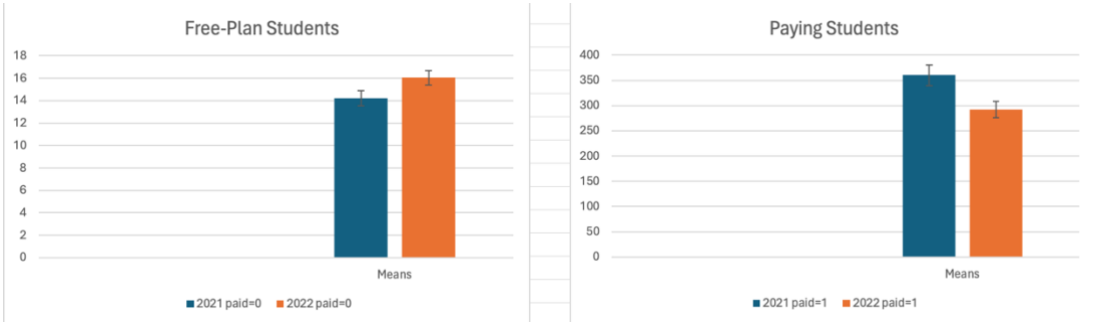
Student Engagement – Final Report

Observations - Confidence Interval

Category	Means	95% Confidence Intervals		Difference	
2021 paid=0	14.20828	13.54806	14.86851	0.66022	0.66022
2022 paid=0	16.03596	15.41480	16.65712	0.62116	0.62116

Category	Means	95% Confidence Intervals		Difference	
2021 paid=1	360.10380	339.60022	380.60739	20.50359	20.50359
2022 paid=1	292.22098	276.53903	307.90293	15.68195	15.68195



For students on the free-plan, there was a noticeable increase in engagement from Q2 2021 to Q2 2022. The confidence interval for Q2 2022 (15.41 to 16.66 minutes) was slightly higher than that for Q2 2021 (13.55 to 14.87 minutes). Students with paid memberships showed significantly higher engagement compared to those on the free plan. This is evident from the confidence intervals for Q2 2021, where non-subscribers had an interval of 13.55 – 14.87 minutes, while subscribers had a much higher interval of 339.60 – 380.61 minutes. However, among paid students, there was a decrease in engagement from Q2 2021 to Q2 2022. The confidence interval for Q2 2022 (276.54 to 307.90 minutes) was lower than that for Q2 2022 (339.60 to 380.61 minutes).

It is essential to recognize that these interpretations are based on observed confidence intervals, and establishing actual cause-effect relationships requires more thorough investigations. For example, higher engagement levels among paid students do not necessarily imply that having a paid subscription causes increased viewing. It could be that students who are already more engaged are more inclined to purchase a subscription. Similarly, the observed decrease in

engagement among paid students from Q2 2021 to Q2 2022 could be influenced by various factors that need to be examined individually.

Hypothesis Testing

For conducting the hypothesis testing to check whether that the engagement in Q2 2021 is higher than Q2 2022 between the two student types, we need to use two-sample t-test which assuming equal or unequal variances between two groups.

It essential we run a F-test to check the equality of the variances first, before we determine which testing method we are going to use.

Null Hypothesis: The variances of the two population variances are equal.

Alternative Hypothesis: The variances of the two population are not equal.

Free-Plan Students

F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	14.20828409	16.03596096
Variance	599.1057104	602.0177856
Observations	5280	5994
df	5279	5993
F	0.995162809	
P(F<=f) one-tail	0.428157921	
F Critical one-tail	0.957004439	

Conclusion: Variances can be assumed equal

So we need to use the T-statistic for Population Variances Unknown but Assumed to be Equal

Paying Students

F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	360.103801	292.220979
Variance	249616.2086	176555.8897
Observations	2281	2758
df	2280	2757
F	1.413808449	
P(F<=f) one-tail	2.04E-18	
F Critical one-tail	1.067950666	

Conclusion: Variances cannot be assumed equal

So we need to use the T-statistic for Population Variances Unknown and Assumed to be Unequal

The p-value for free-plan students is 0.428, indicating that we failed reject the null hypothesis that the variances of the two population variances are equal (variances can be assumed equal), so we are using two-sample t-test assuming equal variances. As for paying students, the p-value is well below the 0.05 threshold, thus we reject the null hypothesis, the variances of the two population are not equal, we are using two-sample t-test assuming unequal variances.

Null Hypothesis: The engagement (minutes watched) in Q2 2021 is higher than or equal to the one in Q2 2022. That is, μ1 ≥ μ2.

Alternative Hypothesis: The engagement (minutes watched) in Q2 2021 is lower than the one in Q2 2022. That is μ1 ≤ μ2.

Free-Plan Students	Count	Mean	Pooled Variance	T-Statistics	Critical Value
Q2 2021	5280	14.20828	600.6539775	-3.951150319	1.645
Q2 2022	5994	16.03596			

Paying Students	Count	Mean	T-Statistics	Critical Value
Q2 2021	2281	360.10380	5.15435529	1.645
Q2 2022	2758	292.22098		

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	14.208284	16.035961
Variance	599.10571	602.01779
Observations	5280	5994
Pooled Variance	600.6539775	
Hypothesized Mean Difference	0	
df	11272	
t Stat	-3.951150319	
P(T<=t) one-tail	3.91253E-05	
t Critical one-tail	1.64498882	
P(T<=t) two-tail	7.82506E-05	
t Critical two-tail	1.960174464	

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	360.1038	292.220979
Variance	249616.21	176555.89
Observations	2281	2758
Hypothesized Mean Difference		0
df		4464
t Stat		5.15435529
P(T<=t) one-tail		1.3275E-07
t Critical one-tail		1.645195044
P(T<=t) two-tail		2.65501E-07
t Critical two-tail		1.960495549

For free-plan students, comparing the one-tailed t-statistics to the critical t-value, a t-statistic of -3.951, which is less than the critical value of 1.645, leads to rejecting the null hypothesis. This negative t-statistic implies that the average minutes watched by free-plan students in Q2 2021 (μ1) is significantly lower than the average minutes watched by free-plan students in Q2 2022 (μ2). This contradicts the null hypothesis, prompting its rejection. However, it's important to note that rejecting the null hypothesis does not prove the alternative hypothesis; it merely indicates that there is sufficient evidence to oppose the null hypothesis.

For paying students, the t-statistic to the critical t-value, a t-statistic of 5.161, which is greater than the critical value of 1.645, means you would fail to reject the null hypothesis. This indicates there isn't sufficient evidence to conclude that the average minutes watched by paying students in Q2 2021 (μ1) is lower than the average minutes watched in Q2 2022 (μ2). Thus, the data supports the null hypothesis that μ1 is larger than or equal to μ2.

Regarding errors in hypothesis testing:

- **Type I Error (False Positive)**: This occurs when you reject a true null hypothesis. In this context, it would mean incorrectly concluding that engagement in 2022 is higher when it is not. The consequence for the company might be over-investing in certain features or becoming complacent about improving existing features.
- **Type II Error (False Negative)**: This occurs when you fail to reject a false null hypothesis. In our case, it would mean incorrectly concluding that engagement in 2022 is

not higher when it is. The impact on the company might include missing out on recognizing successful features or failing to identify areas that need enhancement.

The cost of each type of error depends on the consequences of these incorrect conclusions. Overestimating engagement could lead to misallocated resources, while underestimating it might result in missed opportunity for improvement and growth.

Dependency Analysis

In this part, we will determine if watching a content in Q2 2021 and Q2 2022 are dependent or independent events, to come up with marketing strategies accordingly.

First let's define two events we are interested in:

- Event A: a student watched a lecture in Q2 2021.
- Event B: a student watched a lecture in Q2 2022.

Two events are considered independent if the occurrence of one does not affect the occurrence of the other. In terms of probability, this is represented as:

$P(A \cap B) = P(A) \times P(B)$

Where:

- $P(A \cap B)$ is the probability that a student watched a lecture in both Q2 2021 and Q2 2022.
- P(A) is the probability that a student watched a lecture in Q2 2021.
- P(B) is the probability that a student watched a lecture in Q2 2022.

To determine these probabilities, we will use the given data that we queried using SQL, and the following formulas:

1. $P(A) = \frac{Number\ of\ students\ who\ watched\ a\ lecture\ in\ Q2\ 2021}{Total\ number\ of\ students\ who\ have\ watched\ a\ lecture} = \frac{7{,}639}{15{,}840}$
2. $P(B) = \frac{Number\ of\ students\ who\ watched\ a\ lecture\ in\ Q2\ 2022}{Total\ number\ of\ students\ who\ have\ watched\ a\ lecture} = \frac{8{,}841}{15{,}840}$
3. $P(A \cap B) = \frac{Number\ of\ students\ who\ watched\ a\ lecture\ in\ both\ periods}{Total\ number\ of\ students\ who\ have\ watched\ a\ lecture} = \frac{640}{15{,}840}$

Probabilities	Values
P(A)	0.482260101
P(B)	0.558143939
P(A) × P(B)	0.269170553
P(A ∩ B)	0.04040404

By calculating these probabilities, $P(A \cap B) \neq P(A) \times P(B)$, thus we conclude that the two events (A and B) are dependent, which means that the occurrence of one event has some influence on the occurrence of the other. In addition, $P(A \cap B)$ (0.040) is smaller than $P(A) \times P(B)$ (0.269), it suggests that those who watched a lecture in Q2 2021 were less likely to watch a lecture in Q2 2022 than anticipated if the two events were independent. This result is anticipated, a student who achieved their goals with the program in 2021 may not be as inclined to return in 2022 with the same level of engagement. This phenomenon, often termed as 'good churn,' occurs when users leave after successfully utilizing the platform to meet their objectives.

Despite this, we actively run marketing campaigns to re-engage students who have been registered on the platform for some time but have not been active recently. The purpose of these campaigns is to reintroduce these students to the platform's new features and the expanded course library. By continuously adding fresh and relevant content, we believe that these students can still find value in the program, even after a period of inactivity. Our goal is to demonstrate that the platform evolves with their learning needs, encouraging them to return and benefit from the latest offerings.

Probability Analysis

To confirm the result that we deducted from previous part, we need to prove the opposite direction of the theory to be true as well. To achieve this, we need to determine the probability of Event A occurring given that Event B has occurred, this is represented as $P(A|B)$. This can be solved using Bayes' Rule:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Where:

- P(B|A) is the probability that a student watched a lecture in Q2 2022, given that they watched a lecture in Q2 2021. (This is the same as $P(A \cap B)$ that we calculated previously).

Using the available data, we can calculate these probabilities:

1. $P(A) = \frac{Number\ of\ students\ who\ watched\ a\ lecture\ in\ Q2\ 2021}{Total\ number\ of\ students\ who\ have\ watched\ a\ lecture} = \frac{7,639}{15,840}$
2. $P(B) = \frac{Number\ of\ students\ who\ watched\ a\ lecture\ in\ Q2\ 2022}{Total\ number\ of\ students\ who\ have\ watched\ a\ lecture} = \frac{8,841}{15,840}$
3. $P(B|A) = \frac{Number\ of\ students\ who\ watched\ a\ lecture\ in\ both\ periods}{Total\ Number\ of\ students\ who\ watched\ a\ lecture\ in\ Q2\ 2021} = \frac{640}{7,639}$

Substituting these values into Bayes' Rule gives us:

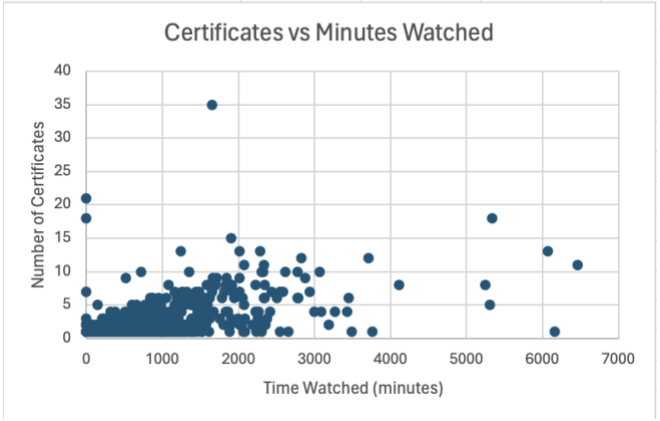
$$P(A|B) = \frac{640}{7,639} \times \frac{7,639}{15,840} \div \frac{8,841}{15,840} \approx 7\%$$

This means that among the students who watched a lecture in Q2 2022, approximately 7% of them had also watched a content in Q2 2021. This result indicates that students who watched a lecture in Q2 2022 were not likely to have watched one in the same quarter of the previous year, which proves our previous conclusion.

