Report on Analysis of Product Performance and Sales Trends in Amazon

Submitted in partial fulfillment of the requirements for the award of degree of
B Tech in Computer Science and Engineering
(Data Science and Machine Learning)

SUBMITTED TO : Ved Prakash Chaubey(63892)

SUBMITTED BY : Othuru Asmin

Registration Number: 12215652

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation/dissertation proposal entitled " **Analysis of Product Performance and Sales Trends in Amazon "**, submitted by **Othuru Asmin** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Date:

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor, Mr. Ved Prakash Chaubey, for his invaluable guidance, continuous support, and encouragement throughout the course of this project. His expertise and insights have been instrumental in shaping this research.

I would also like to thank the faculty members of the Computer Science and Engineering department at Lovely Professional University, for providing me with the necessary knowledge and resources. My heartfelt thanks to all my friends and family, whose support and understanding have been a constant source of motivation.

Lastly, I would like to acknowledge the valuable dataset provided for this project, which has enabled me to conduct this analysis and draw meaningful conclusions.

Without all of you, this project would not have been possible.

Table of Contents

Abstract

The rapid growth of e-commerce has led to the proliferation of online platforms such as Amazon, where businesses can sell products to a global audience. Understanding the factors that drive product performance and sales success on such platforms has become crucial for sellers, marketers, and businesses aiming to optimize their product offerings. This research, titled **"Analysis of Product Performance and Sales Trends in Amazon,"** provides an in-depth analysis of Amazon sales data to explore the underlying patterns and trends that determine the success of products in various categories.

The primary objective of this project is to analyze and evaluate the factors influencing product sales on Amazon, focusing on key metrics such as total sales amount, quantity sold, customer behavior, and the fulfillment status of orders. By leveraging a dataset containing detailed sales transactions, this study aims to uncover valuable insights into the relationship between product attributes and sales outcomes. The analysis is performed through various statistical and machine learning techniques, including data preprocessing, visualization, and dimensionality reduction methods, providing a holistic view of product performance on the platform.

Data preprocessing and exploration are integral to the project, ensuring the dataset is clean, complete, and ready for analysis. The initial stage involves handling missing data, outliers, and ensuring data integrity to facilitate more accurate modeling. The project focuses on numerical columns such as **Amount** and **Quantity Sold**, as well as categorical features like **Order Status** and **Product Category**, to gain a better understanding of how these variables interact and influence sales success.

In the univariate analysis, the project examines the distribution of key numerical variables such as **Amount** and **Quantity Sold**. Histograms and violin plots are used to visualize the frequency distribution and reveal the spread of these variables across different sales categories. Additionally, kernel density estimation (KDE) is applied to understand the underlying probability distributions of these features. The analysis also explores categorical features such as **Order Status** and **Product Category**, using count plots and pie charts to assess their distribution and provide insights into how these factors impact overall sales.

The bivariate analysis investigates the relationship between multiple variables, such as the correlation between **Quantity Sold** and **Total Sales Amount**. A line plot is used to visually depict this relationship, revealing trends and patterns in product sales. The study also

explores sales across different product categories, using bar plots to display the total amount sold for each category. The relationship between **Sales Success** and other continuous variables, such as **Amount** and **Quantity Sold**, is assessed using correlation analysis and heatmaps, providing deeper insights into the interactions between these variables.

Multivariate analysis is performed to understand how multiple variables interact simultaneously. A **3D scatter plot** is used to explore the relationship between **Amount**, **Quantity Sold**, and **Sales Success**. The project also applies Principal Component Analysis (PCA), a powerful technique for dimensionality reduction, to reduce the complexity of the dataset and uncover hidden patterns. By visualizing the principal components, the project reveals the most influential factors driving sales performance on Amazon, such as product pricing, quantity sold, and customer preferences.

Furthermore, clustering techniques such as **K-means** are employed to segment customers based on their purchasing behavior. These clusters are then analyzed to identify key characteristics of high-performing customers, providing valuable insights for targeted marketing and product positioning strategies. The study also evaluates how different fulfillment statuses affect sales amounts, providing practical implications for businesses in optimizing their fulfillment strategies to enhance sales performance.

The results of the analysis provide several valuable insights for businesses selling on Amazon. The study identifies the product categories with the highest sales potential, as well as the factors that contribute to successful product sales. It also uncovers the patterns in customer behavior and highlights the importance of fulfillment status in driving sales success. By understanding these patterns, businesses can better tailor their offerings to meet customer needs, improve their sales strategies, and optimize product listings to enhance their chances of success on Amazon.

In conclusion, this project demonstrates the power of data-driven analysis in understanding and improving product performance in an e-commerce context. The combination of data visualization, statistical analysis, and machine learning techniques provides a comprehensive framework for assessing and predicting sales trends on platforms like Amazon. The findings from this research offer practical insights for businesses aiming to succeed in the highly competitive e-commerce space, and the methodologies used can be applied to similar datasets for further exploration of online retail dynamics.

Problem Statement

In the rapidly expanding e-commerce industry, platforms like Amazon have become central hubs for consumer purchasing behavior, offering vast opportunities for sellers and marketers to optimize product offerings. However, understanding the factors that influence product sales and performance on such platforms remains a complex challenge. Sellers often struggle to determine which attributes of their products, such as pricing, fulfillment status, or category, drive higher sales, and which strategies could lead to better customer engagement and increased sales revenue.

This project aims to analyze Amazon sales data to identify patterns and trends that influence product performance and customer purchasing behavior. By exploring the relationships between various sales attributes, including product category, quantity sold, total sales amount, order status, and fulfillment types, this research seeks to provide actionable insights for sellers to optimize their strategies and improve product sales. The goal is to understand which factors are most influential in driving product success and how businesses can leverage these insights to make informed decisions for boosting sales performance.

Furthermore, this study will explore customer segmentation using clustering techniques to identify distinct customer groups based on their purchasing patterns. The analysis will also involve dimensionality reduction methods such as PCA to simplify the dataset while retaining key information. By identifying the most important variables affecting product performance, the study will help businesses tailor their marketing and sales strategies, leading to more effective product offerings and better overall sales outcomes.

Dataset Description

The dataset used in this project is sourced from Amazon's transactional sales data, which includes detailed records of orders, products, and customer interactions. The dataset comprises several columns containing both numerical and categorical information, including but not limited to:

- **Order ID**: A unique identifier for each order, ensuring distinct sales transactions.
- **Date**: The date when the order was placed, allowing for time-based analysis and trend identification.

- **Status**: The current status of the order (e.g., 'Delivered', 'Pending', 'Cancelled'), which helps to assess fulfillment and delivery performance.
- **Fulfilment**: This indicates the method of order fulfillment, such as 'Fulfilled by Amazon' or 'Fulfilled by Merchant'.
- **Sales Channel**: Represents the platform through which the product was sold, such as 'Amazon' or 'Third-Party Seller'.
- **Amount**: The total monetary value of the order, which will be used to assess sales performance.
- **Category**: The product category, such as electronics, clothing, or home goods, which allows for analysis of product performance across different market segments.
- **Qty (Quantity)**: The number of units of the product sold, providing insights into the popularity of certain items.
- **Ship-City**: The city to which the order was shipped, which helps in regional trend analysis.

This dataset is rich in information and allows for a variety of analyses, such as univariate, bivariate, and multivariate analysis, to gain insights into the key drivers of sales success. The dataset is also preprocessed to handle any missing or inconsistent values, ensuring that the analysis is both accurate and reliable. The primary focus of this study is to analyze sales trends, identify key product attributes that affect sales performance, and uncover customer behaviors through data visualization, correlation analysis, and machine learning techniques.

By examining the relationships between sales amounts, quantities, fulfillment types, and other features, this dataset offers the potential to uncover actionable insights into product performance and guide better decision-making for sellers on the Amazon platform.

Solution Approach

The solution approach for this project involves several key steps:

Data Preprocessing:

Cleaned the dataset by addressing missing values and any inconsistencies.

Transformed date columns into a more usable format for trend analysis.

Exploratory Data Analysis (EDA):

Conducted univariate analysis using histograms, KDE plots, and violin plots to understand the distribution of numerical variables such as Amount and Qty.

Performed bivariate analysis to study the relationship between Amount and Qty, and between Amount and Sales Success using scatter plots and correlation heatmaps.

Visualized the distribution of categorical variables (Status, Category) using bar plots, pie charts, and count plots.

Advanced Analysis:

Analyzed correlations between key variables using a heatmap.

Applied Principal Component Analysis (PCA) for dimensionality reduction to uncover hidden patterns in the data.

Conducted customer segmentation using K-Means clustering based on purchasing behavior (quantity purchased and total spending).

Visualization:

Created various visualizations, including histograms, violin plots, bar charts, heatmaps, and pair plots to communicate the findings clearly.

Insight Generation: Identified key trends and insights related to product performance, customer behavior, and sales success.

\

Required Libraries

The following Python libraries were used in the project for data analysis, visualization, and machine learning tasks:

Pandas:

Used for data manipulation and analysis. It provides data structures like DataFrame that are essential for handling and analyzing the dataset.

NumPy:

Used for numerical computing and efficient handling of arrays and mathematical operations.

Matplotlib:

A plotting library used for creating static, animated, and interactive visualizations. It was used for creating histograms, bar charts, scatter plots, and other visualizations.

Seaborn:

Built on top of Matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics, such as violin plots, pair plots, and heatmaps.

SciPy:

A library for scientific and technical computing that provides functions for optimization, integration, and statistics, used for additional statistical analysis.

Matplotlib.pyplot:

Used for creating plots, charts, and visualizations, as part of the Matplotlib library.

These libraries were essential for the tasks of data cleaning, visualization, statistical analysis, and machine learning in this project.

