Data cleaning and analyzing

Insights:

- General Insights:
  - Insight 0: There is an irrelevant column (“Unnamed: 22”)
  - The column “fulfilled-by” has the most NULLS and is most likely an irrelevant column because the only entry other than NULL is Easy Ship
  - Columns related to the order itself such as (Size, style, category) [are most likely irrelevant columns
  - ship-postal-code is also a row identifier
  - Country is all India so it is absolutely an un-needed piece of information

- **Check distribution of and generate insights:**

  a) **Numerical values (2 cols) [Qty and amount]**

     - Noticed that the column “Qty” exhibits a strong concentration in the first 2 categories. Fig.1
     - The “Amount” had some outliers and after removing them and drew a line plot to see any change of Status with its decrease or increase but found **no significance** only that the orders themselves increases in Amount range 200-400

  b) **Categorical values (21 cols): [The rest]**
     - Each city has a unique **ship-postal code**
     - **Date** is only 4 months category with March being an outlier
     - Days 28, 29 of June shows great number of pending orders and this might be because they are the most recent
     - **Shipped status** contribute to more than ~ 60% of the data while shipped delivered to buyer ~25% and cancelled are ~15% and the others are almost 2% except in June it rises to 6%
     - Behavior “Status distribution” seems the same over the 3 months ***(Fig 3-5)*** **Date** is insignificant and dropped
     - Ship-state have redundancy due to letter case differences unique values are {47}

Data Preprocessing:

a) **Date:**

   i. Although it seemed startling but to prove more that it is a non-informative variable I have visualized the trend over the three months using a line graph and through all the days the Shipped is always more than the unshipped and changes together with none of them dropping or increasing significantly in any time interval as shown in the sample of 10 days in April and in May in ***Figs.12-13.***

b) **Categorical data:**

   i. **Promotions** *(Fig.21)*

      1. I Altered the promotions column to Boolean of having a promotion or not to calculate the canceled orders percentage and I found out there is actually a big tendency of Canceling when not having a promotion
      2. And also, by calculating the P-value and the Chi-Square statistic it indicated there is a strong relation after creating the Contingency table between “Has Promotion” and non-cancelled status categories ***(Fig.2)***

   ii. **Ship-state**

      1. Unified the case of the words and removed the nans
      2. Grouped all the states with orders that are below the average state orders to “others” and this reduced dimensionality from 47 to 12 only
      3. After using bar-graph as shown in ***Fig.14*** , it turned out that states possess undistinctive behavior that reflects upon the status so I decided to **Drop** it
      4. Although it had a high chi-score and low p-value , when I encoded it and the “Category” section and used it in my model it had almost the same accuracy as shown in ***Fig.20.***

   iii. **Ship-service-level (Fig.10)**

      1. First two categories possess an important differentiating value while *easy ship* is un-needed category with only 2 records so I removed it.

   iv. **Style**

      1. Reduced the styles to only 4 categories ,which was 14 *after removing the number codes*, by considering the category having records below average as others
      2. No differential behavior across categories as shown in ***Fig.16*** so it got **dropped**

v. **Category**

   1. Done the same as the style column and reduced the identifiers to 4 including the “others” reaching a dimensionality reduction of 55.5% as shown in **Fig.17** the top 10 categories before the reduction.
   2. No differential behavior across categories as shown in ***Fig.15*** so it got **dropped**
   3. **Note: style and category are almost a mapping for each other**

vi. **Fulfilled by**

   1. Altered the Column to Easy-ship and the blanks as false and Easy-ship as true
   2. It is a mapping for the Fulfilment (totally redundant information for the Fulfilment and the ship-service)
      a. Example (Fulfilment = Merchant means ship-service = standard means Easy-ship)
   3. **Dropped**

vii. **Fulfilment**

   1. Is the same mapping for the “ship service” so I decided to keep it and drop the ship-service and the Fulfilled by as proven in ***Figs.8 to 11.***
   2. *Renamed to “Amazon” and type [Boolean]*

viii. **B2B**

   1. Almost all the data is False (non-significant identifier); so, it is **dropped**

ix. **SKU**

   1. This is the column representing the products; **Fig.18** shows the top 10 sold products and **Fig.19** represents the similar behavioral pattern of the status which led it to be insignificant to the model and **dropped**

b. **Numerical Data:**

i. Qty:

   1. Based on the insights decided on the dimensionality reduction to be 3 categories
   2. After reducing the dimensions to 3 I found out that each category is not a differentiating variable as in **Fig.7** for the status at all except that the ratio of cancelled in the 3-15 bins is 0 and this is because this bin is only 0.4% of the data so its not informative
   3. Based on this the Qty is **dropped**

ii.

Amount:

1. **Dropped** due to the fact of irrelevance to status behavior as in ***Fig.6***

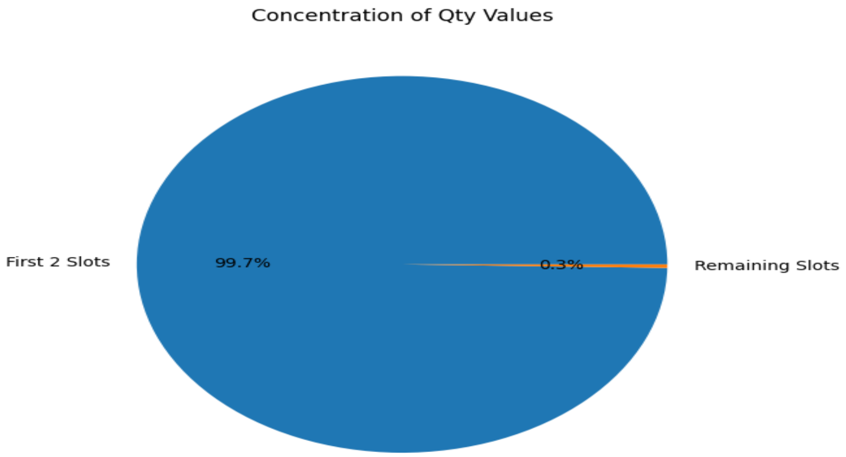
Decision Summary:

Feature	Decision
Promotions	Data transformation (True or False)
Fulfilment	Altered to (Amazon) and is boolean
Date	Dropped
Ship-City	Dropped
Ship-postal-code	Dropped

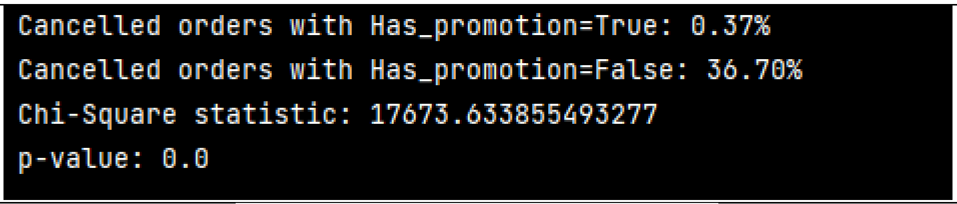
Ship-state	Dropped
Ship-service level	Dropped
SKU ,Size	Dropped
fulfilled-by	Dropped
Style, Category	Dropped
Qty	Dropped
The rest	Dropped for obvious reasons Redundancy, non-differentiating variables

Figures:

<table>
<tr>
<td>1</td>
<td>

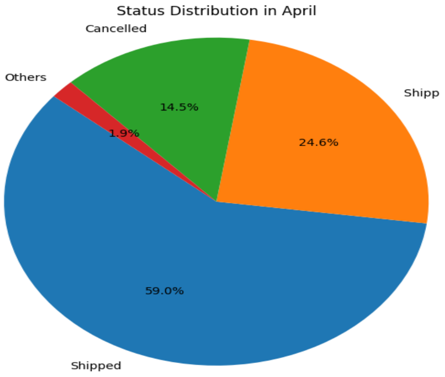


</td>
</tr>
<tr>
<td>2</td>
<td>

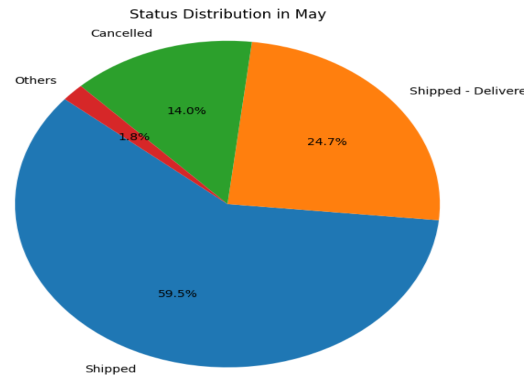
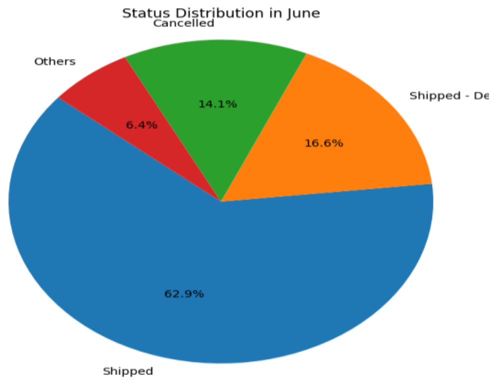
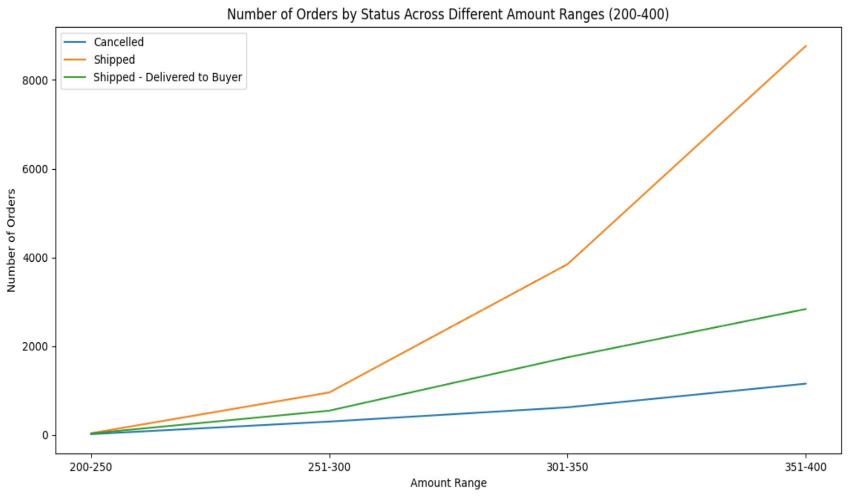


