sales\_prediction.R

user

2023-06-16

#The purpose of the studying is to investigate the association $ effects and prediction of Revenue From Sales\_data with Linear Regression

# Loading the data in two way;by copying the path or use file choosing format  
 #Reading the path by copying the path   
pd=read.csv('F:\\Sales\_predictions\\sales\_data.csv')  
 #Reading the path using file.choose  
pd=read.csv(file.choose())  
#Checking the first 1,2,3,4,5,,,,, but default is first five  
head(pd)

## Date Day Month Year Customer\_Age Age\_Group Customer\_Gender  
## 1 2013-11-26 26 November 2013 19 Youth (<25) M  
## 2 2015-11-26 26 November 2015 19 Youth (<25) M  
## 3 2014-03-23 23 March 2014 49 Adults (35-64) M  
## 4 2016-03-23 23 March 2016 49 Adults (35-64) M  
## 5 2014-05-15 15 May 2014 47 Adults (35-64) F  
## 6 2016-05-15 15 May 2016 47 Adults (35-64) F  
## Country State Product\_Category Sub\_Category Product  
## 1 Canada British Columbia Accessories Bike Racks Hitch Rack - 4-Bike  
## 2 Canada British Columbia Accessories Bike Racks Hitch Rack - 4-Bike  
## 3 Australia New South Wales Accessories Bike Racks Hitch Rack - 4-Bike  
## 4 Australia New South Wales Accessories Bike Racks Hitch Rack - 4-Bike  
## 5 Australia New South Wales Accessories Bike Racks Hitch Rack - 4-Bike  
## 6 Australia New South Wales Accessories Bike Racks Hitch Rack - 4-Bike  
## Order\_Quantity Unit\_Cost Unit\_Price Profit Cost Revenue  
## 1 8 45 120 590 360 950  
## 2 8 45 120 590 360 950  
## 3 23 45 120 1366 1035 2401  
## 4 20 45 120 1188 900 2088  
## 5 4 45 120 238 180 418  
## 6 5 45 120 297 225 522

#Checking the last1,2,3,4,5,,,,, but default is last five  
tail(pd)

## Date Day Month Year Customer\_Age Age\_Group Customer\_Gender  
## 113031 2014-04-12 12 April 2014 41 Adults (35-64) M  
## 113032 2016-04-12 12 April 2016 41 Adults (35-64) M  
## 113033 2014-04-02 2 April 2014 18 Youth (<25) M  
## 113034 2016-04-02 2 April 2016 18 Youth (<25) M  
## 113035 2014-03-04 4 March 2014 37 Adults (35-64) F  
## 113036 2016-03-04 4 March 2016 37 Adults (35-64) F  
## Country State Product\_Category Sub\_Category  
## 113031 United Kingdom England Clothing Vests  
## 113032 United Kingdom England Clothing Vests  
## 113033 Australia Queensland Clothing Vests  
## 113034 Australia Queensland Clothing Vests  
## 113035 France Seine (Paris) Clothing Vests  
## 113036 France Seine (Paris) Clothing Vests  
## Product Order\_Quantity Unit\_Cost Unit\_Price Profit Cost Revenue  
## 113031 Classic Vest, S 6 24 64 225 144 369  
## 113032 Classic Vest, S 3 24 64 112 72 184  
## 113033 Classic Vest, M 22 24 64 655 528 1183  
## 113034 Classic Vest, M 22 24 64 655 528 1183  
## 113035 Classic Vest, L 24 24 64 684 576 1260  
## 113036 Classic Vest, L 23 24 64 655 552 1207

#Checking the variables names  
names(pd)

## [1] "Date" "Day" "Month" "Year"   
## [5] "Customer\_Age" "Age\_Group" "Customer\_Gender" "Country"   
## [9] "State" "Product\_Category" "Sub\_Category" "Product"   
## [13] "Order\_Quantity" "Unit\_Cost" "Unit\_Price" "Profit"   
## [17] "Cost" "Revenue"

#Check the variable type  
str(pd)

