GloBox A/B Test - Data Analysis

Table of contents

GloBox A/B Test - Data Analysis

By: Data Analyst Team
Date: July 20, 2023,

1. Summary

Recommendation: Continue iterating

Based on our analysis, we recommend that the team continue iterating the new homepage design.

—--

GloBox is a reputable e-commerce company that specializes in curating exceptional and high-quality products sourced globally. While it has gained recognition for its exclusive boutique fashion items and upscale decor products, the company has recently experienced significant growth in its food and drink offerings. Consequently, GloBox is focused on enhancing awareness of this product category to drive revenue growth.

The Growth team conducted an A/B test to assess the effectiveness of a banner feature on the mobile version of the website's main page. The purpose was to highlight key products in the food and drink category.

After analyzing the test results, we found no compelling evidence of an increase in revenue per user. However, statistically significant success metrics were observed specifically related to conversion rates. As a result, we recommend that the team continues refining the new homepage design through iterative improvements. This approach will enable GloBox to capture the attention and interest of its customers more effectively, ultimately driving increased revenue from the food and drink category.

This recommendation is aimed at maximizing GloBox's potential to leverage its expanded product offerings and further establish itself as a premier online marketplace for exceptional and high-quality goods. By continually enhancing the user experience and highlighting the food and drink category, GloBox can solidify its position in the market and drive sustained growth.

2. GloBox A/B Test

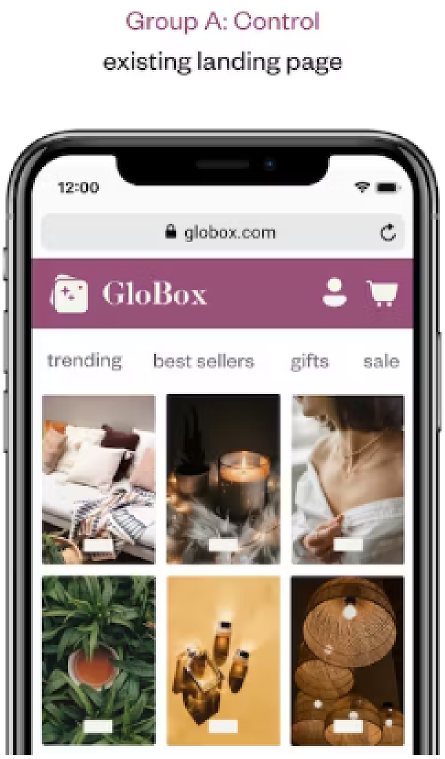
An A/B test is a fundamental experimentation technique employed by businesses to compare and evaluate the performance of two different versions of a webpage, advertisement, or product feature. By randomly assigning customers or users to either the A (control) or B (test) version, businesses can assess which version yields superior outcomes in achieving a specific objective.

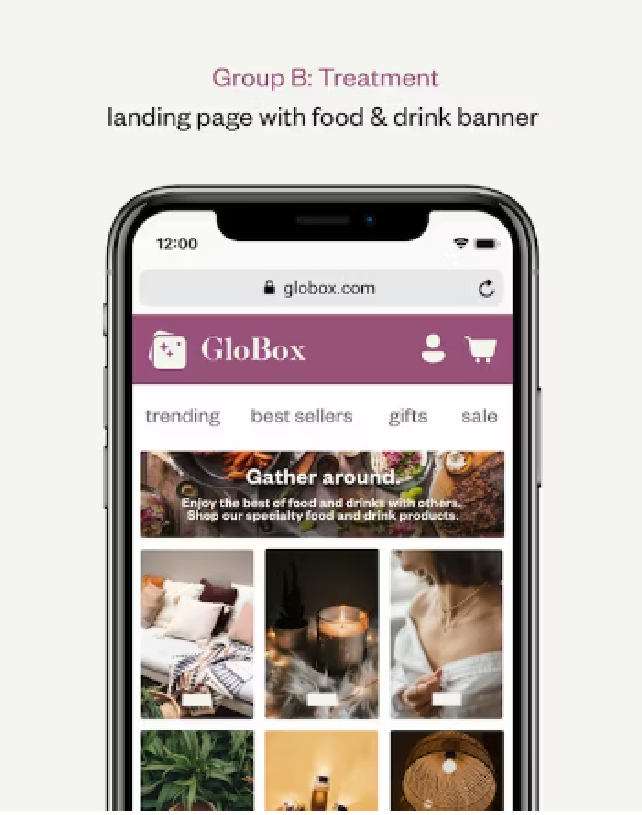
2.1. A/B Test Setup

The Growth team at GloBox conducted an A/B test focusing on the mobile website. The test setup involved introducing a banner at the top of the website, highlighting key products from the food and drink category. The control group did not encounter the banner, while the test group was exposed to it.

Here are the key steps in the A/B test setup:

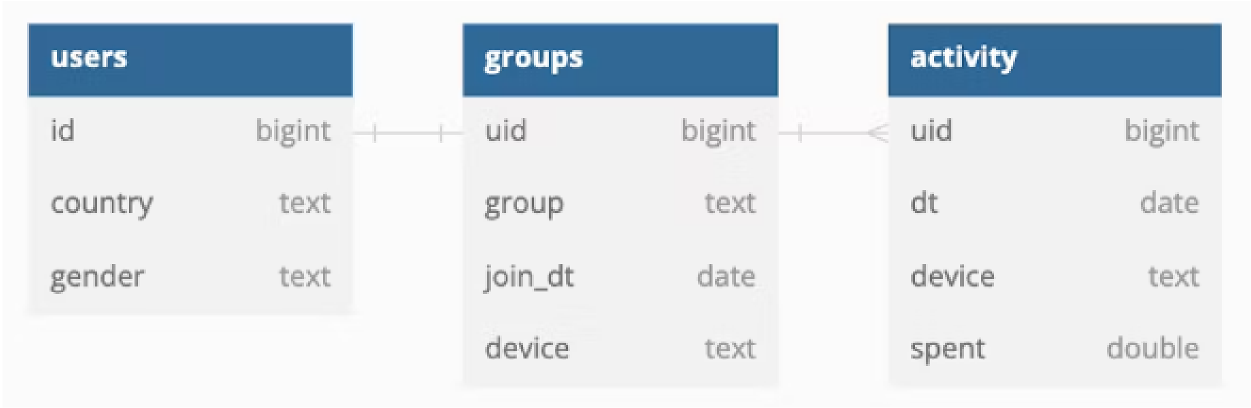
1. User Assignment: Upon visiting the GloBox main page, users were randomly assigned to either the control or test group. This assignment was determined by the user's join date, ensuring a fair and unbiased distribution.
2. Banner Display: The website loaded the banner for users assigned to the test group, while users in the control group did not encounter the banner.
3. Purchase Behavior: Subsequently, users had the opportunity to make purchases from the website. These purchases, whether made on the same day as the user joined the experiment or at a later date, were considered "conversions."




