Name: Abdikarim Jimale

Date: 01-20-2025

HW03 Hypothesis Testing / AB Testing

Part 01:

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

- The t1 file contains the following data:
- Uid: the user ID number
- dt: data when the user registered
- active_mins: how many minutes the user spend.

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

- The t2 have similar columns in t1 like uid and have unique columns like the following:
- dt: the data the user enters the experiment and, in this file, the dt in all rows are 2019-02-06
- signup_date: the date the user signed up
- variant_number: which is which group the user is in. There are two group A and B. This column is either 0 or 1.

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

- This file is before the experiment starts and you can find that if we look at the date columns. The file has similar columns in other files like uid (user ID) and the following columns:
- dt: The date the user registers and this before the expiration date.
- active_mins: how many minutes the user spend.

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

- This file has the following date:
- Uid: user ID
- Gender: the user gender Male, Female, and unknown.
- User_type: it could be one of the following: new_user, non_reader, reader, and contributor

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

- The file contains information about all the files. It does give you a lot of details about how other files work.

Part 02:

1. What is the overall objective of this study?

- The object of this study is to see if the change on the layer has good or bad effect. This study will use the data that has been collected from some users. The data that was collected has information like how many times each user logs in to the program and how many minutes

spend on it each day and each time. By comparing these data, we can find out the effect of this change.

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

- We need to compare the data we have. We need to find out how much time the user spends on data.

3. How is the data in t1 currently organized?

- The data on file t1 is sorted by the user ID and the date. The same user can have many columns for different logins on the same day or different day.

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

- The objective of this study is if the new change takes more time, so we care more about active minutes. We don't need to focus on how many times the use login in. The best way to organize t1 file is by merge the data from t1 and t2 together. We need uid, variant_number, and active_mins. We need to get the sum of active mins for each user.

5. Organize it.

Data > t1_organized_data.csv > data

```
1   uid,variant_number,active_mins
2   0,0,43.0
3   1,0,15205.0
4   2,0,17.0
5   3,0,77.0
```

```
40000   39998,0,1.0
40001   39999,0,102.0
40002   40000,1,25.0
40003   40001,1,299.0
40004   40002,1,183.0
40005   40003,1,0.0
40006   40004,1,56.0
```

Part 3: Statistical Analysis (10 Points)

You can now start running some statistical analysis now that you hopefully organized the data from part 2 in a way that can be useful. Answer the following questions based only on the data from t1 and t2:

code result:

```
The number of element or rows at t1 is:  50000
group A size:  40000
group B size:  10000
Mean of active minutes for group A:  783.719625
Median of active minutes for group A:  45.0
Mean of active minutes for group B:  722.094
Median of active minutes for group B:  59.0
T-statistic:  0.40182764684510935
P-vlaue:  0.687812591494914
There is no statically difference between group 1 and group 2
```

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

- Based on T-test and P-value results we can say there is no difference between the two groups. In P-value test if (P > 0.05) that means there is no statically different and we can reject the null hypothesis case.

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

- Mean of active minutes for group A: 783.719625
- Median of active minutes for group A: 45.0
- Mean of active minutes for group B: 722.094
- Median of active minutes for group B: 59.0

3. What can you conclude based on that data?

Active time spent = mean

Consistently active = median

- Control group (group 1) has a bigger meaning compared to Treatment group (group 02), so we can say control group was more active. The median tells us that the user in Treatment group are more consistently active compare to control group. This experiment didn't show a big difference between the groups, so we can say that there wasn't any increase or decrease in active time spent.

Part 4: Digging a Little Deeper (25 Points)

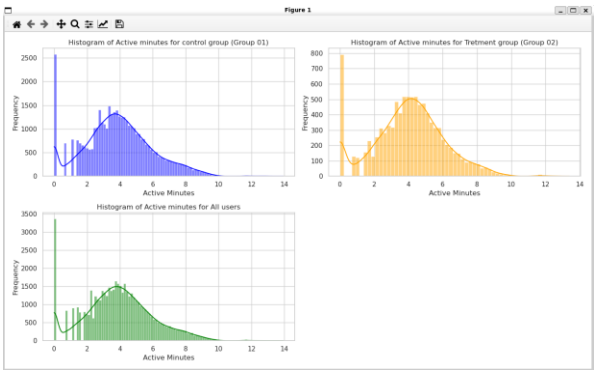
Just because you came to one conclusion does not mean that it is necessarily correct. There can be many different things that are impacting the results of your analysis. Answer the following questions:

