NAAN MUDHALVAN

PHASE 3 PROJECT SUBMISSION

**PRODUCT SALES ANALYSIS**

**TEAM MEMBERS:**

PARVATHY A - 2021504533

PAVITHRAN E - 2021504535

SHAMEEM AHAMED S - 2021504542

THUSHIYANTH K - 2021504555

**PROBLEM DEFINITION:**

This project involves using IBM Cognos to analyze sales data and extract insights about top-selling products, peak sales periods, and customer preferences. The objective is to help businesses improve inventory management and marketing strategies by understanding sales trends and customer behavior. This project includes defining analysis objectives, collecting sales data, designing relevant visualizations in IBM Cognos, and deriving actionable insights.

**DATABASE LINK:**

[**https://www.kaggle.com/datasets/ksabishek/product-sales-data**](https://www.kaggle.com/datasets/ksabishek/product-sales-data)

**OBJECTIVES:**

* Analysing data related to product sales.
* Generating valuable insights from the data.
* Based on the insights, recommendations must be formulated to address issues and optimize sales and profitability.

**Cleaning and Preprocessing:**

**Importing the important packages:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# To ignore warnings

import warnings

warnings.filterwarnings("ignore")

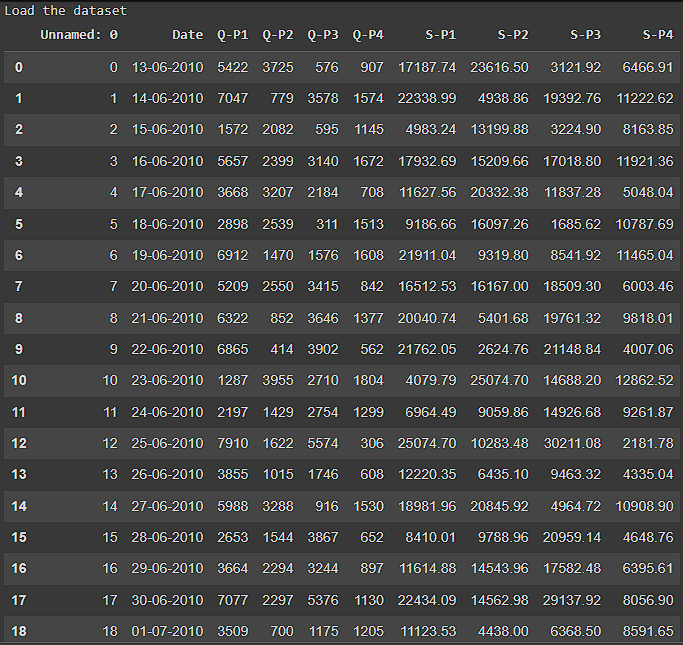
**Loading the dataset:**

print("Load the dataset")

import pandas as pd

data = pd.read\_csv('/statsfinal.csv', low\_memory=False) data.shape

data.head(30)

****

**No. of rows and columns:**

print("Load the dataset")

import pandas as pd

data = pd.read\_csv('/statsfinal.csv', low\_memory=False)

data.shape

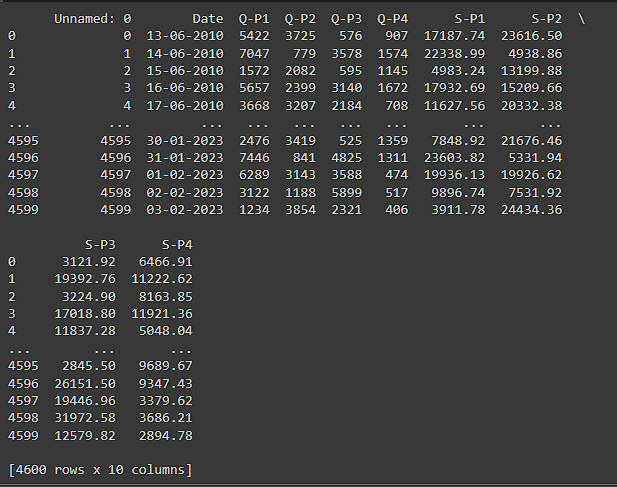
data.head(30)

input\_file = "/statsfinal.csv"

df = pd.read\_csv(input\_file)

print(df)

