UNIVERSITY OF CALIFORNIA

Los Angeles

Application of Experimental Design and Analysis in a Social Network Mobile App

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Muxin Shang

2019

ABSTRACT OF THE THESIS

Application of Experimental Design and Analysis in a Social Network Mobile App

by

Muxin Shang
Master of Applied Statistics
University of California, Los Angeles, 2019
Professor Hongquan Xu, Chair

The objective of this research is to apply techniques of randomized controlled experimentation including experimental design and analysis in user experience optimization and mobile app development. Online controlled experiments started to be used in the late 1990s with the growth of the Internet. Nowadays, many Internet companies leverage controlled experiments, especially A/B testing, to understand and make decisions at every step of product development. Large sites, including Facebook, Google, and Airbnb, run hundreds of A/B testing every month on features from UX design to algorithms for growth.

The power of running controlled experiments is the ability to establish causal inference and quickly validate new product ideas with statistical evidence of success. This process can ultimately help organizations, especially Agile software, optimize products with iterations by better understanding the corresponding impact on the user experience and recognizing the best performer from a list of variations.

The thesis of Muxin Shang is approved.

Nicolas Christou

Frederic R. Paik Schoenberg

Hongquan Xu, Committee Chair

University of California, Los Angeles

2019

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Professor Hongquan Xu for his guidance and advice in the preparation and finalization of this thesis. I would like to thank my thesis committee members for all of their guidance through this process.

CHAPTER 1

Introduction

This paper aims to describe the entire process and methodology of how a controlled experimentation or A/B testing is applied at a mobile app to improve a specific feature from experimental design to result based decision making. The mobile app is a social network app for meeting new people and available for download in Apple store and Android store. It allows users to rate other users′ profiles, and chat with them once both parties like each other. The A/B testing process follows the scientific method and tests on multiple variations by showing different UX design variations to users to determine the best performer based on pre-defined key metrics. Effective experimental design requires the business problems to be answerable and the results to be measurable. A reasonable A/B testing process usually follows these steps in the cycle chart below.

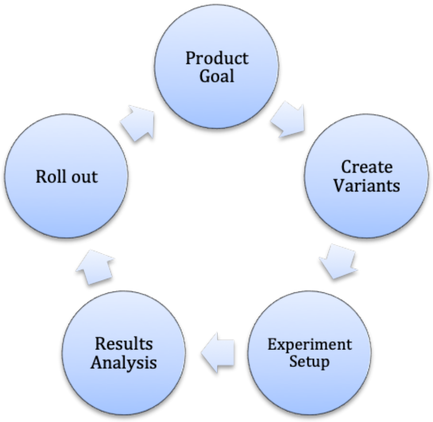


Figure 1.1: A/B Testing Framework

The term A/B testing has been used to refer to controlled experimentation. There are multiple test methods to structure experiments, such as A/B testing, multivariate testing (factorial design), and fractional multivariate testing. Traditional A/B tests recommend testing one factor at a time because it requires fewer samples and hence runs faster, whereas multivariate testing lets you test combinations of variables by isolating the impact of external factors and comparing them against each other in every possible combination. Comparing to running a series of A/B tests one at a time, multivariate testing studies effects of multiple factors and understands the impact on the product ecosystem individually and interactively. The combined effect of two factors may be different from the sum of two individual effects. It is also efficient since it runs in a shorter time to find optimal variables from a list of factors at a single test. Fractional multivariate testing should be considered when strategically there is no interest in a specific set of combinations of variables for a test. It is more practical for high-traffic sites or apps to use the multivariate testing method since it is relatively fast for them to get enough samples for a broad set of variations.

This paper will describe a real application of multivariate A/B testing in the social network app for a project that I led the analytic work providing actionable insights and recommendations based on scientific methods including A/B testing. For this project, I collaborated with both product team and engineering team completing a series of A/B testing experiments using the framework in the Figure 1.1.

CHAPTER 2

Methodology

2.1 Hypothesis Testing

In an ideal world, we would know everything about the population, hence no estimation is necessary. However, in real-world businesses, there are limitations we need to work around. In A/B testing we are limited by the time, resources and number of users we can assign to any given test. Thanks to hypothesis testing methodology, we can make statistical inferences of population parameters on a random sample.

Hypothesis testing is making inference about the relationship between two populations to determine if there is a statistically significant relationship or not. There are two types of hypotheses, the null (H_0) and alternative hypotheses (H_1). Statistical significance means that the observed difference between two samples is caused by something other than chance or random sampling error only. We choose to reject the null hypothesis to declare statistical significance when the p-value is below a pre-specified value α, the probability of conducting Type I error. There are two possible correct decisions and two possible errors in hypothesis testing, as shown in Table 2.1 below.

	H_0 is True	H_0 is False
Reject H_0	Type I error	Correct decision
Do not Reject H_0	Correct decision	Type II error

Table 2.1: Error Types of Hypothesis Testing

The p-value gives the probability of observing an effect from a sample under the null hypothesis. It provides a quantitative measure of statistical significance to determine the observed difference in hypothesis testing. A p-value of less than 0.05 is considered as declaring statistical significance under the conventional threshold. [4] How do we compare two samples in terms of their distributions? The test parameters we choose determine the test statistic for testing the difference in two populations. In general, we would compare the difference in sample means or sample proportions for statistical testing. A sample mean is the average value of a sample and the sample proportion is the amount of the sample that results in success. The proportion takes a value from 0 to 1 measuring probability metrics such as conversion rate. The sample mean and sample proportion allow to compare changes of variables. In general, an A/B testing comparing sample means would define a null hypothesis assuming two sample means equal, and an alternate hypothesis assuming two sample means not equal.

2.2 Multiple Comparisons

Hypothesis testing is useful to prove hypotheses about our data. This method runs into the problem of multiple comparisons when we want to test a set of m hypotheses simultaneously. In reality, to better understand what elements have significant impacts to the product when running an experiment, breaking the samples into segments is necessary in order to dive deeper and get more actionable insights of the data, for instance, what specific group of people adopts different behaviors due to the product changes. Examples of segments are gender, age, and regions of the users. In addition to the numbers of metrics tested, this segmentation will further introduce more simultaneous comparisons.

