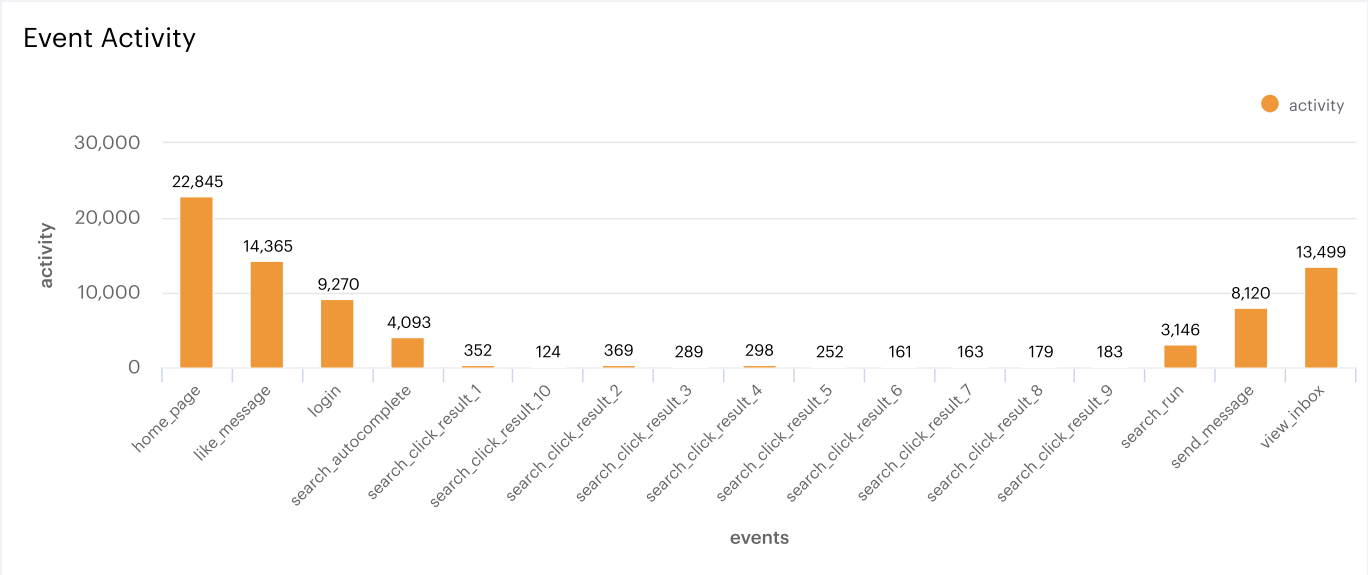
ANALYTICS TRAINING: Validate A/B Test

We will validate the test results for the A/B test by reviewing several parameters. They include (a) test calculations, (b) other metrics, and (c) other anomalies. After this, we will recommend to move forward with the new feature or not.

