Analyzing Amazon Sales data:

- Importing the necessary Libraries to perform the EDA.

In [5]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [6]:
```
pd.set_option('display.max_rows', 1000)  # This will display all rows
pd.set_option('display.max_columns', 200)  # This will display all columns
```

- Importing the CSV file from the system.

In [7]:
```
Data_Frame= pd.read_csv('/Users/sanjay/Desktop/Working UMProject/Project -1/
```

In [8]:
```
Data_Frame.head()
```

Out[8]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	U
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2
2	Europe	Russia	Office Supplies	Offline	L	2/5/2014	341417157	8/5/2014	
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	5/7/2014	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	1/2/2013	115456712	6/2/2013	5

In [9]:
```
Data_Frame.shape
```

Out[9]:
```
(100, 14)
```

In [10]:
```
Data_Frame.ndim
```

Out[10]:
```
2
```

Performing the ETL for the dataset.

Seeing if there is any duplicates, missing valves in dataset and the adjusting the dtpyes of the columns to correct dtypes and adding columns.

In [11]:
```
Data_Frame.drop_duplicates(inplace=True)
```

-As we can see there is no duplicates in the dataset.

In [12]:
```
Data_Frame.head()
```

Out[12]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	U
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2
2	Europe	Russia	Office Supplies	Offline	L	2/5/2014	341417157	8/5/2014	
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	5/7/2014	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	1/2/2013	115456712	6/2/2013	5

In [13]:
```
Data_Frame.isna().sum()
```

Out[13]:
```
Region            0
Country           0
Item Type         0
Sales Channel     0
Order Priority    0
Order Date        0
Order ID          0
Ship Date         0
Units Sold        0
Unit Price        0
Unit Cost         0
Total Revenue     0
Total Cost        0
Total Profit      0
dtype: int64
```

- There is also no missing values in the dataset

In [14]:
```
Data_Frame.dtypes
```

Out[14]:
```
Region             object
Country            object
Item Type          object
Sales Channel      object
Order Priority     object
Order Date         object
Order ID            int64
Ship Date          object
Units Sold          int64
Unit Price        float64
Unit Cost         float64
Total Revenue     float64
Total Cost        float64
Total Profit      float64
dtype: object
```

- As we can see the dtypes of the column are not defined correct for Order Date & Ship Date, so adjusting to correct dtypes.

In [15]:
```
Data_Frame['Order Date'] = pd.to_datetime(Data_Frame['Order Date'])
```

In [16]:
```
Data_Frame['Ship Date']= pd.to_datetime(Data_Frame['Ship Date'])
```

In [17]:
```
Data_Frame.dtypes
```

Out[17]:
```
Region                    object
Country                   object
Item Type                 object
Sales Channel             object
Order Priority            object
Order Date        datetime64[ns]
Order ID                   int64
Ship Date         datetime64[ns]
Units Sold                 int64
Unit Price               float64
Unit Cost                float64
Total Revenue            float64
Total Cost               float64
Total Profit             float64
dtype: object
```

Engineering column to extract years & months from Order Date column for showing sales trends.

In [18]:
```
Data_Frame['Order Year']=Data_Frame['Order Date'].dt.year
Data_Frame['Order Month']=Data_Frame['Order Date'].dt.month_name()
```

In [19]:
```
Data_Frame['Year - Month']=Data_Frame['Order Year'].astype(str) + '-' + Data
```

In [20]:
```
Data_Frame.head()
```

Out[20]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	9925	2
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	2804	2
2	Europe	Russia	Office Supplies	Offline	L	2014-02-05	341417157	2014-08-05	1779	6
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-05-07	8102	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2013-01-02	115456712	2013-06-02	5062	6

To know the describe of the numerical data in the dataset.

