PROJECT:PRODUCT SALES ANALYSIS

PHASE 3- DEVELOPMENT PHASE 3

PREPROCESSING AND CLEANSING OF DATA

CLEANING OF DATASET:

Cleaning of the dataset includes removing duplicates,handling the missing values, handling outliers,data scaling and normalization, data visualization,data splitting and data balancing if needed.

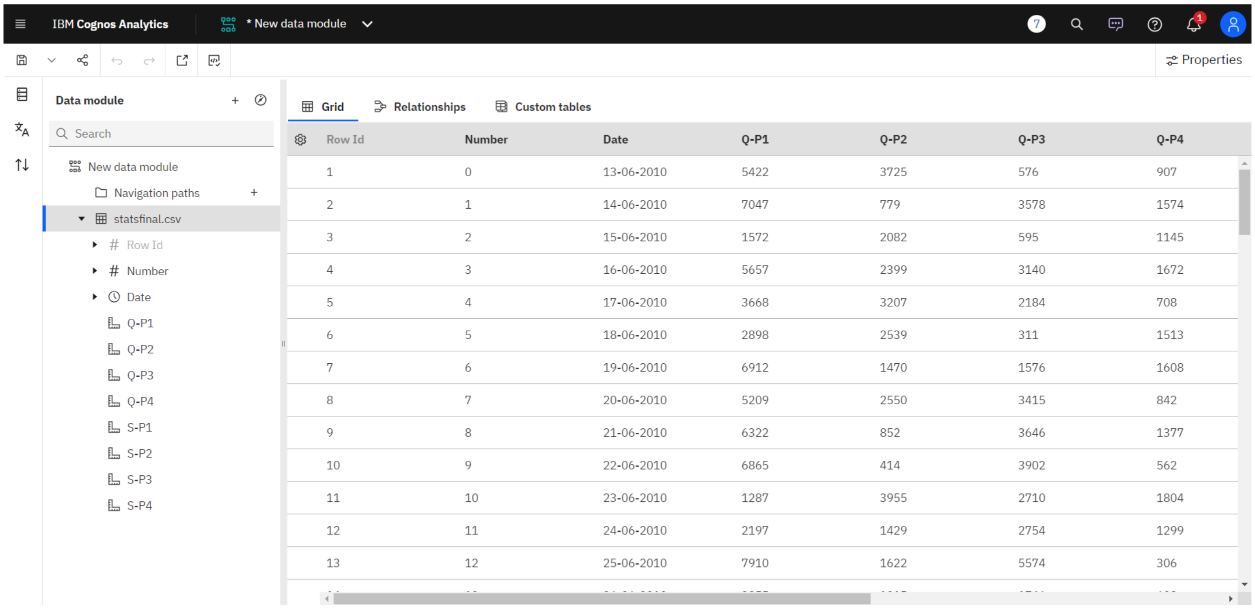
1.Removing duplicates:

data=data.dropna()

Out[8]:

	Unnamed: 0	Date	Q-P1	Q-P2	Q-P3	Q-P4	S-P1	S-P2	S-P3	S-P4
0	0	13-06-2010	5422	3725	576	907	17187.74	23616.50	3121.92	6466.91
1	1	14-06-2010	7047	779	3578	1574	22338.99	4938.86	19392.76	11222.62
2	2	15-06-2010	1572	2082	595	1145	4983.24	13199.88	3224.90	8163.85
3	3	16-06-2010	5657	2399	3140	1672	17932.69	15209.66	17018.80	11921.36
4	4	17-06-2010	3668	3207	2184	708	11627.56	20332.38	11837.28	5048.04
5	5	18-06-2010	2898	2539	311	1513	9186.66	16097.26	1685.62	10787.69
6	6	19-06-2010	6912	1470	1576	1608	21911.04	9319.80	8541.92	11465.04
7	7	20-06-2010	5209	2550	3415	842	16512.53	16167.00	18509.30	6003.46
8	8	21-06-2010	6322	852	3646	1377	20040.74	5401.68	19761.32	9818.01
9	9	22-06-2010	6865	414	3902	562	21762.05	2624.76	21148.84	4007.06

