Metrics Testing

10/23

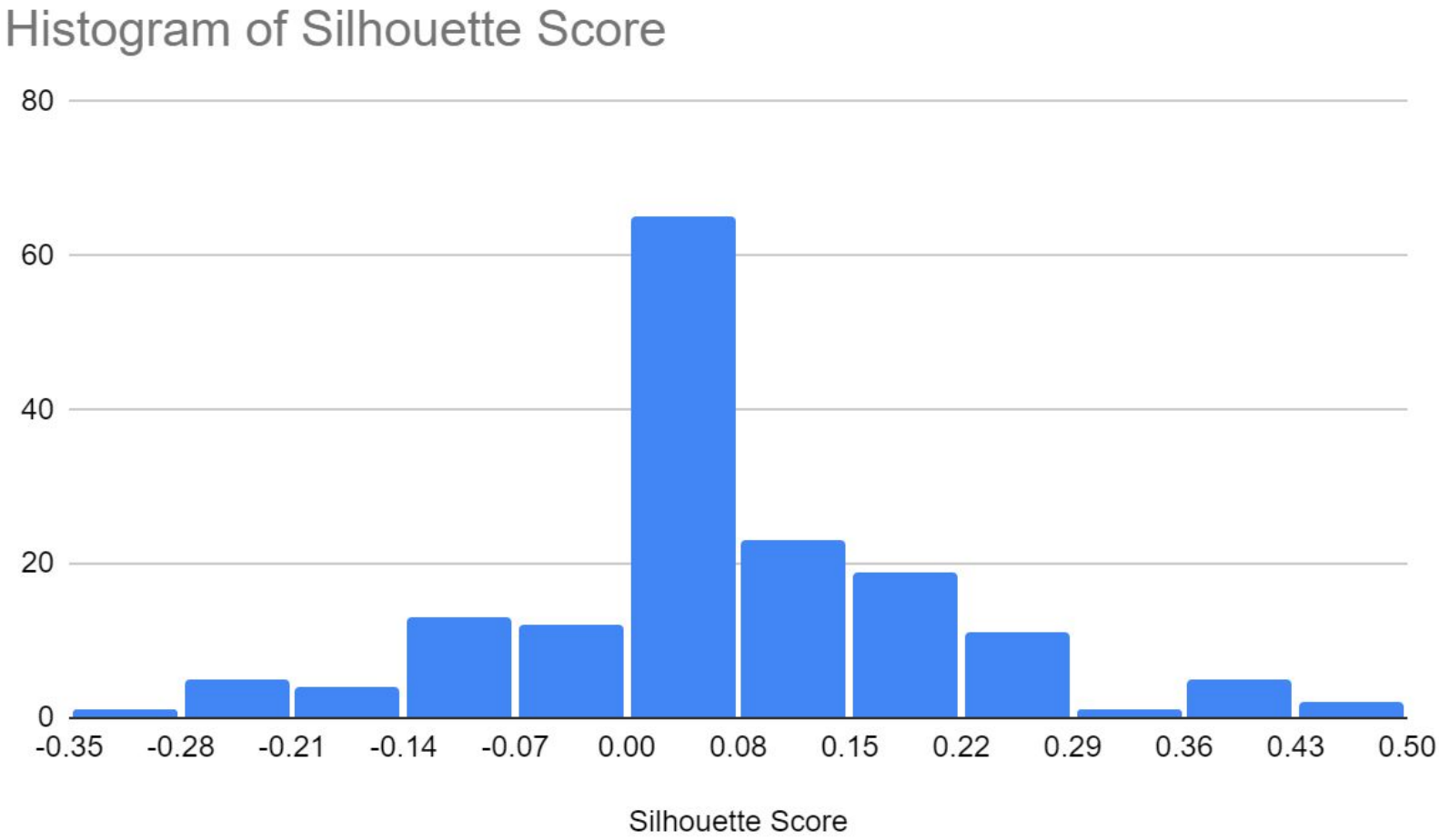
Procedure

- Ran testing on GovInfo, HCPSS, and General datasets
- Algorithms: Birch, OPTICS, K-Means, Agglomerative, Spectral, Mean Shift, DBSCAN, HDBSCAN
- Min clusters: 5
- Max clusters: 15
- Copied and saved data to Google Drive, created graphs using tools

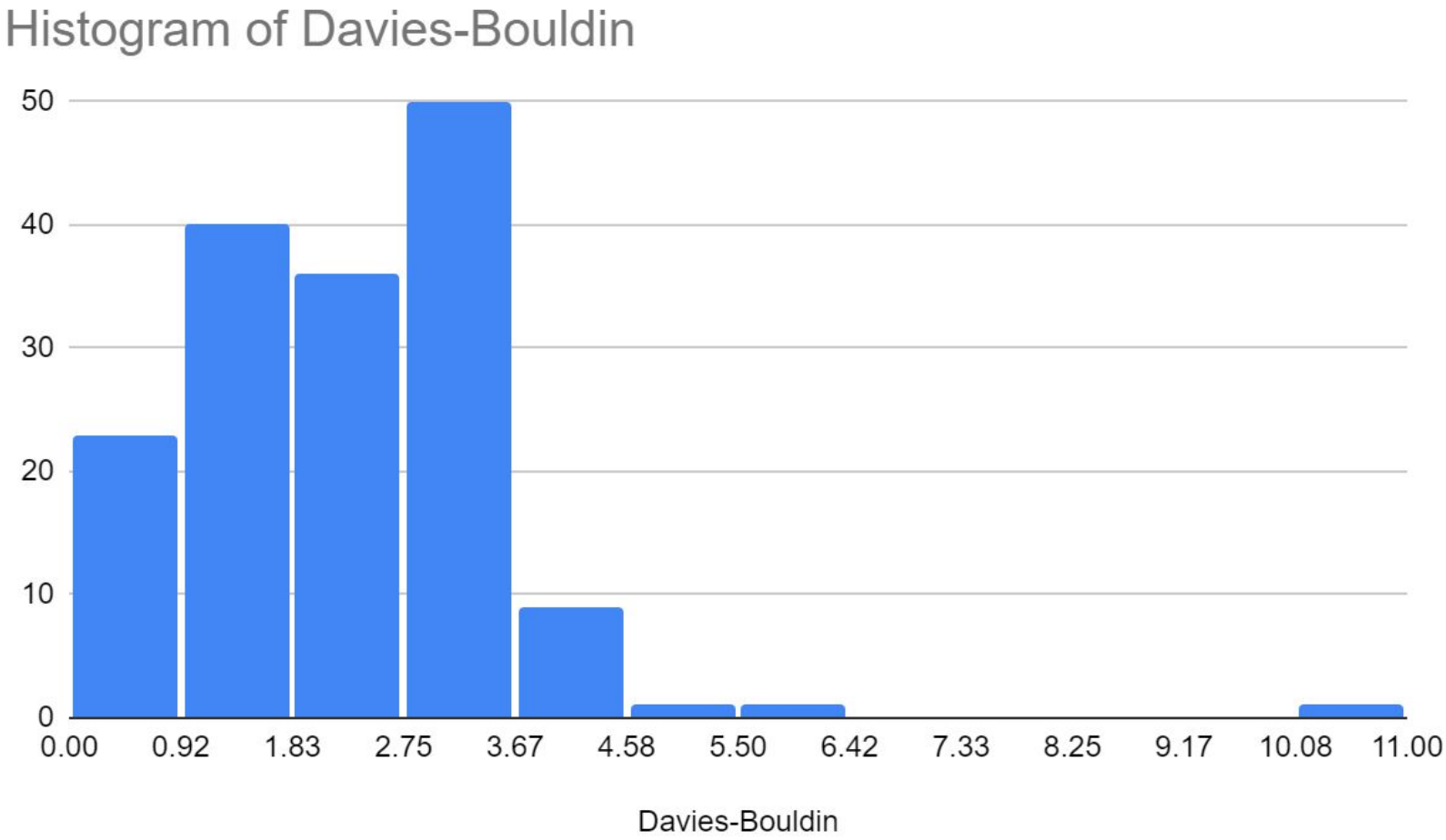
Goal

- Compare scores across datasets and files
- Update benchmarks for metrics
- Create grading boundaries that fit data scoring
- Decide which clustering algorithms are best for our datasets

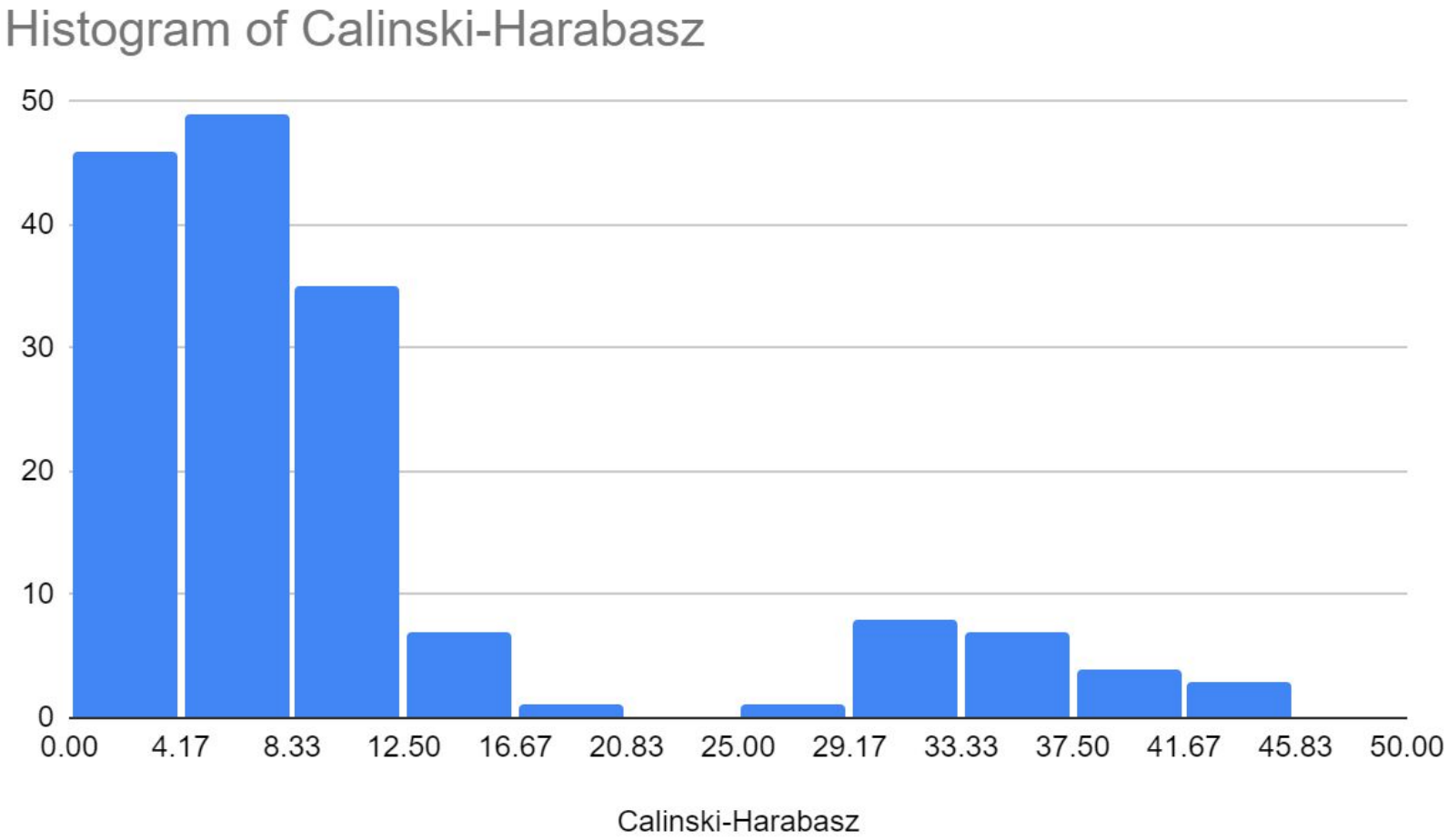
All Datasets & Algorithms



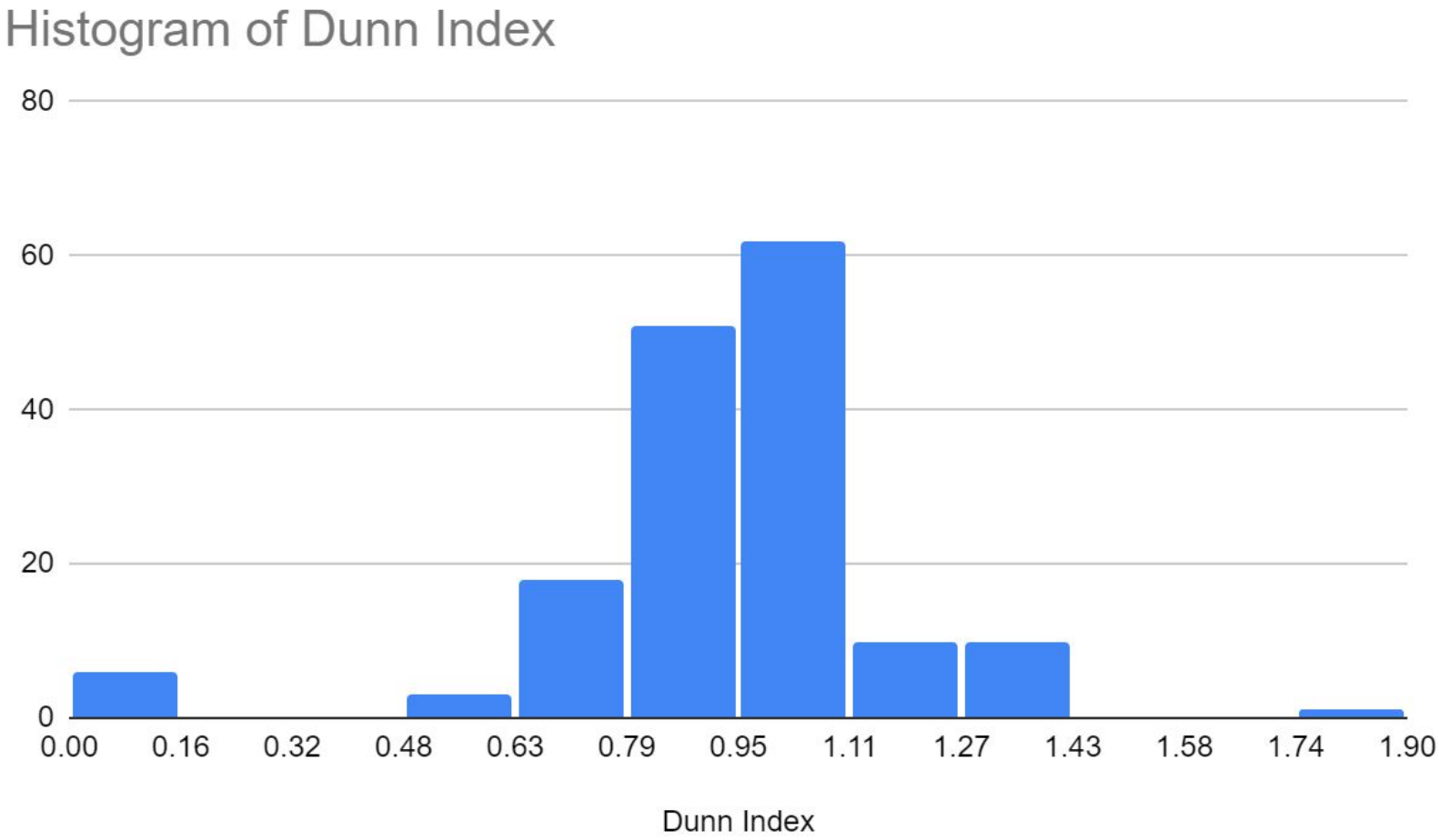
All Datasets + Algorithms



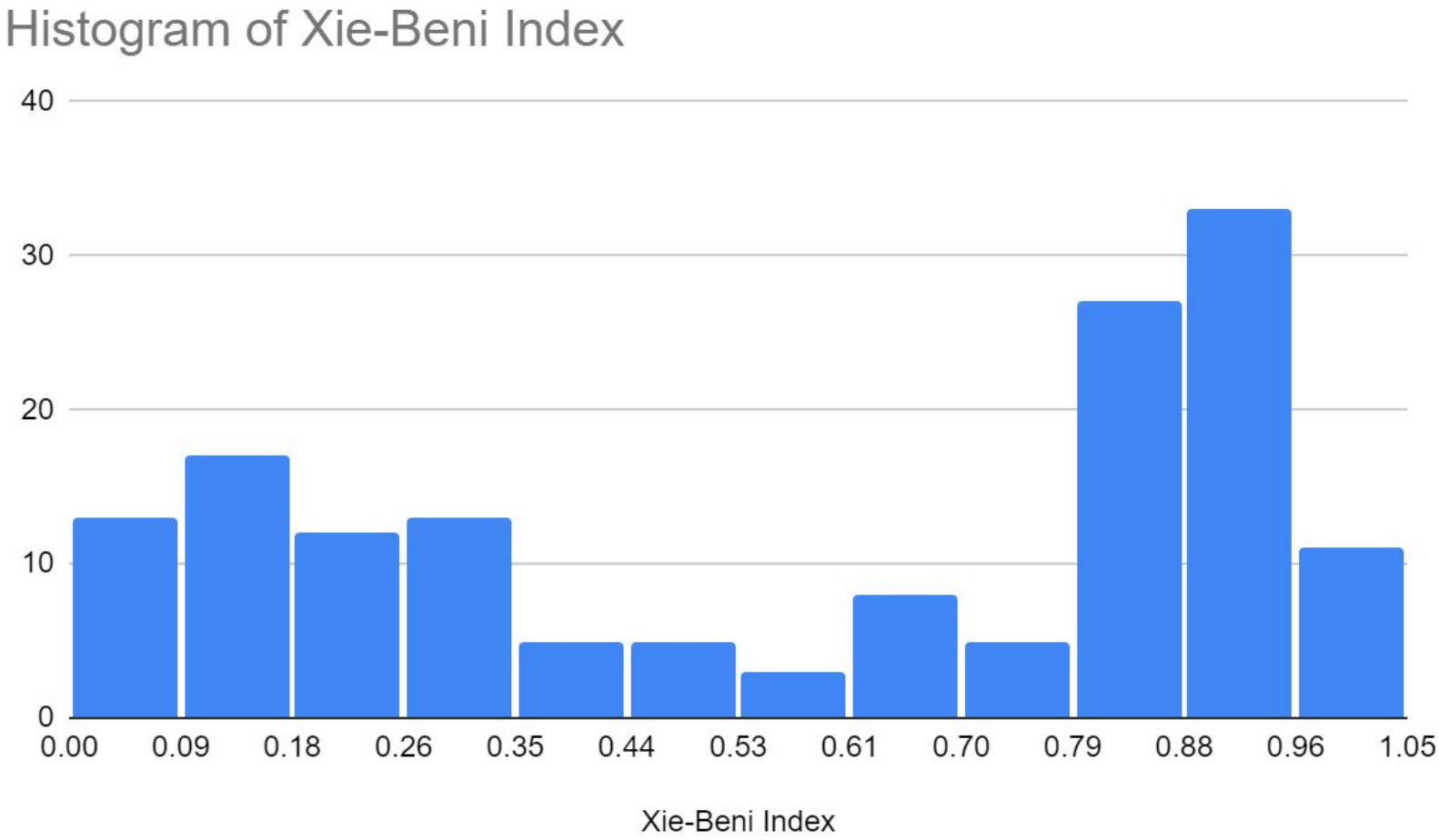
All Datasets + Algorithms



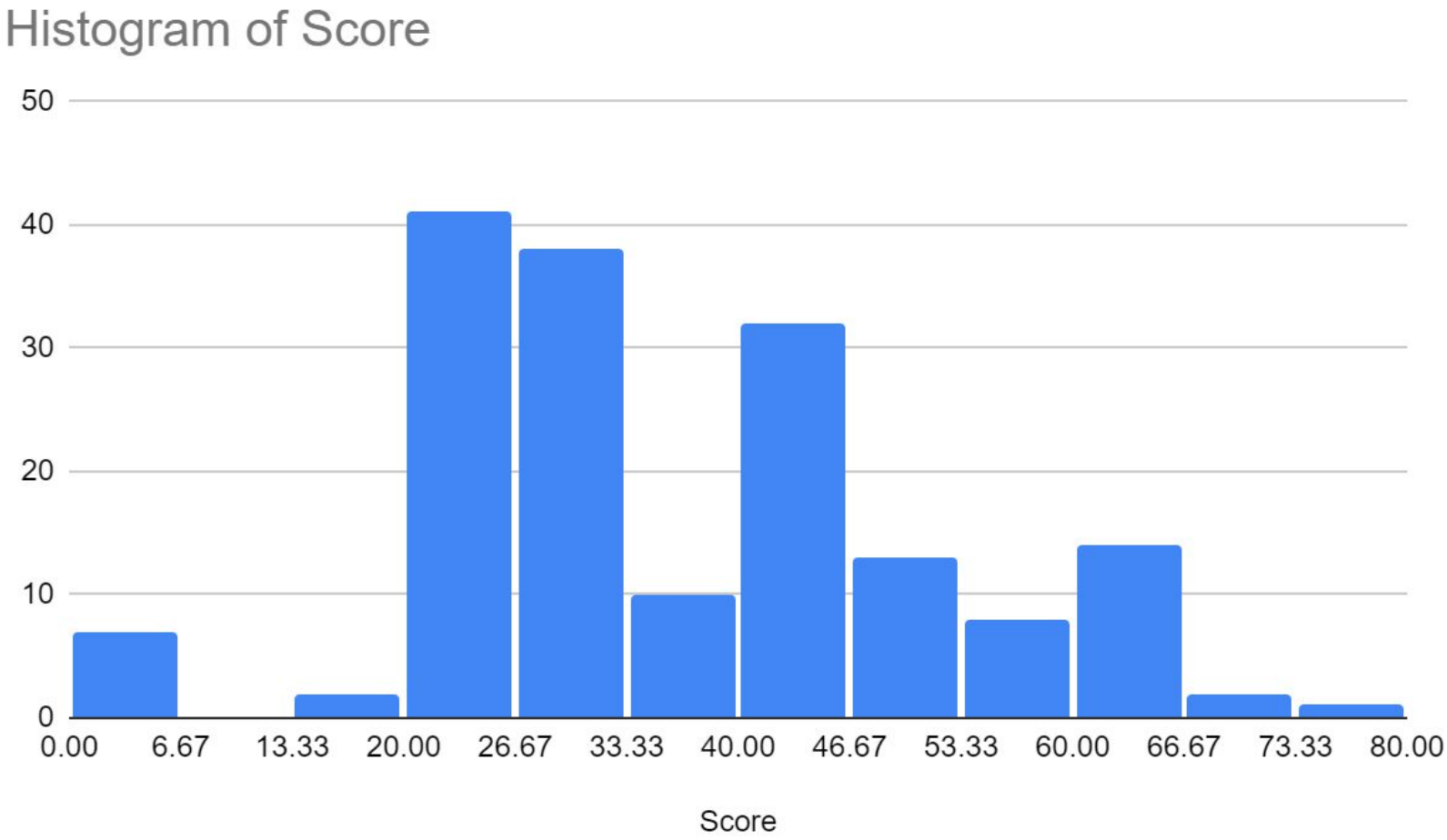
All Datasets + Algorithms



All Datasets + Algorithms

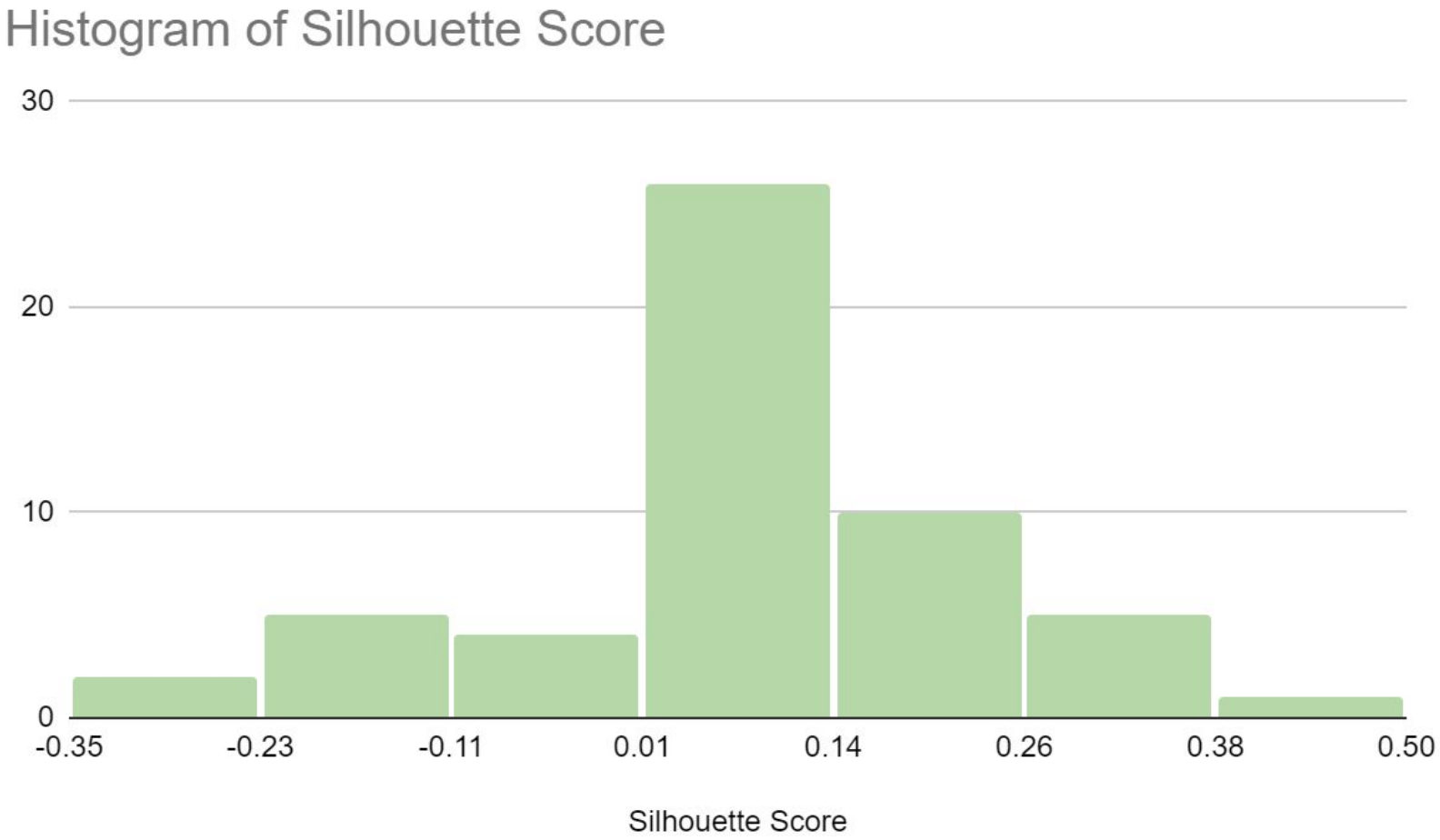


All Datasets + Algorithms

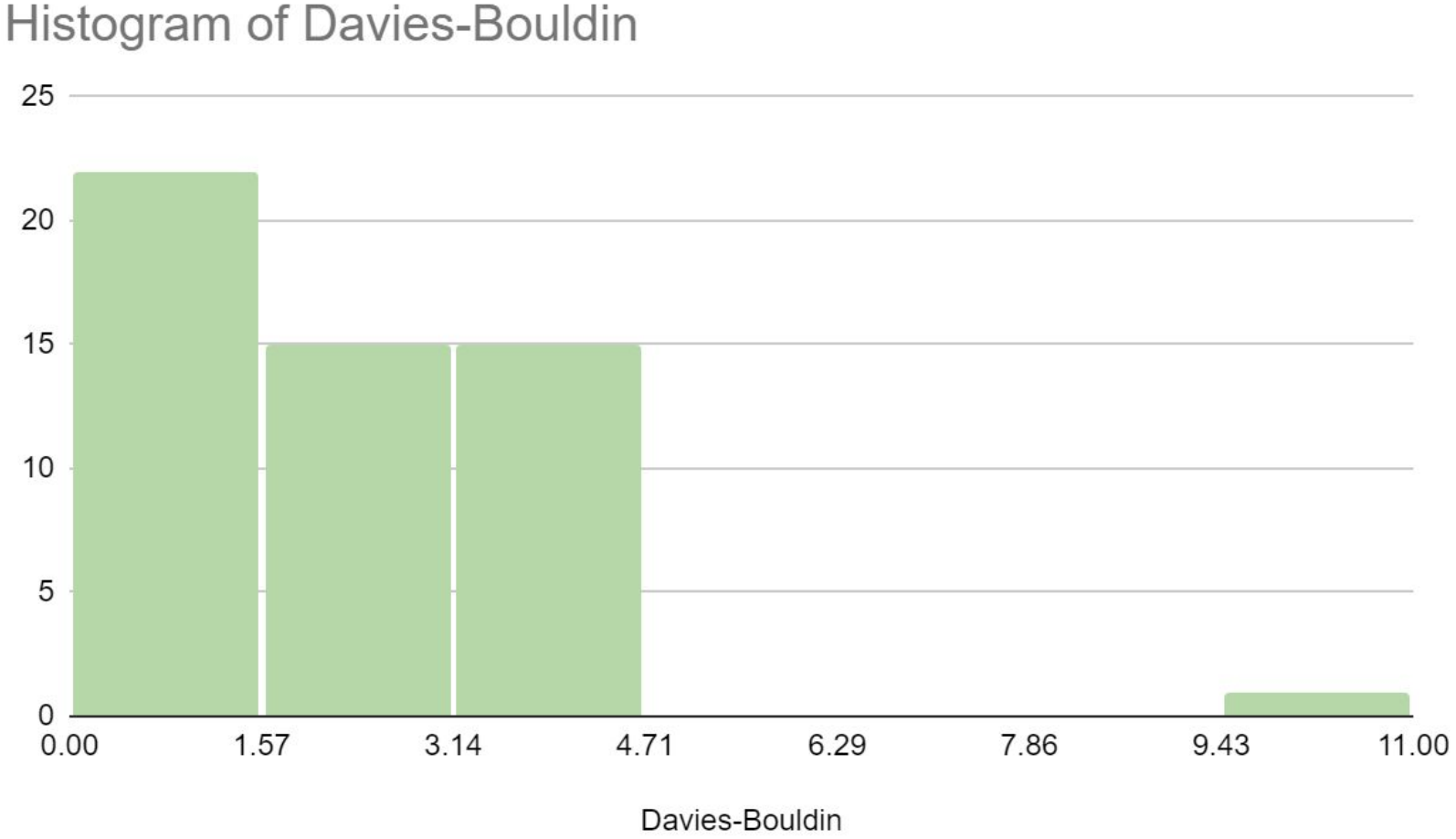


All Datasets + Algorithms

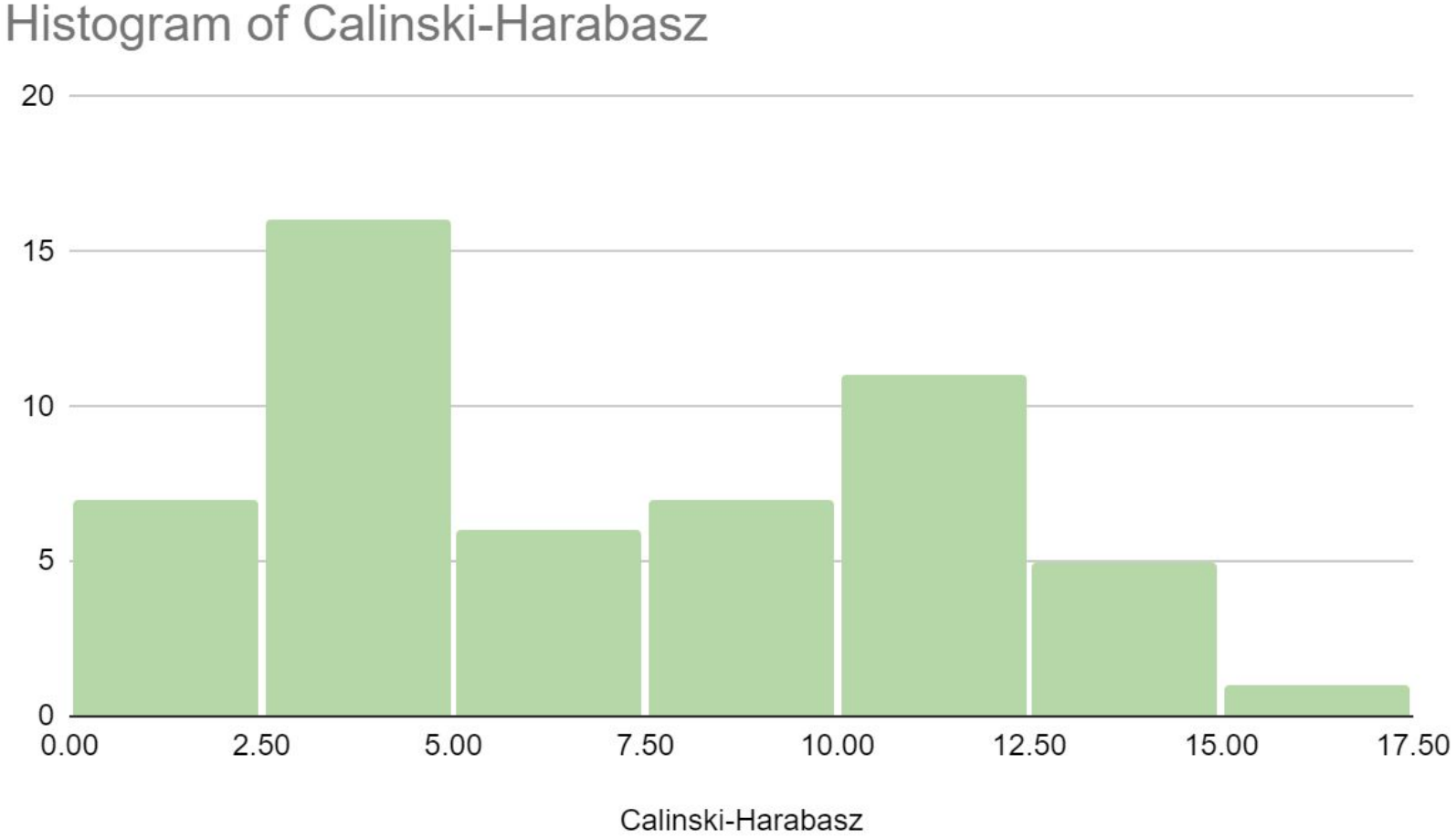
Privacy Law Dataset


Histogram of Silhouette Score
30
20
10
0
-0.35
-0.23
-0.11
0.01
0.14
0.26
0.38
0.50
Silhouette Score

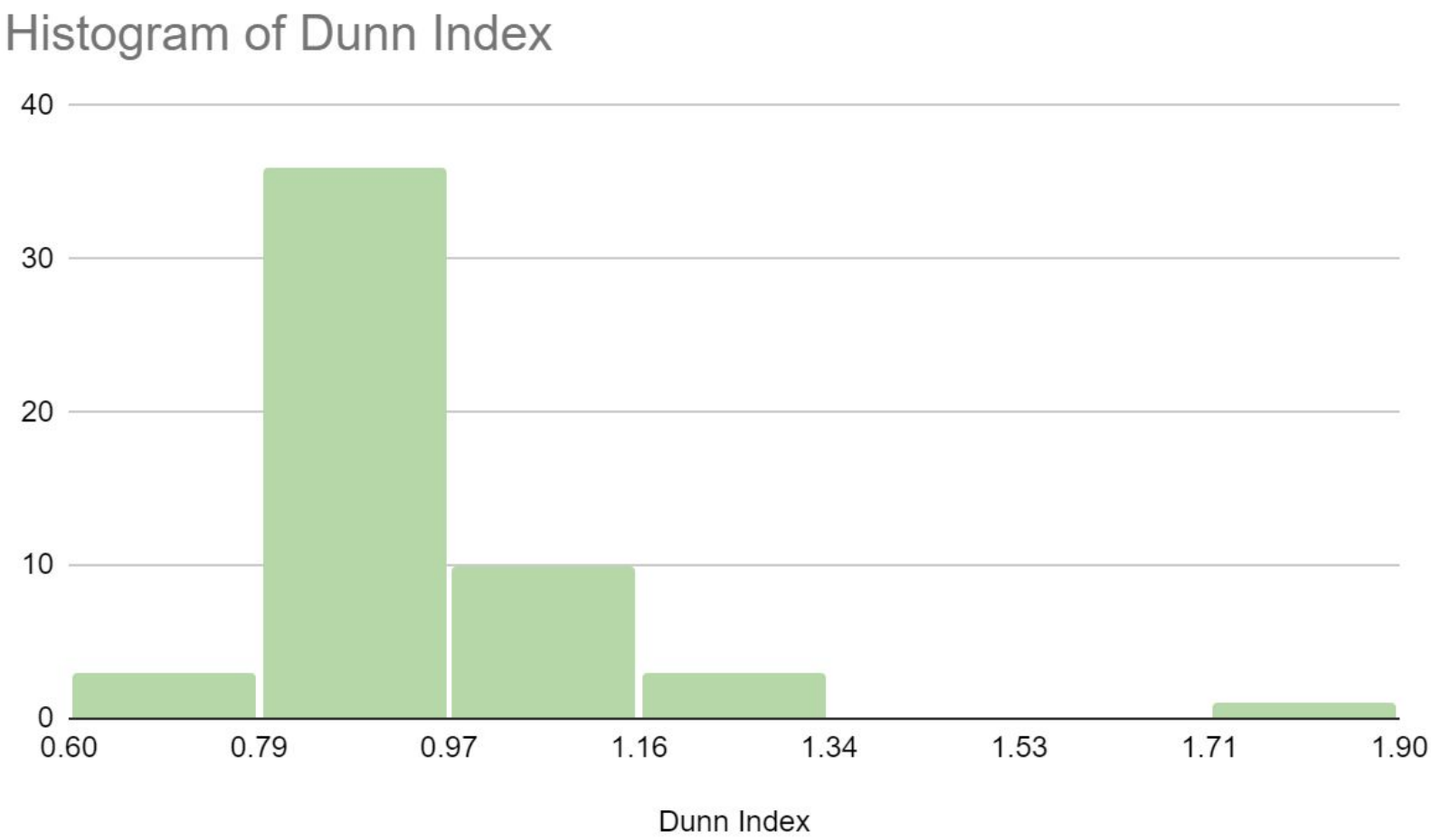
Privacy Law

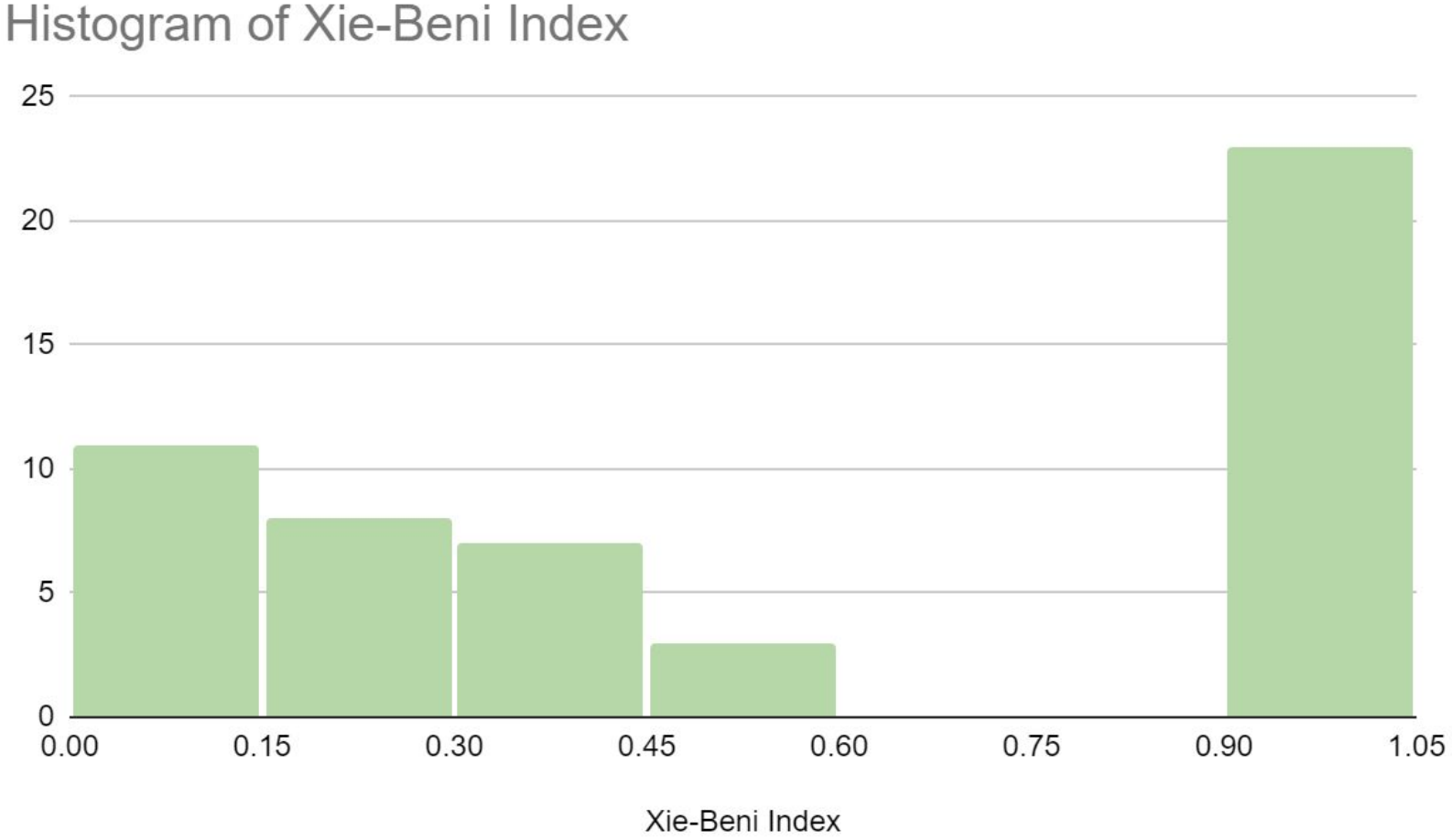

Histogram of Davies-Bouldin
25
20
15
10
5
0
0.00
1.57
3.14
4.71
6.29
7.86
9.43
11.00
Davies-Bouldin

Privacy Law

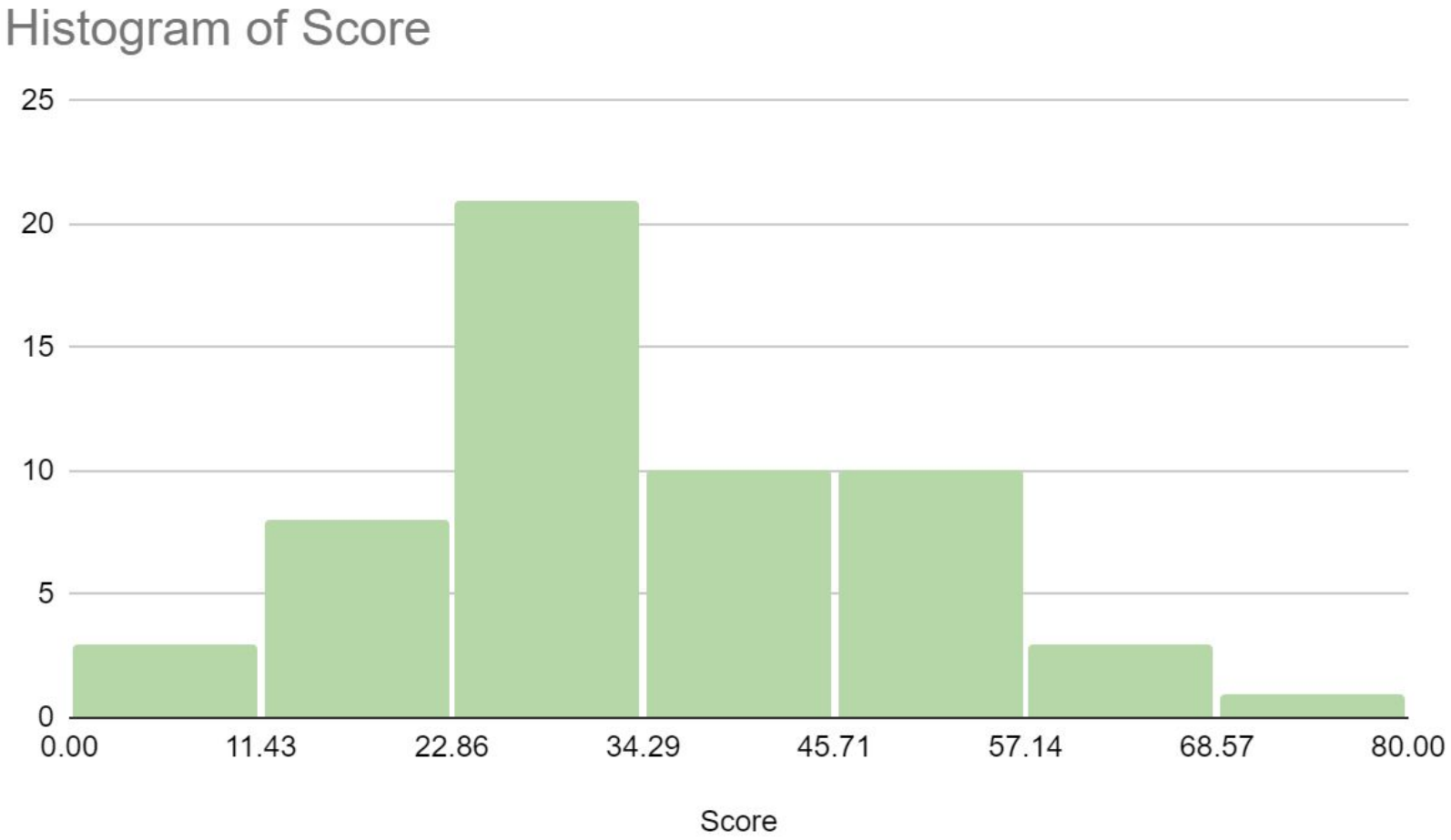

Histogram of Calinski-Harabasz
20
15
10
5
0
0.00
2.50
5.00
7.50
10.00
12.50
15.00
17.50
Calinski-Harabasz

Privacy Law


Histogram of Dunn Index
40
30
20
10
0
0.60
0.79
0.97
1.16
1.34
1.53
1.71
1.90
Dunn Index

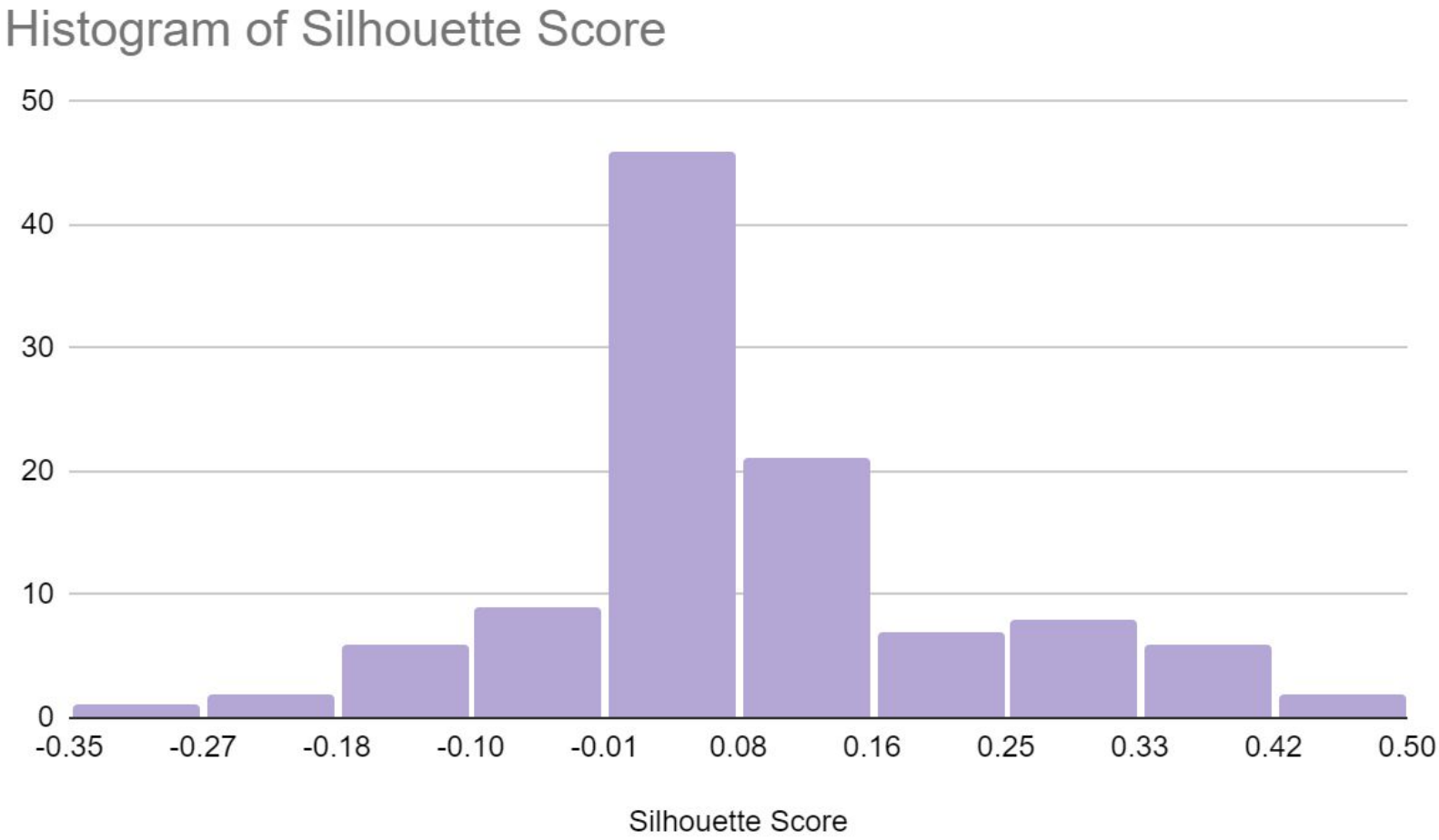

Histogram of Xie-Beni Index
25
20
15
10
5
0
0.00
0.15
0.30
0.45
0.60
0.75
0.90
1.05
Xie-Beni Index

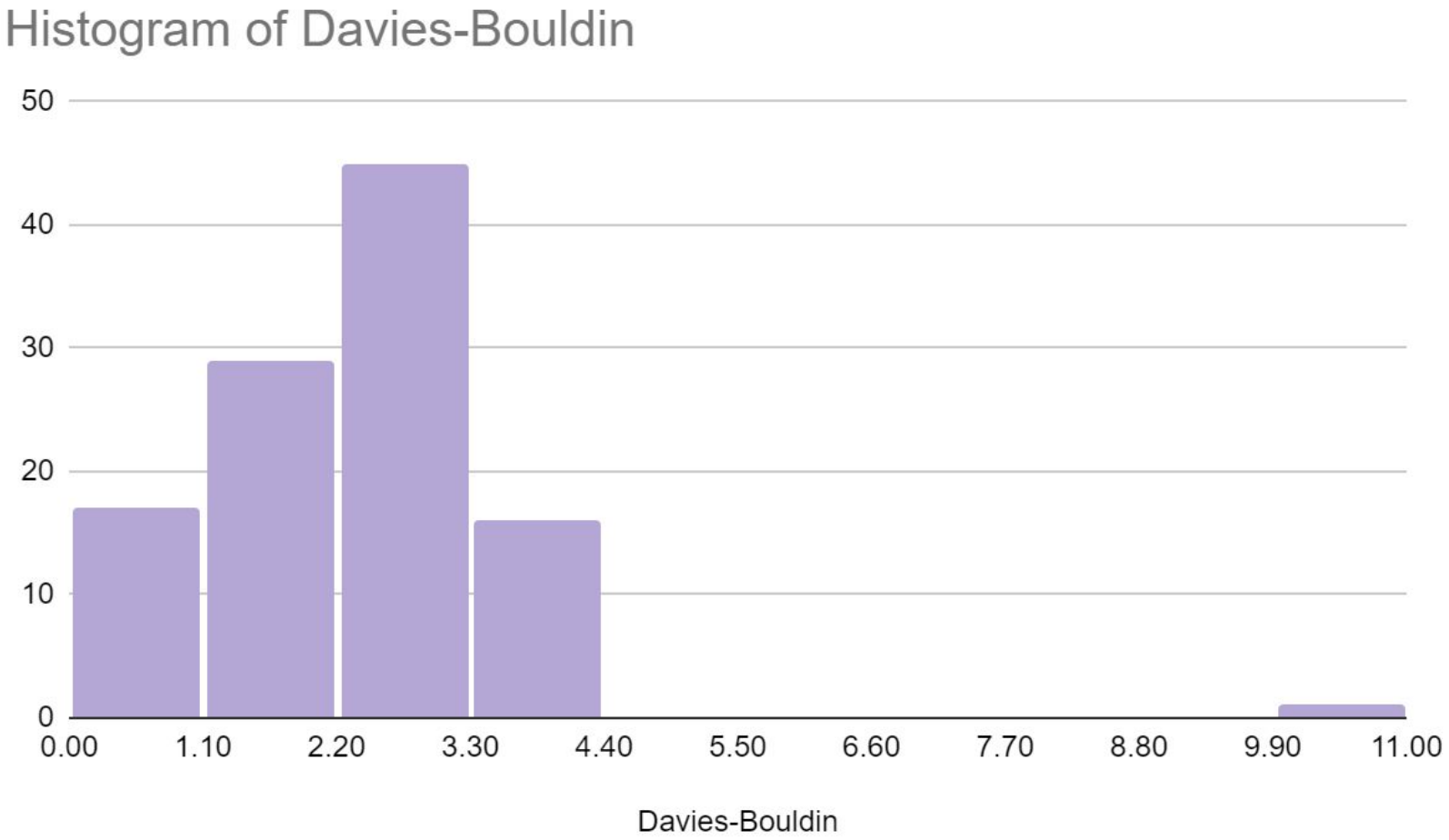
Privacy Law

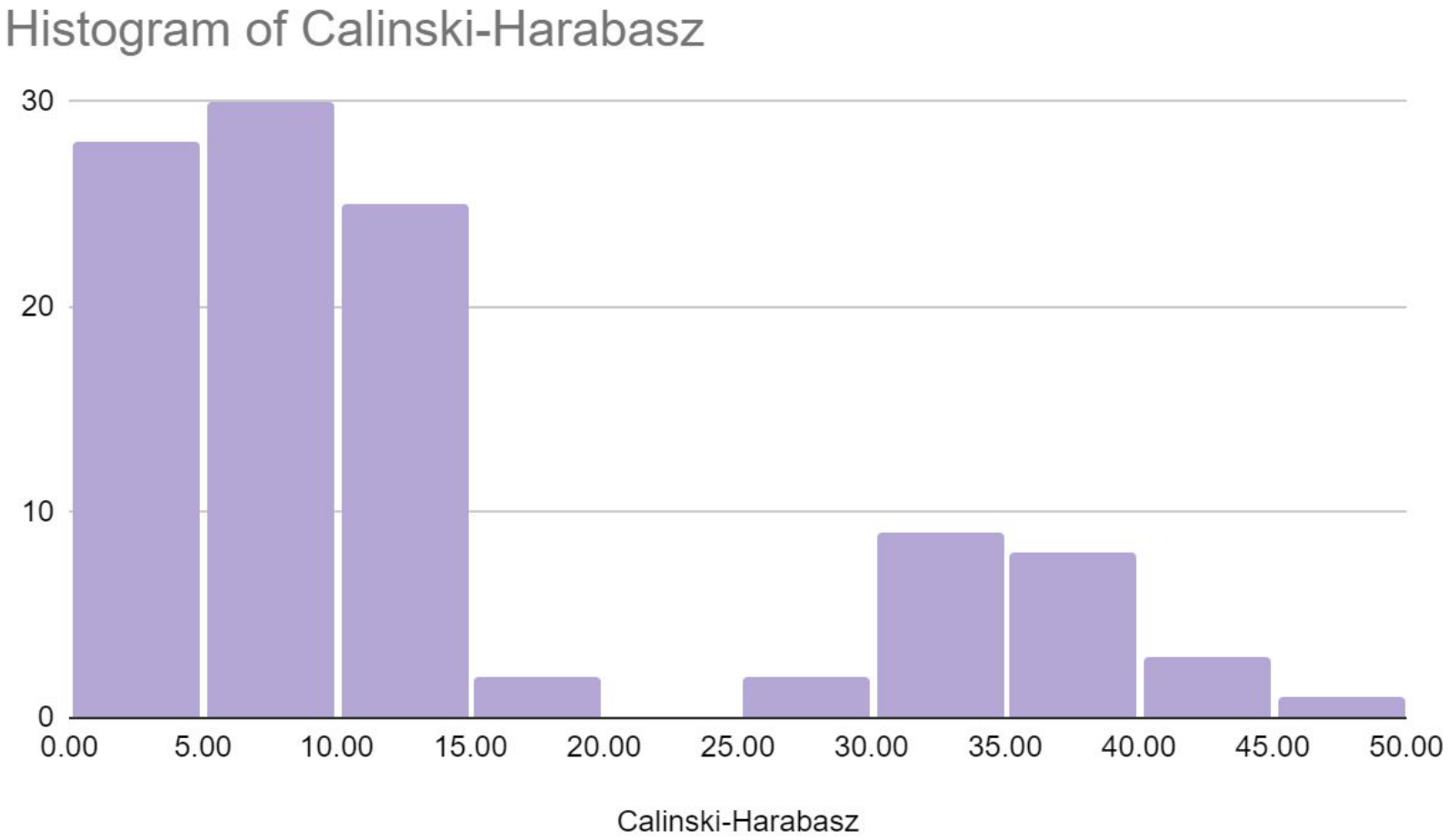

Histogram of Score
25
20
15
10
5
0
0.00
11.43
22.86
34.29
45.71
57.14
68.57
80.00
Score

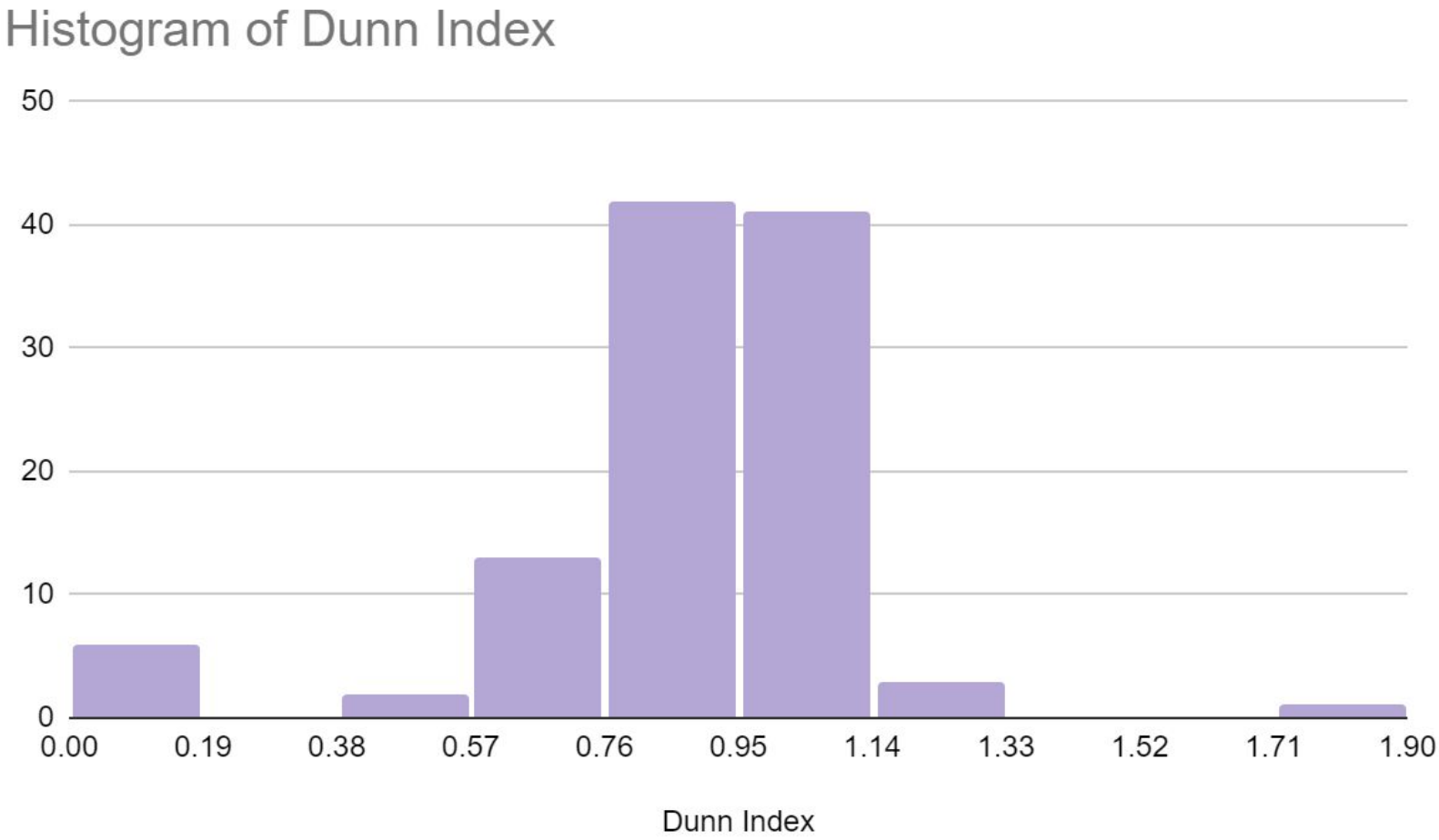
Privacy Law

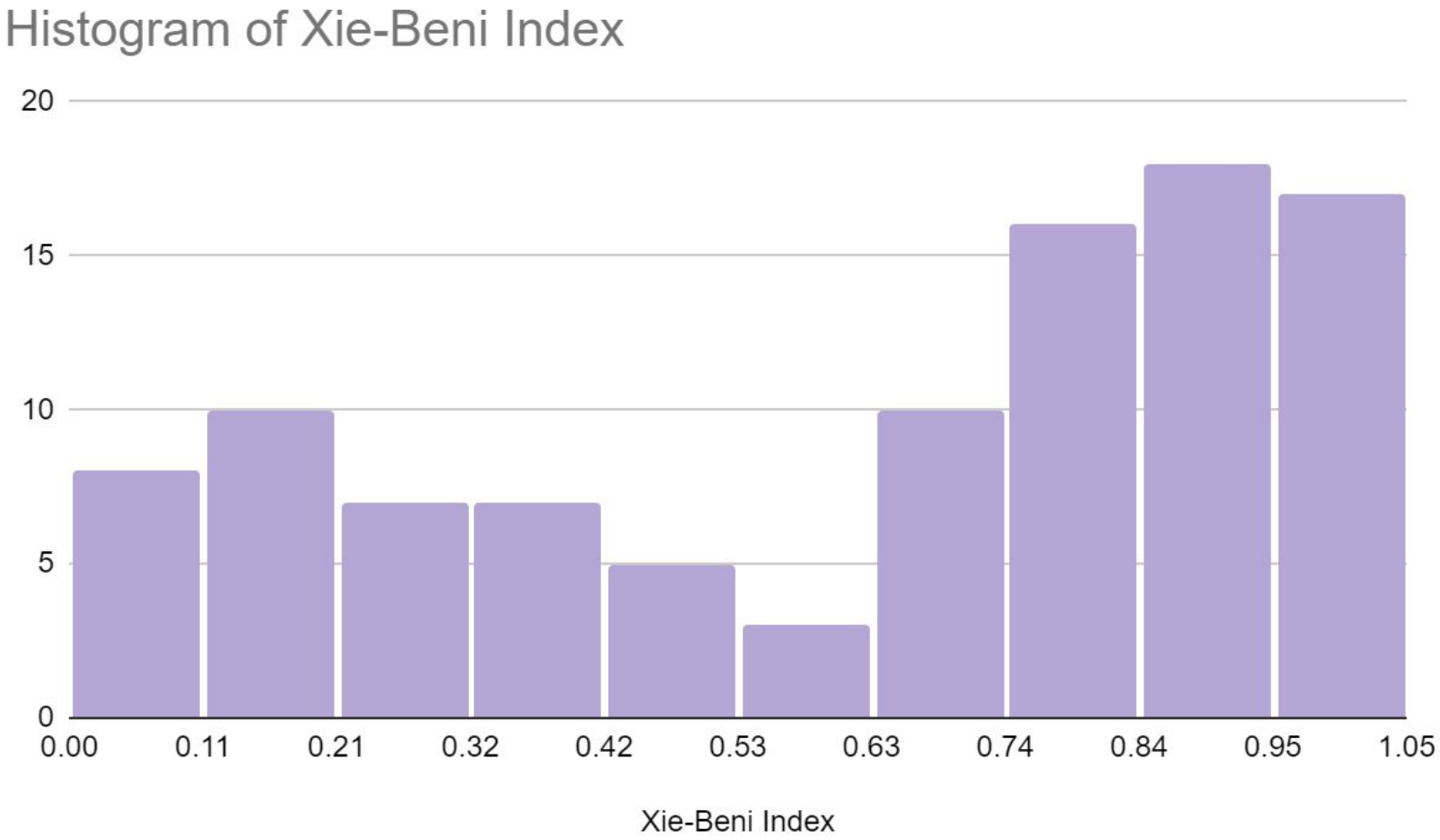
GovInfo Dataset

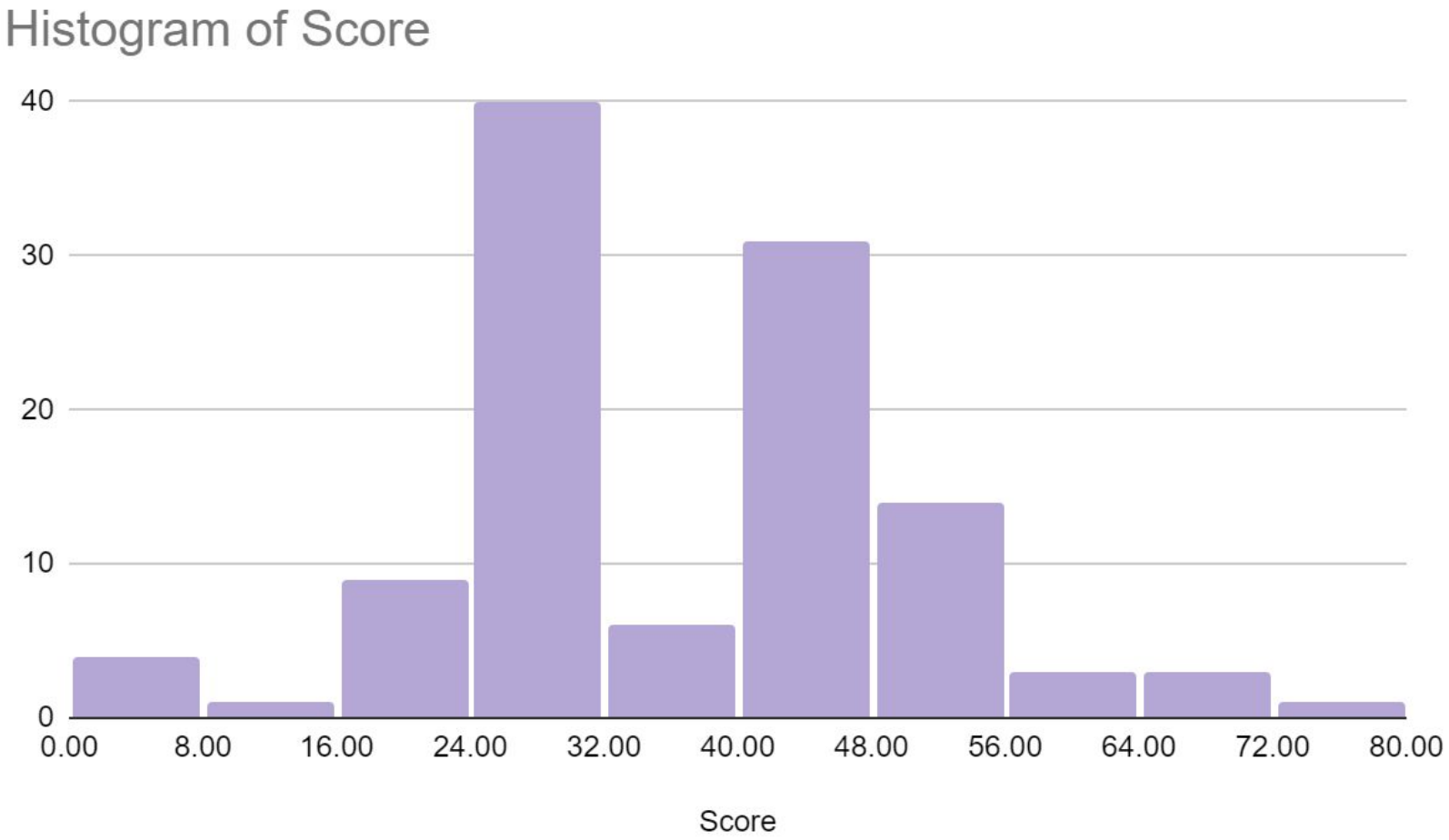

Histogram of Silhouette Score
50
40
30
20
10
0
-0.35
-0.27
-0.18
-0.10
-0.01
0.08
0.16
0.25
0.33
0.42
0.50
Silhouette Score


Histogram of Davies-Bouldin
50
40
30
20
10
0
0.00
1.10
2.20
3.30
4.40
5.50
6.60
7.70
8.80
9.90
11.00
Davies-Bouldin

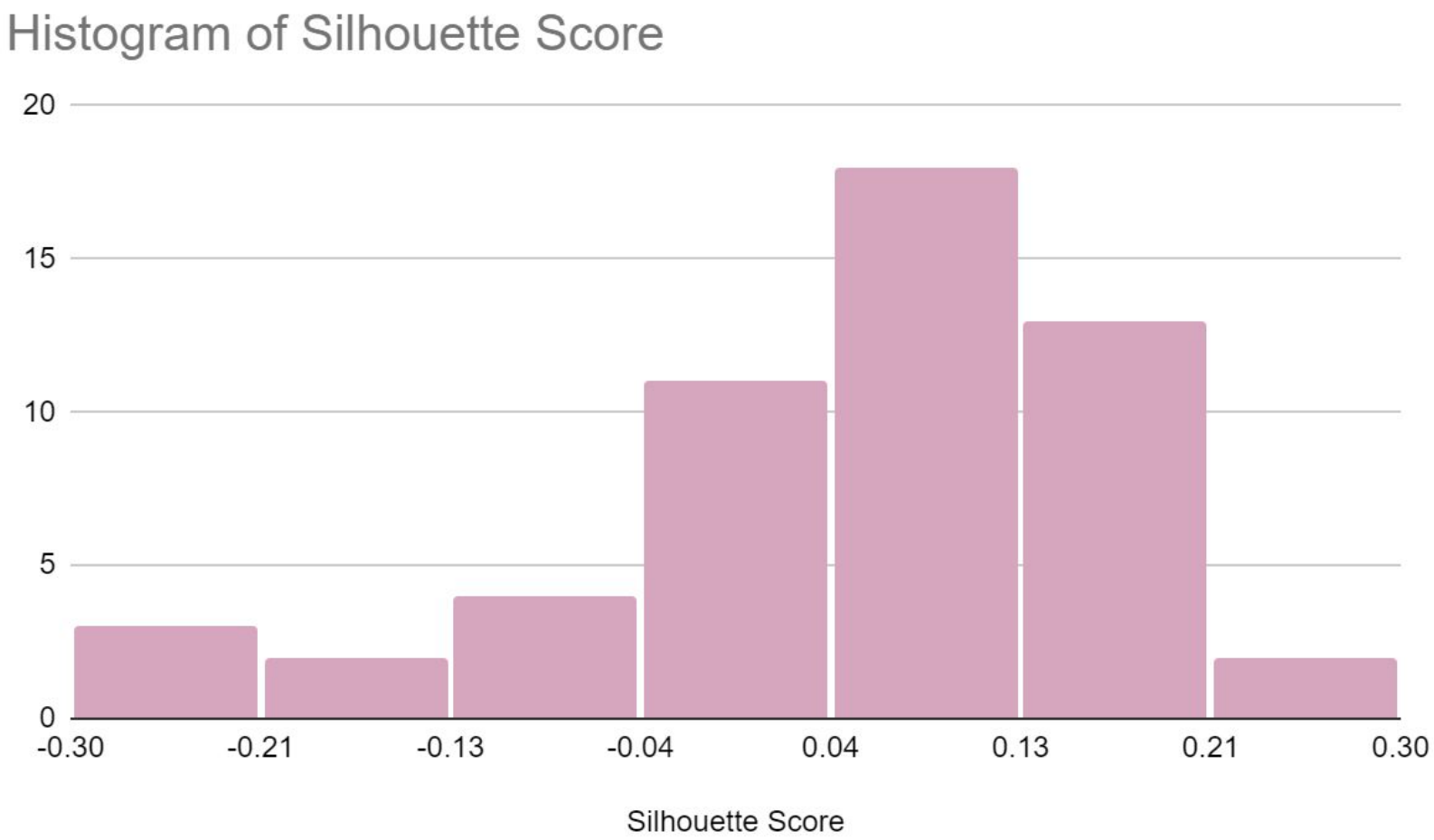

Histogram of Calinski-Harabasz
30
20
10
0
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
45.00
50.00
Calinski-Harabasz

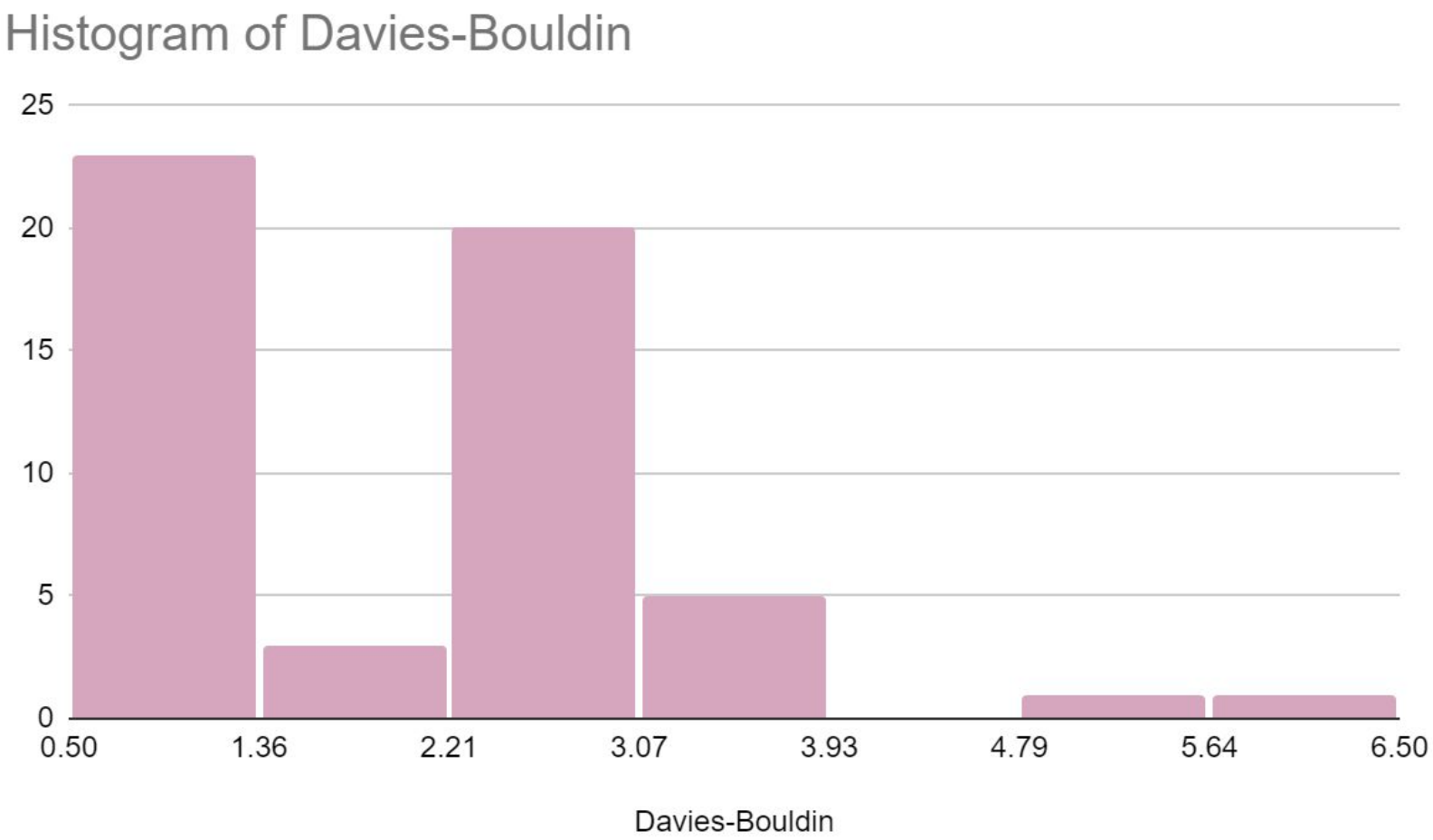

Histogram of Dunn Index
50
40
30
20
10
0
0.00
0.19
0.38
0.57
0.76
0.95
1.14
1.33
1.52
1.71
1.90
Dunn Index

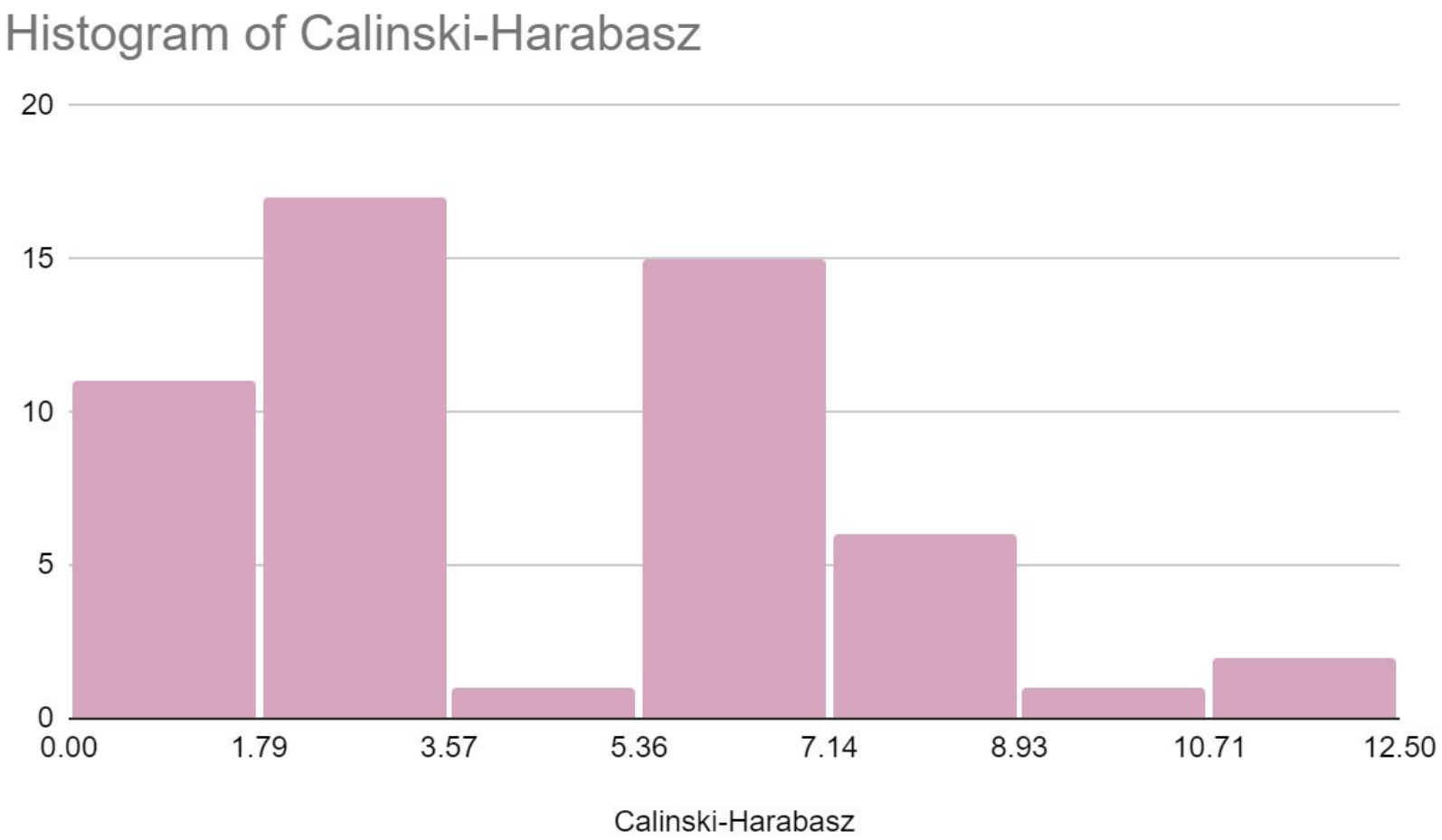

Histogram of Xie-Beni Index
20
15
10
5
0
0.00
0.11
0.21
0.32
0.42
0.53
0.63
0.74
0.84
0.95
1.05
Xie-Beni Index

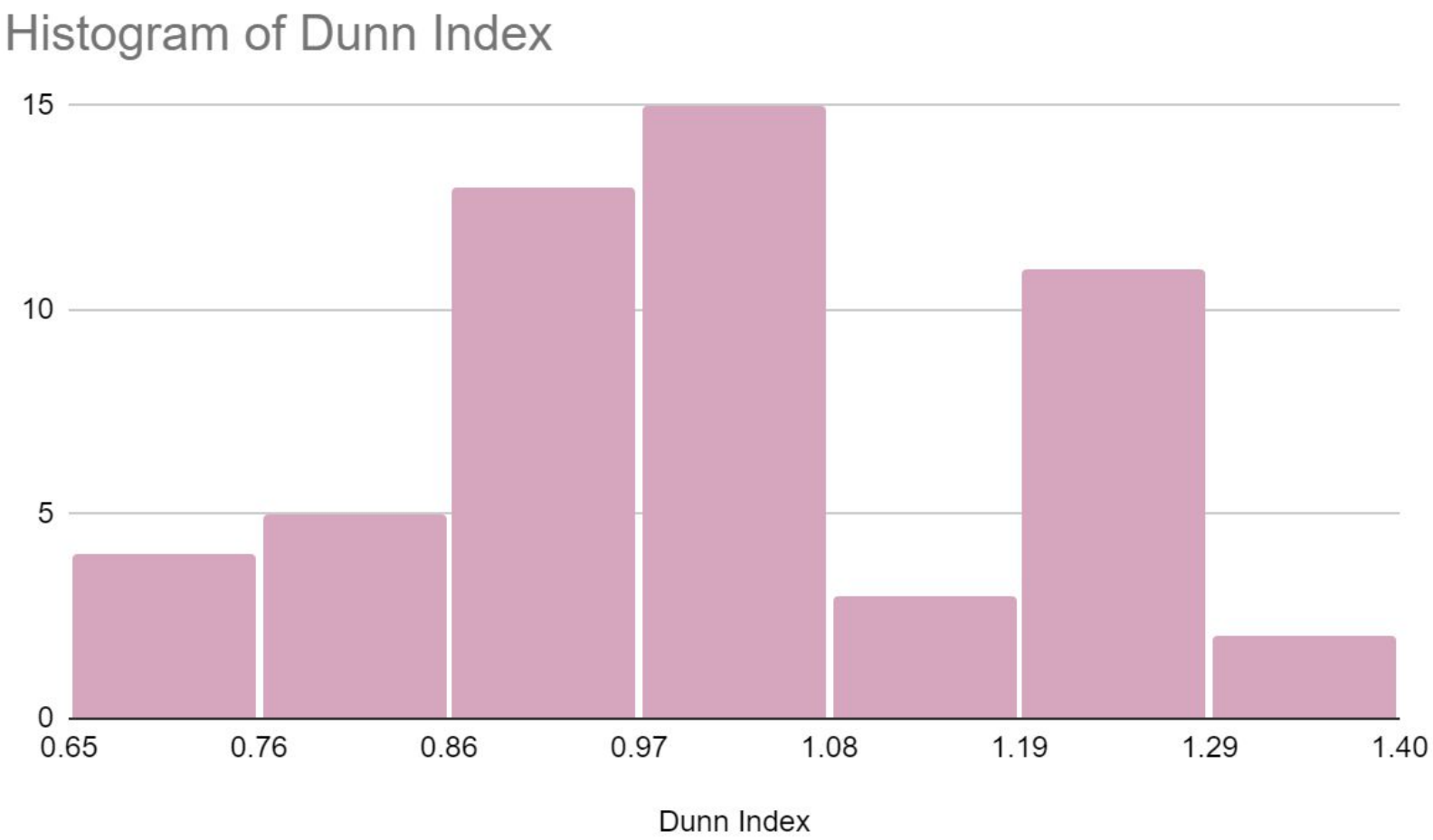

Histogram of Score
40
30
20
10
0
0.00
8.00
16.00
24.00
32.00
40.00
48.00
56.00
64.00
72.00
80.00
Score

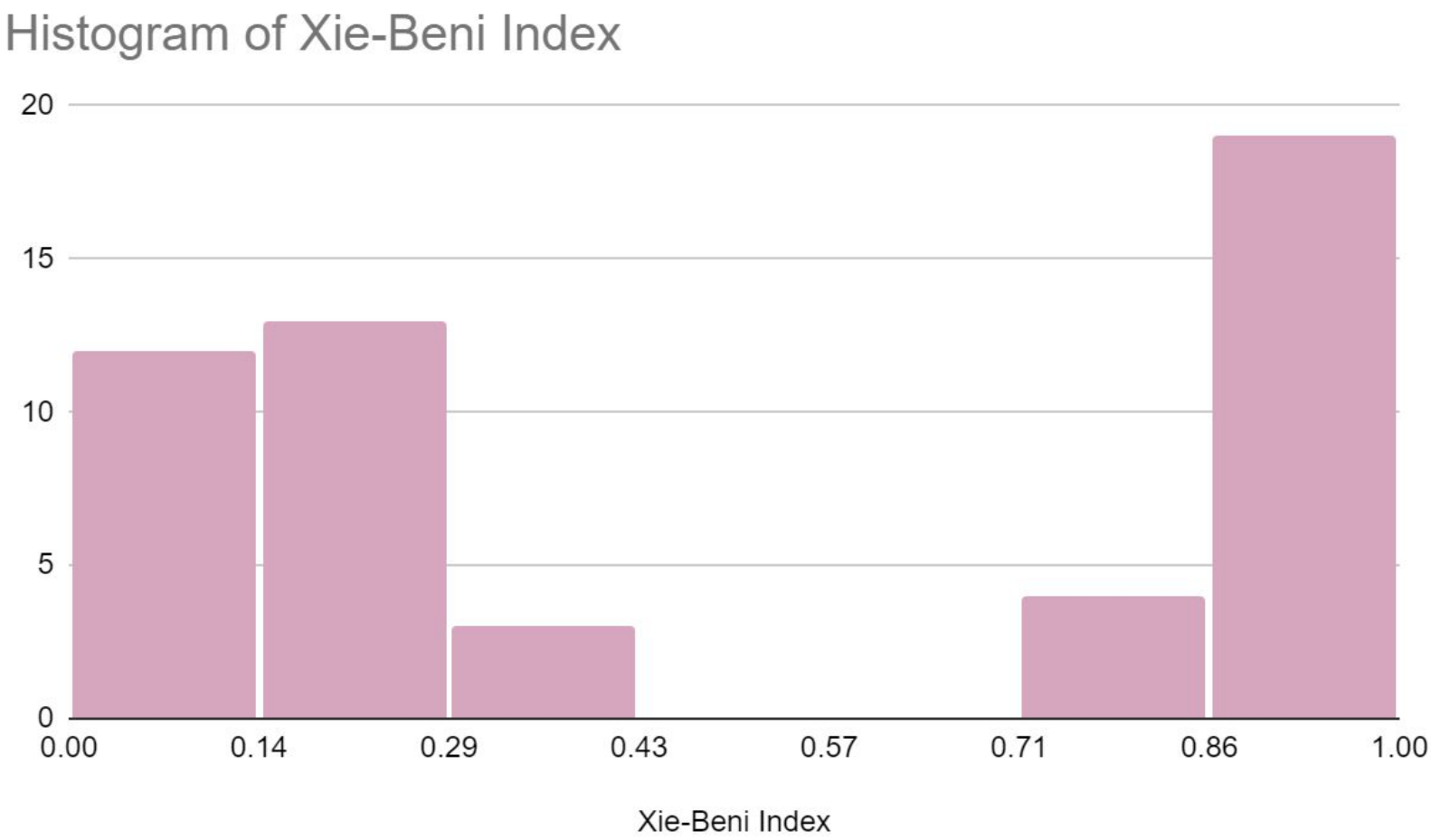
HCPSS Dataset

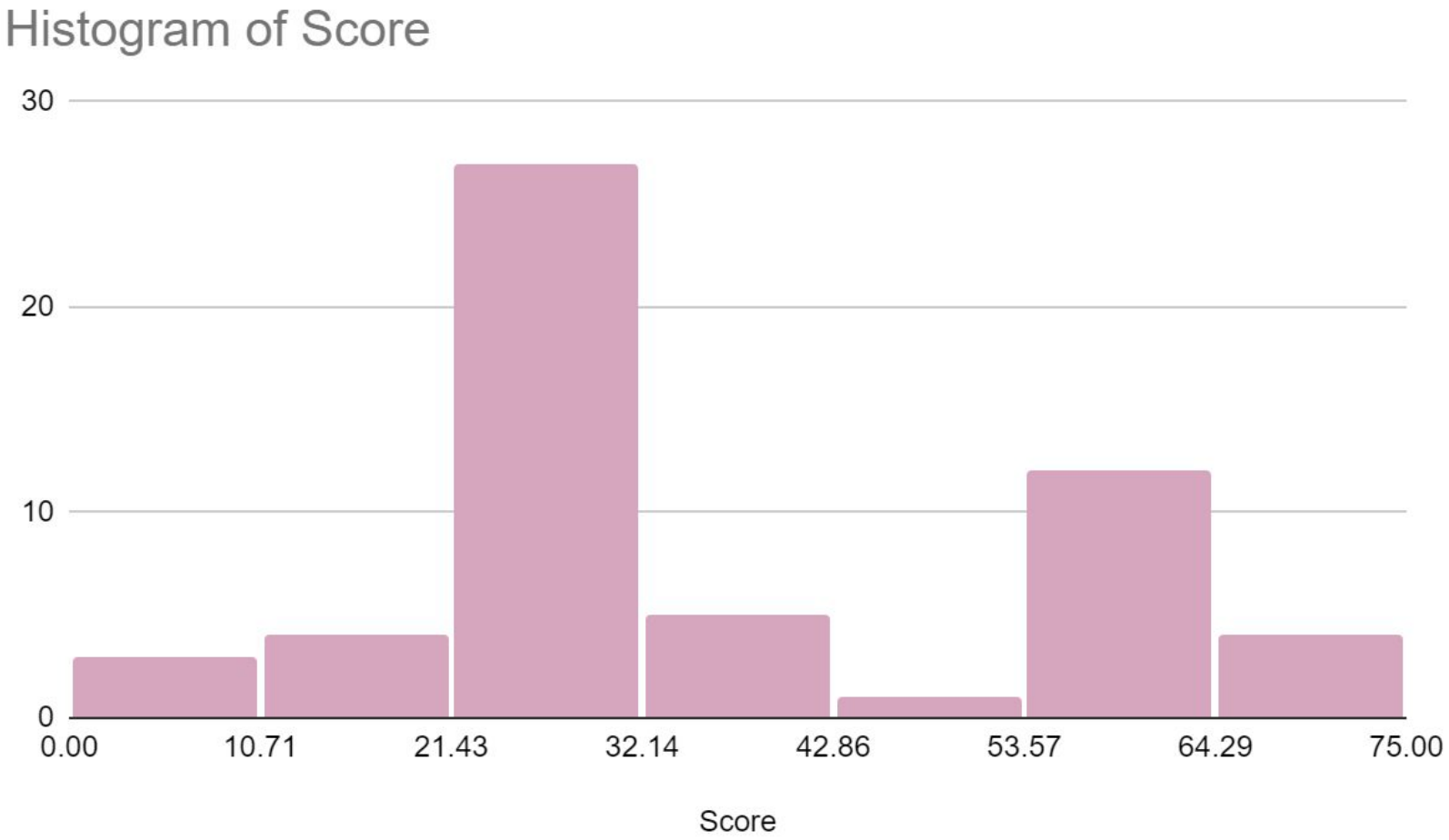

Histogram of Silhouette Score
20
15
10
5
0
-0.30
-0.21
-0.13
-0.04
0.04
0.13
0.21
0.30
Silhouette Score


Histogram of Davies-Bouldin
25
20
15
10
5
0
0.50
1.36
2.21
3.07
3.93
4.79
5.64
6.50
Davies-Bouldin


Histogram of Calinski-Harabasz
20
15
10
5
0
0.00
1.79
3.57
5.36
7.14
8.93
10.71
12.50
Calinski-Harabasz

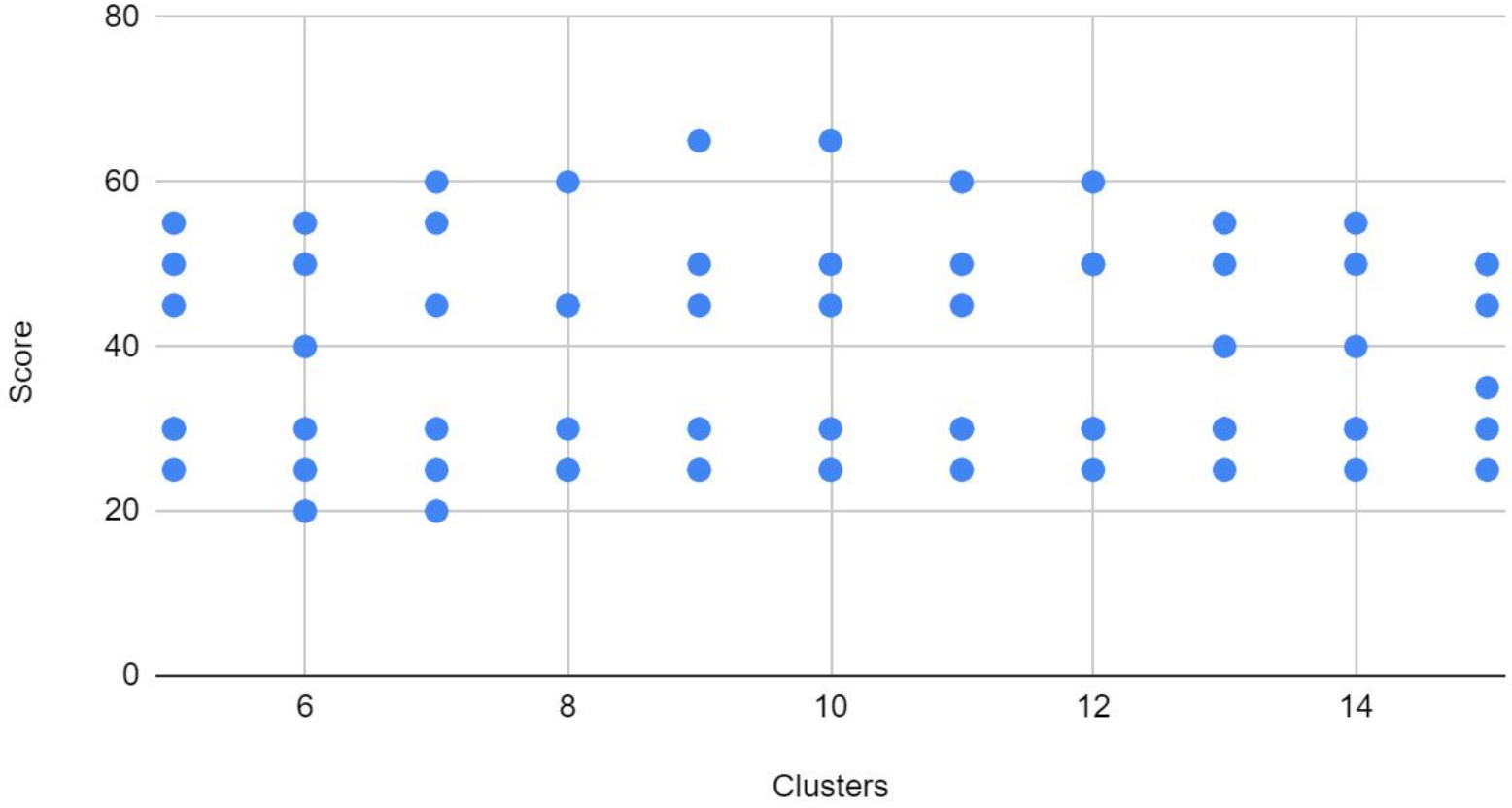

Histogram of Dunn Index
15
10
5
0
0.65
0.76
0.86
0.97
1.08
1.19
1.29
1.40
Dunn Index

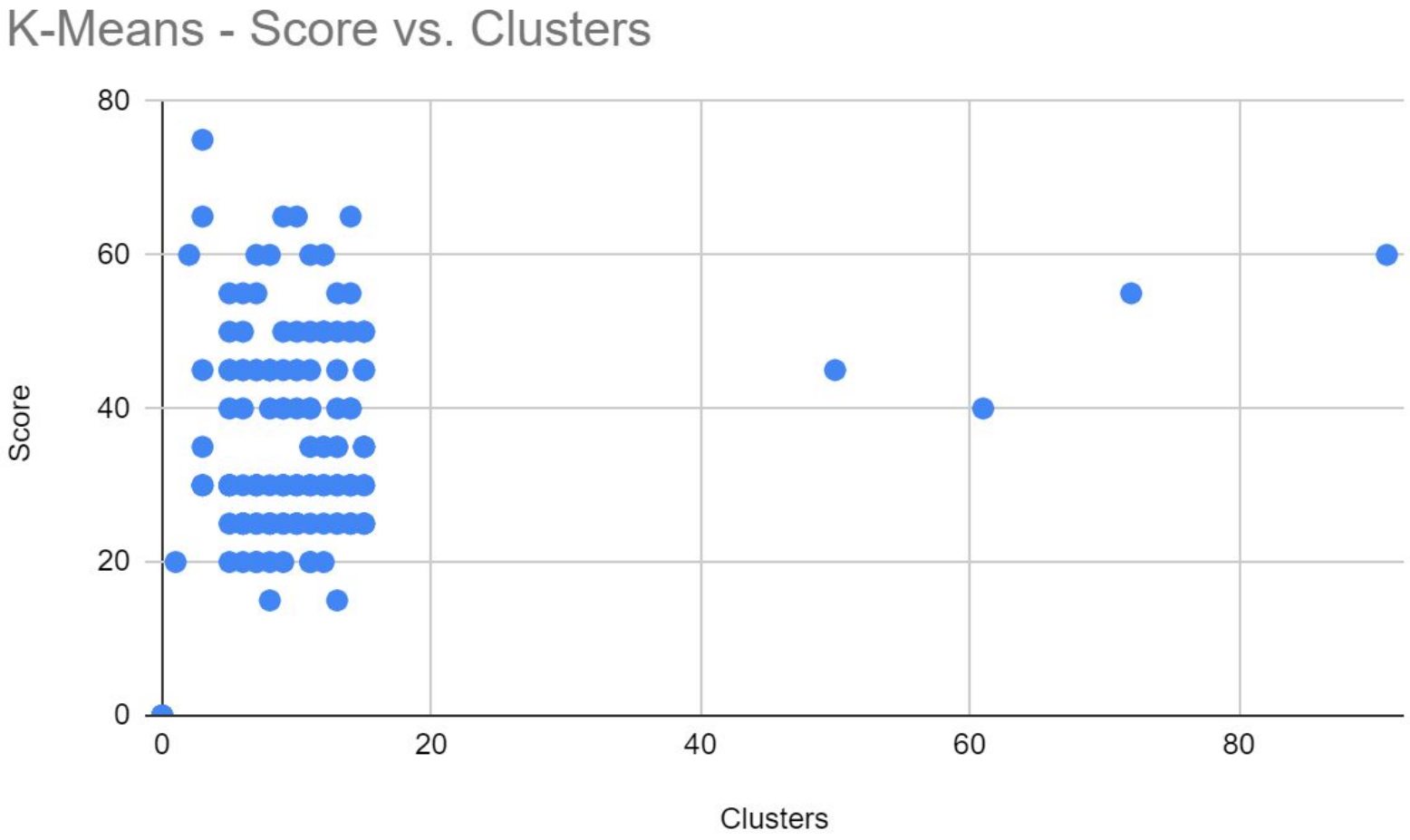

Histogram of Xie-Beni Index
20
15
10
5
0
0.00
0.14
0.29
0.43
0.57
0.71
0.86
1.00
Xie-Beni Index


Histogram of Score
30
20
10
0
0.00
10.71
21.43
32.14
42.86
53.57
64.29
75.00
Score