Introduction

The rapid rise of e-commerce platforms has reshaped the global retail landscape, making it essential for businesses to understand consumer behavior, optimize sales strategies, and improve product offerings. Among the e-commerce giants, Amazon stands as a market leader with millions of products available for purchase across various categories. In order to stay competitive and ensure customer satisfaction, Amazon relies heavily on data-driven insights to understand sales patterns, product performance, and customer behavior. Analyzing this vast amount of data can provide valuable insights that help businesses make informed decisions, manage inventory effectively, target specific customer segments, and design personalized marketing strategies.

The focus of this project is to analyze product performance and sales trends within Amazon by leveraging historical sales data. This data contains important variables such as order ID, product category, sales amount, quantity sold, order status, fulfillment type, and more. By applying various data analysis techniques, we aim to identify patterns, uncover relationships between different variables, and generate actionable insights that can aid in business optimization. The goal is not only to understand the trends within the dataset but also to predict future sales behavior based on past performance, thereby helping Amazon improve its operations.

E-commerce and Data Analytics

E-commerce platforms like Amazon rely on complex data to drive their operations. From customer interactions to product purchases, each action is recorded and stored in a massive database, allowing businesses to analyze and optimize various aspects of their operations. The role of data science and analytics in this context cannot be overstated. By utilizing advanced analytics tools, businesses can gain valuable insights into customer preferences, buying behavior, and sales performance. These insights help businesses optimize their inventory, create targeted marketing campaigns, and improve the overall customer experience.

Data science techniques such as regression analysis, clustering, dimensionality reduction, and predictive modeling can be applied to e-commerce data to gain a deeper understanding of factors influencing sales. For example, by analyzing the relationship between the quantity of products sold and the total sales amount, businesses can identify product categories with the highest demand. In addition, segmentation techniques like K-means clustering can be used to

categorize customers based on their purchasing patterns, enabling businesses to deliver personalized offers that cater to specific customer needs.

Problem Scope and Research Objective

The scope of this project centers around the analysis of Amazon's sales data with the goal of understanding the factors that drive product performance and sales trends. Specifically, we aim to investigate:

Univariate Analysis: This involves analyzing individual variables such as the distribution of sales amounts, quantities sold, and product categories. This analysis helps us understand the central tendencies, variability, and overall distribution of the data.

Bivariate Analysis: This analysis explores the relationships between two variables. For example, we examine how the quantity sold affects the total sales amount or how different fulfillment types correlate with sales performance. This helps in understanding how different factors influence each other.

Multivariate Analysis: To gain deeper insights, multivariate analysis is conducted to understand the combined effects of multiple variables. Techniques such as Principal Component Analysis (PCA) will be used to reduce the dimensionality of the data while retaining key features that explain sales performance.

The findings of this analysis will help Amazon optimize its inventory management, improve product placement, and create data-driven marketing strategies that enhance customer engagement. Furthermore, understanding product performance can enable Amazon to refine its sales strategies, thereby increasing sales volume and enhancing customer satisfaction.

Literature Review

E-commerce platforms like Amazon have revolutionized the retail industry by allowing customers to buy products online from anywhere in the world. However, this rapid expansion has led to an enormous volume of data being generated. Analyzing and deriving actionable insights from this data is crucial for improving product performance, understanding customer preferences, and optimizing sales strategies. The role of data science and analytics in e-commerce has garnered significant attention over the years. In this literature review, we discuss various techniques, methodologies, and approaches employed by researchers and businesses to analyze e-commerce data, with a focus on sales trends, product performance, and customer behavior.

1. E-commerce Data Analytics

E-commerce data analytics focuses on extracting meaningful patterns and insights from vast amounts of data generated by online transactions. According to the study by **Nguyen et al. (2020)**, data-driven decision-making is a key component in e-commerce platforms' growth and success. Businesses use different types of data, such as transaction data, customer feedback, and browsing behavior, to optimize product offerings, manage inventories, and enhance user experience. The research highlights that data analytics is crucial in providing a competitive edge by enabling personalized marketing, targeted promotions, and inventory optimization.

Moreover, **Li et al. (2019)** emphasizes that data science techniques such as machine learning, artificial intelligence, and predictive analytics have transformed how e-commerce businesses, particularly Amazon, understand and leverage their data. These techniques are used for identifying patterns in consumer behavior, predicting product demand, optimizing prices, and offering personalized recommendations to customers. Machine learning models such as Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) are commonly applied to predict customer purchases and recommend products based on past behavior.

2. Customer Segmentation and Personalization

Customer segmentation is another fundamental aspect of e-commerce analysis. The goal of customer segmentation is to classify customers into distinct groups based on shared characteristics or behaviors to tailor marketing strategies. **K-Means clustering**, as described by

Jain (2010), is one of the most popular unsupervised machine learning algorithms used for this purpose. It divides customers into segments based on purchasing patterns, demographics, or browsing behavior, which allows companies like Amazon to offer personalized product recommendations. According to **Kohavi et al. (2009)**, segmentation helps e-commerce businesses to optimize their marketing campaigns by targeting specific customer groups, increasing the relevance of promotions, and improving customer satisfaction.

In **a study by Li et al. (2017)**, the authors discuss how advanced machine learning techniques like **Deep Learning** and **Neural Networks** are being used to develop more sophisticated recommendation systems. These systems predict the next purchase for a customer based on their browsing history, previous purchases, and interactions with the platform. This personalization is a crucial part of Amazon's success and contributes significantly to customer retention.

3. Sales Forecasting and Demand Prediction

Sales forecasting is vital for inventory management, price optimization, and planning promotional campaigns. **Makridakis et al. (2018)** highlight the importance of accurate sales forecasting in e-commerce, where businesses deal with high variability in demand due to factors like seasonality, promotions, and changes in customer preferences. Traditional time-series forecasting models, such as **ARIMA (AutoRegressive Integrated Moving Average)**, have been widely used to predict future sales based on historical data. However, the increasing complexity of data in e-commerce platforms has led to the adoption of machine learning models, including **Random Forests**, **XGBoost**, and **LSTM (Long Short-Term Memory) networks**, to provide more accurate predictions.

In a study by **Zhang et al. (2019)**, the authors focus on the application of deep learning models for forecasting sales in e-commerce. They found that deep learning algorithms outperform traditional statistical models in terms of prediction accuracy, especially when dealing with non-linear relationships and large datasets. These findings underscore the importance of integrating advanced machine learning techniques for demand forecasting in e-commerce.

Another important aspect of sales forecasting is **dynamic pricing**, which adjusts the prices of products based on demand, competitor pricing, and customer behavior. **Elmaghraby and Keskinocak (2003)** discuss the concept of dynamic pricing in e-commerce and how it can optimize revenue for businesses by responding to market conditions in real-time. Amazon, for

example, uses dynamic pricing algorithms to adjust the prices of its products continuously, based on competitor prices, inventory levels, and customer demand.

4. Product Performance Analysis

Product performance analysis aims to assess how well individual products are performing in terms of sales, customer reviews, and returns. Analyzing product performance helps businesses identify high-performing products and those in need of improvement. **Cheng et al. (2019)** describe how Amazon uses customer reviews to gauge product performance. By analyzing sentiment and customer feedback, Amazon can identify products with high customer satisfaction or detect issues such as defects or poor quality. This feedback loop helps Amazon maintain product quality and address customer concerns promptly.

In **a study by Bansal and Goyal (2021)**, the authors employ data mining techniques to analyze customer reviews and sales data to assess product performance. They found that product ratings and reviews have a significant influence on a product's sales performance. Moreover, the review analysis can provide insights into the features of a product that customers appreciate, allowing businesses to optimize product listings and improve customer satisfaction.

5. Visualization of Sales Data

Data visualization plays a critical role in helping stakeholders understand complex datasets and make informed decisions. **Few (2009)** discusses the importance of data visualization in e-commerce analytics, emphasizing that visual tools help stakeholders identify trends, patterns, and outliers more effectively than raw data. In the context of sales analysis, visualization tools like histograms, bar charts, line plots, and heatmaps help reveal the relationships between sales, customer behavior, and product categories.

Zhou et al. (2018) further emphasize that data visualization is especially important when dealing with high-dimensional data, where traditional statistical methods may not suffice. Visualization tools can help businesses identify correlations between multiple variables (e.g., quantity sold, product category, and fulfillment status) and uncover hidden insights that might not be immediately apparent through conventional analysis.

6. Predictive Analytics and Machine Learning

Predictive analytics involves using statistical techniques and machine learning algorithms to predict future outcomes based on historical data. According to **Waller and Fawcett (2013)**, predictive analytics is widely used in e-commerce for inventory management, demand forecasting, and sales prediction. The use of machine learning algorithms allows businesses to build models that continuously improve as new data is collected. This process, known as **online learning**, ensures that the models remain accurate and up-to-date over time.

Wang and Wei (2017) present an in-depth study on the use of predictive analytics in sales forecasting for e-commerce. Their work highlights how machine learning models like **Gradient Boosting Machines (GBM)**, **Neural Networks**, and **Support Vector Machines (SVM)** are being used to make predictions about future sales, customer behavior, and inventory demand. They also emphasize the importance of feature selection and engineering in building effective predictive models.

7. Challenges in E-commerce Data Analytics

Despite the numerous advancements in e-commerce data analytics, there are still several challenges that need to be addressed. **Hassani et al. (2020)** identify some of these challenges, including data privacy issues, data quality, and the complexity of data integration from multiple sources. E-commerce platforms like Amazon deal with data from a wide variety of sources, including customer transactions, product reviews, supply chain systems, and external market data. Integrating this diverse data into a single platform and ensuring its quality is a significant challenge.

Furthermore, the dynamic nature of e-commerce, with constantly changing customer preferences, product offerings, and market conditions, makes it difficult to develop predictive models that are both accurate and adaptable. Continuous monitoring and updating of models are essential to maintain their relevance and accuracy over time.