From IBM Cognos:



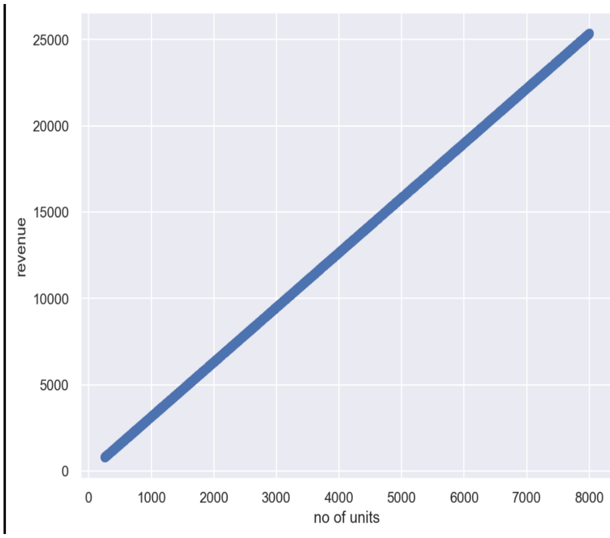
Row Id	Number	Date	Q-P1	Q-P2	Q-P3	Q-P4
1	0	13-06-2010	5422	3725	576	907
2	1	14-06-2010	7047	779	3578	1574
3	2	15-06-2010	1572	2082	595	1145
4	3	16-06-2010	5657	2399	3140	1672
5	4	17-06-2010	3668	3207	2184	708
6	5	18-06-2010	2898	2539	311	1513
7	6	19-06-2010	6912	1470	1576	1608
8	7	20-06-2010	5209	2550	3415	842
9	8	21-06-2010	6322	852	3646	1377
10	9	22-06-2010	6865	414	3902	562
11	10	23-06-2010	1287	3955	2710	1804
12	11	24-06-2010	2197	1429	2754	1299
13	12	25-06-2010	7910	1622	5574	306

2. Handling outliers:

On checking outliers by scatter plot .

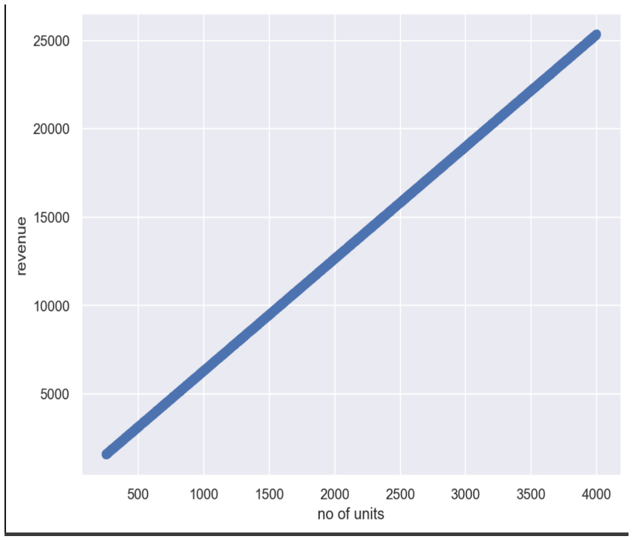
For product1:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Set some default parameters of matplotlib
plt.rcParams['figure.figsize'] = (8, 6)
plt.rcParams['figure.dpi'] = 150

# Use style froms seaborn. Try to comment the next line and see the difference in graph
sns.set()


# A regular scatter plot
plt.scatter(x=data["Q-P1"], y=data["S-P1"])

# Create labels for axises
plt.xlabel('no of units')
plt.ylabel('revenue')

# Display the plot on the screen
plt.show()
```



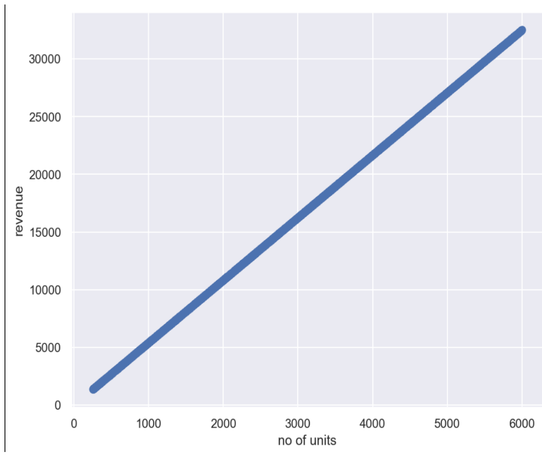
For product 2:

```
plt.rcParams['figure.figsize'] = (8, 6)
plt.rcParams['figure.dpi'] = 150

# Use style froms seaborn. Try to comment the next line and see the difference in graph
sns.set()


# A regular scatter plot
plt.scatter(x=data["Q-P2"], y=data["S-P2"])

# Create labels for axises
plt.xlabel('no of units')
plt.ylabel('revenue')

# Display the plot on the screen
plt.show()
```



For product 3:

```
plt.rcParams['figure.figsize'] = (8, 6)
plt.rcParams['figure.dpi'] = 150

# Use style froms seaborn. Try to comment the next line and see the difference in graph
sns.set()


