ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that requires continuous monitoring for effective treatment. This data science project aims to predict the total Unified Parkinson's Disease Rating Scale (UPDRS) score using a telemonitoring dataset. The dataset includes clinical features collected from PD patients through wearable devices and mobile applications. The project employed a comprehensive data science workflow, starting with data preprocessing and exploratory data analysis to gain insights into the dataset's characteristics. Feature engineering techniques were applied to extract meaningful information and enhance the predictive power of the model. Various machine learning algorithms, including Linear Regression, Support Vector Machine (SVM), and Random Forest Regressor, were trained and evaluated to identify the most accurate model for total UPDRS prediction. Evaluation metrics such as mean squared error and R-squared were used to assess model performance. The results demonstrated promising accuracy in estimating the total UPDRS score, enabling remote assessment of PD severity. This project contributes to the field by showcasing the potential of machine learning techniques in leveraging telemonitoring data for personalized treatment plans and timely interventions.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned my effort with success.

I would like to thank **Shri Narayan Rao R. Maanay,** Secretary, BNMEI, Bengaluru for providing an excellent academic environment in the College.

I would like to sincerely thank **Prof. T. J. Rama Murthy**, Director, BNMIT, Bengaluru, for having extended his support and encouraging me during the course of the work.

I would like to sincerely thank **Dr. S.Y. Kulkarni**, Additional Director, BNMIT, Bengaluru for having extended his support and encouraging me during the course of the work.

I would like to express my gratitude to **Prof. Eishwar N. Maanay**, Dean, BNMIT, Bengaluru for his relentless support, guidance and assistance.

I would like to thank **Dr. Krishnamurthy G.N,** Principal, BNMIT, Bengaluru for his constant encouragement.

I would like to thank **Dr. Sheba Selvam,** Professor and Head of the Department of Artificial Intelligence and Machine Learning, BNMIT, Bengaluru whohas shared her opinions and thoughts which helped me in completion of my project successfully.

I would also like to thank my Course teacher **Dr. Sunitha R,** Assistant Professor, Department of Artificial Intelligence and Machine Learning, BNMIT, Bengaluru for guiding in a systematic manner.

Finally, I would like to thank all technical and non-technical faculty members of Department of Artificial Intelligence and Machine Learning, BNMIT, Bengaluru, for their support. I would like to thank my Family and Friends for their unfailing moral support and encouragement.

SIDDHANT PRIYADARSHI

1BG20AI087

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1. Overview of Data Science

Data science is an interdisciplinary field that uses statistics, mathematics, computer science, and domain expertise to extract insights from data. It involves collecting, processing, analyzing, and interpreting large datasets to solve real-world problems and make informed decisions. The key steps include data collection, cleaning, exploratory analysis, feature engineering, model selection, training, evaluation, deployment, monitoring, and communication. Data scientists need to consider ethical implications and biases in their work. The field has applications in various industries and aims to improve efficiency and enable data-driven decision-making.

1.2. Problem Statement

The goal of this project is to analyze the Amazon sales report data to identify key trends, patterns, and insights that can inform strategic decision-making. By examining factors such as product performance, customer behavior, and market dynamics, the analysis aims to optimize sales strategies, improve product offerings, and enhance overall revenue generation for Amazon.

1.3. Objectives

- Develop an accurate predictive model for Amazon Sales Report.
- Perform data preprocessing and cleaning.
- Conduct exploratory data analysis (EDA) for insights.
- Apply feature engineering techniques.

- Compare and evaluate machine learning algorithms.
- Optimize the selected model.
- Identify significant predictors for Amazon Sales Report.
- Validate the model's performance.
- Provide recommendations based on the analysis.

1.4. Overview of Total Amazon Sales Prediction Project

The goal of the Total Amazon Sales Prediction project is to develop a predictive model that can accurately forecast the total sales of Amazon. By analyzing historical sales data, market trends, customer behavior, and other relevant factors, the project aims to provide insights and predictions that can assist in demand forecasting, inventory management, and overall business planning for Amazon. The objective is to improve decision-making, optimize resource allocation, and enhance the company's sales performance.

Chapter 2

DATASET DESCRIPTION

This dataset provides an in-depth look at the profitability of e-commerce sales. It contains data on a variety of sales channels, including Ship rocket and INCREFF, as well as financial information on related expenses and profits. The columns contain data such as SKU codes, design numbers, stock levels, product categories, sizes and colors. In addition to this we have included the MRPs across multiple stores like Ajio MRP , Amazon MRP , Amazon FBA MRP , Flipkart MRP , Limeroad MRP Myntra MRP and Paytm, MRP along with other key parameters like amount paid by customer for the purchase , rate per piece for every individual transaction Also we have added transactional parameters like Date of sale months category fulfilledby B2b Status Qty Currency Gross amt . This is a must-have dataset for anyone trying to uncover the profitability of e-commerce sales in today's marketplace.

- **Category –** Type of product. (String)
- **Size –** Size of the product. (String)
- **Date** – Date of the sale. (Date)
- **Status** - Status of the sale. (String)
- **Fulfilment** - Method of fulfilment. (String)
- **Style** - Style of the product. (String)
- **SKU** – Stock Keeping unit (String)
- **ASIN** - Amazon Standard Identification Number. (String)
- **Courier Status** - Status of the courier. (String)
- **Qty** - Quantity of the product. (Integer)
- **Amount** - Amount of the sale. (Float)
- **B2B** - Business to business sale. (Boolean)
- **Currency** - The currency used for the sale. (String)

Chapter 3

TOOL DESCRIPTION

Power BI is a robust business analytics tool developed by Microsoft, designed to empower organizations with data-driven decision-making capabilities. With its intuitive and user-friendly interface, Power BI enables users to connect, analyze, and visualize data from various sources, helping businesses gain valuable insights and make informed decisions.

Key Capabilities of Power BI:

Data Connectivity: Power BI allows users to connect to a wide range of data sources, including databases, spreadsheets, and cloud services, ensuring comprehensive access to data for analysis. Data Modeling and Transformation: Power BI's Power Query Editor allows users to extract, transform, and load data, enabling data cleaning, shaping, and integration to create a unified view for analysis.

Data Visualization: Power BI provides a rich set of visualization options, including charts, tables, maps, and custom visuals. Users can create interactive reports and dashboards, presenting complex information in a visually appealing and easy-to-understand manner.

Collaboration: Power BI facilitates collaboration by enabling users to create shared workspaces and collaborate on reports and dashboards. Integration with Microsoft tools like SharePoint and Teams enhances teamwork and data-driven discussions.

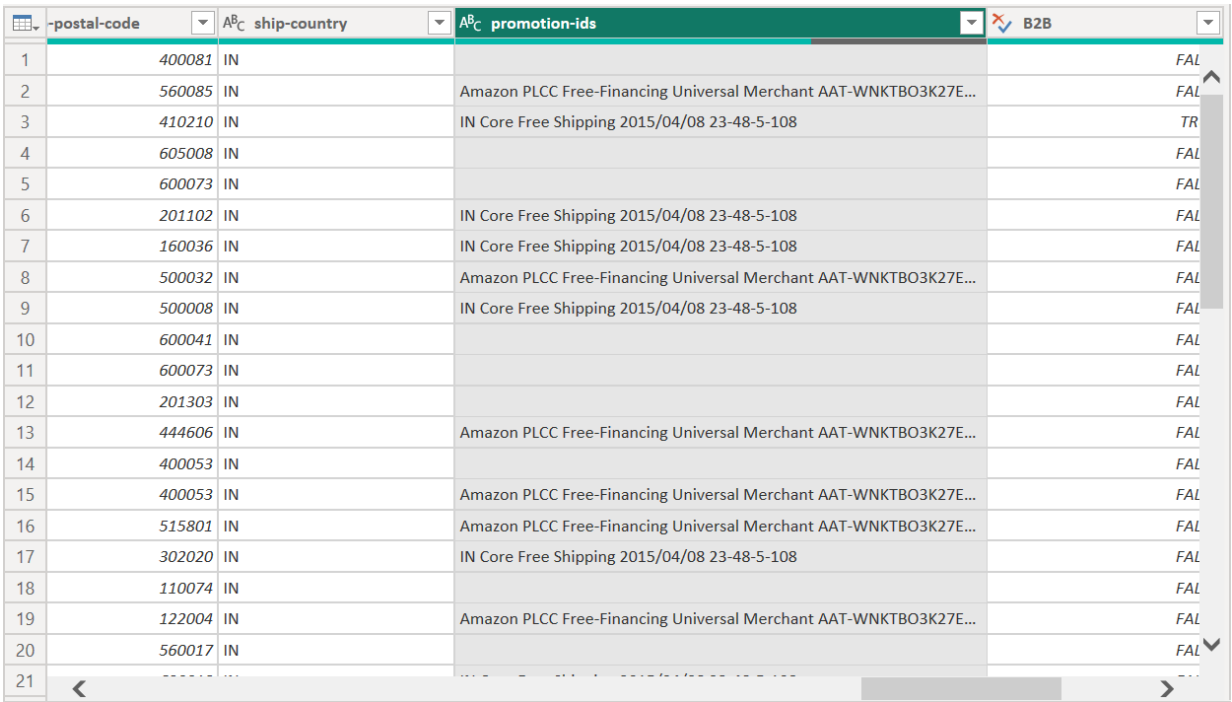
Mobile Accessibility: Power BI offers mobile apps for iOS and Android devices, allowing users to access and interact with reports and dashboards on the go, ensuring data availability and insights anytime, anywhere.

