**Discovering Statistics Using R – Tasks**

Akila Selvaraj

Masters of Data Science,

Bellevue University

DSC 520 T301

January 08,2022

 # Assignment: ASSIGNMENT 4

# Name: Selvaraj, Akila

# Date: 2022-01-07

install.packages('rlang')

install.packages('tidyr')

install.packages("ggplot2")

install.packages("xlsx")

install.packages("xlsx2")

install.packages("plyr")

install.packages("reshape2")

library(xlsx)

library(xlsx2)

library(readxl)

library(plyr)

library(dplyr)

library(reshape2)

getwd()

setwd("G:/Users/a162940/Akila/Work/R/Projects/dsc520-master")

**#Read the excel into dataframe**

Housing\_excel <- read\_excel("data/week-7-housing.xlsx")

Housing\_df <- data.frame(Housing\_excel)

**#Use the apply function on a variable in your dataset**

sum\_of\_sales <- apply(Housing\_df[c('Sale.Price')], 2, sum)

sum\_of\_sales

max\_of\_all\_fields <- apply(Housing\_df, 2, max)

max\_of\_all\_fields

max\_salePrice <- apply(Housing\_df[c('Sale.Price')], 2, max)

max\_salePrice

**#Use the aggregate function on a variable in your dataset**

**#Sum of Sale Price by Sale Date**

aggregate(Housing\_df$Sale.Price, by=Housing\_df["Sale.Date"], FUN=sum)

**#Creating new variable Sale Year**

Housing\_df$Saleyear <- as.integer(format(Housing\_df$Sale.Date, "%Y"))

**#Maximum Sale Price by Sale Year**

aggregate(Housing\_df$Sale.Price, by=Housing\_df["Saleyear"], FUN=max)

**#Sum of sale price by year**

aggregate(Housing\_df$Sale.Price, by=Housing\_df["Saleyear"], FUN=sum)

**#Use the plyr function on a variable in your dataset - more specifically, I want to see you split some data, perform a modification to the data, and then bring it back together**

**#Sum of Sale Price using transform**

ddply(Housing\_df,.(Saleyear), transform, sum.SalePrice = sum(Sale.Price))

**#Mean of Sale Price by Year**

ddply(Housing\_df,.(Saleyear), summarise, mean.SalePrice = mean(Sale.Price))

Housing\_df %>% group\_by(Saleyear) %>% summarize(mean\_Price = mean(Sale.Price))

**#Check distributions of the data**

ggplot(Housing\_df, aes(Saleyear,Sale.Price)) + geom\_point()

ggplot(data=subset(Housing\_df,Saleyear='2006'), aes(Saleyear,Sale.Price)) + geom\_point()

**#Identify if there are any outliers**

Yes, sale.Price is very less for some of the rows which is deviating from the distribution.

For example, sale price less than 1000 like 998,873,698 are outliners.

**#Create at least 2 new variables**

Housing\_df %>% mutate(PriceRange = case\_when(Sale.Price <= 100000 ~ 'low',

Sale.Price <= 1000000 ~ 'med',

Sale.Price > 1000000 ~ 'high'))

Housing\_df %>% mutate(HouseSize = case\_when(sq\_ft\_lot < 1000 ~ 'Small House',

Sale.Price <= 5000 ~ 'Medium House',

Sale.Price > 5000 ~ 'Big House'))