Amazon Project

By Faiz Khan

First we have to import basic Libraries

In [9]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
```

Now we have to load DataSet

```
In [10]: df=pd.read_csv("Amazon Sales data.csv")
```

```
In [11]: df
```

Out[11]:

	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
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654.00	1582243.50	951410.50
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.70	117.11	576782.80	328376.44	248406.36
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.21	524.96	1158502.59	933903.84	224598.75
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33	6.92	75591.66	56065.84	19525.82
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.21	524.96	3296425.02	2657347.52	639077.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	Sub-													

Applying Python's Basic Funtions

```
In [20]: df.isna().sum()
```

```
Out[20]: 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
```

```
In [18]: df.Country.value_counts()
```

```
Out[18]: The Gambia                4
```

```
In [15]: df.describe()
```

Out[15]:

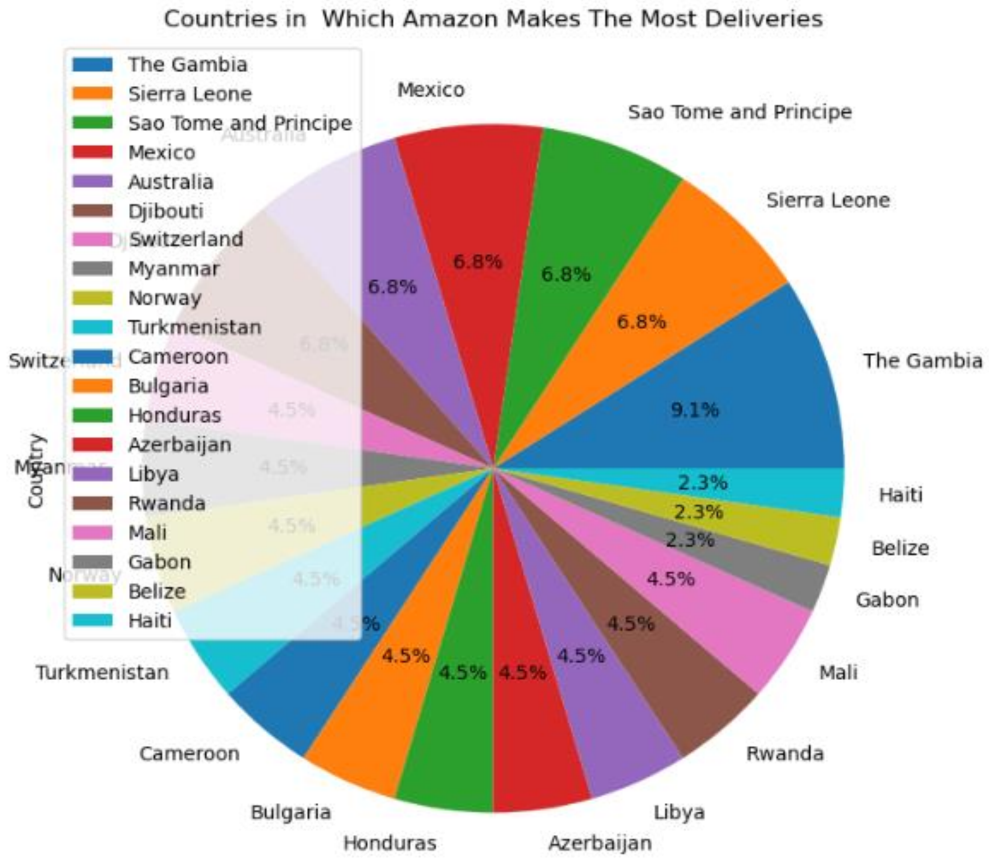
	Order ID	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total P
count	1.000000e+02	100.000000	100.000000	100.000000	1.000000e+02	1.000000e+02	1.000000e
mean	5.550204e+08	5128.710000	276.761300	191.048000	1.373488e+06	9.318057e+05	4.416820e
std	2.606153e+08	2794.484562	235.592241	188.208181	1.460029e+06	1.083938e+06	4.385379e
min	1.146066e+08	124.000000	9.330000	6.920000	4.870260e+03	3.612240e+03	1.258020e
25%	3.389225e+08	2836.250000	81.730000	35.840000	2.687212e+05	1.688680e+05	1.214436e
50%	5.577086e+08	5382.500000	179.880000	107.275000	7.523144e+05	3.635664e+05	2.907680e
75%	7.907551e+08	7369.000000	437.200000	263.330000	2.212045e+06	1.613870e+06	6.358288e
max	9.940222e+08	9925.000000	668.270000	524.960000	5.997055e+06	4.509794e+06	1.719922e

```
In [17]: df.mean().round(2)
```

```
Out[17]: Order ID          5.550204e+08
         Units Sold        5.128710e+03
         Unit Price        2.767600e+02
         Unit Cost         1.910500e+02
         Total Revenue     1.373488e+06
         Total Cost        9.318057e+05
         Total Profit      4.416820e+05
         dtype: float64
```

Visualisation With Python

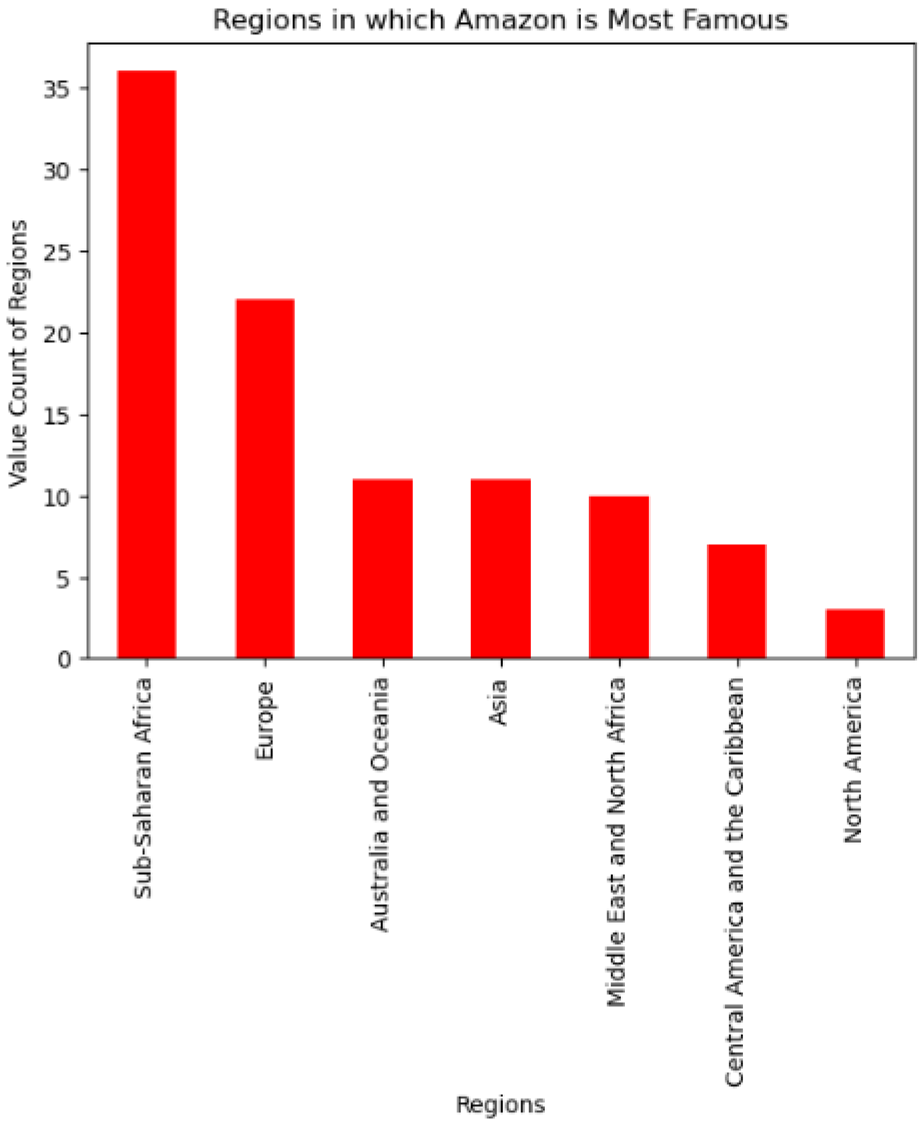
```
In [31]: plt.figure(figsize=(12,8))
         df.Country.value_counts().head(20).plot(kind="pie",autopct="%1.1f%%")
         plt.legend()
         plt.title("Countries in  Which Amazon Makes The Most Deliveries")