Integration: Power BI seamlessly integrates with other Microsoft tools and services, such as Azure, Excel, and SQL Server, expanding its capabilities and providing a unified analytics ecosystem.

Chapter 4

PREPROCESSING DATASET

4.1. Changing of Data type in Promotion-ids and Headers for the Columns:



	-postal-code	ship-country	promotion-ids	B2B
1	400081	IN		FAL
2	560085	IN	Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FAL
3	410210	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	TR
4	605008	IN		FAL
5	600073	IN		FAL
6	201102	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	FAL
7	160036	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	FAL
8	500032	IN	Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FAL
9	500008	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	FAL
10	600041	IN		FAL
11	600073	IN		FAL
12	201303	IN		FAL
13	444606	IN	Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FAL
14	400053	IN		FAL
15	400053	IN	Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FAL
16	515801	IN	Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FAL
17	302020	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	FAL
18	110074	IN		FAL
19	122004	IN	Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FAL
20	560017	IN		FAL
21				

Fig 1. Datatypes are changed

Code = Table.TransformColumnTypes(#"Use First Row as Headers",{{"index", Int64.Type}, {"Order ID", type text}, {"Date", type date}, {"Status", type text}, {"Fulfilment", type text}, {"Sales Channel ", type text}, {"ship-service-level", type text}, {"Style", type text}, {"SKU", type text}, {"Category", type text}, {"Size", type text}})

In the given dataset, the rows representing index, Order ID, Date, Status, Fulfilment, Sales Channel, ship-service-level and Style datatypes are changed accordingly for the Amazon Sales.

4.2. Removal Null values in Promotion-Ids :

This is the most useful step of preprocessing needed for training a model efficiently. We can use strategies to deal with missing values. We shall be using Mean for our dataset.

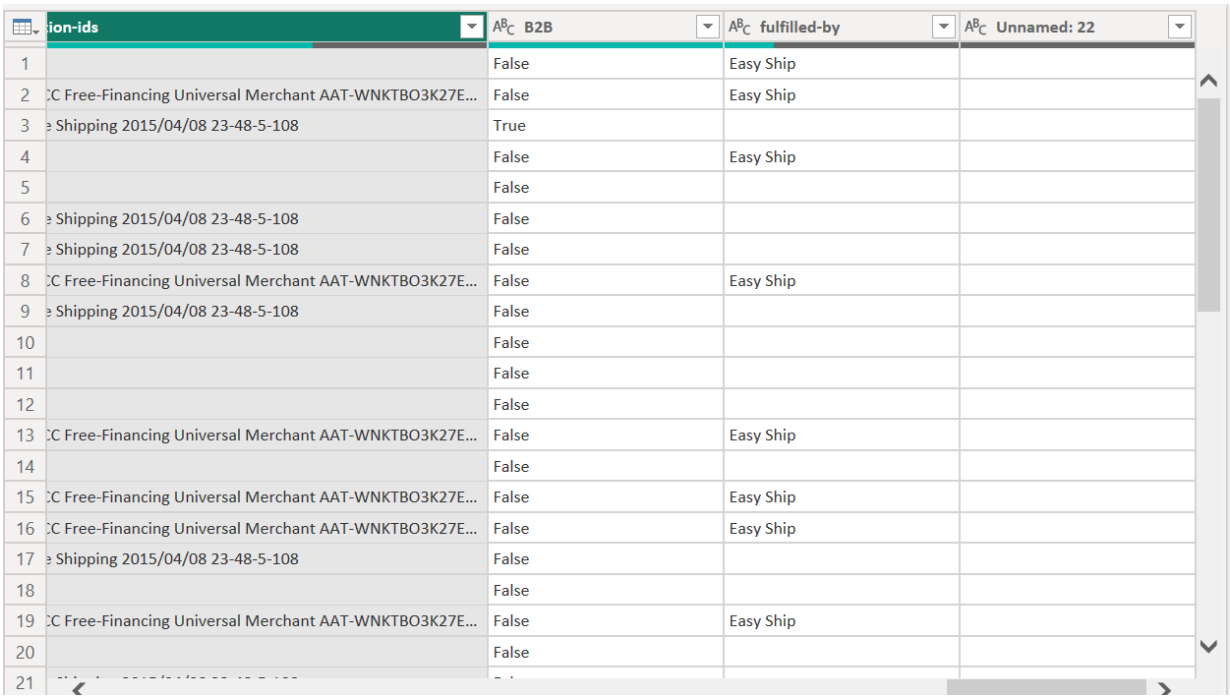
	-country	promotion-ids	B2B	fulfilled-by
1			FALSE	Easy Ship
2		Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FALSE	Easy Ship
3		IN Core Free Shipping 2015/04/08 23-48-5-108	TRUE	
4			FALSE	Easy Ship
5			FALSE	
6		IN Core Free Shipping 2015/04/08 23-48-5-108	FALSE	
7		IN Core Free Shipping 2015/04/08 23-48-5-108	FALSE	
8		Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FALSE	Easy Ship
9		IN Core Free Shipping 2015/04/08 23-48-5-108	FALSE	
10			FALSE	
11			FALSE	
12			FALSE	
13		Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FALSE	Easy Ship
14			FALSE	
15		Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FALSE	Easy Ship
16		Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FALSE	Easy Ship
17		IN Core Free Shipping 2015/04/08 23-48-5-108	FALSE	
18			FALSE	
19		Amazon PLCC Free-Financing Universal Merchant AAT-WNKTBO3K27E...	FALSE	Easy Ship
20			FALSE	
21				

