import libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

import dataset

In [2]:
```
df = pd.read_csv(r'C:\Users\karan\Downloads\Suraj_Work\amazon_sales.csv')
```

call datset

In [3]:
```
df
```

Out[3]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category
0	0	405-8078784-5731545	4/30/2022	Cancelled	Merchant	Amazon.in	Standard	T-shir
1	1	171-9198151-1101146	4/30/2022	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shir
2	2	404-0687676-7273146	4/30/2022	Shipped	Amazon	Amazon.in	Expedited	Shir
3	3	403-9615377-8133951	4/30/2022	Cancelled	Merchant	Amazon.in	Standard	Blazze
4	4	407-1069790-7240320	4/30/2022	Shipped	Amazon	Amazon.in	Expedited	Trouser
...	...	...	...	...	...	...	...	..
128971	128970	406-6001380-7673107	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	Shir
128972	128971	402-9551604-7544318	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	T-shir
128973	128972	407-9547469-3152358	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	Blazze
128974	128973	402-6184140-0545956	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	T-shir
128975	128974	408-7436540-8728312	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	T-shir

128976 rows × 19 columns

In [4]:
```
df.info()              # to check dataframe from the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 128976 entries, 0 to 128975
Data columns (total 19 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
dtypes: bool(1), float64(2), int64(2), object(14)
memory usage: 17.8+ MB
```

In [5]:
```
df.head()                    # show top row from dataset
```

Out[5]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category	Size
0	0	405-8078784-5731545	4/30/2022	Cancelled	Merchant	Amazon.in	Standard	T-shirt	S
1	1	171-9198151-1101146	4/30/2022	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	3XL
2	2	404-0687676-7273146	4/30/2022	Shipped	Amazon	Amazon.in	Expedited	Shirt	XL
3	3	403-9615377-8133951	4/30/2022	Cancelled	Merchant	Amazon.in	Standard	Blazzer	L
4	4	407-1069790-7240320	4/30/2022	Shipped	Amazon	Amazon.in	Expedited	Trousers	3XL

In [6]:
```
df.tail()                    # show bottom row from dataset
```

Out[6]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category
128971	128970	406-6001380-7673107	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	Shirt
128972	128971	402-9551604-7544318	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	T-shirt
128973	128972	407-9547469-3152358	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	Blazzer
128974	128973	402-6184140-0545956	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	T-shirt
128975	128974	408-7436540-8728312	5/31/2022	Shipped	Amazon	Amazon.in	Expedited	T-shirt

In [7]:
```
pd.isnull(df)                    # to check any values are null or not in the
```

Out[7]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category	Size	Courie Statu
0	False	False	False	False	False	False	False	False	False	Fals
1	False	False	False	False	False	False	False	False	False	Fals
2	False	False	False	False	False	False	False	False	False	Fals
3	False	False	False	False	False	False	False	False	False	Fals
4	False	False	False	False	False	False	False	False	False	Fals
...	...	...	...	...	...	...	...	...	...	
128971	False	False	False	False	False	False	False	False	False	Fals
128972	False	False	False	False	False	False	False	False	False	Fals
128973	False	False	False	False	False	False	False	False	False	Fals
128974	False	False	False	False	False	False	False	False	False	Fals
128975	False	False	False	False	False	False	False	False	False	Fals

128976 rows × 19 columns

In [ ]:
```
pd.isnull(df).sum()                    # check/calculate the total null values
```

Out[ ]:
```
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
dtype: int64
```

In [9]:
```
df.shape                    # gave count/totaL of no.of columns and rows from datse
```

Out[9]:
```
(128976, 19)
```

In [10]:
```
df.dropna(inplace=True)             # remove/drop null values
```

In [11]:
```
df.shape
```

Out[11]:
```
(37514, 19)
```

In [12]:
```
pd.isnull(df).sum()
```

Out[12]:
```
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             0
Amount               0
ship-city            0
ship-state           0
ship-postal-code     0
ship-country         0
B2B                  0
fulfilled-by         0
dtype: int64
```

In [13]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 37514 entries, 0 to 128892
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   index               37514 non-null  int64
 1   Order ID            37514 non-null  object
 2   Date                37514 non-null  object
 3   Status              37514 non-null  object
 4   Fulfilment          37514 non-null  object
 5   Sales Channel       37514 non-null  object
 6   ship-service-level  37514 non-null  object
 7   Category            37514 non-null  object
 8   Size                37514 non-null  object
 9   Courier Status      37514 non-null  object
 10  Qty                 37514 non-null  int64
 11  currency            37514 non-null  object
 12  Amount              37514 non-null  float64
 13  ship-city           37514 non-null  object
 14  ship-state          37514 non-null  object
 15  ship-postal-code    37514 non-null  float64
 16  ship-country        37514 non-null  object
 17  B2B                 37514 non-null  bool
 18  fulfilled-by        37514 non-null  object
dtypes: bool(1), float64(2), int64(2), object(14)
memory usage: 5.5+ MB
```

In [14]:
```
df.head()
```

Out[14]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category	Size
0	0	405-8078784-5731545	4/30/2022	Cancelled	Merchant	Amazon.in	Standard	T-shirt	S
1	1	171-9198151-1101146	4/30/2022	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	3XL
3	3	403-9615377-8133951	4/30/2022	Cancelled	Merchant	Amazon.in	Standard	Blazzer	L
7	7	406-7807733-3785945	4/30/2022	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	S
12	12	405-5513694-8146768	4/30/2022	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	XS

In [15]:
```
df['ship-postal-code']=df['ship-postal-code'].astype('int')            # change a da
```

In [ ]:
```
df['ship-postal-code'].dtype                              # check data type of perticul
```

Out[ ]:
```
dtype('int64')
```

In [18]:
```
df['Date']=pd.to_datetime(df['Date'])                  # chane data type of date column
```

In [19]:
```
df['Date'].dtype
```

Out[19]:
```
dtype('<M8[ns]')
```

In [20]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 37514 entries, 0 to 128892
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   index               37514 non-null  int64
 1   Order ID            37514 non-null  object
 2   Date                37514 non-null  datetime64[ns]
 3   Status              37514 non-null  object
 4   Fulfilment          37514 non-null  object
 5   Sales Channel       37514 non-null  object
 6   ship-service-level  37514 non-null  object
 7   Category            37514 non-null  object
 8   Size                37514 non-null  object
 9   Courier Status      37514 non-null  object
 10  Qty                 37514 non-null  int64
 11  currency            37514 non-null  object
 12  Amount              37514 non-null  float64
 13  ship-city           37514 non-null  object
 14  ship-state          37514 non-null  object
 15  ship-postal-code    37514 non-null  int64
 16  ship-country        37514 non-null  object
 17  B2B                 37514 non-null  bool
 18  fulfilled-by        37514 non-null  object
dtypes: bool(1), datetime64[ns](1), float64(1), int64(3), object(13)
memory usage: 5.5+ MB
```

