

## Milestone 2: Data Summary/Visualization

Sahil Shah

6/23/2020

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

URL : <https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

### Importing Libraries

This are the important libraries that are to be installed for the execution of the file.

```
library(ggplot2)
library(tidyverse)
library(gmodels)
library(dplyr)
library(ggmosaic)
library(corrplot)
library(caret)
library(rpart)
library(rpart.plot)
library(cluster)
library(fpc)
library(data.table)
library(knitr)
library(kableExtra)
```

### Importing the Dataset

The `read.csv()` command is used to import 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.

```
str(dataset)
```

```
## '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 : int 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 : Factor w/ 10 levels "Aug","Dec","Feb",...: 3 3 3 3 3 3 3 3 3 3 ...
## $ OperatingSystems : int 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 : Factor w/ 3 levels "New_Visitor",...: 3 3 3 3 3 3 3 3 3 3 ...
## $ Weekend : logi FALSE FALSE FALSE FALSE TRUE FALSE ...
## $ Revenue : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
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
##
```

The purchasing intention model is designed as a classification problem which measures the purchasers' commitment to finalize purchase intent. Hence we have the session data of the users which has two categories : users who purchased the item and who didn't. The dataset consists of both numerical data and categorical data, and thus the target value is categorical. Table 1 refers to the numerical features and Table 2 refers to the categorical features used in the prediction model respectively. There are a total of 12,330 rows where each row represents session data of one particular user.

```
tab1 <- read.csv("table1.csv", header = TRUE)
kable(tab1) %>%
  kable_styling(full_width = T)
```

i.Feature.Name	Description	Min..value	Max..value	SD
Administrative	Number of pages visited by the visitor about account management	0	27.0	3.322e+00
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages	0	3399.0	1.768e+02
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site	0	24.0	1.270e+00
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages	0	2549.4	1.407e+02
Product related	Number of pages visited by visitor about product related pages	0	705.0	4.448e+01
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages	0	63974.0	1.914e+03
Bounce rates	Average bounce rate value of the pages visited by the visitor	0	0.2	4.849e-02
Exit rate	Average exit rate value of the pages visited by the visitor	0	0.2	4.860e-02
Page value	Average page value of the pages visited by the visitor	0	361.8	1.857e+01
Special day	Closeness of the site visiting time to a special day	0	1.0	1.989e-01

```

tab2 <- read.csv("table2.csv", header = TRUE)
kable(tab2) %>%
  kable_styling(full_width = T)

```

i.Name	Description	Values
OperatingSystems	Operating system of the visitor	8
Browser	Browser of the visitor	13
Region	Geographic region from which the session has been started by the visitor	9
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	20
VisitorType	Visitor type as New Visitor, Returning Visitor, and Other	3
Weekend	Boolean value indicating whether the date of the visit is weekend	2
Month	Month value of the visit date	10
Revenue	Class label indicating whether the visit has been finalized with a transaction	2

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))
```

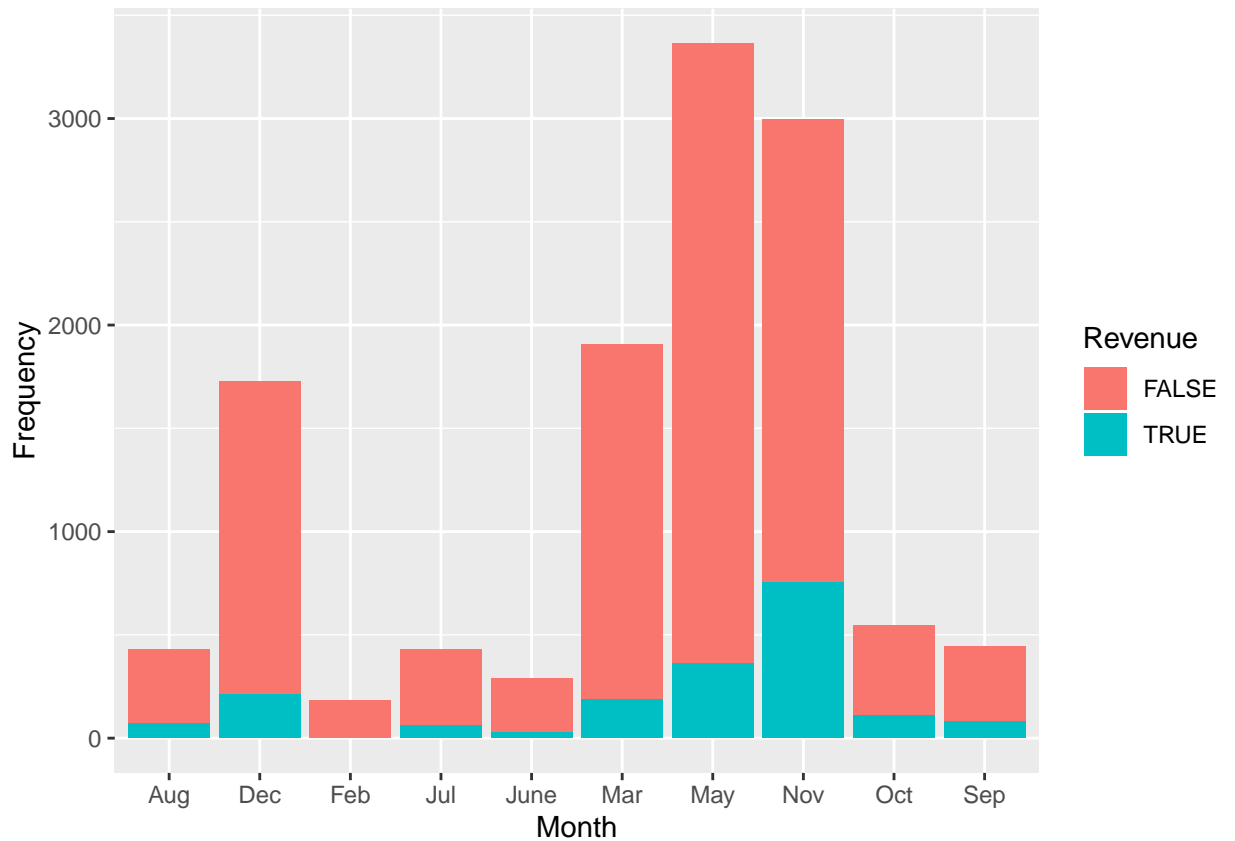
Checking the dataset if it has any missing values.

```
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
```

## Visualizations

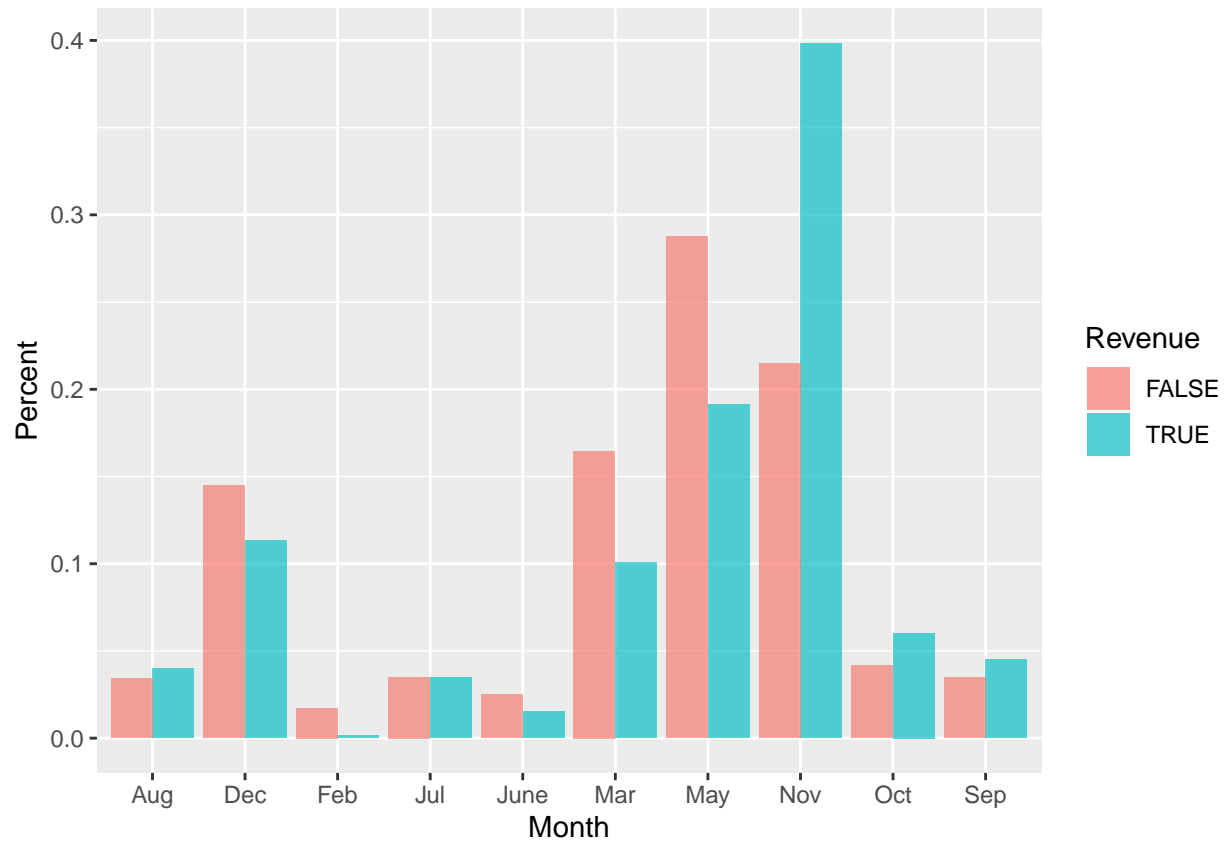
```
dataset %>%
  ggplot() +
  aes(x = Month, Revenue = ..count../nrow(dataset), fill = Revenue) +
  geom_bar() +
  ylab("Frequency")
```



Month

The plot describes the frequency of the revenue generated over the months.

