



Ds – 807

PROJECT REPORT

Online Shopping Retention

Shawn Bedard & Prem Kumar Boini

Contents

| | |
|--|--|
| 1. Introduction and Overview..... | |
| 2. Data Preparation..... | |
| 3. Exploratory Data Analysis..... | |
| 4. Clustering Model..... | |
| <i>K-Means</i> | |
| <i>K-Medoids</i> | |
| <i>Hierarchical Model</i> | |
| 5. Mixture Model..... | |
| <i>Mclust</i> | |
| 6. Neural Network Model..... | |
| <i>Feed Forward Neural Network</i> | |
| 7. Model Results..... | |
| 8. Conclusion..... | |
| 9. Future Work..... | |

1. Introduction and Overview

In recent years, the advancement of technology and the internet especially, has transformed the way the global population shops for their products or services, leading to a significant rise in online retail stores. Due to the convenience and accessibility of offered from e-commerce platforms has led to an increasing number of consumers who are now utilizing these online channels for all their purchasing needs. As a result, the need to understand the intentions and behaviors of these online shoppers has become a crucial aspect for all e-commerce businesses that aim to optimize their business strategies and create a more personalized experience. By analyzing the different factors of the user's experience such as their browsing patterns, search results, and even their transaction history allows a business to gain a more in-depth understanding of their consumers' tendencies and preferences.

This paper aims to explore the classification of online shoppers' intentions by employing different clustering and mixture models to gain meaningful patterns and insights to make informed decisions. By utilizing advanced machine learning techniques and having access to a large dataset, our team aims to develop an accurate and robust classification model that can be used to efficiently determine a customer's shopping intention. The goal with this analysis is to help inform the e-commerce website to optimize their marketing campaigns and product recommendations.

Overview of Data

The dataset used in this analysis was taken from the University of California, Irvine's (UCI) machine learning repository and is titled Online Shopping Purchasing Intention. The dataset was donated to the repository back on October 31st, 2018. The dataset was created by C. Okan Sakar, who works in the department of Computer Engineering at the Bahcesehir University, and Yomi Kastro who is an Information Technology consultant. The dataset consists of 12,330 observations which belong to individual sessions where each session belongs to a different user over the course of a one-year period. The selection of the same date range is to avoid any tendencies of the data favoring any specific campaign, special day, and user profiles. The variables utilized in this analysis were 'Administrative', 'Informational', 'ProductRelated', 'Administrative_duration', 'Informational_duration', 'Page_duration', 'Bounce_rate', 'Exit_rate', 'Special_day', 'Page_values', 'OperatingSystems', 'Browser', 'TrafficType', 'Region', 'Month', 'Weekend'.

Variables:

| <i>Variable Name</i> | <i>Values</i> | <i>Definition</i> |
|----------------------|-------------------------|----------------------------|
| Administrative | Categorical: 1,2,3,4... | Administrative Page Number |
| Informational | Categorical: 1,2,3,4... | Informational Page Number |

DS 807 UNSTRUCTURED DATA

| | | |
|-------------------------|------------------------------------|--|
| Product Related | <i>Categorical: 1,2,3,4...</i> | Product Page Number |
| Administrative Duration | Continuous: 0.0 - 3,339 | Time spent on administrative pages in seconds |
| Informational Duration | Continuous: 0.0 - 2,950 | Time spent on informational pages in seconds |
| Product Duration | Continuous: 0.0 - 63,974 | Time spent on administrative pages in seconds |
| Bounce Rate | Continuous: 0.0-0.20% | % of viewers who leave after viewing 1 page |
| Exit Rate | Continuous: 0.0-0.20% | % of viewers who leave on a specific page |
| Special Day | Categorical: 0.2, 0.4, 0.6, 0.8, 1 | Closeness of the site visiting time to a specific special day (e.g., Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction |
| Page Values | Categorical: 1,2,3,4,5... | Represents the average value for a web page that a user visited before completing an e-commerce transaction |
| Operating Systems | Categorical: 1,2,3,4... | Operating system type of user when on website |
| Browser | Categorical: 1,2,3,4... | Browser type being used to access website |
| Traffic Type | Categorical: 1,2,3,4 | The sources from which users come to the site |
| Region | Categorical: 1,2,3,4... | Region where the consumer is using website |
| Month | Categorical: 1,2,3,4... | The time in months when user was active |
| Weekend | Binary: 0,1 | Whether the user accessed the website on a weekend |

2. Data Preparation

For our analysis of the classification of online shoppers to be accurate and usable, the data needed to be prepared to function properly with the models our team wanted to explore. The variables in the dataset were cleaned and modified to simplify the data and we began with determining whether the dataset consisted of any N/A values which would require some type of data imputation. Based on the figure below, we were able to determine that the dataset did not consist of any N/A values which enabled us to be able to freely use all the observations within the dataset with no issues.

Figure 1. N/A Values

| | | | |
|----------------|-------------------------|---------------|------------------------|
| Administrative | Administrative_Duration | Informational | Informational_Duration |
| 0 | 0 | 0 | 0 |
| ProductRelated | ProductRelated_Duration | BounceRates | ExitRates |
| 0 | 0 | 0 | 0 |
| PageValues | SpecialDay | Month | OperatingSystems |
| 0 | 0 | 0 | 0 |
| Browser | Region | TrafficType | VisitorType |
| 0 | 0 | 0 | 0 |
| Weekend | Revenue | | |
| 0 | 0 | | |

After exploring the possibility of N/A values within the data, our team then modified the data to properly represent their respective meanings. The categorical variables in the dataset were

transformed into factors and the continuous variables were converted to numeric values. The code that was created to make these changes can be seen below in figure 2.

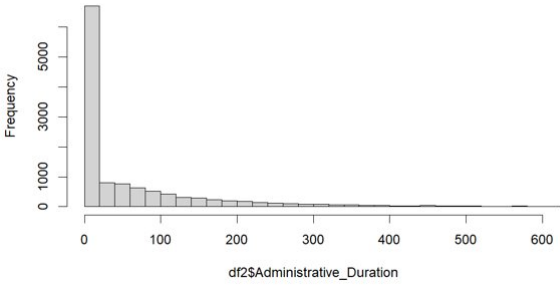
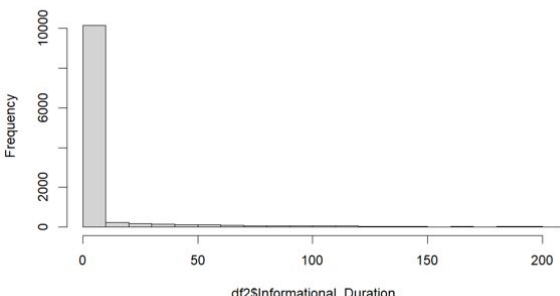
Figure 2. Data Prep

```
df$Administrative <- as.numeric(as.factor(df$Administrative))
df$Informational <- as.numeric(as.factor(df$Informational))
df$ProductRelated <- as.numeric(as.factor(df$ProductRelated))
df$Month <- as.numeric(as.factor(df$Month))
df$OperatingSystems <- as.numeric(as.factor(df$OperatingSystems))
df$Browser <- as.numeric(as.factor(df$Browser))
df$Region <- as.numeric(as.factor(df$Region))
df$TrafficType <- as.numeric(as.factor(df$TrafficType))
df$VisitorType <- as.numeric(as.factor(df$VisitorType))
df$Weekend <- as.numeric(as.factor(df$Weekend))
df$Revenue <- as.factor(df$Revenue)
```

