GloBox A/B Test Analysis

Result and Recommendation Report

LIEZYL JUGALBOT

NOVEMBER (2023)

CONTENTS

EXECUTIVE SUMMARY

After conducting an A/B test for GloBox's food and drink banner on the mobile website, it was observed that there was a statistically significant increase in the conversion rate. However, there was no significant change in the average amount spent per user. Based on the findings, it is recommended not to launch the experiment in its current form. Instead, we should continue iterating on the banner experience and conduct further analysis. This will help us better understand its impact on revenue and user experience before considering a full-scale launch.

CONTEXT

GloBox is an e-commerce company sourcing unique and high-quality products worldwide. While it is renowned for boutique fashion items and high-end decor products, the company has experienced significant growth in its food and drink offerings over the last few months.

MOTIVATION

The primary motivation behind conducting an A/B test is to increase awareness of the food and drink product category and, consequently, boost revenue. GloBox's mission is to bring the world's treasures to its customers' doorsteps, offering everything from exotic spices to handmade jewelry.

A/B TEST SETUP

The A/B test aimed to assess the impact of prominently featuring critical products in the food and drink category on the website. Users who visited the GloBox main page were randomly assigned to either the control or treatment group, marking their join date. The landing page displayed the banner highlighting food and drink items for users in the treatment group, whereas the control group saw the standard website without this feature.

Group A: Control
existing landing page

Group B: Treatment
landing page with food & drink banner

The experiment ran for 13 days, from January 25 to February 6, 2023. There were 24,343 users in the control group and 24,600 users in the treatment group, totaling 48,943.

The subsequent actions of users, such as making purchases on the same day they joined the experiment or in the days that followed, were tracked as conversions.

DATA OVERVIEW

The dataset used in this analysis contained users' demographic information, their group assignments, and whether or not they made purchases after viewing the website with or without the food and drink banner.

RESULTS

UNDERSTANDING THE GLOBOX DATABASE

To understand the database better, let us look at some key points. Please refer to the References page for SQL queries.

- A user can purchase on multiple days.
- To combine the users table with the activity table, we use a `LEFT JOIN`.
- We use the `COALESCE()` function to fill in any null values.
- The experiment ran from January 25, 2023, to February 6, 2023.
- There were 48,943 users - 24,343 in the control and 24,600 in the treatment.
- The conversion rate of all users was 4.28%, with the control at 3.93% and the treatment at 4.63%.
- On average, users in the control group spent $3.37, and the treatment group spent $3.39.
- To measure the impact on total revenue, we cannot solely average the users who converted since fewer users in the treatment group may have converted.

