GLOBOX PROJECT

Abstract

An A\B test for a new feature in GloBox

Robert Tiger

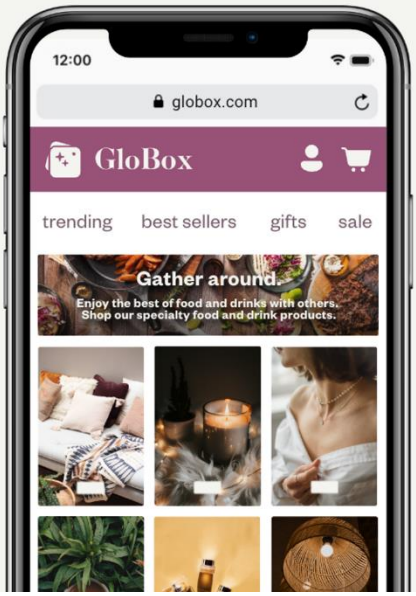
August 23

Table of Contents

Background

Group A: Control
existing landing page

Group B: Treatment
landing page with food & drink banner



In this project, we embark on a journey with GloBox, an innovative e-commerce company known for its unique selection of boutique fashion items and high-end decor products.

As the Data Analyst for GloBox, we are tasked with dissecting the results of a crucial A/B test designed to elevate the visibility of their burgeoning food and drink category. This test involves showcasing key products in the food and drink section through a banner on the mobile website.

The control group is spared this banner, while the test group is exposed to it. Our role extends to collaborating with the Growth Product & Engineering Team, led by Leila and Alejandro, along with Mei, the Head of Marketing, to derive data-driven insights. As we delve into the user-level aggregated dataset extracted using SQL, we will uncover patterns in user behavior, conversions, and purchases, drawing upon statistical methods and visualization tools.

With the impending decision on whether to implement the new experience across the board, our findings and recommendations will serve as the cornerstone of this pivotal choice, bridging the realms of data analysis, product development, and growth strategy.

TL;DR \ Summary

This project involves analysing an A/B test conducted by GloBox, an e-commerce company, to evaluate the impact of a new banner promoting food and drink products on their mobile website. The analysis covers ***(Click on the title to jump to the section)***:

1. **Hypothesis Test for Conversion Rate:** There is a significant difference in conversion rates between the two groups (control and treatment) in favor of the treatment group. This suggests the new banner is effective.
2. **Confidence Interval for Conversion Rate:** The 95% confidence interval shows that the true difference in conversion rates lies between 0.3% and 1.06%. This supports the positive impact of the treatment.
3. **Hypothesis Test for Average Amount Spent:** There's no significant difference in average spending between the two groups. The new banner didn't influence spending significantly.
4. **Confidence Interval for Average Amount Spent:** The 95% confidence interval ranges from approximately -0.44 to 0.47, indicating no statistically significant difference in spending.
5. **Visualizations:** Visual insights include conversion rate, spending distribution, country-wise analysis, gender-based analysis, and device-based analysis.
6. **Recommendations:** Despite a significant rise in conversion rates, the analysis indicates that launching the new feature may not result in increased revenue or average user spending, warranting further consideration of its implementation.
7. **Checking Novelty Effect:** A decline in spending over time indicates a possible novelty effect that should be monitored.
8. **Power Analysis:** The A/B test was conducted with a larger sample size than needed, ensuring robust and reliable results.

Overall, the analysis **does not support launching the new feature**, as despite the positive impact on conversion rates, there is no evidence of significant improvements in revenue or average user spending, and there is a concern about potential novelty effects.

Difference in Conversion Rate Hypothesis Test

Null Hypothesis (H0) μ1 = μ2: There is no significant difference in conversion rate between the two landing page groups. Any observed differences are due to random chance.

Alternative Hypothesis (Ha) μ1 ≠ μ2: There is a significant difference in conversion rate between the groups. The observed difference is not likely to be due to random chance alone, but rather indicate a true effect of the different landing pages.

1. Proportions or Means: Conversion rate is a proportion since it represents the percentage of individuals who take a desired action out of the total number of individuals in a group.

2. One Sample or Two Samples: In an A/B test, we are comparing two different groups. Therefore, we are working with two samples.

3. One-sided or Two-sided: The alternative hypothesis we mentioned earlier is μ1 ≠ μ2, which indicates a two-tailed alternative hypothesis. This means we are interested in whether there is any significant difference in either direction between the conversion rates of the two landing page groups.

Considering these factors, the appropriate statistical test for this A/B test comparing conversion rates with two proportions would be the **two-sample Z-test for proportions**. This test is used to determine if the observed differences in conversion rates between the two groups is statistically significant. It compares the proportions of successes (conversions) in two independent samples.

The formula for the test statistic in this case is calculated as:

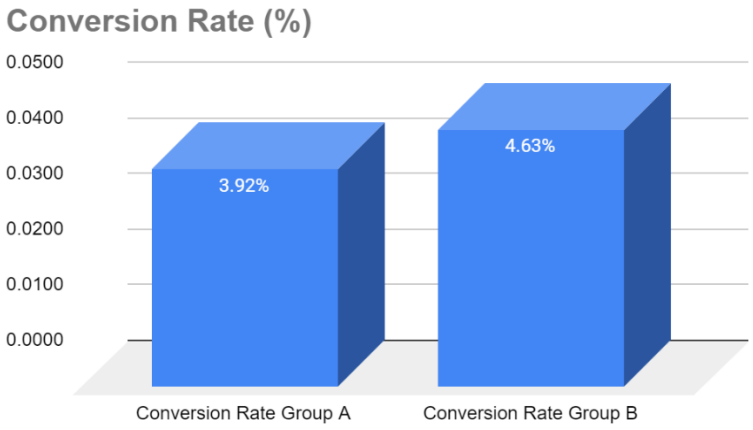
$$Z = \frac{(p_A - p_B)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

- p1 = 0.0392
- *p*2 = 0.0463
- *p_hat* = 0.0427
- *n*1 = 24,343
- *n*2 = 24,600

Z ≈ 3.86

Conclusion

1. **Z-test Statistic and p-value Analysis**:
   The calculated Z score, $Z = 3.86$, indicates a considerable deviation from expected proportions under the null hypothesis. Correspondingly, the p-value of 0.000111 underscores its statistical significance, firmly surpassing the designated significance level (α=0.05).
2. **Interpreting Statistical Significance:**
   The p-value demonstrates that the probability of observing a difference in conversion rates as substantial as this, assuming the null hypothesis, is exceptionally low. This means that the outcome could not have been due to random chance.
3. **Implications and Strategic Impacts:**
   Our analysis highlights the emergence of statistically significant divergence in conversion rates, particularly **favouring Group B**. This suggests that the strategies implemented in Group B have contributed to a discernible alteration in user behaviour.



4. **Guidance and Future Actions:**
   In light of these statistical findings, we recommend a more in-depth exploration of the underlying mechanisms driving the observed difference. It's important to recognize that the p-value is a foundational element guiding our analytical path. Our future actions involve refining strategies based on comprehensive insights gained from deeper exploration.

