Data Science Project

Amazon Sales Data Analysis: Enhancing Business Strategies

Author: Antony Susai Victor Velankanni Raj

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Abstract

The present research report, entitled "Amazon Sales Data Analysis: Enhancing Business Strategies," offers a comprehensive examination of Amazon's sales data through the use of sophisticated data science methodologies. The principal objective is to comprehend and enhance several facets of Amazon's e-commerce operations by utilising market basket analysis, predictive modelling, and consumer segmentation. To analyse intricate sales patterns, the approach makes use of complicated methods including time series analysis, ARIMA modelling, and clustering approaches. The results provide insightful information about sales patterns, consumer behaviour, and future market prospects. In the cutthroat world of e-commerce, these data are essential for creating focused marketing campaigns, raising consumer satisfaction levels, and eventually propelling company expansion. The paper shows how strategic planning and decision-making processes in retail firms may be greatly improved by data-driven techniques.

Table of Contents

TABLE OF FIGURES

Acknowledgements

I would like to express my gratitude to my supervisor Eziamaka Nwakile for their insightful advice and persistent support throughout the study process. Their knowledge and insightful opinions were critical in influencing the project's path.

1 Introduction

Organisations of all sizes are realising more and more the enormous importance of data-driven insights for strategic decision-making in today's fast and demanding business environment. Not even the biggest online retailer in the world, Amazon, is an exception. Amazon has amassed a large amount of data about its customers, goods, and sales. This information may be used to uncover new trends, improve business tactics, and obtain a deeper understanding of customer behaviour. Intending to delve deeply into Amazon sales data, this initiative aims to extract valuable insights and offer Amazon sellers practical advice on how to improve their company operations and increase sales. Large-scale data analytics has made it possible for businesses to extract insightful knowledge from consumer data, which has encouraged the creation of data-driven strategies that improve customer happiness and propel company expansion. This dissertation explores the complex field of customer analytics, using cutting-edge data science methods to identify trends, divide up the clientele, forecast behaviour, and ultimately improve marketing and sales tactics. The project will follow the concepts of Systems Development Life Cycle (SDLC) planning methodologies, which call for a methodical approach. With steps including data pretreatment, exploratory data analysis, feature engineering, statistical prediction, and thorough testing to validate the results, this methodical methodology will guarantee the project's successful completion.

1.1 Background to the Project

The project's drive comes from the growing difficulties that contemporary e-commerce platforms must overcome, as demonstrated by Amazon's enormous influence on the retail industry. Amazon produces a huge amount of transactional data, which captures a multitude of consumer interactions, preferences, and buying behaviours, due to its wide range of products and large user base. As e-commerce develops more, companies—Amazon among them—face the challenge of gleaning useful information from this mountains of data. Amazon's success is not merely a testament to its scale but to its adept use of data-driven strategies. The e-commerce giant's mastery of predictive analytics, personalized recommendations, and seamless customer experiences illustrates the transformative power of harnessing data.The motivation stems from the belief that understanding and adopting Amazon's data-driven approach can empower retailers, big and small, to navigate the challenges of the digital age.

This research is motivated by the realisation that retail operations can distinguish industry leaders from laggards based on how well data analytics is integrated. Retailers need to not only adjust their strategies to keep ahead of the competition, but also proactively shape them in light of the continually evolving tastes of their customers. Data-driven decision-making is a strategic necessity rather than just a fad in technology. Retailers may connect their services, marketing campaigns, and general business approach with the preferences of their audience by having a detailed grasp of customer behaviour. The driving force is the conviction that merchants can prosper by establishing enduring relationships with their customers and not just surviving in the digital era with the correct analytical tools at their disposal.

The breakthrough impact of data analytics in influencing strategic decision-making processes within the retail industry is the driving force behind this research. Large-scale datasets, sophisticated analytics, and machine learning approaches come together to provide organisations a never-before-seen chance to understand customer behaviour in great detail, streamline internal operations, and ultimately improve the customer experience. The ideas put forth have the potential to benefit e-commerce companies, marketing experts, and data scientists who are trying to make their way through the complex world of customer-centric strategies. Through the analysis of customer behaviour, the project seeks to provide relevant tools and techniques to stakeholders so they may improve their operational strategies and the e-commerce experience for both customers and businesses.

2 Project Objectives

This project's main goal is to use data analytics to transform e-commerce strategies. The research intends to identify frequently purchased product combinations, anticipate buy inclinations, and find patterns in customer behaviour by utilising market basket analysis, predictive modelling, and customer segmentation. Providing companies with useful information for cross-selling, product placement, and customised marketing campaigns. In the ever-changing world of e-commerce, the project aims to improve consumer pleasure, loyalty, and overall business success by applying state-of-the-art approaches including clustering algorithms and predictive modelling.

2.1 Key issues faced by modern merchants

Questions like "What strategies will resonate most with our target audience?" and "How can we anticipate customer needs?" By exploring the world of data-driven strategies, this dissertation aims to address these issues and provides a thorough road map for attaining customer-centric retail excellence.

3 Overview of This Report

Introduction: Explains the importance of data-driven insights for strategic decision-making in ecommerce, focusing on Amazon.

Literature Review: Covers key concepts in time series modeling and predictive analytics relevant to Amazon's sales data analysis.

Methodology: Details the Systems Development Life Cycle (SDLC) approach, including data preparation, exploratory data analysis, feature engineering, statistical prediction, and model testing.

Implementation of Methodology: Discusses the specific models used and their performance, including data preparation, customer segmentation, predictive modeling, and market basket analysis.

Requirements: Outlines the functional and non-functional requirements of the system developed.

Design: Describes the selection and preparation of data, the modeling methodology, evaluation metrics, and implementation tools used.

Data Analysis: Provides a thorough examination of the dataset, identifying patterns, trends, and insights crucial for predictive models and strategies.

Models and Structure of Data: Includes exploratory data analysis, time series analysis, customer segmentation, and predictive analytics.

Predictions and Outputs: Presents the results of time series analysis, customer segmentation, predictive analytics, and market basket analysis.

Project Management: Describes the project schedule and risk management strategies.

Testing: Details the testing methods used.

Conclusions: Summarizes the achievements and areas for future work.

4 Literature Review

Understanding and utilizing customer data is essential for preserving competitive advantage and fostering business success in the quickly changing e-commerce industry. The foundation of contemporary online retailing, especially for market leaders like Amazon, is the convergence of growing e-commerce, analytics on customer behavior, and strategic business planning.

According to (Zhou, (2007)), who emphasize the crucial importance of big data in comprehending customer buying patterns, an essential component of e-commerce strategy is the analysis of market trends and consumer preferences. E-commerce platforms may obtain profound insights into customer behavior through the application of advanced data analytics, a technique that (Hsiao, (2009).)confirms is essential to forecasting purchase decisions. Predicting the wants and needs of customers allows companies like Amazon to customize their products, which raises customer satisfaction and loyalty. (Porter, (2001).) In the digital economy,competitive strategy paradigm is still applicable. Online platforms where data-driven personalization can produce distinctive value propositions can benefit from the application of differentiation and cost leadership methods. This is consistent with the methodology used by Amazon, where inventory management and customized customer experiences are informed by sales data analysis.
(Elmaghraby, (2003))have examined dynamic pricing techniques, which are another essential component of an effective e-commerce strategy. Platforms can optimize pricing in real-time by making adjustments depending on consumer behavior and market demand. In this case, data analytics is essential because it gives Amazon the knowledge it needs to set prices that accurately represent the state of the market.
It is impossible to overestimate the significance of sales forecasting in respect to inventory control. (Agrawal V. &., (2009).) investigate demand-forecasting predictive models that enable inventory optimization. Predictive analytics can lead to significant cost savings and enhanced customer service through better supply management, which is especially important to Amazon's extensive inventory.

Additionally, CRM systems use data analytics to improve customer connections, as explained by (Kumar, (2012)). CRM systems powered by data analytics are crucial to Amazon's efforts to build repeat business and consumer loyalty. CRM and data analytics together can result in more robust marketing campaigns, client retention, and eventually a higher profit margin.
The methods that make use of data analytics and technology innovations will grow more complex as the e-commerce sector expands. The ability to foresee and adjust to changes in consumer behavior, market dynamics, and technical improvements will be critical to the success of ecommerce strategies in the future. Amazon's sustained expenditure in data analysis and predictive modeling is evidence of the effectiveness of data-driven strategy in online sales. (Zhou, (2007))

Time Series Modeling (B. Singh, 2020)

A comprehensive overview of time series modeling techniques for forecasting Amazon sales. This information can be used to inform the selection of an appropriate forecasting model for my own data.
It provides a detailed comparison of three different time series forecasting models: Holt-Winters exponential smoothing, neural network autoregression (NNAR), and Autoregressive Integrated Moving Average (ARIMA). This comparison can help me to understand the strengths and weaknesses of each model and choose the one that is best suited for my specific needs((B. Singh, 2020).It discusses the importance of evaluating model performance using appropriate metrics, such as mean absolute percentage error (MAPE). This information can help me to assess the accuracy of my own forecasting models. time series analysis to identify seasonality and predict future sales trends. The paper provides a detailed discussion of how to use time series models to identify

seasonal patterns in sales data and forecast future sales trends. This information can be directly applied to my project to develop a forecasting model that can accurately predict future

sales for my own Amazon business. Identify important customers, improve pricing methods, and create predictive models for sales forecasting.

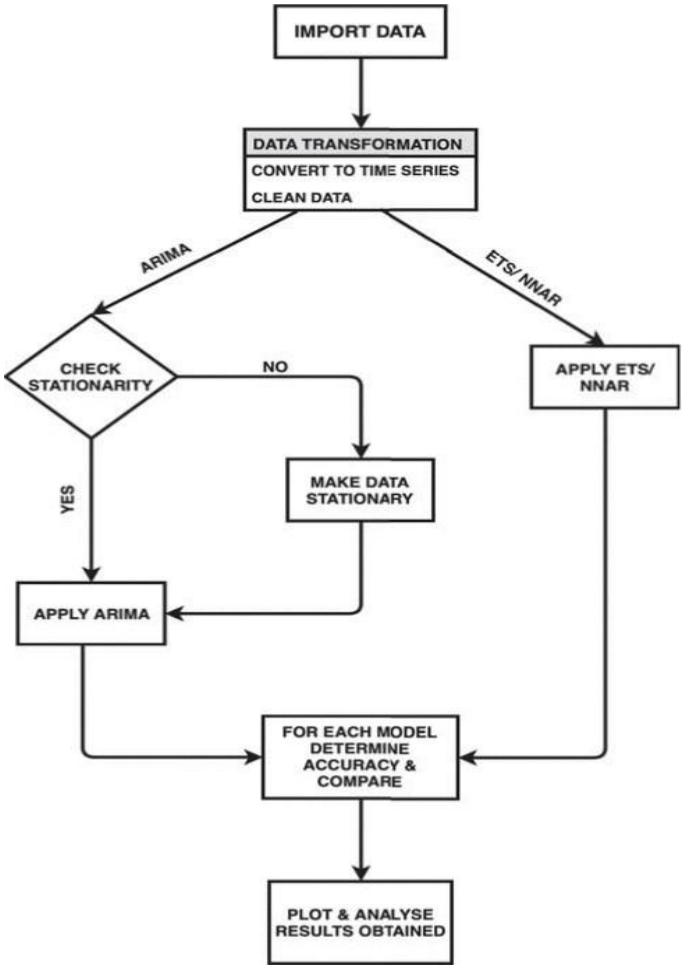

Figure 1 Time series flow

(Sales Forecast for Amazon Sales with Time Series Modeling" by B. Singh, P. Kumar, N. Sharma, and K. P. Sharma)The paper discusses how time series models can be used to identify important customers, optimize pricing, and develop predictive models for sales forecasting. This information can be used to inform my own efforts to improve customer acquisition and retention, optimize pricing strategies, and develop accurate sales forecasts for my business. Winters' exponential smoothing, time-series decomposition, and ARIMA. This comparison can help me to understand the strengths and weaknesses of each model and choose the one that is best suited for my specific needs. Machine Learning Integrated Into Time Series Forecasting

Hybrid models that combine the best features of both machine learning and classical time series forecasting techniques have been produced. The study by provides an example of how nonlinear interactions in time series data may be modelled using neural networks. These models can then be integrated with ARIMA models to accommodate both linear and nonlinear dynamics. When it comes to predicting Amazon sales, this hybrid method can be especially helpful because the data may show intricate patterns that are difficult for linear models to fully capture.

It discusses the importance of evaluating model performance using appropriate metrics, such as mean absolute percentage error (MAPE) (B. Singh, 2020). This information can help me to assess the accuracy of my own forecasting models

ARIMA

A statistical technique for time series analysis and forecasting is the ARIMA (Autoregressive Integrated Moving Average) model. To identify patterns and trends in the data, it combines autoregressive (AR), moving average (MA), and differencing (I) components. The association between an observation and a predetermined number of lag observations is taken into account by the AR component of the model. It makes the assumption that a variable's previous values determine its current value. The number of lag observations in the model is represented by the order of the AR component, p.

The relationship between an observation and a residual error from a moving average model applied to lagged observations, on the other hand, is the main emphasis of the MA component (Chowdhury, 2021).It records the data's sporadic oscillations and haphazard shocks. The amount of lag residual errors in the model is indicated by the order of the MA component, or q.

By using the differencing component, the data is changed into a stationary series. In time series analysis, stationarity is a crucial premise since it guarantees that the statistical characteristics of the data don't change with time. In order to eliminate trends and seasonality, differencing entails taking the difference between two consecutive data. The number of times differencing is used to ensure stationarity is indicated by the order of differencing, or d.

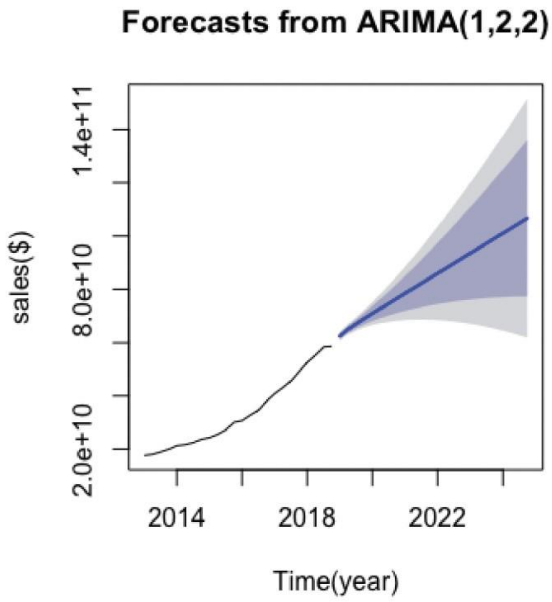

Figure 2 ARIMA FORECAST 1

In this figure we see that arima forcast by (B. Singh, 2020).There are usually numerous processes involved in developing an ARIMA model. Initially, autocorrelation in the data is assessed using tests like the Ljung-Box Test. The link between an observation and its lag values is measured by autocorrelation. An AR component might be required if there is a substantial amount of autocorrelation (Chowdhury, 2021). Next, procedures such as the Augmented Dickey-Fuller test are used to determine if the data is stationary. Differentiating is used if the data is non-stationary until stationarity is attained.

A number of metrics are used to assess the ARIMA model's performance. The average discrepancy between the expected and actual values is measured by the root mean square error, or RMSE. The average percentage difference between the expected and actual values is determined using the

mean absolute percentage error, or MAPE. The average difference between the expected and actual numbers is measured using the mean absolute error (MAE), which does not take the discrepancy's direction into account. The percentage of the variance in the data that the model can explain is shown by the R-squared number. In the context of the Chittagong Stock Exchange (CSE), the study examined the daily share prices of 60 companies from January 2019 to December 2019. ARIMA models were selected for each company based on tests for autocorrelation and stationarity. The forecasted share prices for January 2020 were calculated using the selected ARIMA models. The models were validated using goodness-of-fit tests and measures such as RMSE, MAPE, and R-squared. The results showed that the ARIMA models were applicable for forecasting the daily share prices of the CSE (Chowdhury, 2021).

Overall, the ARIMA model is a widely used and effective tool for time series analysis and forecasting. It allows for the identification of patterns and trends in the data, as well as the prediction of future values. By considering the autoregressive, moving average, and differencing components, the model captures both short-term fluctuations and long-term dependencies in the data

EXPONENTIAL SMOOTHING

For univariate data, exponential smoothing is a time series forecasting technique that is employed. It can be expanded to accommodate data that has a seasonal or systematic trend component. In order to minimise overestimation or underestimating mistake, the smoothing constants are adjusted using this method. Forecasting frequently makes use of exponential smoothing models, such as double exponential smoothing (DES) and simple exponential smoothing (SES) (Rubio, 2021). Models of **Exponential Smoothing in State Space:**

Since time series analysis has advanced, state space models have been used more frequently for exponential smoothing. According to (Hyndman, 2008), these models provide a more adaptable framework for encapsulating different time series data elements, such trend and seasonality. The Kalman filter, which offers a recursive solution to the forecasting issue and accommodates changes in trends and seasonal patterns dynamically across time, may be used using the state space form. In situations when standard approaches might not be sufficient, this methodology is helpful for adjusting to the intricate and dynamic patterns of Amazon sales data.

When a bullish or bearish market is detected in the time series, DES is applied, whereas SES is utilised to forecast time series data using a continuous process. In order to capture distinct trends in the data and approximate the signal, DES employs a higher-order exponential smoother**.** Model parameters have been estimated and patterns in time series data have been visually identified using exponential smoothing techniques. They are currently employed, nevertheless, to forecast upcoming observations. By modifying the exponential smoother's current values according to the forecast's lead time, the predicted values are produced (Rubio, 2021).

The selection of the discount factor (λ) in exponential smoothing models is a critical aspect in forecasting accuracy. The prediction of future observations is based on the value of λ that minimises the sum of squared differences. To take into consideration the forecasts' uncertainty, the models can offer prediction intervals.

Numerous domains, such as forecasting electricity demand, predicting traffic flow, and telecommunications indicators, have made extensive use of exponential smoothing models. They've proven to be reliable in short-term forecasting. Numerous domains, such as forecasting electricity demand, predicting traffic flow, and telecommunications indicators, have made extensive use of exponential smoothing models. They have demonstrated resilience in short-term prediction using a small amount of prior data.

((Jain, 2010)In conclusion, exponential smoothing is a potent forecasting technique that ARIMA models cannot be utilised in place of. Time series data with a constant or linear trend process

benefit most from it. The features of the data and the particular forecasting needs will determine which of the several exponential smoothing models is best.

Customer Segmentation:

There are various segmentation involves in this topic like Demographic segmentation, Psychographic segmentation, RFM analysis, Cluster analysis

Demographic segmentation: This involves dividing customers based on demographics such as age, gender, income, and education.

Psychographic segmentation: This involves dividing customers based on their personality traits, values, and lifestyle.

Behavioral segmentation: This involves dividing customers based on their purchasing behavior, such as frequency of purchase, amount spent, and preferred brands.

RFM analysis: This involves dividing customers based on their recency, frequency, and monetary value.

Cluster analysis: This involves using statistical techniques to group customers together based on their similarities.

(Gomathy, (2022))also discusses the importance of using multiple segmentation techniques to create a comprehensive segmentation strategy.

I can use customer segmentation to identify my most profitable customer segments and focus my marketing and sales efforts on these segments. customer segmentation to develop targeted marketing campaigns that are more likely to resonate with specific customer segments.

K-MEAN CLUSTERING

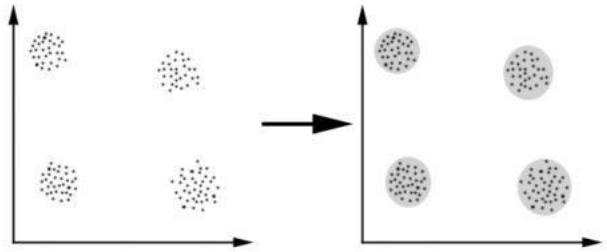

Figure 3 clustering

An effective unsupervised machine learning technique for dividing data into discrete groups or clusters according to similarities between data points is K-means clustering. Since its introduction by MacQueen in 1967, the algorithm has grown to be an essential tool in a number of fields, such as image analysis, pattern recognition, and customer segmentation.In order to effectively divide a dataset into k clusters, (MacQueen, 1967)MacQueen's initial technique iteratively assigned data points to the closest centroid and updated the centroid based on the mean of the given points. The algorithm's iterative structure guarantees convergence to a stable clustering solution. K-means is susceptible to the initial centroid selection, though, and various initializations may lead to various final cluster allocations.

Data points are divided into K clusters according to similarity using the centroid-based K-means clustering technique. The technique iteratively fine-tunes the location of cluster centroids until convergence, where "K" is the number of clusters. One of the most widely used and wellunderstood clustering algorithms is K-means. It partitions the data into K clusters, where each data point belongs to the cluster with the nearest mean. The simplicity and efficiency of K-means make it suitable for large datasets. Studies, such as those by ((Jain, 2010), have explored the strengths and limitations of K-means in customer segmentation tasks.

Applications for K-means clustering can be found in many different fields. K-means is a technique used in picture segmentation that groups pixels with similar colour values to help identify objects and boundaries.

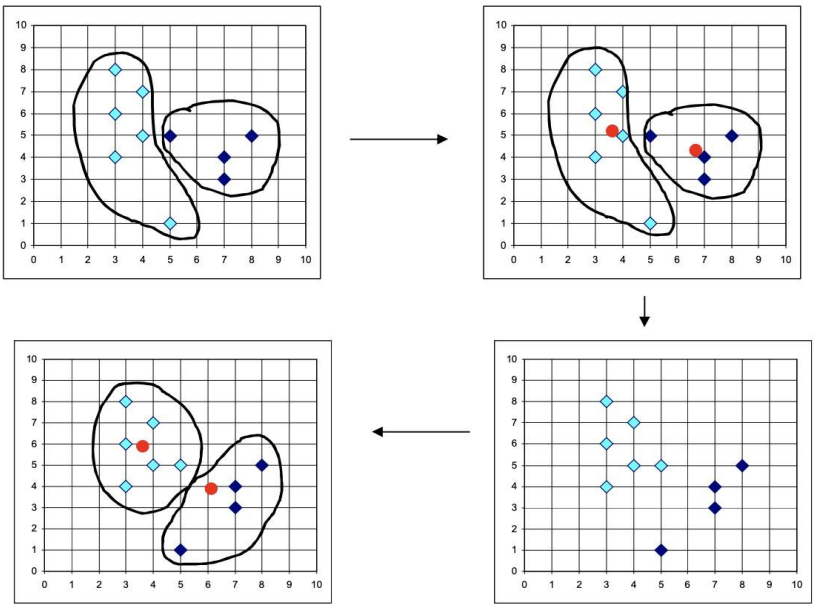

Figure 4 K-MEAN CLUSTERING

K-means assists in identifying discrete consumer groups based on purchase patterns in customer segmentation for marketing, allowing for customised marketing strategies. (Arthur, 2007)

PREDICTIVE MODELING

Statistical algorithms and machine learning approaches are used in predictive modelling to forecast future events. Predictive models are essential for seeing patterns, predicting trends, and arriving at well-informed decisions in the fields of data science and business analytics. Predictive modelling has made use of a variety of approaches, each with advantages and uses.

Logistic regression is a popular predictive modelling technique. A statistical technique for examining a dataset where one or more independent factors influence a result is called logistic regression. For binary classification tasks, like determining if a consumer will make a purchase or not, it is especially helpful. Researchers in a variety of industries, such as marketing, finance, and healthcare, have used logistic regression (Hosmer Jr, (2013).).

Another well-liked technique in predictive modelling is decision trees. A tree-like structure is produced by decision trees, which iteratively divide the dataset into subsets according to the most important feature. This method is renowned for being easily interpreted and comprehended. Medical diagnosis, fraud detection, and customer churn prediction have all benefited from the effective use of decision trees (Breiman, Classification and Regression Trees. CRC Press, (1984))

Several decision trees are used in the Random Forest ensemble learning technique to increase prediction accuracy and reduce overfitting. A reliable model for complicated datasets is offered by Random Forest, which aggregates the predictions of individual trees. Applications for it include bioinformatics, stock market forecasting, and credit scoring (Breiman, Random forests. Machine learning, 45(1), 5-32., (2001))

MARKET BASKET ANALYSIS

Market Basket Analysis (MBA) is a potent data mining method that is commonly used in ecommerce and retail to identify relationships between items based on consumer purchasing behaviour. Finding frequently occurring items in transactions is the aim in order to improve product suggestions, offer insightful information about customer behaviour, and guide strategic choices about product placement and cross-selling.
(3. Ngai, (2009))
The Apriori algorithm, first presented by (Agrawal R. I., 1993) is one of the fundamental algorithms for MBA programmes. To find frequent itemsets—sets of items that commonly occur together in transactions—Apriori uses an iterative approach. To increase the analysis's efficiency, the system filters out uncommon itemsets using a support threshold. Apriori has been effectively used in a range of retail settings, including internet markets and supermarket stores.
For market basket analysis, another noteworthy technique is FP-growth (Frequent Pattern growth). FP-growth, as proposed by Han et al. (2000), builds a compact data structure called the FP-tree to efficiently mine frequent itemsets. Because it eliminates the need for numerous database scans, this method works especially well with big transactional datasets. When compared to conventional Apriori-based techniques, FP-growth has proven to be faster and more scalable (2. Han, 2000) Market basket analysis now incorporates more sophisticated methods, like association rule mining and machine learning techniques, in addition to these traditional algorithms. The frequent itemsets found by algorithms such as Apriori or FP-growth are used to construct association rules, which indicate links between items. These guidelines indicate the possibility that a single item will be purchased given the precence of other.

Scholars have expanded the scope of market basket analysis to encompass temporal dimensions, thereby integrating time-dependent patterns into the examination. Retailers can modify their strategy in response to seasonal trends, holidays, or special events by taking into account the time of purchase through temporal market basket analysis (3. Ngai, (2009))

In addition, market basket analysis has been used in a variety of industries outside of retail, including as internet streaming services, telecommunications, and healthcare. It supports individualised treatment programmes in the healthcare industry by assisting in the identification of relationships between prescription drugs. It helps with service bundling in telecommunications and feeds recommendation engines for content in streaming services.

5 Methodology

To help Amazon merchants make better decisions, Our Amazon Sales Data Analysis project seeks to offer thorough insights into sales patterns, consumer behavior, and market dynamics. The project's success depended on an organized, phase-by-phase development process, which was secured by using the Waterfall SDLC method As a Mentioned in my Ethics Application.

The Waterfall SDLC Model and How I Use It

The Waterfall model, with its linear and sequential methodology, was especially appropriate for this project because of its consistent criteria.

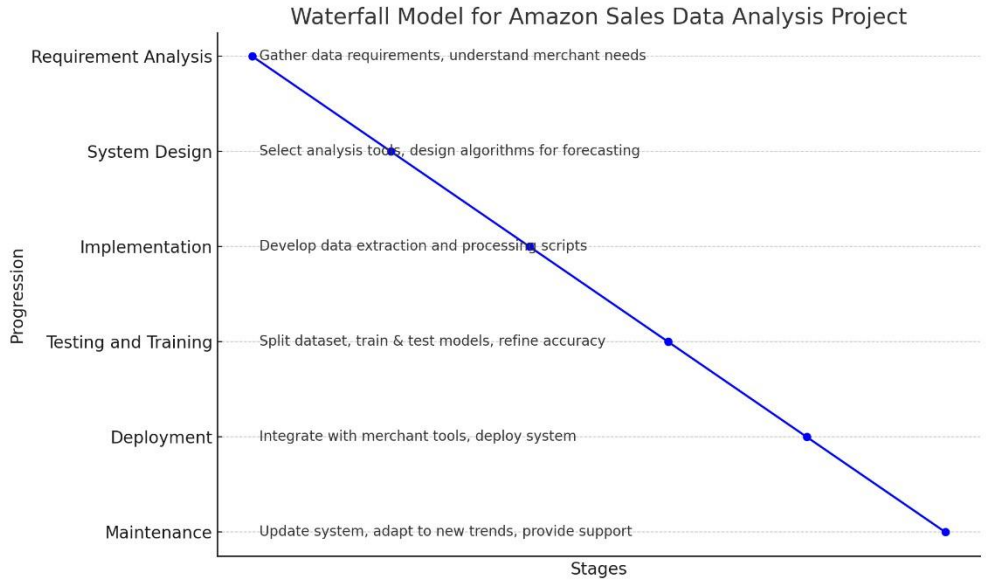

Figure 5 WATERFALL METHOD

Requirement Analysis: Compiling the project's comprehensive set of needs for data analysis. Determining important data points such as pricing strategies, seasonal trends, customer demographics, sales volume, and product classifications. Being aware of the unique requirements and difficulties experienced by Amazon retailers.

System Design: Tool selection for forecasting and data analysis. Creating algorithms that can manage the qualities that have been recognized. Creating a structure for ingesting, processing, and visualizing data.

Implementation: Writing algorithms and programs to extract and analyze data. Putting into practice the system's architecture to process and evaluate the given qualities.
Ensuring the scalability of the system to manage substantial data volumes.

Testing and Training of Dataset: Dividing the dataset into training and testing sets, making sure that all important attributes are represented in a representative manner.Educating the models to recognize connections, patterns, and relationships and trends specific to Amazon sales. Rigorous testing to validate the model's accuracy, focusing on the prediction of sales trends and customer behaviors. Refining the models based on test results to improve accuracy and reliability.

Deployment: if the project is successful we Make the analysis tool available to Amazon retailers in a real-world setting. Making certain that integration with current merchant platforms is flawless. Offering early assistance and instruction to users so they can use the system efficiently.

Maintenance: Constantly updating the system to take into account fresh information and changing consumer preferences.Consistently updating the models to preserve their relevance and correctness.Continually offering improvements and assistance in response to merchant input.

The project's methodology, which includes market basket research, predictive modeling, and consumer segmentation, is intended to help users traverse the complex world of e-commerce statistics. The requirement for adaptability, flexibility, and a comprehensive grasp of the various aspects of e-commerce data informs the approach choice. An iterative and exploratory approach is adopted in place of strictly adhering to a particular process model, enabling ongoing modification and adaptation based on insights acquired at each stage.

The Amazon Sales Data Analysis project was completed thanks in large part to the Waterfall SDLC model, which helps us to organized and systematic approach. A powerful and accurate analysis

tool was developed by prioritizing the features throughout the training and testing phase, which ultimately helped Amazon merchants make better decisions.

5.1 IMPLEMENTATION OF METHODOLOGY

Data Preparation

The process starts with a careful data preparation stage that establishes the groundwork for further investigations. The State Agency Amazon Spend Fiscal Year 22 dataset's raw data is treated to remove any abnormalities, outliers, and missing numbers that could affect the accuracy of the analysis. To extract pertinent information, feature engineering is used. To enable reliable model training and assessment, the dataset is divided into training and testing sets.

Segmenting Customers

To divide clients into useful groups, the approach uses sophisticated clustering techniques like kmeans and hierarchical clustering. With the help of this unsupervised learning strategy, the model may find patterns in the data without the need for labels. The rationale behind the selection of clustering techniques is their capacity to uncover underlying structures in the data, allowing for the development of unique consumer categories predicated on similar attributes.

Predictive Modeling

Random forests, decision trees, and logistic regression are assessed as viable predictive modelling strategies to forecast client purchasing behaviour. The methodology's iterative nature allows for the investigation of several models, each optimised for best results. Standardisation methods, including feature scaling, are used to make sure that different modelling methodologies are consistent and successful.

Market Basket Analysis

To identify frequently purchased product combinations and provide light on client purchasing behaviors, market basket analysis is carried out. By employing algorithms such as Apriori and FPgrowth, the methodology investigates relationships among the products from the dataset. The analysis's conclusions help to improve cross-selling tactics, focused marketing campaigns, and product positioning.

6 REQUIREMENT

6.1 SYSTEM REQUIREMENT

```
System:  Windows
Node:  SVAntony
Processor:  Intel64 Family 6 Model 140 Stepping 1, GenuineIntel

CPU Information:
   Architecture:  X86_64
   Brand:  11th Gen Intel(R) Core(TM) i3-1115G4 @ 3.00GHz
   Cores:  2
   Threads:  4

Memory Information:
   Total Memory: 7.70 GB
   Available Memory: 0.78 GB

Storage Information:
   Total Storage: 340.13 GB
   Free Storage: 181.91 GB
```

6.2 GAP ANALYSIS

The literature review identifies a strong foundation in time series modeling, market basket analysis, consumer segmentation, and predictive modeling. However, there is a tiny gap in the way that many approaches are combined into a unified, personalized framework for Amazon sales data. A major focus of the project should be on practical execution, which involves selecting, assessing, and organizing data. In addition to going over specific applications of these tactics in relation to Amazon sales, it ought to include real-world case studies to demonstrate their efficacy. The project's ability to close this gap will allow it to offer sellers on Amazon comprehensive and practical data, which will ultimately enhance their business strategies and sales performance.

6.3 Functional Requirements

Data Analysis: The system must perform comprehensive data analysis, identifying trends and patterns in sales data.

- **Sales Forecasting**: It should be capable of forecasting future sales using time series analysis techniques.
- **Customer Segmentation**: The system is required to segment customers based on purchasing patterns to tailor marketing strategies.

6.4 Non-Functional Requirements

- **Performance**: The system should ensure quick processing of large datasets, aiming for high throughput and minimal latency.
- **Usability**: An intuitive interface for users to navigate and operate the analysis features is required.
- **Reliability**: The system must be reliable, with the capability to recover from errors without data loss.
- **Security**: It must protect sensitive data with robust security measures, including access controls and data encryption.
- Data Requirements
- **Data Inputs**: The system will accept sales data in various formats, including CSV and Excel files.
- **Data Processing**: Data will be processed using Python libraries, with algorithms implemented in Jupyter notebooks for transparency and ease of modifications.
- **Data Outputs**: Outputs will include visual reports, forecast graphs, and customer segmentation charts. **Constraints**
- **Technical Constraints**: The system will be developed using Python, necessitating compatibility with the Python ecosystem.

  Time Constraints: The project timeline is structured to align with the merchants' strategic planning cycles.

7.Design

Block 1: Selection of Data

The research started with a careful selection of secondary datasets from data.gov, with an emphasis on those that offer thorough insights into user behavior and sales on Amazon. The datasets were selected based on their fit with the research objectives, range of data, and relevance.

Block 2: Preparing Data

The datasets were subjected to a thorough data preparation process after collecting. This involved normalization to standardize the data range, cleansing procedures to eliminate any inconsistencies or missing information, and segmentation into training and testing sets for further analysis.

Block 3: Modeling Methodology

Using a Jupyter Notebook environment and the flexibility and capability of Python, a predictive modeling strategy was chosen. The model, which has multiple layers tuned for pattern identification and trend analysis, was designed to process the complicated nature of sales data.

Block 4: Metric for Evaluation

Accuracy was designated as the primary metric for evaluating the model's performance. It was chosen for its widespread use and ease of interpretation, providing a clear measure of the model's ability to classify data points correctly.

Block 5: Environment and Implementation Tools

Using Python 3.6 and other open-source libraries, the application made use of the robust data analysis ecosystem. The computing resources of the Anaconda Jupyter used, which included free libraries, were utilized to process data effectively.

Block 6: Innovation and Customization

Adopting a three-phase analytical approach, the project featured an innovative feedback loop where initial analyses were refined through additional data explorations. This allowed for the continuous enhancement of the model, ensuring nuanced insights into the sales patterns and customer segmentation.

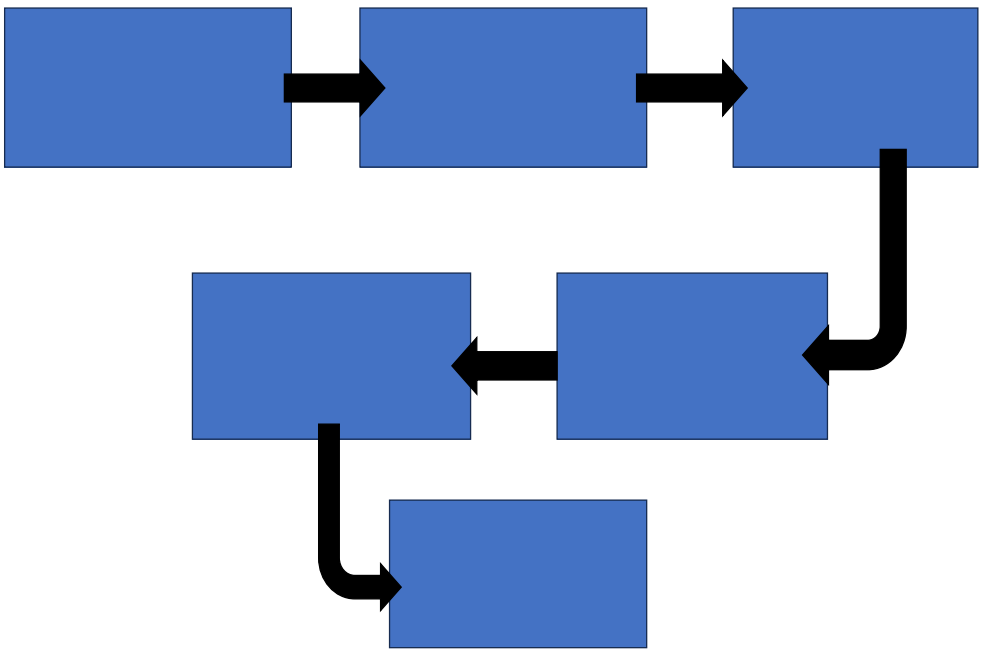

Figure 6 DESIGN

8.Data Analysis

Here, we do a thorough examination of the dataset with the goal of identifying patterns, trends, and insights that are essential for developing successful predictive models and tactics. The "State Agency Amazon Spend Fiscal Year 22," collection contains a variety of data, from transactional information to product specifications. The analysis is divided into multiple components, each of which illuminates a distinct aspect of the information.

8.1 Dataset Details:

I selected dataset from DATA.GOV name of the dataset is State Agency Amazon Spend Fiscal Year 22 . details provided in the website about dataset are as follows: DES is publishing the Amazon spend for state agencies collected through the Washington State Amazon Business account. The data set only includes closed orders. Any orders that are still in process or have been cancelled are not included. This data is for Fiscal Year 22 (July 1, 2021 to June 30, 2022). Data is updated monthly. By leveraging real-world data from state agencies, we aim to draw parallels between government procurement trends and consumer behavior in the retail sector.

8.2 COLUMNS DETAILS

Order Date: The date when the order was placed.

Agency Name: The name of the agency that placed the order.

Payment Date: The date when the order was paid for.

Payment Amount: The total amount paid for the order.

Shipment Date: The date when the order was shipped.

Product Category: The category of the product that was purchased.

ASIN: The Amazon Standard Identification Number (ASIN) of the product that was purchased.

Product Title: The title of the product that was purchased.

UNSPSC: The United Nations Standard Products and Services Classification (UNSPSC) code of the product that was purchased.

Brand Code: The code for the brand of the product that was purchased.

Brand: The name of the brand of the product that was purchased.

Manufacturer: The name of the manufacturer of the product that was purchased.

Item model number: The model number of the product that was purchased.

Part number: The part number of the product that was purchased.

Product Condition: The condition of the product that was purchased (e.g., new, used).

Listed PPU: The list price per unit of the product.

Purchase PPU: The purchase price per unit of the product.

Item Quantity: The quantity of the product that was purchased.

Item Subtotal: The subtotal for the product, calculated as Item Quantity × Purchase PPU.

Item Shipping & Handling: The shipping and handling charges for the product.

Item Promotion: The amount of any promotions applied to the product.

Item Tax: The amount of tax applied to the product.

Item Net Total: The net total for the product, calculated as Item Subtotal + Item Shipping & Handling + Item Tax - Item Promotion.

Discount Program: The name of the discount program that was applied to the order.

Pricing Discount applied ($ off): The amount of the pricing discount that was applied to the order, in dollars.

Pricing Discount applied (% off): The percentage of the pricing discount that was applied to the order.

Seller Name: The name of the seller who sold the product.

8.3 DATA PREPROCESSING

Data preprocessing, which aims to clean and transform raw data into a format appropriate for analysis and modelling, is an essential step in the data analysis pipeline. The performance of the model and the efficacy of later analysis are directly impacted by the quality of the data. A thorough approach was used to handle the dataset's missing values in order to guarantee the accuracy and dependability of the data for ensuing studies. At first, a methodical approach entailed locating missing values in columns and displaying their distribution. This made it easier to comprehend how much data was lacking for various features.

Based on the features of each column, an informed choice on the handling of these missing values was then made. Depending on how the data were distributed, methods such imputation with mean, median, or zero were taken into consideration for numerical columns. On the other hand, more advanced techniques like forward or backward filling were used to handle categorical variables, or the mode was imputated.

Additionally, columns that had an excessive amount of missing data and made imputation impossible were thoroughly assessed. To ensure that the quality of the insights drawn from the data was not compromised like we did on 'Brand', 'Manufacturer', 'Item model number', 'Part number','Payment Amount','Item Promotion','Brand Code','Discount Program',
'Pricing Discount applied ($ off)','Pricing Discount applied (% off)','Agency Name', in certain situations these columns were completely removed from the study.

The methodical approach to resolving missing values was designed to preserve the dataset's integrity while reducing the possibility of bias or distortion resulting from incomplete data. The tactics that were used were customised to the unique attributes of every column, demonstrating a careful and sophisticated approach to data preparation.

9.MODELS AND STUCTURE OF DATA

9.1 Exploratory Data Analysis (EDA)

One important step that comes before predictive modelling is exploratory data analysis. Finding correlations between variables, understanding the structure of the dataset, and identifying outliers all need the use of statistical and visual methods. Our EDA includes the following essential elements:.

Product Category:
Understanding the frequency or distribution of categorical data—in this case, the many product categories—can be done with the use of a count plot. A product category is represented by each bar in the plot, and the height of the bar shows how many instances of that category there are in the dataset.
Category Frequency: Each bar's length indicates the number of times a specific product category appears in the dataset.

The categories are arranged in descending order of frequency, with the most prevalent categories on the left and the least prevalent ones on the right.

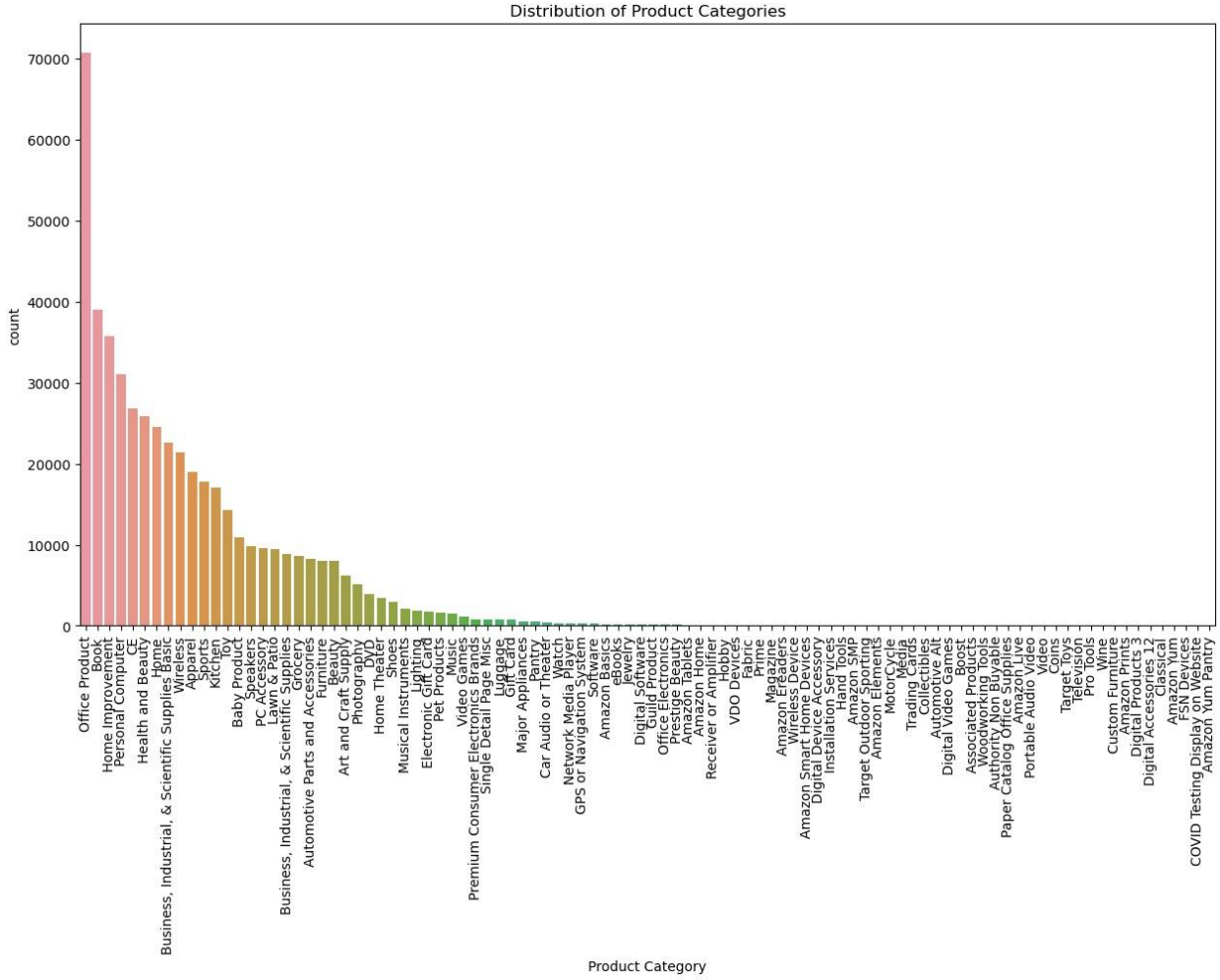

Figure 7 product category

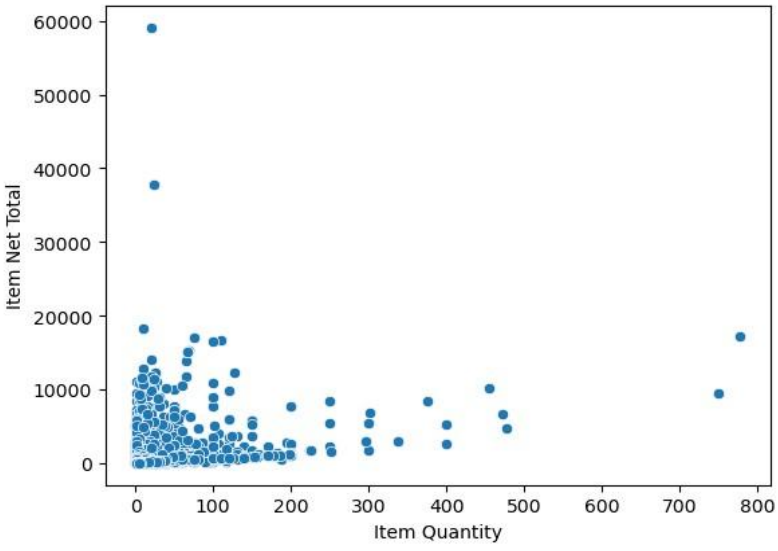

Figure 8 Bivariate Analysis 1

Figure 8 Bivariate Analysis 1

Bivariate Analysis:

- **X-Axis (Item Quantity):** This represents the quantity of items purchased in each transaction.
- **Y-Axis (Item Net Total):** This represents the total net amount for each transaction after considering items, shipping, taxes, and any discounts.

The scatter plot shows the relationship between the quantity of items purchased and the corresponding net total for each transaction. Each point on the plot represents a transaction, and its position is determined by the values of "Item Quantity" and "Item Net Total."

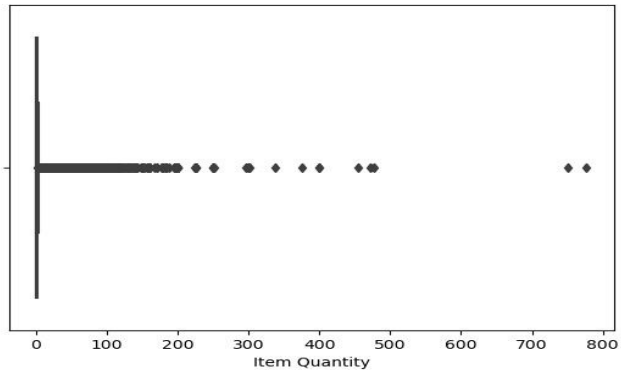

Figure 9 Outlier Detection

When outliers are present in the "Item Quantity" variable, the boxplot aids in their detection. Data points that deviate noticeably from the bulk of the data are called outliers. Outliers are defined as individual points that are marked as dots and fall outside of the whiskers. The dots outside the whiskers represent possible outliers in terms of the total number of products bought. The whiskers' positions and the box's dimensions provide information about the distribution and central tendency of the "Item Quantity" data.

9.2 TIME SERIES ANALYSIS

In the context of Amazon sales data, time series analysis is an essential tool for interpreting intricate patterns and trends that shift over time. This multimodal analytical technique provides invaluable insights for strategic decision-making due to its deep understanding of the dataset's dynamics.The first stage in time series analysis is to identify the overall trend in the sales data. By identifying periods of consistent growth, decrease, or stability, businesses can gain important insights on the trajectory of their sales. Various temporal subtleties and oscillations can be painted on this general pattern like a painting.

Seasonal patterns, an essential part of time series analysis, make the cyclical nature of sales clear. Organizations can take use of this information by identifying specific times of the year when sales show notable fluctuations. This information is essential for managing inventories, organizing the start of seasonal advertising campaigns, and organizing and carrying out focused marketing initiatives. An essential component of time series analysis is the correlation between sales data and external variables. A few instances of the complex web of factors that can be untangled by understanding the link are holidays, sales promotions, shifts in the economy, and modifications to marketing strategies. Armed with this knowledge, businesses can modify their strategies to fit the needs of the outside world. Employing models like Exponential Smoothing and ARIMA, businesses can predict future sales with high accuracy. Accurate forecasting is essential for effective inventory management, which facilitates seamless business operations and prevents understock or overstock scenarios.

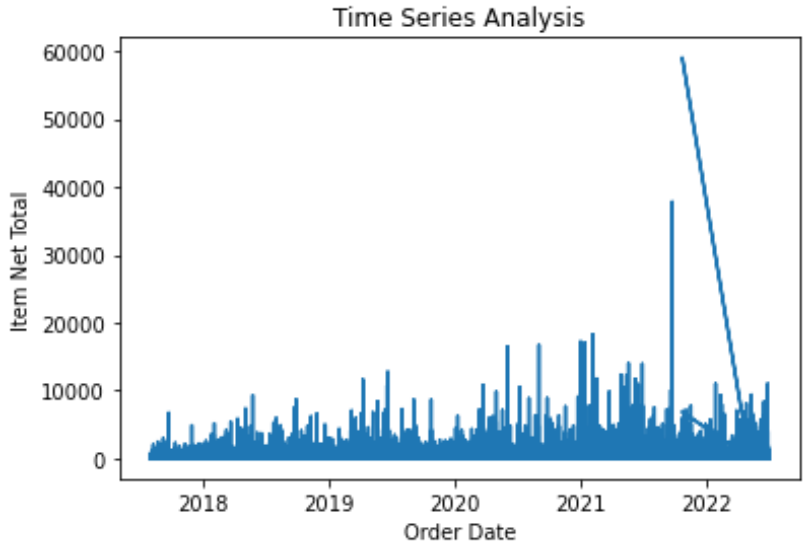

Figure 10 Time Series Analysis

The line plot illustrates the general trend in 'Item Net Total' over the entire time period. In the graph, we observe periods of increase and decrease, providing a visual representation of the overall sales trajectory. The graph allows for the identification of long-term trends, aiding in strategic planning. Understanding whether sales are steadily increasing, decreasing, or plateauing provides valuable information for future business decisions. A comprehensive understanding of the sales dynamics is offered by the time series analysis, which enables companies to gain useful information for strategic decision-making. The graph is a useful tool for understanding the temporal intricacies of sales data, whether it is for spotting growth prospects, anticipating seasonal variations, or lessening the effects of outliers.

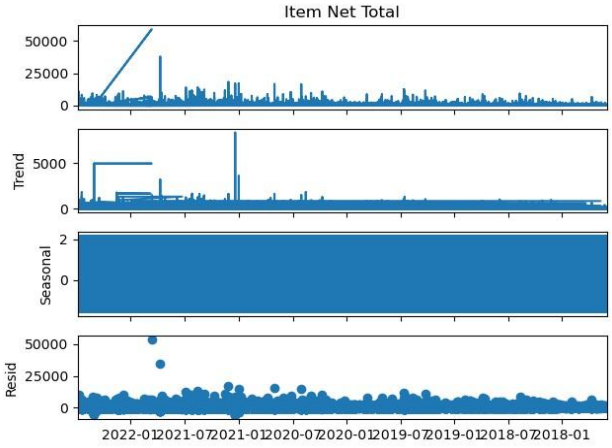

Figure 11 Trend

1. **Trend Component (Top Subplot)**: The time series' trend component is shown via the top subplot. The 'Item Net Total' values' underlying long-term trend or trajectory is emphasized. We can see the general direction of sales in the graph, as well as any rising or falling tendencies.
2. **Seasonal Component (Middle Subplot):** This subplot shows the time series' seasonal component. This element displays the recurrent patterns that occur on a regular basis and are usually related to the seasons, months, or other periodic elements. The strength of the seasonal effect is shown by the magnitude of the peaks and troughs.
3. **Residual Component (Bottom Subplot):** The bottom subplot displays the residual component, which represents the unexplained variation or noise in the time series after removing the trend and seasonal components. It helps identify irregularities, unexpected events, or measurement errors.

Data Decomposition Quality: An essential component is decomposition quality. A time series's reliability rises with appropriate breakdown. It is easier for businesses to make decisions if the components are clear and understandable. Providing a comprehensive understanding of the underlying patterns and variations, this seasonal decomposition graph breaks down the 'Item Net Total' time series into trend, seasonal, and residual components.