# A regular scatter plot
plt.scatter(x=data["Q-P3"], y=data["S-P3"])

# Create labels for axises
plt.xlabel('no of units')
plt.ylabel('revenue')

# Display the plot on the screen
plt.show()
```



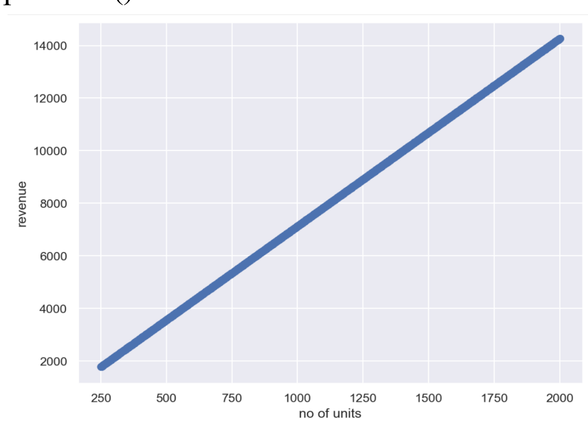
For product 4:
plt.rcParams['figure.figsize'] = (8, 6)
plt.rcParams['figure.dpi'] = 150

# Use style froms seaborn. Try to comment the next line and see the difference in graph
sns.set()

# A regular scatter plot
plt.scatter(x=data["Q-P4"], y=data["S-P4"])

# Create labels for axises
plt.xlabel('no of units')
plt.ylabel('revenue')

# Display the plot on the screen
plt.show()



SUMMARY OF OUR CLEANSED DATA:

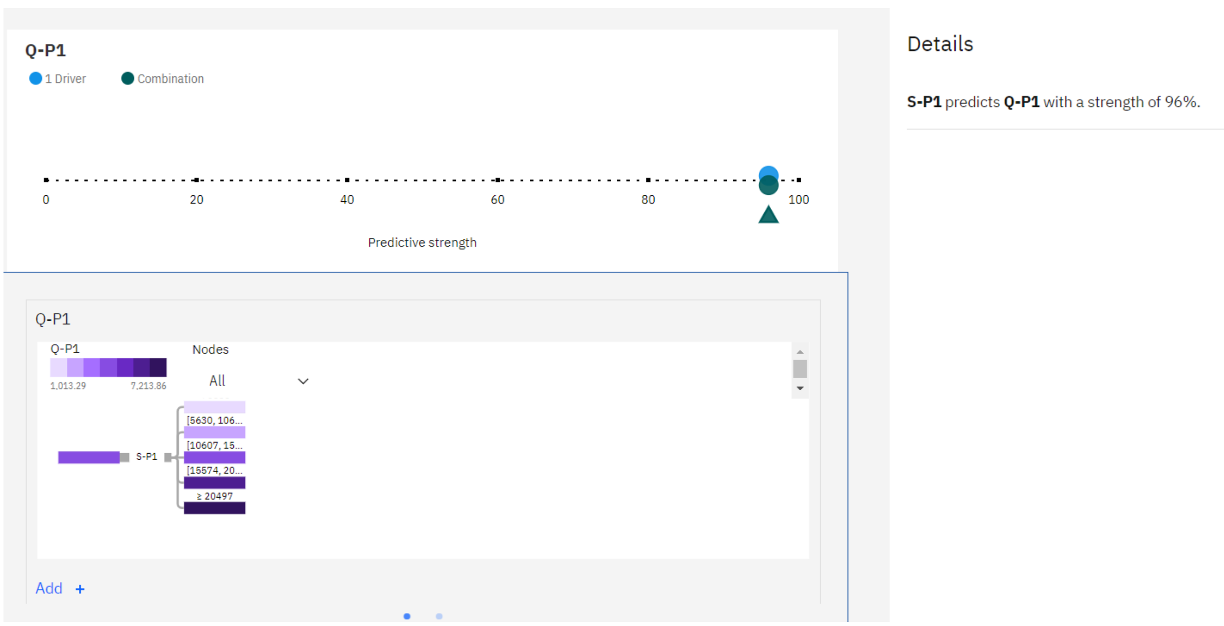
data.info()

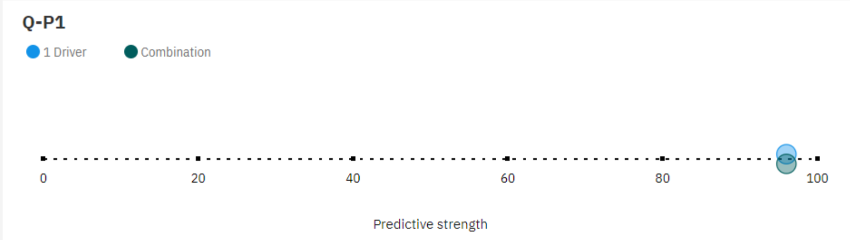
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   4600 non-null   int64
 1   Date         4600 non-null   object
 2   Q-P1         4600 non-null   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
```

