966.7\*x + 257.2\*a + 1303.1\*b + 8240.4\*c

print(Y) #74091.8 USD

#-966.73 is the slope for age of house. There is a negative correlation between age and house. or for every 1 year the age of the house, the value of the home drops by 966.73 usd.

#for every 1 increase in square meters, there is a 257.16 usd increase in the price of the home.

Lets

Modelage <- lm(sub\_house$SalePrice ~ sub\_house$Agehouse)

Modelage

summary(Modelage)

plot(sub\_house$SalePrice ~ sub\_house$Agehouse) #this is the exepcted price of home and increasing age of home.

abline(Modelage)

#lets find out more about these relationships.

summary(Model1) # 0.6581 approximately, 65% of variation in house price, can be explained by the variables.

confint(Model1, conf.level = 0.95) # 95 sure that our real slope for x is these results.

plot(Model1)

sd(house$Agehouse) #32.43289

mean(house$Agehouse) #60.88388

hist(house$Agehouse)

print(Y) #74091.8 USD

sdpricehome <- sd(house$SalePrice) #79291.72

meanprice <- mean(house$SalePrice) #135172

# z score

(Y - meanprice)/sdpricehome

0.7703225\*sdpricehome