PHASE 3 : DEVELOPMENT – PART 1

PROJECT TITLE : PRODUCT SALES ANALYSIS

INTRODUCTION

In phase3 we start building the product sales analysis using IBM Cognos for visualization by loading and preprocessing the provided data set from kaggle

Dataset Link: https://www.kaggle.com/datasets/ksabishek/product-sales-data

We define the analysis objectives and collect sales data from source shared.

Process and clean the collected data to ensure its accuracy and reliability.

Steps Followed :

Step 1 : Import the libraries

We need to Import the required Libraries to load and Preprocess the dataset . we have used libraries like Pandas , Matplotlib , Seaborn and Numpy

In [1] : import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

pd.options.display.max\_columns=50

sns.set(style="darkgrid")

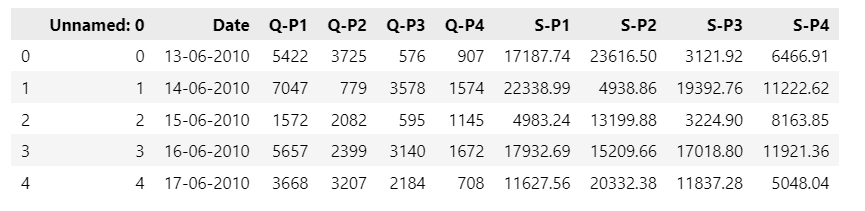
Step 2 : Load dataset

To use the data set for our analysis we need to import the dataset. We can import it using pandas read\_csv() build in function

In [2]: df=pd.read\_csv("./statsfinal.csv")

df.head(5)

out [2] :



In [3] :df.shape

out [3] :

(4600, 10)

In [4] :df.columns

out [4] :

Index(['Unnamed: 0', 'Date', 'Q-P1', 'Q-P2', 'Q-P3', 'Q-P4', 'S-P1', 'S-P2',

'S-P3', 'S-P4'],

dtype='object')

In [5] :df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 4600 entries, 0 to 4599

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 4600 non-null int64

1 Date 4600 non-null object

2 Q-P1 4600 non-null int64

3 Q-P2 4600 non-null int64

4 Q-P3 4600 non-null int64

5 Q-P4 4600 non-null int64

6 S-P1 4600 non-null float64

7 S-P2 4600 non-null float64

8 S-P3 4600 non-null float64

9 S-P4 4600 non-null float64

dtypes: float64(4), int64(5), object(1)

memory usage: 359.5+ KB

In [6] :df.isnull().sum()

Unnamed: 0 0

Date 0

Q-P1 0

Q-P2 0

Q-P3 0

Q-P4 0

S-P1 0

S-P2 0

S-P3 0

S-P4 0

dtype: int64

In [7] :df.dtypes

Unnamed: 0 int64

Date object

Q-P1 int64

Q-P2 int64

Q-P3 int64

Q-P4 int64

S-P1 float64

S-P2 float64

S-P3 float64

S-P4 float64

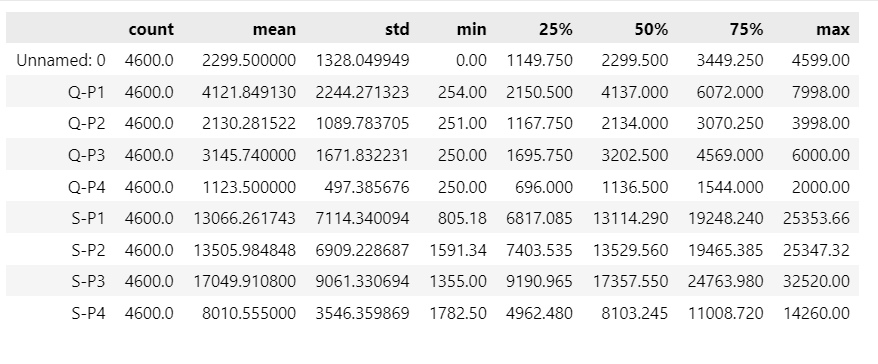
dtype: object

In [8] :df.duplicated().sum()

