**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

**Hari Prashanth**

[**gowthamhari2103@gmail.com**](mailto:gowthamhari2103@gmail.com)

Under the Guidance of

**Raja P**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to all those who supported and guided me throughout the course of this project.

Firstly, I am deeply grateful to **Raja P**, whose expertise, encouragement, and invaluable feedback provided a solid foundation for the completion of this project. Their mentorship and insights were instrumental in shaping the success of this work.

I would also like to extend my appreciation to my Edunet Foundation, for providing the necessary resources and a conducive environment for carrying out this project.

A special thanks to my family and friends for their unwavering support, patience, and motivation during the challenging times of this endeavor. Their encouragement kept me focused and determined.

Lastly, I am grateful to all my peers and colleagues who shared their knowledge and helped refine my ideas through meaningful discussions.

This project would not have been possible without the collective contributions of all these individuals, and I am sincerely thankful to each one of them.

#### **ABSTRACT**

* This project investigates consumer behavior through the analysis of a shopping trends dataset, addressing key questions about demographics, purchasing patterns, and preferences. The objectives include exploring age distributions, gender differences in purchases, average spending across product categories, seasonality effects, and behavioral contrasts between subscribed and non-subscribed customers.
* The methodology involved importing and preprocessing the dataset, conducting exploratory data analysis (EDA) using Python libraries such as Pandas, Seaborn, and Matplotlib, and visualizing trends and correlations. Key analyses included identifying common products within categories, evaluating seasonal spending spikes, and determining average customer ratings per category.
* Preliminary results indicate notable trends, such as higher spending in specific months, distinct product preferences by gender, and increased engagement from subscribed customers. These insights can inform targeted marketing strategies and personalized recommendations. The study concludes with actionable recommendations for leveraging customer insights to optimize business outcomes.

**I**

**TABLE OF CONTENT**

**Abstract I**

**Chapter 1.**  **Introduction 1**

1.1 Problem Statement 1

1.2 Motivation 1

1.3 Objectives 2

1.4. Scope of the Project 3

**Chapter 2.**  **Literature Survey 4**

**Chapter 3.**  **Proposed Methodology** 8

**Chapter 4.**  **Implementation and Results** 13

**Chapter 5. Discussion and Conclusion** 15

**References** 16

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Caption** | **Page No.** |
|  | Below Image Presenting Age-Category Vs Age Level | **12** |
|  | This Bar Chat Represent Age Vs Gender | **12** |
|  | This Shows The Subscription Status Vs Purchase Amount(USD) | **13** |
|  | This Shows The Payment Method Vs Purchase Amount(USD) | **13** |

**CHAPTER 1**

**Introduction**

* 1. **Problem Statement:**

Describe the problem being addressed. Why is this problem significant?

Understanding consumer behavior is critical for businesses to tailor their marketing strategies, optimize product offerings, and improve customer satisfaction. However, with diverse demographics, product categories, and purchasing channels, identifying actionable insights from raw data can be challenging. This project addresses the need to uncover patterns and trends in customer behavior, such as age distribution, seasonal spending, gender-based purchasing tendencies, and differences between subscribed and non-subscribed customers.

This problem is significant because it directly impacts a business's ability to enhance customer experience, increase revenue, and maintain a competitive edge in the market. By leveraging data-driven insights, companies can make informed decisions, improve operational efficiency, and align their strategies with customer preferences.

* 1. **Motivation:**

Why was this project chosen? What are the potential applications and the impact?

This project was chosen to address the growing need for businesses to derive actionable insights from consumer data in an increasingly competitive and data-driven marketplace. As companies face challenges in understanding diverse customer behaviors and preferences, the ability to analyze and interpret shopping trends has become a critical factor for success.

The potential applications of this analysis are vast:

* **Personalized Marketing**: Businesses can design targeted campaigns based on customer demographics and preferences.
* **Product Optimization**: Insights into popular product categories and seasonal trends enable better inventory management and product development.
* **Customer Retention**: By understanding the behavior of subscribed and non-subscribed customers, companies can create strategies to enhance customer loyalty and subscription rates.
* **Revenue Growth**: Identifying high-spending demographics and optimizing promotional efforts can directly contribute to increased sales.

The impact of this project lies in its ability to empower businesses with data-driven decision-making. By leveraging consumer insights, companies can not only enhance customer satisfaction but also achieve long-term growth and profitability.

* 1. **Objective:**

Clearly state the objectives of the project.