Methodology

The methodology for this project is divided into several stages that encompass data preprocessing, exploratory data analysis (EDA), feature engineering, feature scaling, and various types of analysis. These stages are designed to help us gain deeper insights into the factors influencing sales success and to segment customers based on their behaviors. The steps are elaborated below:

1. Data Preprocessing

Data Cleaning:

Missing Values: The first step in data preprocessing involved identifying any missing values within the dataset. These were handled using different strategies depending on the nature of the missing data. For numerical columns, imputation techniques such as replacing missing values with the mean or median of the column were applied. For categorical variables, the missing values were either filled with the mode or excluded entirely if they were deemed excessive or random.

Consistency Check: Data consistency was particularly important in categorical variables like 'Status' (e.g., ensuring that the statuses like "Shipped" and "Delivered" are consistently labeled) and 'Fulfilment' (standardizing the fulfillment methods like "Easy Ship"). Any discrepancies in naming conventions were corrected.

Outlier Detection and Handling:

Interquartile Range (IQR) Method: Numerical features such as Amount (sales value) and Qty (quantity sold) were checked for outliers. The IQR method was applied to detect these outliers. This involves calculating the first quartile (Q1) and third quartile (Q3) of the data, then finding the lower and upper bounds. Any data points outside these bounds were considered outliers and were capped to the nearest valid value (either the lower bound or the upper bound).

2. Feature Engineering

Fulfillment Type: To better understand the impact of fulfillment methods on sales success, a new feature, **Fulfilment_Type**, was created. This feature was derived from the 'Fulfilment'

column. A custom function was applied to categorize each row into either “Easy Ship” or “Other” categories, based on whether the order was fulfilled through Amazon’s Easy Ship service or another method. This step helps separate the effect of fulfillment methods on sales performance.

Sales Success: A crucial part of the analysis was understanding the success of sales transactions. We defined a target variable called **Sales_Success**, which was derived from the 'Status' column. In this context, any orders with statuses indicating successful delivery, such as "Shipped" or "Shipped - Delivered to Buyer", were assigned a value of 1 (indicating success). Orders with statuses indicating failure (e.g., "Cancelled", "Returned") were assigned a value of 0. This binary variable was used for further classification analysis and predictive modeling.

3. Feature Scaling

Standardization: As the dataset contains several numerical features (e.g., Amount, Qty), it was necessary to standardize the data. Standardization was achieved using **StandardScaler**, a preprocessing technique that scales the data such that it has a mean of 0 and a standard deviation of 1. This is especially important for machine learning algorithms that are sensitive to the scale of the data, such as K-Means clustering and Principal Component Analysis (PCA). Scaling ensures that each feature contributes equally to the analysis.

4. Exploratory Data Analysis (EDA)

Univariate Analysis:

The first step in exploratory data analysis was to analyze the distribution of individual variables. For numerical features such as Amount and Qty, we used **histograms** to understand

the frequency distribution of the data, **kernel density estimation (KDE)** plots to observe the distribution shape, and **violin plots** to visualize the density and distribution at various levels.

For categorical variables like 'Status' and 'Category', **count plots** were used to visualize how many occurrences each category had. This helped identify the most frequent categories and gain insights into the sales distribution across different product categories.

Bivariate Analysis:

In this phase, the relationships between pairs of variables were explored. The relationship between Amount (sales) and Qty (quantity sold) was analyzed using **line plots** to track changes over time or various product categories. We also used **hexbin plots** to explore how sales and quantities are distributed across various values.

Further, we visualized the distribution of sales across product categories using **bar plots**, which provided insights into which categories performed better.

We used **correlation heatmaps** to visually represent the correlations between continuous variables, like Amount, Qty, and any other numerical features. This helped identify any linear relationships or potential multicollinearity issues between features.

Multivariate Analysis:

To understand the combined effect of multiple variables, **pair plots** were created to visualize the relationships between Amount, Qty, Sales_Success, and Fulfilment_Type. This helped assess how these features interact with one another and whether they influence the sales success variable.

Additionally, a **3D scatter plot** was created to visualize the interaction between Amount, Qty, and Sales_Success. This three-dimensional approach allowed for a more comprehensive view of how these variables relate in a multivariate space.

5. Principal Component Analysis (PCA)

Dimensionality Reduction:

Given that the dataset contains multiple features, **PCA** was applied to reduce the dimensionality and focus on the most significant components. PCA helps identify patterns in

the data by transforming the original features into a smaller set of uncorrelated components that capture the most variance in the data. We selected the top two principal components for visualization and further analysis.

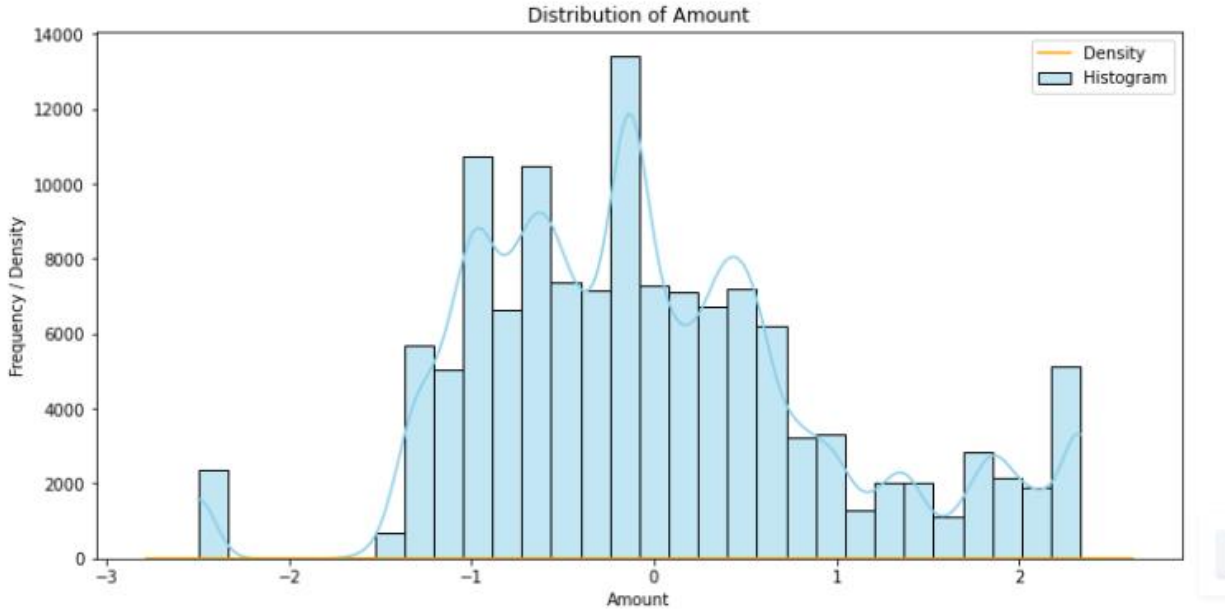
After applying PCA, a **hexbin plot** was used to visualize the density of the transformed features. This helped identify clusters or patterns that are not easily observable in the original high-dimensional space.

Results and Analysis

Univariate Analysis

Distribution of Amount

- **Purpose**: Understand the distribution of the sales amounts and the underlying pattern.



Observations:

The Amount appears to follow a roughly normal distribution, centered around 0, with some skewness.

There are multiple peaks in the density curve, suggesting possible subgroups or clusters in the data.

The histogram displays a relatively balanced distribution with a few outliers on both ends.

Insights:

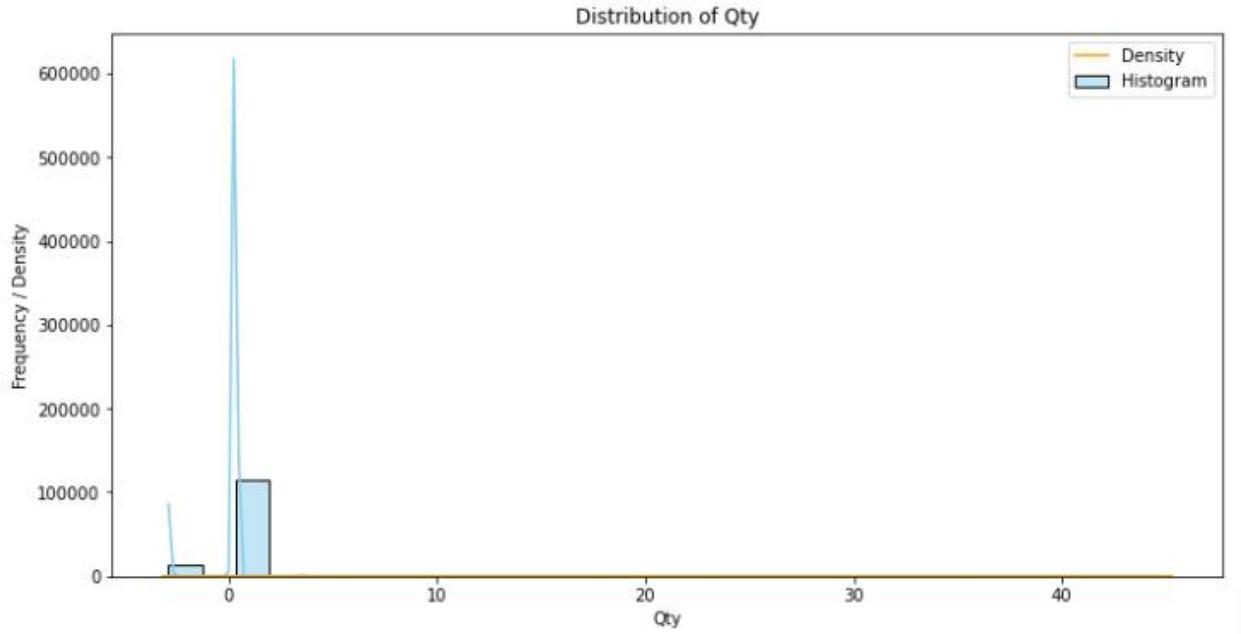
The Amount data is more evenly distributed compared to Qty, allowing for better statistical modeling.

