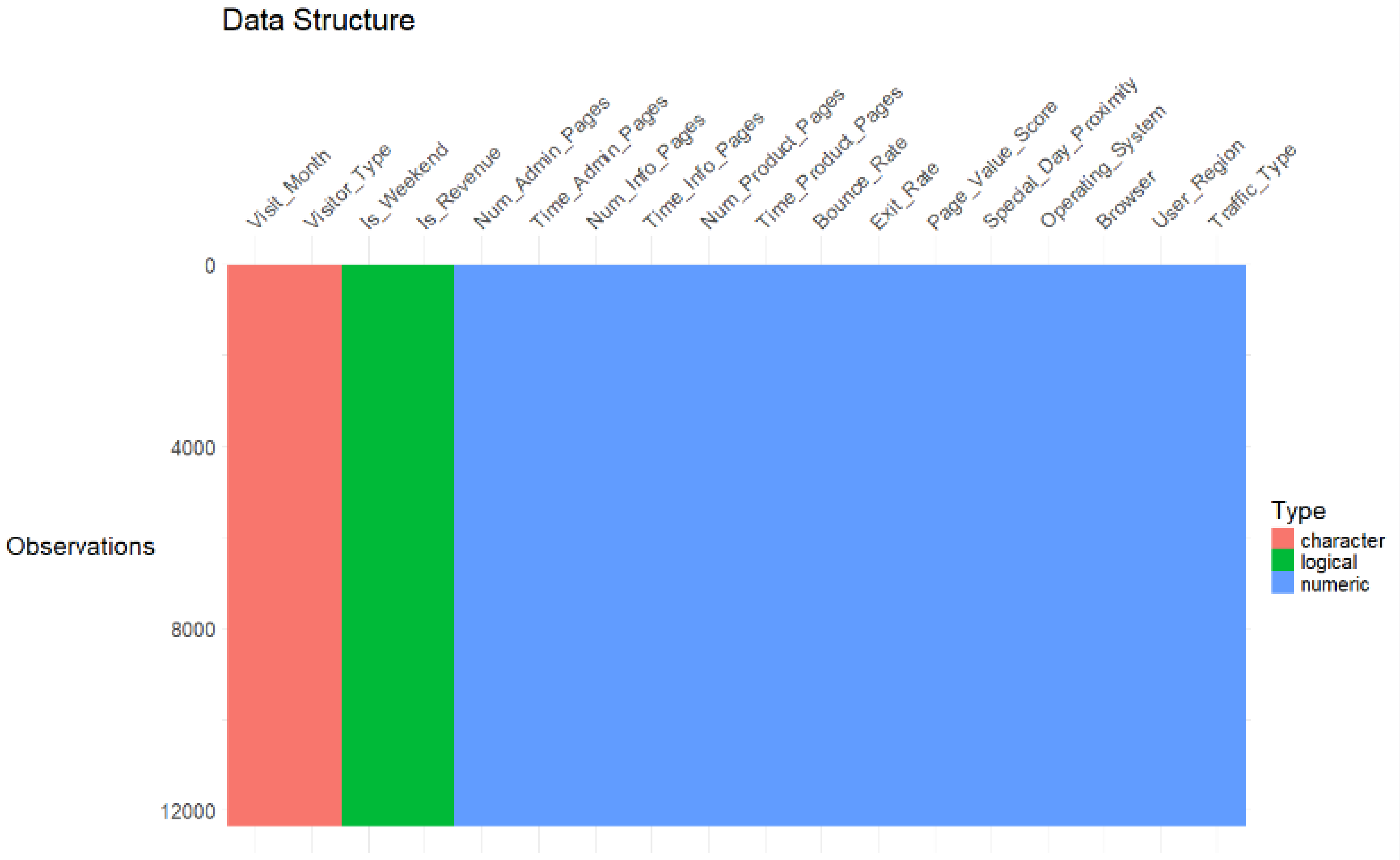
