Machine Learning in E-Commerce: Predicting Purchase Conversion

Overview

Machine learning is a powerful tool for predicting customer purchase conversion in e-commerce. By leveraging data-driven insights, businesses can better understand customer behavior and optimize their marketing strategies.

Research Question

Research questions are essential for understanding the potential of machine learning in e-commerce. By asking the right questions, we can gain insights into how to best predict purchase conversion.

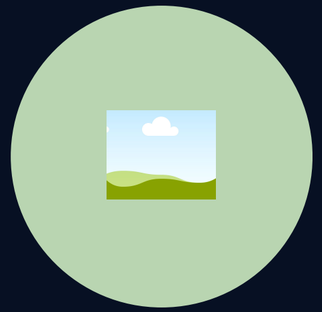
Data Set

Data sets are essential for machine learning in e-commerce. They provide the information needed to accurately predict customer purchase conversion. The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

Understanding Attribute Information in an E-commerce Site

- **10 numerical attributes**
  The dataset consists of 10 numerical attributes
- **8 categorical attributes**
  The dataset consists of 8 categorical attributes
- **Revenue attribute as class label**
  The 'Revenue' attribute can be used as the class label
- **Page visits and time spent**
  Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories
- **Bounce Rate**
  Bounce rate refers to the percentage of visitors who leave a website after viewing only one page. It indicates that these visitors did not find the content or resources they were seeking on the site. A high bounce rate suggests that the website may not be meeting visitor expectations or that there could be issues with the relevancy or usability of the content.
- **Exit Rate**
  On the other hand, the exit rate measures the percentage of visitors who exit a website from a specific page rather than continuing to view other pages. It helps identify pages where visitors are more likely to leave the website. A high exit rate on a specific page may indicate that the page is not providing the desired information or resources, or that there are design or functionality issues affecting user experience.
- **Special Day feature**
  Special Day feature indicates the 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 the transaction
- **Operating system, browser, region, traffic type, visitor type**
  Operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

Exploratory Data Analysis

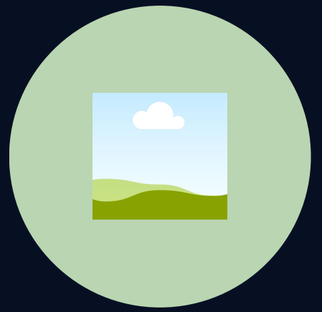


Define the Problem

Identify the goal of the analysis and the data needed to achieve it.

Data Collection

Gather and organize the data from various sources.



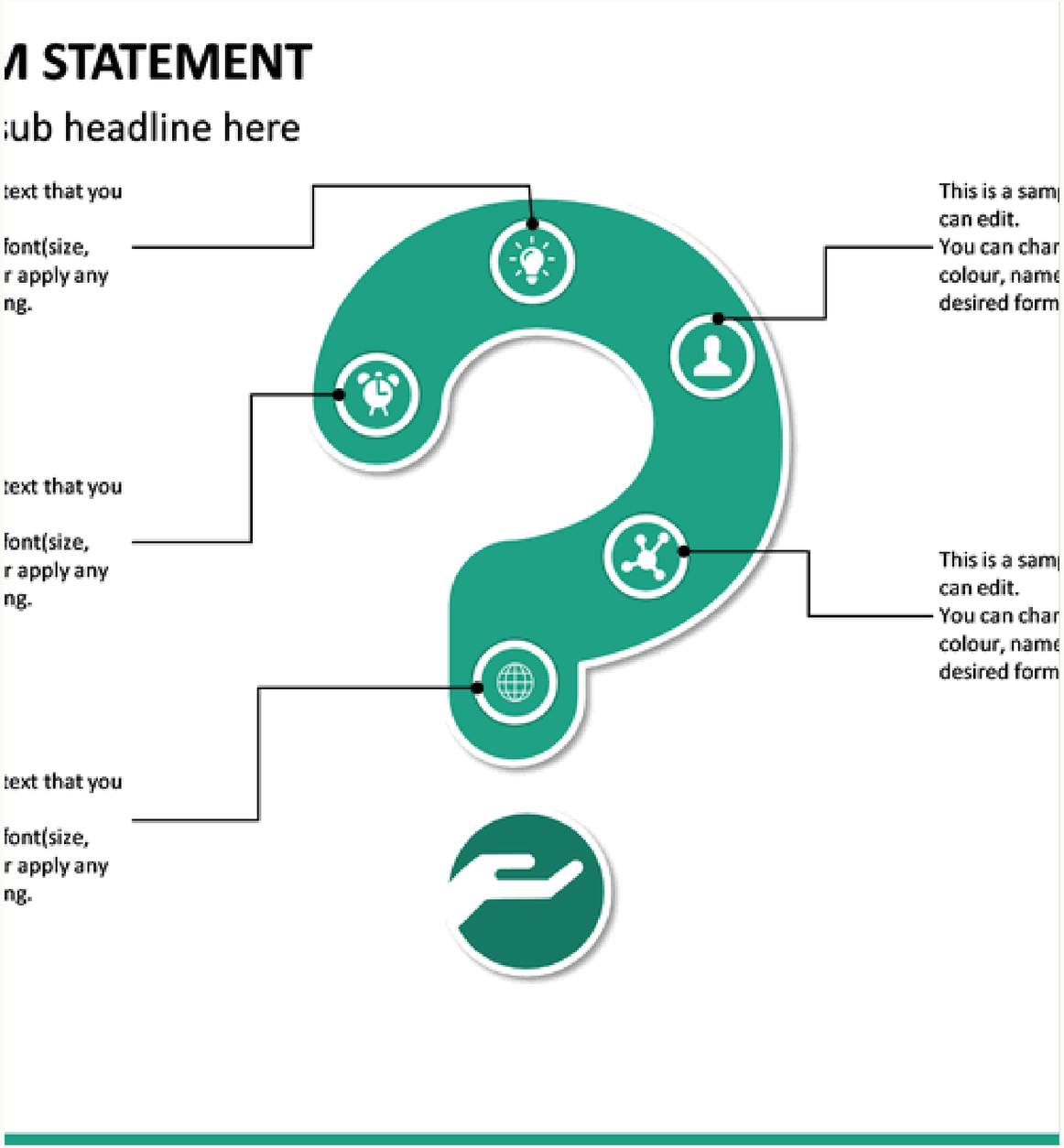
Data Cleaning

Remove any outliers or missing values from the data.

Data Exploration

Analyze the data to identify patterns and trends.

Exploratory Data Analysis is an important step in Machine Learning in E-Commerce, as it helps to identify patterns and trends in the data that can be used to predict purchase conversion.



Our Problem Statement

The online retailing company aims to identify online shoppers who are most likely to generate revenue based on their activity on the company's website. The company wants to leverage the available data on shoppers' online behavior to develop a predictive model that can accurately classify potential revenue-generating shoppers. By solving this problem, the company aims to optimize its marketing strategies, personalize user experiences, and maximize revenue generation from its online platform.

Eliminating NULL Values & Duplicates: A Data Cleaning Guide

Identifying NULL Values

The dataset has good data quality in general. It is in a tidy format, and does not contain NULL values.

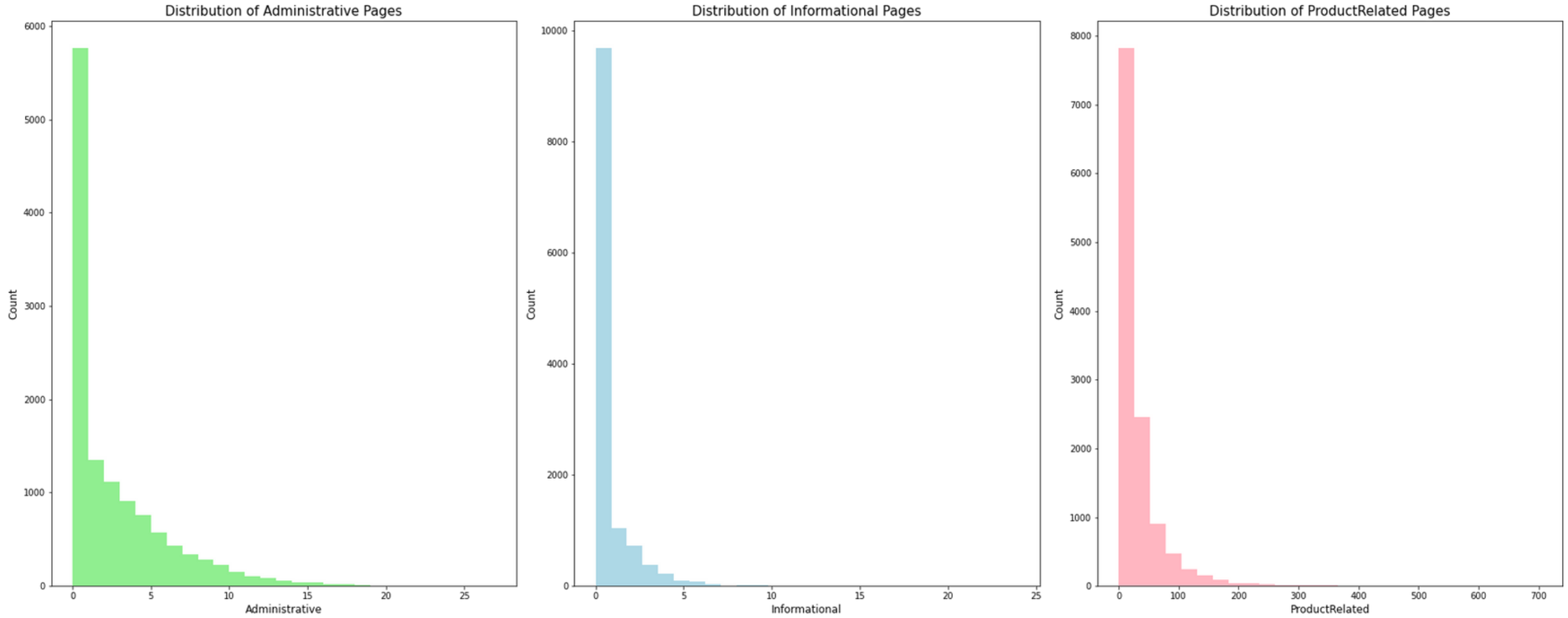
Handling NULL Values

The dataset we received was clean as such there is no need for handling NULL Values.

