Dataset Description and PCA Analysis

Your Name

2024-11-18

Contents

1. Dataset Description

1.1 Overview

The **Online Shoppers Purchasing Intention Dataset** is sourced from the **UCI Machine Learning Repository** and provides valuable insights into online shopping behavior. This dataset tracks user interactions on an e-commerce website, with the primary goal of predicting whether a shopper will make a purchase during their visit.

It includes features such as the number of pages visited, time spent on various types of pages, bounce rates, and exit rates. Additionally, the dataset captures demographic information about the shoppers, such as whether they are new or returning visitors. The target variable indicates whether the shopper made a purchase, making the dataset a valuable resource for understanding and modeling purchasing behavior.

The original dataset contains **12,330 records** and **18 features**, including both numerical and categorical variables. For our analysis, we've taken a subset of the original dataset which contains **2,000 records** and **18 features**.

1.2 Features

Below is a table representation with all the key features:

Feature Name	**Type**	**Description**
Administrative	Quantitative	Number of administrative pages visited
AdministrativeDuration	Quantitative	Time spent on administrative pages in seconds
Informational	Quantitative	Number of informational pages visited
InformationalDuration	Quantitative	Time spent on informational pages in seconds
ProductRelated	Quantitative	Number of product-related pages visited
ProductRelatedDuration	Quantitative	Time spent on product-related pages in seconds
BounceRates	Quantitative	Percentage of visitors leaving after one page
ExitRates	Quantitative	Percentage of sessions exiting from each page
PageValues	Quantitative	Average value attributed to a page
SpecialDay	Quantitative	Metric indicating proximity to significant holidays
Weekend	Binary	Indicates if the session occurred on a weekend (1 = Yes, 0 = No)
Revenue	Binary	Indicates if the session resulted in a purchase (1 = Yes, 0 = No)
VisitorType	Categorical	Visitor category (e.g., Returning, New, Other)
Month	Categorical	Month of the visit (e.g., Jan, Feb)
OperatingSystems	Categorical	Visitor operating system (e.g., Windows, MacOS)
Browser	Categorical	Browser used by the visitor (e.g., Safari, Chrome)
Region	Categorical	Visitor geographical region
TrafficType	Categorical	Type of traffic source leading to the visit

To obtain this structure a preprocessing step was performed, which included the following transformations: - The categorical variables were converted to factors. - The target variable (0 for no purchase, 1 for purchase) and the binary variables were converted to a binary format - Name changes were made to some variables for better readability and naming consistency. - A random subset of 2,000 observations was selected from the original dataset.

The code to preprocess the dataset is as follows:

set.seed(100) data <- read_csv(paste('../../data/online+shoppers+purchasing+intention+dataset/', 'online_shoppers_intention.csv', sep = ''))

data <- data %>% mutate(Month = as.factor(Month), Region = as.factor(Region), TrafficType = as.factor(TrafficType), VisitorType = as.factor(VisitorType), OperatingSystems = as.factor(OperatingSystems), Browser = as.factor(Browser), Weekend = as.factor(as.numeric(Weekend == 'TRUE')), Revenue = as.factor(as.numeric(Revenue == 'TRUE')), SpecialDay = as.factor(SpecialDay)) %>% rename(AdministrativeDuration = Administrative_Duration, InformationalDuration = Informational_Duration, ProductRelatedDuration = ProductRelated_Duration) %>% sample_n(2000)

1.3 Data Integrity

We begin by assessing the integrity of the dataset by analyzing the presence of missing data across its features. The table below presents the count of missing values for each feature. As shown, there are **no missing values** in any of the features, which indicates that the dataset is complete and does not require any imputation for missing data.

Additionally, we check for duplicate records in the dataset. The analysis revealed **6 duplicate rows**, which may be due to several factors. These duplicates could represent unintentional repetitions of the same individual's data, or they could arise from issues during data collection, such as individuals' data being recorded more than once under slightly different conditions. As the amount is relatively small, we will proceed with the analysis without removing these duplicates.

Given the absence of missing values and the limited number of duplicate rows, the dataset does not require any further cleaning or preprocessing steps before proceeding with our initial analysis.

Table 2: Missing Values Count for Each Feature

	x
Administrative	0
AdministrativeDuration	0
Informational	0
InformationalDuration	0
ProductRelated	0
ProductRelatedDuration	0
BounceRates	0
ExitRates	0
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0

```
#> Number of duplicate rows: 6
```

2. Univariate Analysis

2.1 Numeric Variables

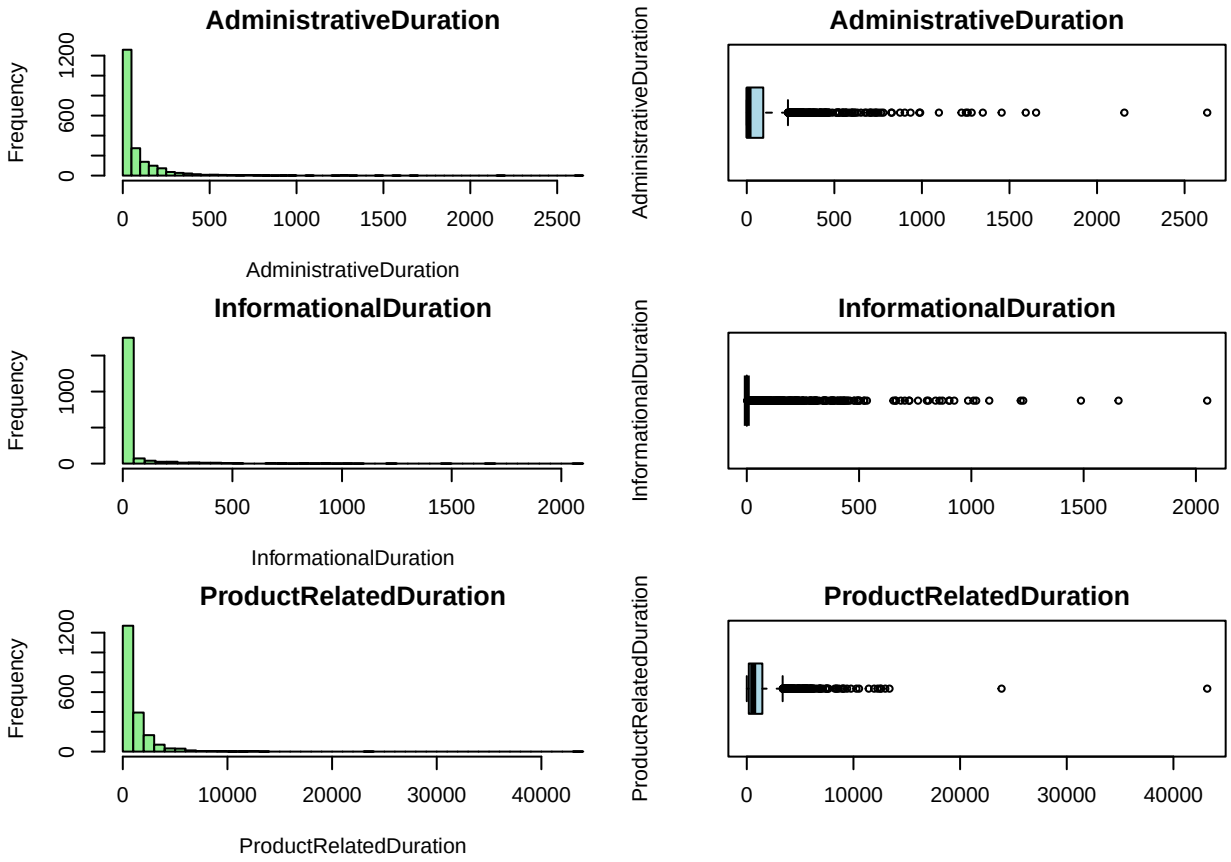
2.1.1 Time Related Variables

The duration variables (e.g., `AdministrativeDuration`, `InformationalDuration`, `ProductRelatedDuration`) exhibit values for the mean significantly higher than the median, indicating a heavily right-skewed distribution. This suggests that there are a few sessions with very high durations that are pulling the mean upwards. The range of values for these variables is quite large, with the maximum values being several times higher than the 75th percentile.

Table 3: Summary of Numeric Time Related Variables

AdministrativeDuration	InformationalDuration	ProductRelatedDuration
Min. : 0.00	Min. : 0.00	Min. : 0.0
1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 187.7
Median : 10.05	Median : 0.00	Median : 602.0
Mean : 79.24	Mean : 36.28	Mean : 1161.3
3rd Qu.: 94.00	3rd Qu.: 0.00	3rd Qu.: 1459.5
Max. :2629.25	Max. :2050.43	Max. :43171.2

The boxplots and histograms provide clear evidence of the right-skewed nature of these variables, characterized by a high concentration of values near zero and a long tail of extreme values. This distribution highlights the presence of numerous outliers, which are not isolated anomalies but rather a significant portion of the data. These extreme values likely represent important behavioral patterns or user interactions that could offer valuable insights. Therefore, instead of removing or treating them as noise, we choose to retain these outliers in our analysis to ensure a comprehensive understanding of the dataset and its implications.



2.1.2 Page Interaction Variables

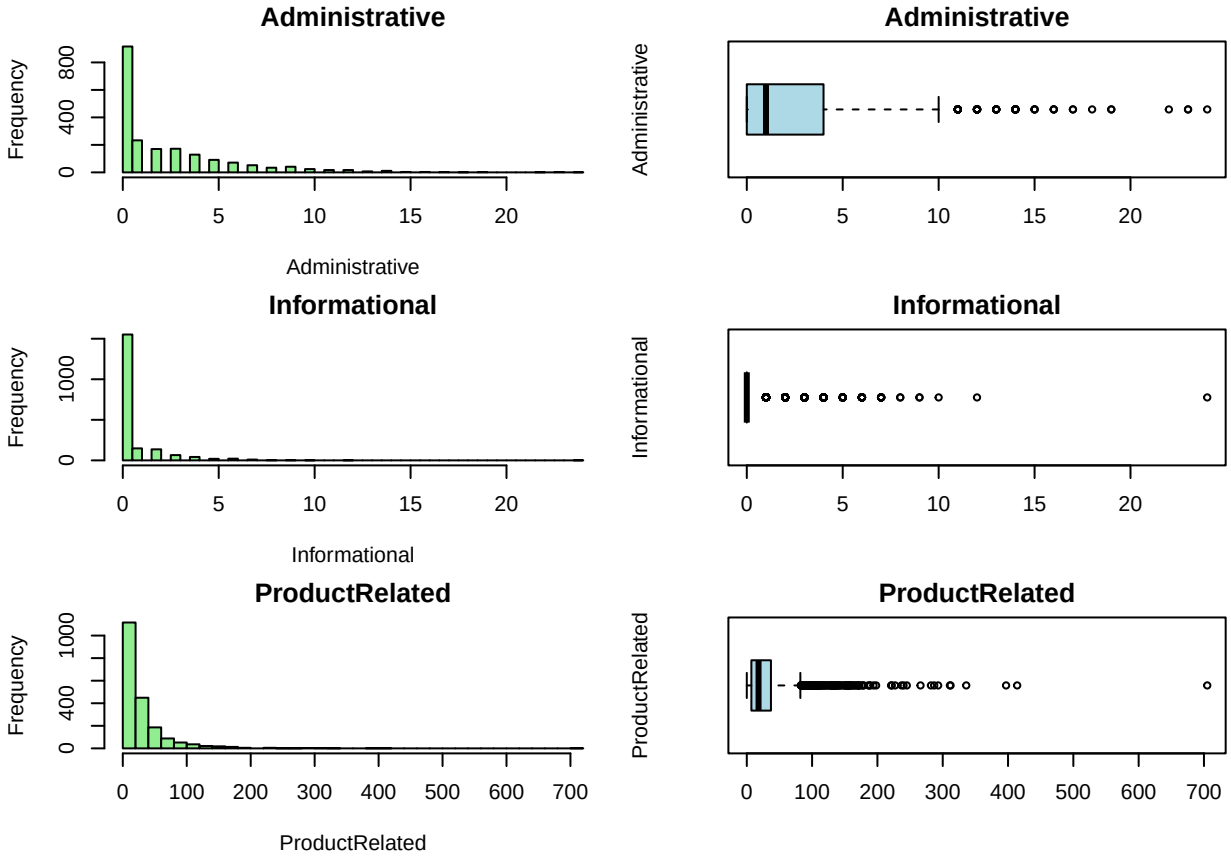
The variables related to page interactions (e.g., `Administrative`, `Informational`, `ProductRelated`) represent the number of pages visited by the user during the session. From all 3 variables, `ProductRelated` has the highest mean and median values while `Informational` and `Administrative` have lower but similar values. The summary statistics also suggest a right-skewed distribution for these variables, with the mean exceeding the median.

Table 4: Summary of Page Interaction Variables

Administrative	Informational	ProductRelated
Min. : 0.000	Min. : 0.000	Min. : 0.00
1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.: 7.00
Median : 1.000	Median : 0.000	Median : 18.00
Mean : 2.321	Mean : 0.577	Mean : 30.53
3rd Qu.: 4.000	3rd Qu.: 0.000	3rd Qu.: 37.00
Max. :24.000	Max. :24.000	Max. :705.00

The boxplots and histograms for these variables confirm the right-skewed distribution, with a large number of sessions having low page interaction counts and a few sessions with very high counts. While all three variables exhibit similar patterns, the `Informational` variable has a more pronounced skewness, followed by `ProductRelated` and `Administrative`. The presence of outliers in these variables is expected, as user behavior can vary significantly, with some users exploring multiple pages while others may exit quickly after

viewing a few pages. We believe this outlier behavior is essential for understanding user engagement and purchase intent and should be retained in the analysis.



2.1.3 Other Numeric Variables

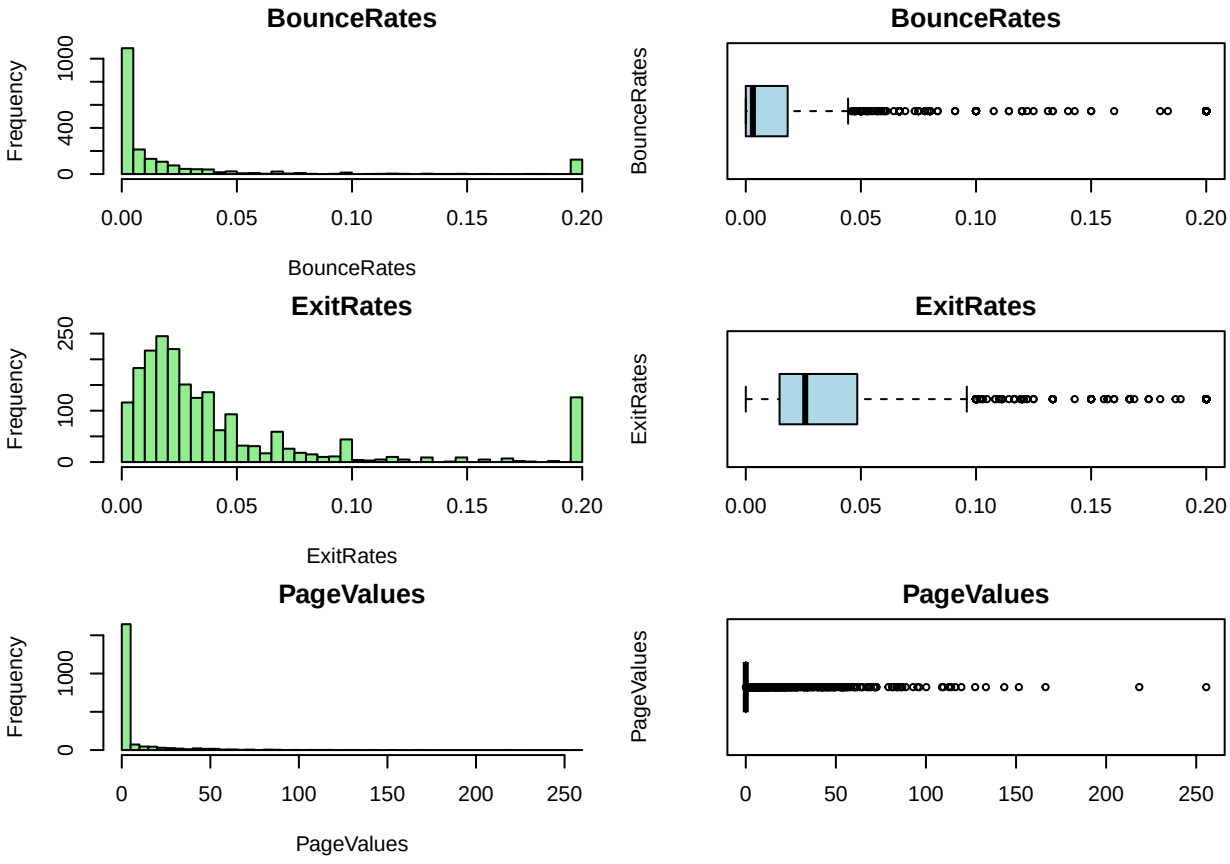
The following variables are derived from Google Analytics, providing insights into user behavior on the website:

- **BounceRates**: Represents the percentage of visitors who leave the site after viewing only one page. Summary statistics show values ranging from 0 to 0.2, with a mean of 0.023 and a median of 0.003. The distribution is right-skewed, as the mean is higher than the median, indicating a concentration of low values and a few higher values.
- **ExitRates**: Represents the percentage of visitors who exit the site from a specific page. While the dataset doesn't specify which page this refers to, it may be a critical page or one leading to a purchase. The values range from 0 to 0.2, with a mean of 0.043 and a median of 0.025. Like **BounceRates**, this variable is also right-skewed although not as extreme, with a concentration of low values and a few higher values.
- **PageValues**: Represents the average value of a page that a user visits before completing a transaction. The values range from 0 to 255, with a mean of 5.89 and a median of 0.0. This variable is heavily right-skewed, indicating that while most pages have little to no assigned value. There are a few pages with high values that significantly impact the mean.

Table 5: Summary of Page Interaction Variables

BounceRates	ExitRates	PageValues
Min. :0.000000	Min. :0.00000	Min. : 0.000
1st Qu.:0.000000	1st Qu.:0.01467	1st Qu.: 0.000
Median :0.003089	Median :0.02581	Median : 0.000
Mean :0.023104	Mean :0.04380	Mean : 5.899
3rd Qu.:0.018182	3rd Qu.:0.04839	3rd Qu.: 0.000
Max. :0.200000	Max. :0.20000	Max. :255.569

The visualizations for these variables further confirm the right-skewed distribution, with a concentration of low values and a few high values. The histograms and boxplots provide a clear representation of the distribution of these variables, highlighting the presence of outliers and the need to consider these extreme values in the analysis. We've observed outliers in all variables which may suggest specific user behaviors or interactions that could be crucial for understanding purchasing intent and user engagement on the website. As done with other variables, we choose to retain these outliers in our analysis to ensure a comprehensive understanding of the dataset.



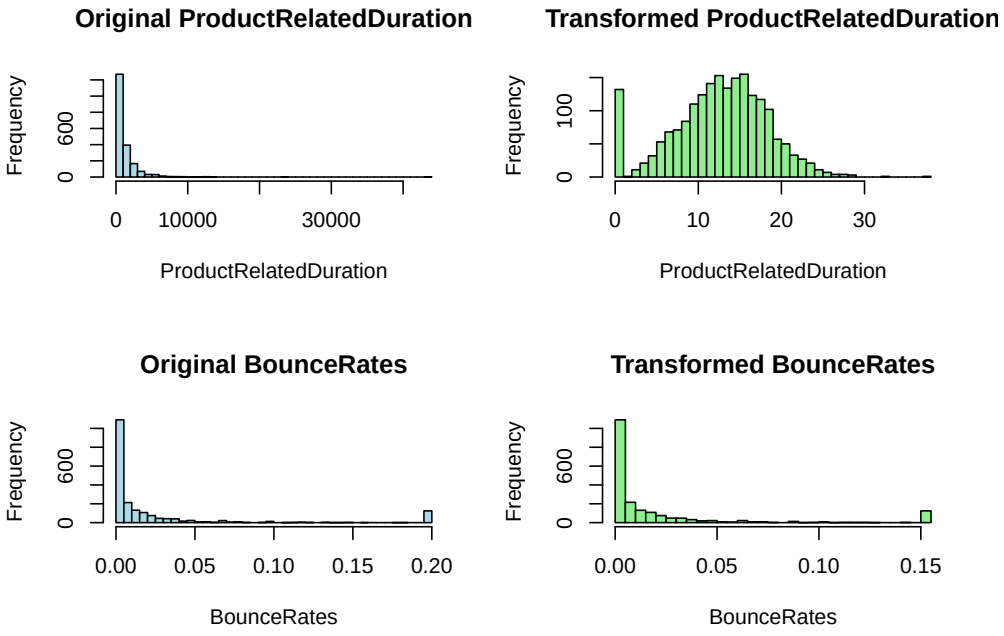
2.1.4 Transformations

Given the right-skewed nature of the numeric variables, we decided to try and apply transformation to achive a more normal distribution by reducing the positive skewness. Due to the extreme values and presence of zeros in the data, we opted for the **Box-Cox transformation** along with shifting the data by adding a constant

value to avoid issues with zero values. This transformation did not result in a significant improvement in normality, as the original data was already heavily skewed.

The 2 histograms below show a comparison of the original and transformed data for the `ProductRelatedDuration` and `BounceRates` variables. The histograms illustrate that 2 scenarios we observed in all our numerical variables. In one case, the presence of a large number of zero values results in a spike at zero in the transformed data. In the other case, the transformation does not significantly alter the distribution of the data, as the original data is extremely skewed.

