CDSA1010 Final Project, Section 4, Group 3

Yrysguli Kaireden, Amina Sheik-Ahmed, Lily Ye

10/05/2020

Abstract: In this assignment, the dataset is a user data from an e-commerce website with feature vectors for 12,330 individual online sessions. The goal of this assignment is to develop a model that can effectively predict a shopper's purchasing intention to finalize a transaction. The data is carefully explored, adjusted and transformed prior to model deployment. The methods used to explore the feasibility of such prediction are clustering and classification models, in which logistic regression, decision tree, random forest, k-NN, PCA, K-means clustering, PAM and hierarchical clustering are used. It is important to note that clustering models added the extra feature in researching the user base. Next, the prediction models were evaluated using precision, recall, overall accuracy, balanced accuracy, specificity, F1 and AUC while internal and external validations were conducted on clustering models. The models generated from this dataset and project can be deployed to be used by e-commerce businesses to 1) improve the prediction of their website's likelihood in yielding a purchase and 2) improve website metrics.

Introduction

Online shopping in Canada alone generated 39.9 billion dollars in 2019, this value is twice what Canadians spent on online shopping in 2014 (Sheldon et al., 2014). It is no surprise that online shopping is the leading online activity across the world. Given this massive opening in the online market especially in recent years, many businesses have made the shift to move their business online. This shift is attributed to website visitor's preference in making purchases online as this experience is more convenient than frequenting physical stores. This shift in website visitor behaviour is excellent for online businesses as this generates user-data which can be used in e-commerce to better understand their users and therefore make better strategic decisions that drive the business to cater to user's wants. Furthermore, e-commerce businesses can use this data to reduce costs, create cost-effective ways to sell their products and services, and most importantly improve user experience. While online shopping has allowed businesses to gain insight from user data to improve operational strategies, it is best that businesses employ early detection and behavioural prediction systems that mimic the behaviour of virtual sales-person so as to improve conversion rates (Sakar et al., 2019).

Background

The dataset used in this assignment was retrieved from UCI machine learning repository. The data was collected from user data from an online store in which metrics were measured by Google Analytics for each page visited in the e-commerce site. The data consists of feature vectors for 12,330 online sessions in which each session is specific to a user in a one-year period (Sakar et al., 2019). The features in this dataset are numerical and categorical which are used for the purchasing intention prediction. This dataset contains 10 numerical and 8 categorical attributes; these features will be further defined in the later data understanding section as per the description in Sakar et al., 2019 reference paper.
Objective

The objective of this study is to provide a classification algorithm that can effectively predict the users' purchasing intentions in real time, and also to provide a clustering algorithm that conducts an in-depth

research into the online users' behaviours and trends. The target variable "Revenue" is binary and coded as 'TRUE' or "FALSE', and this value pertains to the users' intention to finalize a transaction. The predictive models that will be deployed are Logistic regression, Random Forest, Decision Tree, and k-NN; while clustering models deployed are K-means, PAM and Hierarchical Clustering. Next, the prediction models were evaluated using precision, recall, overall accuracy, balanced accuracy, specificity, F1 and AUC while the clustering models were evaluated using the silhouette method and Rand Index.

In the next sections outliers are identified; following this, features transformation and selection steps are described to demonstrate how the final set of features were formed. The final dataset is then passed to predictive models and clustering models in which the best performing model is selected based on the set of evaluation metrics.

Business & Analytical Problem Framing

The dataset entails great information and insight on a number of factors, such as the number of web pages visited by users, browser or computer operating system used by site-visitors, and the month of the year when transactions took place. With this information, the ultimate business goal is to determine whether the website visitors will make a purchase, and also to better understand the website visitors base through research.

Therefore, in the context of online shoppers dataset, the following business problem statement is formulated: how can e-commerce business owners increase their revenue from the purchases made in their shopping website?

Then, from this business question, there would be the following analytical questions:

1. which machine learning models can best predict the purchase status?
2. what are the most important factors that influence the users likelihood of making a purchase, and
3. how are different features represented across the clusters of site-visitors?

This project will solve these analytical questions through classification (predictive) and clustering models. Then, the insights gleaned from the models and the analytical questions will provide a set of recommendations that can help address the business problem.

Data

Data Understanding

The dataset contains 12,330 observations and 18 variables.

```
## '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                  : chr  "Feb" "Feb" "Feb" "Feb" ...
```

```
##  $ 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            : chr  "Returning_Visitor" "Returning_Visitor" "Returning_Visitor" "Return
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

The first 10 columns of this dataset contain the numerical features used in the users behaviour analysis. The 'Administrative' column entails the number of pages visited by the user regarding their account management. Following, the 'Administration duration' column denotes the amount of time spent in seconds by the user exploring account management related pages. The 'Informational' column represents the number of pages visited by the user pertaining to website, communication and address information. The 'Informational duration' column shows the total amount spent by the user in seconds on informational pages. Next, the 'Product-related' column demonstrates the number of pages visited by visitors about product related pages. The total amount of time (in seconds) spent by the visitor on product related pages is represented by the 'Product-related duration' column. The 'Bounce rate' is measured by Google Analytics to show the average rate in which visitors enter a page and leave without visiting another page in the same website (Google Analytics, 2020a).

The 'Exit rate' column is also measured by Google analytics and shows the percentage of pages viewed compared to the page that was in the last session before exiting (Google Analytics, 2020b). Following, The 'Page value' column, which is measured by Google Analytics represents the average value of the pages visited by the visitor before completing a purchase. Lastly, the 'Special day' column shows the closeness of the user's visiting time to a special day such as Christmas, Mother's day in which a visit is finalized to a purchase. The numerical value in this column is measured by operational dynamics of the online business, such as the time between when the order is placed and received. These values have a minimum value of "0" and max value of "1", wherein "0" represents before and after the special date and non-zero values are assigned to dates very close to the special day.

Columns 11 - 18 contain the categorical features used in the users' behaviour analysis. The 'Operating Systems' column denotes the operating system used by the visitor. Next, the 'Browser' column describes the browser type of the user. The 'Region' consists of the geographic region in which the website visit originates. Following, the 'Traffic Type' column indicates the traffic source that leads the user to the website (text, advertisement, direct). The 'Visitor Type' column labels users as either a 'New Visitor', a 'Returning Visitor', or 'Other'. The 'Weekend' column consists of "TRUE" or "False" values indicating whether or not the website was visited during a weekend. Following, the 'Month' column indicates the month in which a user visits the website.

The last column is the target variable 'Revenue', indicating whether the website visit was finalized with a transaction, where TRUE means the visit yielded a transaction and FALSE means the visit did not result in a transaction.

Data Exploration

Correlation and Relationship of the Variables

Looking closely at the target variable, it is visible that there is a disportionately large incident of "FALSE" (10,422 incidents) compared to "TRUE" (1,908), wherein "FALSE" means the visit did not yield a transaction and "TRUE" means the visit yielded a transaction.

Furthermore, the table and graph below shows the exact percentage and count distribution of the target variable.

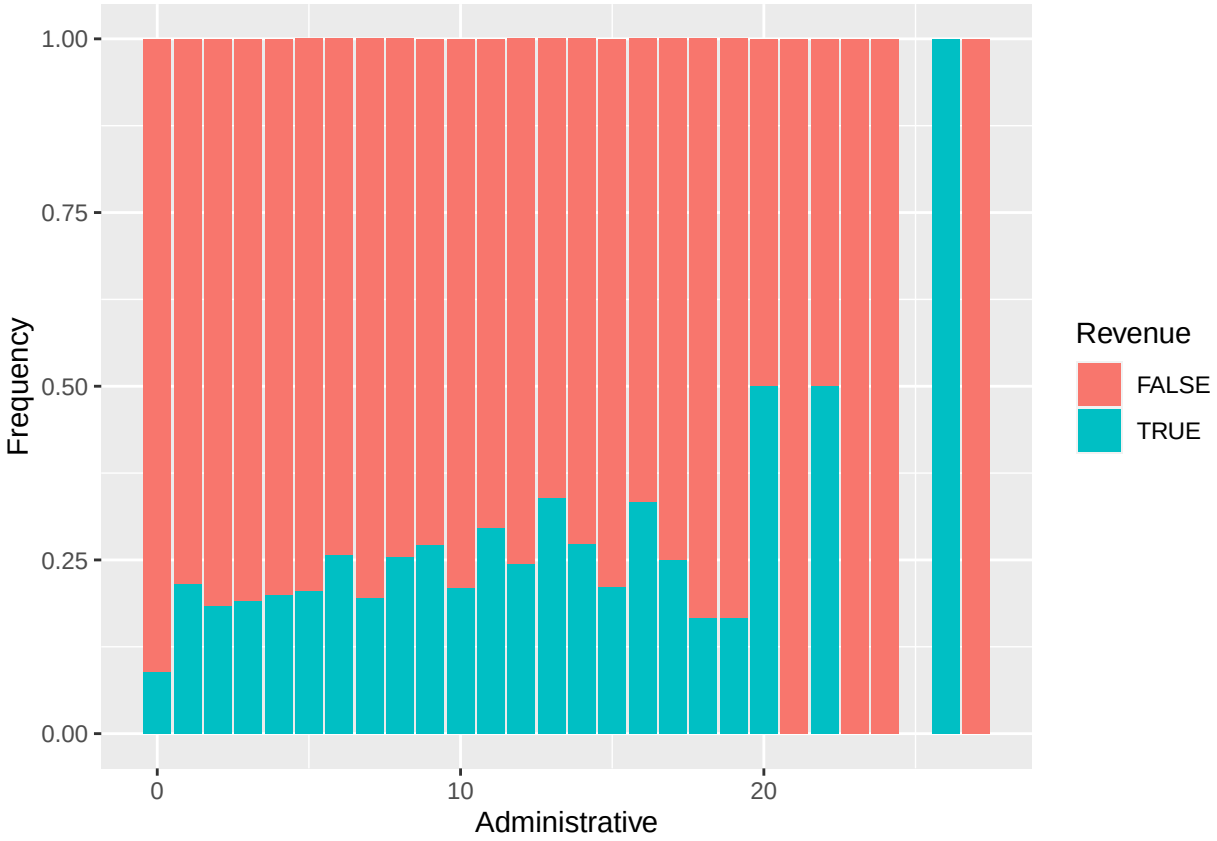
```
##
```

```
##     FALSE      TRUE
## 0.8452555 0.1547445
```

There is an 85% "FALSE" output and a 15% "TRUE" output; this demonstrates that the data is severely unbalanced and will be further discussed when splitting the dataset into test and train.

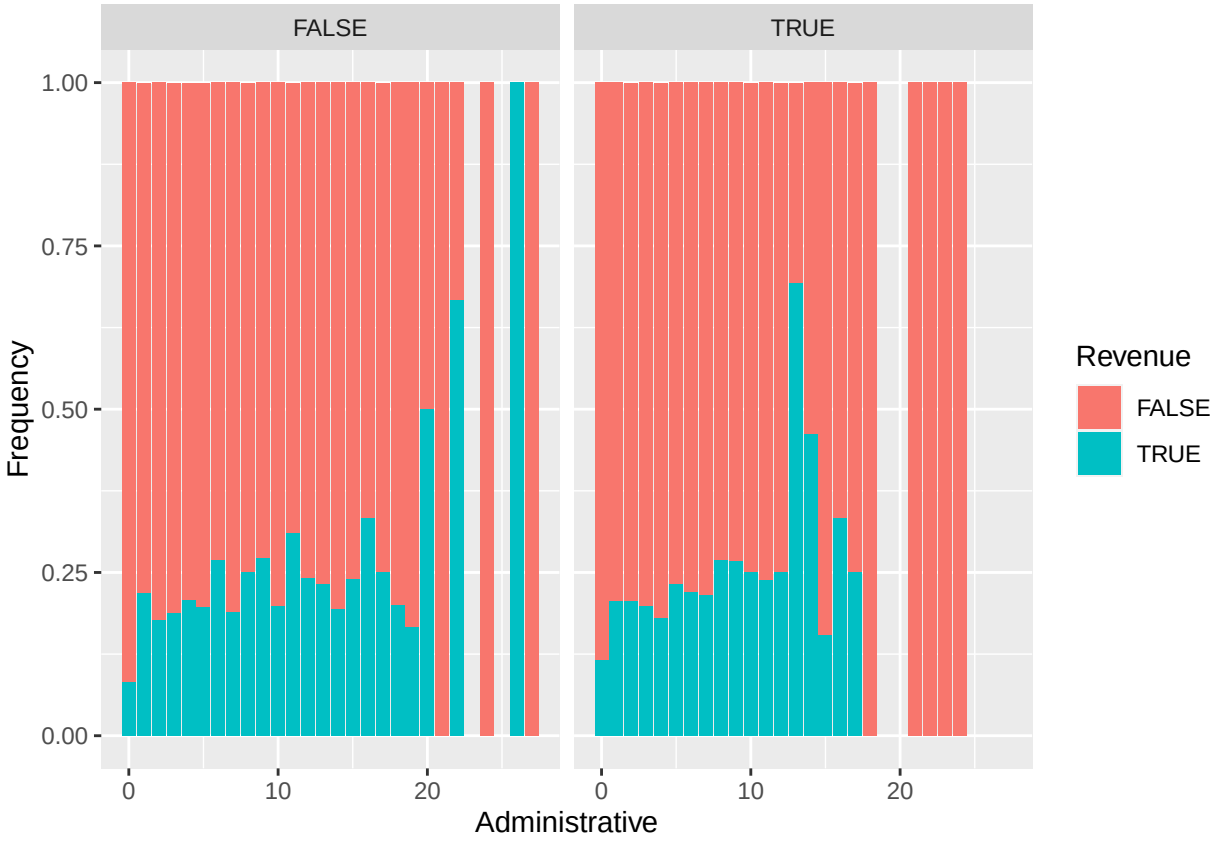
Next, specific features will be discussed in detail.

Relationship between Administrative and Revenue



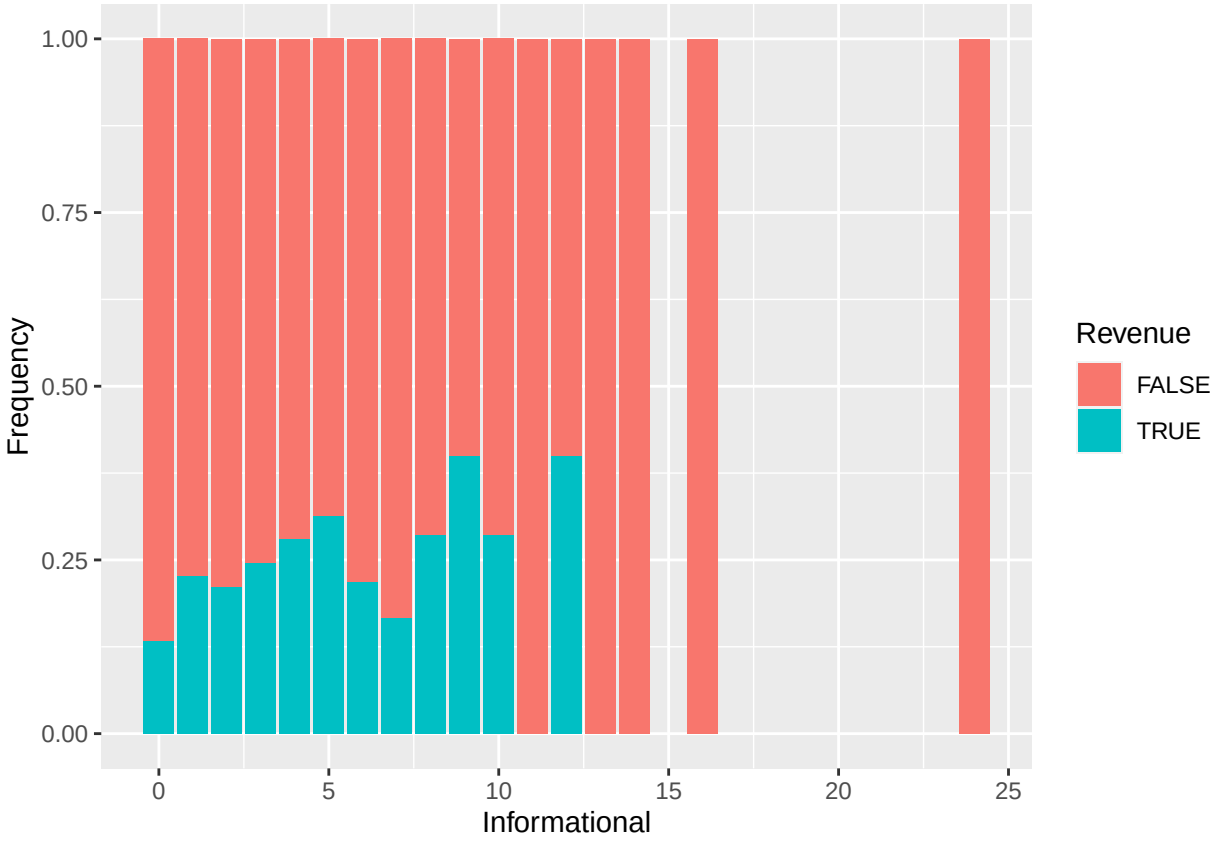
Starting with the administrative attribute, the correlation trend in this graph demonstrates that the users, who visit many pages regarding account management, are more likely to make a purchase than the users who visit less pages regarding account management. This sounds logical, since a frequent page visit to the account management or account setup is indicative of a user who is interested in making a purchase.

Relationship between Administrative and Revenue with respect to Weekend



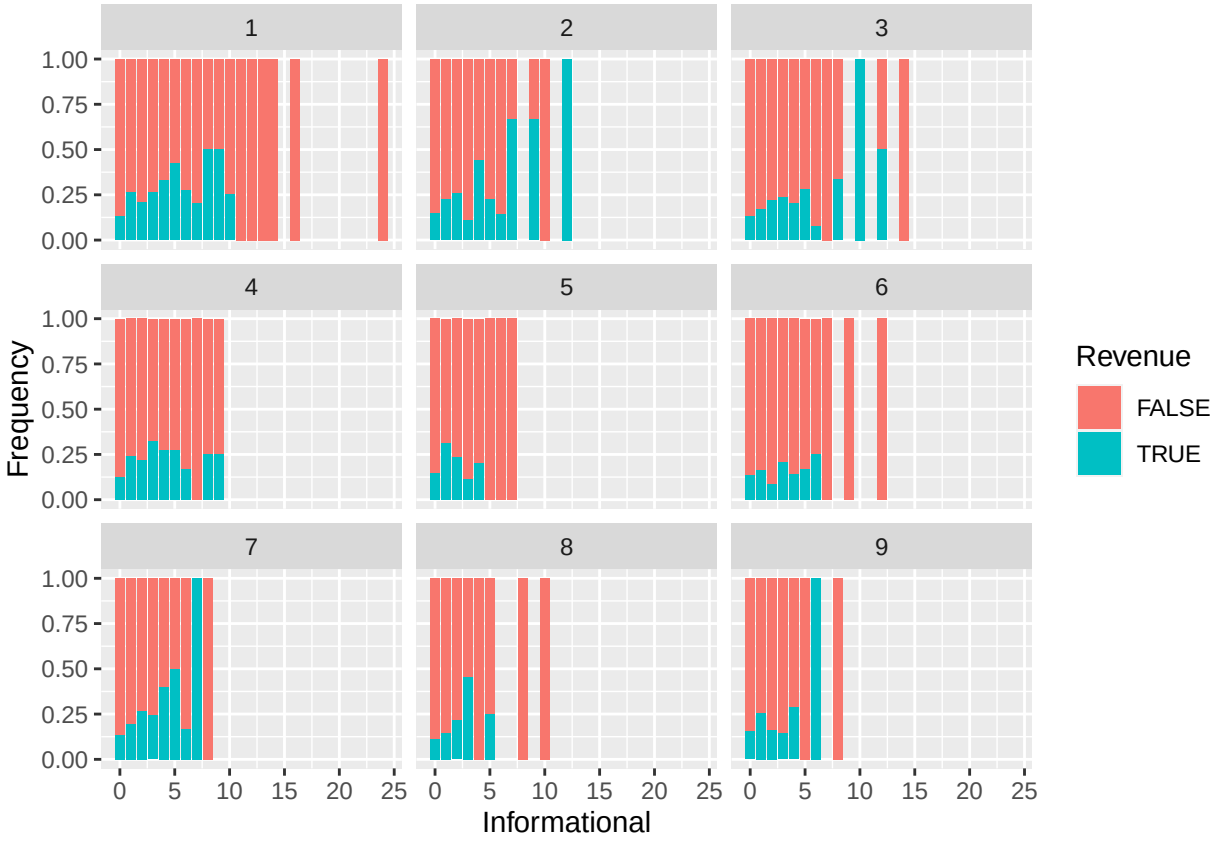
This graph looks at the correlation between the number of pages visited by the user related to account-management and revenue with respect to the weekend. It appears that users are likely to finalize a purchase the more account management related pages they visit. Furthermore, these finalized purchases are more likely to occur during the week rather than the weekend.

Relationship between Informational and Revenue



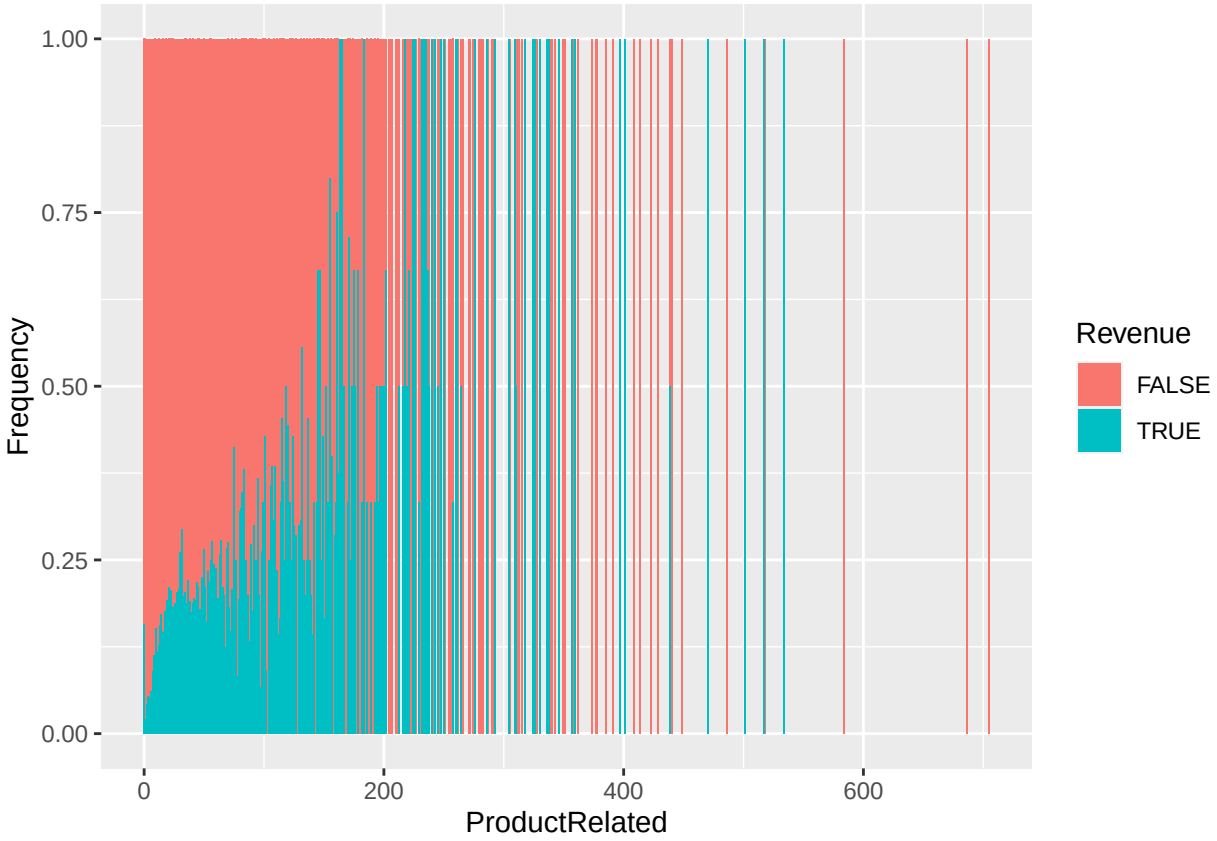
The correlation trend between the informational attribute and revenue suggests that users that frequently visit pages about the company's information, such as address, are likely to finalize a transaction.

Relationship between Informational and Revenue with respect to Region



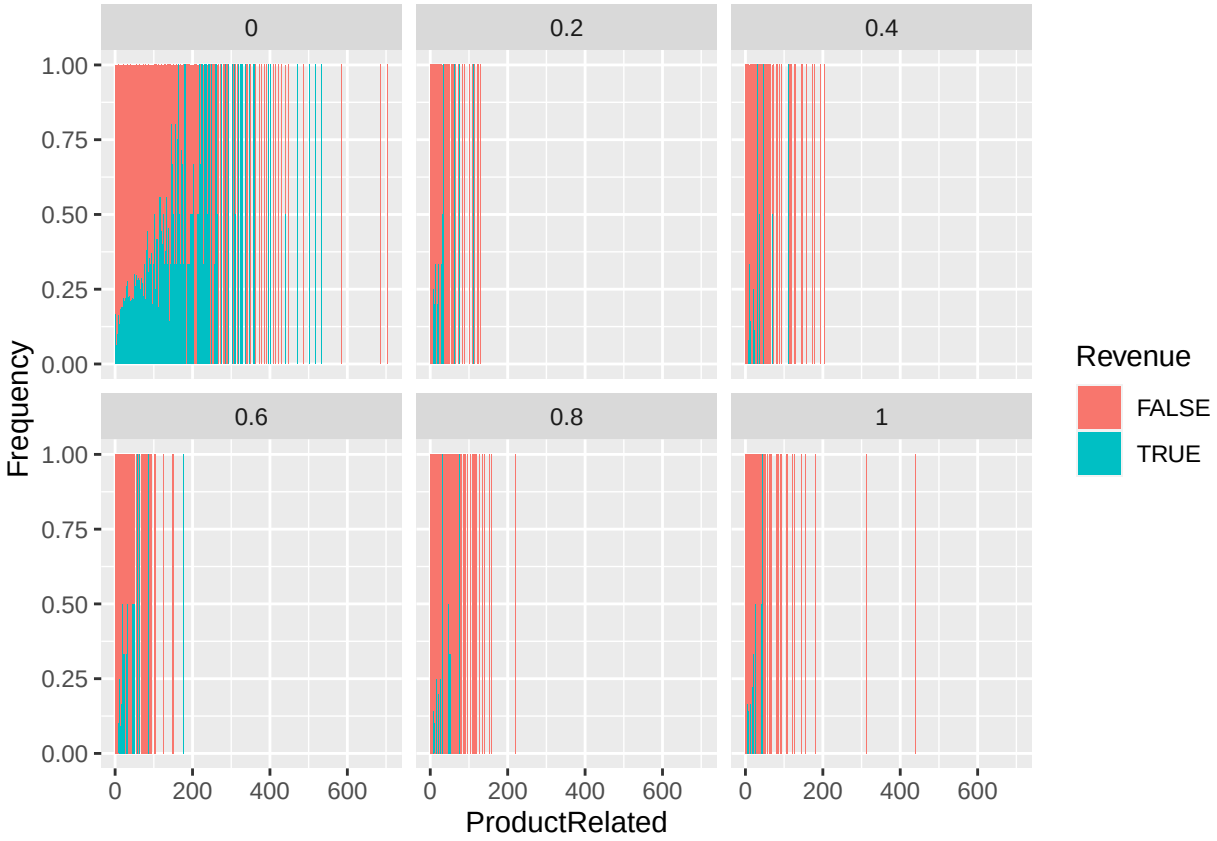
These graphs demonstrate the correlation of the number of pages users visit related to the company's information and the target variable (revenue) with respect to region. The numerical values of these regions were not defined in the reference article (Sakar et al., 2019). It appears that the more informational pages users visit, the greater the likelihood of the user finalizing a transaction. Furthermore, the different regions have varying shopping patterns, with regions 3, 7 and 9 having most purchases.

Relationship between Product-related and revenue



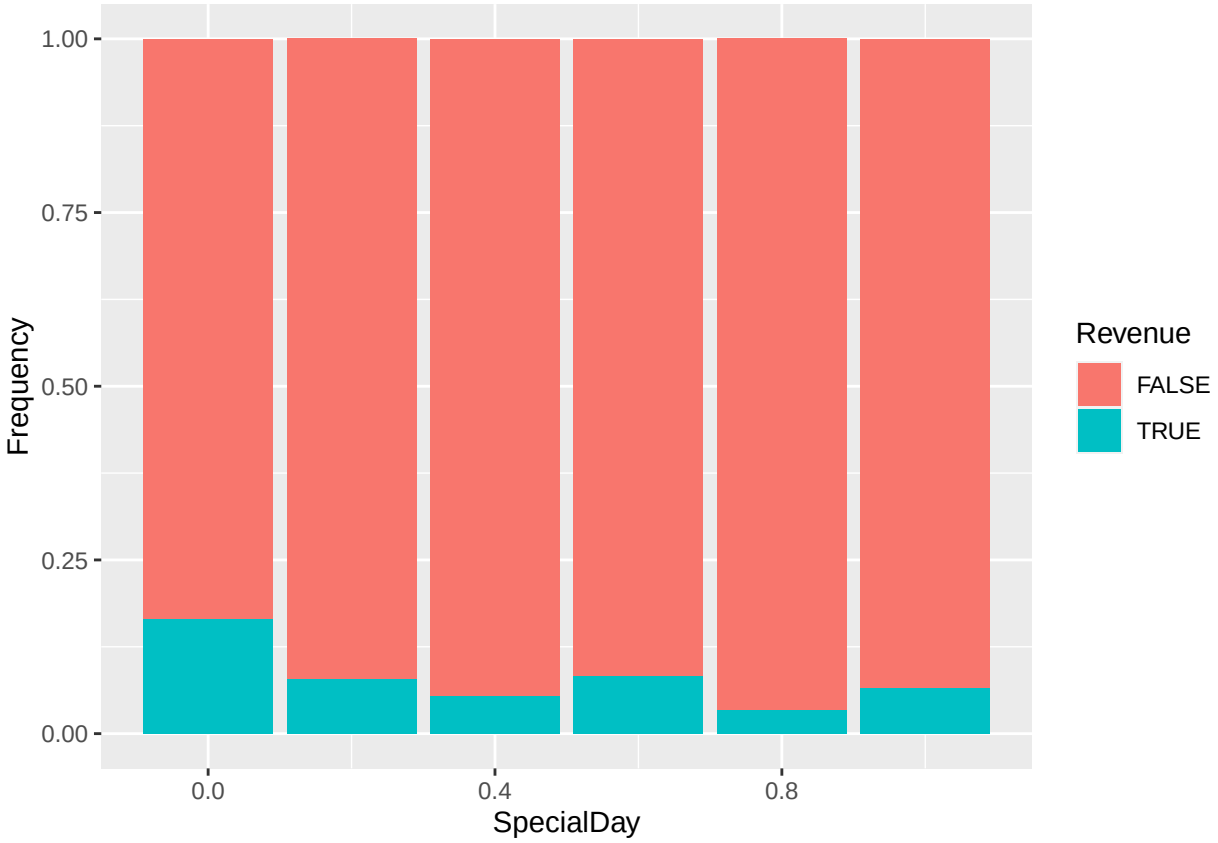
Looking at the product related attribute against the target variable, users that visit more product related pages are likely to finalize a purchase.

Relationship between ProductRelated and Revenue with respect to Special Day



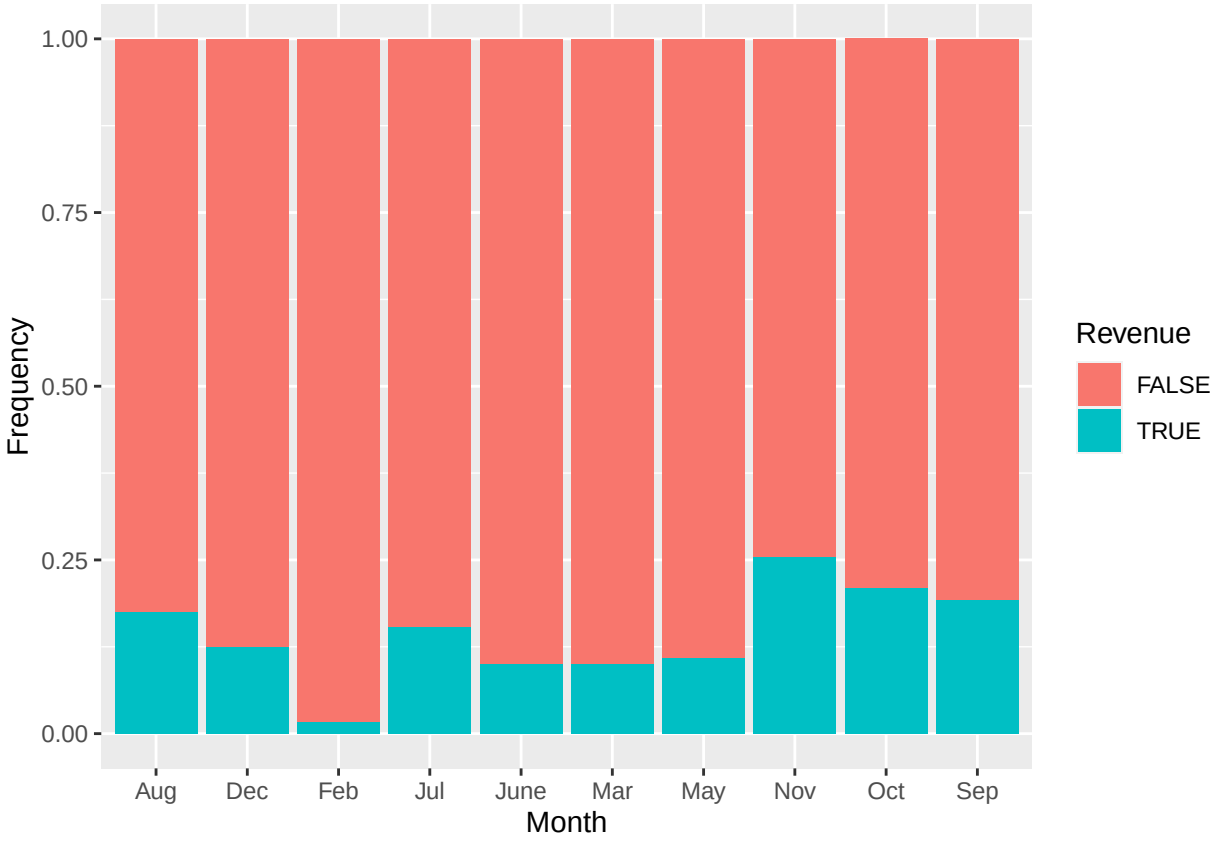
Visitors are more likely to finalize a purchase when they explore more product related pages, this correlation seems to be the strongest during regular days (far before or after holidays); this is represented in the graph with value "0".

Relationship between Special day and revenue



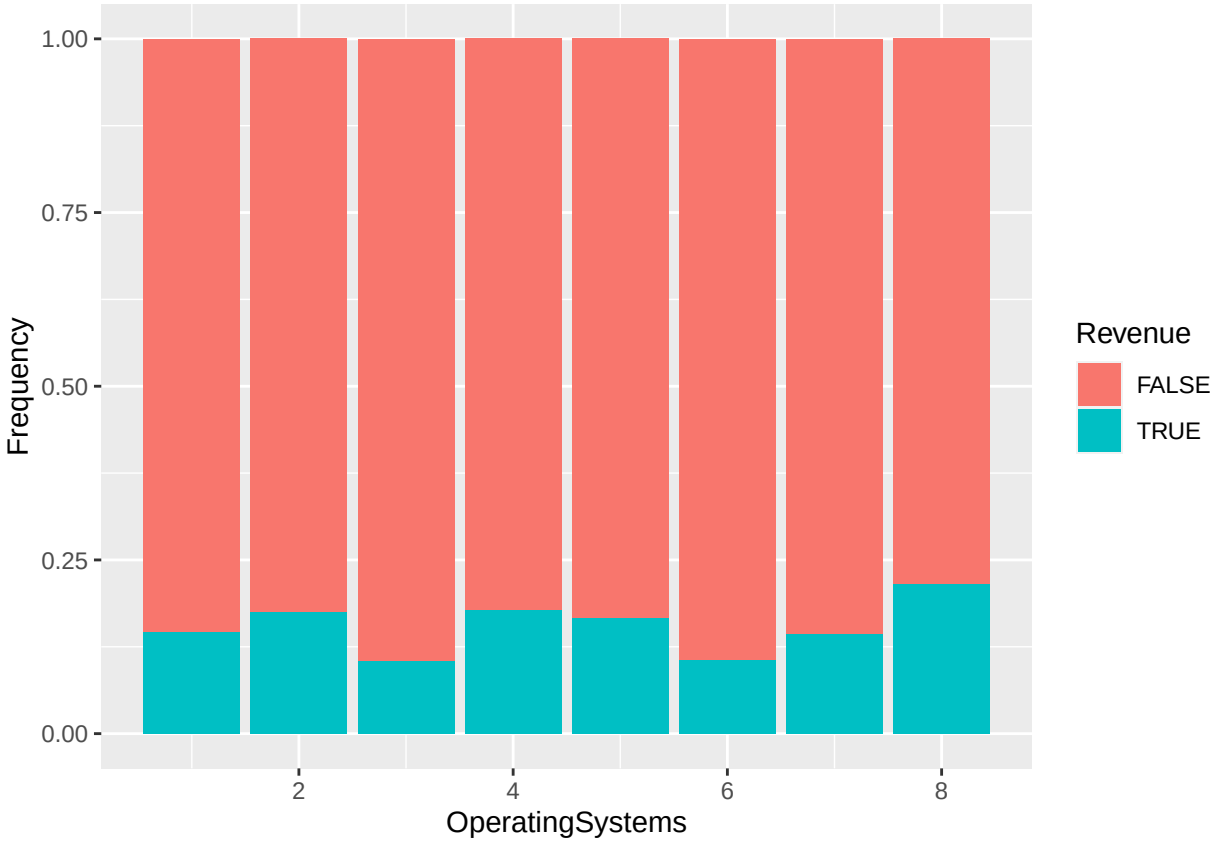
The correlation between the 'Special day' and the target variable is quite poor. Note that a "0" value is assigned to dates visited well before or after the special date, and non-zero values are assigned to the dates visited closest to the special day, wherein 1 would mean the user visited on a date very close to the special day. From this graph, one may infer the e-commerce website was visited the most on dates well before or after the special date (regular days of the year). Also, the portion of the users visiting during this time frame that did make a purchase, are greater than the portion of users making purchases closer to the special day. This might be explained by people taking into account delivery time and purchasing so that an item arrives exactly on that 'Special day'.

Relationship between Month and Revenue



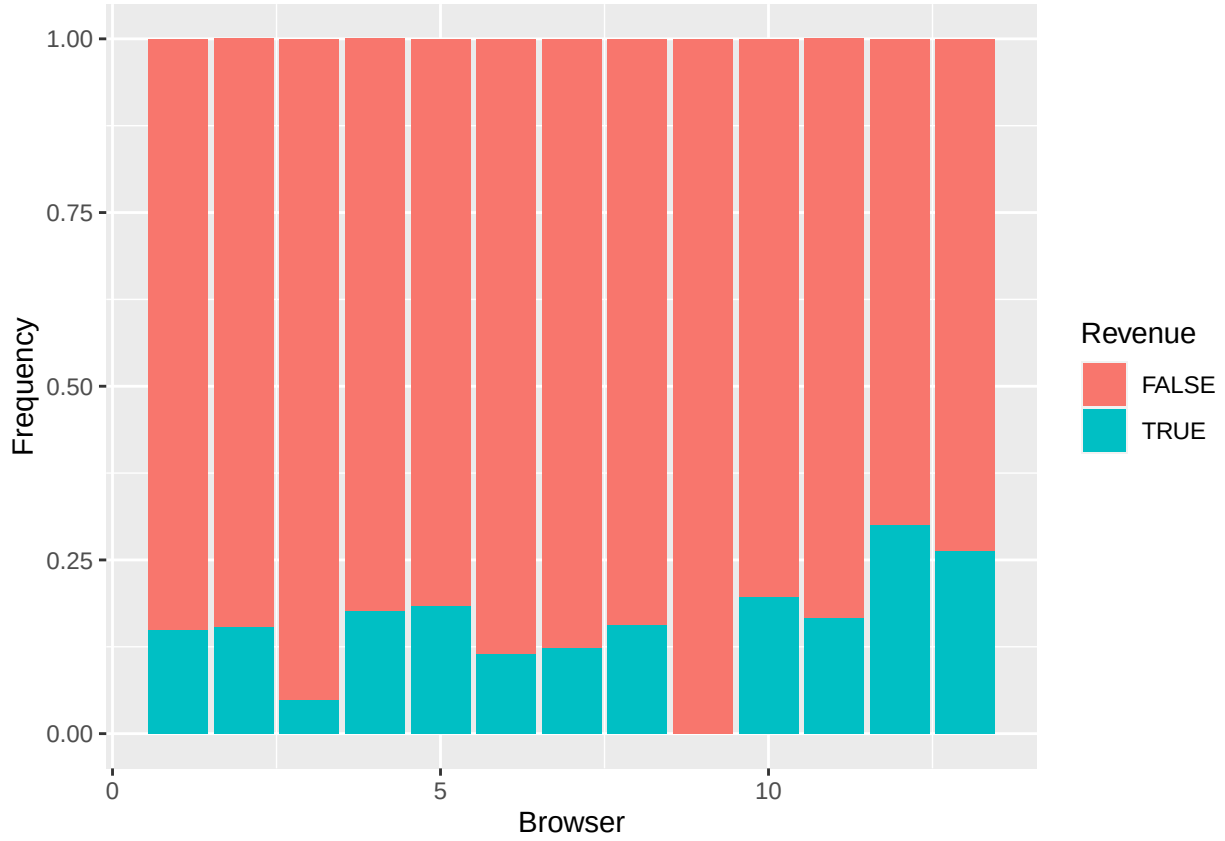
Looking at the month attribute, it appears that users visit the during or a slightly before high peak months more than the holidays and important official dates (i.e., back to school) and thereby make a purchase during these time frames. It looks like visitors tend to purchase during Early Bird Specials before holidays and Clearance after holidays effective.

Relationship between Operating Systems and Revenue



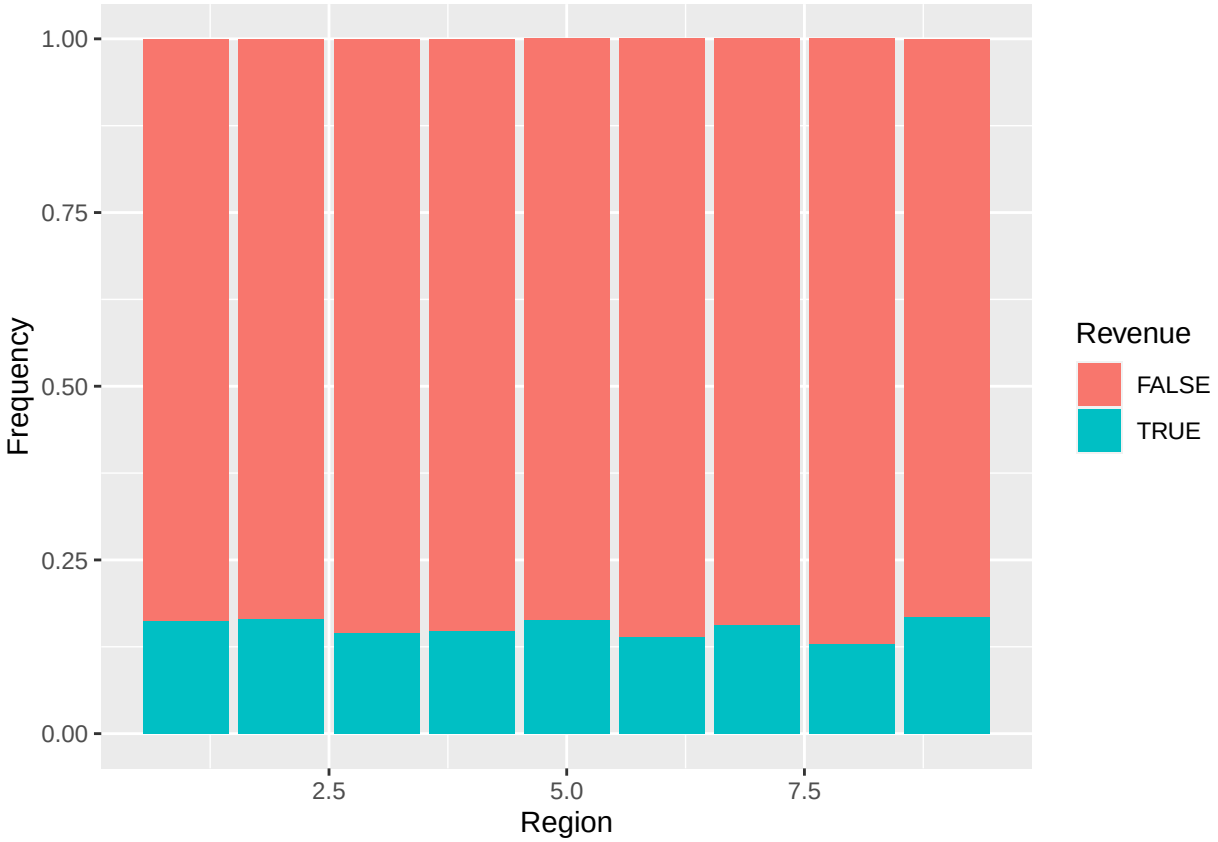
It appears that visitors using operating system "8" have a slightly higher likelihood of finalizing a purchase. There isn't a significant difference in purchasing behaviour and the remaining operating systems. There are 8 different operating systems, however these systems were not defined in the reference article (Sakar et al., 2018).

Relationship between Browser and Revenue



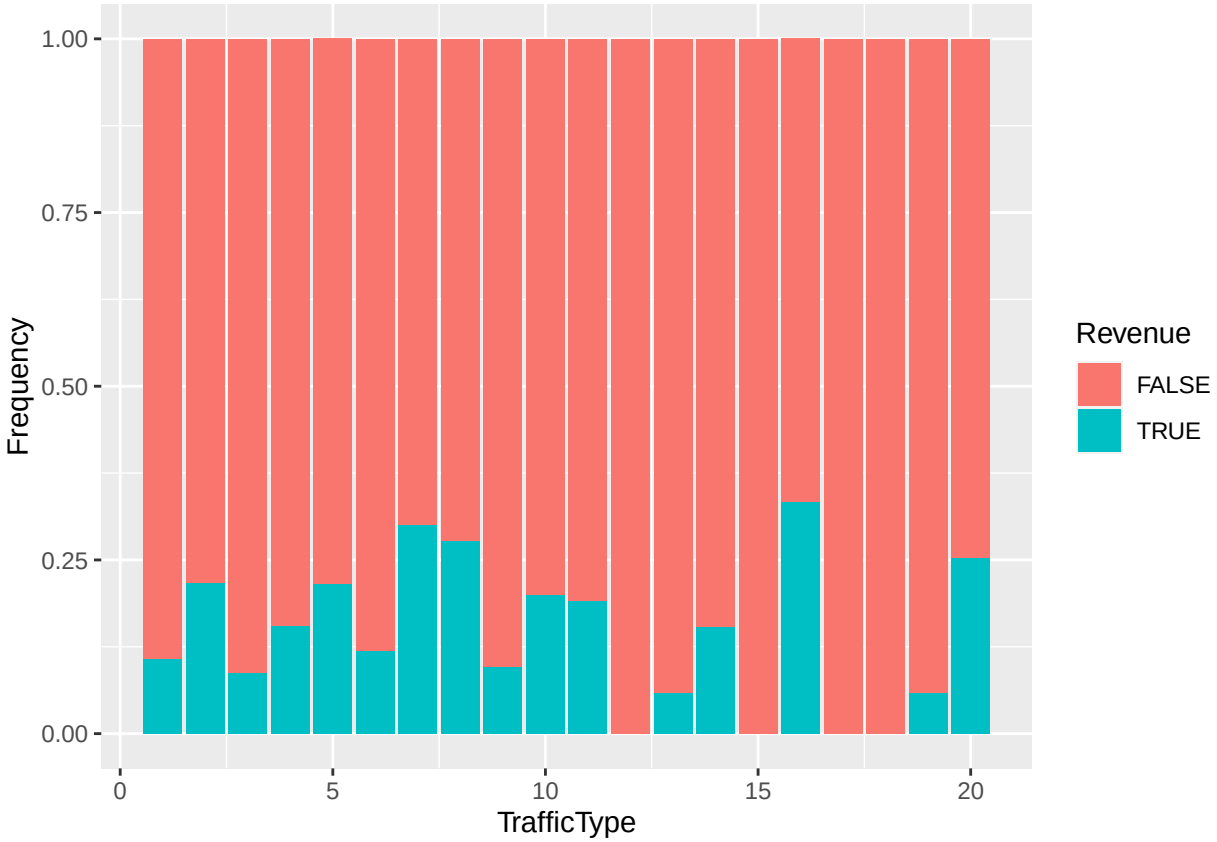
The plot between the user's browser type and the target variable shows the browser type '12' yields the most transactions. There are 13 different browser types; however, browser types were not defined in the reference article (Sakar et at., 2019).

Relationship between Region and Revenue



Users visiting from different server regions appear to show similar shopping behaviour. There are 9 different regions; however they were not defined in the reference article (Sakar et al., 2019).

Relationship between Traffic Type and Revenue

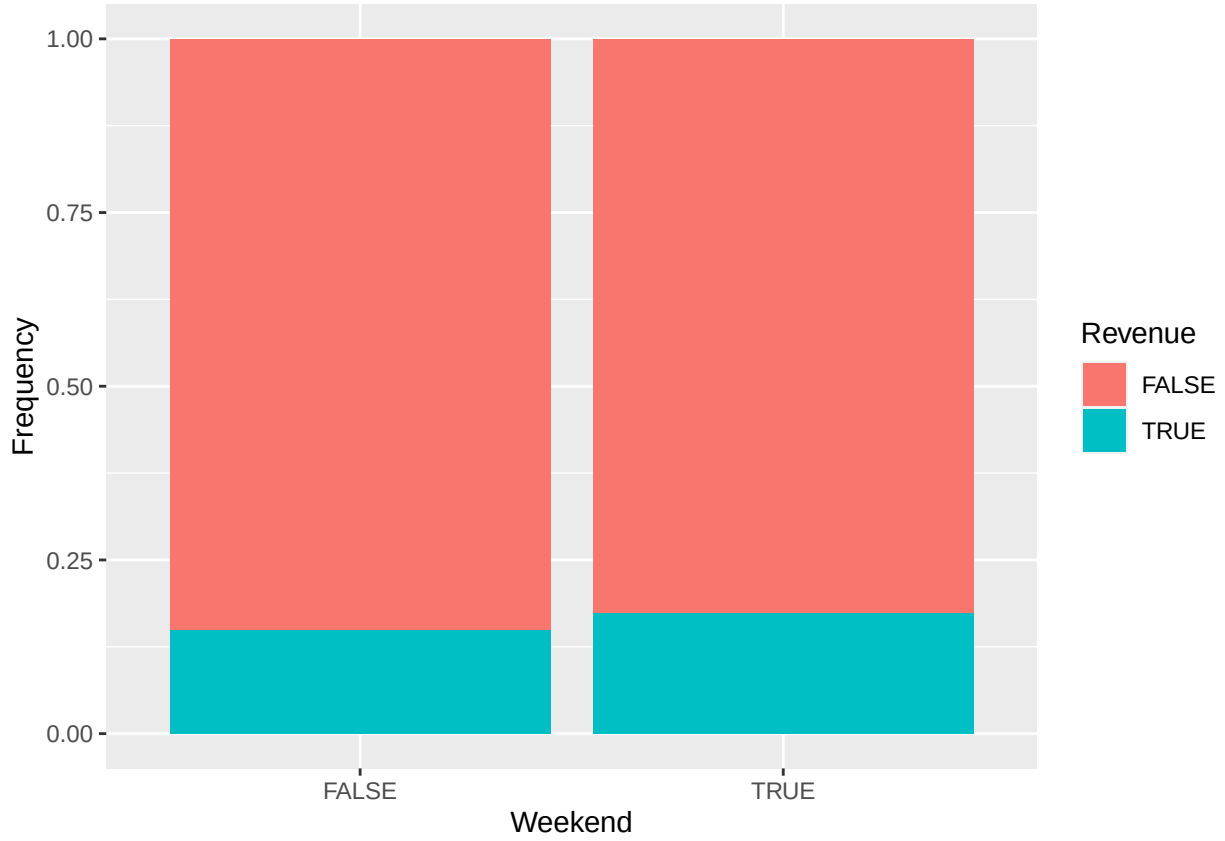


Specific traffic sources by which the visitor arrives from to the website yield different shopping behaviours. It seems that traffic type 16 yields the most buyers. There are 20 traffic sources in the dataset; however they are not defined in the reference paper (Sakar et al., 2019).

Relationship between Visitor Type and Revenue

New visitors compared to "Other" and "Returning visitors" are more likely to finalize a transaction. This could be due to discounts on first order that is usually offered by online stores.

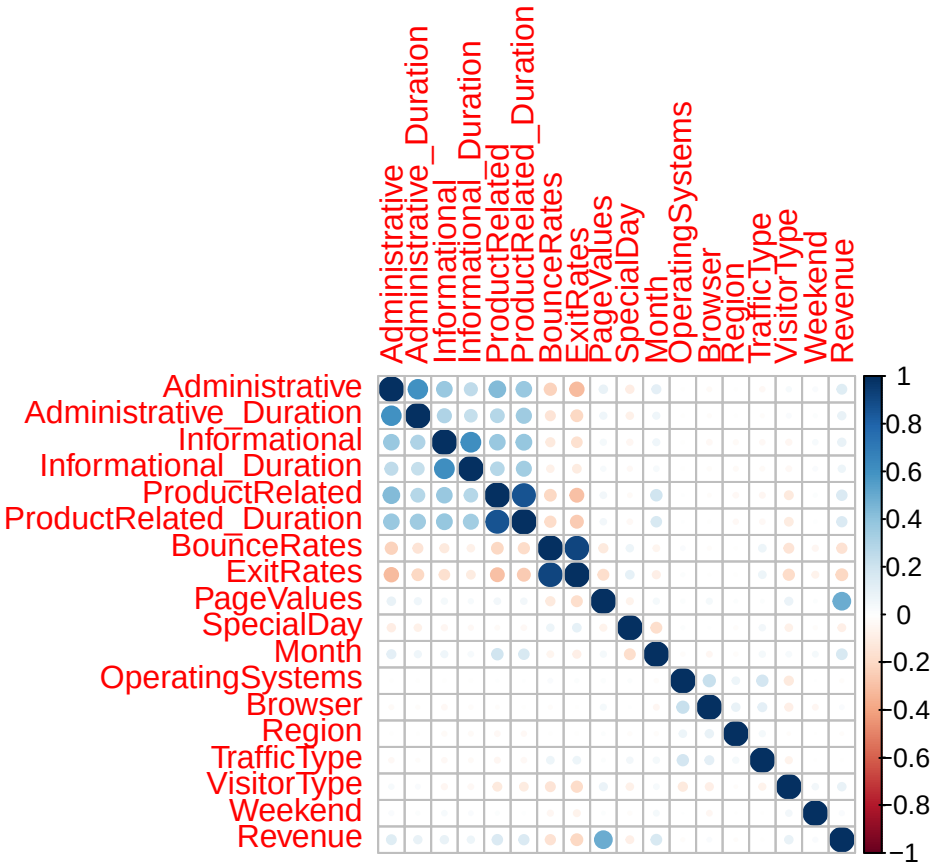
Relationship between Weekend and Revenue



The 'Weekend' column consists of "TRUE" or "False" values indicating whether or not the website was visited during a weekend. Although very close, the user visits that occur during the weekend (TRUE) are more likely to result in a transaction.

Correlation of features with each other