Out[31]: Text(0.5, 1.0, 'Countries in  Which Amazon Makes The Most Deliveries')
```



We can clearly see that in The Gambia amazon deliver most products so they shoould target more to the people of The Gambia and then Djibouti,Australia,Mexico,Sao Tome and Principe, Sierra Leone

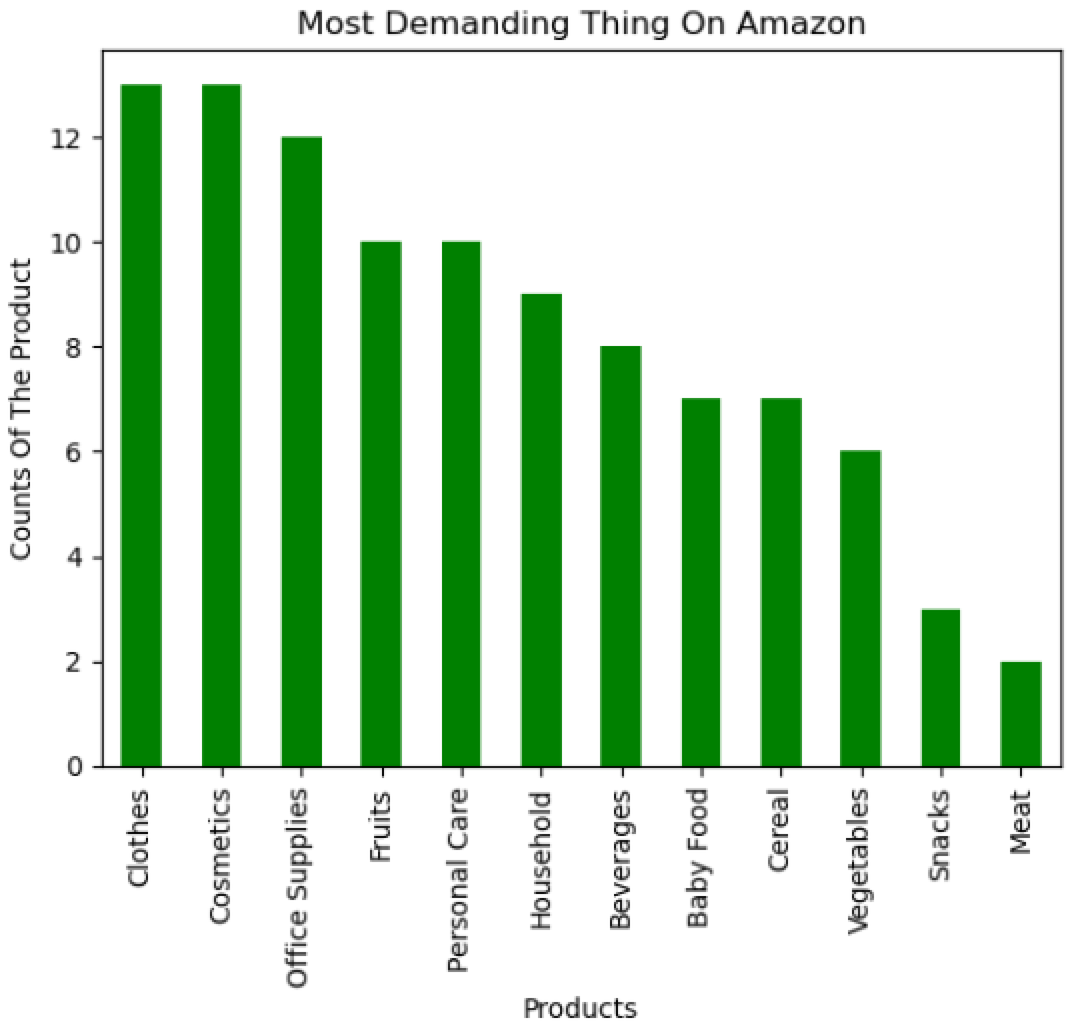
```
In [32]: df.Region.value_counts().head(25).plot(kind="bar",color="Red")
plt.title("Regions in which Amazon is Most Famous")
plt.xlabel("Regions")
plt.ylabel("Value Count of Regions")
```

```
Out[32]: Text(0, 0.5, 'Value Count of Regions')
```



Sub-Saharan Africa is the region where Amazon delivers the most Products.

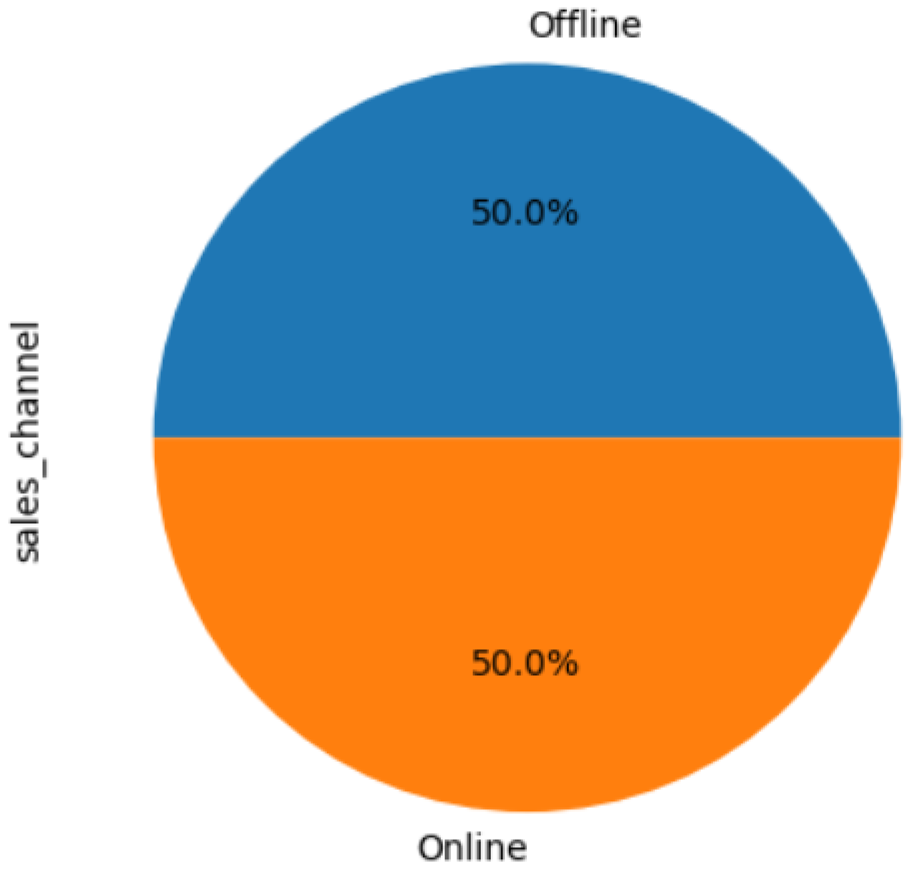
```
In [47]: df.item_type.value_counts().head(15).plot(kind="bar",color="green")
         plt.title("Most Demanding Thing On Amazon")
         plt.ylabel("Counts Of The Product")
         plt.xlabel("Products")