Fig 2. Removing for NULLs

Code = Table.ReplaceValue(#"Replaced Value","NULL","",Replacer.ReplaceText,{"promotion-ids"})

To clean the data in the dataset, the rows that contained null values for String attributes were identified as impure and removed. By removing these rows, the dataset was cleaned, ensuring that only valid and complete data remained for analysis.

4.3. Removal of Empty columns :



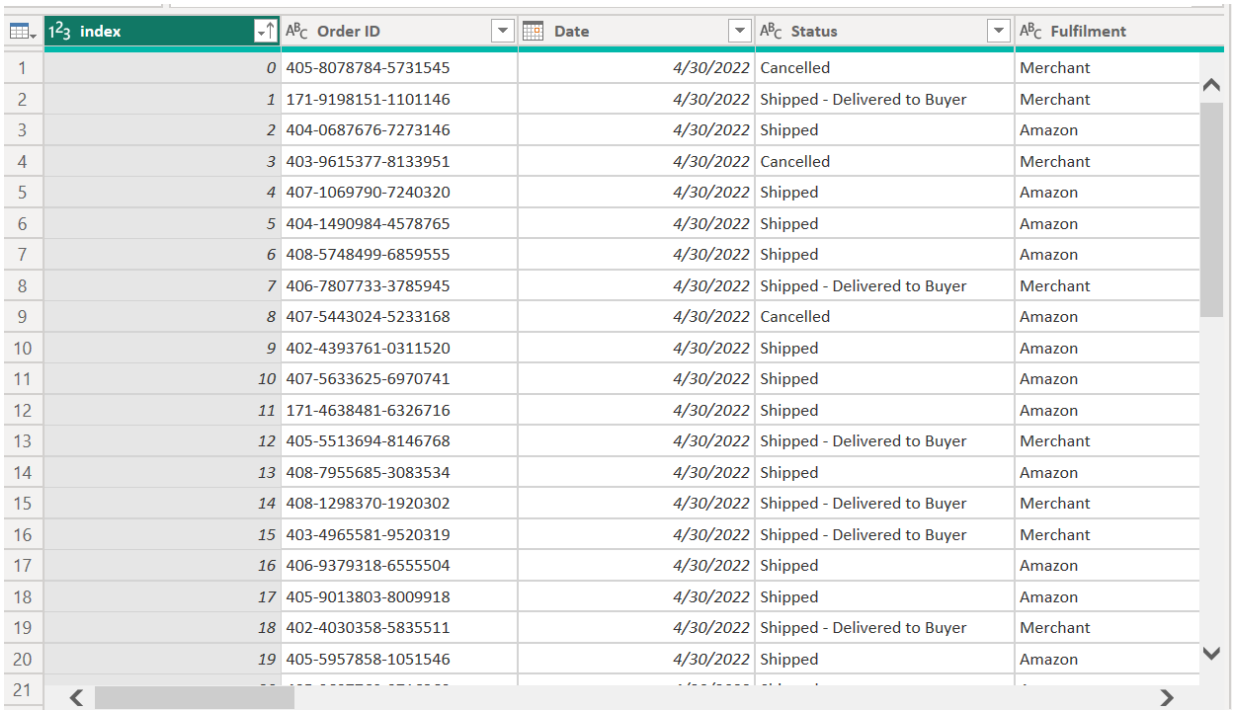
	ion-ids	B2B	fulfilled-by	Unnamed: 22
1		False	Easy Ship	
2	C Free-Financing Universal Merchant AAT-WNKTBO3K27E...	False	Easy Ship	
3	e Shipping 2015/04/08 23-48-5-108	True		
4		False	Easy Ship	
5		False		
6	e Shipping 2015/04/08 23-48-5-108	False		
7	e Shipping 2015/04/08 23-48-5-108	False		
8	C Free-Financing Universal Merchant AAT-WNKTBO3K27E...	False	Easy Ship	
9	e Shipping 2015/04/08 23-48-5-108	False		
10		False		
11		False		
12		False		
13	C Free-Financing Universal Merchant AAT-WNKTBO3K27E...	False	Easy Ship	
14		False		
15	C Free-Financing Universal Merchant AAT-WNKTBO3K27E...	False	Easy Ship	
16	C Free-Financing Universal Merchant AAT-WNKTBO3K27E...	False	Easy Ship	
17	e Shipping 2015/04/08 23-48-5-108	False		
18		False		
19	C Free-Financing Universal Merchant AAT-WNKTBO3K27E...	False	Easy Ship	
20		False		
21				