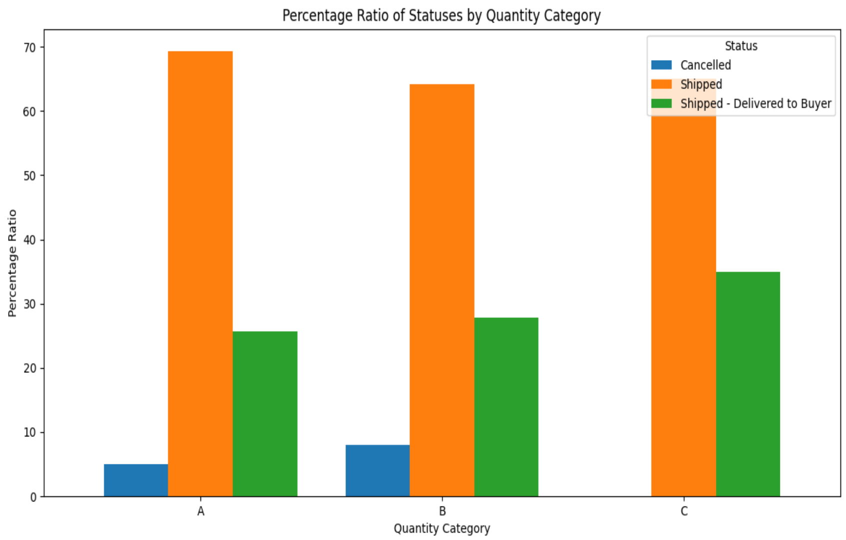
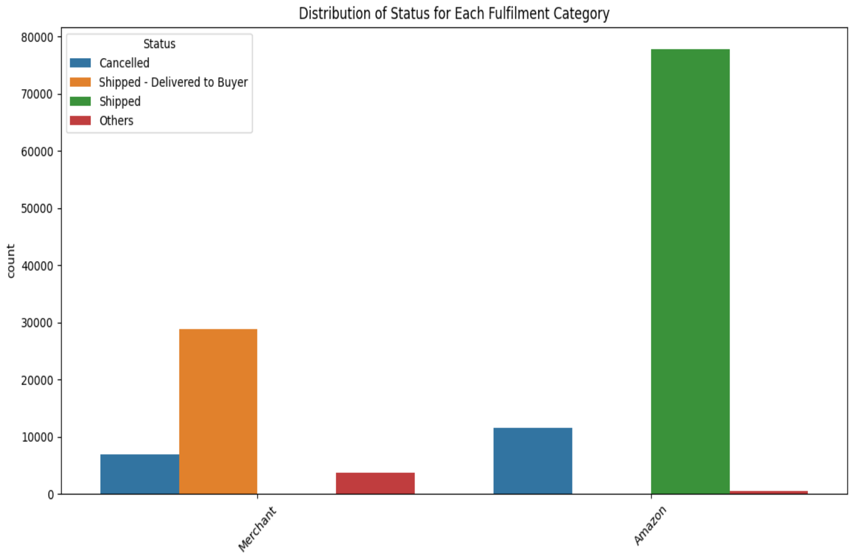
```
Cancelled orders with Has_promotion=True: 0.37%
Cancelled orders with Has_promotion=False: 36.70%
Chi-Square statistic: 17673.633855493277
p-value: 0.0
```

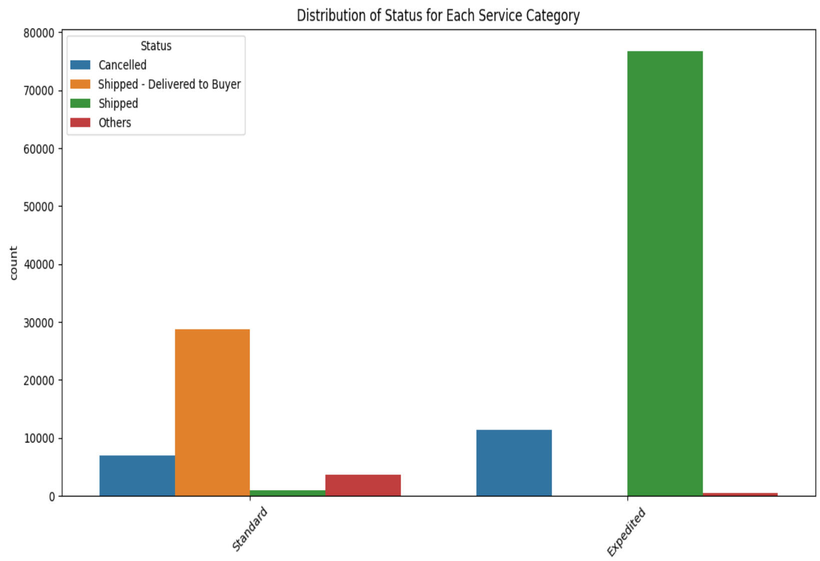
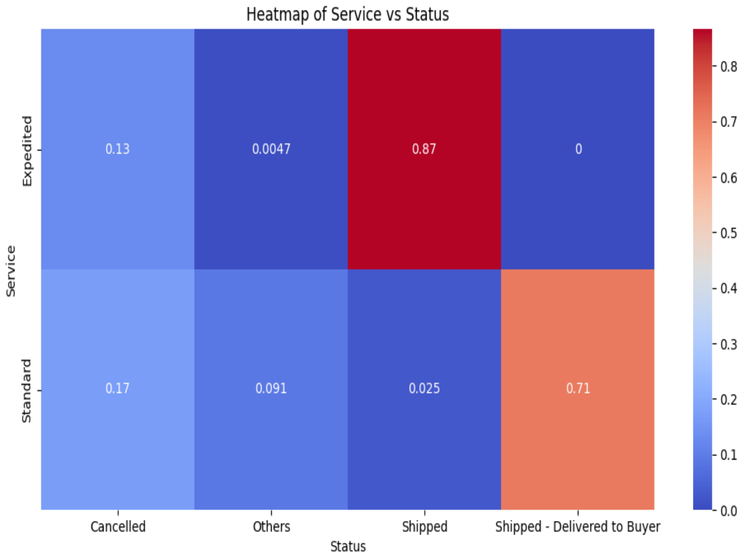
</td>
</tr>
<tr>
<td>3</td>
<td>

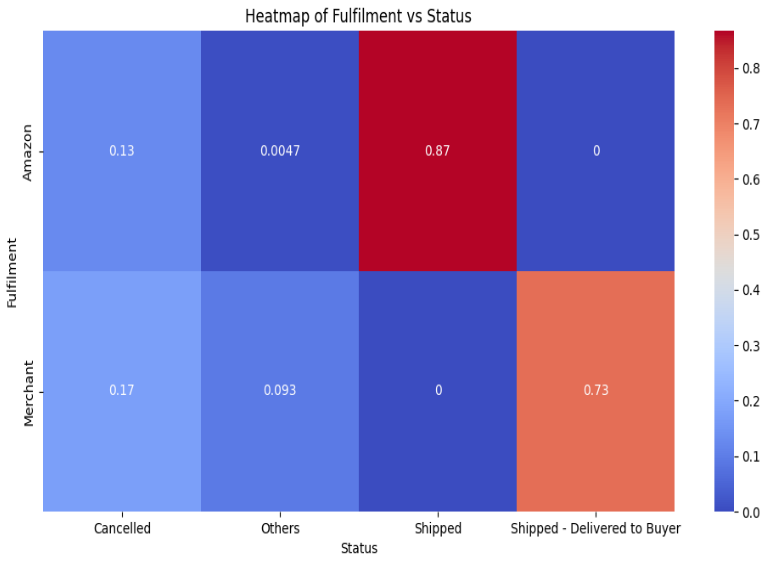
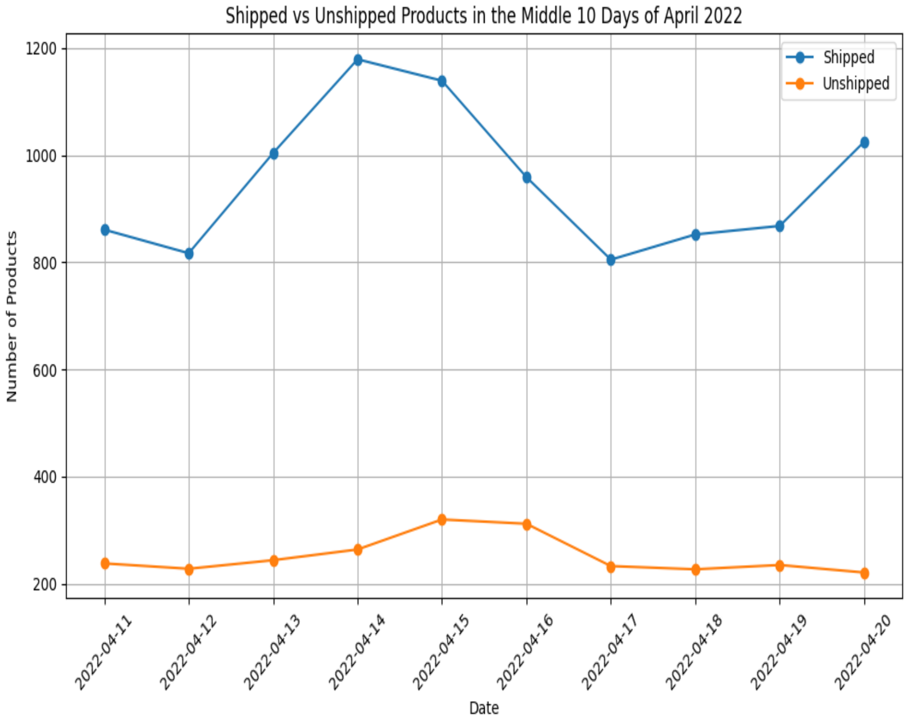


