Online Shoppers Intentions

Overview

- About
- Business Value
- Exploratory Data Analysis
- Feature Engineering
- Model
- Results
- Conclusion

About

The dataset consists of 10 numerical and 8 categorical attributes. The 'Revenue' attribute can be used as the class label. Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.

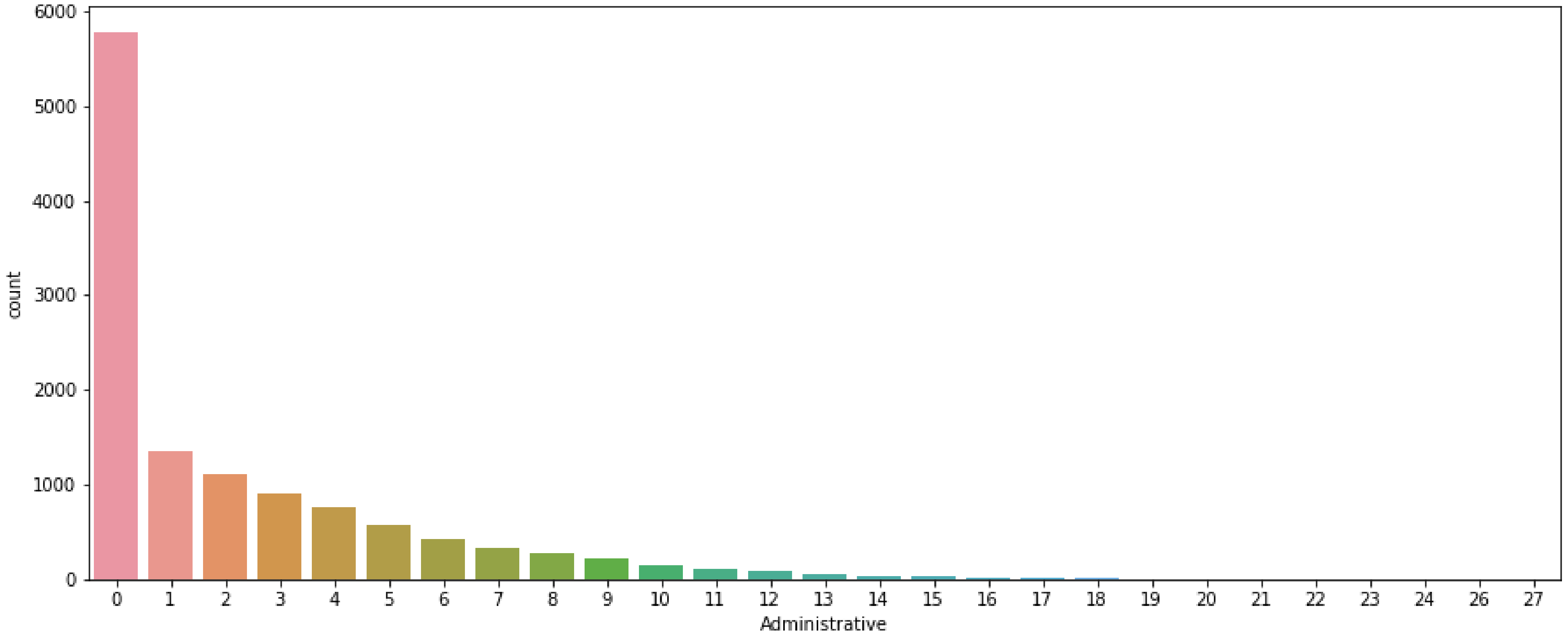
The value of the "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of the "Exit Rate" feature for a specific web page is calculated as the percentage that was the last in the session for all pageviews to the page. The "Page Value" feature represents the average value for a web page a user visited before completing an e-commerce transaction. The "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. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and the delivery date. The dataset also includes the operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is a weekend, and the month of the year.

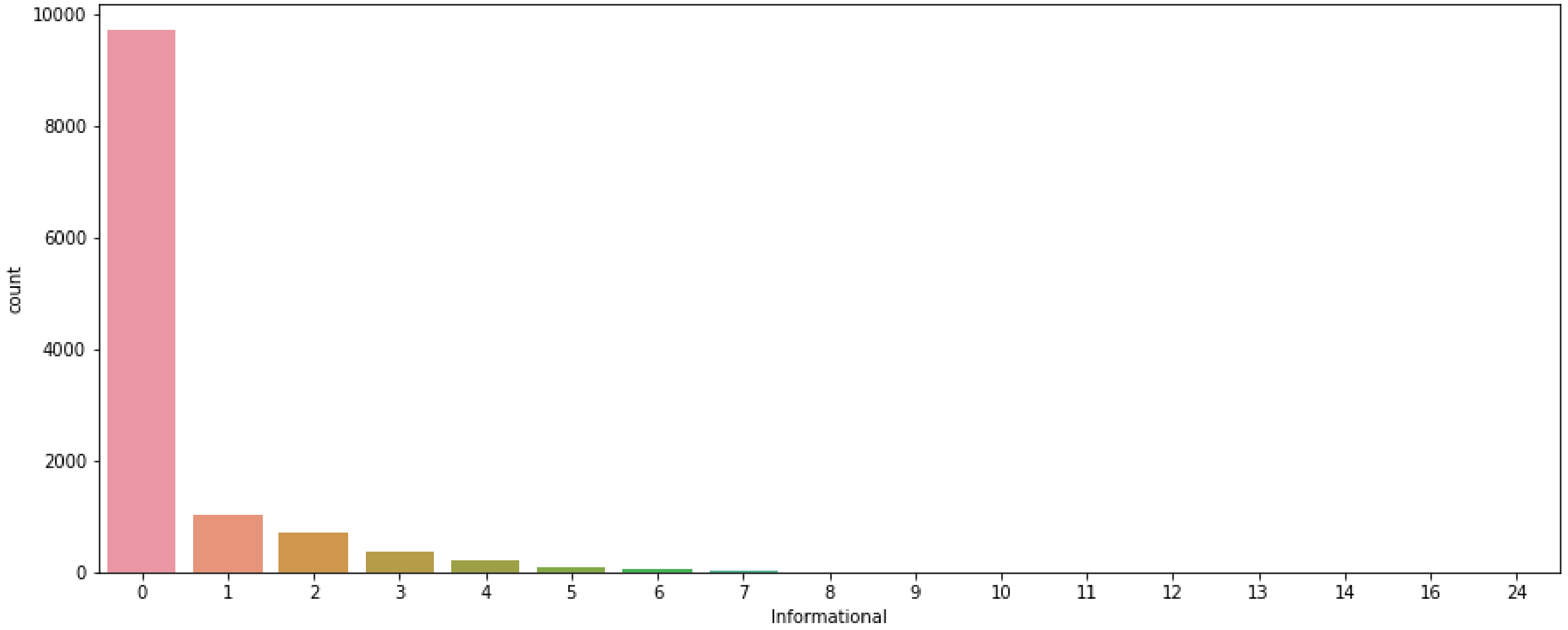
Business Value

The main objective revolved around the identification of key metrics which contributes the most towards predicting a shopper's behavior and to suggest prioritized critical recommendations and performance improvements on the same. Revenue is the attribute of interest which identifies if a purchase was made or not.

