Final Project Report

Data Visualization for the Amazon Sales Dataset

Name of the supervisor: **Miss. Lakni Peiris**

Name of the university: **NSBM Green University**

Batch: **22.1**

Module code & Name: **DS304.3 - Data Visualization**

S_ID & Name of the students: **25228 - MMSS Piyarathna**

26044 - RKNT Rajapaksha

25004 - WSR Withanage

Abstract

This project aims to analyze and visualize the sales and profitability of e-commerce products using the Amazon sale dataset, which contains data on various attributes such as SKU Code, Style, Category, Status, Quantity, Currency, Amount, etc. The project has objectives like, to identify the best-selling and most profitable products and categories, to compare the profitability of sales made through different sales channels and platforms, and to forecast future sales and profits by analyzing the data. The project will use various data preparation and data wrangling techniques, such as data cleaning, data transformation, and data reduction, to ensure the quality and reliability of the data. This project will use Power BI as the main tool to create an interactive and dynamic dashboard that includes charts that present the data in an effective and engaging way. The project will provide valuable insights and recommendations for companies to optimize their product profitability and customer satisfaction by using visualization. This project also can be a help for discussing the implications and limitations of the analysis and suggest directions for future improvements in sales.

Table of Contents

Abstract 3
Chapter 1: Introduction 5
1.1 Background study of the selected sector and reason for selecting the proposed sector. 5
1.2 Problems related to the selected area 5
1.3 Objectives of the Project 6
1.4 Expected Limitations on the visualization process pertaining to selected area. 6
1.5 Proposed work schedule 7
Chapter 2 - Literature Review 8
2.1 Introduction to the dataset selected. 8
2.2 Table for Variables, their definitions, and sources 8
Chapter 3: Data Preparation Process - Data Preprocessing and Data Wrangling 9
3.1 Data Blending and Integration 9
3.2 Data Cleaning 9
3.3 Data Transformation 9
3.4 Data reduction 9
Chapter 4: Methodology 10
4.2 Type of data to be collected and data sources. 10
4.3 Data collection tools and plan 10
4.4 Methods, techniques, and tools used to analyze the data 11
Chapter 5: Final dashboard 12
Order Quantity by the Category 13
Order Quantity by Sizes 15
Order Distribution by Cities 16
Profits by Status 17
Order Fulfilments 19
Statistical Analysis 20
Conclusion 23

Chapter 1: Introduction

1.1 Background study of the selected sector and reason for selecting the proposed sector.

E-commerce is a highly competitive industry with global sales of 5.8 trillion U.S. dollars in 2023 and expected to have a 39 percent growth with expectations to surpass eight trillion dollars by 2027. Despite its rapid growth, e-commerce faces challenges like high customer expectations, low margins, intense competition, and complex logistics.

Data visualization can help e-commerce businesses analyze their profitability and customer insights by identifying profitable products, channels, and platforms, understanding customer segments, monitoring sales performance, inventory, and logistics, comparing performance with competitors, and communicating findings to stakeholders and customers that can help for the related improvements.

1.2 Problems related to the selected area

There are many problems that can be addressed by using visualization in the e-commerce sector.

- Improve customer experience - by providing interactive and engaging visualizations of products, reviews, ratings, recommendations, and feedback.
- Increase sales and conversions - by using visualizations to highlight the benefits, features, and value propositions of products, as well as to create urgency, scarcity, and social proof.
- Optimize marketing campaigns - by using visualizations to analyze the performance, effectiveness, and return on investment of various marketing channels, strategies, and tactics.

- Reduce fraud and theft - by using visualizations to detect and prevent fraudulent transactions, accounts, and activities, as well as to monitor and secure the delivery and shipment of products.
- Enhance inventory and supply chain management - by using visualizations to track and forecast the demand, supply, and availability of products, as well as to optimize the inventory levels, costs, and logistics.

1.3 Objectives of the Project

Data visualization is the process of creating graphical representations of information, such as charts, graphs, maps, and other visual tools. The objectives of data visualization are to help communicate data in a way that is easy to understand, to reveal patterns, trends, and outliers in the data, and to support data-driven decision-making. Data visualization can also make data more accessible and engaging for different audiences.

Using a dataset that includes sales product, amount, quantity, shipped country, and other variables, you can apply various data visualization techniques to achieve different objectives. So therefore, our main objective is to develop a comprehensive dashboard within Power BI that visually presents key metrics like profitability, revenue per unit, total sales, etc.

1.4 Expected Limitations on the visualization process pertaining to selected area.

This project may encounter the following limitations or challenges on the visualization process pertaining to the selected area:

- **Quality, availability, or reliability of the data**: The project will use the Amazon Sale Report.csv data set, which may contain errors, missing values, outliers, or inconsistencies

that could affect the accuracy and validity of the analysis and visualization. The project will need to perform data cleaning, validation, and transformation to ensure the quality, availability, and reliability of the data.