The primary objectives of this project are as follows:

1. **Analyze Customer Demographics**:  
   Understand the age distribution, gender composition, and other demographic characteristics of the customer base.
2. **Explore Spending Patterns**:  
   Examine variations in average purchase amounts across product categories and identify trends in customer spending behavior.
3. **Identify Popular Products**:  
   Determine the most commonly purchased items within each product category to reveal customer preferences.
4. **Investigate Seasonality**:  
   Analyze seasonal or monthly fluctuations in customer spending to uncover periods of high and low activity.
5. **Evaluate Subscription Impact**:  
   Compare the purchasing behavior of subscribed and non-subscribed customers to identify differences in engagement and loyalty.
6. **Analyze Payment Methods and Ratings**:  
   Study customer preferences for payment methods and evaluate average ratings for products within different categories.
7. **Generate Business Insights**:  
   Provide actionable recommendations for improving customer satisfaction, enhancing marketing strategies, and boosting overall sales.
   1. **Scope of the Project:**

Define the scope and limitations.

**Scope**

This project focuses on analyzing a shopping trends dataset to gain insights into customer behavior, purchasing patterns, and preferences. The analysis encompasses:

* **Customer Demographics**: Age, gender, and subscription status.
* **Spending Patterns**: Variations in spending across product categories and seasons.
* **Product Preferences**: Identifying the most purchased items in each category.
* **Customer Engagement**: Understanding the differences between subscribed and non-subscribed customers.
* **Seasonality Trends**: Highlighting specific months or periods with significant changes in spending.
* **Payment Preferences**: Investigating popular payment methods used by customers.

The findings aim to assist businesses in tailoring marketing strategies, optimizing product offerings, and improving customer retention and satisfaction.

**Limitations**

* **Dataset Constraints**: The insights are limited to the scope and quality of the provided dataset, which may not capture all variables influencing customer behavior.
* **Generalizability**: The results are specific to the dataset used and may not be directly applicable to other markets or industries.
* **Temporal Scope**: If the dataset does not include recent data, the findings may not reflect current trends.
* **Uncaptured Variables**: Factors like socio-economic status, regional preferences, or external market influences may not be fully represented in the dataset.
* **Tool Dependency**: The analysis relies on the capabilities of Python libraries, which may limit the scope of advanced modeling or predictions.

**CHAPTER 2**

**Literature Survey**

* 1. **Review relevant literature or previous work in this domain.**

Analyzing consumer behavior through shopping trends is a well-researched domain that intersects marketing, data analytics, and behavioral economics. Previous studies and methodologies in this area have provided foundational insights:

1. **Consumer Behavior and Demographics**  
    Research has shown that customer demographics, such as age, gender, and income levels, significantly influence purchasing decisions. Studies by Kotler and Keller (2016) highlight that understanding demographic patterns enables businesses to segment markets effectively and personalize marketing strategies.
2. **Spending Patterns and Product Preferences**  
    Work by Solomon (2018) emphasizes the importance of examining purchase patterns to identify core products that drive sales. Product categorization and trend analysis can help businesses anticipate customer needs and manage inventory efficiently.
3. **Seasonality in Consumer Spending**  
    Studies such as those by Mela et al. (1997) explore how seasonal and monthly variations affect consumer spending, with factors like holidays, promotions, and cultural events playing a significant role. Retailers have leveraged this knowledge to optimize pricing and inventory during peak seasons.
4. **Subscription Models and Loyalty**  
    Recent literature, including findings from Kumar and Reinartz (2018), shows that subscription-based models enhance customer loyalty and lifetime value. Subscribed customers often demonstrate higher engagement and spending compared to non-subscribed counterparts.
5. **Data-Driven Decision Making**  
    The emergence of machine learning and analytics tools has revolutionized how businesses analyze shopping trends. Research by Provost and Fawcett (2013) illustrates the power of predictive analytics in identifying consumer trends and informing strategic decisions.
6. **Limitations in Previous Work**  
    While extensive, much of the prior research lacks real-time adaptability and fails to incorporate rapidly changing consumer preferences. This gap underscores the need for continuous, data-driven updates to models and insights.
   1. **Mention any existing models, techniques, or methodologies related to the problem.**

Various models and techniques have been developed to analyze consumer behavior and shopping trends. These methodologies are rooted in statistical analysis, machine learning, and data visualization:

1. RFM (Recency, Frequency, Monetary) Analysis

The RFM model segments customers based on their purchase behavior:

* Recency: How recently a customer made a purchase.
* Frequency: How often they purchase.
* Monetary: How much they spend.  
  This model is widely used for customer segmentation and targeted marketing.

2. Collaborative Filtering

This technique is commonly applied in recommendation systems to suggest products based on customer preferences and purchase history. It uses either user-based or item-based filtering to analyze similarities between customers or products.