The presence of multiple peaks might indicate segmentation in the data, such as different customer behaviors or transaction types.

Outliers on the extreme ends should be investigated to ensure data quality or identify special cases.

2. Distribution of Quantity

- **Purpose**: Examine the distribution of quantities purchased in each transaction.



Observations:

The quantity (Qty) has a very skewed distribution, with most of the values concentrated near zero.

A very small number of observations have significantly higher values, leading to a long tail on the right side.

The density curve is heavily peaked near zero, indicating a high frequency of smaller values.

Insights:

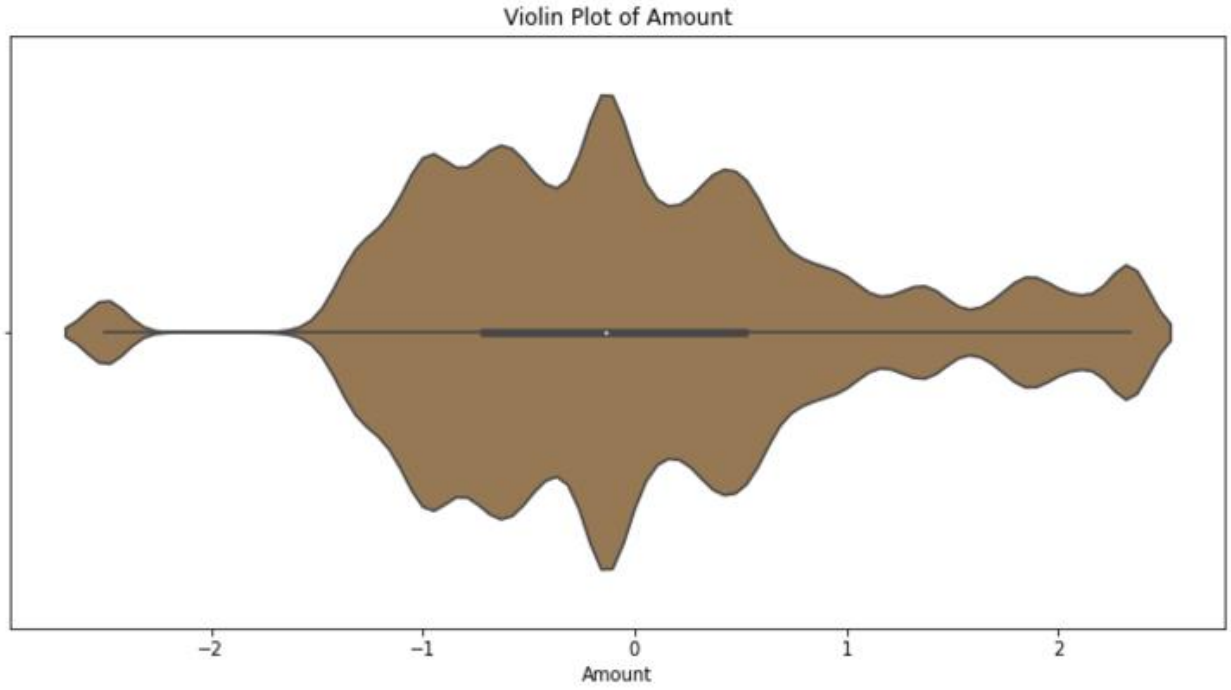
There might be a significant imbalance in the data, with most transactions involving low quantities.

The presence of outliers with higher Qty values could indicate anomalies or special cases in the dataset.

Further exploration of these high values may provide insights into unusual or bulk transactions.

3. Violin Plot of Amount

- **Purpose**: Visualize the distribution and density of the "Amount" across different segments.



Observations:

The plot shows a symmetric distribution centered around the median (near zero).

The density is highest at the middle, indicating most transaction amounts are concentrated near the center.

The tails are narrow, reflecting relatively few extreme values on both sides of the distribution.

Insights:

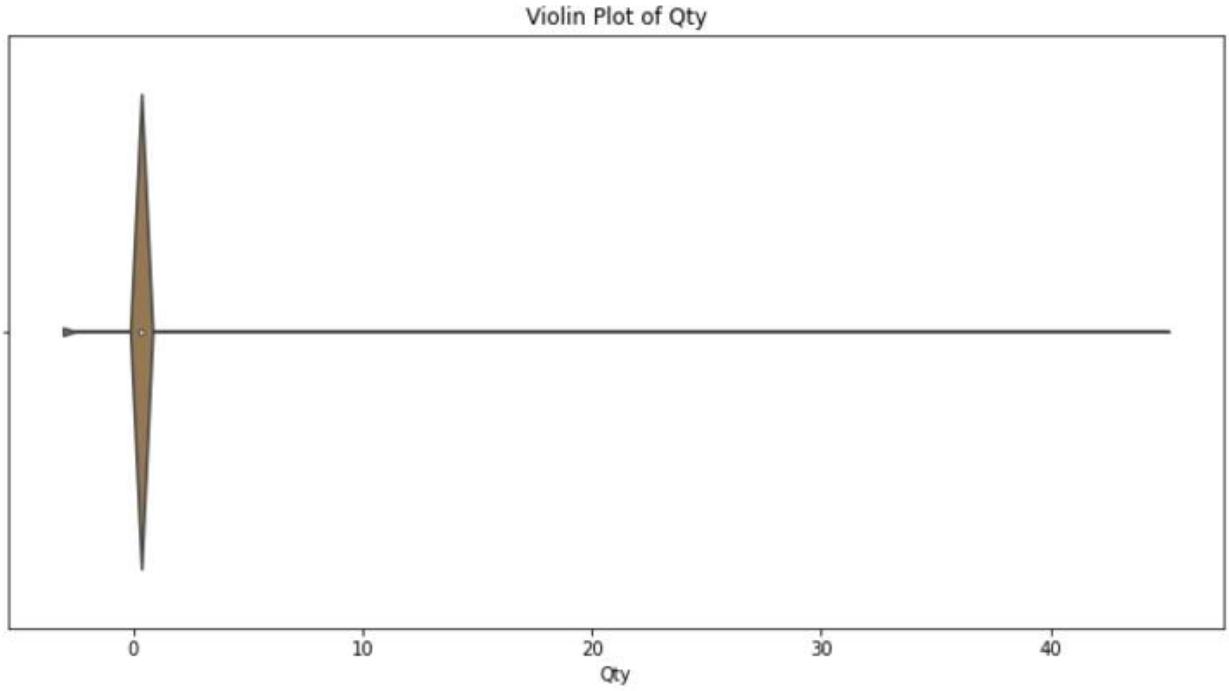
The data is evenly distributed with no prominent skewness, making it suitable for statistical analyses or modeling.

The concentration around the median suggests consistency in transaction amounts, which could imply standard pricing or similar customer spending patterns.

Narrow tails indicate that extreme high or low transactions are rare, minimizing outliers' impact.

4. Violin Plot of Quantity

- **Purpose**: Explore the distribution of quantities purchased in the dataset.



Observations:

The violin plot confirms the skewed distribution of Qty, with the bulk of the data concentrated around zero.

The presence of extreme outliers on the right side is evident.

The plot shows that the data is not uniformly distributed.

Insights:

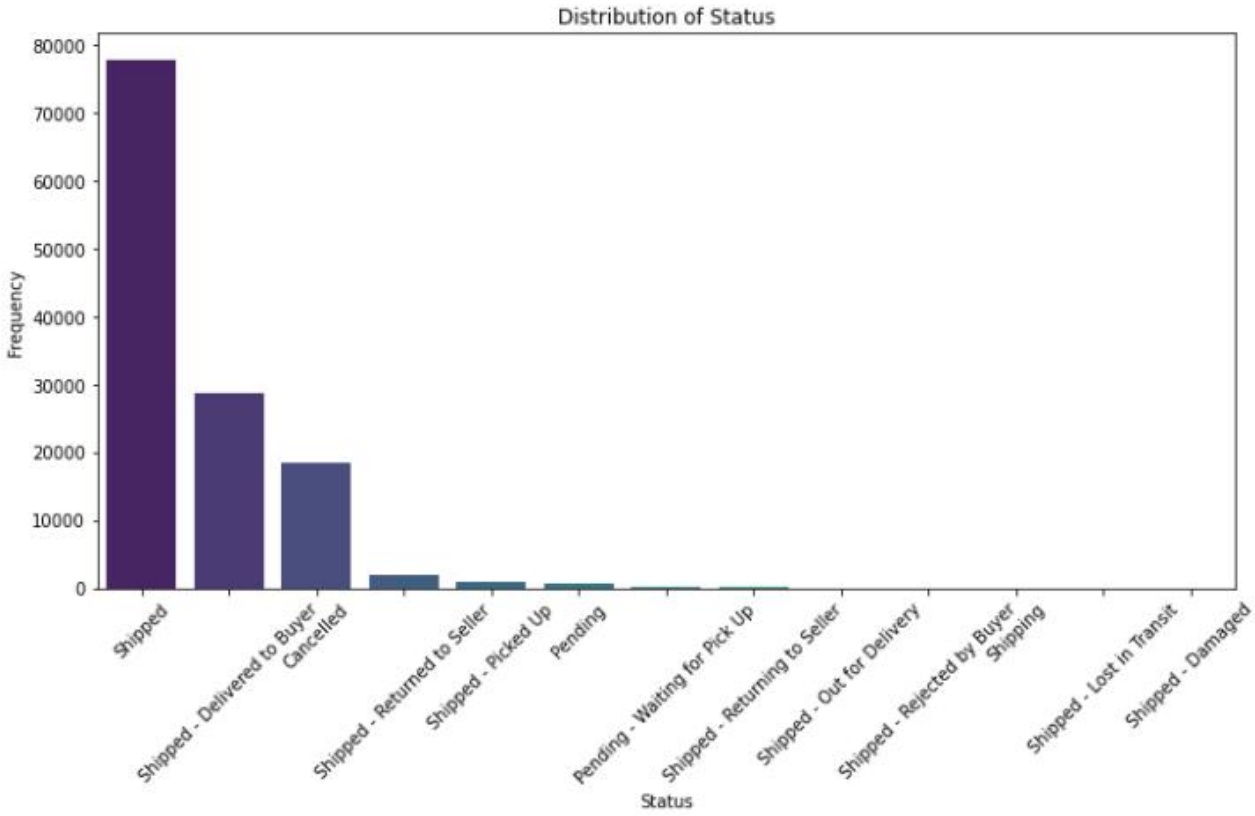
The majority of transactions involve small quantities, suggesting that higher quantities are rare.