Data cleaning is an important step in data analysis, and understanding how to identify and handle NULL values and duplicates is essential for successful data cleaning.

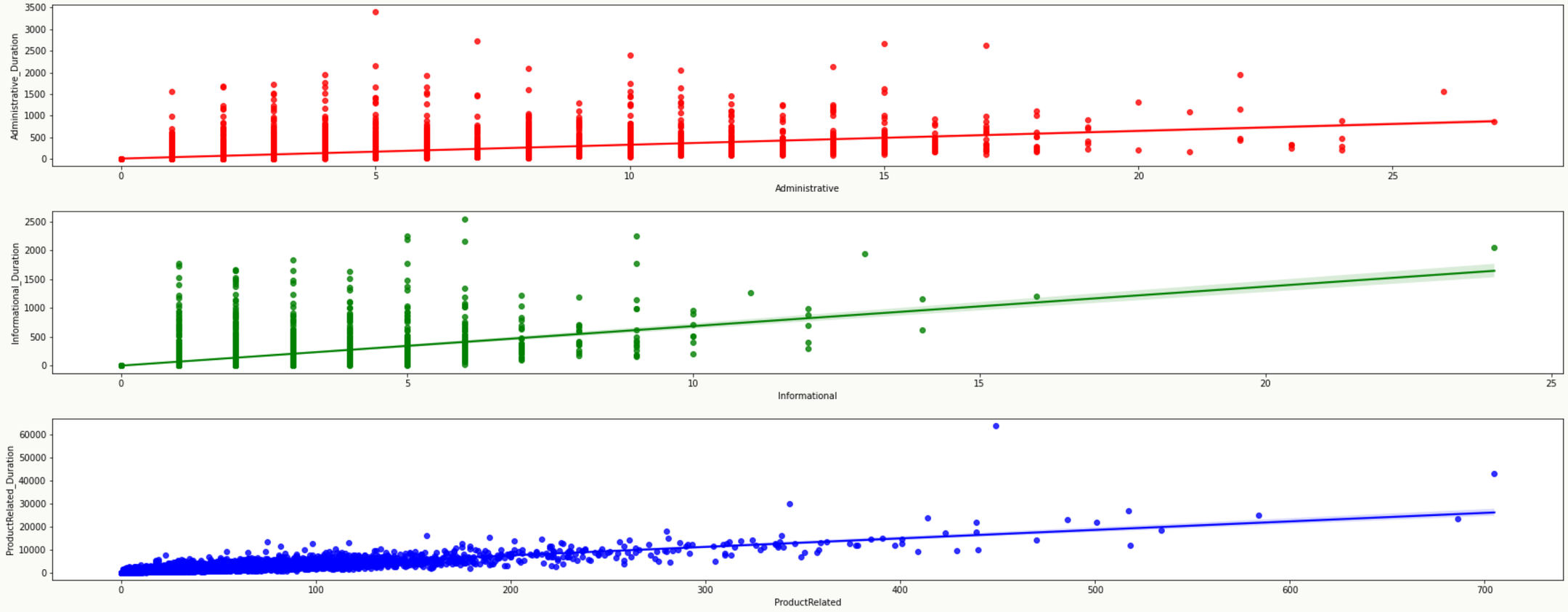
Exploratory Data Analysis

"Administrative", "Administrative Duration", "Informational", and "Informational Duration" represents the number of different types of pages visited by the visitor in that session and the total time spent in each of these page categories.

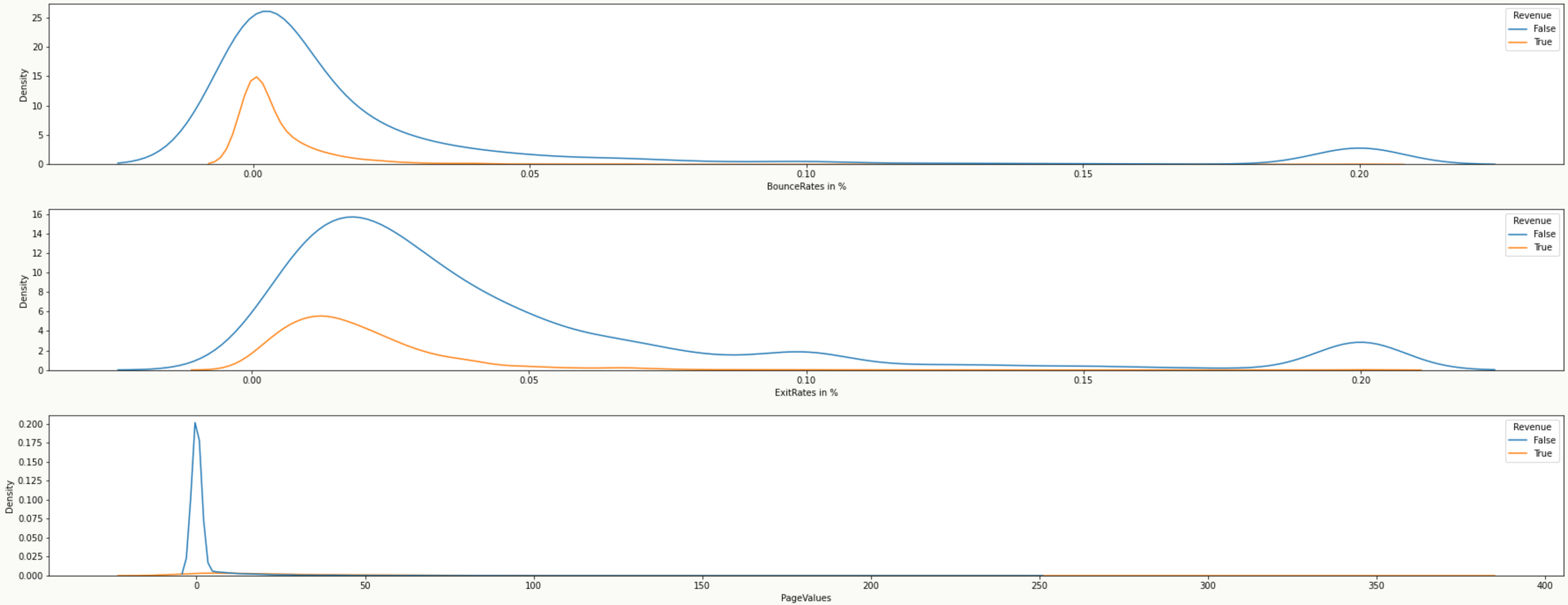


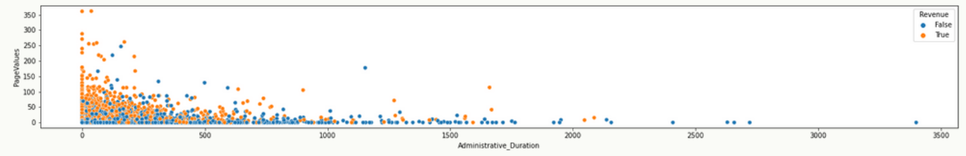
In the plot, the right-skewed distribution of the pages indicates that there are a few pages that are visited very frequently, while the majority of pages have lower visit counts.
For the administrative, informational, and product-related pages, the most frequently visited page is 0, meaning that a large number of users access this page. On the other hand, as the user navigates to more pages the frequency of a page decreases, suggesting that it is visited by very few users.

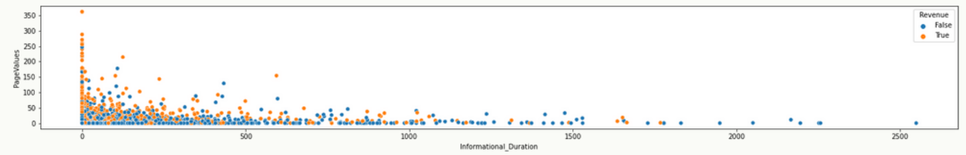
The regplot reveals an interesting relationship between the number of pages visited and the duration of time spent by the user. As the user navigates to more pages, there is a noticeable increase in the duration of time spent, indicating a positive correlation. This suggests that users find the content on these pages useful and engaging, prompting them to spend more time exploring.
The scatter points on the graph show an increasing trend, suggesting that as the number of pages visited increases, the duration of time spent tends to be higher. This observation further supports the notion that users are actively engaging with the content and finding it valuable, as evidenced by their prolonged stay.
The regression line in the plot exhibits a strong positive relationship between the number of pages visited and the duration of time spent. It signifies that as users navigate to a greater number of pages, they tend to spend more time on the website. This finding aligns with our initial hypothesis and reinforces the notion that users are drawn to the content and are willing to invest additional time in exploring it.

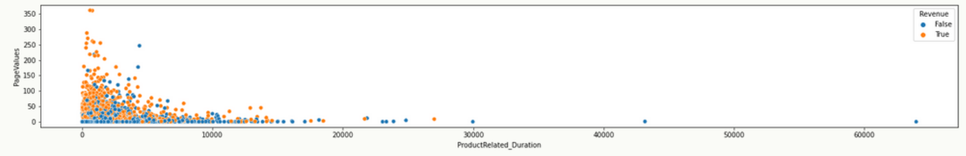


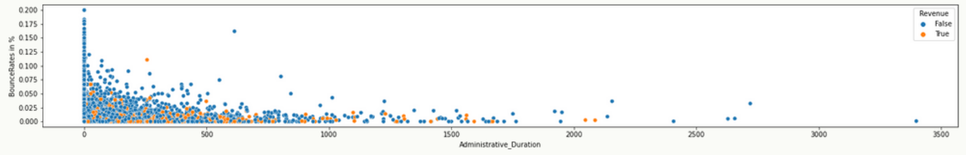
From the kdeplot we can infer that the distribution for exit rates and bounce rates should be low to generate high revenue and the page values should be higher for high revenue. For effective revenue generation low bounce rates, low exit rates, and high page values hold the key. From the below plot, it's clear that Customers with low exit rates and low bounce rates generated more revenue.

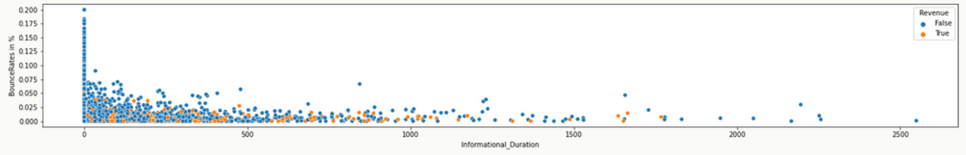


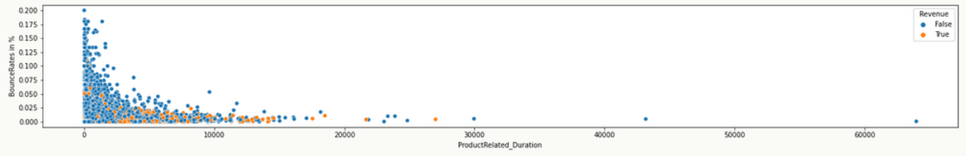


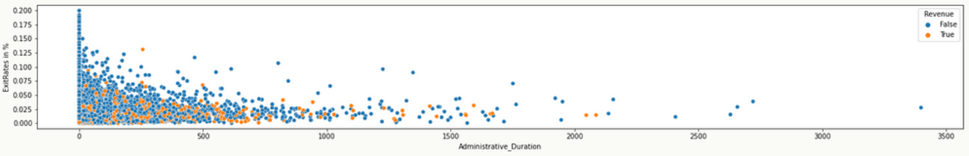


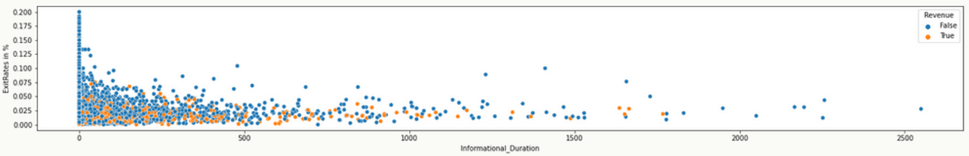


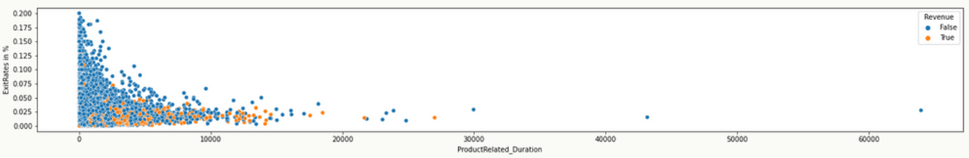










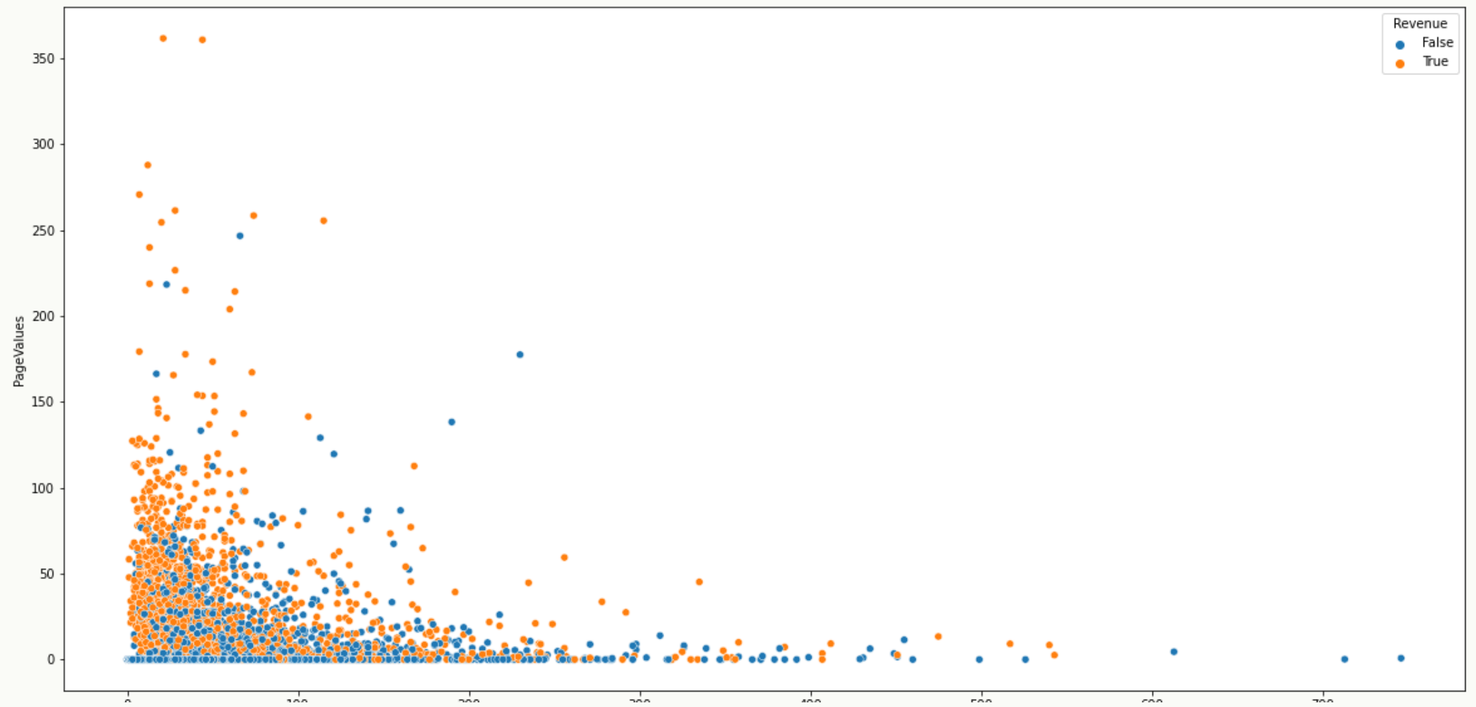


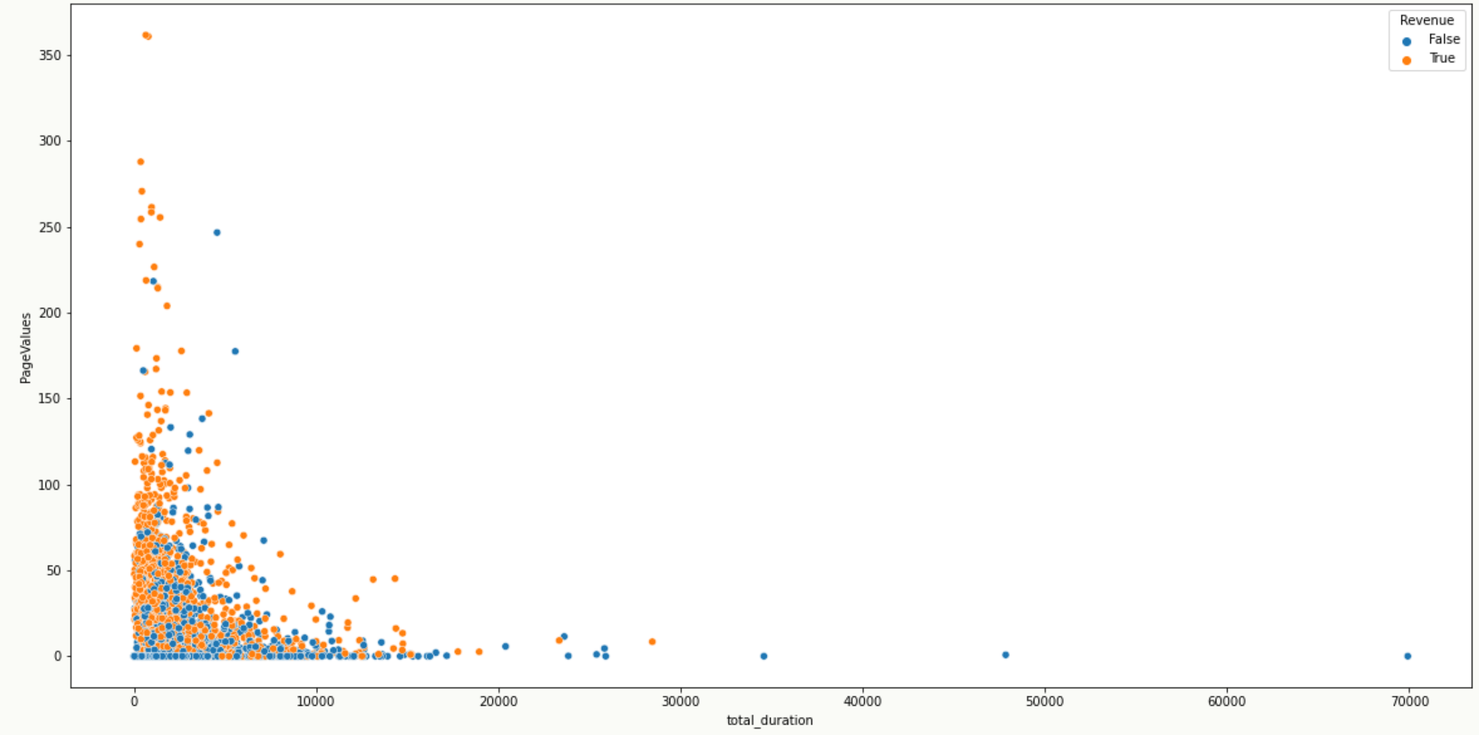
The charts provide insights into the relationship between time spent on each website and the generated revenue. It is observed that individuals who contribute to revenue generation tend to exhibit lower bounce rates and spend a longer duration of time on the website.

