DS5110 Homework 3 – AB Testing - Yueheng Yuan

Part 1. Getting to know your data

1. *What data is in file "t1_users_active_mins.csv"?*

This table contains active minutes data logged after the experiment started, recording the time users spent on the platform on specific dates.

There are three columns: uid (user ID), dt (date when the active minutes were recorded), and active_mins (time users spent on the platform for that date).

2. *What data is in file "t2_users_variant.csv"?*

This table contains user treatment assignments for the experiment, indicating whether a user was in the control or treatment group.

There are four columns: uid (user ID), variant_number (experiment variant – 0 for control and 1 for treatment), dt (date when users entered the experiment, which is consistent for all in this case), and signup_date (the date when a user signed up to the platform).

3. *What data is in file "t3_users_active_mins_pre.csv"?*

This table contains user activity before the experiment by recording the active time users spent on social media on specific dates.

Three columns are the same as those in "t1_users_active_mins.csv".

4. *What data is in file "t4_users_attributes.csv"?*

This table contains user attributes including gender and user types.

There are three columns: uid (user ID), gender, and user_type (new_user, non_reader, reader, or contributor).

5. *What data is in file "table_schema.txt"?*

This file describes the schema of all four csv files above, including the meaning of each column and the context of the data.

Part 2. Organizing the Data

1. *What is the overall objective of this study?*

The objective is to analyze the impact of installing a new feature on user engagement, which is measured by active minutes spent on the platform. In this experiment, users are divided into two subsets by experiencing new features or remaining the same for control.

2. *What data do we need to reach that objective?*

We'd need uid (track individual users), variant_number (identify control vs treatment groups), dt (login date), and active_mins (measure user engagement for each login).

3. *How is the data in t1 currently organized?*

The t1 file contains active minutes data after the experiment organized by uid and dt, which control vs treatment groups cannot be identified

4. *How should the data in t1 be organized to be useful?*

To analyze engagement differences between the control and treatment groups, `t1_user_active_min.csv` and `t2_user_variant.csv` should be merged using `uid` as foreign keys to assign treatment/control labels. The ideal dataset structure after getting merged consists of four columns: uid, dt, variant_number, and active_mins.

5. *Organize it*

The left-join operation is performed by matching values from the t2 file (variant_number) and added into the t1 data based on uid. The organized dataset is saved under `data/merged_user_data_after_exp.csv` for further statistical analysis.

Part 3. Statistical Analysis

Assume control group as group 1 and treatment as group 2, we may perform the statistical analysis under two scenarios: per login (dt) or per user (uid, accumulating the total active time per user).

Scenario #1: If we analyze the user engagement per login,

1. *Is there a statically difference between group 1 and group 2?*

Null Hypothesis (H_0): The new feature doesn't impact the total time spent on the platform. Any difference observed is due to random chance.

Alternative Hypothesis (H_1): The new feature has impacts on the total time spent on the platform.

For the independent t-test, assuming the data in each group approximately follows a normal distribution, T-value = -1.4674, P-value = 0.1423 > 0.05

Therefore, we failed to reject the null hypothesis, and no significant difference was found between the control and treatment groups.

2. *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per login: 35.34, Median active minutes per login: 5.00

Group 2 (Treatment):

Mean active minutes per login: 40.24, Median active minutes per login: 7.00

The treatment mean is slightly higher than the control group, but the median difference is relatively small. However, the large gap between the mean and median for both groups suggests a possible skewed distribution, which may violate the t-test assumption of normality.

3. *What can you conclude based on that data?*

The statistical test indicates that there is no significant difference in user engagement between the control and treatment groups ($p > 0.05$). While the treatment group has a slightly higher mean and median active minutes, this difference is not statistically significant, and the observed increase may be due to random variation rather than the new feature installed.

Scenario #2: If we analyze the user engagement per user,

1. *Is there a statically difference between group 1 and group 2?*

Null Hypothesis and alternative Hypothesis will be the same.

For the independent t-test, assuming the data in each group approximately follows a normal distribution, T-value = 0.4056, P-value = 0.6850 > 0.05

Again, we failed to reject the null hypothesis, and no significant difference was found between the control and treatment groups.

2. *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per user: 837.64, Median active minutes per user: 52.00

Group 2 (Treatment):

Mean active minutes per user: 784.20, Median active minutes per user: 71.00

Like the first scenario, the treatment mean is slightly higher, and a large gap between the mean and median is observed for both groups, suggesting a possible skewed distribution and may violate the t-test assumption of normality.

3. *What can you conclude based on that data?*

Similarly, though the treatment group has a slightly higher mean and median active minutes, the difference in user engagement between the control and treatment groups is not statistically different ($p > 0.05$).

Part 4. Digging a Little Deeper

1. *Can you trust the results? Why or why not?*

The statistical result is less reliable due to the skewed data.

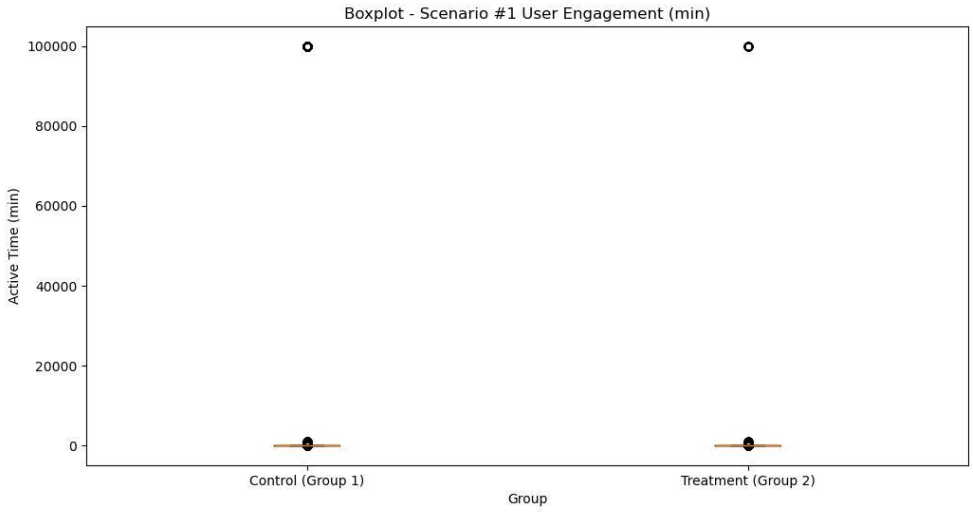
As mentioned in the second question of part 3, there presents large gaps between the mean and median, suggesting that the data in both groups is likely skewed rather than normally distributed. It indicates that the t-test assumption of normality is violated, and both groups may contain outliers or skewed data.

2. *Is the data normally distributed?*

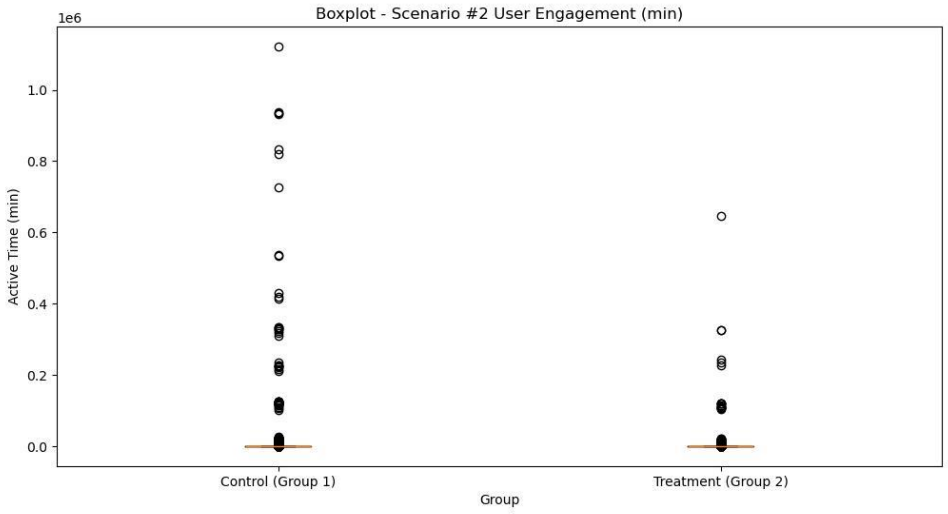
No, the large gap between the mean and median for both groups indicates that they are skewed.

