Analysing Amazon Sales data

Submitted by – Vitarna Sharma

UNID – UMIP10655

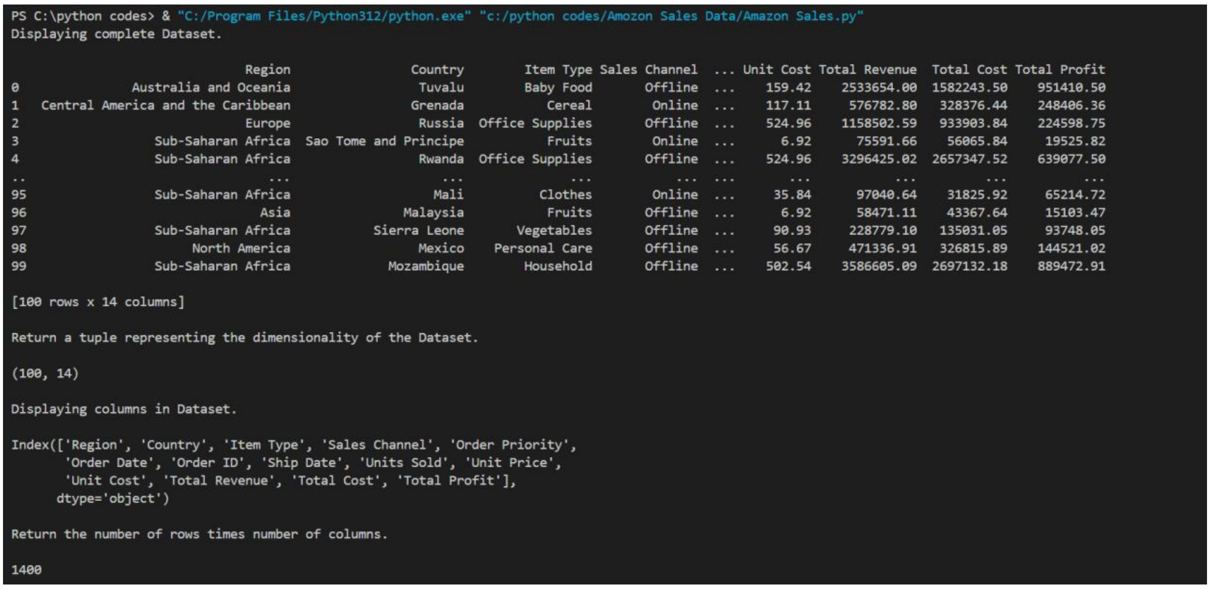
Problem Statement

Sales management has gained importance to meet increasing competition and the need for improved methods of distribution to reduce cost and to increase profits. Sales management today is the most important function in a commercial and business enterprise. Do ETL: Extract-Transform-Load some Amazon dataset and find for me Sales-trend -> month-wise, year-wise, yearly_month-wise Find key metrics and factors and show the meaningful relationships between attributes. Do your own research and come up with your findings