When testing a set of hypotheses simultaneously, we are also dealing with the problem of multiple comparisons, because the probability of observing at least one pair of treatments significantly different at level α is larger than α. For example, we have m=20 hypotheses to test at the significance level of 0.05, the probability of getting at least one false positive just due to chance (experiment-wise error rate) is

$$\begin{aligned}\Pr(\text{at least one significant result}) &= 1 - \Pr(\text{no significant results}) \\ &= 1 - (1-\alpha)^m \\ &= 1 - (1-0.05)^{20} \\ &\approx 0.64\end{aligned}$$

Thus, with 20 hypotheses being tested at the same time, we have a 64% chance of getting at least one significant result, even if all of the comparison results are not significant. As shown in the figure above, this probability increases as more hypotheses being tested simultaneously. It is common for a sophisticated app to run many hypotheses simultaneously at one test, in order to monitor a set of key metrics over multiple segments. One of the approaches to fix this problem is to adjust α, so that the experiment-wise error rate remains below the desired significance level. One convenient method to use to control the experiment-wise error rate is the Bonferroni method. [4]

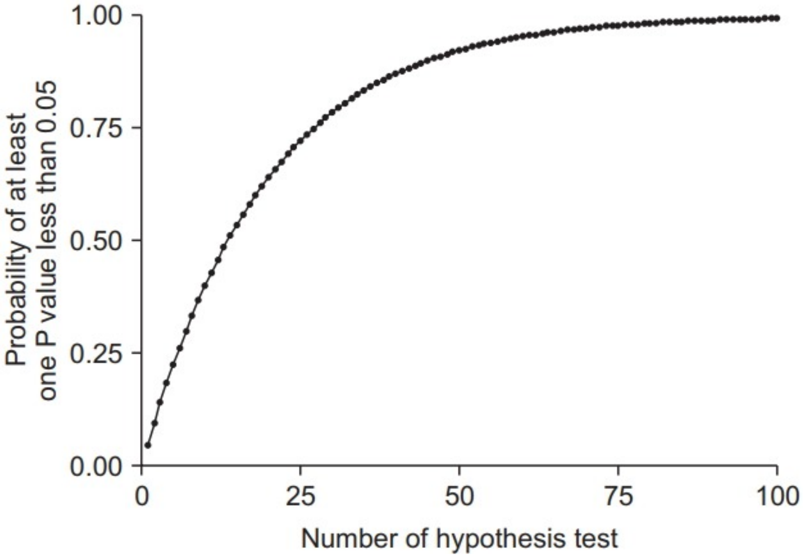


Figure 2.1: Increasing Error Rate of Multiple Comparisons

The Bonferroni correction cuts off the significance α by the number of comparisons m at α/m, so that a null hypothesis should be rejected if the p-value is less than α/m, instead of α. As a result, experiment-wise error rate is corrected as the following such that it is now close to the pre-defined significance level at $\alpha = 0.05$.

$$\begin{aligned}
\Pr(\text{at least one significant result}) &= 1 - \Pr(\text{no significant results}) \\
&= 1 - (1 - \tfrac{\alpha}{m})^m \\
&= 1 - (1 - \tfrac{0.05}{20})^{20} \\
&\approx 0.05
\end{aligned}$$

CHAPTER 3

Experimental Setup

3.1 Set Product Goal

This particular experiment that I designed and analyzed was to focus on improving user interactions with an in-app new feature called "Location". It is a location-based feature collecting public locations, such as restaurants and shopping malls, which a user has visited recently. It allows the user to view profiles and connect with a list of matches who have been to the same locations, which hopefully help them find more common interests and topics effortlessly. The actual UX design for this feature page is shown in the Figure 3.1. A number of locations are ranked based on the recency of users′ visits. The first card is always the most recently visited location, and the user can view more other location cards by swiping the carousel, with one card centered at a time.

The team launched this feature in a small market and planned to iterate the feature with a series of A/B testing to optimize this feature before rolling it out globally. I designed and evaluated all of the experiments for this project. My analyses and recommendations supported the team′s product decisions. This particular test discussed in this paper was one of the series of A/B testing and was designed as multivariate testing. We eventually rolled out the winning variant globally based on my analysis for this test. The following will go through the details behind that experiment using the framework.

The product problem was discovered through one of my exploratory analyses on users′ adoption with this new location feature after launching it in a small market. The analysis pointed out relatively lower usage of the feature, based on location card open rate and the number of location cards opened per user on average.

The goal of the feature is to help users find more relevant matches through viewing other users′ profiles in locations they visited. The relatively low card opens led to lower activities rating other users in those cards and hence undermined the feature effect to the app ecosystem. Thus, the team wanted to find a more effective UX design to improve the funnel drops and enable users to interact with more location cards.

3.2 Create Variants

For the organization, the product problem is the funnel drops on the conversions from the "Location" entry page to card openings. Thus, the team′s objective is to explore factors and factor levels that can get users to open more of location cards from the entry page.