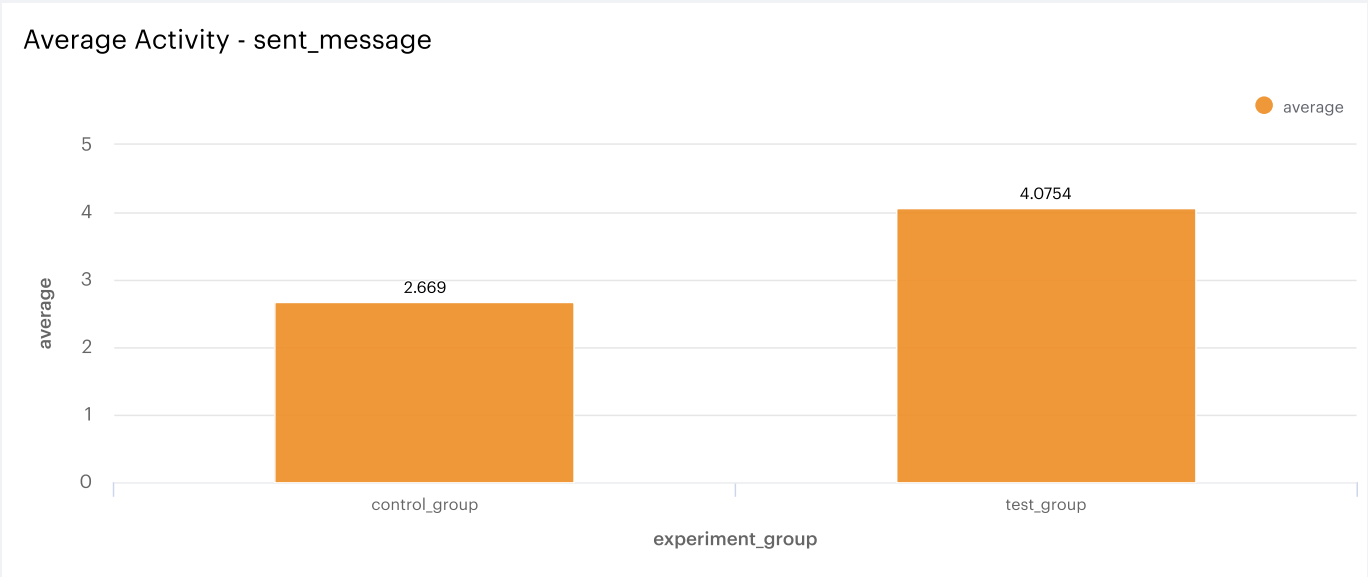
Activities Count

Below shows the events that occurred in Yammer. Higher counts imply more engagement. Here, the two events are 'home_page' and 'view_inbox'. The lowest count are those associated with search. The activities of interest rely on the messages feature. They include 'Sent Message', 'Login', and 'Like Message.' They make sense, because these events show whether the new feature is effective or not. If activity from these events are consistent, then the new feature works.



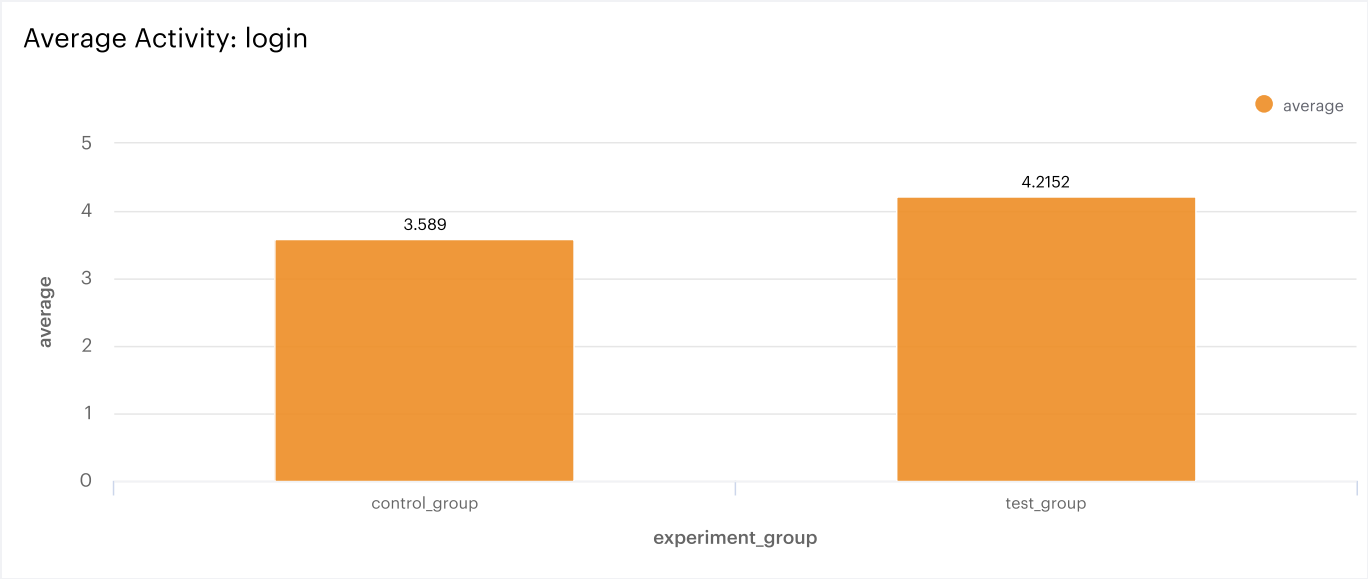
Original Activity: Messages Sent

This is the original query. Here shows the average number of messages sent per user. However, let's explore further.



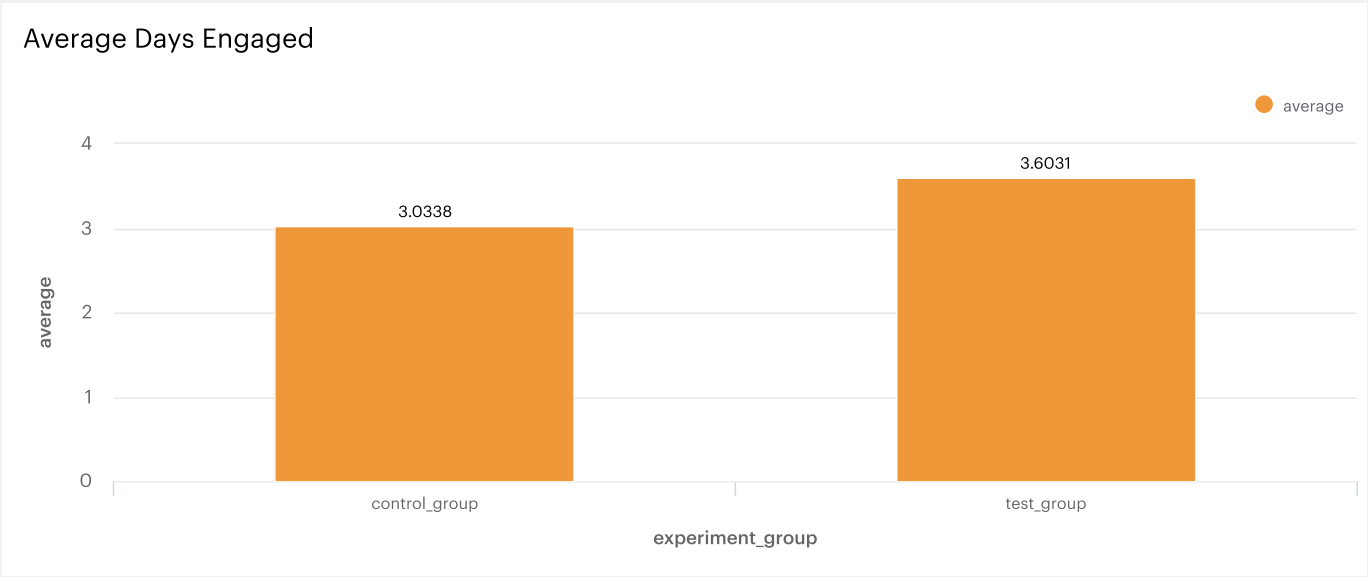
One Recommended Activity: Users Logged In

Mode recommended seeing the 'login' activity. Here, the test group had higher log-in activity. However, this is both good and bad. It's good because people may like the new feature, but it is bad because the user may be kicked out of the system or the new feature may be causing some form of headaches. So let's look at the daily activity.



Suporting the Login Activity

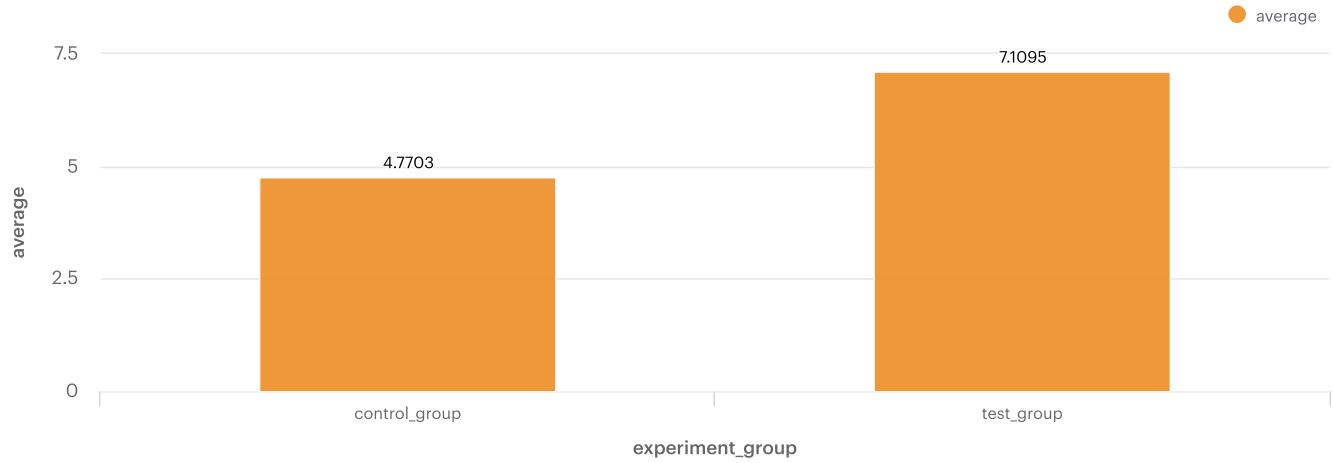
While the treatment group has more log-in activity than the control group, it may speculate whether users are logging back in to the page because of some inconsistency. Let's check the average days each user logs in. Both the control and test group have the same average. During this trial period, the user logins in three times daily. Because the test group has a slightly higher average, it is still insignificant. Therefore, users are not being kicked out of the system because of the new feature. And the log-in activity is normal. This is good.



Interesting Activity: Liked Messages

Because sending messages is one form of activity, users liking a message may support the 'value' one may get from Yammer. Liked messages show that the users are engaged with the product. It's one activity for a user to post, but it is responding to messages that provides the necessary engagement. Below, the test group has a higher average of 'liking' messages. This is good and does support the new feature's need by allowing users to engage with the platform. 'Liking' message may be a quick way to respond to users but it is still an activity worth promoting! This is also what encourages users to send messages too.

Average Activity: like_message



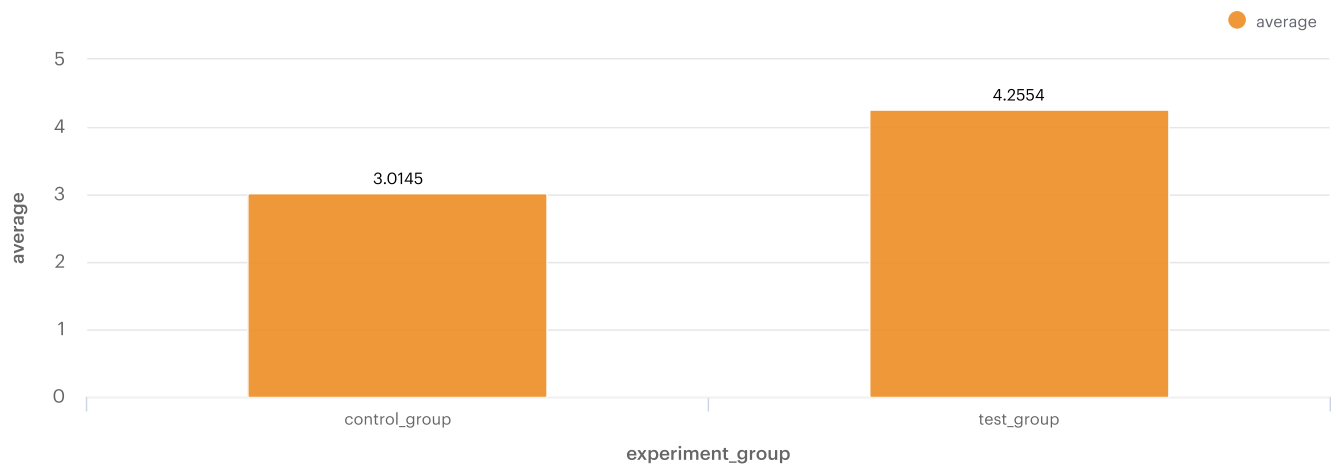
How Did Existing Users Perform?

Now that the activity was reviewed, let's review another aspect: time. Because it is a month of participation, how did the existing users (those who have used the current feature) did well. This query extracted those who signed up one month before the experiment (so this is both veteran users and 'new' users who test-runned the system). The total existing users were 1,337 users, and the division seems fair (about half were part of the control group). This is a good democracy. But moreover, the average number of 'send messages' is higher for the test group. So the new feature is fine among existing users. But what about new ones, or whether there were new users during the trial run?

existing_users

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	stde
1	publisher_upda…	control_group	691	1337	0.52	2083	3.014…	0	0	3.8
2	publisher_upda…	test_group	646	1337	0.48	2749	4.255…	1.24	0.411…	4.8

Average Activity: like_message, Existing Users



What about the New Users?

The new users were all placed in the control group, so they only saw the pre-existing feature. This may be to rely on the experienced users to try the feature as loyalty and experience may push them further. It would have been best for new users to get acquainted with the test feature. Their inexperience (as in how long they have used the platform) may had a minimal impact to the control group's performance (as there are more new than existing users in the control group), but to see a new feature may test how future users may engage with the feature moving forward.

new_users

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	stde
1	publisher_upda…	control_group	873	873	1	3702	4.240…	0	0	5.25

Are test calculations fine?

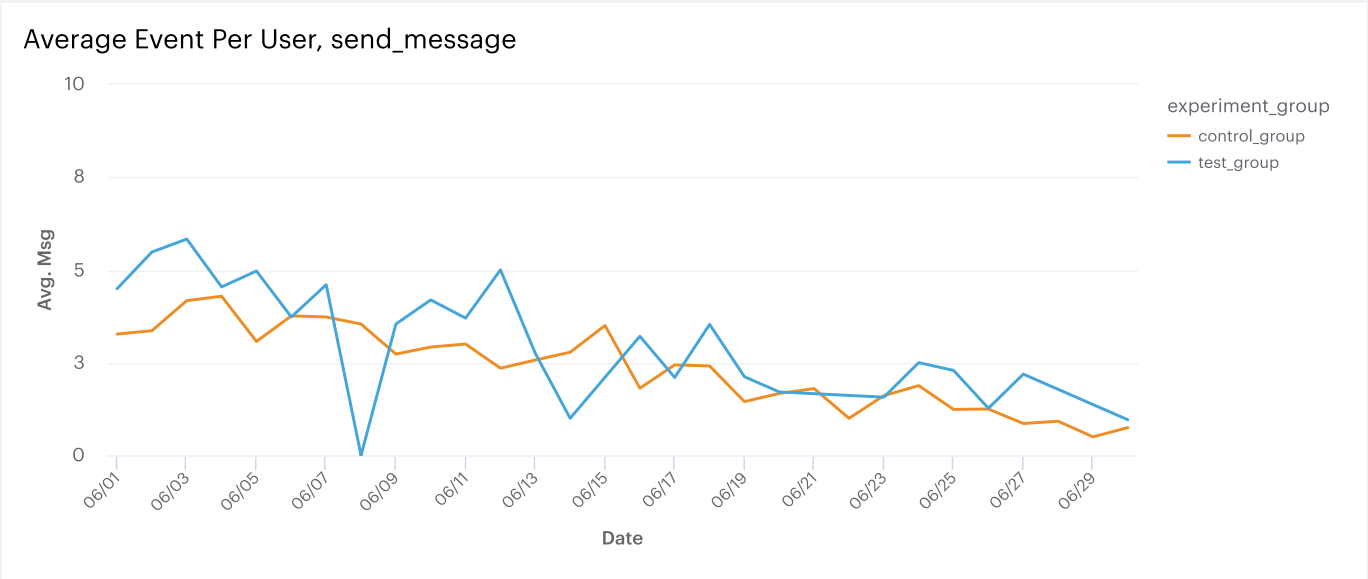
Well, there are several areas worth checking. They include:

- **T-TEST**: (c.average - c.control_average) / SQRT((c.variance/c.users) + (c.control_variance/c.control_users))
- **RATE LIFT**: (c.average - c.control_average)/c.control_average
- **P-VALUE**: (1 - COALESCE(nd.value,1))*2

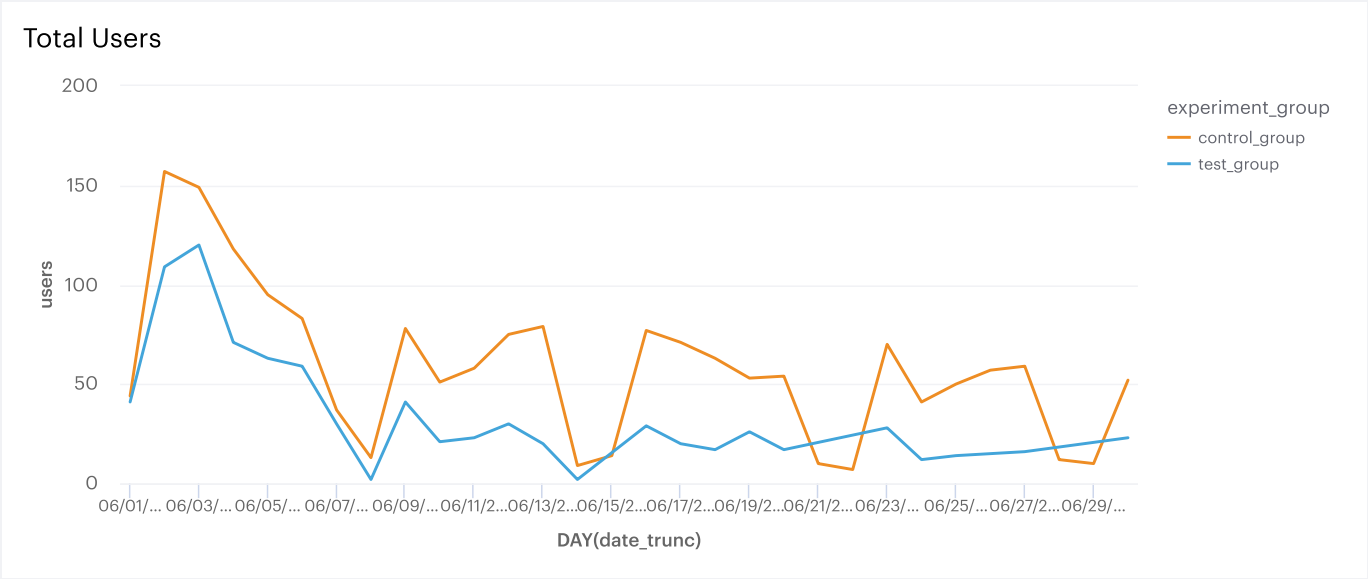
The P-value is taking the t-test statistic's result and providing a significance; here COALESCE is a just-in-case function (if no value, then use 1). A small P-value yields the new feature's success. The rate lift across different tests is also positive yet small. As for the T-test, they used the two-sample T-test. So it is fine. However, as shown before, 100% of those who created Yammer accounts during the trial were placed in the control test. This may have reduced the control group's average. So the testing calcualtions were not done incorrectly?