Confidence Interval for Difference in Conversion Rate

The formula for the confidence interval for the difference in proportions is given by:

$$\text{CI} = (\hat{p}_1 - \hat{p}_2) \pm Z \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Sample Statistic = Sample Proportion (Treatment) - Sample Proportion (Control)

p2_hat - p1_hat = 0.039 - 0.046 = 0.007

To calculate the ***Standard Error***:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Given the data we have:

- *P*1_hat = 0.0463
- *P2_hat* = 0.0392
- *n*1 = 24,600
- *n*2 = 24,343

=SQRT((K10*(1-K10)/K8) + (K11*(1-K11)/K9)) (K10 = p1_hat, K8 = n1, K11 = p2_hat, K9 = n2)

We can calculate the standard error with the formula above and reach the number of: **≈ 0.0018 *Standard Error***

Critical value using the *NORMSINV()* function in Google Sheets will look like this: =NORMSINV(1 - K18/2) where:

- K18 is alpha (0.05)

➔ We reached the number: ***1.96*** = Critical Value

Calculate the Margin of Error (MoE):

Given the **Critical Z Value** of **1.96** and **Standard Error** of **0.0018**:

MoE = Critical Z Value × Standard Error

MoE = 1.96 × 0.0018 ≈ 0.0035

Constructing the interval:

Lower Bound = Sample Statistic - MoE

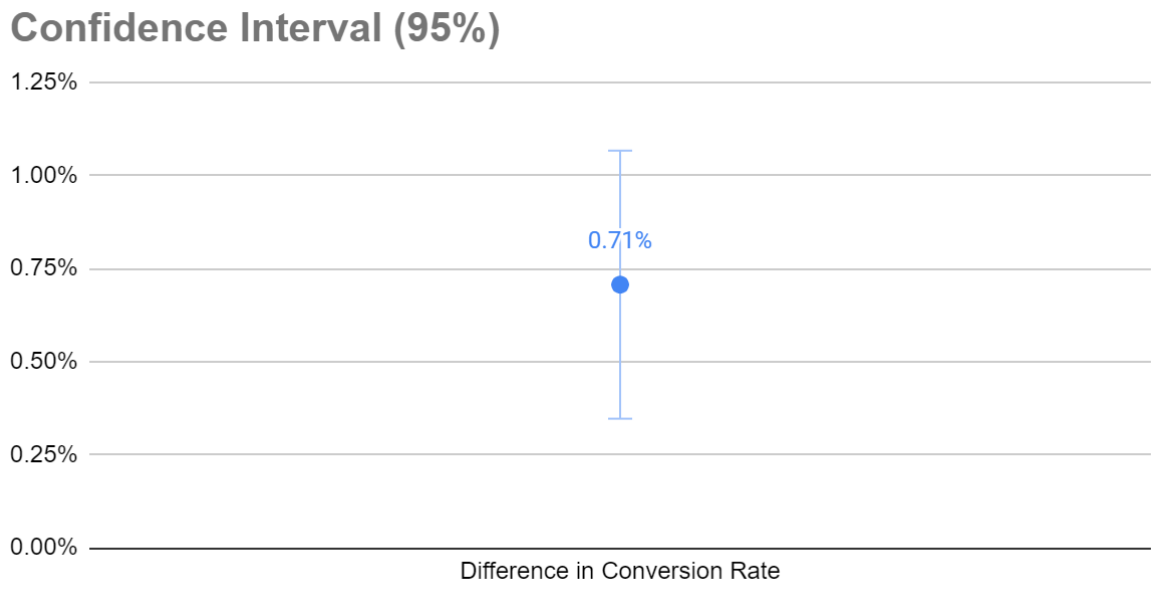
Upper Bound = Sample Statistic + MoE

Given the information provided from the previous steps:

- Sample Statistic: 0.007 (the difference in proportions)
- Critical Value: 1.96
- Standard Error: 0.003

Conclusion

The 95% confidence interval for the difference in conversion rates between the treatment and control groups is approximately **0.003** to **0.0106** (0.3%, 1.06%)



This means that true difference for the conversion rates in either group lies somewhere in between these bounds.

Difference in Average Amount Spent per User Hypothesis Test

Null and Alternative Hypotheses

Null Hypothesis (H0): $H0: \mu 1 = \mu 2$
The average amount spent per user is the same between the two groups.
Alternative Hypothesis (Ha): $Ha: \mu 1 \neq \mu 2$
Average amount spent per user is different between the two groups.

Summarized Data

Calculation	Notation	Value
Sample Size (Control)	n1	24343
Sample Size (Treatment)	n2	24600
Average Sample 1	x1 bar	3.3745
Average Sample 2	x2 bar	3.3909
Difference in Means	x2_bar - x1_bar	0.0163
STD Sample 1	s1	25.9364
STD Sample 2	s2	25.4141

The type of test that we will be using is **two-sample t-test for independent samples with unequal variances.** This test is appropriate when we comparing the means of two independent groups (Group A and Group B) to determine if there is a significant difference between their average values.

Test Statistic Calculation:

$$T = (\bar{x}1 - \bar{x}2) / \sqrt{((s1^2/n1) + (s2^2/n2))}$$
$$T = 0.704$$

Degrees of Freedom:

$$df = =MIN(K38-1,K39-1)$$
$$df = 24342$$

p-Value:

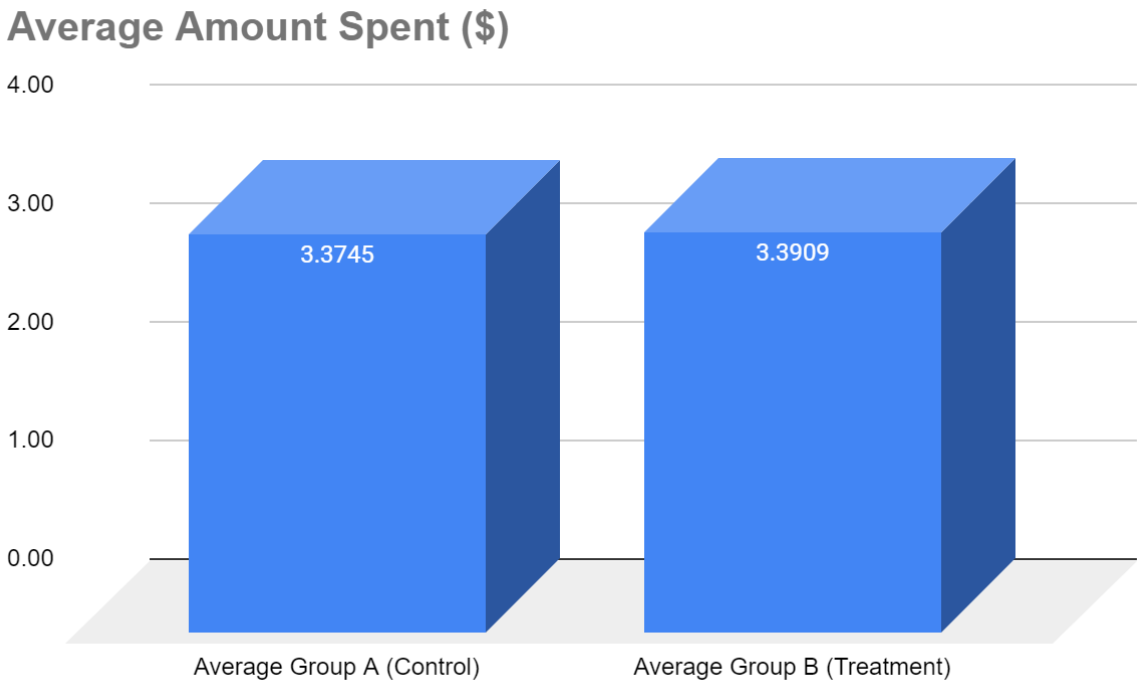
$$P = T.DIST.2T(ABS(K45), K46)$$
$$P = 0.9439$$

Conclusion

After conducting a two-sample t-test for independent samples, the calculated p-value is **0.9439**. Since the p-value is greater than the common significance level of **0.05**, there is not enough evidence to reject the null hypothesis.

Therefore, based on the data and the statistical analysis, **we do not have sufficient evidence to conclude that there is a significant difference in the average amount spent per user between the two groups**. The result suggests that any observed difference in the average amount spent between the groups could likely be due to random variability rather than a true effect caused by the treatment.

