DA106 Mastery Project I – Notes

1 Understanding the GloBox Database

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

- Hints
  - (1) Check if user IDs are repeated
    - Write an SQL query that counts the number of rows per user ID in the activity table. You can use the HAVING clause to see only users who show up more than once. Review the HAVING clause in this DataCamp lesson: Filtering grouped data.
  - (2) Reference the dataset information
    - Return to the Project Overview page in Campus and reference the dataset description to brainstorm why it would make sense for a user to show up only once, or more than once.

- Ad (1) SQL queries

  - (a)
    ```
    SELECT COUNT(uid) AS uid, COUNT (DISTINCT uid) AS count_dist_uid
    FROM activity;
    ```

  - (b)
    ```
    SELECT DISTINCT uid AS dist_uid, COUNT(uid) AS num_rows_uid
    FROM activity
    GROUP BY DISTINCT uid
    HAVING COUNT(uid) > 1;
    ```

  - (c)
    ```
    SELECT u.id, COUNT(a.uid) AS num_rows_uid
    FROM users AS u
    LEFT JOIN activity AS a
    ON u.id = a.uid
    GROUP BY u.id
    HAVING COUNT(a.uid) > 1;
    ```

  - SQL query (a), (b) and (c) outputs/results (cf. CSV files)

- Ad (2) Answer
  - It is said in the description of the dataset (Project Overview), especially regarding the activity table, that „**activity:** user purchase activity, containing 1 row per day that a user made a purchase“. Since the same user can buy several times on different days, the individual users can shop up more than once in the activity table. Users would only show up once in each table, if they did only one purchase during their entire lifetime. Also the database schema or entity relationship diagram respectively indicate that there is a 1:n-relationship between the primary key id in the users table and the foreign key uid in the activity table. Hence, the answer is yes, the reason is given above.

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

- Hints
    - Run queries joining the users and activity tables using an INNER JOIN and LEFT JOIN. Compare the results and see which makes more sense based on your reasoning: Do we expect all users to make a purchase? Do we want to include users regardless of whether they make a purchase?
- Answer
    - As already in the answer to question 1.1 applied we should use a LEFT JOIN to join the users and the activity table. We would only apply an INNER JOIN if we would expect ALL users to make a purchase. But since in the "Key things to note" section it is indicated that "not all users make a purchase", this does not hold true and an INNER JOIN is not applicable. Since only some users make a purchase the set of users is (not strictly/necessarily, but likely) bigger than the set of users making a purchase, we should apply a LEFT JOIN to join the users and the activity table. Since we want to calculate and compare conversion rates of different groups or subsets of users, we want to include ALL users, regardless of whether they make a purchase, hence a LEFT JOIN should be applied.

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

- Hints
    - Review different techniques for dealing with NULL values
        - You can review filtering and replacing NULL values in this DataCamp lesson: Dealing with nulls (https://campus.datacamp.com/courses/reporting-in-sql/cleaning-validation?ex=9). Once you identify the necessary function to fill in NULL values, think about how that might be relevant to this database. You will apply it in a future question.
- Answer
    - Since we want to calculate and compare conversion rates of different groups or subsets of users and some of those users did not make any purchases and hence – according to answer to question 1.2 -, will not show up and not have any values in the activity table, hence will have NULL values in the activity table, we might want to filter those NULLS out or calculate with them as 0 (zero) values. Hence, we would use the CASE WHEN SQL function to fill in the NULL values or replace them with 0 (zeroes) respectively.

1.4 What are the start and end dates of the experiment?

- Hints
    - (1) Determine which table(s) you need
        - Which table(s) contains the dates that a user joined the experiment? Return to the Project Overview page in Campus and reference the dataset description.
    - (2) Find the first and last dates in the data
        - Once you have determined which table(s) you need to query, use the MIN() and MAX() functions to find the first and last dates. You can review these functions in this DataCamp lesson: Summarizing data.
- Ad (1) Answer
    - The groups table contains the dates that a user joined the experiment.
    - Since the last purchase of the experiment could happen after the last join of the experiment, we should check for the latest date a purchase happened. Hence, the activity table is of some importance as well.

- Ad (2) SQL queries

    - (a)
      SELECT MIN(join_dt) AS start_date, MAX(join_dt) AS end_date
      FROM groups;

    - (b)
      SELECT MIN(dt) AS first_purchase, MAX(dt) AS last_purchase
      FROM activity;

    - SQL query (a) and (b) outputs/results (cf. CSV files) and markdowns:

        - (a)
start_date	end_date
2023-01-25	2023-02-06

        - (b)
first_purchase	last_purchase
2023-01-25	2023-02-06

1.5 How many total users were in the experiment?

- Hints
  - (1) Determine which table(s) you need
    - Which table(s) contain the information on which users were in the experiment? Return to the Project Overview page in Campus and reference the dataset description.
  - (2) Count the users
    - Once you have determined which table(s) you need to query, use the COUNT() function to count the number of users in the experiment. Make note of whether or not users are unique in the table(s), checking using DISTINCT. You can review the COUNT() function in this DataCamp lesson: Querying a database.
- Ad (1) Answer
  - The users and the groups tables contain all the users and the database schema or entity relationship diagram respectively indicate that there is a 1:1-relationship of the primary key id in the users table and the foreign key uid in the groups table. Hence, both tables include the information on which users were in the experiment.

- Ad (2) SQL queries

  - (a)
    SELECT COUNT (DISTINCT id) AS dist_users, COUNT (id) AS users
    FROM users;

  - (b)
    SELECT COUNT (DISTINCT uid) AS groups_dist_users, COUNT (uid) AS groups_users
    FROM groups;

  - (c)
    SELECT COUNT (DISTINCT uid) AS activity_dist_users, COUNT (uid) AS activity_users
    FROM activity;

  - SQL query (a), (b) and (c) outputs/results (cf. CSV files) and markdowns:

    - (a)
dist_users	users
48943	48943

    - (b)
groups_dist_users	groups_users
48943	48943

    - (c)
activity_dist_users	activity_users
2094	2233

1.6 How many users were in the control and treatment groups?

- Hints
    - (1) Determine which table(s) you need
        - Review the previous question: How many total users were in the experiment? Determine whether the tables you used contain information on which A/B test group that the user was in. If not, return to the Project Overview page on Campus and reference the dataset description.
    - (2) Count the users
        - You should be able to count the users in the same way as before, but this time you need to GROUP BY the test group. You can review the GROUP BY clause in this DataCamp lesson: Grouping data.
- Ad (1) Answer
    - The groups table contains the information on which A/B test group the individual user was in. As shown in answers and queries on question 1.5 the group table also contains the information on how many total users (48,943) were in the experiment.

- Ad (2) SQL query

```
SELECT "group", COUNT (DISTINCT uid) AS groups_dist_users, COUNT (uid) AS
groups_users
FROM groups
GROUP BY "group";
```

    - SQL query output/result (cf. CSV file) and markdown:

group	groups_dist_users	groups_users
A	24343	24343
B	24600	24600

1.7 What was the conversion rate of all users?

- Hints
    - (1) Review the definition of conversion rate
        - The conversion rate is the number of successful conversions (users who purchased) divided by the total number of users. If you represent conversion as a binary variable (1 for success, 0 for failure), the conversion rate is simply the mean of this variable for the control group.
    - (2) Determine which table(s) you need
        - Which table(s) do you need to find 1) all the users and 2) whether they converted? Return to the Project Overview page in Campus and reference the dataset description.
    - (3) Perform the appropriate join
        - Review the previous question: What type of join should we use to join the users table to the activity table? Which join will allow us to keep all of the users as needed in the denominator of the conversion rate?
    - (4) Determine which users converted

        - First, make sure you are counting each user only once. Review the previous question: Can a user show up more than once in the activity table? Yes or no, and why? If a user shows up more than once, we need to aggregate it down to a single row for each user. You can use the GROUP BY clause, which you can review in this DataCamp lesson: Grouping data.
        - Then, create a column that serves as an indicator for whether the user converted (0 = not converted, 1 = converted). You can use the CASE WHEN statement, covered in this DataCamp lesson: We'll take the CASE.
    - (5) Perform the necessary aggregations
        - Now that you have information about whether each user converted, calculate the overall conversion rate using SUM() / COUNT() or AVG() of the converted column that you created. You can put your query from the previous step into a common table expression (CTE) in order to separate the user-level grouping from the total conversion rate calculation. You can review CTEs in this DataCamp lesson: Common Table Expressions.
- Ad (2) Answers
    - 1) As stated in answer to question 1.5 you can find all users in the users or in the groups table.
    - 2) In the activity table you find whether they converted. As soon as they show up at least once in the activity table, they converted.
- Ad (3) Answer
    - As stated in answer to question 1.2 we should use a LEFT JOIN to join the users table to the activity table. The LEFT JOIN will allow us to keep all the users as needed in the denominator of the conversion rate.

- Ad (4) SQL query

```
SELECT
        g.uid AS user,
        COUNT(a.uid) AS converted_count,
        CASE
                WHEN COUNT(a.uid) = 0 THEN 0
                WHEN COUNT(a.uid) >= 1 THEN 1
                END AS converted_aggr_dwn
FROM groups AS g
LEFT JOIN activity AS a
ON g.uid = a.uid
GROUP BY g.uid
HAVING COUNT(a.uid) > 1;
```

    - SQL query output/result (cf. CSV file)

- Ad (5) SQL query

```
WITH converted_info AS (
    SELECT
        g.uid AS user,
        COUNT(a.uid) AS converted_count,
        CASE
            WHEN COUNT(a.uid) = 0 THEN 0
            WHEN COUNT(a.uid) >= 1 THEN 1
            END AS converted_aggr_dwn
    FROM groups AS g
    LEFT JOIN activity AS a
    ON g.uid = a.uid
    GROUP BY g.uid
    )

SELECT (SUM(converted_aggr_dwn) * 1.0 / COUNT(user)) AS overall_conv_rate
FROM converted_info;
```

  - SQL query output/result (cf. CSV file) and markdown:

overall_conv_rate
0.04278446355965102262

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

- Hints
    - (1) Refer to the previous question: What was the conversion rate of all users?
        - The first steps there will be very similar to what is needed for this question.
    - (2) Split the results by test group
        - You should be able to calculate the conversion rate in the same way as before, but this time you need to GROUP BY the test group and join any additional tables as necessary. You can review the GROUP BY clause in this DataCamp lesson: Grouping data.
- SQL query

```
WITH converted_info AS (
        SELECT
                g.uid AS user,
                g.group,
                COUNT(a.uid) AS converted_count,
                CASE
                        WHEN COUNT(a.uid) = 0 THEN 0
                        WHEN COUNT(a.uid) >= 1 THEN 1
                        END AS converted_aggr_dwn
        FROM groups AS g
        LEFT JOIN activity AS a
        ON g.uid = a.uid
        RIGHT JOIN users AS u
        ON u.id = g.uid
        GROUP BY g.uid
        )

SELECT
        "group", (SUM(converted_aggr_dwn) * 1.0 / COUNT(user))
        AS group_conv_rate
FROM converted_info
GROUP BY "group";
```

    - SQL query output/result (cf. CSV file) and markdown:

group	group_conv_rate
B	0.04630081300813008130
A	0.03923099042845992688

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

- Hints
    - (1) Review the previous question: What is the user conversion rate for the control and treatment groups?
        - The first steps there will be very similar to what is needed for this question.
    - (2) Determine which table(s) you need
        - Which table(s) do you need to find 1) all the users and 2) how much they spent? Return to the Project Overview page in Campus and reference the dataset description.
    - (3) Perform the appropriate join
        - Review the previous question: What type of join should we use to join the users table to the activity table? Which join will allow us to keep all of the users, including users who did not convert?
    - (4) Handle NULL values
        - Rows that do not have a match in the joined table will have NULL values for those columns. In our case, the spent column will be NULL for users who did not convert. Review the previous question: What SQL function can we use to fill in NULL values?
    - (5) Aggregate per user
        - First, make sure you only have one row per user, along with how much they spent in total ($0+). Review the previous question: Can a user show up more than once in the activity table? Yes or no, and why? If a user shows up more than once, we need to aggregate it down to a single row for each user. You can use the GROUP BY clause, which you can review in this DataCamp lesson: Grouping data.
    - (6) Calculate the average amount spent
        - Now that you have information about how much each user spent, calculate the average using AVG(). You can put your query from the previous step into a common table expression (CTE) in order to separate the user-level grouping from the average amount spent calculation. You can review CTEs in this DataCamp lesson: Common Table Expressions.
- Ad (2) Answers
    - 1) As stated in the answer to question 1.5 you can find all users in the users or in the groups table.
    - 2) In the activity table you find how much the users spent.
- Ad (3) Answer
    - As already in the answers to questions 1.1 and 1.2 we should use a LEFT JOIN to join the users table and the activity table. This also allows us to keep all of the users, including the users who did not convert.
- Ad (4) Answer
    - As stated in the answer to question 1.4 we can use the CASE WHEN SQL function to fill in the NULL values.

