A/B test for “GloBox”

Written report about the analysis
by Olga Shuvalov
02.08.2023

1. Summary
2. Context
3. Results
4. Recommendations
5. Appendix

1. Summary:

Was run A/B test and I received the tables with available information about users, the groups (A and B) and their purchases after the test.

I have calculated A/B test statistics using spreadsheets for evaluating the impact of launching the banner. Conversion Rate and Average Amount Spent per user were chosen as the success metrics in the case.

When were conducted Hypothesis tests for this metrics, where was found that there is a statistically significant difference between Conversion Rate of groups A and B. While in Average Amount Spent was not found a significant difference.

I checked the behavior of the users for novelty effect and ran the Power Analysis as well: it was determined that there is no novelty effect, and the amount of data is sufficient for these results.

I saw enough improvement in our metrics of success that I feel confident in releasing the banner to all users. Even though only in one metric I found a noticeable increase. I suppose that the perceived cost of launching the feature is worth the benefits that I saw in the A/B test. So, my recommendation is to launch the banner.

2. Context.

a) Motivation:

GloBox is an online marketplace that specializes in sourcing unique and high-quality products from around the world. GloBox is primarily known amongst its customer base for boutique fashion items and high-end decor products. However, their food and drink offerings have grown tremendously in the last few months, and the company wants to bring awareness to this product category to increase revenue.

We want to look at whether the new food and drink banner is leading to changes in the success metrics. Other words to compare two versions of the webpage for mobile devices: with the banner and without.

b) Parameters:

- The size of the test groups approximately equal: 24,343 users in A-group and 24,600 users in B-group. So total users is 48,943
- The start and the end dates for the experiment are Jan 25th, 2023 — Feb 6th, 2023
- Was run an A/B test that highlights key products in the food and drink category as a banner at the top of the website. The control group (A) does not see the banner, and the test group (B) sees it.
- The experiment is only being run on the mobile website.
- The user subsequently may or may not purchase products from the website. It could be on the same day they join the experiment, or days later. If they do make one or more purchases, this is considered a “conversion”.

c) Overview of the data set:

- There are 3 tables in the GloBox Database:
  users: user demographic information
  - **id**: the user ID
  - **country**: ISO 3166 alpha-3 country code

- **gender**: the user's gender (M = male, F = female, O = other)

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 (visited the page)
- **device**: the device the user visited the page on (I = iOS, A = android)

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 (I = iOS, A = android)
- **spent**: the purchase amount in USD

- All users should be assigned to one A/B test group
- A user visits the GloBox main page and is randomly assigned to either the control or test group. This is the join date for the user in the table Groups.
- Not all users make a purchase
- Purchase activity is for all product categories, not just food and drink

d) Methodology:

- Preprocessing of the data. I handled the data with SQL queries to get a consolidated analysis dataset that joins the tables together and calculates the user-level aggregations. I have dealt with NULL values that appear after joining tables. I removed all zero values in the spent-column and changed them to 0 (zero) for further calculations to avoid errors. I got my analysis dataset I CSV-format.
- To compare two groups, I've used a *statistical method* - ***hypothesis testing*****.** To see whether there is a difference in the conversion rate and in the Average Amount Spent per user between two groups.

In this method we assume a null hypothesis and an alternative and then compute a test statistic assuming that the null hypothesis is true. A test statistic is a statistic calculated from the sample data used as evidence in the hypothesis test.

The tests relies on the assumptions and principles of the Central Limit Theorem. The CLT lays the foundation for the z and t-tests for comparing two samples by providing the principles of normality and standard error estimation necessary for conducting this hypothesis test. The resulting test statistic follows a standard normal distribution under the null hypothesis. CLT also states that the standard deviation of the sample means, known as the standard error, can be estimated using the standard deviation of the population (all data for both groups) and the sample size (size of each group, A or B).

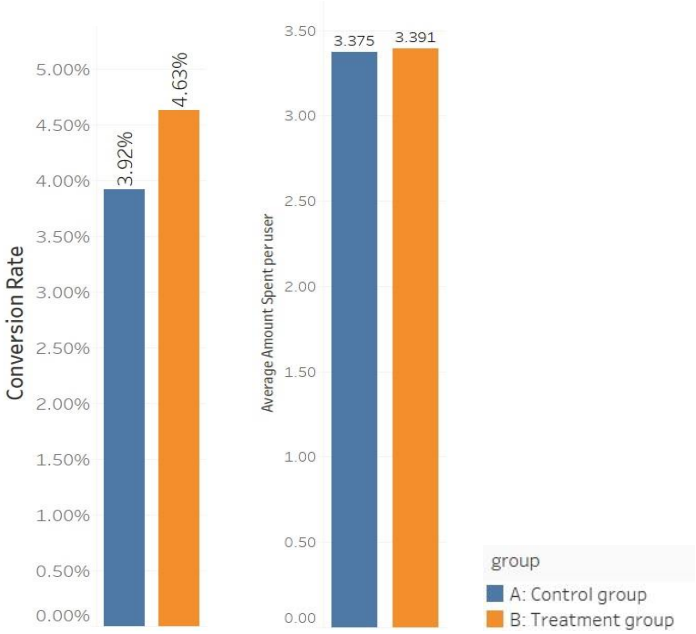
We compare the test statistic to a critical value to determine if the difference between the sample means is statistically significant.

- I calculated the 95% Confidence Interval for both differences – a range of estimations for the differences.
- It was chosen 5% significance level for both differences (α=0.05)

3. Results

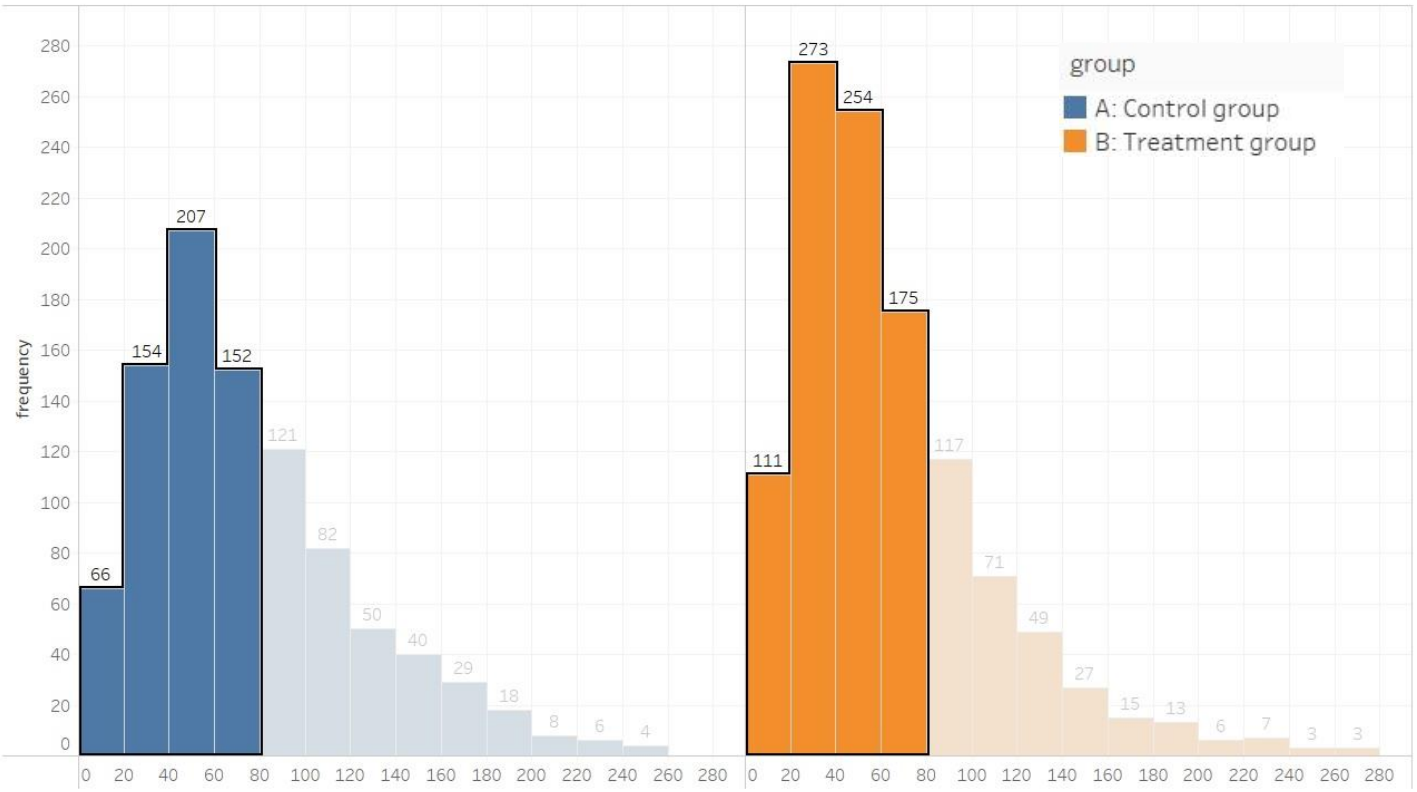
I. Descriptive statistics. Success metrics

- **Conversion Rate and Average Amount Spent per user:**



Here we got the different values for CR in two groups:
In group A (control) CR = 3.92%
In group B (treatment) CR = 4.63%
And almost the same at first glance values for AAS.
Whether this is really a significant difference, we will see later.

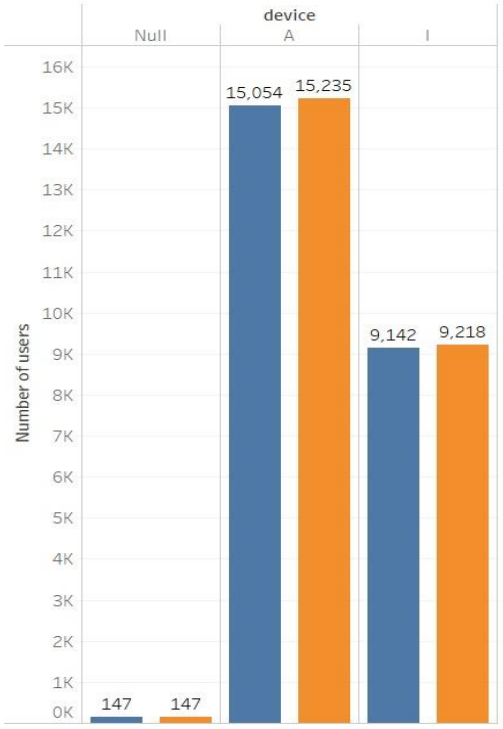
- **The distribution of Amount Spent per user for each group:**



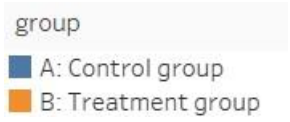
By removing the zeros (by filter: sum of spent per user > 0) and outliers, I aimed to focus solely on the distribution of the remaining data points.

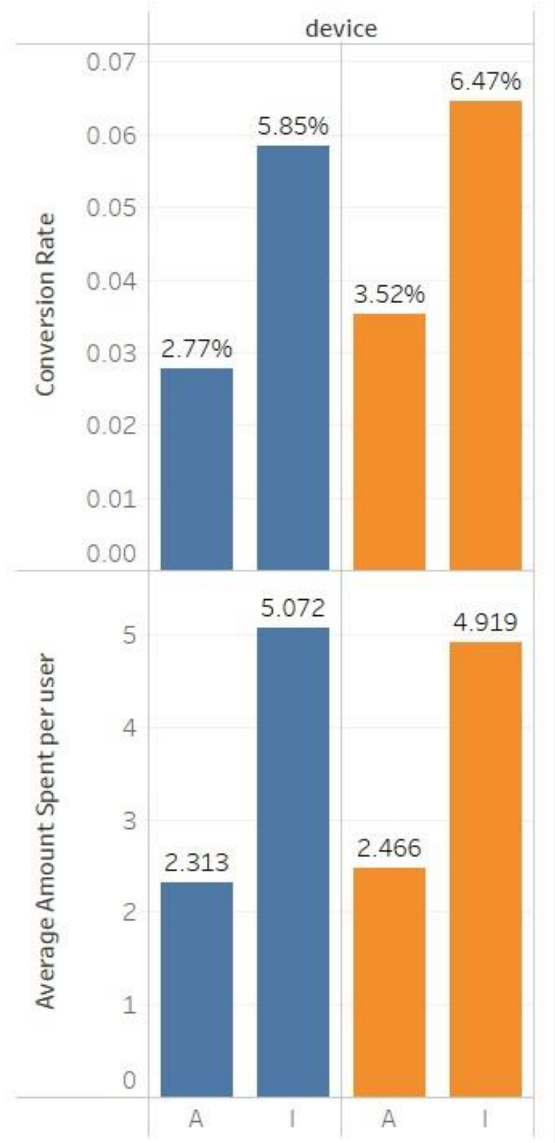
Here we see that was increased frequency of purchasing for smaller amounts (AAS per user: 0 - 80).

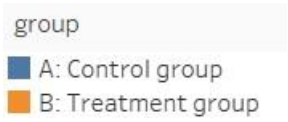
- **The relationship between the test metrics and the user's *device*.**



A – Android and I – ios
Each group has a similar distribution by device.

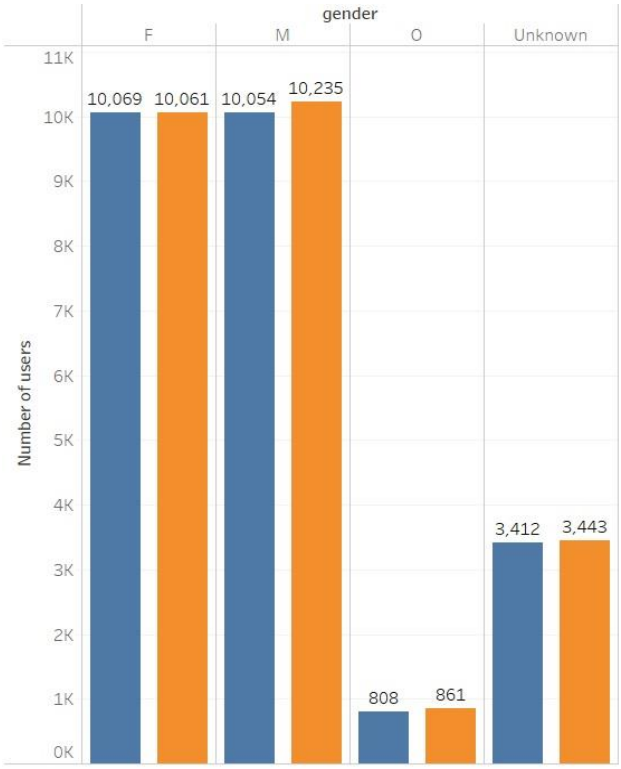




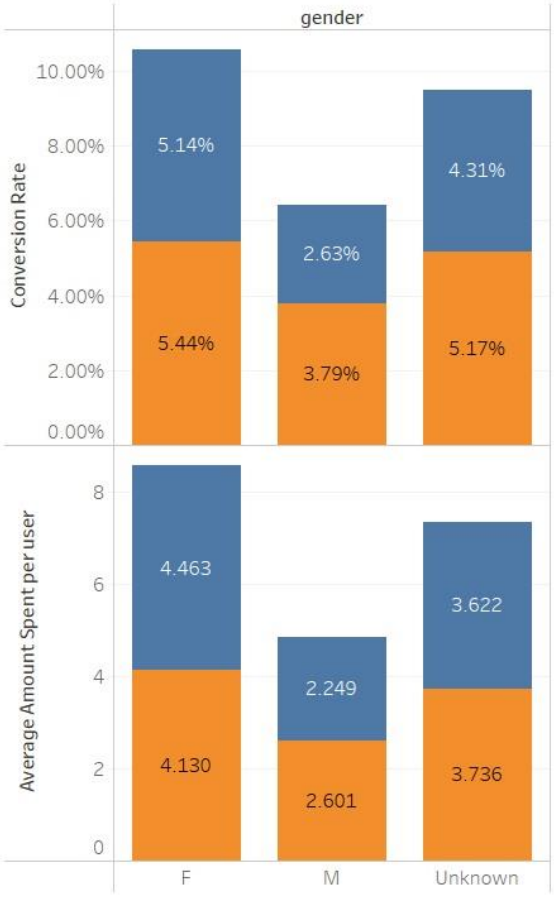


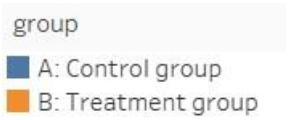
All the metrics have grown except the AAS for ios.

- **The relationship between the test metrics (conversion rate and average amount spent) and the user's *gender***



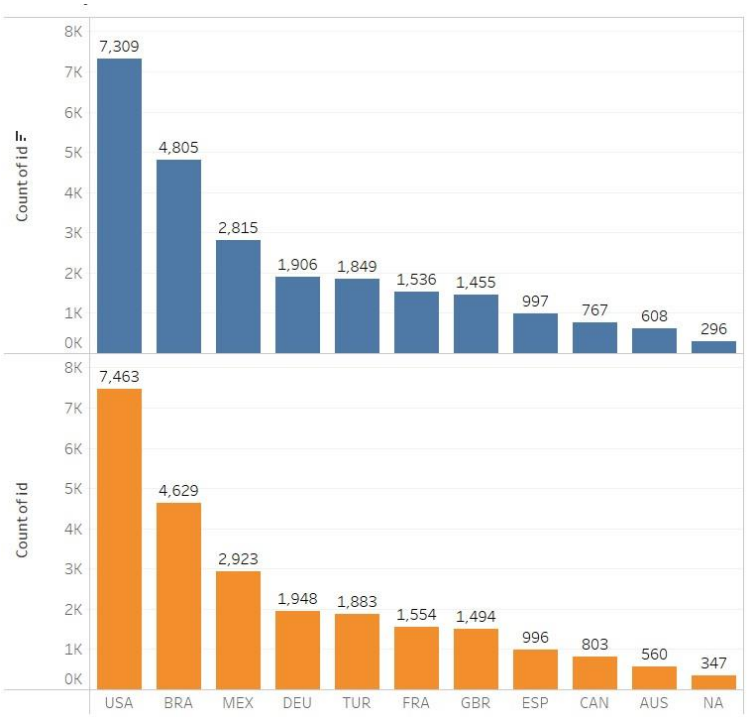
I left here the zero-value and even the unknown gender, since it's clear that both groups have approximately the same characteristics for all options



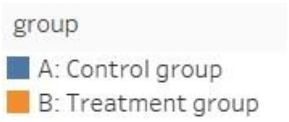


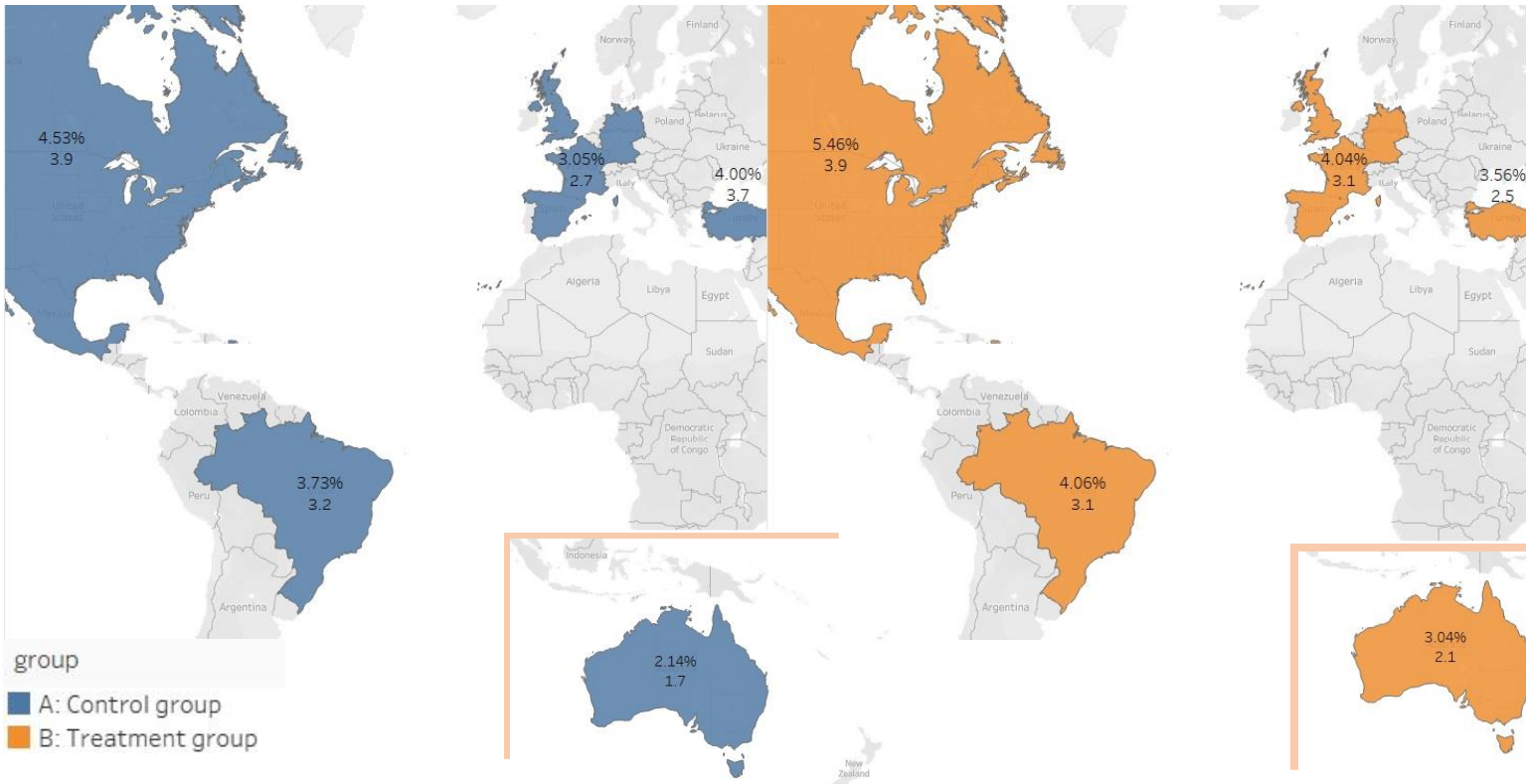
Without 0-gender because it doesn't have many users, then it isn't necessarily meaningful
Here I see that only in AAS for Female-gender has been declining

- **The relationship between the test metrics (conversion rate and average amount spent) and the user's country**



Thus, even though the distribution into groups was carried out automatically and randomly, I saw that the groups are very similar in composition.





There are 5 regions in the experiment with countries:

1. North America: CAN, USA, MEX
2. South America: BRA
3. Europe: GBR, DEU, ESP, FRA
4. West Asia: TUR
5. Australia

For all regions except Turkey, I see an increase in the success metrics. **In Turkey, both success metrics have been declining**

II. Inferential statistic

How we have seen from Descriptive statistic's part there is a difference between Conversion Rate of two groups:

In group A (control) CR(A) = 3.92%

In group B (treatment) CR(B) = 4.63%

So, the difference = 0.71 %

And AAS:

It's time to check how significant the difference between the success metrics values is. And what is the confidence interval for the difference between them. I remind you that the significance level (α) is 0.05.

- **There were used 2 hypotheses in the test.**

1. **Hypothesis 1**: was conducted a hypothesis test to see whether there is a difference in the conversion rate between the two groups. I used the pooled proportion for the standard error.

In the case the Null Hypothesis is "There is no significant difference in conversion rate" or other words the difference equal 0 and the Alternative Hypothesis is "There is a significant difference between two groups".

I performed all the tests in spreadsheets.

For comparing the difference in Conversion Rates, I used Two-sample Z-test (two-tailed) to perform the hypothesis test, because conversion rate is a proportion and I compare two samples

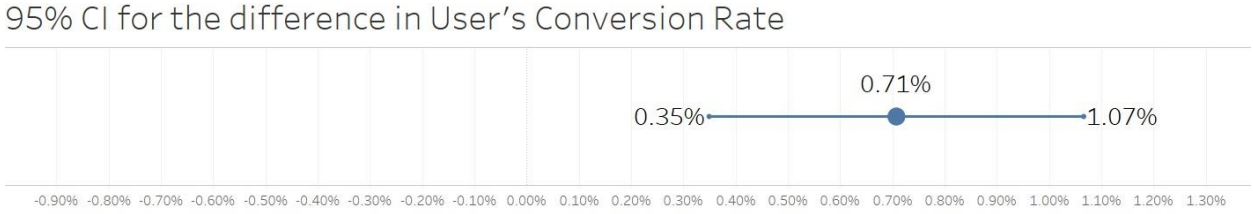
I got these values:

z-test = - 3.86 (Minus because CR(B) larger than CR(A))

p-value = 0.00011

Because 0.00011 < 0.05 I rejected the Null Hypothesis that there is no difference between Conversion Rates for each group. Other words **the difference is significant** for us.

2. I found what is the 95% confidence interval for the difference in the conversion rate between the groups. I used the normal distribution and unpooled proportions for the standard error:



3. Then I conducted a power analysis to ensure that our test is powerful enough to detect that difference:

When I'm choosing Minimum Detectable Effect 10, I need much more sample size, but if to increase this parameter to 11.2% with Statistical Power

= 0.8 I reach my both sample sizes (24343 and 24600) and the CR for B-group (4.63%) is still more than the bound for this MDE (4.39%).

That's mean that my CR for B-group distinguishable from the baseline and make sense when the difference more than 11.2%. So, the sample size is quite enough.

But the Confident Interval is not included in this MDE completely. The range for CR: (4.27%, 5.7%) while the bound for the MDE is 4,39%. Other words in the case only the point estimate (observed difference between the groups in CR) is above the practical significance.

4. **Hypothesis 2:** was conducted a hypothesis test to see whether there is a difference in the average amount spent per user between the two groups.

In the case the Null Hypothesis is "There is no significant difference in mean of spent per user" or other words the difference equal 0 and the Alternative Hypothesis is "There is a significant difference between two groups". Also, $\alpha = 0.05$.

For compare the difference in Average Amount Spent per user I used two-sample T-test (two tailed) for means because AAS is an average of continuous values.

What I got:

t-test =-0.07

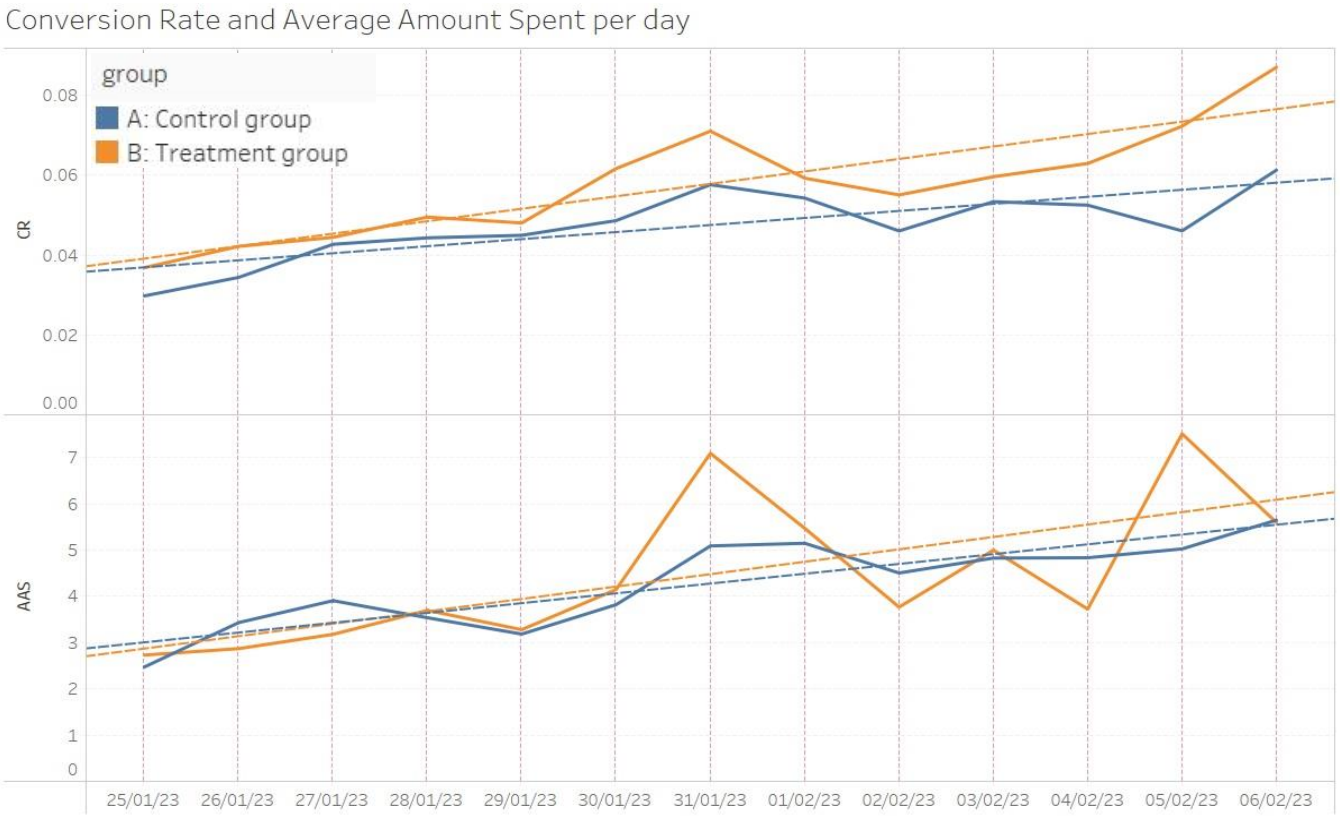
p-value = 0.944

Because 0.944 > 0.05 I failed to reject the Null Hypothesis. Other words **there is no statistically significant difference** between AAS per user in control and treatment groups.

How it looks like with 95% Confidence Interval:

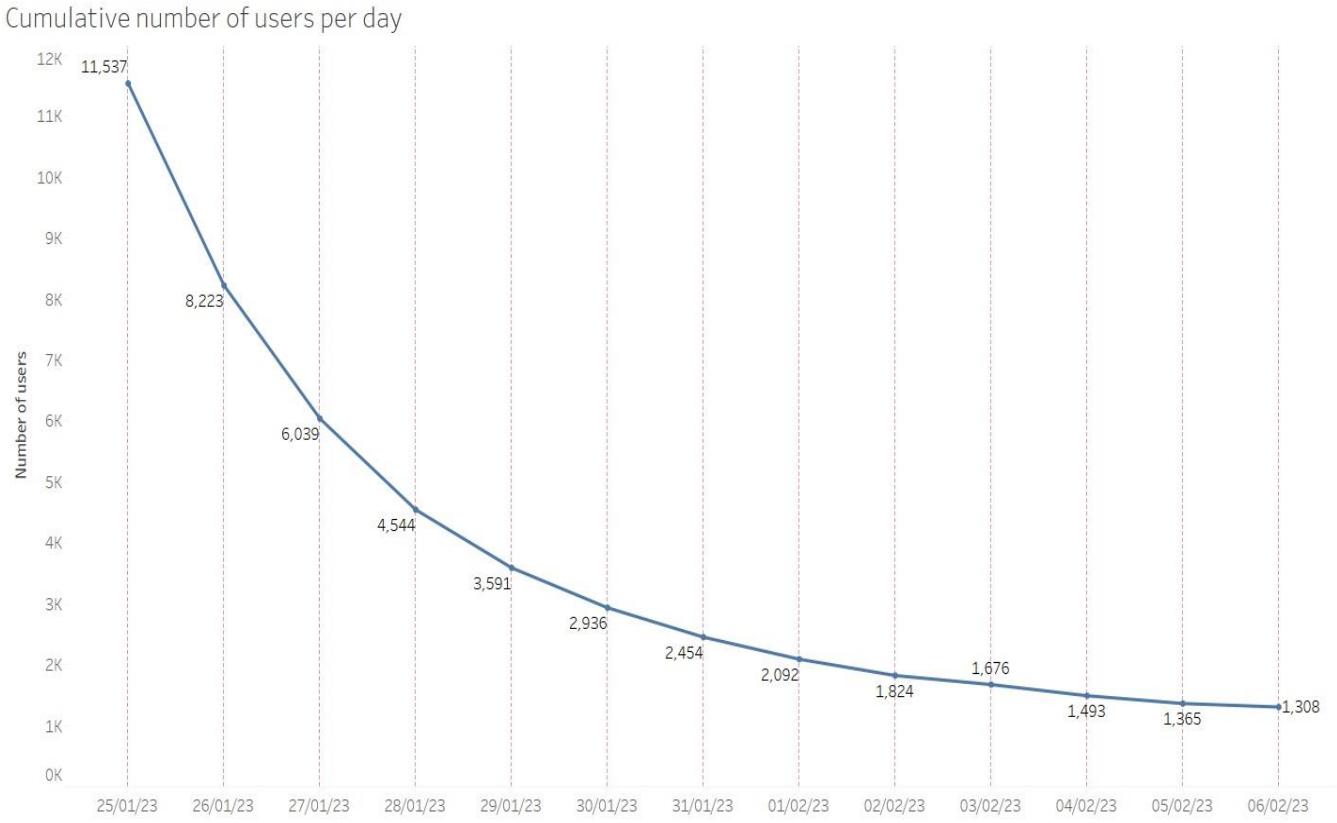
III. Descriptive Statistics for every day.

1) After all I checked what if there is **novelty effect** in the experiment. I wrote another SQL-query and extracted the new one data set with dates. After that in Tableau for each day of the test I computed the success metrics, that's how it looks:



Here I see that there is an increase and decrease in the metrics, but the trend lines have directions of growth. There is no sharp increase at the beginning, but an increase is observed throughout the entire time of observation. **There is no novelty effect.**

2) I checked how the **total number of users** behave **each day**.

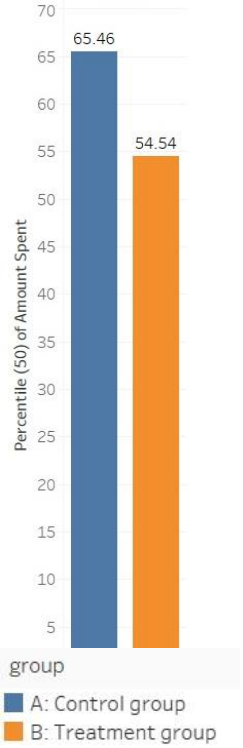


There is decrease of cumulative number of users. With what it can be relative? Maybe advertising campaigns? By the end of the observation period, the intensity of the decline decreases. This shows that if for some reason we want to continue the experiment - we can rely on the number of approximately 1300 users per day.

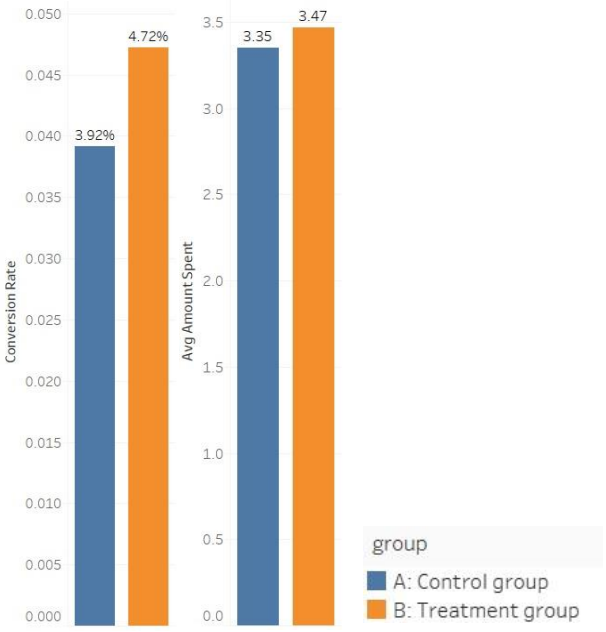
4. Recommendations

My recommendation is **to launch the banner**, because:

1. The difference in conversion rate is statistically significant, but the difference in the average amount spent is not. That's still a good start to getting more paying customers. Maybe we can draw them back to make more purchases later. the median in the group of users who made purchases (conversion is 1) in the experiment group (56.01) is less than in the control group (66.08). Perhaps users made purchases in lower price sections (we just seen that in the distribution of Amount Spent), maybe exactly in the "food and drinks" section. The banner converts a higher percentage of users into customers. And perhaps in the future it will be possible to attract these same customers to purchases in a more expensive product segment. In this case, the increased conversion rate outweighs the lack of impact on the average amount spent, making it reasonable to recommend launching the banner.



2. In all regions, except Turkey, we see an increase in the metrics, or they remain unchanged (AAS). Perhaps due to internal political processes, the earthquake that was on February 6 (exactly the last day of the experiment), in Turkey the picture is not typical for all other regions. But nonetheless its metrics influence the overall picture, because in Turkey more than 1800 users in each group (In USA, the biggest one – 7400). How the success metrics look like without Turkey:



The difference between the metrics is more noticeable.

3. A banner is not typically an expensive feature to launch, in terms of engineering time or operational overhead. So, if we see a significant increase in conversion rate, that's enough justification to go ahead and launch it.

5. Appendix

1. SQL-query: https://docs.google.com/document/d/1fkPu4ve5M8bjVGprggXuEK-t-h55Rw3X/edit?usp=sharing&ouid=101906884593282446689&rtpof=true&sd=true

2. Spreadsheets: https://docs.google.com/spreadsheets/d/13VQtDd1Lix3jKTF1K1Ht9YPXw8ApMSNe5JClW-caaL0/edit?usp=sharing

3. A/B Test Sample Size Calculator - Statsig

4. Another one SQL-query: https://docs.google.com/document/d/1JAo3iC_ij23hYcnK4IOKsA7tmUCHY1ni/edit?usp=sharing&ouid=101906884593282446689&rtpof=true&sd=true