```
##          Administrative Administrative_Duration           Informational
##                      27                    3335                      17
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                    1258                     311                    9551
##              BounceRates              ExitRates             PageValues
##                    1872                    4777                    2704
##              SpecialDay                  Month       OperatingSystems
##                       6                     10                       8
##                 Browser                 Region            TrafficType
##                      13                      9                      20
##             VisitorType                Weekend                Revenue
##                       3                      2                       2

## corrplot 0.84 loaded
```



This graph depicts a strong correlation between bounce rates and exit rates. It is important to note that the 'Bounce rate' is the average rate in which visitors enter a page and leave without visiting another page, while The 'Exit rate' is the percentage of pages viewed compared to the page that was in the last session before exiting (Google Analytics, 2020b). Given that both features calculate the percentage of visitors that leave a page after entering it, this correlation is viable.

Positive correlations are observed between the 'Administrative' and 'Administrative duration' attributes. This correlation suggests that number pages users visit for account management is correlated to the time spent on account management pages. Specifically if a user visits fewer pages for account management purposes they are also spending a short amount of time on such pages.

Also, a positive correlation is observed between the informational and Informational duration attributes. This suggests that the number of pages users visit for the communication and address information of the shopping or service site is correlated with the total time the user spends on the aforementioned pages.

The feature with the strongest correlation to the target variable is 'Page value'. This feature is defined as the average value of the pages visited by the user before completing a purchase, in which the page value is defined by the following equation:

$$\frac{ECommerceRevenue + TotalGoalValue}{UniquePageviewsforGivenPage}$$

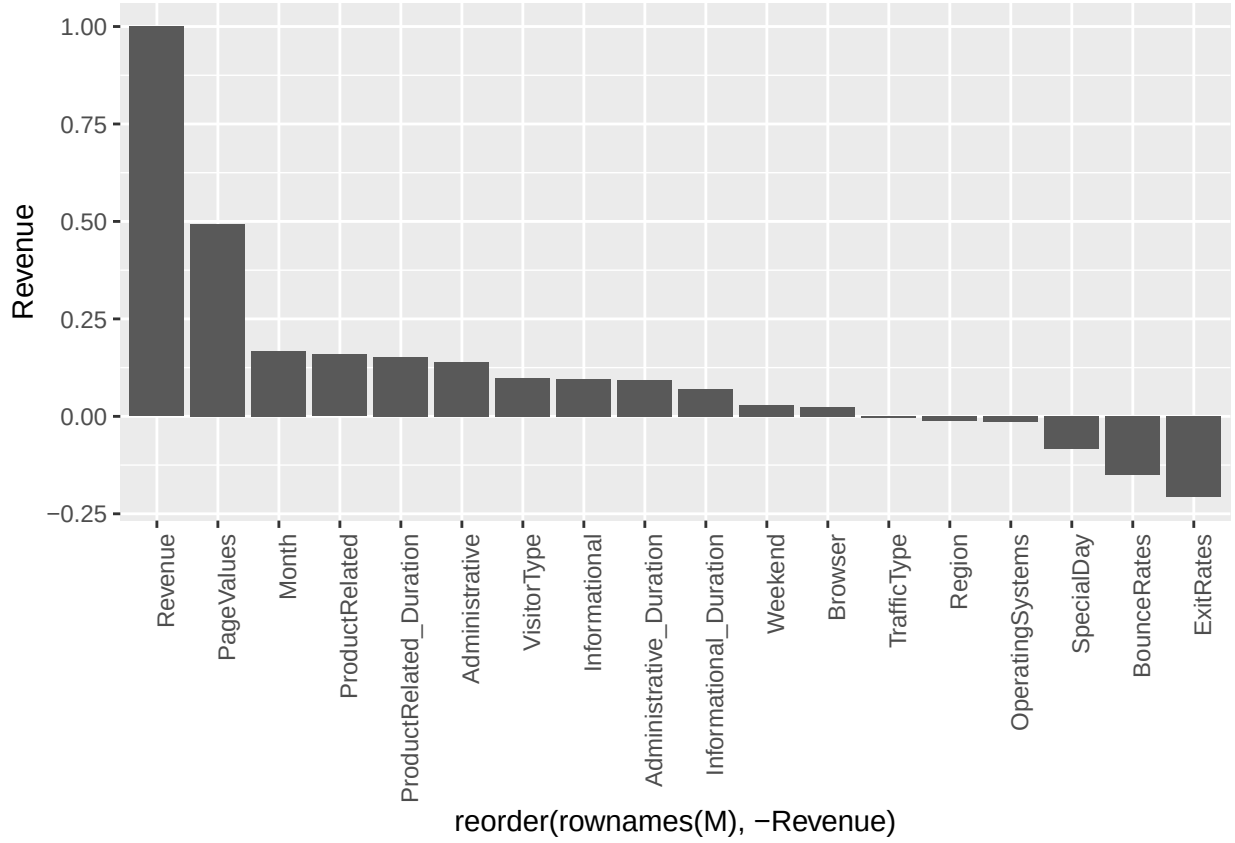
(Google Analytics, 2020c)

Total Goal Value is defined as the dollar amount assigned to the completion of goal in Google Analytics (Koks, 2000).

From this equation, the page value can be increased in two ways. Either the Total Goal Value increases in value or there must be a Unique Pageview for a given page. Ideally, it is best to have a small divisor,

with the pageviews for a given page at 1. Smaller pageviews value can be accomplished if fewer pages are visited before arriving at the transaction or goal page. A high page value is indicative that a user will make a purchase and this will ultimately increase business revenue.

Feature Selection



In this section, important features were selected using correlation values.

The cor() function was used to obtain correlation values. Features containing a very weak correlation strength value falling below 0.02 and -0.02 were detected. It was decided that features falling below a correlation value 0.02 and 0.02 will not be included in the final data set.

In sum, the following were used in the prediction models: Page Values, Month, Product-related, Produce-related duration, Administrative, Visitor type, Informational, Administrative Duration, Informational Duration, Weekend, Browser, Special Date, Bounce Rates and Exit Rates.

Data preparation

The dataset was checked for missing values and outliers.

```
colSums(is.na(datadump))
```

```
##          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(datadump=="")
```

```
##           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
```

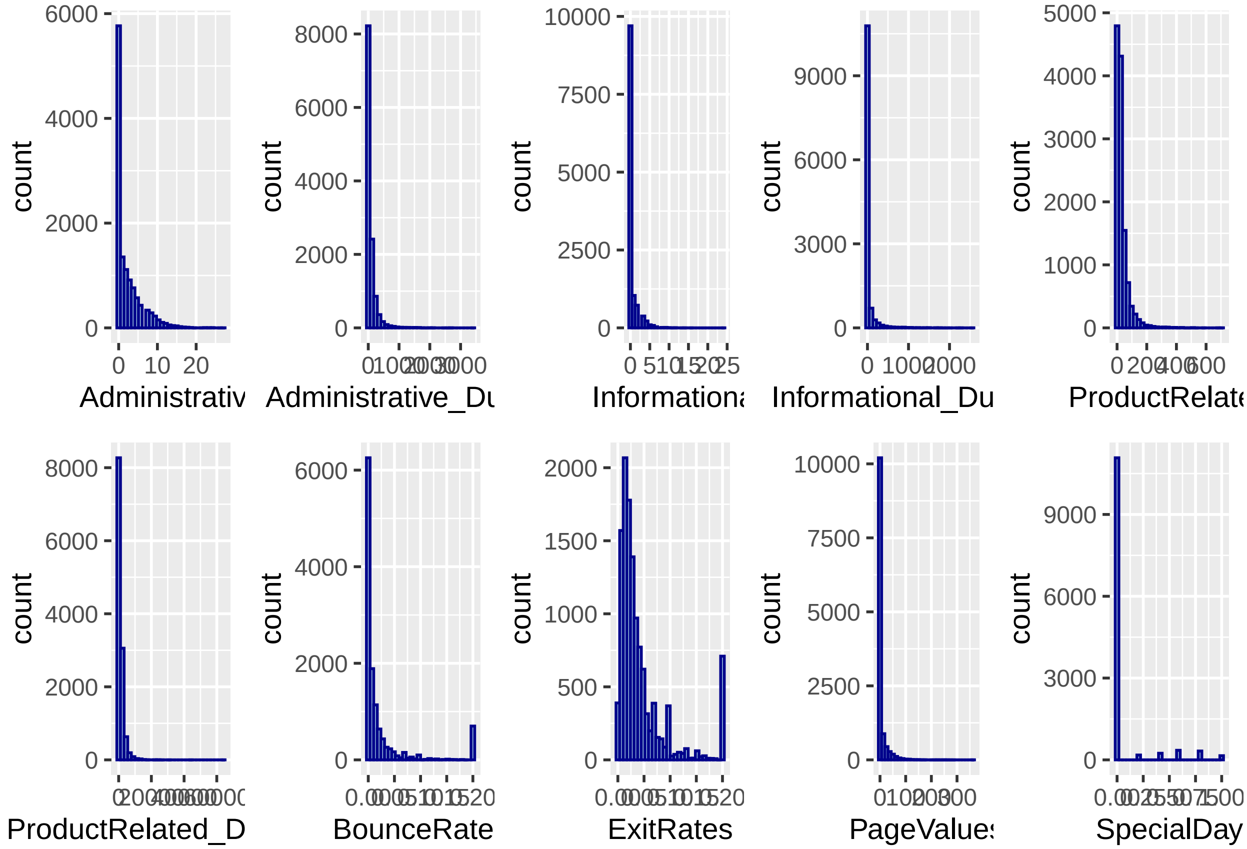
```
#Indetifying Outliers in numeric columns using IQR
num_cols<-
  c("Administrative","Administrative_Duration","Informational",
    "Informational_Duration","ProductRelated","ProductRelated_Duration",
    "BounceRates","ExitRates","PageValues","SpecialDay")
num_outliers<-
  data.frame("Q1"=numeric(10),"Q3"=numeric(10),"IQR"=numeric(10),
             "UL"=numeric(10),"LL"=numeric(10),"Count"=numeric(10))
rownames(num_outliers)=num_cols

for (i in 1:dim(num_outliers)[1]){
  num_outliers[i,"Q1"] = quantile(datadump[,num_cols[i]], 0.25, na.rm=TRUE)
  num_outliers[i,"Q3"] = quantile(datadump[,num_cols[i]], 0.75, na.rm=TRUE)
  num_outliers[i,"IQR"] = IQR(datadump[,num_cols[i]],na.rm=TRUE)
  num_outliers[i,"UL"] = num_outliers[i,"Q3"]+1.5*num_outliers[i,"IQR"]
  num_outliers[i,"LL"] = num_outliers[i,"Q1"]-1.5*num_outliers[i,"IQR"]
  num_outliers[i,"Count"] = length(datadump[,num_cols[i]]
                                   [datadump[,num_cols[i]]>=
                                       num_outliers[i,"UL"] | datadump[,num_cols[i]]<=
                                       num_outliers[i,"LL"]])
}
```

While it had no missing values, the dataset had outliers that made up at least 10% of the total dataset. Due to the significant portion of outliers in our total dataset, they were not excluded.

Then, distribution was observed for all numeric columns, as shown in the below 10 histograms.

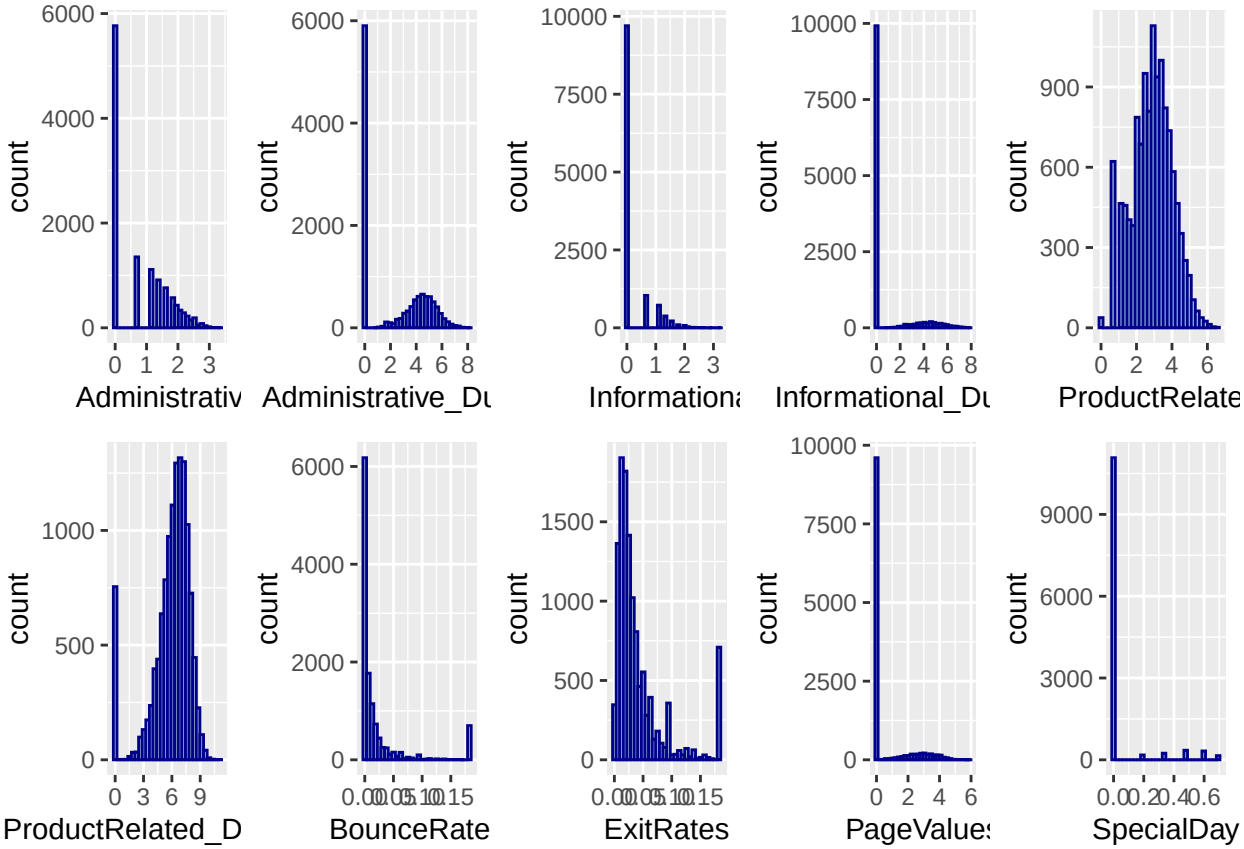
```
## Loading required package: gridExtra
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



All of the above histograms show an extreme positive skewness. Therefore, they were transformed through a log application, which normalized the distribution, as shown in the 10 histograms below.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Models

Predictive Model Development and Evaluation

The dataset was split into 3 subsets for predictive modeling: train, validate, and test. Given that the target variable is imbalanced, createDataPartition function was used to ensure that the target variable is equally and consistently represented in all three data subsets (the below 3 tables show that the target variable 0's and 1's are equally and consistently represented in the train, validate, and test datasets, respectively).

```
## Loaded ROSE 0.0-3

## Loading required package: lattice

##
##         0         1
## 0.5045053 0.4954947

##
##         0         1
## 0.5045045 0.4954955

##
##         0         1
## 0.5048701 0.4951299
```

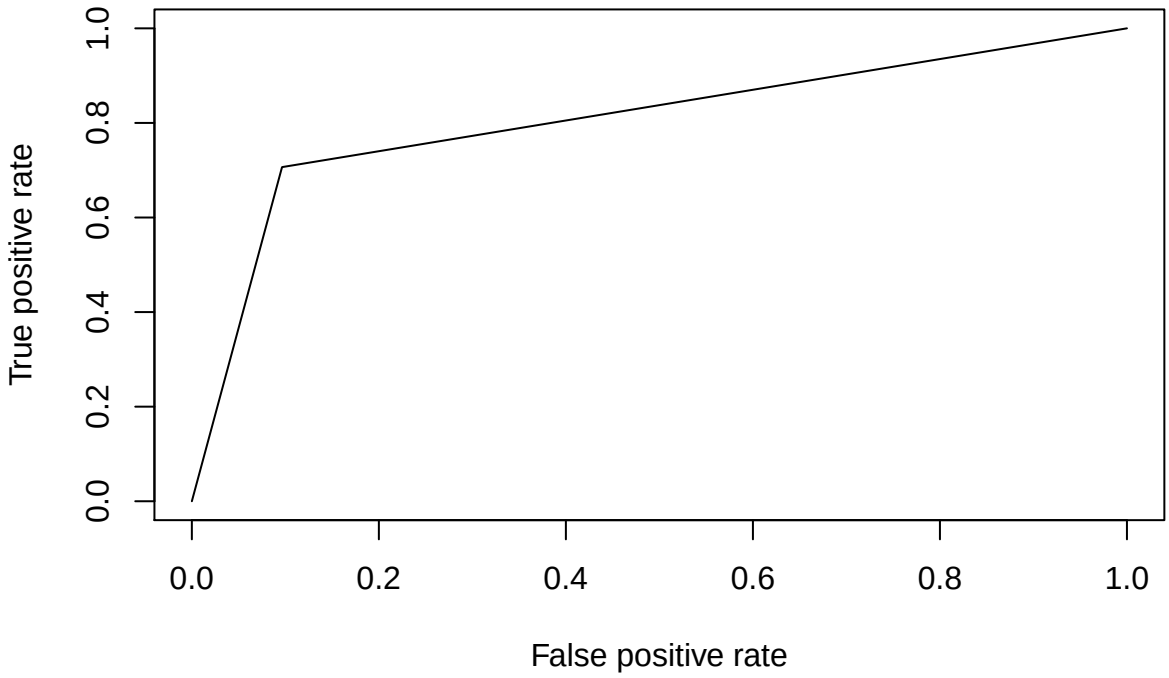
Also, 'ROSE' R package library was used to address the unbalanced data, since it creates balanced synthetic data samples through oversampling, which is accomplished by enlarging the features space of minority and majority class examples (R Documentation, n.d.). Operationally, the new examples are drawn from a conditional kernel density estimate of the two classes, as described in Menardi and Torelli, 2013.

First, Logistic Regression was used to predict online site-visitors' purchase status. It is a machine learning algorithm that models the probabilities for a classification problem by giving two probable outcomes (Molnar, 202). Logistic Regression Models (LRM) provides probabilities for classification problems with two possible outcomes. LRM is a continuation of the Linear Regression Model as it takes classification problems one step further (Molnar, 2020) When compared to other models, this model can be quite advantageous as it provides probilities and can determine the final classification. LRM are not without disadvantage as they perform poorly against other models when predicting performance. Also, LRM poses the challenge of being unable to further train in the event where there is a feature that could separate two classes (Molnar, 2020).

Logistic Regression had a mean prediction score of 81%, and had the following metrics along with ROC curve:

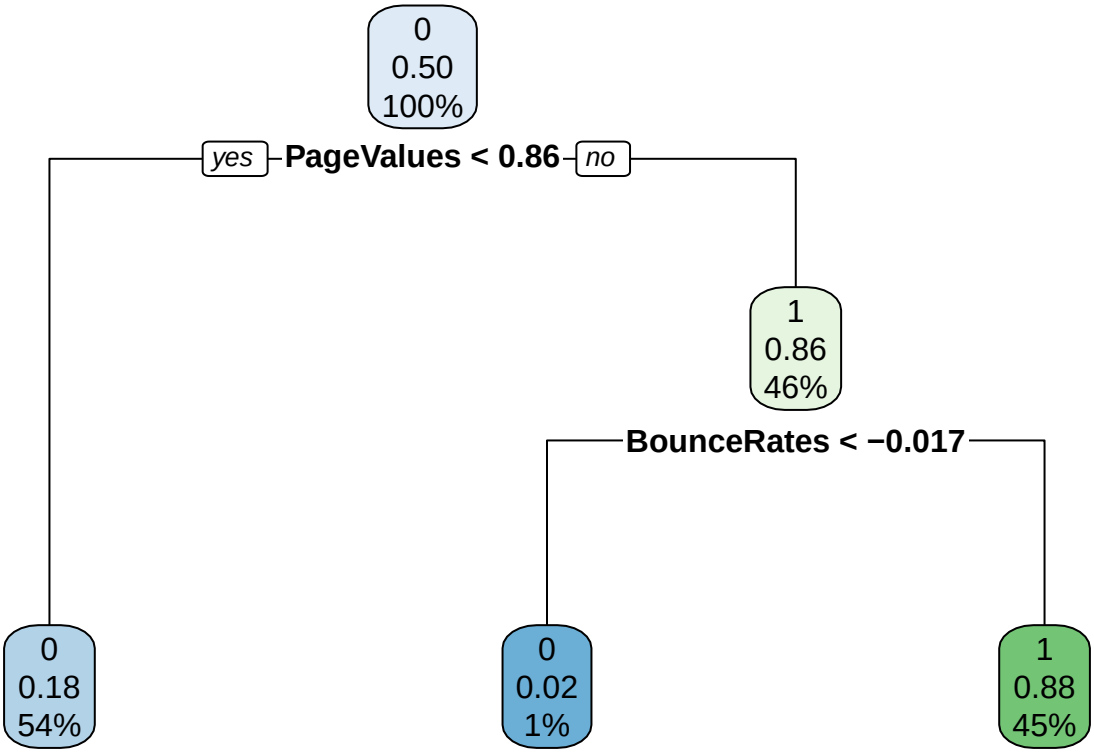
- Precision: 0.758
- Recall: 0.904
- Overall accuracy: 0.806
- Balanced accuracy: 0.818
- Specificity: 0.707
- F1: 0.825
- AUC: 0.805

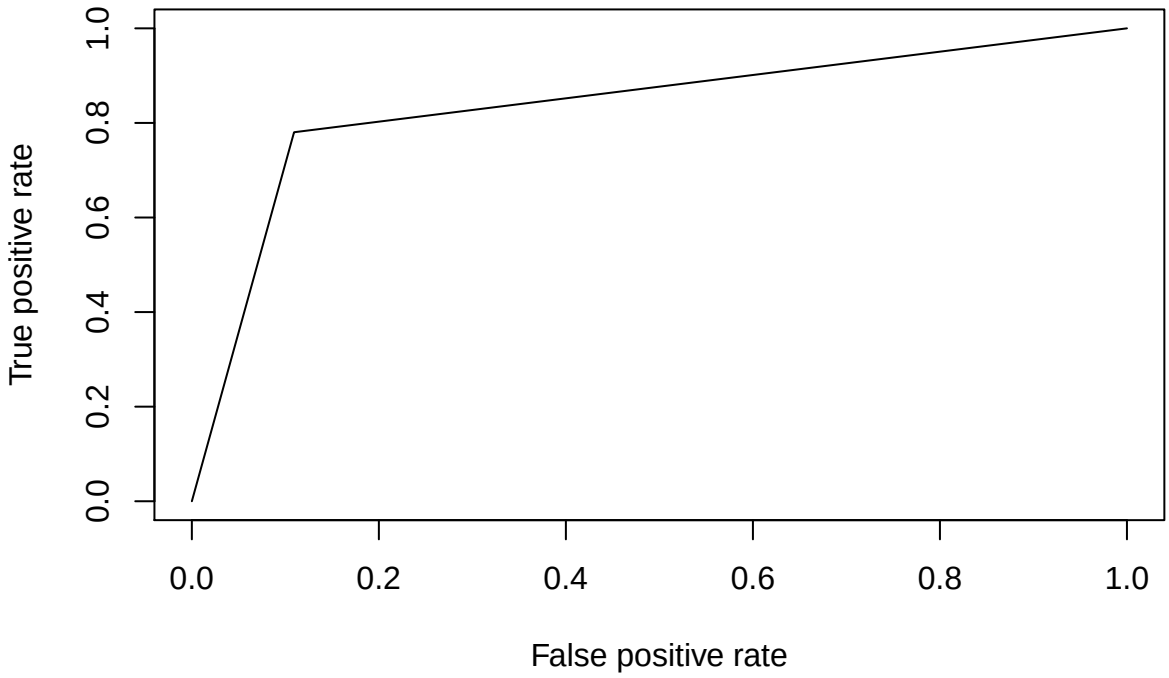
```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall
```



Next, the Decision Tree model was used to predict site-visitors' purchase status. A Decision Tree can be used for classification and regression problems (Molnar, 2020). This model is able to demonstrate interactions between varying features within the data. Decision Trees allow for the data to be represented in specific groups, this allows for easy data interpretation. Another major advantage of decision trees is the model's ability in providing a clear explanation (Molnar, 2020). Decision Tree models are unable to handle linear relationships, this is because the existing relationship between a feature and an outcome is approximated by splits, resulting in a step function (Molnar, 2020). This process is not efficient for this model. Another disadvantage in such a model can prove to be unsteady in that minor change made to the training dataset results in another tree. The model resulted with following metrics:

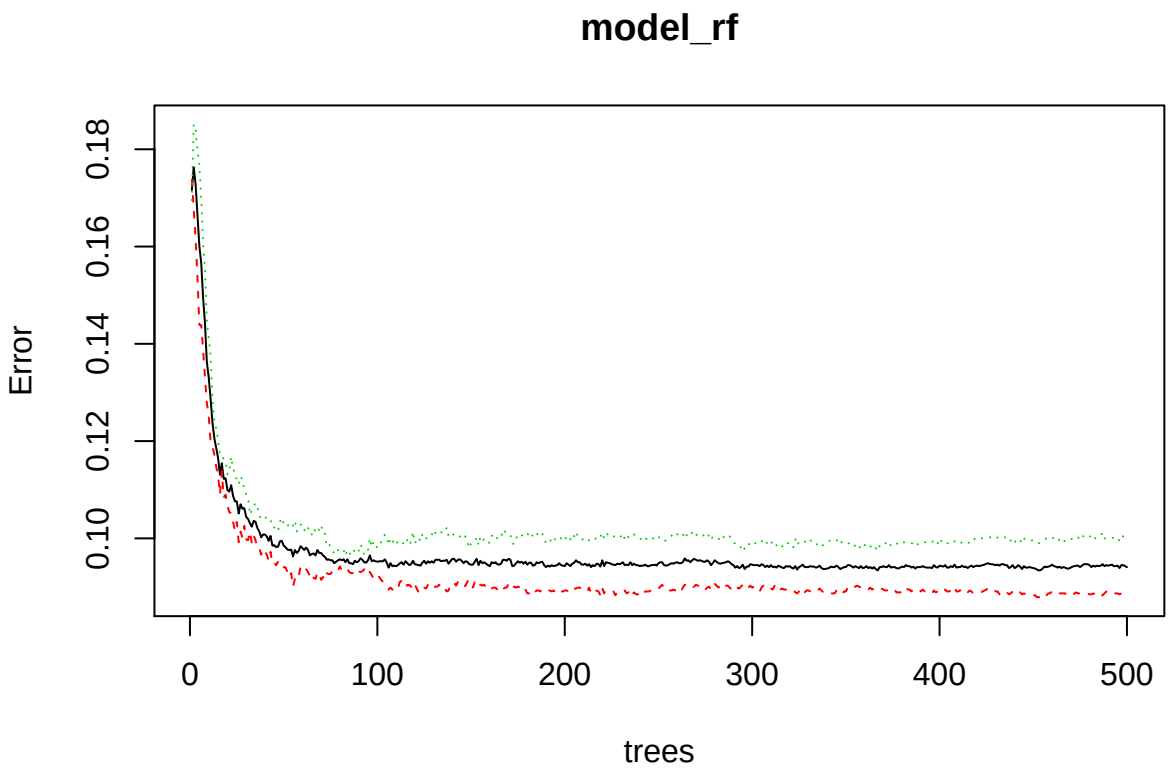
- Precision: 0.805
- Recall: 0.891
- Overall accuracy: 0.836
- Balanced accuracy: 0.840
- Specificity: 0.780
- F1: 0.846
- AUC: 0.836


0
0.50
100%
yes
PageValues < 0.86
no
1
0.86
46%
BounceRates < −0.017
0
0.18
54%
0
0.02
1%
1
0.88
45%



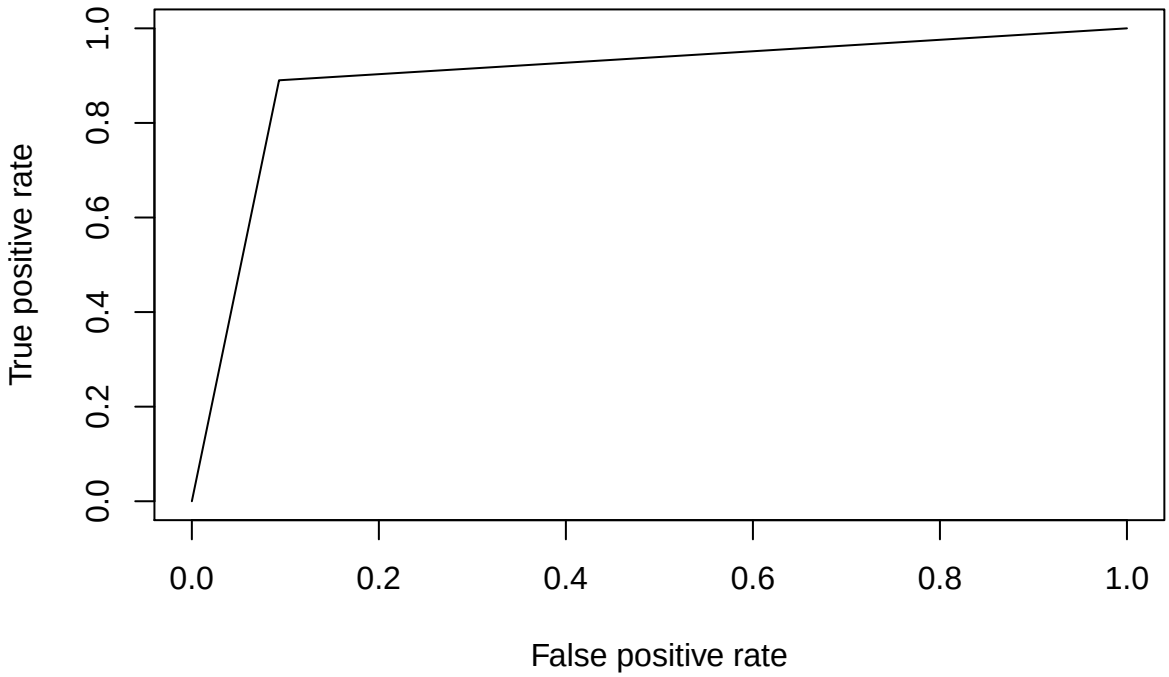
Also, the Random Forest model was used as one of the prediction models. Random forest models are made of decision trees that have no correlation between them. Such a model can determine the significance of a feature as well as the interaction between the varying features ("Random Forest", n.d.). Features do not have to be selected and dimensions do not have to be reduced, this is because such a model can give high dimensional data ("Random Forest", n.d.). The more trees that are present, this model decreases the tendency to overfit. Lastly, this model is very easy to train and implement. One of the major disadvantages of random forest is that it tends to fit specific noisy classification or regression problems ("Random Forest", n.d.). The following shows the random forest plot.

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```



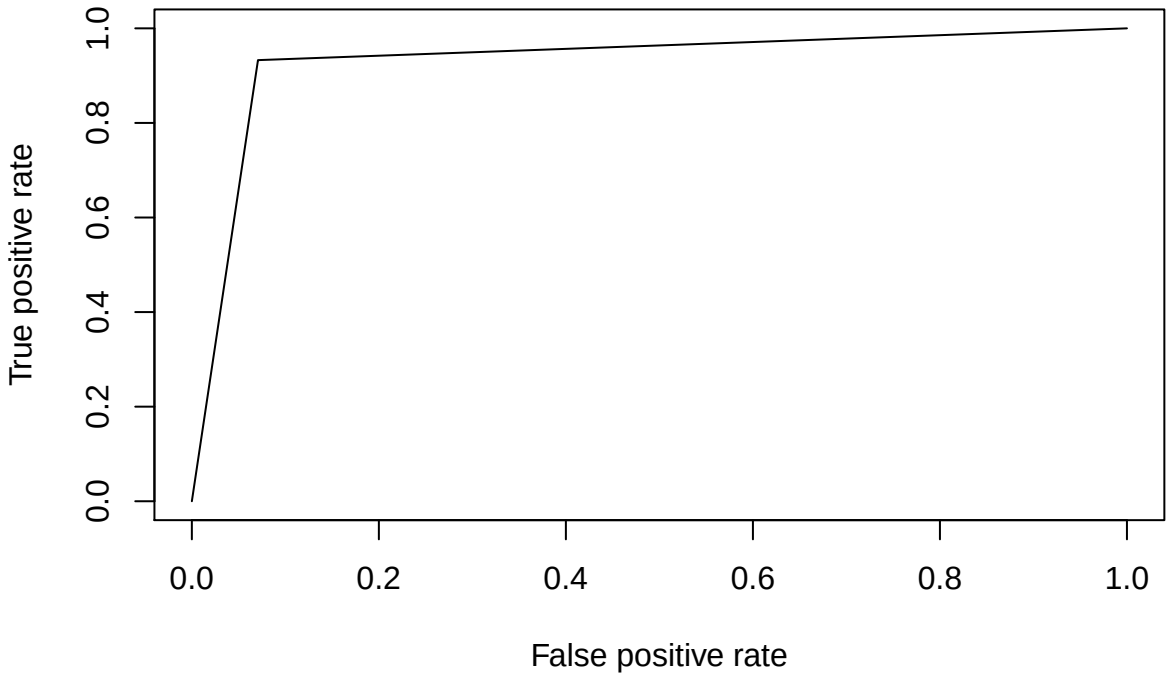
Random Forest model had the following metrics and ROC curve:

- Precision: 0.894
- Recall: 0.907
- Overall accuracy: 0.899
- Balanced accuracy: 0.899
- Specificity: 0.890
- F1: 0.900
- AUC: 0.898



Lastly, the K-Nearest Neighbor (k-NN) model has been implemented. Although it is said to be lazy, it proves to be quite a versatile algorithm. This model can be used to solve classification and regression problems ("What is the K-nearest", n.d.). Furthermore, k-NN can also be used for non-linear classification. Interestingly, k-NN models make no assumptions about the data, they demonstrate high accuracy and are not sensitive to outliers ("What is the K-nearest", n.d.). One of the many challenges this model presents is that it requires a lot of memory. Also, it deploys an ineffective sample balance, in that the number of samples in certain categories can be large, while the number of other samples are small. Additionally, in order to obtain optimal k value selection, the k value size must be combined with k-fold cross validation. Lastly, if a sample is unbalanced, k-NN gives a large prediction bias ("What is the K-nearest", n.d.). This algorithm performed with the below metrics and ROC curve:

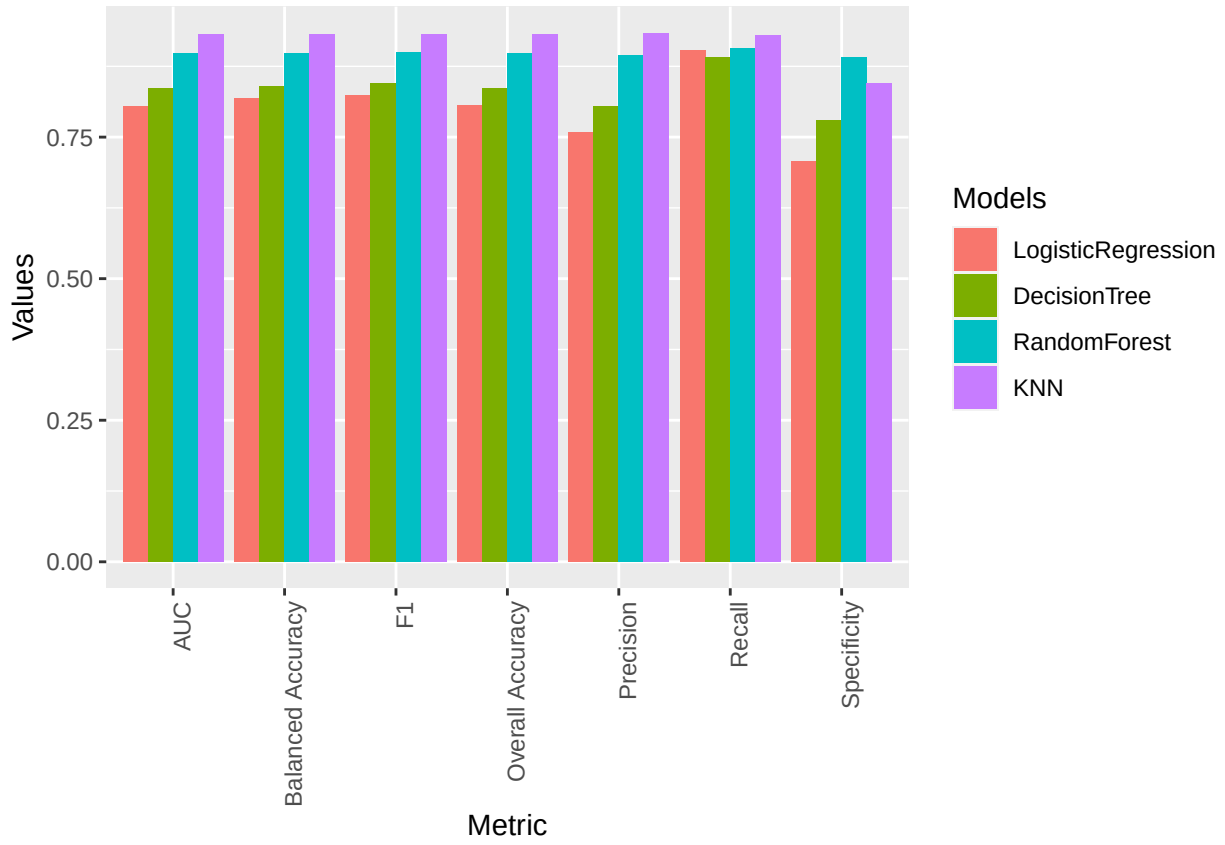
- Precision: 0.934
- Recall: 0.929
- Overall accuracy: 0.931
- Balanced accuracy: 0.928
- Specificity: 0.844
- F1: 0.932
- AUC: 0.931



Putting all the evaluation metrics of the models together, the following visualizations show how the models performed compared to one another.

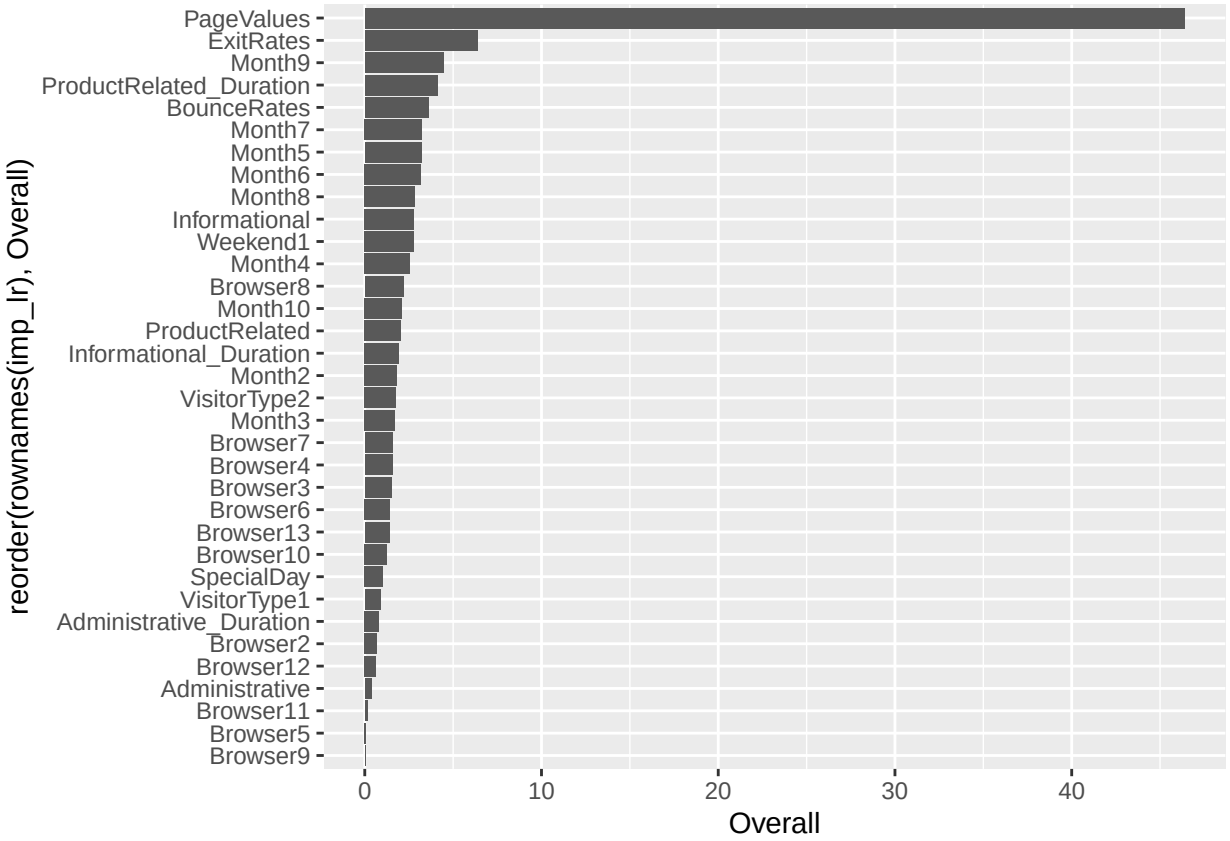
```
##                    LogisticRegression DecisionTree RandomForest       KNN
## Precision                   0.7584345    0.8052326    0.8938193 0.9337641
## Recall                      0.9035370    0.8906752    0.9067524 0.9292605
## Overall Accuracy            0.8060065    0.8360390    0.8985390 0.9310065
## Balanced Accuracy           0.8181175    0.8401163    0.8986568 0.9309930
## Specificity                 0.7065574    0.7803279    0.8901639 0.8442623
## F1                          0.8246515    0.8458015    0.9002394 0.9315068
## AUC                         0.8050472    0.8355016    0.8984582 0.9310237
```

```
## Warning in melt(metrics_v, id.vars = "Metric", value.name = "Values",
## variable.name = "Models"): The melt generic in data.table has been passed
## a data.frame and will attempt to redirect to the relevant reshape2 method;
## please note that reshape2 is deprecated, and this redirection is now
## deprecated as well. To continue using melt methods from reshape2 while both
## libraries are attached, e.g. melt.list, you can prepend the namespace like
## reshape2::melt(metrics_v). In the next version, this warning will become an
## error.
```

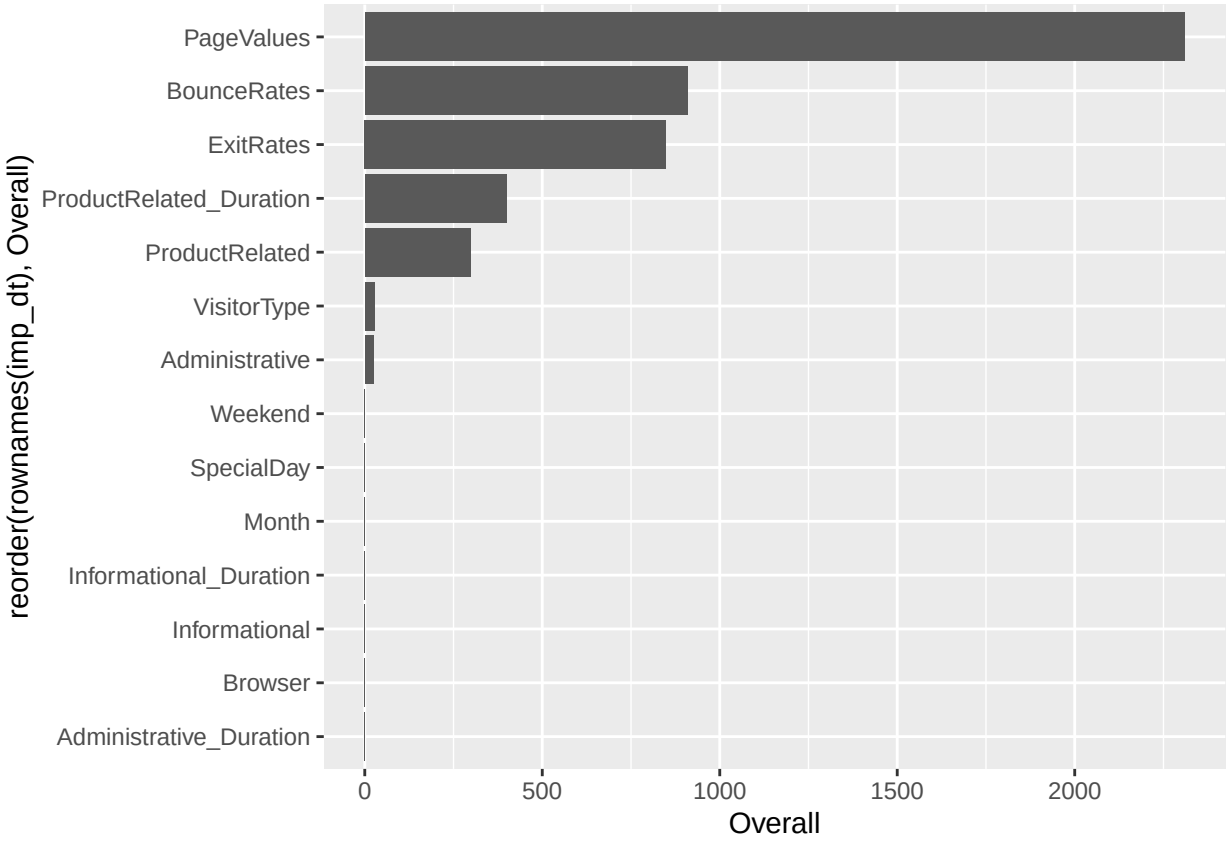


The four models performed are ranked as follows: 1) k-NN was the best model, 2) Random Forest was the second best model, 3) Decision Tree was the third best model, and 4) Logistic Regression performed the worst across the metric scores. Other than specificity, where Random Forest performed the best, k-NN model had the best scores in all metrics.