Out[47]: Text(0.5, 0, 'Products')
```



Clothes are most demanding product on Amazon that people prefer to buy

```
In [51]: df.sales_channel.value_counts().plot(kind="pie",autopct="%1.1f%%")
```

```
Out[51]: <Axes: ylabel='sales_channel'>
```



Amazon uses both Online Offline mode equally as their Sales Channel

In [60]:
```
df["total_sales"]=df.total_revenue-df.total_cost
```

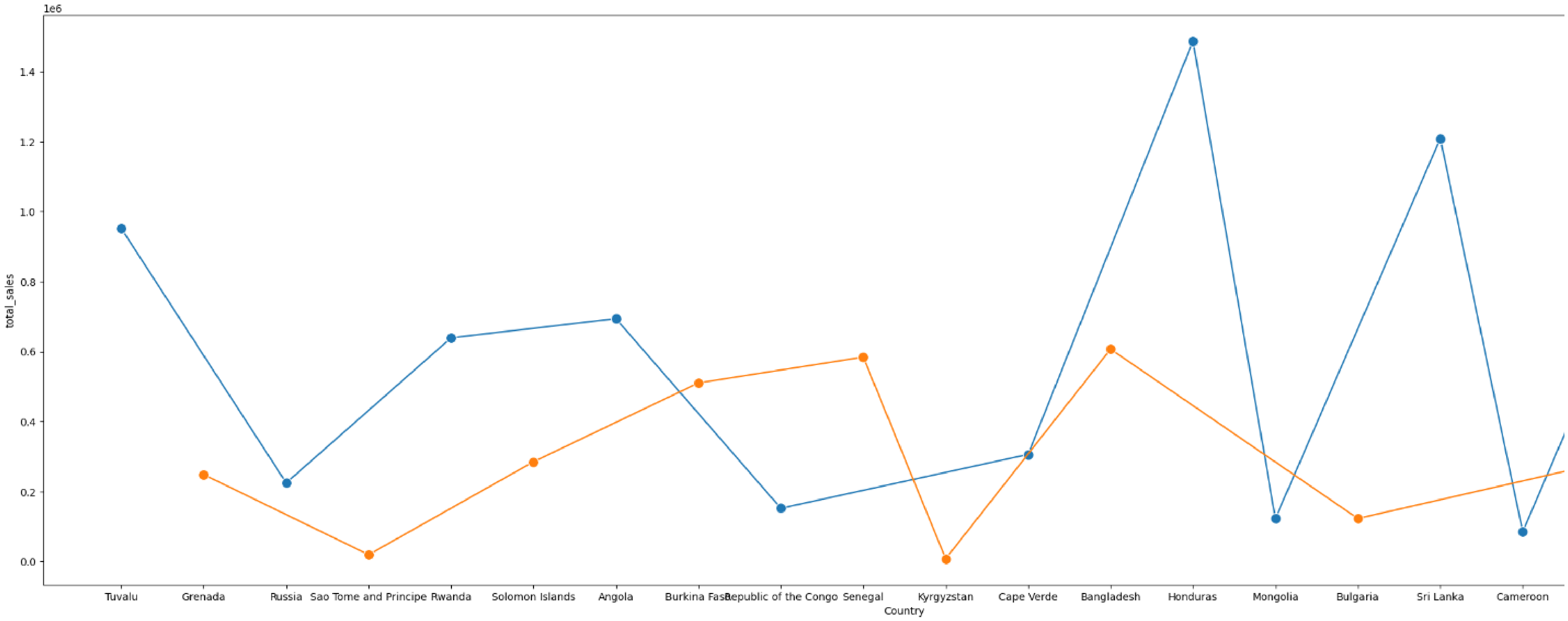
In [61]:
```
df
```

Out[61]:

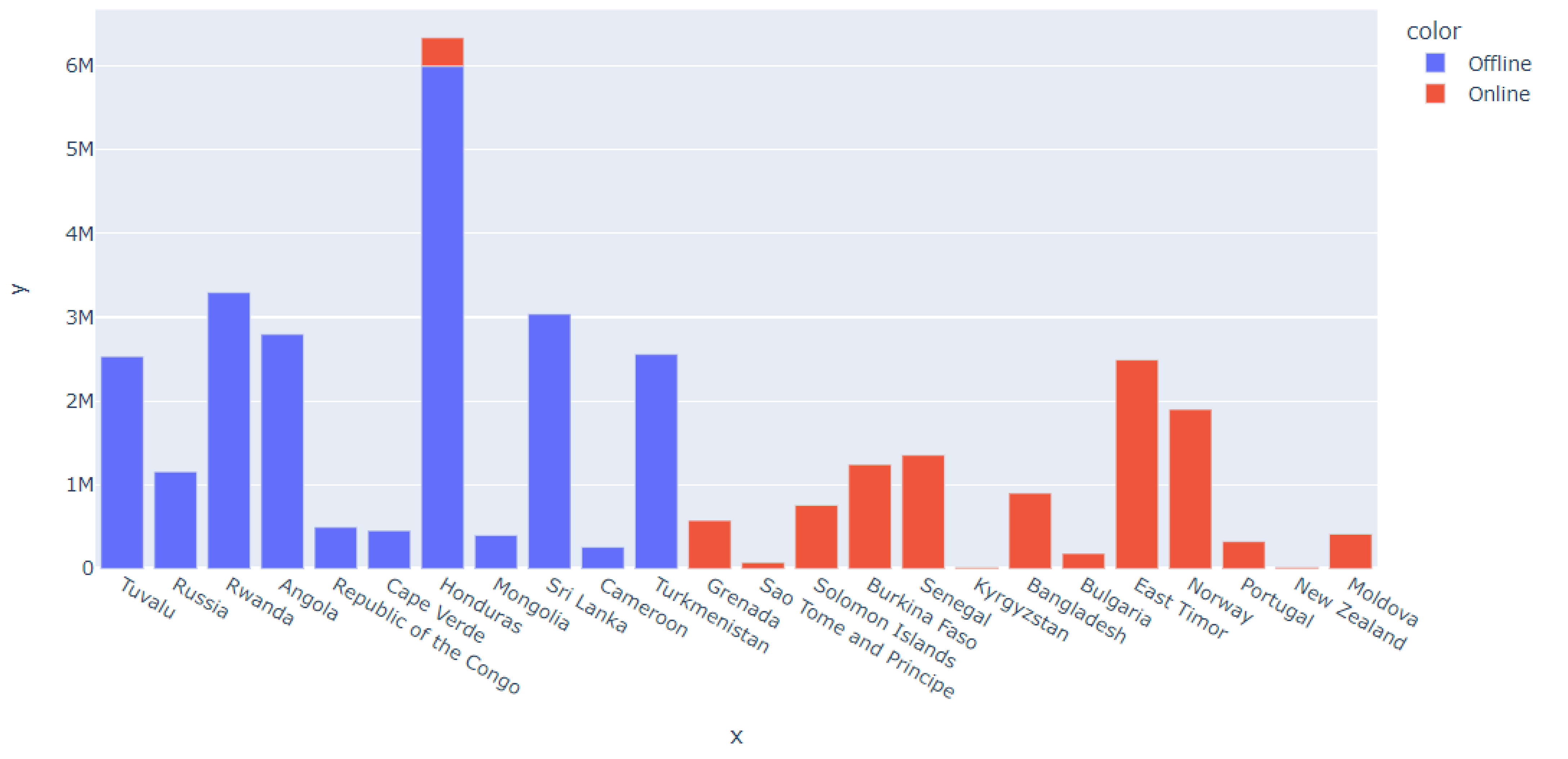
	Region	Country	item_type	sales_channel	order_priority	order_date	order_id	ship_date	units_sold	unit_price	unit_cost	total_revenue	total_
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654.00	158224
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.70	117.11	576782.80	32837
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.21	524.96	1158502.59	93390
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33	6.92	75591.66	5606
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.21	524.96	3296425.02	265734
...	...	...	...	...	...	...	...	...	...	...	...	...	
95	Sub-Saharan Africa	Mali	Clothes	Online	M	7/26/2011	512878119	9/3/2011	888	109.28	35.84	97040.64	3182
96	Asia	Malaysia	Fruits	Offline	L	11/11/2011	810711038	12/28/2011	6267	9.33	6.92	58471.11	4336
97	Sub-Saharan Africa	Sierra Leone	Vegetables	Offline	C	6/1/2016	728815257	6/29/2016	1485	154.06	90.93	228779.10	13503