```
table_month = table(dataset$Month, dataset$Revenue)
tab_mon = as.data.frame(prop.table(table_month,2))
colnames(tab_mon) = c("Month", "Revenue", "perc")
ggplot(data = tab_mon, aes(x = Month, y = perc, fill = Revenue)) +
  geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) +
  xlab("Month")+
  ylab("Percent")
```



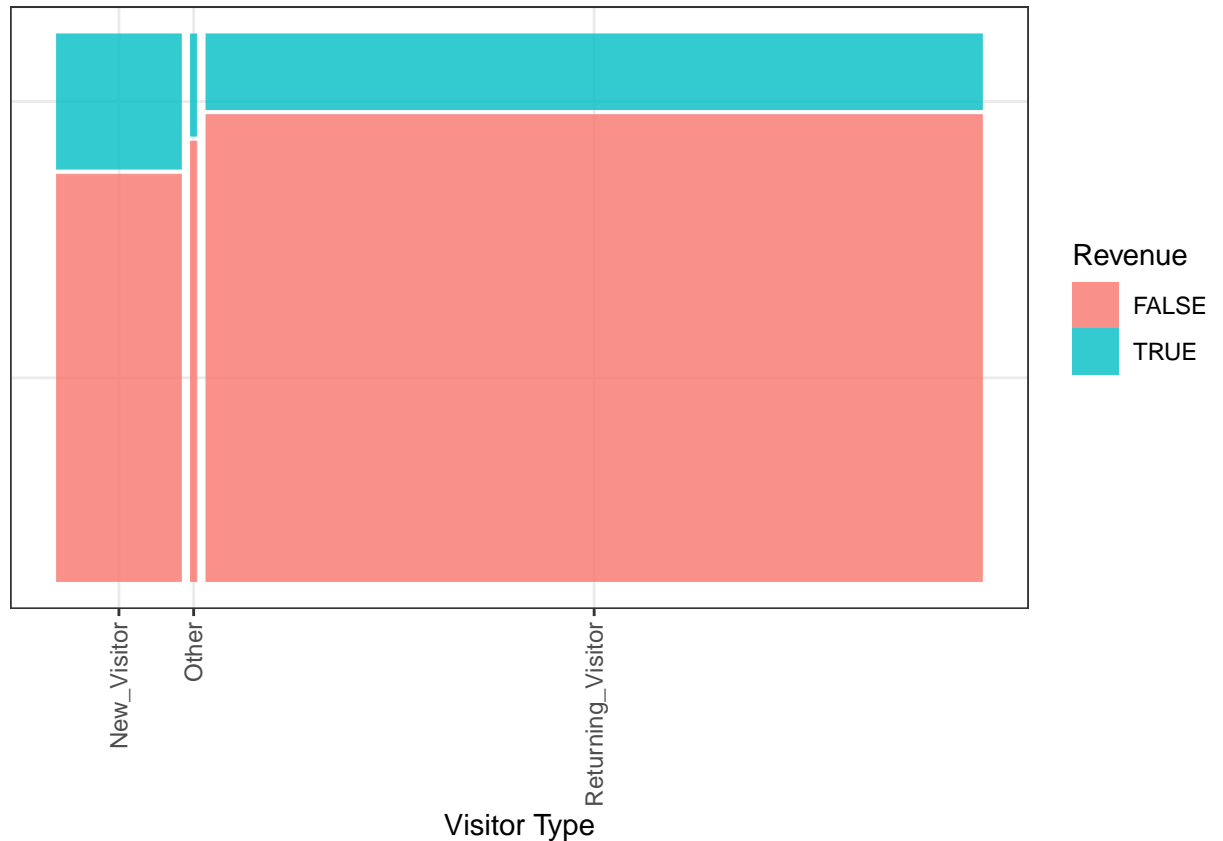
The plot portrays the high shopping rates in the months September, October and November with respect to the customers not buying the products. These months are comparatively considered as the *Holiday Season Months*. Also, there is high hits on the website with positive revenue in the month of may.