EXTRACTING AND CLEANING THE DATASET

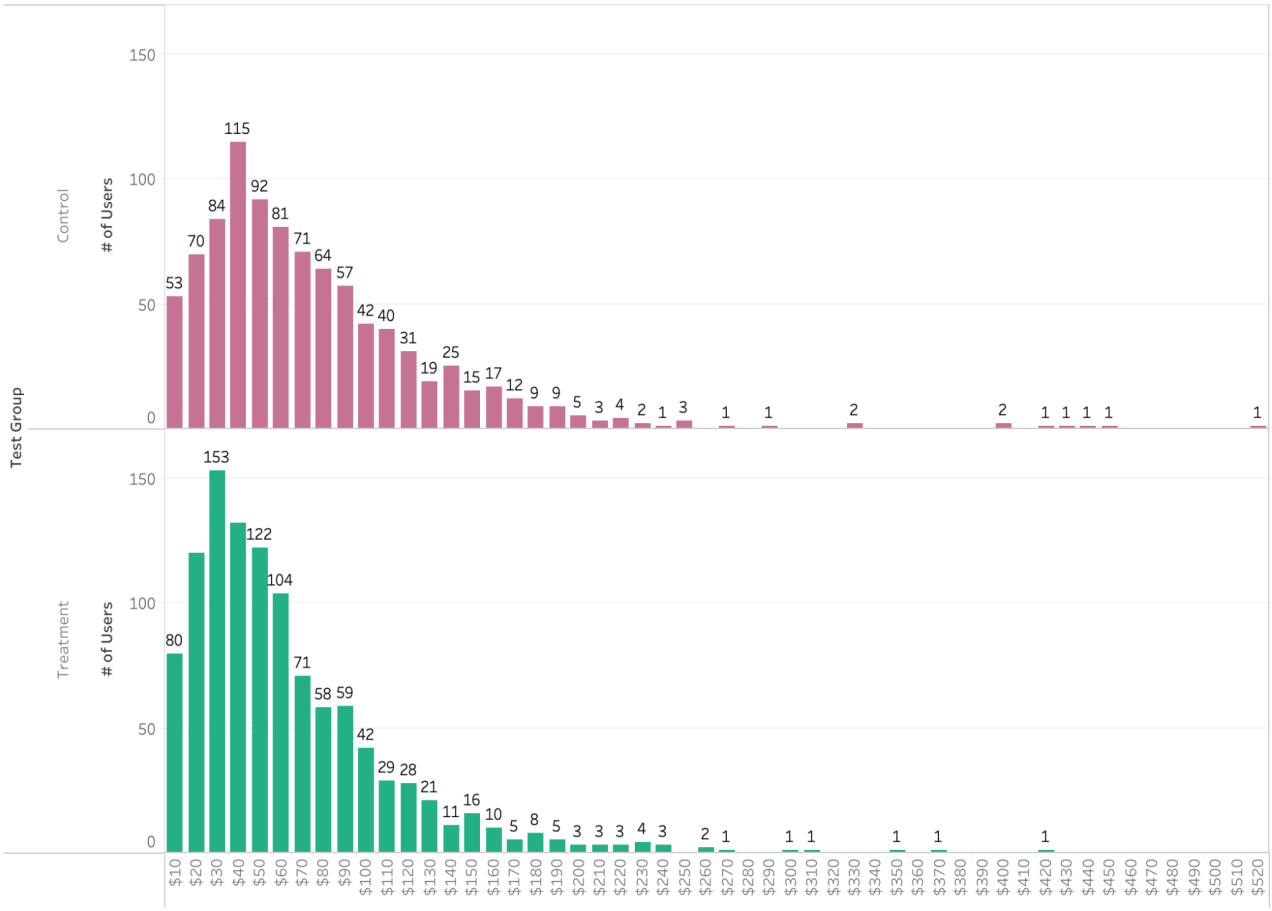
A SQL query code has been written to retrieve the user ID, country, gender, device type, test group, conversion status (whether they spent more than $0), and total spending ($0+). The data was downloaded in CSV and Excel to create visualizations and conduct the hypothesis test.

VISUALIZATIONS

Figure 1. Visualization to compare the test metrics between the test groups.

Figure 1 compares the conversion rate and the average amount spent by the test groups. The graph shows a significant increase of 18% in the conversion rate, which is noteworthy. However, the slight difference in the average spending suggests that the banner had a minimal impact. It indicates that additional efforts might be necessary to influence users' average spending behavior.

Figure 2. Distribution of the amount spent per user for each group.



In Figure 2, we can observe the distribution of the amount spent per user for each group. The graph shows a right-skewed distribution, indicating that most users in the test groups spent relatively lower amounts, with a few users spending significantly more. The long tail on the right side of the graph represents these outliers.

Figure 3. Relationship between the test metrics and the user's device.