```
1]: plt.figure(figsize=(30,10))
    sns.lineplot(x=df.Country.head(20),y=df.total_sales.head(20),hue=df.sales_channel.head(20),marker="o",markersize=
```

```
1]: <Axes: xlabel='Country', ylabel='total_sales'>
```



In [23]:
```python
px.bar(x=df.Country.head(25),y=df.total_revenue.head(25),color=df.sales_channel.head(25))
```



```
In [92]: c=df.groupby("Country")["total_revenue"].unique().reset_index()
```

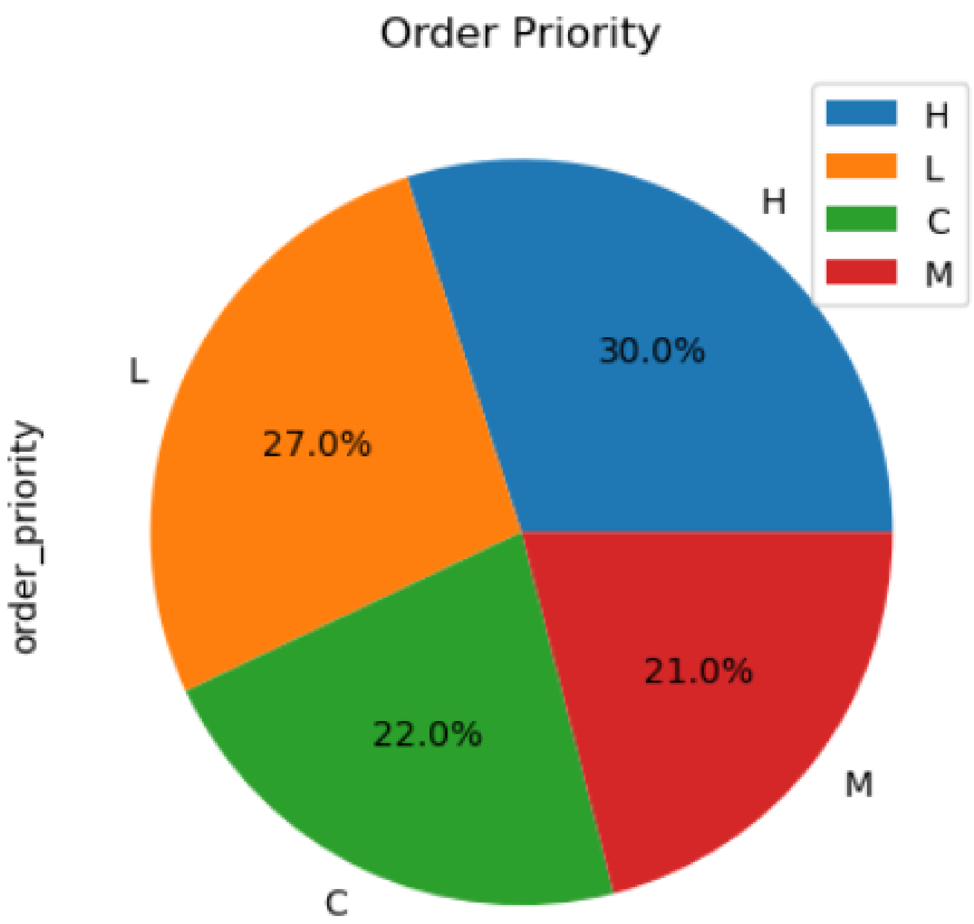
```
In [95]: c
```