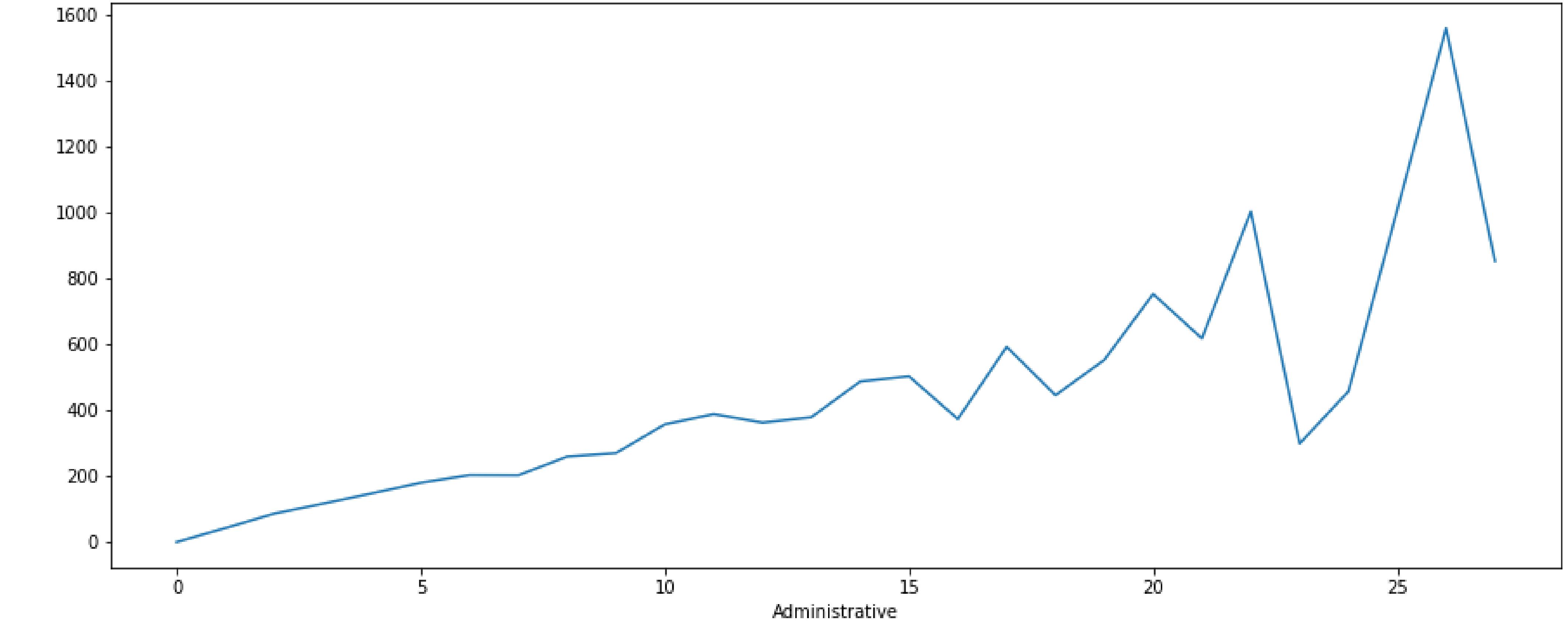
EDA

"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.



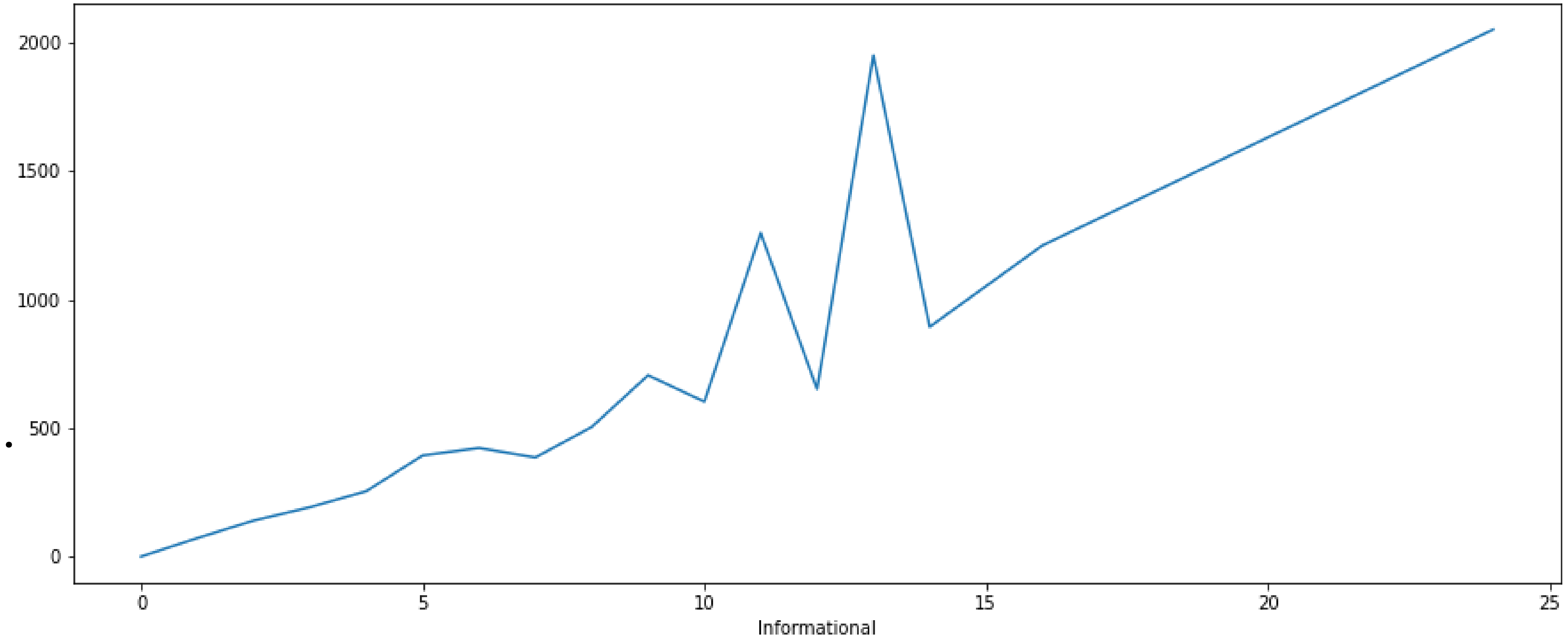


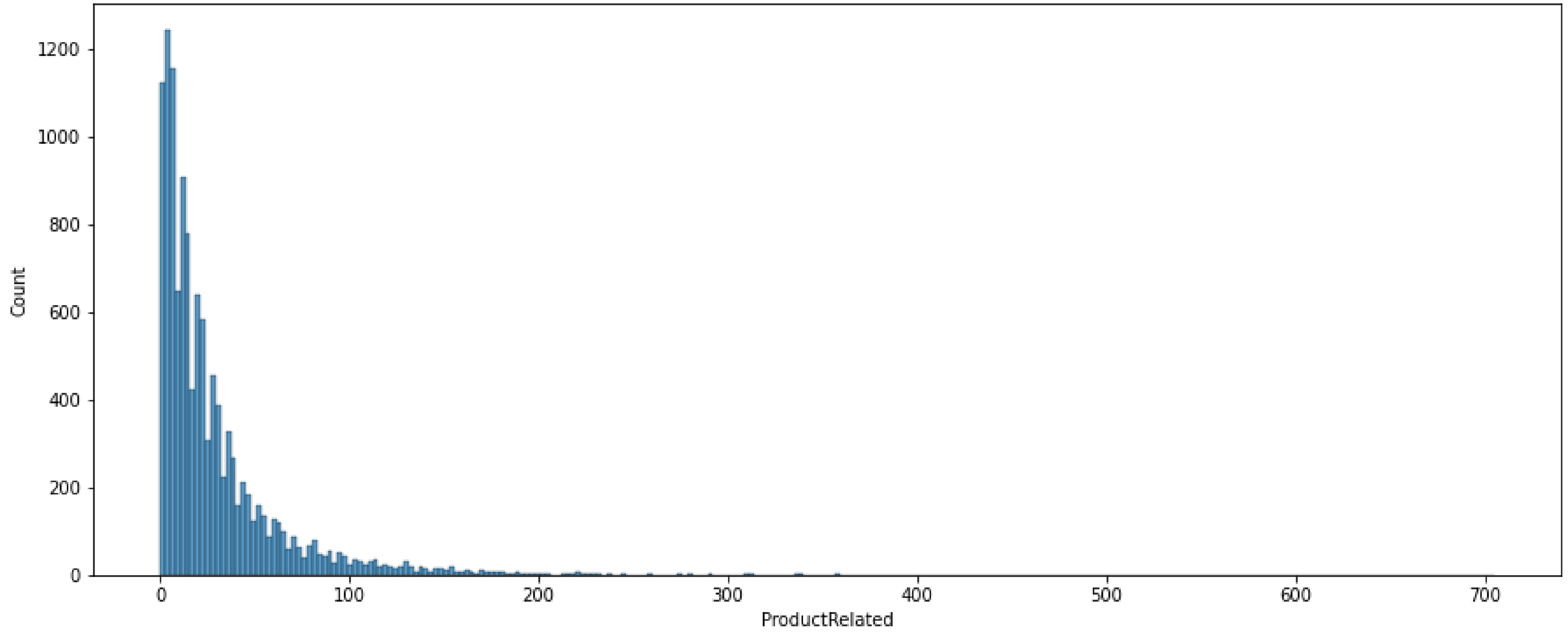
In the plot, the most frequently visited page is 0 and the least frequent is 26 for the administrative page, the most frequently visited page for the informational page is 0 and the least frequent is 13. This indicates user interaction with each pages, with 0 being the most visited page for both Administrative and Informational.



As we can see from the graph, administrative page 0 has an average duration of 0 visits, which may indicate that the user does not find the content useful or informative. The duration of the user's visit increases as he navigates to more pages, which indicates that he finds the content to be more useful. The average duration of page 26 is 1561.717567 seconds.