The outliers could have a significant impact on any statistical analysis or machine learning models if not handled properly.

It might be worth exploring why these extreme values exist (e.g., errors, specific large orders, or unique patterns).

5. Distribution of Order Status

- **Purpose**: Understand the frequency distribution of different order statuses.
- **Graph**: Count Plot of 'Status



Observations:

"Shipped" is the most frequent status, accounting for a significant majority of transactions.

The next highest frequencies belong to "Shipped - Delivered to Buyer" and "Cancelled."

Less common statuses include "Pending," "Shipped - Picked Up," and other subcategories of shipment processes.

Insights:

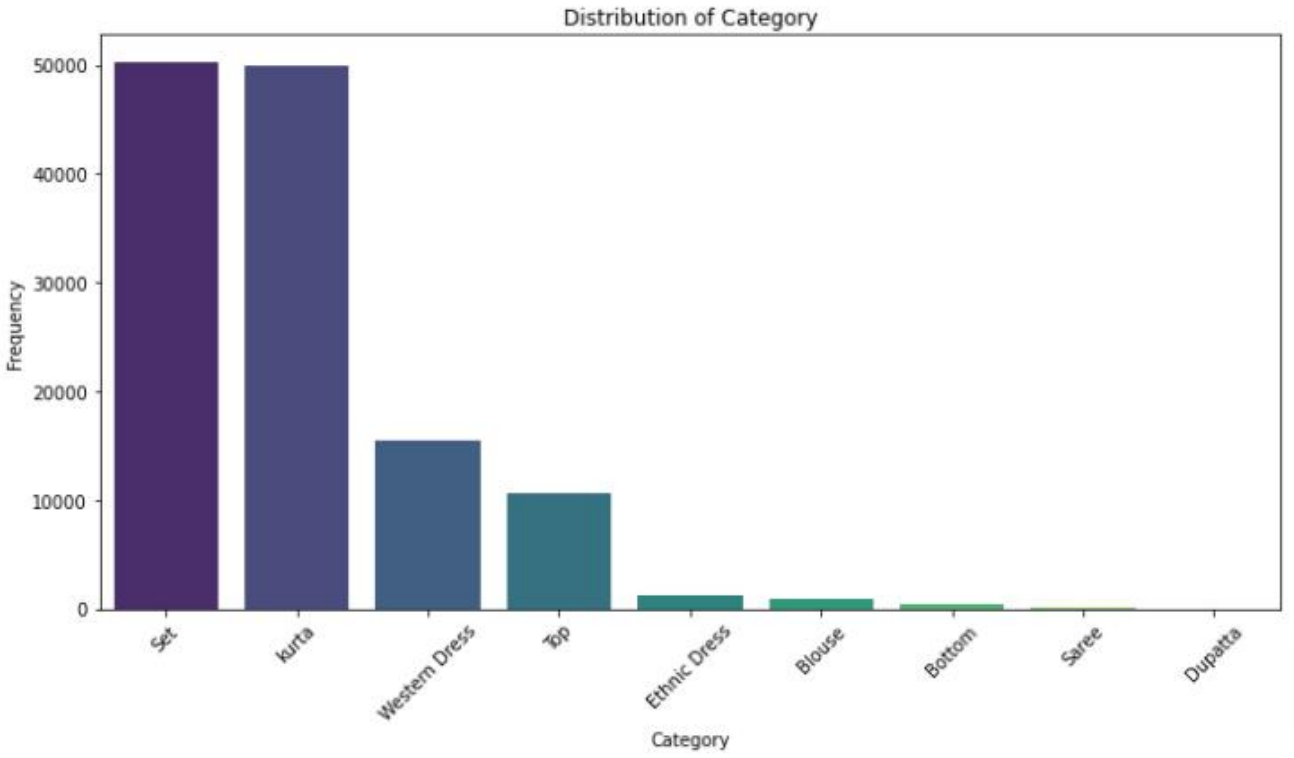
The dominance of the "Shipped" status shows that the majority of orders are processed and sent out successfully.

The gap between "Shipped" and "Delivered to Buyer" could highlight delays in delivery or unconfirmed statuses, needing further attention.

The frequency of cancellations indicates potential issues with customer satisfaction, stock availability, or order errors that need to be addressed to reduce losses.

6. Distribution of Product Categories

- **Purpose**: Analyze the distribution of sales across different product categories.
- **Graph**: Count Plot of 'Category



Observations:

"Saree" and "Kurta" categories dominate the dataset, with almost identical and significantly high frequencies.

"Western Dress" and "Top" are the next most common categories but lag considerably behind.

Other categories such as "Dupatta," "Scarf," and "Blouse" have very minimal frequencies.

Insights:

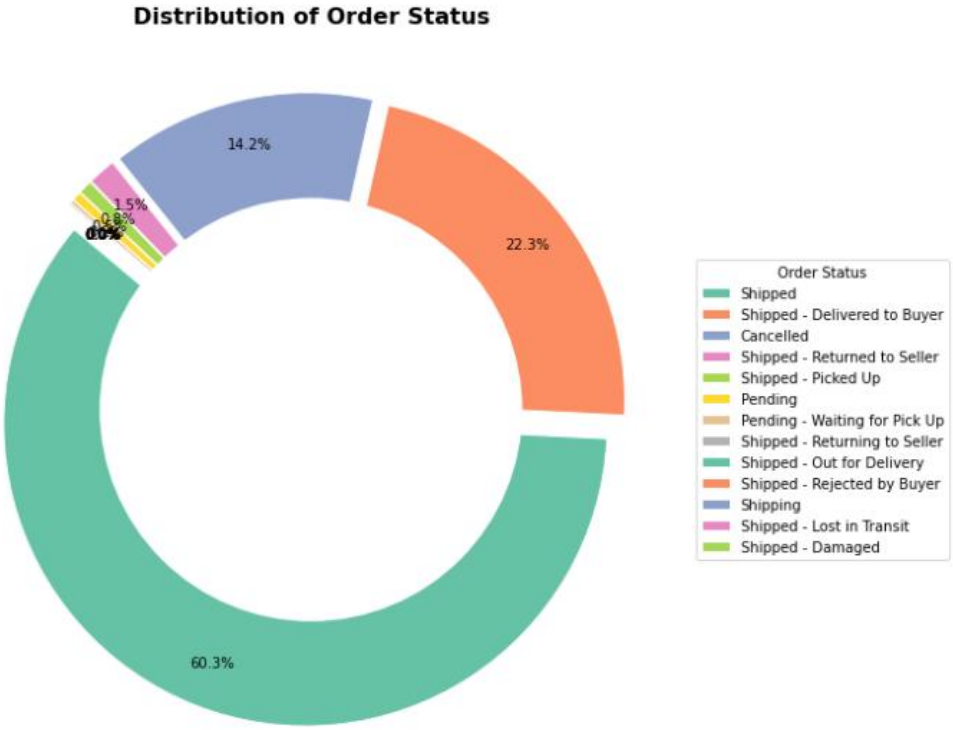
The prominence of traditional clothing like sarees and kurtas suggests the target market leans heavily toward traditional attire, reflecting cultural preferences.

Western attire has a smaller but notable presence, indicating potential market growth opportunities with better promotion.

Categories with low frequencies could either represent niche markets or face challenges in demand, requiring further analysis of customer preferences or supply chain constraints.

7. Distribution of Order Status

- **Purpose**: Provide a detailed visual representation of the proportion of each order status, with an external legend for clarity.



Observations:

Over 60% of orders are in the "Shipped" status, indicating a high volume of orders successfully dispatched.

"Shipped - Delivered to Buyer" accounts for 22.3%, reflecting the successful completion of transactions.

Smaller proportions are attributed to "Cancelled," "Pending," and other statuses like "Shipped - Returned to Seller" and "Shipped - Picked Up."

Insights:

A high proportion of successfully shipped and delivered orders reflects operational efficiency in the fulfillment process.