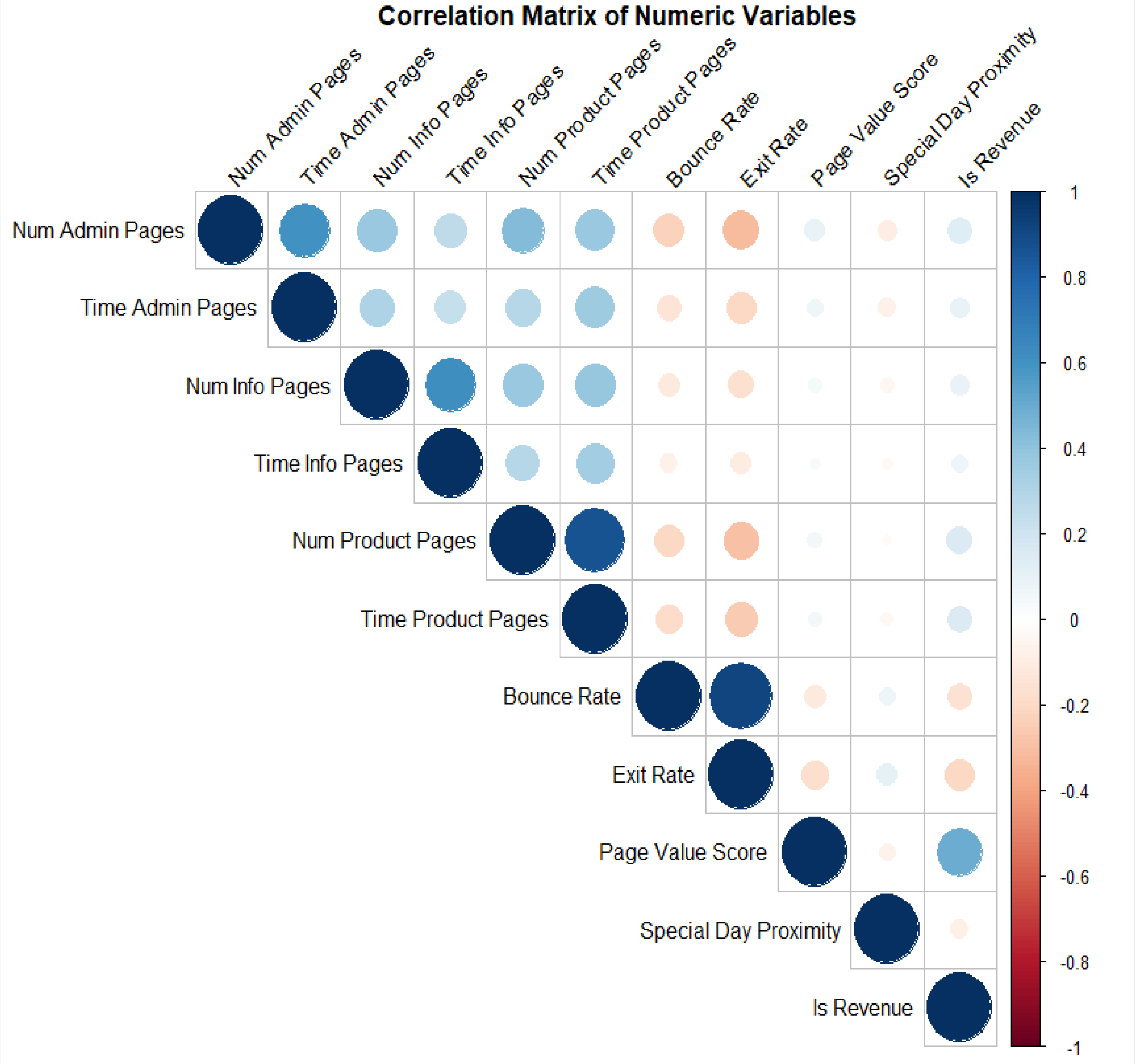
DATA QUALITY ASSESSMENT