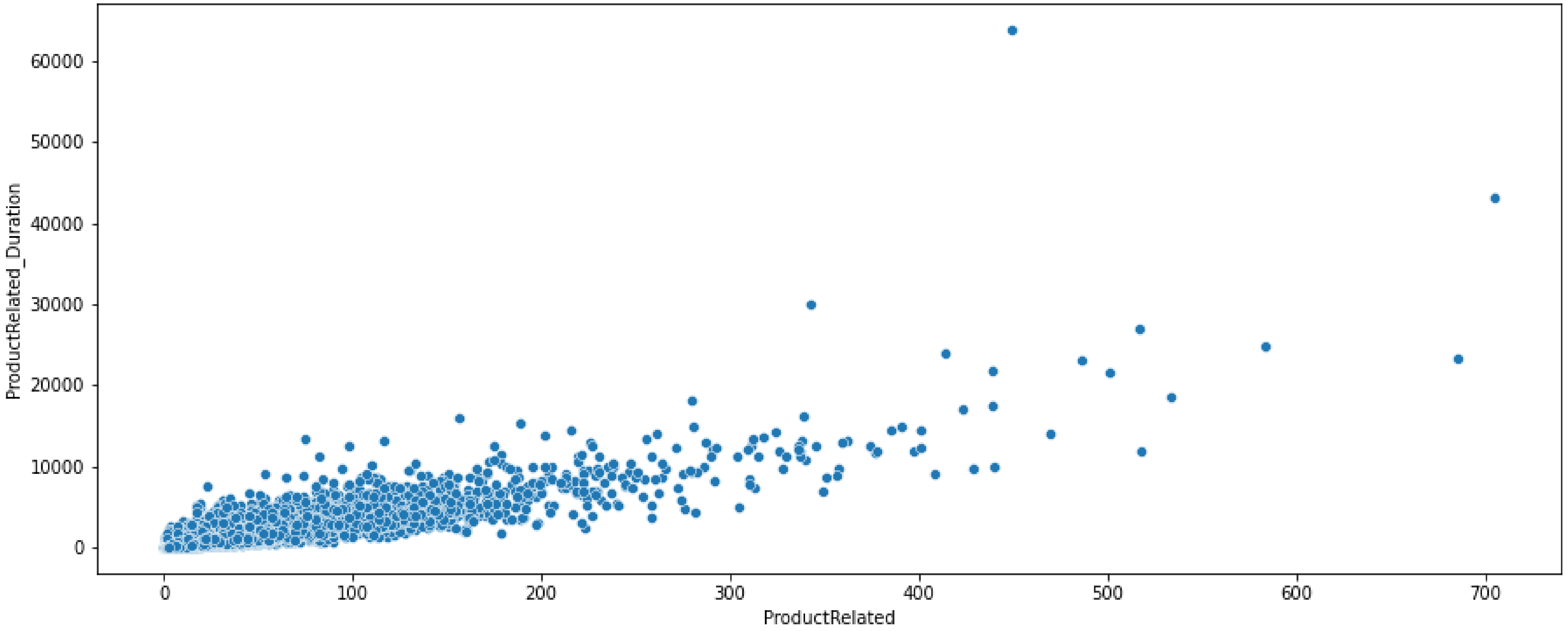
Informative page 0 has an average duration of 0 visits, which is similar to the administrative page, which could indicate that the user is not finding the particular content useful. The duration of the user's visit increases as he navigates to more pages. The average duration of page 24 is 2050.433333 seconds.





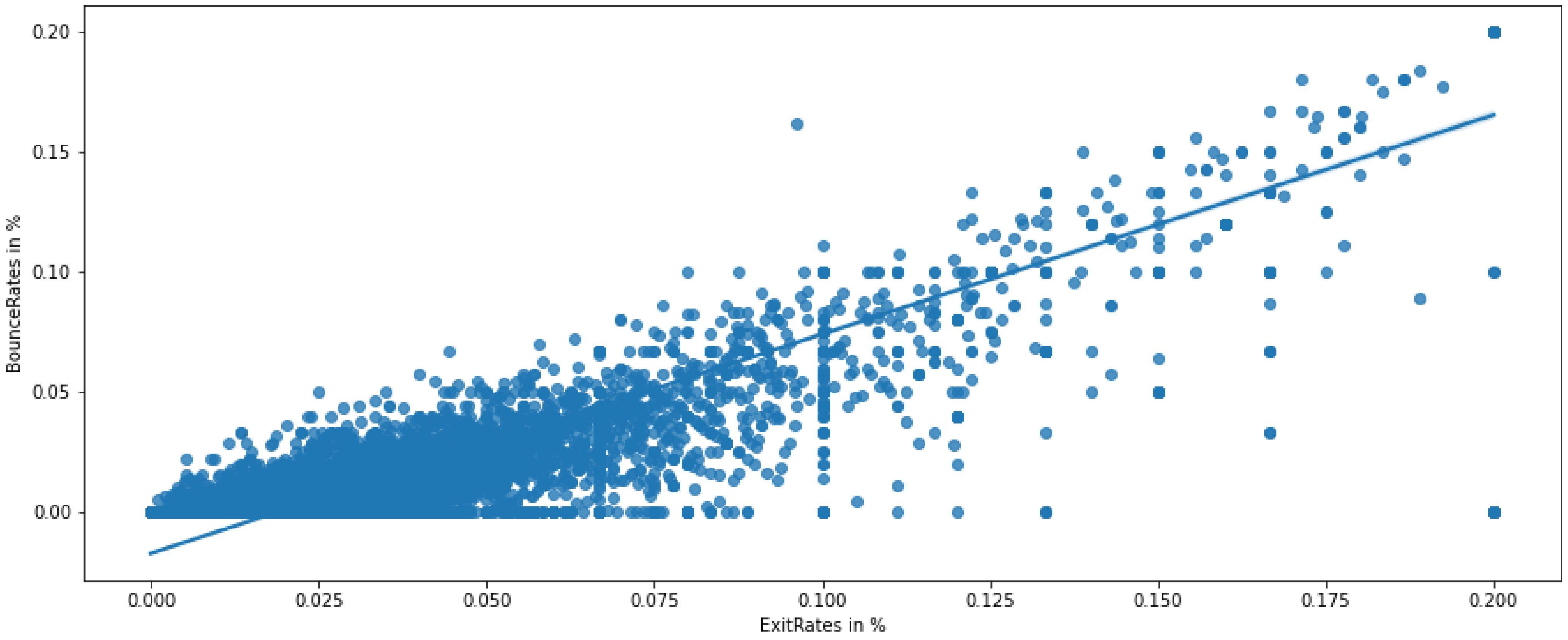
The distribution of product-related pages is right-skewed, with 0 being the most visited page and as we move the page the frequency of users visiting the page also decreases.

A high frequency of visits to a page with a low duration spent could indicate that users are quickly navigating away from the page because they are not finding the content to be relevant or engaging. On the other hand, a low frequency of visits with a high duration could indicate that the page contains content that users find valuable and are spending more time engaging with.