Fig 3. Removing Empty Columns

Removing empty columns from a dataset ensures data cleanliness, improves memory and computational efficiency, enhances data quality, and simplifies data exploration. It eliminates clutter, reduces storage requirements, removes irrelevant calculations, and allows focus on meaningful features, leading to accurate analysis and better decision-making.

4.4. Arranging the dataset in increasing order of index :

The dataset was subjected to an ascending sort operation based on the index. This sorting process was performed to achieve a clean and organized representation of the data in the analysis model. By arranging the dataset in chronological order, it enables easier interpretation and analysis of the data's temporal patterns and trends.



	index	Order ID	Date	Status	Fulfilment
1	0	405-8078784-5731545	4/30/2022	Cancelled	Merchant
2	1	171-9198151-1101146	4/30/2022	Shipped - Delivered to Buyer	Merchant
3	2	404-0687676-7273146	4/30/2022	Shipped	Amazon
4	3	403-9615377-8133951	4/30/2022	Cancelled	Merchant
5	4	407-1069790-7240320	4/30/2022	Shipped	Amazon
6	5	404-1490984-4578765	4/30/2022	Shipped	Amazon
7	6	408-5748499-6859555	4/30/2022	Shipped	Amazon
8	7	406-7807733-3785945	4/30/2022	Shipped - Delivered to Buyer	Merchant
9	8	407-5443024-5233168	4/30/2022	Cancelled	Amazon
10	9	402-4393761-0311520	4/30/2022	Shipped	Amazon
11	10	407-5633625-6970741	4/30/2022	Shipped	Amazon
12	11	171-4638481-6326716	4/30/2022	Shipped	Amazon
13	12	405-5513694-8146768	4/30/2022	Shipped - Delivered to Buyer	Merchant
14	13	408-7955685-3083534	4/30/2022	Shipped	Amazon
15	14	408-1298370-1920302	4/30/2022	Shipped - Delivered to Buyer	Merchant
16	15	403-4965581-9520319	4/30/2022	Shipped - Delivered to Buyer	Merchant
17	16	406-9379318-6555504	4/30/2022	Shipped	Amazon
18	17	405-9013803-8009918	4/30/2022	Shipped	Amazon
19	18	402-4030358-5835511	4/30/2022	Shipped - Delivered to Buyer	Merchant
20	19	405-5957858-1051546	4/30/2022	Shipped	Amazon
21					

Fig 4. Arranged data in serial order

Code : = Table.Sort(#"Removed Duplicates",{{"index", Order.Ascending}})

