Polytechnic University of the Philippines
COLLEGE OF COMPUTER AND INFORMATION SCIENCES
A. Mabini Campus, Sta. Mesa, Manila

In Partial Fulfillment of the Requirements for
INTE-E2 - IT Elective 2 (Data Mining)

Customer Segmentation and Spending Behavior Analysis using Credit Card Transaction Data: Unlocking Insights for Personalized Marketing Strategies

by

LEONA, LUKE MARK
BSIT 3-1

DR. ALETA FABREGAS
Elective 2 Faculty-In-Charge

July 2023

I. Purpose and Objective:

The main purpose of this project is to improve customer engagement and drive business growth by making data-driven marketing decisions. To achieve this, we will use the credit card transaction data to segment customers and understand their spending behavior across various categories. By gaining insights into customer preferences, we can optimize our offerings and tailor marketing strategies to meet their needs effectively. The ultimate goal is to enhance customer satisfaction, increase engagement, and drive business growth by leveraging the power of data analysis.

II. Historical Data and Source:

For this project, we will be working with a dataset that contains records of credit card transactions. The data is stored in a CSV file and includes various details for each transaction, such as the year, month, spending categories (e.g., entertainment, food, health and fitness, home), transaction amounts, transaction frequencies, customer attributes (gender, job description), and date of birth. This dataset is sourced from a financial company. The data from this program will serve as the foundation for our analysis and insights generation in this project.

	year	month	tag	total	Entertainment	Entertainment_Freq	Food	Food_Freq	…	birthdate
1	2019	1	1	2904.31	218.73	4	90.69	3	…	09/03/1988
2	2019	1	2	5513.39	476.5	7	152.08	2	…	09/03/1988
3	2019	2	1	3696.99	244.19	2	22.93	1	…	09/03/1988
4	2019	2	2	2847.65	85.84	2	317.45	6	…	09/03/1988
5	2019	3	1	4785.08	245.7	3	51.88	2	…	09/03/1988
6	2019	3	2	4127.63	303.04	4	176.41	3	…	09/03/1988
7	2019	4	1	2954.68	97.69	3	192.14	3	…	09/03/1988
8	2019	4	2	2857.7	423.72	5	488.68	7	…	09/03/1988
9	2019	5	1	6761.68	494.72	4	18.73	1	…	09/03/1988
…	…	…	…	…	…	…	…	…	…	…
42677	2020	12	1	1606.97	85.68	2	97.22	3	…	19/10/1989

Table 1: Data Overview

Variable	Data	Type Description
year	int	Year of transaction
month	int	Month of transaction
tag	int	Half of Month (HOM) tag. HOM=1 refers to days 1 to 15 while HOM=2 refers to days 16 to 30/31
total	float	Sum of all transactions in one HOM
{category}	float	Subtotal of transactions per spending type/ category (see list of categories below)
{category}_freq	int	Frequency of transactions per category within the HOM period

dob	datetime	Date of birth
gender	str	Sex assigned at birth: M/F
job	str	Job description of customer

Table 2: Data Dictionary

III. Problem Definition:
This project aims to address the following problem statements:

a. Customer Segmentation:

1. *What distinct customer groups can be identified for targeted marketing strategies by segmenting based on spending behavior, job description, gender, and age (derived from date of birth)?*
2. *How can this segmentation be implemented using the credit card transaction data?*

Customer segmentation is crucial for targeted marketing strategies. By identifying distinct customer groups based on spending behavior, job description, gender, and age, we can tailor our marketing efforts to meet their specific needs. By analyzing credit card transaction data, we can uncover patterns and characteristics that enable us to segment customers effectively, leading to personalized marketing approaches.

b. Spending Behavior Analysis:

1. *How can spending patterns across different categories (e.g., entertainment, food, health and fitness, home) be analyzed to gain insights into customer preferences and optimize offerings?*

Understanding customer spending patterns across different categories provides valuable insights into their preferences and priorities. By analyzing the credit card transaction data, we can gain a deeper understanding of customer behavior within each spending category (e.g., entertainment, food, health and fitness, home). This analysis helps optimize offerings, align product/service development, and refine marketing strategies to better cater to customer needs.

c. Predictive Modeling:

1. *Is it possible to develop models using historical data to predict the total spending of customers within a given period for financial planning and targeted marketing campaigns?*
2. *Can predictive modeling using historical transaction data and customer attributes forecast future spending or predict customer churn?*

Predictive modeling using historical transaction data and customer attributes enables us to forecast future customer spending and predict potential churn. By leveraging this data, we can develop models that provide insights into customer behavior and preferences, allowing for more accurate financial planning and targeted marketing campaigns. This helps optimize resource allocation, improve customer retention strategies, and drive business growth.

IV. Method or Process:

a. Customer Segmentation:

- Method: Clustering analysis can be employed to group customers based on their spending behavior, job description, gender, and age.
- Algorithm: K-means clustering or hierarchical clustering algorithms can be utilized for customer segmentation.

b. Spending Behavior Analysis:

- Method: Exploratory data analysis (EDA) techniques can be used to analyze the distribution, trends, and patterns in customer spending across different categories.
- Algorithm: No specific algorithm is required for EDA, but statistical measures and visualizations can be applied to gain insights.

c. Optimization and Personalization:

- Method: Association rule mining or collaborative filtering techniques can be employed to identify relationships between customer attributes, spending patterns, and personalized recommendations.
- Algorithm: Apriori algorithm for association rule mining or collaborative filtering algorithms like user-based or item-based collaborative filtering can be used.

d. Predictive Modeling:

- Method: Supervised machine learning techniques can be applied to build predictive models using historical transaction data and customer attributes.
- Algorithm: Various algorithms can be used based on the specific problem, such as linear regression, decision trees, random forests, gradient boosting, or neural networks.

.

V. Output or Deliverables:

- A report on customer segmentation analysis, identifying distinct customer groups based on spending behavior, job description, gender, and age.
- A report on spending behavior analysis, providing insights into customer preferences and trends within each spending category.
- Strategies for optimizing offerings and personalizing marketing approaches, leveraging the segmentation and spending behavior analysis.
- Predictive models for forecasting future spending or predicting customer churn, along with implementation guidelines.

FINDINGS and CONCLUSIONS

A. Exploratory Data Analysis

Spending Behavior Analysis by Category

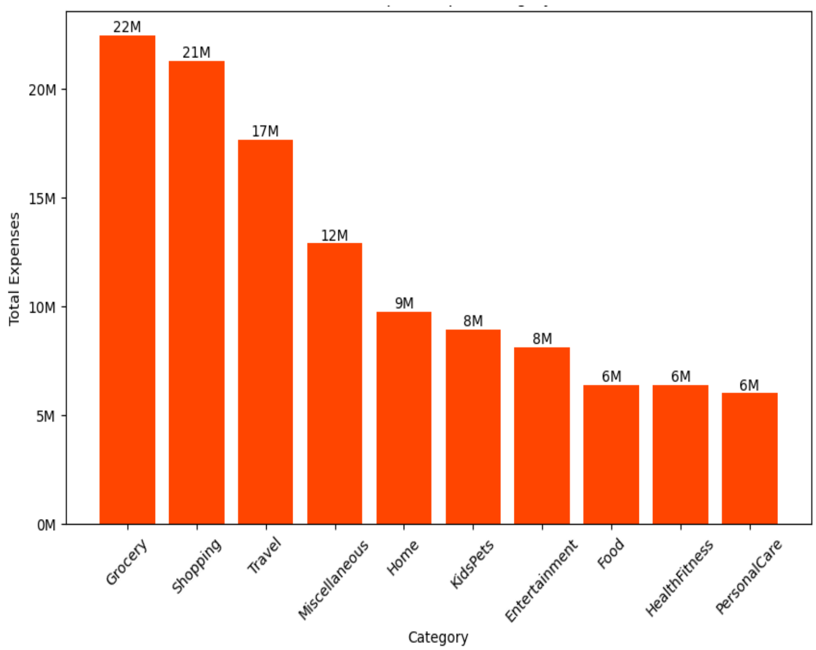


Figure 1: Total Expenses per Category

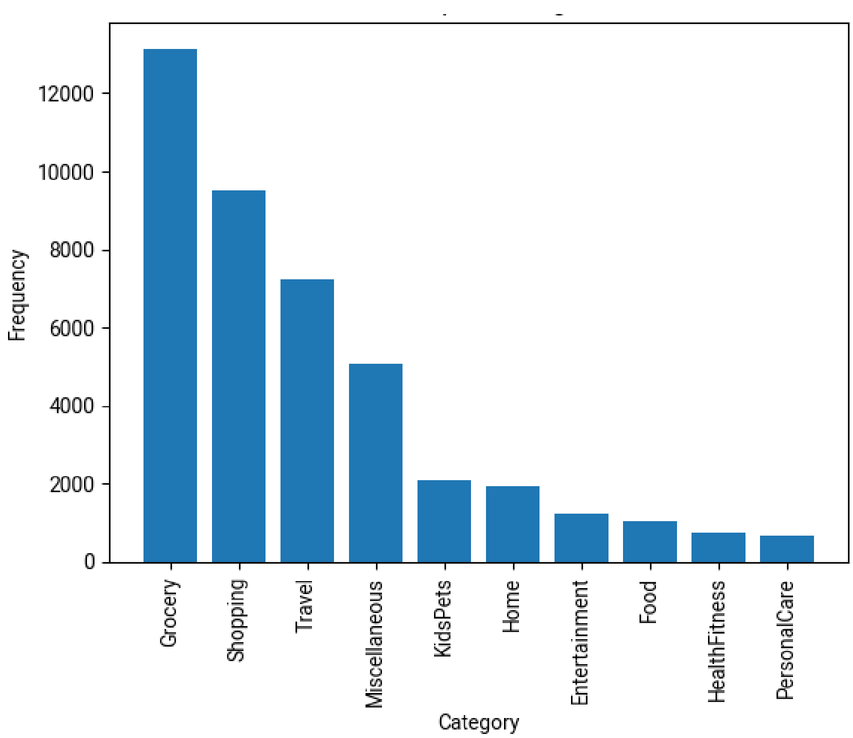


Figure 2: Most Frequent Category

Figure 1 and 2 displays the total expenses per category, revealing the spending patterns of credit card users. The graph highlights that credit card users tend to have higher expenses in certain categories. The category with the highest expenses among credit card users is grocery, indicating that a significant portion of their spending goes towards purchasing groceries.