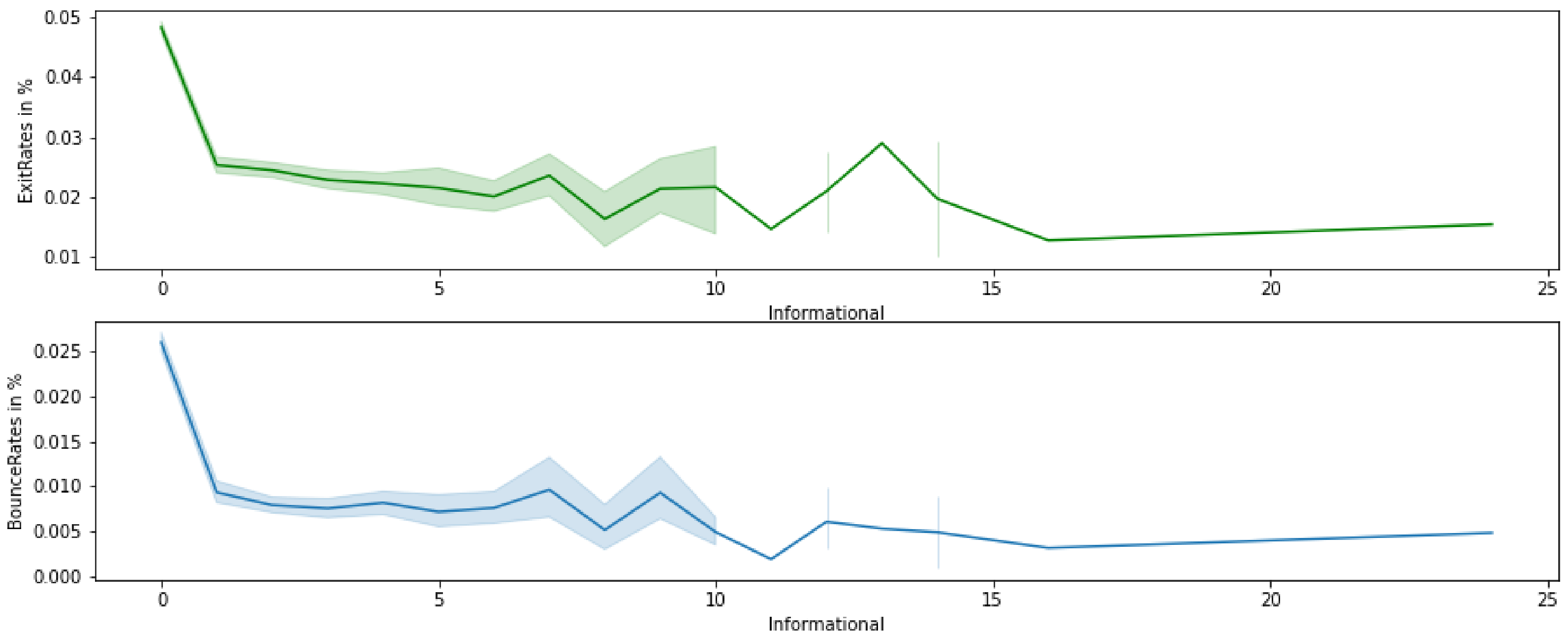


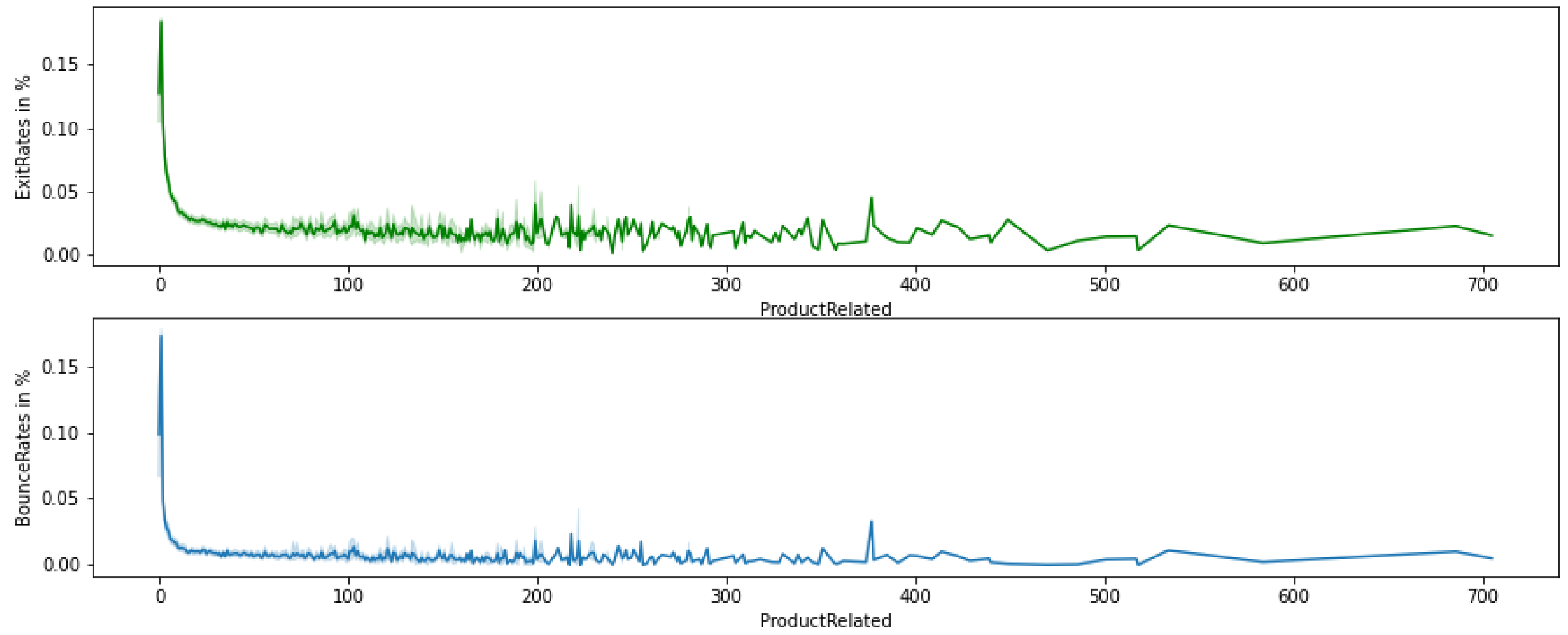
Bounce rate is a metric that measures the percentage of visitors to a website who navigate away from the site after viewing only one page. A high bounce rate means that a high percentage of visitors leave the website after only viewing a single page, which may suggest that the content on the website does not meet their needs or expectations.

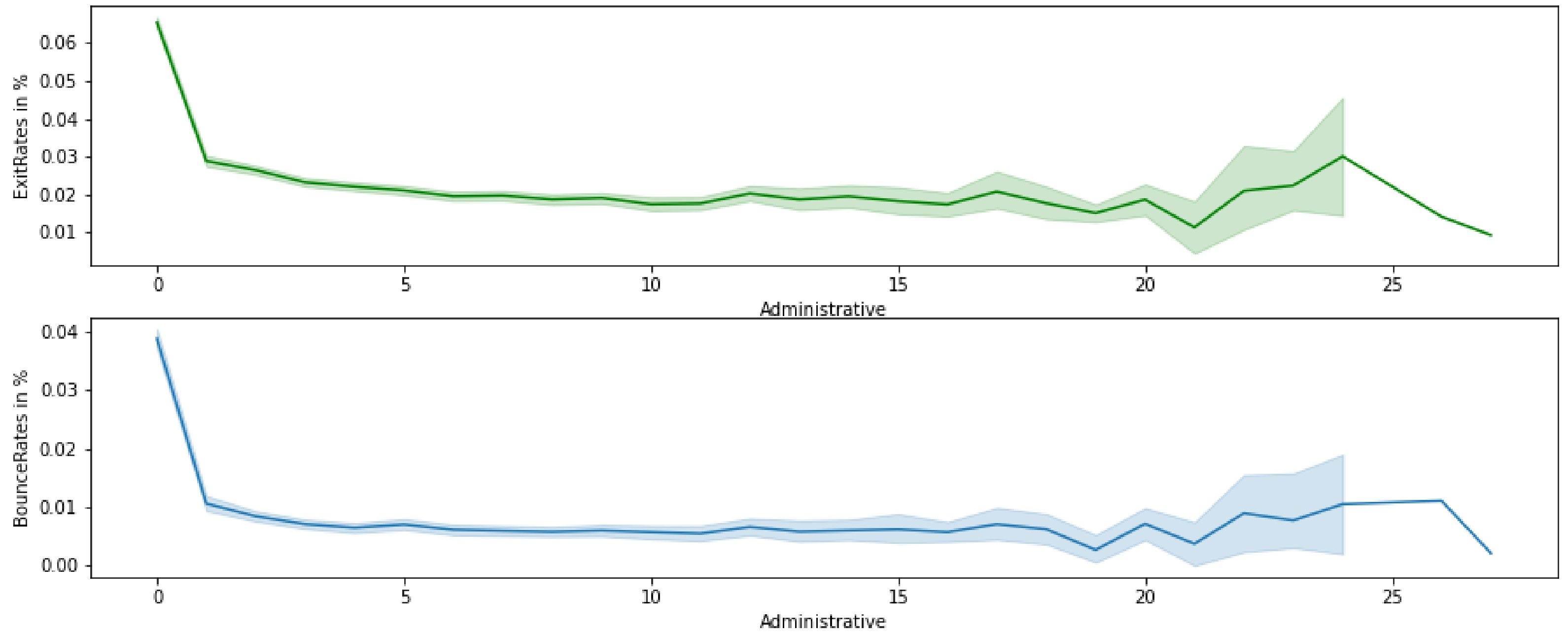
Exit rate is a metric that measures the percentage of visitors who leave a website from a specific page, rather than continuing to view other pages on the site. A high exit rate means that a high percentage of visitors leave the website after visiting multiple pages, which may suggest that the website has usability or navigation issues or that the content is not compelling enough to keep visitors engaged.



As the bounce rate increases exit rate also increases, with a Pearson correlation of 0.913. This indicates a positive linear relationship between the two variables.