This analysis highlights the significance of user engagement and the potential impact it has on revenue generation. The lower bounce rate indicates that visitors are more likely to explore multiple pages and stay engaged with the website's content. Additionally, the longer duration of time spent suggests a higher level of interest and involvement in the offerings, potentially leading to increased conversion and revenue.

By understanding this relationship, businesses can identify patterns and behaviors that contribute to revenue generation. They can focus on improving user engagement, reducing bounce rates, and enhancing the overall user experience to maximize revenue potential.

Overall, this analysis sheds light on the importance of optimizing user engagement and time spent on the website to drive revenue growth.



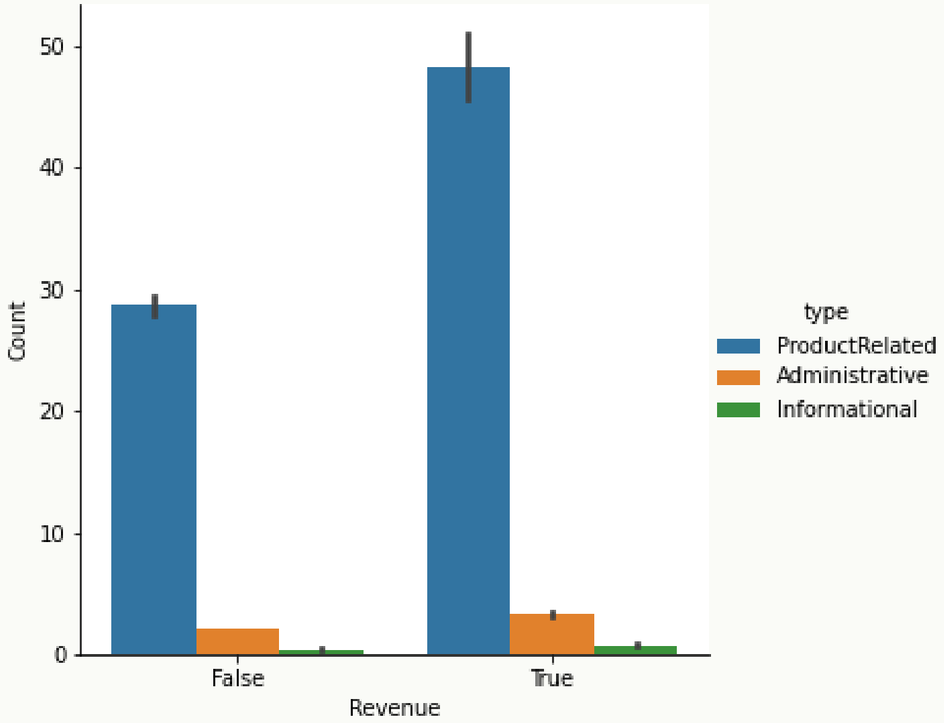


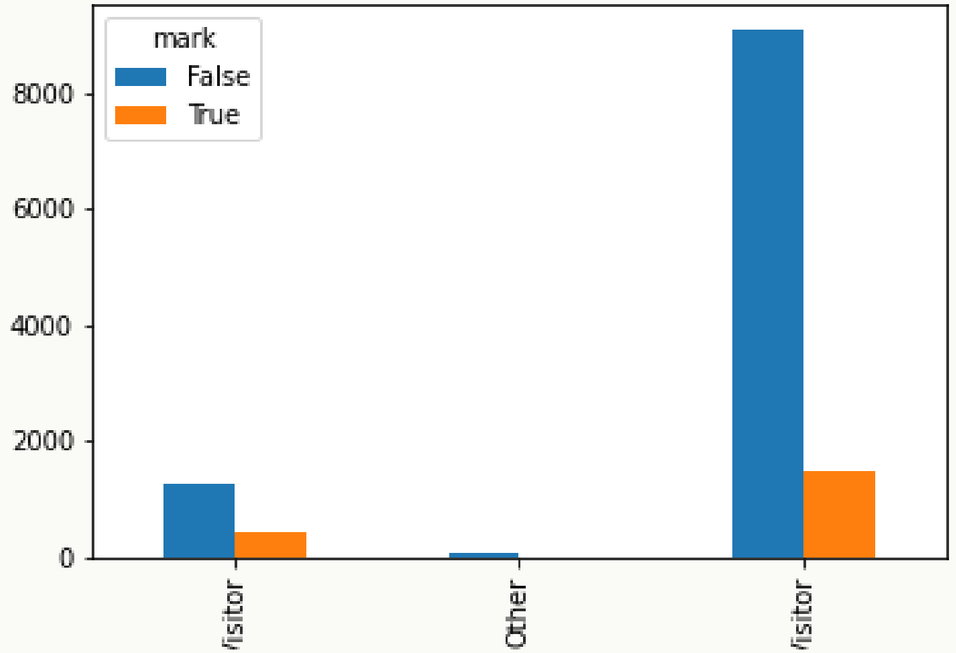
To gain a comprehensive understanding of user behavior on a broader scale, a new feature called "Total Duration" was created. This feature represents the sum of all the durations, encompassing the total time spent by users on the website. By analyzing patterns using this aggregated metric, we can uncover valuable insights.

One interesting finding is the relationship between the "Page Values" metric and the total duration of time spent on the website. It is observed that higher page values correspond to longer durations of user engagement. This suggests that pages with higher value, in terms of their relevance, usefulness, or attractiveness to users, tend to capture their attention and encourage them to spend more time exploring the website.

Additionally, the analysis reveals a correlation between the "Page Values" and the total number of user sessions. Websites that generate higher page values also tend to attract more user sessions, indicating a higher level of user engagement and interaction with the content. This suggests that users find the website valuable and are willing to spend more time actively engaging with its features and offerings.

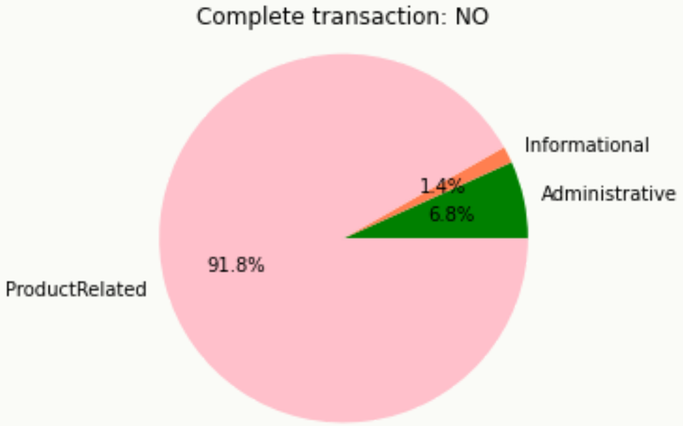
In summary, the creation of the new feature, "Total Duration," allows for a broader perspective on user behavior. The analysis demonstrates a positive association between page values, the total duration of time spent on the website, and the number of user sessions. These findings highlight the importance of providing valuable and engaging content to users, as it can lead to increased user engagement, longer durations of time spent on the website, and higher levels of user interaction.

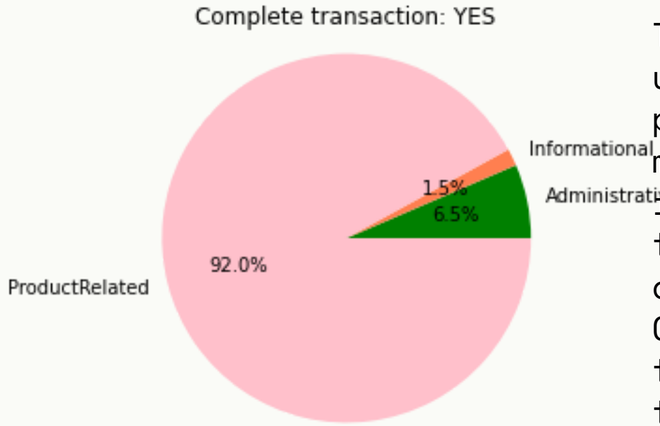


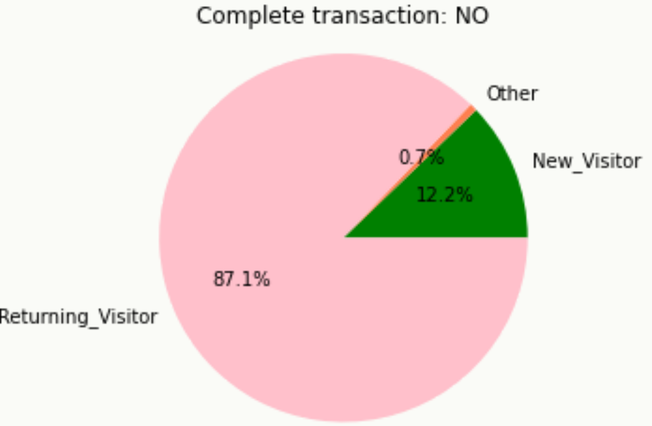


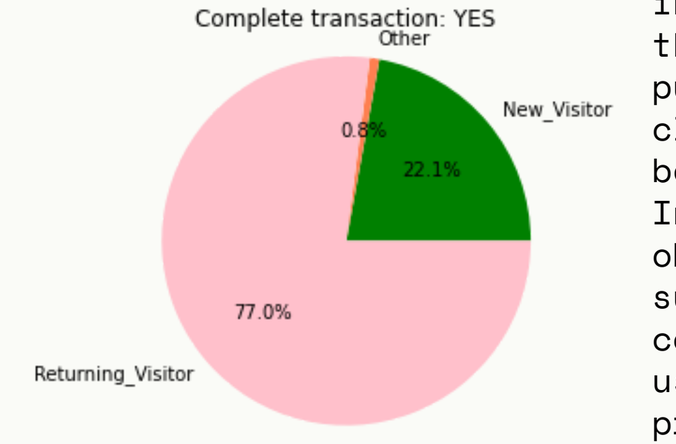
Inference from plots above and below. Customers visited the ProductRelated web page more as compared to Administrative and Informational. At this point of visual EDA it is becoming seemingly clear that the Admin page is the HOME page, Informational is CONTACT US, and the Product related is the actual webpage for the product in interest. This exclusive info was not provided in the

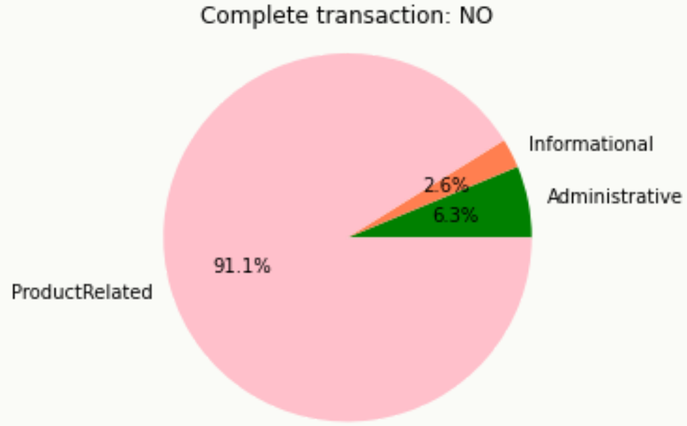
Most of the customers visiting the website are Returning visitors, contributing to the most number of purchases. About 1/4th of New_visitors made the purchase as compared to ~15 % of Returning visitors.