Out[95]:

	Country	total_revenue
0	Albania	[247956.32]
1	Angola	[2798046.49]
2	Australia	[1904138.04, 140287.4, 445508.05]
3	Austria	[1244708.4]
4	Azerbaijan	[3162704.8, 1316095.41]
...	...	...
71	The Gambia	[1583799.9, 2011149.63, 435466.9, 1419101.52]
72	Turkmenistan	[2559474.1, 3262562.1]
73	Tuvalu	[2533654.0]

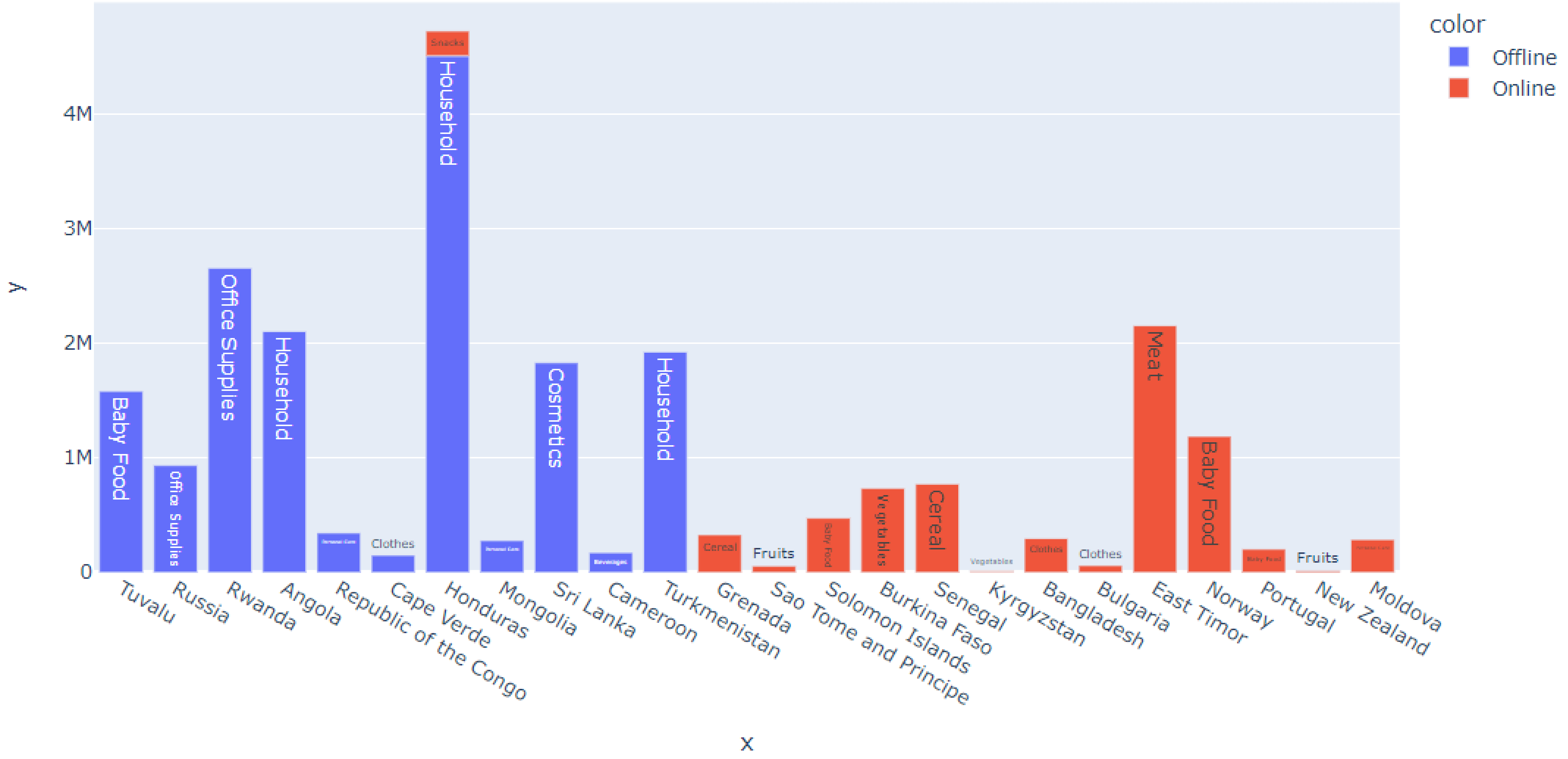
In [113]:
```
df.order_priority.value_counts().plot(kind="pie",autopct="%1.1f%%")
plt.legend()
plt.title("Order Priority")