- **Ethical or legal issues related to data privacy and security**: The project will deal with sensitive and confidential data, such as the customer information, the product details, and the sales transactions, which could pose ethical or legal risks if not handled properly. The project will need to comply with the relevant data protection laws and regulations, such as the General Data Protection Regulation (GDPR) and ensure the data privacy and security of the e-commerce business and its customers. The project will also need to obtain the consent and permission of the data owner and the data subjects before using the data for the project.

- **Technical or design constraints of the visualization tools**: The project will use Power BI as data visualization tool to create and present the data visualizations. The project may face technical or design constraints, such as the compatibility, functionality, or usability of the tools, that could limit the creativity and effectiveness of the visualization.

1.5 Proposed work schedule

As we are done choosing and collecting the dataset and completing the project proposal, now we are focused on completing the dashboard and the final report by 1st of May.

Chapter 2 - Literature Review

2.1 Introduction to the dataset selected.

The e-commerce industry has seen tremendous growth in recent years, with online sales making up an increasingly large portion of the retail market. As a result, understanding the sales transactions of e-commerce companies has become a critical area of research. In this project, we focus on analyzing the sales transactions of an e-commerce company that sells fashion products on Amazon.

Our dataset for this project is the Amazon Sale Report.csv file, which contains 128 976 rows and 24 columns of information about sales transactions. The dataset covers the period from April 2022 to June 2022 and includes sales from India.

2.2 Table for Variables, their definitions, and sources

The variables in the dataset provide a wealth of information about each sale transaction. Category indicates the type of product sold, while Size indicates the size of the product. Date is the date of the sale transaction, and Status indicates whether the transaction was confirmed, cancelled, or returned. Fulfilment provides information about how the sale was fulfilled, and Style indicates the style of the product sold. SKU and ASIN are unique identifiers for each product sold, while Courier Status indicates the status of the courier service that delivered the product. Qty and Amount provide information about the quantity and amount of the sale transaction, respectively. B2B indicates whether the sale was made to another business or to an individual customer, and Currency indicates the currency used for the sale transaction.

Chapter 3: Data Preparation Process - Data Preprocessing and Data Wrangling

We did the whole process of data preparation by using inbuilt features in PowerBI.

3.1 Data Blending and Integration

Since we only used one data source, we do not need to blend or integrate data from various sources, formats, or structures.

3.2 Data Cleaning

We identified and resolved several data quality issues, such as missing values, outliers, duplicates, errors, or inconsistencies.

3.3 Data Transformation

We are going to modify or transform the data to make it suitable for analysis or visualization. We had to change some data types that are not suitable for the columns. and we had to split some columns and extract the values that we only needed.

3.4 Data reduction

It's a crucial step in data analysis and visualization that involves reducing the size or complexity of the data to make it more manageable and efficient. There are different methods of data reduction, including sampling and filtering, each with its purpose and benefits. Filtering involves applying criteria or thresholds to the columns to remove irrelevant, incomplete, or unreliable data. So we did the filtering part by using the PowerBI inbuilt option.

Chapter 4: Methodology

4.1 Introduction

This chapter describes the methods, techniques, and tools used to collect, analyze, and visualize the data for the project, and how they relate to the research questions and hypotheses. The project follows a quantitative and cross-sectional approach, using secondary data from the Amazon sale dataset, which contains data on various sales channels, products, and platforms. The project uses Power BI, a business intelligence, and data visualization tool, to perform the data preparation, analysis, and visualization tasks.

4.2 Type of data to be collected and data sources.

The type of data to be collected for the project is secondary data, which is data that has been collected by someone else for a different purpose. The data source for the project is the Amazon sale dataset, which is available on Kaggle, a platform for data science and machine learning. The dataset provides an in-depth look at the profitability of e-commerce sales. It contains data on various sales channels, including Ship rocket and INCREFF, and financial information on related expenses and profits. The dataset has 128 976 rows and 24 columns and is in CSV format.

4.3 Data collection tools and plan

We used a secondary dataset. Therefore, we do not have any specific information about the tools that the publisher uses to collect the data.

