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

Rishikesh, rishikeshpanditrp@gmail.com

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

ABSTRACT

Retail businesses today generate vast amounts of data, but extracting actionable insights from this data remains a challenge. This project aims to address this issue by analyzing shopping trends using a dataset that includes customer demographics, purchase behaviors, and product details. The objectives were to uncover key insights into customer preferences, seasonal trends, and spending patterns, enabling data-driven decision-making for enhanced business performance.

The methodology involved structured data cleaning, exploratory data analysis (EDA), and visualization using Python and its powerful libraries such as Pandas, Plotly, and Matplotlib. The data was grouped and analyzed to identify trends across demographics, product categories, and seasons.

Key results include:

- **Age Distribution**: Most customers fall between 31 and 57 years, with a median age of 44 years.
- **Category Spending**: Footwear recorded the highest average spending ($60.26), while Outerwear had the lowest ($57.17).
- **Gender-Based Behavior**: Male customers accounted for 68% of purchases and exhibited higher spending patterns than females.
- **Seasonal Trends**: Winter and summer saw the highest customer spending, driven by targeted promotions.
- **Preferred Shipping Types**: Express shipping was dominant, particularly for footwear and high-value products.
- **Minimal Correlation**: Between product size and purchase amount.

The insights derived from this project can help retailers optimize product offerings, target marketing efforts more effectively, and enhance customer satisfaction. Future work includes applying machine learning models for predictive analytics and integrating real-time data for dynamic insights.

This analysis demonstrates the value of data analytics in retail, offering clear and actionable recommendations for businesses aiming to thrive in a competitive market.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

Retail businesses generate vast amounts of data daily, but extracting meaningful insights is often challenging. Understanding customer behavior and shopping trends is crucial for personalized marketing, inventory management, and enhancing customer satisfaction.

1.2 Motivation:

This project was chosen to showcase the value of data analytics in addressing real-world business challenges. Analyzing shopping trends can empower businesses to make data-driven decisions, improve customer retention, and increase profitability.

1.3 Objective:

1) Identify key shopping trends and customer behaviors.
2) Explore relationships between demographic factors and purchasing habits.
3) Provide actionable recommendations for business growth.

1.4 Scope of the Project:

The project focuses on a single dataset, highlighting insights into customer preferences, seasonal trends, and regional differences. Limitations include the scope of data and its applicability to broader contexts.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

Existing literature emphasizes the importance of customer segmentation, seasonal analysis, and purchase behavior in retail. Studies highlight that understanding demographic trends and consumer preferences can lead to improved marketing strategies and inventory management. Key works, such as John W. Tukey's "Exploratory Data Analysis" and Hadley Wickham's contributions to data visualization, provide foundational techniques for identifying patterns in large datasets. Additionally, studies on customer churn and loyalty analysis offer insights into retaining customers through predictive analytics.

2.2 Mention any existing models, techniques, or methodologies related to the problem.

Various methodologies, such as clustering, correlation analysis, and decision trees, are commonly used to uncover patterns in retail data. Clustering techniques like k-means help segment customers based on their purchasing habits, while regression models explore relationships between variables such as price and demand. Visualization tools like ggplot2 and Plotly enhance data interpretation, making complex relationships easier to understand. Furthermore, machine learning models, including random forests and neural networks, have been applied to predict customer behaviors and optimize pricing strategies.

2.3 Highlight the gaps or limitations in existing solutions and how your project will address them.

Despite advancements, several gaps remain in existing solutions:

- Limited integration of real-time data for dynamic trend analysis.
- Underutilization of interactive visualizations to engage stakeholders.
- Challenges in handling heterogeneous datasets with missing or inconsistent values.

This project addresses these limitations by applying a structured data analysis pipeline that includes:

- Comprehensive exploratory data analysis (EDA) to clean and preprocess the data.
- Use of interactive tools like Plotly for better stakeholder engagement.
- Practical insights that can be applied to real-time systems for enhanced decision-making.

By addressing these gaps, the project demonstrates the potential of modern analytics techniques to unlock actionable insights from retail data.

CHAPTER 3

Proposed Methodology

3.1   System Design

The project follows a structured pipeline:

1) Data cleaning and preprocessing.
2) Exploratory Data Analysis (EDA) for identifying trends.
3) Visualization and interpretation of results.

3.2   Requirement Specification

3.2.1   Tools and Technologies:

1) **Python**: Data analysis and visualization.

2) **Plotly, Matplotlib, and Seaborn**: Creating interactive and static visualizations.

3) **Pandas and NumPy**: Data manipulation and analysis.

3.2.2   Hardware Requirements:

1) Modern laptop/desktop with at least 8GB RAM and 2GHz processor.

3.2.3   Software Requirements:

1) Python 3.x, Jupyter Notebook, or any Python IDE.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

1) **Customer Age Distribution**:
   - Median age: 44 years.
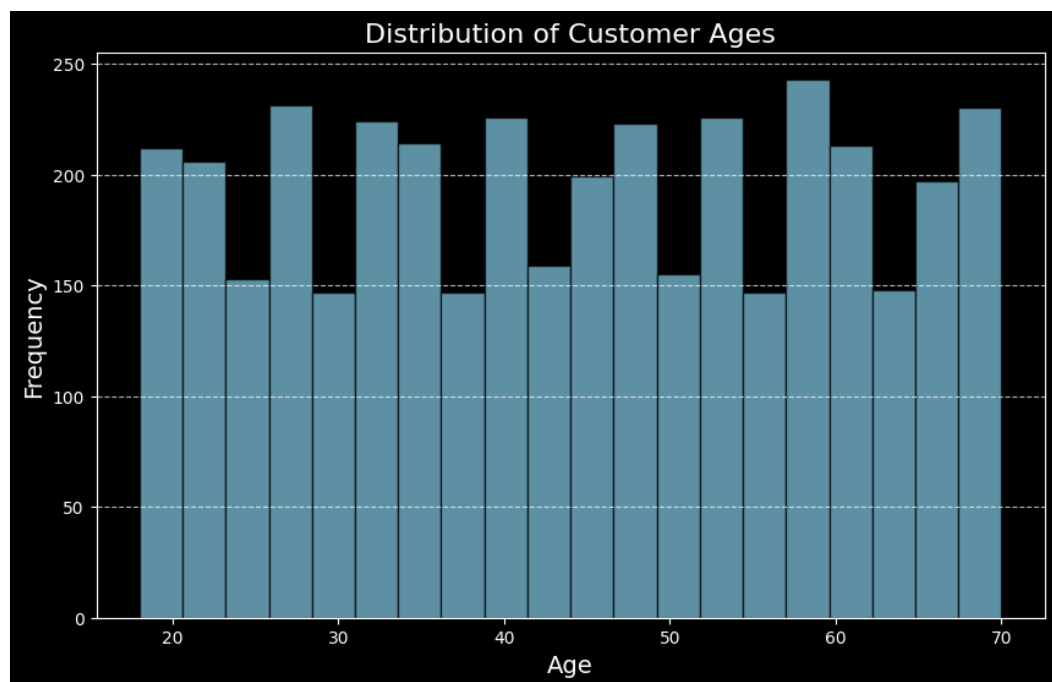   - Most customers fall between 31 and 57 years.

Figure 1: Distribution of Customer Ages

2) **Gender-based Analysis**:

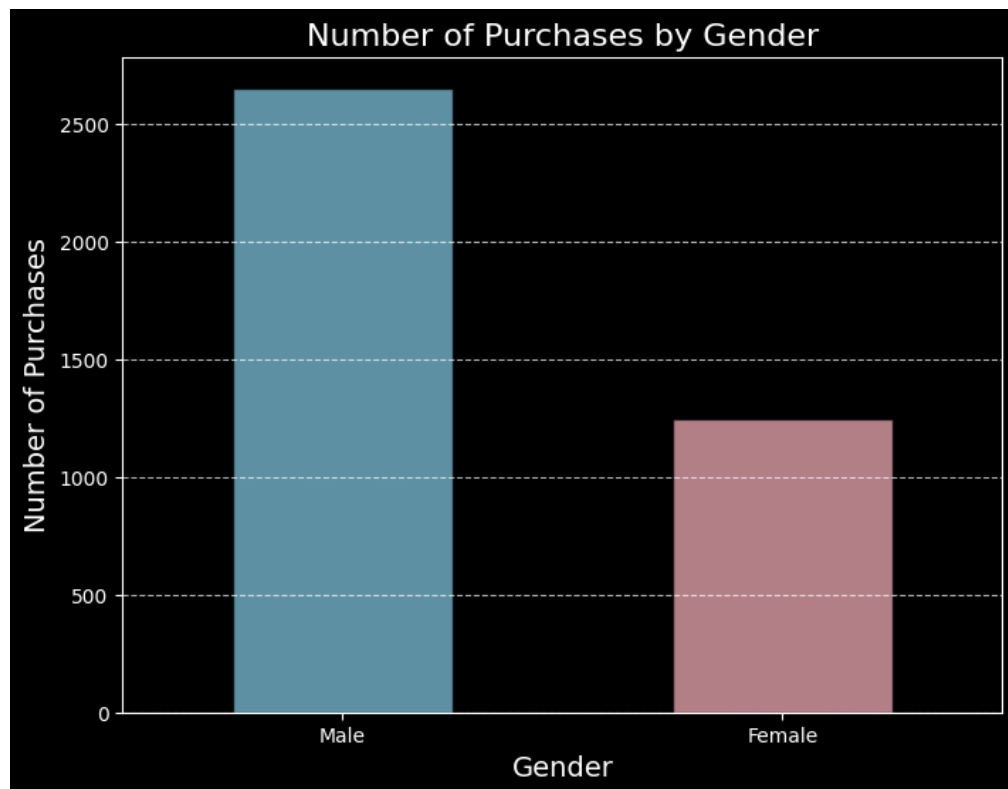- Males account for 68% of purchases, with higher spending than females.

Figure 2: Number of Purchases by Gender

3) Seasonal Trends:

- Highest spending occurs during winter, followed by summer.
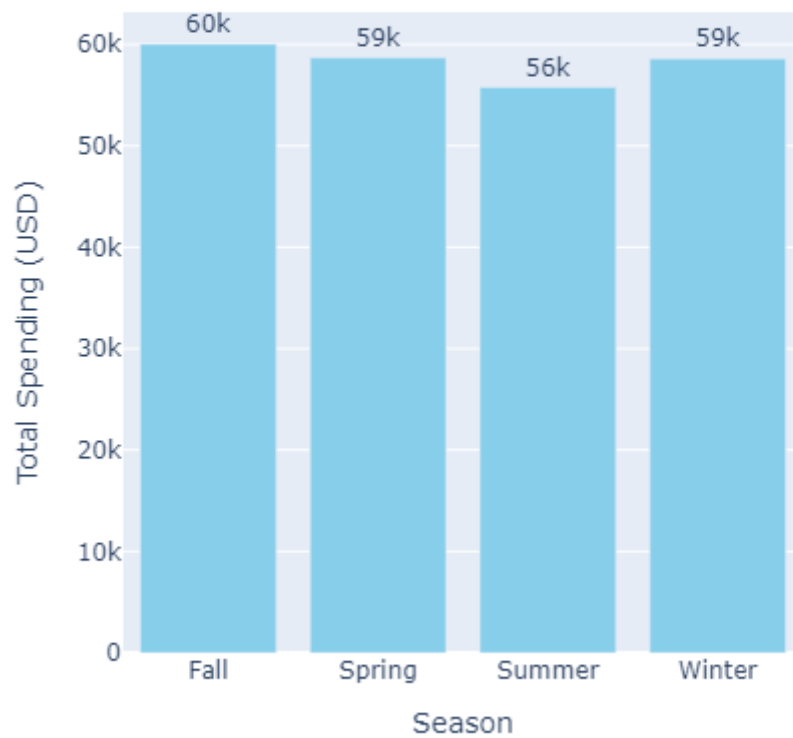

Figure 3: Customer Spending by Season

4.2 GitHub Link for Code: https://github.com/Rishipan/Identifying-Shopping-Trends

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

Future work can include:

- Incorporating real-time data for dynamic trend analysis.

- Applying machine learning models for predictive insights.

- Expanding analysis to additional datasets for broader insights.

5.2 Conclusion:

This project successfully analyzed shopping trends, uncovering critical patterns in customer behavior, product preferences, and seasonal spending. The insights can guide retailers in making strategic decisions, ultimately enhancing customer satisfaction and business performance.

REFERENCES

[1]. Jake VanderPlas, "Python Data Science Handbook", O'Reilly Media, 2016.
[2]. Wes McKinney, "Python for Data Analysis", O'Reilly Media, 2017.
[3]. John W. Tukey, "Exploratory Data Analysis", Addison-Wesley, 1977.
[4]. Hadley Wickham, "ggplot2: Elegant Graphics for Data Analysis", Springer, 2016.