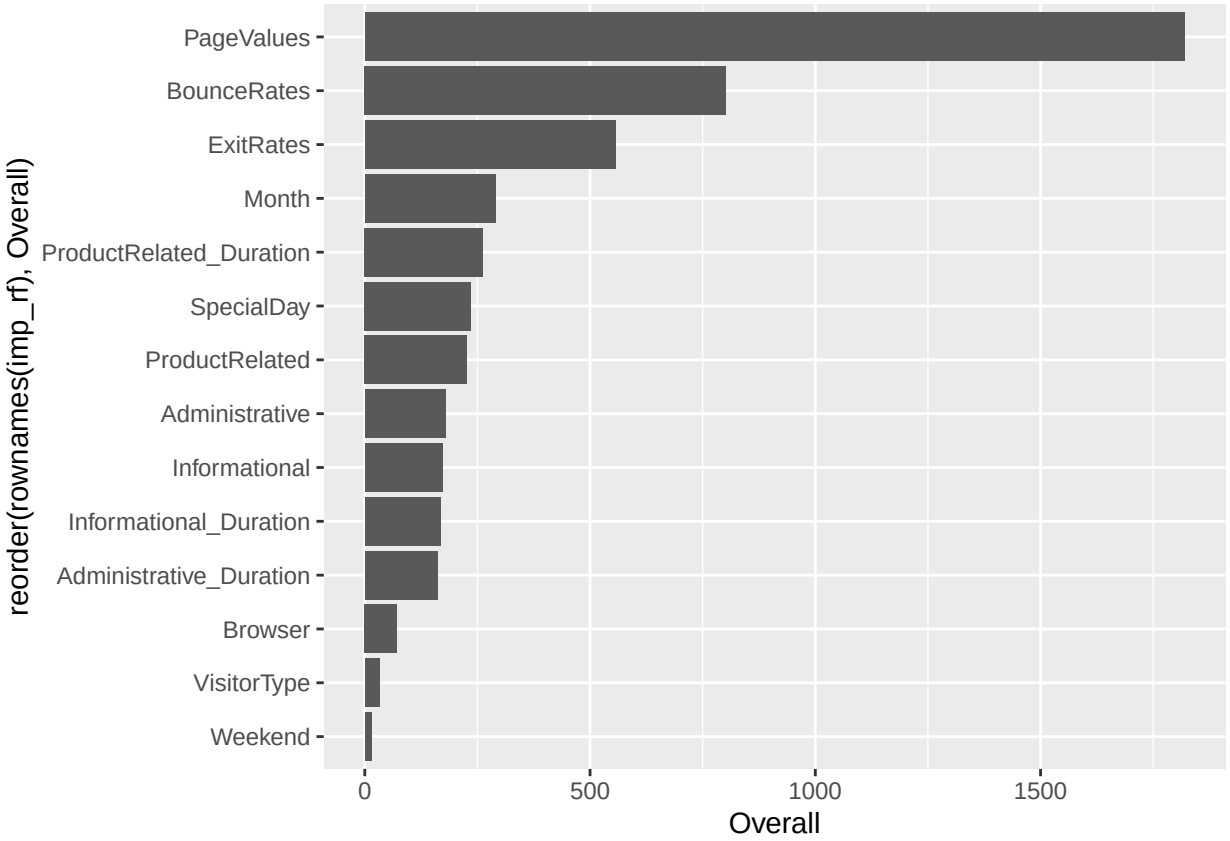
In addition to the model evaluation, further evaluative steps were taken to identify which features were more important and significant in the predictive models. In Logistic regression, it is shown that 'Page Values' was the most important feature, and 'Exit Rate' was the second most important feature. Browser Type was the least important feature.



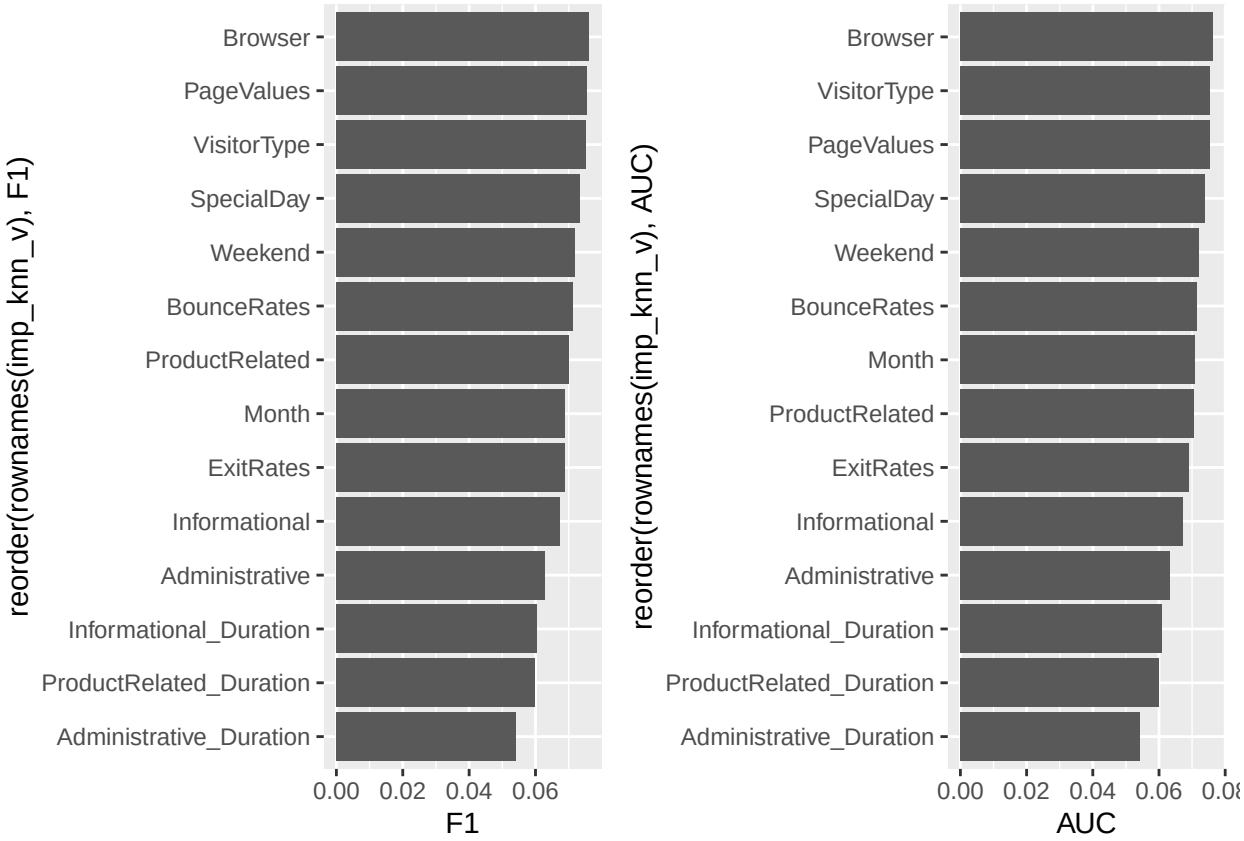
In the Decision Tree model, 'Page Value' was again the most important feature, and 'Bounce Rate' was the second most important feature. Exit Rate, which was ranked as the second important feature in Logistic Regression, came third in the Decision Tree model, which affirms 'Exit Rate' as one of the important features. Administrative Duration was the least important feature, and 'Browser Type' was the second least important feature.



In the random forest model, 'Page Value' was again the most important feature, 'Bounce Rate' as second, and 'Exit Rate' as third most important feature. Weekend was the least important feature. The 'Visitor type' was the second least important, and 'Browser type' was confirmed again as less important by coming to the third place. So far, all models displayed similar features as important and unimportant.



However, k-NN displayed contradicting results. It displayed 'Browser' and 'Visitor Type' as the important features, while those two features were least important features in other models. However, k-NN still showed 'Page Value' as one of the top important features, which is consistent with other models as well.



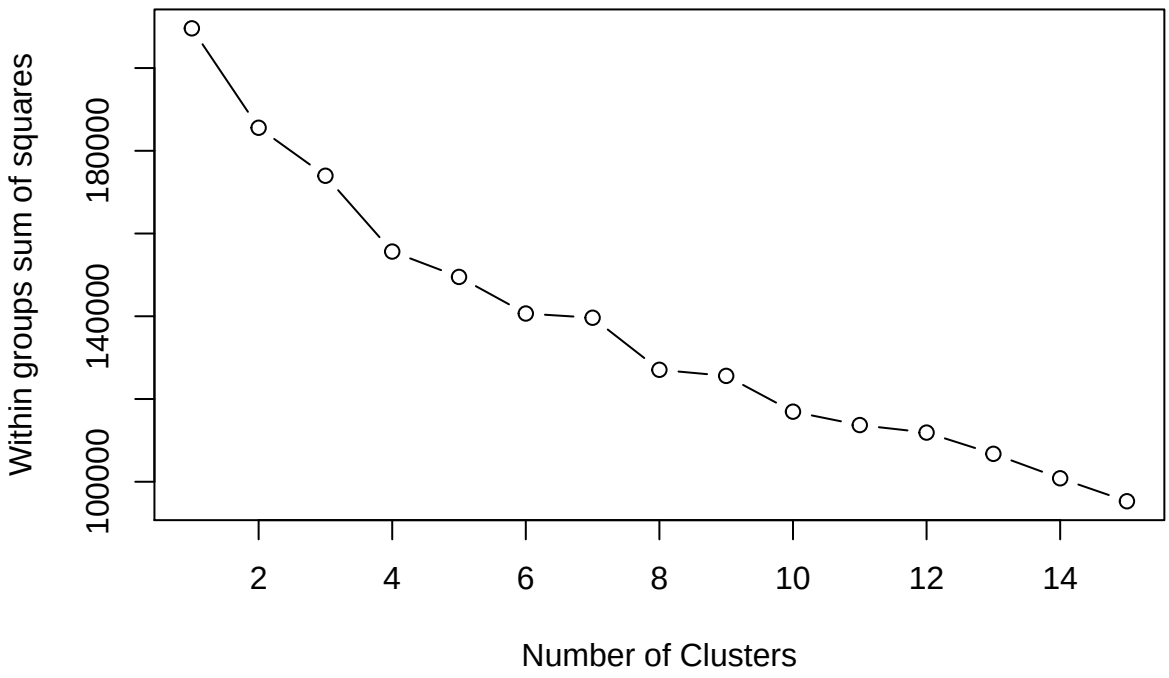
Overall, 'Page Value' seems to be the most important feature, as it was consistently shown in all models to be the most contributing one. The 'Exit Rate' and 'Bounce Rate' closely follow as the second and third most important features. This indicates that the factors that influence website visitors to make a purchase are the value score of the web page and the increased length of stay in the web page.

Clustering Model Development and Evaluation

While classification models were used to predict website visitor's purchase and to evaluate important features that affect the prediction the most, clustering models were used to answer the third analytical question which is two-fold. First, what are the common behaviour trends in our user base/ how are important features represented across our user base? Second, if users are segmented, does this reflect purchasing frequency/power as an alternative to classification models? These analytical questions are designed to provide a thorough understanding of users and their behaviour trends.

Finding optimal K

To prepare for clustering modeling, different methods were used to find the optimal number of clusters (k). First, the Elbow method was used; this method is useful in choosing the optimum value of k as it fits the model with a range of values of k (Franklin, 2019). Unfortunately, with the online shoppers dataset, this method produced a result that was not easily comprehensible.



Therefore, the Calinski-Harabasz method was used as it assesses cluster validity based on the average between and within the cluster sum of squares (Liu et al., 2010). This method indicated the optimal number of K to be 3.

```
library(fpc)
#Calinkski-Harabasz method
tunek_ch <- kmeansruns(clust_data_sd,krange = 1:10,criterion = "ch")
tunek_ch$bestk
```

```
## [1] 3
```

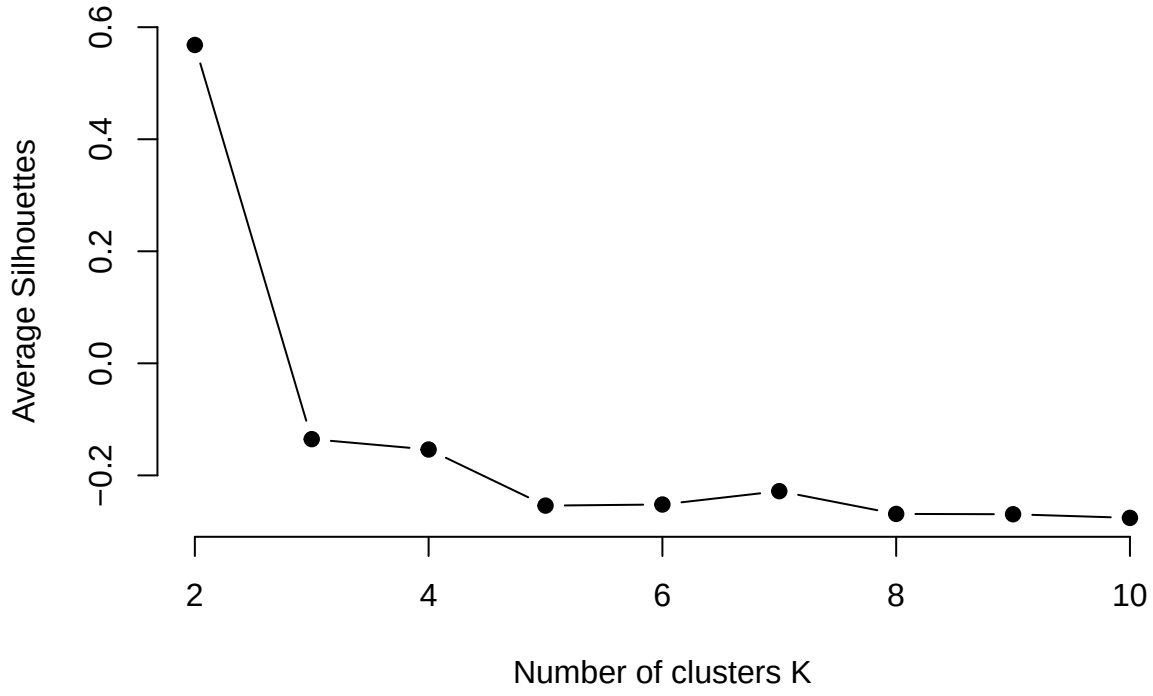
Then, the Silhouette method was used as it can assess the degree of separation between clusters (Dabbura, 2018). In each sample, the Silhouette method will determine the average distance from all data points in the same cluster, determine average distance from all data points in the closest cluse and finally it will determine the coefficient using the following equation:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

In which, the coefficient will take a value between -1 and 1 (Dabbura, 2018). A coefficient value of "0" indicates that the sample is very close to neighbouring clusters, whereas a value of "1" denotes that the sample is far from the neighbouring clusters. Finally, a value of "-1" means that the sample is designated to the wrong cluster (Dabbura, 2018).

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:data.table':
##
##     transpose

## The following object is masked from 'package:caret':
##
##     lift
```



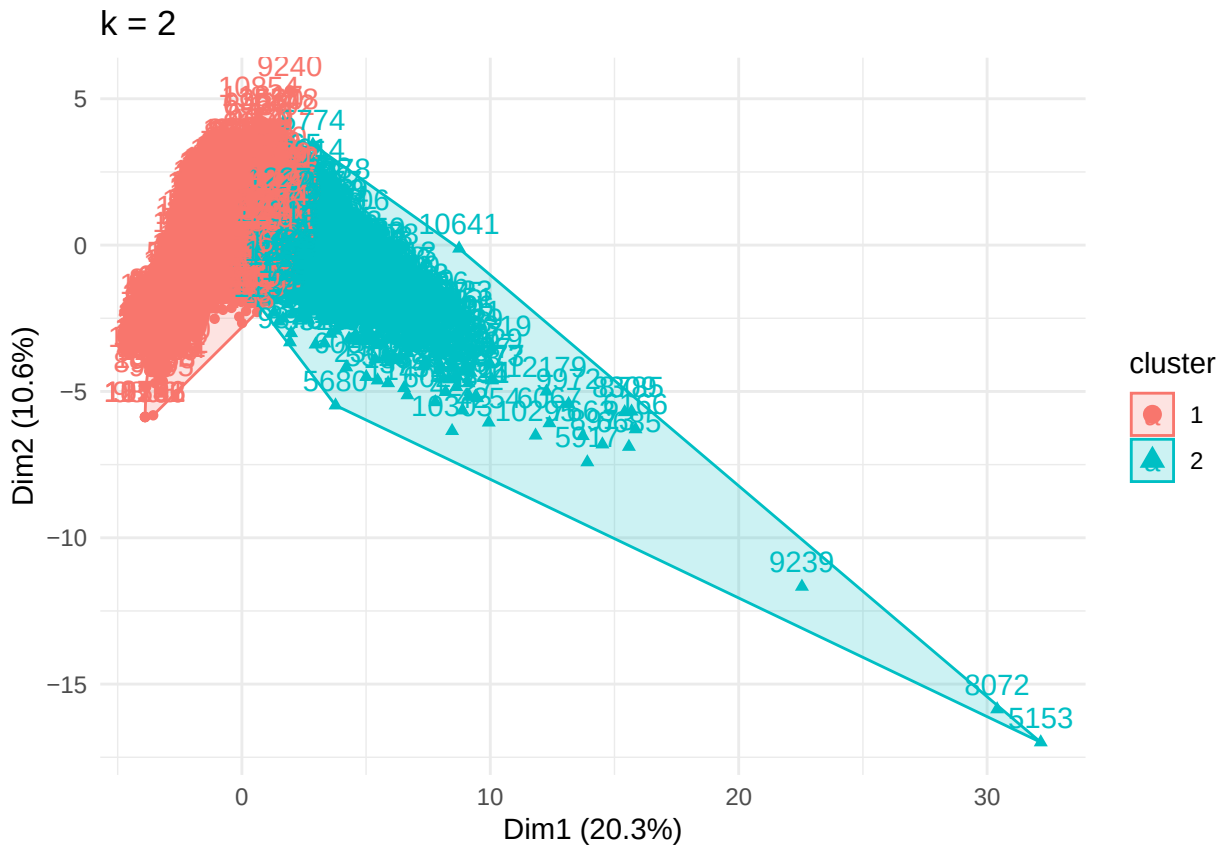
When applied, this method indicated the ideal number of k to be 2. Therefore, both k=2 and k=3 will be used in this clustering analysis. All these methods employed K-means as a base because calculating for optimal k using PAM or Hierarchical clustering was excessively heavy in terms of CPU and memory usage at every attempt.

k=2, Clustering, K-means, PAM, and Hierarchical

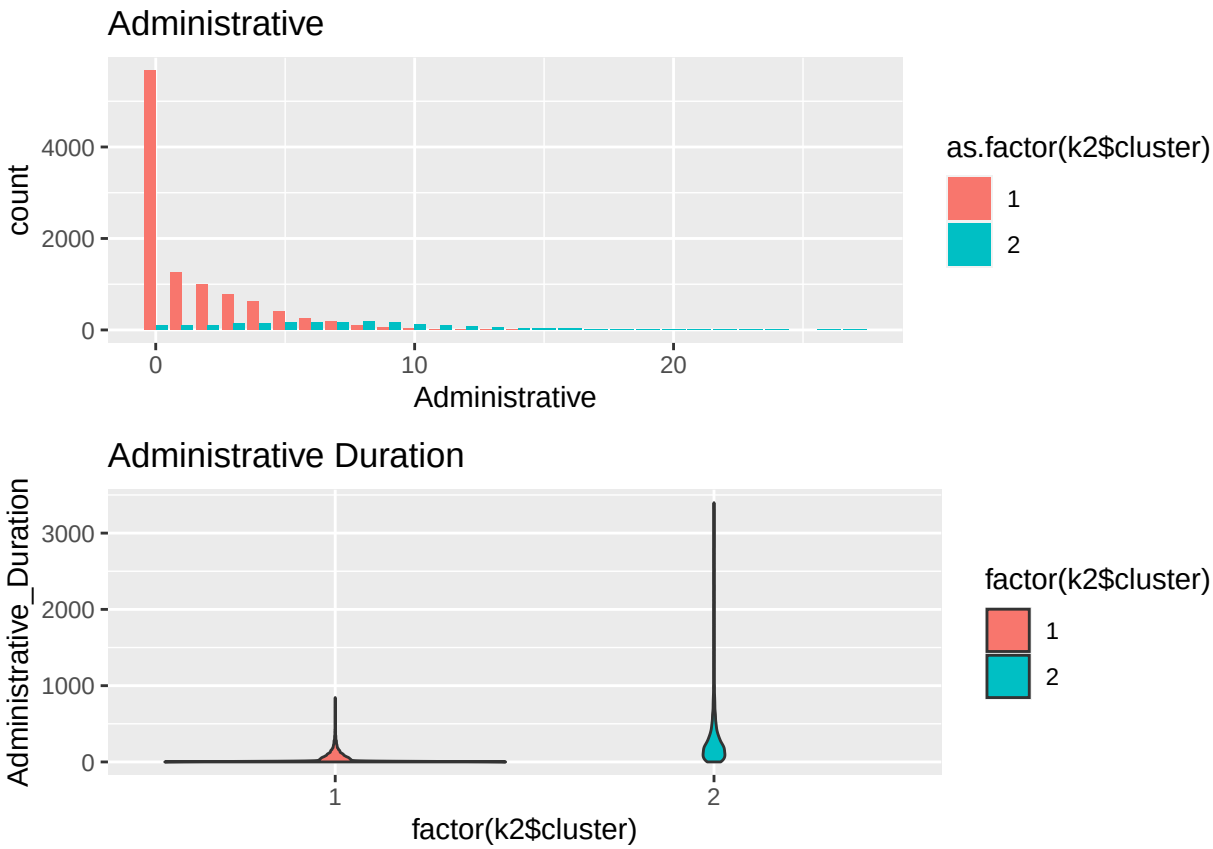
K-means is used to find groups within the data, where the number of groups are denoted by the variable K (Trevino, 2016). This method functions by repeatedly assigning each data point to one of the K groups. Then, data points are clustered according to their feature similarity. The final output of K-means clustering method is 1) the centroids of the k clusters and 2) labels for the training data (Trevino, 2016).

K-means clustering produced the following visual when k=2 was used. There were a total of 10,396 site visitors in cluster 1, and 1,934 site visitors in cluster 2.

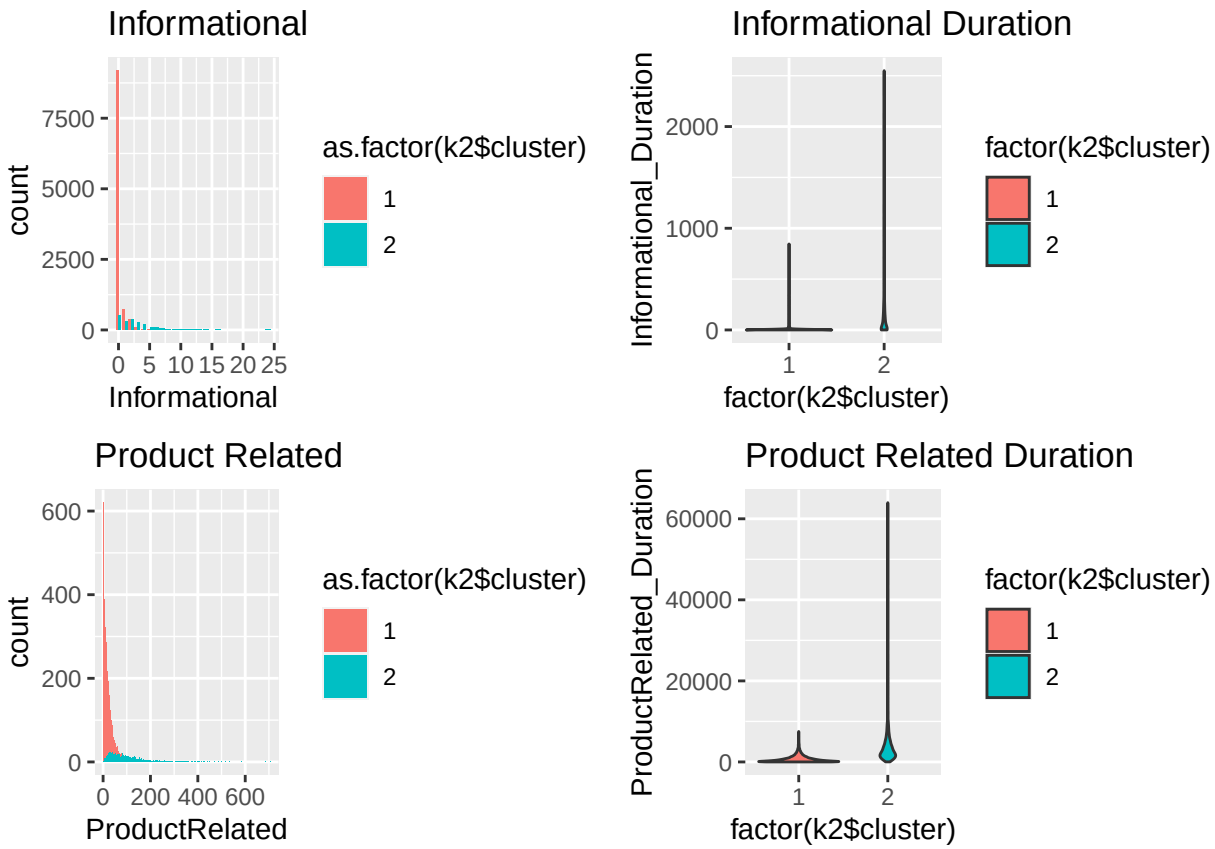
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



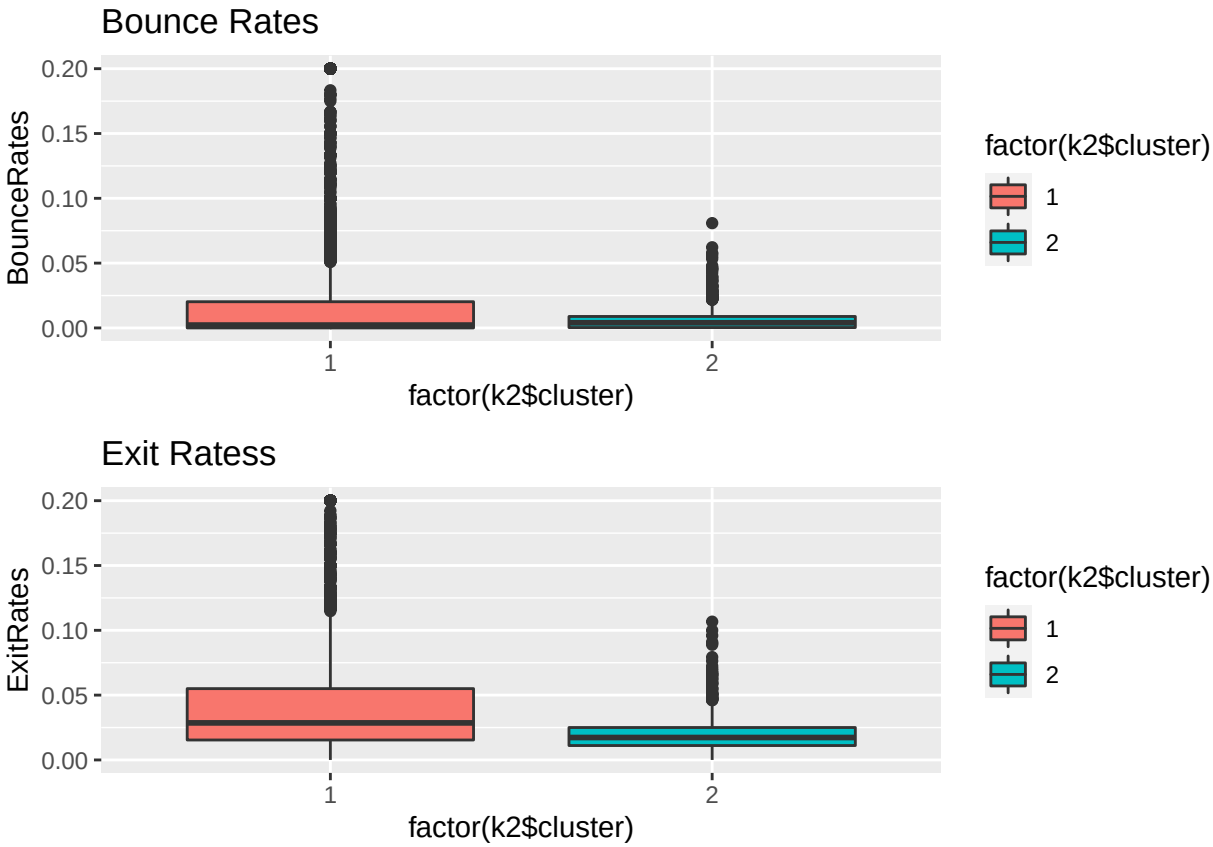

```
##
##     1     2
## 10396  1934
```



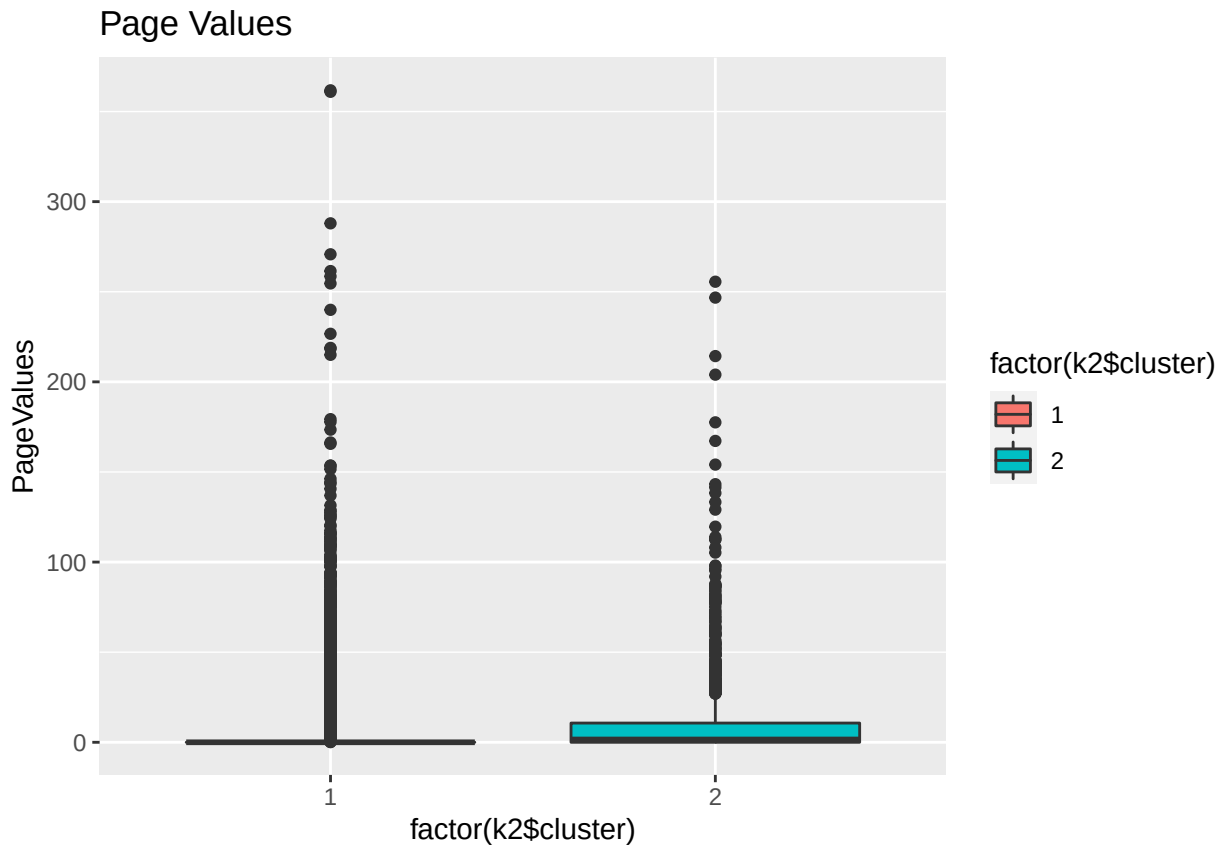
Web page visitors in cluster 1 visited distinctively less administrative (account-related) pages than cluster 2, and spent much less time in the administrative pages than cluster 2.



The exact same behavior pattern was repeated for information (company-related) pages and product-related pages as well, where cluster 1 not only visits less sites, but also spends less time in those sites.

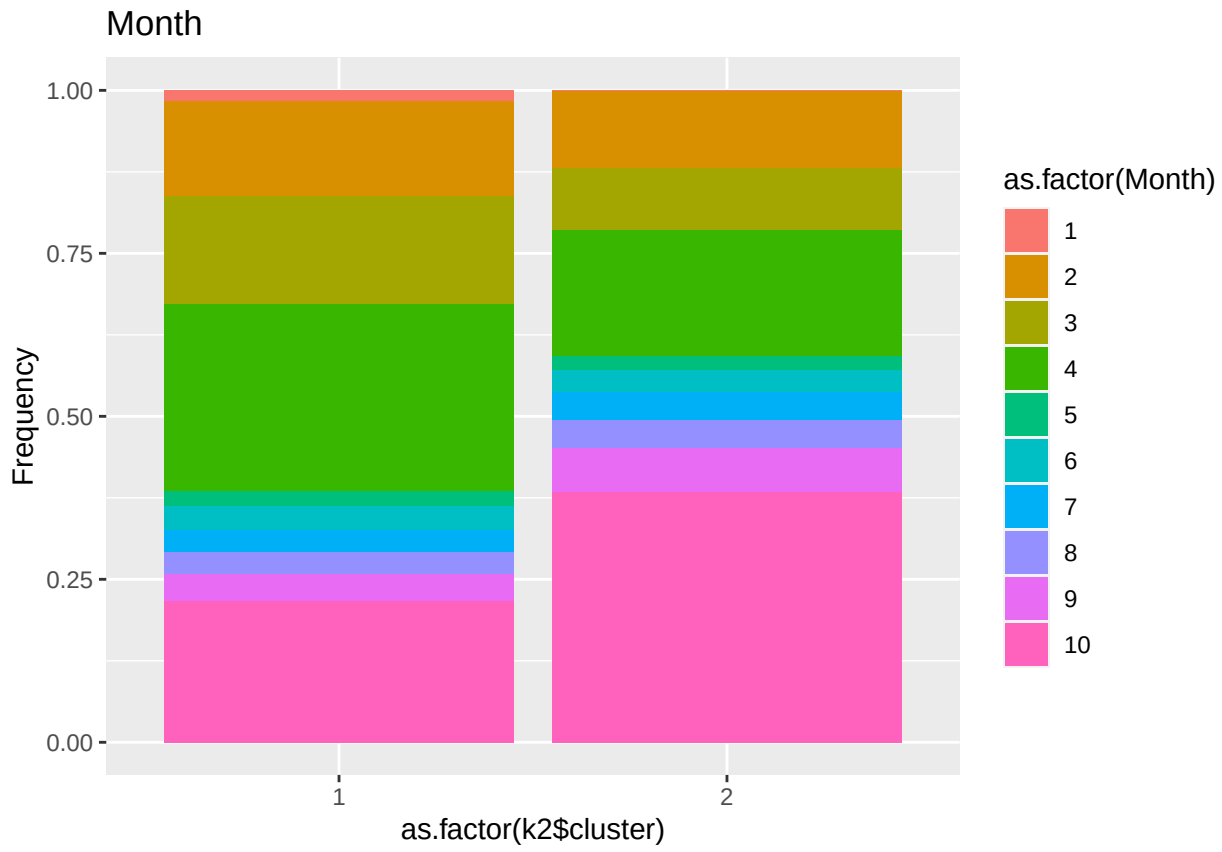


These insights are consistent with the bounce and exit rates as well. Cluster 1 had a much greater bounce and exit rate than cluster 2.

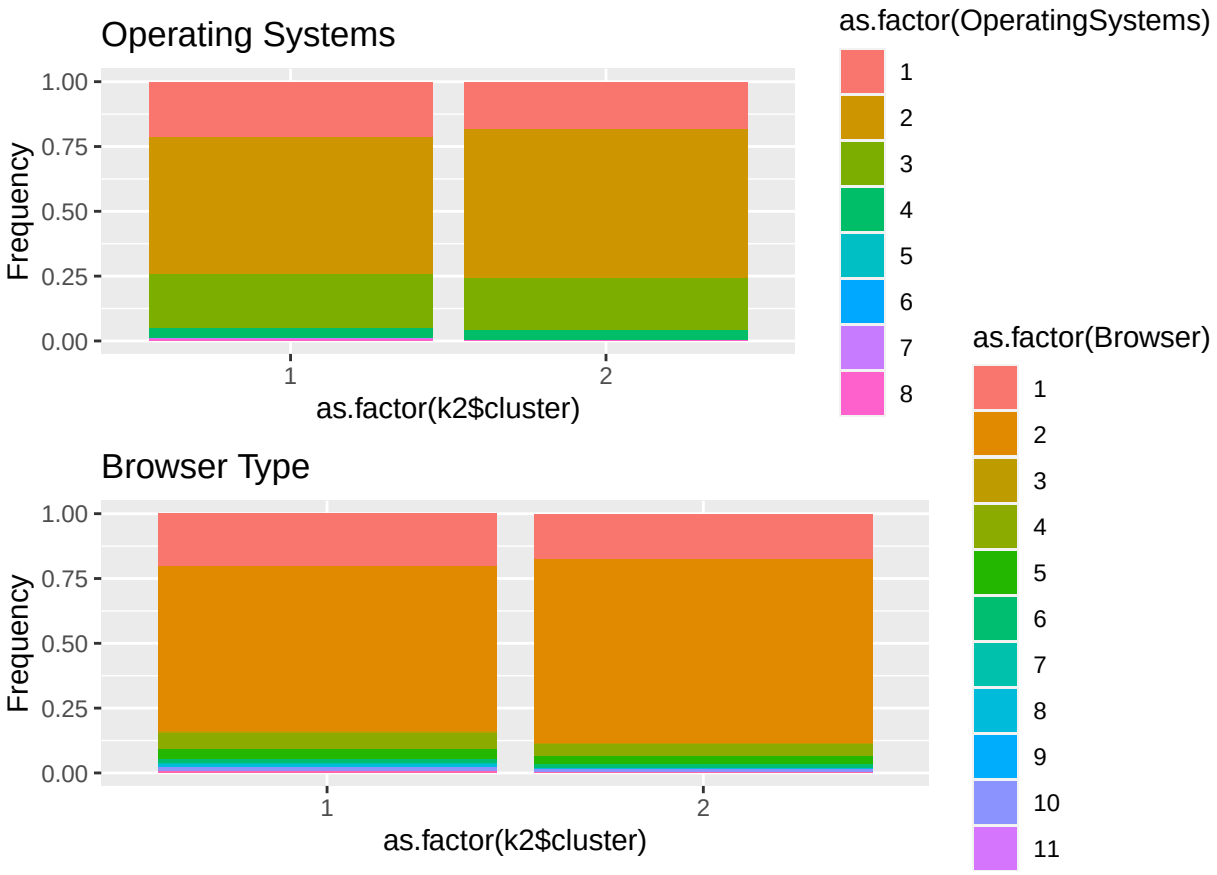


However, page value shows an interesting contradiction. Based on the insights gathered so far, it would make sense if cluster 1 had a lower page value than cluster 2. While the interquartile ranges show that cluster 2 has a higher page value than cluster 1, the outliers show that cluster 1 has a much higher page value than cluster 2. One theory that might explain cluster 1's outliers with a high page value is that site-visitors in cluster 1 generally do not visit many pages in the shopping website and exit right away. This naturally gives a low number divisor in the page value equation (e-commerce revenue + goal value / number of page views). Naturally, site-visitors in cluster 1 are not likely to make purchases since they have much greater exit and bounce rates (this is deduced from the previous sections' results on prediction and feature importance). In the unlikely case where site-visitors in cluster 1 do make a purchase, they will have a very high page value, because they do not visit many websites before making a purchase. The few site-visitors who do make purchases in the cluster 1 may be the outliers represented in the Page Value box-and-whisker plot. But overall, as shown by the interquartile range, cluster 2 has a higher Page Value.

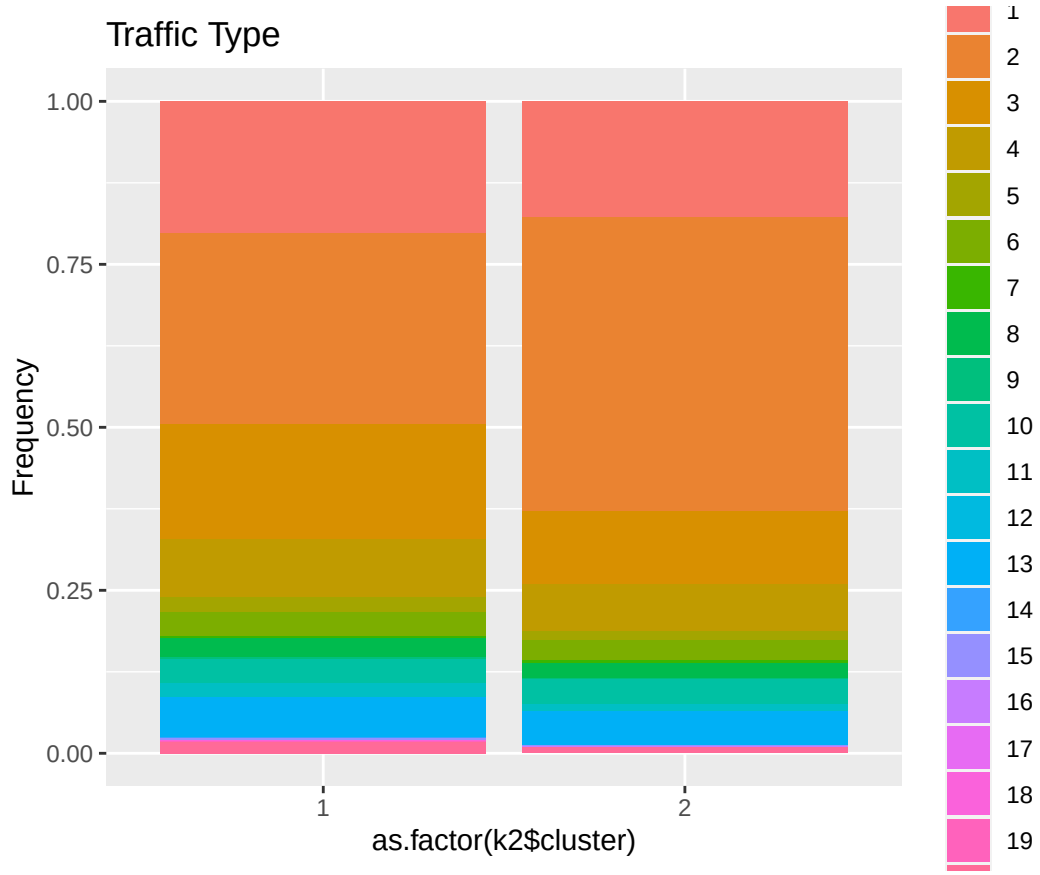
The observations made thus far seem to suggest that the site-visitors in cluster 2 are more likely to make purchases than the site-visitors in cluster 1.



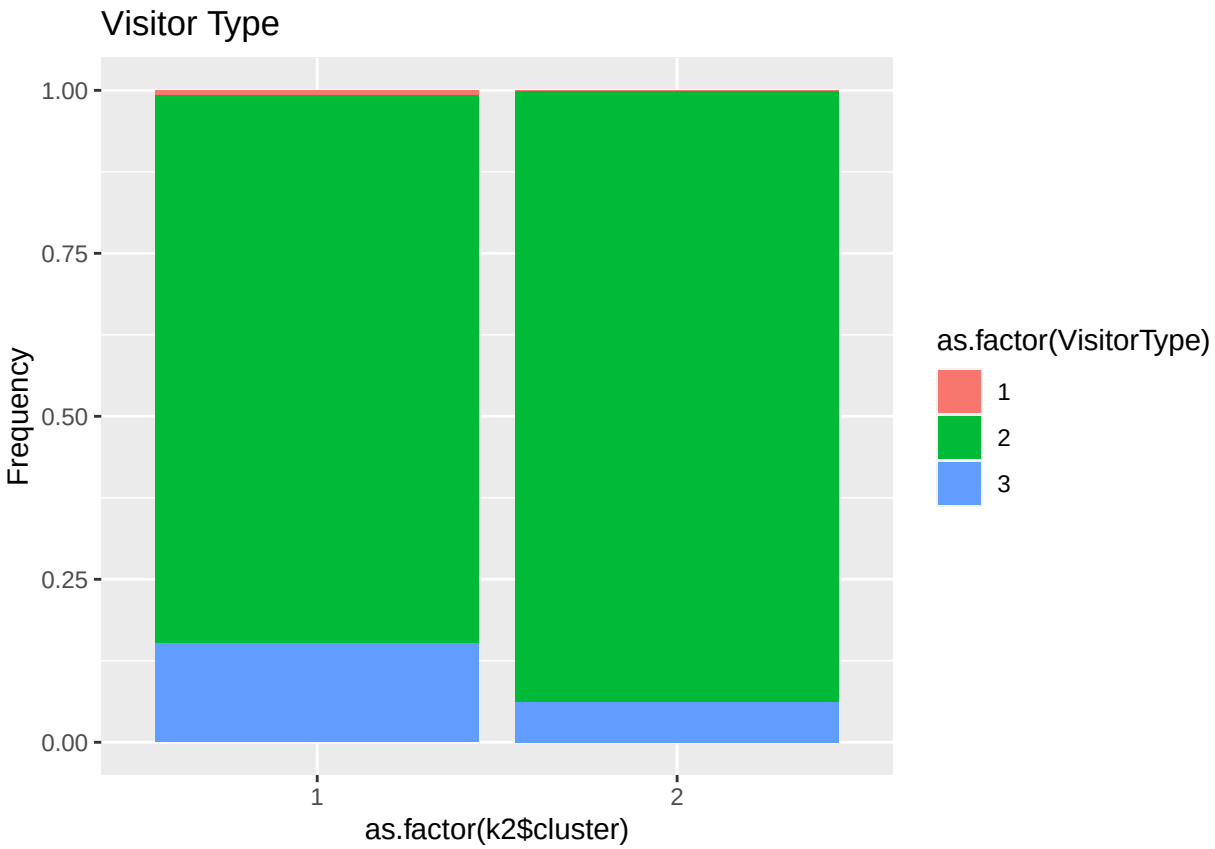
Another interesting behaviour trend among the site-visitors in cluster 2 (the ones with higher likelihood of making purchases) is that they are likely to shop much more in November than site-visitors in cluster 1. From this, one can speculate that site-visitors who are frequent buyers on e-commerce websites, are more likely to prepare for winter holidays seasons by shopping for gifts. In addition, cluster 2 site-visitors are more likely to shop on weekends compared to those in cluster 1.



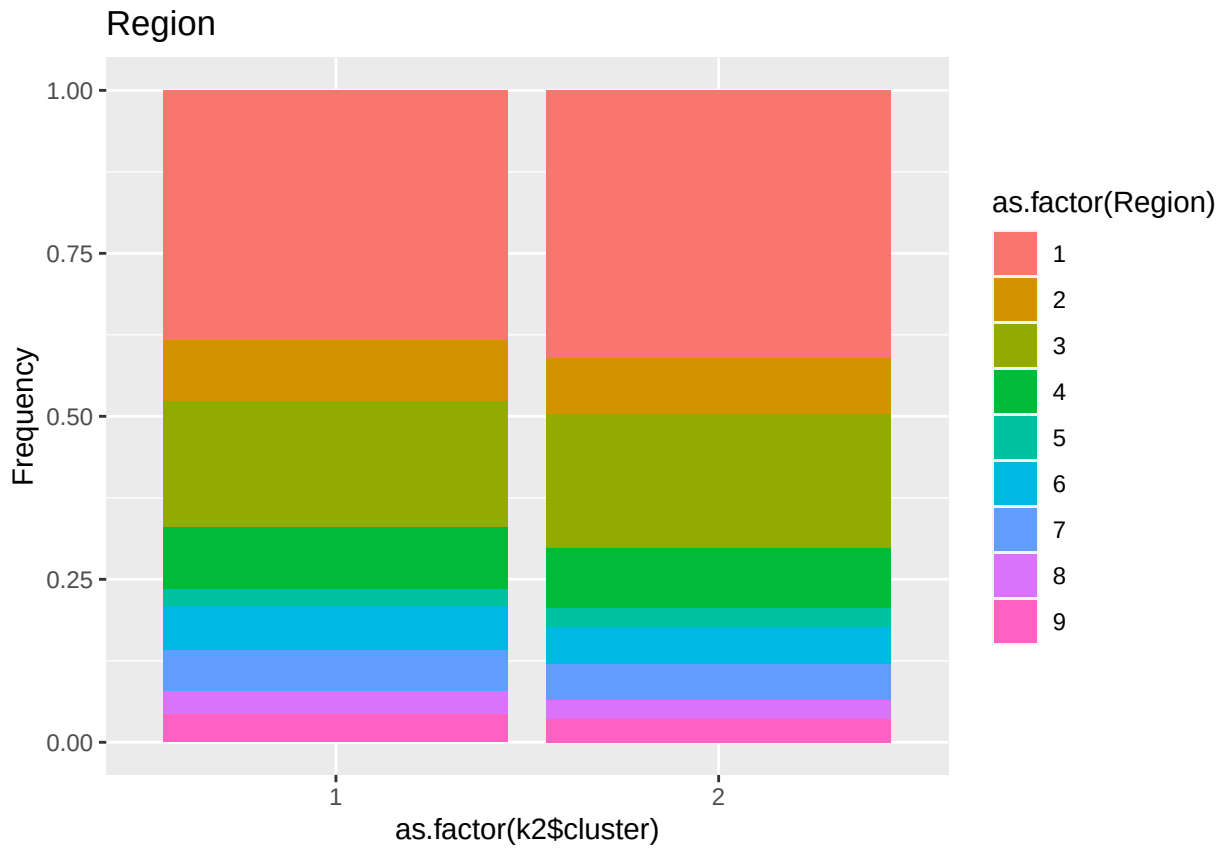
Also, site-visitors in different clusters use different browser types (e.g. Google Chrome, Safari) and computer operating systems (e.g. Mac OS, Windows). Given that the Sakar et al. paper does not define what each numerical value represent in the browser and operating system columns, there is no way of knowing which specific browser and operating system types are popular among cluster 1 and 2; but nevertheless, it's an interesting insight to observe that site-visitors in different clusters have different usage preference in browser and computer operating system.



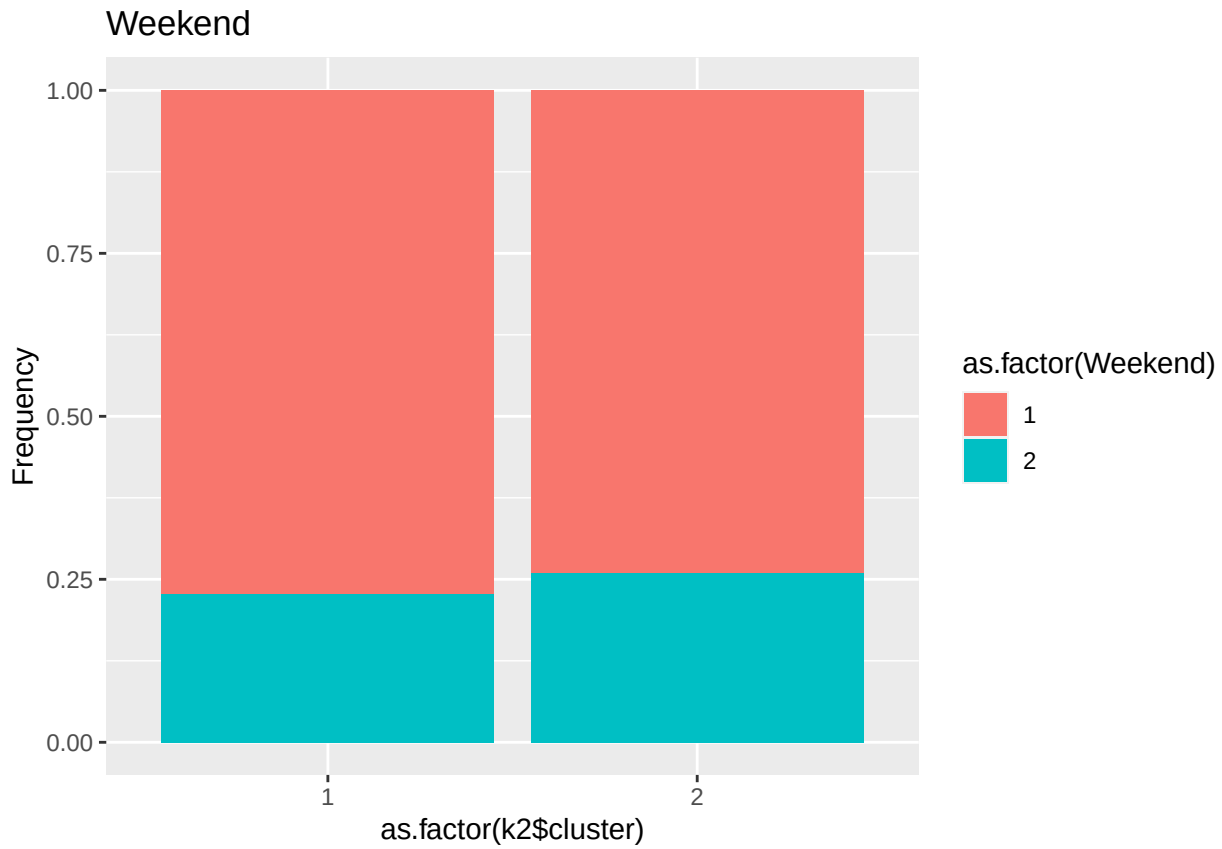
Another interesting observation was made in traffic types. The different traffic type values represent how the visitor arrived at the e-commerce website (e.g. through SMS messaging, email, search engines, etc). Different clusters showed visibly different routes of traffic type. Again, Sakar et al paper does not define the numerical values in the traffic type column, but this still provides an interesting insight that site-visitors who are more likely to make purchases might get triggered to visit e-commerce websites by certain prompts (e.g. website link in the promotional emails) more than the site-visitors who are unlikely to make purchases, which is one of the theories that can explain the different traffic types of the clusters.



In both clusters, there were barely any returning visitors ("1") to the website. Most common visitor type was the new visitor in both clusters, but cluster 1 was more likely to have "Other" visitor type, where cluster 2 was more likely to have "New" visitor type.



While the clusters showed different behaviours in certain columns, there were columns where the clusters did not differ much from each other. Both clusters displayed fairly equal representation for the region. This indicates that the region and location is not significant in characterizing e-commerce visitors.



The clusters were also quite similar for the weekend shopping as well, although the cluster 2 was more likely to make a purchase on weekends.

The following tables show the K centers of cluster 1 and 2, respectively.

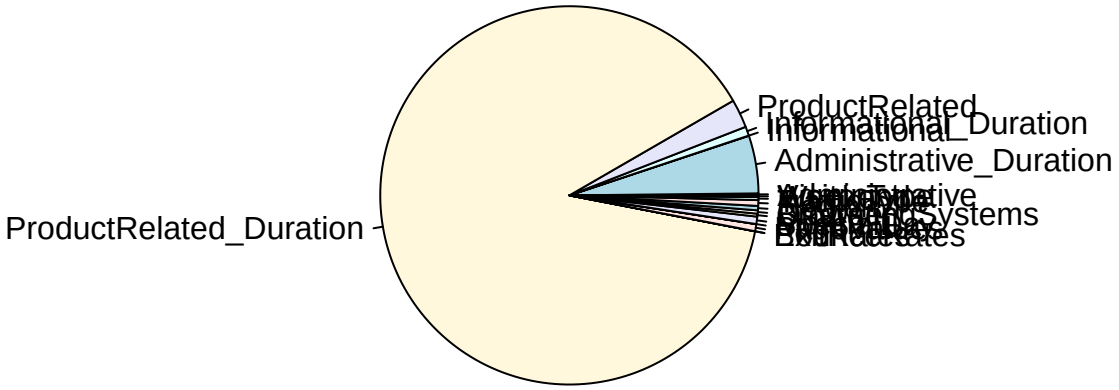
```
##          Administrative Administrative_Duration           Informational
##            -0.266915618            -0.220480434            -0.256282829
##  Informational_Duration          ProductRelated ProductRelated_Duration
##            -0.197974562            -0.249161939            -0.231795334
##              BounceRates              ExitRates              PageValues
##              0.061125977             0.090894032            -0.041449262
##               SpecialDay                   Month        OperatingSystems
##              0.032012771            -0.069664171             0.003676173
##                  Browser                  Region             TrafficType
##              0.015158660             0.012015980             0.021688750
##              VisitorType                 Weekend
##              0.036772401            -0.011871646

