Goblox A/B Test Analysis

By: Satakshi Salaria

Summary

This A/B test assessed a landing page banner that included the product category for food and beverages. The conversion rate significantly increased, going from 3.92% in the control group to 4.63% in the treatment group. But the average amount spent per user, which was $3.37 in the control group and $3.39 in the treatment group, did not significantly alter. The results are broken out in greater depth by device, gender, and continent. Confidence interval, Novelty effects, and Power analysis are performed to further assess the data.

Context

The Growth team chooses to conduct an A/B test that has a banner at the top of the website that promotes key products in the food and beverage category. The treatment group sees the banner while the control group does not.

The two test Group are as follows:

A. Control Group: existing landing page
B. Treatment Group: landing page with food & drink banner

The A/B test is set up as follows:

- Only the mobile website is being used for the trial.
- A user is randomly allocated to the control or test group when they access the GloBox home page.
- If the user is part of the test group, the page loads the banner, if the user is part of the control group, the page does not load the banner.
- The user may then decide whether to buy things from the website. It can happen the day they sign up for the trial or days afterwards. It is referred to as a "conversion" if they make one or more purchase.

Test Parameters

We ran this experiment with the following parameters:

• Date Range: June 12 – July 9’2023
• Total Users: 48,943
• Platform: Mobile website (Android/IOS)
• Countries: AUS, BRA, CAN, DEU, ESP, FRA, GBR, MEX, TUR, USA
• Traffic split: 50/50
• Confidence Interval: 95%
• Normal distribution and Significance level: 5%

Results

We discovered that although the banner enhanced conversion rates, the spending of the converted users in the treatment group prevented an increase in income. This is evident from the users' median expenditure, which was $55 in the treatment group and $65 in the control group for those who converted.

Test Group Results

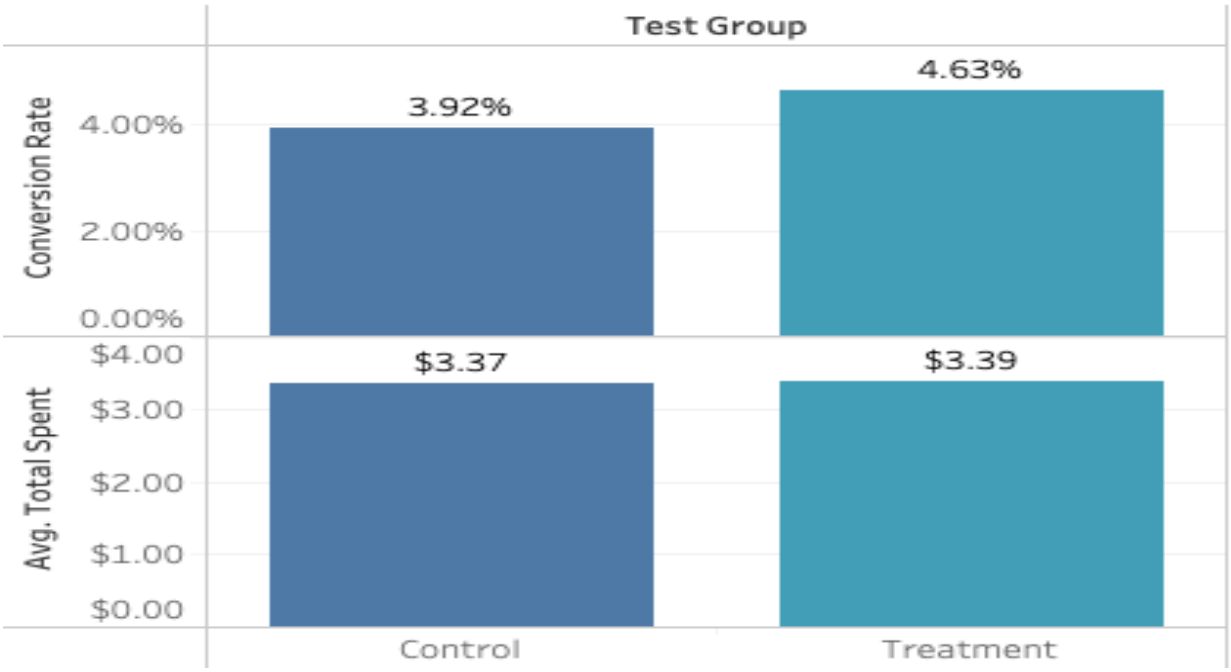
Conversion Rate:

- The conversion rate in the control group was 3.92%.
- The conversion rate in the treatment group was 4.63%.
- This represents an 18% relative change, calculated as (4.63% - 3.92%) / 3.92%.
- The hypothesis test resulted in p = 0.0001, which is less than the significance level of 0.05.
- Since the p-value is less than 0.05, the results are statistically significant.
- Therefore, it can be concluded that the conversion rates between the two groups are not equal.

Average Amount Spent per User:

- The average amount spent per user in the control group was $3.37.
- The average amount spent per user in the treatment group was $3.39.
- This represents a 0.5% relative change, calculated as ($3.39 - $3.37) / $3.37.

- The hypothesis test resulted in p = 0.94, which is greater than the significance level of 0.05.
- Since the p-value is greater than 0.05, the results are not statistically significant.
- Therefore, it cannot be concluded that the average amounts spent per user between the two groups are not equal.



Confidence Interval for the different two groups:

Conversion Rate: We have a 95% confidence level that the difference in conversion rates between the treatment and control is between 0.35% and 1.07%. Since it is evident that this interval does not include zero, the outcome is statistically significant.

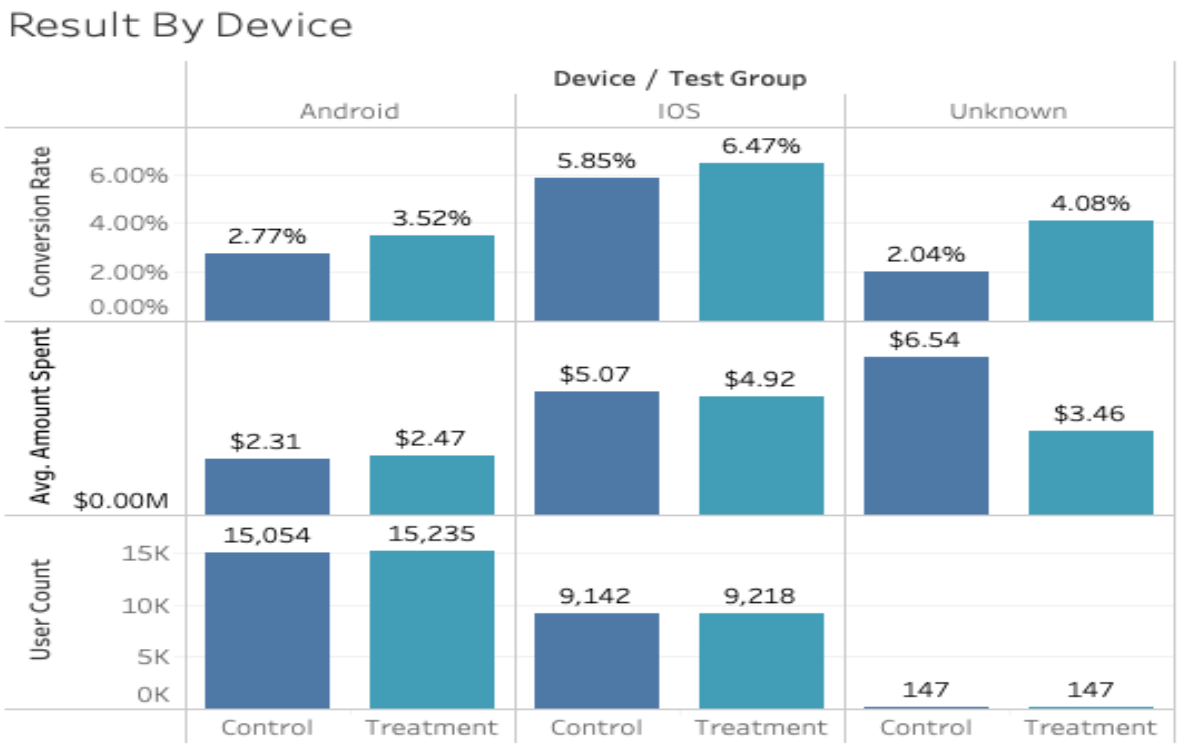
Average Amount Spent per User: We are 95% certain that the difference in the average amount spent per user between the treatment and control is between -$0.44 and $0.47. Our finding is not statistically significant since this interval is virtually perfectly centred around zero.

Result Breakdown

Breakdown of result by device, gender, and continent:

Device

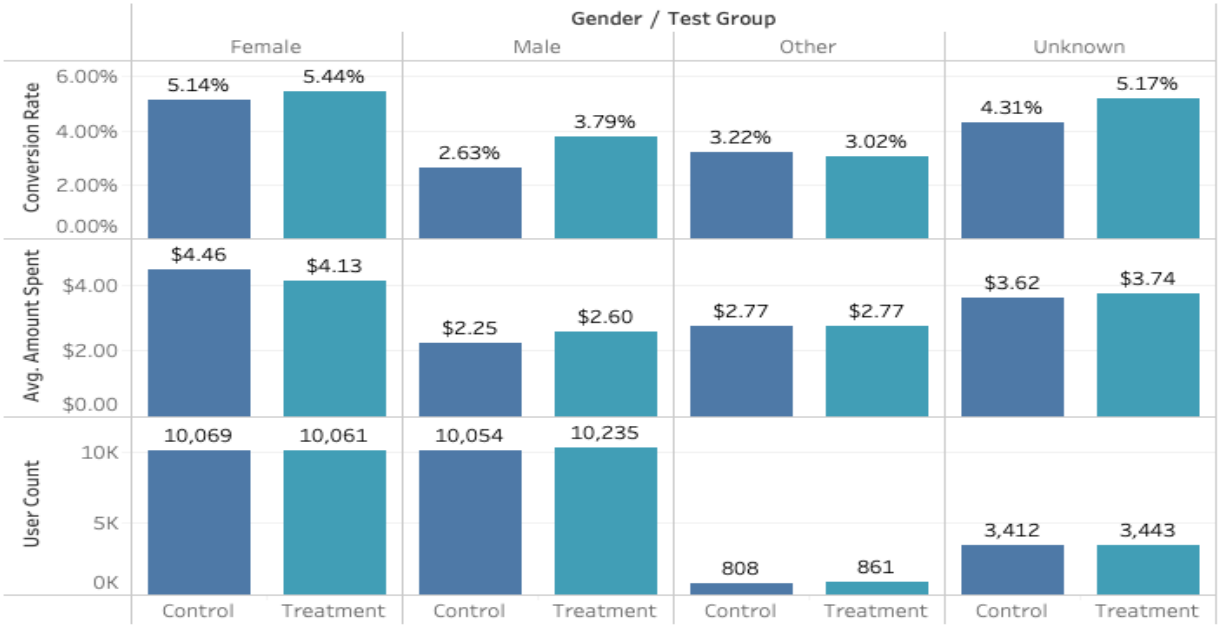
When we segment by device, we see that both IOS and Android had a higher conversion rate with the treatment group. But we observed the average amount spent per user has very slight change in android device whereas with IOS devices the average amount is decreasing with relative change of 3%. There were some users with unknown devices, they also showed the relative decrease of 47% in the amount of spending between control group ($6.54) and treatment group ($3.46).



Gender

The distribution of users for both genders is nearly the same across platforms and countries. When we break down users by gender, we can observe that male users' conversion rates rose the fastest. The relative change for "female" was 6%, "male" increased by 44%, "other" decreased by 6%, and "unknown" increased by 20%. Additionally, we see a 7% decline in average expenditure per user for females and a 16% relative rise for males. As a result, more males than women are seen switching to the banner.

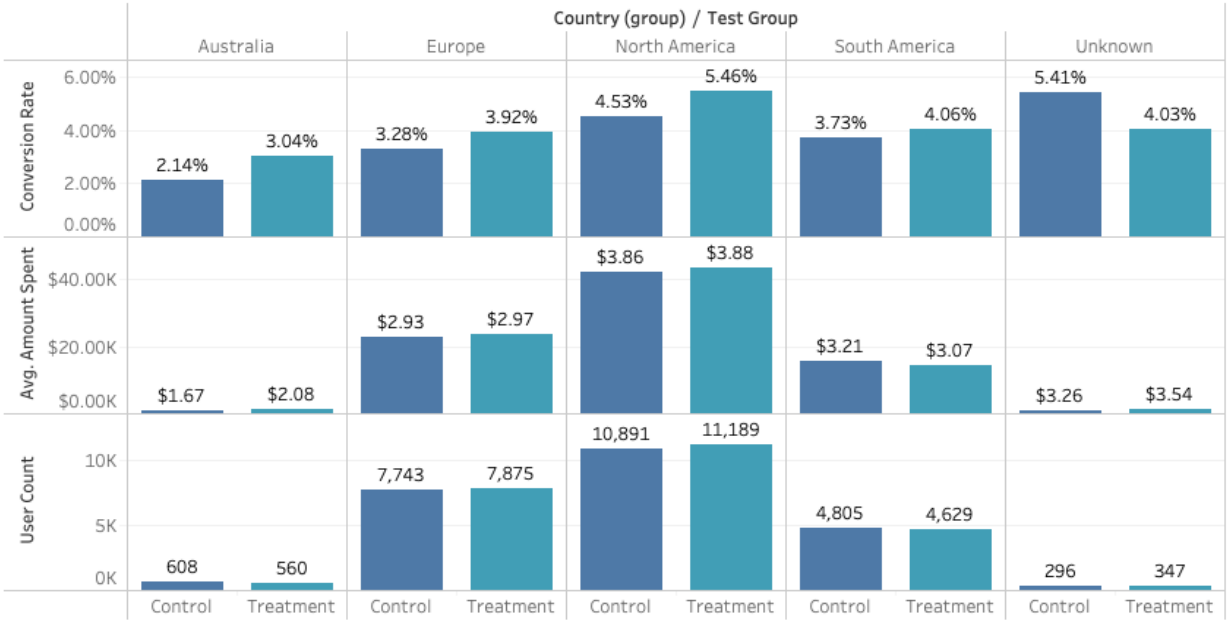
Result By Gender



Continent

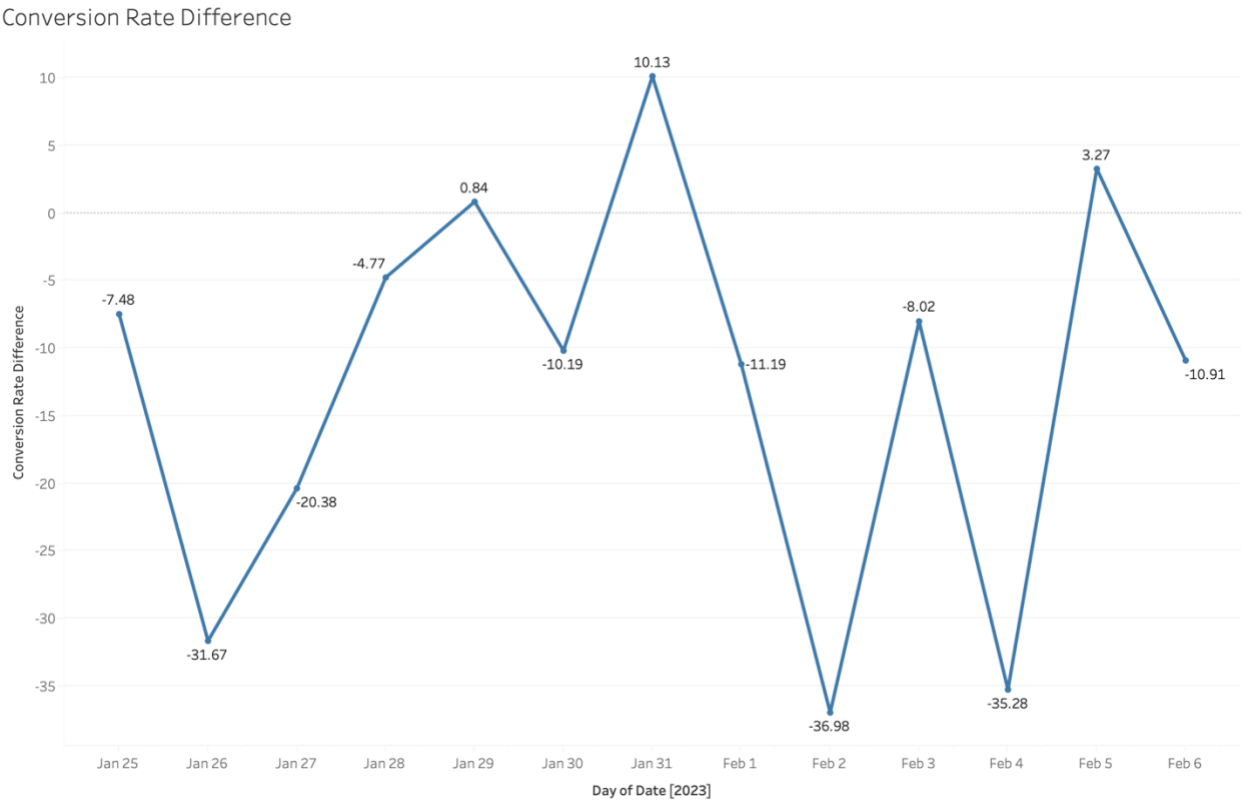
While the average amount spent by a user was steady, conversion rates are rising across the continent. We see a relative shift in conversion rate: 20% increase in North America, 19% increase in Europe, 9% increase in South America, 42% increase in Australia, and 25% decrease in Unknown. Due to the limited sample size, the change for Australia and Unknown is most likely unstable.

Result By Continents



Novelty Effect

Our conversion rate findings do not seem to show a definite novelty impact. Looking at the variation in conversion rates over time between the two groups allows us to quickly verify our hypothesis. As if we had halted the A/B test on that day and examined the overall results, this displays the difference in cumulative conversion rate, which is all the users who had joined and converted up to that point in time.



Power Analysis

According to the power analysis, the sample sizes required to detect a 10% change in both measures were not met. We want to make sure that the sample size we have is large enough to detect the necessary change in conversion rate and average spending. The power analysis suggests a total sample size to increase to minimum Sample size to launch banner in the mobile website. We concluded that to debut the banner, a 10% relative change in each measure would be needed .10% is a high threshold but considering that the banner is placed in a very valuable area of the page, this is the price to pay.

Metrics	Minimum Sample Size	Inputs
Conversion Rate	60.6K	Baseline Conversion Rate: 3.92% Minimum Detectable Effect: 10% Hypothesis: One-sided Test A/B Split Ratio: 0.5

		Significance: 0.05 Statistical Power: 0.8
Average Amount Spent	5,172,010,668	Mean of the Reference Group: .337 Mean of the Test Group: .339 Standard Deviation: 25.67

Calculators used: Statsig, Statulor

Recommendation

We didn't see enough improvement in our metrics of success, so it's not a good idea to release the banner to all users. The perceived cost of launching the feature is not worth it based on the results of the A/B test. If we are very interested in this feature, then we keep running the test for a longer duration to reach the desired sample size of 186k. Even our conversion rate findings do not seem to show a definite novelty impact.

Appendix

SQL Query

Query to get user-level dataset:

```
SELECT u.id AS "User ID",
        u.country AS "Country",
        u.gender AS "Gender",
        g.device AS "Device Type",
        g.group AS "Test Group",

        CASE
        WHEN SUM (a.spent) > 0 THEN 1
        ELSE 0 END AS "Converted",
        COALESCE (SUM (a. spent), 0) AS "Total Spent"
FROM users u
LEFT JOIN
        activity a ON u.id = a.uid
LEFT JOIN
        groups g ON u.id = g.uid
GROUP BY u.id,
        u.country,
        u.gender,
        g.device,
        g.group;
```

Query for Novelty Effect:

```
SELECT a.dt AS date,
        (SUM(CASE WHEN g.group = 'B' THEN a.spent ELSE 0 END) / COUNT(DISTINCT
        CASE WHEN g.group = 'B' THEN u.id END)) AS treatment_conversion_rate,
        (SUM(CASE WHEN g.group = 'A' THEN a.spent ELSE 0 END) / COUNT(DISTINCT CASE
        WHEN g.group = 'A' THEN u.id END)) AS control_conversion_rate,
        (SUM(CASE WHEN g.group = 'B' THEN a.spent ELSE 0 END) / COUNT(DISTINCT CASE
        WHEN g.group = 'B' THEN u.id END))
        - (SUM(CASE WHEN g.group = 'A' THEN a.spent ELSE 0 END) / COUNT(DISTINCT CASE
WHEN g.group = 'A' THEN u.id END)) AS conversion_rate_difference
FROM users AS u
JOIN activity AS a
        ON u.id = a.uid
JOIN groups AS g
        ON u.id = g.uid
GROUP BY a.dt
ORDER BY conversion_rate_difference;
```

File for References:

1.) GloBox_spreadsheet
2.) GloBox Tableau

GloBox

Project

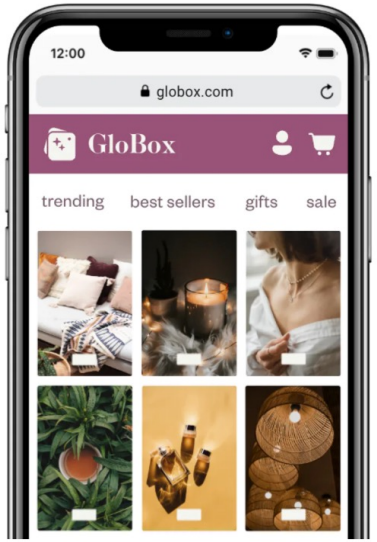
BY – Satakshi Salaria

Experiment

- We tested a new homepage design that we hoped would increase revenue.
- The treatment group sees the banner while the control group does not.
- The experiment ran from 12’June – 8’July 2023
- There were 24600 users in the treatment, 24343 in the control, and 48943 total
- Platform: Mobile website (Android/IOS)
- Countries: AUS, BRA, CAN, DEU, ESP, FRA, GBR, MEX, TUR, USA
- Traffic split: 50/50

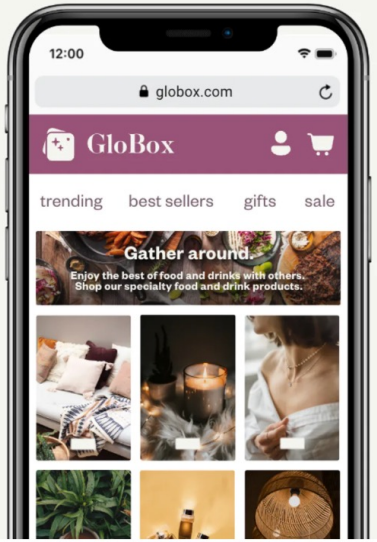
Group A: Control

existing landing page



Group B: Treatment

landing page with food & drink banner



Result

- The conversion rate significantly increased, going from 3.92% in the control group to 4.63% in the treatment group.
- But the average amount spent per user, which was $3.37 in the control group and $3.39 in the treatment group, did not significantly alter.
- We did not see a statistically significant difference between the two groups at the 5% significance level (p=0.94).
- The 95% confidence interval for the difference in revenue per user between the two groups is (-0.44, 0.47). The interval is centered almost around 0.
- Therefore, it does not make sense to launch the treatment because we didn't observe an increase in revenue per user.
- We didn't see enough improvement in our metrics of success, so it's not a good idea to release the banner to all users. The perceived cost of launching the feature is not worth it based on the results of the A/B test.
- If we are very interested in this feature, then we keep running the test for a longer duration to reach the desired sample size of 186k as per Power Analysis.