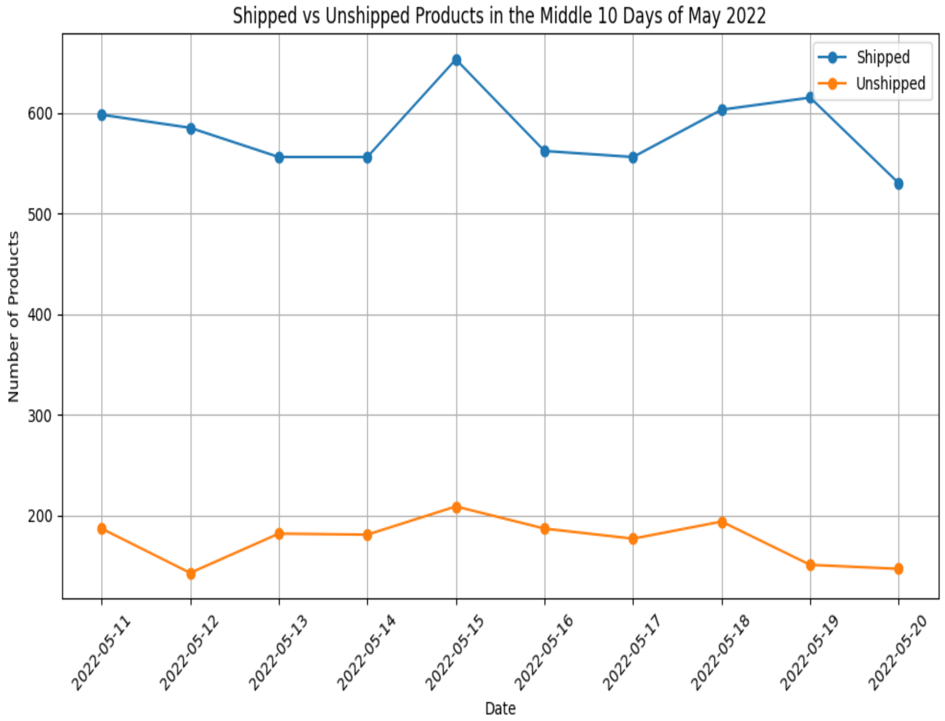
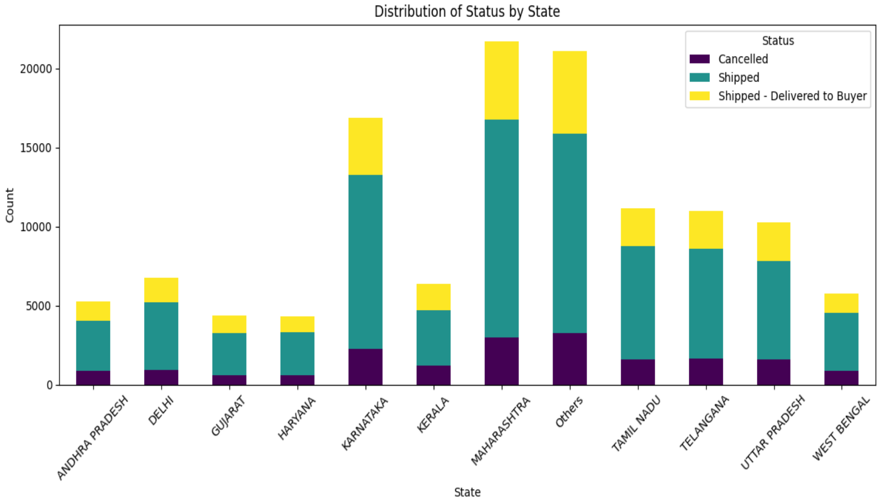
</td>
</tr>
</table>

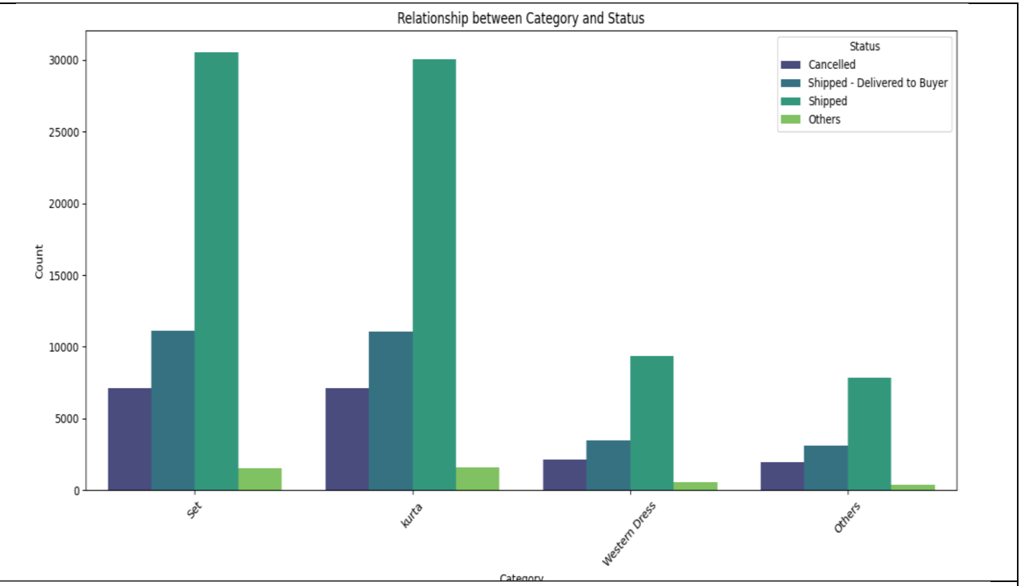
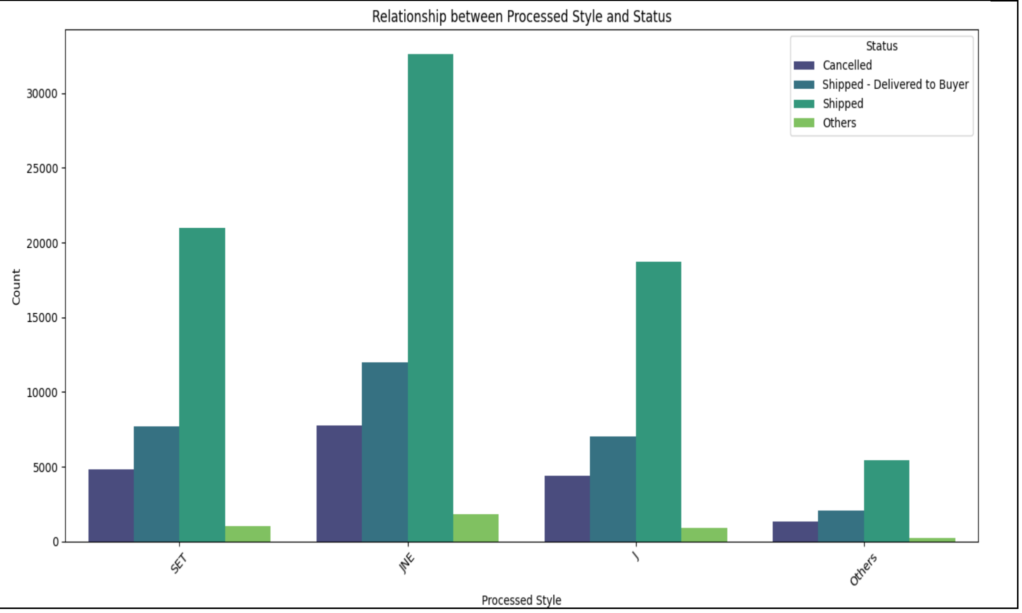
<table>
<tr><td>4</td><td>
Status Distribution in May
Cancelled
Others
Shipped - Delivere
14.0%
1.8%
24.7%
59.5%
Shipped</td></tr>
<tr><td>5</td><td>
Status Distribution in June
Cancelled
Others
Shipped - De
14.1%
6.4%
16.6%
62.9%
Shipped</td></tr>
<tr><td>6</td><td>
Number of Orders by Status Across Different Amount Ranges (200-400)
Cancelled
Shipped
Shipped - Delivered to Buyer
8000
6000
4000
2000
0
Number of Orders
200-250
251-300
301-350
351-400
Amount Range</td></tr>
</table>

7	Percentage Ratio of Statuses by Quantity Category Percentage Ratio 0 10 20 30 40 50 60 70 A B C Quantity Category Status Cancelled Shipped Shipped - Delivered to Buyer
8	Distribution of Status for Each Fulfilment Category count 0 10000 20000 30000 40000 50000 60000 70000 80000 Merchant Amazon Status Cancelled Shipped - Delivered to Buyer Shipped Others

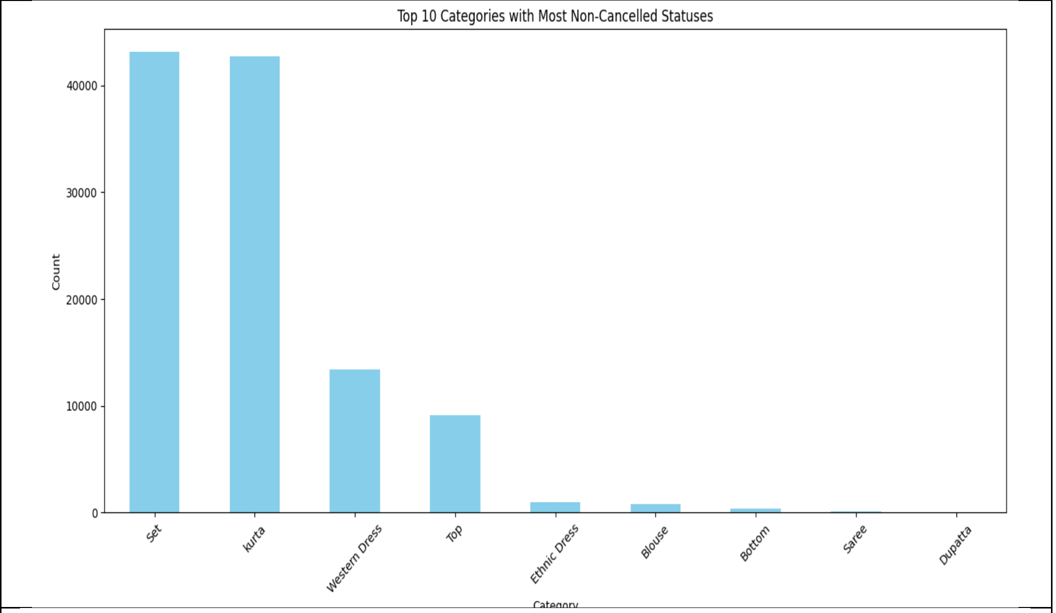
9	 Distribution of Status for Each Service Category Status Cancelled Shipped - Delivered to Buyer Shipped Others count 80000 70000 60000 50000 40000 30000 20000 10000 0 Standard Expedited
10	 Heatmap of Service vs Status Service Expedited Standard 0.13 0.0047 0.87 0 0.17 0.091 0.025 0.71 Cancelled Others Shipped Shipped - Delivered to Buyer Status 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

11	Heatmap of Fulfilment vs Status Fulfilment Amazon Merchant 0.13 0.0047 0.87 0 0.17 0.093 0 0.73 Cancelled Others Shipped Shipped - Delivered to Buyer Status 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
12	Shipped vs Unshipped Products in the Middle 10 Days of April 2022 Shipped Unshipped Number of Products 1200 1000 800 600 400 200 2022-04-11 2022-04-12 2022-04-13 2022-04-14 2022-04-15 2022-04-16 2022-04-17 2022-04-18 2022-04-19 2022-04-20 Date

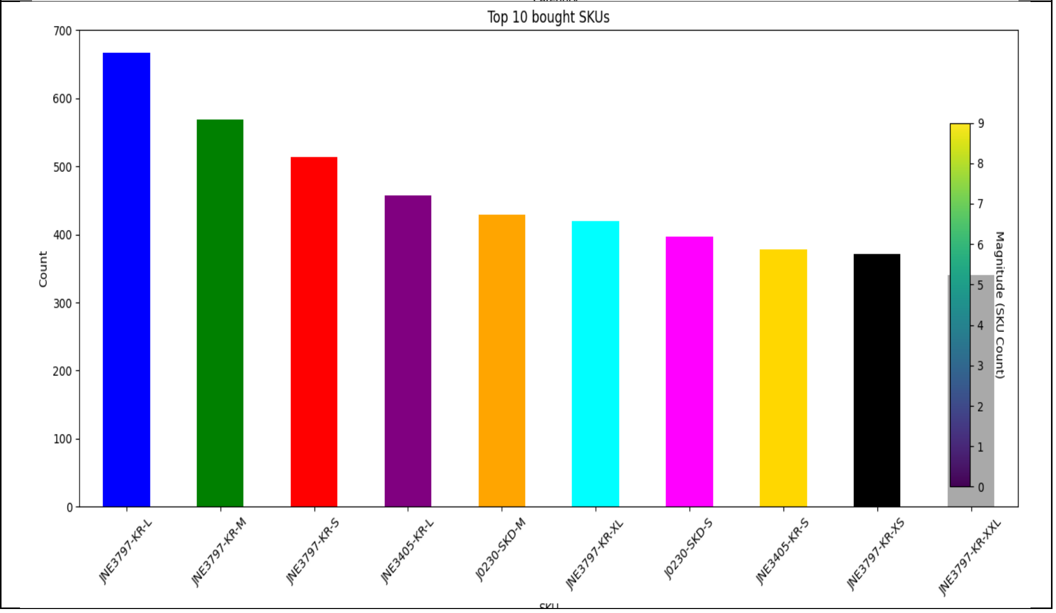
<table>
<tr><td>13</td><td>

Shipped vs Unshipped Products in the Middle 10 Days of May 2022
Shipped
Unshipped
Number of Products
600
500
400
300
200
2022-05-11
2022-05-12
2022-05-13
2022-05-14
2022-05-15
2022-05-16
2022-05-17
2022-05-18
2022-05-19
2022-05-20
Date
</td></tr>
<tr><td>14</td><td>

Distribution of Status by State
Status
Cancelled
Shipped
Shipped - Delivered to Buyer
Count
20000
15000
10000
5000
0
ANDHRA PRADESH
DELHI
GUJARAT
HARYANA
KARNATAKA
KERALA
MAHARASHTRA
Others
TAMIL NADU
TELANGANA
UTTAR PRADESH
WEST BENGAL
State
</td></tr>
</table>

15	Relationship between Category and Status Status Cancelled Shipped - Delivered to Buyer Shipped Others Count 0 5000 10000 15000 20000 25000 30000 Set kurta Western Dress Others Category
16	Relationship between Processed Style and Status Status Cancelled Shipped - Delivered to Buyer Shipped Others Count 0 5000 10000 15000 20000 25000 30000 SET JNE J Others Processed Style

