AMOABENG YEBOAH

PREDICTIVE ANALYSIS FOR WEBSITE USER ENGAGEMENT

- Machine Learning
- EDA

Data description

Dataset Overview

The dataset comprises 14,731 entries across 22 features, focusing on website user behavior and characteristics.

Key Features

- **Numerical Features:** Include HomePage, HomePage_Duration, LandingPage, LandingPage_Duration, ProductDescriptionPage, ProductDescriptionPage_Duration, GoogleMetric:Bounce Rates, GoogleMetric:Exit Rates, GoogleMetric:Page Values, SeasonalPurchase, OS, SearchEngine, Zone, Type of Traffic, and WeekendPurchase.

- **Categorical Features:** Comprise Month_SeasonalPurchase, CustomerType, Gender, Cookies Setting, Education, Marital Status, and the target variable Made_Purchase.

Milestones

EDA

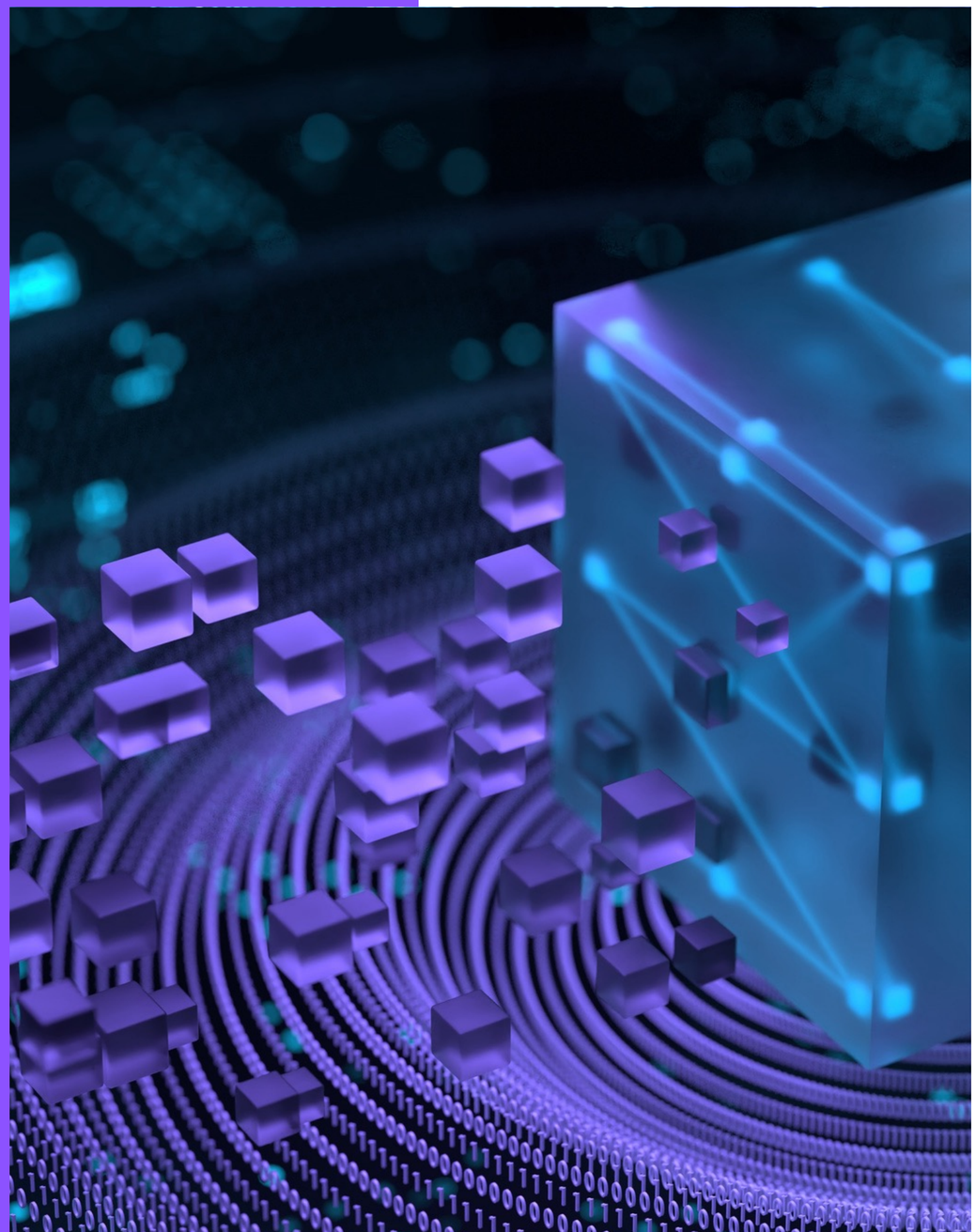
1. Check for missing values

Several features have missing values, ranging from 117 to 167 instances out of approximately 14,600+ entries.

```
HomePage                          153
HomePage_Duration                 150
LandingPage                       153
LandingPage_Duration              135
ProductDescriptionPage            123
ProductDescriptionPage_Duration   167
GoogleMetric:Bounce Rates         151
GoogleMetric:Exit Rates           129
GoogleMetric:Page Values          132
SeasonalPurchase                  150
Month_SeasonalPurchase            144
OS                                134
SearchEngine                      122
Zone                              117
Type of Traffic                   143
CustomerType                      144
Gender                            145
Cookies Setting                   144
Education                         136
Marital Status                    130
WeekendPurchase                   121
Made_Purchase                       0
dtype: int64
```

