Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

ROLL NO: 2021300063

AIM

Create basic charts using Tableau / Power BI / R / Python / D3.js to be performed on the dataset of E-commerce field

Dataset Link

https://www.kaggle.com/datasets/mdsazzatsardar/amazonsalesreport

Data Cleaning

```
df.head()
```

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Style	SKU	Category	...	currency	Amount	ship-city	ship-state	ship-postal-code	ship-country	promotion-ids	B2B	fulfilled-by	Unnamed: 22
0	0	405-8078784-5731545	04-30-22	Cancelled	Merchant	Amazon.in	Standard	SET389	SET389-KR-NP-S	Set	...	INR	647.62	MUMBAI	MAHARASHTRA	400081.0	IN	NaN	False	Easy Ship	NaN
1	1	171-9198151-1101146	04-30-22	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	JNE3781	JNE3781-KR-XXXL	kurta	...	INR	406.00	BENGALURU	KARNATAKA	560085.0	IN	Amazon PLCC Free-Financing Universal Merchant ...	False	Easy Ship	NaN
2	2	404-0687676-7273146	04-30-22	Shipped	Amazon	Amazon.in	Expedited	JNE3371	JNE3371-KR-XL	kurta	...	INR	329.00	NAVI MUMBAI	MAHARASHTRA	410210.0	IN	IN Core Free Shipping 2015/04/08 23-48-5-108	True	NaN	NaN
3	3	403-9615377-8133951	04-30-22	Cancelled	Merchant	Amazon.in	Standard	J0341	J0341-DR-L	Western Dress	...	INR	753.33	PUDUCHERRY	PUDUCHERRY	605008.0	IN	NaN	False	Easy Ship	NaN
4	4	407-1069790-7240320	04-30-22	Shipped	Amazon	Amazon.in	Expedited	JNE3671	JNE3671-TU-XXXL	Top	...	INR	574.00	CHENNAI	TAMIL NADU	600073.0	IN	NaN	False	NaN	NaN

5 rows × 24 columns

The values in the ship-state field are repeating and with different cases or spellings. These need to be corrected.

```
df['ship-state'].unique()
```