In [21]:
```
Numerical_stat=Data_Frame.drop(['Order ID','Order Year','Order Month','Order
Numerical_stat.describe()
```

Out[21]:

	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total P
count	100.000000	100.000000	100.000000	1.000000e+02	1.000000e+02	1.000000
mean	5128.710000	276.761300	191.048000	1.373488e+06	9.318057e+05	4.416820
std	2794.484562	235.592241	188.208181	1.460029e+06	1.083938e+06	4.385379
min	124.000000	9.330000	6.920000	4.870260e+03	3.612240e+03	1.258020
25%	2836.250000	81.730000	35.840000	2.687212e+05	1.688680e+05	1.214436
50%	5382.500000	179.880000	107.275000	7.523144e+05	3.635664e+05	2.907680
75%	7369.000000	437.200000	263.330000	2.212045e+06	1.613870e+06	6.358288
max	9925.000000	668.270000	524.960000	5.997055e+06	4.509794e+06	1.719922

Done with ETL of the dataset, now performing the EDA of the dataset.

Analysing the Catagorical columns of the dataset to get some insights.

In [22]:
```
Data_Frame['Region'].unique()
```

Out[22]:
```
array(['Australia and Oceania', 'Central America and the Caribbean',
       'Europe', 'Sub-Saharan Africa', 'Asia',
       'Middle East and North Africa', 'North America'], dtype=object)
```

In [23]:
```
Data_Frame['Country'].unique()
```

Out[23]:
```
array(['Tuvalu', 'Grenada', 'Russia', 'Sao Tome and Principe', 'Rwanda',
       'Solomon Islands', 'Angola', 'Burkina Faso',
       'Republic of the Congo', 'Senegal', 'Kyrgyzstan', 'Cape Verde',
       'Bangladesh', 'Honduras', 'Mongolia', 'Bulgaria', 'Sri Lanka',
       'Cameroon', 'Turkmenistan', 'East Timor', 'Norway', 'Portugal',
       'New Zealand', 'Moldova ', 'France', 'Kiribati', 'Mali',
       'The Gambia', 'Switzerland', 'South Sudan', 'Australia', 'Myanmar',
       'Djibouti', 'Costa Rica', 'Syria', 'Brunei', 'Niger', 'Azerbaijan',
       'Slovakia', 'Comoros', 'Iceland', 'Macedonia', 'Mauritania',
       'Albania', 'Lesotho', 'Saudi Arabia', 'Sierra Leone',
       "Cote d'Ivoire", 'Fiji', 'Austria', 'United Kingdom', 'San Marino',
       'Libya', 'Haiti', 'Gabon', 'Belize', 'Lithuania', 'Madagascar',
       'Democratic Republic of the Congo', 'Pakistan', 'Mexico',
       'Federated States of Micronesia', 'Laos', 'Monaco', 'Samoa ',
       'Spain', 'Lebanon', 'Iran', 'Zambia', 'Kenya', 'Kuwait',
       'Slovenia', 'Romania', 'Nicaragua', 'Malaysia', 'Mozambique'],
      dtype=object)
```

In [24]:
```
Data_Frame['Item Type'].unique()
```

Out[24]:
```
array(['Baby Food', 'Cereal', 'Office Supplies', 'Fruits', 'Household',
       'Vegetables', 'Personal Care', 'Clothes', 'Cosmetics', 'Beverages',
       'Meat', 'Snacks'], dtype=object)
```

In [25]:
```
Data_Frame['Sales Channel'].unique()
```

Out[25]:
```
array(['Offline', 'Online'], dtype=object)
```

In [26]:
```
Data_Frame['Order Priority'].unique()
```

Out[26]:
```
array(['H', 'C', 'L', 'M'], dtype=object)
```

In [27]:
```
Data_Frame['Order Year'].unique()
```

Out[27]:
```
array([2010, 2012, 2014, 2013, 2015, 2011, 2017, 2016], dtype=int32)
```

In [28]:
```
Data_Frame['Order Month'].unique()
```

Out[28]:
```
array(['May', 'August', 'February', 'June', 'January', 'April', 'July',
       'November', 'December', 'October', 'September', 'March'],
      dtype=object)
```

Finding the Sales-trend -> month-wise, year-wise, yearly_month-wise by ploting an charts for getting the trends from the dataset

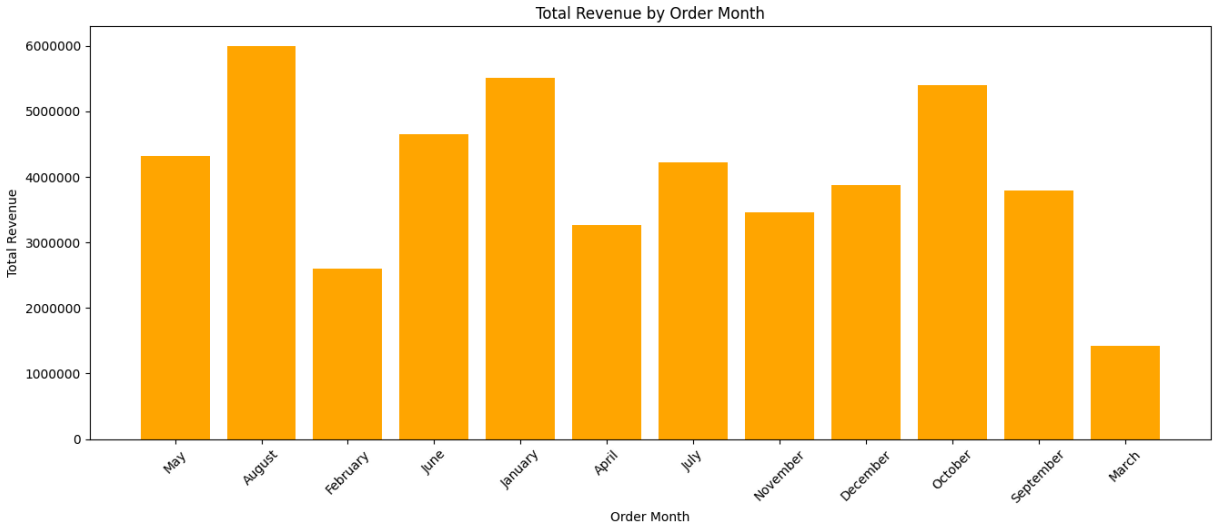
- Month wise Sales or Total Revenue Bar graph

In [29]:
```
Data_Frame.head()
```

Out[29]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	9925	2
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	2804	2
2	Europe	Russia	Office Supplies	Offline	L	2014-02-05	341417157	2014-08-05	1779	6
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-05-07	8102	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2013-01-02	115456712	2013-06-02	5062	6

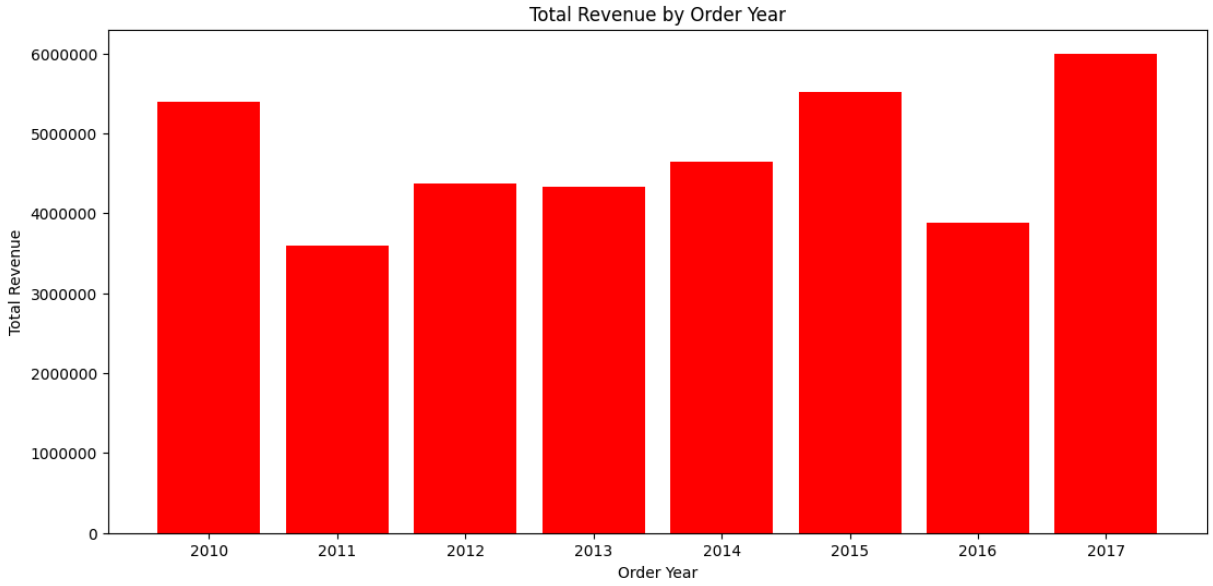
In [30]:
```
plt.figure(figsize=(13,6))

plt.bar(Data_Frame['Order Month'], Data_Frame['Total Revenue'],color='orange

plt.xlabel('Order Month')  # Label for the X-axis
plt.ylabel('Total Revenue')  # Label for the Y-axis
plt.title('Total Revenue by Order Month')  # Title of the plot
plt.xticks(rotation=45) # Rotate the x-axis labels for better readability
plt.tight_layout()
plt.gca().ticklabel_format(style='plain', axis='y') # To remove the scientif
plt.show()
```