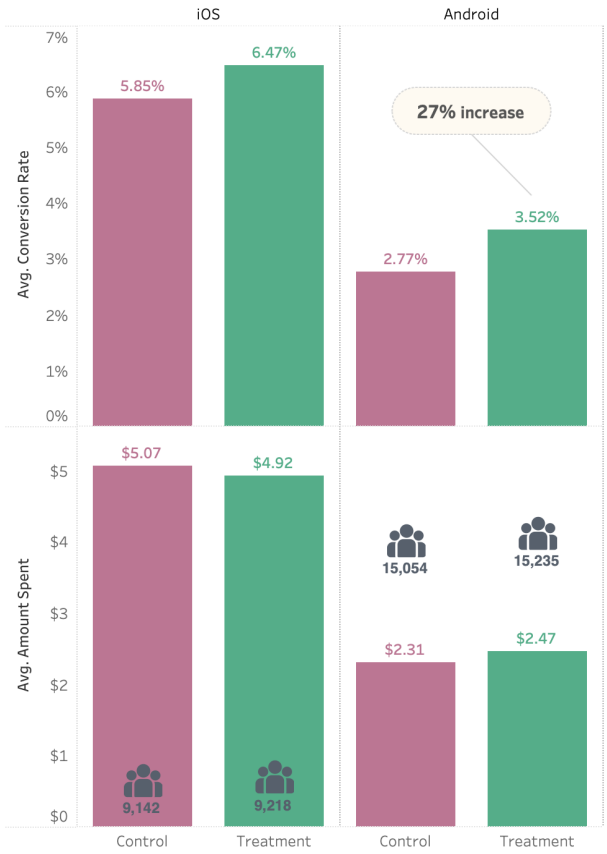
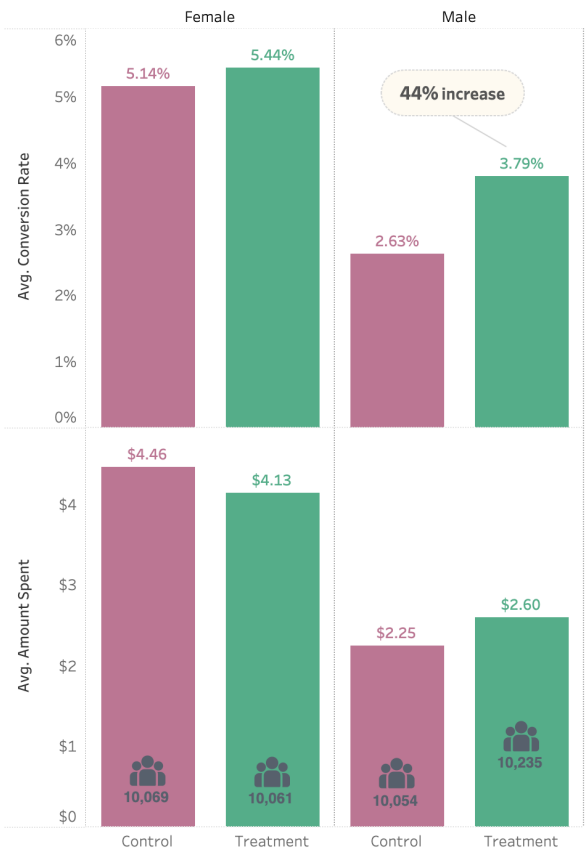


Figure 4. Relationship between the test metrics and the user's gender.



Figures 3 and 4 show the correlation between the test metrics, the user's device, and gender. The data shows that there are significantly more Android users than iOS users, but iOS users show a higher conversion rate and the average amount spent. Notably, there has been a 27% increase in conversion rate among Android users. Female users have higher overall metrics than male users, but it is also worth noting that male users show a significant 44% increase in conversion rate.

Figure 5. Relationship between the test metrics and the user's country.

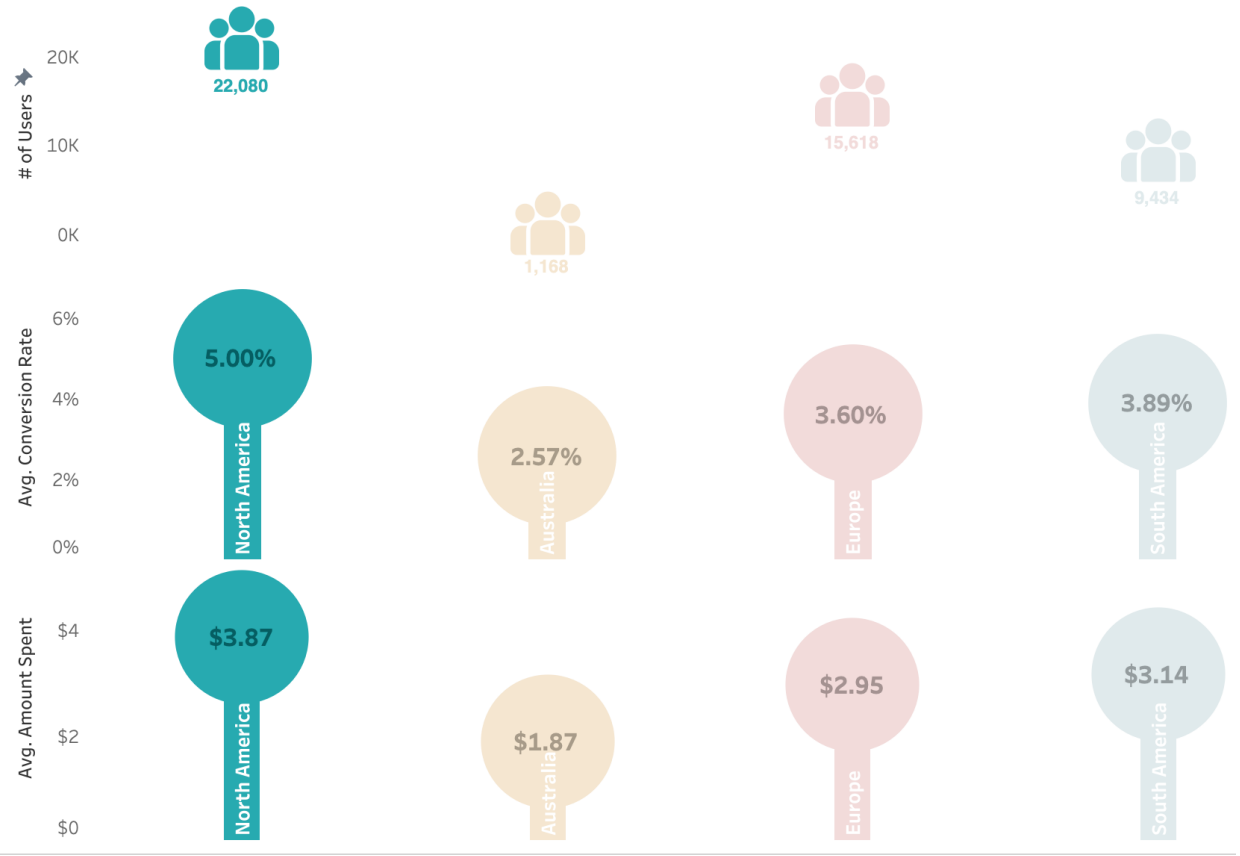


Figure 5 illustrates the relationship between user metrics and their respective countries. North America is the top-performing region regarding conversion rate and average user spending. This analysis suggests that users in North America are more likely to convert and spend more on average than users in other regions.

HYPOTHESIS TEST

- The null hypothesis (H0) stated no difference in the conversion rate/average amount spent between the two test groups.
- The alternative hypothesis (H1) stated that there is a difference in the conversion rate/average amount spent between the two test groups.

Hypothesis Test for the Difference in Conversion Rate

Two-sample **z-test** with pooled proportion

A - Control B - Treatment

test_group	COUNT of user_id	SUM of conversion	AVERAGE of conversion
A	24343	955	0.0392
B	24600	1139	0.0463
Grand Total	**48943**	**2094**	**0.0428**

Calculation	**Notation**	**Value**
sample size (A-control)	n1	24343
sample size (B-treatment)	n2	24600
sample mean (A-control)	x1 bar	0.0392
sample mean (B-treatment)	x2 bar	0.0463
sample proportion	p-hat	0.0428
standard error	SE	0.0018
test statistic	T	-3.8643
p-value	pval	**0.0001**

The hypothesis test results provide statistically solid evidence of a significant difference in conversion rates between the control and the treatment groups. With the p-value = 0.0001 < 0.05, we **reject the null hypothesis** that the conversion rate is the same between the two groups in favor of the alternative hypothesis that there is a difference in the conversion rate between the two groups.

Hypothesis Test for the Difference in Average Amount Spent

Two-sample **t-test** with unpooled variance

A - Control B - Treatment

test_group	COUNT of user_id	AVERAGE of total_spent	STDEV of total_spent
A	24343	3.375	25.936
B	24600	3.391	25.414
Grand Total	**48943**	**$3.38**	**25.675**

Calculation	**Notation**	**Value**
sample size (A-control)	n1	24343
sample size (B-treatment)	n2	24600
sample mean (A-control)	x1 bar	3.375
sample mean (B-treatment)	x2 bar	3.391
sample std dev (A-control)	s1	25.936
sample std dev (B-treatment)	s2	25.414
standard error	SE	0.2321
test statistic	T	-0.07043
degrees of freedom	df	24342
p-value	pval	**0.94385**

The hypothesis test results for the difference in the average amount spent between the two groups indicate **no statistically significant difference** in the average spending of the users. With the p-value = 0.94 > 0.05, we **fail to reject the null hypothesis** that the average amount spent is the same between the two groups in favor of the alternative hypothesis that there is a difference in the average amount spent between the two groups.

