Data Preprocessing:

1. The dataset was organized neatly and did not contain any null values.
2. 125 duplicate values: They must be removed at the beginning of the preprocessing steps.
3. The outliers were NOT handled as important information would have been erased if they were removed.
4. More traffic near special occasions did not result in more revenue generation, so it can be stated that for this particular dataset, the feature ‘‘SpecialDay’’ is not affecting the Revenue.

(Removing the “SpecialDay” column)

Features Engineering:

The purpose of feature engineering is to aggregate common features and record feature interactions.

1. ‘‘Administrative’’, ‘‘Informational’’, and ‘‘ProductRelated’’ were all representing the number of page visits by the user, they were aggregated into one, named ‘‘total\_page\_view’’.
2. ‘‘Administrative\_Duration’’, ‘‘Informational\_Duration’’, and ‘‘ProductRelated\_Duration’’ were summed into a new column ‘‘total\_page\_duration’’.

(After creating these two new features, the individual six features were dropped from the data frame.)

1. After some thorough investigation, it was found that features such as ‘‘SpecialDay’’, ‘‘OperatingSystems’’, ‘‘Browser’’, ‘‘Region’’, ‘‘TrafficType’’, and ‘‘Weekend’’ were not related to the proposed LRFS analysis. Hence, those six columns were removed as well.
2. After that, the remaining categorical values ‘‘Month’’, ‘‘VisitorType’’, and ‘‘Revenue’’ was replaced by numerical values.
   1. The feature ‘‘Month’’ is important for any e-commerce analysis, as it shows the time of the year in which the site might be crowded.
   2. One of the most prominent categorical features in this dataset was the Visitor Type, which tells whether a customer came back to the platform or not. The ‘‘Other’’ visitor type (labelled as 2) holds a very negligible percentage, so for the sake of easier calculation, it was considered with the ‘‘New\_Visitor’’ type.

Due to the Correlation Matrix:

1. All Bounce Rates are considered to be the Exit Rates. So, to avoid redundancy, in this research, only the Exit Rates were taken for analysis purposes.
2. The more the pages are being accessed by the users, the more time they are going to spend. So, the Total Page Duration was removed this time and only the Total Page Views were taken for further calculation for building the LRFS model.
3. Generally, if customers keep exiting the website frequently, it does not lead to better revenue generation. On the other hand, if the user spends more time on the website, it could increase the possibility of more purchases from the website. Hence, the Staying Rate, which is the opposite of Exit Rates was given a vital role to introduce the new dimension of the LRFS model.

Features Extractions (LRFS):

It is not possible to directly calculate each customer’s recency or frequency. However, there are some very important features in this dataset which has the potential to make better customer segmentation. New features were created by combining or tweaking the existing ones and were mapped as components namely ‘‘L’’, ‘‘R’’, ‘‘F’’, and ‘‘S’’.

1. Length (L): Length indicates the interval between two particular visits. So in general, the equation of Length would be, .
2. Recency (R): Recency tells how recently the customer has visited. The highest value for the month of the whole dataset, meaning the most recent month for that particular website (December in this case) was taken and subtracted from the current month for each customer. So the Recency for each customer “*i*” was calculated as, .
3. Frequency (F): Frequency tells the number of times a customer has visited a website, so .

The number of total page visits made by the customer, namely column ‘‘total\_page\_view’’, the summation of Administrative, Informational, and ProductRelated sessions, were taken as the component Frequency.

1. Staying Rate for Revenue (S): It symbolizes the essence of a particular customer staying on the website and eventually contributing to the revenue.

In the definition of ‘‘S’’, the term ‘‘Revenue’’ refers to both Website Revenue and Goal Value.

The ExitRates have a negative relationship with the feature PageValues. So the goal should be to minimize the exit rates as much as possible by consistently checking on it and updating the websites accordingly. That being said, (1 - Exit Rates) is the exact opposite of the Exit Rates, referring to the value with which the user stayed on the page, which is why it was named Staying Rates in this research.

Now, multiplying it with Page Value gives us a ratio of Revenue per Staying Rate, so looking at the component ‘‘S’’ would tell us how much revenue a certain customer is contributing based on how long the customer has been staying on that website. The final equation would be:

1. Next, feature scaling was performed with the help of a data standardization technique, standard scaling; the retrieved features are rescaled to a specific range, where the attribute’s mean value is 0 and the distribution’s standard deviation is 1.

CORRELATION OF COMPONENTS WITH REVENUE:

For identifying the customer segments, all the components of the LRFS model also the Revenue column have to be considered. the newly introduced parameter S has a weak association with each of the other parameters, which makes S a unique trait that cannot be replaced by other features. Also, it has a moderate correlation with the Revenue.

MODEL SPECIFICATION:

Several well-known techniques, including the Elbow Method, the Silhouette Coefficient, the Cumulative Explained Variance Ratio, dimensionality reduction techniques (PCA, t-SNE, Autoencoder), and two of the most widely used clustering algorithms (K-Means and K-Medoids) were used for the model specification. Each of these methods has assisted in the discovery of the independent attributes of the proposed model.