It is because of these reasons that we decided to retain the original data for our analysis, as the transformation did not provide a substantial improvement in normality and also reduced the interpretability of the data. The presence of outliers and extreme values in the original data is essential for understanding user behavior and engagement on the website, and we believe that these values should be retained in our analysis.

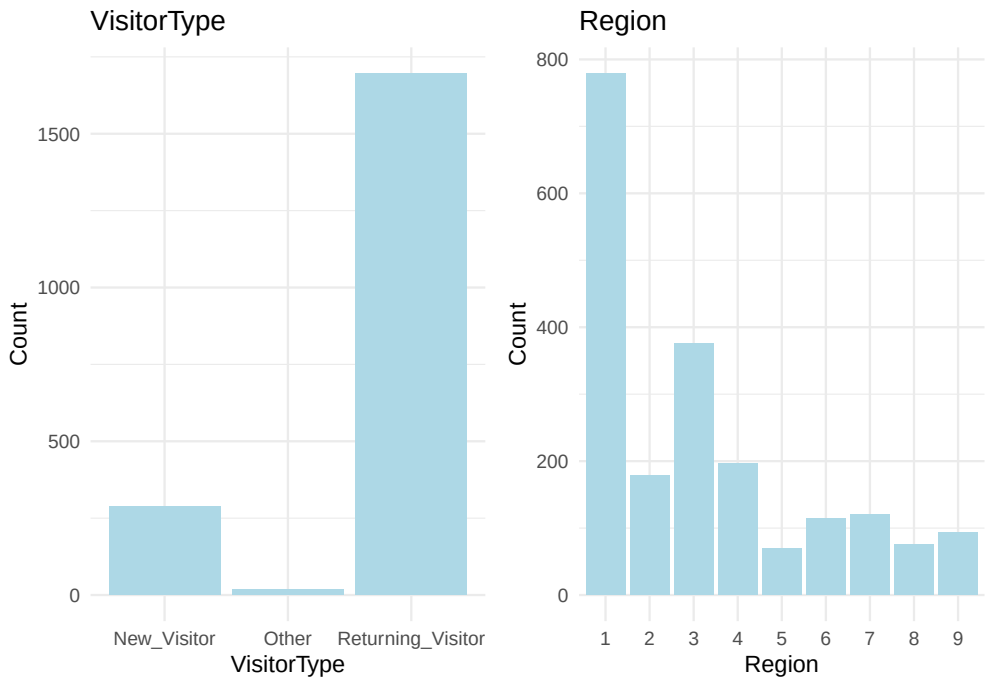


2.2 Categorical Variables

2.2.1 Visitor Demographics/Identity

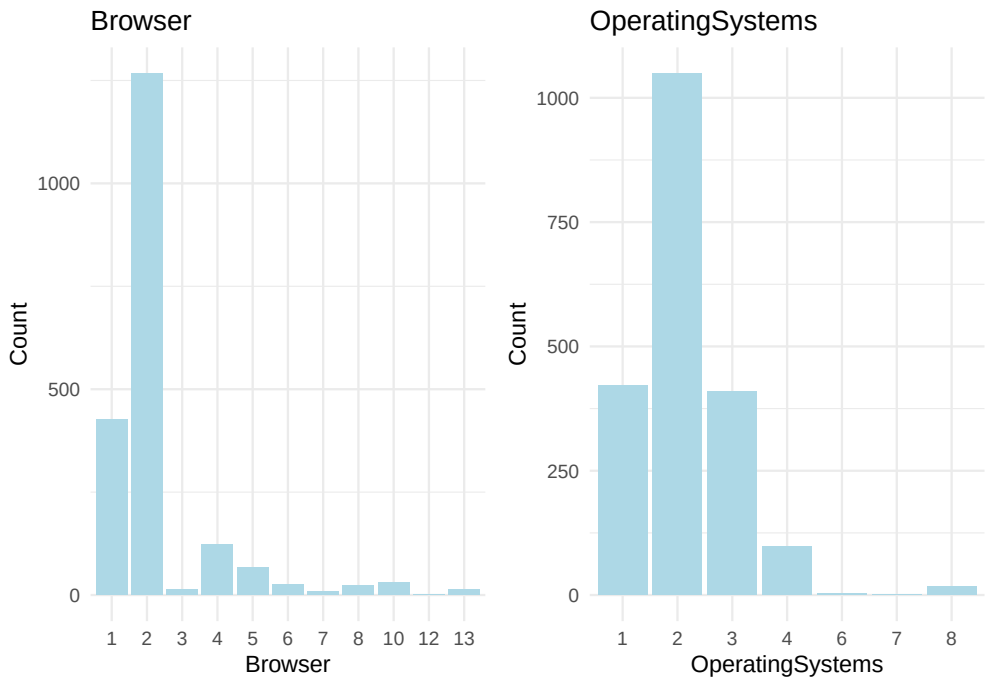
- **VisitorType:**
  The majority of visitors are **Returning Visitors**, indicating a strong base of repeat users. **New Visitors** form a smaller proportion, and the **Other** category is negligible, suggesting minimal contribution from other visitor types.

- **Region:**
  Region **1** has the highest count of visitors, followed by Region **3**. Traffic is highly concentrated in these specific regions, with other regions showing a relatively lower number of visitors.