- Data types are a mix of categorical/symbolic and numeric variables.
- Target variable is binary and originally denoted as Boolean TRUE/FALSE.
- No missing values identified in the dataset.
- Duplicate values were identified, and after reviewing them, they were retained in the main dataset for modelling, since they are not errors and reflect repeated behaviour of different users. The only exception was during t-SNE visualisation, where duplicates were removed to avoid overplotting and to ensure better visual clarity.
- Minority class (TRUE) is imbalanced compared to the FALSE class. Thus, to prevent biased predictions, this variable should be balanced.



DATA CLEANING & FEATURE ENGINEERING

- Column renaming for interpretability.
- Log transformation of some strongly right skewed variables (e.g., Page_Value_Score, Time_Admin_Pages, Time_Info_Pages, Time_Product_Pages) and discretization of (e.g., Num_Info_Pages, Num_Admin_Pages and Num_Product_Pages).
- High cardinality was identified among some categorical variables e.g., Browser, which had 13 levels, and Traffic_Type, which had 20 levels. To reduce the number of unique levels, rare categories (representing less than 5 percent of the data) were grouped into “Other.”
- Standardisation (Z-score) applied to numeric features to ensure that they have a mean of 0 and a standard deviation of 1, to improve comparability across features and better model performance.
- Target variable encoding from TRUE/FALSE to a factor (YES/NO) for logistic regression & random forest, although XGBoost was encoded numerically (1/0) to ensure compatibility.
- Converted categorical variables into numeric format using feature encoding techniques (e.g., one-hot encoding), ensuring compatiability with machine learning algorithms and improving model performance.
- Correlation matrix revealed weak to moderate correlation between majority of the features and the target variable.
- In addition, there were signs of moderate multicollinearity between some variables (e.g., Num_Product_Pages & Time_Product_Pages, Bounce_Rate & Exit_Rate).
- The Variable Inflation Factor (VIF) technique, helped to determine that none of the features were redundant (VIF <5) and therefore no variables were dropped from the dataset.



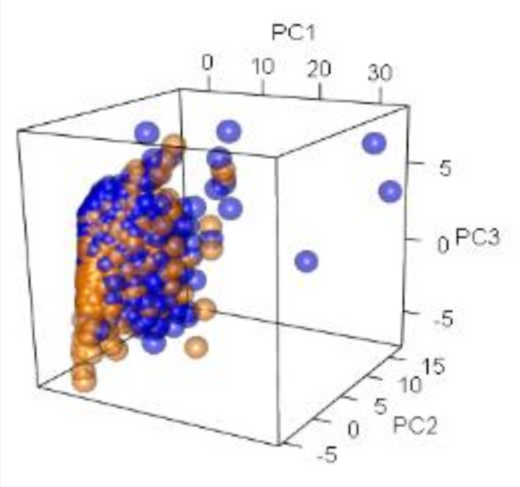
EXPLORATORY DATA ANALYSIS- DATA STRUCTURE