##          Administrative Administrative_Duration           Informational
##              1.43477496              1.18516784              1.37761959
##  Informational_Duration          ProductRelated ProductRelated_Duration
##              1.06419005              1.33934205              1.24598981
##              BounceRates              ExitRates              PageValues
##              -0.32857583            -0.48859067              0.22280586
##               SpecialDay                   Month        OperatingSystems
##              -0.17208106             0.37447194             -0.01976085
```

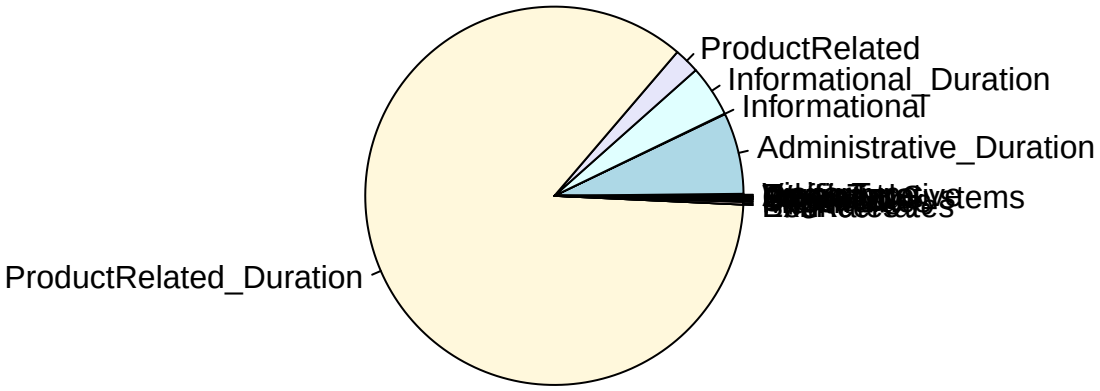
```
##                 Browser                  Region             TrafficType
##             -0.08148368             -0.06459055             -0.11658544
##             VisitorType                 Weekend
##             -0.19766592              0.06381470
```

Moreover, the following shows the differences between the clusters' k-means centers.

```
##          Administrative           Informational          ProductRelated
##              -1.70169057             -1.63390242             -1.58850398
## ProductRelated_Duration Administrative_Duration  Informational_Duration
##              -1.47778515             -1.40564827             -1.26216461
##                ExitRates                   Month             BounceRates
##               0.57948470             -0.44413611              0.38970181
##               PageValues             VisitorType              SpecialDay
##              -0.26425512              0.23443832              0.20409383
##             TrafficType                 Browser                  Region
##               0.13827420              0.09664234              0.07660653
##                 Weekend        OperatingSystems
##             -0.07568634              0.02343703
```

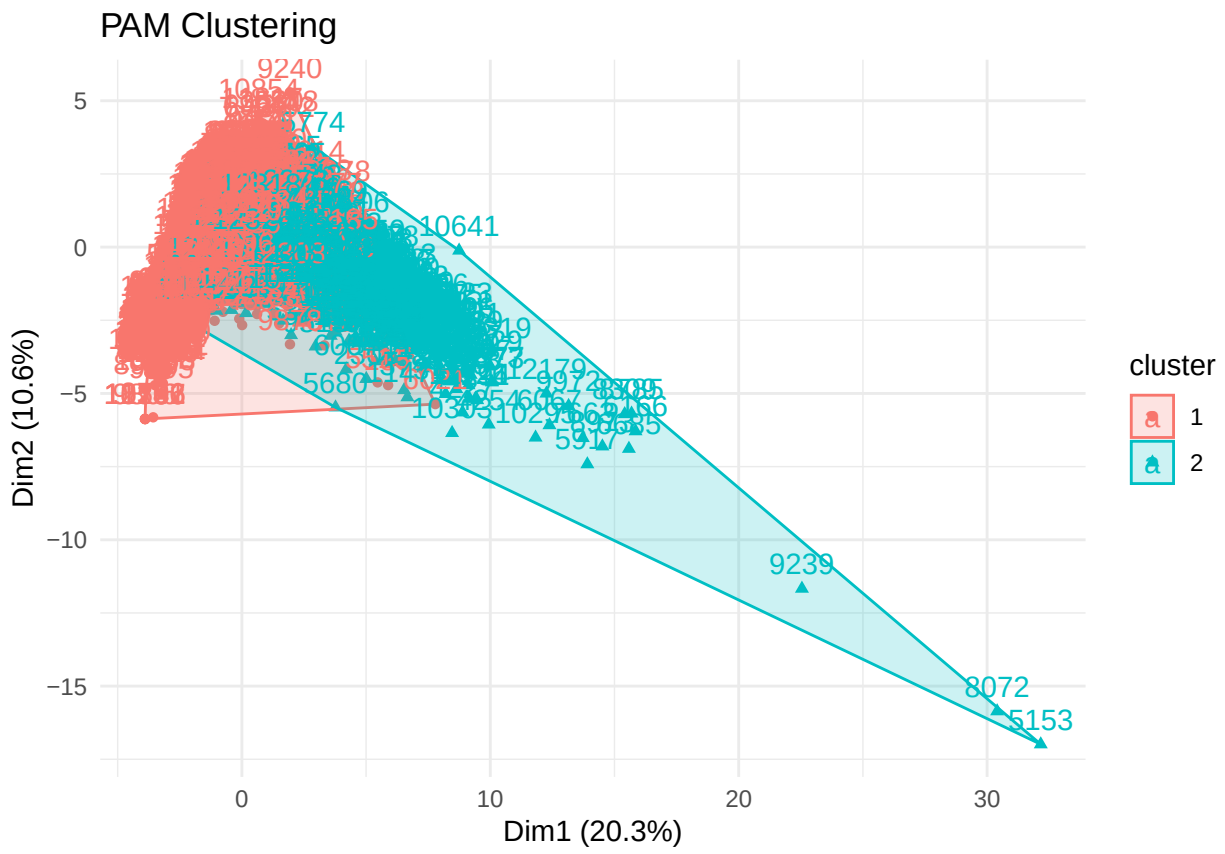
The following pie charts show the cluster representation of 1 and 2, respectively. The pie charts indicate that the major interest of e-commerce visitors seems to be product related duration, which is the amount of time spent by the visitors viewing product related web pages.





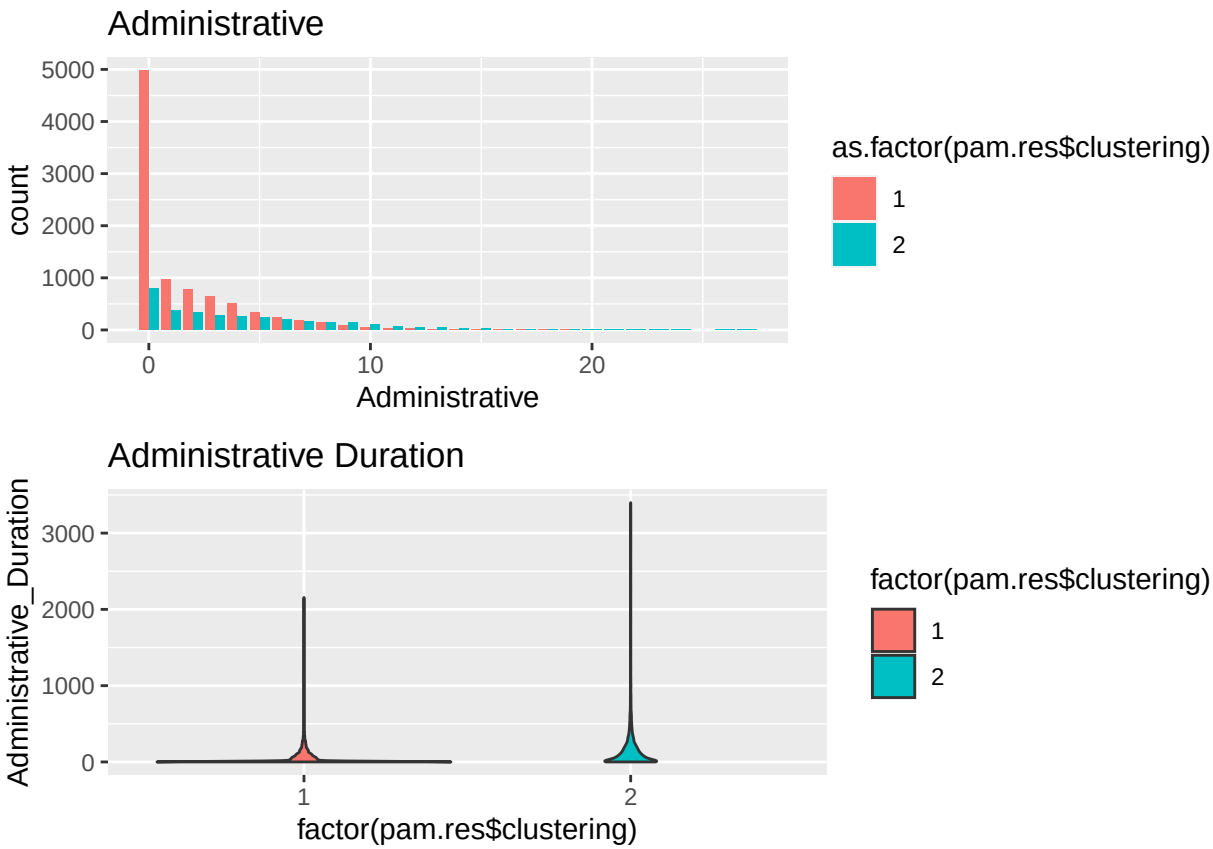
In addition to the K-means clustering, the PAM model was also used, as this algorithm searches for the k representative objects or medoids in the dataset (Kassambara, n.d.). Once a set of k medoids are found, clusters are made by allocating each observation to the closest medoid. Next, the objective function is determined by swapping the selected medoid and non-medoid data point (Kassambara, n.d.).

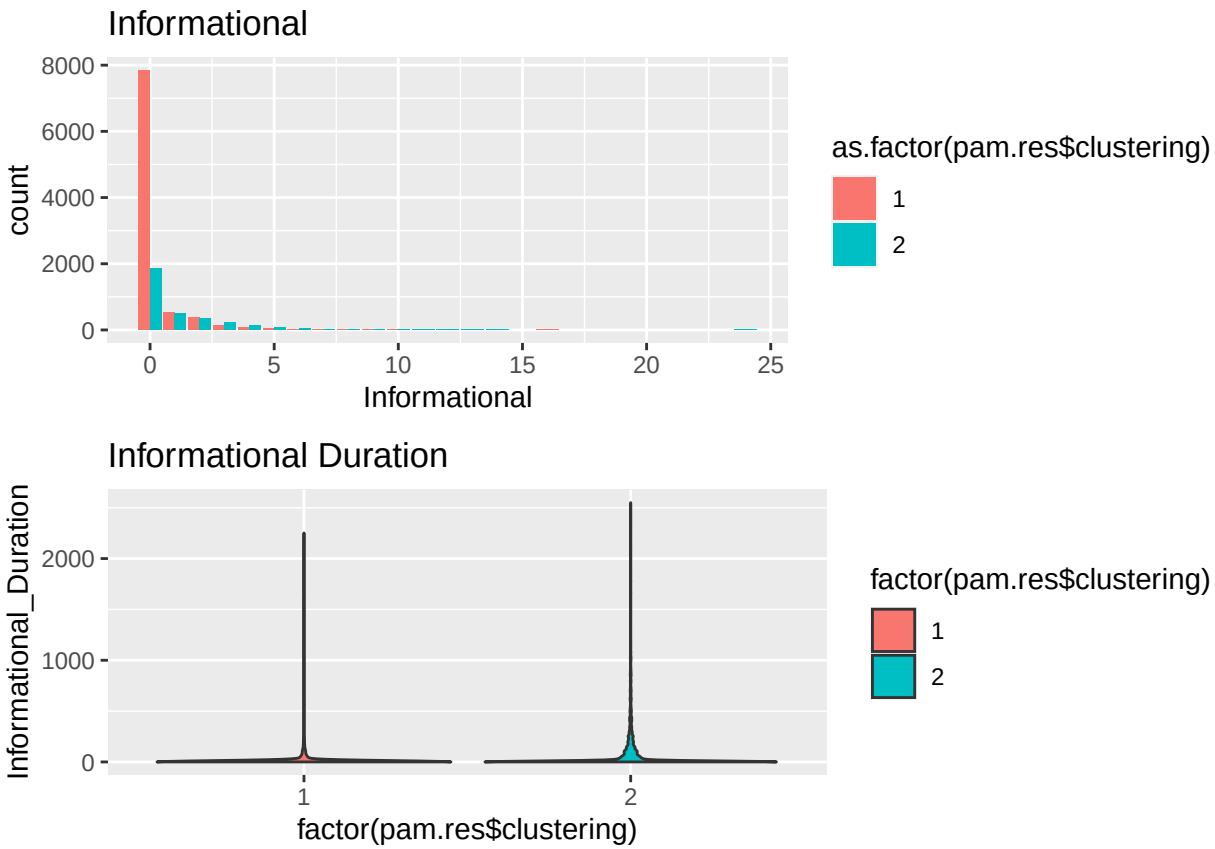
The following graph shows how PAM divided the site-visitors into 2 clusters, in which 9,039 site-visitors were in the first cluster, and 3,291 site-visitors in the second cluster.

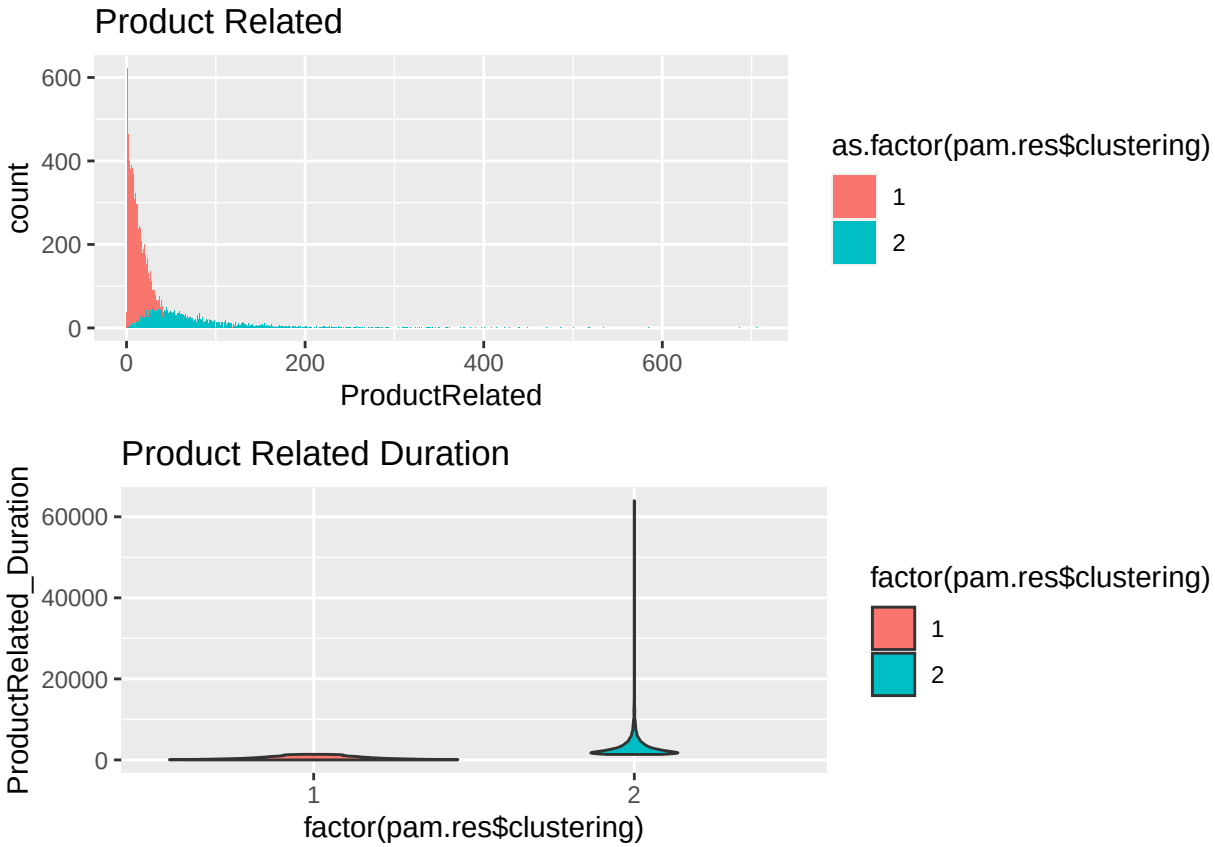


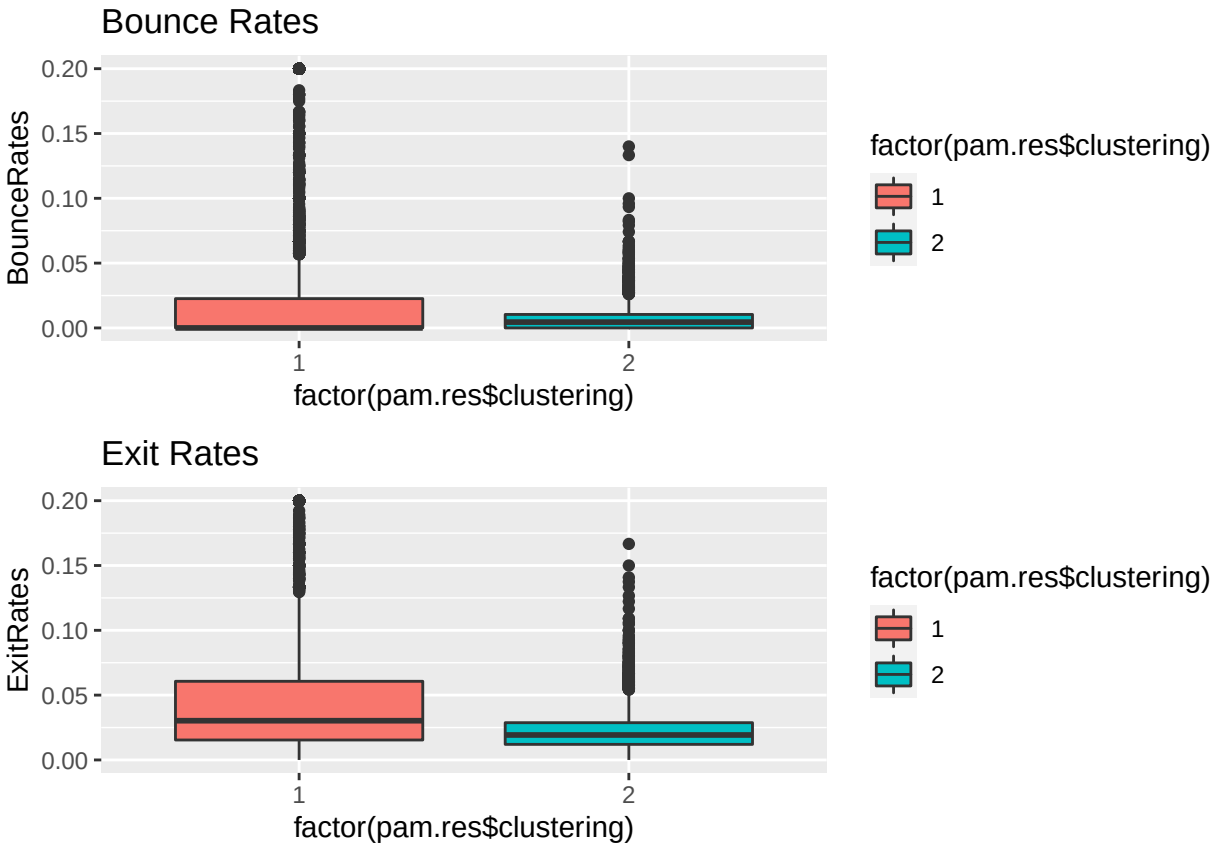

```
##
##    1    2
## 9039 3291
```

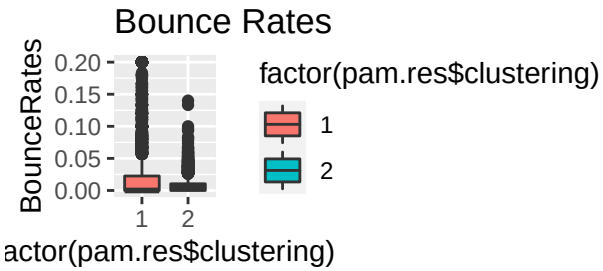
Much of the observation made in the PAM model was similar as was in the K-means model: cluster 2 visitors are more likely to make a purchase by visiting more administrative, informational, and product-related pages, and spending more time overall in the pages visited, had lower bounce and exit rates, had a higher page value in the interquartile range, were more likely to shop around November, used different browser and operating system, and arrived at the shopping website through a different traffic channel. Much like the K-means model, clusters were similar in the region, visitor type, and weekend columns. The following pages will show the plots for the PAM model that correspond with the observation provided in this paragraph.

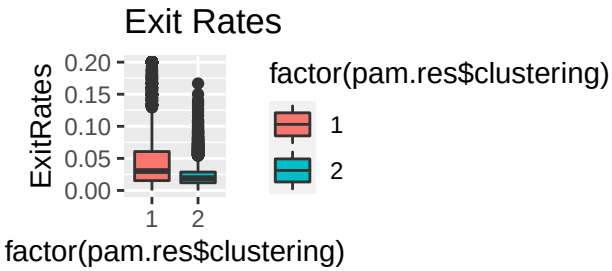

Administrative
5000
4000
3000
2000
1000
0
count
0
10
20
Administrative
as.factor(pam.res$clustering)
1
2
Administrative Duration
3000
2000
1000
0
Administrative_Duration
1
2
factor(pam.res$clustering)
factor(pam.res$clustering)
1
2

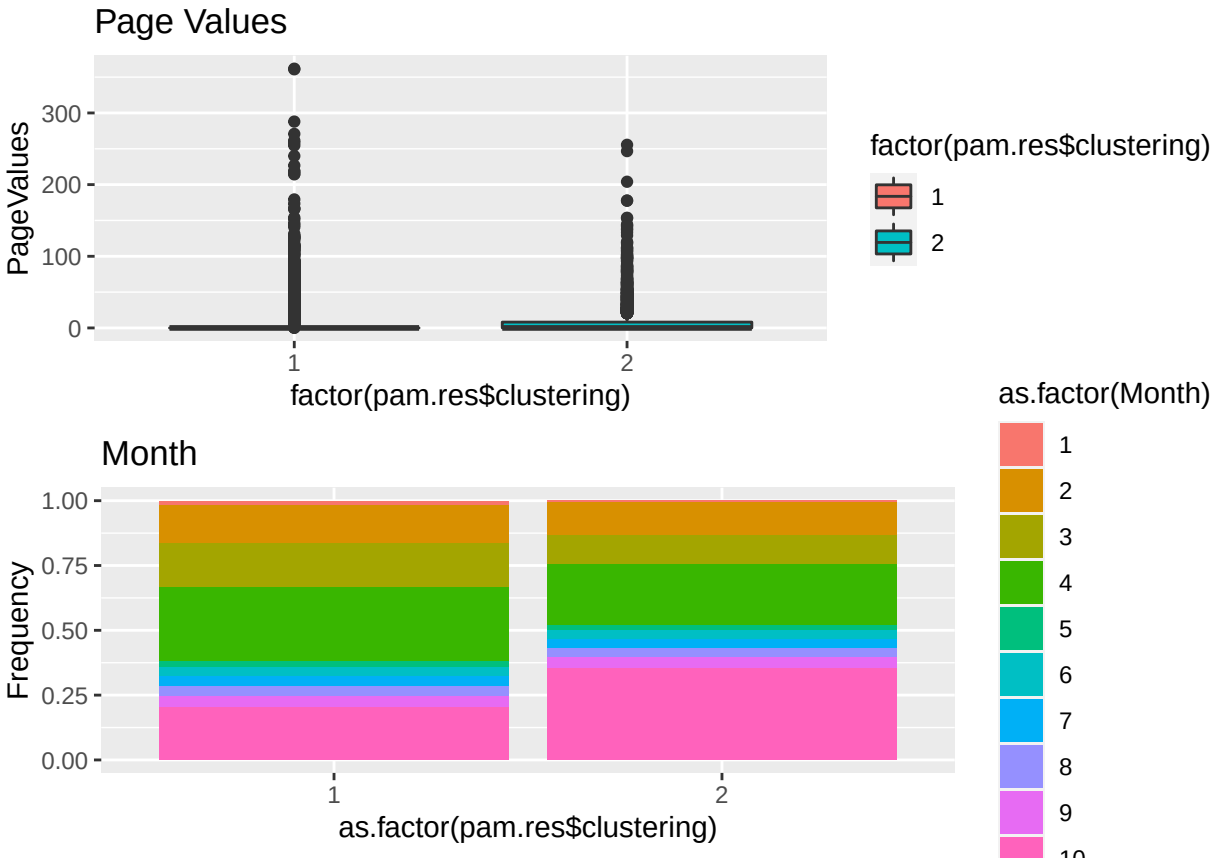

Informational
8000
6000
4000
2000
0
count
0
5
10
15
20
25
Informational
as.factor(pam.res$clustering)
1
2
Informational Duration
2000
1000
0
Informational_Duration
1
2
factor(pam.res$clustering)
factor(pam.res$clustering)
1
2

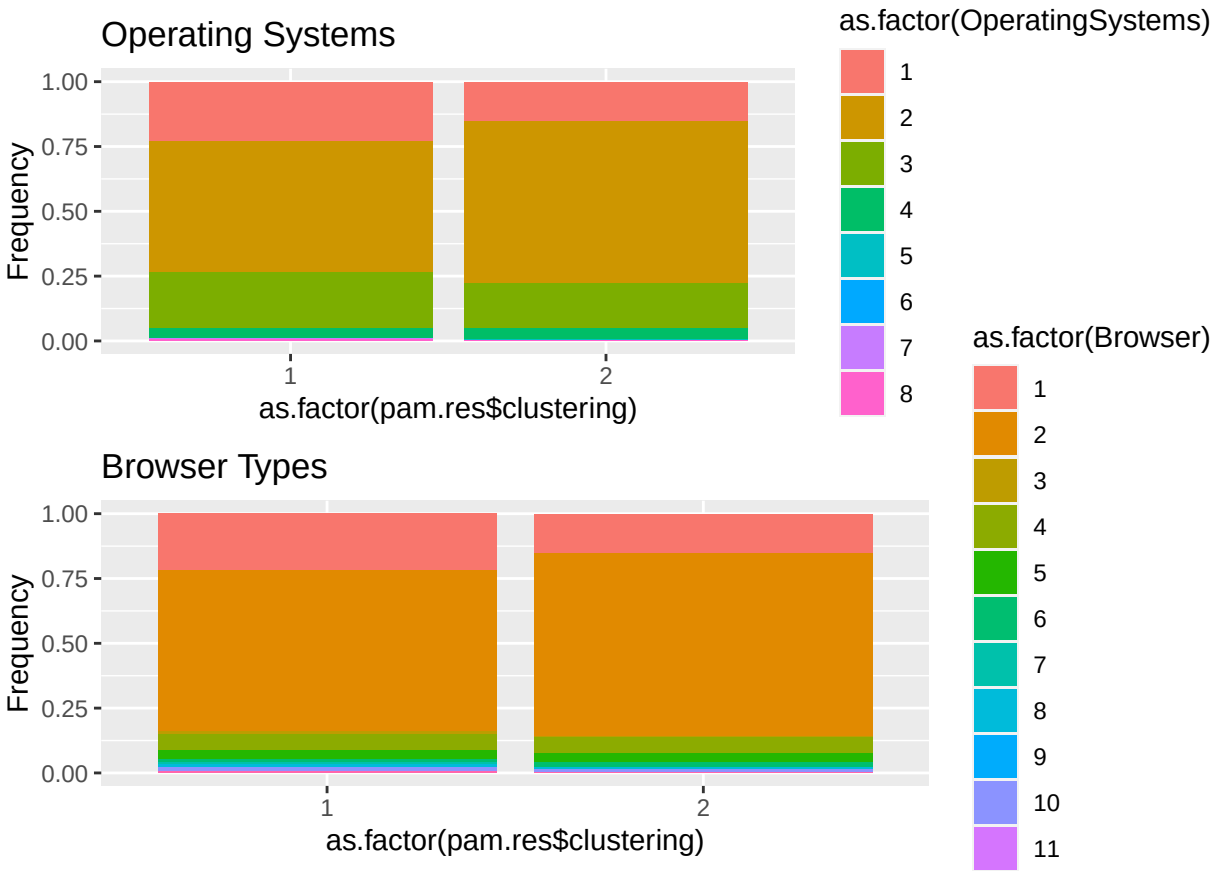

Product Related
600
400
200
0
count
0
200
400
600
ProductRelated
as.factor(pam.res$clustering)
1
2
Product Related Duration
60000
40000
20000
0
ProductRelated_Duration
1
2
factor(pam.res$clustering)
factor(pam.res$clustering)
1
2

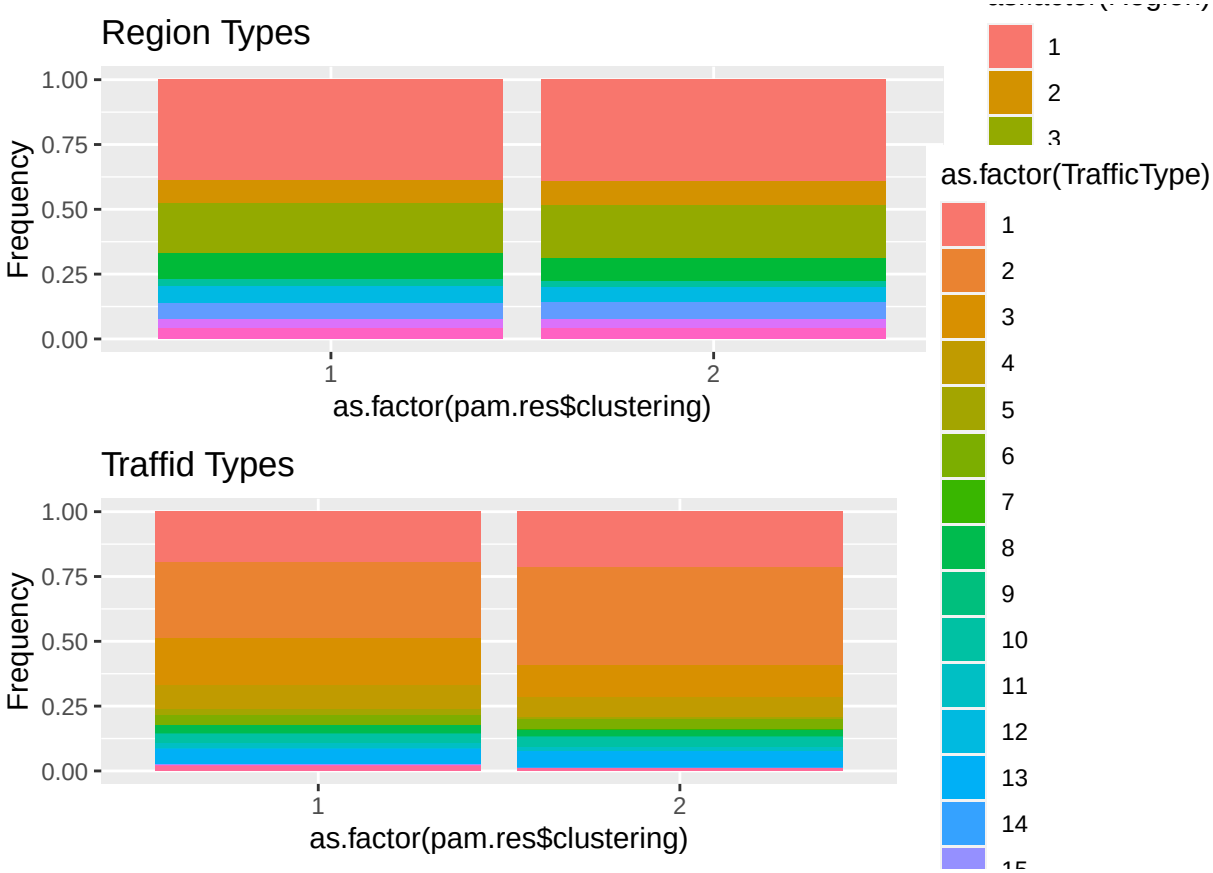

Bounce Rates
BounceRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(pam.res$clustering)
factor(pam.res$clustering)
1
2
Exit Rates
ExitRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(pam.res$clustering)
factor(pam.res$clustering)
1
2

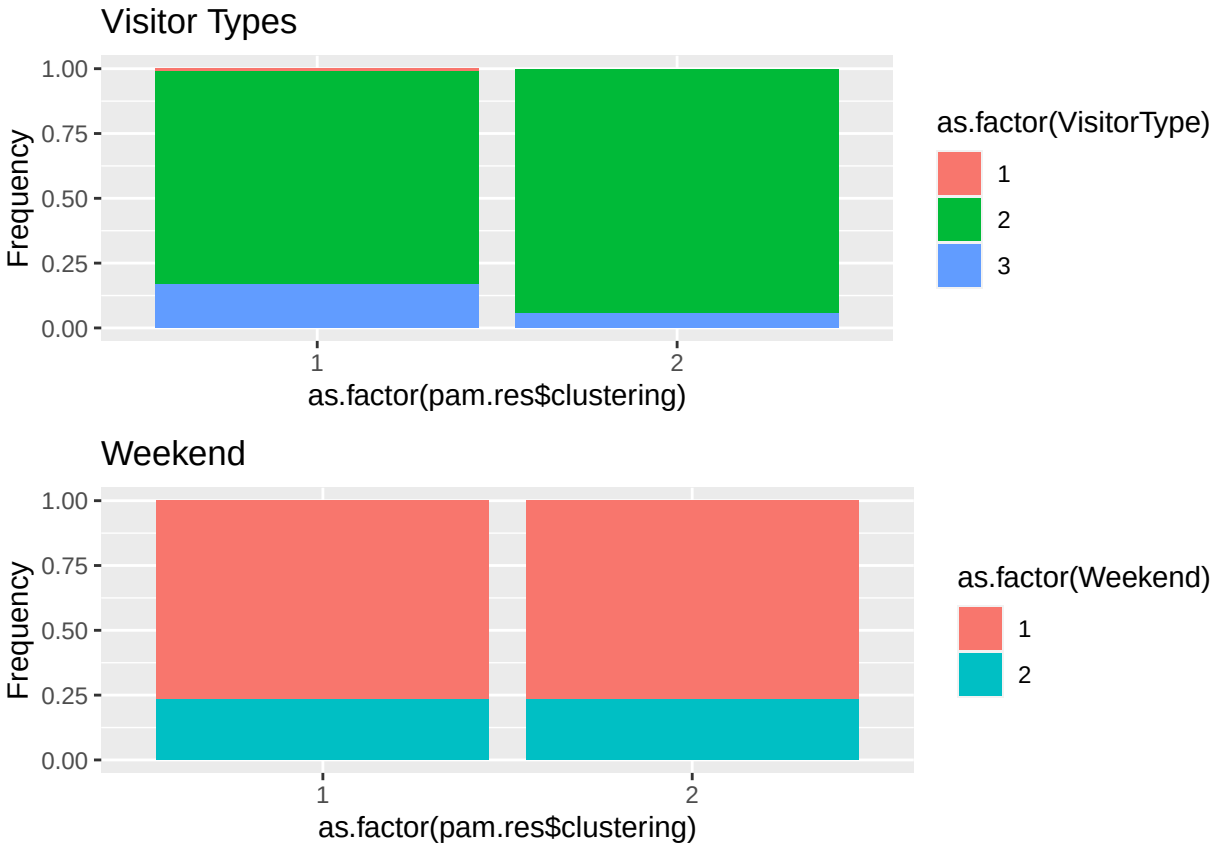

Bounce Rates
BounceRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(pam.res$clustering)
1
2
actor(pam.res$clustering)


Exit Rates
ExitRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(pam.res$clustering)
1
2
factor(pam.res$clustering)


Page Values
PageValues
300
200
100
0
1
2
factor(pam.res$clustering)
factor(pam.res$clustering)
1
2
Month
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(pam.res$clustering)
as.factor(Month)
1
2
3
4
5
6
7
8
9


Operating Systems
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(pam.res$clustering)
as.factor(OperatingSystems)
1
2
3
4
5
6
7
8
Browser Types
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(pam.res$clustering)
as.factor(Browser)
1
2
3
4
5
6
7
8
9
10
11


Region Types
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(pam.res$clustering)
Traffid Types
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(pam.res$clustering)
1
2
3
as.factor(TrafficType)
1
2
3
4
5
6
7
8
9
10
11
12
13
14

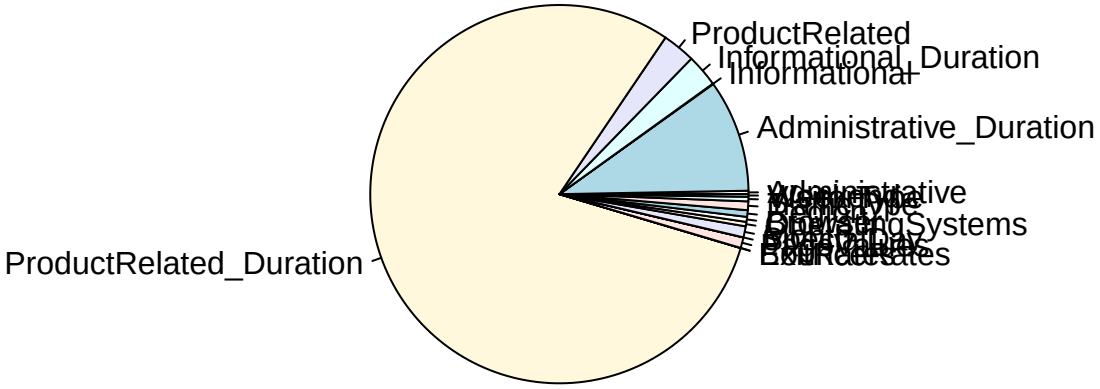


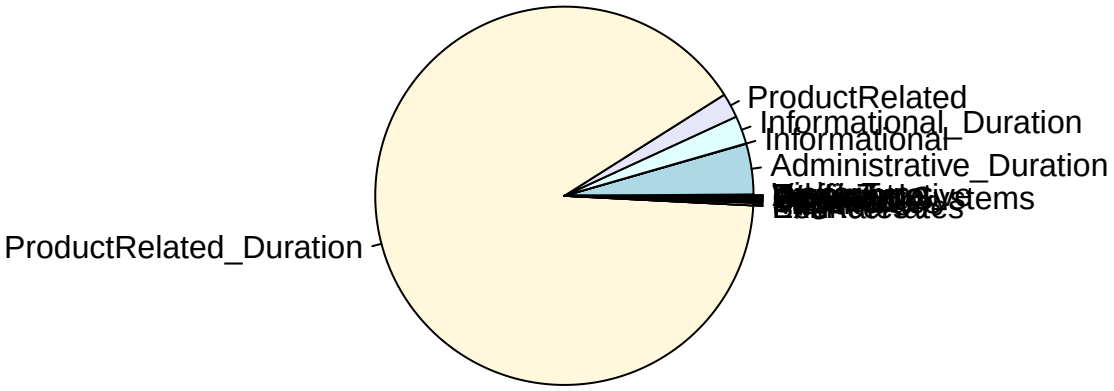
The following shows the medoids of cluster 1 and 2, respectively, in the PAM model.

```
##          Administrative Administrative_Duration           Informational
##              1.00000000              31.00000000              0.00000000
##  Informational_Duration          ProductRelated ProductRelated_Duration
##              0.00000000              10.00000000             347.00000000
##             BounceRates               ExitRates              PageValues
##              0.00000000               0.02222222              0.00000000
##              SpecialDay                    Month        OperatingSystems
##              0.00000000              4.00000000              1.00000000
##                 Browser                   Region            TrafficType
##              1.00000000              1.00000000              1.00000000
##             VisitorType                  Weekend
##              2.00000000              2.00000000

##          Administrative Administrative_Duration           Informational
##            6.000000e+00            9.198485e+01            3.000000e+00
##  Informational_Duration          ProductRelated ProductRelated_Duration
##            5.895238e+01            5.400000e+01            2.407323e+03
##             BounceRates               ExitRates              PageValues
##            3.448276e-03            1.041872e-02            0.000000e+00
##              SpecialDay                    Month        OperatingSystems
##            0.000000e+00            4.000000e+00            3.000000e+00
##                 Browser                   Region            TrafficType
##            2.000000e+00            3.000000e+00            4.000000e+00
##             VisitorType                  Weekend
```

```
##             2.000000e+00            1.000000e+00
```

The pie charts below show the cluster representation of 1 and 2, respectively, in the PAM model. These pie charts are different in PAM compared to the K-means model. In K-means, cluster 1 has a greater product related duration, but in the PAM model, it is cluster 2 that has a greater product related duration. At this point with the current limited dataset, it is difficult to deduce what this signifies.



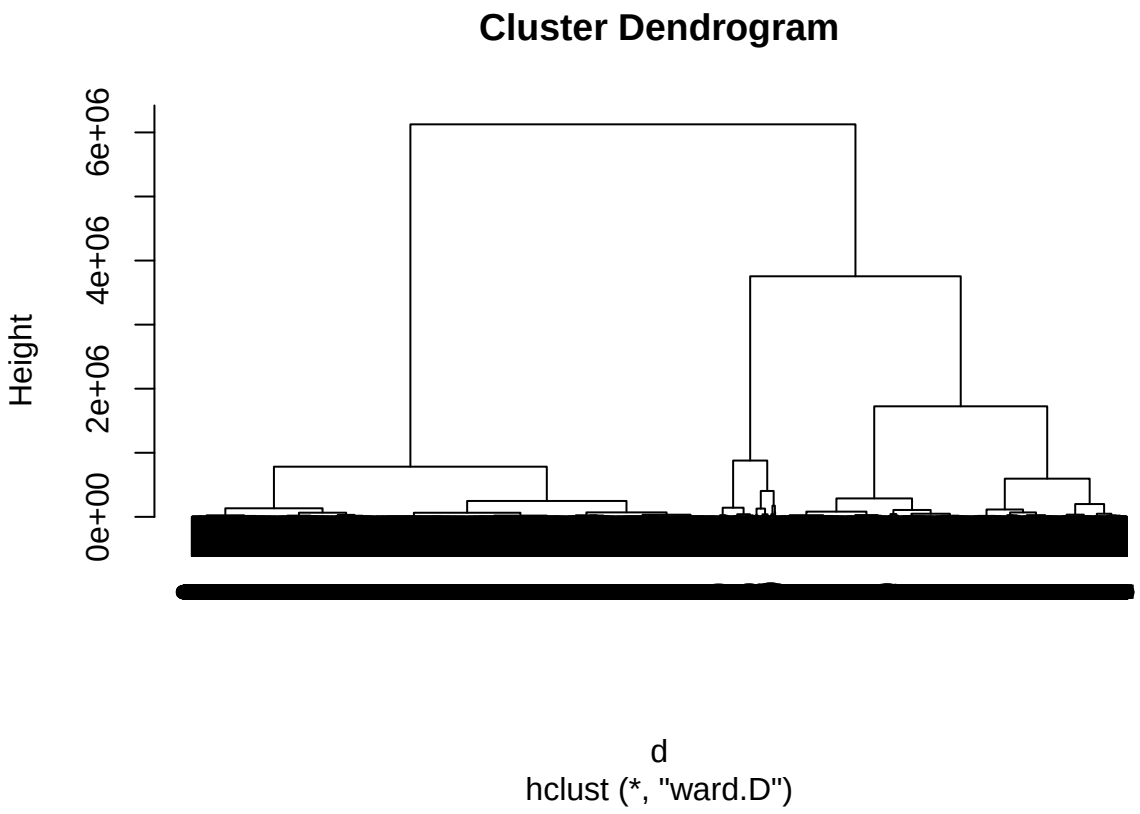


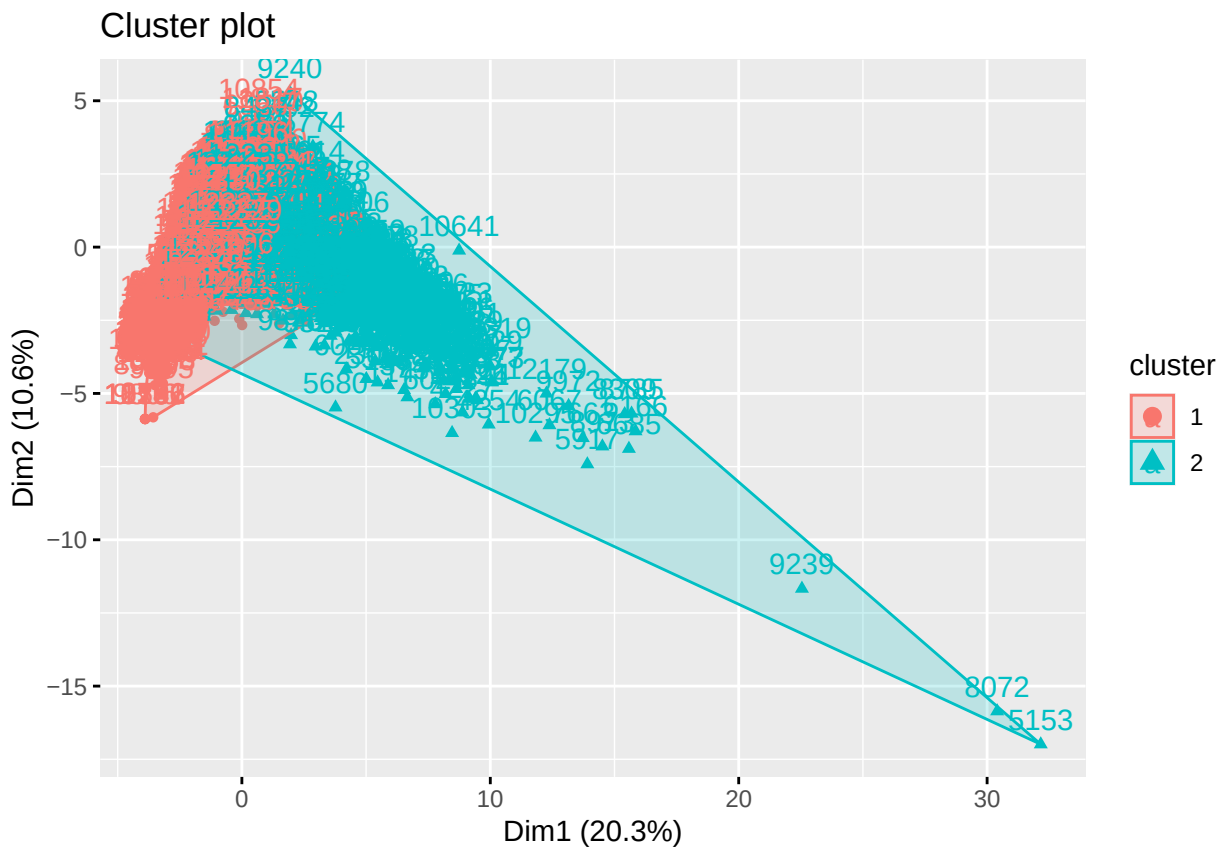
The following table shows the differences between the cluster 1 and 2 medoids.