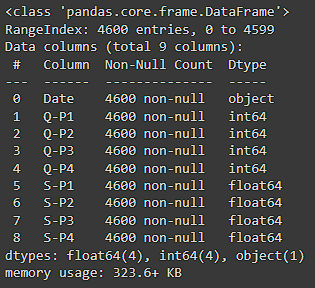
**OUTPUT:**



**DROPPING FIRST COLUMN:**

data = data.drop(columns=['Unnamed: 0'])

data.info()



**Cleaned File:**

output\_file = "cleaned\_file.csv"

df.to\_csv(output\_file, index=False)

print(output\_file)

**Output:**

cleaned\_file.csv

**Loading the cleaned CSV file:**

input\_file = "cleaned\_file.csv"

df = pd.read\_csv(input\_file)

**Histogram visualization:**

plt.figure(figsize=(8, 6))

plt.hist(df['S-P1'], bins=20, color='red')

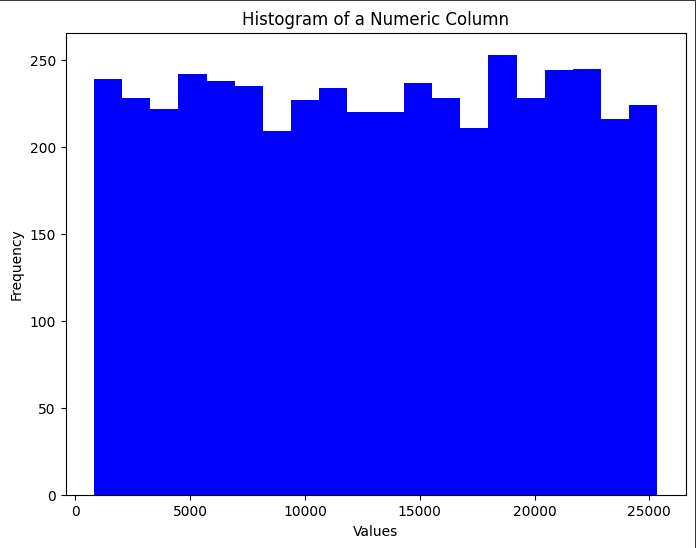
plt.title('Histogram of a Numeric Column')

plt.xlabel('Values')

plt.ylabel('Frequency')

plt.show()

**Output:**

****

**Scatter Plot Visualization:**

plt.figure(figsize=(8, 6))

plt.scatter(df['S-P1'], df['S-P2'], color='blue', alpha=0.5)

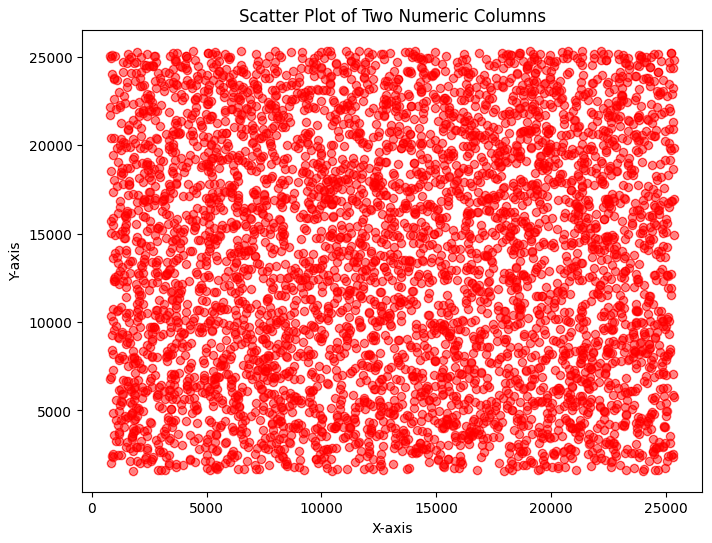
plt.title('Scatter Plot of Two Numeric Columns')

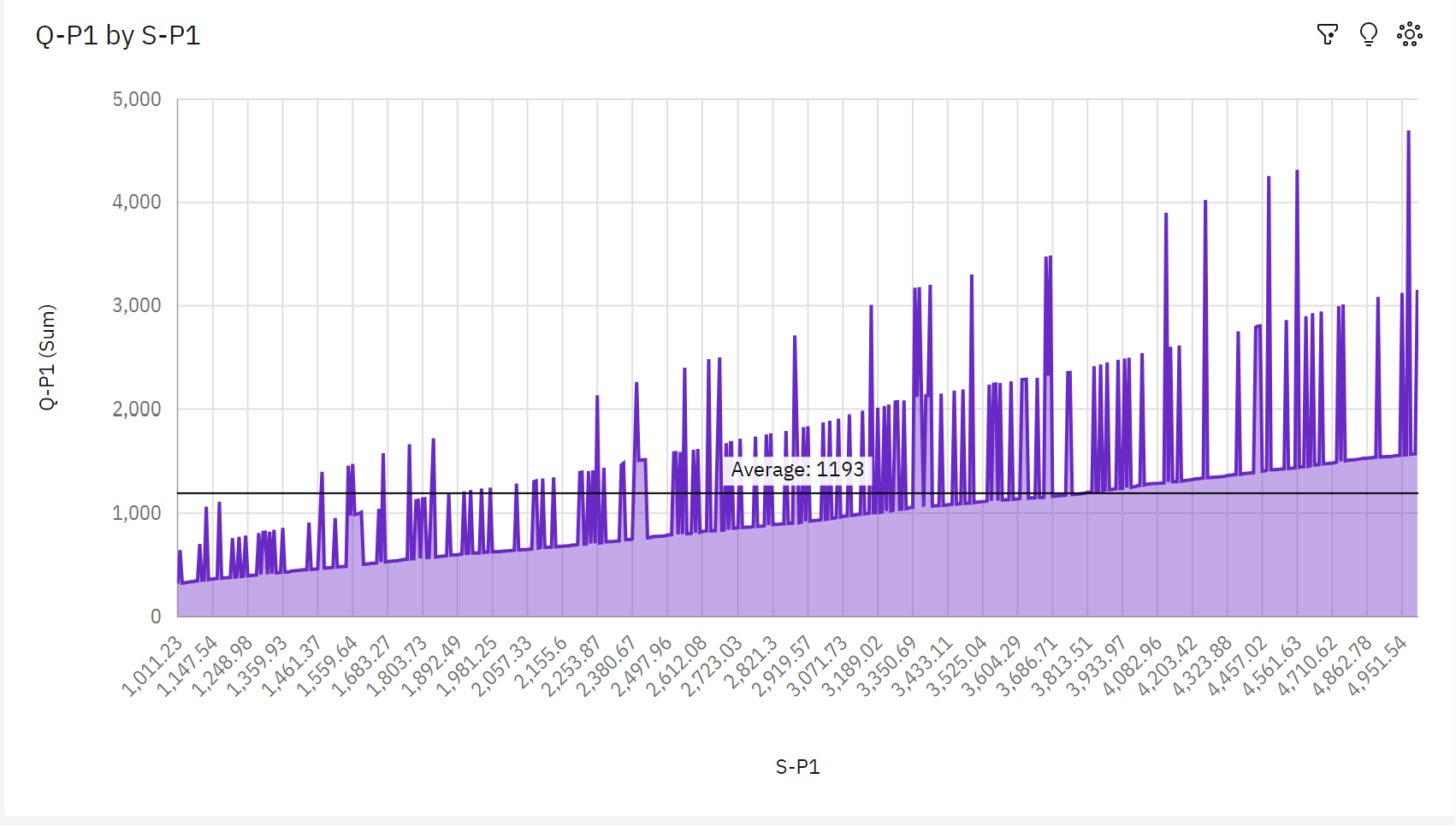
plt.xlabel('X-axis')

plt.ylabel('Y-axis')

plt.show()

**Output:**

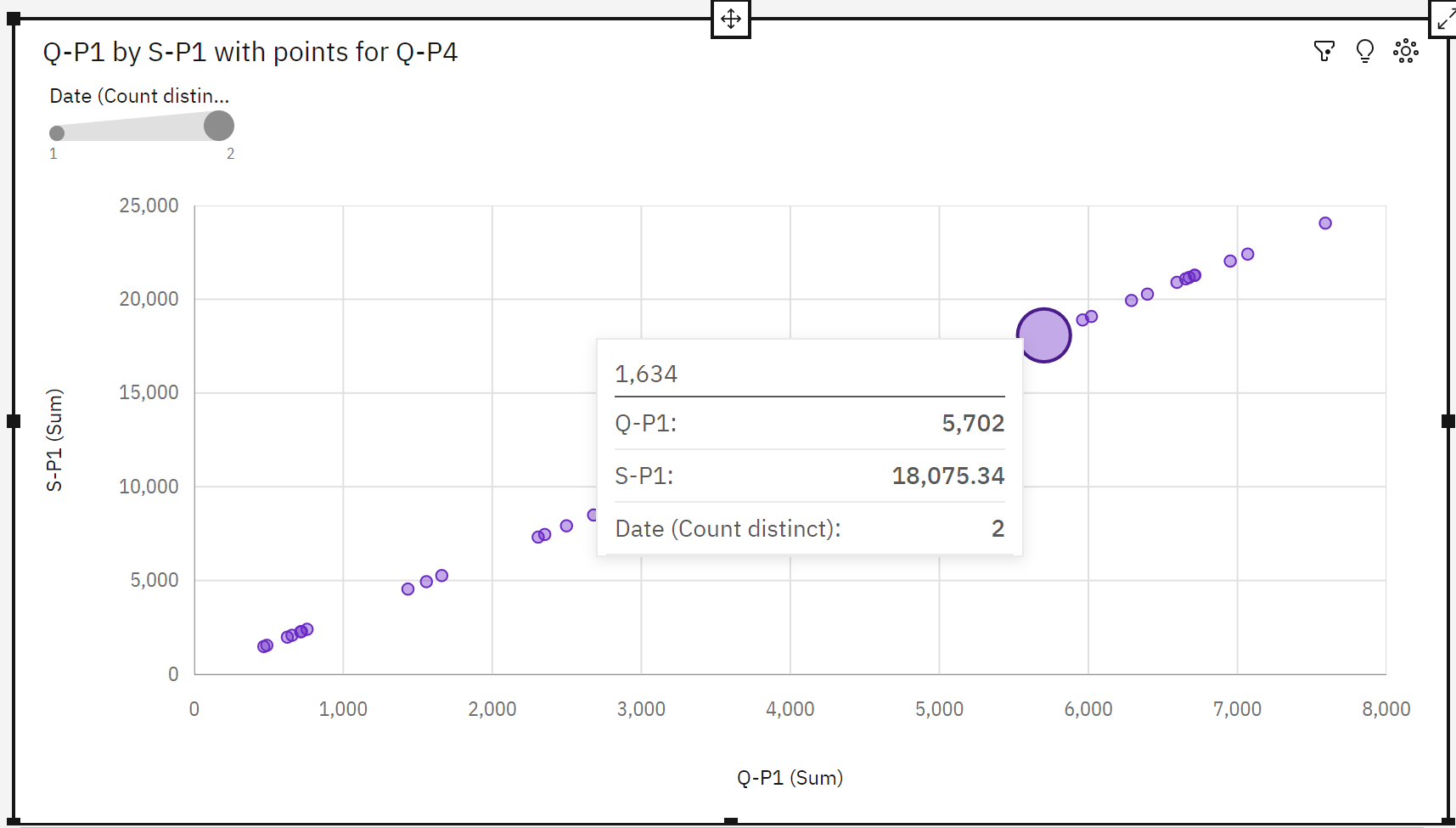
****

**VISUALIZATION 1:**

**Insights:**

* Over all values of S-P1, the sum of Q-P1 is nearly 678 thousand.
* **Q-P1** ranges from **319**, when **S-P1** is **1011.23**, to **over 4500**, when **S-P1** is **4964.22**.
* **S-P1** **3683.54** has the highest **Total Q-P4** but is ranked #**6** in **Total Q-P1**.
* **S-P1** **4964.22** has the highest **Total Q-P1** but is ranked #**238** in **Total Q-P4**.
* **Q-P4** and **Q-P1** diverged the most when **S-P1** is **1122.18**, and when **Q-P4** was **over three thousand** higher than the **Q-P1**.
* **3683.54** **Q-P4** at **over 5 thousand** is **31**% higher than the **Q-P1** of **almost 3500**.

