TechSaksham edunet foundation

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

Shashank C Kotagi , shashankckotagi@gmail.com

Under the Guidance of

P.Raja , Jay Rathod

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who has supported and guided me throughout this project.

First and foremost, I extend my heartfelt thanks to my supervisor, P. Raja sir, for his exceptional mentorship and guidance. His invaluable advice, constant encouragement, and constructive feedback have been a wellspring of inspiration and innovation, playing a crucial role in the successful completion of this project. His insights and teachings have not only enhanced my technical skills but have also shaped me into a more responsible and thoughtful professional.

I am deeply grateful to the TechSaksham initiative by Microsoft and SAP for providing me with this golden opportunity. Their support has been instrumental in making this project possible.

Lastly, I would like to express my appreciation to my family, peers, and friends for their unwavering belief in me and for providing all the essential support I needed throughout this journey. Their encouragement has been a constant source of motivation.

Shashank C Kotagi

ABSTRACT

This report presents a comprehensive analysis of shopping trends through data analysis, titled "Identifying Shopping Trends using Data Analysis." Retail businesses are inundated with vast amounts of data from various sources, including in-store systems and online platforms. This abundance of information often complicates the process of identifying shopping trends, understanding customer preferences, and forecasting seasonal demands. Failure to recognize these trends can lead to significant challenges such as revenue loss, inventory mismanagement, ineffective marketing strategies, and diminished competitiveness.

This project presents a practical, data-driven approach to tackle these challenges. It follows five key steps:

1. **Data Acquisition:** Gathering relevant data from multiple sources, including point-of-sale systems, e-commerce platforms, and customer feedback channels, to create a comprehensive dataset for analysis.

2. **Data Preparation and Transformation:** Cleaning and transforming the collected data to ensure accuracy and consistency. This step involves handling missing values, standardizing formats, and creating new features that enhance the dataset's usability.

3. **Descriptive Analytics:** Conducting descriptive analysis to summarize historical data and identify key trends and patterns. This includes calculating metrics such as sales volume, customer demographics, and seasonal variations through statistical techniques.

4. **Visual Analytics:** Developing interactive visualizations and dashboards that present the analyzed data in an easily interpretable format. This helps stakeholders quickly grasp insights and make informed decisions based on visual cues.

5. **Strategic Insights and Recommendations:** Formulating actionable business strategies based on the analytical findings. This may involve recommendations for inventory optimization, targeted marketing campaigns, pricing adjustments, and enhancing customer engagement strategies.

Utilizing Python along with its powerful libraries—NumPy, Pandas, Seaborn, and Matplotlib—this analysis was conducted within the Jupyter Notebook environment. The insights derived not only provide actionable recommendations but also unveil promising opportunities for future exploration. These include leveraging machine learning for predictive analytics, enabling real-time decision-making through live data streams, and implementing customer segmentation for personalized marketing initiatives.

This project underscores the potential of Python in transforming raw shopping data into

valuable insights that empower businesses to make informed decisions and maintain a competitive edge in the retail landscape. With advancements such as predictive analytics and real-time data processing on the horizon, this approach holds significant promise for enhancing operational efficiency and fostering long-term growth.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

Retail businesses today are inundated with data from various sources, including in-store systems and online platforms. However, many struggle to effectively analyze this wealth of information to extract valuable insights. This challenge leads to a critical gap in understanding customer behavior, identifying market trends, and accurately forecasting demand. As a result, retailers often face significant obstacles in key areas of their operations. They may miss out on revenue opportunities, grapple with inventory issues such as overstocking or stockouts, implement ineffective marketing strategies, and ultimately find it difficult to maintain a competitive edge in the market. The inability to harness the power of data affects crucial decision-making processes across the board, from inventory management to marketing campaigns and overall operational efficiency. Addressing this data analysis challenge is therefore paramount for retail businesses aiming to enhance their performance, make more informed decisions, and secure a strong position in today's competitive retail landscape.