0

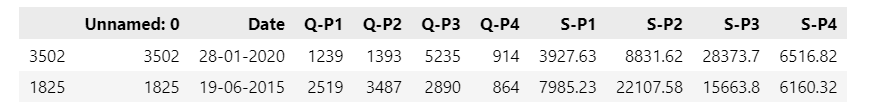
In [9] :df.describe().T

out [9] :



In [10] :df.sample(2)

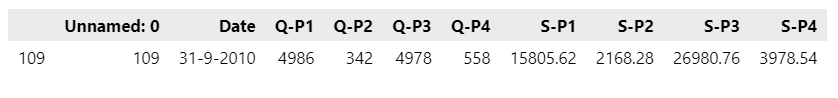
out [10] :



In [11] :from datetime import datetime as dt

df[df["Date"]=="31-9-2010"]

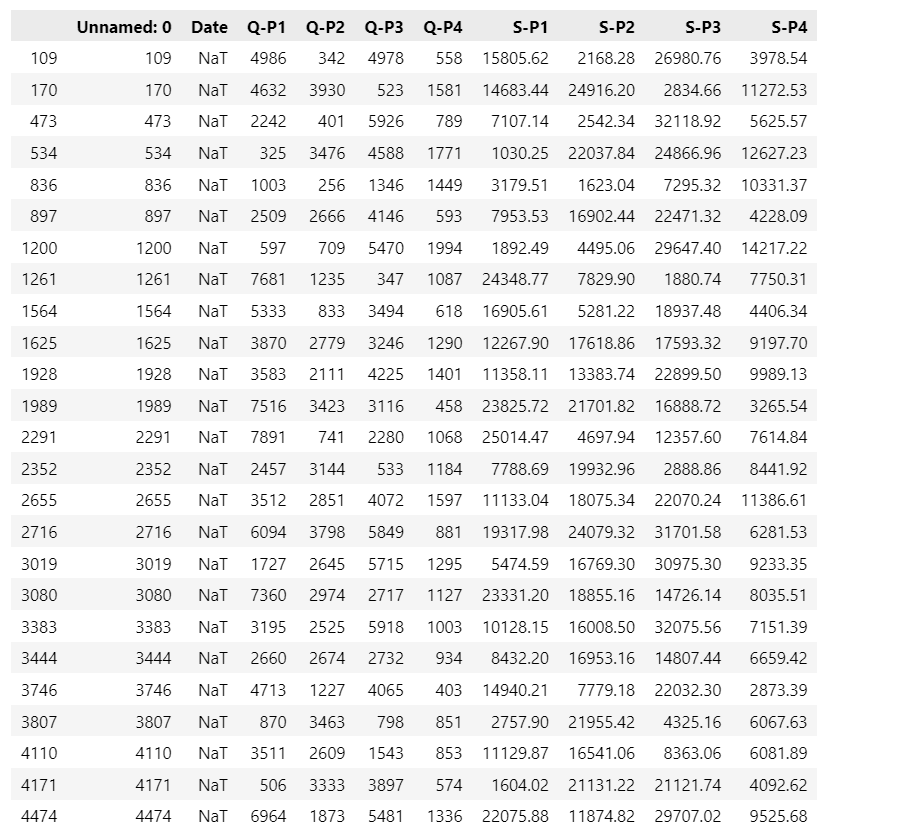
out [11] :



In [12] :df['Date'] = pd.to\_datetime(df['Date'], errors='coerce',dayfirst=True)

In [13] :df[df['Date'].isnull()]

out [13] :



In [14] :df["Date"].fillna(df["Date"].mean(),inplace=True)

In [15] :df['Date'].isnull().sum()

out [15] :

0

In [16] :df.dtypes

Unnamed: 0 int64

Date datetime64[ns]

