Name: Stella Sun
Contact: ss8955@nyu.edu

Case Study: Messaging Button Experiment

Problem Statement

This case study is trying to understand if the change of the button design, from "dark" to "light" will increase the discoverability and usage of message feature.

Experiment Method

To better illustrate the difference of the change, I would like to conduct an experiment, an A/B testing, to address the problem.

Data

- Data Clean

There are three datasets given

experiment_subjects (user_id, audience_name, enrolled_at)
experiment_actions (user_id, action, new_thread, timestamp)
messaging (message_id, thread_id, sender_type, timestamp, sender_id, recipient_id).

I checked a few things to ensure the data quality on all three tables:

missing values/null values: if any column had missing/null values? Was it due to randomness?
duplicated ids(user_id/message_id): there shouldn't be users who were in both control and treatment groups;
outliers (irregular timestamp, actions, etc.): used distribution plot to find outliers.

Summary:
experiment_subjects

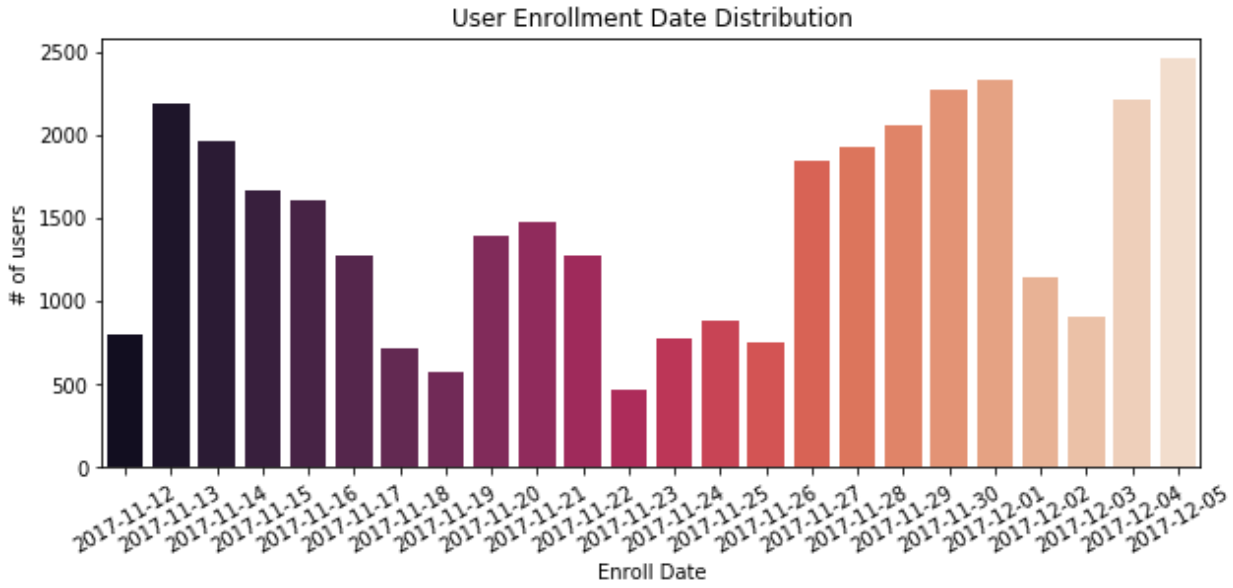
1. Dropped 2 rows without user_id
2. Dropped 16 duplicated user_ids in both control and treatment groups

experiment_actions

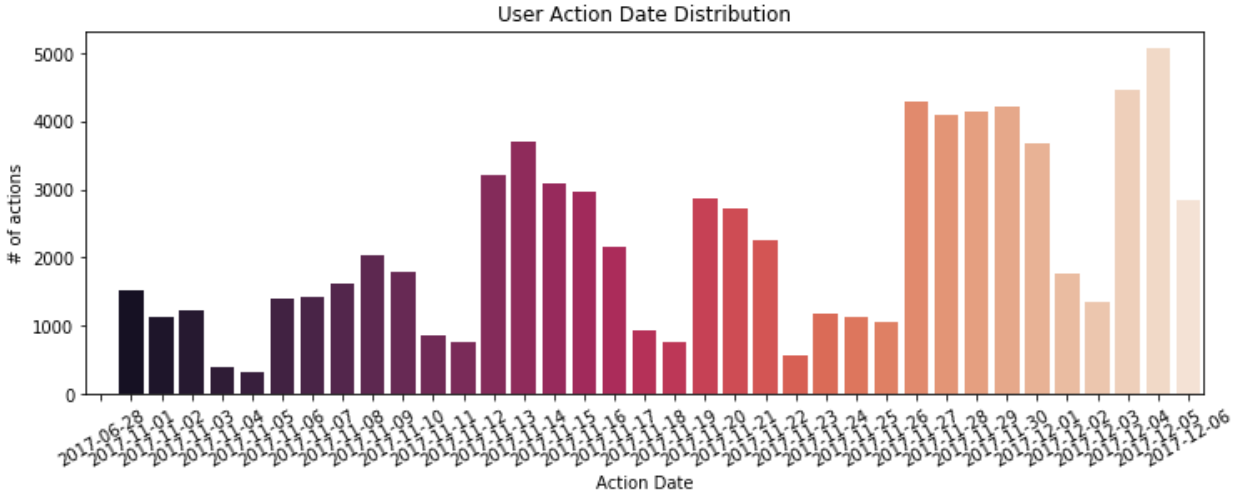
1. Dropped actions on 2017-06-28 as most actions happened between 2017-11-01 and 2017-12-06

messaging: Nothing irregular found in messaging table

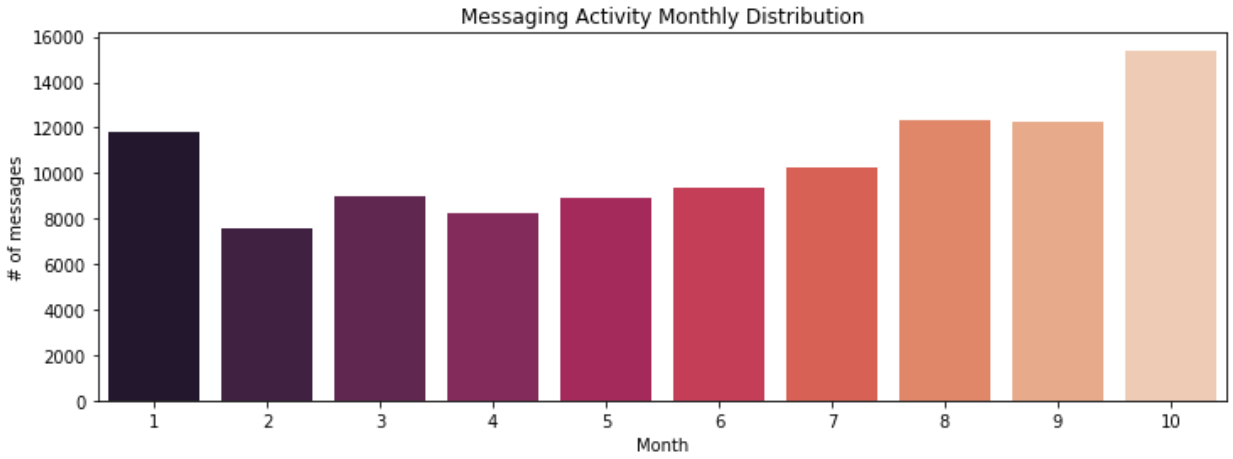
Here were some analysis on distribution I did on three tables:



There were different groups of users who enrolled in experiments on different dates, and we could see that all users were enrolled between 2017-11-12 and 2017-12-05. I would see these groups as different cohort users, and we could capture user behaviors on each cohort later.

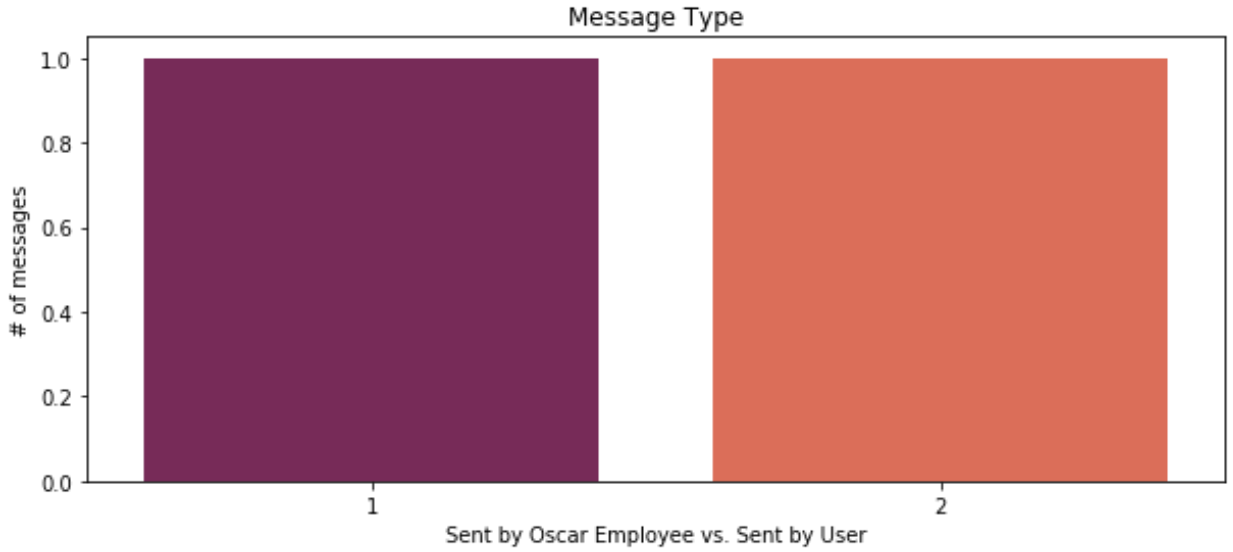


The action table recorded all user actions, click button, view inbox, sent message between 2017-06-28 and 2017-12-06, however, compared to all other action dates, 2017-06-28 seems to be an outliers, so I removed the actions on this day.

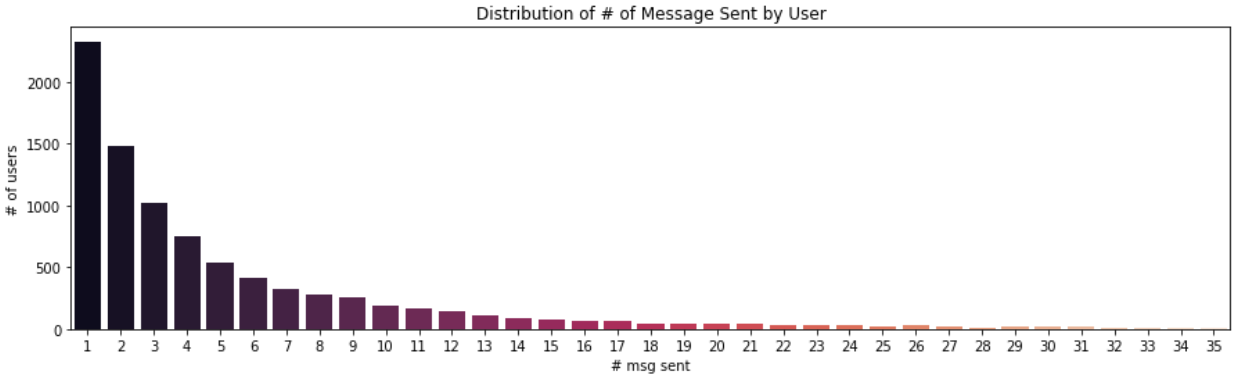


The messaging table recorded the messages sent between users and Oscar employees, since our earliest experiment was on 2017-11-12, I only plotted the message data from 2017-01 to 2017-10 to see the messaging activities before the experiment.

As we can see that, users sent the most messages on October, and then January, while other months have fewer messages.



There were very similar amount of messages sent by users and messages sent by Oscar employees, therefore, it is a good sign to show that Oscar employees were really responsive to user messages.



The distribution of # message sent by user shows that, most of users sent one message, while there is an extreme value when some users sent a huge amount of messages.

Metrics

I am going to evaluate two things to measure the success of the new button design: discoverability and usage.