In summary, the analysis **does not support the hypothesis that the treatment has led to a statistically significant difference in the average amount spent per user between the two groups.**



Confidence Interval for Difference in Average Amount Spent per User

	Group A	Group B
Size	24343	24600
Mean	3.3745	3.3909
STD	25.8099	25.4141

Sample Statistic (Differenece in Means): *3.39 – 3.37 = 0.0163*

Standard Error: *SE = sqrt((s1^2 / n1) + (s2^2 / n2))*

SE = 0.2316

Critical Value: *Z = NORMSINV(1 - alpha/2)*

Z = 1.96

Margin of Error (MoE):

Given the **Critical Z Value** of **1.96** and **Standard Error** of **0.2316**:

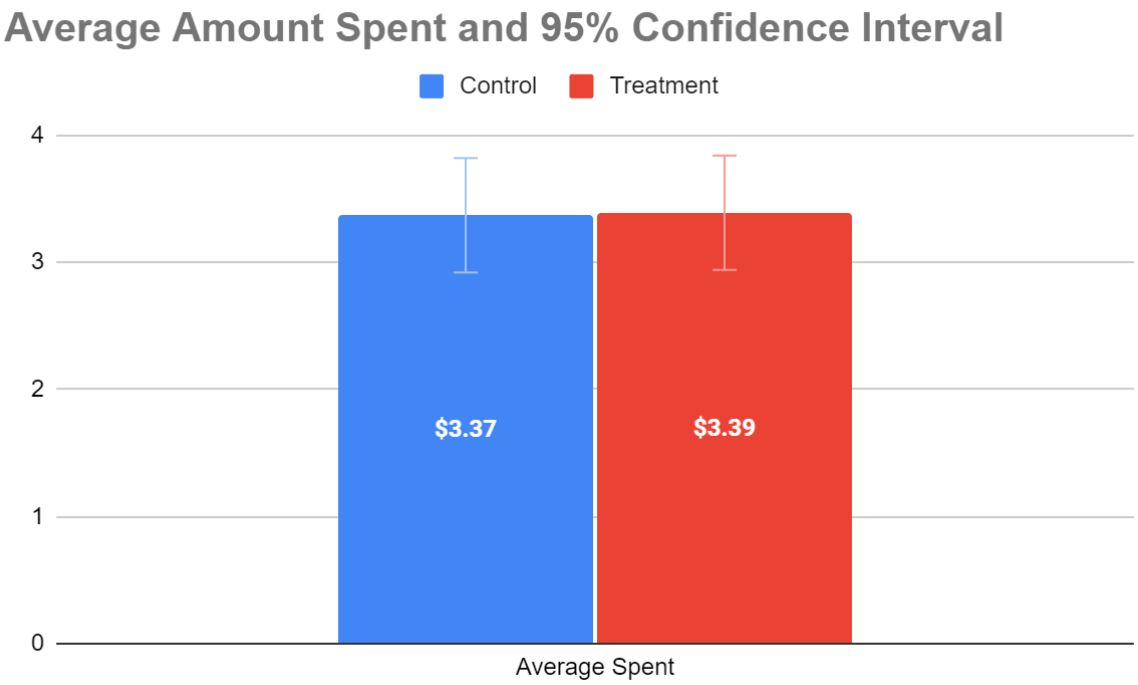
$$MoE = \text{Critical Z Value} \times \text{Standard Error}$$

$$MoE = 1.96 \times 0.2316 \approx 0.4539$$

The Interval: *Sample Statistic +- MoE*

Lower Bound: *0.0163 – 0.4539 = -0.4375*

Upper Bound: *0.0163 + 0.0163 = 0.4702*



Conclusion

We have calculated a 95% confidence interval for the difference in average amount spent per user between Group A (Control) and Group B (Treatment). **The calculated confidence interval ranges from -0.44** to **0.47**. This means that we are 95% confident that the true difference in average spending per user lies somewhere within this interval.

The lower bound of the confidence interval (-0.44) represents the minimum plausible difference in spending between the two groups, while the upper bound (0.47) represents the maximum plausible difference. **Since the interval includes both negative and positive values, we cannot conclude with 95% confidence that there is a statistically significant difference in spending between the two groups**.

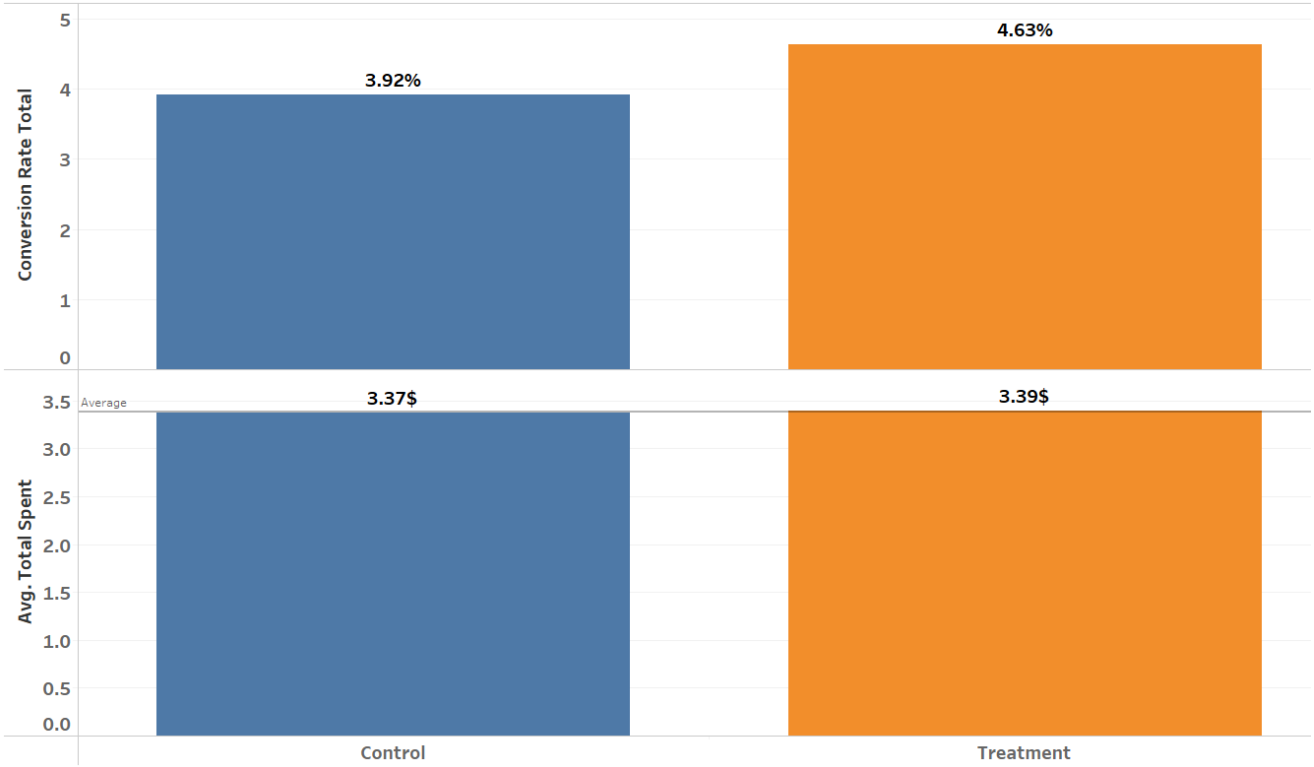
In practical terms, this suggests that the treatment group's average spending could be slightly higher or slightly lower than the control group, but **this difference is not strong enough to be considered statistically significant at the 95% confidence level**. Therefore, based on the results of this analysis, **we cannot confidently say that the treatment has led to a significant change in user spending**.

If we had observed a confidence interval that did not include zero, we would have been more confident in asserting a significant difference between the two groups.

Overall, this analysis provides us with valuable information to consider when evaluating the impact of the treatment on user spending. It's important to remember that while the results may not show a significant difference, they contribute to our understanding of how the treatment influences user behaviour.