Following grocery, the second-highest category of expenses is shopping. This suggests that credit card users frequently make purchases in retail stores or online shopping platforms.

The third-highest category of expenses is travel, indicating that credit card users also spend a considerable amount on travel-related expenses, such as flights, accommodations, and transportation.

On the other hand, the categories of HealthFitness and Personal Care have the least expenses among credit card users. This implies that they tend to spend relatively less on health and fitness-related products or services, as well as personal care items.

In summary, Figure 1 and 2 provides valuable insights into the spending behavior of credit card users. It highlights their higher expenses in grocery and shopping, followed by travel, while showing that health, fitness, and personal care are the categories with the least expenses. This information can be useful for understanding consumer preferences and tailoring marketing strategies accordingly.

Spending Behavior Analysis by Gender

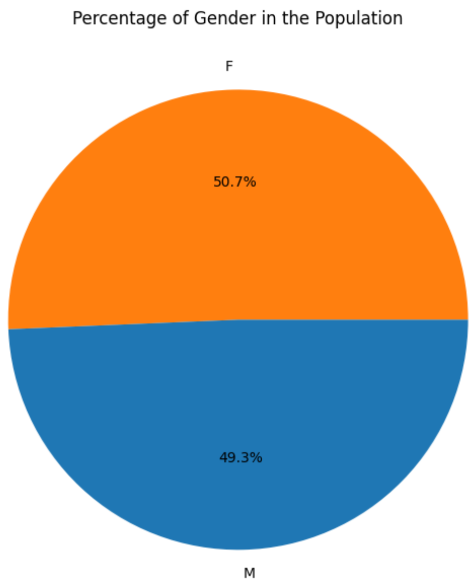


Figure 3

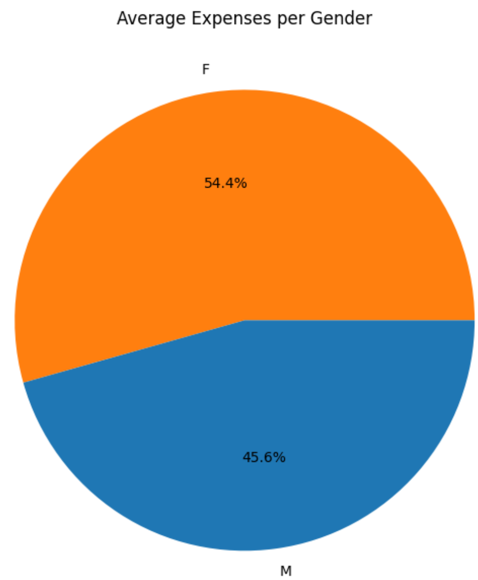


Figure 4

The figure 3 and 4 illustrates the percentage of each gender and their respective total expenses. The data indicates that 50.7% of the individuals are female, while 49.3% are male. In terms of total expenses, females account for a higher proportion, representing 55% of the total, whereas males account for 45% of the total expenses. This clearly demonstrates that females have more expenses compared to males.

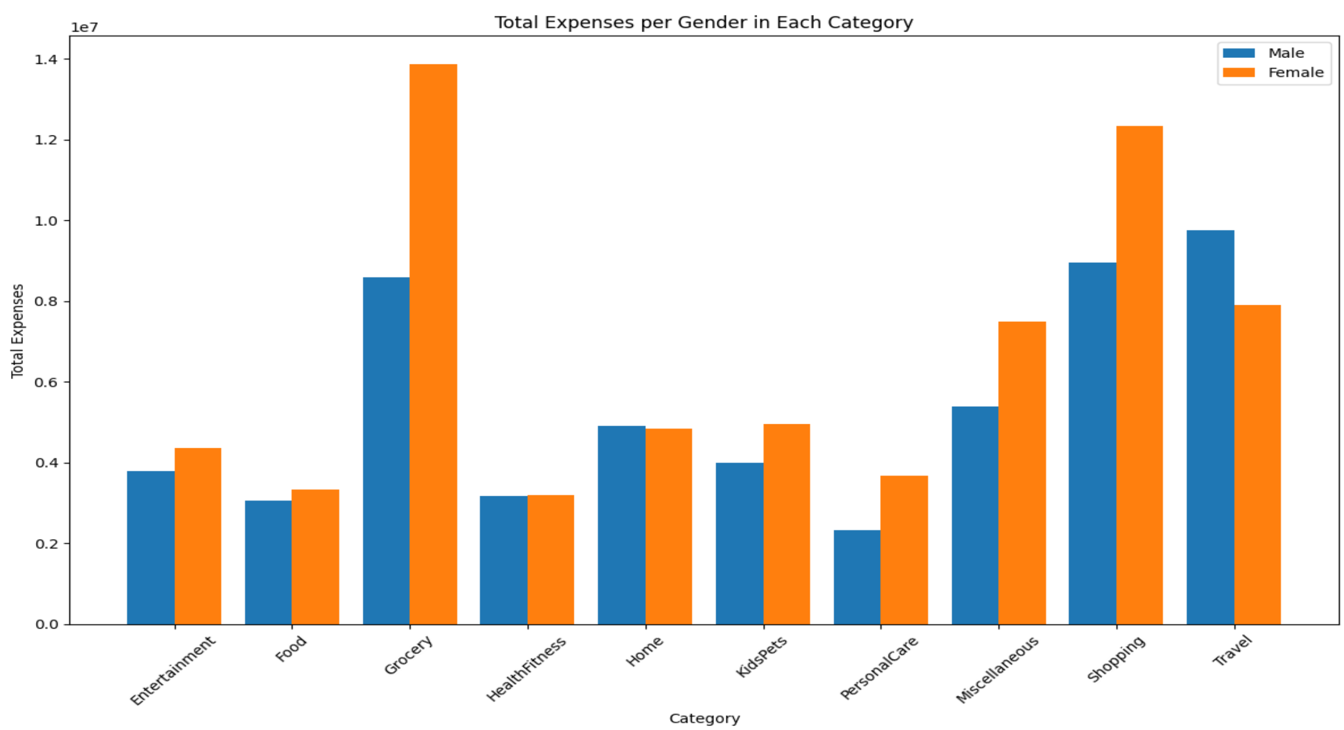


Figure 5: Total Expenses per Gender in Each Category

Figure 5 displays the total expenses for each gender in various spending categories. The graph reveals that females have accumulated higher expenses in the majority of the spending categories. Specifically, the top category where females have the highest expenses is grocery, followed closely by shopping.

However, there are only two categories where males have the highest expenses, which are travel and home expenses. In all other spending categories, females have outspent males, indicating that females tend to spend more across a wide range of expense categories.

The figure clearly illustrates that females tend to have higher expenses than males in most spending categories, with grocery and shopping being the areas where females spend the most, while males lead in travel and home expenses.