```
## ProductRelated_Duration Administrative_Duration  Informational_Duration
##           -2.060323e+03           -6.098485e+01           -5.895238e+01
##          ProductRelated          Administrative           Informational
##           -4.400000e+01           -5.000000e+00           -3.000000e+00
##             TrafficType        OperatingSystems                  Region
##           -3.000000e+00           -2.000000e+00           -2.000000e+00
##                 Browser                 Weekend              ExitRates
##           -1.000000e+00            1.000000e+00            1.180350e-02
##             BounceRates              PageValues             SpecialDay
##           -3.448276e-03            0.000000e+00            0.000000e+00
##                  Month             VisitorType
##            0.000000e+00            0.000000e+00
```

Lastly, the hierarchical clustering model was used in order to find a hierarchy of clusters, this hierarchy creates a dendrogram that resembles a tree structure (Kilitcioglu, 2018). The hierarchical cluster technique used is agglomerative clustering in which each data point initiates its own clusters. Then, the two most similar clusters are joined (Kilitcioglu, 2018).

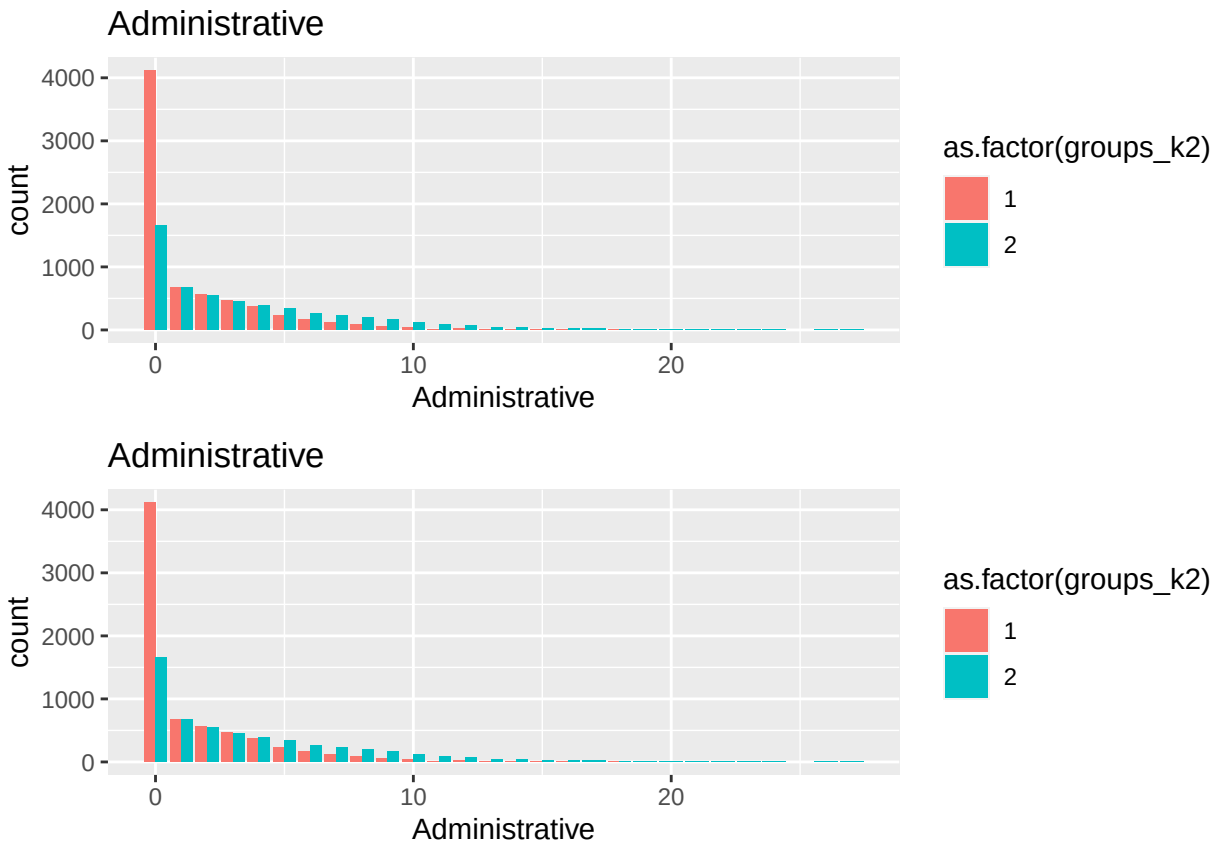
The following graph shows the model's cluster dendrogram and cluster plot. The model had 6,946 site-visitors in cluster 1, and 5,384 site-visitors in cluster 2.

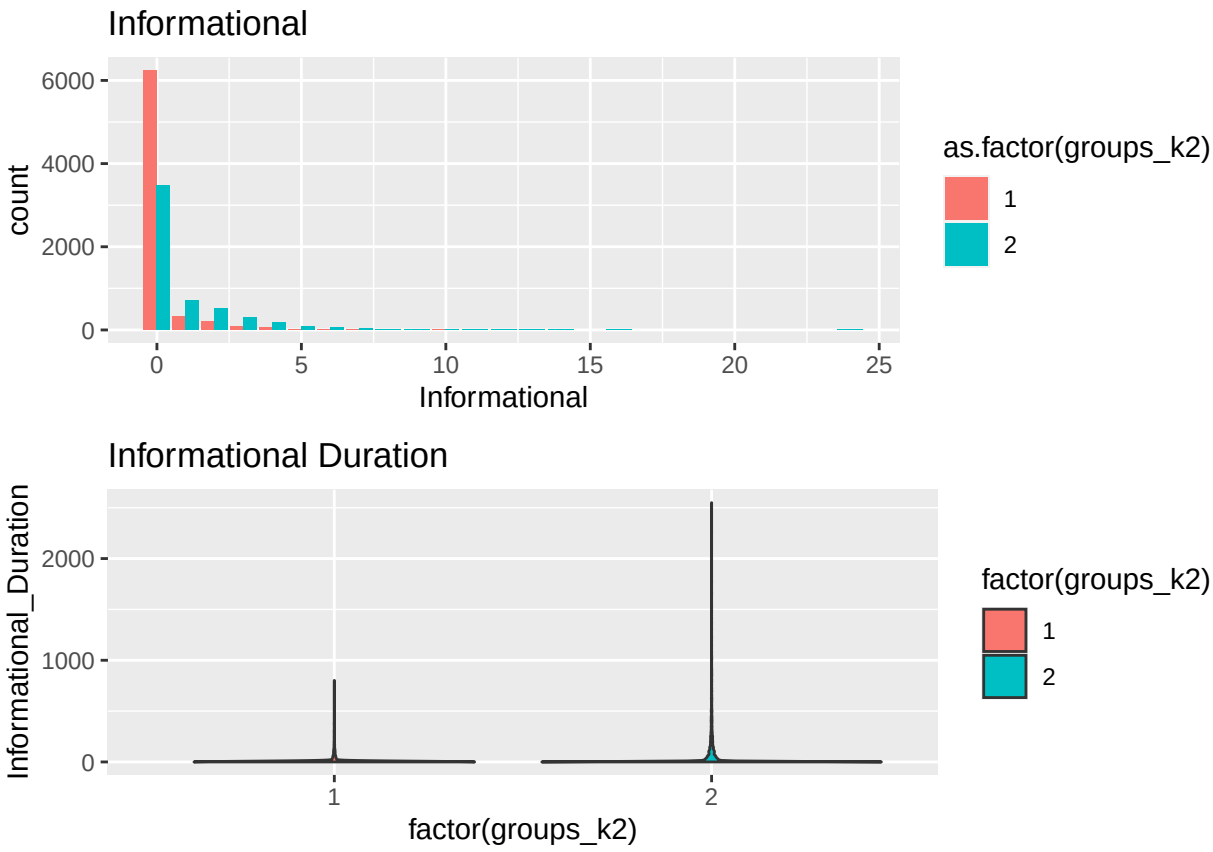

Cluster Dendrogram
Height
0e+00
2e+06
4e+06
6e+06
d
hclust (*, "ward.D")

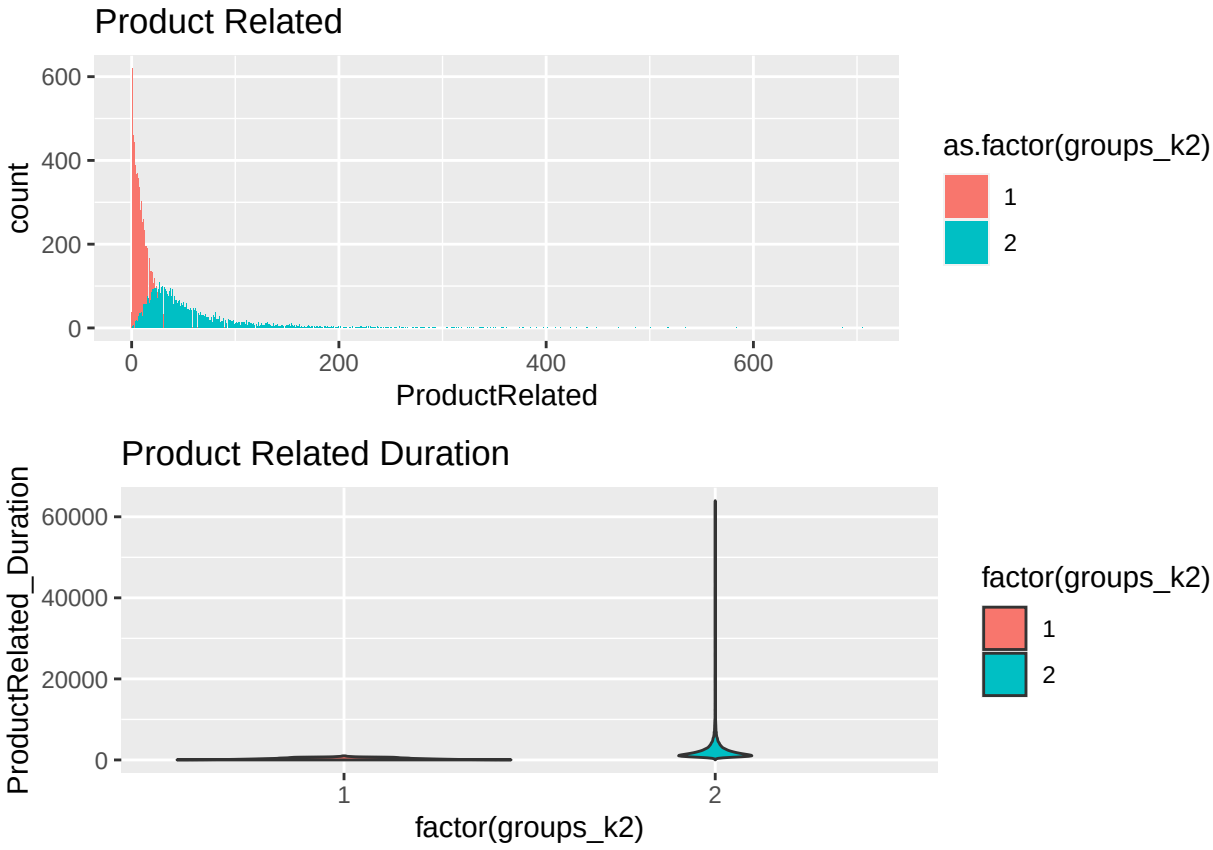


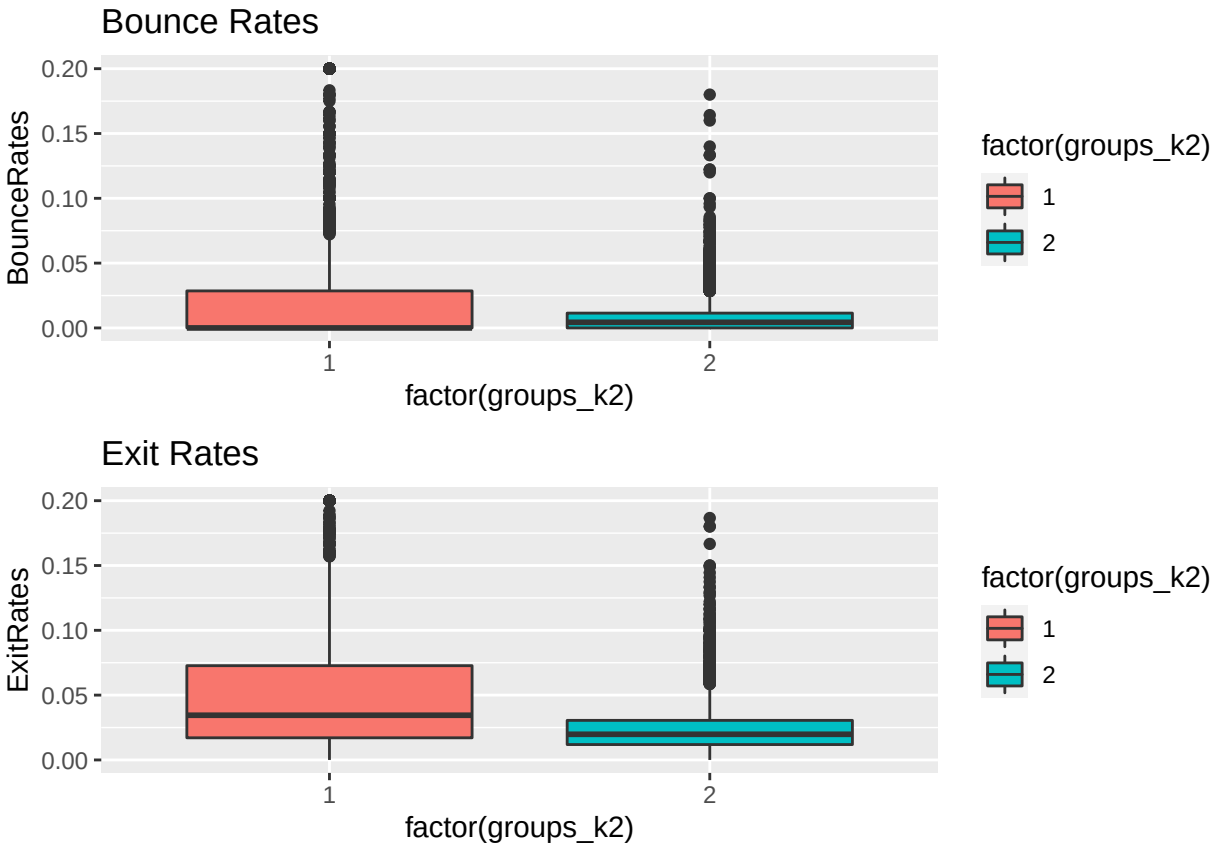

```
##
##    1    2
## 6946 5384
```

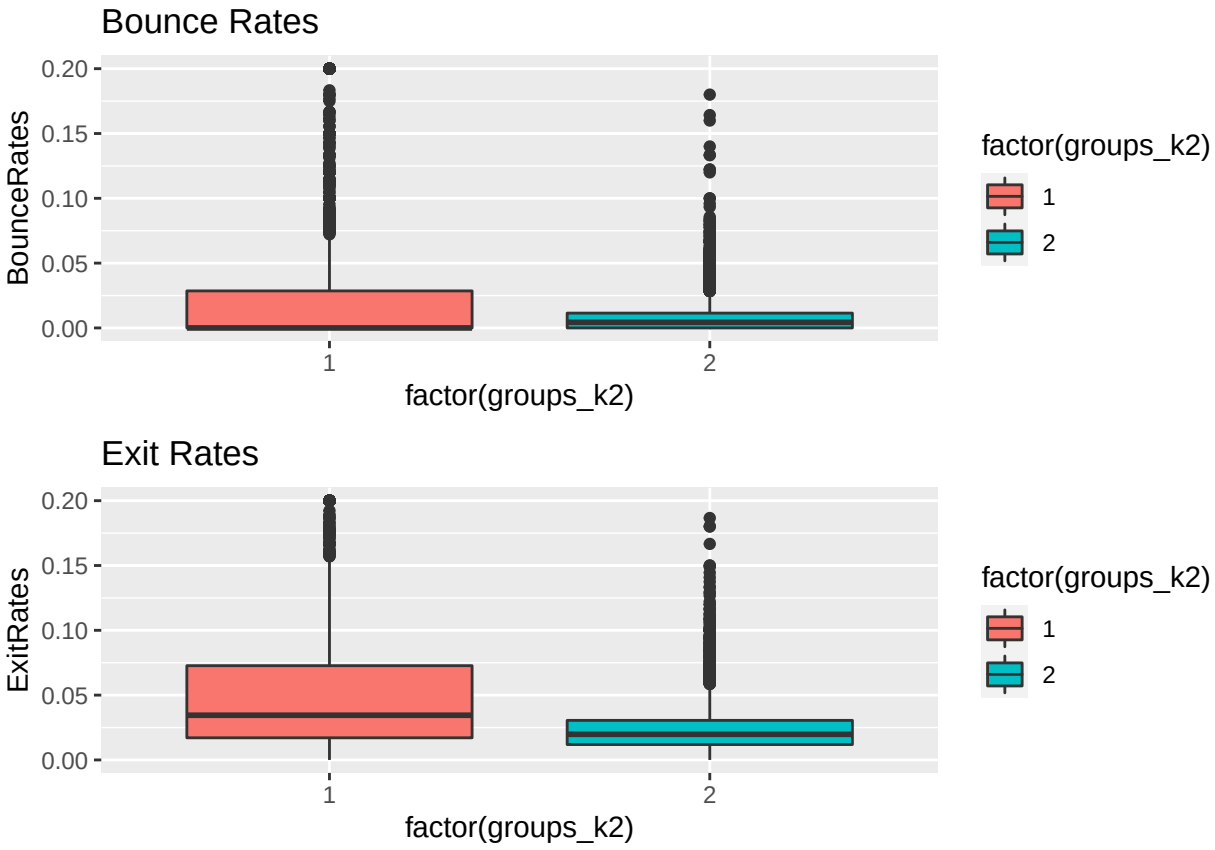
Much of the observation made in hierarchical clustering was also similar to those made for the K-means and PAM clustering: cluster 2 were more likely to make a purchase, spending more time overall in websites, visiting more websites, had lower bounce and exit rates, had higher page value, were more likely to shop around November, used different browser and operating system than cluster 1, arrived at the website through a different traffic type than cluster 1, and were similar to other cluster in terms of region and visitor type. Much like the PAM model, the clusters in the hierarchical model were similar in the weekend shopping. The following pages will show the plots for the hierarchical clustering model that correspond with the observation provided in this paragraph.

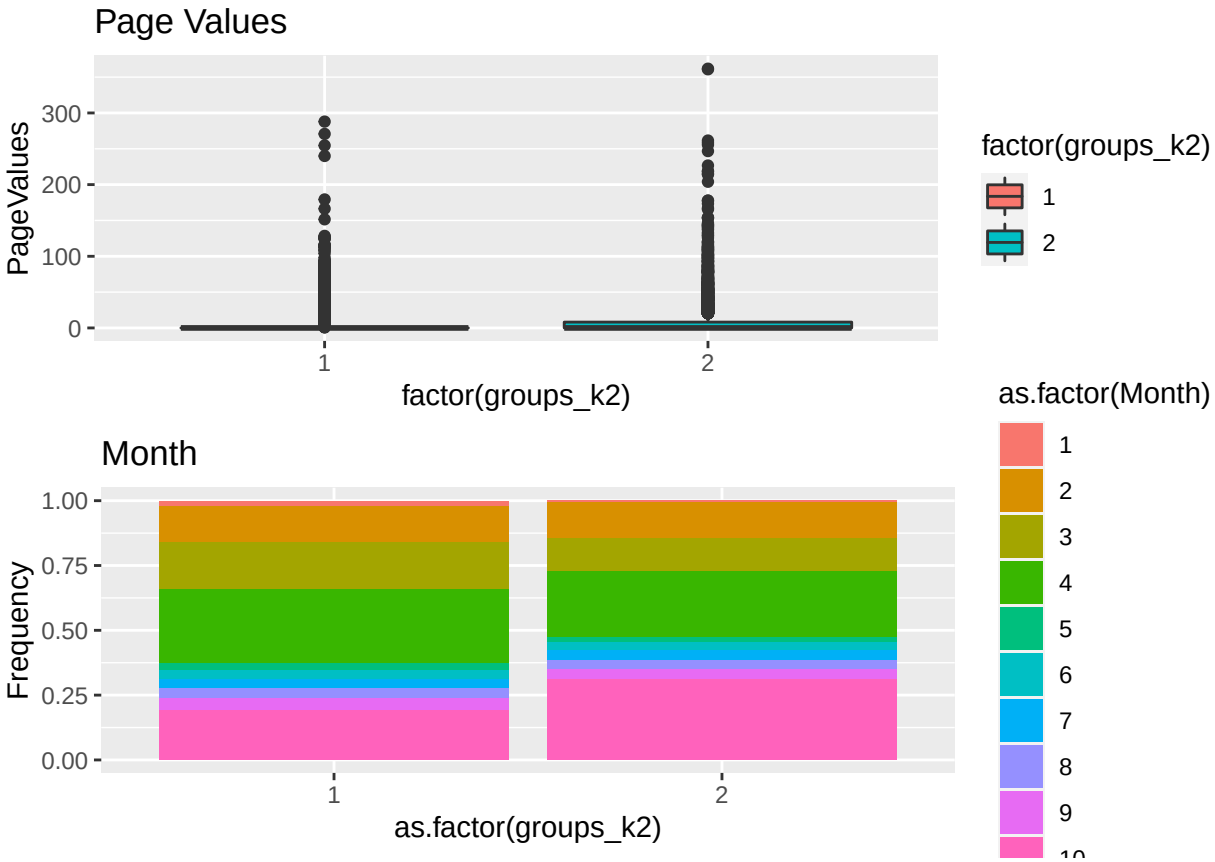

Administrative
4000
3000
2000
1000
0
count
0
10
20
Administrative
as.factor(groups_k2)
1
2
Administrative
4000
3000
2000
1000
0
count
0
10
20
Administrative
as.factor(groups_k2)
1
2

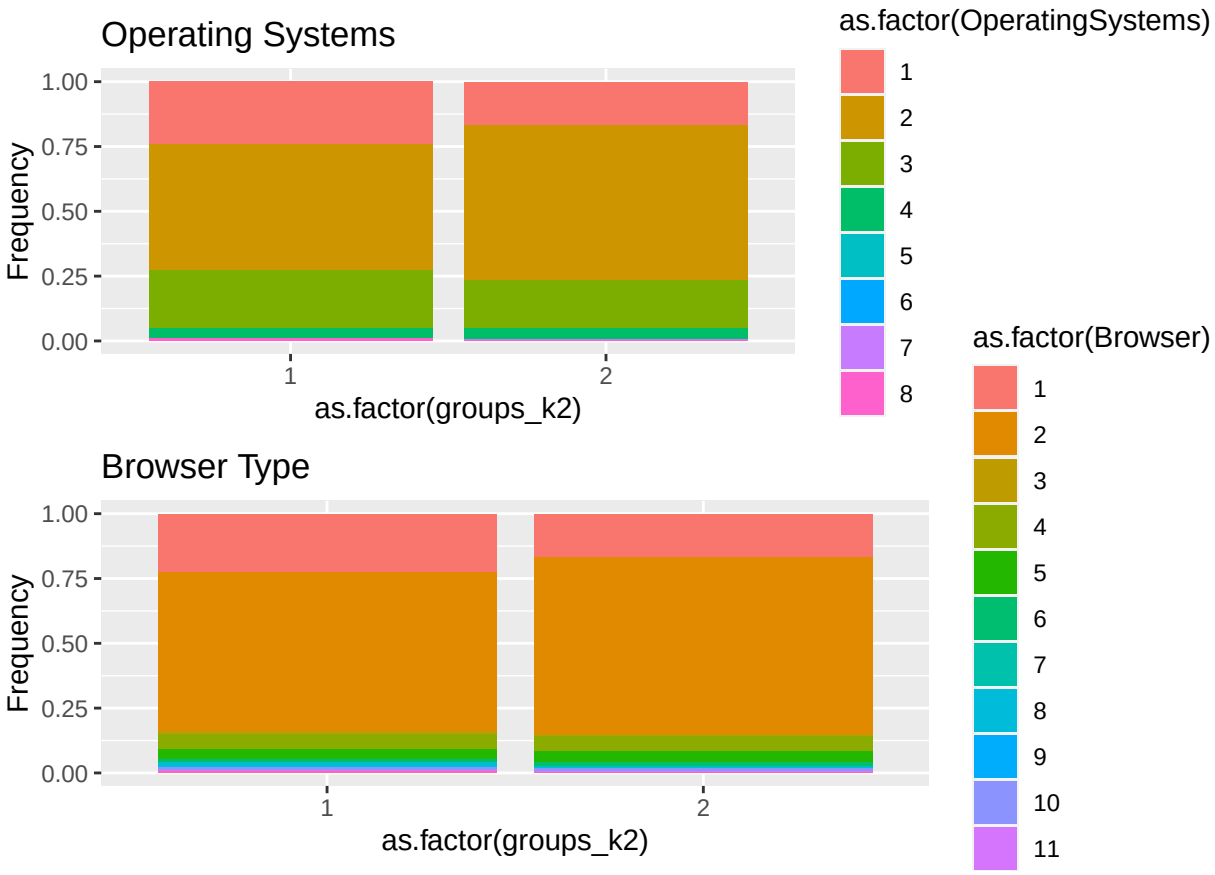

Informational
count
6000
4000
2000
0
0
5
10
15
20
25
Informational
as.factor(groups_k2)
1
2
Informational Duration
Informational_Duration
2000
1000
0
1
2
factor(groups_k2)
factor(groups_k2)
1
2

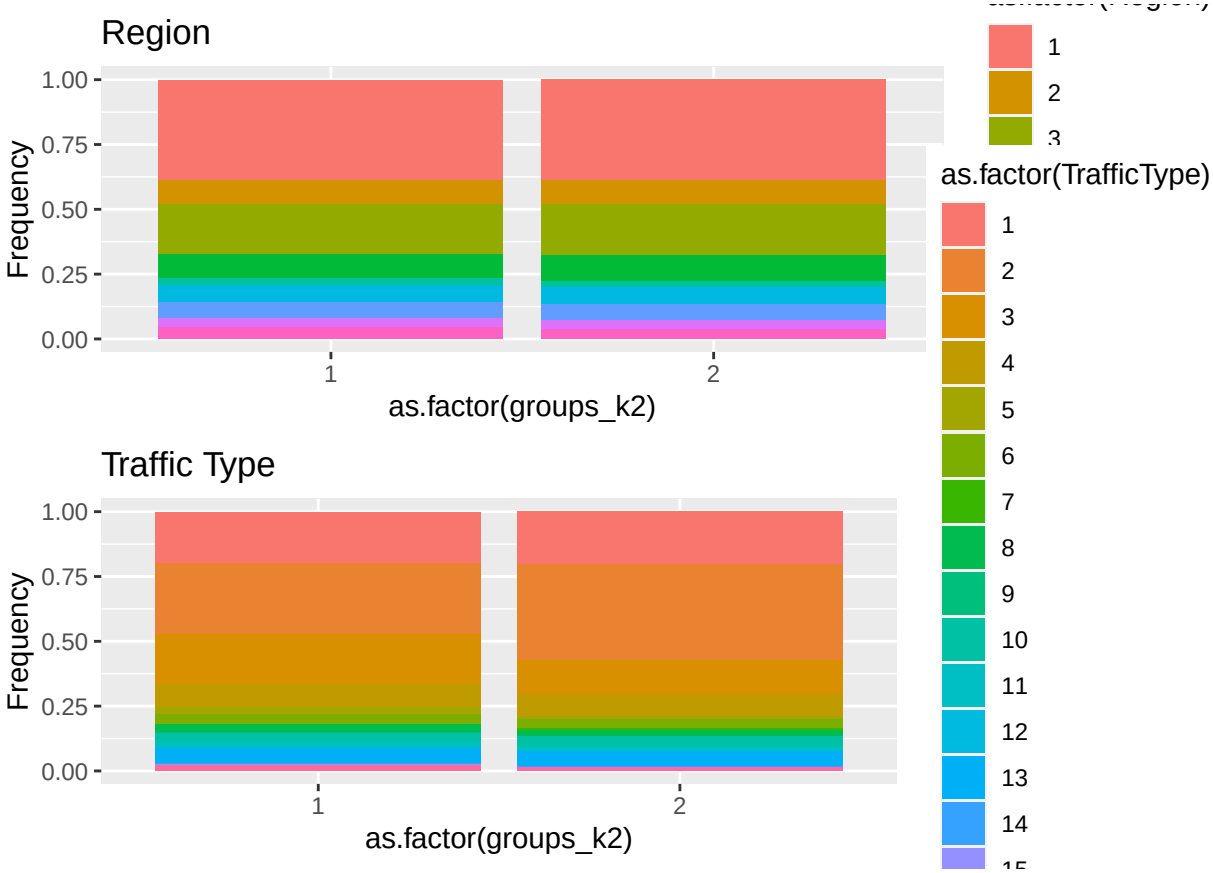

Product Related
600
400
200
0
count
0
200
400
600
ProductRelated
as.factor(groups_k2)
1
2
Product Related Duration
60000
40000
20000
0
ProductRelated_Duration
1
2
factor(groups_k2)
factor(groups_k2)
1
2

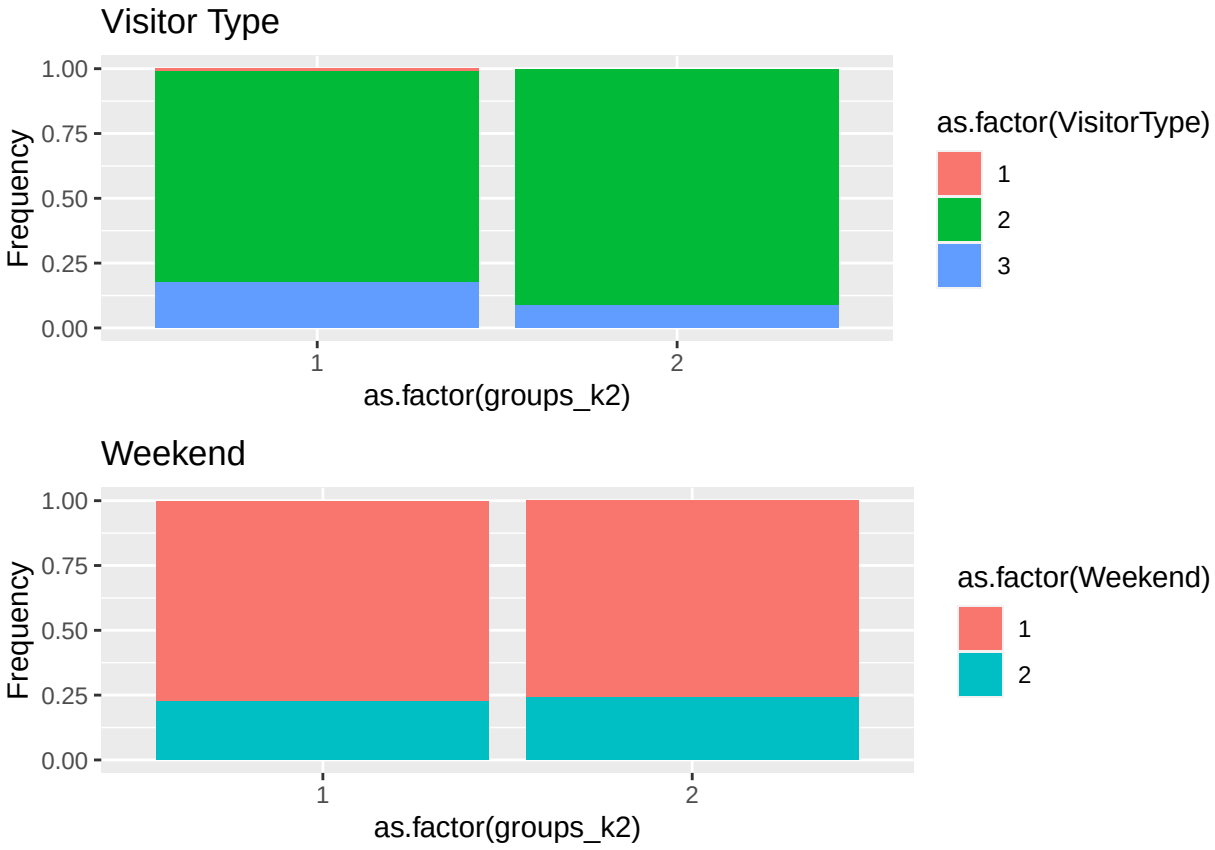

Bounce Rates
BounceRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(groups_k2)
factor(groups_k2)
1
2
Exit Rates
ExitRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(groups_k2)
factor(groups_k2)
1
2


Bounce Rates
BounceRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(groups_k2)
factor(groups_k2)
1
2
Exit Rates
ExitRates
0.20
0.15
0.10
0.05
0.00
1
2
factor(groups_k2)
factor(groups_k2)
1
2


Page Values
PageValues
300
200
100
0
1
2
factor(groups_k2)
factor(groups_k2)
1
2
Month
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(groups_k2)
as.factor(Month)
1
2
3
4
5
6
7
8
9


Operating Systems
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(groups_k2)
as.factor(OperatingSystems)
1
2
3
4
5
6
7
8
Browser Type
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(groups_k2)
as.factor(Browser)
1
2
3
4
5
6
7
8
9
10
11

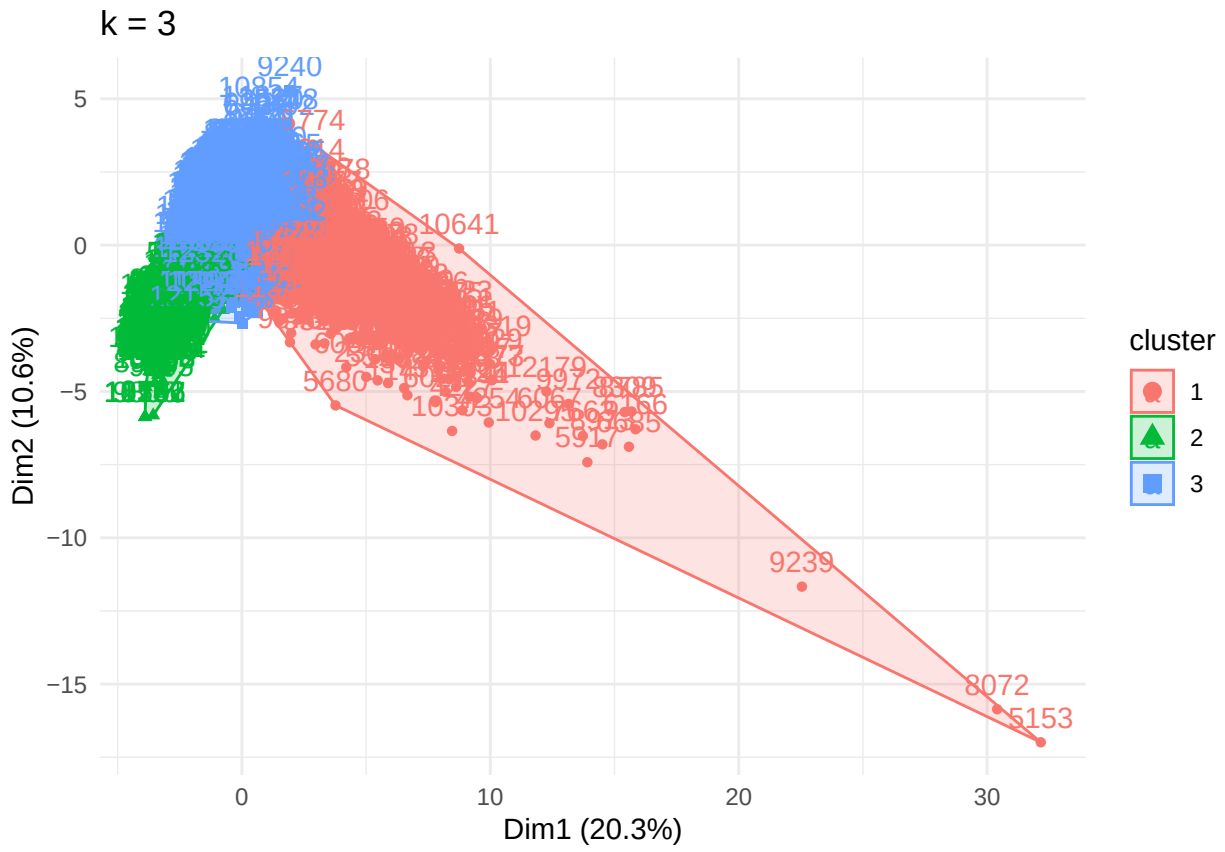

Region
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(groups_k2)
Traffic Type
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
as.factor(groups_k2)
1
2
3
as.factor(TrafficType)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



Further Research: k=3, Clustering, K-means, PAM, and Hierarchical

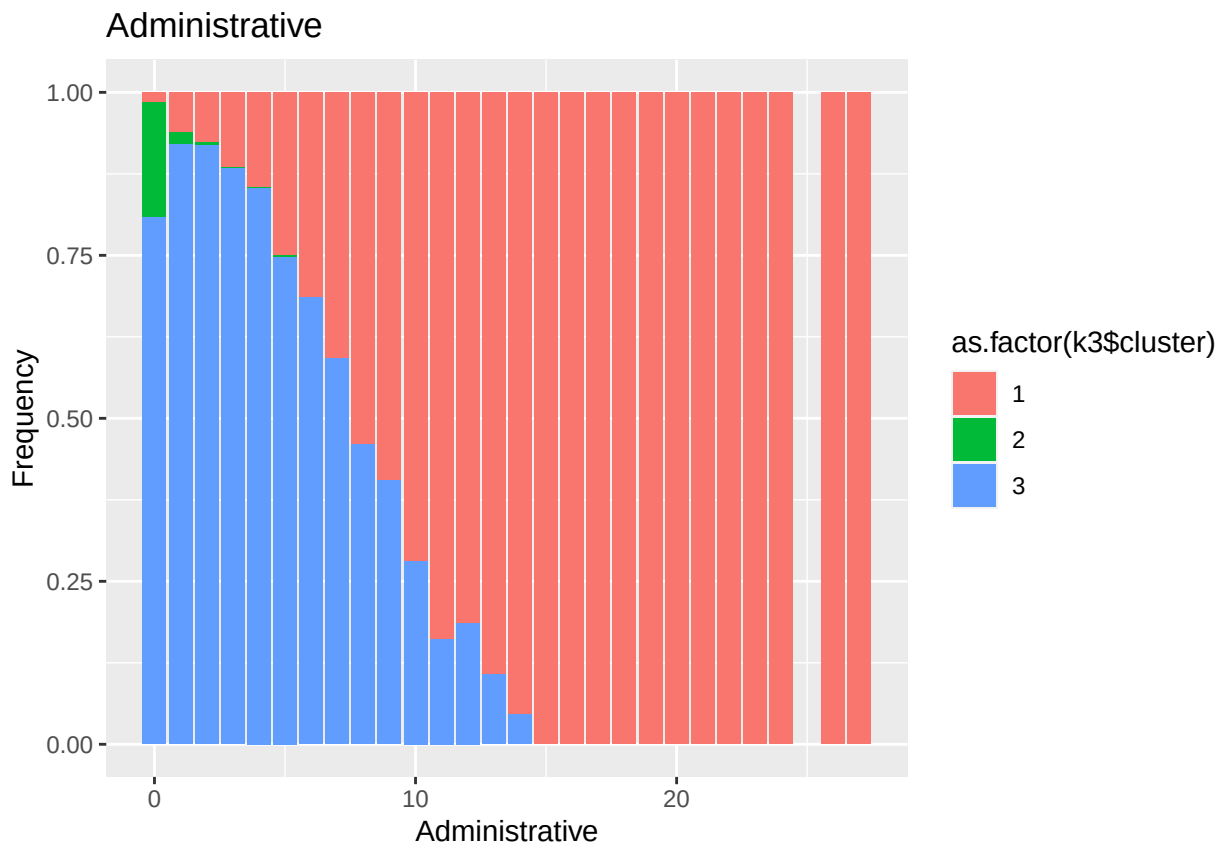
To further analyze the site-visitors and their behaviour trends, k=3 was used for all the models again.

The K-means clustering model produced the following plot for k=3, where there were 1,646 visitors in cluster 1, 1,061 visitors in cluster 2, and 9,623 visitors in cluster 3.

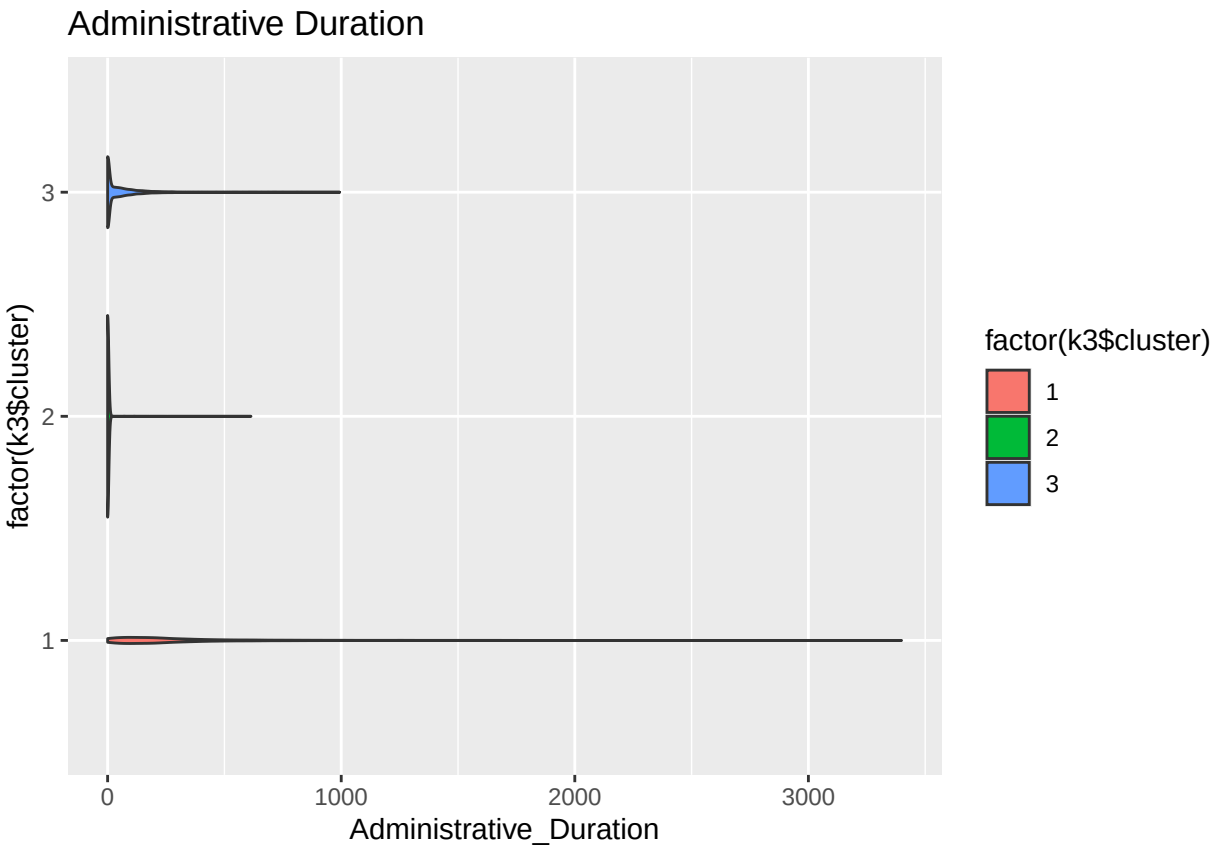



```
##
##    1    2    3
## 1646 1061 9623
```

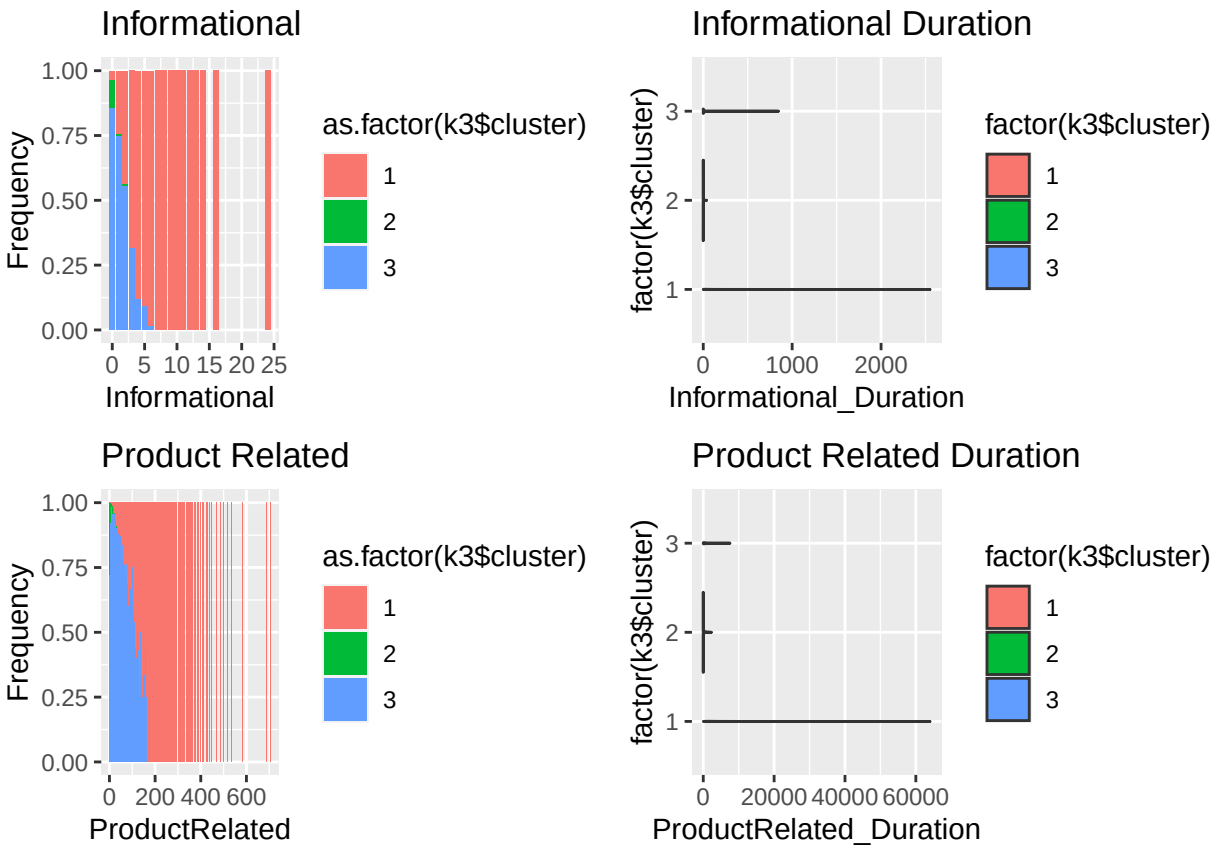
As in k=2 clustering, the k=3 clustering also produced a group that was likely and unlikely to make a purchase. However, it further divided the group that was unlikely to make a purchase into two. While both of those two groups were unlikely to make a purchase (performed similarly in page visit and duration columns), one cluster had a slightly lower bounce and exit rate. Therefore, it seemed that k=3 clustering divided the site-visitors into 1) likely to make purchases (cluster 1), 2) less likely to make a purchase (cluster 3), and 3) very stubborn and least likely to make a purchase (cluster 2).



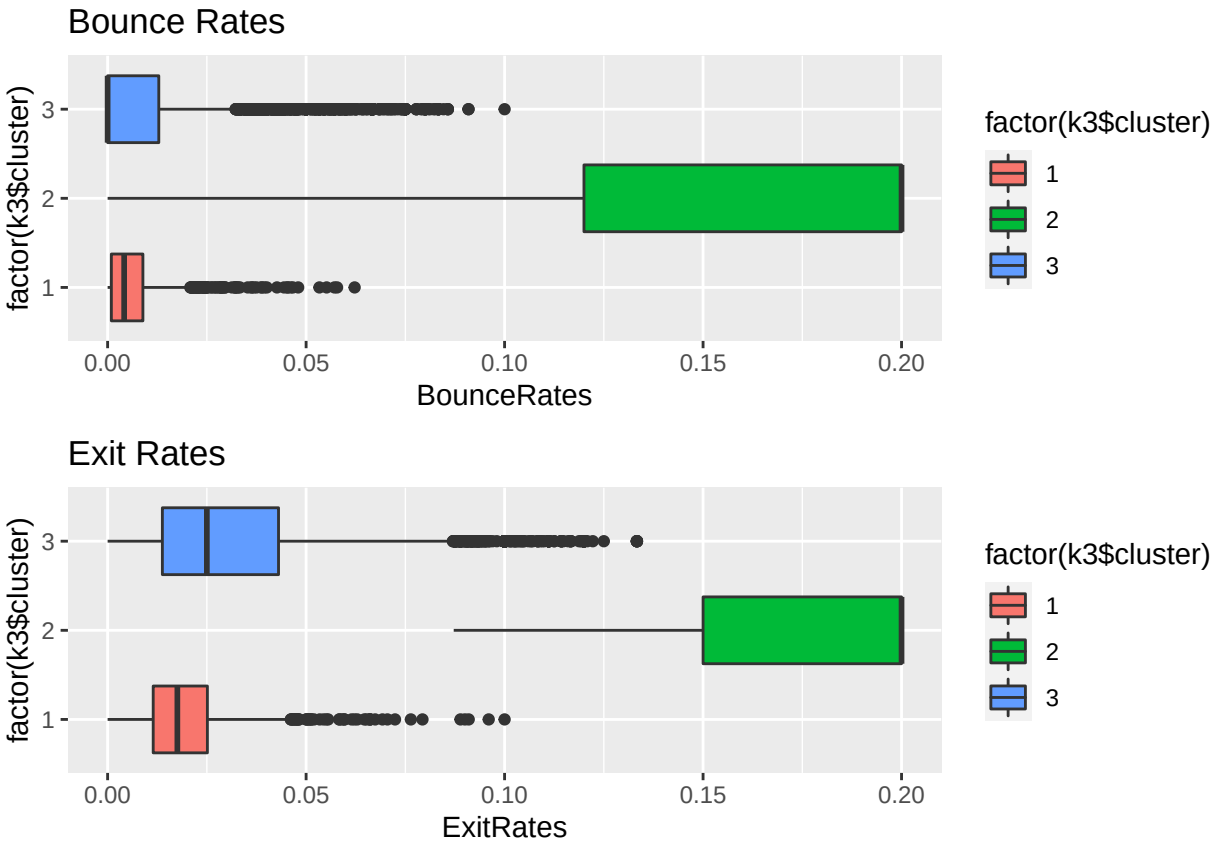
Here, we can see that the cluster 2 has the highest proportion of site-visitors who visited 0 administrative web pages, while the red cluster (the most likely to make a purchase) visited a high number of websites.



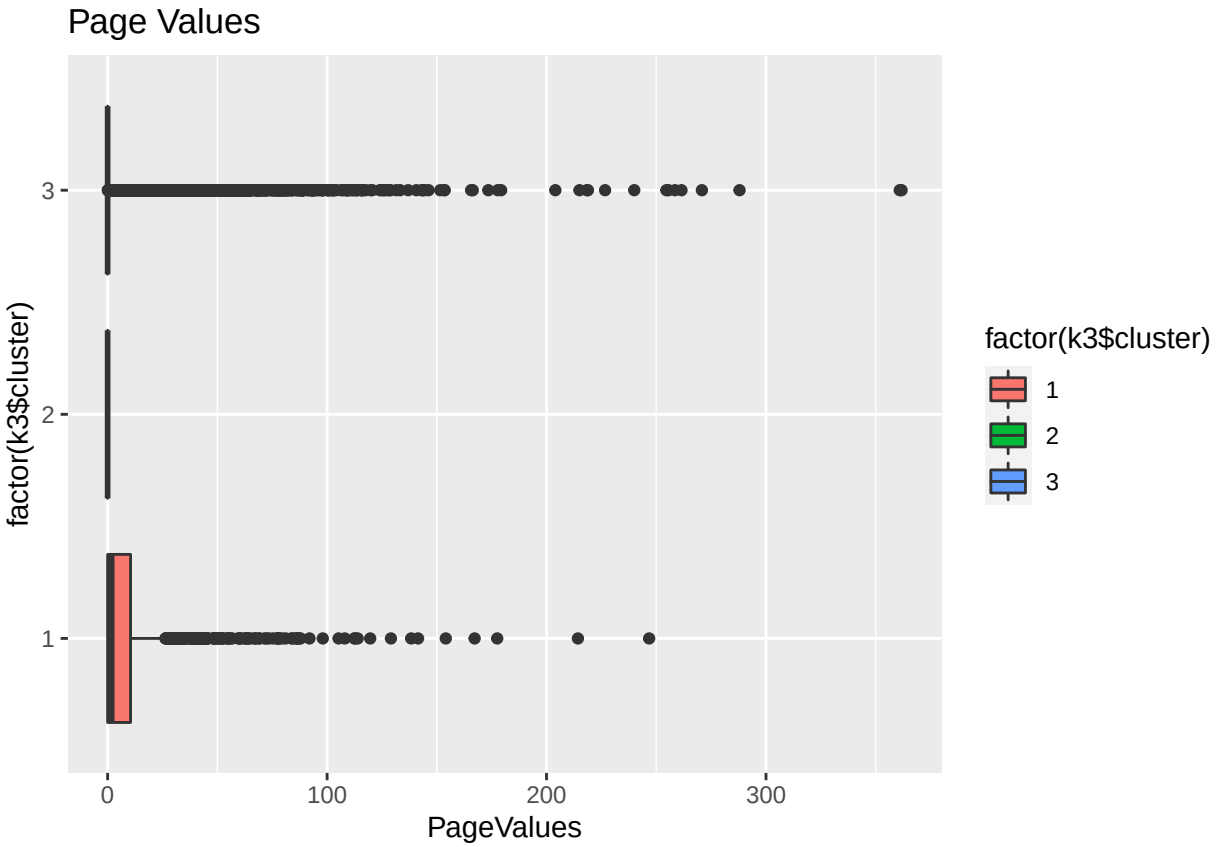
Moreover, cluster 1 spent the most time in the administrative web pages, green cluster spent the least time, and the blue cluster was in-between.



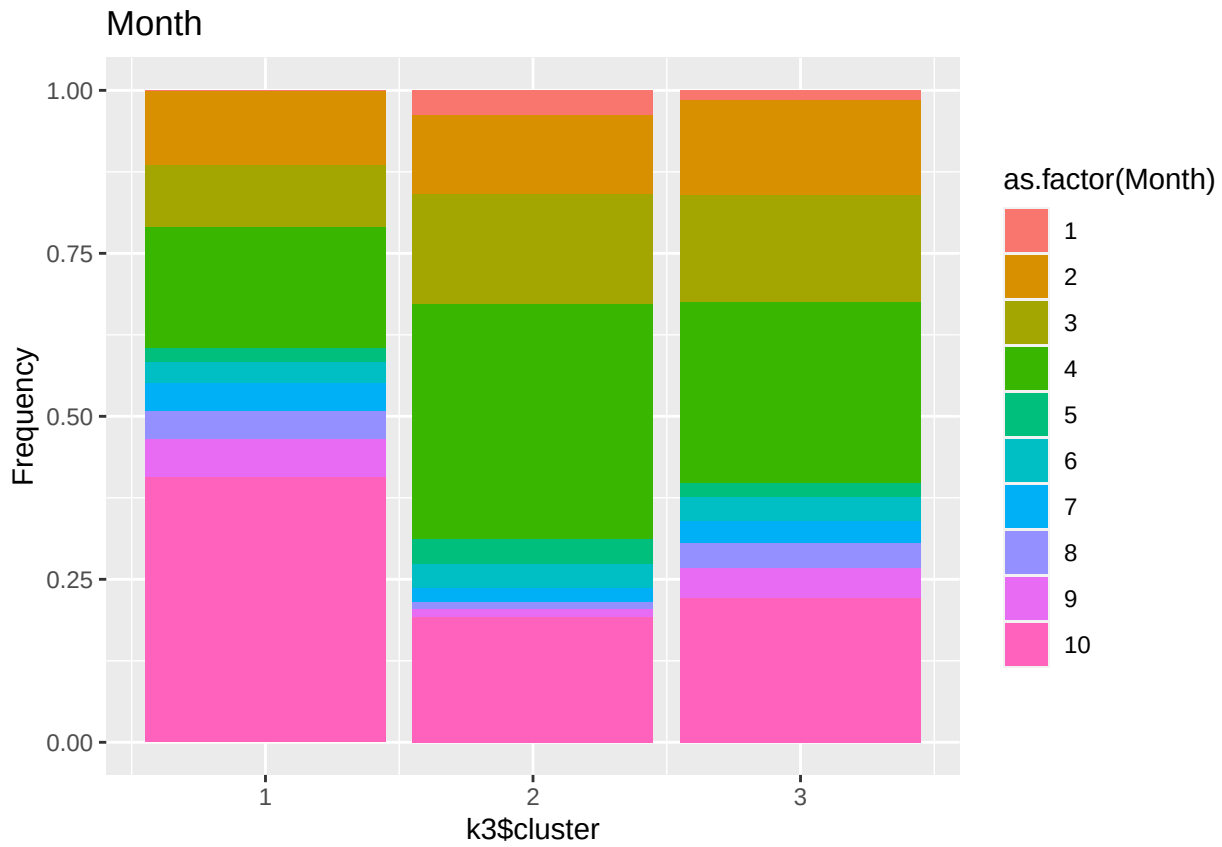
Similar observations were repeated with the informational and product related web pages.



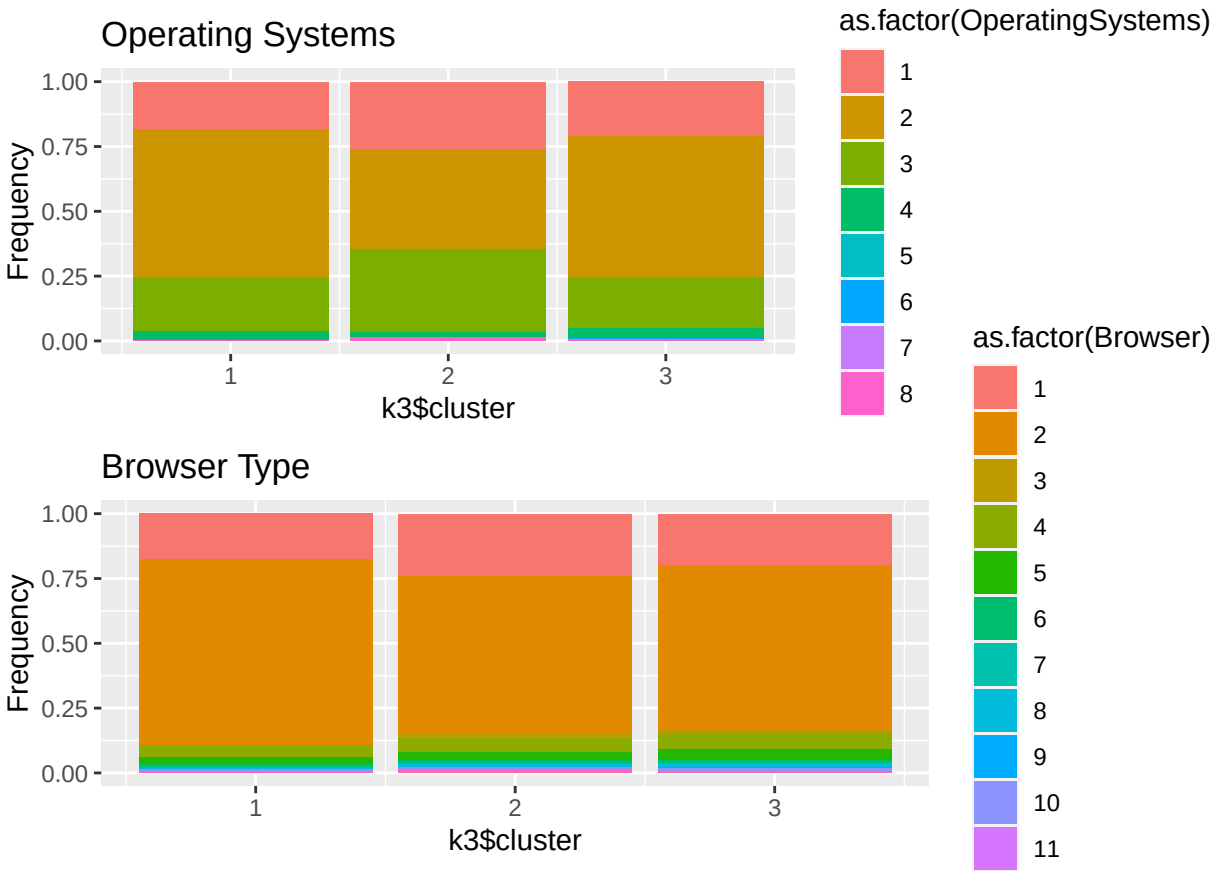
The cluster 2 (least likely to make a purchase) also showed the highest bounce and exit rate, while the red cluster showed the least, and the blue was, again, in-between.



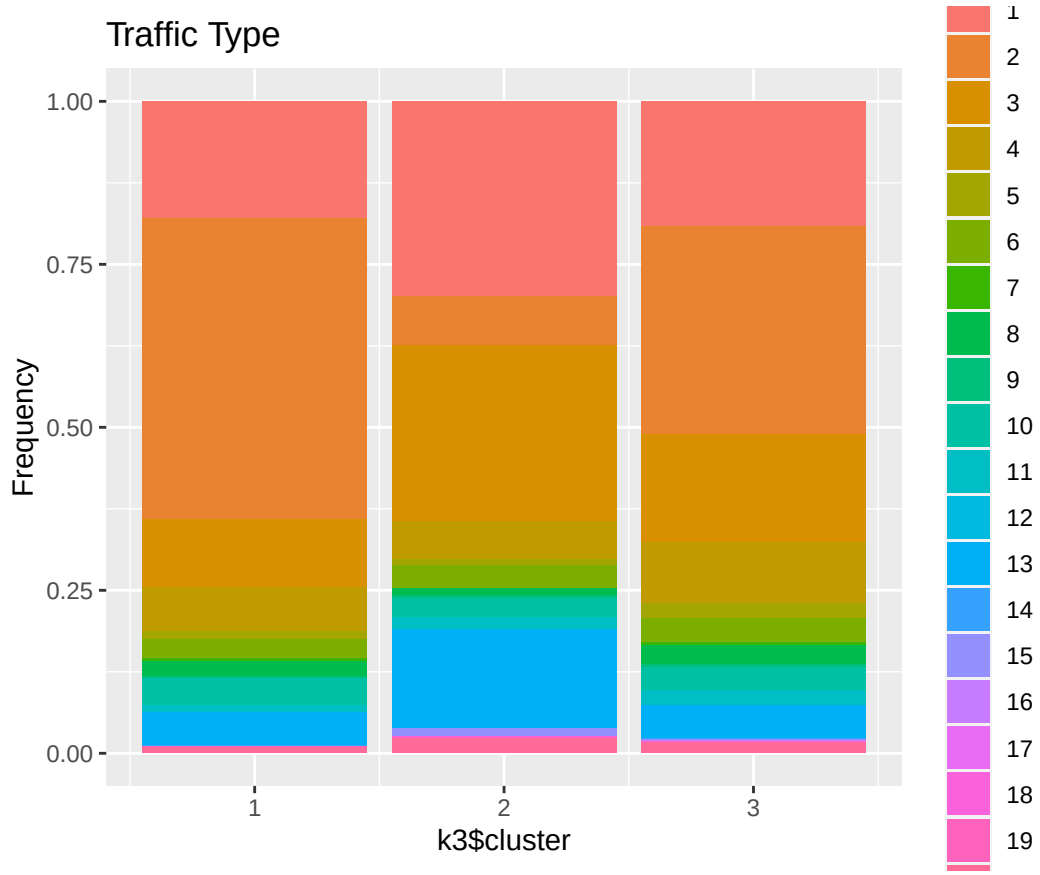
The interquartile range of the cluster 1 had the highest page value, which is in accordance with all the previous observations.



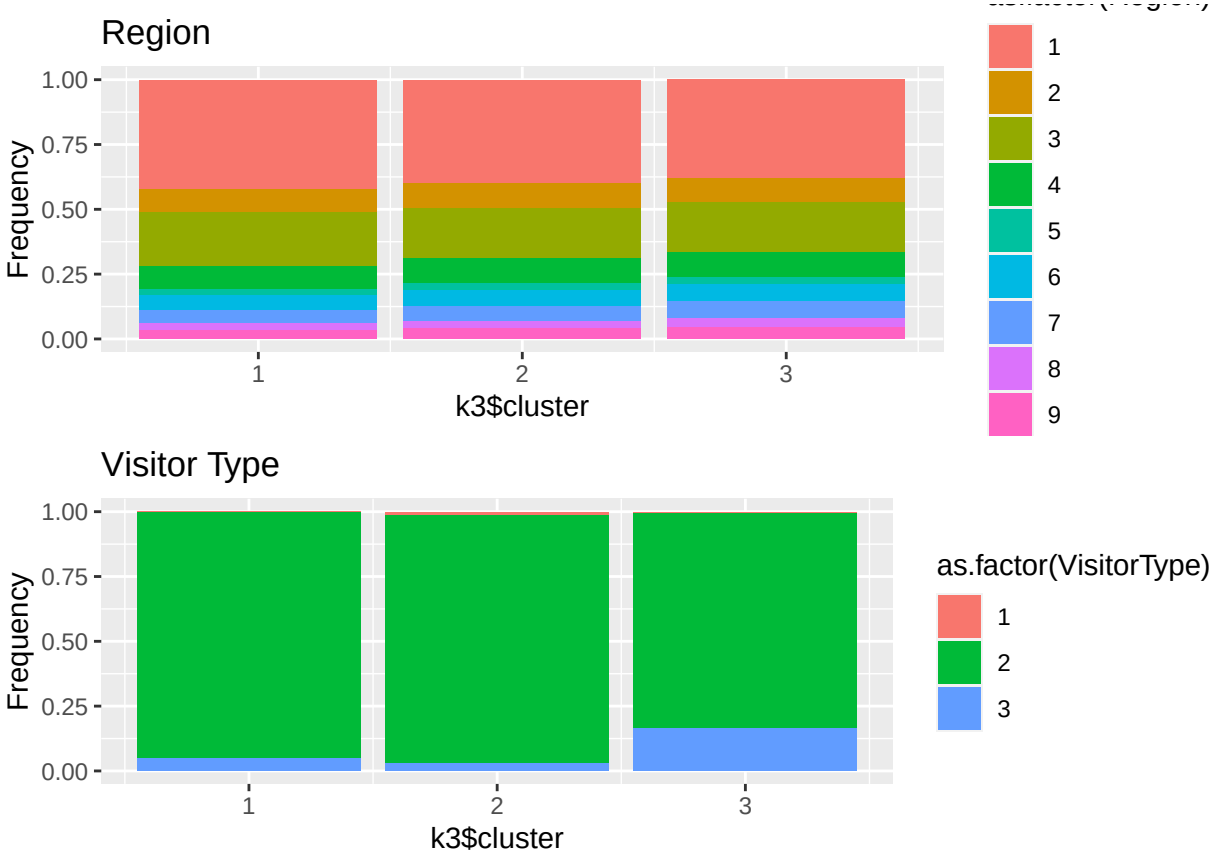
Here, we see an interesting observation among the site-visitors who are least likely to purchase (cluster 2). They shop earlier in the year (February~May) compared to other clusters, while other clusters shop more in the later months of the year (November) when winter holiday is near. It is unclear why early months of the year are popular shopping months for the site-visitors in cluster 2 (the cluster that's least likely to make a purchase). Nevertheless, it is an interesting insight that can inform e-commerce business owners that anyone who shops earlier in the year is unlikely to make a purchase again, as they are not the ideal type of users.



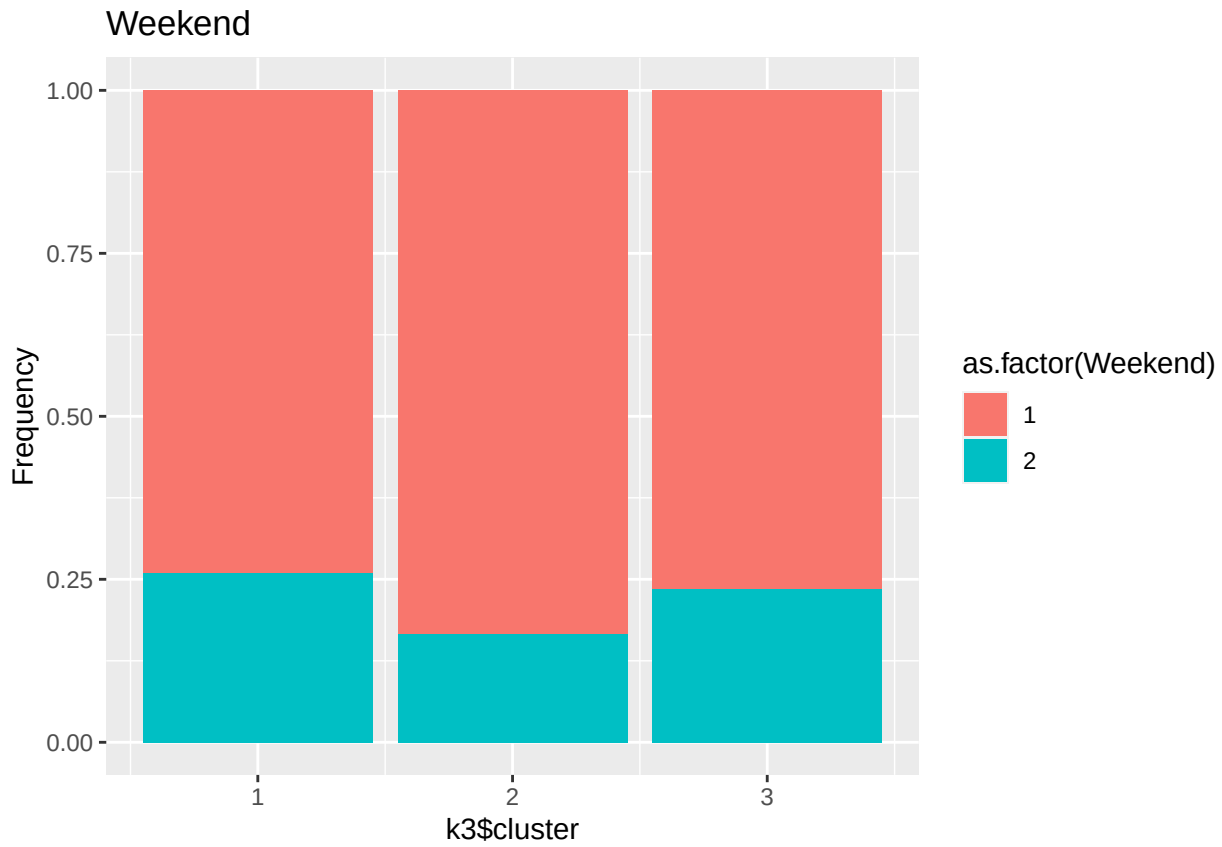
Here, cluster 2 is uniquely represented again for the operating system and browser type. Cluster 2 seems to prefer using the operating system and browser type that is unpopular in other clusters. For the operating system type, it looks like type "2" is the most popular and commonly used computer operating system type. Interestingly, cluster 2 uses it the least, and instead, uses other operating system types that are unpopular in other clusters. The same trend is observed in the browser type as well. It would have been very interesting to find out what those operating systems and browser types were exactly. Perhaps cluster 2 uses an older operating system and browser type, such as Windows 98 and Internet Explorer, because site-visitors using an older system are less likely to be adapted to technology, and therefore less likely to make online shopping purchases.



Traffic type divides the three clusters distinctively. Cluster 1 (most likely to make a purchase) favors traffic type "3", cluster 2 (least likely to make a purchase) favors traffic type "1, 4, and 13", while the cluster 3 (in-between cluster) shows moderate preference for all those traffic types. Since cluster 1 is the most likely to make a purchase, perhaps their traffic types are more relevant to advertisements and promotional emails, while the cluster 2, once again, might prefer unpopular and old traffic types, such as newspaper and print ads.



Similar to the previous trends observed, the cluster did not differ much in region or visitor type.

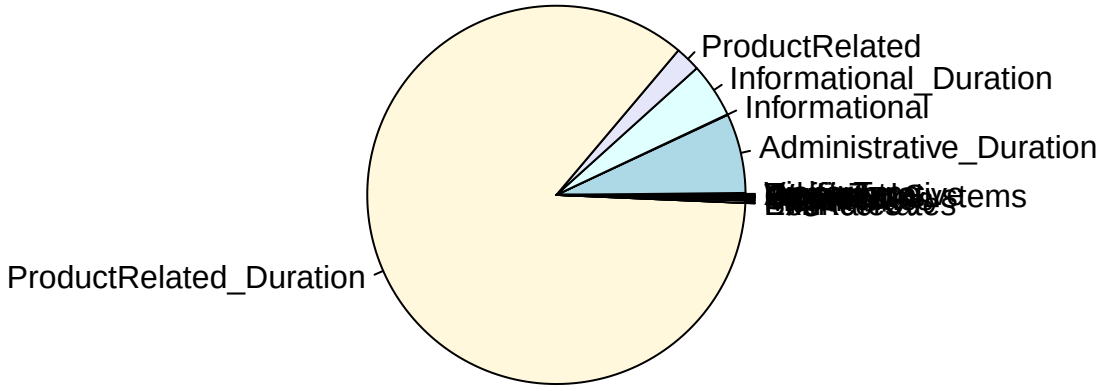


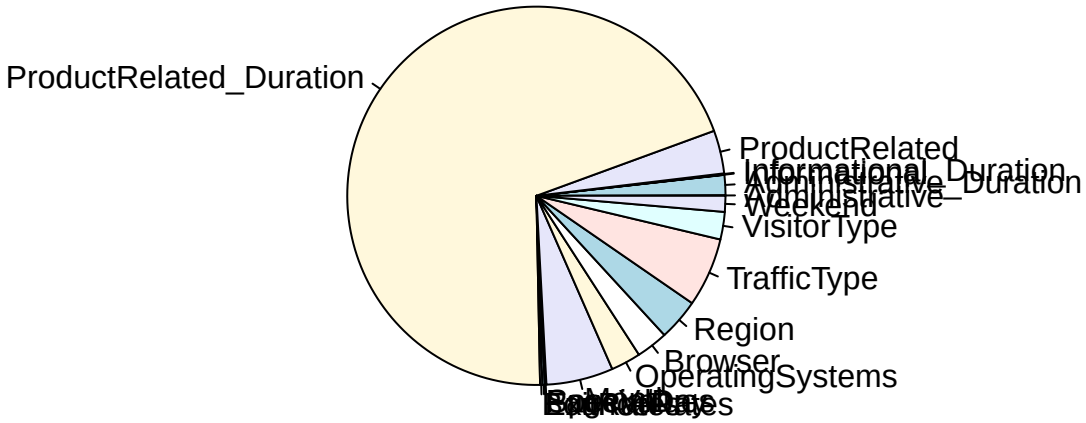
Also similar to the previous trends, the cluster 1 (most likely to make a purchase) shopped more on weekends, while cluster 2 shopped the least on weekends.

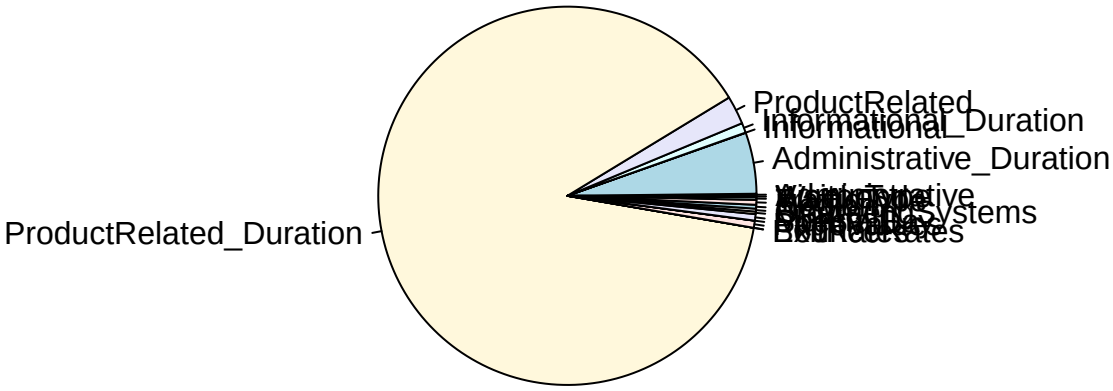
The following table shows the k-mean centers for cluster 1, 2, and 3. Product related duration again had the largest area in the cluster pie charts. Interestingly, cluster 3 (blue cluster) had the smallest product related duration; again, it is insufficient to draw a meaningful conclusion from this observation with the current dataset.

```
##          Administrative Administrative_Duration           Informational
##              1.50672160              1.26746377              1.56271172
##  Informational_Duration          ProductRelated ProductRelated_Duration
##              1.24658246              1.46591822              1.38112888
##             BounceRates              ExitRates              PageValues
##             -0.32538194             -0.48202859              0.19926491
##              SpecialDay                   Month        OperatingSystems
##             -0.17136978              0.40574148             -0.01407607
##                 Browser                  Region             TrafficType
##             -0.08376768             -0.09880078             -0.11901708
##             VisitorType                 Weekend
##             -0.23296249              0.06346228

##          Administrative Administrative_Duration           Informational
##              -0.68192673             -0.44888159             -0.38384714
##  Informational_Duration          ProductRelated ProductRelated_Duration
##              -0.24426969             -0.64248878             -0.59312156
##             BounceRates              ExitRates              PageValues
```

```
##              2.95700965              2.80943526             -0.31716498
##             SpecialDay                   Month        OperatingSystems
##              0.25882586             -0.20960825              0.06766786
##                 Browser                  Region             TrafficType
##             -0.02408269             -0.03977624              0.24262247
##             VisitorType                 Weekend
##             -0.33145152             -0.15791905

##          Administrative Administrative_Duration           Informational
##            -0.1825355379           -0.1673056213           -0.2249778318
##  Informational_Duration          ProductRelated ProductRelated_Duration
##            -0.1862937320           -0.1799044784           -0.1708444518
##              BounceRates               ExitRates              PageValues
##            -0.2703739545           -0.2273087141            0.0008855871
##             SpecialDay                   Month        OperatingSystems
##             0.0007752697           -0.0462907751           -0.0050531424
##                 Browser                  Region             TrafficType
##             0.0169836161            0.0212853246           -0.0063930497
##             VisitorType                 Weekend
##             0.0763926343            0.0065565000
```



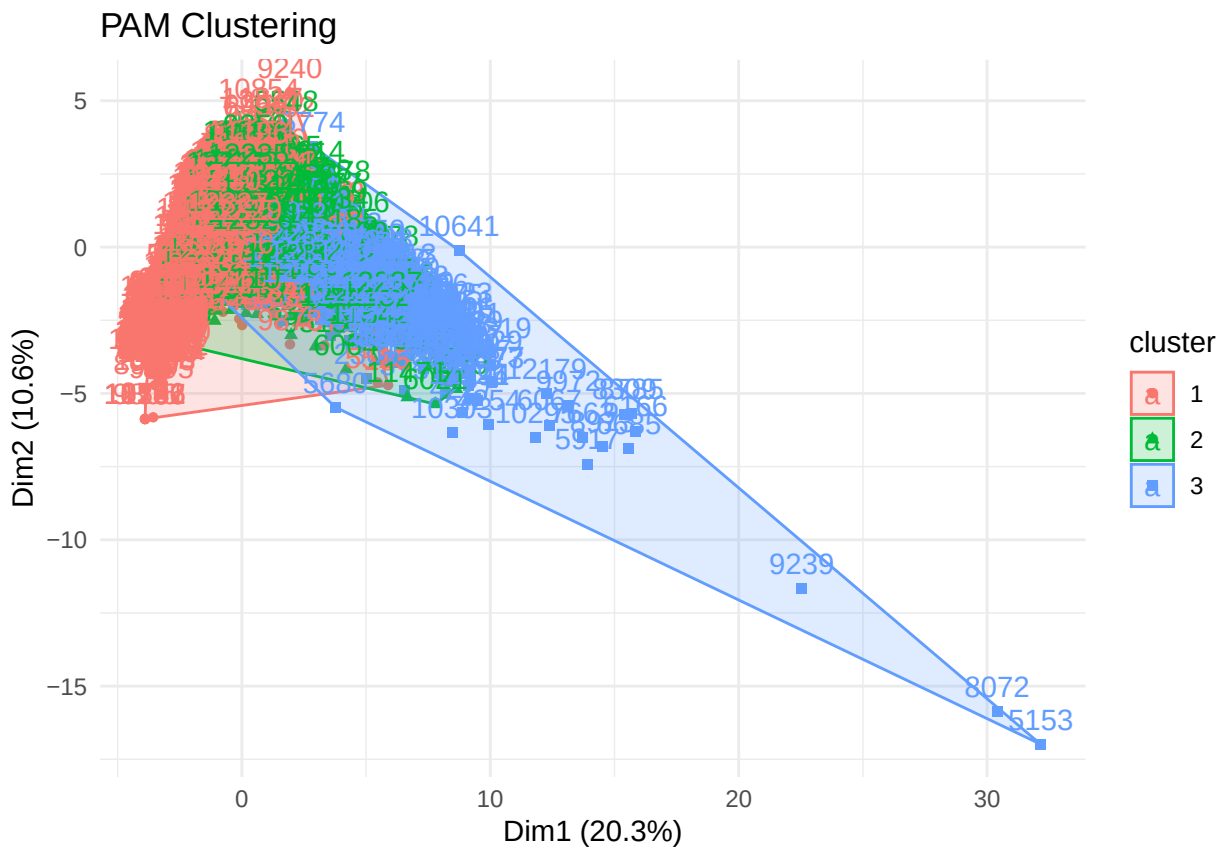

ProductRelated_Duration
ProductRelated
Informational_Duration
Administrative_Duration
Administrative
Weekend
VisitorType
TrafficType
Region
Browser
OperatingSystems



Moreover, the following table shows the k-center differences between the clusters.

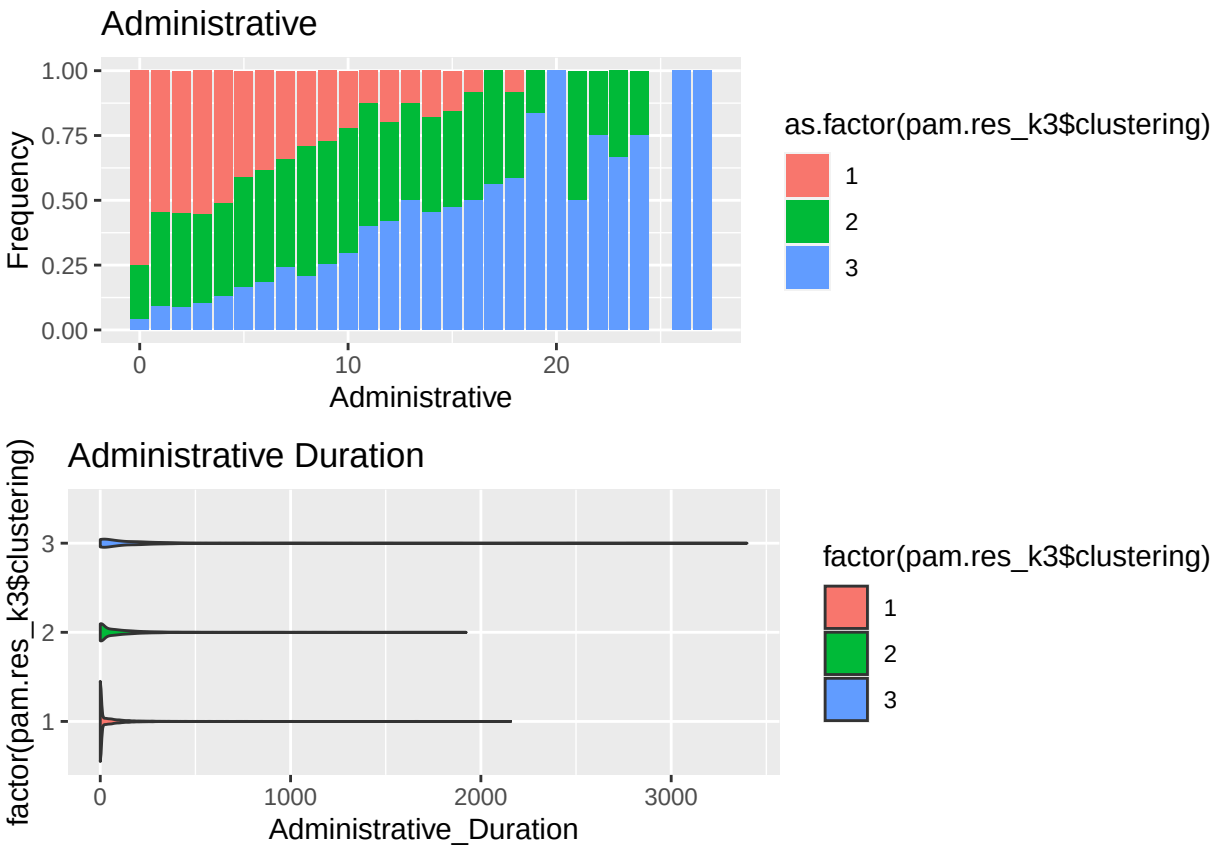
```
##         diff12       diff13      diff23
## 1   2.18864833  1.689257133 -0.49939120
## 2   1.71634535  1.434769387 -0.28157597
## 3   1.94655886  1.787689552 -0.15886931
## 4   1.49085214  1.432876188 -0.05797596
## 5   2.10840700  1.645822698 -0.46258430
## 6   1.97425044  1.551973330 -0.42227711
## 7  -3.28239160 -0.055007989  3.22738361
## 8  -3.29146385 -0.254719876  3.03674398
## 9   0.51642989  0.198379322 -0.31805057
## 10 -0.43019564 -0.172145049  0.25805059
## 11  0.61534974  0.452032260 -0.16331748
## 12 -0.08174394 -0.009022929  0.07272101
## 13 -0.05968499 -0.100751297 -0.04106631
## 14 -0.05902455 -0.120086109 -0.06106156
## 15 -0.36163955 -0.112624034  0.24901552
## 16  0.09848904 -0.309355122 -0.40784416
## 17  0.22138133  0.056905782 -0.16447555
```

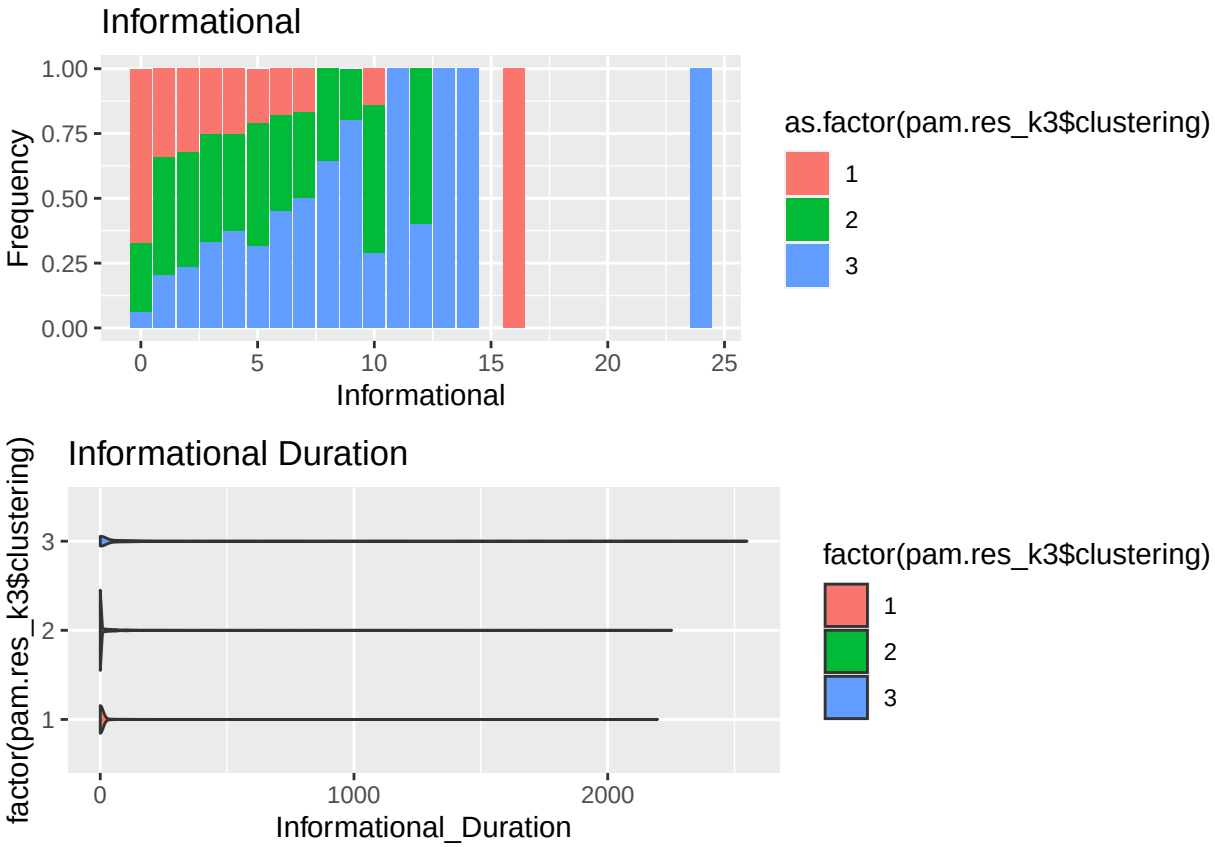
PAM clustering was repeated for k=3, and the following graph shows how the PAM model divided site-visitors into 3 clusters, where 7,326 site-visitors were in cluster 1, 3,737 site-visitors were in cluster 2, and 1,267 site-visitors were in cluster 3.

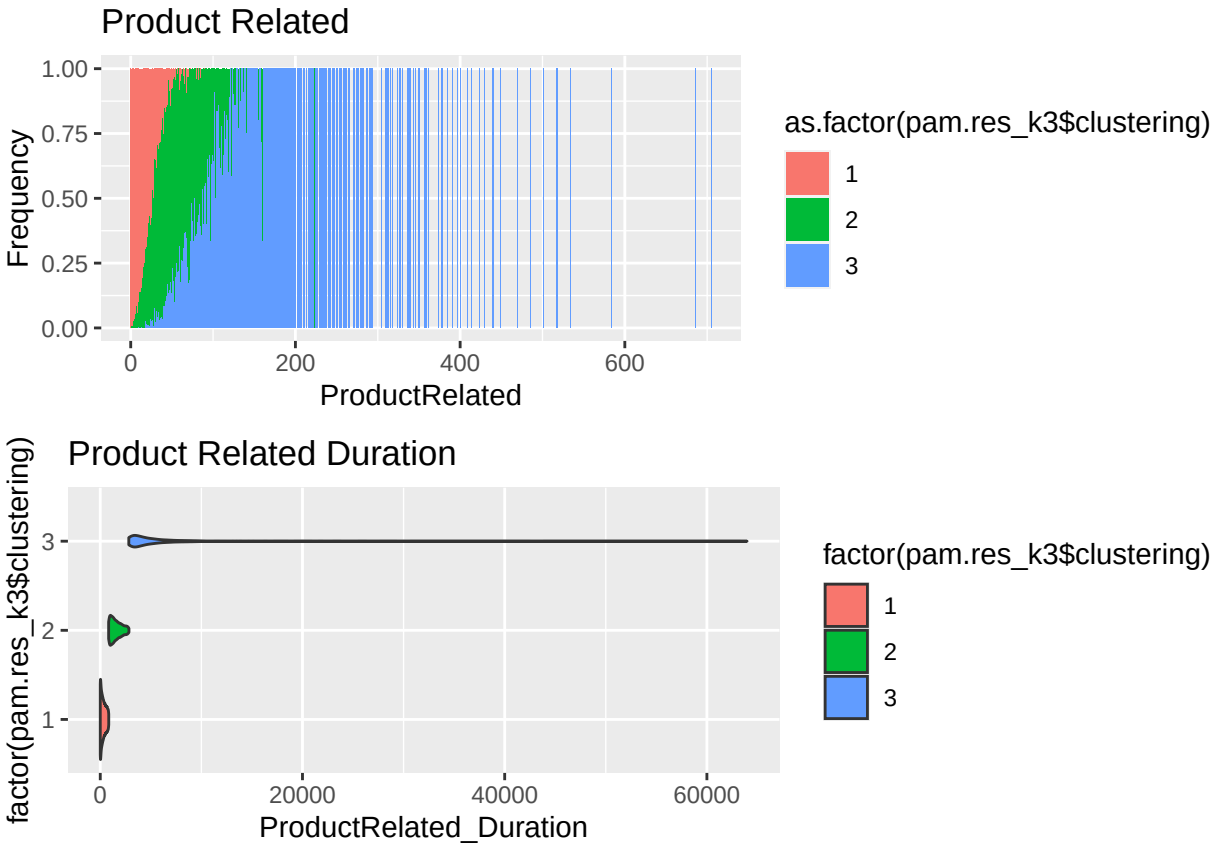


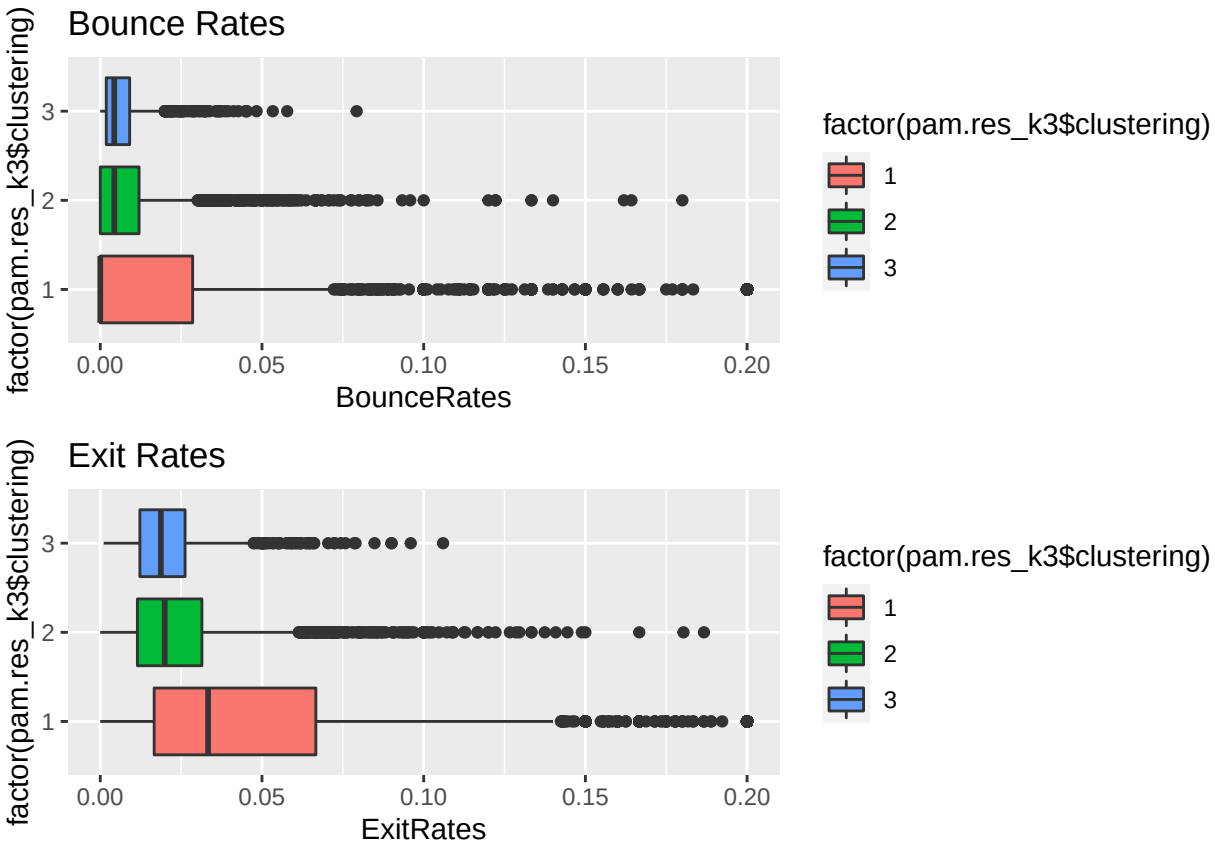