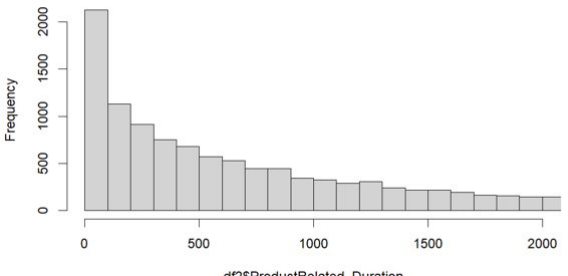
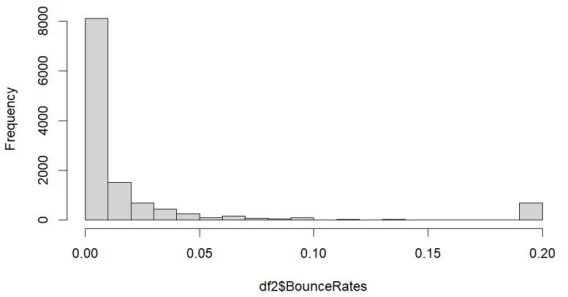
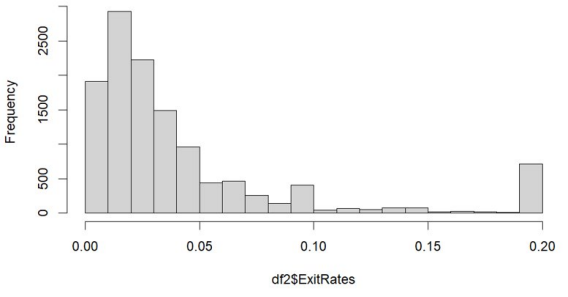
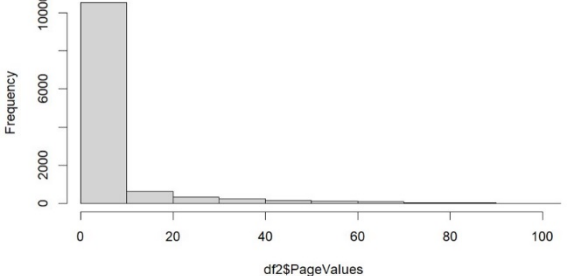
3. Exploratory Data Analysis

Starting the exploratory data analysis, our team performed some visualizations of the variable's distribution by creating histograms for the continuous variables and bar charts for the categorical or factored data. The plots of the continuous variables are shown below.

Table #1:

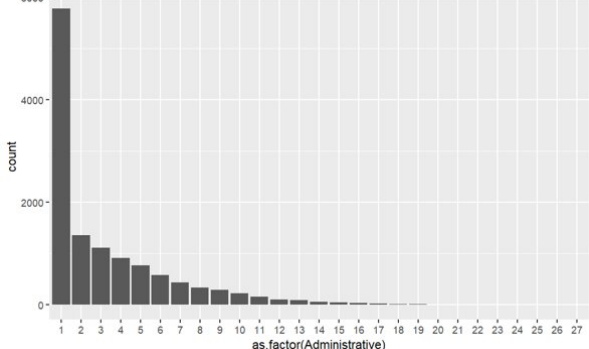
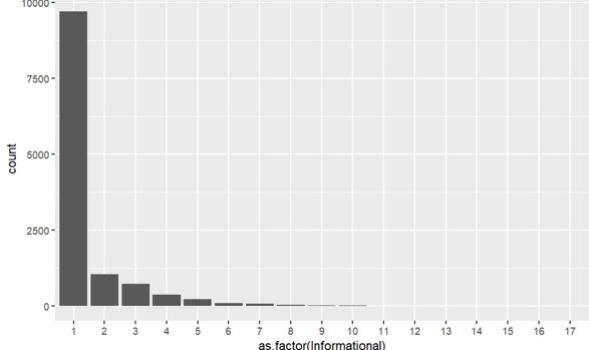
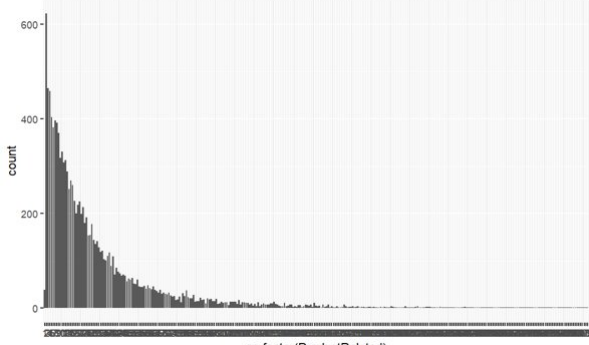
| Histogram | Inference |
|--|---|
| <p>Administrative Duration</p>  <p>The histogram for Administrative Duration shows a very high frequency (over 5000) for the first bin (0-10 seconds). The frequency drops sharply for subsequent bins, with most of the remaining data points having frequencies below 1000. The x-axis ranges from 0 to 600 seconds, and the y-axis (Frequency) ranges from 0 to 5000.</p> | <p>Based on the visualization on the left, most users are only spending 0 to 10 seconds on the administrative pages. After the first initial spike, there is a clear drop off to less than 1000 users spending more time than 10 seconds on these pages. These results are consistent with a typical e-commerce site since their main consumer base is on their webpage to shop for products and are likely uninterested in any administrative pages.</p> |
| <p>Informational Duration</p>  <p>The histogram for Informational Duration shows a very high frequency (over 10000) for the first bin (0-10 seconds). The frequency drops sharply for subsequent bins, with most of the remaining data points having frequencies below 2000. The x-axis ranges from 0 to 200 seconds, and the y-axis (Frequency) ranges from 0 to 10000.</p> | <p>Based on the visualization on the left, most users are only spending under 10 seconds on any of the information pages. After the first initial spike in the first group of 0 to 10 seconds, there is a very low number of users who stay on these pages. These results are consistent with a typical e-commerce site since their main consumer base is on their webpage to shop for products and are likely uninterested in any informational pages unless they are seeking information on the company itself.</p> |