2. Summary statistics for numerical features

HomePage	HomePage_Duration	LandingPage	LandingPage_Duration	Product DescriptionPage	Product DescriptionPage_Duration	GoogleMetric:Bounce Rates	GoogleMetric: Exit Rates	GoogleMetric: Page Values	SeasonalPurchase	OS	SearchEngine	Zone	Type of Traffic	Weekend-Purchase	Purchase Made
count	14578.000000	14581.000000	14578.000000	14596.000000	14608.000000	14564.000000	14580.000000	14602.000000	14599.000000	14581.000000	14597.000000	14609.000000	14614.000000	14588.000000	14610.000000
mean	2.250240	79.300762	0.490739	33.455943	31.559488	1184.346084	0.023366	0.044664	4.812620	0.064083	2.122422	2.356629	3.155673	4.090143	0.234155
std	3.288042	179.374699	1.252376	140.146256	44.897089	2009.496307	0.050011	0.049912	16.887366	0.202583	0.914404	1.721823	2.405155	4.040147	0.423484
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	173.187500	0.000000	0.014502	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000	0.000000
50%	1.000000	5.000000	0.000000	0.000000	17.500000	584.333333	0.003478	0.026406	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000	0.000000
75%	3.000000	91.000000	0.000000	0.000000	38.000000	1434.255128	0.018182	0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000	0.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000	361.763742	1.000000	8.000000	13.000000	9.000000	20.000000	1.000000



3. Distribution of the target variable

False 0.615369
True 0.384631
Name: Made_Purchase, dtype: float64

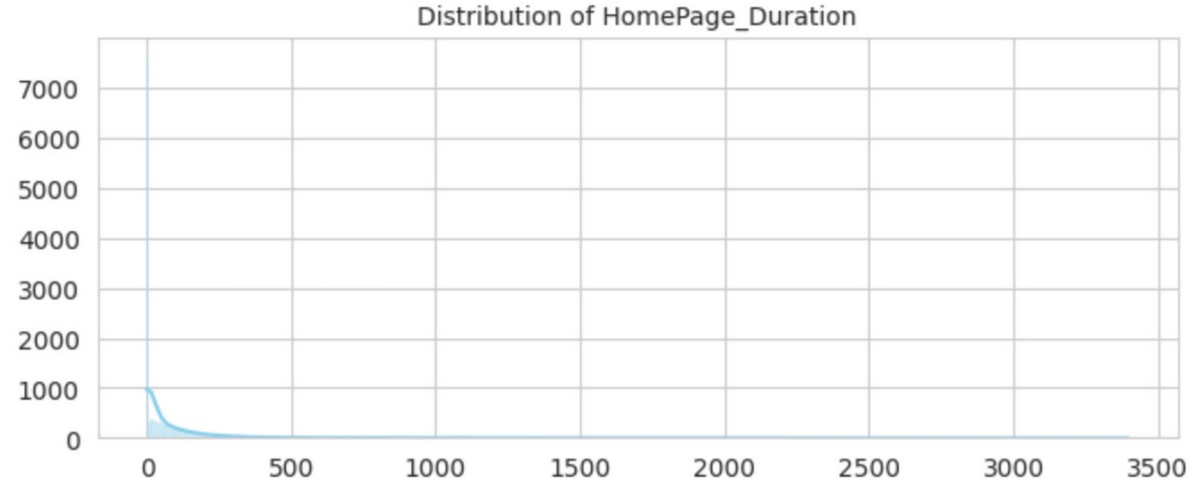
The Made_Purchase variable indicates that
approximately 61.5% of
sessions did not result in a purchase,
while 38.5% did.

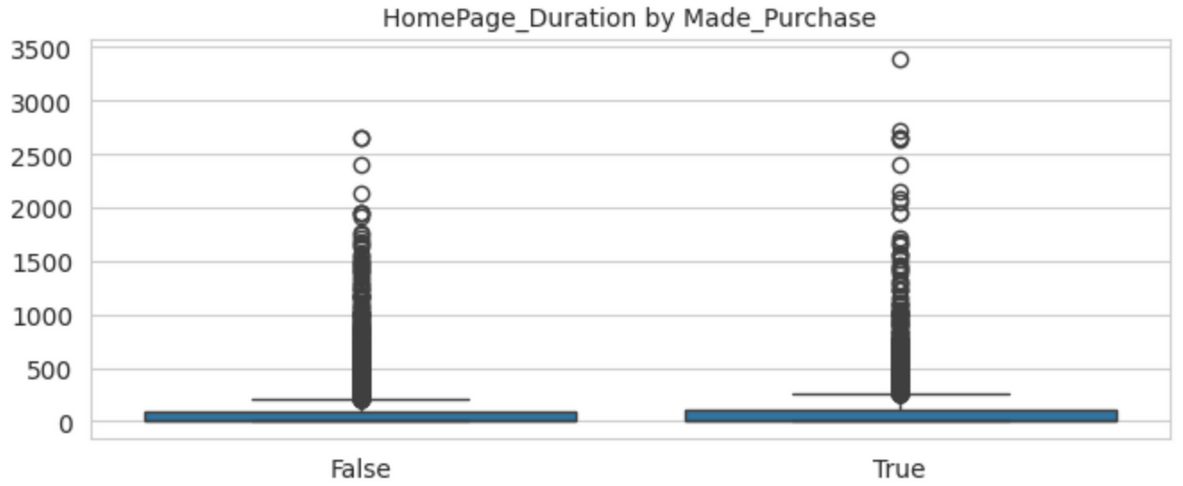
4. Data Cleaning: Impute missing values for numerical features with their median and categorical features with their mode