CONFIDENCE INTERVALS

Confidence Interval for a Difference in Conversion Rate

Two-sample **z-interval** with unpooled variance

A - Control B - Treatment

test_group	COUNT of user_id	SUM of conversion	AVERAGE of conversion
A	24343	955	0.0392
B	24600	1139	0.0463
Grand Total	**48943**	**2094**	**0.0428**

Calculation	**Notation**	**Value**
sample size (A-control)	n1	24343
sample size (B-treatment)	n2	24600
sample mean (A-control)	p1	0.0392
sample mean (B-treatment)	p2	0.0463
degrees of freedom	df	24342
critical value	c	1.96
standard error	SE	0.0018
margin of error	E	0.0036
lower bound	LB	**0.0035**
upper bound	UB	**0.0107**

The calculated confidence interval for the conversion rate is (0.0035, 0.0107). We can say that we are 95% confident that the difference in the conversion rate falls between 0.35% and 1.07%. Moreover, this confidence interval **does not include the value 0**, indicating **a statistically significant difference** in the conversion rates between the two groups.

Confidence Interval for a Difference in Average Amount Spent

Two-sample **t-interval** with unpooled variance

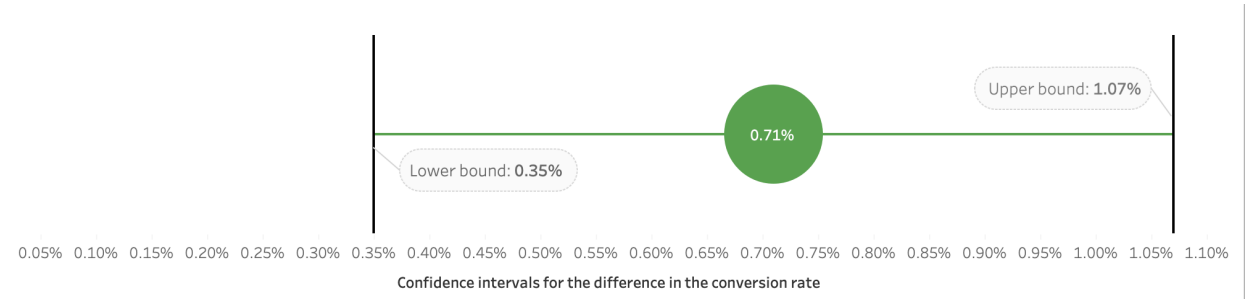
A - Control B - Treatment

test_group	COUNT of user_id	AVERAGE of total_spent	STDEV of total_spent
A	24343	3.375	25.936
B	24600	3.391	25.414
Grand Total	**48943**	**$3.38**	**25.675**

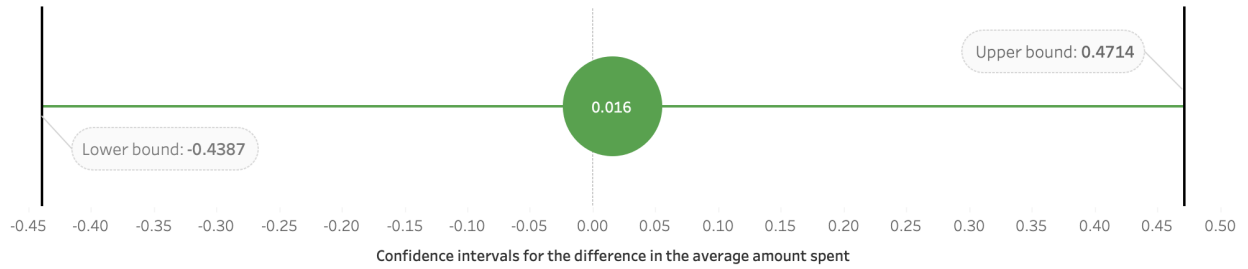
Calculation	**Notation**	**Value**
sample size (A-control)	n1	24343
sample size (B-treatment)	n2	24600
sample mean (A-control)	x1 bar	3.375
sample mean (B-treatment)	x2 bar	3.391
sample std dev (A-control)	s1	25.936
sample std dev (B-treatment)	s2	25.414
degrees of freedom	df	24342
critical value	c	1.96
standard error	SE	0.2321
margin of error	E	0.4550
lower bound	LB	**-0.4387**
upper bound	UB	**0.4714**

The confidence interval for the difference in average amount spent is (-0.4387, 0.4714). With 95% confidence, the difference in the average amount spent falls between -0.4387 and 0.4714. More importantly, the confidence interval **includes the value 0,** indicating a chance that the actual difference in the average amount spent between the two groups is zero or, in other words, there is **no statistically significant difference** in the average amount spent.

