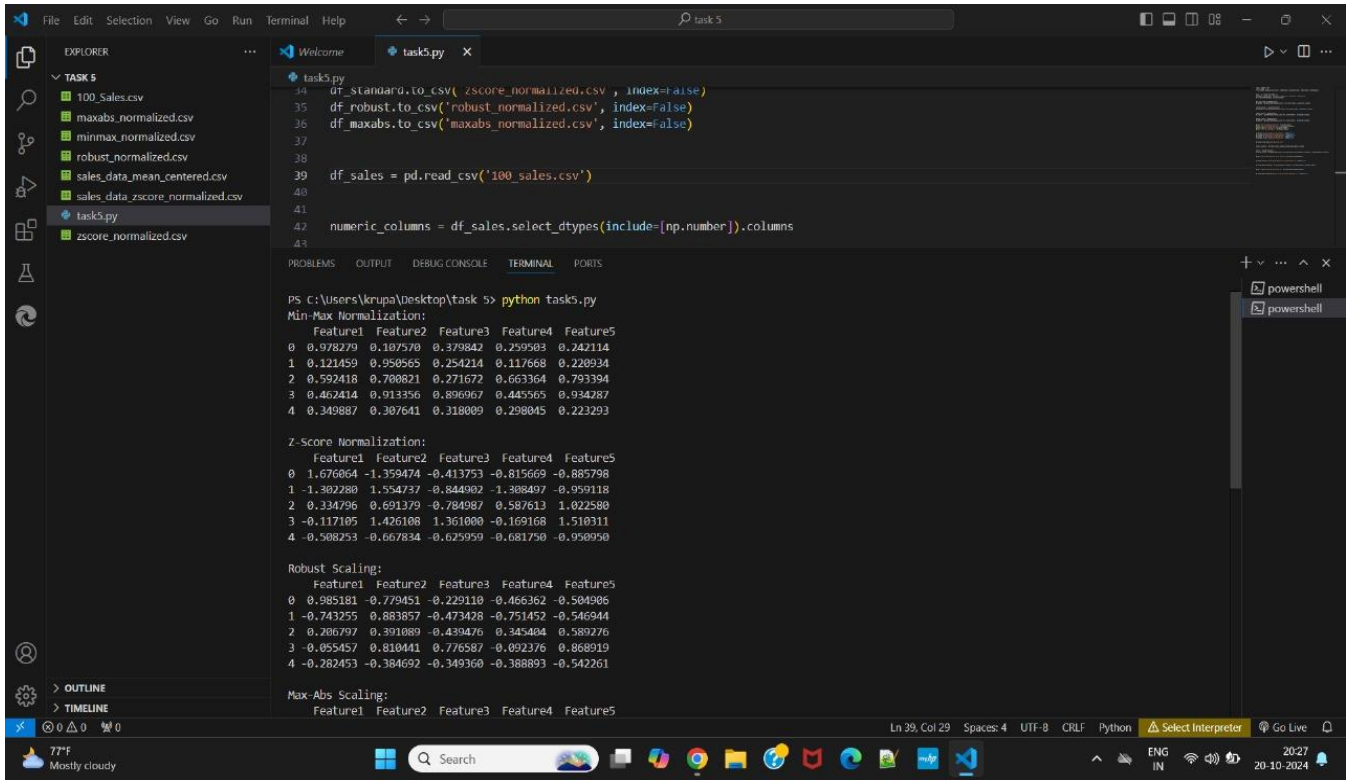
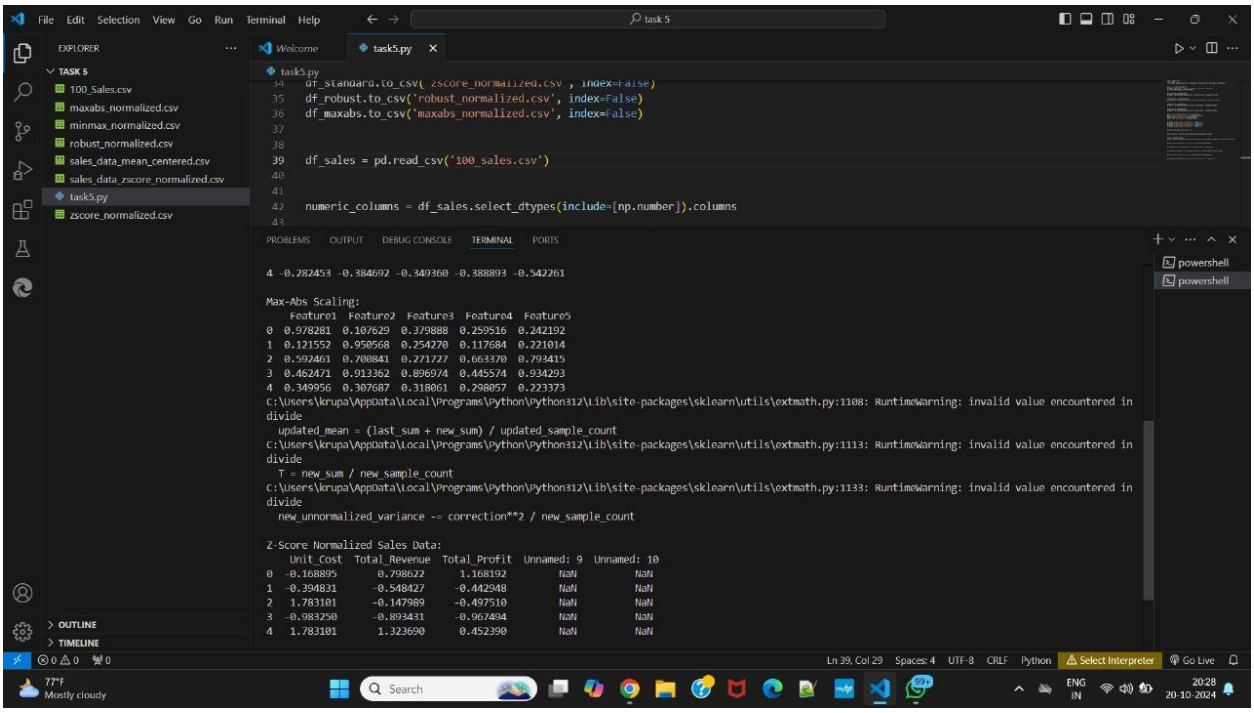
1. Task Description

