A/B Test Analysis for GloBox Website.
By: Anhelina Baftalovska

Summary

In this A/B test analysis of the GloBox website, two versions of a web page were compared: a control group without a banner and a treatment group with a food and drink banner at the top of the website. The analysis focused on the metrics of conversion rate and average amount spent per user.

The results indicate that users who were exposed to the banner had a significantly higher conversion rate compared to the control group. However, there was no statistically significant difference in the average amount spent per user between the two groups.

Further analysis revealed insights based on different factors such as device type, gender, and region/country. These insights provide a preliminary understanding of user behavior, but more comprehensive data on user demographics and campaign targeting would be beneficial.

Additionally, a power analysis and join curve forecast suggest that a larger sample size would be required to achieve sufficient statistical power and accurately capture the desired effects. Therefore, it is recommended to rerun the test at a larger scale to ensure more reliable results.

In summary, based on the significant difference in conversion rate and the potential for acquiring more paying customers, it is recommended to launch the banner feature. However, further data analysis and a larger sample size are necessary to gain a deeper understanding of the banner's impact and optimize its effectiveness in generating higher revenue. Improvements can be made by considering additional user factors and conducting a more extensive power analysis.

Context
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.

Motivation

The Growth team decides to run an A/B test to compare two versions of a web page that highlights key products in the food and drink category as a banner at the top of the website. By randomly assigning users to either the A or B version, the business can determine which version is more effective. All users should be assigned to one A/B test group. The experiment is only being run on the mobile website. The control group (A) does not see the banner, and the treatment(B) group sees it as shown below:

Group A: Control
existing landing page

Group B: Treatment
landing page with food & drink banner
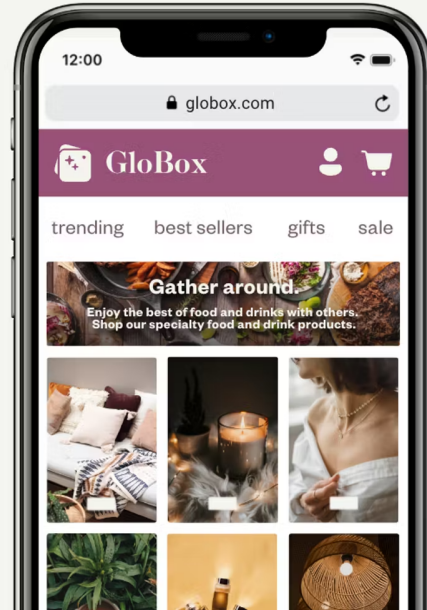
The user subsequently may or may not purchase products from the website. It could be on the same day they join the experiment, or days later. Not all users make a purchase. If they do make one or more purchases, this is considered a "conversion".

Test Parameters

Control group: had the original landing page, 24,343 users
Treatment group: had landing page with food and drink banner, 24600 users
Total number of users: 48,943
Duration: 13 days (January 25th to February 6th, 2023)
Type of mobile devices used: iOS, Android
Countries: AUS, CAN ,BRA, FRA, MEX, TUR, GBR, DEU, ESP, USA

Metrics

- Conversion Rate(%): the number of successful conversions (users who purchased) divided by the total number of users.
- Revenue($): average amount spent per user including users who made and didn't make purchases.

Results

Our analysis found that users who had banners have significantly higher conversion rate, while average spent doesn't have a meaningful difference.