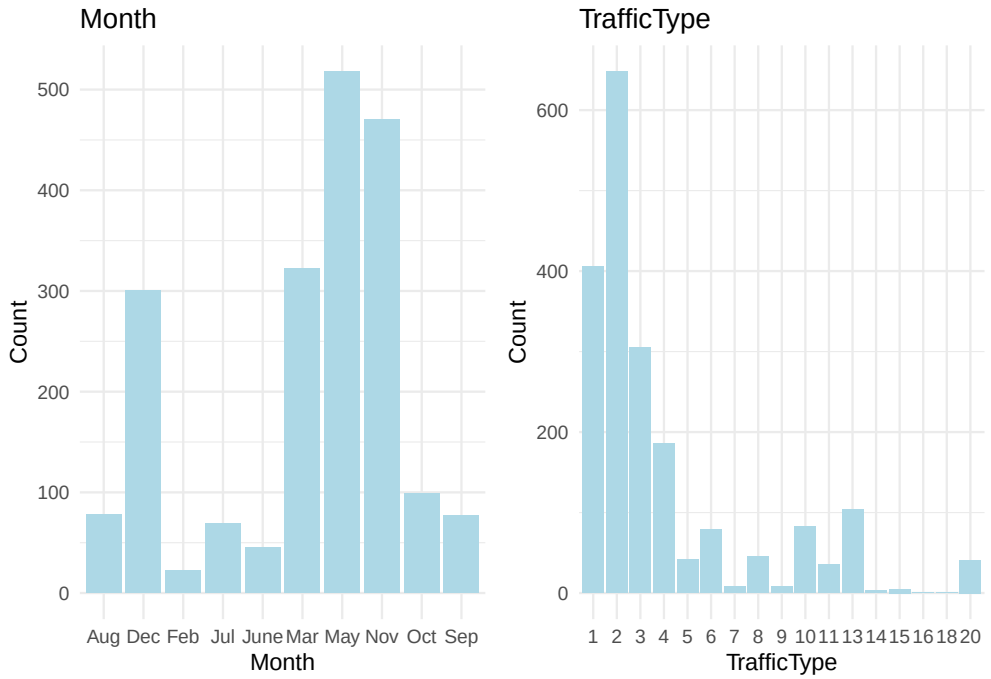
2.2.2 Technical Attributes

- **Browser:**
  Browser **2** is the most commonly used browser, significantly surpassing others in popularity. Browser **1** has a notable user base, while the other browsers are used by only a small fraction of users.
- **OperatingSystems:**
  Operating System **2** dominates usage, followed by Operating System **1**. Other operating systems have a minor presence, highlighting user preference for a few dominant operating systems.



2.2.3 Traffic and Source

- **Month:**
  Traffic peaks in **March**, **May**, and **November**, with the highest activity in **May**. Lower traffic is observed in **June**, **July**, and **February**, potentially reflecting seasonal trends or business cycles.
- **TrafficType:**
  TrafficType **2** is the primary source of traffic, with a significant lead over TrafficType **1**. Other TrafficTypes contribute marginally, indicating that a few referral sources or marketing channels drive the majority of traffic.



2.3 Binary Variables

- **Weekend:**
  The binary variable `Weekend` indicates whether the session occurred on a weekend. We observe that around ~23% of the sessions occurred on weekends, while the majority (~77%) took place on weekdays. This distribution suggests that the website receives a higher volume of traffic on weekdays compared to weekends (which also holds if we took the daily average).

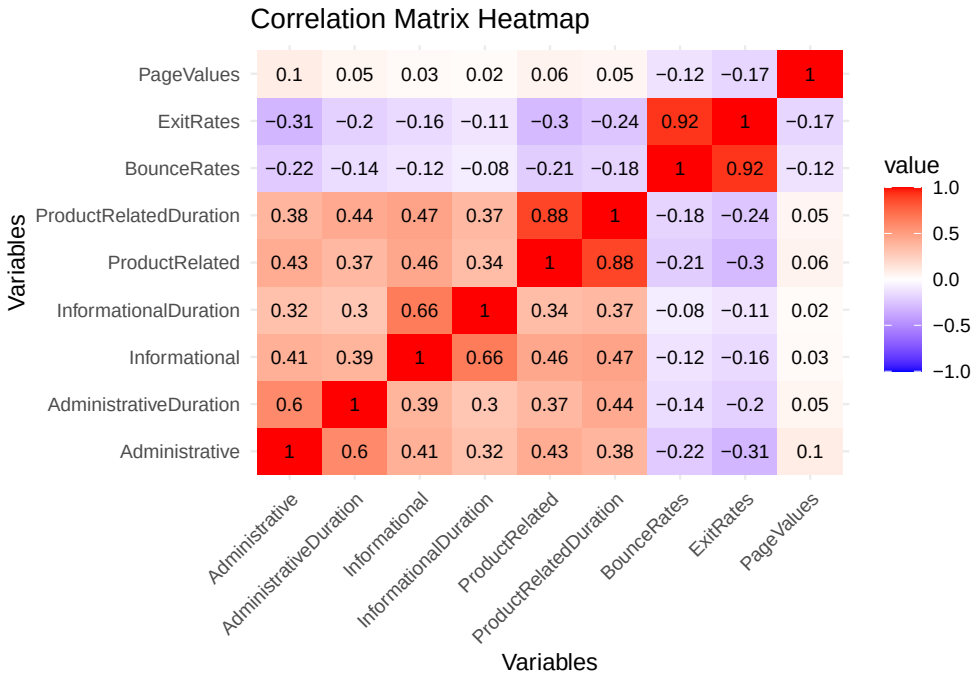
- **Revenue:**
  The target variable `Revenue` indicates whether a session resulted in a purchase. The dataset is imbalanced, with a higher number of sessions where no purchase (~85%) was made compared to sessions resulting in a purchase (~15%). This imbalance is expected in e-commerce datasets, where the conversion rate is typically lower than the non-conversion rate.

