**CAPSTONE 3 PROJECT REPORT:**

**Customer Segmentation and Sales Analysis for Amazon Products**

**INTRODUCTION**

**Problem Statement:**

The goal of this project is to analyze the sales data from an e-commerce platform to identify patterns and insights that can inform business strategies. Specifically, we aim to cluster the sales data to understand different customer segments and their purchasing behavior.

**Objectives**:

1. Perform data preprocessing to clean and prepare the data for analysis.
2. Apply clustering techniques to segment the sales data.
3. Develop machine learning models to predict the identified clusters.
4. Analyze the key factors driving each cluster.
5. Provide actionable recommendations based on the findings.

**DATA OVERVIEW**

**Description of the Data:**

The dataset used in this project is the Amazon Sale Report, which includes details about orders such as order ID, date, status, fulfillment method, sales channel, product category, quantity, amount, and shipping information.

**Data Sources:**

The data was sourced from an Amazon sales report CSV file.

**APPROACH**

**Data Preprocessing Steps:**

* Loaded the data using Dask to handle large file sizes efficiently.
* Selected relevant features for analysis: Amount, Qty, Category, and Cluster.
* Created dummy variables for categorical features.
* Handled missing values using SimpleImputer with a median strategy.
* Standardized the numerical features using StandardScaler.

**Feature Selection:**

Focused on key features: Amount, Qty, and Category to perform clustering and model training.

**Model Selection and Training:**

* Applied KMeans clustering to segment the data into four clusters.
* Developed three machine learning models: Logistic Regression, Random Forest, and SVC.
* Performed hyperparameter tuning to optimize model performance.
* Evaluated models using accuracy, precision, recall, and F1 score metrics.

**Findings**

**Model Performance Summary:**

Logistic Regression:

Accuracy: 0.9852

Precision: 0.9705

Recall: 0.9852

F1 Score: 0.9778

Random Forest:

Accuracy: 0.9851

Precision: 0.9705

Recall: 0.9851

F1 Score: 0.9777

SVC:

Accuracy: 0.9852

Precision: 0.9705

Recall: 0.9852

F1 Score: 0.9778

**Key Insights from Feature Importance:**

The most important feature identified by the models was “Amount” with a feature importance score of 0.968393. Other significant features include Qty, Category\_Set, and Category\_Top.

**Confusion Matrix and Classification Report Results:**

The confusion matrix and classification report for all models indicate high accuracy and recall, but some precision issues due to certain clusters having no predicted samples.

**Recommendations**

**Three Concrete Recommendations for the Client:**

1. Increase Marketing Efforts for Low-Demand Products:

Focus on marketing strategies for products in Cluster 0 (Blouse, Bottom, Dupatta, Saree, Ethnic Dress, and Western Dress) to boost their sales.

1. Maintain Current Strategies for Moderately Performing Products:

Continue existing marketing and sales strategies for products in Cluster 1 (Western Dress), which show balanced demand and price points.

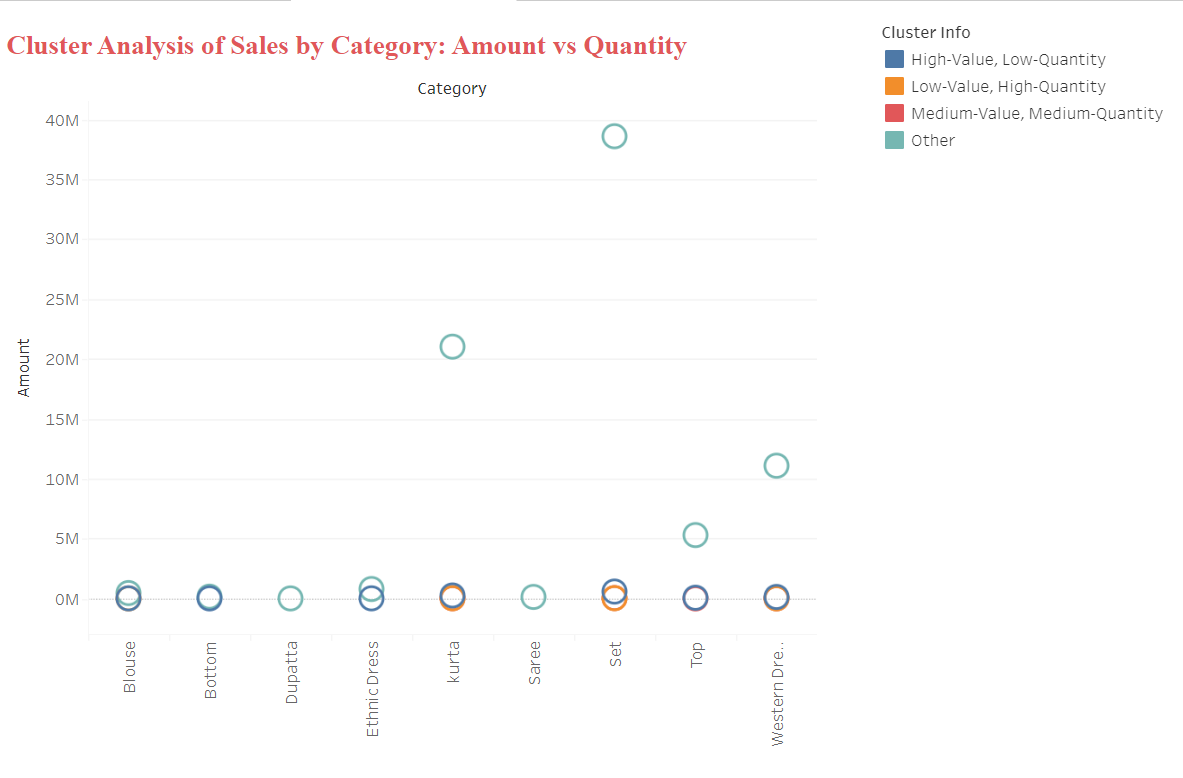
1. Emphasize Premium Quality for High-Priced Items:

Highlight the premium quality and exclusivity of products in Cluster 2 (Ethnic Dress) in marketing campaigns to attract selective customers.