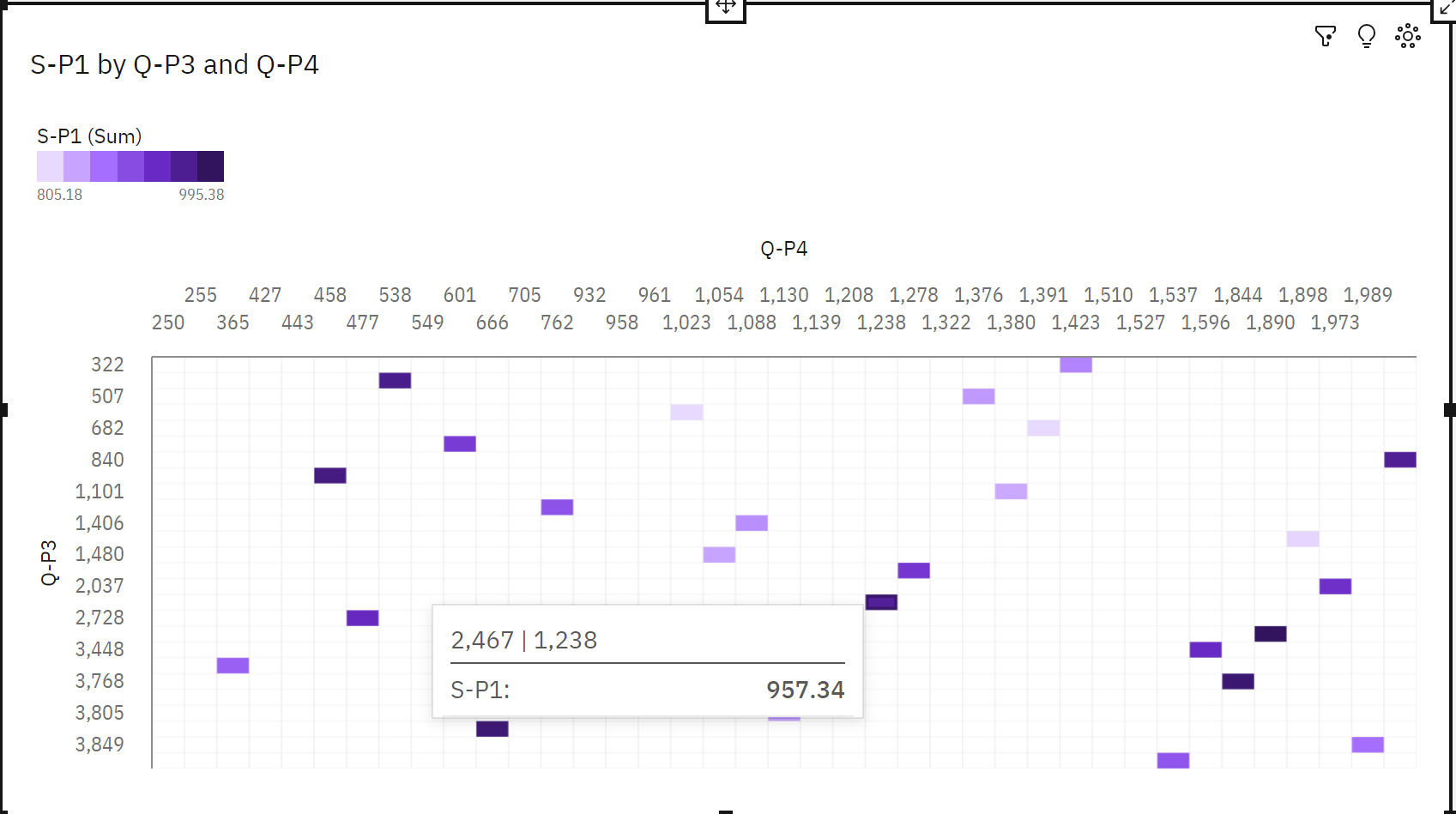
**VISUALIZATION 2:**



**Insights:**

* **S-P4** at **over 23 thousand** is **76**% higher than the **Q-P1** of **over 5500**.
* **S-P4** and **Q-P1** diverged the most when **Q-P4** is **1634**, and when **S-P4** was **nearly 18 thousand** higher than the **Q-P1**.
* **Q-P4** **960** has the highest **Total Q-P1** but is ranked #**30** in **Total S-P4.**
* **Q-P4 1634 has the highest Total S-P4 but is ranked #13 in Total Q-P1.**
* **The total of S-P1 is nearly 575 thousand.**

