In [17]:

```
pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\moham\anaconda3\lib\site-packag
es (2.1.4)
Requirement already satisfied: numpy<2,>=1.23.2 in c:\users\moham\anaconda3\lib\s
ite-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\moham\anaconda3
\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\moham\anaconda3\lib\site-
packages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\moham\anaconda3\lib\sit
e-packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\moham\anaconda3\lib\site-pack
ages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [1]:

```
import pandas as pd
# Load the dataset
df = pd.read_csv(r"C:\Users\moham\my project\Amazon Sales data.csv")
df
```

Out[1]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Shi
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6
...	...	...	...	...	...	...	...	
95	Sub-Saharan Africa	Mali	Clothes	Online	M	7/26/2011	512878119	9/3
96	Asia	Malaysia	Fruits	Offline	L	11/11/2011	810711038	12/28
97	Sub-Saharan Africa	Sierra Leone	Vegetables	Offline	C	6/1/2016	728815257	6/29
98	North America	Mexico	Personal Care	Offline	M	7/30/2015	559427106	8/8
99	Sub-Saharan Africa	Mozambique	Household	Offline	L	2/10/2012	665095412	2/15

100 rows × 14 columns

In [2]:

```
# Print the column names to identify the exact names
print(df.columns)
```

```
Index(['Region', 'Country', 'Item Type', 'Sales Channel', 'Order Priority',
       'Order Date', 'Order ID', 'Ship Date', 'Units Sold', 'Unit Price',
       'Unit Cost', 'Total Revenue', 'Total Cost', 'Total Profit'],
      dtype='object')
```

In [3]:

```
# Specify the relevant columns based on the actual column names
order_date_col = 'Order Date'  # Replace with the correct column name if differe
ship_date_col = 'Ship Date'  # Replace with the correct column name if different
sales_col = 'Total Revenue'  # Replace with the correct column name if different
```

In [4]:

```
# Drop rows with null values for the specified columns
df_cleaned = df.dropna(subset=[order_date_col, ship_date_col, sales_col])
```

In [5]:

```
# Convert 'Order Date' and 'Ship Date' to datetime format
df_cleaned[order_date_col] = pd.to_datetime(df_cleaned[order_date_col])
df_cleaned[ship_date_col] = pd.to_datetime(df_cleaned[ship_date_col])
```

In [6]:

```
# Extract Year, Month, and Month-Year from 'Order Date'
df_cleaned['Year'] = df_cleaned[order_date_col].dt.year
df_cleaned['Month'] = df_cleaned[order_date_col].dt.strftime('%B')
df_cleaned['MonthYear'] = df_cleaned[order_date_col].dt.strftime('%b %Y')
```

In [7]:

```
# Display the cleaned and transformed dataset
print(df_cleaned.head())
```

```
                              Region                Country        Item Type  \
0              Australia and Oceania                 Tuvalu        Baby Food   
1  Central America and the Caribbean                Grenada           Cereal   
2                             Europe                 Russia  Office Supplies   
3                 Sub-Saharan Africa  Sao Tome and Principe           Fruits   
4                 Sub-Saharan Africa                 Rwanda  Office Supplies   

  Sales Channel Order Priority Order Date   Order ID  Ship Date  Units Sold  \
0       Offline              H 2010-05-28  669165933 2010-06-27        9925   
1        Online              C 2012-08-22  963881480 2012-09-15        2804   
2       Offline              L 2014-05-02  341417157 2014-05-08        1779   
3        Online              C 2014-06-20  514321792 2014-07-05        8102   
4       Offline              L 2013-02-01  115456712 2013-02-06        5062   

   Unit Price  Unit Cost  Total Revenue  Total Cost  Total Profit  Year  \
0      255.28     159.42     2533654.00  1582243.50     951410.50  2010   
1      205.70     117.11      576782.80   328376.44     248406.36  2012   
2      651.21     524.96     1158502.59   933903.84     224598.75  2014   
3        9.33       6.92       75591.66    56065.84      19525.82  2014   
4      651.21     524.96     3296425.02  2657347.52     639077.50  2013   

      Month MonthYear  
0       May  May 2010  
1    August  Aug 2012  
2       May  May 2014  
3      June  Jun 2014  
4  February  Feb 2013  
```

In [8]:

```
# Creating the Date Table
date_table = pd.DataFrame({
    'Date': pd.date_range(start=df_cleaned[order_date_col].min(), end=df_cleaned
})
date_table['Year'] = date_table['Date'].dt.year
date_table['Month'] = date_table['Date'].dt.strftime('%B')
date_table['MonthYear'] = date_table['Date'].dt.strftime('%b %Y')
```

In [9]:

```
print(date_table.head())
```

```
        Date  Year     Month MonthYear
0 2010-02-02  2010  February  Feb 2010
1 2010-02-03  2010  February  Feb 2010
2 2010-02-04  2010  February  Feb 2010
3 2010-02-05  2010  February  Feb 2010
4 2010-02-06  2010  February  Feb 2010
```

In [10]:

```
# Calculate Total Sales
total_sales = df_cleaned[sales_col].sum()
```

```
print(f'Total Sales: {total_sales}')
```

```
Total Sales: 137348768.31
```

In [15]:

```
# Save the cleaned dataset to a new CSV file
df_cleaned.to_csv('cleaned_amazon_sales_data.csv', index=False)
df_cleaned
```

Out[15]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Ur S
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	99
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	28
2	Europe	Russia	Office Supplies	Offline	L	2014-05-02	341417157	2014-05-08	17
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-07-05	81
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2013-02-01	115456712	2013-02-06	50
...	...	...	...	...	...	...	...	...	
95	Sub-Saharan Africa	Mali	Clothes	Online	M	2011-07-26	512878119	2011-09-03	8
96	Asia	Malaysia	Fruits	Offline	L	2011-11-11	810711038	2011-12-28	62
97	Sub-Saharan Africa	Sierra Leone	Vegetables	Offline	C	2016-06-01	728815257	2016-06-29	14
98	North America	Mexico	Personal Care	Offline	M	2015-07-30	559427106	2015-08-08	57
99	Sub-Saharan Africa	Mozambique	Household	Offline	L	2012-02-10	665095412	2012-02-15	53

100 rows × 17 columns

In [16]:

```
print(df_cleaned.columns)
```

```
Index(['Region', 'Country', 'Item Type', 'Sales Channel', 'Order Priority',
       'Order Date', 'Order ID', 'Ship Date', 'Units Sold', 'Unit Price',
       'Unit Cost', 'Total Revenue', 'Total Cost', 'Total Profit', 'Year',
       'Month', 'MonthYear'],
      dtype='object')
```

```
In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```