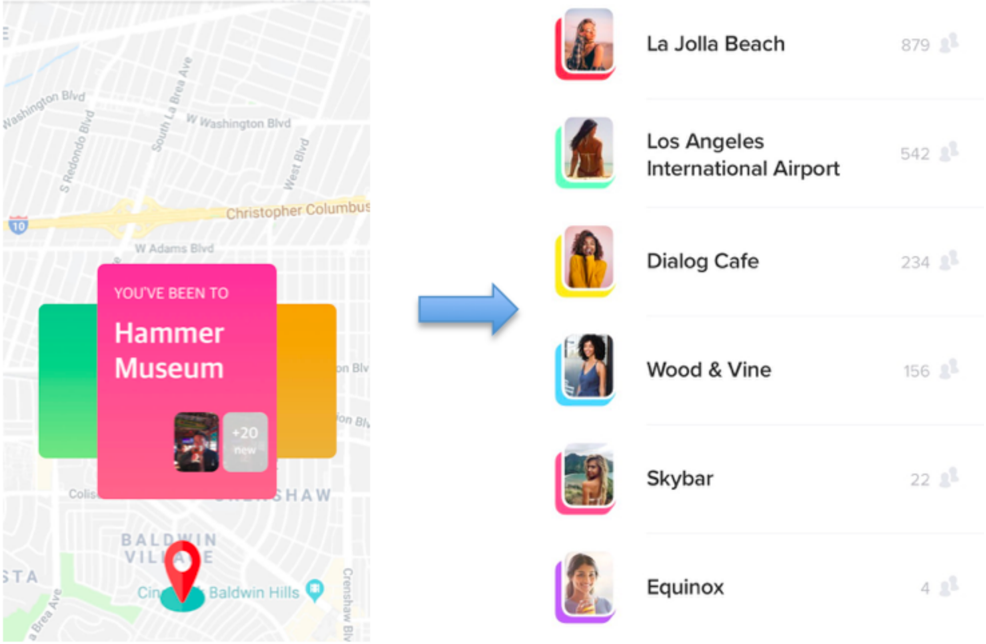
The first step is to find out the main factors that would potentially result in the highest increase in conversions. Theoretically, there are many elements on this entry layout page that we could optimize such as the entire layout, the display of the card carousel, the design of the location card, and even the theme color. We could only test a limited amount of element options under the timeline and the budget cost, including the products roadmap timeline and engineering cost. Thus, the best option for us was to find out the most impactful elements to test. Through discussions with the product team and the design team, we decided to test the page layout UX design and the sorting algorithms displaying the location cards. We have determined that those two factors were the most impactful elements among all the possible ones on this page.

The next step is to then discover appropriate levels for each factor. It is impossible to test large amounts of levels due to limited resources in any organization. Based on the team′s experience and product sense, we strategically selected two levels for the layout UX design and four levels for the card sorting algorithms.

	Level			
Factor	Original	1	2	3
A. UX layout	Map horizontal carousel	Vertical List		
B. Card sorting algorithm	Recency	Recency without empty cards	Distance	Number of profiles in a card

Table 3.1: Factors and Levels

As shown in Table 3.1, the two layout UX designs are the original map carousel design and the new vertical list design. The four card sorting algorithms are the original recency and the three new ones including recency with empty cards removed, distance, and the number of profiles in a card. One example of the actual UX designs from control to variant groups for this test is shown in Figure 3.1 below.



La Jolla Beach	879
Los Angeles International Airport	542
Dialog Cafe	234
Wood & Vine	156
Skybar	22
Equinox	4

Figure 3.1: Design of Variants

As shown in Table 3.1, this experiment consists of two factors with the layout UX at two levels and the sorting algorithm at four levels, and it generates total possible combinations of eight test groups. Full factorial testing or multivariate testing is considered as the experiment method to test on every possible combination available. Instead of running two separate single factor experiments, running the experiment for the two factors with a combination of all levels at once would allow us to learn the interaction effects over all levels chosen. The goal is to figure out what specific elements on the product page play the most important role in achieving business objectives. Overall, this experiment will have 8 test groups. Unlike traditional in-person experimentation associated with the pharmaceutical industry, Internet companies are usually not limited by the expense of getting enough samples. They can run tests and assign users randomly into a test group when the users log into the app or site successfully. Given that this mobile app has enough traffic, the team is open to running a number of variations without worrying about running insufficient samples and slowing down the test as a result.

	Factors	
Group	UX Layout	Sorting Algorithm
Control	Map horizontal carousel	Recency
Variant 1	Map horizontal carousel	Recency without Empty Cards
Variant 2	Map horizontal carousel	Distance
Variant 3	Map horizontal carousel	Number of Profiles in a Card
Variant 4	Vertical List	Recency
Variant 5	Vertical List	Recency without Empty Cards
Variant 6	Vertical List	Distance
Variant 7	Vertical List	Number of Profiles in a Card

Table 3.2: Planning Matrix

3.3 Define Key Metrics

The first step of planning A/B testing is to define the key metrics measuring the success of the test. It comes down to the product goal. Whether an experiment should be taken as successful or not largely depends on the product objectives. They are the factors the decisions are based upon. So a determination of objectives is necessary before any experiments. Clear objectives help decide clear KPIs (key performance indicator) or key metrics for the test.

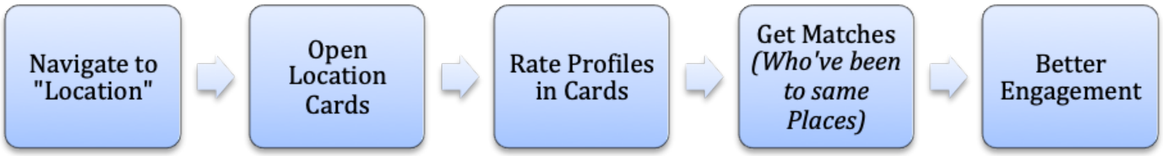


Figure 3.2: Feature Engagement Flow

Figure 3.2 is the product funnel flow for this new “Location” feature. In order to increase the result of getting more location relevant matches, we would have to improve the conversions at the upper funnel, which is getting more users to open and then swipe in the “Location” cards. Since we have got the optimal variations to test based on our capacity, the key metrics should measure the results from the variations. The metrics that matter to the team are:

- Average number of location card opens
- Location card open coverage
- Average days active in “Location”

The two goals we want to achieve are more users open a location card and more location cards opened on average. We believe that the list layout design will naturally increase more cards viewed than the horizontal carousel design with one card being centered at one time. Location card coverage allows us to see if more users open at least one location card, and the average number of location cards opened allows us to see if users open more location cards during a period of time. Average days active in “Location” attempts to measure the ecosystem impact to this feature from the variations. It allows us to see if users are using this “Location” feature more frequently. The optimal variation would ultimately lead to significant lifts in all of the three key metrics above. We would not want to measure and compare more metrics than what we need because more hypotheses tested simultaneously will decrease the statistical power and lead to longer test length.

3.4 Determine Minimum Sample Size

An appropriate sample size is crucial to an A/B testing. We expect large samples to achieve more reliable results and we also want to run experiments faster due to resources limitation. However, small sample size decreases statistical power. Power is the probability of rejecting the null hypothesis when it is false. The power of a test is its ability to detect an effect when there is one to be detected. Power analysis can be used to estimate the minimum sample size required for A/B testing with pre-defined significance level, effect size, and statistical power. In general, for every hypothesis test or A/B testing, we'll want to do the following to estimate effective sample size.

- Minimize the probability of committing a Type I error, α. Typically, $\alpha = 0.05$ is used as the convention threshold for significance level
- Maximize the power, $1-\beta$. A power of 0.80 or greater is typically the convention threshold

The power analysis can be used to estimate minimum sample size, and it is calculated using the following formula: [1]

- Minimum sample size n for difference in means: n = $\frac{2\sigma^2(Z_{1-\beta}+Z_{\alpha/2})^2}{(\bar{X}_1-\bar{X}_0)^2}$
- Minimum sample size n for difference in proportions: n = $\frac{2\bar{p}(1-\bar{p})(Z_{1-\beta}+Z_{\alpha/2})^2}{(p_1-p_0)^2}$

Based on the formulas above, we can get minimum sample size for each of the key metrics, based on our pre-defined significance level of 5% and power of 95%. We take the largest sample size in order to get enough samples for any of the metrics. From the Table 5 below, we can conclude that each experiment group requires at least 5,000 samples to get enough power for statistical significance for the key metrics.

Key Metric	Historical performance	Expected lift	Effect size	Minimum sample size
Average number of location card opens	5	+10%	0.5	2000
Location card open coverage	50%	+5%	0.025	5000
Average days active in Location	3	+2%	0.06	5000

Table 3.3: Key Metrics and Sample Size

3.5 Launch Experiment

One of the most essential parameters to know when launching a test is the test duration, which is based on the calculated minimum sample size. To avoid calling conclusive results too quickly, we need to estimate how long the test should run given the sample size based on the historical traffic. In addition to be able to reach right conclusion, it is also important for the team to know the test length and collaborate with other teams on A/B testing calendar so that we do not run multiple related tests in the same market simultaneously. In this case, the test consists of 8 variants, and each of the variants requires at least 5,000 samples to detect statistical significance with enough power. This gives us of 40,000 users as a minimum sample size. As a result, we decided to run this A/B testing for 14 days.

When turning on the test on the setup page, one important configuration is equal size distribution for each variant group, which is 1/8 in this test. In general, the best approach to split the traffic for each variant group in A/B testing is to split it equally across the test groups. Equal split not only helps in reaching the statistical significance faster but also gives a more reliable result. We are sometimes concerned about allocating huge traffic to a variant as it could lead to a negative impact, so it is safer to only assign the number of users based on the minimum sample size if we want to be more conservative about the test.

When choosing a target experimental group, we would usually want to test users globally, since the mobile app is used worldwide. Sampling users globally would help reduce bias towards certain regions and be more representative of all population. It is important to make sure to isolate this experiment from all others that test on the same feature page at the same time. It is common for a large organization to run hundreds of experiments on any given day. We just need to make sure to communicate and coordinate ahead with other teams for a clean experiment environment. It is also very important to sample randomly for statistical inferences.

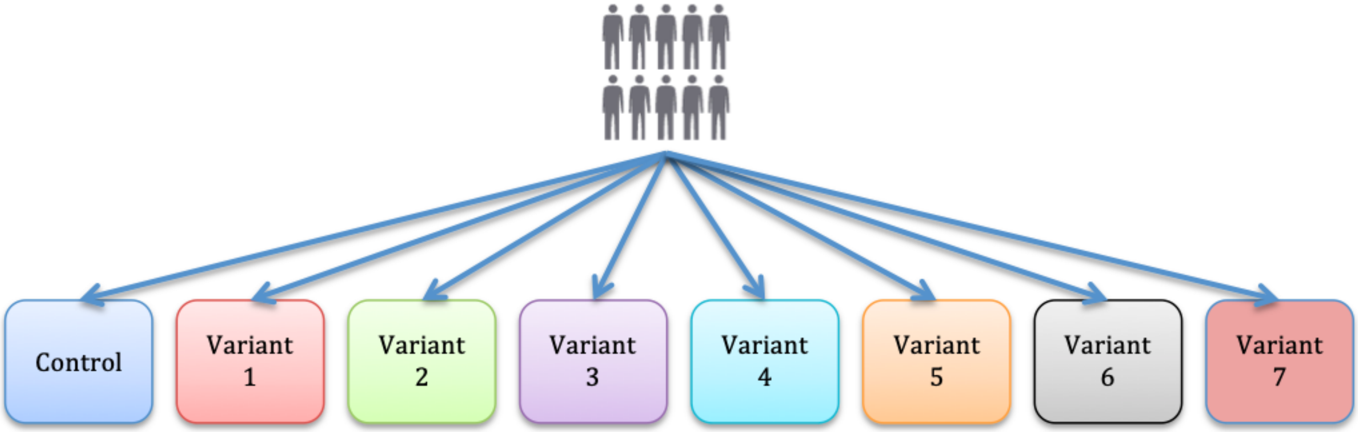


Figure 3.3: Equal Split Test Assignment

CHAPTER 4

Results Analysis

4.1 Results Summary

The analysis script is set to run two weeks from the test start day. Users were assigned equally across the eight test groups. As expected, we had enough users for each test group at the end of two weeks. For the test parameters, we used the standard significance level of 0.05. Since we ran 21 comparisons for the seven variants and three key metrics simultaneously, we corrected the significance level to be 0.05/21=0.0025 based on the Bonferroni Correction method. The following is the summary tables of the significant changes from this test.

variant	baseline	metric	significant changes	winner (Y/N)
Variant 1	Control	Mean location cards opens	—	N
Variant 2	Control	Mean location cards opens	+1.5%	N
Variant 3	Control	Mean location cards opens	+4%	N
Variant 4	Control	Mean location cards opens	+6%	N
Variant 5	Control	Mean location cards opens	+10%	N
Variant 6	Control	Mean location cards opens	+12%	N
Variant 7	Control	Mean location cards opens	+15%	**Y**

Table 4.1: Location Card Opens

variant	baseline	metric	significant changes	winner (Y/N)
Variant 1	Control	Location card open Coverage	+1%	N
Variant 2	Control	Location card open Coverage	+1%	N
Variant 3	Control	Location card open Coverage	+3%	N
Variant 4	Control	Location card open Coverage	+4%	N
Variant 5	Control	Location card open Coverage	+6%	N
Variant 6	Control	Location card open Coverage	+8%	N
Variant 7	Control	Location card open Coverage	+10%	**Y**

Table 4.2: Location Card Open Coverage

variant	baseline	metric	significant changes	winner (Y/N)
Variant 1	Control	Mean days active in “Location”	—	N
Variant 2	Control	Mean days active in “Location”	—	N
Variant 3	Control	Mean days active in “Location”	—	N
Variant 4	Control	Mean days active in “Location”	+1%	N
Variant 5	Control	Mean days active in “Location”	+1.5%	N
Variant 6	Control	Mean days active in “Location”	+2%	N
Variant 7	Control	Mean days active in “Location”	+4%	**Y**

Table 4.3: Days Active in “Location”

4.2 Learnings

As we can see from the significance summary table above, the clear winner variant is the variant 7, UX list + sort result by the number of profiles. Location cards open coverage is a metric measuring how good the UX entices users to interact with the location cards at least once. The goal is to drive more users to engage and drive users to engage more. The large increase on the cards open coverage in variants indicates that users are more interested in exploring the location cards in depth when they come to the feature. This would confirm our hypotheses about this product that more relevant location cards being shown first will increase users engagement. More specifically, we can see that the lifts among variant 4 to variant 8 are much higher than the ones among variant 1 to variant 3. It seems that the new sort algorithms alone performed slightly better than the original, and they performed much better when combing the UX list layout design. By comparing variant 4 and the control, where the only difference is the UX layout change, the lift is 4% just due to the UX list. This seems to suggest that the list layout is effective at getting users to interact with location cards. It is probably more efficient to find cards that users show high interest in from a long list. This seems to verify our hypothesis about the designs.

The significant lifts on days active in “Location” seem to suggest that users who have experienced the new designs and changes become stickier with the feature and retain more frequently. This suggests positive impacts on the app ecosystem because stickier with the “Location” feature will likely ultimately get them to use the app more often. The 4% lift from the variant 7 suggests a big success, as increase on retention is usually hard to achieve. These new designs not only improve the feature usage significantly but also make the feature work better for users. We believe that matches received through “Location” should be more relevant than ones received through the regular core stack. More user engagement with “Location” could lead to more relevant matches received, and it will hopefully ultimately lead to better conversations. So the increase on this days-active metric shows a strong signal that this test is successful in optimizing the current feature in terms of better engagement.

Overall, based on the key metrics above, we have learned that both UX list and sort results together did lead to much higher lifts than each of the two factors alone. Comparing to the original recency sort result, recency without empty cards is slightly more effective for engagement. It seems to suggest that users would be engaged more with this feature by removing the empty cards. The sort result by distance is more effective than both the original and the recency without empty cards. It seems to suggest that users are more interested in engaging with closer distance locations than more recent visited locations. Most importantly, users are interested most in cards that have many profiles to view and rate. This would verify our hypothesis that the most important factor influencing users behaviors in this feature remains the number of profiles, not recency and distance. It creates a more efficient product environment that users can still consume large volumes of profile content in a session and the location information of the profiles makes this experience more interesting.

4.3 Additional Analysis: Two Way ANOVA and Interactions

A factorial experiment with two factors A and B, with a levels and b levels, respectively, has $a \times b$ treatments. The general ANOVA model, with r replicates for each treatment, can be written as [2, 4]

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, r$$

where μ is the overall mean response, τ_i is the treatment effect due to the i-th level of factor A, β_j is the treatment effect due to the j-th level of factor B, γ_{ij} is the interaction effect between the i-th level of A and the j-th level of B, and ϵ_{ijk} is the error residual. The resulting ANOVA table for a two-way factorial experiment is the following table

Source	Degrees of Freedom	SS (Sum of Squares)	MS (Mean Square)	F Statistic
A: UX Layout	a-1	SS_A	MS_A	MS_A/MS_{within}
B: Sort Result	b-1	SS_B	MS_B	MS_B/MS_{within}
A×B	(a-1)(b-1)	SS_{AB}	MS_{AB}	MS_{AB}/MS_{within}
Within	ab(r-1)	SS_{within}	MS_{within}	
Total	abr-1	SS_{total}		

Table 4.4: ANOVA Table

The total sum of squares can be partitioned as [2, 4]:

$$SS_{total} = SS_A + SS_B + SS_{AB} + SS_{within}$$

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(Y_{ijk}-\bar{Y}_{...})^2 &= r\cdot b\cdot\sum_{i=1}^{a}(\bar{Y}_{i..}-\bar{Y}_{...})^2+r\cdot a\cdot\sum_{j=1}^{b}(\bar{Y}_{.j.}-\bar{Y}_{...})^2\\ &+ \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...})^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(Y_{ijk}-\bar{Y}_{ij.})^2\end{aligned}$$

Since this experiment has quantitative outcomes and two categorical factors with subjects being exposed to any combination of one level of the two variables, two-way ANOVA is an appropriate method to analyze the test data. This experiment has two factors at two levels for UX layout and four levels for sort result. ANOVA will be conducted for each of the three metrics for this factorial experiment.

4.3.1 Number of Location Card Opens

There are three types of null hypotheses of interest as follows:

- H_0: Number of location card opens does not depend on the type of UX layout
- H_0: Number of location card opens does not depend on the type of sort result
- H_0: There is no interaction between UX layout and sort result

There are 2×4=8 different combinations of UX layout and sort result. We take 100 replicates for each combination of the factors randomly from the entire samples. The following is the resulting summary table for the number of location card opens.

Source	Degrees of Freedom	SS	MS	F value	Pr(>F)
A: UX Layout	1	224	223.66	43.88	<6e-11***
B: Sort Result	3	158	52.67	10.33	<1e-06***
A×B	3	41	13.51	2.65	0.0477*
Residuals	792	4037	5.1		

Table 4.5: ANOVA Table for Number of Location Card Opens

As we can see from the ANOVA table above, the p-values of UX layout and sort result are both very significant, which indicate that the levels of UX layout are significantly associated with number of card opens, and the levels of sort result are significantly associated with number of card opens. The p-value for the interaction between UX layout and sort result is nearly 0.05. This significance will most likely be gone once adjusting significance level based on the number of comparisons.

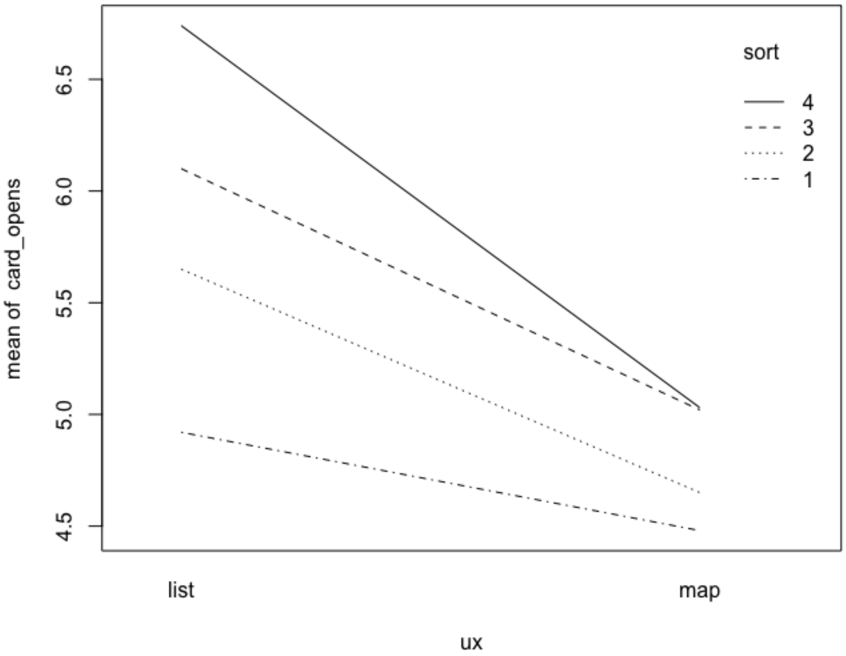


Figure 4.1: Interaction Plot for Card Opens

From the interaction plot, we can see that the lines are nearly parallel, suggesting that there is no interaction between UX layout and sort result. This aligns with what we see in the summary table.

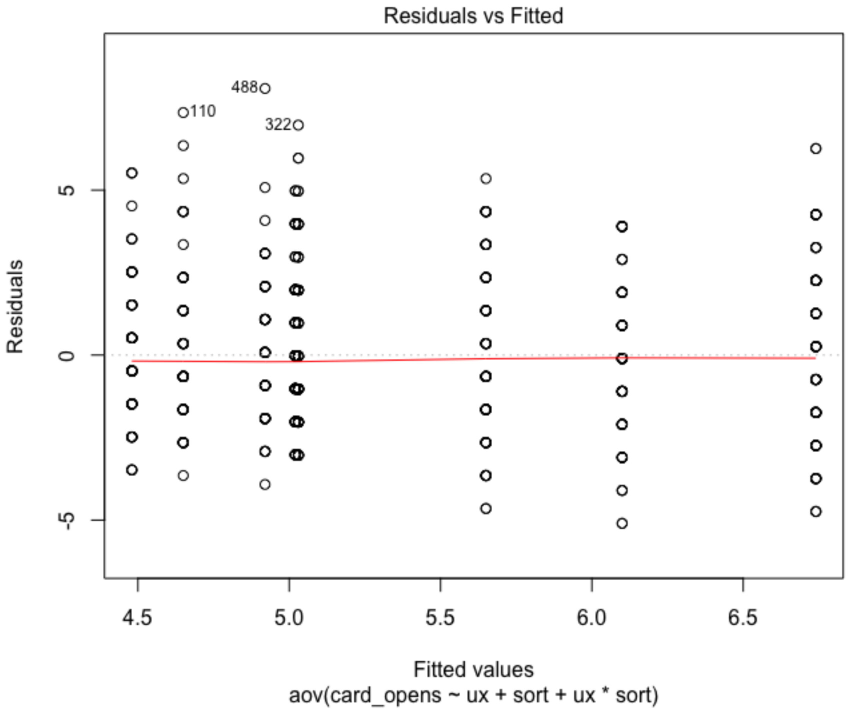


Figure 4.2: Residual Plot for Card Opens

The residual plot seems to suggest that there are no clear outliers and the variances of the residuals seem equal, so this model is acceptable.

4.3.2 Location Card Open Coverage

There are three types of null hypotheses of interest as follows:

- H_0: The coverage of opening a location card does not depend on the type of UX layout
- H_0: The coverage of opening a location card does not depend on the type of sort result
- H_0: There is no interaction between UX layout and sort result

Similarly, this time we take 100 replicates to get higher power for each combination of the factors randomly from the entire samples. The following is the resulting summary table.

Source	Degrees of Freedom	SS	MS	F value	Pr(>F)
A: UX Layout	1	2.2	2.205	9.218	0.00248**
B: Sort Result	3	3.07	1.0233	4.278	0.00524**
A×B	3	0.15	0.0483	0.202	0.895
Residuals	792	189.46	0.2392		

Table 4.6: ANOVA Table for Location Card Open Coverage

As we can see from the ANOVA table above, the p values are significant, and it seems to suggest that the levels of UX layout and the levels of sort are significantly associated with number of card open coverage. The p-value for the interaction between UX layout and sort result is not significant, suggesting that there is no interaction between UX layout and sort result.

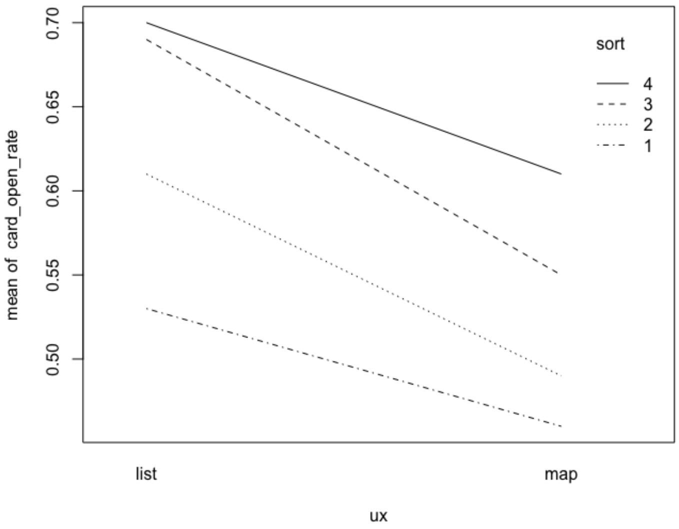


Figure 4.3: Interaction Plot for Card Open Coverage

From this interaction plot, we can see that the lines are nearly parallel, suggesting that there is no interaction between the two factors. List UX clearly performs better than map, and sort 3 and sort 4 perform better than sort 1 and sort 2. This aligns with what we expect because sort 2 is slight change based on sort 1 and sort 3 and sort 4 are visually different.

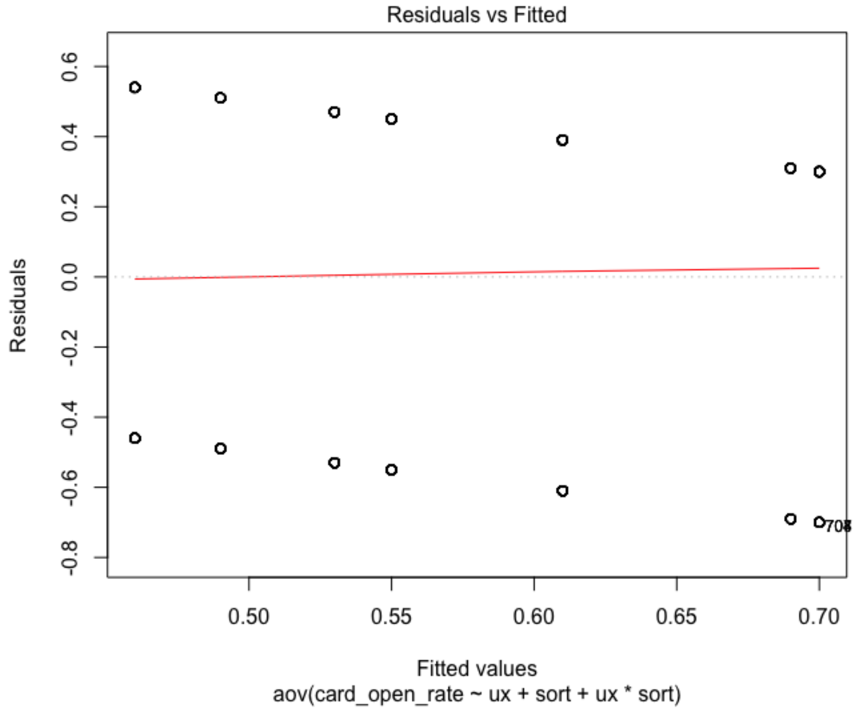


Figure 4.4: Residual Plot for Card Open Coverage

In this residual plot, there are no clear outliers and the variances of the residuals seem equal. So this model is acceptable.

4.3.3 Days Active in "Location"

There are three types of null hypotheses of interest as follows:

- H_0: The number of days active in "Location" does not depend on the type of UX layout
- H_0: The number of days active in "Location" does not depend on the type of sort result
- H_0: There is no interaction between UX layout and sort result

Similarly, we take 100 replicates for each combination of the factors randomly from the entire samples. The following is the resulting summary table.

Source	Degrees of Freedom	SS	MS	F value	Pr(>F)
A: UX Layout	1	15.7	15.68	5.22	0.0226*
B: Sort Result	3	29.3	9.77	3.25	0.0212*
A×B	3	6.9	2.31	0.77	0.51
Residuals	792	2377.7	3		

Table 4.7: ANOVA table for Days Active in “Location”

The ANOVA table suggests that both the levels of UX layout and the levels of sort result are significantly associated with days active in Location, whereas there is no interaction between UX layout and sort result.

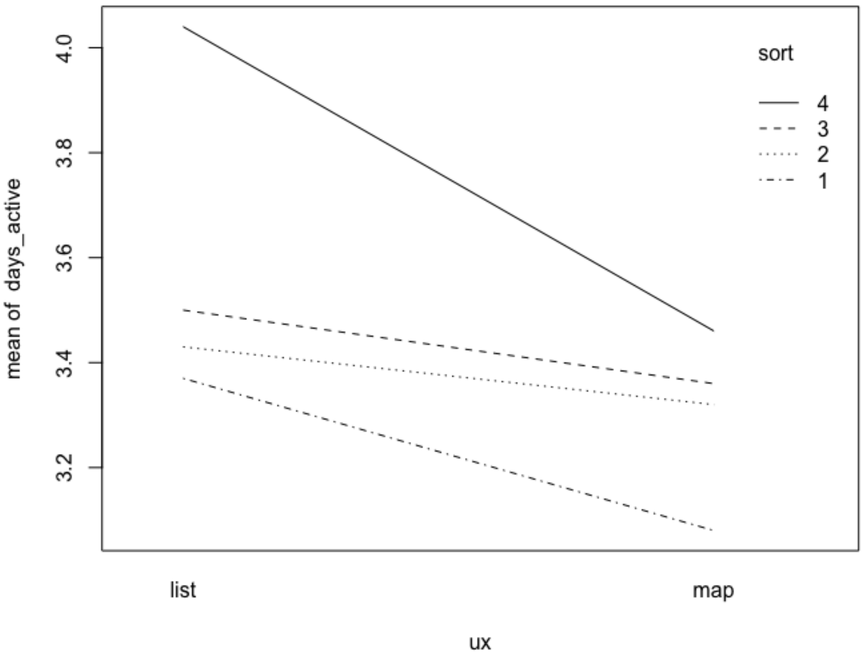


Figure 4.5: Interaction Plot for Days Active

The interaction plot shows that the lines are nearly parallel, suggesting that there is no interaction between the two factors. This aligns with the result from the ANOVA table for the interaction effect. For the same sort result, such as sort 1 and sort 4, the responses change significantly more when the level of UX is list. The changes between sort 1 and sort 4 are larger than the ones between sort 2 and sort 3 because sort 2 removes empty location cards and sort 3 is more similar to sort 2 than sort 1. This seems to still suggest that the combination of UX list and sort by the number of profiles is still the winner in terms of the days active metric.

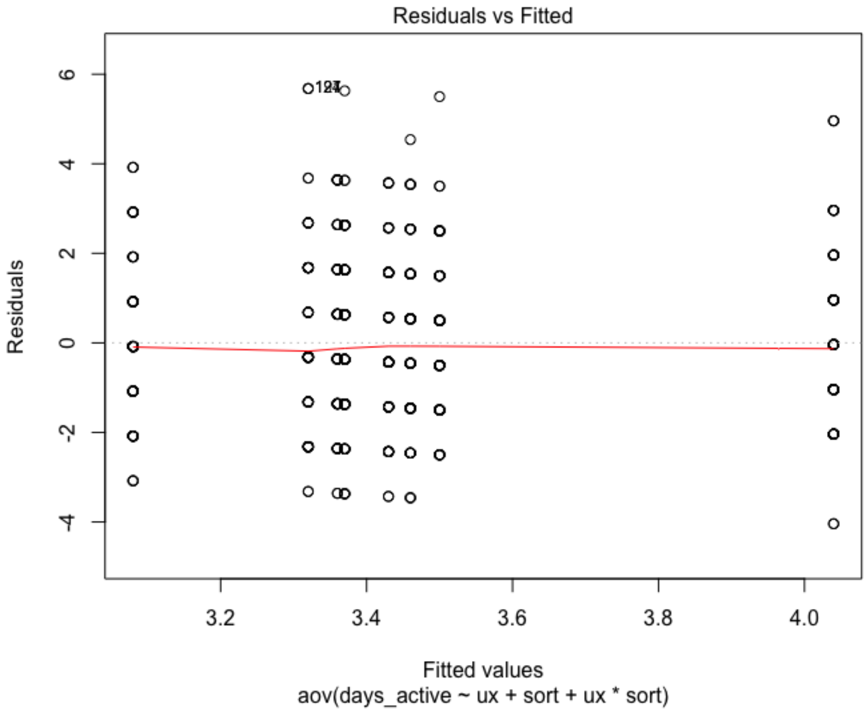


Figure 4.6: Residual Plot for Days Active

In this residual plot, there are no clear outliers and the variances of the residuals seem equal, suggesting that this model is acceptable.

CHAPTER 5

Summary

5.1 Challenges

Typical A/B testing assumes that individual units are independent, which is known as Stable Unit Treatment Value Assumption (SUTVA) [3]. However, for A/B testing in social network products, the assumption of individual independence might be violated inexplicitly due to interactions between users, and this could result in inaccurate inference of the causal effect of treatments. For example, users might chat about their experience using the Location feature after they become matches through the feature. The users might feel confused about hearing completely different UX designs in the same feature, and result in fewer interactions or more interactions with the feature. There are different possible methods to deal with this issue. Since we have not yet settled on one specific method that works effectively in our mobile app context and this specific test does not suffer as much as a chatting feature does, we decide to ignore network effect factor in this A/B testing analysis. There seems to be no single simple solution yet to deal with the network issue, but it is worth putting more efforts into research and explore the best approach to understand the impact of the network issue for future testing.

In addition, due to our technical limitation, we only assign users when they log into the app. The "Location" feature requires users to navigate to the feature main page, which would result in only a portion of the assigned test users ever entering into this feature. This would undermine the statistical power if we directly compare test results for all of the test users, since many of them in this test never even tried the feature. If the true effect size is small, the effect could be diluted so that it is harder to detect statistical significance when there is. A cleaner way to run this test is assigning users at random when they navigate to the feature page. In order to better estimate the treatment effect treated, we only analyzed users who have viewed the feature at least once during this test. This approach is based on the randomized assignment (users get assigned into a test group at random when they log in) so that users in any of the test groups navigate to “Location” feature at random. This method seems to work well for us in this case after I have investigated the filtered user cohort and ensured that each group still maintains the same proportions of subgroup, such as gender and city.

5.2 Conclusion

All of the analyses suggest that the two factors have positive impacts on the key metrics. As a result, the team has decided to roll out the new changes, UX list and sort result by the number of profiles, globally. These studies from this experiment are impactful on the team′s product strategies towards the feature of future iterations. We are positive that we will reach our product goals through a series of experimentations. Statistics certainly play a very important role in making informed decisions in new feature developments in the organization. As the data volumes grow rapidly every day, more data scientific methods are needed for better exploring the large datasets, understanding users′ preferences and ultimately identifying business opportunities.

Bibliography

[1] S K Lwanga and S Lemeshow. *Sample Size Determination in Health Studies: A Practical Manual*. Geneva: World Health Organization, 1991.

[2] D C Montgomery. *Design and Analysis of Experiments, 8th ed*. John wiley & sons, 2013.

[3] S Schwartz, N M Gatto, and U B Campbell. Extending the sufficient component cause model to describe the stable unit treatment value assumption (sutva). *Epidemiologic Perspectives & Innovations*, 9(1):3, 2012.

[4] C F J Wu and M S Hamada. *Experiments: Planning, Analysis and Optimization, 2nd ed*. John Wiley & Sons, 2009.