## 'data.frame': 113036 obs. of 18 variables:  
## $ Date : chr "2013-11-26" "2015-11-26" "2014-03-23" "2016-03-23" ...  
## $ Day : int 26 26 23 23 15 15 22 22 22 22 ...  
## $ Month : chr "November" "November" "March" "March" ...  
## $ Year : int 2013 2015 2014 2016 2014 2016 2014 2016 2014 2016 ...  
## $ Customer\_Age : int 19 19 49 49 47 47 47 47 35 35 ...  
## $ Age\_Group : chr "Youth (<25)" "Youth (<25)" "Adults (35-64)" "Adults (35-64)" ...  
## $ Customer\_Gender : chr "M" "M" "M" "M" ...  
## $ Country : chr "Canada" "Canada" "Australia" "Australia" ...  
## $ State : chr "British Columbia" "British Columbia" "New South Wales" "New South Wales" ...  
## $ Product\_Category: chr "Accessories" "Accessories" "Accessories" "Accessories" ...  
## $ Sub\_Category : chr "Bike Racks" "Bike Racks" "Bike Racks" "Bike Racks" ...  
## $ Product : chr "Hitch Rack - 4-Bike" "Hitch Rack - 4-Bike" "Hitch Rack - 4-Bike" "Hitch Rack - 4-Bike" ...  
## $ Order\_Quantity : int 8 8 23 20 4 5 4 2 22 21 ...  
## $ Unit\_Cost : int 45 45 45 45 45 45 45 45 45 45 ...  
## $ Unit\_Price : int 120 120 120 120 120 120 120 120 120 120 ...  
## $ Profit : int 590 590 1366 1188 238 297 199 100 1096 1046 ...  
## $ Cost : int 360 360 1035 900 180 225 180 90 990 945 ...  
## $ Revenue : int 950 950 2401 2088 418 522 379 190 2086 1991 ...

#Checking columns  
ncol(pd)

## [1] 18

#checking rows  
nrow(pd)

## [1] 113036

#To determine if there is missing values in between variables  
sum(is.na(pd))# In general

## [1] 0

sum(is.na(pd$Product))

## [1] 0

sum(is.na(pd$Order\_Quantity))

## [1] 0

#We are going to conduct descriptive analysis that is central tendency and dispersion  
summary(pd)

## Date Day Month Year   
## Length:113036 Min. : 1.00 Length:113036 Min. :2011   
## Class :character 1st Qu.: 8.00 Class :character 1st Qu.:2013   
## Mode :character Median :16.00 Mode :character Median :2014   
## Mean :15.67 Mean :2014   
## 3rd Qu.:23.00 3rd Qu.:2016   
## Max. :31.00 Max. :2016   
## Customer\_Age Age\_Group Customer\_Gender Country   
## Min. :17.00 Length:113036 Length:113036 Length:113036   
## 1st Qu.:28.00 Class :character Class :character Class :character   
## Median :35.00 Mode :character Mode :character Mode :character   
## Mean :35.92   
## 3rd Qu.:43.00   
## Max. :87.00   
## State Product\_Category Sub\_Category Product   
## Length:113036 Length:113036 Length:113036 Length:113036   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Order\_Quantity Unit\_Cost Unit\_Price Profit   
## Min. : 1.0 Min. : 1.0 Min. : 2.0 Min. : -30.0   
## 1st Qu.: 2.0 1st Qu.: 2.0 1st Qu.: 5.0 1st Qu.: 29.0   
## Median :10.0 Median : 9.0 Median : 24.0 Median : 101.0   
## Mean :11.9 Mean : 267.3 Mean : 452.9 Mean : 285.1   
## 3rd Qu.:20.0 3rd Qu.: 42.0 3rd Qu.: 70.0 3rd Qu.: 358.0   
## Max. :32.0 Max. :2171.0 Max. :3578.0 Max. :15096.0   
## Cost Revenue   
## Min. : 1.0 Min. : 2.0   
## 1st Qu.: 28.0 1st Qu.: 63.0   
## Median : 108.0 Median : 223.0   
## Mean : 469.3 Mean : 754.4   
## 3rd Qu.: 432.0 3rd Qu.: 800.0   
## Max. :42978.0 Max. :58074.0

# we will need to select relevant variable for our analysis by loading tidyverse and dplyr  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.3

