HW3: Experiment Analysis Report

Student Name: Bini Chandra

Date: 02/07/2025

Part 1: Getting to know your data

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

This file contains data about the users' active minutes after the experiment started. It has three columns: uid, dt, and active_mins. Each row represents the total active minutes(active_mins) spent by a user(uid) on a specific date (dt) after the experiment began. If a user does not visit the site on a given date, they will not have an entry for that date.

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

This file contains data about the users' treatment assignments. It has four columns: uid, variant_number, dt, and signup_date. Every user belongs to either the control group(variant_number=0) or the treatment group(variant_number=1). Each row represents the experiment assignment of a unique user. The column 'dt' is 2019-02-06 for all users, while 'signup_date' varies per user.

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

This file contains data about the users' active minutes before the experiment started. It has the same structure as "t1_users_active_mins.csv," with three columns: uid, dt, and active_mins. However, the dates in this file (dt) are from before the experiment started. This will help us compare user activity before and after the update.

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

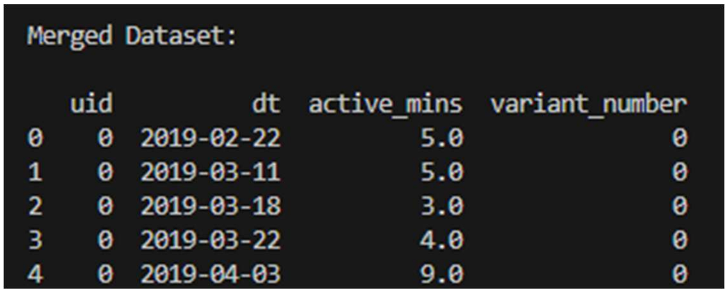
This file contains data about some user attributes. It has three columns: uid, gender, and user_type. Each row represents a unique user's attributes, such as gender and activity level ('new_user', 'non_reader', 'reader', or 'contributor').

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

This file explains the structure of all the dataset files and helps us understand what each column represents.

Part 2: Organizing the Data

The overall objective of this study is to see if the new layout increases user engagement. For this, we need data on how long users stay active and which group they are in (control or treatment). The file *t1_user_active_min.csv* has user activity but doesn't say which group a user belongs to. That information is in *t2_user_variant.csv*, so I have merged both files using *uid*. The final dataset is now structured with userID, date, active minutes, and experiment group and ready for statistical analysis.



```
Merged Dataset:

   uid          dt  active_mins  variant_number
0    0  2019-02-22          5.0               0
1    0  2019-03-11          5.0               0
2    0  2019-03-18          3.0               0
3    0  2019-03-22          4.0               0
4    0  2019-04-03          9.0               0
```

Part 3: Statistical Analysis

Control Group (A) = Users who did not receive the new platform update.
Treatment Group (B) = Users who received the new platform update.
Metric = Total active minutes spent on the platform.

Let's now perform hypothesis testing to check if the update significantly increased engagement.

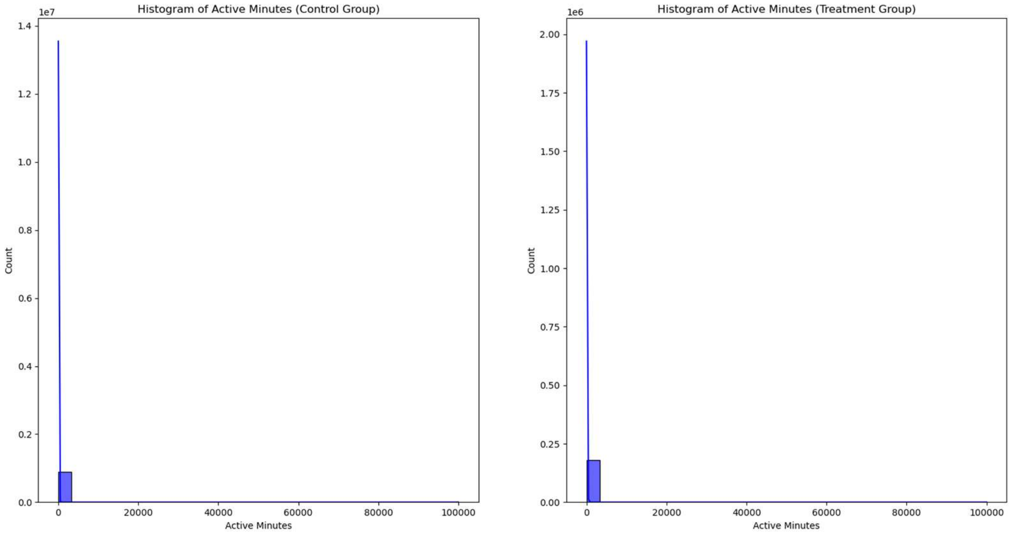
Null Hypothesis: The new platform update does not increase active minutes.