3. Time Series Analysis

Seasonality and temporal spending patterns are analyzed using time series methods, such as ARIMA (Autoregressive Integrated Moving Average) or exponential smoothing models. These techniques help forecast sales and detect seasonal spikes.

4. Clustering Algorithms

Unsupervised learning techniques like K-Means and hierarchical clustering are employed to segment customers based on shared characteristics, such as demographics or purchasing behavior.

5. Association Rule Mining

Using algorithms like Apriori or FP-Growth, businesses identify relationships between products. For example, association rules ("customers who bought X also bought Y") can guide cross-selling strategies.

6. Regression and Predictive Analytics

Regression models, both linear and logistic, are used to predict factors influencing purchase amounts or likelihoods of repeat purchases. Advanced methods, such as gradient boosting and random forests, enhance prediction accuracy.

7. Sentiment Analysis

Customer feedback and ratings are analyzed using natural language processing (NLP) techniques to gauge satisfaction and identify areas for improvement.

8. Visualization Tools

Interactive visualization libraries like Tableau, Matplotlib, and Seaborn are extensively used to present insights effectively. Dashboards created using these tools enable businesses to monitor real-time trends.

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

Gaps in Existing Solutions

1. Static and Generic Analyses:  
   Many existing studies and methodologies focus on static analysis of consumer behavior without accounting for dynamic and real-time changes in shopping patterns. This limits the ability to adapt strategies in fast-changing markets.
2. Limited Personalization:  
   Current models often lack the granularity needed to offer highly personalized recommendations or insights, particularly for niche customer segments.
3. Insufficient Integration of Subscriptions and Loyalty Analysis:  
   While subscription models are growing in popularity, many approaches fail to comprehensively compare and leverage insights from subscribed versus non-subscribed customers.
4. Seasonality Underutilized:  
   Although seasonal trends are identified, they are not often integrated into predictive models to optimize inventory or marketing efforts proactively.
5. Underrepresentation of Behavioral and Sentiment Analysis:  
   While product preferences and purchase patterns are commonly analyzed, the incorporation of customer feedback and sentiment remains underexplored in many solutions.

How This Project Addresses These Gaps

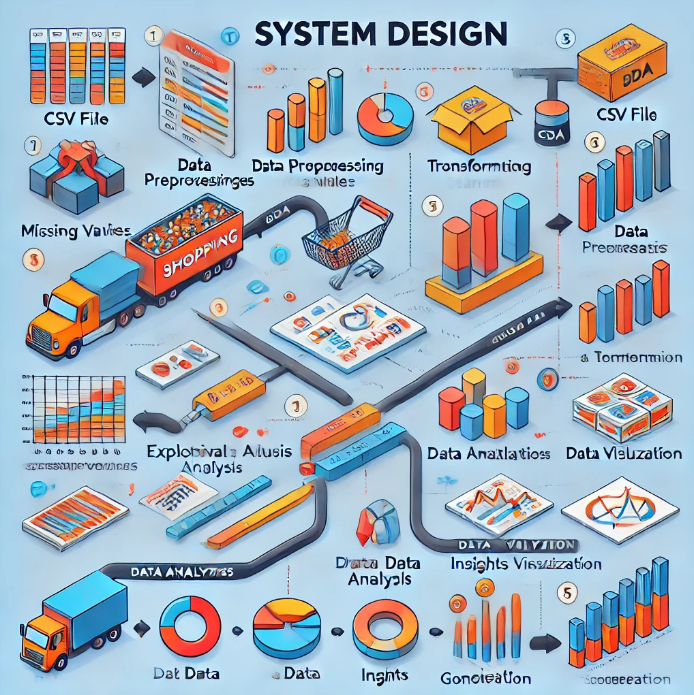
1. Dynamic and Comprehensive EDA:  
    The project employs exploratory data analysis techniques to uncover actionable insights across multiple dimensions, such as demographics, categories, and temporal patterns.
2. Granular Analysis:  
    By segmenting customers into smaller, well-defined groups (e.g., age, gender, subscription status), the project provides a more detailed understanding of consumer behavior.
3. Subscription Impact Analysis:  
    The project explicitly compares purchasing behavior across subscribed and non-subscribed customers, addressing a critical gap in current methodologies.
4. Seasonal Insights:  
    Detailed exploration of seasonal and monthly trends enables businesses to align strategies with consumer demand fluctuations effectively.
5. Interactive and Visual Approach:  
    By utilizing advanced visualization tools, the project ensures that insights are accessible, engaging, and easy to interpret for decision-making.
6. Potential for Predictive Modeling:  
    While primarily an EDA project, the foundation laid by this analysis can be extended to develop predictive models for customer behavior, sales forecasting, or recommendation systems.