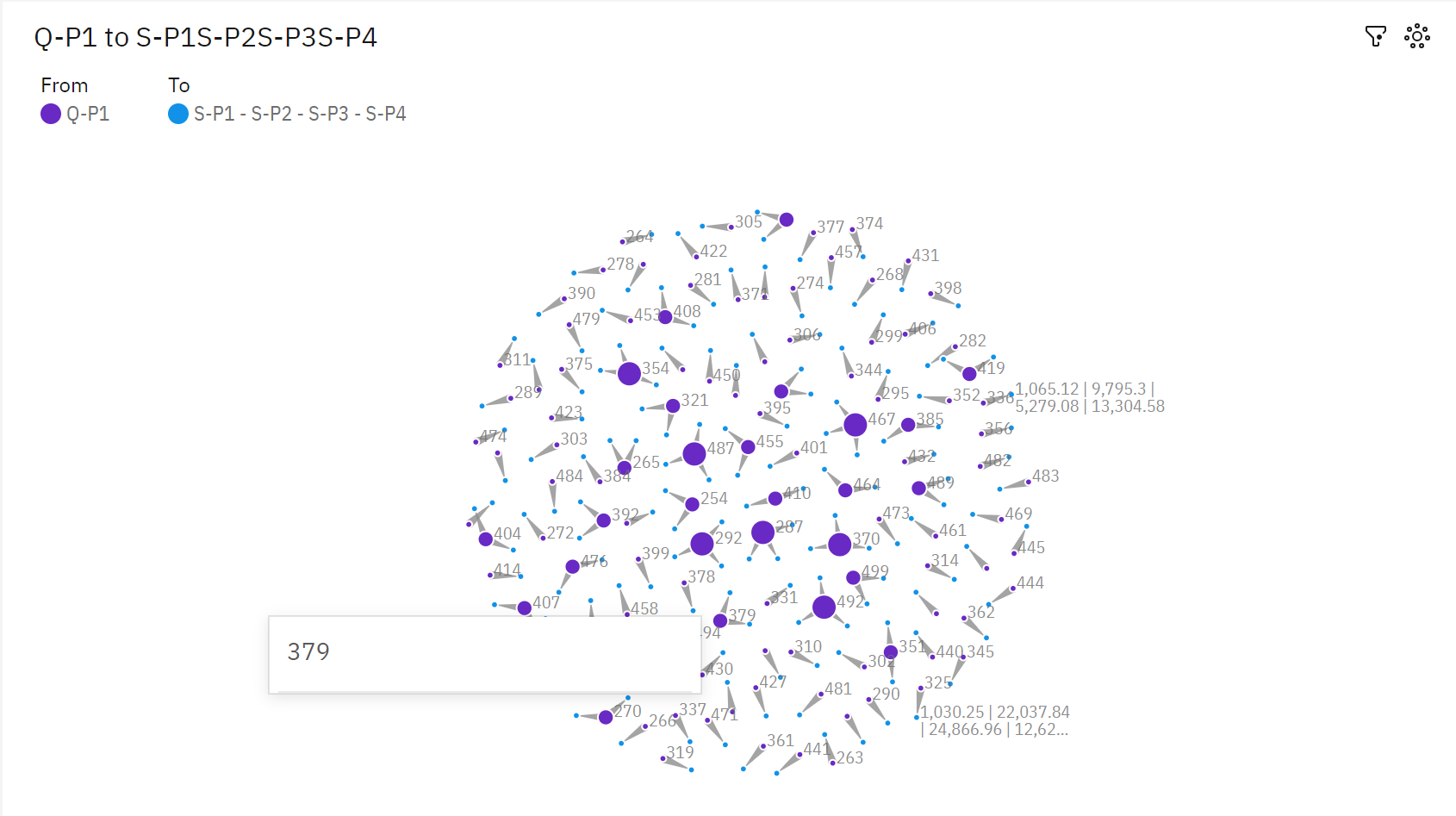
**VISUALIZATION 3:**



**Insights:**

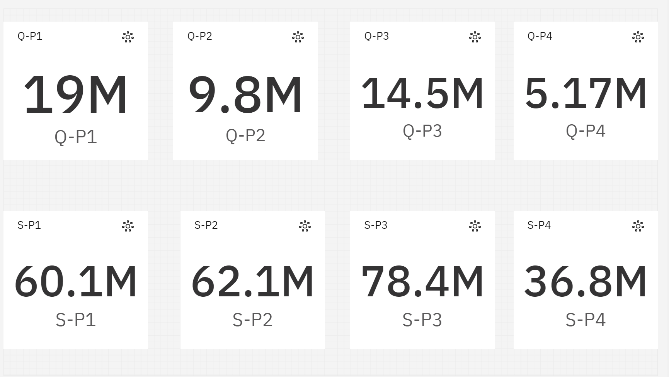
* The summed values of **S-P1** range from **805.2** to **995.4**.
* For **S-P1**, the most significant value of **Q-P4** is **1890**, whose respective **S-P1** values add up to **995.4**, or **2.8** % of the total.
* For **S-P1**, the most significant value of **Q-P3** is **2773**, whose respective **S-P1** values add up to **995.4**, or **2.8** % of the total.
* Across all values of **Q-P3** and **Q-P4**, the sum of **S-P1** is **over 35 thousand**
* **2773** **S-P4** at **13,476** is **93**% higher than the **S-P1** of **995.4**.
* **250** **S-P1** at **985.9** is **68**% higher than the **Q-P1** of **311**.
* **Q-P4** **250** has the highest values of both **S-P1** and **Q-P1**.
* **Q-P3** **2773** has the highest values of both **S-P1** and **S-P4**.
* **Q-P3** **2773** has the highest total **S-P1** due to **Q-P4** **1890**.

**VISUALIZATION 4:**



INSIGHTS:

* **S-P1** **1445.52** has the highest **Unaggregated Q-P1** but is ranked #**99** in **Total Q-P4**
* **S-P1** **1122.18** has the highest **Total Q-P4** but is ranked #**108** in **Unaggregated Q-P1**.

**SUMMARY:**