```
theme_set(theme_bw())

## setting default parameters for mosaic plots
mosaic_theme = theme(axis.text.x = element_text(angle = 90,
                                                  hjust = 1,
                                                  vjust = 0.5),
                     axis.text.y = element_blank(),
                     axis.ticks.y = element_blank())
```

```
dataset %>%
  ggplot() +
  geom_mosaic(aes(x = product(Revenue, VisitorType), fill = Revenue)) +
  mosaic_theme +
  xlab("Visitor Type") +
  ylab(NULL)
```





## Visitor

The comparison of the VisitorType which are New\_Visitors, Returning\_Visitor and Others with Revenue generated. There are many returning visitors in the contrast to less new visitors. Although, the new visitors have high probability of purchasing the product and help the revenue than the returning visitors.

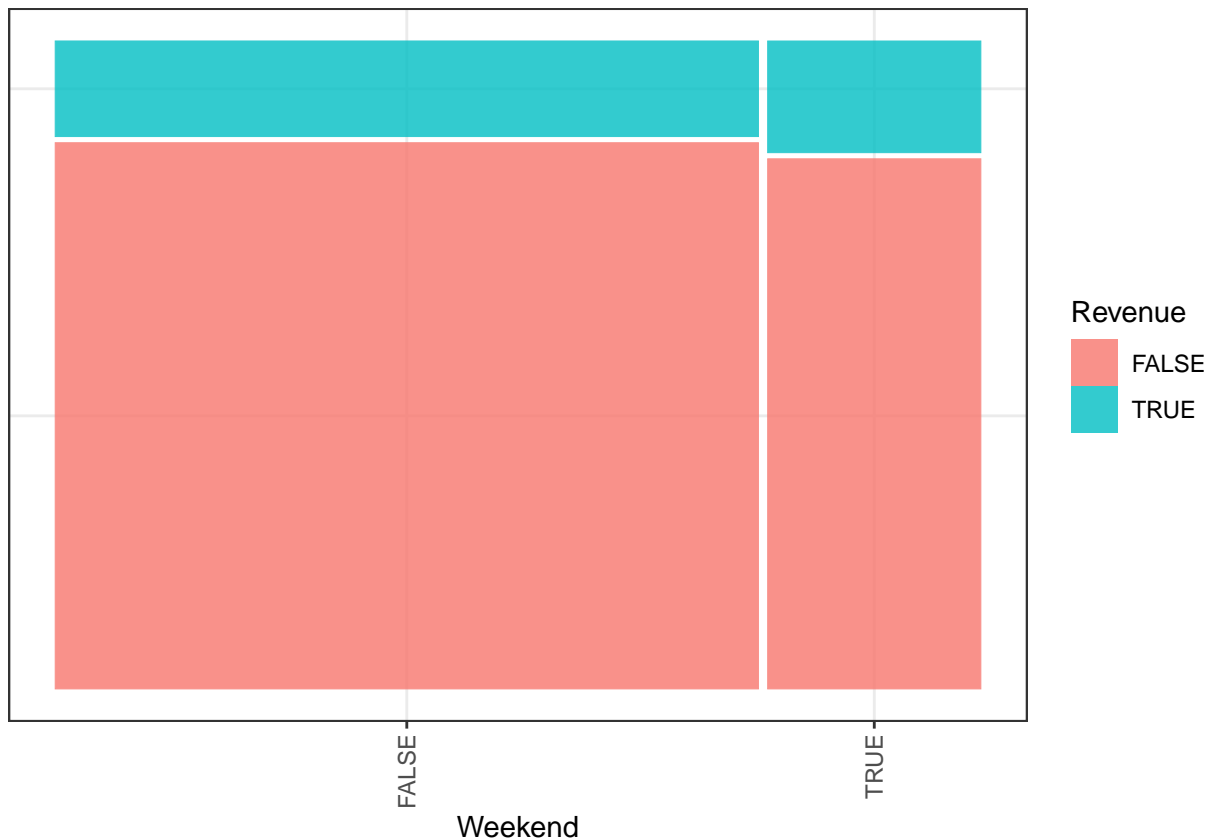
```
CrossTable(dataset$Weekend, dataset$Revenue)
```

## Weekend

```
##
##
##   Cell Contents
## |-----|
## |                      N |
## | Chi-square contribution |
## |      N / Row Total |
## |      N / Col Total |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table: 12330
##
##
```

```
##          | dataset$Revenue
## dataset$Weekend |      FALSE |      TRUE | Row Total |
## -----|-----|-----|-----|
##          FALSE |      8053 |      1409 |      9462 |
##          |      0.381 |      2.080 |          |
##          |      0.851 |      0.149 |      0.767 |
##          |      0.773 |      0.738 |          |
##          |      0.653 |      0.114 |          |
## -----|-----|-----|-----|
##          TRUE |      2369 |      499 |      2868 |
##          |      1.257 |      6.864 |          |
##          |      0.826 |      0.174 |      0.233 |
##          |      0.227 |      0.262 |          |
##          |      0.192 |      0.040 |          |
## -----|-----|-----|-----|
##      Column Total |      10422 |      1908 |      12330 |
##          |      0.845 |      0.155 |          |
## -----|-----|-----|-----|
##
##
```

```
dataset %>%
  ggplot() +
  mosaic_theme +
  geom_mosaic(aes(x = product(Revenue,Weekend), fill = Revenue)) +
  xlab("Weekend") +
  ylab(NULL)
```



The **Weekend** analysis shows that more than 70% of visitors are visiting the site on weekdays, with 15% chance of actually buying the products. The rest 30% visit on the weekend and there is 17% speculation of buying.

## Appendix—Code

```
knitr::opts_chunk$set(echo= TRUE, warning=FALSE, message=FALSE)
library(ggplot2)
library(tidyverse)
library(gmodels)
library(dplyr)
library(ggmosaic)
library(corrplot)
library(caret)
library(rpart)
library(rpart.plot)
library(cluster)
library(fpc)
library(data.table)
library(knitr)
library(kableExtra)
dataset <- read.csv("online_shoppers_intention.csv", header = TRUE)
attach(dataset)
ncol(dataset)
```

```

nrow(dataset)
str(dataset)
summary(dataset)
tab1 <- read.csv("table1.csv", header = TRUE)
kable(tab1) %>%
  kable_styling(full_width = T)
tab2 <- read.csv("table2.csv", header = TRUE)
kable(tab2) %>%
  kable_styling(full_width = T)
library(gmodels)
summary(dataset$Revenue)
CrossTable(dataset$Revenue)
dataset <- dataset %>%
  mutate(Revenue_binary = ifelse(dataset$Revenue == "TRUE", 1, 0))
colSums(is.na(dataset))
dataset %>%
  ggplot() +
    aes(x = Month, Revenue = ..count../nrow(dataset), fill = Revenue) +
    geom_bar() +
    ylab("Frequency")
table_month = table(dataset$Month, dataset$Revenue)
tab_mon = as.data.frame(prop.table(table_month, 2))
colnames(tab_mon) = c("Month", "Revenue", "perc")
ggplot(data = tab_mon, aes(x = Month, y = perc, fill = Revenue)) +
  geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) +
  xlab("Month") +
  ylab("Percent")
theme_set(theme_bw())

## setting default parameters for mosaic plots
mosaic_theme = theme(axis.text.x = element_text(angle = 90,
                                                  hjust = 1,
                                                  vjust = 0.5),
                     axis.text.y = element_blank(),
                     axis.ticks.y = element_blank())

dataset %>%
  ggplot() +
    geom_mosaic(aes(x = product(Revenue, VisitorType), fill = Revenue)) +
    mosaic_theme +
    xlab("Visitor Type") +
    ylab(NULL)
CrossTable(dataset$Weekend, dataset$Revenue)
dataset %>%
  ggplot() +
    mosaic_theme +
    geom_mosaic(aes(x = product(Revenue, Weekend), fill = Revenue)) +
    xlab("Weekend") +
    ylab(NULL)

```