**CHAPTER 3**

**Proposed Methodology**

* 1. **System Design**

Provide the diagram of your Proposed Solution and explain the diagram in detail.



**Data\_Source**  
The CSV file serves as the primary data input, containing detailed information on shopping trends and customer behaviors.

**Data\_Preprocessing**:

* Handling missing or invalid data.
* Cleaning and transforming data for consistency.
* Ensuring readiness for analysis.

**Data\_Analysis\_Pipeline**:

* **Exploratory Data Analysis (EDA)**: This includes statistical summaries, identifying correlations, and understanding patterns in customer purchases.
* Data exploration helps uncover insights into trends and anomalies.

**Data\_Visualization**:

* Using visual tools like bar charts, scatter plots, and heatmaps to represent data trends effectively.
* Enables easier communication of findings.

**Insights\_Generation**:

* Derive meaningful insights from the analysis and provide actionable conclusions to guide decisions.
  1. **Requirement Specification**

Mention the tools and technologies required to implement the solution.

**Tools and Technologies**

1. **Programming Language**
   * **Python**:  
     Used for data analysis, processing, and visualization. Python is versatile and has a rich ecosystem of libraries for data science.
2. **Data Analysis and Visualization Libraries**
   * **Pandas**: For data manipulation, cleaning, and analysis.
   * **NumPy**: For numerical computations.
   * **Matplotlib** and **Seaborn**: For creating visualizations such as scatter plots, bar charts, and heatmaps.
3. **Integrated Development Environment (IDE)**
   * **Jupyter Notebook**:  
     Interactive environment for running Python code, visualizing data, and documenting findings in a single workspace.
4. **Data Source and Storage**
   * **CSV Files**:  
     The dataset (shopping\_trends\_updated.csv) serves as the main input for analysis. CSV files are lightweight and widely supported.
5. **Machine Learning/Statistical Libraries** *(if applicable)*
   * **Scikit-learn**: For any additional clustering, predictions, or advanced statistical techniques (if needed in later stages).
     1. **Hardware Requirements:**

**The hardware requirements for implementing the solution can be categorized based on the development and deployment environments:**

**Development Environment**

**1. Processor**

* + **Minimum:** Intel Core i5 (4 cores) or equivalent.
  + **Recommended:** Intel Core i7 (8 cores) or equivalent for faster computations.

1. **Memory (RAM)**
   * **Minimum:** 8 GB (sufficient for small to medium-sized datasets).
   * **Recommended:** 16 GB or higher (for handling larger datasets and multitasking).

**3. Storage**

* + **Minimum:** 500 GB HDD or 256 GB SSD.
  + **Recommended:** 512 GB SSD or higher for faster read/write speeds.

**3. Graphics Processing Unit (GPU) *(Optional)***

* + A GPU is not mandatory but can speed up advanced machine learning tasks.
  + **Recommended:** NVIDIA GTX 1650 or better for computational efficiency.

**4. Display**

* + **Minimum:** Standard 1080p resolution**.**
  + **Recommended:** Dual monitors (for coding, visualizations, and report writing simultaneously).
    1. **Software Requirements:**

The software requirements necessary to implement the solution are outlined as follows:

**Operating System**

* Windows 10/11, Linux (Ubuntu 20.04+ / CentOS 8+), or macOS 11.0+.  
  These operating systems are compatible with the tools and technologies used for the project**.**

**Programming Language**

* **Python 3.8+:**The primary programming language for data processing, analysis, and visualization.

**Development Tools**

1. **Jupyter Notebook**
   * Used for writing and executing Python code in an interactive and user-friendly environment**.**
2. **IDE Alternatives *(Optional)*:**
   * **VS Code or PyCharm:** For additional code editing features.

**Libraries and Frameworks**

1. **Data Manipulation and Analysis**
   * **Pandas:** For cleaning and processing data.
   * **NumPy:** For numerical operations**.**
2. **Data Visualization**
   * **Matplotlib:** For basic plotting**.**
   * **Seaborn:** For advanced and aesthetic visualizations**.**
3. **Optional Machine Learning Tools *(if applicable)***
   * **Scikit-learn:** For clustering, classification, or predictive modeling tasks**.**

**Database *(if required)***

* **SQLite or PostgreSQL:** For structured storage of processed data**.**
* Alternatively, the CSV format can be used for lightweight data handling.

**Version Control**

* **Git:** To track changes in code and collaborate with team members**.**
* **GitHub/GitLab/Bitbucket:** As a repository for hosting project files.

