Driving User Engagement and Revenue: Assessing the Influence of a Banner on Average Spend and Conversion Rate at Globox

Lauren Kemmy-Adani - 22/07/2023

Summary:

An A/B test was carried out on a sample of Globox customers, during which participants in the treatment group were introduced to a banner advertising the expanding range of food and drinks on offer. These results were collected, and statistical analysis was conducted to determine whether or not there was a difference between average spend and conversion rate between the control and treatment groups.The results of the analysis suggest that while there was a significant difference between conversion rates, there was no significant difference between the average spend of the two groups. The final recommendation is to launch the banner to all users, regardless of the lack of difference in average spend. This is because the banner is designed to draw attention to the food and drink offerings on the website, and based on the knowledge that the website primarily sells boutique fashion items and high end decor, it is reasonable to assume that these big ticket items are more expensive than the food and drink advertised in the banner. Therefore it would be expected to observe higher conversion rates in the treatment group, without seeing an elevated average spend, if the banner was drawing attention to these items. The results of the analysis reflect this.

Context:

The motivation for the project was to test whether a banner advertising the key food and drinks products at the top of the landing page for the Globox website, would increase revenue, and draw attention to these products to encourage customers to purchase them. To do this an A/B test was conducted between the 25th of January 2023 and the 6th of February 2023, in which participants were placed into either the control or treatment group. The control group experienced the website in its usual form, and their interactions were collected, alternatively the treatment group were introduced to the new banner, and their interactions were also collected, and then compared to the control group.

Data Gathering and Analytical Approach:

To conduct the analysis an SQL code was written (see appendix 1), to collate the data. There were a total of 48,943 participants, the control group contained 24,343 participants and the treatment group contained 24,600 participants. The data collected included total spend, user id, gender, device type, and country information. The next step was to perform the actual analysis, so the data was imported into google sheets (see appendix 2), and a z-test with a threshold of 5% was performed to determine whether or not there was a significant difference in conversion rates between groups, and a t-test with a threshold of 5% was performed to determine whether or not there was a significant difference in average spend between groups. Calculations to determine the confidence intervals for both average spend and conversion rate were conducted, and the data was imported to tableau for further analysis through a number of visualisations (see appendix 3).

In order to gain further insights into the results of the test a second SQL query was written, this query contained the data relating to the users join date, and the the date the user made a purchase (see appendix 4). This was then imported into tableau for additional visualisations (see appendix 3).

Finally, calculators were used to determine the power analysis for both the conversion rate and the average spend. This was to better understand how the sample sizes would have affected the results (see appendix 5).

Results:

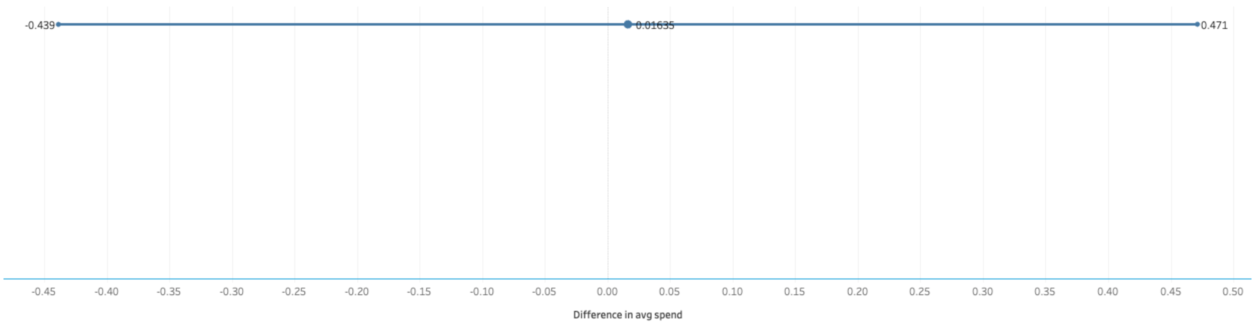
Average Spend:

A paired samples t-test was conducted to examine the impact of the banner on the average spend of customers between groups. The average spend of users in Group A (M = 3.37, SD = 25.94) was compared to Group B (M = 3.39, SD = 25.41). The results suggested no statistically significant difference between the two groups, t(48942) = -0.070, p = 0.944, indicating that the mean difference in average spend was not significant (p > 0.05, two-tailed). The mean difference in average spend was 0.02, with a confidence interval ranging from -0.439 to 0.471. Based on the analysis we fail to reject the null hypothesis. In summary, there is no evidence of a difference in average spend between groups.

Null Hypothesis (H0): The average spend does not differ significantly between the two groups.
Alternative Hypothesis (H1): There is a significant difference in the average spend between the two groups.

Group	Sample Size	Mean Spend	Standard Deviation	Difference in Mean Spend	t-value	p-value	95% Confidence Interval
Control	24343	3.37	25.94	0.020	-0.070	0.944	-0.439 - 0.471
Treatment	24600	3.39	25.41	**Standard Error**	0.232	**Degrees of freedom**	48942

The confidence interval spans from -0.439 to 0.471. This range represents the potential difference in average spend between the two groups. In simpler terms, the confidence interval suggests that the true difference in average spend between the two groups is likely to fall somewhere between a little less than $0.44 lower for one group and a little more than $0.47 higher for the other group when compared to the point estimate difference of $0.02. This range allows for fluctuations in average spending differences, but we can be 95% confident that the true difference lies within this margin of uncertainty.You can visualise the confidence interval for average spend in the graphic below:



Conversion Rate:

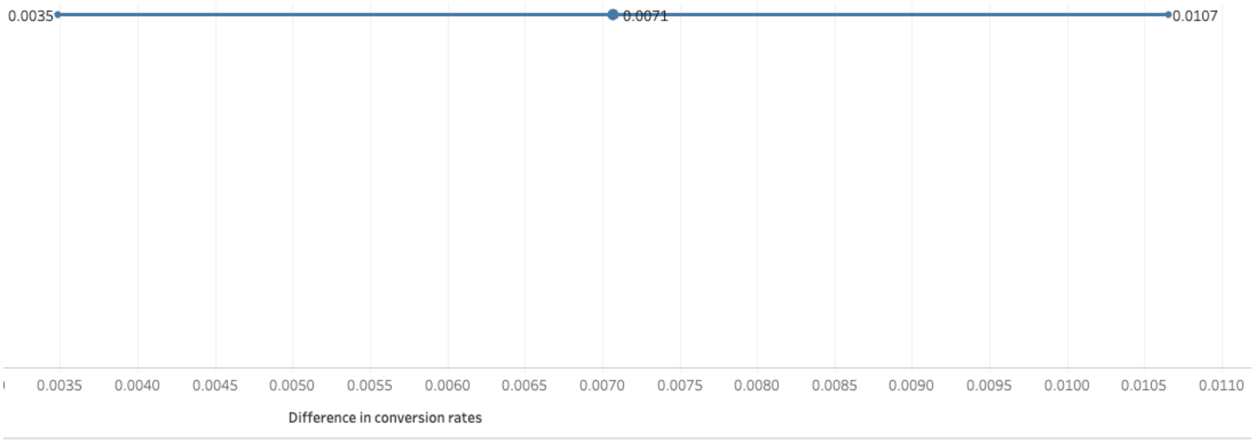
A paired sample z-test was conducted to examine the difference in conversion rates between Group A and Group B. The conversion rate for Group A was 0.0392, while the conversion rate for Group B was 0.0463. The combined conversion rate across both groups was 0.0428. The calculated difference in conversion rates was 0.0071, with a standard deviation of 0.0018. The z-score was computed as 3.864, which exceeded the critical value at a significance level of 0.05. The corresponding p-value was determined to be 0.0001. These results suggest a statistically significant difference in conversion rates between the two groups.

The confidence interval for the difference in conversion rates ranged from 0.0035 to 0.0107. As the p-value is less than the significance level, we reject the null hypothesis, suggesting that there is a significant difference in the conversion rates of Group A and Group B. In summary, there is evidence to support the conclusion that the conversion rates differ significantly between the two groups, with Group B demonstrating a higher conversion rate than Group A.

Null Hypothesis (H0): There is no significant difference in the conversion rate between the two groups.
Alternative Hypothesis (H1): There is a significant difference in the conversion rate between the two groups.

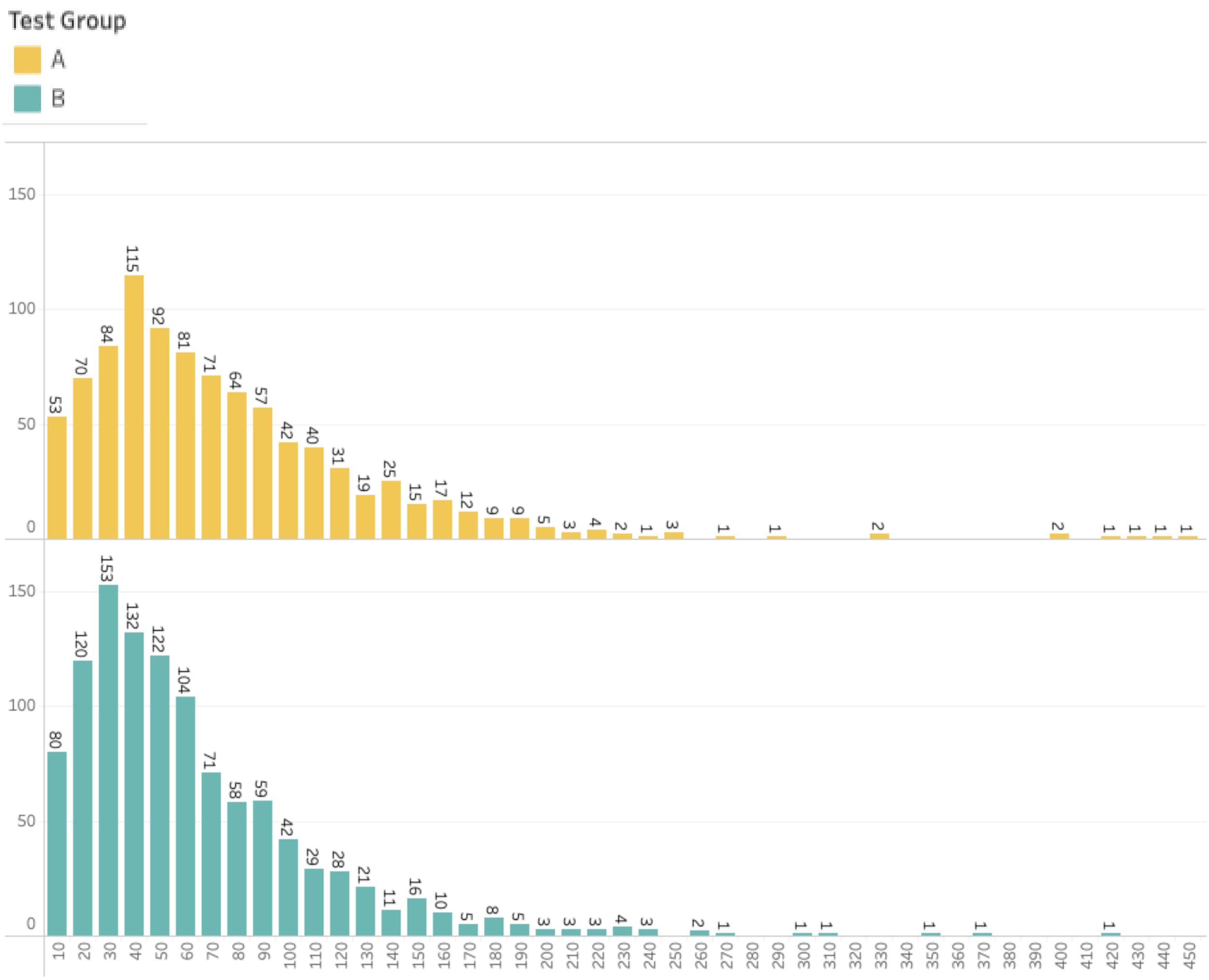
Group	Sample Size	Conversion Rate	Number of Conversions	Conversion Rate Difference	z-value	p-value	95% Confidence Interval
Control	24343	0.0392	955	0.0071	3.864	0.0001	0.0035 - 0.0107
Treatment	24600	0.0463	1139	**Combined Conversion Rate**	0.0428	**Standard Deviation**	0.0018

The confidence interval spans from 0.0035 to 0.0107. This range represents the potential difference in conversion rates between the two groups. In simpler terms, the confidence interval suggests that the true difference in conversion rates between the two groups is likely to fall somewhere between 0.35% and 1.07% when compared to the point estimate difference of 0.71%. This range allows for fluctuations in conversion rate differences, but we can be 95% confident that the true difference lies within this margin of uncertainty.You can visualise the confidence interval for conversion rate in the graphic below:



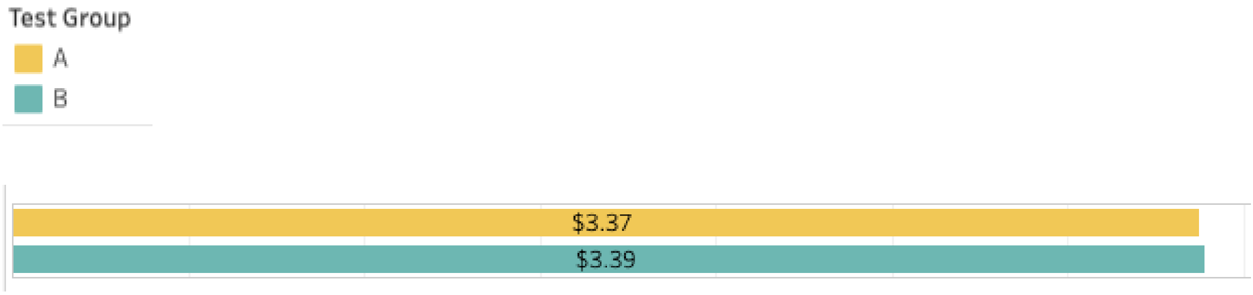
Visualisations:

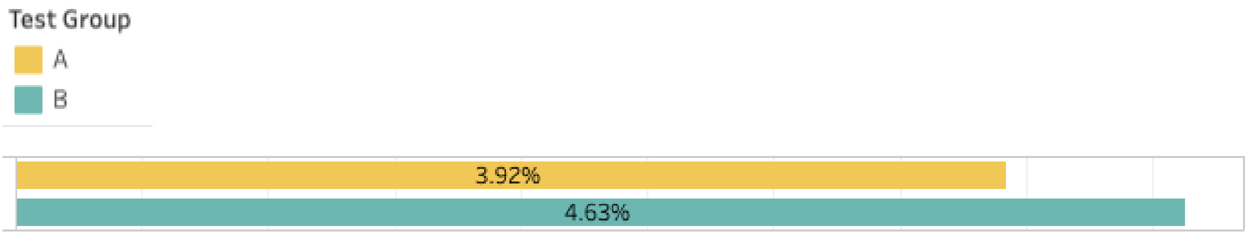
Spend distribution between groups:



Here we can see a higher number of users spending in smaller amounts in the treatment group when compared to the control group.

The graphics below display the average spend and conversion rates for each group:

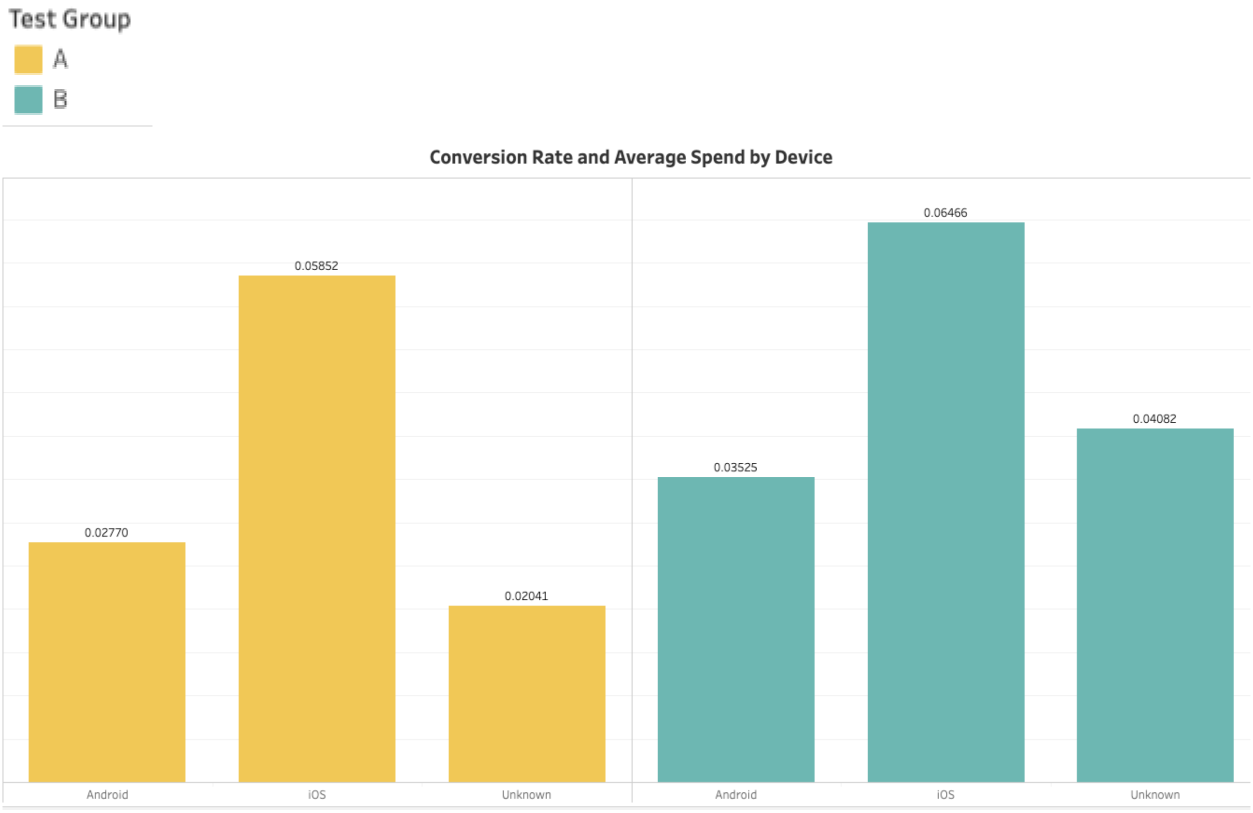




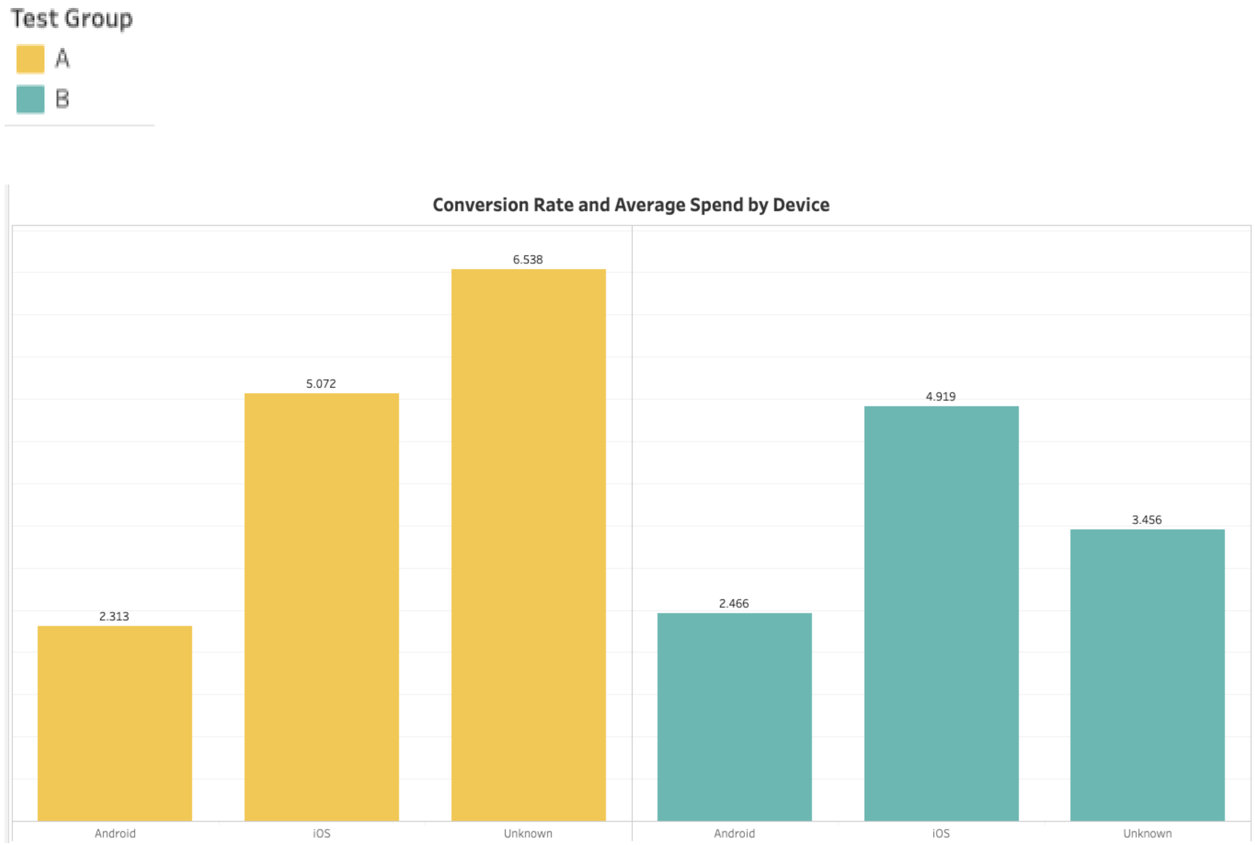
We can see that while average spend is very similar between groups, there is a significant difference in conversion rate. This supports the conclusion that more users are converting in the treatment group, but spending in smaller amounts, which suggests that the users in the treatment group are being drawn to those food and drink items, and are purchasing them as a result.

Device:

Conversion rate:



Average spend:



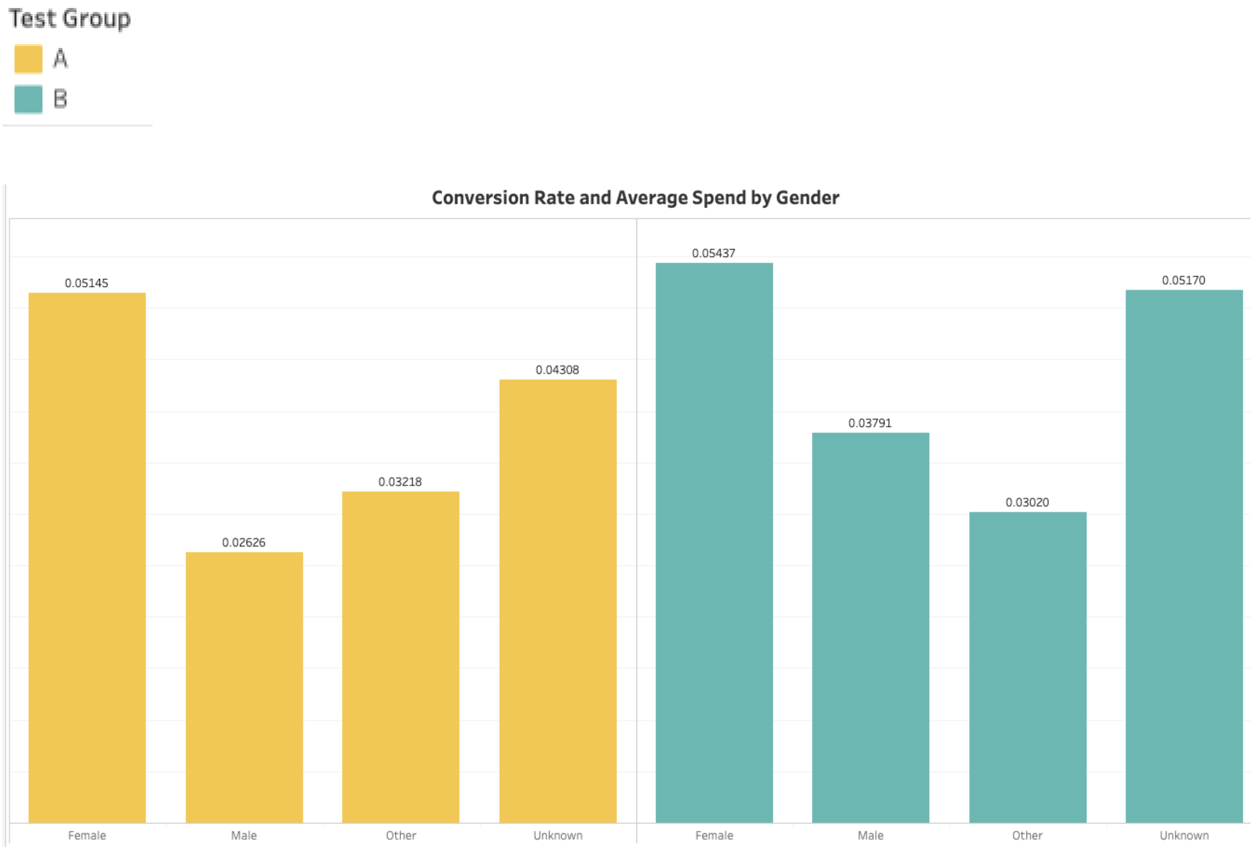
Here we can see that between iOS and Android devices, users on an iOS device had higher conversion rates and average spend overall, but that while the conversion rates for users on an iOS device were higher in the treatment group (6.47%) than the control group (5.85%), the average spend for users on an iOS device was actually higher in the control group ($5.07) than in the treatment group ($4.92).

The information on device type was not available for all users, this could be because the user was on a device other than an iOS or Android, however the decision to keep the Unknown device values in the visualisation was made, because they accounted for a large percentage of both average spend and conversion rate. For the control group the Unknown device type had a conversion rate of 2.04%, but an average spend of $6.54, which was the highest average spend for all devices across both groups.

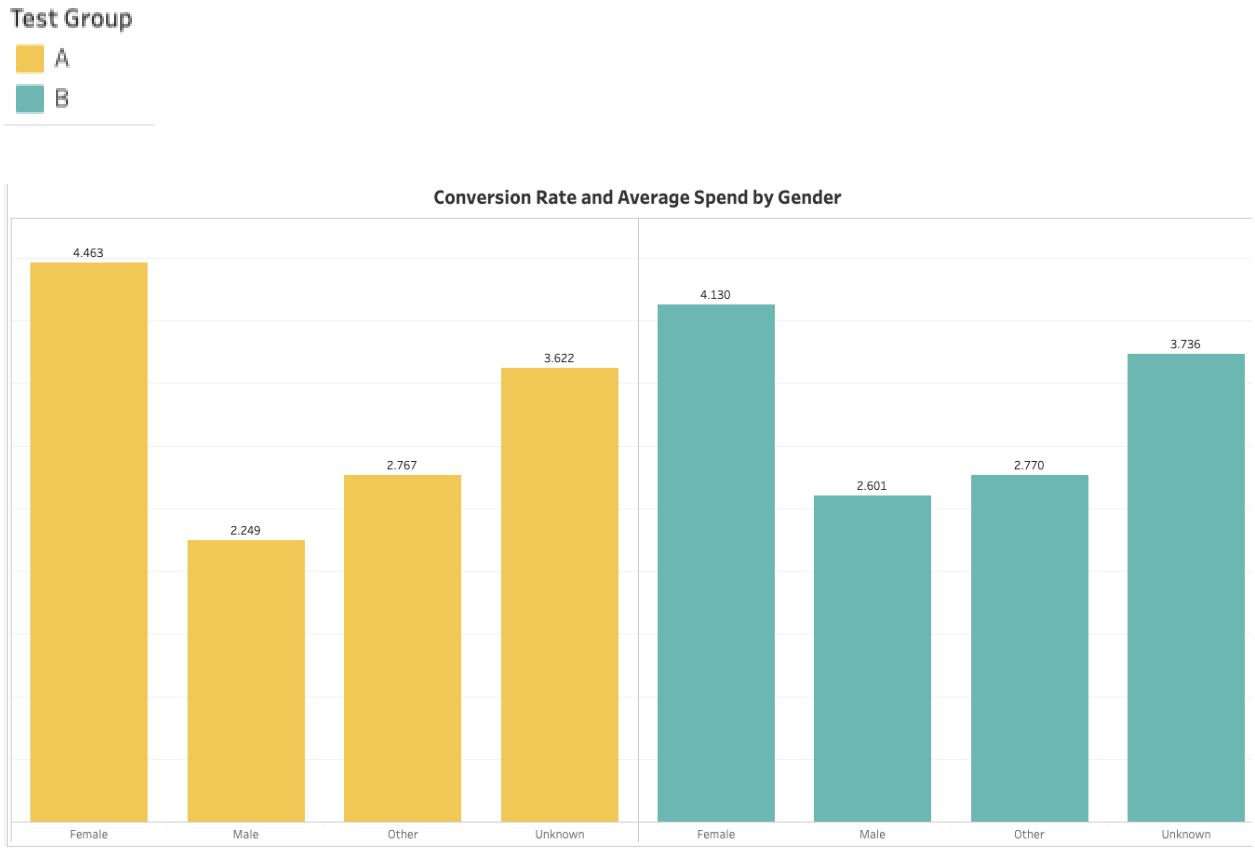
For the treatment group the Unknown device type had the conversion rate of 4.08% and an average spend of $3.46. Although including the Unknown device types doesn't provide much insight into the data obtained via the A/B test, it does suggest that it might be useful to obtain this information. This missing information would be useful to have because it would allow for more in depth analysis that could lead to an increase in revenue. Particularly if it demonstrates an issue with the way the banner displays on particular devices, or highlights a better, more interactive experience that could be replicated across all devices.

Gender:

Conversion Rate:



Average Spend:



Here we can see that The Female gender group had higher conversion rates and average spend overall, but average spend for the Female gender group was actually higher in the control group ($4.46) compared to the treatment group ($4.13), while the conversion rate for the Female gender group is higher in the treatment group (5.44%) than the control group (5.14%). The Male gender group had higher average spend in the treatment group ($2.60), when compared to the control group ($2.25), similarly the conversion rates for the male gender group were higher in the treatment group (3.79%) than the control group (2.63%). Some users indicated their gender as ‘Other’, and the analysis found that the Other gender group had the same average spend value in both the control and treatment groups ($2.77), although the conversion rate was slightly higher in the control group (3.22%), when compared to the treatment group (3.02%).

Not all users opted to include their gender information, so there are a number of Unknown values. The decision was made to include the Unknown values in the visualisation, because they do make up a large proportion of both the conversion rate and average spend, in both the control and treatment groups. For example, the Unknown gender in the treatment group has the second highest conversion rate overall

(5.17%), and the third highest average spend ($3.74). This isn’t going to add further insights into the current analysis, but it does suggest that it may be useful to see if users can be encouraged to include this information in future tests, because it could provide key insights into the different demographics. The most interesting takeaway from the analysis on gender is that while conversion rates for both the male and female gender groups are increasing in the treatment group when compared to the control group, there is a significant increase of 44% for the male gender group when compared to just a 5% increase from the female gender group. This indicates that the banner is encouraging more men to purchase from the website than normal, so it could be argued that the banner should be launched to all users if Globox wanted to attract more male customers.

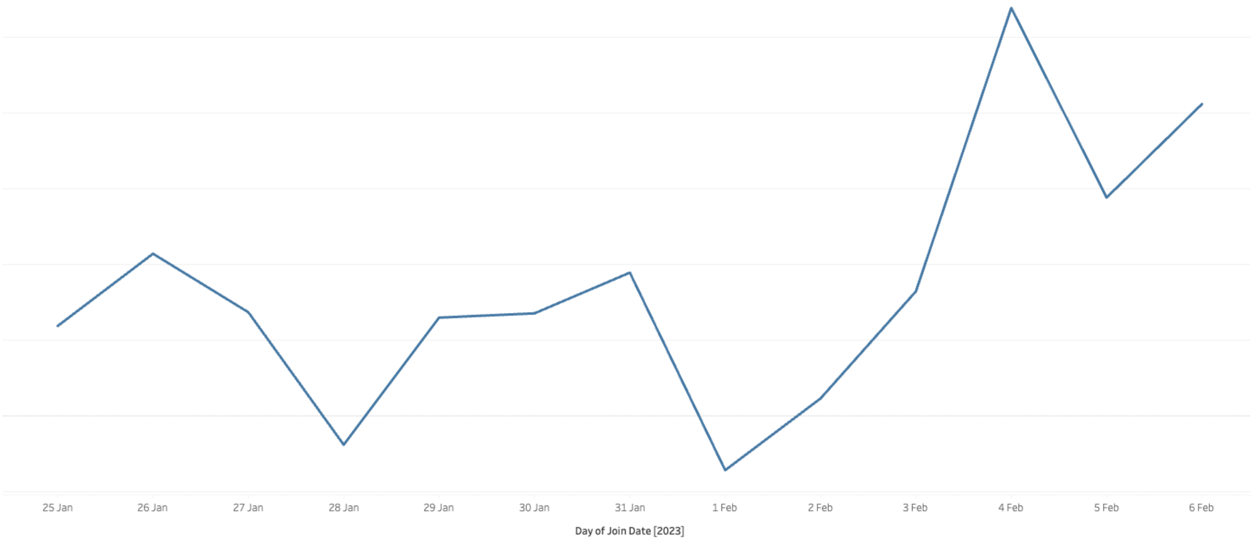
Country:

An interactive map was created to visualise the difference in conversion rates and average spend between countries, this visualisation has not been included in this report because the information relating to the users country isn’t particularly helpful to us in this particular test. If we had more information relating to the products available on the website, and if any of these were not available in certain countries, we might have found this data more useful. The country information was not available for all users, and a decision was made not to include the unknown values in the visualisation for the country. This is because the country visualisation is an interactive map, and including a pie chart for the unknown values would have been visually confusing. It doesn’t particularly aid the analysis to include those unknown values, however it might be helpful to consider encouraging that information on future tests, where information on product availability by location is also available. A link to all of the tableau visualisations can be found at appendix 3 should it be necessary to explore regional differences.

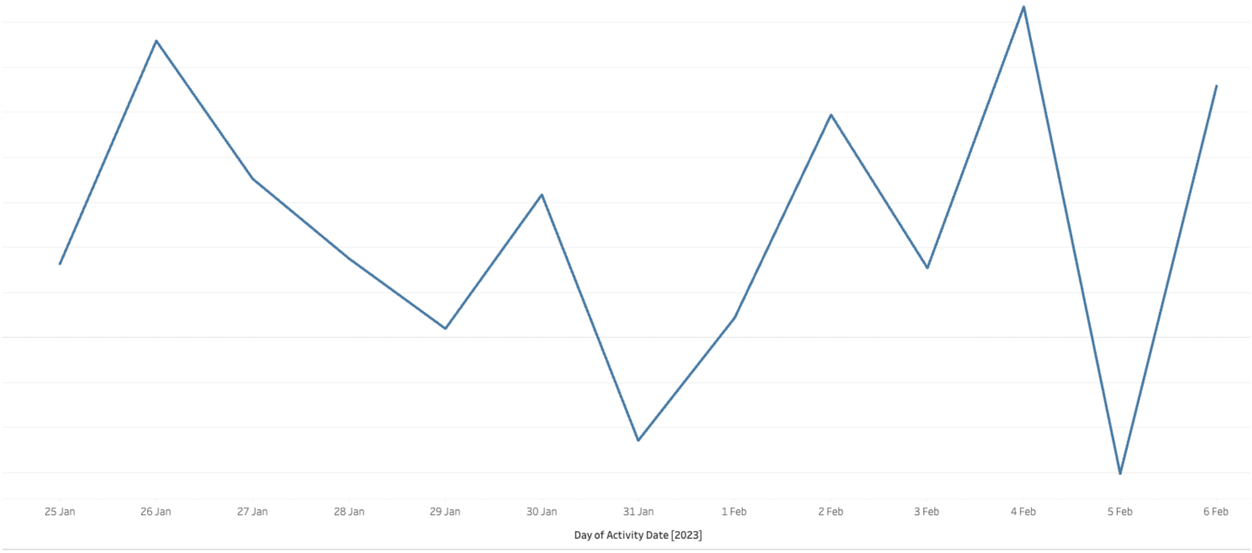
Novelty Effect:

A novelty effect refers to the initial excitement or curiosity that people may experience when they encounter something new or different. In the context of this A/B test, it means that when we introduce the banner, some users may show different behaviours simply because they are curious about the change. It is a temporary response to something new, like the initial excitement of trying a new flavour of ice cream, and how you’re likely to feel less excited about the flavour once you have sampled it a few times. It is essential to consider because it may influence the early results of our A/B test. As the novelty wears off, user behaviour could return to normal levels, so it is a good idea to look at the results over a longer period of time to understand the lasting impact of the change. If we were to agree that we are observing a novelty effect we would expect to see an initial spike, followed by a decline.

The difference in conversion rates between the two groups over the course of the test:



The difference in average spend between the two groups over the course of the test:



While we do observe an initial spike and decline that is indicative of a novelty effect, this is followed by a series of spikes and dips for both the difference in conversion rate and the difference in average spend, and this is likely to indicate other factors at play.
The test only runs over 12 days, so it is possible that the spikes and dips are unrelated to a novelty effect, and in an experiment of this length it wouldn't be unusual for daily fluctuations in behaviour to have an impact on the results. These types of fluctuations can include day of the week, mood, news, or other external events.

Based on this, the initial spike and decline in both the difference in conversion rates and the difference in average spend between groups is possibly indicative of a novelty effect. But in the context of the short time frame, and given the other potential factors that can be observed in relation to the significant spikes and dips, it seems more likely that external factors are at play. In order to obtain a definitive answer in relation to novelty effect, it may be beneficial to run the experiment over a longer period of time, to assess the metrics and see if the patterns persist or change.

Power Analysis:

When conducting a power analysis on the difference in conversion rates between groups, using the conversion rate for the control group as a baseline, with the minimal detectable effect set to 10%, the suggested total sample size is 60.6k (30.3k per group). The A/B test was able to observe a significant difference in conversion rates with a slightly smaller sample size of just under 50k, a 10% difference was initially targeted for detection of a significant difference, however a larger difference of 18% was observed instead. This difference of 18% was so large that the sample size we already have is more than adequate, and we wouldn't need to conduct a new test with a larger sample size.

When conducting a power analysis on the difference in average spend between groups, using the average amount spent for the control group as a baseline, the pooled standard deviation and the minimal detectable effect set to 10%, the suggested total sample size is 182,164k (91,082k per group). A 10% difference was initially targeted for detection of a significant difference, however a much smaller difference of 0.5% was observed instead. This difference is so small that it would present a challenge to obtain a sample size large enough to satisfy the statistical power, and since the difference is so small, even with a larger sample size it may not reach the required level of statistical significance. Moreover, the small effect size means that the observed difference is not practically meaningful or significant for our purposes. In other words, it doesn't have a significant impact on our goals or objectives. To make the results more meaningful we would need to focus on increasing the effect size itself.

Recommendation:

The final recommendation is to launch the banner to all users. To recap, the motive for carrying out the A/B test was to discover whether introducing a new banner advertising the ever growing selection of food and drinks offerings, would increase revenue, and encourage users to shop those items.

It was understood from the brief that the main goal was to draw users to these items, as they were relatively new to the website, and because they were different from the website's usual offerings, they hadn't received much traffic - possibly because users were not aware of their introduction. The analysis observed overall higher conversion rates in the treatment group when compared to the control group across all metrics, which does support the conclusion that the users in the treatment group were being drawn to, and buying these products.

The analysis did not observe a significant difference in average spend between the two groups, which would mean that the banner did not increase revenue. Based on the knowledge that the website primarily offers boutique fashion items and high end decor, it would be reasonable to assume that the cost of the food and drinks items is significantly less than the usual fare. In this instance it would be expected to observe more users converting in the treatment group, but spending less money overall.

When considering business relevance over statistical significance, the overall user experience should be taken into consideration. The treatment group showed an increase in conversion rate, and even without a significant difference in average spend, this suggests that the banner is engaging users, and positively impacting their experience on the website. This can lead to increased brand awareness, improved user satisfaction, and therefore the potential for increased revenue. This means that even though the A/B test did not demonstrate a statistically significant difference in average spend between the two groups, the analysis as a whole, along with the business relevance suggests that running a new A/B test, over a longer period of time with a larger sample size is not necessary.

Although the banner occupies valuable real estate on the website, the cost of launching the banner itself is relatively low and other factors such as the ease of implementation, potential revenue gains over time, increased brand awareness, and the potential for iterative improvements, make launching the banner to all users both cost effective and practical, and for this reason the recommendation is to launch the banner to all users.

Appendix:

Initial SQL query:

```
SELECT
  u.id AS user_id,
  u.country,
  u.gender,
  COALESCE(g.device, 'Unknown') AS device,
  g.group AS test_group,
  CASE WHEN total_spent > 0 THEN true ELSE false END AS converted,
  COALESCE(total_spent, 0) AS total_spent,
  COALESCE(group_a_total_spent, 0) AS group_a_total_spent,
  COALESCE(group_b_total_spent, 0) AS group_b_total_spent
FROM
  users u
LEFT JOIN
  groups g ON u.id = g.uid
LEFT JOIN
  (
    SELECT uid, COALESCE(SUM(spent), 0) AS total_spent
    FROM activity
    GROUP BY uid
  ) a ON u.id = a.uid
LEFT JOIN
  (
    SELECT g.uid, COALESCE(SUM(a.spent), 0) AS group_a_total_spent
    FROM groups g
    LEFT JOIN activity a ON g.uid = a.uid
    WHERE g.group = 'A'
    GROUP BY g.uid
  ) group_a ON u.id = group_a.uid
LEFT JOIN
  (
    SELECT g.uid, COALESCE(SUM(a.spent), 0) AS group_b_total_spent
    FROM groups g
    LEFT JOIN activity a ON g.uid = a.uid
    WHERE g.group = 'B'
    GROUP BY g.uid
  ) group_b ON u.id = group_b.uid;
```

Google sheets link:

https://docs.google.com/spreadsheets/d/12hPxjCrgWbP1wI_ybGCOzE7IalmMgH8D5Ki1Rw6sKCc/edit#gid=430977844

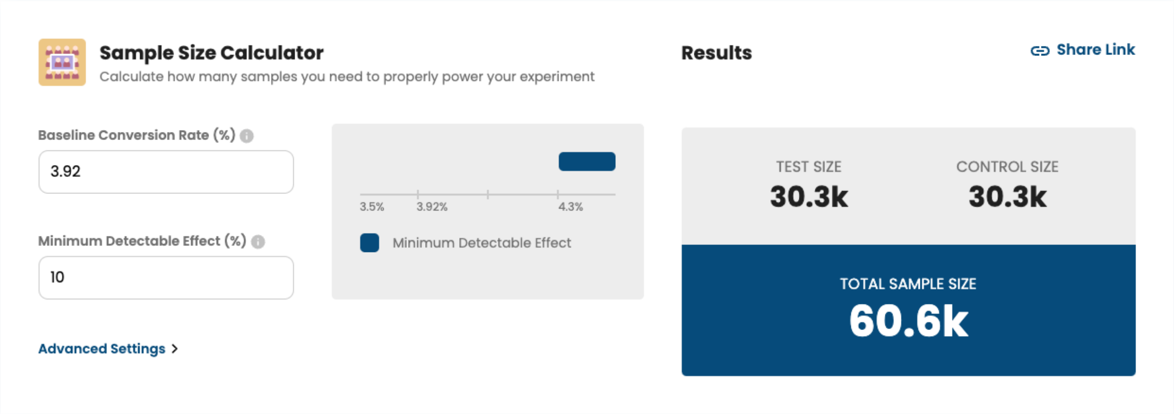
Tableau public link:

https://public.tableau.com/app/profile/lauren.kemmy/viz/Globoxvis-LaurenKemmy/ConversionRatevsAverageSpend

Second SQL query:

```
SELECT
  u.id AS user_id,
  COALESCE(u.country, 'Unknown') AS country,
  COALESCE(u.gender, 'Unknown') AS gender,
  COALESCE(g.device, 'Unknown') AS device,
  COALESCE(g.group, 'Unknown') AS test_group,
  g.join_dt AS join_date,
  CASE WHEN a.spent > 0 THEN true ELSE false END AS converted,
  COALESCE(a.spent, 0) AS total_spent,
  a.dt AS activity_date,
  CASE WHEN g.group = 'A' THEN COALESCE(a.spent, 0) ELSE 0 END AS total_spent_a,
  CASE WHEN g.group = 'B' THEN COALESCE(a.spent, 0) ELSE 0 END AS total_spent_b
FROM
  users u
LEFT JOIN
  groups g ON u.id = g.uid
LEFT JOIN
  (
    SELECT uid, device, SUM(spent) AS spent, MAX(dt) AS dt
    FROM activity
    GROUP BY uid, device
  ) a ON u.id = a.uid;
```

Power Analysis Calculators:



Sample Size Calculator

Calculate how many samples you need to properly power your experiment

Baseline Conversion Rate (%): 3.92

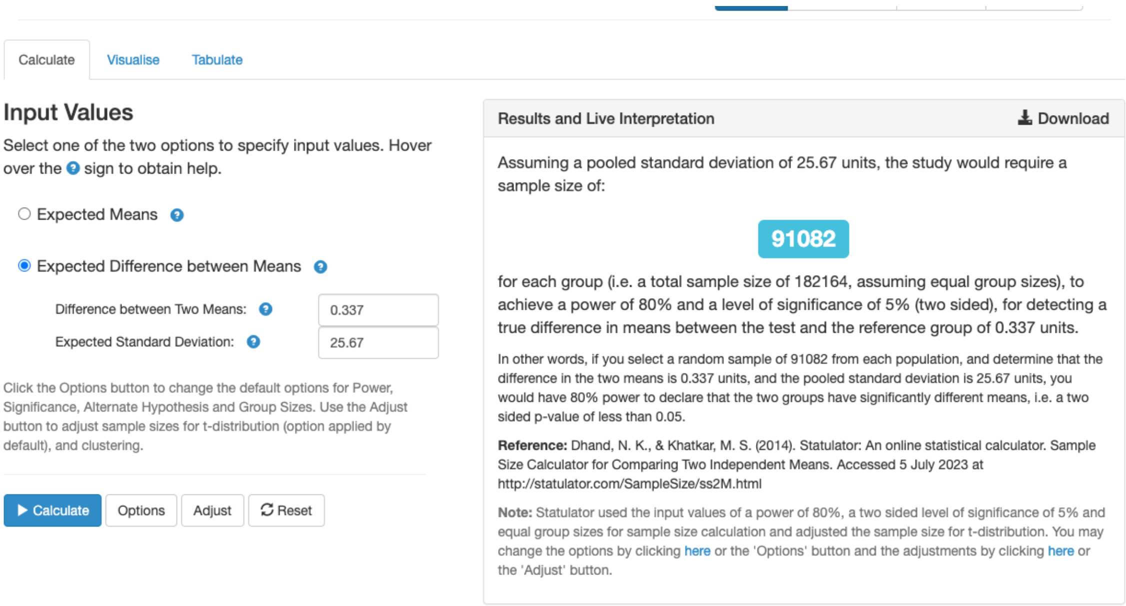
Minimum Detectable Effect (%): 10

Results

TEST SIZE 30.3k

CONTROL SIZE 30.3k

TOTAL SAMPLE SIZE 60.6k



Input Values

Select one of the two options to specify input values. Hover over the ? sign to obtain help.

○ Expected Means

● Expected Difference between Means

Difference between Two Means: 0.337

Expected Standard Deviation: 25.67

Click the Options button to change the default options for Power, Significance, Alternate Hypothesis and Group Sizes. Use the Adjust button to adjust sample sizes for t-distribution (option applied by default), and clustering.

Results and Live Interpretation

Assuming a pooled standard deviation of 25.67 units, the study would require a sample size of:

91082

for each group (i.e. a total sample size of 182164, assuming equal group sizes), to achieve a power of 80% and a level of significance of 5% (two sided), for detecting a true difference in means between the test and the reference group of 0.337 units.

In other words, if you select a random sample of 91082 from each population, and determine that the difference in the two means is 0.337 units, and the pooled standard deviation is 25.67 units, you would have 80% power to declare that the two groups have significantly different means, i.e. a two sided p-value of less than 0.05.

Reference: Dhand, N. K., & Khatkar, M. S. (2014). Statulator: An online statistical calculator. Sample Size Calculator for Comparing Two Independent Means. Accessed 5 July 2023 at http://statulator.com/SampleSize/ss2M.html

Note: Statulator used the input values of a power of 80%, a two sided level of significance of 5% and equal group sizes for sample size calculation and adjusted the sample size for t-distribution. You may change the options by clicking here or the 'Options' button and the adjustments by clicking here or the 'Adjust' button.