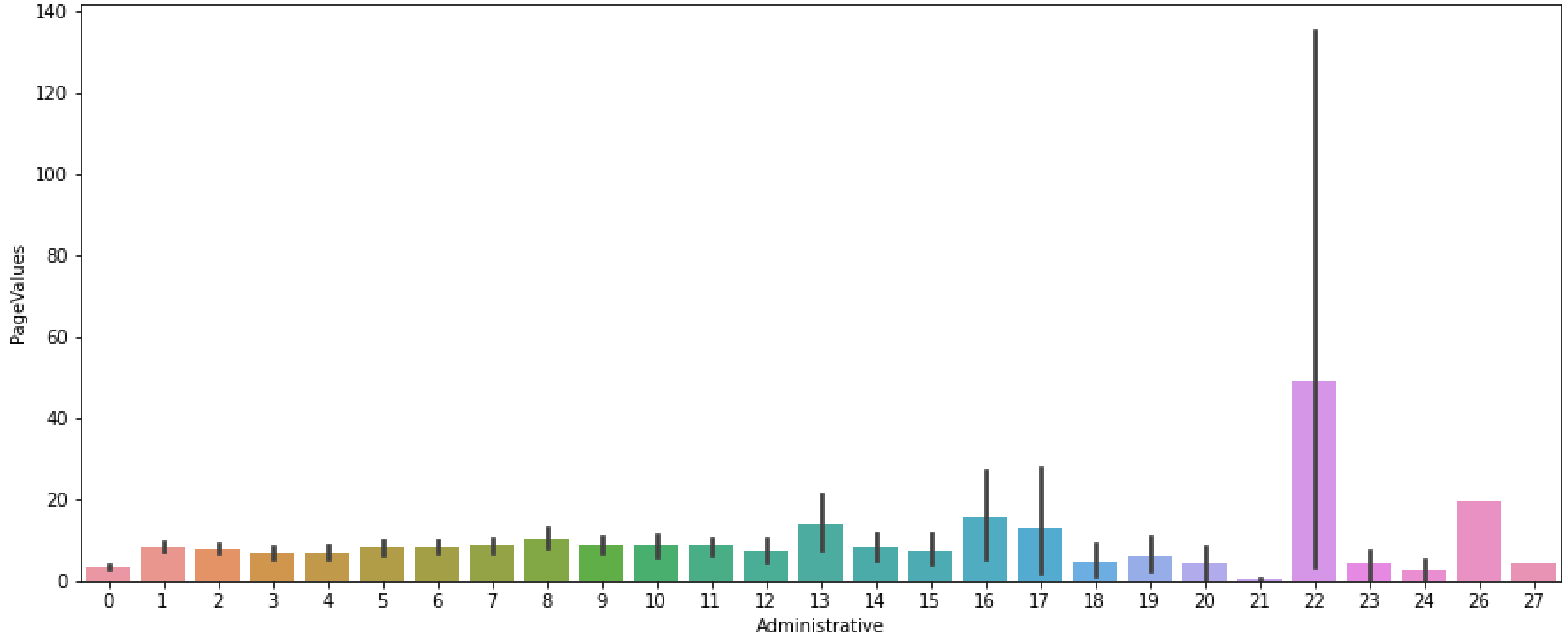




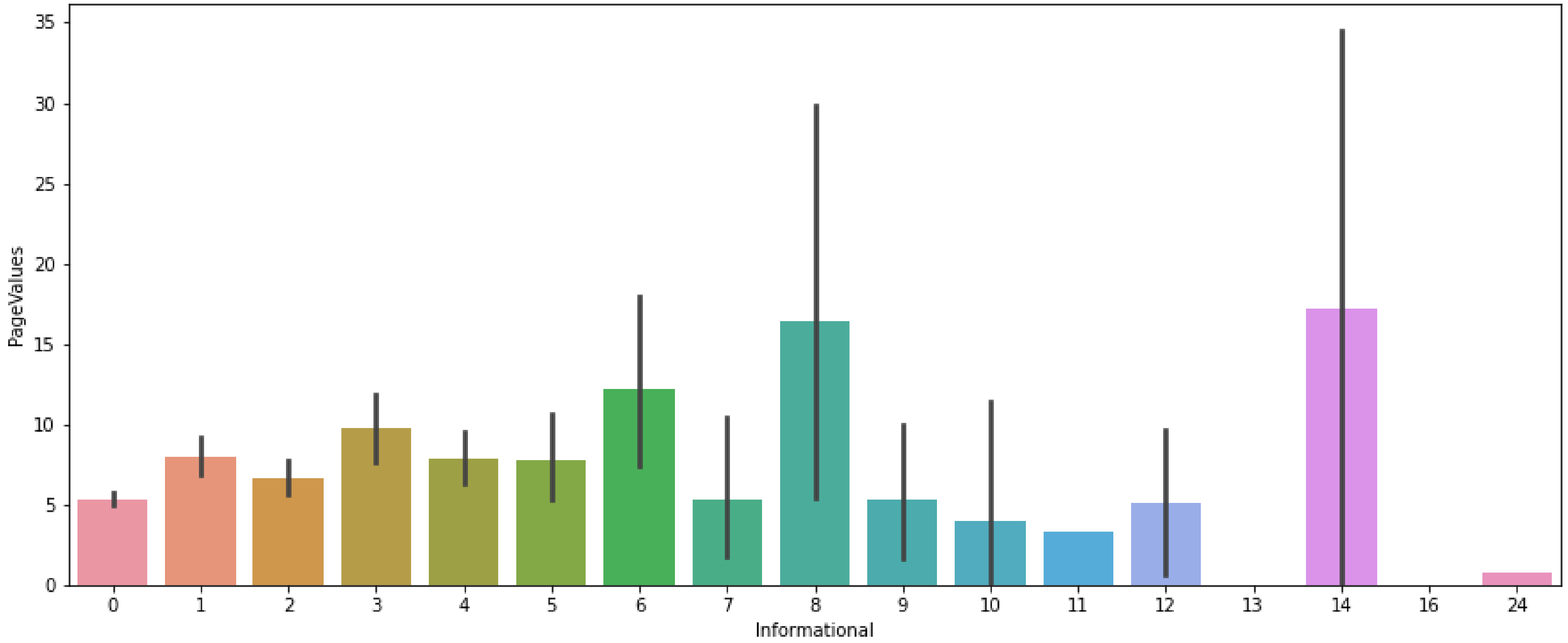


The data suggests that as the number of ProductRelated pages visited increases, the Exit Rates tend to decrease, the Bounce Rates tend to decrease, and the ProductRelated Duration tends to increase.

This pattern of decreasing Exit Rates and Bounce Rates as the number of pages visited increases can indicate that visitors are becoming more engaged with the content on the website as they navigate through more pages. It could mean that visitors are finding more relevant and interesting content as they visit more pages, and as a result, they spend more time on the website and are less likely to leave the website (as seen by the decrease in Exit Rates) and less likely to leave the website after visiting only one page (as seen by the decrease in Bounce Rates).

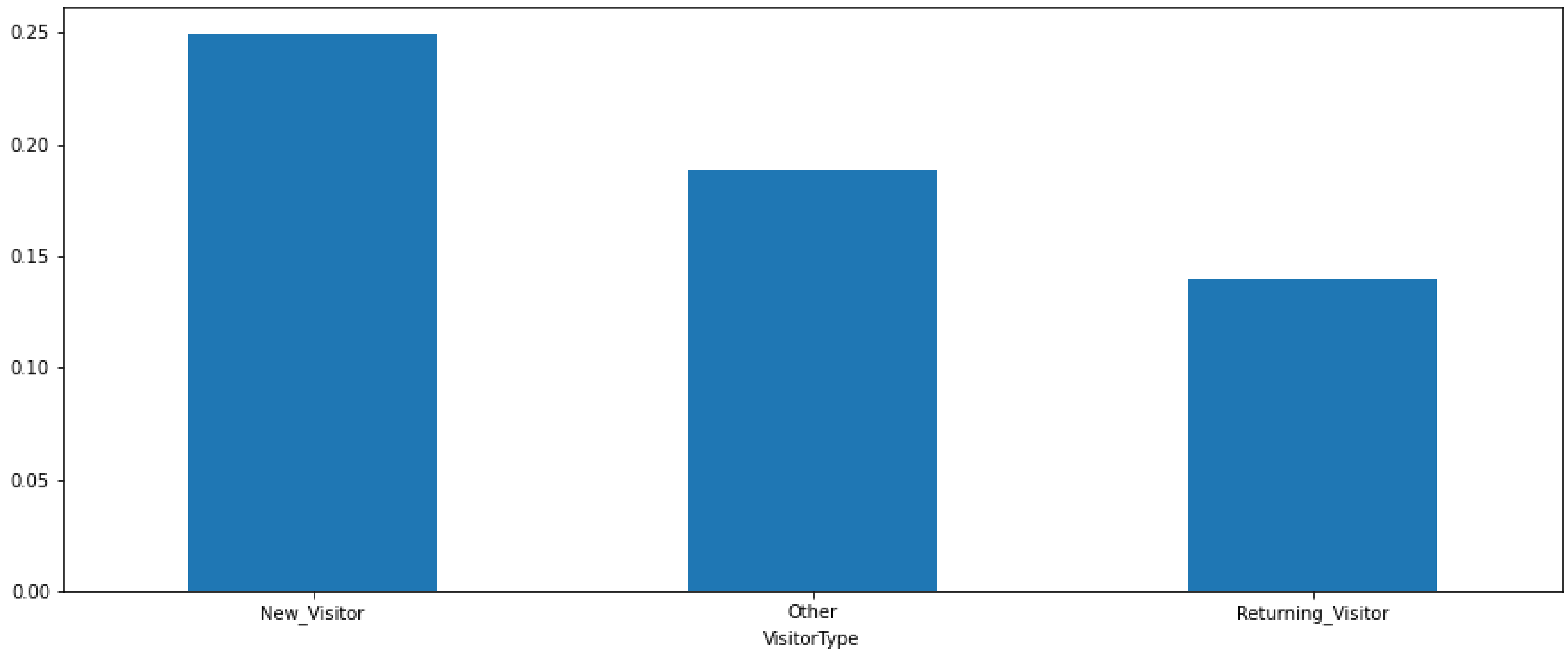


Here we can see the barplot of PageValues with the Administrative Page. From the chart, we can see the high and low page values for a particular administrative page. High page value for a page indicating the importance of that particular page in generating revenue or conversion We can see page 22 with the highest page values and 21 with the lowest.

