A/B Testing Analysis Report of Globox's New Banner

Report by: Nasreen Sultana Date: 3rd Aug 2023

TABLE OF CONTENT:

SUMMARY:

Based on our analysis, we advise against proceeding with the experiment's launch. Despite a considerable increase in the number of users in the Treatment group (B) compared to the Control group (A), the average amount spent between the two groups was found to be notably low. This suggests that releasing the banner to all users might not lead to a significant benefit in terms of revenue generation.

Considering the perceived cost of launching the feature, it does not seem justifiable based on the test outcomes. It would be prudent to reconsider the approach and explore alternative strategies that could yield more promising results before proceeding with a full release.

BACKGROUND:

GloBox is an online marketplace that has built its reputation among customers for its exclusive range of fashionable boutique items and sophisticated, upscale decor products. Customers can explore our thoughtfully curated assortment of one-of-a-kind products, spanning from exotic spices and elusive teas to meticulously crafted jewelry and textiles, each offering a unique and unmatched experience that they won't encounter elsewhere.

MOTIVATION :

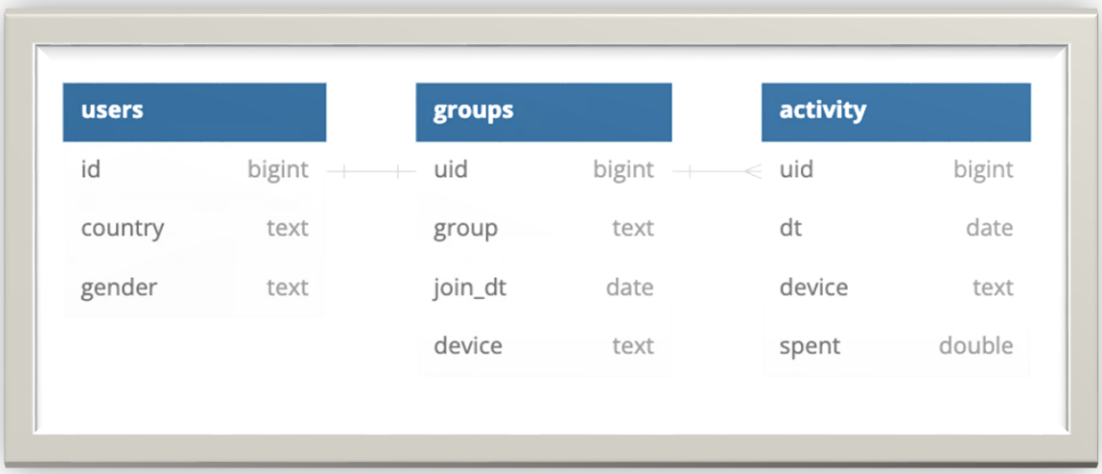
In recent months, their culinary and beverage selections have experienced significant expansion, and the company aims to raise awareness about this product category to boost revenue.

To achieve favorable outcomes, we utilize an A/B test, a technique widely adopted by businesses to experiment with two distinct versions (A and B) of a webpage, advertisement, or product feature to compare their performance. The main objective is to identify which variant performs better in terms of achieving

specific goals. By randomly assigning customers or users to experience either version A or B, businesses can accurately evaluate and determine which option is more successful in meeting their desired objectives. This data-driven approach empowers organizations to make informed decisions and optimize their offerings for better results.

DATA SET UP:

Below is the information of the three tables on which the A/B experiment was done between Jan 25 – Feb 6



A/B Test Setup :

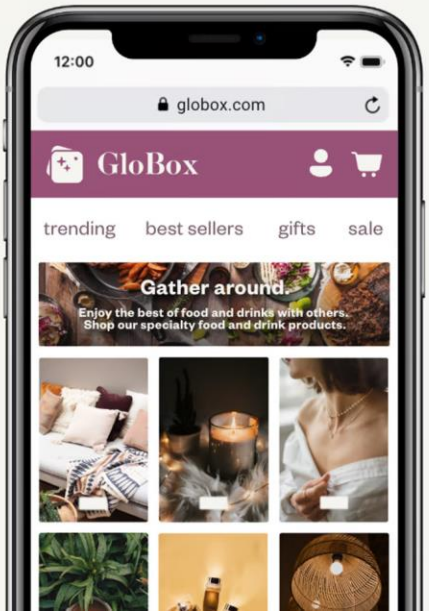
In the A/B experiment, the control group is not exposed to the banner, while the test group is presented with the banner as shown below:

Group A: Control

existing landing page

Group B: Treatment

landing page with food & drink banner



The setup of the A/B test is as follows:

1. The experiment is only being run on the mobile website.
2. 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.
3. The page loads the banner if the user is assigned to the test group and does not load the banner if the user is assigned to the control group.
4. 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". The control group does not see the banner, and the test group sees it as shown below:

PARAMETERS OF THE TEST:

Conversion rate, average amount spent per user

CONTEXT & RESULTS:

Step 1: Data Extraction with SQL

Data extraction with SQL involves retrieving specific information as in the user ID, the user's country, the user's gender, the user's device type, the user's test group, whether or not they converted (spent > $0), and how much they spent in total ($0+).

After extracting the data using SQL, we downloaded it as a CSV file format.