Q-P1 int64

Q-P2 int64

Q-P3 int64

Q-P4 int64

S-P1 float64

S-P2 float64

S-P3 float64

S-P4 float64

dtype: object

In [17] :df["month"]=df["Date"].dt.month\_name()

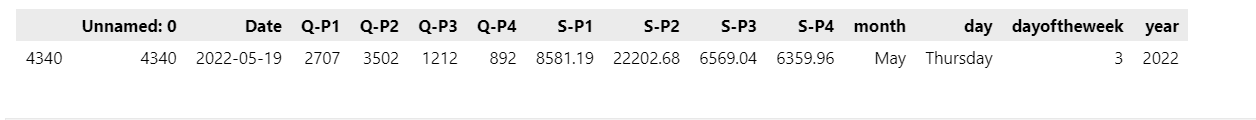
df["day"]=df["Date"].dt.day\_name()

df["dayoftheweek"]=df["Date"].dt.weekday

df["year"]=df["Date"].dt.year

df.sample()

out [17] :



In [18] :df.dtypes

Unnamed: 0 int64

Date datetime64[ns]

Q-P1 int64

Q-P2 int64

Q-P3 int64

Q-P4 int64

S-P1 float64

S-P2 float64

S-P3 float64

S-P4 float64

month object

day object

dayoftheweek int32

year int32

dtype: object

In [19] :df.drop(columns=["Unnamed: 0"],inplace=True)

df.sample()

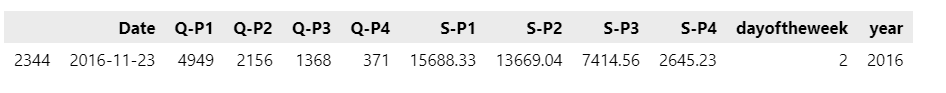
df.drop(columns=["month"],inplace=True)

df.sample()

df.drop(columns=["day"],inplace=True)

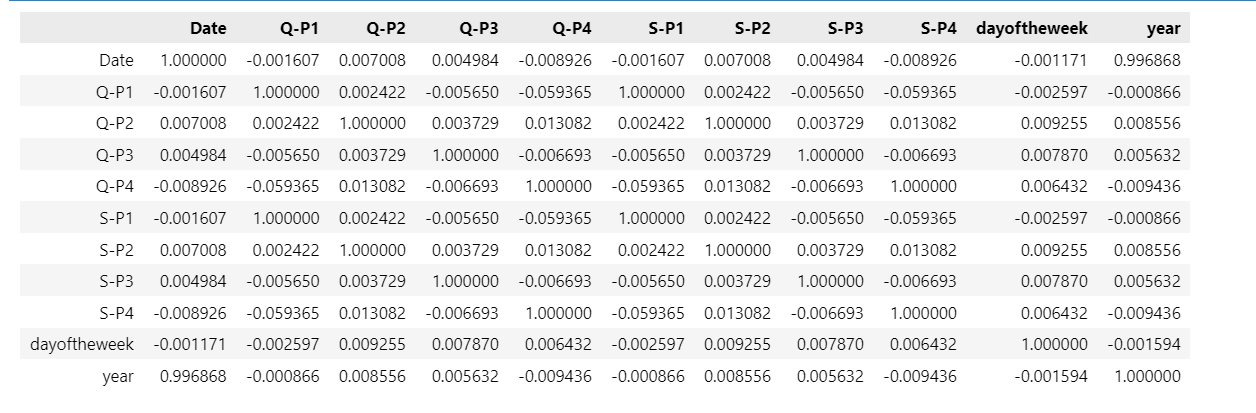
df.sample()

out [19] :