3. *Plot a box plot of group 1 and group 2.*

Scenario #1: If we analyze the user engagement per login,



Scenario #2: If we analyze the user engagement per user,



4. *Are there any outliers?*

Both boxplots observe clear outliers for both control and treatment groups.

5. *What might be causing those outliers? (Hint, look at the data in t1. What is the maximum time a user should possibly have?).*

Both groups consist of outliers such that the maximum recorded active time is 99999.0, while there are only 1440 minutes in a day.

6. *Remove any data point that might be causing outliers.*
7. *Redo part 2 and 3 with the new data without those data points.*

Again, the statistical analysis is performed under two scenarios: per login (dt) or per user (uid, accumulating the total active time per user).

Scenario #1: If we analyze the user engagement per login,

- *Is there a statically difference between group 1 and group 2?*

For the independent t-test, T-value = -30.6868, P-value = 0.0000 < 0.05

Therefore, there is a statistically significant difference between the control and treatment groups to reject the null hypothesis.

- *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per login: 19.34, Median active minutes per login: 5.00

Group 2 (Treatment):

Mean active minutes per login: 23.53, Median active minutes per login: 7.00

The median is much lower than the mean in both groups, indicating that the data is right-skewed, which may imply that some users have very high engagement and therefore increases the mean value.

Scenario #2: If we analyze the user engagement per user,

- *Is there a statically difference between group 1 and group 2?*

For the independent t-test, assuming the data in each group approximately follows a normal distribution, T-value = -0.0093, P-value = 0.9926 > 0.05

In this case, we failed to reject the null hypothesis, and no significant difference was found between the control and treatment groups. When aggregating active time for each user, the treatment doesn't show a significant impact in comparison to the control group.

- *What is the mean and median for group 1 and group 2?*

Group 1 (Control):

Mean active minutes per user: 458.22, Median active minutes per user: 52.00

Group 2 (Treatment):

Mean active minutes per user: 458.40, Median active minutes per user: 71.00

When summing the total active time for each user, the mean is nearly identical. The median for the treatment is slightly higher, but this difference is not statistically significant, which suggests that while individual logins may be longer or more frequent in the treatment group, the impact isn't significant enough when summed across all logins for a user.

- *What can you conclude based on that data?*

According to the output from both scenarios, the treatment seemed to have an effect on engagement per login, but not on total active time per user. This could imply that the treatment led to more engagement per login, but it didn't significantly affect the total active time for each user.

8. *What is the new conclusion based on the new data?*

As mentioned above, the new feature installed tended to result in more engagement per login, but it didn't significantly affect the total active time for each user.

At the login level regardless of user, the treatment had a significant positive effect on user engagement. When looking at total active time per user, there is no statistically significant difference between the control and treatment groups.

Part 5. Digging Even Deeper

1. *Why do we care about the data from t3?*

The data for user active time before the experiment can provide a baseline for comparison, which helps assess whether the treatment had a significant impact on engagement. The absence of this baseline may bring uncertainty such that any observed change in active time was caused by the treatment or part of normal user behavior.

2. *Accounting for the data from t3 rerun part 2 and 3.*

Part 2: Organize data.

After cleaning outliers with active_mins > 1440, t1 -t2 and t3-t2 were merged based on uid to identify control or treatment group. The active time was summed per user, then subtracted to observe the difference on the personal total active time between before and after the experiment.