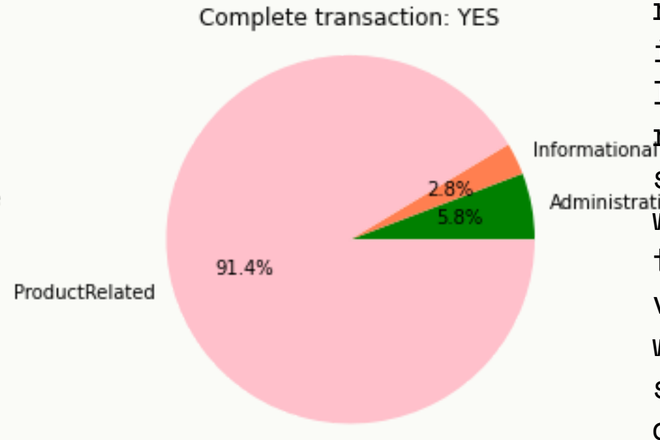










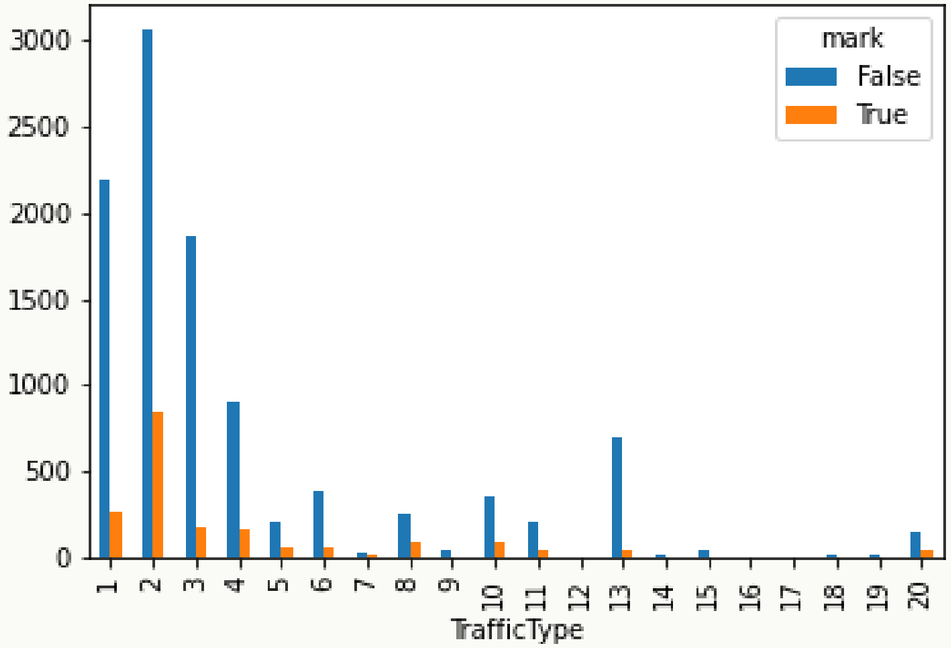


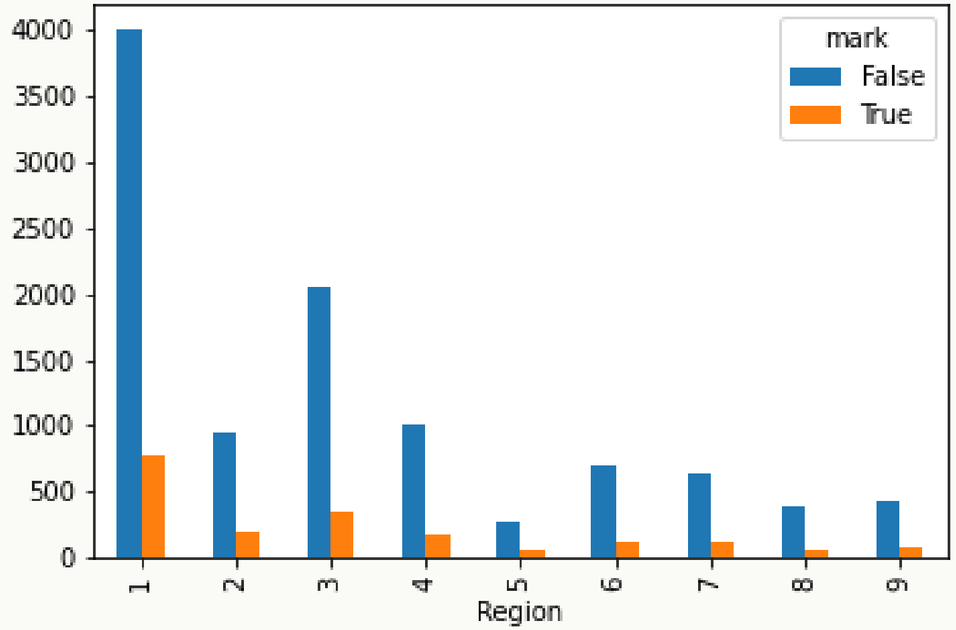
The website analytics reveal some interesting patterns in user behavior. It is evident that users primarily visit product-related pages, which aligns with their intention to make purchases rather than focusing on administrative or informational content. This finding reinforces the notion that users are more inclined towards exploring product offerings and making buying decisions.

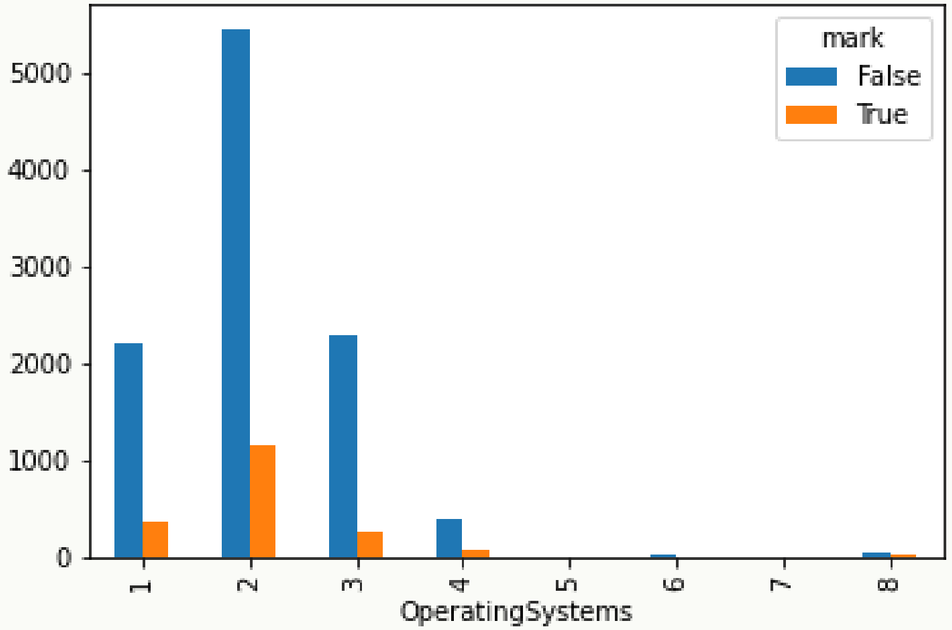
On average, users who successfully complete transactions tend to visit more webpages and spend a longer duration on the website. This indicates a higher level of engagement and involvement in the purchase process. However, when comparing the types of webpages visited between users who complete the purchase and those who do not, there doesn't seem to be a clear distinction. Both groups exhibit similar browsing behaviors in terms of the types of pages accessed.

Interestingly, a higher proportion of new visitors is observed among those who successfully complete purchases, suggesting that the website effectively attracts and converts new users into customers. On the other hand, among users who do not complete purchases, there is a higher prevalence of returning visitors. This implies that returning visitors may require additional strategies or incentives to encourage them to complete the purchase.

In summary, the data indicates a strong focus on product-related pages and higher engagement among users who successfully complete transactions. However, the types of webpages visited do not show a clear differentiation between the two groups. The presence of a larger percentage of new visitors among successful purchasers highlights the website's ability to attract and convert new users. Tailored strategies may be necessary to enhance conversion rates among returning visitors.