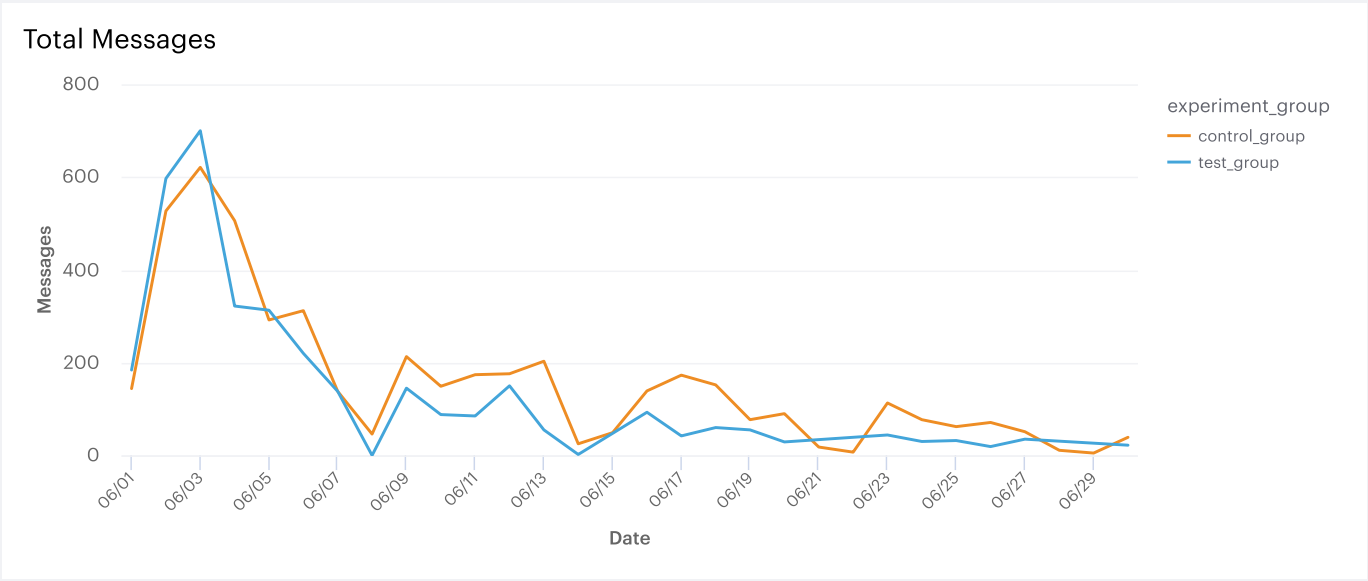
One More Analysis ...

After the test calculations, let's look at how the users posted messages daily. It's one to provide test statistics. It's another to see the feature in user on a daily basis. Otherwise, why promote it?



Keep in mind, this is seeing the average message per day. Both groups experience a lower average. This doesn't mean that activity is low, but it looks as though the test group is sending more messages on average daily. This is supportive of the new feature. This also supports the average days a user is logged in.





Conclusion

The analysis shows that the tests didn't like: people were getting more value in part of the new feature . The statistics were not created or distorted to favor the new feature becuase other metrics promoted the use favorably. However, a flaw in the test was assigning new users to the 'control group' as they may have lowered the average and statistics. If they were randomly placed between the test group and the control group, the results would be different. In our final analysis, just having a higher messages per user for the new feature is supportive. The team should investigate further into installing the new feature, but it would be best to leverage it during a second-popular period and with new users randomized into different groups.