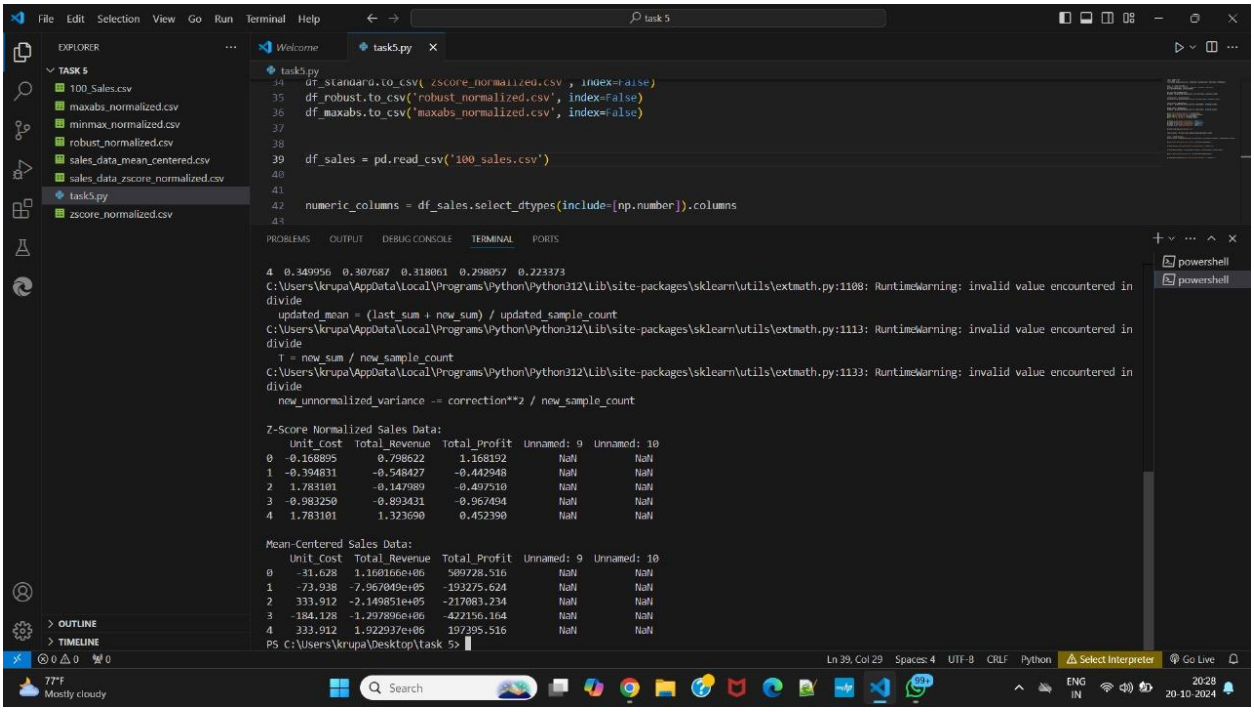
Task : Data Normalization Techniques and Mean-Centering with Z-Score