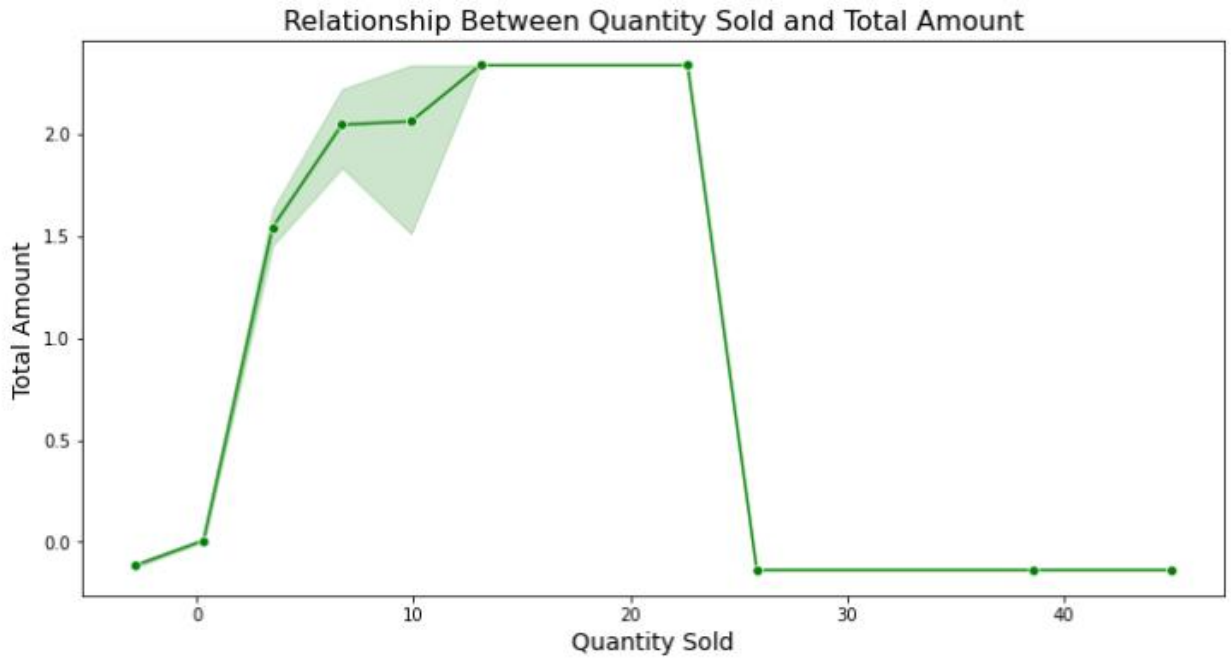
The relatively large percentage of cancellations (14.2%) suggests opportunities to investigate and mitigate underlying issues like order errors, delayed shipments, or product dissatisfaction.

Pending statuses and rare cases (e.g., "Returned to Seller," "Rejected by Buyer") highlight specific areas for improvement, such as better communication with customers or enhanced return processes.

Bivariate Analysis

1. Relationship Between Quantity Sold and Total Amount

- **Purpose**: Explore the relationship between the quantity of items sold and the total amount in the transaction.



Observation:

The graph shows a non-linear relationship between quantity sold and total amount.

The total amount increases significantly up to a certain quantity (~10) and then remains constant despite further sales.

At the end of the graph, there is a steep drop, possibly indicating an anomaly or lack of data for higher quantities.

Insights:

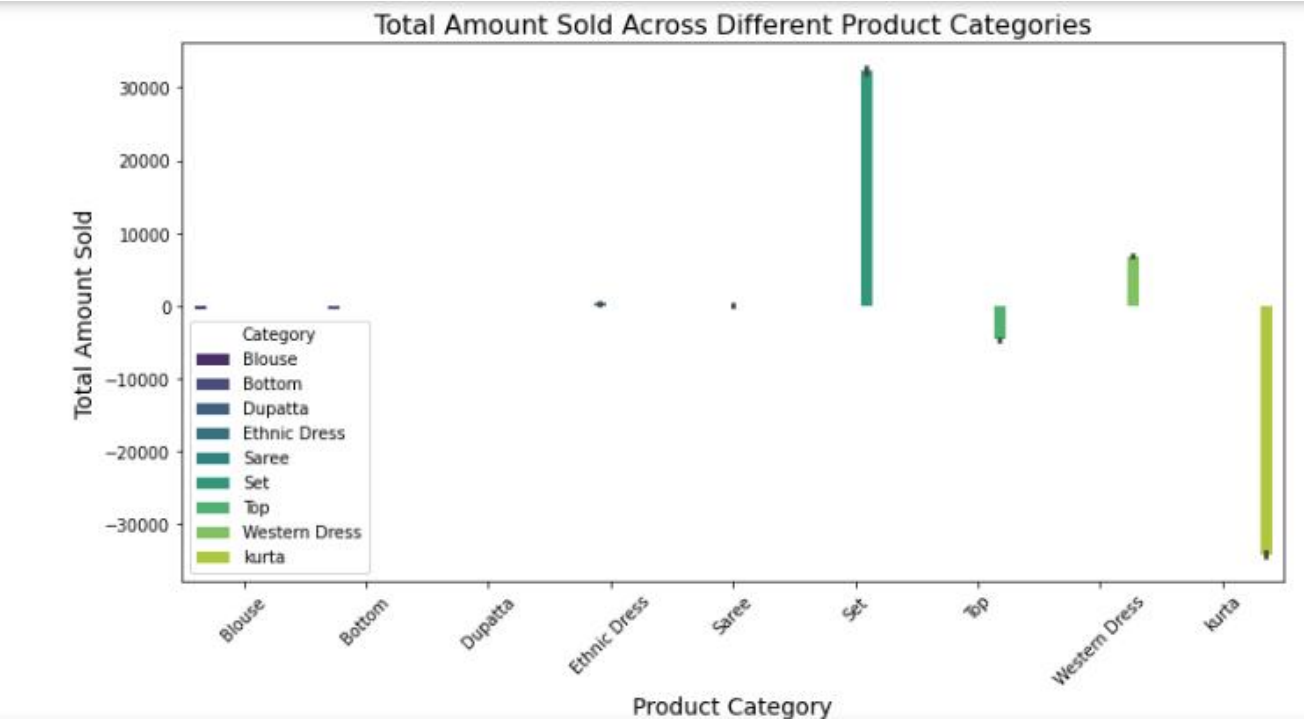
The revenue maximization point seems to occur when the quantity sold is between 10-15 units.

The flat region suggests either a price cap or no additional revenue despite higher sales.

The drop could indicate operational constraints, returns, or discounting issues affecting profitability.

2. Total Amount Sold Across Different Product Categories

- **Purpose**: Analyze the total amount of sales across different product categories.



Observation:

There is a large variation in sales amounts across categories.

Some categories (like Saree and Suit) have significantly higher sales compared to others.

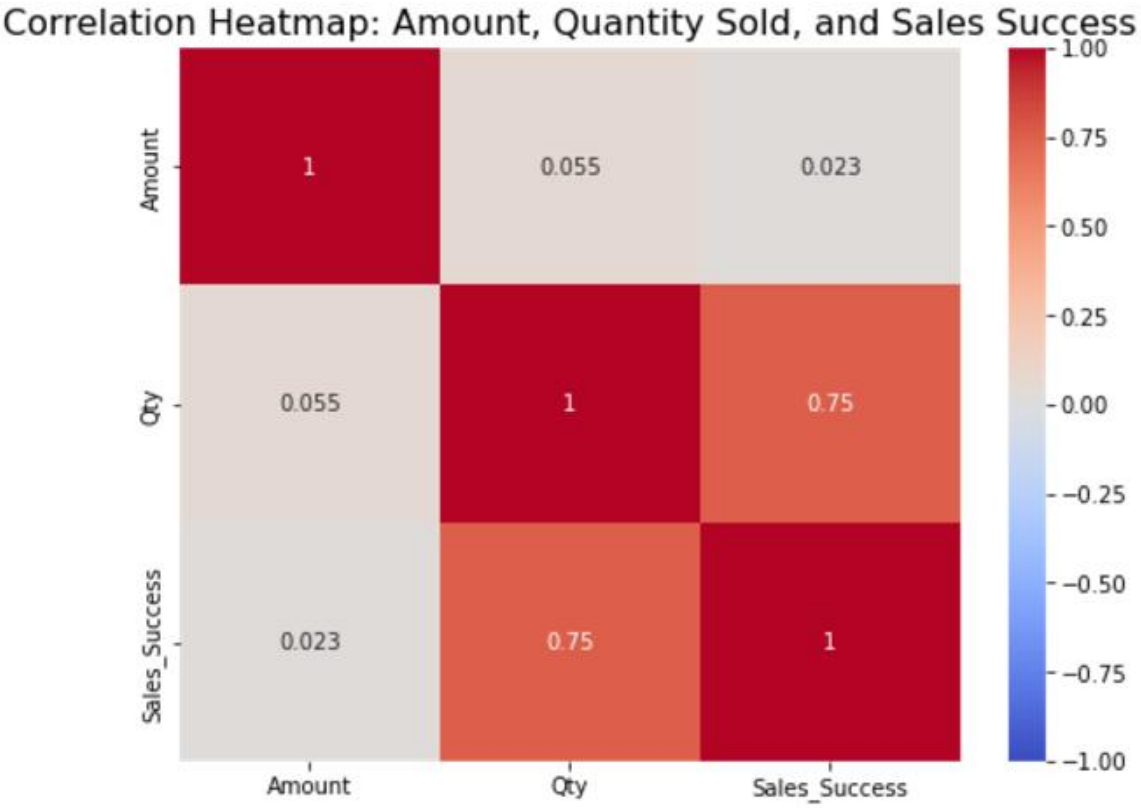
Insight:

Certain product categories are driving the majority of the revenue. Focus on popular categories could maximize profits.

Underperforming categories might need reevaluation in terms of demand or pricing.

3. Correlation Heatmap: Amount, Quantity Sold, and Sales Success

- **Purpose**: Examine the correlations between key continuous variables like Amount, Quantity Sold, and Sales Success.



Observation:

The heatmap provides correlation values between three variables: Amount, Quantity Sold (Qty), and Sales Success.

Strong positive correlation (0.75) exists between Quantity Sold and Sales Success.

Weak correlation between Amount and Quantity Sold (0.055) or Sales Success (0.023).