**Visualization Tools *(Optional)***

* **Power BI or Tableau: For creating dashboards and sharing insights interactively.**

**Browser**

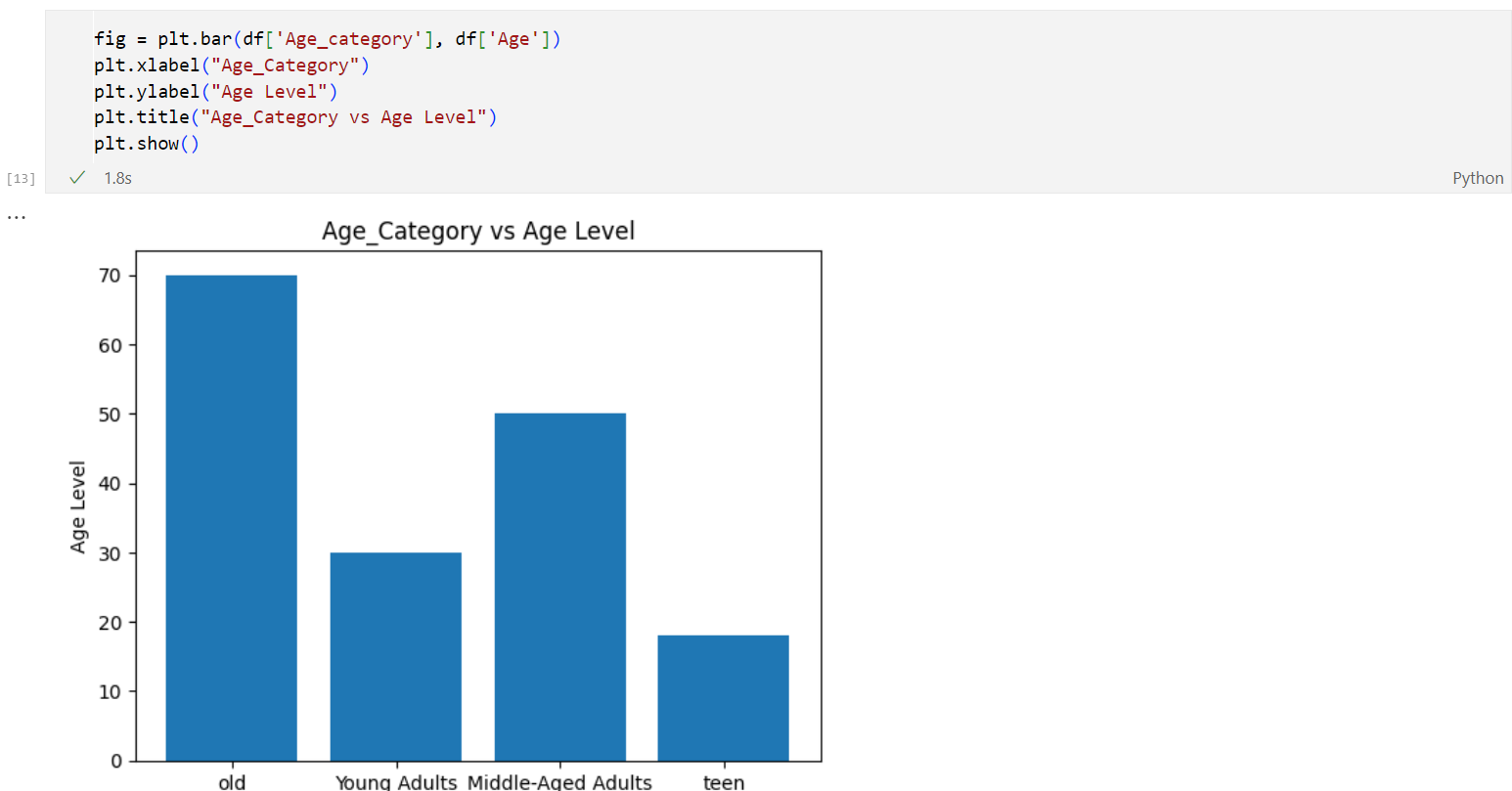
* **Google Chrome, Mozilla Firefox, or any modern web browser for viewing dashboards or online resources.**

**CHAPTER 4**

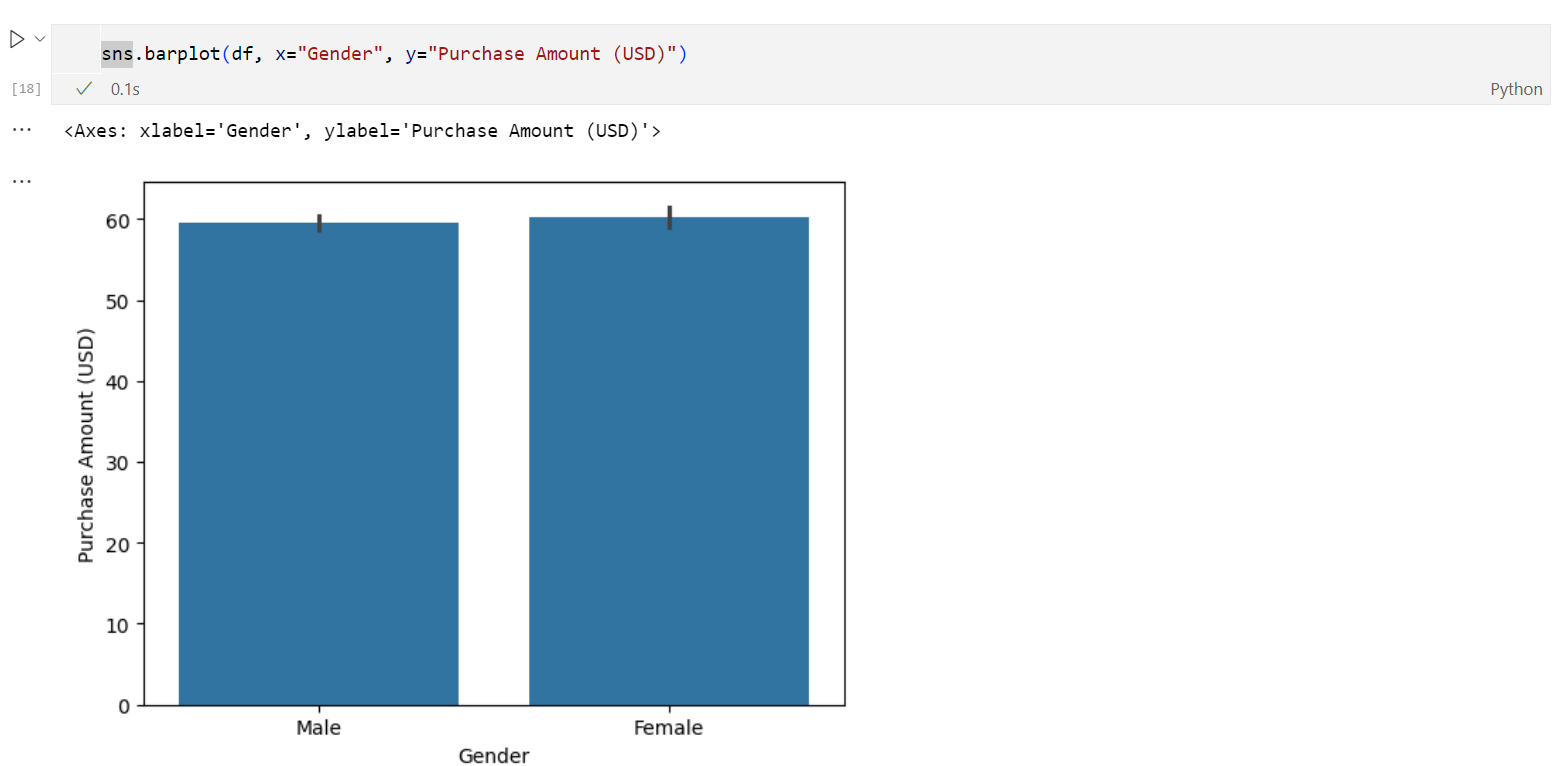
**Implementation and Result**

* 1. **Snap Shots of Result:**

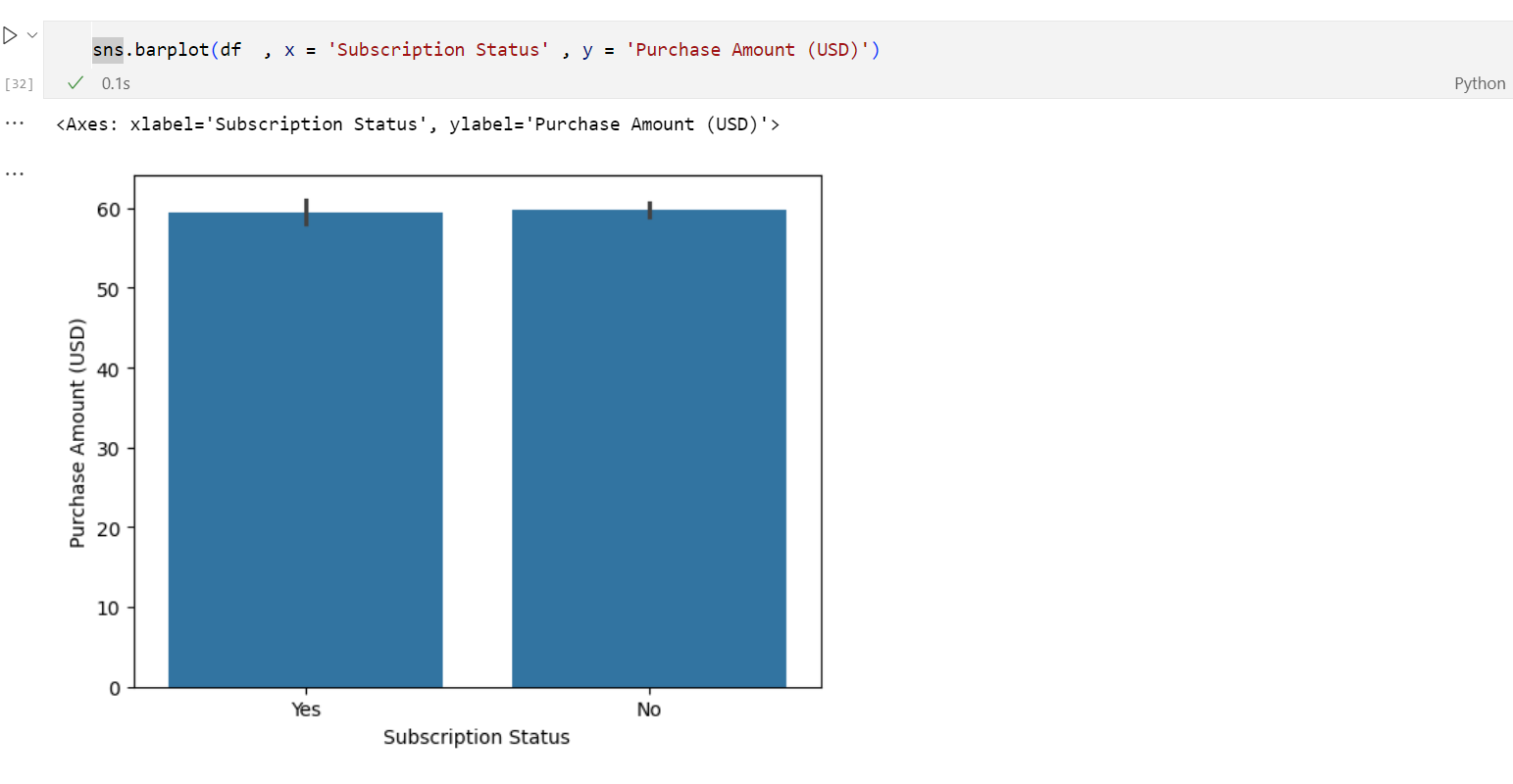
**Below Image Presenting Age-Category Vs Age Level**