Convertion rate and AVG Amount spend by Test group
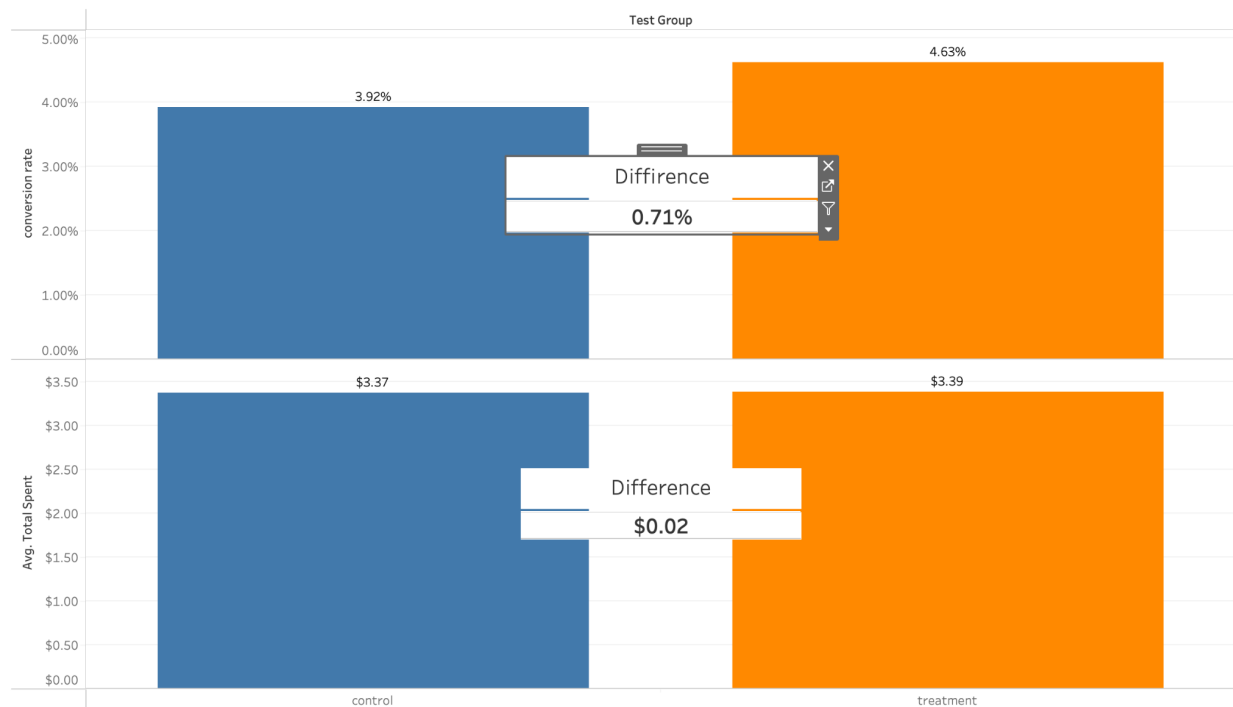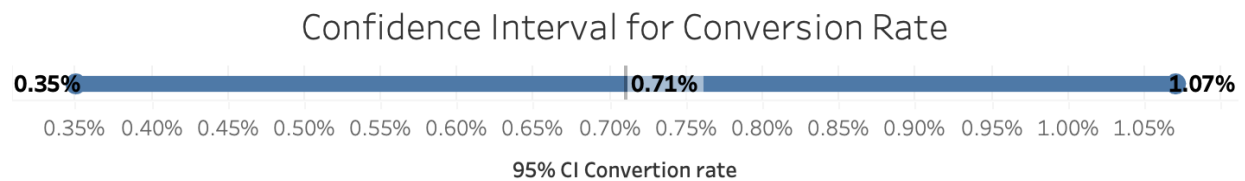

Hypothesis Test (Conversion Rate)

In order to test for a statistically significant difference in conversion rate between control and treatment group, we used a two-sided z-test for proportion, since it is ideal to determine a statistical significance of positive or negative results, or if there is not statistically significant between the proportions. We used the normal distribution and a 5% significance level as well as the pooled proportion for the standard error.

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

Alternative Hypothesis (HA): There is a difference in the conversion rate between the two groups.

After performing the calculation, we found p-value comes out to be extremely small (approximately 0.0001), which is less than the significance level of 0.05, we reject the null hypothesis. This means that there is sufficient evidence to conclude that there is a significant difference in the conversion rate between the two groups.

The 95% confidence interval for the difference in the conversion rate between the treatment and control groups falls approximately (0.00352, 0.01068)

Confidence Interval for Conversion Rate

0.35%	0.71%	1.07%

0.35% 0.40% 0.45% 0.50% 0.55% 0.60% 0.65% 0.70% 0.75% 0.80% 0.85% 0.90% 0.95% 1.00% 1.05%

95% CI Convertion rate

The confidence interval for the difference in proportions 0.35% and 1.07% does not include zero, which further supports the conclusion that there is a significant difference between the two groups.

Hypothesis Test (Average Amount Spent)
To test whether there is a difference in the average amount spent per user between the control and treatment groups, we can perform a two-sample t-test.
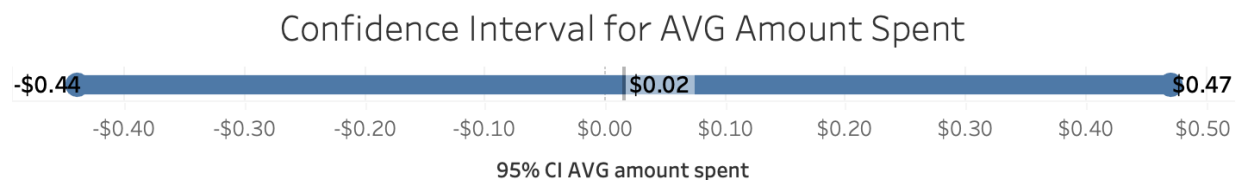We used the t distribution and a 5% significance level, assuming unequal variance.

Null Hypothesis (H0): The average amount spent per user is the same in the control and treatment groups.

Alternative Hypothesis (HA): The average amount spent per user is different between the control and treatment groups.

After performing the calculation, we return a p-value of 0.9439, which is greater than our significance level of 0.05. Therefore, we would not reject the null hypothesis. There is no statistically significant difference in the average amount spent per user.

The 95% confidence interval for the difference in the average amount spent per user between the treatment and control groups is (-0.4387, 0.4714).

Confidence Interval for AVG Amount Spent

-$0.44	$0.02	$0.47

-$0.40   -$0.30   -$0.20   -$0.10 $0.00 $0.10 $0.20 $0.30 $0.40 $0.50

95% CI AVG amount spent

Since the interval contains zero, which represents no difference in average spending between the two groups, we can conclude that there is no statistically significant difference in the average amount spent per user between the treatment and control groups at the 5% significance level.