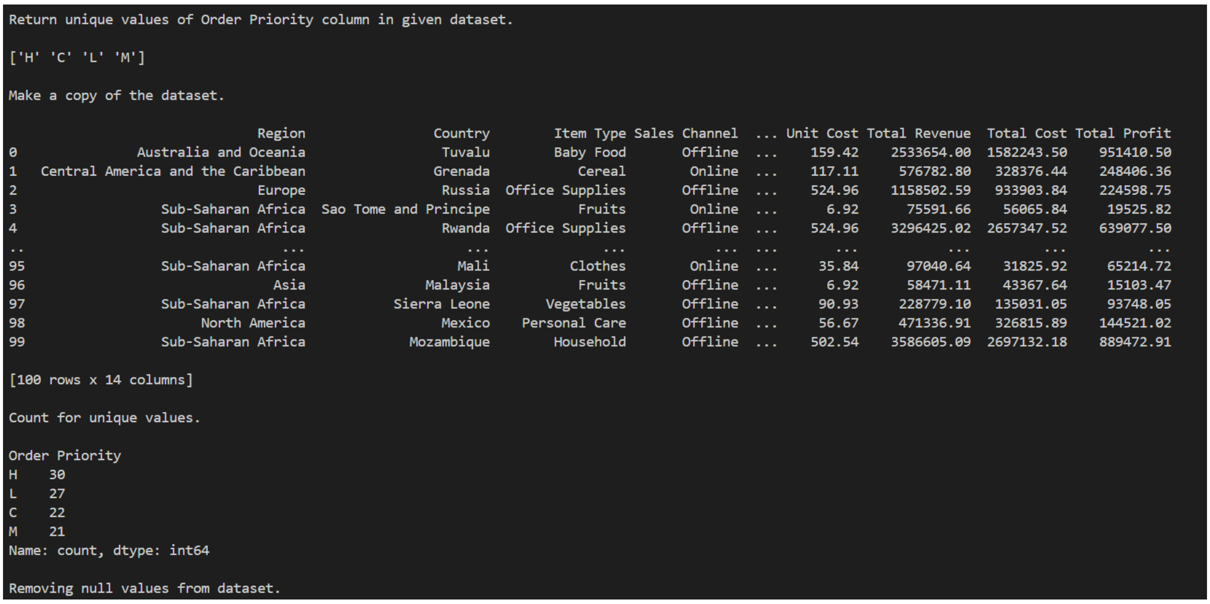
Python Code

```
import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import datetime as dt

# Load the dataset

data = pd.read_csv("C:/python codes/Amozon Sales Data/Amazon Sales data.csv")

# Display the complete dataset

print("Displaying complete Dataset.\n")

print(data)

# Return a tuple representing the dimensionality of the dataset

print("\nReturn a tuple representing the dimensionality of the Dataset.\n")

print(data.shape)

# Display columns in the dataset

print("\nDisplaying columns in Dataset.\n")

print(data.columns)

# Return the number of rows times number of columns

print("\nReturn the number of rows times number of columns.\n")

print(data.size)

# Print a concise summary of the dataset

print("\nPrint a concise summary of a Dataset.\n")

print(data.info())

# Generate descriptive statistics

print("\nGenerate descriptive statistics.\n")

print(data.describe())

# Return an integer value representing the array dimensions

print("\nReturn an integer value representing the array dimensions.\n")
print(data.ndim)

# Detect missing values and return the sum of the values

print("\nDetect missing values and return the sum of the values.\n")

print(data.isnull().sum())

# Return unique values of Order Priority column in the given dataset

print("\nReturn unique values of Order Priority column in given dataset.\n")

print(data['Order Priority'].unique())

# Make a copy of the dataset

print("\nMake a copy of the dataset.\n")

data_copy = data.copy()

print(data_copy)

# Count for unique values

print("\nCount for unique values.\n")

print(data['Order Priority'].value_counts())

# Remove null values from dataset

print("\nRemoving null values from dataset.\n")

data_copy.dropna(subset=['Unit Price', 'Unit Cost', 'Order Priority'], inplace=True)

print(data_copy)

# Generate descriptive statistics for the copied dataset

print("\nGenerate descriptive statistics.\n")

print(data_copy.describe())

# Creating Year, Month, Quarter, Day Columns in dataset

print("Creating Year, Month, Quarter, Day Columns in dataset")

data['Order Date'] = pd.to_datetime(data_copy['Order Date'], errors='coerce')

data_copy['Order_Year'] = data['Order Date'].dt.year

data_copy['Order_Month'] = data['Order Date'].dt.month

data_copy['Order_Quarter'] = data['Order Date'].dt.quarter

data_copy['Order_Day'] = data['Order Date'].dt.day

print(data_copy)

# Save the modified dataframe back to a CSV file

data_copy.to_csv("C:/python codes/Amozon Sales Data/Amazon Sales data_modified.csv", index=False)

# Print a concise summary of the dataset

print("\nPrint a concise summary of a Dataset.\n")

print(data_copy.info())

#Creating Dataframe only with neccessary values

data_selcol = data_copy[['Units Sold', 'Unit Price',  'Unit Cost', 'Total Revenue', 'Total Cost', 'Total Profit','Order_Year', 'Order_Month', 'Order_Quarter', 'Order_Day']]

print("printing a new dataframe with only selected columns")
print(data_selcol)

# Detect missing values and return the sum of the values

print("\nDetect missing values and return the sum of the values.\n")

print(data_selcol.isnull().sum())

#Checking the correaltion

plt.figure(figsize=(16,9))

sns.heatmap(data_selcol.corr(method='pearson'), annot=True, vmin=-1, vmax=1, cmap='YlGnBu')

plt.xticks(rotation=90)

plt.show()

print("OBSERVASTION")

print("1- Units sold highly cause effect on Total Revenue ,Total Cost and Total Profit and moderately related to Unit Price and Unit Cost ")

print("2- Units sold depends on Order Priority List and Country asking for its sales.")

print("3- All the other related observations displayed in tableau.")
```