> INSIGHT : We can see the month wise total revenue is highest in the month of AUGUST

- Year Wise Sales or Total Revenue Bar graph

In [31]:
```
plt.figure(figsize=(13,6))
plt.bar(Data_Frame['Order Year'],Data_Frame['Total Revenue'],color='red')
plt.xlabel('Order Year')
plt.ylabel('Total Revenue')
plt.title('Total Revenue by Order Year')
plt.gca().ticklabel_format(style='plain', axis='y')
```



> INSIGHT : The highest revenue is generated in the year 2017

- Yearly Month Wise Total Revenue Bar Graph

In [32]:
```
Year_Month_Column_df=Data_Frame[['Year - Month','Total Revenue']] # created
```

In [33]:
```
Year_Month_Column_df.head()
```

Out[33]:

	Year - Month	Total Revenue
0	2010-May	2533654.00
1	2012-August	576782.80
2	2014-February	1158502.59
3	2014-June	75591.66
4	2013-January	3296425.02

In [34]:
```
Group_Year_Months=Year_Month_Column_df.groupby(by='Year - Month') #Groupby t
```

In [35]:
```
Year_Month_agg=Group_Year_Months['Total Revenue'].sum() #used the aggeratior
```

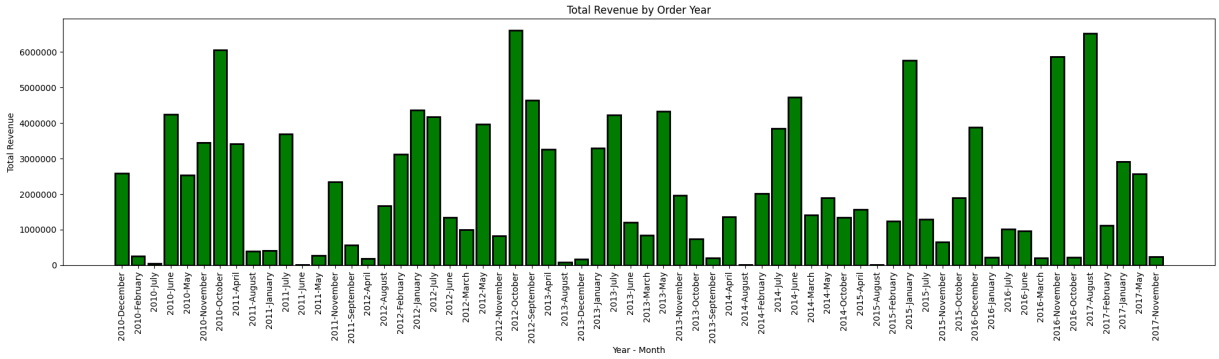
In [36]:
```
Year_Month_df=Year_Month_agg.to_frame().reset_index() # Formed an Data Frame
```

In [37]:
```
Year_Month_df.head() # Data Frame
```

Out[37]:

	Year - Month	Total Revenue
0	2010-December	2581786.39
1	2010-February	247956.32
2	2010-July	54319.26
3	2010-June	4245123.20
4	2010-May	2533654.00

In [38]:
```
# The Graph that shows the unique Year-Month total revenue

plt.figure(figsize=(20,6))

plt.bar(Year_Month_df['Year - Month'],Year_Month_df['Total Revenue'],color='
plt.xlabel('Year - Month')
plt.ylabel('Total Revenue')
plt.title('Total Revenue by Order Year')
plt.xticks(rotation=90)
plt.gca().ticklabel_format(style='plain', axis='y')
plt.tight_layout()
```



> INSIGHT : The higest revenue generated in yearly month wise is 2017-August

Total Revenue Performance by Region:

In [39]:
```
Data_Frame.head()
```

Out[39]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	9925	2
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	2804	2
2	Europe	Russia	Office Supplies	Offline	L	2014-02-05	341417157	2014-08-05	1779	(
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-05-07	8102	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2013-01-02	115456712	2013-06-02	5062	(

In [40]:
```
Group_Region=Data_Frame.groupby('Region')
```

In [41]:
```
Region_agg=Group_Region['Total Revenue'].sum()
```

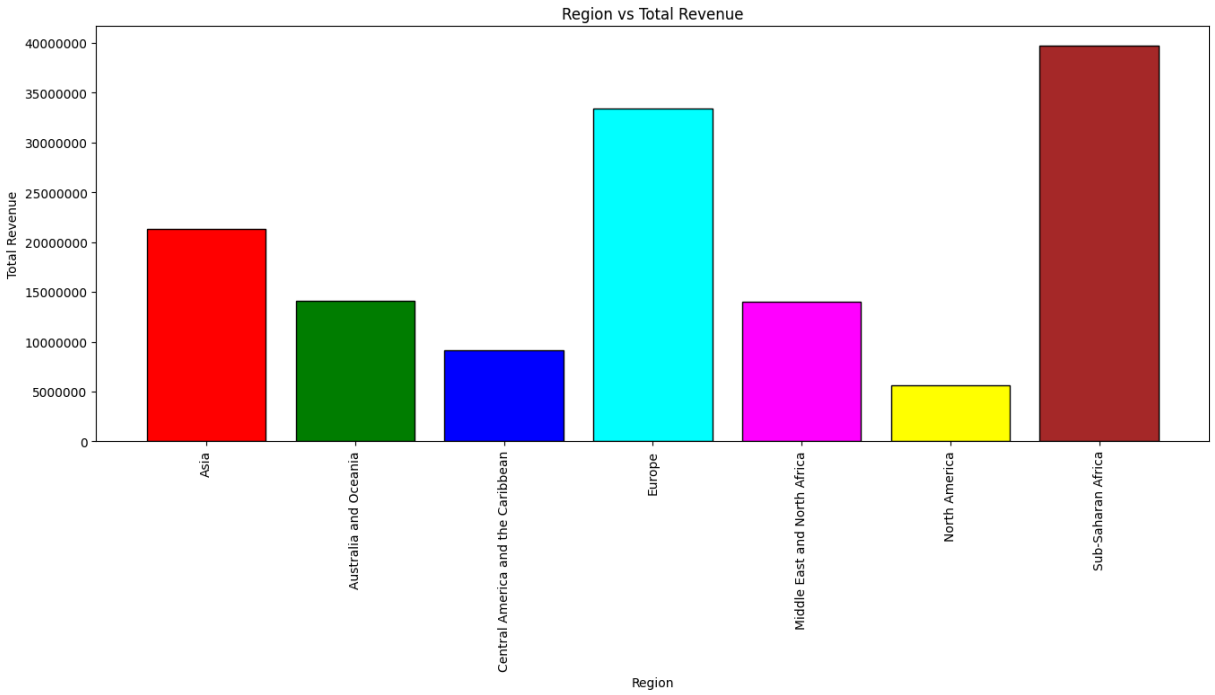
In [42]:
```
Region_df=Region_agg.to_frame().reset_index()
```

In [43]:
```
Region_df
```

Out[43]:

	Region	Total Revenue
0	Asia	21347091.02
1	Australia and Oceania	14094265.13
2	Central America and the Caribbean	9170385.49
3	Europe	33368932.11
4	Middle East and North Africa	14052706.58
5	North America	5643356.55
6	Sub-Saharan Africa	39672031.43

In [44]:
```
plt.figure(figsize=(16,6))
colors = ['red', 'green', 'blue', 'cyan', 'magenta', 'yellow', 'brown'] #lis

plt.bar(Region_df['Region'],Region_df['Total Revenue'],color=colors,edgecolo
plt.xlabel('Region')
plt.ylabel('Total Revenue')
plt.title('Region vs Total Revenue')
plt.gca().ticklabel_format(style='plain', axis='y')
plt.xticks(rotation=90)
plt.show()
```



> INSIGHT : Revenue generated form an Region is higest at SUB-SAHARAN AFRICA

Filtering Total Revenue data based on Online & Offline for a Region

In [45]:
```
Online = Data_Frame[Data_Frame['Sales Channel'] == 'Online'] #filtering the
```

In [46]:
```
Online_agg=Online.groupby('Region')['Total Revenue'].sum()
Online_df=Online_agg.to_frame().reset_index()
Online_df
```

Out[46]:

	Region	Total Revenue
0	Asia	9200993.26
1	Australia and Oceania	9892397.28
2	Central America and the Caribbean	916273.30
3	Europe	15246445.66
4	Middle East and North Africa	9059567.70
5	Sub-Saharan Africa	13938281.91

In [47]:
```
Offline=Data_Frame[Data_Frame['Sales Channel']=='Offline']
```

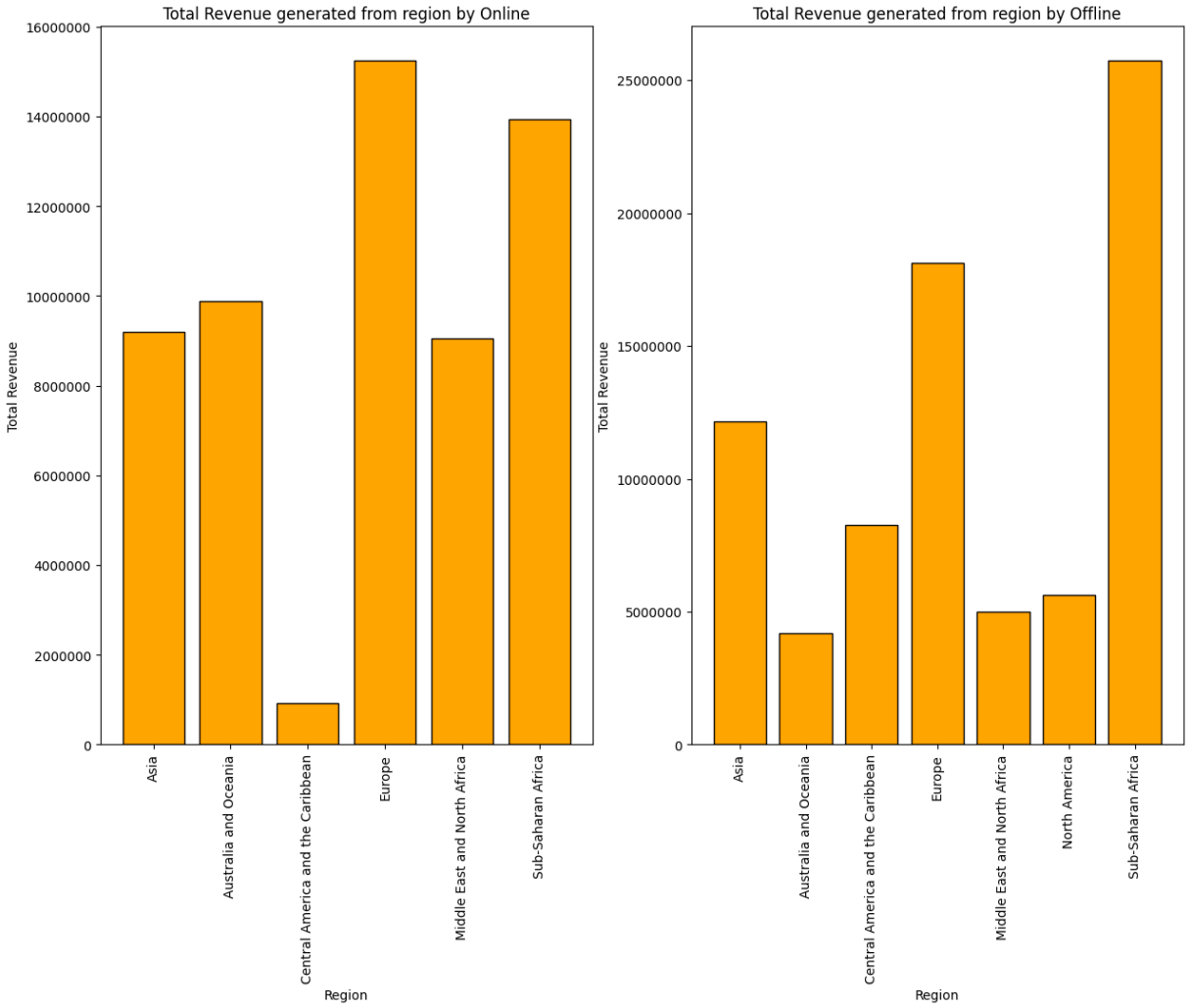
In [48]:
```
Offline_agg=Offline.groupby('Region')['Total Revenue'].sum()
Offline_df=Offline_agg.to_frame().reset_index()
Offline_df
```

Out[48]:

	Region	Total Revenue
0	Asia	12146097.76
1	Australia and Oceania	4201867.85
2	Central America and the Caribbean	8254112.19
3	Europe	18122486.45
4	Middle East and North Africa	4993138.88
5	North America	5643356.55
6	Sub-Saharan Africa	25733749.52

In [49]:
```
#Filtered Online Sales
plt.figure(figsize=(15,10)) #canvas size

plt.subplot(1,2,1) #fitting the plots in the row wise

plt.bar(Online_df['Region'],Online_df['Total Revenue'],color='orange',edgecc
plt.xlabel('Region')
plt.ylabel('Total Revenue')
plt.title('Total Revenue generated from region by Online')
plt.gca().ticklabel_format(style='plain', axis='y')
plt.xticks(rotation=90)


plt.subplot(1,2,2)
```

```
plt.bar(Offline_df['Region'],Offline_df['Total Revenue'],color='orange',edge
plt.xlabel('Region')
plt.ylabel('Total Revenue')
plt.title('Total Revenue generated from region by Offline')
plt.gca().ticklabel_format(style='plain', axis='y')
plt.xticks(rotation=90)


plt.show()
```



> INSIGHT :

As the comparing the plot of Online & Offline, the Revenue generated in Region by Online is highest at EUPORE Region & the Revenue generated in Region by Offline is highest at SUB-SAHARAN AFRICA

Top 3 Most Unit Sold Countries in each Region

In [50]:
```
Data_Frame.head()
```

Out[50]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	9925	2
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	2804	2
2	Europe	Russia	Office Supplies	Offline	L	2014-02-05	341417157	2014-08-05	1779	6
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-05-07	8102	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2013-01-02	115456712	2013-06-02	5062	6

In [51]:
```
Region_Country_df=Data_Frame.groupby(['Region','Country'])['Units Sold'].sum
```

In [52]:
```
Region_Country_df.head()
```

Out[52]:

	Region	Country	Units Sold
0	Asia	Bangladesh	8263
1	Asia	Brunei	6708
2	Asia	Kyrgyzstan	124
3	Asia	Laos	3732
4	Asia	Malaysia	6267

In [53]:
```
Sorted_R_C_df=Region_Country_df.sort_values(['Region','Units Sold'],ascendir
```

In [54]:
```
Sorted_R_C_df.head()
```

Out[54]:

	Region	Country	Units Sold
6	Asia	Myanmar	14180
8	Asia	Turkmenistan	8840
0	Asia	Bangladesh	8263
7	Asia	Sri Lanka	6952
1	Asia	Brunei	6708