Results by different metrics

Distribution of the amount spent per user

Both the control and treatment groups had a substantial number of purchases in this price range, indicating that a significant portion of users spent below $300. On average, the amount spent per purchase is $69.93 across the two groups. Only 31 (1.4%) purchases out of 2,094 spent between $300 and $1546.

Spending Distribution per group
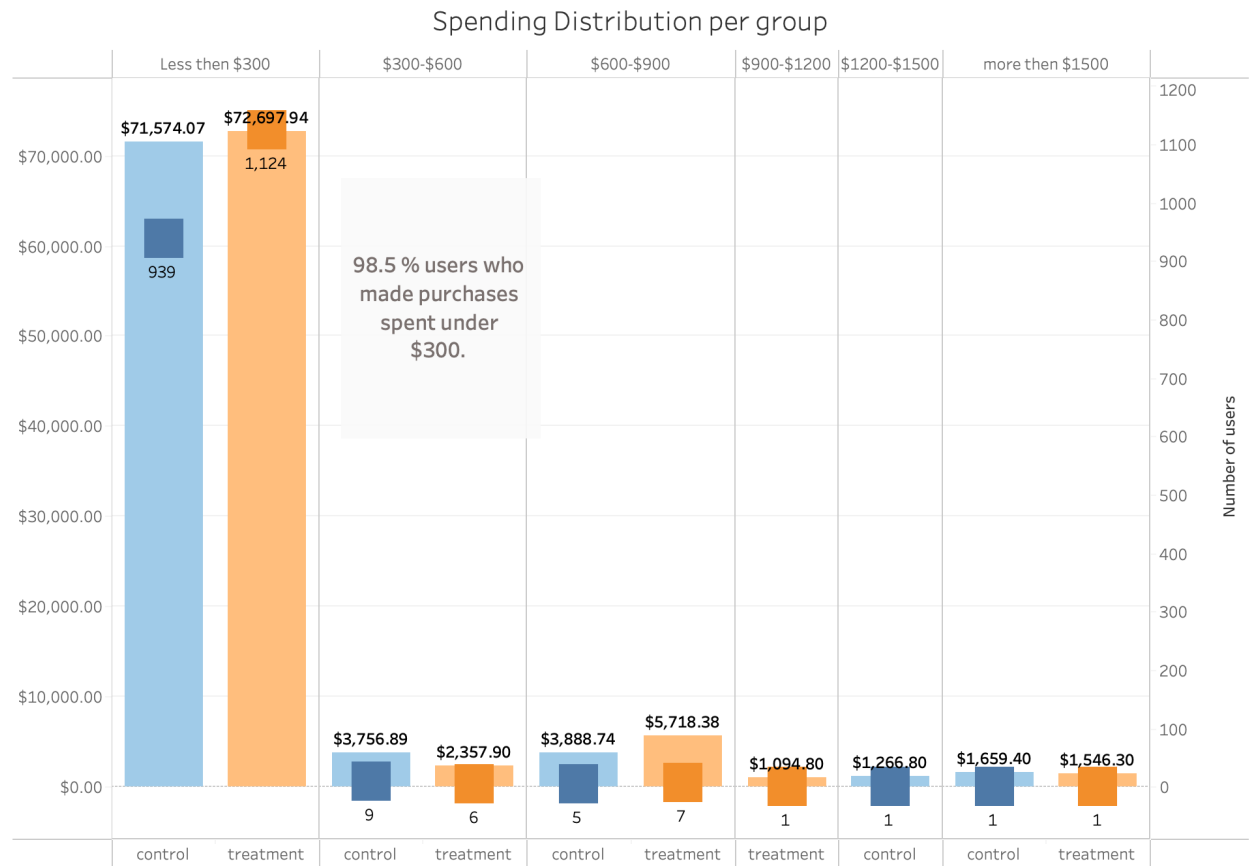
Result by Device

From these observations, we can make the following preliminary conclusions:

The treatment group tends to have higher conversion rates compared to the control group across all device types, except for unknown devices.

The average amount spent varies across device types, with iOS users generally spending more than Android users.

Unknown device users have a lower average amount spent in the treatment group but a higher conversion rate compared to the control group. However, the sample size for the unknown devices is small, so the findings may not be as statistically significant.