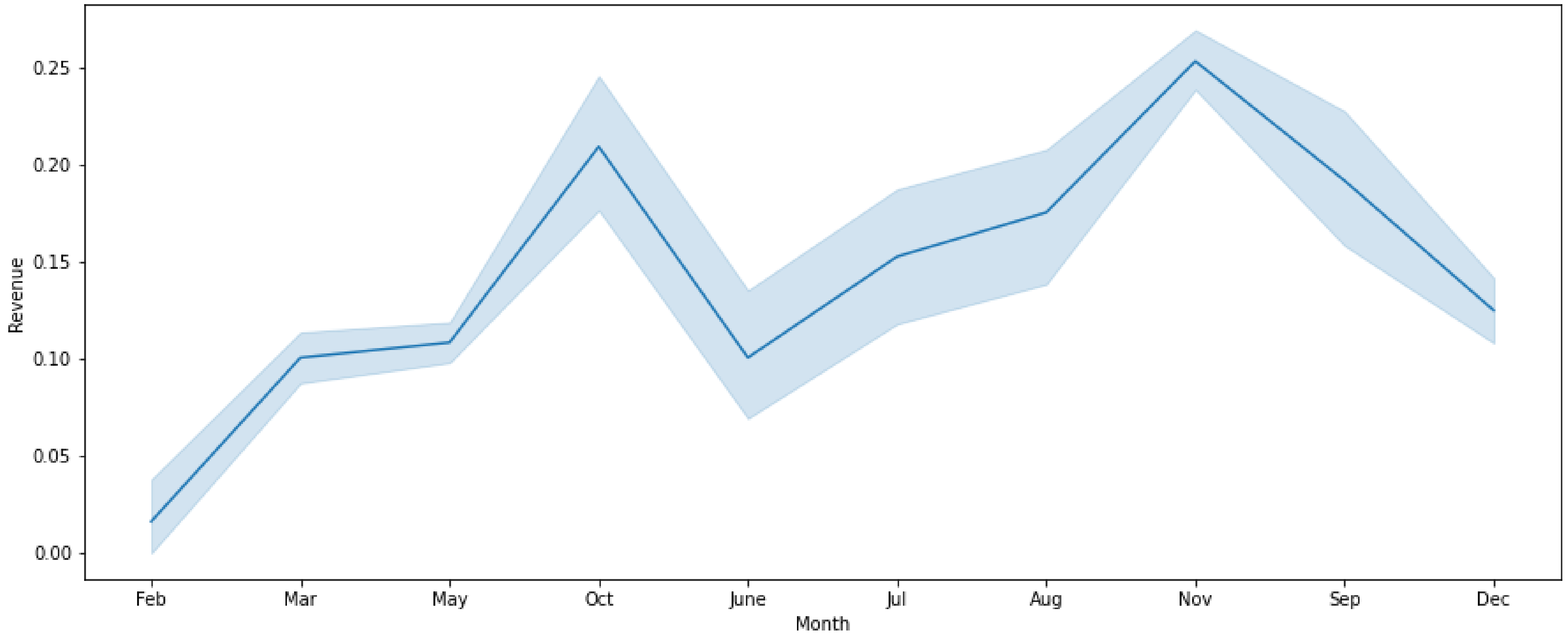


For the Informational Page, we see page 14 giving the highest page values, with 13 and 16 being close to 0.

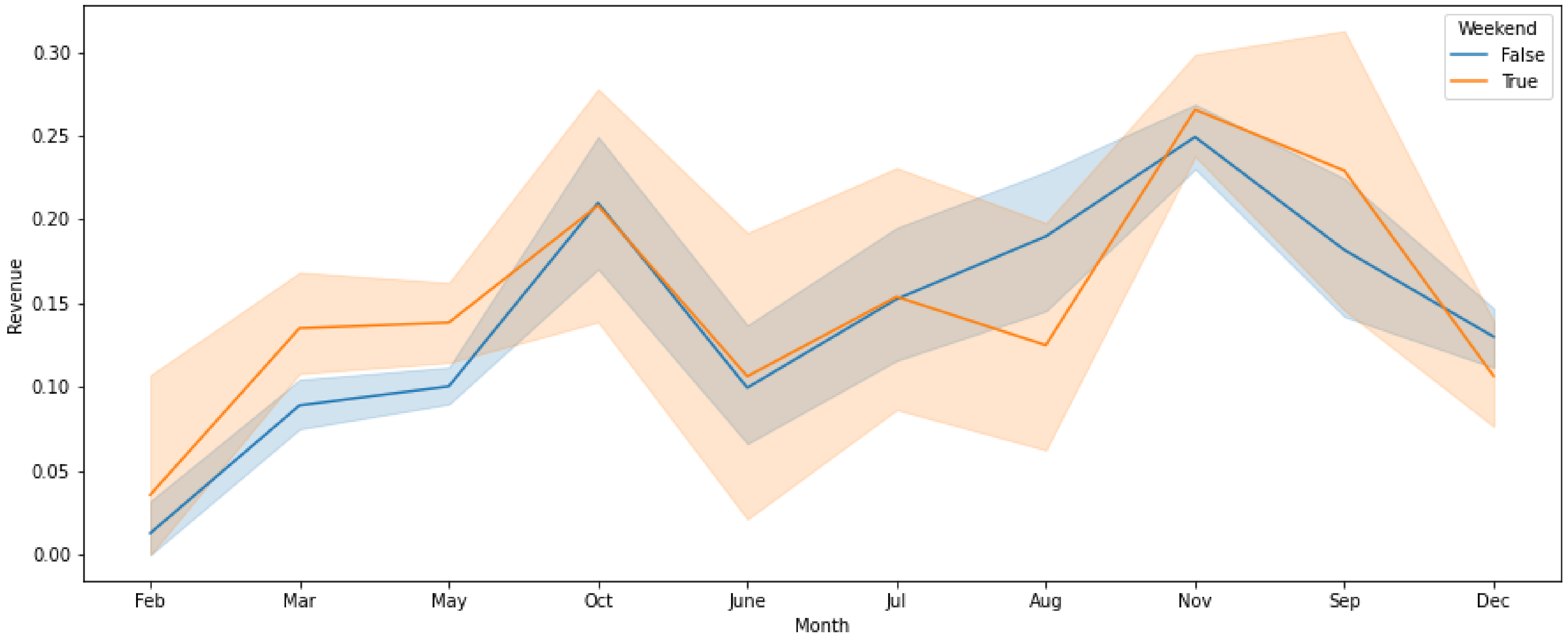
Pages with low page values are underperforming and can be optimized for better revenue generation.



