1. Introduction and Motivation

The purpose of this study is to investigate which mentor and mentoring factors have a meaningful impact on student retention at the Coachify coaching platform.
The main goal is to identify potential improvements in mentor practices to enhance student engagement and retention rates, which ultimately would improve the platform's profitability and effectiveness.

---

2. Data Collection and Preprocessing

The data were collected from two sources:

- The Coachify student database, which included student enrollment information (mentor assignment, membership start month, dropout month).
- A survey administered to **30 mentors**, gathering data about their weekly call frequency, average call duration, messaging habits, and primary communication format (video or voice).

Data Cleaning Steps:

- Students who initially purchased long-term packages were excluded to avoid bias towards artificially high retention.
- Mentors with **fewer than 3 students** were excluded to reduce statistical noise.
- Mentor and student names were anonymized.
- Weekly survey responses were multiplied by 4 to approximate **monthly** activity levels.

The cleaned dataset included **22 mentors** and approximately **116 students**.

---

3. Retention Metric Definition

Retention was defined as **the transition from the 1st month to the 2nd month**.

- **Pass Rate** = (Number of students continuing to 2nd month) / (Total number of students)

This metric was used because:

- It captures early-stage student engagement, which is critical in subscription-based models.
- It provides a clear, binary outcome (continued/dropped) ideal for statistical comparison.

4. Research Questions

The analysis aimed to answer the following research questions:

RQ1: Does the mentor's YKS ranking affect student retention?

- **H_0 (Null Hypothesis):** Mentor's YKS ranking is not correlated with student retention (Pass Rate).
- **H_1 (Alternative Hypothesis):** Mentor's YKS ranking is correlated with student retention (Pass Rate).

RQ2: Does the primary communication format (video or voice) impact student retention?

- **H_0:** There is no difference in student retention (Pass Rate) between mentors using video and those using voice communication.
- **H_1:** There is a difference in student retention (Pass Rate) between mentors using video and those using voice communication.

RQ3: Is there a relationship between the average call duration and student retention?

- **H_0:** Average call duration is not correlated with student retention (Pass Rate).
- **H_1:** Average call duration is correlated with student retention (Pass Rate).

RQ4: Does the frequency of weekly calls affect retention?

- **H_0:** Weekly call frequency is not correlated with student retention (Pass Rate).
- **H_1:** Weekly call frequency is correlated with student retention (Pass Rate).

RQ5: Does the frequency of sending progress monitoring messages affect retention?

- **H_0:** Monthly progress message frequency is not correlated with student retention (Pass Rate).
- **H_1:** Monthly progress message frequency is correlated with student retention (Pass Rate).

RQ6: Does the number of weekly communication days correlate with student retention?

- **H_0:** Weekly communication days are not correlated with student retention (Pass Rate).
- **H_1:** Weekly communication days are correlated with student retention (Pass Rate).

RQ7: Is there a relationship between Pass Rate and Average Normalized Retention?

- **H_0:** Pass Rate is not correlated with Average Normalized Retention.
- **H_1:** Pass Rate is correlated with Average Normalized Retention.

5. Methodology

Two levels of analysis were conducted:

- **Mentor-Based Analysis:**
  Each mentor was treated as one data point (n=22).
  Pass Rate was compared against mentor attributes.
- **Student-Based Analysis:**
  Each student was treated as one data point (n≈116).
  Student pass/fail outcomes were compared against their assigned mentor's characteristics.
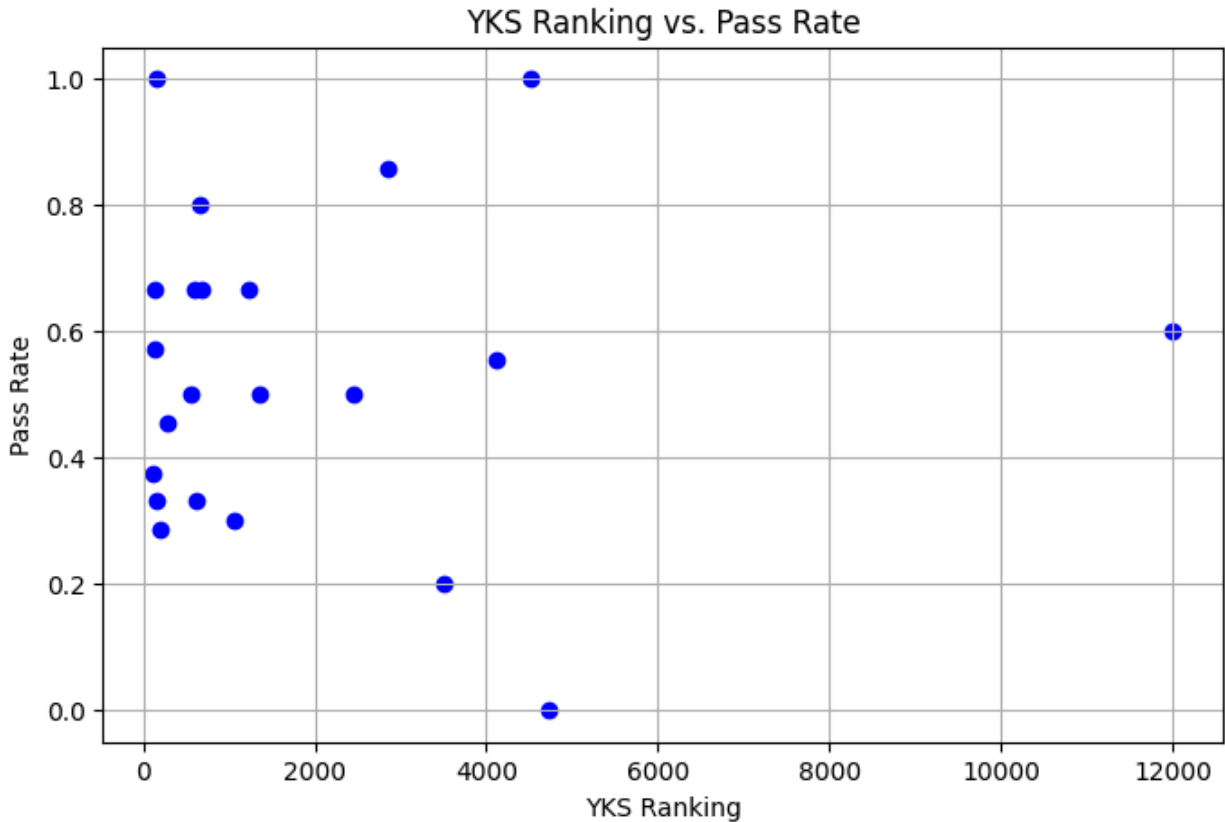
Statistical methods used:

- Pearson Correlation (for continuous variables)
- Independent Samples t-test (for categorical variables)
- Bar plots, boxplots, and stacked bar plots for visualization

All methods were consistent with the techniques covered during coursework.

6. Findings

6.1 Mentor-Based Findings

6.1.1 Effect of Mentor YKS Ranking on Student Retention

(Graph 1: Scatter Plot - YKS Ranking vs Pass Rate)

No significant relationship was found between mentors' YKS ranking and the student pass rate. The Pearson correlation was **r = 0.000** with a **p-value = 0.999**, indicating no linear relationship. Thus, a mentor's national exam ranking does not seem to influence students' short-term retention.

6.1.2 Impact of Communication Format (Video/Voice) on Retention

## Pass Rate by Communication Format
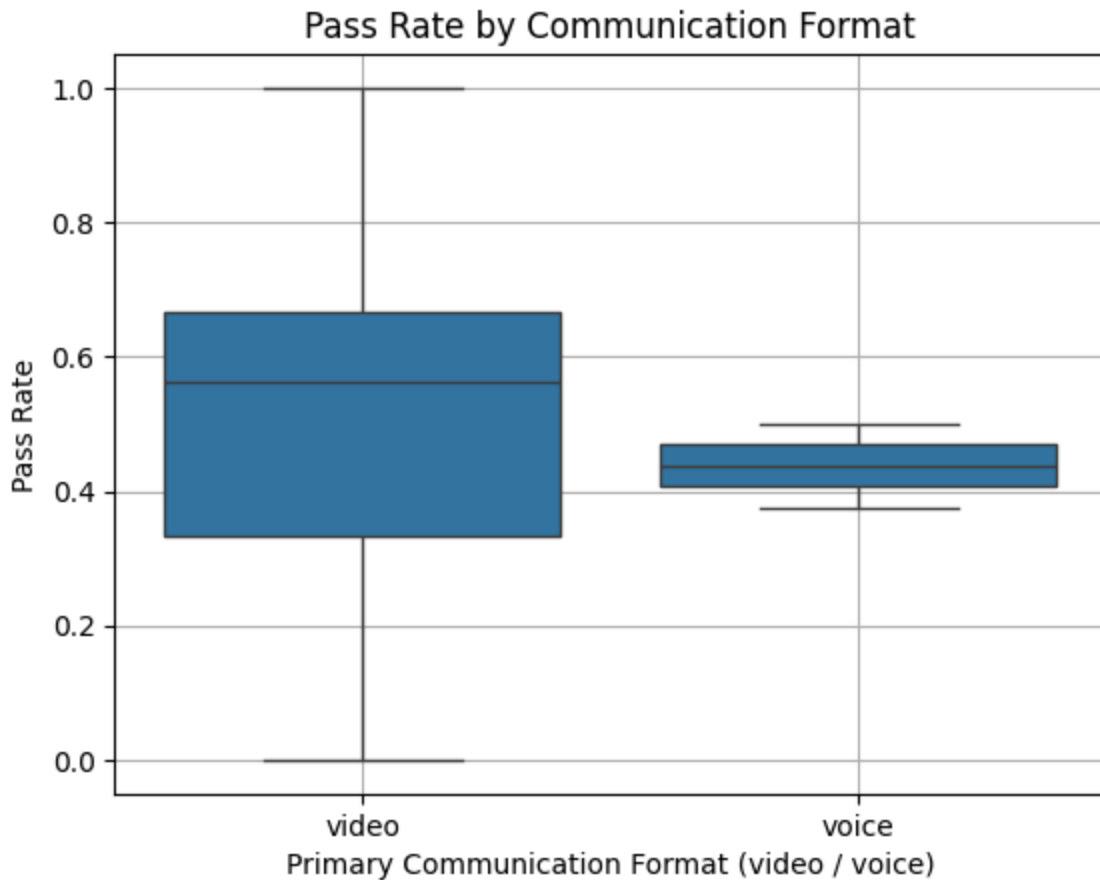
(Graph 2: Boxplot - Communication Format vs Pass Rate)

The comparison between mentors who primarily used video calls and those who used voice calls showed no statistically significant difference in pass rates.
The t-test results were **t = 1.293** and **p-value = 0.278**, suggesting that the communication method does not have a meaningful impact on student retention.

**6.1.3 Relation Between Average Call Duration and Retention**
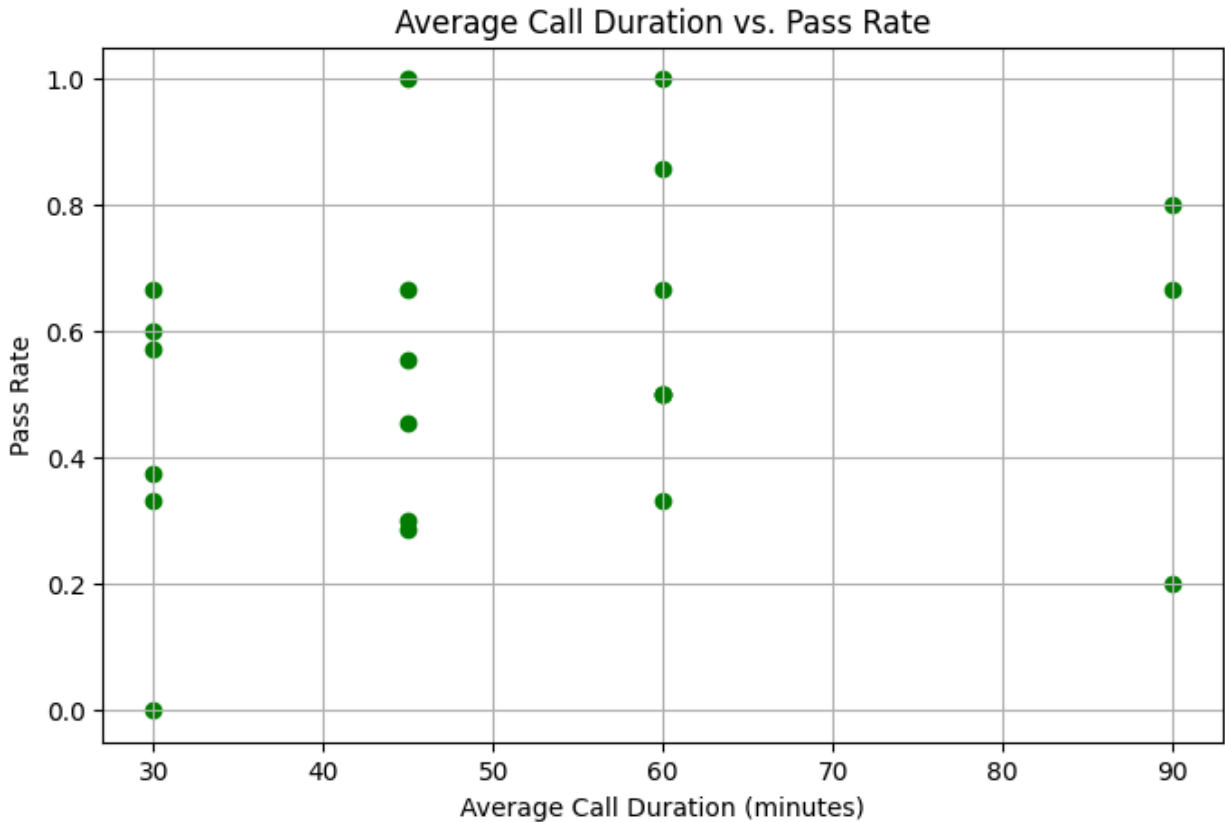
(Graph 3: Scatter Plot - Average Call Duration vs Pass Rate)

No significant correlation was found between the average call duration and the student pass rate.
The Pearson correlation coefficient was **r = 0.210** with a **p-value = 0.3487**, indicating a weak and non-significant relationship.

6.1.4 Effect of Weekly Call Frequency on Retention

Weekly Call Frequency vs. Pass Rate
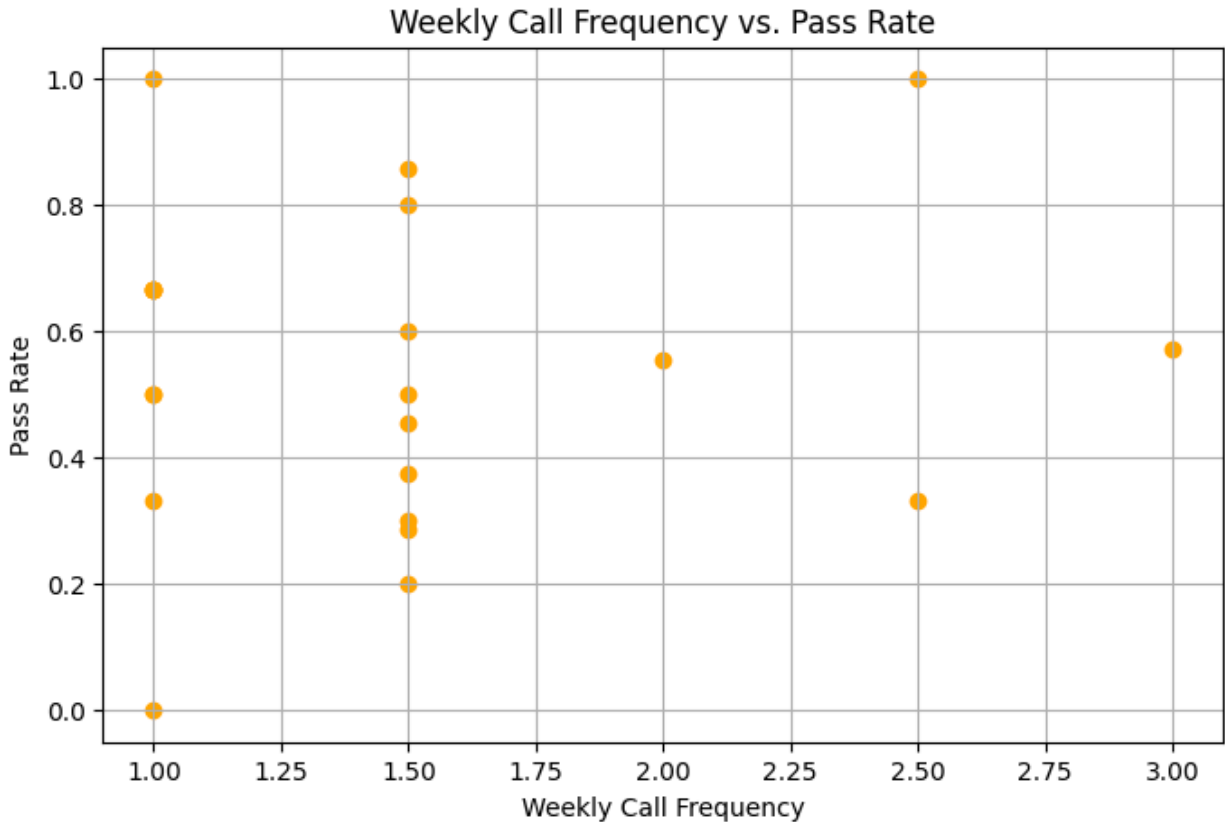
(Graph 4: Scatter Plot - Weekly Call Frequency vs Pass Rate)

Weekly call frequency showed no significant correlation with pass rates.
The Pearson correlation was **r = 0.080** and the **p-value = 0.7244**.
This result suggests that calling students more often does not necessarily increase the likelihood of their retention.

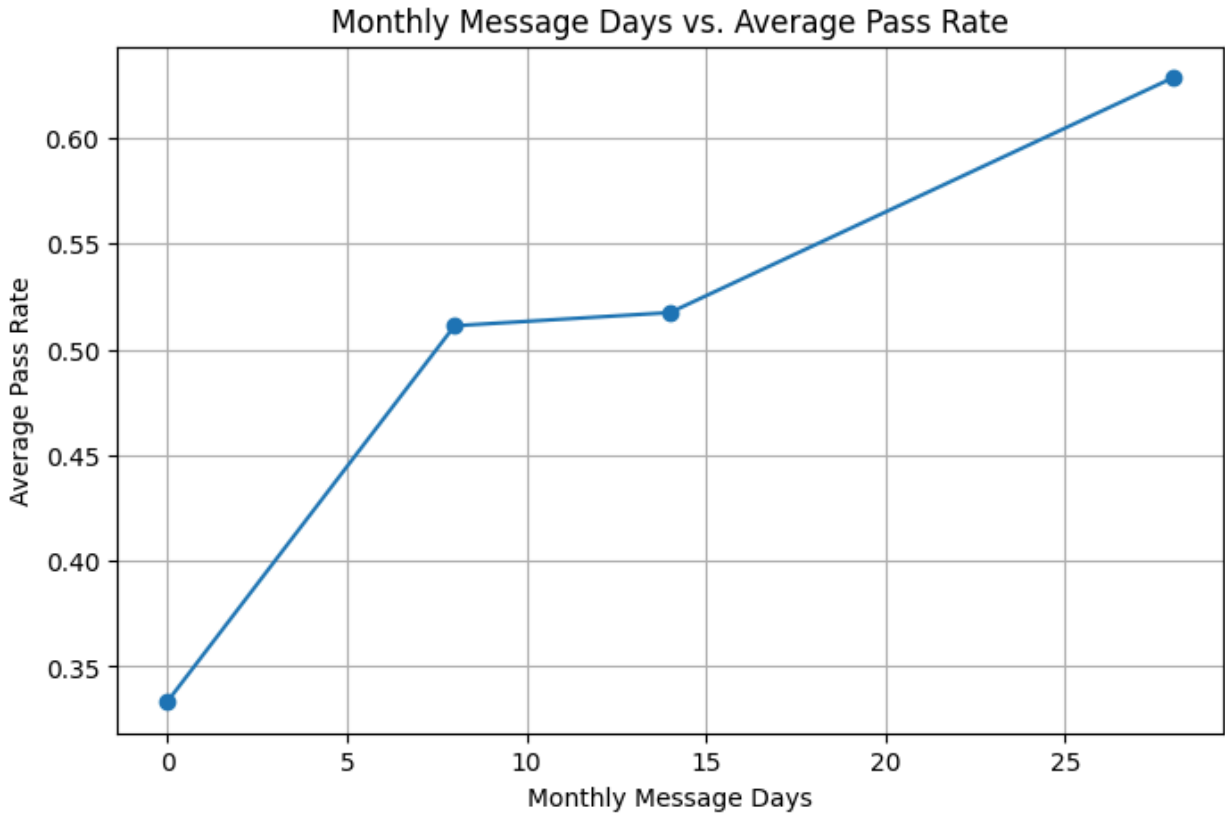### 6.1.5 Effect of Monthly Progress Messages on Retention

(Graph 5: Scatter Plot - Monthly Progress Messages vs Pass Rate)

There was no statistically significant relationship between the number of monthly progress messages sent and the student pass rate.

The Pearson correlation was **r = 0.163** with a **p-value = 0.468**.

**6.1.6 Impact of Monthly Message Days on Retention**

(Graph 6: Scatter Plot - Monthly Message Days vs Pass Rate)

Similarly, no significant relationship was found between the number of days mentors communicated with students via messaging and the student pass rate.
The Pearson correlation was **r = 0.256** with a **p-value = 0.251**.

6.1.7 Relationship Between Pass Rate and Average Normalized Retention
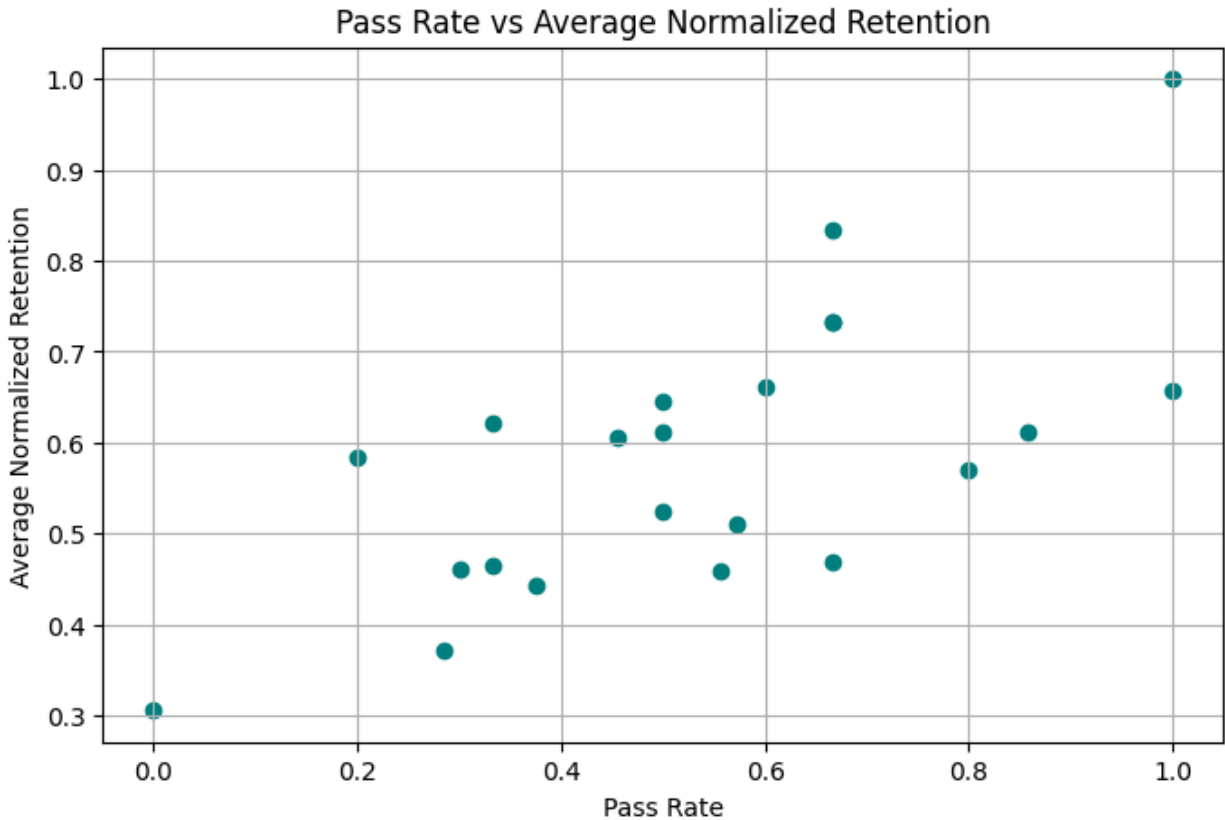
**(Graph 7: Scatter Plot - Pass Rate vs Average Normalized Retention)**

A **strong positive correlation** was found between the Pass Rate and the Average Normalized Retention.

The Pearson correlation was **r = 0.681** with a **p-value = 0.0005**, indicating a statistically significant and meaningful relationship.

This result validates the reliability of Pass Rate as a measure of student retention performance.

6.2 Student-Based Findings

### 6.2.1 Mentor Field (Quantitative vs Equal Weight) vs Student Retention
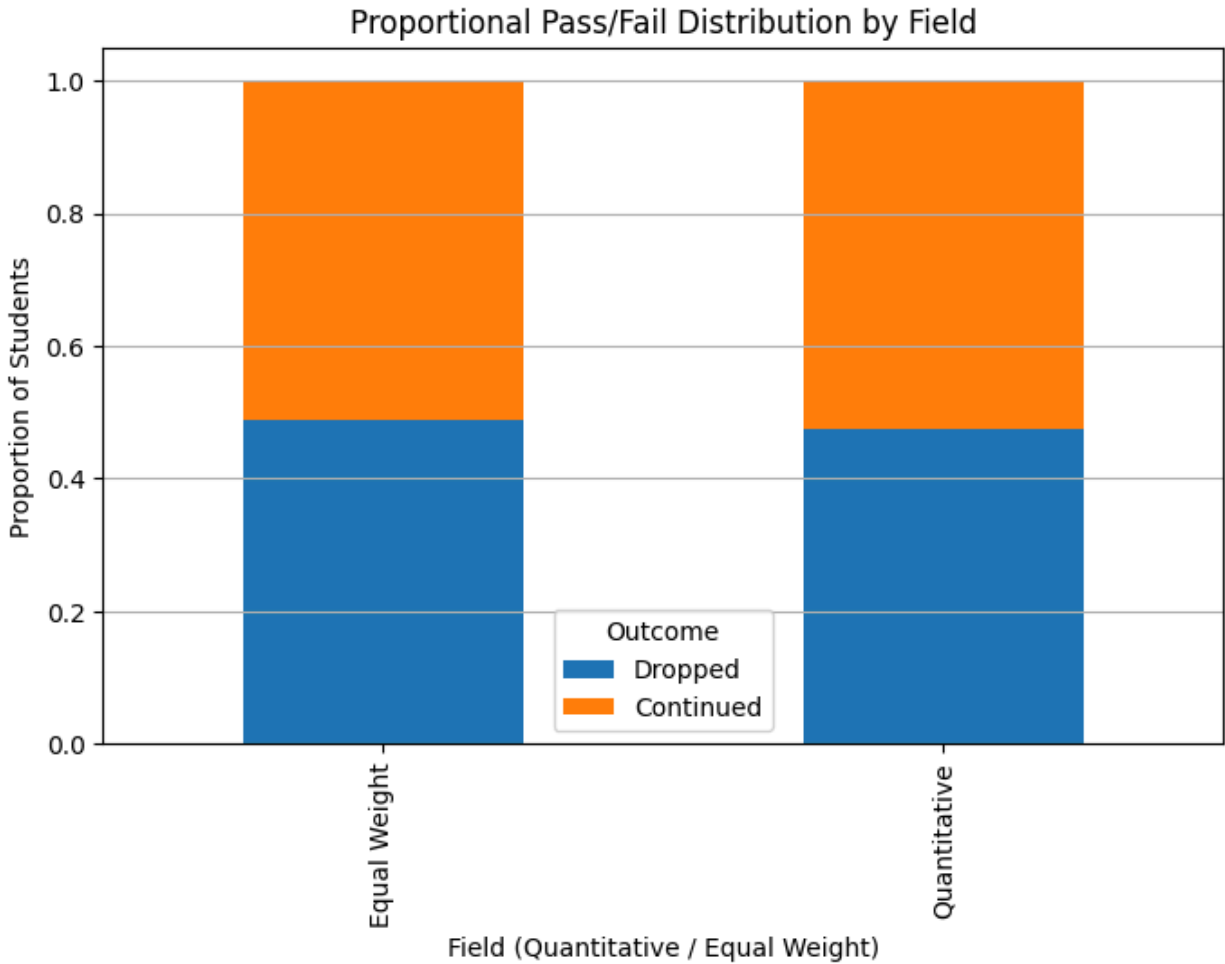
(Graph 8: Proportional Stacked Bar Plot - Field vs Pass Rate)

No significant difference was found in pass rates between students assigned to mentors from the Quantitative or Equal Weight fields.
Visual inspection of the proportional bar plot confirmed the similarity between groups, and statistical tests supported the lack of a meaningful difference.

**6.2.2 Communication Format (Video vs Voice) vs Student Retention**
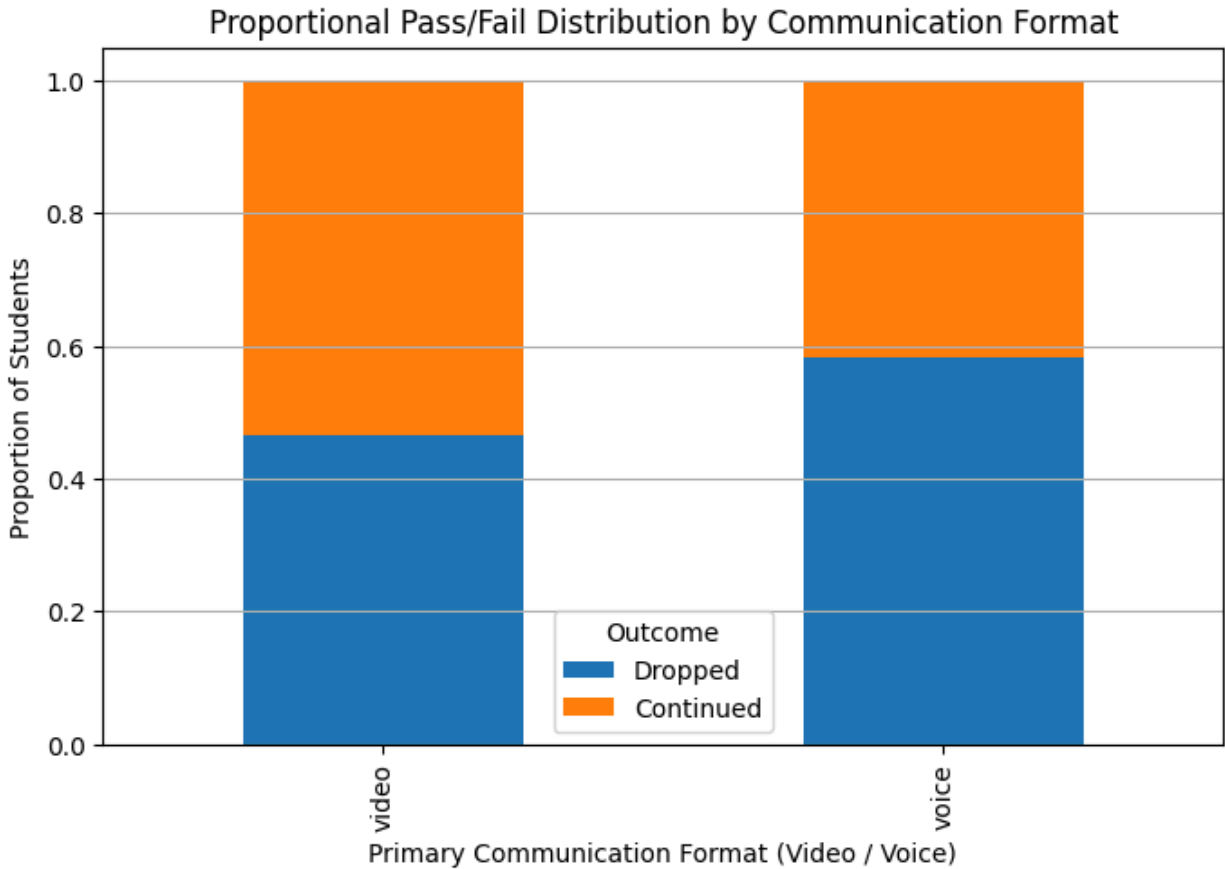
(Graph 9: Proportional Stacked Bar Plot - Communication Format vs Pass Rate)

There was no meaningful difference in student pass rates based on whether the mentor used primarily video or voice communication.
The proportional distribution between pass and fail students remained similar across both groups.

**6.2.3 Weekly Call Frequency vs Student Retention**
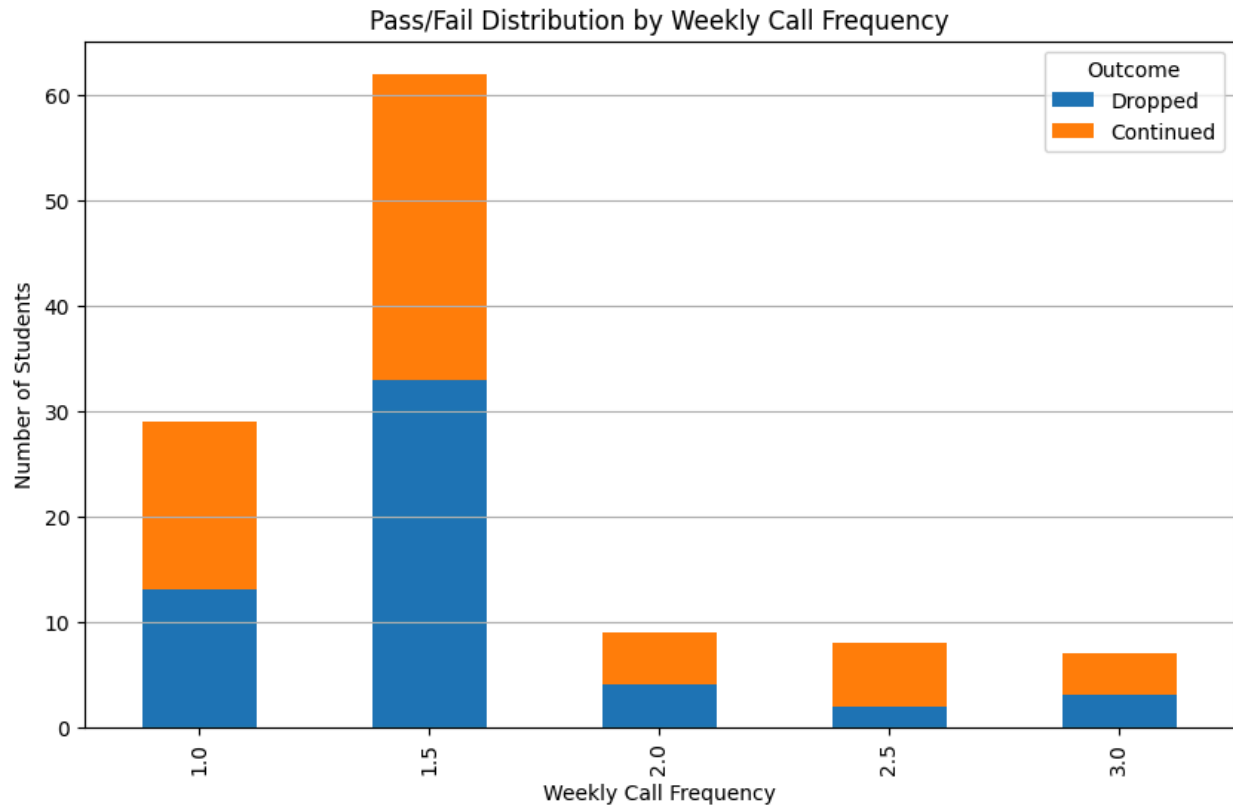
(Graph 10: Stacked Bar Plot - Weekly Call Frequency vs Pass/Fail Counts)

When analyzing the number of weekly calls, no clear trend was observed where higher call frequency led to higher student retention.
Both low and moderate call frequencies had a mixture of passing and failing students, and no strong pattern was visible.