```

Out[113]: Text(0.5, 1.0, 'Order Priority')

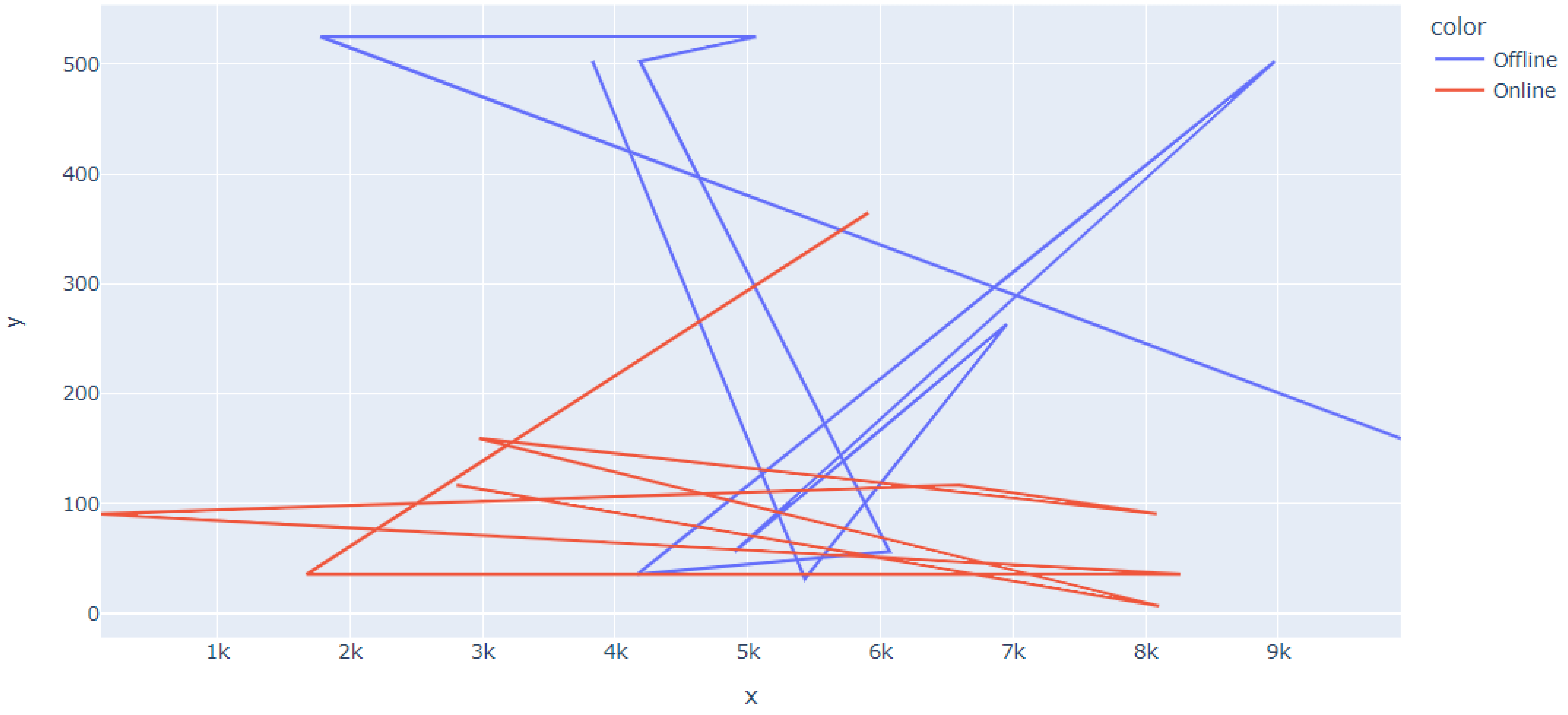


H is the highest order Priopity of People in diffrent Countries

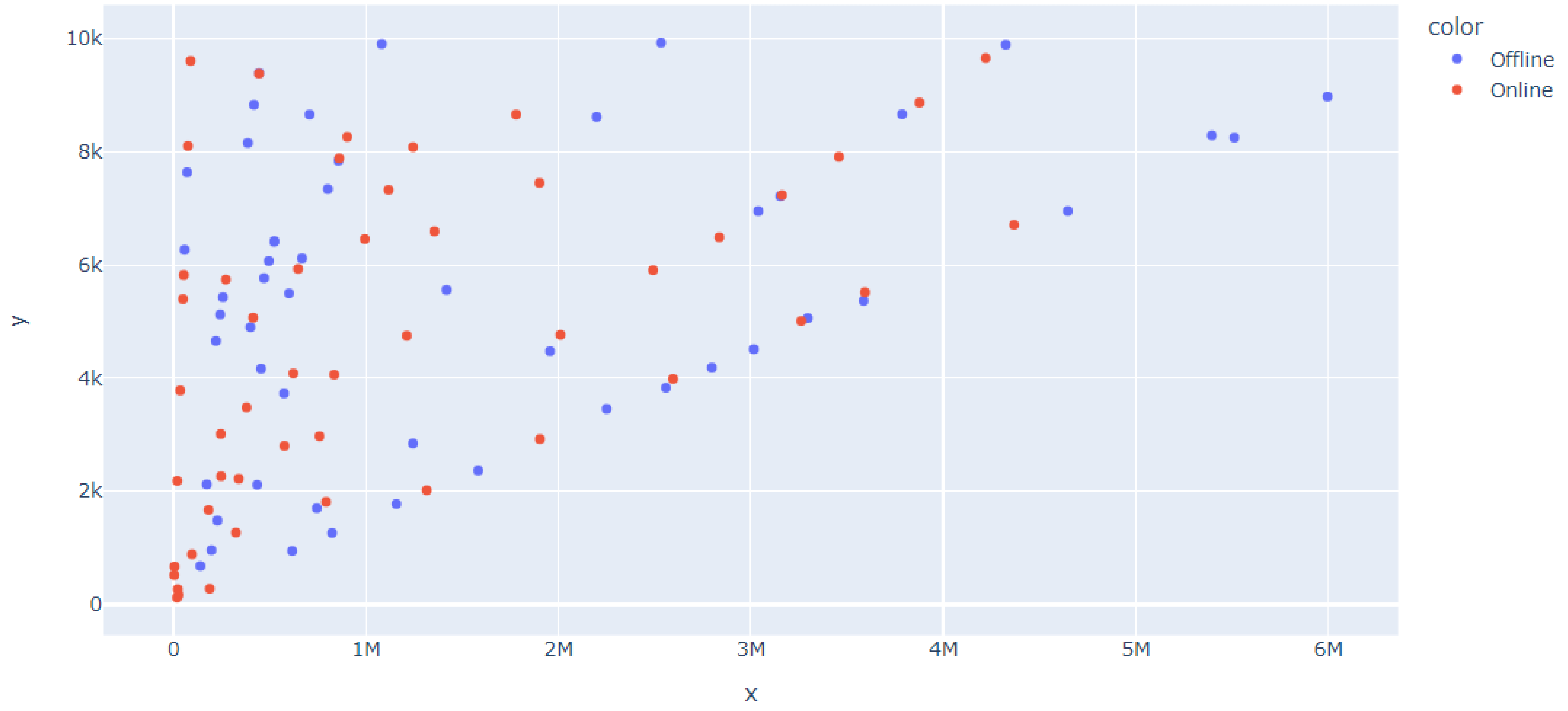
In [24]:
```
px.bar(x=df.Country.head(25),y=df.total_cost.head(25),color=df.sales_channel.head(25),text=df.item_type.head(25))
```



In [25]:
```
px.line(x=df.units_sold.head(20),y=df.unit_cost.head(20),color=df.sales_channel.head(20))
```

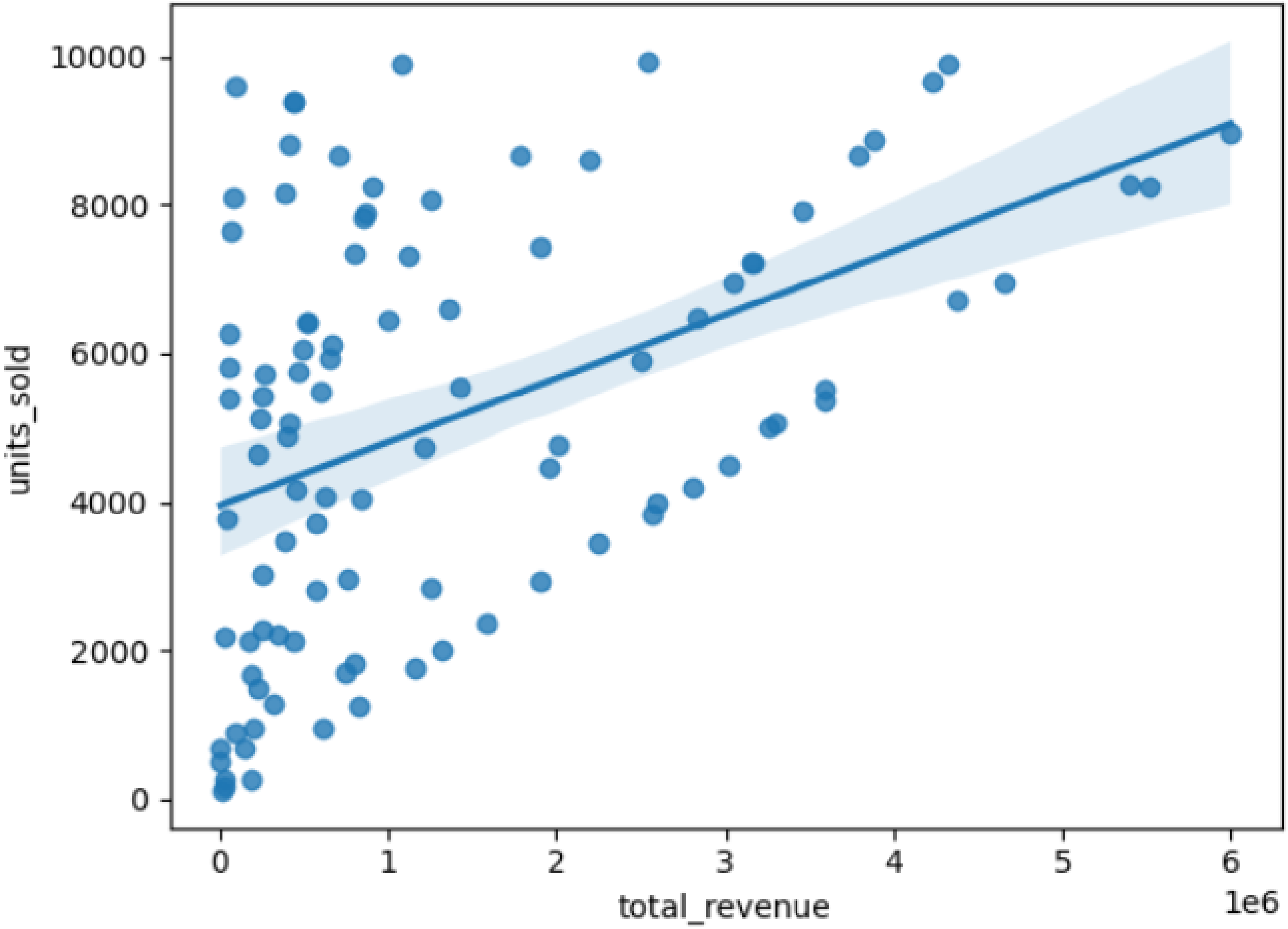


```
In [26]: px.scatter(x=df.total_revenue,y=df.units_sold,color=df.sales_channel)
```

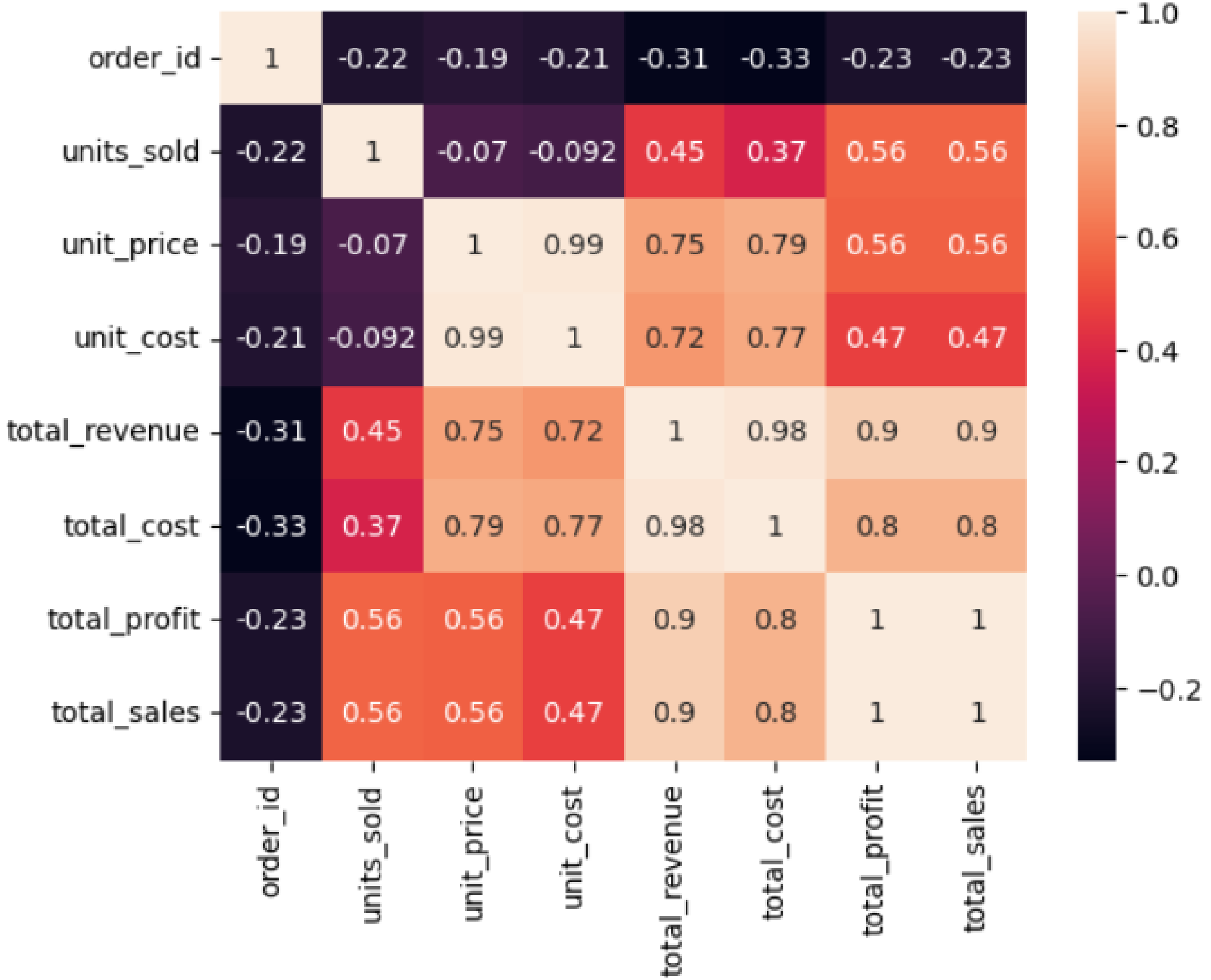


In [140]:
```
sns.regplot(x=df.total_revenue,y=df.units_sold)
```

Out[140]:
```
<Axes: xlabel='total_revenue', ylabel='units_sold'>
```



In [143]:
```
sns.heatmap(df.corr(numeric_only=True),annot=True)
```
Out[143]: <Axes: >



Applying ML Algorithm

In [241]:
```
from sklearn.model_selection import train_test_split
```

In [259]:
```
x=df.drop(["Region","Country","item_type","sales_channel","order_priority","order_id","order_date","ship_date","units_sold","unit
```

In [260]:
```
y=df.total_sales
```

In [262]:
```
x.astype("int")
```

Out[262]:

	total_revenue	total_cost	total_profit	total_sales
0	2533654	1582243	951410	951410
1	576782	328376	248406	248406
2	1158502	933903	224598	224598
3	75591	56065	19525	19525
4	3296425	2657347	639077	639077
...	...	...	...	...
95	97040	31825	65214	65214
96	58471	43367	15103	15103
97	228779	135031	93748	93748
98	471336	326815	144521	144521
99	3586605	2697132	889472	889472

100 rows × 4 columns

In [263]:
```
y.astype("int")
```

In [262]:
```
x.astype("int")
```

Out[262]:

	total_revenue	total_cost	total_profit	total_sales
0	2533654	1582243	951410	951410
1	576782	328376	248406	248406
2	1158502	933903	224598	224598
3	75591	56065	19525	19525
4	3296425	2657347	639077	639077
...	...	...	...	...
95	97040	31825	65214	65214
96	58471	43367	15103	15103
97	228779	135031	93748	93748
98	471336	326815	144521	144521
99	3586605	2697132	889472	889472

100 rows × 4 columns

In [263]:
```
y.astype("int")
```

Out[263]:
```
0     951410
1     248406
2     224598
3      19525
4     639077
       ...
95     65214
96     15103
97     93748
98    144521
99    889472
Name: total_sales, Length: 100, dtype: int32
```

In [264]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=44)
```