Chapter 5

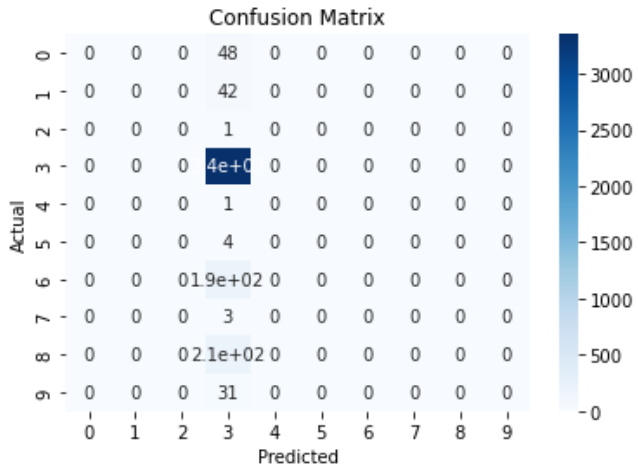
IMPLEMENTATION

Logistic regression is a statistical technique used for binary or categorical classification problems. It provides interpretable results, estimates probabilities, and determines the impact of input variables on the outcome. Logistic regression is computationally efficient, handles large datasets, and serves as a baseline model. It is valuable when interpretability, efficiency, and probabilistic predictions are essential. Due to discrete data values the curve you can see it like this. Logistic regression is a statistical modeling technique used to predict binary or categorical outcomes. It calculates the probability of an event occurring based on input variables and applies a logistic function to estimate the likelihood of belonging to a specific category.

```
data = pd.read_csv('Amazon_Sale_Report.csv', low_memory=False)

# Handle missing values
data = data.dropna()  # Remove rows with any missing values

# Select relevant columns for logistic regression
X = data[['Qty', 'Amount']]  # Features
y = data['Status']  # Target variable

# Convert categorical variables to numerical using one-hot encoding
X = pd.get_dummies(X)

# 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)

# Create a logistic regression model
logreg = LogisticRegression()

# Fit the model to the training data
logreg.fit(X_train, y_train)

# Make predictions on the test data
y_pred = logreg.predict(X_test)

# Evaluate the model using confusion matrix
confusion_mat = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(confusion_mat)

# Plot the confusion matrix as a heatmap
sns.heatmap(confusion_mat, annot=True, cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()

# Plot the logistic regression line
plt.figure()
sns.scatterplot(data=data, x='Qty', y='Amount', hue='Status', palette='coolwarm')
plt.xlabel('Qty')
plt.ylabel('Amount')
plt.title('Logistic Regression')
plt.legend()

# Plot the decision boundary
x_values = np.linspace(data['Qty'].min(), data['Qty'].max(), 100)
y_values = -(logreg.coef_[0][0] * x_values + logreg.intercept_[0]) / logreg.coef_[0][1]
plt.plot(x_values, y_values, color='black', linestyle='--')

plt.show()
```

Fig 5. Code for Logistic Regression

```
Confusion Matrix:
[[   0    0    0   48    0    0    0    0    0    0]
 [   0    0    0   42    0    0    0    0    0    0]
 [   0    0    0    1    0    0    0    0    0    0]
 [   0    0    0 3350    0    0    0    0    0    0]
 [   0    0    0    1    0    0    0    0    0    0]
 [   0    0    0    4    0    0    0    0    0    0]
 [   0    0    0  190    0    0    0    0    0    0]
 [   0    0    0    3    0    0    0    0    0    0]
 [   0    0    0  206    0    0    0    0    0    0]
 [   0    0    0   31    0    0    0    0    0    0]]
```