DS 807 UNSTRUCTURED DATA

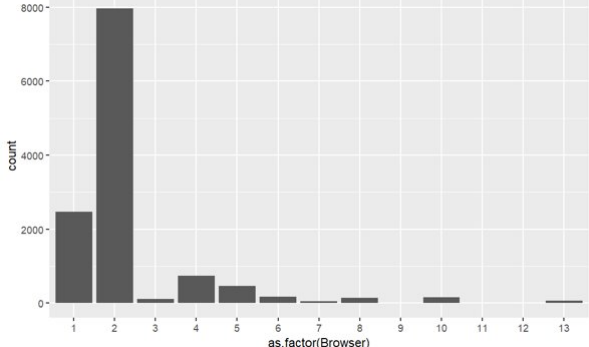
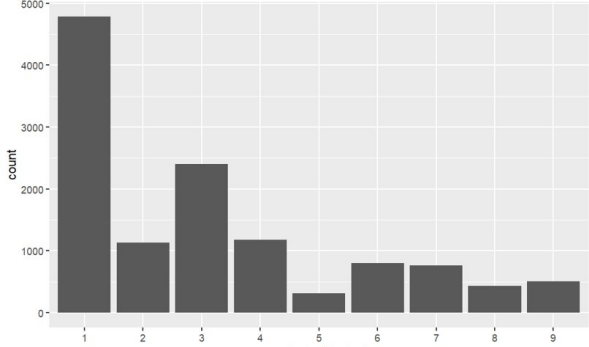
| | |
|---|---|
| <p style="text-align: center;">Product Duration</p>  | <p>Based on the visualization on the left, we can infer that the users are clearly spending most of their time browsing through the different products offered on their e-commerce site. Compared to the first two histograms, this has a very clear distribution that slowly drops off as the length of time on a product related page. This is an expected result as we assume most users who visit an online retailer would be looking at products rather than administrative or informational pages</p> |
| <p style="text-align: center;">Histogram of df2\$BounceRates</p>  | <p>Based on the visualization on the left, the breakdown of the bounce rate of its users shows positive results. The performance of a website's bounce rate is shown by how small the percentage the bounce rate is compared to competitors. From the histogram, we can determine that most users have a very low bounce rate under 5% which is very positive but there is a clear spike of users with a 20% bounce rate. This is a clear outlier that should be further investigated since this is a clear audience which we want to avoid.</p> |
| <p style="text-align: center;">Histogram of df2\$ExitRates</p>  | <p>Based on the visualization on the left, the breakdown of the exit rates for this online retail store also shows positive results. There is a clear distribution with most of the users being under a 5% exit rate. There is a clear spike of users with a 20% exit rate. This is a clear outlier that should be further investigated since this is a clear audience which we want to avoid. It would also be valuable to group the data based on the page type to determine the exit rates for specific web pages to properly determine potential bottlenecks.</p> |
| <p style="text-align: center;">Page Values</p>  | <p>Based on the visualization on the left, we can determine the breakdown of the Page Values variable. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. This variable shows the length of time in number of pages before the average user would purchase a product from the platform.</p> |

Continuing with the exploration of the variables used in this analysis, our team created visualizations using bar charts for each of the categorical variables. This allows us to determine the total counts of each of the different options within the variable. The results of the plots can be seen below.

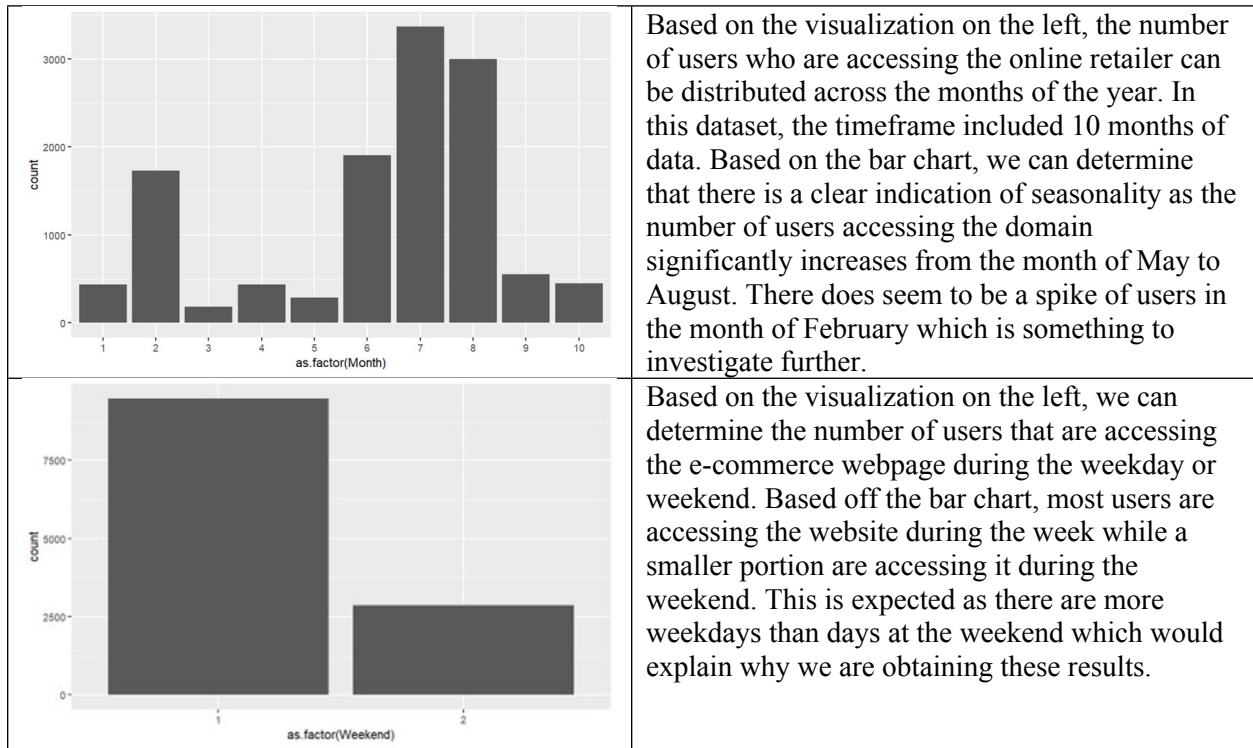
Table #2

| Bar Chart | Inference |
|---|---|
|  | <p>Based on the visualization on the left, we can determine the number of users that are accessing specific web pages that are labeled as administrative. Based on the bar chart, it is clear the most users are only visiting the first page and then leaving this category of pages since the number of viewers that continue past page 2 is just over 1,000. This number for users who view the page continues to decrease significantly up until page 19 where we see that little to no users are viewing any page past 20.</p> |
|  | <p>Based on the visualization on the left, we can determine the number of users that are accessing specific web pages that are labeled as informational. Based on the bar chart, it is clear the most users are only visiting the first page and then leaving this category of pages since the number of viewers that continue past page 2 is drastically lower than the first page. The number of views per page continues to decrease until page 10 and after page 11 there seems to be little or no views at all.</p> |
|  | <p>Based on the visualization on the left, the page of the products is shown in decreasing popularity. From the chart, there seems to be a clear distinction of products that are more favored than others. The online platform does carry a lot of products which are not being seen by most viewers. The product names were not provided and are unknown in this dataset.</p> |