6.2.4 Weekly Message Days vs Student Retention
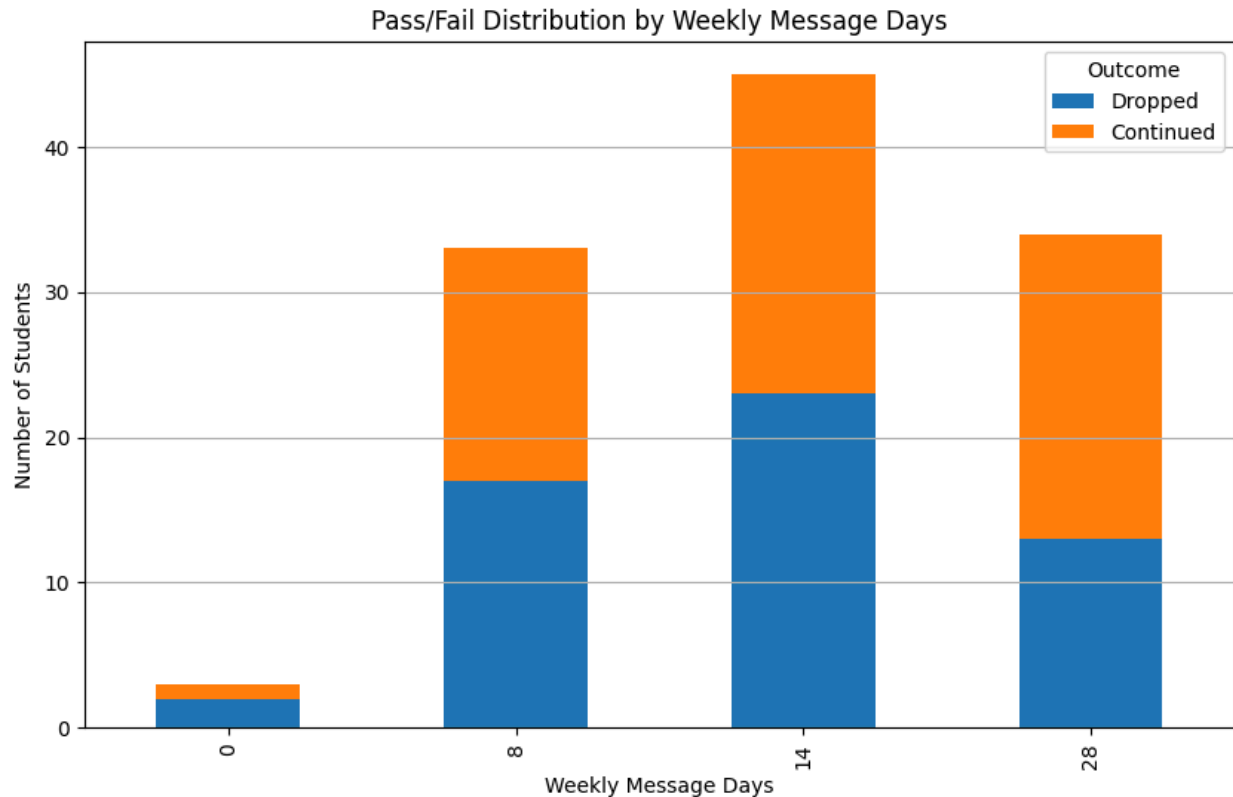
(Graph 11: Stacked Bar Plot - Weekly Message Days vs Pass/Fail Counts)

Similarly, the number of days mentors and students exchanged messages did not show a strong pattern with retention.
While students with more frequent communication tended to have slightly higher pass rates, the relationship was not strong or statistically significant.

7. General Conclusion

Based on the statistical tests conducted for RQ1 through RQ6,

- **No statistically significant relationships** were found between mentor characteristics (such as YKS ranking, communication format, call duration, call frequency, progress messaging, and communication days) and student retention after one month.
- **Thus, we fail to reject the null hypotheses** for RQ1 through RQ6.
- This suggests that **early-stage student engagement** may be influenced more strongly by **internal student factors**(e.g., motivation, study habits) rather than mentor-specific attributes.

However, for RQ7:

- A **strong and statistically significant positive correlation** was found between **Pass Rate** and **Average Normalized Retention**.
- **Thus, we reject the null hypothesis** for RQ7 and accept that **Pass Rate** is reliably associated with longer-term student retention behavior.

This result **validates the use of Pass Rate as a meaningful and reliable short-term metric** for predicting broader retention patterns within the Coachify platform.

Future studies could explore:

- A deeper investigation into **student-driven factors** affecting retention,
- Increasing the **sample size** to improve statistical power,
- Developing models that incorporate both mentor and student characteristics.

8. Additional Concept Explanation: What is Average Normalized Retention?

Average Normalized Retention is a derived metric that measures how long a student stays enrolled relative to a normalized timescale.
In this project:

- Students were tracked month by month.
- Their retention duration was scaled between 0 and 1 to create a **normalized retention score**.
- The **average** of these normalized values across all students under a mentor reflects how well that mentor retains students over time.

This metric provides a **more granular, continuous measure** compared to simple binary retention outcomes (continued/dropped after one month).
It allows for:

- Capturing subtle differences in retention behaviors,
- Validating whether early engagement indicators (like Pass Rate) reliably predict longer-term outcomes.

In this study, the **strong correlation between Pass Rate and Average Normalized Retention** confirms that Pass Rate can serve as an **early, actionable KPI** (Key Performance Indicator) for student engagement success.

9. Machine Learning-Based Prediction Attempt

To further investigate whether mentor-related factors could be used to predict early retention, a logistic regression model was applied using student-level data.

The model aimed to classify whether a student would continue to the second month (Pass = 1) or not (Pass = 0) based on the following four features:

- Mentor's Field (Quantitative or Equal Weight)
- Primary Communication Format (Video or Voice)
- Weekly Call Frequency
- Weekly Message Days

To compensate for the limited dataset (n ≈ 116), we used 10-fold cross-validation for more reliable model evaluation. The logistic regression model achieved an average accuracy of 45.15%, with individual fold results ranging from 25% to 66%.

These results indicate that, although mentor behaviors like communication type and frequency are intuitively related to student engagement, they do not hold strong predictive value for early retention. This supports the earlier findings that mentor-specific factors alone may not sufficiently explain student continuation. Other factors—such as student motivation or personal circumstances—may play a more significant role.

Thus, while the model was properly implemented and evaluated, its low performance further highlights the complexity of retention prediction and the necessity for a more holistic approach in future studies.