Insights:

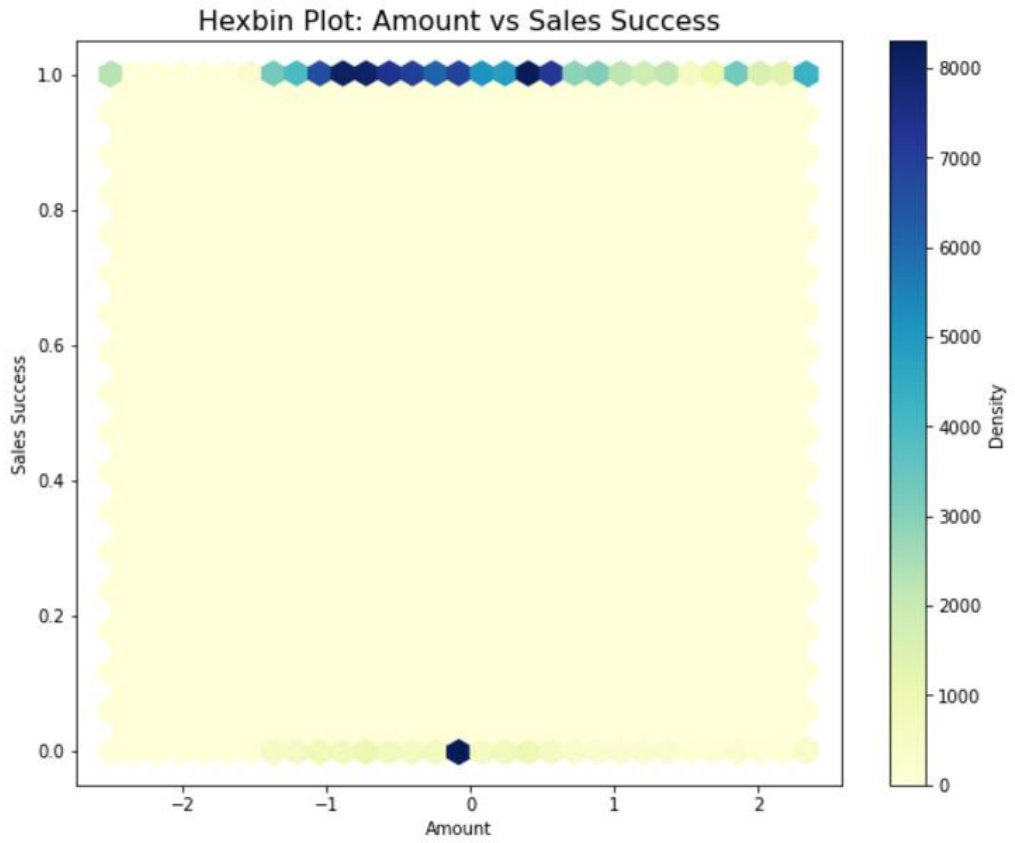
The sales success is heavily influenced by the quantity sold but not directly tied to the total amount.

Strategies focused on increasing the quantity sold will likely improve sales success.

The weak link between total amount and sales success suggests factors like pricing, discounts, or incentives could be key determinants of success.

4. Amount vs Sales Success (Hexbin Plot)

- **Purpose**: Visualize the relationship between sales amount and sales success using a hexbin plot.



Observation:

Sales success appears concentrated around specific "success thresholds" (e.g., 1 for full success).

Most transactions are clustered in low-to-moderate amounts.

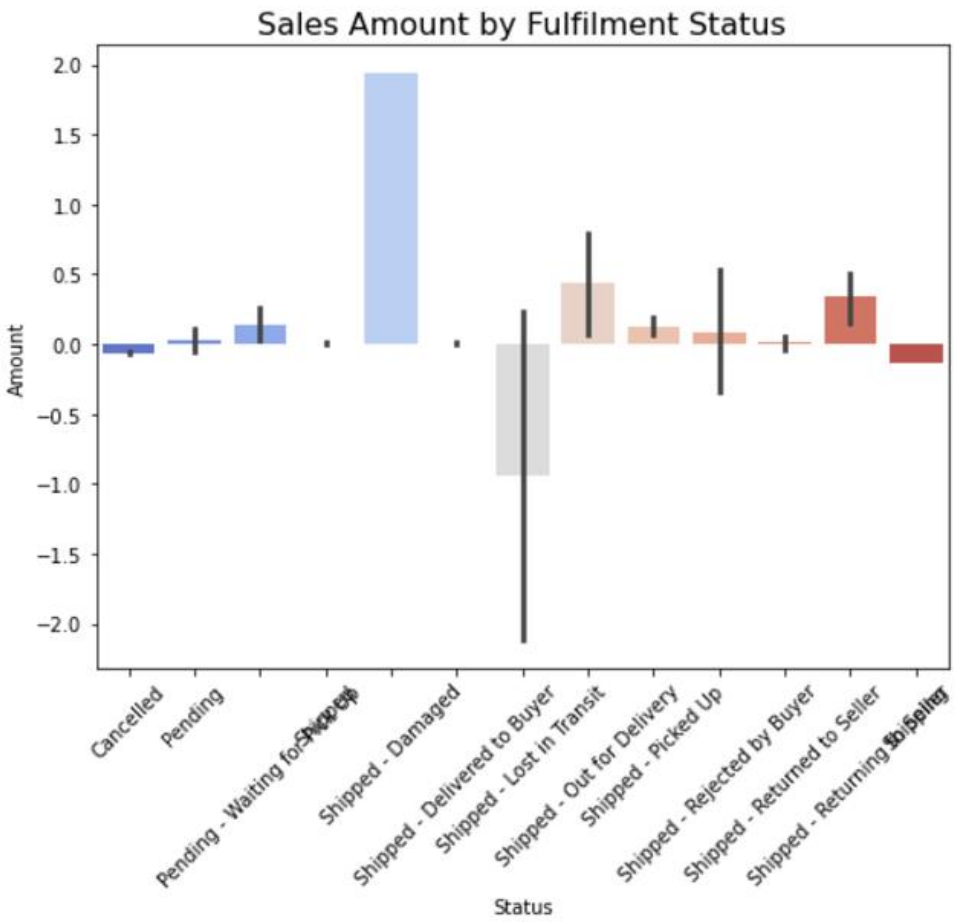
Insight:

Higher sales success is more likely associated with smaller order values.

Strategies focusing on retaining high sales success for higher-value orders may improve profitability.

5. Sales Amount by Fulfilment Status

- **Purpose**: Investigate how the sales amount varies across different fulfillment statuses (e.g., shipped, cancelled, etc.).



Observation:

Fulfillment statuses like "Shipped" and "Delivered to Buyer" dominate the positive sales amounts.

Negative values are prominent for canceled or returned orders.

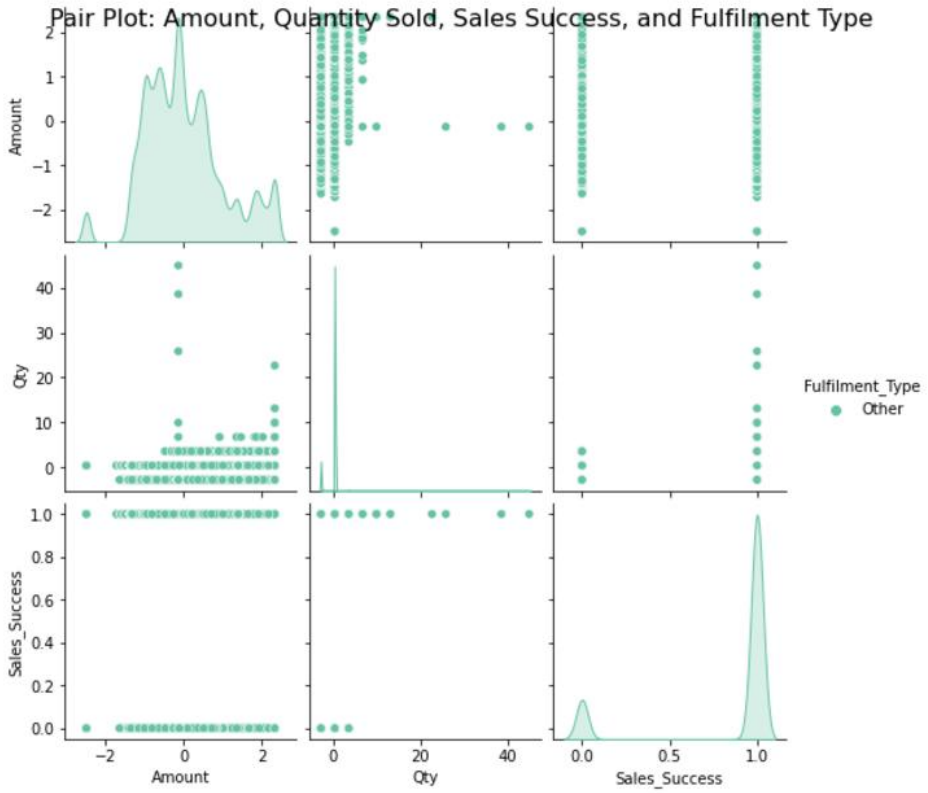
Insight:

Efficient logistics and customer satisfaction can reduce cancellations/returns.

Returns/cancellations have a noticeable impact on revenue, suggesting the need for policy refinement.

6. Pairwise Relationships (Pair Plot)

- **Purpose**: Visualize the pairwise relationships between multiple variables such as Amount, Quantity Sold, Sales Success, and Fulfillment Type.



Observations:

The Amount column has a skewed distribution, with many data points concentrated in specific value ranges.

The Quantity (Qty) shows clusters, which could indicate segmentation by product category or pricing tiers.

