Identifying Shopping Trends using Data Analysis

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Shashwat Dhondyal, shashwatdhondyal812@gmail.com

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

We would like to take this opportunity to express our sincere gratitude to everyone who supported and guided us throughout the course of this internship. The knowledge, skills, and experiences gained during this period have been invaluable and will greatly benefit us in our professional journey.

First and foremost, we extend our deepest appreciation to our mentor, **Mr. Jay Rathod**, for his exceptional guidance, encouragement, and unwavering support. His insightful advice and constructive feedback have been instrumental in shaping the direction and quality of our work. Throughout the internship, his ability to simplify complex problems, provide practical solutions, and inspire innovative thinking has been truly commendable. His mentorship has not only enhanced our technical expertise but has also instilled in us a sense of responsibility and professionalism that will stay with us in the years to come.

We are immensely grateful to Mr. Rathod for his patience and understanding, especially during challenging times. His ability to create a supportive and collaborative environment allowed us to confidently tackle obstacles and seek creative solutions. His emphasis on learning through hands-on experience has helped us gain deeper insights into the industry and its dynamic nature.

Finally, we are thankful to the organization for providing us with the opportunity to work on this internship and for equipping us with the resources and platforms necessary to grow and excel. This experience has not only helped us gain technical skills but has also taught us the importance of collaboration, time management, and adaptability in a professional setting.

In conclusion, this internship has been a remarkable learning journey, and we owe its success to the guidance and support of **Mr. Jay Rathod** and everyone who contributed to it, directly or indirectly. We are truly grateful for the mentorship, encouragement, and opportunities that have shaped our growth and development throughout this experience.

ABSTRACT

This project focuses on Exploratory Data Analysis (EDA) to uncover shopping trends using a structured dataset stored in a CSV file. The primary objective was to analyse customer purchasing behaviour, identify key patterns, and derive actionable insights that can help businesses optimize their strategies and improve customer satisfaction. The analysis was performed using Jupyter Notebook and Visual Studio Code as development environments, leveraging Python libraries for efficient data processing and visualization.

The problem statement revolves around understanding consumer preferences and identifying trends within the dataset, which can often be a challenging task due to the volume and complexity of data. The project aimed to address these challenges by employing EDA techniques to transform raw data into meaningful insights.

The methodology involved multiple stages: data loading, cleaning, and preprocessing using pandas and NumPy, followed by data visualization using libraries such as Matplotlib, Seaborn, and Plotly Express. These tools provided both static and interactive visualizations, enabling a comprehensive exploration of trends like seasonal purchasing habits, popular product categories, and demographic influences on spending patterns.

Key results revealed significant insights, including peak shopping periods, variations in spending across customer segments, and correlations between product categories. Interactive visualizations created using Plotly Express enhanced the interpretation of these findings, offering an intuitive and engaging way to communicate the results.

In conclusion, this project successfully demonstrated the power of EDA in extracting valuable insights from raw data. By utilizing Python libraries and interactive tools, the study provided a clear understanding of shopping trends and offered recommendations for data-driven decision-making. The approach and results showcase the potential of EDA in helping businesses enhance their customer-focused strategies, improve inventory management, and optimize marketing efforts.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem Statement:

Understanding customer shopping trends is critical for businesses to stay competitive in an ever-evolving market. However, the sheer volume and complexity of transactional data can make it challenging to extract actionable insights. This project addresses the need for systematic analysis of shopping behavior to identify key patterns, trends, and factors influencing customer decisions. By leveraging data science tools and techniques, the project aims to uncover insights such as popular products, seasonal variations, and demographic preferences. These insights can help businesses optimize their strategies, enhance customer satisfaction, and improve operational efficiency.

1.2 Motivation:

The motivation behind this project stems from the growing significance of data-driven decision-making in the retail industry. With increasing competition, businesses must understand their customers' preferences and habits to provide personalized experiences and improve their offerings. The ability to analyze large datasets using EDA tools and techniques can bridge this gap effectively. This project was chosen to demonstrate the application of Python libraries for data analysis and visualization, providing valuable experience in working with real-world datasets. The potential applications include optimizing marketing campaigns, forecasting demand, and improving inventory management, ultimately contributing to enhanced business performance and customer satisfaction.

1.3 Objective:

The objectives of this project are as follows:

1. Perform Exploratory Data Analysis (EDA) on a shopping dataset to uncover meaningful insights.
2. Identify key trends, such as popular product categories, seasonal shopping patterns, and demographic influences.
3. Develop static and interactive visualizations to effectively communicate findings.
4. Provide data-driven recommendations to improve business strategies and operations.

1.4 Scope of the Project:

The project focuses on the following aspects:

1.4.1 Data Analysis:

- Performing Exploratory Data Analysis (EDA) on a shopping dataset in CSV format.
- Identifying patterns, trends, and correlations to understand shopping behavior.

1.4.2 Data Visualization:

- Developing static visualizations using Matplotlib and Seaborn for clear representation.
- Creating interactive visualizations with Plotly Express to enhance engagement and interpretation.

1.4.3 Trend Identification:

- Analyzing factors like seasonal shopping habits, demographic influences, and popular product categories to uncover actionable insights.

1.4.4 Decision Support:

- Providing data-driven recommendations for businesses to improve their strategies, inventory management, and customer satisfaction.

1.4.5 Tools and Techniques:

- Utilizing Python libraries such as pandas, NumPy, Matplotlib, Seaborn, and Plotly Express for efficient data processing and visualization.
- Employing Jupyter Notebook and Visual Studio Code as the development platforms for the project.

1.5 Limitations of the Project

The project has the following limitations:

1.5.1 Dataset Dependency:

- The results and insights are specific to the provided dataset and may not be directly applicable to other datasets or industries.

1.5.2 Absence of Predictive Modeling:

- The project focuses solely on Exploratory Data Analysis (EDA) and does not include advanced machine learning or predictive analytics techniques.

1.5.3 Data Quality Constraints:

- The analysis is dependent on the quality and completeness of the dataset. Missing or inaccurate data may affect the reliability of the findings.

1.5.4 Static Insights:

- The analysis is based on historical data and does not account for real-time or dynamic changes in shopping behavior.

1.5.5 Limited Scope of Analysis:

- Advanced topics such as customer segmentation, pricing strategies, or sentiment analysis are beyond the scope of this project.

1.5.6 Resource Constraints:

- The computational resources and libraries available may limit the complexity of the analysis and visualization techniques employed in the project.

CHAPTER 2

Literature Survey

2.1 Review of Relevant Literature

In this section, we review key literature and previous works related to the analysis of shopping trends and Exploratory Data Analysis (EDA). These studies and methodologies have contributed to the development of techniques used in this project. The following literature is relevant to the scope of our work:

2.1.1. "Data Science for Business" by Foster Provost and Tom Fawcett (2013)

- This book offers a comprehensive understanding of data science techniques and how they can be applied to business problems. It emphasizes the use of data to make informed decisions, particularly in areas like customer behavior analysis and trend identification.
- Source: *O'Reilly Media*
- Link:https://www.oreilly.com/library/view/data-science-for/9781449374273/

2.1.2. "Exploratory Data Analysis" by John Tukey (1977)

- John Tukey's foundational work on Exploratory Data Analysis (EDA) introduced methods for analyzing data sets and uncovering underlying patterns without formal modeling. His work has laid the groundwork for modern EDA practices in fields ranging from retail analytics to healthcare.
- Source: *Addison-Wesley*
- Link:https://www.amazon.com/Exploratory-Data-Analysis-Addison-Wesley-Statistics/dp/0201076160

2.1.3. "Retail Data Analytics: A Practical Guide to Data-Driven Retailing" by K. S. Rajasekaran and R. K. Gupta (2019)

- This book focuses on retail data analytics and its application in understanding customer behavior, sales trends, and inventory management. It explores the use of data analytics to drive decisions in the retail sector and optimize various processes.
- Source: *Springer*
- Link: https://link.springer.com/book/10.1007/978-981-13-6639-4

2.1.4. **"Analyzing Consumer Behavior with Data Science"** by L. J. M. P. Schikora and M. J. C. van der Heijden (2016)

- This paper discusses how consumer behaviour can be analysed using data science techniques such as clustering and classification. It highlights how demographic, economic, and product-based factors influence purchasing behaviour.
- Source: *Journal of Retail Analytics*
- Link: https://www.researchgate.net/publication/312430142_Analyzing_consumer_behavior_with_data_science

2.1.5. **"Retail Trend Analysis Using Big Data Analytics"** by M. R. Khusainova and N. F. Abdukhamidov (2020)

- This study examines the role of big data analytics in understanding retail trends. It covers the integration of machine learning algorithms and data visualization techniques to predict shopping behaviors and optimize inventory management.
- Source: *Springer*
- Link: https://link.springer.com/chapter/10.1007/978-3-030-38767-7_11

These works provide a strong foundation for exploring the analysis of shopping trends through EDA, and their methodologies directly align with the approaches used in this project. The literature not only presents theoretical frameworks but also provides practical applications of data science in understanding and optimizing retail operations.

2.2 Existing Models, Techniques, or Methodologies Related to the Problem.

Several models, techniques, and methodologies have been developed and applied to understand shopping trends and consumer behavior using data analytics. The following are some of the prominent methods and models that align with the problem addressed in this project:

2.2.1 Exploratory Data Analysis (EDA)

- EDA is a critical technique used to analyse and summarize the main characteristics of data sets, often with visual methods. It helps to uncover underlying patterns,

detect outliers, test assumptions, and check for data inconsistencies. In retail, EDA is used to identify trends in customer behaviour, seasonal shopping patterns, and product preferences without making assumptions about the underlying data.

- **Methodology**: EDA typically involves steps such as data cleaning, feature selection, data transformation, and visualization (using tools like histograms, box plots, and scatter plots). Libraries such as pandas, NumPy, Matplotlib, Seaborn, and Plotly Express are commonly used in the implementation of EDA.
- **Application**: Identifying seasonal shopping habits, popular product categories, and correlations between customer demographics and purchasing behavior.

2.2.2. Clustering Techniques (K-Means, Hierarchical Clustering)

- Clustering techniques are used to group similar data points together, making it easier to identify patterns. In the context of shopping trends, clustering is applied to segment customers based on their purchasing behavior, demographics, or preferences.
- Methodology: K-means clustering, a popular unsupervised learning technique, divides data into K distinct groups. Hierarchical clustering, another method, builds a tree-like structure to represent data clusters.
- Application: Identifying customer segments such as frequent buyers, seasonal shoppers, or high-spending groups. This can aid in personalized marketing and targeted inventory management.

2.2.3 Association Rule Mining (Apriori Algorithm)

- Association rule mining is a technique used to find relationships between variables in large datasets. The Apriori algorithm is frequently used in retail to analyze product associations and identify items that are often purchased together (market basket analysis).
- Methodology: The Apriori algorithm identifies frequent itemsets and generates association rules with minimum support and confidence thresholds. This helps in understanding which products are typically bought together.

- **Application**: Identifying cross-selling opportunities and optimizing product placement in stores or on e-commerce platforms.

2.2.4 Time Series Analysis (Seasonal Decomposition, ARIMA)

- Time series analysis is used to analyse temporal data and forecast future trends. In retail, it is particularly useful for understanding seasonal trends, sales fluctuations, and predicting future demand.
- **Methodology**: Seasonal decomposition of time series (STL) helps to extract trends, seasonality, and residuals from temporal data. ARIMA (Auto-Regressive Integrated Moving Average) models are used to predict future sales based on past data.
- **Application**: Predicting future shopping patterns, demand forecasting, and inventory management based on historical shopping data.

2.2.5 Sentiment Analysis (Text Mining)

- Sentiment analysis involves using natural language processing (NLP) techniques to analyze customer reviews, feedback, or social media posts. This technique helps understand customer sentiments and opinions about products or brands.
- **Methodology**: Text mining techniques, such as tokenization, stopword removal, and sentiment classification, are used to extract meaningful insights from unstructured textual data.
- **Application**: Analyzing customer feedback to identify preferences, complaints, and sentiments about specific products or shopping experiences.

2.2.6 Decision Trees (CART, Random Forests)

- Decision trees are used for classification and regression tasks. In retail, decision trees can help predict customer behavior, such as whether a customer will purchase a product or how much they will spend.
- **Methodology**: CART (Classification and Regression Trees) is a popular method used to build decision trees. Random forests, an ensemble method, combine multiple decision trees to improve accuracy and prevent overfitting.
- **Application**: Predicting customer purchase behavior, targeting the right products to customers, and optimizing marketing strategies.

These existing models, techniques, and methodologies form the foundation of the analysis conducted in this project, allowing for the identification of meaningful insights and trends in shopping behavior. By leveraging EDA and advanced techniques such as clustering and time series analysis, businesses can optimize strategies, improve customer experiences, and drive data-driven decision-making.

2.3 Gaps and Limitations in Existing Solutions

2.3.1. Lack of Comprehensive Visualization

Most existing studies or solutions focus on static visualizations, which may fail to provide an interactive and user-friendly experience for deeper data exploration.

Addressed by This Project:

- This project incorporates both static visualizations (using Matplotlib and Seaborn) and interactive visualizations (using Plotly Express) to enhance data interpretation and engagement.

2.3.2. Limited Focus on Trend Analysis

Existing solutions often overlook detailed analysis of seasonal shopping trends, demographic influences, and product category preferences.

Addressed by This Project:

- This project identifies specific patterns such as seasonal variations, shopping frequency by demographics, and popular product categories to provide actionable insights.

2.3.3. Generalized Recommendations

Previous methodologies tend to offer broad, generic recommendations without tailoring them to the dataset's specifics.

Addressed by This Project:

This project delivers customized, data-driven recommendations for inventory management, marketing strategies, and improving customer satisfaction based on the dataset's findings.

2.3.4. Limited Use of Interactive Tools

Many existing approaches rely on traditional tools and techniques that are not as flexible or interactive for modern data analysis needs.
Addressed by This Project:

- This project leverages powerful Python libraries like Plotly Express for interactive data exploration and enhances usability for decision-makers.

5. Lack of Scalability for Real-World Applications

Existing solutions may not scale well for integration with real-world business systems or fail to provide outputs in a business-friendly manner.

Addressed by This Project:

- While this project focuses on EDA, it sets a foundation for integrating with business intelligence tools or extending into predictive analytics to meet real-world requirements.

By addressing these gaps, this project enhances the scope and applicability of data analysis in understanding shopping trends and enabling informed business decisions.

CHAPTER 3

Proposed Methodology

3.1 System Design

Proposed Solution Diagram:

Below is a step-by-step flow diagram representing the proposed methodology:

3.1.1. Explanation of the Diagram:

3.1.1.1. Dataset (CSV):

- The project starts by obtaining a shopping dataset in CSV format. This dataset contains key attributes such as customer demographics, shopping behavior, and temporal data (e.g., timestamps for purchases).

3.1.1.1 Data Preprocessing:

- Objective: Clean and prepare the raw data for analysis.
- Steps Involved:
  - Handle missing or null values.
  - Remove duplicate entries to ensure data integrity.
  - Format columns (e.g., converting date strings to a uniform format).

- Create new attributes if needed (e.g., derive total spending from individual transactions).

3.1.1.2 Exploratory Data Analysis (EDA):

- Objective: Understand the data, identify trends, and uncover patterns.
- Steps Involved:
  - Perform descriptive statistical analysis (mean, median, standard deviation).
  - Identify correlations between variables (e.g., age and spending habits).
  - Examine categorical distributions (e.g., popular product categories).

3.1.1.3 Data Visualization:

- Objective: Communicate insights effectively through visual representations.
- Tools Used:
  - Matplotlib and Seaborn: Generate static plots like histograms, scatter plots, and heatmaps for statistical insights.
  - Plotly Express: Create interactive visualizations (e.g., interactive bar charts, pie charts, and line plots) for more dynamic exploration.

3.1.1.4 Insights and Recommendations:

- Objective: Translate findings into actionable insights.
- Output:
  - Identification of seasonal trends and popular product categories.
  - Analysis of customer demographics affecting purchasing decisions.
  - Customized recommendations for businesses to enhance inventory management, marketing strategies, and customer satisfaction.

3.2 Requirement Specification

The following tools and technologies are essential for implementing the solution:

3.2.1 Hardware Requirements:

3.2.1.1 Processor: Intel Core i3 or higher / AMD Ryzen equivalent

3.2.1.2 RAM: Minimum 4 GB (8 GB recommended for smoother operation)

3.2.1.3 Storage: At least 2 GB of free space for datasets, libraries, and output files

3.2.1.4 System Architecture: 64-bit operating system

3.2.1.5 Display: Screen resolution of 1280x720 or higher for optimal visualization

3.2.2 Software Requirements:

3.2.2.1 Operating System: Windows 10/11, macOS, or Linux

3.2.2.2 Programming Language: Python (Version 3.7 or above)

3.2.2.3 Development Platforms/IDEs: Jupyter Notebook (for stepwise exploration and analysis) and Visual Studio Code (for script-based development and debugging)

3.2.2.4 Python Libraries:

- numpy: For numerical computations and array manipulation
- pandas: For data manipulation, cleaning, and exploration
- matplotlib: For creating static visualizations
- seaborn: For statistical data visualization
- plotly.express: For building interactive plots

3.2.2.5 Additional Tools and Resources:

- Git: For version control and collaboration
- Web Browser: To view interactive visualizations if needed

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

4.1.1 Snapshot: Dataset Overview

Code Snippet:

```
# reading the data set
shop = pd.read_csv(r"C:\Users\shash\Downloads\shopping_trends.csv")
shop.shape
#display all columns
pd.set_option('display.max_columns', None)
print(shop.head())
```

Output:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually

Table 1 Dataset Overview

Description: This snapshot provides a preview of the dataset after loading it into the environment. It includes all columns to offer a complete understanding of the data structure. The dataset contains details such as customer demographics, purchase behavior, product categories, and more, which will be used for analysis.

By examining the first few rows, we can identify the types of data available for further exploration and visualization

4.1.2. Gender Distribution (Bar Plot):

Code Snippet:

```
# Bar plot
shop["Gender"].value_counts().plot(kind="bar")
plt.title("Gender Distribution (Bar Plot)")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.show()  # <-- This is necessary
```

Output:

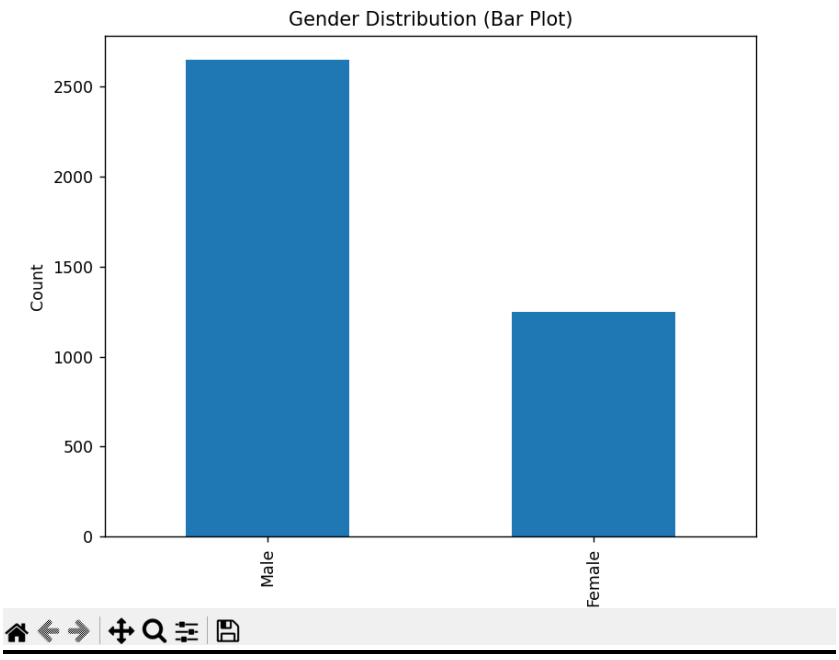

Figure 1: Gender Distribution Bar Plot

Description: This bar plot displays the gender distribution in the dataset. Each bar represents the count of individuals of a specific gender, making it easy to see which gender dominates the shopping trends dataset.

4.1.3. Datatypes of Columns:

Code Snippet:

```
print(shop.dtypes)
```

Output:

```
Customer ID                int64
Age                        int64
Gender                    object
Item Purchased            object
Category                  object
Purchase Amount (USD)      int64
Location                  object
Size                      object
Color                     object
Season                    object
Review Rating            float64
Subscription Status       object
Shipping Type             object
Discount Applied          object
Promo Code Used           object
Previous Purchases         int64
Payment Method            object
Frequency of Purchases    object
dtype: object
```

Table 2 Datatypes of the table elements

Description: *This snapshot displays the data types of all columns in the dataset. It helps identify whether each column is numerical, categorical, or another type. Understanding the data types is crucial for determining the appropriate preprocessing techniques and visualization methods. For instance:*

- Numerical Columns: Used for statistical analysis and continuous data visualizations (e.g., bar charts, histograms).
- Categorical Columns: Used for grouping, aggregation, or categorical comparisons (e.g., pie charts, box plots).
- Datetime Columns: Used for time-series analysis.

4.1.4. Age Distribution by Category:

Code Snippet:

```
shop['Age_category'] = pd.cut(shop['Age'], bins= [0,15, 18 , 30 , 50 , 70] , labels= ['child' , 'teen' , 'Young Adults' ,'Middle-Aged Adults'
                                                                                      , 'old'] )

fig = px.histogram(shop , y = 'Age' , x = 'Age_category')
fig.show()
```

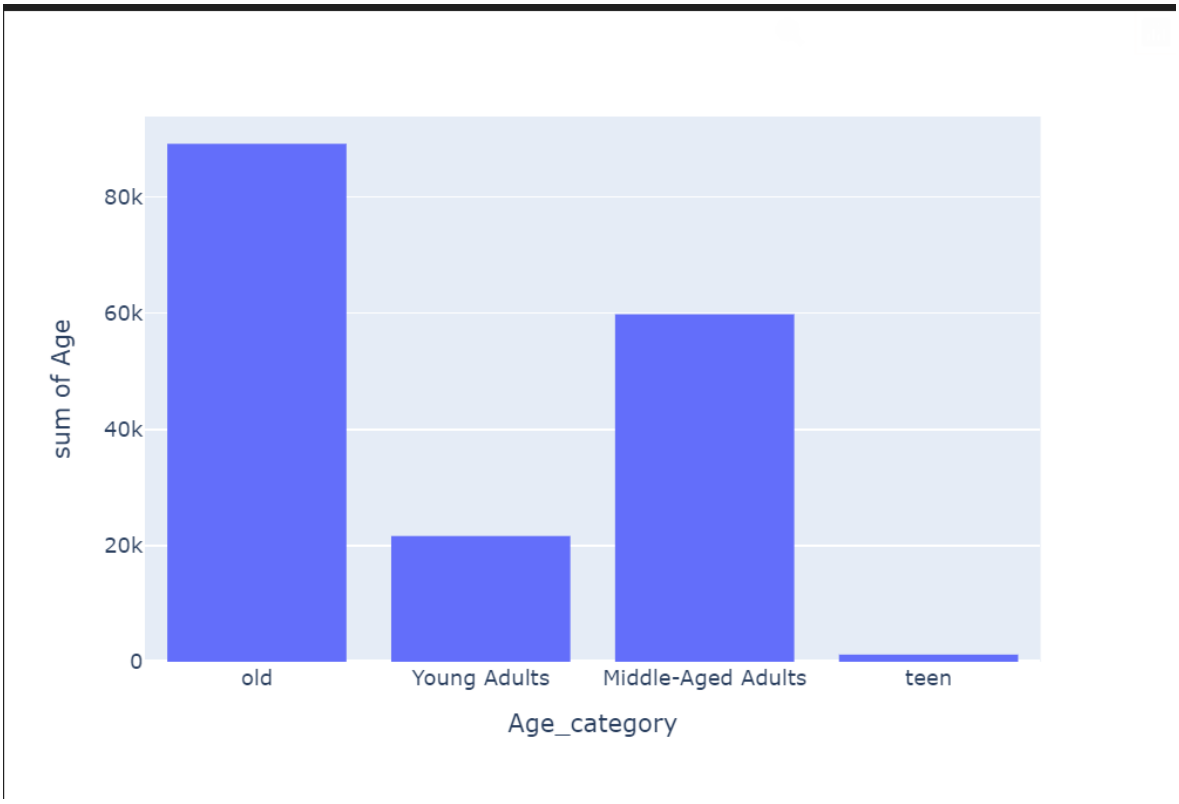

Figure 2 Histogram: Age Distribution by Category

4.1.5. Average Purchase Amount by Gender:

Code Snippet:

```
sns.barplot(data=shop, x='Gender', y='Purchase Amount (USD)', ci=None)
plt.title("Average Purchase Amount by Gender")
plt.xlabel("Gender")
plt.ylabel("Purchase Amount (USD)")
plt.show()
```

Output:

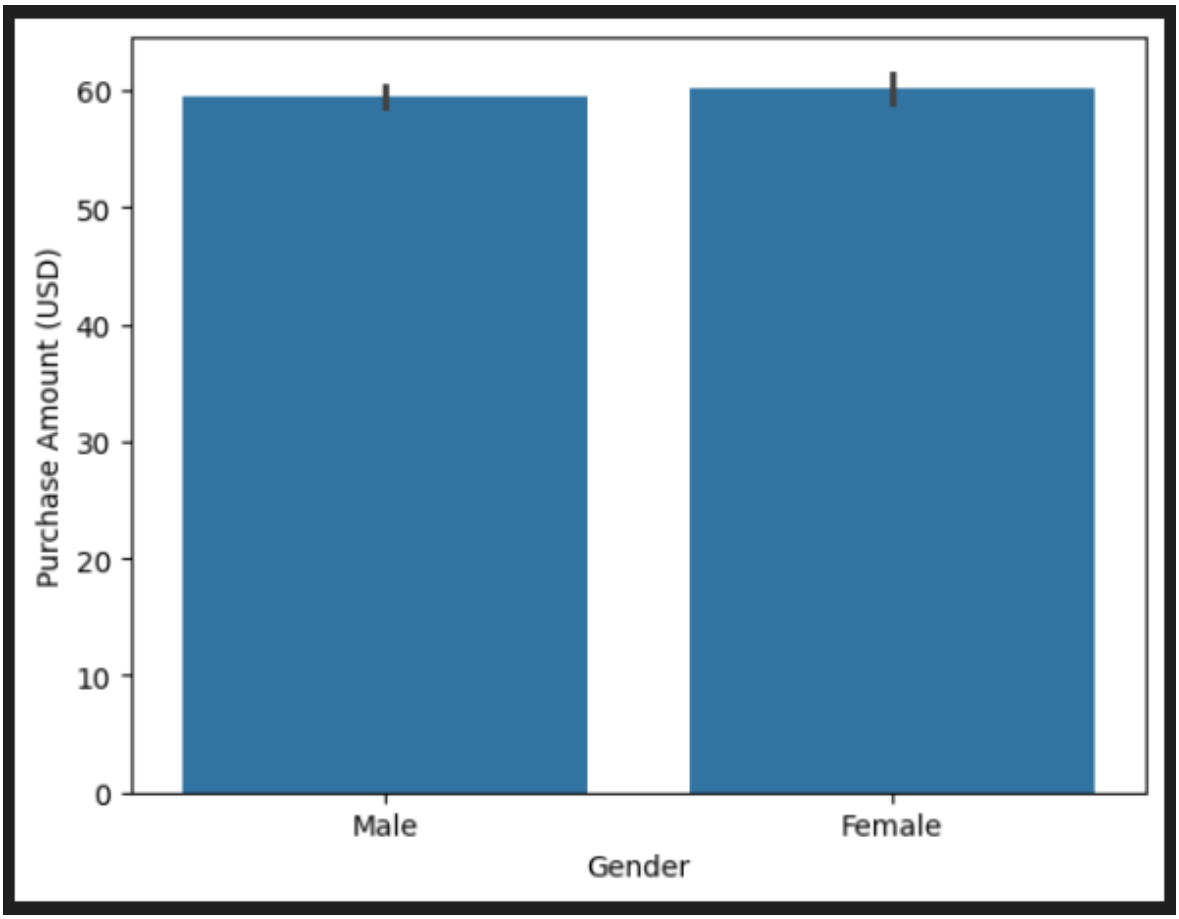

Figure 3: Average Purchase Amount by Gender

Description: This bar plot shows the average purchase amount for each gender in the dataset. The key elements are:

- X-Axis: Gender categories (e.g., Male, Female).

- Y-Axis: Average purchase amount in USD.

Insights:

- Identify which gender spends more on average.
- Determine any significant variations in spending habits.

4.1.5. Average Purchase Amount by Gender:

Code Snippet:

```
shop.columns()
shop.groupby('Category')['Item Purchased'].value_counts()
fig = px.histogram(shop , x = 'Item Purchased' , color = 'Category')
fig.show()
```

Output:

Category	Item Purchased	
Accessories	Jewelry	171
	Sunglasses	161
	Belt	161
	Scarf	157
	Hat	154
	Handbag	153
	Backpack	143
	Gloves	140
Clothing	Pants	171
	Blouse	171
	Shirt	169
	Dress	166
	Sweater	164
	Socks	159
	Skirt	158
	Shorts	157
	Hoodie	151
	T-shirt	147
	Jeans	124
Footwear	Sandals	160
	Shoes	150
	Sneakers	145
	Boots	144
Outerwear	Jacket	163
	Coat	161

Name: count, dtype: int64

Table 3: Item Purchased in each Category

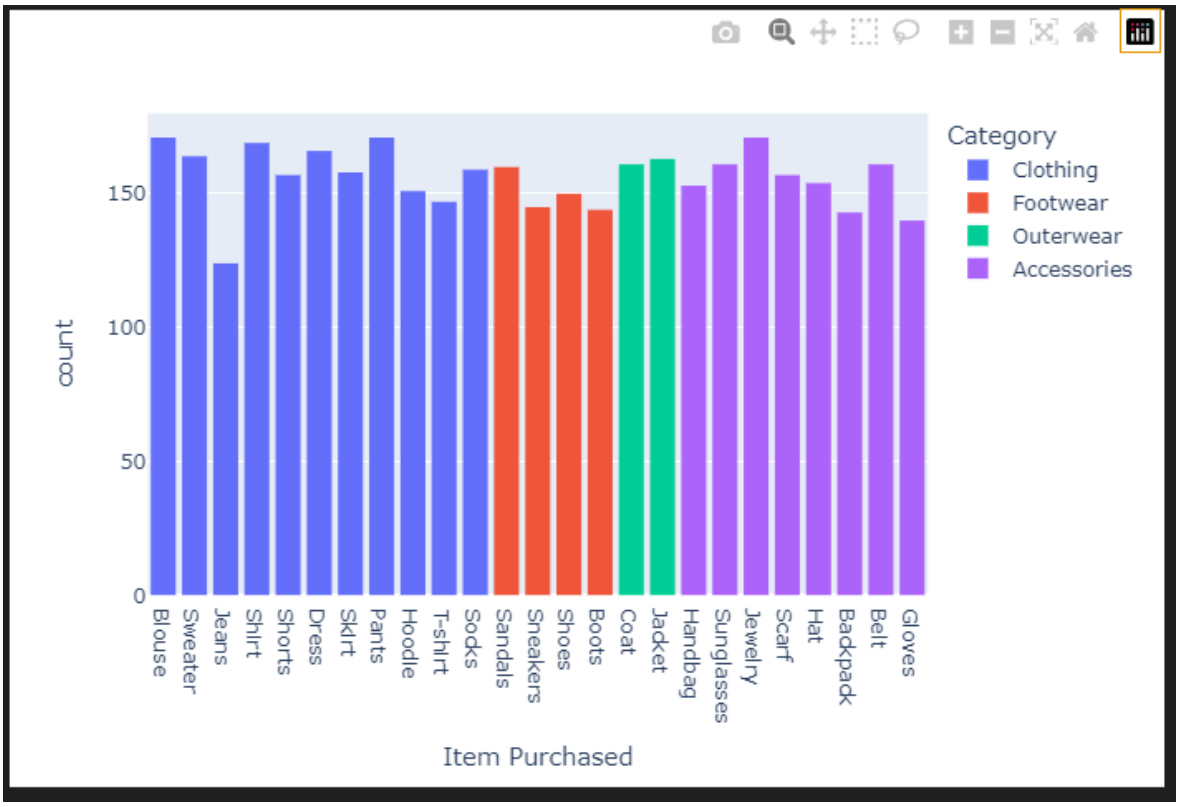

Figure 4: Most Common Items purchased in Each Category

Description: The given table and histogram shows the most common items purchased in each of the given categories. For the ACCESSORY category – Jewellery is the most common item bought while for the CLOTHES category Pants and Bloused has a tie. The data is useful in determining the commonly bought item in each category.

4.1.6. Average Purchase Amount by Gender:

Code Snippet:

```
shop['Subscription Status'].unique()
sns.barplot(shop  , x = 'Subscription Status' , y = 'Purchase Amount (USD)')

shop['Purchase Amount (USD)'].sum()
shop.groupby('Subscription Status')['Purchase Amount (USD)'].mean()
```

OUTPUT:

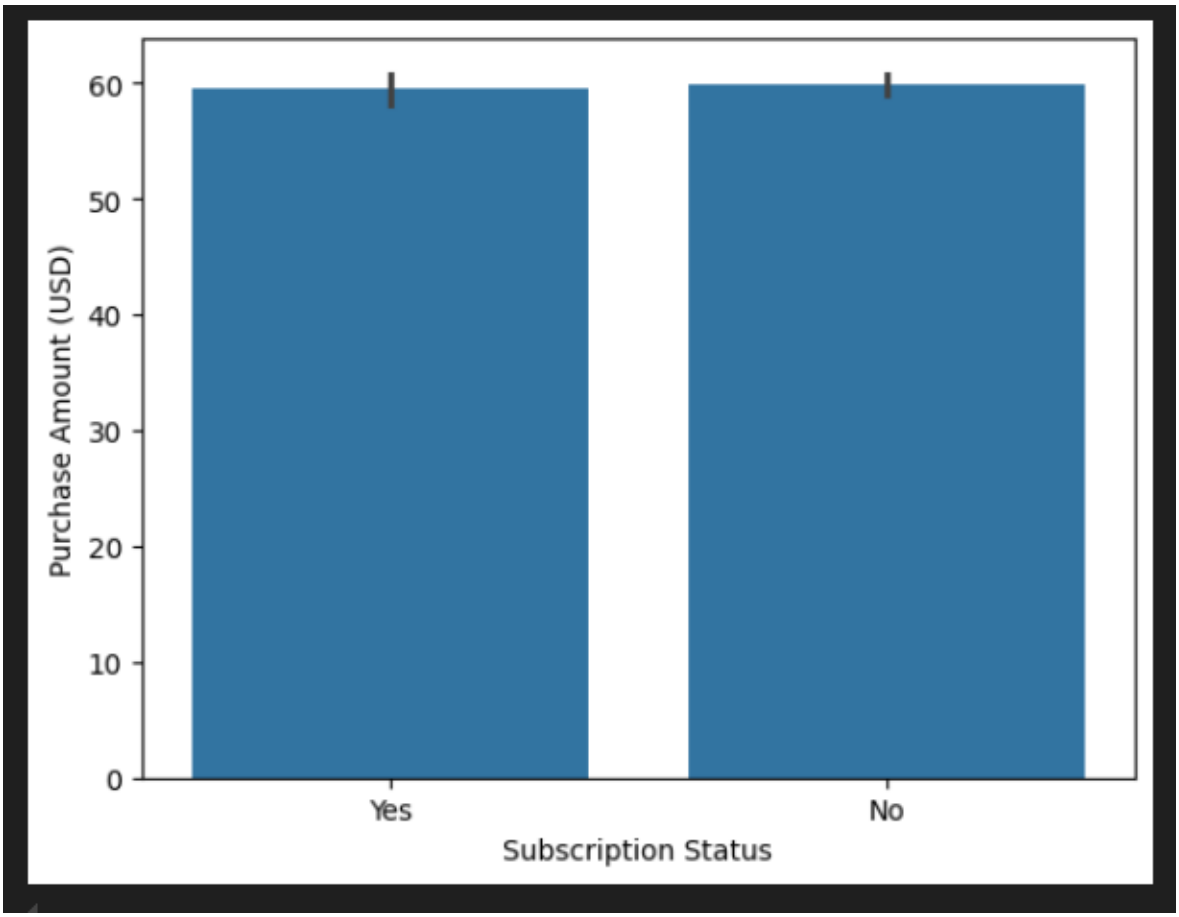

Figure 5: Purchase behavior between subscribed and non-subscribed customers.

```
Subscription Status
No     59.865121
Yes    59.491928
Name: Purchase Amount (USD), dtype: float64
```

Figure 6: Mean of Purchasing Behavior

Description: The bar-plot helps in identifying the difference between the purchasing behaviour of the subscribe and non-subscribed customers. The mean and bar-plot shows almost little to no difference in shopping patterns and behaviour.

4.1.7. Promo code usage:

Code Snippet:

```
shop_groupby  = shop.groupby('Promo Code Used')['Purchase Amount (USD)'].sum().reset_index()
fig = px.sunburst(shop , path=['Gender' , 'Promo Code Used'] , values='Purchase Amount (USD)')
fig.show()
```

Output:

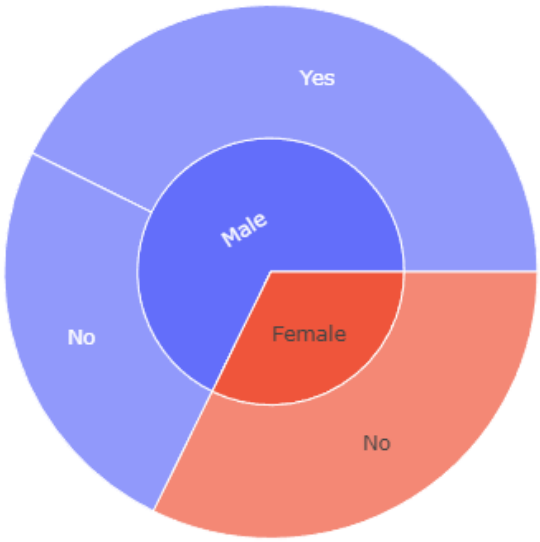

Figure 6: Promo code Usage

Description: The pie chart compares whether the customers using promo code tends to spend or not. From the pie chart it can derived that a vast majority of male customers tend to spend more with a promo code, while female customers spending don't spend more with the promo code.

4.1.8. Purchase Amount Difference:

Code Snippet:

```
shop_group = shop.groupby('Gender')['Purchase Amount (USD)'].sum().reset_index()
fig = px.bar(shop_group , x = 'Gender' , y = 'Purchase Amount (USD)')
fig.show()
px.sunburst(data_frame= shop , path = ['Gender' ,'Age_category'] , values='Purchase Amount (USD)')
```

Output:

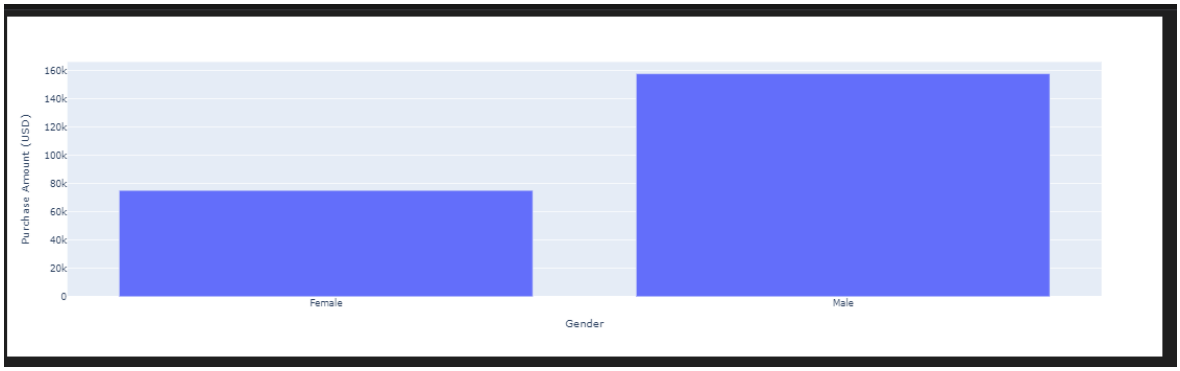

Figure 7: Average purchase amount by Genders

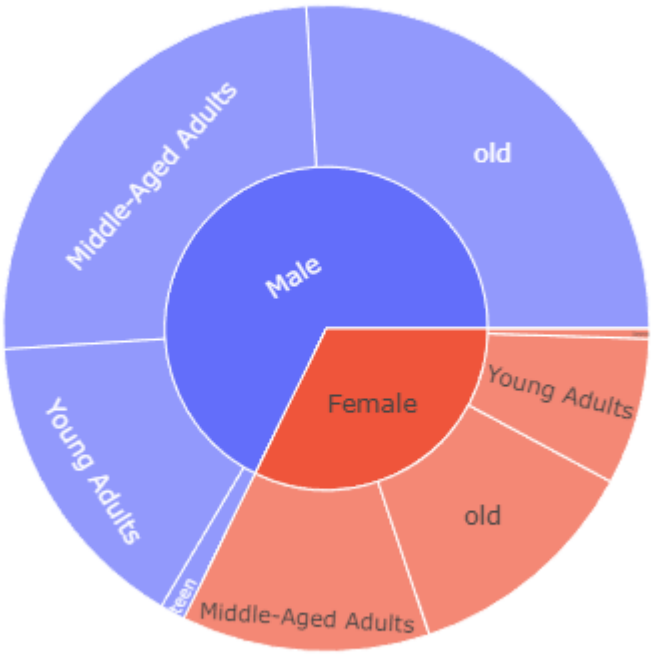

Figure 8: Pie chart: Spending differences based on Gender

Description: The figures illustrate the average spending differences between genders. The bar graph indicates that males tend to spend more than females. Meanwhile, the pie chart highlights which age group within each gender accounts for the largest share of spending. Among male customers, teens are the smallest spenders, while middle-aged and older males are the top spenders. Similarly, for females, older and middle-aged women are the

largest spenders, with young adults spending slightly less, and teens being the smallest spenders.

4.2. GitHub Link for Code:

Link: https://github.com/BinaryBrainss/Internship

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

While the project successfully analyses shopping trends and provides actionable insights, there are several potential areas for improvement and future enhancement:

5.1.1 Incorporation of Predictive Analytics:

5.1..1 Currently, the project performs exploratory data analysis (EDA) and descriptive analysis. In the future, incorporating machine learning models (such as regression, classification, or clustering) could help predict future shopping trends, customer behaviour, or product demand, allowing businesses to anticipate changes and optimize strategies proactively.

5.1.2 Real-Time Data Integration:

5.1.2.1 The analysis relies on static historical data. Integrating real-time data from online shopping platforms, customer interactions, or social media could enhance the model's applicability in fast-moving retail environments. This could help businesses monitor trends as they emerge and adjust their strategies in real-time.

5.1.3 Advanced Data Visualization:

5.1.3.1 While static and interactive visualizations were utilized in this project, incorporating advanced dashboarding tools (e.g., Tableau, Power BI, or custom-built dashboards) would allow decision-makers to interact with the data in more versatile ways, filtering by multiple variables and obtaining insights dynamically.

5.1.4 Personalized Customer Segmentation:

5.1.4.1 A deeper analysis could focus on segmenting customers based on their shopping behavior, demographics, or preferences. Implementing clustering algorithms like k-means or hierarchical clustering could

identify specific customer groups that can be targeted with personalized marketing strategies.

5.1.5 **Integration with Business Systems:**

5.1.5.1 Extending the project to work with real business systems, such as inventory management or customer relationship management (CRM) tools, could further automate decision-making. For instance, sales predictions could be used to adjust stock levels, or targeted promotions could be crafted for specific customer segments.

5.1.6 **Geospatial Analysis:**

5.1.6.1 Adding geographic dimensions to the data could provide insights into regional shopping trends and allow businesses to tailor marketing campaigns to specific locations. Geospatial analysis could also help optimize supply chain management by identifying regions with higher demand.

5.1.7 **Improvement in Data Quality and Volume:**

5.1.7.1 To build more robust insights, it's crucial to use larger and more diverse datasets. Incorporating external data sources, such as market trends, economic factors, or customer sentiment analysis, could provide a more comprehensive view of shopping behavior.

These future improvements will enhance the robustness of the model, its scalability, and its real-time applicability, ultimately driving more impactful business decisions.

5.2 Conclusion:

- This project provides a comprehensive analysis of shopping trends by performing Exploratory Data Analysis (EDA) and utilizing data visualization techniques to identify key patterns and trends in consumer behavior. By leveraging Python's powerful libraries such as Pandas, Matplotlib, Seaborn, and Plotly, this project offers valuable insights for businesses to optimize inventory management, tailor marketing strategies, and enhance customer satisfaction.

- Key findings include the identification of seasonal shopping patterns, correlations between customer demographics and spending habits, and the popularity of different product categories. The interactive visualizations further facilitate deeper exploration, making the analysis more engaging and accessible to decision-makers. The customized recommendations provided in this project empower businesses to adapt and thrive in a competitive market.

- In conclusion, the project has successfully demonstrated the importance of data-driven insights for businesses and the power of EDA in revealing actionable trends. With further enhancement, including predictive modeling and real-time data integration, the scope of the project can be expanded to provide even more value.

REFERENCES

[1] **Wickham, H. (2016).** *ggplot2: Elegant Graphics for Data Analysis.* Springer.
- A foundational text on data visualization, offering methods to improve the effectiveness of charts and plots in data analysis.

[2] **McKinney, W. (2017).** *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* O'Reilly Media.
- This book provides a comprehensive guide to data analysis with Python, focusing on the Pandas library for data manipulation and analysis.

[3] **Hunter, J. D. (2007).** *Matplotlib: A 2D Graphics Environment.* Computing in Science & Engineering, 9(3), 90-95.
- The paper introduces Matplotlib and its applications in creating high-quality visualizations in Python.

[4] **Plotly Technologies Inc. (2015).** *Plotly: Collaborative Data Science.* https://plot.ly
- Plotly provides an overview of interactive visualizations and how they can be used to enhance data exploration.

[5] **Seaborn: Statistical Data Visualization (2021).** *Seaborn Documentation.* https://seaborn.pydata.org
- Seaborn is a Python visualization library that builds on Matplotlib to make complex visualizations simpler and more readable.