Machine Learning Report

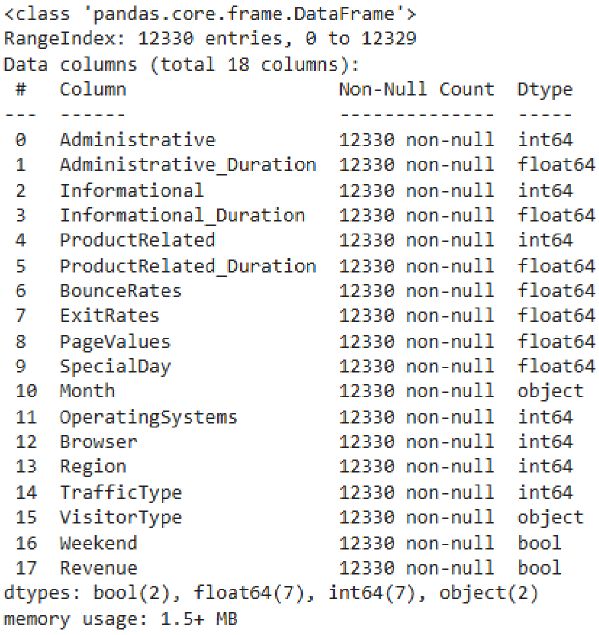
Siddharth Bahekar
Rachita Shah
College of engineering
Northeastern University
Toronto,ON
bahekar.si@northeastern.edu
shah.rachit@northeastern.edu

Abstract

The growth of e-commerce has made understanding online shopping behavior important for improving sales and customer experience. This study examines customer session data to identify patterns that influence purchase decisions. We analysed features such as page visits, session duration, bounce rates, and exit rates to understand their impact. The results show significant links between browsing behavior and purchase likelihood. These findings provide insights into customer segmentation, helping e-commerce platforms improve recommendations and increase conversion rates, contributing to a better understanding of online consumer behavior.

1. Dataset

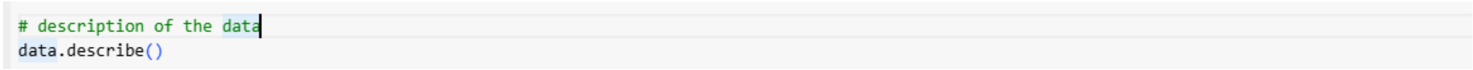
The "Online Shoppers Purchasing Intention Dataset" from the UCI Machine Learning Repository is a suitable choice for this project as it provides important information about user behaviour on online shopping websites. The dataset contains 12330 rows and 18 columns It includes features such as the number of pages visited, time spent on the website, and the type of traffic source. These factors help identify patterns that influence whether a customer makes a purchase or not, which is represented by a binary target variable in the dataset. Using this data, the project can develop predictive models to help improve marketing strategies, such as targeted promotions and personalized campaigns, to increase customer engagement and revenue.



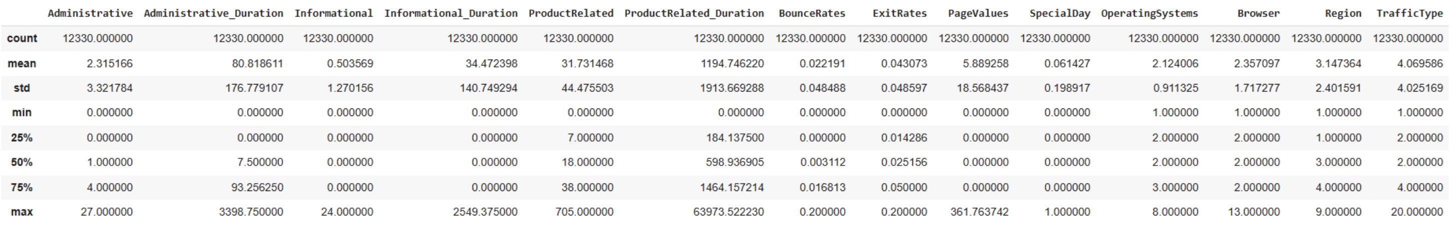
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Administrative           12330 non-null  int64  
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64  
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64  
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object 
 11  OperatingSystems         12330 non-null  int64  
 12  Browser                  12330 non-null  int64  
 13  Region                   12330 non-null  int64  
 14  TrafficType              12330 non-null  int64  
 15  VisitorType              12330 non-null  object 
 16  Weekend                  12330 non-null  bool   
 17  Revenue                  12330 non-null  bool   
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

2. Dataset Description

Online Shoppers Intention:



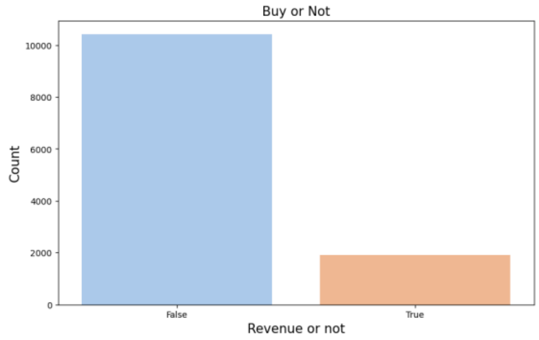
```
# description of the data
data.describe()
```



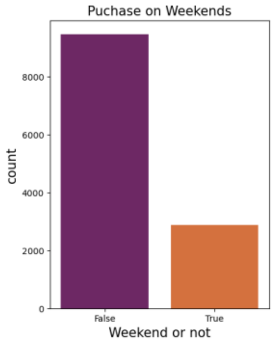
	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.022191	0.043073	5.889258	0.061427	2.124006	2.357097	3.147364	4.069586
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048488	0.048597	18.568437	0.198917	0.911325	1.717277	2.401591	4.025169
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
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.000000	0.014286	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.003112	0.025156	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157214	0.016813	0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000
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

3. Univariate Analysis:

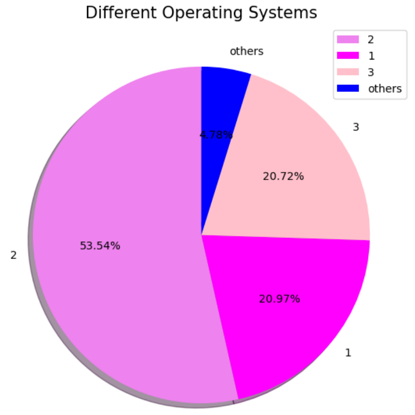
Univariate Analysis is the analysis of a single variable at a time to summarize its main characteristics, patterns, and distribution.