In [265]:
```
x_train.shape,x_test.shape
```

Out[265]:
```
((75, 4), (25, 4))
```

In [266]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [267]:
```
norm=MinMaxScaler()
```

In [268]:
```
norm
```

Out[268]:
```
MinMaxScaler()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [269]:
```
x_train=pd.DataFrame(norm.fit_transform(x_train))
```

In [270]:
```
x_test=pd.DataFrame(norm.fit_transform(x_test))
```

In [249]:
```
x_train[["Region","Country","item_type","sales_channel","order_priority"]]=oe.fit_transform(x_train[["Region","Country","item_typ
```

In [250]:
```
x_test[["Region","Country","item_type","sales_channel","order_priority"]]=oe.fit_transform(x_test[["Region","Country","item_type"
```

In [271]:
```
from sklearn.linear_model import LinearRegression
```

In [272]:
```
lr=LinearRegression()
```

In [273]:
```
lr
```

Out[273]:
```
LinearRegression()
```

In [273]:
```
lr
```

Out[273]:
```
LinearRegression()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

click to expand output; double click to hide output

In [274]:
```
lr.fit(x_train,y_train)
```

Out[274]:
```
LinearRegression()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [275]:
```
y_pred=lr.predict(x_test)
```

In [276]:
```
from sklearn.metrics import r2_score
```

In [277]:
```
r2_score(y_pred,y_test)
```

Out[277]:
```
0.9984731605620133
```

Tableau'Dashboard of Amazon DataSet

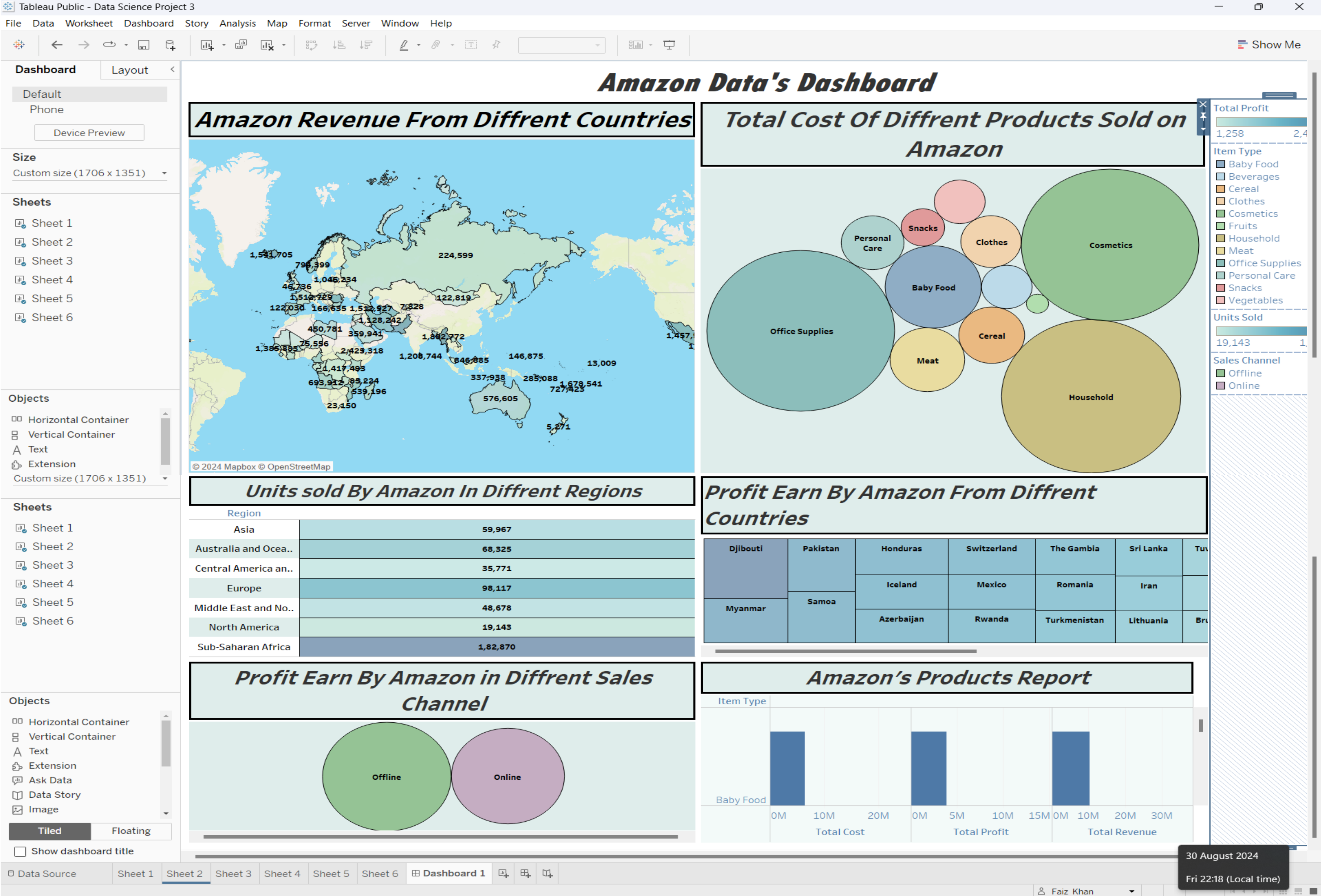


Tableau Public - Data Science Project 3
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Show Me
Dashboard Layout
Default
Phone
Device Preview
Size
Custom size (1706 x 1351)
Sheets
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
Sheet 6
Objects
Horizontal Container
Vertical Container
Text
Extension
Ask Data
Data Story
Image
Tiled
Floating
Show dashboard title
Amazon Data's Dashboard
Amazon Revenue From Diffrent Countries
© 2024 Mapbox © OpenStreetMap
Total Cost Of Diffrent Products Sold on Amazon
Office Supplies
Personal Care
Snacks
Clothes
Cosmetics
Baby Food
Cereal
Meat
Household
Total Profit
1,258
Item Type
Baby Food
Beverages
Cereal
Clothes
Cosmetics
Fruits
Household
Meat
Office Supplies
Personal Care
Snacks
Vegetables
Units Sold
19,143
Sales Channel
Offline
Online
Units sold By Amazon In Diffrent Regions
Region
Asia 59,967
Australia and Ocea.. 68,325
Central America an.. 35,771
Europe 98,117
Middle East and No.. 48,678
North America 19,143
Sub-Saharan Africa 1,82,870
Profit Earn By Amazon From Diffrent Countries
Djibouti
Myanmar
Pakistan
Samoa
Honduras
Iceland
Azerbaijan
Switzerland
Mexico
Rwanda
The Gambia
Romania
Turkmenistan
Sri Lanka
Iran
Lithuania
Profit Earn By Amazon in Diffrent Sales Channel
Offline
Online
Amazon's Products Report
Item Type
Baby Food
0M 10M 20M
Total Cost
0M 5M 10M 15M
Total Profit
0M 10M 20M 30M
Total Revenue
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
Sheet 6
Dashboard 1
30 August 2024
Fri 22:18 (Local time)
Faiz Khan

Important Points

- We can clearly see that in The Gambia amazon deliver most products so they shoould target more to the people of The Gambia and then Djibouti,Australia,Mexico,Sao Tome and Principe, Sierra Leone.
- Sub-Saharan Africa is the region where Amazon delivers the most Products(182870 units sold).
- Clothes are most demanding product on Amazon that people prefer to buy.
- Amazon uses both Online Offline mode equally as their Sales Channel.
- H is the highest order Priopity of People in diffrent Countries.
- Product with the Highest Total Cost Sold on Amazon is Office Supplies.
- Amazon earns highest revenue from the country DJIBOUTI.
- Profit earn by Amazon in offine mode i.e of Rs 2,49,20,727 as compare to offline mode i.e of Rs 1,92,47,472.
- Highest Revenue that the Amazon gets is form Baby Food of Rs 1,03,50,328.