- Numerical Columns
- Categorical Columns
- Imputation

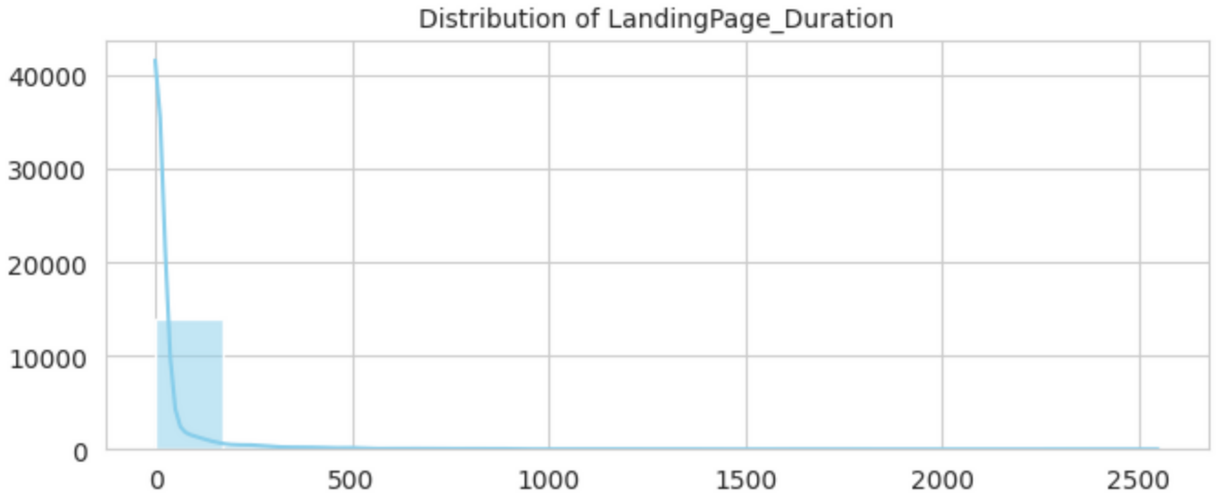
5. Check if any missing values remain
6. Setting the aesthetic style of the plots
7. Define a list of key numerical features for visualization
8. Plotting distributions of key numerical features

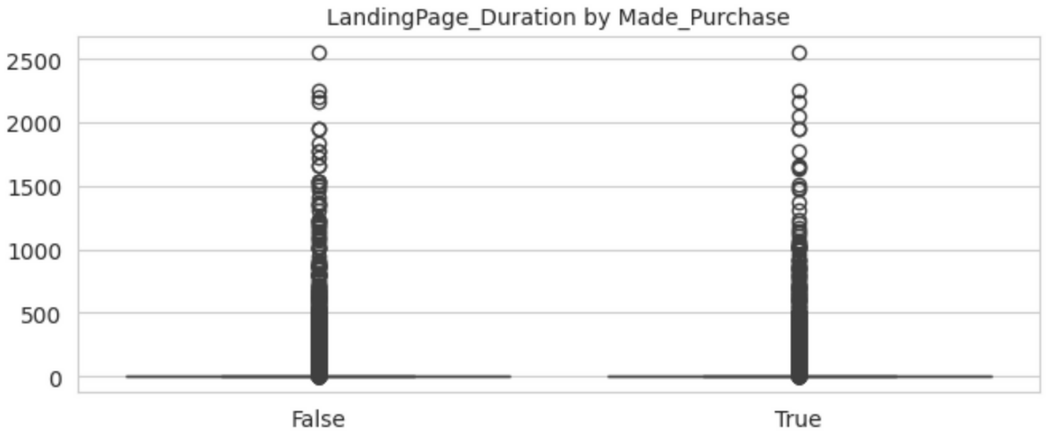
Homepage Duration



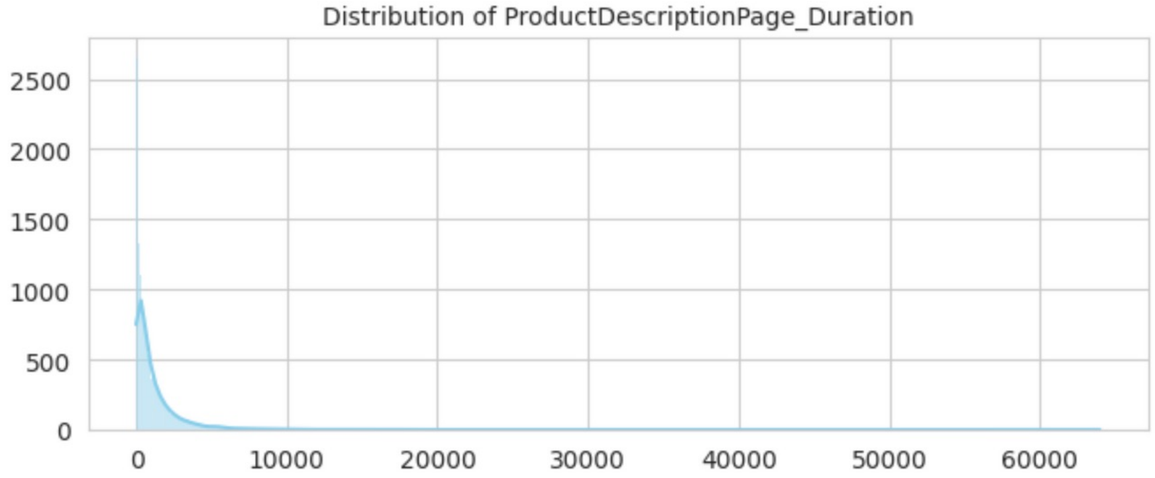


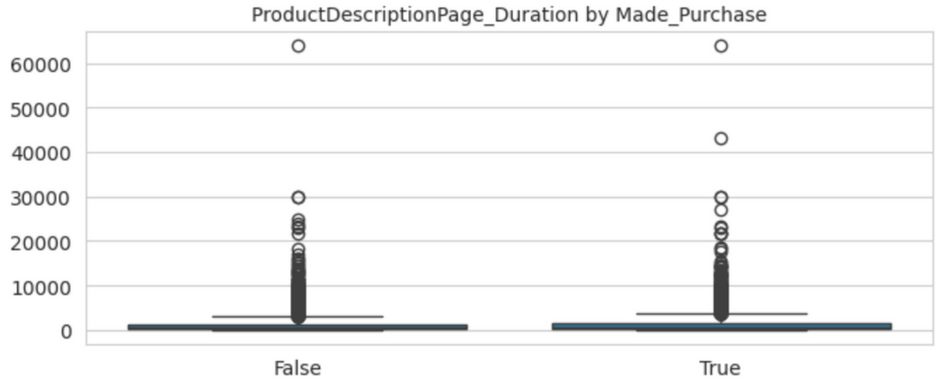
Landingpage Duration



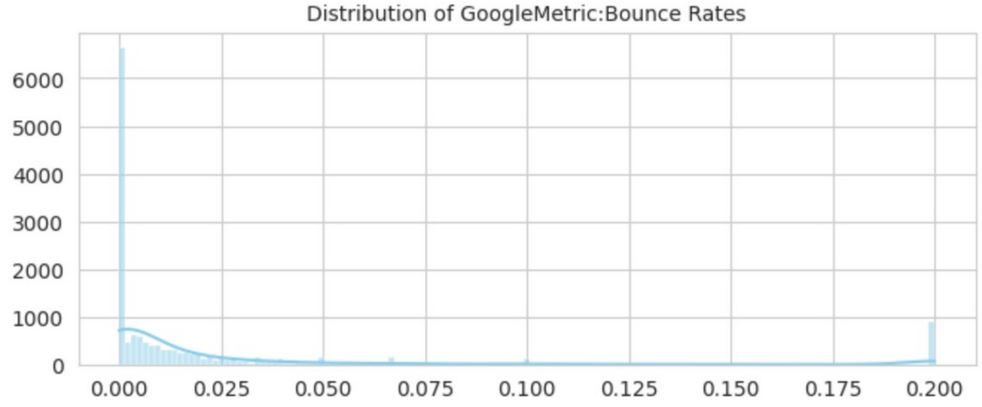


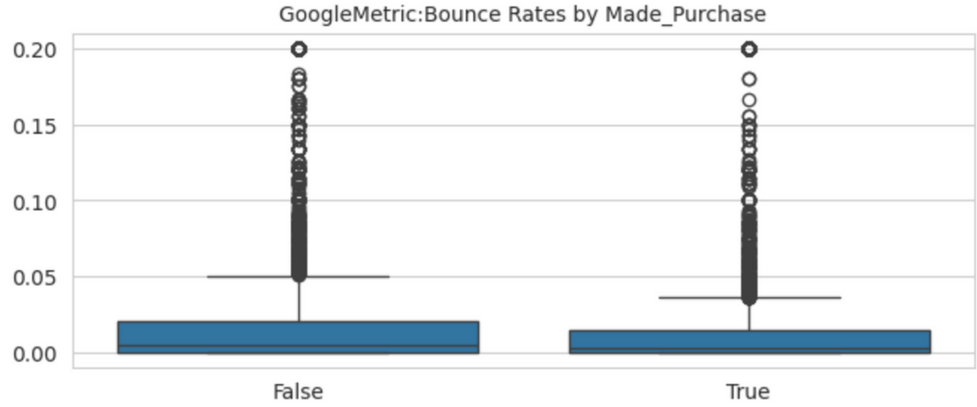
Product Description Page



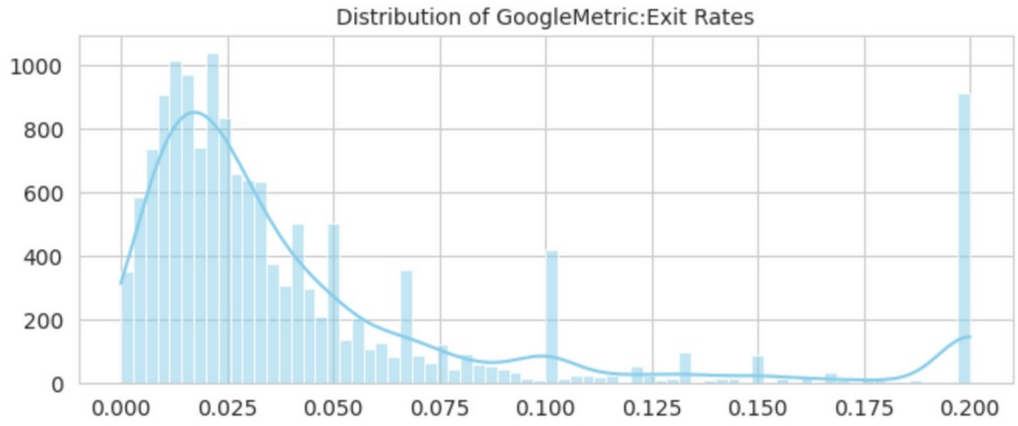


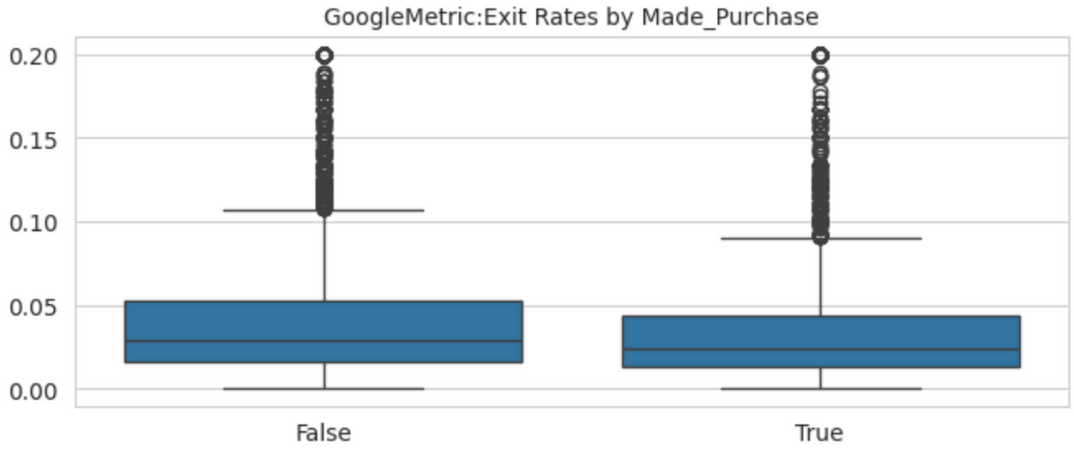
Google Metric



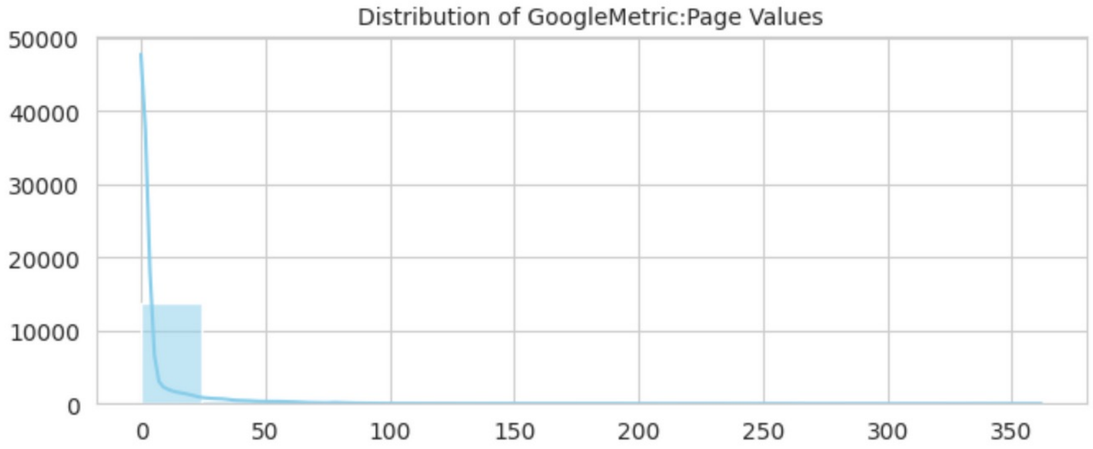


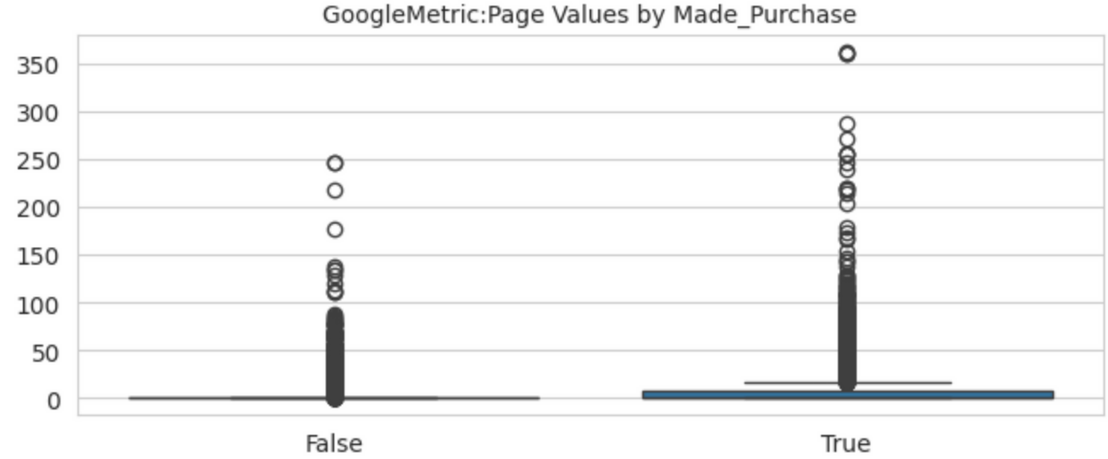
Google Metric : Exit Rates





Google Metric: Page Values





Distributions of Features

- HomePage_Duration, LandingPage_Duration, and ProductDescriptionPage_Duration show skewed distributions with a long tail, indicating that while many sessions have short durations, a few sessions have very long durations on these pages

- GoogleMetric:Bounce Rates and GoogleMetric:Exit Rates also display skewed distributions, with a majority of sessions having lower rates, suggesting that most users navigate beyond their entry page.

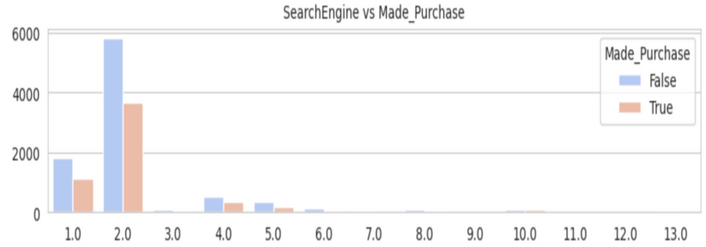
- GoogleMetric:Page Values shows that most pages have a value close to zero, with a few exceptions having higher values, indicating that most pages are not directly preceding purchases or goal completions.

Relationship with Made_Purchase

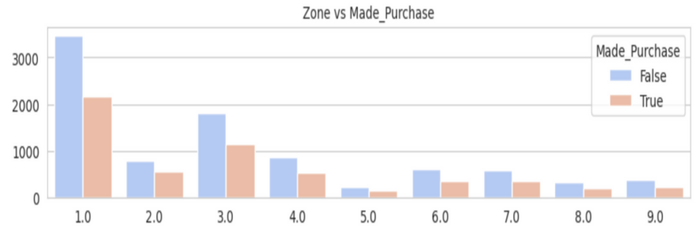
- For all features, there is a noticeable difference in distributions between sessions that resulted in a purchase and those that did not. Particularly, GoogleMetric:Page Values shows a stark contrast, where sessions leading to a purchase tend to have higher page values.

- Duration features (HomePage_Duration, LandingPage_Duration, ProductDescriptionPage_Duration) generally show higher values for sessions that ended in a purchase, suggesting that longer engagement with the site is positively correlated with the likelihood of making a purchase.

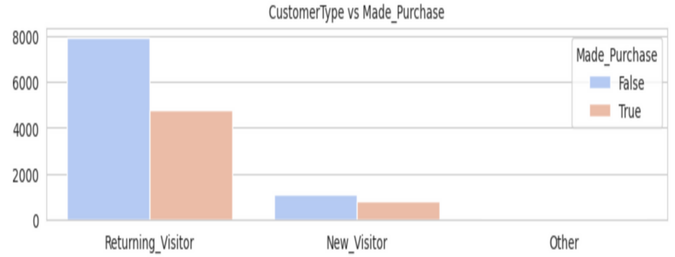
OS (Operating System):
The distribution across different operating systems shows some variation in purchase behavior, suggesting that the user's OS could influence their likelihood of making a purchase.



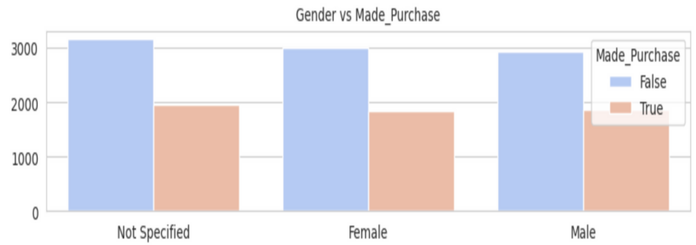
Search Engine:
Different search engines have varying levels of association with purchase outcomes. This might reflect the search engine's effectiveness in bringing potential buyers to the site or differences in user demographics across search engines.



Zone:
The geographical zone of the user shows varied purchase patterns, indicating that regional factors or preferences could affect purchase decisions.

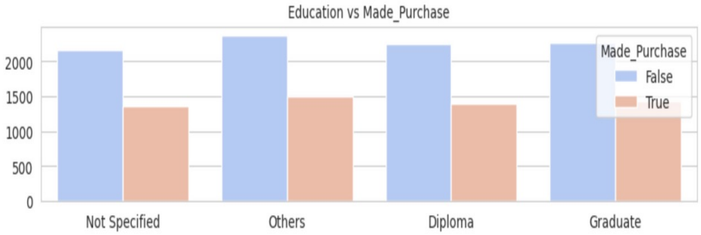


Customer Type: Returning_Visitor:
shows a higher proportion of purchases compared to new visitors, highlighting the importance of retaining customers and the potential higher conversion rate among users familiar with the website.

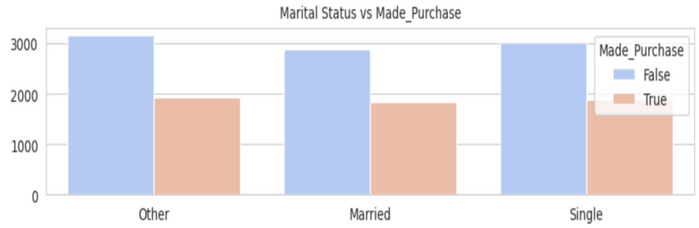


Gender:
Gender distribution indicates differences in purchase behavior, suggesting that male and female users might have different preferences or shopping behavior.

Cookies Setting:
The cookies setting preference ('Deny', 'ALL', etc.) shows differences in purchase outcomes, possibly reflecting user comfort with privacy settings and its impact on personalized shopping experiences.