Part 3:

- *Is there a statically difference between group 1 and group 2?*

T-value = -17.9948, P-value: 0.0000 < 0.05

There is a statistically significant difference between the control and treatment groups. The null hypothesis can be rejected, and the new feature has impacts on user engagement.

- *What is the mean and median for group 1 and group 2?*

Group 1 (Control): Mean = -47.30, Median = -6.00

Group 2 (Treatment): Mean = 164.65, Median = 12.00

The positive mean for the treatment group implies the increase of user engagement after installing the new feature. The median suggests a rising trend, but it is far smaller than the mean value, meaning some users with high engagement significantly influenced the result.

- *What can you conclude based on that data?*

After removing outliers and accounting for pre-experiment engagement, it indicates that there exists statistically significant difference between the treatment and control groups.

Besides, the control group demonstrates the decline in engagement, implying that users might spend less time on the platform over time with the current design. The new feature not only increases user engagement for the treatment group, but also prevents the natural decline observed in the control group.

3. *Are there any new conclusions?*

In the previous analysis for the `per user` scenario, no impact was detected. After accounting for user activity before the experiment, it can be indicated that installing the new feature could help gain more engagement, while keeping the original version might lose engagement. The large t-value also suggests a potentially significant business impact to reinforce the user engagement across platform.

Part 6. Exploring other conclusions

Given the comparison of control vs treatment (overall) above, I'd like to perform a three-way interaction by adding control (male) vs treatment (male) and control (female) vs treatment (female). The goal is to check whether the feature impact males and females differently.

- Male: control vs treatment

Group 1 (Control): Mean = -46.80, Median = -7.00

Group 2 (Treatment): Mean = 192.75, Median = 15.00

T-value = -13.8369, P-value = 0.0000 < 0.05, so there is a statistically significant difference between the control and treatment groups.

- Female: control cs treatment

Group 1 (Control): Mean = -40.71, Median = -5.00

Group 2 (Treatment): Mean = 130.91, Median = 10.00

T-value = -9.1575, P-value = 0.0000 < 0.05, so there is a statistically significant difference between the control and treatment groups.

- Conclusion

This result indicates a statistically significant increase in engagement for both genders after installing the new feature, and the new feature demonstrates a more positive impact on male users. Besides, users in both genders tended to use the platform less over time without the new feature.

Part 7. Summarize Your Results

In part 1, four datasets were analyzed to help define key variables needed for analysis, such as active minutes, variant assignment, and user attributes.

Next, to compare engagement between the control and treatment groups, files t1 and t2 were merged using `uid`, allowing to categorize users into control or treatment groups for AB testing – detecting whether the new feature may bring new business impacts.

For the initial statistical analysis in part 3, two scenarios were analyzed: the per-login analysis showed no statistically significant difference between the control and treatment groups, and the per-user analysis also indicated similar results. However, the large gap between mean and median values suggested a skewed distribution, which violated normality assumptions and implied potential outliers.

In part 4, after removing unrealistic active times (>1440 min/day), the per-login analysis showed that engagement was significantly higher with the new feature. However, the per-user engagement remained unchanged, suggesting the new feature affected engagement per session but not total time spent per user.

For further analysis, the t3 file was merged with pre-experiment data to provide a baseline comparison. In this case, the control group showed a natural decline in engagement, while the treatment group gained engagement. A statistically significant difference was found, confirming that the new feature could not only counteract the natural decline but also increase engagement per user, bringing positive business impacts.

Additionally, a three-way interaction test was conducted with gender-based analysis. It has been revealed that both males and females would like to spend more time on the platform with this new feature, and the impact was much stronger for male users. This feature also successfully reversed engagement decline for both genders.

Overall, the new feature significantly increased engagement, and without the feature, engagement tended to decline over time. The gender analysis showed a stronger effect for male users, but positive impacts for both genders were observed. It should also be mentioned that pre-experiment activity data was critical, as accounting for prior engagement as baseline helped reveal a significant impact.