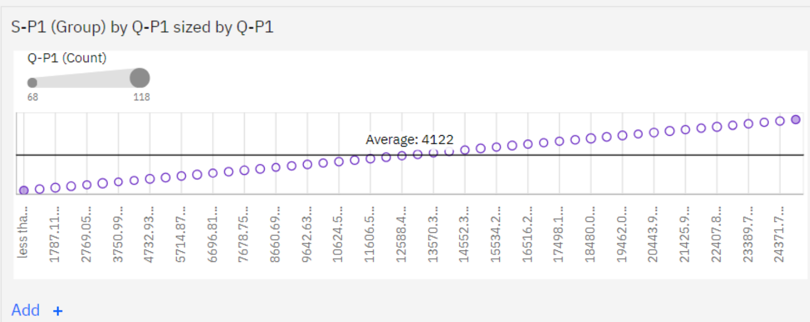
data.describe()

Accuracy of dataset performed by IBM cognos:

Q-P1

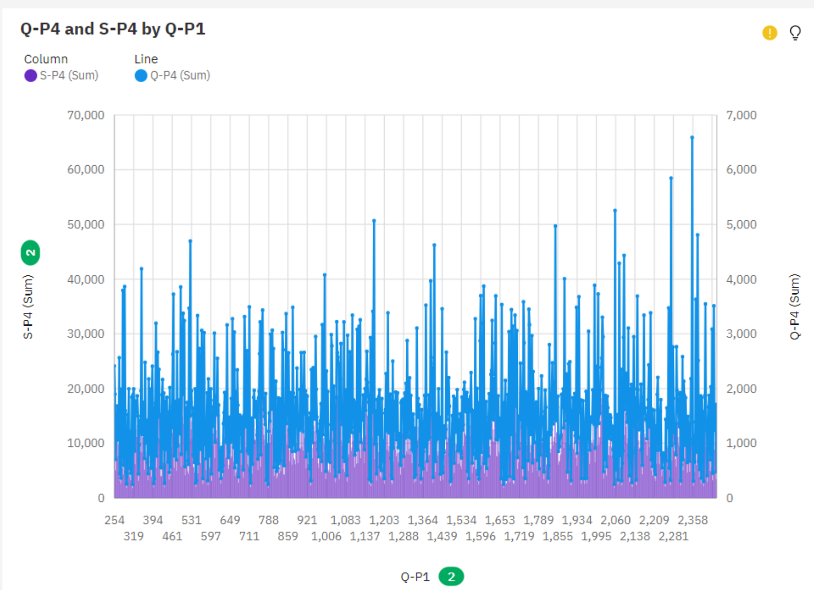






Details

S-P1 predicts **Q-P1** with a strength of 96%

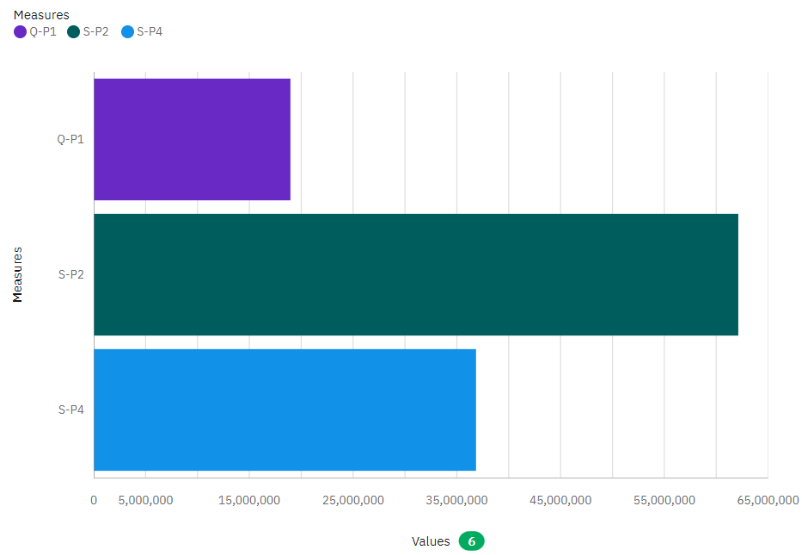


Details

Chart Insights were not computed because this visualization is based on clipped data. Consider applying a filter to reduce the number of records, and to prevent the data from being clipped, before creating the visualization.

Chart Insights were not computed because this visualization is based on clipped data. Consider applying a filter to reduce the number of records, and to prevent the data from being clipped, before creating the visualization.

S-P4, S-P2 and Q-P1



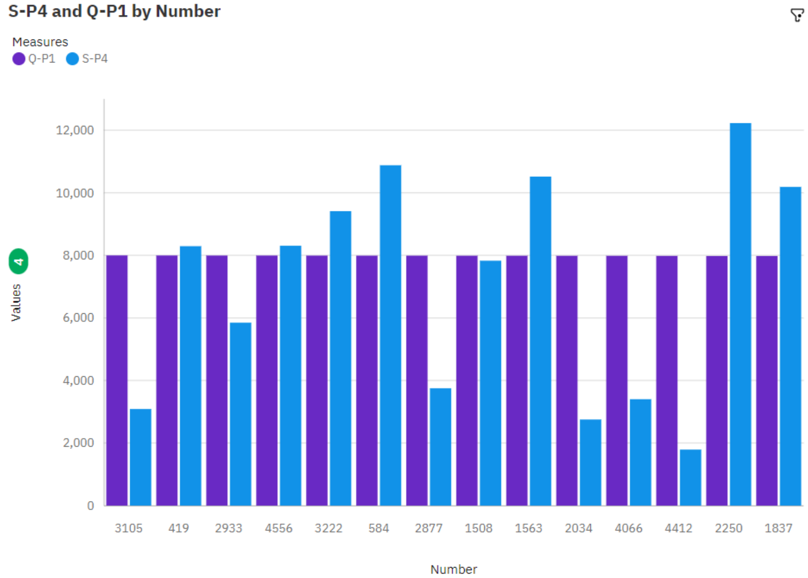
Details

The overall number of results for **S-P4** is over 4500.