4.4 Methods, techniques, and tools used to analyze the data

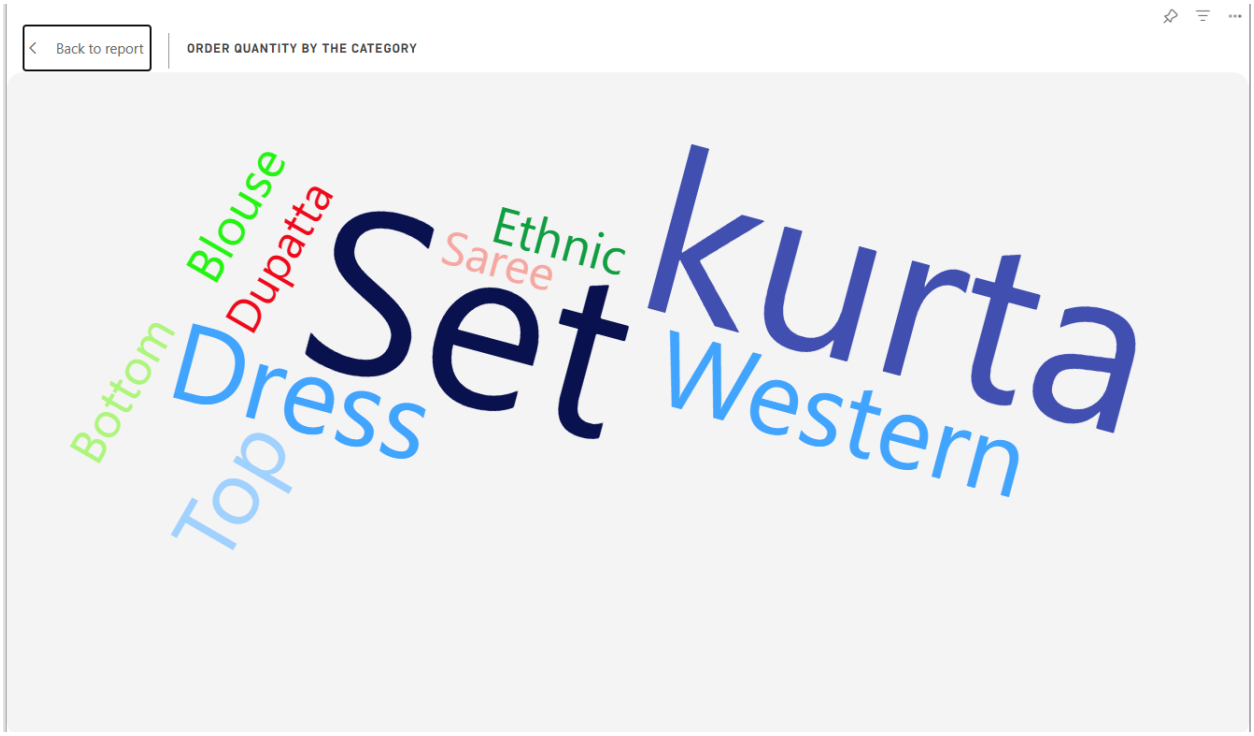
Data analysis involves many methods and techniques such as descriptive statistics, inferential statistics, hypothesis testing, correlation, regression, classification, clustering, and machine learning. These techniques are used to summarize, explore, and model the data to answer research questions and hypotheses. However, it is important to select the appropriate methods and techniques for each scenario based on the type, level, and distribution of the data, and the objectives and assumptions of the analysis.

To apply these methods and techniques, Power BI Desktop and Power BI service tools are used. These tools allow users to perform various data analysis functions, such as calculating, aggregating, filtering, sorting, grouping, or pivoting the data. Once the data has been analyzed, the results and outcomes are interpreted using the same tools, which allow users to visualize and explore the data analysis outputs, such as tables, charts, graphs, maps, or indicators.

Chapter 5: Final dashboard

Dashboard link- https://app.powerbi.com/links/8AAmilNEIV?ctid=9486ac65-39d3-4d25-977c-76d9c31c0046&pbi_source=linkShare

Order Quantity by the Category



Set	kurta	Western dress	Top	Ethnic	Blouse	Bottom	Saree	Dupatta

Possible Interpretations

- **Dominant Category:** "set" is the largest word, indicating it is the most popular category among your customers. This suggests a strong demand for “set” dress products.

- **Other Categories:** The remaining categories ("Kurtha", “ethnic” "Western", "Dress", "Blouse", "Dupatta") represent other categories that customers are ordering, but less than " set".

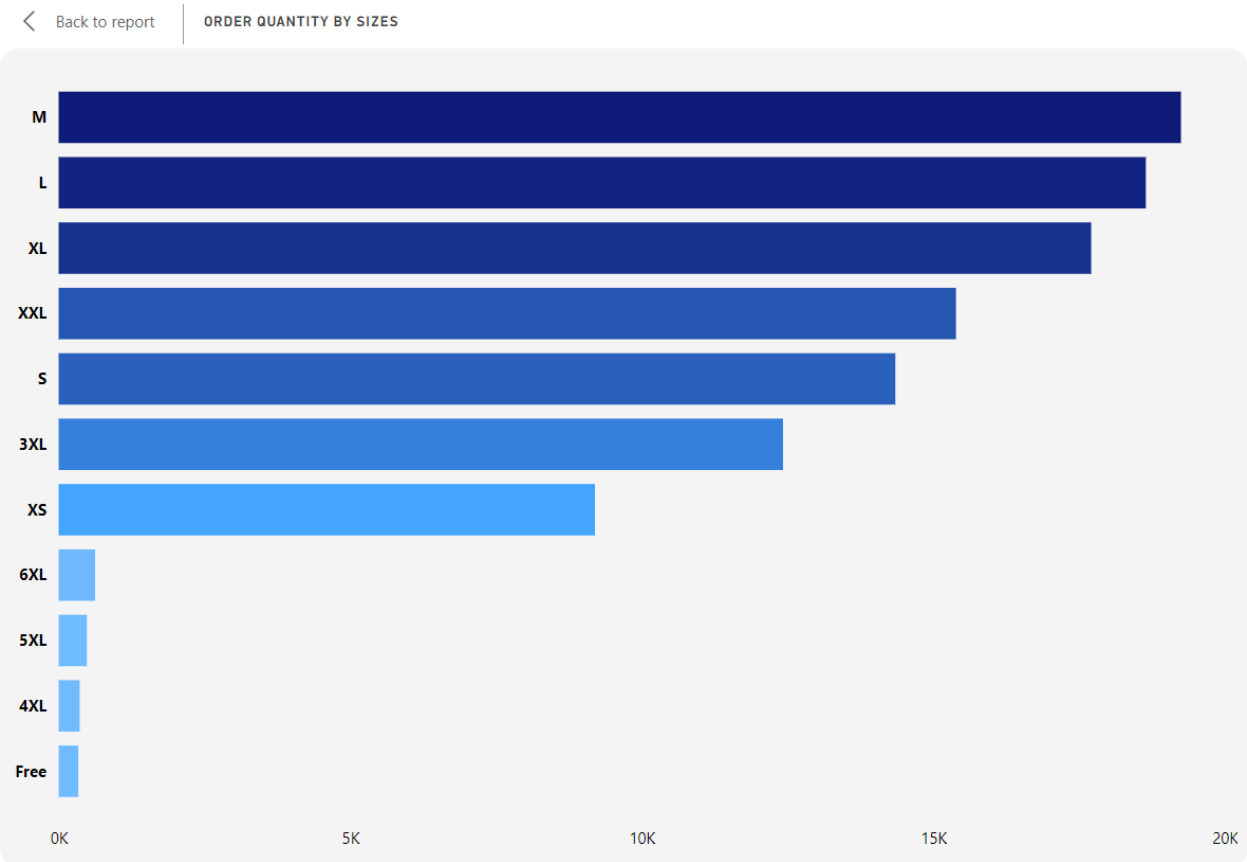
Business Insights

This word cloud offers valuable customer insights to inform the company’s inventory management and marketing strategies:

- **Prioritize Inventory:** Given the high demand for "set", it's crucial to prioritize stocking this category to avoid stockouts and ensure customer satisfaction.
- **Targeted Marketing:** Leverage marketing efforts to promote other categories (like "Western" or "Dress") that might have lower visibility but still generate sales. This could involve targeted social media campaigns or promotions.

By analyzing the word cloud and understanding customer preferences, you can make data-driven decisions to optimize your product offerings and marketing strategies.

Order Quantity by Sizes

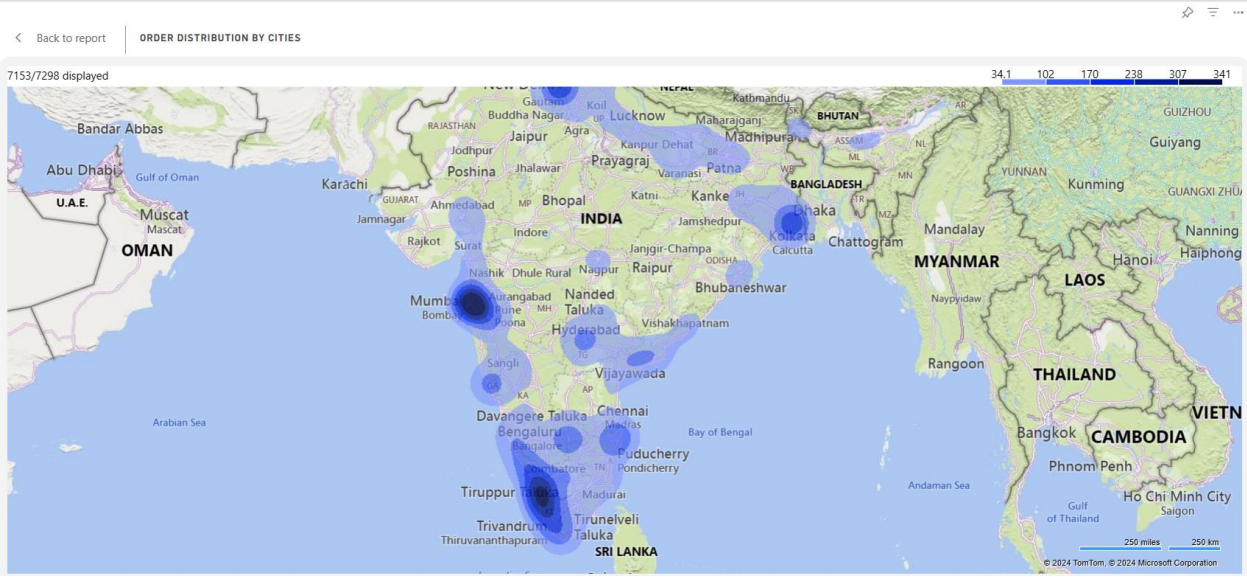


We considered using a horizontal bar chart to implement the size distribution in our dashboard. This chart reveals the variety of sizes that customers are ordering.

Insights:

- **Inventory Management:** This plot data can be valuable for optimizing inventory management. By understanding which sizes are in higher demand, you can:
    - Prioritize stocking these popular sizes to avoid stockouts.
    - Adjust ordering quantities for less popular sizes to minimize excess inventory.

Order Distribution by Cities



We included a heat map to visualize the order distribution.

Business Insights

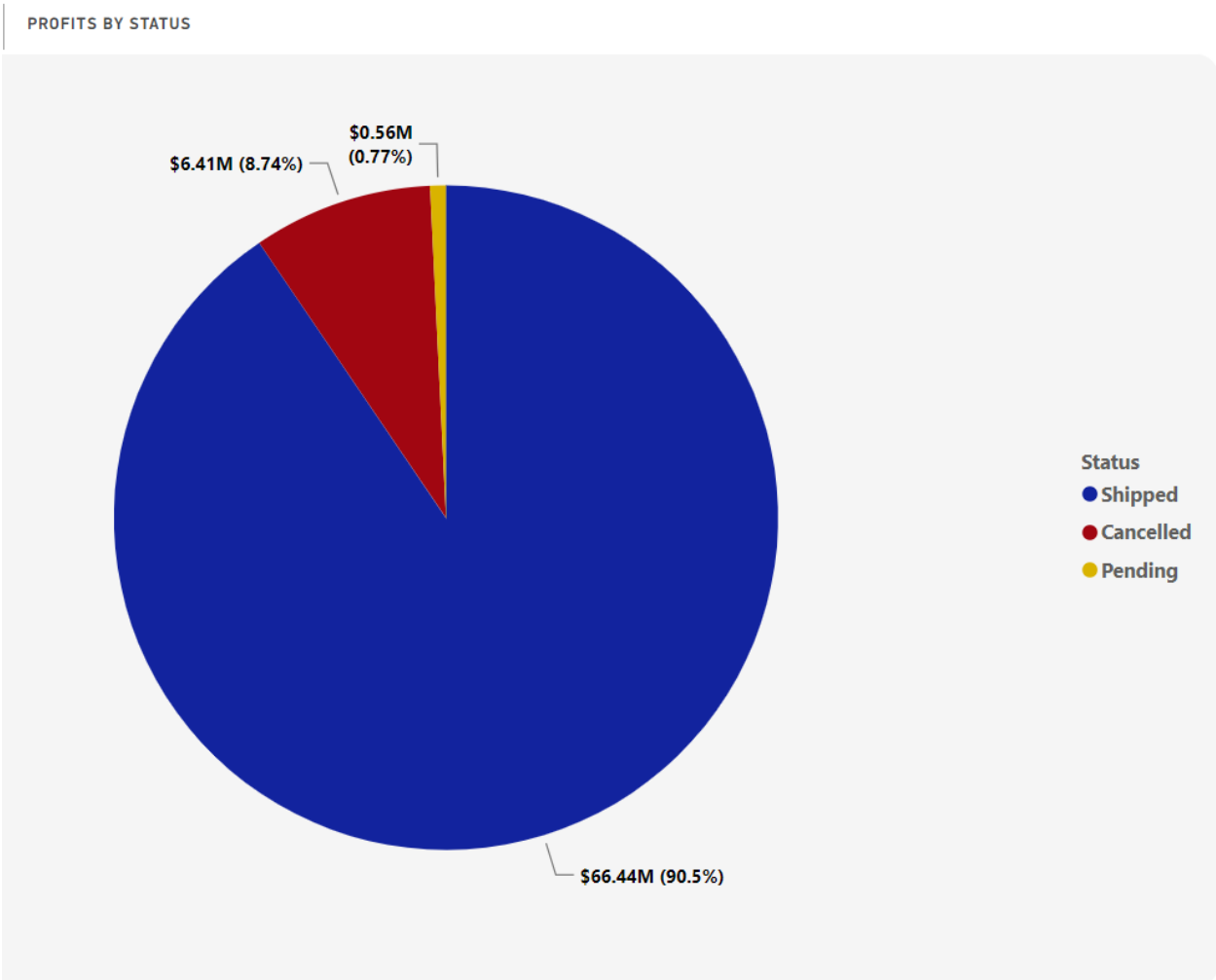
This heatmap provides valuable geographical insights into the customer base.

- Aligning Marketing Efforts: By aligning the marketing efforts with the geographical distribution revealed by the heatmap, we can ensure maximum impact and resonance with our target audience. For cities exhibiting high order volumes, we can deploy targeted social media campaigns, localized promotions, and geotargeted advertisements to capitalize on existing traction and further amplify brand visibility and engagement.

- Customer Acquisition Strategies: For cities characterized by medium or low order volumes, tailored customer acquisition strategies are imperative to unlock untapped market potential. This entails deploying dynamic marketing campaigns tailored to the unique preferences and demographics of each region. Exploring new marketing channels can facilitate deeper market penetration and drive customer acquisition in these regions.

- Resource Allocation Optimization: Strategic allocation of resources based on order volume distribution is paramount for maximizing resource efficiency. Prioritizing marketing investments in high-volume cities, where the likelihood of conversion is inherently higher, enables us to maximize sales potential and capitalize on existing market demand. Simultaneously, adopting cost-effective approaches, such as targeted email marketing or community engagement initiatives, allows us to cultivate brand awareness and stimulate demand in emerging markets with lower order volumes, thereby laying the groundwork for future growth and expansion.

Profits by Status



The pie chart reveals the breakdown of the total profits according to order fulfillment status.

- **Shipped Orders:** The largest slice of the pie chart (colored blue) represents shipped orders, which account for 90.5% or $66.44 million of your total profits. This indicates that most of your profits come from successfully fulfilled orders.
- **Cancelled Orders:** The smaller slice (colored red) represents cancelled orders, contributing 8.74% or $6.41 million to your total profits. While cancelled orders do generate some revenue, they can also signal potential issues with product availability, pricing, or customer service.

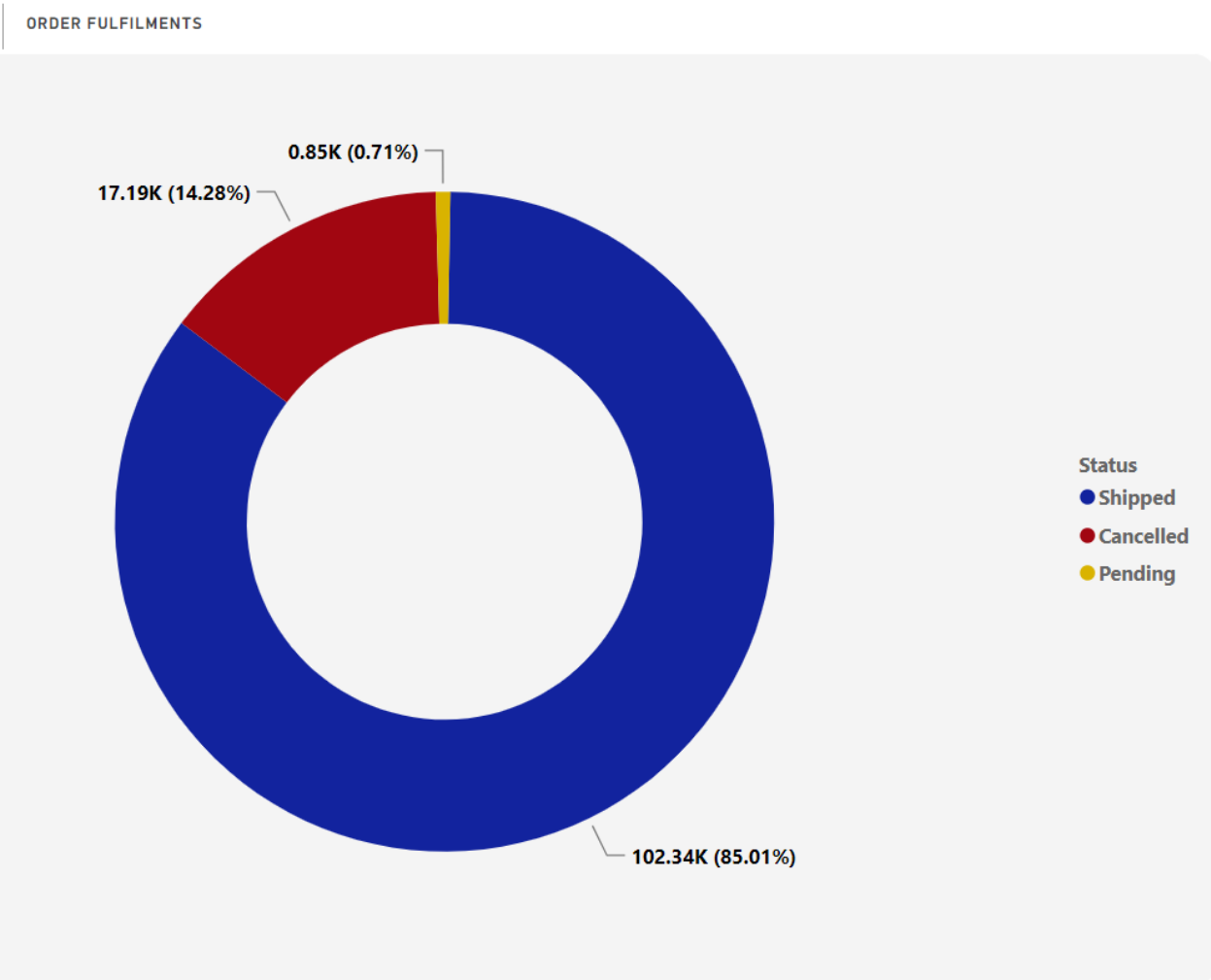
Business Insights

Hence there are considerable amount of order cancellations we can suggest considering the following strategies to keep up the customer satisfaction on a level to get the maximum profitability through the online platform.