Spending Behavior Analysis by Time

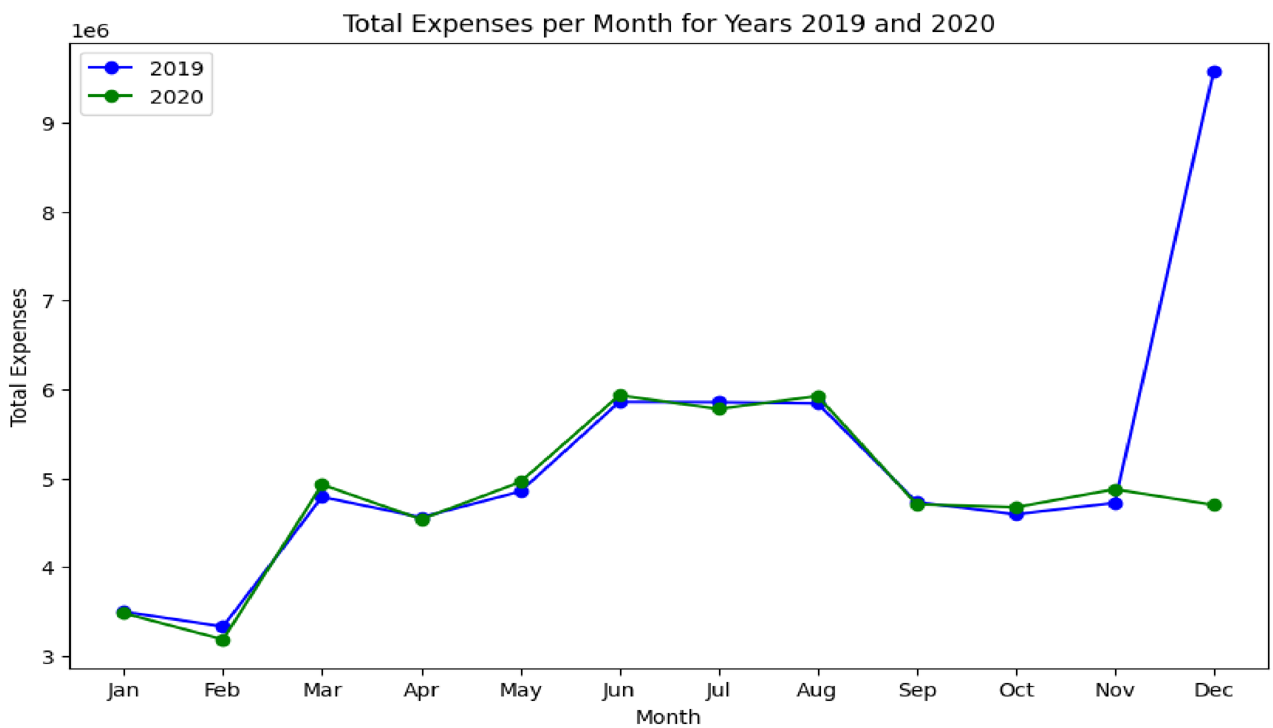


Figure 6: Total Expenses per Month for Years 2019 and 2020

Figure 6 presents a line graph illustrating the expenditure patterns for the years 2019 and 2020. By comparing the expenses per year, we observe that there is generally no substantial difference from January to November in both years. The expenditure trends appear quite similar during these months.

However, during specific months, such as March 2020 through June 2020, there is a noticeable increase in expenses compared to the same months in 2019. These months stand out as having higher expenditure levels in 2020. Moreover, when we analyze the expenses for December, we observe a significant difference between the two years. December 2019 reflects notably higher spending compared to December 2020.

It is crucial to consider external factors that could influence these spending patterns. One significant factor that emerged during 2020 was the global pandemic, Covid-19. This unprecedented event significantly impacted people's lives, resulting in changes to daily routines, travel restrictions, and increased emphasis on staying at home. As a result, there might be shifts in spending behaviors, such as increased online shopping, altered travel plans, and variations in entertainment expenses.

Additionally, the line graph highlights that the months between June to August have experienced a considerable peak in expenses. The significant increase in spending during this period suggests that there might be specific factors contributing to these higher expenditures.

During the summer months, individuals tend to engage in various activities such as vacations, travel, outdoor events, and recreational activities. These factors can lead to increased spending on travel expenses, accommodation, dining out, entertainment, and other leisure-related expenditures.

Overall, the line graph not only captures the general trend of expenses for the years 2019 and 2020 but also points out interesting fluctuations in certain months. These fluctuations may be attributed to a combination of factors, including seasonal influences, economic conditions, and significant events like the Covid-19 pandemic. By understanding these patterns, individuals and businesses can make informed decisions and plan their finances more effectively throughout the year.

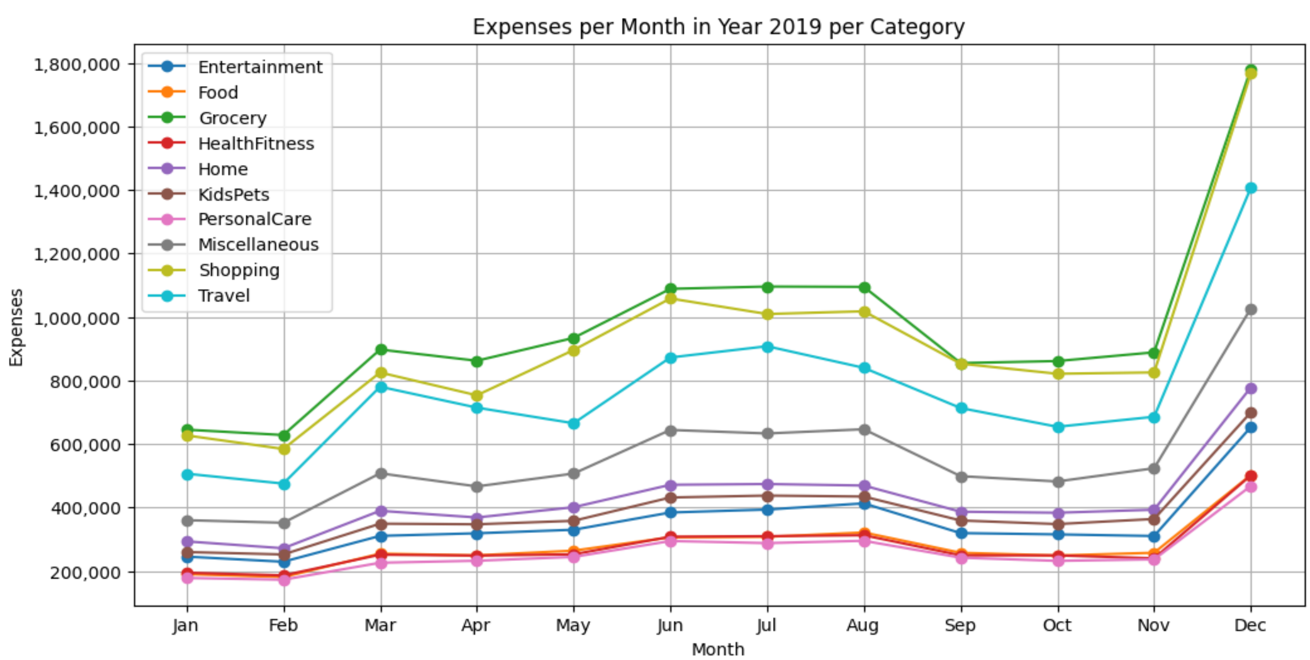


Figure 7: Expenses per Month in Year 2019 per Category

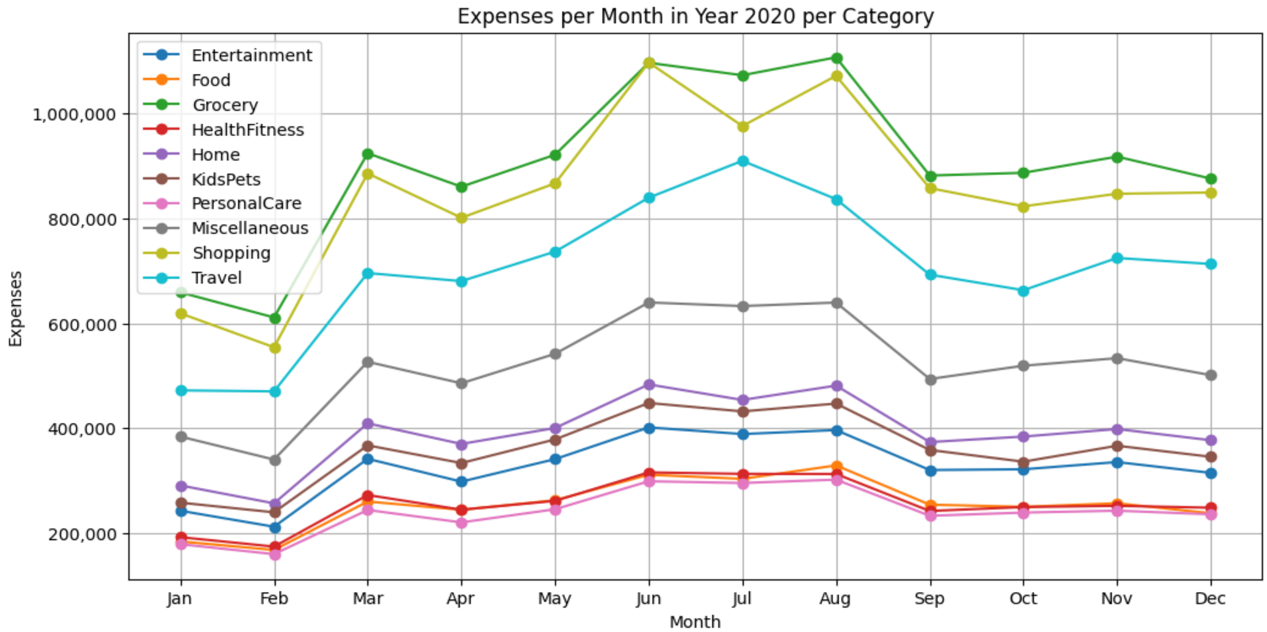


Figure 8: Expenses per Month in Year 2020 per Category

In Figures 7 and 8, we can observe a comparison of the monthly expenses per category for the years 2019 and 2020. The data visualization provides insights into the spending behavior before and during the Covid-19 pandemic. In Figure 7, which represents the expenses for the year 2019, we can see that there is minimal difference between the expenses for the categories HealthFitness and Personal Care. The two categories are relatively close in terms of spending throughout the year. However, in Figure 8, depicting the expenses for the year 2020, there is a noticeable shift. The category HealthFitness shows a slight increase compared to the previous year. This shift suggests that there might be a change in people's spending priorities during the pandemic.