The overall number of results for **S-P2** is over 4500.

The overall number of results for **Q-P1** is over 4500.

S-P4 and Q-P1 by Number



Details

The total number of results for **Q-P1**, across all **numbers**, is 14.

Over all **numbers**, the average of **Q-P1** is nearly eight thousand.

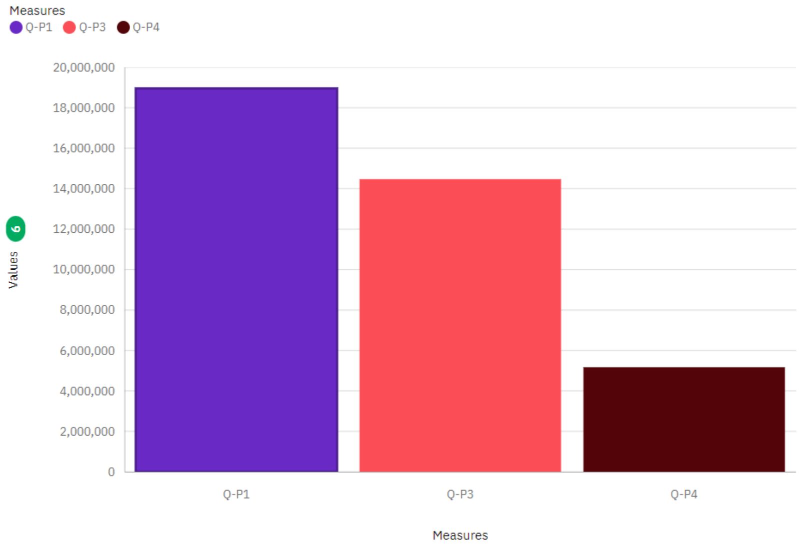
The total number of results for **S-P4**, across all **numbers**, is 14.

Across all **numbers**, the average of **S-P4** is over seven thousand.

Q-P1 ranges from 7979, when **Number** is 2250, to 7998, when **Number** is 3105.

S-P4 ranges from nearly two thousand, when **Number** is 4412, to over twelve thousand, when **Number** is 2250.

Q-P4, Q-P3 and Q-P1



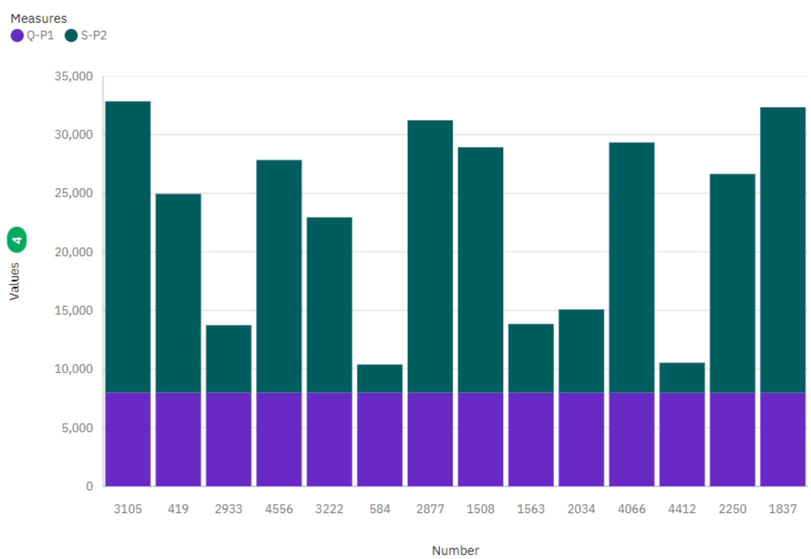
Details

The overall number of results for **Q-P3** is over 4500.

The overall number of results for **Q-P1** is over 4500.