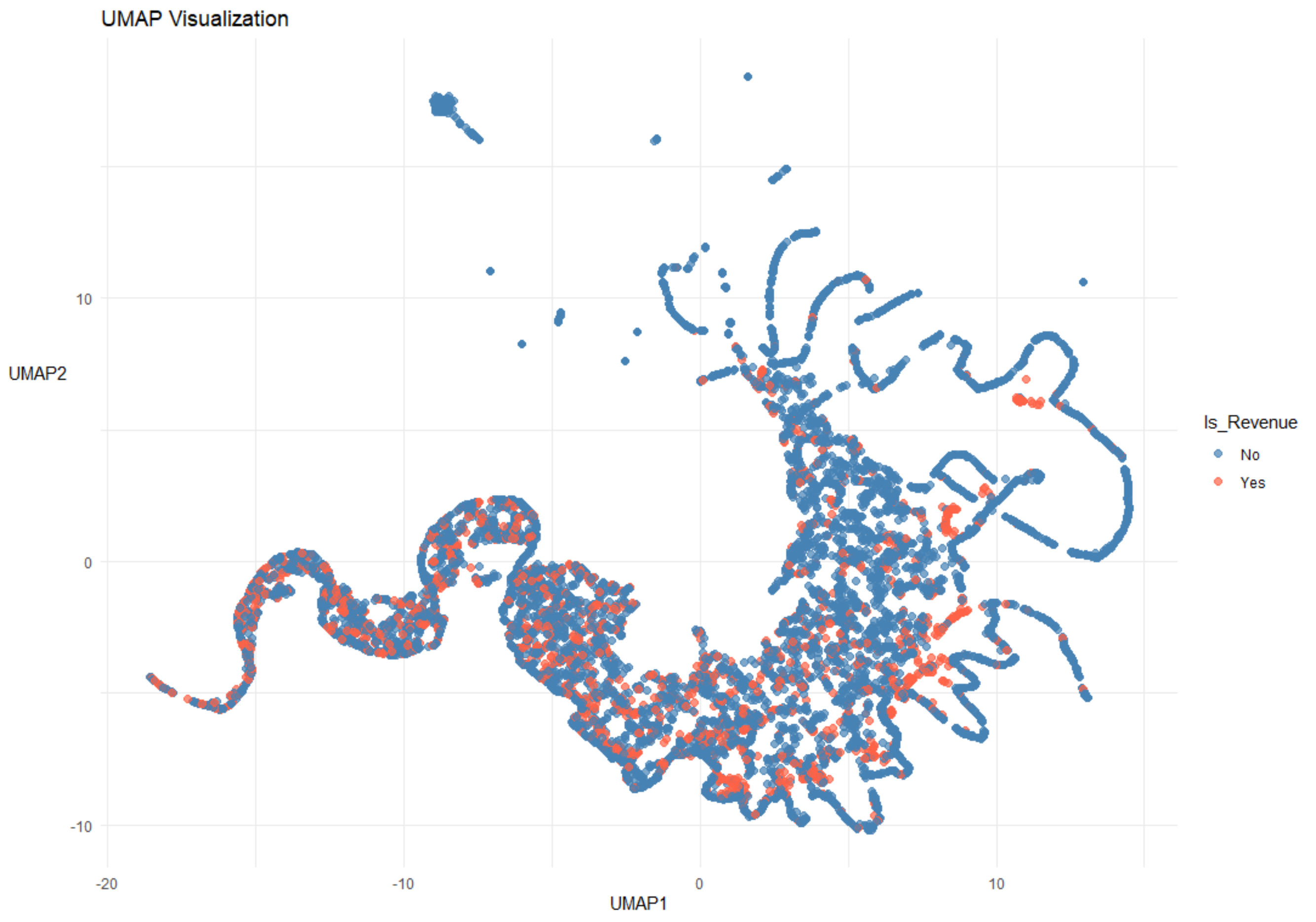
PCA 3D Plot

- Demonstrates the underlying structure of the high-dimensional dataset.
- Shows how observations are distributed across the first three principal components.
- Clear clustering patterns are limited → classes are not easily separable with linear methods.