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## Warning: package 'lubridate' was built under R version 4.1.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --  
## v dplyr 1.1.0 v readr 2.1.4  
## v forcats 1.0.0 v stringr 1.5.0  
## v ggplot2 3.4.1 v tibble 3.1.8  
## v lubridate 1.9.2 v tidyr 1.3.0  
## v purrr 1.0.1   
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

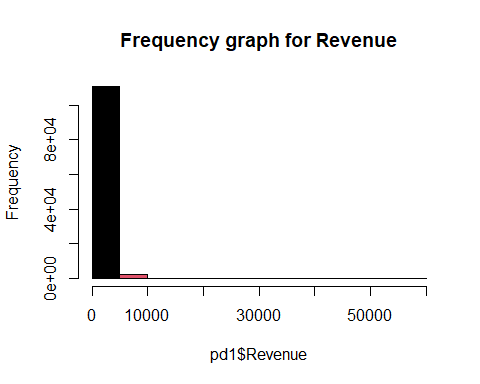
library(dplyr)  
#selecting the variable required leaving others  
pd1=select(pd,Order\_Quantity,Unit\_Cost,Unit\_Price,Cost,Revenue)  
#Checking the associations  
cor(pd1)

## Order\_Quantity Unit\_Cost Unit\_Price Cost Revenue  
## Order\_Quantity 1.0000000 -0.5158350 -0.5159246 -0.3403816 -0.3128950  
## Unit\_Cost -0.5158350 1.0000000 0.9978936 0.8298690 0.8178650  
## Unit\_Price -0.5159246 0.9978936 1.0000000 0.8263011 0.8185218  
## Cost -0.3403816 0.8298690 0.8263011 1.0000000 0.9887584  
## Revenue -0.3128950 0.8178650 0.8185218 0.9887584 1.0000000

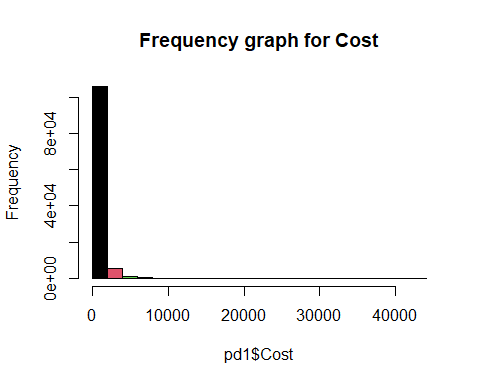
# checking again the summary statistics  
summary(pd1)

## Order\_Quantity Unit\_Cost Unit\_Price Cost   
## Min. : 1.0 Min. : 1.0 Min. : 2.0 Min. : 1.0   
## 1st Qu.: 2.0 1st Qu.: 2.0 1st Qu.: 5.0 1st Qu.: 28.0   
## Median :10.0 Median : 9.0 Median : 24.0 Median : 108.0   
## Mean :11.9 Mean : 267.3 Mean : 452.9 Mean : 469.3   
## 3rd Qu.:20.0 3rd Qu.: 42.0 3rd Qu.: 70.0 3rd Qu.: 432.0   
## Max. :32.0 Max. :2171.0 Max. :3578.0 Max. :42978.0   
## Revenue   
## Min. : 2.0   
## 1st Qu.: 63.0   
## Median : 223.0   
## Mean : 754.4   
## 3rd Qu.: 800.0   
## Max. :58074.0

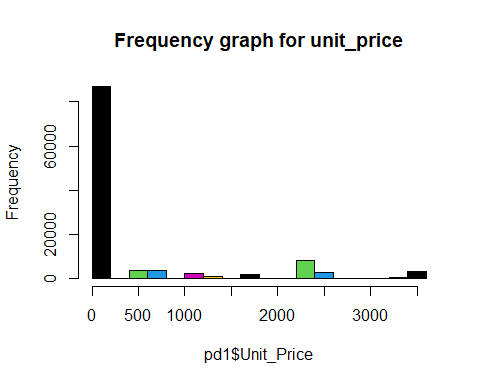
# Conducting EDA through visualization  
hist(pd1$Revenue,col = c(1,2),main = 'Frequency graph for Revenue')