DS 807 UNSTRUCTURED DATA

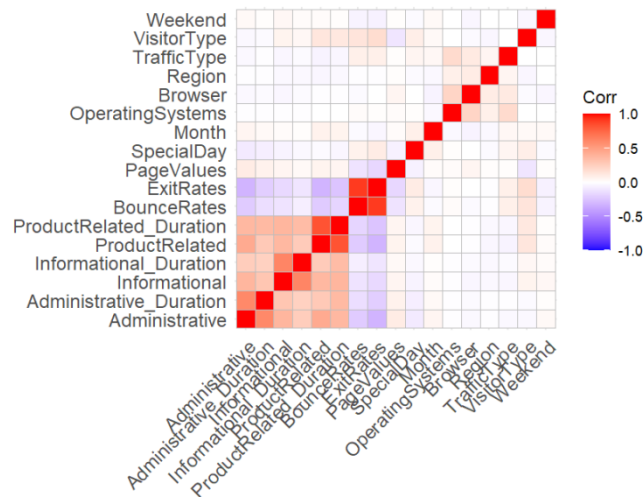
|  <table border="1"> <thead> <tr> <th>Operating System</th> <th>Count</th> </tr> </thead> <tbody> <tr><td>1</td><td>2500</td></tr> <tr><td>2</td><td>6500</td></tr> <tr><td>3</td><td>2500</td></tr> <tr><td>4</td><td>500</td></tr> <tr><td>5</td><td>0</td></tr> <tr><td>6</td><td>0</td></tr> <tr><td>7</td><td>0</td></tr> <tr><td>8</td><td>100</td></tr> </tbody> </table> | Operating System | Count | 1 | 2500 | 2 | 6500 | 3 | 2500 | 4 | 500 | 5 | 0 | 6 | 0 | 7 | 0 | 8 | 100 | <p>Based on the visualization on the left, we can determine the number of users who are using different types of operating systems. Based on the data used in this analysis, we can determine that there are three main operating systems that are accessing the online platform. There are some users who appear to be using a fourth operating system with a fair number of users and then there are very few who use any of the four other systems. These systems could possibly be IOS, Windows, or Linux but the information regarding the representation of this data was not provided.</p> | | | | | | | | | | | | | | | | | | | | | | | | |
|--|------------------|-------|---|------|---|------|---|------|---|------|---|-----|---|-----|---|-----|---|-----|---|-----|---|-----|----|-----|----|---|----|-----|---|---|----|---|----|---|----|---|----|---|----|---|----|-----|--|
| Operating System | Count | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 6500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 2500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|  <table border="1"> <thead> <tr> <th>Browser</th> <th>Count</th> </tr> </thead> <tbody> <tr><td>1</td><td>2500</td></tr> <tr><td>2</td><td>8000</td></tr> <tr><td>3</td><td>0</td></tr> <tr><td>4</td><td>800</td></tr> <tr><td>5</td><td>500</td></tr> <tr><td>6</td><td>200</td></tr> <tr><td>7</td><td>0</td></tr> <tr><td>8</td><td>200</td></tr> <tr><td>9</td><td>0</td></tr> <tr><td>10</td><td>200</td></tr> <tr><td>11</td><td>0</td></tr> <tr><td>12</td><td>0</td></tr> <tr><td>13</td><td>100</td></tr> </tbody> </table> | Browser | Count | 1 | 2500 | 2 | 8000 | 3 | 0 | 4 | 800 | 5 | 500 | 6 | 200 | 7 | 0 | 8 | 200 | 9 | 0 | 10 | 200 | 11 | 0 | 12 | 0 | 13 | 100 | <p>Based on the visualization on the left, we can determine the count of users who are utilizing the e-commerce platform. Based on the data used in this analysis, there are a large number of users who are using browsers 1 and 2 with some users using browsers 4 and 6. The other remaining browsers have very few to no users. The representation of the data was not provided so the names of the browsers are not known.</p> | | | | | | | | | | | | | | |
| Browser | Count | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 8000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 800 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | 200 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 200 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10 | 200 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 11 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13 | 100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|  <table border="1"> <thead> <tr> <th>Traffic Type</th> <th>Count</th> </tr> </thead> <tbody> <tr><td>1</td><td>2500</td></tr> <tr><td>2</td><td>3800</td></tr> <tr><td>3</td><td>2100</td></tr> <tr><td>4</td><td>1100</td></tr> <tr><td>5</td><td>300</td></tr> <tr><td>6</td><td>500</td></tr> <tr><td>7</td><td>0</td></tr> <tr><td>8</td><td>400</td></tr> <tr><td>9</td><td>0</td></tr> <tr><td>10</td><td>500</td></tr> <tr><td>11</td><td>300</td></tr> <tr><td>12</td><td>0</td></tr> <tr><td>13</td><td>800</td></tr> <tr><td>14</td><td>0</td></tr> <tr><td>15</td><td>0</td></tr> <tr><td>16</td><td>0</td></tr> <tr><td>17</td><td>0</td></tr> <tr><td>18</td><td>0</td></tr> <tr><td>19</td><td>0</td></tr> <tr><td>20</td><td>200</td></tr> </tbody> </table> | Traffic Type | Count | 1 | 2500 | 2 | 3800 | 3 | 2100 | 4 | 1100 | 5 | 300 | 6 | 500 | 7 | 0 | 8 | 400 | 9 | 0 | 10 | 500 | 11 | 300 | 12 | 0 | 13 | 800 | 14 | 0 | 15 | 0 | 16 | 0 | 17 | 0 | 18 | 0 | 19 | 0 | 20 | 200 | <p>Based on the visualization on the left, we can understand the breakdown of the users based on their traffic type. Most users appear to be coming from 4 to 5 different sources. Traffic source #2 appears to be the most popular followed by 1,3,4, and 13. These traffic sources may be indirect or direct sources such as web searches, social media, or a marketing campaign. The traffic type representation was not provided and is unknown in this dataset.</p> |
| Traffic Type | Count | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 3800 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 2100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 1100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 300 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | 500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 400 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10 | 500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 11 | 300 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13 | 800 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 14 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 15 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 16 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 17 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 18 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 19 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 20 | 200 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|  <table border="1"> <thead> <tr> <th>Region</th> <th>Count</th> </tr> </thead> <tbody> <tr><td>1</td><td>4800</td></tr> <tr><td>2</td><td>1100</td></tr> <tr><td>3</td><td>2400</td></tr> <tr><td>4</td><td>1100</td></tr> <tr><td>5</td><td>300</td></tr> <tr><td>6</td><td>800</td></tr> <tr><td>7</td><td>700</td></tr> <tr><td>8</td><td>400</td></tr> <tr><td>9</td><td>500</td></tr> </tbody> </table> | Region | Count | 1 | 4800 | 2 | 1100 | 3 | 2400 | 4 | 1100 | 5 | 300 | 6 | 800 | 7 | 700 | 8 | 400 | 9 | 500 | <p>Based on the visualization on the left, the numbers of users who are accessing the e-commerce platform based on their different regions. Based on the bar chart, the platform is most popular in region 1 followed by region 3 who has roughly half the number of users. The online platform seems to be used decently across all regions with region 5 having the least number of users. The area of the regions was not provided and is unknown in our analysis.</p> | | | | | | | | | | | | | | | | | | | | | | |
| Region | Count | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 4800 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 1100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 2400 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 1100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 300 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | 800 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | 700 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 400 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | 500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

DS 807 UNSTRUCTURED DATA



Continuing to understand the data, a correlation matrix for the entire dataset was created to determine which variables were correlated with each other. It allows our team to determine any relationships while also helping in identifying any patterns or dependencies within the data. In the figure below, we can identify strong positive correlation based on the coefficient's closeness to 1. The opposite can be said for negatively correlated variables based on their closeness to -1 . The positive correlations are represented in the color red, and the negative correlated variables are shown in purple. The matrix can also play a key role in identifying potential multicollinearity issues as well as aid in the feature selection process. Overall, the correlation matrix provides valuable insights that can help contribute to the overall understanding of the underlying data. The results of the matrix can be seen below.

Figure 3. Correlation Matrix



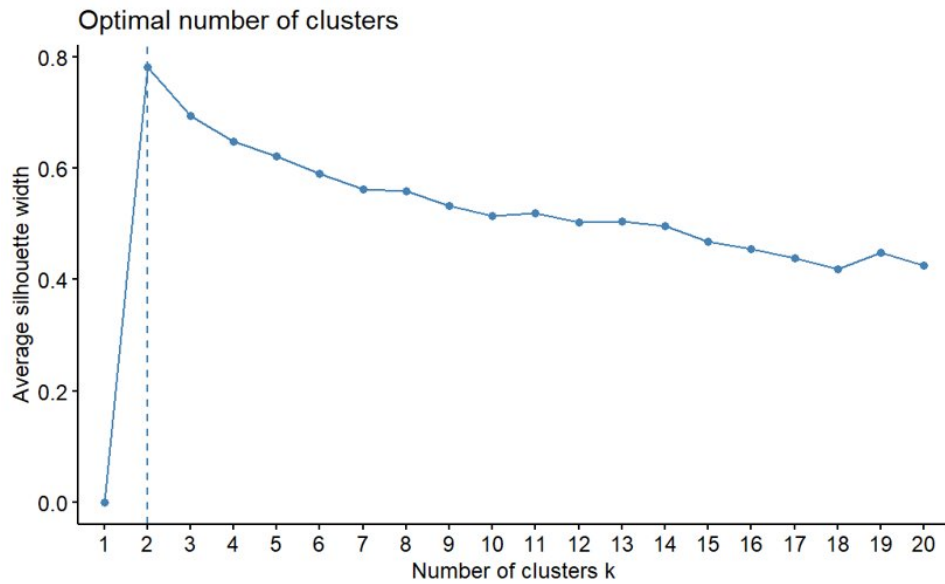
4. Clustering Models

K-Means

The K-means clustering mode is a very popular and used unsupervised machine learning algorithm which aims to partition the dataset into one of the predefined clusters. The data points are then classified into separate clusters based on the cluster with the closest mean value. The assignment is based on the Euclidean distance metric which calculates the distance between each point and each cluster. After initially assigning the data points, the model recalculates the cluster based on the new mean of all the data points. This process continues until the optimal cluster location no longer significantly changes and the minimization of the within-cluster sum of squares converges.

The creation of the 'K'-means model involved determining the pre-defined number of clusters that would be provided as a parameter. To come up with a 'K' value, our team utilized the silhouette method to come up with a value that can be argued is optimal statistically. The silhouette method is a popular approach commonly used in clustering models and it can evaluate the quality of the clusters based on the average silhouette coefficient. This allows the model to measure and compare the results of a specific 'K' value compared to the other given 'K' values. Below in figure 3, the visualization of the silhouette model can be seen along with the optimal 'K' value, which was equal to '2' for this model.

Figure 4. Silhouette Method



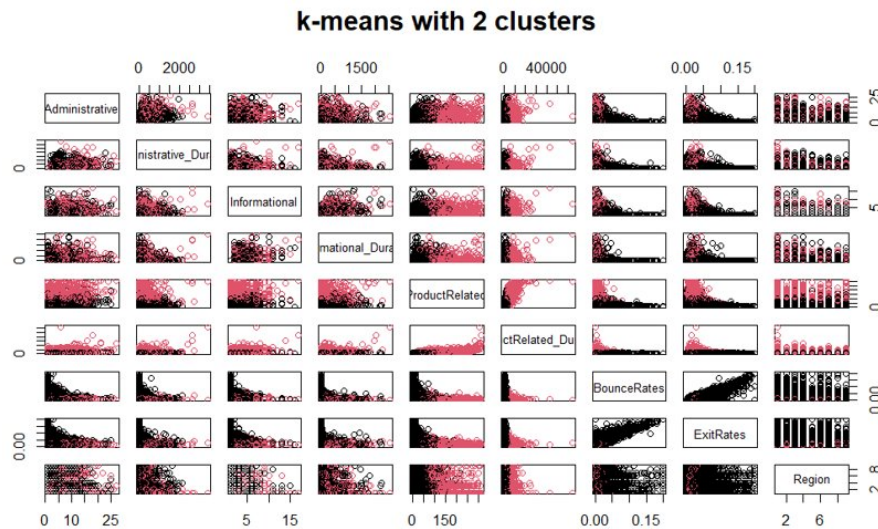
Based on the results of the silhouette method, our team was then able to proceed with the creation and understanding of the K-means model for our given dataset. The code and output of the model can be seen below in figure 4.

Code: `km.out = kmeans(df2, 2, nstart = 20)`

Results:

| Cluster 1 | Cluster 2 |
|-----------|-----------|
| 11,397 | 933 |

Figure 5. K-Means Model



Based on the results provided from the variable paired clusters, we can visually analyze the difference between both clusters to identify their key differences. The K-means model was predetermined to create two clusters and appears to have struggled with determining the data points within cluster two. The clusters are heavily imbalanced with over 11,000 user sessions being grouped within cluster 1 while only having 933 observations within the second cluster. Based off the plots, it appears that cluster 1 contains users who are only browsing on the online platform as they make up most users who exit or ‘bounce’ from the website while also not spending much time going through the various products. The first cluster also makes up a large majority of users who are viewing pages such as the Administrative and informational type pages much more than the product related pages. Cluster two seems to be comprised of users who purchase products from the platform as they can be seen viewing more products for a larger amount of time while also not exiting the website as often as the first cluster. Based off our team’s results, we would recommend focusing on cluster two as they seem to play a more important role when it comes to purchasing products while cluster one appears to me more interested in the company or platform more than the actual products.

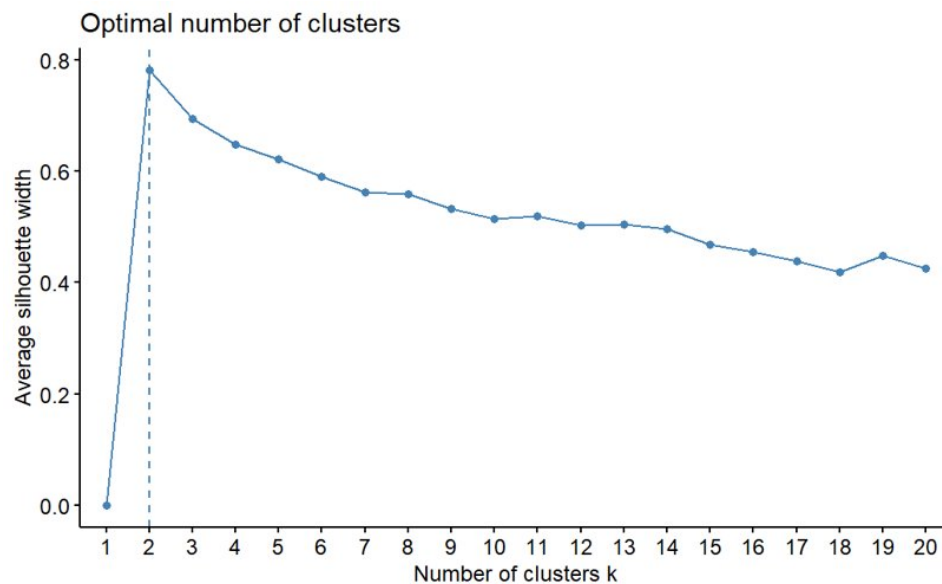
K-Medoids

The K-medoids model, also known as the Partition Around Medoids (PAM) model, is a very similar clustering algorithm to the K-means but assigns the data points based on their distance to the central most point in a specific cluster which is referred to as a medoid. This model is designed to work well when the mean does not represent the underlying data such as