In [55]:
```
Top_country_by_region= Sorted_R_C_df.groupby('Region').head(3)
```

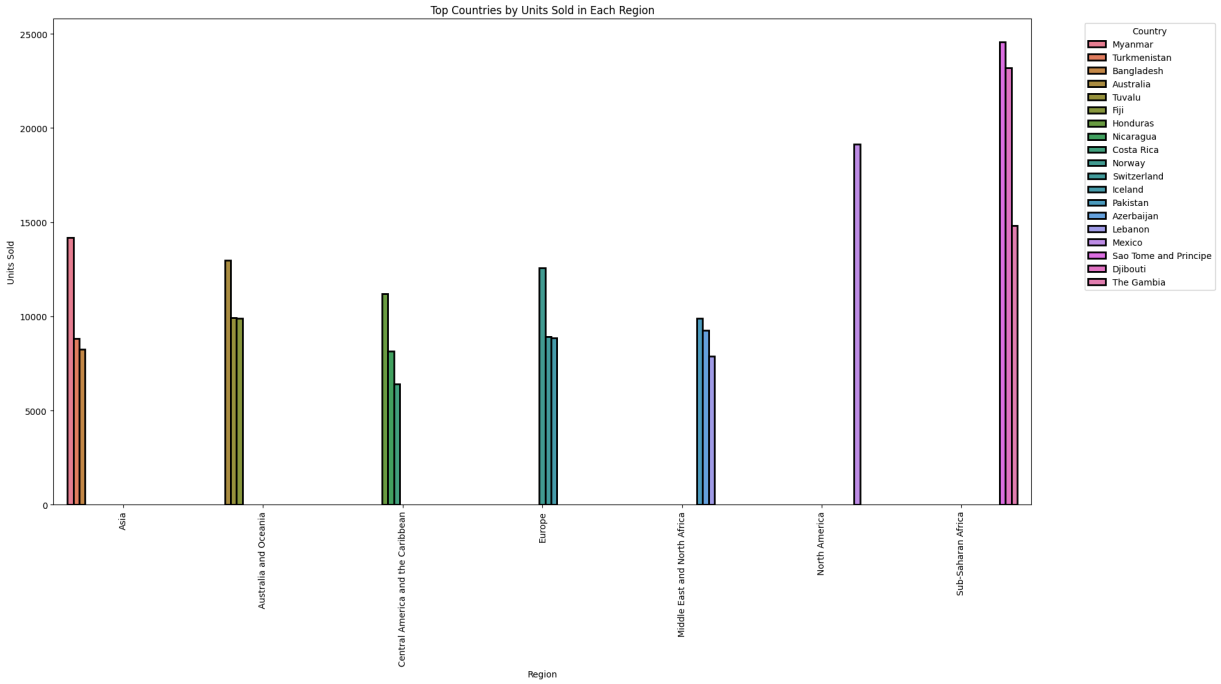
In [58]:
```
Top_country_by_region
```

Out[58]:

	Region	Country	Units Sold
6	Asia	Myanmar	14180
8	Asia	Turkmenistan	8840
0	Asia	Bangladesh	8263
9	Australia and Oceania	Australia	12995
17	Australia and Oceania	Tuvalu	9925
12	Australia and Oceania	Fiji	9905
22	Central America and the Caribbean	Honduras	11199
23	Central America and the Caribbean	Nicaragua	8156
19	Central America and the Caribbean	Costa Rica	6409
33	Europe	Norway	12574
41	Europe	Switzerland	8934
28	Europe	Iceland	8867
48	Middle East and North Africa	Pakistan	9892
43	Middle East and North Africa	Azerbaijan	9255
46	Middle East and North Africa	Lebanon	7884
51	North America	Mexico	19143
70	Sub-Saharan Africa	Sao Tome and Principe	24568
59	Sub-Saharan Africa	Djibouti	23198
74	Sub-Saharan Africa	The Gambia	14813

In [61]:
```
plt.figure(figsize=(20,10))

sns.barplot(
    data=Top_country_by_region,
    x='Region',
    y='Units Sold',
    hue='Country',edgecolor='black',
    linewidth=2)

# Improve the readability
plt.xticks(rotation=90)
plt.xlabel('Region')
plt.ylabel('Units Sold')
plt.title('Top Countries by Units Sold in Each Region')

# Adjust the legend
```

```
plt.legend(title='Country', bbox_to_anchor=(1.05, 1), loc='upper left')  # M
plt.show()
```



INSIGHT :

As the top 3 with most Unit Sold countries in each region is as followed

- Asia : Myanmar, Turkmenistan, Bangladesh
- Australia and Oceania : Australia, Tuvalu, Fiji
- Central America and the Caribbean : Honduras, Nicaragua, Costa Rica
- Europe : Norway, Switzerland, Iceland
- Middle East and North Africa : Pakistan, Azerbaijan, Lebanon
- North America : Mexico
- Sub-Saharan Africa : Sao Tome and Principe, Djibouti, The Gambia