In [21]:
```
df.describe()
```

Out[21]:

	index	Date	Qty	Amount	ship-postal-code
count	37514.000000	37514	37514.000000	37514.000000	37514.000000
mean	60953.809858	2022-05-11 07:56:47.303939840	0.867383	646.553960	463291.552754
min	0.000000	2022-03-31 00:00:00	0.000000	0.000000	110001.000000
25%	27235.250000	2022-04-20 00:00:00	1.000000	458.000000	370465.000000
50%	63470.500000	2022-05-09 00:00:00	1.000000	629.000000	500019.000000
75%	91790.750000	2022-06-01 00:00:00	1.000000	771.000000	600042.000000
max	128891.000000	2022-06-29 00:00:00	5.000000	5495.000000	989898.000000
std	36844.853039	NaN	0.354160	279.952414	194550.425637

In [ ]:
```
df.describe(include='object')       # Describe our data with object data type
```

Out[ ]:

	Order ID	Status	Fulfilment	Sales Channel	ship-service-level	Category	Size	Courier Status	cu
count	37514	37514	37514	37514	37514	37514	37514	37514	
unique	34664	11	1	1	1	8	11	3	
top	171-5057375-2831560	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	T-shirt	M	Shipped	
freq	12	28741	37514	37514	37514	14062	6806	31859	

In [23]:
```
df[['Qty','Amount']].describe()           # Describe for specific columns
```

Out[23]:

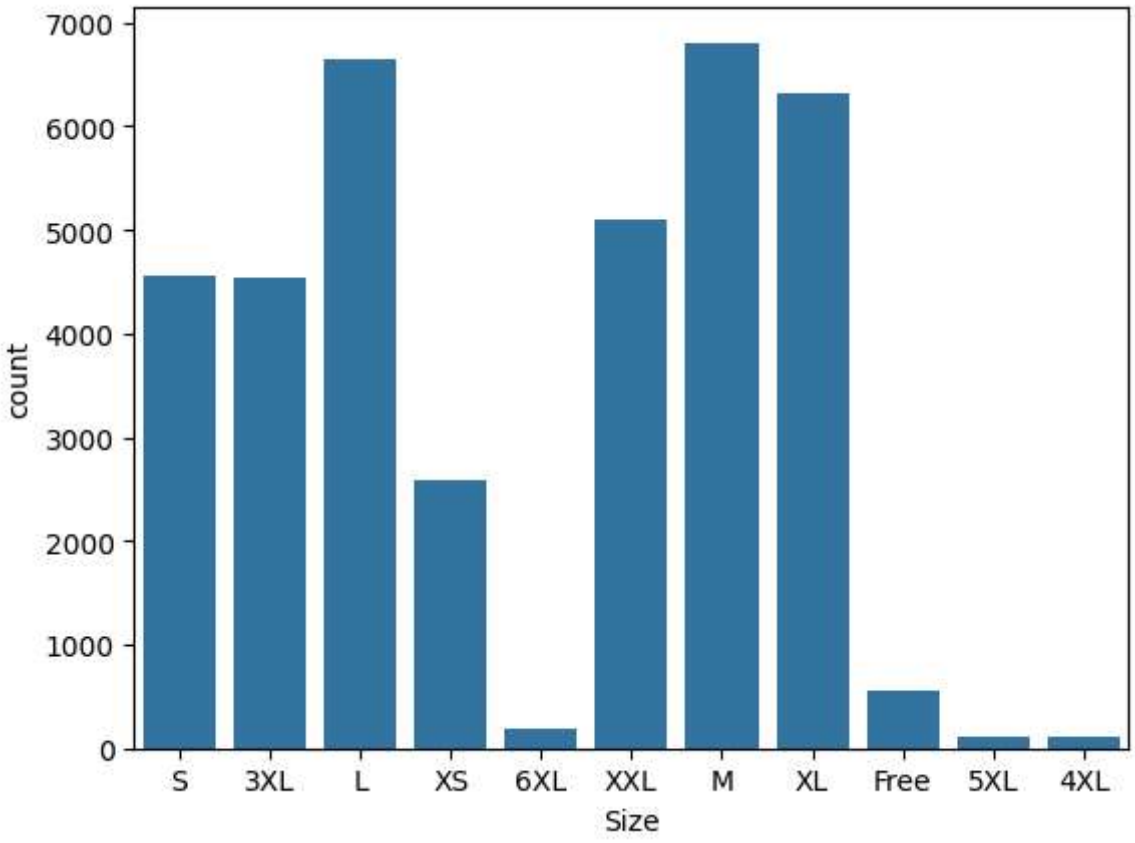
	Qty	Amount
count	37514.000000	37514.000000
mean	0.867383	646.553960
std	0.354160	279.952414
min	0.000000	0.000000
25%	1.000000	458.000000
50%	1.000000	629.000000
75%	1.000000	771.000000
max	5.000000	5495.000000

