Globox A/B Test Experiment Report

Ethan Golledge

4th of August 2023

Summary

The basis of the A/B test was to see if there was any effect of introducing a food and drinks banner to the home page. This test was run between 2023-01-25 and 2023-02-06, a course of 12 days. When users were introduced to the home page, they were randomly assigned to either the control or treatment group. The control group sample was a size of 24343 users and the treatment group sample was a size of 24600 users, totalling to a sample size of 48943 users. We measured the success of the banner via test metrics including the average amount spent by a user and the average rate of conversion.

We did not see a significant increase in the average amount spent between the groups, a $0.02 increase for those who saw the banner.

Although we did see a significant increase in the average rate of conversion, a 0.71% increase for those who saw the banner.

I recommend we move forward with the use of the banner, but it is imperative we make adjustments in a variety of ways, I will go into more detail later in the report but as a team there are aspects of our strategy and data mining processes that we all need to consider, so that we are able to make more informed decisions.

Context

During a 12-day A/B test, we aimed to assess the influence of introducing a food and drinks banner on the mobile website. Users visiting the website were randomly assigned to either the control group (no banner) or the treatment group (with banner). Our objective in this report is to analyse the results and determine whether the presence of the banner led to a noticeable change in user behaviour while browsing and using the mobile website.

We gathered data on users in a variety of ways, listed in the overview below.

Overview of the tables and their subsequent fields.

- **users**: user demographic information
  - **id**: the user ID
  - **country**: ISO 3166 alpha-3 country code
  - **gender**: the user's gender
- **groups**: user A/B test group assignment
  - **uid**: the user ID
  - **group**: the user’s test group
  - **join_dt**: the date the user joined the test

    - **device**: the device the user visited the page on
- **activity**: user purchase activity, containing 1 row per day that a user made a purchase.
    - **uid**: the user ID
    - **dt**: date of purchase activity
    - **device**: the device type the user purchased on
    - **spent**: the purchase amount in USD

Results

During the period from **January 25th 2023 to February 6th 2023** [1.], a total of **48,934** users interacted with the website, with **24,343** users in the control group and **24,600** users in the treatment group [4.]

Notably, the treatment group showed a statistically significant increase in user conversion [7.], with a rate of **4.63%**, compared to **3.92%** in the control group, resulting in a **0.71%** increase [5.]. I am 95% confident that the true proportion difference of customers in group control and treatment converting due to the banner is between **0.35%** and **1.07%** [8.].

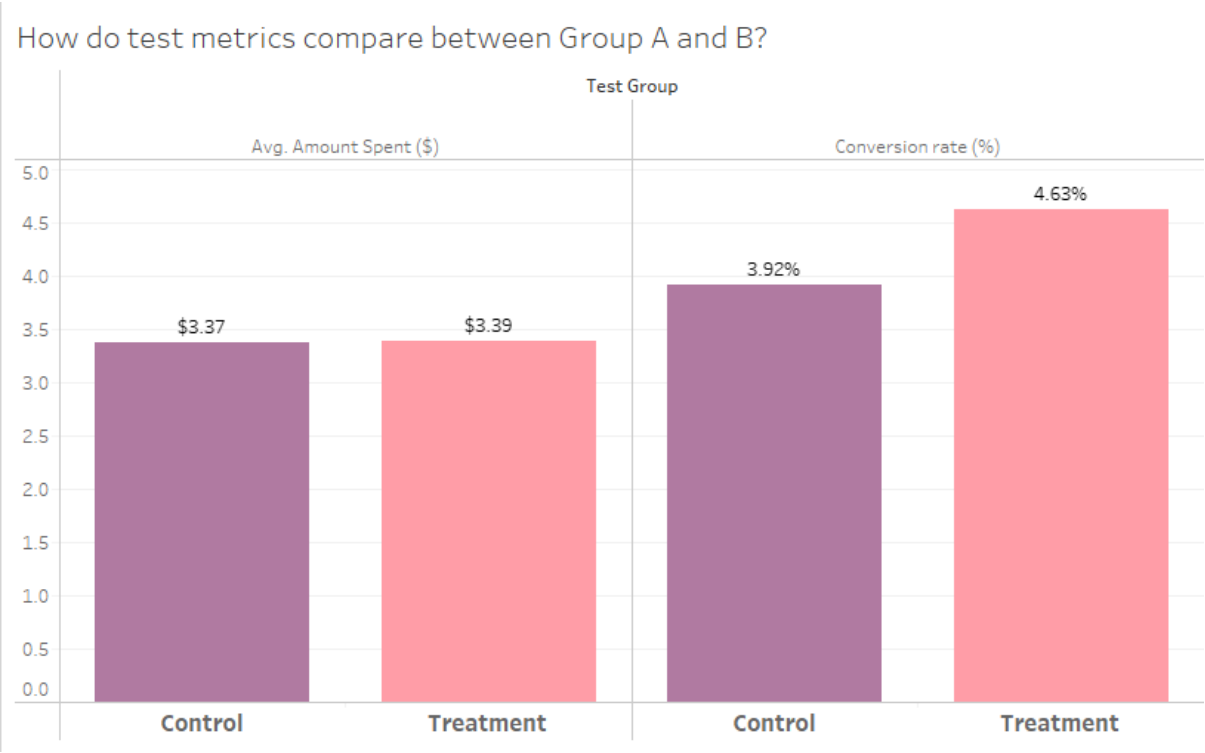
However, the treatment group showed a statistically insignificant increase in the average amount spent [9.], with a marginal increase in the average amount spent by users in the treatment group, with an average of **$3.39**, compared to **$3.37** in the control group, representing a **$0.02** increase [6.]. I am 95% confident that the true mean difference of the average amount spent is between group A and group B is between **$-0.439** and **$0.471** [10.].

To ensure the validity of our findings, we conducted statistical tests to formalize and provide more confidence in the results and their implications.

For detailed insights on how these metrics were gathered using SQL, please refer to the appendix (1-6). For insights on how statistical tests were performed, please refer to the appendix (7-10). Methods of data manipulation and cleaning are present here.

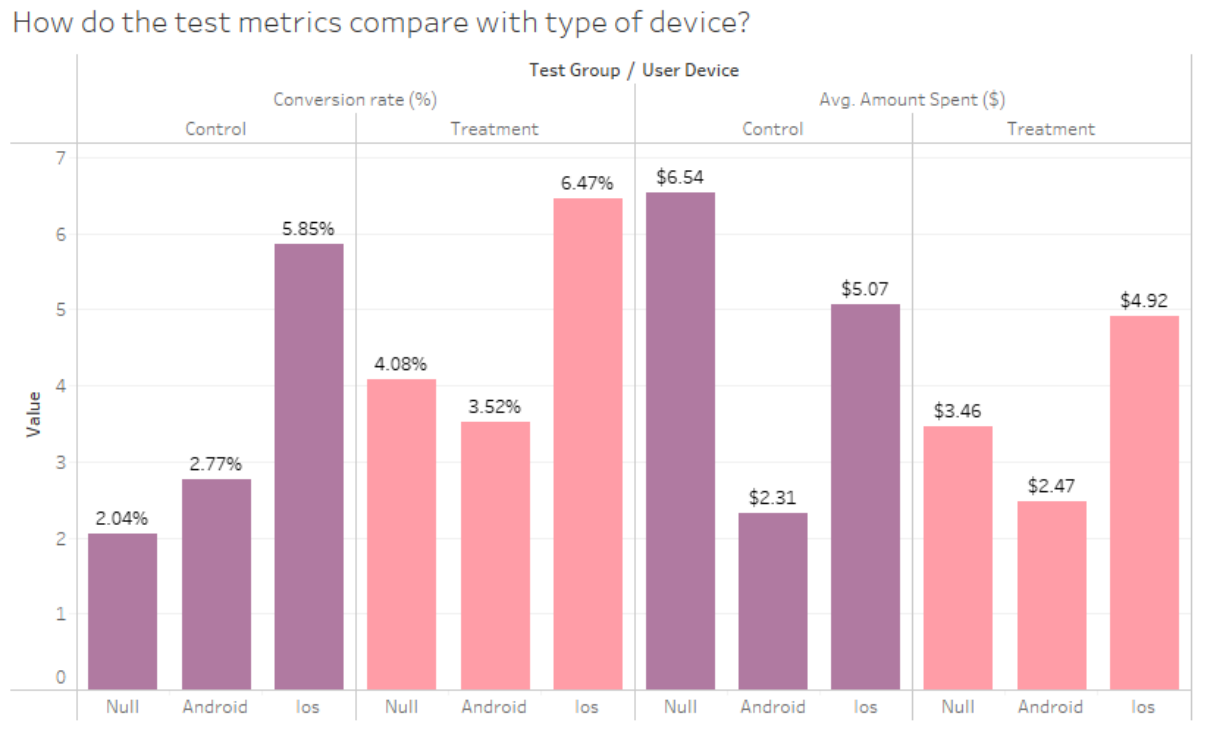
Visualisations using Tableau

Comparing test metrics



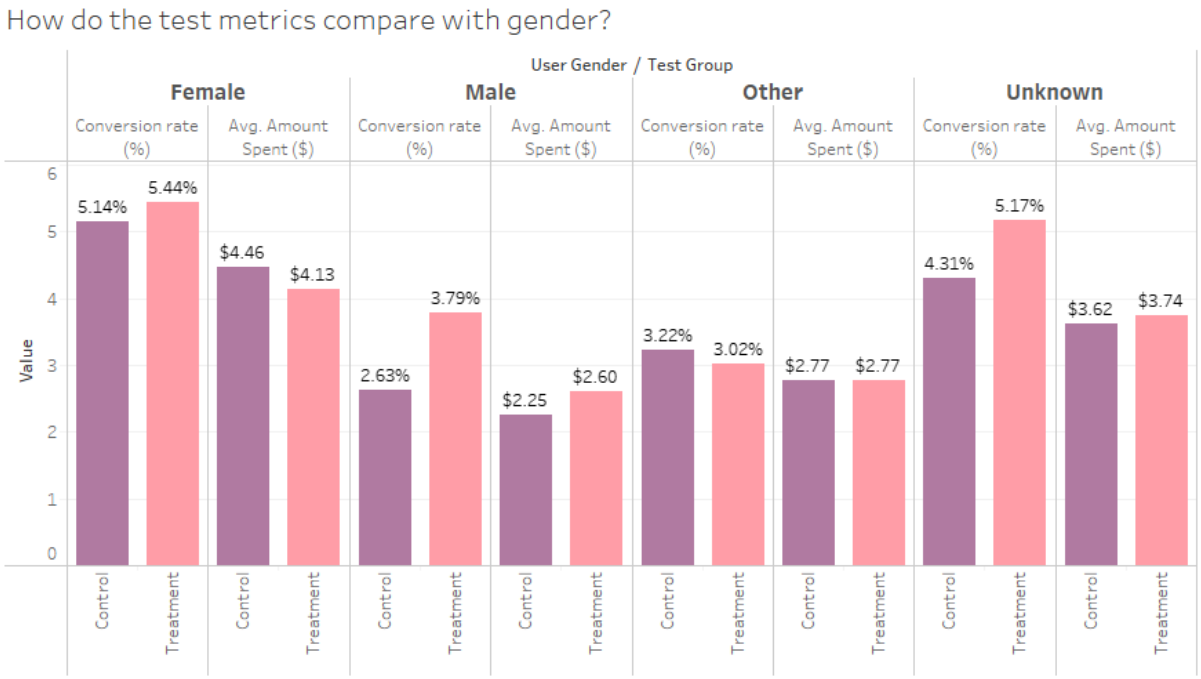
LINK to tableau worksheet test metrics - [final project | Tableau Public]

Comparing test metrics on device used



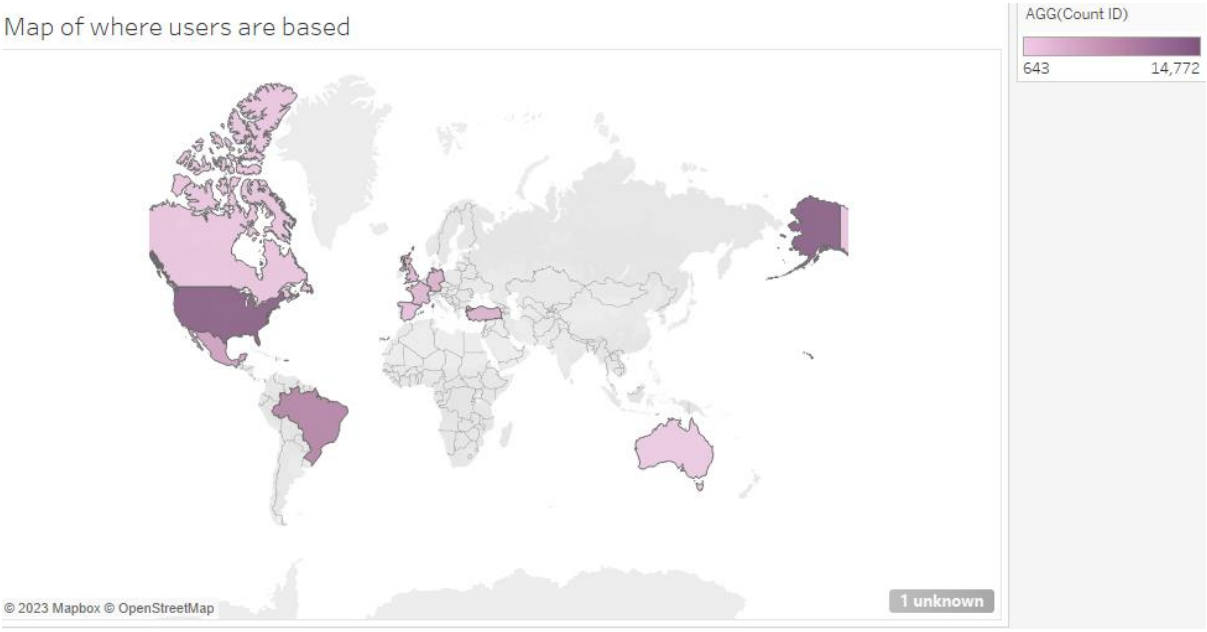
LINK to tableau worksheet device - [final project | Tableau Public]

Comparing test metrics between genders



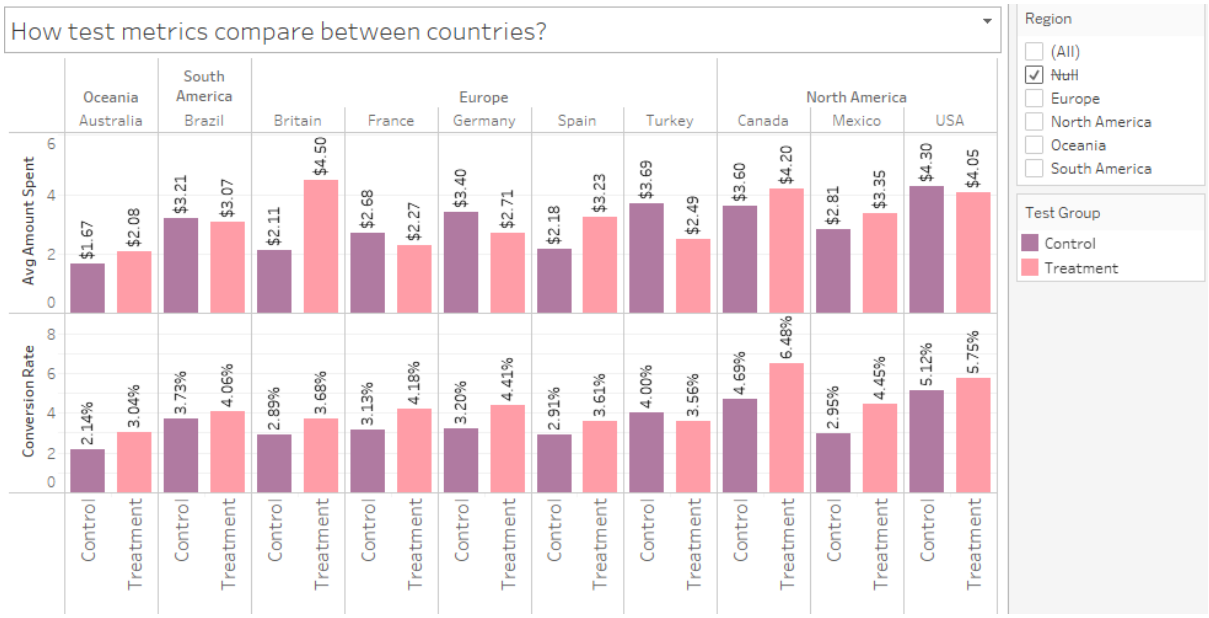
LINK to tableau worksheet gender - [final project | Tableau Public]

Map of user countries



LINK to tableau worksheet on where users are based - [final project | Tableau Public]

Comparing test metrics between regions and countries

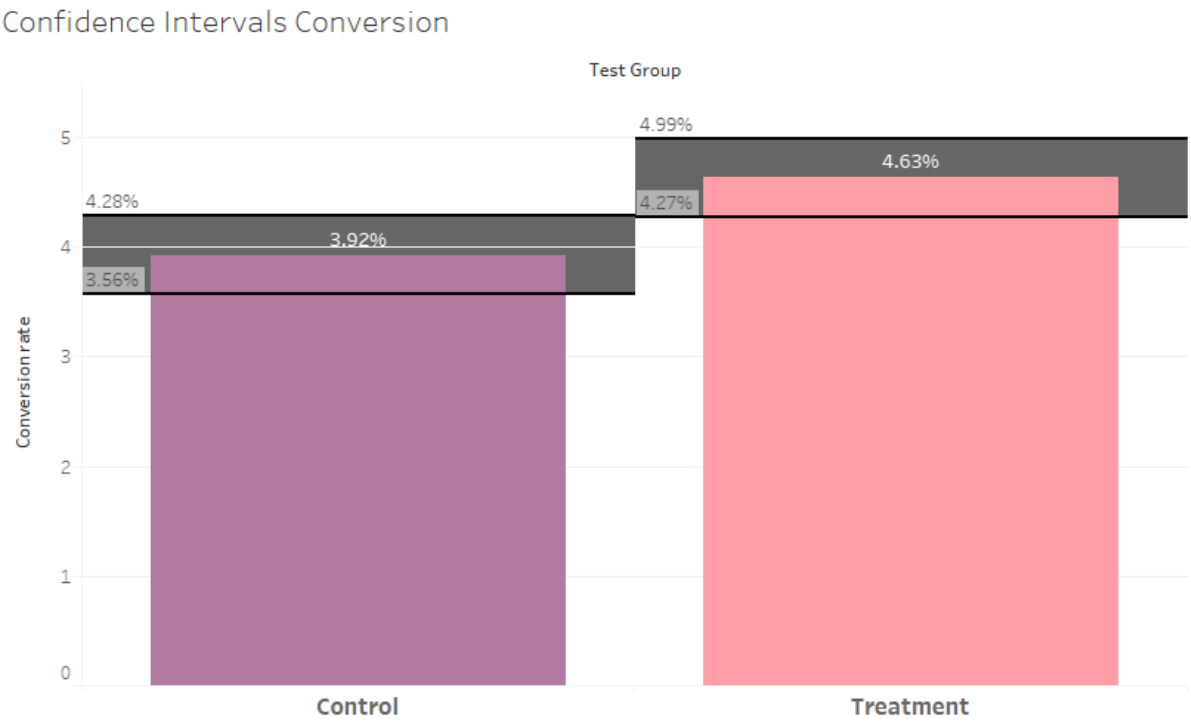


LINK to tableau worksheet on metrics across regions and countries - [final project | Tableau Public]

Confidence level visual

Conversion rates

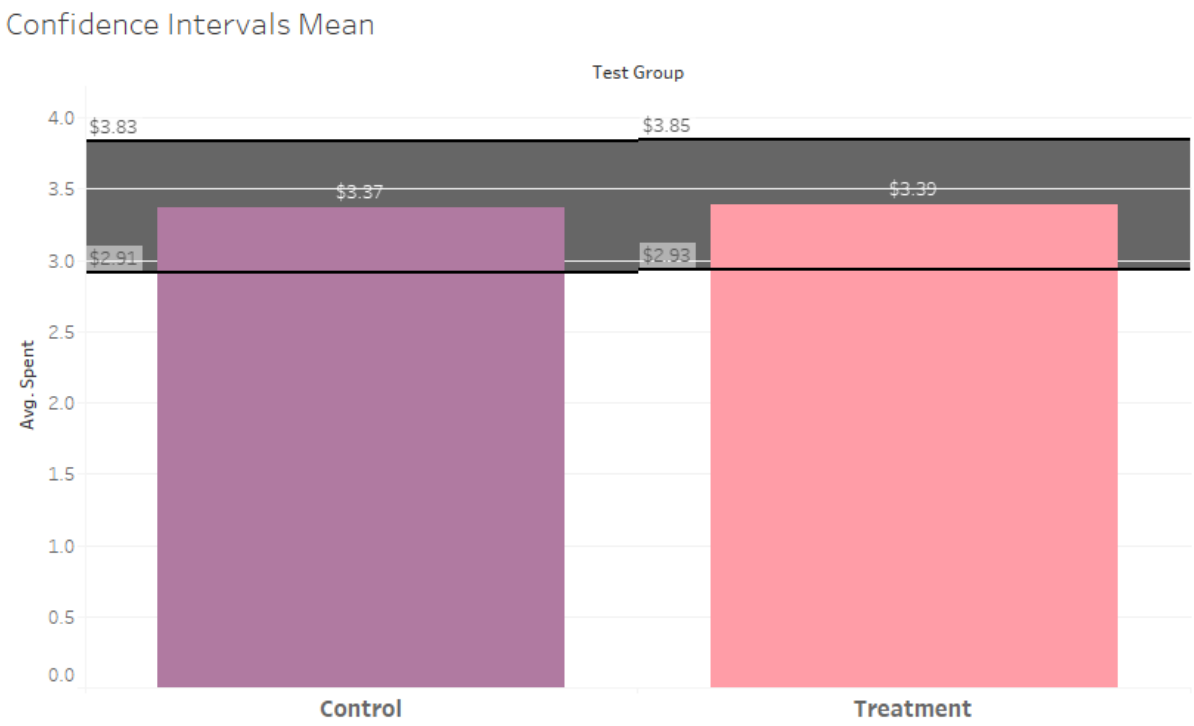
Based on the results, we can confidently state that the true proportion difference of customers converting between the control and treatment group, due to the banner, is estimated to be between **0.35%** and **1.07%** with a 95% confidence level. The margin of error for this estimation is **0.36%**. To visually understand the impact of the banner, we can refer to the bar chart provided in the Tableau link. The chart clearly shows the proportions for each of the groups +/- the margin of error, visualising the confidence intervals, concluding that statistical significance is evident.



LINK to table worksheet confidence intervals conversion - [final project | Tableau Public]

Average amount spent

With a 95% confidence level, we estimate that the true mean difference in the average amount spent between the control and treatment group is between **-$0.439** and **$0.471**. The margin of error for this estimation is **$0.46**. To gain a clear visual understanding of the difference between the groups in terms of average amount spent, you can refer to the bar chart in the Tableau link. The chart clearly shows the average amount spent for each of the groups +/- the margin of error, visualising the confidence intervals, indicating that the results are statistically insignificant.



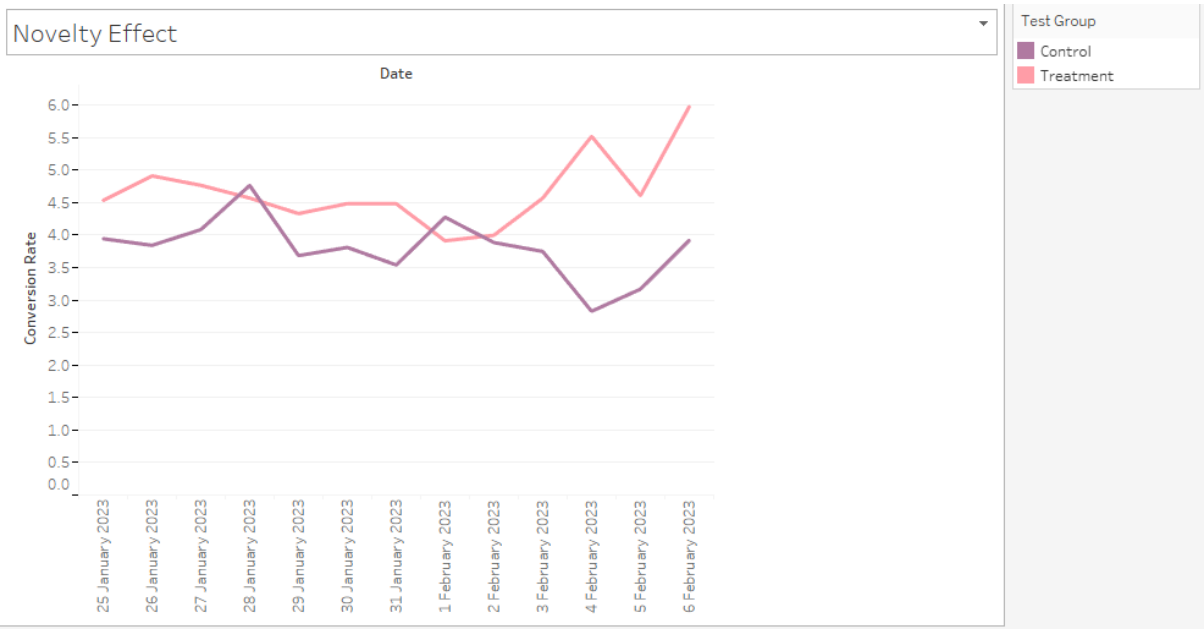
LINK to tableau worksheet confidence levels average amount spent [final project | Tableau Public]

Novelty Effect

When analysing user behaviour, it is important to consider the potential influence of the novelty effect, where customers may actively engage or disengage with a new feature. We need to distinguish short-term fluctuations from lasting changes in test metrics. To assess this effect in our experiment, we focused on conversion rates, observing early trends and comparing them with later ones.

In the initial six days of the experiment, we observed relatively small fluctuations of approximately 0.5% in the treatment group and larger fluctuations of around 1.5% in the control group. However, from January 31st to February 6th, the treatment group showed a positive trend, reaching a peak of around 6%, while the control group exhibited a slighter negative trend fluctuating around 4%. These fluctuations are normal in A/B tests, but the key insight is that the banner had a positive impact on user conversion towards the end of the experiment when comparing both groups.

Although the 12-day duration makes it challenging to conclusively determine the presence of a novelty effect, I am confident in stating that no significant novelty effect was observed, as there was no dramatic increase or decrease in conversion rates during the first six days of the treatment group. The sustained positive trend in the treatment group supports the notion of a meaningful impact of the banner on user conversion.



LINK to tableau worksheet on the novelty effect - [final project | Tableau Public]

Power Analysis

When conducting inferential statistics to draw conclusions about a population, understanding the sample size's relevance is key. A larger sample size is more likely to provide a representative picture of the true population. However, gathering data from the entire population is often impractical due to time and cost constraints. To gain confidence in the generalisability of our findings, I explored the relationship between the sample size and its ability to represent the wider population. To achieve this, I used sample size calculators for both the conversion rate and the average amount spent metrics.

After conducting power analysis it was revealed that we would need a minimum sample size of 182,164 to detect a 5% change in the average amount spent per user. Also, through power analysis it was recommended that we would need a sample size of 60,600 users to detect an effect of a 10% conversion rate. Therefore, while our findings provide valuable insights, we must be cautious in generalising them to the entire population without further data collection and analysis on a larger scale.

Recommendations

Having the right data

Our current analysis focuses on conversion rates and average spending across the groups, providing us with valuable insights. However, to gain a more comprehensive understanding of user behaviour and the impact of the banner, we would benefit from gathering more specific data points in the future.

Examples of data we could collect to enhance our analysis include:

- More specific purchase activity timestamps, including minute and hour information. This would allow us to investigate if customers are checking out more quickly with the food and drinks banner.
- Detailed data on the items customers are buying. Understanding their preferences and purchase patterns would provide valuable insights for decision-making.
- User interactions data, such as click-through rates, funnel conversion, and the specific parts of the website users are engaging with. This information would help us identify user preferences and optimise website design.
- Additional demographic data, such as age brackets, income levels, education, and occupation, would enable us to segment users more effectively and tailor strategies accordingly.
- User feedback would provide direct insights into customer satisfaction, pain points, and suggestions for improvement.

By incorporating these specific data points into our analysis, we can gain a deeper understanding of user behaviour and make more informed decisions to optimize overall business strategy.

Conclusion

The first thing to consider at this point is the sample size, after doing power analysis it was evident that we would need a much larger sample size to make a definite conclusion about the banners effect on the population. The test was simply not run for a long enough time to make hardline conclusions about the banners effect, but regardless we can comment on the success of the sample.

I would consider the A/B test a success, as we saw at least one of the two metrics increase without the other decreasing and we did not see a novelty effect. If we are to continue with the banner, we need to determine what our goals are as a business before deciding whether it is worth incorporating the banner permanently. With our current data we can make decisions based on the conversion rate and average amount spent, but being more abstract we can ask ourselves if changing the brand perception is one of our goals. After all we are known for our fashion and high-end décor items, but our food and drinks offerings have grown in the past few months, incorporating the banner may have a long-term effect on what customers buy from our website and how we are perceived which could impact the rates of conversion and our revenue.

Unfortunately, it is difficult to make conclusions about customer perception and behaviour without the right data, the data I have been given is too vague to understand what is really happening with our customers experience. I have made a list of recommendations with how we could be more effective with our data gathering in the future.

Some recommendations for the team–

- If there are variations between different demographics, perhaps Alejandro will have to create a set of banners that will target the demographics with a more refined approach.
- Leila may have to change management priorities to make sure that we are all working toward the same goals.
- Mei and the marketing team may have to orientate their work around targeting the right audiences in terms of advertising and introduction of the banner depending on how we can boost metrics based on demographics.

- The engineering team may have to update requests on data mining techniques so that we can make sure we have the right data to best optimize our strategy.

The main point that I will make is that we will have to work as a team, our goals will have to be aligned and before taking on this project we must consider all aspects that can be leveraged and make sure we are all on the same page to optimise our results.

How does this affect us as a whole and how will we work together?

There is the potential of the banner being unsuccessful and it should be mentioned so that we can prepare ourselves for the event that we do not like the results, how will we adapt and make sure we keep to our goals and do not go astray with our project to maintain an increase in conversion rates and try to increase overall revenue. An example may be that we have a significant increase in user conversion, but our revenue may decrease as people are spending less on average, how will we battle this? We need to make sure we are set up in the best possible way to make sure we always have the same set of goals and prepare for any possible outcome.

In conclusion, I strongly recommend that we continue iterating and refining our approach, keeping in mind the importance of clear team goals before proceeding with the banner for a longer duration. The experiment has already shown promising results and there are potential ways to increase revenue. By working together effectively, we can optimize our strategies and ensure that both conversion rates and average spending improve.

Despite the challenges and limitations, we faced in this initial test, I am confident that we have laid a solid foundation for success. By addressing the concerns and incorporating more specific data gathering in the future, we can make even more informed decisions and further enhance our outcomes.

Overall, I am optimistic about the future of the food and drinks banner and its impact on our business. Let's seize this opportunity to work collaboratively and capitalize on the positive trends we've observed. With a unified and data-driven approach, we can achieve significant growth and solidify GloBox's position as a leading e-commerce platform in our niche. Together, we can continue to innovate and delight our customers, bringing the world's unique and high-quality products right to their doorstep.

Appendix

Analysis using SQL

1. Determining the start and end date of the experiment

   The **start date** was **2023-01-25** and the **end date** was **2023-02-06**.

Context
In the analysis of the A/B test, it is essential to consider the time duration during which the experiment was conducted.
To identify the start and end dates of the test, I have utilised the 'groups' table, which contains information on whether each user was randomly assigned to group A or B on specific dates. By employing the MAX and MIN aggregate functions, I extracted the earliest start date and the latest end date for the A/B test.

SQL Code

```
SELECT
   MIN(join_dt) AS start_date,
   MAX(join_dt) AS end_date
FROM groups;
```

Start date: **2023-01-25**
End date: **2023-02-06**

Relevance
Determining the start and end date of the A/B test is crucial for analysing and understanding the impact of any changes or interventions during the experiment. These dates help to precisely assess the results and provide context to the stakeholders when interpreting the cost of running the banner and weigh up the potential of a novelty effect.

2. Unique user IDs that appear more than once

Context
This report primarily focuses on the analysis of the activity table, where users appear multiple times.
The activity table specifically includes users who have completed a purchase, resulting in unique a user ID appearing more than once. By executing the provided SQL code, we can generate a list of unique user IDs who have made more than one conversion.

SQL Code

```
SELECT
   uid,
   COUNT(uid) AS count_of_conversions
FROM activity
GROUP BY uid
HAVING COUNT(uid) > 1;
```

Relevance
If we did not factor in user id's appearing multiple times, we would not get an accurate representation of desired test metrics. Specifically, the conversion rate would appear much higher than the true value, misrepresenting findings to stakeholders.

3. Handling NULL values

Context
During data analysis, it is crucial to consider the presence of null values in tables, which may arise from missing or incomplete data. When performing joins between tables, new data combinations can lead to the creation of null values. To handle this, the COALESCE function effectively manages null values without modifying the original data.
This is particularity relevant for the 'spent' field when joining the 'activity' with the 'users' or 'groups' table, where null values may represent users who did not make a conversion. By using the COALESCE function with a default value of 0, I ensured that null values are replaced, allowing analysis of the average amount spent by a user in each group.

SQL Code
```
COALESCE(activity.spent, 0)
```

Relevance
If we did not factor in null values, we would not get an accurate representation of desired test metrics. Specifically, the average amount spent by a user would appear much higher than the true value, misrepresenting findings to stakeholders.

4. Exploring the sample size

The results show a total sample size of **48943** users, **24343** users in the control group (A), and **24600** users in the treatment group (B).

i) Context, total sample size
First, we need to know the size of the total sample for both the control and treatment group. The provided SQL Code calculates the total number of unique users in the 'groups' table by counting the distinct values of the 'uid' column.

SQL Code
```
SELECT DISTINCT
   COUNT(uid) AS count_of_users
FROM groups.
```

Total sample size: **48943** users

ii) Context, sample size between groups
Next, we need to know the sample size for each group individually: the control group (A) and the treatment group (B).
In this SQL code, we group the data in the 'groups' table based on the 'group' column which separates the control and treatment groups. The query then counts the number of the unique users in the group using the 'uid' column.

SQL Code

```
SELECT DISTINCT
   "group",
   COUNT(uid) AS group_count
FROM groups
GROUP BY "group";
```

Control group (A) sample size: **24343** users
Treatment group (B) sample size: **24600** users

Relevance
Understanding the total, control and treatment sample sizes is crucial when evaluating the reliability of the experiment and when conducting subsequent statistical tests. These sample size values will be essential for assessing the experiments validity and statistical significance, helping to guide further interpretations on relevance to the wider population.

5. Exploring rates of conversion

The results show a **4.28%** conversion rate for all users, a **3.92%** conversion rate for the control group and a **4.63%** conversion rate for the treatment group.

i) Context, conversion rate across all users
In the initial step of our analysis, we aim to calculate a test metric - conversion rate. This SQL code results in a metric with insight into the rate of conversion across the entire dataset, including both the control and treatment group.

The SQL code calculates the conversion rate across all users by first counting the total number of distinct ‘uid’ values in the ‘activity’ table (representing users who made a conversion), then dividing it by the total number of ‘id’ values in the ‘users’ table (representing all users in the dataset) to finally be multiplied by 100 to represent a percentage.

SQL Code
```
SELECT
   CONCAT(((SELECT CAST(COUNT(DISTINCT uid) AS float)
   FROM activity)
   /
   (SELECT CAST(COUNT(id) AS float)
   FROM users))*100, '%')
   AS conversion_rate;
```

Overall conversion rate: **4.28%.**

ii) Context, conversion rate between groups
Next, we need to consider the test metric conversion rate for each group individually, the control group (A) and the treatment group (B).
In these SQL queries, we calculate the conversion rate for each group (control and treatment) individually. The conversion rate is obtained by dividing the total number of distinct ‘uid’ values in the ‘activity’ table for each group (representing users who have made

a conversion) by the total number of distinct ‘uid’ values in the ‘groups’ table for each group (representing all users in each group).

SQL Code

```
SELECT
  CONCAT(((SELECT CAST(COUNT(DISTINCT activity.uid) AS float)
  FROM activity
  RIGHT JOIN groups
  ON activity.uid = groups.uid
  WHERE spent IS NOT NULL
  AND "group" = 'A')
  /
  (SELECT CAST(COUNT(DISTINCT groups.uid) AS float)
  FROM activity
  RIGHT JOIN groups
  ON activity.uid = groups.uid
  WHERE "group" = 'A'))*100, '%')
  AS conversion_rate_for_group_a,

  CONCAT(((SELECT CAST(COUNT(DISTINCT activity.uid) AS float)
  FROM activity
  RIGHT JOIN groups
  ON activity.uid = groups.uid
  WHERE spent IS NOT NULL
  AND "group" = 'B')
  /
  (SELECT CAST(COUNT(DISTINCT groups.uid) AS float)
  FROM activity
  RIGHT JOIN groups
  ON activity.uid = groups.uid
  WHERE "group" = 'B'))*100, '%')
  AS conversion_rate_for_group_b;
```

Control group (A) conversion rate: **3.92%**
Treatment group (B) conversion rate: **4.63%**

Relevance

Understanding the conversion rates for the overall sample and each group is crucial for evaluating the experiment’s effectiveness. These calculated conversion rates will help us compare the performance of the control and treatment group, enabling us to make informed decisions.

6. Exploring the average amount spent

The average amount spent by users in control group (A) was **$3.37**, the average amount spent by the treatment group (B) was **$3.39**.

Context
The next test metric we need to calculate is the average amount spent by a user. This SQL code gives us insight spending patterns between the control and treatment groups.
The SQL code calculates the average amount spent by a user for both the control (A) and treatment (B) groups. It achieves this by summing the total spending ('spent' column), in the 'activity' table and then dividing it by the distinct count of 'uid' values in the groups table for each group. By performing a right join between the 'activity' and 'groups' tables, the query combines the relevant information for calculating the average amount spent per user, including those who did not convert.

SQL Code

```
SELECT
   (SELECT SUM(spent)/COUNT(DISTINCT groups.uid)
   FROM activity
   RIGHT JOIN groups
   ON activity.uid = groups.uid
   WHERE "group" = 'A')
   AS average_spent_group_a,
   (SELECT SUM(spent)/COUNT(DISTINCT groups.uid)
   FROM activity
   RIGHT JOIN groups
   ON activity.uid = groups.uid
   WHERE "group" = 'B')
   AS average_spent_group_b
```

Average amount spent by a user in the control (A): **$3.37**
Average amount spent by a user in treatment (B): **$3.39**

Relevance
Understanding the average amount spent per user is crucial in assessing the impact of the A/B test on user spending behaviour. By including all users, irrespective of their conversion status, the metric provides an accurate representation of spending trends, which will help us compare the performance of the control and treatment group, enabling us to make informed decisions.

Gathering relevant data for visualisations

SQL Code

```
WITH user_activity AS (
  SELECT
    users.id AS user_id,
    users.country AS user_country,
    COALESCE(users.gender, 'Unknown') AS user_gender,
    groups."group" AS test_group,
    groups.device AS user_device,
    activity.spent,
    groups.join_dt AS date,
```

```
    CASE WHEN activity.spent > 0 THEN TRUE ELSE FALSE END AS converted
  FROM users
  LEFT JOIN groups ON users.id = groups.uid
  LEFT JOIN activity ON users.id = activity.uid
)

SELECT
  user_id,
  user_country,
  user_gender,
  test_group,
  user_device,
  converted,
  COALESCE(SUM(spent), 0) AS total_spent,
  date
FROM user_activity
GROUP BY user_id, user_country, user_gender, test_group, user_device, converted, date;
```

Statistical Tests

Link to excel sheet, specifics on methodology, key inputs and results

- **Globox Hypothesis Statistical Tests.xlsx**

First before we begin, we need to consider conditions so that we can make assumptions on the following statistical tests. As we are using the same data, both hypothesis tests are to be assumed that the following conditions are met.

1. **Is the sample random?**

Yes, when customers are introduced to the home page on Globox, they are randomly assigned to either group A (control) or Group B (treatment). This random assignment ensures that each customer has an equal chance of being assigned to either group.

2. **Are the observations in the sample independent?**

Yes, each customer's visit to the home page can be considered independent of others. This means that the observations within each group (A and B) and between the groups are independent.

3. **Is the sample size large enough to assume that a normal distribution will occur?**

Yes, the total sample size is 48,943 users, with 24,343 users in Group A and 24,600 users in Group B. The sample sizes for both groups are sufficiently large, meaning a normal distribution can be assumed.

Based on these conditions we can proceed with conducting a hypothesis test to compare the following test metrics between Group A and Group B on the Globox home page.

Conversion rate

Hypothesis test (7.)
Two-sample z-test with pooled proportion with a 5% significance level.

Does introducing the food and drink banner on the home page introduce a difference to the rate of conversion between the control group and the treatment group?

H_o (null): $\hat{p}a = \hat{p}b$
There is no difference in the user conversion rate between the control and treatment group.

H_A (alternative): $\hat{p}a \neq \hat{p}b$
There is a difference in the user conversion rate between the control and treatment group.

p = 0.0001, statistically significant. We REJECT THE NULL HYPOTHESIS that there is no difference in the user conversion rate between the control and treatment as p value < α.

Confidence level (8.)
Two-sample z-interval

I am 95% confident that the true proportion difference of customers in group A and group B converting due to the banner is between 0.35% and 1.07%.

Average amount spent

Hypothesis test (9.)
Two-sample t-test with un-pooled variance (simplified Walsh t-test).

Does introducing the food and drink banner change on the home page introduce a difference on the average amount spent between the treatment and control group?

H_o (null hypothesis):
$\bar{x}a - \bar{x}b = \bar{x}_o$
There is no difference in the average amount spent between the control and treatment groups.

H_A (alternative hypothesis):
$\bar{x}a - \bar{x}b \neq \bar{x}_o$
There is a difference in the average amount spent between the control and treatment groups.

p = 0.944, statistically insignificant. We FAIL TO REJECT THE NULL HYPOTHESIS that there is no difference in the mean amount spent between the control group and the treatment group as p value > α.

Confidence level (10.)
Two-sample t-interval with un-pooled variance.

I am 95% confident that the true mean difference of the average amount spent is between group A and group B is between $-0.439 and $0.471.

False Positive and False Negative during hypothesis testing

As we pre-determined our significance level before conducting any statistical tests, there was no bias in the analysis. However, it is essential to consider the possibility of false positives or false negatives in hypothesis testing. These errors can occur due to an incorrectly set confidence level, which determines the significance of the results by comparing the alpha value.

In our analysis, we obtained a p-value of 0.01% for the proportions test on the conversion rate and a p-value of 94.39% for the mean difference. The significant difference between the p-values and the chosen 5% alpha significance level provides confidence that there were no false positives or false negatives in our results.

By adhering to the pre-defined significance level and observing substantial differences between the p-values and the alpha value, we can be confident that our conclusions from the statistical tests are reliable and not subject to false results. This demonstrates the importance of carefully selecting and understanding the significance level in hypothesis testing to draw accurate and meaningful conclusions.