GloBox

A/B Testing Analysis Report

25 Aug 2023

By MEKDES ASFAW

DA106 Mastery Project

Summary

GloBox is an online market place that specializes in sourcing unique and high-quality product from around the world. GloBox is primarily known amongst its customer base for boutique fashion items and high-end décor 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. The company decided to run a test that highlights key products in the food and drink category as a banner at the top of the website.

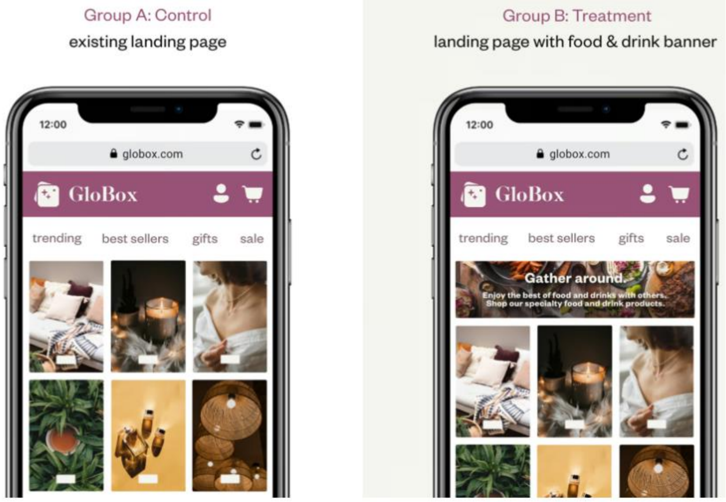
However, the result of the test shows that there is no substantial difference on the company revenue due to the banner on the top of the website. Therefore, we don’t recommend to launch the new website with the banner.

Context:

A/B testing method is used to carry out the test with two groups. First group, GroupA, is the control group which does not see the banner, and the second group, GroupB, is the group which sees the website with the banner.

The test is run for 11 days in the last week of Jan’23 and first week of Feb’23. The test only run on mobile website platform. There are a total of 48,943 users tested, the group randomly assigned to either A or B groups, with 24,343 and 24,600 sample size respectively.

GloBox team has set the metrics of the test which is the conversion rate and average spent per user.


Group A: Control
existing landing page
Group B: Treatment
landing page with food & drink banner
12:00
globox.com
GloBox
trending
best sellers
gifts
sale
Gather around.
Enjoy the best of food and drinks with others.
Shop our specialty food and drink products.

A/B Test Analysis Result

1. Test of Conversation Rate

Conversion rate is one of the metrics to be measured. And we use hypothesis testing to see if there is a difference in the conversion rate between the two groups.

We used hypothesis testing with Ho is True,
Ho -which means the banner has no effect in the conversion rate.
H1 -is the banner has effect on the conversion rate.
Significant value (α) – is 0.05(5%) Confidence level is 95%

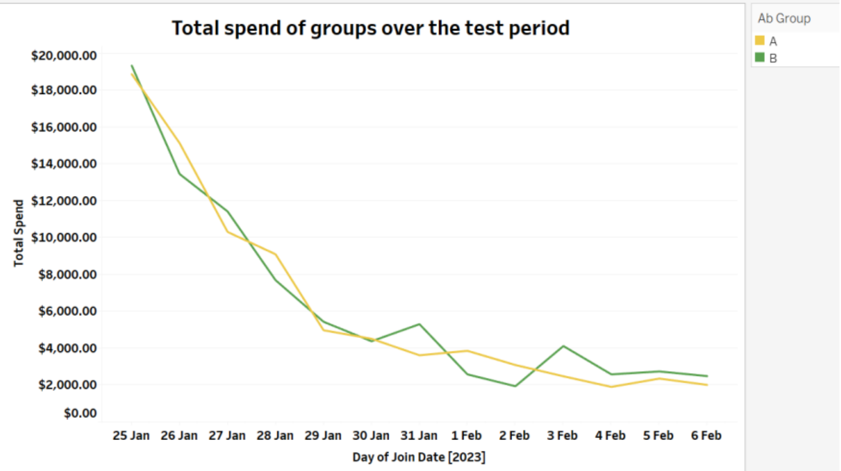
This test is a two-sided test because we check the banner has weather positive or negative effect and use normal distribution Z-Test. Since this A/B test has two groups, this is a two-sample test.

Basic findings from GloBox Database:

As we see in the table and time series graph below. The conversion rate change between Group A and B is 0.7% (from 3.92% to 4.63%). This is an 18% conversation rate lift. In addition, the total spend trend is relatively similar showing higher spend in the first week of the test and lower spend in the second week of the test for both groups. There doesn't seem noticeable big difference in the total spend or in the average spend per user.
(Detailed SQL query results can be found in Appendix: A.)

Test Group	No of group(n)	*No of converted*	Proportion	Conversion Rate %	Total spent($)	Average Spent($)
A	24,343	954	0.039	3.92%	82,145.9	3.37
B	24,600	1,140	0.046	4.63%	83,402.9	3.39
Total	48,943	2,094	0.043	4.28%	165,548.8	3.38

Total spent of groups from 25th Jan'23 to 06 Feb'23

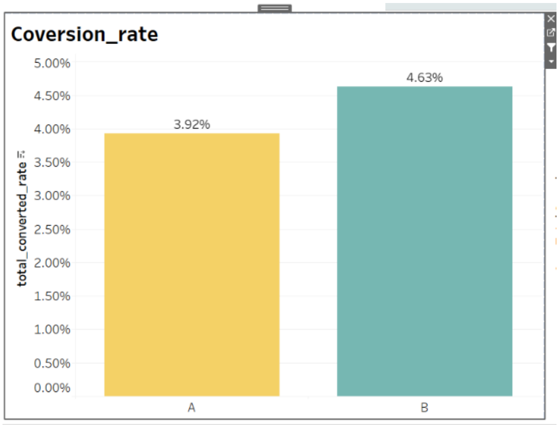


➢ Calculation result of the Test statistic and P-Value

Calculation	Notation	Value
Sample size Group A(Control)	n1	24,343
Sample size Group B(Treatment)	n2	24,600
Sample proportion(Control)	p1	0.0392
Sample proportion(Treatment)	p2	0.0463
Pooled proportion	p hat	0.0428
Test Statistic	T	3.8801
P-Value	Pval	0.0001

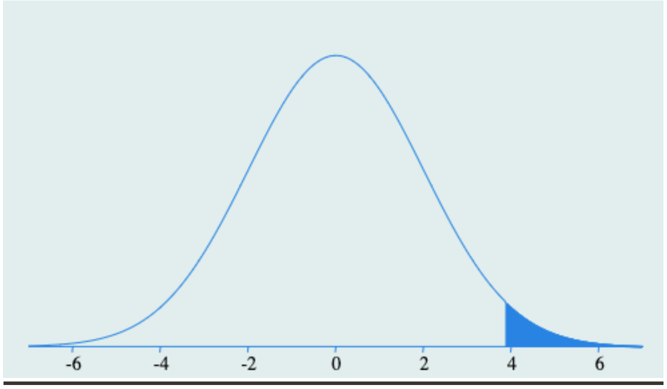
(Detailed calculation to arrive the above figures can be found in Appendix: B)

Comparison of conversion rate per user between groups



Conclusion: With a p-value = 0.0001 < 0.05, we reject the null hypothesis that the conversion rate is the same between the two groups. We are in favour of the alternative hypothesis that there is a difference on conversion rate between the two groups.

The following figure shows the rejected area on the normal probability distribution graph**.**



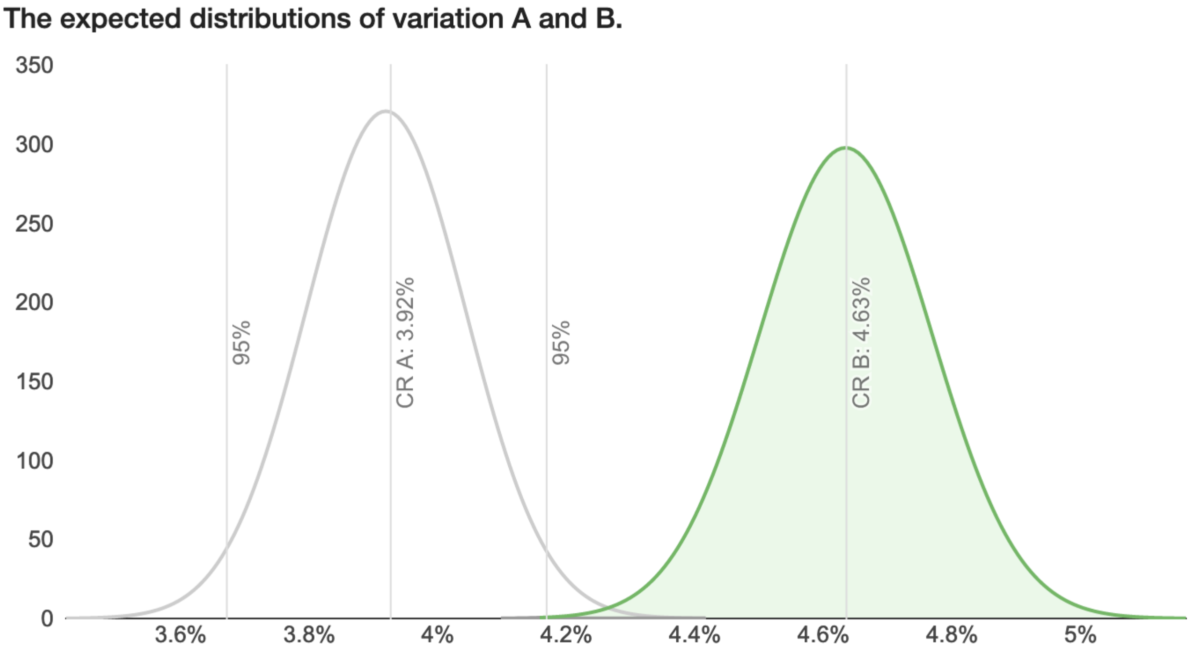
Graph output is used from online calculator: **https://www.hackmath.net/en/calculator/normal-distribution?mean=0&sd=2&area=above&above=3.88&below=&ll=&ul=&outsideLL=&outsideUL=&draw=Calculate**

- **Calculation result of confidence interval**

Calculation	Notation	Value
Sample size Group A(Control)	n1	24,343
Sample size Group B(Treatment)	n2	24,600
Sample proportion(Control)	p1 hat	0.039
Sample proportion(Treatment)	p2 hat	0.046
Sample statistic/Point estimate	stat	0.007
Standard Error	SE	0.0018
Critical Value	z	1.9600
Margin of error	moe	0.0036
Lower bound	CI lower	0.0034
Upper bound	CI upper	0.0106

(Detailed calculation to arrive the above figures can be found in Appendix: C)

Therefore, we are 95% confident that the conversion rate difference will be above 0.0034 and below 0.0106. This means that, we are 95% confident that the conversion rate of Group B ranges from 4.24% to 4.96%



Graph output is used from online calculator: https://abtestguide.com/calc/

Power analysis – Conversion Rate

Let’s first check the conversion rate practical significance using the following given variables β=20%, i.e. the statistical power (1-β) will be 80%, and Alpha=0.05, and with GroupA sample size ratio to Group B sample size ratio of ~ 0.5 in to the statsig.com calculator.

Selecting Minimum Detectable Effect(MDE) of equals 10%, the total sample size required is around 60,000 but our sample size is 48,943 which is a little bit lower. As per the above conversion ratio calculation we have seen a 0.7% conversion ratio change which is 18% relative change from the controlled group due to the addition of the banner.

This relationship shows us that when the sample size decreases the detectable effect increases. By using the calculator to find out what will be the MDE, with 48,000 sample size, we find MDE value of 11.25%.

This show us that our 18% result is higher than the MDE we can get from our the test sample size. Therefore, we can say with statistical power of 80% confidence that this change is practically significant.

2.Test of Average amount Spend

Average amount spend per user is the second key metric that we need to test.

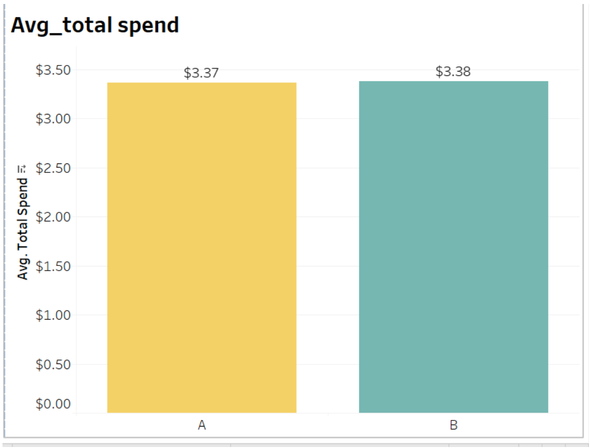
We used hypothesis testing with Ho is True,
Ho -which means the banner has no effect in the average amount spent per user.
H1 - means the banner has effect on the average amount spent per user.
Significant value (α) – is 0.05(5%)
Confidence level is 95%
Since values are mean we will use t distribution test to calculate the test statistics and p-values.

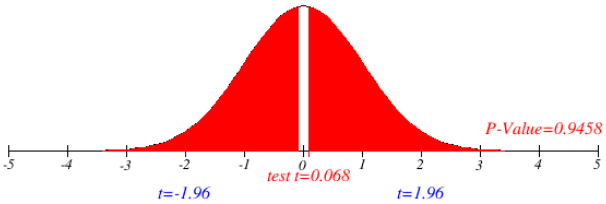
- **Calculation result of the Test statistic and P-Value**

Calculation	Notation	Value
Sample size Group A(Control)	n1	24,343
Sample size Group B(Treatment)	n2	24,600
Average_spent(control)	x1	3.37
Average_spent(treatment)	x2	3.39
Standard devation(control)	s1	25.94
Standard devation(treatment)	s2	25.41
Standard error	s	0.232
Test statistic	T	0.068
Degree of freedom	df	24,342
P-value	pval	0.946

(Detailed calculation to arrive the above figures can be found in Appendix D)

Average spent per user between groups





Graph output is used from online calculator: http://www.imathas.com/stattools/norm.html

Conclusion: With a p-value 0.946 > 0.05 we fail to reject the null hypothesis that the average amount the users spent in both groups is similar or there is no significant change on average spending.

- **Calculation result of confidence interval**

Calculation	Notation	Value
Sample size Group A(Control)	n1	24,343
Sample size Group B(Treatment)	n2	24,600
Sample average(control)	x1	3.37
Sample average(treatment)	x2	3.39
Standard deviation(control)	s1	25.94
Standard deviation(treatment)	s2	25.41
Standard error	s	0.232
Sample statistic /point estimate	stat (x2-x1)	0.016
Degree of freedom	df	24,342
Test statistic	T	0.068
Critical value	t	1.960
Margin of error	moe	0.455
Lower bound	lower	-0.439
Upper bound	upper	0.471

(Detailed calculation to arrive the above figures can be found in Appendix: E)

Therefore, we are 95% confident that the difference that the user average spent before and after the banner is between (-0.439 to 0.471). Using this range we say we are 95% confident that the average spent per user in the Group B will range from $2.94 to $3.85. The change on customer spending is inconsiderable, it is has changed only 0.016. Hence, we could say that the banner does not have substantial impact on revenue.

Power analysis – Average Spent

Testing the deference of average user's spending and the significance level using the following given variables: β=20%, that means the statistical power (1-β) will be 80%, and Alpha=0.05, and with GroupA sample size ratio to Group B sample size ratio of ~ 0.5.

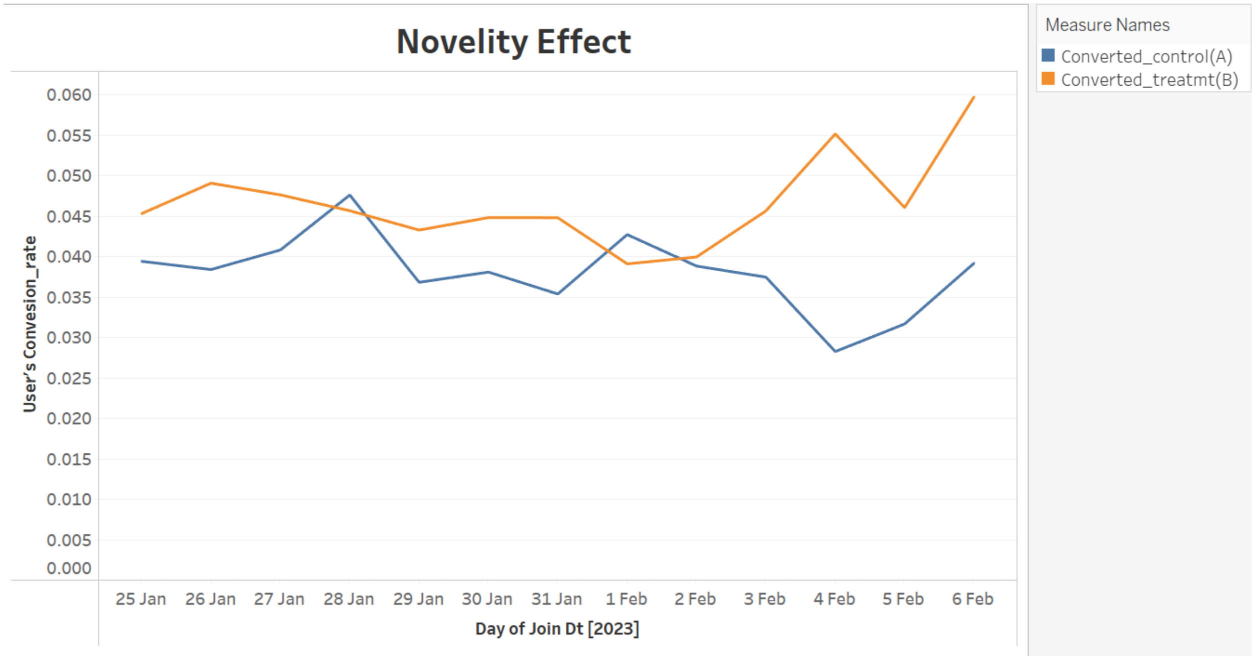
By using the online calculator, we would find the following result:

With a pooled standard deviation of 25.67 units, the study would require a sample size of 41,435,753 for each group (i.e. Total sample size of 82,871,506, 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 ~ 0.02 (i.e.3.39 - 3.37) units.

However, in our case with 48,943 sample size, we should have (~ 0.65) minimum detectable effect (MDE) but what we have got on the test (~ 0.02) which is lower than expected and it is inconsiderable to statistical power of 80%. This result drive us to conclude not to launch the banner as it is practically insignificant

Novelty Effects

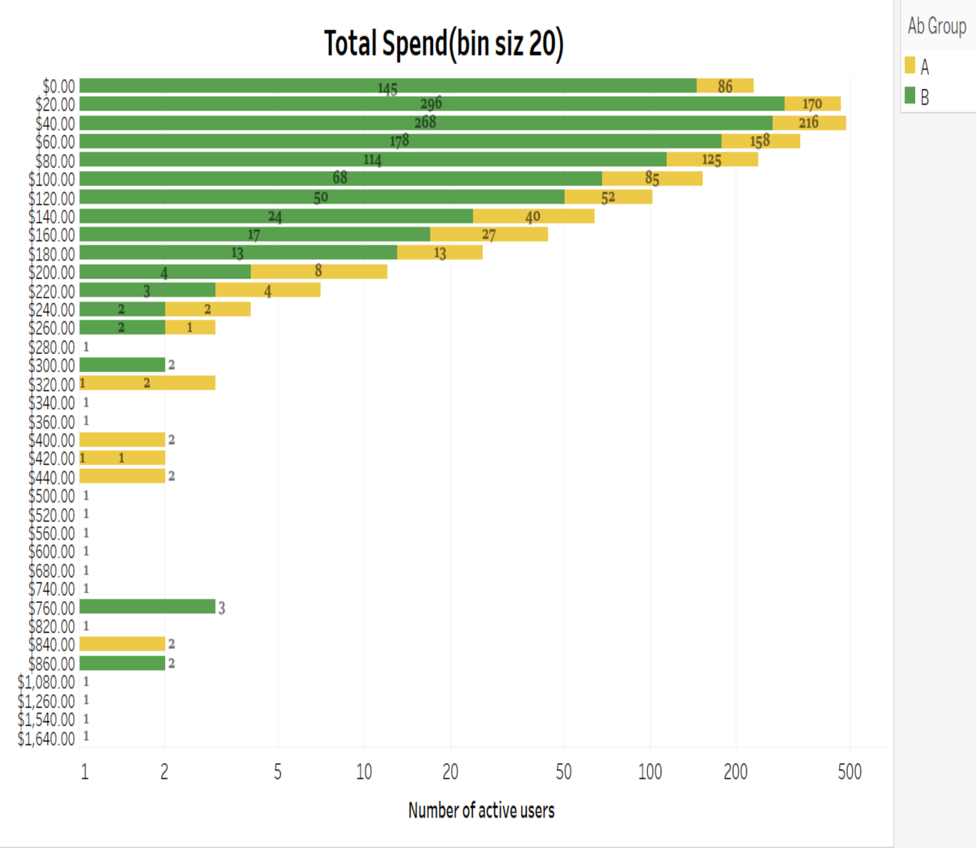
The graph shows nearly the same curve on the conversion rate over the first week of the test, and then it picks up towards the last week of the test. This shows that there is no obvious novelty effect for the test.



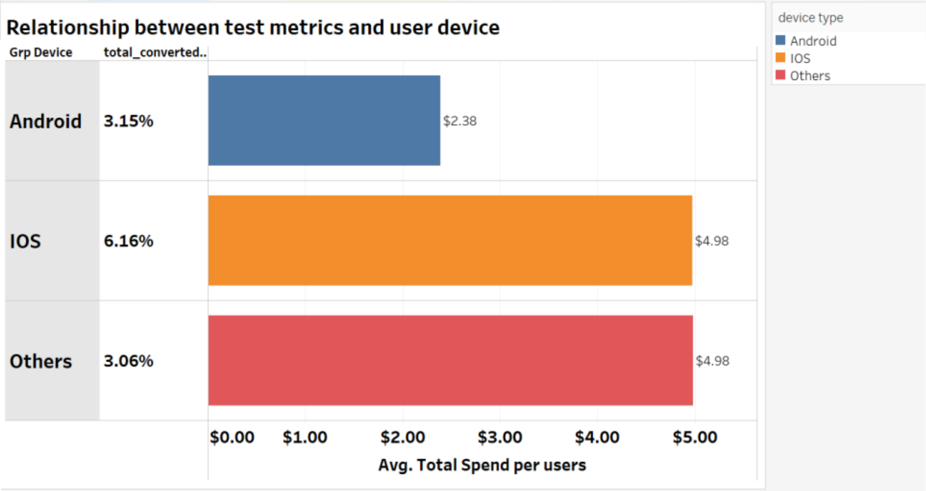
Visualizing the test metrics

All the dashboard can be found in the Tableau public page linked in Appendix F.

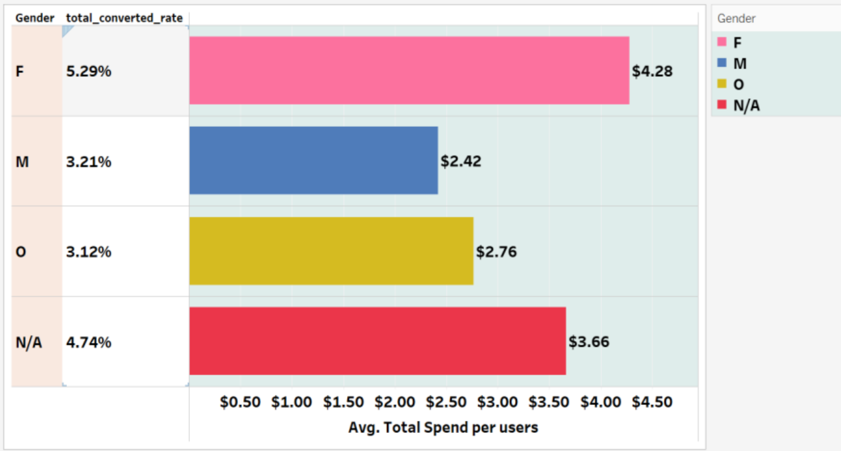
Distribution of the amount spent per user for each group



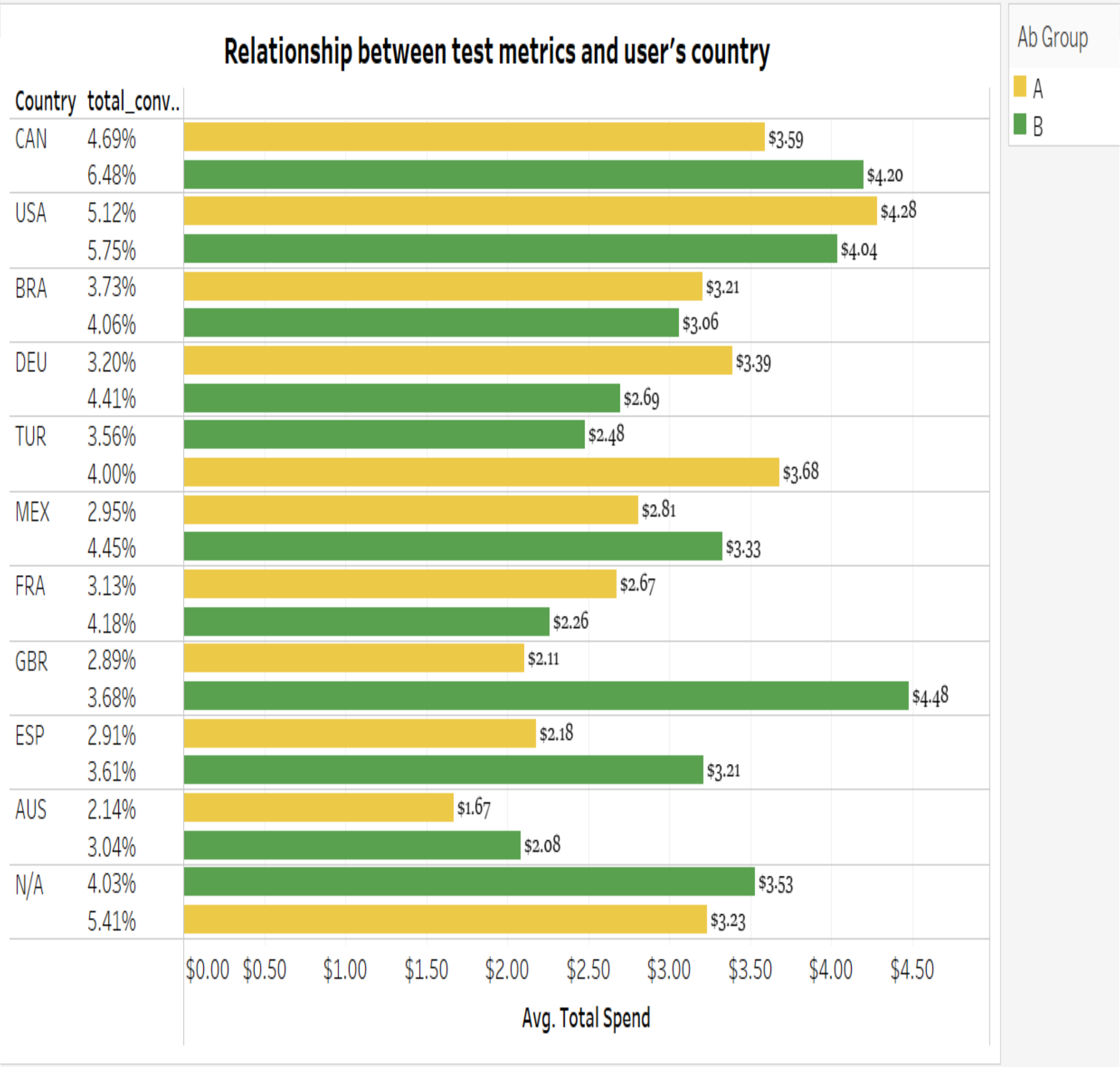
The relationship between the test metrics and user’s devices



Relationship between the test metrics and user's gender



Relationship between the test metrics and user's country


Relationship between test metrics and user's country
Country
total_conv..
Ab Group
🟨 A
🟩 B
CAN
4.69%
$3.59
6.48%
$4.20
USA
5.12%
$4.28
5.75%
$4.04
BRA
3.73%
$3.21
4.06%
$3.06
DEU
3.20%
$3.39
4.41%
$2.69
TUR
3.56%
$2.48
4.00%
$3.68
MEX
2.95%
$2.81
4.45%
$3.33
FRA
3.13%
$2.67
4.18%
$2.26
GBR
2.89%
$2.11
3.68%
$4.48
ESP
2.91%
$2.18
3.61%
$3.21
AUS
2.14%
$1.67
3.04%
$2.08
N/A
4.03%
$3.53
5.41%
$3.23
$0.00
$0.50
$1.00
$1.50
$2.00
$2.50
$3.00
$3.50
$4.00
$4.50
Avg. Total Spend

Recommendation

We do not recommend to launch the website with the banner because we did not see enough statistical evidence to show it has a positive change on the revenue. Even though there is a positive conversion rate change between the groups, there is no significant difference on average spent per user to have a positive impact on total revenue.

However, when we see the time series behavior on conversation rate, we see a significant increase on conversion rate towards the end of the test period. This could continue if we test it for longer period. And also, these converted users will have a potential that they may increase their spending because, there will be a repetitive purchase on food and beverage products in the future as it is not a one-time usage if they like the product.

In addition, in the distribution of amount spending and number of users graph shows that there are higher purchases from the treatment group that indicate that there is a customer who spent more after the banner, if we give more time, we could see a positive impact on the revenue. Therefore, if the cost of the banner design and test is low, we would recommend to run the test with in at lease 4 to 8 weeks to see if these assumptions are true to have a positive impact on revenue.

Appendix

A:
https://docs.google.com/document/d/1mmxdvEc41sFk36QLSla1b83br5UDYwvtGzwXAoXgXmE/edit

B:
https://docs.google.com/spreadsheets/d/19_942Uebco409WPA1uR5YDtYQnWtkpZRnNyWnUprQ5k/edit#gid=0

C:
https://docs.google.com/spreadsheets/d/19_942Uebco409WPA1uR5YDtYQnWtkpZRnNyWnUprQ5k/edit#gid=947160096

D:
https://docs.google.com/spreadsheets/d/19_942Uebco409WPA1uR5YDtYQnWtkpZRnNyWnUprQ5k/edit#gid=1610229971

E:
https://docs.google.com/spreadsheets/d/19_942Uebco409WPA1uR5YDtYQnWtkpZRnNyWnUprQ5k/edit#gid=107587036

F: Tableau dashboards:
https://public.tableau.com/app/profile/mekdes.asfaw

References:

I. https://abtestguide.com/calc/
II. https://cxl.com/ab-test-calculator/
III. https://statisticsbyjim.com/hypothesis-testing/confidence-interval/
IV. https://blog.analytics-toolkit.com/2022/a-b-testing-statistics-a-concise-guide/
V. http://www.imathas.com/stattools/norm.html
VI. https://onlinestatbook.com/2/estimation/t_distribution.html
VII. http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pt.html
VIII. https://campus.datacamp.com/courses/statistical-techniques-in-tableau/measures-of-spread-and-confidence-intervals?ex=10
IX. https://blog.hubspot.com/marketing/a-b-testing-experiments-examples
X. https://www.abtasty.com/blog/learn-from-5-ab-test-case-studies/
XI. https://cxl.com/blog/visualize-ab-test-results/
XII. https://cxl.com/ab-test-calculator/
XIII. https://abtestguide.com/calc/
XIV. https://www.abtasty.com/sample-size-calculator/
XV. https://www.datacamp.com/cheat-sheet/data-viz-cheat-sheet
XVI. https://statulator.com/SampleSize/ss2M.html
XVII. https://www.crazyegg.com/blog/ab-testing-examples/