

Exploratory Data Analysis of Quercus Activity and First-Year Retention

Amanda Ng

2025-01-06

1. Data Collection and Preparation

1.1 Dataset source and brief description

Our data set was collected from 2022-2023 University of Toronto first year students between September 1, 2022 and December 7, 2022. The data set was extracted from two sources: Quercus Records Store and Institutional Retention.

Quercus is a U of T's learning management system (LMS) and all clicks are tracked in the transactional data set Quercus Records Store (QRS). QRS was collected with primary use to provide individual instructors insight into patterns of activity in their courses. It was updated daily and indicated what was clicked on and which items were viewed and when. However, in our extracted data set, we only have information about whether Quercus was accessed on a day. In other words, we do not have information about the items and the exact time of the student's interaction with Quercus. Note that students who never interacted in Quercus at all during the term are not recorded in the data.

Institutional Retention is a database that tracks the retention status (either active, inactive, or graduated) of students by cohort (defined as the year when they started at the University of Toronto). It gets updated once a year on November 1st. In our extracted data set, all students are from the same cohort (i.e. 2022) and we have information on the students' second-year retention status (either active or inactive), first-year academic performance, and some demographic information.

The data collected for this study aims to investigate the association between students' engagement patterns in Quercus and their likelihood of dropping out. The research is guided by two primary questions:

1. Does the trajectory of Quercus access throughout the academic term influence a student's decision to drop out? By analyzing the cumulative access data over time, we seek to identify any trends in engagement that correlate with the decision to drop out.
2. At what point (in terms of access frequency or specific dates/weeks), should we be particularly concerned about a student potentially dropping out? By tracking and correlating the timing of changes in access behavior with drop-out decisions, we hope to uncover critical indicators that could serve as early warning signs for student disengagement.

These analyses will inform interventions, such as assigning college registrars to conduct outreach to support students who are at risk of dropping out based on their interaction with the course platform.

1.2 Raw data dimensions

The raw data is in wide format, meaning that each row represents information collected from a student and no more than one row contains information from the same student. It consists of 3311 observations and 109 variables.

1.3 Data types of variables

Below is a summary description of all 109 variables in the dataset.

- Randomized student ID: Numerical variable
- Incoming GPA before students began at UofT: Categorical variable
 - No admission average
 - less than 76%
 - 76-80%
 - 81-85%
 - 86-90%
 - 91-95%
 - greater than 95%
- Gender: Categorical variable
 - Male
 - Female
 - Unreported
- Immigration Category: Categorical variable
 - Domestic
 - International
- First Term Course Count as of November 4, 2022: Numerical variable ranging from 0 to 6
- Students registration status as of November 4, 2022: Categorical variable
 - Registered
 - Financially cancelled
 - Deregistered
 - Cancelled
 - Invited
- Year 1 CGPA: Numerical variable ranging from 0 to 4.0
- Year 1 Final Academic Standing: Categorical variable
 - In good standing
 - On academic probation
 - No Registration
 - No assessment data
 - Suspended for one year
- Year 2 Active: Binary variable with 1 indicating active for Fall 2023
- Year 2 Inactive: Binary variable with 1 indicating inactive for Fall 2023
- *1/9/2022 - 7/11/2022*: Binary variables with 1 indicating student accessed Quercus that day

2. Data Cleaning

2.1 Missing value

There are 3 observations with missing values in the “Incoming GPA before students began at UofT” variable, we placed these observations under the “No admission average” category. After cleaning, there are 347 observations with no admission average. Due to the significant size, we did not remove these observations to ensure the representativeness of the sample.

2.2 Data inconsistencies

No discrepancies were identified while examining the data for inconsistencies between the “Year 2 Active” and “Year 2 Inactive” variables. Specifically, there are no cases where both variables were simultaneously marked as 1 (active and inactive) or both marked as 0 (neither active nor inactive). This confirms that the variables are mutually exclusive and consistently coded.

2.2 Remove non-registered students

We are only interested in students who were properly registered in Fall 2022 and wish to predict their probability of dropping out in the coming year. Therefore, we excluded any student who were not properly registered in Fall 2022, as their data may not be relevant to the research question. This process removed data from an insignificant amount of observations (12 students), resulting in a total of 3299 observations after cleaning.

2.3 Remove unnecessary variables

The “Yr 2 Active” variable was removed because it is redundant when paired with the “Yr 2 Inactive” variable. Since these variables are mutually exclusive and collectively exhaustive (i.e., if one is 1, the other must be 0, and vice versa), retaining only “Yr 2 Inactive” provides all necessary information while reducing duplication and simplifying the dataset. The “Students registration status as of November 4, 2022” variable was also removed because it is irrelevant to this study after data cleaning.

2.4 Duplicate entries

There are no duplicated observations in the data set.

2.5 Data transformation

In the raw data’s original wide format, each date is represented as a separate column, which can make it cumbersome to identify trends across all dates and perform trend comparisons between different student subgroups. We reshaped the data into a long format to make it more efficient for later visualization and analysis, since the data includes repeated measurements over time. By reorganizing the data so that all dates are listed in a single column, with their corresponding binary Quercus interaction values in another column, we created a data structure in which each row corresponds to an occurrence or missing Quercus interaction event on a specific date. Note that multiple rows in the long format data may refer to the same student if they accessed Quercus on multiple days. This format makes it easier to filter, group, and analyze data based on specific periods and simplifies the process of generating visualizations, such as the time-series plots in section 3.2.

3. Descriptive Statistics

In total, there are 3311 students in this study. Among them, 2953 students remained active in Fall 2023, while 346 students dropped out after their first year. This distribution highlights that the majority (nearly 90%) of first-year students in the data set remained enrolled at the University of Toronto in Year 2.

This study’s observation window is 1/9/2022 - 7/11/2022, totaling to 98 days or 15 weeks.

3.1 Summary statistics and Visualizations of variables determined before Fall 2022

3.11 Incoming GPA

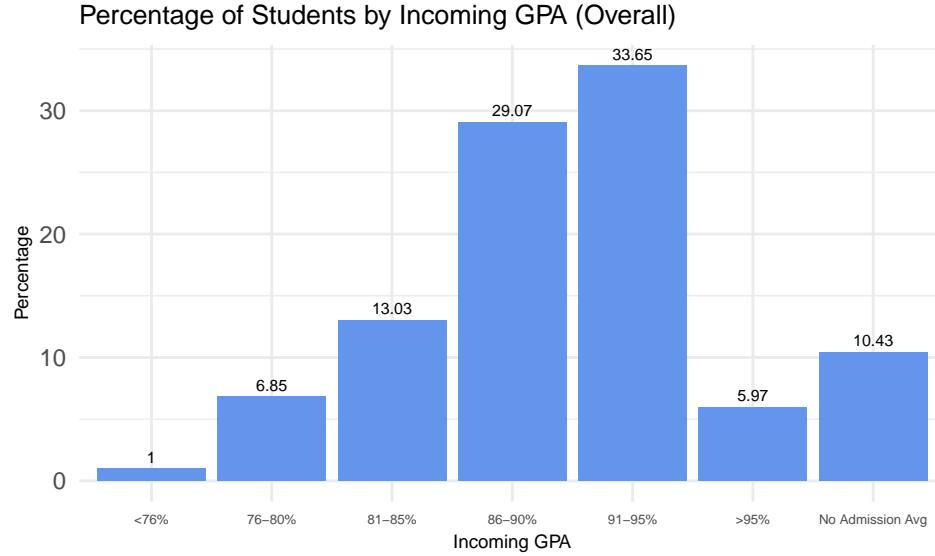


Figure 1: Percentage of Students by Incoming GPA (Overall). This bar plot illustrates the distribution of all students based on their incoming GPA, categorized into six groups: less than 76%, 76-80%, 81-85%, 86-90%, 91-95%, and above 95%. An additional category, “No Admission Avg,” represents students without an admission GPA recorded. The value on top of each bar represents the percentage of students falling into what category.

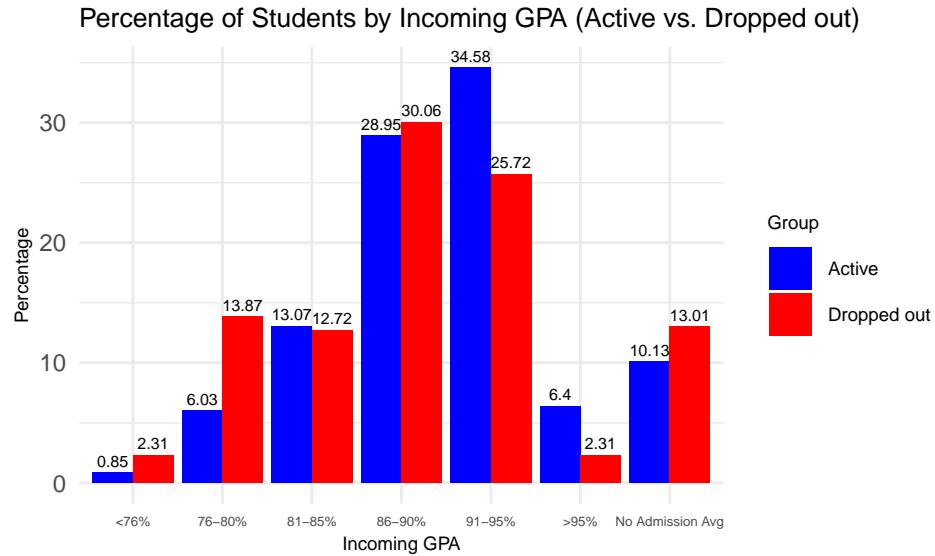


Figure 2: Percentage of Students by Incoming GPA (Active vs Dropped out). This bar plot illustrates the distributions of active and dropped-out students based on their incoming GPA, categorized into six groups: less than 76%, 76-80%, 81-85%, 86-90%, 91-95%, and above 95%. An additional category, “No Admission Avg,” represents students without an admission GPA recorded. The blue bars correspond to active students, while the red bars represent dropped-out students. The value on top of each bar indicates the percentage of students within each category.

Figures 1-2 illustrate the distribution of students across different incoming GPA categories for all students, students who remained enrolled in their second year, and those who dropped out. For active students, the majority fall into the 91–95% and 86–90% GPA ranges, accounting for 34.58% and 28.95%, respectively. Dropouts show a similar trend, with 30.06% in the 86–90% range and 25.72% in the 91–95% range, though the distribution is more spread across lower GPA brackets compared to active students. Notably, we observe the percentage of students who had >95% incoming GPA among active students is 2.84 times higher than those who dropped out.

The greater proportion of students in lower GPA brackets among dropouts highlights that lower incoming academic performance may predispose students to disengagement and eventual withdrawal. Furthermore, the significantly higher proportion of active students with >95% incoming GPA suggests that high academic achievers may be more resilient to disengagement, which could influence the interpretation of engagement behaviors. This finding indicates that incoming GPA could be an important covariate in the model when analyzing the effect of Quercus engagement patterns on drop-out likelihood, ensuring that the model accounts for the baseline differences in academic preparedness.

3.12 Gender

Table 1: Gender Distribution Across Student Groups

Group	Female (%)	Male (%)	Unreported (%)
Overall	54.80	42.19	3.00
Active	55.50	41.65	2.84
Dropped out	48.84	46.82	4.34

According to Table 1, the gender distribution across student groups shows only slight variation in the proportions of females and males. In the overall population, females constitute 54.80%, compared to 42.19% males. Among active students, the proportion of females is slightly higher at 55.50%, 13.85% higher than that of males. For students who dropped out, the distribution is more balanced, with 48.84% females and only 2.02% higher than that of males. These differences suggest relatively minor shifts in gender composition between the active and dropped-out groups.

The results of the gender distribution analysis suggest that gender composition differences across all students, active students, and dropouts are relatively minor and may not play a significant role in explaining the likelihood of dropping out based on engagement patterns in Quercus. These findings imply that gender may not be a strong predictor of drop-out likelihood when compared to other factors, such as engagement patterns in Quercus or academic performance. Nonetheless, we may still consider gender in the modeling process as a covariate, to ensure that any subtle baseline associations between gender and engagement patterns are accounted for.

3.13 Immigration Category

Table 2: Percentage of Students by Immigration Category Across Student Groups

Group	Domestic (%)	International (%)
Overall	77.02	22.98
Active	76.97	23.03
Dropped out	77.46	22.54

According to Table 2, the distribution of students by immigration category is fairly consistent across all groups. The majority of students in the overall population, as well as in the active and dropped-out groups, are domestic, accounting for around 77%. Meanwhile, around 23% of the students were international. There are no significant differences in the proportions of domestic and international students between the groups. However, note that the percentage of domestic students within the dropped-out group is slightly higher than the active group.

This suggests that the immigration category possibly only plays a minor role in explaining differences in drop-out likelihood. If it is included as a covariate in the model, we expect that the effect of this variable in predicting drop-out likelihood is small. Yet, it can capture underlying factors related to engagement patterns, such as differences in access to resources, cultural adjustment, or language barriers, hence it is still worth including in the model.

3.14 First Term Course Count

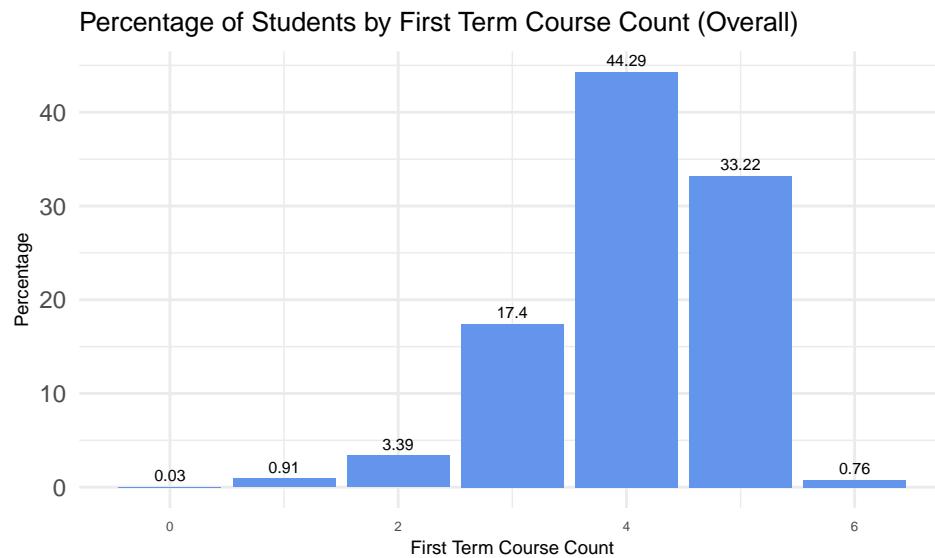


Figure 3: Percentage of Students by First Term Course Count (Overall). This bar plot illustrates the distribution of all students based on the number of courses they took during the 2022 Fall, ranging from 0 to 6. The value on top of each bar indicates the percentage of students within each category.

Percentage of Students by First Term Course Count (Active vs. Dropped out)

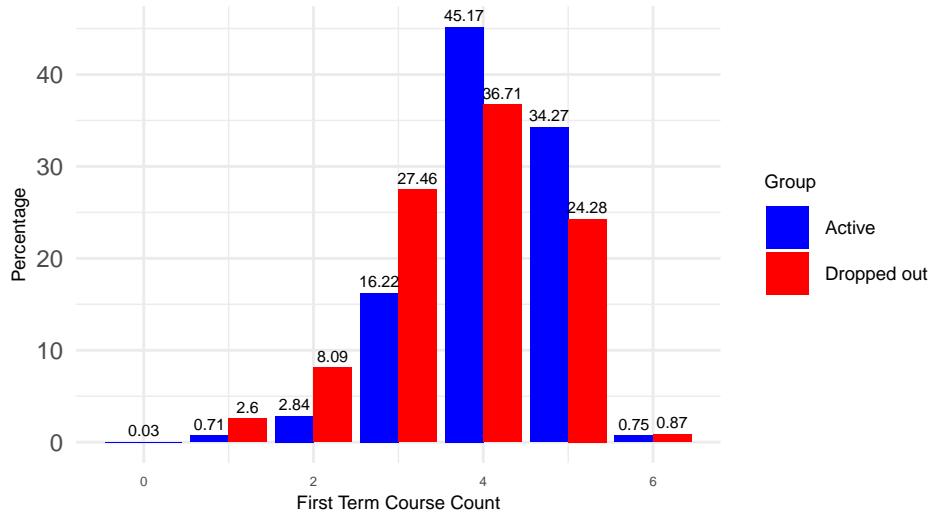


Figure 4: Percentage of Students by First Term Course Count (Active vs Dropped out). This bar plot illustrates the distributions of active and dropped-out students based on the number of courses they took during the 2022 Fall, ranging from 0 to 6. The blue bars correspond to active students, while the red bars represent dropped-out students. The value on top of each bar indicates the percentage of students within each category.

Figures 3-4 highlight the distribution of students by their first-term course count across overall, active, and dropped-out students. A consistent trend found across all groups is that the largest proportion of students concentrated in the middle range of course counts (e.g., around 3-5 courses), though the exact percentages vary. In the overall and active groups, the majority of students are clustered around course counts of 4 and 5, with noticeable tails at lower and higher counts. For students who dropped out, the distribution is somewhat more evenly spread and shows a higher percentage at lower course counts compared to active students. This suggests that a lower course load might be associated with higher dropout rates.

These results suggest that course load is a potentially important factor in understanding the likelihood of dropping out. The higher proportion of students with lower course counts among those who dropped out indicates that a reduced course load might be associated with disengagement or challenges that increase the likelihood of leaving. On the other hand, lower course loads could influence engagement patterns in Quercus, such as less frequent logins, since lower degree of participation is expected when students are enrolled to fewer courses. These findings highlight the need to include first-term course count as a covariate since it could serve as a proxy for workload and academic commitment affecting students' ability to manage their studies.

3.2 Summary statistics and Visualizations of variables determined after Fall 2022

3.21 Year 1 CGPA

Histogram of End of year CGPA (Overall)

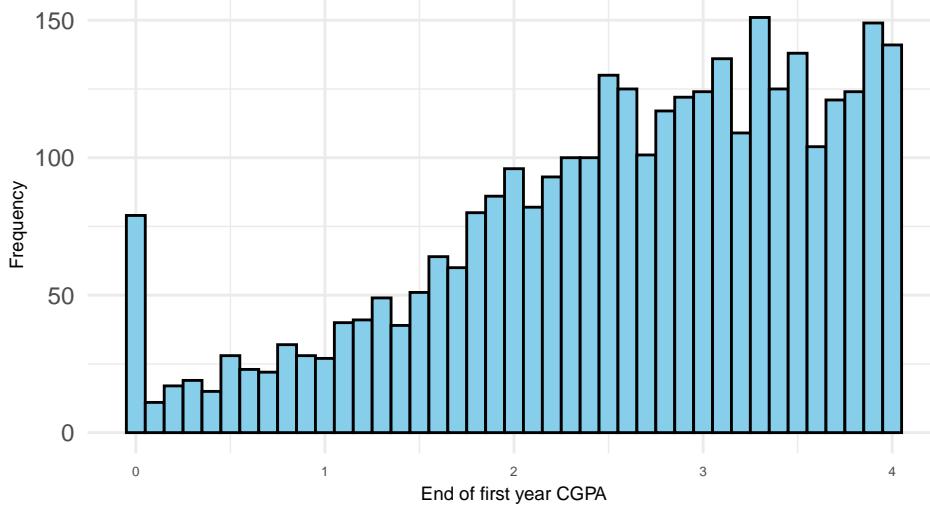


Figure 5: Histogram of End of first year CGPA (Overall). This histogram shows the frequency distribution of first-year CGPA across all students, regardless of whether they dropped out after year 1. It is approximately left-skewed, with most CGPA values clustering at the higher end of the GPA range, i.e. closer to 4.0.

Histogram of End of Year CGPA (Active vs. Dropped out)

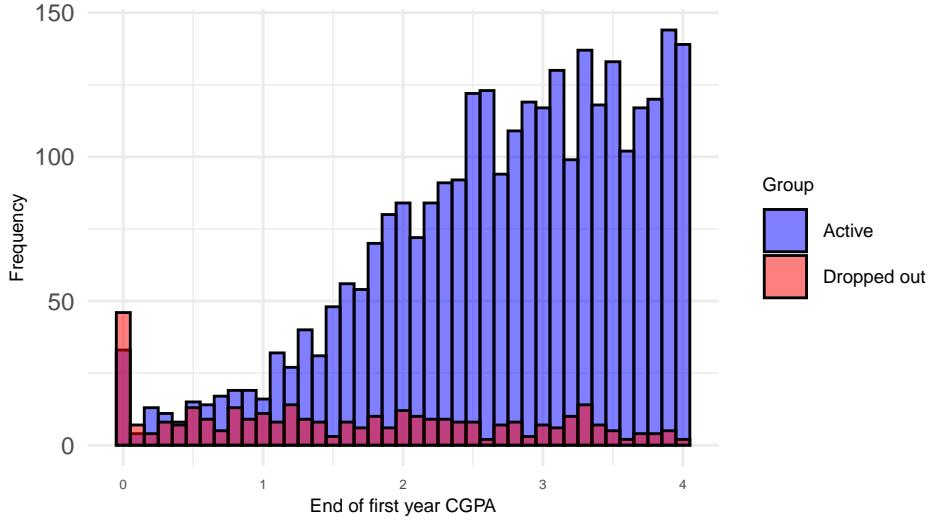


Figure 6: Histogram of End of first year CGPA (Active vs Dropped out). This histogram compares the frequency distributions of first-year CGPA between active and dropped-out students. The active students' histograms are approximately left-skewed, with most CGPA values clustering at the higher end of the GPA range, i.e. closer to 4.0. The dropped-out students' histogram, on the other hand, has a peak at 0 and is approximately uniform across other non-zero values.

Figures 5-6 display the distribution of students' end-of-first-year CGPA across overall, active, and dropped-out students. The overall and active student histograms are highly left-skewed, suggesting that a majority of students overall and within the active group achieved CGPAs between 2 and 4. There is a peak near the upper end of the scale, indicating a higher frequency of students with CGPAs closer to 4. There are

significantly fewer students who struggled academically as reflected by the low frequencies at the lower end of the scale. Conversely, the dropped-out students histogram is highly right-skewed, with a distinct peak at the lower end of the scale, indicating a higher frequency of students with near 0 CGPA. The frequencies of students with non-zero CGPAs are roughly uniform and much lower than that of near 0 CGPA. In short, there is a significant difference in the distribution of end-of-first-year CGPA between active and dropped-out students.

Table 3: Summary table of Year 1 CGPA

Group	Mean	Median	Standard deviation
Overall	2.596847	2.8	1.0177752
Active	2.711243	2.8	0.9318671
Dropped out	1.620520	1.5	1.1849103

Table 3 highlights notable differences in academic performance between the overall group, active students, and students who dropped out. Active students tend to perform better academically, with higher average grades and less variation in their performance. In contrast, students who dropped out show significantly lower academic achievement on average, with more variability in their grades. This suggests that students who remained active after first year tend to perform better and consistently academically and those who dropped out had lower and more varied performance. Additionally, greater variability in first-year CGPA among students who dropped out might reflect diverse underlying challenges, including academic, personal, or external factors, that impact both engagement and the decision to leave.

3.22 Year 1 Final Academic Standing

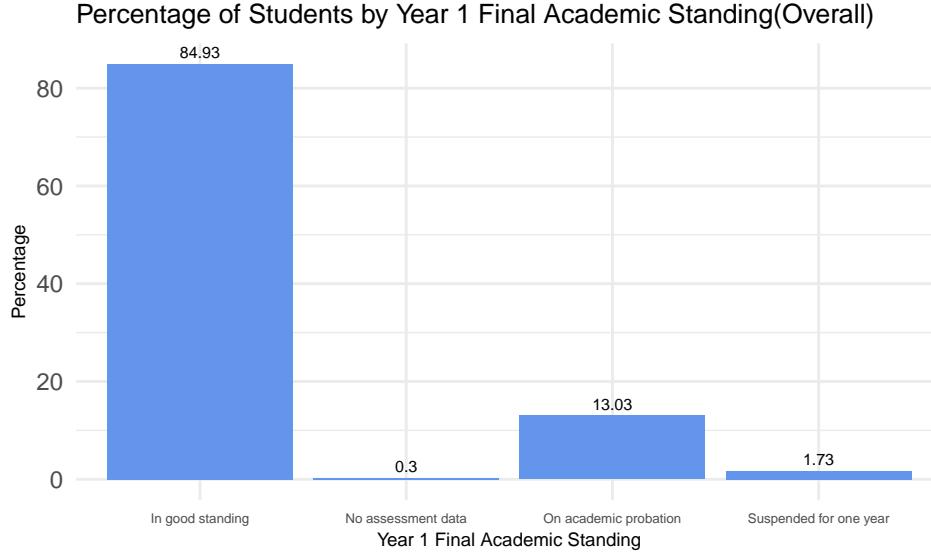


Figure 7: Percentage of Students by Year 1 Final Academic Standing (Overall). This bar plots illustrate the distribution of all students based on their incoming GPA, categorized into six groups: In good standing, No assessment data, No registration, On academic probation, and Suspended for one year. The value on top of each bar represents the percentage of students falling into what category.

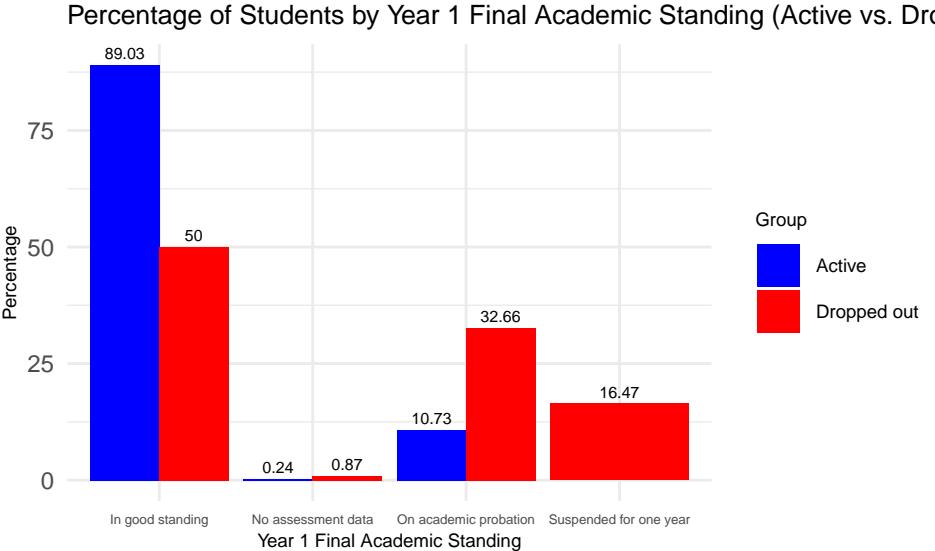


Figure 8: Percentage of Students by Year 1 Final Academic Standing (Active vs Dropped out). This bar plot illustrates the distributions of active and dropped-out students based on their incoming GPA, categorized into six groups: In good standing, No assessment data, No registration, On academic probation, and Suspended for one year. The blue bars correspond to active students, while the red bars represent dropped-out students. The value on top of each bar indicates the percentage of students within each category.

Figures 7-8 show the percentage distribution of students' final academic standing after year 1 between the overall group, active students, and students who dropped out. A common pattern across both plots is that the majority of students are "In good standing," though the proportion varies. Overall speaking, 84.93% of students are in good standing, while this percentage increases to 89.03% among "Active" students. Conversely, the "Dropped Out" group has only 50.00% in good standing, showing a significant contrast. Students "On academic probation" form a notable portion in each group but are highest in the "Dropped Out" group at 32.66%, compared to 13.03% overall and 10.73% among active students. Another distinctive pattern is seen in the percentage of students "Suspended for one year," which is far higher for those who dropped out (16.47%) than for the overall group (1.73%), with none recorded in the active group. Negligible percentages are observed across all categories for "No assessment data" and "No Registration".

In short, students in good academic standing are predominantly active, while those who dropped out are more likely to be on academic probation or suspended, suggesting that poor academic performance status strongly correlates with dropping out. This suggests that academic standing serves as a critical factor in dropout decisions and may confound the relationship between engagement patterns and dropout likelihood.

However, both the "End-of-year GPA" and "Year 1 Final Academic Standing" variables should not be included as predictors in the model for this study, as they are based on information that becomes available only after the fall term. These variables may, nonetheless, be useful for future analysis once full-year data is available, as they could provide a more comprehensive understanding of the factors influencing student outcomes throughout the entire academic year.

3.3 Visualization of Quercus engagement patterns

To analyze the cumulative Quercus visits over time, we generated separate spaghetti plots for the active and inactive student groups.

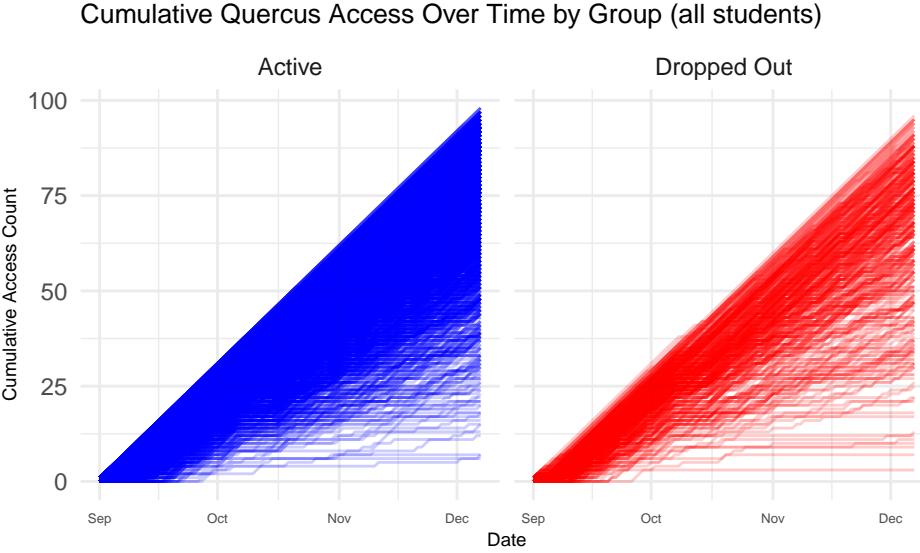


Figure 9: Cumulative Quercus access over time by group (Active/ Dropped out). The spaghetti plots compare the cumulative Quercus access counts over time for two groups: students who remained active (left, in blue; $n = 2953$) and students who dropped out (right, in red; $n = 346$) during 1/9/2022 - 7/11/2022. Each line represents the Quercus access visits trajectory of an individual student from a sample of each group. The diagonal lines in the plots represent students who visited Quercus daily, while horizontal lines indicate periods where students did not visit the platform.

Given the substantial group size difference between the active and dropout groups, we drew random samples of size 100 from each group to ensure a fair comparison in the visualization.

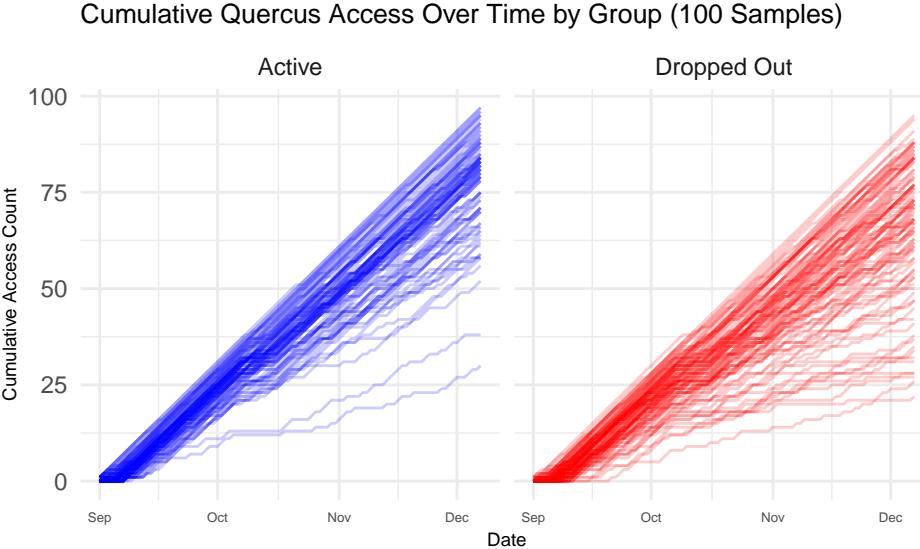


Figure 10: Cumulative Quercus access over time by group (Active/ Dropped out) of random samples $n = 100$. The spaghetti plots compare the cumulative Quercus access counts over time for two groups: students who remained active (left, in blue) and students who dropped out (right, in red) during 1/9/2022 - 7/11/2022.

Each line represents the Quercus access visits trajectory of an individual student from a sample of 100 in each group.

According to figures 9 and 10, both the active and dropped-out groups show an overall increase in cumulative access counts over time, but the trajectories differ in their spread and consistency. The active group exhibits tightly clustered lines, indicating more consistent and uniformly high engagement across students. Conversely, the dropped-out group shows more variability, with more students displaying slower growth in cumulative access counts or flatter trajectories, suggesting lower or irregular engagement. This difference implies that students with lower or irregular engagement are more likely to drop out, reinforcing the idea that cumulative engagement is a potential predictor of dropout risk.

When modeling the association between Quercus engagement and dropout risks, the findings above suggest that cumulative access counts can serve as a key predictor, potentially modeled as a time-varying covariate to capture Quercus engagement patterns. The variability in the engagement trajectories of the dropped-out group indicates that the timing and magnitude of access could play a significant role in identifying at-risk students.

