```
import pandas as pd

# Load the dataset (replace 'amazon_sales_data.csv' with your actual file)
df = pd.read_csv('/content/Amazon Sale Report.csv')

# Display the first few rows of the dataset
print(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  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]
<ipython-input-1-db299168e1a2>:4: DtypeWarning: Columns (17) have mixed types. Specify dtype option on
  df = pd.read_csv('/content/Amazon Sale Report.csv')
```

```
# Check for missing values in each column
missing_values = df.isnull().sum()

# Display columns with missing values
print("Columns with missing values:")
print(missing_values[missing_values > 0])
```

```
Columns with missing values:
currency              6440
Amount                6440
ship-city               28
ship-state              28
ship-postal-code        29
ship-country            29
B2B                      1
fulfilled-by         72673
New                 106443
PendingS            106443
```

```
dtype: int64


# Remove rows with any missing values
df_dropped = df.dropna()

# Alternatively, remove columns with any missing values
df_dropped_cols = df.dropna(axis=1)



# Fill missing values with the mean of the column (for numeric columns only)
numeric_columns = df.select_dtypes(include=['number']).columns
df_filled_mean = df.copy()  # Create a copy to avoid modifying the original DataFrame
df_filled_mean[numeric_columns] = df_filled_mean[numeric_columns].fillna(df_filled_mean[numeric_columns].mea

# Display the first few rows of the DataFrame with filled means
print(df_filled_mean.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  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]
```

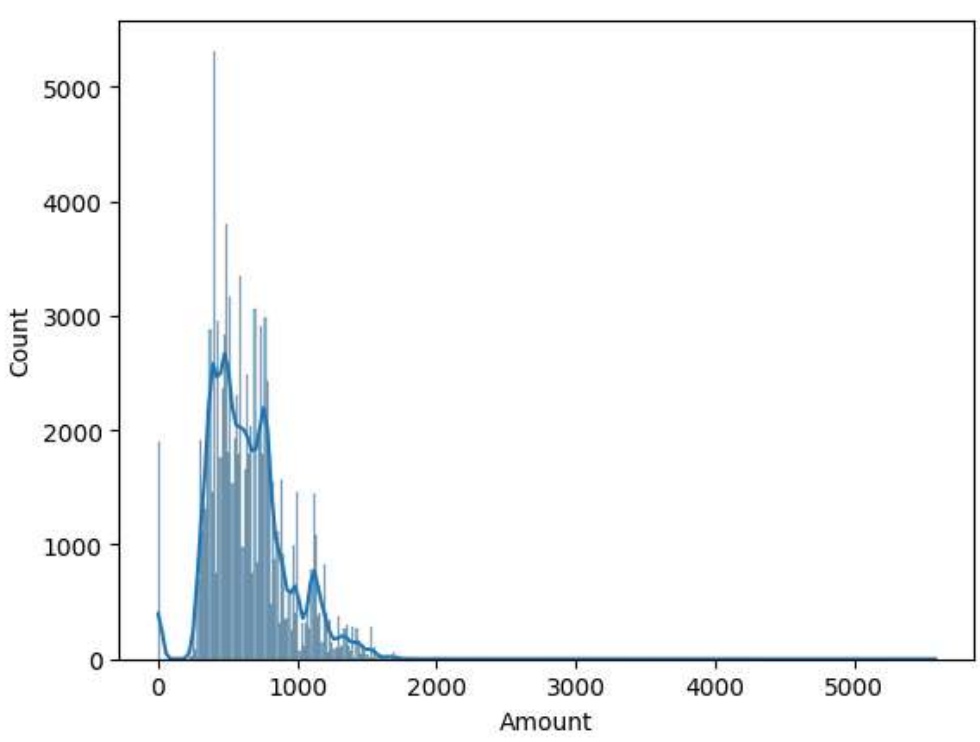
```
# Fill missing values with the mean of the column (for numeric columns only)
numeric_columns = df.select_dtypes(include=['number']).columns
df_filled_mean = df.copy()  # Create a copy to avoid modifying the original DataFrame
df_filled_mean[numeric_columns] = df_filled_mean[numeric_columns].fillna(df_filled_mean[numeric_columns].me

# Display the first few rows of the DataFrame with filled means
print(df_filled_mean.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  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]
```

```
# Save the cleaned dataset to a new CSV file
# Replace 'df_filled_mean' with the appropriate DataFrame you want to save
df_cleaned = df_filled_mean
df_cleaned.to_csv('amazon_sales_data_cleaned.csv', index=False)
```

```
import seaborn as sns
import matplotlib.pyplot as plt

# Example: Distribution of a numerical column
sns.histplot(df['Amount'], kde=True)
plt.show()

# Example: Relationship between two variables
sns.scatterplot(x='Fulfilment', y='Amount', data=df)
plt.show()
```




```
# Example: Creating a new feature based on existing ones
# Convert 'Fulfilment' column to numeric type, handling potential errors
df['Fulfilment'] = pd.to_numeric(df['Fulfilment'], errors='coerce')

# Calculate the new 'Category' feature
df['Category'] = df['Amount'] / df['Fulfilment']
```