Management Development Institute Murshidabad

Murshidabad – 742235

Project on

"Customer engagement Analysis of
365 Financial Analyst"

With:

Ivan Kitov

Submitted By:

Ankesh Kr. Saurabh

Table of Content

Case Description

In 2022, there were high expectations for the growth of the 365 company and increased student engagement based on the introduction of new website platform features. Some of these features included an XP system that enabled students to track their progress, level up, and earn rewards by completing various learning objectives. The platform also offered in-app coins that could be exchanged for special awards, a leaderboard where students could compete for top positions in different divisions, earning weekly rewards and advancing up the ladder, and streaks to motivate students to maintain consistent learning habits. Additionally, the company expanded its course library, covering a broader range of topics to provide its students with a richer set of skills and attract a larger audience. These enhancements were anticipated to positively impact the student experience, create an effective strategy for customer engagement, and contribute to the company's success in the coming year. With this Customer Engagement Analysis in Excel project, you must analyze whether the new additions to the platform have increased student engagement.

Project files

During this Customer Engagement Analysis in Excel project, we have to analyze a dataset from the 365 company. Nevertheless, we have to use this dataset to accurately represent the company's operations—providing a realistic and relevant context for your analysis.

In addition, consider the following information about the column values while working with the data:

1. **Student_id** – the unique identifier for each student in the dataset. The field contains IDs for students who used the 365 Data Science platform with free or paid accounts in **Q4 2021** (October 1, 2021 – December 31, 2021, both included) and **Q4 2022** (October 1, 2022 – December 31, 2022, both included).

1. **Student_country** – identifies the country of each student. The field provides information about students' geographic location and can help analyze regional differences or conduct country-specific analyses.
2. **Paid** – indicates whether a student had a paid account during the specified period. It is a binary variable, where '1' represents a paid account and '0' represents a free or unpaid

account. It helps differentiate between students who have access to additional features or content through a paid subscription.

3. **minutes_watched_21** – represents the student's engagement level, as expressed by the number of minutes a student has watched in Q4 2021.
4. **minutes_watched_22** – denotes the student's engagement level, as expressed by the number of minutes a student has watched in Q4 2022.

Analysis and Interpretation

Part 1

Task 1

You can find the solution to this problem in the Descriptive Statistics – tasks 1-2.xlsx file:

- *Sheet: Task 1*

Interpreting the Results

- **Paid-plan Students**

	minutes_watched_21	minutes_watched_22
Mean	33.80	273.02
Median	26.33	40.28
Standard Deviation	28.21	854.58

  - **Mean**: Among students who watched between 1 and 100 minutes in 2021, the average minutes watched by paid-plan students increased significantly from Q4 2021 to Q4 2022, from approximately 33.80 minutes to about 273.02 minutes. This suggests a substantial increase in engagement among this group of initially low-engagement-paid-plan students.
  - **Median**: The median minutes these low-engagement-paid-plan students watched increased from Q4 2021 to Q4 2022, from 26.33 minutes to 40.28 minutes. While this increase is not as dramatic as the increase in the mean, it indicates that the typical student in this group (i.e., the student in the middle of the distribution) also increased their engagement. This suggests that the increase

in engagement was more widespread among paid-plan students and not solely driven by a few outliers.

- **Standard Deviation**: The standard deviation for these low-engagement-paid-plan students increased substantially from 28.21 minutes in Q4 2021 to 854.58 minutes in Q4 2022. This indicates a much larger variability in the minutes watched by these students in Q4 2022 compared to Q4 2021. This could be due to a broader range of engagement levels among the students in Q4 2022, with some students watching very little content and others watching a lot of content.

These results suggest that paid-plan students who were initially low-engagement in 2021 significantly increased their engagement in 2022. But the increased standard deviation indicates a broader range of engagement levels among these students in 2022. Understanding the reasons behind this variability could provide valuable insights for further boosting engagement. For instance, the factors that motivated the students who significantly increased their engagement might be leveraged to encourage increased engagement among other students.

- **Free-Plan Students**

	minutes_watched_21	minutes_watched_22
Mean	25.39	117.64
Median	14.17	11.83
Standard Deviation	26.23	468.93

-
  - **Mean:** Among students who watched between 1 and 100 minutes in 2021, the average minutes watched by free-plan students increased from about 25.39 minutes in Q4 2021 to about 117.64 minutes in Q4 2022. This suggests that overall engagement among these initially low-engagement-free-plan students increased during this period. But the extent of this increase is less than what was observed for similar low-engagement-paid-plan students, suggesting that while these free-plan students are watching more content, they're still not as engaged as the equivalent group of paid-plan students.
  - **Median:** Interestingly, the median minutes watched by these low-engagement-free-plan students decreased from Q4 2021 to Q4 2022, from 14.17 minutes to

11.83 minutes. This indicates that engagement decreased for the typical student in this group (i.e., the student in the middle of the distribution). The increase in the mean might be driven by a small number of free-plan students who significantly increased their engagement in Q4 2022, while the majority did not increase their engagement or even reduced it.

- **Standard Deviation:** The standard deviation for the low-engagement-free-plan students increased from 26.23 minutes in Q4 2021 to 468.93 minutes in Q4 2022. This indicates a more significant variability in the minutes watched by these students in Q4 2022 compared to Q4 2021. The behavior of these students then became more diverse in Q4 2022, with some watching a lot of content and others watching very little.