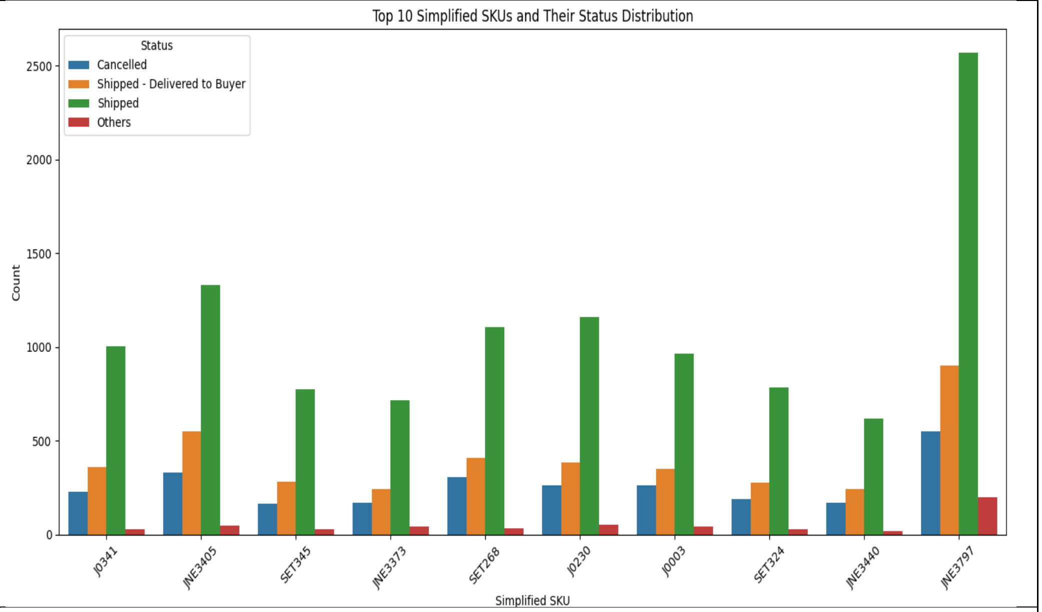
17


Top 10 Categories with Most Non-Cancelled Statuses
Count
0
10000
20000
30000
40000
Set
kurta
Western Dress
Top
Ethnic Dress
Blouse
Bottom
Saree
Dupatta
Category

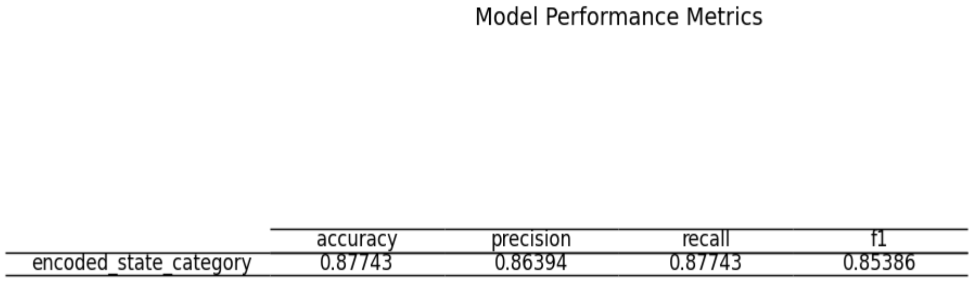
18


Top 10 bought SKUs
Count
0
100
200
300
400
500
600
700
JNE3797-KR-L
JNE3797-KR-M
JNE3797-KR-S
JNE3405-KR-L
J0230-SKD-M
JNE3797-KR-XL
J0230-SKD-S
JNE3405-KR-S
JNE3797-KR-XS
JNE3797-KR-XXL
SKU
Magnitude (SKU Count)
0
1
2
3
4
5
6
7
8
9

19

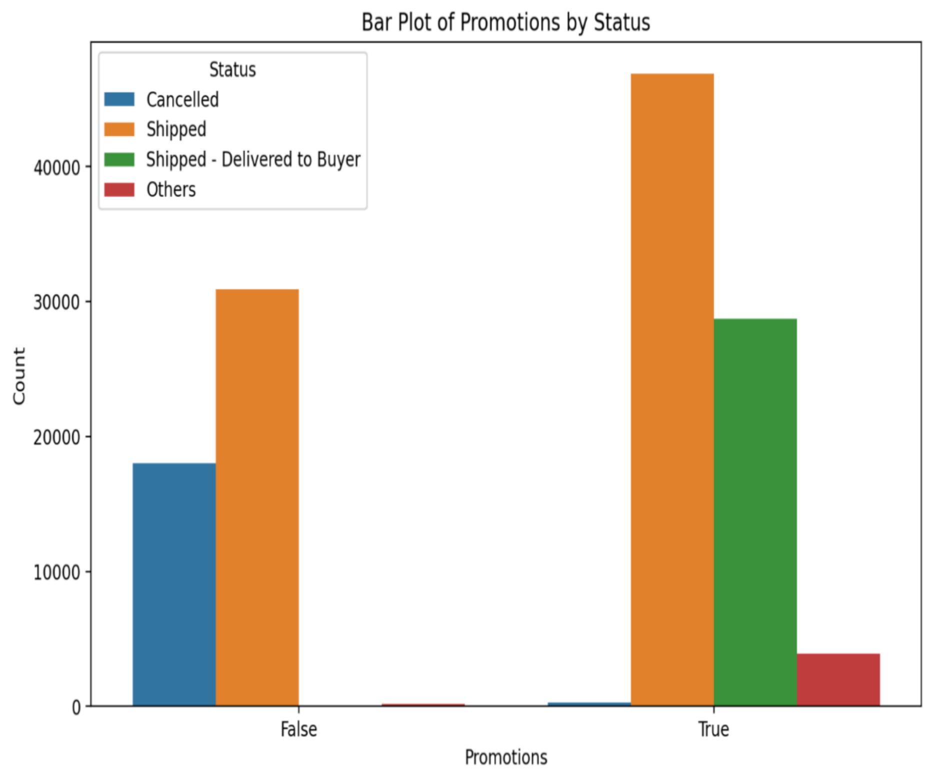


20



Model Performance Metrics

	accuracy	precision	recall	f1
encoded_state_category	0.87743	0.86394	0.87743	0.85386

21	Bar Plot of Promotions by Status Status 🟦 Cancelled 🟧 Shipped 🟩 Shipped - Delivered to Buyer 🟥 Others 40000 30000 20000 10000 0 Count False True Promotions