1.2 Motivation:

This project aims to address a critical challenge in the retail industry: bridging the gap between the vast amount of available data and its effective utilization for strategic decision-making. By leveraging data analytics, the project seeks to unlock the immense potential of shopping trend analysis to optimize inventory levels, enhance customer satisfaction, and improve marketing efforts. The primary motivation is to demonstrate how data-driven insights can transform retail operations, enabling businesses to make informed decisions, reduce costs, and increase profitability. Through this initiative, we aim to showcase the power of analytics in helping organizations adapt to the evolving market landscape, stay competitive, and drive sustainable growth. By providing a practical framework for turning raw data into actionable strategies, this project aspires to equip retailers with the tools and knowledge needed to thrive in today's data-rich environment, ultimately reshaping how businesses operate and succeed in the modern retail sector.

1.3 Objective:

- The primary objective of this project is to analyze shopping data to identify patterns, trends, and actionable insights that can drive better business decisions. The key goals include:

- Developing a robust process for collecting, cleaning, and integrating retail data from multiple sources

- Using statistical analysis and visualizations to uncover shopping trends and customer preferences

- Providing actionable recommendations to improve inventory management, marketing strategies, and operational efficiency

- Creating automated reports and dashboards for real-time insights and decision-making

- These goals aim to address the challenges faced by retail businesses in effectively leveraging their data to gain competitive advantages and optimize operations

.

1.4 Scope of the Project:

This project focuses on analyzing shopping data using Python's data analysis and visualization libraries. The scope includes:

- Data collection and preprocessing to ensure accuracy and consistency.
- Exploratory Data Analysis (EDA) to identify trends and insights
- Building automated visualizations and reports to present findings.
- Providing actionable recommendations based on the analysis.

The limitations include reliance on historical data for analysis and the exclusion of predictive analytics or real-time data integration in the initial phase. While the project lays a strong foundation for data-driven decision-making, further enhancements could involve advanced machine learning models and live data streams for real-time trend analysis.

CHAPTER 2

Literature Survey

2.1 Review of Relevant Literature

The retail industry has increasingly embraced data analytics to drive decision-making and gain deeper insights into consumer behavior. Numerous studies have demonstrated the effectiveness of statistical methods and machine learning techniques in analyzing shopping data and predicting trends. Research has highlighted the importance of data preprocessing techniques, such as handling missing values and removing duplicates, to ensure data reliability. Exploratory Data Analysis (EDA) has been emphasized for its role in identifying key patterns and relationships in datasets, while visualizations have been widely adopted to present findings in a user-friendly manner. The integration of business intelligence tools like Tableau and Power BI has also been extensively explored, enabling businesses to create interactive dashboards that provide real-time, actionable insights.

Despite these advancements, there are notable gaps in existing solutions. Many methods involve complex machine learning algorithms that require specialized knowledge and substantial computational power, making them inaccessible to smaller businesses. There is often a focus on predictive analytics at the expense of actionable insights derived from historical data. Additionally, while some tools emphasize real-time data integration, they may not combine this with detailed historical analysis for comprehensive trend identification. Cost constraints associated with advanced tools and platforms can also be prohibitive for some businesses. To address these gaps, this project adopts a cost-effective and straightforward methodology using Python's open-source libraries to analyze historical shopping data and generate actionable insights, balancing simplicity with effectiveness

2.2 Existing Models, Techniques, and Methodologies

Several existing models and methodologies have been developed to address similar challenges in retail data analysis. These include:

ETL Pipelines: Widely used for data collection and integration, ETL pipelines extract data from various sources, transform it into a standardized format, and load it into a centralized database for analysis.

Machine Learning Models: Algorithms such as linear regression, clustering, and classification are commonly employed to predict customer behavior and segment audiences for targeted marketing.

Statistical Methods: Techniques like correlation analysis and hypothesis testing are applied during Exploratory Data Analysis (EDA) to identify patterns and relationships in shopping data.

Visualization Libraries: Tools like Matplotlib, Seaborn, and Tableau are frequently utilized to create intuitive visualizations that facilitate data interpretation and decision-making.

While these approaches offer significant advantages, many require advanced expertise, substantial computational resources, or focus heavily on predictive modeling rather than deriving actionable insights from historical data.

2.3 Gaps and Limitations in Existing Solutions

Despite the progress in the field, there are notable gaps and limitations in existing solutions:

- **High Complexity:** Many methods involve sophisticated machine learning algorithms that require specialized knowledge and substantial computational power, making them inaccessible to smaller businesses.
- **Focus on Prediction:** Existing solutions often emphasize predictive analytics, while ignoring the importance of actionable insights derived from historical data.
- **Limited Real-Time Capabilities:** Although some tools focus on real-time data integration, they are not always combined with detailed historical analysis for comprehensive trend identification.
- **Cost Constraints:** Many advanced tools and platforms, such as Tableau and Power BI, may not be affordable for all businesses.