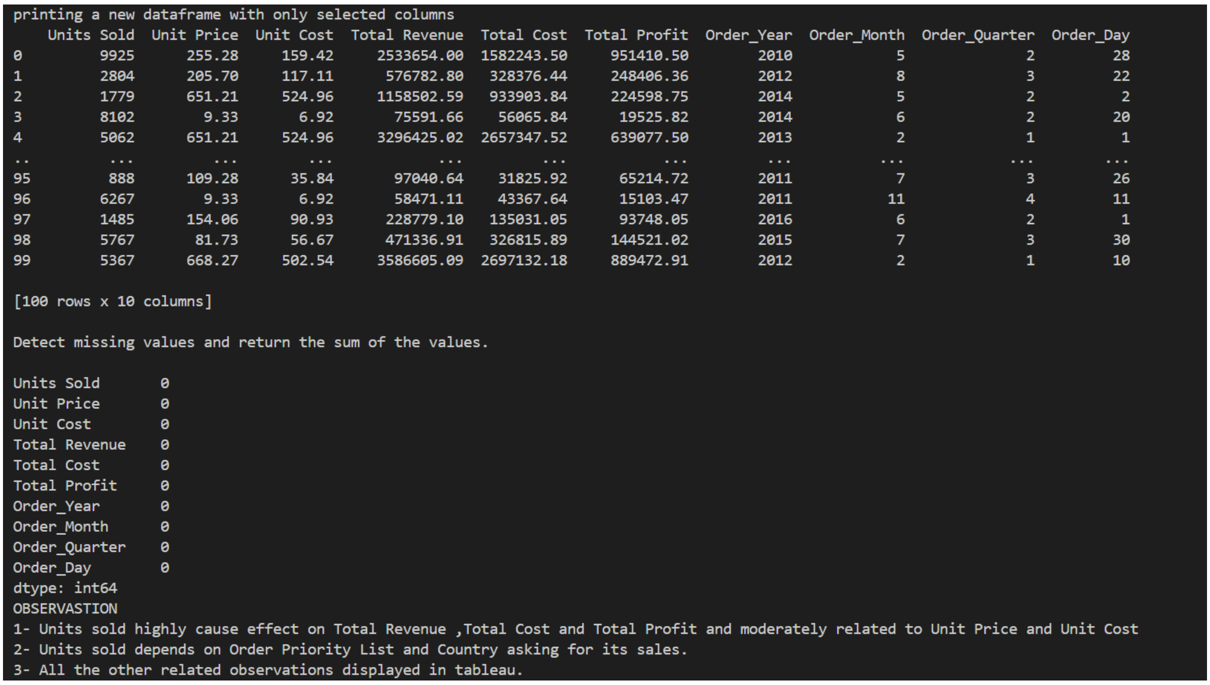
Output(snippets)



```
PS C:\python codes> & "C:/Program Files/Python312/python.exe" "c:/python codes/Amozon Sales Data/Amazon Sales.py"
Displaying complete Dataset.

                               Region                Country        Item Type Sales Channel  ... Unit Cost Total Revenue  Total Cost Total Profit
0               Australia and Oceania                 Tuvalu        Baby Food       Offline  ...    159.42    2533654.00  1582243.50    951410.50
1   Central America and the Caribbean                Grenada           Cereal        Online  ...    117.11     576782.80   328376.44    248406.36
2                              Europe                 Russia  Office Supplies       Offline  ...    524.96    1158502.59   933903.84    224598.75
3                  Sub-Saharan Africa  Sao Tome and Principe           Fruits        Online  ...      6.92      75591.66    56065.84     19525.82
4                  Sub-Saharan Africa                 Rwanda  Office Supplies       Offline  ...    524.96    3296425.02  2657347.52    639077.50
..                                ...                    ...              ...           ...  ...       ...           ...         ...          ...
95                 Sub-Saharan Africa                   Mali          Clothes        Online  ...     35.84      97040.64    31825.92     65214.72
96                               Asia               Malaysia           Fruits       Offline  ...      6.92      58471.11    43367.64     15103.47
97                 Sub-Saharan Africa           Sierra Leone       Vegetables       Offline  ...     90.93     228779.10   135031.05     93748.05
98                      North America                 Mexico    Personal Care       Offline  ...     56.67     471336.91   326815.89    144521.02
99                 Sub-Saharan Africa             Mozambique        Household       Offline  ...    502.54    3586605.09  2697132.18    889472.91

[100 rows x 14 columns]

Return a tuple representing the dimensionality of the Dataset.

(100, 14)

Displaying columns in Dataset.

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')

Return the number of rows times number of columns.

1400
```



```
Return unique values of Order Priority column in given dataset.

['H' 'C' 'L' 'M']

Make a copy of the dataset.

                           Region                Country        Item Type Sales Channel  ... Unit Cost Total Revenue  Total Cost Total Profit
0               Australia and Oceania                 Tuvalu        Baby Food       Offline  ...    159.42    2533654.00  1582243.50    951410.50
1   Central America and the Caribbean                Grenada           Cereal        Online  ...    117.11     576782.80   328376.44    248406.36
2                              Europe                 Russia  Office Supplies       Offline  ...    524.96    1158502.59   933903.84    224598.75
3                  Sub-Saharan Africa  Sao Tome and Principe           Fruits        Online  ...      6.92      75591.66    56065.84     19525.82
4                  Sub-Saharan Africa                 Rwanda  Office Supplies       Offline  ...    524.96    3296425.02  2657347.52    639077.50
..                                ...                    ...              ...           ...  ...       ...           ...         ...          ...
95                 Sub-Saharan Africa                   Mali          Clothes        Online  ...     35.84      97040.64    31825.92     65214.72
96                               Asia               Malaysia           Fruits       Offline  ...      6.92      58471.11    43367.64     15103.47
97                 Sub-Saharan Africa           Sierra Leone       Vegetables       Offline  ...     90.93     228779.10   135031.05     93748.05
98                      North America                 Mexico    Personal Care       Offline  ...     56.67     471336.91   326815.89    144521.02
99                 Sub-Saharan Africa             Mozambique        Household       Offline  ...    502.54    3586605.09  2697132.18    889472.91

[100 rows x 14 columns]

Count for unique values.

Order Priority
H    30
L    27
C    22
M    21
Name: count, dtype: int64

Removing null values from dataset.
```



```
printing a new dataframe with only selected columns
    Units Sold  Unit Price  Unit Cost  Total Revenue  Total Cost  Total Profit  Order_Year  Order_Month  Order_Quarter  Order_Day
0         9925      255.28     159.42     2533654.00  1582243.50     951410.50        2010            5              2         28
1         2804      205.70     117.11      576782.80   328376.44     248406.36        2012            8              3         22
2         1779      651.21     524.96     1158502.59   933903.84     224598.75        2014            5              2          2
3         8102        9.33       6.92       75591.66    56065.84      19525.82        2014            6              2         20
4         5062      651.21     524.96     3296425.02  2657347.52     639077.50        2013            2              1          1
..         ...         ...        ...            ...         ...           ...         ...          ...            ...        ...
95         888      109.28      35.84       97040.64    31825.92      65214.72        2011            7              3         26
96        6267        9.33       6.92       58471.11    43367.64      15103.47        2011           11              4         11
97        1485      154.06      90.93      228779.10   135031.05      93748.05        2016            6              2          1
98        5767       81.73      56.67      471336.91   326815.89     144521.02        2015            7              3         30
99        5367      668.27     502.54     3586605.09  2697132.18     889472.91        2012            2              1         10

[100 rows x 10 columns]

Detect missing values and return the sum of the values.

Units Sold       0
Unit Price       0
Unit Cost        0
Total Revenue    0
Total Cost       0
Total Profit     0
Order_Year       0
Order_Month      0
Order_Quarter    0
Order_Day        0
dtype: int64
OBSERVASTION
1- Units sold highly cause effect on Total Revenue ,Total Cost and Total Profit and moderately related to Unit Price and Unit Cost
2- Units sold depends on Order Priority List and Country asking for its sales.
3- All the other related observations displayed in tableau.
```

