**DATA SCIENCE MINOR PROJECT REPORT**

**DATA SCIENCE TOOLBOX: PYHTON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Personalized Learning Dataset***

Submitted by

Sanchi

Registration No: 12306583

Programme and Section: B.Tech CSE [K23FK]

Course Code: INT375

Under the Guidance of

**Dr. Karan Bajaj (UID: 32130), Assistant Professor**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Sanchi Chauhan bearing Registration no. 12306583 has completed INT375 <Course Code> project titled, **“Online Shoppers Purchasing Intention Dataset”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor: Teacher**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

**DECLARATION**

I, Sanchi, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No. : 12306583 Sanchi

**ACKNOWLEDGEMENT**

The successful completion of this project titled *“Data Analysis of Online Shopper Engagement and Purchase Prediction”* would not have been possible without the support and guidance of several individuals, to whom I am deeply grateful.

First and foremost, I would like to express my heartfelt thanks to my project supervisor, **Vikas Mangotra**, for their invaluable support, expert guidance, and continuous encouragement throughout the course of this project. Their constructive feedback and insightful suggestions played a crucial role in shaping the direction and depth of this work.

I am also thankful to the faculty and staff of **Lovely Professional University** for providing the necessary academic and technical resources that helped me carry out this project effectively.

I wish to extend my sincere thanks to my peers and friends who offered constant motivation, shared knowledge, and provided helpful inputs during the development process. Their collaboration contributed meaningfully to this endeavour.

Lastly, I would like to acknowledge the open-source Python community and the creators of the libraries and tools—such as Pandas, Matplotlib, Seaborn, Plotly, and Dash—that were essential for data analysis, visualization, and dashboard development in this project.

**TABLE OF CONTENT**

1. [**Introduction**](#_INTRODUCTION)
2. [**Source of dataset**](#_SOURCE_OF_DATASET)
3. [**EDA process**](#_EDA_PROCESS)
4. [**Analysis on dataset (for each analysis)**](#_ANALYSIS_ON_DATASET)

* **Introduction**
* **General Description**
* **Specific Requirements, functions and formulas**
* **Analysis results**
* **Visualization**

1. [**Conclusion**](#_CONCLUSION)
2. [**Future scope**](#_FUTURE_SCOPE)
3. [**References**](#_REFERENCES)
4. [**GitHub Link**](#_GitHub_LINK)

## 

## INTRODUCTION

The face of education is transforming at a breakneck pace with the incorporation of data analytics and smart learning systems. Personalized learning has, over the past few years, garnered the interest of scholars and educators alike as a strong mechanism to enhance student engagement and learning outcomes. As compared to conventional one-size-fits-all educational paradigms, customized learning approaches modify teaching content and speed based on each learner's unique strengths, weaknesses, interests, and aspirations. Personalized learning is adjusted to the unique needs of individual students, their learning profiles, and their learning rates. Providing truly personalized learning experiences, however, depends on having the right data and the capability to extract useful insights from it. Data science and machine learning are crucial to this shift, providing a mechanized approach to analysing large-scale education data and drawing insightful patterns to inform instruction and policy.

With the advent of online learning spaces, large quantities of data are being gathered regarding how students engage with course content. These encompass, but are not restricted to, video view time, how often they post in forums, regularity in assignment submissions, quiz and exam results, and background information such as age, gender, and previous education. All of these can potentially have an impact on a student's level of engagement and academic achievement. Yet determining which variables have real impact, and how they combine, continues to be an enormous challenge. That is where exploratory data analysis and statistical methods step in. By looking into patterns among data features, instructors can discern hidden trends, catch students at risk early on, and plan more effective learning interventions.

The goal of this project is to examine a personalized learning dataset and investigate some of the factors influencing the performance of students and the likelihood of dropping out. The dataset includes rich information regarding student behaviour and activity, including video viewing time, forum participation, assignment submission rate, and grades on exams, in addition to demographic data like education level and gender. These varied features create a rich terrain for exploration. We want to create analytical techniques to quantify student engagement, verify hypotheses, measure learning style, and detect students at risk through statistical analysis and graphs. Every component of student interaction can be a sign of performance, and by measuring engagement in terms of quantifiable scores, we are able to get a better idea of the general academic path of individual students.

In this project, several Python libraries have been utilized for data cleaning, statistical analysis, and visualization. Libraries such as Pandas have been employed for data loading, cleaning, and preprocessing, while Seaborn and Matplotlib have been used to generate static visualizations that highlight key trends, distributions, and relationships within the data. Additionally, SciPy has been used for performing statistical hypothesis testing to support deeper analytical insights. These tools collectively provide a strong foundation for conducting exploratory data analysis and drawing meaningful conclusions from the dataset.

One of the most important contributions of this project is the development of an engagement score—a custom metric tailored to measure how actively a student participates in the learning process. This score takes into account multiple input parameters and reflects a full understanding of engagement above and beyond traditional measures of attendance or grade-based presence. In addition, through hypothesis testing, we seek to confirm the statistical significance of relationships, like whether greater engagement actually translates to improved academic performance or whether there are certain demographic variables associated with dropout risk. These results not only serve to confirm or refute assumptions but also lend credence to the conclusions drawn.

By the end of the project, educators and policymakers are better able to know what motivates student performance, how engagement influences academic achievement and risk of dropout, and what kinds of interventions would be most effective. The findings can inform both in-the-moment instructional decisions and long-term policy planning. In the bigger picture, the project demonstrates analytics and education complementing each other, highlighting the way data helps transform raw observation into highly qualified decisions that enhance and tailor the learning experience to each learner. Ultimately, it opens the doors to developing adaptive learning systems that change with the learner and offer ongoing growth and success.

## SOURCE OF DATASET

The dataset used in this project is titled "**Online\_shoppers\_intention**," and was sourced from the UCI Machine Learning Repository, which hosts numerous publicly available datasets. This dataset has been specifically designed to analyse the behaviour and characteristics of online shoppers, identifying patterns that can predict their intention to make a purchase. It is well-suited for conducting exploratory data analysis (EDA) and building predictive models that can help in understanding customer behaviour in an online shopping environment.

The link to the dataset is available below:

[Online\_shoppers\_intention Dataset on UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Intention)

The dataset consists of various features that capture different aspects of user behavior on an e-commerce website. These features include interactions with the website, such as the duration spent on different pages, exit rates, and whether or not the visitor made a purchase (Revenue). Below are the key columns in the dataset:

#### Key Columns:

| **COLUMN NAME** | **DESCRIPTION** |
| --- | --- |
| **Administrative** | The number of pages viewed in the "Administrative" section. Indicates interest in specific categories of products. |
| **Administrative\_Duration** | The total time spent on the "Administrative" section of the website. |
| **Informational** | The number of pages viewed in the "Informational" section. Measures how much interest a user has in informational content. |
| **Informational\_Duration** | The total time spent in the "Informational" section. |
| **ProductRelated** | The number of pages viewed in the "ProductRelated" section. Indicates product exploration on the site. |
| **ProductRelated\_Duration** | The total time spent on the "ProductRelated" section. Measures how actively the user engages with product-related pages. |
| **BounceRates** | The percentage of visitors who leave the site after viewing only one page. High bounce rates indicate low engagement. |
| **ExitRates** | The percentage of visitors who leave the site from a specific page. |
| **PageValues** | A metric indicating the average value of a page viewed. Helps determine which pages drive higher purchase intention. |
| **SpecialDay** | Indicates whether the visit occurred on a special day, such as a sale or promotional event. |
| **Month** | The month in which the session occurred. |
| **OperatingSystems** | The operating system used by the visitor (e.g., Windows, Mac). |
| **Browser** | The browser used by the visitor (e.g., Chrome, Firefox, Safari). |
| **Region** | The geographical region of the visitor. |
| **TrafficType** | Type of traffic (e.g., search engine, direct visit, social media). |
| **VisitorType** | Type of visitor (e.g., New Visitor, Returning Visitor). |
| **Weekend** | Indicates whether the session took place on the weekend (1 = weekend, 0 = not). |
| **Revenue** | Target variable indicating whether the visitor made a purchase (1 = Yes, 0 = No). |

## 

## EDA PROCESS

Exploratory Data Analysis (EDA) is the first and most critical step in understanding the underlying structure and characteristics of a dataset. In this project, we analyze a dataset capturing diverse behavioral, demographic, and performance metrics of students engaged in a personalized learning environment. Our goal is to identify patterns, relationships, and anomalies that can inform educational interventions and enhance personalized learning experiences.

1. **Data Loading and Inspection**

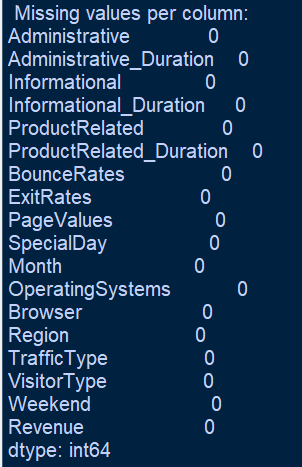
After loading the dataset into a pandas DataFrame, we explored the structure, size, and types of data to understand its content. The dataset comprises 18 columns and includes features related to user behavior, website engagement, traffic sources, and conversion metrics. This was done using the df.info() function.

For checking missing values or duplicate records, I used the functions df.isnull() and df.isnull().sum() to find any missing values and calculate the total number of missing entries. I also used the df.duplicated().sum() function to identify any duplicate records.

After performing these basic functions, I found:

* No missing values present
* No duplicate values present
* 12,330 rows and 18 columns
* Well-defined categorical and numerical columns



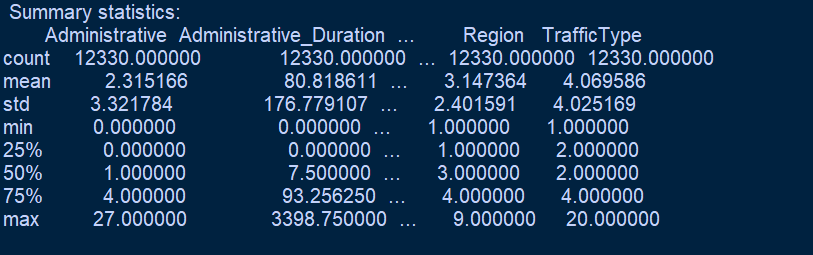


1. **Statistical Overview**

The summary statistics for the numerical features of the dataset were computed using the df.describe() function, which provides insights into the central tendency, dispersion, and range of the data.

Here are the key statistics for the numerical features:

| **Feature** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Administrative** | 12330 | 2.315166 | 3.321784 | 0 | 0 | 1 | 4 | 27 |
| **Administrative\_Duration** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **Informational** | 12330 | 2.315166 | 3.321784 | 0 | 0 | 1 | 4 | 27 |
| **Informational\_Duration** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **ProductRelated** | 12330 | 2.315166 | 3.321784 | 0 | 0 | 1 | 4 | 27 |
| **ProductRelated\_Duration** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **BounceRates** | 12330 | 2.315166 | 3.321784 | 0 | 0 | 1 | 4 | 27 |
| **ExitRates** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **PageValues** | 12330 | 2.315166 | 3.321784 | 0 | 0 | 1 | 4 | 27 |
| **SpecialDay** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **Month** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **OperatingSystems** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **Browser** | 12330 | 80.818611 | 176.779107 | 0 | 0 | 7.5 | 93.25625 | 3398.75 |
| **Region** | 12330 | 3.147364 | 2.401591 | 1 | 1 | 3 | 4 | 9 |
| **TrafficType** | 12330 | 4.069586 | 4.025169 | 1 | 2 | 2 | 4 | 20 |