How This Project Addresses These Gaps

This project adopts a cost-effective and straightforward methodology to analyze historical shopping data and generate actionable insights. By leveraging Python's open-source libraries—such as NumPy for numerical computations, Pandas for data manipulation, and Matplotlib and Seaborn for data visualization—it provides an accessible solution that balances simplicity with effectiveness.

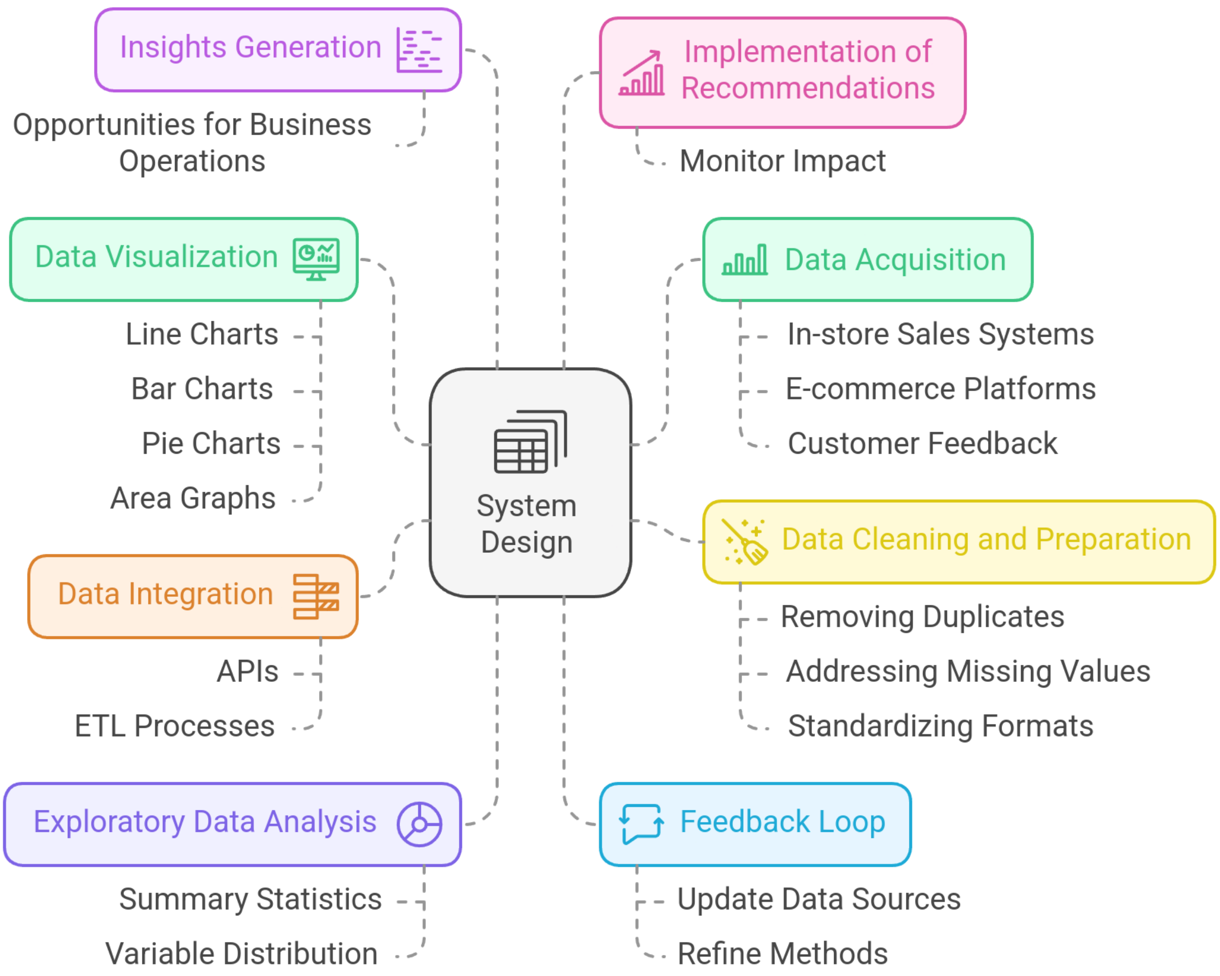
The focus on Exploratory Data Analysis (EDA) ensures a thorough examination of past trends, helping businesses understand customer preferences and behavior through statistical summaries and visualizations. Automated dashboards and visualizations eliminate the need for expensive tools like Tableau while offering real-time insights for decision-making. This approach bridges the gap between predictive analytics and retrospective evaluation, ensuring businesses of all sizes can benefit from data-driven strategies to enhance operations, optimize inventory, and improve marketing efforts.