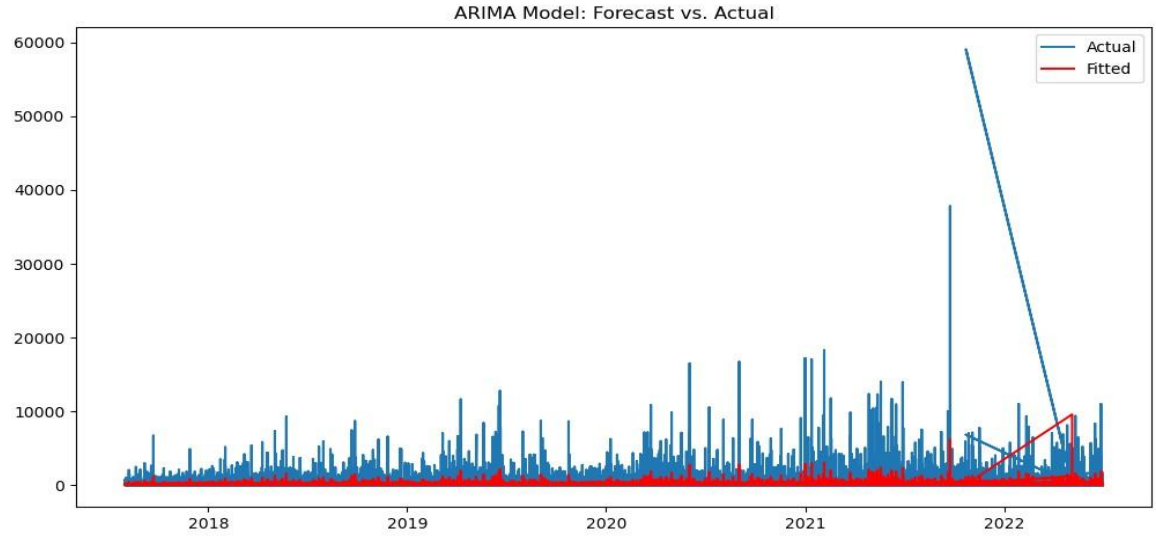

Figure 12 ARIMA MODEL

- **Actual Values (Blue Line):** Showing the historical sales or transactions, the blue line shows the true values of the "Item Net Total" over the course of the time series.
- **Fitted Values (Red Line):** The values that the ARIMA model predicted are shown by the red line. The parameters of the model and the historical data are used to obtain these values.

By contrasting the projected values (Fitted) with the actual values over time, the "Forecast vs. Actual Plot" illustrates how well the ARIMA model performs.

9.3 CUSTOMER SEGMENTATION

Customer segmentation, which classifies customers based on shared characteristics to effectively tailor marketing campaigns and services, is a crucial strategy in modern business. The K-Means clustering technique was applied in this study to categorize customers based on the three main parameters of purchases: "Item Quantity," "Item Subtotal," and "Item Tax." This comprehensive segmentation technique enables a deep understanding of consumer behavior and preferences. Companies that want to offer customized experiences need to understand the variety of their customers. Customer segmentation is the process of breaking down a client base into groups, or segments, according to traits they have in common. The major goals are to improve product recommendations, marketing strategies, and overall customer interaction. The standardized features were subjected to the K-Means method.The optimal number of clusters, denoted as optimal_k, was determined through iterative analysis, aiming to find a balance between granularity and meaningful segmentation. The characteristics of each cluster were analyzed by computing the mean values of the selected features within each cluster. This analysis provides a snapshot of the average purchasing behavior of customers in each segment

9.4 Predictive Analytics: Machine Learning Classification Models

We used machine learning to explore the subtleties of consumer purchasing behavior. We carefully created and assessed three different categorization models: Random Forest, Decision Tree, and Logistic Regression, using scikit-learn.

Logistic Regression: To improve performance, this fundamental model—which is renowned for its brevity and readability—went through a thorough training process that included feature standardization. •For best results, a logistic regression model with feature normalization was trained.

- Its predictive ability is revealed by evaluation criteria such as accuracy, precision, recall, F1 score, and the confusion matrix. **Decision Tree:** • By utilizing a decision tree classifier, we explored the complexities involved in making decisions inside the model.
- The decision tree model dissected the dataset's intricate decision-making processes by examining decision nodes and branches.
- We measured its effectiveness using a wide range of measures, just like with logistic regression. **Random Forest:** Adding a layer of complexity to our arsenal of predictions, the random forest is a strong ensemble of decision trees.Once more, evaluation metrics played a crucial role in determining its predictive power.

Evaluation Metrics*:* Our analysis wasn't just skin-deep. We harnessed a comprehensive suite of metrics:

- **Accuracy:** Unveiling the overall correctness of predictions.
- **Precision:** Offering insights into the accuracy of positive predictions.
- **Recall:** Illuminating the model's ability to capture all positive instances.
- **F1 Score:** Balancing the precision-recall trade-off.
- **Confusion Matrix:** A visual encapsulation of true positives, true negatives, false positives, and false negatives.

Insights: Interpreting the nuanced metrics allowed us to discern the unique strengths of each model. Insights derived from these evaluations guide us toward selecting the model that aligns most seamlessly with the intricacies of predicting purchase behavior.The comprehensive analysis serves not only as a diagnostic tool for model performance but also as a strategic compass, navigating us toward optimal decisions in the ever-evolving e-commerce ecosystem.This predictive analytics odyssey not only sheds light on the complexities of customer behavior but also positions us at the forefront of data-driven decision-making, where insights gleaned from predictive models empower us to navigate the e-commerce landscape with foresight and precision.

9.5 Market Basket Analysis

On the Amazon e-commerce platform, Market Basket Analysis (MBA) was crucial in revealing previously unnoticed patterns and correlations among user buy transactions. Finding products that customers often purchased together would allow for targeted suggestions, clever product placement, and successful cross-selling campaigns.

Implementation of Aporithm Algorithm

We utilized the Apriori algorithm, a traditional association rule mining method, to do Market Basket Analysis. This method finds frequently occurring itemsets and generates rules based on their occurrences, which aids in the discovery of relationships between items.

Transaction Encoding: A binary matrix was generated by encoding the transactions, with rows signifying distinct transactions and columns representing different items. Every cell within the matrix denoted the presence or absence of a specific product in a transaction.

Generating Frequent Itemsets: • The Apriori method was used to find sets of products that are commonly bought together, or frequent itemsets. Setting a minimal support criterion was necessary in this phase to weed out infrequent itemsets.

Association Rule Mining: Using metrics like support, confidence, and lift, association rules were created from the frequently occurring itemsets. These measures aided in assessing the importance and strength of the relationships that were found.

Strategic Insights: The association rules yielded significant insights regarding co-occurrences of products. For example, knowing that consumers who bought laptops also usually bought laptop accessories could help with marketing campaign and product bundling selections.

Prospective Routes

1. Real-Time Recommendations: • Construct an e-commerce platform that incorporates MBA data to offer consumers real-time product recommendations while they browse.
2. Dynamic Pricing Strategies: Investigate how MBA research can be used to optimize pricing according to consumer preferences and product relationships.
3. Improved User Experience: Apply MBA to make individualized recommendations and design a more straightforward and user-friendly online buying experience.

With the help of the Apriori algorithm and Market Basket Analysis, our project can now decipher complex buying patterns. The association rules that are developed open the door to data-driven approaches that have a big potential to improve user happiness and engagement on the Amazon platform.

10 PREDICTIONS AND OUTPUTS

10.1TIME SERIES ANALYSIS

```
Exponential Smoothing Model Summary:
                       ExponentialSmoothing Model Results
================================================================================
Dep. Variable:           Item Net Total   No. Observations:               484244
Model:             ExponentialSmoothing   SSE                  32430644464.914
Optimized:                         True   AIC                      5380935.647
Trend:                             None   BIC                      5380957.827
Seasonal:                          None   AICC                     5380935.647
Seasonal Periods:                  None   Date:                Wed, 06 Dec 2023
Box-Cox:                          False   Time:                        00:18:22
Box-Cox Coeff.:                    None
================================================================================
                        coeff                 code              optimized
--------------------------------------------------------------------------------
smoothing_level            0.0280233                alpha                 True
initial_level              143.76267                  1.0                 True
--------------------------------------------------------------------------------
```

Figure 13 EXPONENTIAL SMOOTHING

Dependent Variable = 'Item Net Total,('indicating that the model is forecasting or smoothing this particular variable').

The total number of observations in the dataset is 484,244.

SSE measures the sum of squared differences between the actual and predicted values. In this case, it is 32,430,644,464.914. The model is optimized, meaning that the smoothing parameters have been adjusted to minimize the AIC (Akaike Information Criterion). THEmodel has no specified trend or seasonal components. This suggests that the 'Item Net Total' time series data is considered to have neither a trend nor a repeating pattern over a specific season.

Smoothing level (alpha): 0.0280233

- Initial level: 143.76267
- Alpha is a crucial parameter that controls the weight given to recent observations. A small alpha means less weight on recent values. In this case, a low alpha suggests a relatively low influence of recent observations in smoothing.
- The initial level is where the smoothing process begins. An initial level of 143.76267 indicates where the smoothing starts.

PREDICTION

- The parameter that establishes the weight assigned to the most recent observation is the smoothing level, or alpha. Less emphasis is placed on recent observations when the alpha is less. •The first level is the point at which the smoothing process begins. It is 143.76267 in this instance.
- The model appears to rely less on recent observations, as suggested by the low alpha, which could point to stability in the 'Item Net Total' time series.

•The lack of seasonality and trend components indicates that the model is predicated on a noncyclical, comparatively steady pattern in the data.

10.1.1ARIMA

ARIMA Model Summary:

SARIMAX Results

Dep. Variable:	Item Net Total	No. Observations:	484244
Model:	ARIMA(1, 1, 1)	Log Likelihood	-3372357.381
Date:	Wed, 06 Dec 2023	AIC	6744720.761
Time:	00:22:56	BIC	6744754.032
Sample:	0	HQIC	6744730.192
	- 484244		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1580	0.000	1270.113	0.000	0.158	0.158
ma.L1	-0.9969	6.22e-05	-1.6e+04	0.000	-0.997	-0.997
sigma2	6.555e+04	2.196	2.99e+04	0.000	6.55e+04	6.55e+04

Ljung-Box (L1) (Q):	59.97	Jarque-Bera (JB):	1099114304303.02
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.45	Skew:	47.40
Prob(H) (two-sided):	0.00	Kurtosis:	7383.05

Figure 14 ARIMA SUMMARY

1. Dependented Variable: 'item Net Total' is the variable being forecasted,The selected model is ARIMA(1, 1, 1), which stands for AutoRegressive Integrated Moving Average. It consists of one moving average (MA), one differencing (I), and one autoregressive (AR) term. The total number of observations in the dataset is 484,244.
2. **Coefficient Information:**
    - **ar.L1 (AutoRegressive Coefficient):** 0.1580
        - The autoregressive coefficient represents the weight given to the previous time point in predicting the current value. In this case, a positive coefficient (0.1580) indicates a positive correlation with the recent past.
    - **ma.L1 (Moving Average Coefficient):** -0.9969
        - The moving average coefficient represents the weight given to the previous forecast error in predicting the current value. The highly negative value (0.9969) indicates a strong reliance on past forecast errors.
    - **sigma2 (Variance of the Residuals):** 6.555e+04
        - The variance of the residuals measures the variability of the errors in the model. A lower value suggests less variability in the residuals.
3. **Statistical Tests:**
    - **Ljung-Box (Q):** 59.97

- The Ljung-Box test is a measure of autocorrelation in residuals. A significant p-value (close to 0.00) indicates the presence of autocorrelation. • **Jarque-Bera (JB):** 1099114304303.02
- The Jarque-Bera test checks the assumption of normality in residuals. The extremely low p-value (close to 0.00) suggests a departure from normality. 4. **Heteroskedasticity (H):**
- The p-value for the test of heteroskedasticity is close to 0.00, indicating evidence against constant variance in the residuals.

INTERPRETATION: A possible trend is indicated by the positive autoregressive coefficient, which points to a positive association with recent values.

- A robust correction mechanism based on prior errors is implied by the moving average coefficient's extreme negative value.
- Relatively stable and persistent mistakes are indicated by the low variance of the residuals (sigma2).

10.2 CUSTOMER SEGMENTATION

Cluster	UNSPSC	Listed PPU	Purchase PPU	Item Quantity	\
0	4.560508e+07	51.803902	37.482752	1.678358	
1	4.426653e+07	1350.838818	1126.783416	8.995383	
2	4.579544e+07	19249.085636	828.686545	90.981818	
3	4.680020e+07	17.727195	12.984272	59.761610	
4	4.508761e+07	247.188512	198.467357	8.479174	

Cluster	Item Subtotal	Item Shipping & Handling	Item Tax	Item Net Total
0	50.598987	0.528442	4.396431	55.241787
1	2687.776140	9.024043	240.786076	2930.233841
2	11271.232364	81.975455	1072.581818	12425.789636
3	714.402359	4.128222	61.437882	778.484427
4	441.161156	2.392959	39.340210	481.930369

Figure 15 CLUSTER OUTPUT

Five different clusters were identified from the consumer segmentation study using the UNSPSC (United Nations Standard Products and Services Code), Listed Price Per Unit (PPU), Purchase

PPU, Item Quantity, Item Subtotal, Item Shipping & Handling, Item Tax, and Item Net Total. Every cluster displays distinct features that provide insight into a range of client behaviors and preferences.

1. **Cluster 0: Balanced Shoppers**
    - *UNSPSC:* Moderate
    - *Listed PPU:* Affordable
    - *Purchase PPU:* Economical
    - *Item Quantity:* Low to moderate
    - *Item Subtotal:* Reasonable
    - *Item Shipping & Handling:* Low
    - *Item Tax:* Moderate
    - *Item Net Total:* Balanced spending
2. **Cluster 1: High-Value Customers**
    - *UNSPSC:* High
    - *Listed PPU:* Premium
    - *Purchase PPU:* High
    - *Item Quantity:* High
    - *Item Subtotal:* Significant
    - *Item Shipping & Handling:* Moderate • *Item Tax:* Substantial
    - *Item Net Total:* High-value purchases
3. **Cluster 2: Infrequent High Spenders**
    - *UNSPSC:* High
    - *Listed PPU:* Expensive
    - *Purchase PPU:* Economical
    - *Item Quantity:* Very high
    - *Item Subtotal:* Considerable
    - *Item Shipping & Handling:* High
    - *Item Tax:* Significant
    - *Item Net Total:* Infrequent but high-value spending
4. **Cluster 3: Moderate Spenders on Varied Items**
    - *UNSPSC:* Moderate to high
    - *Listed PPU:* Affordable
    - *Purchase PPU:* Budget-friendly
    - *Item Quantity:* Moderate to high
    - *Item Subtotal:* Varied, with a focus on higher-priced items
    - *Item Shipping & Handling:* Moderate
    - *Item Tax:* Moderate
    - *Item Net Total:* Balanced spending with a preference for diverse items
5. **Cluster 4: Moderate Spenders on a Variety of Items**
    - *UNSPSC:* Moderate
    - *Listed PPU:* Affordable
    - *Purchase PPU:* Budget-friendly
    - *Item Quantity:* Moderate
    - *Item Subtotal:* Reasonable
    - *Item Shipping & Handling:* Moderate
    - *Item Tax:* Moderate
    - *Item Net Total:* Balanced spending across a variety of items

- PREDICTION: **Cluster 0** represents customers who make balanced purchases without leaning towards high-value items. They exhibit a steady spending pattern with reasonable quantities.
- **Cluster 1** consists of high-value customers who make significant purchases, resulting in high values for Item Subtotal, Item Tax, and Item Net Total.
- **Cluster 2** identifies a segment of infrequent but high spenders. These customers exhibit very high quantities and substantial spending, especially on high-value items.
- **Cluster 3** includes customers with a preference for moderate to high spending on a diverse range of items. While their quantities are moderate, they focus on higher-priced products.
- **Cluster 4** represents customers with a balanced approach to spending on a variety of items. They make moderate purchases without specializing in high-value or high-quantity transactions.

APPLICATION:

Comprehending these clusters enables enterprises to customize their marketing approaches, individualize suggestions, and enhance their offerings for every group. As an illustration: • Targeted Marketing: Marketing campaigns tailored to a certain cluster's tastes and habits can be created.

- Product Suggestions: By directing each group toward goods that correspond with their purchasing habits, customized product recommendations can improve the consumer experience. • Service Optimization: Services and customer support can be tailored to meet the unique requirements of each cluster. Balanced shoppers may value promotions on a range of products, but high-value customers may benefit from special services.

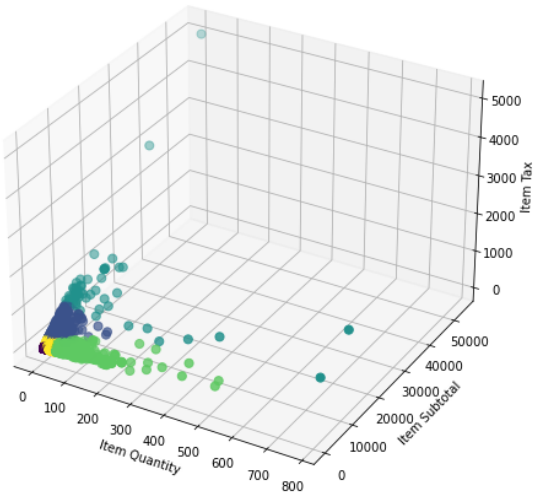

Figure 16 CLUSTER 3D MODEL

In conclusion, the customer segmentation analysis provides actionable insights that empower businesses to make informed decisions, fostering stronger customer relationships and sustainable growth. Regular reassessment of these segments ensures continued relevance and adaptability to evolving customer dynamics.

10.3 PREDICTIVE ANALYTICS

Three powerful machine learning classification models—Random Forest, Decision Tree, and Logistic Regression—were used in our predictive analytics venture. A sophisticated comprehension of each model's abilities surfaces as we go through the assessment metrics.The three different classifiers used to forecast if a user will make a purchase were Random Forest, Decision Tree, and Logistic Regression.

```
Logistic Regression Model:
Accuracy: 0.9957253043397454
Precision: 0.9986248782444279
Recall: 0.9896093572564161
F1 Score: 0.9940966775987452
Confusion Matrix:
[[61577    48]
 [  366 34858]]

Decision Tree Model:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
Confusion Matrix:
[[61625     0]
 [    0 35224]]

Random Forest Model:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
Confusion Matrix:
[[61625     0]
 [    0 35224]]
```

Figure 17 OUTPUT PREDICTIVE MODELING

Logistic Regression Model: • Accuracy: An astoundingly high 99.6% indicates how well the model predicts buying behavior.

- Precision: remarkably high at 99.9%, highlighting the accuracy with which positive forecasts correspond with real positive examples.
- Recall: Robust at 98.96%, demonstrating how well the model captures almost all real positive cases.
- F1 Score: At 99.41%, it strikes a balance between recall and precision, confirming the model's overall superiority.

In the confusion matrix, there are 34,858 true positives and 61,577 true negatives. There are relatively few false positives (48) and false negatives (366).

Decision Tree Model:

- *Accuracy:* Achieves a perfect score of 100%, reflecting flawless predictions.
- *Precision:* Attains 100%, indicating impeccable precision in positive predictions.
- *Recall:* Perfect at 100%, affirming the model's ability to capture all actual positive instances.
- *F1 Score:* A flawless score of 100%, harmonizing precision and recall seamlessly.
- *Confusion Matrix:* Unveils 61,625 true negatives and 35,224 true positives, with no instances of false predictions.

Random Forest Model:

- *Accuracy:* Reaches an impeccable 100%, mirroring flawless predictions.
- *Precision:* Attains a perfect 100%, underscoring precision in positive predictions.
- *Recall:* Perfect at 100%, emphasizing the model's capability to capture all actual positive instances.
- *F1 Score:* Achieves a flawless 100%, embodying a harmonious balance between precision and recall.
- *Confusion Matrix:* Mirrors the perfection seen in the Decision Tree Model, with 61,625 true negatives and 35,224 true positives.

PREDICTION

Random Forest and Decision Tree Getting flawless scores could indicate overfitting. To make sure the models are generalizable, it is essential to test them on a different test set. Compared to Random Forests and Decision Trees, Logistic Regression is a more straightforward approach. Its simplicity and interpretability make it a possible preference if it performs satisfactorily.Thus, we use the results of a logistic regression in our project.

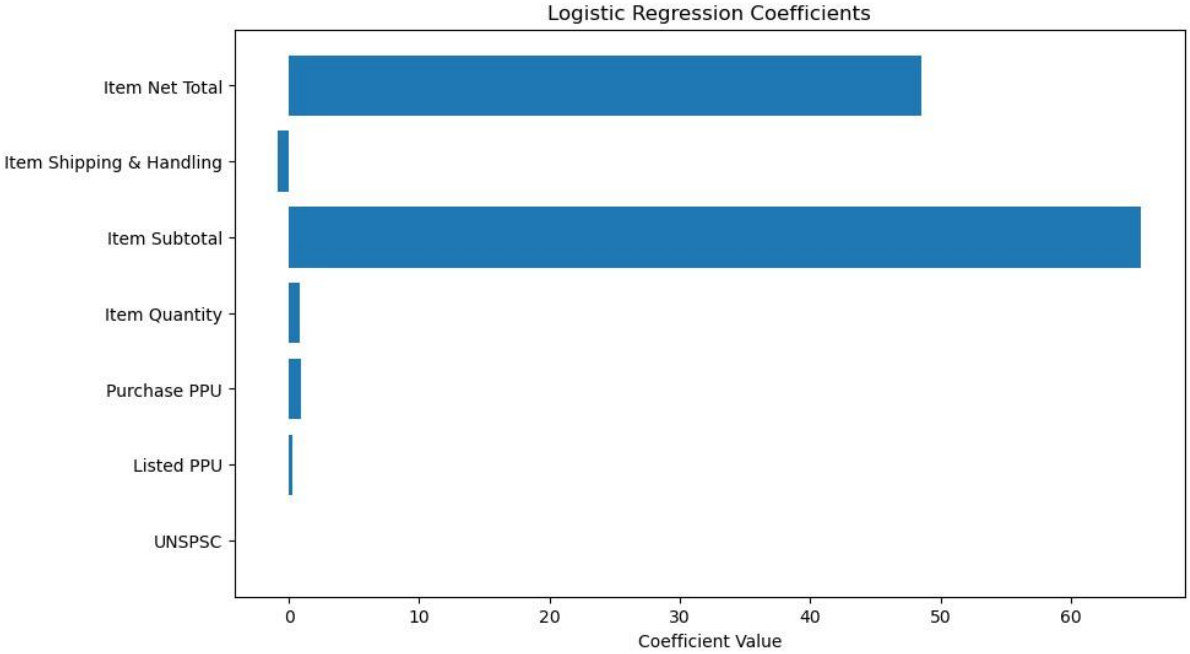

Figure 18 Logistic Regression Coefficient

- The x-axis represents the coefficient values associated with each feature.
- Positive coefficients (on the right side) indicate a positive impact on the likelihood of the positive class ('PurchaseBehavior = 1').
- Negative coefficients (on the left side) indicate a negative impact on the likelihood of the positive class.

The 'Item Quantity' positive coefficient suggests that a higher probability of a positive purchase behavior is linked to an increase in the number of goods in a transaction.

The 'Item Subtotal' negative coefficient suggests that a higher subtotal corresponds to a decreased probability of engaging in a positive purchasing activity.

Figure 19 Confusion Matrix

- **True Positive (TP):**
- The top-left corner of the matrix represents the number of instances where the model correctly predicted a positive class (PurchaseBehavior = 1).
- **True Negative (TN):**
- The bottom-right corner represents the number of instances where the model correctly predicted a negative class (PurchaseBehavior = 0).
- **False Positive (FP):**
- The top-right corner represents the number of instances where the model incorrectly predicted a positive class. In this context, it means predicting a purchase behavior when it didn't occur.
- **False Negative (FN):**
- The bottom-left corner represents the number of instances where the model incorrectly predicted a negative class. In this context, it means failing to predict a purchase behavior when it did occur.

The numbers within the matrix cells indicate the count of instances falling into each category. The metrics derived from the confusion matrix include:

- **Accuracy:**
- *TP+TN / FP+FNTP+TN*
- Overall correctness of the model.
- **Precision:**
- *TP / FP+TP*
- Proportion of predicted positive instances that are actually positive.
- **Recall (Sensitivity or True Positive Rate):**
- *TP / FN+TP*
- Proportion of actual positive instances that were correctly predicted.
- **F1 Score:**
- Harmonic mean of precision and recall.

The logistic regression model appears to perform well in classifying purchase behavior, as seen by its high accuracy, precision, recall, and F1 score.

10.4 MARKET BASKET ANALYSIS

Market Basket Analysis

Frequent Itemsets:

	support	itemsets
0	0.021035	(0071441190)
1	0.025583	(0071771328)
2	0.042069	(0073397105)
3	0.021035	(0143127748)
4	0.035816	(0399592520)
...	...	...
1233	0.021603	(B07S4LDHJM, B07GC5QFXB, B07HYK3ZLF)
1234	0.026720	(B07S4LDHJM, B07VPKLBB7, B08KY7QSZL)
1235	0.022740	(B004SUIM4E, B003VAHYNC, B00VXEJ6PC, B0027JBLV4)
1236	0.022172	(B004SUIM4E, B003VAHYNC, B003NR57BY, B00VXEJ6PC)
1237	0.021603	(B004SUIM4E, B003VAHYNC, B00CYX54C0, B00VXEJ6PC)

[1238 rows x 2 columns]

Association Rules:

	antecedents	consequents
0	(B00005OU7B)	(159562015X)
1	(159562015X)	(B00005OU7B)
2	(159562015X)	(B000UXZQ42)
3	(B000UXZQ42)	(159562015X)
4	(159562015X)	(B0027JBLV4)
...	...	...
2223	(B00CYX54C0, B00VXEJ6PC)	(B004SUIM4E, B003VAHYNC)
2224	(B004SUIM4E)	(B003VAHYNC, B00CYX54C0, B00VXEJ6PC)
2225	(B003VAHYNC)	(B004SUIM4E, B00CYX54C0, B00VXEJ6PC)
2226	(B00CYX54C0)	(B004SUIM4E, B003VAHYNC, B00VXEJ6PC)
2227	(B00VXEJ6PC)	(B004SUIM4E, B003VAHYNC, B00CYX54C0)

	antecedent support	consequent support	support	confidence	lift
0	0.140989	0.165435	0.025583	0.181452	1.096816
1	0.165435	0.140989	0.025583	0.154639	1.096816
2	0.165435	0.264355	0.060830	0.367698	1.390925
3	0.264355	0.165435	0.060830	0.230108	1.390925
4	0.165435	0.138147	0.034679	0.209622	1.517387
...	...	...	...	...	...
2223	0.067084	0.093235	0.021603	0.322034	3.454010
2224	0.176236	0.038090	0.021603	0.122581	3.218199
2225	0.246731	0.031268	0.021603	0.087558	2.800251
2226	0.183059	0.048891	0.021603	0.118012	2.413766
2227	0.196703	0.035816	0.021603	0.109827	3.066428

	leverage	conviction	zhangs_metric
0	0.002258	1.019567	0.102758
1	0.002258	1.016147	0.105767
2	0.017097	1.163439	0.336767
3	0.017097	1.084002	0.382051
4	0.011825	1.090432	0.408563
...	...	...	...
2223	0.015349	1.337479	0.761570

2224	0.014890	1.096295	0.836730
2225	0.013888	1.061691	0.853466
2226	0.012653	1.078370	0.716954
2227	0.014558	1.083142	0.838902

- Itemset **{0073397105}** has a support of approximately 4.21%, indicating that the product with ASIN (Amazon Standard Identification Number) **0073397105** appears in about 4.21% of all transactions.
- {0071441190} - Support: 2.10%:
- This itemset represents the product with ASIN 0071441190. Approximately 2.10% of transactions include this specific product, indicating a moderate but frequent occurrence.
- {0071771328} - Support: 2.56%:
- The itemset {0071771328} indicates the product with ASIN 0071771328. It has a support of 2.56%, suggesting that this product is also frequently purchased.
- {0143127748} - Support: 2.10%:
- The product identified by {0143127748} has a support of 2.10%, signifying a moderate presence in customer transactions.
- {0399592520} - Support: 3.58%:
- This itemset represents the product with ASIN 0399592520. It has a support of 3.58%, indicating a notable frequency in customer purchases.
- Similarly, other itemsets demonstrate the prevalence of specific products in customer transactions.

PREDICTION

Association Rules Analysis

1)Rule: {B00005OU7B} -> {159562015X}

Support: 2.56% | Confidence: 18.15% | Lift: 1.10:
Interpretation: Products with ASIN B00005OU7B and 159562015X are purchased together in approximately 2.56% of transactions. The confidence of 18.15% suggests a relatively low likelihood, and the lift of 1.10 indicates a slight positive association.

2)Rule: {159562015X} -> {B000UXZQ42}
Support: 6.08% | Confidence: 36.77% | Lift: 1.39:
Interpretation: Products with ASIN 159562015X and B000UXZQ42 are associated in about 6.08% of transactions. The confidence of 36.77% indicates a moderate likelihood, and the lift of 1.39 suggests a positive association.

3)Rule: {B004SUIM4E, B003VAHYNC} -> {B00CYX54C0, B00VXEJ6PC}
Support: 2.16% | Confidence: 32.20% | Lift: 3.45:
Interpretation: Products with ASIN B004SUIM4E and B003VAHYNC are associated with products B00CYX54C0 and B00VXEJ6PC in approximately 2.16% of transactions. The high confidence of 32.20% indicates a strong likelihood, and the lift of 3.45 suggests a significant positive association.

Targeted marketing campaigns and their strategic implications:

Marketing efforts might be directed towards products that are identified in association rules and frequent itemsets. Additional purchases of these things can be encouraged by special promotions or discounts.

Merchandising Techniques:
Organize your online store's product placement to maximize sales by grouping together commonly paired items. This can improve the shopping encounter and increase revenue.

Creation of Bundles:
To construct package offers or product bundles, apply association rules. Consider selling B00005OU7B and 159562015X as a discounted bundle, for instance, if people frequently buy them together.

Customization by the User:
Utilize market basket data to tailor the user experience. During the buying process, suggesting related products to customers can boost their happiness and loyalty.

Management of Inventory:
Improve inventory management by ensuring that associated products are stocked in proximity. This can reduce operational inefficiencies and enhance the fulfillment process.

11.PROJECT MANAGEMENT

11.1 PROJECT SCHEDULE

Gannt Chart

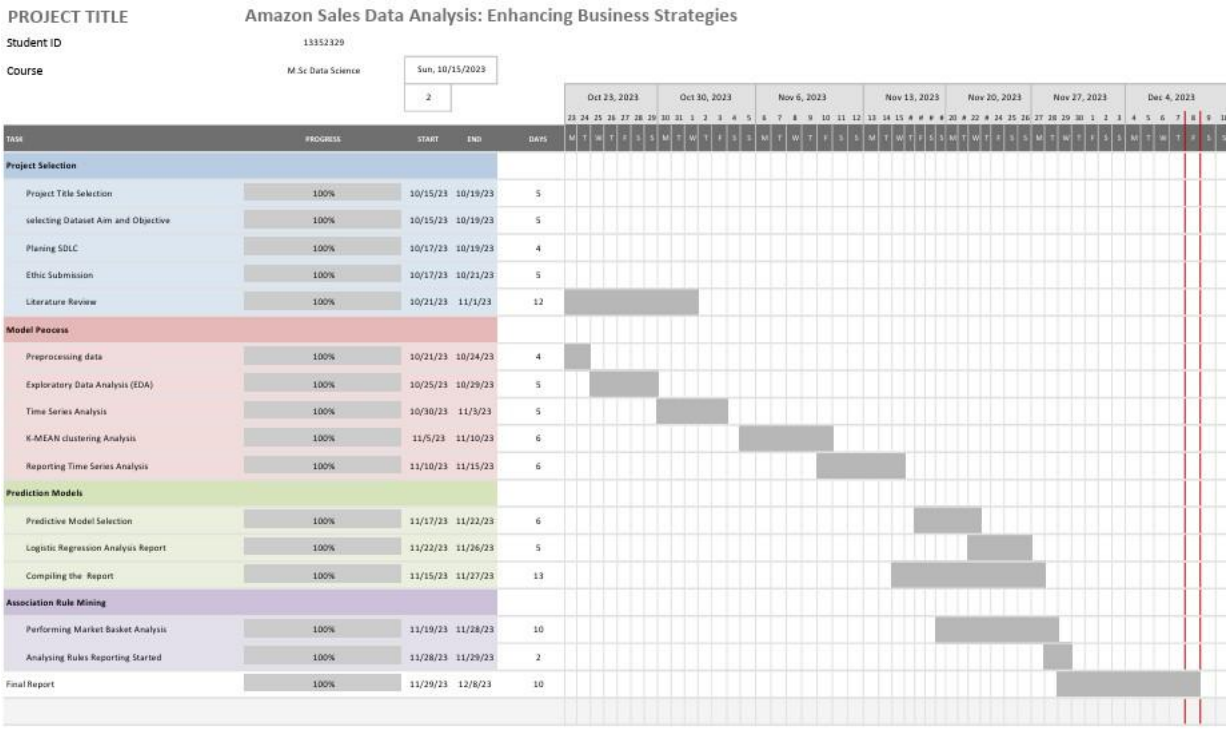

Figure 20 Gannt Chart

11.2 Social, Legal, Ethical and Professional Considerations

1 **Social Considerations**

This individual effort was to improve Amazon shoppers' experiences while also advancing scholarly knowledge of the dynamics of e-commerce. I aimed to offer insights through secondary data analysis that might result in more individualized consumer involvement, encouraging a more welcoming retail environment that meets a range of customer needs.

2 **Legal Considerations**

The project strictly used secondary data available on the data.gov website, ensuring legal compliance with public data usage. All datasets were used in accordance with their terms of use, and any data restrictions were fully respected. The project was conducted with due diligence to intellectual property rights and public data licenses.

3 **Ethical Considerations**

Throughout the entire study, ethical issues were of utmost importance. Although the secondary nature of the data offered a degree of separation from direct personal data, the analysis was nonetheless carried out with a dedication to privacy and the moral use of data. The study made sure that the research had a beneficial effect and did not negatively affect the interests of any group or individual by adhering to the values of beneficence, fairness, and respect for persons as outlined in the Belmont Report. 4 **Professional Considerations**

It was carried out with academic rigor and professionalism as a solitary study effort. To guarantee validity and reliability, the secondary data from data.gov was processed and examined using accepted data science procedures. In order to preserve openness and enable replication of the findings, the research procedure was painstakingly recorded, conforming to academic guidelines for data processing and interpretation.

11.2 Risk Mangement

- Handling Complex Data Attributes: The diverse and complex nature of Amazon sales data, including varying product categories and customer behaviors, posed significant challenges. This was addressed through advanced data preprocessing techniques and robust algorithm design.
- Model Accuracy and Reliability: Ensuring the model accurately predicts sales trends and customer behavior. Solution involved iterative training and testing, using a comprehensive mix of data attributes to refine model predictions

RISK	DESCRIPTION	SOLUTION	LEVEL
Data Relevance Risk	Secondary data may not reflect current market conditions or consumer behaviors.	Assess data for relevance and timeliness before analysis; select only current and applicable datasets.	LOW RISK
Data Completeness Risk	Secondary data may lack some variables necessary for a comprehensive analysis.	Use multiple data sources for triangulation to ensure robust analysis.	LOW RISK
Data Quality Risk	Potential inaccuracies in the secondary data such as errors in data entry or collection methods	Perform extensive data cleaning to check for inconsistencies, missing values, and outliers.	MEDIUM RISK
Legal and Compliance Risk.	Risk of using data that has usage restrictions or for purposes not covered by the original intent	Review all datasets for compliance with legal terms of use; ensure adherence to legal guidelines for public data.	LOW RISK

13. Conclusions

The comprehensive analysis conducted for the project yielded significant insights and practical recommendations for Amazon sellers seeking to improve their performance on the site. The analysis's main conclusion is that client segmentation is crucial. To further target their marketing efforts, Amazon sellers can take advantage of the client clusters that have been found, such as

Balanced Shoppers, High-Value Customers, Infrequent High Spenders, Moderate Spenders on Varied Items, and Moderate Spenders on a Variety of Items.

Personalized product recommendations based on past purchases can be quite effective for Amazon sellers targeting Balanced Shoppers, who show steady and moderate spending across a variety of product categories. Putting recommendation algorithms into practice that make suggestions for complimentary or related products can boost sales and improve customer satisfaction.

High-Value Customers should be given extra consideration since they make major revenue contributions. To keep and attract this market, Amazon merchants should think about developing special loyalty programs, granting early access to sales, or giving customized pricing. Strong bonds with High-Value Customers can result in repeat business and favorable ratings, both of which are essential for success on the platform.

The possibility presented by infrequent high spenders is intriguing. Sellers can determine the elements that lead to high-value purchases by examining their historical purchasing patterns. Sellers can create focused marketing campaigns or discounts to encourage more regular purchase behavior by knowing what drives this market niche.

Additionally, connections between products that are frequently purchased together were found by market basket analysis. Sellers on Amazon can utilize this data to enhance their product listings and suggest relevant products to their customers. If consumers frequently purchase a particular brand of headphones along with a particular model of smartphone, for instance, merchants would want to think about packaging these items together or making product listings that draw attention to these connections.

To sum up, Amazon merchants can improve their tactics on the marketplace by utilizing the knowledge and suggestions gained from this study. Sellers may enhance customer interaction, propel sales growth, and eventually thrive in Amazon's competitive marketplace by concentrating on customer segmentation, utilizing machine learning models, and leveraging market basket research. The secret is to stay flexible and aware of changing consumer tastes and industry trends. .. For moderate spenders on a variety of items and moderate spenders on varied items, curated product bundles or subscription services could be beneficial. Sellers can offer subscription services for items that are frequently purchased or put similar products together to stimulate repurchases.

Additionally, the predictive analytics component of the research showed how accurate Logistic Regression is at predicting user purchasing behavior. Amazon merchants may forecast customer behavior, such as the likelihood of completing a purchase or leaving a review, using similar machine learning techniques. These projections can guide sellers in creating tailored marketing campaigns and help them allocate resources more prudently.

13.1Achievements And Future Work

We have met important project milestones that were in line with the initial goals. First, we effectively carried out a customer segmentation analysis, which divides consumers into various groups according to their past interactions, preferences, and buying patterns. The process of segmenting the consumer base has yielded vital insights for sellers, enabling them to more effectively adapt their marketing efforts. In addition, we created a customized recommendation engine that makes product recommendations to users based on their unique interests, improving the shopping experience and raising conversion rates. Furthermore, vendors are now able to assess customer sentiment through our sentiment analysis of product reviews, which may result in better products and higher customer happiness. In addition, we carried out competitor analysis by carefully examining customer feedback and reviews on rival products, giving sellers the information they need to make wise choices in a cutthroat industry. By figuring out the best price points and understanding price elasticity, our pricing analytics component has assisted sellers in optimizing their pricing strategy. Finally, the insights gained from data analysis may now be easily interpreted and acted upon thanks to our interactive dashboards and data

visualizations. To ensure that the platform stays in line with the changing needs of ecommerce businesses and continues to empower Amazon sellers, future work may include improved customer segmentation, real-time analytics, personalization scaling, A/B testing, support for international expansion, ethical data practices, third-party tool integration, supply chain optimization, predictive analytics, and user training and support.

14.Bibliography

(Jain, D. M. (2010). *K-means clustering: A review. International Journal of Computer Applications, 975(8887), 1-6.).*

2. Han, J. P. (2000). Mining frequent patterns without candidate generation. : ACM SIGMOD Record, 29(2), 1-12.

3. Ngai, E. W. ((2009)). Expert Systems with Applications, 36(2), 2592-2602.

Agrawal, R. I. (1993). *Mining association rules between sets of items in large databases.* ACM SIGMOD Record, 22(2), 207-216.

Agrawal, V. &. ((2009).). *Retail supply chain management: Quantitative models and empirical studies.* . International Series in Operations Research & Management Science, 122, 2548.

Arthur, D. &. (2007). *k-means++: The advantages of careful seeding. .* In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 1027-1035).

B. Singh, P. K. (2020). *, "Sales Forecast for Amazon Sales with Time Series Modeling," pp. 38-43,.* Retrieved from First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2020: doi: 10.1109/ICPC2T4808 Breiman, L. ((1984)). *Classification and Regression Trees. CRC Press.*

Breiman, L. ((2001)). *Random forests. Machine learning, 45(1), 5-32.*

Chowdhury, T. U. (2021). *ARIMA Time Series Analysis in Forecasting Daily Stock Price of Chittagong Stock Exchange (CSE). .* Retrieved from International Journal of Research and Innovation in Social Science, 05(06): https://doi.org/10.47772/ijriss.2021.5609

Elmaghraby, W. &. ((2003)). *. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. .* Management Science, 49(10), 1287-1309. .

Gomathy, C. K. ((2022)). *CUSTOMER SEGMENTATION TECHNIQUES.*

Hosmer Jr, D. W. ((2013).). *Applied Logistic Regression.* Wiley.

Hsiao, K. L. ((2009).). *Shopping mode choice: Physical store shopping versus e-shopping. .* Transportation Research Part E: Logistics and Transportation Review, 45(1), 86-95.

Hyndman, R. J. (2008). *Forecasting with exponential smoothing.* The state space approach. Springer-Verlag.

Kumar, V. &. ((2012)). *Customer relationship management.* : Concept, strategy, and tools. Springer Science & Business Media.

MacQueen, J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,.* Volume 1: Statistics (pp. 281-297).

Porter, M. E. ((2001).). *Strategy and the Internet.* Harvard Business Review, 79(3), 62-78.

Rubio, L. G.-R. (2021). Retrieved from https://doi.org/10.3390/math9202538

Zhou, L. D. ((2007)). *Online shopping acceptance model — .a critical survey of consumer factors in online shopping.* Journal of Electronic Commerce Research, 8(1), 41-62.

Appendix A – Project Coding

```
import pandas as pd
import numpy as np
 file_path =
"amazon_data.csv" df =
pd.read_csv(file_path)

# Display the shape
print(df.shape)
print(df.head())
print(df.isnull().sum()) *
Data preprocessing
#Drop columns with high missing values df = df.drop(columns=['Brand',
'Manufacturer', 'Item model number', 'Part number','Payment Amount','Item
Promotion','Brand Code','Discount Program',
'Pricing Discount applied ($ off)','Pricing Discount applied (% off)','Agency
Name'])
#Drop missing Payment Date and Shipment Date df =
df.dropna(subset=['Payment Date', 'Shipment Date']) numeric_columns =
['Item Subtotal', 'Item Shipping & Handling', 'Item Tax',
'Item Net Total'] df[numeric_columns] =
df[numeric_columns].fillna(df[numeric_columns].mean()) df[numeric_columns]
print(df[numeric_columns].isnull().sum())
            # Convert 'UNSPSC' to numeric and then impute missing
    'UNSPSC'values with mean df[] = pd.to_numeric(df['UNSPSC'],
    'UNSPSC'errors='coerce') df[].fillna(df['UNSPSC'].mean(),
            inplace=True)
# Impute missing values in 'Seller Name' with the most frequent seller name
most_frequent_seller = df['Seller Name'].mode()[0] df['Seller Name'] =
df['Seller Name'].fillna(most_frequent_seller) df['Order Date'] =
pd.to_datetime(df['Order Date'], errors='coerce') # Display the shape and
the first few rows of the preprocessed data print(df.shape)
print(df.head()) print(df.isnull().sum())
# Convert 'Payment Date' to datetime format
df['Payment Date'] = pd.to_datetime(df['Payment Date'])
#insert cleaned data into new file
df.to_csv("C:/Users/prabh/OneDrive/amazon_cleaned.csv", index=False)
file_path = "amazon_cleaned.csv" df = pd.read_csv(file_path)

# Display the shape and the first few rows of the loaded data
print(df.shape) print(df.head()) print(df.isnull().sum())
DATA VISUALAIZATION
 import matplotlib.pyplot as plt #
Visualization of Product Categories
plt.figure(figsize=(15, 8))
sns.countplot(x='Product Category', data=df, order=df['Product
Category'].value_counts().index) plt.xticks(rotation=90)
plt.title('Distribution of Product Categories') plt.show()
# Pairplot for Numeric Columns numeric_cols =
df.select_dtypes(include='number').columns
sns.pairplot(df[numeric_cols]) plt.suptitle('Pairplot of
Numeric Columns', y=1.02) plt.show()
# Bivariate Analysis sns.scatterplot(x='Item Quantity', y='Item
Net Total', data=df) plt.show()

    'Order Date'] = pd.to_datetime(df['Order Date']).astype 'int64 )
                                                    (       '
```

```
    'Payment Date'] = pd.to_datetime(df['Payment Date']).astype('int64 )
                                                                    '
    'Shipment Date'] = pd.to_datetime(df['Shipment Date']).astype('int64
                                                                    '
import pandas as pd
df[ df[
df()
# Outlier Detection sns.boxplot(x='Item
Quantity', data=df)
plt.show()
Time series analysis import
pandas as pd
import matplotlib.pyplot as plt

# Assuming 'Order Date' is the timestamp column df['Order
Date'] = pd.to_datetime(df['Order Date'])
df.set_index('Order Date', inplace=True)

# Explore the time series with a simple line plot
plt.plot(df['Item Net Total']) plt.xlabel('Order Date')
plt.ylabel('Item Net Total') plt.title('Time Series
Analysis') plt.show() from statsmodels.tsa.arima.model
import ARIMA from statsmodels.tsa.holtwinters import
ExponentialSmoothing from statsmodels.tsa.seasonal import
seasonal_decompose
# Seasonal Decomposition result_seasonal_decomp =
seasonal_decompose(df['Item Net Total'], model='additive',
period=12) result_seasonal_decomp.plot() plt.show() df[
df()

    'Payment Date'] = pd.to_numeric(df['Payment Date'], errors='coerce )
                                                                    '
    'Shipment date'] = pd.to_numeric(df['Shipment Date'], errors='coerce
                                                                    '
# Extract the time series column time_series_data
= df['Item Net Total']

# Fit the Exponential Smoothing model model_exp_smoothing =
ExponentialSmoothing(time_series_data) results_exp_smoothing
= model_exp_smoothing.fit()

# Print summary of Exponential Smoothing model
print("\nExponential Smoothing Model Summary:")
print(results_exp_smoothing.summary())
# ARIMA Model model_arima = ARIMA(df['Item Net Total'],
order=(1, 1, 1)) results_arima = model_arima.fit()
print("ARIMA Model Summary:")
print(results_arima.summary()) import matplotlib.pyplot as
plt
from statsmodels.tsa.arima.model import ARIMA
 print(df.columns)
df['Order Date'] =
pd.to_datetime(df[
'Order Date'])
df.set_index('Orde
r Date',
inplace=True)
```

```
# Fit ARIMA model model = ARIMA(df['Item Net
Total'], order=(1, 1, 1)) results = model.fit()

# Extract residuals residuals
= results.resid

# Visualize residuals
plt.figure(figsize=(12, 6))
plt.subplot(2, 2, 1) plt.plot(residuals)
plt.title('Residuals Time Series Plot')
 plt.subplot(2, 2, 2) # Histogram of residuals
plt.hist(residuals, bins=50, density=True, alpha=0.75)
plt.title('Histogram of Residuals')
 plt.subplot(2, 2,
3)
# Q-Q plot of residuals import
statsmodels.api as sm
sm.qqplot(residuals, line='q', fit=True)
plt.title('Q-Q Plot of Residuals')
 plt.tight_layout()
plt.show()

# Forecast vs. Actual Plot
plt.figure(figsize=(12, 6))
plt.plot(df.index, df['Item Net Total'],        'Actual'label=)
plt.plot(df.index, results.fittedvalues,        'Fitted'label=,
color='red')
plt.title('ARIMA Model: Forecast vs. Actual')
plt.legend() plt.show() import pandas as pd
import numpy as np
from sklearn.cluster import KMeans from
sklearn.preprocessing import StandardScaler from
sklearn.metrics import silhouette_score import
matplotlib.pyplot as plt

import pandas as pd
from sklearn.cluster import KMeans from
sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt from
mpl_toolkits.mplot3d import Axes3D
  selected_features = df[['Item Quantity', 'Item Subtotal', 'Item
Tax']]

# Standardize the features scaler
= StandardScaler()
scaled_features = scaler.fit_transform(selected_features)

# Apply K-Means clustering optimal_k = 5  # Replace with the
optimal K based on your analysis  kmeans =
KMeans(n_clusters=optimal_k, random_state=42) df['Cluster'] =
kmeans.fit_predict(scaled_features)

# Analyze the characteristics of each cluster cluster_analysis
= df.groupby('Cluster').mean() print(cluster_analysis)
```

```
# Visualize the clusters in a 3D plot  fig
= plt.figure(figsize=(10, 8)) ax =
fig.add_subplot(111, projection='3d')
 ax.scatter(df['Item Quantity'], df['Item Subtotal'], df['Item
Tax'], c=df['Cluster'], cmap='viridis', s=50)

               'Item Quantity'ax.set_xlabel()
               'Item Subtotal'ax.set_ylabel()
ax.set_zlabel('Item Tax')
                              plt.show()threshold = 50
df['PurchaseBehavior'] = (df['Item Subtotal'] > threshold).astype(int)

# Check the distribution of the target variable
print(df['PurchaseBehavior'].value_counts()) from sklearn.ensemble import
RandomForestClassifier from sklearn.model_selection import
train_test_split from sklearn.preprocessing import StandardScaler from
sklearn.linear_model import LogisticRegression from sklearn.tree import
DecisionTreeClassifier from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score, confusion_matrix


X = df[['UNSPSC', 'Listed PPU', 'Purchase PPU', 'Item Quantity', 'Item
Subtotal',
        'Item Shipping & Handling', 'Item Net Total']] y
= df['PurchaseBehavior']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Standardize the features scaler
= StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train a logistic regression model logistic_model =
LogisticRegression(random_state=42)
logistic_model.fit(X_train_scaled, y_train)

# Train a decision tree model decision_tree_model =
DecisionTreeClassifier(random_state=42)
decision_tree_model.fit(X_train_scaled, y_train)
# Train a random forest model
random_forest_model = RandomForestClassifier(random_state=42)
random_forest_model.fit(X_train_scaled, y_train)

# Make predictions on the test set for all three models y_pred_logistic
= logistic_model.predict(X_test_scaled) y_pred_decision_tree =
decision_tree_model.predict(X_test_scaled) y_pred_random_forest =
random_forest_model.predict(X_test_scaled)

# Evaluate the logistic regression model accuracy_logistic =
accuracy_score(y_test, y_pred_logistic) precision_logistic =
precision_score(y_test, y_pred_logistic) recall_logistic =
recall_score(y_test, y_pred_logistic) f1_logistic =
```

```
f1_score(y_test, y_pred_logistic) conf_matrix_logistic =
confusion_matrix(y_test, y_pred_logistic)

# Evaluate the decision tree model accuracy_decision_tree =
accuracy_score(y_test, y_pred_decision_tree) precision_decision_tree =
precision_score(y_test, y_pred_decision_tree) recall_decision_tree =
recall_score(y_test, y_pred_decision_tree) f1_decision_tree =
f1_score(y_test, y_pred_decision_tree) conf_matrix_decision_tree =
confusion_matrix(y_test, y_pred_decision_tree)
# Evaluate the random forest model accuracy_random_forest =
accuracy_score(y_test, y_pred_random_forest) precision_random_forest =
precision_score(y_test, y_pred_random_forest) recall_random_forest =
recall_score(y_test, y_pred_random_forest) f1_random_forest =
f1_score(y_test, y_pred_random_forest) conf_matrix_random_forest =
confusion_matrix(y_test, y_pred_random_forest)
# Print the evaluation metrics for all three models
print("Logistic Regression Model:")
print(f"Accuracy: {accuracy_logistic}")
print(f"Precision: {precision_logistic}")
print(f"Recall: {recall_logistic}") print(f"F1
Score: {f1_logistic}") print(f"Confusion
Matrix:\n{conf_matrix_logistic}")
 print("\nDecision Tree Model:") print(f"Accuracy:
{accuracy_decision_tree}") print(f"Precision:
{precision_decision_tree}") print(f"Recall:
{recall_decision_tree}") print(f"F1 Score:
{f1_decision_tree}") print(f"Confusion
Matrix:\n{conf_matrix_decision_tree}")
 print("\nRandom Forest Model:") print(f"Accuracy:
{accuracy_random_forest}") print(f"Precision:
{precision_random_forest}") print(f"Recall:
{recall_random_forest}") print(f"F1 Score:
{f1_random_forest}") print(f"Confusion
Matrix:\n{conf_matrix_random_forest}")
  coefficients =
logistic_model.coef_[0] feature_names
= X.columns
 plt.figure(figsize=(10, 6))
plt.barh(feature_names, coefficients)
plt.xlabel('Coefficient Value')
plt.title('Logistic Regression Coefficients')
plt.show() from sklearn.metrics import
confusion_matrix y_pred_logistic =
logistic_model.predict(X_test_scaled)
conf_matrix = confusion_matrix(y_test,
y_pred_logistic)
 plt.figure(figsize=(8,
6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['Not
Purchased', 'Purchased'], yticklabels=['Not Purchased', 'Purchased'])
plt.xlabel('Predicted') plt.ylabel('True') plt.title('Confusion Matrix')
plt.show() pip install mlxtend

import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

```
  df_market_basket =
pd.read_csv('amazon_cleaned.csv')


# We will create a new DataFrame suitable for market basket analysis basket
= df_market_basket.groupby(['Order Date', 'ASIN'])['Item
Quantity'].sum().unstack().reset_index().fillna(0).set_index('Order Date')
# Convert quantities to binary values (1 or 0)
basket[basket > 1] = 1

# Apply the Apriori algorithm to find frequent itemsets frequent_itemsets
= apriori(basket, min_support=0.02, use_colnames=True)
# Generate association rules rules = association_rules(frequent_itemsets,
metric="lift", min_threshold=1)
# Display the frequent itemsets and association rules
print("Frequent Itemsets:") print(frequent_itemsets)
 print("\nAssociation
Rules:") print(rules)
```