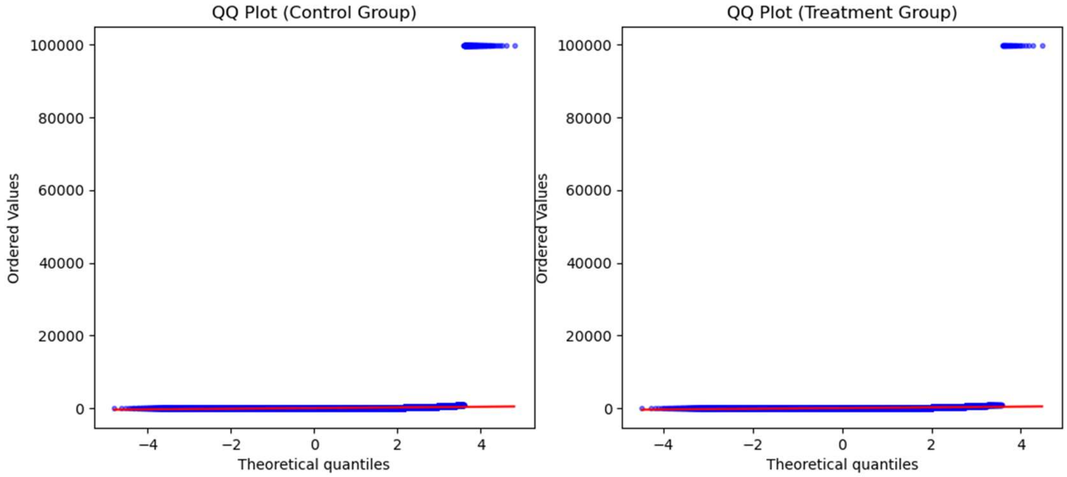
Alternative Hypothesis: The new platform update does increase active minutes.

Let's set α = 0.05. If p-value < 0.05, we will reject the Null Hypothesis and conclude that the update improved engagement.

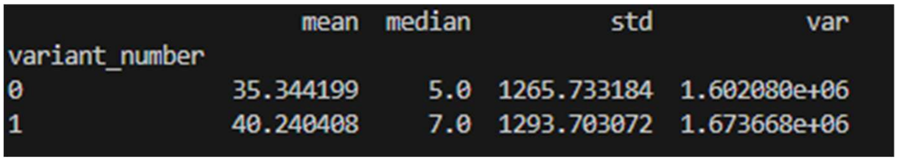
Now, let's check whether our data is distributed normally by plotting Histogram and QQ plots.

If it is normal, we will use a T-test; otherwise, we will use a Mann-Whitney U-Test.





The histograms show a right-skewed distribution and also the QQ plots indicate that the data is not normally distributed. Therefore, we will use the Mann-Whitney U-Test instead on T-test.



```
                     mean  median          std           var
variant_number
0               35.344199     5.0  1265.733184  1.602080e+06
1               40.240408     7.0  1293.703072  1.673668e+06
```

Th results of the Mann-Whitney U-Test shows that p-value is less than 0.05, so we reject the Null Hypothesis. This means the new platform significantly increased user engagement.

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

Yes, there is a significant difference. The p-value = 0.0 which is less than 0.05, so we reject the Null Hypothesis.

This means the new platform significantly increased user engagement.

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

The treatment group (new platform) shows a higher average and median active time compared to the control group. However, the high variance and standard deviation suggest that the data has large fluctuations. This indicates the presence of outliers that might be skewing the results.

3. What can you conclude based on that data?

The treatment group has higher active minutes than the control group, suggesting that the new platform is encouraging users to stay longer. The p-value is statistically significant (0.0), so we reject the Null Hypothesis.

