

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Load Data
```

```
df=pd.read_csv('Amazon Sale Report.csv', low_memory=False)
```

```
df.head()
```

	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	Style
SKU \				
0	Merchant	Amazon.in	Standard	SET389 SET389-KR-NP-S
1	Merchant	Amazon.in	Standard	JNE3781 JNE3781-KR-XXXL
2	Amazon	Amazon.in	Expedited	JNE3371 JNE3371-KR-XL
3	Merchant	Amazon.in	Standard	J0341 J0341-DR-L
4	Amazon	Amazon.in	Expedited	JNE3671 JNE3671-TU-XXXL

	Category	...	currency	Amount	ship-city	ship-state
\						
0	Set	...	INR	647.62	MUMBAI	MAHARASHTRA
1	kurta	...	INR	406.00	BENGALURU	KARNATAKA
2	kurta	...	INR	329.00	NAVI MUMBAI	MAHARASHTRA
3	Western Dress	...	INR	753.33	PUDUCHERRY	PUDUCHERRY
4	Top	...	INR	574.00	CHENNAI	TAMIL NADU

	ship-postal-code	ship-country
\		
0	400081.0	IN
1	560085.0	IN
2	410210.0	IN
3	605008.0	IN
4	600073.0	IN

	promotion-ids	B2B
\		
fulfilled-by		
0	NaN	False Easy

```

Ship
1 Amazon PLCC Free-Financing Universal Merchant ... False Easy
Ship
2 IN Core Free Shipping 2015/04/08 23-48-5-108 True
NaN
3 NaN False Easy
Ship
4 NaN False
NaN

```

```

    Unnamed: 22
0      NaN
1      NaN
2      NaN
3      NaN
4      NaN

```

```
[5 rows x 24 columns]
```

```
df.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 128975 entries, 0 to 128974
Data columns (total 24 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   index                                128975 non-null  int64
1   Order ID                             128975 non-null  object
2   Date                                 128975 non-null  object
3   Status                               128975 non-null  object
4   Fulfilment                           128975 non-null  object
5   Sales Channel                         128975 non-null  object
6   ship-service-level                   128975 non-null  object
7   Style                                128975 non-null  object
8   SKU                                  128975 non-null  object
9   Category                             128975 non-null  object
10  Size                                 128975 non-null  object
11  ASIN                                 128975 non-null  object
12  Courier Status                        122103 non-null  object
13  Qty                                  128975 non-null  int64
14  currency                             121180 non-null  object
15  Amount                               121180 non-null  float64
16  ship-city                             128942 non-null  object
17  ship-state                           128942 non-null  object
18  ship-postal-code                     128942 non-null  float64
19  ship-country                         128942 non-null  object
20  promotion-ids                        79822 non-null  object
21  B2B                                  128975 non-null  bool
22  fulfilled-by                         39277 non-null  object
23  Unnamed: 22                         79925 non-null  object

```

```
dtypes: bool(1), float64(2), int64(2), object(19)
memory usage: 22.8+ MB
```

```
# Display summary statistics for all categorical (object-type) columns
# including count, number of unique values, most frequent value, and
its frequency
```

```
df.describe(include='O')
```

	Order ID	Date	Status	Fulfilment	Sales
Channel \					
count	128975	128975	128975	128975	
128975					
unique	120378	91	13	2	
2					
top	171-5057375-2831560	05-03-22	Shipped	Amazon	
Amazon.in					
freq	12	2085	77804	89698	
128851					

	ship-service-level	Style	SKU	Category	Size
ASIN \					
count	128975	128975	128975	128975	128975
128975					
unique	2	1377	7195	9	11
7190					
top	Expedited	JNE3797	JNE3797-KR-L	Set	M
B09SDXFFQ1					
freq	88615	4224	773	50284	22711
773					

	Courier	Status	currency	ship-city	ship-state	ship-country	\
count	122103	122103	121180	128942	128942	128942	
unique	3	3	1	8955	69	1	
top	Shipped	Shipped	INR	BENGALURU	MAHARASHTRA	IN	
freq	109487	109487	121180	11217	22260	128942	

	promotion-ids	fulfilled-by
Unnamed: 22		
count	79822	39277
79925		
unique	5787	1
1		
top	IN Core Free Shipping	2015/04/08 23-48-5-108
False		Easy Ship
freq	46100	39277
79925		

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

index	0
Order ID	0
Date	0
Status	0
Fulfilment	0
Sales Channel	0
ship-service-level	0
Style	0
SKU	0
Category	0
Size	0
ASIN	0
Courier Status	6872
Qty	0
currency	7795
Amount	7795
ship-city	33
ship-state	33
ship-postal-code	33
ship-country	33
promotion-ids	49153
B2B	0
fulfilled-by	89698
Unnamed: 22	49050
dtype:	int64