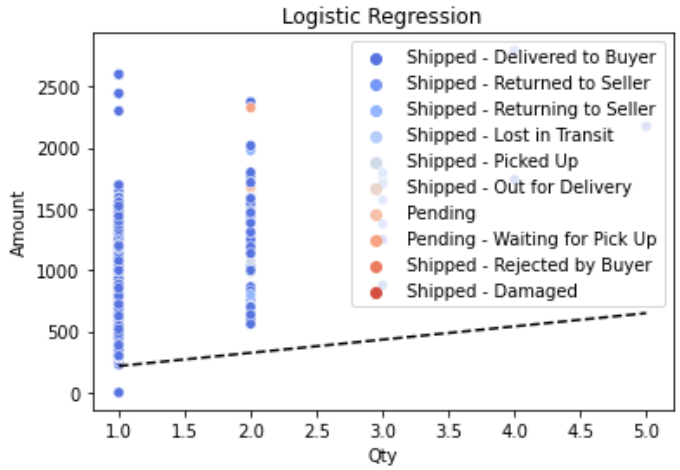


Fig 6. Output for Logistic Regression

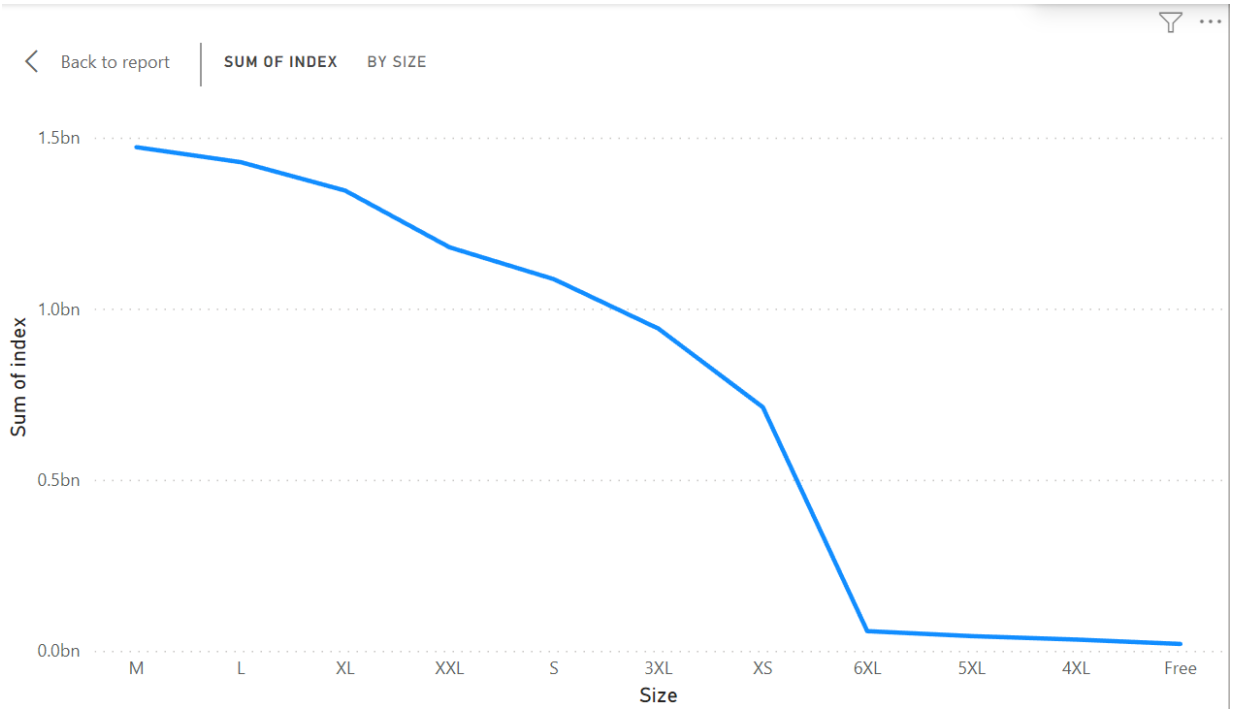


Fig 7. Graph on Sum of index and Size

The efficacy of the preprocessing step was substantiated by implementing a model and acquiring a visual representation in the form of a plot, showcasing the interplay between Sum of Index and Size. This plot serves as concrete evidence, affirming that the preprocessing stage adeptly refined the dataset for subsequent analysis.

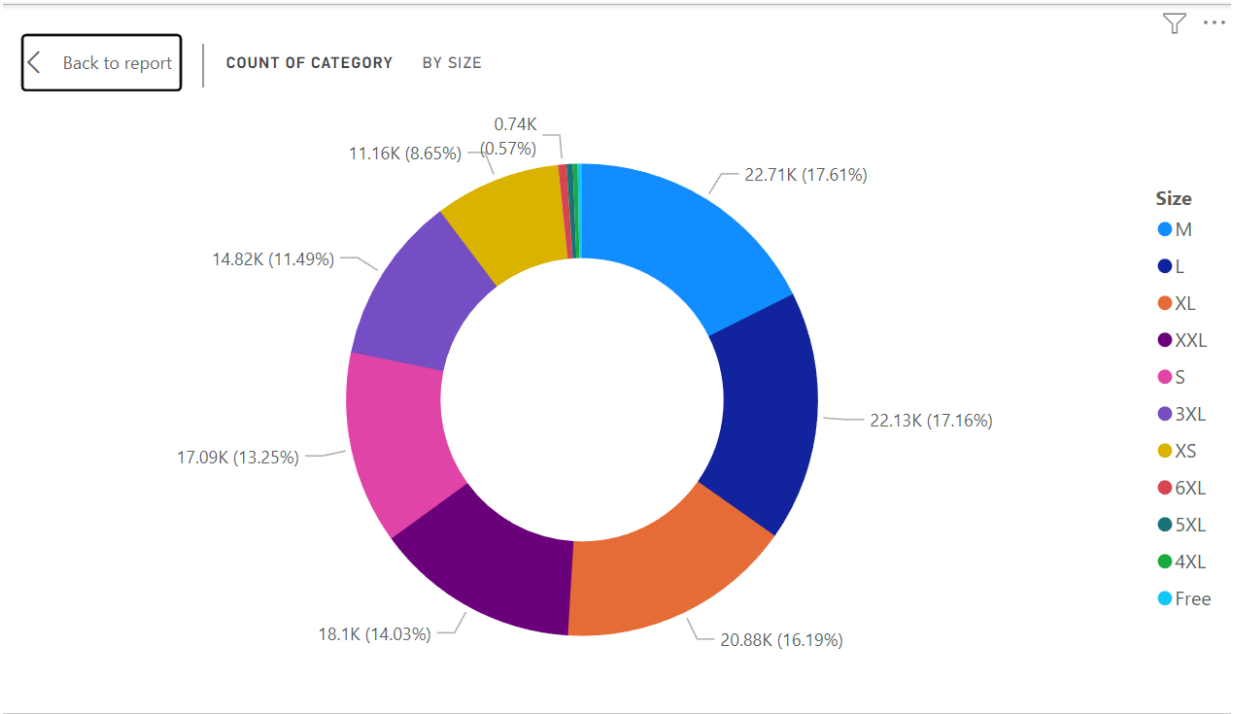


Fig 8. Count of category vs Sizes

This represents the Count of Category vs Size. The size which is taken in account is M, L, XL, XXL, S, 3XL, XS, 6XL, 5XL, 4XL, Free and for Categories are Set, Kurta, Western, Ethnic and Top. It shows the sales of category of dresses with respect to the size.