However, the high variance and standard deviation as well graphs indicate the presence of outliers, which may be affecting the results.

Part 4: Digging a Little Deeper

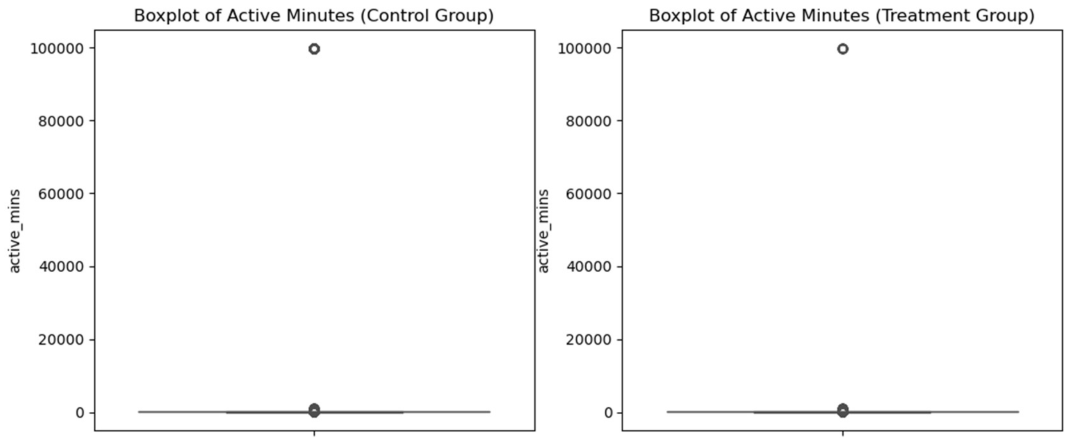
1. Can you trust that the results? Why or why not?

No, we cannot trust our results in Part3 because we noticed that some users have extremely high active minutes, which means outliers. Outliers can distort the mean and give misleading results. Hence, before making a final conclusion, we need to check if removing these outliers changes our conclusion.

2. Is the data normally distributed?

No, the data is not normally distributed. We already checked this in Part 3 by plotting histograms and QQ plots. The histograms show right-skewed distributions with long tails. The QQ plots also confirmed that data is not normally distributed.

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



4. Are there any outliers?

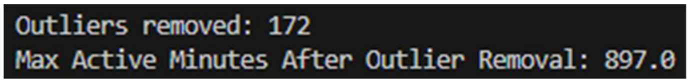
Yes! The boxplots confirm that there are many outliers in both the control and treatment groups. Some users have active minutes exceeding 1,440 per day, which is impossible because a day has only 1,440 minutes. So, they should be considered as outliers.

5. What might be causing those outliers?

These outliers might be caused by data entry errors or by some automated systems like bots. Theoretically, no user should have more than 1,440 minutes in a day (because 1,440 minutes = 24 hours).

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

We will filter out users with more than 1,440 active minutes per day and see how many outliers are removed. Even 1,440 active minutes per day is also unusual, but there may be some folks who are addicted to the new feature, so we can't take a risk by removing it.

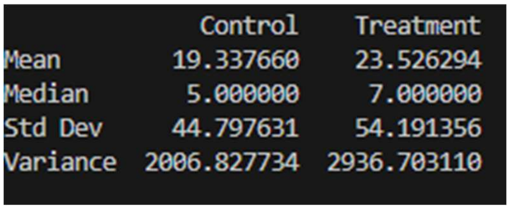


```
Outliers removed: 172
Max Active Minutes After Outlier Removal: 897.0
```

A total of 172 extreme outliers were removed from the data.