(A)(Task for normalization and z-score) Find 10000 row dataset and apply all type of normalization in dataset . 20(B) Apply z-score using sklearn library and do mean-centering of Sales dataset(https://www.kaggle.com/datasets/nishathakkar/100-sales)

2. Task Output Screenshot




File Edit Selection View Go Run Terminal Help
task 5
EXPLORER
TASK 5
100_Sales.csv
maxabs_normalized.csv
minmax_normalized.csv
robust_normalized.csv
sales_data_mean_centered.csv
sales_data_zscore_normalized.csv
task5.py
zscore_normalized.csv
Welcome
task5.py
df_standard.to_csv('zscore_normalized.csv', index=False)
df_robust.to_csv('robust_normalized.csv', index=False)
df_maxabs.to_csv('maxabs_normalized.csv', index=False)
df_sales = pd.read_csv('100_sales.csv')
numeric_columns = df_sales.select_dtypes(include=[np.number]).columns
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
4 -0.282453 -0.384692 -0.340360 -0.388893 -0.542261
Max-Abs Scaling:
Feature1 Feature2 Feature3 Feature4 Feature5
0 0.978281 0.107629 0.379888 0.259516 0.242192
1 0.121552 0.950568 0.254270 0.117684 0.221014
2 0.592461 0.700841 0.271727 0.663370 0.793415
3 0.462471 0.913362 0.896974 0.445574 0.934293
4 0.349956 0.307687 0.318061 0.298057 0.223373
C:\Users\krupa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\utils\extmath.py:1108: RuntimeWarning: invalid value encountered in divide
updated_mean = (last_sum + new_sum) / updated_sample_count
C:\Users\krupa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\utils\extmath.py:1113: RuntimeWarning: invalid value encountered in divide
T = new_sum / new_sample_count
C:\Users\krupa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\utils\extmath.py:1133: RuntimeWarning: invalid value encountered in divide
new_unnormalized_variance -= correction**2 / new_sample_count
Z-Score Normalized Sales Data:
Unit_Cost Total_Revenue Total_Profit Unnamed: 9 Unnamed: 10
0 -0.168895 0.798622 1.168192 NaN NaN
1 -0.394831 -0.548427 -0.442948 NaN NaN
2 1.783101 -0.147989 -0.497510 NaN NaN
3 -0.983250 -0.893431 -0.967494 NaN NaN
4 1.783101 1.323690 0.452390 NaN NaN
OUTLINE
TIMELINE
Ln 39, Col 29 Spaces: 4 UTF-8 CRLF Python Select Interpreter Go Live
77°F Mostly cloudy
Search
ENG IN
20:28 20-10-2024


