OYIEYO SILAS
First Mastery Project!

Company: **Data GloBox**

Sprint 1. Project Tasks WEEK 01-08/10/2023

In this first sprint, you will become familiar with the dataset and extract the data to use for the rest of the project using SQL queries. Then, you'll import that data into Tableau to visualize the results.

Answer the quiz questions to understand the database

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

```
Yes, the user can visit the app several times to make purchases.
Total users in the experiment are 49032 and the distinct users are
48943
SELECT u.id, COUNT(u.id)
FROM users U
LEFT JOIN activity a
ON u.id = a.uid
LEFT JOIN groups g
ON g.uid = u.id
GROUP BY u.id
ORDER BY 2 DESC
```

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

Left joint. Since we need not to lose the data from the left table.

- A LEFT JOIN returns all the rows from the left table (the table mentioned before the LEFT JOIN keyword) and the matched rows from the right table (the table mentioned after the LEFT JOIN keyword).

- If there is no match for a particular row in the right table, the result will still include that row from the left table with NULL values in columns from the right table.
- It preserves all the data from the left table and includes matching data from the right table.

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

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

```
Coalecse FUNCTION:
SELECT COALESCE(column_name, 'Default Value') AS new_column
FROM your_table;

SELECT *, COALESCE(a.uid,10) AS n_uid
FROM users u
LEFT JOIN activity a
ON u.id = a.uid
```

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

```
WITH experiment AS(
SELECT *
FROM users U
LEFT JOIN activity a
ON u.id = a.uid
LEFT JOIN groups g
ON g.uid = u.id
  )
SELECT MIN(dt) AS experiment start_date, MAX(dt) AS
experiment_end_date
FROM experiment;
```

5. How many total users were in the experiment?

```
SELECT COUNT(DISTINCT u.id)
FROM users u
LEFT JOIN groups g
ON g.uid = u.id
WHERE g.join_dt IS NOT NULL;
```

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

```
SELECT COUNT(DISTINCT u.id), COUNT(DISTINCT g.uid), g.group
FROM users u
LEFT JOIN activity a
ON a.uid = u.id
LEFT JOIN groups g
ON g.uid = u.id
WHERE g.join_dt IS NOT NULL
GROUP BY g.group;
OR
SELECT g.group,COUNT(DISTINCT u.id)
FROM users u
LEFT JOIN groups g
ON g.uid = u.id
WHERE g.join_dt IS NOT NULL
GROUP BY 1
```

7. What was the conversion rate of all users?

```
SELECT ROUND(CAST(COUNT(DISTINCT a.uid)AS
DECIMAL(10,4))/COUNT(DISTINCT u.id),4)
FROM users u
LEFT JOIN activity a
ON a.uid = u.id
LEFT JOIN groups g
ON g.uid = u.id ON g.uid = u.id
```

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

```
SELECT g.group,ROUND(CAST(COUNT(DISTINCT a.uid)AS
DECIMAL(10,4))/COUNT(DISTINCT u.id),4)
FROM users u
LEFT JOIN activity a
ON a.uid = u.id
LEFT JOIN groups g
```

```
ON g.uid = u.id
GROUP BY 1
```

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

```
SELECT SUM(a.spent)/COUNT(DISTINCT u.id)
FROM users u
LEFT JOIN activity a
ON a.uid = u.id
LEFT JOIN groups g
ON g.uid = u.id
GROUP BY g.group
```

10. Why does it matter to include users who did not convert when calculating the average amount spent per user?

Yes. It contains useful information when conclusion and decision is drawn about the experiment.

Extract the analysis dataset

1. Write a SQL query that returns: 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+).

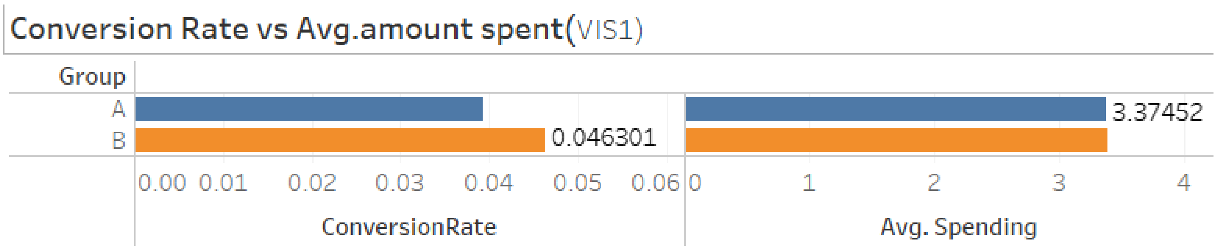
```
Un-Coalesced data
SELECT u.id, u.country, u.gender, a.device,
       g.group, SUM(a.spent) AS total_spent
FROM users U
LEFT JOIN activity a
ON u.id = a.uid
LEFT JOIN groups g
ON g.uid = u.id
GROUP BY 1,4,5
```

2. Download the data as a CSV. This is what you will use in the next phase of analysis during the second sprint.

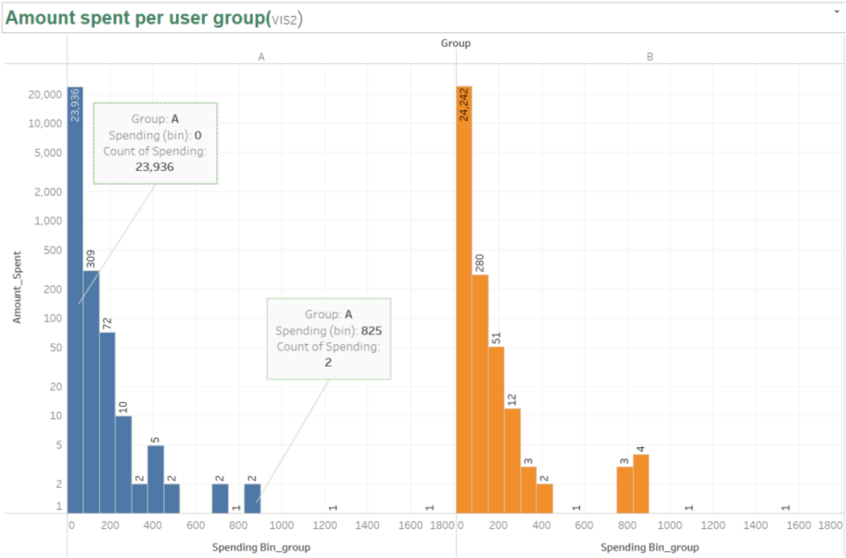
```
Downloaded Grouped CSV
SELECT u.id, COALESCE(u.country,'NO_REG') AS country,
COALESCE(u.gender, 'NO_G') AS gender,
       COALESCE(a.device,'NO_DEV') AS device_type, g.group,
       CASE WHEN a.spent > 0 THEN 'YES' ELSE 'No' END AS
conversion,
       COALESCE(SUM(a.spent), 0) AS Spending
FROM users U
LEFT JOIN activity a
ON u.id = a.uid
LEFT JOIN groups g
ON g.uid = u.id
GROUP BY 1,4,5,6
```

Visualize the results in Tableau

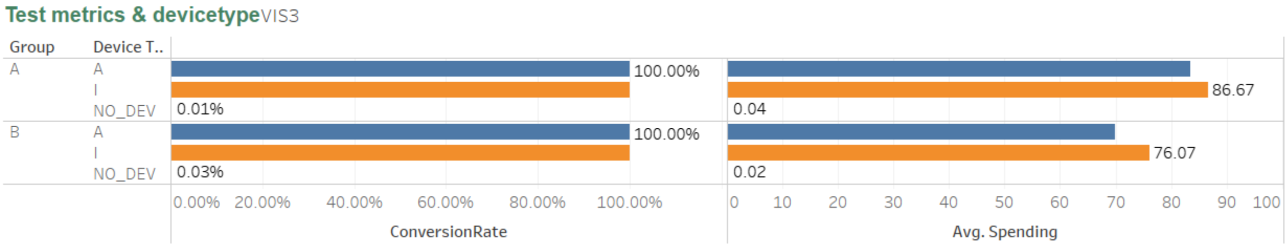
1. Create a visualization to compare the conversion rate and average amount spent between the test groups.



2. Visualize the distribution of the amount spent per user for each group.
   Used the Logarithmic scale to stabilize the axis. Exclude the bins with no values.

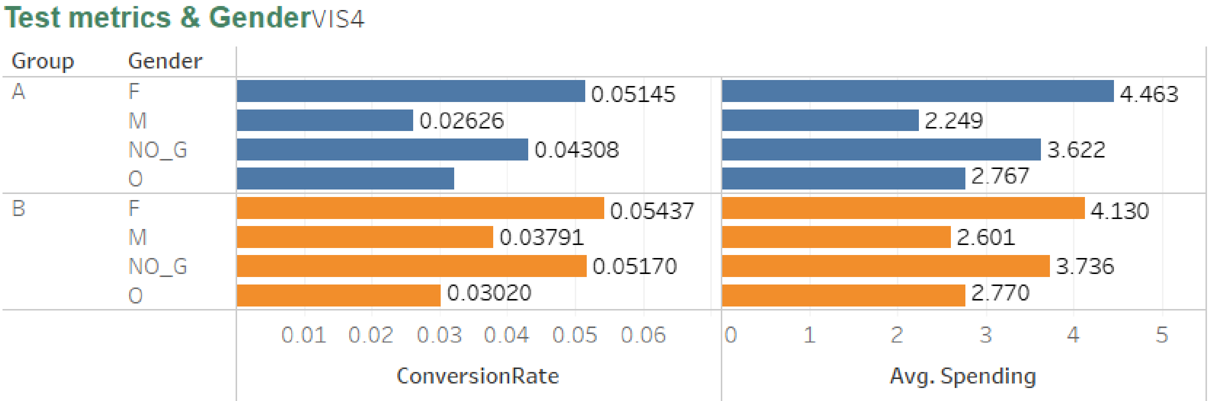


3. Create visualizations to explore the relationship between the test metrics (conversion rate and average amount spent) and the user's device.

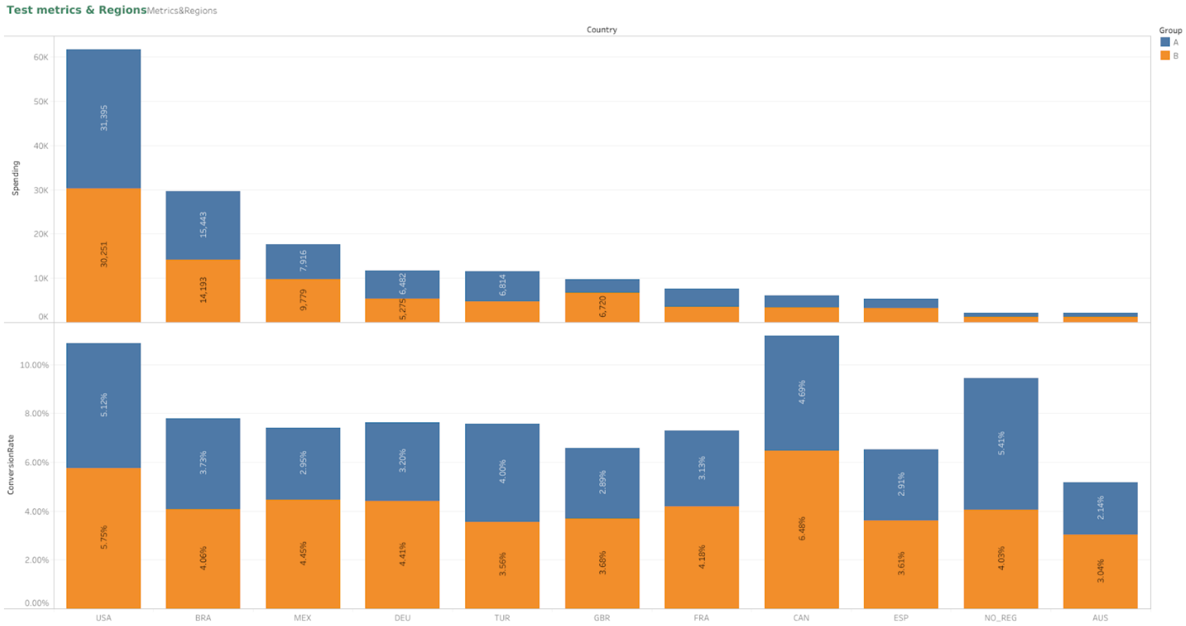


4.

5. Create visualizations to explore the relationship between the test metrics (conversion rate and average amount spent) and the user's gender.



6. Create visualizations to explore the relationship between the test metrics (conversion rate and average amount spent) and the user's country.



Sprint 2. Project Tasks WEEK 09-16/10/2023

Importing your analysis dataset CSV from the previous sprint into a new spreadsheet.

1. **Conduct a hypothesis test to see whether there is a difference in the conversion rate between the two groups. What are the resulting p-value and conclusion?**

   Use the normal distribution and a 5% significance level. Use the pooled proportion for the standard error.

   Steps:

   ‣

   Determine the null and alternative hypothesis

   H0; Conversion rate of A equals to conversion rate of B, (H0:Ac=Bc)
   H1. Conversion rate of A is not equals to conversion rate B,(H1: Ac!=Bc)

   Determine what type of test you are using

   Conversion rate is in proportion form. Two samples and with two tailed test.
   ‣Z-test is used. Produces the p-values. P-values cut-off 0.01 or 0.05.When less than 5% reject the H0.

   When testing only > or < then test has 1 tail. Only one side is expected.
   H0 equals to /= then 2 tail (Above or below the mean. Either +- is expected.
   NOTE: Measuring same observation at different times = type =1
   Measuring different observations with same variance =type =2
   Measuring different observations with different variances. Type =3

   Calculate the test statistic

   ‣ *Based on the formula on the cheat sheet. Test Statistic= 3.864*

$$T = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$\hat{p} = \frac{\hat{p}_1 * n_1 + \hat{p}_2 * n_2}{n_1 + n_2}$$

Calculate the p-value

$$2 * P(Z > |T|)$$

Answer; 0.000111545221

Draw a conclusion about the hypothesis

The value (0.000111545221) is statistically significant. The null hypothesis (H0; Conversion rate of A equals to conversion rate of B, (H0:Ac=Bc)) is then rejected.

2. **What is the 95% confidence interval for the difference in the conversion rate between the treatment and control (treatment-control)?**

Determine what type of interval you are computing

‣*The of interval for computation is (Two sample Interval). Two different samples in proportion.*

Calculate the sample statistic

‣*Sample statistic = ^p1=^p2. (0.03923099043- 0.04630081301)*

Answer; 0.00707

Calculate the standard error

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

‣

N1 = 24343, n2 = 24600

Answer; 0.00183

Calculate the critical value

‣*Code for critical value in spreadsheet is (=NORMINT($\frac{alpha}{2}$, 0,1)).Limits of >0,<1*

Answer; 196

Construct the interval

$$Interval = (Sample\ standard \pm (crtical * standard\ Error))$$

Answer; (0.00347, 0.01066)

3. **Conduct a hypothesis test to see whether there is a difference in the average amount spent per user between the two groups. What are the resulting p-value and conclusion?**

Use the t distribution and a 5% significance level. Assume unequal variance.

Determine the null and alternative hypothesis

H0: Avgspending(A) = Avgspending(B)

H1: Avgspending(A) != Avgspending(B)

Determine what type of test you are using

Two-sample t-test for mean difference

Calculate the test statistic

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Answer; -0.0707. Values are available on spreadsheet.

Calculate the p-value
*=(2 * (1 - NORMDIST(ABS(-0.0707), 0, 1, TRUE))).*
Answer; 0.9436365208

Draw a conclusion about the hypothesis
P-Value is statistically insignificance. Null hypothesis is accepted. There is no different in mean amount spent.

4. **What is the 95% confidence interval for the difference in the average amount spent per user between the treatment and the control (treatment-control)?**

Use the t distribution and assume unequal variance.
▸
Determine what type of interval you are computing
Two- sample t-interval for mean difference in proportions.
Calculate the sample statistic
Sample statistics is the difference between two sample means. Answer; 0.0164
Calculate the standard error.

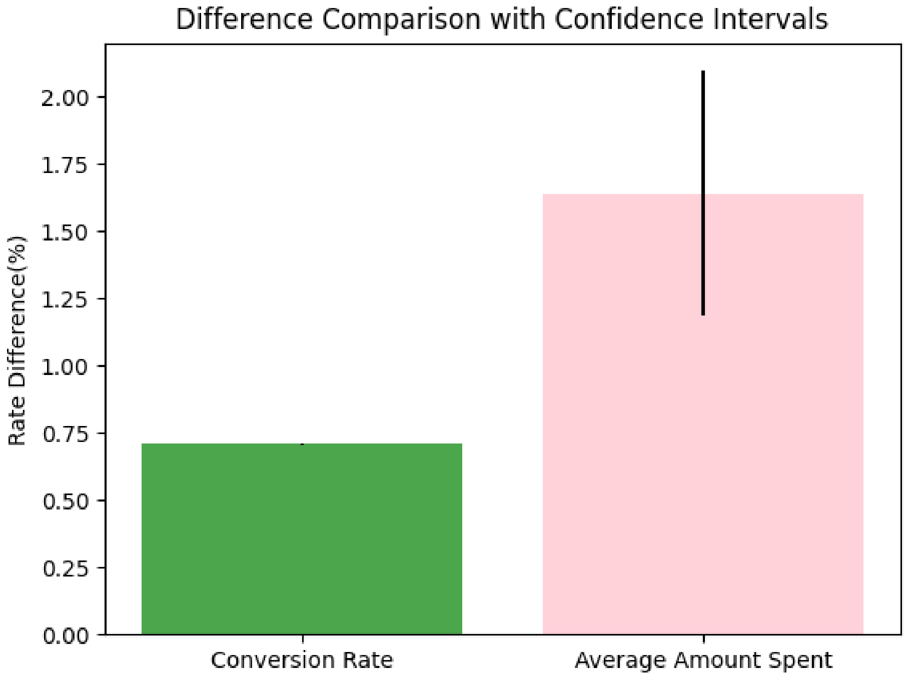
From the formula $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ Answer; 0.2321

Calculate the critical value.
Critical values for two-sample test can be calculated in the spreadsheet by the formula.
=T.INV.2T(0.05, 48941) Answer; 1.960012458
Construct the interval

$$critical_{interval} = (sample\ statistic \pm (s - error * criticalvalue))$$

Answer; (-0.4385188915, 0.4713188915)

Advanced Tasks
Visualize the Confidence Intervals
Plot the confidence intervals for the difference in conversion rate and the difference in the average amount spent between the two groups



Check for Novelty Effects

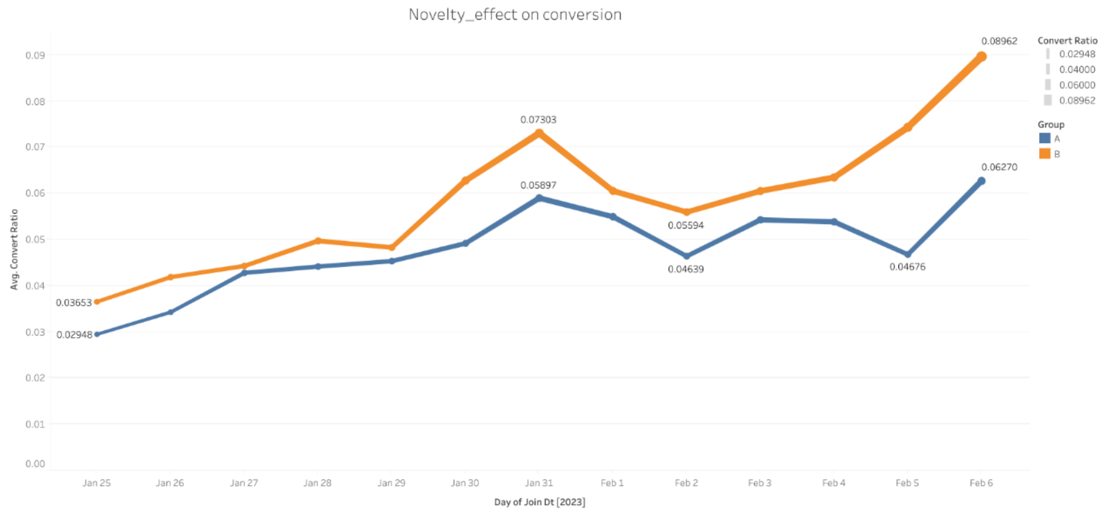
Users might behave differently when the treatment is new, which is called a novelty effect.

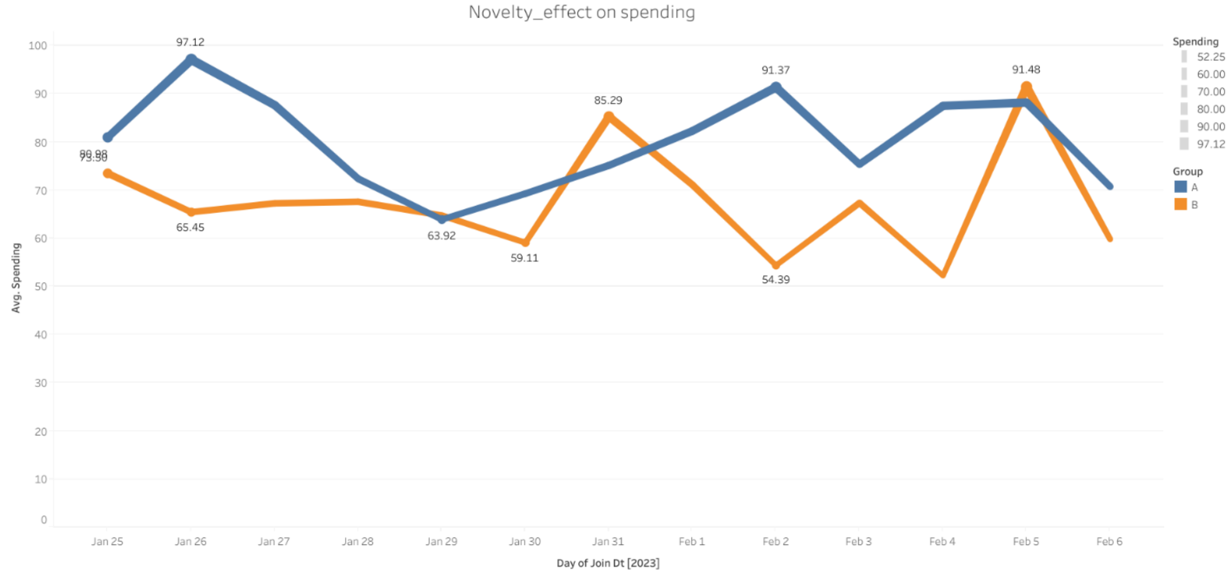
```
-- Step 1: Creating join_dt_agg
CREATE TEMPORARY TABLE join_dt_agg AS
SELECT g.join_dt, g.group, COUNT(DISTINCT u.id) AS user_count
FROM users u
INNER JOIN groups g
ON u.id = g.uid
GROUP BY 1, 2;


-- Step 2: Creating convert_dt_agg
CREATE TEMPORARY TABLE convert_dt_agg AS
SELECT a.dt, g.group, AVG(a.spent) AS spending, COUNT(DISTINCT
a.uid) AS converted_user_count
FROM groups g
LEFT JOIN activity a
ON g.uid = a.uid
GROUP BY 1, 2;

--step 3
SELECT j.join_dt,j.group,
CAST(c.converted_user_count AS DECIMAL(10,2)) / j.user_count AS
convert_ratio,
c.spending
FROM join_dt_agg j
```

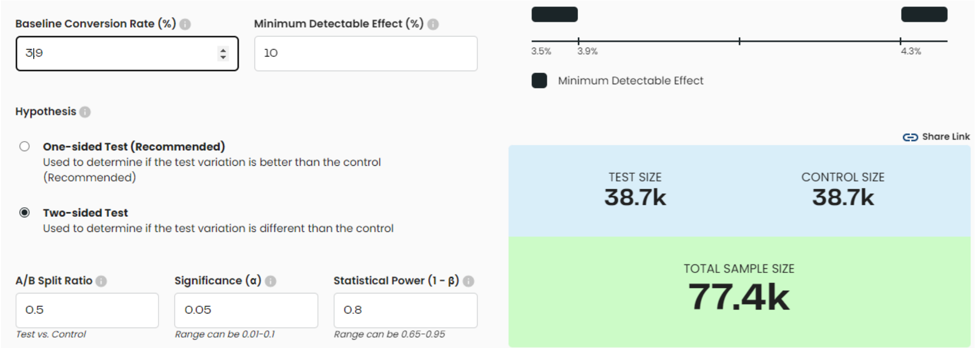
```
INNER JOIN convert_dt_agg c
ON j.join_dt = c.dt AND j.group=c.group;
```





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



Baseline Conversion Rate (%)

3|9

Minimum Detectable Effect (%)

10

Hypothesis

○ **One-sided Test (Recommended)**
Used to determine if the test variation is better than the control (Recommended)

● **Two-sided Test**
Used to determine if the test variation is different than the control

TEST SIZE 38.7k

CONTROL SIZE 38.7k

TOTAL SAMPLE SIZE 77.4k

A/B Split Ratio: 0.5 — *Test vs. Control*

Significance (α): 0.05 — *Range can be 0.01–0.1*

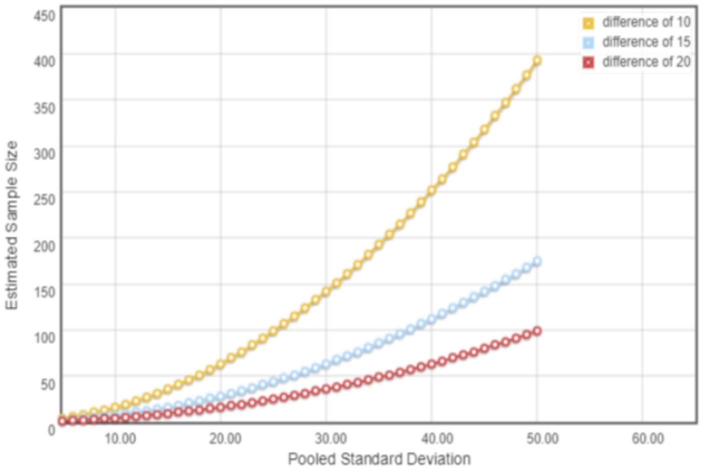
Statistical Power (1 – β): 0.8 — *Range can be 0.65–0.95*

The baseline conversion rate is 3.92%, control group. Selection of Minimum detectable effect of 15%. It’s realistic to have at least 12% increase in usage.

Sample Sizes for Comparing Two Means

Influence of Changing Input values on Sample Size Estimates

Figure created by Statulator beta: www.statulator.com/SampleSize/ss2M.html


difference of 10
difference of 15
difference of 20
Estimated Sample Size
Pooled Standard Deviation