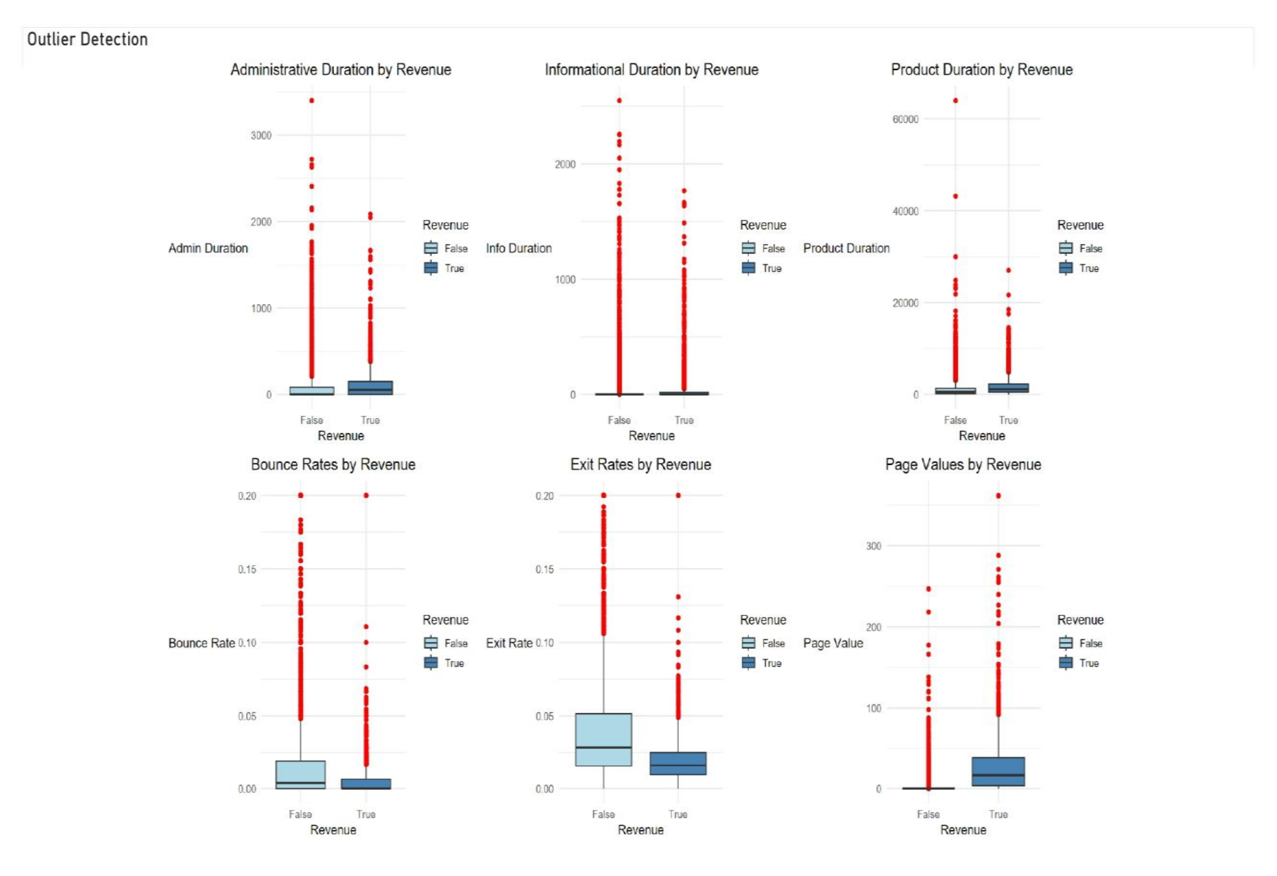
UMAP Plot

- Example run with fixed parameters.
- Captures **non-linear relationships** and complex data structure.
- Reveals overlapping clusters → classification is challenging.
- Indicates that advanced, non-linear models may be more effective.





EXPLORATORY DATA ANALYSIS- DISTRIBUTION ANALYSIS

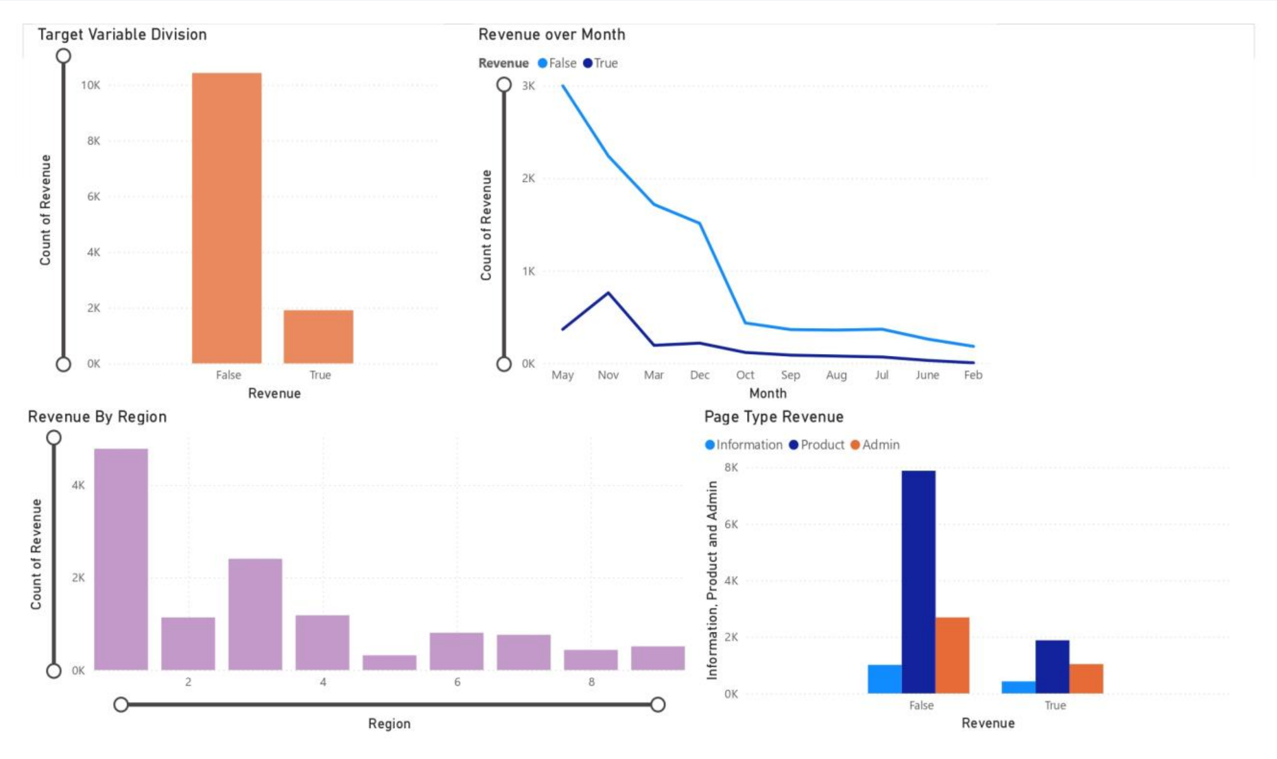


Revenue division has more of false that shows class imbalance

November and May has a peak in revenue suggesting seasonal hikes

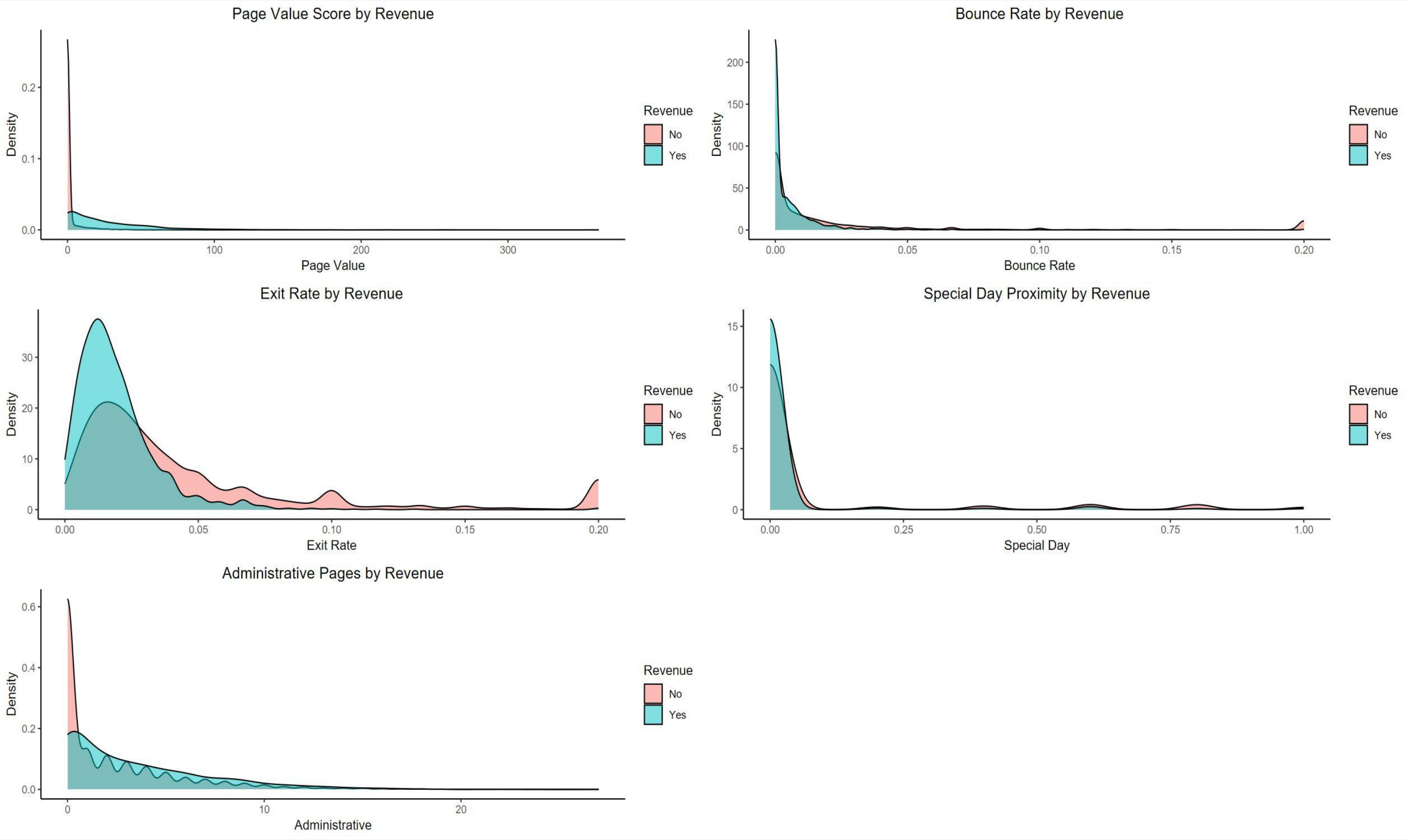
Region 1 has highest revenue followed by 3

Product related duration is higher than Informational and Administration as expected due to the website being a retail page



This set of boxplots visualises the distribution of key numeric features—such as page duration, bounce rate, exit rate, and page value—grouped by session revenue outcome (Yes vs. No). Notably, several variables show significant differences in spread and central tendency between converting and non-converting sessions. For instance, sessions that generated revenue tend to have higher values across engagement-related features, suggesting stronger user intent and interaction.

EXPLORATORY DATA ANALYSIS- DISTRIBUTION ANALYSIS



These plots compare distributions **between sessions that did and did not result in revenue generation**

- **Page Values**:

Revenue-generating sessions have a **notably higher density** at elevated Page Values. This confirms that **higher page value correlates with revenue**.

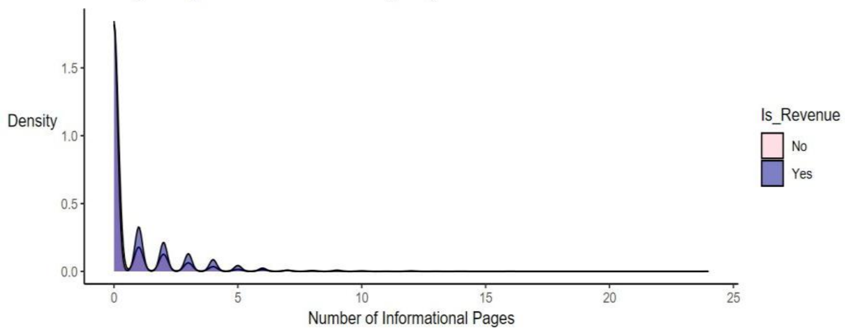
- **Exit Rate & Bounce Rate**:

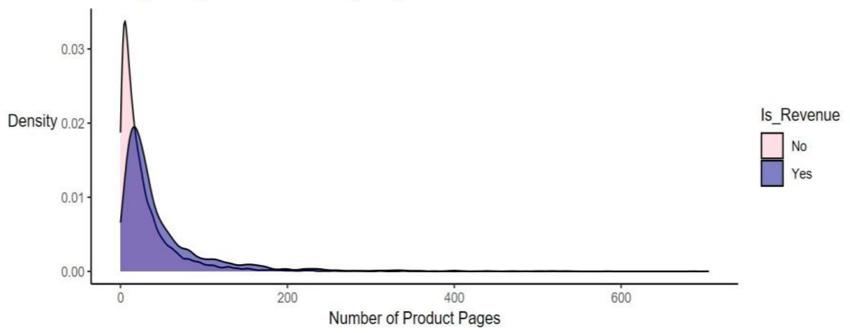
Sessions without revenue tend to have **higher bounce and exit rates**.
Revenue sessions cluster more at **lower bounce and exit rates**, suggesting better engagement.

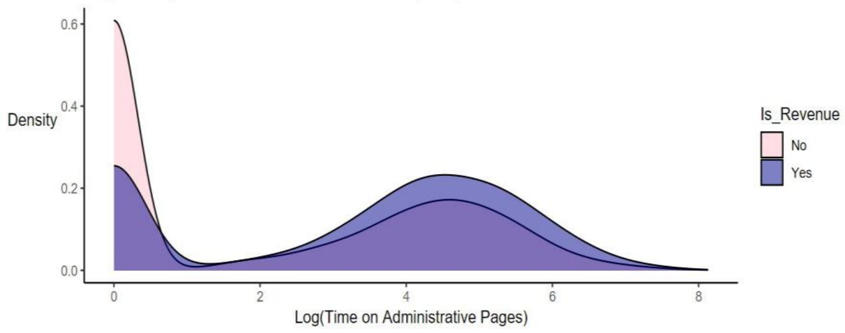
- **Administrative Pages & Special Day Proximity**:

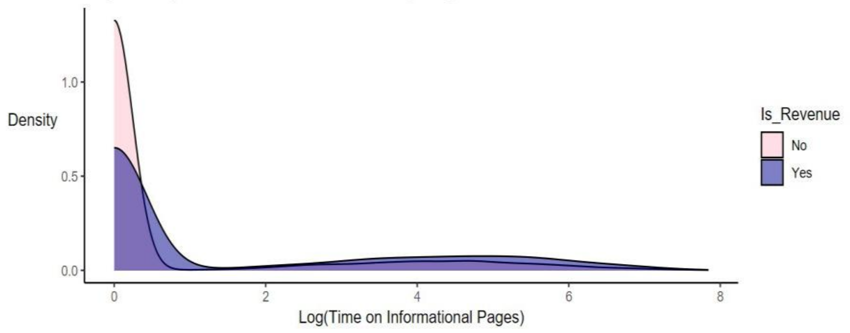
Revenue sessions generally have a **slightly higher density** at low Administrative values.
Special Day shows almost **no clear distinction** between revenue and non-revenue sessions — this might not be a strong predictor.

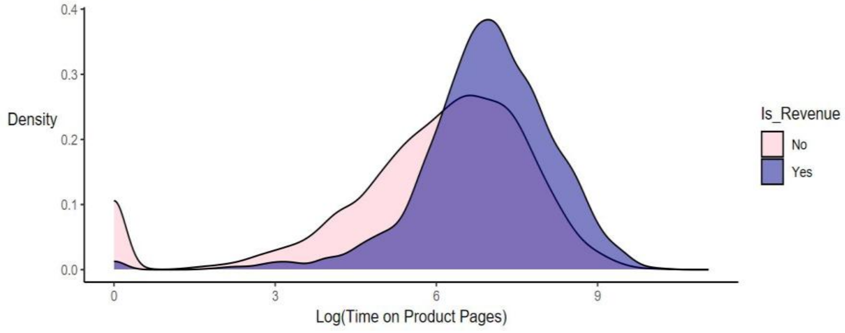
EXPLORATORY DATA ANALYSIS- DISTRIBUTIONANALYSIS


Density Histogram of Informational Pages by Revenue


Density Histogram of Product Pages by Revenue


Log-Density of Time on Administrative Pages by Revenue


Log-Density of Time on Informational Pages by Revenue


Log-Density of Time on Product Pages by Revenue

- **Product Page Views**:

Revenue users view **20–100+ product pages**.
Non-revenue users mostly <10 pages

- **Time on Product Pages (log seconds)**:

Revenue users: peak at **log(6–7)**
Non-revenue: peak at **log(4–5)**

Informational Pages:

- Slightly higher count and time for revenue users.
- Modest impact.

Administrative Pages:

- Revenue users show **second peak at log(4–5)** suggesting deeper engagement.

Summary of findings:

- Most continuous variables are **skewed** → log-transform required for modelling.
- **High PageValue, low BounceRate/ExitRate** correlate with revenue generation.
- **Product engagement** (page views & time) is the **strongest revenue signal**.
- Informational/Admin content plays a **supporting role**, but less predictive.

DATA BALANCING APPROACH

Why Balancing was needed

- Approximately 15% of user sessions resulted in revenue generation, while 85% did not result in revenue generation.
- Class imbalance can cause models to favour the majority class, which weakens their ability to predict outcomes for the minority class.

ROSE Method (Random Over-Sampling Examples)

- Generates synthetic samples for the minority class.
- Helps create a more balanced dataset without simply duplicating existing rows.
- Preserves underlying feature distribution for better generalisation.

Random Stratified Over Sampling Advantages

- Ensures proportional representation of both classes in the resampled dataset.
- Reduces sampling and representation bias in classification models.
- Chosen to improve the model's ability to detect *Revenue* = *Yes* cases.

Before and After balance check (see table below for balancing results)

- Result: A more balanced dataset, leading to fairer representation of both classes and better model performance.

Yes	No
1,908 (15%)	10,422 (85%)

Yes	No
6,056 (49.6%)	6,149 (50%)

SUMMARY OF CLEAN DATASET

Dataset Balance (Post-Cleaning)

- Records: 12,205
- Revenue = Yes: ~49.6%
- Revenue = No: ~50%

Feature Transformation

- Applied log transformations to reduce skewness in time and page variables.
- Discretized selected continuous features to capture non-linear patterns.
- Standardised variables for consistency.

Feature Encoding

- Categorical variables such as *Operating System*, *Browser*, *Traffic Type*, *Visitor Type*, and *Month* were transformed into factors for analysis.
- High-cardinality categorical variables such as *Traffic Type* and *Browser* were grouped into an “Other” category to reduce sparsity.
- For logistic regression and random forest models, categorical factors were handled internally through dummy coding.
- For models such as logistic regression and XGBoost, categorical variables were explicitly one-hot encoded using model.matrix().

Target Variable Encoding

- The target variable Is_Revenue was originally stored as booleans (TRUE/FALSE).
- It was converted to a factor variable with “No” set as the reference category to ensure consistent interpretation in classification models.
- For some analyses and correlation checks, Is_Revenue was temporarily converted to a numeric variable (0 = No, 1 = Yes) which enabled ROC curve generation, probability predictions, and logistic regression analysis.
- Consistent factor encoding ensured that cross-validation and resampling methods (e.g., K-fold CV) worked correctly across all models.

Prepared for Modelling

- Cleaned, balanced, transformed, and validated for stable model training.