Test metrics by Device
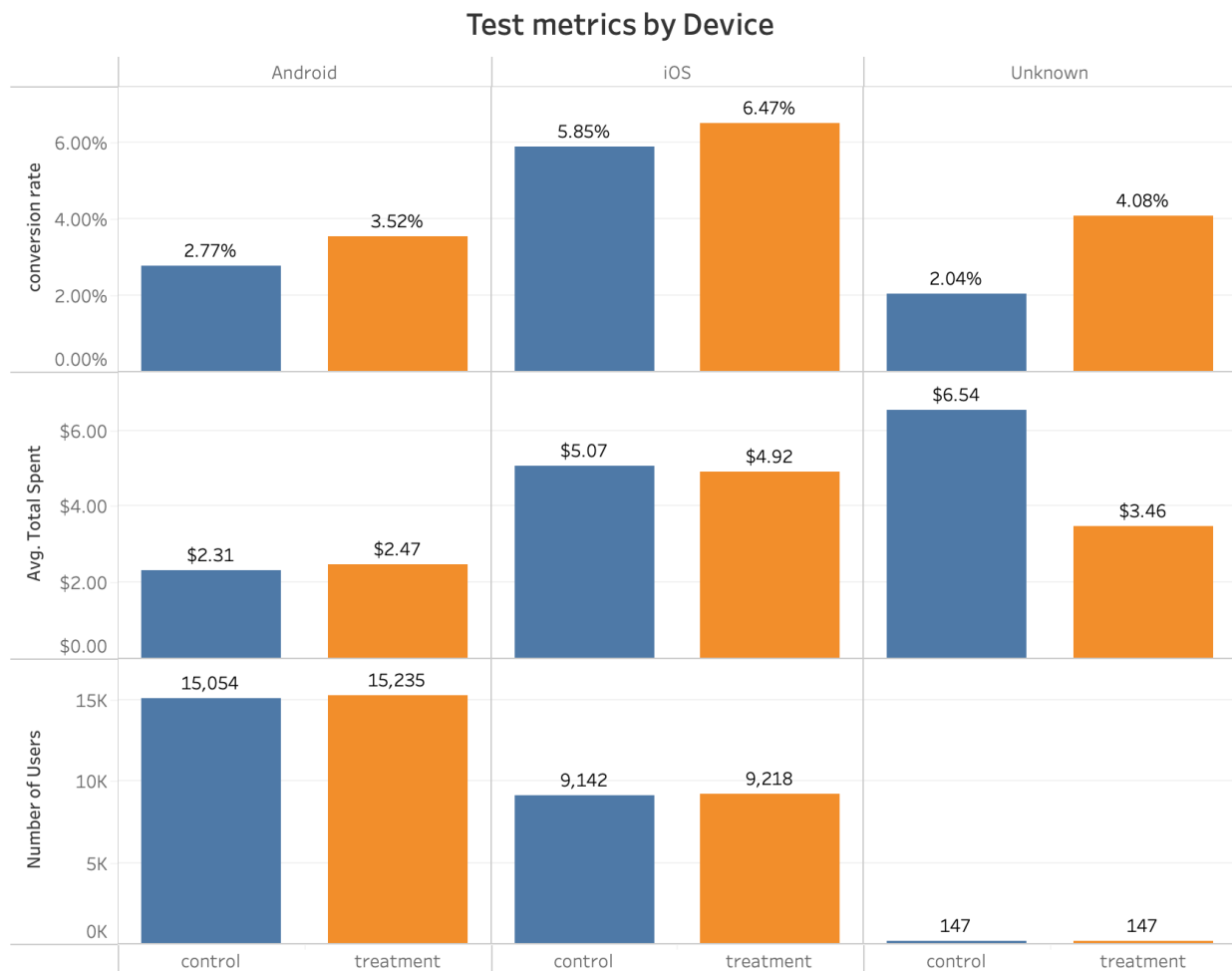


Result by Gender

Female users tend to have higher conversion rates compared to male users, but their average amount spent is slightly lower.

Male users, on the other hand, have a lower conversion rate but a higher average amount spent.

Users with "Other" gender and those with unknown gender show similar average amount spent and conversion rates in both the treatment and control groups.

Test metrics by Gender
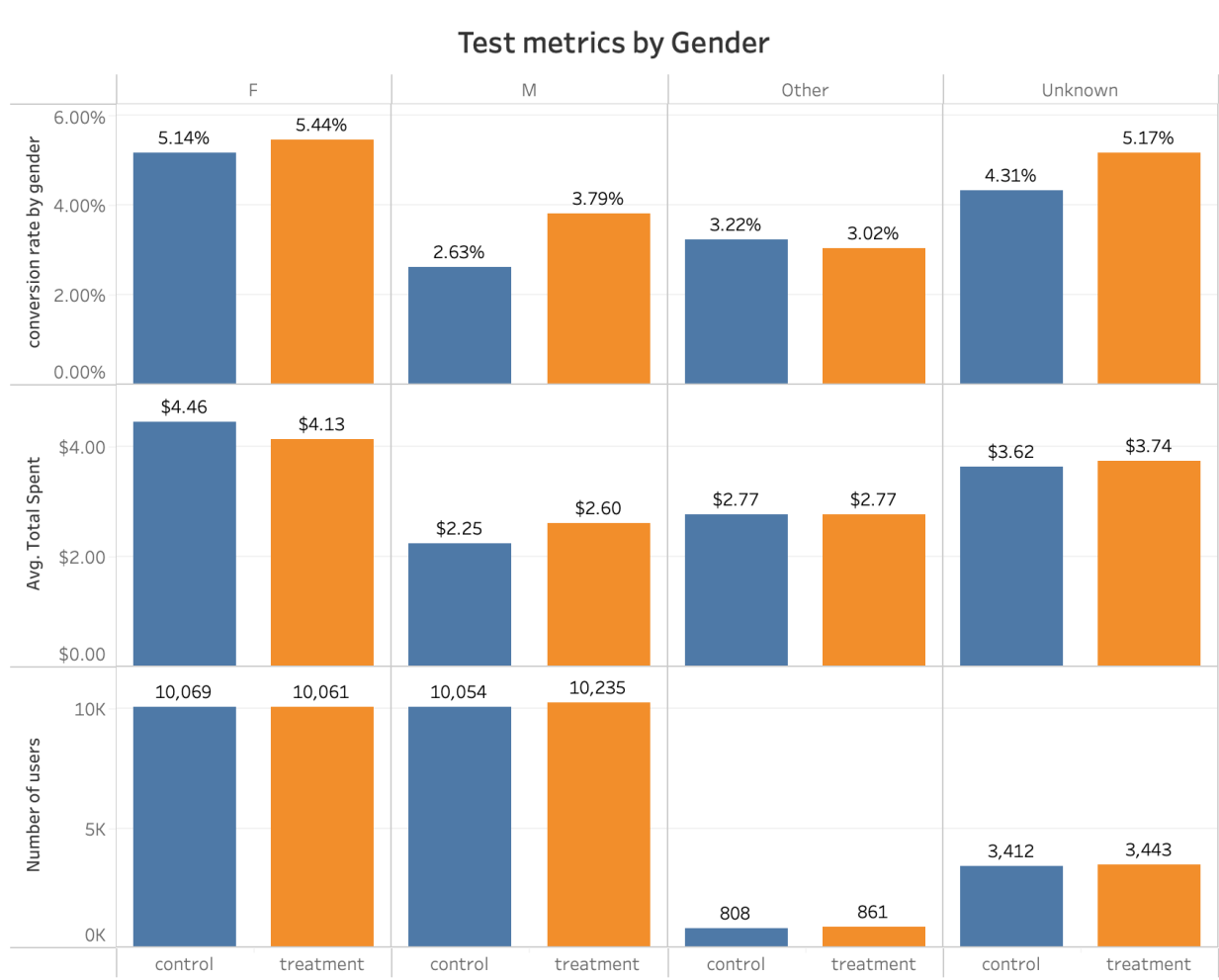
Result by Region/Country