```
##
##    1    2    3
## 7326 3737 1267
```

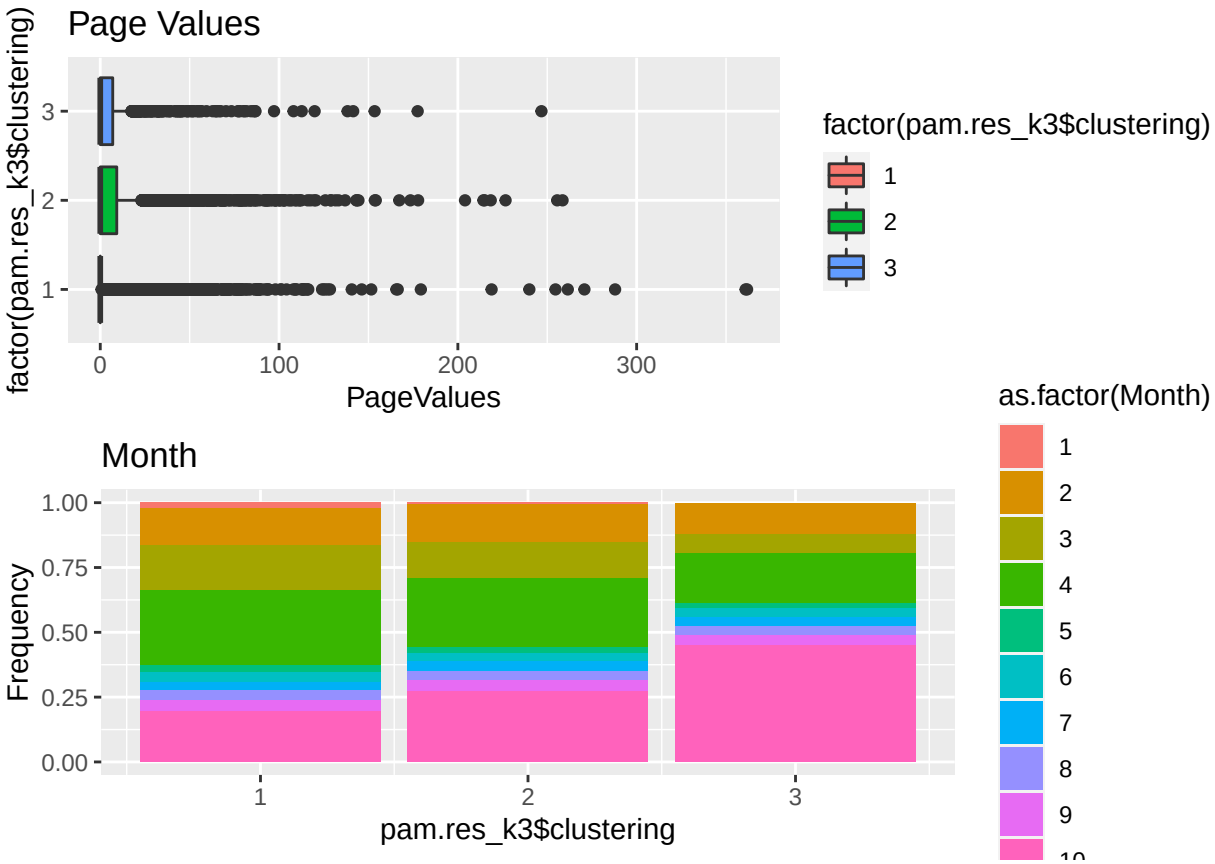
Many observations made in the PAM models were similar with the observations made in the K-means model for k=3, where one cluster was most likely to make a purchase, another cluster least likely to make a purchase, and the remaining cluster that was in-between. The following pages will show the plots for the PAM, k=3, clustering model that correspond with the observation provided in this paragraph.

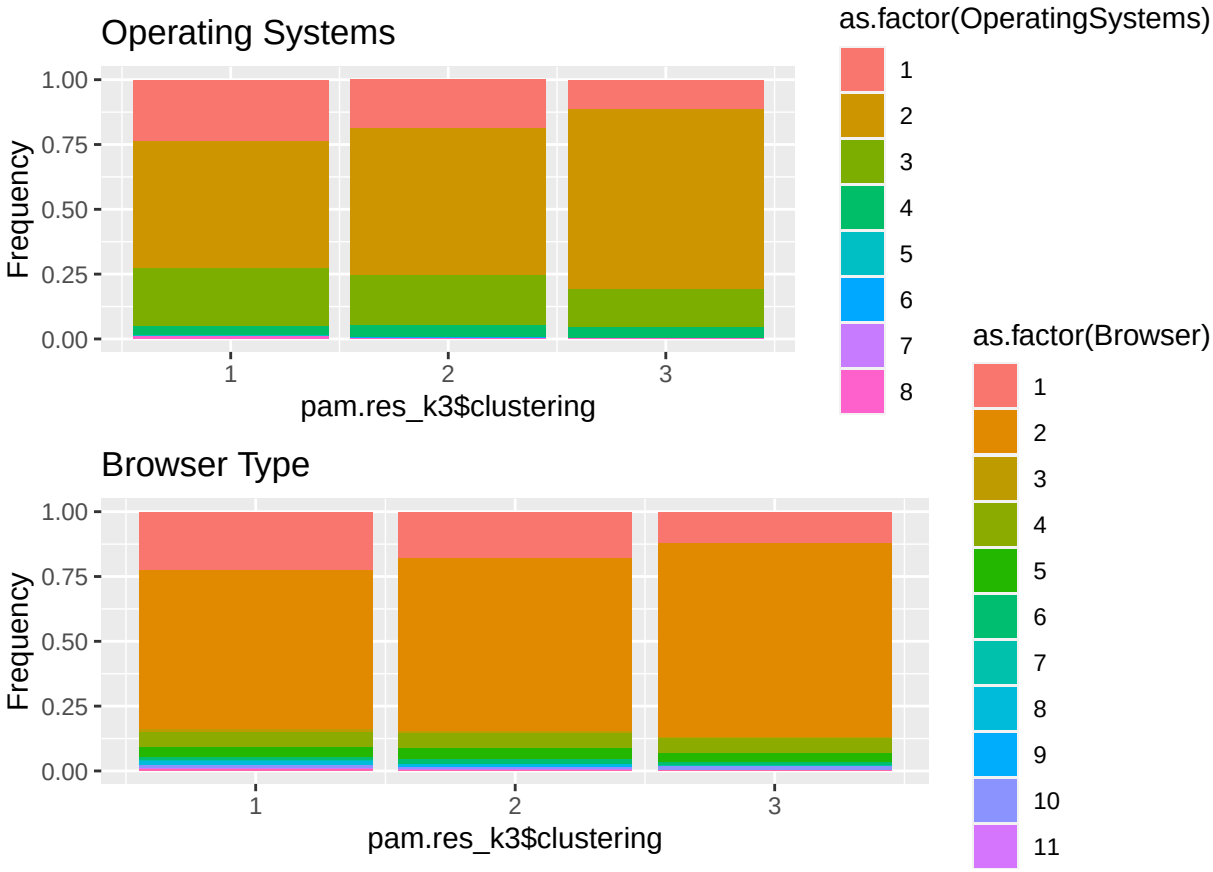

Administrative
Frequency
1.00
0.75
0.50
0.25
0.00
0
10
20
Administrative
as.factor(pam.res_k3$clustering)
1
2
3
Administrative Duration
factor(pam.res_k3$clustering)
3
2
1
0
1000
2000
3000
Administrative_Duration
factor(pam.res_k3$clustering)
1
2
3

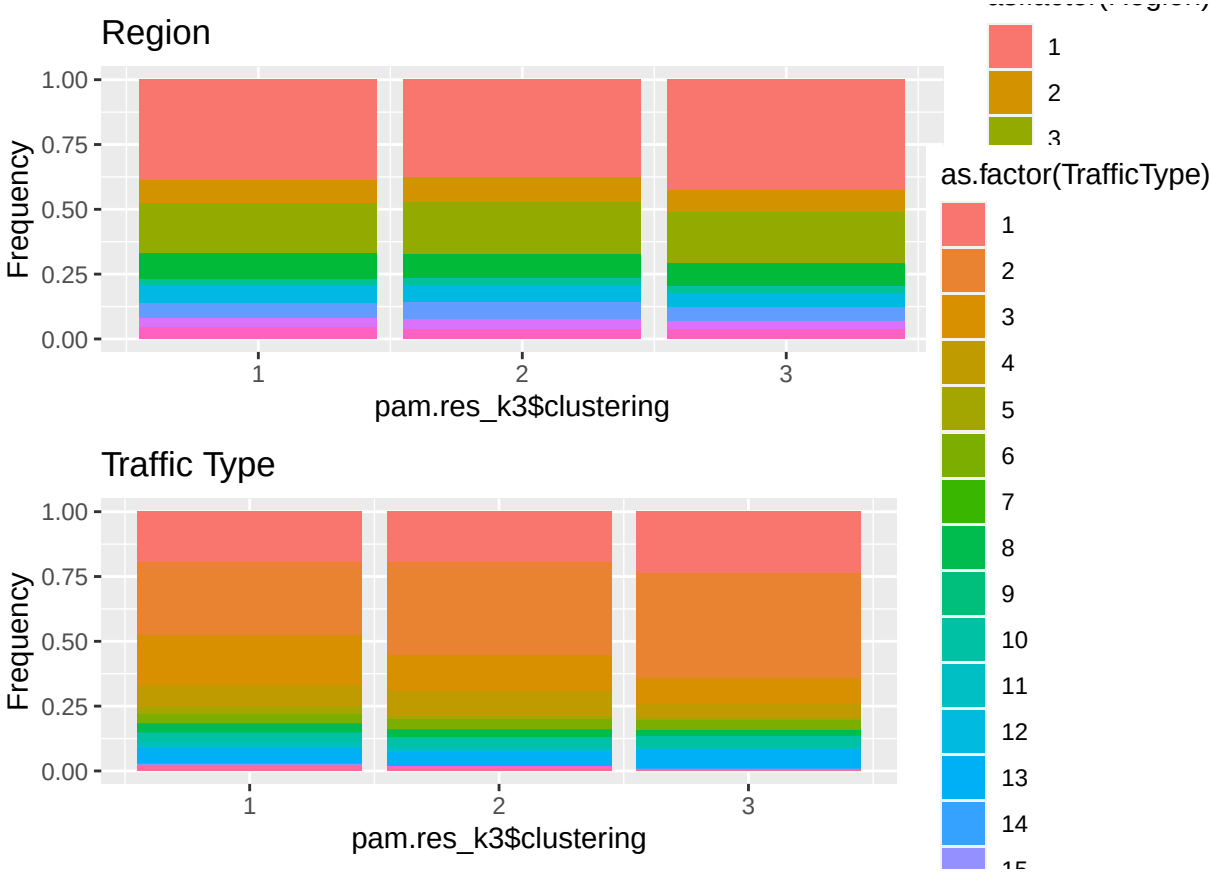

Informational
Frequency
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
Informational
as.factor(pam.res_k3$clustering)
1
2
3
Informational Duration
factor(pam.res_k3$clustering)
3
2
1
0
1000
2000
Informational_Duration
factor(pam.res_k3$clustering)
1
2
3

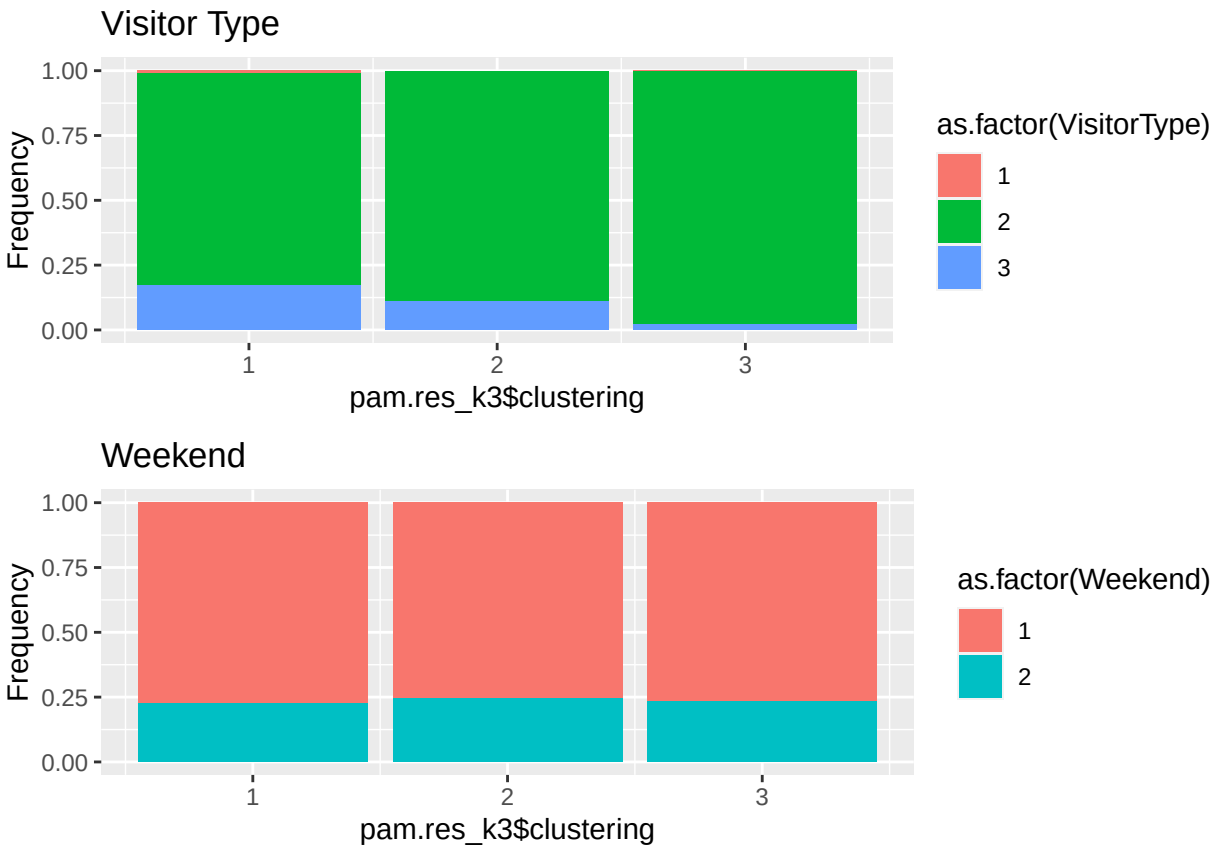

Product Related
Frequency
1.00
0.75
0.50
0.25
0.00
0
200
400
600
ProductRelated
as.factor(pam.res_k3$clustering)
1
2
3
Product Related Duration
factor(pam.res_k3$clustering)
3
2
1
0
20000
40000
60000
ProductRelated_Duration
factor(pam.res_k3$clustering)
1
2
3


Bounce Rates
factor(pam.res_k3$clustering)
3
2
1
0.00
0.05
0.10
0.15
0.20
BounceRates
factor(pam.res_k3$clustering)
1
2
3
Exit Rates
factor(pam.res_k3$clustering)
3
2
1
0.00
0.05
0.10
0.15
0.20
ExitRates
factor(pam.res_k3$clustering)
1
2
3

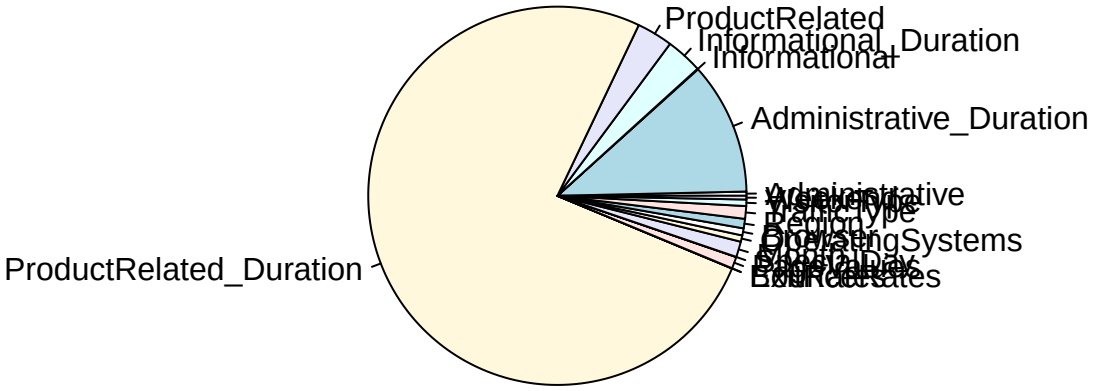

Page Values
factor(pam.res_k3$clustering)
3
2
1
0
100
200
300
PageValues
factor(pam.res_k3$clustering)
1
2
3
Month
as.factor(Month)
1.00
0.75
0.50
0.25
0.00
Frequency
1
2
3
pam.res_k3$clustering
1
2
3
4
5
6
7
8
9

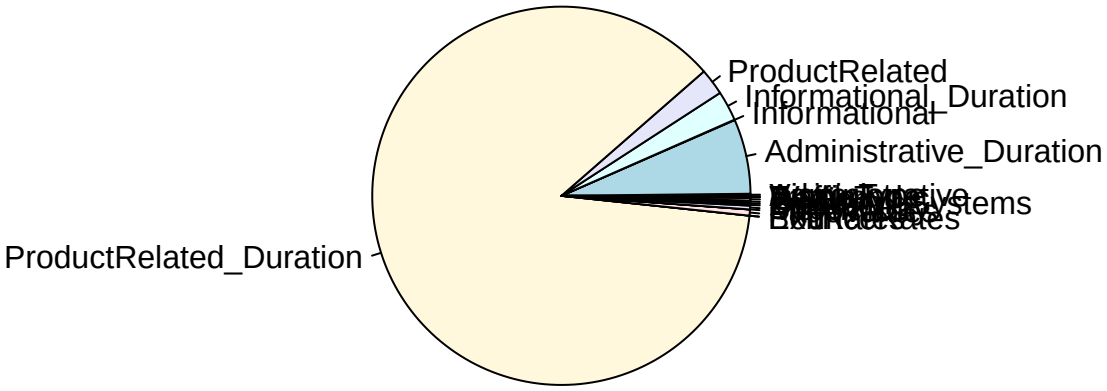

Operating Systems
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
pam.res_k3$clustering
as.factor(OperatingSystems)
1
2
3
4
5
6
7
8
Browser Type
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
pam.res_k3$clustering
as.factor(Browser)
1
2
3
4
5
6
7
8
9
10
11

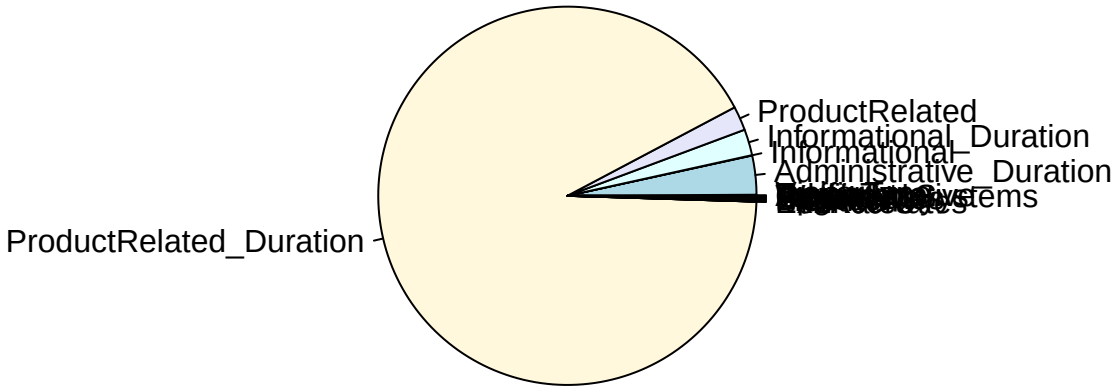

Region
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
pam.res_k3$clustering
Traffic Type
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
pam.res_k3$clustering
1
2
3
as.factor(TrafficType)
1
2
3
4
5
6
7
8
9
10
11
12
13
14



The following cluster pie charts for the PAM model (shown in the order of cluster 1, 2, and 3) also displayed similar trends as with the K-means model, where product related duration had the highest proportion in the chart.


ProductRelated
Informational_Duration
Informational
Administrative_Duration
ProductRelated_Duration


ProductRelated
Informational_Duration
Informational
Administrative_Duration
ProductRelated_Duration

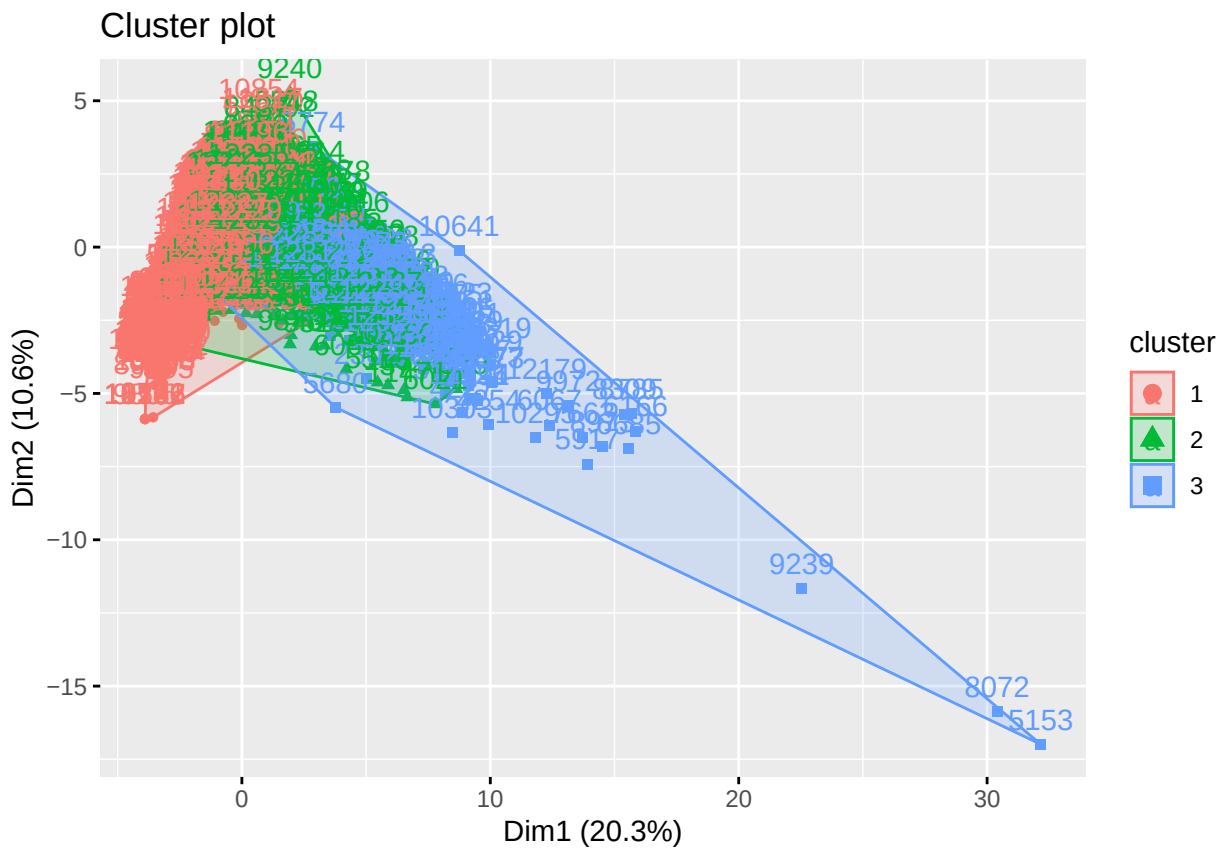


The following shows the medoid centers for cluster 1, 2, and 3 for the PAM model, respectively, as well as the medoids difference between the clusters.

```
##          Administrative Administrative_Duration           Informational
##              1.00000000             24.00000000              0.00000000
##  Informational_Duration          ProductRelated ProductRelated_Duration
##              0.00000000             10.00000000            238.50000000
##             BounceRates               ExitRates              PageValues
##              0.00000000              0.01111111              0.00000000
##              SpecialDay                   Month        OperatingSystems
##              0.00000000              4.00000000              2.00000000
##                 Browser                  Region             TrafficType
##              4.00000000              3.00000000              3.00000000
##             VisitorType                 Weekend
##              3.00000000              1.00000000

##          Administrative Administrative_Duration           Informational
##               4.0000000              64.3333333               1.0000000
##  Informational_Duration          ProductRelated ProductRelated_Duration
##              13.0000000              27.0000000            1433.9166670
##             BounceRates               ExitRates              PageValues
##               0.0000000               0.0182266               3.6913448
##              SpecialDay                   Month        OperatingSystems
##               0.0000000               3.0000000               1.0000000
##                 Browser                  Region             TrafficType
##               4.0000000               1.0000000               8.0000000
```

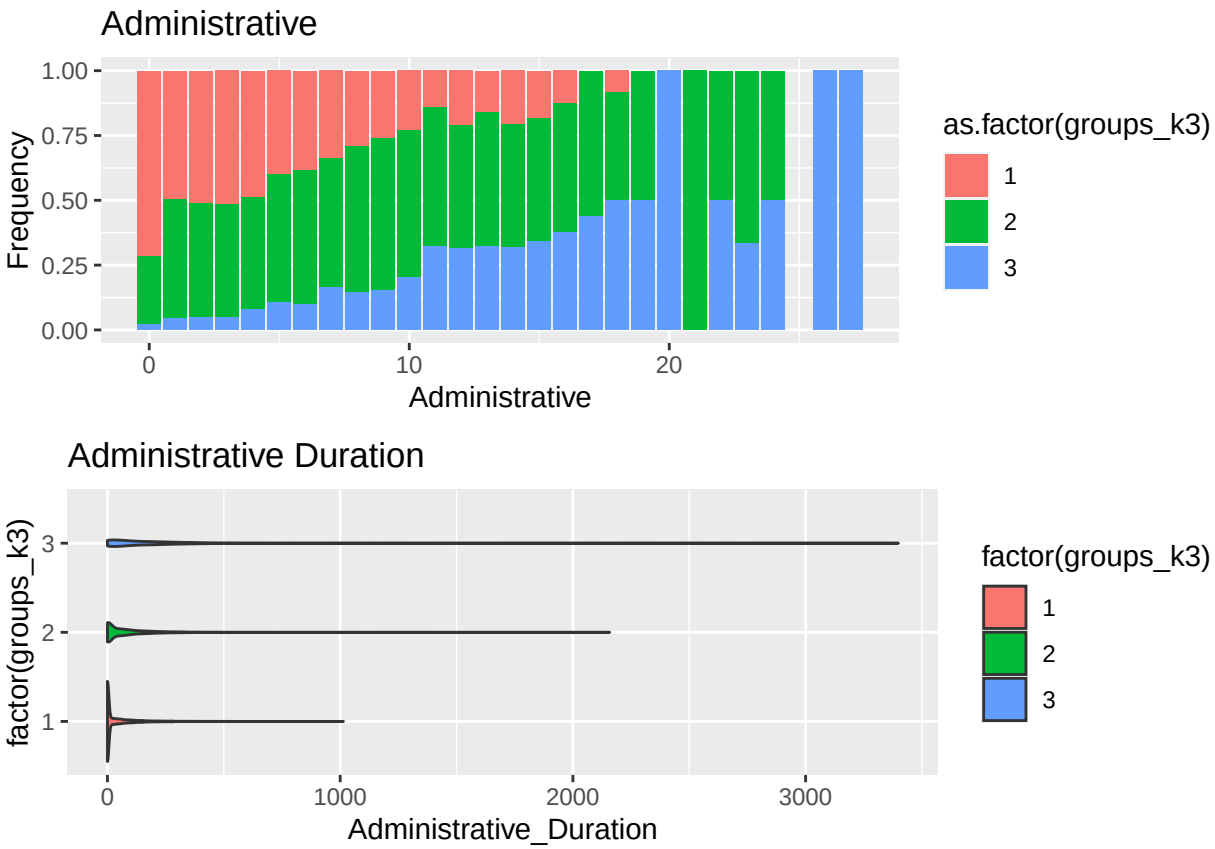
```
##             VisitorType                 Weekend
##               3.0000000               1.0000000

##          Administrative Administrative_Duration           Informational
##            5.000000e+00            8.434545e+01            3.000000e+00
##  Informational_Duration          ProductRelated ProductRelated_Duration
##            1.068000e+02            8.000000e+01            4.240036e+03
##             BounceRates               ExitRates              PageValues
##            1.696930e-04            1.163554e-02            1.257696e+00
##              SpecialDay                   Month        OperatingSystems
##            0.000000e+00            6.000000e+00            3.000000e+00
##                 Browser                  Region             TrafficType
##            2.000000e+00            1.000000e+00            2.000000e+00
##             VisitorType                 Weekend
##            2.000000e+00            1.000000e+00

##           diff12        diff13        diff23
## 1  -3.000000e+00 -4.000000e+00 -1.000000e+00
## 2  -4.033333e+01 -6.034545e+01 -2.001212e+01
## 3  -1.000000e+00 -3.000000e+00 -2.000000e+00
## 4  -1.300000e+01 -1.068000e+02 -9.380000e+01
## 5  -1.700000e+01 -7.000000e+01 -5.300000e+01
## 6  -1.195417e+03 -4.001536e+03 -2.806120e+03
## 7   0.000000e+00 -1.696930e-04 -1.696930e-04
## 8  -7.115490e-03 -5.244260e-04  6.591064e-03
## 9  -3.691345e+00 -1.257696e+00  2.433649e+00
## 10  0.000000e+00  0.000000e+00  0.000000e+00
## 11  1.000000e+00 -2.000000e+00 -3.000000e+00
## 12  1.000000e+00 -1.000000e+00 -2.000000e+00
## 13  0.000000e+00  2.000000e+00  2.000000e+00
## 14  2.000000e+00  2.000000e+00  0.000000e+00
## 15 -5.000000e+00  1.000000e+00  6.000000e+00
## 16  0.000000e+00  1.000000e+00  1.000000e+00
## 17  0.000000e+00  0.000000e+00  0.000000e+00
```

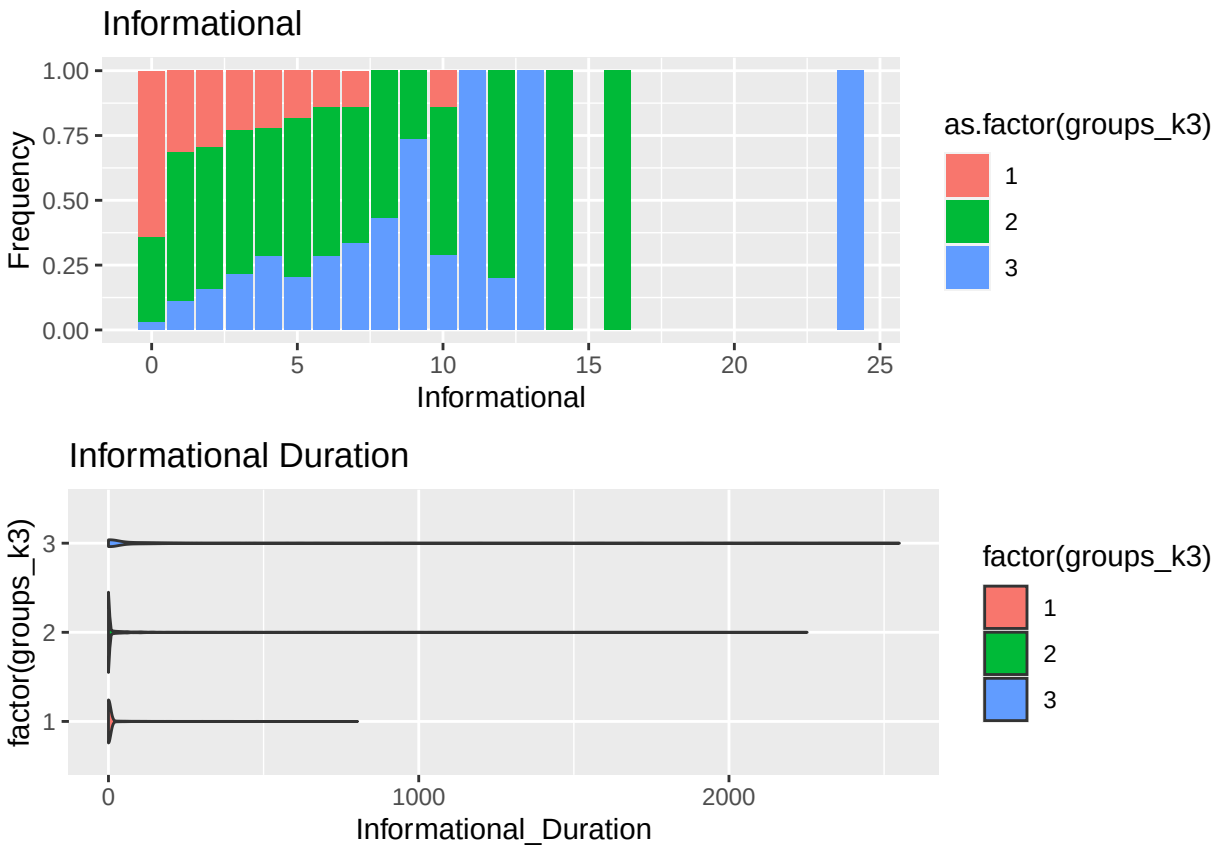
Finally, the hierarchical clustering model was used once more for k=3 user research. The following graph shows the model's cluster plot. The model had 6,946 site-visitors in cluster 1, 4,638 site-visitors in cluster 2, and 746 site-visitors in cluster 3.

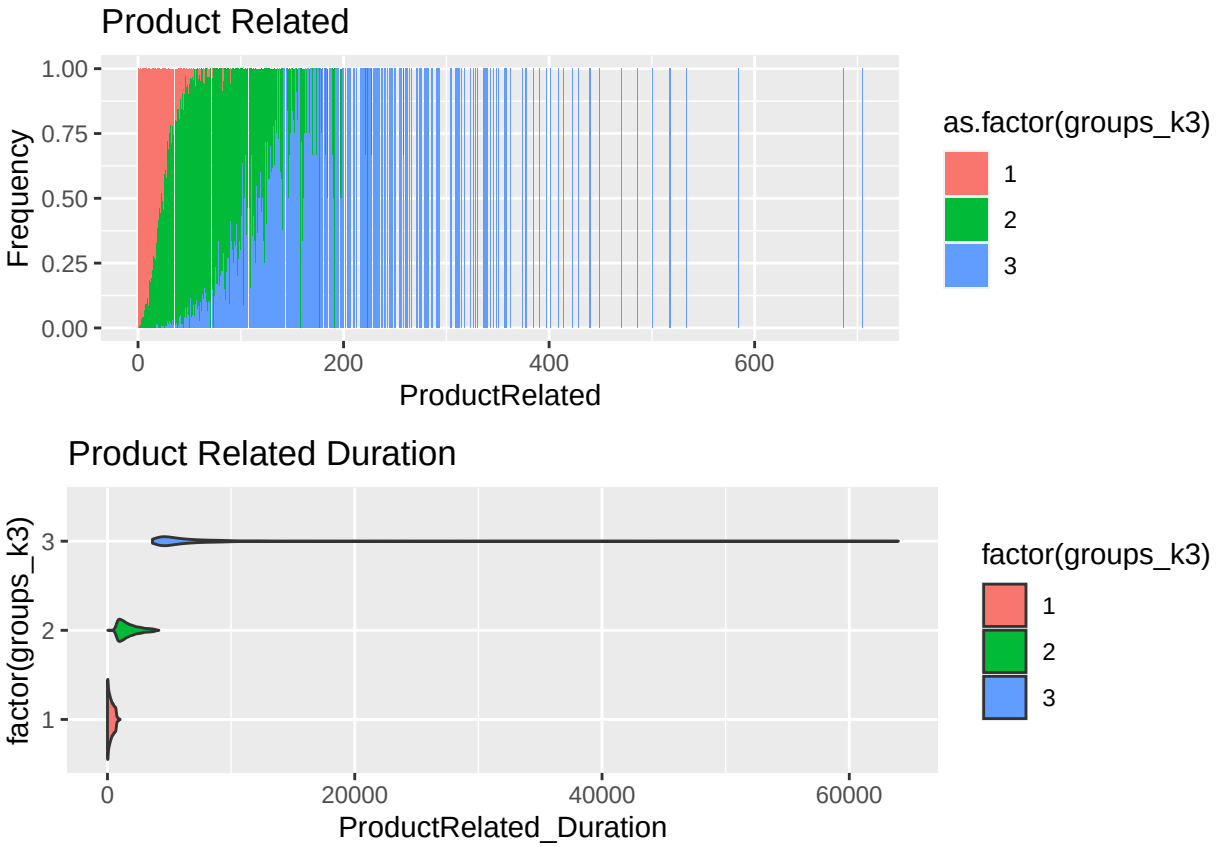


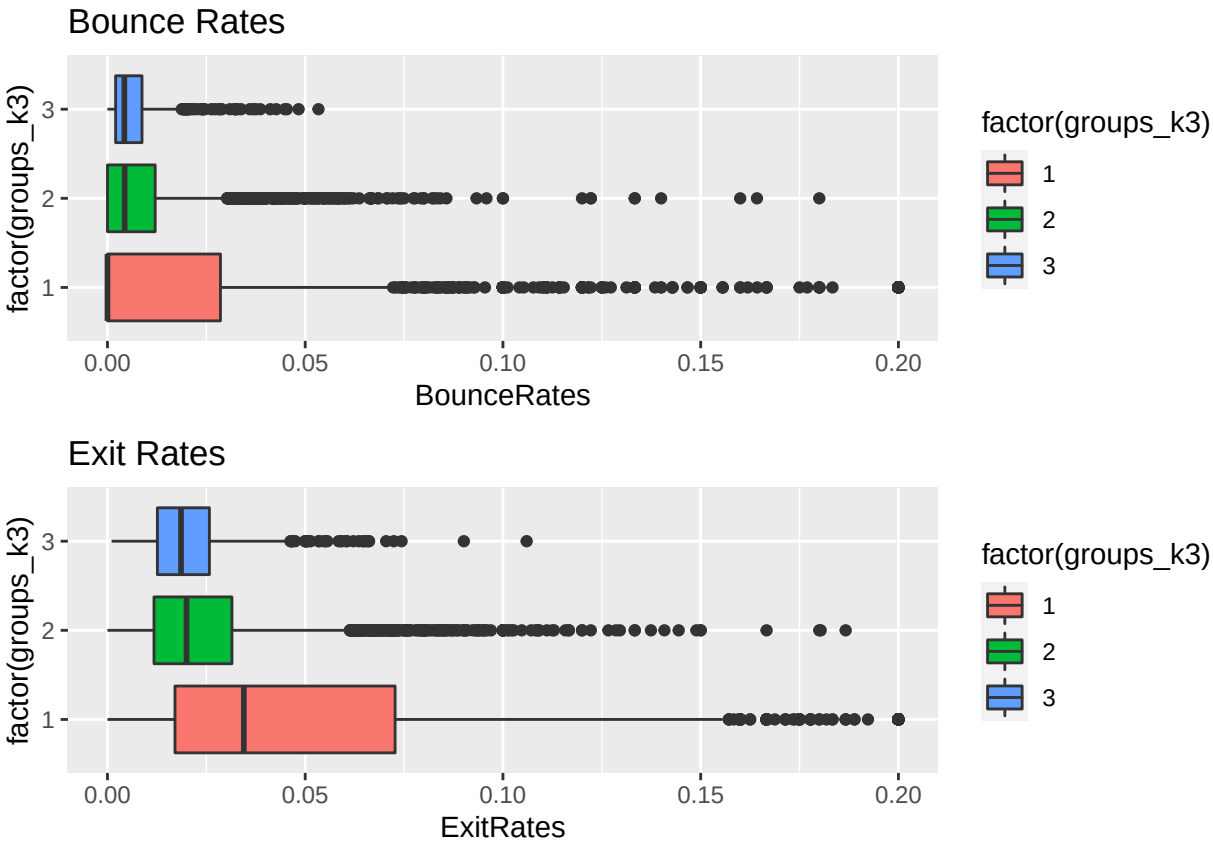

```
##
##    1    2    3
## 6946 4638  746
```

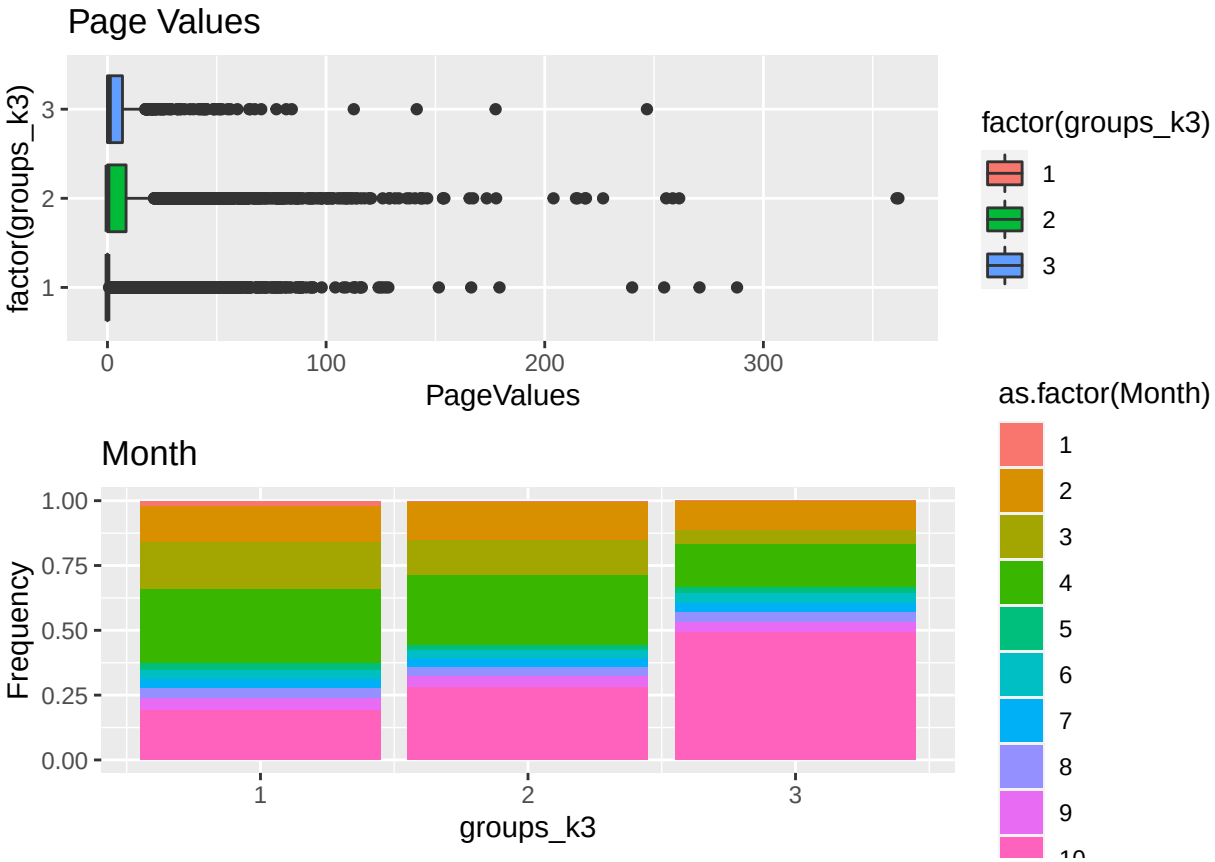
Much of the observations made in hierarchical clustering were similar to those made for k-means and PAM clustering: there was one cluster that was more likely to make a purchase (cluster 3 in this case), one cluster that was the least likely to make a purchase (cluster 1 in this model), and the cluster that was in-between (cluster 2). The cluster that was the most likely to make a purchase spent the most time in websites, visited the most websites, had the lowest bounce and exit rates, had the highest page value, were the most likely to shop around November, used different browser and operating system than other clusters, arrived at the website through a different traffic type than other clusters. Again, all clusters were similar in region, visitor type, and weekend shopping. The following pages will show the plots for the hierarchical clustering, k=3, that correspond with the observation provided in this paragraph.

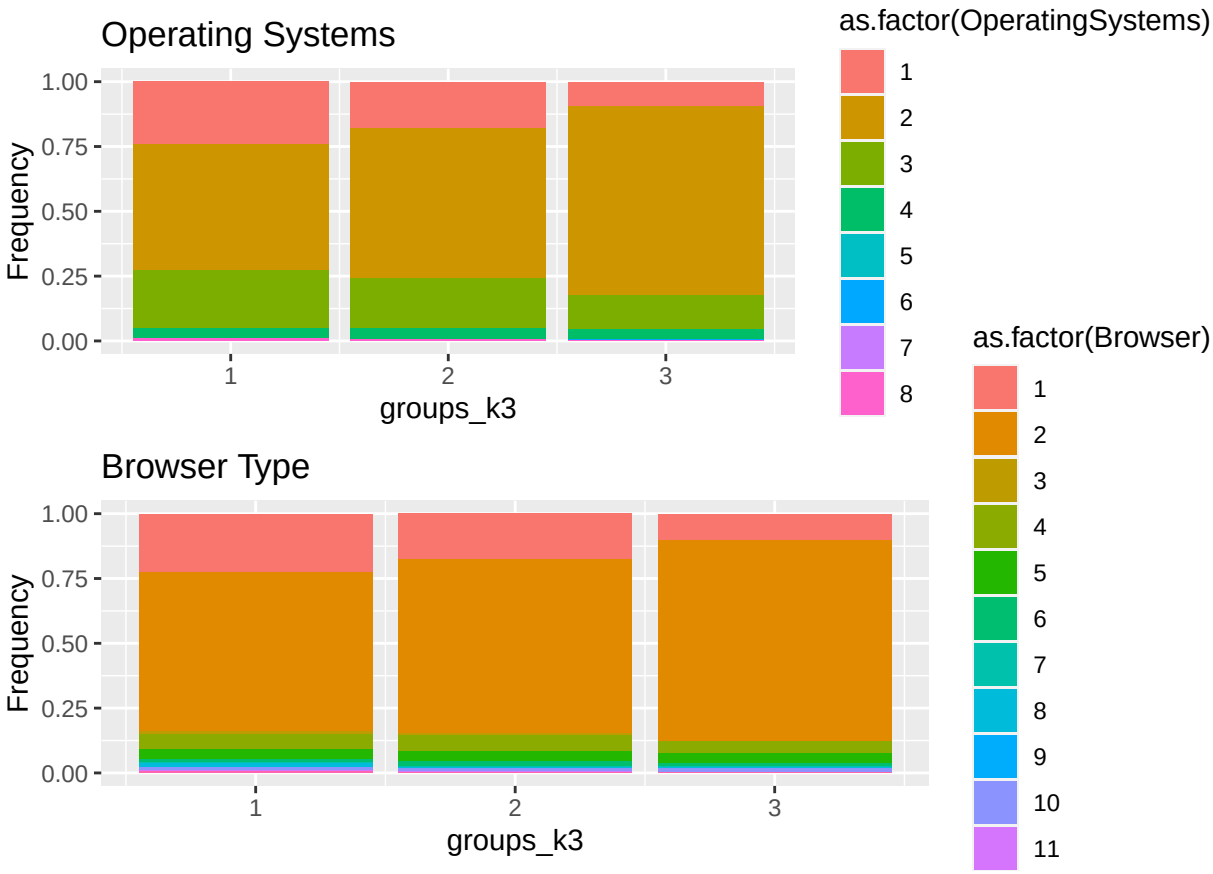

Administrative
Frequency
1.00
0.75
0.50
0.25
0.00
0
10
20
Administrative
as.factor(groups_k3)
1
2
3
Administrative Duration
factor(groups_k3)
3
2
1
0
1000
2000
3000
Administrative_Duration
factor(groups_k3)
1
2
3

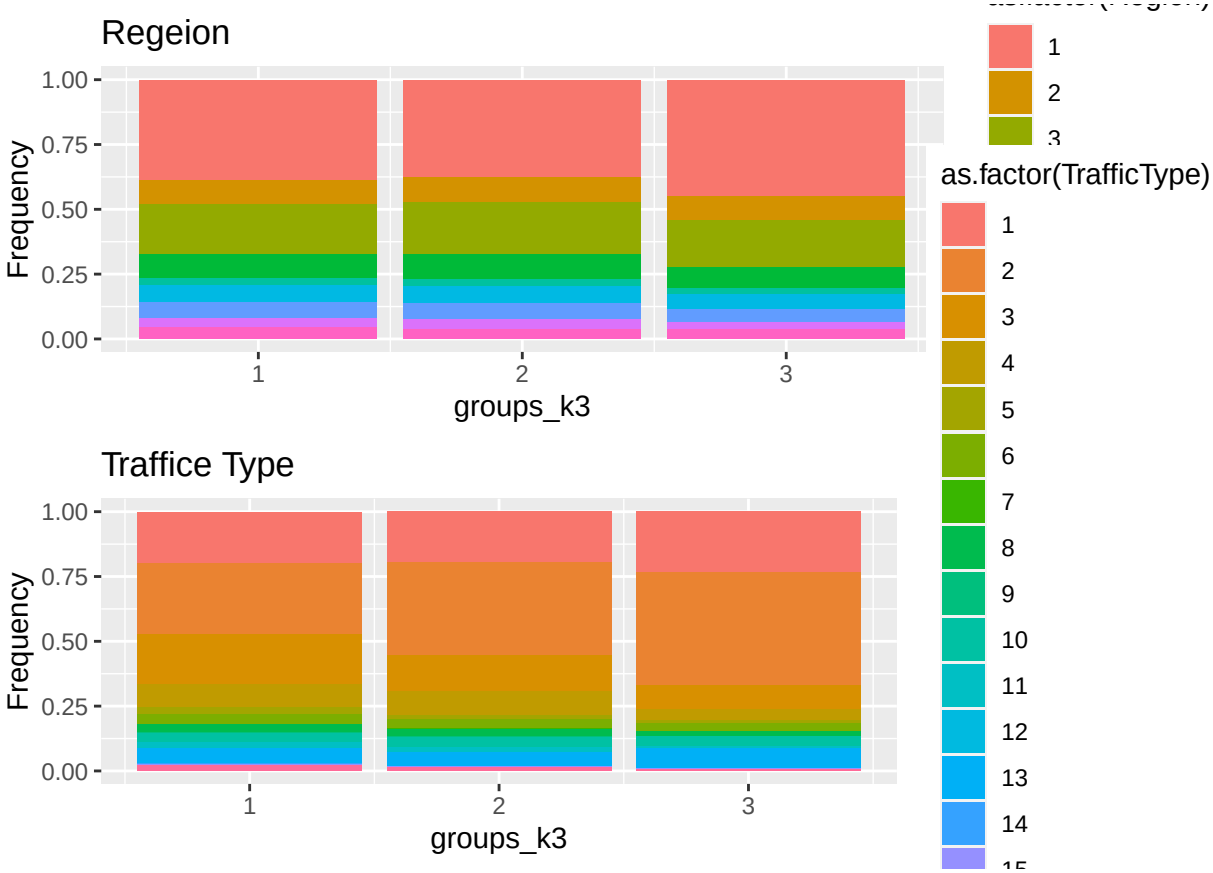

Informational
Frequency
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
Informational
as.factor(groups_k3)
1
2
3
Informational Duration
factor(groups_k3)
3
2
1
0
1000
2000
Informational_Duration
factor(groups_k3)
1
2
3

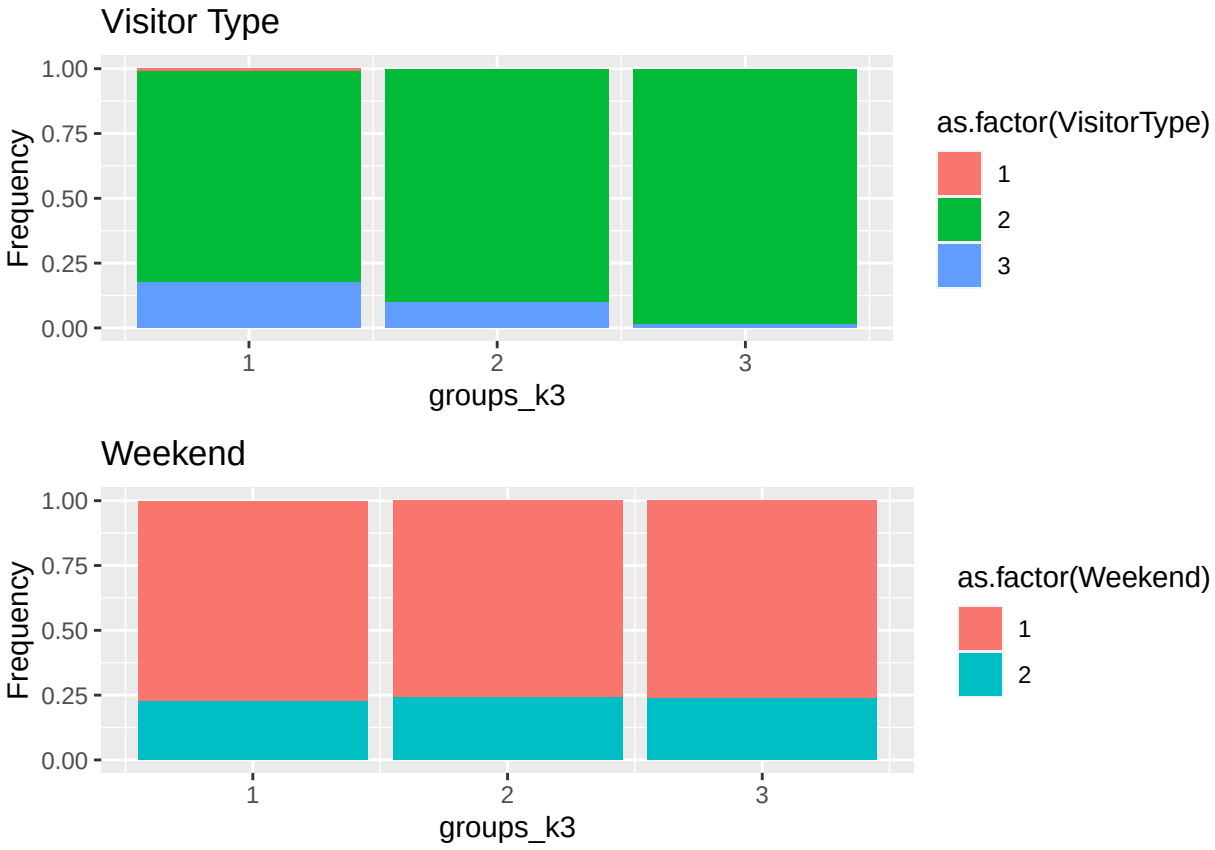

Product Related
Frequency
1.00
0.75
0.50
0.25
0.00
0
200
400
600
ProductRelated
as.factor(groups_k3)
1
2
3
Product Related Duration
factor(groups_k3)
3
2
1
0
20000
40000
60000
ProductRelated_Duration
factor(groups_k3)
1
2
3


Bounce Rates
factor(groups_k3)
3
2
1
0.00
0.05
0.10
0.15
0.20
BounceRates
factor(groups_k3)
1
2
3
Exit Rates
factor(groups_k3)
3
2
1
0.00
0.05
0.10
0.15
0.20
ExitRates
factor(groups_k3)
1
2
3


Page Values
factor(groups_k3)
3
2
1
0
100
200
300
PageValues
factor(groups_k3)
1
2
3
Month
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
groups_k3
as.factor(Month)
1
2
3
4
5
6
7
8
9


Operating Systems
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
groups_k3
as.factor(OperatingSystems)
1
2
3
4
5
6
7
8
Browser Type
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
groups_k3
as.factor(Browser)
1
2
3
4
5
6
7
8
9
10
11


Regeion
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
groups_k3
Traffice Type
Frequency
1.00
0.75
0.50
0.25
0.00
1
2
3
groups_k3
1
2
3
as.factor(TrafficType)
1
2
3
4
5
6
7
8
9
10
11
12
13
14



Model Evaluation - Classification vs Clustering

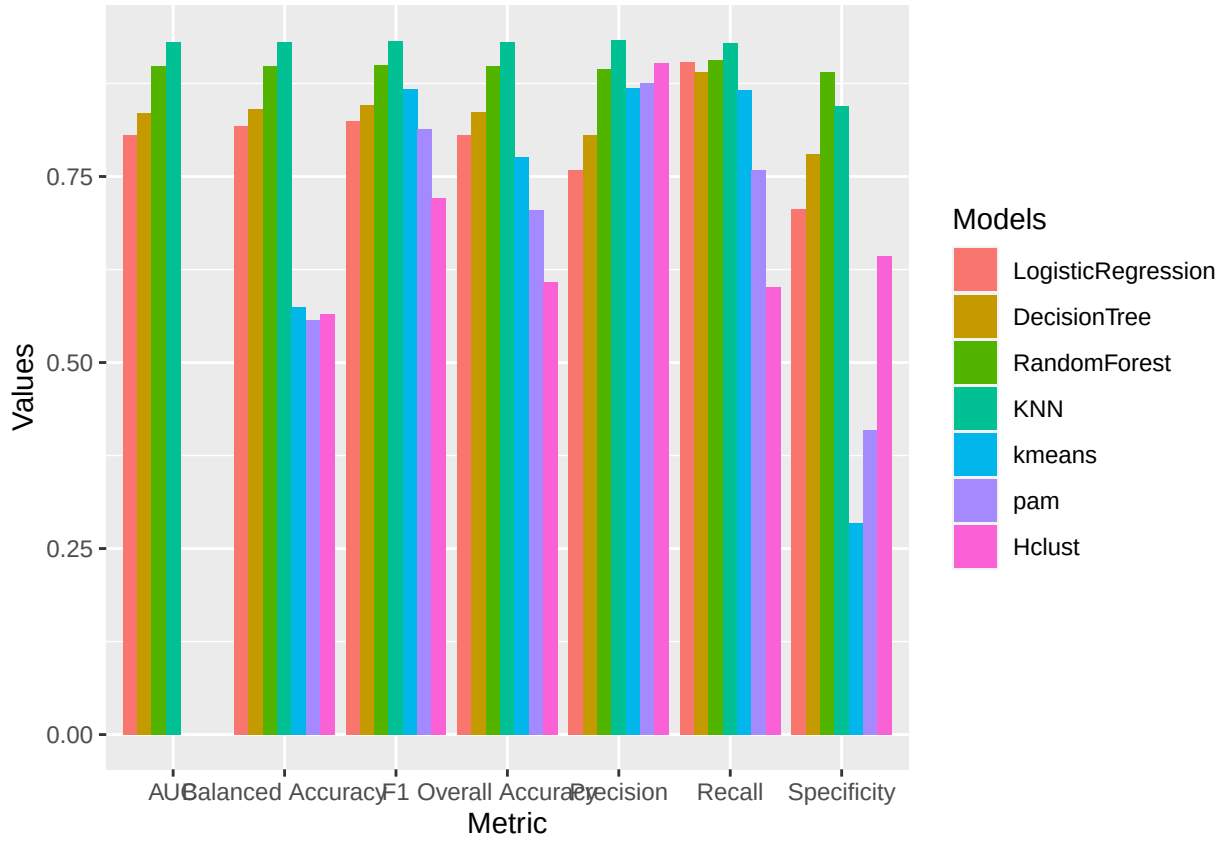
The clustering models were compared against the classification models in predicting online visitors' purchase. The following table and graph shows the evaluation metrics of all classification and clustering models.