In [24]:
```
df.head()
```

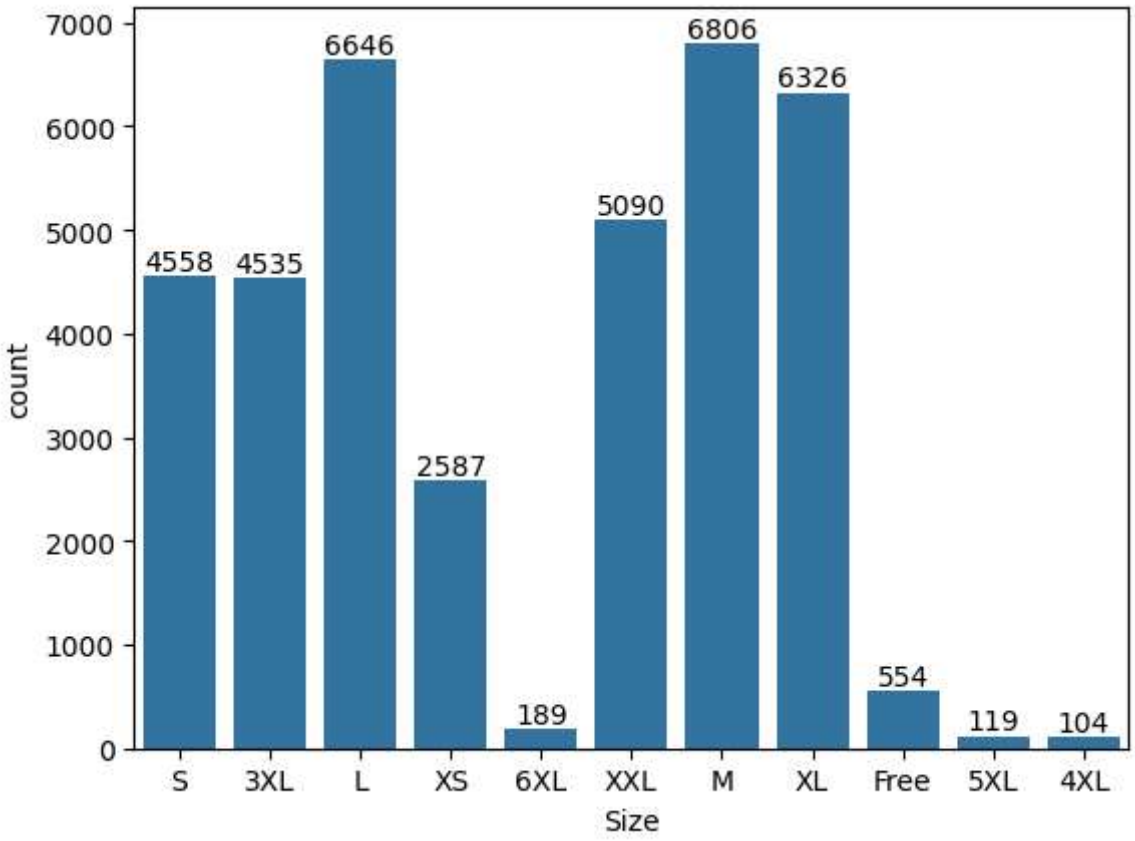
Out[24]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Category	Size	Cc S
0	0	405-8078784-5731545	2022-04-30	Cancelled	Merchant	Amazon.in	Standard	T-shirt	S	C
1	1	171-9198151-1101146	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	3XL	Sh
3	3	403-9615377-8133951	2022-04-30	Cancelled	Merchant	Amazon.in	Standard	Blazzer	L	C
7	7	406-7807733-3785945	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	S	Sh
12	12	405-5513694-8146768	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	Shirt	XS	Sh

In [25]:
```
ax= sns.countplot(x='Size', data = df)          # Analysis on Size
```



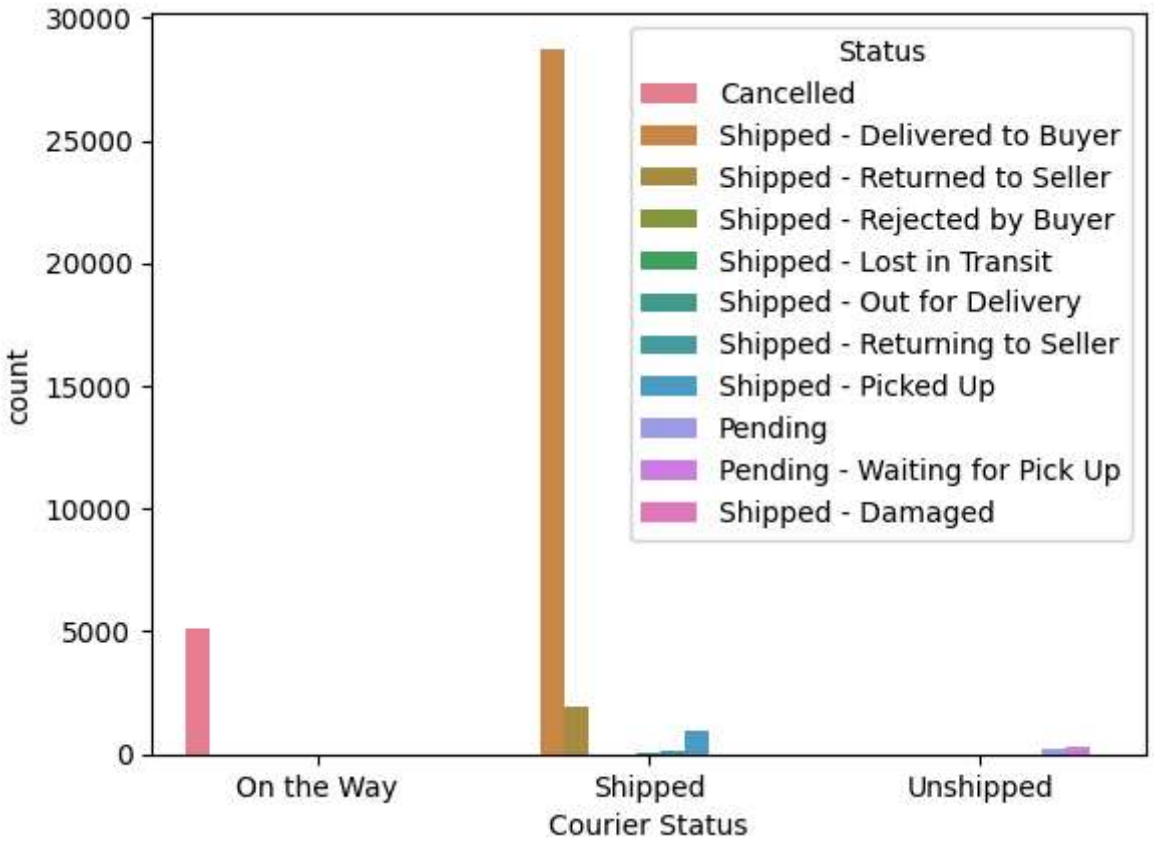
In [27]:
```
ax= sns.countplot(x='Size', data = df)
for bars in ax.containers:                    # checking data label in ax(Size)
    ax.bar_label(bars)
```