Group B: Treatment
landing page with food & drink banner
12:00
globox.com
GloBox
trending
best sellers
gifts
sale
Gather around.
Enjoy the best of food and drinks with others.
Shop our specialty food and drink products.

Dataset and Database:

The A/B test data for GloBox is stored in a relational database. The dataset's entity-relationship diagram (ERD) provides an overview of the database structure and relationships among the different data components.



users	
id	bigint
country	text
gender	text

groups	
uid	bigint
group	text
join_dt	date
device	text

activity	
uid	bigint
dt	date
device	text
spent	double

Important Considerations:

To ensure a comprehensive understanding of the A/B test, it is essential to note the following key aspects:

- Group Assignment: Every user participating in the A/B test was assigned to either the control or test group, guaranteeing complete coverage across all users.
- Purchase Activity: It is crucial to recognize that not all users made purchases during the test period. Purchase activity encompassed all product categories available on the GloBox website, rather than being limited to the food and drink category alone.

2.2. Data extraction

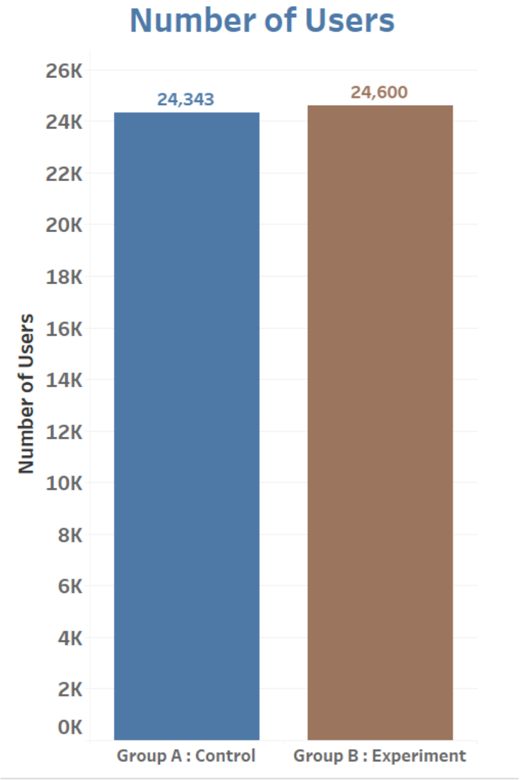
In order to conduct rigorous statistical analysis, data were extracted from the relational database system utilizing SQL queries. These queries were specifically designed to retrieve the necessary data for analysis. Furthermore, initial exploratory data analysis was performed using SQL queries, providing preliminary insights into the dataset. Detailed information regarding the SQL queries can be found in the APPENDIX section of this document.

The extracted data was downloaded in CSV format to facilitate subsequent statistical analysis. Google spreadsheets and Python were utilized as a tool for statistical calculations, while Tableau was employed for data visualizations. This combination of tools allowed for a comprehensive examination of the dataset, enabling the generation of meaningful statistical insights.

By employing SQL queries for data extraction, conducting statistical analysis, and leveraging the capabilities of Google spreadsheets and Tableau for data manipulation and visualization, a robust and systematic approach was adopted to analyze the dataset and derive valuable conclusions.

2.3. Data Analysis

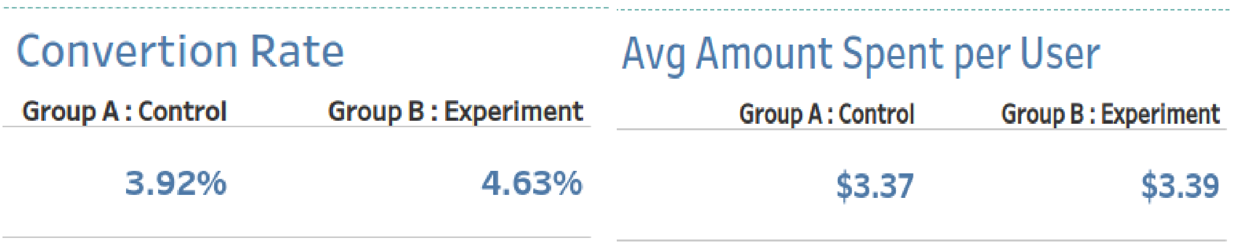
The A/B test was conducted over a period of **2 weeks** in Q1 2023, involving a total of **48,943** users. Among them, **24,343** users were assigned to the control group (Group A), while **24,600** users were assigned to the treatment group (Group B).



To assess the impact of the banner feature on the success metrics, a hypothesis test was performed. The objective was to determine if there was a statistically significant difference in the **average revenue per** user and **conversion rate** between the two groups.