Education:
Education levels show some variations in purchasing behavior, which could be useful for segmenting users and tailoring marketing strategies.



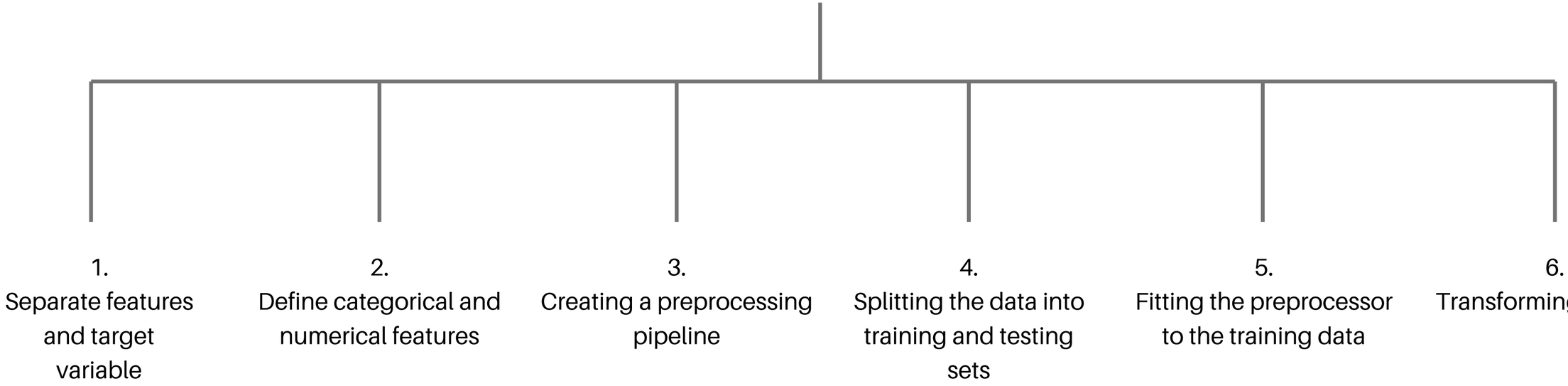
Marital Status:
sMarital status also shows variations in purchase behavior, suggesting that relationship status might influence shopping patterns or needs.

Machine Learning

1. **Separate features and target variable**
2. **Define categorical and numerical features**
3. **Creating a preprocessing pipeline**
4. **Splitting the data into training and testing sets**
5. **Fitting the preprocessor to the training data**
6. **Transforming the data**

The data has been successfully preprocessed and split into training and testing sets, with the training set containing 11,784 samples and the testing set containing 2,947 samples. The preprocessing steps included scaling numerical features and one-hot encoding categorical features, resulting in a total of 41 features after transformation.

Machine Learning



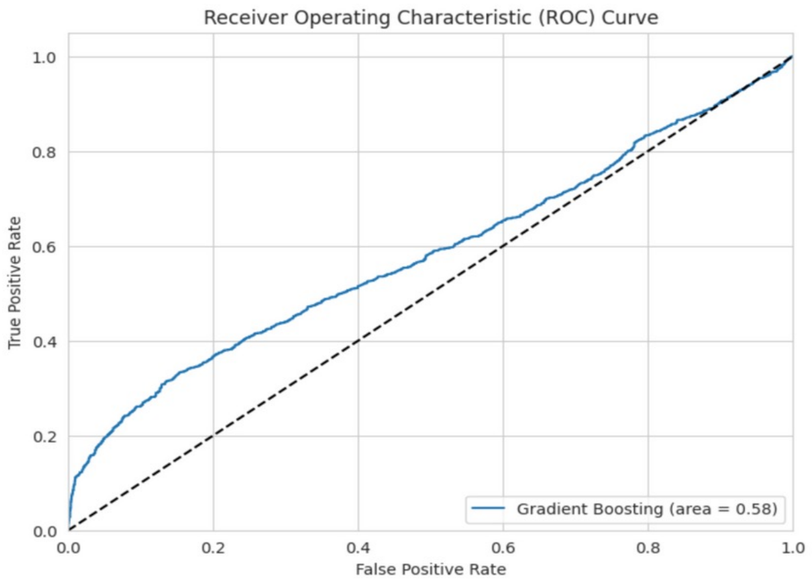
The data has been successfully preprocessed and split into training and testing sets, with the training set containing 11,784 samples and the testing set containing 2,947 samples. The preprocessing steps included scaling numerical features and one-hot encoding categorical features, resulting in a total of 41 features after transformation.

Regression

Logistic, RandomForest & Gradient Boosting

	Accuracy	AUC-ROC
Logistic Regression	0.648117	0.596862
Random Forest	0.521547	0.370473
Gradient Boosting	0.657618	0.582710

Logistic Regression achieved an accuracy of approximately 64.8% and an AUC-ROC of 0.597.
Random Forest had a lower performance with an accuracy of about 51.7% and an AUC-ROC of 0.373.
Gradient Boosting performed slightly better than Logistic Regression with an accuracy of around 65.8% and an AUC-ROC of 0.583.



The Receiver Operating Characteristic (ROC) curve for the Gradient Boosting model shows its performance in distinguishing between sessions that resulted in a purchase and those that did not.
With an area under the curve (AUC) of approximately 0.58, the model demonstrates a modest ability to differentiate between the two classes.
While the AUC is not close to 1, which would indicate perfect classification, it suggests that the model has learned some patterns from the data that are predictive of purchase behavior

Summary of Findings and Recommendations

The Gradient Boosting model showed the best initial performance among the evaluated models, suggesting that ensemble methods that combine multiple weak learners can effectively capture complex relationships in the data.

The AUC-ROC score indicates that there's room for improvement. Model performance could potentially be enhanced through more sophisticated feature engineering, including creating interaction features, more nuanced handling of categorical variables, and addressing the class imbalance directly (e.g., through oversampling, undersampling, or advanced techniques like SMOTE).

Model interpretability could also be improved. Reviewing feature importances generated by the Gradient Boosting model can provide insights into which factors are most influential in predicting purchase behavior, offering actionable insights for business strategy (e.g., website layout optimizations, targeted marketing campaigns).

LASSO & RIDGE REGRESSION

1. Ridge Regression with L2 Regularization
2. Evaluating Ridge Regression model

.....→ {'Accuracy': 0.6470987444859179, 'AUC-ROC': 0.6002005406767311}

1. Lasso Regression with L1 Regularization
2. Evaluating Lasso Regression model

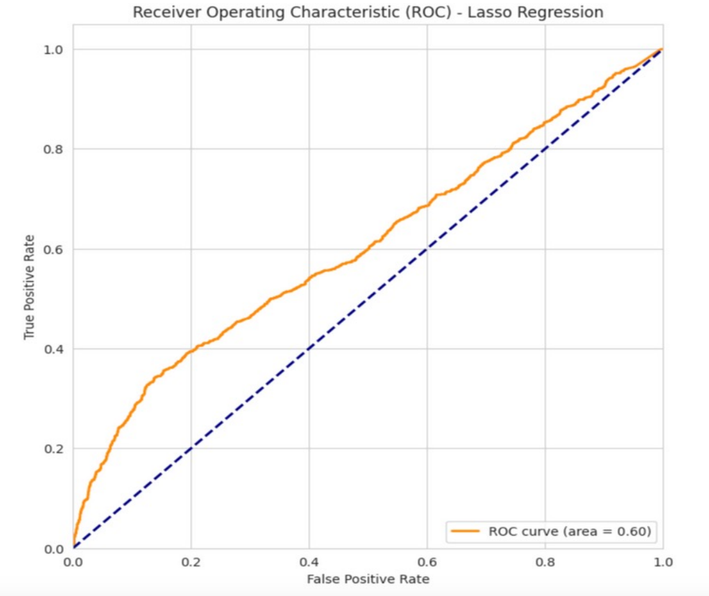
.....→ {'Accuracy': 0.6498133695283339, 'AUC-ROC': 0.6047556302658343}

Both Ridge and Lasso regressions have shown improvements in model performance compared to the initial evaluations of Logistic Regression and Random Forest, with Gradient Boosting being closely competitive. The Lasso regression, with its feature selection capability, slightly outperformed Ridge regression in terms of both accuracy and AUC-ROC.

Gradient Boosting still holds the highest accuracy but does not lead in terms of AUC-ROC. This might indicate that while Gradient Boosting is good at making correct predictions, it might not be as confident in its probabilities as the Lasso model, especially in a balanced dataset scenario.

LASSO TOP FEATURES & ROC

	Feature	Importance
8	Month_SeasonalPurchase_Oct	0.516847
27	HomePage_Duration	-0.105370
5	Month_SeasonalPurchase_Mar	0.039345
4	Month_SeasonalPurchase_June	0.027179
0	Month_SeasonalPurchase_Aug	0.000000
31	ProductDescriptionPage_Duration	0.000000
24	Marital Status_Other	0.000000
25	Marital Status_Single	0.000000
26	HomePage	0.000000
28	LandingPage	0.000000



- **Month_SeasonalPurchase_Oct (Importance: 0.516847):** Indicates that visits in October have a significant positive impact on the likelihood of making a purchase. This could be related to seasonal buying patterns, such as holiday shopping.

- **HomePage_Duration (Importance: -0.105370):** The negative coefficient suggests that longer durations spent on the HomePage are associated with a lower likelihood of making a purchase. This might indicate that users who find what they're looking for quickly are more likely to convert.

- **Month_SeasonalPurchase_Mar (Importance: 0.039345):** Visits in March also positively influence purchase decisions, though it's lesser to the extent of October. This could be related to specific sales or seasonal events.

- **Month_SeasonalPurchase_June (Importance: 0.027179):** Similar to March, visits in June have a positive but relatively small impact on purchasing. This again might relate to seasonal factors or promotions.

The curve illustrates the trade-off between the True Positive Rate and False Positive Rate at various threshold levels, and with an AUC of 0.60, the model has a fair ability to distinguish between the positive and negative classes.

Conclusion

- The project's analysis underscores the importance of thorough data preparation, including cleaning and feature engineering, to address issues like missing values and outliers. This foundation is critical for building reliable predictive models.

- The insights derived from feature distributions suggest that user engagement on the website is not uniform, with significant variations in how different pages contribute to overall engagement or conversion metrics. This variability highlights opportunities for targeted improvements to website content and layout to enhance user engagement and conversions.

- The choice of machine learning models (Logistic Regression, RandomForestClassifier, GradientBoostingClassifier) reflects an approach tailored to capture both linear relationships and complex non-linear patterns in the data, which is appropriate given the varied nature of the dataset.

In summary, the analysis indicates that there's potential to use machine learning to predict user engagement and identify key areas for website optimization. The findings from the data preparation and exploratory data analysis provide a solid basis for developing models that could help in enhancing user experience and potentially increasing conversions on the website.