The overall number of results for **Q-P4** is over 4500.

S-P2 and Q-P1 by Number



Details

Q-P1 ranges from 7979, when **Number** is 2250, to 7998, when **Number** is 3105.

S-P2 ranges from almost 2500, when **Number** is 584, to almost 25 thousand, when **Number** is 3105.

The total number of results for **S-P2**, across all **numbers**, is 14.

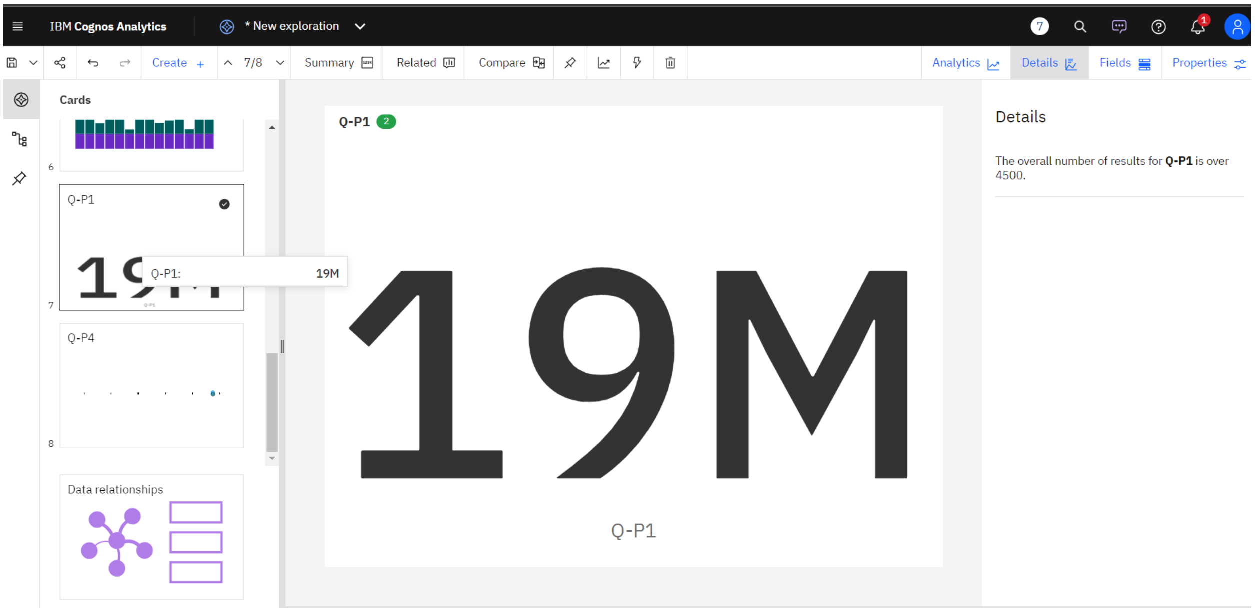
Over all **numbers**, the average of **S-P2** is nearly fifteen thousand.

The total number of results for **Q-P1**, across all **numbers**, is 14.