Dynamic Pricing Optimization: Utilizing data analytics and competitor benchmarking, we can refine dynamic pricing strategies to adapt swiftly to market dynamics while ensuring profitability. This approach will help mitigate potential customer dissatisfaction arising from perceived pricing discrepancies.

Streamlined Online Customer Support: Enhancing the online customer support infrastructure, like responsive email support, will enable swift resolution of customer queries and concerns, thereby reducing the likelihood of order cancellations due to unresolved issues.

Order Fulfilments



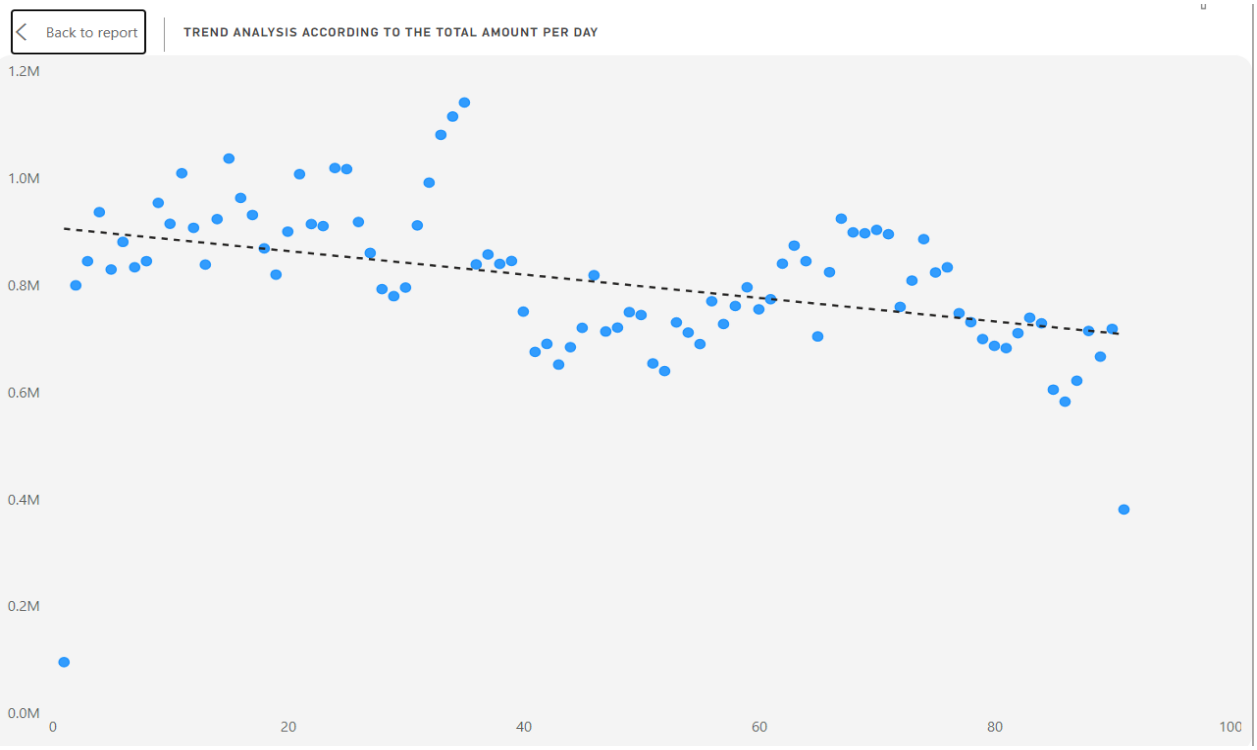
- **Shipped** orders: 85.01% (102,340 out of 120,000 total orders)
- **Cancelled** orders: 14.28% (17,190 orders)
- **Pending** orders: 0.71% (850 orders)

Order Fulfillment Analysis

- High Fulfillment Rate: The substantial portion of the pie chart, representing shipped orders at 85.01%, underscores the efficiency of our fulfillment process. This high fulfillment rate is indicative of the commitment to promptly delivering products to customers, thereby fostering positive experiences and bolstering customer satisfaction.