with categorical or binary values. The K-medoids model also relies on the Euclidean distance metric to assign the data points and like the K-means model by recalculating the medoid at each iteration until the total dissimilarity is minimized. Once the model converges, the final clusters are obtained, and each data point would be assigned to its respective cluster.

The creation of the K-medoids model does rely on the pre-selection of the number of clusters for the model to create. This problem was solved by again using the silhouette method to compare different numbers of clusters to determine the most optimal K value. The output of the silhouette method can be seen below in figure 5 with the optimal value being equal to two.

Figure 6. Silhouette Method



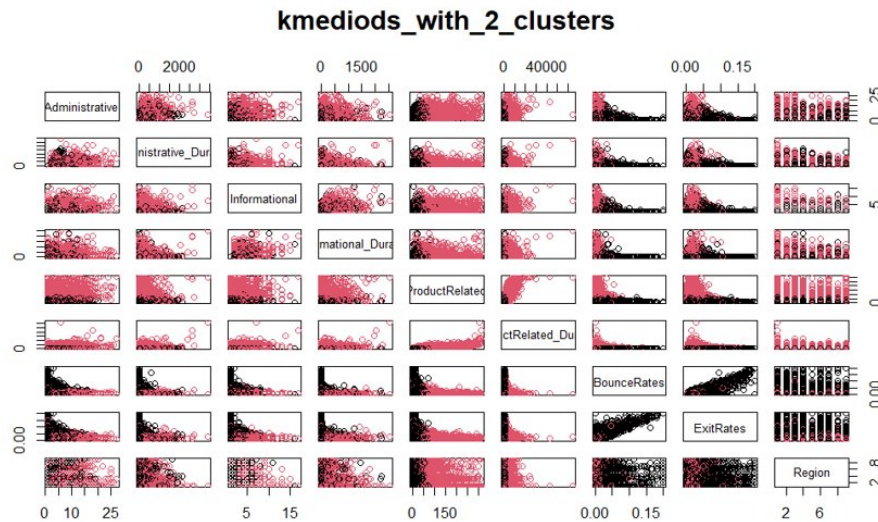
Based on the results from the chart above, we were able to determine our optimal number of clusters to create the K-medoids model. The creation and the resulting output of the model can be seen below in figure 6.

Code: `km = pam(df2,2)`

Results:

| Cluster 1 | Cluster 2 |
|-----------|-----------|
| 9,036 | 3,291 |

Figure 7. K-Medoids Model

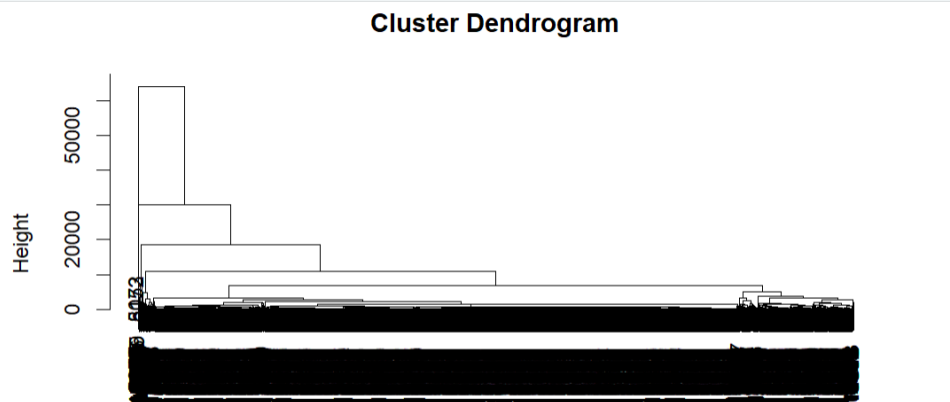


Based on the results of this model, the K-medoids model was able to classify two separate clusters that capture the behaviors of two groups of users. This model was able to create a more balanced set of clusters having 9,036 users in cluster 1 and 3,291 users in cluster 2. This model was able to better split the data into even groups and it allows us to visually determine the differences between the two groups. Like the K-means algorithm, the K-medoid's model was able to split the data into two groups that can be best described as users who are likely to purchase and those who are not. We can determine that the red cluster representing cluster 2, spends much more time on the site and browsing through more pages while the black cluster tends to stay for the first few pages and exit the webpage much more. Through this analysis, we can claim that the cluster two should be more focused on as they appear to be their main consumer base who are more likely to purchase a product while using the online platform.

Hierarchical Model

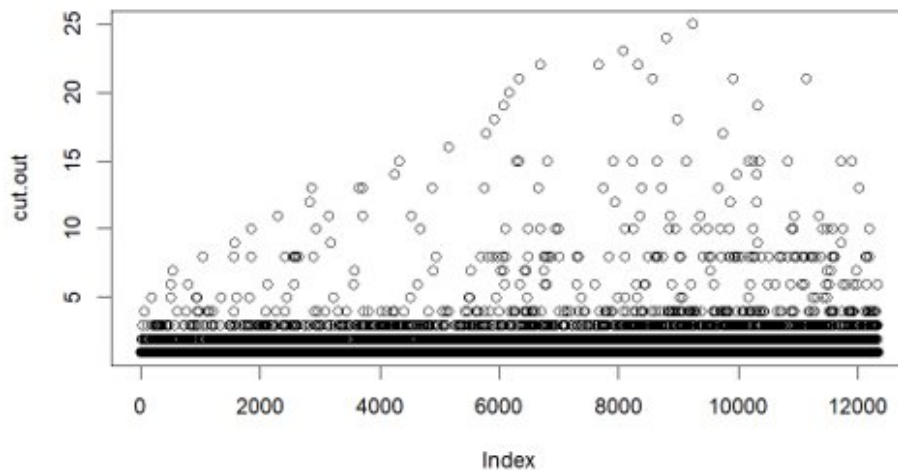
The Hierarchical clustering model is a different approach to clustering data points since it does not rely on a predefined number of clusters in its parameters. This model creates a tree-like structure called a dendrogram which visualizes the relationship between the clusters and the decision-making process. The linkage criterion that was used when creating the hierarchical model was selected to be the average linkage. This determines the similarity between two clusters as the average similarity between all data points from each cluster. Below we can see the visualization of the model's dendrogram.

