A/B TEST REPORT

GloBox

July 28, 2023

Subasini Ravindran
Data Analyst

Summary:

An A/B test was conducted to find the effects of introducing a new banner in the GloBox app to bring awareness to the food and drink category to increase their revenue. The users were randomly grouped into two groups. The difference in the conversion rate and spending averages between the groups was observed. We discovered that there is sufficiently strong evidence that the conversion rates of the group with the banner are higher.

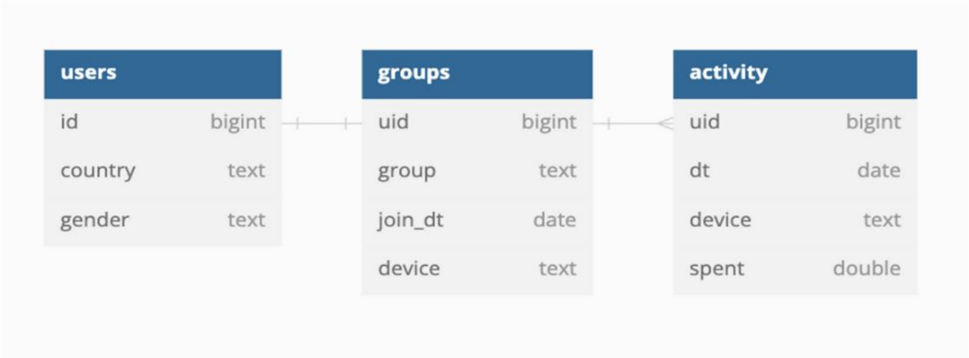
While some promising results indicate launching the banner could potentially improve the customer acquisition and revenue aspects of the business, we didn't see enough improvements to be confident to launch the feature. I recommend to re-iterate the experiment with a large enough sample size and analyzing the data to get better confident results.

Context:

GloBox is primarily known amongst its customer base for boutique fashion items and high-end decor products. However, their food and drink offerings have grown tremendously in the last few months, and the company wants to bring awareness to this product category to increase revenue. So, the Software Engineers of the Growth Product and Engineering Team have developed a banner that highlights their key products from the food and drinks category. Before launching the feature, the team wants to conduct an A/B test to decide whether the feature is worth launching. The setup of the A/B test is as follows:

1. The experiment is only being run on the mobile website.
2. A user visits the GloBox main page and is randomly assigned to either the control(A) or treatment(B) group. This is the join date for the user.
3. The page loads the banner if the user is assigned to the treatment group and does not load the banner if the user is assigned to the control group.
4. The user subsequently may or may not purchase products from the website. It could be on the same day they join the experiment, or days later. If they do make one or more purchases, this is considered a "conversion".

GloBox stores its data in a relational database, which can be accessed using this link in a SQL editor like Beekeeper Studio. The test was conducted from Jan 25, 2023, to Feb 6, 2023. The structure of the database follows:



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

Results:

Data Extraction:

The data was extracted from the database using the following SQL script.

```
SELECT
    u.id,
    u.country,
    u.gender,
    g.device,
    g.group,
    CASE
        WHEN Sum(a.spent)>0 THEN 'Converted'
        ELSE 'Not Converted'
    END AS has_converted,
    SUM(a.spent) AS tot_spending
FROM users u
JOIN groups g ON u.id = g.uid
FULL JOIN activity a ON g.uid = a.uid
GROUP BY u.id, u.country, u.gender, g.device, g.group;
```

Then the data was downloaded into an Excel spreadsheet for further cleaning and analysis. The dataset had some null values in the columns country, gender, and device. Since it did not affect the conversion rates and average spending calculation, the null values were not removed. The data was checked for duplicates and anomalies. Now, the data is ready for analysis.

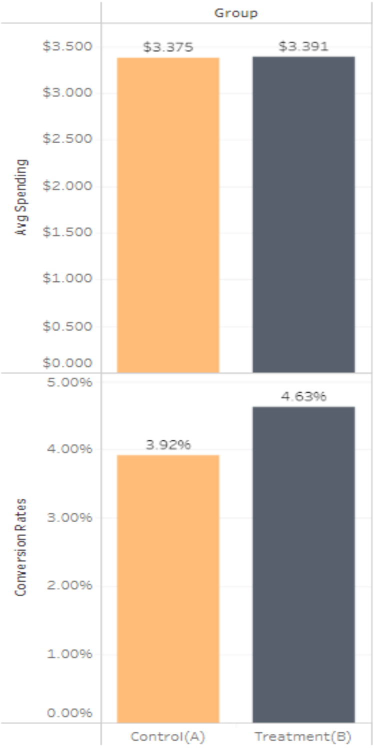
Hypothesis Testing:

I conducted a hypothesis test to find the difference in the conversion rates between the control and treatment groups. Please refer to the calculations tab of the attached Excel workbook. The null hypothesis, H_0, is that there is no difference in the conversion rates between the groups. The alternative hypothesis, H_a, is that the conversion rate of one group is more than the other. The significance level considered here is 0.05. The calculated p-value of this sample is 0.0001, which is less than the significance level. Thus, concluding that the result is statistically significant, and rejecting the null hypothesis.

The 95% confidence interval of the difference in conversion rate between the treatment-control group is 0.35% to 1.07% with a mean of 0.71%. The low variability or low spread of the data implies that the individual data points in the treatment group are tightly clustered around the mean conversion rate. This suggests that the observed effect is consistent and stable, making the results more reliable.

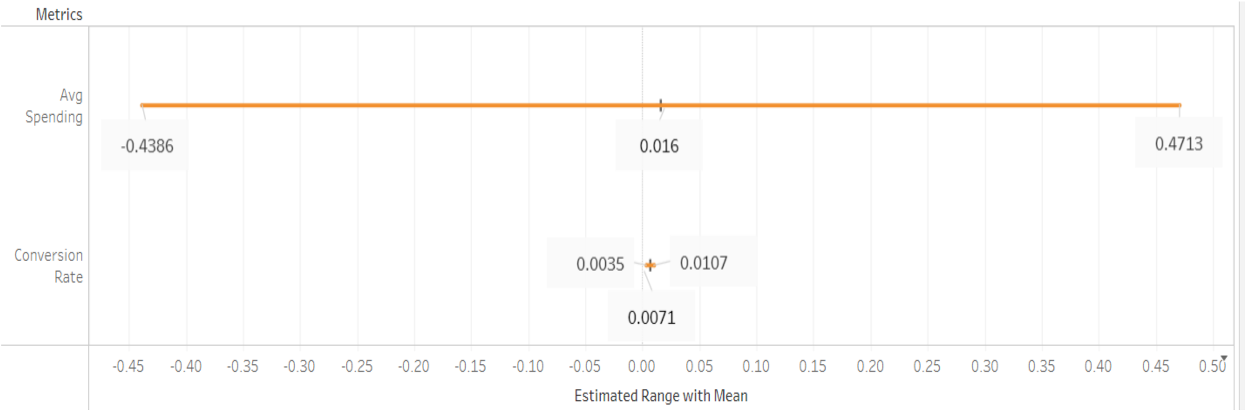
Based on the statistical significance, higher conversion rate, and low variability in the treatment group, the conclusion is that implementing the banner in the treatment group has a positive effect on the conversion rate compared to the control group.

Comparison of Conversion Rate and Spending Between Test Groups:



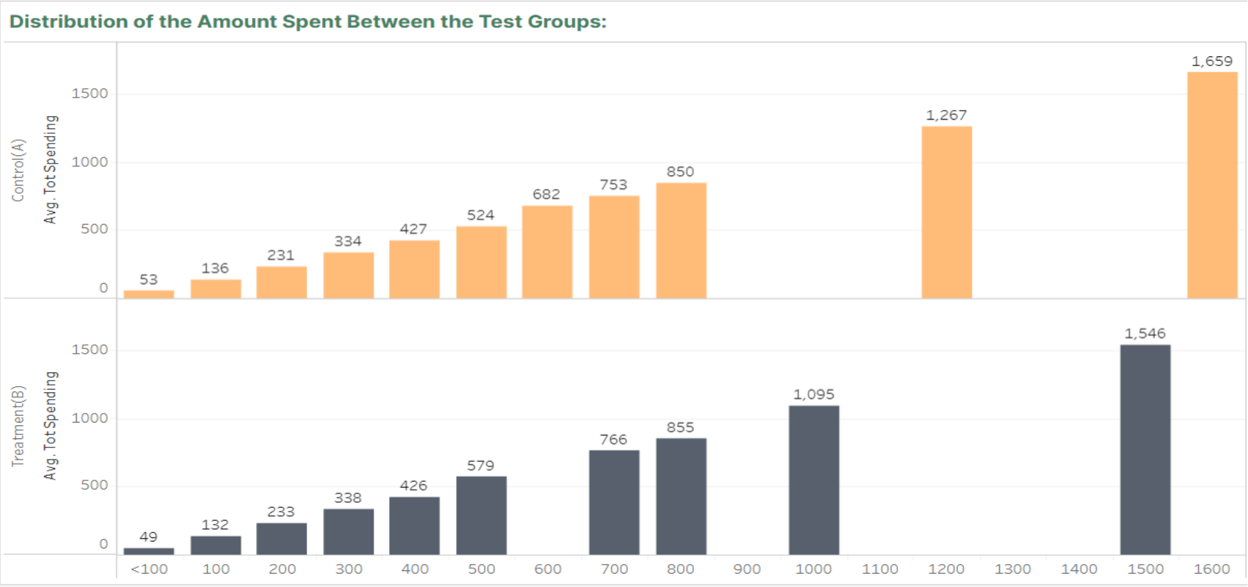
Next, a second hypothesis test was conducted to find whether there is a difference in the average amount spent per user between the two groups. Please refer to the same calculations tab of the attached Excel workbook. The null hypothesis, H_0, is that there is no difference in the average amount spent per user between the two groups. The alternative hypothesis, H_a, is the average amount spent per user in one group is more than the other. The significance level is 0.05. The calculated p-value is 0.944, which is higher than the significance level. Thus, concluding that the result is statistically insignificant, and we fail to reject the null hypothesis.

The 95% confidence interval of the difference in the average amount spent per user between the groups is -0.439 to 0.471 with a mean of 0.016. Even though the difference is statistically insignificant, the spread is large. If we assume a practical significance of 2%, though not statistically proven, is practically meaningful and relevant. So, we should consider conducting replication studies with larger sample sizes to increase the power to detect smaller effects.



Distribution of total amount spent between the groups:

Here the average amount spent by the users is visualized in increments of 100 as a histogram. Both the control and treatment groups show the same trend.



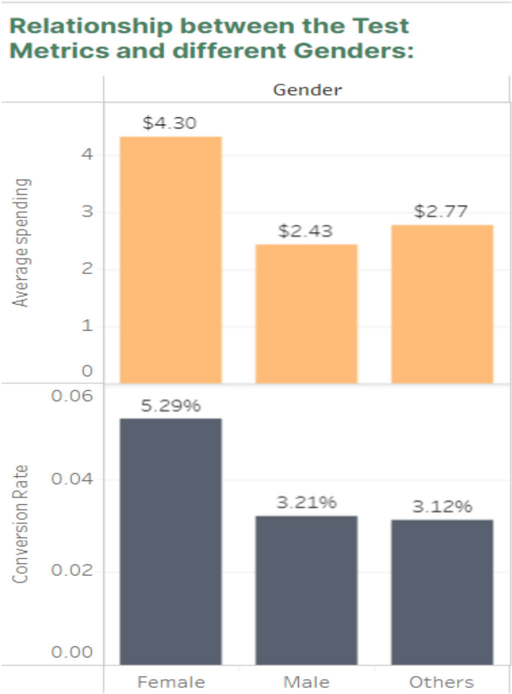
Distribution of the Amount Spent Between the Test Groups:

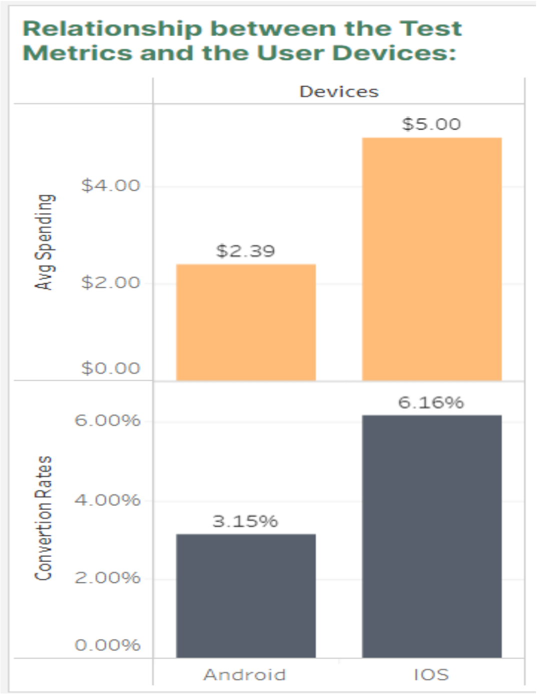
Comparison of the key metrics among different Genders, User Devices, and Countries:

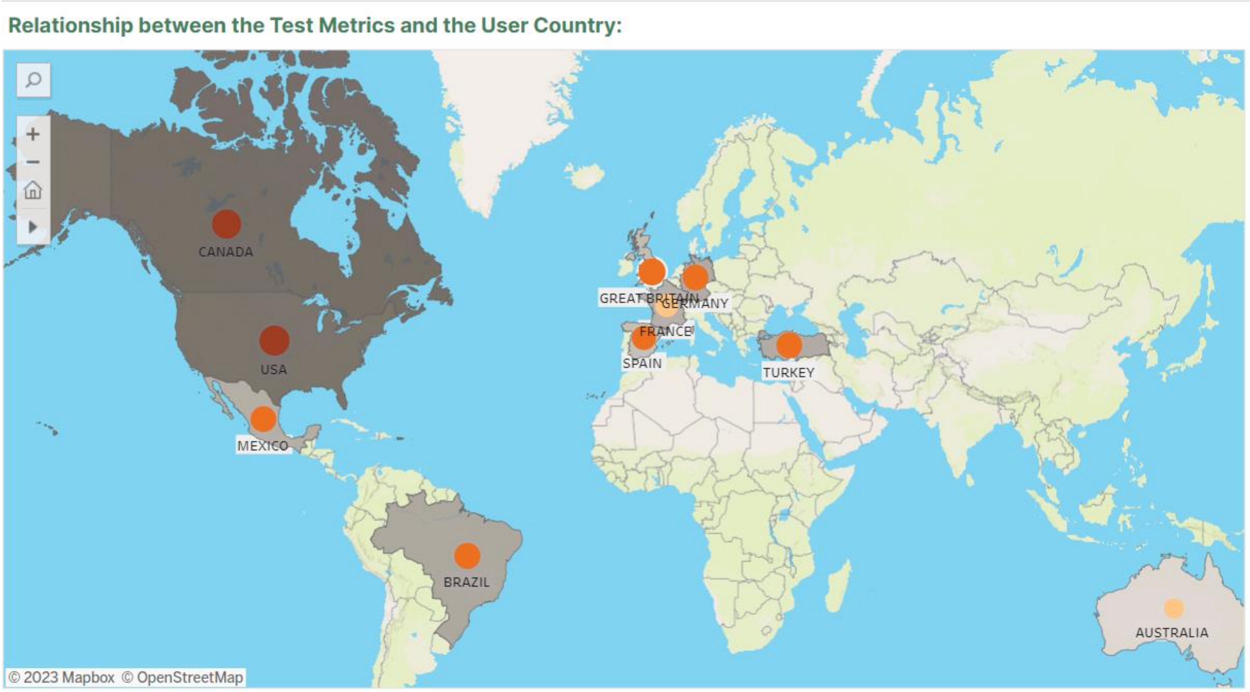
Unlike the test groups, the number of users per gender, device, or country is not split evenly. We might observe a big difference between the control and treatment groups for any device, gender, or country, but if it doesn’t have many users, then it isn’t necessarily meaningful.

The female users came at the top for both conversion rates and average spending. The difference between the male and other users is minor but female users are exceptionally high in this sample. This could imply that the majority of their customer base is female.

Likewise, Ios users have high spending averages and conversion rates than Android users. This could also imply that most of their users use their IOS app to shop for their products.


Relationship between the Test Metrics and different Genders:

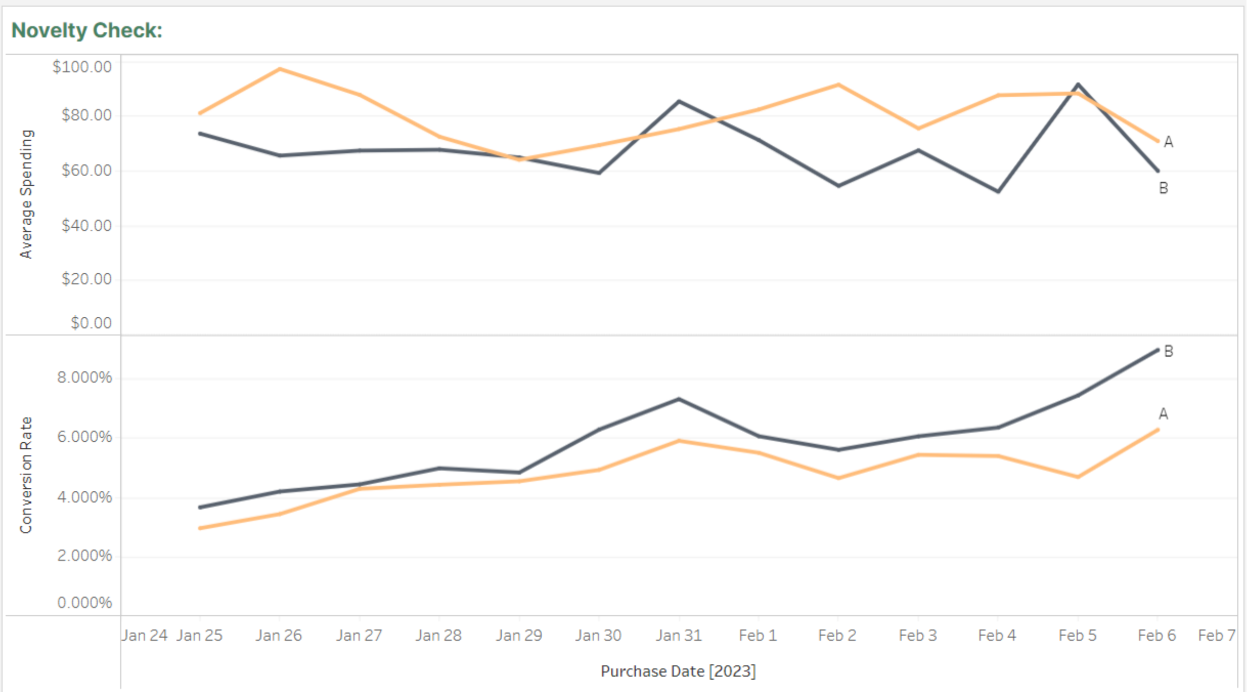

Relationship between the Test Metrics and the User Devices:


Relationship between the Test Metrics and the User Country:

The users from the USA and Canada have high conversion rates and spending averages. Australia is the least in both. This could imply that these two countries are the primary sources of our user base.

Novelty Check:

I inspected the difference in the key metrics throughout our experiment for the novelty effect. The experiment was conducted only for a short period. Data with dates are extracted from the database.



It is clear from the above graph that control and treatment groups have similar trends over the period. Except the control group almost always performed better than the treatment group in spending average and vice versa in conversion rates.

Power Analysis:

Let’s check if the data is statistically powerful enough for our results to be sufficiently sensitive. The power is calculated using this link. With the conversion rate of the control group as the baseline, 10% MDE, 5% significance level, and 80% statistical power as our metrics, the sample size should be 76900 to be powerful enough. Our current sample size is 48943. Our sample does not have enough statistical power to produce the correct results. The risk of a Type II error is inversely related to the statistical power of a study. So, the chances of producing a false negative result are higher here.

Recommendations:

The difference in the conversion rates between the treatment-control groups is statistically significant. Assuming a practical significance of 2%, the confidence interval is not practically significant. The difference in the spending average between the treatment-control groups is statistically significant. With the same assumed practical significance, the confidence interval spread is high, and it extends beyond the practical significance. The period of the experiment is too short to notice a novelty effect. Finally, the sample is not statistically powerful enough to produce sufficiently sensitive results.

While some promising results indicate launching the banner could potentially improve the customer acquisition and revenue aspects of the business, we didn't see enough improvements to be confident in launching the banner. I recommend to re-iterate the experiment with a large enough sample size and analyzing the data to get better confident results.

Appendix:

1. Refer to the Tableau dashboard.
2. Excel and SQL query sheets are attached to the report.