*Clustering*

1A

* Cleaned the dataset used for clustering by removing any rows (Universities) that have missing values (NaN) in any of the variables.
* Clustering algorithms require complete data to compute distances accurately, missing values would break this.
* Before Vs. After Cleaning

|  |  |
| --- | --- |
| Step | Count of Records |
| Original Dataset | 1302 Universities |
| After dropping missing rows | 471 Universities |

* This step ensures:
* Accurate distance calculations between universities.
* Prevents errors in clustering algorithms that cannot handle NaNs.
* Creates a solid, clean foundation for meaningful clustering.

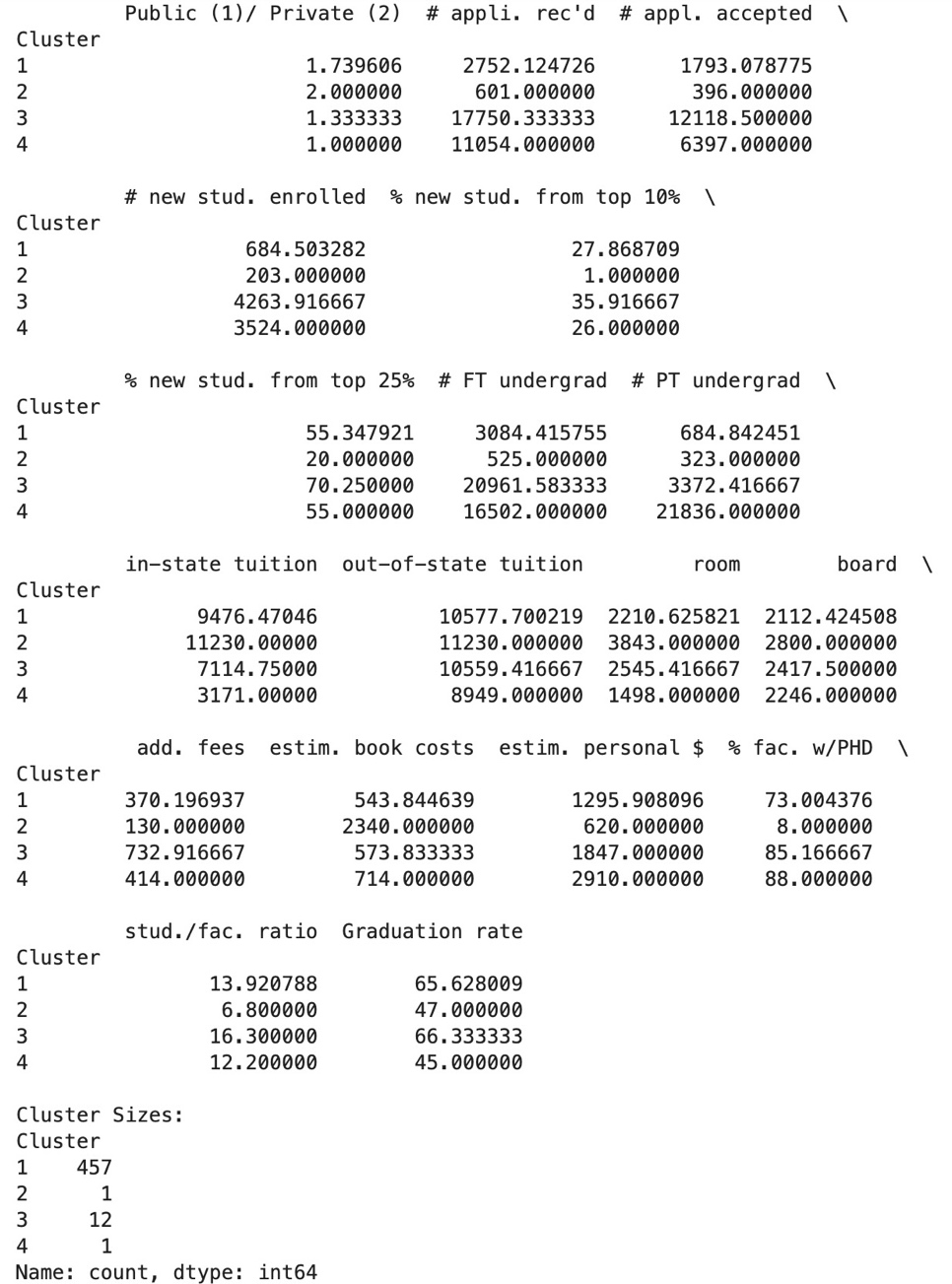
1B

* Extracted Continuous Variables:
* Selected only the numerical (float) columns from the cleaned dataset, as these are appropriate for clustering based on distance.
* Number of numerical columns used: 17.
* Normalization:
* Applied Z-Score Normalization to scale each feature to a mean of 0 and standard deviation of 1. This step is crucial because features like tuition and faculty ratio to have vastly different scales.
* Hierarchical Clustering:
* Used SciPy’s linkage () method to perform hierarchical clustering with:
* Method=’complete’ 🡪 uses the farthest distance between points in clusters.
* Metric=’euclidean’ 🡪 standard straight-line distance.
* Dendrogram:
* A graph showing a clustering diagram

  AI-generated content may be incorrect.
* The vertical lines indicate the distance between the merged clusters.
* Large jumps in height suggest good cut-off points for clusters.
* Around height 15-20, there is a clear big jump, splitting the data into a few well-separated groups.
* Went ahead with 4-cluster as it is the most reasonable option based on the dendrogram, based on it capturing distinct structures without overfitting.

1C

* After hierarchical clustering was complete and clusters were assigned, I grouped the dataset by cluster and computed the mean of each numerical variable.
* Cluster 1 – 457 Universities:
* Tuition: Moderate ($9,476 in-state / $10,576 out-of-state)
* Size: Mid-sized (3,000 FT undergrads)
* Selectivity: Moderate (28% from top 10%)
* Faculty: 73% with PhD’s
* Graduation Rate: 65%
* Student-Faculty Ratio: 14
* This cluster represented a mix of public and private colleges offering reasonably good academics at moderate cost. These are likely regional public universities and smaller private colleges.
* Cluster 2 – 1 University (Outlier)
* Tuition: High ($11,230 both in- and out-of-state) 🡪 suggest Private.
* Faculty: Very low (only 8% with PhD’s)
* Book Costs: Extremely high ($2340)
* Graduation Rate: Low (47%)
* A private outlier university with low faculty credentials and high student expenses. Possibly a for-profit or under-resourced institution.
* Cluster 3 – 12 Universities
* Tuition: Slightly lower in-state ($7,115) but same out-of-state as others ($10,559).
* Size: Large (20,961 FT undergrads)
* Selectivity: High (35.9% from top 10%)
* Faculty: Highest PhD % (85%)
* Graduation Rate: Highest (66.3%)
* Student-Faculty Ratio: 16.3
* These are prestigious and selective institutions, likely including Ivy League schools, top private universities, and flagship public institutions with excellent academic outcomes.
* Cluster 4 – 1 University (Outlier)
* Tuition: Very Low ($3,171 in-state)
* Size: Huge (16,500 FT / 21,836 PT)
* Faculty: 88% PhD’s
* Graduation Rate: 45%
* High Personal Costs: $2910
* A large, public, commuter-style university – probably located in an urban setting with minimal on-campus life.
* May serve non-traditional students or have open enrollment.
* **Cluster 1** *is the “typical” college group.*
* **Cluster 3** *includes top-tier institutions.*
* **Clusters 2 and 4** *are unique cases/outliers.*



1D

* Public vs Private Breakdown by Cluster:
* Grouped universities by cluster and counted how many were public or private.
* **Cluster 1 (Main group)**
* 119 Public and 338 Private 🡪 mostly private.
* Top States: PA, NY, OH, NC, MA, TX, IA, MO, IN, VA.
* This cluster reflects a broad national representation, mostly private colleges but with public institutions as well (typically regional colleges and small private universities)
* **Cluster 2 (1 University)**
* Private
* State: Michigan (MI)
* A small, isolated private college in MI with unusual characteristics. Could be a for-profit or non-traditional institution.
* **Cluster 3 (12 Universities)**
* 8 Public and 4 private.
* States: MA, TX, AZ, CA, IL, MI, NJ, NY, PA.
* These are selective and reputable institutions spread across the U.S., showing a balanced public-private split.
* **Cluster 4 (1 University)**
* Public
* State: Minnesota (MN)
* A massive public university – probably a commuter-heavy, urban campus with high part-time enrollment.

A screenshot of a computer

AI-generated content may be incorrect.

* This image displays the breakdown of public vs. private institutions and the top contributing states for each of the four clusters.

1E

* **Cluster 1 – “Typical Colleges” (457 Universities)**
* Many of these are smaller-to-mid-sized private liberal arts colleges and regional public universities.
* States like PA, NY, OH, NC are known for having a high concentration of such institutions.
* These schools typically serve local populations with moderate selectivity and affordability.
* May rely more on tuition revenue than large endowments or public subsidies.
* This cluster captures the majority of U.S. colleges, which focus on accessibility and affordability, rather than prestige.
* They offer practical degrees and draw students regionally.
* **Cluster 2 – “Private Outlier” (1 University)**
* Located in Michigan.
* Very low % of PhD faculty (8%).
* Extremely high book costs and low graduation rate.
* This school is likely a for-profit or unranked private institution with limited faculty qualifications.
* These institutions often have a business or tech-skewed curriculum, fewer full-time faculty, and serve non-traditional students.
* **Cluster 3 – “Elite a & Flagship Schools” (12 Universities)**
* High % of top 10% students.
* Large enrollments with highly qualified faculty.
* Includes public and private schools from prestigious states (CA, NY, MA, IL, TX).
* These are likely Ivy League Schools, top private colleges, and flagship public universities. They benefit from:
* Large endowments.
* High national/international rankings.
* Strong research programs.
* Competitive admissions.
* **Cluster 4 – “Big Public Commuter School” (1 University)**
* Located in Minnesota.
* Massive part-time enrollment, low tuition, high personal expenses.
* Likely a large public commuter university, serving a wide demographic including working adults, transfer students, and commuters.
* Probably part of a state university system, with a focus on accessibility rather than elite academics.
* The clusters align with real-world segments of U.S. higher education:
* Elite schools driven by prestige and performance.
* Public flagship balancing access and research.
* Smaller private/public institutions that focus on regional service.
* Outlier institutions that reflect niche or alternative models.

1F

* Filtered the dataset and confirmed that Tufts is in row 457 and found that only missing value is:
* # PT Undergrad
* Used only columns where Tufts has complete values for comparison.
* These were passed into a Euclidean distance calculation against the centroid (mean vector) of each cluster.
* Tufts was closest to Cluster 1, which contains 457 universities mostly regional public and private institutions.
* Based on similarity across available variables, Tufts fits best with Cluster 1, a group of well-rounded, mid-sized private/public schools.
* Using the cluster average makes the imputation statistically and contextually sound.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.