Figure 8. Hierarchical Dendrogram



The Hierarchical cluster dendrogram of the project data with 18 variables looks unclear but the use of a greater number of variables doesn't bring out a clear dendrogram. To determine the optimal number of clusters for this model, our team utilized the silhouette method to determine the most optimal cut off point in the dendrogram. Using this method, the results came out to be K is equal to 2 but when performing this on the model the results showed all but one data point going into cluster 1 while a single data point was classified into cluster 2. This not being ideal, our team determined the cutoff point manually stopping at the height of 1,500. This created 25 different clusters with varying number of users in each cluster. Overall, this model did not perform the best and did not add any real insights into our analysis compared to the previous models.

Figure 9. Hierarchical Clusters



5. Mixture Model

Mclust

The Mclust model is a commonly used clustering algorithm used in classification analysis. The model works effectively when dealing with complex datasets that may have elements that may not be possible to capture with simpler models such as the K-means. Mclust provides a probabilistic approach when designing the clusters and allows for a more accurate and flexible assignment of the data points. The model employs different models to determine the optimal number of clusters within the data as well as the optimal covariance as a parameter. It utilizes the Expectation-Maximization Algorithm (EM) to optimize the parameters Gaussian distribution of the mixture model. The Mclust model allows for different covariances within each individual cluster which allows the model to be more flexible to different possibilities to reflect the data more accurately. The model also utilizes the Bayesian Information Criterion (BIC) to determine the performance of the various models which then allows it to select the model with the best fit. Once the model is selected, the data points are then assigned to a cluster which has the highest probability of fitting with the other assigned points.

In the creation of the Mclust model, the model was given the option of determining the optimal number of clusters given a range from 2 to 20. Based on the model's selection, it determined that the optimal number of clusters was equal to 4 which had a BIC of $-763,428$. The function includes the parameter called parameters which is equal to 'True'. This allows the model to determine its own estimates for the mixture model creating more useful and interpretable results. The creation and results of the Mclust model is shown below.

Figure 9. BIC Plot

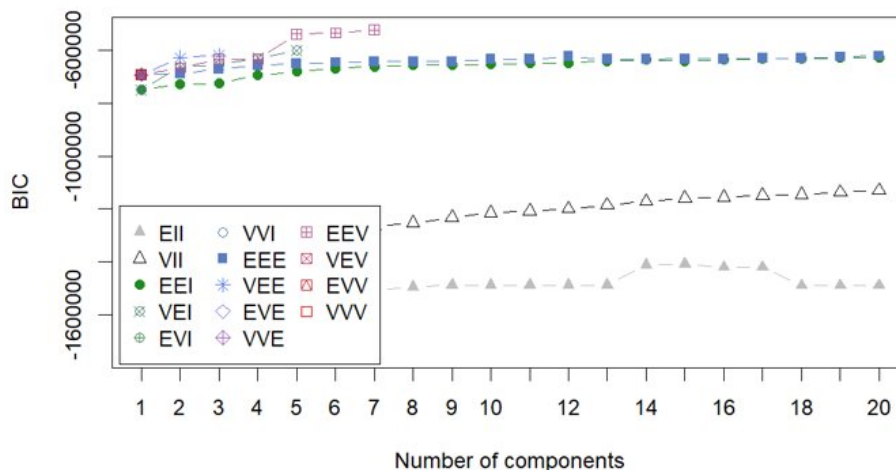
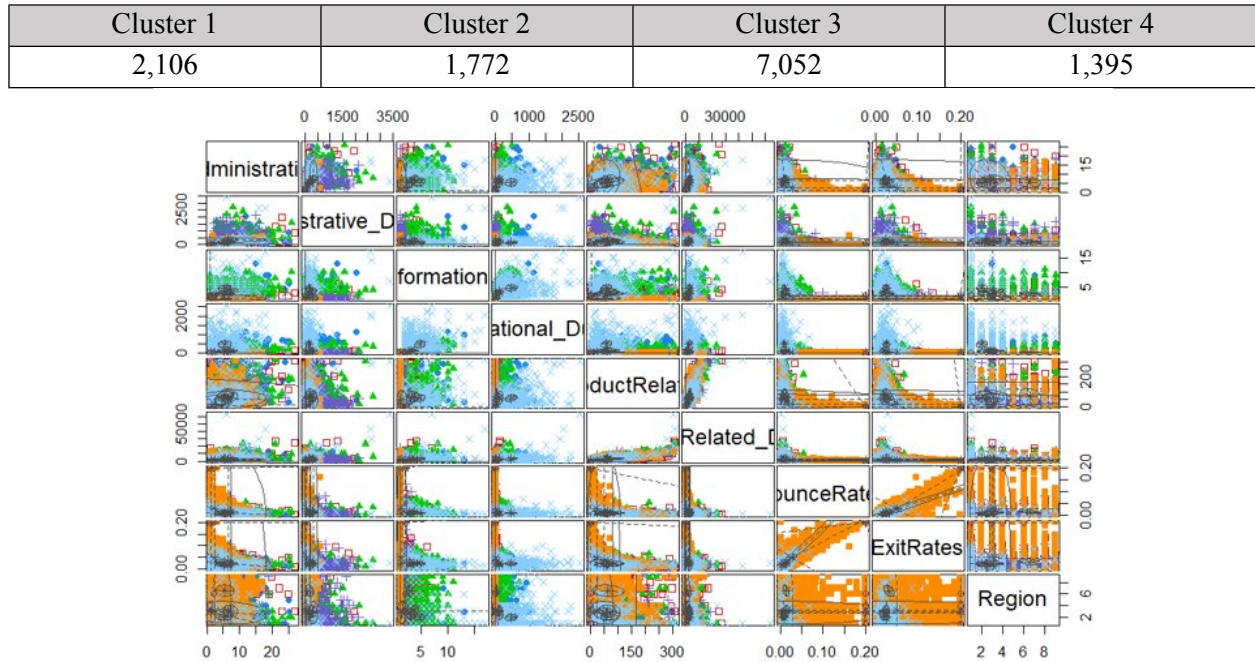


Figure 10. Mclust Model



Based on the visualization of the clusters based on the variable pairings, we can visually determine the four separate clusters created by the model. The results group the users into four even clusters which capture the different behaviors and tendencies of their customer base. The largest cluster represented as orange, appears to be the group of users who are simply browsing the e-commerce platform and not purchasing any products. This group takes most of the users who high a high exit and bounce rate while often not spending large amounts of time on the product related pages. The blue cluster appears to spend most of their time on non-product pages such as Administrative or Informational since they view the most pages for the longest amount of time. This group does still look at products and overall spends the longest time on the platform. The green cluster we believe represents the main customers who are spending time on the website while also consisting of the users who spend the most time viewing products. Our recommendations based off this model would be to focus ideally on the green and blue clusters to generate ads or marketing campaigns while attempting to avoid users who are like the orange cluster.

6. Neural Network Model

Feed Forward Neural Network (FFNN):

FFNN stands for Feedforward Neural Network. It is a sort of artificial neural network in which information flows unidirectionally from the input to the output layer. The network structure of this architecture has no loops or cycles, and each layer is fully connected to the next layer.