3. Multivariate Analysis

3.1 Correlation Analysis

The correlation matrix provides insights into the relationships between the numeric variables in the dataset. The heatmap below visualizes the correlation matrix, with values ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. We can observe 3 distinct clusters of variables:

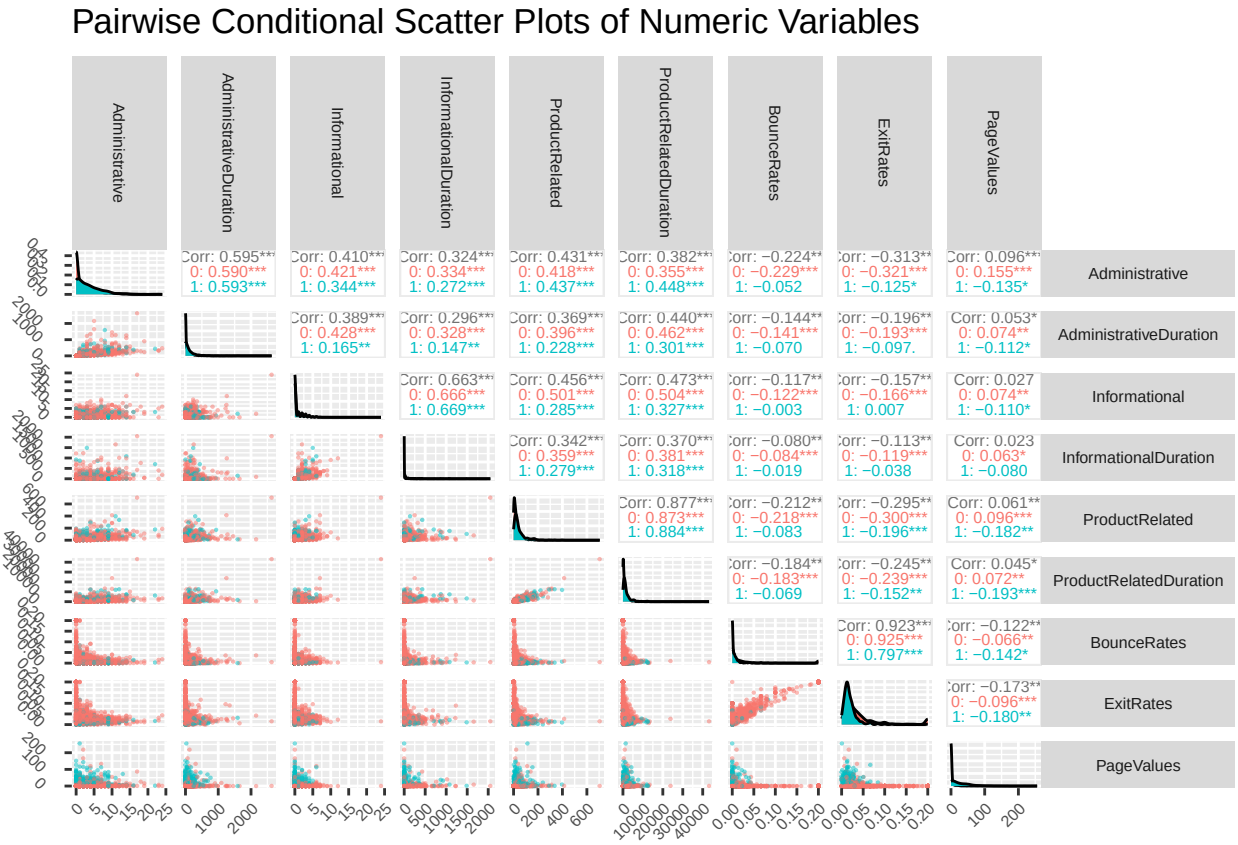
- Page Related Variables like `Administrative`, `Informational`, `ProductRelated`, `AdministrativeDuration`, `InformationalDuration` and `ProductRelatedDuration`are moderately positively correlated with each other.
- `BounceRates` and `ExitRates` are highly positively correlated with each other and negatively correlated with the rest of the variables.
- `PageValues` shows very little to no correlation with the rest of the variables.



3.2 Pairwise Conditional Scatter Plots

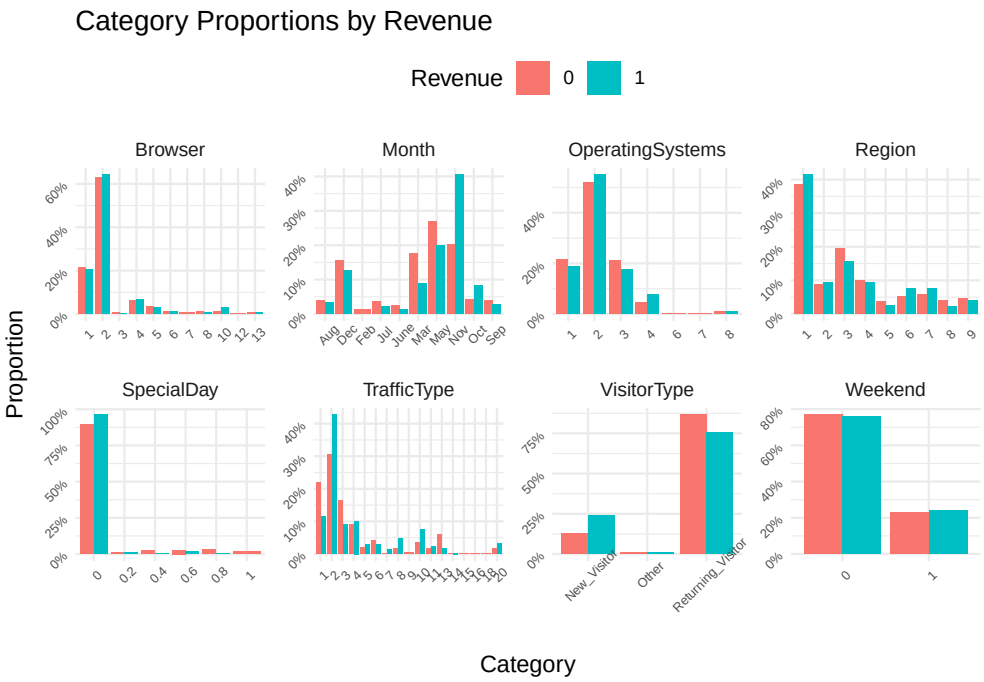
This conditional pairwise scatter plots provide a visual representation of the relationships between numeric variables, colored by the `Revenue` variable. We observe what we've already seen in the correlation matrix and in the univariate analysis. We see clearly skewed distributions for all variables due to the nature of the data, and as seen in the correlation matrix, we do not observe any clear linear relationships between the variables except for `BounceRates` and `ExitRates` which are significantly positively correlated between them.

