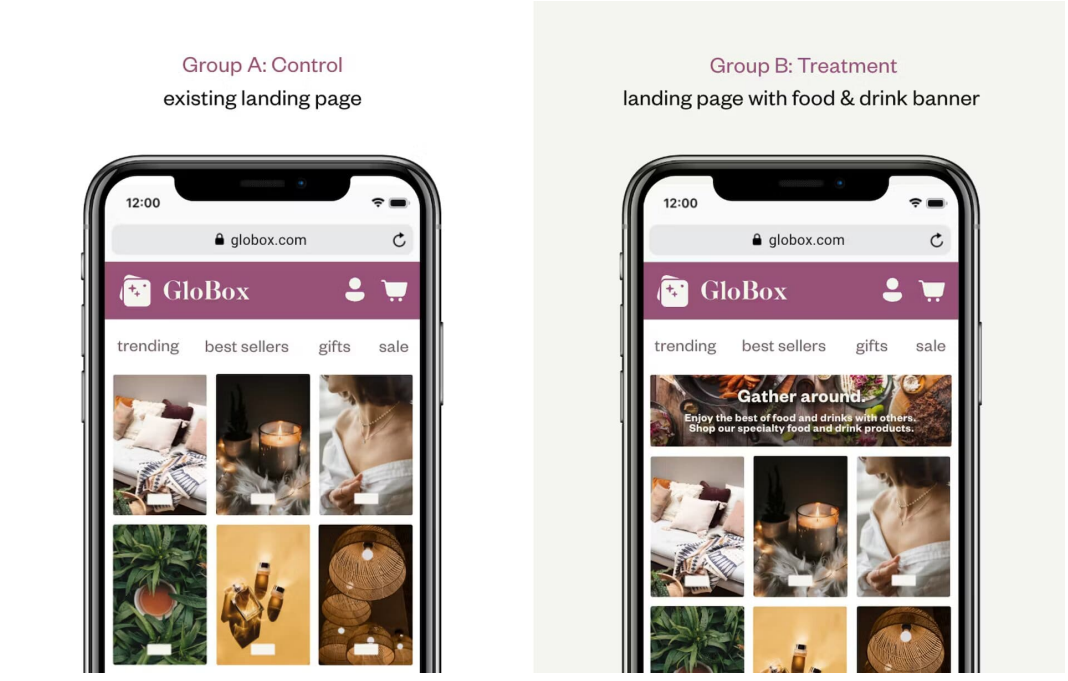
A/B testing for GloBox main webpage: written report

By Fedor Dorokhov

Introduction

Globox company is mostly known among its customers as an online marketplace which is sourcing high-quality fashion items and décor products. However, in light of sufficient growth of food and drink offerings, the company would like to increase customer base's awareness of that category of products by implementing the special banner, which is highlighting key products from the "Food and drinks" category on top of the mobile webpage. The goal of current report is to demonstrate findings of A/B test conducted in order to decide, will that banner actually impact the sales' revenue.



The purpose of the experiment is to compare performance of Control group (Group A) with Treatment group (Group B) to determine if the changes in conversion rate and average amount spent per user are statistically significant.

Also, the further analyses were conducted for deeper and more elaborated understanding of experiment's outcome:

1) Visualization of the distribution of average amount spent per user for each group.
2) Exploration of relationships between user's gender/country/used device and key test metrics (conversion rate and average amount spent per user).

Methodology

The experiment was conducted only on the mobile website. When user visited the Globox main page, he was randomly assigned to the control or treatment group. In other words, randomized controlled trial design (RCT) was used. Control group interacted with a website version contained no banner, while treatment group experienced the new version with food and drink category-related banner on top of the page. The sample size consisted of **48943** participants (**24343** for control group and **24600** for the treatment one) and the experiment was conducted over a time period of **13** days (**25/01/2023 — 06/02/2023**).

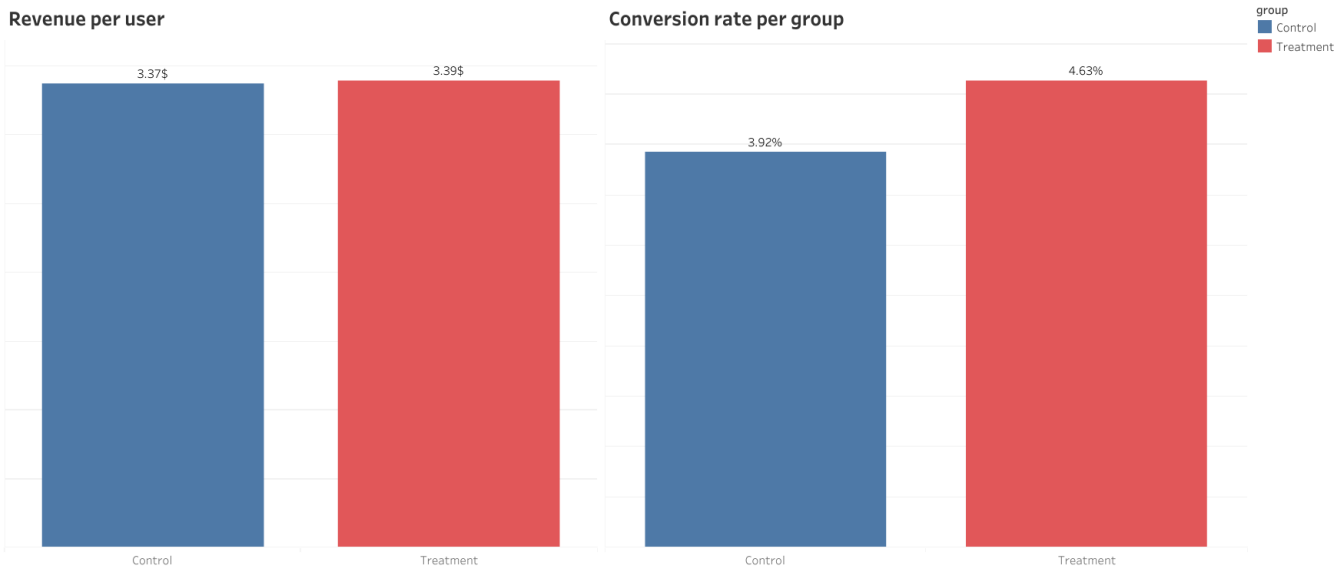
Results

Does banner make customers buy more often?

The control group demonstrated conversion rate of **3.92%**, which means the percentage of website visitors that interacted with an "old" version of the website (without a banner) and made a conversion (bought something). The treatment group, however, which experienced the interaction with a banner-implemented version of the webpage reached the conversion rate of **4.63%**. The p-value of the difference between conversion rates of both groups was calculated to be 0.0001, which shows strong evidence for statistically significant difference. The 95% confidence interval for that difference in conversion rates is **[0.0035, 0.0107]**.

Do customers actually spend more (in average)?

With respect to another key metric measured – mean amount spent per user, the picture is different. The control group resulted with the amount of 3.37$, when treatment group's one was 3.39$. The p-value of the difference was calculated to be 0.944, which clearly indicates statistical insignificance of the difference between two groups. The 95% confidence interval for the difference in average amount spent is **[-0.439, 0.471]**.

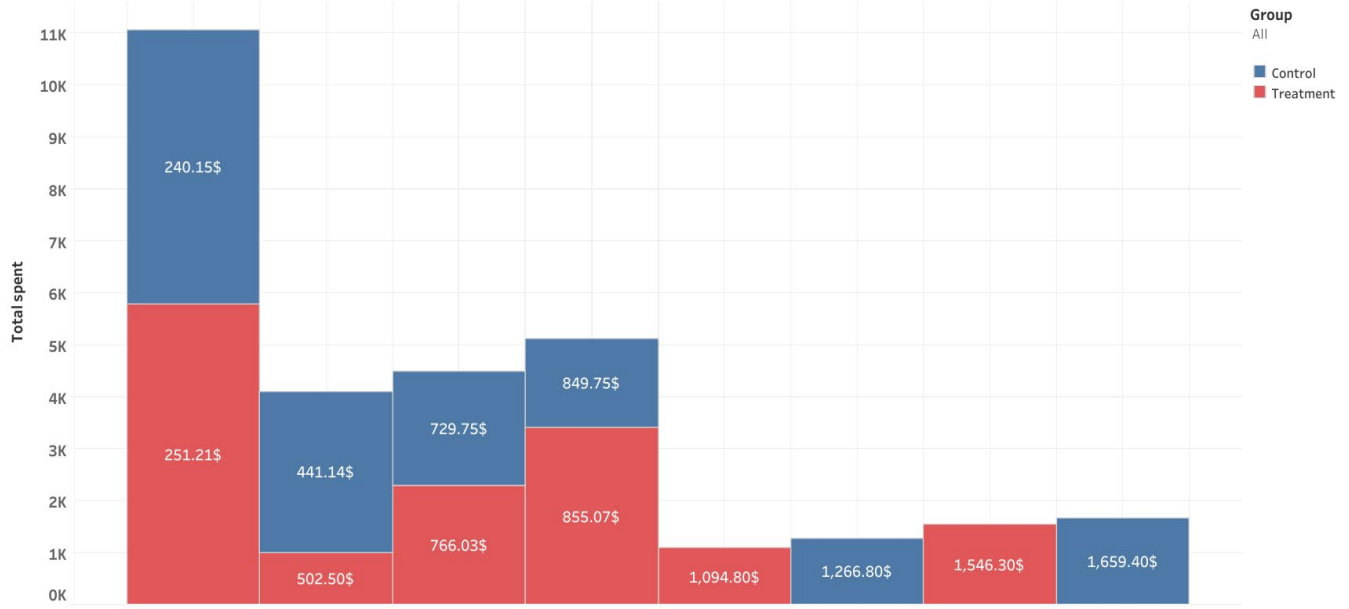


If sample size is unreachable, is there a point?

Also, power analysis was made to understand if sample size is sufficient for our experiment. For difference in conversion rates testing **48500** participants was detected as sample size required for our test to be sufficiently sensitive, in other words, our sample was proven to be powerful enough (detected percent difference between groups is **18%**). However, for difference in average spent by user, detected sample size required was **25860054** for each group (**51720108** in total), which is quite an enormous number unachievable from the business perspective!

Unveiling insights through histogram

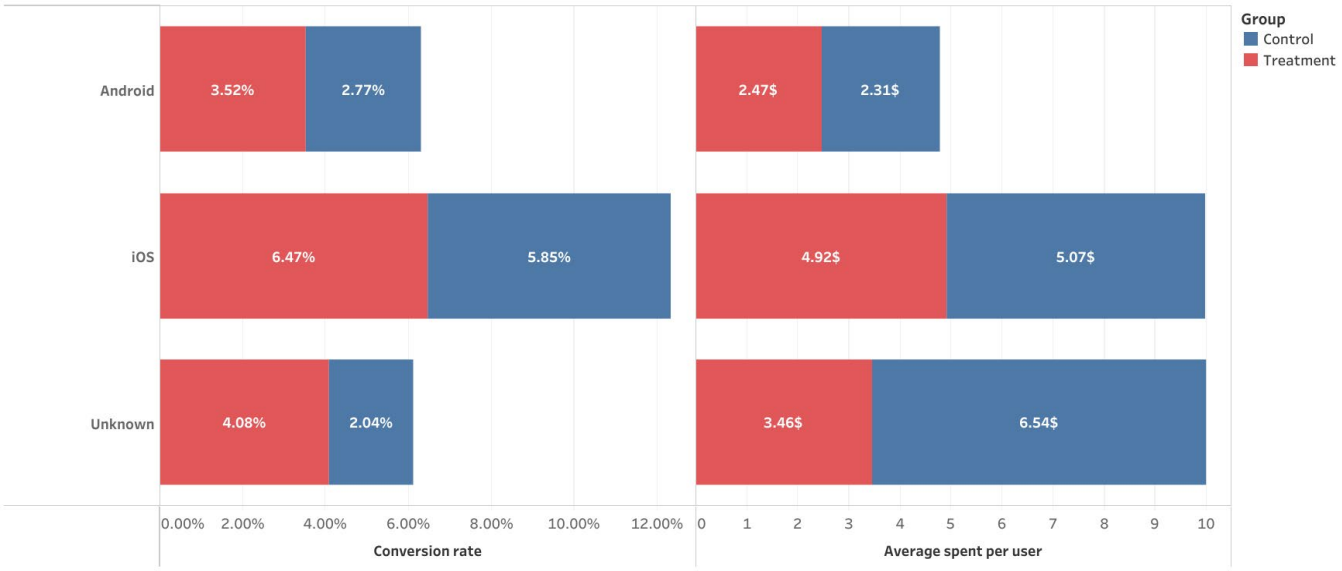
To understand distribution of average amount spent per user for each group, histogram with 8 bins was created. The horizontal axis represents the range of mean amount spent, 200$ for each bin, while vertical axis shows the total amount spent within each range. The first range of 0-200 was excluded from the analysis as clear outlier that affects the final result. The histogram revealed the highest revenue in total was received from the customers whose average amount spent was in range 200-400, with control group's mean of 240.15$ and treatment group's of 251.21$. The second insight is related to skewness's difference per group – if control group distribution is a classic right-skewed, the treatment one is asymmetrical unimodal, even if it still shows the same skewness in some extent. That difference, in my opinion, demonstrates the lack of distribution's predictability with respect to group, that received a treatment.



Android or iOS?

Several relationships between conversion rate and average amount spent per user on one side and a few specific variables on another one, in particular, user's device, gender and country, were analyzed for better and 'wider' understanding of A/B

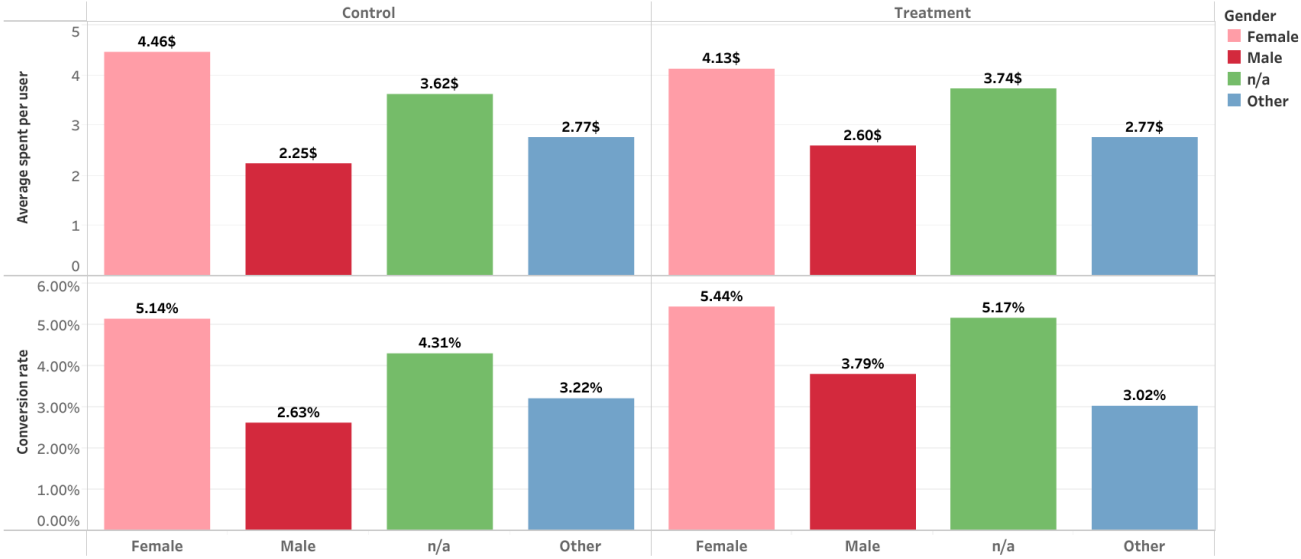
testing's outcome. Regarding customer's device-test metrics relationship couple interesting insights were revealed. First, the only device demonstrated growth with respect to both metrics was Android, however, iOS seems to be much more popular among converted users and approximately doubles mean amount spent per user compare to it's rival. Nevertheless, regarding to the average spent, iOS users performed worse in 2.96% after the treatment, but the situation with customers, whose devices remained unknown is even more peculiar. In fact, that group doubled the conversion rate after the treatment, but at the same time demonstrated 47% decrease in average amount spent! That controversy should definitely makes us doubt the expediency of experiment's launching and emphasize the necessity of further researches.



Gender matters!

Relationship between key test metrics and user's gender was explored, which allowed us to reveal few valuable insights. First of all, females appeared to be a gender group with the highest scores with respect to both metrics for control and treatment groups. Secondly, the overall performance generally follows the tendency shown in the 'main analysis' – if regarding to conversion rate the raise was demonstrated for all categories, except the 'Other' group, with respect to amount average spent it is not so smooth, where only the male group showed relatively

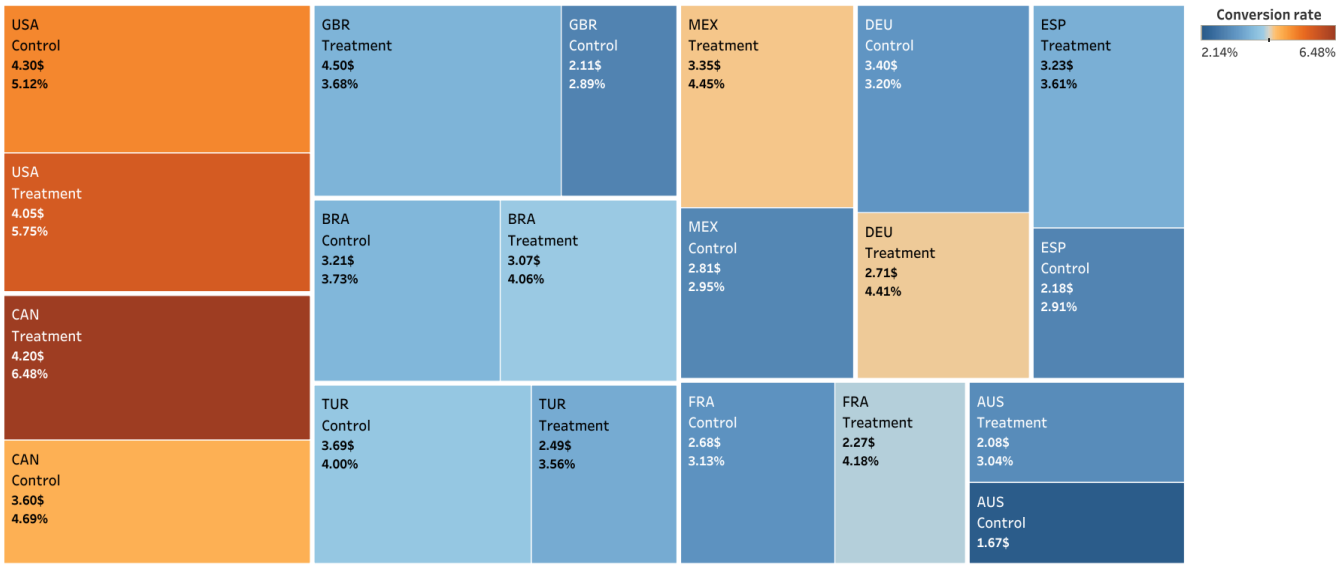
sufficient growth of 15.5%. Compare it to the same gender category in the conversion rate part, where growth was more than 44%! Also, the most "active" female group demonstrated average amount spent decrease after the treatment, which should be the warning signal, considering, how important that gender category is for our customer base.



Country-level zoom-in

The last relationship explored between test metrics and customer's country was definitely not least. The first insight further illustrates divergence of ways, how conversion rate and average amount spent per user change after the treatment. If 9 of 10 observed countries clearly demonstrated, that number of converted customers grows, it is not the case for the second key metric, where customers from 5 of 10 countries showed decrease of amount spent in average. Great Britain example can be noted as a curious anomaly, where numbers didn't just grow in both metrics, but customers expenses were more than doubled (113% raise)!

Nevertheless, the examples of customers' performance from US and continental European countries (France and Germany) demonstrates instability with respect to the average revenue and it should be noted in light of making a final decision regarding to launching the experiment.



Recommendations

Based on the described results I would like to make the following recommendations: **I don't recommend to launch the experiment.** The A/B test didn't show a statistically significant increase of average amount spent per user, the difference between two groups is slightly higher than half percent. In my opinion, that metric should be prioritized, since the whole point of the experiment was to determine, if revenue will be raised in a relatively large extent. Even though, a statistically significant difference in conversion rate was detected and power analysis showed that test is powerful enough it is not so crucial if actual user spends don't raise in a sufficient extent! The banner takes a lot of space on the main webpage, and we can't be confident, that it really worth it. The relationships between key test metrics and user's device/gender/country also don't seem optimistic and look 'instable' regarding to received revenue per user. Furthermore, the power analysis conducted to understand, if sample size for average spent per user testing is large enough revealed, that sufficient sample size for that test to be sensitive enough is too large (millions) and very unlikely to be achieved. And even though it is not the case for conversion rate, which is increased – what is the point of more customers making a

conversion if it doesn't lead to increase of a revenue? Is it the case, that our banner distracts users' attention from more expensive goods to cheaper ones?

Appendix

SQL queries:

What are the start and end dates of the experiment?

```
SELECT Min(join_dt) AS start_date,
       Max(join_dt) AS end_date
FROM   (SELECT u.id,
               u.country,
               u.gender,
               COALESCE(a.uid, 0) AS activity_uid,
               g.join_dt
        FROM   users u
               LEFT JOIN groups g
                      ON u.id = g.uid
               LEFT JOIN activity a
                      ON g.uid = a.uid) sub
```

How many total users were in the experiment?

```
SELECT Count(id) AS total_sample
FROM   users
```

How many users were in the control and treatment groups?

```
SELECT groups.group,
       Count(*) AS group_sample
FROM   groups
GROUP  BY 1
```

What was the conversion rate of all users?

```
SELECT Count(*) AS user_counter,
       conversion
FROM   (SELECT u.id,
               u.country,
               u.gender,
               g.group,
               g.device,
               CASE
```

```
                  WHEN a.spent > 0 THEN 'True'
                  ELSE 'False'
                END           AS Conversion,
                Sum(a.spent) AS total_spent
        FROM    users u
                LEFT JOIN groups g
                       ON g.uid = u.id
                LEFT JOIN activity a
                       ON a.uid = g.uid
        GROUP   BY 1,
                   2,
                   3,
                   4,
                   5,
                   6) t1
GROUP   BY 2
```

Query calculates the total number of users who made a conversion and who didn't. Final answer for the question is in spreadsheet: https://docs.google.com/spreadsheets/d/1N2PLL6hWrMbhP80xIMxAFjB9Hy8WMjzjik2b5nILrrg/edit?usp=sharing

Or in Tableau story:

https://public.tableau.com/views/VisuzalizationsfortheGloboxproject/VisualizationoftheABtestingresults?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

What was the conversion rate for the control and treatment group?

```
SELECT Count(*) AS user_counter,
       t1.group,
       conversion
FROM   (SELECT u.id,
                u.country,
                u.gender,
                g.group,
                g.device,
                CASE
                  WHEN a.spent > 0 THEN 'True'
                  ELSE 'False'
                END           AS Conversion,
                Sum(a.spent) AS total_spent
        FROM    users u
                LEFT JOIN groups g
                       ON g.uid = u.id
                LEFT JOIN activity a
                       ON a.uid = g.uid
        GROUP   BY 1,
                   2,
```

```
                  3,
                  4,
                  5,
                  6) t1
WHERE  conversion = 'True'
GROUP  BY 2,
          3
```

As in a previous question, the query shows how many users per group made a conversion. The final result can be found in Tableau story attached:

https://public.tableau.com/views/VisuzalizationsfortheGloboxproject/VisualizationoftheABtestingresults?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

or in spreadsheet:
https://docs.google.com/spreadsheets/d/1N2PLL6hWrMbhP80xIMxAFjB9Hy8WMjzjik2b5nILrrg/edit?usp=sharing

What is the average amount spent per user for the control and treatment groups, including users who did not convert?

https://docs.google.com/spreadsheets/d/1N2PLL6hWrMbhP80xIMxAFjB9Hy8WMjzjik2b5nILrrg/edit?usp=sharing

SQL query used to extract A/B test data from the database for further analysis:

```
WITH t1
     AS (SELECT u.id,
                u.country,
                u.gender,
                g.group,
                g.device,
                CASE
                  WHEN a.spent > 0 THEN 'True'
                  ELSE 'False'
                END          AS Conversion,
                Sum(a.spent) AS total_spent
         FROM   users u
```

```
                LEFT JOIN groups g
                       ON g.uid = u.id
                LEFT JOIN activity a
                       ON a.uid = g.uid
          GROUP  BY 1,
                    2,
                    3,
                    4,
                    5,
                    6)
SELECT id,
       COALESCE(country, 'N/A')  upd_country,
       COALESCE(gender, 'N/A')   upd_gender,
       t1.group,
       COALESCE (device, 'N/A')  upd_device,
       conversion,
       COALESCE (total_spent, 0) total_spent
FROM   t1

ORDER BY 4
```

A/B Test Statistics calculations:

https://docs.google.com/spreadsheets/d/1N2PLL6hWrMbhP80xIMxAFjB9Hy8WMjzjik2b5nILrrg/edit?usp=sharing

Tableau visualizations for the project:

https://public.tableau.com/views/VisuzalizationsfortheGloboxproject/VisualizationoftheABtestingresults?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Power analysis of the sample size for conversion rates difference testing:

https://www.statsig.com/calculator?mde=12.6&bcr=3.92&twoSided=true&splitRatio=0.5&alpha=0.05&power=0.8

Power analysis of the sample size for two independent means difference testing:

See attached PDF file “Sample size calculation result for two independent means”