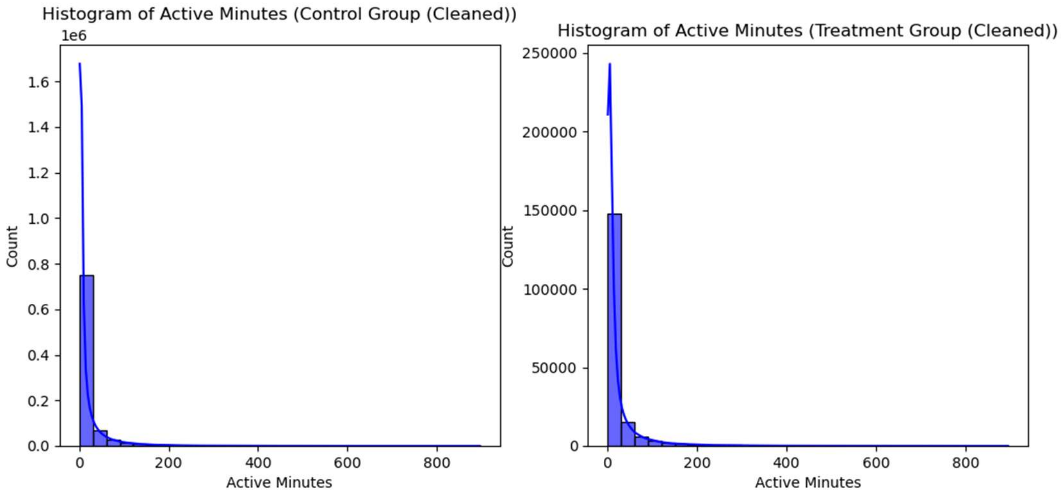
7. Redo part 2 and 3 with the new data without those data points.

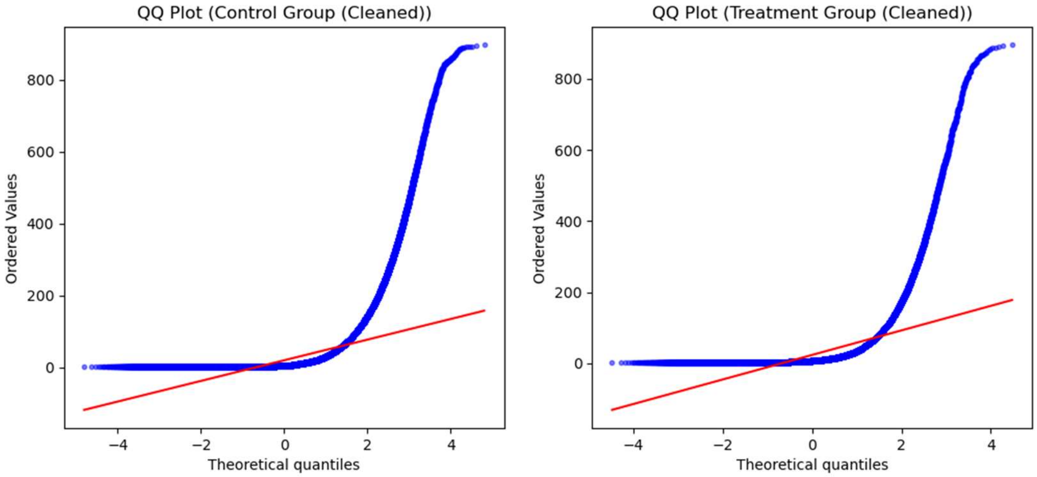
Now that we have removed the outliers, let's recompute the mean and median.

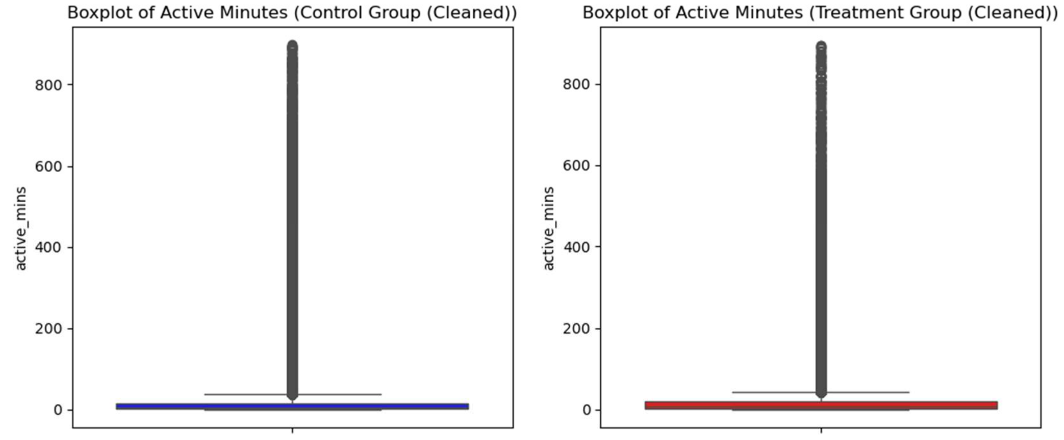


```
              Control    Treatment
Mean        19.337660    23.526294
Median       5.000000     7.000000
Std Dev     44.797631    54.191356
Variance  2006.827734  2936.703110
```

