Final Project-Team 1

Cole Bailey-Rudy Fasano-Tyler Wolff

12/13/2020

#Section 1: Data Importing and Pre-Processing

```r
#Import libraries
library(ggplot2)
library(knitr)
```

```r
#Import raw data file
data = read.csv("online_shoppers_intention.csv", header =T, sep=",")
head(data,5)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                       0             0                      0
## 4              0                       0             0                      0
## 5              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000        0.20      0.20          0
## 2              2               64.000000        0.00      0.10          0
## 3              1                0.000000        0.20      0.20          0
## 4              2                2.666667        0.05      0.14          0
## 5             10              627.500000        0.02      0.05          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   False   False
## 2 Returning_Visitor   False   False
## 3 Returning_Visitor   False   False
## 4 Returning_Visitor   False   False
## 5 Returning_Visitor    True   False
```

```r
#Changing null values to zero. There were initially 382 rows with one null
#value, and four rows with more than one null value.
data[is.na(data)] <- 0
```

```r
#This is selecting only a select portion of the columns
data1 = data[, c(7,8)]
summary(data1)
```

```
##   BounceRates         ExitRates
##  Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429
##  Median :0.003112   Median :0.02516
##  Mean   :0.022191   Mean   :0.04307
##  3rd Qu.:0.016813   3rd Qu.:0.05000
##  Max.   :0.200000   Max.   :0.20000
```

```
#Assigning columns to variables in order to work with them independently
x = data1$BounceRates
y = data1$ExitRates
```

#Section 2: Data Analysis and Visualization

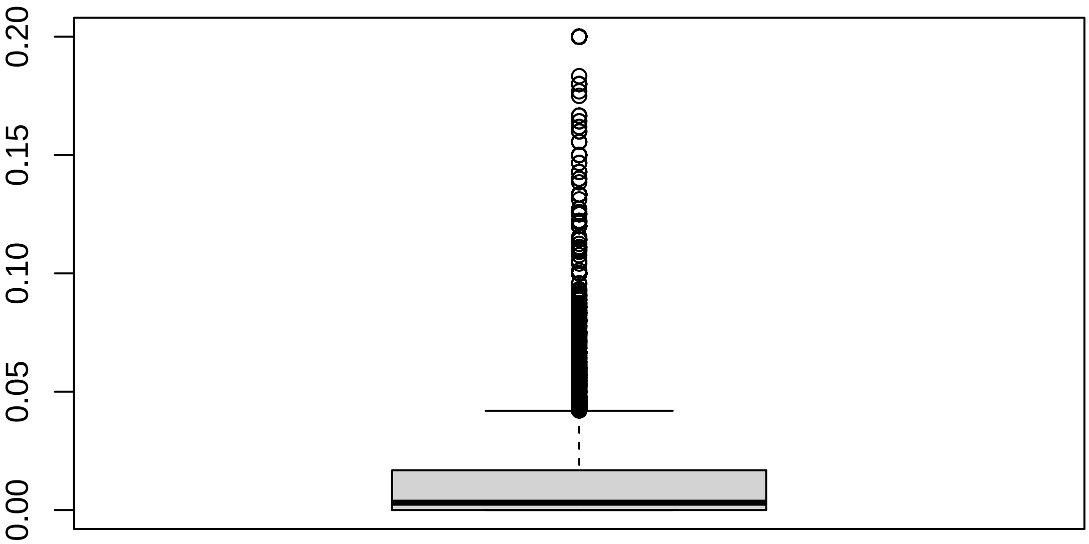
```
#The "str" function identifies categorical, ordinal, and numerical variables
#within data. It also provides the dimensions of the data.

#The raw data file was a comma separated value (CSV file) consisting of 12,330
#Rows and 18 Columns that was imported by the read table function after setting
#the working directory.
str(data)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : chr  "Feb" "Feb" "Feb" "Feb" ...
##  $ OperatingSystems       : num  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : chr  "Returning_Visitor" "Returning_Visitor" "Returning_Visitor" "Returni
##  $ Weekend                : chr  "False" "False" "False" "False" ...
##  $ Revenue                : chr  "False" "False" "False" "False" ...
```

```
#The summary function displays the mean and other measures of centrality
summary(data)
```

```
##  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   : 0.4985
##  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   :  5.846   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
##  Length:12330       Min.   :0.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.102    Mean   : 2.357   Mean   :3.147
##                     3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##                     Max.   :8.000    Max.   :13.000   Max.   :9.000
##   TrafficType    VisitorType          Weekend             Revenue
##  Min.   : 1.00   Length:12330       Length:12330       Length:12330
##  1st Qu.: 2.00   Class :character   Class :character   Class :character
##  Median : 2.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
```

```
#We used a boxplot to visualize potential outliers. Upon further review, we
#decided not to omit the outliers as they reflect a percentage of visitors who
#enter the site then leave without utilizing the page.
BPlotx = boxplot(x)
```