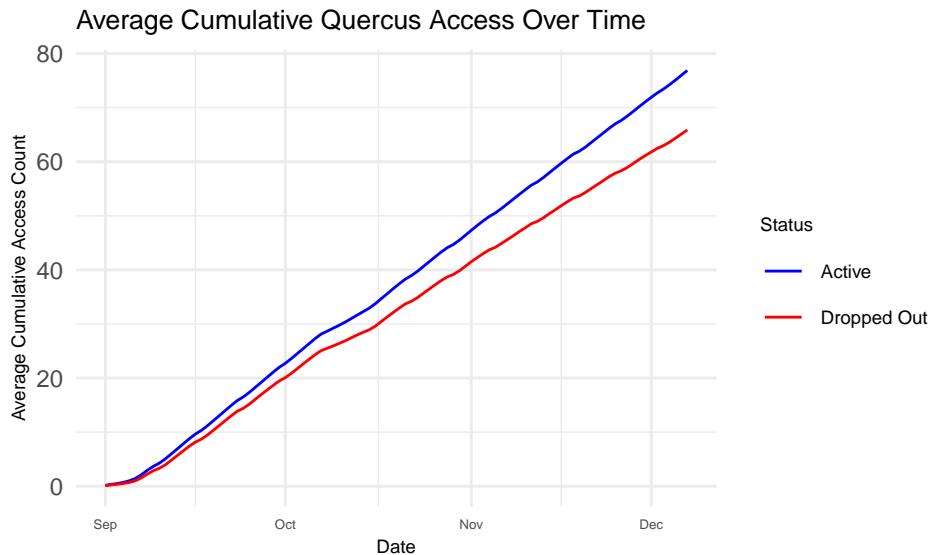


Figure 11: Average Cumulative Quercus access over time by group (Active/ Dropped out). The spaghetti plot displays the averaged cumulative Quercus visits of each group during 1/9/2022 - 7/11/2022. The blue represents the averaged cumulative Quercus access distribution among students who remained after the first year while the red line represents that among students who dropped out.

Additionally, we calculated and plotted the **average** cumulative Quercus visit line for each group. The steeper trajectory observed in the active group indicates sustained or increasing engagement over time, while the flatter trajectory in the dropout group suggests lower or declining engagement. This pattern implies that students who engage more frequently with Quercus are less likely to drop out, highlighting the potential association of Quercus engagement with dropout risks.

3.4 Correlation analysis

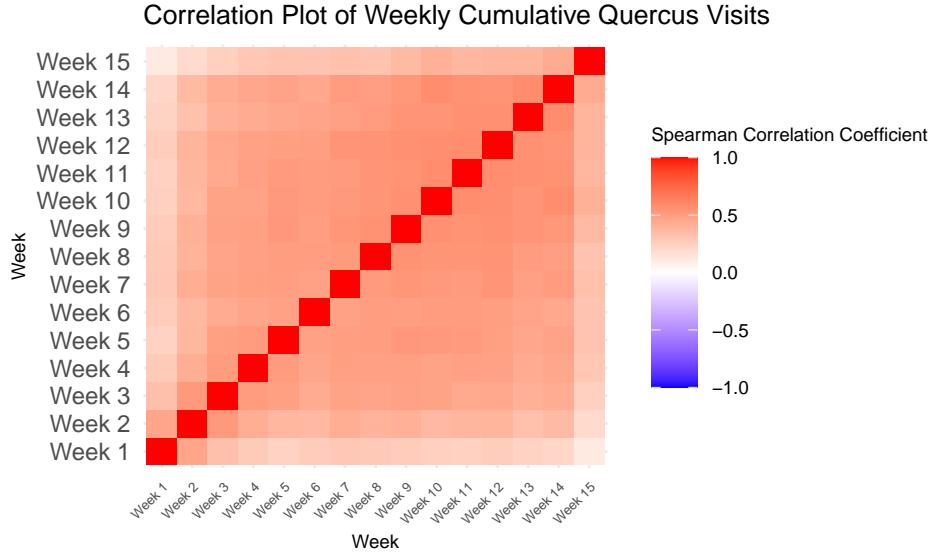


Figure 12: Correlation Plot of Weekly Cumulative Quercus Visits. This plot shows the correlation of the cumulative Quercus visits value between each pair of weeks. Darker red grids represent stronger positive correlations, while darker blue grids indicate stronger negative correlations.

Figure 12 reveals that all pairwise correlations cumulative Quercus visit counts between weeks are positive, as indicated by the red tones across the grid. As expected, the strongest correlations are observed along the diagonal, where each week is perfectly correlated with itself (correlation coefficient = 1.0). As the time gap between weeks increases, the correlation weakens slightly, reflected by lighter shades of red further from the diagonal. These results suggest that cumulative Quercus visits are consistently and positively correlated over time, with stronger relationships between consecutive or temporally closer weeks.

The observed correlation pattern, where correlations are strongest between adjacent weeks and gradually weaken as weeks become farther apart, suggests that an autoregressive structure of order 1 (AR(1)) may be appropriate for modeling the covariance structure. The AR(1) structure assumes that the covariance between observations decreases exponentially with increasing time lag, aligning with the gradual reduction in correlation observed in the plot. This structure effectively captures the temporal dependence seen in our dataset, which consists of repeated measures of cumulative Quercus visits from the same set of students over time. In this study, the AR(1) covariance structure is particularly suitable because it models the intuitive idea that a student's Quercus engagement behavior in a given week is strongly influenced by their behavior in the immediately preceding week, with this influence diminishing as the temporal gap increases. Specifically, the AR(1) structure assumes that the covariance between visits in weeks t and $t + k$ is proportional to ρ^k , where ρ is a parameter reflecting the strength of the temporal dependency.

Depending on our final model, we may need to specify the covariance structure of the outcome variable (e.g. dropout likelihoods) and capture the within-student correlation over time. Just like the cumulative Quercus visits exhibit temporal dependence, we expect the likelihood/intention of dropping out at different time-points of the same student to follow a similar pattern. By incorporating the AR(1) structure into our model, we account for this decaying temporal dependence, ensuring that the repeated measures structure of the data is appropriately modeled.