- Ad (5) SQL query

```
SELECT
        g.uid AS user,
        g.group,
        SUM(a.spent) AS sum_spent,
        CASE
                WHEN SUM(a.spent) IS NULL THEN 0
                ELSE SUM(a.spent)
                END AS sum_spent_clean,
        COUNT(a.uid) AS converted_count,
        CASE
                WHEN COUNT(a.uid) = 0 THEN 0
                WHEN COUNT(a.uid) >= 1 THEN 1
                END AS converted_aggr_dwn
FROM groups AS g
LEFT JOIN activity AS a
ON g.uid = a.uid
RIGHT JOIN users AS u
ON u.id = g.uid
GROUP BY g.uid;
```

  - SQL query output/result (cf. CSV file)

- Ad (6) SQL query

```
WITH converted_info AS (
    SELECT
        g.uid AS user,
        g.group,
        SUM(a.spent) AS sum_spent,
        CASE
            WHEN SUM(a.spent) IS NULL THEN 0
            ELSE SUM(a.spent)
            END AS sum_spent_clean,
        COUNT(a.uid) AS converted_count,
        CASE
            WHEN COUNT(a.uid) = 0 THEN 0
            WHEN COUNT(a.uid) >= 1 THEN 1
            END AS converted_aggr_dwn
    FROM groups AS g
    LEFT JOIN activity AS a
    ON g.uid = a.uid
    RIGHT JOIN users AS u
    ON u.id = g.uid
    GROUP BY g.uid
    )

SELECT
    "group",
    AVG(sum_spent_clean) AS avg_sum_spent_clean
FROM converted_info
GROUP BY "group";
```

  - SQL query output/result (cf. CSV file) and markdown:

group	avg_sum_spent_clean
B	3.39086694588578326
A	3.3745184679288412

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

- Hints
    - Consider what contributes to total revenue
        - When we think about how much revenue a company is making, two things matter:
            - 1. How many users/customers are making purchases
            - 2. How much money users/customers are spending, when they do
        - The experiment could result in more users converting, but spending less when they do. Or, it could result in fewer users converting, but spending more when they do. How will we know if the tradeoff is worth it in terms of total revenue?
- Answers
    - Including users who did not convert (i.e., did not make a purchase) when calculating the average amount spent per user is vital for obtaining a comprehensive understanding of customer behavior and the overall performance of a business or specific marketing campaigns. Here's why considering both converting and non-converting users is crucial:
    - 1. Holistic View
        - Balanced Perspective: Including all users, whether they converted or not, provides a balanced view of the overall performance and customer engagement with the business or campaign.
    - 2. Analyzing Revenue Impacts
        - Revenue Composition: The total revenue is derived from not just how much each converting user spends but also from how many users convert. An increase in average spend per user does not necessarily translate to an increase in total revenue if fewer users are converting.
        - Tradeoff Analysis: Sometimes, a marketing strategy might boost the average spend but decrease the conversion rate, or vice versa. By considering all users in the calculation, businesses can better understand and analyze these tradeoffs to optimize for total revenue.
    - 3. Customer Behavior Insight
        - Insight into Non-converters: Understanding the behavior and characteristics of users who do not convert can unveil opportunities for improved targeting, better personalization, and enhanced customer experience.
        - Enhancing Strategy: Analyzing both converters and non-converters helps to refine marketing strategies and product offerings to potentially convert non-buyers into buyers.
    - 4. Metric Authenticity
        - Accurate Averages: Calculating the average amount spent per user using all users ensures the metric accurately reflects the average across your entire user base, not just those who convert.
        - Realistic Performance Assessment: It helps in evaluating the real impact and success of campaigns or business strategies on the whole user base and not just a segment.

- 5. Effective Resource Allocation
    - Budget Optimization: Insights into the conversion and non-conversion rates assist in optimizing marketing budgets, ensuring resources are effectively utilized to improve overall performance.
    - Focus Areas Identification: Understanding patterns and behaviors of both converting and non-converting users can help identify areas that require more focus and investment.
- Conclusion:
    - In the context of total revenue and understanding the complete picture, considering both converting and non-converting users in calculations and analyses is essential. It helps in not only getting accurate metrics and insights but also aids in crafting strategies that enhance both user conversion and the amount spent, thereby optimizing total revenue.
    - Ensuring that the strategies and experiments conducted consider both conversion rate and average spend will pave the way to establishing a solid base for revenue optimization and business growth.

2 Extracting the Analysis Dataset

- Instruction
    - 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+).
- Hints
    - Understand how the tables join together
        - You will need all three database tables for this task. Return to the Project Overview page in Campus and reference the dataset description to understand the contents of each table. This will also help you determine which columns you need from each base table to get the columns listed above.
        - Refer to the questions and hints on Understanding the GloBox Database page:
            - Can a user show up more than once in the activity table? Yes or no, and why?
            - What type of join should we use to join the users table to the activity table?
    - Generate a user-level aggregate table
        - Once you have an idea of how the tables join together, you'll need to ensure that you end up with a dataset that has one row per user. If a user appears more than once in any table, you will need to summarize all of their purchases (if any) into one row using GROUP BY.
        - Refer to the questions and hints on Understanding the GloBox Database page:
            - What is the user conversion rate for the control and treatment groups?
            - What is the average amount spent per user for the control and treatment groups, including users who did not convert?

- SQL query

```
SELECT
        u.id,
        u.country,
        u.gender,
        g.device AS device_visit,
        g.group AS user_group,
        COALESCE(SUM(a.spent), 0) AS sum_spent,
        CASE
                WHEN COALESCE(SUM(a.spent), 0) > 0 THEN 1
                ELSE 0
                END AS is_converted
FROM users AS u
LEFT JOIN groups AS g
ON u.id = g.uid
LEFT JOIN activity AS a
ON g.uid = a.uid
GROUP BY u.id, u.country, u.gender, g.device, g.group;
```

  - SQL query output/result (cf. CSV file)

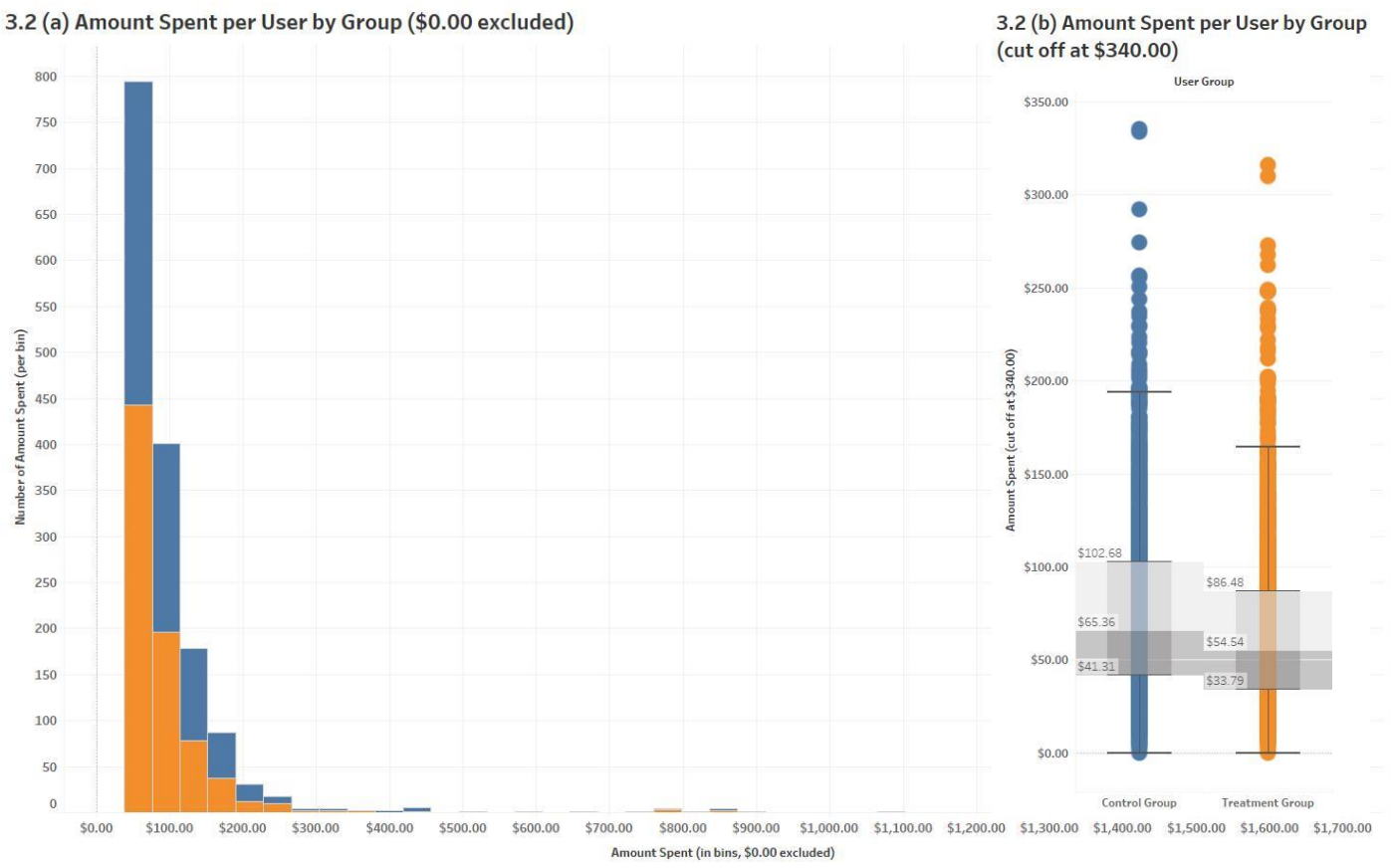
3 Visualize the Results in Tableau

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

- Hints
    - Decide which type of chart is appropriate
        - In this case, we are visualizing the relationship between the test group (categorical) and the metrics (numeric). Use the Data Visualization Cheat Sheet to determine which types of charts would be applicable here.
- Visualizations (cf. Tableau Public Workbook online and here)

3.2 Visualize the distribution of the amount spent per user for each group.

- Hints
    - Experiment with the chart
        - This variable average amount spent contains mostly zeros. How does this impact the visualization? Does it make it hard to see any additional patterns? Try making adjustments to the chart until it tells a clearer story about the difference in distributions between the two groups. You could try removing the zeros, adjusting the axis range, or adjusting the bin size (for histograms).
- Visualizations (cf. Tableau Public Workbook online and here)

