CS 439 25F
Student Assessments Graph Report

Group XX

Member Connor Fisher(cjf192)

Member Justin Shaw (jfs199)

Member Jonathan Maxwell (jhm141)

Date Oct 23, 2025

From Data Frames to Graphs

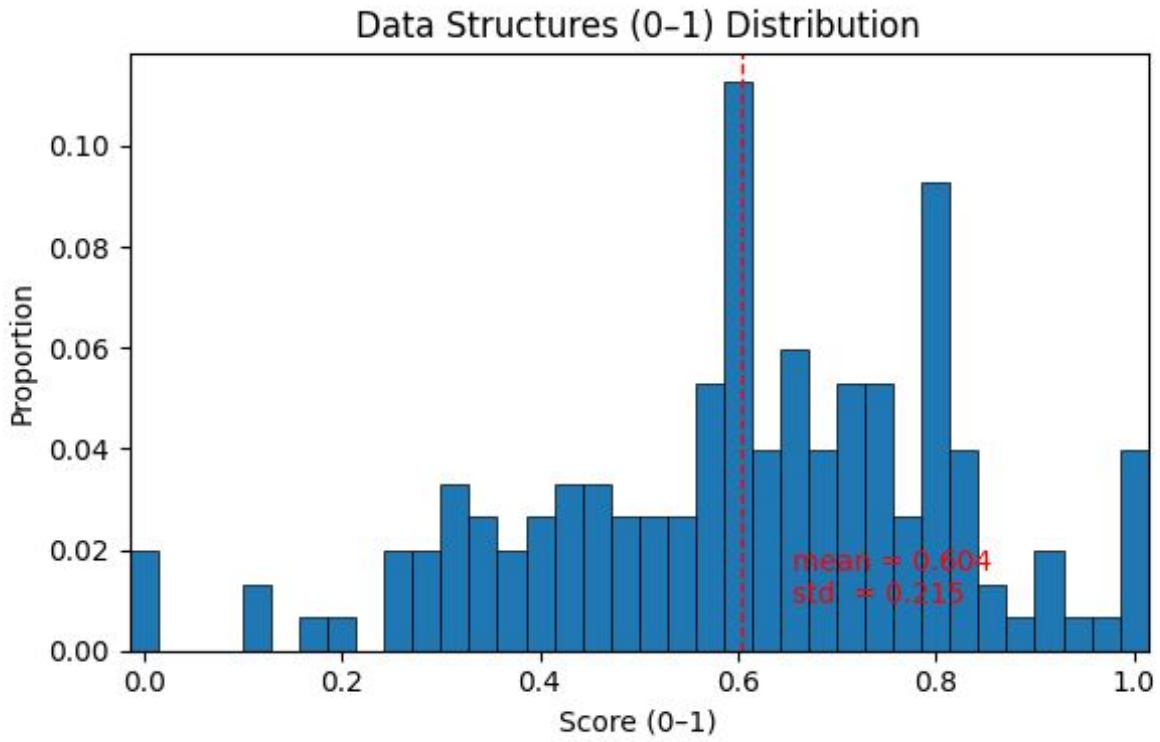
- **Input:** A Data Frame containing records of students' assessment scores in 15 Computer Science & Mathematics topics.
- **Output:** A collection of Graphs with different thresholds. The number of vertices corresponds to the number of students. Two students are connected by an edge if their distance is less than a prespecified threshold.
- **Method:**
  - **z-Score Normalization**: Scores in each of the 15 assessment topics are transformed to their corresponding z-scores in that topic, $z_x = (x - \mu)/\sigma$
  - **Euclidean Distance**: A pair of students is weighted by the Euclidean Distance between their records' z-scores, $d(A, B)^2 = \sum(A_i - B_i)^2$
  - **Weight Distance Distribution**: Compute the mean and standard deviation, and select the number of standard deviations to be used as a guide for edge selection.
  - **Edge Selection**: A pair of students is connected by an edge if their Euclidean weight distance is less than a specified number of standard deviations less than the mean.

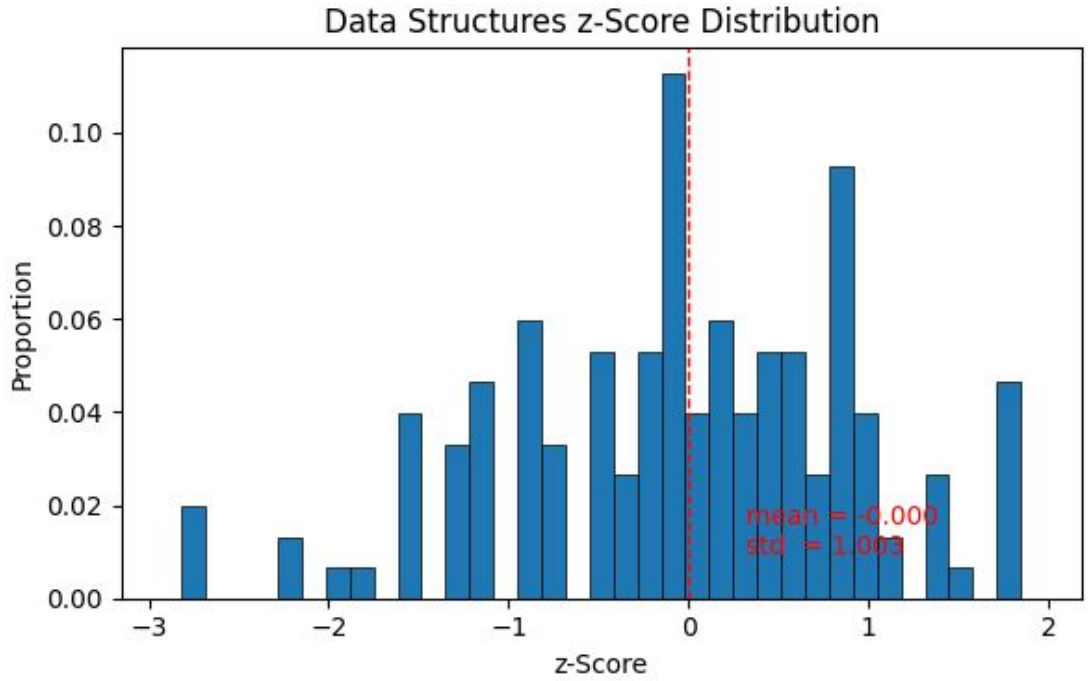
Data Set Description

- The Student Assessment csv file is a flat table where each record is one student's assessment submission at a specific time.
- Each row = one student's performance on one submission (one student, one timestamp).
- Columns:
  - Identifiers & context: timestamp, netid, ruid, section, role, major
  - Skill scores (numeric): data_structures, calculus_and_linear_algebra, probability_and_statistics, data_visualization, sql, python_scripting, jupyter_notebook, regression, programming_languages, algorithms, complexity_measures, visualization_tools, massive_data_processing
  - Aggregate: total_score (sum of the skill scores)
- Values: IDs/context = text/categorical; timestamp = datetime; skill fields = non-negative numbers; total_score = numeric row sum.

Data Set Description

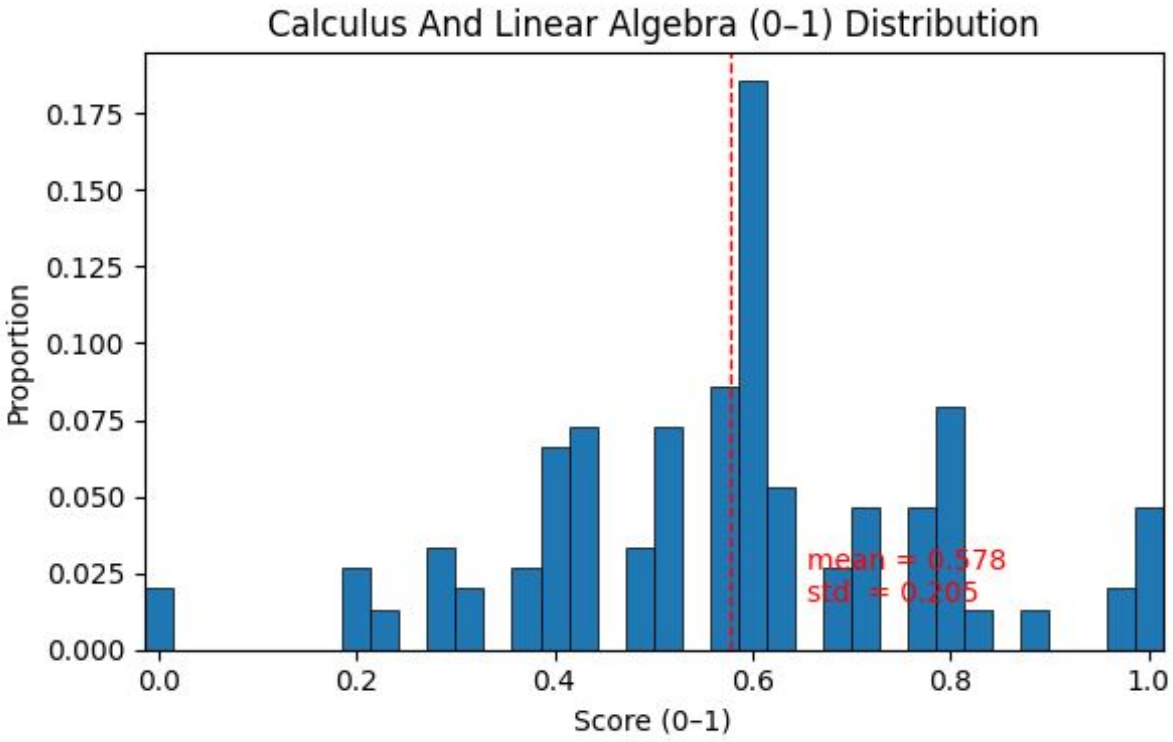
Column: Data Structures

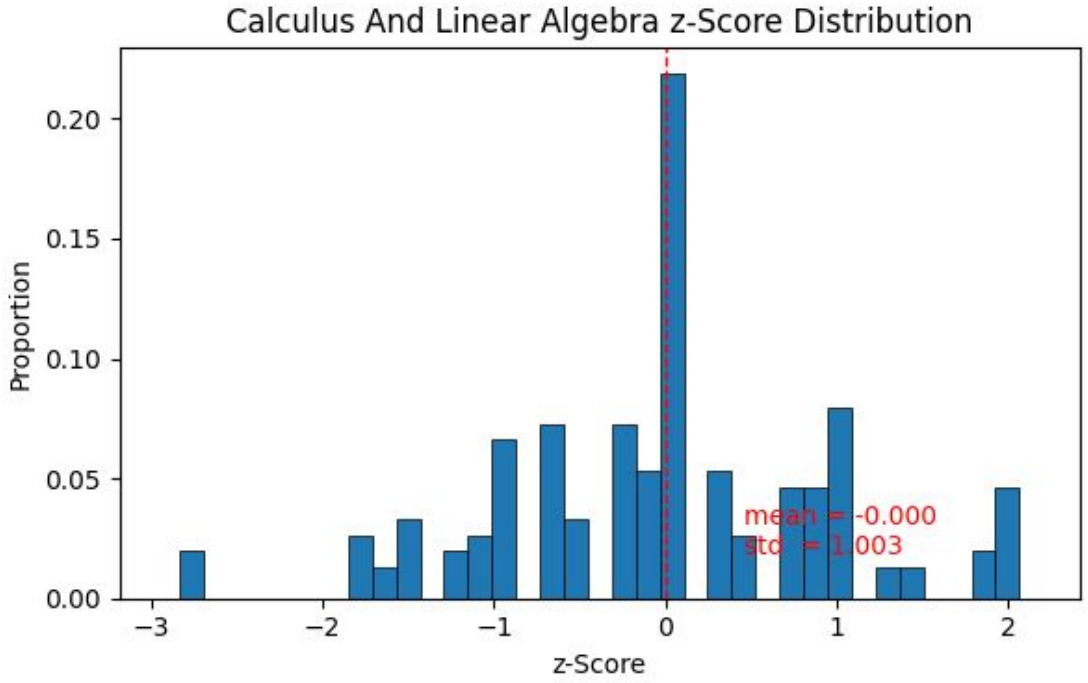




Data Set Description

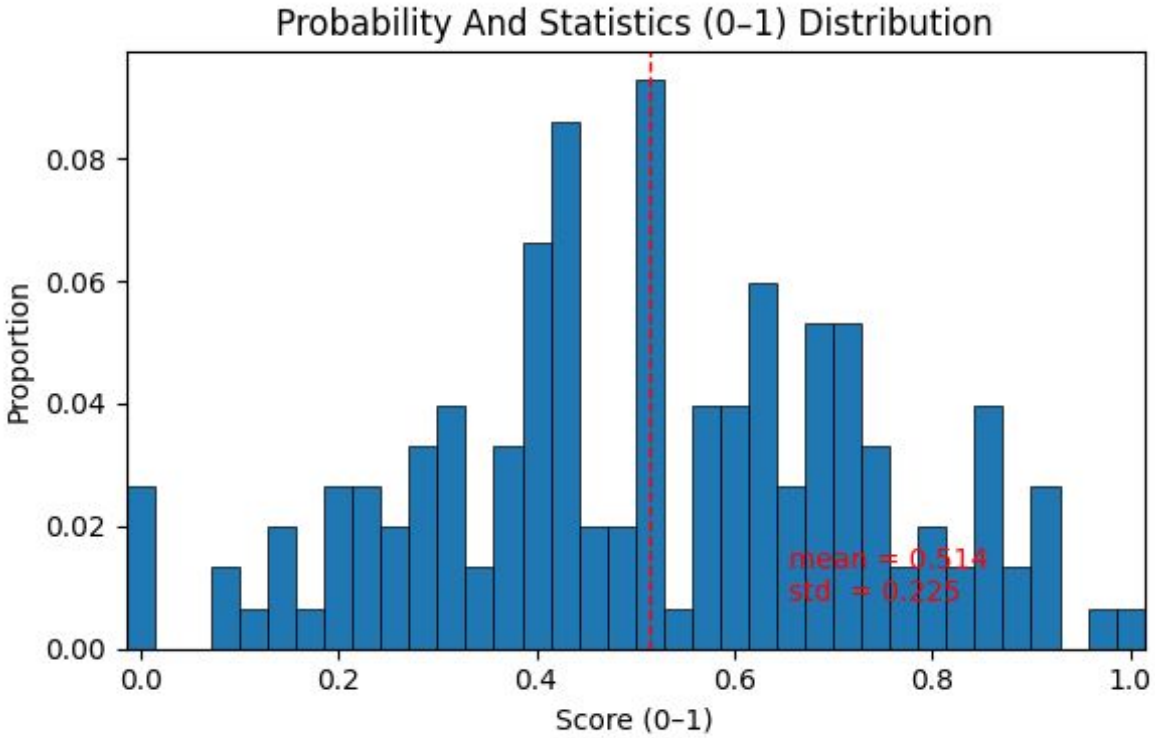
Column: Calculus And linear Algebra

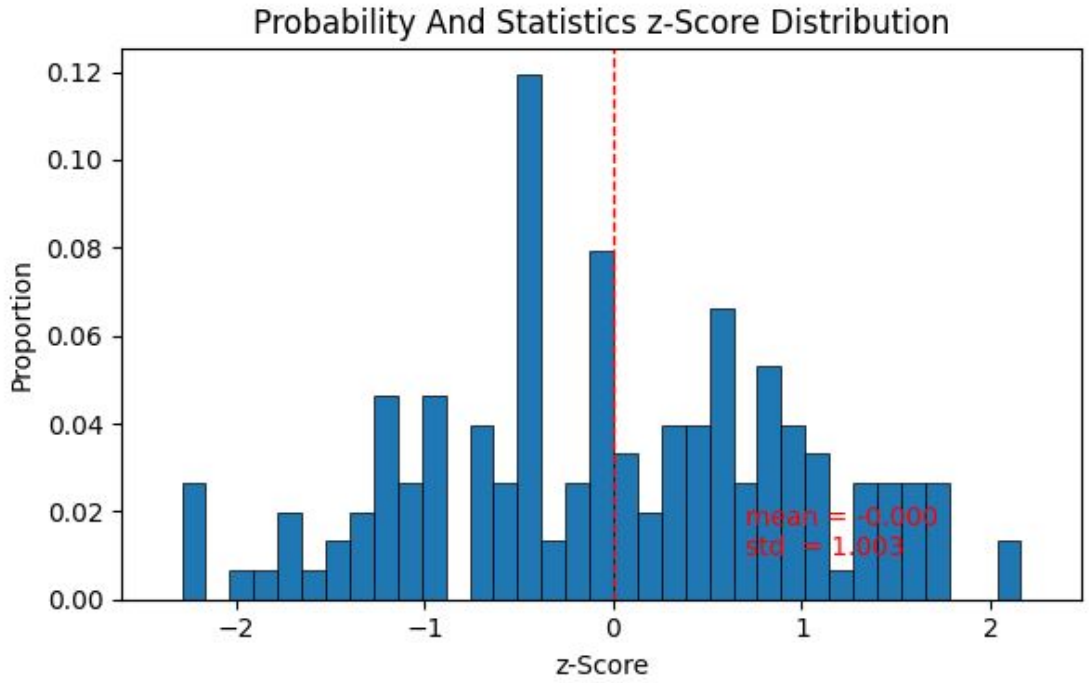




Data Set Description

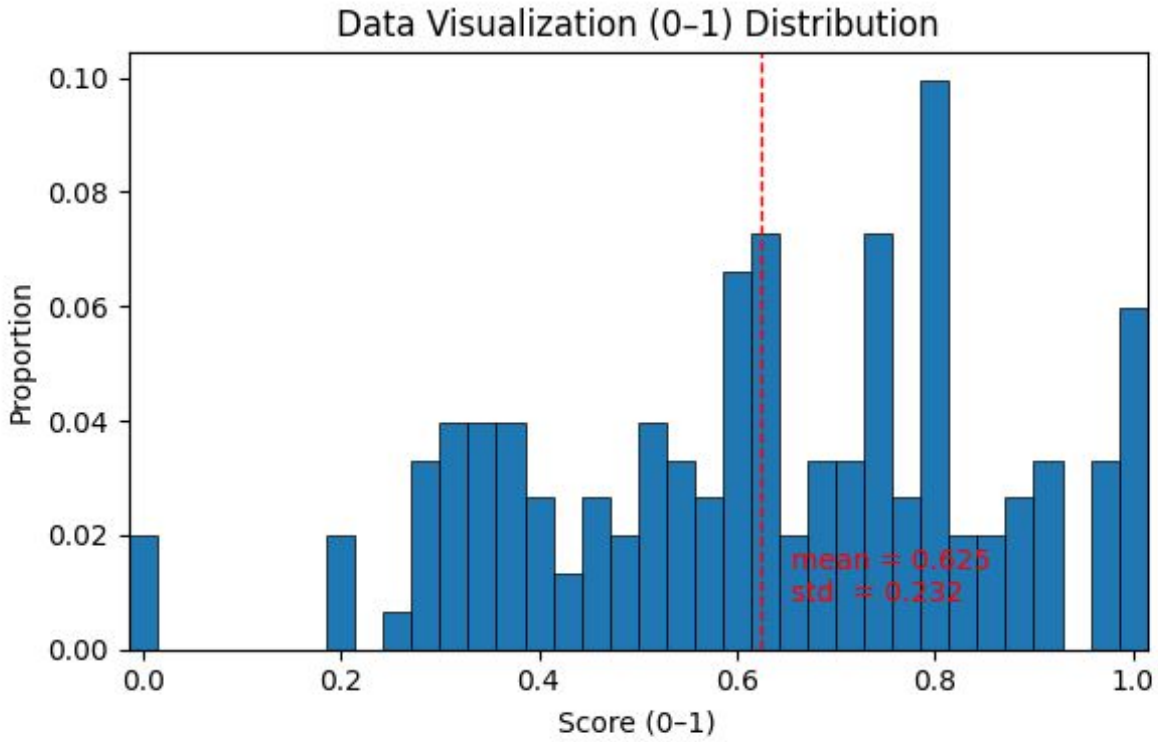
Column: Probability And Statistics

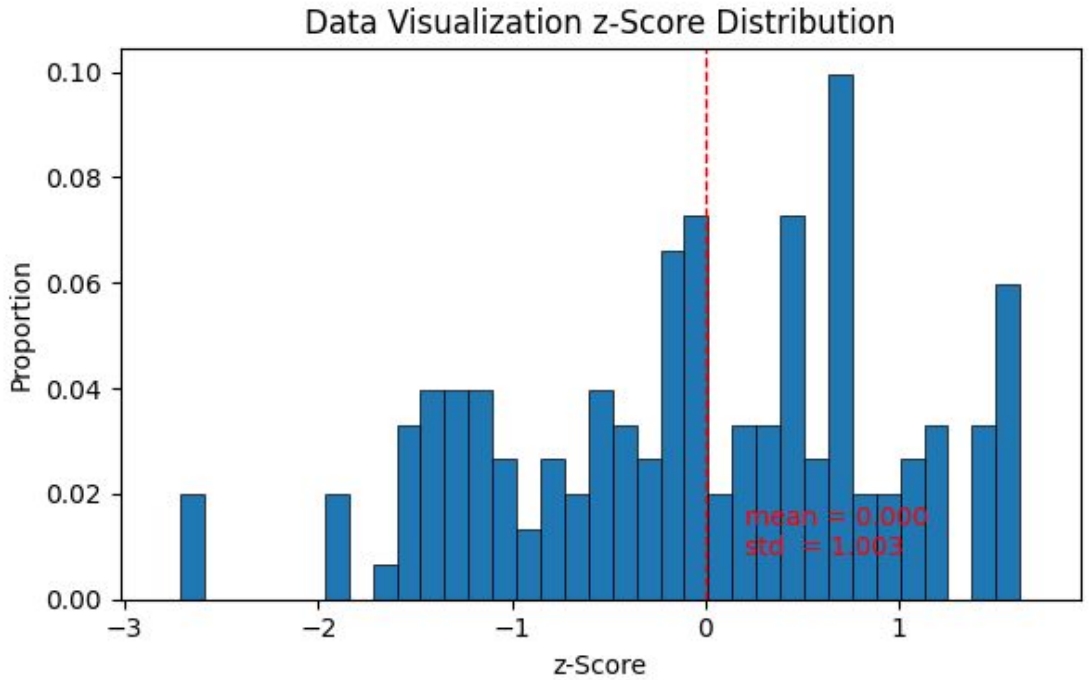




Data Set Description

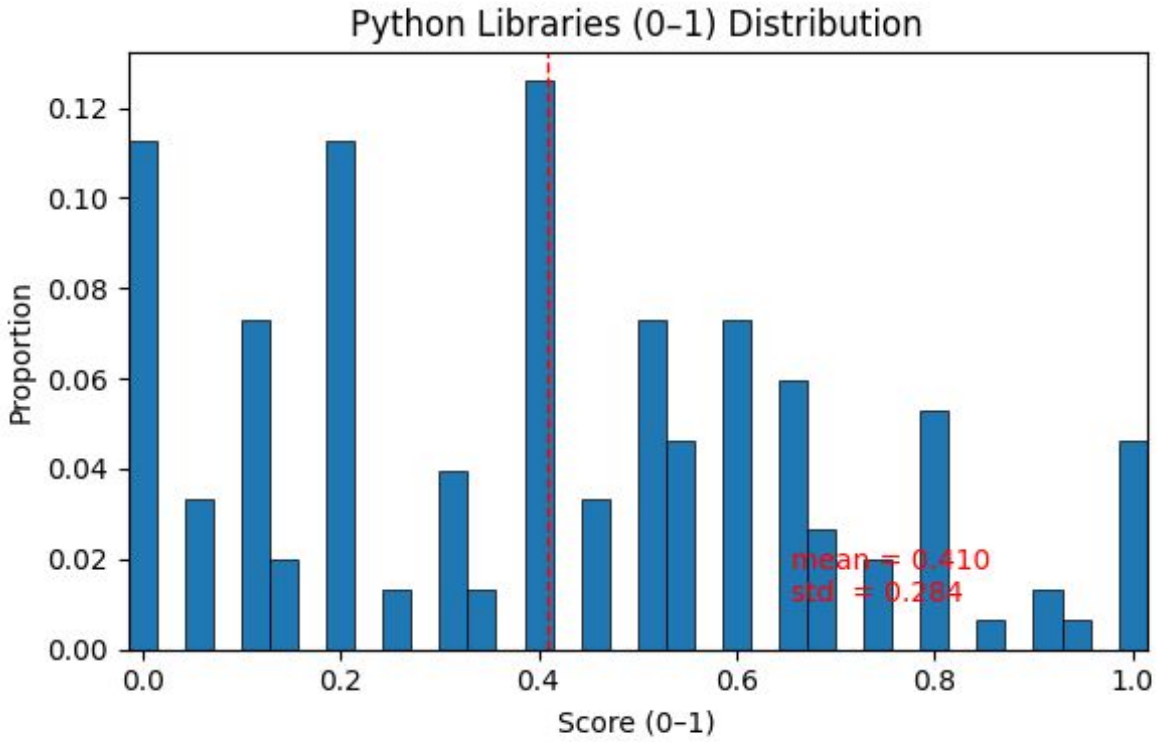
Column: Data Visualization

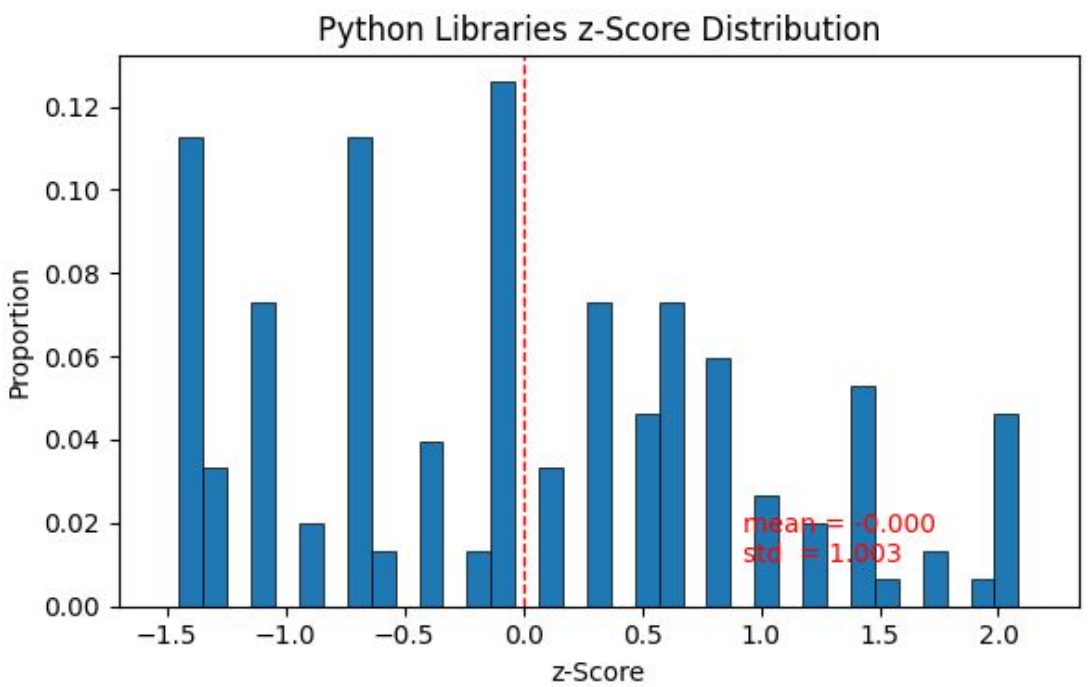




Data Set Description

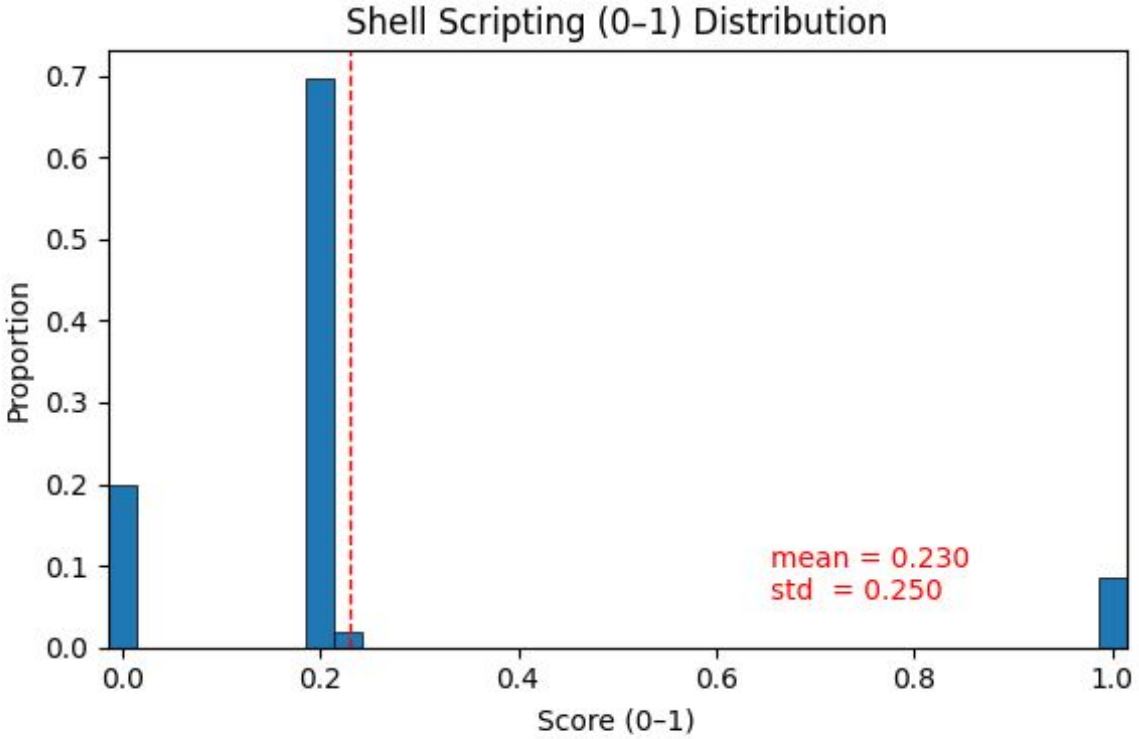
Column: Python Libraries

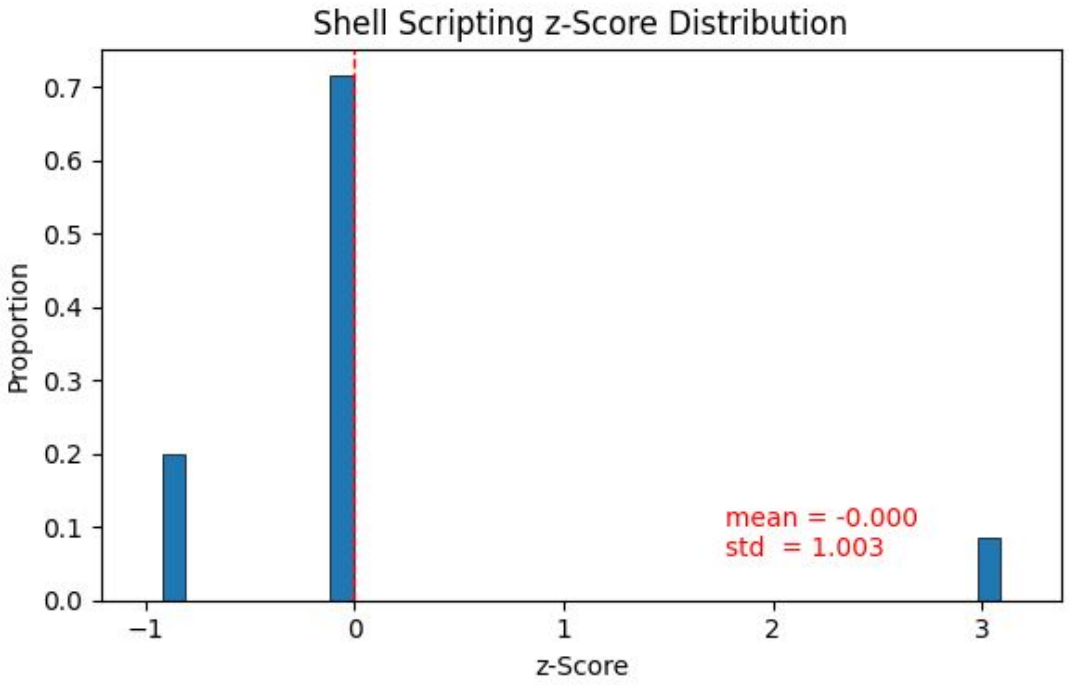




Data Set Description

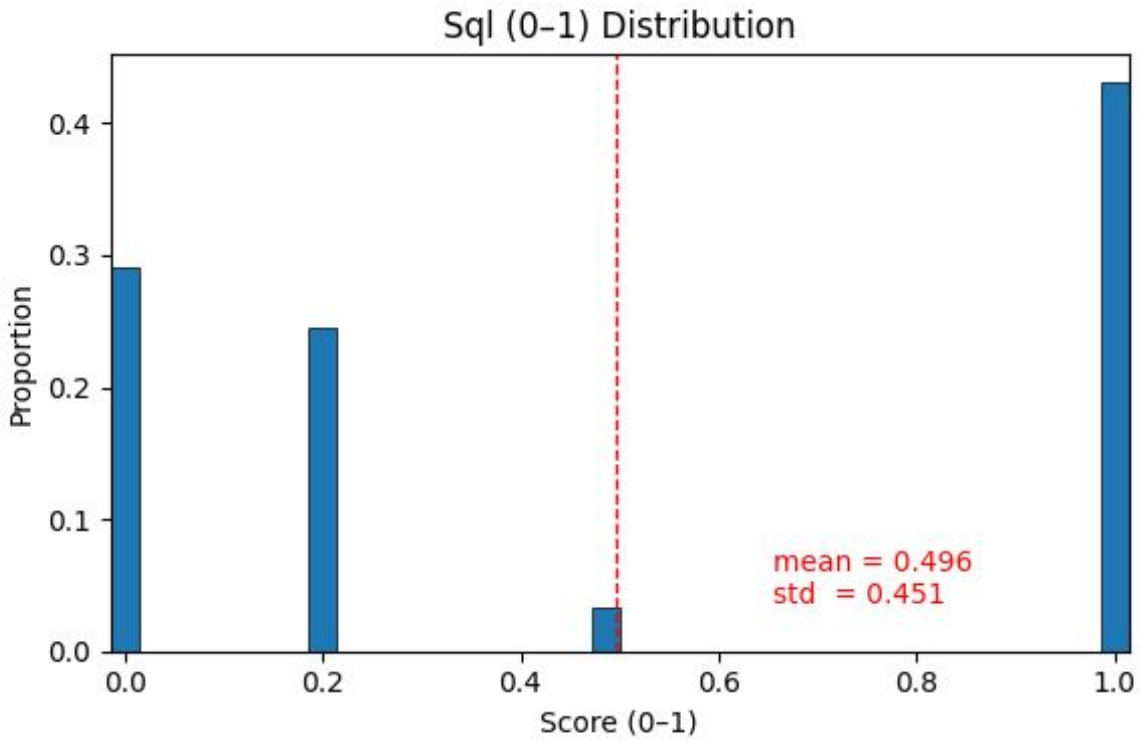
Column: Shell Scripting

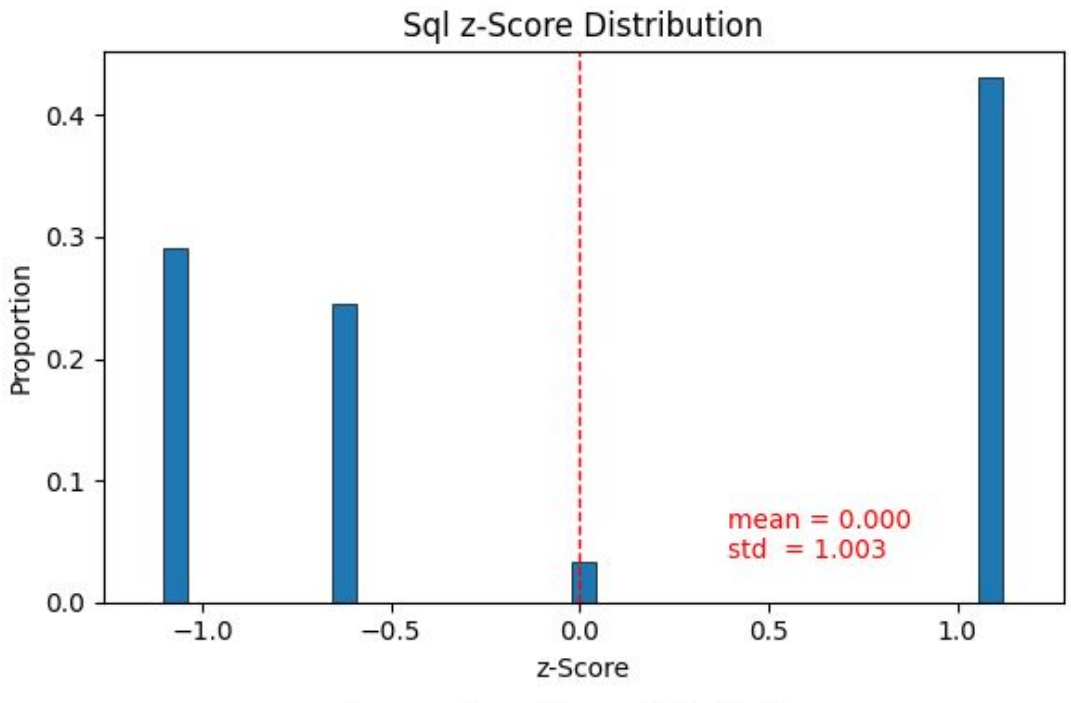




Data Set Description

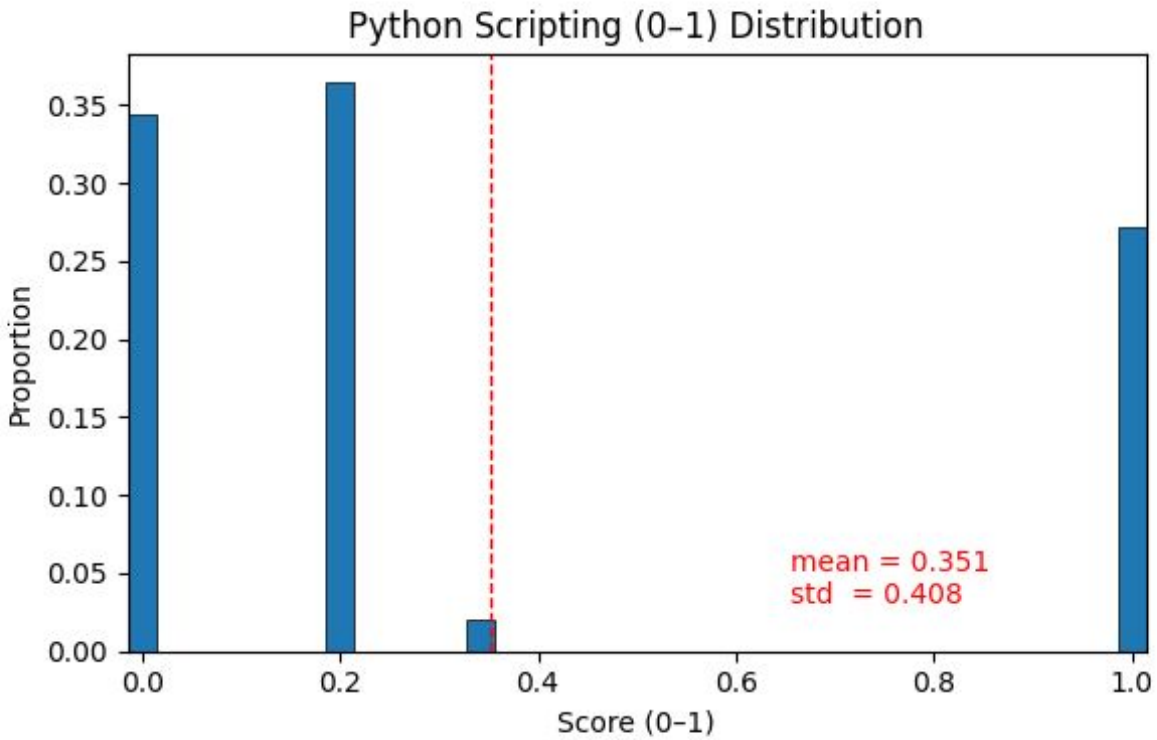
Column: Sql

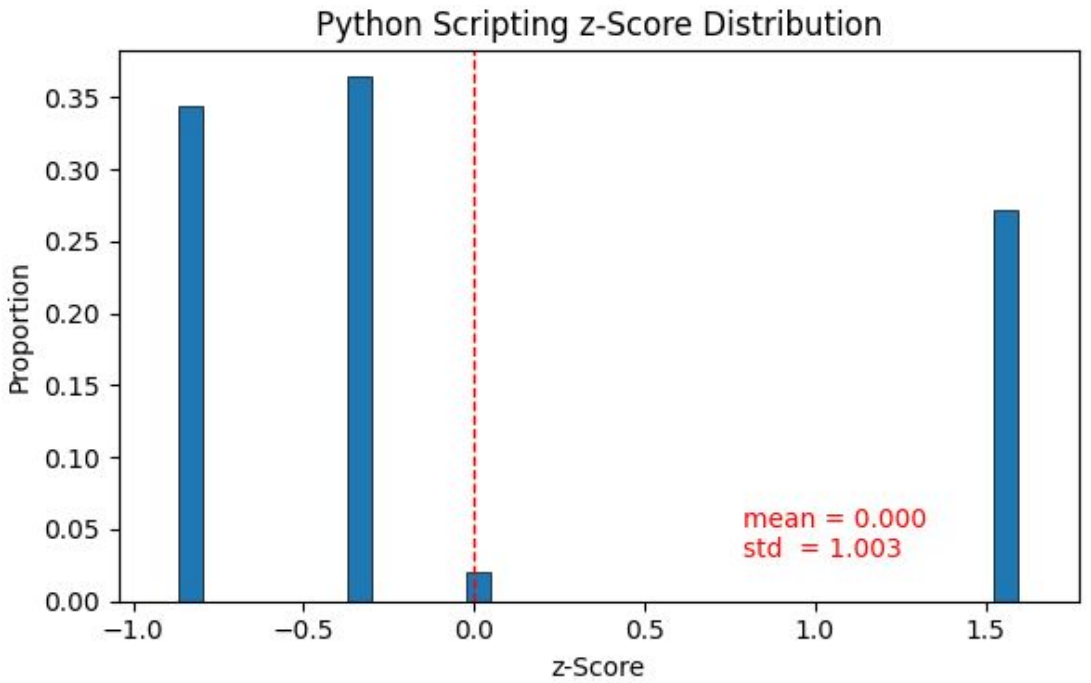




Data Set Description

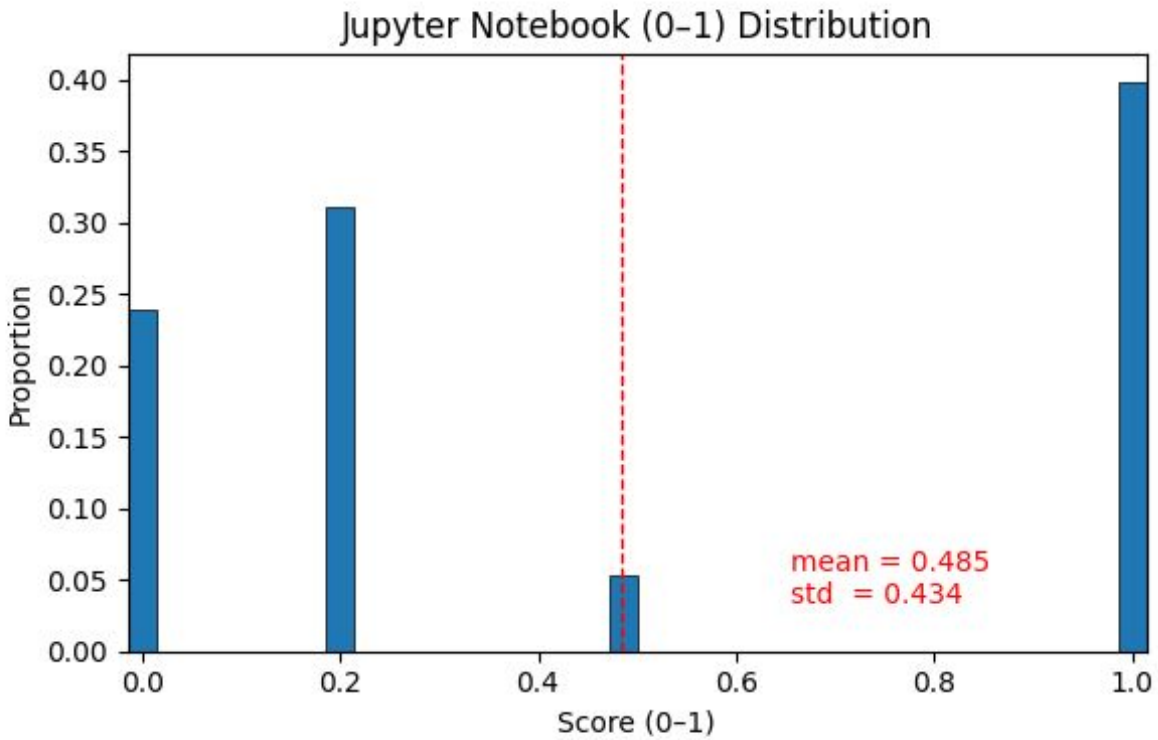
Column: Python Scripting

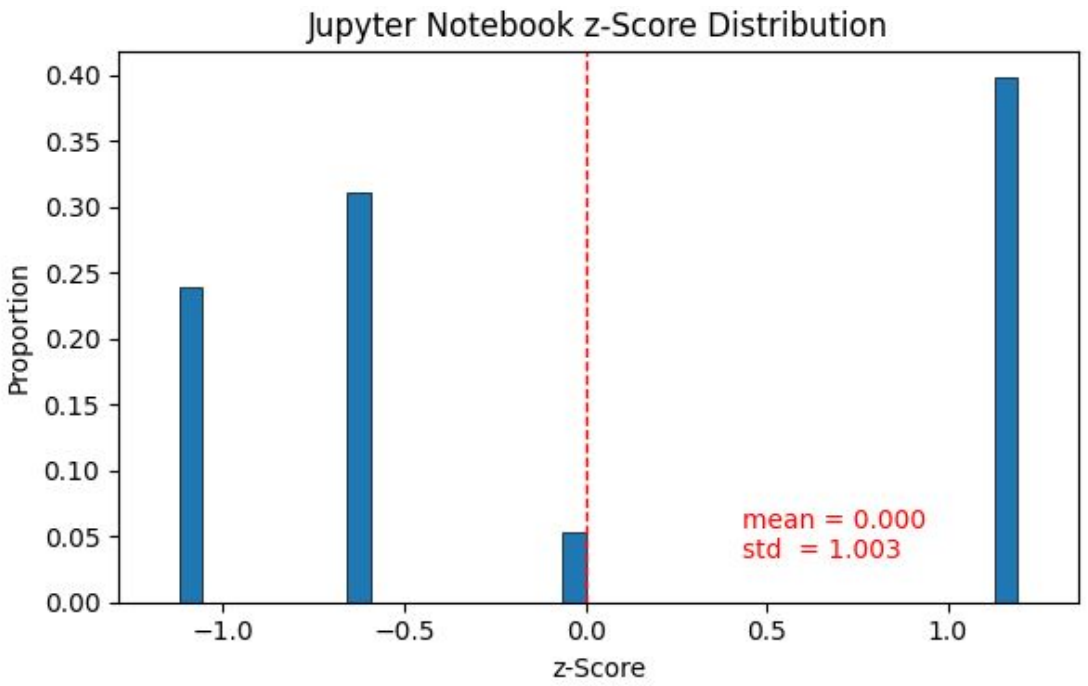




Data Set Description

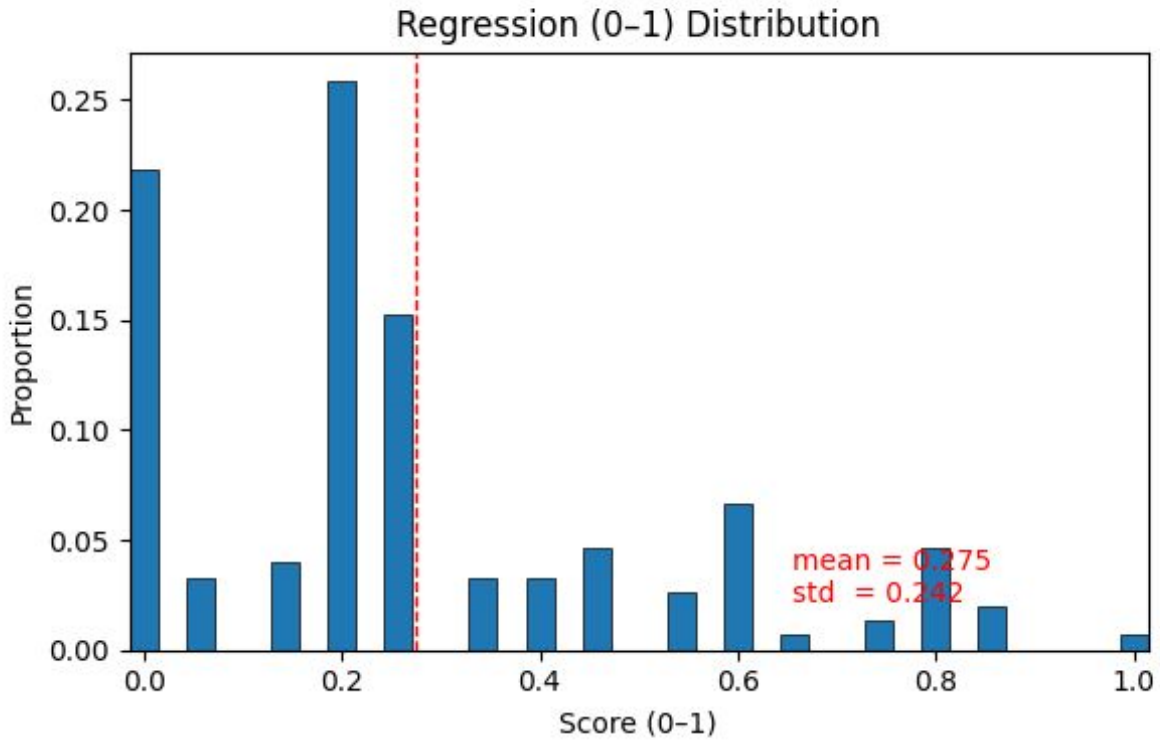
Column: Jupyter Notebook

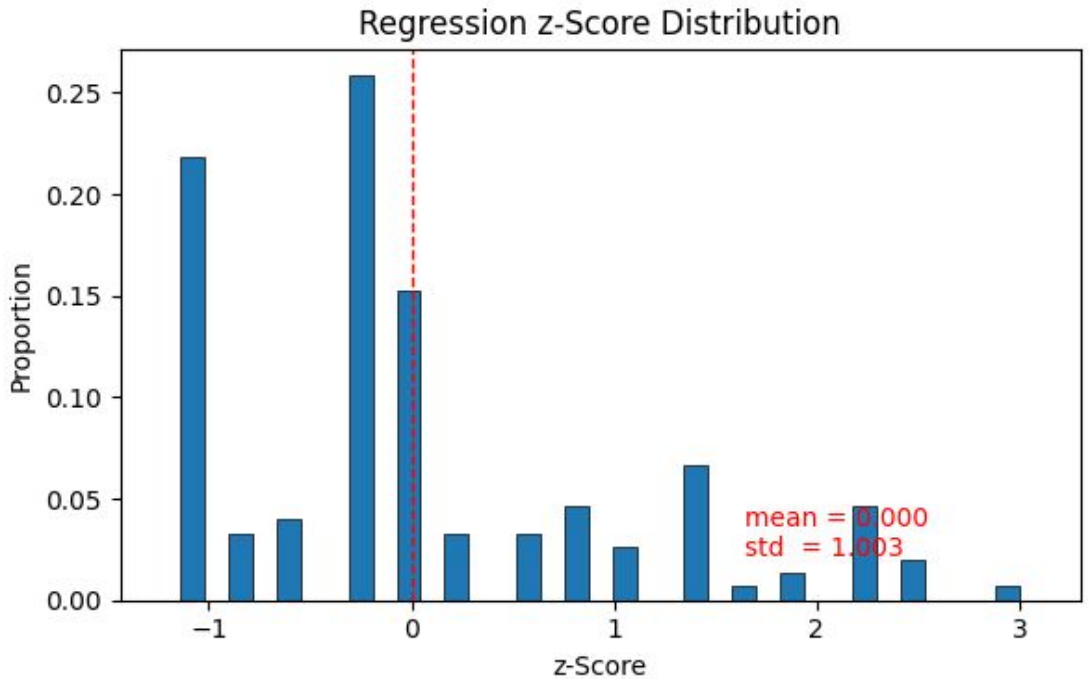




Data Set Description

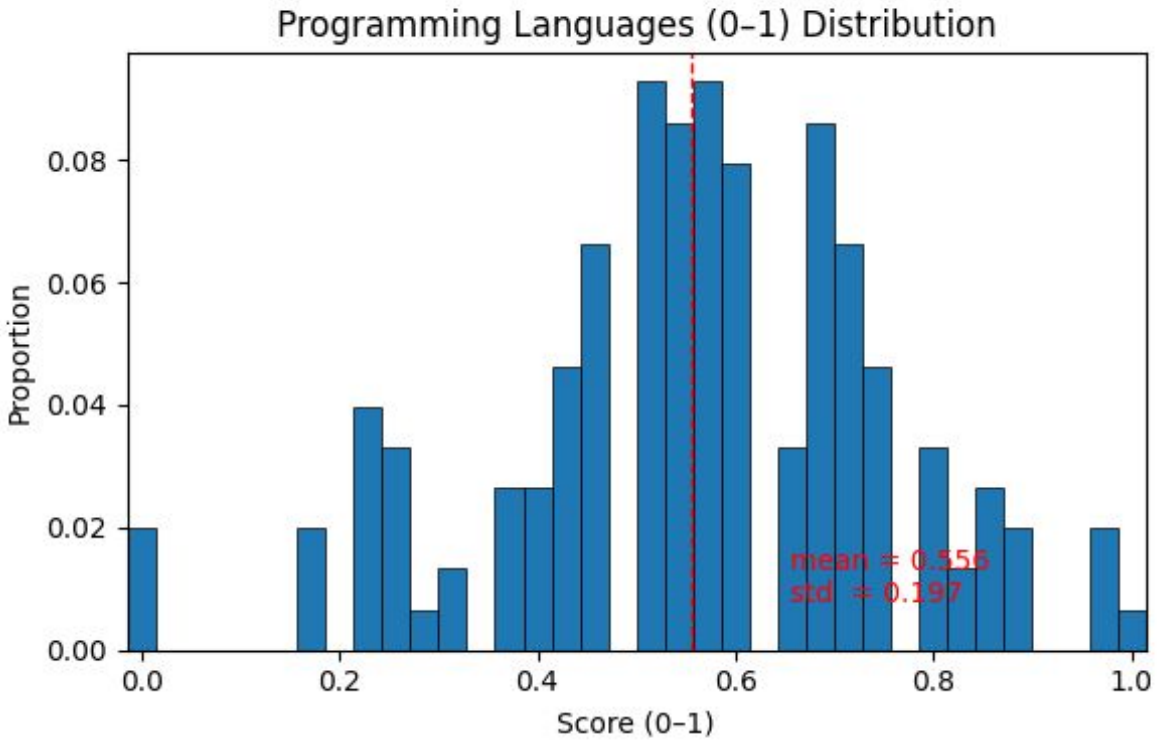
Column: Regression

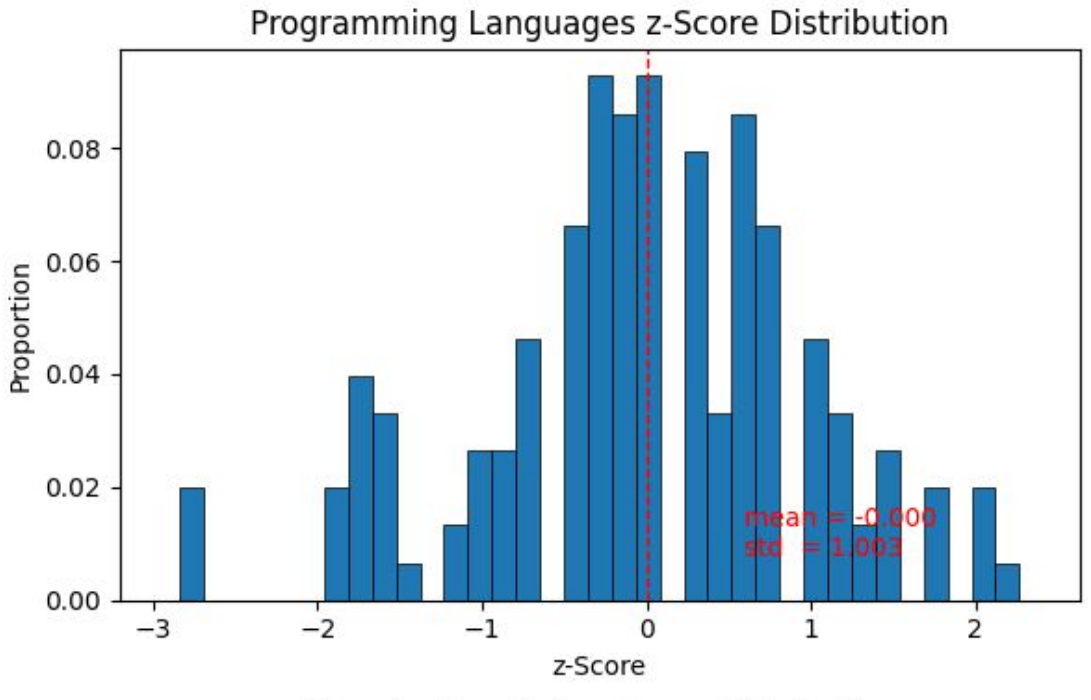




Data Set Description

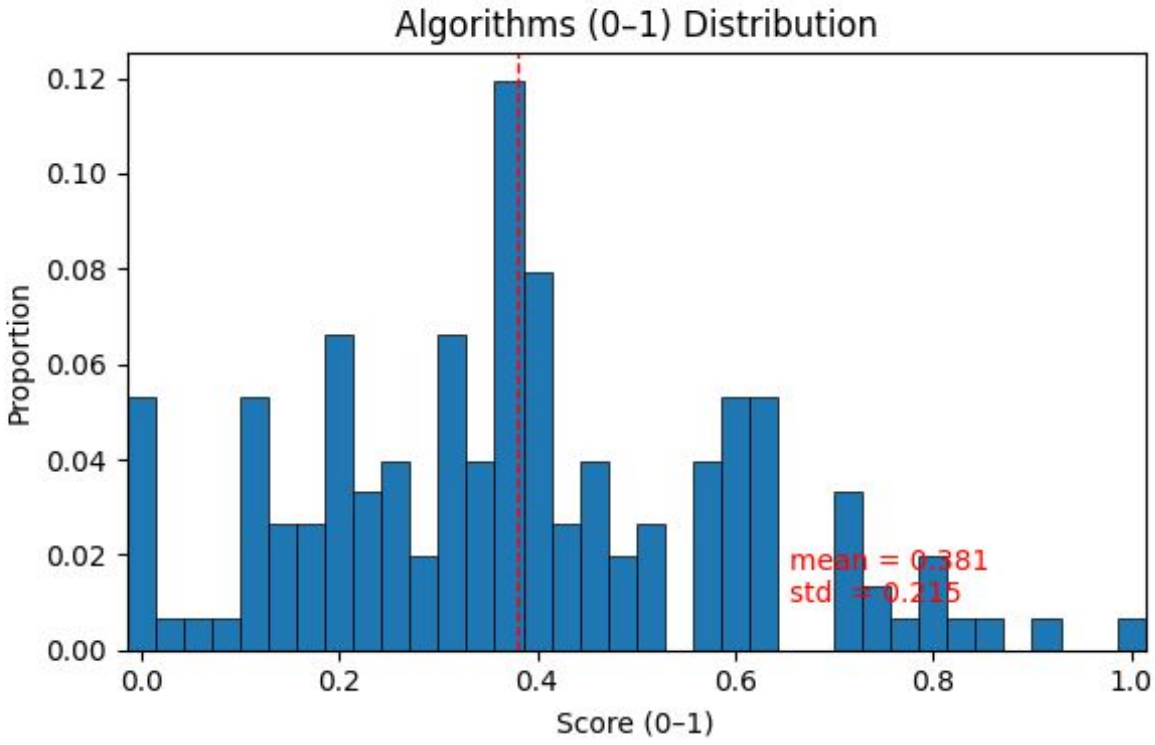
Column: Programming Languages

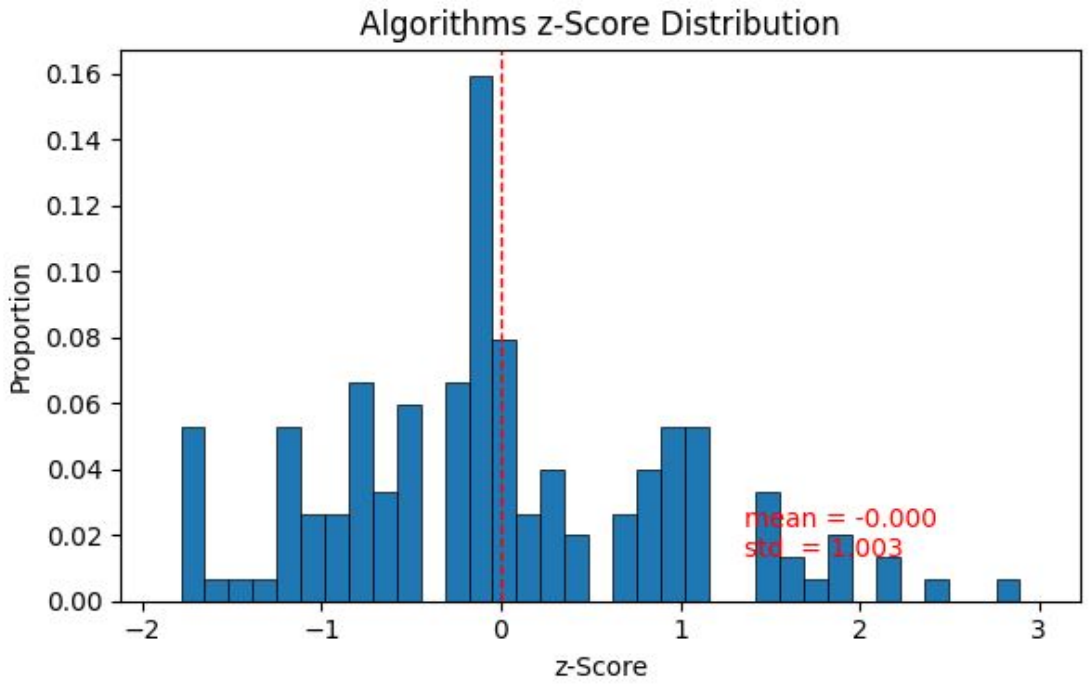




Data Set Description

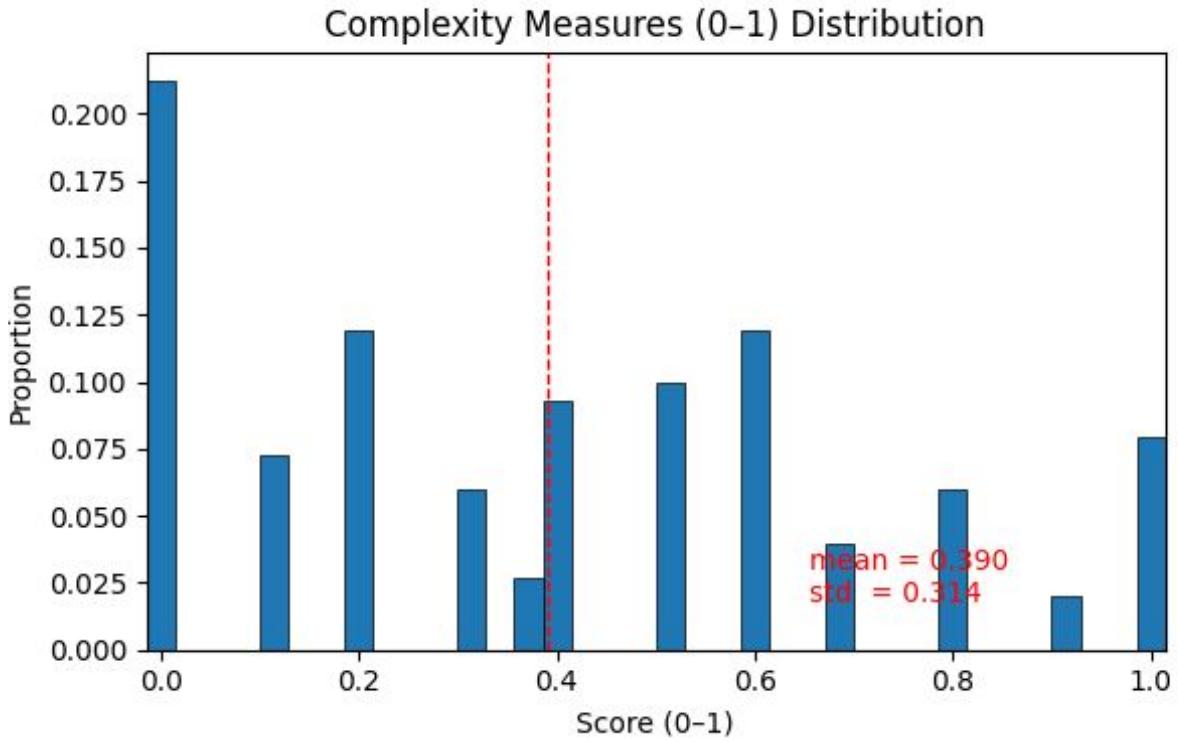
Column: Algorithms

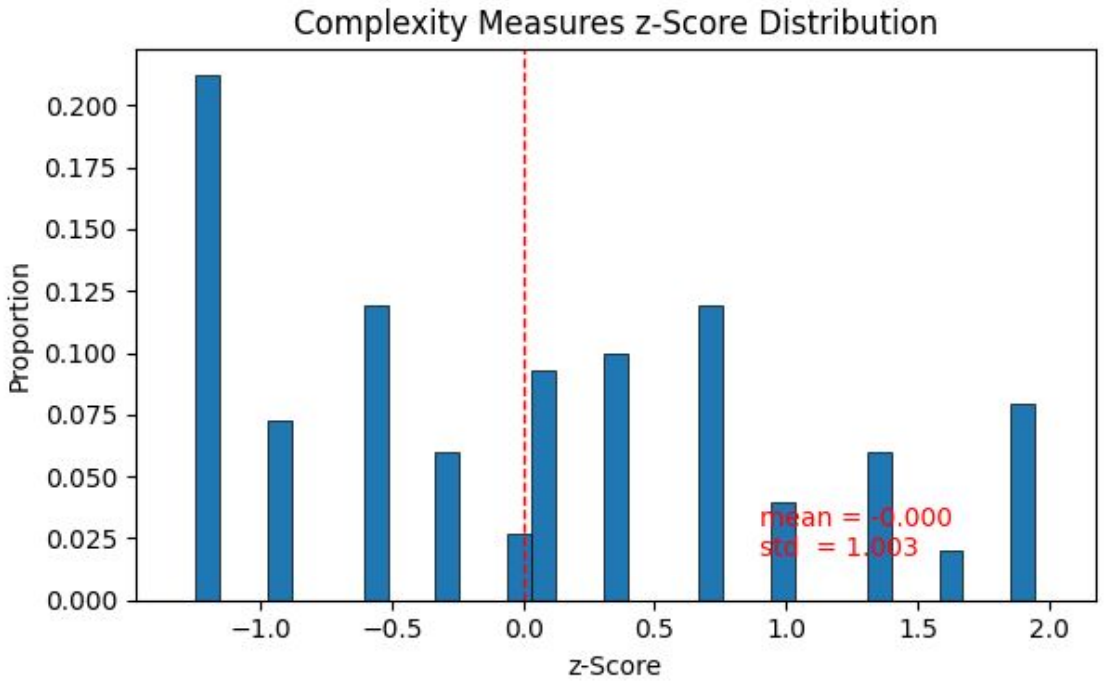




Data Set Description

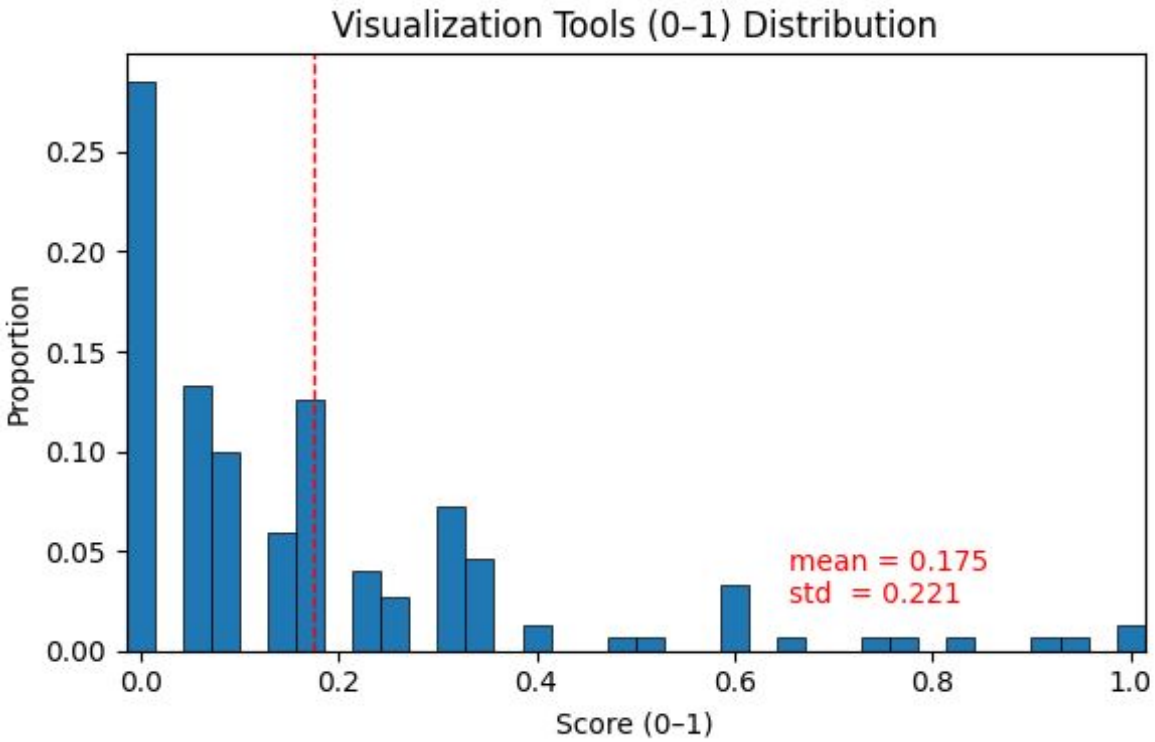
Column: Complexity Measures

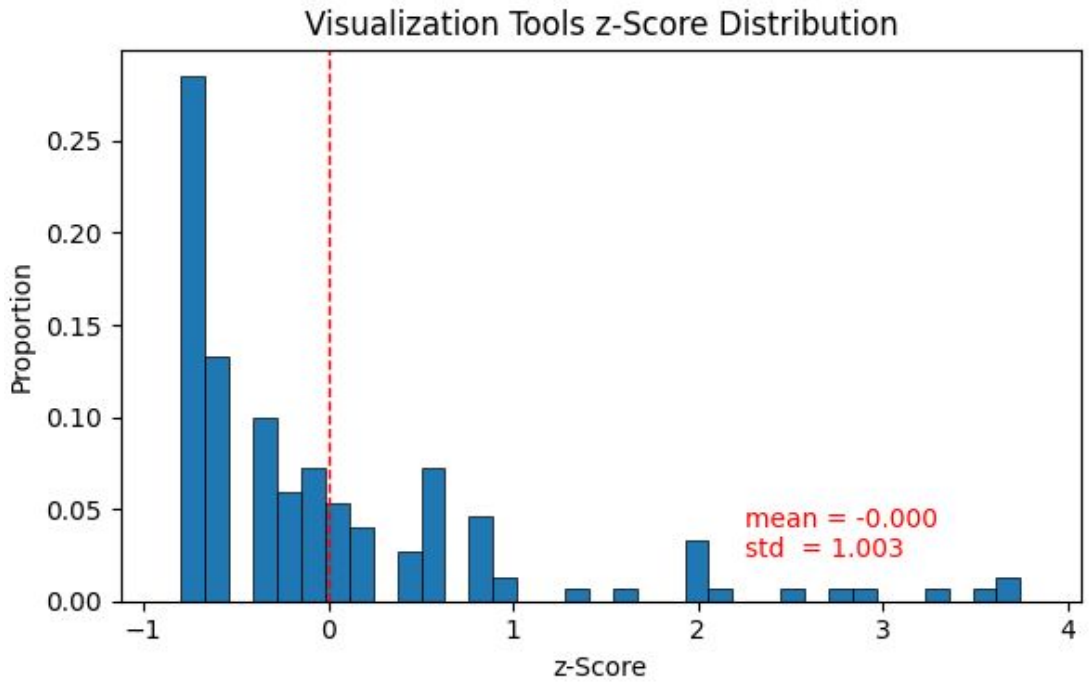




Data Set Description

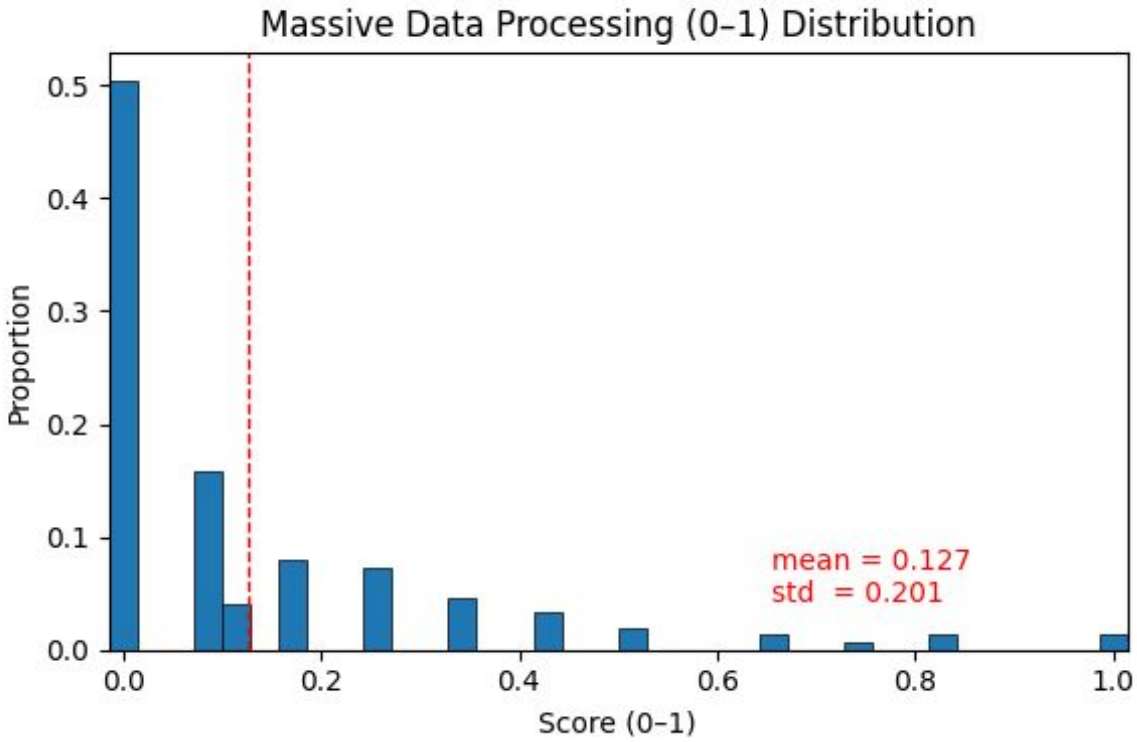
Column: Visualization Tools

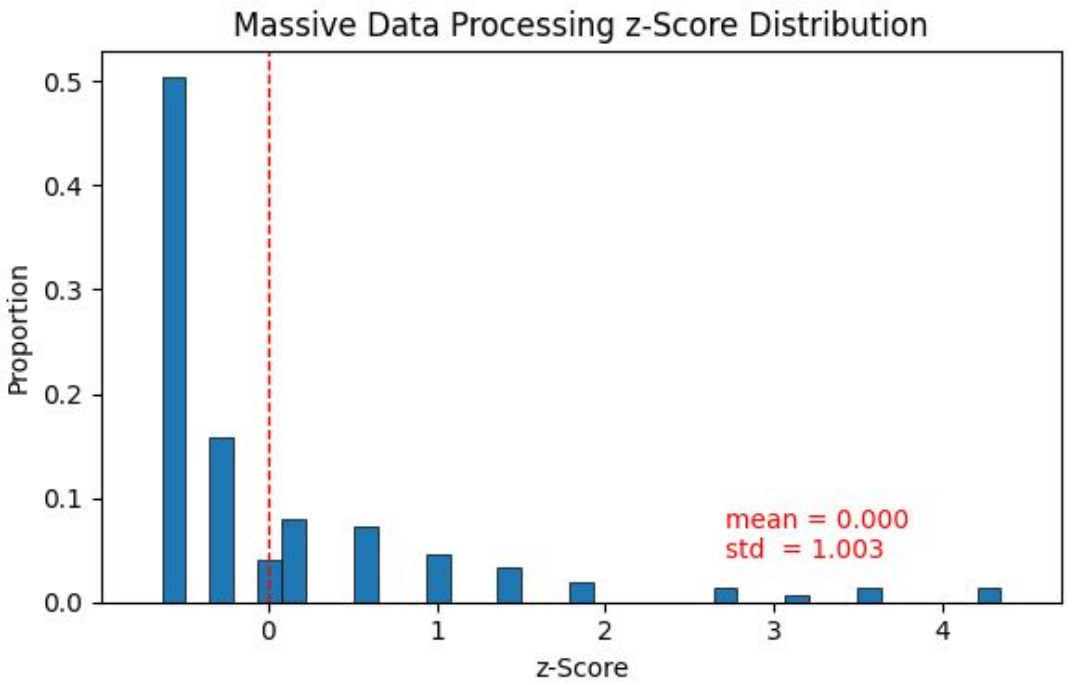




Data Set Description

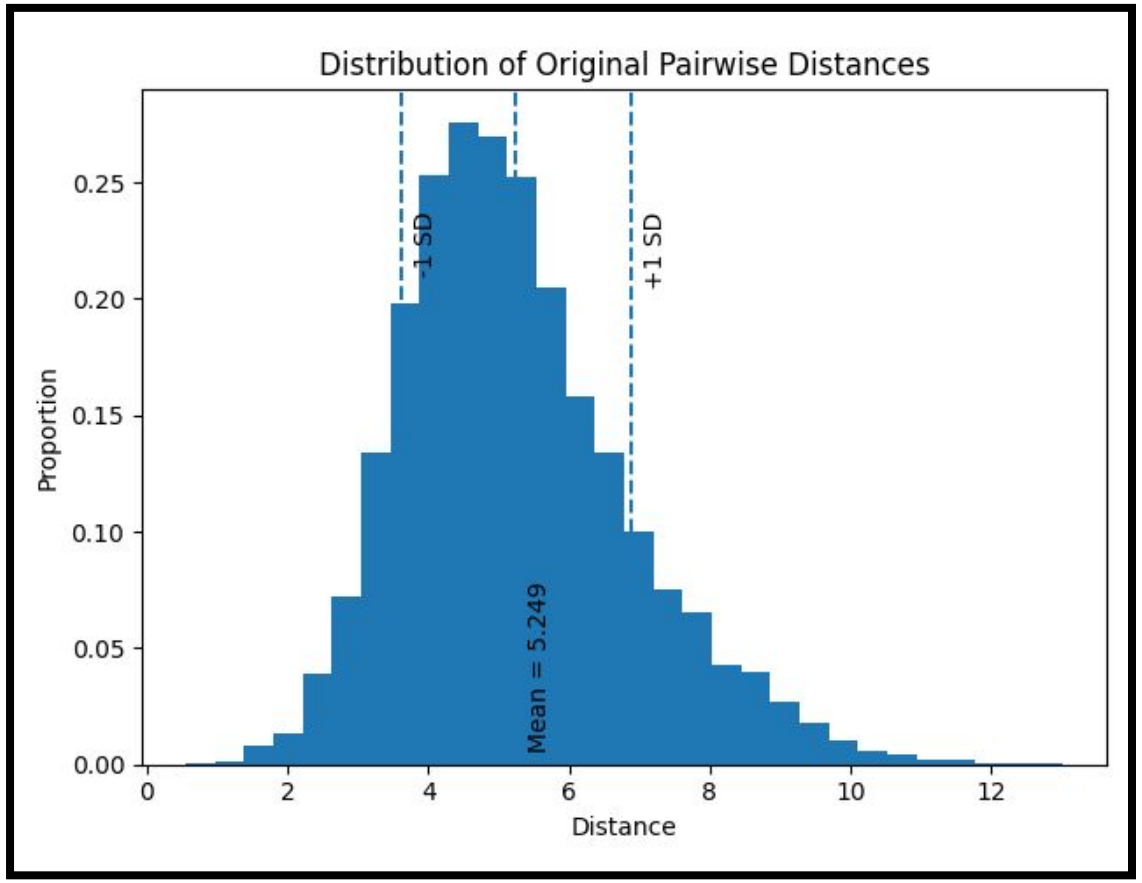
Column: Massive Data Processing

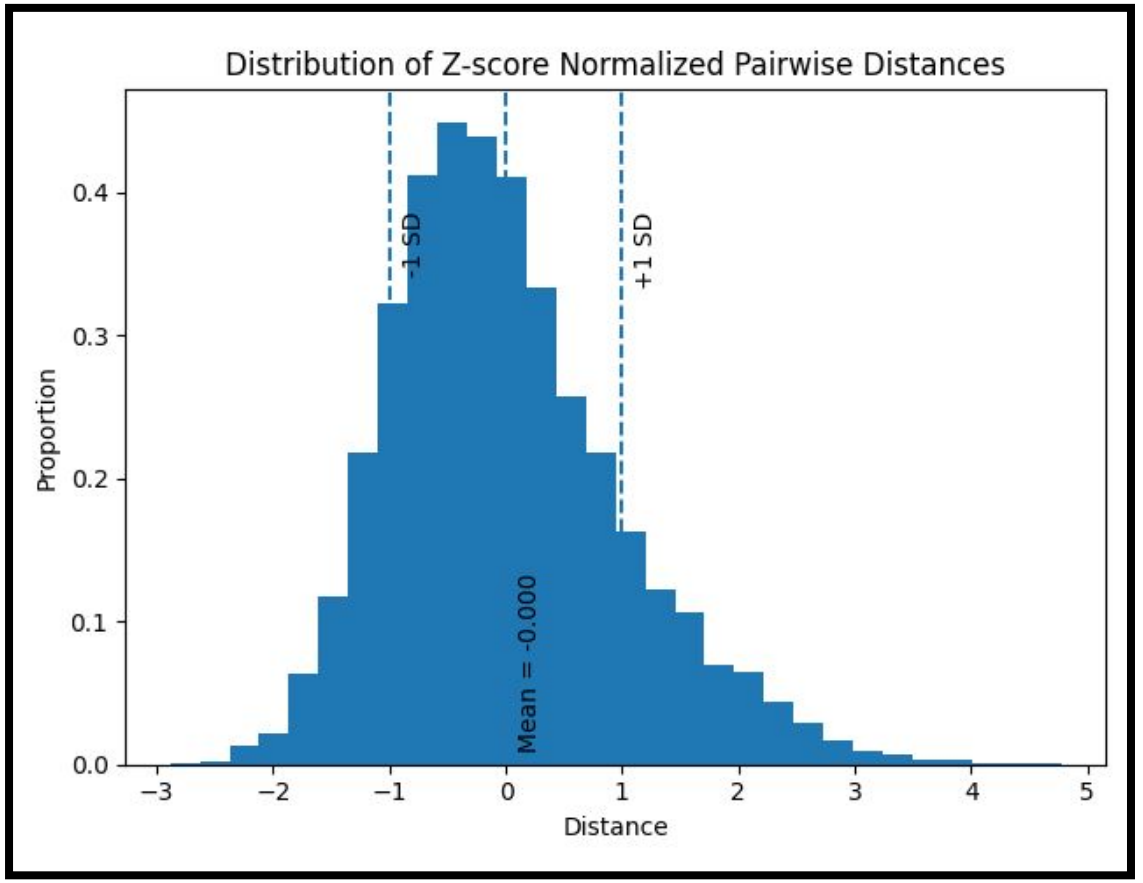




Pairwise Euclidean Students Distances

- Global histogram of pairwise Euclidean Students distances
- Global histogram of pairwise Euclidean Students distances, z-Score normalized

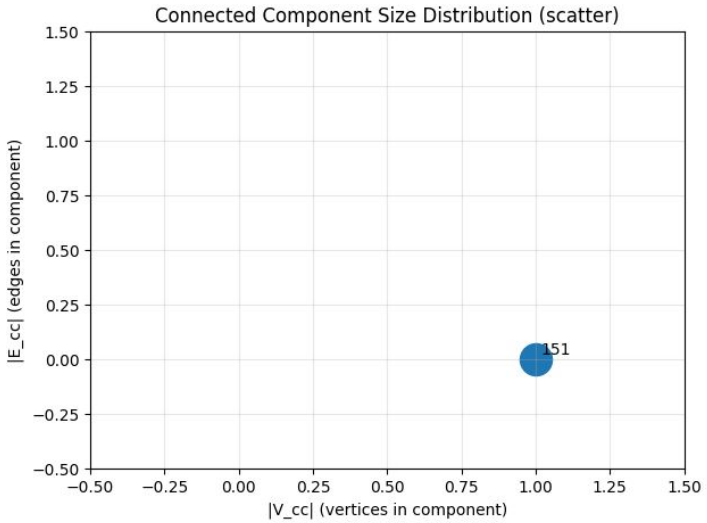




Student_Assessment_Graph_-3SD

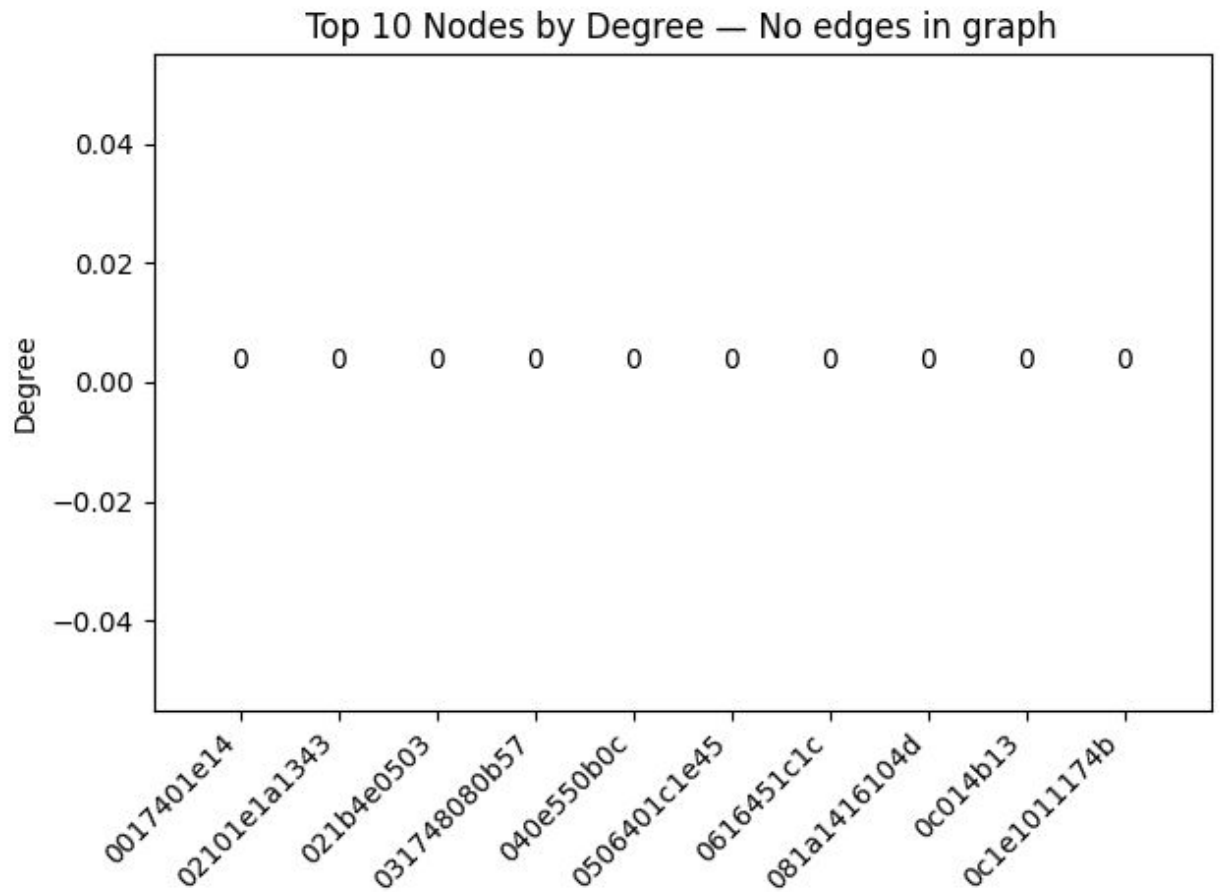
|V| = **151**
|E| = **0**
Does this graph degree distribution look like a Power Law Graph? **Cannot Plot**

|cc|= **151**
Create a scatter plot, where each dot represents the number of cc with a fixed vertex size and edge size.



Student_Assessment_Graph_-3SD

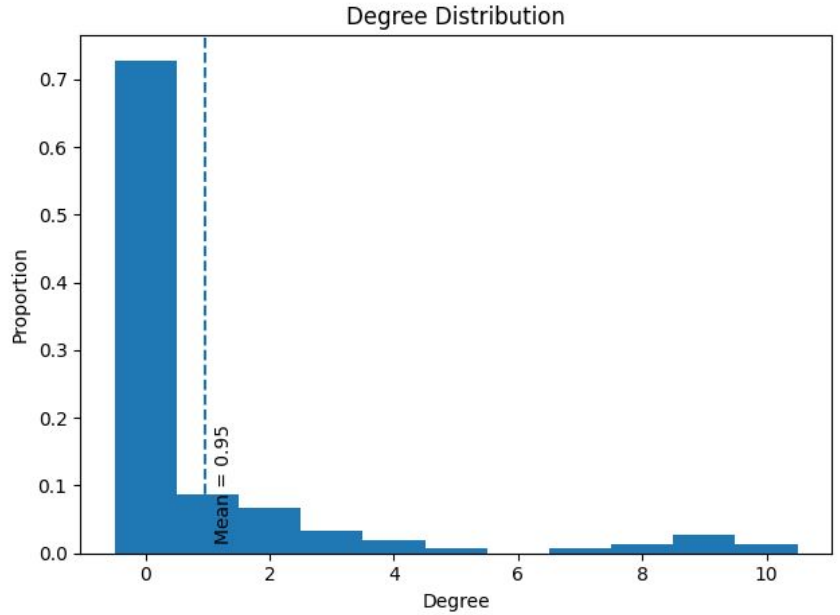
Top 10% students’ NetIDs and Total_Score

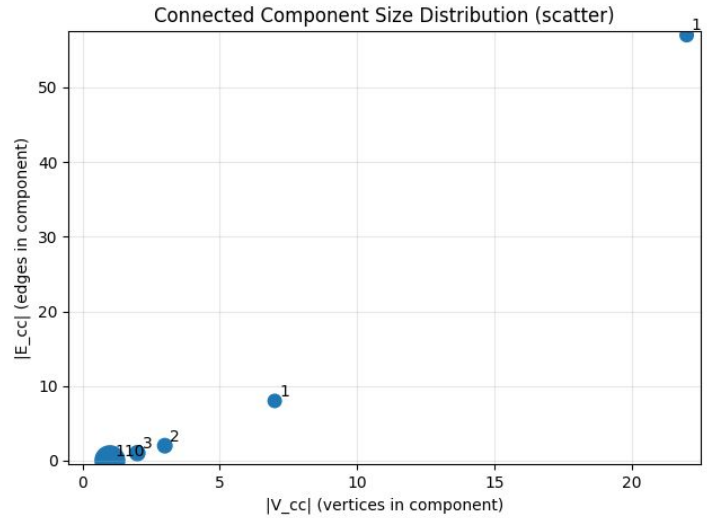


Student_Assessment_Graph_-2SD

|V| = **151**
|E| = **72**
Does this graph degree distribution look like a Power Law Graph? **Yes**

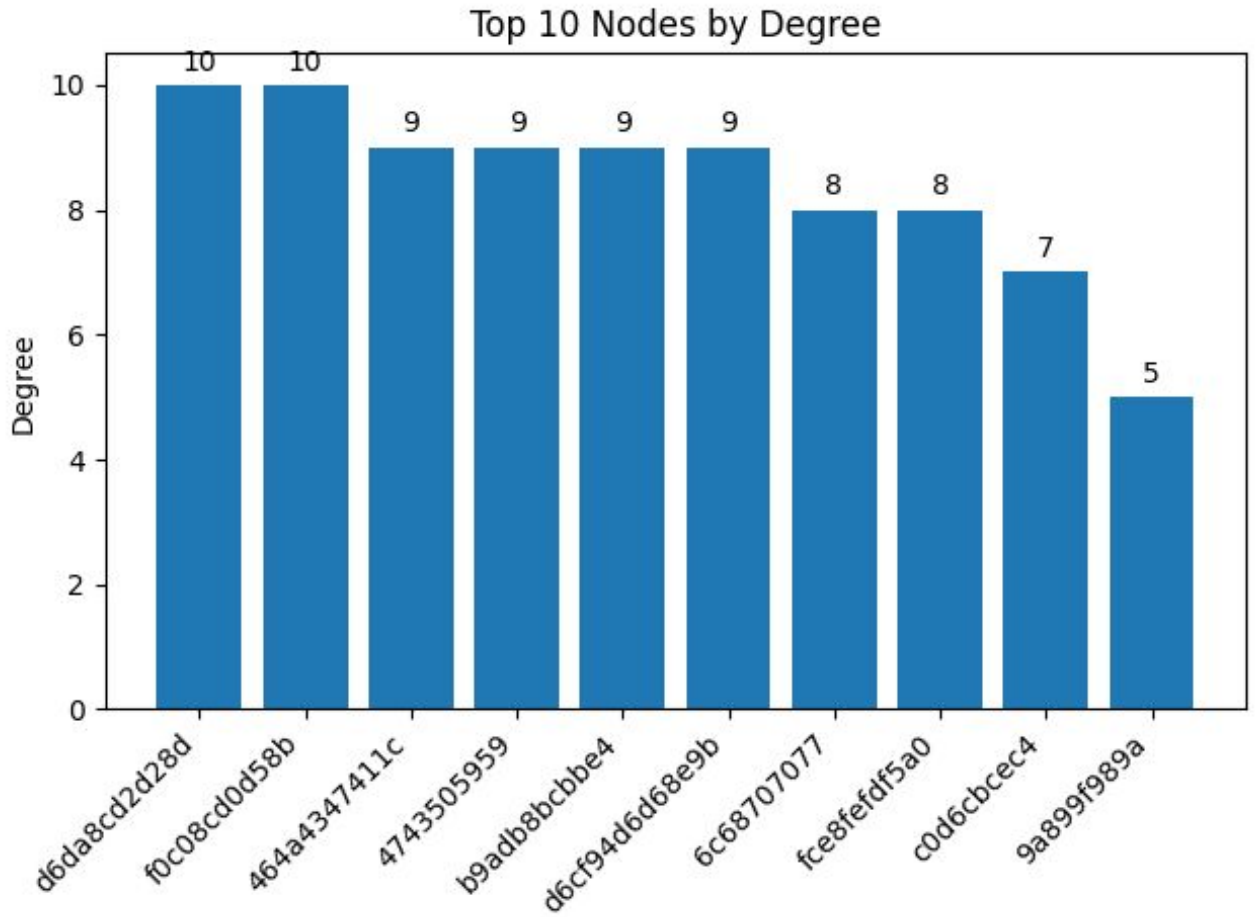
|cc|= **117**
Create a scatter plot, where each dot represents the number of cc with a fixed vertex size and edge size.





Student_Assessment_Graph_-2SD

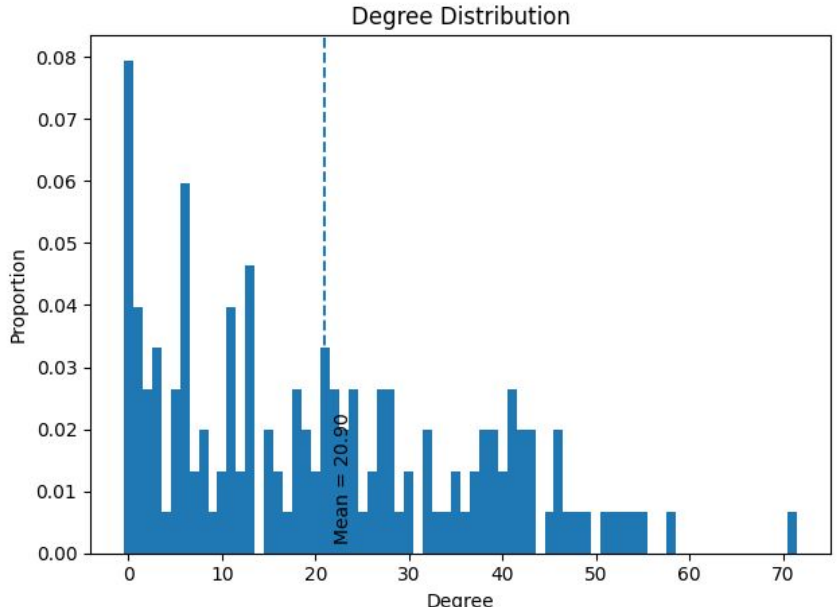
Top 10% students’ NetIDs and Total_Score

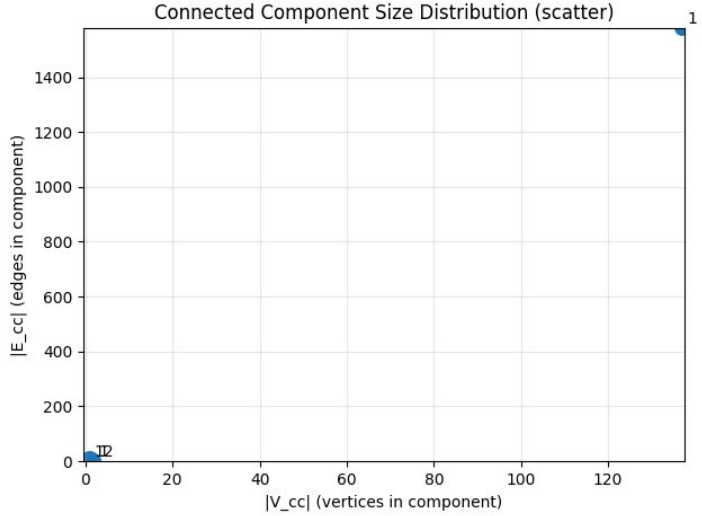


Student_Assessment_Graph_-1SD

|V| = **151**
|E| = **1578**
Does this graph degree distribution look like a Power Law Graph? **Yes**

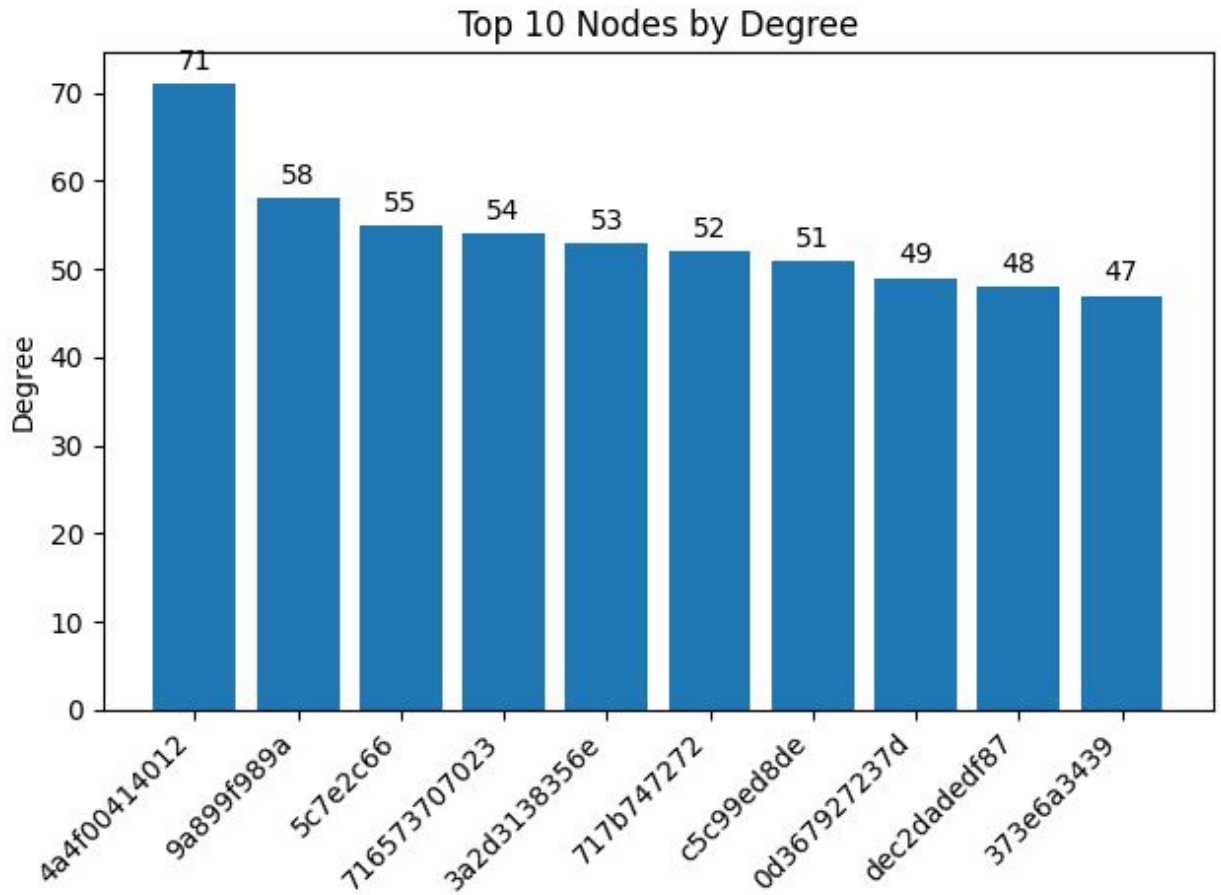
|cc|= **14**
Create a scatter plot, where each dot represents the number of cc with a fixed vertex size and edge size.





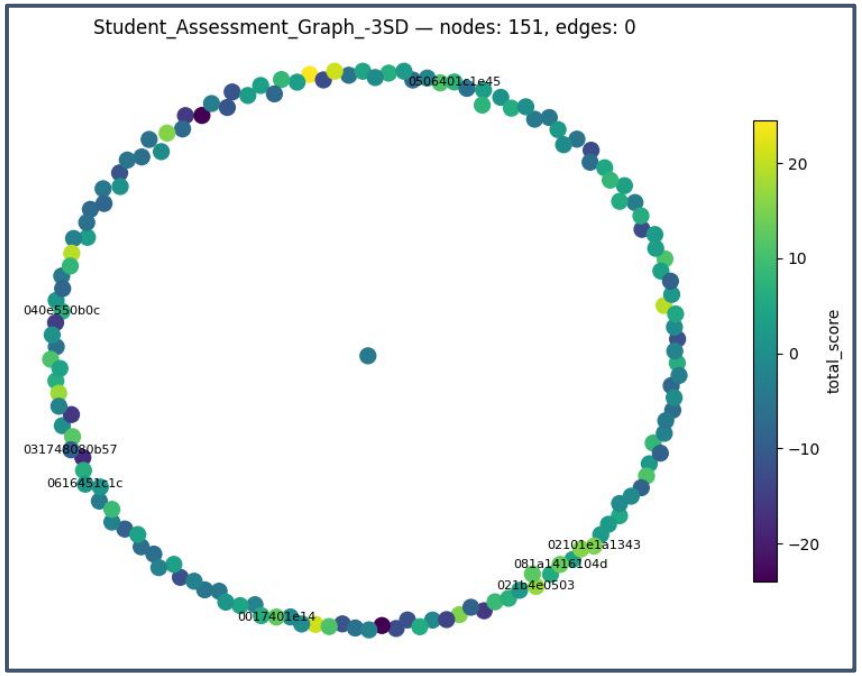
Student_Assessment_Graph_-1SD

Top 10% students’ NetIDs and Total_Score

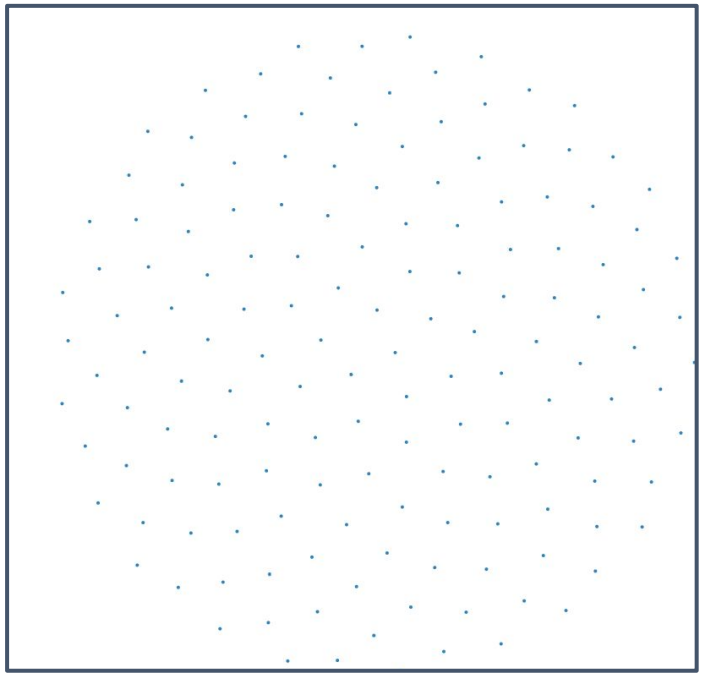


Layouts of Student_Assessment_Graph_-3SD

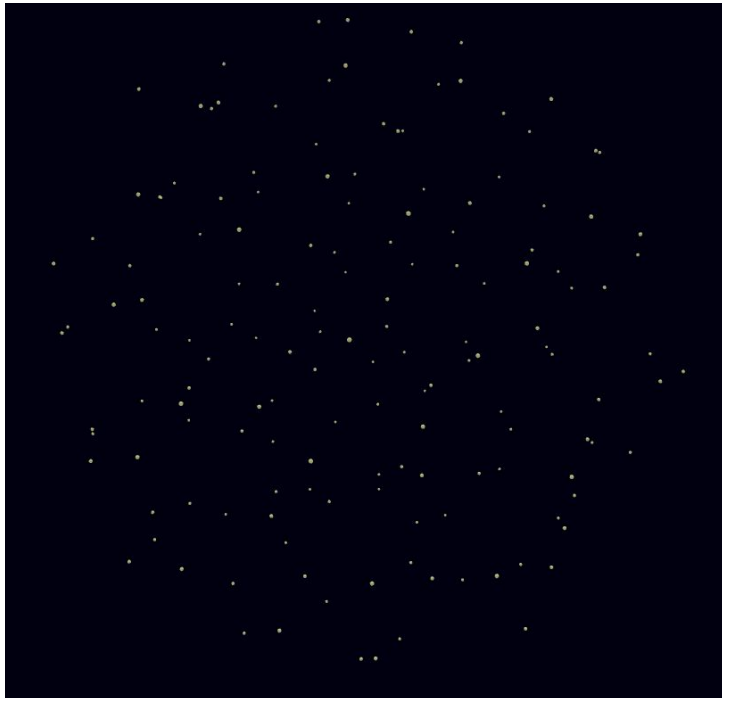
NetworkX.Spring_Layout()



JavaScript 2D Force Graph

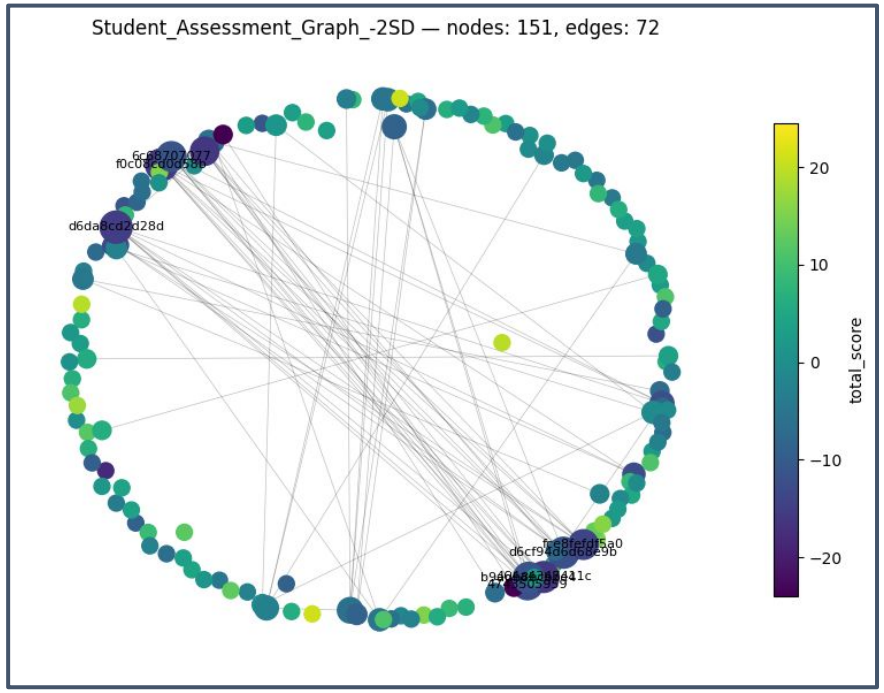


JavaScript 3D Force Graph

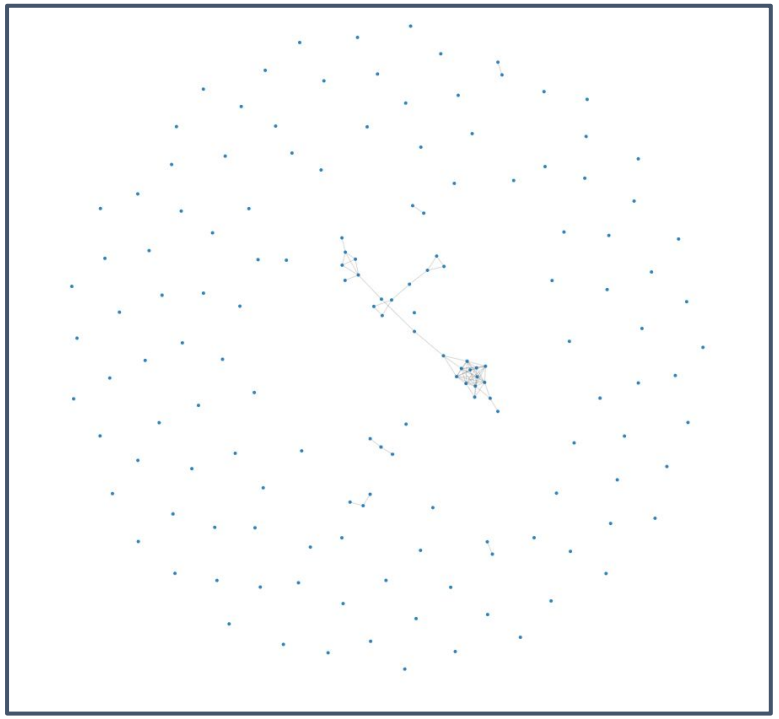


Layouts of Student_Assessment_Graph_-2SD

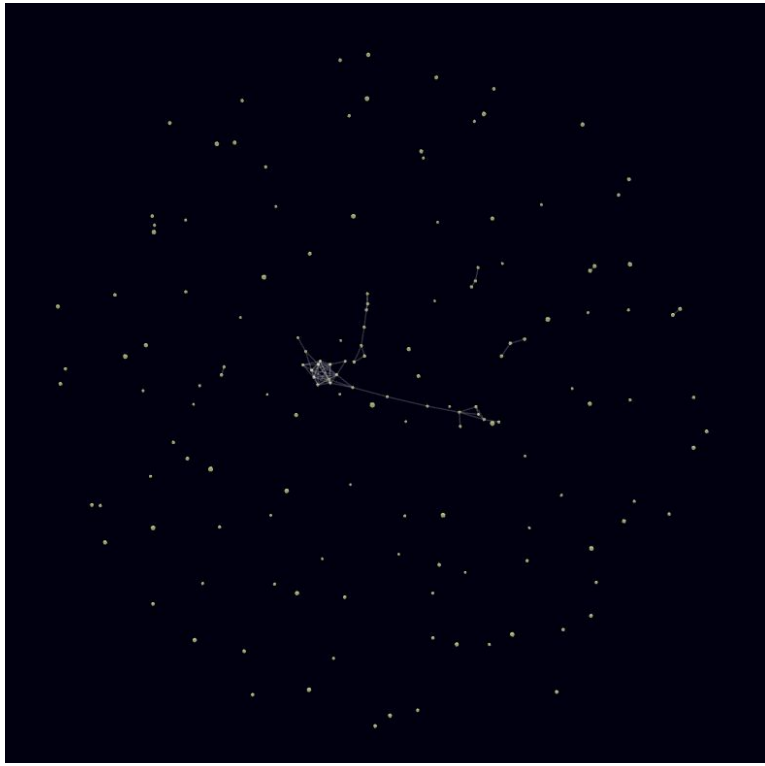
NetworkX.Spring_Layout()



JavaScript 2D Force Graph

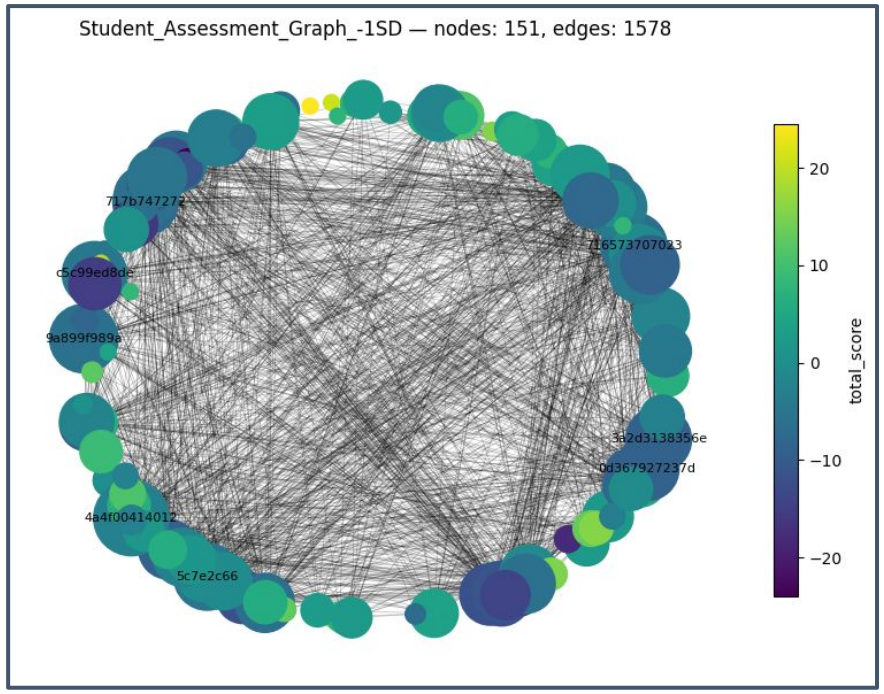


JavaScript 3D Force Graph

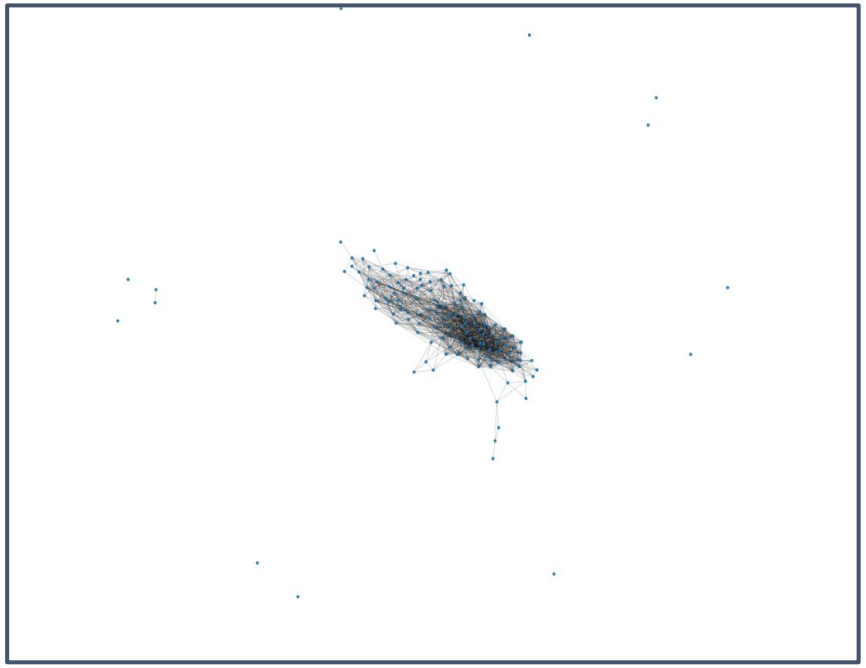


Layouts of Student_Assessment_Graph_-1SD

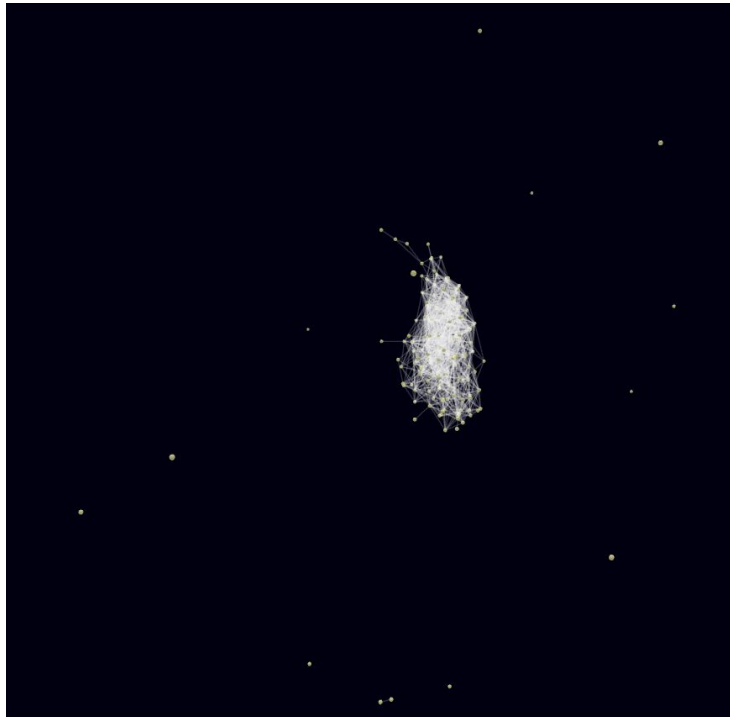
NetworkX.Spring_Layout()



JavaScript 2D Force Graph



JavaScript 3D Force Graph



Observations

Clusters in 1SD

	NetID 1	NetID 2	NetID 3	NetID 4	NetID 5	NetID 6	NetID 7	NetID 8	NetID 9
Cluster 1 (n=137)	0017401e14	02101e1a1343	031748080b57	040e550b0c	0506401c1e45	0616451c1c	081a1416104d	0c014b13	0c1e1011174b
Cluster 2 (n=2)	53420a	70686a6263							

Clusters in 2SD

	NetID 1	NetID 2	NetID 3	NetID 4	NetID 5	NetID 6	NetID 7	NetID 8	NetID 9
Cluster 1 (n=22)	0c014b13	0d367927237d	1d38202e27	3a2d3138356e	464a4347411c	4743505959	47545e565704	4a4f00414012	6c68707077
Cluster 2 (n=7)	383e22212178	5c7e2c66	5d504543461b	717b747272	d6dcd0d0d3	dec3ddd2d3	fee3bdfffbac		
Cluster 3 (n=3)	6e7832636338	c5c99ed8de	cadb82d9d5						
Cluster 4 (n=3)	0c1e1011174b	a5aae0bcbbef	dec2dadedf87						
Cluster 5 (n=2)	6a676c6d6c	87							
Cluster 6 (n=2)	322d323a346a	ebfbbde0e8b8							
Cluster 7 (n=2)	8193df8b88	ebf5e2e0e7be							

Findings

Does 1SD have “similar” high scores in certain subsets of skill proficiencies?

- *Yes*, but the threshold is loose: there’s one giant component (n=137) and one small cluster (n=2) with clear common strengths.
- This suggests −1SD links many students together; use −2SD to see finer groups.

Cluster A (n=137): 0017401e14, 02101e1a1343, …, fee3f0f0f0ab

Signature skills: no single subset stands out; very broad mix since it is a giant cluster which likely is not meaningful for profiling

Cluster B (n=2): 53420a, 70686a6263

Signature skills: algorithms (mean z≈1.96, d≈2.02), visualization_tools (≈2.26, 2.36), massive_data_processing (≈3.93, 4.45), plus python_libraries, regression, jupyter_notebook, probability/stats, complexity (all ≥0.8 mean z and ≥0.8 d)

Why meaningful: both students are jointly high on the same skills of algorithms, MDP, viz, Python, with large effect sizes vs. the rest.

Findings

Does 2SD have “similar” high scores in certain subsets of skill proficiencies?

- *Yes*. We see multiple mid-size clusters plus a few pairs/triads. Several clusters share distinctive skill “signatures.”

Cluster C (n=22): 0c014b13, 0d367927237d, …, fce8fefdf5a0
Signature skills: no single subset clears the mean-z≥0.7 & d≥0.5 bar; weaker evidence

Cluster D (n=7): 383e22212178, 5c7e2c66, 5d504543461b, 717b747272, d6dcd0d0d3, dec3ddd2d3, fee3bdfffbac
Signature skills: SQL (mean z≈1.12, d≈1.20)
Why meaningful: coherent shared strength in SQL with clear separation from others.

Cluster 3 (n=3): 6e7832636338, c5c99ed8de, cadb82d9d5
Signature skills: jupyter_notebook_zscore (mean z≈1.19, d≈1.23)
Why Meaningful: shared notebook/tooling strength above class.

Cluster 4 (n=3): 0c1e1011174b, a5aae0bcbbef, dec2dadedf87
Signature skills: no signature skills

Cluster 5 (n=2): 6a676c6d6c, 87
Signature skills: no signature skills

Cluster E (n=2): 322d323a346a, ebfbbde0e8b8
Signature skills: jupyter_notebook (≈1.19, 1.21), python_scripting (≈1.60, 1.64), SQL (≈1.12, 1.14)
Why meaningful: both students share a tooling-heavy profile on Python, Jupyter, and SQL at clearly above-average levels.

Cluster F (n=2): 8193df8b88, ebf5e2e0e7be
Signature skills: algorithms (≈1.22, 1.24), calculus/linear algebra (≈1.09, 1.10), complexity (≈0.99, 1.00), jupyter_notebook (≈1.19, 1.21), probability/stats (≈1.14, 1.16), regression (≈1.35, 1.37)
Why meaningful: math, algorithms, and stats pattern consistent across both, with moderate-to-large effect sizes.

Findings

Therefore, as the threshold changes from 3SD to 2SD to 1SD the number of similar skills amongst each student can occur.

For example, Threshold 3 does not have any similar skills amongst students. Threshold 2 increases the number of similar skills amongst students to 22+7+3+3+2+2+2=41 students. And threshold 1 increases the number of similar skills amongst students to 137+2=139 students.

In conclusion, **2SD** is the best fit. This is because 3SD is too strict (no edges and no clusters), while 1SD is too loose (a giant, heterogeneous component of 137 plus a pair). And At 2SD we get 7 interpretable clusters (41 students total) that show coherent “signature” skill subsets.

References

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, 51–56.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. 2020. Array programming with NumPy. *Nature* 585, 7825 (2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

John D. Hunter. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, 11–15.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, et al. 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90.