The Covid-19 pandemic significantly impacted daily routines, lifestyles, and consumer behaviors worldwide. With restrictions on social activities, travel limitations, and health concerns, individuals might have adjusted their spending habits. As a result, there could be an increased focus on health-related expenses, such as fitness equipment, online fitness classes, and personal care products during the pandemic.

The visual representation in the figures lends support to the statement that Covid-19 could be a factor influencing people's expenses. The data reveals a subtle yet distinct change in spending patterns, indicating that external events like the pandemic can have an impact on individual financial choices.

Spending Behavior Analysis by Generation

Legend:

Generation	Born on	Age
Silent Generation	1925 - 1945	77 to 97 years old
Baby Boomers	1946 - 1964	57 to 75 years old
Generation X	1965 - 1980	41 to 56 years old
Millennials	1981 - 1996	25 to 40 years old
Generation Z	1997 - 2012	9 to 24 years old

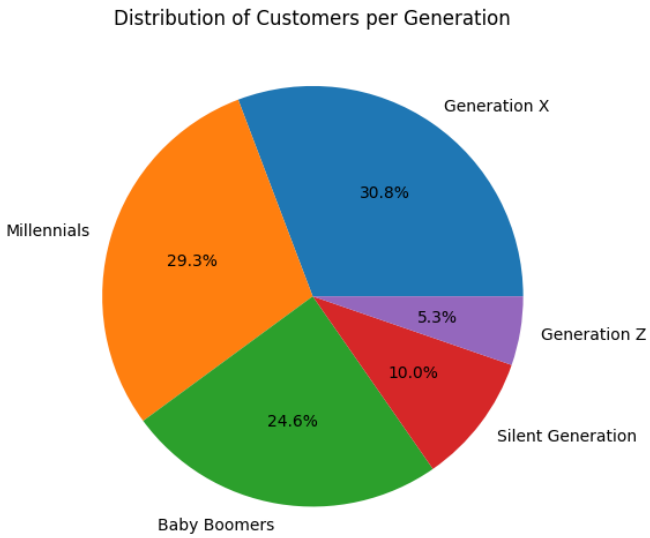


Figure 9

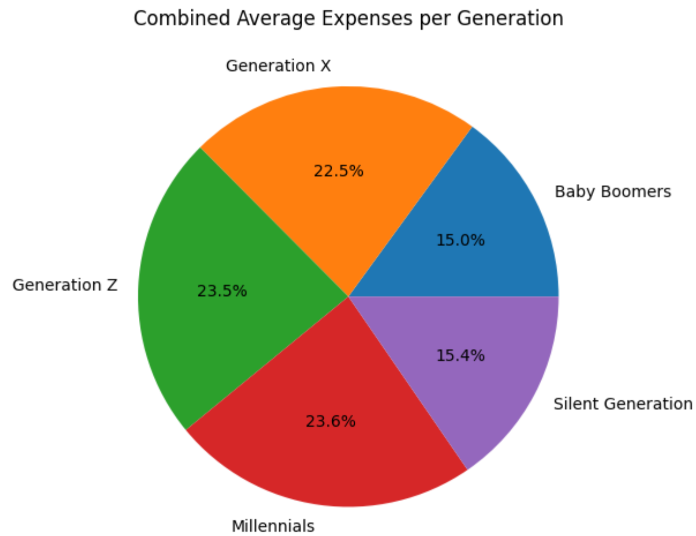


Figure 10

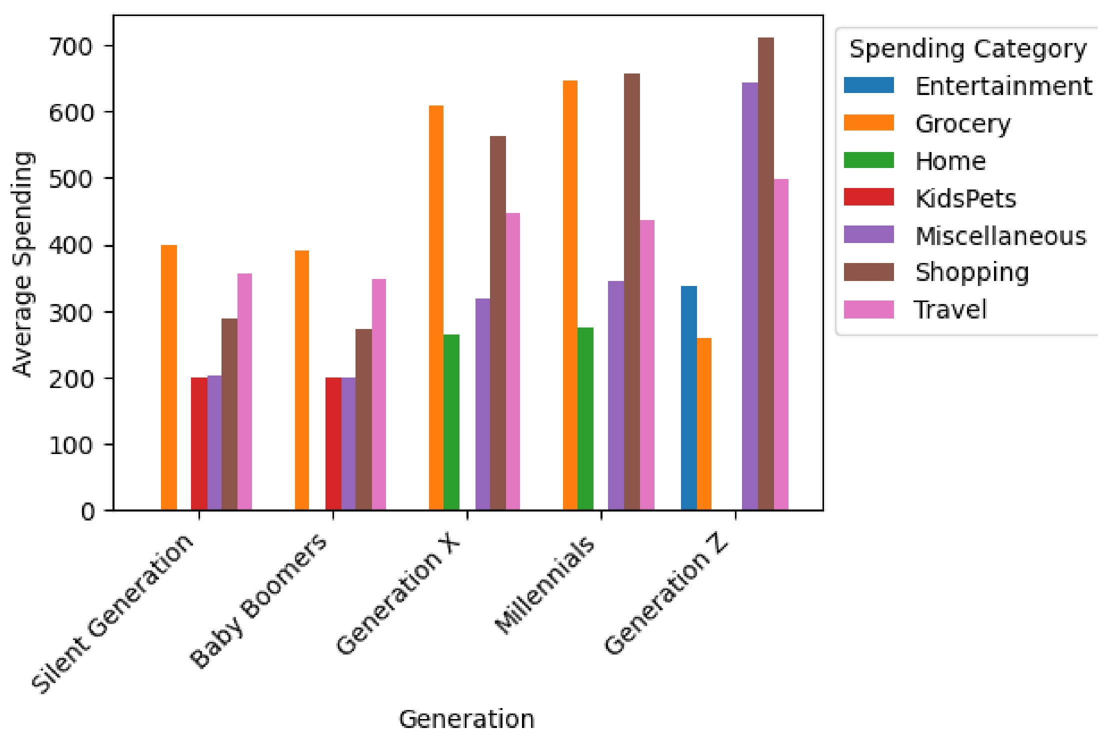


Figure 11: Top 5 Average Spending Categories for Each Generation

The figure 9, 10, and 11 present a visual representation of the distribution of average expenses across different generations and their top 5 spending categories. The data reveals interesting insights into the spending behaviors of each generation.

The bar chart displays the percentage distribution of average expenses among different generations. According to the chart, Millennials have the highest average expenses, accounting for 23.6% of the total. Generation Z closely follows with 23.5% of the average expenses. Generation X comes next with 22.5% of the total, while both the Silent Generation and Baby Boomers have the lowest share, each contributing 15% of the average expenses.

The grouped bar chart provides a breakdown of the top 5 spending categories for each generation. It shows that the Silent Generation, Baby Boomers, and Generation X prioritize Grocery as their top expense category. On the other hand, Millennials and Generation Z prioritize Shopping as their top expense category.

Various factors can influence these spending patterns among different generations. Economic conditions play a significant role, with job opportunities, wage levels, and overall economic growth impacting the purchasing power of each generation. Societal and lifestyle trends can also shape spending behaviors, as younger generations may be more influenced by social media and digital marketing, leading to increased spending on trendy products and experiences.

Technological advancements and the ease of online shopping might also influence the preferences of younger generations, leading to higher expenditures on shopping-related categories. On the other hand, the Silent Generation and Baby Boomers, who may be less tech-savvy, could have different priorities and spending habits.

Moreover, generational characteristics and life stages can impact spending behaviors. For instance, the Silent Generation and Baby Boomers, being older, might prioritize more practical expenses related to healthcare and housing. Millennials and Generation Z, being younger and potentially more mobile, may allocate more of their expenses to leisure activities and experiences.

In conclusion, the visual representations offer valuable insights into the spending behaviors of different generations and the top expense categories for each group. Understanding these patterns and the factors that influence them can help businesses and policymakers tailor their products and services to effectively meet the needs and preferences of each generation.

Spending Behavior Analysis by Job and Salary Range

To better understand the relationship between income and spending, we decided to perform a feature engineering step. Feature engineering involves creating new features or modifying existing ones to extract more meaningful insights from the data.

In our case, we introduced a new feature called "Average_Salary_Range" to the dataset. This feature categorizes individuals into different income brackets based on their salaries. The income brackets are defined as follows:

Salary Range	Average Total Spending
15000-25000	6076.43
25000-40000	5557.11
40000-70000	5854.50
70000-120000	5129.76

Upon analyzing the data, we can observe the following trends:

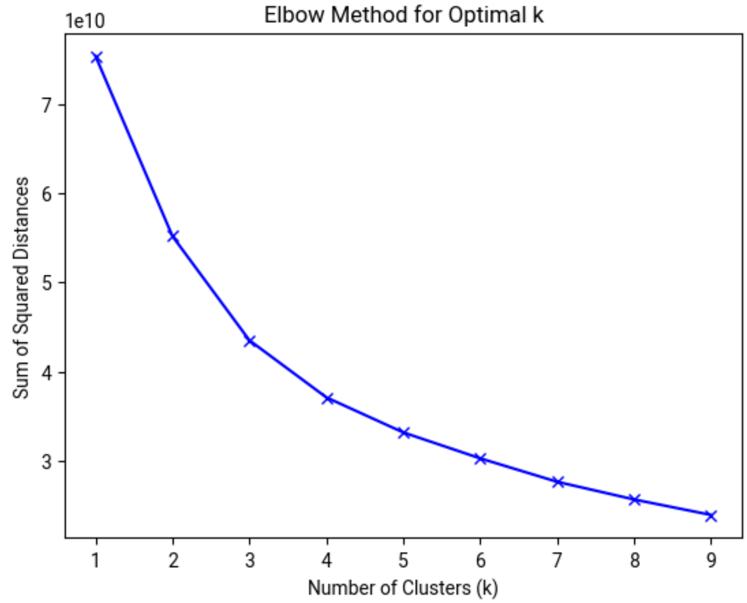
- Firstly, it is observed that people with lower incomes, falling within the "15000-25000" salary range, tend to have the highest average credit card expenses. This could be attributed to various factors, including their specific lifestyle and spending habits. Such individuals may need to allocate a significant portion of their income to cover essential expenses, such as basic necessities and medical costs, which could lead to higher credit card usage.
- On the other hand, as we move to the "25000-40000" salary range, the average total spending slightly decreases. This suggests that individuals with slightly higher incomes may exhibit more cautious spending behavior or prioritize savings and investments over excessive credit card usage.
- Interestingly, the trend changes again in the "40000-70000" salary range, where the average total spending increases. This could be indicative of individuals with moderate incomes having greater financial flexibility to engage in discretionary spending or enjoy a higher standard of living.
- However, the data takes an intriguing turn as we reach the "70000-120000" salary range, where the average total spending significantly decreases. People in this income bracket might adopt a more prudent approach to credit card usage, focusing on efficient debt management and possibly emphasizing long-term financial planning.

Multiple factors could be influencing these spending behaviors. For instance, economic conditions and inflation rates could play a crucial role in shaping the financial choices of individuals across different income groups. Regional variations in the cost of living might also contribute to the observed spending trends.

B. CLUSTERING TECHNIQUES

Clustering techniques are essential tools in data analysis and machine learning that enable us to group similar data points together based on their characteristics or attributes. Market segmentation is a critical application of clustering techniques in marketing and business analysis. It involves dividing a large and diverse market into smaller, more manageable segments or subgroups that share similar characteristics, needs, preferences, or behaviors. The primary goal of market segmentation is to identify distinct customer segments to tailor marketing strategies, products, and services to meet the unique needs and preferences of each segment.

In this project we use Elbow Method to find the optimal number of clusters (k). The Elbow Method helps in determining the number of clusters that provides the best balance between clustering performance and simplicity. For each value of k (ranging from 1 to 9 in this case), the sum of squared distances (SSE) from each data point to its assigned cluster centroid is computed. The SSE is stored in the sse list.



We then plot the SSE values against the number of clusters (k). The plot shows an elbow-shaped curve. The optimal number of clusters is typically the value of k at the "elbow" point, where the SSE starts to level off. In this case, it appears to be around 3 clusters.

In this case, the optimal number of clusters is manually set to 3 based on the Elbow Method results.

Result on the average spending values for each spending category within each cluster after performing K-Means clustering with 3 clusters:

Cluster 1	
Category	**Average Spend**
Entertainment	120.546872
Food	110.092618
HealthFitness	101.084853
Home	157.915180
KidsPets	158.027464
PersonalCare	98.283576
Grocery	313.682963
Miscellaneous	166.808799
Shopping	173.432347
Travel	240.872862

- This cluster has relatively lower average spending across all categories compared to the other clusters.
- The spending in Grocery is the highest among all categories, followed by Shopping and Travel.
- This cluster may represent customers who are more conservative in their spending and have lower overall expenditures.
- Offer budgeting and financial planning tools to help customers manage their expenses effectively.
- Provide incentives for using the credit card in specific categories where their spending is relatively lower.
- Introduce reward programs that align with their spending habits and offer cashback or discounts on essential categories like groceries or home expenses.

Cluster 2	
Category	Average Spend
Entertainment	453.686927
Food	292.445645
HealthFitness	287.40472
Home	428.222937
KidsPets	381.212233
PersonalCare	294.461909
Grocery	1398.031341
Miscellaneous	756.931455
Shopping	2662.424352
Travel	447.544952

- This cluster shows significantly higher spending in almost all categories compared to other clusters.
- The spending in Shopping is the highest among all categories, followed by Grocery and Miscellaneous.
- This cluster may represent customers who are frequent shoppers and have higher overall expenditures.
- Collaborate with retail partners to offer exclusive discounts or rewards for using the credit card at their stores.
- Create personalized shopping experiences by recommending products based on their past purchases and preferences.
- Develop loyalty programs that provide additional benefits for frequent shoppers, such as early access to sales or special promotions.

Cluster 3	
Category	Average Spend
Entertainment	263.045556
Food	201.06
HealthFitness	204.566257
Home	301.15924
KidsPets	280.687368
PersonalCare	196.361696
Grocery	705.123158
Miscellaneous	390.908889
Shopping	564.797544
Travel	8915.798772

- This cluster shows very high spending in the Travel category compared to other clusters.
- The spending in Travel is exceptionally high, while other categories have relatively lower spending.
- This cluster may represent customers who have a specific preference for travel-related expenses and spend significantly more in this category compared to others.
- Collaborate with airlines, hotels, and travel agencies to offer exclusive travel rewards, discounts, or complimentary travel-related services.
- Provide travel insurance or assistance services as additional benefits for using the credit card for travel expenses.
- Offer specialized travel-focused credit card features, such as no foreign transaction fees or enhanced rewards for travel-related purchases.

Market segmentation is a critical aspect of our project as it allows us to understand and target distinct groups of customers with tailored strategies and offerings. By dividing the market into smaller, homogenous segments based on demographic, behavioral, or psychographic characteristics, we gain valuable insights into the diverse needs, preferences, and behaviors of our customer base. This knowledge enables us to develop more effective marketing campaigns, product designs, and pricing strategies that resonate with specific customer groups, ultimately driving higher customer satisfaction and loyalty. Market segmentation also helps optimize resource allocation and budgeting by focusing on the most profitable segments. By implementing a well-defined market segmentation approach, we can efficiently address the unique requirements of our customers, stay ahead of competitors, and foster long-term success for our business.

C. PREDICTIVE MODELING

APRIORI ALGORITHM

In this project, we utilized the Apriori algorithm to analyze customer behavior and uncover valuable patterns in our transactional data. The Apriori algorithm played a vital role in identifying meaningful associations and relationships between different items and categories within our dataset. By conducting market basket analysis, we discovered which products were frequently purchased together, allowing us to optimize product placements and implement targeted marketing strategies. This customer behavior analysis provided invaluable insights into our customers' preferences, interests, and buying habits, enabling us to offer personalized recommendations and enhance the overall customer experience. The use of Apriori empowered us to make data-driven decisions, tailor our services, and improve customer satisfaction. These findings have significant implications for our business, guiding us in making informed decisions and driving growth and success in our organization.

After running the Apriori algorithm, we filtered the results to focus on the top 20 high significant correlations.

- To determine the significance of these associations, we established specific criteria for filtering. We set a minimum confidence level of 0.9, which means that the consequent category must be present in at least 90% of transactions where the antecedent category is present. This high confidence threshold ensures that the associations we consider are very reliable and strong.
- Additionally, we set a minimum lift of 1.2, indicating that the presence of the antecedent category makes the consequent category 1.2 times more likely to be present than if the two categories occurred independently. This condition helps us identify meaningful and interesting relationships between categories.
- Furthermore, we set a minimum conviction of 1.0, which measures the extent to which the consequent category relies on the antecedent category. A conviction of 1.0 indicates that the consequent category does not rely on the antecedent category, and higher values imply a stronger dependence.
- By applying these filtering criteria, we obtained the top 20 high significant correlations between categories. These insights provide valuable information about which categories are frequently associated with each other and can be used to optimize product placements, marketing strategies, and personalized recommendations, ultimately enhancing customer experiences and driving business growth.

Results:

Antecedents	Consequents	Support	Confidence	Lift	Conviction
Miscellaneous, Food	Entertainment	0.812	1	1	inf
Miscellaneous, Grocery	KidsPets	0.89	1	1	inf
HealthFitness	KidsPets, Shopping	0.89	1	1	inf
Shopping	KidsPets, HealthFitness	0.89	1	1	inf
Home, Miscellaneous	Shopping	0.81	1	1	inf
Miscellaneous	Shopping, HealthFitness	0.89	1	1	inf
Travel, Home	Shopping	0.831	1	1	inf
Travel, HealthFitness	KidsPets, Grocery	0.902	1	1	inf
Travel, Grocery	KidsPets, HealthFitness	0.902	1	1	inf
KidsPets, PersonalCare	Grocery	0.919	1	1	inf
PersonalCare, Grocery	KidsPets	0.919	1	1	inf
PersonalCare	KidsPets, Grocery	0.919	1	1	inf
Miscellaneous, PersonalCare	KidsPets	0.813	1	1	inf
KidsPets, PersonalCare	Shopping	0.919	1	1	inf
PersonalCare, Shopping	KidsPets	0.919	1	1	inf
PersonalCare	KidsPets, Shopping	0.919	1	1	inf
Travel, PersonalCare	KidsPets	0.829	1	1	inf
Travel, KidsPets	Grocery, HealthFitness	0.902	1	1	inf
Shopping, HealthFitness	KidsPets, Grocery	1	1	1	inf
Travel, Home	Grocery	0.831	1	1	inf

FORCASTED SPENDING

In this project, the ARIMA (AutoRegressive Integrated Moving Average) model has been employed in the field of machine learning to predict the total spending for the upcoming year. By analyzing historical data, the ARIMA model is capable of identifying patterns and trends, thus providing valuable insights into future spending behavior.

Upon conducting the forecast, an interesting observation has been made. Some of the predicted "next_HOM" data points are lower than the actual "HOM_total" values. This discrepancy indicates that the model predicts a potential decrease in spending for certain periods, which might be due to various factors, such as changes in consumer behavior, economic conditions, or external events.

However, it is essential to interpret these results in context. The observed decline in predicted spending may not necessarily indicate a gloomy outlook for the business. Instead, it could be attributed to uncertainties caused by prevailing economic conditions and restrictions in place during the prediction period.

The report further suggests that the global economy's recovery and the eventual easing of restrictions could significantly impact consumer spending patterns in the forecasted year. If these positive developments take place, there is a possibility of a rebound in consumer spending for 2021 compared to the previous years, 2020 and 2019.

As businesses adapt to changing circumstances and consumers regain confidence, the predicted spending trends may evolve. The ARIMA model's forecasts offer valuable insights to help businesses plan ahead and make informed decisions in anticipation of potential market fluctuations.

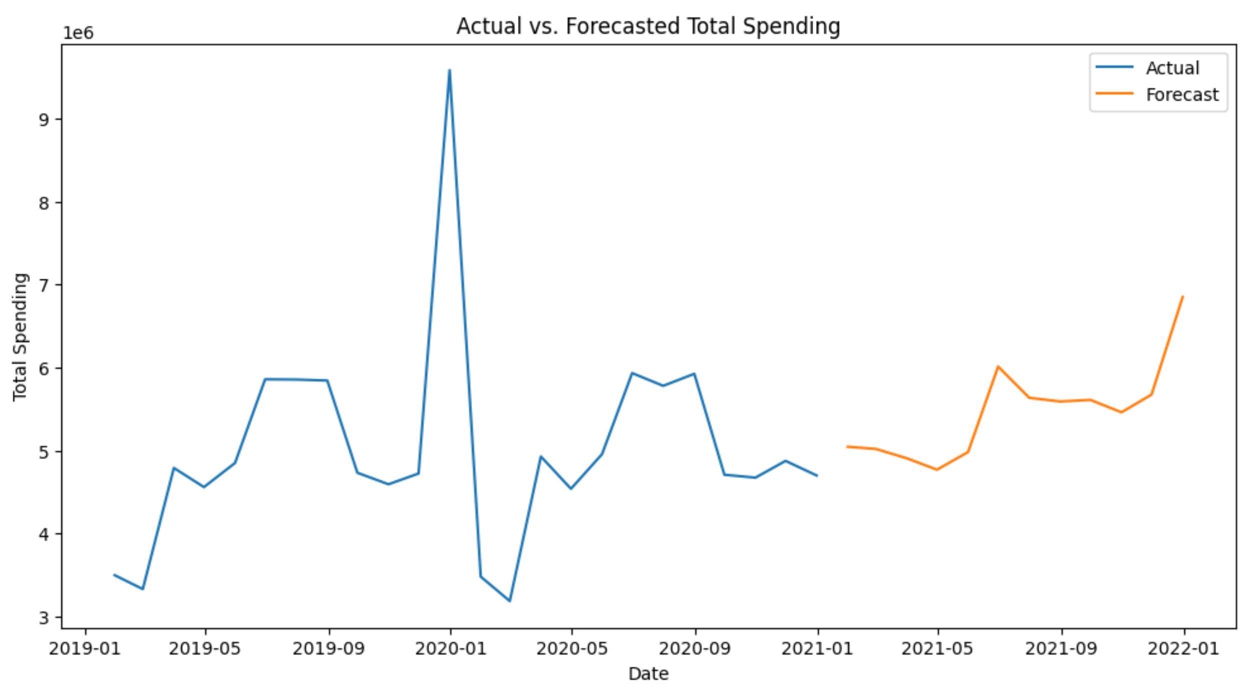


Figure 12: Actual Vs Total Spending

Figure 12 Actual Vs Total Spending provides crucial insights into the financial health and performance of our business. This graph compares the actual spending of our customers with their total spending capacity. By analyzing this comparison, we can gauge the effectiveness of our marketing strategies, pricing decisions, and overall customer engagement.

The significance of this graph lies in its ability to highlight any disparities between customer spending behavior and their potential spending capacity. If the actual spending is consistently below the total spending capacity, it may indicate missed opportunities to capture more revenue from existing customers. On the other hand, if the actual spending exceeds the total spending capacity, it could signal potential financial risks for both our customers and the business.

Understanding these patterns allows us to fine-tune our marketing efforts and align them with customer needs and preferences. By tailoring promotions, discounts, and offers to customers' spending capacity, we can increase their likelihood of making purchases while fostering a positive shopping experience.

Moreover, Figure 12 aids in identifying customer segments with untapped potential or those experiencing financial constraints. Armed with this knowledge, we can implement targeted strategies to cater to specific segments, such as offering installment plans, loyalty rewards, or personalized financial guidance.

In addition to benefiting our business, this graph also provides valuable insights to our customers. By being mindful of their total spending capacity, we can avoid pushing them into financial strain and cultivate long-term customer loyalty and trust.

REPORT AND SIGNIFICANCE

In the modern business landscape, data-driven decision-making has become paramount for organizations aiming to stay competitive and successful. Data Exploratory Analysis, Clustering Techniques, and Predictive Modeling are three powerful tools that play pivotal roles in unlocking the value of data. This report aims to highlight the importance of these techniques and their applications in enhancing business strategies, driving growth, and optimizing operations.

Data Exploratory Analysis:
Data Exploratory Analysis is the initial step in understanding and summarizing the characteristics of a dataset. This technique involves identifying patterns, trends, and potential relationships within the data. By visualizing and exploring the data through charts, graphs, and statistical measures, businesses gain valuable insights that facilitate better decision-making.

Importance:

a) Identify Data Quality Issues: Data Exploratory Analysis helps uncover data anomalies, inconsistencies, or missing values, ensuring data accuracy and reliability.
b) Business Insights: It provides an overview of customer behavior, market trends, and product preferences, enabling businesses to tailor their offerings to meet customer demands.
c) Identify Opportunities: By understanding customer preferences and market dynamics, businesses can identify new opportunities for growth and innovation.
d) Risk Mitigation: Early identification of potential risks and challenges allows businesses to take proactive measures and mitigate potential threats.

Clustering Techniques:
Clustering Techniques involve grouping similar data points together based on their characteristics. It is a vital method for market segmentation, customer profiling, and pattern recognition. By classifying data into clusters, businesses can better understand their target audience and optimize marketing strategies.

Importance:

a) Market Segmentation: Clustering helps businesses divide their customer base into distinct groups based on shared characteristics, allowing targeted marketing campaigns and personalized offerings.
b) Customer Retention: By identifying loyal and high-value customers, businesses can develop retention strategies to enhance customer satisfaction and loyalty.
c) Product Customization: Clustering assists in identifying customer preferences and demands, enabling businesses to develop tailored products and services.
d) Resource Allocation: Clustering helps organizations allocate resources efficiently by understanding which customer segments require more attention and investment.

Predictive Modeling:
Predictive Modeling leverages historical data to make predictions about future outcomes. It involves using statistical algorithms to identify patterns and trends, enabling businesses to make informed forecasts and strategic decisions.

Importance:

a) Demand Forecasting: Predictive Modeling assists businesses in accurately forecasting demand for products or services, optimizing inventory management and supply chain operations.
b) Customer Churn Prediction: By predicting customer churn, businesses can implement targeted retention strategies and reduce customer attrition.
c) Risk Management: Predictive Modeling aids in assessing potential risks and making data-driven decisions to minimize financial losses.
d) Sales and Revenue Optimization: Businesses can optimize pricing strategies and sales tactics by using predictive models to identify factors that drive revenue growth.

Conclusion:
Data Exploratory Analysis, Clustering Techniques, and Predictive Modeling are invaluable tools that empower businesses to harness the power of data effectively. By gaining deeper insights into customer behavior, market trends, and future outcomes, organizations can make informed decisions, enhance operational efficiency, and stay ahead of the competition. Integrating these techniques into business strategies enables businesses to unlock new opportunities, improve customer satisfaction, and achieve sustainable growth in an increasingly data-centric world.

WORKBOOK LINK: https://bit.ly/WorkBookCompilation

OR SCAN THIS QR CODE:

SOURCE CODE

PLOTTING FOR DATA VISUALIZATION

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker

data = pd.read_csv('cleaneddata.csv')

category_columns = ['Entertainment', 'Food', 'HealthFitness', 'Home', 'KidsPets', 'PersonalCare',
'Grocery', 'Miscellaneous', 'Shopping', 'Travel']

total_expenses = data[category_columns].sum()
sorted_expenses = total_expenses.sort_values(ascending=False)

# Define the desired color for the bars
bar_color = '#FF4500'  # Red-Orange color

plt.figure(figsize=(10, 6))
bars = plt.bar(sorted_expenses.index, sorted_expenses.values, color=bar_color)
plt.xlabel('Category')
plt.ylabel('Total Expenses')
plt.title('Total Expenses per Category')
plt.xticks(rotation=45)

# Create a formatter to convert the y-axis values to a shortened format
formatter = ticker.FuncFormatter(lambda x, pos: f"{int(x/1e6)}M")

for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, f"{int(yval/1e6)}M", ha='center', va='bottom')

plt.gca().yaxis.set_major_formatter(formatter)

plt.show()
```

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('cleaneddata.csv')

# Calculate the frequency of each category
category_counts = df['Most_Frequent_Category'].value_counts()

# Sort the categories by frequency in descending order
sorted_categories = category_counts.sort_values(ascending=False)

# Get the category names and their corresponding frequencies
categories = sorted_categories.index.tolist()
frequencies = sorted_categories.values.tolist()

# Create a bar graph
plt.bar(categories, frequencies)

# Set the title and labels
plt.title('Most Frequent Categories')
plt.xlabel('Category')
plt.ylabel('Frequency')

# Rotate x-axis labels for better readability
plt.xticks(rotation='vertical')

# Display the bar graph
plt.show()
```

```
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Data Preprocessing

# Load the CSV file into a pandas DataFrame
file_path = "cleaneddata.csv"  # Replace 'cleaneddata.csv' with the actual file path
df = pd.read_csv(file_path)

# Convert the 'dob' column to datetime type, handling inconsistent date formats
df['dob'] = pd.to_datetime(df['dob'], format="%d/%m/%Y", errors='coerce')
# Plot the grouped bar chart with custom x-axis tick labels
```

```
# Step 2: Extract Generations and Group Data

# Define the generation categories based on the provided date ranges
generations = {
    'Silent Generation': (1925, 1945),
    'Baby Boomers': (1946, 1964),
    'Generation X': (1965, 1980),
    'Millennials': (1981, 1996),
    'Generation Z': (1997, 2012),
}

# Create a function to assign the generation category based on the 'dob' column
def get_generation(dob):
    if pd.isnull(dob):  # Handle NaT values resulting from inconsistent date formats
        return None
    year = dob.year
    for gen, (start, end) in generations.items():
        if start <= year <= end:
            return gen

# Apply the get_generation function to create a new 'Generation' column
df['Generation'] = df['dob'].apply(get_generation)

# Step 3: Define Spending Categories and Create the Grouped Bar Chart

# Define the spending categories that you want to consider in the analysis
categories = ['Entertainment', 'Food', 'Grocery', 'HealthFitness', 'Home', 'KidsPets', 'PersonalCare',
'Miscellaneous', 'Shopping', 'Travel']

# Calculate the average spending for each spending category and each generation
grouped_data = df.groupby('Generation')[categories].mean()

# Sort the data by the top 5 average spending categories for each generation
sorted_data = grouped_data.apply(lambda row: row.sort_values(ascending=False).head(5),
axis=1)

# Define the desired sequence of generations for the x-axis tick labels
generation_sequence = ['Silent Generation', 'Baby Boomers', 'Generation X', 'Millennials',
'Generation Z']
```

```
# Plot the grouped bar chart with custom x-axis tick labels
plt.figure(figsize=(12, 8))
sorted_data = sorted_data.reindex(generation_sequence)  # Reorder the data based on the
generation_sequence
sorted_data.plot(kind='bar', width=0.8)  # Adjust width as per your preference

plt.xlabel("Generation")
plt.ylabel("Average Spending")
plt.title("Top 5 Average Spending Categories for Each Generation")
plt.legend(title="Spending Category", bbox_to_anchor=(1, 1))
plt.tight_layout()

# Set the x-axis tick labels to display generations in the desired sequence
plt.xticks(range(len(generation_sequence)), generation_sequence, rotation=45, ha='right')

plt.show()
```

```
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Data Preprocessing

# Load the CSV file into a pandas DataFrame
file_path = "cleaneddata.csv"  # Replace 'cleaneddata.csv' with the actual file path
df = pd.read_csv(file_path)

# Convert the 'dob' column to datetime type, handling inconsistent date formats
df['dob'] = pd.to_datetime(df['dob'], format="%d/%m/%Y", errors='coerce')

# Step 2: Extract Data for Years 2019 and 2020

# Filter the data for years 2019 and 2020
df_2019_2020 = df[(df['year'] == 2019) | (df['year'] == 2020)]

# Calculate the total spending for each month in 2019 and 2020
total_expenses_per_month_2019_2020 = df_2019_2020.groupby(['year',
'month'])[categories].sum().sum(axis=1).unstack()

# Step 3: Create the Line Graph for Years 2019 and 2020
```

```python
# Step 3: Create the Line Graph for Years 2019 and 2020

# Plot the line graph for years 2019 and 2020
plt.figure(figsize=(10, 6))
months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']
plt.plot(range(1, 13), total_expenses_per_month_2019_2020.loc[2019], marker='o', linestyle='-',
color='b', label='2019')
plt.plot(range(1, 13), total_expenses_per_month_2019_2020.loc[2020], marker='o', linestyle='-',
color='g', label='2020')

plt.xticks(range(1, 13), months)  # Set x-axis tick labels to months
plt.xlabel("Month")
plt.ylabel("Total Expenses")
plt.title("Total Expenses per Month for Years 2019 and 2020")
plt.legend()

plt.show()
```

```python
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Data Preprocessing

# Load the CSV file into a pandas DataFrame
file_path = "cleaneddata.csv"  # Replace 'cleaneddata.csv' with the actual file path
df = pd.read_csv(file_path)

# Convert the 'dob' column to datetime type, handling inconsistent date formats
df['dob'] = pd.to_datetime(df['dob'], format="%d/%m/%Y", errors='coerce')

# Step 2: Extract Data for Year 2019

# Filter the data for year 2019
df_2019 = df[df['year'] == 2019]
```

```
# Step 3: Calculate the Expenses per Month in Year 2019 per Category

# Define the spending categories that you want to consider in the analysis
categories = ['Entertainment', 'Food', 'Grocery', 'HealthFitness', 'Home', 'KidsPets', 'PersonalCare',
'Miscellaneous', 'Shopping', 'Travel']

# Calculate the expenses per month in year 2019 per category
expenses_per_month_2019 = df_2019.groupby('month')[categories].sum()

# Step 4: Create the Line Graph

plt.figure(figsize=(12, 6))

# Plot each category as a separate line on the graph
for category in categories:
    plt.plot(expenses_per_month_2019.index, expenses_per_month_2019[category], marker='o',
linestyle='-', label=category)

plt.xlabel("Month")
plt.ylabel("Expenses")
plt.title("Expenses per Month in Year 2019 per Category")
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.legend()
plt.grid(True)

plt.show()
```

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.ticker import StrMethodFormatter

# Step 1: Data Preprocessing

# Load the CSV file into a pandas DataFrame
file_path = "cleaneddata.csv"  # Replace 'cleaneddata.csv' with the actual file path
df = pd.read_csv(file_path)

# Convert the 'dob' column to datetime type, handling inconsistent date formats
df['dob'] = pd.to_datetime(df['dob'], format="%d/%m/%Y", errors='coerce')

# Step 2: Extract Data for Year 2020

# Filter the data for year 2020
df_2020 = df[df['year'] == 2020]
```

```
# Step 3: Calculate the Expenses per Month in Year 2020 per Category

# Define the spending categories that you want to consider in the analysis
categories = ['Entertainment', 'Food', 'Grocery', 'HealthFitness', 'Home', 'KidsPets', 'PersonalCare',
'Miscellaneous', 'Shopping', 'Travel']

# Calculate the expenses per month in year 2020 per category
expenses_per_month_2020 = df_2020.groupby('month')[categories].sum()

# Step 4: Create the Line Graph

plt.figure(figsize=(12, 6))

# Plot each category as a separate line on the graph
for category in categories:
    plt.plot(expenses_per_month_2020.index, expenses_per_month_2020[category], marker='o',
linestyle='-', label=category)

plt.xlabel("Month")
plt.ylabel("Expenses")
plt.title("Expenses per Month in Year 2020 per Category")
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.legend()

# Set y-axis tick format to display exact digits
plt.gca().yaxis.set_major_formatter(StrMethodFormatter('{x:,.0f}'))

plt.grid(True)

plt.show()
```

```
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Data Preprocessing

# Load the CSV file into a pandas DataFrame
file_path = "cleaneddata.csv"  # Replace 'cleaneddata.csv' with the actual file path
df = pd.read_csv(file_path)

# Convert the 'dob' column to datetime type, handling inconsistent date formats
df['dob'] = pd.to_datetime(df['dob'], format="%d/%m/%Y", errors='coerce')
```

```python
# Step 2: Calculate Total Expenses per Category for Each Gender

# Define the spending categories that you want to consider in the analysis
categories = ['Entertainment', 'Food', 'Grocery', 'HealthFitness', 'Home', 'KidsPets', 'PersonalCare',
'Miscellaneous', 'Shopping', 'Travel']

# Calculate the total expenses per category for each gender
grouped_data = df.groupby('gender')[categories].sum()

# Step 3: Create the Grouped Bar Chart

plt.figure(figsize=(10, 6))

# Get the number of categories and genders for positioning the bars
num_categories = len(categories)
num_genders = len(grouped_data)

# Set the width of the bars
bar_width = 0.35

# Set the positions of the bars on the x-axis
positions = list(range(1, num_categories + 1))

# Transpose the grouped_data DataFrame to swap rows and columns
transposed_data = grouped_data.transpose()

# Plot the grouped bar chart for male and female
plt.bar(positions, transposed_data['M'], width=bar_width, label='Male')
plt.bar([pos + bar_width for pos in positions], transposed_data['F'], width=bar_width,
label='Female')