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

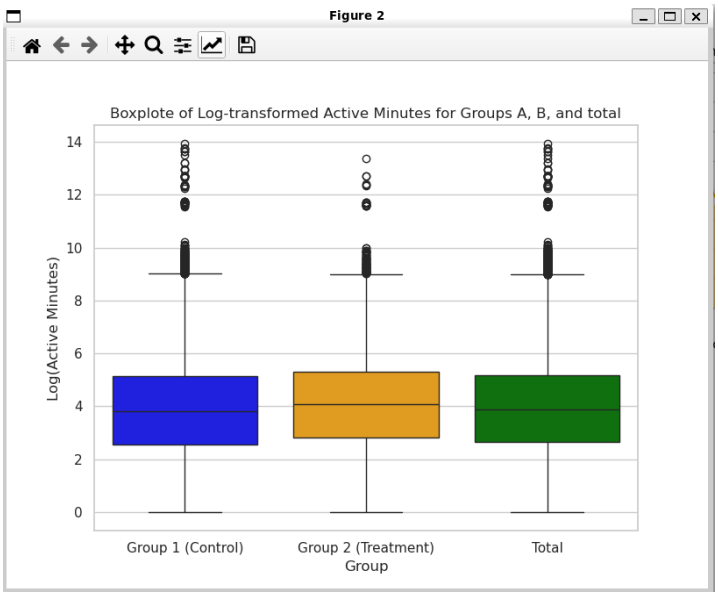
- Yes, I can trust these results because they are based on a large number of users. One thing to consider is the difference between the group sizes. The control group size is way bigger than treatment group size. This difference can affect the statistical power of the test and reliability of the result.

2. Is the data normally distributed?

- Based on the P-value results we get which is P > 0.05 and is fail to reject the null hypothesis, the data is normally distributed.



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



4. Are there any outliers?

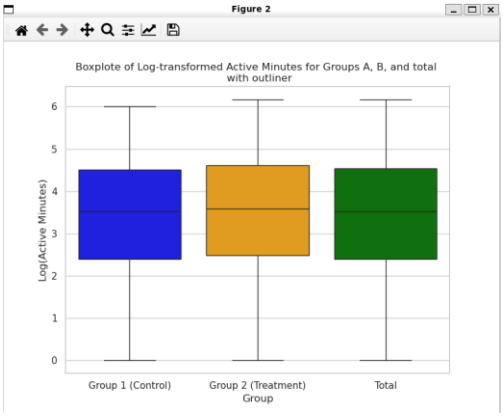
- Yes, there some outliers for both groups.

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

The reason for these outliners could be because of some users. Some users might have unusually high activity. This is far above the average user. Another reason for this could be error in data collection process.

```
MAximum active minutes recorded:  1121783.0
Whisker low (minimum in bixplot):  -253.5
Whisker high (maximum in bixplot):  470.5
(DS5110) (base) abdikarim@DESKTOP-K3E9SL8:~/hw3-ajimale$
```

6. Remove any data point that might be causing outliers.



7. Redo part 2 and 3 with the new data without those data points.

```
**************************************************************************************************
Filtered Group A size after remove outliner:  42561
Filtered Group B size after remove outliner:  43242
Filtered Group total size after remove outliner:  85803
**************************************************************************************************
Mean of active minutes for group A without outliners :  68.92453184840582
Median of active minutes for group A without outliners:  33.0
Mean of active minutes for group B without outliners:  74.78039868646223
Median of active minutes for group B without outliners:  35.0
**************************************************************************************************
T-statistic:  -9.281351809858139
P-vlaue:  1.710823033908345e-20
There is a statically difference between group 1 and group 2
Conclusion: There difference in active minute between the two group.
**************************************************************************************************
```

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

The data in t1 is now organized by used, variant_number, and active_mins (total). I did but the new data without the outliners in the new variable. After doing the t-test and p-value we noticed that there is a difference between group 1 and group 2.

Part 5: Digging Even Deeper (25 Points)

Now is the time to account for the data from t3. Answer the following questions:

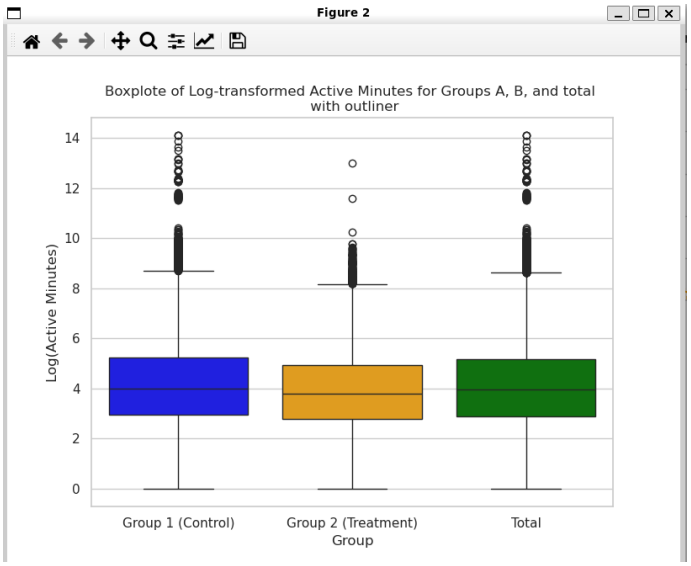
1. Why do we care about the data from t3?

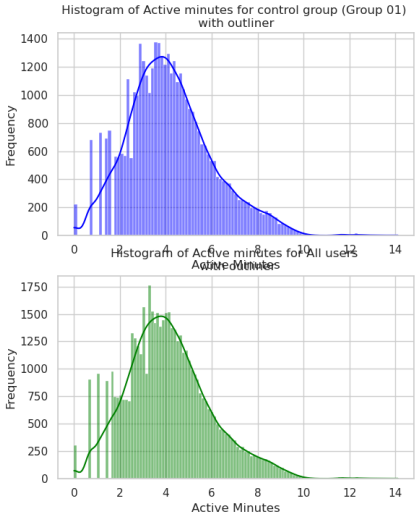
Understanding the data in t3 will help us know how the users acted before the experiment. It helps us know if the change in user behavior was due to the experiment or just natural variations. It will make the statistical analyses more accurate.

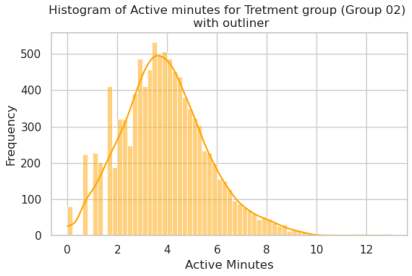
2. Accounting for the data from t3 rerun part 2 and 3.

```
Data is organized and save.
The number of element or rows at t3 is:  50000
**********************************************************************************************************
group A size:  40000
group B size:  10000
Mean of active minutes for group A:  877.481025
Median of active minutes for group A:  54.0
Mean of active minutes for group B:  322.5503
Median of active minutes for group B:  44.0
**********************************************************************************************************
T-statistic:  3.3490024169572163
P-vlaue:  0.0008116298164030166
There is a statically difference between group 1 and group 2
Conclusion: There difference in active minute between the two group.
**********************************************************************************************************
MAximum active minutes recorded:  1345714.0
Whisker low (minimum in bixplot):  -227.5
Whisker high (maximum in bixplot):  424.5
**********************************************************************************************************
Filtered Group A size after remove outliner:  43147
Filtered Group B size after remove outliner:  41469
Filtered Group total size after remove outliner:  84616
**********************************************************************************************************
Mean of active minutes for group A without outliners :  76.49025424710872
Median of active minutes for group A without outliners:  39.0
Mean of active minutes for group B without outliners:  64.20543056258892
Median of active minutes for group B without outliners:  36.0
**********************************************************************************************************
T-statistic:  21.487266156851696
P-vlaue:  3.846773532599824e-102
There is a statically difference between group 1 and group 2
Conclusion: There difference in active minute between the two group.
**********************************************************************************************************
```

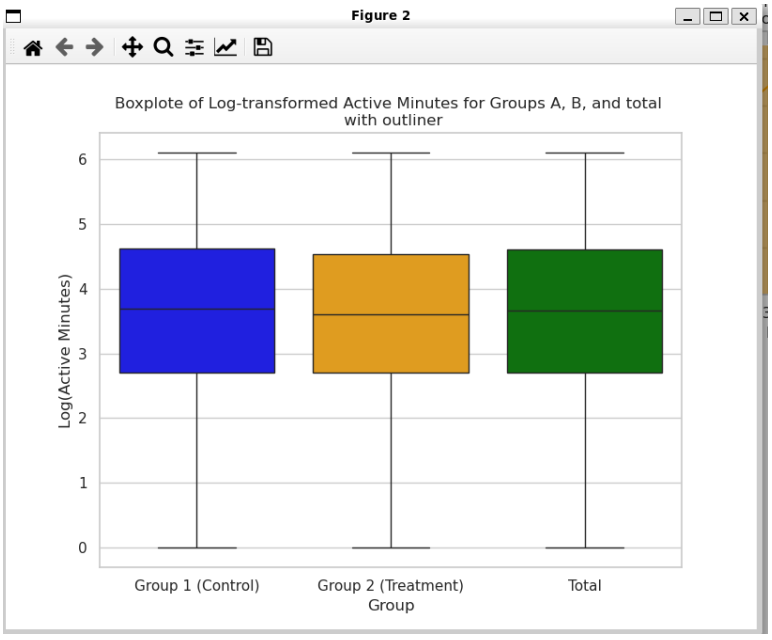
With outliners:

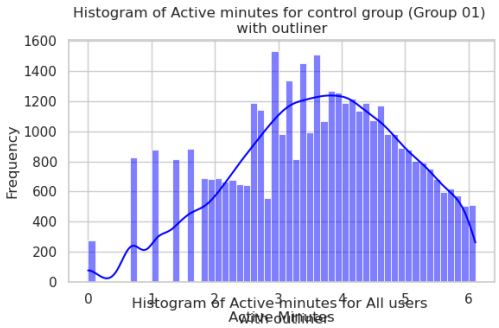


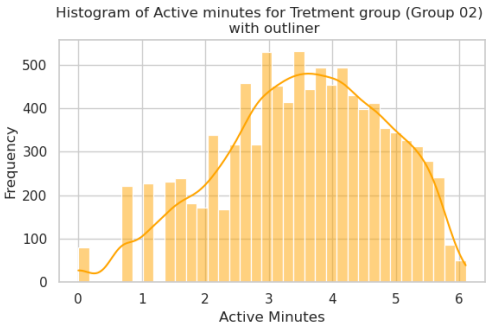


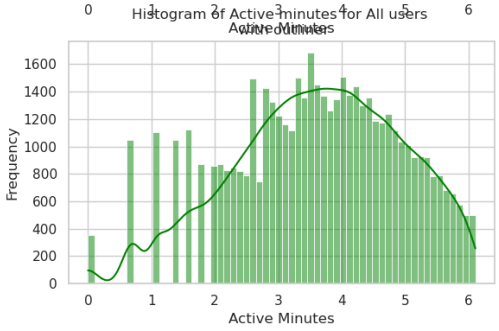


Without outliners:







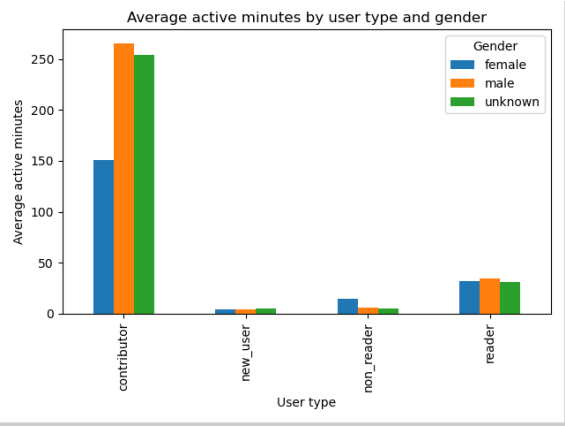


3. Are there any new conclusions?

The treatment group (group B) had lower active minutes compared to control group (group A) in both outliner and without outliners. From the data we get from the t3 file there are differences between the two group.

Part 6: Exploring other conclusions (10 Points)

Can you come up with any other conclusion with the data given in t4? If So, what are they? This is an open ended. This is left open ended to allow you to further explore the data that is given.



The t4 has gender uid, gender, and user_type. We can use active minutes for each user with gender. This will help us know if there is one gender active more than the other. This can be helpful to know the best target we can focus on to increase the amount of time the user on the program.

Part 7: Summarize Your Results (10 Points)

Write a summary for each part of this assignment and how it impacted your results.

The object of this object is to study the impact of a layout change on user engagement. This first thing we did is to identified what information each file contains. We did re-organize some of the files to make it easy to use them and analysis the data. We notice that having outliners may effect on the result on that study. In t1 with the outliners we come to conclusion that there is no difference in active minutes between the two groups. After we removed the outliners which were due to some unusual activity of users, the conclusion was there is a difference in active time between the groups. Studying the data in t3 which is before the experiment gives a clear give a clear picture of user engagement before and after the experiment. Analyzing the data in t4 shows how different user types and genders can help know potential targets.