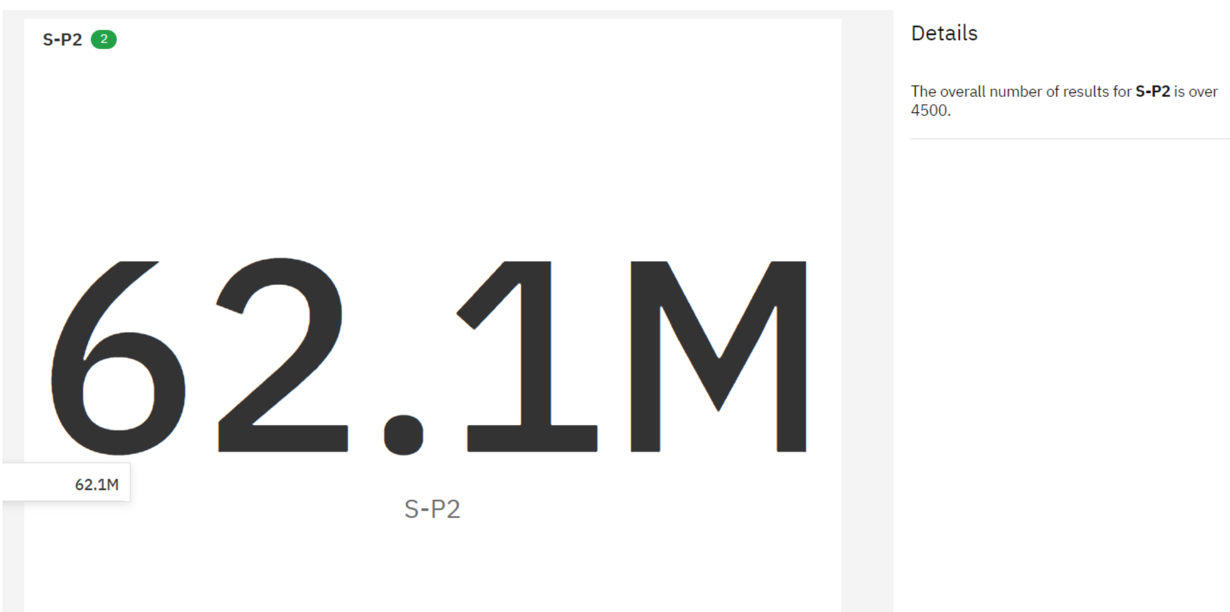
Over all **numbers**, the average of **Q-P1** is nearly eight thousand.

OVERALL SUMMARY IN IBM COGNOS:

Q-P1

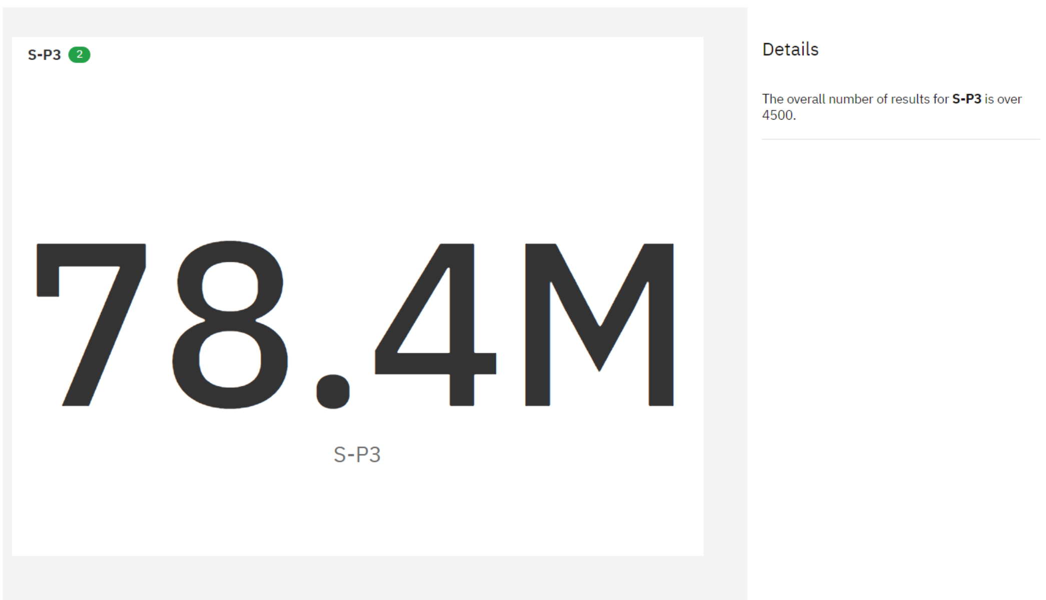


Q-P2



Q-P3

Q-P4



CONCLUSION:

In conclusion, cleaning and preprocessing a dataset are essential steps in the data analysis and machine learning process. These steps help ensure that your data is accurate, consistent, and ready for analysis or modeling. Clean, well-preprocessed data is the foundation for meaningful and actionable insights