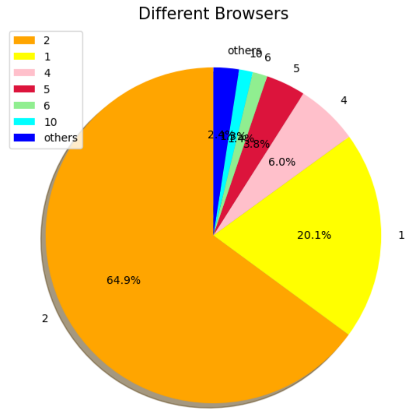
1. Revenue



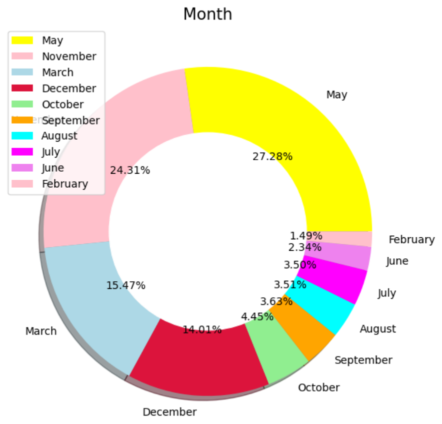
2. Weekend



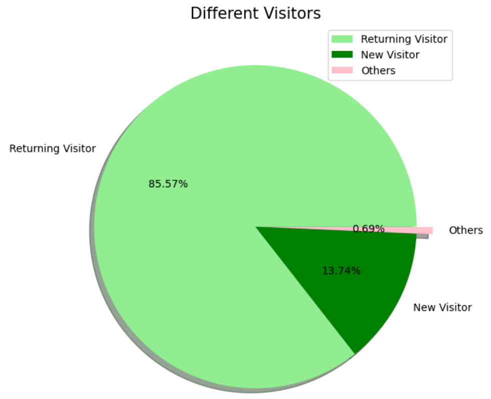
3. Operating System



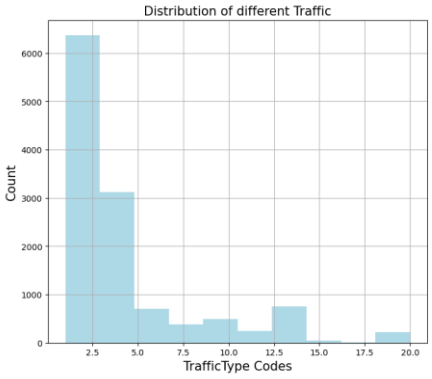
4. Browser



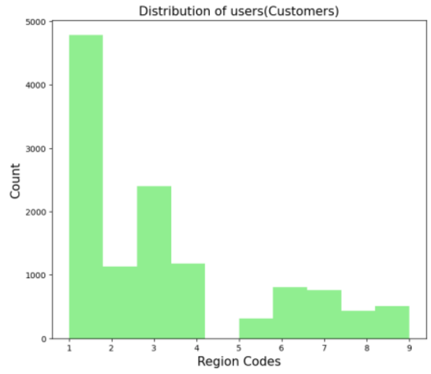
5. Month



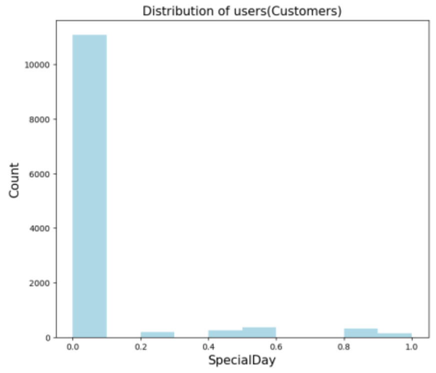
6. Visitor Type



7. Traffic Type



8. Region



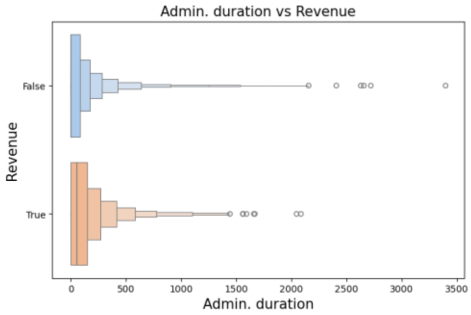
9. Special Days

Observations:

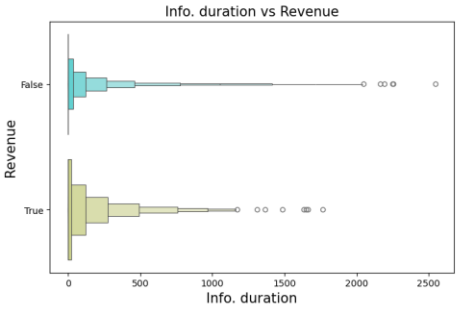
1. For Revenue it is observed that only 15.5% of instances indicate revenue generation, showing a high imbalance in the Revenue feature.
2. Only 23.3% of user activity occurs on weekends, indicating an imbalance in the Weekend feature.
3. For operating System, the top 3 operating systems account for 95% of user activity, with OS 2 being the most used.
4. The top 3 browsers constitute 90% of usage, with Browser 2 leading significantly.
5. According to the graph we can see that user activity peaks in May, November, March, and December, showing seasonal trends.
6. It is observed that 85% of visitors are returning users, making them a key demographic for targeted marketing.
7. It is observed that traffic types are exponentially distributed, with the top 5 types covering most of the dataset.
8. Regional data is also exponentially distributed, with the top 4 regions accounting for the majority of users.
9. It is observed that 85% of instances have no special day association, though special days may still influence user behavior.

These observations highlight key imbalances and concentrated distributions that can guide targeted strategies and preprocessing efforts.