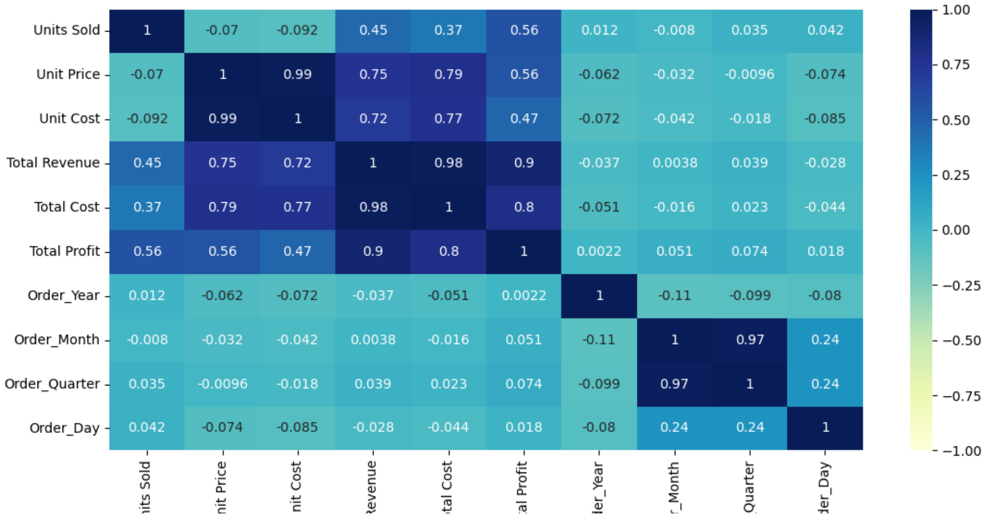


Tableau Link

Dashboard 1

Dashboard 2

Story