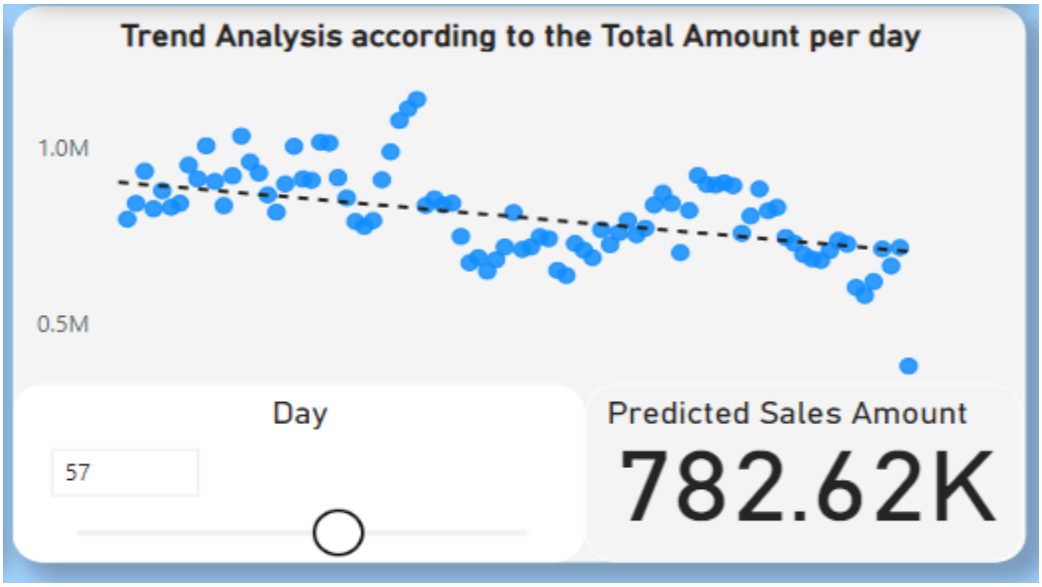
- Cancellations Analysis: Contrary to the initial assessment, the revised data reveals that cancelled orders constitute a substantial portion of our order fulfillment breakdown, accounting for 14.28% of total orders or 17,190 orders. While this figure is higher than previously indicated, it's imperative to delve into the underlying reasons for these cancellations. By analyzing customer feedback and identifying recurring issues, such as inventory discrepancies or fulfillment errors, we can implement corrective measures to minimize cancellations and preserve revenue streams.
- Addressing Pending Orders: With pending orders now comprising a minimal 0.71% of total orders or 850 orders, it's evident that our focus should primarily be directed towards mitigating cancellations and optimizing order fulfillment efficiency. Strategies to achieve this may include refining inventory management practices, enhancing communication channels with customers regarding order status, and implementing measures to expedite order processing and shipment.

Statistical Analysis



Trend Analysis According to The Total Amount Per Day

We did the statistical analysis by using linear regression. This model considered ordered date count as the independent variable and the daily order amount as the dependent variable. By this, we tried to predict the daily sales amount over the given time period.



And we implemented a feature to predict the future sales amount according to the day count.

Sales Trends

In reviewing the sales trends depicted by the line graph, a more nuanced understanding emerges when applying linear regression analysis, revealing insights pivotal for informed decision-making.

Overall Decreasing Sales Trend: Contrary to the initial interpretation, the linear regression analysis indicates a downward slope in total daily sales over the specified period. This signifies a consistent decline in total sales rather than an increase, highlighting the need for proactive measures to address this negative trend.

Business Insights and Strategic Considerations:

Marketing Campaign Evaluation: Assessing the impact of past marketing campaigns on sales performance becomes crucial in optimizing future marketing strategies. By correlating periods of

increased sales with corresponding marketing initiatives, we can discern the effectiveness of different promotional tactics and allocate resources more efficiently in future campaigns.

Forecasting Future Sales: Leveraging historical sales data in conjunction with linear regression analysis enables us to extrapolate future sales trends and anticipate demand fluctuations. This facilitates proactive inventory management, ensuring optimal stock levels to meet customer demand while minimizing excess inventory costs.

Conclusion

In conclusion, our project has delved into the realm of e-commerce, specifically focusing on analyzing and visualizing sales data from Amazon using advanced data visualization techniques. Through the development of an interactive dashboard powered by Power BI, we've not only uncovered valuable insights but also provided actionable recommendations for businesses seeking to optimize their strategies.

Our journey wasn't without its challenges; we encountered data quality issues and navigated ethical considerations along the way. However, by persevering through these obstacles, we emerged with a clearer understanding of product profitability, sales performance, and customer behavior.

The insights gleaned from our analysis have far-reaching implications for e-commerce businesses. By identifying best-selling products, discerning order fulfillment patterns, and tracking sales trends, we've equipped businesses with the knowledge needed to enhance customer experiences, increase sales, and refine inventory management practices.

Looking ahead, the significance of leveraging data visualization tools and techniques in the e-commerce landscape cannot be overstated. In an era defined by rapid digital transformation, the ability to extract meaningful insights from complex datasets is paramount for making informed decisions and sustaining growth.

Ultimately, our project serves as a testament to the power of data visualization in unraveling the intricacies of e-commerce operations. By shedding light on the nuances of sales data, we've provided a pathway for businesses to navigate the ever-evolving landscape of online commerce with confidence and foresight. As we conclude this endeavor, we do so with a sense of accomplishment, knowing that our contributions will catalyze greater understanding and success within the e-commerce industry.