**Ideas for Further Research:**

* Investigate customer demographics and purchase behaviors to refine cluster definitions.
* Explore time-series analysis to predict future sales trends.
* Analyze the impact of promotions and discounts on different product categories.

**Figures and Visualizations**

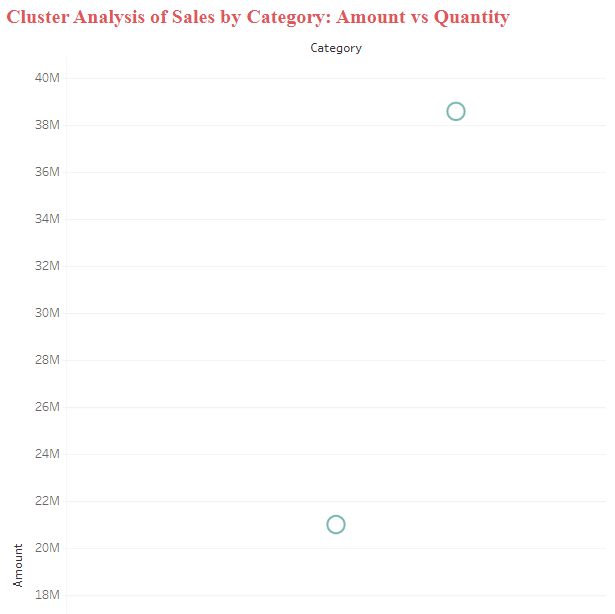


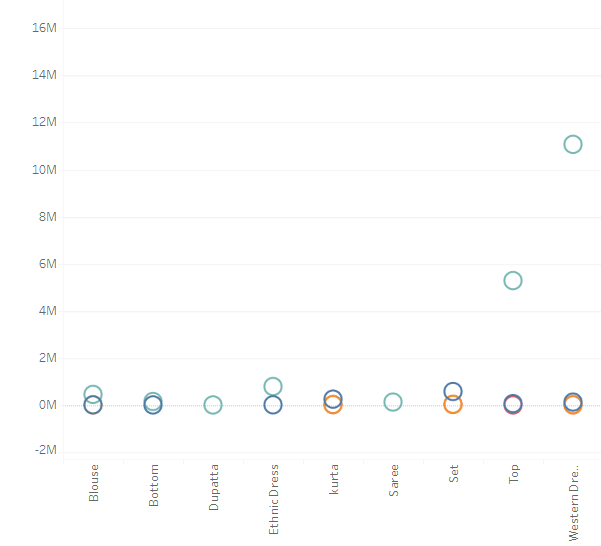
#### **1. Scatter Plot of Clusters by Amount and Qty**

This scatter plot shows the clusters formed in the data, with Amount on the x-axis and Qty on the y-axis. The points are colored by cluster, providing a clear visualization of how different clusters are distributed based on these two features.

2. **Cluster Analysis of Sales by Category: Amount vs Quantity**

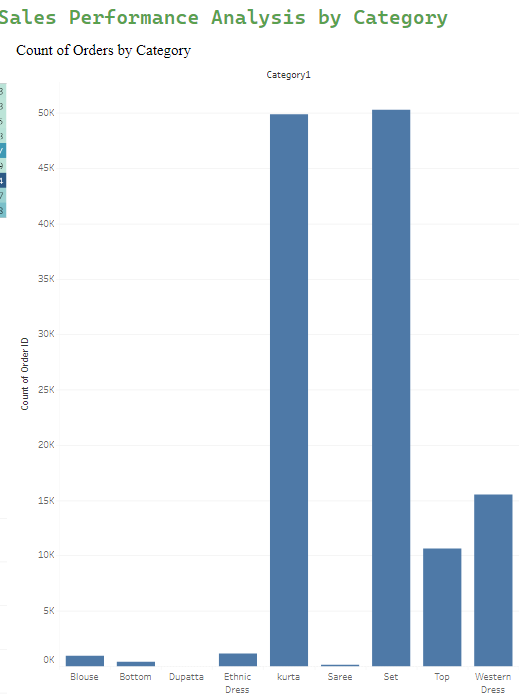
This scatter plot shows the clusters formed in the data, with Amount on the y-axis and Category on the x-axis. The points are colored by cluster, providing a clear visualization of how different clusters are distributed based on these two features. This helps in identifying patterns and differences among the clusters.





3. **Sales Performance Analysis by Category: Count of Orders by Category**

This bar chart illustrates the count of orders for each product category, with Kurta and Set leading in sales volume, each surpassing 50,000 orders. Western Dress also shows substantial sales, followed by Top with moderate order counts. Ethnic Dress, Blouse, Bottom, Dupatta, and Saree have lower order counts, indicating more selective demand. This visualization highlights the most and least popular product categories, guiding inventory and marketing strategies.



**Conclusion**

**Summary of Work Done:**

In this project, I performed data preprocessing, clustering, and model training to analyze Amazon sales data. Key features driving each cluster were identified, and model performance was thoroughly evaluated. Based on the analysis, actionable recommendations were provided to the client to enhance their sales strategies and operational efficiency.

**Final Thoughts:**

The project successfully identified distinct customer segments and their purchasing behaviors. Further research and more granular data analysis can provide deeper insights to optimize sales strategies.