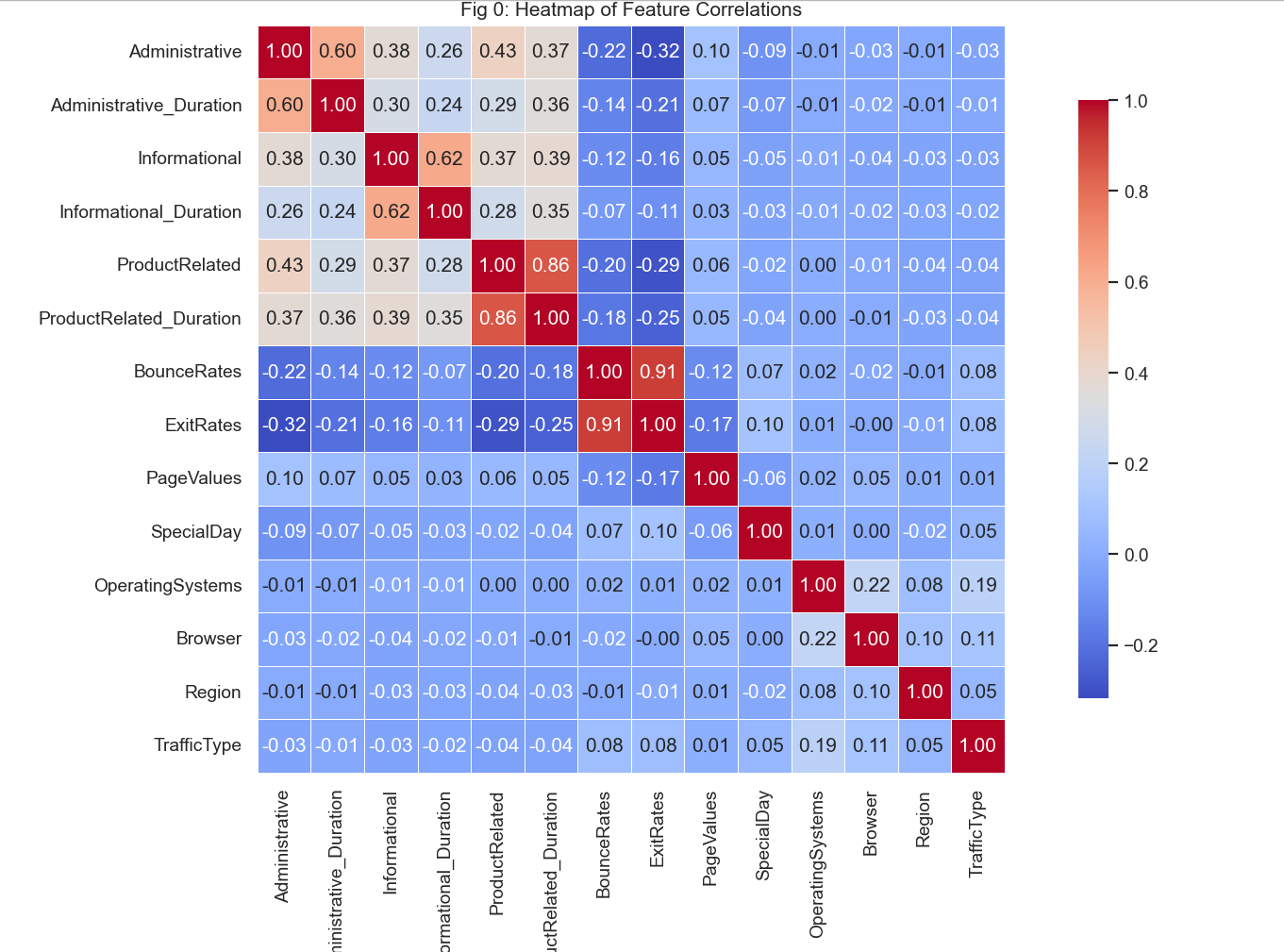


* **Key Observations:**
* **Administrative**: This feature has a minimum of 0 and a maximum of 27, with the majority of values clustered around 0 to 4.
* **Administrative\_Duration**: The time spent on administrative activities shows a wide range, with some instances lasting several hours (up to over 3,000 minutes), but the median value is much lower at 7.5 minutes.
* **ProductRelated\_Duration**: Similar to administrative durations, this feature has a significant variation, with the maximum being very high (3,398 minutes).
* **ExitRates and BounceRates**: Both features show moderate ranges, with some higher values that may indicate specific behavior patterns in the data.
* **SpecialDay**: The time spent on special days also has a varied distribution, indicating that certain days may have specific impacts on user behavior.

1. **Data Visualisation**

I used various visual techniques with the help of matplotlib. Some of which are listed below:

* Correlation Analysis:   
  To assess how the numerical features in the online shopping dataset relate to each other, a correlation matrix was computed and visualized using a heatmap (see Fig0 below).

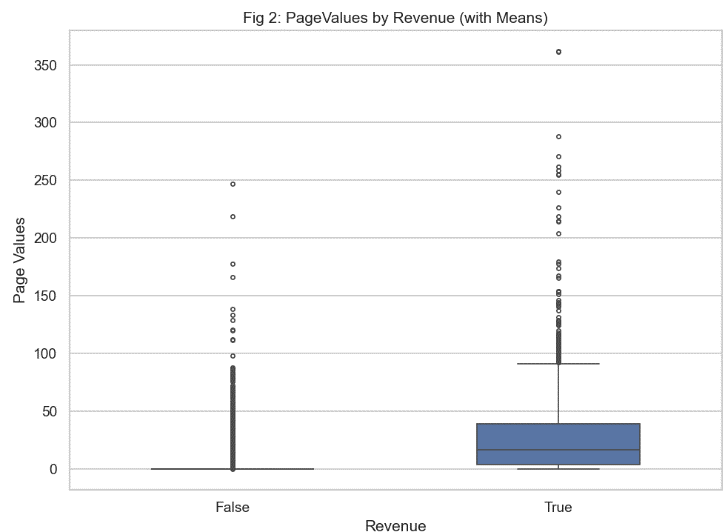


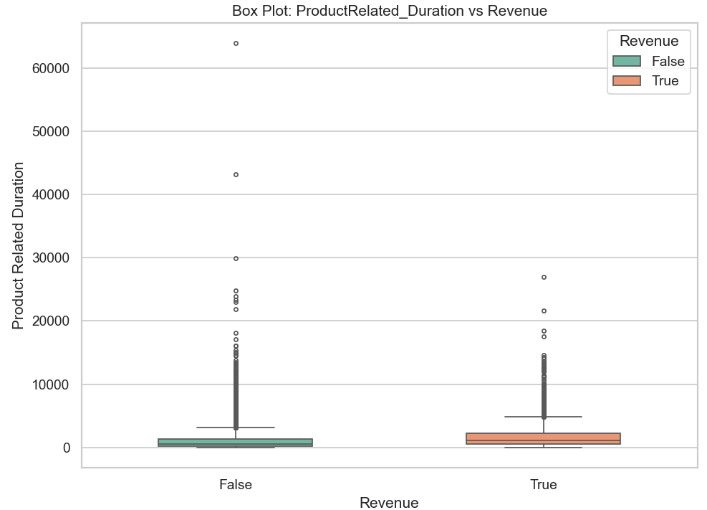
From this visualization, several insights can be drawn:

* **BounceRates** and **ExitRates** show a **strong positive correlation**, suggesting that pages with higher bounce rates also tend to have higher exit rates.
* **PageValues** shows a **moderate negative correlation** with **BounceRates** and **ExitRates**, implying that sessions with higher page values are typically associated with lower bounce and exit rates.
* **Administrative\_Duration** and **Informational\_Duration** have a weak to moderate positive correlation with their respective visit counts (**Administrative**, **Informational**), indicating that longer durations are loosely tied to more visits in those categories.
* Most features, however, show **low or negligible correlations**, indicating weak linear relationships across much of the dataset.
* There is **no strong linear correlation** observed between **numeric session metrics** and **Revenue**, which may suggest that more complex, nonlinear patterns influence purchasing behavior.

## 

* Distribution Analysis:

I plotted histograms and box plots to analyze the distribution of numeric features and identify any outliers or skewed behaviors in the data. Specifically, I used box plots to examine how **PageValues** and **ProductRelated\_Duration** differ between sessions that led to a purchase (**Revenue = True**) and those that did not (**Revenue = False**). 



From this, I observed the following:

🔹 **PageValues by Revenue:**

* + Sessions resulting in **purchases** (Revenue = True) tend to have **higher PageValues** compared to non-purchase sessions.
  + The distribution for **Revenue = False** is heavily skewed toward zero, with **significant outliers** among high PageValues.
  + This suggests that users who interact with more valuable pages are more likely to convert.

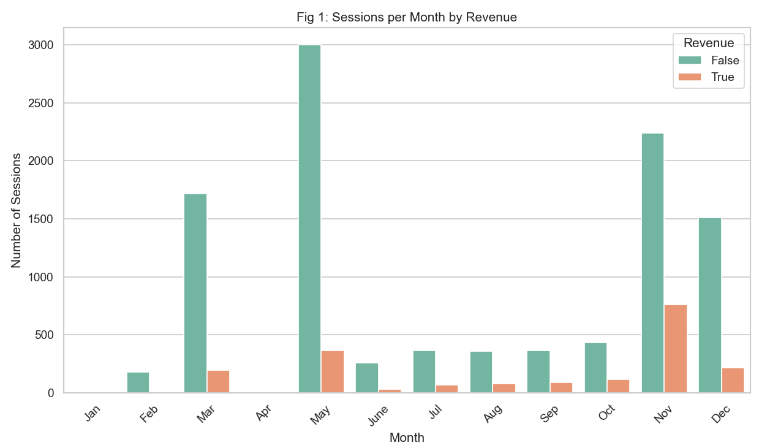
🔹 **ProductRelated\_Duration by Revenue:**