```
##                   LogisticRegression DecisionTree RandomForest       KNN
## Precision                  0.7584345    0.8052326    0.8938193 0.9337641
## Recall                     0.9035370    0.8906752    0.9067524 0.9292605
## Overall Accuracy           0.8060065    0.8360390    0.8985390 0.9310065
## Balanced Accuracy          0.8181175    0.8401163    0.8986568 0.9309930
## Specificity                0.7065574    0.7803279    0.8901639 0.8442623
## F1                         0.8246515    0.8458015    0.9002394 0.9315068
## AUC                        0.8050472    0.8355016    0.8984582 0.9310237
##                       kmeans       pam    Hclust
## Precision          0.8686033 0.8754287 0.9018140
## Recall             0.8664364 0.7592593 0.6010363
## Overall Accuracy   0.7763179 0.7051906 0.6074615
## Balanced Accuracy  0.5744257 0.5565232 0.5647629
## Specificity        0.2840671 0.4098532 0.6425577
## F1                 0.8675185 0.8132162 0.7213266
## AUC                       NA        NA        NA
```

```
## Warning in melt(metrics_v, id.vars = "Metric", value.name = "Values",
## variable.name = "Models"): The melt generic in data.table has been passed
```

```
## a data.frame and will attempt to redirect to the relevant reshape2 method;
## please note that reshape2 is deprecated, and this redirection is now
## deprecated as well. To continue using melt methods from reshape2 while both
## libraries are attached, e.g. melt.list, you can prepend the namespace like
## reshape2::melt(metrics_v). In the next version, this warning will become an
## error.
```

```
## Warning: Removed 3 rows containing missing values (geom_bar).
```



As shown above, clustering models overall did much poorly than classification models in predicting online visitors' purchase. The clustering models did especially poorly in the balanced accuracy and specificity scores.

While clustering models were very useful in researching and learning more about the online visitors' behaviours and trends, they are not very useful models for predicting user purchase.

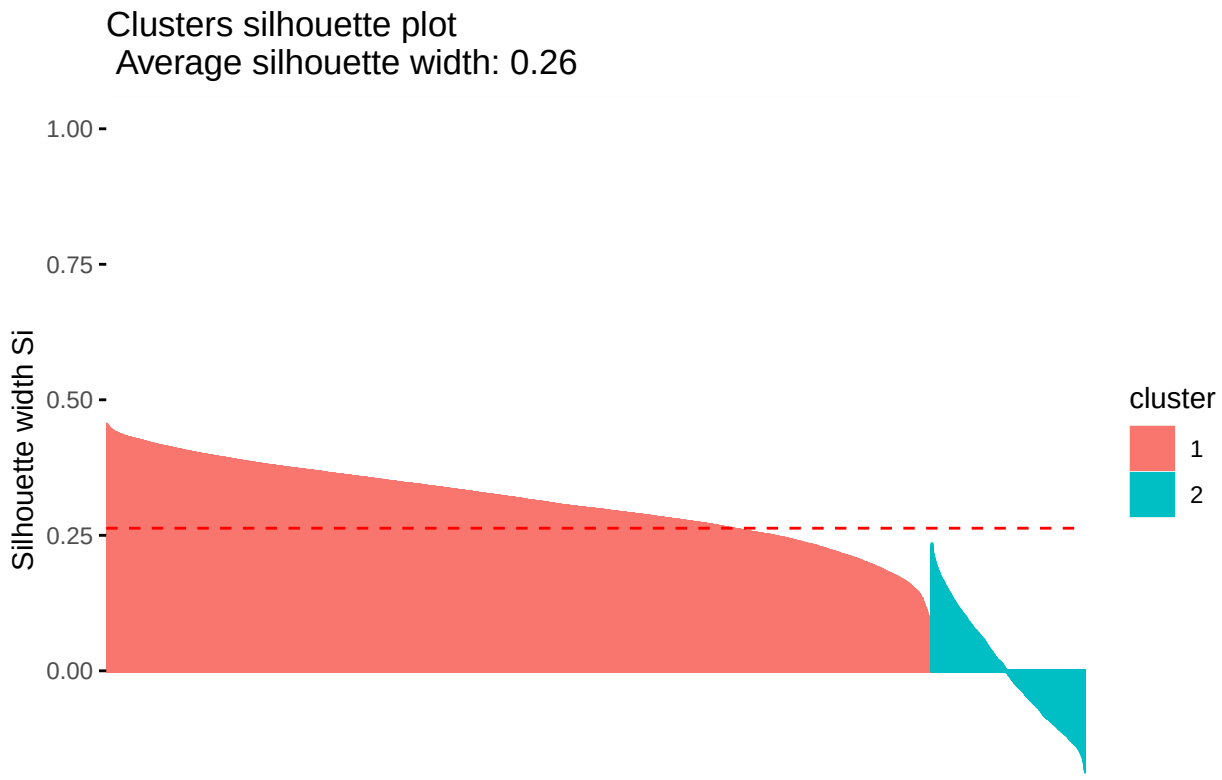
Model Evaluation - Clustering, k=2

To compare which models performed better in clustering for k=2, two metrics were used: internal and external. Silhouette is an internal evaluation method, where the score ranges from 0-1. The closer a score is to 1, the better the model is. Also, the silhouette method produces a visual plot that shows a "tail" that may run beneath the x-axis. The longer this tail is, and deeper it runs beneath the x-axis, the worse the model is, since that tail represents the amount of data points that are misplaced. For the external evaluation method, Rand Index score was used.

The following list shows how all three clustering models performed in the silhouette and Rand index score.

```
## Loading required package: gtools
```

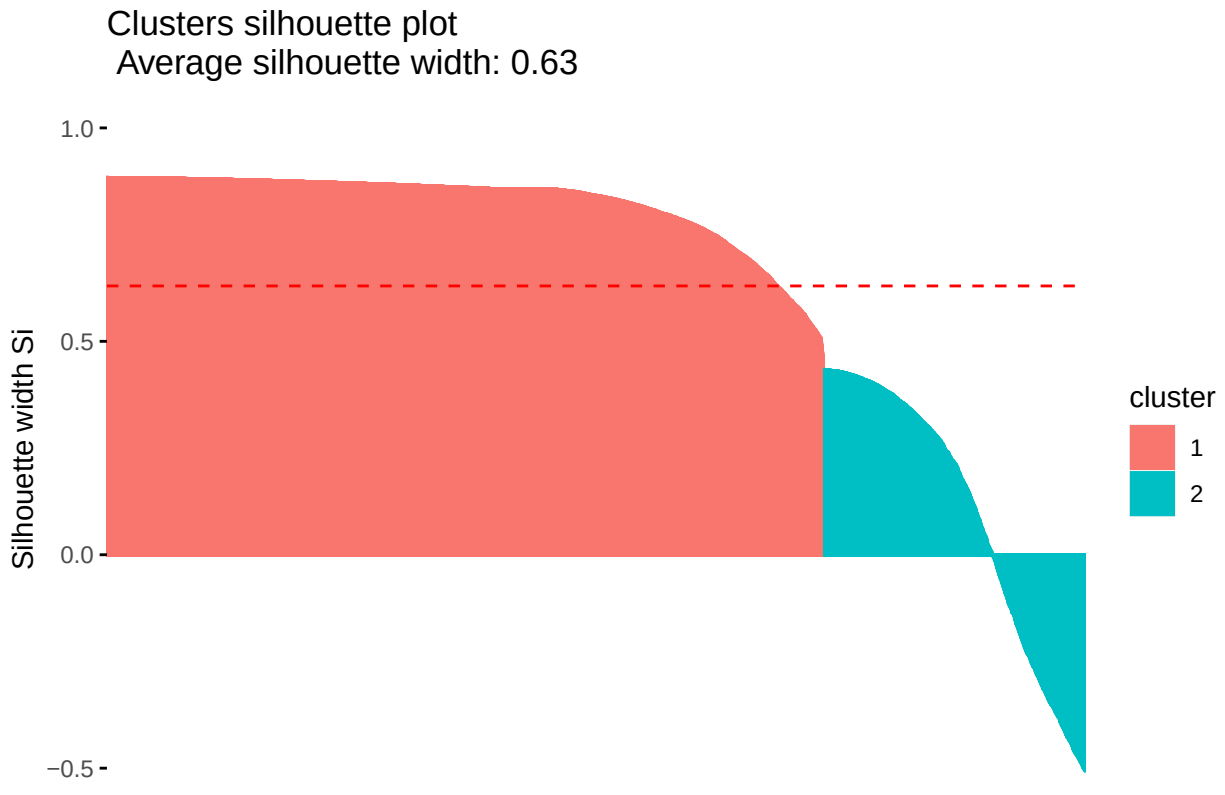
```
##   cluster  size ave.sil.width
## 1       1 10396          0.31
## 2       2  1934          0.00
```



K-means

- k=2 Silhouette score: 0.26
- k=2 misplaced: 997
- Rand Index: 0.653

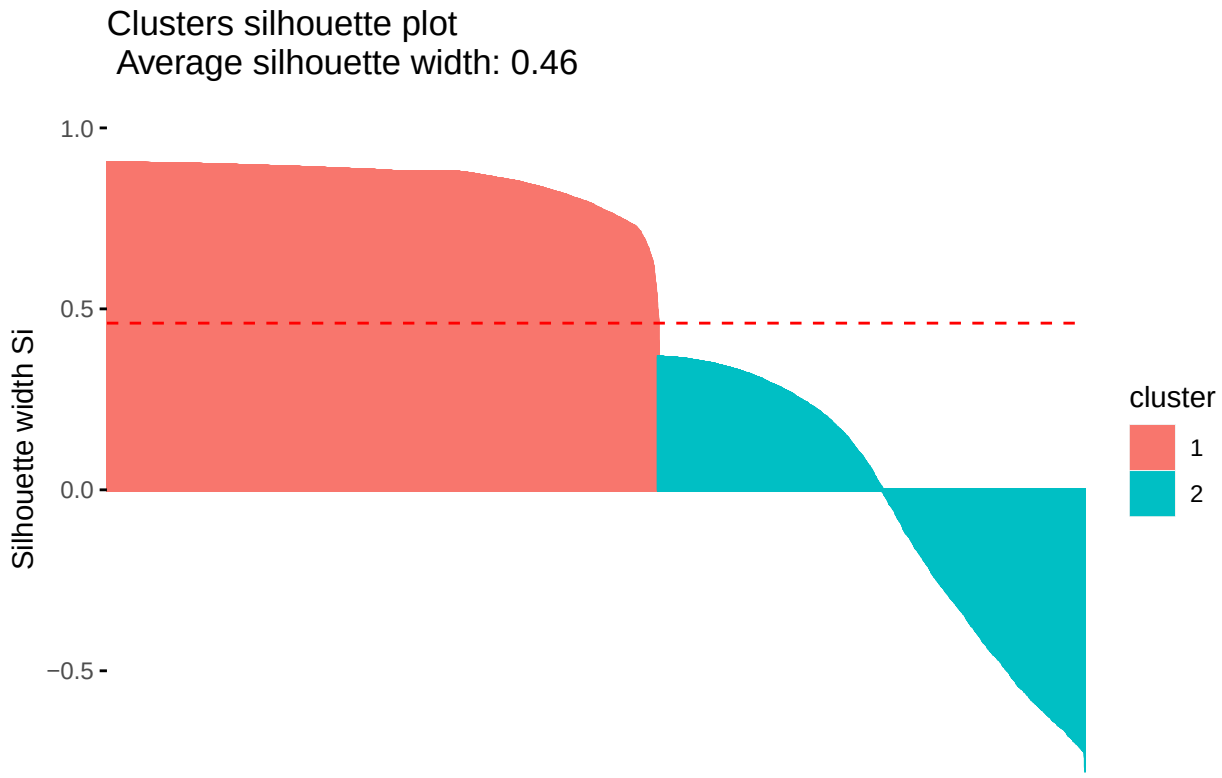
```
##   cluster size ave.sil.width
## 1       1 9039          0.82
## 2       2 3291          0.10
```



PAM

- k=2 Silhouette score: 0.63
- k=2 misplaced: 1,172
- Rand Index: 0.584

```
##   cluster size ave.sil.width
## 1       1 6946          0.86
## 2       2 5384         -0.05
```



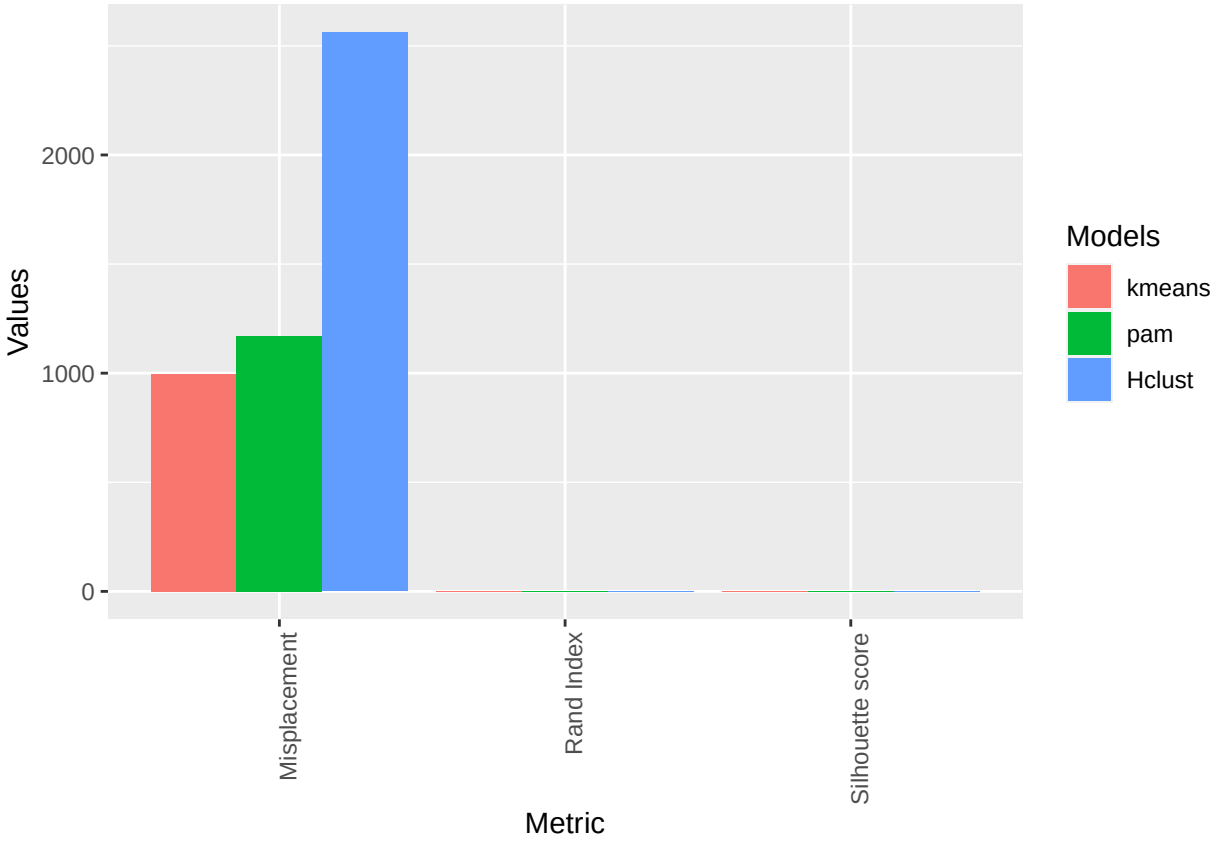
Hierarchical

- k=2 Silhouette score: 0.46
- k=2 misplaced: 2,561
- Rand Index: 0.523

The metrics were summarized into a table and a visual plot.

```
##                       kmeans          pam       Hclust
## Rand Index          0.652675    0.5841726    0.5230573
## Silhouette score    0.260000    0.6300000    0.4600000
## Misplacement      997.000000 1172.0000000 2561.0000000
```

```
## Warning in melt(metrics_k2clustering2, id.vars = "Metric", value.name =
## "Values", : The melt generic in data.table has been passed a data.frame and will
## attempt to redirect to the relevant reshape2 method; please note that reshape2
## is deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(metrics_k2clustering2). In the
## next version, this warning will become an error.
```



It seems that PAM performed the best in Silhouette score, k-means clustering performed the best in Rand index score, and PAM had the least misplacement.

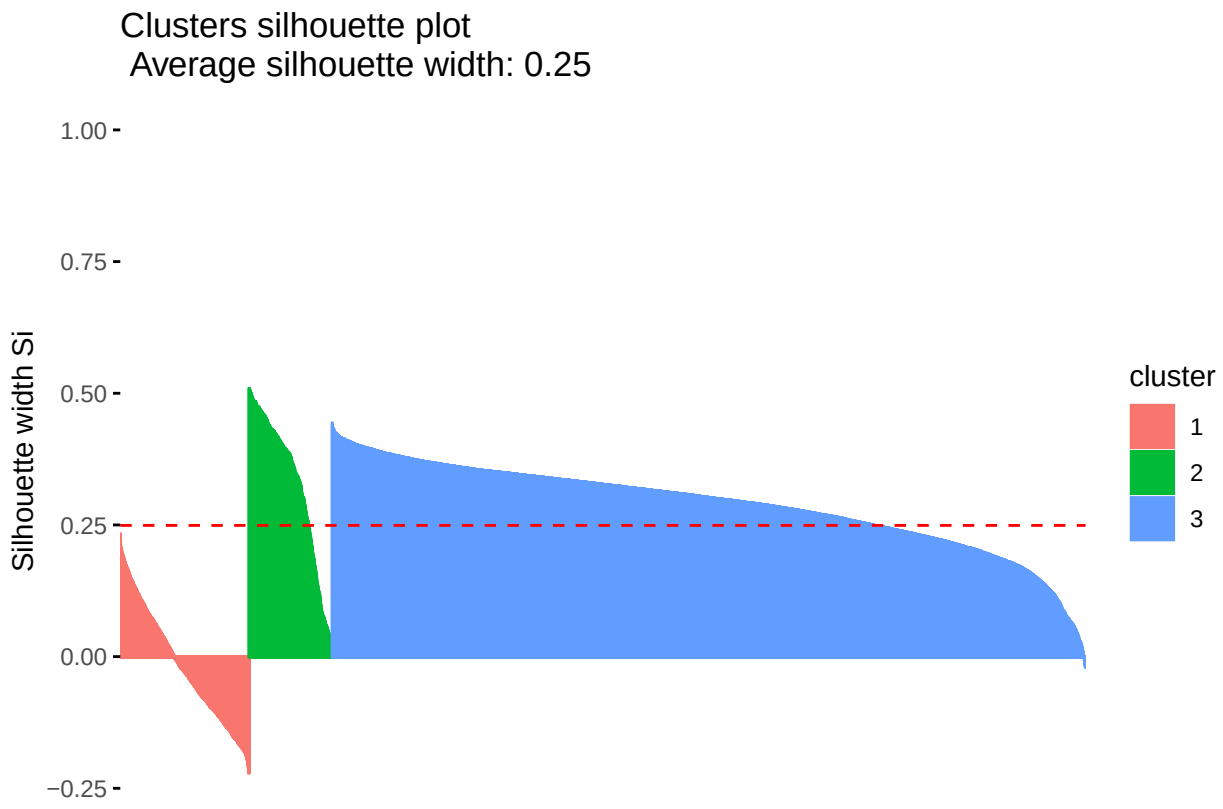
Overall, it seems like the PAM model performed the best due to its consistent performance across the different evaluation metrics.

Model Evaluation - Clustering, k=3

Most of the same metrics were used again to compare which clustering model performed better in clustering for k=3. Here, Silhouette score and its displacement value were used again to evaluate clustering models.

The following list shows how the three clustering models performed in the metrics.

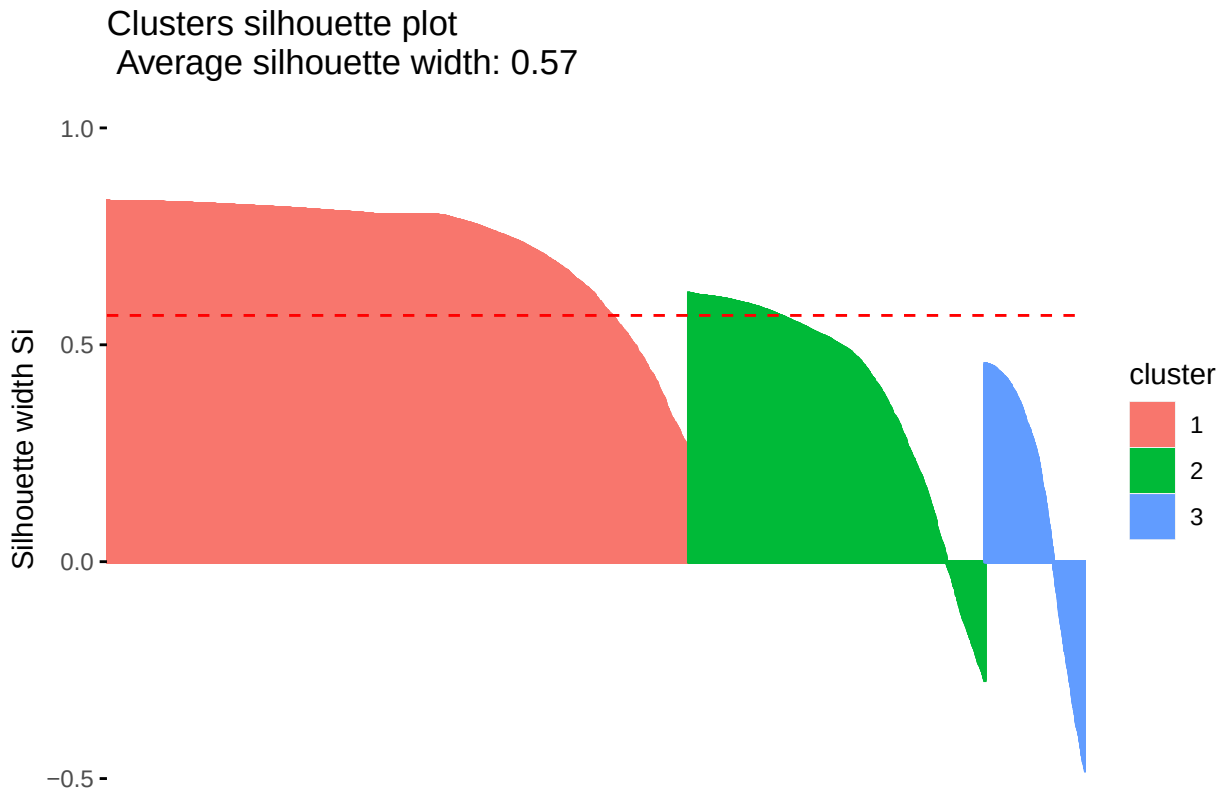
```
##   cluster size ave.sil.width
## 1       1 1646         -0.02
## 2       2 1061          0.32
## 3       3 9623          0.29
```



K-means

- k=3 Silhouette score: 0.25
- k=3 misplaced: 982

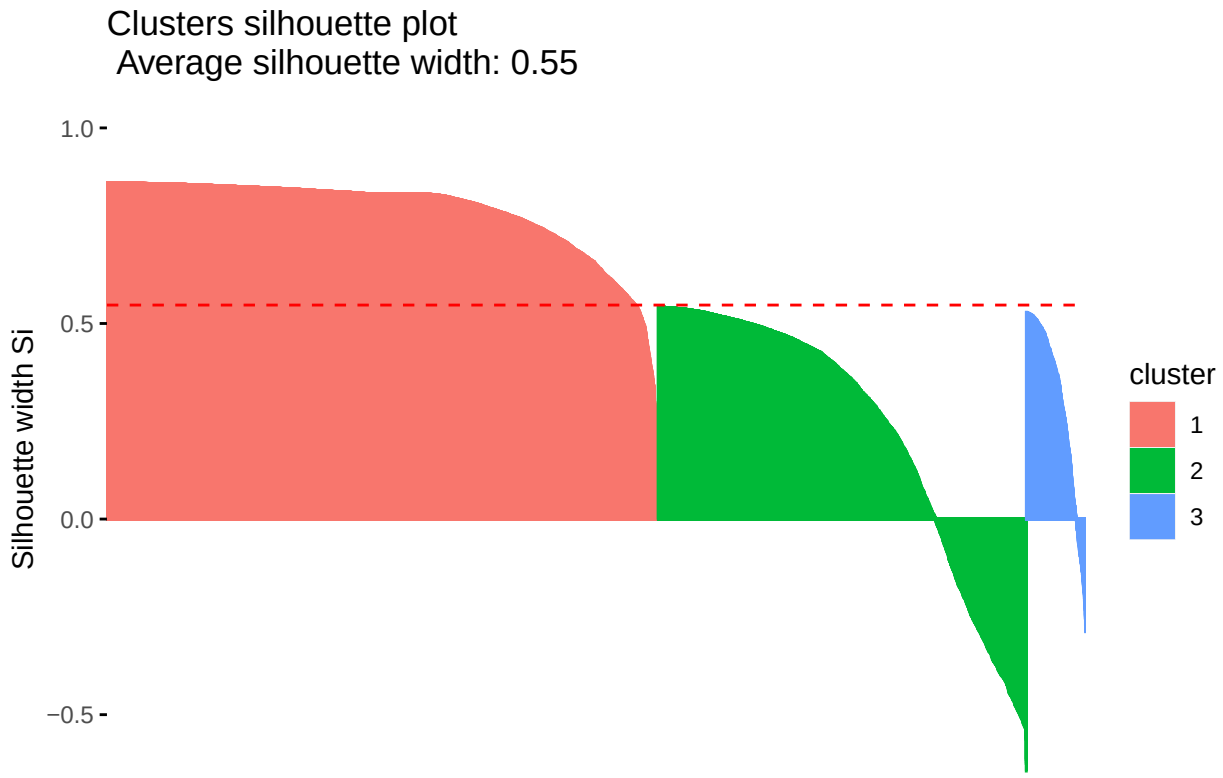
```
##   cluster size ave.sil.width
## 1       1 7326          0.73
## 2       2 3737          0.39
## 3       3 1267          0.14
```



PAM

- k=3 Silhouette score: 0.57
- k=3 misplaced: 883

```
##   cluster size ave.sil.width
## 1       1 6946          0.79
## 2       2 4638          0.23
## 3       3  746          0.30
```



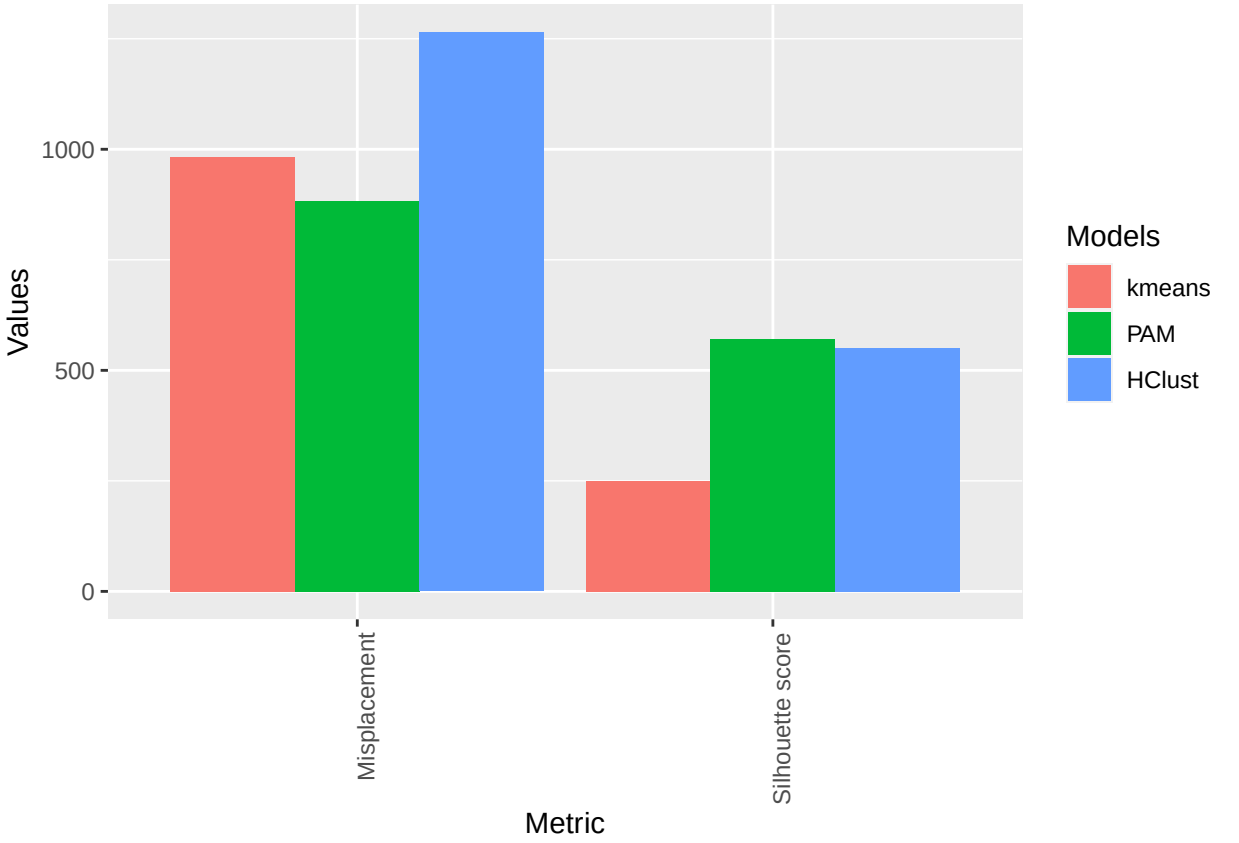
Hierarchical

- k=3 Silhouette score: 0.55
- k=3 misplaced: 1,264

These metrics were also made into a table and a visual plot

```
##                  kmeans    PAM  HClust
## Silhouette score   0.25   0.57    0.55
## Misplacement     982.00 883.00 1264.00
```

```
## Warning in melt(metrics_k3_clustering, id.vars = "Metric", value.name =
## "Values", : The melt generic in data.table has been passed a data.frame and will
## attempt to redirect to the relevant reshape2 method; please note that reshape2
## is deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(metrics_k3_clustering). In the
## next version, this warning will become an error.
```



It seems that the PAM model again performed the best in Silhouette score, and also misplaced the least.

The PAM model once again performed the best in the evaluation metrics for k=3 as well.

Moreover, the calculation for the higest average silhouette score also results in the PAM model.

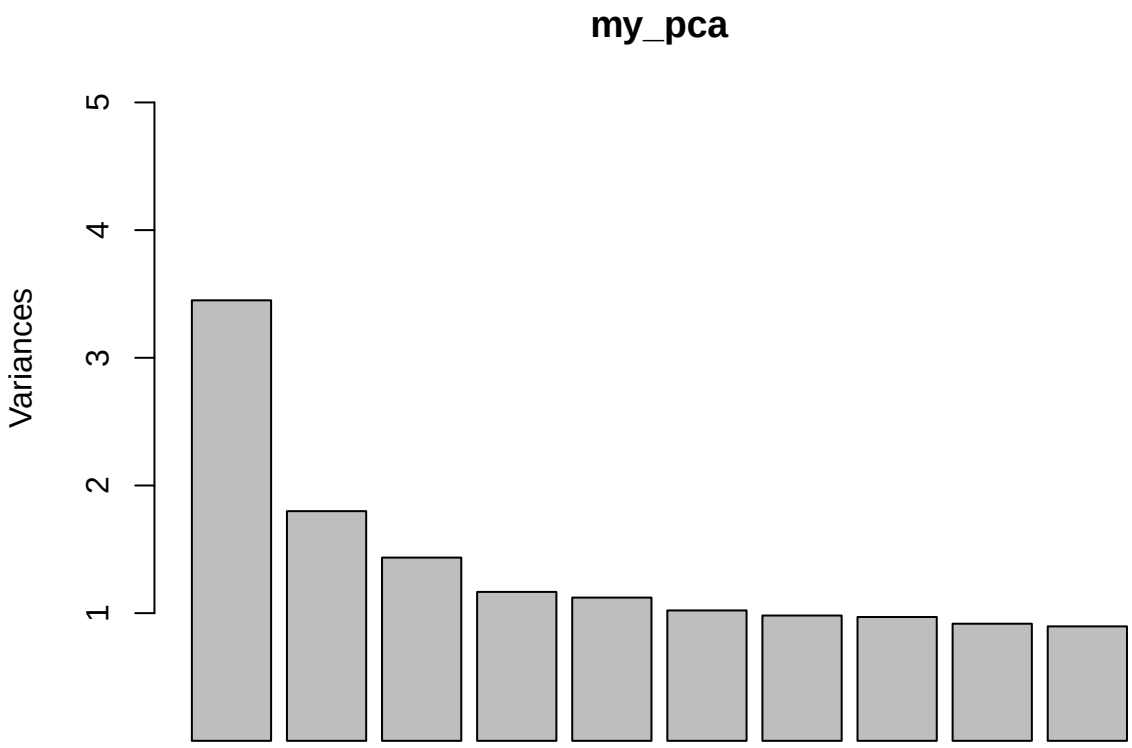
```
#The highest average silhouette
best_sil=max(mean(k3_sil[,3]),pam.res_k3$silinfo$avg.width,mean(hc3_sil[,3]))
#The name of the best clustering method for customer base research
if (mean(k3_sil[,3])==best_sil) {"kmeans"} else if (pam.res_k3$silinfo$avg.width==best_sil) {"pam"} else
```

```
## [1] "pam"
```

PCA

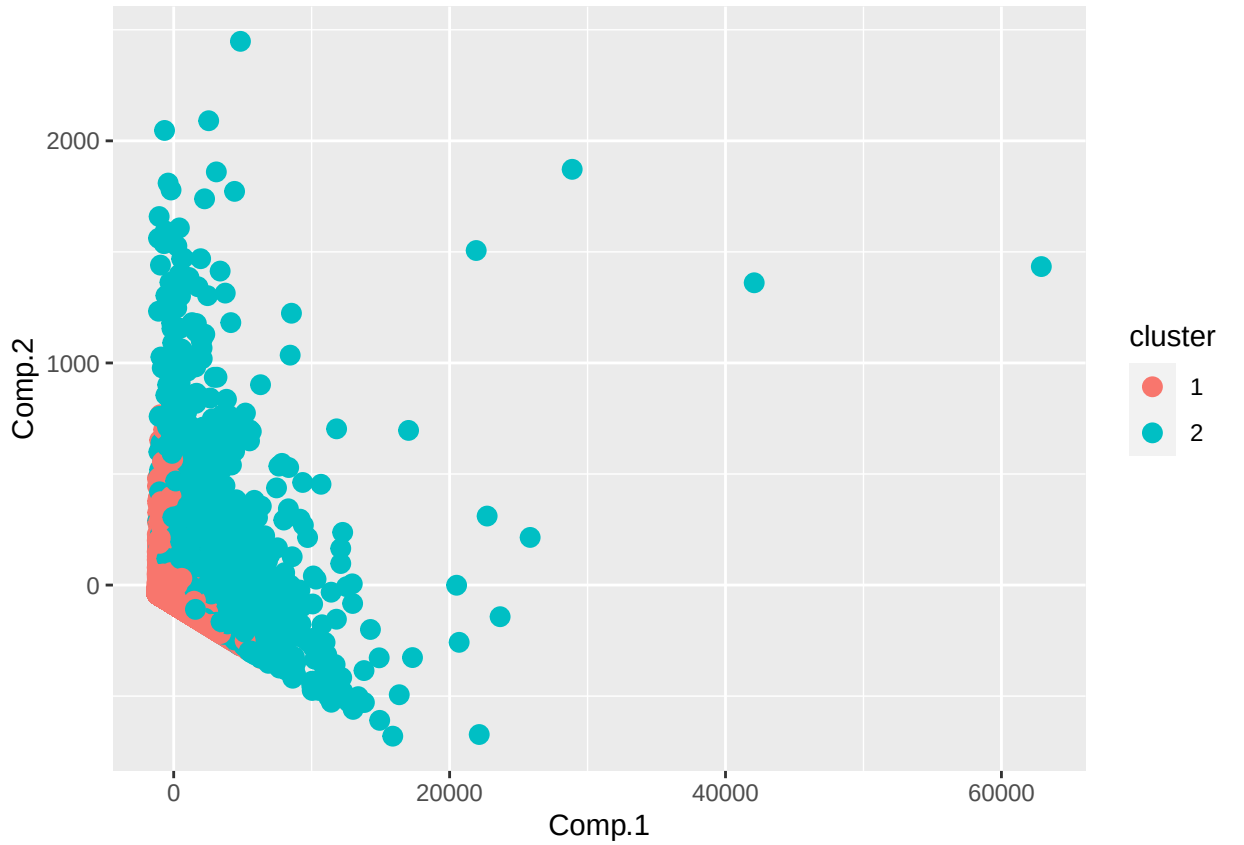
Principal component analysis (PCA) is used to extract the important information from a multivariate data table and to express this information as a set of few new variables called principal components. These new variables correspond to a linear combination of the originals. The goal of PCA is to identify directions (or principal components) along which the variation in the data is maximal.

PCA was performed, and produced the following graph, which shows the PCA number in the x-axis, and the variation representation in the y-axis.



This graph shows that PCA1 and PCA2 represent the majority of data variation.

PCA1 and PCA2 were plotted on the graph as follows.



There seems to be a great deal of overlap between cluster 1 and 2 on this graph. This graph might indicate that the variation between cluster 1 and 2 are not enough to make a clear distinction between the e-commerce visitors who are likely to make a purchase and the visitors who are unlikely to make a purchase. This supports the conclusion that was made in the previous section, where the clustering models' metric performed much worse than the predictive classification models.

Motivations and behaviours of online visitors are complex, and the current dataset might not be sufficient enough to capture the variance between cluster 1 and 2. There may be a need for a more complex type of data collection that delves further into the thoughts, motivations, and behaviours of site-visitors to be able to show more variance between the cluster 1 and 2.

Discussion

Both classification (predictive) and clustering models were used in this project to assess how well the models can predict the purchase status of site-visitors. Within the classification models, the k-NN model performed the best, and random forest a close second. Overall, all of the classification models performed well in predicting online visitors' purchase, while clustering models did not perform well in the same evaluation metrics. Therefore, classification models, specifically the k-NN model, would be best suited for predicting the purchase status of site-visitors.

From the feature importance results, the best recommendations for e-commerce business owners would be to focus on three important features: Page Value, Bounce Rate, and Exit Rate. Online shopping web pages should aim for a high page value, this can be achieved if users go through a maxium of one page before arriving to the target or transaction page. Also, the shopping pages should be designed to capture site-visitors' interest to increase their stay, and thereby decreasing the Bounce and Exit Rate.

The clustering models were used as a research method to better understand the user base. From the clusters, it was observed that a meaningful way to segment the users was by diving either into 2 or 3 clusters.

Dividing into 2 clusters showed one cluster that had the following set of behaviours: quick exit and abandonment (exit and bounce rate) from the website, less likely to be affected by special holidays to make purchases, and spends less time setting up their account information, looking up product related pages, or reading upon the company's information. Also, this particular cluster arrived at the shopping website through a different traffic compared to another cluster. Although the traffic types were not defined in the Sakar et al paper, it can be assumed that the traffic type of this cluster would be irrelevant to the company's advertisements or promotions. On the other hand, the other cluster displayed these behaviours in the opposite direction: more likely to spend time setting up their account, looking at the company's information and product pages, more likely to make purchase on special days/ holidays, and had a different traffic type, which is likely to be relevant to the company's advertisements and promotions. Overall, it seems that the second cluster is representative of those who are likely to make a purchase. Another interesting insight was that the first cluster used a different browser and computer operating system than the second cluster. Perhaps this might indicate the generation gaps between the clusters, since older site-visitors are likely to use certain types of computer operating system and browser than younger site-visitors (Agarwal, 2018; Bursztein, 2012; Thubron, 2019). Moreover, location of visitors (region) had no effect on the likelihood of purchase, which supports the notion that the online shopping in the age of Internet-of-Things (IoT) is unaffected by locations and regions like the brick-and-mortar stores.

Dividing the user base into 3 clusters mainly produces the similar insights as dividing into 2 clusters. However, it further divides to create following levels of clusters: most likely to make a purchase, less likely to make a purchase, and the least likely to make a purchase. The cluster that spends the most time in setting up accounts, and looking at the product and company related pages were the most likely to purchase out of the three clusters. This cluster was also more likely to shop on special days/ holidays. Perhaps this cluster represents the ideal type of users that e-commerce business owners can target to send more advertisements and promotional materials.

Deployment

A deployment interface is created and launched in this shinyapp.io: https://lily-ye.shinyapps.io/DeploymentFinalProject/. The objective of the deployment would be to inform the e-commerce business owners of their website's metrics (e.g. bounce and exit rates), as well as the prediction of their website visitor's likelihood of making a purchase.

The website's bounce rate and exit rate metrics would inform the business owners on how well their websites are designed to influence the visitor's stay (e.g. designing a website with a great slogan or pictures to capture their interests to stay longer). The deployment interface allows it's user to observe metrics and its changes over any time period via its sidebar's date range selection. With this interface, business owners can observe the effects of their operational and business decisions on the website's metrics. For instance, if the business owner changed the layout of the company's shopping website in January 2020, they can observe whether this might increase or decrease the bounce and exit rates.

Moreover, using the k-NN model that performed the best among all predictive models, the second tab in the deployment interface has the list of website visitors, and the prediction output of whether they are likely to make a purchase. Then, the business owners can see these prediction outputs, and can use the information for targeted marketing and promotions to further encourage those target users to make purchases.

Overall, this interface provides business owners to observe website metrics, purchase predictions of site-visitors and thorough understanding of their user base. With this information, business owners are better equipped to improve shopping websites and to identify target users easily, which can ultimately contribute towards increasing sales and revenue.

Conclusion

Within the classification models, k-NN had the best overall performance in the evaluation metrics, while PAM performed the best within the clustering models. However, clustering models overall performed worse than classification in predicting site-visitors' purchase, and therefore, k-NN model is the ideal machine learning model for prediction.

Clustering was an excellent research method to look further into the trends and behaviours of the site-visitor base. k=2 and k=3 were both used in the clustering models, and all clustering models divided clusters in a similar way: a cluster that is more likely to make a purchase, and a cluster that is less likely to make a purchase. The two clusters had a different set of behaviours and trends, and this result was consistently replicated in all clustering models.

References

Agarwal, D. (2018). What web browser do old site-visitors use? Quora. https://www.quora.com/What-web-browser-do-old-site-visitors-use

Bursztein, E. (2012). Survey: Internet explorer users are older, chrome seduces youth. Elie Bursztein's Site. https://www.elie.net/blog/web/survey-internet-explorer-users-are-older-chrome-seduces-youth

Dabbura, I. (2018, September 17). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Retrieved May 8, 2020, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Franklin, J. S. (2019, November 23). Elbow method of K-means clustering using Python - Analytics Vidhya - Medium. Retrieved May 8, 2020, from https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540

Google Analytics. (2020a). Bounce rate—Analytics Help. Google. https://support.google.com/analytics/answer/1009409?hl=en

Google Analytics. (2020b). Exit Rate vs. Bounce Rate—Analytics Help. Google. https://support.google.com/analytics/answer/2525491?hl=en&ref_topic=6156780

Google Analytics. (2020c). How Page Value is calculated—Analytics Help. Google. https://support.google.com/analytics/answer/2695658?hl=en

Kassambara, A. (n.d.). K-Medoids in R: Algorithm and Practical Examples - Datanovia. Retrieved May 8, 2020, from https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/#pam-algorithm

Kassambara, A. (n.d.) (2017, September 23). PCA - Principal Component Analysis Essentials - STHDA. Retrieved May 8, 2020, from http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/#pca-data-format

Kilitcioglu, D. (2018, October 26). Hierarchical Clustering and its Applications - Towards Data Science. Retrieved May 8, 2020, from https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6

Koks, P. (March 10, 2020). How Page Value in Google Analytics Can Improve Your Insights. Retrieved from https://online-metrics.com/page-value/

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. https://doi.org/10.1109/ICDM.2010.35

Molnar, C. (2020.). 4.4 Decision Tree | Interpretable Machine Learning. Retrieved April 24, 2020, from https://christophm.github.io/interpretable-ml-book/tree.html

R Documentation. (n.d.). ROSE function. Retrieved May 8, 2020, from https://www.rdocumentation.org/packages/ROSE/versions/0.0-3/topics/ROSE

Read Random Forest-Random Forest (4 implementation steps + 10 advantages and disadvantages). (n.d.). Retrieved April 24, 2020, from https://easyai.tech/en/ai-definition/random-forest/

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31(10), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0

Sheldon, P., Wigder, Z. D., Wray, J., Varon, L., & Katz, R. (October 14, 2014). Canadian Online Retail Forecast, 2014 To 2019. Retrieved May 7, 2020, from https://www.forrester.com/report/Canadian+Online+Retail+Forecast+2014+To+2019/-/E-RES115497

Thubron, R. (2019). 71% of students own or would prefer a Mac, claims survey. TechSpot. https://www.techspot.com/news/80220-71-students-own-or-would-prefer-mac-claims.html

Trevino, A. (2016, December 6). Introduction to K-means Clustering | Oracle Data Science. Retrieved May 7, 2020, from https://blogs.oracle.com/datascience/introduction-to-k-means-clustering

Xiaoqiang. (2019). What is the K-nearest neighbors|KNN? - Product Manager's Artificial Intelligence Learning Library. Easy AI. https://easyai.tech/en/ai-definition/k-nearest-neighbors/