CHAPTER 3

Proposed Methodology

3.1 System Design

System Design for Identifying Shopping Trends



3.2 Requirement Specification

3.2.1 Hardware Requirements:

Processor: Intel Core i7 or higher (or equivalent AMD Ryzen processor)
Memory: Minimum 16 GB RAM
Storage: 512 GB SSD (Solid State Drive) or higher
Operating System: Windows 10, macOS, or Linux

3.2.2 Software Requirements:

Programming Language: Python 3.x for data analysis and processing.

Libraries:

- NumPy: For handling numerical data operations and array manipulation.
- Pandas: To manage, preprocess, and analyze datasets efficiently.
- Seaborn: For creating informative statistical data visualizations that enhance understanding of data patterns.
- Matplotlib: For generating additional plots and custom visualizations, providing flexibility in data representation.

Integrated Development Environment (IDE):

- Jupyter Notebook: For interactive coding and testing of Python scripts, allowing for real-time feedback during development.

Visualization Tools:

- Tableau: A leading tool for creating advanced dashboards that provide intuitive visual insights into retail data.
- Power BI: For interactive reporting and business analytics, enabling users to visualize data from various sources seamlessly.

Additional Tools:

- MicroStrategy Analytics: For comprehensive business intelligence capabilities, assisting in data analysis and digital security.
- Cloud Storage Solutions: Such as AWS or Google Cloud for scalable storage and processing power, facilitating collaboration and remote access to datasets

CHAPTER 4

Implementation and Result

4.1 : Snap Shots of Result:

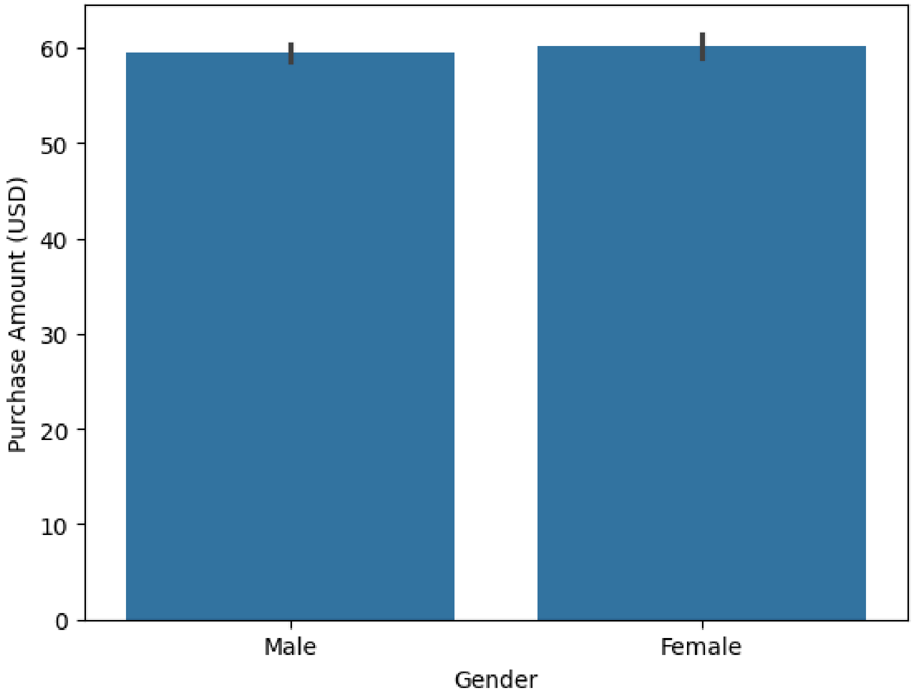
Analysis of Purchases by Gender

Question

This analysis aims to determine which gender has the highest number of purchases based on the total purchase amount in USD. Understanding gender-based purchasing behavior can help businesses tailor their marketing strategies effectively.