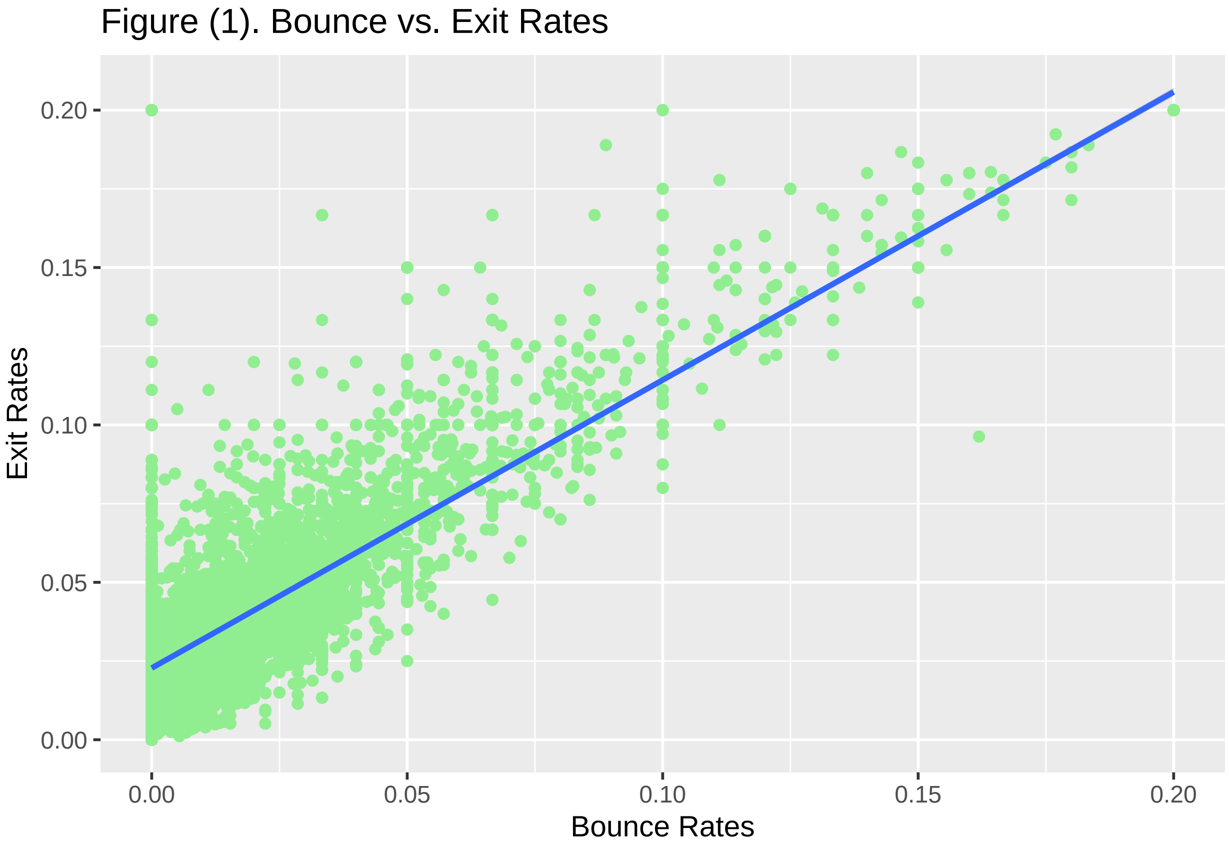
```
#The following is the correlation between the variables
cor.test(x,y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 248.49, df = 12328, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.9100187 0.9158954
## sample estimates:
##       cor
## 0.9130044
```

```
#A visualization with a regression line between bounce rates and exit rates.
#There is a strong positive correlation (0.913) between Bounce and Exit Rates
ggplot(data1, aes(x,y))+geom_point(col ="Light Green")+ggtitle("Figure (1). Bounce vs. Exit Rates")+
  xlab("Bounce Rates")+ylab("Exit Rates") +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



#Section 3: Data Analytics

```
data3 = read.csv("online_shoppers_intention.csv", header =T, sep=",")

data3$Revenue[data3$Revenue == "True"] <-1
data3$Revenue[data3$Revenue == "False"] <-0
data3$Revenue <- as.integer(data3$Revenue)
head(data3, 5)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                       0             0                      0
## 4              0                       0             0                      0
## 5              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000        0.20      0.20          0
## 2              2               64.000000        0.00      0.10          0
## 3              1                0.000000        0.20      0.20          0
## 4              2                2.666667        0.05      0.14          0
## 5             10              627.500000        0.02      0.05          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   False       0
## 2 Returning_Visitor   False       0
## 3 Returning_Visitor   False       0
## 4 Returning_Visitor   False       0
## 5 Returning_Visitor    True       0
```

```
dataRev = data3$Revenue
#Data3 is the original data but transforming "Revenue" to binary integer values.
```

```
#Multiple Linear regression values between "Revenue" generated and "Bounce/Exit
#Rates"

fit_1 <- lm(dataRev ~ x + y , data = data3)
summary(fit_1)
```

```
##
## Call:
## lm(formula = dataRev ~ x + y, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27274 -0.20298 -0.15954  0.00637  1.06038
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.250451   0.005044   49.66   <2e-16 ***
## x            1.720354   0.160366   10.73   <2e-16 ***
## y           -3.108302   0.160009  -19.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3522 on 12327 degrees of freedom
## Multiple R-squared:  0.05173,    Adjusted R-squared:  0.05158
## F-statistic: 336.2 on 2 and 12327 DF,  p-value: < 2.2e-16
```

```
#The adjusted R-Squared value is 0.05158 (5.1%). This indicates there is not a
#significance between revenue and bounce/exit rates.
```

Predictive Modeling

```
#We can use the standard error and other coefficients to create a predictive
#model.
```

```
#Y =  B0 + B1(x) + E Where Y and X variables are the independent and dependent
#variables where the relation is being evaluated, B0 is the model intercept, B1
#represents the model slope, and E is the standard error.
```