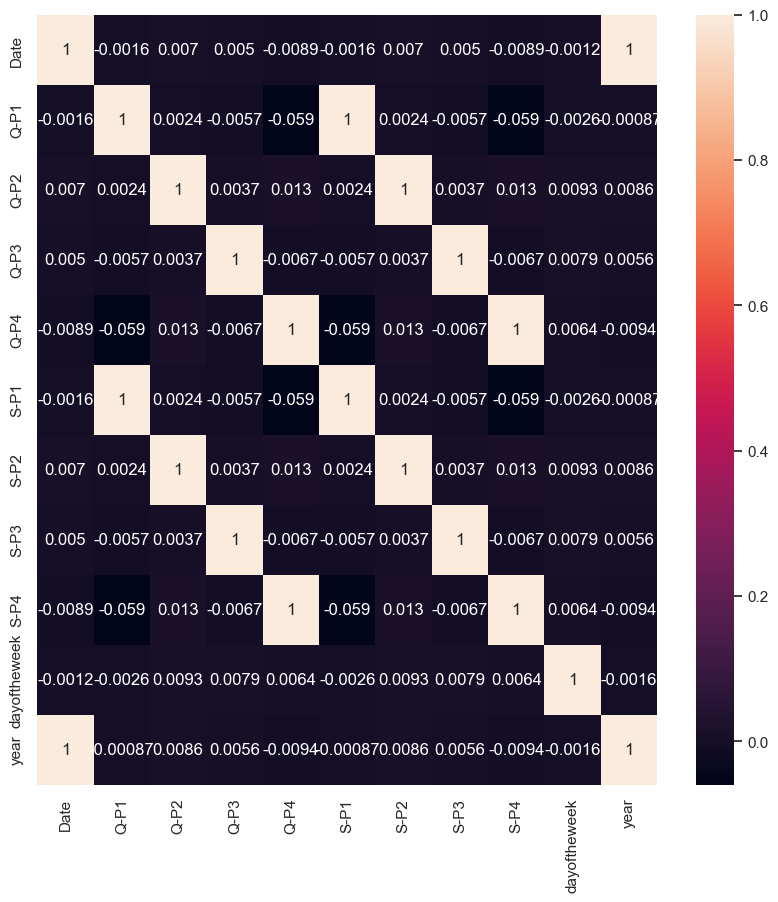
In [22] :df.corr().T

out [22] :



In [23] :plt.figure(figsize=(10,10))

sns.heatmap(df.corr(),annot=True)



In [24]:for i in df.columns:

    print(i,"---------",df[i].unique())

Date --------- <DatetimeArray>

['2010-06-13 00:00:00', '2010-06-14 00:00:00', '2010-06-15 00:00:00',

'2010-06-16 00:00:00', '2010-06-17 00:00:00', '2010-06-18 00:00:00',

'2010-06-19 00:00:00', '2010-06-20 00:00:00', '2010-06-21 00:00:00',

'2010-06-22 00:00:00',

...

'2023-01-25 00:00:00', '2023-01-26 00:00:00', '2023-01-27 00:00:00',

'2023-01-28 00:00:00', '2023-01-29 00:00:00', '2023-01-30 00:00:00',

'2023-01-31 00:00:00', '2023-02-01 00:00:00', '2023-02-02 00:00:00',

'2023-02-03 00:00:00']

Length: 4575, dtype: datetime64[ns]

Q-P1 --------- [5422 7047 1572 ... 1227 3122 1234]

Q-P2 --------- [3725 779 2082 ... 3404 841 3143]

Q-P3 --------- [ 576 3578 595 ... 4825 3588 5899]

Q-P4 --------- [ 907 1574 1145 ... 1161 1151 1112]

S-P1 --------- [17187.74 22338.99 4983.24 ... 3889.59 9896.74 3911.78]

S-P2 --------- [23616.5 4938.86 13199.88 ... 21581.36 5331.94 19926.62]

S-P3 --------- [ 3121.92 19392.76 3224.9 ... 26151.5 19446.96 31972.58]

S-P4 --------- [ 6466.91 11222.62 8163.85 ... 8277.93 8206.63 7928.56]

dayoftheweek --------- [6 0 1 2 3 4 5]

year --------- [2010 2016 2011 2012 2013 2014 2015 2017 2018 2019 2020 2021 2022 2023]