



# Lab 2 Report: Predicting behavior of online shoppers

August 2020

---

York University - CSDA 1010 - Group 1

Melisa Theodore, Filipe Soares, Gabriel Fearon, Sina Aligholizadeh, Jagos Radovic

## Abstract

As the world entered into the Covid-19 pandemic earlier this year, we witnessed a massive surge of online shopping due to the lockdown measures that changed consumer behaviors. Many brick and mortar businesses, specifically small businesses, scrambled to develop their online sales and customer acquisition funnels. In that context, understanding the behavior of online shoppers became crucially important. In this project, Decision Tree, Linear Regression and Logistic Regression models were used to answer and predict the purchasing behaviors of visitors of online shoppers. To that end, we used the Online Shoppers Purchasing Intention Dataset published by Sakar et al. in 2018.

## 1. Business Understanding

### 1.1. Background

From its introduction into the consumer market, E-commerce has steadily increased over the years. In 2018, 84% of internet users completed online shopping transactions. This resulted in a total spending of \$57.4 billion, (Statistics Canada, 2018). During the COVID-19 lockdown (February to May 2020), online shopping sales soared to 99.3%. April recorded the most significant change with sales increasing from 3.8% in 2019 to 11.4% in 2020 resulting in \$33.9 billion. In May, consumers continued to purchase online at 5x times the rate as in-store purchases, (Statistics Canada, 2020).

As COVID-19 lingers, people are choosing to stay out of malls and stores and continue to shop online, even as the lockdown measures eased, (The Globe & Mail, 2020). Prior to the lockdown, the reasons why 16% of internet users chose not to shop online would now see a huge shift: “no need or no interest”, and “wanted the opportunity to see, hold, or try on the product before purchasing”. It would be suffice to say that online shopping may remain as the primary shopping option.

In March and April 2020, Shopify experienced a growth of 62%, as many small businesses rushed to create online shopping experiences to stay afloat, (Shopify, 2020). Since small businesses are the largest employers in the private sector, it is important to study the behaviours of online shoppers and be able to predict their purchasing intentions for businesses, to assist in their creation of e-commerce platforms.

Window shopping is a common behaviour in the 21st century which is just for individuals to fill up little gaps in their day. Also considering our products over a good range of different ethnicities or backgrounds. All of these different backgrounds together contain a lot of special holidays and etc. for example, Middle east and North America don't share the same day for mothers day or father's day. If we take into consideration how many different

special days we have and how we can focus on those days with different specials, discounts or even advertisements, there could be a lot of improvement that could be made to maximize revenue.

## 1.2. Objective and Hypothesis

The objective of this project is to evaluate factors that are contributing to the online purchasing behavior or website visitors, namely their decision to finalize their visit to the website with a transaction.

We hypothesized that certain variables such as the time spent on certain pages, weekends, days around holidays, type of visitors (e.g. new or returning), geographic location, were the biggest determinants for a website visitor to engage in an online purchase.

In order to achieve our objective, a most optimal model will be selected to predict whether or not a website visitor is likely to engage in a transaction on the website based on factors such as Product Related Duration, Month, Special Day, Page Value, Browser, Region, Traffic Type, Visitor Type, Exit Rates, Bounce Rates, and Weekend. We will build, analyze and compare three models, including Linear Regression, Logistic Regression and Decision Tree models.

## 1.3. Assumptions

Key assumption is behavioral independence of each unique website visitor, i.e. visitors do not influence each other's shopping decision. Since we don't know the user demographic data, we will also assume that their key socio-economic characteristics of website visitors are representative of the general population, i.e. are not heavily skewed to any particular feature, such as age, gender, income bracket, etc.

# 2. Data understanding

We selected the "Online Shoppers Purchasing Intention" dataset from the UCI Machine Learning Repository. Original source of the data is a paper by Sakar et al. (2018). The dataset consists of 18 features vectors belonging to 12,330 website sessions. According to the dataset description, each session corresponds to a different user, within a 1-year period. In this way, any bias towards a specific campaign, special day, user profile, or period has been minimized.

**Table 1.** List of dataset features with their descriptions (modified from Sakar et al., 2018).

Feature	Feature description
<b><i>Numerical</i></b>	
Administrative	Number of pages visited by the visitor about account management
Administrative	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages
Product related	Number of pages visited by visitor about product related pages
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce rate	Average bounce rate value of the pages visited by the visitor
Exit rate	Average exit rate value of the pages visited by the visitor
Page value	Average page value of the pages visited by the visitor
Special day	Closeness of the site visiting time to a special day
<b><i>Categorical</i></b>	
OperatingSystems	Operating system of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)
VisitorType	Visitor type as “New Visitor,” “Returning Visitor,” and “Other”
Weekend	Boolean value indicating whether the date of the visit is weekend
Month	Month value of the visit date
Revenue	Class label indicating whether the visit has been finalized with a transaction

As can be seen in the Table 1, dataset features include 10 numerical and 8 categorical attributes. Feature descriptions provided in the dataset source are fairly self-explanatory, except the “Page value” feature. Upon further investigation, we found out that this is a metric calculated by Google Analytics, and defined as the average dollar value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction, or both (<https://yoast.com/what-is-page-value-in-google-analytics/>).

We examined the structure of the dataset using the *str* function in R. As mentioned, it can be observed that the majority of variables are numeric (integer or floats), “Month” and “Visitor Type” are factors, and “Weekend” and “Revenue” are logical values (“True” or “False”), Fig. 1.

```
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 : 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 ...
```

**Figure 1.** Structure of the investigated dataset.

Next, we checked for the presence of missing data, using the *colSums* function, and observed that the dataset does not contain any missing values, Fig 2.



```
colSums(is.na(data))

##      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

colSums(data=="")

##      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
```

**Figure 2.** Summary of missing data evaluation.

Finally, we reviewed some basic descriptive statistics of dataset features using the *summary* function, Fig. 3.

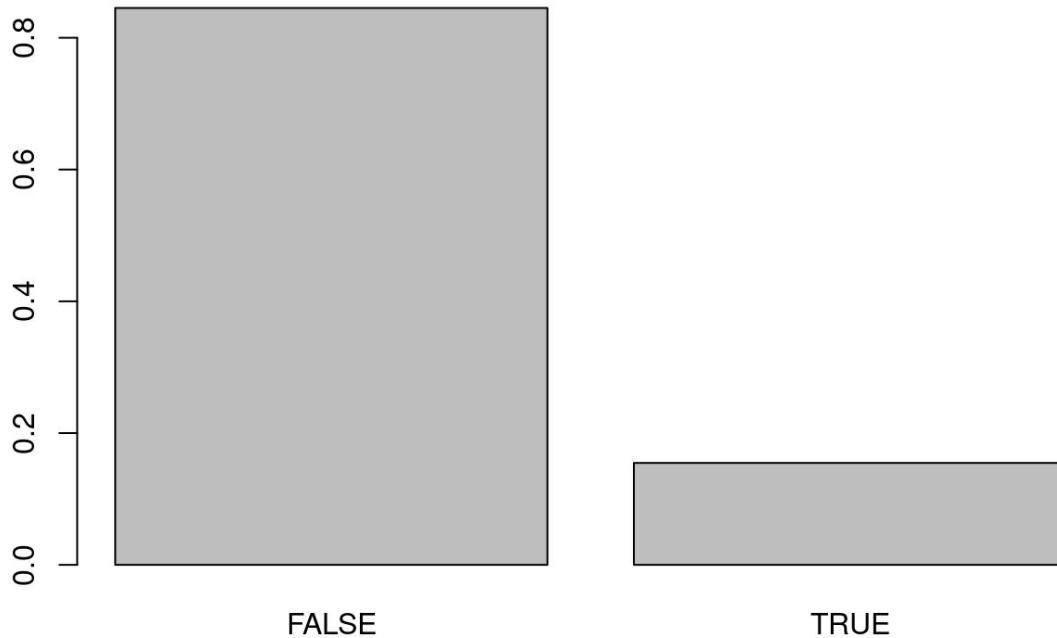
```
summary(data)

## 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   : 0.5036
## 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
##
```

**Figure 3.** Basic statistic descriptors of the dataset features.

Some relevant observations inferred from descriptive statistics include the fact that visitors spend significantly more time on product related website pages, on average 1195 seconds, compared to administrative and informational pages, with average duration of ~81 and 34 seconds; next, we can observe that website has approx. an order of magnitude more returning visitors than new visitors (10551 versus 1694); finally we can deduce that only a smaller proportion of visitors finalizes their website session with a transaction/purchase ("Revenue" attribute).

