# Patterns/ Connections Among different activities of students that show their technical skills and abilities?

Kevin Duran
Dustin Enriquez
Angel Contreras

Instructor:
Liuliu Fu

California State University, Los Angeles

December 10, 2024

**Abstract**
This study aims to explore patterns in student technology use and technical ability through an integrated analysis of both pre- and post-survey data. The pre-survey evaluated familiarity with web platforms and technical systems such as social media, UI familiarity, and other web sources, with responses scored to compute "technical skill" levels based on weighted question categories. This was further explored through Linear Regression Analysis for tech-related questions. The post-survey(After) assessed engagement with a tagging system introduced during an experiment, by employing Correlation and *k*-means clustering Analysis to uncover performance patterns. Insights from these analyses revealed how specific questions can influence instances such as overall technical scores and how students grouped into distinct clusters based on their engagement. Analysis performed through Excel and Tableau highlighted connections between technical ability and system usability, offering actionable strategies for improving system design and tailoring educational tools to diverse skill levels.

## I. Data Analysis: Exploring Student Technology Use Through Means of Data Mining and Linear Regression: Pre Survey

### A. Introduction

In this analysis, responses from a technology-focused survey were looked at and analyzed in order to reach findings regarding the technological skills of those who took the survey. In specific, this analysis takes a look at the “Pre Survey”, which mainly consists of questions and responses that relate to the user’s previous experience on the Internet of

Things and various web applications/social media platforms. Each question is in a short-response format that tries to capture the user's current level of technical knowledge and skill. Afterward, through methods of data mining and regression in Microsoft Excel, we are able to see, not only the technical skills of the users in the process but what specific questions benefitted more towards the scores of the users.

### B. Data Preparation

With the dataset from this survey containing 35 mainly short-answer questions and answers, the data was largely raw, meaning there had to be some organization and structure in place to continue the analysis. For every question, each response was read through and scored on a scale from 1 to 5, with 1 being the lowest and 5 being the highest as it felt appropriate with each question asking the user about their familiarity with a certain system. After scores were given, the sum of scores for each user was calculated and, for this analysis, the top 10 and bottom 10 users in terms of the sum of their scores were chosen.

For this analysis, to find out the technical skill of the top and bottom 10 scores, questions labeled "knowing" or "advanced" by the determination of our team were used. In specific, these were questions 1, 4, 7, 10, 11, 12, 13, 27, 28, 29, and 30. Using these questions, we were able to find out the sum of "technologically advanced" points that users got compared to their overall sum of points. This is reflected in the image below:

Figure I. *Top and bottom 10 Pre Survey takers showing their points gained from technologically advanced questions, total points, and percentage of total points that are from technologically advanced questions, Ordered by the least to greatest total points.*

| | G | H | I | J |
|---|---|---|---|---|
| 1 | | Points Gained From "Technologically Advanced Questions" | Total Points | Percentage of Total Points that are from "Technologically Advanced Questions" |
| 2 | cfloy005 | 29 | 104 | 27.88% |
| 3 | jvene002 | 33 | 103 | 32.04% |
| 4 | rmurt002 | 37 | 103 | 35.92% |
| 5 | slupt001 | 27 | 101 | 26.73% |
| 6 | kcrou006 | 36 | 98 | 36.73% |
| 7 | slueh001 | 26 | 98 | 26.53% |
| 8 | mbrow094 | 28 | 97 | 28.87% |
| 9 | cjone122 | 31 | 95 | 32.63% |
| 10 | ltupm001 | 30 | 95 | 31.58% |
| 11 | psilv002 | 30 | 95 | 31.58% |
| 12 | dharv007 | 15 | 57 | 26.32% |
| 13 | esled001 | 21 | 56 | 37.50% |
| 14 | mrowe007 | 15 | 56 | 26.79% |
| 15 | alins003 | 19 | 55 | 34.55% |
| 16 | mslus002 | 15 | 52 | 28.85% |
| 17 | norne001 | 22 | 52 | 42.31% |
| 18 | mbajr555 | 12 | 44 | 27.27% |
| 19 | kchis002 | 12 | 43 | 27.91% |
| 20 | jturn041 | 15 | 41 | 36.59% |
| 21 | ctuck020 | 11 | 34 | 32.35% |

### C. Data Mining

From the results seen in Figure I, we can see that, even though some point totals are higher than others, someone with a lower point total can have more “technologically advanced” points. An example can be the user with the most amount of points overall, cfloy005, having fewer “technologically advanced” points than the runner-up, rmurt002 with 29 and 37 “technologically advanced” points respectively. Also, with “knowing” and “advanced” being two different levels of technical knowledge or intensity, we made a formula to calculate technical knowledge on this survey based on the two types of questions:

$$\text{□□□}h\text{□□□□□ □□□□□} = \text{□□□ □□ "□□□□□□□□" □□□□□□} + (2 * \text{□□□ □□ "□□□□□□□□□" □□□□□□})$$

Using this formula, we were able to find and sort the technical skills of users based on responses from the survey. The list of technical scores from highest to lowest is as follows:

| | Highest/Lowest Scores Ranked on "Technical Skill" Score (Formula: (Sum of scores from "Knowing" questions + (2 x (Sum of scores from "advanced" questions)) |
|---|---|
| rmurt002 | 55 |
| kcrou006 | 51 |
| jvene002 | 47 |
| psilv002 | 43 |
| cjone122 | 41 |
| ltupm001 | 39 |
| cfloy005 | 36 |
| mbrow094 | 35 |
| slueh001 | 35 |
| slupt001 | 35 |
| norne001 | 31 |
| esled001 | 25 |
| alins003 | 24 |
| dharv007 | 20 |
| mrowe007 | 19 |
| jturn041 | 18 |
| mslus002 | 18 |
| kchis002 | 15 |
| mbajr555 | 15 |
| ctuck020 | 14 |

*Figure II. “Technical Skill” of users based on formula from greatest to least.*

From the cells in Figure II, we’re able to see that even though user cfloy005 had the most amount of points overall, their score on “technical skill” pales in comparison to the user rmurt002, who was second in overall points but with the most amount of “technical skill” points. Suddenly, because of the change made in counting the points, the bottom 10 scores seem a lot closer to the top 10 scores. In terms of the bigger picture, we’re able to see a more contextualized score that puts more weight on the questions that require more skill for a higher score.

**D. Regression Analysis**

Through the use of the "technical skill" scores, we're able to conduct a regression analysis on the scores for both the "knowing" and "advanced" questions in order to find the impact of every question on someone's technical score. Using the number of points each user got from each "knowing" and "advanced" question, we were able to conduct the following regression analyses in Excel:

| SUMMARY OUTPUT | |
|---|---|
| | |
| *Regression Statistics* | |
| Multiple R | 0.88919937 |
| R Square | 0.790675519 |
| Adjusted R Square | 0.715916776 |
| Standard Error | 6.809431907 |
| Observations | 20 |
| | |
| ANOVA | |
| | *df* |
| Regression | 5 |
| Residual | 14 |
| Total | 19 |
| | |
| | *Coefficients* |
| Intercept | 2.007074494 |
| 1 | 2.141600471 |
| 4 | 4.662600775 |
| 7 | 3.329451776 |
| 10 | -8.718489669 |
| 13 | 0.07475421 |

*Figure III. Regression Analysis for dependent value ("technical skill" score) and independent values (scores for questions 1, 4, 7, 10, 13; "knowing" questions)*

| SUMMARY OUTPUT | | |
|---|---|---|
| | | |
| *Regression Statistics* | | |
| Multiple R | | 0.976682835 |
| R Square | | 0.95390936 |
| Adjusted R Square | | 0.741618522 |
| Standard Error | | 3.086921807 |
| Observations | | 20 |
| | | |
| ANOVA | | |
| | | *df* |
| Regression | | 7 |
| Residual | | 15 |
| Total | | 22 |
| | | |
| | | *Coefficients* |
| Intercept | | 6.582238333 |
| | 8 | 3.061531491 |
| | 11 | 7.086542818 |
| | 12 | 0 |
| | 27 | 3.623742637 |
| | 28 | 5.030403262 |
| | 29 | 0 |
| | 30 | 0 |

*Figure IV. Regression Analysis for dependent value ("technical skill" score) and independent values (scores for questions 8, 11, 12, 27, 28, 29, 30; "advanced" questions)*

With the information above from Figure III and Figure IV, we're able to analyze and contextualize a few things: Coefficients and R Square.

1.) **Coefficients**
In the regression analyses done, the coefficients mean, on average, how much the "technical skill" score (Y value) is going to increase with every singular point increase to the points from the individual questions (X value). That means that, on average, questions 12, 13, 29, and 30 have essentially zero effect on the "technical skill" score of the users. That being said, for "knowing" questions, question 4 seems to have the most positive impact, and question 10 has the most, and only, negative impact. Question 4 had a coefficient of about 4.7 points and question 10 had a coefficient of about -8.7 points. This means that for every point a user goes up in question 4, on average, their "technical skill" points as a whole is likely to go up by about 4.7 points. On the other hand, for question 10, every point you gain for that question averages about 8.7 points loss, on average, to someone's "technical skill" score. This implies that those who got overall low "technical

skill" scores scored highly on question 10, whereas those who got high "technical skill" scores didn't. For the "advanced" questions, question 11 has the most positive impact, with "technical skill" scores going up by about 7.1 for every point gained on question 11. Within the "advanced" questions, there are no questions that have a negative impact on an increase in points for that respective question.

2.) **R-Square**
For the "knowing" and "advanced" regression analyses, they received R-squared values of about 0.79 and 0.95 respectively. This means that the "knowing" questions are very closely tied to the independent variable, which is the "technical skill" score. The approximate 0.79 R-squared value means that about 79% of all variation that goes on with the "technical skill" score can be explained by the regression model. That being said, with a value of 0.95, the "advanced" questions can almost be said to fit the model perfectly as they would explain about 95% of the variation in the regression model. With both values being so high (since R-squared ranges from 0 to 1), we can conclude that learnings from this analysis can be applied to the rest of the users with sufficient accuracy.

E. **Key Findings**

Throughout the data mining and regression analysis sections of this analysis, a few findings were able to be made:

1.) Having the highest overall score doesn't mean having the most technical prowess. When only considering the number of points people gained from "knowing" and "advanced" questions, the gap between the top 10 and bottom 10 overall scores closes, showing the potential difference on the entire dataset if scores are to be weighted the entire way through.
2.) Despite questions being labeled and weighted based on "knowing" and "advanced", the regression analysis still shows the impact of certain technical skills in this context. For example, in the regression analyses that were done, the skills from being familiar with online multimedia repositories (knowing question) have more value in this dataset than having skills with any virtual reality sites (advanced question). Also, there

are values that are very much NOT valued like being familiar with social bookmarking sites as they seemed to negatively impact “technical skill” point totals more than positively. That being said, however, it’s interesting to note that being a lot more experienced with social bookmarking sites is more valued than simply being familiar. Figure IV’s regression analysis shows that being far more experienced than familiar with social bookmarking sites is very valuable, projecting about a 7-point increase to a user’s “technical skill” total for every point gained. Even though they were labeled as “knowing” or “advanced”, it’s good to note that some skills weren’t very valued in the regression analysis. Of these skills, being familiar with blogs or forums (knowing) and being familiar with online games (advanced) had no effect on the “technical skill” scores of users.

## II. Data Analysis Exploring Student Technology Use through Correlation and Clustering Post(After) Survey

This analysis investigates patterns in student responses to a technology-focused survey, aiming to uncover connections between their technical ability and their perceptions of system usability. The dataset, titled “After Survey,” comprises responses to twelve multiple-choice questions rated on a scale (1–5), along with a short-answer question. Each question evaluates aspects of a tagging system that was introduced during an experiment. According to Mingle and Adams (2015), students often engage with social media in ways that may enhance communication but may also detract from their academic focus. This observation aligns with the results of our study, where varying degrees of technical ability and engagement were evident among students as they interacted with the tagging system. Tóth et al. (2021) also look into this by exploring how "digital technology can help modernize the education process and make it more effective, assuming that digital technology is used to motivate students to learn or to make learning more time efficient." This perspective aligns with the findings of this analysis, where the integration of technology-focused questions and data-driven clustering methods revealed insights into student engagement and technical skills development. By employing correlation analysis and clustering techniques, we seek to highlight relationships between specific questions and overall performance while identifying distinct student groups based on their responses. The clustering results, particularly the high-scoring students that we found also align with Mingle and Adams' (2015) findings that focused use of technology can promote better academic engagement, provided distractions are minimized. The analysis was conducted using Excel for data preparation and Tableau for advanced visualization.

### Data Preparation

The dataset initially contained raw responses to twelve questions, usernames, and an overall score calculating the sum of responses to Questions 1 through 12. However, to continue further analysis additional fields had to be generated. This includes fields like assigned clusters, distances to centroids, and the sum of squared errors (SSE) which were all generated using Excel. These fields were essential for implementing the clustering analysis process and evaluating its performance and effectiveness.

*Original dataset columns*

| | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Username | Score | Answer 1 | Answer 2 | Answer 3 | Answer 4 | Answer 5 | Answer 6 | Answer 7 | Answer 8 | Answer 9 | Answer 10 | Answer 11 | Answer 12 | Answer 13 | | |
| 2 | aalph002 | 19 | 1 | 3 | 1 | 1 | 1 | 1 | 3 | 3 | 1 | 1 | 2 | 1 | The system and assignment | | |
| 3 | abaxx004 | 55 | 4 | 5 | 3 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | More than getting knowledg | | |
| 4 | alule005 | 54 | 5 | 5 | 3 | 5 | 4 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | I learned how the librarians | | |
| 5 | ametc003 | 36 | 3 | 4 | 3 | 2 | 4 | 2 | 4 | 3 | 2 | 3 | 4 | 2 | I thought that i was 1 makin | | |
| 6 | aneel004 | 53 | 4 | 5 | 3 | 5 | 4 | 4 | 5 | 5 | 4 | 4 | 5 | 5 | At first, the system was kind | | |
| 7 | arees002 | 43 | 3 | 4 | 2 | 4 | 5 | 5 | 5 | 3 | 2 | 3 | 4 | 3 | <Unanswered> | | |
| 8 | asing053 | 49 | 4 | 5 | 3 | 3 | 5 | 2 | 4 | 5 | 4 | 5 | 5 | 4 | When creating associations, | | |
| 9 | awils031 | 52 | 4 | 5 | 3 | 3 | 5 | 5 | 5 | 5 | 4 | 4 | 5 | 4 | <Unanswered> | | |
| 10 | b1006 | 49 | 5 | 5 | 3 | 5 | 5 | 2 | 5 | 5 | 3 | 1 | 5 | 5 | <p class="MsoNormal" style | | |
| 11 | cbull055 | 45 | 4 | 4 | 4 | 3 | 4 | 3 | 5 | 4 | 4 | 3 | 3 | 4 | <Unanswered> | | |
| 12 | ccoxx058 | 51 | 4 | 4 | 4 | 4 | 5 | 4 | 5 | 3 | 5 | 3 | 5 | 5 | <Unanswered> | | |
| 13 | ccrea005 | 53 | 4 | 5 | 4 | 4 | 5 | 3 | 5 | 4 | 5 | 5 | 5 | 4 | this is a very good system | | |
| 14 | cespu005 | 45 | 4 | 5 | 3 | 3 | 3 | 4 | 5 | 4 | 3 | 2 | 5 | 4 | N5 | | |
| 15 | cfloy001 | 45 | 3 | 3 | 4 | 5 | 5 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | <p class="MsoNormal" style | | |
| 16 | chend049 | 39 | 4 | 2 | 3 | 3 | 4 | 3 | 5 | 3 | 3 | 4 | 3 | 2 | In phase 5 I had issue while | | |

