GloBox - Introducing a new banner A/B test analysis

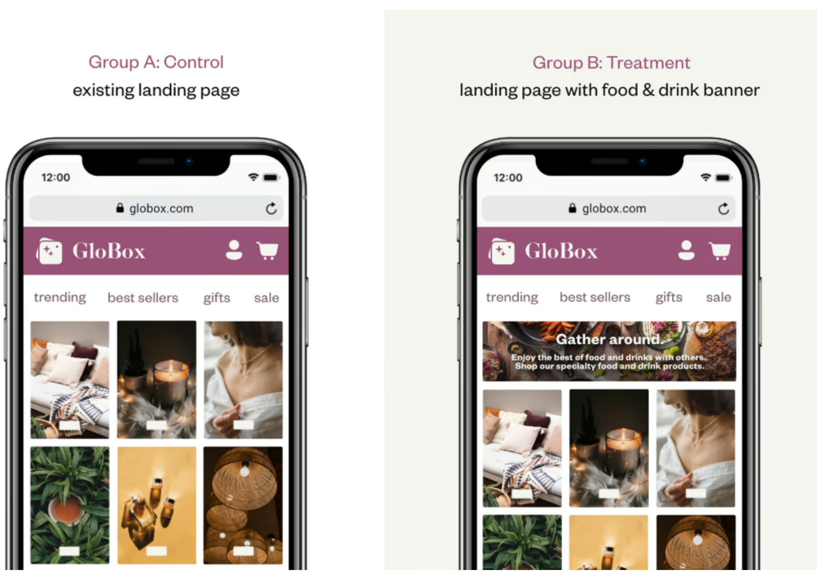
By: Anat Filossof

Summary

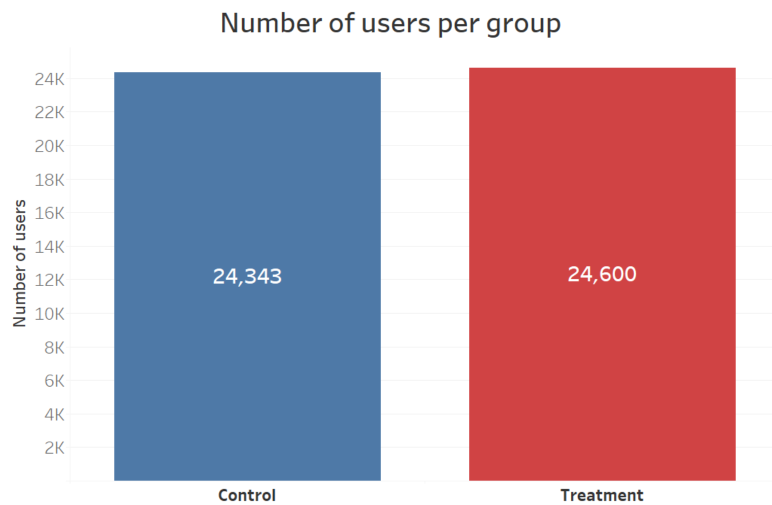
My recommendation is not launching the current banner because we did not observe strong evidence that there was an increase in revenue per user. The analysis revealed an insignificant difference in the average amount spent per user between the tested groups. However, the banner improved the conversion rate metric. A sufficiently strong evidence for a higher conversion rate was observed for the treatment group but it doesn’t lead to more revenue.

Context

The food and drink offerings have grown tremendously in the last few months, so GloBox wants to bring awareness to this product category in order to increase revenue.
We ran an A/B test with highlights of key products in the food and drink category, as a banner at the top of the landing page to see if it would increase revenue. The banner is presented only for the treatment group, as follows:

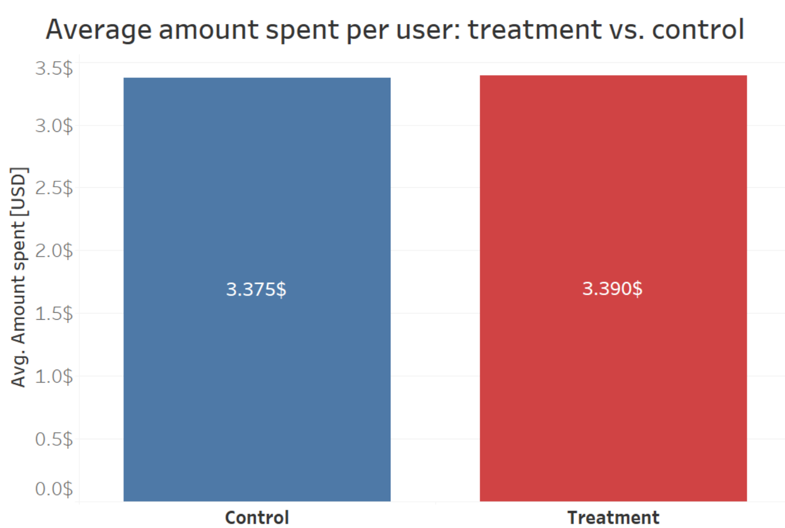


The experiment is only being run on the mobile website. A user visits the GloBox main page and is randomly assigned to either the control or test group.
The experiment ran for 13 days in Q1 2023. There were 24,600 users in the treatment, and 24,343 users in the control (total of 48,943).

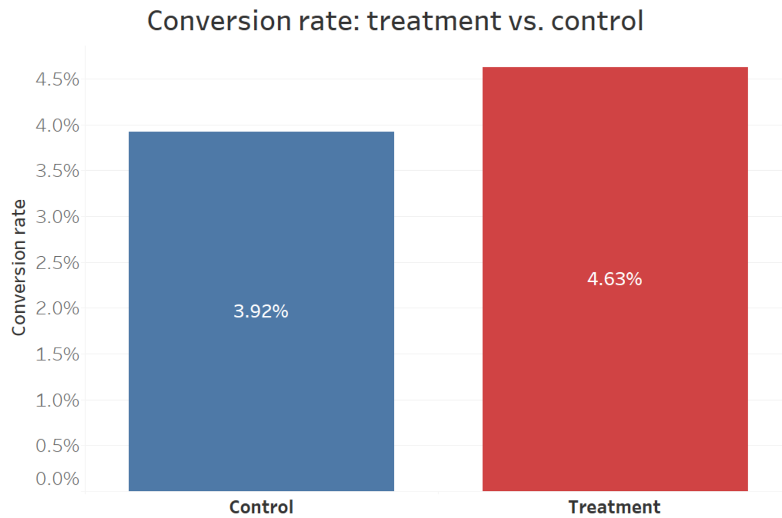


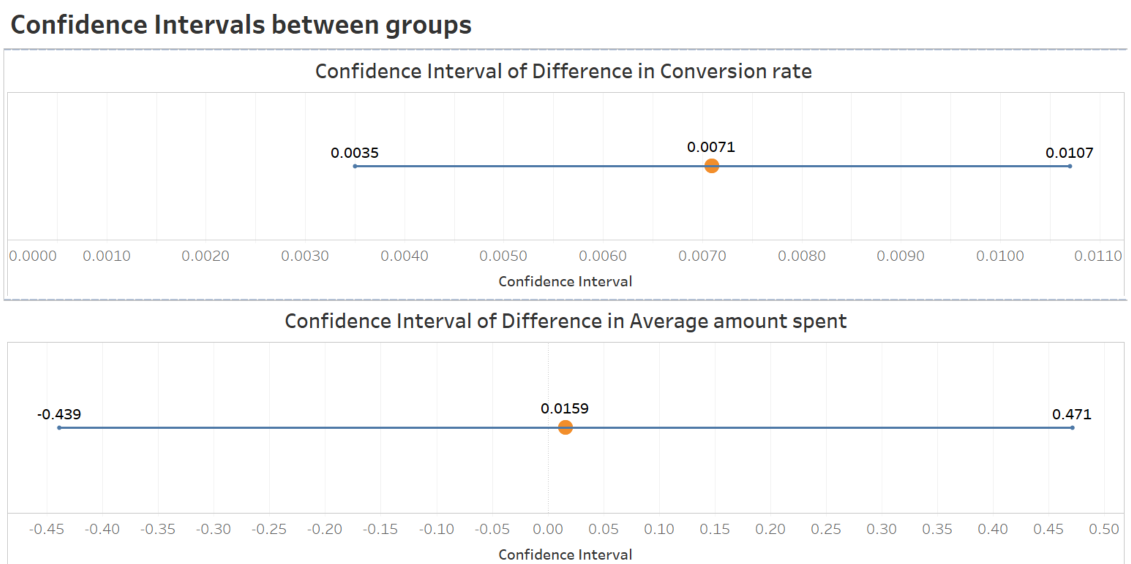
Results

In order to determine whether there was a difference in revenue per user between the two groups, we ran a hypothesis test. We did not see a statistically significant difference between the two groups at the 5% significance level ($p=0.948$). The 95% confidence interval for the difference in revenue per user between the two groups is (-0.439, 0.471). Note that the interval includes 0, meaning that there is strong evidence that there is no significant difference in revenue per user between the two groups (calculated as average amount spent per user).

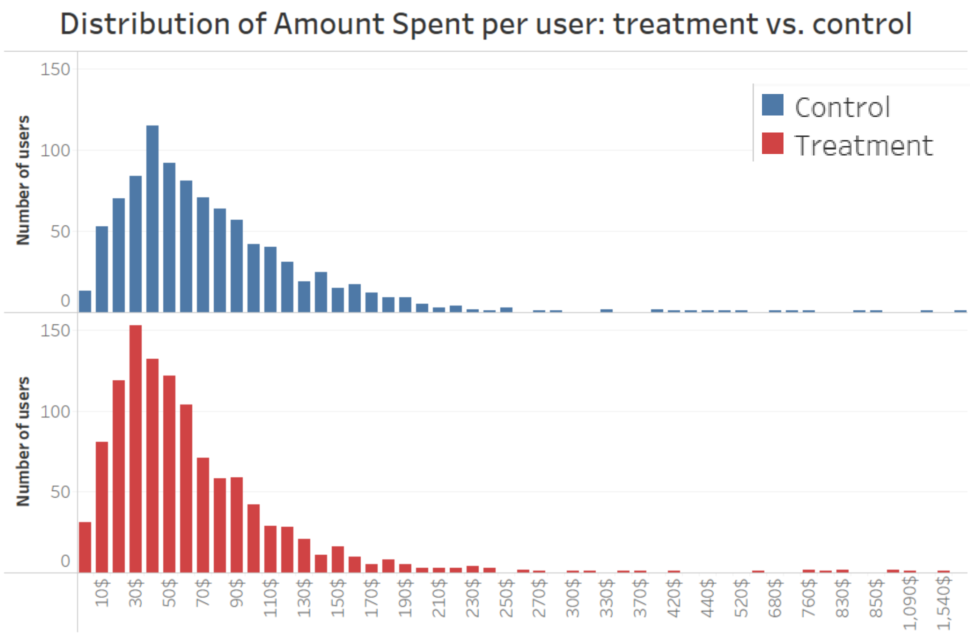


Also, we ran a hypothesis test for the difference in conversion rate between the two groups. In this case, a statistically significant difference between the two groups at the 5% significance level ($p=0.0001$) was observed. The 95% confidence interval for the difference in conversion rate between the two groups is (0.35%, 1.07%). This is strong evidence for a higher conversion rate in the treatment group.

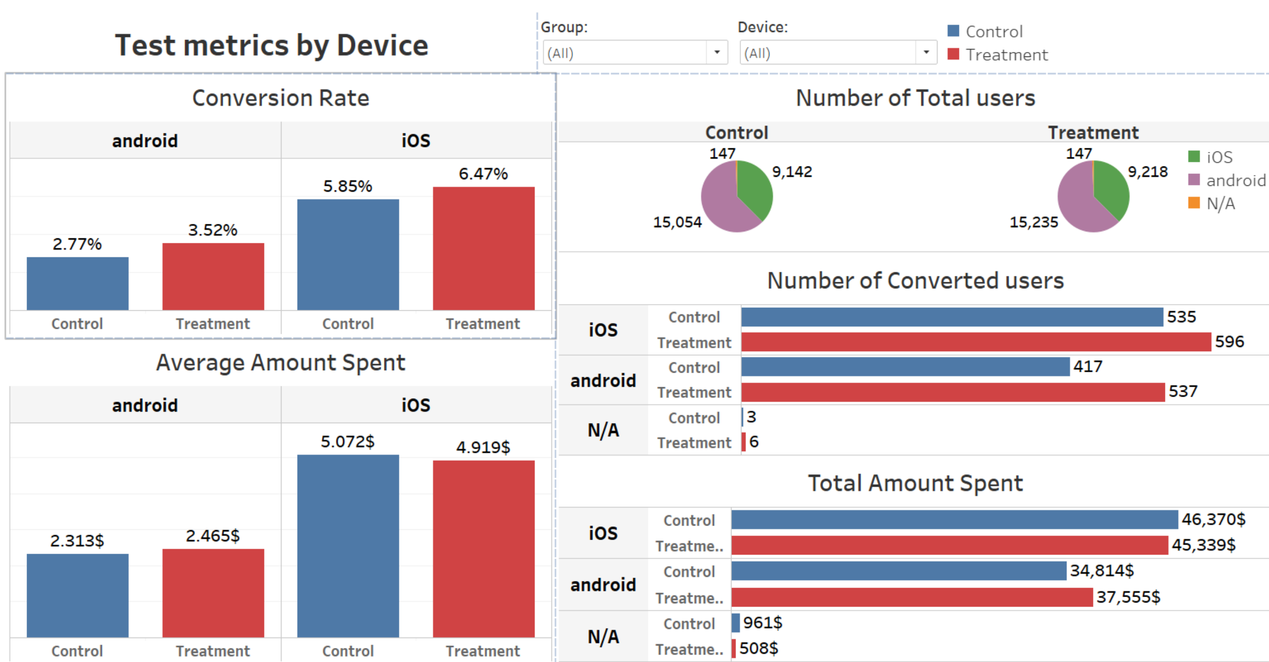




We also analyzed the distribution of total amount spent per user for each group. Only users who made purchases were counted, in order to get a clear distribution pattern. Comparing the two group distributions, the most significant number of purchases, in the control group, occurred within the price range of 40-50$, while in the treatment group it ranged between 30-40$.
On one hand, the conversion rate was enhanced (in the treatment group), but on the other hand, the amount spent per user in the treatment group has decreased - causing no meaningful revenue disparity between the two groups.



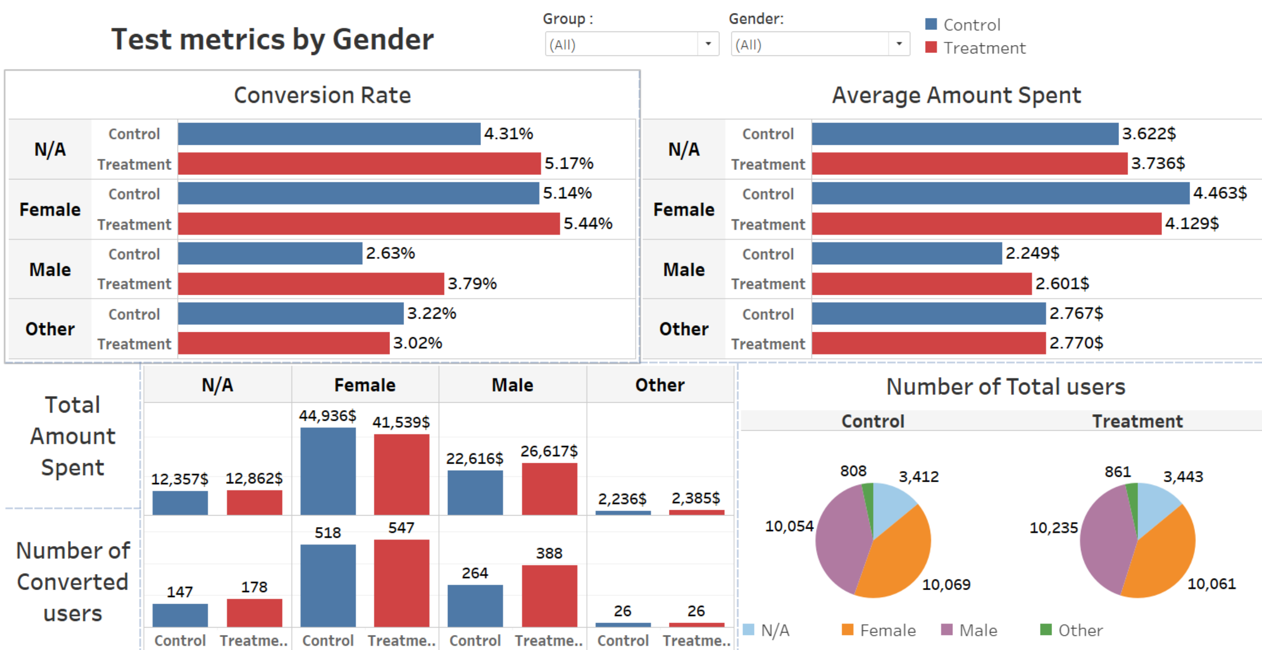
To explore the relationship between the test metrics and user's device, the following visualizations were created:



The distribution of the number of total users by device is presented in the pie charts for each group. The charts exhibit similar distribution between the groups, in which the majority of users are using android devices. In addition to the test metrics (average amount spent per user and

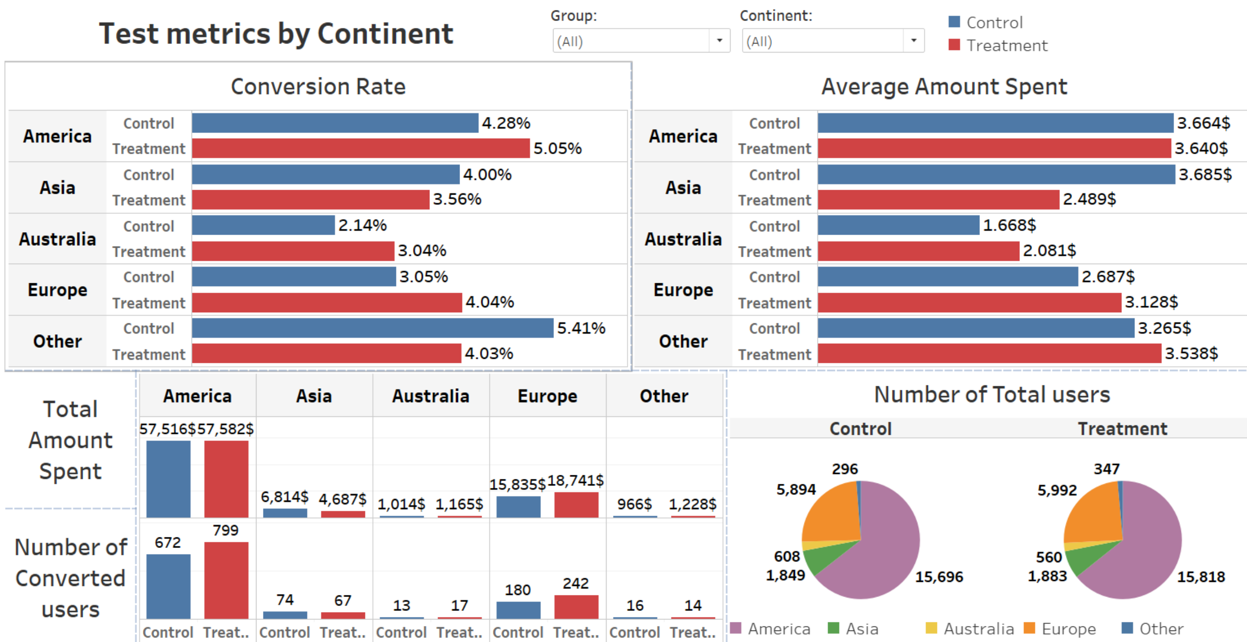
conversion rate), the number of converted users and their total amount spent were also visualized. We can see that we have a small number of missing values for the user's device relative to the total number of users per test group. Thus, a decision to omit this missing values column in the test metric visualizations was taken.
Despite the higher number of Android users, more purchases were made by iOS users, compared to android users in both control and treatment groups. The same pattern was observed in the total amount spent bar chart. As a consequence, significantly higher results for conversion rate and average amount spent were obtained for iOS compared to Android. When comparing the control and treatment group for each device, we may observe the same trend as before (for all users) - higher conversion rate for the treatment group and no evidence for increased average amount spent.

To explore the relationship between the test metrics and user's gender, the following visualizations were created:



As before, the pie charts for the distribution of the number of total users by gender display a similar distribution between the test groups. The largest segments of users are females and males, which are roughly equivalent in size. Additionally, there are segments labeled as 'other' and users with missing gender values. While these latter segments are not insignificant in number, they have also been visualized to provide a comprehensive representation.
We can clearly see from the total amount spent and number of converted users graphs that females tend to purchase more than males. Similarly, The test metrics (conversion rate and average amount spent) indicated the same pattern. All genders, except 'other' gender, demonstrated enhanced conversion rate, and there was no consistent trend for the average amount spent between the groups.

To examine the correlation between the test metrics and the user's country, a decision was made to categorize countries into continents. This approach aimed to identify any broad trends. The visualizations illustrating the relationship between the test metrics and the user's continent are provided below.

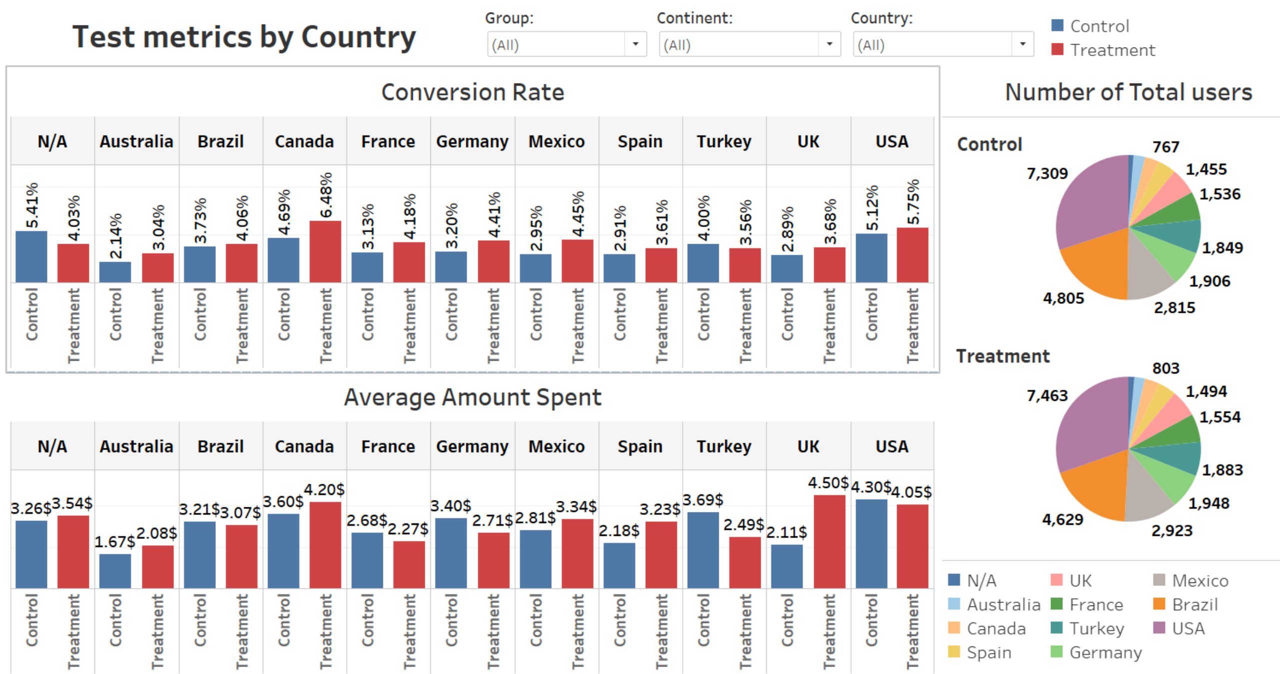


Once again, a consistent distribution of total user numbers across continents can be observed within the test groups. The majority of users originate from the Americas, followed by Europe. Additionally, there exists a segment categorized as 'other,' representing users with missing country information. Despite their relatively smaller proportion, these users have been retained in the visualization for comprehensive representation.

As expected, the continents with the most money spent and the most users who made purchases match the order of user numbers we saw earlier. Users from the Americas made the most purchases on the website, then came Europe, and so on. This pattern holds for the treatment group in the tests, except for the 'other' group, which is missing information.

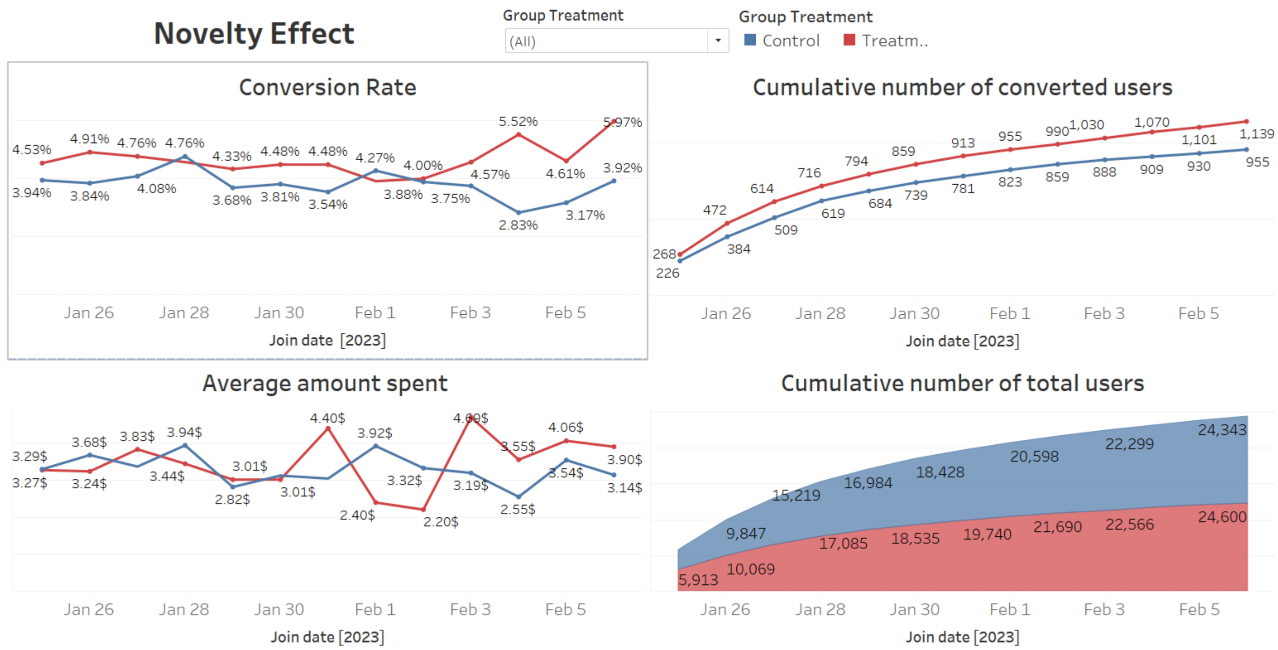
All continents, except from Asia and the ‘other’ label, demonstrated enhanced conversion rate. As for Asia, which includes a relatively significant number of users, a drop in the conversion rate was observed. Moreover, we couldn’t observe any consistent trend for the average amount spent between the two groups in any of the continents.

The visualizations for the relationship between the test metrics and the user's country are presented below:



A similar pattern in the number of total users was noticed across countries within the test groups. The United States stands out as the country with the highest user count, followed by Mexico in third place. These two countries make up the majority of users from the Americas. The United Kingdom, situated in Europe, holds the second spot in terms of user count.
In terms of conversion rates, all countries exhibited an increase in the treatment group compared to the control group, except for Turkey in Asia, and apart from countries with missing data.
However, when considering the average amount spent, no clear connection was observed based on the user's country. These findings align with what we saw in the continent visualizations, maintaining consistency across the results.

We also examined variations in the key metrics between the groups over time to determine whether user behavior differs when exposed to the new treatment (referred to as a novelty effect):



We haven't noticed any novelty effect in the test metrics (conversion rate and average amount spent) visualizations above, that means that the effectiveness of the banner isn't short-lived.

Having obtained strong evidence supporting a higher conversion rate in the treatment group, the practical significance of this metric was also sought to be verified.
To achieve this, the "statsig sample size calculator for conversions" was employed to determine the required sample size for achieving the desired minimum detectable effect and statistical power. For the minimum detectable effect, a 10% relative change was chosen. The calculation yielded a total sample size of 77,000 users for the A/B test.
Subsequently, the cumulative number of users (both total users and converted users) in the experiment at each time point was visualized. This visualization helps to forecast how much longer we have needed to run the experiment in order to reach the desired sample size. However, as evident from the trend in the graph, extending the experiment duration will not contribute to achieving the necessary sample size for the conversion rate metric to hold practical significance.

Recommendation

Derived from the aforementioned findings, proceeding with the launch of the current banner appears unjustified, as we have not observed a notable enhancement in revenue per user. Consequently, I recommend that we do not launch it.
We should only be comfortable launching the banner if it leads to a significant increase in revenue per user. Therefore, I suggest shifting to a higher-priced, profitable category for the banner's content. Subsequently, we should initiate a fresh iteration of the experiment with the updated version. Drawing from our findings, a category that caters to a female audience might prove beneficial, considering our observation of females as significant consumers. Leveraging the promising outcome of the conversion rate in this A/B test, we anticipate an improvement in the revenue per user metric.

Appendix

Link to Tableau dashboards:
[GloBox - A/B Test | Tableau Public]

Link to spreadsheet file:
https://docs.google.com/spreadsheets/d/1wP79xPwp15tn822XR9IwroIKIhG-d1j_79ZKtrtpua4/edit#gid=198983625

SQL code:

```
SELECT Min(join_dt)
FROM   groups

SELECT Max(join_dt)
FROM   groups

SELECT Count(DISTINCT id)
FROM   users

SELECT "group" AS group_treatment,
       Count(DISTINCT uid)
FROM   groups
GROUP  BY 1

WITH activity_users
     AS (SELECT Cast(Count(DISTINCT uid) AS DECIMAL) conversion_users
         FROM   activity),
     all_users
     AS (SELECT Cast(Count(DISTINCT id) AS DECIMAL) total_users
         FROM   users)
SELECT conversion_users / total_users * 100 conversion_rate
FROM   activity_users
```

```sql
       CROSS JOIN all_users


WITH activity_users
     AS (SELECT "group"                                  AS
treatment_group,
                Cast(Count(DISTINCT act.uid) AS DECIMAL)
conversion_users
         FROM   activity act
                JOIN groups gr
                  ON act.uid = gr.uid
         GROUP  BY 1),
     all_group_users
     AS (SELECT "group"                              AS
treatment_group,
                Cast(Count(DISTINCT uid) AS DECIMAL) group_users
         FROM   groups
         GROUP  BY 1)
SELECT act.treatment_group,
       conversion_users / group_users * 100 group_conversion_rate
FROM   activity_users act
       JOIN all_group_users allg
         ON act.treatment_group = allg.treatment_group


WITH user_spent
     AS (SELECT uid,
                Sum(spent) spent_by_user
         FROM   activity
         GROUP  BY 1)
SELECT "group"                                          AS
group_treatment,
       Sum(spent_by_user) / Count(DISTINCT gro.uid) avg_spent_by_user
FROM   groups gro
       LEFT JOIN user_spent spe
              ON gro.uid = spe.uid
GROUP  BY 1
```

```sql
SELECT id,
       country,
       gender,
       gr.device,
       "group"                AS group_treatment,
       CASE
         WHEN spent > 0 THEN true
         ELSE false
       END                    converted,
       COALESCE(Sum(spent), 0) total_spent
FROM   users us
       LEFT JOIN groups gr
              ON us.id = gr.uid
       LEFT JOIN activity act
              ON us.id = act.uid
GROUP  BY 1,
          2,
          3,
          4,
          5,
          6

SELECT id,
       country,
       gender,
       gr.device,
       "group"                AS group_treatment,
       join_dt,
       CASE
         WHEN spent > 0 THEN true
         ELSE false
       END                    converted,
       COALESCE(Sum(spent), 0) total_spent
FROM   users us
       LEFT JOIN groups gr
              ON us.id = gr.uid
       LEFT JOIN activity act
              ON us.id = act.uid
GROUP  BY 1,
          2,
          3,
          4,
          5,
          6,
          7
```