# Set the x-axis labels to represent the spending categories
plt.xticks([pos + bar_width / 2 for pos in positions], categories, rotation=45)

plt.xlabel("Category")
plt.ylabel("Total Expenses")
plt.title("Total Expenses per Category for Each Gender")
plt.legend()

plt.tight_layout()
plt.show()
```

```python
# Step 2: Calculate Total Expenses per Category for Each Gender

# Define the spending categories that you want to consider in the analysis
categories = ['Entertainment', 'Food', 'Grocery', 'HealthFitness', 'Home', 'KidsPets', 'PersonalCare',
'Miscellaneous', 'Shopping', 'Travel']

# Calculate the total expenses per category for each gender
grouped_data = df.groupby('gender')[categories].sum()

# Step 3: Create the Grouped Bar Chart

plt.figure(figsize=(10, 6))

# Get the number of categories and genders for positioning the bars
num_categories = len(categories)
num_genders = len(grouped_data)

# Set the width of the bars
bar_width = 0.35

# Set the positions of the bars on the x-axis
positions = list(range(1, num_categories + 1))

# Transpose the grouped_data DataFrame to swap rows and columns
transposed_data = grouped_data.transpose()

# Plot the grouped bar chart for male and female
plt.bar(positions, transposed_data['M'], width=bar_width, label='Male')
plt.bar([pos + bar_width for pos in positions], transposed_data['F'], width=bar_width,
label='Female')

# Set the x-axis labels to represent the spending categories
plt.xticks([pos + bar_width / 2 for pos in positions], categories, rotation=45)

plt.xlabel("Category")
plt.ylabel("Total Expenses")
plt.title("Total Expenses per Category for Each Gender")
plt.legend()

plt.tight_layout()
plt.show()
```

```python
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Data Preprocessing

# Load the CSV file into a pandas DataFrame
file_path = "cleaneddata.csv"  # Replace 'cleaneddata.csv' with the actual file path
df = pd.read_csv(file_path)

# Convert the 'dob' column to datetime type, handling inconsistent date formats
df['dob'] = pd.to_datetime(df['dob'], format="%d/%m/%Y", errors='coerce')

# Step 2: Calculate the Distribution of Customers per Generation

# Define the generation categories based on the provided date ranges
generations = {
    'Silent Generation': (1925, 1945),
    'Baby Boomers': (1946, 1964),
    'Generation X': (1965, 1980),
    'Millennials': (1981, 1996),
    'Generation Z': (1997, 2012),
}

# Create a function to assign the generation category based on the 'dob' column
def get_generation(dob):
    if pd.isnull(dob):  # Handle NaT values resulting from inconsistent date formats
        return None
    year = dob.year
    for gen, (start, end) in generations.items():
        if start <= year <= end:
            return gen

# Apply the get_generation function to create a new 'Generation' column
df['Generation'] = df['dob'].apply(get_generation)

# Calculate the distribution of customers per generation
customer_distribution_per_generation = df['Generation'].value_counts()
```

```
# Step 3: Calculate the Combined Average Expenses per Generation

# Calculate the average expenses for each generation
average_expenses_per_generation = df.groupby('Generation')['Total_Spending'].mean()

# Step 4: Create the Pie Charts

# Create the figure and axes for the two pie charts
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 6))

# Define custom colors for the pie charts
colors_generation = ['#1f77b4', '#ff7f0f', '#2ca02c', '#d62728', '#9467bd']

# Create the first pie chart for the distribution of customers per generation
ax1.pie(customer_distribution_per_generation,
labels=customer_distribution_per_generation.index, autopct='%1.1f%%',
colors=colors_generation)
ax1.set_title("Distribution of Customers per Generation")

# Create the second pie chart for the combined average expenses per generation
ax2.pie(average_expenses_per_generation, labels=average_expenses_per_generation.index,
autopct='%1.1f%%', colors=colors_generation)
ax2.set_title("Combined Average Expenses per Generation")

# Display the two pie charts
plt.tight_layout()
plt.show()
```

```
import pandas as pd

# Load the CSV file
file_path = "top_50_jobs_average_spending.csv"
df = pd.read_csv(file_path)

# Convert 'Average_Salary_Range' column to numeric
df['Average_Salary_Range'] = df['Average_Salary_Range'].str.replace(',', '').str.extract(r'(\d+)',
expand=False).astype(float)

# Define the job salary ranges
salary_ranges = [
    (15000, 25000),
    (25000, 40000),
```

```
    (40000, 70000),
    (70000, 120000)
]

# Calculate the average total spending for each job salary range
result = []
for salary_range in salary_ranges:
    lower_bound, upper_bound = salary_range
    mask = (df['Average_Salary_Range'] >= lower_bound) & (df['Average_Salary_Range'] <
upper_bound)
    average_spending = df.loc[mask, 'Total_Spending'].mean()
    result.append((f"{lower_bound}-{upper_bound}", average_spending))

# Display the results
for range_str, average_spending in result:
    print(f"Salary Range: {range_str}")
    print(f"Average Total Spending: ${average_spending:.2f}")
    print()
```

```
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
# Read the dataset
df = pd.read_csv('apriori.csv')

# Drop any rows with missing values (NaN)
df.dropna(inplace=True)

# Convert categorical columns to string, as Apriori requires string values
df['gender'] = df['gender'].astype(str)
df['job'] = df['job'].astype(str)

# Convert the 'dob' column to datetime type
df['dob'] = pd.to_datetime(df['dob'])

# Convert the 'month' and 'HOM_tag' columns to object (string) type
df['month'] = df['month'].astype(str)
df['HOM_tag'] = df['HOM_tag'].astype(str)

# Create a list of columns containing item categories
category_columns = [col for col in df.columns if col.endswith('_Freq')]
```

```python
# Convert the item category columns to binary format (1 if frequency > 0, 0 otherwise)
for col in category_columns:
    df[col] = df[col].apply(lambda x: 1 if x > 0 else 0)

# Apriori algorithm to find frequent itemsets
frequent_itemsets = apriori(df[category_columns], min_support=0.05, use_colnames=True)

# Association rules generation
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)

# Display frequent itemsets
print("Frequent Itemsets:")
print(frequent_itemsets)

# Display association rules
print("\nAssociation Rules:")
print(rules)
```

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.statespace.sarimax import SARIMAX

data = pd.read_csv('cleaneddata.csv')

# Convert 'year' and 'month' columns to datetime format
data['date'] = pd.to_datetime(data['year'].astype(str) + '-' + data['month'].astype(str))

# Set the 'date' column as the index
data.set_index('date', inplace=True)

# Resample the data to monthly frequency
data = data.resample('M').sum()

# Plot the total spending over time
plt.figure(figsize=(12, 6))
plt.plot(data['Total_Spending'])
plt.xlabel('Date')
plt.ylabel('Total Spending')
plt.title('Total Spending Over Time')
plt.show()
```

```
# Decompose the time series
decomposition = seasonal_decompose(data['Total_Spending'], model='additive')

# Plot the decomposed components
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(data['Total_Spending'], label='Original')
plt.legend(loc='best')
plt.subplot(412)
plt.plot(decomposition.trend, label='Trend')
plt.legend(loc='best')
plt.subplot(413)
plt.plot(decomposition.seasonal, label='Seasonality')
plt.legend(loc='best')
plt.subplot(414)
plt.plot(decomposition.resid, label='Residuals')
plt.legend(loc='best')
plt.tight_layout()
plt.show()

num_data_points = len(data['Total_Spending'])
print("Number of data points:", num_data_points)

from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Plot ACF
plt.figure(figsize=(12, 6))
plot_acf(data['Total_Spending'], lags=12, alpha=0.05)
plt.xlabel('Lag')
plt.ylabel('Autocorrelation')
plt.title('Autocorrelation Function (ACF)')
plt.show()

# Plot PACF
plt.figure(figsize=(12, 6))
plot_pacf(data['Total_Spending'], lags=6, alpha=0.05)
plt.xlabel('Lag')
plt.ylabel('Partial Autocorrelation')
plt.title('Partial Autocorrelation Function (PACF)')
plt.show()
```

```python
# Define the SARIMAX model parameters
order = (1, 1, 1)  # Defin4 the order of autoregressive (p), differencing (d), and moving average (q) terms
seasonal_order = (1, 1, 1, 6)  # Define the seasonal order of autoregressive (P), differencing (D), moving average (Q) terms, and the period (s)

# Create and fit the SARIMAX model
model = SARIMAX(data['Total_Spending'], order=order, seasonal_order=seasonal_order)
model_fit = model.fit()

# Optionally, print the model summary for details
print(model_fit.summary())

# Forecast future spending trends
forecast_periods = 12  # Number of periods to forecast
forecast = model_fit.get_forecast(steps=forecast_periods)

# Extract the forecasted values
forecasted_values = forecast.predicted_mean

# Visualize the forecasted values
plt.figure(figsize=(12, 6))
plt.plot(data['Total_Spending'], label='Actual')
plt.plot(forecasted_values, label='Forecast')
plt.xlabel('Date')
plt.ylabel('Total Spending')
plt.title('Actual vs. Forecasted Total Spending')
plt.legend(loc='best')
plt.show()

# Print the forecasted values
print(forecasted_values)
```