Another interesting observation is that `Revenue` does not seem to have a clear separation between the two classes in the scatter plots. This suggests that the numeric variables alone may not be sufficient to predict the `Revenue` variable.



3.3 Conditional Barplots

The conditional barplots below show the proportion of sessions resulting in a purchase (`Revenue`) for each categorical variable. The plots reveal little variation in the proportion of purchases across different categories within each variable. This suggests that the categorical variables alone may not be sufficient to predict the purchase intent of a session. However, these variables could still provide valuable insights when combined with other features in a predictive model. The only variables that seem to have some variation in the proportion of purchases are `VisitorType`, `Month` and `TrafficType` where some of the categories have a higher proportion of purchases than others. For example, in the `VisitorType` variable, `Returning_Visitor` shows a lower proportion of purchases while the `New_Visitor` category has a higher proportion of purchases.



4.0 Principal Component Analysis

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining as much variance as possible. This technique is instrumental in identifying key patterns in the data and visualizing its underlying structure.

4.1 Methodology

After initial data preprocessing and analysis, we observed that the dataset contains numerical variables with differing scales. Given the nature of the data, we opted not to apply additional transformations, as explained previously. To address the issue of differing scales, we applied PCA using the correlation matrix instead of the covariance matrix. This approach standardizes the variables to have a mean of 0 and a standard deviation of 1, ensuring comparability across variables.

While we previously noted that the numerical variables are not highly correlated pairwise, it is also valuable to evaluate intercorrelation measures for the entire dataset. PCA performs optimally when the data exhibits strong intercorrelation, so these measures will provide insight into how effective PCA is likely to be for this dataset.

Table 6: Intercorrelation Measures

Measure	Value
Multivariate Dispersion	0.8098
KMO-like Measure	0.7298
Bartlett's Determinant Test	0.9298
Multivariate Kurtosis	0.2507
Multicollinearity Index	0.9628
Average Variable Dependency	0.5815

The intercorrelation measures indicate that the dataset exhibits moderate to high intercorrelation among variables. This suggests that applying PCA is appropriate, as it will effectively reduce dimensionality while preserving most of the variance in the data. The results highlight a sufficient level of correlation to justify PCA, with some redundancy among variables that PCA can address.

4.2 Results

The following table displays the principal components along with the variance explained by each component. The first principal component explains the highest proportion of variance, followed by subsequent components.

Table 7: Principal Components with Variance Explained

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Administrative	0.3711	0.0330	0.2195	-0.0522	0.5251	0.7137	-0.0452	-0.1458	0.0444
AdministrativeDuration	0.3471	0.1235	0.1949	0.0154	0.6028	-0.6627	-0.0494	0.1449	0.0104
Informational	0.3735	0.2610	0.0453	-0.3945	-0.2381	-0.0248	0.7598	0.0008	-0.0116
InformationalDuration	0.3187	0.2662	0.0750	-0.5430	-0.3380	-0.0269	-0.6413	0.0300	0.0046
ProductRelated	0.4174	0.1120	-0.2210	0.4742	-0.2021	0.1538	-0.0409	0.6701	0.1560
ProductRelatedDuration	0.4171	0.1564	-0.2178	0.4594	-0.1877	-0.1235	-0.0711	-0.6889	-0.1143
BounceRates	-0.2501	0.6305	0.1651	0.1667	0.0441	0.0985	-0.0104	0.1339	-0.6747
ExitRates	-0.2902	0.6081	0.1115	0.1155	0.0289	0.0067	0.0073	-0.1235	0.7099
PageValues	0.0761	-0.2007	0.8776	0.2638	-0.3335	-0.0391	0.0110	-0.0101	0.0344
Variance Explained	**39.7595**	**19.1530**	**10.7535**	**10.3421**	**9.8792**	**4.5367**	**3.5797**	**1.2411**	**0.7552**

4.3 Principal Component Selection

5. Packages and Tools

The analysis was performed using **R**, with the report generated in **RMarkdown**. For data wrangling, the `dplyr` and `reshape2` libraries were utilized. Visualizations were created using `ggplot2`, `GGally`, and `gridExtra`. The report was compiled and formatted with `knitr` and `pander`, while `caret` was employed for the Principal Component Analysis.

6. References

- UCI Machine Learning Repository. (n.d.). **Online Shoppers Purchasing Intention Dataset**. Retrieved from https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset.