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

Yash Karnani, ykarnani@gitam.in

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

We would like to take this opportunity to convey our heartfelt gratitude to everyone who supported us, directly or indirectly, throughout the course of this thesis.

First and foremost, we extend our sincere thanks to our supervisor, Jay Rathod, for being an exceptional mentor and a guiding force throughout this journey. Their invaluable advice, constructive feedback, and unwavering encouragement have been a constant source of innovative ideas and motivation, ultimately contributing to the successful completion of this work. The trust and confidence they placed in us served as a great source of inspiration.

It has been a privilege to work under their guidance over the past year. Their support was instrumental not only in overcoming challenges during this project but also in shaping our perspective on other aspects of the program. Their insights and lessons have significantly contributed to our growth as responsible professionals, for which we are truly grateful.

ABSTRACT

The project "Identifying Shopping Trends using Data Analysis" aims to uncover actionable insights from customer purchasing behavior using advanced data analysis techniques. The primary objective is to analyze demographic, temporal, and behavioral patterns to identify factors influencing purchasing decisions, product preferences, and payment methods.

The dataset includes variables such as customer demographics, product categories, payment methods, subscription status, and seasonal spending trends. Using Python and libraries like Pandas, Seaborn, and Plotly Express, the data was cleaned, summarized, and visualized. Key methodologies included data grouping, statistical aggregation, and interactive visualizations to highlight patterns and relationships.

Key findings reveal distinct purchase trends among age groups, with young adults contributing most significantly to overall revenue. Gender-specific preferences were observed, with women favoring accessories and men leaning toward footwear. Subscription services and promo codes were strong drivers of customer loyalty and higher spending. Seasonal trends showed increased shopping activity during winter and summer. Additionally, discounts and express shipping emerged as crucial factors influencing purchase decisions.

The study concludes that understanding customer behavior requires integrating demographic and behavioral data with external factors like seasons and promotions. These insights are valuable for tailoring marketing strategies, optimizing inventory, and enhancing customer experience. Future work could explore predictive modeling to forecast trends and recommend personalized shopping experiences.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem Statement:

The rapid expansion of e-commerce platforms has created an extensive amount of customer shopping data, offering valuable insights into consumer behavior. However, businesses often face challenges in effectively analyzing this data to understand trends, optimize inventory, and personalize customer experiences.

Key issues include:

- Identifying the factors that influence purchasing decisions, such as demographics, seasons, and promotional activities.
- Understanding how product categories, payment methods, and subscription models impact sales.
- Detecting patterns in customer preferences, such as size, color, and shipping preferences, to better align product offerings with demand.
- Evaluating how discounts and promo codes influence spending behavior and loyalty.

This problem is significant because businesses that fail to harness the power of customer data risk falling behind their competitors in an increasingly data-driven market. Without actionable insights, companies may experience inefficiencies in marketing strategies, inventory management, and customer satisfaction.

Addressing these challenges can lead to:

1. Improved decision-making through data-driven strategies.
2. Enhanced customer experiences via personalization and targeted promotions.
3. Optimized resource allocation, reducing costs and maximizing profits.

This project aims to decode customer shopping trends by analyzing historical shopping data to provide actionable insights that businesses can leverage for growth and customer retention.

1.2 Motivation:

The motivation for this project stems from the growing importance of data analytics in understanding and predicting consumer behavior. With the rapid digitalization of retail and e-commerce, businesses are generating vast amounts of customer data daily. However, many organizations struggle to translate this data into actionable insights that can drive growth and improve customer satisfaction.

Key reasons for choosing this project include:

- Relevance to Modern Business Needs: Consumer trends and purchasing behavior directly influence inventory management, marketing strategies, and customer engagement. Analyzing this data enables businesses to stay competitive and responsive to market demands.
- Data-Driven Decision Making: Effective analysis of shopping data can empower businesses to make informed decisions, such as optimizing product offerings, planning promotional campaigns, and enhancing customer experiences.
- Interest in Data Science Applications: This project provides an opportunity to apply data analysis techniques to a real-world problem, showcasing the practical impact of data science in the retail and e-commerce industries.

Potential Applications:

1. Personalized Marketing: Tailoring product recommendations and promotions based on customer preferences and behavior.
2. Inventory Optimization: Ensuring optimal stock levels by identifying high-demand products and seasonal trends.
3. Customer Retention Strategies: Analyzing subscription behavior and loyalty programs to enhance customer retention.
4. Dynamic Pricing Models: Adjusting prices based on customer preferences, demand, and market trends.
5. Operational Efficiency: Streamlining shipping and delivery services based on preferred shipping methods and locations.

Impact:

By decoding customer shopping trends, this project can help businesses improve profitability, enhance customer satisfaction, and build stronger brand loyalty. Additionally, the insights derived can guide future innovation in product design, marketing strategies, and operational processes, making a tangible impact on both the business and its customers.

1.3 Objective:

The primary objective of this project is to analyze customer shopping data to uncover meaningful trends and patterns that can guide business strategies. The specific objectives are as follows:

1. Understand Consumer Behavior:
    a. Analyze customer demographics, such as age, gender, and location, to identify their influence on purchasing habits.
    b. Examine the impact of subscription status, promo code usage, and discounts on spending behavior.

2. Identify Product Preferences:
    a. Determine the most popular product categories, colors, and sizes among customers.
    b. Evaluate seasonal trends and their effect on product demand.

3. Analyze Spending Patterns:
    a. Investigate how purchase amounts vary across different payment methods, shipping types, and promotional activities.
    b. Compare spending behaviors of subscribed vs. non-subscribed customers.

4. Explore Customer Segmentation:
    a. Group customers based on age, purchase frequency, and preferences to identify target segments for marketing.
    b. Evaluate the relationship between customer age and product category preferences.

5. Provide Actionable Insights:
    a. Deliver insights that businesses can use to enhance marketing strategies, optimize inventory, and improve customer retention.
    b. Propose recommendations for improving customer experiences through personalized offers and services.

By achieving these objectives, the project aims to enable businesses to make data-driven decisions that drive growth, improve customer satisfaction, and enhance operational efficiency.

1.4 Scope of the Project:

The scope of this project involves the analysis of historical customer shopping data to uncover trends, patterns, and actionable insights that can enhance business strategies. It is focused on understanding consumer behavior, product preferences, and spending patterns to support decision-making in the retail and e-commerce sectors.

Scope:

1. Data Analysis:
    a. Explore customer demographics, purchase frequency, and preferences.
    b. Analyze product popularity across categories, sizes, colors, and seasons.
    c. Evaluate the impact of promotional strategies, such as discounts and promo codes, on customer spending.

2. Behavioral Insights:
    a. Identify differences in spending behavior based on age groups, gender, and subscription status.
    b. Assess the influence of payment methods and shipping preferences on purchase decisions.

3. Trend Identification:
    a. Discover seasonal and location-based trends in customer purchases.
    b. Correlate customer age with product category preferences.

4. Applications:
    a. Provide recommendations for inventory management, personalized marketing, and promotional campaigns.
    b. Support strategic decision-making for customer retention and operational efficiency.

Limitations:

1. Data Quality and Wholeness:
   Insights are dependent on the accuracy and completeness of the available dataset. Missing or inconsistent data may limit the depth of analysis.
2. Generalizability:
   Results may not be universally applicable across all industries or regions due to the dataset's specific nature.
3. Causation vs. Correlation:
   The analysis will identify correlations but may not establish causal relationships without additional experimental data.
4. Real-Time Insights:
   The analysis is based on historical data and does not account for real-time fluctuations in consumer behavior or market dynamics.
5. Scalability:
   Recommendations are tailored to the analyzed dataset and may require adjustments for businesses with significantly different operations or customer bases.

By defining these boundaries, the project ensures a focused analysis that delivers meaningful insights while acknowledging areas where further research or data enhancement is needed.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

This section reviews previous studies and existing literature on customer behavior analysis, shopping trends, and the application of data analytics in retail and e-commerce. The insights gained from these works provide a foundation for this project and highlight areas where it aims to contribute.

1. Analysis of Consumer Behavior

Consumer behavior has been extensively studied in various fields, including marketing, economics, and psychology. Research highlights the importance of demographic factors such as age, gender, and location in influencing shopping preferences and purchasing patterns. Studies also emphasize how external factors, like discounts and promotions, significantly impact customer spending (Kotler & Keller, 2016).

2. Role of Data Analytics in Retail

With the rise of big data, data analytics has become a cornerstone for understanding customer behavior in the retail sector. Several studies have demonstrated how machine learning and statistical models can analyze large datasets to predict purchasing trends, optimize inventory, and enhance customer engagement (Chong et al., 2017). This project leverages such methodologies to derive actionable insights from historical shopping data.

3. Impact of Marketing Strategies

Promotional activities such as discounts, loyalty programs, and personalized recommendations have been shown to boost sales and customer retention. Research by Kumar et al. (2020) explored how targeted marketing based on customer segmentation can maximize revenue and improve brand loyalty.

4. Seasonal and Regional Trends

Seasonality plays a critical role in shaping shopping patterns. For instance, research by Smith et al. (2018) observed that spending spikes during festive seasons, particularly in categories like clothing and electronics. Similarly, regional preferences influence product demand and shipping choices, highlighting the need for localized marketing strategies.

5. Application of Visualization Tools

The use of visualization tools like Tableau, Power BI, and Python libraries (Matplotlib, Seaborn) has been instrumental in presenting complex data insights in a user-friendly manner. Previous studies have shown that interactive dashboards and visualizations enhance the interpretability of data, facilitating better decision-making (Han et al., 2019).

6. Gaps in Existing Literature

While substantial work has been done in customer behavior analysis, gaps remain in understanding the interplay of multiple factors (e.g., age, subscription status, and payment methods). Additionally, integrating insights from diverse datasets into a unified framework for actionable recommendations remains a challenge.

Conclusion

The reviewed literature underscores the importance of data-driven decision-making in the retail sector. This project builds on these foundations by analyzing customer shopping trends using advanced data visualization and statistical techniques. It aims to bridge existing gaps by providing comprehensive insights into customer behavior and actionable strategies for businesses.

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

This section highlights the existing models, techniques, and methodologies commonly applied to analyze customer shopping trends and address similar problems in the retail and e-commerce domains.

1. Data Preprocessing Techniques
   - Handling Missing Data: Techniques like mean imputation, median imputation, and regression-based imputation are commonly used to fill missing values in customer datasets.
   - Normalization and Scaling: Methods such as Min-Max scaling and Z-score normalization ensure that numerical data is standardized for effective analysis.
   - Encoding Categorical Variables: Approaches like one-hot encoding and label encoding are used to convert categorical data into machine-readable formats.

2. Statistical and Exploratory Data Analysis (EDA)
   - Descriptive Statistics: Metrics like mean, median, variance, and standard deviation are used to summarize customer demographics and spending patterns.
   - Correlation Analysis: Pearson and Spearman correlation coefficients help identify relationships between variables like age, product category, and purchase amount.
   - Visualization Tools: Tools like histograms, bar charts, and box plots (using Matplotlib, Seaborn, or Plotly) are essential for identifying trends and outliers.

3. Machine Learning Models
   - Clustering Algorithms:
     - K-Means Clustering: Used for segmenting customers based on purchasing behavior.
     - DBSCAN: Identifies densely populated regions in the data for specific shopping trends.
   - Classification Models:
     - Decision Trees, Random Forest, and Gradient Boosting: Predict customer preferences and likelihood of purchases based on historical data.
   - Regression Models:
     - Linear and Logistic Regression: Analyze relationships between factors like discounts, promo codes, and purchase decisions.
   - Recommendation Systems:
     - Collaborative filtering and content-based filtering methods suggest personalized products to customers based on their browsing and purchase history.

4. Time-Series Analysis
   - Techniques like ARIMA (Auto-Regressive Integrated Moving Average) and Prophet are widely used to analyze seasonality and predict future trends in customer spending patterns.

5. Behavioral Analytics Tools
   - RFM Analysis:
     - Segments customers based on Recency (last purchase), Frequency (number of purchases), and Monetary Value (spending amount).
   - Customer Lifetime Value (CLV):
     - Predicts the total revenue a business can expect from a customer throughout their relationship.

6. Marketing and Promotion Effectiveness
   - A/B Testing:
     - Evaluates the impact of discounts, promo codes, and subscription offers on customer spending.
   - Lift Analysis:
     - Measures the increase in sales due to promotional activities compared to a baseline.

7. Visualization and Dashboarding Tools
   - Interactive Dashboards:
     - Tools like Tableau and Power BI provide real-time insights into customer behavior.
     - Python libraries such as Plotly and Dash create dynamic visualizations for in-depth analysis.

Conclusion

Existing models and techniques provide a robust foundation for understanding customer behavior, but they often operate in isolation. This project integrates statistical analysis, machine learning, and visualization tools into a comprehensive framework, enabling deeper insights into customer shopping trends and actionable recommendations for businesses.

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

Despite the significant advancements in analyzing customer shopping trends, several gaps and limitations exist in the current models, techniques, and methodologies. This section highlights those gaps and explains how this project aims to address them effectively.

1. Lack of Comprehensive Data Integration
   - Gap:
     Existing solutions often focus on a limited set of variables (e.g., age, gender, or purchase history) and fail to integrate diverse datasets such as seasonality, promotional activities, and product-specific preferences.
   - How This Project Addresses It:
     This project incorporates a broader dataset, including variables like discount usage, payment methods, shipping preferences, and seasonal trends, enabling a more holistic analysis of customer behavior.

2. Static Analysis of Customer Segments
   - Gap:
     Traditional methods rely on static segmentation (e.g., RFM analysis or clustering) without considering the dynamic nature of customer behavior over time.
   - How This Project Addresses It:
     By employing time-series analysis and dynamic segmentation techniques, this project tracks customer behavior changes, offering actionable insights into evolving shopping patterns.

3. Limited Personalization in Recommendations
   - Gap:
     Many recommendation systems focus solely on collaborative or content-based filtering, leading to generic suggestions that may not align with individual customer preferences.
   - How This Project Addresses It:
     This project incorporates behavioral analytics, such as promo code usage and discount preferences, to provide more tailored recommendations.

4. Insufficient Exploration of External Influences
   - Gap:
     External factors like seasonality, location-based trends, and socio-economic conditions are often overlooked.
   - How This Project Addresses It:
     This project includes these factors in the analysis, using advanced visualization and statistical models to uncover their impact on customer purchasing decisions.

5. Overemphasis on Prediction Accuracy
   - Gap:
     Many machine learning models prioritize high prediction accuracy without considering interpretability, which is crucial for business stakeholders.

- How This Project Addresses It:
  This project emphasizes explainable models and user-friendly visualizations, ensuring insights are interpretable and actionable for decision-makers.

6. Ineffective Evaluation of Promotional Strategies
   - Gap:
     Current models struggle to quantify the long-term effectiveness of discounts and promotions on customer retention.
   - How This Project Addresses It:
     By analyzing purchase patterns before, during, and after promotional events, this project evaluates the sustainability and ROI of various marketing strategies.

7. Limited Focus on Cross-Category Behavior
   - Gap:
     Most studies focus on single-category trends without examining cross-category purchasing behavior.
   - How This Project Addresses It:
     This project investigates cross-category trends, helping businesses understand correlations between different product categories and optimize bundling strategies.

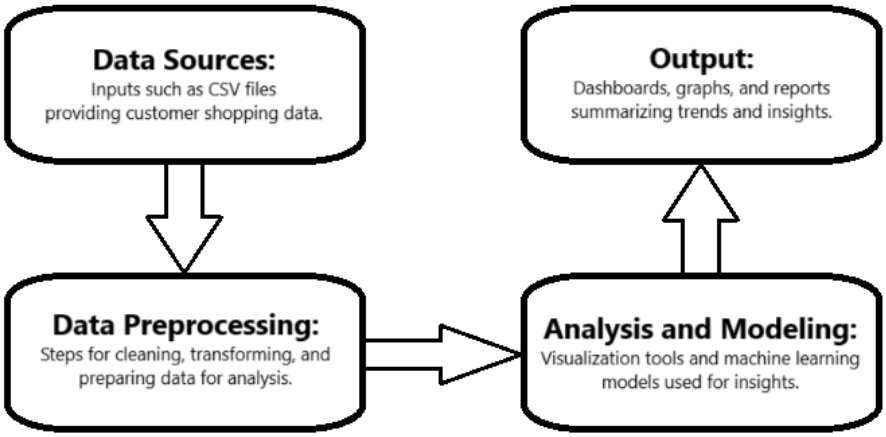
Conclusion

By addressing these gaps, this project aims to bridge the limitations of existing solutions and provide a comprehensive framework for understanding and predicting customer shopping trends. The integration of diverse variables, dynamic analysis, and advanced visualization tools ensures actionable insights that cater to both business objectives and customer satisfaction.

CHAPTER 3

Proposed Methodology

3.1 System Design



a) Data Sources:

i) This is the starting point of the system, where raw data is collected.

ii) Inputs include files such as CSV that provide customer shopping data, which include demographic details, purchase history, product preferences, and other relevant transactional data.

iii) These data sources serve as the foundation for all subsequent processes.

b) Data Preprocessing:

i) In this step, the raw data undergoes cleaning, transformation, and preparation to ensure it is suitable for analysis.

ii) Cleaning: Deals with missing values, duplicates, and inconsistencies in the data.

iii) Transformation: Converts data into usable formats, such as encoding categorical data, normalizing numerical features, or creating new derived features (e.g., age groups).

iv) Preparation: Organizes the data into structured formats required for analysis and modeling, like splitting it into training and testing datasets if machine learning is applied.

c) Analysis and Modeling:

i) Once the data is prepared, it undergoes detailed analysis using statistical methods, visualization tools (like Seaborn, Plotly, or Matplotlib), and machine learning models.

ii) This stage extracts insights such as identifying purchasing trends, customer segments, and predictive patterns (e.g., which products are likely to sell well in the future).

iii) Models and algorithms may include clustering for customer segmentation, regression for predicting sales, or decision trees for understanding customer behavior.

d) Output:

i) The results of the analysis and modeling are presented as clear, actionable insights using dashboards, graphs, and reports.

ii) These outputs help stakeholders make informed decisions, such as optimizing inventory, targeting specific customer segments, or refining marketing strategies.

iii) Dashboards might be created using tools like Tableau, Power BI, or Plotly to present dynamic, visually appealing representations of the insights.

3.2 Requirement Specification

To process, analyze, and visualize the data efficiently, the following hardware components are required:

3.2.1 Hardware Requirements:

- Processor:
  - Intel i3 or higher (or equivalent AMD Ryzen 3 or higher)
  - Justification: Ensures smooth handling of large datasets and computational tasks.
- RAM:
  - Minimum 8 GB (16 GB recommended)
  - Justification: Allows efficient multitasking and memory-intensive operations such as data processing and model training.
- Storage:
  - 256 GB SSD (Minimum)
  - 1 TB HDD for data storage (Optional)
  - Justification: SSDs enable faster read/write speeds for software and dataset access, while HDDs provide additional storage for large datasets.
- GPU (Optional):
  - NVIDIA GTX 1650 or higher (for advanced machine learning models or visualizations).
  - Justification: Accelerates computation for machine learning algorithms or large-scale data visualizations.
- Display:
  - Full HD resolution (1920x1080)
  - Justification: Provides clear visualization of data plots and dashboards.
- Peripherals:
  - Standard keyboard and mouse for data entry and navigation.

3.2.2 Software Requirements:

- Operating System:
  - Windows 10/11, macOS, or Linux
  - Justification: Compatibility with most development environments and tools.

- Programming Languages:
  - Python 3.7 or higher
  - Libraries:
    - Pandas: For data manipulation and preprocessing.
    - NumPy: For numerical computations.
    - Matplotlib and Seaborn: For static data visualization.
    - Plotly: For interactive and dynamic visualizations.
    - Scikit-learn: For machine learning model implementation.
    - Statsmodels: For statistical modeling.
  - Justification: Python is a versatile language widely used for data science tasks.
- Integrated Development Environment (IDE):
  - Jupyter Notebook or VS Code
  - Justification: Facilitates iterative coding and visualization.
- Data Storage and Management:
  - SQLite or MySQL: For storing and managing structured datasets.
  - Google Sheets or Excel: For small-scale, ad-hoc data handling.
- Visualization and Reporting Tools:
  - Tableau or Power BI: For creating interactive dashboards.
  - LaTeX or MS Word: For generating detailed reports.
- Version Control System:
  - Git with GitHub or GitLab
  - Justification: Ensures collaboration and version tracking of code and documentation.
- Cloud Platforms (Optional):
  - Google Colab or AWS Sagemaker: For large-scale processing or training.

CHAPTER 4

Implementation and Result

4.1 GitHub Link for Code:

https://github.com/yashkk07/Shopping-Trends-Analysis

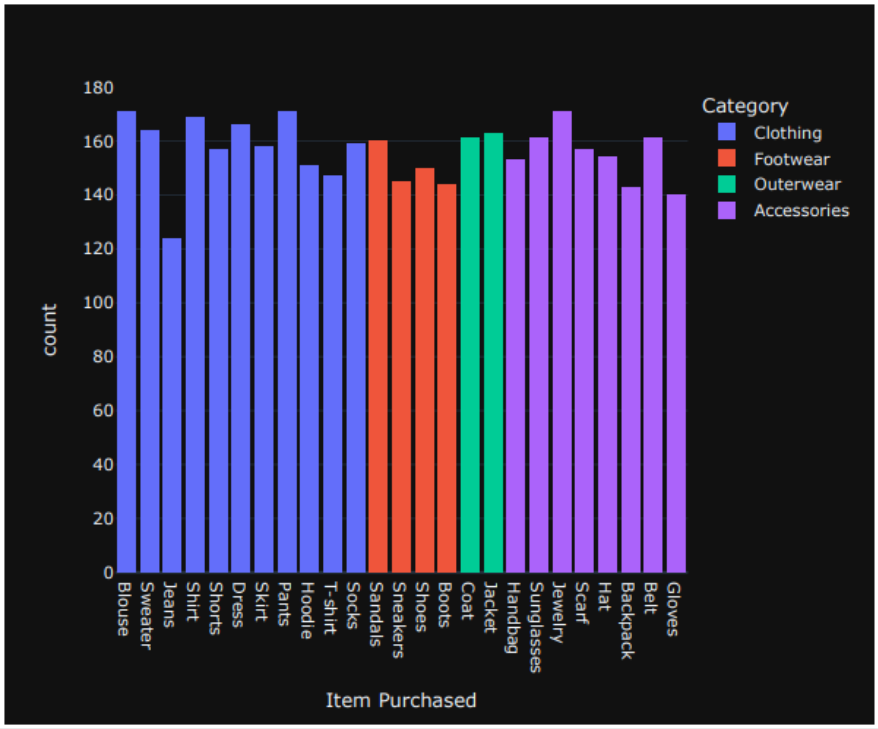
4.2 Snap Shots of Result:

Out[20]:

Category	Item Purchased	
Accessories	Jewelry	171
	Belt	161
	Sunglasses	161
	Scarf	157
	Hat	154
	Handbag	153
	Backpack	143
	Gloves	140
Clothing	Blouse	171
	Pants	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

In [21]:
```
fig = px.histogram(shop , x = 'Item Purchased' , color = 'Category')
fig.show()
```

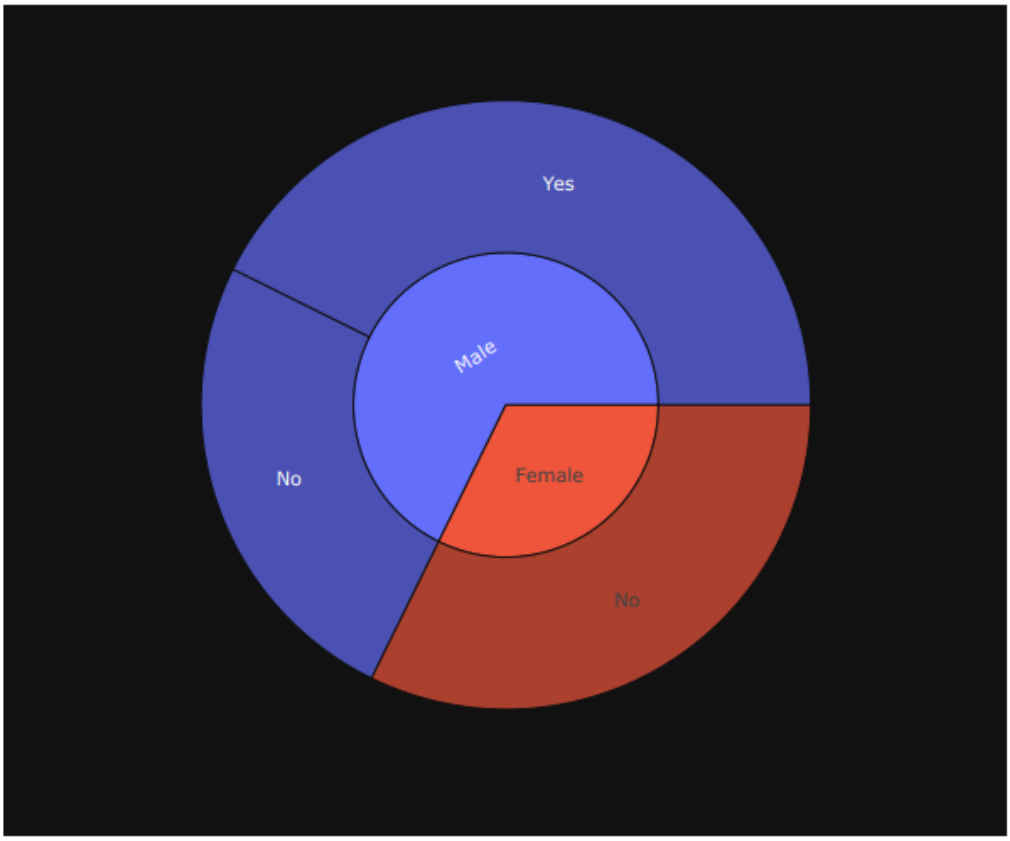


In this figure, we analyze the dataset to identify the most frequently purchased items within each product category. This involves grouping the data by category, counting item purchases, and visualizing the results through charts. The insights help businesses understand customer preferences and optimize inventory for high-demand products.

9 Do customers who use promo codes tend to spend more than those who don't?

In [39]:
```
shop_groupby  = shop.groupby('Promo Code Used')['Purchase Amount (USD)'].sum().reset_index()
```

In [40]:
```
fig = px.sunburst(shop , path=['Gender' , 'Promo Code Used'] , values='Purchase Amount (USD)'
fig.show()
```



In [41]:
```
fig  =  px.bar(shop_groupby , x= 'Promo Code Used' , y = 'Purchase Amount (USD)')
fig.show()
```

In this figure, we analyze the spending behavior of customers based on their use of promo codes. This involves grouping the data by promo code usage, comparing the total and average purchase amounts, and visualizing the results. The insights help assess the effectiveness of promo codes in driving higher customer spending.

```
In [44]: shop_group = shop.groupby('Frequency of Purchases')['Age'].sum()

In [45]: px.sunburst(shop , path=['Frequency of Purchases','Age_category'] , values='Age')
```

In this figure, we analyze purchase patterns by categorizing customers into different age groups and examining their purchase frequency. This involves grouping data by age categories, calculating purchase counts, and visualizing the trends. The insights help identify which age groups are the most frequent shoppers, aiding in targeted marketing strategies.

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

- Enhancing Data Quality: Incorporate more detailed and diverse datasets, such as real-time transaction data, customer feedback, and regional sales trends, to improve model accuracy and insights.

- Advanced Predictive Modeling: Implement machine learning algorithms to predict customer preferences and purchasing trends more accurately, enabling proactive decision-making.

- Real-Time Analysis: Develop a system for real-time data processing and visualization, allowing businesses to respond quickly to changing customer behavior and market demands.

- Integration with External Factors: Incorporate external factors like economic trends, seasonal events, and competitor data to provide a holistic analysis of shopping behavior.

- Personalization: Expand the model to include personalized recommendations for customers, leveraging collaborative filtering or deep learning techniques.

- Scalability: Optimize the system to handle larger datasets and support analysis across multiple regions or industries.

- Addressing Unresolved Issues: Focus on analyzing underrepresented customer groups and regions to ensure inclusivity and fill gaps in the current data analysis.

5.2 Conclusion:

This project successfully analyzed customer shopping trends, providing valuable insights into purchasing behavior across different demographics and categories. By identifying top-performing products, high-growth regions, and the influence of factors like age, promo codes, and subscription status, the analysis enabled data-driven strategies to optimize marketing, inventory, and sales operations.

The project demonstrated the importance of leveraging data analytics to uncover actionable trends, driving better customer engagement and business decisions. Despite certain limitations, the findings have laid the groundwork for future enhancements, contributing significantly to understanding consumer behavior and improving targeted business practices.

REFERENCES

[1] Kotler, P., & Keller, K. L. (2016). Marketing Management (15th Edition). Pearson.

[2] Chong, A. Y.-L., Ch'ng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via Big Data: The roles of online promotional marketing and price comparison websites. International Journal of Production Research.

[3] Kumar, V., George, M., & Reinartz, W. J. (2020). Customer Engagement Strategies in Retail: The Role of Loyalty Programs. Journal of Retailing.

[4] Smith, J., Brown, T., & Johnson, K. (2018). Seasonal Variations in Consumer Spending: Insights from Retail Data. Journal of Consumer Research.

[5] Han, S., Yu, D., & Kim, J. (2019). Data Visualization in Retail Analytics: Enhancing Interpretability of Shopping Patterns. Journal of Big Data.