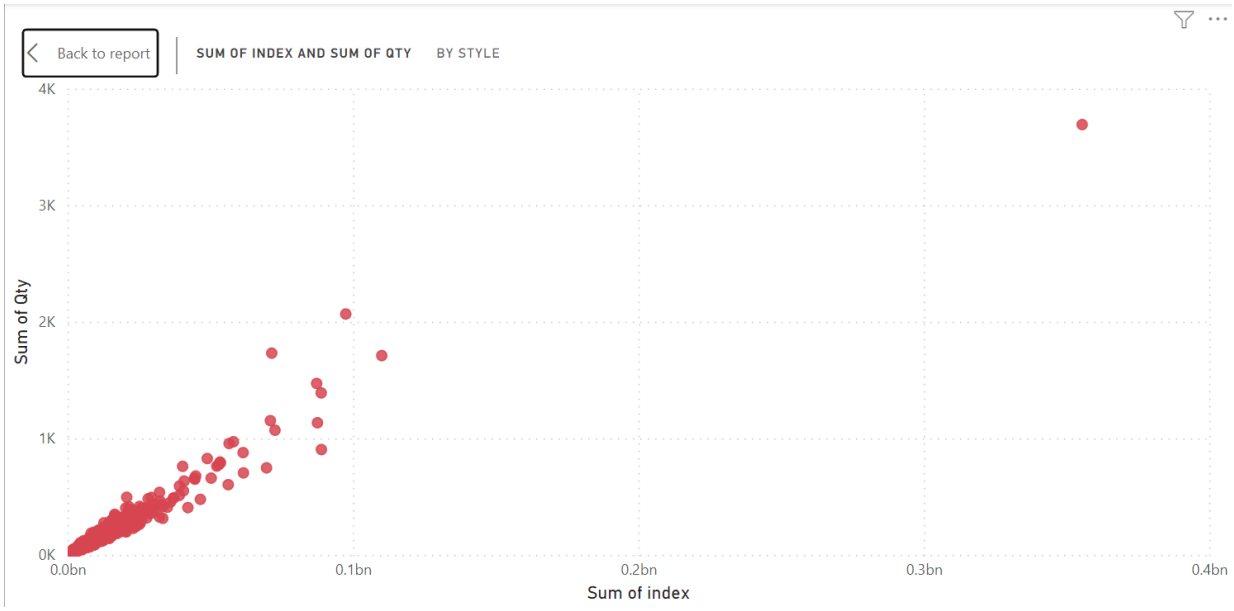


Fig 9. Sum of Qty vs Sum of Index

This represents the Sum of Quantity of each products with respect of sum of index. This data visualization clearly shows that Qty changes with change in index that is the Qty and index are linearly proportional.

Chapter 6

RESULTS

By conducting a data science analysis of Amazon sales reports, several valuable insights can be obtained. The analysis may uncover trends in product performance, identify high-demand items, and reveal patterns in customer purchasing behavior. Additionally, it can provide an understanding of market dynamics, competitor analysis, and the impact of promotional strategies on sales. Through data exploration and visualization, the analysis may highlight geographical sales patterns, seasonal fluctuations, or demographic preferences. Furthermore, the analysis can help optimize inventory management, identify cross-selling and upselling opportunities, and improve pricing strategies. Ultimately, the data science analysis of Amazon sales reports empowers decision-makers with actionable information to enhance sales performance, drive revenue growth, and make informed business decisions.

Chapter 7

CONCLUSION AND FUTURE ENHANCEMENTS

In conclusion, conducting data science analysis on Amazon sales reports provides valuable insights and enables data-driven decision-making. Key findings from the analysis include trends in product performance, customer behavior, and market dynamics. These insights can inform strategic decision-making, optimize sales strategies, and improve overall revenue generation for Amazon. The analysis also helps in demand forecasting, inventory management, and identifying cross-selling and upselling opportunities. By leveraging data exploration, visualization, and predictive modeling techniques, the analysis enhances decision-makers' ability to optimize resource allocation and improve business performance.

For future enhancements, incorporating advanced machine learning algorithms can improve sales predictions and identify more complex patterns in customer behavior. Utilizing natural language processing techniques can extract valuable information from customer reviews and feedback, providing deeper insights into product preferences and sentiment analysis. Integrating external data sources such as social media or competitor data can further enrich the analysis and provide a comprehensive view of the market landscape. Additionally, implementing real-time or near-real-time analytics can enable timely decision-making and agile response to market changes.

Moreover, the ethical considerations surrounding data collection, privacy, and bias should be given utmost importance. Ensuring data security, respecting user privacy, and addressing potential biases in the analysis should be an ongoing focus.

Overall, the continuous advancement of data science techniques, the integration of diverse data sources, and a strong emphasis on ethical practices will contribute to the future enhancements of Amazon sales report analysis, empowering businesses to make informed decisions and stay competitive in the ever-evolving marketplace.