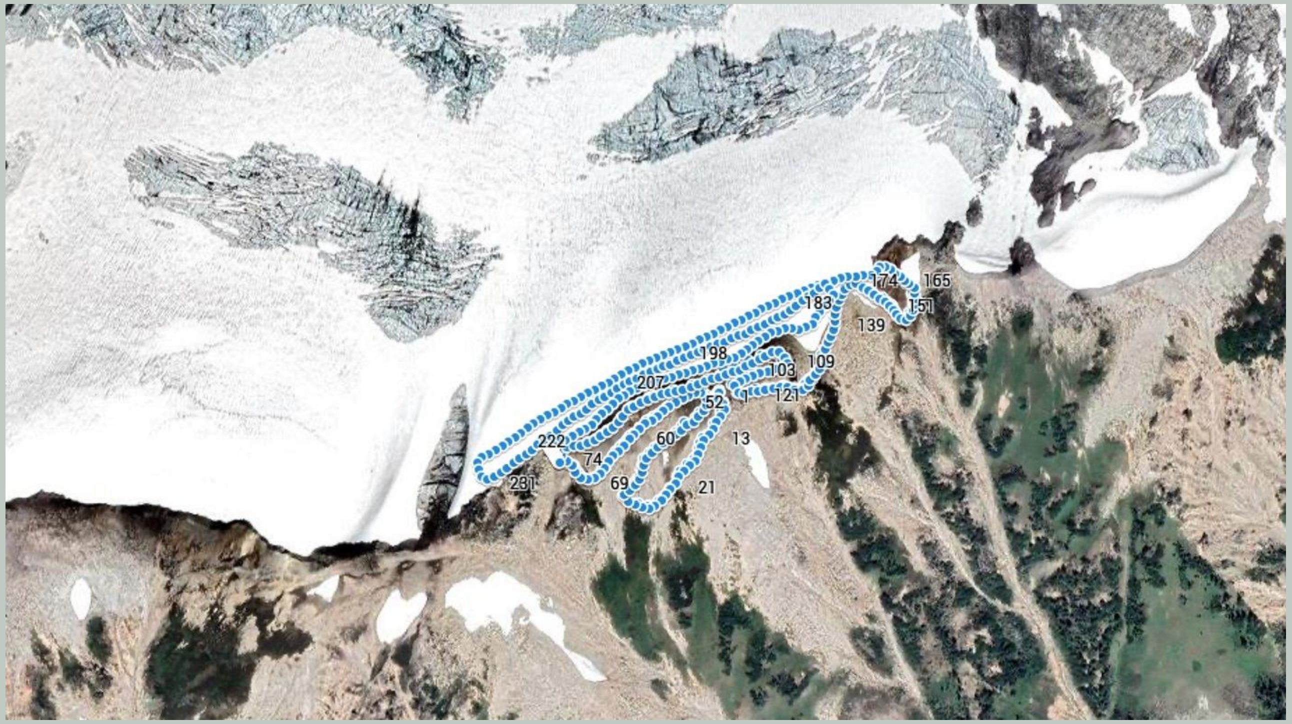
MoMacMo Data Analysis Project



Mk Maharana

May – August 2023

Data Science Intern

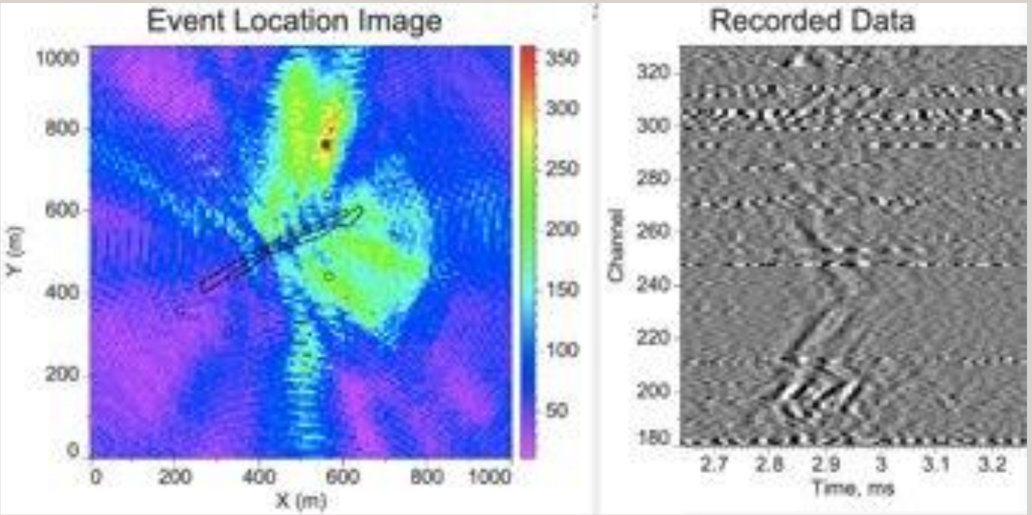
Company Goals

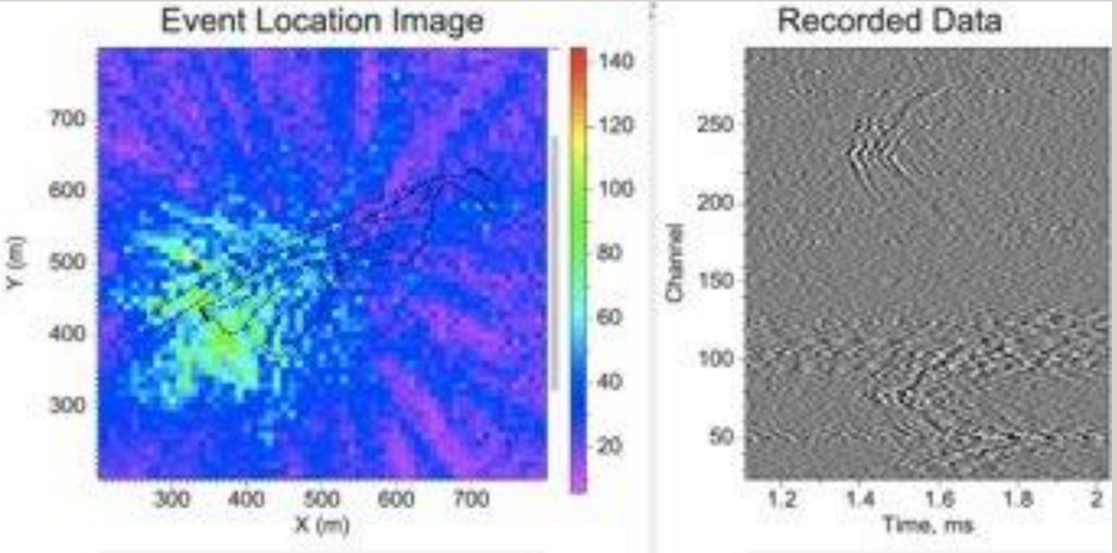
- Apply clustering analysis to microseismic events from the Meager DAS Experiment
- Design a Labeling Scheme to allow machine learning to predict a more efficient event location assignment
- How can we use machine learning and classification to locate events?
- Establish a labeling methodology that captures the geographic distribution of events

01 Introduction

The company is in the process of manually locating micro-seismic events using earthquake epicenter location concepts.

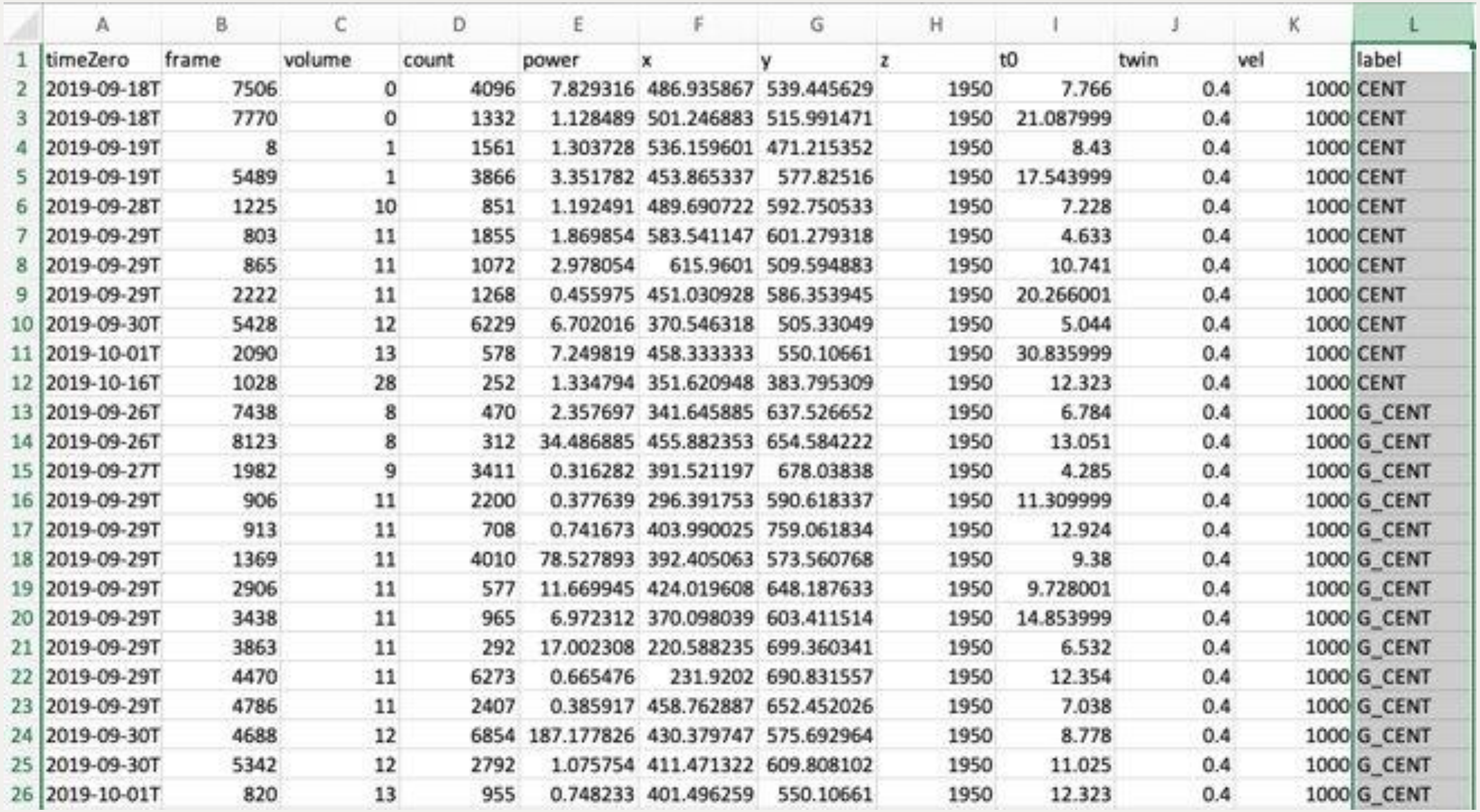
The company learned that the events are focused in particular geographic areas.





02 Original Data

Spreadsheet with Picked Events



	A	B	C	D	E	F	G	H	I	J	K	L
1	timeZero	frame	volume	count	power	x	y	z	t0	twin	vel	label
2	2019-09-18T	7506	0	4096	7.829316	486.935867	539.445629	1950	7.766	0.4	1000	CENT
3	2019-09-18T	7770	0	1332	1.128489	501.246883	515.991471	1950	21.087999	0.4	1000	CENT
4	2019-09-19T	8	1	1561	1.303728	536.159601	471.215352	1950	8.43	0.4	1000	CENT
5	2019-09-19T	5489	1	3866	3.351782	453.865337	577.82516	1950	17.543999	0.4	1000	CENT
6	2019-09-28T	1225	10	851	1.192491	489.690722	592.750533	1950	7.228	0.4	1000	CENT
7	2019-09-29T	803	11	1855	1.869854	583.541147	601.279318	1950	4.633	0.4	1000	CENT
8	2019-09-29T	865	11	1072	2.978054	615.9601	509.594883	1950	10.741	0.4	1000	CENT
9	2019-09-29T	2222	11	1268	0.455975	451.030928	586.353945	1950	20.266001	0.4	1000	CENT
10	2019-09-30T	5428	12	6229	6.702016	370.546318	505.33049	1950	5.044	0.4	1000	CENT
11	2019-10-01T	2090	13	578	7.249819	458.333333	550.10661	1950	30.835999	0.4	1000	CENT
12	2019-10-16T	1028	28	252	1.334794	351.620948	383.795309	1950	12.323	0.4	1000	CENT
13	2019-09-26T	7438	8	470	2.357697	341.645885	637.526652	1950	6.784	0.4	1000	G_CENT
14	2019-09-26T	8123	8	312	34.486885	455.882353	654.584222	1950	13.051	0.4	1000	G_CENT
15	2019-09-27T	1982	9	3411	0.316282	391.521197	678.03838	1950	4.285	0.4	1000	G_CENT
16	2019-09-29T	906	11	2200	0.377639	296.391753	590.618337	1950	11.309999	0.4	1000	G_CENT
17	2019-09-29T	913	11	708	0.741673	403.990025	759.061834	1950	12.924	0.4	1000	G_CENT
18	2019-09-29T	1369	11	4010	78.527893	392.405063	573.560768	1950	9.38	0.4	1000	G_CENT
19	2019-09-29T	2906	11	577	11.669945	424.019608	648.187633	1950	9.728001	0.4	1000	G_CENT
20	2019-09-29T	3438	11	965	6.972312	370.098039	603.411514	1950	14.853999	0.4	1000	G_CENT
21	2019-09-29T	3863	11	292	17.002308	220.588235	699.360341	1950	6.532	0.4	1000	G_CENT
22	2019-09-29T	4470	11	6273	0.665476	231.9202	690.831557	1950	12.354	0.4	1000	G_CENT
23	2019-09-29T	4786	11	2407	0.385917	458.762887	652.452026	1950	7.038	0.4	1000	G_CENT
24	2019-09-30T	4688	12	6854	187.177826	430.379747	575.692964	1950	8.778	0.4	1000	G_CENT
25	2019-09-30T	5342	12	2792	1.075754	411.471322	609.808102	1950	11.025	0.4	1000	G_CENT
26	2019-10-01T	820	13	955	0.748233	401.496259	550.10661	1950	12.323	0.4	1000	G_CENT

Project Plan

- At the end of the project we will have analysis that tells us:
  - What ML/DA techniques are applicable for clustering ?
  - Are events that we pick clustering in geographic areas of interest ?
  - Can we say anything about the quality of the manually picked events ?
- What do we need to do to accomplish this ?
  - Review and select ML/DA approaches for clustering
  - Apply clustering analysis to the picked data
  - Export cluster model so that it can be used to label the data
  - Apply labels to the picked data and report on results

The Program

- In the end I produced a program in Python that performs k-means clustering on a dataset, plots the clusters in different dimensions, and computes the silhouette score to evaluate the quality of the clustering.

Program Breakdown

- The necessary libraries are imported, including matplotlib, KMeans from sklearn.cluster, StandardScaler, SimpleImputer from sklearn.preprocessing, and silhouette_score from sklearn.metrics.
- The data is loaded from a CSV file named "trainingEventsDistributed.csv" using pandas.
- Missing values in the numeric columns are imputed (filled) with the mean using SimpleImputer.
- Three columns ('x', 'y', and 'power') are selected to be used for clustering.
- The selected columns are standardized to have a mean of 0 and variance of 1 using StandardScaler.
- The Elbow method is used to determine the optimal number of clusters. The within-cluster sum of squares (WCSS) is calculated for different numbers of clusters (ranging from 1 to 10) and plotted. The point where the plot starts to level off is chosen as the optimal number of clusters.

Program Breakdown cont.

- After determining the optimal number of clusters (in this case, 3), the k-means clustering model is created with n_clusters=3 and fit to the standardized data.

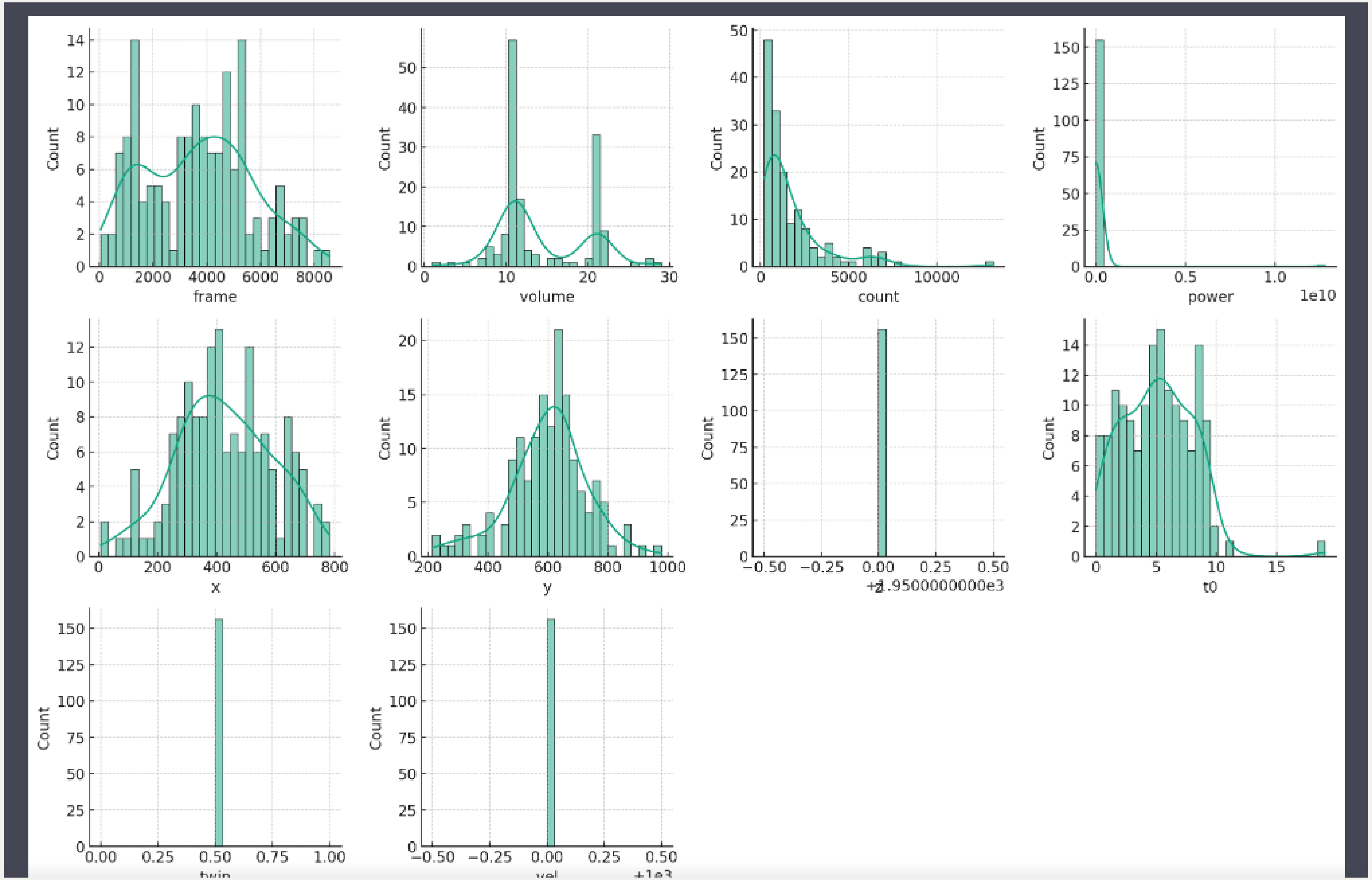
- The cluster labels are assigned to each data point based on the k-means clustering model.

- The cluster labels are added to the original data and saved to a new CSV file named "trainingEventsDistributed_with_clusters.csv".

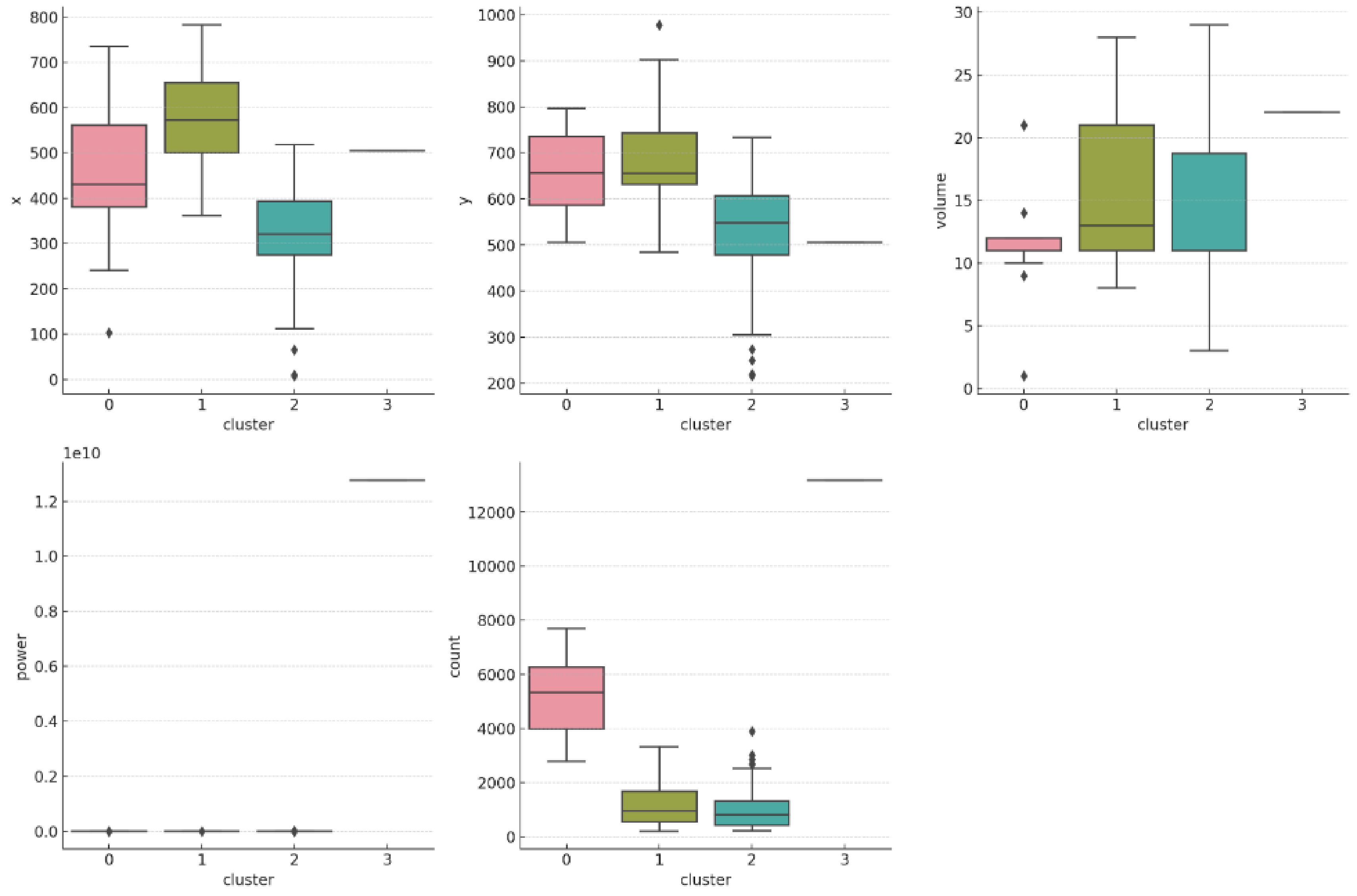
- Three scatter plots are created to visualize the clusters in different dimensions: 'x' vs 'y', 'x' vs 'power', and a 3D plot of 'x', 'y', and 'power'.

- The silhouette score is computed to evaluate the quality of the clustering. The silhouette score measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters.

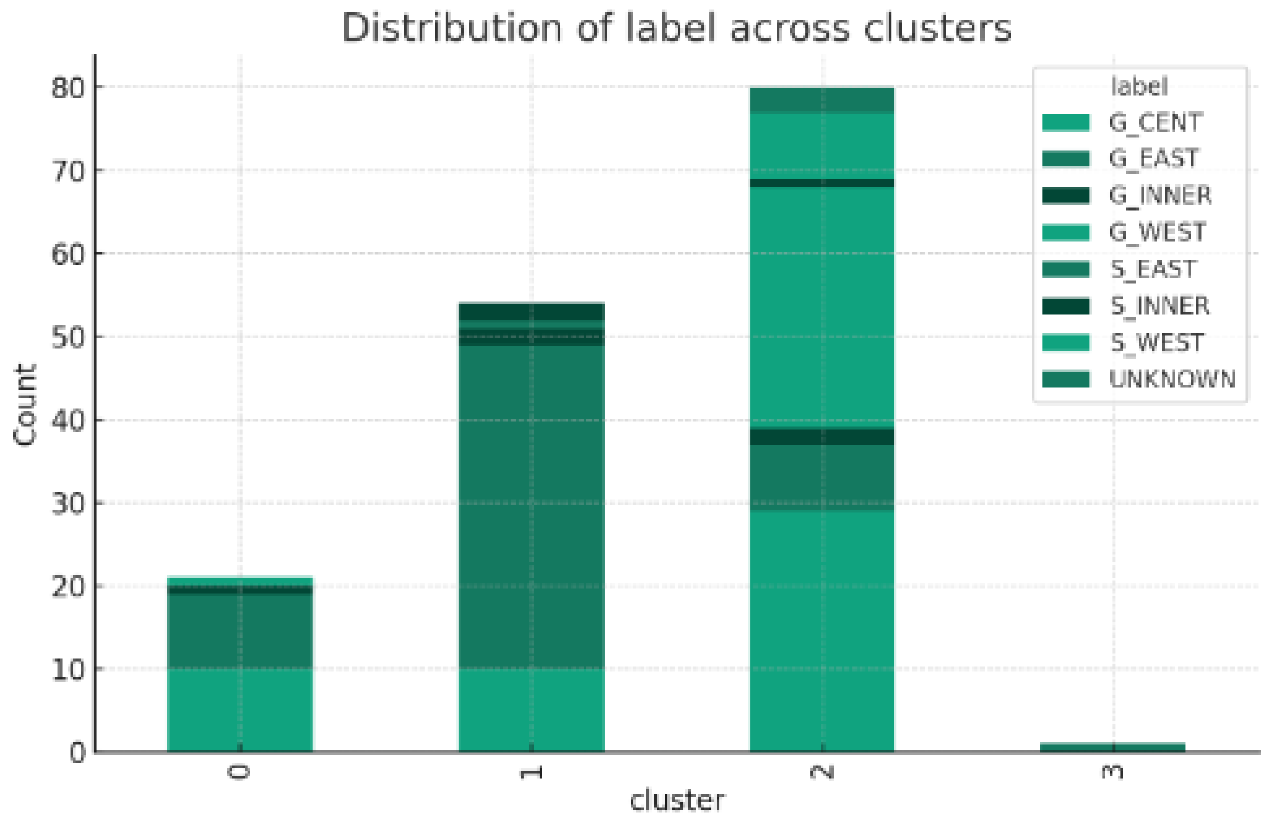
Distribution of Numerical Variables

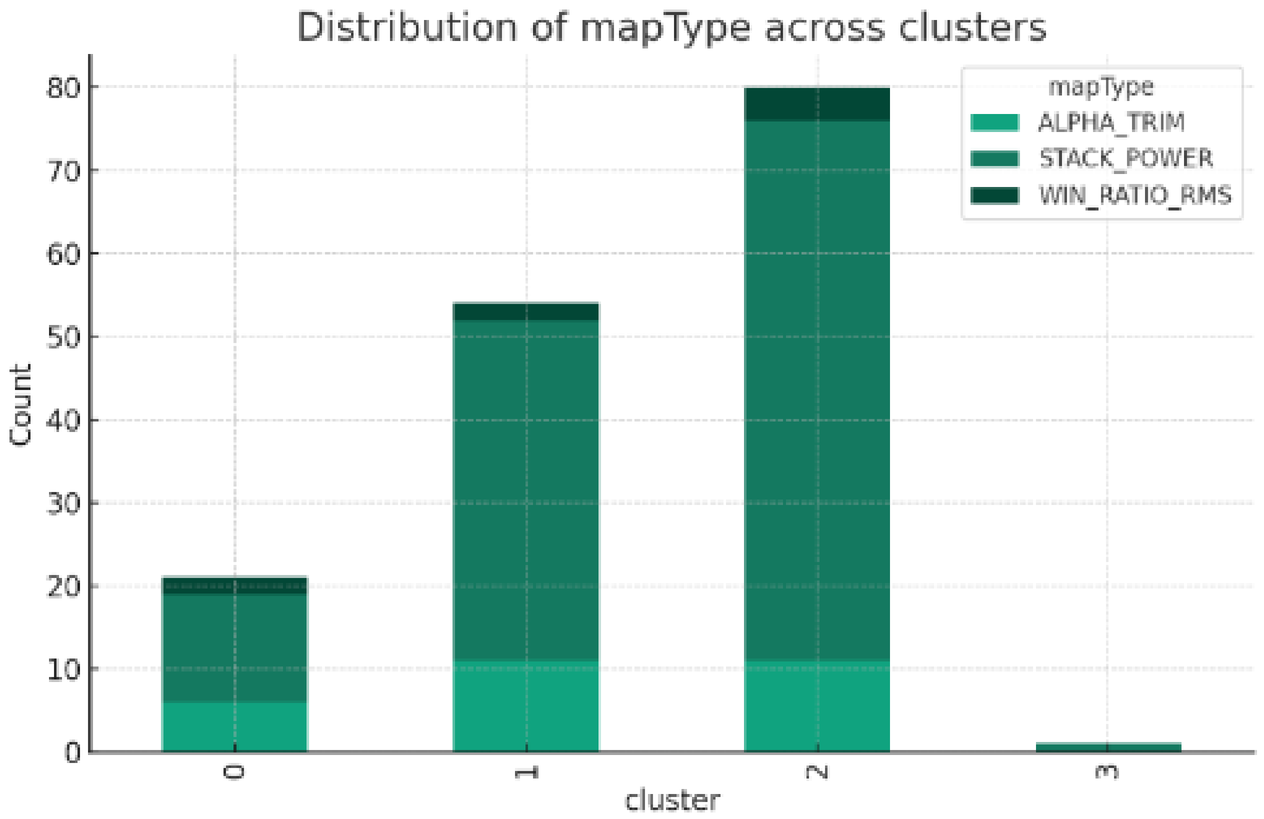


Distribution of x, y, volume, power

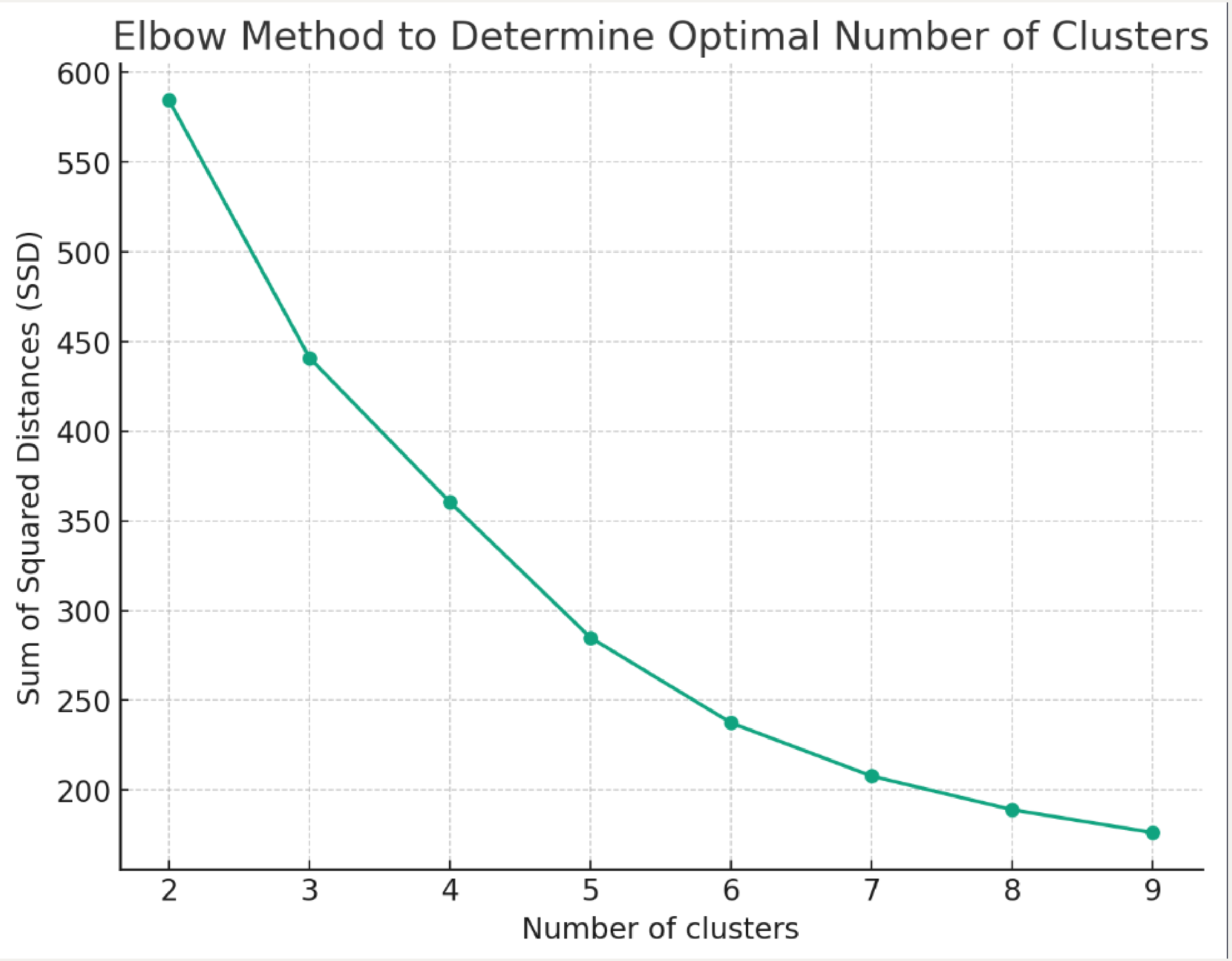


Label Across Clusters





Results


Elbow Method to Determine Optimal Number of Clusters
Sum of Squared Distances (SSD)
600
550
500
450
400
350
300
250
200
2
3
4
5
6
7
8
9
Number of clusters

Results

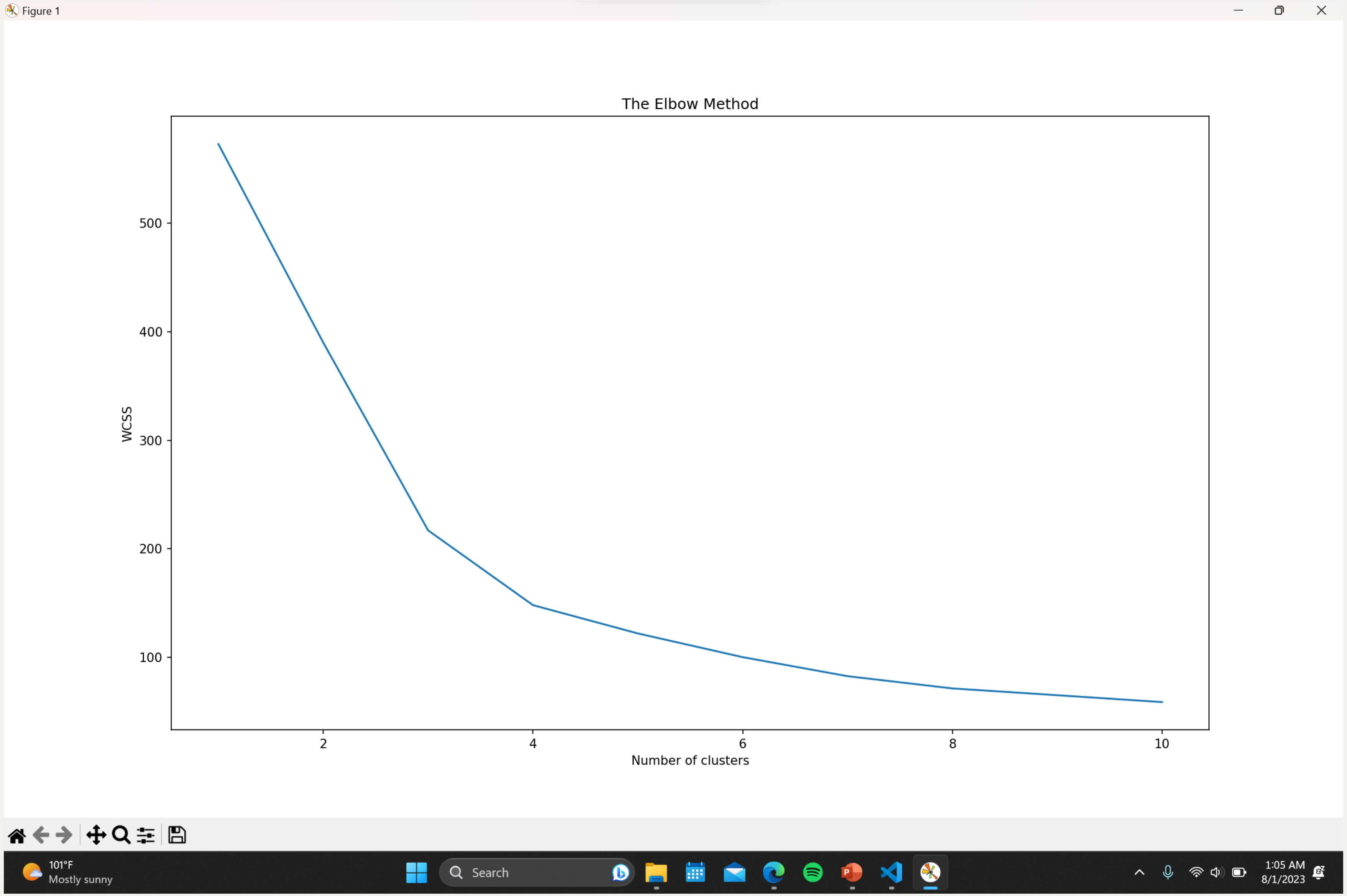


Figure 1
The Elbow Method
WCSS
500
400
300
200
100
2
4
6
8
10
Number of clusters
101°F
Mostly sunny
Search
1:05 AM
8/1/2023

Results

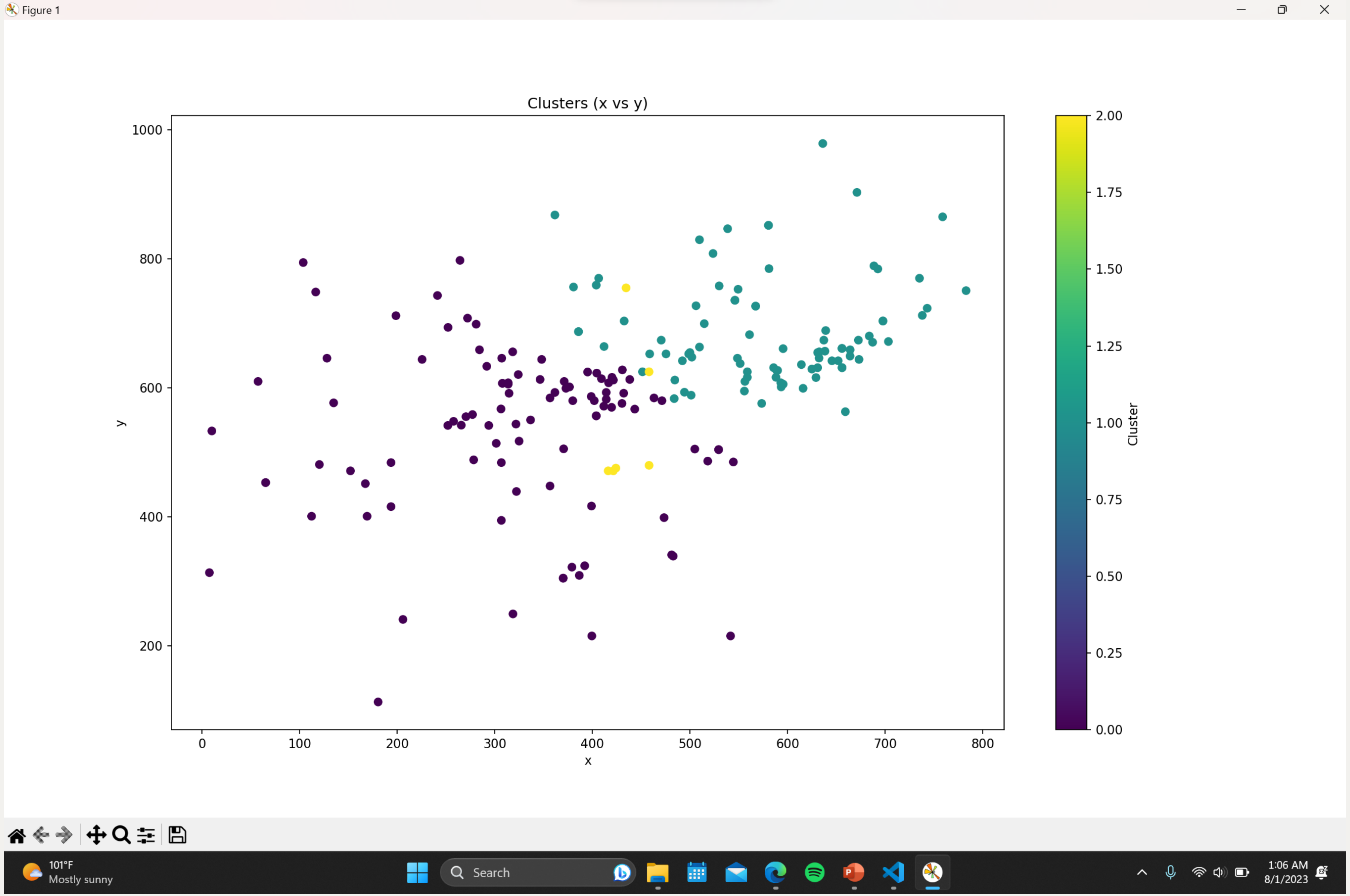


Figure 1
Clusters (x vs y)
1000
800
600
400
200
y
0
100
200
300
400
500
600
700
800
x
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Cluster
101°F
Mostly sunny
Search
1:06 AM
8/1/2023

Results

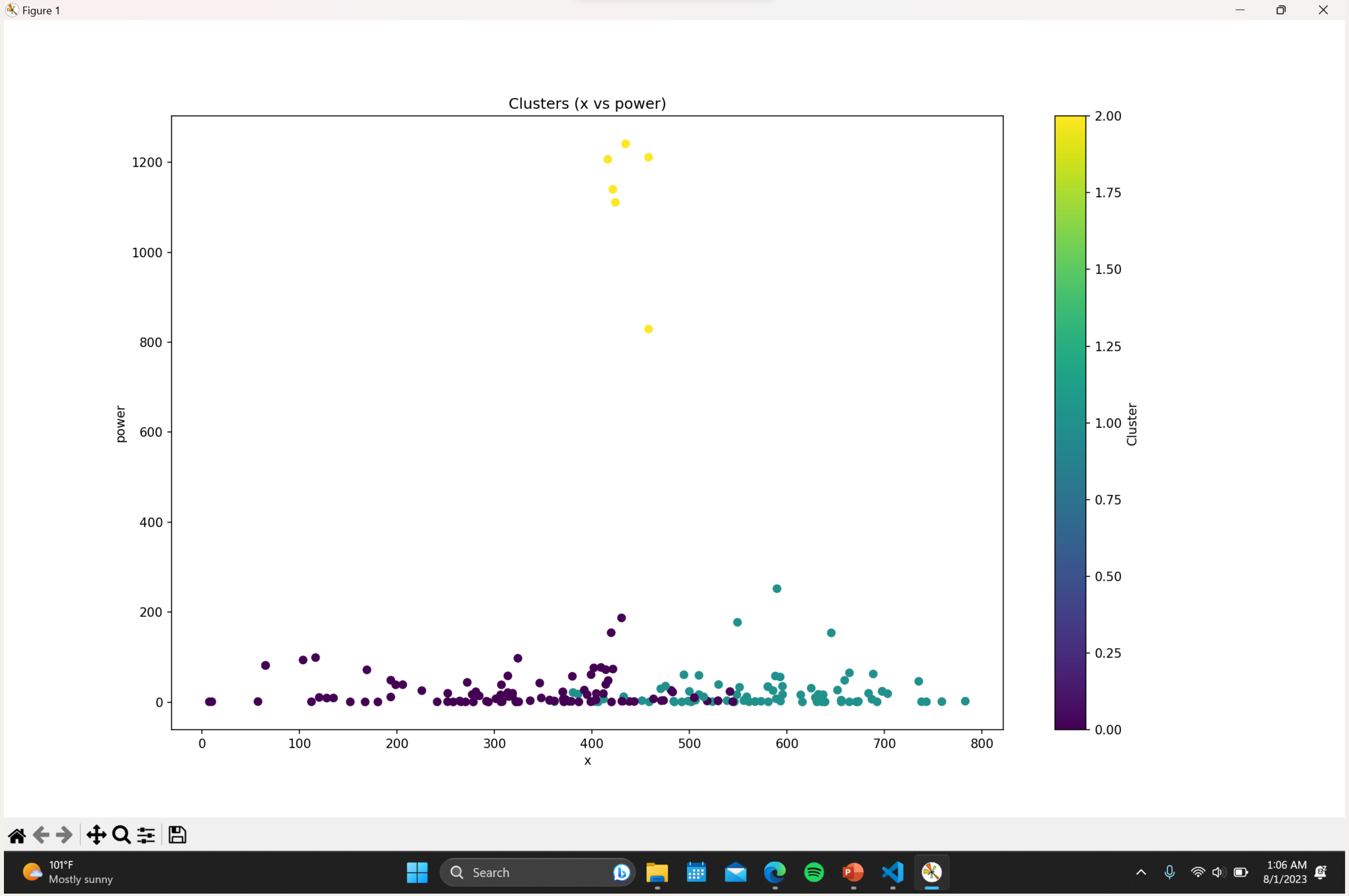
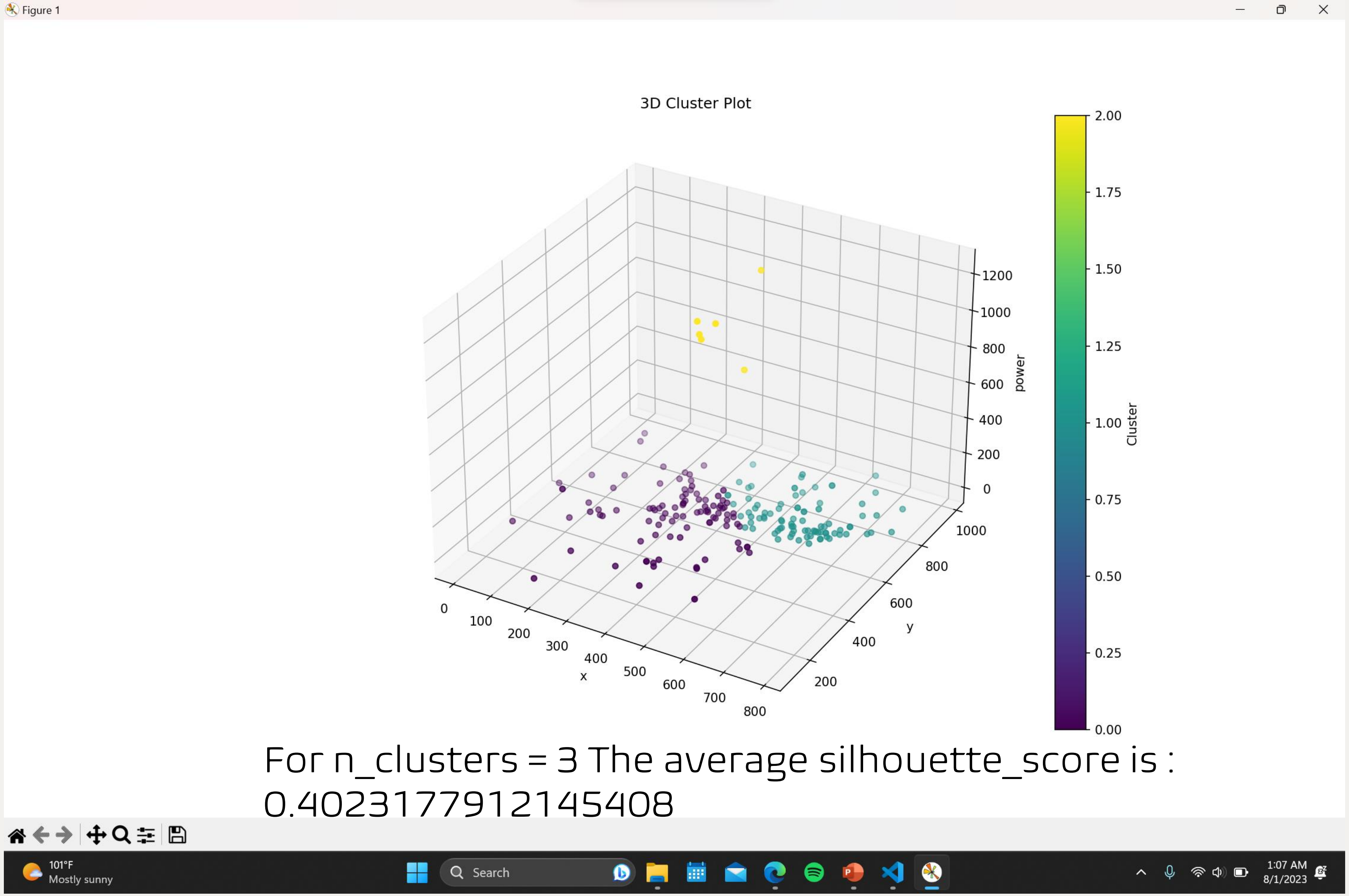


Figure 1
Clusters (x vs power)
power
1200
1000
800
600
400
200
0
0
100
200
300
400
500
600
700
800
x
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Cluster
101°F
Mostly sunny
Search
1:06 AM
8/1/2023

Results

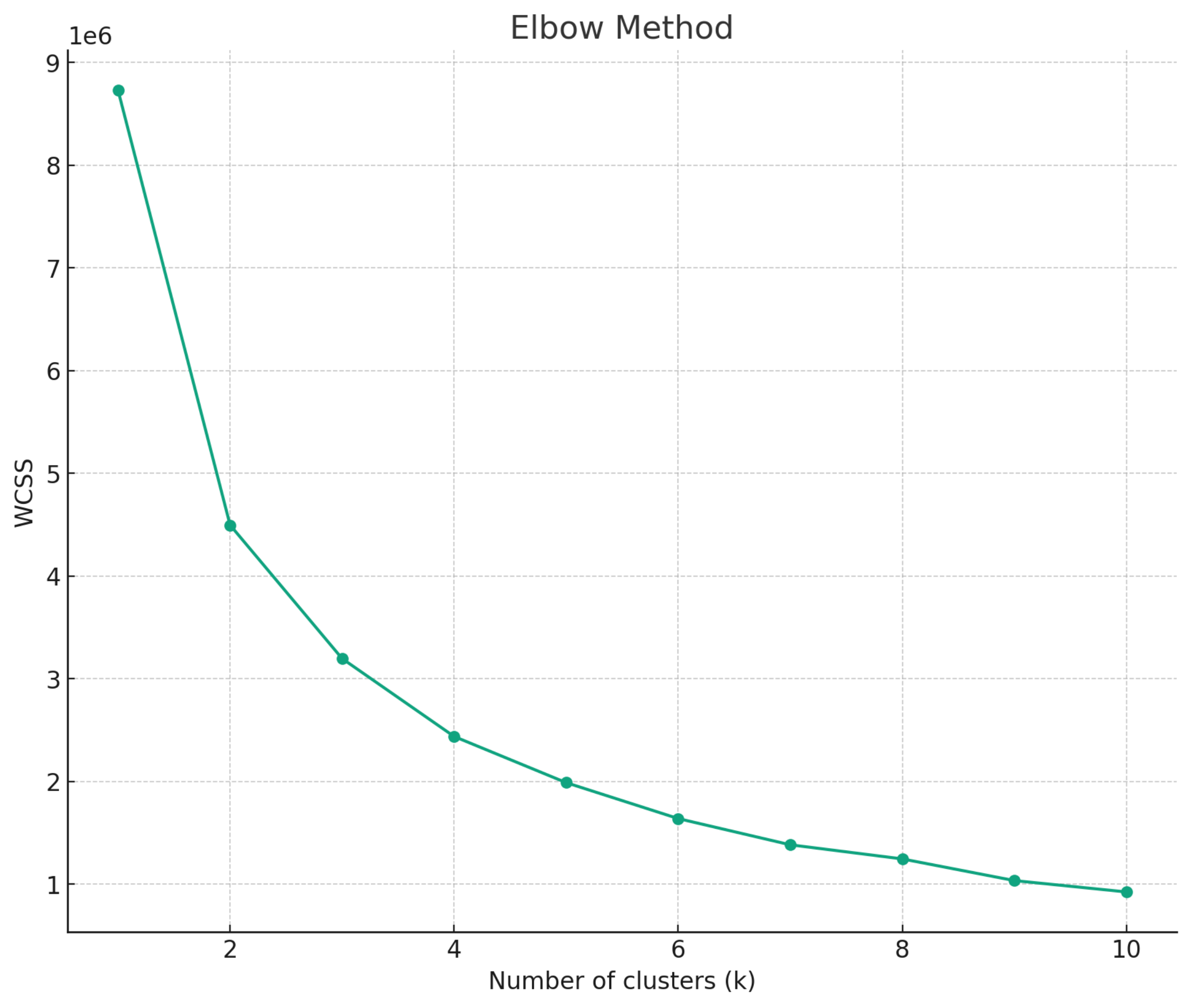


For n_clusters = 3 The average silhouette_score is : 0.4023177912145408

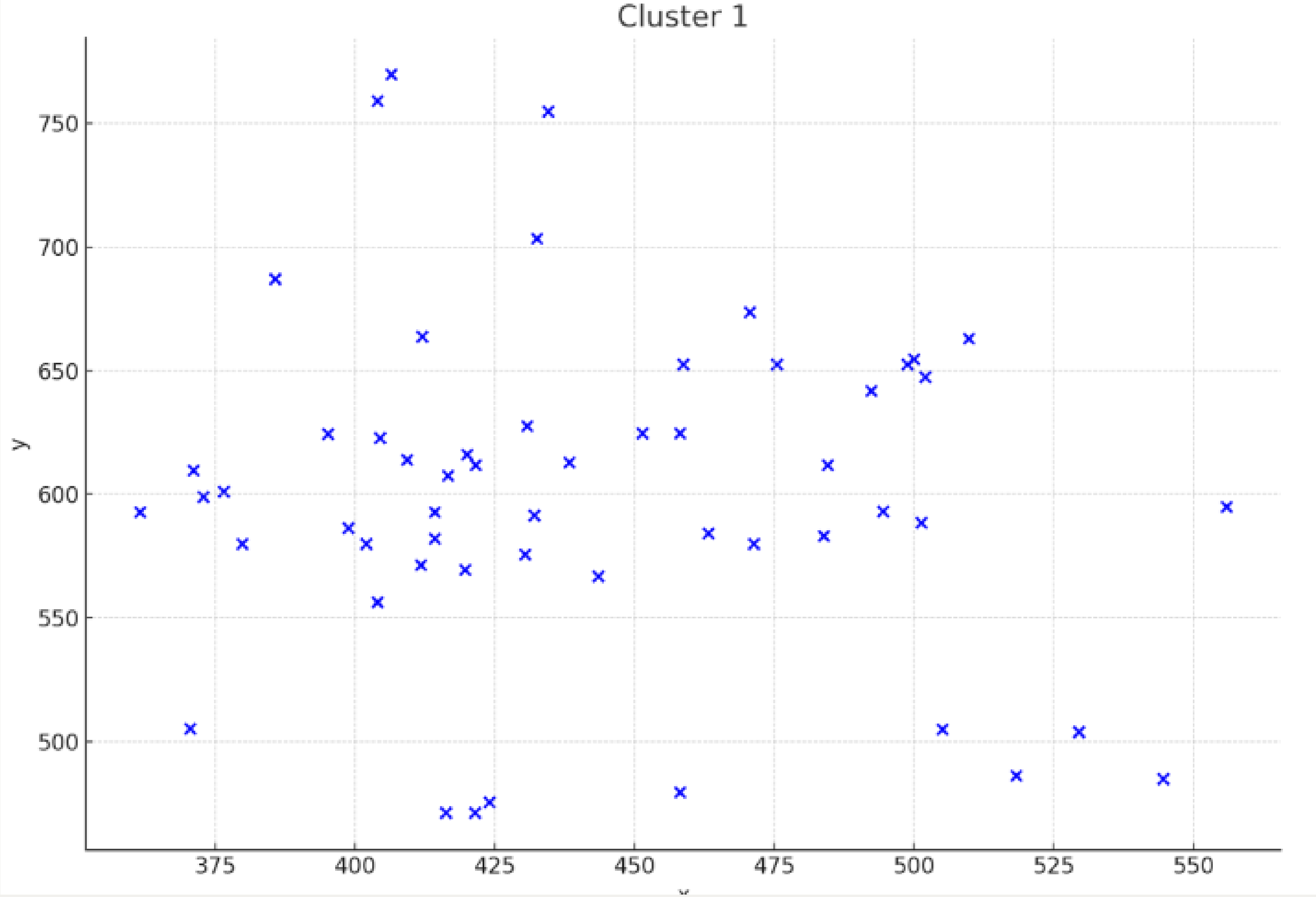
X and y clustering

- Additionally, I developed a program that would perform k-means clustering with 5 clusters on the 'x' and 'y' variables.

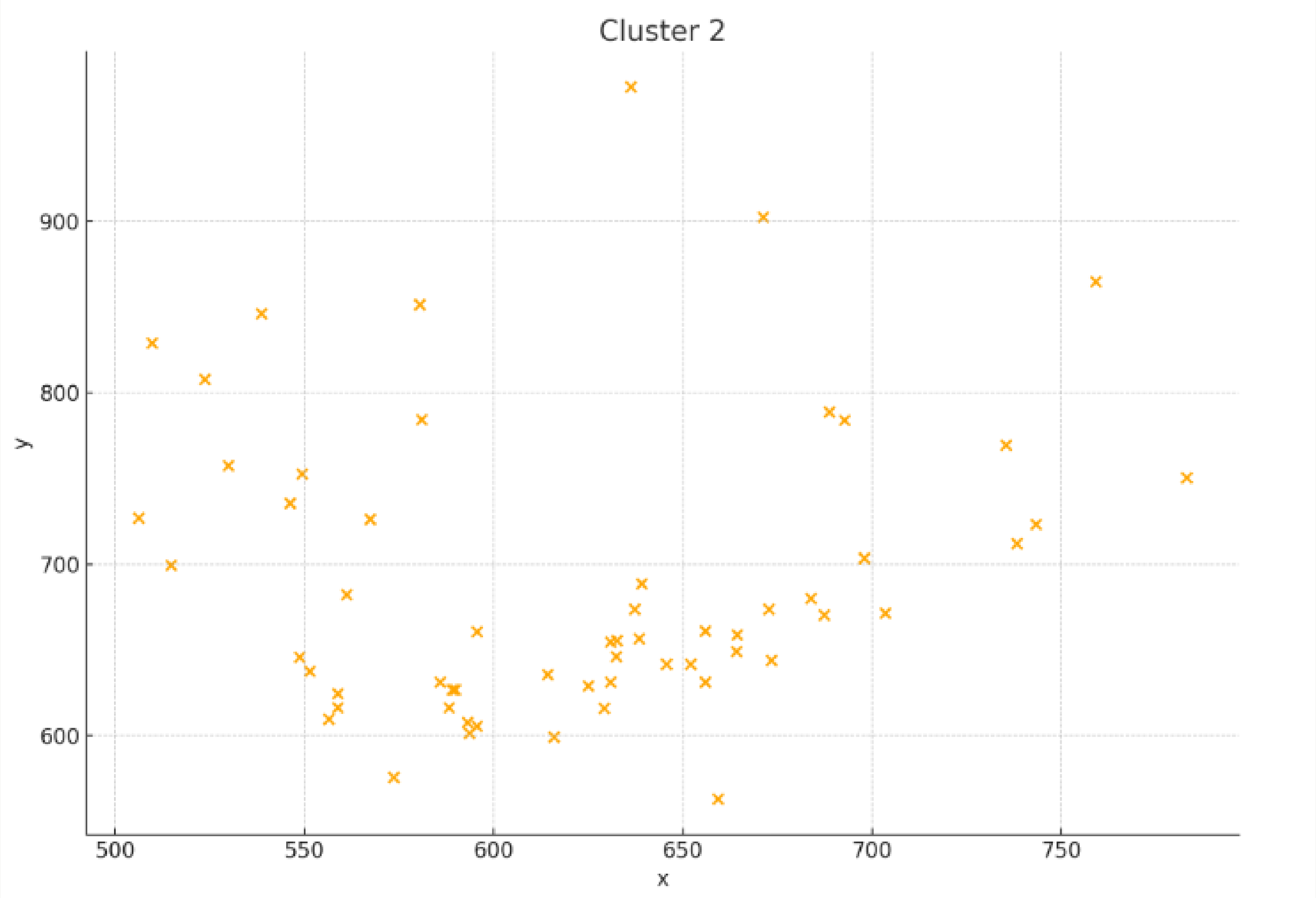
Elbow Plot


Elbow Method
1e6
9
8
7
6
5
4
3
2
1
WCSS
2
4
6
8
10
Number of clusters (k)

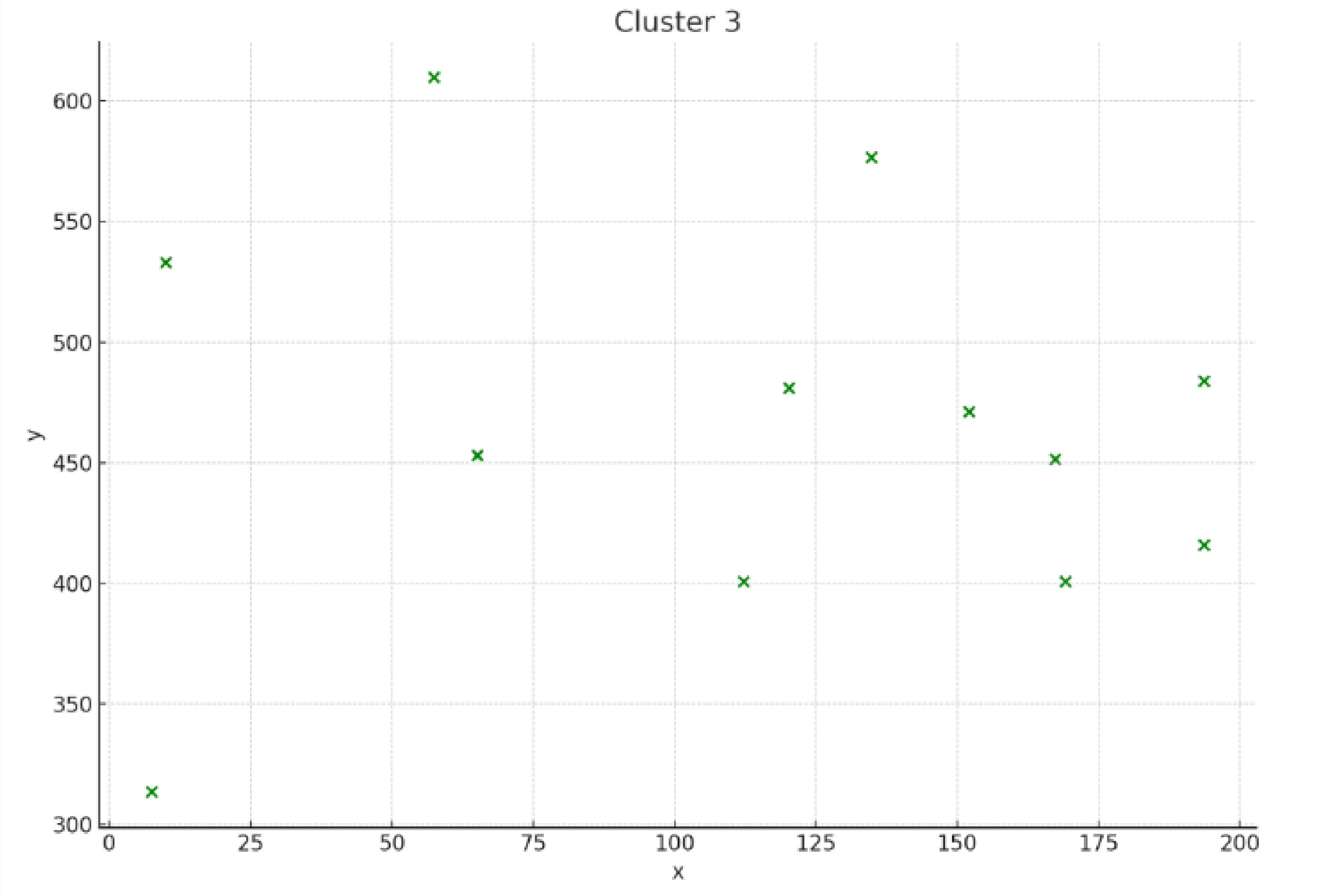
Cluster Data Points: Cluster 1



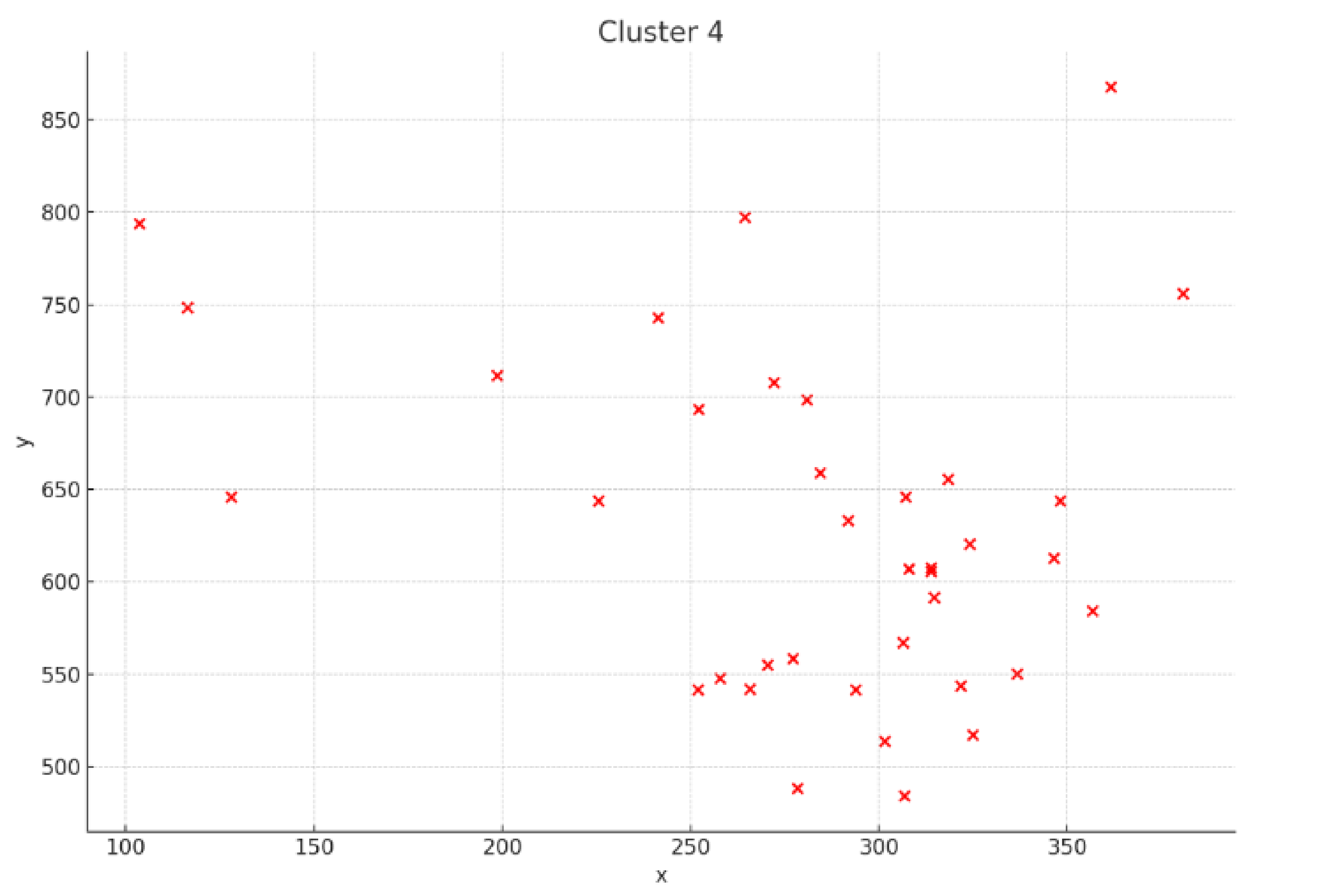
Cluster Data Points: Cluster 2



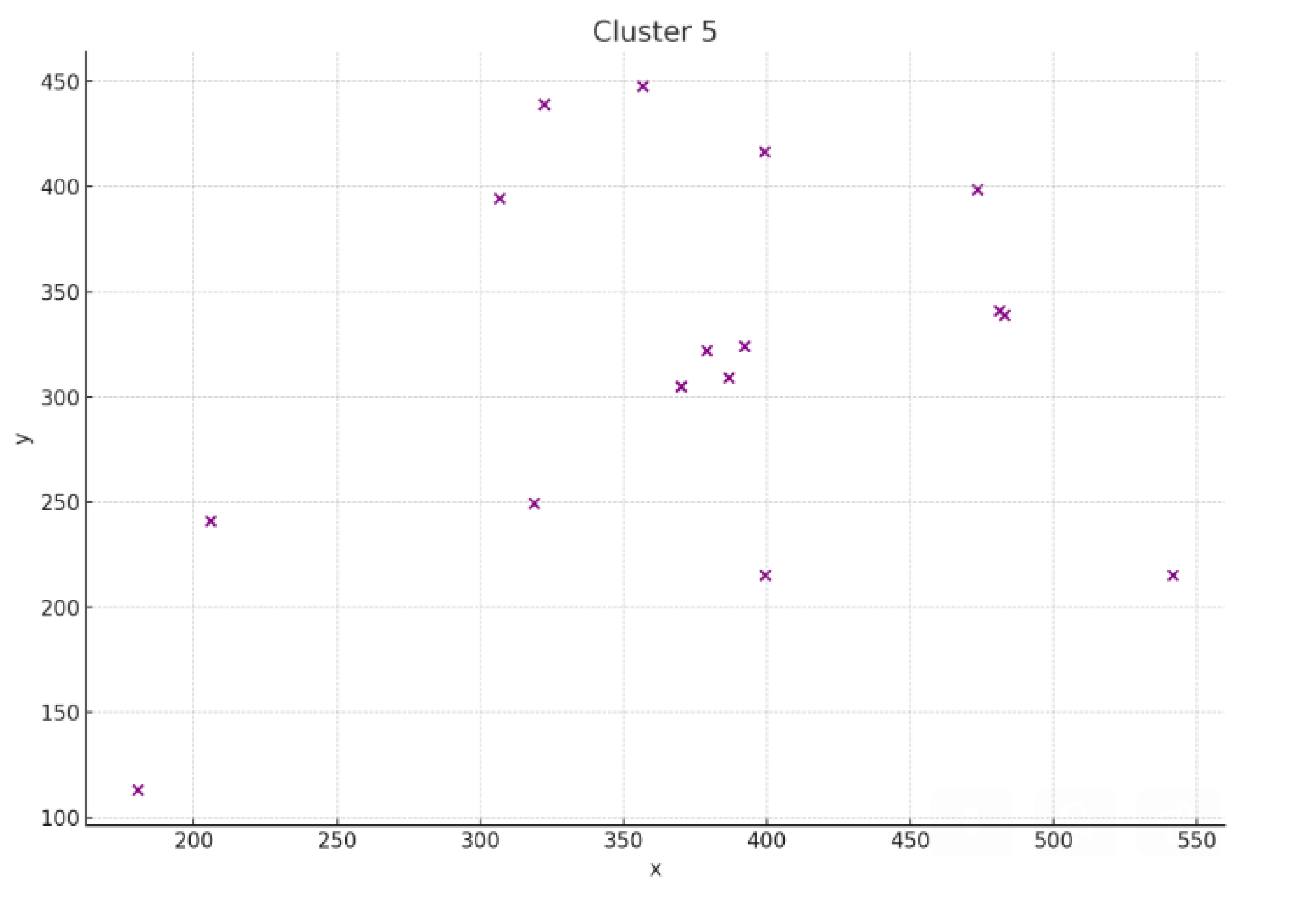
Cluster Data Points: Cluster 3



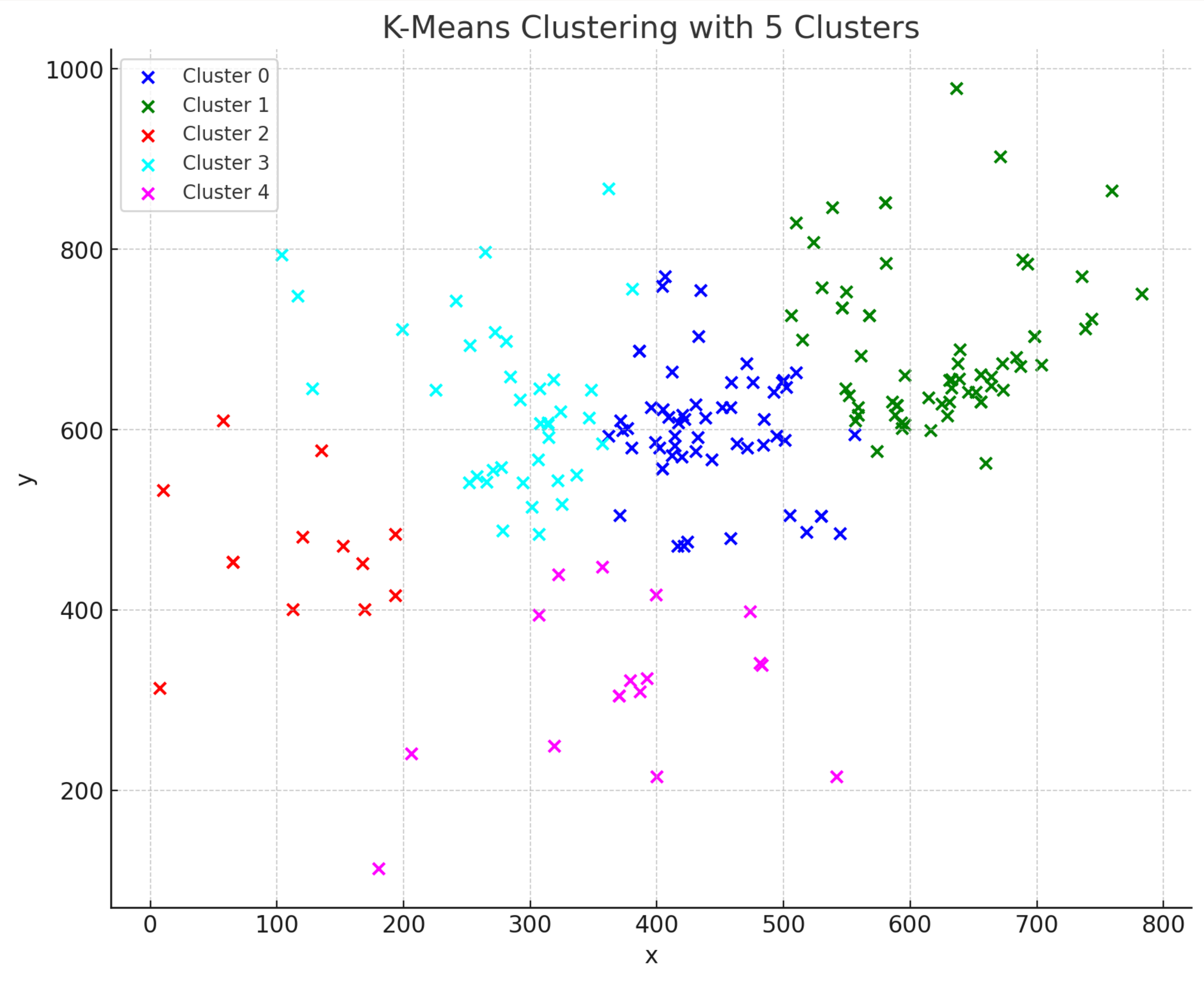
Cluster Data Points: Cluster 4



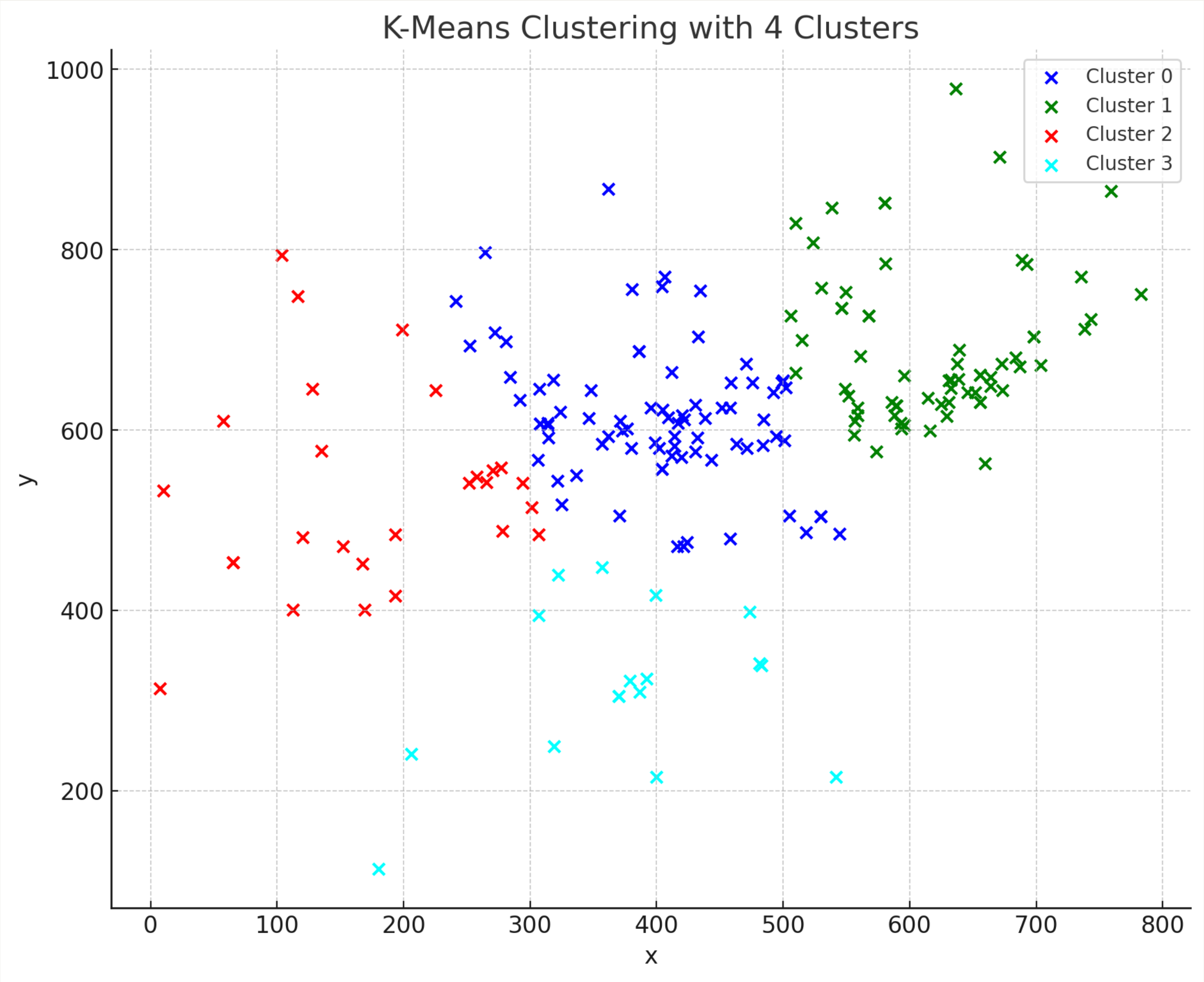
Cluster Data Points: Cluster 5



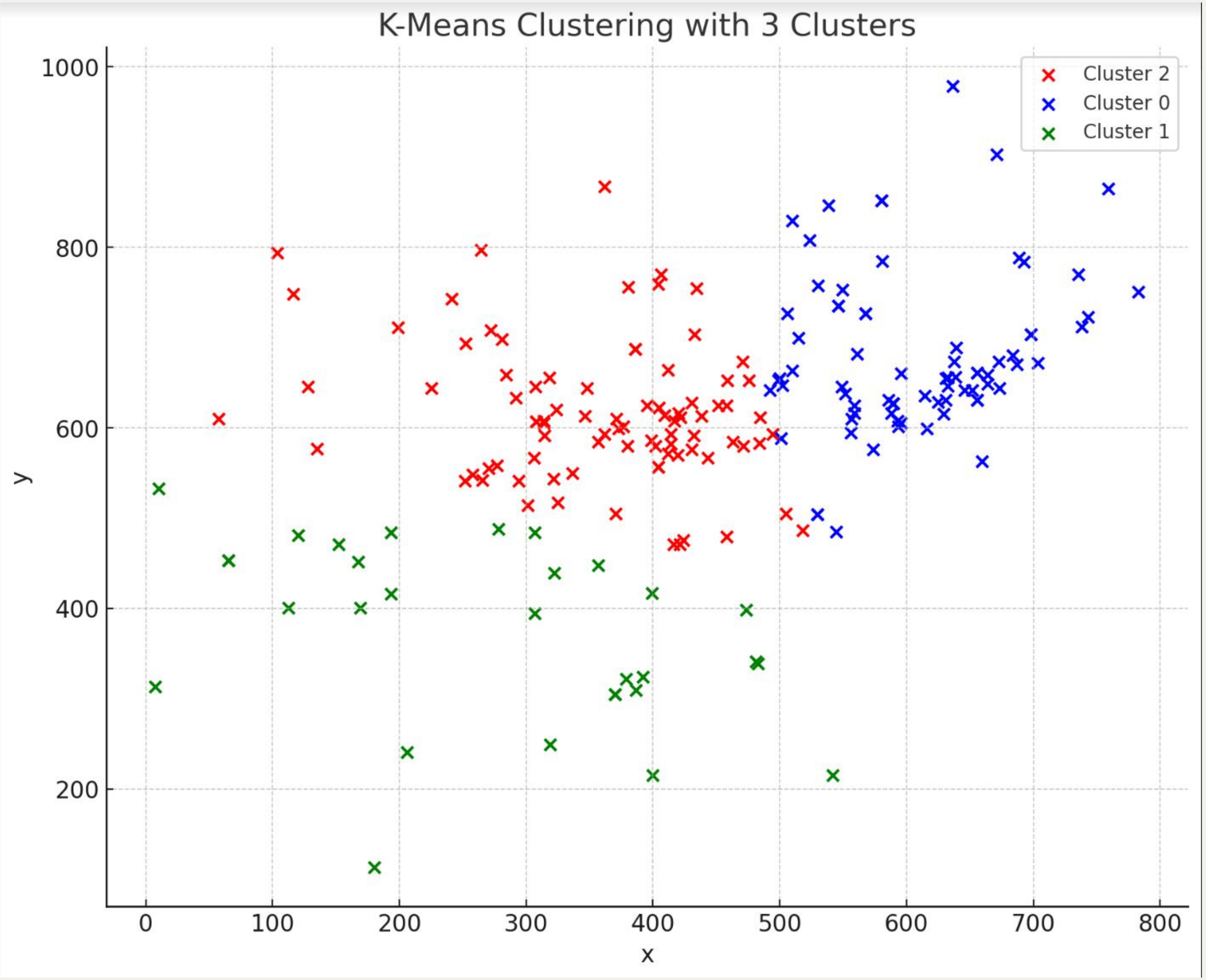
5 Cluster Plot



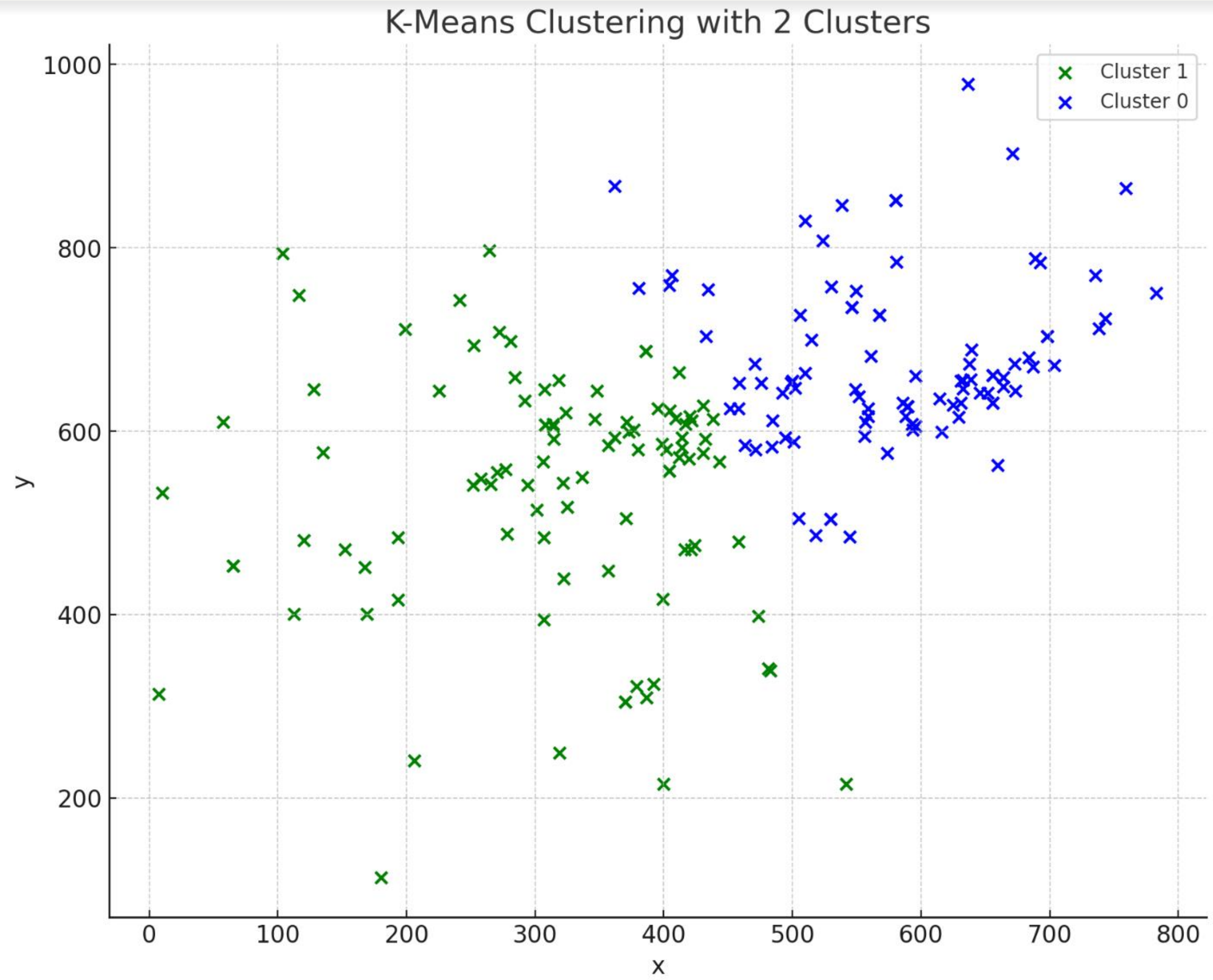
4 Cluster Plot



3 Cluster Plot

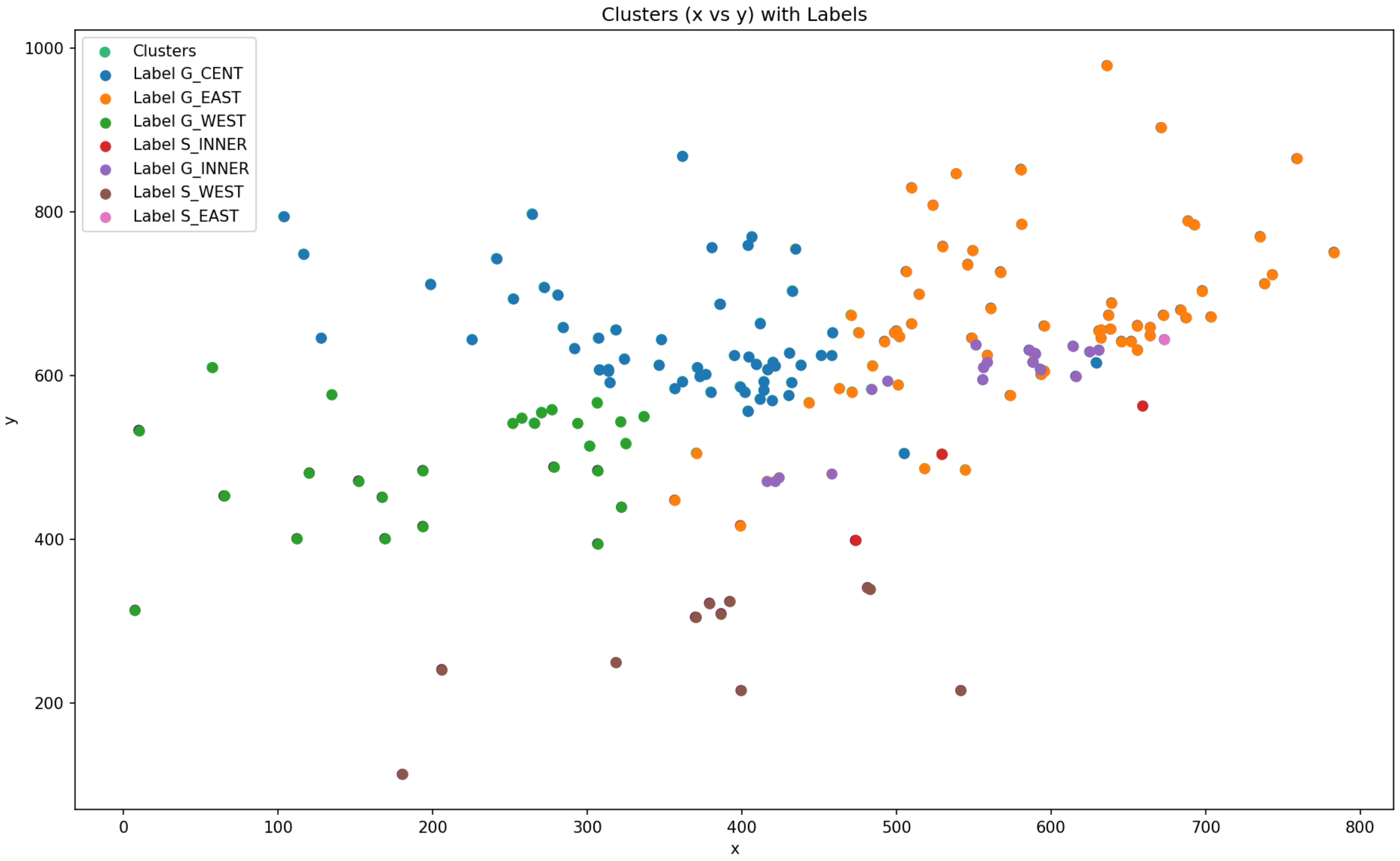


2 Cluster Plot

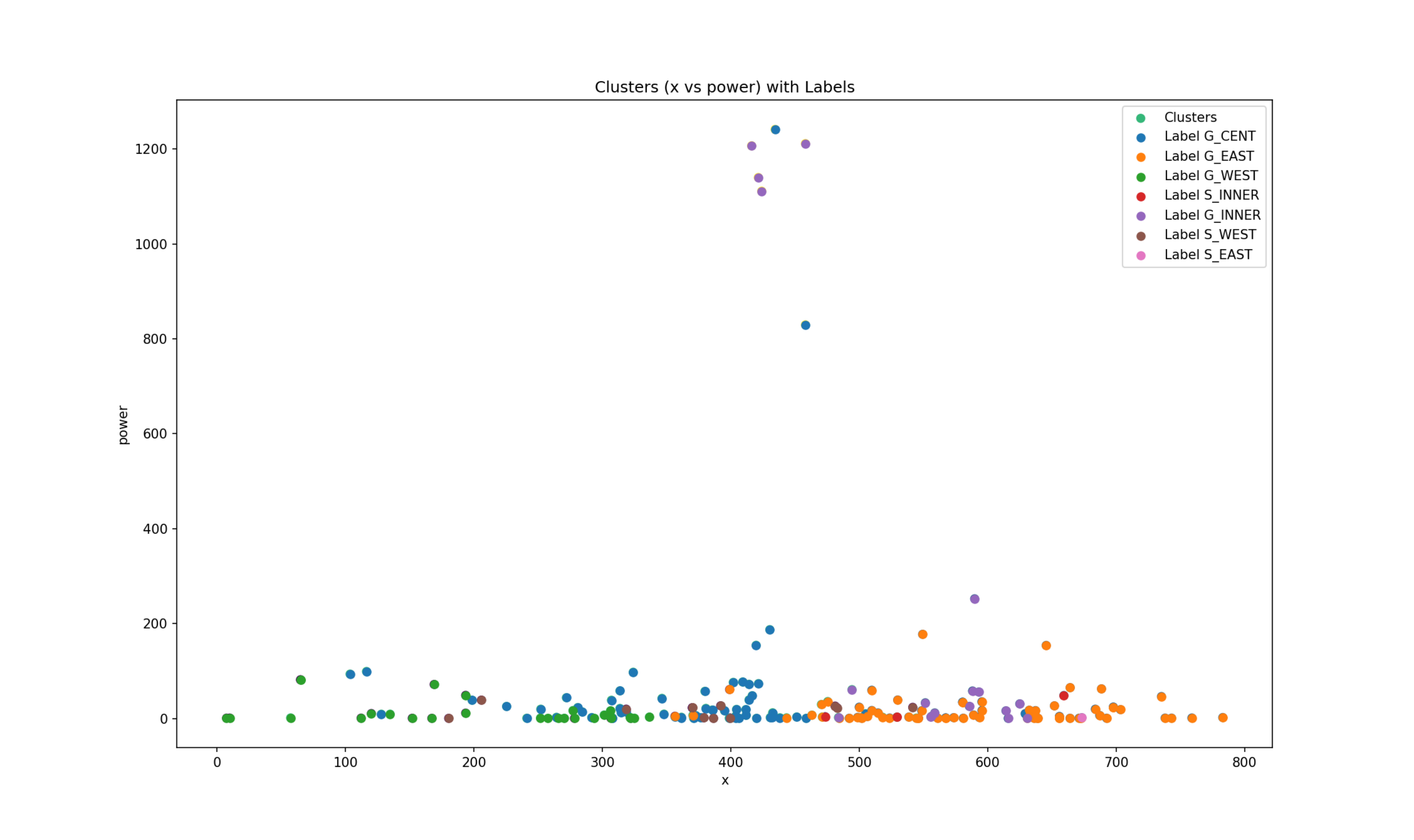


Clusters (x vs y) with Given Labels

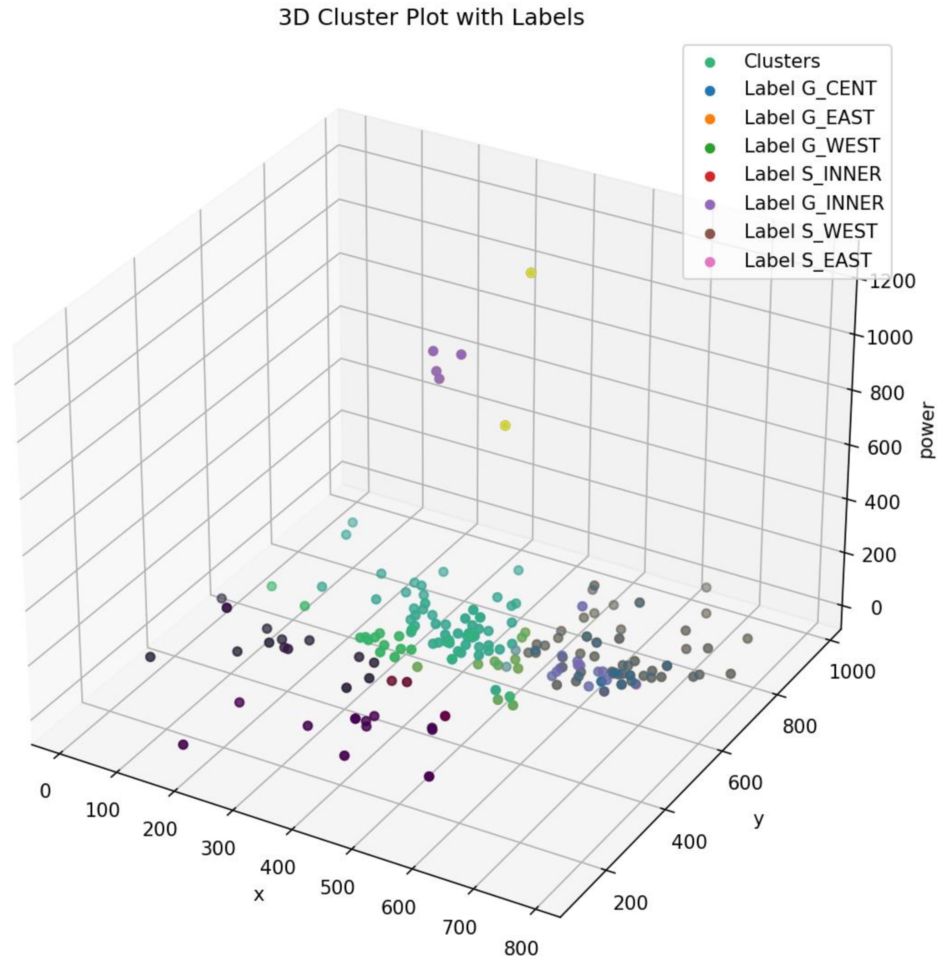
Results


Clusters (x vs y) with Labels
Clusters
Label G_CENT
Label G_EAST
Label G_WEST
Label S_INNER
Label G_INNER
Label S_WEST
Label S_EAST
1000
800
600
400
200
y
0
100
200
300
400
500
600
700
800
x

Results

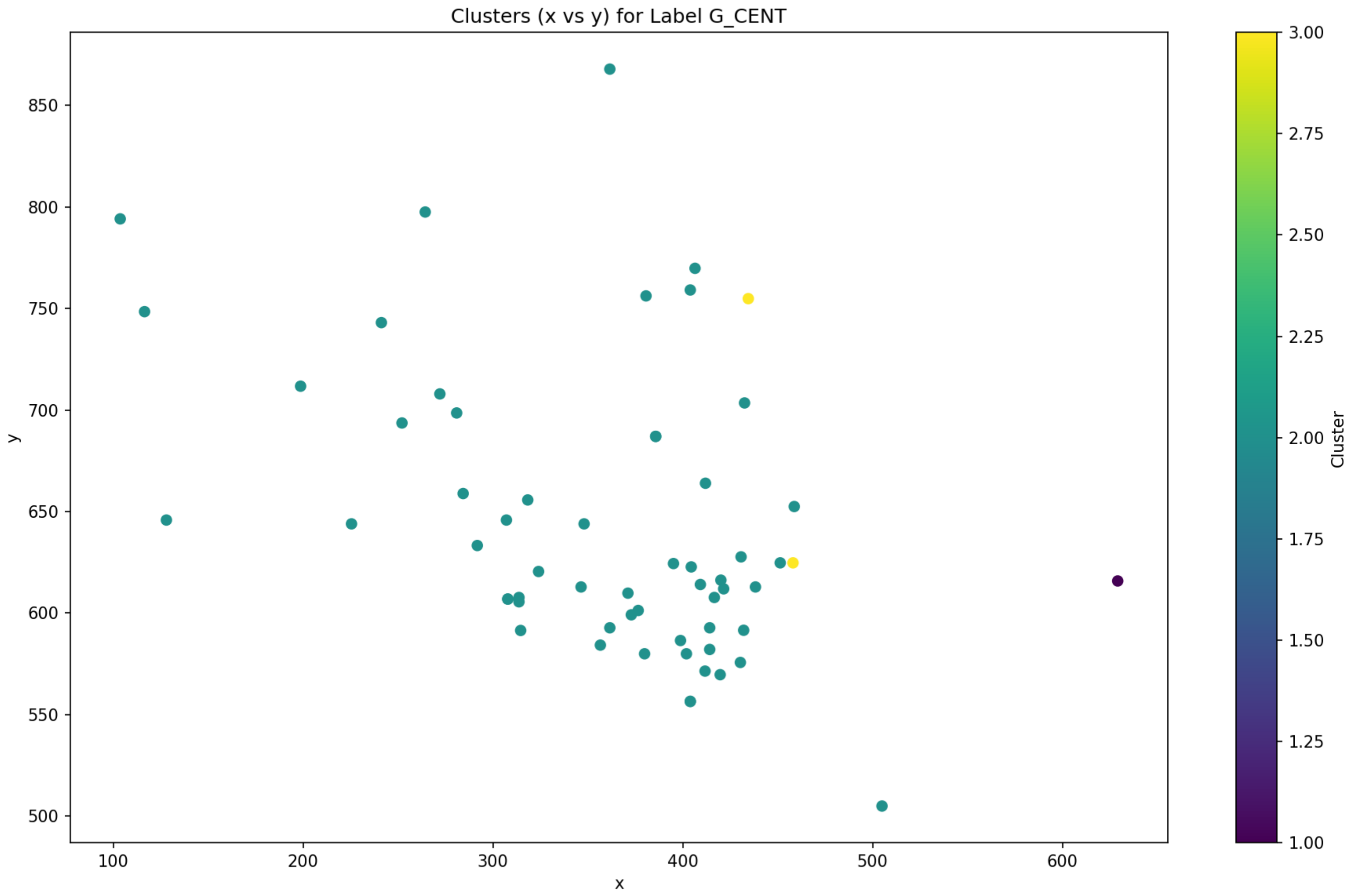

Clusters (x vs power) with Labels
Clusters
Label G_CENT
Label G_EAST
Label G_WEST
Label S_INNER
Label G_INNER
Label S_WEST
Label S_EAST
power
0
200
400
600
800
1000
1200
x
0
100
200
300
400
500
600
700
800

Results

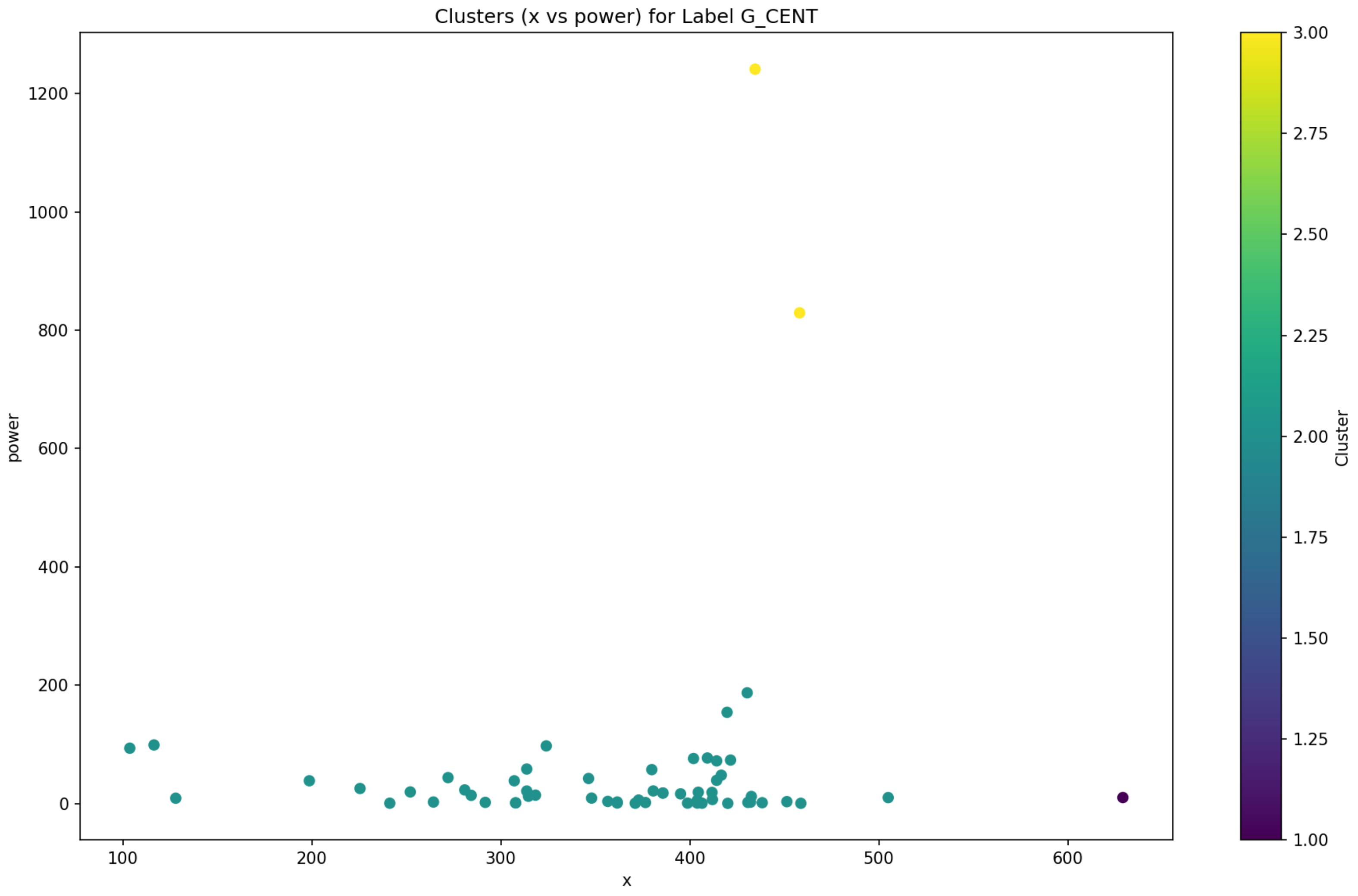

3D Cluster Plot with Labels
Clusters
Label G_CENT
Label G_EAST
Label G_WEST
Label S_INNER
Label G_INNER
Label S_WEST
Label S_EAST
power
0
200
400
600
800
1000
1200
y
200
400
600
800
1000
x
0
100
200
300
400
500
600
700
800

Detailed Results of
Clusters (x vs y) with Given Labels

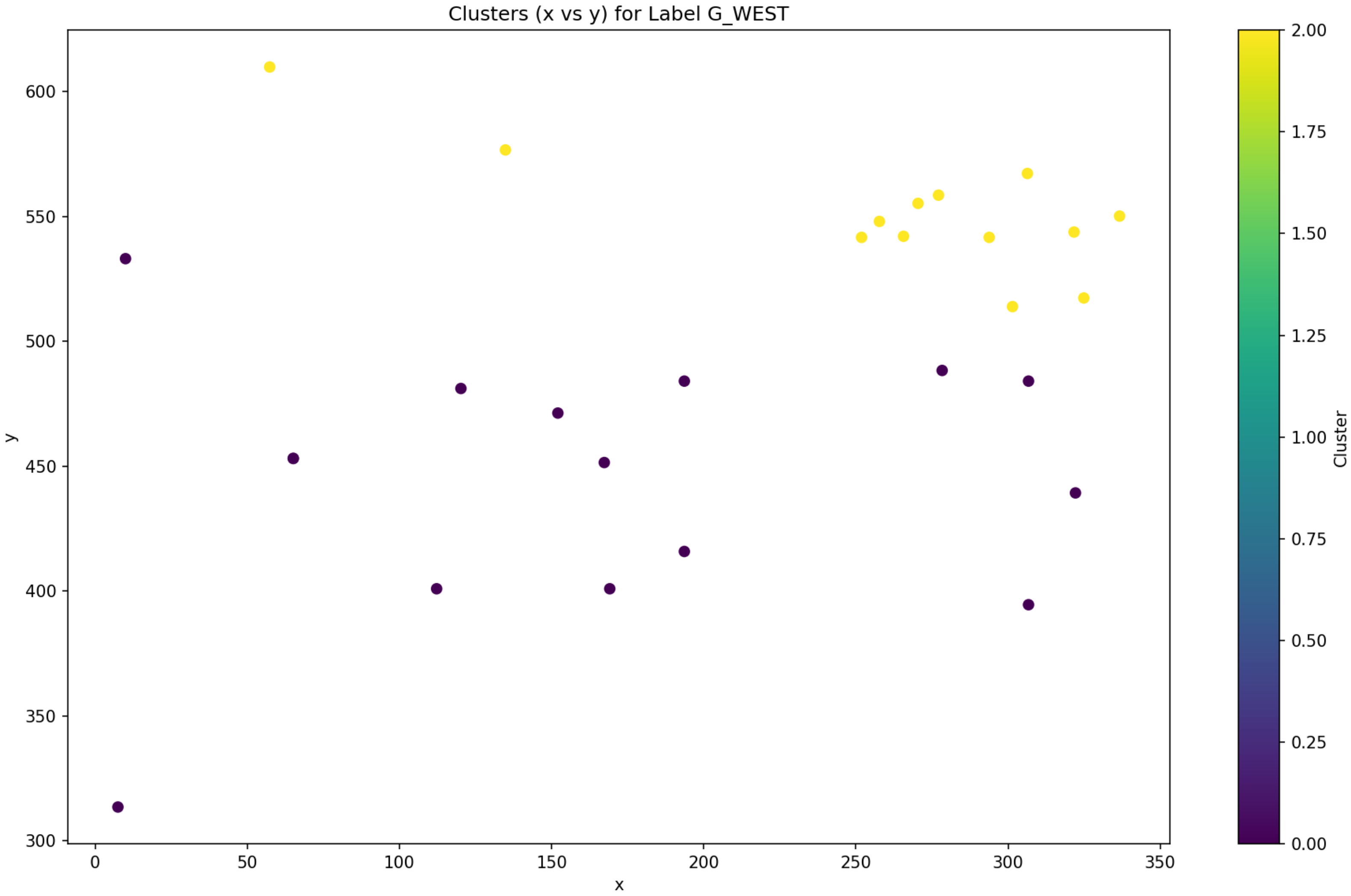
Results


Clusters (x vs y) for Label G_CENT
y
850
800
750
700
650
600
550
500
x
100
200
300
400
500
600
Cluster
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00

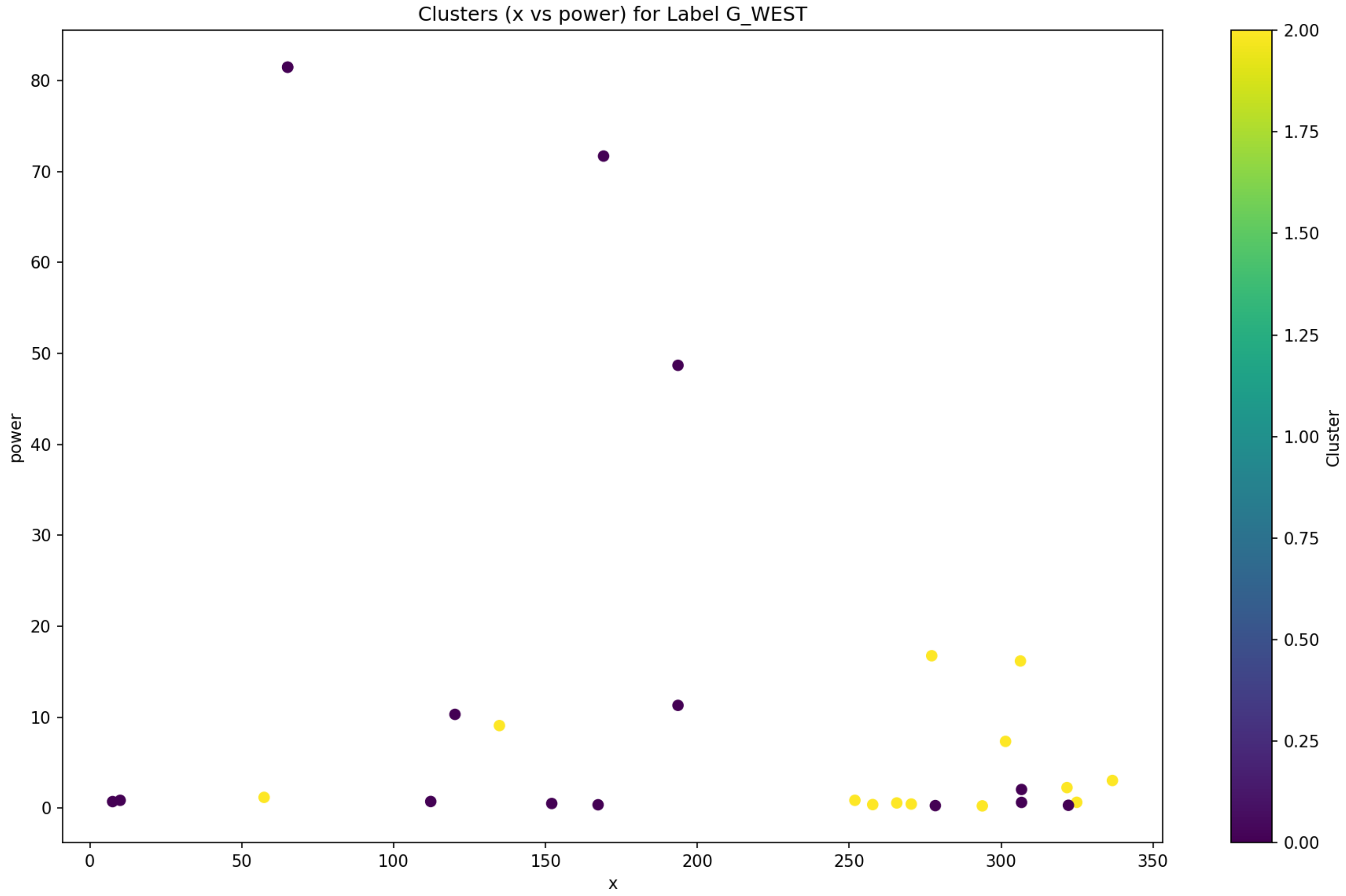
Results


Clusters (x vs power) for Label G_CENT
power
1200
1000
800
600
400
200
0
x
100
200
300
400
500
600
Cluster
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00

Results

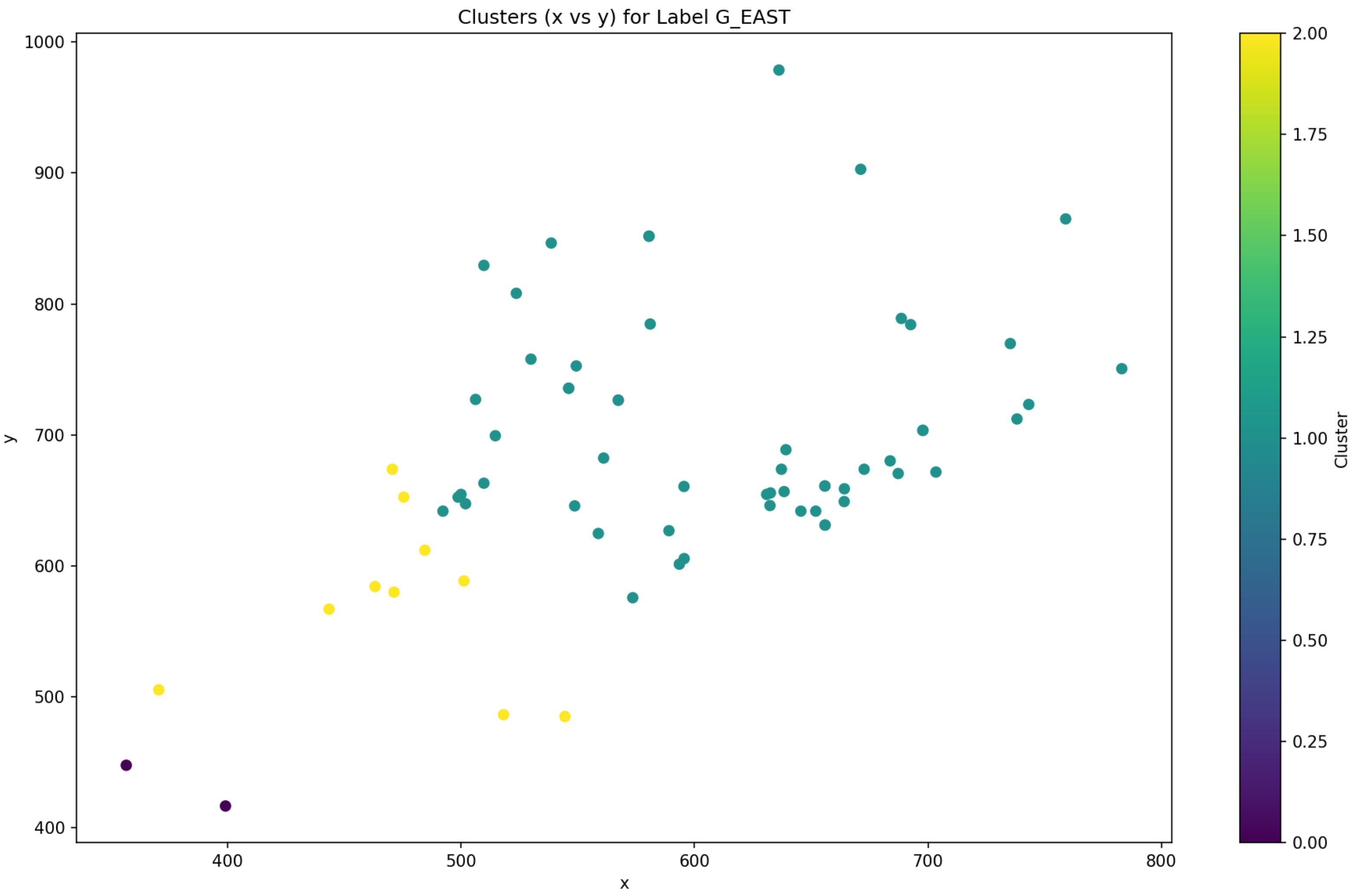

Clusters (x vs y) for Label G_WEST
y
600
550
500
450
400
350
300
x
0
50
100
150
200
250
300
350
Cluster
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

Results

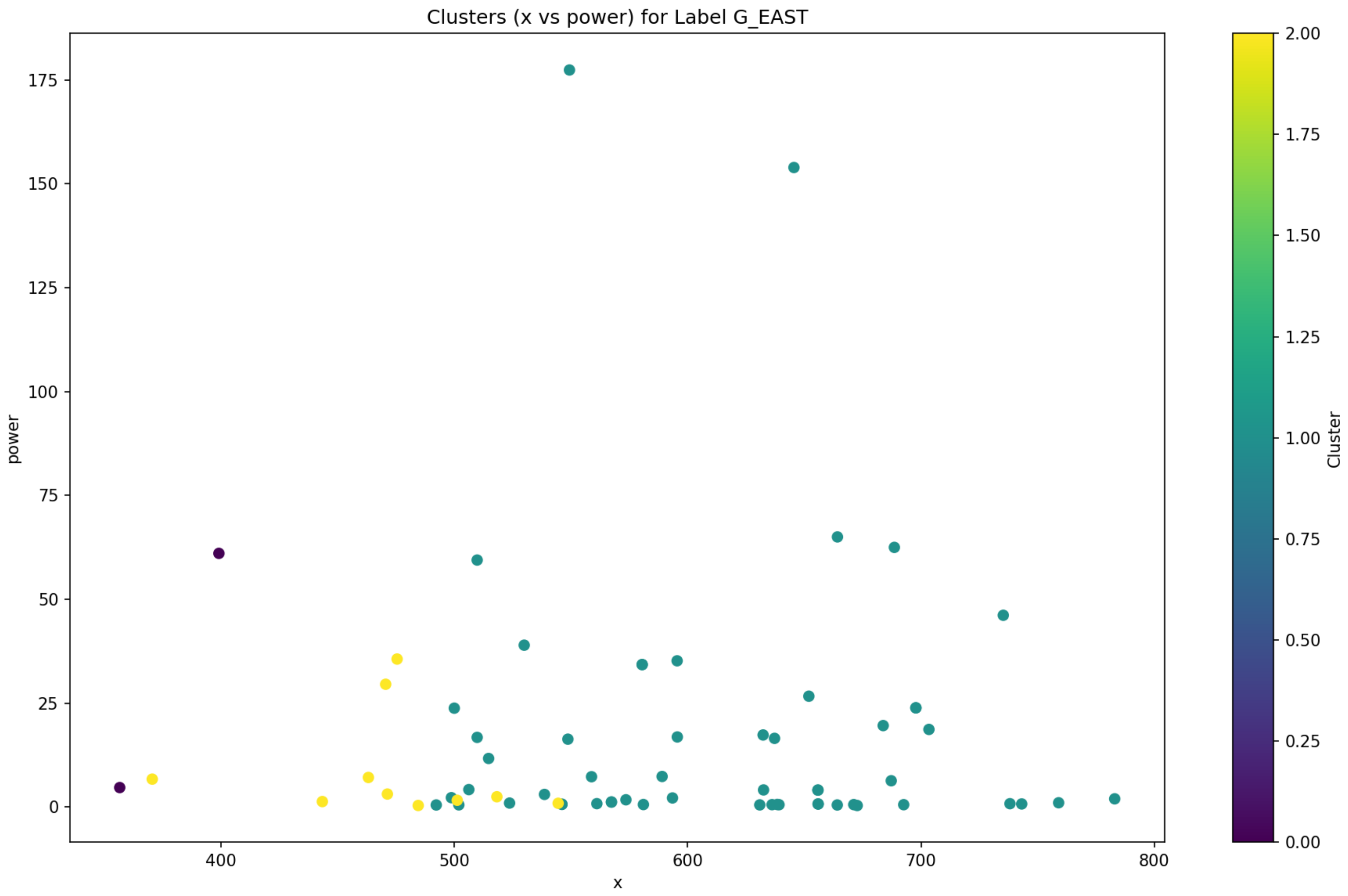


Clusters (x vs power) for Label G_WEST
power
80
70
60
50
40
30
20
10
0
x
0
50
100
150
200
250
300
350
Cluster
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

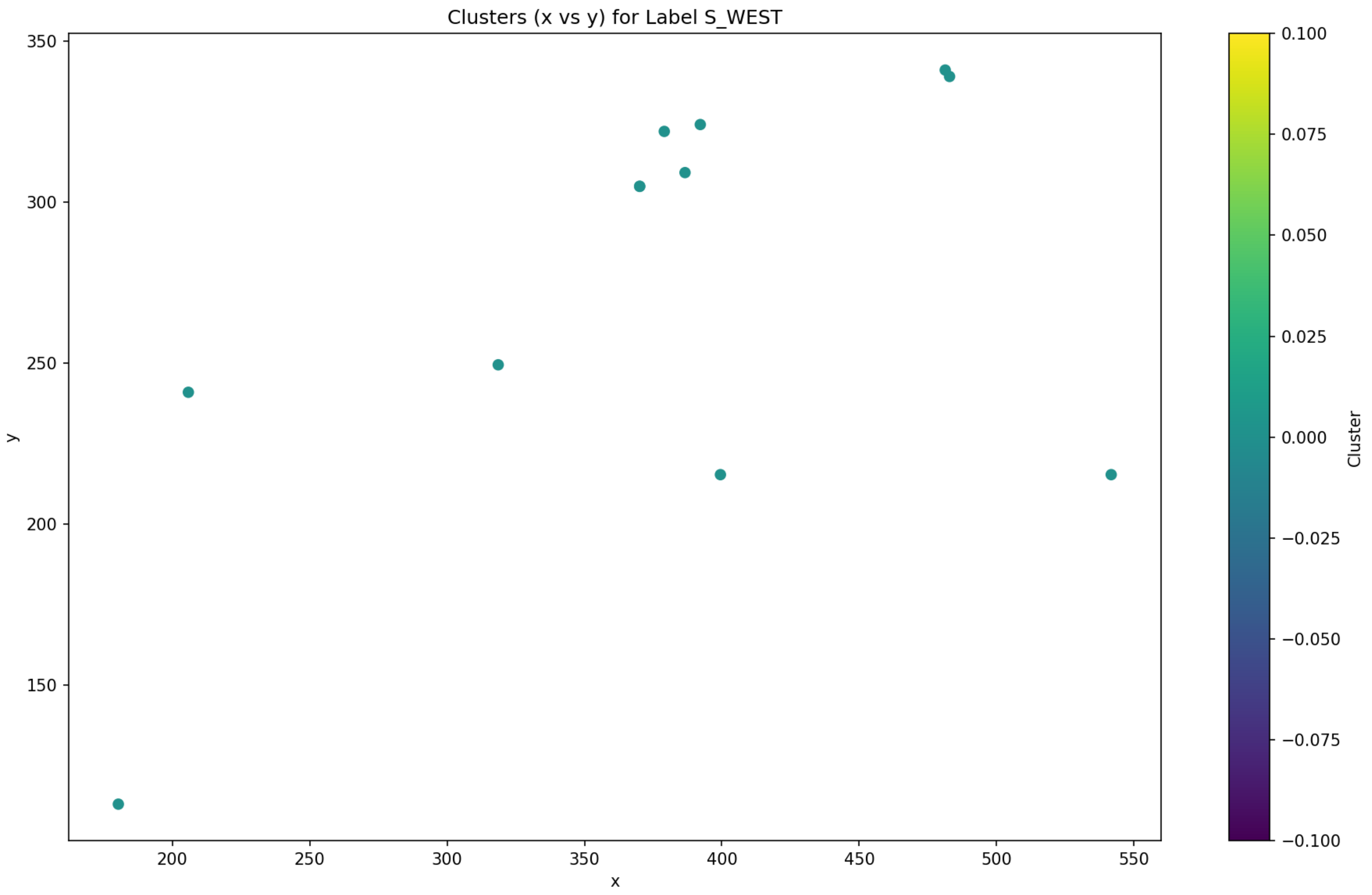
Results


Clusters (x vs y) for Label G_EAST
y
1000
900
800
700
600
500
400
x
400
500
600
700
800
Cluster
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

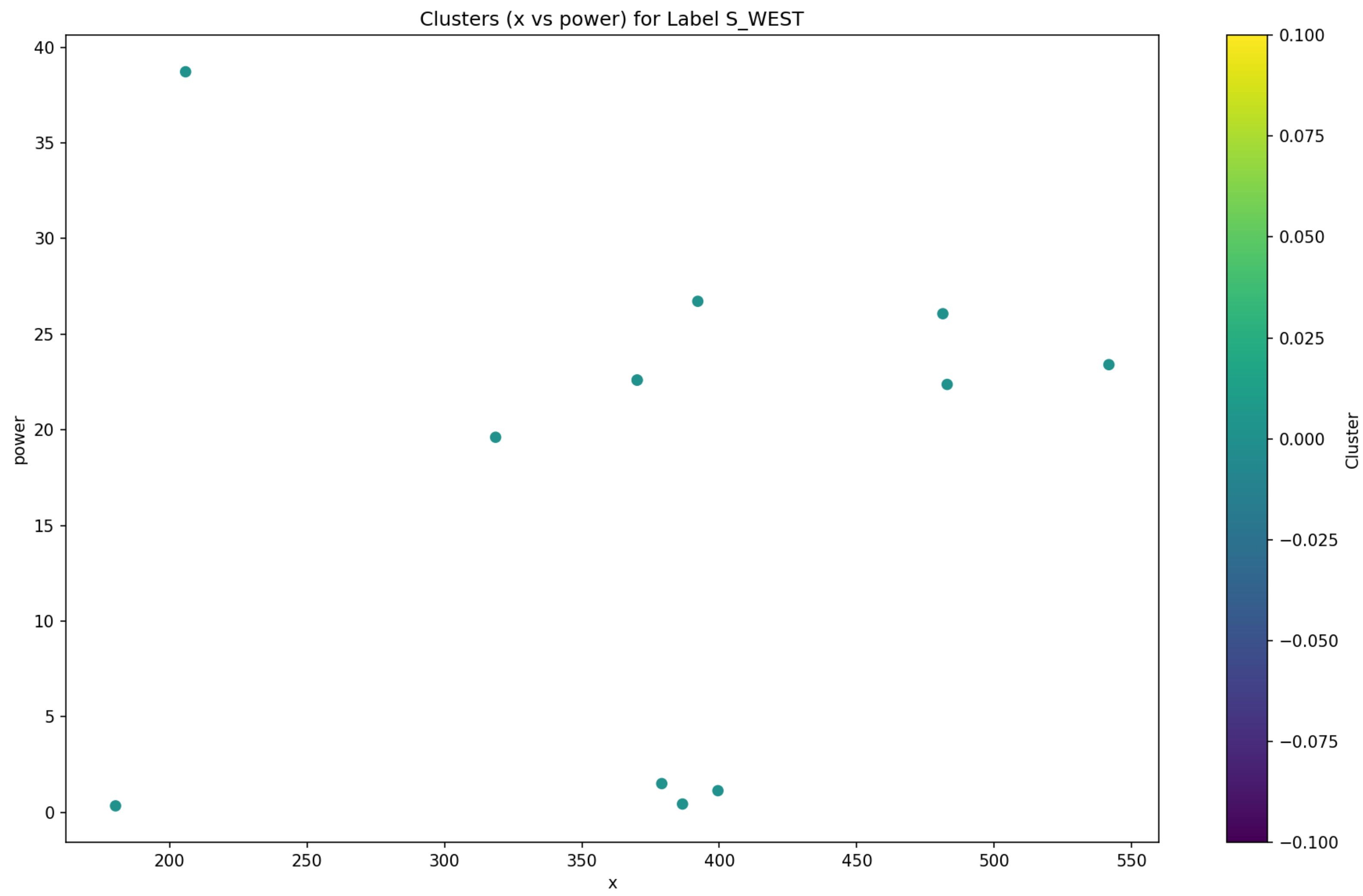
Results


Clusters (x vs power) for Label G_EAST
power
175
150
125
100
75
50
25
0
x
400
500
600
700
800
Cluster
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

Results

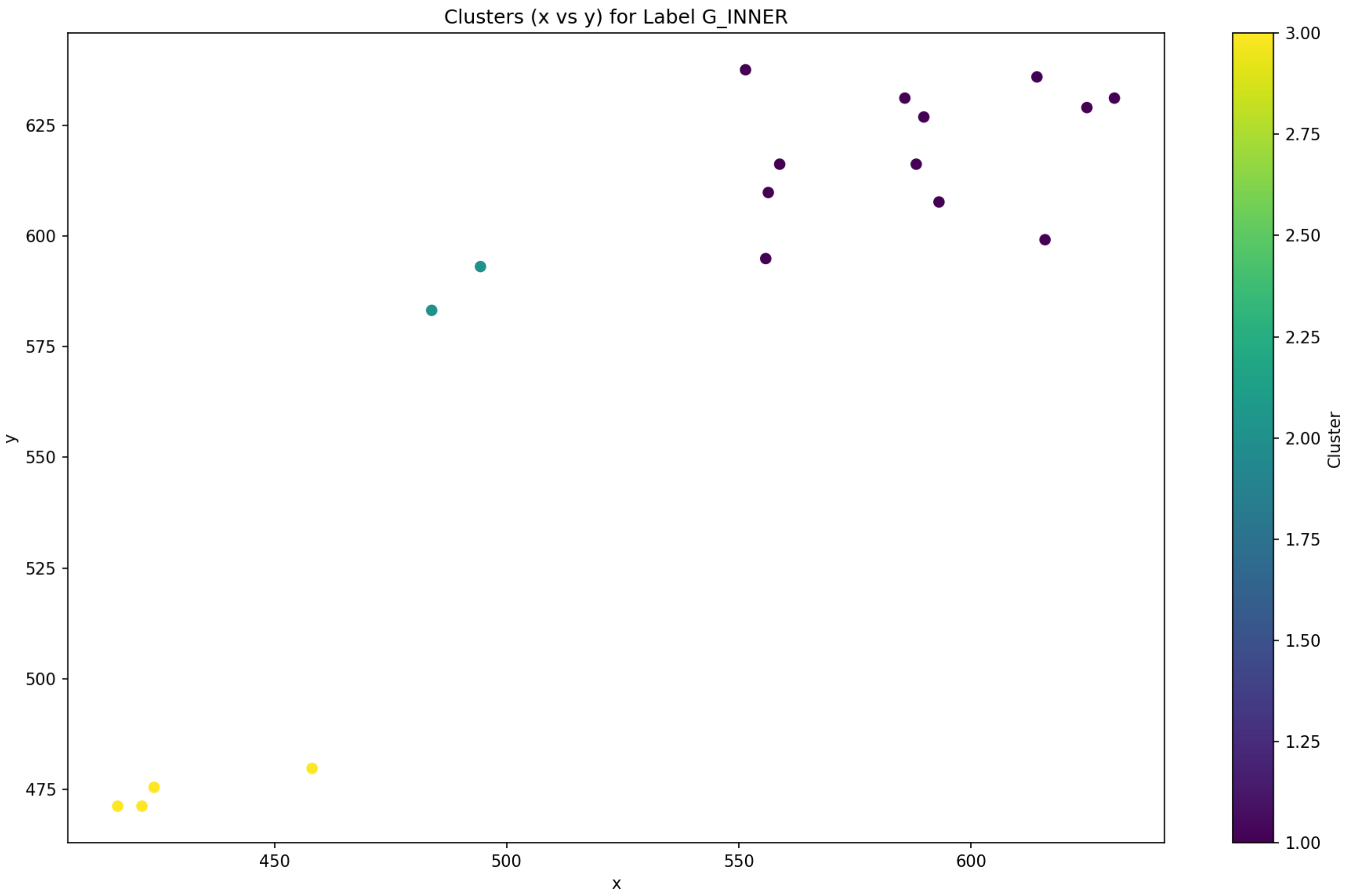

Clusters (x vs y) for Label S_WEST
350
300
250
200
150
y
200
250
300
350
400
450
500
550
x
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
Cluster

Results



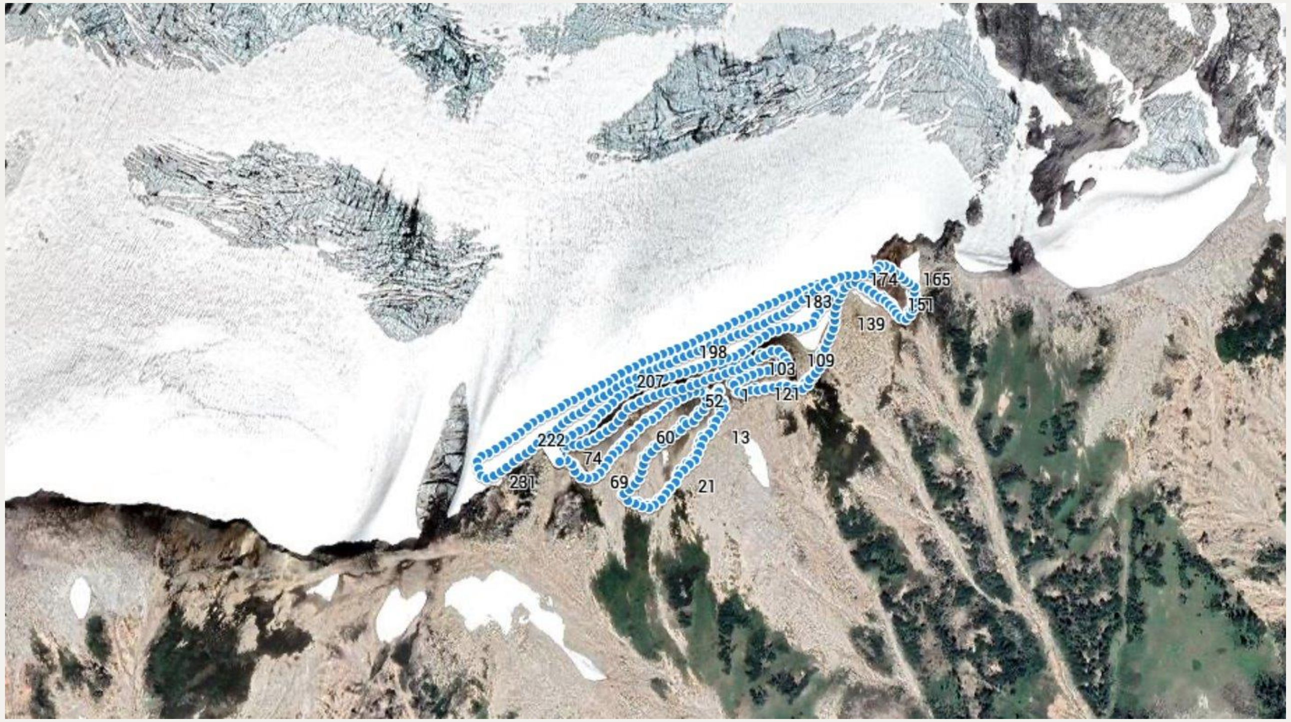
Clusters (x vs power) for Label S_WEST
power
x
0
5
10
15
20
25
30
35
40
200
250
300
350
400
450
500
550
Cluster
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

Results


Clusters (x vs y) for Label G_INNER
y
625
600
575
550
525
500
475
x
450
500
550
600
Cluster
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00

05 Closing

Implemented a Python program for preprocessing, clustering, and visualization of geophysical microseismic event data, aimed to uncover underlying patterns in multi-dimensional event parameters.



Mk Maharana

www.linkedin.com/in/mrigank-maharana-67a07020a