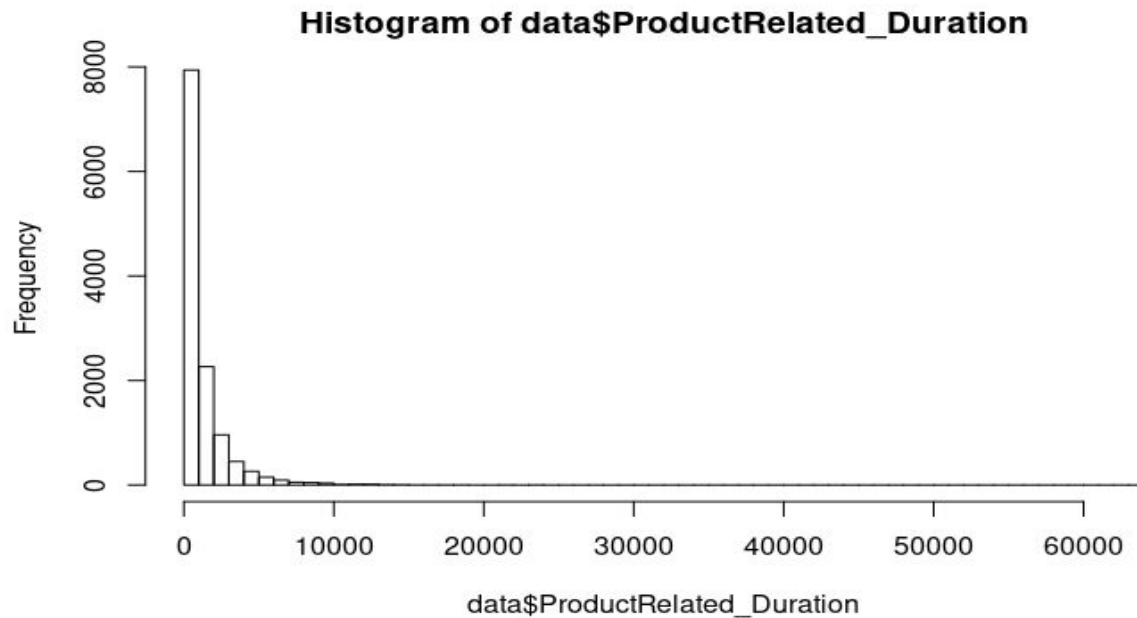
More precisely, by using *prop.table* function on the “Revenue” feature, we discovered that ~15.5% of website visitors engage in a transaction, while 84.5% of users only visit the website, as shown in the barplot, Fig 4.



**Figure 4.** Barplot showing the proportion of website visitors that complete the transaction during the website session (“True”), compared to the fraction of visits that do not result in a purchase (“False”).

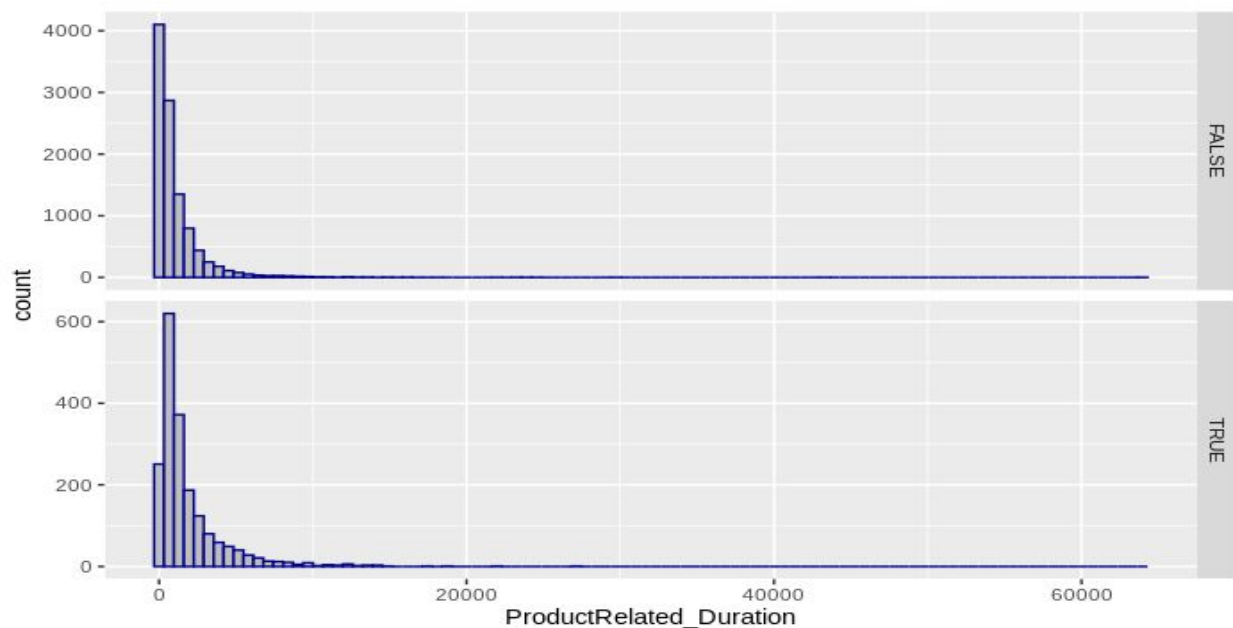
In addition to “Revenue”, another feature of interest is the duration of sessions on product related pages. We already saw that users spend significantly more time browsing those pages. More specifically, most of the users will spend 1,000 seconds on product pages.





**Figure 5.1.** Histogram regarding duration of sessions on product related pages

If we focus on revenue generating sessions, most of them lasted ~1280 seconds, i.e. approx. 21 minutes, compared to ~640 seconds (approx. 11 minutes) which was spent on most of the sessions that did not result in a transaction.

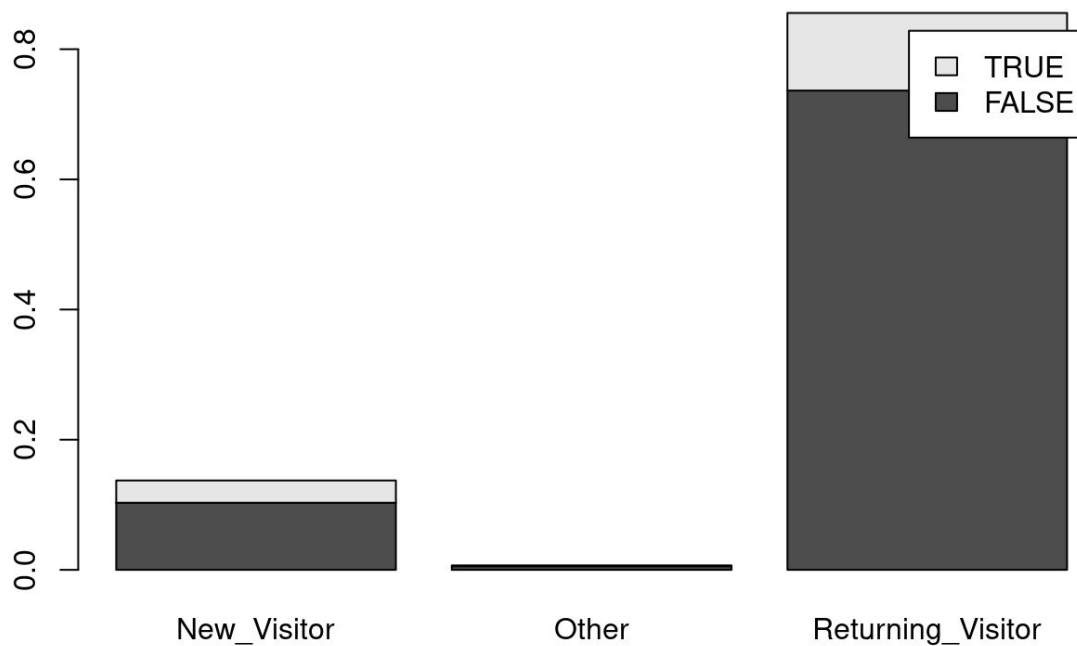


**Figure 5.2.** Histogram regarding product related duration split into true and false outcomes

In terms of visitor type, somewhat counterintuitively, it can be observed that there is a higher proportion (~0.33) of sessions ending up in transactions among the “New\_Visitor” category, compared to the “Returning\_Visitor” (~0.16), Fig. 6.

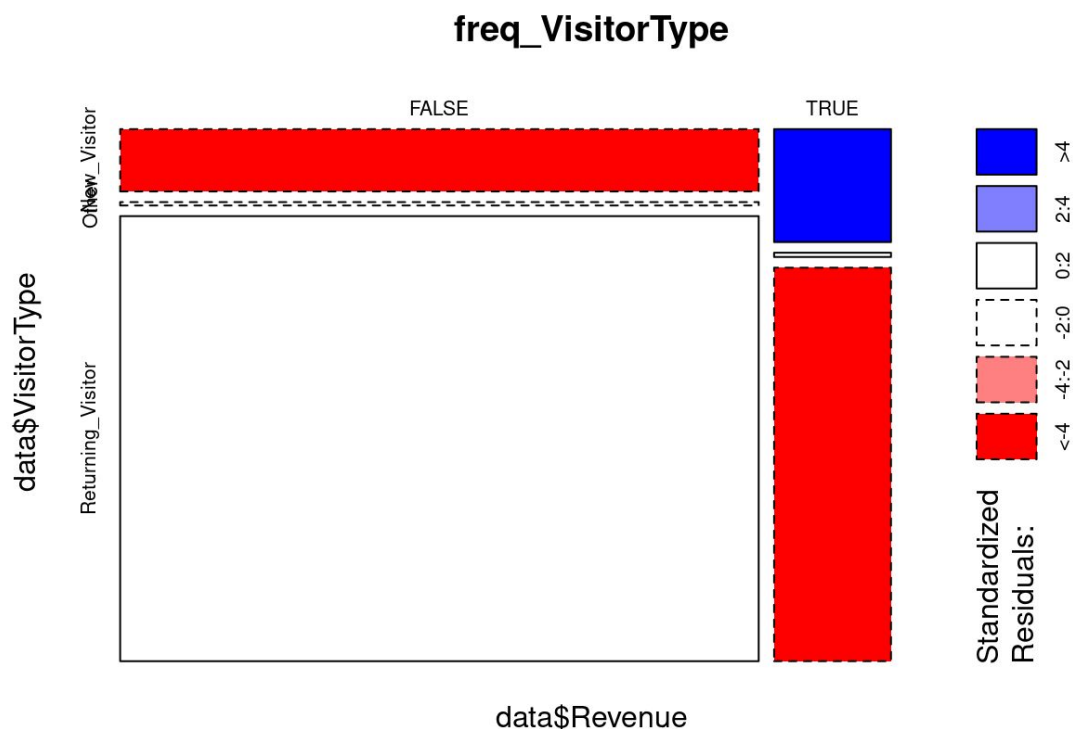
```
freq_VisitorType=xtabs(~data$Revenue+data$VisitorType)
prop.table(freq_VisitorType)
```

```
##          data$VisitorType
## data$Revenue New_Visitor   Other Returning_Visitor
##      FALSE 0.103163017 0.005596107 0.736496350
##       TRUE 0.034225466 0.001297648 0.119221411
```



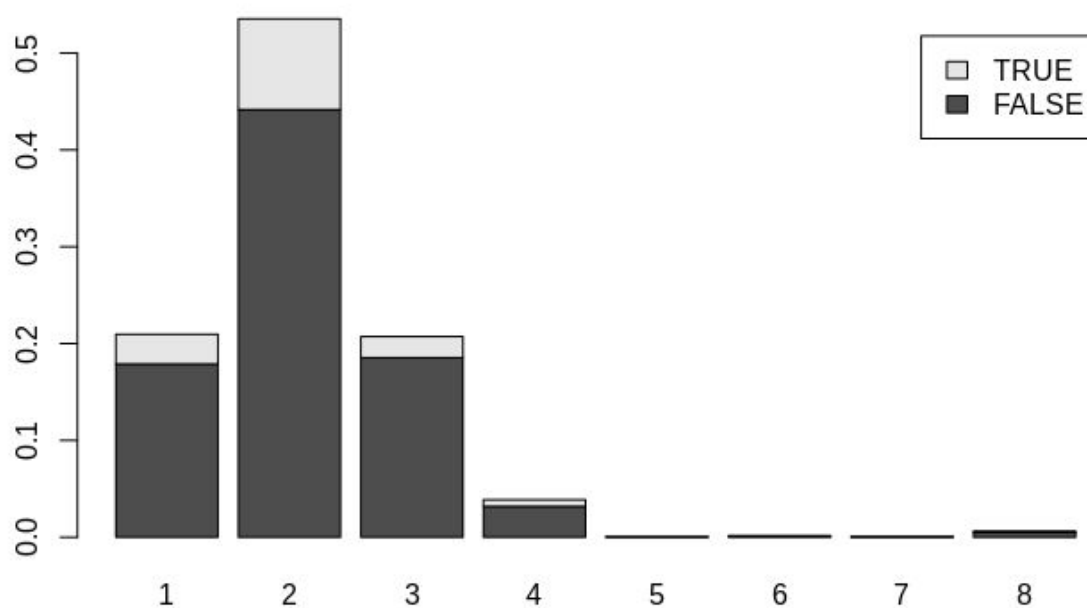
**Figure 6.** Barplot showing the purchasing behavior of different types of users.

We confirmed this observation by using the mosaic plot with standardized residuals, which is able to show whether certain observations are over (blue shades) or under (red shades) represented in a given feature cell, more that it would be the case for null hypothesis, i.e. if variables were independent.



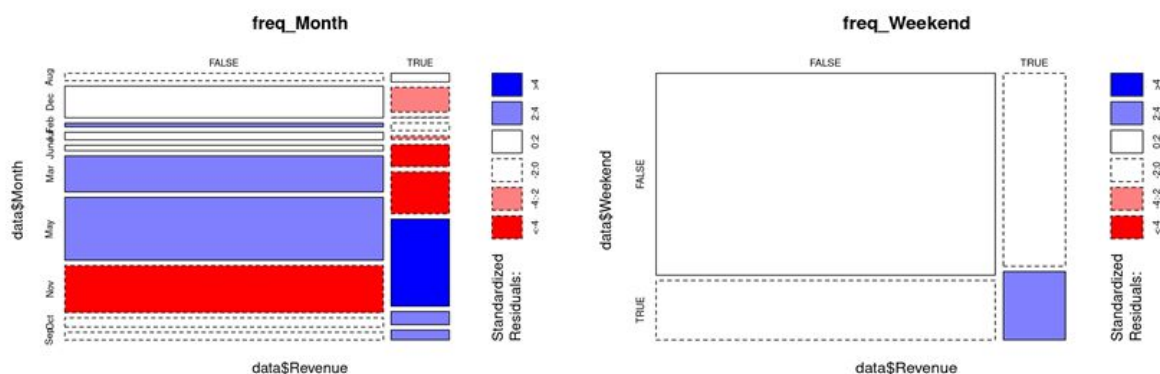
**Figure 7.** Mosaic plot with standardized residuals for VisitorType and Revenue feature.

Majority of the users are surfing the web with the operating system (OS) number 2, which is also the system from which the highest proportion of website transactions are made. We do not have more detailed descriptions of the operating systems in this dataset, but we can assume that the operating system 2 corresponds to one of the most ubiquitous OS types - Windows or Mac OS.



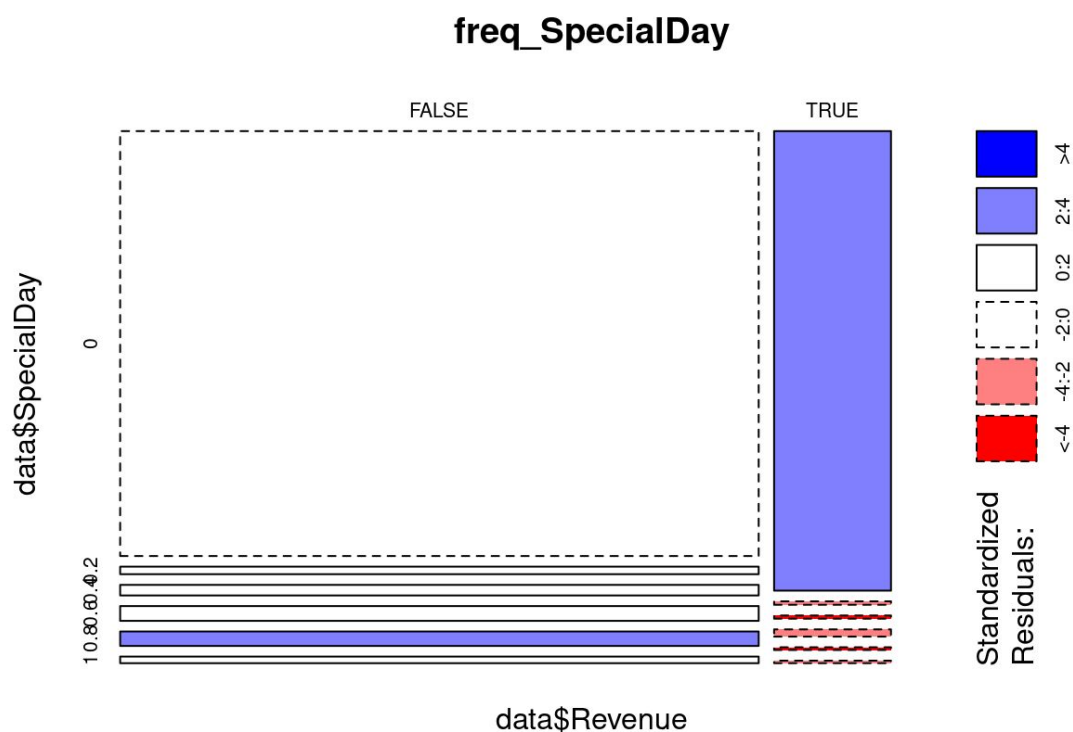
**Figure 8.** Barplot showing purchasing proportions based on the operating system.

Mosaic plots were also useful to determine that during weekends there are more revenue-generating sessions, as well as during the months of September, October and November, likely indicating the fall shopping season, related to the start of the school year and various year-end holidays, Figs. 9.1 and 9.2.



