da-amazon-sale-project

June 23, 2024

[11]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

[119]:
```
df=pd.read_csv(r"C:\Users\HP\Downloads\Amazon Sale Report.csv")
```

[123]:
```
df.shape
```

[123]:
```
(128976, 21)
```

[125]:
```
df.head()
```

[125]:
```
   index             Order ID      Date                        Status  \
0      0  405-8078784-5731545  04-30-22                     Cancelled
1      1  171-9198151-1101146  04-30-22  Shipped - Delivered to Buyer
2      2  404-0687676-7273146  04-30-22                       Shipped
3      3  403-9615377-8133951  04-30-22                     Cancelled
4      4  407-1069790-7240320  04-30-22                       Shipped

  Fulfilment Sales Channel ship-service-level  Category Size Courier Status  \
0   Merchant     Amazon.in           Standard   T-shirt    S     On the Way
1   Merchant     Amazon.in           Standard     Shirt  3XL        Shipped
2     Amazon     Amazon.in          Expedited     Shirt   XL        Shipped
3   Merchant     Amazon.in           Standard   Blazzer    L     On the Way
4     Amazon     Amazon.in          Expedited  Trousers  3XL        Shipped

   …  currency  Amount    ship-city   ship-state ship-postal-code  \
0  …       INR  647.62       MUMBAI  MAHARASHTRA         400081.0
1  …       INR  406.00    BENGALURU    KARNATAKA         560085.0
2  …       INR  329.00  NAVI MUMBAI  MAHARASHTRA         410210.0
3  …       INR  753.33   PUDUCHERRY   PUDUCHERRY         605008.0
4  …       INR  574.00      CHENNAI   TAMIL NADU         600073.0

   ship-country    B2B  fulfilled-by New  PendingS
0            IN  False     Easy Ship NaN       NaN
```

```
1            IN  False     Easy Ship NaN       NaN
2            IN   True           NaN NaN       NaN
3            IN  False     Easy Ship NaN       NaN
4            IN  False           NaN NaN       NaN

[5 rows x 21 columns]
```

[127]:

```python
df.tail()
```

[127]:

```
         index             Order ID      Date   Status Fulfilment   \
128971  128970  406-6001380-7673107  05-31-22  Shipped     Amazon
128972  128971  402-9551604-7544318  05-31-22  Shipped     Amazon
128973  128972  407-9547469-3152358  05-31-22  Shipped     Amazon
128974  128973  402-6184140-0545956  05-31-22  Shipped     Amazon
128975  128974  408-7436540-8728312  05-31-22  Shipped     Amazon

       Sales Channel  ship-service-level Category Size Courier Status  …   \
128971     Amazon.in           Expedited    Shirt   XL         Shipped  …
128972     Amazon.in           Expedited  T-shirt    M         Shipped  …
128973     Amazon.in           Expedited  Blazzer  XXL         Shipped  …
128974     Amazon.in           Expedited  T-shirt   XS         Shipped  …
128975     Amazon.in           Expedited  T-shirt    S         Shipped  …

        currency  Amount  ship-city    ship-state ship-postal-code   \
128971       INR   517.0  HYDERABAD     TELANGANA         500013.0
128972       INR   999.0   GURUGRAM       HARYANA         122004.0
128973       INR   690.0  HYDERABAD     TELANGANA         500049.0
128974       INR  1199.0      Halol       Gujarat         389350.0
128975       INR   696.0     Raipur  CHHATTISGARH         492014.0

        ship-country    B2B  fulfilled-by New  PendingS
128971            IN  False           NaN NaN       NaN
128972            IN  False           NaN NaN       NaN
128973            IN  False           NaN NaN       NaN
128974            IN  False           NaN NaN       NaN
128975            IN  False           NaN NaN       NaN

[5 rows x 21 columns]
```

[129]:

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 128976 entries, 0 to 128975
Data columns (total 21 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   index               128976 non-null  int64
 1   Order ID            128976 non-null  object
 2   Date                128976 non-null  object
 3   Status              128976 non-null  object
 4   Fulfilment          128976 non-null  object
 5   Sales Channel       128976 non-null  object
 6   ship-service-level  128976 non-null  object
 7   Category            128976 non-null  object
 8   Size                128976 non-null  object
 9   Courier Status      128976 non-null  object
 10  Qty                 128976 non-null  int64
 11  currency            121176 non-null  object
 12  Amount              121176 non-null  float64
 13  ship-city           128941 non-null  object
 14  ship-state          128941 non-null  object
 15  ship-postal-code    128941 non-null  float64
 16  ship-country        128941 non-null  object
 17  B2B                 128976 non-null  bool
 18  fulfilled-by        39263 non-null   object
 19  New                 0 non-null       float64
 20  PendingS            0 non-null       float64
dtypes: bool(1), float64(4), int64(2), object(14)
memory usage: 19.8+ MB
```

[131]:

```
df.isnull().sum()
```

[131]:

```
index                      0
Order ID                   0
Date                       0
Status                     0
Fulfilment                 0
Sales Channel              0
ship-service-level         0
Category                   0
Size                       0
Courier Status             0
Qty                        0
currency                7800
Amount                  7800
ship-city                 35
ship-state                35
ship-postal-code          35
ship-country              35
B2B                        0
fulfilled-by           89713
New                   128976
PendingS              128976
dtype: int64
```

[133]:
```
df.drop(['New','PendingS'],axis=1,inplace=True)
```

[135]:
```
df.columns
```

[135]:
```
Index(['index', 'Order ID', 'Date', 'Status', 'Fulfilment', 'Sales Channel',
       'ship-service-level', 'Category', 'Size', 'Courier Status', 'Qty',
       'currency', 'Amount', 'ship-city', 'ship-state', 'ship-postal-code',
       'ship-country', 'B2B', 'fulfilled-by'],
      dtype='object')
```

[137]:
```
df.dropna(inplace=True)
```

[139]:
```
df.shape
```

[139]:
```
(37514, 19)
```

[141]:
```
df['ship-postal-code']=df['ship-postal-code'].astype('int')
```

[143]:
```
df.describe()
```

[143]:
```
               index           Qty        Amount  ship-postal-code
count   37514.000000  37514.000000  37514.000000      37514.000000
mean    60953.809858      0.867383    646.553960     463291.552754
std     36844.853039      0.354160    279.952414     194550.425637
min         0.000000      0.000000      0.000000     110001.000000
25%     27235.250000      1.000000    458.000000     370465.000000
50%     63470.500000      1.000000    629.000000     500019.000000
75%     91790.750000      1.000000    771.000000     600042.000000
max    128891.000000      5.000000   5495.000000     989898.000000
```

[148]:
```
df.describe(include='object')
```

[148]:
```
                   Order ID      Date                         Status  \
count                 37514     37514                          37514
unique                34664        91                             11
top     171-5057375-2831560  04-25-22  Shipped - Delivered to Buyer
freq                     12       697                          28741

       Fulfilment Sales Channel ship-service-level Category   Size  \
count       37514         37514              37514    37514  37514
unique          1             1                  1        8     11
top      Merchant     Amazon.in           Standard  T-shirt      M
freq        37514         37514              37514    14062   6806

       Courier Status currency ship-city   ship-state ship-country  \
count           37514    37514     37514        37514        37514
unique              3        1      4698           58            1
```

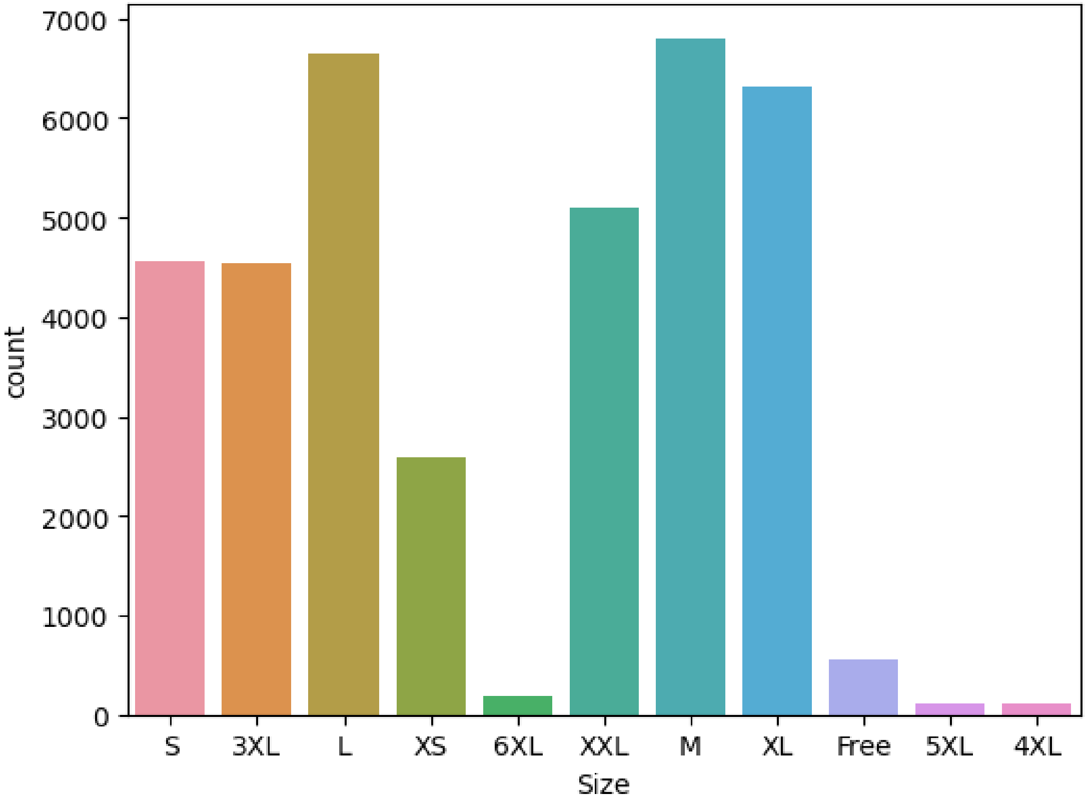
```
top            Shipped      INR  BENGALURU  MAHARASHTRA           IN
freq             31859    37514       2839         6236        37514

       fulfilled-by
count         37514
unique            1
top       Easy Ship
freq          37514
```

EDA

```
[153]: df.columns
```

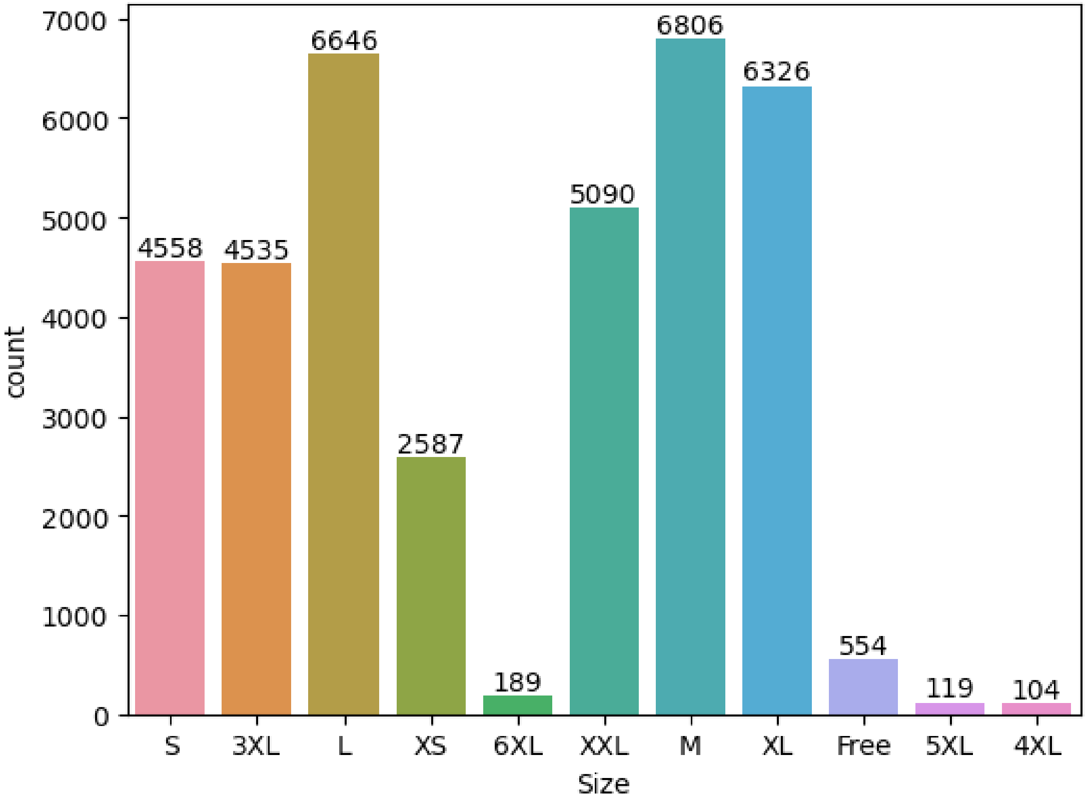
```
[153]: Index(['index', 'Order ID', 'Date', 'Status', 'Fulfilment', 'Sales Channel',
              'ship-service-level', 'Category', 'Size', 'Courier Status', 'Qty',
              'currency', 'Amount', 'ship-city', 'ship-state', 'ship-postal-code',
              'ship-country', 'B2B', 'fulfilled-by'],
             dtype='object')
```

```
[157]: ax=sns.countplot(x='Size',data=df)
```



[159]:

```
ax=sns.countplot(x='Size',data=df)
for bars in ax.containers:
    ax.bar_label(bars)
```



[175]:

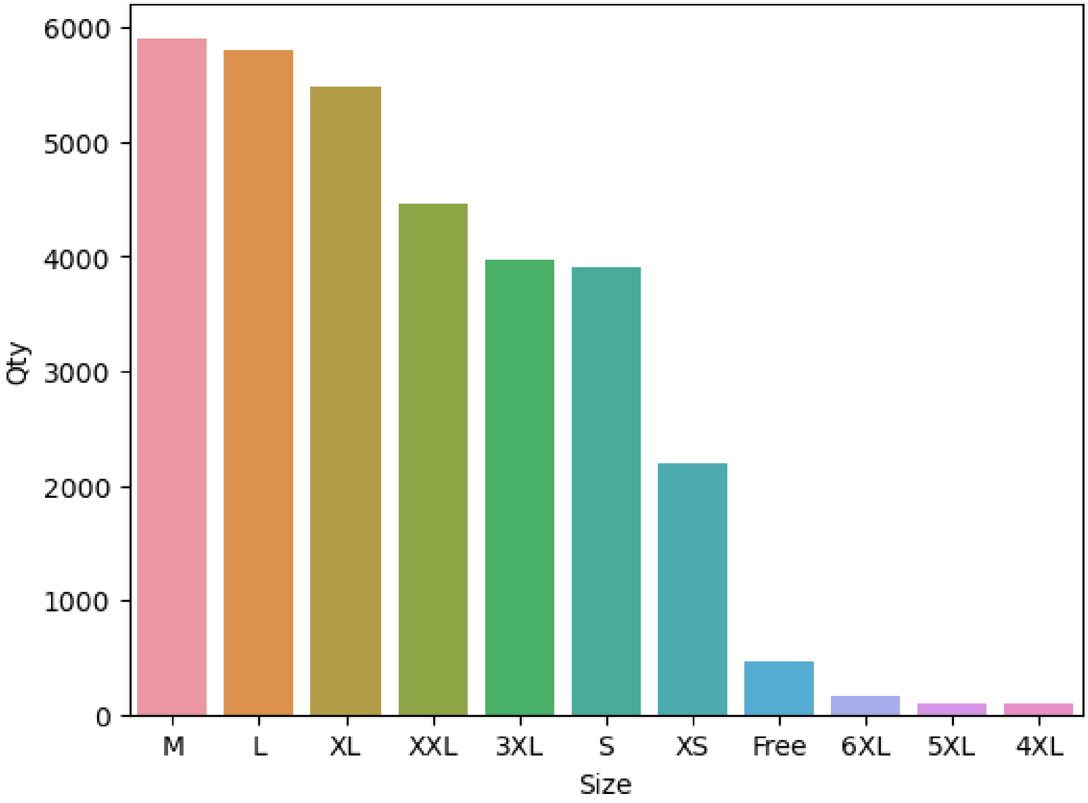
```
df.groupby(['Size'],as_index=False)['Qty'].sum().
 ↪sort_values(by='Qty',ascending=False)
```

[175]:

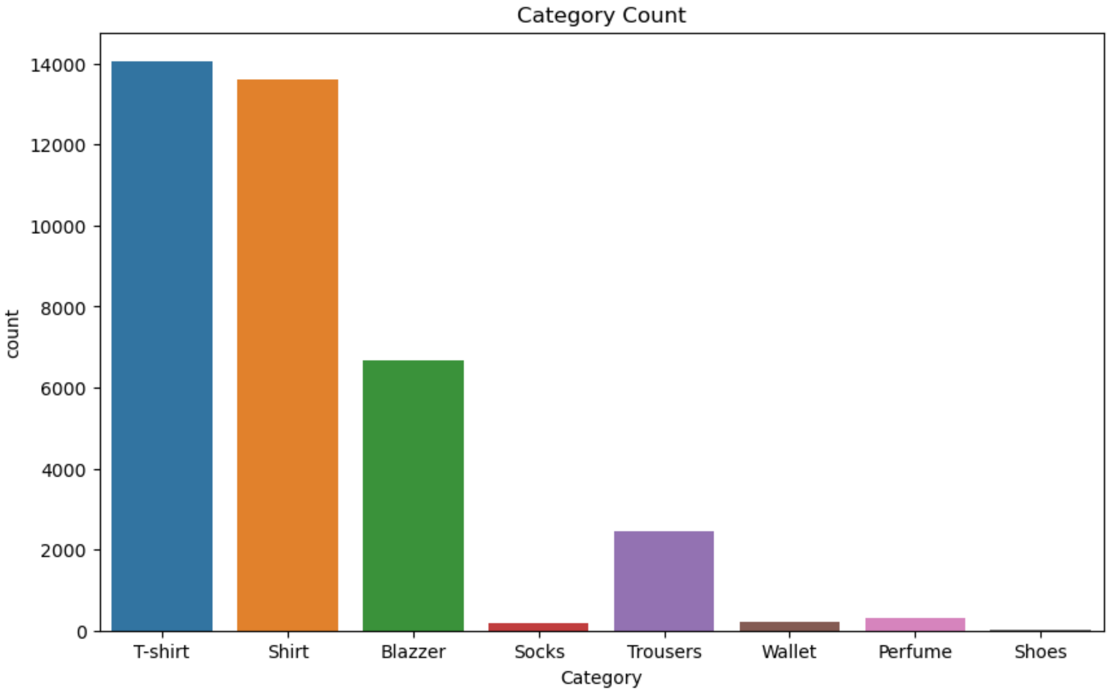
```
    Size   Qty
6      M  5905
5      L  5795
8     XL  5481
10   XXL  4465
0    3XL  3972
7      S  3896
9     XS  2191
4   Free   467
3    6XL   170
2    5XL   104
1    4XL    93
```

[179]:
```
S_Qty=df.groupby(['Size'],as_index=False)['Qty'].sum().
 ↪sort_values(by='Qty',ascending=False)
sns.barplot(x='Size',y='Qty',data=S_Qty)
```

[179]:
```
<Axes: xlabel='Size', ylabel='Qty'>
```

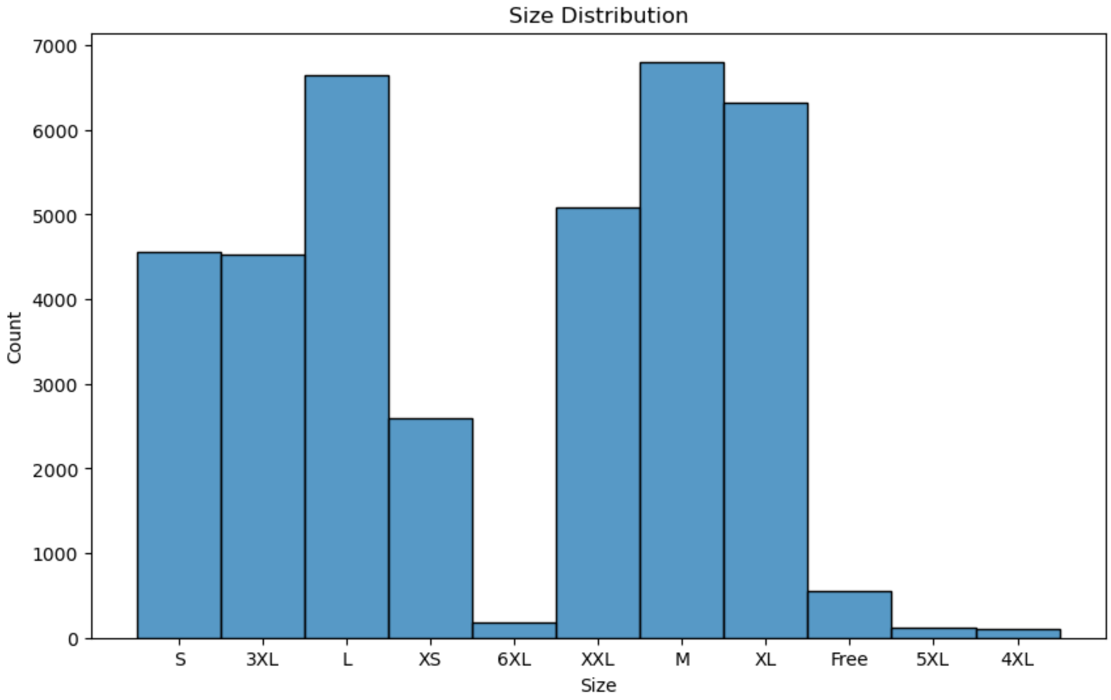


[191]:
```
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='Category')
plt.title('Category Count')
plt.show()
```

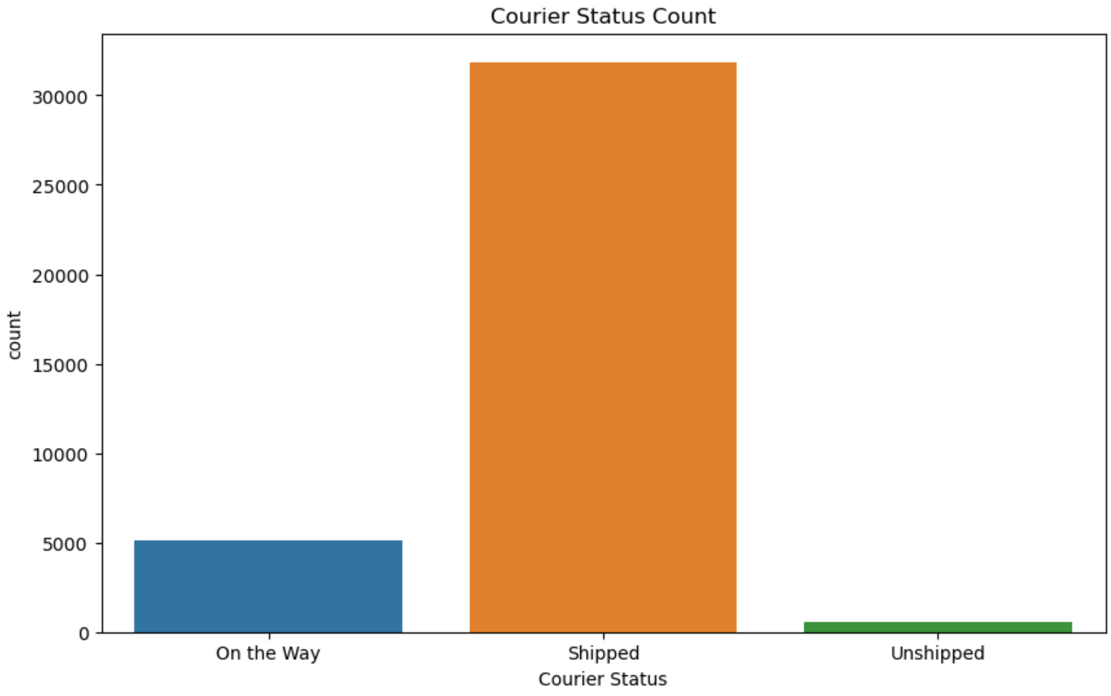


```
[193]: plt.figure(figsize=(10, 6))
       sns.histplot(data=df, x='Size')
       plt.title('Size Distribution')
       plt.show()
```

```
C:\Users\HP\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning:
use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

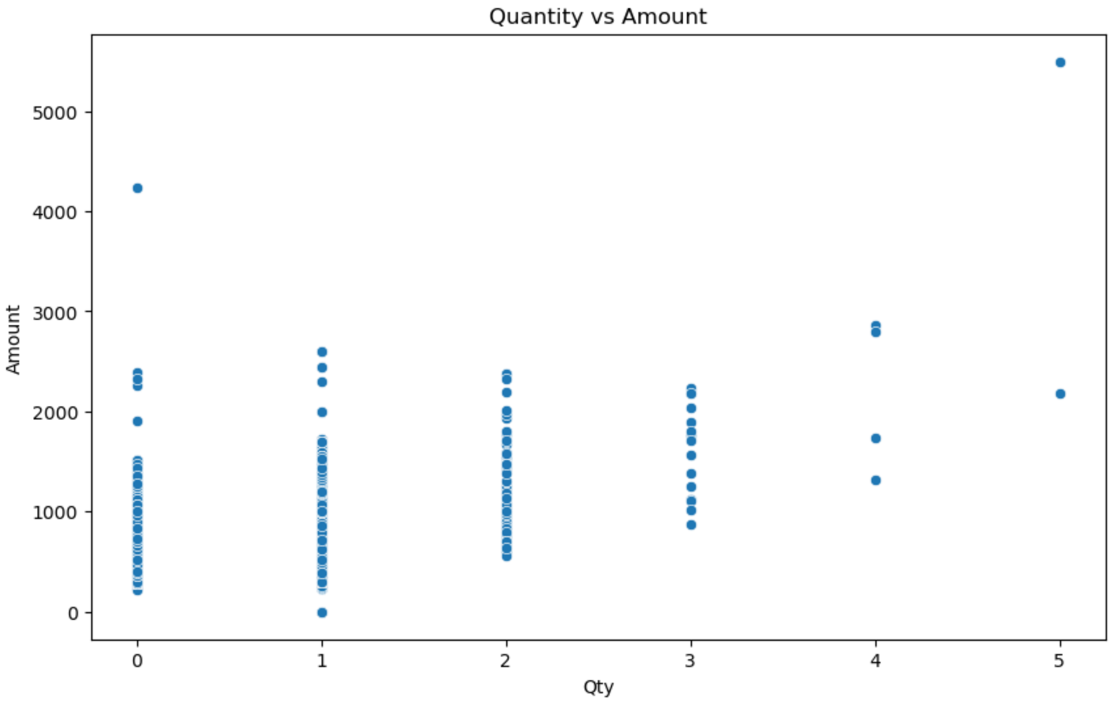


[195]:
```
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='Courier Status')
plt.title('Courier Status Count')
plt.show()
```



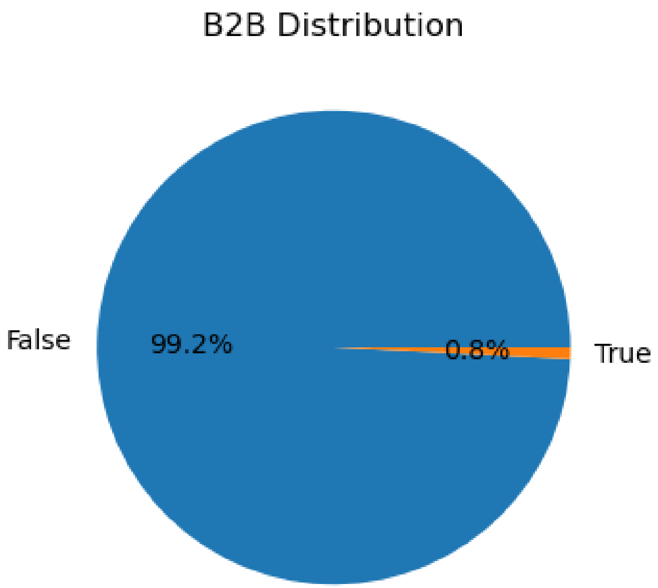
[197]:

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Qty', y='Amount')
plt.title('Quantity vs Amount')
plt.show()
```

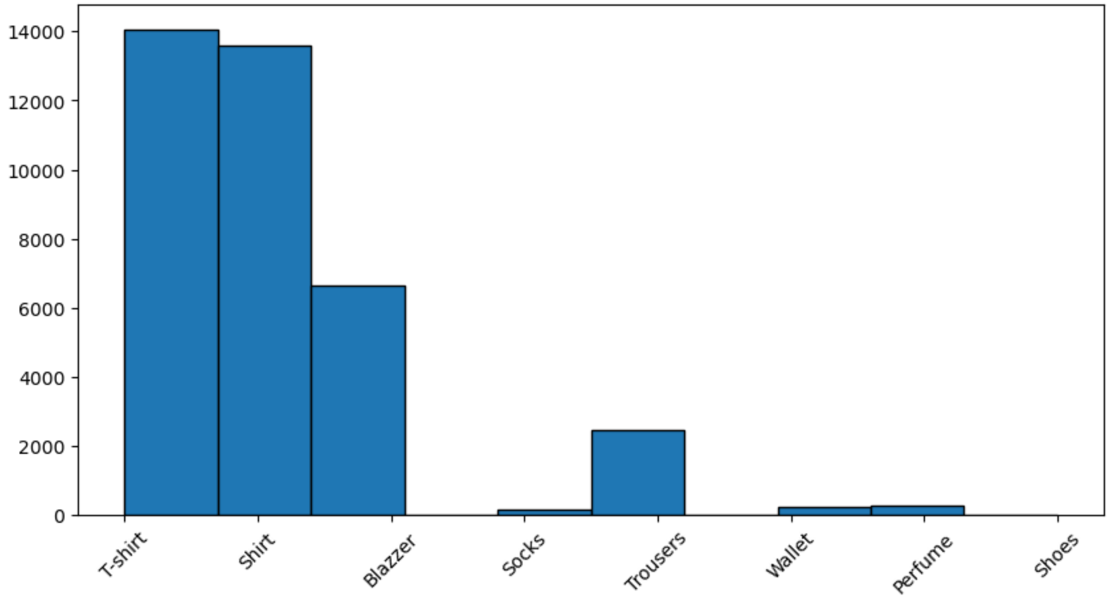


[208]:

```
plt.figure(figsize=(4, 4))
df['B2B'].value_counts().plot.pie(autopct='%1.1f%%')
plt.title('B2B Distribution')
plt.ylabel('')
plt.show()
```

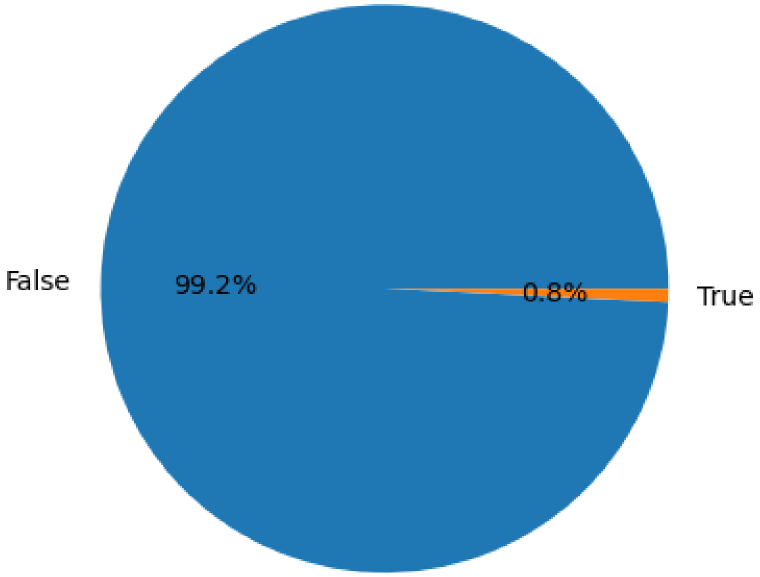


[222]:
```
df['Category']=df['Category'].astype(str)
column_data=df['Category']
plt.figure(figsize=(10, 5))
plt.hist(column_data, bins=10,edgecolor='Black')
plt.xticks(rotation=45)
plt.show()
```



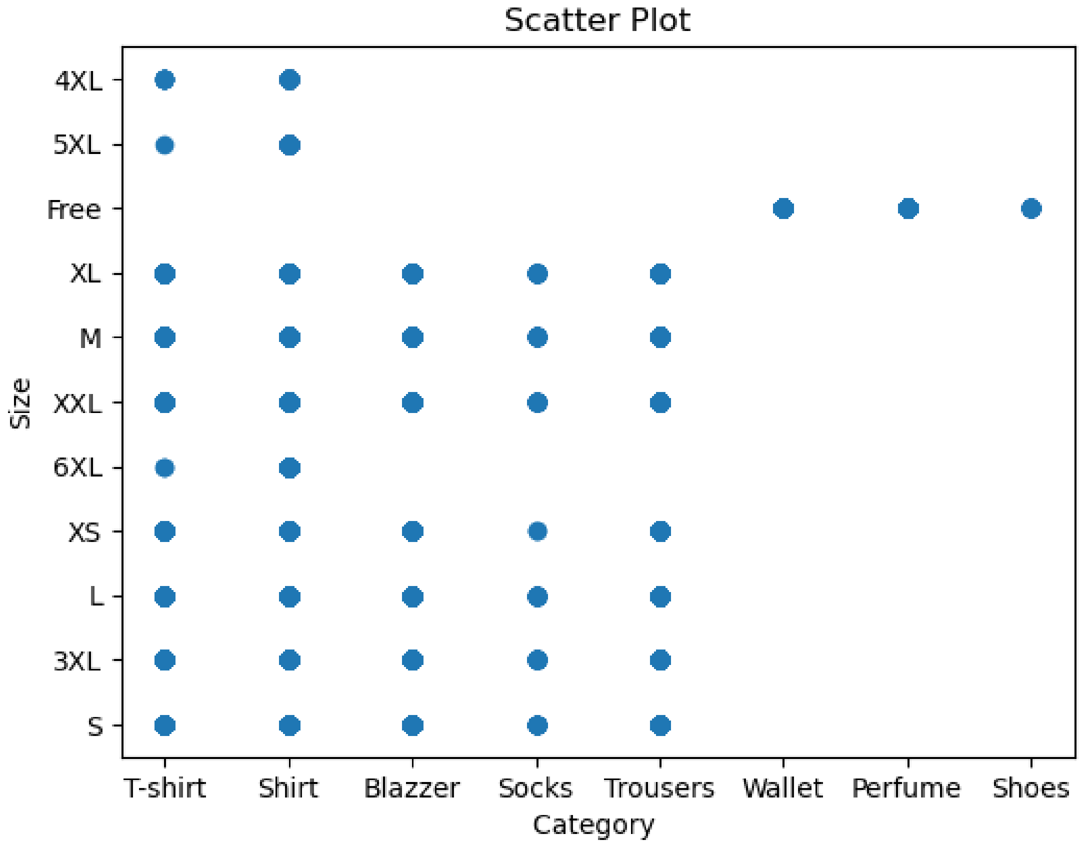
[226]:

```
B2B_Check =df['B2B'].value_counts()
plt.pie(B2B_Check, labels=B2B_Check.index,autopct='%1.1f%%')
plt.show()
```



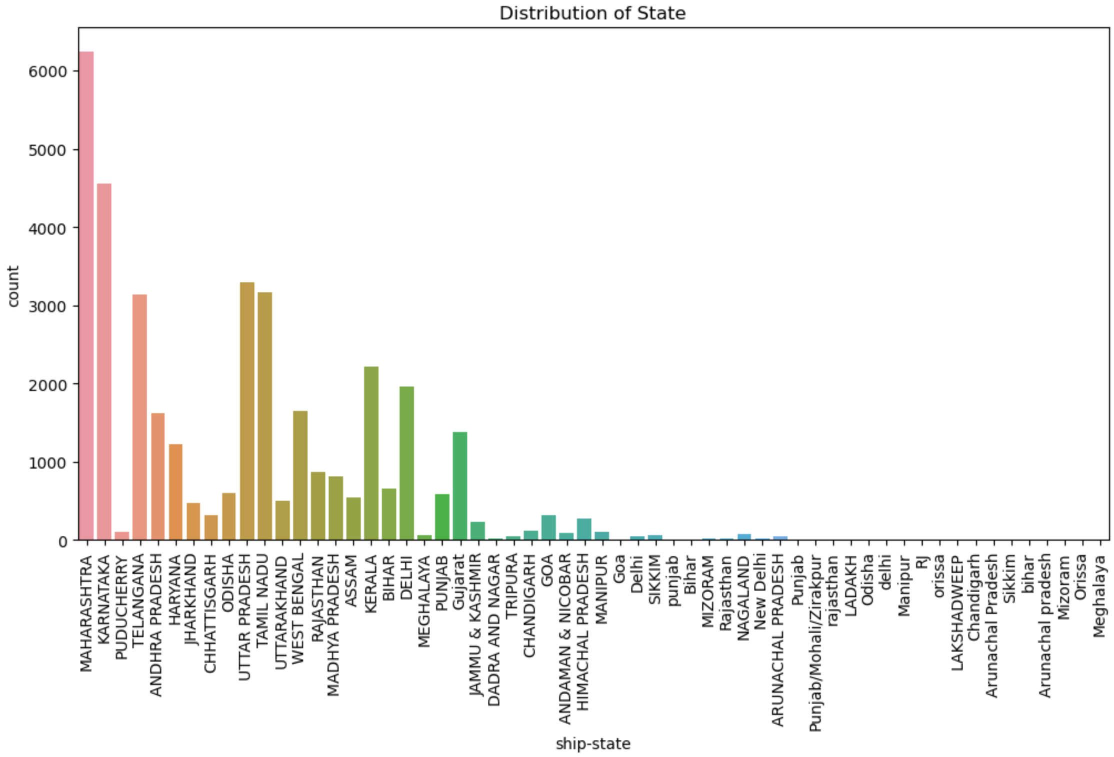
[230]:

```
x_data = df['Category']
y_data = df['Size']

plt.scatter(x_data, y_data)
plt.xlabel('Category ')
plt.ylabel('Size')
plt.title('Scatter Plot')
plt.show()
```



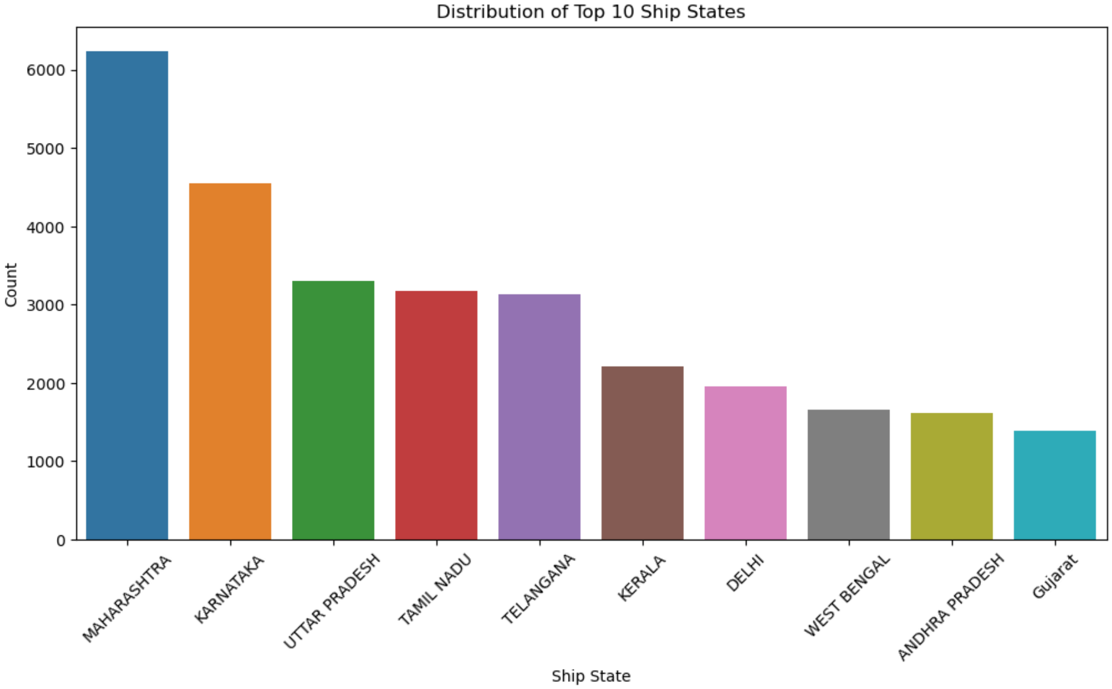
[232]:

```
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x='ship-state')
plt.xlabel('ship-state')
plt.ylabel('count')
plt.title('Distribution of State')
plt.xticks(rotation=90)
plt.show()
```



[242]:

```
top_10_state = df['ship-state'].value_counts().head(10).index

filtered_df = df[df['ship-state'].isin(top_10_state)]

state_order = filtered_df['ship-state'].value_counts().index

sns.countplot(data=filtered_df, x='ship-state', order=state_order)
plt.xlabel('Ship State')
plt.ylabel('Count')
plt.title('Distribution of Top 10 Ship States')
plt.xticks(rotation=45)
plt.show()
```


Distribution of Top 10 Ship States
Count
6000
5000
4000
3000
2000
1000
0
MAHARASHTRA
KARNATAKA
UTTAR PRADESH
TAMIL NADU
TELANGANA
KERALA
DELHI
WEST BENGAL
ANDHRA PRADESH
Gujarat
Ship State

[ ]: