GloBox A/B Testing

Result and Recommendation Report

Joseph Udo

October 2023

Image courtesy of freepik

Table of Contents

Executive Summary

This report provides details of an A/B testing experiment conducted for GloBox, a rapidly expanding online store. The aim is to bring awareness to the new food and product category through the use of advert banner thereby increasing revenue. The report presents the context, planning and execution, analysis and recommendation based on the outcome of the test.

The experiment was conducted over a 13-day period for 48948 mobile users with a randomized control test into effectively even sampling control and treatment groups. The metrics under consideration were conversion rates and spend per user.

The initial analysis shows that the sample size is not adequately sufficient to draw a concrete conclusion based on the test outcome and there is a need to review the volume. Insights showed different user behaviour across different segmentations.

The test results showed that the new banner had significant impact on conversion rates but no substantial effect on the spend per user. The observations warranted the recommendation that to enhance the statistical power of the analysis and identify smaller differences, it is advisable to consider additional research, potentially with an expanded sample size. Extending the duration of the test is also recommended, all based on the assumption of no financial or service costs to the company.

Context

Introduction of Client

Globox is a fast pace growing online marketplace that specializes in sourcing unique and high-quality products from around the world.

The company is primarily known for boutique fashion items and high-end décor products. Nonetheless, their culinary and beverage selections have expanded significantly over the past few months, and the company aims to raise awareness about this product category to boost their income.

Group A: Control
existing landing page

Group B: Treatment
landing page with food & drink banner

Business Objective

The aim of the business objective is to ascertain if increasing awareness of the new product category will have an impact on the key metrics managed by the Growth Team i.e., user conversion and revenue.

Globox intends to launch a user experience with a new banner that shows the key food and drink products (placed at the top of its mobile homepage for maximum visibility) to impact

user behaviour over time. Initially, it will conduct a randomized controlled test (RCT) in the form of an A/B test for the new banner.

The A/B test plays a pivotal role in this strategy. Should the test reveal that the new banner consistently generates increased revenue, GloBox will allocate substantial resources to this domain and further broaden its range of growth endeavours within the food and beverage category.

Stakeholder Analyses

The table below itemizes the extent and prerequisites expected by the principal stakeholders participating in the new banner program, establishing the context for the A/B test.

Role	Domain	Function
Growth Product & Engineering Team (Data Analyst)	Data and the analysis. The team develops features for the GloBox website that drive growth in users and revenue.	Collect and analyse data to understand user behaviour. Run hypotheses testing. Communicate and make recommendations
Product Manager	High level insights. KPIs (metrics)	Setting goals for projects. Measuring their success against defined KPIs. Communicating results to other company leaders
UX Designer	Detail user experience design.	Conduct user research. Design experience currently being evaluated
Head of Marketing	Marketing and research	Identifying target audience to drive product usage and conversion. Communicate marketing plans

Methodology

Planning and Setup of A/B Test

Objectives of A/B Test

To acquire empirical evidence supported by statistical analysis from the test subjects and address the following queries:

- Will the new banner significantly and confidently result in an increase in conversion rate (CR) and spend per user ($/user)?
- Are there any learning effects such as change aversion or novelty effect?

Duration

The timeframe of the test was within a 13-day period, from January 25th to February 6th, inclusive.

Conversion

This occurs when a user makes a purchase (or more). A converted user subject must be unique and make one or mor purchases.

Metric setup

Propensity measure:

- Conversion rate (CR) = total conversions/ total user count.

Volume measure:

- Spend per user ($ / user) = total spend ($) / total user count.

Test Setup

- The experiment is only being run on the mobile website.
- A user visits the GloBox main page and is randomly assigned to either the control (Group A) or test group (Group B) also known as treatment group. Control group A will be the baseline.
- The point at which a user visits the Globox main page is the join date for the user.
- Only users in group B will see the new banner (that displays food and drinks items). The control group will only view the default page featuring the primary products.
- Conversion is not product category dependent.

Data Understanding

Data Collection

GloBox stores its data in a relational database. This data contains the purchasing activity data and personal data such as device type, gender and country.

Data Description

The database contains three tables (users, groups and activity) with different variables required for the test. The table below show a description of the different data tables.

users			**groups**			**activity**		
Field Name	*Description*	*Variable Type*	*Field Name*	*Description*	*Variable Type*	*Field Name*	*Description*	*Variable Type*
id	User ID	bigint	uid	User ID	bigint	uid	User ID	bigint
country	ISO 3166 alpha-3 country code	text	group	User's test group	text	dt	Date of purchase	date
gender	User's gender (M = male, F = female, O = other)	text	join_dt	Date the user visited the page	date	device	Device type user purchased on (I = iOS, A = android)	text
			device	Device type user purchased on (I = iOS, A = android)	text	spent	Purchase amount in USD	double

Data Preparation

The data collected is in the format that will address the key experiment questions:

- What is the user conversion rate for the control and treatment groups?
- What is the average amount spent per user for the control and treatment groups, including users who did not convert?

The resulting table used for the analysis was extracted from the database using SQL query, *see Appendix for code,* that contains the user ID, the user's country, the user's gender, the user's device type, the user's test group, whether they converted (spent > $0), and how much they spent in total ($0+).

Null values were detected in certain fields, and the missing data was consolidated to create distinct variables that did not overlap or get excluded from the dataset.

Recommendation Guidance

The limitations and trade-offs inherent in the A/B test create ambiguity that should be addressed through prudent and relatively rigorous recommendation choices.

Launch the experience!

- If confident the banner shows a statistically significant difference on all metrics.

Alternative outcomes

- Iterate on the experience design and conduct a fresh A/B test, launching the banner only if there is statistical significance in at least one metric, with a focus on conversion rate. Avoid launching the banner if confidence is lacking in its significant impact on both metrics.

Analysis

Inferential statistics play a crucial role in determining whether the alterations being tested through A/B testing result in substantial and significant shifts in the relevant metrics. For GloBox, our focus is on assessing whether the new food and drink banner is driving variations in both the user conversion rate and the average spending per user.

Significance level and tail type, novelty check and power of the test.

Significance level and tail type

Two-tailed z-test and t-test will be run as appropriate based on 0.05 or 5% significance level.

Novelty check

By inspecting the difference in key metrics between over time, we aim to check for a novelty effect meaning that the effectiveness of the banner is temporary. This will impact our conclusion and recommendations.

Power analysis

Considering a relatively low marginal cost of conducting the A/B test, the strength of the test is set at 0.8 statistical power and 0.05 significance level. A power analysis assists us in determining the required sample size to attain our target minimum detectable effect and desired statistical power.

Test Integrity

A/B test is appropriate for this exercise due to the conditions listed below.

- The test involves a single variable, the new banner, which serves to differentiate the test group from the baseline and is not intricate.
- Being mindful of potential risks related to the test thereby ensuring the results are properly interpreted.

Preview Analysis

Power Analysis

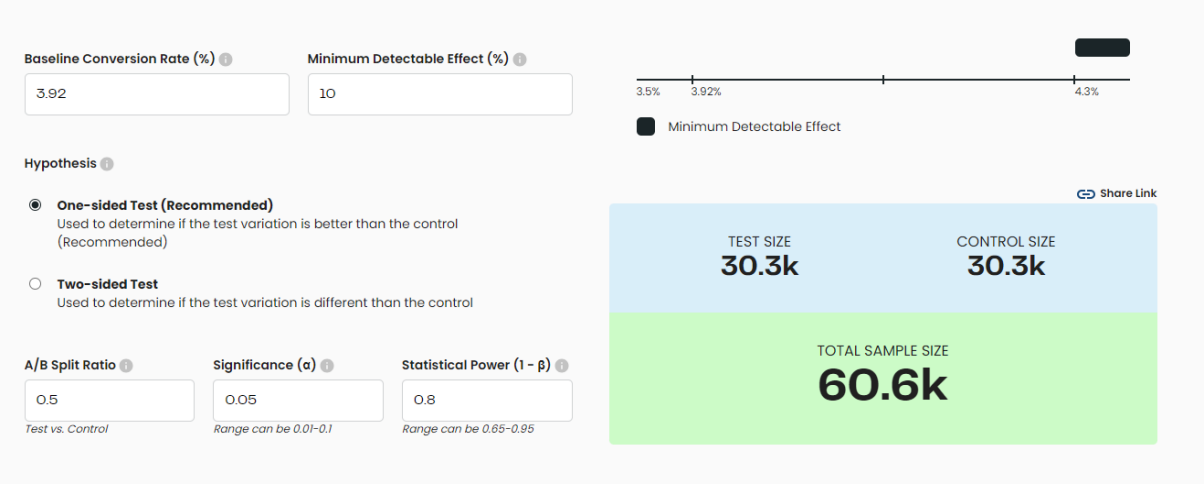
48943 of mobile users to webpage are registered as test subjects.

Effective 50:50 split size between baseline (24343) and treatment (24600).

Predefined significance level (0.05) and statistical power (0.8). The minimum detectable effect adopted here is 10%.