The treatment groups generally have higher conversion rates compared to the control groups across different countries, except in the unknown category.

Average amount spent varies across countries, with Europe having the highest average amount spent and Australia having the lowest.

The relationship between average amount spent and conversion rate varies across countries, with some countries showing similar trends and others showing variations.

Test metrics per Continent

	North America		South America		Europe		AUS		Uknown	
conversion rate by country	9.81% (control)	12.22% (treatment)	6.67% (control)	8.51% (treatment)	16.12% (control)	19.45% (treatment)	2.14% (control)	3.04% (treatment)	5.41% (control)	4.03% (treatment)
Avg. Total Spent	$4.23	$4.07	$3.07	$3.17	$2.93	$2.97	$1.67	$2.08	$3.26	$3.54
Count of Id	8,076	8,266	7,620	7,552	7,743	7,875	608	560	296	347
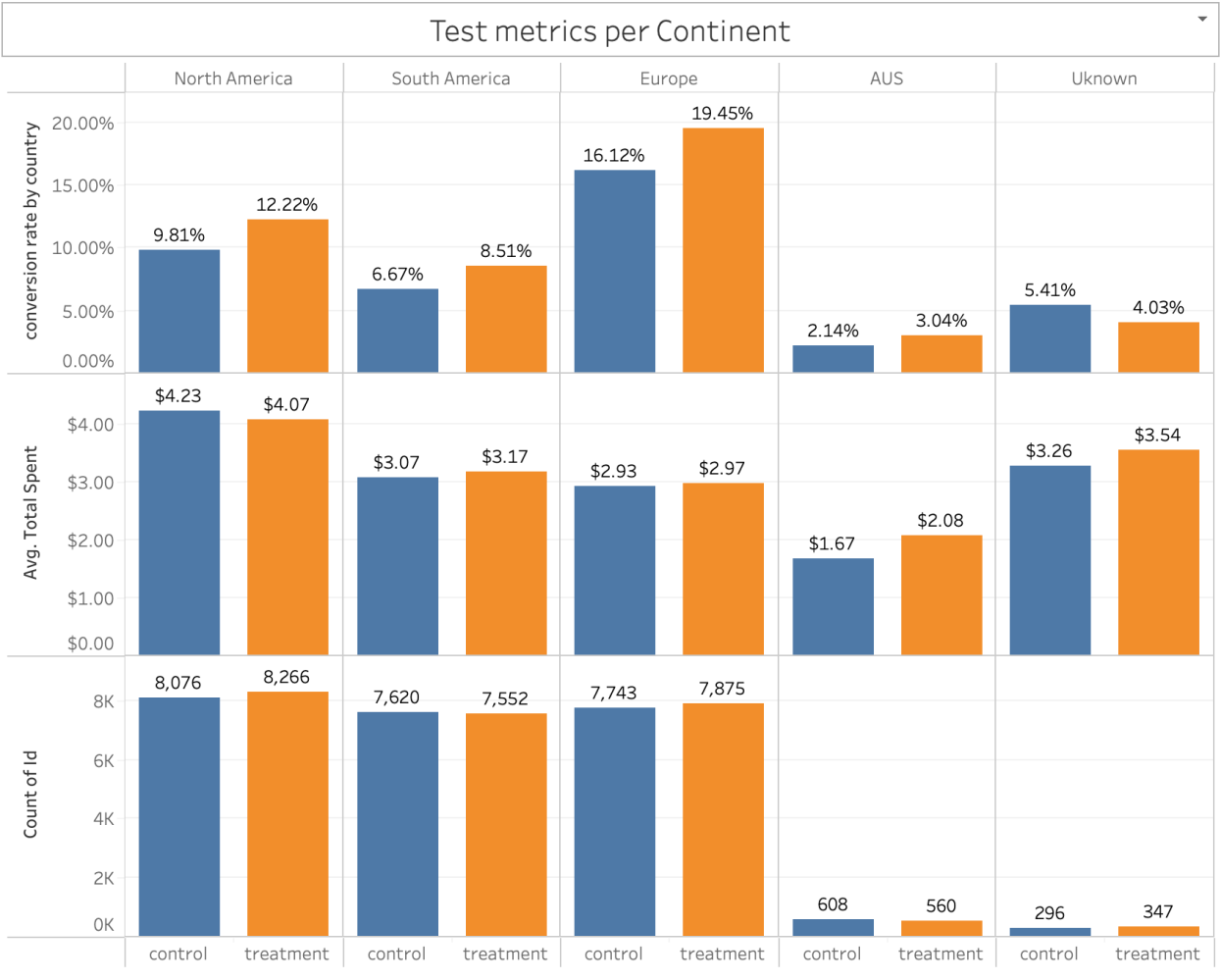
Test Metrics by Country

Count.. ≟	Number of users		Conversion Rate by country		Avg. Total Spent		Total Spent	
	control	treatment	control	treatment	control	treatment	control	treatment
AUS	608	560	2.14%	3.04%	$1.67	$2.08	$1,014.38	$1,165.15
Uknown	296	347	5.41%	4.03%	$3.26	$3.54	$966.40	$1,227.84
ESP	997	996	2.91%	3.61%	$2.18	$3.23	$2,171.86	$3,221.30
CAN	767	803	4.69%	6.48%	$3.60	$4.20	$2,762.66	$3,371.45
FRA	1,536	1,554	3.13%	4.18%	$2.68	$2.27	$4,113.21	$3,524.63
GBR	1,455	1,494	2.89%	3.68%	$2.11	$4.50	$3,068.22	$6,720.02
TUR	1,849	1,883	4.00%	3.56%	$3.69	$2.49	$6,814.26	$4,686.71
DEU	1,906	1,948	3.20%	4.41%	$3.40	$2.71	$6,481.52	$5,275.35
MEX	2,815	2,923	2.95%	4.45%	$2.81	$3.35	$7,915.63	$9,778.91
BRA	4,805	4,629	3.73%	4.06%	$3.21	$3.07	$15,442.95	$14,193.06
USA	7,309	7,463	5.12%	5.75%	$4.30	$4.05	$31,394.81	$30,250.91

To find out more about the relationship between the test metrics (conversion rate and average amount spent) and the user's device, gender or country/region we should consider other factors, such as user behavior, demographics, or campaign targeting, which can provide a more comprehensive understanding.

Novelty Effect

Initially, on January 25th, the Treatment Group had a higher average spent and total spent compared to the Control Group.

However, from January 26th to January 30th, the Control Group consistently showed higher average spent and total spent compared to the Treatment Group .