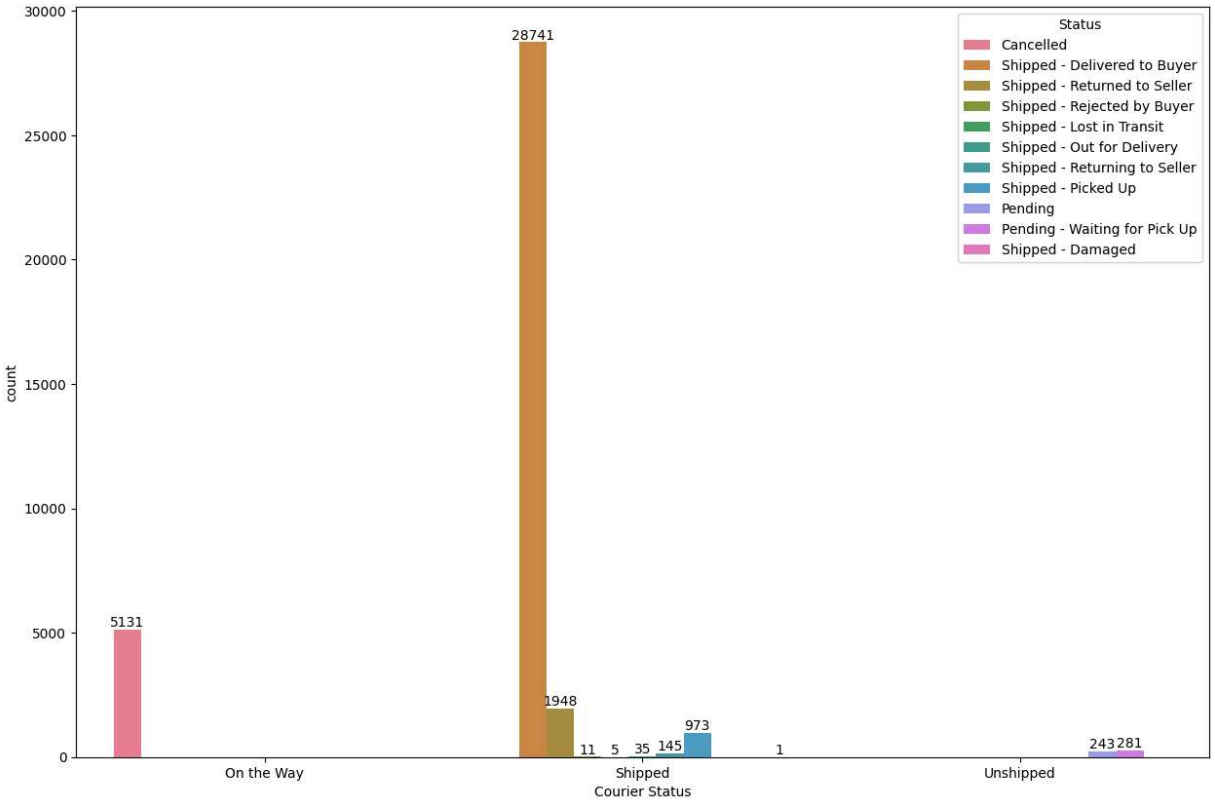
Most of the people buys "M" Size and the least people buys "4XL" size.

In [28]:
```
sns.countplot(x='Courier Status', data = df, hue='Status')     #  check Courior Sta
```

Out[28]:
```
<Axes: xlabel='Courier Status', ylabel='count'>
```



In [ ]:
```
plt.figure(figsize=(15,10))                                    # increase or decreas
ay = sns.countplot(x='Courier Status', data = df, hue='Status')
for bars in ay.containers:                                     # using for condition
    ay.bar_label(bars)
plt.show()                                                      # shows all the figu
```



Majority of the order are shipped through the courier.

In [34]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 37514 entries, 0 to 128892
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   index               37514 non-null  int64
 1   Order ID            37514 non-null  object
 2   Date                37514 non-null  datetime64[ns]
 3   Status              37514 non-null  object
 4   Fulfilment          37514 non-null  object
 5   Sales Channel       37514 non-null  object
 6   ship-service-level  37514 non-null  object
 7   Category            37514 non-null  object
 8   Size                37514 non-null  object
 9   Courier Status      37514 non-null  object
 10  Qty                 37514 non-null  int64
 11  currency            37514 non-null  object
 12  Amount              37514 non-null  float64
 13  ship-city           37514 non-null  object
 14  ship-state          37514 non-null  object
 15  ship-postal-code    37514 non-null  int64
 16  ship-country        37514 non-null  object
 17  B2B                 37514 non-null  bool
 18  fulfilled-by        37514 non-null  object
dtypes: bool(1), datetime64[ns](1), float64(1), int64(3), object(13)
memory usage: 5.5+ MB
```

In [41]:
```
df['Category']= df['Category'].astype('string')
```

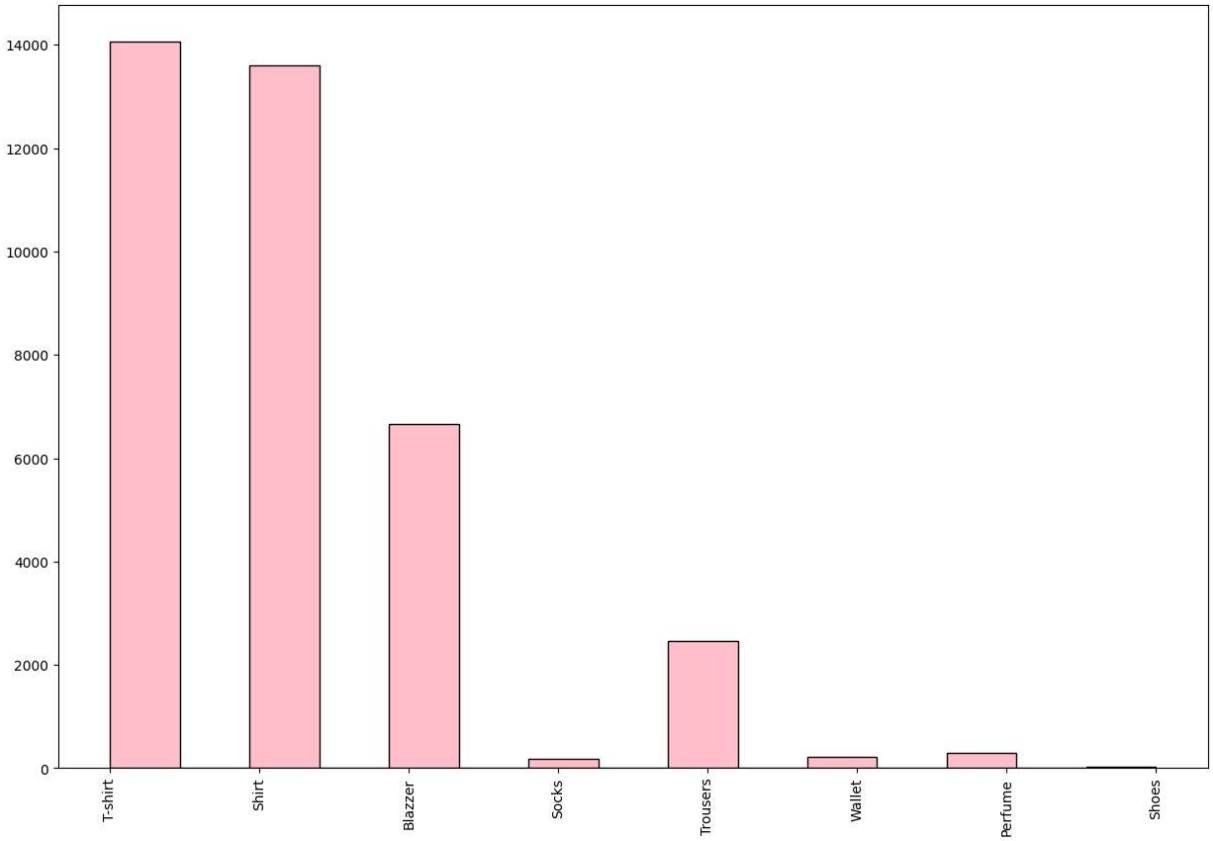
In [42]:
```
df['Category'].dtype
```

Out[42]: string[python]

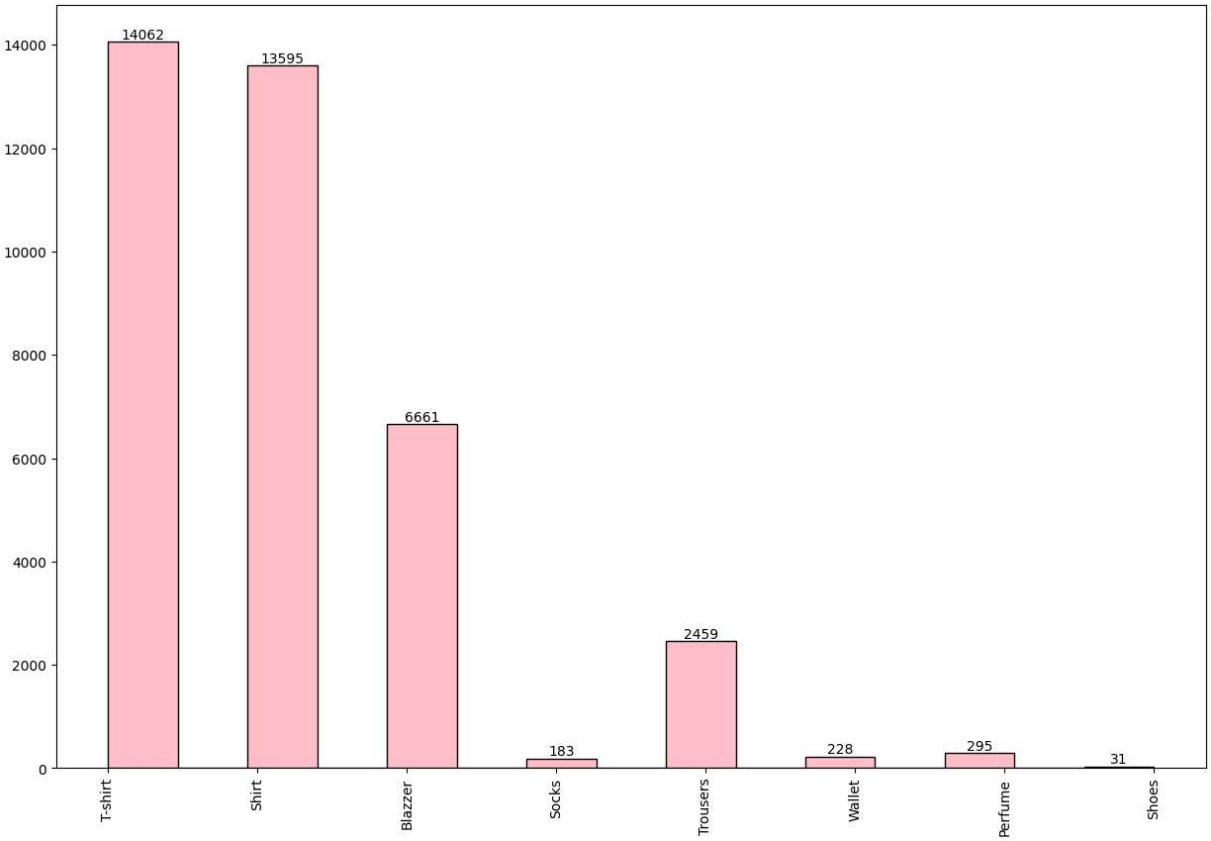
In [43]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 37514 entries, 0 to 128892
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype         
---  ------              --------------  -----         
 0   index               37514 non-null  int64         
 1   Order ID            37514 non-null  object        
 2   Date                37514 non-null  datetime64[ns]
 3   Status              37514 non-null  object        
 4   Fulfilment          37514 non-null  object        
 5   Sales Channel       37514 non-null  object        
 6   ship-service-level  37514 non-null  object        
 7   Category            37514 non-null  string        
 8   Size                37514 non-null  object        
 9   Courier Status      37514 non-null  object        
 10  Qty                 37514 non-null  int64         
 11  currency            37514 non-null  object        
 12  Amount              37514 non-null  float64       
 13  ship-city           37514 non-null  object        
 14  ship-state          37514 non-null  object        
 15  ship-postal-code    37514 non-null  int64         
 16  ship-country        37514 non-null  object        
 17  B2B                 37514 non-null  bool          
 18  fulfilled-by        37514 non-null  object        
dtypes: bool(1), datetime64[ns](1), float64(1), int64(3), object(12), string(1)
memory usage: 5.5+ MB
```

In [56]:
```
c_d=df['Category']
plt.figure(figsize=(15,10))
plt.hist(c_d,bins=15,edgecolor="Black",color='pink')
plt.xticks(rotation=90)                                  # show topic name rotation on
plt.show()
```

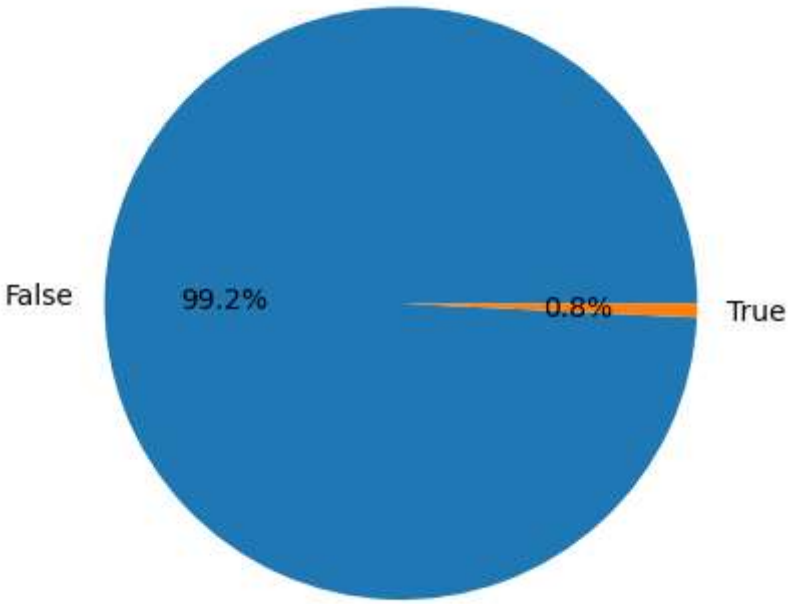


In [57]:
```
c_d=df['Category']
plt.figure(figsize=(15,10))
n, bins, patches = plt.hist(c_d,bins=15,edgecolor="Black",color='pink')
for count, patch in zip(n,patches):
    if count > 0:                                   # Only add labels to bars with a
        x = patch.get_x() + patch.get_width() / 2  # Center of the bar
        y = patch.get_height()                     # Height of the bar
        plt.text(x, y, int(count), ha='center', va='bottom', fontsize=10)
plt.xticks(rotation=90)                             # show topic name rotation on
plt.show()
```



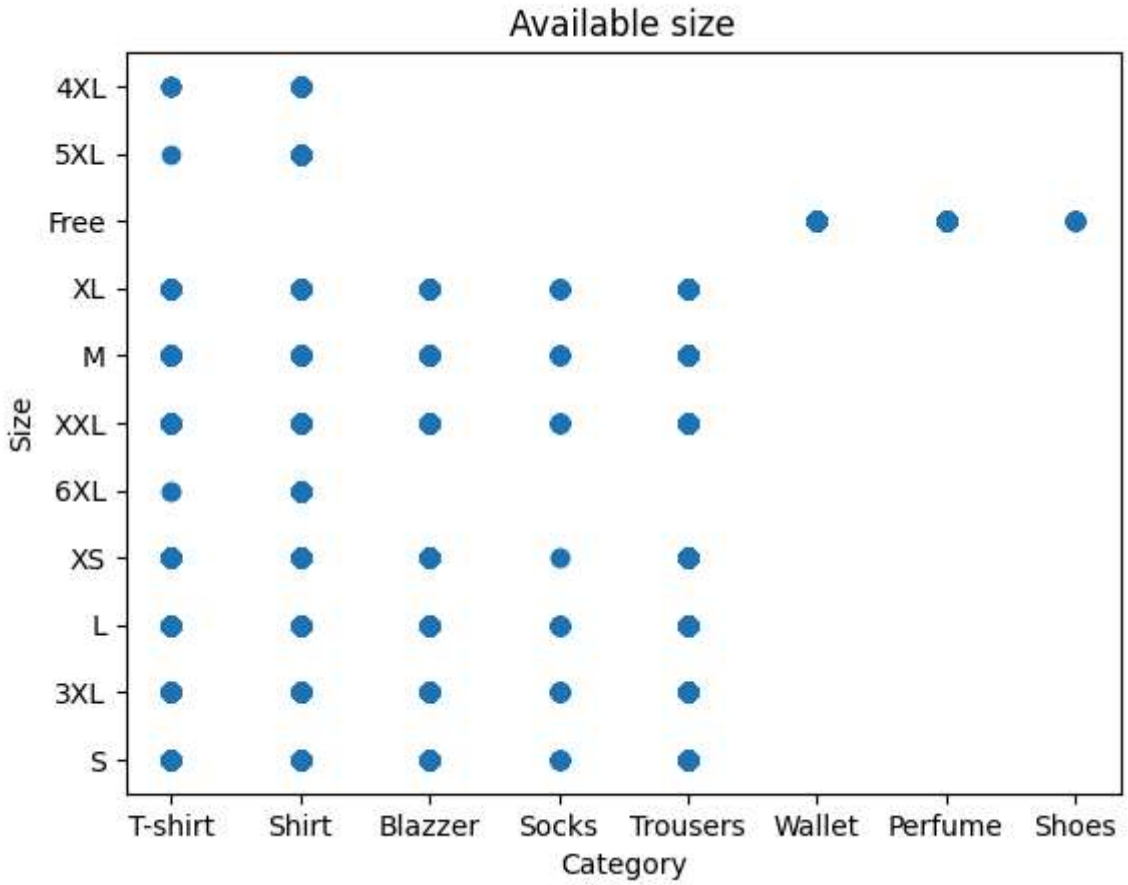
Most of the people buys T-shirt

In [62]:
```
check_B2B = df['B2B'].value_counts()
plt.pie(check_B2B,labels= check_B2B.index, autopct='%1.1f%%')
plt.show()
```

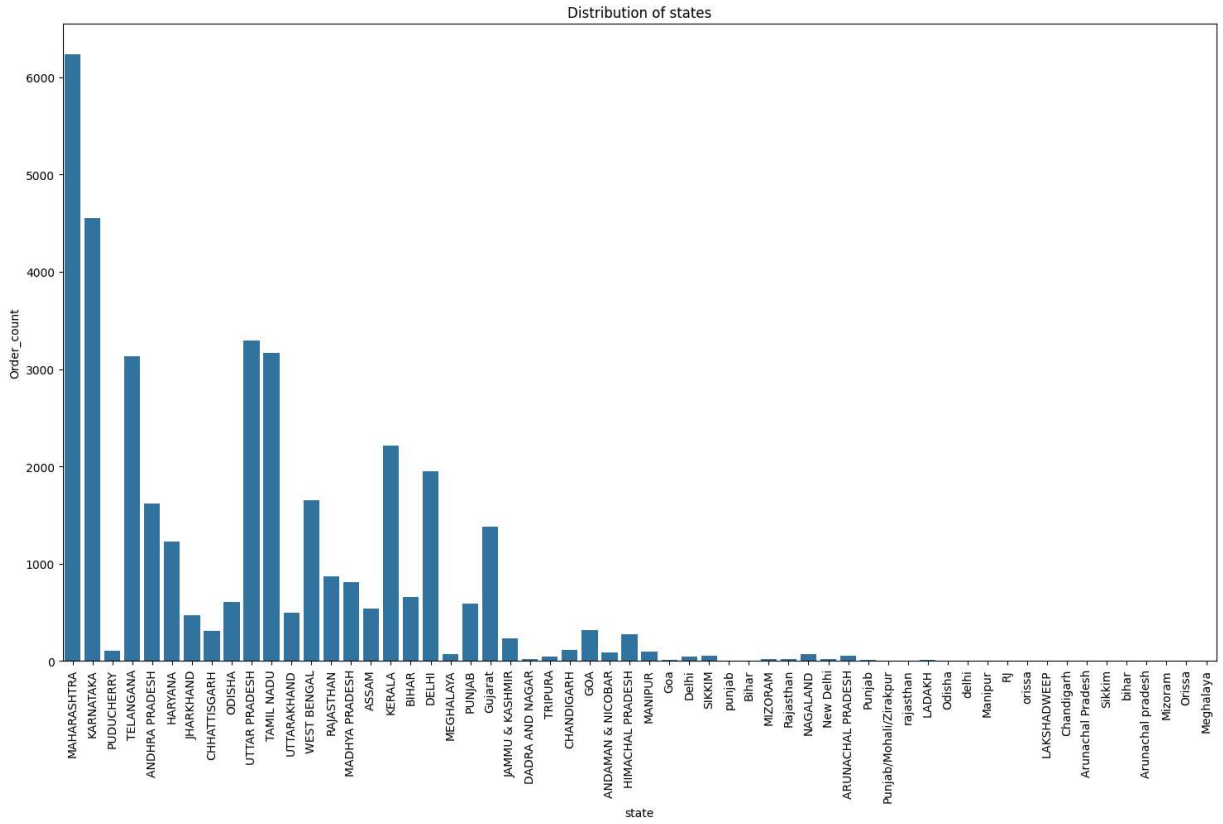


Up to 99.2% of buyers are retailers, while 0.8% are wholesalers.