Code Explanation

- sns.barplot: This function from the Seaborn library creates a bar plot for visualizing data.
- shop: This is the DataFrame containing the shopping data.
- x = 'Gender': Sets the x-axis to represent different genders.
- y = 'Purchase Amount (USD)': Sets the y-axis to represent the total purchase amount in USD.
- Result: The bar plot visually compares the total purchases made by each gender, allowing for easy identification of which gender has the highest spending.



Analysis of Purchase Frequency Across Age Groups

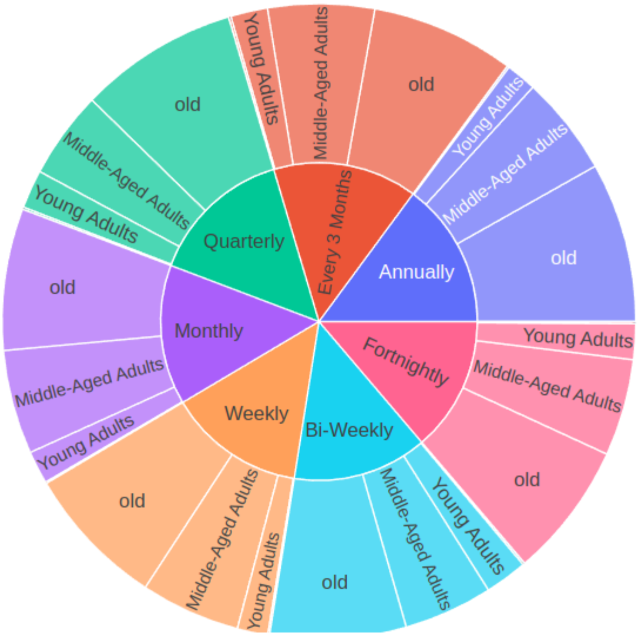
This analysis aims to explore how the frequency of purchases varies across different age groups. Understanding these patterns can help retailers tailor their marketing strategies and product offerings to specific demographics

Code

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

Code Explanation

- import plotly.express as px: Imports the Plotly Express library for creating interactive visualizations.
- px.sunburst: This function generates a sunburst chart, which visualizes hierarchical data.
- shop: Refers to the DataFrame containing the shopping data.
- path=['Frequency of Purchases', 'Age_category']: Specifies the hierarchical structure for the sunburst chart, with "Frequency of Purchases" as the root and "Age_category" as the child segments.
- values='Age': Indicates that the size of each segment in the sunburst chart will be based on the frequency of purchases corresponding to each age category.



Analysis of the Impact of Discounts on Purchase Decisions

This analysis investigates how the presence of a discount influences customer purchase decisions. Understanding this relationship is crucial for retailers to optimize their promotional strategies and maximize sales.

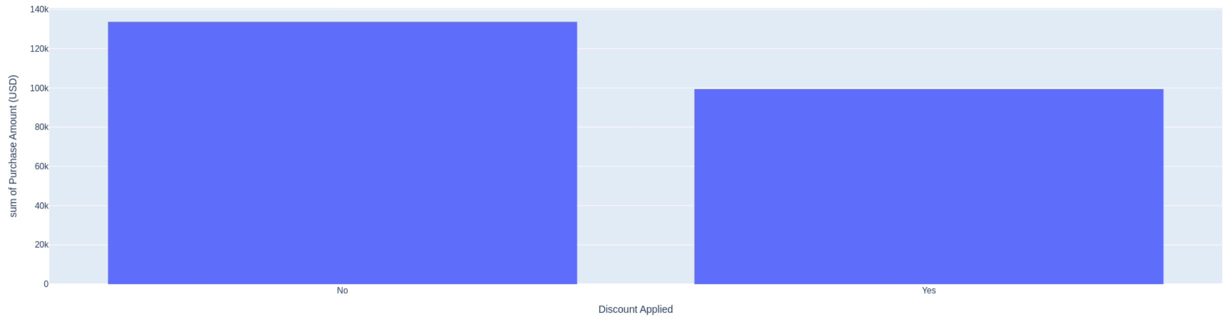
Code

```
shop_group = shop.groupby('Discount Applied')['Purchase Amount (USD)'].sum().reset_index()

px.histogram(shop_group, x='Discount Applied', y='Purchase Amount (USD)')
```

