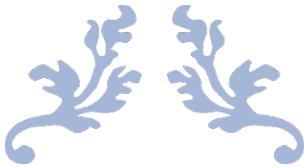


MASTERY PROJECT 1

Fereshteh Ranjbar

MAY 12, 2023

MASTER SCHOOL

Table of Contents

Table of Figures

Introduction

The purpose of this analysis is to evaluate and compare two variations of the app/website to determine the most viable course of action, whether it is launching the new version, refraining from launching, or re-evaluating the decision to change the front end.

This decision will be based on a thorough examination of conversion rates, which serve as a key metric for assessing user engagement and success. By analyzing and comparing the conversion rates between the two variations, we can derive valuable insights regarding user behavior and preferences, allowing us to make an informed decision regarding the future direction of the app/website.

Calculations

In this section, we will present the calculations, SQL code, and visualizations that were used to analyze the data. By employing these techniques, we aim to gain a comprehensive understanding of the underlying patterns and trends.

Conversion Rate

To begin the analysis, we extracted the relevant data from the database using SQL queries.

1. Conversion Rate Calculation: I computed the conversion rate by dividing the number of successful conversions by the total number of visitors or users. The formula used was:

$$Conversion\ Rate\ = (\frac{Successful\ Conversions}{Total\ Visitors}) * 100$$

$$Conversion\ Rate = \frac{Number\ of\ users\ in\ "activity"\ table}{Number\ of\ users\ in\ the\ "group"\ table}$$

Group A

This section includes all the necessary calculations for coming to a reasonable conclusion for group A.

Control Group's Conversion Rate (A/Old version)

The following is the SQL code for calculating the conversion rate for the control group (A).

```sql
with control as(
  select count(distinct uid) as numcont
  from groups as g
  where g.group='A'
),
numactA as(
  select count(distinct activity.uid)
 from activity
 left join groups
 on activity.uid=groups.uid
 where groups.group = 'A'
)
SELECT CAST((SELECT * FROM numactA) AS FLOAT) /
(SELECT * FROM control) AS controlconvrate;
```

Figure 1 Conversion rate for group A.

Conversion Rate Calculation Results

The following is the result of the SQL query which shows the conversion rate (in percentage)

Query Results
1 ROWS

controlconvrate
FLOAT8
3.9230990428459926

Figure 2 Results for group A.

Control Group's Conversion Rate 95% Confidence Interval (A/Old version)

$$confidence\_level = 0.95 \ \#\ 95\%\ confidence\ level$$
$$Z\_value = 1.96 \ \#\ Z\ value\ for\ a\ 95\%\ confidence\ interval$$
$$lower\_bound\_A = conversion\_rate\_A - Z\_value * standard\_error\_A$$
$$upper\_bound\_A = conversion\_rate\_A + Z\_value * standard\_error\_A$$
$$confidence\_interval\_A = (lower\_bound\_A, upper\_bound\_A)$$

The code snippet provided in Figure 2 demonstrates the calculation of the 95% confidence interval for the conversion rate of group A in a dynamic manner.
It is important to note that the Z value of 1.96, obtained from the Z-value table, is used in this calculation. Also, the equation box below is the theoretical steps behind calculating the interval.

```sql
WITH control AS(
  SELECT COUNT(DISTINCT uid) AS numcont
  FROM groups AS g
  WHERE g.group='A'
),
numactA AS(
  SELECT COUNT(DISTINCT activity.uid)
  FROM activity
  LEFT JOIN groups ON activity.uid=groups.uid
  WHERE groups.group = 'A'
),
conversion_rate AS (
  SELECT CAST((SELECT * FROM numactA) AS FLOAT) / (SELECT * FROM control) AS rate
),
se AS (
  SELECT SQRT(rate*(1-rate)/control.numcont) AS se
  FROM conversion_rate, control
)
SELECT (rate - 1.96*se)*100 AS lower_bound, (rate + 1.96*se)*100 AS upper_bound
FROM conversion_rate, se;
```

Figure 3 95% Confidence interval calculation for group A.

Confidence Interval Results

The following is the result of the SQL query which shows the conversion rate 95% confidence interval.

Query Results
1 ROWS

lower_bound FLOAT8	upper_bound FLOAT8
3.679209516027415	4.1669885696645705

Figure 4 Confidence interval results group A.

Group B

This section includes all the necessary calculations for coming to a reasonable conclusion for group B.

Treatment Group's Conversion Rate (B/New version)

The following is the SQL code for calculating the conversion rate for the treatment group (B).

```
with treat as(
  select count(distinct uid) as numcont
  from groups as g
  where g.group='B'
),
numactB as(
  select count(distinct activity.uid)
 from activity
 left join groups
 on activity.uid=groups.uid
 where groups.group = 'B'
)
SELECT CAST((SELECT * FROM numactB) AS FLOAT)*100 /
(SELECT * FROM treat) AS treatconvrate;
```

Figure 5 Conversion rate for group A

Conversion Rate Calculation Results

The following is the result of the SQL query which shows the conversion rate (in percentage)

Query Results
1 ROWS

controlconvrate FLOAT8
4.630081300813008

Figure 6 Results for group B.

Treatment Group's Conversion Rate 95% Confidence Interval (B/New version)

The code snippet provided in Figure 5 demonstrates the calculation of the 95% confidence interval for the conversion rate of group B in a dynamic manner.

```sql
WITH treat AS(
  SELECT COUNT(DISTINCT uid) AS numcont
  FROM groups AS g
  WHERE g.group='B'
),
numactB AS(
  SELECT COUNT(DISTINCT activity.uid)
  FROM activity
  LEFT JOIN groups ON activity.uid=groups.uid
  WHERE groups.group = 'B'
),
conversion_rate AS (
  SELECT CAST((SELECT * FROM numactB) AS FLOAT) / (SELECT * FROM treat) AS rate
),
se AS (
  SELECT SQRT(rate*(1-rate)/treat.numcont) AS se
  FROM conversion_rate, treat
)
SELECT (rate - 1.96*se)*100 AS lower_bound, (rate + 1.96*se)*100 AS upper_bound
FROM conversion_rate, se;
```

Figure 7 Confidence interval for group B.

Confidence Interval Results

The following is the result of the SQL query which shows the conversion rate 95% confidence interval.

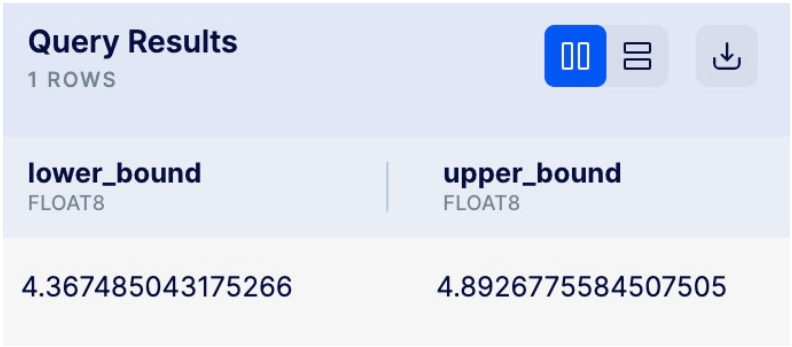


Figure 8 Confidence interval results for group B.

Comparison Between Groups

This section encompasses the calculation of crucial values required for conducting a hypothesis test, which forms the basis for drawing a reasonable conclusion. By performing this analysis, we aim to assess the significance of any differences between the groups under investigation. The calculated values play a pivotal role in guiding our decision-making process.

Pooled Probability of Conversion

The pooled probability is the total probability of conversion across groups.

$$P_{pool} = \frac{Successful\ conversion_A + Successful\ conversion_B}{Total_A + Total_B}$$

SQL code & Results

The following is the SQL code and query results for calculating the pooled probability.

```sql
WITH total AS(
  SELECT COUNT(DISTINCT uid) AS numcont
  FROM groups
),
conversion AS(
  SELECT COUNT(DISTINCT activity.uid)
  FROM activity
  LEFT JOIN groups ON activity.uid=groups.uid
),
pooledprob AS (
  SELECT CAST((SELECT * FROM conversion) AS FLOAT)*100 /
   (SELECT * FROM total) AS rate)
   select * from pooledprob
```

Figure 10 SQL code for pool probability.

Query Results
1 ROWS

rate FLOAT8
4.2784463559651025

Figure 9 Pooled probability result.

Pooled Standard Error

The pooled standard error which is given by this formula:

$$SE_{pool} = \sqrt{P_{pool} * (1 - P_{pool}) * ({}^{1}/_{Total_A} + {}^{1}/_{Total_B})}$$

SQL code & Results

The following is the SQL code and query results for calculating the pooled standard error.

```sql
with control as(
  select count(distinct uid) as totalA
  from groups as g
  where g.group='A'
),
numactA as(
  select count(distinct activity.uid)
 from activity
 left join groups
 on activity.uid=groups.uid
 where groups.group = 'A'),
 treat as(
  select count(distinct uid) as totalB
  from groups as g
  where g.group='B'),
  numactB as(
  select count(distinct activity.uid)
 from activity
 left join groups
 on activity.uid=groups.uid
 where groups.group = 'B'),
  ppool AS (
  SELECT CAST((SELECT * FROM numactA)+(SELECT * FROM numactB) AS FLOAT) /
   ((SELECT * FROM control)+(SELECT * FROM treat)) AS rate)
  SELECT SQRT(rate*(1-rate)*((1.0/totalA)+(1.0/totalb))) AS se
  FROM ppool, control,treat
```

Figure 11 Pooled standard error code.

Query Results

1 ROWS

se FLOAT8
0.001829526081285274

12 Calculated pooled standard error.

Hypothesis Test

To assess the impact of the new version, a hypothesis test was conducted to determine if there is a significant difference between the conversion rates of the two groups. This test plays a crucial role in informing our decision on whether to proceed with the launch or not. If a noticeable difference is observed in the conversion rates, it provides valuable evidence to support the decision-making process. On the other hand, if the difference is found to be statistically insignificant, it suggests that further investigation and additional data are necessary to arrive at a conclusive decision. In such cases, it is advisable to revisit the problem and gather more data to make a well-informed conclusion. The hypothesis test serves as a critical tool in evaluating the effectiveness of the new version and guiding the decision-making process for the future direction of the project.

$$\hat{d} = Conversion\ Rate_B - Conversion\ Rate_A$$

Hypothesis:

$$H_0: \widehat{d} = 0\,, \qquad \hat{d} \sim \mathrm{N}(0, SE_{pool})$$

$$H_a: \hat{d} \neq 0$$

If the following condition is true, then reject the null hypothesis. If not, then we cannot conclude anything.

$$\hat{d} > 1.96 * SE_{pool}\ or\ \hat{d} < -1.96 * SE_{pool}$$

$$1.96 * SE_{pool} = 0.003528$$

Difference Calculation

In the previous calculations we had the following results:

$$Conversion\ rate_B = 4.630$$

$$Conversion\ rate_A = 3.923$$

$$\hat{d} = 4.630 - 3.923 = 0.707$$

Data Visualization

I used Jupyter notebooks and Pandas to connect to the database and acquire more information about the data. The Jupyter notebook code is submitted in the zipped file.

Firstly, I plotted the normal distribution curves for the 3 observations (group A and group B and difference between groups' conversion rates) using their standard deviations and means. As observed, changing between the two groups resulted in a normal distribution curve that does not have a mean of 0.

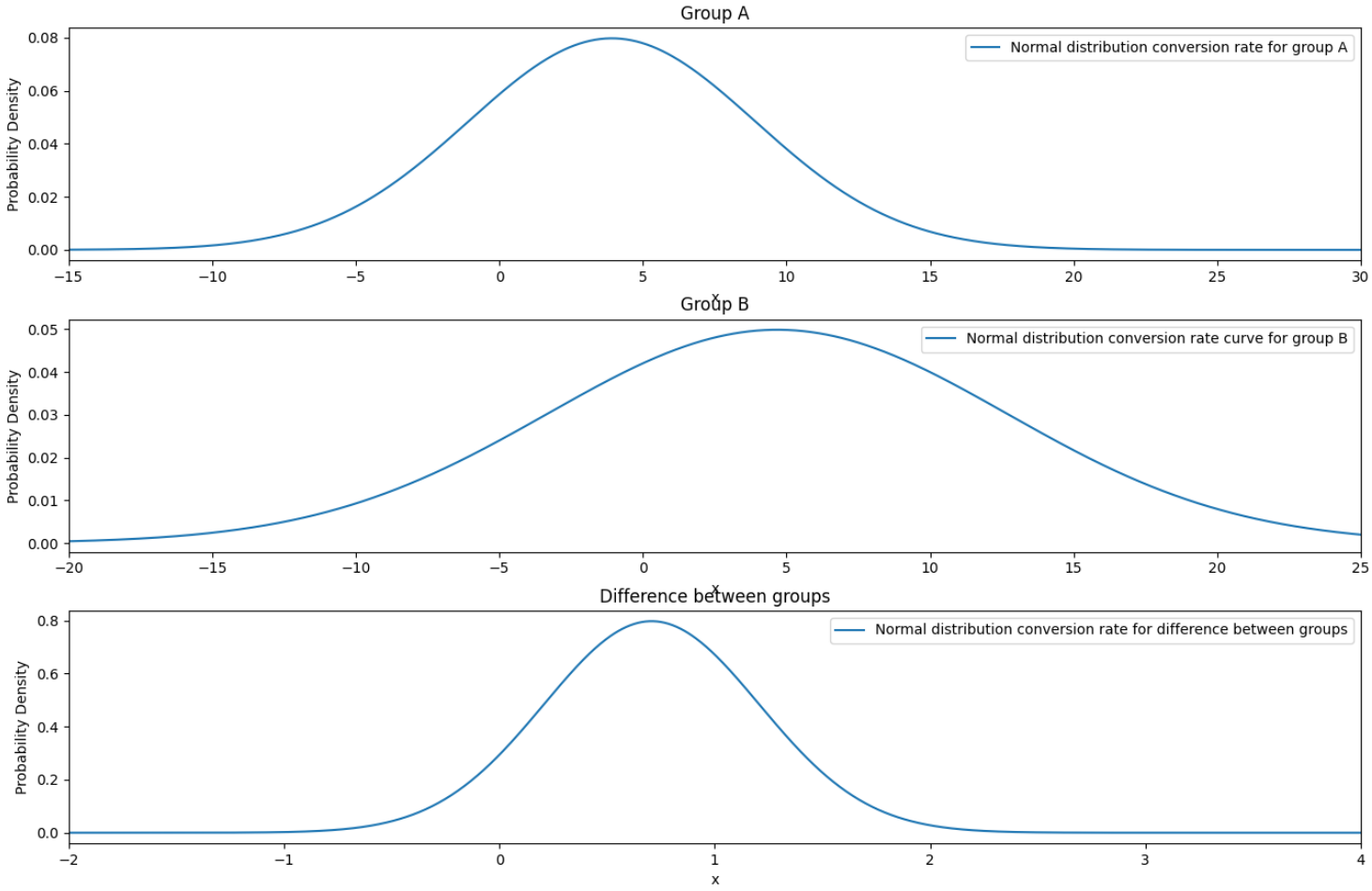
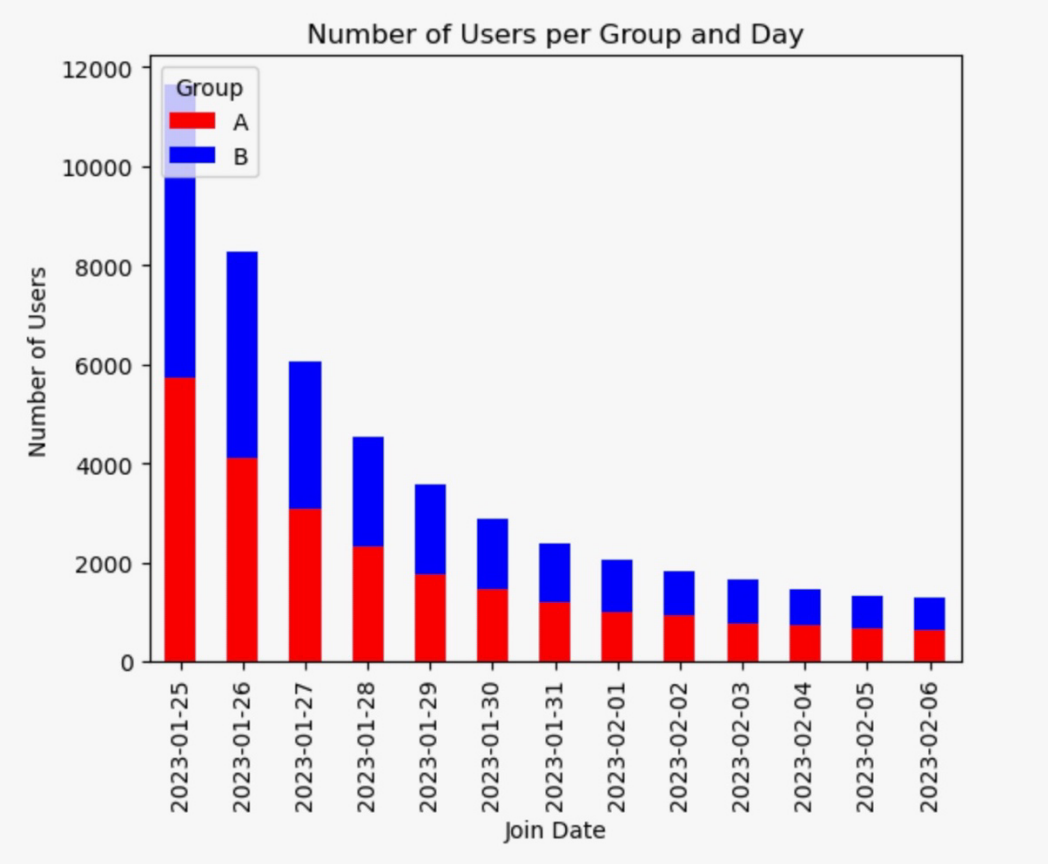
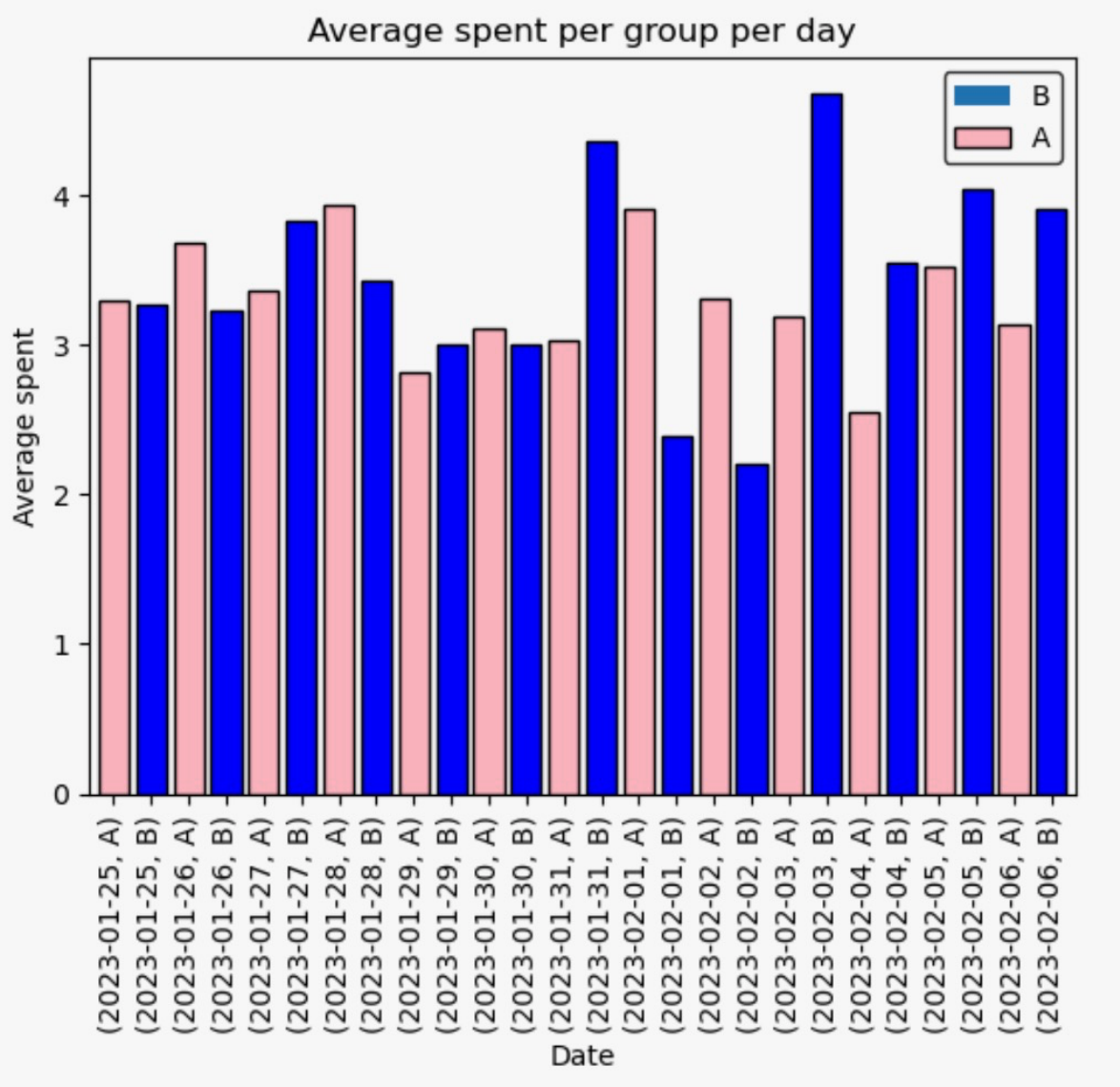


Figure 13 Normal Distribution Curves.

Then, I plotted the number of visitors to the website on each day for each group. Which is shown in Figure 14 below. By analysing this Figure, we can see that the number of visitors per day have increased for Group B. Which means that group B has more visitors and therefore it would be better to launch the new version.



Finally, I used the simplest factor of all “average amount spend per day for each group”, and as expected group B has overall higher average of amount spent per day.



Conclusion & Recommendations

The calculated difference ($\hat{d}$) between the conversion rates of the two groups, which is found to be greater than the value of 0.003528, indicates a statistically significant distinction. Consequently, we reject the null hypothesis that assumes the equality of the conversion rates. Moreover, this result provides compelling evidence to support the decision to proceed with the launch of the new version. By demonstrating a notable difference between the groups, the analysis reinforces the rationale behind moving forward with the implementation of the updated version, thereby reinforcing the potential benefits it may bring. Furthermore, the Pandas visualizations for different factor also were in favor of lunching the new version. So based on our statistical analysis we can confidently conclude that the new version should be launched.