Discoverability

1. % users who clicked the button
2. Average clicks per user
3. % users who sent message
4. Average messages per user

Usage

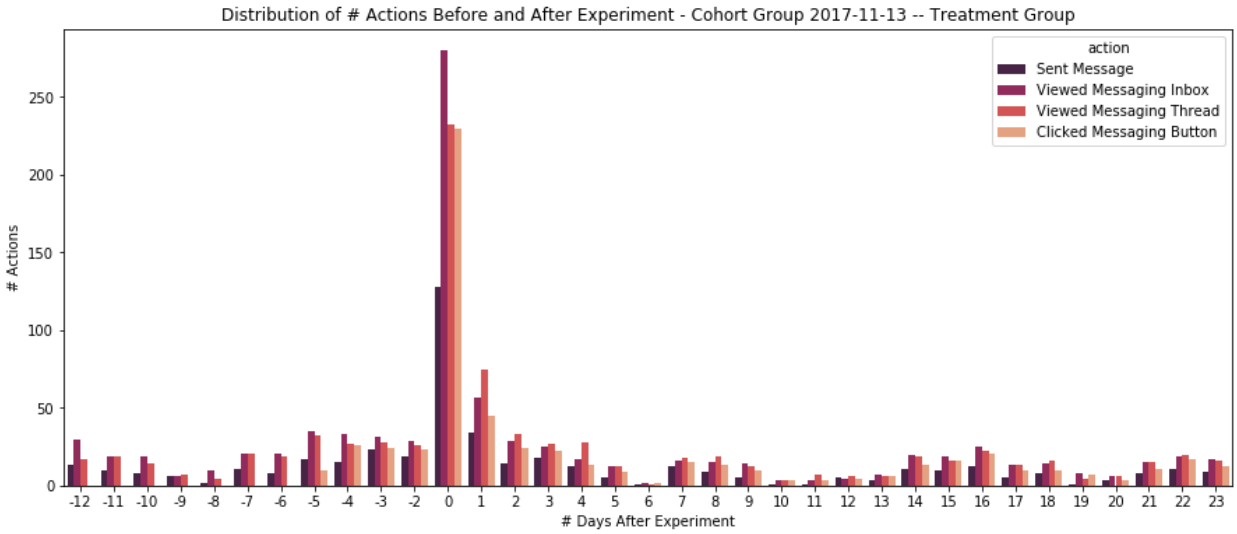
1. % users who sent at least 1 message
2. Average messages sent per user
3. % users who received at least 1 message
4. Average messages received per user

Now as I defined the metrics, I would start looking at user data, and defined sample data. As we had different cohort users who joined the experiment at different date, there were two problems to solve:

1. Can I aggregate different cohort users as one sample group to do A/B testing?
2. How long should I run the experiment; should I keep the experiment time consistent for all cohort users?

Understand User Behaviors in Different Cohort Group

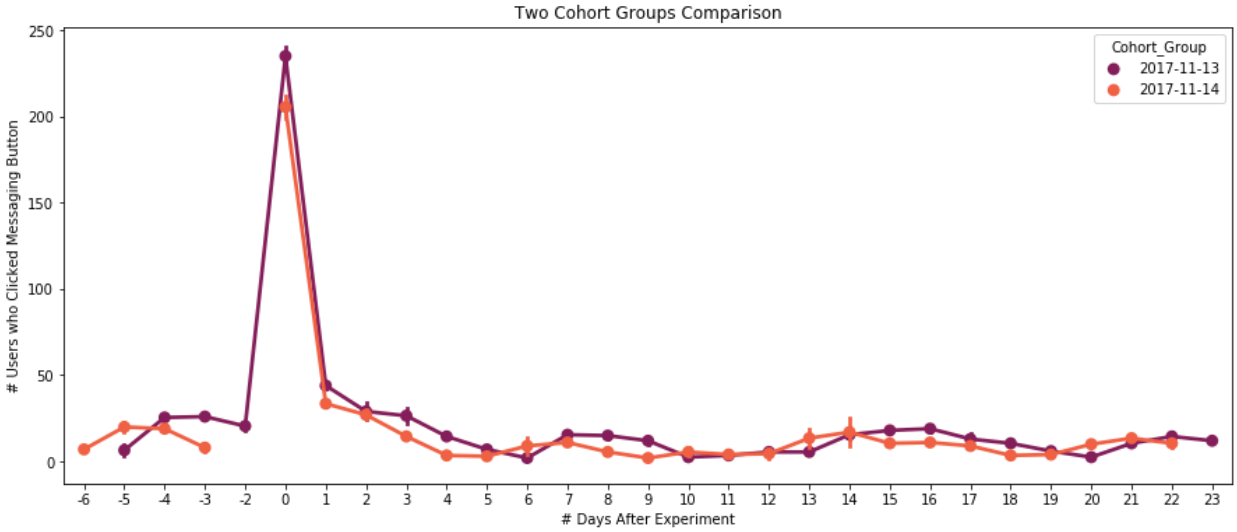
I picked one cohort group, users who enrolled in the experiment on 2017-11-13, and checked user actions before and after the experiment on four actions: clicked button, viewed inbox, viewed thread, sent message.



The chart above showed the trend of actions(clicked messaging button, viewed messaging inbox, viewed messaging thread, sent message) of the cohort group 2017-11-13. X-axis refers to "number of days away from the experiment start date", and y-axis refers to "number of users who take actions".

We could see that on the day of enrollment, there was a huge number of users who clicked the button; in the following 7 days, number of users who clicked on the button dropped. As there might be some weekly trend in user behaviors, I was going to use "7 day experiment" to make sure that there was enough time for each cohort users take actions.

Also, I wanted to find out action behaviors with different cohort groups: if both groups have similar trend, I could aggregate them together as one sample data.



The chart above only refers to "clicked messaging button" action; x-axis refers to "number of days away from the experiment start date" and y-axis refers to "number of users who clicked on the button".

We could easily tell from the chart above that, different cohort groups behave very similar in the same amount of time after the experiments, even though they enrolled on different dates, therefore, I would aggregate different cohort groups without concerns.

Hypothesis

Null Hypothesis:
There is no difference between the group who used the "dark" designed button and the group who used "light" designed button

Alternative Hypothesis:
There is difference between the group who used the "dark" designed button and the group who used "light" designed button

Experiment

Result

Discoverability Measure	*Control Group (dark button)*	*Treatment Group (light button)*
Sample Size	**12904**	**12987**
% users who clicked the button	2836/12904	2497/12987
Average button clicked	5209/12904	4787/12987
% users who sent msg	1191/12904	1472/12987
Average msg sent	2246/12904	2518/12987

Usage Measure	*Control Group (dark button)*	*Treatment Group (light button)*
Sample Size	**12904**	**12987**
% users who sent at least 1 msg	1633/12904	1885/12987
average msg sent to Oscar	3482/12904	3839/12987
average msg received by users	3633/12904	3882/12987
% users who received at least 1 msg	6824/12904	7154/12987

Significance Analysis

	Metric	z-score	p-value	if significant(5% significant level)	Bonferroni Correction (5%/8 = 0.00625)
Discoverability	% user clicked button	5.47	4.44E-08	TRUE	TRUE
	% user sent msg	-5.57	2.48E-08	TRUE	TRUE
	average clicks	2.57	0.01	TRUE	FALSE
	average msg sent	-2.2	0.028	TRUE	FALSE
Usage	% user sent at least 1 msg	-4.27	1.27E-05	TRUE	TRUE
	% user received at least 1 msg	-3.55	0.00E+00	TRUE	TRUE
	average msg sent	-1.98	0.047	FALSE	FALSE
	average msg received	-1.63	0.1	TRUE	FALSE

(All results are considered "significant" under the significant level 5%)

Discoverability

1. "light" button has higher proportion of users who sent messages and higher average message sent;
2. "dark" button has higher proportion of users who clicked, and higher average user clicks;
3. All results are statistical significant (without Bonferroni Correction)

Usage

1. "light" button has higher proportion of users who sent at least 1 message, higher proportion of users who received at least 1 message, and higher average message received.

Summary

- "Light" button doesn't increase the discoverability of the message button, but it does result in more users who sent messages, and we also have more messages sent using "light" button
- "Light" button does increase the usage of message, as it increased the number of users who sent and received messages, and also increase the total number of message sent.

**Sample size could be calculated using significant level, power and minimum detective difference, however, I was not able to do it as I don't have the minimum detective difference.*
**Bonferroni Correction is also need to implemented as I was testing a group of metric using one sample data, but I didn't know how to implemented correctly, so I won't report the results using Bonferroni Correction here.*