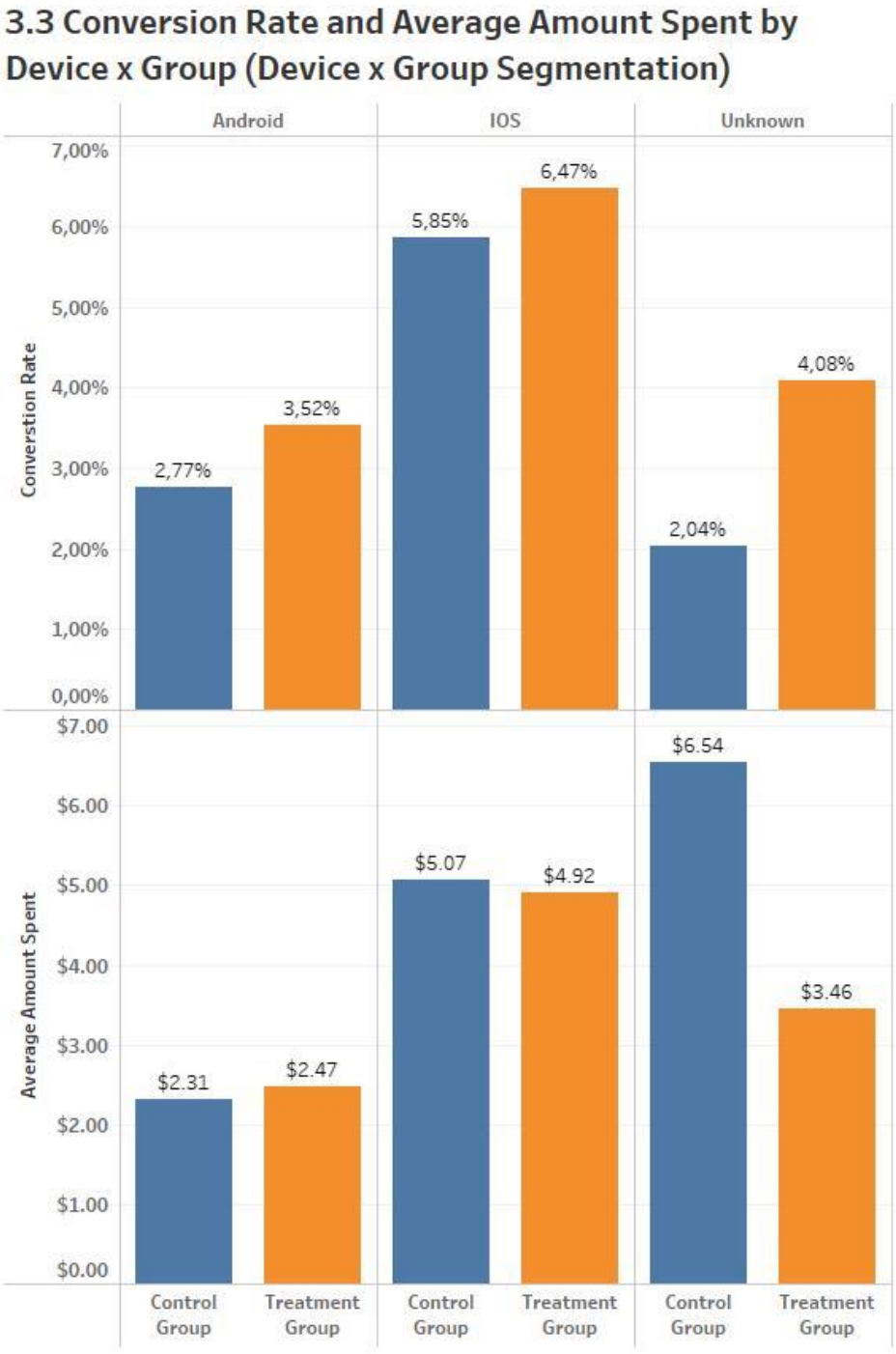


3.3 Create visualizations to explore the relationship between the test metrics (conversion rate and average amount spent) and the user's device.Hints

- Hints
    - Decide which type of chart(s) is appropriate
        - You will likely need to create more than one chart or a faceted chart to understand these relationships. Facets are just "small multiples" of charts that show the same data sliced by a dimension: Tableau 201: How to Make Small Multiples.
    - Don't forget about the sample size!
        - Unlike the test groups, the number of users per device is not split evenly. We might observe a big difference between the control and treatment for a particular device, but if it doesn't have many users, then it isn't necessarily meaningful. This is important when drawing conclusions based on what you see in the visualization.

- Handling missing values
  - This column contains missing values. When deciding how to handle this in your visualization, you can choose to either filter them out or treat them as their own category. This choice should be largely based on how many observations are missing.

- Visualizations (cf. Tableau Public Workbook online and here)

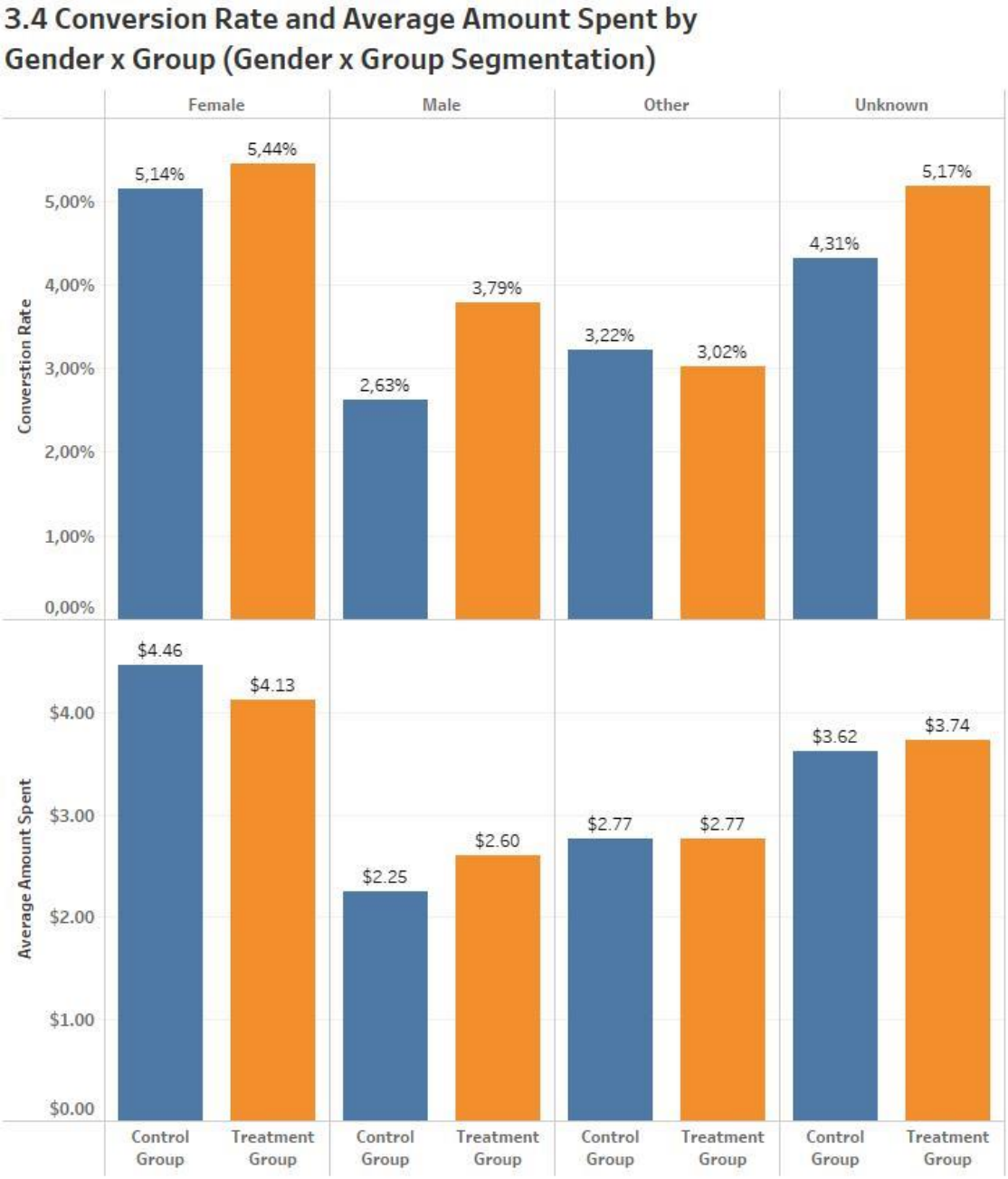


3.3 Sample Sizes by Device x Group

User Group	Android	IOS	Unknown
Control Group	15.054	9.142	147
Treatment Group	15.235	9.218	147

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

- Hints
    - Same hints as above
        - Similar to the device charts, you will need to consider:
            - Which chart is appropriate
            - The sample size per gender
            - How to handle missing values
- Visualizations (cf. Tableau Public Workbook online and here)

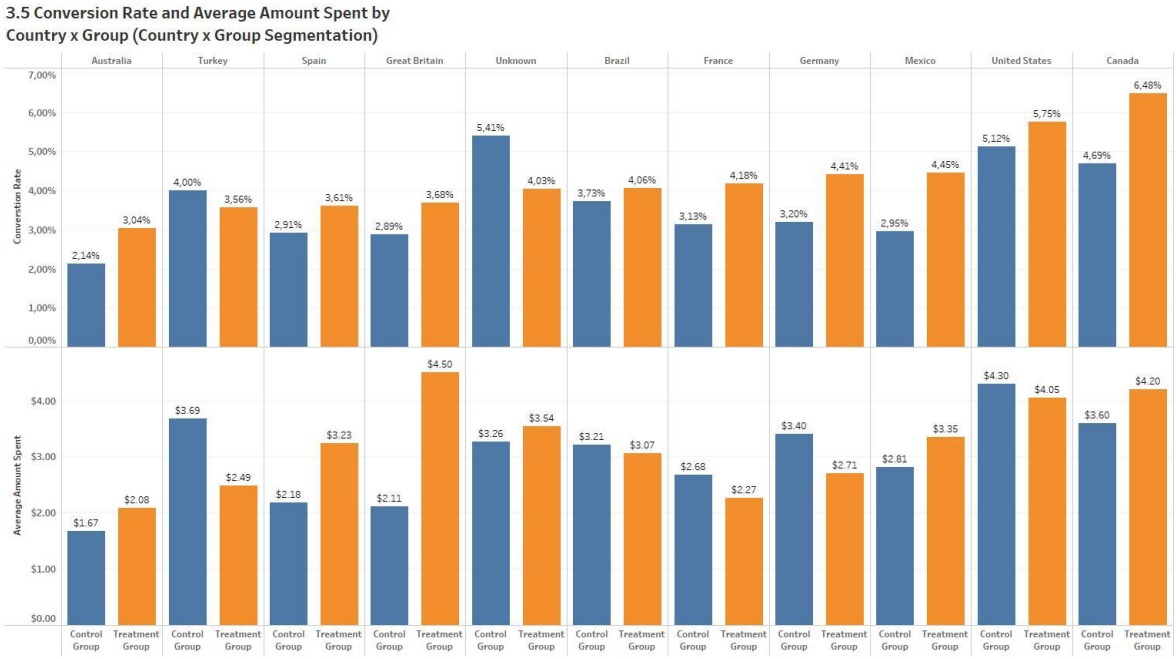


3.4 Sample Sizes by Gender x Group

User Group	Female	Male	Other	Unknown
Control Group	10.069	10.054	808	3.412
Treatment Group	10.061	10.235	861	3.443

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

- Hints
    - Same hints as above
        - Similar to the device charts, you will need to consider:
            - Which chart is appropriate
            - The sample size per gender
            - How to handle missing values
    - Consider grouping by region
        - There are 10+ countries to consider, which can be hard to interpret in a visualization. You could consider grouping the countries into regions or continents so that there are fewer, and the high-level trends will be easier to identify. You can still additionally inspect each individual country to see if there are any anomalies.
    - Be careful with maps
        - Since country is a geographic variable, you could explore using maps here. However, you'll need to be careful to make sure that they don't become overly complex. Be thoughtful about the story you want to tell, and whether a map is helpful to do that.
- Visualizations (cf. Tableau Public Workbook online and here)



3.5 Sample Sizes by Country x Group

User Group	Australia	Turkey	Spain	Great Britain	Unknown	Brazil	France	Germany	Mexico	United States	Canada
Control Group	608	1.849	997	1.455	296	4.805	1.536	1.906	2.815	7.309	767
Treatment Group	560	1.883	996	1.494	347	4.629	1.554	1.948	2.923	7.463	803

4 Calculate A/B Test Statistics using Spreadsheets

4.1 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
    - (1) Determine the null and alternative hypothesis
        - You can review the definition and notation of these hypotheses in the following DataCamp lessons: Statistics: Hypothesis Testing and Statistics in Spreadsheets: Hypothesis Testing.
        - Try writing the null and alternative hypotheses in both (a) plain language and (b) statistical notation. (c) When choosing between a one-sided or two-sided alternative hypothesis, consider the phrasing above: "whether there is a difference".
    - (2) Determine what type of test you are using
        - We can decide which test to use based on 3 factors:
            - (a) Are we working with proportions or means?
                - A proportion can be represented as a percentage. A mean is an average of continuous values. Is conversion rate a proportion or mean?
            - (b) Are we evaluating one sample or comparing two samples?
                - Remember that we are comparing groups of an A/B test.
            - (c) Is the test one-sided or two-sided?
                - This is also referred to as "one-tailed" or "two-tailed". Refer to your alternative hypothesis from the previous hint.
        - (d) Once you answer these questions, follow the flow chart above to find the right set of formulas to use on the Confidence Interval and Hypothesis Testing Cheat Sheet for the following steps.
    - (3) Calculate the test statistic
        - (a) You can find the appropriate formula for the test statistic on the Confidence Interval and Hypothesis Testing Cheat Sheet. You can use formulas or pivot tables to get the summary statistics for each group (counts, averages, variances, etc), then input them into another formula for the test statistic.
        - (b) A test statistic is a statistic calculated from the sample data used as evidence in the hypothesis test. Our test statistic is usually a standardized version of a sample statistic that follows its own distribution, called the sampling distribution. You can review sampling distributions in this DataCamp lesson: The central limit theorem.
        - Alternatively, if you know you are using a two-sample t-test for means, you can use the T.TEST() function in Google Sheets that calculates the test statistic and p-value for you. However, it is only applicable to one of these questions 😉. You can see how to use this function in this DataCamp lesson: Statistics in Spreadsheets: Hypothesis Testing.
    - (4) Calculate the p-value
        - The p-value is the probability we would have observed our evidence/data (or more extreme) if the null hypothesis is true. You can review p-values in this DataCamp lesson: Interpreting hypothesis testing results.

        - Knowing that our test statistic follows a particular distribution under the null hypothesis, we can calculate the tail probability using cumulative distribution functions in spreadsheets. In Google Sheets, we can use NORMSDIST() for z-tests and T.DIST() with cumulative=TRUE for t-tests.
    - (5) Draw a conclusion about the hypothesis
        - We draw a conclusion about the null hypothesis by comparing the p-value to the significance level. You can review significance levels in this DataCamp lesson: Interpreting hypothesis testing results (same video as above).
        - We decide to either reject the null hypothesis or fail to reject the null hypothesis in favor of the alternative hypothesis. If we reject the null hypothesis, our results are statistically significant.
        - A statistically precise written conclusion may look like the following: "With [$p<0.05$ | $p\geq0.05$], we [reject | fail to reject] the null hypothesis that [null hypothesis in words] in favor of the alternative hypothesis that [alternative hypothesis in words]."
- Ad (1) Answers
    - (a) Plain Language
        - Null Hypothesis (H_0): There is NO difference in the conversion rate between the two groups.
        - Alternative Hypothesis (H_1): There is A difference in the conversion rate between the two groups.
    - (b) Statistical Notation
        - Let p_1 be the conversion rate for group 1 and p_2 be the conversion rate for group 2.
        - Null Hypothesis (H_0): $p_1 = p_2$
        - Alternative Hypothesis (H_1): $p_1 \neq p_2$
    - (c) Given the phrasing "whether there is a difference", we can conclude that this is a two-sided hypothesis test.
- Ad (2) Answers
    - Ad (a): We are working with proportions.
    - Ad (b): We are comparing two samples.
    - Ad (c): The test is two sided ("two-tailed")
    - Ad (d): Two-sample z int/test for difference in proportions

- Ad (3) Answers
  - Ad (a)

Hypothesis Test for a Difference in Proportions

Two-sample z-test with pooled proportion

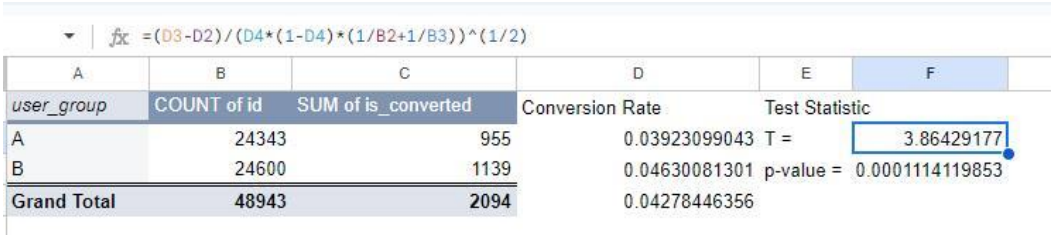
Use the unpooled proportions above if p_0 is something other than 0.

hypotheses	$H_0 : p_1 - p_2 = p_0$ $H_1 : p_1 - p_2 <> p_0$
test statistic	$T = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ $\hat{p} = \frac{\hat{p}_1 * n_1 + \hat{p}_2 * n_2}{n_1 + n_2}$
p-value	$2 * P(Z > \|T\|)$

▼ Notation

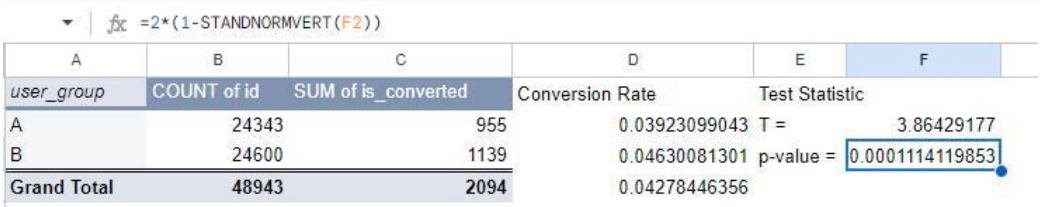
- H_0, H_1 = null and alternative hypotheses, respectively
- p_1, p_2 = proportions for populations 1 and 2, respectively
- p_0 = null value, the value we think the difference in population proportions might be
- $\hat{p}_1$, $\hat{p}_2$ = sample proportions for samples 1 and 2, respectively
- $\hat{p}$ = the pooled sample proportion
- T = test statistic
- n_1, n_2 = sample size for samples 1 and 2, respectively
- Z = the standard normal distribution

  - Ad (b) Spreadsheets results (cf. online Spreadsheet and here)



	A	B	C	D	E	F
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Test Statistic	
2	A	24343	955	0.03923099043	T =	3.86429177
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853
4	Grand Total	48943	2094	0.04278446356		

- Ad (4) Answer
  - Spreadsheets results (cf. online Spreadsheet and here)



	A	B	C	D	E	F
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Test Statistic	
2	A	24343	955	0.03923099043	T =	3.86429177
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853
4	Grand Total	48943	2094	0.04278446356		

- Ad (5) Answer
  - Conclusion about the hypothesis
    - With $p < 0.05$ and even $p < 0.01$ we reject the null hypothesis (H_0) that "There is NO difference in the conversion rate between the two groups" in favor of the alternative hypothesis (H_1) that "There is A difference in the conversion rate between the two groups."

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

- Use the normal distribution and unpooled proportions for the standard error.
- Steps
    - (1) Determine what type of interval you are computing
        - We can decide which interval to use based on 2 factors that you should have determined in the previous question:
            - 1. Are we working with proportions or means?
            - 2. Are we evaluating one sample or comparing two samples?
        - Once you answer these questions, you can follow the flow chart above to find the right set of formulas to use for the next steps on the Confidence Interval and Hypothesis Testing Cheat Sheet.
    - (2) Calculate the sample statistic
        - We center our interval around a sample statistic that serves as a "point estimate" (as opposed to an "interval estimate").
        - You can find the appropriate formula for the sample statistic on the Confidence Interval and Hypothesis Testing Cheat Sheet. You can use formulas or pivot tables to get the summary statistics for each group (counts, averages, variances, etc).
    - (3) Calculate the standard error
        - The distribution of the sample statistic is the sampling distribution. The standard error is the standard deviation of the sampling distribution. You can review sampling distributions in this DataCamp lesson: The central limit theorem.
        - The more variation there is in the sampling distribution, the larger the standard error will be, and the wider the interval will be. This Khan Academy video dives a little deeper into standard error: Standard error of the mean.
        - You can find the appropriate formula for the sample statistic on the Confidence Interval and Hypothesis Testing Cheat Sheet.
    - (4) Calculate the critical value
        - The critical value determines "how many standard errors wide" is our interval. It is determined by our confidence level and our sampling distribution. The higher our confidence, the higher the critical value will be, and the wider our interval will be. You can learn more about critical values in this Khan Academy video: Critical value (z*) for a given confidence level.
        - We can use inverse cumulative distribution functions in spreadsheets to calculate the critical value. In Google Sheets, we can use NORMSINV() for z-intervals and T.INV() for t-intervals.
    - (5) Construct the interval
        - A confidence interval consists of a lower bound and an upper bound that define a range of estimation for the sample statistic. First, we compute the margin of error = critical value * standard error. Then, we take the sample statistic and subtract the margin of error to get the lower bound, and add the margin of error to get the upper bound.

- Ad (1) Answer

Confidence Interval for a Difference in Proportions

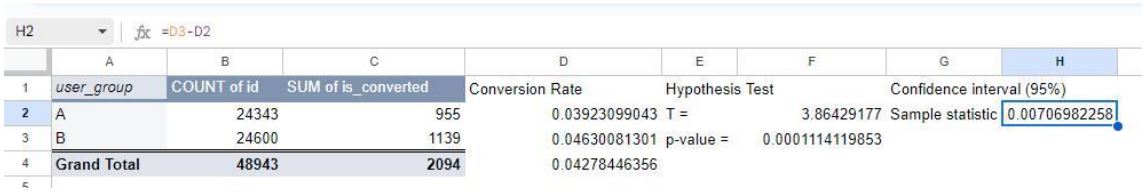
Two-sample z-interval with unpooled proportions

sample statistic	$\hat{p}_1 - \hat{p}_2$
critical value	$z_{1-\frac{\alpha}{2}}$
standard error	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

▼ Notation

- $\hat{p}_1$, $\hat{p}_2$ = sample proportions for samples 1 and 2, respectively
- n_1, n_2 = sample size for samples 1 and 2, respectively
- $z_{1-\frac{\alpha}{2}}$ = z-value, the value of the standard normal distribution at the $(1-\frac{\alpha}{2})th$ percentile

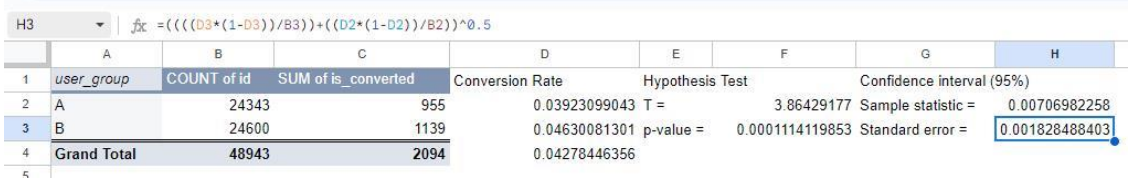
- Ad (2) Answer
  - Spreadsheets results (cf. online Spreadsheet and here)



H2 | =D3-D2

	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Hypothesis Test		Confidence interval (95%)	
2	A	24343	955	0.03923099043	T =	3.86429177	Sample statistic	0.00706982258
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853		
4	Grand Total	48943	2094	0.04278446356				

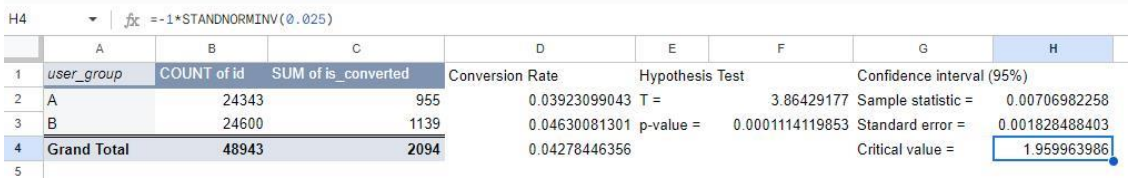
- Ad (3) Answer
  - Spreadsheets results (cf. online Spreadsheet and here)



H3 | =((((D3*(1-D3))/B3))+((D2*(1-D2))/B2))^0.5

	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Hypothesis Test		Confidence interval (95%)	
2	A	24343	955	0.03923099043	T =	3.86429177	Sample statistic =	0.00706982258
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853	Standard error =	0.001828488403
4	Grand Total	48943	2094	0.04278446356				

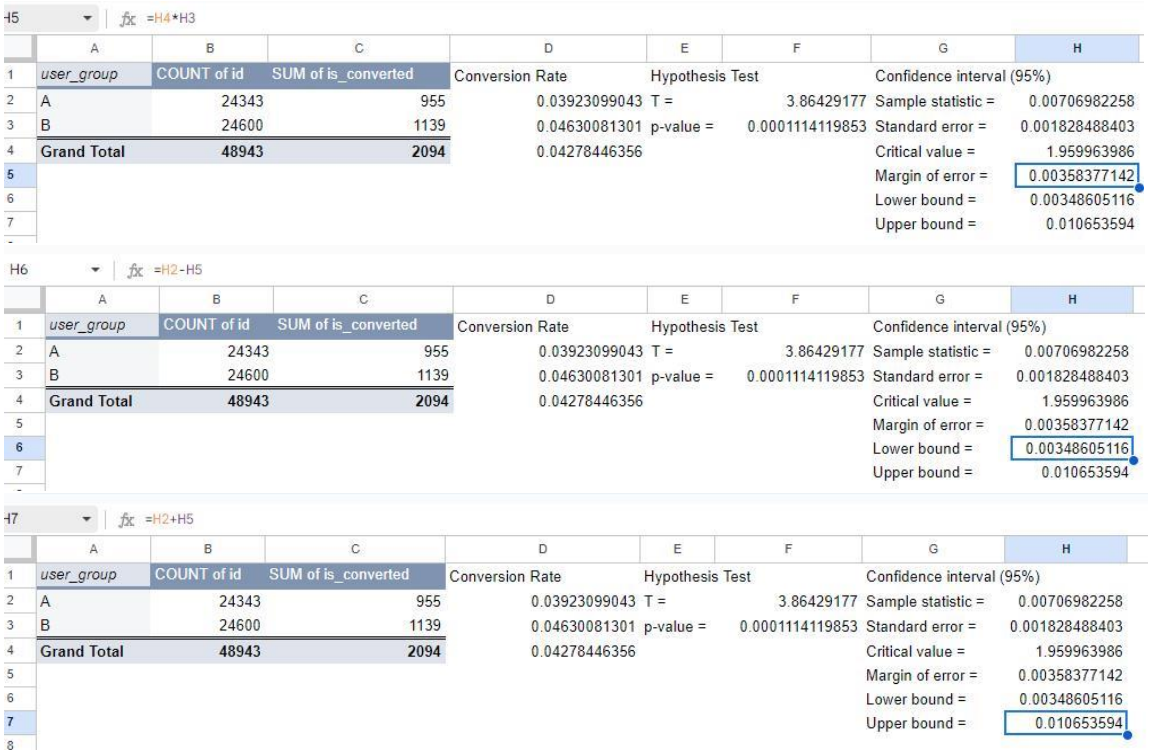
- Ad (4) Answer
  - Spreadsheets results (cf. online Spreadsheet and here)



H4 | =-1*STANDNORMINV(0.025)

	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Hypothesis Test		Confidence interval (95%)	
2	A	24343	955	0.03923099043	T =	3.86429177	Sample statistic =	0.00706982258
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853	Standard error =	0.001828488403
4	Grand Total	48943	2094	0.04278446356			Critical value =	1.959963986

- Ad (5) Answer
    - Spreadsheets results (cf. online Spreadsheet and here)



H5 =H4*H3

	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Hypothesis Test		Confidence interval (95%)	
2	A	24343	955	0.03923099043	T =	3.86429177	Sample statistic =	0.00706982258
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853	Standard error =	0.001828488403
4	Grand Total	48943	2094	0.04278446356			Critical value =	1.959963986
5							Margin of error =	0.00358377142
6							Lower bound =	0.00348605116
7							Upper bound =	0.010653594

H6 =H2-H5

	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Hypothesis Test		Confidence interval (95%)	
2	A	24343	955	0.03923099043	T =	3.86429177	Sample statistic =	0.00706982258
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853	Standard error =	0.001828488403
4	Grand Total	48943	2094	0.04278446356			Critical value =	1.959963986
5							Margin of error =	0.00358377142
6							Lower bound =	0.00348605116
7							Upper bound =	0.010653594

H7 =H2+H5

	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	SUM of is_converted	Conversion Rate	Hypothesis Test		Confidence interval (95%)	
2	A	24343	955	0.03923099043	T =	3.86429177	Sample statistic =	0.00706982258
3	B	24600	1139	0.04630081301	p-value =	0.0001114119853	Standard error =	0.001828488403
4	Grand Total	48943	2094	0.04278446356			Critical value =	1.959963986
5							Margin of error =	0.00358377142
6							Lower bound =	0.00348605116
7							Upper bound =	0.010653594
8								

4.3 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.
- See the steps from the previous hypothesis test
    - Follow the same steps as before:
        - (1) Determine the null and alternative hypothesis
        - (2) Determine what type of test you are using
        - (3) Calculate the test statistic
        - (4) Calculate the p-value
        - (5) Draw a conclusion about the hypothesis
    - Keeping in mind that the process will be the same, but using different formulas on the Confidence Interval and Hypothesis Testing Cheat Sheet and corresponding spreadsheet functions.
- Ad (1) Answers
    - (a) Plain Language
        - Null Hypothesis (H_0): There is NO difference in the average amount spent per user between the two groups.
        - Alternative Hypothesis (H_1): There is A difference in the average amount spent between the two groups.
    - (b) Statistical Notation
        - Let μ_1 be the average amount spent for group 1 and μ_2 be the average amount spent for group 2.
        - Null Hypothesis (H_0): $\mu_1 = \mu_2$
        - Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$
    - (c) Given the phrasing “whether there is a difference”, we can conclude that this is a two-sided hypothesis test.

- Ad (2) Answers
  - Ad (a): We are working with means.
  - Ad (b): We are comparing two samples.
  - Ad (c): The test is two sided (“two-tailed”)
  - Ad (d): Two-sample t int/test for difference in means
- Ad (3) Answers
  - Ad (a)

Hypothesis Test for a Difference in Means

Two-sample t-test with unpooled variance

This is what you might call “simplified Welch’s t-test” used in Khan Academy here.

hypotheses	$H_0 : \mu_1 - \mu_2 = \mu_0$ $H_1 : \mu_1 - \mu_2 <> \mu_0$
test statistic	$T = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
p-value	$2 * P(t_{df} > \lvert T \rvert)$ $df = min(n_1 - 1, n_2 - 1)$

▼ Notation

- H_0, H_1 = null and alternative hypotheses, respectively
- μ_1, μ_2 = means for populations 1 and 2, respectively
- μ_0 = null value, the value we think the difference in population means might be
- $\bar{x}_1$, $\bar{x}_2$ = sample means for samples 1 and 2, respectively
- T = test statistic
- n_1, n_2 = sample size for samples 1 and 2, respectively
- s_1, s_2 = sample standard deviation for samples 1 and 2, respectively
- t_{df} = the t distribution with degrees of freedom (df)
- df = degrees of freedom

  - Ad (b) Spreadsheets results (cf. online Spreadsheet and here)

F2 =((C3-C2)/(((D3/B3)+(D2/B2)))^(1/2))

	A	B	C	D	E	F
1	*user_group*	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test	
2	A	24343	3.374518468	672.6687211	T =	0.07042634919
3	B	24600	3.390866946	645.8507116	p-value =	0.9438548982
4	**Grand Total**	**48943**	**3.38273563**	**659.1893724**		

- Ad (4) Answer
  - Spreadsheets results (cf. online Spreadsheet and here)

F3 =2*(1-T.DIST(F2,B2-1,WAHR))

	A	B	C	D	E	F
1	*user_group*	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test	
2	A	24343	3.374518468	672.6687211	T =	0.07042634919
3	B	24600	3.390866946	645.8507116	p-value =	0.9438548982
4	**Grand Total**	**48943**	**3.38273563**	**659.1893724**		

- Ad (5) Answer
    - Conclusion about the hypothesis
        - With p ≥0.05 we fail to reject the null hypothesis (H_0) that "There is NO difference in the average amount spent between the two groups" in favor of the alternative hypothesis (H_1) that "There is A difference in the average amount spent between the two groups."

4.4 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.
    - See the steps from the previous confidence interval
    - Follow the same steps as before:
        - (1) Determine what type of interval you are computing
        - (2) Calculate the sample statistic
        - (3) Calculate the standard error
        - (4) Calculate the critical value
        - (5) Construct the interval
    - Keeping in mind that the process will be the same, but using different formulas on the Confidence Interval and Hypothesis Testing Cheat Sheet and corresponding spreadsheet functions.
- Ad (1) Answer

Confidence Interval for a Difference in Means

Two-sample t-interval with unpooled variance

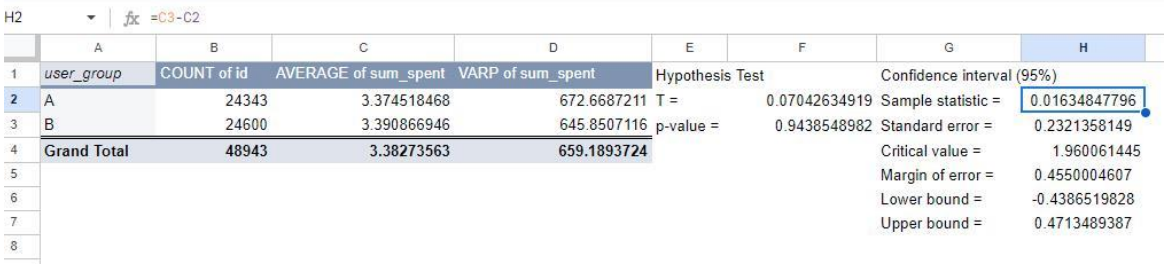
This is what you might call "simplified Welsh's t-interval" used in Khan Academy here.

sample statistic	$\bar{x}_1 - \bar{x}_2$
critical value	$t_{(1-\frac{\alpha}{2},\, df)}$ $df = min(n_1 - 1, n_2 - 1)$
standard error	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

▼ Notation

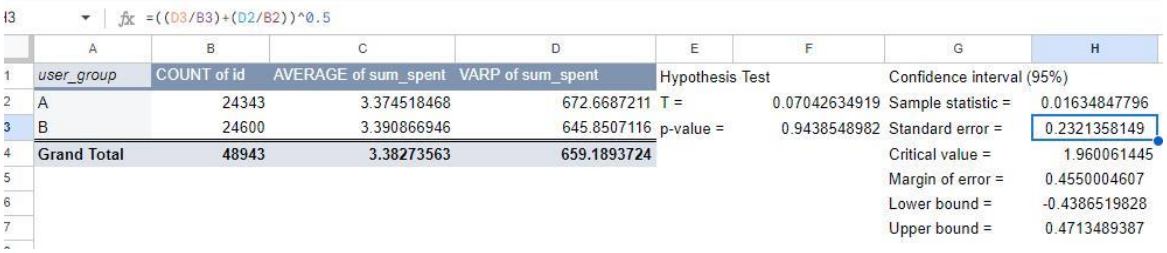
- $\bar{x}_1$, $\bar{x}_2$ = sample means for samples 1 and 2, respectively
- n_1, n_2 = sample size for samples 1 and 2, respectively
- s_1, s_2 = sample standard deviation for samples 1 and 2, respectively
- $t_{(1-\frac{\alpha}{2},\, df)}$ = t-value, the value of the t distribution with a specific degrees of freedom at the $(1 - \frac{\alpha}{2})th$ percentile
- df = degrees of freedom

- Ad (2) Answer
    - Spreadsheets results (cf. online Spreadsheet and here)



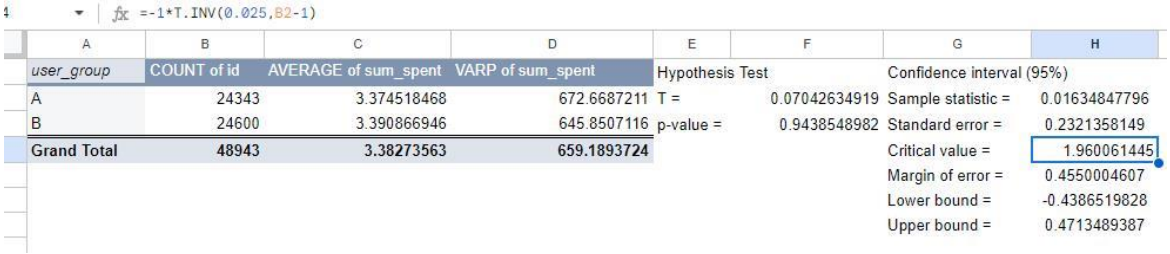
	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test		Confidence interval (95%)	
2	A	24343	3.374518468	672.6687211	T =	0.07042634919	Sample statistic =	0.01634847796
3	B	24600	3.390866946	645.8507116	p-value =	0.9438548982	Standard error =	0.2321358149
4	Grand Total	48943	3.38273563	659.1893724			Critical value =	1.960061445
5							Margin of error =	0.4550004607
6							Lower bound =	-0.4386519828
7							Upper bound =	0.4713489387
8								

- Ad (3) Answer
    - Spreadsheets results (cf. online Spreadsheet and here)



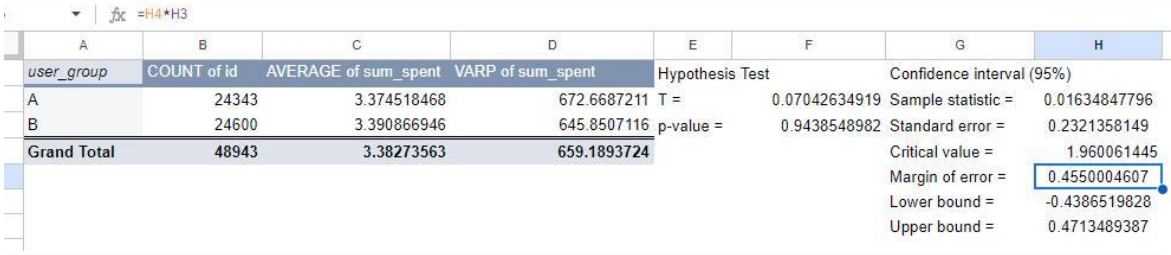
	A	B	C	D	E	F	G	H
1	user_group	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test		Confidence interval (95%)	
2	A	24343	3.374518468	672.6687211	T =	0.07042634919	Sample statistic =	0.01634847796
3	B	24600	3.390866946	645.8507116	p-value =	0.9438548982	Standard error =	0.2321358149
4	Grand Total	48943	3.38273563	659.1893724			Critical value =	1.960061445
5							Margin of error =	0.4550004607
6							Lower bound =	-0.4386519828
7							Upper bound =	0.4713489387

- Ad (4) Answer
    - Spreadsheets results (cf. online Spreadsheet and here)

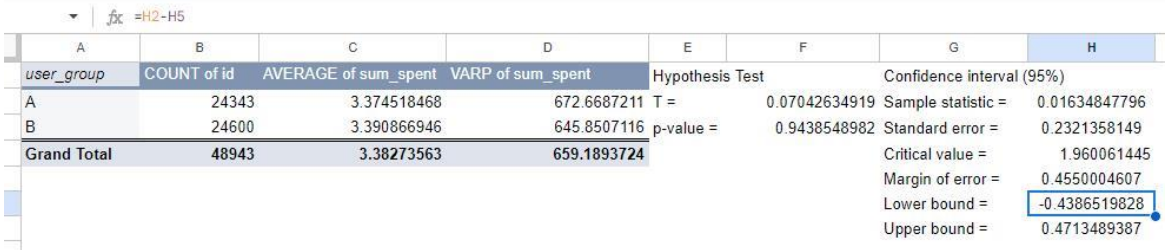


	A	B	C	D	E	F	G	H
	user_group	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test		Confidence interval (95%)	
	A	24343	3.374518468	672.6687211	T =	0.07042634919	Sample statistic =	0.01634847796
	B	24600	3.390866946	645.8507116	p-value =	0.9438548982	Standard error =	0.2321358149
	Grand Total	48943	3.38273563	659.1893724			Critical value =	1.960061445
							Margin of error =	0.4550004607
							Lower bound =	-0.4386519828
							Upper bound =	0.4713489387

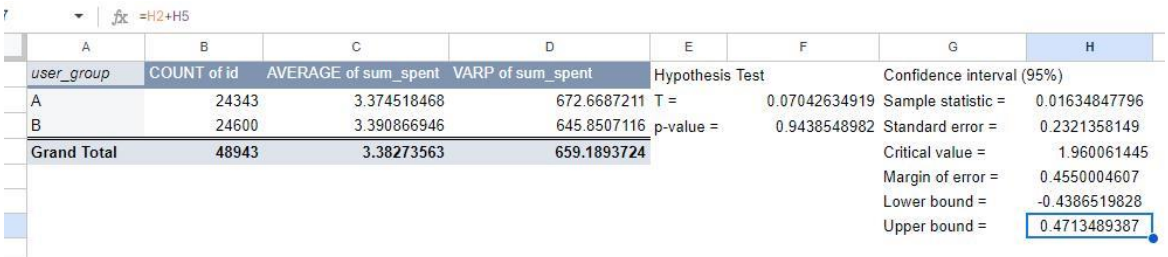
- Ad (5) Answer
    - Spreadsheets results (cf. online Spreadsheet and here)



	A	B	C	D	E	F	G	H
	user_group	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test		Confidence interval (95%)	
	A	24343	3.374518468	672.6687211	T =	0.07042634919	Sample statistic =	0.01634847796
	B	24600	3.390866946	645.8507116	p-value =	0.9438548982	Standard error =	0.2321358149
	Grand Total	48943	3.38273563	659.1893724			Critical value =	1.960061445
							Margin of error =	0.4550004607
							Lower bound =	-0.4386519828
							Upper bound =	0.4713489387



	A	B	C	D	E	F	G	H
	user_group	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test		Confidence interval (95%)	
	A	24343	3.374518468	672.6687211	T =	0.07042634919	Sample statistic =	0.01634847796
	B	24600	3.390866946	645.8507116	p-value =	0.9438548982	Standard error =	0.2321358149
	Grand Total	48943	3.38273563	659.1893724			Critical value =	1.960061445
							Margin of error =	0.4550004607
							Lower bound =	-0.4386519828
							Upper bound =	0.4713489387



	A	B	C	D	E	F	G	H
	user_group	COUNT of id	AVERAGE of sum_spent	VARP of sum_spent	Hypothesis Test		Confidence interval (95%)	
	A	24343	3.374518468	672.6687211	T =	0.07042634919	Sample statistic =	0.01634847796
	B	24600	3.390866946	645.8507116	p-value =	0.9438548982	Standard error =	0.2321358149
	Grand Total	48943	3.38273563	659.1893724			Critical value =	1.960061445
							Margin of error =	0.4550004607
							Lower bound =	-0.4386519828
							Upper bound =	0.4713489387

5 Advanced Tasks

5.1 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. There are several acceptable options for the exact appearance, with some references shown below:
    - Single point with error bars
    - Bar chart with error bars
- Hints
    - (1) Create a new data source
        - This will be the most straightforward if you begin by creating a new data source in your Tableau workbook that contains the lower bound, point estimate, and upper bound for both intervals.
        - You can create the calculated fields with the user-level analysis dataset in Tableau, but that would be more complex than creating this new data source with the values pre-calculated.
    - (2) Use a dual axis
        - You can reference this video on how to visualize the confidence intervals using a dual axis: Understanding Confidence Interval in Tableau.
    - (3) Tableau's built-in confidence interval calculations
        - Another option is to explore Tableau's built-in confidence interval calculations. You can review them in this DataCamp lesson: Tableau: adding lines and distribution bands.
        - However, it is notable that these computations are quite slow with our data volume. In addition, it is not always clear exactly what type of confidence interval it is using. If you choose to explore this route, make sure to double-check that the intervals it calculates are the same as what you have in your spreadsheet results.
- Answers
    - Ad (1) Separate New Data Source (cf. CSV)
    - Ad (2) Visualizations

    - Ad (3) not applied

5.2 Check for Novelty Effects

- Users might behave differently when the treatment is new, which is called a novelty effect. You can learn more about this in the following DataCamp video: Sanity checks: external validity.
- Inspect the difference in the key metrics between the groups over time. If we notice a novelty effect, that means that the effectiveness of the banner is short-lived, which may lead us to conclude that it isn't worth launching.
- Hints
    - Write another SQL query
        - This will require you to write (1) a new SQL query since we did not include any date columns in our analysis dataset from sprint 1. This new query should return the date and the metrics for each group in separate columns. Then, we can import this using a new data source in Tableau to do a visual inspection.
- SQL queries
    - (a)

```
SELECT
        u.id,
        u.country,
        u.gender,
        g.device AS device_visit,
        g.group AS user_group,
        g.join_dt AS join_date,
        COALESCE(SUM(a.spent), 0) AS sum_spent,
        CASE
        WHEN COALESCE(SUM(a.spent), 0) > 0 THEN 1
        ELSE 0
        END AS is_converted
FROM users AS u
LEFT JOIN groups AS g
ON u.id = g.uid
LEFT JOIN activity AS a
ON g.uid = a.uid
GROUP BY u.id, u.country, u.gender, g.device, g.group, g.join_dt;
```

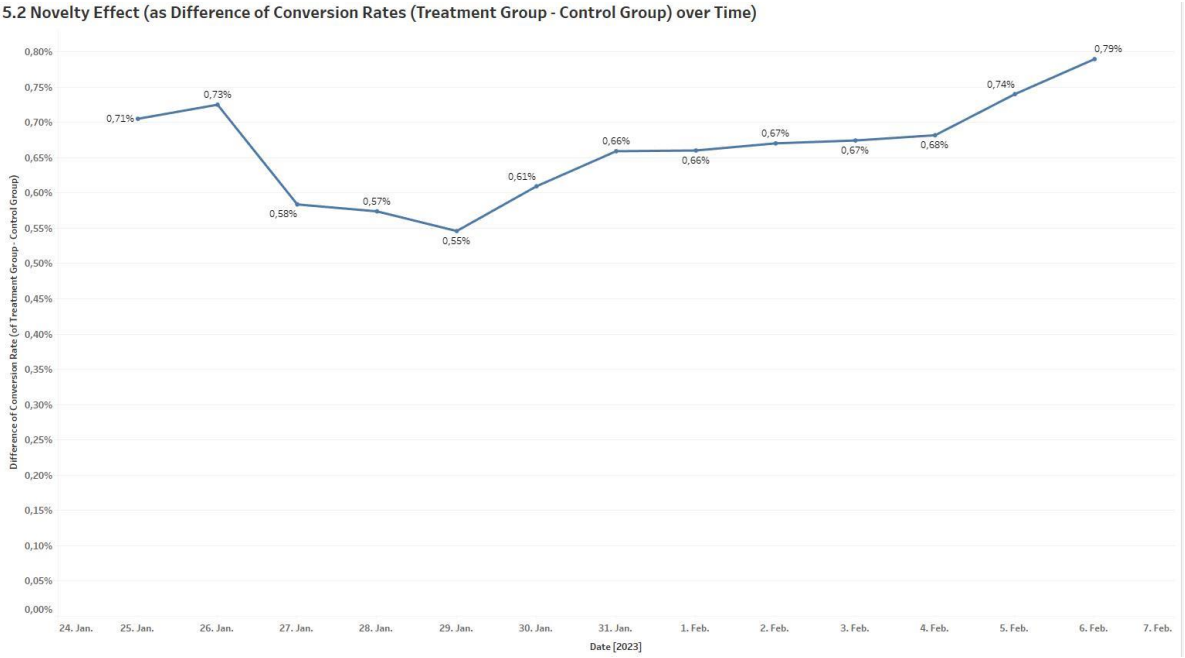
    - (b)

```
-- Step 1: Creating join_dt_agg
WITH join_dt_agg AS (
        SELECT
                g.join_dt,
                g.group AS test_group,
                COUNT(DISTINCT g.uid) AS user_count
        FROM groups g
        WHERE g.group IN ('A', 'B')
        GROUP BY g.join_dt, g.group
        ),
```

```
-- Step 2: Creating convert_dt_agg
convert_dt_agg AS (
        SELECT
                a.dt,
                g.group AS test_group,
                COUNT(DISTINCT a.uid) AS converted_user_count
        FROM groups g
        JOIN activity a ON g.uid = a.uid AND a.spent > 0
        WHERE g.group IN ('A', 'B')
        GROUP BY a.dt, g.group
        ),

-- Step 3: Creating cumulative_users
cumulative_users AS (
        SELECT
                COALESCE(a.dt, j.join_dt) AS dt,
                COALESCE(a.test_group, j.test_group) AS test_group,
                COALESCE(SUM(j.user_count) OVER (PARTITION BY j.test_group
                ORDER BY j.join_dt), 0) AS cum_users,
                COALESCE(SUM(a.converted_user_count) OVER (PARTITION BY
                a.test_group ORDER BY a.dt), 0) AS cum_converted_users
        FROM join_dt_agg j
        FULL JOIN convert_dt_agg a ON j.join_dt = a.dt AND j.test_group =
        a.test_group
        ),

-- Step 4: Creating cumulative_conversion
cumulative_conversion AS (
        SELECT
                dt,
                test_group,
                CASE
                        WHEN cum_users = 0 THEN 0
                        ELSE cum_converted_users * 1.0 / cum_users
                        END AS cum_conversion_rate
         FROM cumulative_users
        )

-- Step 5: Final Result
SELECT
        a.dt,
        a.cum_conversion_rate AS conversion_rate_A,
        b.cum_conversion_rate AS conversion_rate_B,
        b.cum_conversion_rate - a.cum_conversion_rate AS cum_diff
FROM cumulative_conversion a
JOIN cumulative_conversion b ON a.dt = b.dt
WHERE a.test_group = 'A' AND b.test_group = 'B'
ORDER BY a.dt;
```

- SQL query output/result (cf. CSV file)

- Visualizations (cf. Tableau Public Workbook online and here)

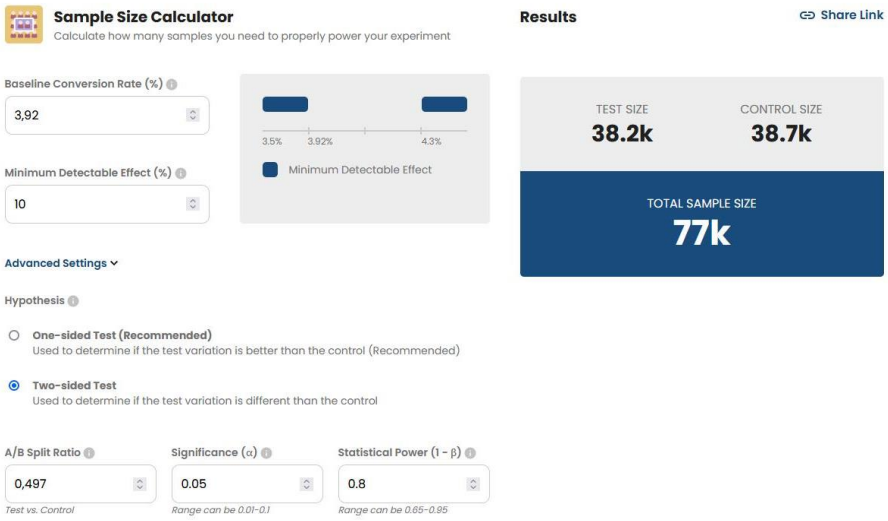


  - Visualization of the Novelty Effect (as Difference of Average Amount Spent (Treatment Group – Control Group) over Time) has not been conducted because of insignificant differences per date.

5.3 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. You can learn more about power analysis in this StatQuest video: Power Analysis, Clearly Explained!!!
- Explore the sample size calculators below to determine how many users we should have aimed to have in our test before making a conclusion.
    - Statsig Sample Size Calculator for Conversions
    - Statulator Sample Size Calculator for Means
- Hints
    - Practical significance and minimum detectable effect
        - Practical significance refers to what change is actually meaningful to the business if we are interested in launching this feature. This is different than statistical significance. You can learn more about the difference in this video: Statistical vs. Practical Significance Compared.
        - The minimum detectable effect (MDE) is the smallest difference (effect size) between the groups that we would be able to find statistically significant. The smaller the MDE, the more sensitive our test is and the more data we need.
        - We want our minimum detectable effect to be smaller than our practical significance level so that if we observe a difference that we are interested in, we will be able to detect it as statistically significant.
    - Choosing a minimum detectable effect
        - You are free to choose an MDE that you think is appropriate based on your business understanding. When it comes to a website banner like the one in this experiment, the engineering cost of launching the feature is typically very low. However, the business cost of not using that high-value page space for something else might not be worth it for a very small increase in conversions or revenue. If you are unsure of what to choose, you could use a 10% relative change.
    - (1) Using the calculators
        - The control group can serve as your baseline. If the calculator accepts a relative percent change MDE, you can input 10% or whatever value that you choose. If it expects an absolute difference, you can input the control group's value multiplied by the relative MDE.
    - (2) (Optional) Visualize the join curve
        - Optionally, you can visualize (a) the "join curve", which is the cumulative number of users in the experiment at each point in time. This would help you (b) forecast how much longer you would have needed to run the experiment in order to reach the desired sample size.

- Answers
    - Ad (1) Calculating Minimum Sample Size (taking MDE into account)
        - (a) Regarding Conversion Rate (CR)
            - Two-sided Test
            - Baseline CR = CR_A = 3.92
            - MDE = 10% (according to Hints)
            - A/B Split Ratio = n_A/N = 24343/48943 = 0.497
            - Significance (α) = 5% = 0.05
            - Statistical Power (1-β) = 0,8 (according to the video in Hints)
            - Results for Minimum Sample Size for CR ≥ 10%
                - $N_{CR \geq 10\%}$= 76900
                - $n_{CR \geq 10\%, A}$ = 38700
                - $n_{CR \geq 10\%, B}$ = 38200

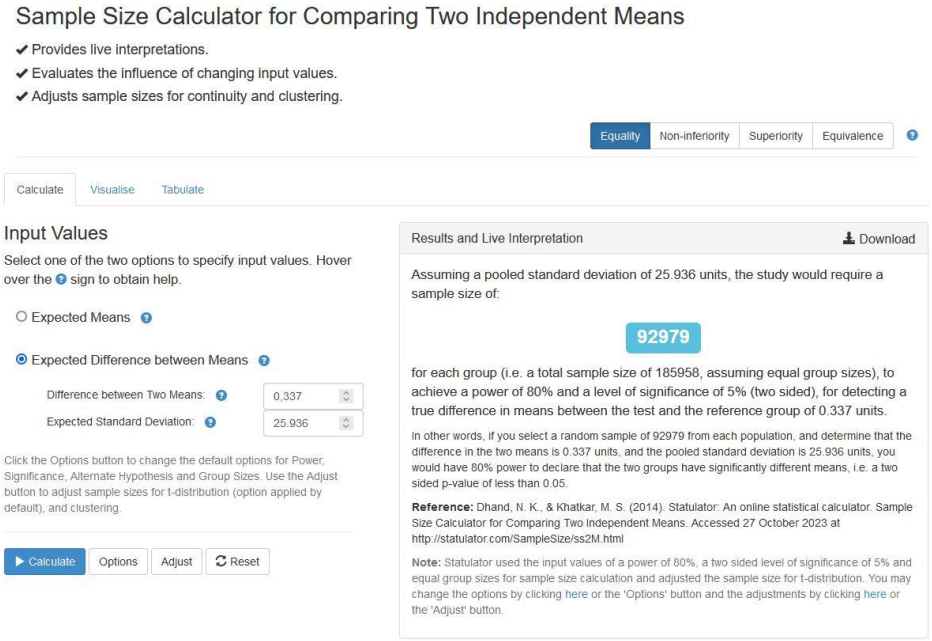


- Source: https://www.statsig.com/calculator?mde=10&bcr=3.92&twoSided=true&splitRatio=0.497&alpha=0.05&power=0.8

- (b) Regarding Average Amount Spent (AAS)
  - Expected Difference between Means = 10% of ASS_A
    - ASS_A = 3.37 $ → 10% of ASS_A = 0,1 * 3.37 $ = 0.337 $
  - Expected Standard Deviation
    - Baseline is s_A = $STDEVP_A$ = 25.936 (cf. CSV)

	A	B	C	D
1	user_group	COUNT of id	AVERAGE of sum_spent	STDEVP of sum_spent
2	A	24343	3.374518468	25.93585782
3	B	24600	3.390866946	25.41359305
4	Grand Total	48943	3.38273563	25.67468349

  - Results for Minimum Sample Size for AAS ≥ 10%
    - $N_{AAS \geq 10\%}$= 185958
    - $n_{AAS \geq 10\%, A}$ = $n_{AAS \geq 10\%, B}$ = 92979



Sample Size Calculation Results

Results and Live Interpretation

Assuming a pooled standard deviation of 25.936 units, the study would require a sample size of:

92979

for each group (i.e. a total sample size of 185958, 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 0.337 units.

In other words, if you select a random sample of 92979 from each population, and determine that the difference in the two means is 0.337 units, and the pooled standard deviation is 25.936 units, you would have 80% power to declare that the two groups have significantly different means, i.e. a two sided p-value of less than 0.05.

Reference: Dhand, N. K., & Khatkar, M. S. (2014). Statulator: An online statistical calculator. Sample Size Calculator for Comparing Two Independent Means. Accessed 28 October 2023 at http://statulator.com/SampleSize/ss2M.html

Note: Statulator used the input values of a power of 80%, a two sided level of significance of 5% and equal group sizes for sample size calculation and adjusted the sample size for t-distribution. You may change the options by clicking here or the 'Options' button and the adjustments by clicking here or the 'Adjust' button.

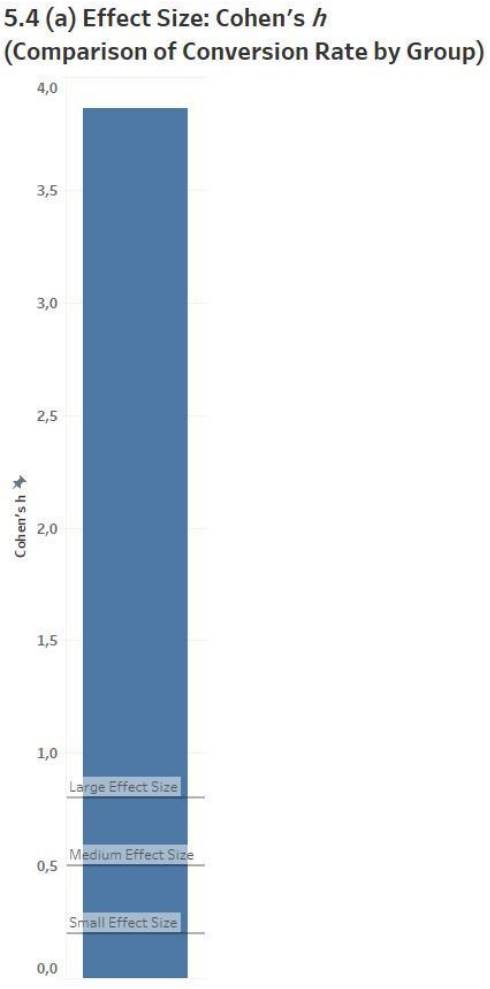
  - Source: Results produced by Statulator beta: http://statulator.com/SampleSize/ss2M.html

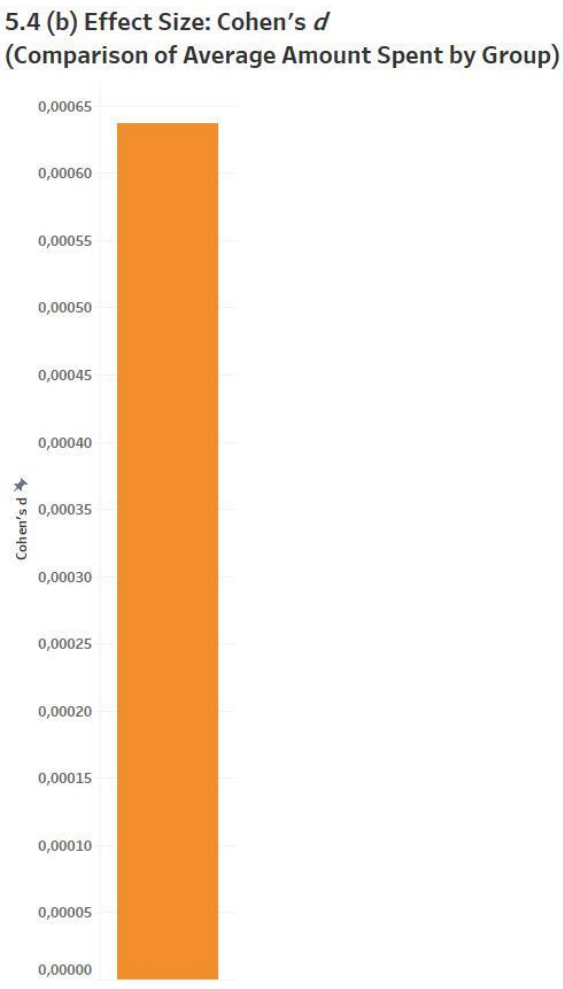
5.4 Calculation of Effect Sizes (voluntary)

- Hints
    - So, in this A/B test analysis, the effect size can be calculated for the observed differences in the conversion rate and average amount spent per user between the control and treatment groups. Effect size helps you determine the practical significance of the differences. One commonly used measure of effect size for proportions (conversion rates) is Cohen's *h*, while for means (average amount spent), Cohen's *d* is often used.
    - The calculated Cohen's *h* and Cohen's *d* will give you a measure of the effect size for the differences observed in conversion rate and average amount spent, respectively. A larger effect size suggests a more significant practical impact.
    - Keep in mind that effect size interpretation may vary depending on your domain, so you should consider what constitutes a meaningful effect size for your specific case. Cohen's guidelines are often used as a reference, with values like 0.2, 0.5, and 0.8 indicating small, medium, and large effect sizes, respectively.
    - (1) Cohen's *h* for Conversion Rate
        - (a) First, calculate the pooled proportion (P) by combining the conversion rates of both groups. It's the total number of conversions divided by the total number of users
            - $P = (n_{Converted\_Control} + n_{Converted\_Treatment}) / N$
        - (b) Calculate the standard error (SE) for the difference in proportions. This is a measure of the variability in your data
            - $SE = (P * (1 - P) * (1 / n_{Control} + 1 / n_{Treatment}))^{½}$
        - (c) Calculate the observed difference in proportions (P_diff) between the control and treatment groups
            - $CR_{Diff} = CR_{Treatment} - CR_{Control}$
        - (d) Calculate Cohen's *h* using the formula
            - Cohen's *h* = CR_{Diff} / SE
    - (2) Cohen's *d* for Average Amount Spent
        - (a) Calculate the pooled standard deviation (SD) for the two groups. It's a measure of the spread of the data
            - $SD_{Total} = (((n_{Control} - 1) * SD_{Control}^2 + (n_{Treatment} - 1) * SD_{Treatment}^2) / (n_{Control} + n_{Treatment}t - 2))^{½}$
        - (b) Calculate the observed difference in means (Mean_diff) between the control and treatment groups
            - $Mean_{Diff}$ = Average_Amount_Spent_Treatment - Average_Amount_Spent_Control
        - (c) Calculate Cohen's *d* using the formula
            - Cohen's *d* = $Mean_{Diff}$ / SD_{Total}

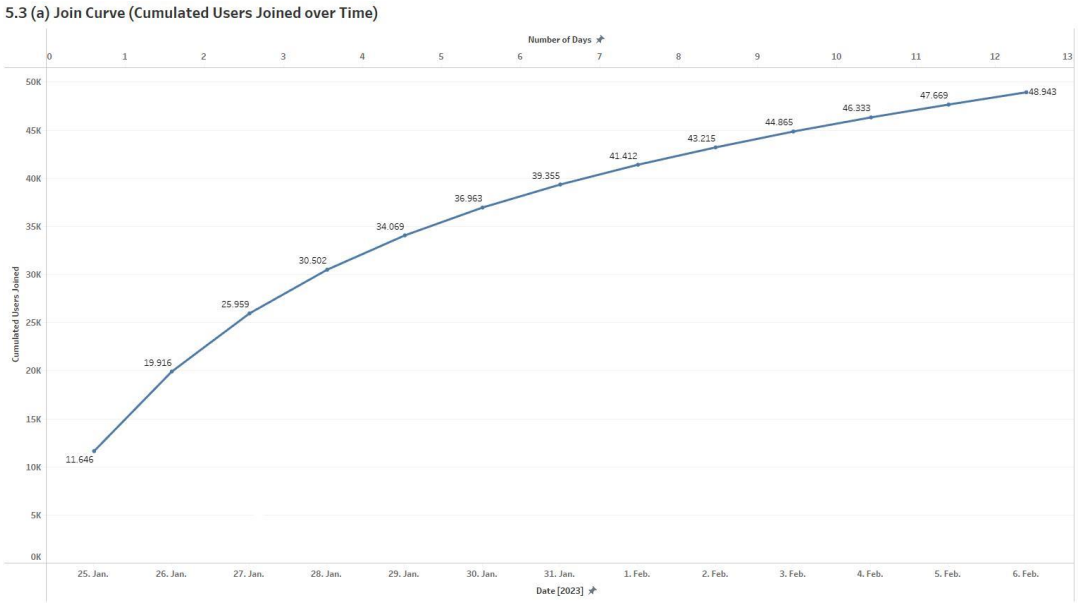
- Answers
    - (1) Cohen's *h* for Conversion Rate (CR)
        - (a)
            - $n_{Converted\_Control}$ = 955
            - $n_{Converted\_Treatment}$ = 1139
            - $n_{Converted\_Total}$ = 2094
            - N = 48943
            - P = (955 + 1139) / 48943 = 2094 / 48943 = 0.0427845
        - (b)
            - $n_{Control}$ = 24343
            - $n_{Treatment}$ = 24600
            - SE = $(0.0427845 * (1 - 0.0427845) * (1 / 24343 + 1 / 24600))^{1/2}$ = 0.00182952682558212
        - (c)
            - $CR_{Treatment}$ = 0.04630081300813008130
            - $CR_{Control}$ = 0.03923099042845992688
            - $CR_{Diff} = CR_{Treatment} - CR_{Control}$ = 0.0070698225796701
        - (d)
            - Cohen's *h* = 0.0070698225796701 / 0.00182952682558212 = 3.86429019832525
        - (e) Interpretation of the magnitude of Cohen's *h*
            - Thresholds
                - |*h*| = 0.20: Small effect size
                - |*h*| = 0.50: Medium effect size
                - |*h*| = 0.80: Large effect size
            - Given that your Cohen's *h* value is 3.864290198325253, it's substantially larger than the 0.80 threshold for a large effect size. This suggests a very large difference between the two compared proportions. In practical terms, a Cohen's *h* value of this magnitude suggests that there's a substantial difference between the two compared groups.

- (2) Cohen's *d* for Average Amount Spent
  - (a)
    - SD_{Total} = 25.67468349
  - (b)
    - $Mean_{Diff}$ = 3.39086694588578326 - 3.3745184679288412
      = 0.0163484779569401
  - (c)
    - Cohen's *d* = 0.0163484779569401 / 25.67468349
      = 0.000636754800241555
  - (d) Interpretation of the magnitude of Cohen's *d*
    - Thresholds
      - |*d*|=0.20: Small effect size
      - |*d*|=0.50: Medium effect size
      - |*d*|=0.80: Large effect size
    - Given that your Cohen's *d* value is 0.0006367548002415550, it's very close to 0. This suggests a very small or negligible effect size. In other words, the standardized difference between the means of the two compared groups is extremely small. In practical terms, a Cohen's *d* value of this magnitude implies that the difference between the two groups, in terms of their means, is trivial or negligible. When making decisions or interpretations based on this result, the two groups could be treated as having essentially the same mean.
- (3) Visualizations





- Ad (2) Visualization of Join Curve and Extrapolation
  - (a) Join Curve



  - (b) Forecast