Code Explanation

- import plotly.express as px: Imports the Plotly Express library for creating interactive visualizations.
- shop.groupby('Discount Applied')['Purchase Amount (USD)'].sum().reset_index(): Groups the dataset by whether a discount was applied, summing the total purchase amounts for each group, and resets the index for easier plotting.
- px.histogram: This function generates a histogram to visualize the distribution of purchase amounts based on discount application.
- x='Discount Applied': Sets the x-axis to represent whether discounts were applied.
- y='Purchase Amount (USD)': Sets the y-axis to represent the total purchase amount in USD, allowing for comparison of spending with and without discounts.



Analysis of the Impact of Promo Codes on Customer Spending

This analysis seeks to determine whether customers who use promo codes tend to spend more than those who do not. Understanding this relationship can provide valuable insights for businesses looking to optimize their promotional strategies and enhance customer engagement.

Code

```
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()
```

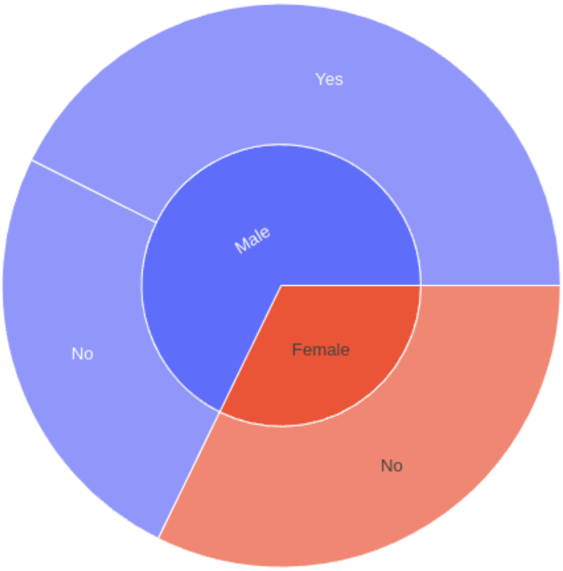
Code Explanation

- shop.groupby('Promo Code Used')['Purchase Amount (USD)'].sum().reset_index(): Groups the dataset by whether a promo code was used, summing the total purchase amounts for each group, and resets the index for easier plotting.

- px.sunburst: This function generates a sunburst chart to visualize hierarchical data.

- path=['Gender', 'Promo Code Used']: Specifies the hierarchical structure for the sunburst chart, with "Gender" as the root and "Promo Code Used" as the child segments.

- values='Purchase Amount (USD)': Indicates that the size of each segment in the sunburst chart will be based on the total purchase amount in USD.



Analysis of Purchase Behavior Across Different Locations

This analysis aims to identify any noticeable differences in purchase behavior among various locations. Understanding these differences can help retailers tailor their marketing strategies and optimize their product offerings based on regional preferences.

Code

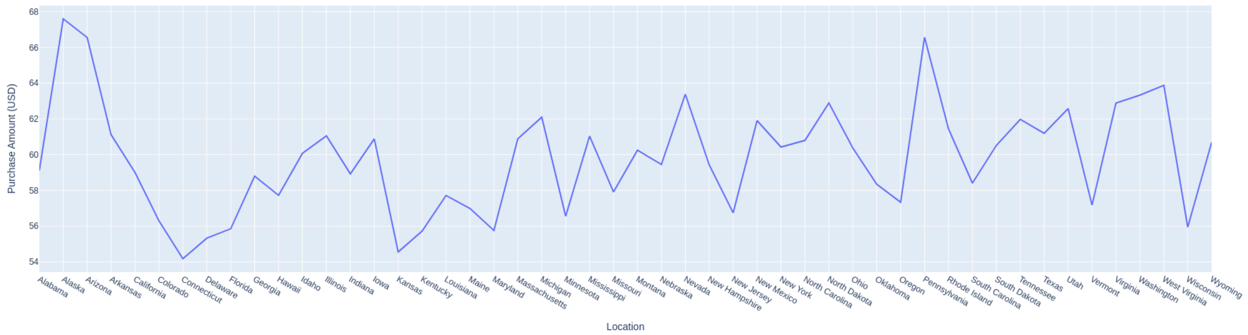
```
shop_group = shop.groupby('Location')['Purchase Amount (USD)'].mean().reset_index()

fig = px.line(shop_group, x='Location', y='Purchase Amount (USD)')

fig.show()
```