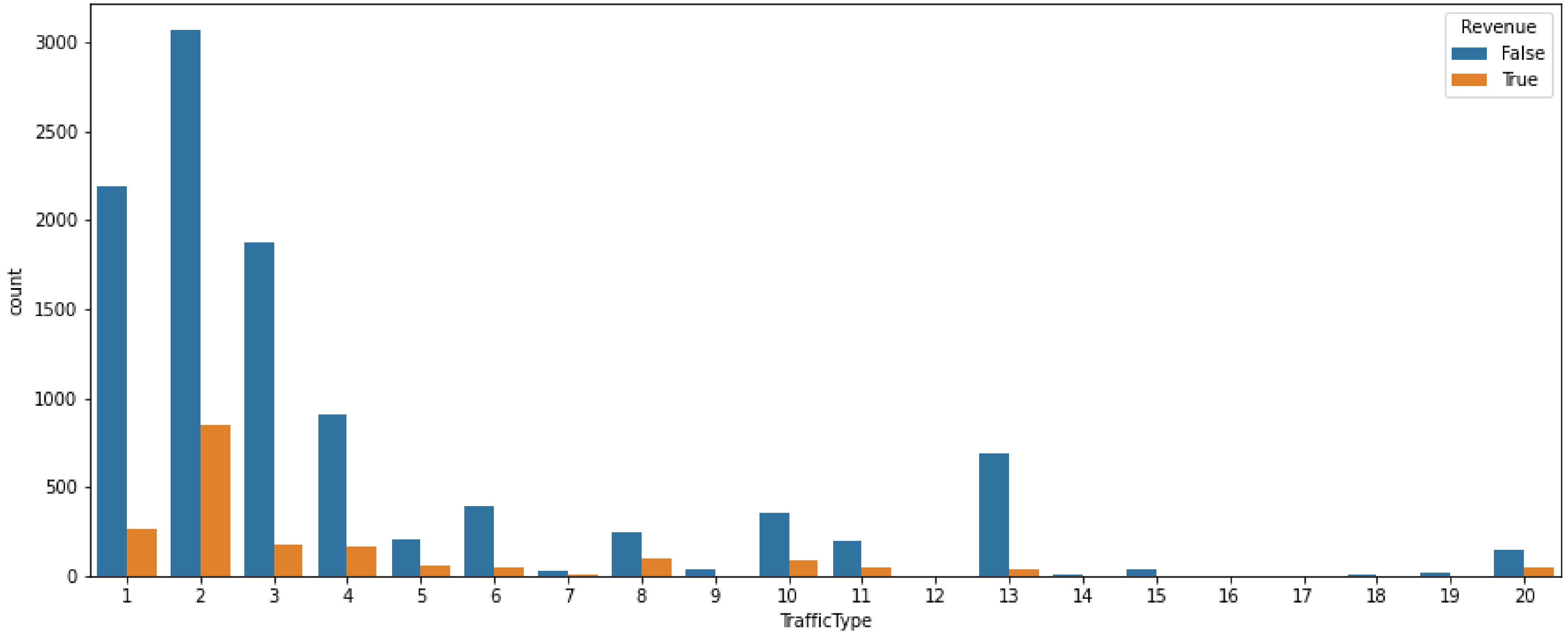
This chart is a barplot showing the relationship between the VisitorType and Revenue. The y-axis represents the average Revenue by each VisitorType and the x-axis represents the different VisitorType. The chart provides a visual representation of the average revenue generated for each visitor type and how they vary based on the level of Revenue. From the chart, it is clear that New Visitors generate the highest revenue followed by Others and Returning Visitors.



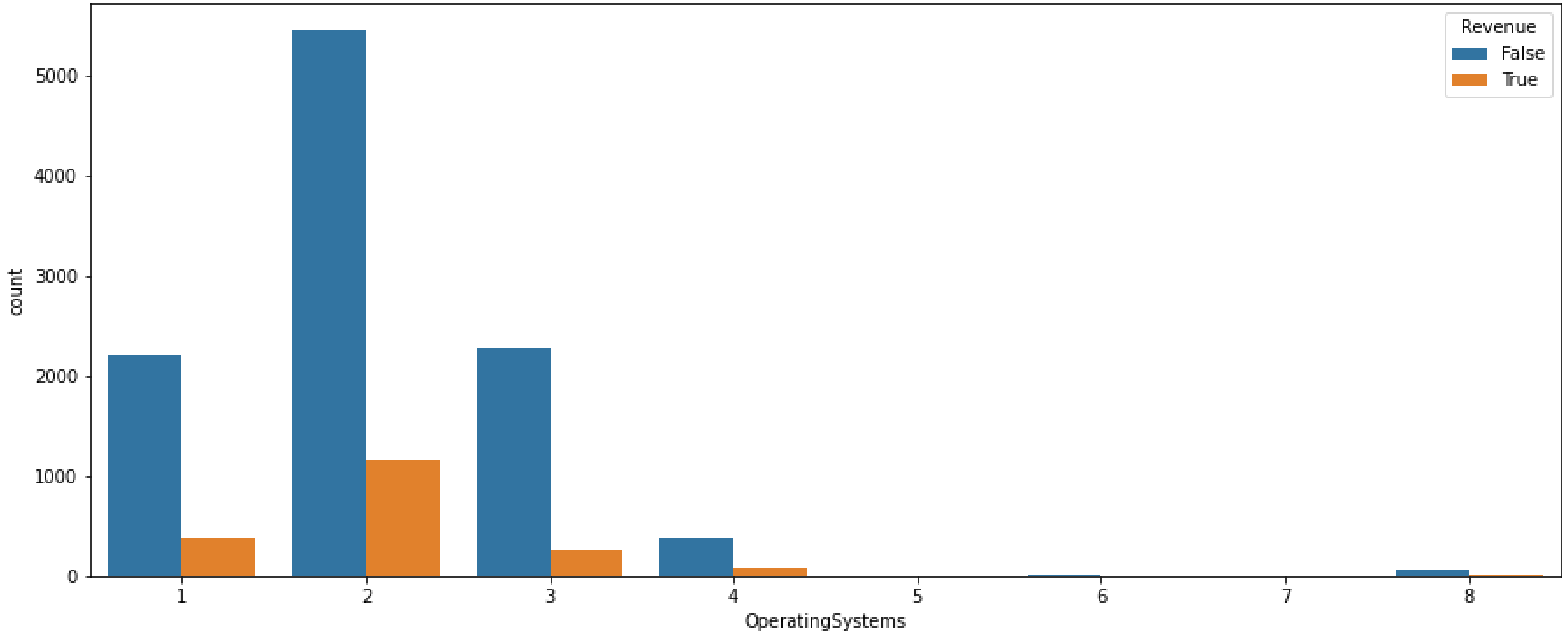
This plot shows the monthly revenue generation over time. We can see spikes in the month of October and an upward trend from June to November followed by a downtrend. This indicates the impact of how a user interaction each month affects revenue.

