Milestone 2: Data Summary/Visualization

Kaarthik Sundaramoorthy, Sahil Shah and Vidhi Shah

6/23/2020

The dataset used in the project is based on *“online shoppers purchasing intention”* available on UCI Machine Learning dataset. The dataset consist of the vectors

#### Importing Libraries

library(ggplot2)  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------ tidyverse 1.3.0 --

## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0  
## v purrr 0.3.3

## -- Conflicts --------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(gmodels)  
library(dplyr)  
library(ggmosaic)  
library(corrplot)

## corrplot 0.84 loaded

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(rpart)  
library(rpart.plot)  
library(cluster)  
library(fpc)  
library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

#### Importing the Dataset

dataset <- read.csv("online\_shoppers\_intention.csv", header = TRUE)  
attach(dataset)

Checking the number of columns and rows of the dataset.

ncol(dataset)

## [1] 18

nrow(dataset)

## [1] 12330

Looking at the dataset data structure.

summary(dataset)

## Administrative Administrative\_Duration Informational   
## Min. : 0.000 Min. : 0.00 Min. : 0.0000   
## 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.0000   
## Median : 1.000 Median : 7.50 Median : 0.0000   
## Mean : 2.315 Mean : 80.82 Mean : 0.5036   
## 3rd Qu.: 4.000 3rd Qu.: 93.26 3rd Qu.: 0.0000   
## Max. :27.000 Max. :3398.75 Max. :24.0000   
##   
## Informational\_Duration ProductRelated ProductRelated\_Duration  
## Min. : 0.00 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.: 7.00 1st Qu.: 184.1   
## Median : 0.00 Median : 18.00 Median : 598.9   
## Mean : 34.47 Mean : 31.73 Mean : 1194.8   
## 3rd Qu.: 0.00 3rd Qu.: 38.00 3rd Qu.: 1464.2   
## Max. :2549.38 Max. :705.00 Max. :63973.5   
##   
## BounceRates ExitRates PageValues SpecialDay   
## Min. :0.000000 Min. :0.00000 Min. : 0.000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.01429 1st Qu.: 0.000 1st Qu.:0.00000   
## Median :0.003112 Median :0.02516 Median : 0.000 Median :0.00000   
## Mean :0.022191 Mean :0.04307 Mean : 5.889 Mean :0.06143   
## 3rd Qu.:0.016813 3rd Qu.:0.05000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :0.200000 Max. :0.20000 Max. :361.764 Max. :1.00000   
##   
## Month OperatingSystems Browser Region   
## May :3364 Min. :1.000 Min. : 1.000 Min. :1.000   
## Nov :2998 1st Qu.:2.000 1st Qu.: 2.000 1st Qu.:1.000   
## Mar :1907 Median :2.000 Median : 2.000 Median :3.000   
## Dec :1727 Mean :2.124 Mean : 2.357 Mean :3.147   
## Oct : 549 3rd Qu.:3.000 3rd Qu.: 2.000 3rd Qu.:4.000   
## Sep : 448 Max. :8.000 Max. :13.000 Max. :9.000   
## (Other):1337   
## TrafficType VisitorType Weekend Revenue   
## Min. : 1.00 New\_Visitor : 1694 Mode :logical Mode :logical   
## 1st Qu.: 2.00 Other : 85 FALSE:9462 FALSE:10422   
## Median : 2.00 Returning\_Visitor:10551 TRUE :2868 TRUE :1908   
## Mean : 4.07   
## 3rd Qu.: 4.00   
## Max. :20.00   
##

Taking the look at the **REVENUE** column which is the target column. The datatype of the REVENUE column is Logical which holds the value **TRUE** and **FALSE**

library(gmodels)  
summary(dataset$Revenue)

## Mode FALSE TRUE   
## logical 10422 1908

CrossTable(dataset$Revenue)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 12330   
##   
##   
## | FALSE | TRUE |   
## |-----------|-----------|  
## | 10422 | 1908 |   
## | 0.845 | 0.155 |   
## |-----------|-----------|  
##   
##   
##   
##

Adding the new *Revenue\_binary* column by using Logical Data of Shopper’s Revenue into binary dependent variable that will helpful for potential regression models. The data will be converted with values 0 and 1, i.e. If it is false the value is 0 and if true it will be 1.

dataset <- dataset %>%  
 mutate(Revenue\_binary = ifelse(dataset$Revenue == "TRUE", 1, 0))

Checking the dataset if it has any missing values.

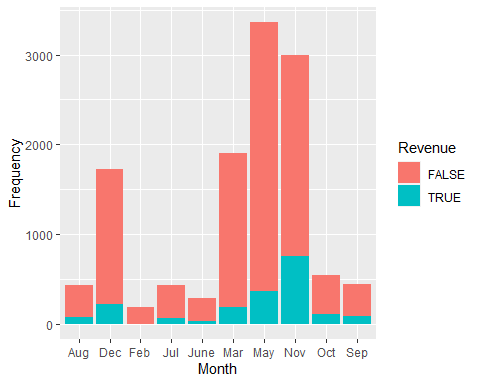
colSums(is.na(dataset))

## Administrative Administrative\_Duration Informational   
## 0 0 0   
## Informational\_Duration ProductRelated ProductRelated\_Duration   
## 0 0 0   
## BounceRates ExitRates PageValues   
## 0 0 0   
## SpecialDay Month OperatingSystems   
## 0 0 0   
## Browser Region TrafficType   
## 0 0 0   
## VisitorType Weekend Revenue   
## 0 0 0   
## Revenue\_binary   
## 0

#### sdjks Visualizations

sjkdkiasasdj

dataset %>%  
 ggplot() +   
 aes(x = Month, Revenue = ..count../nrow(dataset), fill = Revenue) +  
 geom\_bar() +  
 ylab("Frequency")

 ```