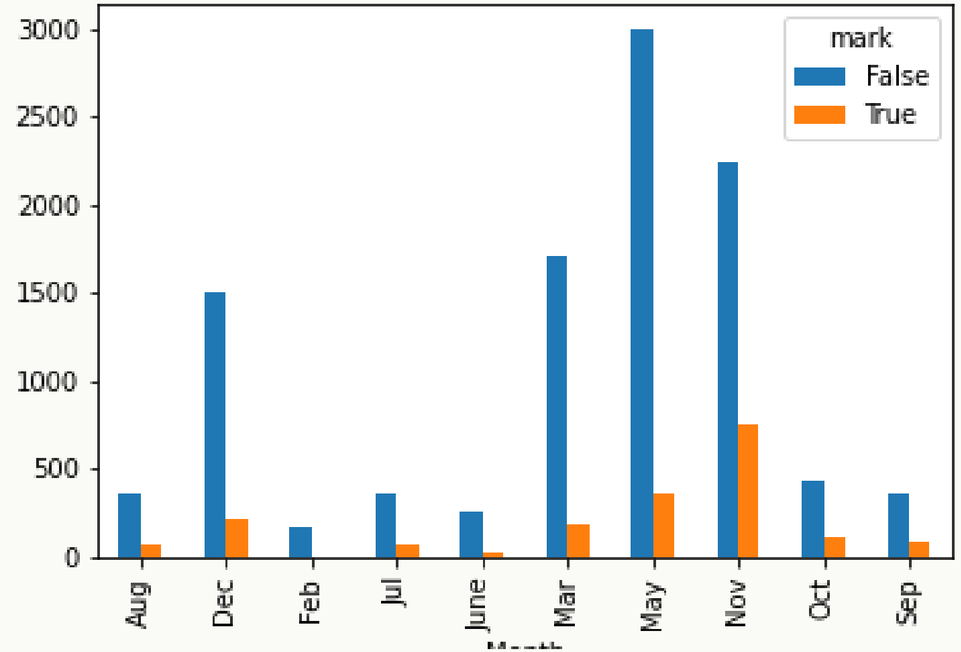


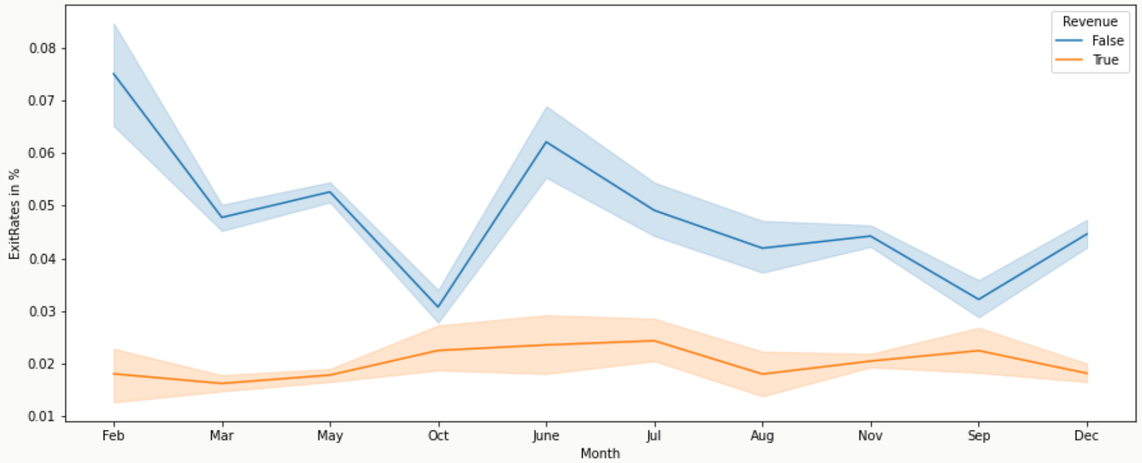
The revenue generation analysis based on different factors reveals interesting insights. Users operating on OS type 2 contribute significantly to the overall revenue, indicating the importance of targeting and optimizing experiences for this specific operating system.

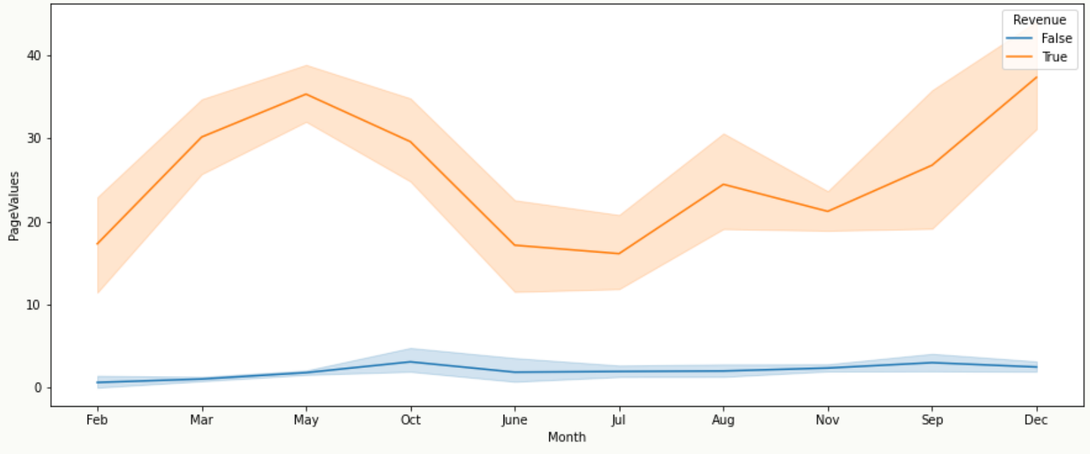
When considering traffic types, it is evident that traffic type 2 generates the highest revenue, followed by types 1 and 3. These traffic types should be prioritized in marketing and promotional strategies to maximize revenue generation. Conversely, traffic types 14 to 20 exhibit minimal contributions to revenue, suggesting a potential opportunity for improvement or reconsideration of targeting strategies for these types of traffic.

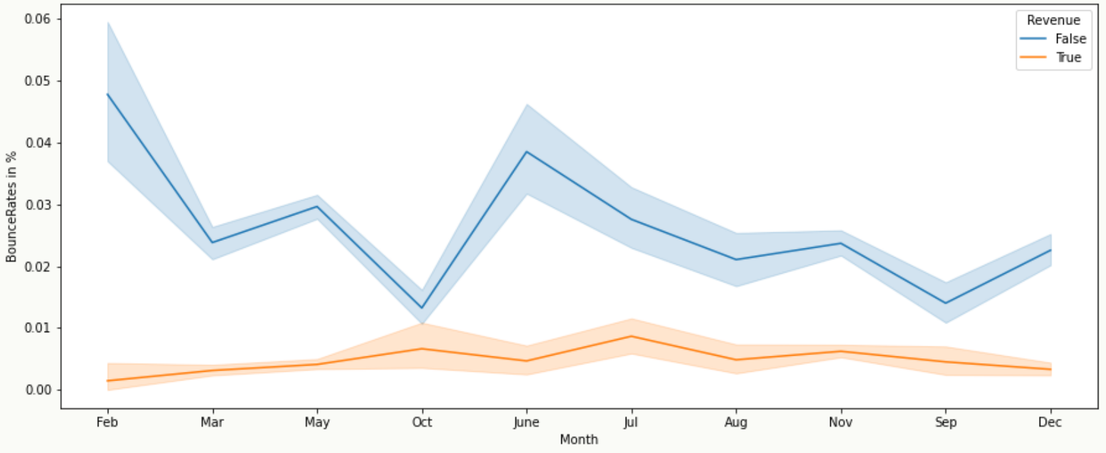
Analyzing revenue distribution by regions, Region 1 emerges as the highest revenue generator, highlighting its significance in driving business outcomes. Following Region 1, Region 3 contributes notably to revenue generation. However, Regions 5 and 8 demonstrate negligible contributions, suggesting that efforts and resources could be reallocated to more fruitful regions in order to optimize revenue generation.

By understanding these patterns and trends, businesses can focus their efforts on OS type 2 users, prioritize traffic types 2, 1, and 3, and concentrate resources in revenue-rich regions such as Region 1 and Region 3. This strategic approach can help maximize revenue and optimize resource allocation for enhanced business outcomes.