Following is the results from the query:

```sql
SELECT distinct g.uid as userid,COALESCE(u.country,'notknown') as country,COALESCE(u.gender,'notknown') as gender,
COALESCE(g.device,'notknown') as device,g.group,g.join_dt as join_dt,a.dt as purchase_dt,
    COUNT(DISTINCT g.uid) AS total_users,
    COUNT(DISTINCT a.uid) AS converted_users,
    COUNT(DISTINCT g.uid) - COUNT(DISTINCT a.uid) AS non_converted_users,
    CAST(COUNT(DISTINCT a.uid) * 100 / COUNT(DISTINCT g.uid) AS INT) AS conversion_rate,
    COALESCE(SUM(a.spent), 0) AS total_amount_spent
FROM groups g
LEFT JOIN activity a ON g.uid = a.uid
LEFT JOIN users u ON g.uid = u.id
GROUP BY 1,2,3,4,5,6,7
order by total_amount_spent desc
```

Step 2: Calculate and analyze A/B Test Statistics using spreadsheets

After exporting the data as a CSV file, we imported it into a spreadsheet, where we performed two hypothesis tests to analyze the A/B test results.

The **first hypothesis test** conducted was a **two-sample z-test** for a Difference in Proportions. Its purpose was to assess whether there is a statistically significant difference in the conversion rate between the control and test groups.

1. Define the null and alternative hypothesis

Null Hypothesis (H0): NO difference in conversion rate between A (Control) and B (Treatment) group. ie pControl = pTest

Alternative Hypothesis(H1): There will be a difference in conversion rate between A (Control) and B (Treatment) group. ie pControl ≠ pTest

In other words,

Null Hypothesis(H0), p1-p2=p0

Alternative Hypothesis(H1), p1-p2<>p0

Where

p1=conversion rate of group A

p2=conversion rate of group B

2.Evaluate pooled proportion and Standard error

p1=conversion rate of group A		**0.03923**			
p2=conversion rate of group B		**0.04630**			
n1=total users in group A		**24343**			
n1=total users in group B		**24600**			
pooled proportion (p^)		**(p1*n1+p2*n2)/(n1+n2)**		**0.04278**	
Standard Error (SE)		**√p^(1-p^) (1/n1+1/n2)**		**0.0018295**	

3.Calculating Test Statistic

p0 in this formula is the true population proportion, which is 0 (null value)				**0**

(treatment - control) = (p2-p1)				**0.00707**
SE				**0.0018295**
Test Statistic	**(p1 - p2) -p0/ Standard Error**			**3.86429177**

4.Calculating p-value and Confidence intervals

p-value = 0.0001114

For a 95% confidence level						
Alpha α		=	**0.05**			
the critical value is approximately		=	**1.96**			
Sample statistic		=	**0.00707**			
Standard Error	**(SE)**	**=**	**0.0018295**			
The margin of error (MOE) for the difference in proportions:			**= critical value * SE**			
			≈ 1.96*0.00182			
			0.003585			
Finally, we can construct the confidence interval:						
95% Confidence Interval = (p2 - p1) ± MOE						

95% Confidence Interval		**0.01065**		**Upper Bond = Sample Stats + Margin of Error**		
		0.00348		**Lower Bond = Sample Stat - Margin of Error**		

5.Conclusion:

Since the p-value is much smaller than the significance level (α = 0.05), **we reject the null hypothesis (H0)**. There is strong evidence to suggest that there is a statistically significant difference in the conversion rates between Group A (Control group)and Group B(Treatment group). The conversion rate for one group is significantly different from the other group.

The **second hypothesis test** conducted was a **two-sample t-test** for a Difference in Means. Its purpose was to assess whether there is a statistically significant difference in the average amount spent per user between the two groups.

1. Define the null and alternative hypothesis

Null Hypothesis (H0): NO difference in average amount spent between A (Control) and B (Treatment) group. ie pControl = pTest

Alternative Hypothesis(H1): There will be a difference in average amount spent between A (Control) and B (Treatment) group. ie pControl ≠ pTest

In other words,

Null Hypothesis(H0), p1-p2=p0

Alternative Hypothesis(H1), p1-p2<>p0

2.Calculating Standard error

Control Users - A		Test/Treatment Users - B	
Mean (x̄1)	3.375	Mean (x̄2)	3.391
Std Dev (s1)	25.94	Std Dev (s2)	25.414
n1	955	n2	1139
Mean difference(x̄1- x̄2)	=.016		
Standard Error	(SE)	0.232141	

3. Calculating Test Statistic

Test Statistic (T)	0.0144987

4. Calculating p-value and Confidence intervals

the p-value for a t-distribution:		0.988435

For a 95% confidence level											
Alpha α			0.05								
Degrees of Freedom DF =min (N1 - 1) , (N2 - 1)) = 955-1			24342			(since between the two number of users group A has less users compared to group B)					
(degrees of freedom is the minimum n-1 between the groups)											

Using a t-table or statistical software, the critical value for a 95% confidence level and 954 degrees of freedom is approximately			**1.961**								
The margin of error (MOE) for the difference in proportions:			**=' critical value * SE**								
			0.455								
Finally, we can construct the confidence interval:											
95% Confidence Interval = (x̄2 - x̄1) ± MOE											
95% Confidence Interval			**-0.439**		**Lower Bond = Sample Stat - Margin of Error**						
			0.471		**Upper Bond = Sample Stats + Margin of Error**						

5.Conclusion

Since p-value > α, i.e.0988 > 0.05 ,H0 cannot be rejected.

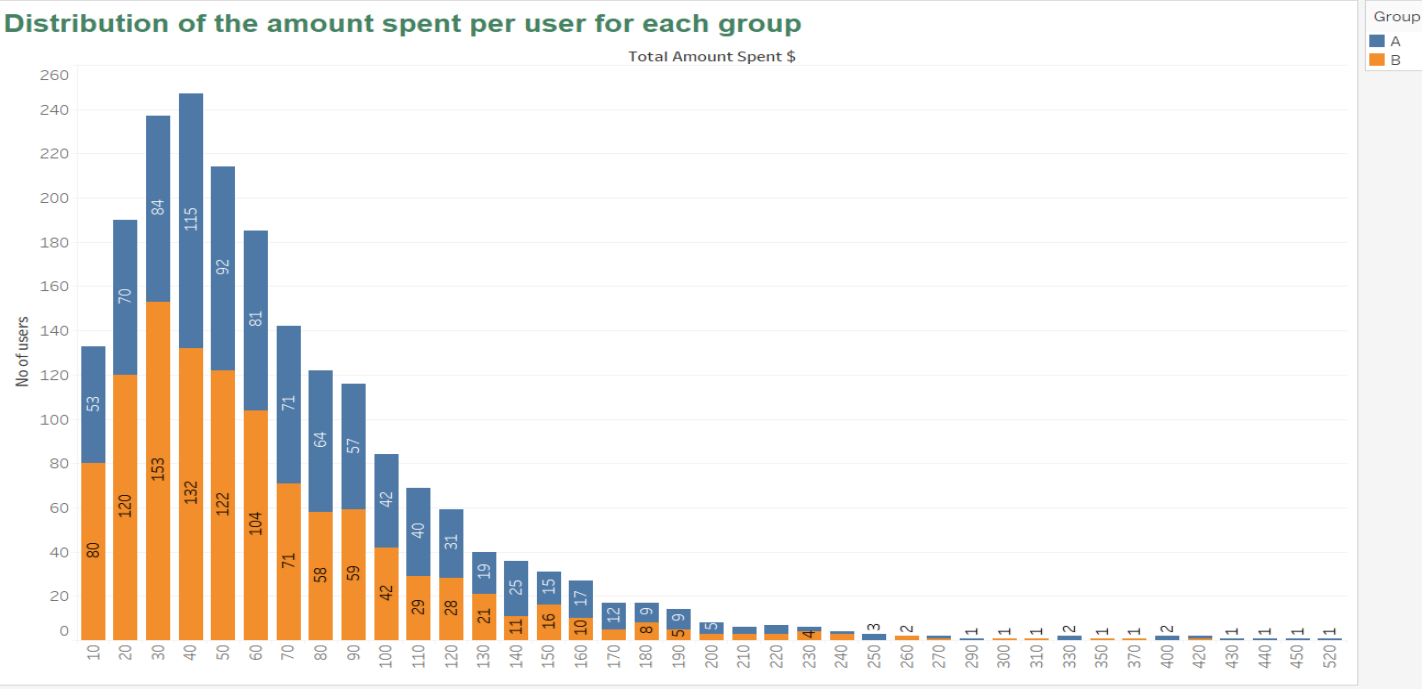
Therefore, we fail to reject the null hypothesis, which states that the average amount spent per user in Group A is equal to the average amount spent per user in Group B.

Step 3: Visualize the Results in Tableau

By leveraging Tableau, a powerful data visualization tool, we delved deeper into user behavior and spending patterns. Through additional analysis, we explored user behavior across various dimensions, including device usage, gender, and region. The visualizations created using Tableau, such as interactive charts, graphs, and dashboards, facilitated a comprehensive understanding of user interactions and spending habits, enabling us to make more informed decisions and optimize the platform's performance to enhance user experiences and drive revenue growth.

Here are some insights we observed:

Insight 1: Distribution of average amount spent per user for each group

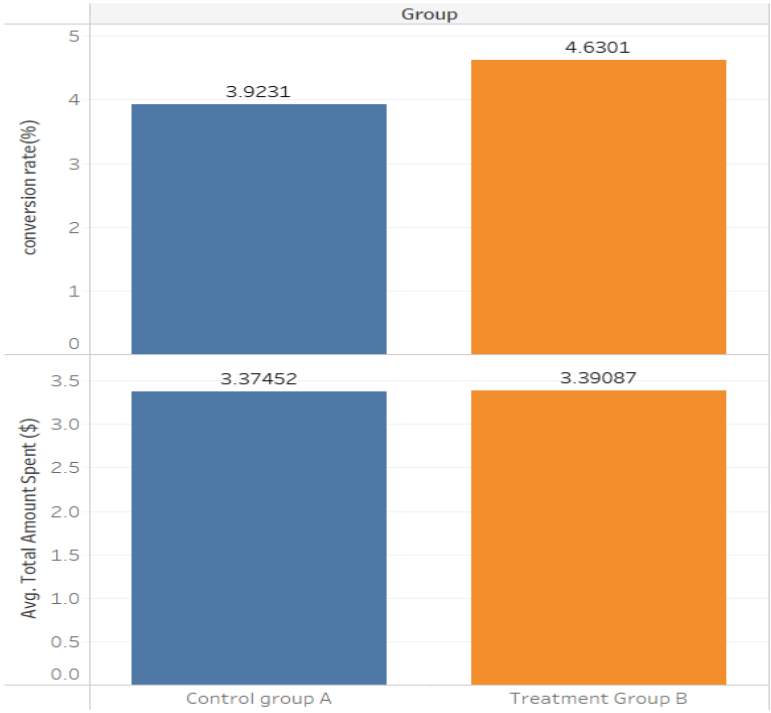


• A significant proportion of users in the Control group, specifically around 115 individuals, spent an amount ranging from $40 to $50

• A significant proportion of users in the Treatment group, specifically around 153 individuals, spent an amount ranging from $30 to $40

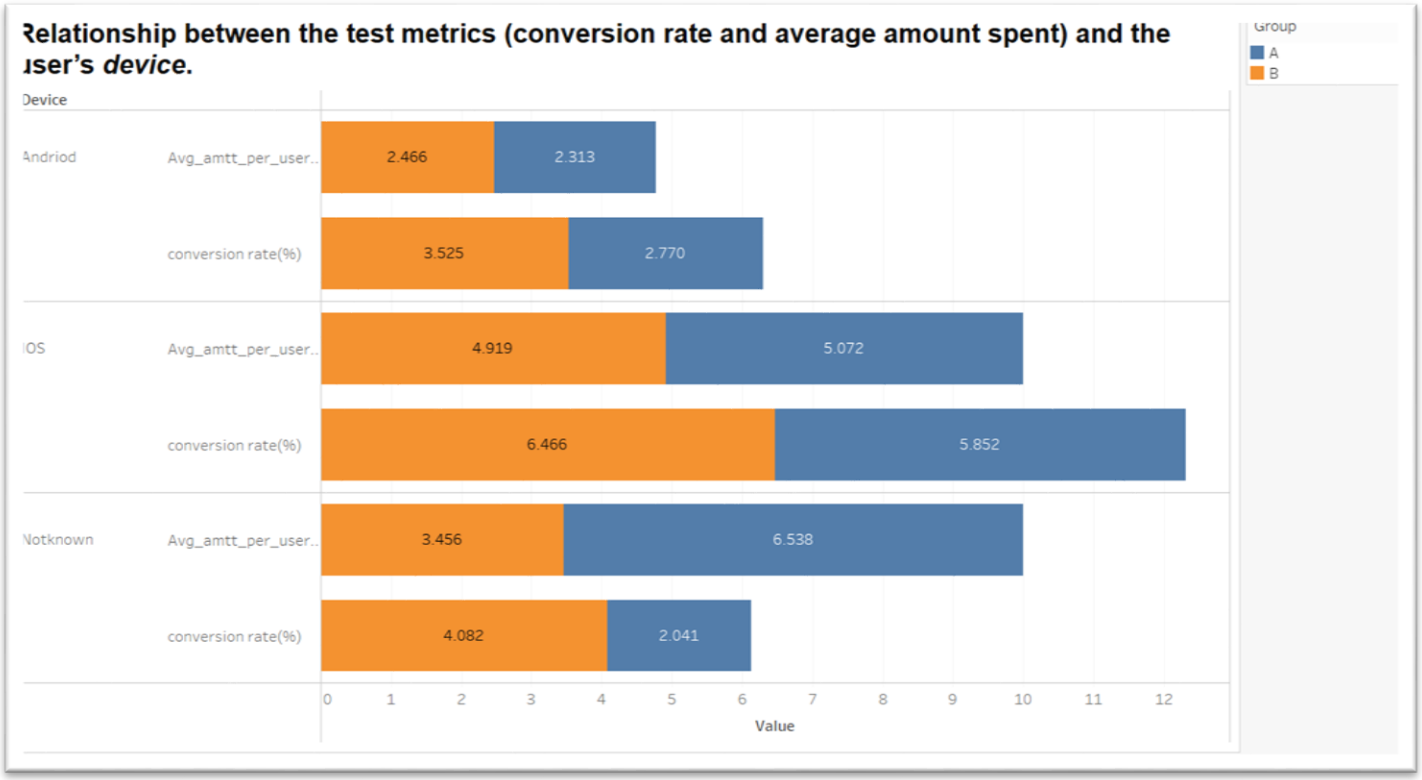
Insight 2: Comparing the conversion rate and average amount spent between the test groups.

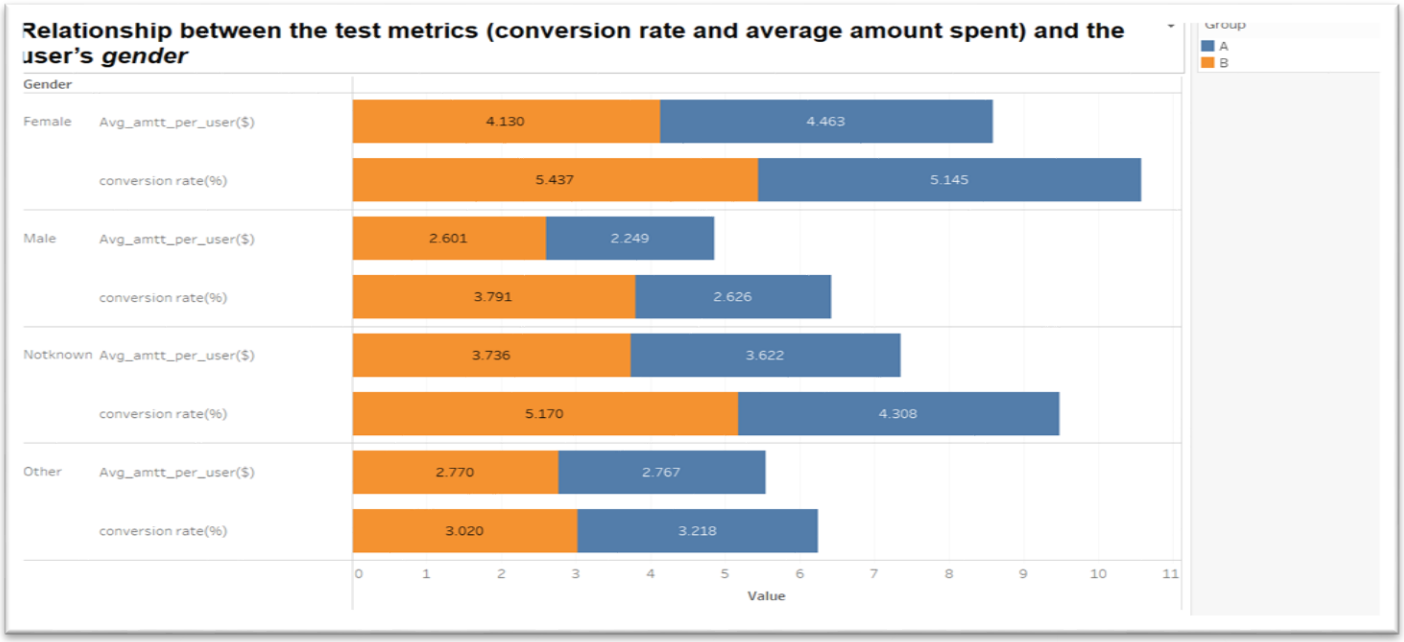
(Q1)Comparing the conversion rate and average amount spent between the test groups.

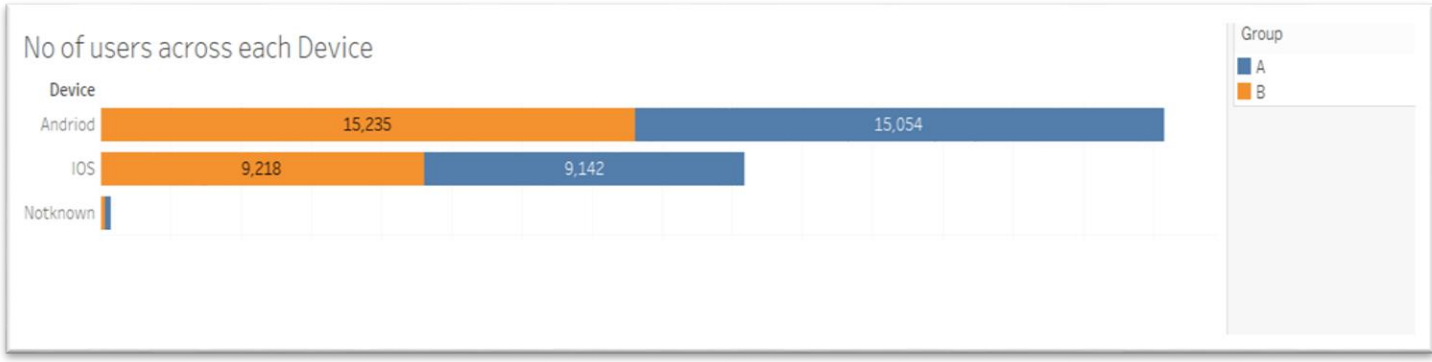


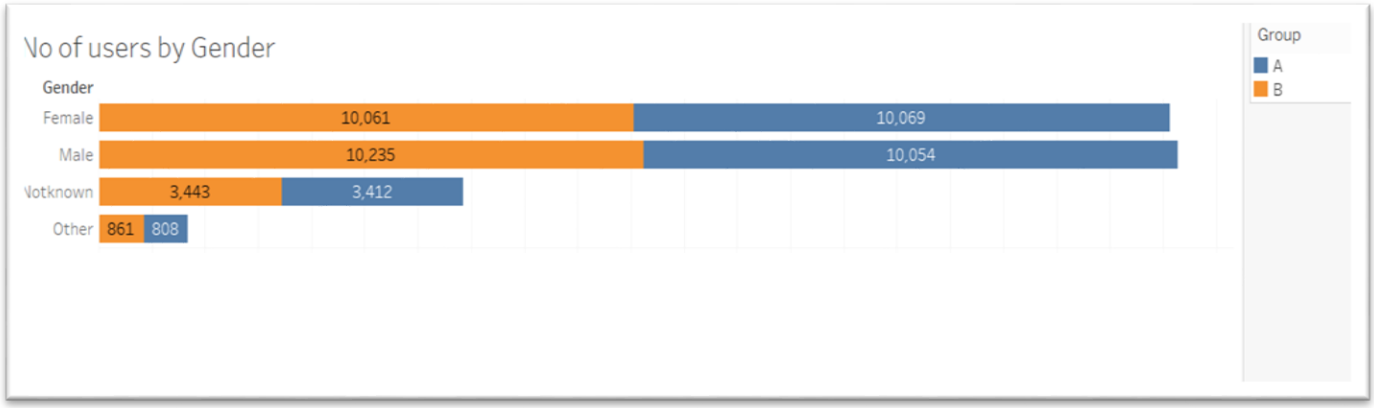
• The difference in conversion rates and average spending between the Treatment group (B) and Control group (A) could be attributed to the new banner's focus on food and drinks, which deviates from Globox's usual high-end and luxury items. As a result, the Treatment group may have a higher conversion rate due to the appeal of food and drinks, but their average spending might be lower as these items tend to have lower price points. On the other hand, the Control group, with a focus on high-valued items, may have fewer conversions but higher average spending per individual, aligning with Globox's typical customer preferences for luxury products.

Insight 3: Relationship between the test metrics (conversion rate and average amount spent) and the user’s device and gender


Relationship between the test metrics (conversion rate and average amount spent) and the user's device.
Group
A
B
Device
Andriod
IOS
Notknown
Avg_amtt_per_user..
conversion rate(%)
2.466
2.313
3.525
2.770
4.919
5.072
6.466
5.852
3.456
6.538
4.082
2.041
0 1 2 3 4 5 6 7 8 9 10 11 12
Value

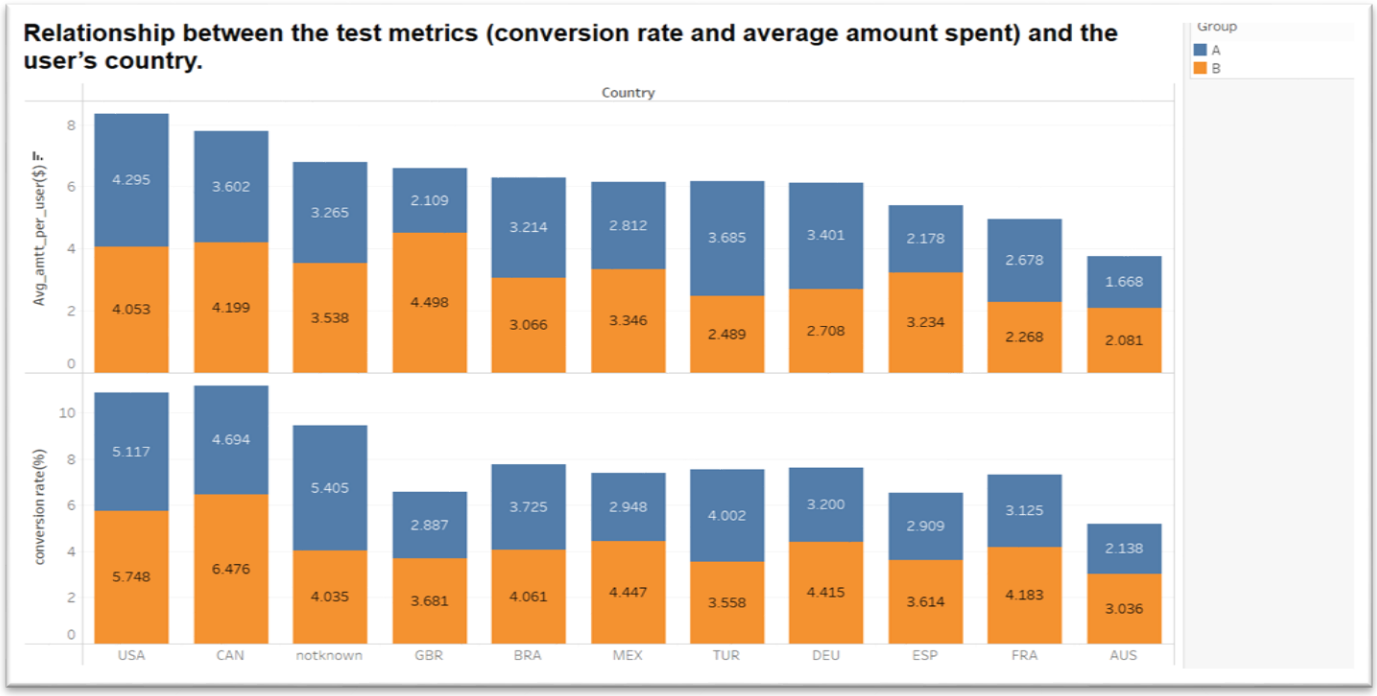

Relationship between the test metrics (conversion rate and average amount spent) and the user's gender
Group
A
B
Gender
Female
Male
Notknown
Other
Avg_amtt_per_user($)
conversion rate(%)
4.130
4.463
5.437
5.145
2.601
2.249
3.791
2.626
3.736
3.622
5.170
4.308
2.770
2.767
3.020
3.218
0 1 2 3 4 5 6 7 8 9 10 11
Value





• In the Treatment group (B), we observed consistent improvements in the number of users and conversion rate across various user devices compared to the Control group (A). However, despite these positive trends, there was no statistically significant increase in the amount spent by the users. This suggests that while the banner may have positively impacted user engagement and conversion rates across devices, it did not lead to a notable increase in the average spending behavior of users in the Treatment group.

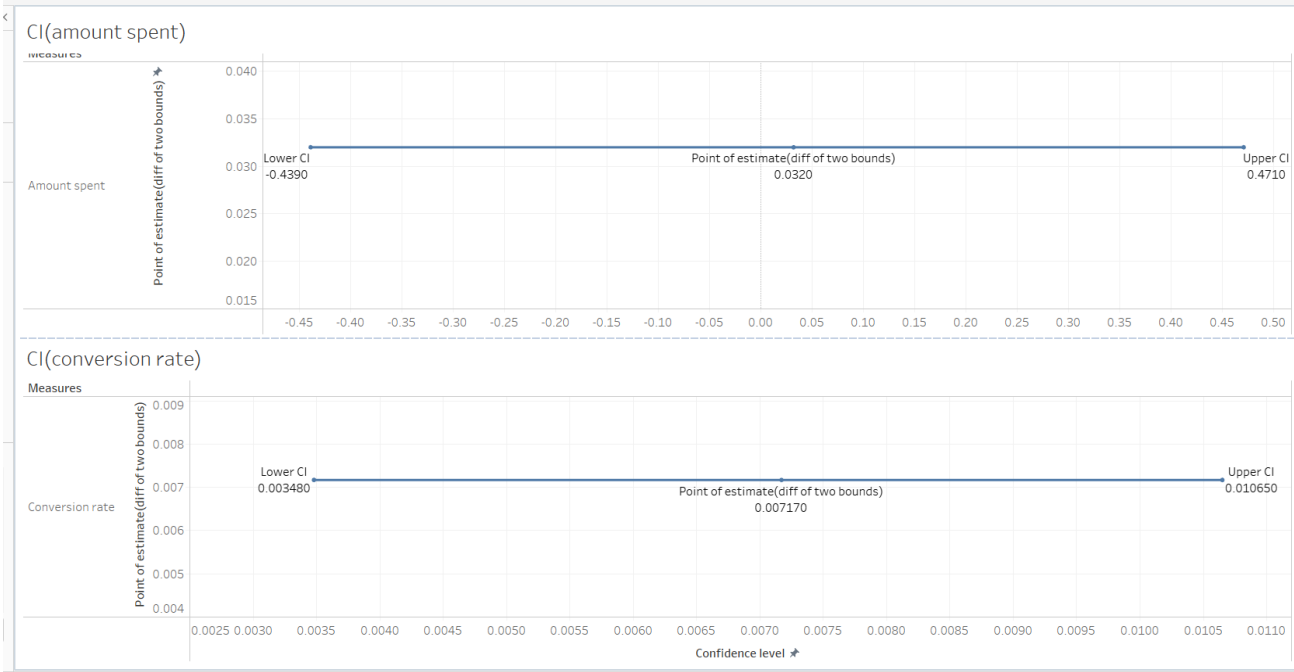
Insight 4: the relationship between the test metrics (conversion rate and average amount spent) and the user's country



• While Treatment group B consistently exhibits a higher conversion rate compared to Control group A, there is a notable and significant improvement in both conversion rate and amount spent per user in the European region.

• However, despite these positive developments in Europe, the overall revenue across all continents does not show promising improvement.

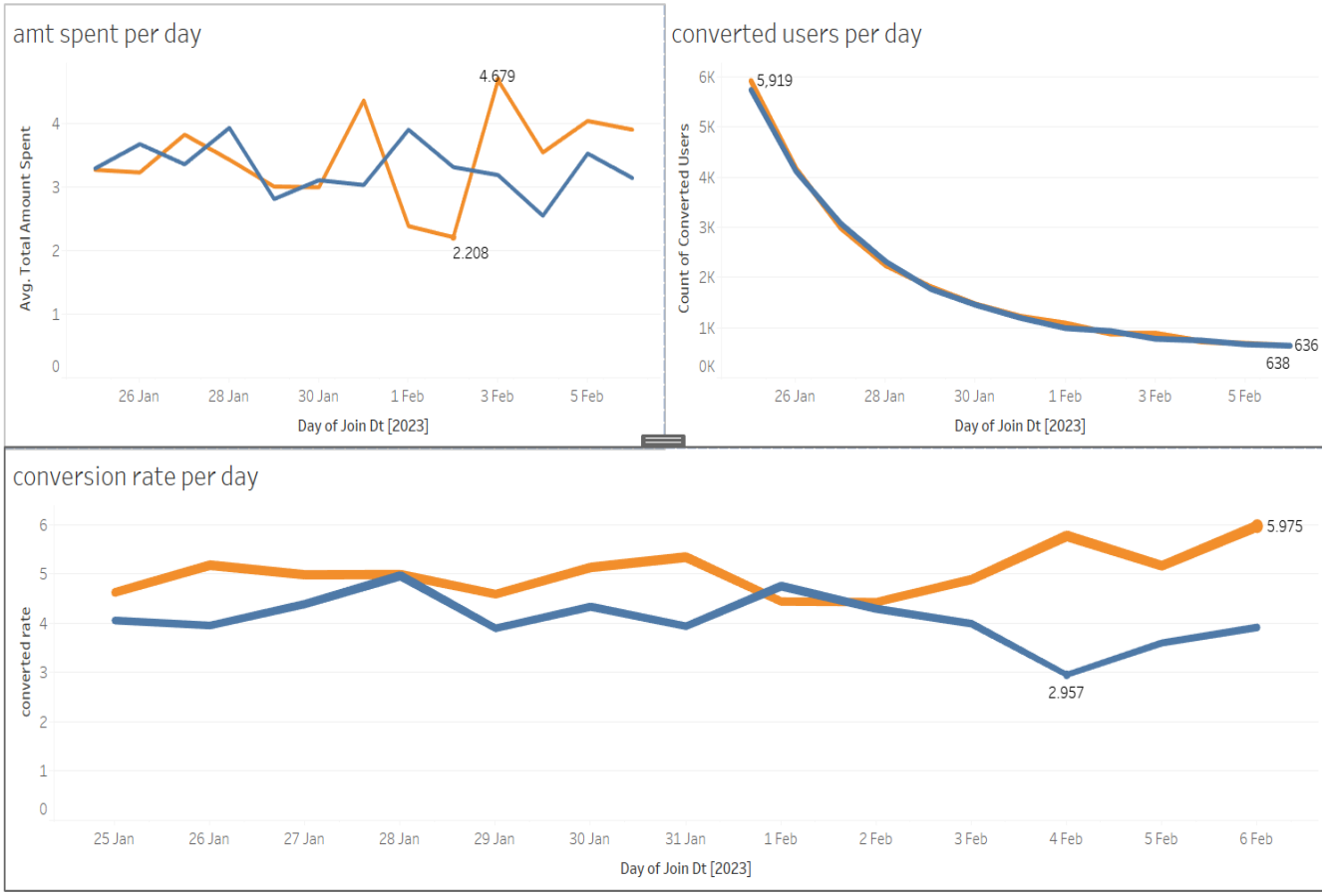
Insight 5: Visualize of 95% Confidence Intervals



• Based on the visualization of the 95% confidence intervals, we can be highly confident that the new banner has a significant impact on the conversion rate, as evidenced by Treatment group B consistently outperforming Control group A. The data shows that the difference in conversion rates between the two groups is estimated to be around 0.007, with a range of 0.0035 to 0.0107, indicating a clear positive effect.

•On the other hand, when it comes to the average amount spent per user, the 95% confidence interval indicates that the difference between the two groups falls within a broader range, from -0.439 to 0.471, with an estimated point of 0.016. This suggests that there is no significant difference between the two groups regarding the average amount customers spend per order. In other words, while the new banner effectively boosts conversions, it may not have a significant impact on the average amount customers spend per transaction.

Insight 6: Check for Novelty Effects



•To check novelty effect, we rewrote the query with joining ang purchase date

Upon analyzing the difference in the average amount spent and conversion rate of all users between the groups over time, we can say that the effectiveness of the banner is short-lived, contrary to the novelty effect. While the banner did show positive results during the initial novelty period, like no of users were higher at the beginning of the banner launch but its impact diminishes significantly over time.

• While the amount of money spent shows some fluctuations, the differences between the Control and Treatment groups are not consistently significant enough to indicate a novelty effect.

Insight 7: Power Analysis

Sample Size Calculator

Calculate how many samples you need to properly power your experiment

Baseline Conversion Rate (%)	3.92
Minimum Detectable Effect (%)	18
A/B Split Ratio	0.5
Significance (α)	0.05
Statistical Power (1 - β)	0.8
Results	
TEST SIZE	**CONTROL SIZE**
11.9k	**11.9k**
TOTAL SAMPLE SIZE	
23.8k	

Expected Difference between Means

Calculate how many samples you need to properly power your experiment

Difference between Two Means:	0.016
Expected Standard Deviation:	25.94
Significance (α)	0.05
Statistical Power (1 - β)	0.8
Alternate Hypothesis	Two sided
A/B Split Ratio	0.5
Results	
TOTAL SAMPLE SIZE	
82,521,602	

• To ensure we could detect our desired minimum effect size and achieve appropriate statistical power, we conducted a power analysis to determine the required sample size. The results showed that the sample size was sufficient to detect differences in the conversion rate metric effectively.

• However, for the average amount spent per user metric test, the initial sample size was found to be inadequate in providing enough sensitivity to detect meaningful differences accurately. To address this limitation, we plan to repeat the test on a larger scale. By increasing the sample size, we aim to improve the reliability and precision of our findings regarding the average amount spent per user, ensuring a more robust and comprehensive analysis.

RECOMMENDATION

We advise against launching the new banner currently. Although the analysis indicates a higher conversion rate, the fact that the average amount spent per user did not increase significantly is a cause for concern. The focus on food and drink products, which generate lower revenue per order compared to your typical boutique fashion and high-end offerings, might lead to lower overall revenue despite the higher conversion rate.

While the higher conversion rate is encouraging, it does not guarantee sustained revenue growth. The potential for future purchases relies on engaging customers effectively, and the current data does not show substantial evidence of this potential. Moreover, additional testing and analysis of user behavior across different devices, genders, and regions would be necessary to validate the effectiveness of the banner in attracting and retaining paying customers.

Launching the banner without a clear strategy to address the discrepancy in revenue per order and without solid evidence of sustained customer engagement may not yield the desired results. It is essential to thoroughly assess the potential risks and drawbacks associated with the banner's launch, especially given the importance of maximizing revenue and customer value.

APPENDIX

1.SQL Queries

The initial phase of the analysis involved formulating queries to extract relevant data for test statistics and preparing it for use in Google Spreadsheets.

Here is the link to all the queries

2.Spreadsheets

We used spreadsheet to extract data to perform hypothesis testing

Here is the the link

3.Tableau

The final step was to visualize the data using Tableau, and all the insights and findings can be explored in Tableau accessible through the provided link below.

Here is the link

Here is the link for visualization of Confidence level

Here is the link for visualization of Novelty Effect