****

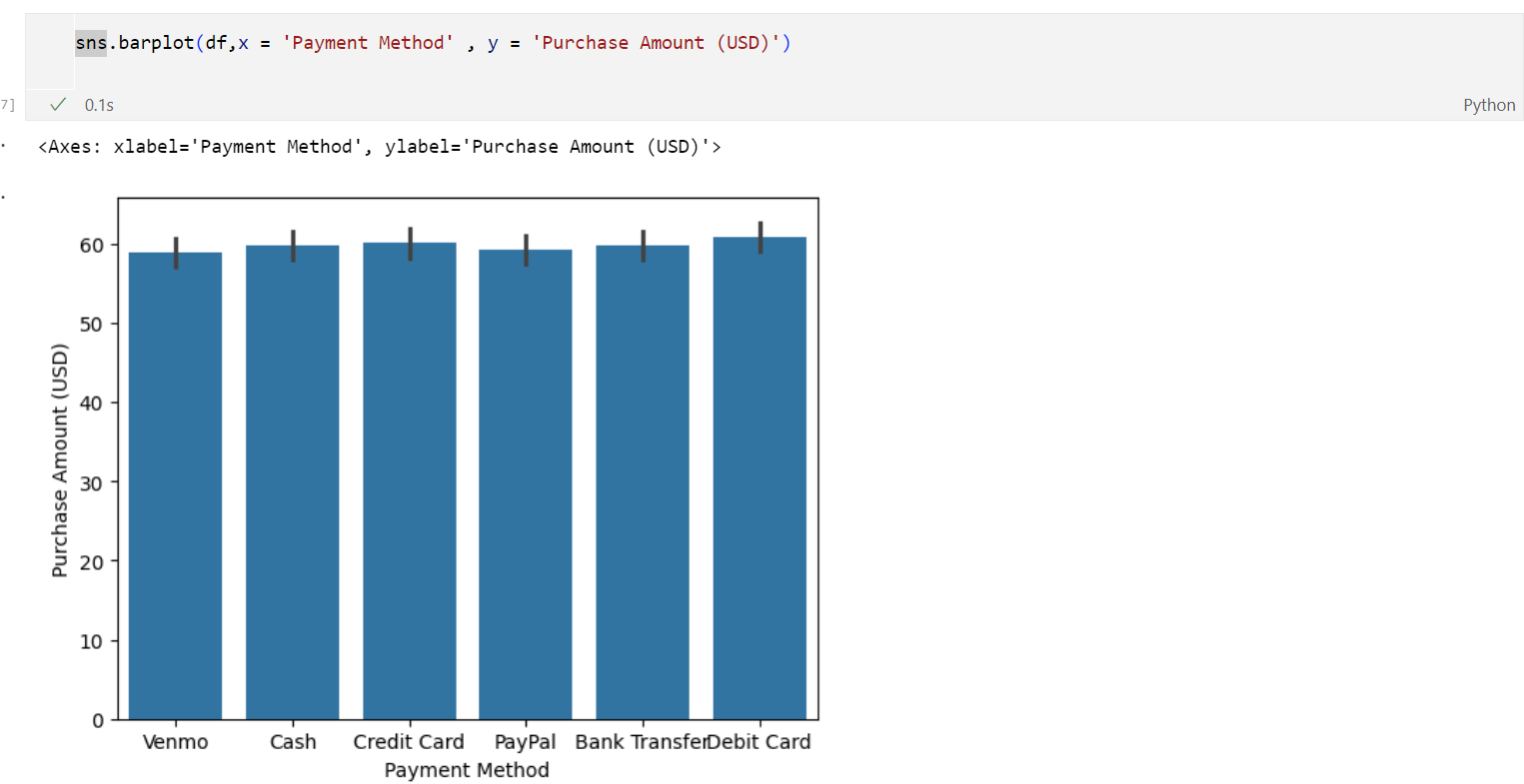
**This bar chat represent Age Vs Gender**

****

**This Shows The Subscription Status Vs Purchase Amount(USD)**

****

**This Shows The Payment Method Vs Purchase Amount(USD)**

****

* 1. **GitHub Link for Code:**

[**Project link**](https://github.com/Prashanth-HP/Internship-Project)

**CHAPTER 5**

**Discussion and Conclusion**

* 1. **Future Work:**

**Enhanced Data Integration**

* Incorporate additional data sources (e.g., APIs, social media, real-time data).
* Automate data pipelines for dynamic and scalable workflows.

**Improved Data Processing**

* Use advanced outlier detection techniques.
* Perform feature engineering to derive new, insightful features.

**Advanced Analytical Methods**

* Implement machine learning models for customer segmentation and recommendations.
* Explore time-series analysis for predicting shopping trends and seasonality.

**Interactive Visualizations**

* Build dynamic dashboards with tools like Tableau, Power BI, or Plotly for better insights presentation.

**Performance Optimization**

* Utilize big data tools like Apache Spark for handling large datasets.
* Deploy the system on cloud platforms for better scalability and processing power.

**Addressing Unresolved Issues**

* Apply robust methods for missing data imputation.
* Assess and mitigate biases in data and analysis.

**Expanding Scope**

Adapt the analysis for cross-domain applications in retail, healthcare, or other sectors.

* 1. **Conclusion:**

**Insight into Shopping Trends**

Provided a comprehensive analysis of customer behavior, purchasing patterns, and preferences, enabling businesses to better understand their target audience.

**Data-Driven Decision Making**

Equipped stakeholders with actionable insights, allowing them to optimize marketing strategies, product offerings, and customer engagement.

**Enhanced Customer Experience**

Identified key factors influencing purchases, paving the way for personalized recommendations and improved customer satisfaction.

**Scalability for Future Applications**

Established a reusable framework for analyzing shopping trends that can be expanded to other domains like e-commerce or retail banking

**Educational Contribution**

Demonstrated the application of data science techniques in real-world scenarios, serving as a learning resource for data analysis and visualization.

**Foundation for Further Research**

Highlighted unresolved challenges and opportunities for advanced analysis, such as predictive modeling and time-series forecasting.

**REFERENCES**

1. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.