East West University
Department of CSE

CSE 488
Fall 2024
Date : 01/01/2025

Mini Project Name: Analyzing the Course Advising Dataset

Name of Student & Id:	**Course Instructor:**
Suraiya Nusrat Tanha 2021-2-60-030 Prinom Mojumder 2021-2-60-098 Tasnim Israk Synthia 2021-2-60-097	Dr. Mohammad Rezwanul Haq Associate Professor Department of Computer Science & Engineering East West University

Introduction:

This project analyzes student course selection patterns and credit trends by focusing on course co-occurrence to provide actionable insights. Based on a course advising dataset, where we get 2960 students advising information based on their complete credits, taken credits and taken courses. Now we are trying to analyze the information to understand how students choose their courses and identify the courses they have not taken. Our analysis focused on uncovering patterns in course selection, as well as exploring the demand for specific courses and the relationships between different courses.
We aimed to present these insights in a clear and structured manner. The project was implemented using Python, with a particular emphasis on utilizing Spark to handle and analyze the data effectively.

Dataset Characteristics and Exploratory Data Analysis:

Here in this Dataset there are 2960 row entries and 11 Column entries.

Attribute	**Description**	**Null Count**	**Data Type**
StudentId	Unique identifier for each student	0	Integer
CreditsCompleted	Total credits completed by the student	0	Float
takencredit	Total credits currently taken by the student	0	Float
takennocourse	Number of courses currently taken	0	Integer
C1	Course	0	Object
C2	Course	22	Object
C3	Course	83	Object
C4	Course	356	Object
C5	Course	2117	Object
C6	Course	2836	Object
C7	Course	2958	Object

Table 01: Summery

There are a total of 60 Courses which are offered.

Course Name	Total Offered Number	Taken
ACT	1	31
BUS	2	24
CE	1	1
CSE	32	7145
CHE	1	423
ECO	2	352
ENG	3	955
FIN	1	486
GEN	7	486
MAT	5	1535
MGT	1	8
MKT	1	45
PHY	2	774
STA	1	436

table 02: Course Category Distribution

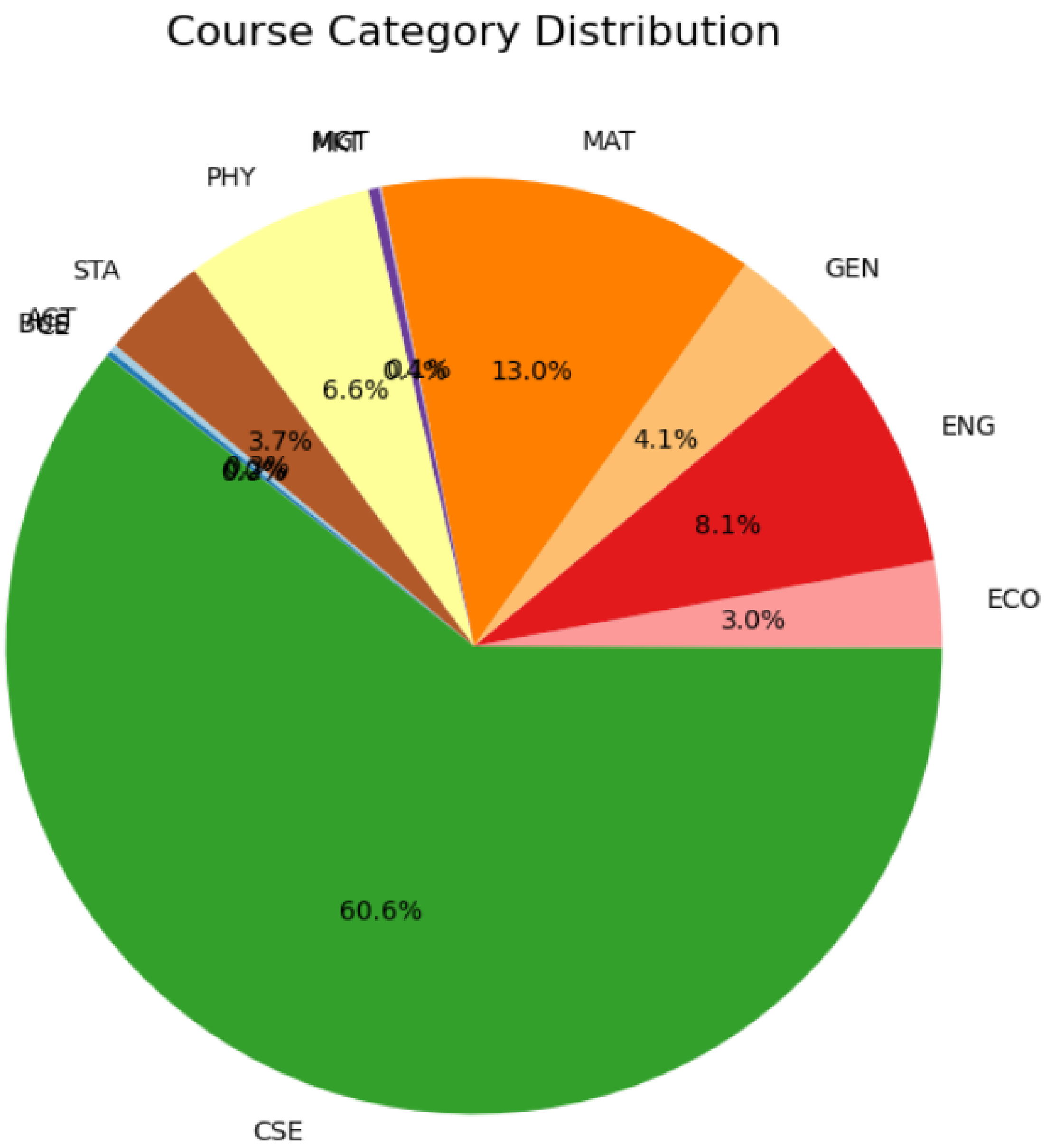


Fig 01: Pie Chart of Course Category Distribution

Courses Taken By Students:

Course	Occ	Course	Occ	Course	Occ	Course	Occ	Course	Occ	Course	Occ
ACT 101	31	**CSE 251**	226	**CSE 407**	104	**CSE 487**	149	**GEN 202**	1	**MAT 110**	1
BUS 101	4	**CSE 302**	279	**CSE 412**	100	**CSE 488**	33	**GEN 203**	98	**MAT 205**	355
CE 200	1	**CSE 303**	235	**CSE 420**	35	**CSE 489**	57	**GEN 205**	1	**MGT 337**	8
CHE 109	423	**CSE 325**	254	**CSE 430**	43	**CSE 495**	113	**GEN 209**	2	**MKT 101**	45
CSE 103	800	**CSE 345**	258	**CSE 438**	101	**ECO 101**	350	**GEN 210**	52	**PHY 109**	422
CSE 106	913	**CSE 347**	203	**CSE 453**	3	**ECO 102**	2	**GEN 214**	77	**PHY 209**	352
CSE 110	538	**CSE 350**	69	**CSE 464**	36	**ENG 099**	35	**GEN 226**	255	**STA 102**	436
CSE 200	395	**CSE 360**	199	**CSE 475**	94	**ENG 101**	709	**MAT 101**	532	**CSE 209**	422
CSE 207	398	**CSE 400**	211	**CSE 477**	52	**ENG 102**	211	**MAT 102**	274	**BUS 231**	20
CSE 246	357	**CSE 405**	169	**CSE 479**	99	**FIN 101**	133	**MAT 104**	373	**CSE 366**	199

table 03: Course Taken by Students

From this table we can clearly see that most popular course is **CSE106 with 913 occurrences**
And least popular courses are
CE200- with 1 occurrences
GEN202 with 1 occurrences
GEN205 with 1 occurrences
MAT110 with 1 occurrences

Most Popular Courses & Least Popular Course Visualization

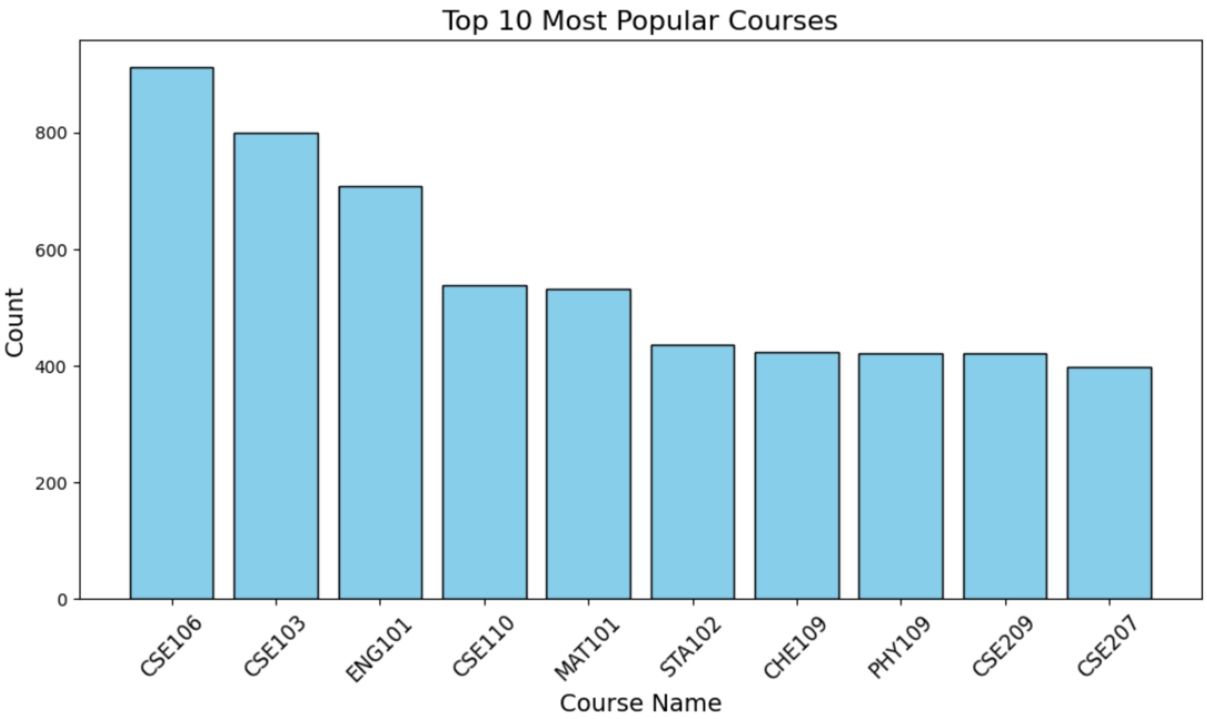


Fig 02: Most Popular Courses

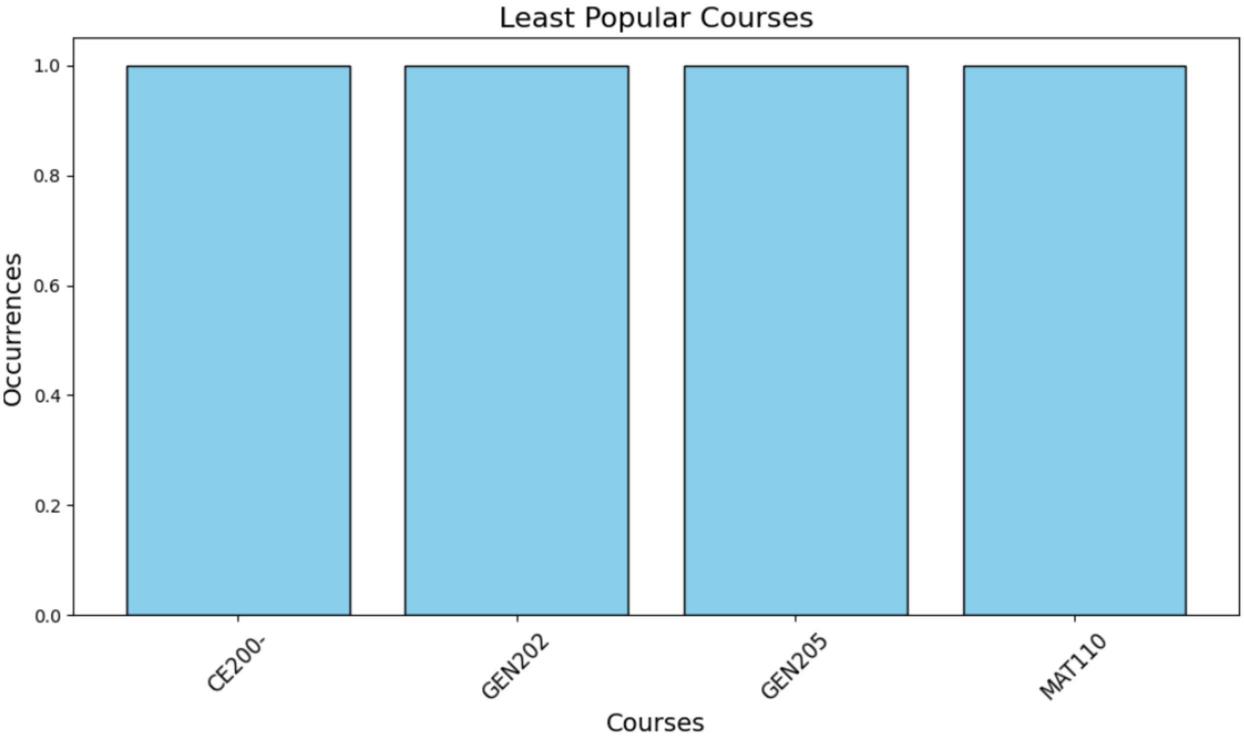


Fig 03: Least Popular Courses

Maximum, Minimum and Average Course Taken By a Student

Maximum	6
Average	4
Minimum	1

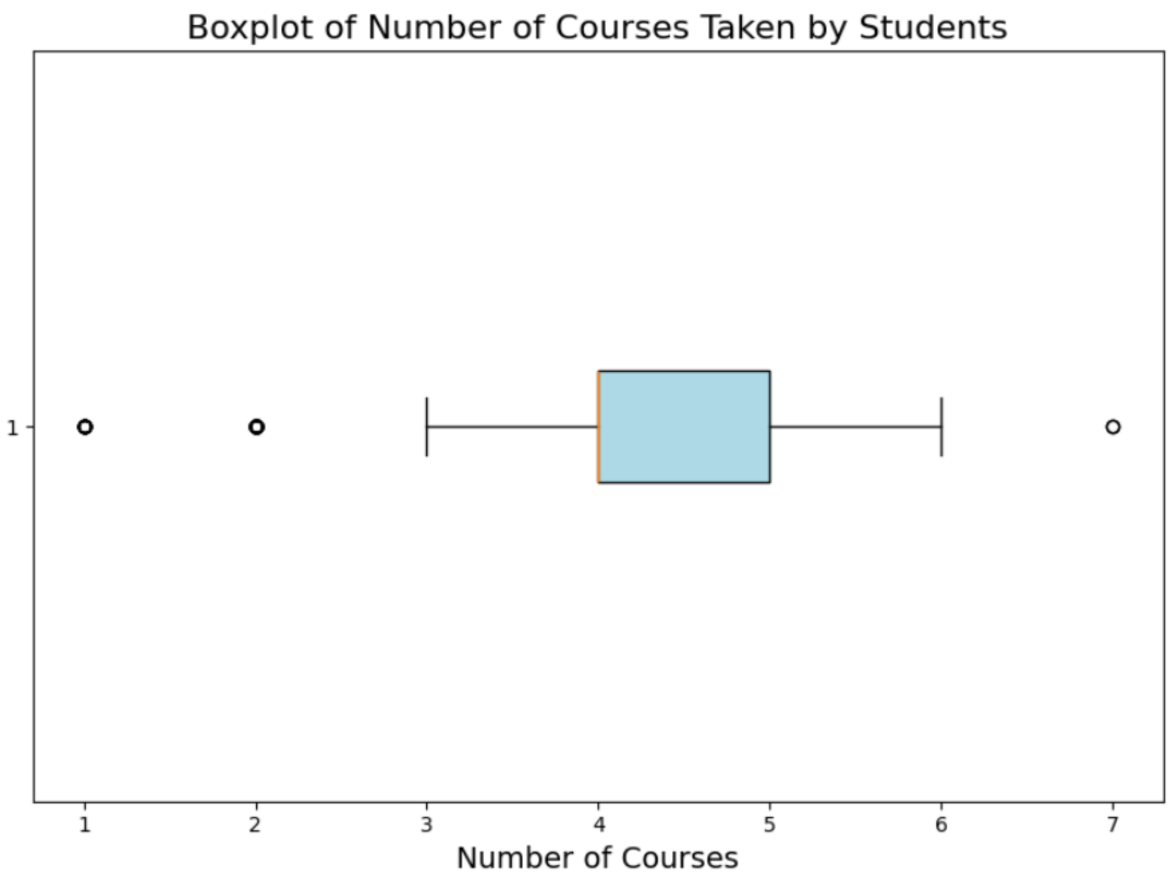


Fig 04: Boxplot of Number of Courses Taken by Students

Distribution of Courses among Student

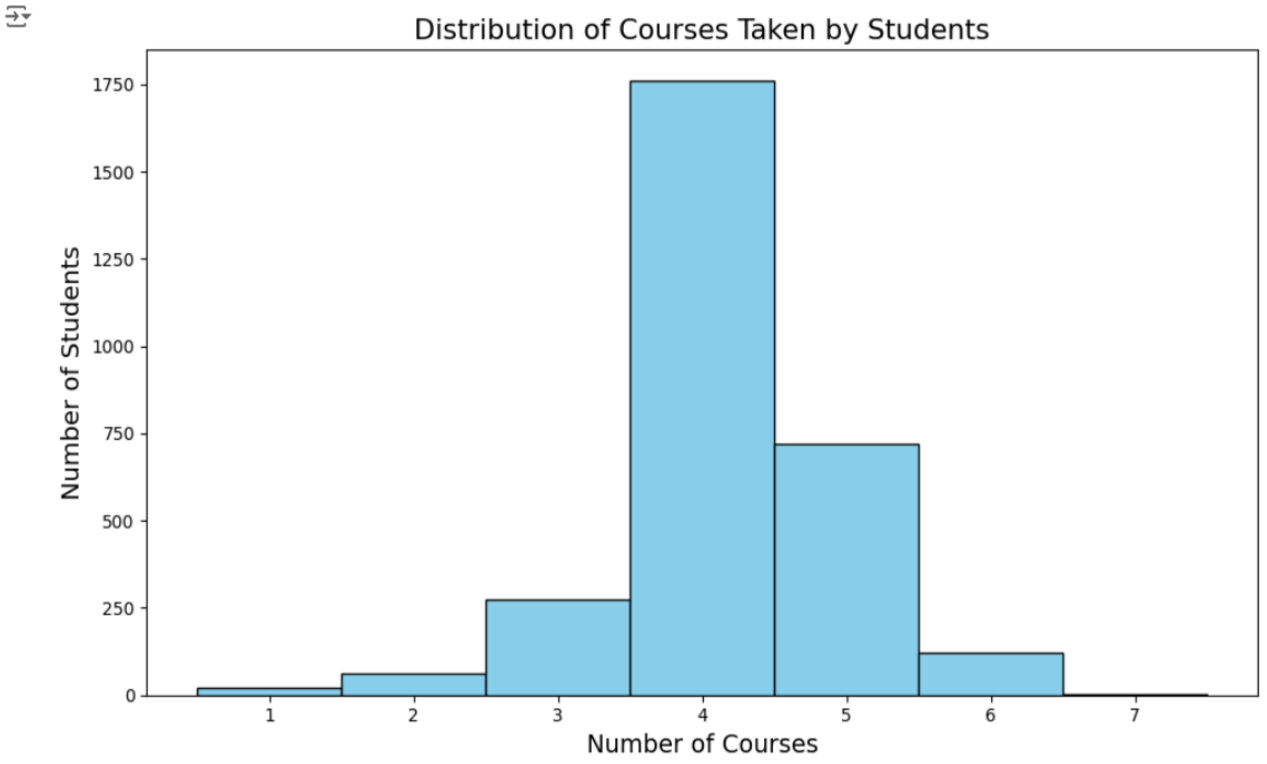


Fig 05: Number of Courses Taken by Students

This histogram shows the distribution of the number of courses taken by students. The majority of students take **4 courses**, as indicated by the highest bar, followed by a smaller group taking **5 courses**. Few students take **1, 2, 3 or 6 courses**, suggesting that 4 courses is the most common academic load for this dataset.

Most frequent combinations of courses taken together (pair)

Pairs	Occurrence
CSE 103, CSE 106	756
CSE 103, ENG 101	696
CSE 106, ENG 101	693
CSE 106, MAT 101	509
CSE 103, MAT 101	508
ENG 101, MAT 101	460
CHE 109, CSE 106	307
CHE 109, CSE 103	283
CSE 110, STA 102	243
CHE 109, ENG 101	238

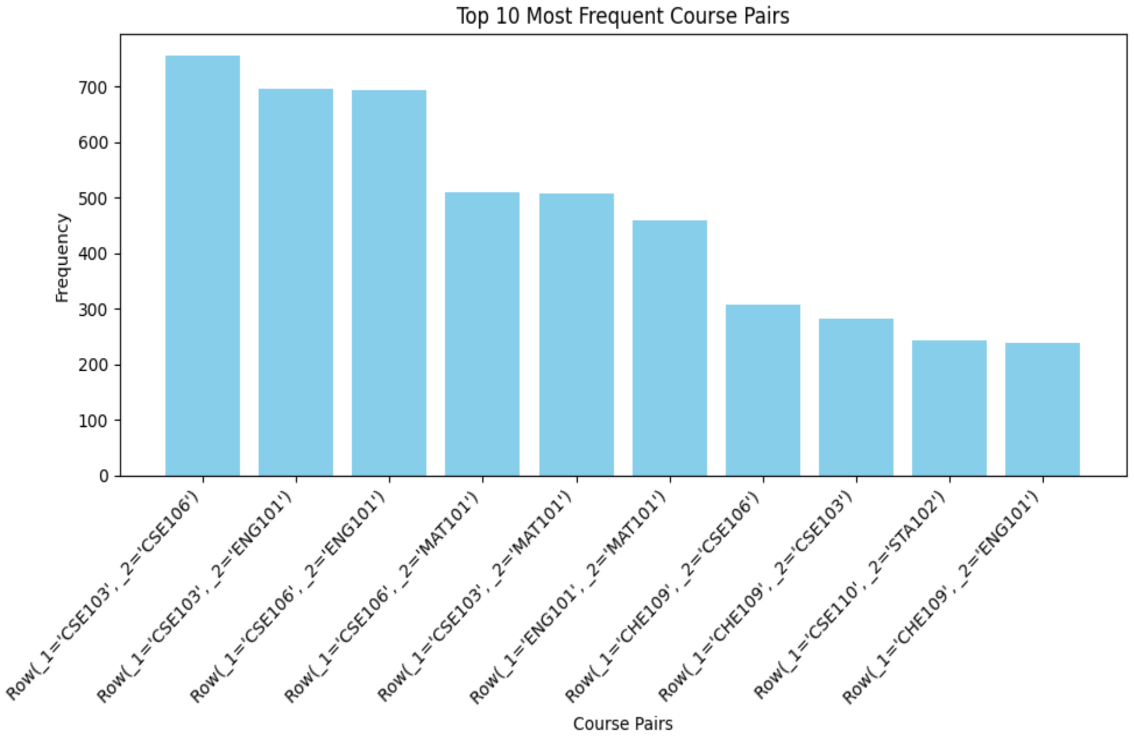


Fig 06: bar chart of Most Frequent Pair

This table shows the most frequent course pairs from a dataset, showing the courses often taken together and their respective occurrence counts. This data helps identify strong relationships between courses, possibly indicating shared curriculums or student preferences. The top pair is 'CSE103' and 'CSE106' with 756 occurrences.

Most frequent combinations of courses taken together (Triples)

Triples	**Occurrence**
CSE103', 'CSE106', 'ENG101'	691
'CSE103', 'CSE106', 'MAT101'	501
'CSE103', 'ENG101', 'MAT101'	459
'CSE106', 'ENG101', 'MAT101'	457
'CHE109', 'CSE103', 'CSE106'	269
'CHE109', 'CSE103', 'ENG101'	234
'CHE109', 'CSE106', 'ENG101'	230
'CSE110', 'PHY109', 'STA102'	93
'CSE207', 'CSE209', 'MAT205'	88
'CSE110', 'MAT104', 'STA102'	86

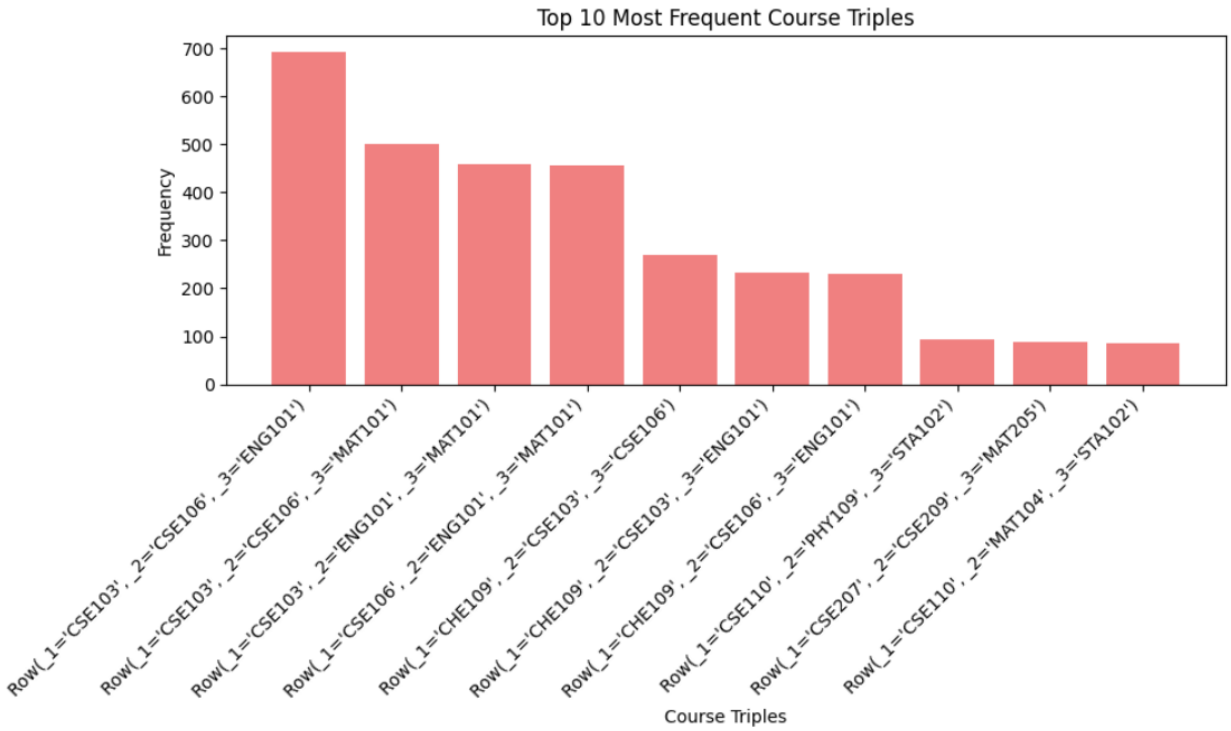


Fig 07: bar chart of Most Frequent Triple

The data lists the most frequent course triples, showing groups of three courses often taken together along with their occurrence counts. These triples reveal common course combinations that likely represent shared pathways or popular curricula.

- The most frequent triple is ('CSE103', 'CSE106', 'ENG101') with 691 occurrences,

Most frequent combinations of courses taken together (Quadruples)

Quadruples	**Occurrence**
'CSE103', 'CSE106', 'ENG101', 'MAT101'	457
'CHE109', 'CSE103', 'CSE106', 'ENG101'	230
'CSE207', 'CSE209', 'ECO101', 'MAT205'	36

'CHE109', 'CSE103', 'CSE106', 'MAT101'	35
'CSE110', 'MAT102', 'PHY109', 'STA102'	34
'CSE103', 'CSE106', 'ENG099', 'MAT101'	31
'CSE110', 'ENG102', 'MAT102', 'PHY109'	27
'CSE110', 'ECO101', 'MAT104', 'STA102'	25
'CSE110', 'MAT104', 'PHY109', 'STA102'	25
'CSE207', 'CSE209', 'GEN226', 'MAT205')	24

The data lists the most frequent course quadruples, showing groups of four courses frequently taken together along with their occurrence counts. These quadruples highlight common course combinations, potentially indicating structured curricula or popular course sets.

- The most common quadruple is ('CSE103', 'CSE106', 'ENG101', 'MAT101') with 457 occurrences.

Statistical Measures:

Mean, Median, and Mode

Statistic	Completed Credits	Taken Credit	Taken No Course
Mean	42.56	14.47	4.17
Median	33.50	14.50	4.00
mode	0.00	14.50	4.00

Correlation between Completed Credits and Taken Credit: 0.005571148471276273

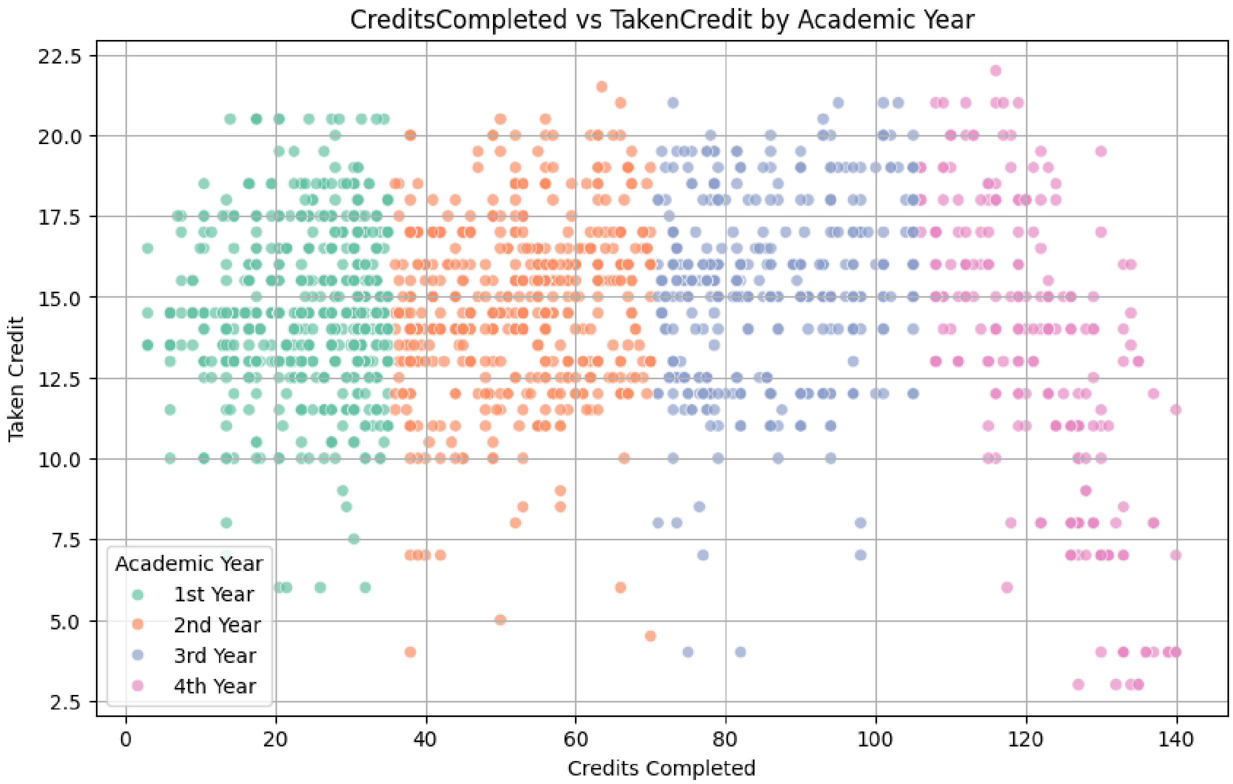


Fig 08: Correlation between completed credit and taken credit

This scatter plot shows the relationship between total `Credits Completed` and `taken credit`. Most students consistently take 10–15 credits per semester, regardless of how many credits they've completed overall. The data reveals no strong correlation between these variables, as points are widely scattered. A dense cluster is visible for students with 20–80 completed credits taking 10–15 credits. Outliers include students with very low or very high completed credits taking unusually low or high current credit loads, possibly indicating exceptional cases like graduation. Also we see at first years students have more course load than the last. Load decrease while completing credit increase.

2.Visualization Tasks:

Single Course popularity

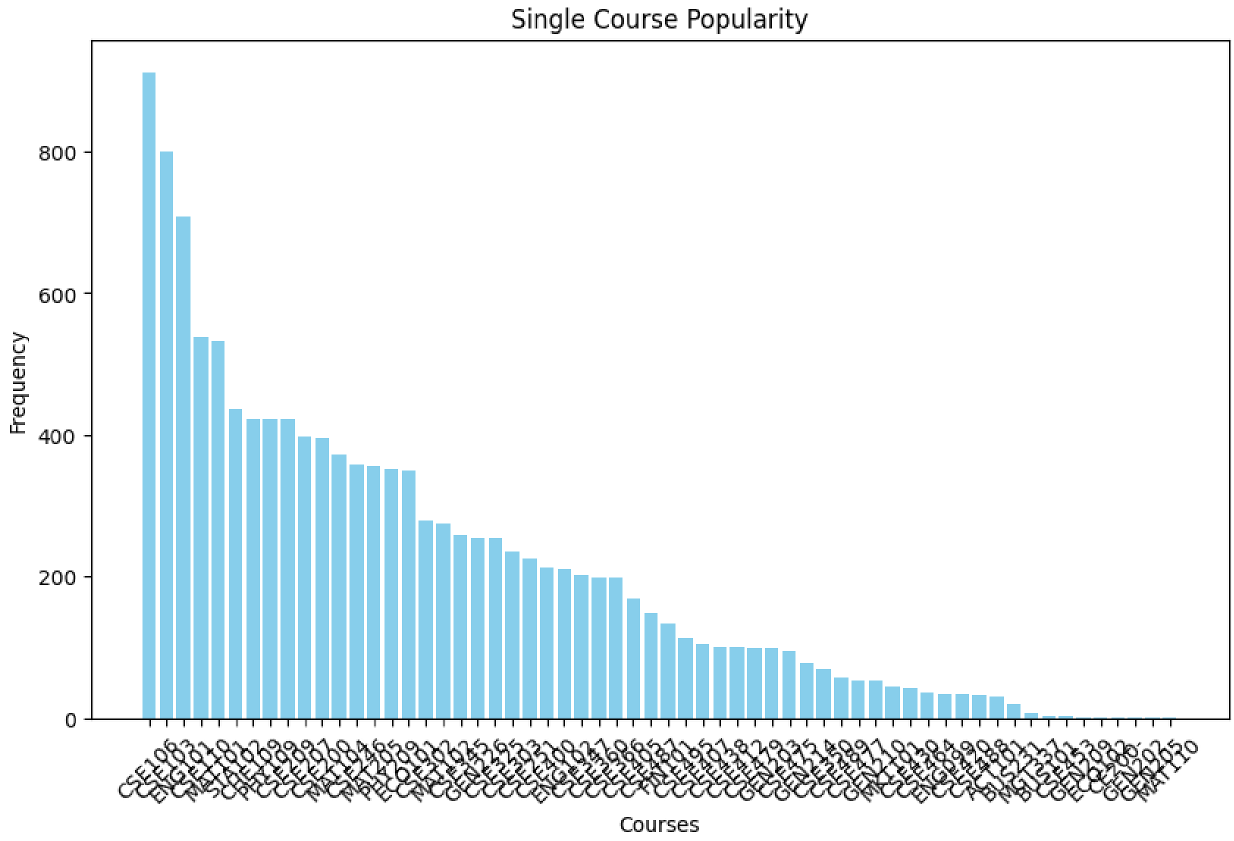


Fig 09: bar chart of Single Course Popularity

Bar chart to show the frequency of Single Course Popularity. And we see how famous CSE 106 course in between student.

Combination Patterns

- Heatmap of visualize the relationships between courses frequently taken together.

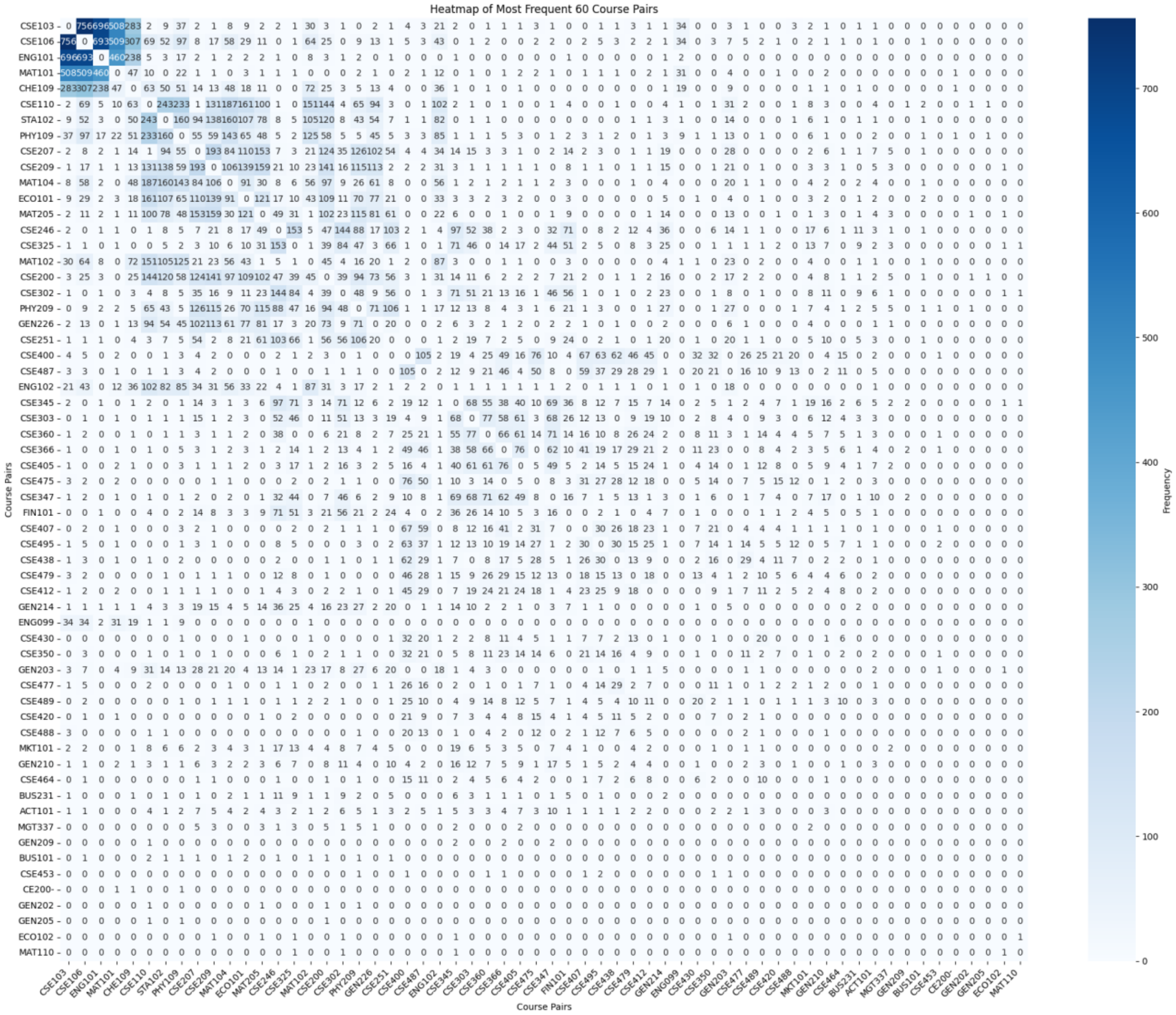
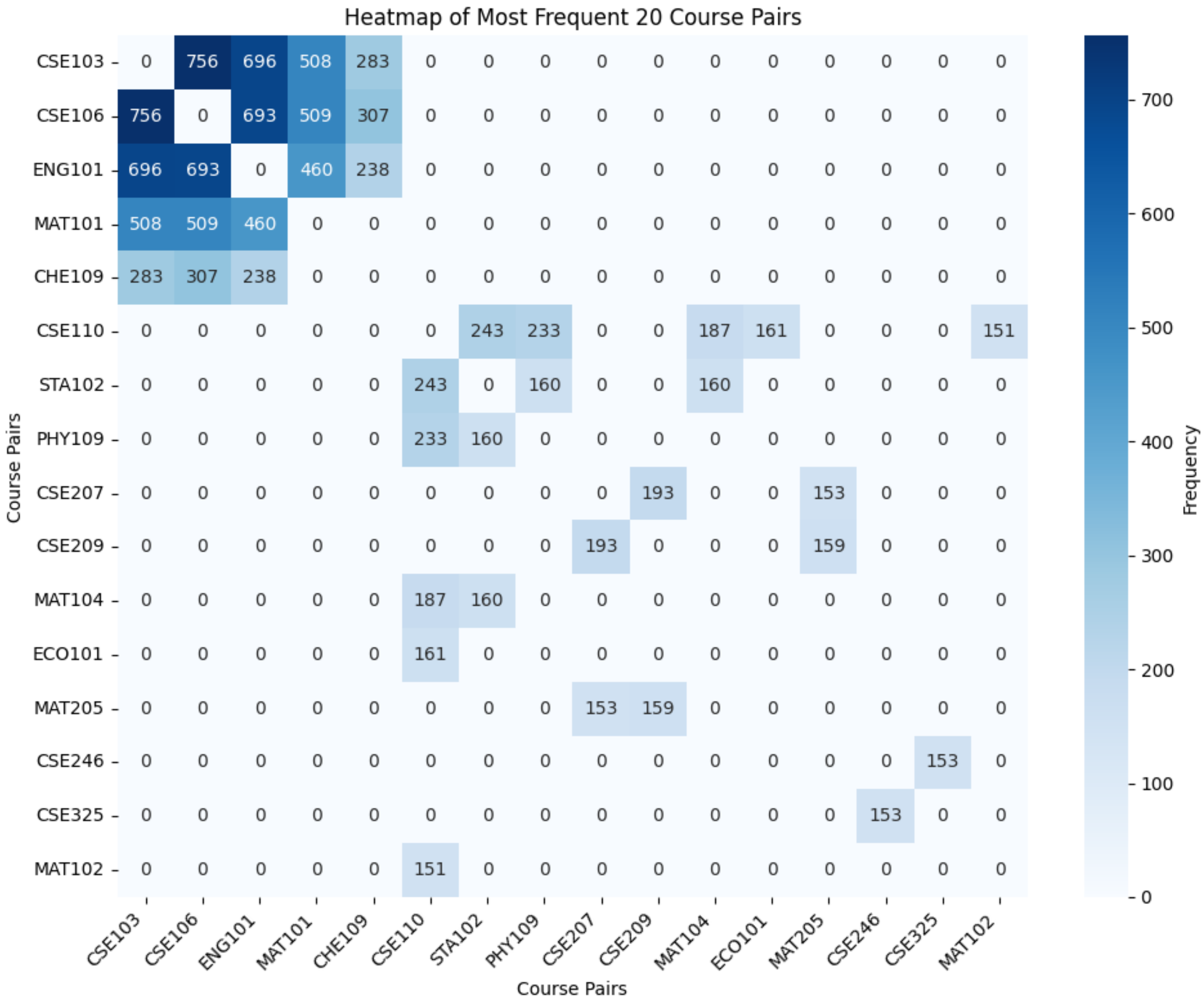


Fig 10: Heat map for all course pairs



	CSE103	CSE106	ENG101	MAT101	CHE109	CSE110	STA102	PHY109	CSE207	CSE209	MAT104	ECO101	MAT205	CSE246	CSE325	MAT102
CSE103	0	756	696	508	283	0	0	0	0	0	0	0	0	0	0	0
CSE106	756	0	693	509	307	0	0	0	0	0	0	0	0	0	0	0
ENG101	696	693	0	460	238	0	0	0	0	0	0	0	0	0	0	0
MAT101	508	509	460	0	0	0	0	0	0	0	0	0	0	0	0	0
CHE109	283	307	238	0	0	0	0	0	0	0	0	0	0	0	0	0
CSE110	0	0	0	0	0	0	243	233	0	0	187	161	0	0	0	151
STA102	0	0	0	0	0	243	0	160	0	0	160	0	0	0	0	0
PHY109	0	0	0	0	0	233	160	0	0	0	0	0	0	0	0	0
CSE207	0	0	0	0	0	0	0	0	0	193	0	0	153	0	0	0
CSE209	0	0	0	0	0	0	0	0	193	0	0	0	159	0	0	0
MAT104	0	0	0	0	0	187	160	0	0	0	0	0	0	0	0	0
ECO101	0	0	0	0	0	161	0	0	0	0	0	0	0	0	0	0
MAT205	0	0	0	0	0	0	0	0	153	159	0	0	0	0	0	0
CSE246	0	0	0	0	0	0	0	0	0	0	0	0	0	0	153	0
CSE325	0	0	0	0	0	0	0	0	0	0	0	0	0	153	0	0
MAT102	0	0	0	0	0	151	0	0	0	0	0	0	0	0	0	0

Fig 11: Heat map of top 20 course pairs

Here each row and column represents a unique course (CSE110, ENG101 etc.). The intersection of a row and column indicates the frequency with which two courses are taken together. Courses with high co-occurrence (darker colors) are likely core or mandatory courses taken together by many students on the other hand lighter areas indicate electives or less frequently taken course combinations.

Credits Distribution:

- Histogram for the distribution of Credits Completed.

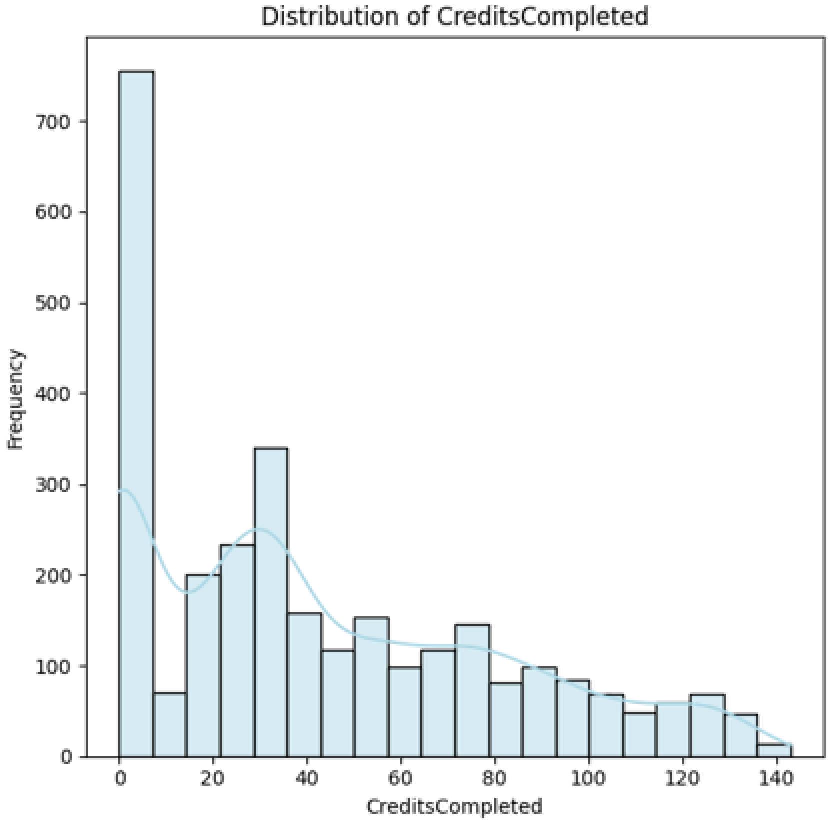


Fig 12: Histogram for the distribution of Credits Completed.

The histogram shows the distribution of credits completed by students. The x-axis represents the number of credits, and the y-axis indicates how many students fall into each credit range. Most students have completed fewer than 20 credits, as shown by the tallest bar at the start. A smooth density curve is overlaid to highlight the general trend, which reveals a right-skewed distribution. This means the number of students decreases steadily as the completed credits increase.

- Histogram for the distribution of taken Credit.

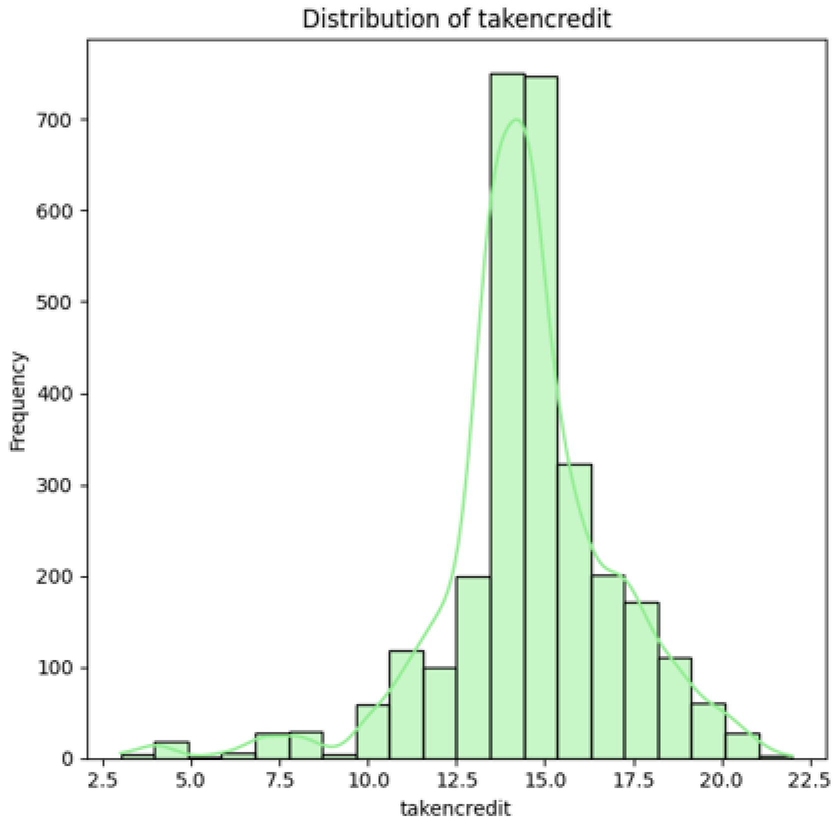


Fig 13: Histogram for the distribution of taken Credit

The histogram displays the distribution of **takencredit**, which represents the number of credits students take. The x-axis shows the credit values, and the y-axis indicates the number of students in each range. Most students take between 14 and 15 credits, as shown by the highest bars in this range. The smooth density curve overlaid on the histogram highlights a symmetrical, bell-shaped distribution, suggesting that the majority of students take a similar number of credits, with fewer students taking significantly more or fewer credits.

Course Co-occurrence

Network graph to display courses that are frequently taken together.

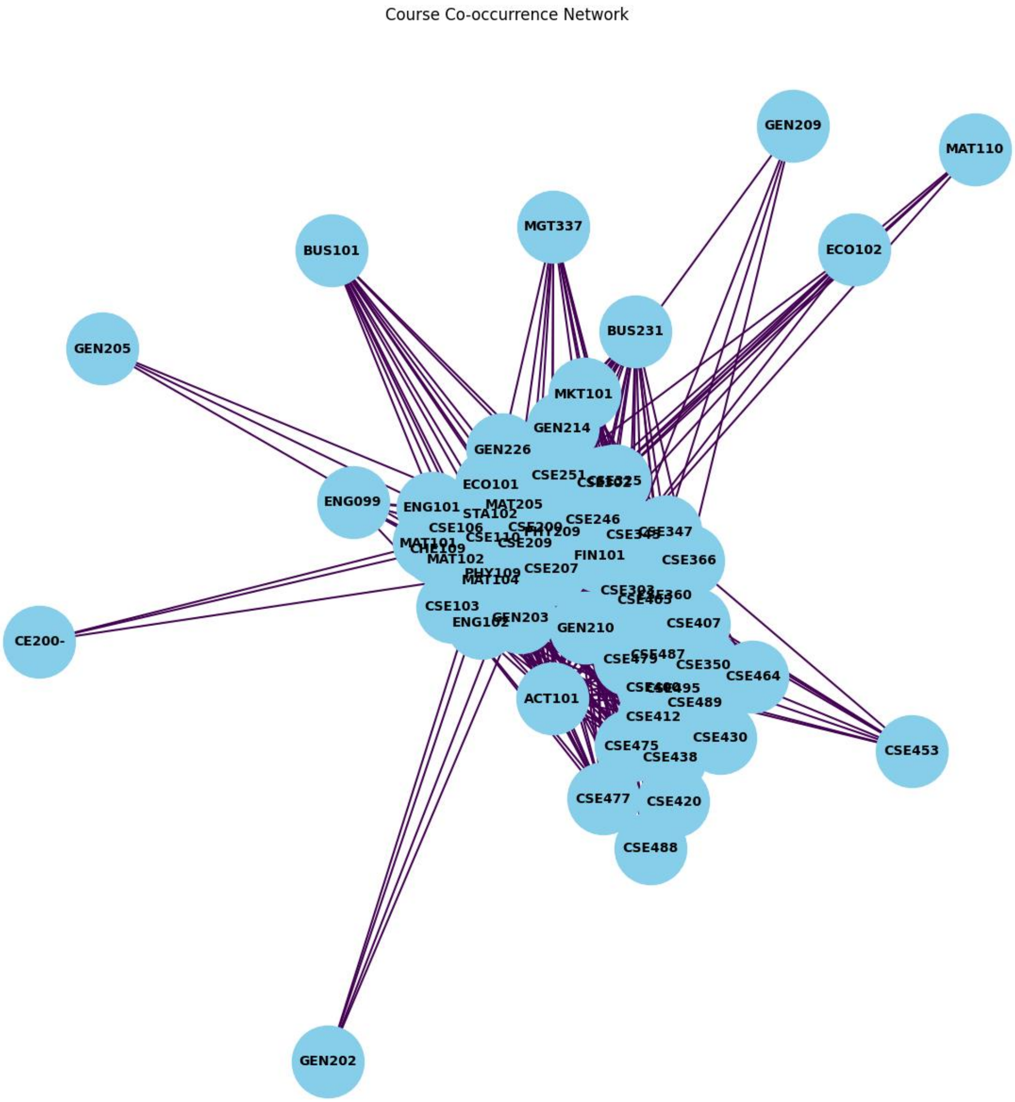


Fig 14: Course Co-Occurrence Network

This diagram shows a **course co-occurrence network**, where each blue circle represents a course labeled with its code (e.g., MAT110, CSE103). The size of the circles indicates how frequently a course is taken—larger circles mean higher enrollment or importance. The purple lines connecting the circles represent courses often taken together, with thicker lines showing stronger relationships.

In the center, there’s a dense cluster of courses that are closely connected, meaning students frequently enroll in them together. In contrast, some courses, like CE200- and GEN202, are more isolated, suggesting they are less commonly taken with others.

This network highlights patterns in course enrollment, making it useful for planning, identifying relationships between courses, and optimizing recommendations for students.

- For top 30 most common pairs Co-occurrence network

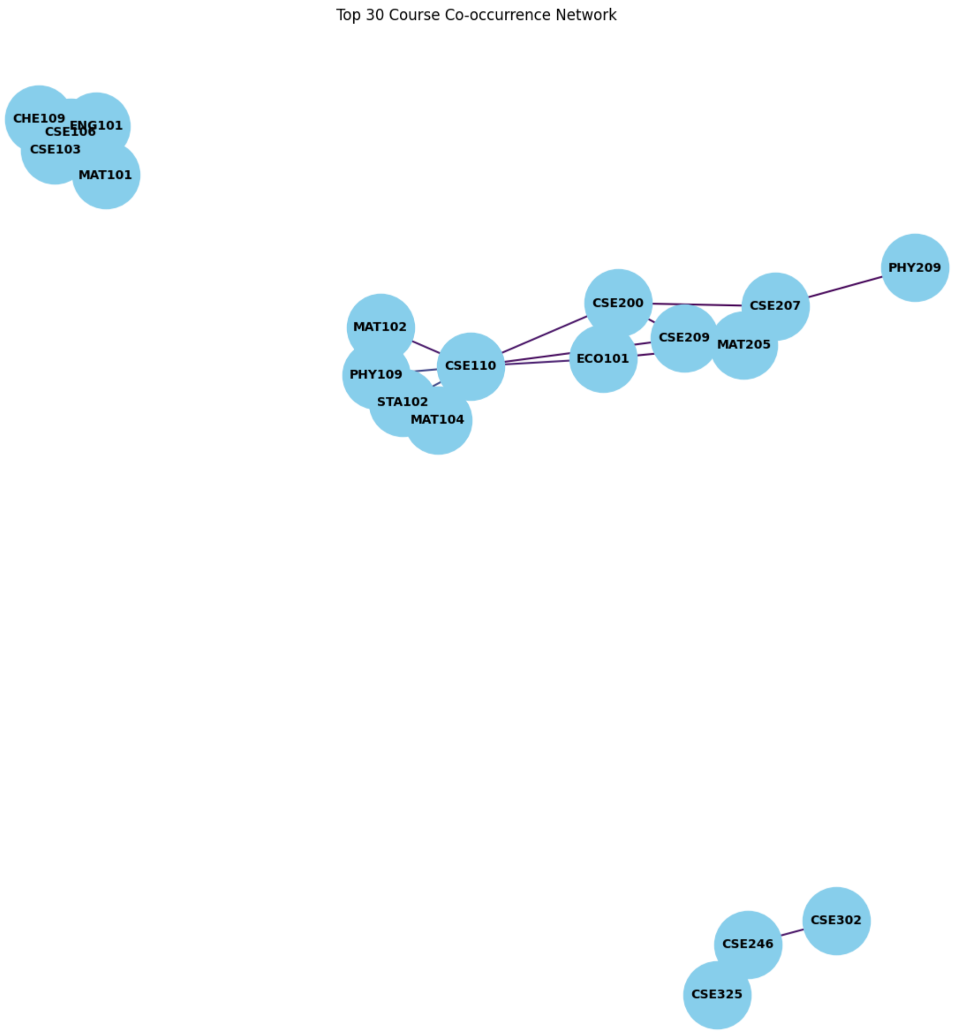


Fig 15: Top 30 Most Common Course Co-Occurrence Network

This diagram represents a **Top 30 Course Co-occurrence Network**, where each blue circle is a course, and the purple lines show how often courses are taken together. Thicker lines mean stronger relationships.

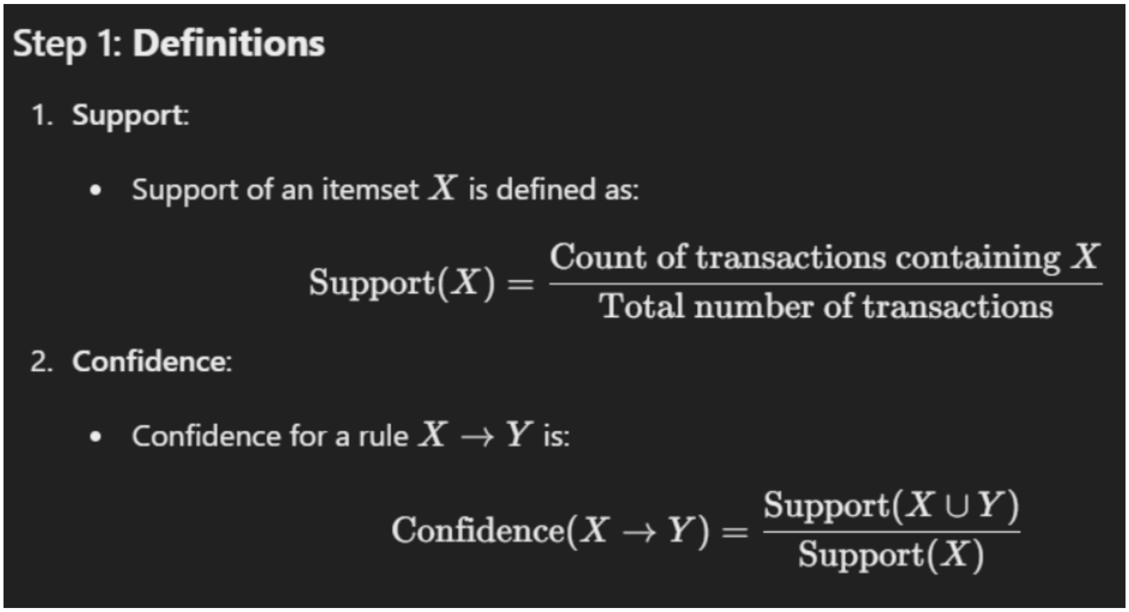
The top-left cluster includes foundational courses like CHE109, ENG101, and CSE103, while the central cluster connects core courses such as MAT102, PHY109, and CSE110. In the bottom-right, advanced courses like CSE246, CSE302, and CSE325 form a separate group. PHY209 stands out as loosely connected, showing limited co-enrollment.

3. Predictive Analytics

Association Rule Mining:

Here to identify relationships between courses frequently taken together, we applied the **Apriori algorithm** to generate **association rules**. The algorithm analyzed course combinations (pairs, triples, etc.) to compute:

- **Support**: The proportion of students who enrolled in a specific course combination.
- **Confidence**: The likelihood of taking one course, given that another course is already taken.



Step 1: Definitions

1. **Support**:
   - Support of an itemset X is defined as:

$$\text{Support}(X) = \frac{\text{Count of transactions containing } X}{\text{Total number of transactions}}$$

2. **Confidence**:
   - Confidence for a rule $X \to Y$ is:

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

By this we find out support and confidence. Explain this by taking one output from each is given below.

Support Explanation:

- **Example: ('CHE109', 'CSE103', 'CSE106', 'ENG101'): 0.08**
  - **Meaning**: 8% of all transactions (students) include this exact combination of courses (CHE109, CSE103, CSE106, and ENG101). This is a relatively rare combination in the dataset.

Confidence Explanation:

- **Example: ('CSE407',) -> ('CSE487',): 0.57**
  - **Meaning**: When a student takes CSE407, there is a **57% probability** that they also take CSE487. This shows a moderate association between the two courses.

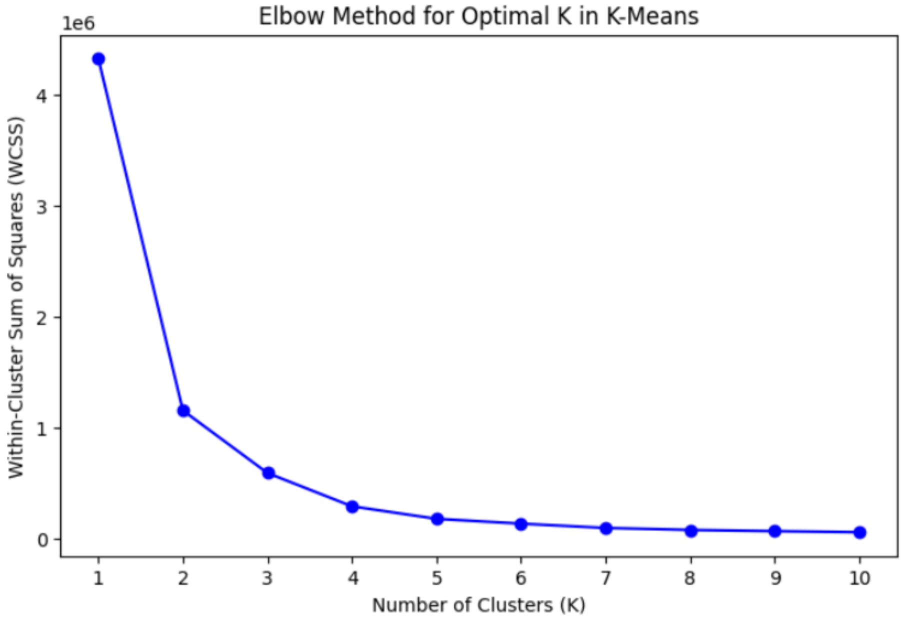
Clustering

Data Preprocessing:

The dataset was transformed into a binary format suitable for association rule mining. All unique courses were extracted from the original columns (C1 to C7), and new binary columns were created for each course. Each column indicates whether a student has taken a specific course (1 for taken, 0 for not taken). The original course columns were dropped, leaving a binary matrix where each row represents a student, and each column represents a course.

K-Means Clustering

K-means clustering is applied to group students based on their course enrollment patterns. Using the binary dataset (where 1 indicates a course taken and 0 indicates not taken), the algorithm divided students into distinct clusters, each representing similar course-taking behaviors. This approach helps identify groups of students with shared academic preferences or pathways, enabling insights into program structure, student interests, and potential course recommendations.



From elbow method, we see the optimal K is 4

Run Cluster Over Courses

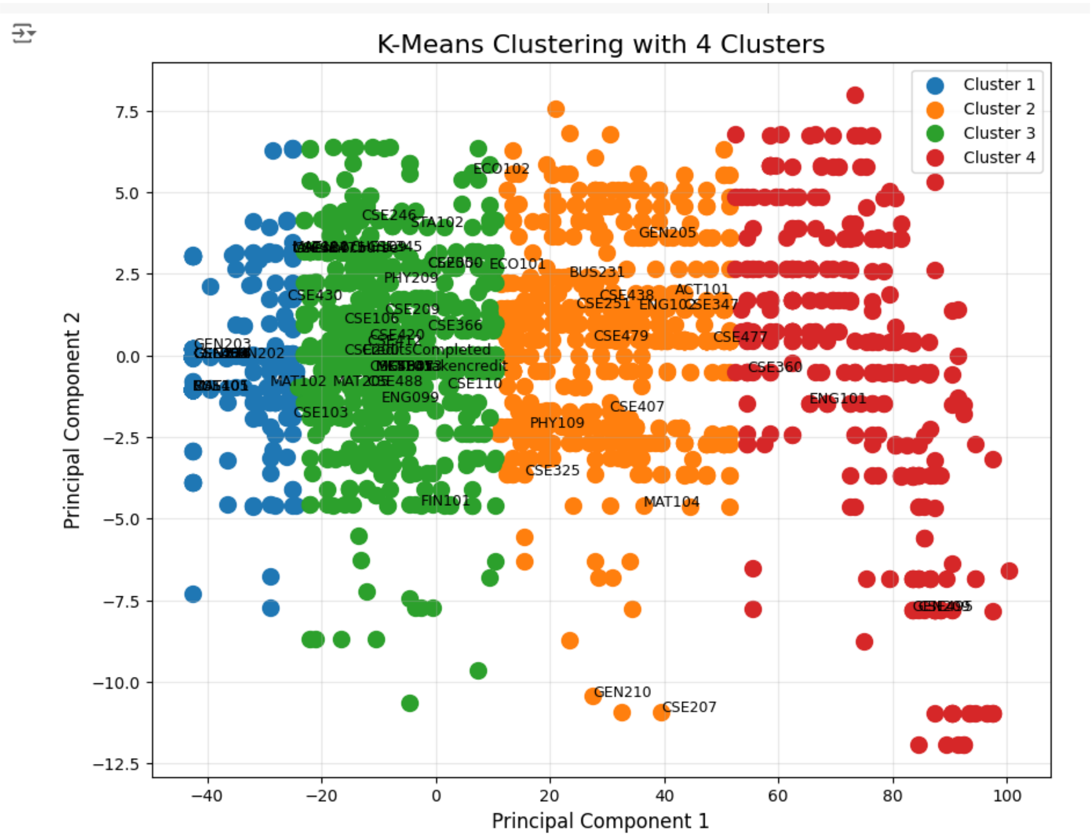
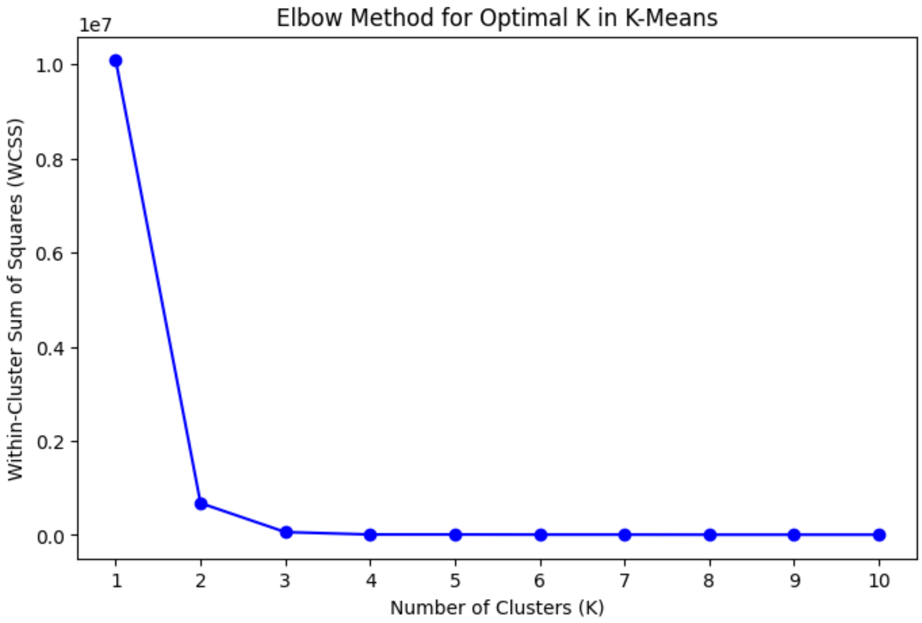


Fig 16: K Means clustering represents students with similar course-taking patterns,

Cluster 1	'takennocourse', 'CSE303', 'CSE103', 'CSE302', 'CSE405', 'CSE487', 'GEN226', 'CSE464', 'CSE430', 'MAT101', 'MAT102', 'BUS101', 'GEN203', 'CSE489', 'GEN214', 'GEN202', 'MAT110'
Cluster 2	'CSE207', 'CSE347', 'CSE251', 'CSE325', 'MAT104', 'ACT101', 'BUS231', 'CSE477', 'ENG102', 'CSE438', 'CSE479', 'PHY109', 'CSE407', 'GEN210', 'GEN205'
Cluster 3	'CreditsCompleted', 'takencredit', 'CSE110', 'CHE109', 'CSE246', 'CSE209', 'CSE200', 'CSE106', 'CSE366', 'CSE412', 'CSE400', 'CSE350', 'CSE475', 'CSE345', 'ECO101', 'CSE420', 'CSE488', 'CE200-', 'MAT205', 'FIN101', 'STA102', 'ECO102', 'ENG099', 'PHY209', 'CSE453', 'MKT101', 'MGT337'
Cluster 4	'CSE360', 'CSE495', 'ENG101', 'GEN209'

Run Cluster Over Students:



From Elbow fig, we see K is 3 here.

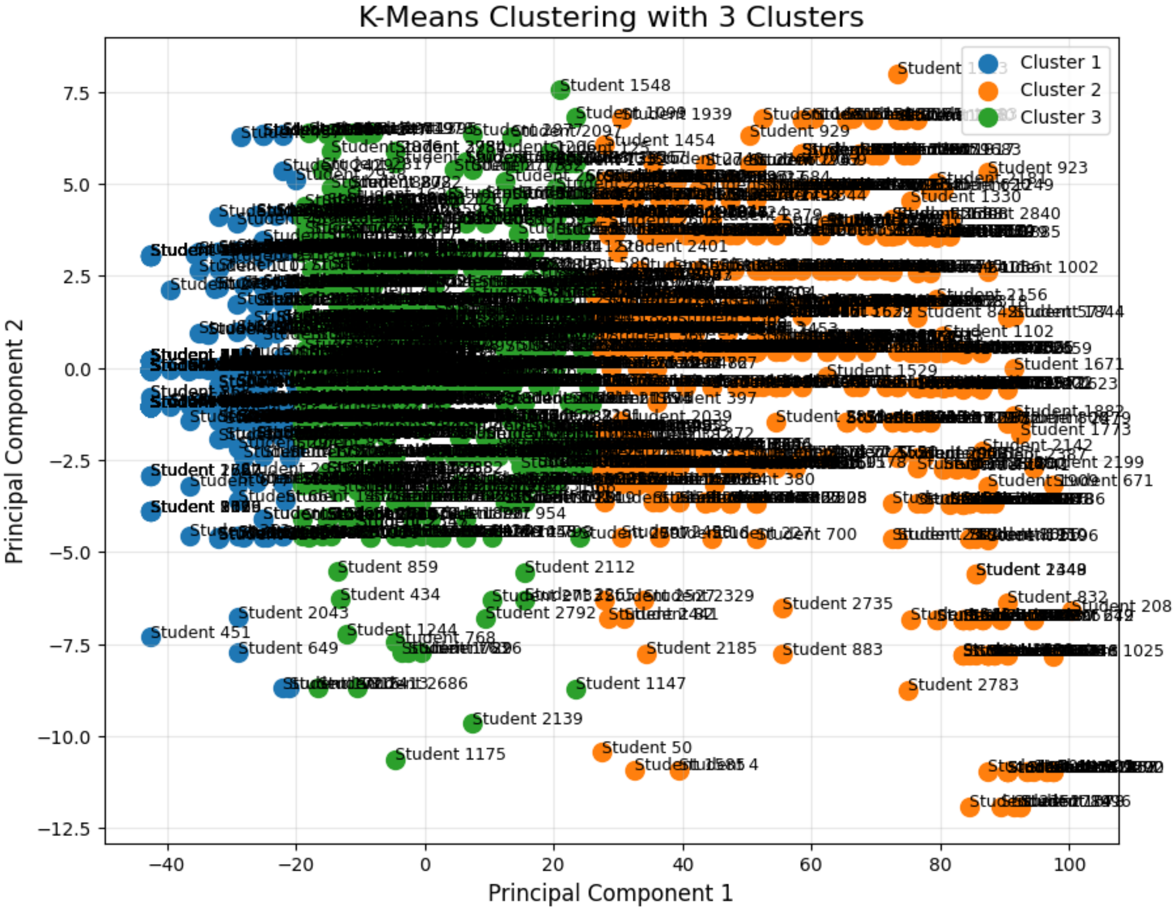


Fig 17: K Means Clustering applied to group students based on their course-taking behavior

4. Advanced Analytics

Recommendation System

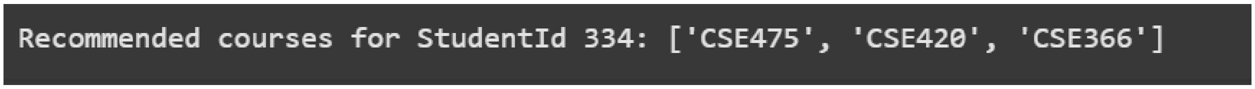
The code processes a dataset containing student course information, where each row represents a student and their enrolled courses. It consolidates the individual course columns (`C1` to `C7`) into a single string for each student in the new `Courses` column, combining all the courses a student has taken into one field. This transformation simplifies the data and makes it easier to analyze students' course histories.

Example :

StudentId	C1	C2	C3	C4	C5	C6	C7	Courses
1	CSE110	ECO101	MAT205	PHY109	NaN	NaN	NaN	CSE110 ECO101 MAT205 PHY109
2	CSE207	CSE209	MAT205	PHY209	NaN	NaN	NaN	CSE207 CSE209 MAT205 PHY209
3	CSE110	ENG102	MAT102	MAT104	PHY109	NaN	NaN	CSE110 ENG102 MAT102 MAT104 PHY109
4	CSE110	GEN203	MAT101	MAT205	NaN	NaN	NaN	CSE110 GEN203 MAT101 MAT205
5	CSE303	NaN	NaN	NaN	NaN	NaN	NaN	CSE303

The code recommends courses to a student by comparing their course choices with those of other students. It calculates the similarity between students based on their courses and suggests courses that similar students have taken, which the target student hasn't yet enrolled in. The result is a list of recommended courses for the student.

Like as:

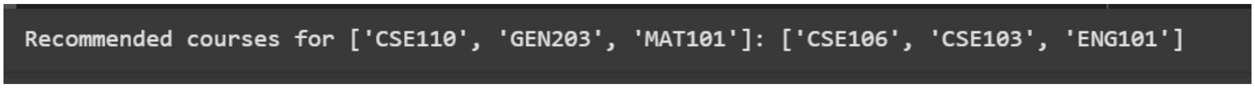


```
Recommended courses for StudentId 334: ['CSE475', 'CSE420', 'CSE366']
```

- Simple Searching by courses

The code is designed to recommend courses to a student based on their previous course selections. It starts by loading a CSV file containing the course data and removing unnecessary columns. The courses each student has taken are then combined into a list, excluding any missing values. The recommendation function takes a list of courses the student has already completed and identifies other students who have taken at least one of those courses. It collects the courses these similar students have taken, excluding the ones the target student has already completed, and calculates the most frequent courses among them. Finally, it suggests the top 3 most recommended courses for the student.

Like as:



```
Recommended courses for ['CSE110', 'GEN203', 'MAT101']: ['CSE106', 'CSE103', 'ENG101']
```

Comparison

The code compares the course-taking habits of students who have completed more than 50 credits with those who have completed 50 or fewer credits. It divides the students into two groups and identifies the 20 most popular courses taken by each group. The findings are presented both in a printed list and visually through bar charts, allowing for an easy comparison of course preferences between the two groups based on their total credits completed.

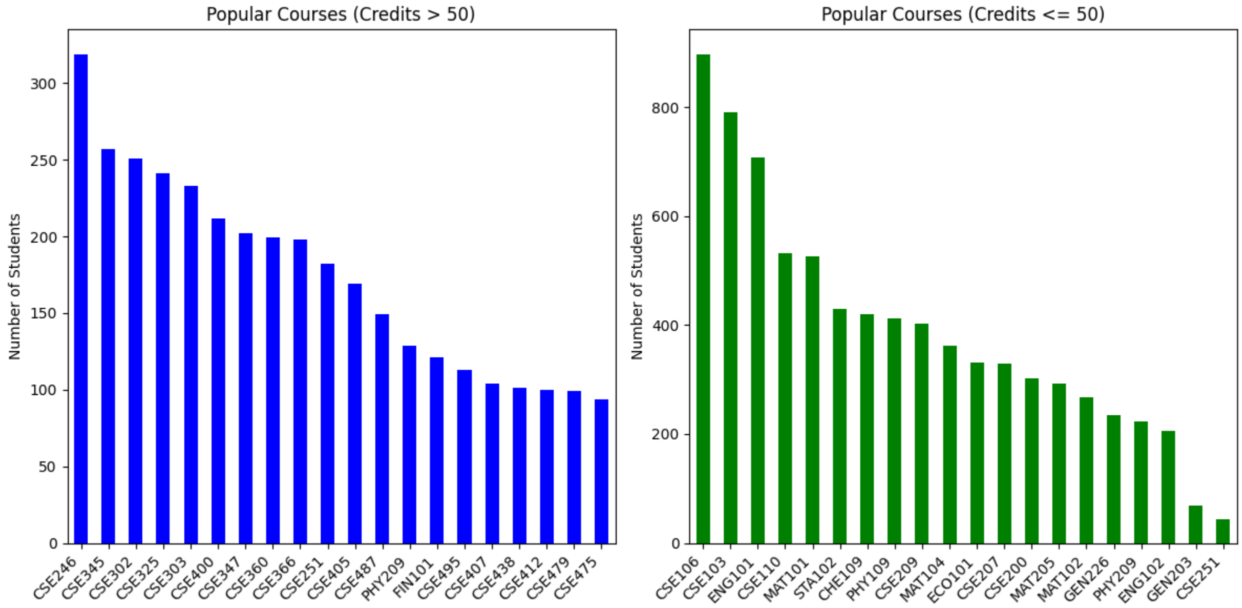


Fig 18: Comparison of Popular University Courses Based on Credit Hours and Enrollment

The **left chart** is titled "Popular Courses (Credits > 50)" and displays data for courses that require higher credit hours. The bars are colored blue. The most popular course in this category is **CSE246**, with around 300 students enrolled. This is followed by courses such as **CSE345**, **CSE209**, and **CSE303**, which each have over 200 students. Enrollment numbers gradually decrease for other courses like **CSE360**, **CSE365**, **CSE495**, and **PHI109**, reaching less than 100 students for **CSE475**, which is the least popular in this group.

The **right chart** is titled "Popular Courses (Credits <= 50)" and features courses with lower credit hour requirements. The bars are colored green. The most popular course in this category is **CSE106**, with approximately 800 students enrolled, followed by **ENG103** and **GEN101**, each with over 600 students. Other popular courses include **MAT101**, **STA102**, **CSE101**, and **PHI109**, with enrollment numbers ranging between 200 and 500. Towards the lower end, courses like **GEN226**, **ENG209**, and **CSE251** have less than 100 students, with **CSE251** being the least popular in this category.

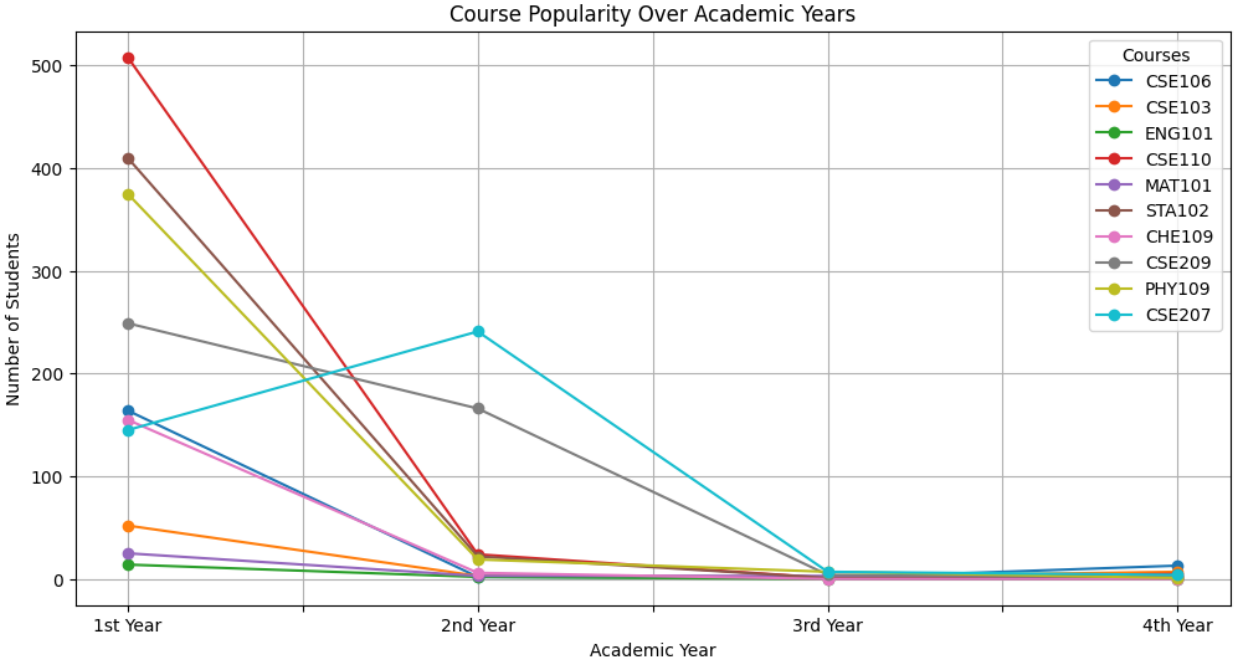


Fig 19: Analysis of Course Enrollment Trends over Academic Years

This code analyzes course enrollment trends over academic years. It separates courses into individual rows, groups them by year to count enrollments, and reshapes the data into a table with years as rows and courses as columns. It identifies the 10 most popular courses and plots a line chart to show how their enrollments change over time. The reshaped table is also printed for inspection.

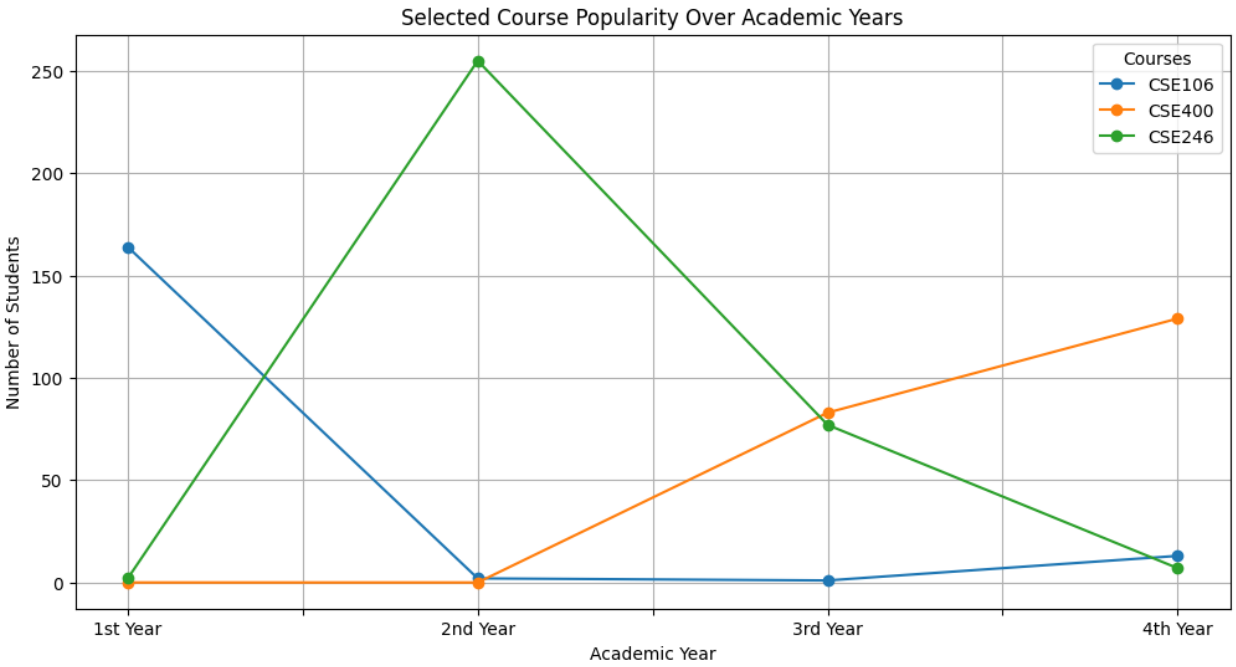


Fig 20: Enrollment Trends of Courses Over Four Academic Years

The chart displays the enrollment trends for three courses—CSE106, CSE400, and CSE246—over four academic years. CSE106 starts with around 150 students in the 1st year but drops to zero in the 2nd year, slowly picking up again in the following years without returning to its original numbers. CSE400 has no enrollments in the first two years, but its popularity rises in the 3rd year and peaks at 80 students in the 4th year. Meanwhile, CSE246 shows no enrollments initially, surges to 250 students in the 2nd year, and then steadily decreases, nearing zero by the 4th year.

5. Creative and Open-Ended Projects

Student Profile

The process involves filtering the dataset using the provided student ID and visualizing the student's academic details. It generates two visualizations: one compares the credits completed by the student with the credits taken, while the other shows the list of courses the student has enrolled in. Additionally, the academic year of the student is displayed.

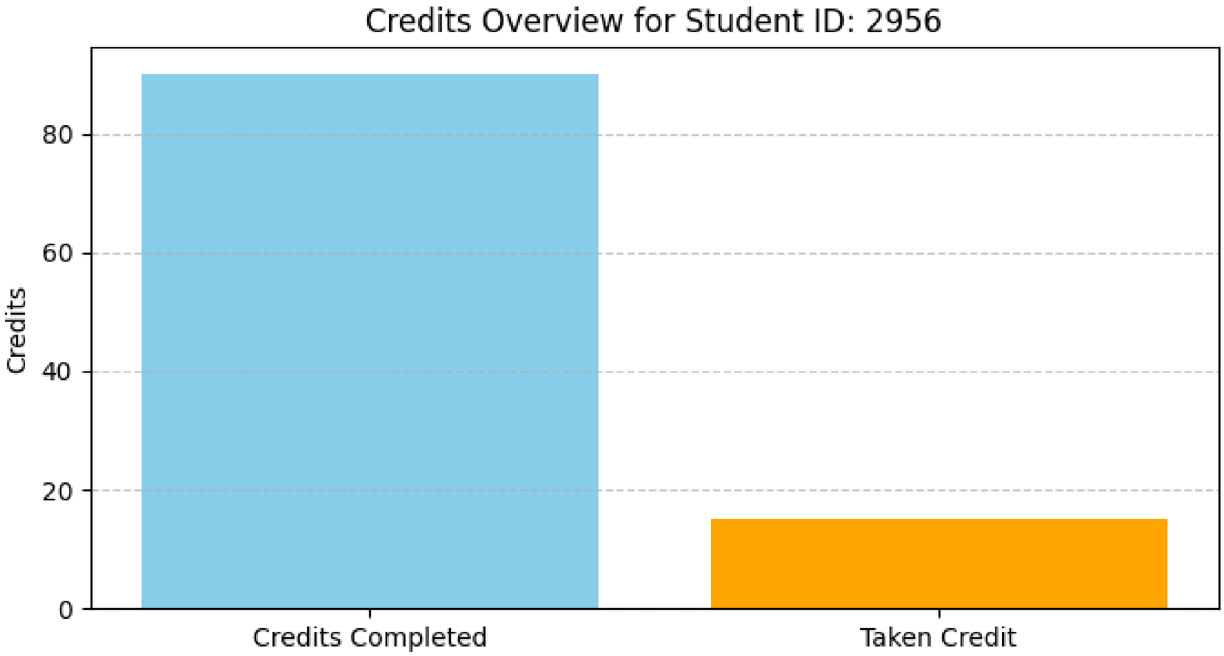


Fig 21: Credits Overview for Student ID : 2956

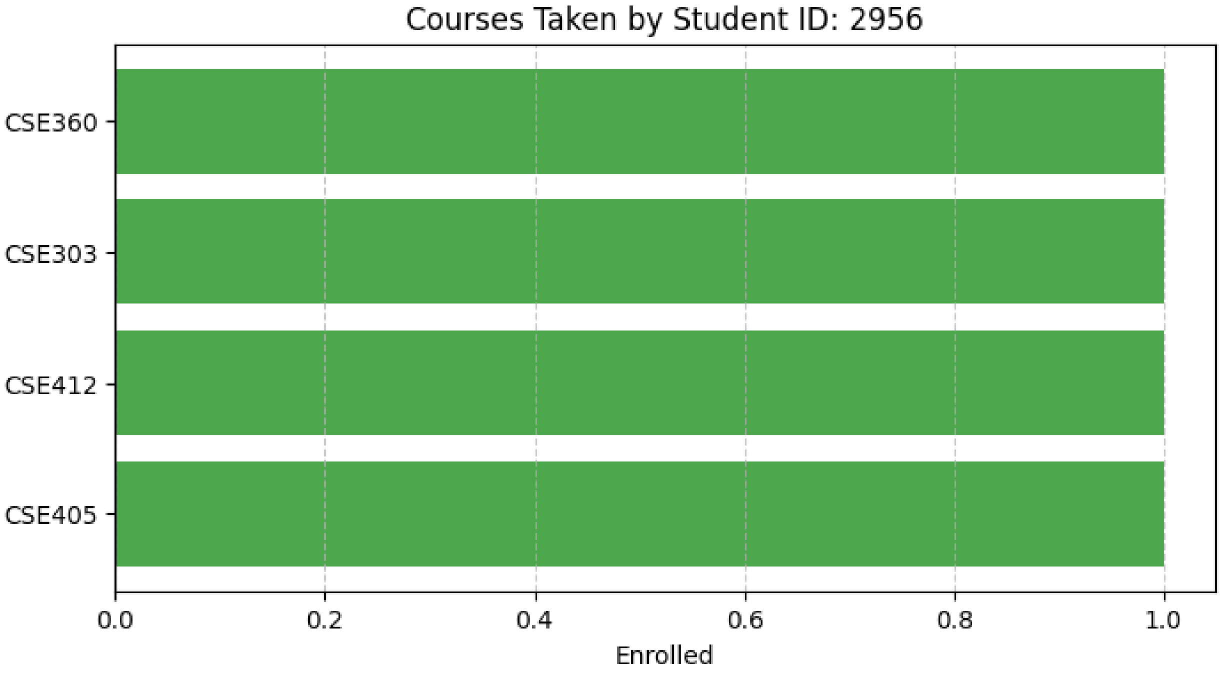


Fig 22: Courses Taken by Student ID : 2956

Some Extra Visualization:

- Number of Students per Academic Year

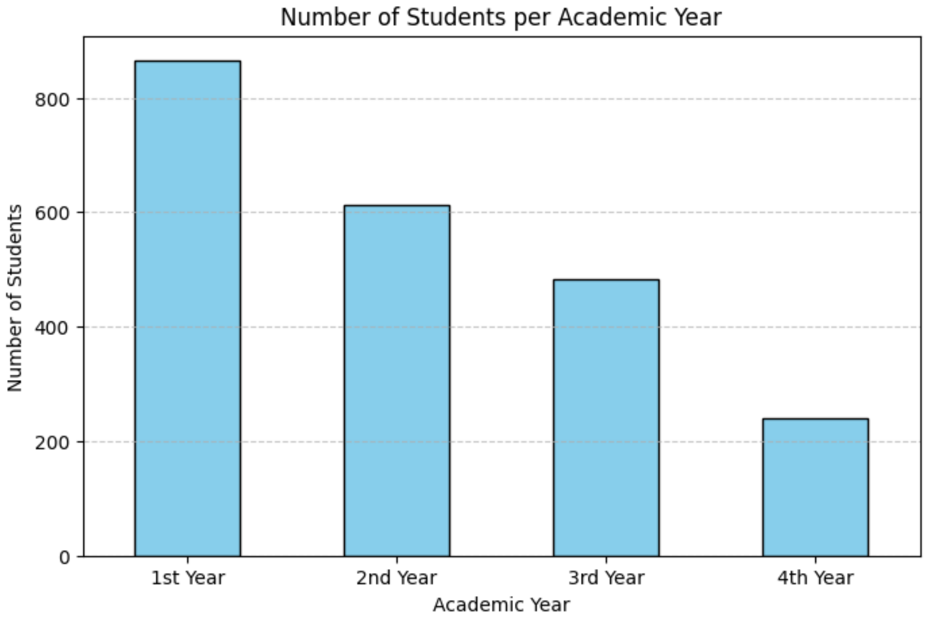


Fig 23: Number of Students in Each Academic Year

The bar graph shows the distribution of students across different academic years. The horizontal axis represents the academic years (1st, 2nd, 3rd, and 4th), while the vertical axis displays the number of students. From the graph, it’s evident that the number of students decreases as the years progress. The 1st Year has the highest number, with over 800 students, followed by the 2nd Year with about 600. The 3rd Year has approximately 400 students, and the 4th Year has the lowest, with around 200. This decline suggests that fewer students remain enrolled as they move to higher academic years, which might be due to dropouts, transfers, or other factors.

- Top 10 Course Combinations

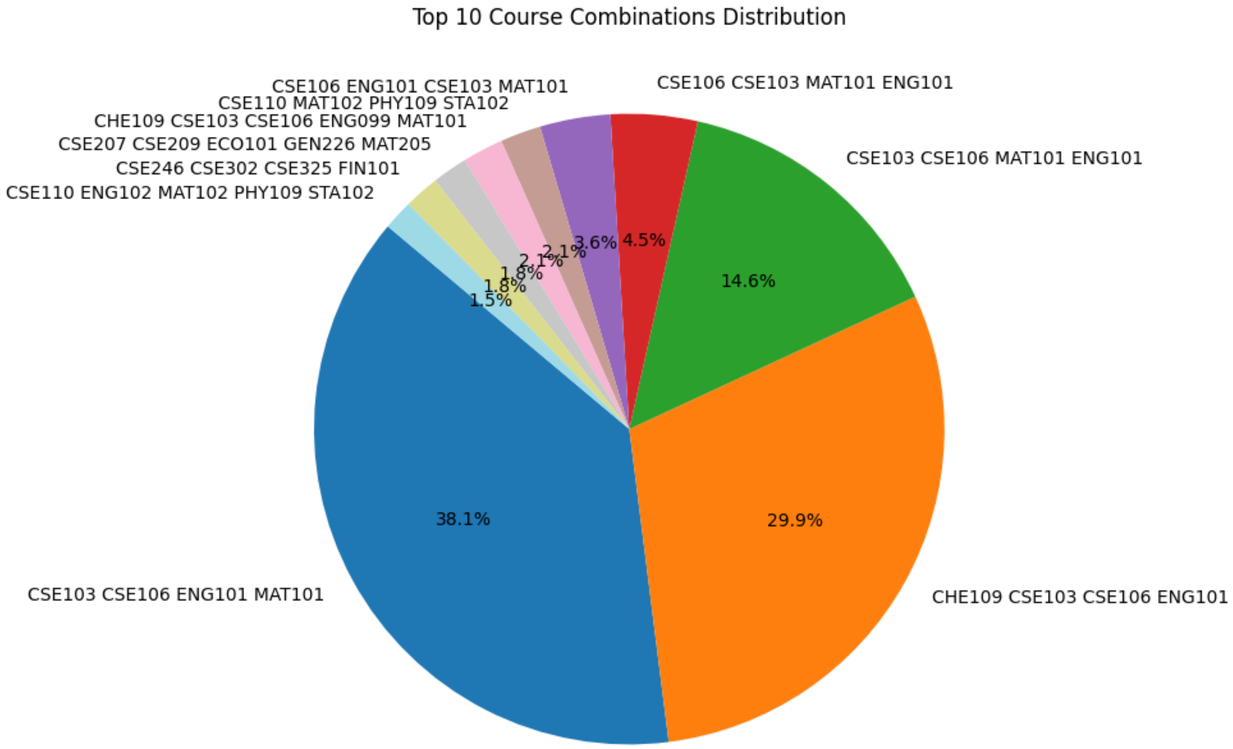


Fig 24: Most Popular Course Combinations

The pie chart shows the top 10 most popular course combinations among students. Each section of the chart represents a different combination, with the size of the section showing how many students chose it. The most popular combination, "CSE103, CSE106, ENG101, MAT101," makes up 38.1% of the total, followed by "CHE109, CSE103, CSE106, ENG101," which accounts for 29.9%. The third most chosen combination is "CSE103, CSE106, MAT101, ENG101," at 14.6%. The remaining combinations are less common, each making up smaller percentages of the total.

6. Conclusion:

This mini project analyzes the course selection and credit trends of 2960 students based on a dataset of 60 courses. Key findings reveal that CSE106 is the most popular course, while courses like CE200 and GEN202 are less frequently chosen. Common course combinations, such as CSE103, CSE106, ENG101, and MAT101, were also identified. Most students typically take 4 courses per semester, with some opting for 1 or 6. Credit trends show that students generally complete 14-15 credits per semester, with no clear correlation between credits taken and credits completed. Various visualizations, including histograms, bar charts, and heatmaps, provided a clear view of course preferences and credit distributions. Additionally, a recommendation system was developed to suggest courses based on previous selections. Trends over time showed how course popularity shifted across academic years. In conclusion, this project uncovers valuable insights into student course selection patterns, offering potential improvements for course offerings, advising strategies, and student planning.