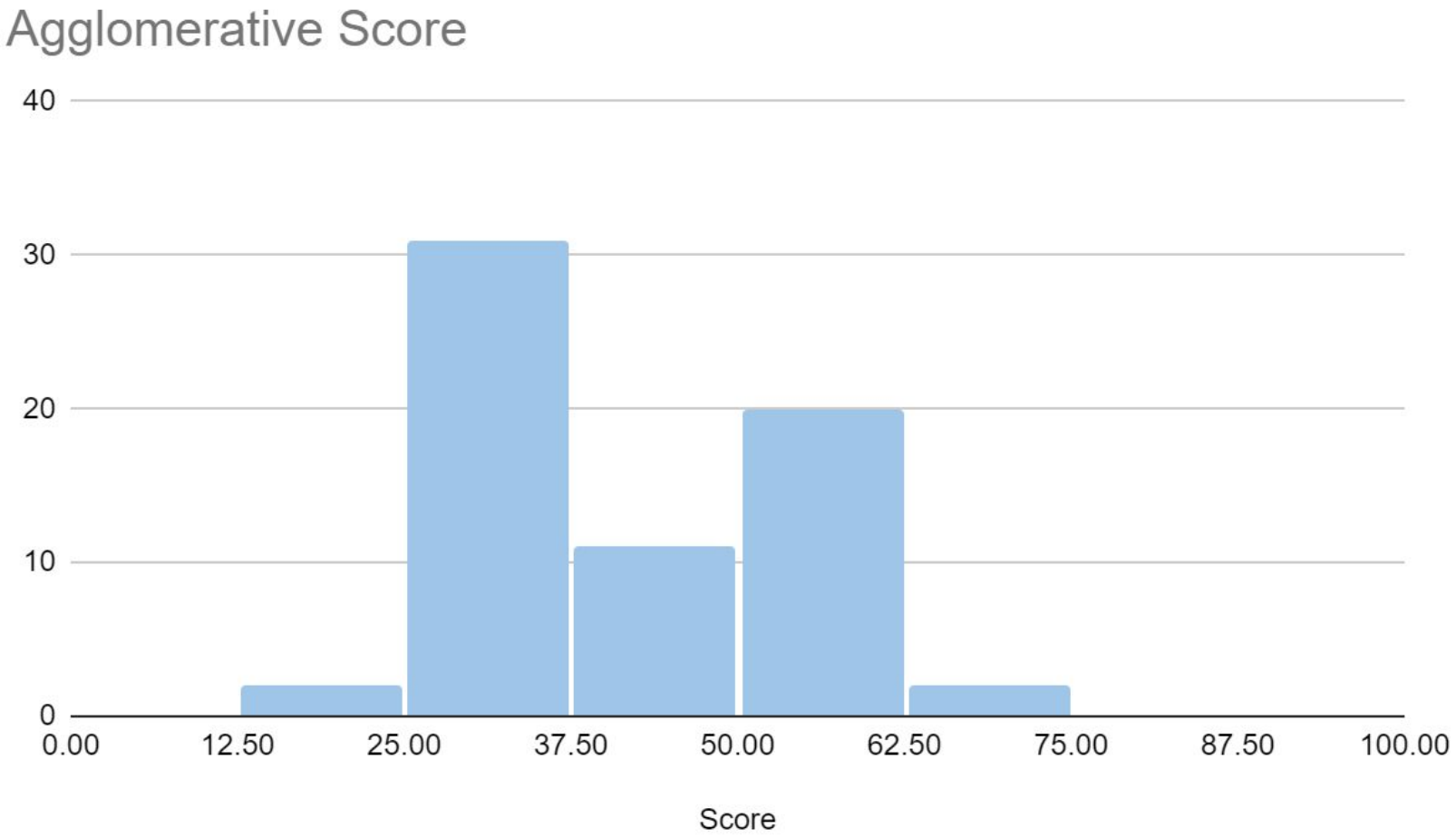
Num of Clusters x Score
By Algorithm

Agglomerative - Score vs. Clusters

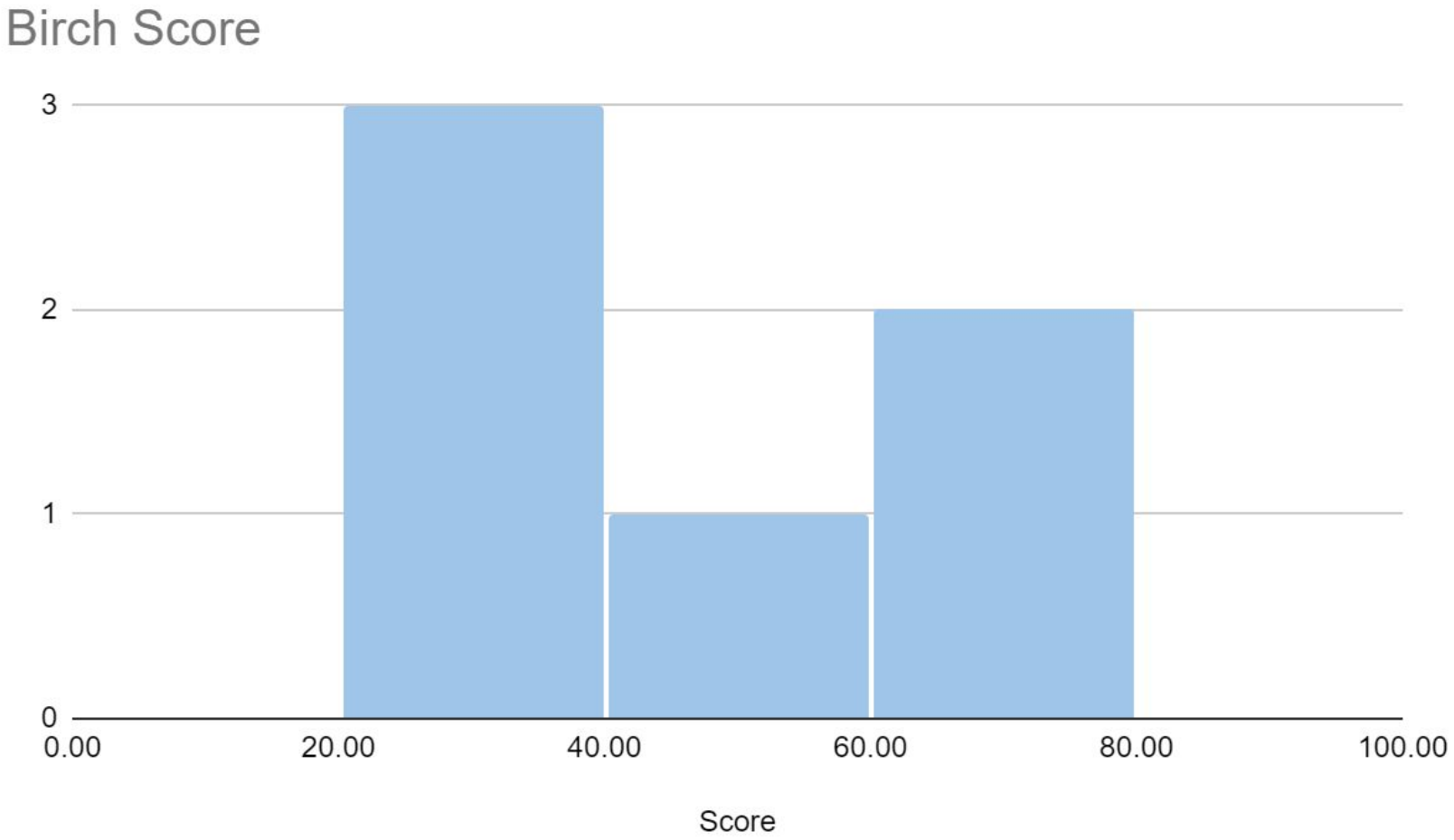



K-Means - Score vs. Clusters
Score
Clusters
0
20
40
60
80
0
20
40
60
80

Score x Algorithm

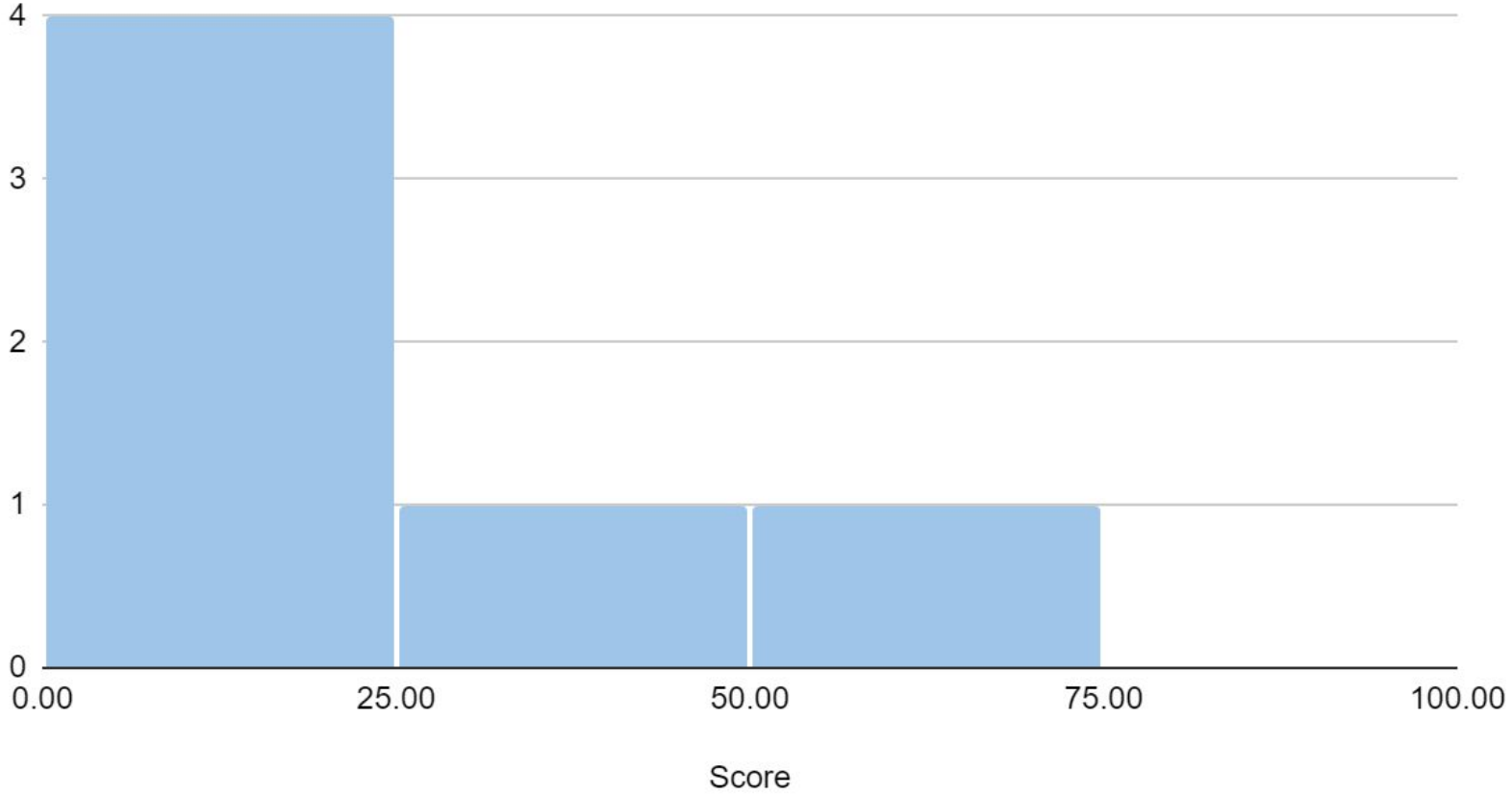


All Datasets

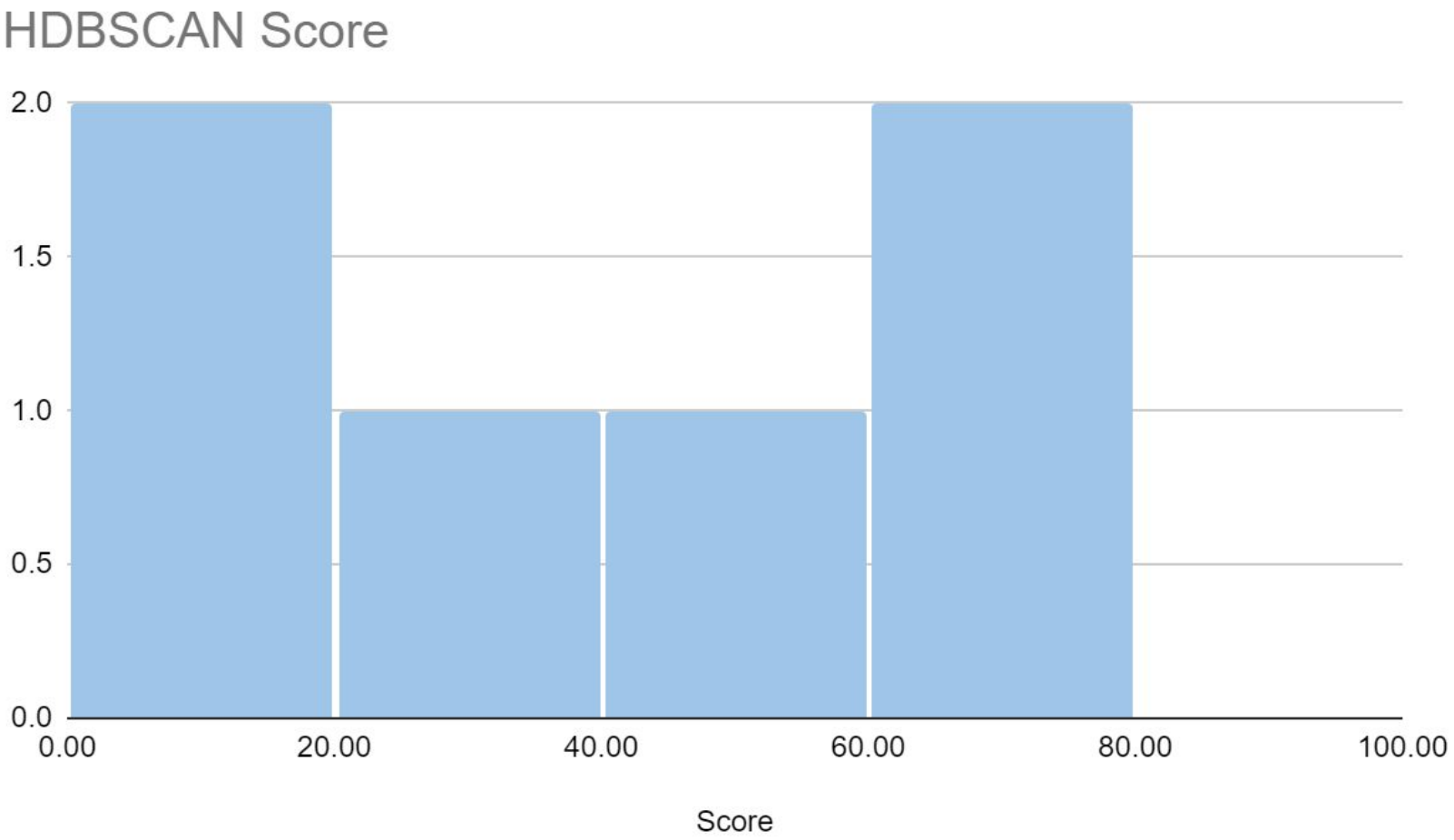


All Datasets

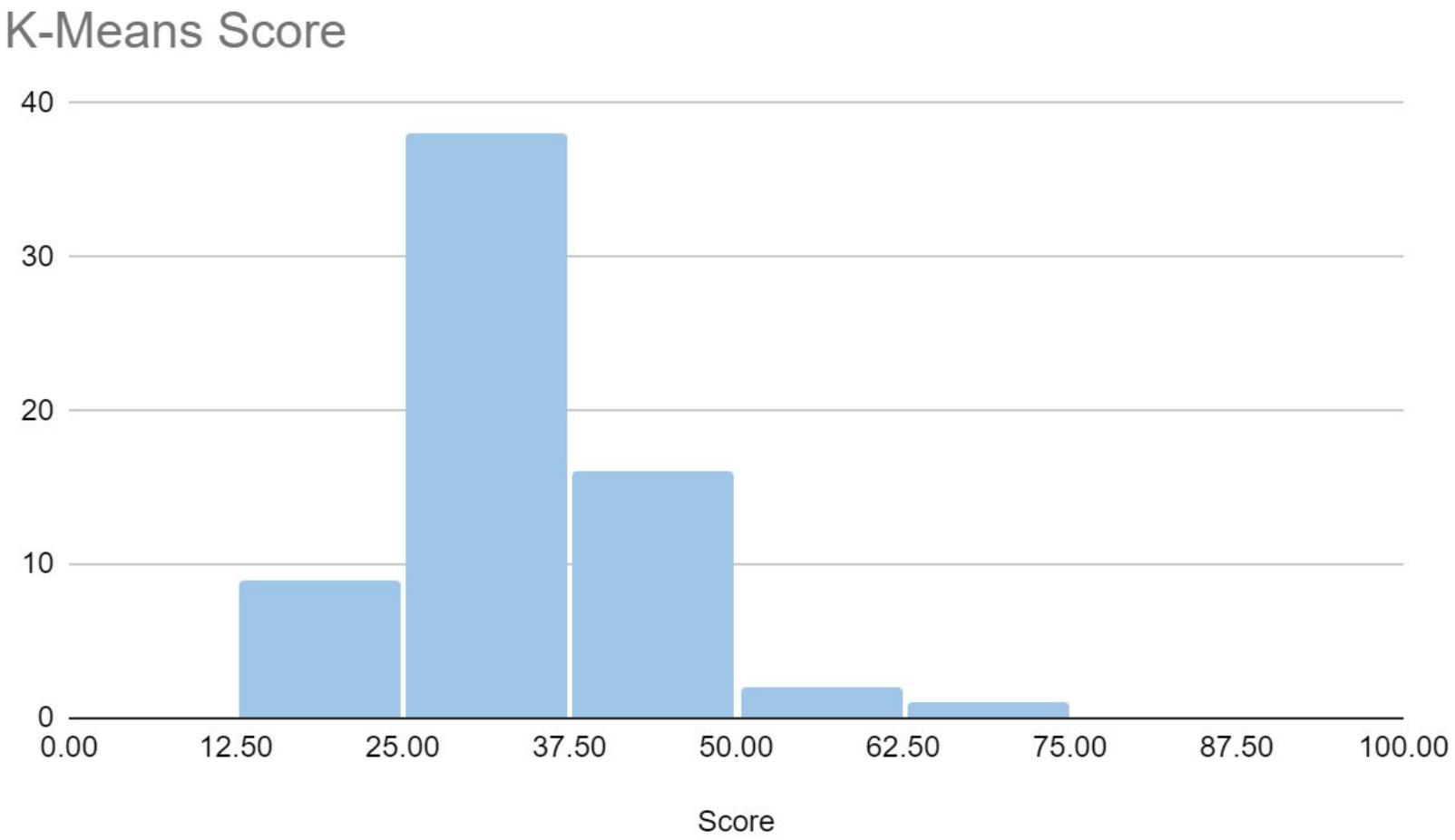
DBSCAN Score



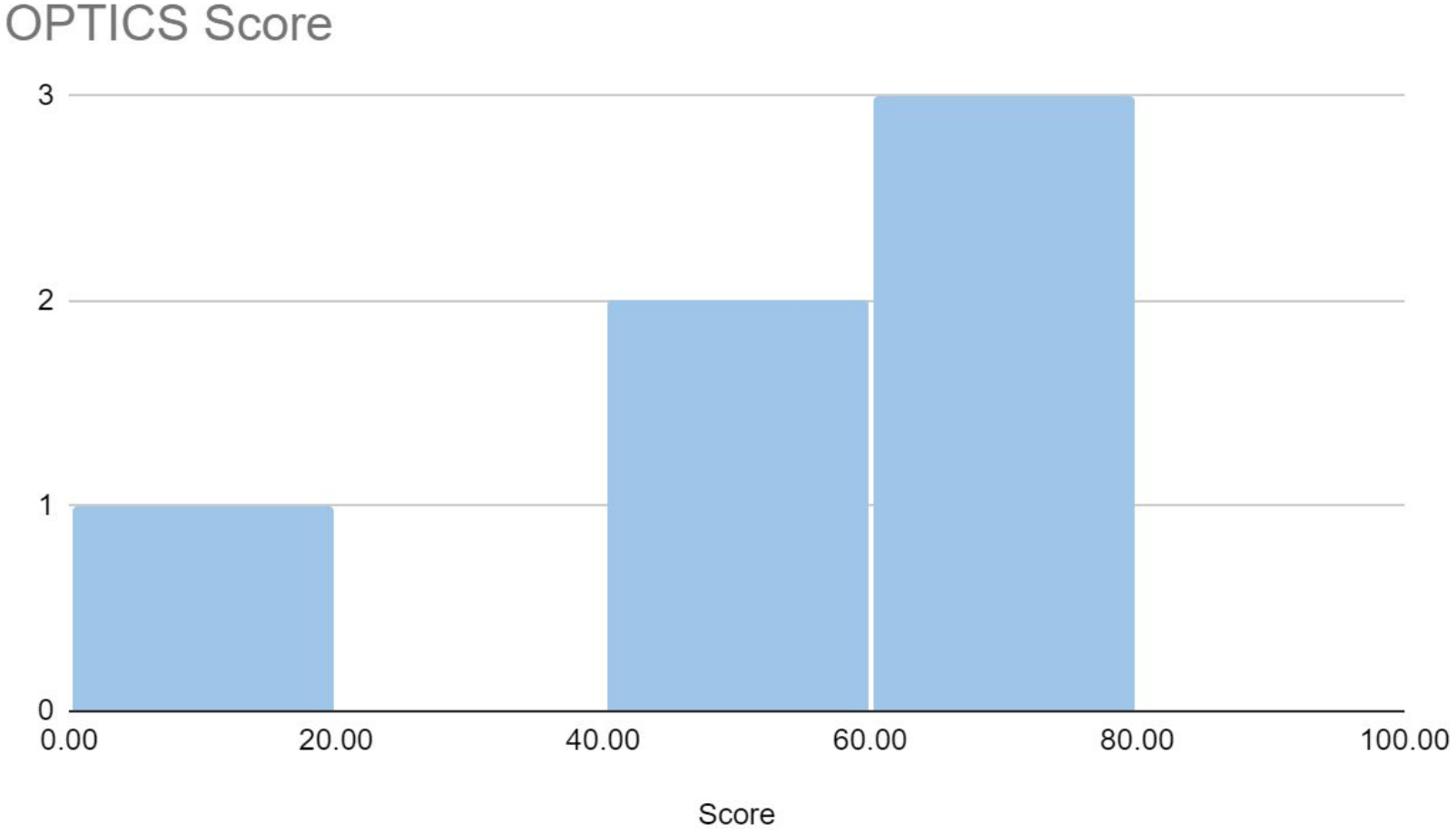
All Datasets



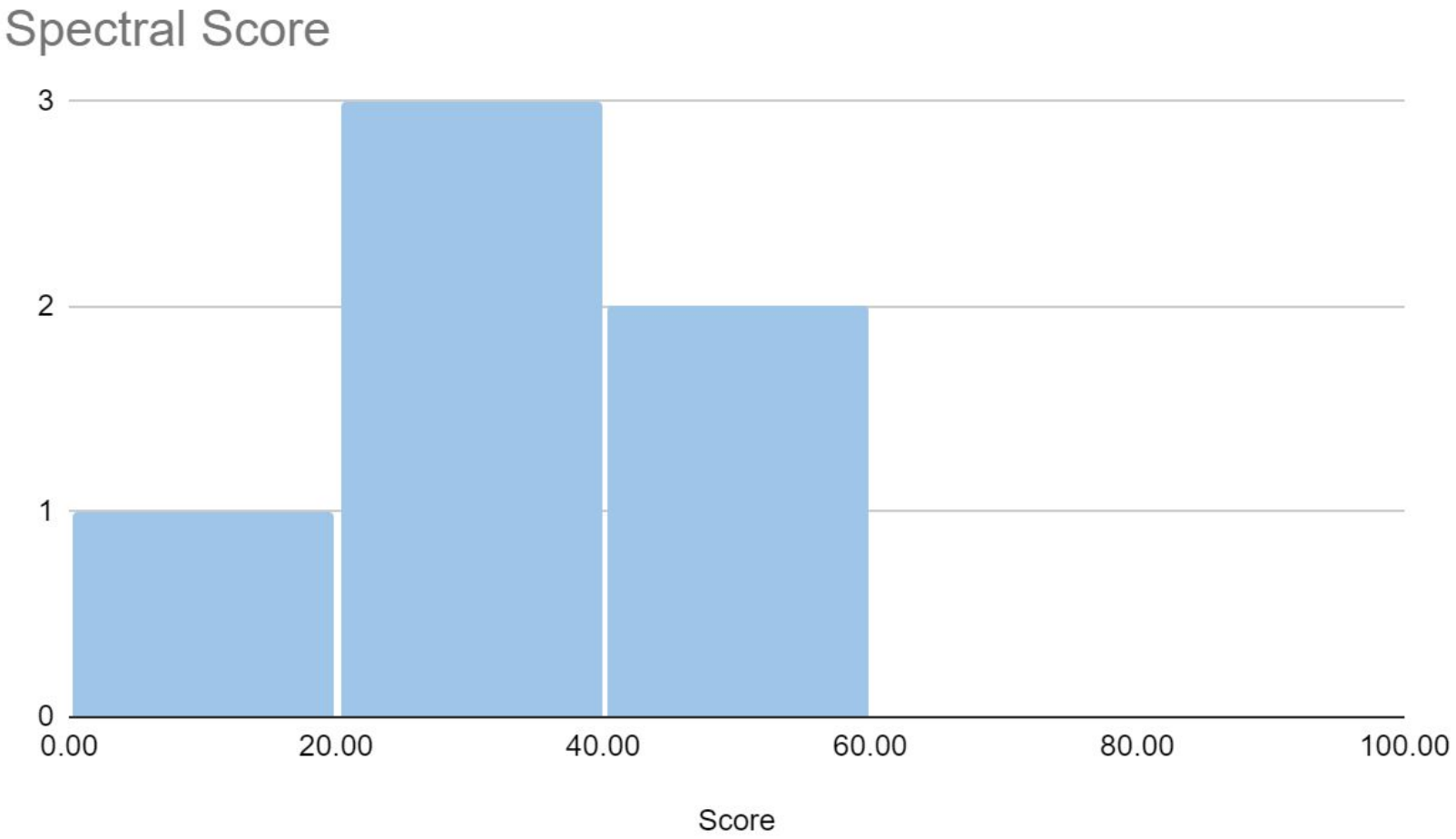
All Datasets



All Datasets



All Datasets



All Datasets