### Correlation Analysis

To begin, since the overall score was calculated by summing the values for each respondent's answers we can go on next to correlation analysis. This was performed to identify which survey questions had the strongest relationship with the overall score. Using Excel's CORREL function, a correlation matrix was created to measure linear relationships between the questions and the overall score. This helped reveal that Question 9 ("How easy was it to understand the assignment directions?") and Question 12 ("In general, rate the user-friendliness of the system") had the highest correlation coefficients with the overall score, at 0.746 and 0.744. With these findings, these two questions were selected as the primary variables for clustering and further exploration.

*Correlation Matrix*

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | *Score* | *Answer 1* | *Answer 2* | *Answer 3* | *Answer 4* | *Answer 5* | *Answer 6* | *Answer 7* | *Answer 8* | *Answer 9* | *Answer 10* | *Answer 11* | *Answer 12* |
| 2 | Score | 1 | | | | | | | | | | | | |
| 3 | Answer 1 | 0.623441 | 1 | | | | | | | | | | | |
| 4 | Answer 2 | 0.586573 | 0.313852 | 1 | | | | | | | | | | |
| 5 | Answer 3 | 0.561722 | 0.296775 | 0.222479 | 1 | | | | | | | | | |
| 6 | Answer 4 | 0.588995 | 0.321084 | 0.406071 | 0.409335 | 1 | | | | | | | | |
| 7 | Answer 5 | 0.537967 | 0.415341 | 0.237693 | 0.353464 | 0.3962 | 1 | | | | | | | |
| 8 | Answer 6 | 0.599246 | 0.19626 | 0.248598 | 0.363383 | 0.382512 | 0.361154 | 1 | | | | | | |
| 9 | Answer 7 | 0.511486 | 0.326407 | 0.284862 | 0.177734 | 0.171563 | 0.136157 | 0.286159 | 1 | | | | | |
| 10 | Answer 8 | 0.608727 | 0.292463 | 0.348614 | 0.153448 | 0.180861 | 0.19287 | 0.288498 | 0.322703 | 1 | | | | |
| 11 | Answer 9 | 0.74568 | 0.403968 | 0.299096 | 0.504591 | 0.268453 | 0.268641 | 0.405065 | 0.261568 | 0.391941 | 1 | | | |
| 12 | Answer 10 | 0.679858 | 0.37407 | 0.356735 | 0.270569 | 0.293861 | 0.179806 | 0.281356 | 0.270613 | 0.35491 | 0.644297 | 1 | | |
| 13 | Answer 11 | 0.630672 | 0.376699 | 0.276133 | 0.198392 | 0.263567 | 0.255231 | 0.231238 | 0.290937 | 0.370747 | 0.409762 | 0.461287 | 1 | |
| 14 | Answer 12 | 0.743999 | 0.420458 | 0.424823 | 0.211998 | 0.323261 | 0.269321 | 0.326277 | 0.403061 | 0.621332 | 0.487556 | 0.414992 | 0.632765 | 1 |

## Clustering Analysis

In order to better segment students into meaningful groups, the *k*-means clustering algorithm was implemented in Excel. This algorithm aims to partition data into *k*-clusters, where each data point belongs to the cluster relating with the nearest centroid. The process involved many different steps before completion and visualization:

1. **Initialization**: Centroids for three clusters were initialized with initial centroids for three clusters (Cluster 1, Cluster 2, and Cluster 3) which were three users who scored differently overall within the survey as well as corresponding to the selected questions (Questions 9 and 12).

2. **Distance Calculation**: Using the Euclidean distance formula, the distance of each student's responses to each centroid was computed:

   $$d=\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
   **Where:** x2 and $y_2$ are the coordinates of the data point (Question 9 and 12 scores.
   $x_1$ and $y_1$ are the coordinates of the centroid.
   Excel's **SQRT** and **SUMXMY2** functions were used to automate these calculations.
   Excel Formula: **=SQRT(SUMXMY2()) =SQRT((Answer9 - CentroidX)^2 + (Answer12 - CentroidY)^2)**
   A table was generated to store these distances for each user across all centroids, enabling the assignment of each user to their nearest cluster.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Username | Score | Answer 1 | Answer 2 | Answer 3 | Answer 4 | Answer 5 | Answer 6 | Answer 7 | Answer 8 | Answer 9 | Answer 10 | Answer 11 | Answer 12 |
| 2 | aalph002 | 19 | 1 | 3 | 1 | 1 | 1 | 1 | 3 | 3 | 1 | 1 | 2 | 1 |
| 83 | revan040 | 42 | 3 | 4 | 2 | 3 | 2 | 2 | 5 | 5 | 2 | 4 | 5 | 5 |
| 50 | jvene004 | 55 | 4 | 5 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 5 |

*Users chosen*

3. **Cluster Assignment**
   Each user was assigned to a cluster corresponding to the smallest distance calculated. The clusters were labeled as Cluster 1, Cluster 2, and Cluster 3. Once assigned, new centroids were calculated based on the mean values of the Question 9 and Question 12 scores within each cluster.
4. **Iteration and Stabilization**
   The process of reassigning clusters and recalculating centroids was repeated until the clusters stabilized. Stability was defined as the point when the users no longer switched between clusters in repeated subsequent iterations. This is how I knew that the values were stabilized.
5. **SSE Evaluation**
   To validate the clustering results, the Sum of Squared Errors (SSE) was also computed overall and for each cluster. This measure helps with quantifying the compactness of clusters. For example, lower SSE values usually indicate better defined clusters. After multiple iterations, the SSE stabilized at a value that reflected the optimal clustering solution.

| T | U | V |
|---|---|---|
| SUM of SSE by Cluster | Total SSE | Cluster Numbers |
| 38.5 | 761.633 | 1 |
| 462.133781 | | 2 |
| 260.9992604 | | 3 |

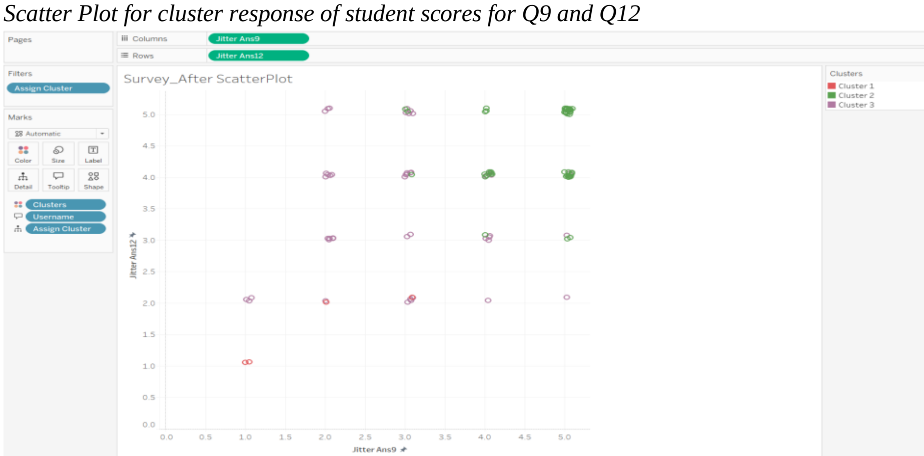
*SSE generated data*

**SSE and Scatterplot Insights**

Following the SSE evaluation, a scatter plot was created to further visualize the results. This plot used a created calculated measure "Jitter Ans9" and "Jitter Ans12" on the axes to minimize overlapping points, ensuring every user was visible. It is important to note however the random influxations from these Clusters are not imperative to their scores but rather only used for visibility of all users since there was much overlapping. The clusters were distinguished by colors, and usernames were added as a tooltip for detailed analysis. This plot provided clarity on the relationship between the responses to Questions 9 and 12, particularly in identifying distinct groupings. The separation between clusters in the scatter plot aligned with the SSE results, indicating that the clustering process effectively captured variations in user responses.


*Created calculated field*

*Scatter Plot for cluster response of student scores for Q9 and Q12*

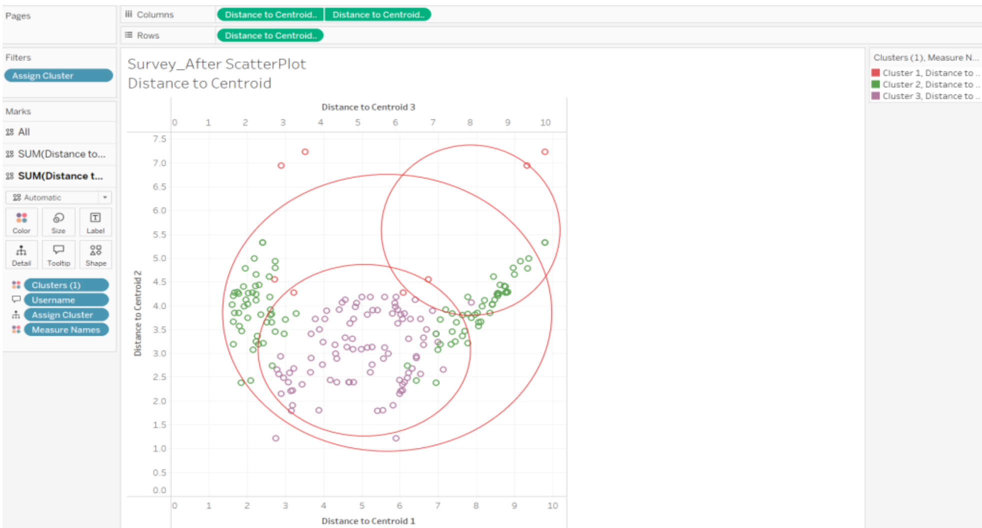



**Additional Visualization: Distance to Centroids Scatterplot**

This scatterplot uses the calculated distances to centroids to visualize the relationship between each user's proximity to the cluster centroids. The axes represent the distances to Centroid 1 and Centroid 2, and the colors correspond to the assigned clusters (Cluster 1, Cluster 2, and Cluster 3). By plotting these distances, the visualization helps evaluate the relative compactness and spread of each cluster, revealing how closely users align with their assigned clusters**.**

1. **Cluster 1 (Red):** Points in Cluster 1 are distinctly separated from the other clusters. These users show minimal overlap with other clusters, suggesting that their responses are highly distinct and closely aligned to Centroid 1. This separation likely reflects unique response patterns, such as low technical ability or minimal familiarity with the system.

2. **Cluster 2 (Green):** While Cluster 2 has a strong concentration of points near Centroid 2, it also demonstrates some spread toward Centroid 3. This indicates variability within this cluster, as some users' responses align closely with Centroid 2, while others are more distributed. This spread suggests that Cluster 2 represents users with a mix of intermediate or moderate technical skills and system perceptions.
3. **Cluster 3 (Purple):** Points in Cluster 3 are primarily clustered near Centroid 3, but there is a moderate spread across the central region of the plot. This suggests a moderate degree of compactness for Cluster 3, with responses that are relatively consistent compared to Cluster 2. Users in this cluster might represent those who fall in the middle range of technical ability and system usability.
4. **Overlapping Points:** The overlap between Clusters 2 and 3, as seen in areas where green and purple points are close together, indicates some shared response patterns. This overlap might reflect users with borderline characteristics that place them between moderate and intermediate technical skills or system usability.




*Distance to Centroids Scatterplot*

## Insights and Connection to the Main Objective

This visualization complements the previous scatterplot by illustrating how users' responses align with the centroids of their respective clusters.

**Key findings:**

- The clear separation of Cluster 1, showing that this group represents a distinct subset of users.
- The variability within Cluster 2, reflecting a diverse set of responses and technical abilities.
- The moderate compactness of Cluster 3, which suggests consistent response patterns but some overlap with Cluster 2.

**Restate Main Discoveries**

This analysis provided valuable insights into how students' survey responses, technical skills, and perceptions of the tagging system are interconnected. By leveraging correlation analysis and *k*-means clustering, distinct patterns and relationships were identified, shedding light on the diversity of student interactions with the system. The correlation analysis revealed that Questions 9 and 12 had the strongest relationships with the overall score, suggesting that these questions play a pivotal role in assessing students' familiarity and satisfaction with the system.
These findings served as the foundation for clustering, as they represented key indicators of technical ability and system usability.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Username | Score | Answer 1 | Answer 2 | Answer 3 | Answer 4 | Answer 5 | Answer 6 | Answer 7 | Answer 8 | Answer 9 | Answer 10 | Answer 11 | Answer 12 | Distance to Centroid 1 | Distance to Centroid 2 | Distance to Centroid 3 | Assign Cluster | Squared Distance | SUM of SSE by Cluster | Total SSE | Cluster Numbers |
| 2 | aalph002 | 19 | 1 | 3 | 1 | 1 | 1 | 1 | 3 | 3 | 1 | 1 | 2 | 1 | 3.51781182 | 7.224163139 | 9.985899823 | Cluster 1 | 12.375 | 38.5 | 761.633 | 1 |
| 3 | abaxx004 | 55 | 4 | 5 | 3 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | 8.595056719 | 4.243431969 | 1.815246083 | Cluster 3 | 3.295118343 | 462.133781 | | 2 |
| 4 | alule005 | 54 | 5 | 5 | 3 | 5 | 4 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | 8.624094155 | 4.432051276 | 2.358887261 | Cluster 3 | 5.564349112 | 260.9992604 | | 3 |
| 5 | ametc003 | 36 | 3 | 4 | 3 | 2 | 4 | 2 | 4 | 3 | 2 | 3 | 4 | 2 | 3.65718471 | 2.594992794 | 5.326541225 | Cluster 2 | 6.733987603 | | | |
| 6 | aneel004 | 53 | 4 | 5 | 3 | 5 | 4 | 4 | 5 | 5 | 4 | 4 | 5 | 5 | 8.054501847 | 3.562096722 | 1.836311907 | Cluster 3 | 3.37204142 | | | |
| 7 | arees002 | 43 | 3 | 4 | 2 | 4 | 5 | 5 | 5 | 3 | 2 | 3 | 4 | 3 | 5.689903338 | 2.985934659 | 4.410098014 | Cluster 2 | 8.915805785 | | | |
| 8 | asing053 | 49 | 4 | 5 | 3 | 3 | 5 | 2 | 4 | 5 | 4 | 5 | 5 | 4 | 7.373940602 | 3.451927836 | 2.791973915 | Cluster 3 | 7.795118343 | | | |
| 9 | awils031 | 52 | 4 | 5 | 3 | 3 | 5 | 5 | 5 | 5 | 4 | 4 | 5 | 4 | 7.705517504 | 3.465070688 | 1.898107123 | Cluster 3 | 3.602810651 | | | |
| 10 | b1006 | 49 | 5 | 5 | 3 | 5 | 5 | 2 | 5 | 5 | 3 | 1 | 5 | 5 | 7.866066361 | 4.068435746 | 4.581203558 | Cluster 2 | 16.55216942 | | | |
| 11 | cbull055 | 45 | 4 | 4 | 4 | 3 | 4 | 3 | 5 | 4 | 4 | 3 | 3 | 4 | 5.989574275 | 2.144207072 | 2.952659016 | Cluster 2 | 4.597623967 | | | |
| 12 | ccoxx058 | 51 | 4 | 4 | 4 | 4 | 5 | 4 | 5 | 3 | 5 | 3 | 5 | 5 | 7.640353395 | 3.365248385 | 2.309454455 | Cluster 3 | 5.333579882 | | | |
| 13 | ccrea005 | 53 | 4 | 5 | 4 | 4 | 5 | 3 | 5 | 4 | 5 | 5 | 5 | 4 | 8.11634154 | 3.656549089 | 1.683324057 | Cluster 3 | 2.833579882 | | | |
| 14 | cespu005 | 45 | 4 | 5 | 3 | 3 | 3 | 4 | 5 | 4 | 3 | 2 | 5 | 4 | 6.072478901 | 2.34663906 | 3.506439554 | Cluster 2 | 5.506714876 | | | |
| 15 | cfloy001 | 45 | 3 | 3 | 4 | 5 | 5 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 6.62382065 | 3.177675875 | 4.392620897 | Cluster 2 | 10.09762397 | | | |
| 16 | chend049 | 39 | 4 | 2 | 3 | 3 | 4 | 3 | 5 | 3 | 3 | 4 | 3 | 2 | 4.987484336 | 3.083296106 | 4.866182665 | Cluster 2 | 9.506714876 | | | |
| 17 | chend034 | 53 | 3 | 5 | 3 | 4 | 3 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 8.448372624 | 4.603100375 | 2.628783551 | Cluster 3 | 6.910502959 | | | |
| 18 | cholm05C | 47 | 5 | 5 | 5 | 3 | 4 | 3 | 5 | 2 | 4 | 4 | 4 | 3 | 6.955213872 | 3.405528442 | 3.75024654 | Cluster 2 | 11.59762397 | | | |
| 19 | cj5544 | 55 | 5 | 5 | 4 | 5 | 5 | 3 | 4 | 4 | 5 | 5 | 5 | 5 | 8.739279146 | 4.411491849 | 2.258940161 | Cluster 3 | 5.102810651 | | | |
| 20 | cmoor026 | 51 | 3 | 5 | 4 | 4 | 3 | 5 | 5 | 4 | 5 | 5 | 4 | 4 | 7.770135134 | 4.184102637 | 2.650639088 | Cluster 3 | 7.025887574 | | | |
| 21 | cnorm005 | 56 | 5 | 5 | 4 | 5 | 4 | 4 | 5 | 5 | 5 | 4 | 5 | 5 | 8.824681297 | 4.280758892 | 1.793932905 | Cluster 3 | 3.218195266 | | | |
| 22 | cowus005 | 49 | 4 | 3 | 3 | 3 | 5 | 4 | 5 | 3 | 5 | 5 | 5 | 4 | 7.474958194 | 3.644096891 | 2.700947675 | Cluster 3 | 7.295118343 | | | |
| 23 | cpico005 | 52 | 5 | 5 | 3 | 3 | 4 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | 8.20822758 | 4.112882904 | 2.275902838 | Cluster 3 | 5.179733728 | | | |
| 24 | cschn002 | 45 | 4 | 3 | 3 | 4 | 3 | 4 | 5 | 4 | 3 | 3 | 5 | 4 | 6.031169041 | 2.206887229 | 3.202601568 | Cluster 2 | 4.87035124 | | | |
| 25 | ctuck040 | 53 | 3 | 5 | 4 | 4 | 5 | 5 | 5 | 5 | 4 | 3 | 5 | 5 | 7.960841664 | 3.808768157 | 2.407305204 | Cluster 3 | 5.795118343 | | | |
| 26 | cvagn004 | 41 | 4 | 4 | 2 | 2 | 4 | 3 | 3 | 3 | 3 | 4 | 5 | 4 | 5.27967802 | 2.764611819 | 4.050987899 | Cluster 2 | 7.643078512 | | | |
| 27 | cwhit088 | 49 | 4 | 5 | 4 | 3 | 4 | 3 | 5 | 4 | 4 | 4 | 5 | 4 | 6.955213872 | 2.375516473 | 1.877734692 | Cluster 3 | 3.525887574 | | | |
| 28 | ddomi001 | 46 | 4 | 5 | 3 | 3 | 5 | 4 | 5 | 5 | 2 | 3 | 3 | 4 | 6.432340165 | 2.893155931 | 3.755370907 | Cluster 2 | 8.37035124 | | | |
| 29 | ddurr005 | 42 | 4 | 3 | 5 | 3 | 5 | 2 | 3 | 3 | 5 | 2 | 5 | 2 | 5.905505906 | 4.178667291 | 5.112853625 | Cluster 2 | 17.46126033 | | | |
| 30 | dreed009 | 56 | 5 | 5 | 4 | 5 | 5 | 4 | 5 | 5 | 5 | 3 | 5 | 5 | 8.796306043 | 4.270127352 | 2.180976113 | Cluster 3 | 4.756656805 | | | |
| 31 | dtayl026 | 37 | 3 | 3 | 2 | 2 | 4 | 2 | 4 | 5 | 2 | 2 | 4 | 4 | 4.623310502 | 3.134469125 | 5.380424193 | Cluster 2 | 9.824896694 | | | |
| 32 | dwatk055 | 43 | 4 | 3 | 3 | 3 | 4 | 3 | 5 | 4 | 4 | 3 | 4 | 3 | 5.419870847 | 1.785646398 | 3.226531103 | Cluster 2 | 3.188533058 | | | |
| 33 | ecare002 | 43 | 4 | 5 | 3 | 3 | 4 | 3 | 5 | 3 | 2 | 3 | 5 | 3 | 5.556527693 | 1.798329114 | 3.964617922 | Cluster 2 | 3.233987603 | | | |
| 34 | econn009 | 45 | 4 | 3 | 3 | 4 | 5 | 3 | 5 | 3 | 4 | 4 | 4 | 3 | 6.11350963 | 2.385064581 | 3.148095132 | Cluster 2 | 5.688533058 | | | |
| 35 | emart059 | 33 | 3 | 4 | 3 | 2 | 3 | 2 | 2 | 4 | 3 | 3 | 2 | 2 | 3.221024682 | 4.270127352 | 6.210020057 | Cluster 1 | 10.375 | | | |
| 36 | eroma004 | 39 | 4 | 3 | 3 | 2 | 4 | 1 | 5 | 3 | 2 | 4 | 5 | 3 | 5.601339126 | 3.112640849 | 5.246511944 | Cluster 2 | 9.688533058 | | | |
| 37 | esled005 | 60 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 9.791578014 | 5.317841869 | 2.454768334 | Cluster 3 | 6.025887574 | | | |
| 38 | hlixx008 | 39 | 4 | 4 | 2 | 2 | 2 | 1 | 5 | 3 | 4 | 4 | 5 | 3 | 6.031169041 | 3.808768157 | 5.242845223 | Cluster 2 | 14.50671488 | | | |
| 39 | iadug004 | 46 | 3 | 5 | 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 5 | 5 | 6.31466547 | 2.681143983 | 3.291440396 | Cluster 2 | 7.188533058 | | | |
| 40 | jcoop046 | 54 | 5 | 4 | 3 | 4 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | 8.41872912 | 4.023497615 | 1.648695019 | Cluster 3 | 2.718195266 | | | |

*Newly formatted dataset*

**Implications**

This analysis highlights the diversity of technical abilities among students, providing actionable insights for improving the tagging system and supporting students. For instance:

- **Cluster 1:** Students may benefit from targeted tutorials and system enhancements to address their challenges.

- **Cluster 2:** Insights from this group can inform best practices for system design and usability, leveraging their advanced technical skills.
- **Cluster 3:** This group may require a combination of support and system enhancements to bridge the gap between moderate and high technical abilities.

**Key Takeaways**

The final clustering results, combined with visualizations from Excel and Tableau, demonstrate how data analysis can help uncover meaningful patterns in student performance and technical ability. By using the inclusion of users with varied scores in cluster initialization, this allowed for a better comprehensive representation of the “After Survey” dataset for data analysis. The clustering results, particularly the high-scoring students in Cluster 2, align with Mingle and Adams' (2015) findings that focused use of technology can promote better academic engagement, provided distractions are minimized. This highlights the potential for targeted system designs to foster student engagement and improve outcomes. Future analyses could explore additional questions or incorporate qualitative feedback to build on these findings. This approach provides a strong foundation for developing tailored strategies to enhance both system design and the student experience.

**Conclusion**

This study set out to determine patterns in student technology use focusing on their technical skills and engagement with systems, as evaluated through our pre- and post-survey analyses. The initial analysis of the Pre Survey helped demonstrate the significance of questions in which we categorized as "knowing" or "advanced," (corresponding to technical relevancy) revealing how specific question types contribute to users' technical skill via their scores. Through data mining techniques, such as weighted scoring, we found that higher overall scores did not always equate to greater technical proficiency, providing an understanding of student capabilities. In the Post-Survey(After), clustering methods such as *k*-means and visualization tools like Excel and Tableau were used to uncover distinct groupings of students. These clusters helped illuminate relationships between specific survey responses and overall performance, suggesting that students with varying scores demonstrated distinct technical behaviors and engagement patterns. The insights gathered from this clustering analysis, particularly focused on the use of questions 9 and 12, suggest that certain aspects of technology use may correlate strongly with technical skill development.This study also raised new questions. Could external factors, such as prior access to technology or educational background, further contextualize the clusters? If the dataset provided more categorical data on the users we can explore how these student’s traits could affect their responses and scores. Ultimately, this project shows the value of data driven methods for analyzing student technology use. By identifying specific questions and patterns that align with technical skills, the findings provide a foundation for tailoring educational interventions, enhancing system usability, and supporting student learning especially in technology focused environments.

# References

**Mingle, J., & Adams, M. (2015).** Social media usage and academic performance: A study on university students. *International Journal of Social Media Studies, 3*(1), 45–54. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3093275

**Tóth, T., Virágh, R., Hallová, M., Stuchlý, P., & Hennyeyová, K. (2021).** Digital competence of digital native students as prerequisite for digital transformation of education. *International Journal of Emerging Technologies in Learning, 17*(16), 91–108. https://doi.org/10.3991/ijet.v17i16.31791