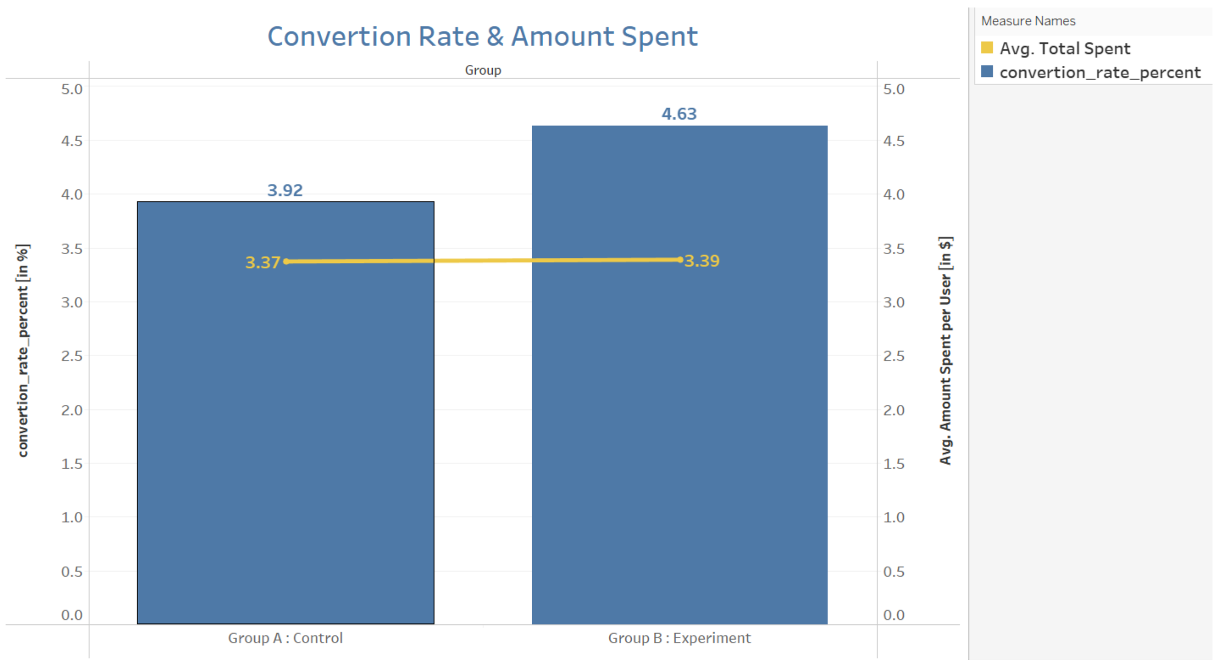
The analysis of the A/B test results revealed no substantial improvement in the success metrics, raising concerns about releasing the banner feature in its current state. However, a statistically significant increase was observed in one of the success metrics, specifically the User Conversion Rate. The conversion rate for the control group was **3.92**%, while it was **4.63**% for the treatment group. Furthermore, the average amount spent for the control group was $**3.37**, compared to $**3.39** for the treatment group.

Success metrics:

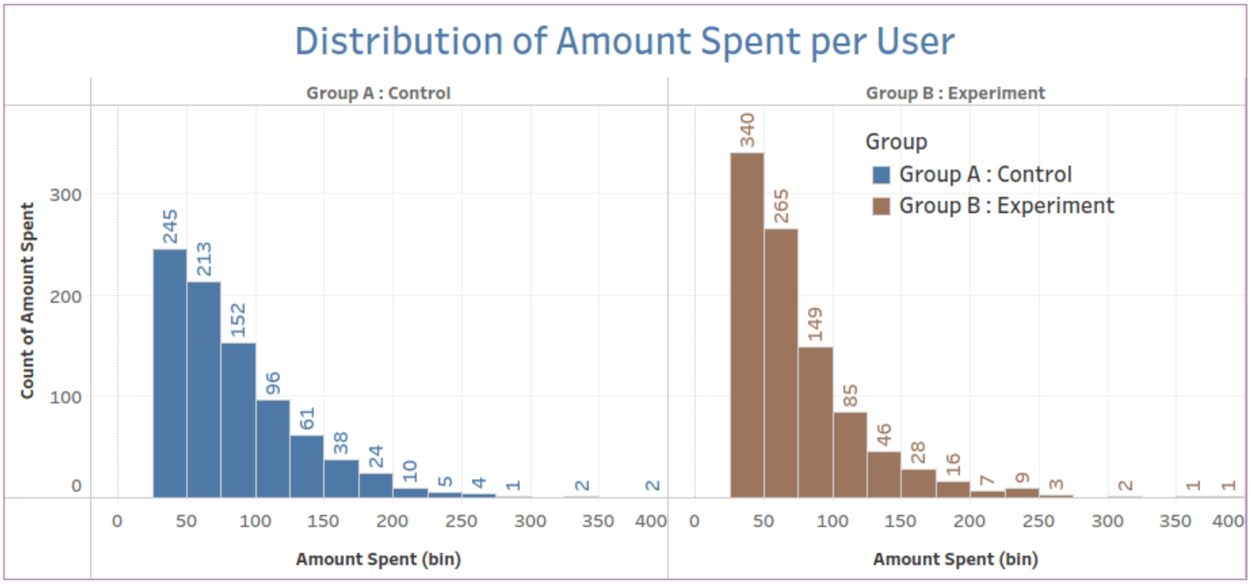


Convertion Rate	
Group A : Control	Group B : Experiment
3.92%	4.63%

Avg Amount Spent per User	
Group A : Control	Group B : Experiment
$3.37	$3.39



Distribution of amount spent per user:



Statistics: Hypothesis testing & Confidence intervals

Inferential statistics were used in establishing if the changes that we are A/B testing are leading to a real and meaningful change in the metrics of interest. For GloBox, we wanted to determine whether the new food and drink banner is leading to changes in the user conversion rate and the average amount spent per user.

Two Hypothesis tests were used to determine whether or not there is a statistically significant difference between the two test groups for our success metrics.

For Hypothesis Testing of user conversion rate, normal distribution, a 5% significance level, and pooled proportion for the standard error was used. Z-Test resulted in **p-value = 0.0001**, **statistically significant**. We **reject** the null hypothesis that there is no difference in the user conversion rate between the control and treatment.

Hypothesis Testing - User Conversion rate	
H0 (accepted fact, or status quo):	
User Conversion rate of Control group (A) is EQUAL to User Conversion rate of treatment group (B)	
H1: (everything else)	
User Conversion rate of Control group (A) is NOT EQUAL to User Conversion rate of treatment group (B)	
Z-Test	0.0001
H0 Results	REJECT
	Statistically significant
alpha	0.05
Note: REJECT if the p-value is less than 0.05. If not, FAIL TO REJECT.	

For Hypothesis Testing of Average amount spent per user, t-distribution, and a 5% significance level were used and unequal variance was assumed. T-Test resulted in **p-value = 0.944**, **statistically insignificant**. We **fail to reject** the null hypothesis that there is no difference in the mean amount spent per user between the control and treatment.

Hypothesis Testing - Average amount spent per user	
H0 (accepted fact, or status quo):	
Average amount spent per user of Control group (A) is EQUAL to Average amount spent per user of treatment group (B)	
H1: (everything else)	

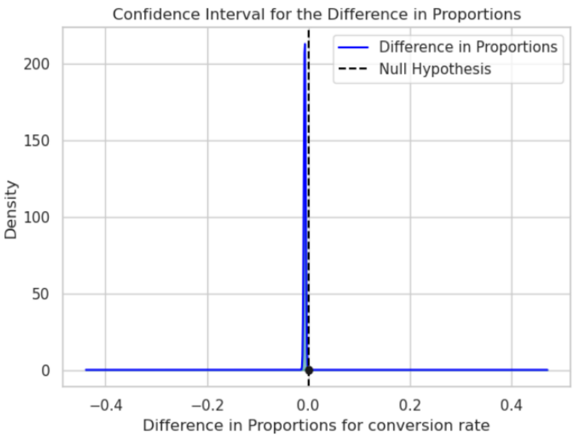
Average amount spent per user of Control group (A) is NOT EQUAL to Average amount spent per user of treatment group (B)	
T-Test	0.9438
H0 Results	FAIL TO REJECT
	Statistically insignificant
alpha	0.05
Note: REJECT if the p-value is less than 0.05. If not, FAIL TO REJECT.	

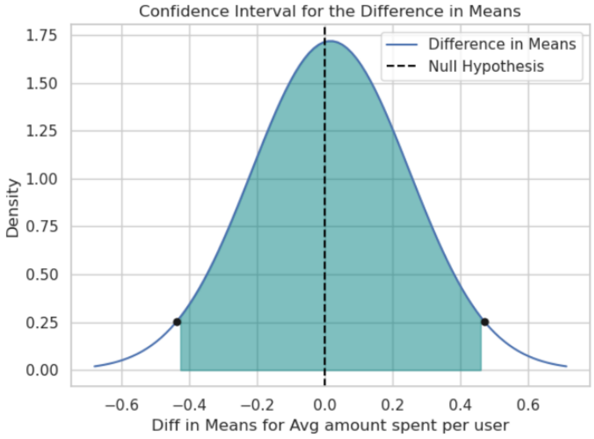
95% Confidence intervals were used to understand the magnitude of the difference between the two groups for our success metrics.

For Confidence Interval of User Conversion rate, normal distribution and unpooled proportions for the standard error were used. 95% Confidence interval for User Conversion rate was **(0.0035, 0.0107).**

For Confidence Interval of Average amount spent per user, t-distribution was used and assumed unequal variance. 95% Confidence interval for Average amount spent per user was **(-0.439, 0.471)**

Confidence Interval - User Conversion rate	
pooled proportion (p_hat)	0.04278446356
Standard error(SE)	0.001829526081
Test stat	3.86429177
Critical value	0.003585688167
Lower bound	0.0107
upper bound	0.0035
Confidence Interval:	(0.0035, 0.0107)
Confidence Interval - Average amount spent per user	
Sample statistic	-0.016
Standard error	0.2321358149
Degrees of freedom	24342
Critical value and Margin of error	0.4550094107
Lower bound	-0.439
upper bound	0.471
Confidence Interval:	(-0.439, 0.471)



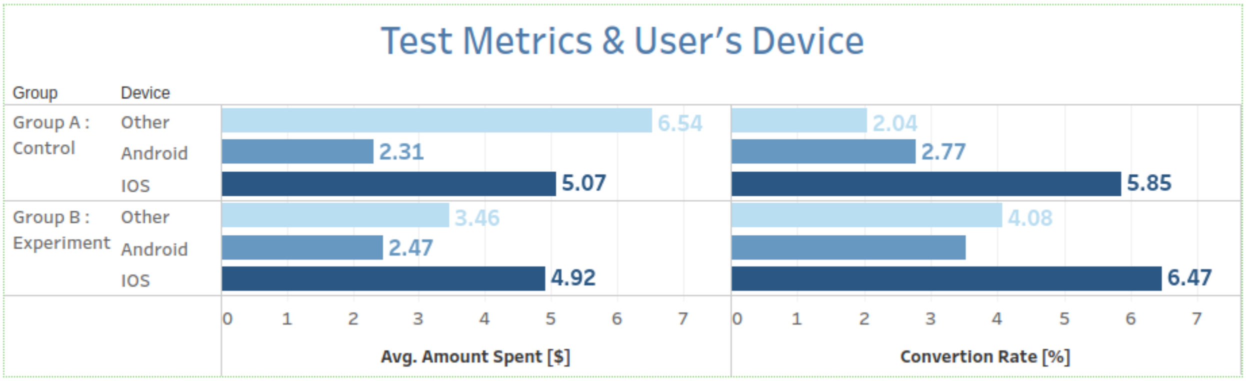


More details are available in APPENDIX section 5.2 - Formulas: Hypothesis testing & Confidence intervals. The link to a Google spreadsheet with all statistical calculations is available in APPENDIX section 5.3 - Reference Links.

Test metrics and the user's device:

Note: While looking at relationships between test metrics and the user's device, gender, and country, any missing values, were either filtered out or were treated as their own category largely based on how many observations were missing in a sample.

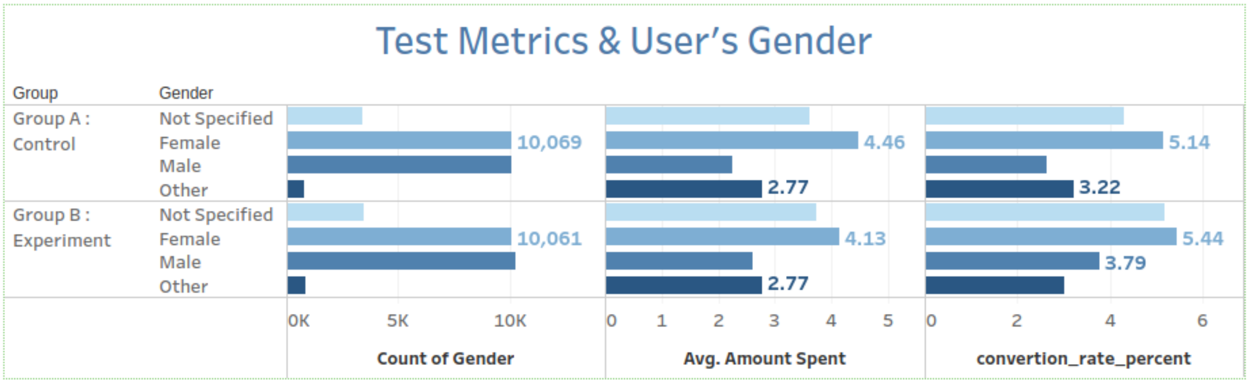
The below chart shows the relationship between the test metrics (conversion rate and average amount spent) and the user's device. The conversion rate for each test group on Android was Control: **2.77**%, Treatment: **3.52**%, whereas the same on IOS was Control: **5.85**%, Treatment: **6.47**%.



An important point to note is that, unlike the test groups, the number of users per device is not split evenly for the test. Hence, a difference between the control and treatment for a particular device, when it doesn't have many users, isn't necessarily meaningful. That resulted in giving a bit lower importance to this metric when drawing conclusions based on what we see in the visualization.

Test metrics and the user’s gender:

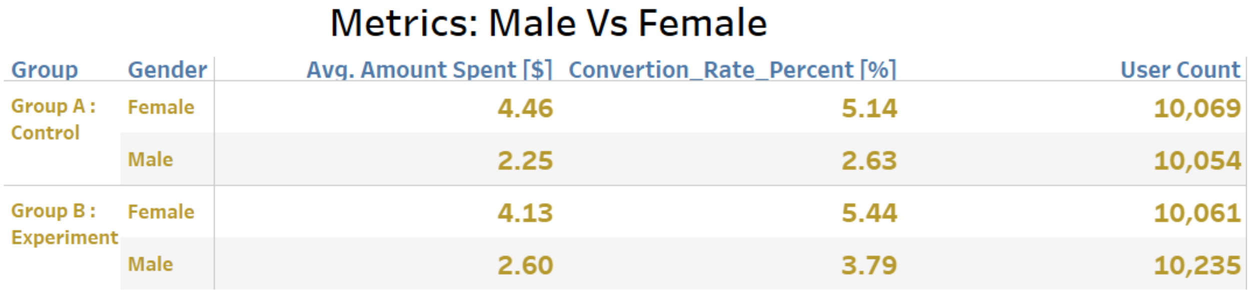
Missing values in the user’s gender column were significant, hence were treated as their own category - ‘Not Specified’.



The number of users that have missing gender in each group: Control: **808**, Treatment: **861**
Male and female users in each group were approximately the same. Both success metrics were slightly higher for the Female users. ‘Other’ gender has an average amount spent of $**2.77** for both test groups.

Female users:
Avg Amount Spent: Control: **$4.46**, Treatment: **$4.13**
Conversion Rate: Control: **5.14%**, Treatment: **5.44%**

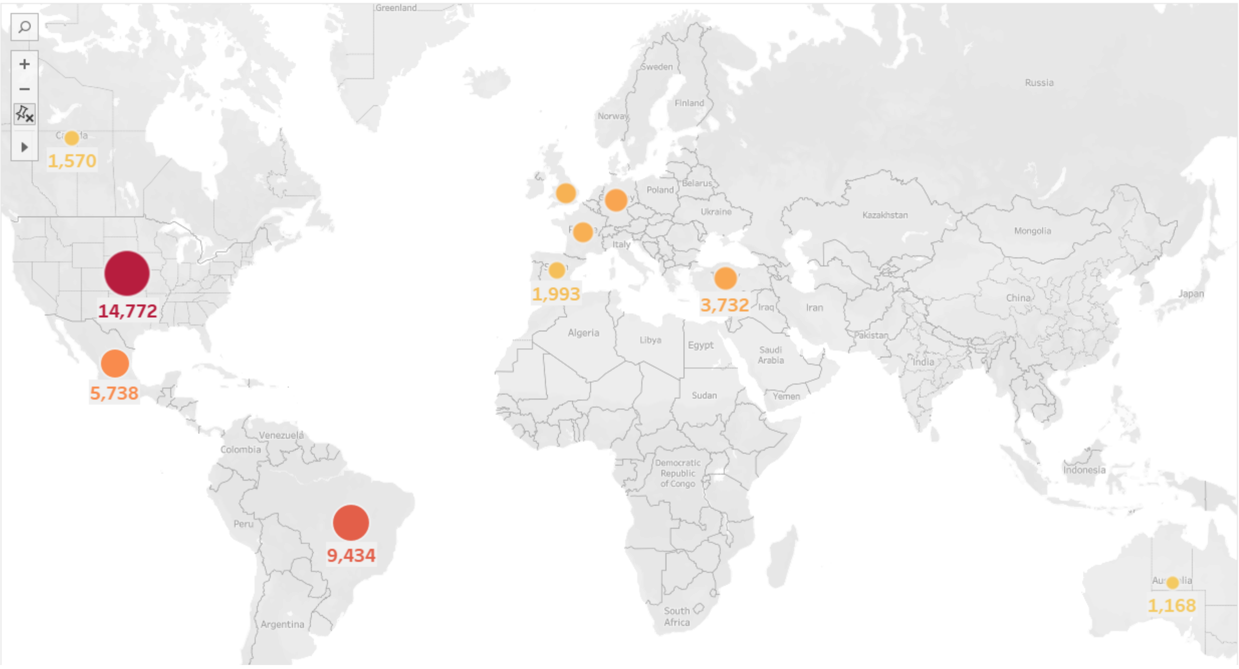
Male users:
Avg Amount Spent: Control: **$2.25**, Treatment: **$2.60**
Conversion Rate: Control: **2.63%**, Treatment: **3.79%**



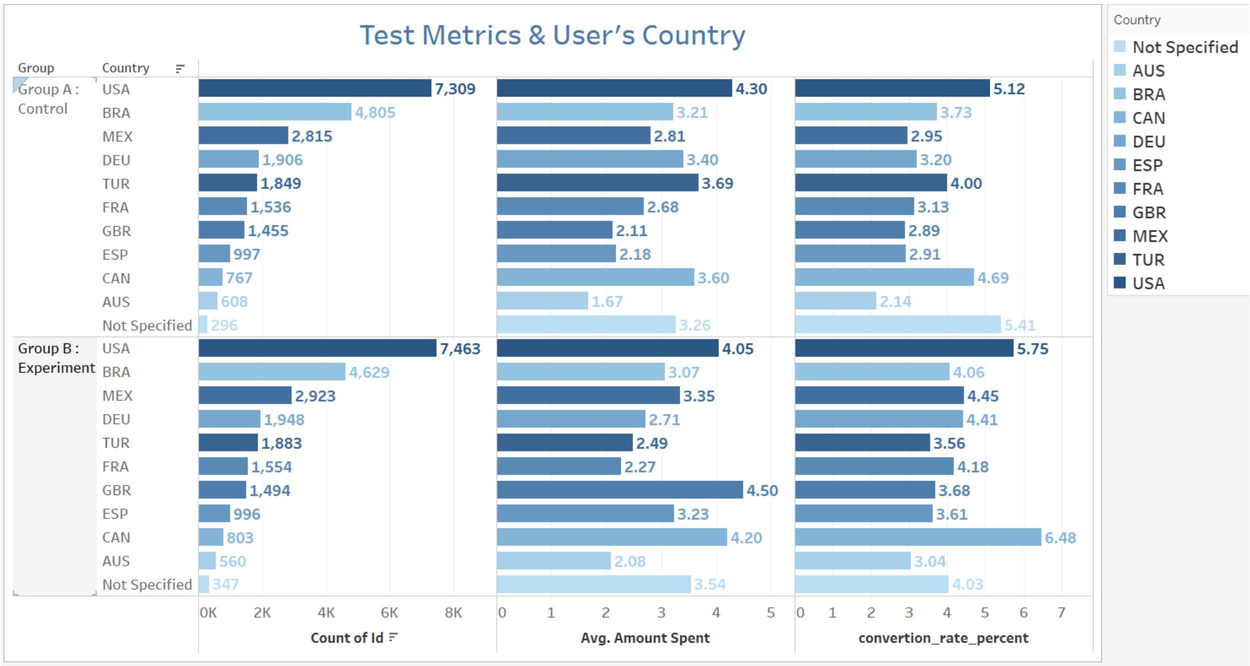
Metrics: Male Vs Female

Group	Gender	Avg. Amount Spent [$]	Convertion_Rate_Percent [%]	User Count
Group A : Control	Female	4.46	5.14	10,069
	Male	2.25	2.63	10,054
Group B : Experiment	Female	4.13	5.44	10,061
	Male	2.60	3.79	10,235

Test metrics and the user’s country:

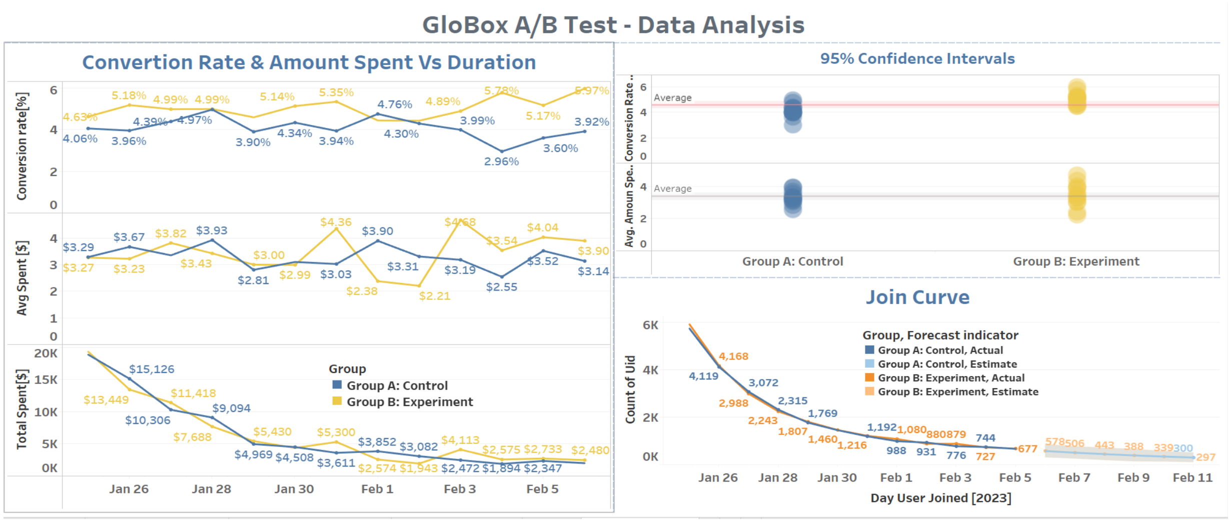


The USA had the most users in this experiment. The average amount spent for each test group in the USA was - Control: $**4.30**, Treatment: $**4.05.** The conversion rate for each test group in the USA was - Control:**5.12**%, Treatment:**5.75**%

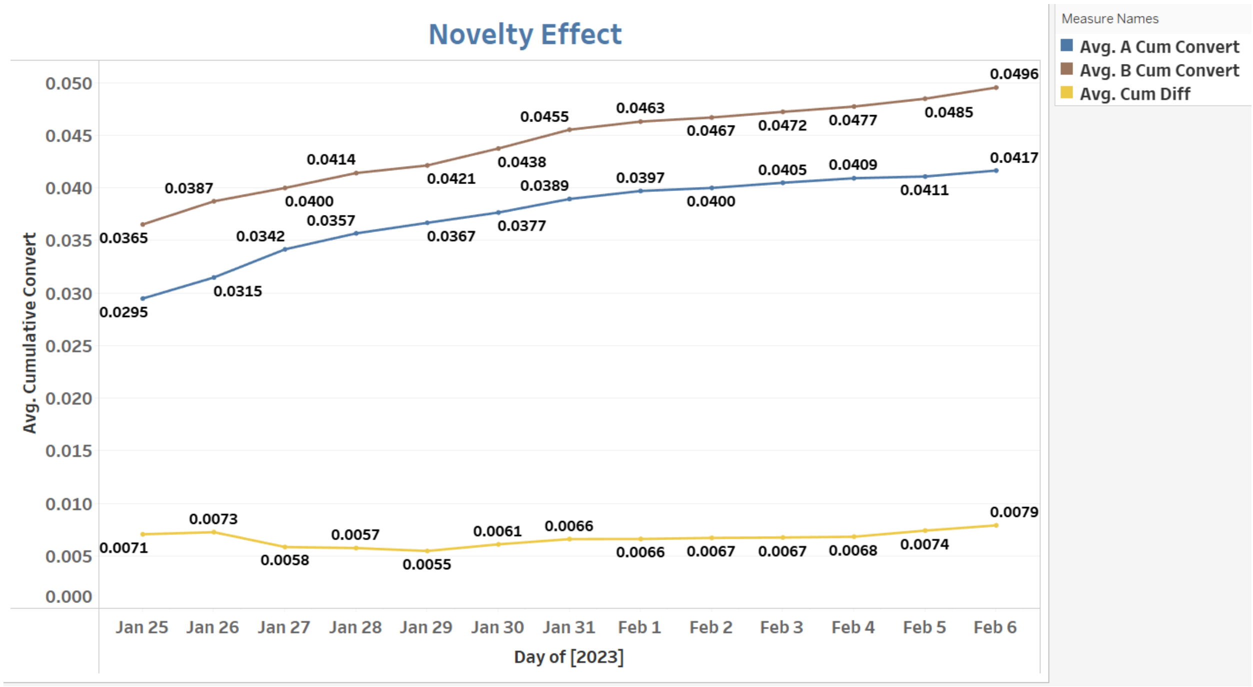


Novelty/Seasonality effect:

Moreover, during the analysis, a potential novelty or seasonality effect was identified. It indicated a decline in the total amount spent over the duration of the A/B testing. This finding suggests that external factors or user behavior might have influenced the overall spending patterns during the test period.



Novelty Effect Analysis: Cumulative



These statistical findings provide insights into the performance of the banner feature and its impact on user conversion rates and average revenue per user. The observed increase in the conversion rate highlights the potential effectiveness of the feature, while the absence of substantial improvement in other metrics suggests the need for further refinement. The identified novelty/seasonality effect also underscores the importance of considering external factors that may influence the success of the A/B test.

Power analysis

To ensure the statistical rigor of the A/B test, a power analysis was conducted using the Statsig and Statulator sample calculators. The goal was to determine the required sample size based on specific parameters.

The power analysis was performed with the following parameters:

Baseline Conversion Rate (%): 5
Minimum Detectable Effect (%): 5

Based on the calculations, the following results were obtained:
Total Sample Size: 188,000
Control Group Size: 94,000
Test Group Size: 94,000

These sample sizes were determined to provide sufficient statistical power to detect a minimum detectable effect of 5% with a baseline conversion rate of 5%.

For more detailed power calculations and further reference, the link to the **Statsig** calculator can be found in the APPENDIX - Reference Links section.

By conducting power analysis and determining the appropriate sample sizes, the A/B test can be conducted with confidence, ensuring that the results will be statistically meaningful and reliable.

2.4. Revenue Impact

To evaluate the effectiveness of the banner feature and its impact on revenue, it is essential to assess its practical significance. Considering the valuable real estate occupied by the banner, especially on smartphone screens, it becomes crucial to ensure that its implementation leads to a practically significant increase in revenue per user, surpassing the threshold of 5%.

2.5. Practical Significance

At present, the A/B test results have not achieved this target, indicating the need for further improvements. However, it is important to note that even a modest increase of 3% to 5% in average revenue per user holds practical significance for an online e-commerce company. Therefore, investing in another iteration of the A/B test is highly recommended.

By conducting another data-driven iteration, we can gather additional conclusive results that will enable us to make informed decisions. It is important to consider the potential costs and efforts associated with the iterative process. However, the potential benefits of achieving a practically significant increase in revenue per user outweigh these considerations, making further improvements and tests a worthwhile investment.

This approach ensures that decisions are driven by data and are aligned with the goal of maximizing revenue and optimizing the user experience on the platform.

3. Recommendations

Based on the analysis of the current A/B test results, although the success was limited, there are promising insights that suggest potential enhancements to the banner experience. In order to achieve better results in subsequent tests, the following recommendations are proposed:

1) Enhance Clickability: Update the banner design to make it visually clickable and inviting. Consider adding a button to prompt user engagement, increasing the likelihood of interaction and conversions.
2) Conduct Another A/B Test: Design a new A/B test with a larger sample size of **188,000**, increased by three times from the previous test. This larger sample size will enhance the statistical power of the test, providing more reliable and robust results for analysis.
3) Extend Test Duration: Increase the duration of the test by **one additional week**. This extension will allow for the identification and analysis of long-term trends, particularly related to the novelty/seasonality effect. Understanding these trends can provide valuable insights for further decision-making.
4) Perform Comprehensive Data Analysis: Upon completion of the extended test, conduct a thorough data analysis to gain deeper insights into user behavior and the impact of the banner feature. Explore additional success metrics, segment the data based on user characteristics, and analyze the overall performance of the feature.

4. Conclusion

In conclusion, while the current A/B test results did not provide sufficient confidence to proceed with the release of the banner feature, the recommended improvements and further iterations of the test present a significant opportunity for achieving the desired impact on key success metrics. The potential increase in revenue and the practical significance of these improvements make it worthwhile to invest in additional testing, analysis, and refinement. By following these recommendations, GloBox can enhance its user experience and drive improved outcomes in revenue generation.

5. APPENDIX

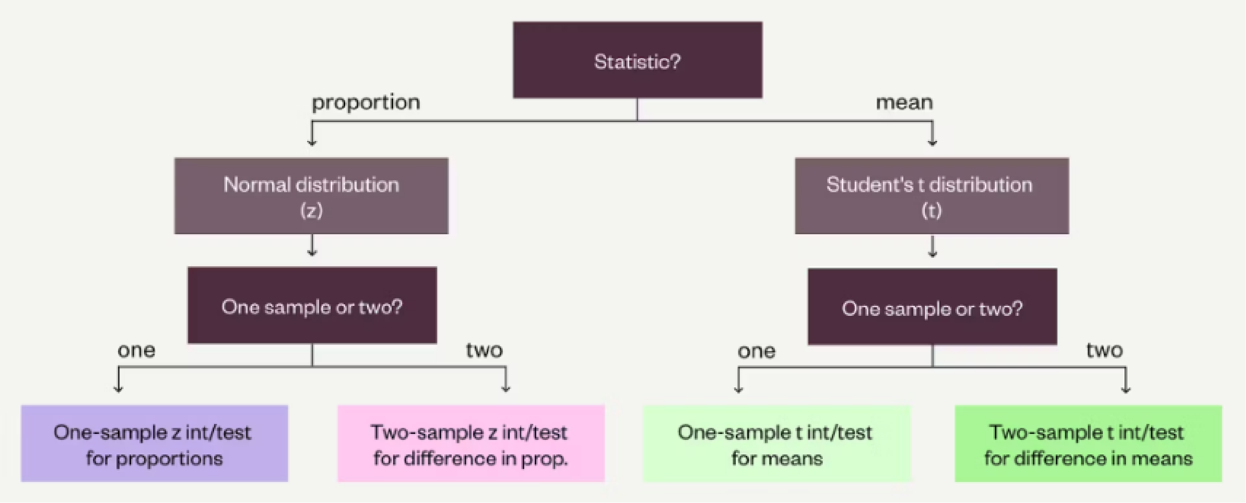
5.1. SQL data extraction

```sql
-- SQL query that returns: the user ID, the user’s country, the user’s gender, the user’s device type, the user’s test group, whether or not they converted (spent > $0), and how much they spent in total ($0+).

DROP VIEW IF EXISTS globox;

CREATE VIEW globox
AS
(SELECT u.id, u.country, u.gender,
g.device, g.group, SUM(COALESCE(a.spent::numeric, 0)) AS total_spent,
CASE
WHEN SUM(COALESCE(a.spent::numeric, 0)) > 0 THEN 'Yes'
ELSE 'No'
END
AS is_converted

FROM "groups" AS g
LEFT JOIN "users" AS u
ON g.uid = u.id
LEFT JOIN "activity" AS a
ON g.uid = a.uid
GROUP BY u.id, u.country, u.gender,
g.device, g.group
ORDER BY u.id);

SELECT *
FROM globox;

-------------

-- alternate way
select u.id as user_id,
g.group as test_group,
u.country, u.gender, g.device,
sum(coalesce(a.spent,0)) as amount_spent,
case when sum(a.spent) > 0 then 1 else 0 end
as converted
from users as u
inner join groups as g
on u.id = g.uid
left join activity as a
on u.id = a.uid
group by 1,2,3,4,5;

-------------

--alternate way with dates included (only first date per user)
SELECT g.uid, g.join_dt AS date_joined, u.country, u.gender, a.dt AS
date_converted,
g.device, g.group, SUM(COALESCE(a.spent::numeric, 0)) AS total_spent,
CASE
WHEN SUM(COALESCE(a.spent::numeric, 0)) > 0 THEN 1
ELSE 0
END
AS is_converted

FROM "groups" AS g
INNER JOIN "users" AS u
ON g.uid = u.id
LEFT JOIN "activity" AS a
ON g.uid = a.uid
GROUP BY g.uid, g.join_dt, a.dt, u.country, u.gender,
g.device, g.group
ORDER BY g.uid;

-------------

-- Query for checking novelty effect

with join_dt_agg as (
select g.join_dt as dt, g.group as test_group,
count(distinct u.id) as user_count
from users as u
```

```sql
inner join groups as g
on u.id = g.uid
group by 1,2
),

convert_dt_agg as (
select a.dt, g.group as test_group,
count(distinct a.uid) as converted_user_count
from groups as g
left join activity as a
on g.uid = a.uid
group by 1,2
),

cumulative_users as (
select j.test_group, j.dt,
j.user_count, c.converted_user_count,
sum(j.user_count) over (partition by j.test_group order by j.dt) as
cum_users,
sum(c.converted_user_count) over (partition by j.test_group order by j.dt)
as cum_converted_users
from join_dt_agg as j
inner join convert_dt_agg as c
on j.dt = c.dt and j.test_group = c.test_group
order by 1,2
),

cumulative_conversion as (
select *, cum_converted_users/cum_users as cum_conversion_rate
from cumulative_users
)

select a.dt,
a.cum_conversion_rate as a_cum_convert,
b.cum_conversion_rate as b_cum_convert,
b.cum_conversion_rate - a.cum_conversion_rate as cum_diff
from cumulative_conversion as a
inner join cumulative_conversion as b
on a.dt = b.dt
and a.test_group = 'A' and b.test_group = 'B';
```

5.2. Formulas: Hypothesis testing & Confidence intervals

Flow chart below serves as a quick reference for terminology and formulas for both confidence intervals and hypothesis testing.


Statistic?
proportion
mean
Normal distribution (z)
Student's t distribution (t)
One sample or two?
one
two
One sample or two?
one
two
One-sample z int/test for proportions
Two-sample z int/test for difference in prop.
One-sample t int/test for means
Two-sample t int/test for difference in means

Formulas used:

1) **Hypothesis Testing - User Conversion rate:**

n1 = total users in group A

n2 = total users in group B

p1 = conversion rate for group A

p2 = conversion rate for group B

pooled proportion (p_hat) = ((p1*n1)+(p2*n2))/(n1+n2)

SE = sqrt(p_hat * (1 - p_hat) * (1/n1 + 1/n2))

Test stat = (p2-p1) / SE

p-value = 2 * NORM.S.DIST(-ABS(Test stat))

2) **Hypothesis Testing - Average amount spent per user:**

n1 and n2 are the same

s1 = stdev for group A

s2 = stdev group B

x1 = mean group A

x2 = mean group B

diff in means (x_hat) = x1 - x2

SE = sqrt((s1^2/n1) + (s2^2/n2))

Test stat = (abs(x_hat) - u0) / SE

p-value =T.DIST.2T(test stat, degrees of freedom)

5.3. Reference Links

- Globox test statistics in Google Spreadsheet
- GitHub repository

- Globox A/B test data analysis visualizations in Tableau:
    - GloBox-Data Story
    - GloBox-Advanced analysis
    - Novelty effect: cumulative

- Power analysis using 'Statsig' calculator