Correlation Analysis

The company also wants to assess the effectiveness of their contents, identifying a correlation between minutes watched and number of certificates issued helps in understanding which courses or content format are more engaging and effective. This information can be used to refine and improve the quality of course materials. If high engagement leads to more certificates, it suggests that the courses are achieving their educational objectives.

Correlation Coefficient
0.512615114



The correlation calculated is approximately 0.513, indicates a moderate positive relationship, it does not necessarily mean that students will receive more certificates if they spend more time watching contents. Let's build a linear regression model using only minutes watched to predict the number of certificates for deeper interpretations.

After training the linear regression model, the value of the slop and y-intercept are represented as follows:

```
# get the interception and the coefficient of the model
model.intercept_, model.coef_
```

```
(1.2906568291230498, array([0.00146215]))
```

The linear equation of the model can then be expressed as:

$$y = 0.00146215x + 1.296568$$

```
# get the coefficient of determination r2
model.score(X_train,y_train)
```

```
0.24026809272839034
```

The calculated R^2 is approximately 0.24. This means that around 24% of the variability in the target variable (number of certificated issued) can be explained by the predicting variable (minutes watched). Consequently, this model leaves 76% of the variability unaccounted for:

An R^2 value of 0.24 indicates a moderate level of explanatory power. While it is a meaningful result, it also highlights the presence of other influential factors. Let's delve into some of these factors in detail:

1. **Course Length and Structure**: Students might complete multiple short courses, each resulting in a certificate, versus a single, lengthy course. For example, two short courses may yield two certificates, while one long course , even if it requires similar overall study time, results in just one certificate.
2. **Prior Knowledge**: Some students may possess prior knowledge of the subject matter, allowing them to pass exams with minimal additional study. These students might watch fewer minutes of selected course contents but still achieve a high number of certificates. Such students are primarily motivated by obtaining the certificates as proof if proficiency rather than learning new material, which decouples the relationship between minutes watched and certificates issued.
3. **Engagement and Learning Method**: Students have varied learning styles and engagement levels. Some may prefer in-depth learning and therefore watch more content without necessarily aiming for multiple certificates, while others might focus on quickly completing courses to earn certificates.

While the number of minutes watched is a significant predictor of the number of certificates issued, it accounts for only 24% of the variability in the outcome. This indicates that other factors play crucial roles in determining certificate issuance. Therefore, while the model provides valuable insights, it is important to incorporate additional variables to create a more comprehensive and accurate model. Relying solely on minutes watched may lead to an incomplete understanding of the factors driving certificate completion.

Business Recommendations

Based on the analyses we conducted in this project, here are some recommendations for the platform:

- **Improving Course Content and Structure**: If certain lengths of content are correlated with higher completion rates, the company can optimize course durations to maximize student success.
- **Personalized Learning Paths**: Tailored Recommendations: Knowing the correlation allows the platform to provide personalized recommendations to students. For example, if students who watch more minutes tend to earn more certificates, the platform can encourage less engaged students to increase their watching time through personalized messages or suggestions.
- **Intervention Strategies**: For students showing low engagement, the platform can implement targeted interventions, such as reminders, additional resources, or motivational messages, to boost their participation and course completion rates.
- **Targeted Marketing**: By understanding engagement patterns, the company can design targeted marketing campaigns to attract new users and retain existing ones. Highlighting the success stories of highly engaged students who earn many certificates can be an effective marketing strategy.
- **Retention Efforts**: Student who achieved their goals are very likely to churn in the following year. The company can focus on retention strategies that keep students engaged, such as regular updates, new course offerings, and community-building activities.
- **Spotting Gaps**: A weak or non-existent correlation might indicate areas where the platform needs improvement. For instance, it might suggest that students are watching content but not finding it valuable enough to complete courses and earn certificates.