Power Analysis for Conversion rate

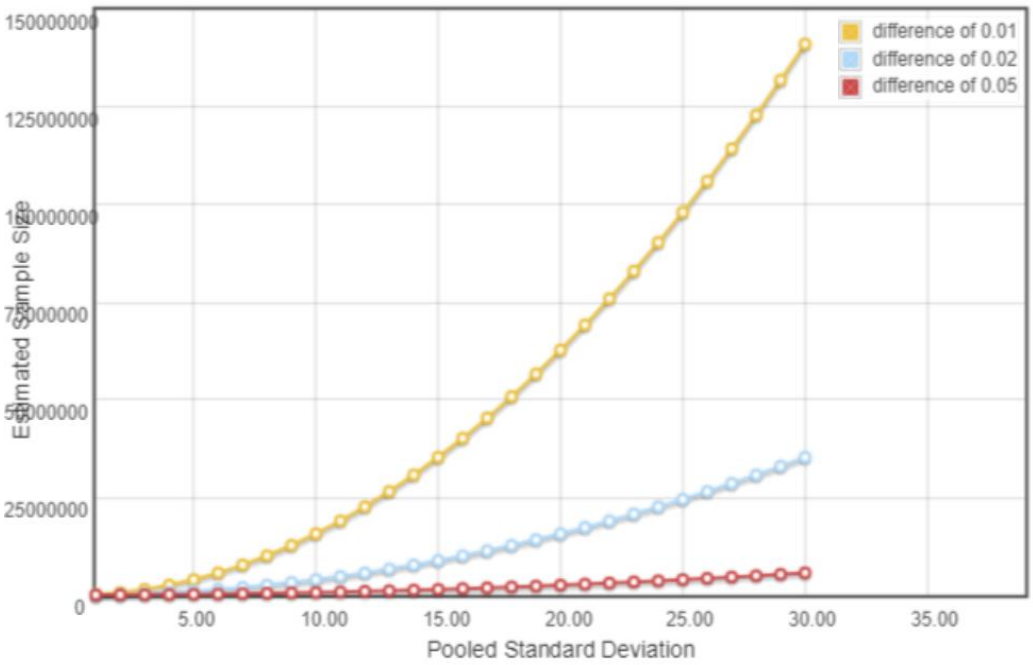
Test with sample size is sufficiently sensitive to detect conversion rate. A one-sided test was used to determine if the test variation is better than the control. Using the baseline conversion rate of 3.92 percent, there must be sample size of at least 60k users for both test groups to achieve the 10% MDE where the experiment is adequately powered to achieve a statistically significant result.



Power Analysis for Difference in Mean

Assuming a pooled standard deviation of 25.67 units, the study would require a sample size of at least 26 million users for each group (i.e., a total sample size of at least 52 million, assuming equal group sizes), to achieve a power of 80% and a level of significance of 5% (two sided), for detecting a true difference in means between the test and the reference group of 0.02 units. The sample size was insufficient to detect a statistically significant result confirming difference in spend per user metric for both groups.

The graph below shows a plot of sample sizes for a range of pooled Standard Deviations (min = 1 and max = 30) and for three values of Difference of means (0.01, 0.02 and 0.03) between groups.

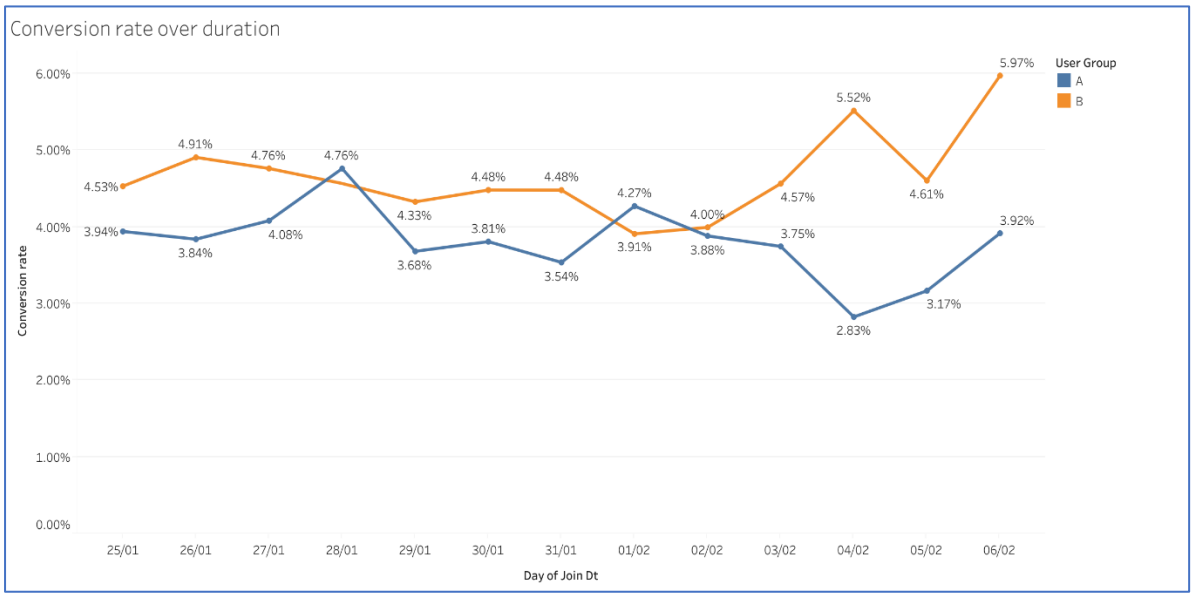


Influence of sample size on difference of means

The power analysis confirms that the sample size used in A/B test was sufficient to detect significant effect in conversion rates but not practically significant for detecting the minimal effect of average spending.

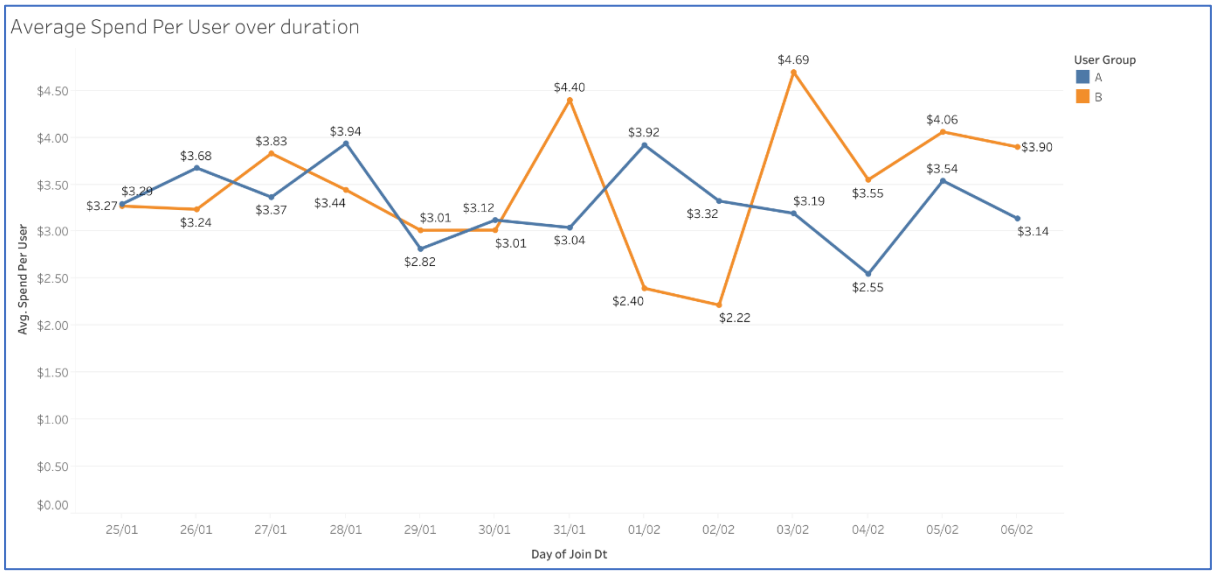
Novelty Effects from the perspective of metric transformations

- There not enough activity to indicate change aversion as there is not significant transient activity to indicate a novelty effect.
- The duration of the test maybe too short and there is a risk of missed opportunity to adequately study stabilisation or novelty effect in the conversion rate.



Novelty effect not detected.

- Much noise in both groups for the spend per user ($ / user) metric.



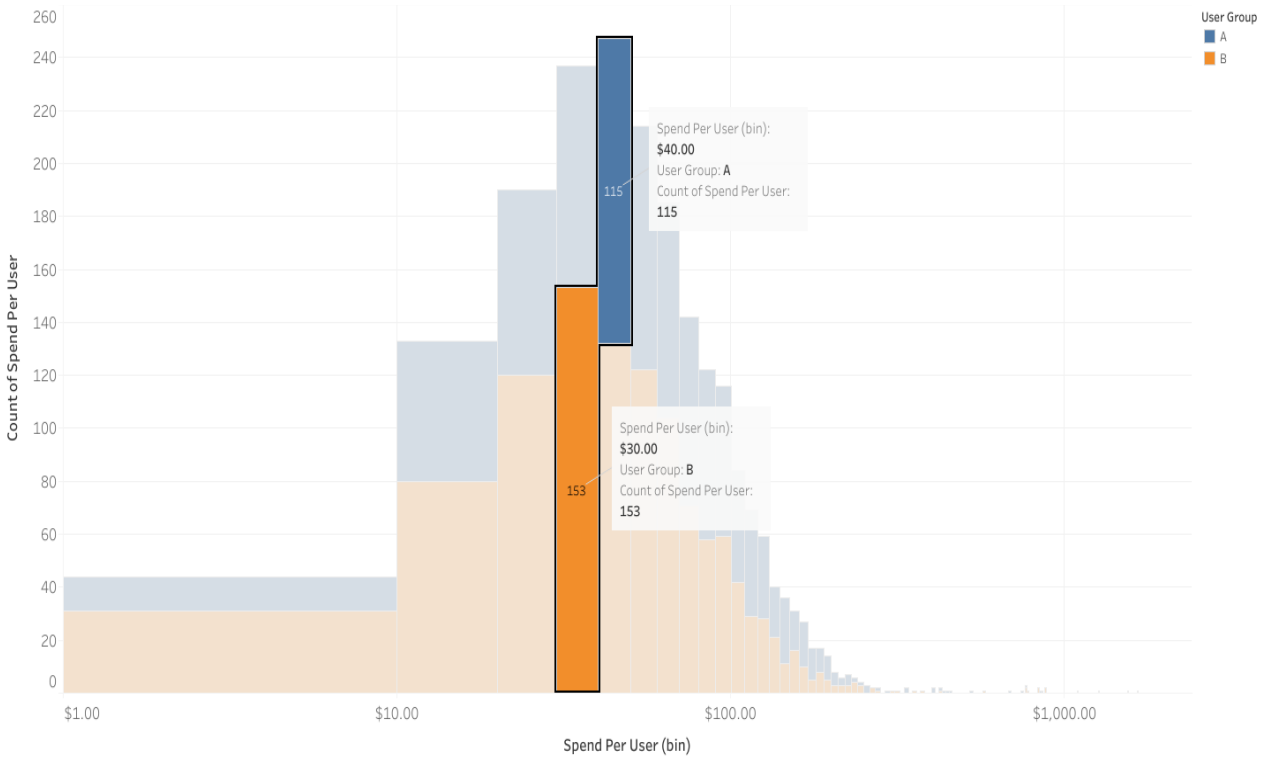
New banner may influence erratic behaviour.