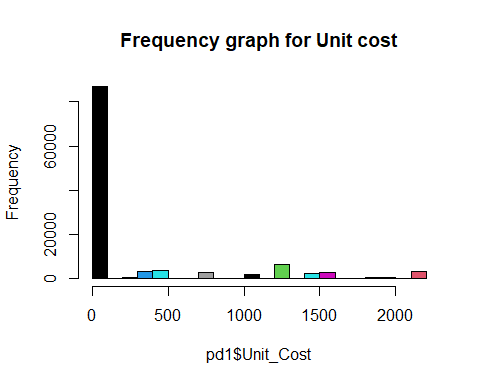
hist(pd1$Cost,col = c(1,2,3,4,6),main = 'Frequency graph for Cost')



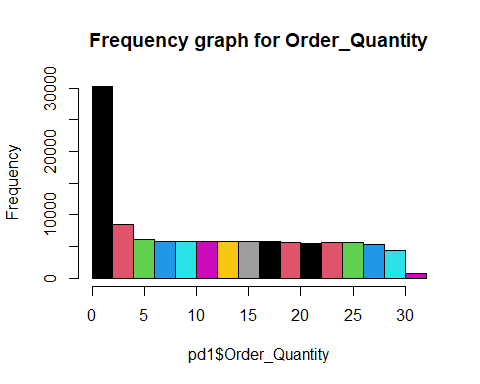
hist(pd1$Unit\_Price,col = c(1,2,3,4,5,6,7,8,9),main = 'Frequency graph for unit\_price')



hist(pd1$Unit\_Cost,col = c(1,2,3,4,5,6,7,8,9,10),main = 'Frequency graph for Unit cost')



hist(pd1$Order\_Quantity,col = c(1,2,3,4,5,6,7,8,9,10),main = 'Frequency graph for Order\_Quantity')



# Here we are going to utilize regression analysis in fitting the model  
fit=lm(pd1$Revenue~pd1$Order\_Quantity+pd1$Unit\_Cost+pd1$Unit\_Price+pd1$Cost)  
#Viewing the intercepts  
fit

##   
## Call:  
## lm(formula = pd1$Revenue ~ pd1$Order\_Quantity + pd1$Unit\_Cost +   
## pd1$Unit\_Price + pd1$Cost)  
##   
## Coefficients:  
## (Intercept) pd1$Order\_Quantity pd1$Unit\_Cost pd1$Unit\_Price   
## -0.7649 4.3089 -2.3602 1.4248   
## pd1$Cost   
## 1.4689

#To check all regressions coefficient  
summary(fit)

##   
## Call:  
## lm(formula = pd1$Revenue ~ pd1$Order\_Quantity + pd1$Unit\_Cost +   
## pd1$Unit\_Price + pd1$Cost)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5079.8 -60.6 -13.6 59.5 1732.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.764901 1.067526 -0.717 0.474   
## pd1$Order\_Quantity 4.308885 0.063873 67.461 <2e-16 \*\*\*  
## pd1$Unit\_Cost -2.360241 0.014588 -161.794 <2e-16 \*\*\*  
## pd1$Unit\_Price 1.424808 0.008609 165.495 <2e-16 \*\*\*  
## pd1$Cost 1.468908 0.001061 1384.655 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 172.9 on 113031 degrees of freedom  
## Multiple R-squared: 0.9826, Adjusted R-squared: 0.9826   
## F-statistic: 1.592e+06 on 4 and 113031 DF, p-value: < 2.2e-16

# Running Analysis of variance  
anova(fit)

## Analysis of Variance Table  
##   
## Response: pd1$Revenue  
## Df Sum Sq Mean Sq F value Pr(>F)   
## pd1$Order\_Quantity 1 1.8965e+10 1.8965e+10 634327 < 2.2e-16 \*\*\*  
## pd1$Unit\_Cost 1 1.1374e+11 1.1374e+11 3804440 < 2.2e-16 \*\*\*  
## pd1$Unit\_Price 1 3.0043e+08 3.0043e+08 10049 < 2.2e-16 \*\*\*  
## pd1$Cost 1 5.7322e+10 5.7322e+10 1917270 < 2.2e-16 \*\*\*  
## Residuals 113031 3.3794e+09 2.9898e+04   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1