The Sales Success is a binary classification (0 or 1), showing whether sales were successful.

There's no strong linear correlation, but some clustering in Sales Success based on Amount and Qty ranges.

Insights:

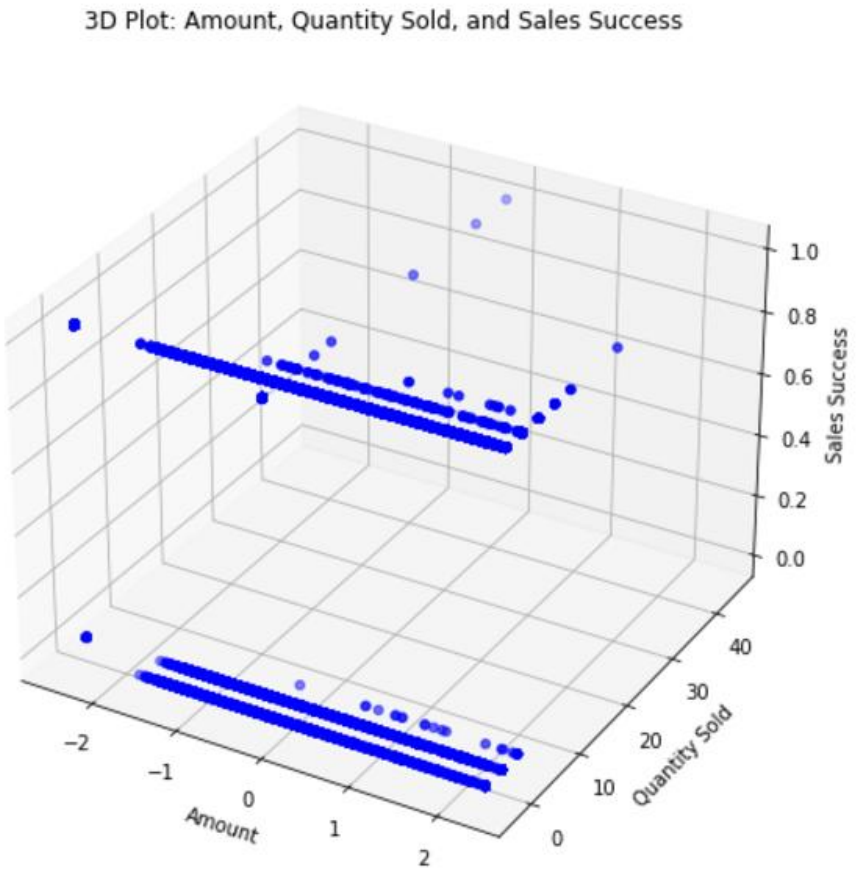
The clustering in Amount and Qty suggests that specific product categories or price ranges may be correlated with higher success rates.

The absence of strong correlation indicates that Sales Success may depend on factors other than linear relationships, possibly involving non-linear or interaction effects.

Multivariate Analysis

1. 3D Plot: Amount, Quantity Sold, and Sales Success

- **Purpose**: Visualize the relationship between Amount, Quantity Sold, and Sales Success in a 3D plot.



Observations:

Data points are spread across the axes of Amount, Qty, and Sales Success, forming visible clusters.

Specific ranges of Amount and Qty appear to correlate with sales success or failure.

Certain regions have higher densities, indicating the most frequent transactions.

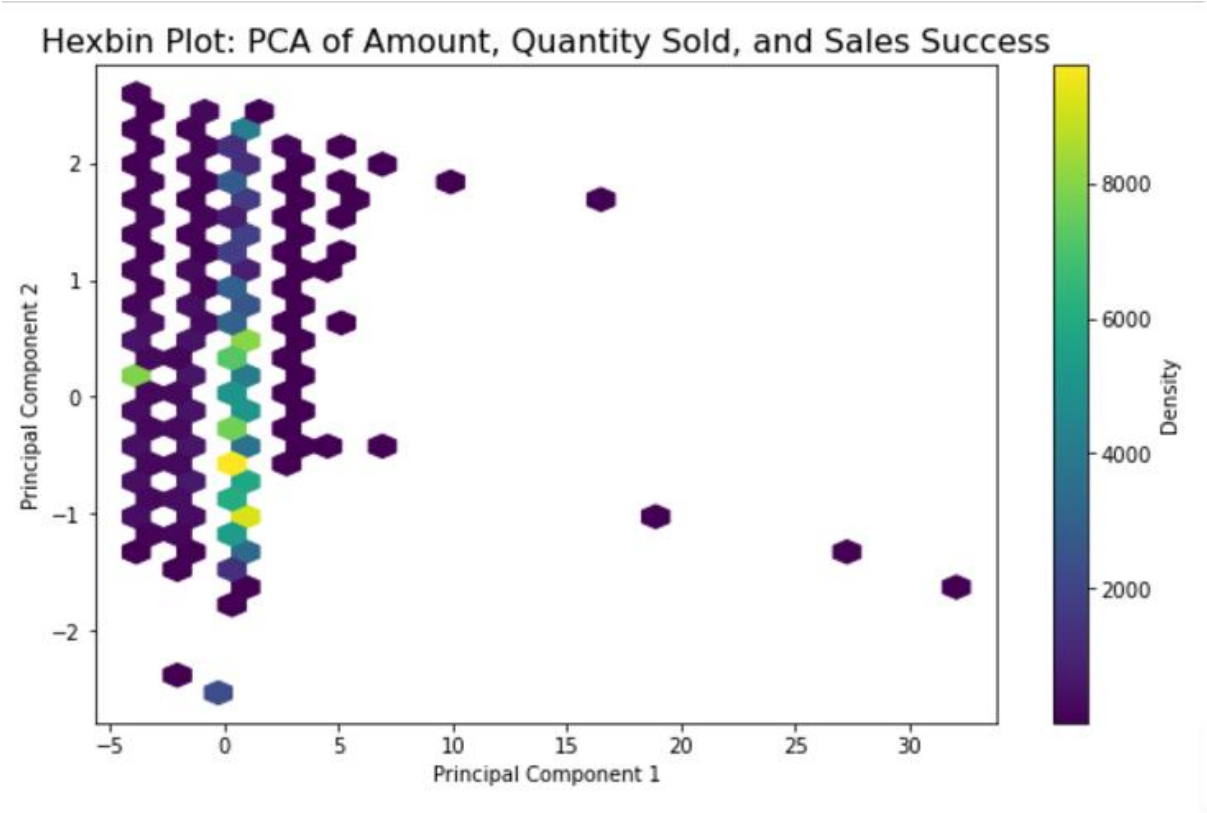
Insights:

High-density areas may indicate key thresholds for sales success, such as certain price points or quantities that lead to more successful sales.

There seems to be a relationship between Quantity and Amount when predicting Sales Success, suggesting that sales volume and pricing are important predictive features.

2. PCA: Hexbin Plot of Amount, Quantity Sold, and Sales Success

- **Purpose**: Use PCA for dimensionality reduction and visualize the principal components of Amount, Quantity Sold, and Sales Success.



Observations:

After reducing dimensions via PCA, Principal Component 1 captures the majority of variance in the data, with dense regions along this component.

Sparse regions in the plot represent outliers or less frequent data patterns.

Dense hexagons indicate high occurrence and similarity of data points.

Insights:

Principal Component 1 is likely influenced by a combination of Amount and Qty, which explains the bulk of sales trends.

Dimensionality reduction simplifies the dataset, helping to reveal dominant patterns that might be otherwise hidden.

The sparse areas could point to outliers or unique transaction patterns that might require further investigation.

Conclusion

The analysis of the dataset reveals significant insights into the sales dynamics, customer behavior, and operational performance. Univariate analysis highlights the distribution patterns of *Amount* and *Quantity Sold*, where the skewness in *Quantity* suggests most transactions involve smaller quantities, while *Amount* shows a more balanced distribution. The presence of outliers in both variables signals potential anomalies or special cases. Bivariate analysis shows that *Sales Success* is closely linked to the quantity sold, while the relationship with *Amount* is weaker. The correlation heatmap emphasizes the importance of quantity in driving sales success. Multivariate analysis, including 3D and PCA visualizations, uncovers patterns in the relationship between *Amount*, *Quantity*, and *Sales Success*, indicating that certain price ranges and quantities contribute to higher success rates. Overall, the findings suggest that focusing on optimizing quantities sold and addressing the operational challenges related to cancellations and returns can significantly improve sales performance. The insights into product category distribution and fulfillment status provide actionable strategies for improving customer satisfaction and revenue maximization.

References:

Python Programming Language

Python Software Foundation. https://www.python.org/

Pandas: Python Data Analysis Library

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, pp. 51-56. https://pandas.pydata.org/

NumPy: Fundamental Package for Scientific Computing with Python

Harris, C.R., et al. (2020). Array programming with NumPy. Nature, 585(7825), pp. 357-362. https://numpy.org/

Matplotlib: Visualization with Python

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), pp. 90-95. https://matplotlib.org/

Seaborn: Statistical Data Visualization

Waskom, M.L. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021. https://seaborn.pydata.org/

Amazon Dataset for Data Analysis

Kaggle: Amazon Sales Dataset. https://www.kaggle.com/ (Dataset source if applicable).

Principal Component Analysis (PCA)

Jolliffe, I.T. (1986). Principal Component Analysis. Springer Series in Statistics. https://doi.org/10.1007/978-1-4757-1904-8

Jupyter Notebook for Analysis and Visualization

Project Jupyter. https://jupyter.org/

Online Documentation and Tutorials

Towards Data Science Blog: https://towardsdatascience.com/

GeeksforGeeks: Python and Data Science Tutorials. https://www.geeksforgeeks.org/

Github Repositry Link:

https://github.com/AsminOthuru/Amazon-Sales-Trends-Analysis