Results

Data Insights

Distribution of Average Amount Spent

Insights reveals that majority of the control group spent between $40 and $50 while majority of the treatment group spent between $30 and $40.

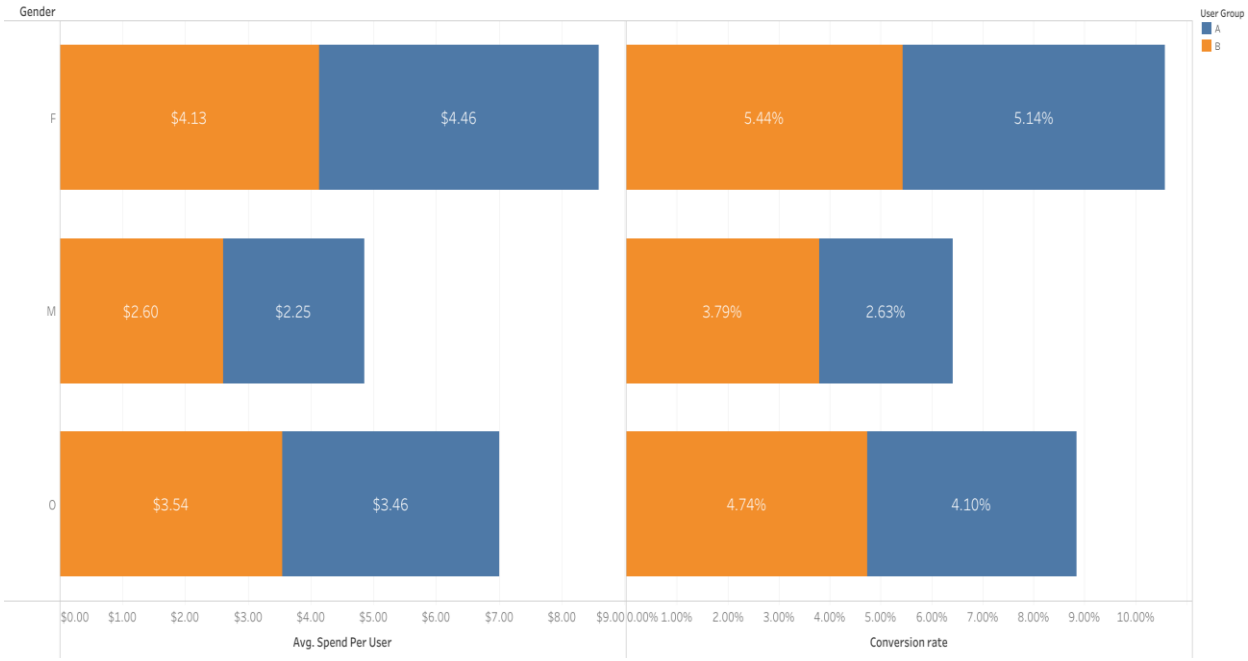


Proposals:

- Investigate reasons behind the spending patterns observed in each group.
- Monitor spending trends over time and adjust marketing strategies as needed.

Gender Analysis

Comparing all genders in the control group, males have the lowest conversion rate and average spending. However, there is a significant 44% increase in conversion rate when focusing on the treatment group. The female users while having a relatively high conversion rate in the control group, saw an insignificant 5% increase in conversion rate in the treatment group.

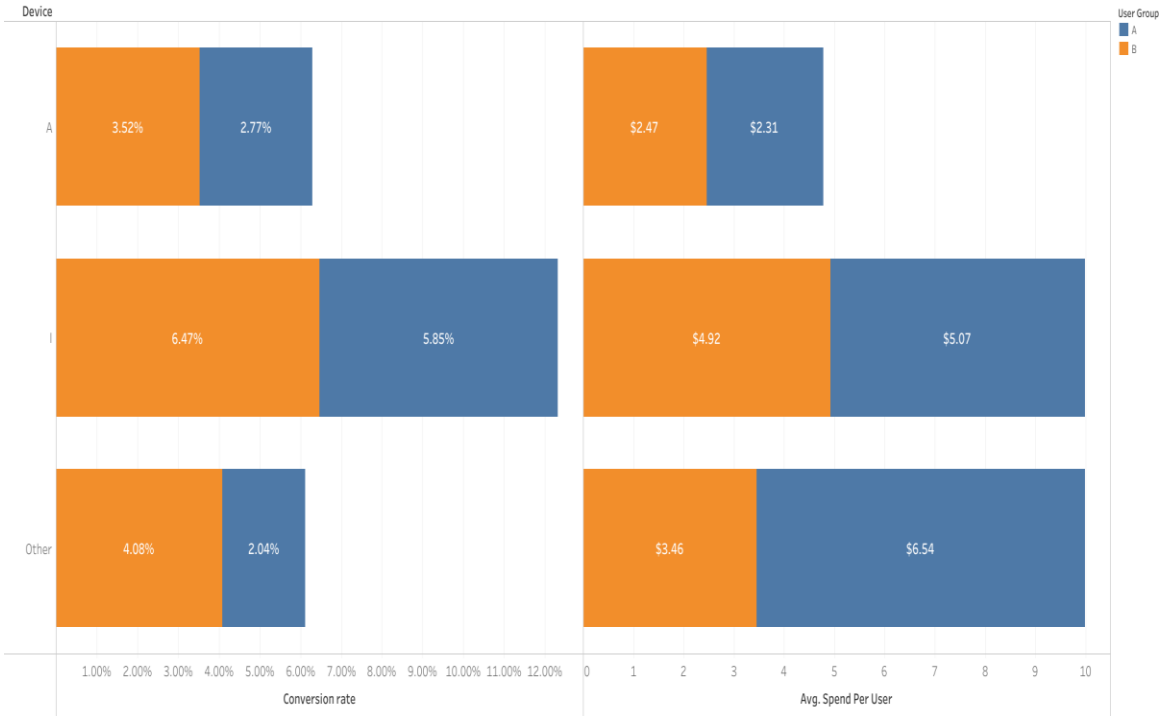


Proposals:

- Current focus should be in creating targeted marketing campaign to male preferences.
- Further investigate the female spending pattern and gather insights on preferences.
- Additional data should be gathered for the "**Other**" group to understand behaviour and preferences.

Device Analysis

The banner had a great impact on iOS users with the treatment group experiencing about 10% increase in conversion rate. Other device type had a significant positive change in the treatment group for both metrics. Android device type was the least affect by the new banner in both metrics.

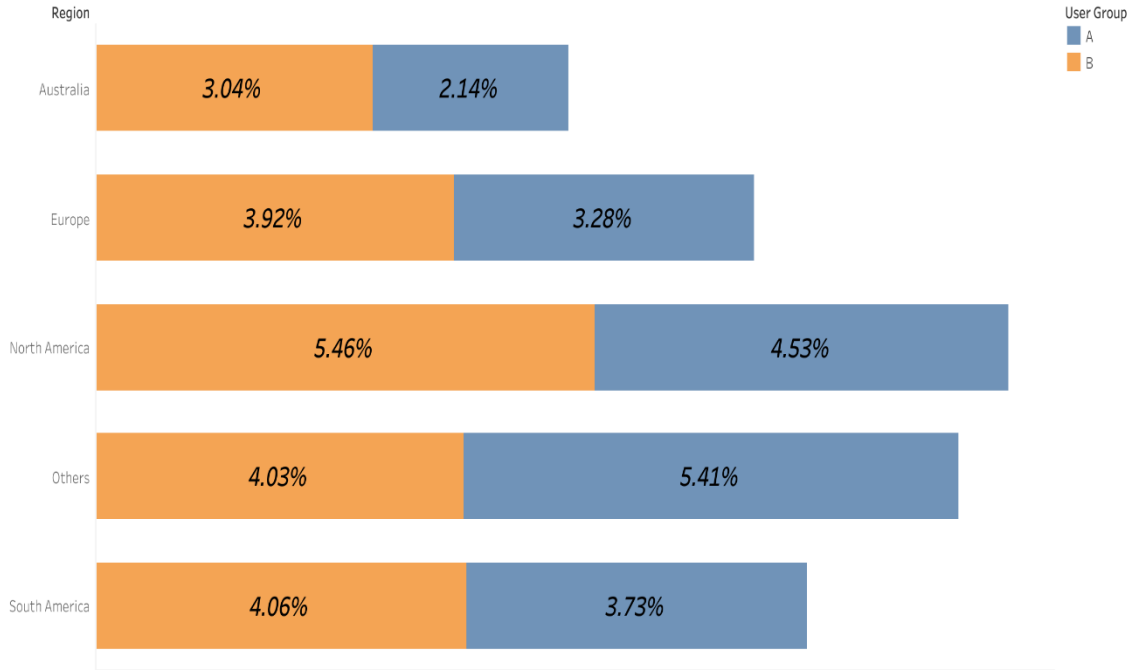


Proposals:

- To maximize the conversion rate among iOS users, the growth team should concentrate on enhancing the conversion funnel specifically tailored for iOS users.
- More data should be collected for the "**Other**" device type users to understand behaviour and preferences.
- A deep dive and continuous iteration of the Android device type user journey will help understand the low impact in both metrics from the new banner.

Global Analysis

There is a significant increase in conversion rate in Australia of about 42%. The other regions considering only the conversion rate metrics are North America and Europe.



Proposal:

- Capitalize on the significant improvements in Australia, North America and Europe.
- Gather more data and investigate the reverse trend for the “Others” region. Conduct in depth analysis for South America to identify opportunities for improvement in conversion.

Results of A/B Tests

Synopsis

The default stance, also known as the null hypothesis (H_0), assumes the absence of any evidence indicating that the new banner results in a substantial alteration in the conversion rate and per-user spending in the test group compared to the baseline. Conversely, the alternative hypothesis (H_1) contradicts H_0, proposing that there is evidence to support the notion that the new banner has an impact on these metrics within the test group.

Conversion Rate (CR)

Metric type: Proportion %.

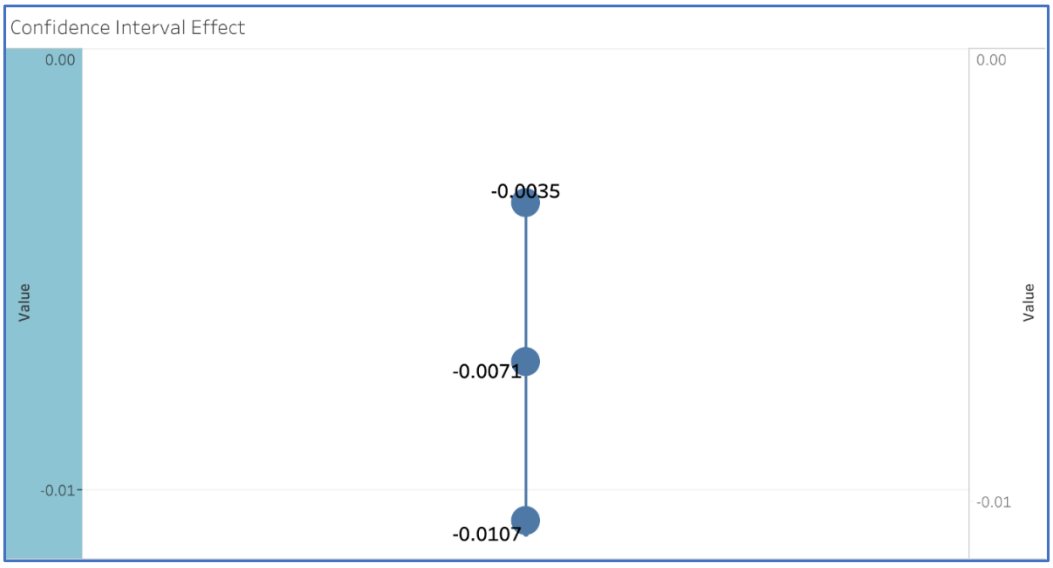
Test type: Two-sample z-interval with unequal variance

Hypothesis:

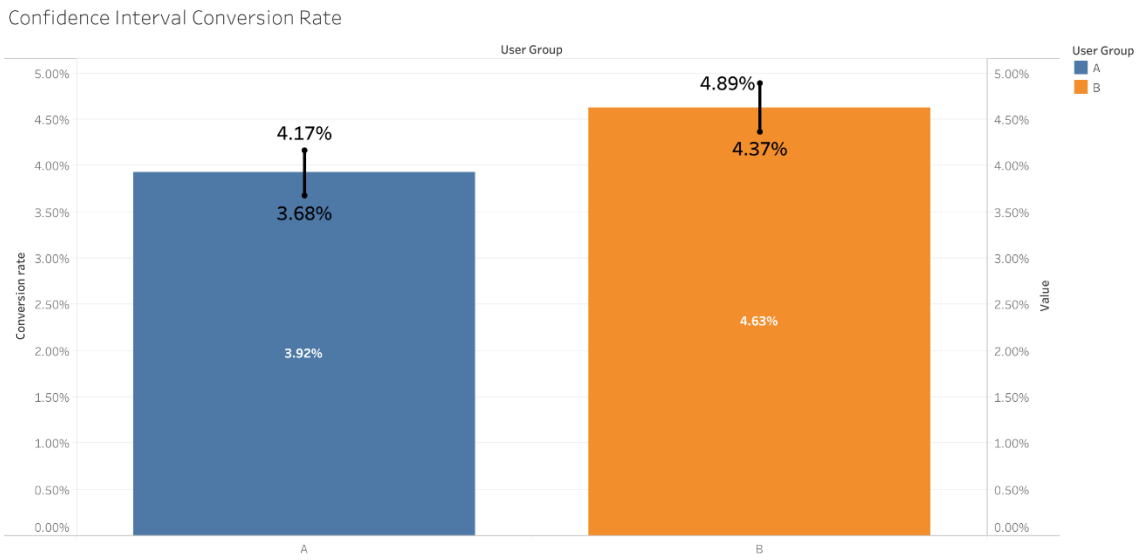
- H_0: $p_{control} = p_{test}$ i.e., Diff (if any) observed is due to chance.
- H_1: $p_{control} \neq p_{test}$ i.e., Diff is due to change in behaviour.

Conclusion:

- With p <0.05 (0.0001 < 0.05), we reject the null hypothesis that the is no difference in the average conversion rate of Globox mobile user customers in the control and treatment groups.
- Difference in conversion rate estimated to be in interval below 95% confidence interval is between -1.07% and -0.35%. This shows that the test group has a higher conversion rate than the baseline. The point estimate is ~0.71%. The standard error of the difference is ~0.0018, and the margin of error is ~0.00358.



- There is evidence of statistical significance as such **REJECT H_0**



Average Spend per User ($/user)

Metric type: Average Spend ($).

Test type: 2-tail t-test, Degree of freedom (df) is 24342.

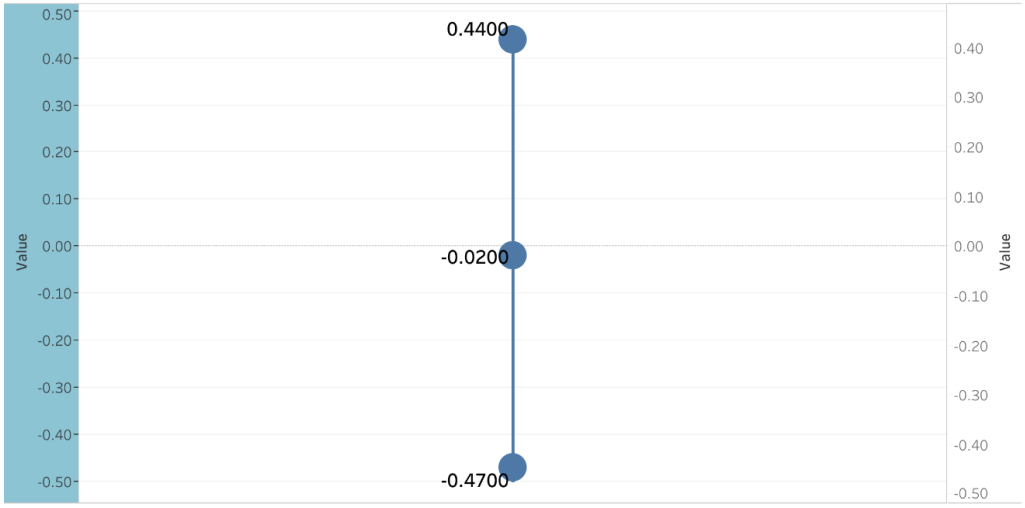
Hypothesis:

- H_0: $p_{control} = p_{test}$ i.e., Diff (if any) observed is due to chance.
- H_1: $p_{control} \neq p_{test}$ i.e., Diff is due to change in behaviour.

Conclusion:

- With p > 0.05 (0.94 > 0.05), we fail to reject the null hypothesis that the is no difference in the average spend of Globox mobile user customers in the control and treatment groups.
- Difference in conversion rate estimated to be in interval below 95% confidence interval is between -0.47 and 0.44. The point estimate is ~$0.02. The standard error of the difference is ~0.232, and the margin error is ~0.455.

Confidence Interval Effect



- The confidence interval straddles zero as such **FAIL TO REJECT H_0**

Confidence Interval for Avg Spend

Conclusion and Recommendation

In its entirety, the evidence suggests that the new banner program has a significant impact in conversion rate but almost negligible effect on average spend between both test groups.

The benefit of the statistically significant effect in increasing conversion rate is offset by the much insignificant revenue value being converted which nullifies the practical significance of embarking on the experience.

There is also a recurring observation of the need for further investigation into the preferences and user behaviour of different user groupings.

To conduct further research, possibly with a larger sample size, to increase the statistical power of the analysis and detect smaller differences. Also, increase the duration of the test to allow time to gather more information, conduct further analyses to identify different user behaviour and preferences that may influence purchasing decision.

It is therefore not advisable to launch the experience in full scale. Considering the recommendation guidance, we recommend **continue iterating with design of the experience and run a new A/B test.** This is recommended on the premise of zero financial and service cost to the company.

Appendix

Tableau Visualization

Test Statistics Calculations

Data Extraction Query