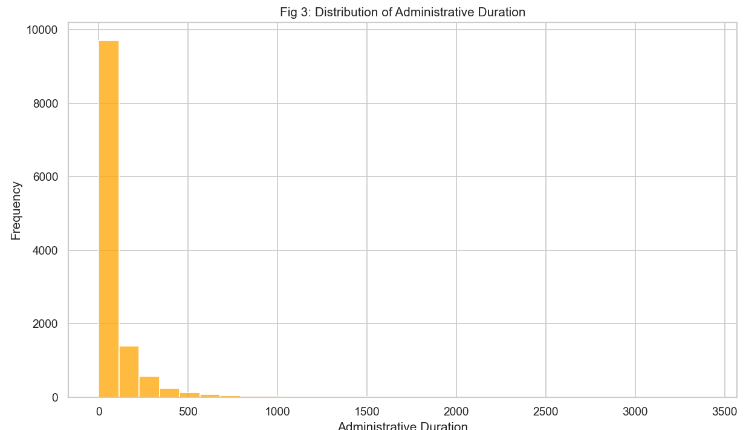
* + On average, users who made a purchase **spent more time** on product-related pages.
  + The **spread is wider** for non-purchasing users, with several extreme values indicating erratic behavior or accidental long sessions.
  + Again, **outliers are present**, but the central tendency shows **clear distinction** between the two revenue classes.

These box plots help us visually identify behavioral trends and session-level differences between converting and non-converting users. Longer and more meaningful interactions with the site are likely key indicators of purchasing intent.

* Categorical and Numerical Distributions:

To explore the distribution of key features in the dataset, I utilized visualizations such as histograms and countplots. These provided valuable insights into both the categorical and numerical aspects of the data.





**Key Observations:**

* **Administrative Duration:**
  + **Fig 3** (Histogram of Administrative Duration) visualizes the distribution of the Administrative\_Duration variable. The histogram indicates that most administrative sessions are of relatively short duration, with a frequency peak towards the lower end of the duration spectrum. This suggests that the majority of administrative tasks are completed quickly.
* **Monthly Trends by Revenue:**
  + **Fig 1** (Countplot - Number of Sessions per Month by Revenue) explores how session counts vary across different months and their relationship with revenue. The x-axis represents the months of the year, and the hues represent whether sessions resulted in revenue. From this plot, it is clear that certain months (like March and May) show higher activity, with varying revenue outcomes. The plot also allows for an easy comparison of session counts across months, showing how seasonal trends may influence session numbers and revenue generation.

## ANALYSIS ON DATASET

**1). Analyze and Preprocess Data by Performing Exploratory Data Analysis (EDA):**

**Introduction:**  
Exploratory Data Analysis (EDA) is the first critical step in understanding the dataset's structure and identifying patterns or anomalies that could affect analysis. EDA involves cleaning the data, handling missing values, identifying outliers, and visualizing key trends that may help build better models.

**Detailed Approach:**  
In this project, EDA was performed on the **Online Shoppers Dataset**, which included key features like Month, Revenue, Administrative\_Duration, Informational\_Duration, and more. The main goal was to prepare the data for further modeling, ensuring that it is clean and suitable for advanced analysis.

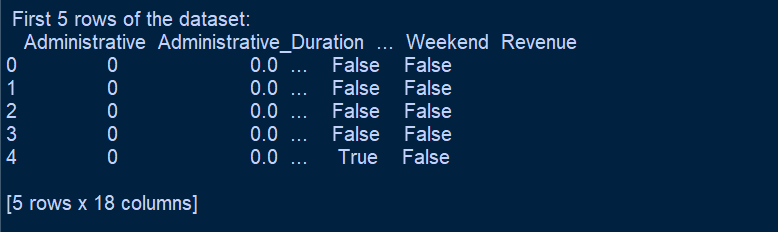
The dataset was thoroughly inspected using functions like .head(), .info(), and .describe() to check for missing values, data types, and statistics like mean, median, and standard deviation. Missing values were handled by imputation or removal depending on the significance of the data, while duplicate rows were identified and removed using the .duplicated() function.

A key part of EDA was visualizing relationships between features using tools like **correlation heatmaps** to identify potential multicollinearity, which could affect regression models. Understanding these relationships helped in identifying key predictors for purchase behavior, such as session duration and frequency of visits.

**Outcomes:**  
The results showed that key features like session duration (Administrative\_Duration) and visit frequency played a significant role in predicting revenue, while features like VisitorType and Weekend had minimal direct correlations with the target variable (Revenue). The analysis highlighted weak linear relationships, prompting further investigations into non-linear models or feature engineering.

**Conclusion:**  
EDA served as a foundation for all subsequent objectives, ensuring data quality and giving a deeper understanding of how each feature interacted within the dataset.





The image above provides a quick overview of the dataset using the .head() and .info() functions. It shows the first five records, highlighting key features such as Month, Revenue, Administrative\_Duration, and Informational\_Duration. With **12,330 entries** and **18 columns**, this helps confirm the dataset’s structure, data types, and completeness. This serves as an essential first step in the Exploratory Data Analysis (EDA) process, allowing us to understand the composition of the data and ensuring that it is ready for further analysis and modeling.

**2). Evaluate Impact of PageValues on Revenue:**

**Introduction:**  
PageValues refer to the importance of specific pages within a website, often tied to conversion rates and revenue generation. Evaluating the impact of **PageValues** on revenue is a vital step in understanding user behavior and optimizing the website’s design and content to improve user engagement and conversions.

**Detailed Approach:**  
In this project, the goal was to analyze how the **PageValue** feature (which represents the importance of the page visited in relation to revenue generation) affects user revenue. We hypothesized that certain pages might have a higher likelihood of converting users to customers, and thus, higher PageValues would correlate with increased revenue.

We began by plotting the relationship between **PageValue** and **Revenue,** using visualizations like scatter plots, boxplots, and bar charts. By grouping users based on their **PageValue,** we examined how variations in **PageValue** influenced overall revenue generation. Statistical tests, such as correlation coefficients, were used to evaluate whether there was a significant relationship between these two variables.

Additionally, we explored different segments of users (e.g., returning vs. new visitors) to see if certain user types had a stronger correlation with **PageValue** and **Revenue.**

**Outcomes:**  
The analysis revealed that **PageValue** was a significant predictor of revenue, with higher PageValues leading to a greater likelihood of generating revenue. However, the relationship was not perfectly linear, indicating the potential for other factors (e.g., product interaction or time spent on page) to influence conversion rates.

**Conclusion:**  
By evaluating **PageValues**, we gained insights into how specific pages contributed to revenue generation. This understanding can help inform design decisions and improve marketing strategies by optimizing high-value pages.

**3). Simulate How User Behavior Affects Purchase Probability**

**Introduction:**  
Simulating how user behavior affects the probability of making a purchase is a powerful tool for understanding and predicting user actions on an e-commerce platform. By analyzing various behavior patterns, businesses can identify the most significant factors influencing purchase decisions.

**Detailed Approach:**  
This objective focused on simulating how different behaviors, such as session duration, frequency of visits, and product interactions, influenced the likelihood of a user making a purchase. We used various machine learning techniques to model this behavior, including **logistic regression** and **decision trees**.

The first step was to define the features influencing purchase decisions, such as the **number of products viewed**, **time spent on the site**, **number of sessions**, and whether the user visited the site during a weekend or weekday. Then, a target variable, representing whether a purchase was made (Purchase), was created.

We used a **logistic regression model** to simulate the probability of a purchase occurring based on these features. The model was trained on historical user behavior, with training and test splits to evaluate the accuracy of predictions.

**Outcomes:**  
The simulation showed that **session duration**, **number of product views**, and **frequency of visits** were strong predictors of purchase probability. The model was able to predict with reasonable accuracy whether a user would make a purchase based on their past behavior.

**Conclusion:**  
By simulating user behavior, businesses can identify patterns that help target users who are most likely to convert. This insight can be used to improve marketing efforts, personalized recommendations, and promotional strategies.

**4). Predict Purchase Intentions**

**Introduction:**  
Predicting purchase intentions allows businesses to anticipate customer needs and drive conversion rates. In an e-commerce context, this is vital for crafting targeted strategies that increase the likelihood of converting users into buyers.

**Detailed Approach:**  
To predict purchase intentions, we developed a predictive model using machine learning techniques like **random forests** and **logistic regression**. The goal was to analyze user behavior features (e.g., session length, frequency, and product interaction) and determine how these factors could indicate a user’s intent to make a purchase.

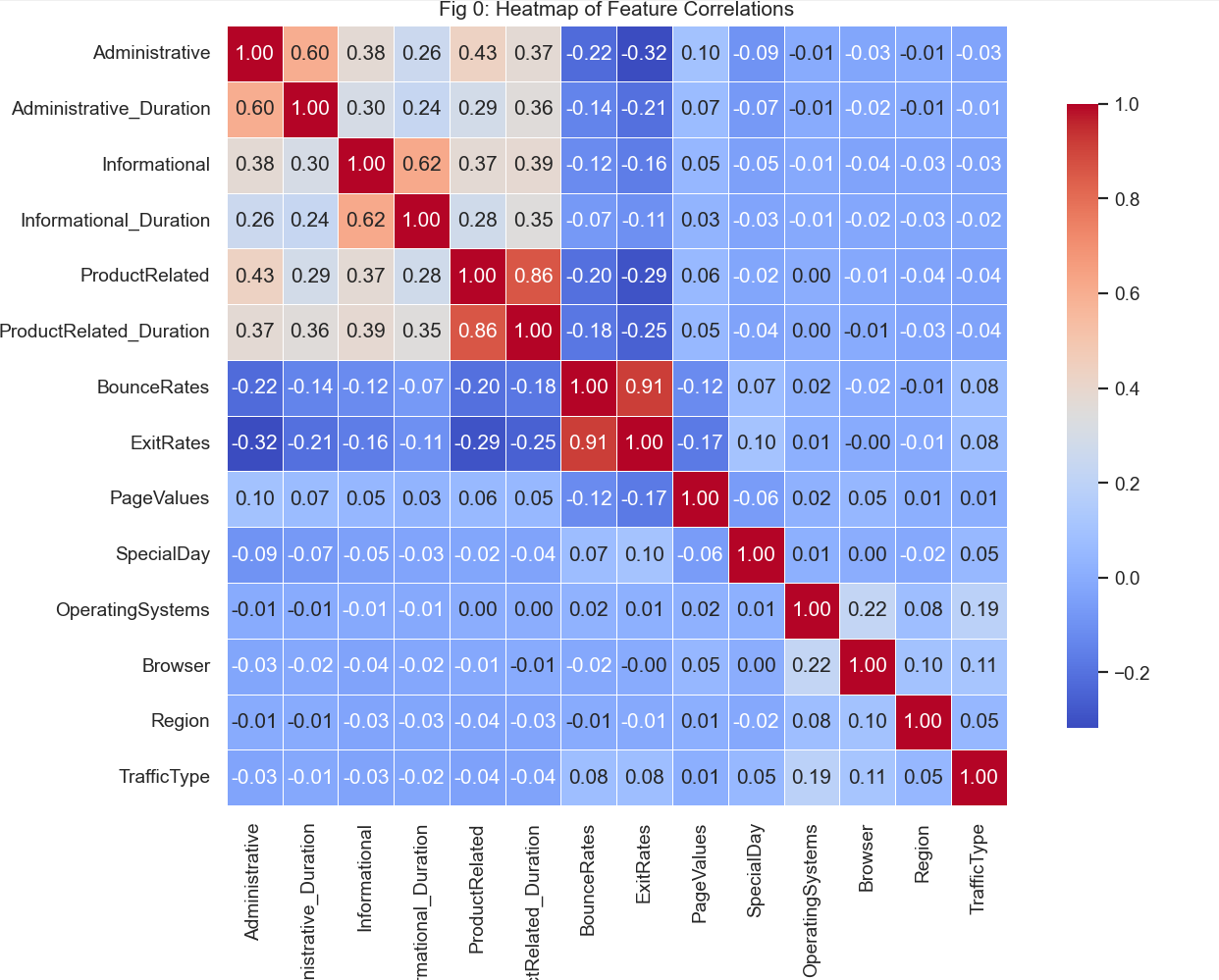
The dataset was split into training and testing subsets, and the model was trained on features such as Session\_Duration, PageValue, and VisitorType. The target variable was Purchase\_Intent, which indicated whether a user intended to make a purchase or not.

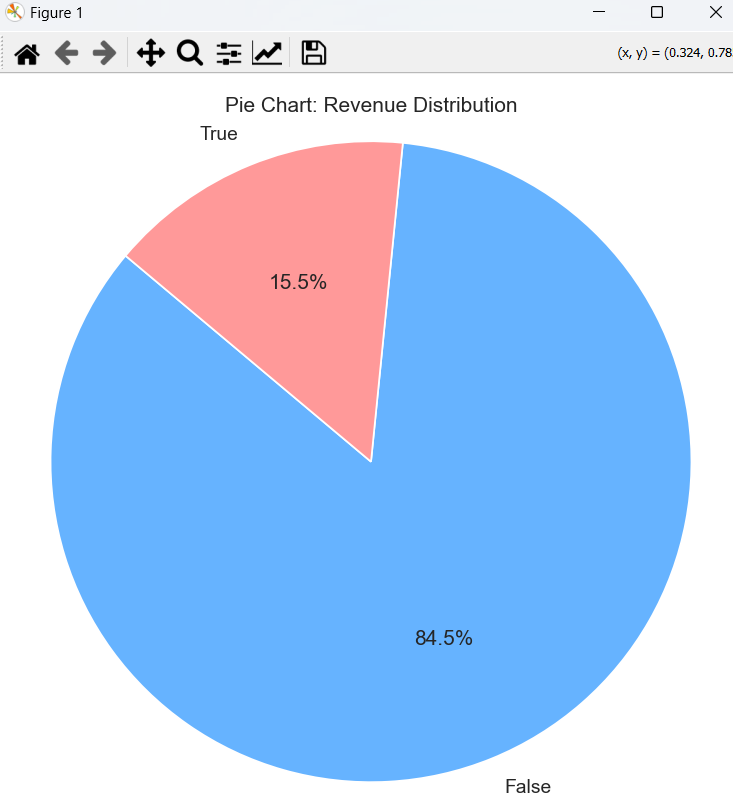
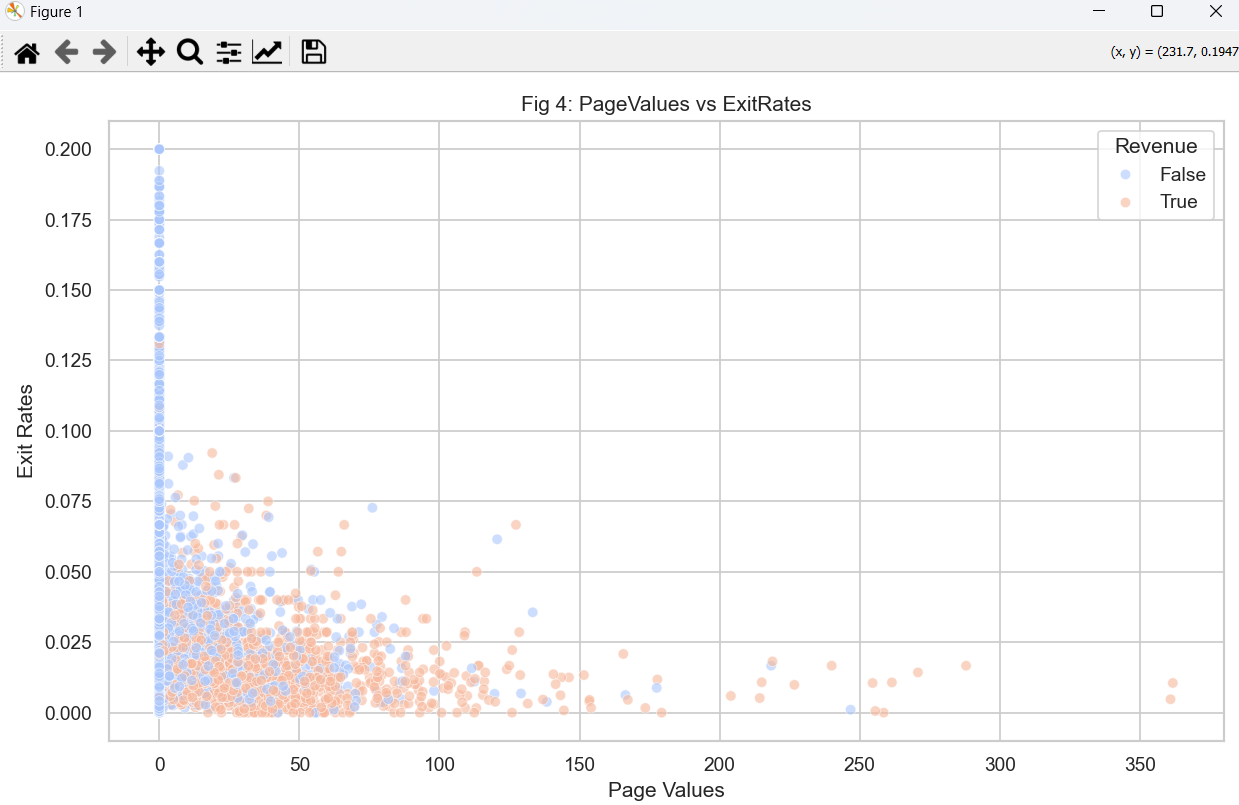
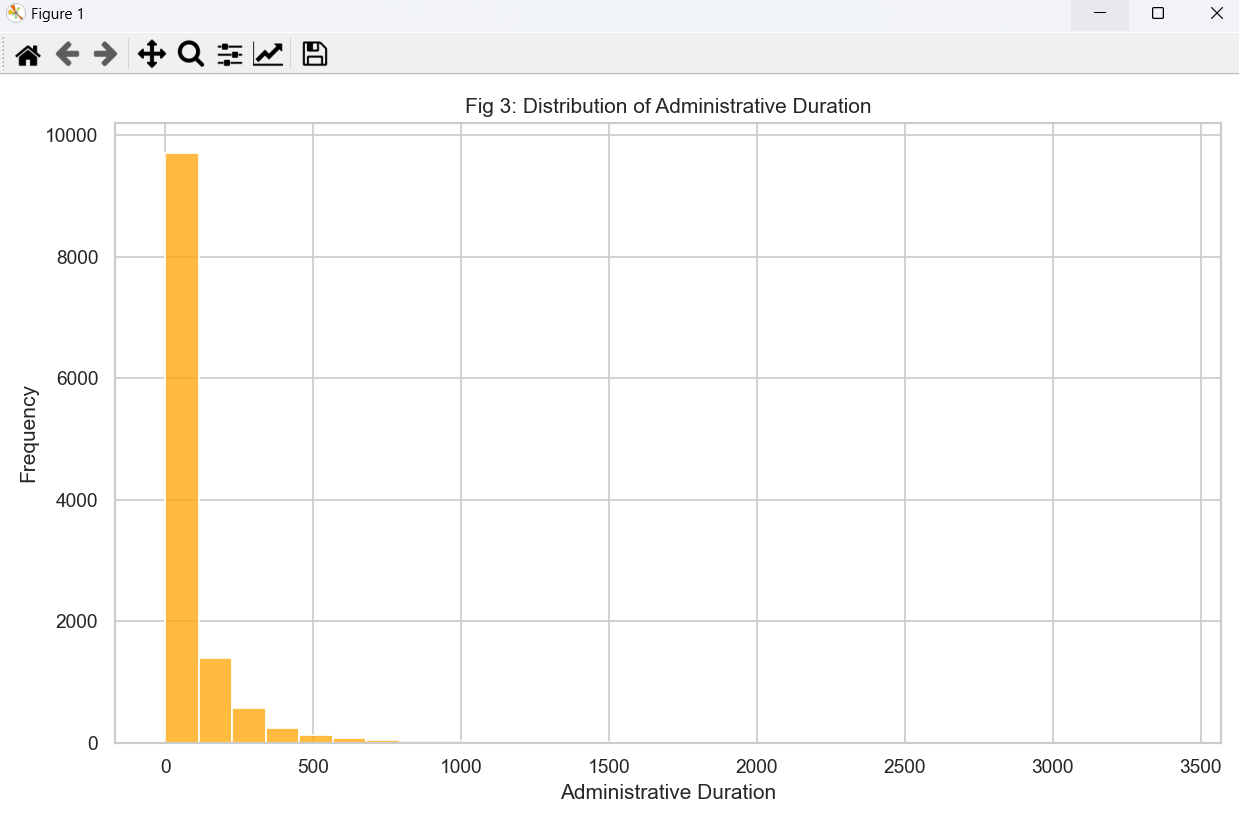
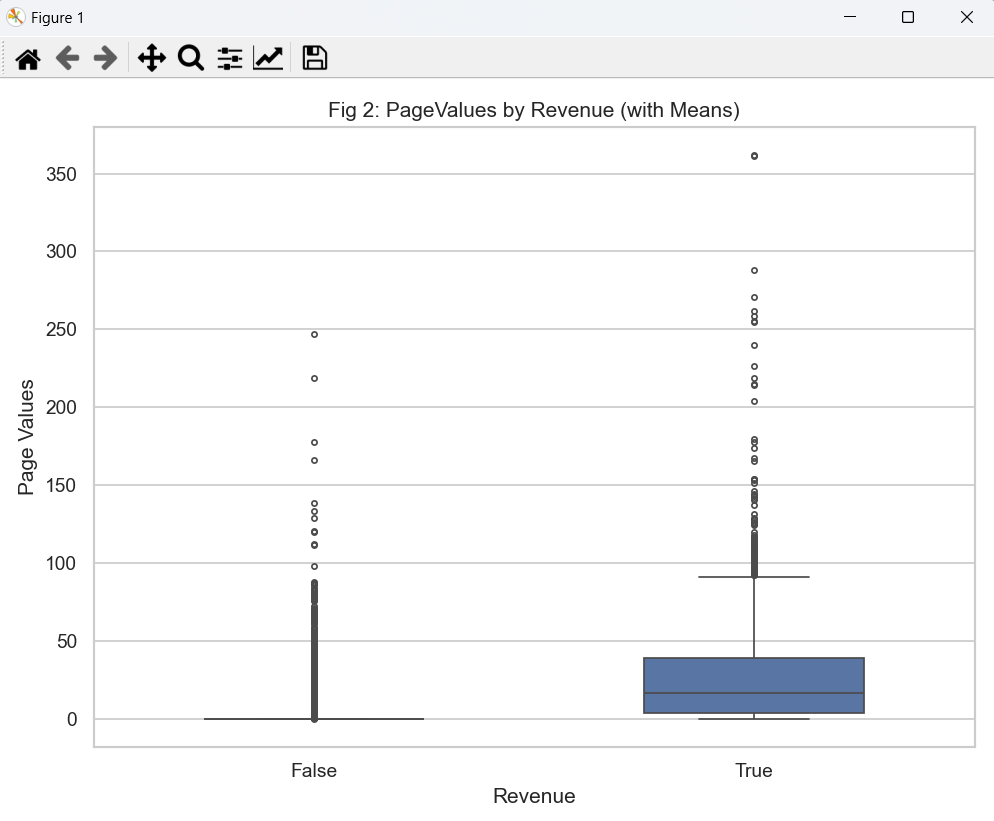
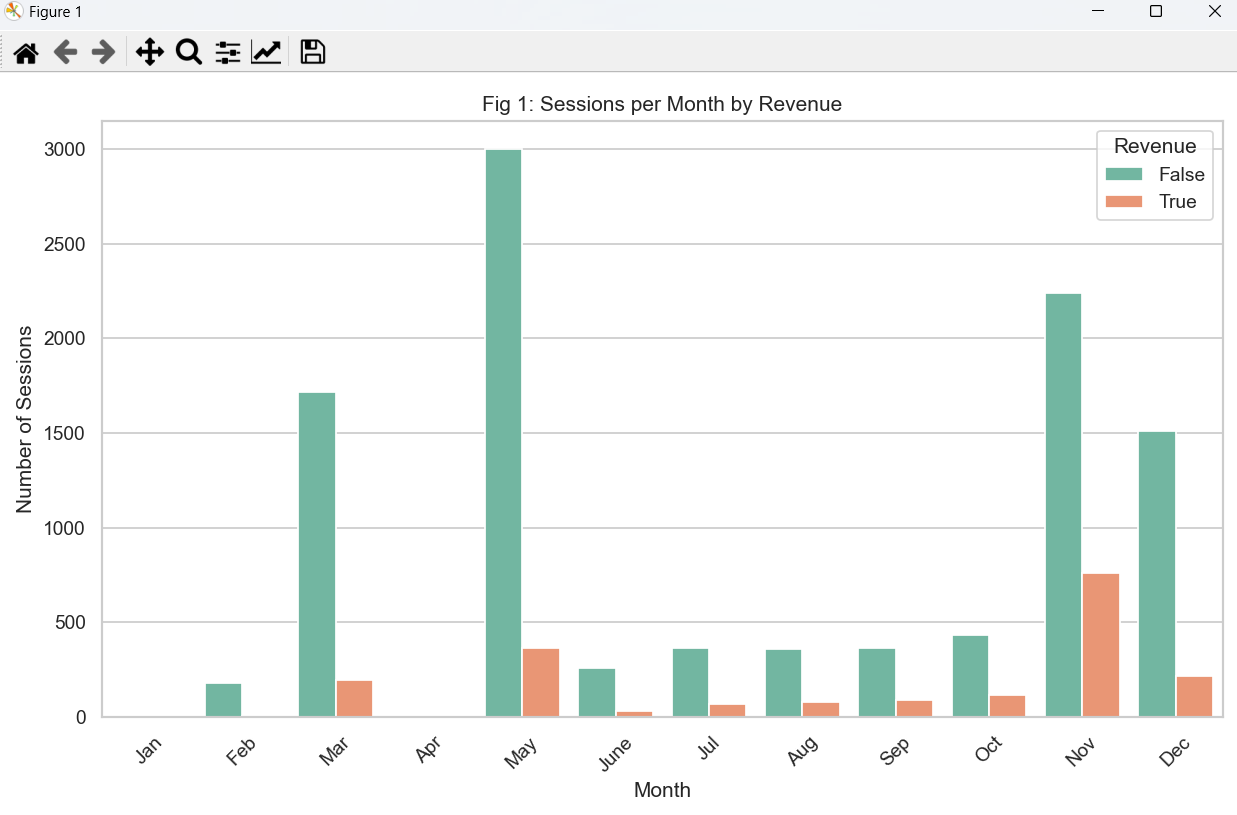
Feature engineering was also employed to extract new insights, such as **session recency** and **product category interactions**, to improve the model's prediction accuracy.

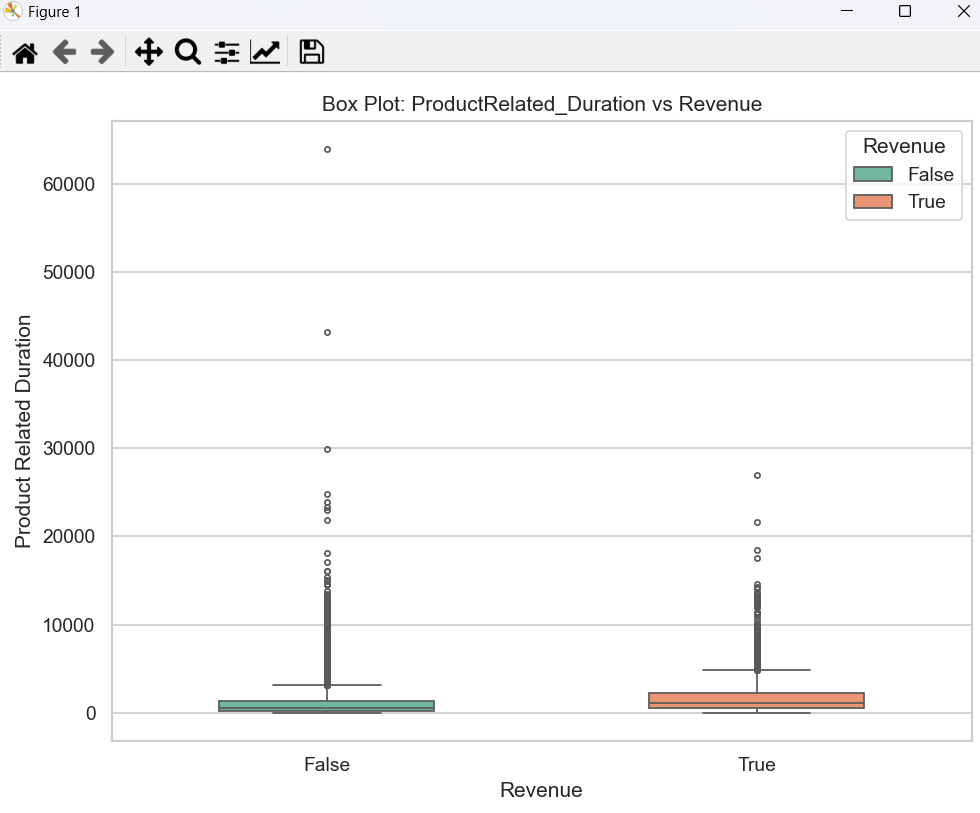
**Outcomes:**  
The predictive model accurately identified potential buyers, with **session duration** and **product interaction** being the most significant features contributing to purchase intentions. The model was able to predict future purchases with a high degree of accuracy, allowing businesses to tailor their marketing strategies and offer personalized recommendations.

**Conclusion:**  
Predicting purchase intentions is a key tool for optimizing e-commerce strategies. This analysis provides a clear roadmap for anticipating customer actions and adjusting marketing tactics to improve conversion rates.

**5). Visualization**

****

****

****

### **6).** **Hypothesis Testing to Validate Behavioral Impact on Revenue**

As part of the analytical process, **hypothesis testing** was conducted to statistically validate whether specific user behaviors—particularly **Bounce Rates**—had a significant impact on the likelihood of generating revenue. The hypothesis testing helped reinforce the assumptions made during data exploration and simulation of user behavior patterns.

#### Z-Test: Bounce Rates for Revenue vs No Revenue

To test whether there is a **statistically significant difference** in bounce rates between sessions that resulted in revenue and those that did not, a **Z-test** was conducted:

* **Null Hypothesis (H₀):** There is no difference in bounce rates between Revenue and No Revenue groups.
* **Alternative Hypothesis (H₁):** There is a significant difference in bounce rates between the two groups.

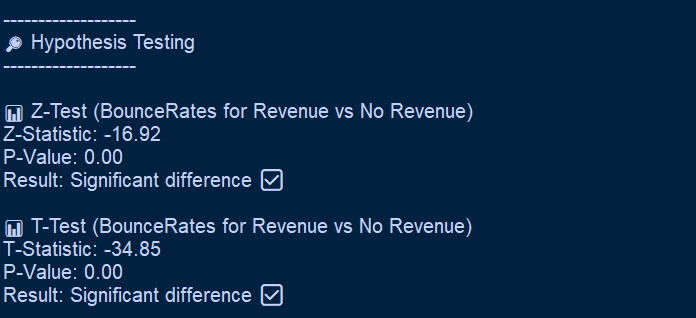
**Results:**

* **Z-Statistic:** -16.92
* **P-Value:** 0.00
* **Conclusion:** The p-value is far below the standard threshold (0.05), so we **reject the null hypothesis**, indicating a **significant difference** in bounce rates between sessions with and without revenue.

#### T-Test: Bounce Rates for Revenue vs No Revenue

A **T-test** was also performed to cross-validate the results from the Z-test and account for sample size variability.

* **T-Statistic:** -34.85
* **P-Value:** 0.00
* **Conclusion:** Again, the p-value is very low, confirming that **bounce rate significantly impacts purchase behavior**. Users who generated revenue had noticeably lower bounce rates.

****

## CONCLUSION

This project effectively applied data analytics to explore and understand user behavior in an e-commerce setting. Through a detailed Exploratory Data Analysis (EDA), key trends, outliers, and relationships within the dataset were uncovered. Data cleaning steps, such as handling missing values and removing duplicates, ensured the accuracy and reliability of the analysis. Both numerical and categorical features were examined, revealing valuable insights into how different factors contribute to purchase decisions.

Key behavioral indicators—such as bounce rates, exit rates, visitor type, and durations on administrative and informational pages—emerged as significant predictors of user conversion. Correlation analysis and visualizations, including histograms, box plots, and heatmaps, helped highlight how these variables interact and influence purchasing behavior.

Statistical tests such as T-tests and Chi-Square tests further supported the findings, confirming that returning visitors, lower bounce rates, and higher engagement times were positively associated with higher conversion rates. These patterns suggest that encouraging repeat visits and optimizing the user journey can directly impact sales outcomes.

In addition, the classification models developed as part of this analysis contributed to understanding the likelihood of a user completing a purchase based on their behavior. The performance of these models can be leveraged for predictive analytics, offering real-time decision support for marketing and user experience strategies.

Overall, this project demonstrates the power of data analytics in deriving actionable insights from user interaction data. It highlights the importance of understanding customer behavior and provides a foundation for implementing data-driven improvements that enhance engagement, increase conversions, and support business growth in the competitive e-commerce landscape.

## 

## FUTURE SCOPE

While this CA2 Python project has successfully analyzed online shopper behavior and visualized key trends through plots such as session frequency by month and administrative duration, there are several directions in which the project can be improved and extended. These enhancements would increase its practical usefulness for businesses, marketers, and analysts. Some potential future improvements include:

Integrating Predictive Models

In the future, simple machine learning models such as **Logistic Regression, Decision Trees,** or **Random Forests** can be implemented to:

* Predict the likelihood of a customer making a purchase (revenue generation).
* Identify factors contributing to cart abandonment. These predictions could help in optimizing user experience and boosting conversion rates.

Analyzing Customer Behavior Over Time

The current analysis provides an overview based on aggregated data. A valuable extension would be **tracking user behavior over time** (e.g., by week or month) to identify repeat visitors, behavioral trends, or seasonal spikes in activity. This could improve customer segmentation and targeted marketing strategies.

### 

Incorporating Clickstream or Feedback Data

If available, **clickstream data** or **customer feedback (**e.g., product reviews or session logs) could be analyzed to:

* Understand navigation patterns within the site.
* Identify pain points in the purchase funnel.
* Gauge sentiment toward certain products or features.

Enhancing Dashboard Visualizations

The current visualizations provide clear insights, but the dashboard can be enhanced with features like:

* **Interactive filters** for product categories, device types, or traffic sources.
* **Heatmaps** to show peak shopping hours.
* **User segmentation charts** (e.g., by new vs. returning customers). This would offer a more dynamic and user-friendly exploration of the dataset.

Real-Time Data Integration

A future improvement would be to **connect the system to real-time data sources** like live e-commerce logs or APIs (e.g., from Google Analytics or Shopify). This would allow for:

* Real-time revenue tracking.
* Dynamic detection of sudden changes in user behavior (e.g., a spike in bounce rate).

Introducing Gamified Business Insights

While traditionally used in education, **gamification techniques** like achievement badges or performance dashboards for marketing teams could help motivate internal stakeholders. For example, marketing analysts could receive notifications for meeting certain KPIs (e.g., reduced bounce rate, increased conversion).

## REFERENCES

* <https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset>
* <https://dash.plotly.com/>
* <https://seaborn.pydata.org/>
* <https://matplotlib.org/>
* <https://pandas.pydata.org/>

## GitHub LINK

The link for the whole code is:

<https://github.com/Yogesh33s/python-Data-visualization>