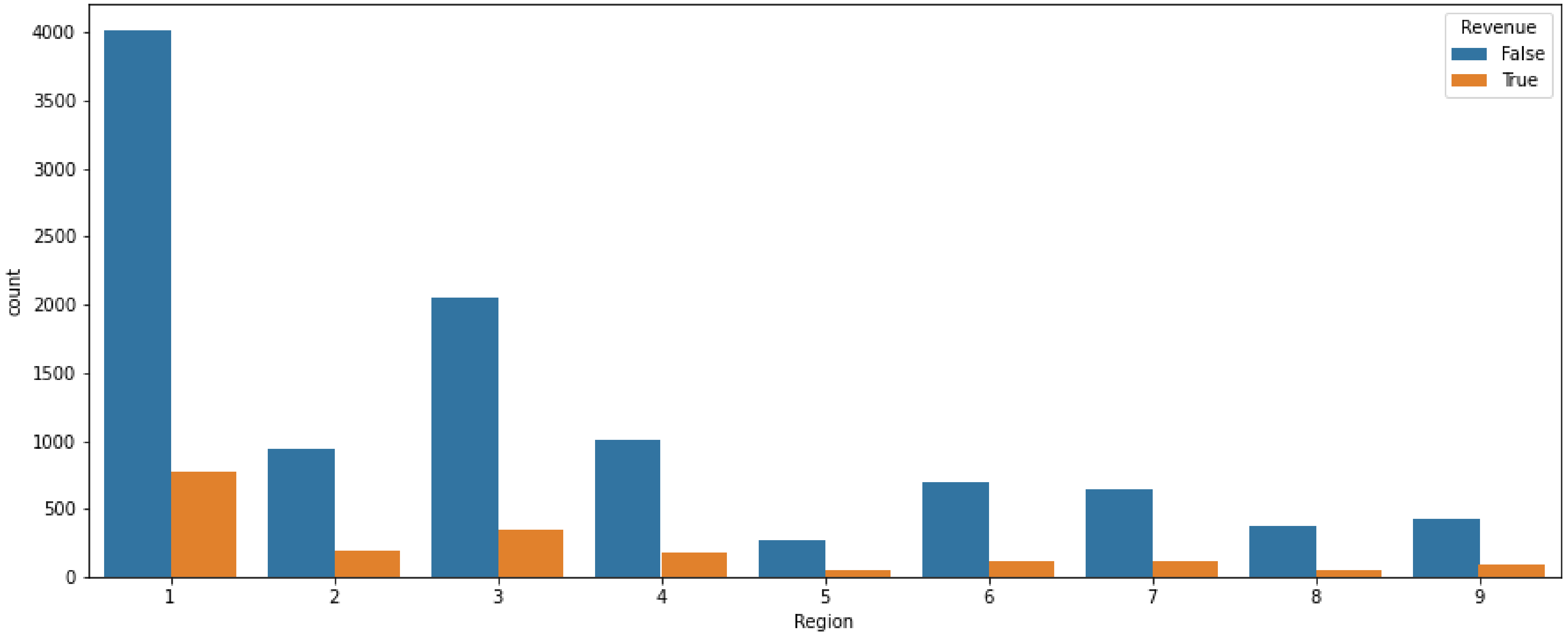


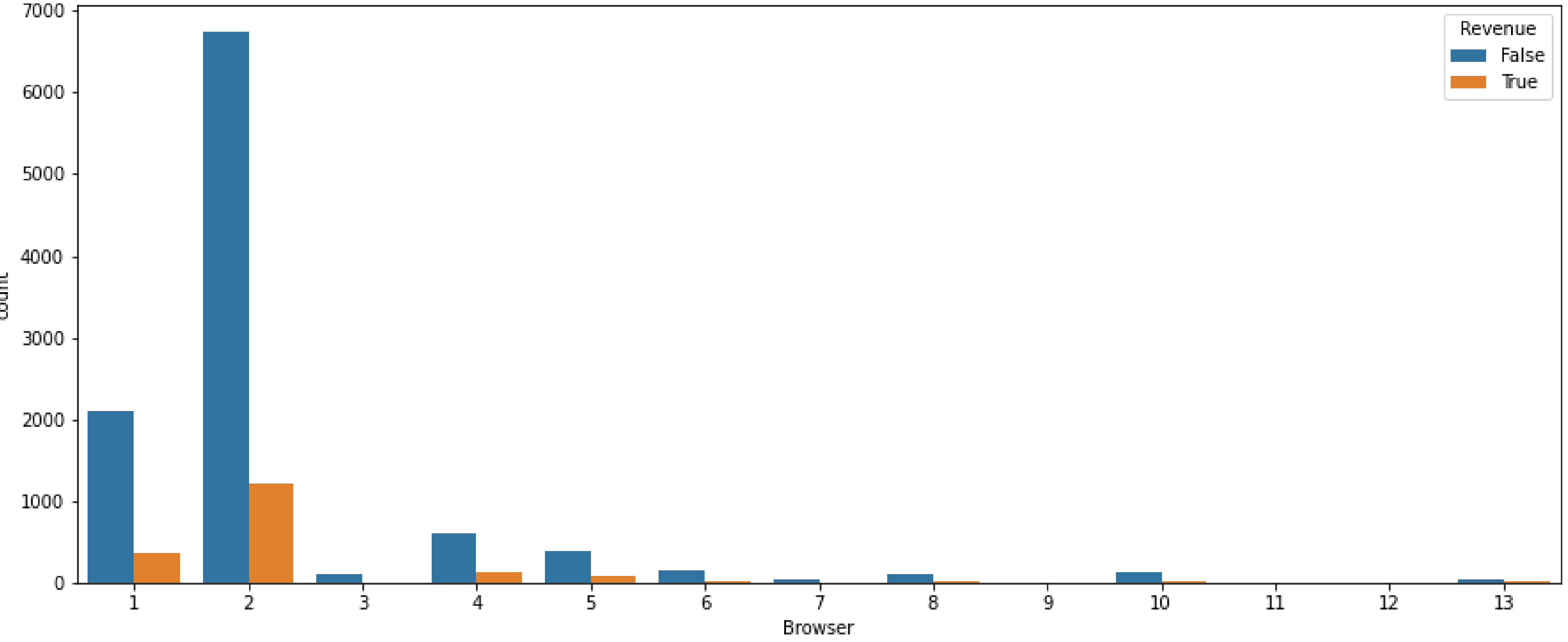
This chart is a line plot showing the relationship between the Months and Revenue. The y-axis represents the Revenue and the x-axis represents the Months. The charts are split by Weekend, with different colors representing the different levels of Weekend. The chart provides a visual representation of the Revenue generated monthly and how they vary based on the level of Weekend.



we can see that traffic type 2 has the highest count and generates the most revenue. followed by traffic type 1. Operating system 2 has the highest count with the highest Revenue followed by operating system 1.







we can see that traffic type 2 has the highest count and generates the most revenue. followed by traffic type 1. Operating system 2 has the highest count with the highest Revenue followed by operating system 1.

Revenue Visitor Type	False	True
New Visitor	1272	422
Others	69	16
Returning Visitor	9081	1470

In this case, the chi2-statistics is 135.2519228192047 and the p-value is extremely low (4.269904152293867e-30) which means that the probability of observing the dataset given the null hypothesis (that the two variables are independent) is extremely low. So we can reject the null hypothesis and conclude that there is a significant relationship between the variables 'VisitorType' and 'Revenue'.

Chi2_Value	p-value
135.2519228192047	4.269904152293867e-30

Revenue Visitor Type	False	True
New Visitor	1431.86277372	262.13722628
Others	71.84671533	13.15328467
Returning Visitor	8918.29051095	1632.7094895

The chi-squared test of independence compares the observed frequencies of the different categories of the two variables to the frequencies that would be expected if the variables were independent. In this case, the test found that the observed frequencies of the "Returning_Visitor" and "Revenue" categories were much higher than the expected frequencies, indicating that there is a strong association between the two variables. The expected frequencies are calculated based on the marginal distributions of the two variables, and the calculated values are compared with the observed values. If the observed values deviate significantly from the expected values, it suggests that there is an association between the two variables. In this case, the test found that the observed frequency of the "Returning_Visitor" and "True" (revenue generated) categories was much higher than the expected frequency, indicating that returning visitors are more likely to generate revenue as compared to new visitors.

Revenue Weekend	False	True
False	8053	1409
True	2369	499

In this case, the chi2-statistics is 10.390978319534856 and the p-value is extremely low (0.0012663251061221968) which means that the probability of observing the dataset given the null hypothesis (that the two variables are independent) is extremely low. So we can reject the null hypothesis and conclude that there is a significant relationship between the variables 'Weekend' and 'Revenue'.

Chi2_Value	p-value
10.390978319534856	0.0012663251061221968

Revenue Weekend	False	True
False	7997.80729927	1464.19270073
True	2424.19270073	443.80729927

The chi-squared test of independence compares the observed frequencies of the different categories of the two variables to the frequencies that would be expected if the variables were independent. In this case, the test found that the observed frequencies of the "False" and "Revenue" categories were much higher than the expected frequencies, indicating that there is a strong association between the two variables. The expected frequencies are calculated based on the marginal distributions of the two variables, and the calculated values are compared with the observed values. If the observed values deviate significantly from the expected values, it suggests that there is an association between the two variables. In this case, the test found that the observed frequency of the "False" and "True" (revenue generated) categories was much higher than the expected frequency, indicating that returning visitors are more likely to generate revenue as compared to new visitors.

Revenue Month	False	True
Aug	357	76
Dec	1511	216
Feb	181	3
Jul	366	66
June	259	29
Mar	1715	192
May	2999	365
Nov	2238	760
Oct	434	115
Sep	362	86

In this case, the chi2-statistics is 384.93476153599426 and the p-value is extremely low (2.2387855164805443e-77) which means that the probability of observing the dataset given the null hypothesis (that the two variables are independent) is extremely low. So we can reject the null hypothesis and conclude that there is a significant relationship between the variables 'VisitorType' and 'Revenue'.

Chi2_Value	**p-value**
384.93476153599426	**2.2387855164805443e77,**

Pre-Processing

Within the exploratory data analysis step, I found that there were no missing or duplicate values. The next step is data pre-processing where I observed that in columns Month, VisitorType were of object and columns Weekend and Revenue(target) were of type bool. Features like Month and VisitorType need to be converted into values as they were of categorical type. For VisitorType, I used Label Encoding and for Month I mapped the values in order from January to December. Binary encoding was used for boolean features. For a better understanding of how variables affect the target, I used the statsmodels library logit function. The summary table of Logit() gives us a descriptive summary of the regression results.

Features like Administrative, Administrative_Duration, Informational, Informational_Duration, BounceRates in %, Region, TrafficType, and Weekend had a p-value greater than 0.05, which indicates that these variables are not significant in predicting the target variable.

My next step was to calculate the Variance Inflation factor to determine whether the variables were multi-collinear. The variables BounceRates in % and ExitRates in % caused multi-collinearity with values of 6.407180 and 7.139377, respectively.

For feature importance and feature selection, I used multiple techniques like PermutationImportance, RandomForestClassifier, and XGBoost Calssifier.

After the trial and error method, I decided to keep all the features as my model was performing better and gave better results.

Model Building

Split the data using train-test-split with a test size of 0.2. I experimented with the basic algorithms and found Random Forest Classifier to give the best score. Therefore, I went ahead with the said model to test which of the hyperparameters gave the highest score as per the ground truth. I also experimented with Feed Forward neural Network which gave amazing results but was not as close to Random Forest Classifier. The last part was handling class imbalance, for that, I used multiple techniques like using random forests parameter class_weights, and ADASYN. ADASYN with Random Forest Classifier gave the best results. The metrics used were ROC AUC Score, Classification Report, and Confusion Matrix.

Random Forest Classifier Result

```
AUC ROC Score: 0.934
Classification Matrix: [[1915  185]
                        [  90 1975]]
Classification Report:          precision    recall  f1-score   support

                  0       0.96      0.91      0.93      2100
                  1       0.91      0.96      0.93      2065

           accuracy                           0.93      4165
          macro avg       0.93      0.93      0.93      4165
       weighted avg       0.93      0.93      0.93      4165
```