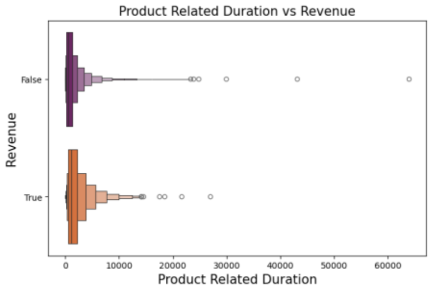
4. Bi-Variate Analysis



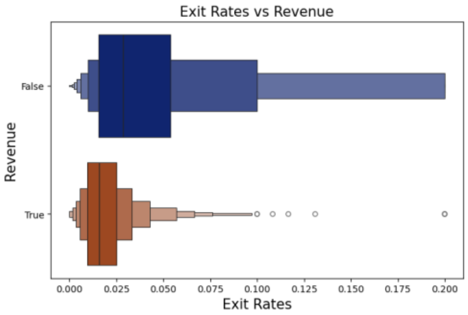
1.Admindurationn vs Revenue



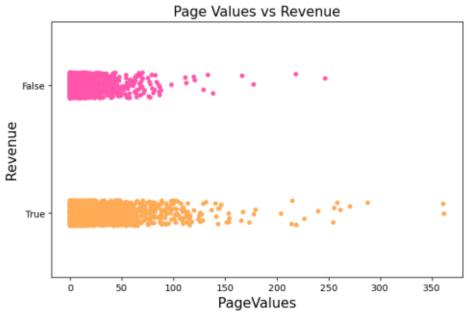
2. Info Duration vs. Revenue



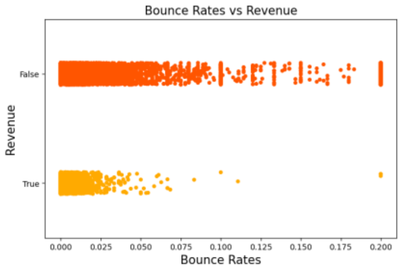
3. Product Related Duration vs. Rev



4. Exit Rates vs. Revenue

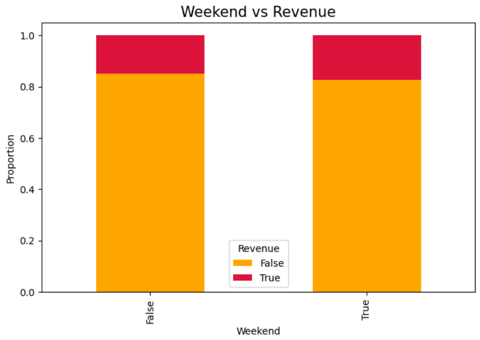


5. Page Values vs. Revenue



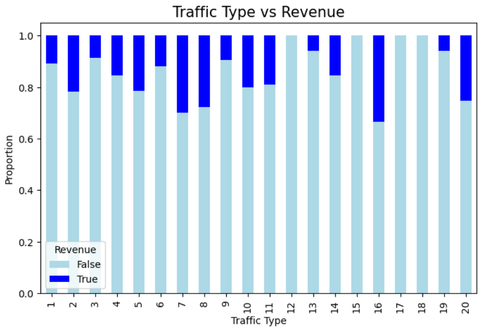
6. Bounce Rates vs. Revenue

```
<Figure size 1500x500 with 0 Axes>
```

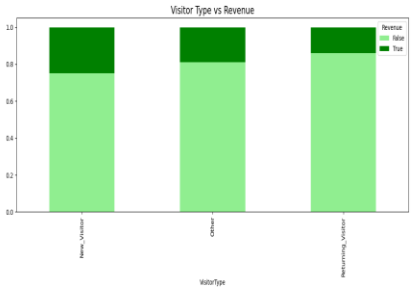


7. Weekend vs. Revenue

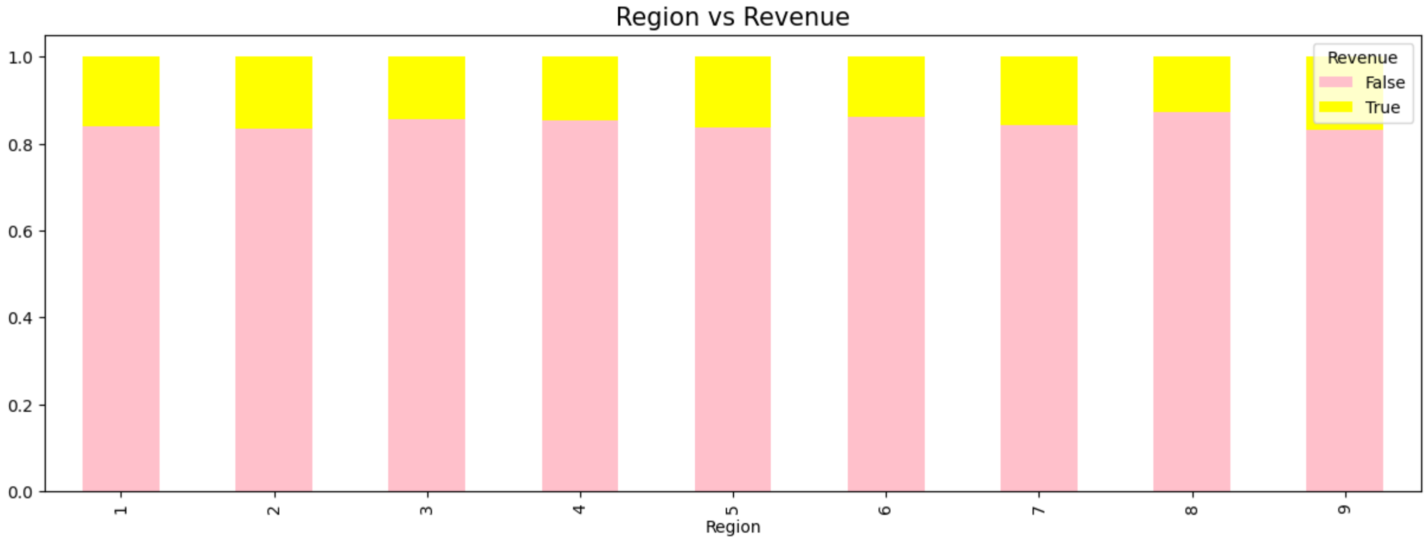
```
<Figure size 1500x500 with 0 Axes>
```