These results suggest a complex picture for the initially low-engagement-free-plan students. While the mean minutes watched increased—signifying an increase in overall engagement—the median minutes watched decreased, indicating that the typical student in this group did not increase their engagement. This discrepancy and the increased standard deviation suggest that a small number of students within this group might significantly increase their engagement while the majority did not. This might imply the need for targeted strategies to boost engagement among the broader population of initially low-engagement-free-plan students.

- **Paid vs Free-Plan Students**

On average, low-engagement-paid students initially increased their watching time more significantly than the free-plan students from Q4 2021 to Q4 2022. This could suggest that paid-plan students find more value in the platform, possibly due to premium features or content that are available to them. In contrast, the median watch time decreased for free-plan students, suggesting that the typical free-plan student in this group did not increase their engagement. This discrepancy might indicate that the strategies or features designed to increase engagement are more effective for paid-plan students. It could also suggest that the monetary investment leads to increased usage due to a desire to get their money's worth.

Based on the findings, the platform is more successful in increasing engagement among students who make a monetary investment (i.e., paid-plan students). But the increased variability, especially among paid-plan students, indicates that there are likely differences in how individual students are responding to the platform's offerings. Therefore, personalized

approaches might be beneficial in boosting engagement, and further analysis could help understand the factors that drive increased engagement among paid- and free-plan students.

Task 2

You can find the solution to this problem in the Descriptive Statistics – tasks 1-2.xlsx file:

- *Sheet: Task 2*

Skewness is a fundamental measure of probability distribution asymmetry in a dataset. It reveals whether the observations are concentrated more on one side of the distribution. This metric helps us understand how the data deviates from a normal distribution and provides insights into its underlying structure. A positive skewness value (higher than 0) indicates a right-skewed distribution, while a negative skewness value (lower than 0) points to a left-skewed distribution. A symmetrical distribution has a skewness value of 0, indicating a balanced data spread around the mean.

For **paid-plan students**, the skewness increased from 0.63 in Q4 2021 to 7.07 in Q4 2022.

Paid Students

student_id	paid	minutes_watched_21	minutes_watched_22
16979	1	13.32	260.72
207114	1	40.12	387.98
156680	1	17.57	128.78
149601	1	42.95	7417.4
251499	1	4.92	10.47
179664	1	45.07	628.05
145813	1	16.98	949.9

Skewness	0.63	7.07

The skewness for **free-plan students** increased from 1.17 in Q4 2021 to 15.06 in Q4 2022, indicating positive skewness.

Free Students

student_id	paid	minutes_watched_21	minutes_watched_22
238865	0	1.43	157.28
247592	0	3.1	0.1
195373	0	8.45	12.57
229324	0	44.87	1
198040	0	61.88	0.23
14672	0	55.05	114.17
182954	0	3.13	0.07

Skewness	1.17	15.06

Positive skew (right-skew) occurs when the data is not symmetrical around the mean, forming a long tail on its right side. This signifies that most of the distribution's observations are concentrated to the left of the peak. Positive skewness can have several implications.

The mean is larger than the median in a right-skewed distribution because the distribution tail pulls the mean to the right. This observation is confirmed by the mean and median values in the two years. An increasing skewness suggests that more students watch significantly more content than most over time, pulling the mean upwards.

In both cases, the mean is higher than the median (33.80 > 26.33 in 2021 and 273.02 > 40.28 in 2022).

Paid Students

student_id	paid	minutes_watched_21	minutes_watched_22
16979	1	13.32	260.72
207114	1	40.12	387.98
156680	1	17.57	128.78
149601	1	42.95	7417.4
251499	1	4.92	10.47
179664	1	45.07	628.05

	minutes_watched_21	minutes_watched_22
Mean	33.80	273.02
Median	26.33	40.28

As a result, the mean is no longer a good central tendency indicator, and it cannot accurately reflect the typical value of the dataset. Note that skewness tells us the direction of outliers but doesn't indicate the number that occurs.

Kurtosis measures the degree of tailedness—the weight of the tails relative to the rest of the distribution. In other words, it shows how much of the data is in the tails compared to the center. Located farthest from the center, the tails represent the regions where data points are more dispersed—suggesting the presence of more extreme values. If a distribution is heavy-tailed—i.e., more data in the tails—it exhibits high kurtosis. Meanwhile, a low kurtosis occurs when the data is more evenly distributed between the tails and the center or the distribution is light-tailed.

For paid-plan students, the kurtosis increased from -0.85 in Q4 2021 to 58.48 in Q4 2022.

Paid Students

student_id	paid	minutes_watched_21	minutes_watched_22
16979	1	13.32	260.72
207114	1	40.12	387.98
156680	1	17.57	128.78
149601	1	42.95	7417.4
251499	1	4.92	10.47

	minutes_watched_21	minutes_watched_22
Kurtosis	-0.85	58.48

The kurtosis increased from free-plan students—from 0.36 in Q4 2021 to 315.76 in Q4 2022.

Free Students

student_id	paid	minutes_watched_21	minutes_watched_22
238865	0	1.43	157.28
247592	0	3.1	0.1
195373	0	8.45	12.57
229324	0	44.87	1
198040	0	61.88	0.23

	minutes_watched_21	minutes_watched_22
Kurtosis	0.36	315.76

Kurtosis values greater than 0 indicate that the data has heavier tails and a sharper peak than the normal distribution (leptokurtic). A leptokurtic distribution has a high positive kurtosis, suggesting that it's very peaked and has a relatively large number of outliers. This type has a higher frequency of extreme values or outliers. The increase in kurtosis over time suggests more extreme cases in the data in Q4 2022 than in Q4 2021, particularly for free-plan students.

Overall, the increasing skewness and kurtosis for both groups from Q4 2021 to Q4 2022 suggest a growing number of students watching significantly more content than the majority. This is especially true for free-plan students with a higher skewness and kurtosis in Q4 2022 than paid-plan students.
Descriptive Statistics - tasks 1-2.xlsx

Part 2: Confidence Intervals

Task 3

You can find the solution to this problem in the Confidence Intervals – task 3xlsx file:

- *Sheet: Task 3*

Comparing the four groups, you can observe the following:

- **Paid-Plan Students**:

Paid Students

minutes_watched_21	minutes_watched_22
2973.67	4110.17
2939.48	4099.42
2860.78	4085.2
2853.73	4064.35
2830.2	4024.33
2809.67	3948.85
2803.17	3909.85
2797.55	3908.57
2782.08	3879.82
2741.9	3866.8
2703.03	3828.88
2699.63	3776.67

	minutes_watched_21
Mean	332.50
Standard Deviation	485.86
Standard Error	8.29
95% CI, $z_{0.025}$	1.96

Z	CI low	CI high
95%	316.25	348.76

	minutes_watched_22
Mean	368.35
Standard Deviation	596.41
Standard Error	8.35
95% CI, $z_{0.025}$	1.96

Z	CI low	CI high
95%	351.99	384.72

For paid-plan students, there's an increase in engagement from Q4 2021 to Q4 2022. The confidence interval for the average minutes watched by paid-plan students increased from Q4

2021 (316.25 to 348.76 minutes) to Q4 2022 (351.91 to 384.72 minutes). This suggests that we can be 95% confident that the true average minutes watched by all paid-plan students in the population increased from Q4 2021 to Q4 2022.

- **Free-Plan Students**:

Free Students

minutes_watched_21	minutes_watched_22
4716.68	6338.07
4670.7	6280.12
4622.35	6250.32
4617.75	6208.8
4599.53	6204.55
4581.45	6099.35
4568.73	6091.17
4564.67	6073.37
4481.85	6072.77
4464.67	6071.8
4439.47	6068.17
4388.53	6043.12

	minutes_watched_21
Mean	133.93
Standard Deviation	367.26
Standard Error	2.05
95% CI, $z_{0.025}$	1.96

Z	CI low	CI high
95%	129.92	137.95

	minutes_watched_22
Mean	69.15
Standard Deviation	255.62
Standard Error	0.74
95% CI, $z_{0.025}$	1.96

Z	CI low	CI high
95%	67.71	70.59

Among free-plan students, there's a decrease in engagement from Q4 2021 to Q4 2022. The confidence interval for the average minutes watched decreased from Q4 2021 (129.92 to 137.95 minutes) to Q4 2022 (67.71 to 70.59 minutes). We then can be 95% confident that the true average minutes watched by all free-plan students in the population decreased from Q4 2021 to Q4 2022.

- **Comparison between Paid- and Free-Plan Students (Q4 2022).**

Students with a paid-plan subscription watch substantially more than those without. The confidence interval for the average minutes watched in Q4 2022 was 61.71 to 70.59 minutes for free-plan students and 351.99 to 384.72 minutes for paid-plan students. We then can be 95% confident that paid-plan students watched significantly more minutes than free-plan students in Q4 2022. This aligns with the expectation that paid-plan students who have invested in the platform tend to be more engaged than free-plan users.

Please note that these are just interpretations based on the confidence intervals. Actual cause-effect relationships must be examined further to understand the causes behind these engagement changes.

The fact that paid-plan subscribers watch more doesn't necessarily mean that having a paid-plan subscription encourages them to watch more. For example, the higher engagement among paid-plan students may be due to the additional features or content available or because more engaged students are more likely to choose a paid-plan subscription.

Similarly, the decrease in engagement among free-plan students could be due to various factors, such as changes in the platform, competition from other platforms, or changes in the user base. Confidence Intervals - task 3.xlsx

Part 3: Hypothesis Testing

Task 4

You can find the solution to this problem in the Hypothesis Testing – task 4.xlsx file:

- *Sheet: Task 4*
- *Sheet: Test for Variances*

First, you must perform a two-sample f-test for variances to prove that assumption of unequal variances between the samples for free- and paid-plan subscribers:

Two-Sample F-Test for Variances

Paid Students

	minutes_watched_21	*minutes_watched_22*
Mean	332.502508	368.3547139
Variance	236063.3116	355699.1148
Observations	3433	5104
df	3432	5103
F	0.663660104	
P(F<=f) one-tail	0	
F Critical one-tail	0.949796198	

Free Students

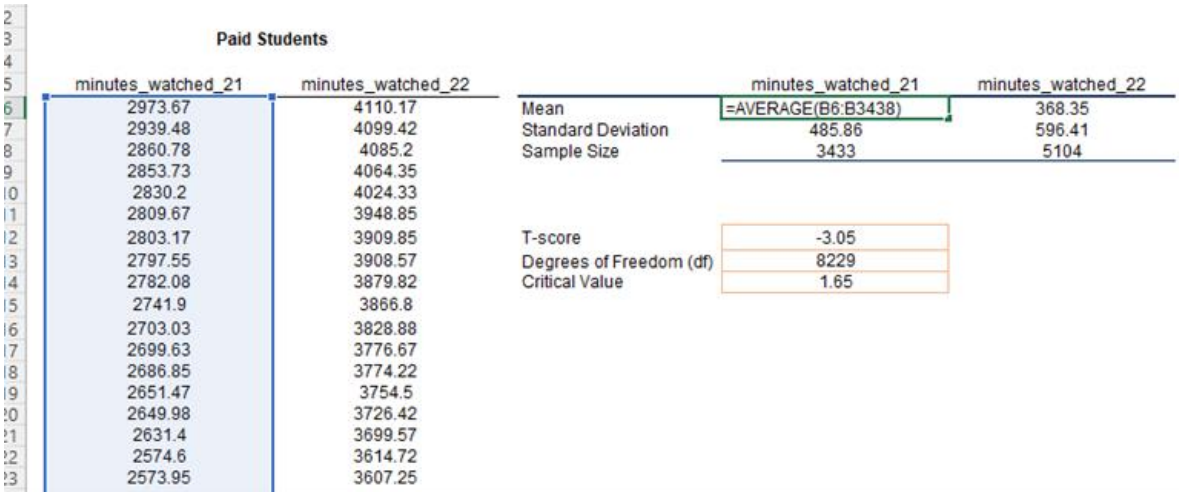
	minutes_watched_21	*minutes_watched_22*
Mean	133.9333129	69.14765544
Variance	134881.7038	65343.34428
Observations	32171	120658
df	32170	120657
F	2.06419958	
P(F<=f) one-tail	0	
F Critical one-tail	1.014667161	

The p-value indicates the probability of obtaining the observed f-value if the null hypothesis (equal variances) were true. The sample variances are not identical since the p-value in both cases is 0.

Next, we use a left-tailed t-test assuming unequal variances for paying and free-plan students.

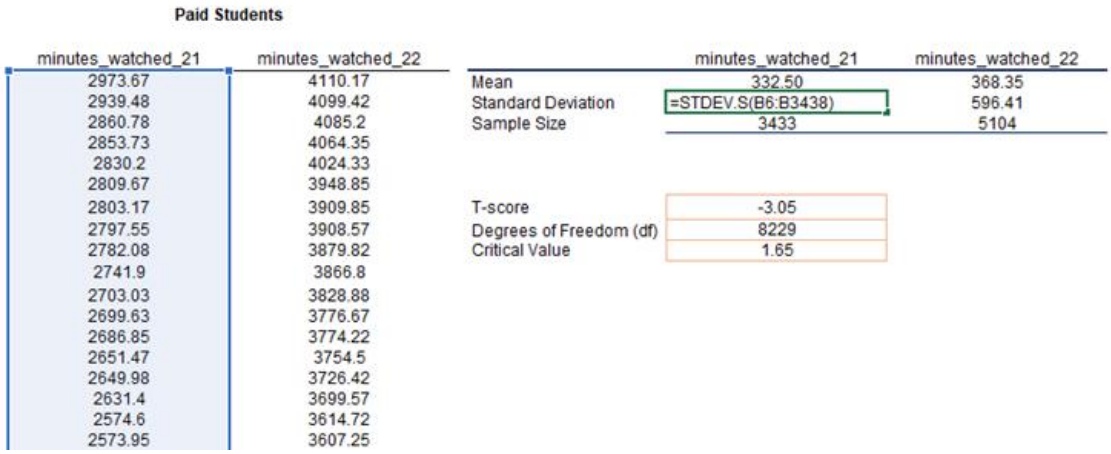
Paid-Plan Students

Calculate the mean, standard deviation, and sample size for paid-plan students:



	Paid Students	
	minutes_watched_21	minutes_watched_22
6	2973.67	4110.17
7	2939.48	4099.42
8	2860.78	4085.2
9	2853.73	4064.35
10	2830.2	4024.33
11	2809.67	3948.85
12	2803.17	3909.85
13	2797.55	3908.57
14	2782.08	3879.82
15	2741.9	3866.8
16	2703.03	3828.88
17	2699.63	3776.67
18	2686.85	3774.22
19	2651.47	3754.5
20	2649.98	3726.42
21	2631.4	3699.57
22	2574.6	3614.72
23	2573.95	3607.25

	minutes_watched_21	minutes_watched_22
Mean	=AVERAGE(B6:B3438)	368.35
Standard Deviation	485.86	596.41
Sample Size	3433	5104
T-score	-3.05	
Degrees of Freedom (df)	8229	
Critical Value	1.65	



Paid Students	
minutes_watched_21	minutes_watched_22
2973.67	4110.17
2939.48	4099.42
2860.78	4085.2
2853.73	4064.35
2830.2	4024.33
2809.67	3948.85
2803.17	3909.85
2797.55	3908.57
2782.08	3879.82
2741.9	3866.8
2703.03	3828.88
2699.63	3776.67
2686.85	3774.22
2651.47	3754.5
2649.98	3726.42
2631.4	3699.57
2574.6	3614.72
2573.95	3607.25

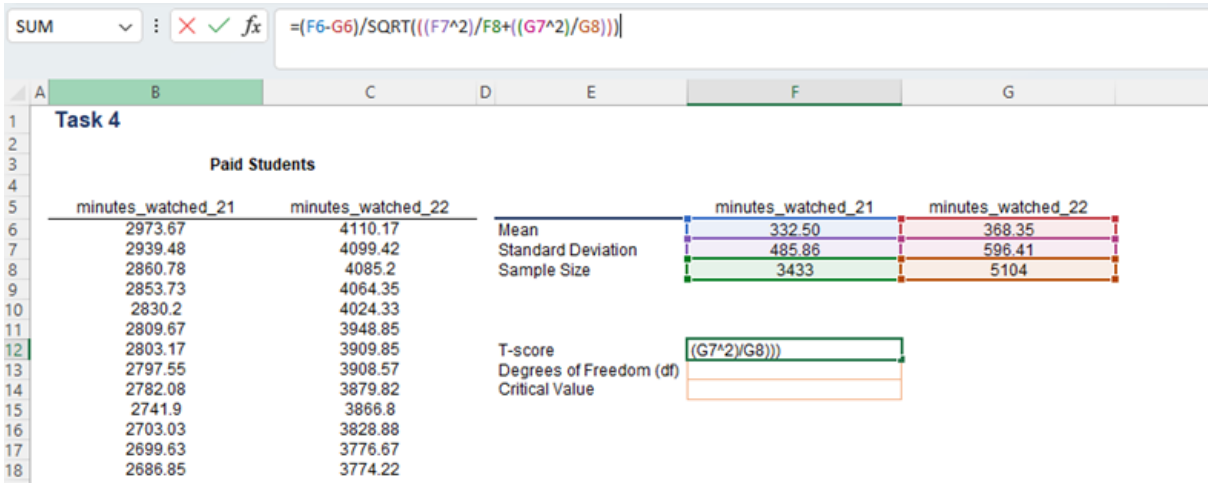
	minutes_watched_21	minutes_watched_22
Mean	332.50	368.35
Standard Deviation	=STDEV.S(B6:B3438)	596.41
Sample Size	3433	5104
T-score	-3.05	
Degrees of Freedom (df)	8229	
Critical Value	1.65	

Paid Students

minutes_watched_21	minutes_watched_22
2973.67	4110.17
2939.48	4099.42
2860.78	4085.2
2853.73	4064.35
2830.2	4024.33
2809.67	3948.85
2803.17	3909.85
2797.55	3908.57
2782.08	3879.82
2741.9	3866.8
2703.03	3828.88
2699.63	3776.67
2686.85	3774.22
2651.47	3754.5
2649.98	3726.42
2631.4	3699.57
2574.6	3614.72
2573.95	3607.25

	minutes_watched_21	minutes_watched_22
Mean	332.50	368.35
Standard Deviation	485.86	596.41
Sample Size	=COUNTA(B6:B3438)	5104
T-score	-3.05	
Degrees of Freedom (df)	8229	
Critical Value	1.65	

Calculate the t-statistic:



Look up the critical t-value using a t-distribution table to correspond to your chosen significance level (commonly 0.05) and calculated degrees of freedom.

Paid Students

minutes_watched_21	minutes_watched_22
2973.67	4110.17
2939.48	4099.42
2860.78	4085.2
2853.73	4064.35
2830.2	4024.33
2809.67	3948.85
2803.17	3909.85
2797.55	3908.57
2782.08	3879.82
2741.9	3866.8
2703.03	3828.88
2699.63	3776.67
2686.85	3774.22
2651.47	3754.5
2649.98	3726.42
2631.4	3699.57

	minutes_watched_21	minutes_watched_22
Mean	332.50	368.35
Standard Deviation	485.86	596.41
Sample Size	3433	5104
T-score	-3.05	
Degrees of Freedom (df)	8229	
Critical Value	1.65	

Compare the t-statistic to the critical t-value:

If $-3.35 \leq -1.65$, then reject H0$If\ -3.35 \leq -1.65\ ,\ then\ reject\ H0$

Conclusion: Reject because the calculated **t-statistic** is lower than the critical value.

Alternatively, use the Two-Sample t-Test Assuming Unequal Variances that is part of the Data Analysis ToolPak:

Paid Students

	minutes_watched_21	*minutes_watched_22*
Mean	332.502508	368.3547139
Variance	236063.3116	355699.1148
Observations	3433	5104
Hypothesized Mean Differ	0	
df	8229	
t Stat	-3.046942872	
P(T<=t) one-tail	0.001159572	
t Critical one-tail	1.645038819	
P(T<=t) two-tail	0.002319144	
t Critical two-tail	1.960252308	

Decision Rule: If $p-value \leq 0.05$, Reject H0$Decision\ Rule: If\ p-value \leq 0.05\ ,\ Reject\ H0$

Conclusion: Reject because the p-value is lower than the specified significance level α (0.05).

Summary: With a t-statistic of -3.05 (less than the critical value of -1.645), you would reject the null hypothesis because the negative t-statistic indicates that (the mean minutes watched by students in Q4 2021) is significantly smaller than (the mean minutes watched by students in Q4 2022). This is contrary to the null, so we reject it. Of course, rejecting the null hypothesis does not confirm the alternative hypothesis; it suggests that the data provide enough evidence against the null hypothesis.

Free-Plan Students

Calculate the mean, standard deviation, and sample size for free-plan students:

Free Students

minutes_watched_21	minutes_watched_22
4716.68	6338.07
4670.7	6280.12
4622.35	6250.32
4617.75	6208.8
4599.53	6204.55
4581.45	6099.35
4568.73	6091.17
4564.67	6073.37
4481.85	6072.77
4464.67	6071.8
4439.47	6068.17
4388.53	6043.12
4385.1	5998.4

	minutes_watched_21	minutes_watched_22
Mean	=AVERAGE(I6:I32176)	69.15
Standard Deviation	367.26	255.62
Sample Size	32171	120658
T-score		
Degrees of Freedom (df)		
Critical Value		

Free Students

minutes_watched_21	minutes_watched_22
4716.68	6338.07
4670.7	6280.12
4622.35	6250.32
4617.75	6208.8
4599.53	6204.55
4581.45	6099.35
4568.73	6091.17
4564.67	6073.37
4481.85	6072.77
4464.67	6071.8
4439.47	6068.17
4388.53	6043.12
4385.1	5998.4

	minutes_watched_21	minutes_watched_22
Mean	133.93	69.15
Standard Deviation	=STDEV.S(I6:I32176)	255.62
Sample Size	32171	120658
T-score		
Degrees of Freedom (df)		
Critical Value		

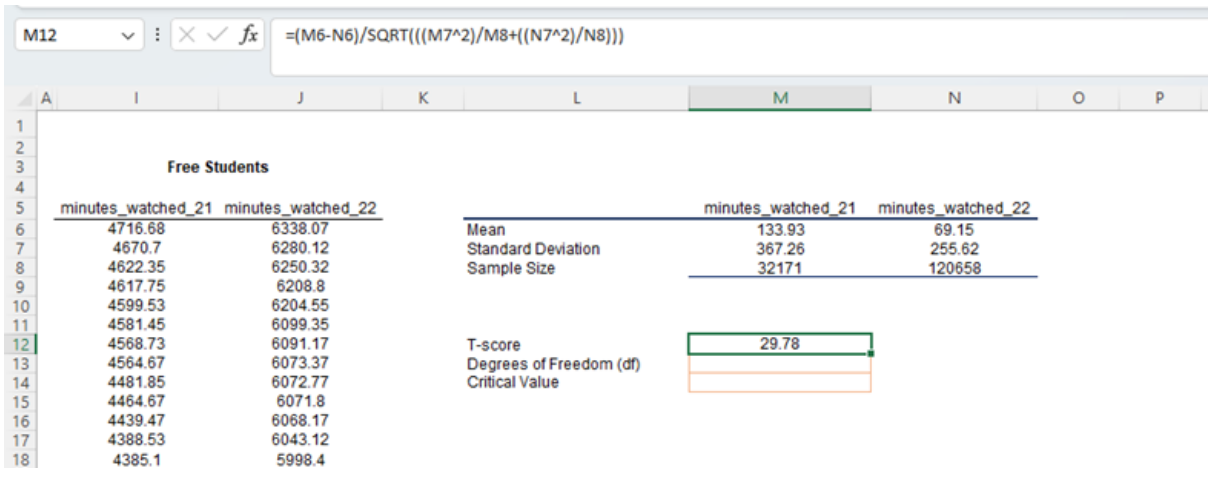
Free Students

minutes_watched_21	minutes_watched_22
4716.68	6338.07
4670.7	6280.12
4622.35	6250.32
4617.75	6208.8
4599.53	6204.55
4581.45	6099.35
4568.73	6091.17
4564.67	6073.37
4481.85	6072.77
4464.67	6071.8
4439.47	6068.17
4388.53	6043.12
4385.1	5998.4
4131	5949.77

	minutes_watched_21	minutes_watched_22
Mean	133.93	69.15
Standard Deviation	367.26	255.62
Sample Size	=COUNTA(I6:I32176)	120658
T-score		
Degrees of Freedom (df)		
Critical Value		

Free Students

Calculate the t-statistic:



Look up the critical t-value using a t-distribution table to correspond to your chosen significance level (commonly 0.05) and calculated degrees of freedom.

Free Students

minutes_watched_21	minutes_watched_22
4716.68	6338.07
4670.7	6280.12
4622.35	6250.32
4617.75	6208.8
4599.53	6204.55
4581.45	6099.35
4568.73	6091.17
4564.67	6073.37
4481.85	6072.77
4464.67	6071.8
4439.47	6068.17
4388.53	6043.12
4385.1	5998.4

	minutes_watched_21	minutes_watched_22
Mean	133.93	69.15
Standard Deviation	367.26	255.62
Sample Size	32171	120658
T-score	29.78	
Degrees of Freedom (df)	40836	
Critical Value	1.65	

Compare the t-statistic to the critical t-value:

If $29.78 \leq -1.65$, then reject H_0 $If\ 29.78 \leq -1.65$, $then\ reject\ H0$

Conclusion: Fail to Reject because the calculated **t-statistic** is higher than the critical value.

Alternatively, use the Two-Sample t-Test Assuming Unequal Variances that is part of the Data Analysis ToolPak to confirm the result:

Free Students

	minutes_watched_21	*minutes_watched_22*
Mean	133.9333129	69.14765544
Variance	134881.7038	65343.34428
Observations	32171	120658
Hypothesized Mean Difference	0	
df	40836	
t Stat	29.77523819	
P(T<=t) one-tail	4.7441E-193	
t Critical one-tail	1.644890942	
P(T<=t) two-tail	9.4881E-193	
t Critical two-tail	1.960022079	

For free-plan students: With a t-statistic of 29.78 (greater than the critical value of -1.645), you would fail to reject the null hypothesis. This means there's not enough evidence to conclude that μ_1 is smaller than μ_2. So, the data supports the null hypothesis that μ_1 is larger than or equal to μ_2.

These results align with previous findings from the confidence intervals and further underscore the difference in engagement patterns between paid- and free-plan students.

Regarding the second part of the question, a Type I error (false positive) occurs when you reject the null hypothesis, which is true. In our case, this would mean concluding that engagement in 2022 is higher when it's not. The probability of making this error is the level of significance, α. Since you (the researcher) choose the significance level of the hypothesis test, the responsibility for making this error lies solely on you.

Note that the significance level is closely related to the confidence level, representing our degree of certainty in the estimated results. It's equal to $(1 - \alpha)$. For example, a significance level of 5% for a hypothesis test means a 5% probability of rejecting a true null hypothesis, corresponding to a 95% confidence level.

A Type II error (false negative) occurs when you fail to reject the null hypothesis, but it's false. In our case, this would mean that the engagement in 2022 is not higher than it is.

The cost to the company of each type of error would depend on the implications of incorrectly concluding that engagement has increased—potentially leading to over-investment in certain features or complacency about needing to improve features—versus incorrectly concluding that engagement has not increased—potentially missing out on recognizing successful features or identifying areas that need improvement.

Task 5

You can find the solution to this problem in the Hypothesis Testing - task 5.xlsx file:

- *Sheet: Task 5*
- *Sheet: Test for Variances*

First, you must perform a two-sample f-test for variances to prove that assumption of unequal variances between the samples for free- and paid-plan subscribers:

Free-plan Students

	minutes_watched_22_US	*minutes_watched_22_IN*
Mean	73.07053569	78.42208628
Variance	95208.64187	101975.5527
Observations	6459	21210
df	6458	21209
F	0.933641833	
P(F<=f) one-tail	0.000347535	
F Critical one-tail	0.967314359	

The p-value indicates the probability of obtaining the observed f-value if the null hypothesis (equal variances) were true. The sample variances are not identical since the p-value is lower than 0. We must perform a left-tailed t-test assuming unequal variances:

Calculate the mean, standard deviation, and sample size for both samples:

Free-plan Students

minutes_watched_22_US	minutes_watched_22_IN
35.75	27.13
71.2	0.37
45.63	0.07
37.98	0.1
0.65	0.37
58.65	9.12
4.82	4.67
41.05	3.73
35.95	4.18
20.4	73.17
1.45	6.43
62.5	1.47
191.6	2188.4
11.83	0.05
0.48	0.18
162.85	12.75
0.1	24.13
1.27	0.73
6.22	15.63

	minutes_watched_22_US	minutes_watched_22_IN
Mean	=AVERAGE(	78.42
Standard Deviation	AVERAGE(**number1**, [number2], ...)	319.34
Sample Size	6459	21210
T-score	-1.21	
Degrees of Freedom (df)	11001	
Critical Value	1.65	

Free-plan Students

minutes_watched_22_US	minutes_watched_22_IN
35.75	27.13
71.2	0.37
45.63	0.07
37.98	0.1
0.65	0.37
58.65	9.12
4.82	4.67
41.05	3.73
35.95	4.18
20.4	73.17
1.45	6.43
62.5	1.47
191.6	2188.4
11.83	0.05
0.48	0.18
162.85	12.75
0.1	24.13
1.27	0.73
6.22	15.63

	minutes_watched_22_US	minutes_watched_22_IN
Mean	73.07	78.42
Standard Deviation	=STDEV.S(	319.34
Sample Size	STDEV.S(**number1**, [number2], ...)	21210
T-score	-1.21	
Degrees of Freedom (df)	11001	
Critical Value	1.65	

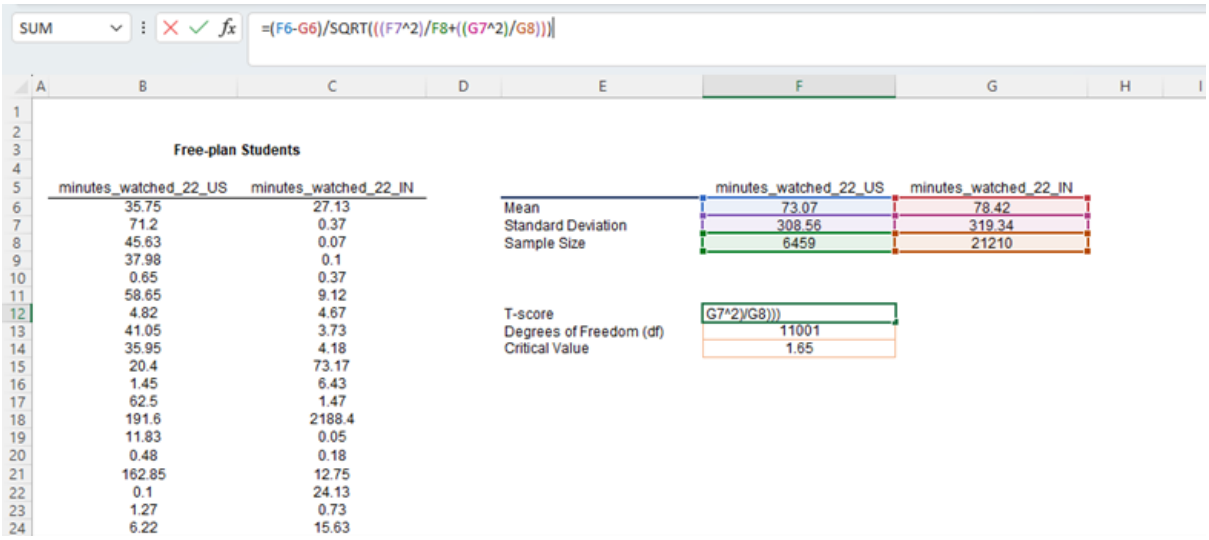
Free-plan Students

minutes_watched_22_US	minutes_watched_22_IN
35.75	27.13
71.2	0.37
45.63	0.07
37.98	0.1
0.65	0.37
58.65	9.12
4.82	4.67
41.05	3.73
35.95	4.18
20.4	73.17
1.45	6.43
62.5	1.47
191.6	2188.4
11.83	0.05
0.48	0.18
162.85	12.75
0.1	24.13
1.27	0.73
6.22	15.63

	minutes_watched_22_US	minutes_watched_22_IN
Mean	73.07	78.42
Standard Deviation	308.56	319.34
Sample Size	=COUNTA(	21210
T-score	-1.21	
Degrees of Freedom (df)	11001	
Critical Value	1.65	

COUNTA(**value1**, [value2], ...)

Calculate the t-statistic:



Look up the critical t-value using a t-distribution table to correspond to your chosen significance level (commonly 0.05) and calculated degrees of freedom.

Free-plan Students

minutes_watched_22_US	minutes_watched_22_IN
35.75	27.13
71.2	0.37
45.63	0.07
37.98	0.1
0.65	0.37
58.65	9.12
4.82	4.67
41.05	3.73
35.95	4.18
20.4	73.17
1.45	6.43
62.5	1.47
191.6	2188.4
11.83	0.05
0.48	0.18
162.85	12.75
0.1	24.13
1.27	0.73
6.22	15.63

	minutes_watched_22_US	minutes_watched_22_IN
Mean	73.07	78.42
Standard Deviation	308.56	319.34
Sample Size	6459	21210
T-score	-1.21	
Degrees of Freedom (df)	11001	
Critical Value	1.65	

Compare the t-statistic to the critical t-value:

If $-1.21 \leq -1.65$, then reject H_0 $If\ -1.21 \leq -1.65\ ,\ then\ reject\ H0$

Conclusion: Fail to Reject because the calculated **t-statistic** is higher than the critical value.

Alternatively, use the Two-Sample t-Test Assuming Unequal Variances that is part of the Data Analysis ToolPak:

Free Students

	minutes_watched_22_US	*minutes_watched_22_IN*
Mean	73.07053569	78.42208628
Variance	95208.64187	101975.5527
Observations	6459	21210
Hypothesized Mean Difference	0	
df	11001	
t Stat	-1.210387573	
P(T<=t) one-tail	0.113078106	
t Critical one-tail	1.644992151	
P(T<=t) two-tail	0.226156213	
t Critical two-tail	1.960179649	

Decision Rule: If $p-value \leq 0.05$, Reject H_0 $Decision\ Rule:\ If\ p-value \leq 0.05\ ,\ Reject\ H0$

Conclusion: Fail to Reject because the p-value is higher than the specified significance level α (0.05).

If the hypothesis that US students watch more or an equal amount of content as Indian students is rejected, this suggests that US students watch less content on average than students in India.

Conclusions:

- Paid-plan students showed a significant increase in engagement from Q4 2021 to Q4 2022, as indicated by the increase in mean minutes watched and confidence intervals.
- Free-plan students, on average, showed a decrease in engagement from Q4 2021 to Q4 2022, as evidenced by the decrease in mean minutes watched and confidence intervals.
- Paid-plan students watched substantially more minutes than free-plan students in Q4 2022, suggesting that paid subscriptions contribute to higher engagement.
- There was an increasing variability in engagement levels among both paid and free-plan students from Q4 2021 to Q4 2022, as indicated by the increasing standard deviation and kurtosis.

Recommendations:

1. **Personalized Strategies**: Implement personalized approaches to boost engagement, considering the increasing variability in engagement levels among students.
2. **Paid-Plan Features**: Analyze the features, content, or other factors that contribute to higher engagement among paid-plan students, and consider enhancing or promoting them further.
3. **Free-Plan Engagement**: Evaluate the reasons behind the decreasing engagement among free-plan students and implement strategies to improve their experience and involvement with the platform.
4. **Regional Differences**: Investigate potential regional or country-specific differences in engagement levels, as suggested by the analysis comparing US and Indian students.

Implications:

- **Resource Allocation**: Allocate resources strategically based on engagement patterns, focusing on regions or student segments with higher potential for growth and engagement.
- **Content and Feature Development**: Invest in developing content and features that cater to the preferences and needs of paid-plan students, as they demonstrate higher engagement levels.
- **Monetization Strategies**: Explore strategies to convert free-plan students to paid plans, potentially by offering attractive features or content exclusively for paid subscribers.
- **Market Expansion**: Identify regions or countries with higher engagement levels and consider expanding marketing efforts or tailoring offerings to those markets.
- **User Segmentation**: Develop a deeper understanding of user segments based on engagement levels, demographics, and other relevant factors to optimize the platform's offerings and user experience.