On January 31st, there was a shift where the Treatment Group had a higher average spent and total spent compared to the Control Group.
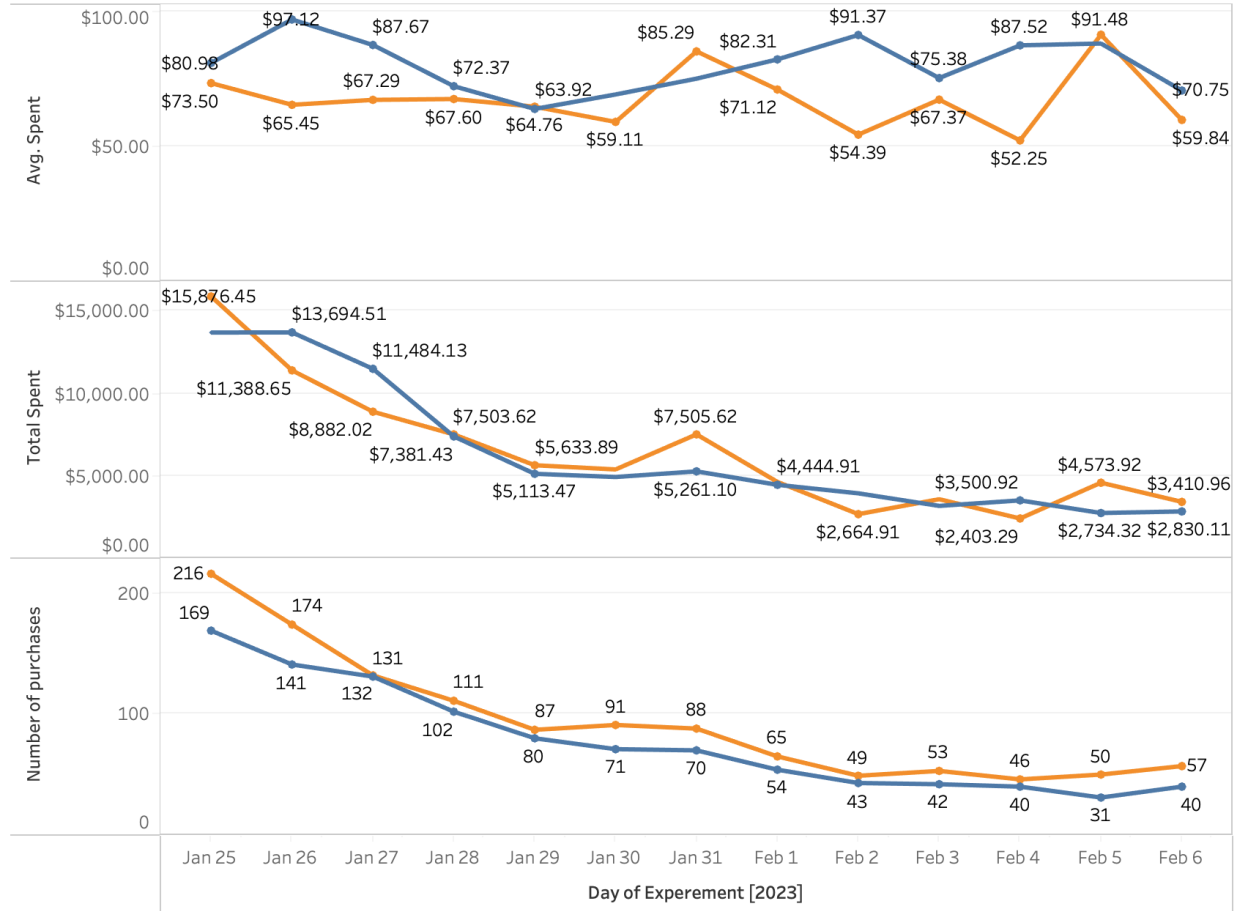
From February 1st to February 6th, the Control Group consistently delivered higher average spent and total spent compared to the Treatment Group.

While the average amount spent may vary between the groups, Treatment Group consistently has a higher number of purchases or transactions. This could indicate that Treatment Group has a higher user engagement, leading to more purchases overall.

Overall, this data suggests that the presence of the banner (Treatment group) may have an impact on user behavior, resulting in differences in the average amount spent and count of purchases between the two groups.

Based on this analysis, it does not appear that there is a clear novelty effect observed where the effectiveness of the banner is short-lived. The behavior of the groups fluctuates throughout the period, and there is no consistent pattern indicating a significant novelty effect.

Novelty Effects

Avg. Spent

$80.98, $97.12, $87.67, $72.37, $63.92, $85.29, $82.31, $91.37, $75.38, $87.52, $91.48, $70.75
$73.50, $65.45, $67.29, $67.60, $64.76, $59.11, $71.12, $54.39, $67.37, $52.25, $59.84

Total Spent

$15,876.45, $13,694.51, $11,484.13, $7,503.62, $5,633.89, $7,505.62, $4,444.91, $3,500.92, $4,573.92, $3,410.96
$11,388.65, $8,882.02, $7,381.43, $5,113.47, $5,261.10, $2,664.91, $2,403.29, $2,734.32, $2,830.11

Number of purchases

216, 174, 131, 111, 91, 88, 65, 53, 46, 50, 57
169, 141, 132, 102, 80, 71, 70, 54, 43, 42, 40, 31, 40

Day of Experement [2023]

Power Analysis / Join Curve forecast

Based on the sample size calculations and the join curve forecast provided, we can draw the following conclusions:

For the conversion rate metric, the power analysis suggests that a total sample size of approximately 60.6k users (both groups combined) would be needed to achieve a desired minimum detectable effect (MDE) of 10% with a statistical power of 0.8. This indicates that the test should have been conducted with a larger sample size to ensure sufficient sensitivity in detecting the desired effect.

Similarly, for the average amount spent metric, the power analysis suggests a total sample size of approximately 182k users (both groups combined) to detect a difference of $0.337 (10% of control group mean) with a significance level (α) of 0.05 and a power of 0.8. Again, this indicates that a larger sample size should have been aimed to increase the sensitivity of the test.

The join curve forecast provides an estimate of the cumulative number of users joining the experiment over time. It shows a gradual increase in the number of users joining from February 6, 2023 to March 25, 2023.

In conclusion, both the sample size calculations and the Join curve forecast indicate the need for a larger sample size to achieve sufficient statistical power and to capture the desired effects accurately. It would be recommended to run the test again at a larger scale, ensuring that the sample size meets the recommended requirements determined by the power analysis.

Join Curve Forecast
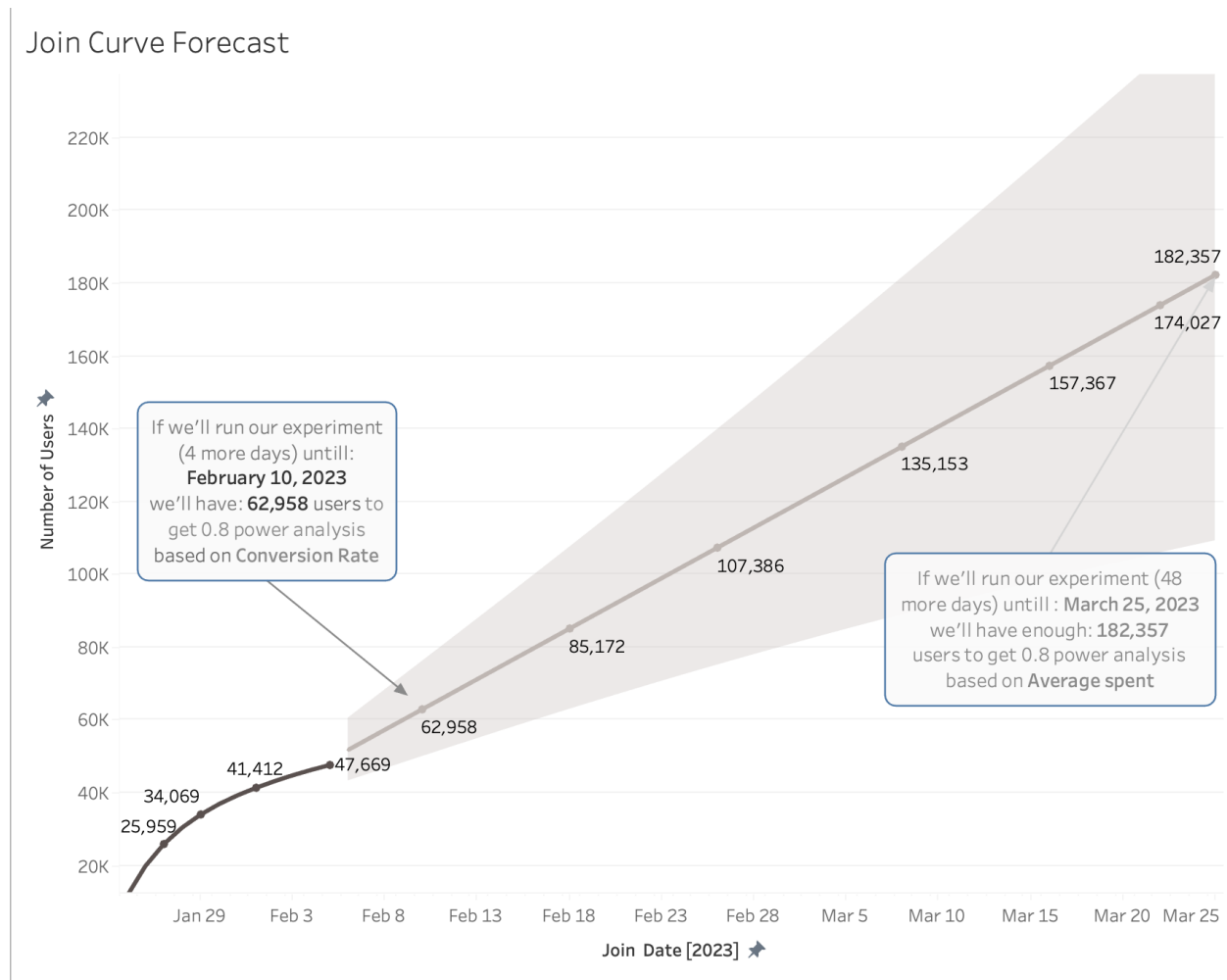
Recommendation

This banner feature should be launched. Based on the conclusions of the hypothesis tests, we find that there is a statistically significant difference in the conversion rate

between the groups. In such cases, it is justifiable to go ahead and launch the banner feature.

Even if we observe a significant increase in the conversion rate, without a corresponding increase in average amount spent, it is still a positive outcome. It indicates a potential improvement in acquiring more paying customers or generating higher revenue from existing customers.

However, to make a more confident recommendation, further data analysis and a larger sample size are needed. This will allow us to gain a deeper understanding of the impact of the banner feature on user behavior and evaluate its effectiveness more accurately. Launching the feature initially and monitoring its impact will provide valuable insights for future decisions and optimizations.

Appendix

SQL Data Extraction:

Extract full relevant data from three tables:

```
1 SELECT u.id, u.country, u.gender,
2       COALESCE(sum(a.spent), 0) as total_spent,
3       g.group, g.join_dt, g.device,CASE
4         WHEN SUM(a.spent) > 0 THEN 1
5         ELSE 0 END AS "Converted"
6 FROM users as u
7 LEFT JOIN activity as a
8 ON u.id = a.uid
9 LEFT JOIN groups as g
10 ON u.id = g.uid
11 GROUP BY u.id, u.country, u.gender, g.group, g.join_dt, g.device
```

Extract Data to check Novelty Effect:

```
1 Select a.Uid as ID , "group", dt as date, spent
2 from activity as a
3 left join "groups" as g
4 ON a.uid = g.uid
5
6
```

Extract running total users over time data to build Join curve forecast:

```
1 select join_dt as join_date,count(uid) as number_of_users,
2        sum(count(uid)) OVER (ORDER BY join_dt) as running_total_users
3 From "groups"
4 group by join_dt
5 order by join_dt
6
7
8
9
```

[Google Sheets. Data and Analysis Calculation](#)

[Tableau Story](#)