FFNNs are widely utilized in a variety of applications, including image recognition, natural language processing, and time series prediction.

In this Project, we have applied FFNN to our data for determining the predictor variable 'Revenue'. So, we considered the variable 'Revenue' as a binary variable. In the first step, we converted the predictor variable into numeric. We created two values x & y and data splitting process done with 70% train data and 30% test data. After then, conversion of x and y variables into matrix form. The data is then divided into x test and y test forms also.

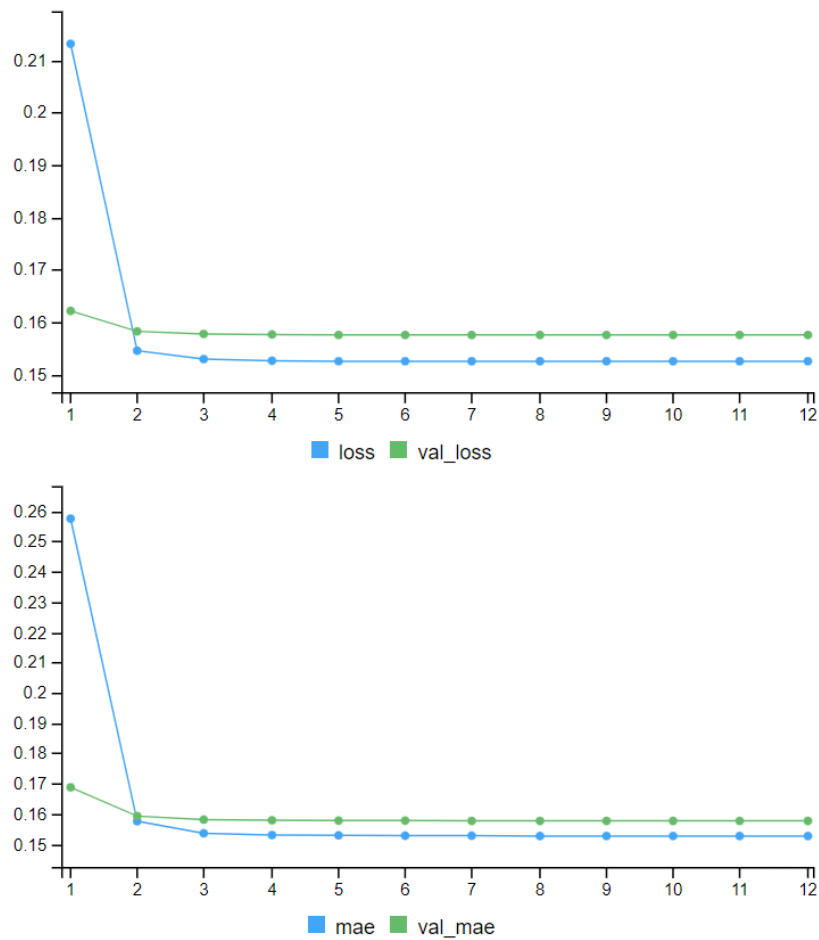
The next step is to build a keras sequential model with the help of keras library. The model comprises of 3 dense layers and a drop out parameter. The dropout parameter helps in reducing the overfitting of data and two dense layers with 'relu' activation & a dense layer using sigmoid activation is added. From running the Summary(model) function we see the total parameters distributed among the dense layers as output.

Figure 11. FFNN Layers

| Model: "sequential_8" | | |
|--------------------------|--------------|---------|
| Layer (type) | Output Shape | Param # |
| dense_26 (Dense) | (None, 256) | 4608 |
| dropout_8 (Dropout) | (None, 256) | 0 |
| dense_25 (Dense) | (None, 128) | 32896 |
| dense_24 (Dense) | (None, 1) | 129 |
| Total params: 37,633 | | |
| Trainable params: 37,633 | | |
| Non-trainable params: 0 | | |

In the next step of R code, we compile the model by using loss, optimizer and metrics. We substitute loss function as 'MSE', optimizer as 'RMSPROP' and Metrics as 'MAE' respectively. And further we create a fitted model with history by giving the parameters such as batch_size, epochs, validation_split, x_train and y_train etc., The final plot is obtained which compares the values of loss Vs valence loss and mae vs valence mse at given number of epochs.

Figure 12. FFNN Plotted History



In this analysis, we understood the steps required in order to create a neural net model that is used for prediction of numeric data. We then created a model which was then tested and used to predict the median house value variable. We learned how to create different layers within the model. The model created the neural net model which had a Loss value of 0.1551934, MAE value of 0.1552021 and accuracy value of 0.8448229 (~84.4%).

7. Findings

In this analysis, utilizing different classification algorithms our team has been able to create valuable insights on the consumer base of the e-commerce platform. Based on the results of all

the models, we have determined that the optimal model was the Mclust model due to its ability to determine more than two distinct customer groups which we believe would better represent the true grouping of its users. The K-medoids model was also found to be interesting since it allowed for a separation between two groups of customers that we can claim consists of those who are likely to purchase a product and those who are not. This provides a much simpler and broader approach that can still be used to specifically target or avoid a group of users. Both the K-means and Hierarchical clustering models still managed to provide some valuable insights but not to the ability of the Mclust model.

8. Conclusion

The rapid growth of technology and the internet has changed the way the public shops for their everyday items. To remain competitive in the e-commerce landscape, a business must be able to understand the intention and shopping behavior of their customers. The goal of this analysis was to explore the different types of online shoppers by analyzing various factors which include browsing patterns and duration of page visits. By utilizing the abilities of machine learning algorithms, our team has been able to classify the users of the online retail store in multiple different fashions which allowed us to narrow down the model in which we believe best represents the underlying data. The creation of an accurate classification model based on a large dataset can help aid in the e-commerce business decision making process allowing them to efficiently optimize their strategies and provide a more personalized experience for all users.

In summary, our team was able to determine a classification model that could be utilized in a real e-commerce platform that is still in need of understanding their consumer base. The models included variables on the pages viewed, such as `Administrative` and `Informational`, variables on the time spent on the different page types, such as `Administrative_Duration` and `Informational_Duration`, and finally variables on their users such as `Region` and `Exit_Rate`. Clustering algorithms, K-means and K-medoids, have been shown to be a useful tool in splitting the consumers into two separate clusters which can be identified as those who have purchased a product and those who have not purchased. The Mclust algorithm was a valuable tool that utilized the advantages of a mixture model and after concluding our analysis it has proven to be one of the best models in determining the various types of consumers. Our team has been able to provide valuable insights that can be further analyzed to best inform decision makers.

9. Future Scope

1. We would recommend the use of this model on small cap ecommerce platforms to increase sales.
2. Dealing with a few more variables could be useful in generating the optimal revenue.
3. Helpful in suggesting these clustering models to a product-based company to make profits.
4. Diving deeper into details of individual clusters and the users that are within them.
5. Focusing on website bottlenecks to optimize website organization minimizing exit rates.
6. Introduce a causal model approach to determine patterns within clusters.
7. Longer timeline with more detailed data regarding dates to identify seasonal or yearly trends.