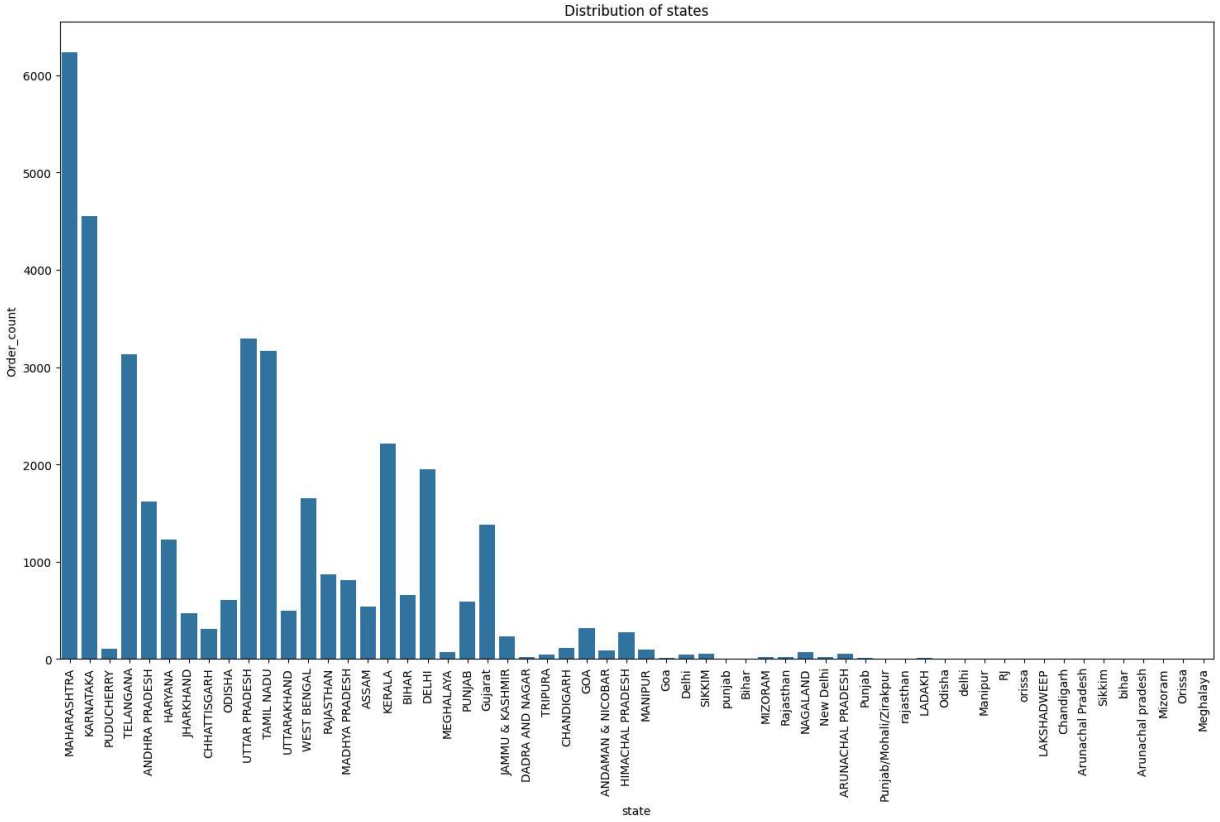
In [63]:
```
#scatter plot
x_data = df['Category']
y_data = df['Size']

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



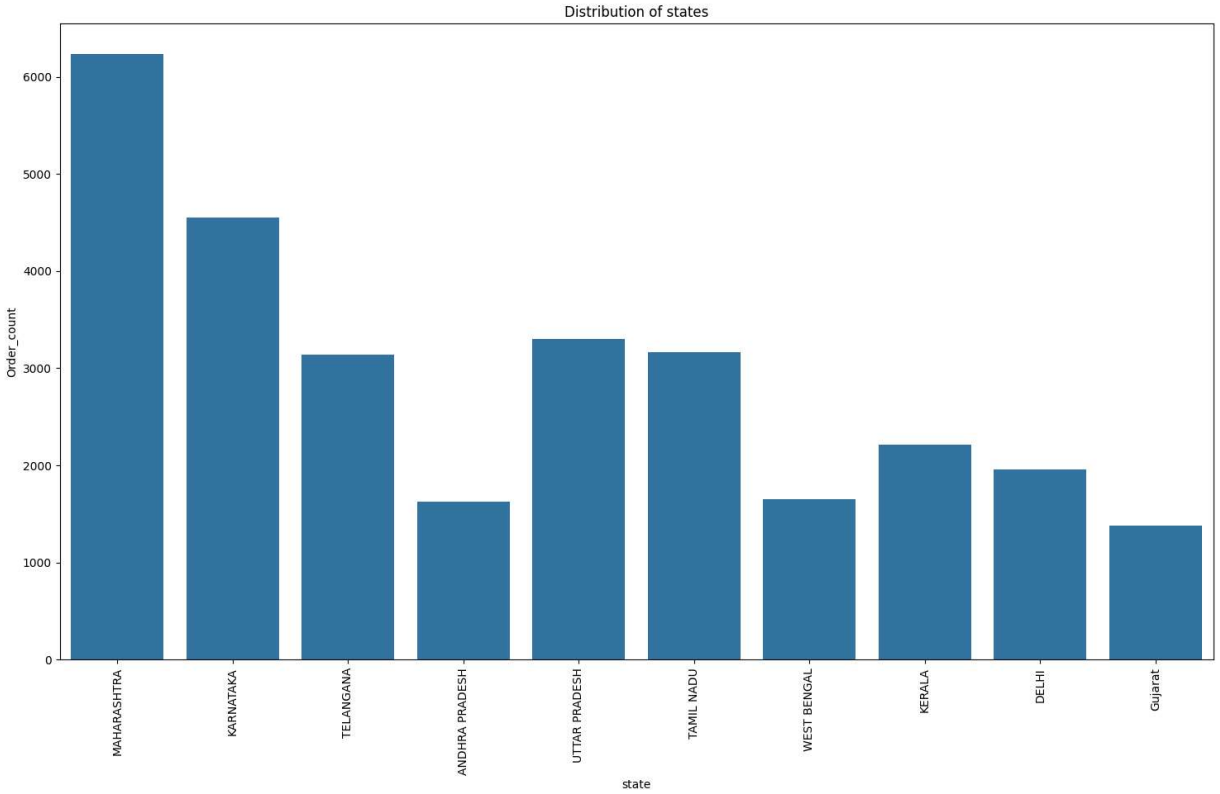
In [72]:
```
plt.figure(figsize=(18,10))
sns.countplot(data= df,x='ship-state')
plt.xlabel('state')
plt.ylabel('Order_count')
plt.title('Distribution of states')
plt.xticks(rotation=90)
plt.show()
```



In [ ]:
```
top10_state = df['ship-state'].value_counts().head(10)                    ## Top Ten
plt.figure(figsize=(18,10))
sns.countplot(data= df,x='ship-state')
plt.xlabel('state')
plt.ylabel('Order_count')
plt.title('Distribution of states')
plt.xticks(rotation=90)
plt.show()
```



In [75]:
```
top10_state = df['ship-state'].value_counts().head(10)                    ## Top Ten
plt.figure(figsize=(18,10))
sns.countplot(data= df[df['ship-state'].isin(top10_state.index)],x='ship-state')
plt.xlabel('state')
plt.ylabel('Order_count')
plt.title('Distribution of states')
plt.xticks(rotation=90)
plt.show()
```



Most of the buyers are from the state of Maharashtra.

conclusion :

Size Preferences: The Medium size is the most purchased (6,806 units), followed closely by Large (6,646 units) and XL (6,326 units). Small and 3XL sizes have lower demand, with 4,558 and 4,535 units sold, respectively.

Shipping Performance: A total of 28,741 items were successfully delivered to buyers.Only 1,948 items were returned to the seller, indicating a low return rate of approximately 6.3%.

Product Popularity: T-shirts are the most popular product, with 14,062 units sold.Shirts are the second most purchased product (13,595 units), followed by Blazers (6,631 units) and Trousers (2,459 units).

Geographical Insights: Most buyers are located in Maharashtra, Karnataka, Uttar Pradesh, Tamil Nadu, and Telangana, highlighting these states as key markets for Amazon sales.

Demand Concentration: Medium-sized products and T-shirts dominate the sales, reflecting their broad appeal across buyer demographics. Maharashtra and Karnataka likely contribute significantly to the overall sales volume.

In [ ]: