Customer Engagement Analysis in Excel Project

By

Pranoy Chakraborty

Descriptive Statistics

Task 1

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.

Please open the 'Engagement project.xlsx' file and navigate to the 'Task 1 and 2' sheet using Microsoft Excel. Your first task is to provide insights into the relative engagement levels in Q4 2021 and Q4 2022. You will focus on low-engagement users (those who watched between 1 and 100 minutes in 2021). Low-engagement users often represent the most significant potential for growth. If 365 can find ways to increase its usage, it could significantly impact the overall use of the platform.

If there are repeated students who watched in Q4 2021 and Q4 2022, how does their engagement compare between the two periods? Compute the mean, median, and standard deviation for these groups. Is there a difference in engagement between paid- and free-plan subscribers?

Solution

1. Open the 'Engagement project.xlsx' file and navigate to the 'Task 1 and 2' sheet using Microsoft Excel.
2. Apply the AVERAGE, MEDIAN, and STDEV.S Excel functions to the ‘minutes_watched’ column to compute the mean, median, and standard deviation for both groups (free- and paid-plan students).

Paid-Plan 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
160274	1	61.97	2480.43
9305	1	72.33	715.95
211124	1	1.12	5.7
172631	1	88.8	217.65
240248	1	42.17	392.63
233048	1	12.3	47.92
1436	1	4.88	1045.55
150663	1	41.52	253.78
156542	1	19.33	9214.13
862	1	57.8	162.23
240591	1	77.25	293.13
230669	1	72.52	230.67

Free-plan 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
245547	0	63.03	28.9
37976	0	35.17	30.28
231774	0	48.85	0.05
180503	0	16.02	6.3
117871	0	2.48	0.07
87495	0	3.9	0.05
186757	0	8.02	57.82
247599	0	20.48	20.42
225846	0	66.57	24.17
59019	0	6.25	0.13
228556	0	3.65	137.2
172658	0	97.2	13.03

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

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

Task 2

Please open the 'Engagement project.xlsx' file and navigate to the 'Task 1 and 2' sheet using Microsoft Excel. Calculate the skewness and kurtosis of students who watched content in Q4 2021 and Q4 2022. Consider paid- and free-plan students. Does the result contradict the mean and median values previously obtained?

Solution

1. **Finding the skewness:** To calculate the skewness of a set of numbers in Excel, you can use the SKEW () function. Select the range of cells to calculate the skewness and put it inside the brackets.
2. **Finding the kurtosis:** Calculate the kurtosis of a set of numbers in Excel using the KURT() function. Like the SKEW () function, select the range of cells to calculate the kurtosis.

Paid-Plan 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
160274	1	61.97	2480.43
9305	1	72.33	715.95
211124	1	1.12	5.7
172631	1	88.8	217.65
240248	1	42.17	392.63
233048	1	12.3	47.92
1436	1	4.88	1045.55
150663	1	41.52	253.78
156542	1	19.33	9214.13
862	1	57.8	162.23

Free-plan 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
245547	0	63.03	28.9
37976	0	35.17	30.28
231774	0	48.85	0.05
180503	0	16.02	6.3
117871	0	2.48	0.07
87495	0	3.9	0.05
186757	0	8.02	57.82
247599	0	20.48	20.42
225846	0	66.57	24.17
59019	0	6.25	0.13

Paid-Plan Students

	minutes_watched_21	minutes_watched_22
Skewness	0.63	7.07
Kurtsis	-0.85	58.48

Free-Plan Students

	minutes_watched_21	minutes_watched_22
Skewness	1.17	15.06
Kurtsis	0.36	315.76

Confidence Intervals

Task 3

Please open the 'Engagement project.xlsx' file and navigate to the 'Task 3 sheet using Microsoft Excel.

- Students who haven't had a paid-plan subscription
  - Engaged in Q4 2021
  - Engaged in Q4 2022
- Students who have been paid-plan subscribers
  - Engaged in Q4 2021
  - Engaged in Q4 2022

For each of the four groups, determine the minute interval within which you can be 95% confident that a randomly selected individual will be situated.

Solution

Paid-Plan 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
2571.68	3589.12
2571.35	3585.32
2517.88	3572.52
2511.82	3566.3
2505.03	3550.82
2489.27	3516.85
2465.6	3501.77
2453.7	3488.13
2415.2	3479.73
2414.2	3477.15
2401.58	3462.83

Free-plan 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
4123.17	5918.68
4119.53	5875.6
4115.7	5852.32
4114.38	5849.75
4112.93	5821.23
4075.7	5804.3
4024.75	5779.78
4022.77	5776.73
4003.25	5727.63
3983.32	5649.58
3975.67	5574.72
3956.37	5541.45
3954.93	5384.4
3947.18	5360.8
3940.65	5293.23

Paid-Plan Students

minutes_watched_21	
Confidence Level(95.0%)	16.26

minutes_watched_22	
Confidence Level(95.0%)	16.37

Paid-Plan Students

minutes_watched_21	
Confidence Interval(95.0%)	[316.24,348.76]

minutes_watched_22	
Confidence Interval(95.0%)	[351.98,384.72]

Free-plan Students

minutes_watched_21	
Confidence Level(95.0%)	4.01

minutes_watched_22	
Confidence Level(95.0%)	1.44

Free-plan Students

minutes_watched_21	
Confidence Interval(95.0%)	[129.92,137.94]

minutes_watched_22	
Confidence Interval(95.0%)	[67.71,70.59]

Mean

Paid-Plan Students

minutes_watched_21	minutes_watched_22
332.50	368.35

Free-plan Students

minutes_watched_21	minutes_watched_22
133.93	69.15

The confidence intervals reveal distinct engagement patterns between paid-plan and free-plan students. For paid-plan students, the minutes watched in Q4 2021 (316.24 to 348.76) and Q4 2022 (351.98 to 384.72) indicate consistently higher engagement with slightly increased variability in 2022. In contrast, free-plan students showed significantly lower engagement, with tighter confidence intervals in Q4 2021 (129.92 to 137.94) and Q4 2022 (67.71 to 70.59), reflecting reduced and more uniform engagement over time. This suggests that paid-plan students maintain higher and more diverse engagement levels, while free-plan students exhibit declining and increasingly predictable behavior.

Hypothesis Testing

Task 4

Use the data in Task 4 of the 'Engagement project.xlsx' to solve the following task.
You want to reach a data-driven customer engagement decision on whether the platform's new features contribute to the increase of minutes watched on the platform for both free-plan and paying students—i.e., the rise in student engagement in their study process. To do that, use hypothesis testing on both groups (free-plan and paying) for 2021 and 2022.
Additionally, make the following assumptions:

- For free-plan students, perform a two-sample t-test assuming unequal variances.
- For paying students, conduct a two-sample t-test assuming unequal variances.

What conclusion can you draw from this test? Comment on the results of committing a Type I or a Type II error in this study. Which one would result in higher costs for the company?

Paid-Plan Students		Free-plan Students	
minutes_watched_21	minutes_watched_22	minutes_watched_21	minutes_watched_22
2973.67	4110.17	4716.68	6338.07
2939.48	4099.42	4670.7	6280.12
2860.78	4085.2	4622.35	6250.32
2853.73	4064.35	4617.75	6208.8
2830.2	4024.33	4599.53	6204.55
2809.67	3948.85	4581.45	6099.35
2803.17	3909.85	4568.73	6091.17
2797.55	3908.57	4564.67	6073.37
2782.08	3879.82	4481.85	6072.77
2741.9	3866.8	4464.67	6071.8
2703.03	3828.88	4439.47	6068.17
2699.63	3776.67	4388.53	6043.12
2686.85	3774.22	4385.1	5998.4
2651.47	3754.5	4131	5949.77
2649.98	3726.42	4123.17	5918.68
2631.4	3699.57	4119.53	5875.6
2574.6	3614.72	4115.7	5852.32
2573.95	3607.25	4114.38	5849.75
2571.68	3589.12	4112.93	5821.23
2571.35	3585.32	4075.7	5804.3
2517.88	3572.52	4024.75	5779.78
2511.82	3566.3	4022.77	5776.73
2505.03	3550.82	4003.25	5727.63
2489.27	3516.85	3983.32	5649.58
2465.6	3501.77	3975.67	5574.72
2453.7	3488.13	3956.37	5541.45
2415.2	3479.73	3954.93	5384.4

t-Test: Two-Sample Assuming Unequal Variances (for Paid-Plan Students)

	minutes_watched_21	*minutes_watched_22*
Mean	332.502508	368.3547139
Variance	236063.3116	355699.1148
Observations	3433	5104
Hypothesized Mean Difference	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	

t-Test: Two-Sample Assuming Unequal Variances (for Free-plan 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	

Conclusion from the t-tests

1. **Paid-Plan Students**:
    - The **t-statistic** is −3.05-3.05−3.05, and the **p-value (two-tailed)** is 0.00230.00230.0023, which is less than the significance level (α=0.05\alpha = 0.05α=0.05).
    - **Conclusion**: There is a statistically significant difference in the mean minutes watched between Q4 2021 and Q4 2022 for paid-plan students. This indicates that engagement has increased for this group.
2. **Free-Plan Students**:
    - The **t-statistic** is 29.7829.7829.78, and the **p-value (two-tailed)** is 9.4881×10−1939.4881 \times 10^{-193}9.4881×10−193, which is much smaller than the significance level (α=0.05\alpha = 0.05α=0.05).
    - **Conclusion**: There is a highly significant difference in the mean minutes watched between Q4 2021 and Q4 2022 for free-plan students. Engagement for free-plan students has dropped considerably.

Type I and Type II Errors in the Study

- **Type I Error (False Positive)**: Concluding there is a significant difference in engagement between years when no real difference exists.
    - **Implications**: If the company incorrectly believes that engagement patterns have changed and implements unnecessary interventions, it could incur costs for unnecessary strategy changes or campaigns.
- **Type II Error (False Negative)**: Failing to detect a significant difference in engagement when one actually exists.
    - **Implications**: The company might miss out on identifying areas where engagement is dropping (e.g., free-plan users) or opportunities where engagement is improving (e.g., paid-plan users). This could result in lost revenue or customer dissatisfaction.

Which Error is More Costly?

In this context, **Type II Error** would likely result in higher costs for the company. Missing the decline in engagement for free-plan users could lead to a loss of potential paid-plan conversions, while ignoring improved engagement among paid-plan users might prevent leveraging these insights for retention strategies or marketing efforts. Recognizing true engagement trends is critical for optimizing business strategies.

Task 5

Your last task is determining whether the average number of minutes watched in the US is similar to that in India.
Understanding the differences in usage patterns can help in product localization. The platform might need to tailor its content, features, or user interface to better fit the preferences or needs of users in different regions.
You'll focus only on free-plan students in 2022. Use the Excel sheet Task 5 to perform your calculations.
Your null hypotheses should (respectively) include the following:

- The engagement (minutes watched) in the US is higher than or equal to that in India (μ1 ≥ μ2). We test only free-plan students.
- The engagement (minutes watched) in the US is lower than that in India (μ1 < μ2). We test only free-plan students.

Additionally, perform a two-sample t-test assuming unequal variances.
Optional: *Perform a two-sample f-test for variances to support the assumptions.*
What conclusion can you draw from this test? Is the engagement in the US higher than that in India?

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
278.33	30.27
446.95	0.67
5.62	2.4
1.62	0.2
239.73	86.08
0.18	42.83
0.75	72.87
362.5	50.67

t-Test: Two-Sample Assuming Unequal Variances

	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	

F-Test Two-Sample for Variances

	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	

Analysis of the Two-Sample t-Test

Final Conclusion

The two-sample t-test indicates that the average engagement (minutes watched) in the US is not significantly lower than in India. However, the F-test confirms a significant difference in the variances between the two groups, suggesting more variability in minutes watched in the Indian group compared to the US.

While no significant difference was detected in mean engagement, the platform might still explore localization strategies for India, considering the higher variance in engagement patterns, which may point to diverse user behaviors.

Question 1:

Consider your solution to Descriptive Statistics (Task 1). Which of the following statements is correct?

- (●) The mean and median of minutes watched for free-plan students increased from 2021 to 2022.
- ( ) The standard deviation for free- and paid-plan students decreased from 2021 to 2022.
- ( ) The mean and median of minutes watched by paid-plan students increased from 2021 to 2022.
- ( ) The median of minutes watched for paid-plan students in Q4 2022 equals approximately 273.

Question 2:

Consider your solution to Descriptive Statistics (Task 2). Which of the following statements is most likely correct?

- ( ) The data is approximately normally distributed for free- and paid-plan students, with the mean close to the median.
- (●) The skewness for paid-plan students in 2022 is lower than that for free-plan students.
- ( ) The kurtosis of minutes watched for free-plan students is similar in Q4 2021 and Q4 2022.
- ( ) The skewness of minutes watched for free-plan students equals 0.63 in Q4 2022.

Question 3:

Consider your solution to Confidence Intervals (Task 3). What conclusion can we draw based on the confidence intervals of the free-plan students who watched in Q4 2021 and Q4 2022?

- ( ) The average number of minutes students watched in Q4 2021 is higher than in Q4 2022. This is based on the observation that the lower confidence interval limit for Q4 2021 is less than the lower limit for Q4 2022.
- ( ) The average minutes students watched in Q4 2022 are significantly lower than those in Q4 2021 because the entire confidence interval for Q4 2022 is below the one for Q4 2021.
- (●) The average minutes students watched in Q4 2022 are lower than those in Q4 2021. This is inferred from the fact that the upper limit of the confidence interval for Q4 2022 is higher than the one for Q4 2021.
- ( ) There is no significant difference in the average minutes watched by students in Q4 2021 and Q4 2022 because the confidence intervals for both periods overlap.

Question 4:

Consider your solution to Confidence Intervals (Task 3). What conclusion can we draw based on the confidence intervals of the paying students who watched in Q4 2021 and Q4 2022?

- ( ) The students who watched a course in Q4 2021 had significantly more minutes watched on average than those who watched a course in Q4 2022.
- (●) The students who watched a course in Q4 2022 had substantially more minutes watched on average than students who watched a course in Q4 2021.
- ( ) The students who watched a course in Q4 2021 and those who watched a course in Q4 2022 had the same average minutes watched.
- ( ) We cannot make a valid conclusion about the differences in the average minutes watched between students who watched a course in Q4 2021 and those who watched a course in Q4 2022.

Question 5:

Consider your solution to Hypothesis Testing (Task 4). What is the value of the t-statistic you obtain for free-plan students? Do you accept or reject the null hypothesis?

- ○ -3.05 – We accept the null hypothesis.
- ○ -3.05 – We reject the null hypothesis.
- ◉ 29.78 – We accept the null hypothesis.
- ○ 15.15 – We reject the null hypothesis.

Question 6:

Consider your solution to Hypothesis Testing (Task 4). What is the value of the t-statistic you obtain for paying students? Do you accept or reject the null hypothesis?

- ○ -3.05 – We accept the null hypothesis.
- ◉ -3.05 – We reject the null hypothesis.
- ○ 15.15 – We accept the null hypothesis.
- ○ 15.15 – We reject the null hypothesis.

Question 7:

Consider your solution to Hypothesis Testing (Task 5). What is the value of the t-statistic you obtain? Do you accept or reject the null hypothesis?

- ◉ -1.21 – We accept the null hypothesis.
- ○ -1.21 – We reject the null hypothesis.
- ○ 12.25 – We accept the null hypothesis.
- ○ 12.25 – We reject the null hypothesis.

Question 8:

Consider your solution to Hypothesis Testing (Task 5). How would you interpret the result?

- ○ On average, US students watch more video content than Indian students.
- ○ We cannot make a valid conclusion.
- ○ On average, students in India watch more video content than students in the US.
- ◉ Both student groups watch approximately the same number of minutes.