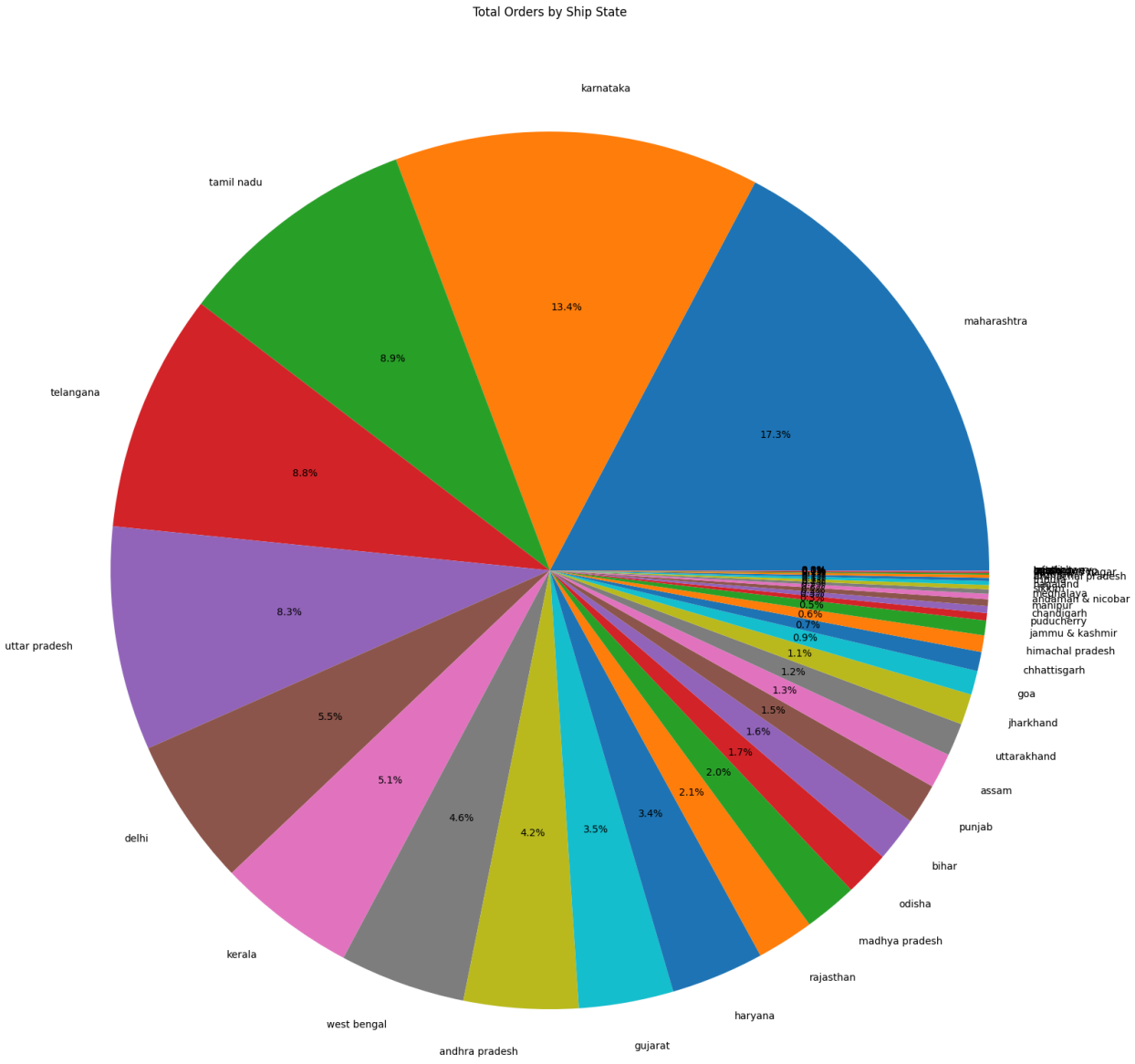
```
array(['MAHARASHTRA', 'KARNATAKA', 'PUDUCHERRY', 'TAMIL NADU',
       'UTTAR PRADESH', 'CHANDIGARH', 'TELANGANA', 'ANDHRA PRADESH',
       'RAJASTHAN', 'DELHI', 'HARYANA', 'ASSAM', 'JHARKHAND',
       'CHHATTISGARH', 'ODISHA', 'KERALA', 'MADHYA PRADESH',
       'WEST BENGAL', 'NAGALAND', 'Gujarat', 'UTTARAKHAND', 'BIHAR',
       'JAMMU & KASHMIR', 'PUNJAB', 'HIMACHAL PRADESH',
       'ARUNACHAL PRADESH', 'MANIPUR', 'Goa', 'MEGHALAYA', 'GOA',
       'TRIPURA', 'LADAKH', 'DADRA AND NAGAR', 'SIKKIM', 'Delhi', nan,
       'ANDAMAN & NICOBAR ', 'Punjab', 'Rajshthan', 'Manipur',
       'rajasthan', 'Odisha', 'NL', 'Bihar', 'MIZORAM', 'punjab',
       'New Delhi', 'Rajasthan', 'Punjab/Mohali/Zirakpur', 'Puducherry',
       'delhi', 'RJ', 'Chandigarh', 'orissa', 'LAKSHADWEEP', 'goa', 'PB',
       'APO', 'Arunachal Pradesh', 'AR', 'Pondicherry', 'Sikkim',
       'Arunachal pradesh', 'Nagaland', 'bihar', 'Mizoram', 'rajsthan',
       'Orissa', 'Rajsthan', 'Meghalaya'], dtype=object)
```

```
df['ship-state'] = df['ship-state'].str.replace('rj','rajasthan')
df['ship-state'] = df['ship-state'].str.replace('pb','punjab')
df['ship-state'] = df['ship-state'].str.replace('punjab/mohali/zirakpur','punjab')
df['ship-state'] = df['ship-state'].str.replace('nl','nagaland')
df['ship-state'] = df['ship-state'].str.replace('new delhi','delhi')
df['ship-state'] = df['ship-state'].str.strip()
df['ship-state'] = df['ship-state'].str.replace('rajshthan','rajasthan')
```

Visualizations

1. State wise overall sales

```
[ ]  state_wise_orders = df.groupby('ship-state')['Order ID'].count().sort_values(ascending=False)
     plt.figure(figsize=(20,20))
     plt.pie(state_wise_orders.values, labels=state_wise_orders.index, autopct='%1.1f%%')
     plt.title('Total Orders by Ship State')
     plt.show()
```

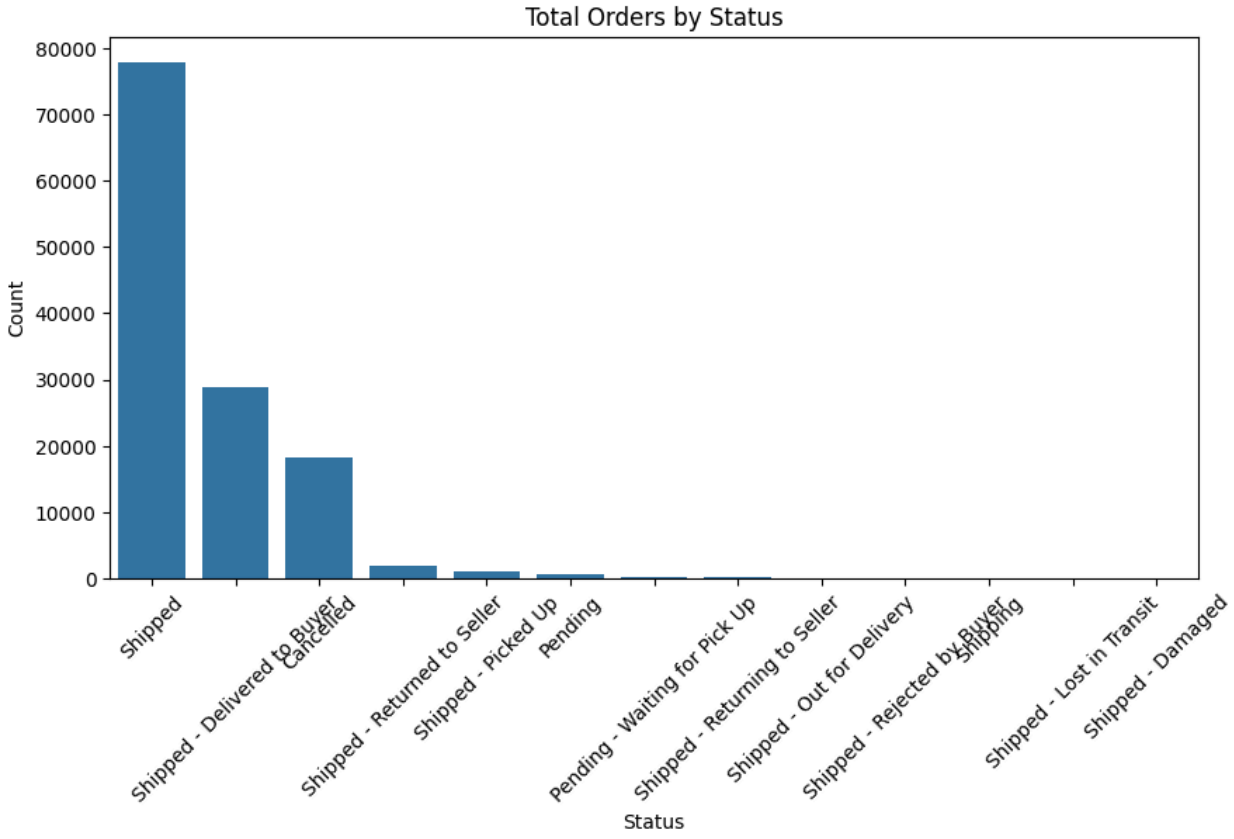


This pie chart shows the distribution of total orders by ship state. Maharashtra accounts for the largest share, contributing 17.3% of the total orders, followed by Karnataka at 13.4%. Tamil Nadu and Telangana also hold significant shares, with 8.9% and 8.8%, respectively.

Uttar Pradesh follows closely with 8.3%, while Delhi has a 5.5% share. This shows that over 50% of the sales are contributed by the top 5 states. The remaining states have smaller contributions, with Kerala at 5.1%, West Bengal at 4.6%, Andhra Pradesh at 4.2%, Gujarat at 3.5%, and Haryana at 3.4%. The chart shows that a large number of states have minimal order shares, each below 2%.

2. Overall status of orders

```
status_wise_orders = df.groupby('Status')['Order ID'].count().sort_values(ascending=False)
plt.figure(figsize=(10, 5))
sns.barplot(x=status_wise_orders.index, y=status_wise_orders.values)
plt.title('Total Orders by Status')
plt.xlabel('Status')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

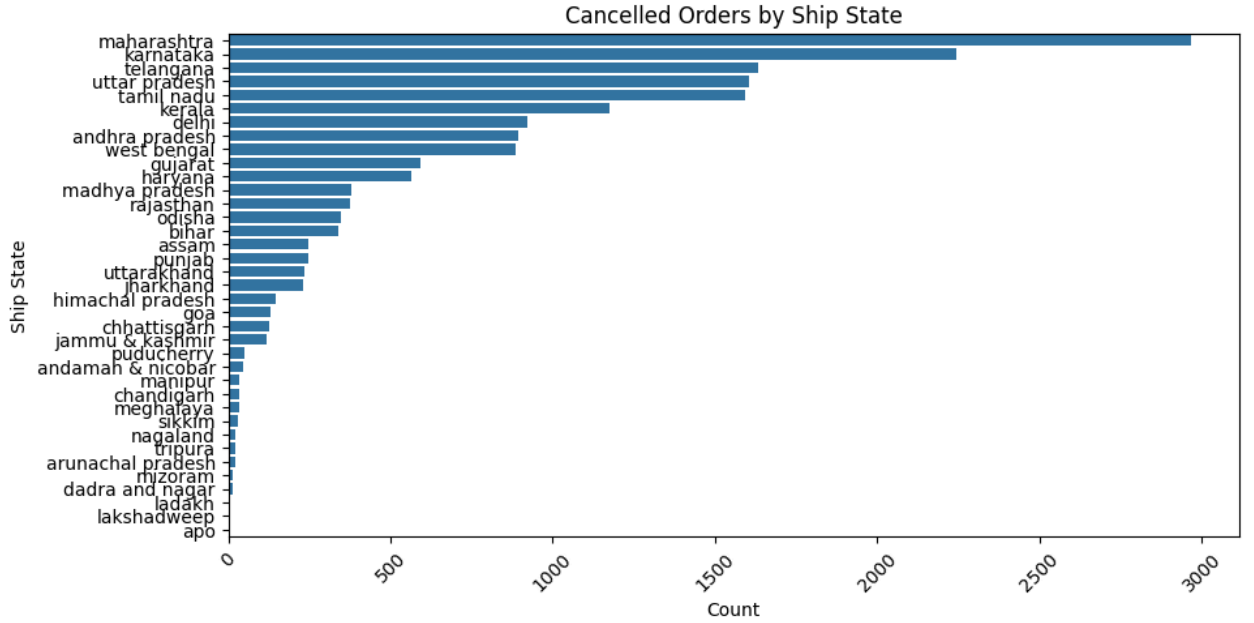


The overwhelming majority of orders are in the "Shipped" status, indicating a high percentage of successful order fulfillments. A substantial number of orders have progressed to the "Shipped - Delivered to Buyer" stage, suggesting a smooth delivery process. The categories "Shipped - Returned to Seller" and "Pending - Waiting for Pick Up" have significantly lower counts, implying that returns and cancellations are less common. The

remaining statuses, such as "Shipped - Picked Up," "Shipped - Out for Delivery," "Shipped - Rejected by Buyer," "Shipped Lost in Transit," and "Shipped - Damaged," have relatively small numbers of orders, suggesting that these situations are less frequent. Overall it can be observed that the orders have a high success rate.

3. State wise order cancellations

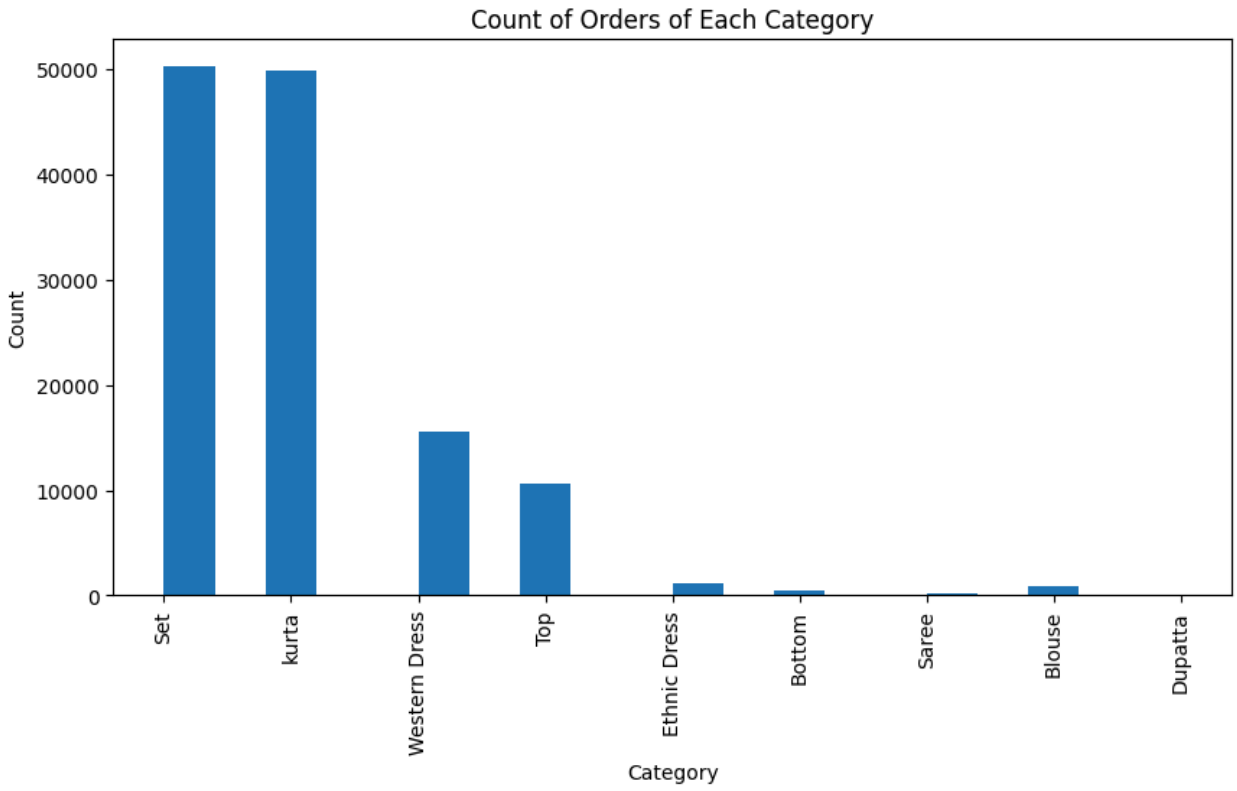
```
cancelled_orders = df[df['Status'] == 'Cancelled']
state_wise_cancelled_orders = cancelled_orders.groupby('ship-state')['Order ID'].count().sort_values(ascending=False)
plt.figure(figsize=(10, 5))
sns.barplot(y=state_wise_cancelled_orders.index, x=state_wise_cancelled_orders.values)
plt.title('Cancelled Orders by Ship State')
plt.ylabel('Ship State')
plt.xlabel('Count')
plt.xticks(rotation=45)
plt.show()
```



The states with higher orders are more likely to have proportionally equal number of cancellations in orders which can be seen in this bar graph. This graph indicates that there is no state with abnormally higher cancellation rates. Also it can be said that karnataka has proportionally lower cancellation rates as compared to order count.

4. Item wise order count

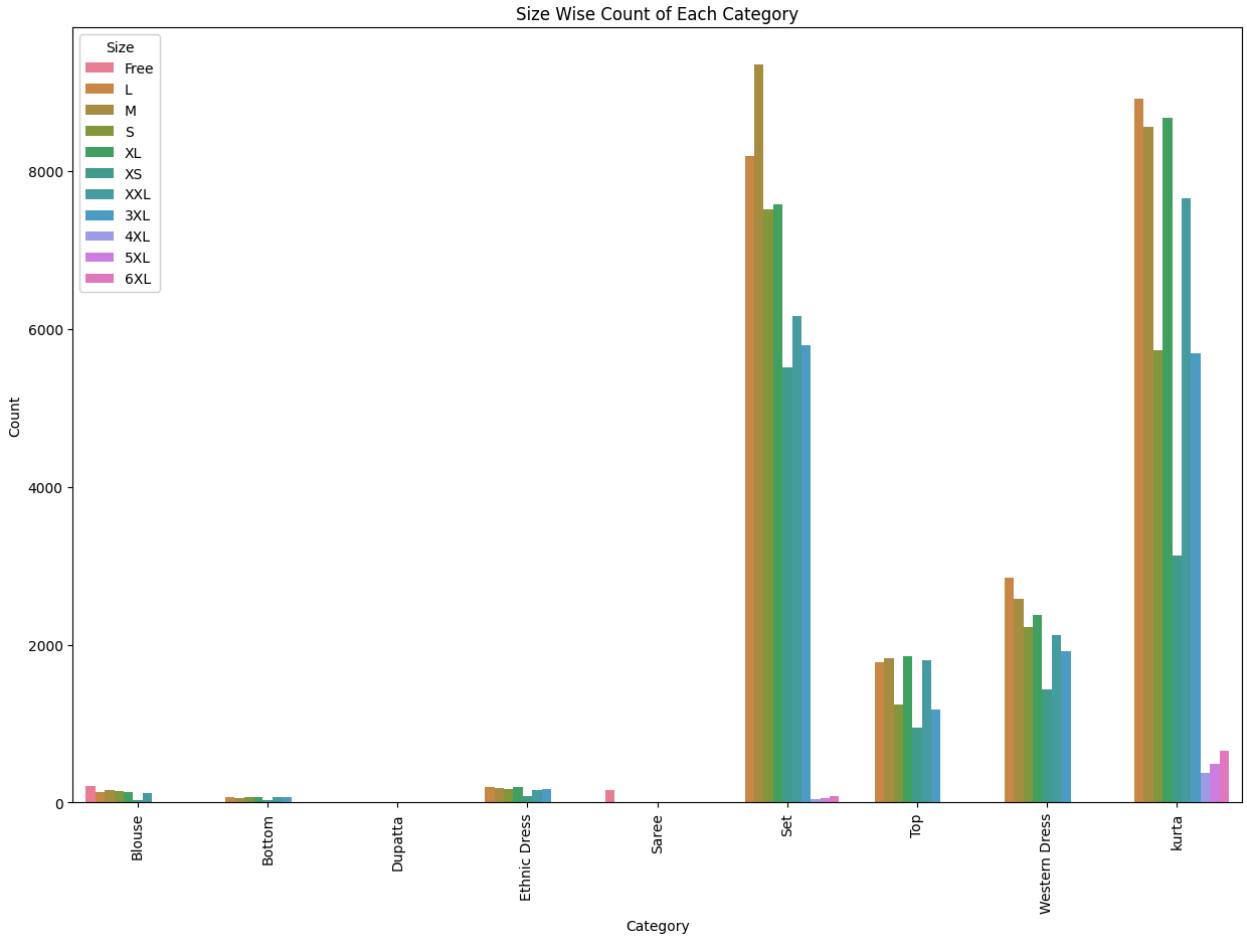
```
category_wise_orders = df.groupby('Category')['Order ID'].count().sort_values(ascending=False)
plt.figure(figsize=(10, 5))
plt.hist(df['Category'], bins=20)
plt.title('Count of Orders of Each Category')
plt.xlabel('Category')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.show()
```



The histogram of the count of orders of each category reveals a clear disparity in product popularity. "Set" and "Kurta" emerged as the most sought-after categories, significantly outselling the others. Western Dress and Top followed with a considerable number of orders, while Ethnic Dress and Bottom had noticeably lower demand. Saree, Blouse, and Dupatta, on the other hand, received minimal interest.

5. Size wise count of each item

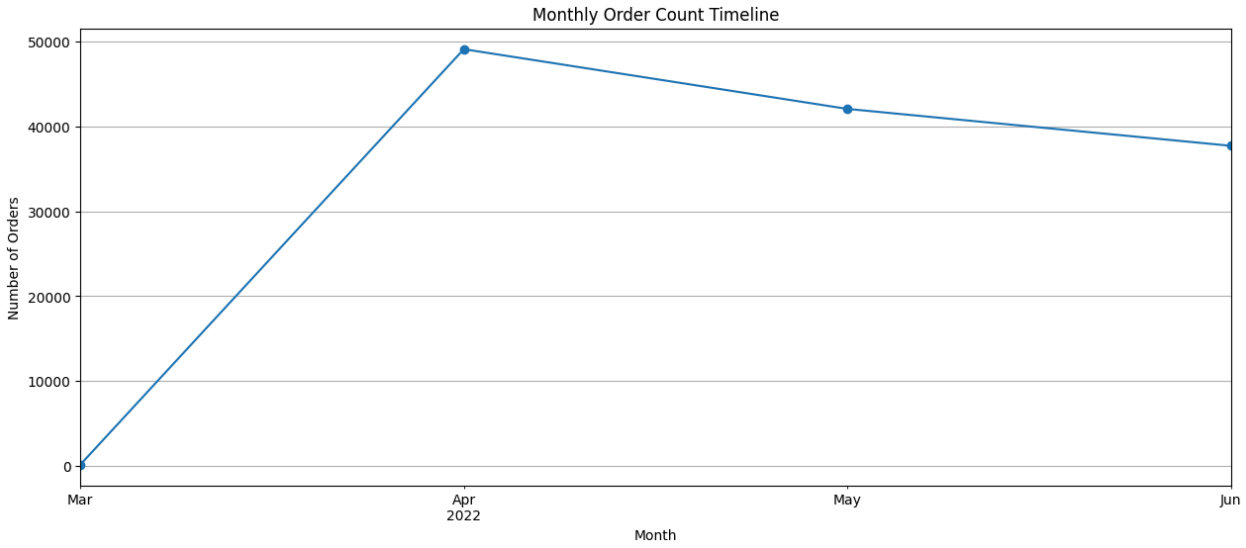
```
size_wise_count = df.groupby(['Category', 'Size'])['Order ID'].count().reset_index()
plt.figure(figsize=(15, 10))
sns.barplot(x='Category', y='Order ID', hue='Size', data=size_wise_count)
plt.title('Size Wise Count of Each Category')
plt.xlabel('Category')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.show()
```



The chart reveals that the most popular sizes vary significantly among different categories. For instance, "Set" and "Kurta" have a higher demand for larger sizes like XL, XXL, and 3XL, while "Blouse" and "Dupatta" primarily attract orders for smaller sizes like S and M. Each category exhibits distinct size preferences. "Saree" and "Top" have a relatively even distribution across various sizes, suggesting a broader range of customer preferences. In contrast, "Western Dress" and "Ethnic Dress" show a higher demand for medium-sized options. The chart also highlights potential gaps in size availability. If certain categories consistently have a higher demand for specific sizes that are not well-stocked, it could lead to lost sales and customer dissatisfaction.

6. Monthly order count timeline

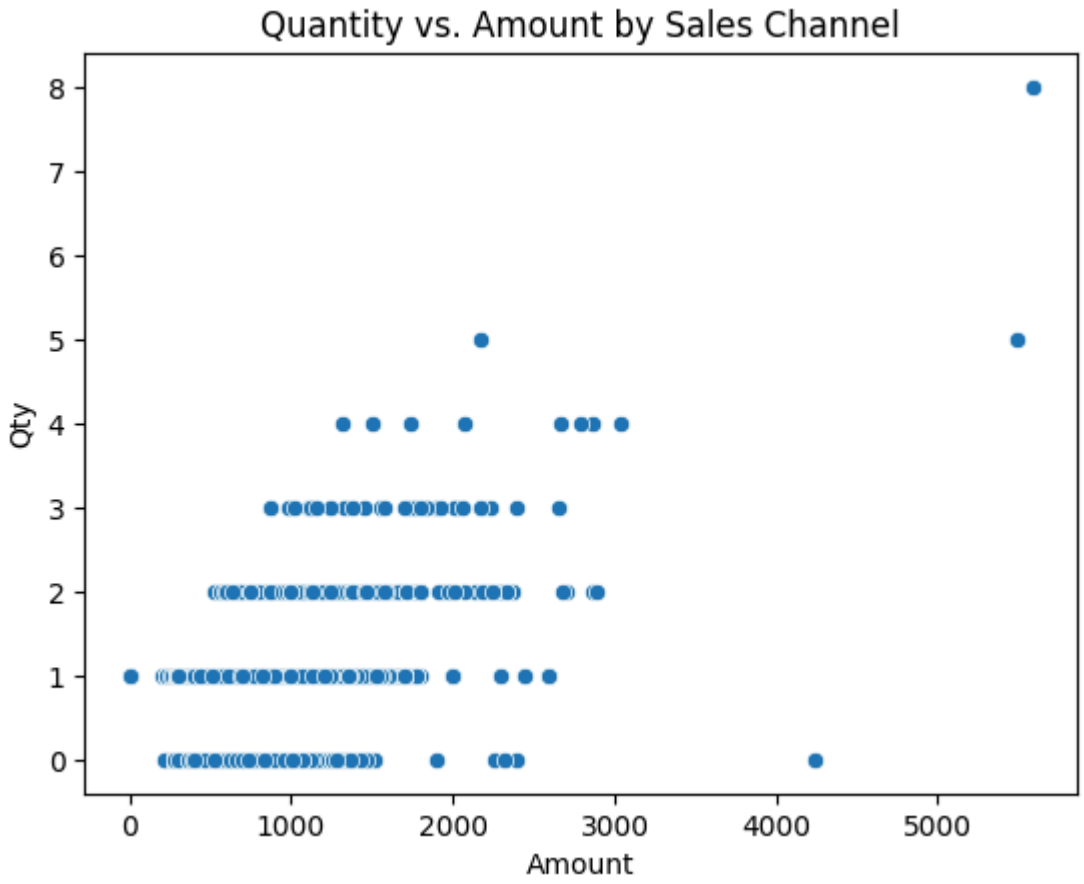
```
df['Date'] = pd.to_datetime(df['Date'])
df.set_index('Date', inplace=True)