```
# Drop unnecessary columns
df.drop(columns= ['Order ID',
                  'Date',
                  'Sales Channel ',
                  'Style',
                  'SKU',
                  'Size',
                  'ASIN',
                  'Courier Status',
                  'Qty',
                  'currency',
                  'ship-city',
                  'ship-state',
                  'ship-postal-code',
                  'ship-country',
                  'B2B',
                  'fulfilled-by',
                  'index', 'Unnamed: 22'
                ],
        inplace = True)

df
```

	Status	Fulfilment	ship-service-level	\
0	Cancelled	Merchant	Standard	
1	Shipped - Delivered to Buyer	Merchant	Standard	
2	Shipped	Amazon	Expedited	
3	Cancelled	Merchant	Standard	
4	Shipped	Amazon	Expedited	
...	
128970	Shipped	Amazon	Expedited	
128971	Shipped	Amazon	Expedited	
128972	Shipped	Amazon	Expedited	
128973	Shipped	Amazon	Expedited	
128974	Shipped	Amazon	Expedited	

	Category	Amount	\
0	Set	647.62	
1	kurta	406.00	
2	kurta	329.00	
3	Western Dress	753.33	
4	Top	574.00	
...	
128970	kurta	517.00	
128971	Set	999.00	
128972	Western Dress	690.00	
128973	Set	1199.00	
128974	Set	696.00	

	promotion-ids
0	NaN
1	Amazon PLCC Free-Financing Universal Merchant ...
2	IN Core Free Shipping 2015/04/08 23-48-5-108
3	NaN
4	NaN
...	...
128970	NaN
128971	IN Core Free Shipping 2015/04/08 23-48-5-108
128972	NaN
128973	IN Core Free Shipping 2015/04/08 23-48-5-108
128974	IN Core Free Shipping 2015/04/08 23-48-5-108

[128975 rows x 6 columns]

```
# Replace missing values in the 'promotion-ids' column with 'no'
df['promotion-ids'].fillna('no', inplace=True)
df["promotion-ids"]
```

0	no
1	Amazon PLCC Free-Financing Universal Merchant ...
2	IN Core Free Shipping 2015/04/08 23-48-5-108
3	no
4	no

```

...
128970                                     no
128971      IN Core Free Shipping 2015/04/08 23-48-5-108
128972                                     no
128973      IN Core Free Shipping 2015/04/08 23-48-5-108
128974      IN Core Free Shipping 2015/04/08 23-48-5-108
Name: promotion-ids, Length: 128975, dtype: object

# Apply promotion classification and overwrite the original
# 'promotion-ids' column with the new categories
def classify_promotion(promo):
    if promo.startswith('Amazon PLCC Free-Financing Universal'):
        return 'PLCC'
    elif promo.startswith('IN Core Free Shipping'):
        return 'FreeShipping'
    elif promo.strip() == 'no':
        return 'no'
    else:
        return 'Other'

df['promotion-ids'] = df['promotion-ids'].apply(classify_promotion)

df['promotion-ids'].value_counts()

promotion-ids
no                49153
FreeShipping      46198
PLCC               32348
Other              1276
Name: count, dtype: int64

df['Status'].value_counts()

Status
Shipped                77804
Shipped - Delivered to Buyer  28769
Cancelled              18332
Shipped - Returned to Seller   1953
Shipped - Picked Up           973
Pending                 658
Pending - Waiting for Pick Up  281
Shipped - Returning to Seller  145
Shipped - Out for Delivery     35
Shipped - Rejected by Buyer    11
Shipping                 8
Shipped - Lost in Transit      5
Shipped - Damaged              1
Name: count, dtype: int64

# Filter the DataFrame to keep only completed or cancelled orders
# Encode 'Cancelled' as 1 and others as 0 in the 'Status' column

```

```

completed_status = [
    "Shipped",
    "Shipped - Delivered to Buyer",
    "Shipped - Picked Up",
    "Pending",
    "Shipping",
    "Pending - Waiting for Pick Up"]
cancelled_status = ["Cancelled"]
df = df[df["Status"].isin(completed_status + cancelled_status)].copy()
df["Status"] = df["Status"].apply(lambda x: 1 if x == "Cancelled" else
0)

df['Status'].value_counts()

Status
0    108493
1     18332
Name: count, dtype: int64

# Encode categorical variables numerically
df["Fulfilment"] = df["Fulfilment"].map({"Amazon": 0, "Merchant": 1})
df["ship-service-level"] = df["ship-service-level"].map({"Standard":
0, "Expedited": 1})
df["promotion-ids"] = df["promotion-ids"].map({"no": 0,
"FreeShipping": 1, "PLCC": 2, "Other": 3})

top_n = 10
top_categories = df["Category"].value_counts().nlargest(top_n).index
df["Category_name"] = df["Category"].apply(lambda x: x if x in
top_categories else "Other")
df["Category"] = df["Category_name"].astype("category").cat.codes

df["Amount"] = df["Amount"].fillna(0)

df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 126825 entries, 0 to 128974
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Status                126825 non-null  int64
1   Fulfilment            126825 non-null  int64
2   ship-service-level    126825 non-null  int64
3   Category              126825 non-null  int8
4   Amount                126825 non-null  float64
5   promotion-ids         126825 non-null  int64
6   Category_name         126825 non-null  object
dtypes: float64(1), int64(4), int8(1), object(1)
memory usage: 6.9+ MB

```

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

```
Status      0
Fulfilment   0
ship-service-level  0
Category     0
Amount       0
promotion-ids 0
Category_name 0
dtype: int64
```

```
df
```

	Status	Fulfilment	ship-service-level	Category	Amount	\
0	1	1	0	5	647.62	
1	0	1	0	8	406.00	
2	0	0	1	8	329.00	
3	1	1	0	7	753.33	
4	0	0	1	6	574.00	
...	
128970	0	0	1	8	517.00	
128971	0	0	1	5	999.00	
128972	0	0	1	7	690.00	
128973	0	0	1	5	1199.00	
128974	0	0	1	5	696.00	

	promotion-ids	Category_name
0	0	Set
1	2	kurta
2	1	kurta
3	0	Western Dress
4	0	Top
...
128970	0	kurta
128971	1	Set
128972	0	Western Dress
128973	1	Set
128974	1	Set

```
[126825 rows x 7 columns]
```