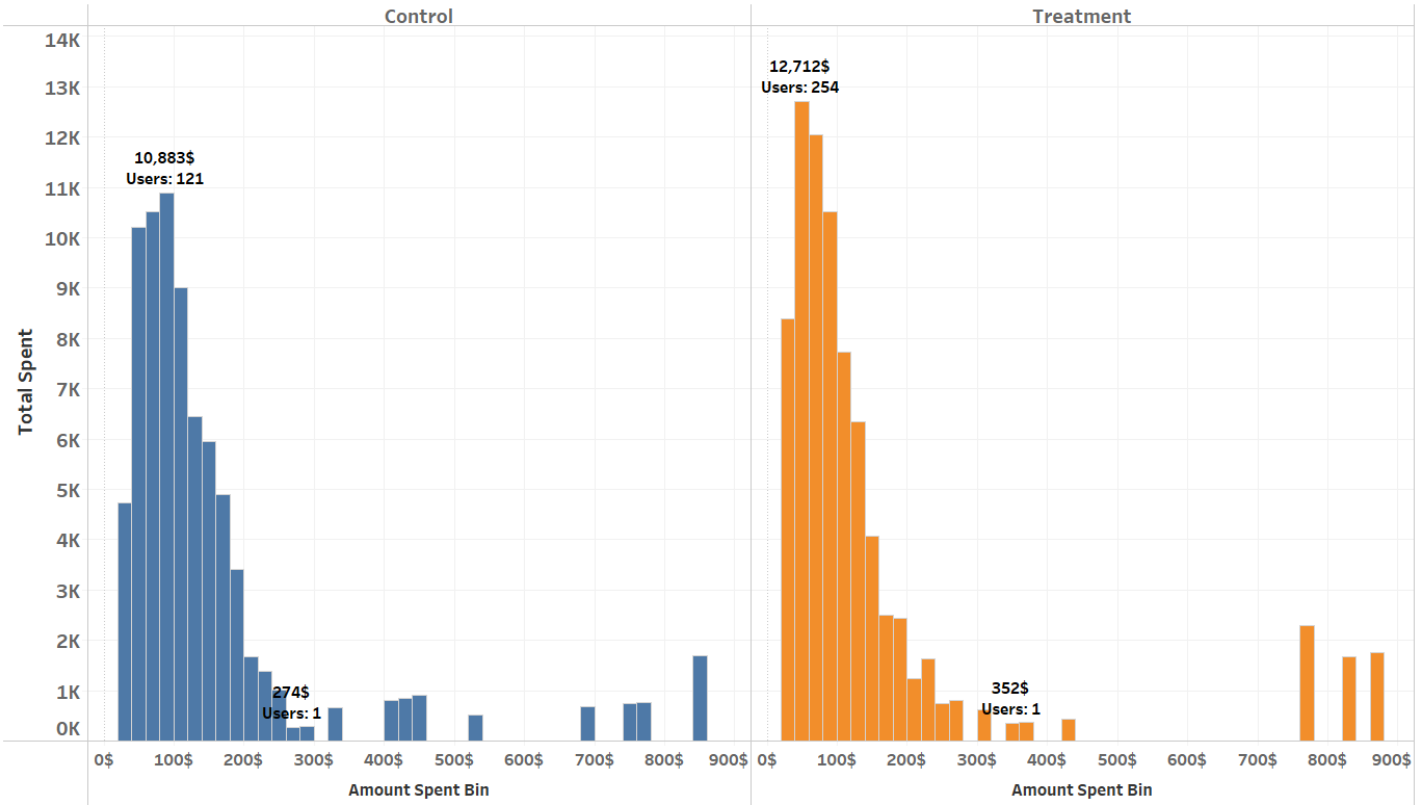
Visualizations

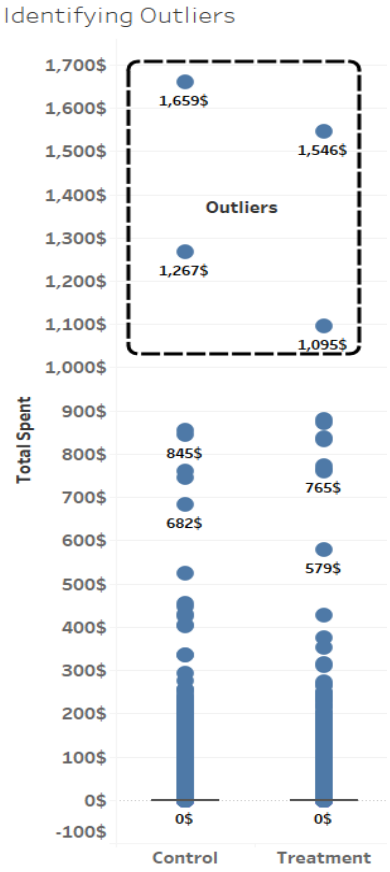
Conversion Rate, Average Amount Spent Between Groups



- Similar average spending
- Big difference in the conversion rates among the groups (favors the treatment group) where **A: 3.9%** and **B: 4.6%** conversion rates.
- Though there is a significant difference in the conversion rate, this difference did not led to a significant increase in the total revenue or average amount spent per user.

Amount Spent Distribution Between Groups

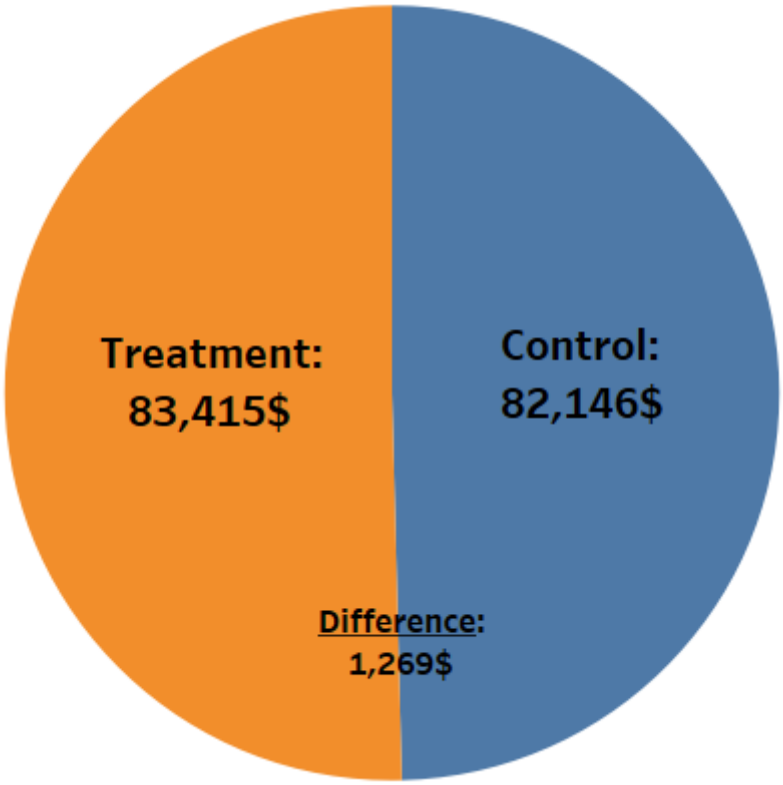




To analyze the amount spent between the groups we excluded outliers where the amount spent was greater than 1,000$. We chose to exclude > 1,000$ because only 2 users from group A and B (4 total) spent over 1,000$ in a purchase. The majority of the amount spent falls under 1,000$.

- Both groups follow similar distribution, with slight difference in group B where it seems they were spending more than group A. However, the A\B test results shows that the average amount spent between the groups is no statistically significant.
- The majority of the spending amounts for **group A** lies in **between 80-100.**
  - **10,883$** were spent by **121 users**.
- The majority of the spending amounts for **group B** lies in between **40-60.**
  - **12,712$** were spent by **254 users**.

Distribution of Revenue Between Groups



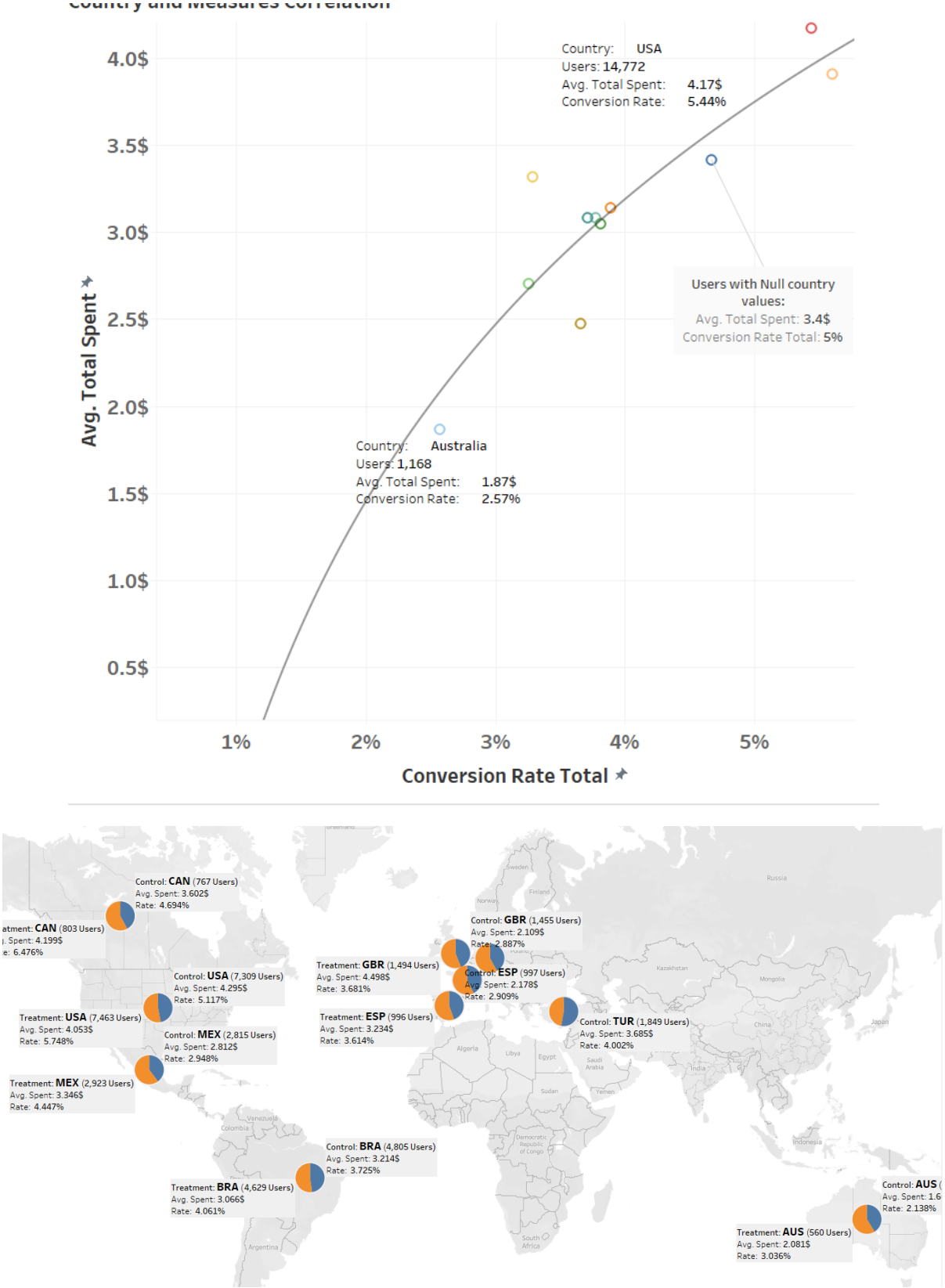
- Group A (Control Group) Revenue: **$82,146**
- Group B (Treatment Group) Revenue: **$83,415**

Key Observations:

- Group B (Treatment) generated slightly higher revenue compared to Group A (Control) in the A/B test.
- The **revenue difference** between the two groups is **$1,269** in favor of Group B.
- A hypothesis test was conducted to determine the statistical significance of this difference.
- The **p-value** obtained from the test was **0.9439**, indicating that the observed revenue difference is **not** statistically significant.
- Therefore, there is insufficient evidence to conclude that the treatment applied to Group B had a significant effect on revenue when compared to Group A.
- The result suggests that any difference in revenue may be due to random variation rather than the treatment.

In summary, while Group B showed higher revenue in the A/B test, the statistical analysis did not find this difference to be statistically significant.

Correlation Between Country and Avg. Amount Spent, Conversion Rate

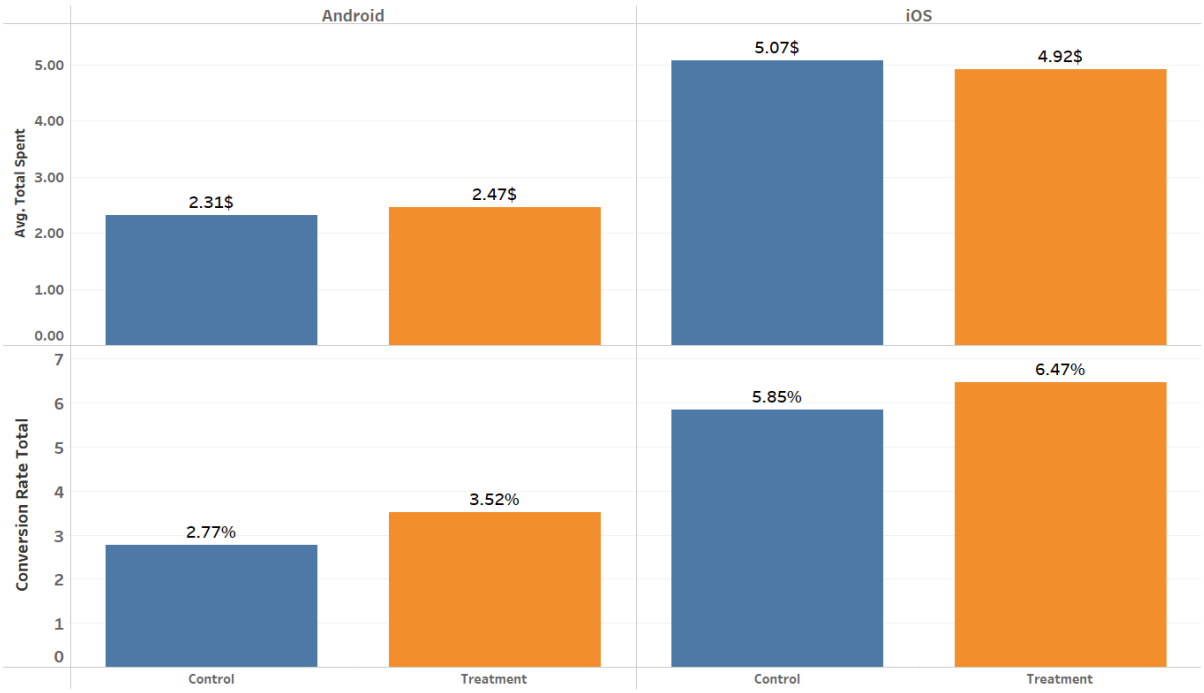


- **Australia** has the lowest conversion rate and average amount spent: **2.5% and 1.86$** respectively.
- **Canada** and the **USA** are the top performers both in conversion rates and average amount spent.
- **The US** had a higher **average spent** per user of **4.1$** (oppose to **Canada's: 3.9$**)
- **Canada** has a higher **conversion rate** of **5.6%** (oppose to the **US's 5.4%**)
- **634** had **NULL** values in country. We might want to delve into why this happened.

Measures by Gender

- In both groups the conversion rate favors the Females. With above 5%, women are more convertible than men and “other”.
- Avg. amount spent per user was lower with females.
- Average amount spent for Other gender was roughly the same between the groups.
- **Conversion rate** for men was higher in **group B (3.7%)** where in **group A** it was **2.6%.** a 1.1% difference.
- Male users had the lowest average spending in both groups.
- Overall, it seems like women tend to purchase more frequently and spend more average wise.

Measures by Device



- On Android devices we saw similar results in average amount spent per user in both groups.
- Treatment group had a higher **conversion rate** in both android and iOS users. The difference was **0.07% on android** and on **iOS it was 0.6%.**
- iOS devices in all metrics if we compare their results with android users. An average of 5$ spent per users in both groups. The conversion rate is a lot higher in both control and treatment groups (5.8% and 6.4% respectively)
- Looking at the data, it's safe to assume that iOS users tend to spend and convert more on the site regardless of the A\B test.
- iOS users in the treatment group (**4.92$**) were spending less the control group (**5.07$**) on average. Meaning that the banner perhaps negatively impacted the user's spending behavior.

Recommendations, Further Guidance, Final Thoughts

After conducting a comprehensive analysis of the A/B test, **it is recommended that the we should refrain from launching the new banner** on the mobile website. The analysis revealed several key findings that indicate the new feature **did not significantly impact revenue and average amount spent per user**:

1. **Conversion Rate Effect:** The A/B test did show a significant increase in conversion rates in favor of the treatment group exposed to the new banner. **However**, this increase in conversion rates did not translate into a substantial difference in revenue or average spending.

2. **Average Amount Spent:** The analysis found **no statistically significant difference in the average amount spent per user** between the treatment and control groups. This suggests that while more users may have converted in the treatment group, they did not spend significantly more on average.

3. **Visualizations:** Visualizations of the data, including spending distributions, country-wise analysis, gender-based analysis, and device-based analysis, consistently showed that the new feature did not lead to a meaningful increase in revenue or average spending.

4. **Novelty Effect:** The decline in spending over time, as indicated by a clear pattern in the data, suggests the possibility of a novelty effect. **Users may have initially engaged with the new feature out of curiosity but did not sustain higher spending levels**.

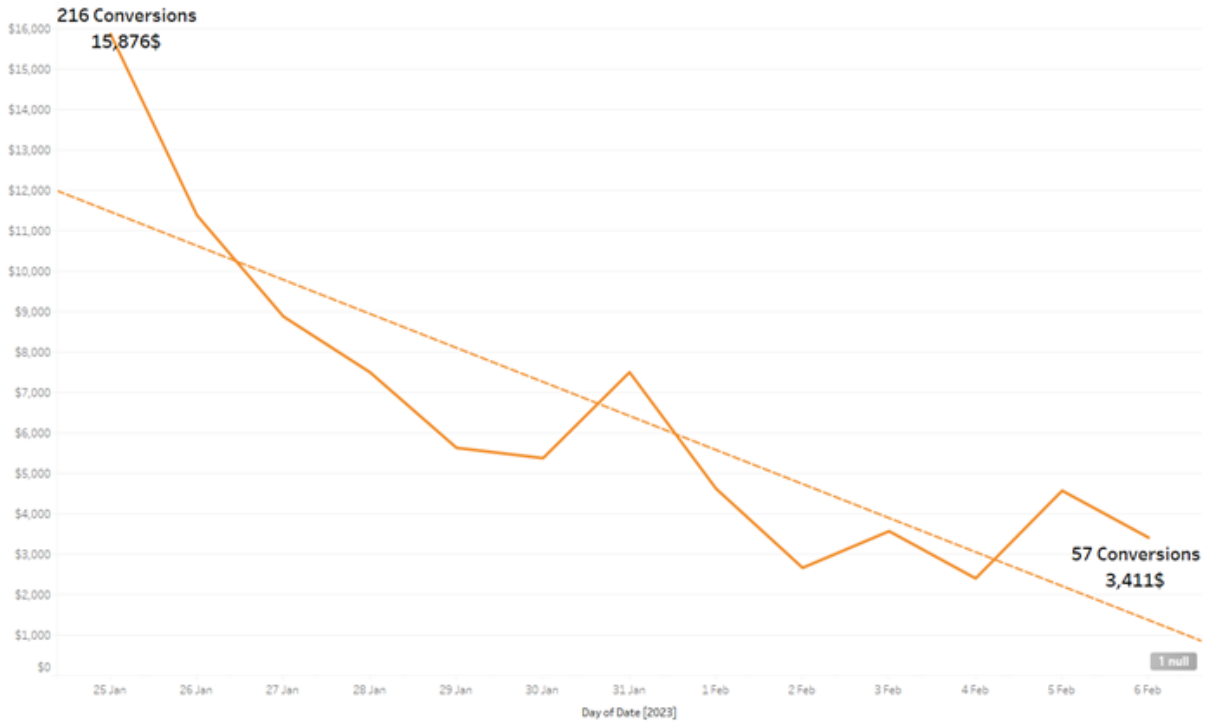
5. **Power Analysis:** The A/B test was conducted with a much larger sample size than needed, enhancing the reliability of the results. Despite the large sample size, **no significant impact on revenue or average spending was observed**.

In conclusion, while the A/B test did show an increase in conversion rates, this increase did not lead to a substantial improvement in revenue or average amount spent per user. Considering the lack of a significant financial benefit and the potential presence of a novelty effect, it is advisable to hold off on implementing the new feature across the board. Instead, the company may want to explore alternative strategies for driving revenue growth that are more likely to yield a substantial return on investment.

Novelty Effect

Users might behave differently when the treatment is new, which is called a novelty effect.

While inspecting the key metric of amount spent per day, we noticed a clear decline over the days of the experiment.



- **We can see a clear pattern where group B had a decrease in the amount spent for a day as well as # of conversions.**
- Trend line was added to confirm that the trend is clearly declining.
- The best days were **15,876$** of total sales and **216** conversions in the very start of the experiment.
- The worst days were at the end of the experiment, about a month after where group B had 57 conversions with a total of 3,411$ in sales.
- This is a clear sign that a novelty might be involved in this situation.

Power Analysis

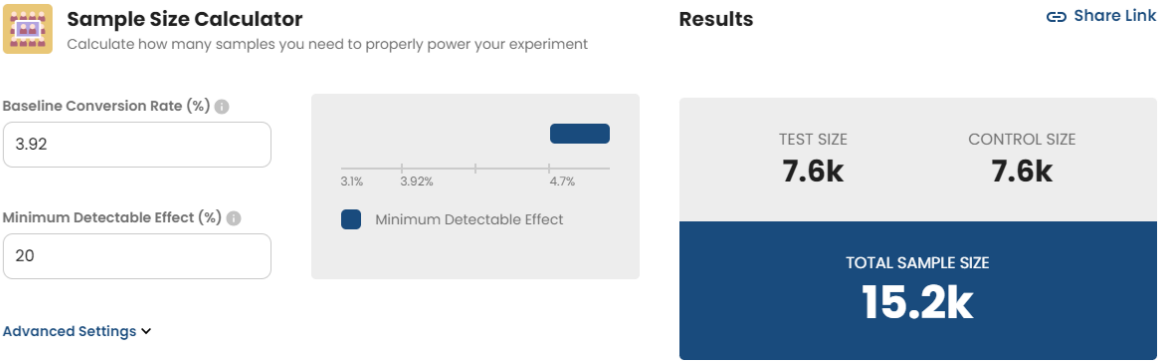
A power analysis helps us understand the necessary sample size in order to achieve our desired minimum detectable effect and statistical power. If we find that we did not have enough sample size for our test to be sufficiently sensitive, we could recommend that we run the test again at a larger scale.

Parameters Used:

- Baseline Conversion Rate of Group A: **3.92%**
- Minimum Detectable Effect (MDE): **20%**
- Statistical Power (1-β): **0.80**
- Significance Level (α): **0.05**

Results of Power Analysis:

The power analysis determined that a sample size of **7,600 participants** is needed for each group (Group A and Group B) for a total of **15,200 participants** to achieve a statistical power of 0.80 at a significance level of 0.05. This sample size calculation ensures that the test has an 80% probability of detecting a minimum detectable effect of 20% if it truly exists.



Actual Sample Sizes in the A/B Test:

In practice, the A/B test was conducted with a significantly larger sample size than initially calculated. Each group (Group A and Group B) consisted of **over 20,000 participants**, resulting in a total sample size exceeding **40,000 participants**.

Comparison:

The actual sample sizes used in the A/B test far exceeded the minimum requirements established by the power analysis. While the power analysis recommended a sample size of **15,200 participants**, the test was conducted with a total sample size of over **40,000 participants** for both groups. This sizable sample size not only ensures the desired statistical power but also enhances the reliability and robustness of the test results.

Conclusion:

The use of a substantially larger sample size in the A/B test underscores our commitment to achieving highly reliable and accurate results. This larger sample size enhances the test's ability to detect even subtle effects and reduces the risk of false negatives. As a result, we can have increased confidence in the validity of the conclusions drawn from the A/B test.

Appendix

Tableau:

- **Main Workbook with Dashboard:** https://public.tableau.com/views/TableauWorkbook_16927024093220/Dashboard?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link
- **Novelty Effect Workbook:** https://public.tableau.com/views/TableuNoveltyEffect/Sheet2?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

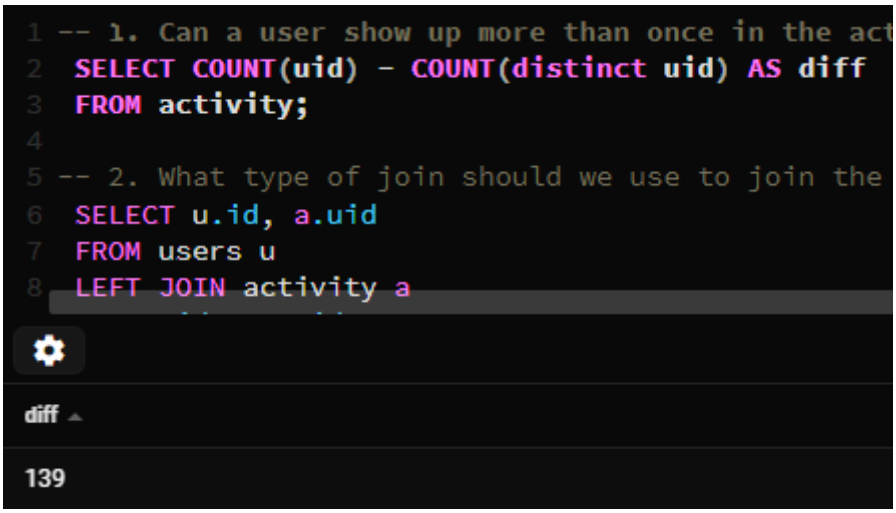
Calculations with Sheets:

- https://docs.google.com/spreadsheets/d/1eZ9BVNC21zsEK_5u_PFZuTQNqxDtw7QPDVwoddy0VPg/edit?usp=sharing

SQL Queries:

-- 1. Can a user show up more than once in the activity table? Yes or no, and why?

```
SELECT COUNT(uid) - COUNT(distinct uid) AS diff
 FROM activity;
```

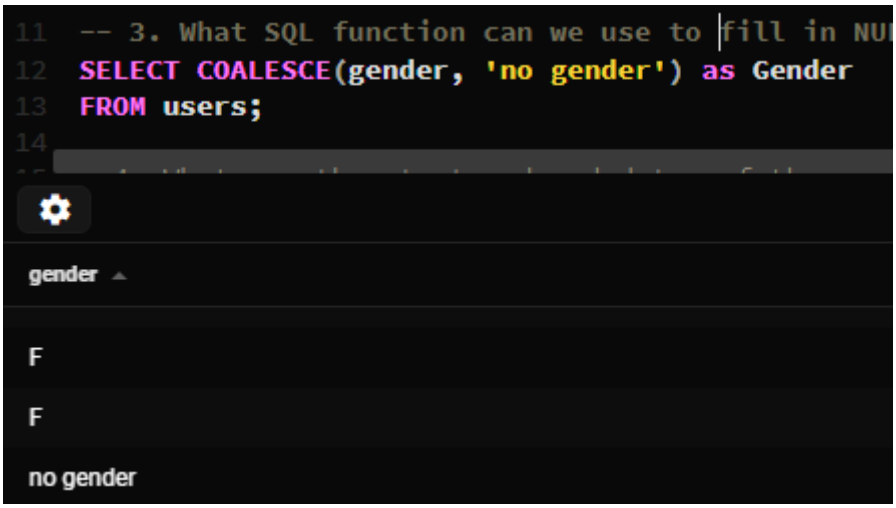


-- 2. What type of join should we use to join the users table to the activity table?

```
SELECT u.id, a.uid
FROM users u
LEFT JOIN activity a
        ON u.id = a.uid;
```

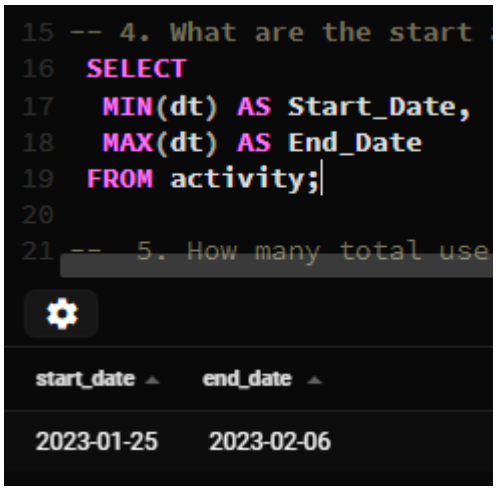
-- 3. What SQL function can we use to fill in NULL values?

```
SELECT COALESCE(gender, 'no gender') as Gender
FROM users;
```



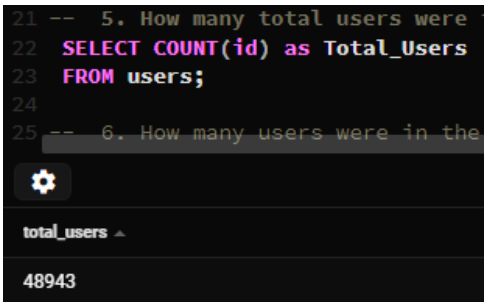
-- 4. What are the start and end dates of the experiment?

```
SELECT
        MIN(dt) AS Start_Date,
 MAX(dt) AS End_Date
FROM activity;
```



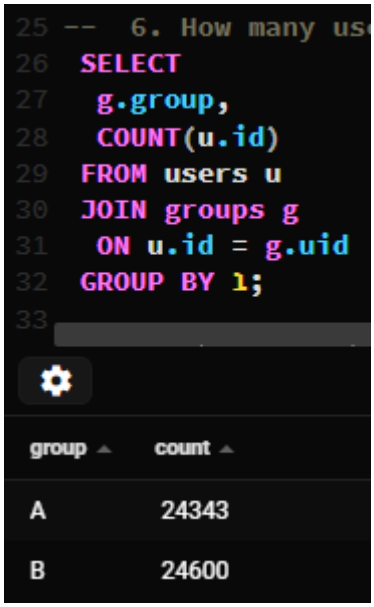
-- 5. How many total users were in the experiment?

```
SELECT COUNT(id) as Total_Users
FROM users;
```



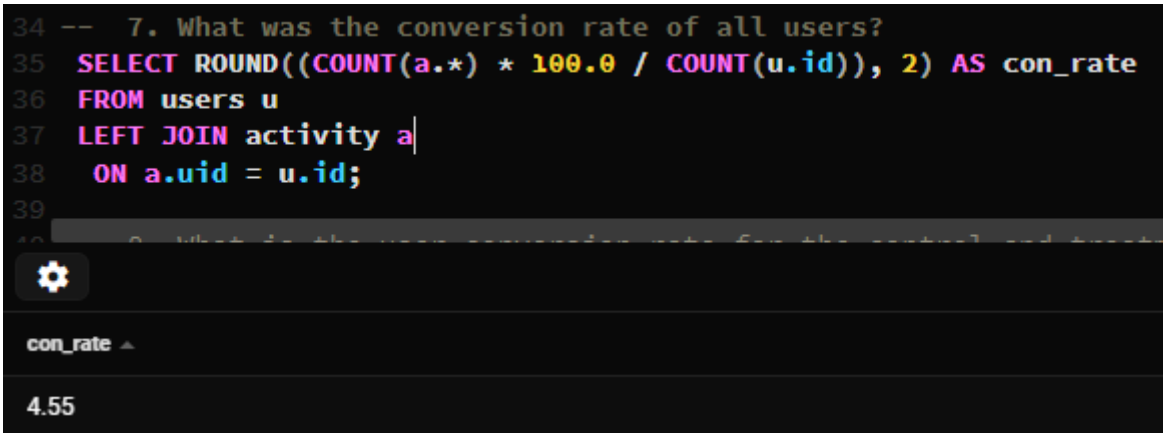
-- 6. How many users were in the control and treatment groups?

```
SELECT
        g.group,
        COUNT(u.id)
FROM users u
JOIN groups g
        ON u.id = g.uid
GROUP BY 1;
```



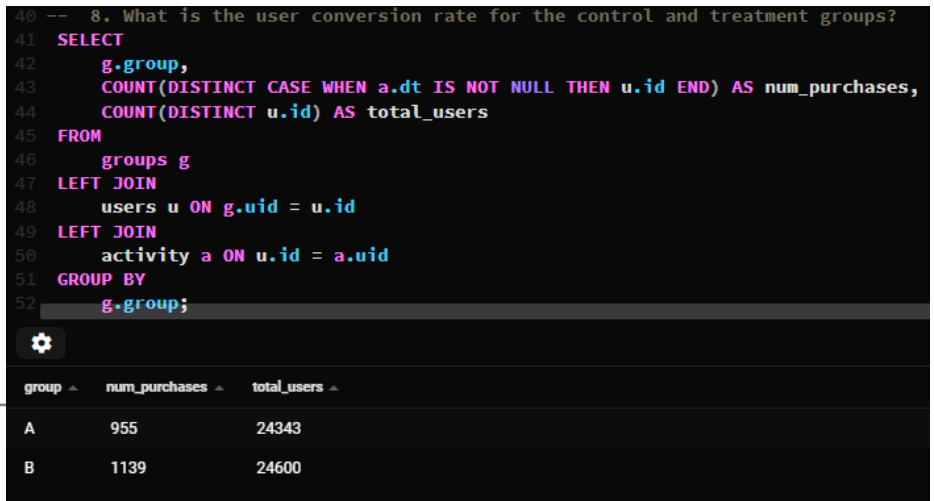
-- 7. What was the conversion rate of all users?

```
SELECT ROUND((COUNT(a.*) * 100.0 / COUNT(u.id)), 2) AS con_rate
FROM users u
LEFT JOIN activity a
        ON a.uid = u.id;
```



-- 8. What is the user conversion rate for the control and treatment groups?

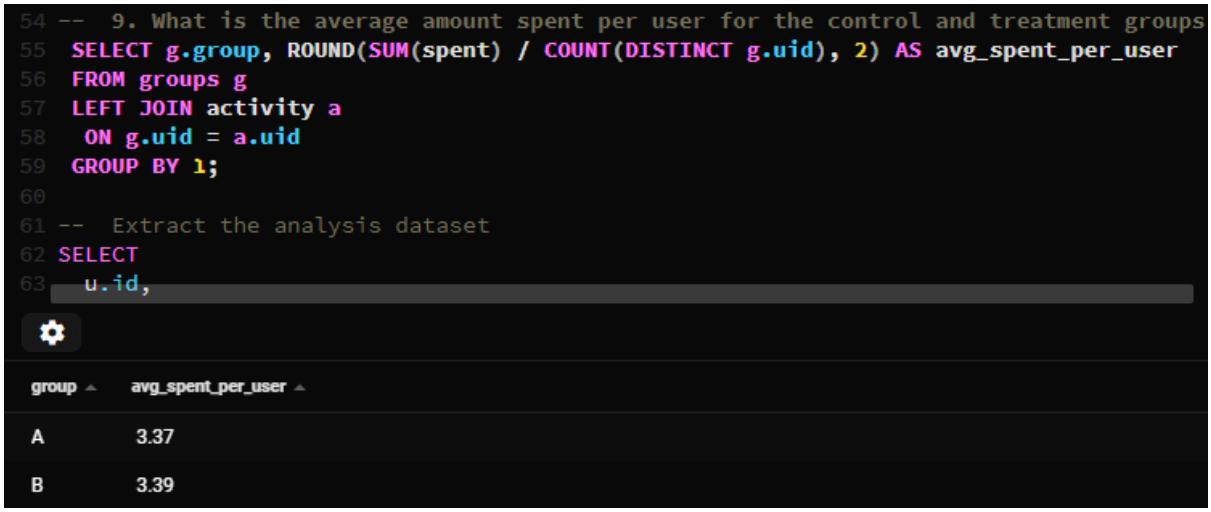
```
SELECT
    g.group,
    COUNT(DISTINCT CASE WHEN a.dt IS NOT NULL THEN u.id END) AS num_purchases,
    COUNT(DISTINCT u.id) AS total_users
FROM
    groups g
LEFT JOIN
    users u ON g.uid = u.id
LEFT JOIN
    activity a ON u.id = a.uid
GROUP BY
    g.group;
```



g.group;

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

```
SELECT g.group, ROUND(SUM(spent) / COUNT(DISTINCT g.uid), 2) AS avg_spent_per_user
FROM groups g
LEFT JOIN activity a
        ON g.uid = a.uid
GROUP BY 1;
```



-- Extract the analysis dataset

```
SELECT
    u.id AS user_id,
    g.group AS test_group,
    CASE WHEN u.country IS NULL THEN 'No Country' ELSE u.country END,
    CASE WHEN u.gender IS NULL THEN 'No Gender' ELSE u.gender END,
    CASE WHEN g.device IS NULL THEN 'No Device Type' ELSE g.device END,
    CASE WHEN a.spent > 0 THEN 'TRUE' ELSE 'FALSE' END AS converted,
    CASE WHEN SUM(a.spent) IS NULL THEN 0 ELSE SUM(a.spent) END
FROM users u
LEFT JOIN
    groups g ON u.id = g.uid
LEFT JOIN
                activity a ON u.id = a.uid
GROUP BY
    u.id, g.device, g.group, a.spent
```

```sql
--  Extract the analysis dataset
SELECT
    u.id AS user_id,
    g.group AS test_group,
    CASE WHEN u.country IS NULL THEN 'No Country' ELSE u.country END,
    CASE WHEN u.gender IS NULL THEN 'No Gender' ELSE u.gender END,
    CASE WHEN g.device IS NULL THEN 'No Device Type' ELSE g.device END,
    CASE WHEN a.spent > 0 THEN 'TRUE' ELSE 'FALSE' END AS converted,
    CASE WHEN SUM(a.spent) IS NULL THEN 0 ELSE SUM(a.spent) END
FROM users u
LEFT JOIN
    groups g ON u.id = g.uid
LEFT JOIN
    activity a ON u.id = a.uid
GROUP BY
    u.id, g.device, g.group, a.spent
```

user_id	test_group	country	gender	device	converted	sum
1000041	A	CAN	M	A	FALSE	0
1000042	A	BRA	F	I	FALSE	0
1000043	B	No Cou...	F	A	FALSE	0
1000044	B	USA	No Gender	A	FALSE	0
1000045	B	USA	F	I	TRUE	51.58
1000046	A	BRA	M	I	FALSE	0
1000047	A	USA	F	A	FALSE	0
1000048	A	CAN	F	I	FALSE	0

-- Novelty Effect CSV

```sql
SELECT u.id,
        a.dt,
    g. group,
    ROUND(a.spent, 2) AS amount_spent
FROM users u
JOIN
        groups g ON u.id = g.uid
JOIN
        activity a ON u.id = a.uid
```