Code Explanation

- import plotly.express as px: Imports the Plotly Express library for creating interactive visualizations.
- shop.groupby('Location')['Purchase Amount (USD)'].mean().reset_index(): Groups the dataset by location, calculating the average purchase amount for each location, and resets the index for easier plotting.
- px.line: This function generates a line chart to visualize trends over a continuous variable.
- x='Location': Sets the x-axis to represent different locations.
- y='Purchase Amount (USD)': Sets the y-axis to represent the average purchase amount in USD, allowing for comparison of spending across locations.



4.2 : GitHub Link for Code:

https://github.com/Shashankckotagi/shopping-data-analysis

CHAPTER 5

Discussion and Conclusion

5.1 Future Work

To expand upon the outcomes of this project, the following areas are recommended for further exploration:

- **Advanced Predictive Modeling**: Utilize machine learning techniques, such as ensemble methods, to improve the accuracy of trend predictions.
- **Real-Time Analytics**: Develop systems capable of processing data in real time, enabling immediate responses to changing consumer behaviors.
- **Customer Segmentation**: Apply clustering algorithms to identify distinct customer groups, facilitating personalized marketing strategies.
- **A/B Testing**: Conduct controlled experiments to evaluate the effectiveness of various marketing and inventory management approaches.
- **Improved Dashboards**: Refine user interfaces to make insights more accessible, with a focus on mobile-friendly designs for broader usability.

5.2 Conclusion

This project demonstrates the effective use of data analytics to derive actionable insights from historical shopping data.

- By utilizing Python's open-source libraries (NumPy, Pandas, Matplotlib, and Seaborn), it provides a cost-effective and accessible solution for retail businesses.
- The focus on Exploratory Data Analysis (EDA) enables the identification of shopping trends and customer behavior, supporting informed decision-making in inventory management, marketing strategies, and operational efficiency.
- Automated visualizations and dashboards bridge the gap between predictive analytics and retrospective evaluation, eliminating the need for expensive tools.
- The project lays a foundation for future enhancements, such as advanced predictive modeling, real-time analytics, and customer segmentation.

Overall, this approach empowers businesses of all sizes to adopt data-driven strategies, adapt to market changes, and achieve sustainable growth.

REFERENCES

- **Han, J., Kamber, M., & Pei, J.** (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

  This book provides an in-depth explanation of data preprocessing, data integration, and exploratory data analysis techniques.

- **Wickham, H.** (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

  A reference for visualization techniques and tools for effective data presentation.

- **McKinney, W.** (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.

  This paper discusses the use of Pandas for data manipulation and its role in data preprocessing and analysis.

- **Seaborn Documentation.** (n.d.). Available at: https://seaborn.pydata.org

  A guide to statistical data visualization and its applications in analyzing large datasets.

- **Power BI Documentation.** (n.d.). Microsoft Power BI. Available at: https://powerbi.microsoft.com

  Provides details about building dashboards for real-time analytics and visualization.

- **Matplotlib Documentation.** (n.d.). Available at: https://matplotlib.org

  Discusses plotting capabilities for creating clear and actionable visualizations.

- **Chen, H., Chiang, R. H., & Storey, V. C.** (2012). *Business Intelligence and Analytics: From Big Data to Big Impact*. MIS Quarterly, 36(4), 1165-1188.

  This paper explores the use of business intelligence tools for data-driven decision-making.

- **ETL Process Overview.** (n.d.). Talend. Available at: https://www.talend.com/resources/what-is-etl

  Provides an overview of ETL pipelines for data collection and integration.

- **Das, S. R., & Chen, M. Y.** (2007). *Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web*. Management Science, 53(9), 1375-1388.

  Focuses on the use of statistical and machine learning techniques for identifying trends in data.

- **Kelleher, J. D., Mac Namee, B., & D'Arcy, A.** (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

  Covers machine learning models and methodologies relevant to customer behavior prediction and clustering.