monthly_orders = df['Order ID'].resample('M').count()

plt.figure(figsize=(15, 6))
monthly_orders.plot(kind='line', marker='o')
plt.title('Monthly Order Count Timeline')
plt.xlabel('Month')
plt.ylabel('Number of Orders')
plt.grid(True)
plt.show()
```



The most notable trend is a substantial increase in orders from March to April. This indicates a significant boost in sales or customer activity during this period. Following the peak in April, the number of orders experienced a steady decline in May and June. This could be attributed to various factors such as seasonality, promotional activities, or changes in market conditions.

7. Item cost versus Quantity Analysis

```
sns.scatterplot(data=df, y='Qty', x='Amount')
plt.title('Quantity vs. Amount')
plt.show()
```

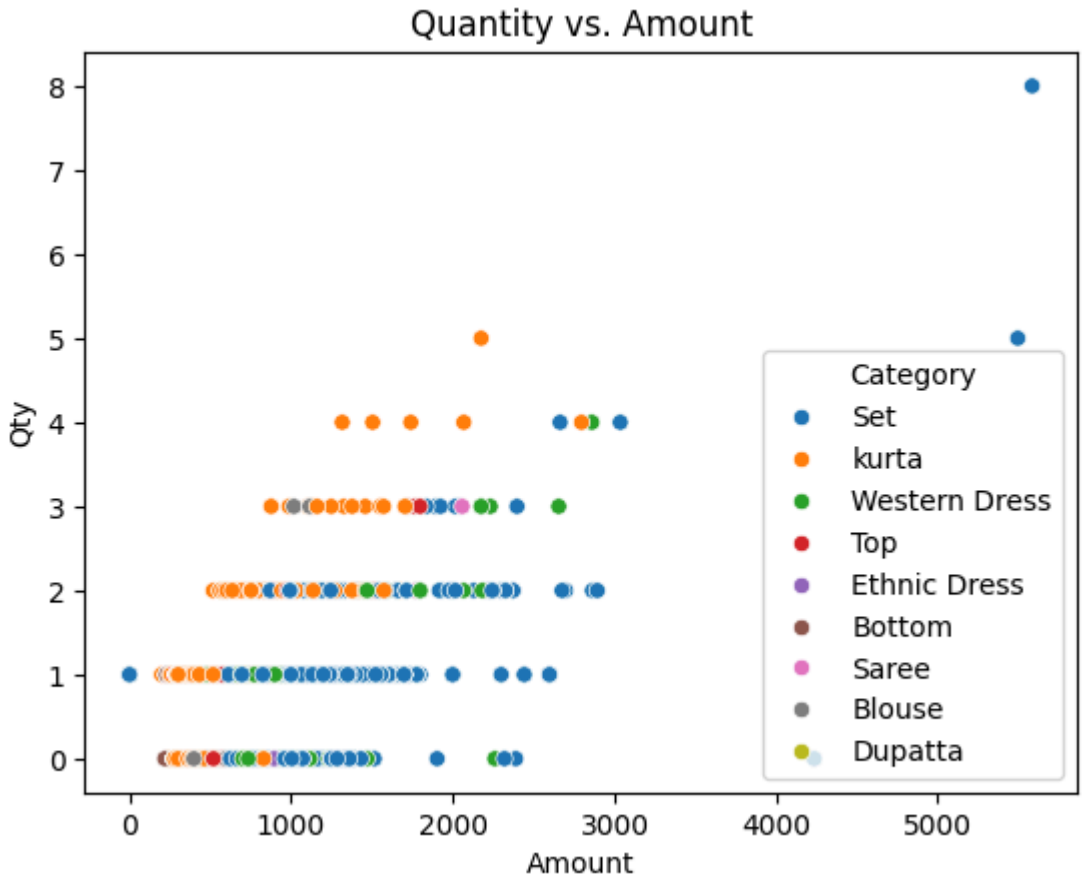


There appears to be a general positive correlation between quantity and amount, indicating that as the quantity of products sold increases, the total amount generated also tends to rise.

Outliers: A few data points, particularly those with high quantities and amounts, stand out as outliers. These could represent large bulk orders or special deals that significantly impact the overall sales figures.

8. Item cost versus Quantity Analysis by Category

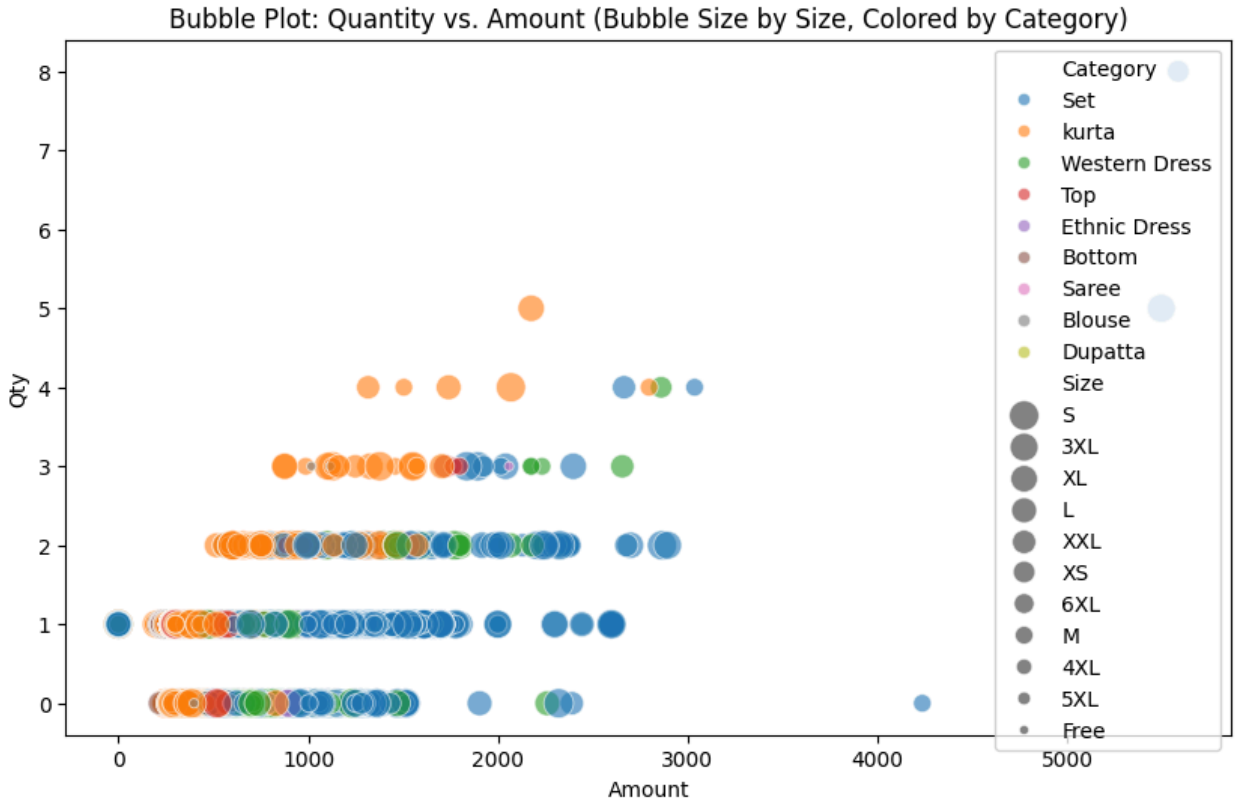
```
sns.scatterplot(data=df, y='Qty', x='Amount',hue='Category',legend='auto')
plt.title('Quantity vs. Amount')
plt.show()
```



There appears to be a general positive correlation between quantity and amount, indicating that as the quantity of products sold increases, the total amount generated also tends to rise. The different colors representing different categories reveal distinct patterns in the data. For example, "Kurta" and "Set" tend to have higher quantities and amounts sold compared to other categories.

9. Bubble Plot: Quantity vs. Amount (Bubble Size by Size, Colored by Category)

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, y='Qty', x='Amount', size='Size', hue='Category', legend=True, alpha=0.6, sizes=(20,
plt.title('Bubble Plot: Quantity vs. Amount (Bubble Size by Size, Colored by Category)')
plt.show()
```



There appears to be a general positive correlation between quantity and amount, indicating that as the quantity of products sold increases, the total amount generated also tends to rise. The different colors representing different categories reveal distinct patterns in the data. For example, "Kurta" and "Set" tend to have higher quantities and amounts sold compared to other categories. The bubble sizes, which represent the product size, show that certain categories have a higher demand for specific sizes. For instance, "Set" and "Kurta" tend to have larger bubble sizes, indicating a preference for larger sizes. A few data points, particularly those with high quantities and amounts, stand out as outliers. These could represent large bulk orders or special deals that significantly impact the overall sales figures.