**Figure 9.1 and 9.2.** Mosaic plots illustrating revenue generated based on month and weekend.

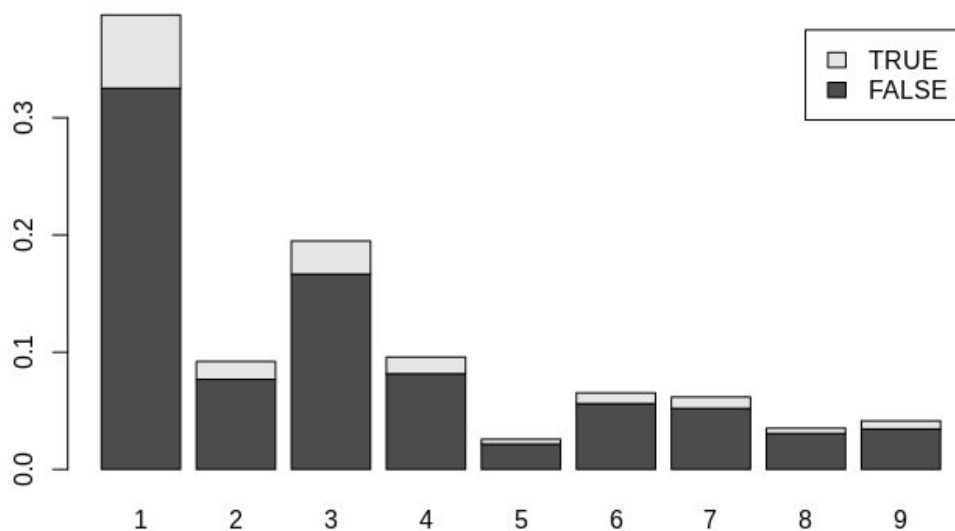
In relation to holidays, and other special days, there is an apparent tendency to commit to purchase online transactions on the day of the special occasion, Fig. 9.3.



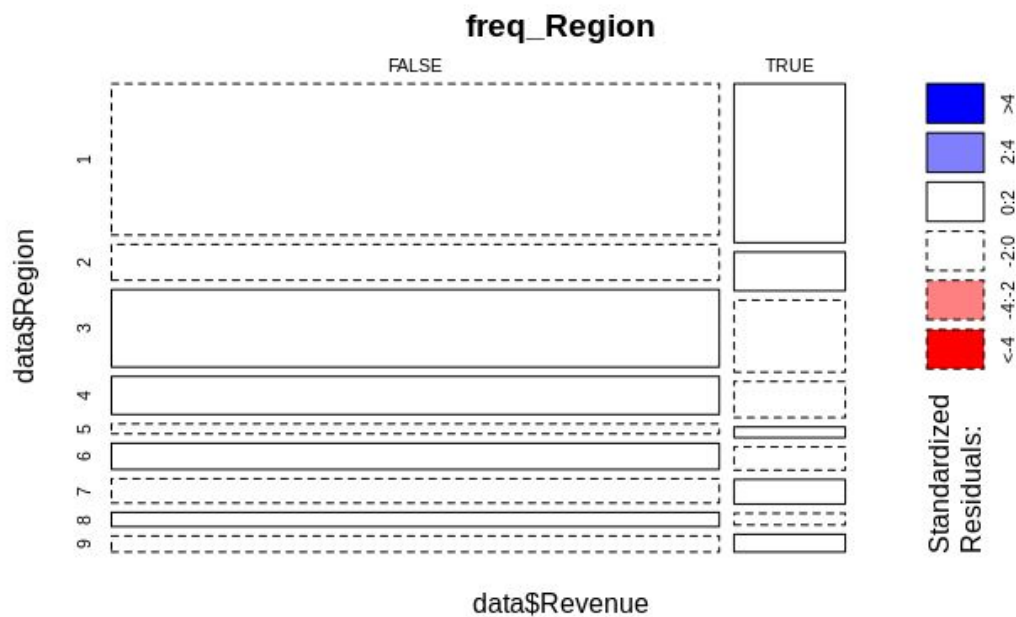
**Figure 9.3.** Mosaic plot showing revenue with respect to holidays.

Geographically, out of the 9 regions, most of the users are located in Region 1; however, mosaic plot residuals suggest that there is no strong relationship between the users' location and the propensity to commit to the transaction, Figs 10.1 and 10.2.





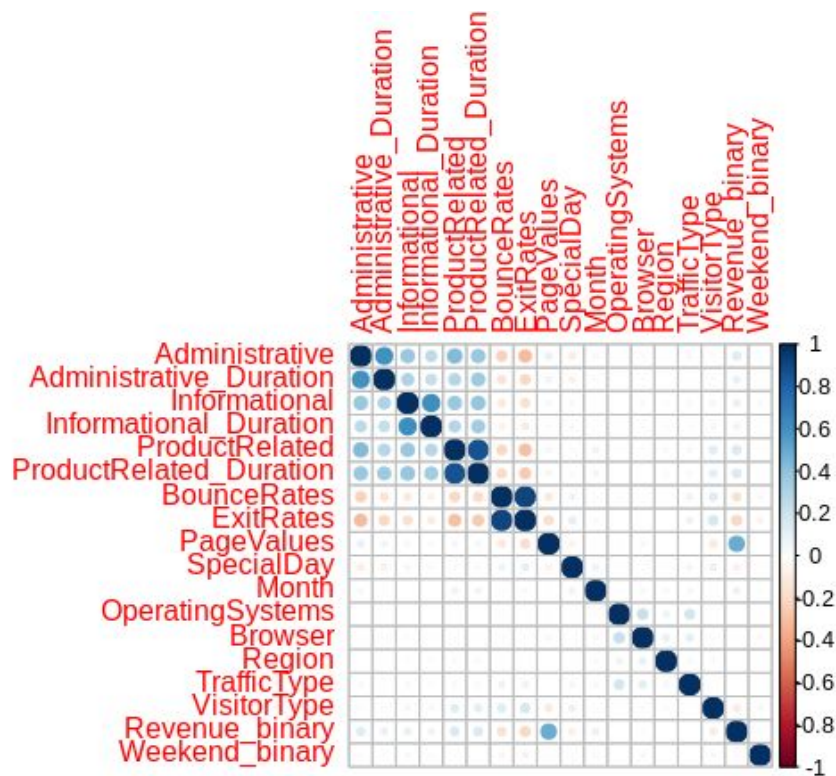
**Figure 10.1.** Barplot showing proportion of purchases based on region.



**Figure 10.2.** Mosaic plot showing revenue with respect to region.

### 3. Feature Engineering and Modelling

As an initial evaluation, we computed and plotted a matrix of correlation coefficients for all possible pairs of features in the dataset, using *rcorr* and *corrplot* functions. The output suggests the strongest positive correlation ( $\sim 0.49$ ) between the “Revenue” and “Page Values” features, Fig. 11. On the other hand, the “Revenue” feature is most negatively correlated with exit and bounce rates. Another strong negative correlation is found between the ProductRelated\_Duration and ExitRates, meaning that users that browse product related content on the website are less likely to quickly exit the session.



**Figure 11.** Correlation matrix of dataset features.

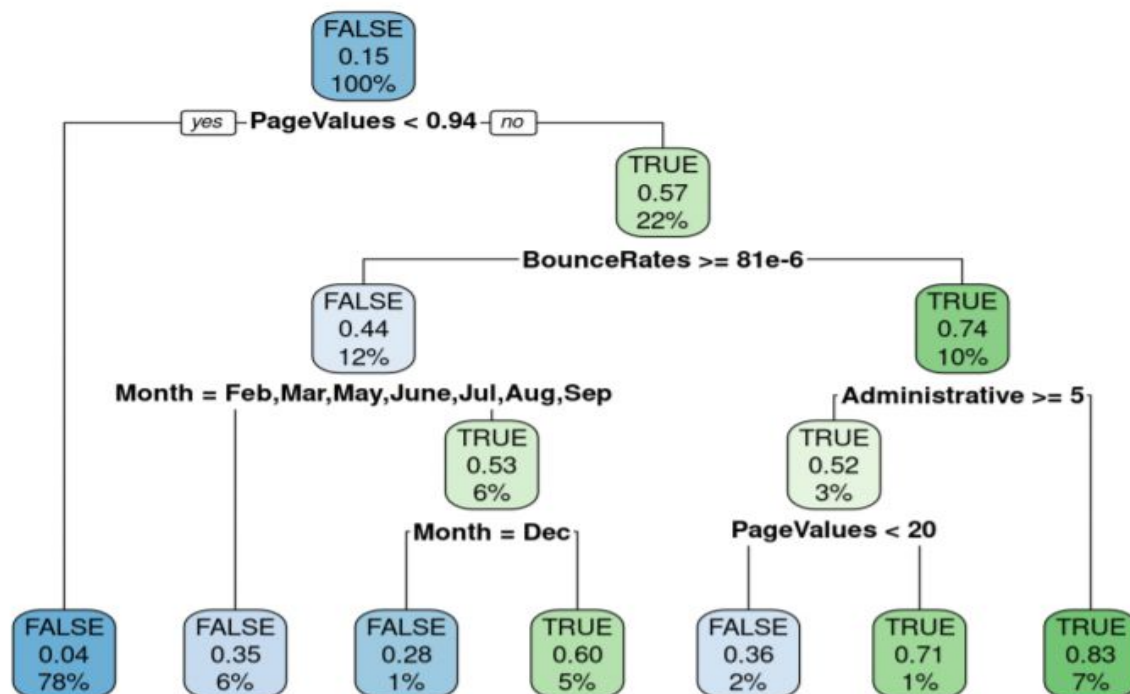
### Model development and results

#### Decision Tree:

One of the selected models was the decision tree because this model is generally easy to understand due to its simple presentation. Decision trees are also effective when dealing


with multicollinearity, so it is a useful model when using a dataset that has many variables that appear to be related. Similar to how human beings make decisions, the decision tree weighs the most important variables and considers the consequences at each point (node) with a yes/no (True/False) response. This process continues until decisions with respect to all significant variables until we are left with the final decisions (end nodes).

It was decided to create a decision tree that looked at what factors would contribute the most to determining whether or not a sale was made. The decision tree determined that page values, bounce rates, month, the time spent on administrative pages and the product related pages were the most important variables in determining if a customer was going to purchase a product. The multiple appearances of page values indicates that it is most likely the biggest factor in identifying who will be a paying customer.



**Figure 12.** Decision tree created with respect to the revenue variable.

The figure above illustrates the predicted process in whether or not revenue will be generated by customers visiting the website. The first node splits the training data with respect to page values. If the value is less than 0.94, we see that 78% of the online shoppers will not be purchasing. The remaining 22% are split by the bounce rate variable. 12% of shoppers go down the left side, 10% to the right. The left side of this split leads to the third node which is split on whether the shoppers visited in February, March, May, June,



July, August or September. If the answer is yes, they did not make a purchase. If the answer is no, the remaining customers are split into groups who visited in December or not. If the answer was yes, there was no sale and the remaining 5% purchased an item. The right branch of the bounce rate node leads to a decision regarding administrative pages. 3% went to the left to be further split by page values one more time. The remaining 7% purchased an item. The decision tree concludes with 7 end nodes each containing the term “True” or “False” with a percentage inside. When adding the False nodes together it is understood that the tree predicts 87% of shoppers will not be buying an item.

While the visual representation of the data is easy to understand, decision trees can be volatile at times and may be altered based on the data it samples. They also have a tendency to overfit the data so it may be difficult to use decision trees to predict outcomes in the world outside of the model’s dataset. However, a decision tree may be useful in this situation because the majority of the visitors were returning customers so the model may be able to predict whether or not a sale will occur based on their previous behaviour. These are some of the key factors that must be evaluated when deciding which model to use.

## **Linear regression**

One potential way of analyzing this data set is through an ordinary least squares (OLS) linear regression. The use of a linear regression will provide several different forms of information that could be relevant in determining how different variables impact individuals' online shopping behavior. This can include displaying how different variables affect our chosen dependent variables and depending on the effectiveness of the model, the regression will provide some predictive capability in determining individuals' online shopping behavior. As well, in ensuring the assumptions needed to make an OLS regression, we can determine a number of details regarding the nature of the data set.

The linear regression will focus on the Product Related Duration variable in the data set. This variable describes how long individuals spend looking at a particular type of product. Product Related Duration could be important to many online businesses for the purposes of designing their website to increase sales, maintain high levels of site traffic and other business challenges. This paper will detail two OLS regressions examining the Product Related Duration, the first will use all other variables in the regression as independent variables, while the second will take a more focused approach based only on the results that are statistically significant in the first model. When running the first regression there are a number of interesting results that are found. The first is that variables such as the operating system, browser type and the individual's geographical region was not statistically significant in determining the Product related duration rate. As well, while the special day (how close it was the web activity was to a holiday) variable was statistically significant, the month that the web activity occurred was not statistically significant. The negative value of Special day also suggests that individuals who shop closer to a holiday

they spend less time looking at related products to what they are looking for. This could potentially mean that individuals when online shopping have a clearer idea of what they are looking for or spend less time looking for alternatives.

```
##
## Call:
## lm(formula = ProductRelated_Duration ~ ., data = dataReg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6390  -278    -78    158   39083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.445e+02  4.841e+01  -2.986 0.002832 **
## Administrative -5.869e+01  3.438e+00 -17.073 < 2e-16 ***
## Administrative_Duration 1.642e+00  5.912e-02  27.780 < 2e-16 ***
## Informational    1.131e+01  8.828e+00   1.282 0.199992
## Informational_Duration 1.360e+00  7.526e-02  18.070 < 2e-16 ***
## ProductRelated  3.574e+01  2.237e-01 159.744 < 2e-16 ***
## BounceRates    -2.199e+03  4.322e+02  -5.087 3.69e-07 ***
## ExitRates       2.156e+03  4.559e+02   4.728 2.29e-06 ***
## PageValues     -3.136e-01  5.177e-01  -0.606 0.544687
## SpecialDay     -1.273e+02  4.239e+01  -3.004 0.002674 **
## Month          2.684e+00  3.541e+00   0.758 0.448438
## OperatingSystems 1.514e+00  9.488e+00   0.160 0.873182
## Browser        4.790e+00  4.999e+00   0.958 0.338018
## Region         1.514e+00  3.480e+00   0.435 0.663481
## TrafficType     4.395e-01  2.115e+00   0.208 0.835420
## VisitorType     2.858e+01  1.259e+01   2.270 0.023238 *
## Revenue_binary  9.105e+01  2.695e+01   3.379 0.000731 ***
## Weekend_binary -3.461e+01  1.970e+01  -1.757 0.078924 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 919.8 on 12312 degrees of freedom
## Multiple R-squared:  0.7693, Adjusted R-squared:  0.769
## F-statistic: 2415 on 17 and 12312 DF, p-value: < 2.2e-16
```

**Image 1.** First linear regression model.

The first linear regression provides a fairly encompassing model for the data set with an adjusted R-square of 0.769. In an attempt to create an even more accurate model, a second linear regression had been constructed that would remove the variables that were not statistically significant in determining the Product Related Duration. Following this logic, to create this model the variables: informational, page values, months, operating systems, browser, region and traffic type were not included. As well this model also breaks down the months into individuals' binary variables. This may provide more accuracy and predictive power as each month may have its own individual impact rather than a constant rate of impact as the year continues.



```
##
## Call:
## lm(formula = ProductRelated_Duration ~ Administrative + Administrative_Duration +
##      Informational_Duration + ProductRelated + BounceRates + ExitRates +
##      SpecialDay + factor(Month) + VisitorType + Weekend_binary +
##      Revenue_binary, data = dataReg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6257   -284    -79    162   38948
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.691e+02  5.661e+01  -4.754 2.02e-06 ***
## Administrative -5.646e+01  3.408e+00 -16.565 < 2e-16 ***
## Administrative_Duration 1.641e+00  5.897e-02  27.832 < 2e-16 ***
## Informational_Duration 1.404e+00  6.238e-02  22.512 < 2e-16 ***
## ProductRelated  3.581e+01  2.241e-01 159.757 < 2e-16 ***
## BounceRates     -2.193e+03  4.302e+02  -5.097 3.51e-07 ***
## ExitRates       2.164e+03  4.534e+02   4.773 1.84e-06 ***
## SpecialDay     -1.322e+02  4.790e+01  -2.760 0.005796 **
## factor(Month)2  1.988e+02  4.944e+01   4.021 5.82e-05 ***
## factor(Month)3  2.328e+02  8.198e+01   2.839 0.004528 **
## factor(Month)4  6.625e-01  6.248e+01   0.011 0.991540
## factor(Month)5  6.342e+01  6.995e+01   0.907 0.364593
## factor(Month)6  1.926e+02  4.909e+01   3.924 8.75e-05 ***
## factor(Month)7  1.618e+02  4.807e+01   3.367 0.000763 ***
## factor(Month)8  1.872e+02  4.732e+01   3.956 7.67e-05 ***
## factor(Month)9  1.607e+01  5.908e+01   0.272 0.785630
## factor(Month)10 1.772e+02  6.192e+01   2.862 0.004212 **
## VisitorType    2.809e+01  1.262e+01   2.225 0.026080 *
## Weekend_binary -3.505e+01  1.966e+01  -1.783 0.074658 .
## Revenue_binary  8.545e+01  2.392e+01   3.573 0.000354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 918.1 on 12310 degrees of freedom
## Multiple R-squared:  0.7702, Adjusted R-squared:  0.7698
## F-statistic: 2171 on 19 and 12310 DF, p-value: < 2.2e-16
```

**Image 2.** Second linear regression model.

While the new model provides a number of similar results to the first regression, there are several areas in which this model differs. The first includes a lower intercept and a number of variables having a slightly larger impact on the product related duration. As well, most of the variables retain their statistical significance in the model that the variables had in the first regression. The largest difference in this second model compared to the previous model was the inclusion of the month binary variables. While most of the months are statistically significant, there are several months that are not. As there are some months (January & April) that were not included in the data, the numbers on the result table do not represent the correct month (see *Appendix Table 1*). The months that are not statistically significant are June, July and November. The focus of the other types of models in this lab, the revenue binary variable, was statistically significant. The estimate was also a positive value of 8.545e+01 this would suggest that individuals who did purchase a product generating revenue for the site spent longer on product related sites. Notably in the second

linear regression, the impact that the revenue has is smaller than the original model. This model also provides a higher R-square then the first model. Even when taking into account an increased R-squared value due to the second model having a larger number of variables the second model has a higher adjusted R-squared value.

### Logistic regression - Generalized Linear Model (GLM)

Logistic regression is an extension of simple linear regressions which can be used for cases when the target variable is not continuous, as is the case with the binary “Revenue” feature in our dataset. Thus, we built a binomial generalized linear model for the revenue against the rest of the dataset features.

```
glm = glm(Revenue_binary ~ ., data=trainReg, family=binomial)
summary(glm)
```

```
##
## Call:
## glm(formula = Revenue_binary ~ ., family = binomial, data = trainReg)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -6.0703  -0.4778  -0.3551  -0.1768   3.4122
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.300e+00  2.110e-01 -10.899  < 2e-16 ***
## Administrative    1.586e-02  1.299e-02   1.221  0.222051
## Administrative_Duration -2.429e-04  2.349e-04  -1.034  0.301044
## Informational    -2.892e-03  3.225e-02  -0.090  0.928529
## Informational_Duration  2.551e-04  2.620e-04   0.974  0.330232
## ProductRelated    3.410e-03  1.427e-03   2.389  0.016876 *
## ProductRelated_Duration  5.813e-05  3.465e-05   1.678  0.093404 .
## BounceRates     -3.177e+00  3.865e+00  -0.822  0.411074
## ExitRates       -1.585e+01  2.864e+00 -5.533  3.15e-08 ***
## PageValues      7.993e-02  2.816e-03  28.380  < 2e-16 ***
## SpecialDay     -8.736e-01  2.602e-01  -3.357  0.000788 ***
## Month          1.125e-01  1.634e-02   6.884  5.83e-12 ***
## OperatingSystems -5.815e-02  4.516e-02  -1.288  0.197811
## Browser         3.308e-02  2.188e-02   1.512  0.130636
## Region        -1.538e-02  1.571e-02  -0.979  0.327808
## TrafficType     9.045e-03  9.791e-03   0.924  0.355585
## VisitorType    -1.836e-01  5.041e-02  -3.643  0.000270 ***
## Weekend_binary   1.628e-01  8.425e-02   1.932  0.053323 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 7438.7  on 8630  degrees of freedom
## Residual deviance: 5112.3  on 8613  degrees of freedom
## AIC: 5148.3
##
## Number of Fisher Scoring iterations: 7
```

**Image 3.** Logistic regression - Generalized Linear Model (GLM).

The model output suggests high level of significance for ExitRates, PageValues, SpecialDay, Month, and VisitorType (denoted by 3 asterisks, i.e. \*\*\*, corresponding to p value < 0.001).

We performed the initial evaluation of predictive power of the model on the train dataset, which resulted in an average prediction probability of ~45% for revenue-generating transactions.

```
predictglm = predict(glm, type="response")
summary(predictglm)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
## 0.0008732 0.0458509 0.0850265 0.1547909 0.1504548 1.0000000
```

```
tapply(predictglm, trainReg$Revenue_binary, mean)
```

```
##      0      1
## 0.1014172 0.4462285
```

**Image 4.** Evaluation of predictive power of GLM

In addition, the predictive power for the desired outcome, i.e. revenue occurring, is higher than for the negative outcome, i.e. session ended without transaction, which is an indicator of favorable model performance.

## 4. Model evaluation and comparison

### Decision Tree:

When creating the decision tree, we used the 'rpart' function from the 'rpart.plot' library in R. This function already uses k-fold cross-validation when building the tree, so pruning was not necessarily required in this situation. However, we still manually used 'complexity parameter' (CP) and cross-validation error rate (xerror) to confirm the correct data was used. After implementing the pruning technique, the resulting decision tree was identical to the original. The code for the pruned tree can be found below.

```
ptree<- prune(dtree, cp= dtree$scptable[which.min(dtree$scptable[, 'xerror']), 'CP'])
rpart.plot(ptree)
```

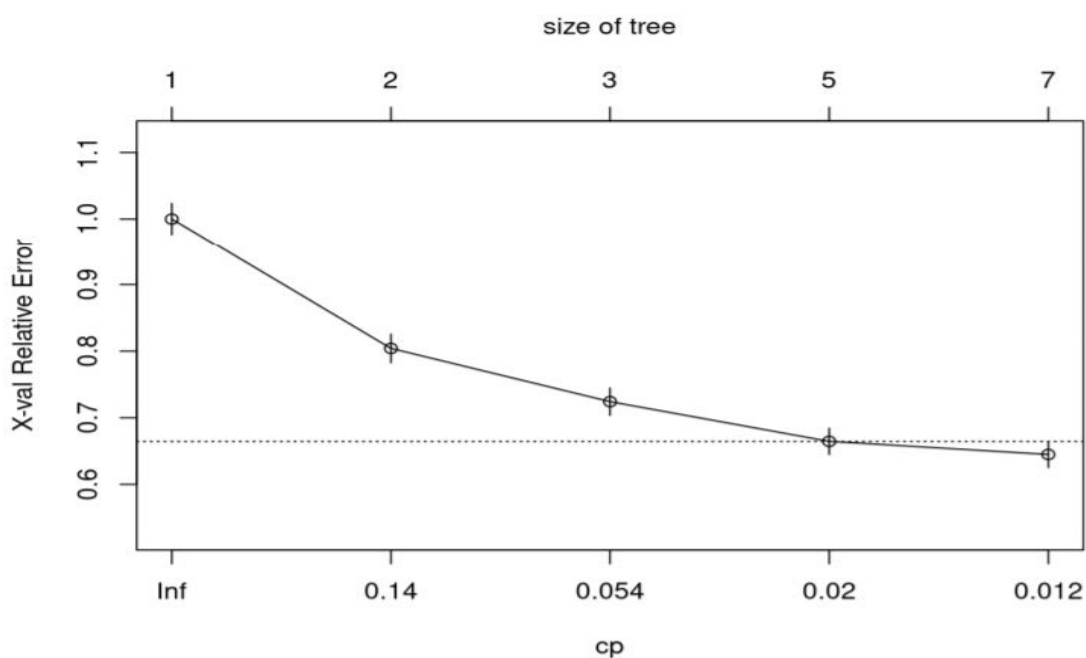
**Image 5.** Code for pruning decision tree.

The “cptable” in the code above refers to the table below. The code selects the CP with the lowest value, which also has the lowest xerror.

```
## Classification tree:
## rpart(formula = Revenue ~ ., data = train_tree, method = "class")
##
## Variables actually used in tree construction:
## [1] Administrative BounceRates Month PageValues
##
## Root node error: 1526/9864 = 0.1547
##
## n= 9864
##
##      CP nsplit rel error  xerror  xstd
## 1 0.196592    0  1.00000 1.00000 0.023536
## 2 0.101573    1  0.80341 0.80406 0.021479
## 3 0.029161    2  0.70183 0.72412 0.020527
## 4 0.014089    4  0.64351 0.66448 0.019766
## 5 0.010000    6  0.61533 0.64482 0.019504
```

**Image 6.** Complexity parameter table

The previous table can also be visualized in the plot below to determine which CP should be used when creating the decision tree.



**Figure 13.** Complexity parameter plot

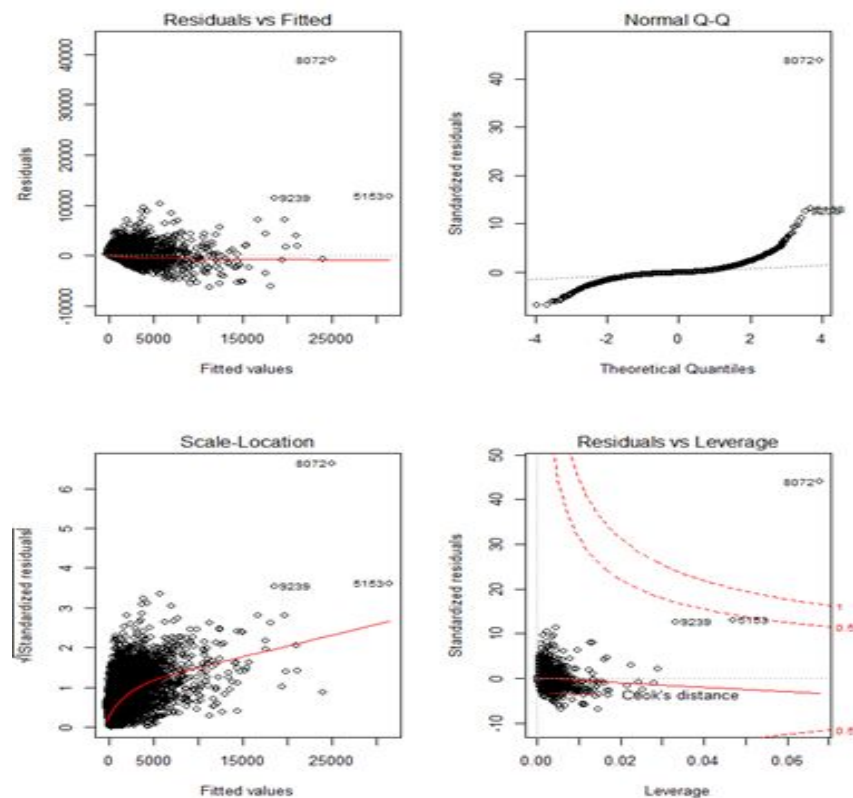
We also used accuracy, precision, recall and balanced accuracy to determine which model to use. This was done using a confusion matrix on the test data.

predTree				
	FALSE	TRUE		
FALSE	1988	96	Accuracy: 0.8941606	Precision: 0.6932907
TRUE	165	217	Balanced Accuracy: 0.7609988	Recall: 0.5680628

**Image 7.** Confusion matrix for decision tree with accompanying values for model metrics.

## Linear Regression

Before it can be determined which of the two OLS models are the better model, it must be proven that these models meet the necessary assumptions for an OLS regression. As shown in the previous section from the correlation plot the model does not exhibit multicollinearity. To determine if the models violate any of the other assumptions several plots will need to be created.