8. Traffic Type vs. Revenue



9.Visitor Type vs. Revenue



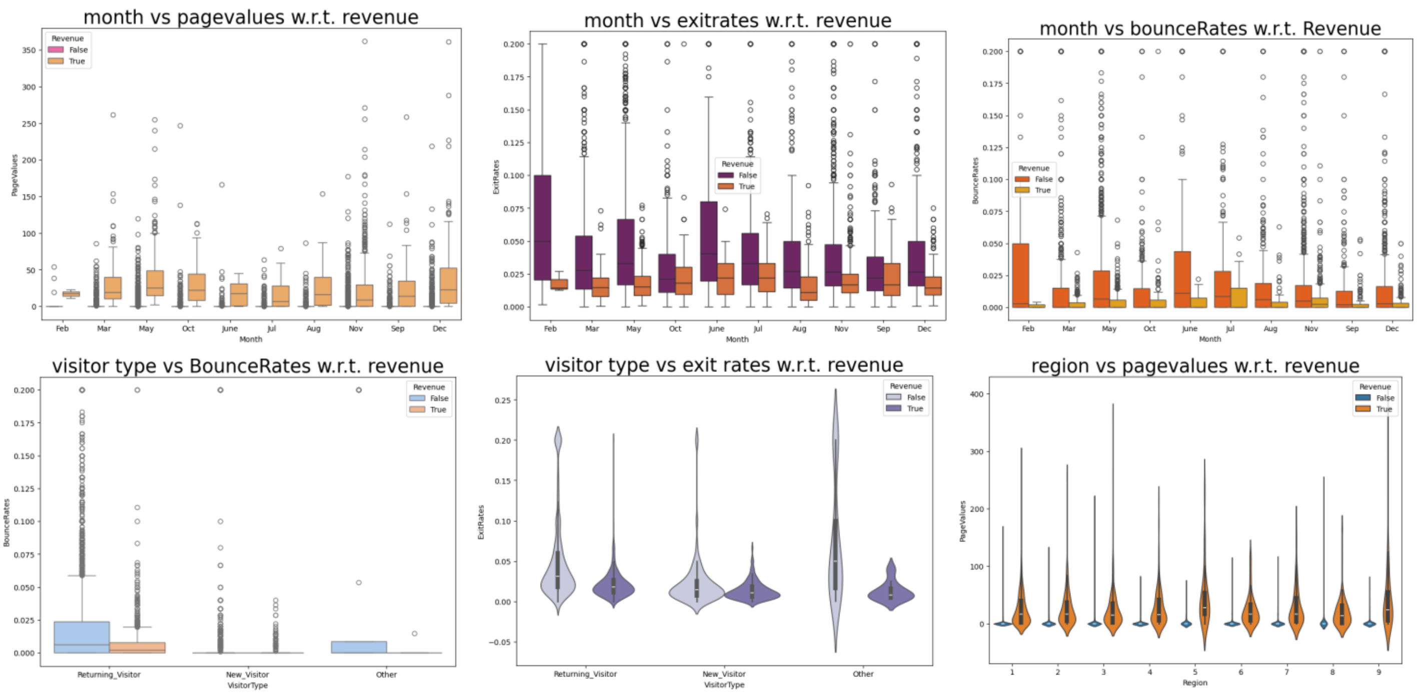
10. Region vs. Revenue

Observations:

1. Administrative Duration vs. Revenue: Exponentially distributed for both groups with many outliers in the non-purchased group.
2. Informational Duration vs. Revenue: Exponentially distributed with many outliers in the non-purchased group.
3. Product Related Duration vs. Revenue: Exponentially distributed with many outliers in the non-purchased group.
4. Exit Rates vs. Revenue: Normally distributed with many outliers in the non-purchased group.
5. Page Values vs. Revenue: Exponentially distributed; outliers exist in the purchased group, and PageValues strongly influence purchasing.
6. Bounce Rates vs. Revenue: Exponentially distributed with outliers in the non-purchased group; highly related to purchasing behavior.
7. Weekend vs. Revenue: No significant trend observed.
8. Traffic Type vs. Revenue: Different categories show varied influence on purchases, with categories 2, 7, 16, and 20 being highly influential.
9. Visitor Type vs. Revenue: New Visitors show a higher likelihood of purchasing.
10. Region vs. Revenue: All regions show similar revenue patterns with no major differences.

5. Multivariate Analysis:

Multivariate analysis helps in understanding how multiple variables are related, identifying patterns, and making predictions by analyzing their combined effects.

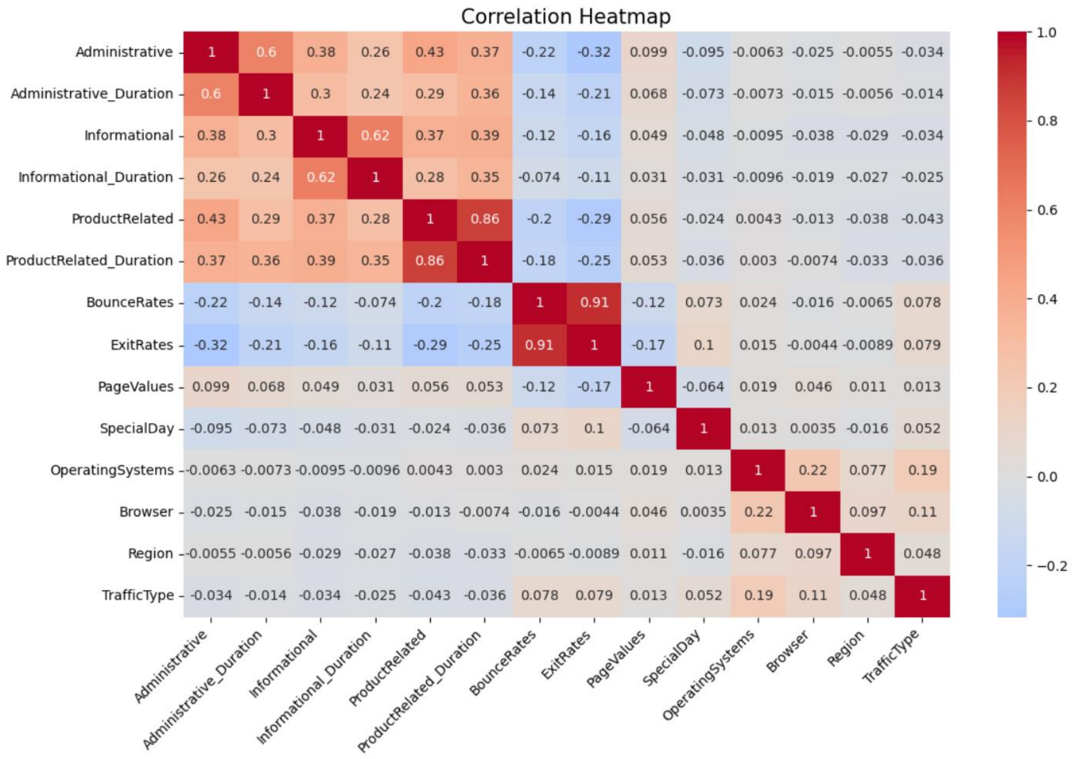


Observations:

1. Month vs. PageValues: PageValues are normally distributed during purchases, showing stable user behavior, but outliers suggest occasional high-value sessions.
2. Month vs. ExitRates: ExitRates are consistently normal across months, indicating stable exit patterns, with some outliers showing unusual exits.
3. Month vs. BounceRates: BounceRates are normal during purchase months but exponential in others, suggesting predictable engagement during purchases and lower interest in non-purchase months.
4. VisitorType vs. BounceRates: Returning visitors have consistent bounce behavior during purchases, while new users show variable behavior, indicating less engagement.
5. VisitorType vs. ExitRates: ExitRates are exponential for all visitor types, but returning visitors have more outliers, showing diverse navigation behavior.
6. Region vs. PageValues: PageValues are normally distributed across regions, indicating stable performance, with outliers showing unique regional user behavior.