File Edit Selection View Go Run Terminal Help
task 5
EXPLORER
TASK 5
100_Sales.csv
maxabs_normalized.csv
minmax_normalized.csv
robust_normalized.csv
sales_data_mean_centered.csv
sales_data_zscore_normalized.csv
task5.py
zscore_normalized.csv
Welcome
task5.py
df_standard.to_csv('zscore_normalized.csv', index=False)
df_robust.to_csv('robust_normalized.csv', index=False)
df_maxabs.to_csv('maxabs_normalized.csv', index=False)
df_sales = pd.read_csv('100_sales.csv')
numeric_columns = df_sales.select_dtypes(include=[np.number]).columns
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
4 0.349956 0.307687 0.318061 0.298057 0.223373
C:\Users\krupa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\utils\extmath.py:1108: RuntimeWarning: invalid value encountered in divide
updated_mean = (last_sum + new_sum) / updated_sample_count
C:\Users\krupa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\utils\extmath.py:1113: RuntimeWarning: invalid value encountered in divide
T = new_sum / new_sample_count
C:\Users\krupa\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\utils\extmath.py:1133: RuntimeWarning: invalid value encountered in divide
new_unnormalized_variance -= correction**2 / new_sample_count
Z-Score Normalized Sales Data:
Unit_Cost Total_Revenue Total_Profit Unnamed: 9 Unnamed: 10
0 -0.168895 0.798622 1.168192 NaN NaN
1 -0.394831 -0.548427 -0.442948 NaN NaN
2 1.783101 -0.147989 -0.497510 NaN NaN
3 -0.983250 -0.893431 -0.967494 NaN NaN
4 1.783101 1.323690 0.452390 NaN NaN
Mean-Centered Sales Data:
Unit_Cost Total_Revenue Total_Profit Unnamed: 9 Unnamed: 10
0 -31.628 1.160166e+06 589728.516 NaN NaN
1 -73.938 -7.967049e+05 -193275.624 NaN NaN
2 333.912 -2.149851e+05 -217083.234 NaN NaN
3 -184.128 -1.297896e+06 -422156.164 NaN NaN
4 333.912 1.922937e+06 197395.516 NaN NaN
PS C:\Users\krupa\Desktop\task 5>
OUTLINE
TIMELINE
Ln 39, Col 29 Spaces: 4 UTF-8 CRLF Python Select Interpreter Go Live
77°F Mostly cloudy
Search
ENG IN
20:28 20-10-2024

3) Algorithm Used In Task

1) **Min-Max Normalization:**

   Algorithm: This technique scales the data to a fixed range, typically between 0 and 1. It preserves the relationships between values while ensuring they all fit within a common scale.

2) **Z-Score Normalization (Standardization):**

   Algorithm: Standardization (Z-Score normalization) scales data to have a mean of 0 and a standard deviation of 1, which helps remove the effects of scale from the data.

3) **Robust Scaling:**

   Algorithm: This algorithm scales the data using the **median** and the **interquartile range (IQR)**, making it less sensitive to outliers compared to Min-Max scaling

4) **Max-Abs Scaling:**

   Algorithm: This algorithm scales data based on the **absolute maximum value** in each feature, transforming the data such that the maximum absolute value is 1. This is useful when the data contains both positive and negative values.

5) **Mean-Centering:**

   Algorithm: Mean-centering is the process of subtracting the mean of each feature from the values, shifting the dataset such that the mean of each feature becomes 0. This is not technically normalization but helps in some machine learning algorithms