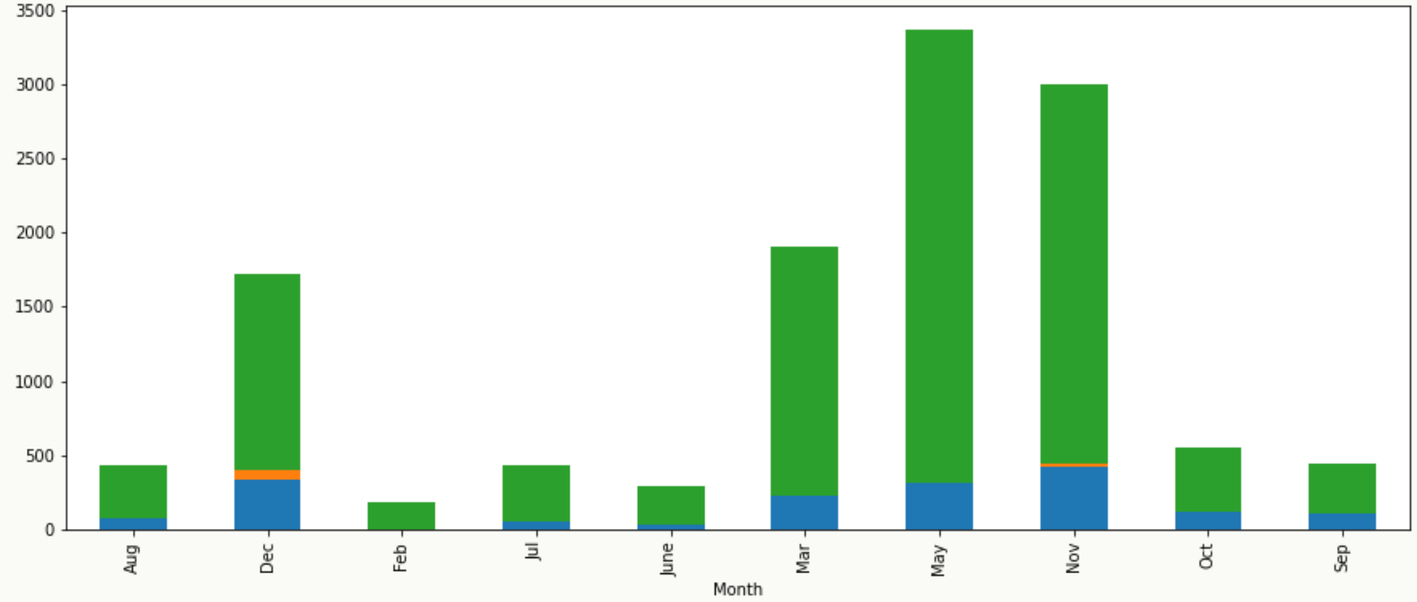


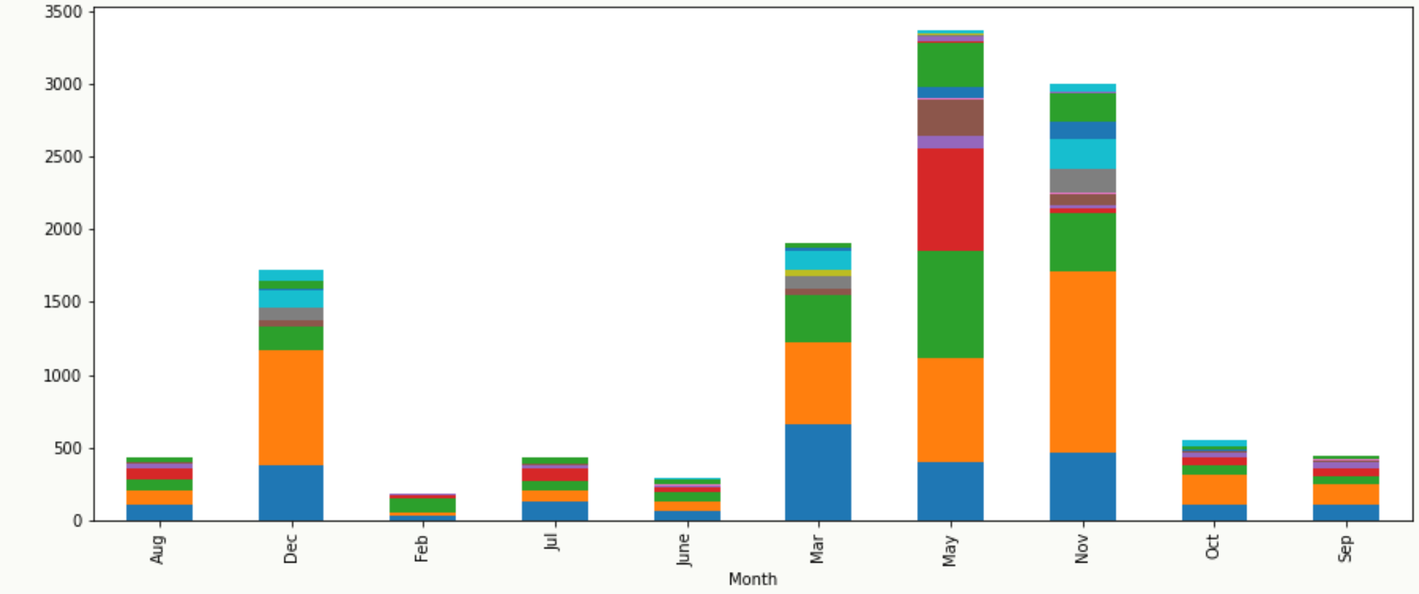
The line chart depicts the website visits and user conversion trends across different months. It reveals that May and November attracted the highest number of visitors, indicating their popularity among users. Interestingly, March, May, and November stand out as months with a higher rate of user conversion, suggesting a potential correlation between these months and successful conversions.

Analyzing the chart, we observe notable patterns in bounce rates and exit rates. In November, both bounce rates and exit rates decrease significantly, indicating positive user engagement and retention. On the other hand, February exhibits the highest bounce and exit rates, implying potential challenges in user retention during that month.

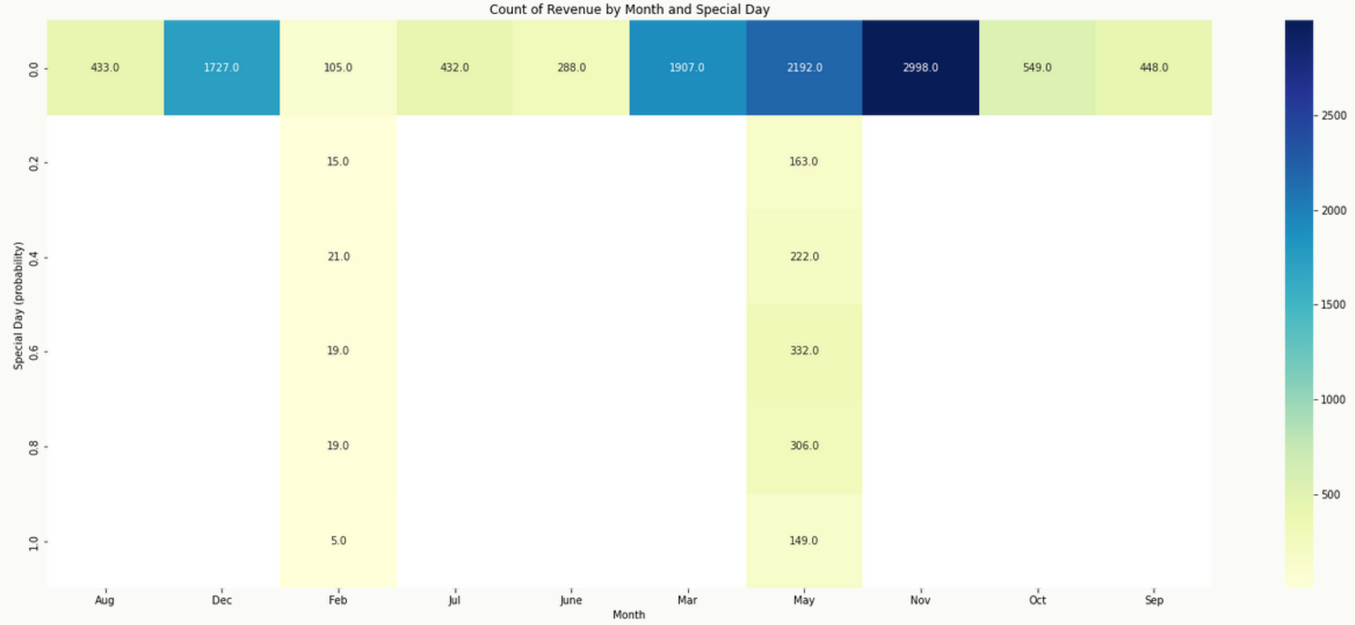
Examining the Page Values, we notice that November records the highest values, indicating that users tend to engage with more valuable content during this month. Conversely, February has the lowest Page Values, suggesting a comparatively lower user engagement and interaction with valuable content.

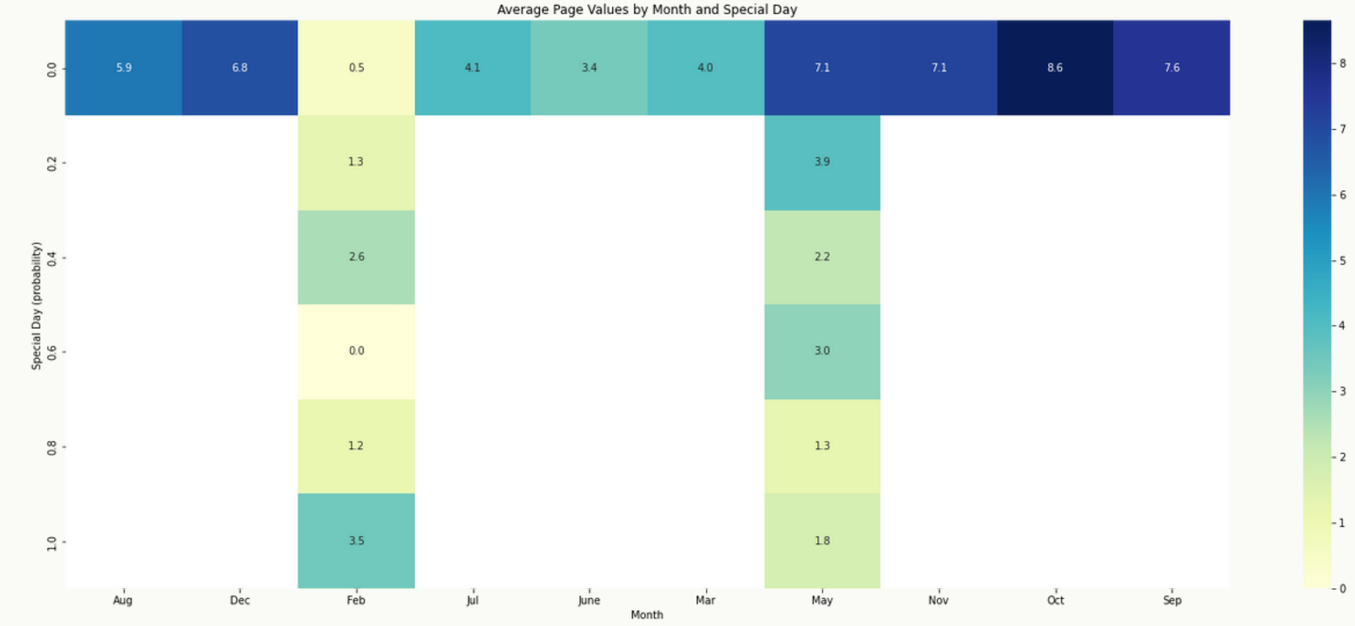
These insights align with our initial hypothesis, supporting the notion that the months of March, May, and November exhibit favorable conditions for user conversion. By considering the observed trends in bounce rates, exit rates, and Page Values, we can further refine our understanding of user behavior during these months and optimize strategies to enhance conversion rates and user engagement.





This stacked bar chart provides insights into the distribution of traffic and visitor types across different months. It reveals that the months of March, May, and November tend to have a higher volume of traffic compared to other months. Interestingly, the month of November stands out with both high traffic and a significant conversion rate. Furthermore, the chart highlights the impact of visitor types on revenue generation. Specifically, returning visitors are found to contribute significantly to the overall revenue. This observation aligns with the high conversion rates observed in November, suggesting a potential connection between the presence of returning visitors and increased conversion during that month. Overall, the chart suggests that the months with higher traffic, especially when accompanied by a higher proportion of returning visitors, have the potential for increased conversion rates and revenue generation. This insight can be valuable in optimizing marketing and promotional strategies to capitalize on the patterns observed in visitor behavior and maximize conversion opportunities.





The first chart displays the count of revenue based on the combination of month and special day. It reveals that November has the highest count of special days, followed by February. This indicates a higher level of user interaction and potentially increased opportunities for revenue generation during these months.

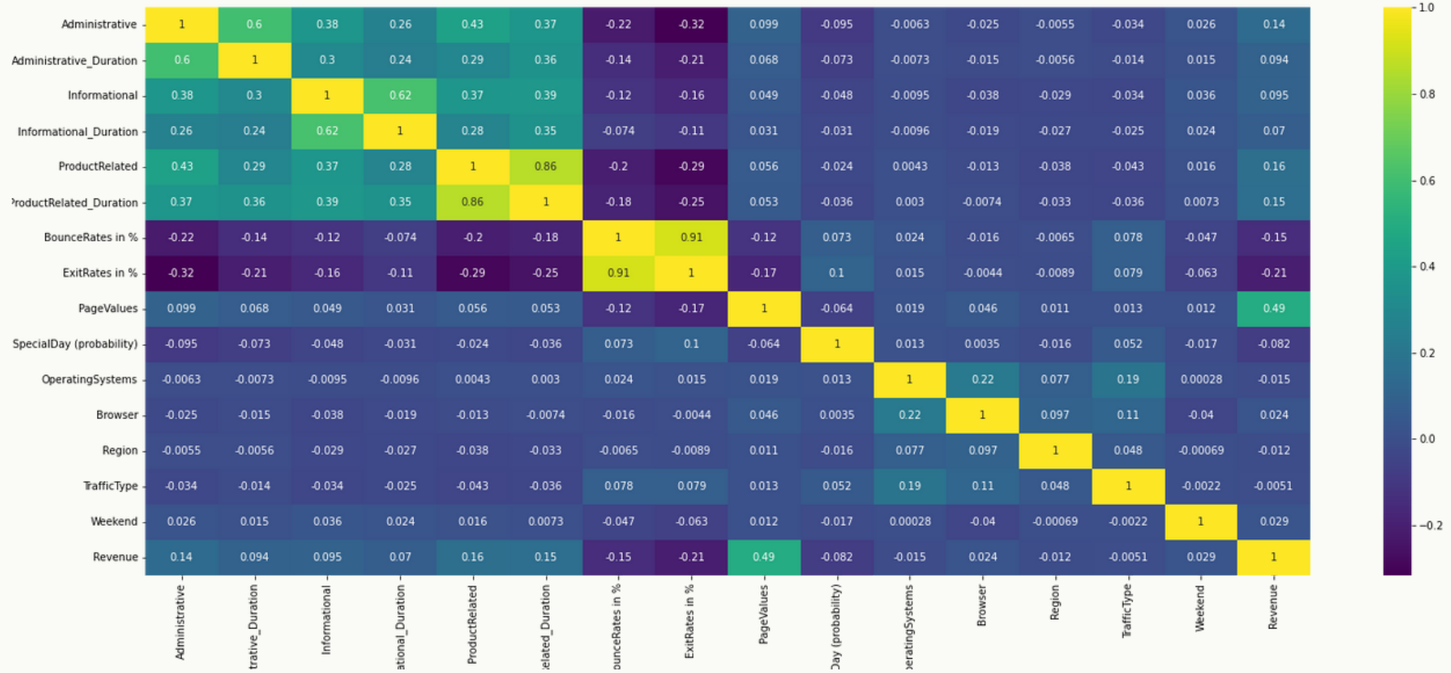
To provide a visual representation of the count data, a heatmap is used. The heatmap shows the count of revenue for each month and special day combination, with color intensity representing the count value. The annotation on the heatmap provides the exact count values for better interpretation.

The second chart focuses on the average page values by month and special day. Page values reflect the monetary value assigned to web pages based on user interactions and conversions. The heatmap showcases the average page values for each month and special day combination. The color gradient in the heatmap represents the average page value, with higher values indicated by darker shades.

Both charts provide valuable insights into the relationship between month, special day, revenue count, and average page values. They offer a visual understanding of how these factors contribute to revenue generation and the potential impact of different time periods on user behavior and engagement.

Overall, these charts help analyze the patterns and dynamics of revenue generation in relation to month and special day, providing insights that can inform strategic decision-making and optimization of marketing efforts.

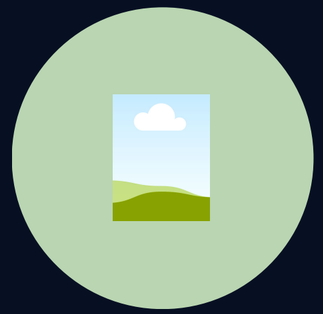
In this particular heatmap, each cell represents the correlation coefficient between two variables. The correlation coefficient measures the strength and direction of the linear relationship between two variables. The values range from -1 to 1, where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation.
The colors in the heatmap represent the magnitude of the correlation. Darker colors, such as dark blue or dark green, indicate a strong correlation (either positive or negative), while lighter colors indicate a weaker correlation or no correlation.
The annotation within each cell of the heatmap displays the correlation coefficient value, providing a numerical representation of the correlation strength. The variable page value is having a high positive correlation coefficient with our target variable which will be an important predictor of our target variable. We see that variables like Exit Rates and Bounce Rates, Product Related, and Duration are having a high correlation coefficient with itself indicating the presence of multicollinearity in the dataset.



Model Selection

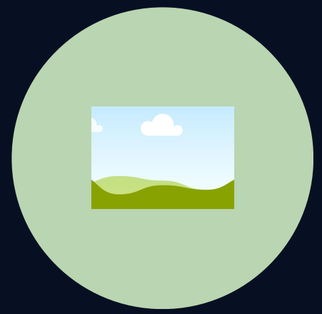
Overview of Model Selection

Explain the process of selecting the best model for predicting purchase conversion in e-commerce



Types of Models

Discuss the different types of models available for machine learning in e-commerce



Model Evaluation

Describe the evaluation process for determining the best model for predicting purchase conversion

Model selection is an important step in machine learning in e-commerce, as it helps to determine the best model for predicting purchase conversion.

Data Preprocessing

Categorical Variable Handling: Various techniques were employed to handle categorical variables, including frequency encoding, label encoding, binary encoding, and one-hot encoding. Among these, one-hot encoding yielded the best performance in terms of predictive accuracy and model fit.
Low Variance Column Handling: To address columns with low variance, a strategy was implemented to aggregate features with a low count into a single feature labeled as 'Others'. This consolidation helps to reduce noise and improve the overall stability of the model.
Continuous Feature Binning and Transformation: To tackle skewness in continuous features, binning techniques were applied to group data into intervals. Additionally, transformations such as logarithmic, square, and Yeo-Johnson transformations were utilized to achieve a more normal distribution shape and mitigate the impact of extreme values.
Outlier Handling: Outliers, which can significantly skew results, were addressed by clipping continuous features. This involved setting a predetermined threshold to cap extreme values and bring them within a more reasonable range.
Feature Engineering: Several new features were created to enhance the predictive power of the model. These included total duration and total session, which capture the cumulative duration and session count for website interactions. Additionally, the percentage of views per page, average duration per page, and interaction terms involving page value, bounce rate, and duration of web pages were generated. These engineered features aim to provide deeper insights and capture more complex relationships within the data.
By implementing these techniques and feature engineering strategies, the data preprocessing stage enhances the quality and relevance of the input data, contributing to improved model performance and more accurate predictions.

During the feature selection process, Recursive Feature Elimination (RFE) was employed in combination with built-in feature selection techniques offered by Random Forest and XGBoost models. This approach helps identify the most relevant features for the predictive task.

Multiple models were evaluated, including Baseline Logistic Regression, Logistic Regression with Principal Component Analysis (PCA), Random Forest, XGBoost, and Light GBM. Among these models, XGBoost demonstrated superior performance compared to Random Forest when considering the baseline metrics.

Based on the evaluation results, XGBoost was selected as the preferred model for further analysis and modeling tasks. Its ability to provide a stronger baseline and improved predictive power made it the optimal choice for the given dataset and problem at hand.

I used Stratified K fold split for splitting the data.

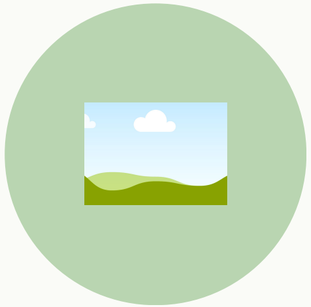
Hyperparameter Tuning

What is Hyperparameter Tuning?

Hyperparameter tuning is the process of optimizing the hyperparameters of a model to improve its performance.

Why is Hyperparameter Tuning Important?

Hyperparameter tuning is important for improving the accuracy and performance of a machine learning model.



How to Perform Hyperparameter Tuning?

Hyperparameter tuning can be done using grid search, random search, or Bayesian optimization.

Hyperparameter tuning is an important step in the machine learning process that can help improve the accuracy and performance of a model.

For hyperparameter tuning, Optuna was utilized. The tuning process focused on optimizing the learning rate, maximum depth, and maximum number of leaves.
The hyperparameter search space was defined using Optuna's param_space dictionary, which included parameters such as 'eta' (learning rate), 'max_depth' (maximum depth), and 'max_leaves' (maximum number of leaves). These parameters were constrained within specific ranges using hp. uniform to ensure effective tuning. I used SHAP and the yellow brick library for model evaluation and interpretation. The results I achieved are a 73% recall score. I focused on recall because we are mostly interested if the user generates the revenue.