6. Correlation Matrices

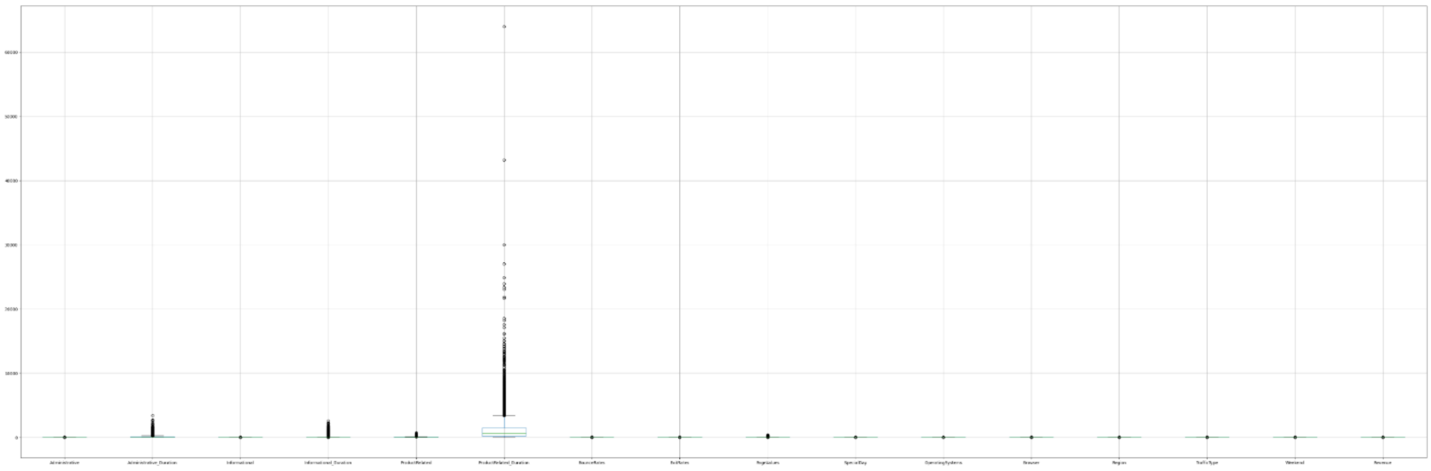
We generated a correlation heatmap to visualize the relationships between different features in the Online Shoppers Purchasing Intention Dataset. A strong positive correlation was observed between ProductRelated_Duration and ProductRelated pages (correlation coefficient > 0.75), suggesting that customers who spend more time on product-related pages tend to view more products. Additionally, there are moderate positive correlations between Administrative_Duration and Administrative pages, as well as between Informational_Duration and Informational pages, indicating consistent user behavior patterns across different page types.

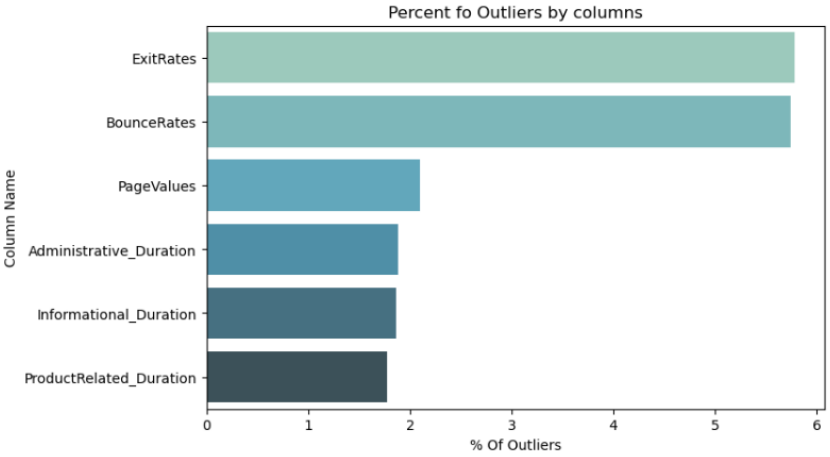

Correlation Heatmap

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
Administrative	1	0.6	0.38	0.26	0.43	0.37	-0.22	-0.32	0.099	-0.095	-0.0063	-0.025	-0.0055	-0.034
Administrative_Duration	0.6	1	0.3	0.24	0.29	0.36	-0.14	-0.21	0.068	-0.073	-0.0073	-0.015	-0.0056	-0.014
Informational	0.38	0.3	1	0.62	0.37	0.39	-0.12	-0.16	0.049	-0.048	-0.0095	-0.038	-0.029	-0.034
Informational_Duration	0.26	0.24	0.62	1	0.28	0.35	-0.074	-0.11	0.031	-0.031	-0.0096	-0.019	-0.027	-0.025
ProductRelated	0.43	0.29	0.37	0.28	1	0.86	-0.2	-0.29	0.056	-0.024	0.0043	-0.013	-0.038	-0.043
ProductRelated_Duration	0.37	0.36	0.39	0.35	0.86	1	-0.18	-0.25	0.053	-0.036	0.003	-0.0074	-0.033	-0.036
BounceRates	-0.22	-0.14	-0.12	-0.074	-0.2	-0.18	1	0.91	-0.12	0.073	0.024	-0.016	-0.0065	0.078
ExitRates	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1	-0.17	0.1	0.015	-0.0044	-0.0089	0.079
PageValues	0.099	0.068	0.049	0.031	0.056	0.053	-0.12	-0.17	1	-0.064	0.019	0.046	0.011	0.013
SpecialDay	-0.095	-0.073	-0.048	-0.031	-0.024	-0.036	0.073	0.1	-0.064	1	0.013	0.0035	-0.016	0.052
OperatingSystems	-0.0063	-0.0073	-0.0095	-0.0096	0.0043	0.003	0.024	0.015	0.019	0.013	1	0.22	0.077	0.19
Browser	-0.025	-0.015	-0.038	-0.019	-0.013	-0.0074	-0.016	-0.0044	0.046	0.0035	0.22	1	0.097	0.11
Region	-0.0055	-0.0056	-0.029	-0.027	-0.038	-0.033	-0.0065	-0.0089	0.011	-0.016	0.077	0.097	1	0.048
TrafficType	-0.034	-0.014	-0.034	-0.025	-0.043	-0.036	0.078	0.079	0.013	0.052	0.19	0.11	0.048	1

7. Outliers

The analysis reveals significant outliers across several key metrics. Page Values exhibit notable upper-range outliers, indicating exceptional engagement from certain visitors. Duration-related variables, such as ProductRelated_Duration and Administrative_Duration, show multiple outliers, reflecting diverse user engagement patterns. Additionally, Bounce Rates and Exit Rates display outliers primarily in the upper quartile, representing sessions with unusually high abandonment rates.

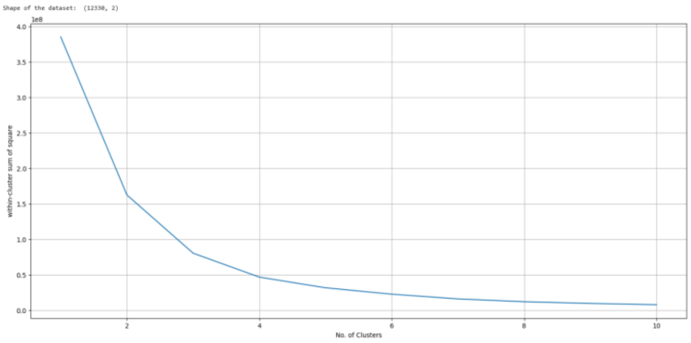




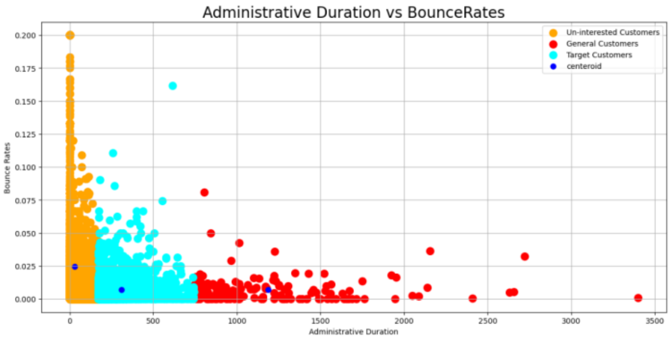
8. Clustering Analysis

The Elbow Method was used to determine the optimal number of clusters by plotting the within-cluster sum of squares against different values of k. The bend in the plot indicates the appropriate number of clusters.

Administrative Duration vs Bounce Rates:

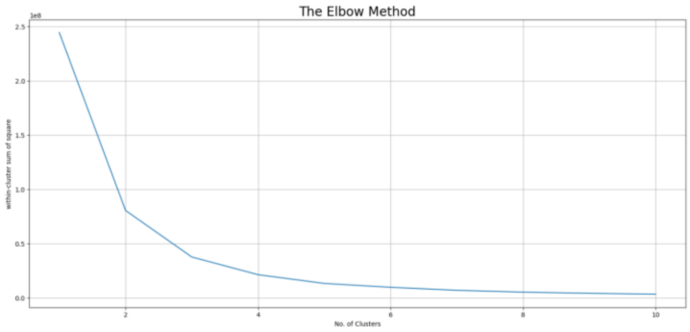


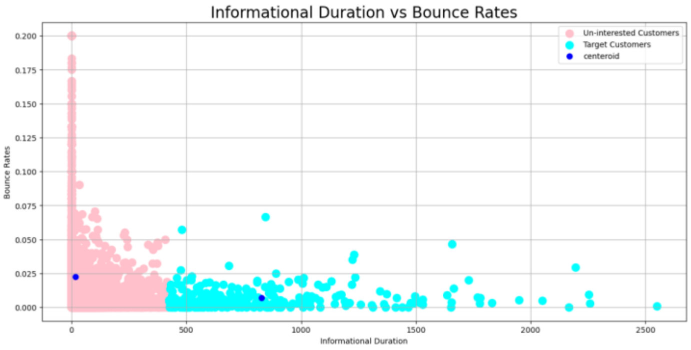
```
Shape of the dataset:  (12330, 2)
```



The analysis identified 3 optimal clusters, with longer administrative durations correlating to lower bounce rates. Customers in the shortest-duration cluster showed the highest bounce rates, indicating a need for strategies to engage this group.

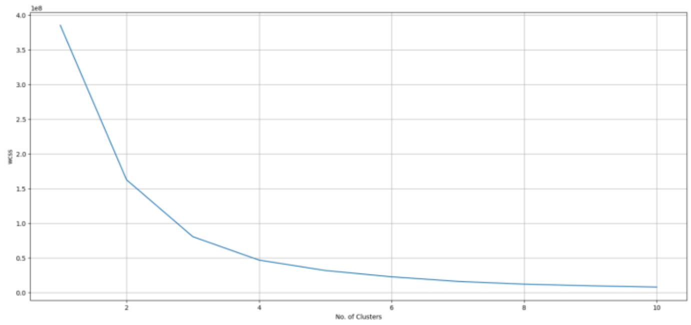
Informational Duration vs Bounce Rates:

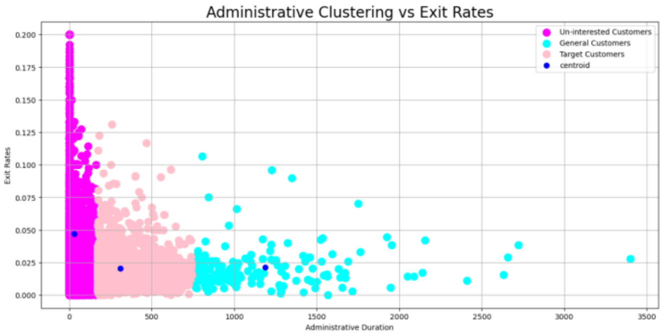




2 optimal clusters were found, showing that customers who spend more time on informational pages are less likely to bounce. The shortest-duration cluster exhibited the highest likelihood of bouncing.

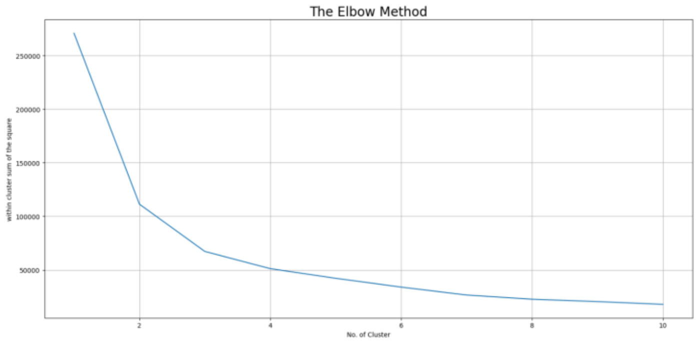
Administrative Duration vs Exit Rates:

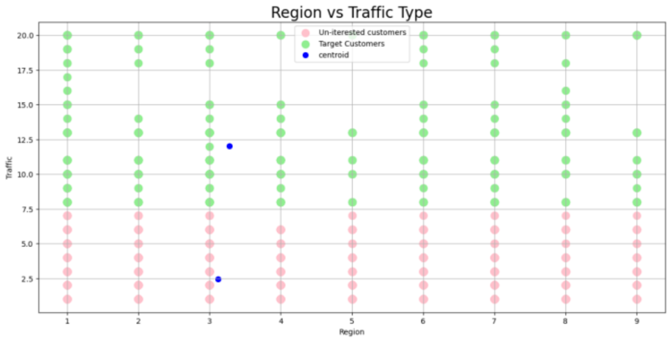




3 clusters were identified, revealing that customers spending more time on administrative activities are less likely to exit the website. The shortest-duration cluster displayed the highest exit rates.

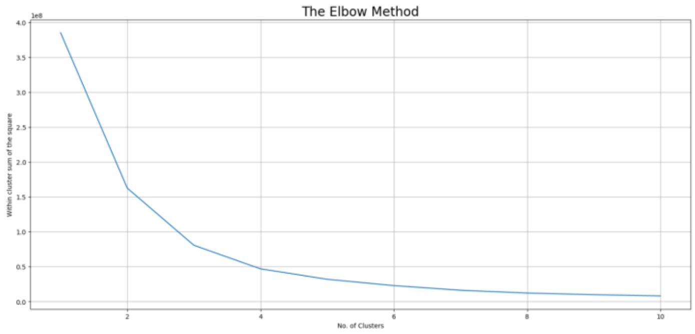
Region vs Traffic Type:

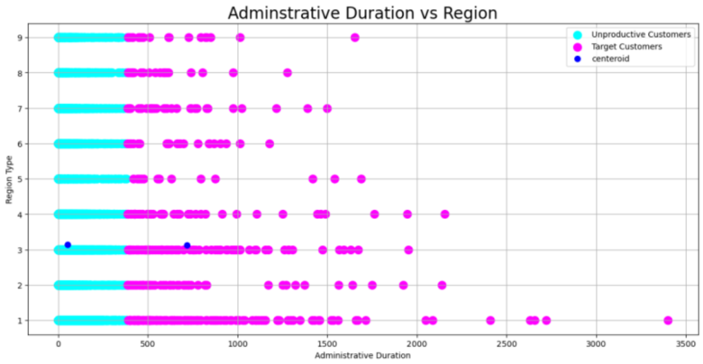




The clustering analysis determined two optimal clusters, indicating that regions 2, 4, and 5 experience lower traffic compared to other regions, which may suggest regional differences in user engagement.

Administrative Duration vs Region





2 clusters were identified, showing that customers with longer administrative durations are less likely to come from regions 2 and 4. This suggests regional factors may influence user behavior.

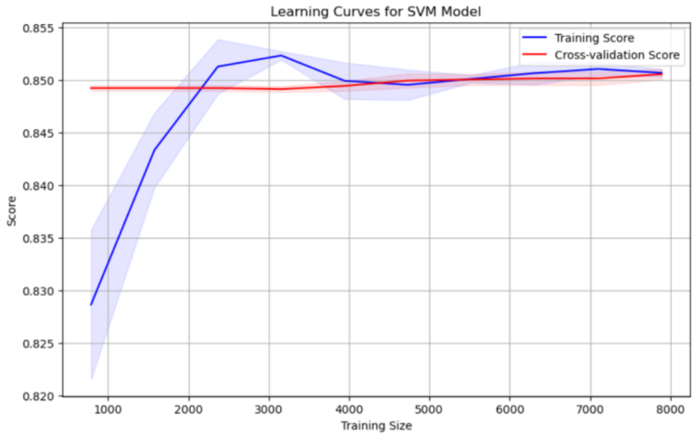
9. Why are these datasets interesting?

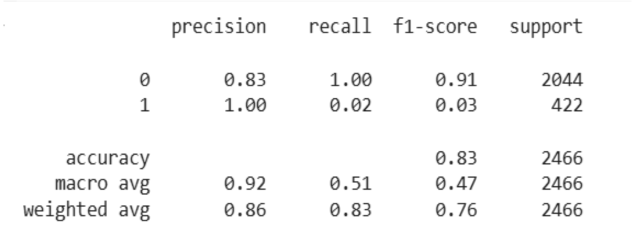
We find this dataset interesting because they provide valuable insights into user behavior and decision-making processes, enabling practical applications in real-world scenarios. For instance, the "Online Shoppers Purchasing Intention Dataset" captures detailed information about customer interactions with an e-commerce platform, such as browsing patterns, time spent on the site, and traffic sources. This data allows for the development of predictive models to identify factors influencing purchase decisions, helping businesses optimize marketing strategies like personalization, promotions, and pricing. The insights gained from analysing such datasets can significantly impact customer acquisition, retention, and revenue generation, making them highly relevant and engaging for study.

10. Machine Learning Model

Support Vector Machine:

We implemented a Support Vector Machine (SVM), a supervised learning algorithm that classifies data by finding an optimal hyperplane separating different classes in the feature space. Using the sklearn.svm. SVC class, we trained the model with default parameters, including the RBF kernel, C=1.0 and gamma='scale'\text{gamma='scale'}gamma=scale'. The performance was evaluated using metrics such as accuracy, precision, recall, and F1 score to ensure a comprehensive assessment of the model. Additionally, a learning curve was plotted to analyze the model’s performance across different training set sizes, helping identify potential overfitting or underfitting issues

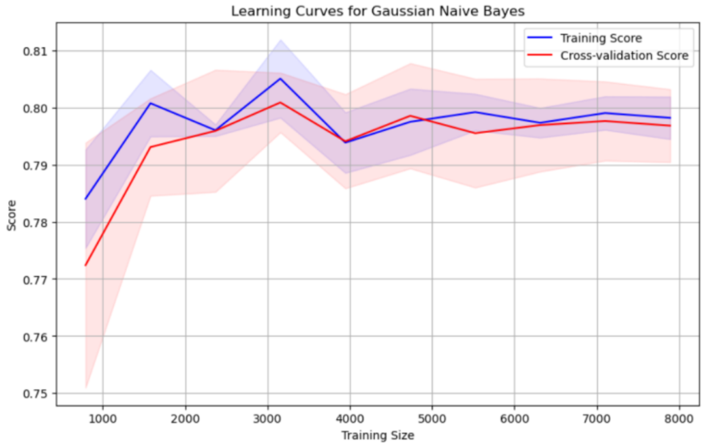


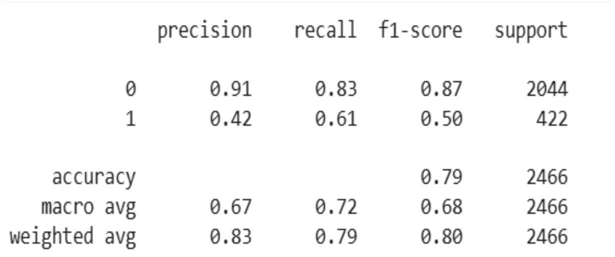


```
              precision    recall  f1-score   support

           0       0.83      1.00      0.91      2044
           1       1.00      0.02      0.03       422

    accuracy                           0.83      2466
   macro avg       0.92      0.51      0.47      2466
weighted avg       0.86      0.83      0.76      2466
```

Naive Bayes:

We implemented Gaussian Naive Bayes, a supervised classification algorithm based on Bayes' Theorem that assumes feature independence and models continuous data using a Gaussian distribution. Using the sklearn.naive_bayes.GaussianNB class with default parameters (priors=None, var_smoothing=1e-09), the model was trained on the data. The priors=None setting allows class priors to be learned from the data, while var_smoothing=1e-09 prevents numerical instability. The model's performance was evaluated using accuracy, precision, recall, and F1 score, and a learning curve was plotted to assess performance across different training set sizes, identifying potential overfitting or underfitting. This efficient probabilistic approach serves as a strong baseline for classification tasks.

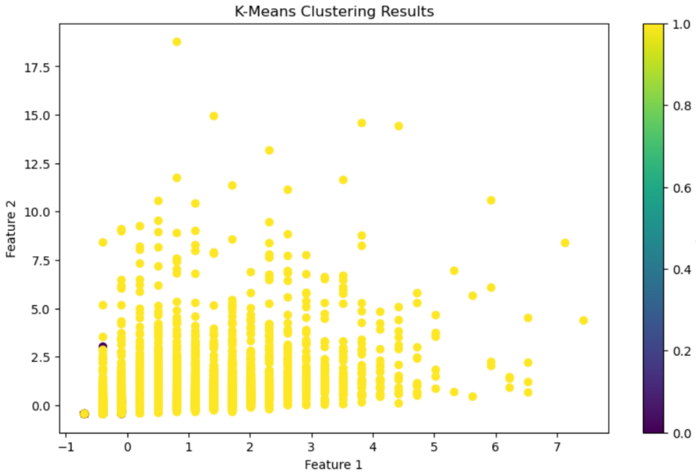




```
              precision    recall  f1-score   support

           0       0.91      0.83      0.87      2044
           1       0.42      0.61      0.50       422

    accuracy                           0.79      2466
   macro avg       0.67      0.72      0.68      2466
weighted avg       0.83      0.79      0.80      2466
```

K-Means Clustering:

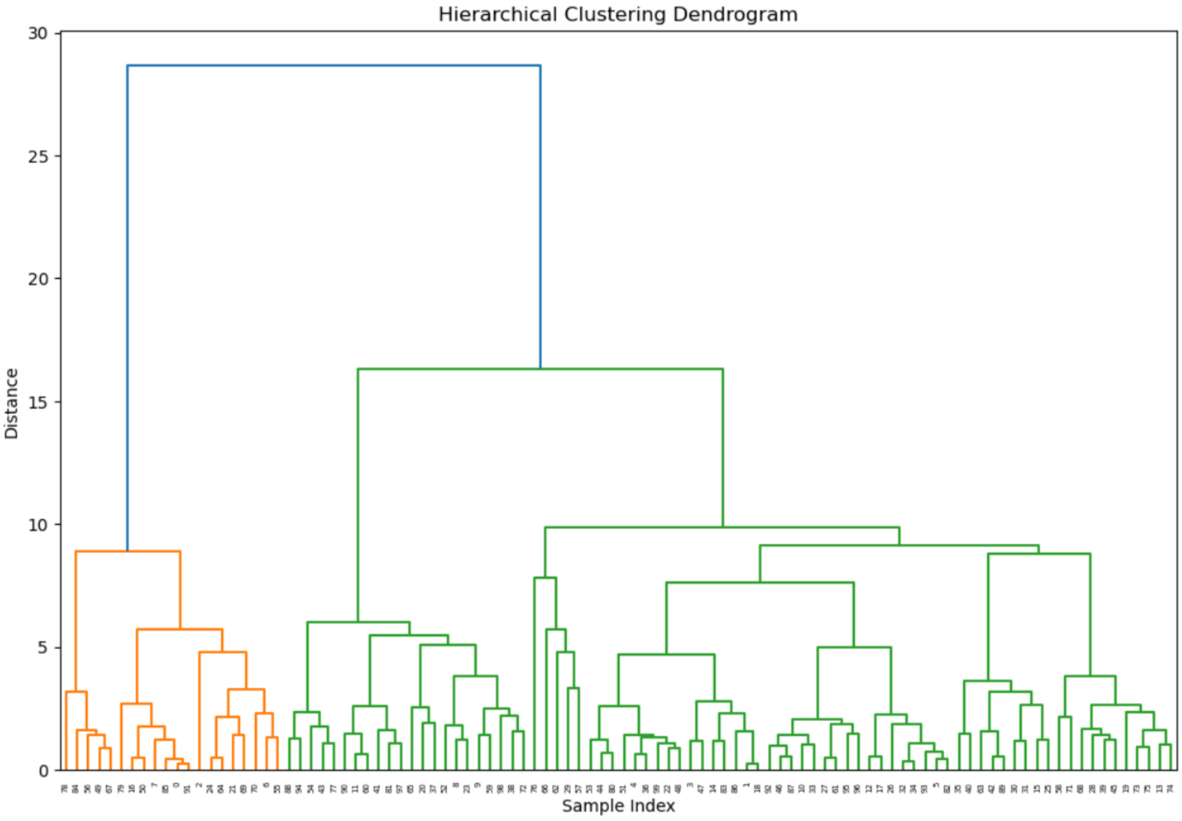
We used K-Means Clustering, an unsupervised machine learning algorithm that groups data into clusters based on similarity between features. Using k=2k=2 clusters with the KMeans class from scikit-learn, we selected numeric features from the dataset and scaled them using StandardScaler to ensure fair comparison between all features. The K-Means algorithm divided the data into two clusters, and we visualized the results with a scatter plot. This visualization helped identify patterns and relationships in the data, focusing on customer behavior based on session duration and page interaction metrics.



Hierarchical Clustering:

We implemented Hierarchical Clustering, an unsupervised machine learning technique that groups data based on similarity without predefining the number of clusters. Specifically, we used **Agglomerative Hierarchical Clustering**, a bottom-up approach where each data point starts as its own cluster and clusters are merged step-by-step based on similarity until the desired number of clusters is achieved. In this case, we set n=2 clusters using the Agglomerative

Clustering class from scikit-learn. To visualize the clustering process, we plotted a dendrogram using the first 100 samples of scaled data and the 'ward' method, which minimizes the variance within clusters at each step of merging. The dendrogram illustrated the hierarchical merging of clusters and the distance between them, providing insights into how clusters were formed.



11. Conclusion

Among all the algorithms tested, Support Vector Machine (SVM), Naive Bayes, K-Means Clustering, and Hierarchical Clustering showed different performance patterns. The SVM model performed well in identifying "no purchase" behavior but struggled with detecting "purchase" events due to class imbalance, achieving only 0.02 recall for the "purchase" class despite an overall accuracy of 83%.Naive Bayes, while slightly less accurate, proved to be the most practical choice due to its simplicity, speed, and balanced performance across classes.K-Means and Hierarchical Clustering provided meaningful insights by uncovering patterns and segments within the data, contributing to strategic decision-making.

In conclusion, Naive Bayes emerged as the most effective and reliable choice for this classification task, while the clustering models offered valuable behavioral insights.

12. References

- https://www.kaggle.com/datasets/henrysue/online-shoppers-intention
- https://archive.ics.uci.edu/
- https://scikit-learn.org/1.5/index.html
- https://chat.openai.com/