VISUALIZING THE CONFIDENCE INTERVALS

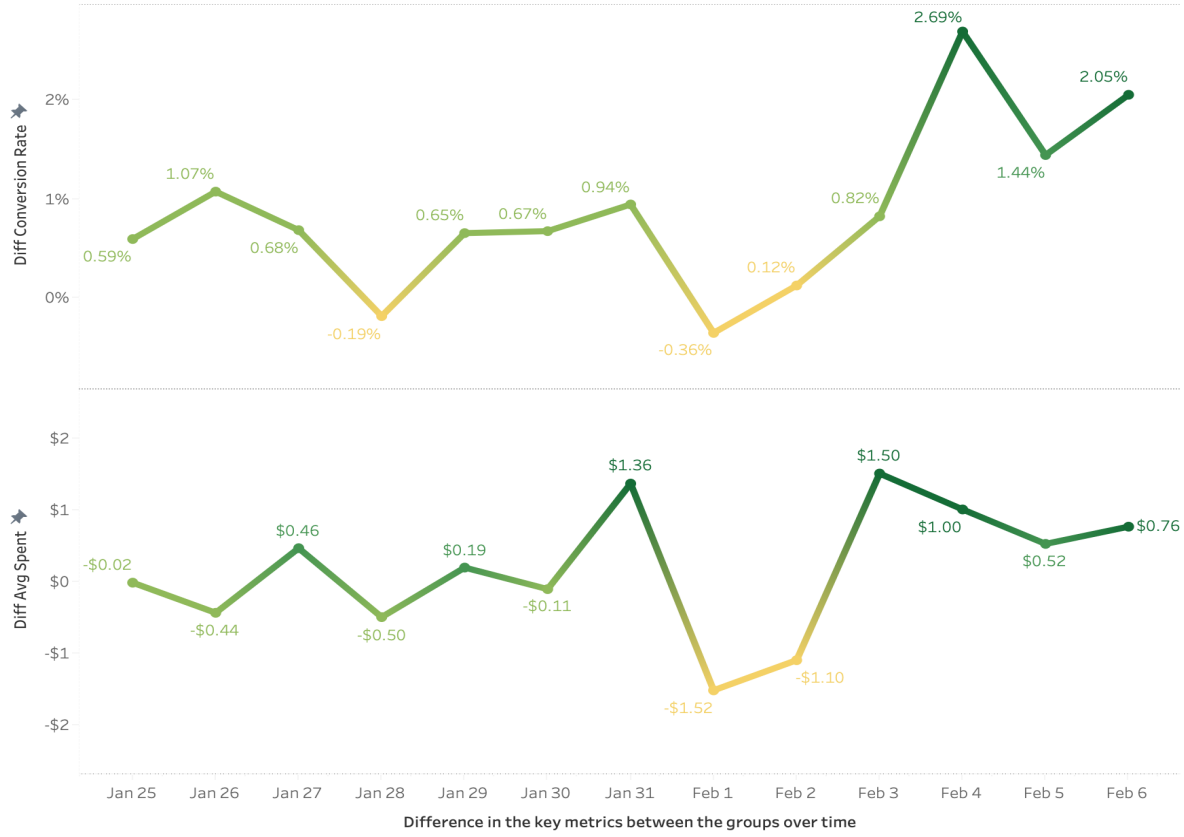


The CI is positive and statistically significant since 0 is not in the bound. However, the lower bound of the CI is less than the practical significance at +10% MDE.



There is a wide CI for the difference in the average amount spent, and there's no statistical significance since 0 is in the bound. The upper bound exceeding the practical significance at +10% suggests that the valid parameter may lie equal to or above the MDE.

NOVELTY EFFECT



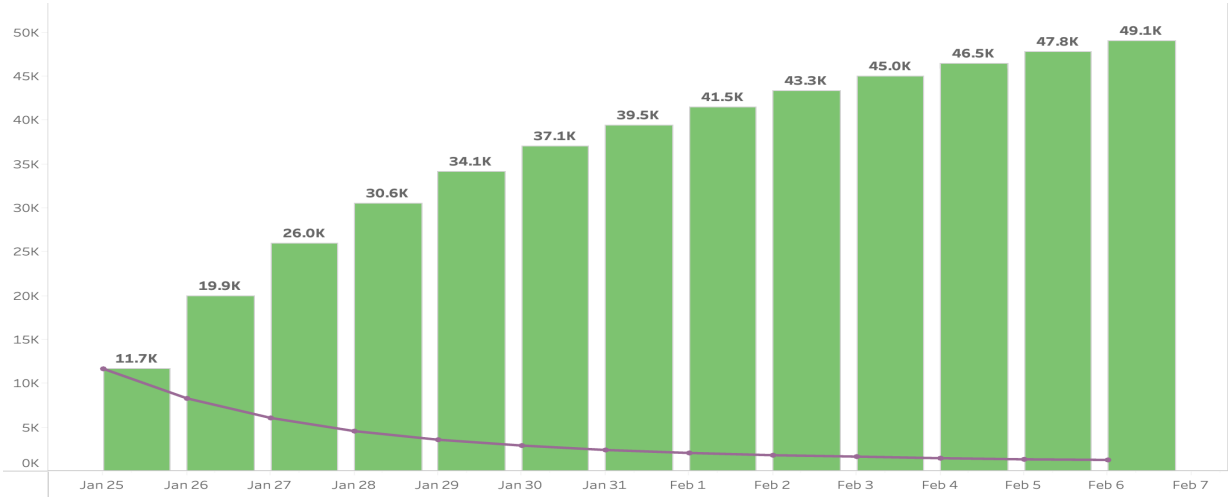
Introducing the food and drink banner led to higher conversion rates, particularly in the final days of the experiment. Analysis reveals no detectable novelty effect from the introduction of the banner, indicating that immediate and significant changes in user behavior, typically associated with novelty, were not observed. However, the banner may have an impact on unpredictable behavior in the average amount spent without a noticeable novelty effect.

Overall, the lack of unexpected effects and the banner's seamless integration into the user experience suggest that it is effective and well-received, matching user preferences and minimizing disruptions. and minimizing disruptions.

POWER ANALYSIS

After analyzing the results of our hypothesis test, we rejected the null hypothesis that there is no difference in the conversion rate between the two groups but we did not find enough evidence to reject the null hypothesis in the average amount spent, which suggests that the average amount spent is the same for both groups. However, we must conduct a power analysis to determine the required sample size to obtain a significant difference in the average amount spent between the two groups.

With a significance level set at 0.05, a statistical power of 0.80, and a 10% minimum detectable effect, a sample size of roughly 25K users is necessary to observe significant changes in conversion rate. However, **to achieve 80% power and declare significant differences between the two groups in terms of the average amount spent, a total sample size of approximately 170K users is required.**



The line on the chart indicates the number of users who joined daily. The desired total sample size reaches sufficiency in 3 days to detect significant changes in conversion rate. We need to ramp up more users to detect the significant difference in the average amount spent.

RECOMMENDATION

I recommend that we **do not launch the experiment** in its current form. Instead, we continue iterating and take more time to fine-tune the experiment before making it available to all users. We need to see more improvement in our success metrics to be confident in releasing the feature.

Based on our analysis, we have observed a positive trend in the conversion rate, which indicates that there is a potential for improving the banner experience in the future. However, the lower bound of the confidence interval falls below our practical significance threshold of +10% MDE. Furthermore, we conducted a power analysis and concluded that our current sample size is not enough to detect a significant change in the average amount spent. Hence, we recommend that the test be re-run with an increased sample size to ensure that the results are practically significant and sufficient to detect significant changes.

In addition, we should consider geographic targeting in marketing campaigns to take advantage of North America's strong performance and to improve results in other regions.

The banner positively affects user engagement, but launching based on one success metric would take time and effort. In addition, we should evaluate costs before launching the banner to determine if there are clear financial benefits and to assess any negative impact on user experience or resource requirements. It is essential to evaluate its impact on profit in addition to revenue.

REFERENCES

1. SQL Query Codes for Globox Project_liez
2. Statistics in Spreadsheets_liez
3. Confidence Interval and Hypothesis Testing Cheat Sheet
4. Tableau - Metrics and Test Groups
5. Tableau - Distribution of Amount Spent
6. Tableau - Metrics and User's Device
7. Tableau - Metrics and User's Gender
8. Tableau - Metrics and User's Country
9. Tableau - Confidence Intervals
10. Tableau - Novelty Effect
11. Tableau - GloBox Dashboard
12. Calculator - Estimated sample size for proportions
13. Calculator - Estimated sample size for means
14. Using confidence intervals to make an informed decision