**Figure 14.** Comparing the OLS models



The first plot of interest is the Residual vs Fitted Plot. While the beginning of the plot opens out in a funnel shape, this shape does not persist for most of the fitted values. From this plot it can be interpreted that this data is linear though it is also worth noting that there are some outliers. The next plot is the Normal Q-Q plot, the plot for the regression suggests that there is a non-normal residual distribution. This would be problematic as this breaks the assumption that the error terms must have a normal distribution. The final plot to examine is the Scale-Location plot. While the plot does appear to be somewhat skewed to the left it doesn't appear to be following a particular pattern suggesting that there is not any present heteroskedasticity.

Another area of testing for this model is to perform cross validation for the second OLS regression model and to compare it with the first to determine how effectively the models are able to predict values not in their data set. To do this test, the data set has been split into two sections, the first is the train data set with 70% of the original data set with the remaining data being used as a test data set. The test and train data set are the same ones used to cross validate the other models in the report. The results from the RMSE suggest that the second model is more accurate as a predictor. This would mean that the second OLS model would be a better model to use to predict Product Related Duration. While the value of 1056.793 may seem like a large value, the range of values for product related duration spans from 0-63973.52 as a result an RMSE of 1056.793 could be seen as a relatively small value.

*Testing the model*

```
testlm2 <- predict(trainlm2, testReg)
lm2rmse <- rmse(actual = testReg$ProductRelated_Duration, predicted = testlm2)
print(lm2rmse)
```

```
## [1] 1056.793
```

*Testing the new model against the first model using the same train and test data*

```
trainlm1 <- lm(ProductRelated_Duration ~., data = trainReg)

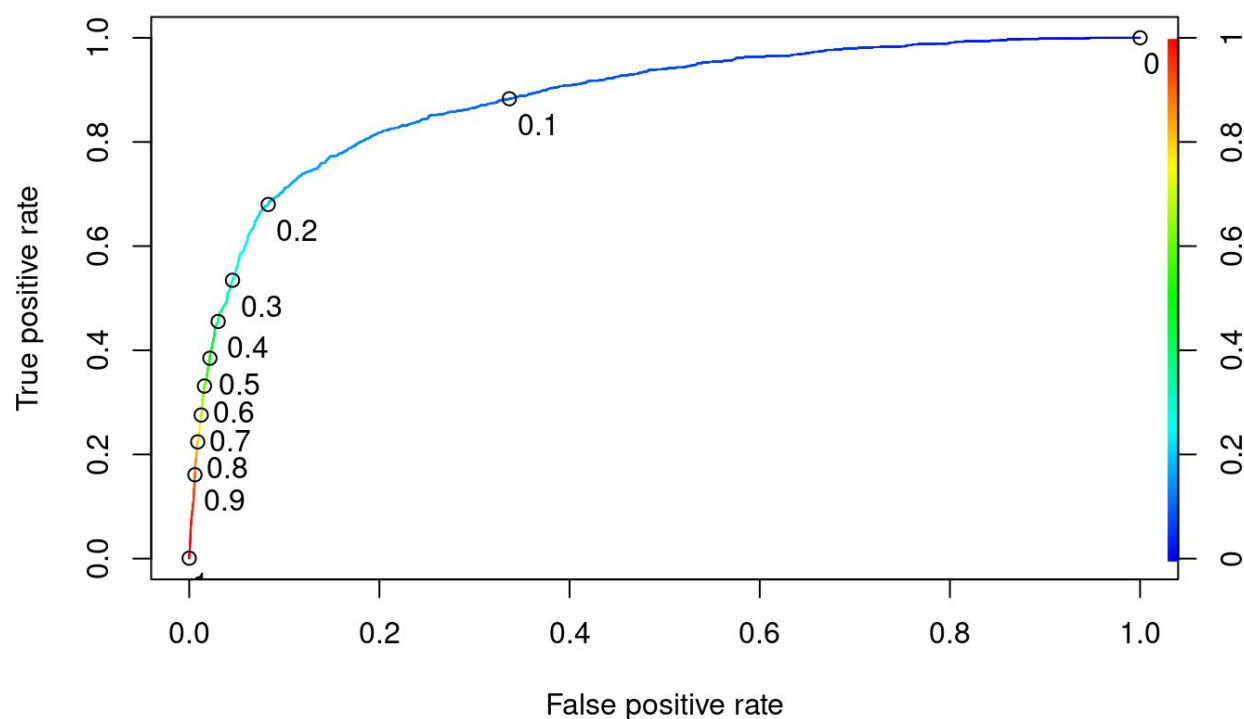
testlm1 <- predict(trainlm1, testReg)
lm1rmse <- rmse(actual = testReg$ProductRelated_Duration, predicted = testlm1)
print(lm1rmse)
```

```
## [1] 1066.44
```

**Image 8.** Root mean square error comparison for both linear models.

## Logistic regression - Generalized Linear Model (GLM):

In order to evaluate the model performance on the test dataset, first we had to select the appropriate threshold value, using the Receiver Operator Characteristic curve, i.e. ROC curve, Fig. 15.



**Figure 15.** ROC curve for the GLM model.

Selecting the threshold is a tradeoff between having an optimal true positive rate, while keeping the false positive rate low. In our case, we opted for a threshold value of 0.3, to optimize the overall accuracy of the model. Using that value we evaluated the model on the test dataset:

```
predictTest = predict(glm, type = "response", newdata = testReg)
```

```
table(testReg$Revenue_binary, predictTest >= 0.3)
```

and obtained the confusion matrix below.

```
##
##      FALSE TRUE
##    0   2959  168
##    1    275  297
```

Based on it we calculated the following model metrics:

Precision: 0.6387097

Recall: 0.5192308

Accuracy: 0.8802379

Balanced accuracy: 0.7327526

which suggests that the selected threshold was appropriate to achieve a reasonably good accuracy of 0.88.

Finally we estimated the root mean square error of the model predictions:

```
testReg = testReg %>%
  mutate(predictions_quad = predict(glm, testReg))

sqrt(testReg %>%
  summarise(RMSE_glm = mean((Revenue_binary-predictions_quad)^2)))
```

	RMSE_glm <dbl>
1 row	3.085286

**Image 9.** Estimated root mean square error of GLM

In sum, based on the evaluation, GLM model would be appropriate for the prediction of binary “Revenue” feature.

## 5. Conclusions/Recommendations

### Conclusions

As the E-commerce market substantially grows, the behavior patterns of online shoppers are an influential aspect for businesses. As a result, it would be valuable for businesses to understand what characteristics are most important in generating site traffic and revenue. Our group hypothesised that variables such as the time spent on certain pages, weekends, days around holidays, type of visitors, geographic location would be the most important variables in determining revenue and site traffic. In this report three models were presented, each designed to test this hypothesis. The models included: a decision tree, an

OLS linear regression and a logistic regression. Through the logistic regression and decision tree models, it has been determined that many of these variables where features affecting the purchasing behavior (i.e. Revenue feature) are Page value, Month (+December), Bounce Rates, Exit rate, Admin Duration, Product Related, Special Day and Visitor type. Through the linear regression, it was determined that in terms of time spent on product related pages Variables such as product related, visitor type, revenue, month, exit rate have a positive impact on the product related duration. It was also determined contrary to what may be expected weekends and proximity to holidays has a negative impact on product related duration. During the testing section, the decision tree model performed the best, with an accuracy of 90.3% and a precision of 71%, while the logistic regression tested satisfactorily. The linear regression however failed to meet all the assumptions needed for the OLS regression to be accurate.

With these results there are a number of implications that would be useful for online businesses. First implication is that correctly configured e-commerce metrics that track page value performance are essential to optimize website development in a way that maximizes purchasing decisions. As well, months and holidays play a statistically significant impact on revenue businesses can use these results as the optimal times for promotions and when the largest times of revenue generation. The statistical significance of revenue in the OLS suggested that those who purchased items also spent a longer time looking at the products pages.

## Recommendations

1. Increase the efficiency in the way that products are visualized and make the moving between different pages easier to encourage users to browse a wider range of products.
2. The delta frequency was greater in May and November, which shows potential of increase in sales. A steady market shows a well stabilized amount of sales and any noticeable frequency means there is room for improvement.
3. Obtaining live and real world data is a necessity for effectiveness. Having access to live data from the real-world e commerce platforms allows us to further analyze the market every day which will allow us to predict the market behaviour which will allow us to adapt to maximize sales.

## References

Sakar, C.O., Polat, S.O., Katircioglu, M. and Kastro, Y., 2018sup. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10), pp.6893-6908.

Statistics Canada (2018). Online shopping in Canada, 2018.

<https://www150.statcan.gc.ca/n1/pub/89-28-0001/2018001/article/00016-eng.htm> (2018)

Jason Aston, Owen Vipond, Kyle Virgin, Omar Youssouf. Statistics Canada (2020). Retail e-commerce and COVID-19: How online shopping opened doors while many were closing.

<https://www150.statcan.gc.ca/n1/pub/45-28-0001/2020001/article/00064-eng.htm>

Susan Krashinsky Robertson. The Globe and Mail (2020). The post-lockdown consumer: How the global crisis has changed Canadians' spending patterns.

<https://www.theglobeandmail.com/business/article-the-post-lockdown-consumer-how-the-global-crisis-has-changed/>

Shopify (2020). Shopify Announces First-Quarter 2020 Financial Results

[https://s23.q4cdn.com/550512644/files/doc\\_financials/2020/Q1/Press-Release-Q1-2020.pdf](https://s23.q4cdn.com/550512644/files/doc_financials/2020/Q1/Press-Release-Q1-2020.pdf)

## APPENDIX

**Appendix table 1:** OLS month to number table

Month	Number
February	1
March	2



May	3
June	4
July	5
August	6
September	7
October	8
November	9
December	10

**Appendix table 2:** Cross-correlation matrix of all the dataset features

##	Administrative	Administrative_Duration	Informational
## Administrative	1.000000000	0.601583342	0.376850429
## Administrative_Duration	0.601583342	1.000000000	0.302709709
## Informational	0.376850429	0.302709709	1.000000000
## Informational_Duration	0.255848140	0.238030789	0.618954862
## ProductRelated	0.431119340	0.289086621	0.374164291
## ProductRelated_Duration	0.373939013	0.355421954	0.387505306
## BounceRates	-0.223562630	-0.144170410	-0.116113616
## ExitRates	-0.316482998	-0.205797757	-0.163666061
## PageValues	0.098989585	0.067608481	0.048631692
## SpecialDay	-0.094777598	-0.073303725	-0.048219254
## Month	0.048560251	0.029061426	0.019742688
## OperatingSystems	-0.006347063	-0.007343418	-0.009526668
## Browser	-0.025034572	-0.015391527	-0.038234678
## Region	-0.005486805	-0.005560563	-0.029168638
## TrafficType	-0.033560713	-0.014376431	-0.034490754
## VisitorType	-0.025819710	-0.023939717	0.055827573
## Revenue_binary	0.138917094	0.093586719	0.095200343
## Weekend_binary	0.026416750	0.014990142	0.035784725
##	Informational_Duration	ProductRelated	
## Administrative	0.255848140	0.431119340	
## Administrative_Duration	0.238030789	0.289086621	
## Informational	0.618954862	0.374164291	
## Informational_Duration	1.000000000	0.280046268	
## ProductRelated	0.280046268	1.000000000	
## ProductRelated_Duration	0.347363577	0.860926836	
## BounceRates	-0.074066610	-0.204577633	
## ExitRates	-0.105275683	-0.292526283	
## PageValues	0.030860874	0.056281794	
## SpecialDay	-0.030576549	-0.023958175	
## Month	0.005987214	0.070298510	
## OperatingSystems	-0.009578676	0.004289621	
## Browser	-0.019284981	-0.013145721	
## Region	-0.027144112	-0.038121842	
## TrafficType	-0.024674908	-0.043064304	
## VisitorType	0.044676760	0.126655811	
## Revenue_binary	0.070344502	0.158537984	
## Weekend_binary	0.024078486	0.016091964	

##	ProductRelated_Duration	BounceRates	ExitRates	
## Administrative	0.373939013	-0.223562630	-0.316482998	
## Administrative_Duration	0.355421954	-0.144170410	-0.205797757	
## Informational	0.387505306	-0.116113616	-0.163666061	
## Informational_Duration	0.347363577	-0.074066610	-0.105275683	
## ProductRelated	0.860926836	-0.204577633	-0.292526283	
## ProductRelated_Duration	1.000000000	-0.184541115	-0.251984097	
## BounceRates	-0.184541115	1.000000000	0.913004396	
## ExitRates	-0.251984097	0.913004396	1.000000000	
## PageValues	0.052823063	-0.119386026	-0.174498310	
## SpecialDay	-0.036379845	0.072702253	0.102241802	
## Month	0.061185682	-0.023762666	-0.039049283	
## OperatingSystems	0.002975790	0.023823182	0.014566735	
## Browser	-0.007380440	-0.015772209	-0.004442355	
## Region	-0.033090520	-0.006485347	-0.008907006	
## TrafficType	-0.036377170	0.078285541	0.078616331	
## VisitorType	0.119329172	0.135536393	0.179143931	
## Revenue_binary	0.152372611	-0.150672912	-0.207071082	
## Weekend_binary	0.007310614	-0.046513997	-0.062587048	
##	PageValues	SpecialDay	Month	OperatingSystems
## Administrative	0.09898959	-0.094777598	0.048560251	-0.0063470633
## Administrative_Duration	0.06760848	-0.073303725	0.029061426	-0.0073434175
## Informational	0.04863169	-0.048219254	0.019742688	-0.0095266679
## Informational_Duration	0.03086087	-0.030576549	0.005987214	-0.0095786764
## ProductRelated	0.05628179	-0.023958175	0.070298510	0.0042896206
## ProductRelated_Duration	0.05282306	-0.036379845	0.061185682	0.0029757898
## BounceRates	-0.11938603	0.072702253	-0.023762666	0.0238231825
## ExitRates	-0.17449831	0.102241802	-0.039049283	0.0145667353
## PageValues	1.000000000	-0.063541272	0.021780268	0.0185079466
## SpecialDay	-0.06354127	1.000000000	0.079341098	0.0126522347
## Month	0.02178027	0.079341098	1.000000000	-0.0295799600
## OperatingSystems	0.01850795	0.012652235	-0.029579960	1.0000000000
## Browser	0.04559192	0.003498747	-0.045913324	0.2230128882
## Region	0.01131530	-0.016097975	-0.032530328	0.0767754856
## TrafficType	0.01253169	0.052301443	0.041839131	0.1891536121
## VisitorType	-0.11122783	0.085556612	0.026481310	0.0015042220
## Revenue_binary	0.49256930	-0.082304598	0.080150468	-0.0146675596
## Weekend_binary	0.01200164	-0.016767155	0.029131513	0.0002842506

##	Browser	Region	TrafficType	VisitorType
## Administrative	-0.025034572	-0.0054868053	-0.033560713	-0.025819710
## Administrative_Duration	-0.015391527	-0.0055605628	-0.014376431	-0.023939717
## Informational	-0.038234678	-0.0291686379	-0.034490754	0.055827573
## Informational_Duration	-0.019284981	-0.0271441124	-0.024674908	0.044676760
## ProductRelated	-0.013145721	-0.0381218417	-0.043064304	0.126655811
## ProductRelated_Duration	-0.007380440	-0.0330905198	-0.036377170	0.119329172
## BounceRates	-0.015772209	-0.0064853474	0.078285541	0.135536393
## ExitRates	-0.004442355	-0.0089070060	0.078616331	0.179143931
## PageValues	0.045591919	0.0113152995	0.012531693	-0.111227826
## SpecialDay	0.003498747	-0.0160979746	0.052301443	0.085556612
## Month	-0.045913324	-0.0325303281	0.041839131	0.026481310
## OperatingSystems	0.223012888	0.0767754856	0.189153612	0.001504222
## Browser	1.000000000	0.0973928492	0.111938224	-0.021866988
## Region	0.097392849	1.000000000	0.047520231	-0.036190794
## TrafficType	0.111938224	0.047520231	1.000000000	-0.002839178
## VisitorType	-0.021866988	-0.0361907939	-0.002839178	1.000000000
## Revenue_binary	0.023984289	-0.0115950678	-0.005112971	-0.104725722
## Weekend_binary	-0.040260864	-0.0006906703	-0.002221229	-0.043679249
##	Revenue_binary	Weekend_binary		
## Administrative	0.138917094	0.0264167503		
## Administrative_Duration	0.093586719	0.0149901419		
## Informational	0.095200343	0.0357847251		
## Informational_Duration	0.070344502	0.0240784862		
## ProductRelated	0.158537984	0.0160919642		
## ProductRelated_Duration	0.152372611	0.0073106138		
## BounceRates	-0.150672912	-0.0465139965		
## ExitRates	-0.207071082	-0.0625870480		
## PageValues	0.492569295	0.0120016392		
## SpecialDay	-0.082304598	-0.0167671553		
## Month	0.080150468	0.0291315131		
## OperatingSystems	-0.014667560	0.0002842506		
## Browser	0.023984289	-0.0402608638		
## Region	-0.011595068	-0.0006906703		
## TrafficType	-0.005112971	-0.0022212292		
## VisitorType	-0.104725722	-0.0436792493		
## Revenue_binary	1.000000000	0.0292953680		
## Weekend_binary	0.029295368	1.000000000		