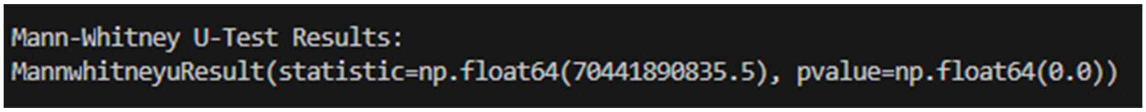
Let's again check normality of data by plotting histogram, and boxplots.







The treatment group still has a higher mean and median than the control group.



```
Mann-Whitney U-Test Results:
MannwhitneyuResult(statistic=np.float64(70441890835.5), pvalue=np.float64(0.0))
```

The Mann-Whitney U-Test still resulted in p-value = 0.0.

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

In Part 3, before removing outliers, the Treatment group had significantly higher engagement (p-value = 0.0).
However, even after removing outliers, the p-value remains 0.0, meaning the difference is still statistically significant.

This means the new platform update has significantly improved engagement.

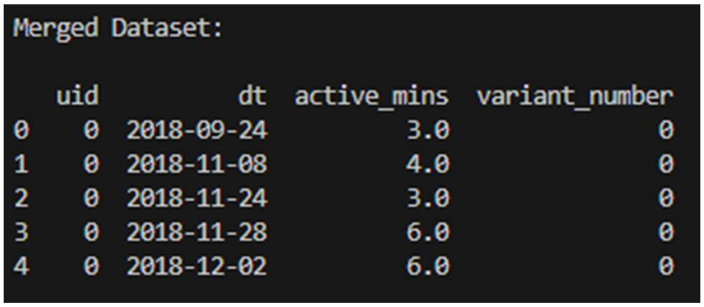
Part 5: Digging Even Deeper

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

We care about this file because it contains user engagement before the experiment started. We need to check if the treatment and control groups were equal before the update. If one group was already using the platform less, then their increase in engagement might not be fully caused by the update.

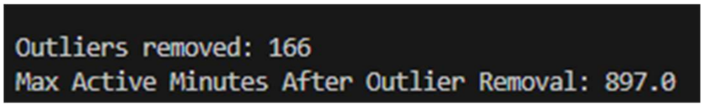
2. Accounting for the data from t3 rerun part 2 and 3.

- Merged t3 (pre-experiment data) with t2 (user assignments)
  This allowed us to see which users were in the control and treatment groups before the update.



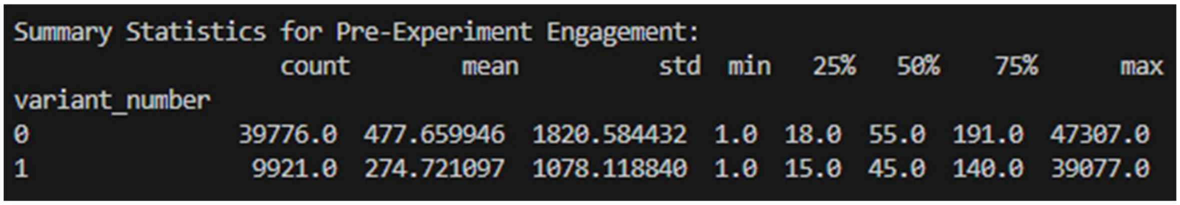
```
Merged Dataset:

   uid          dt  active_mins  variant_number
0    0  2018-09-24          3.0               0
1    0  2018-11-08          4.0               0
2    0  2018-11-24          3.0               0
3    0  2018-11-28          6.0               0
4    0  2018-12-02          6.0               0
```

- Removed outliers (users with more than 1,440 active minutes in a single day).
  This ensured our pre-experiment data was clean and realistic.

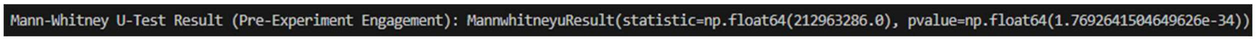


```
Outliers removed: 166
Max Active Minutes After Outlier Removal: 897.0
```

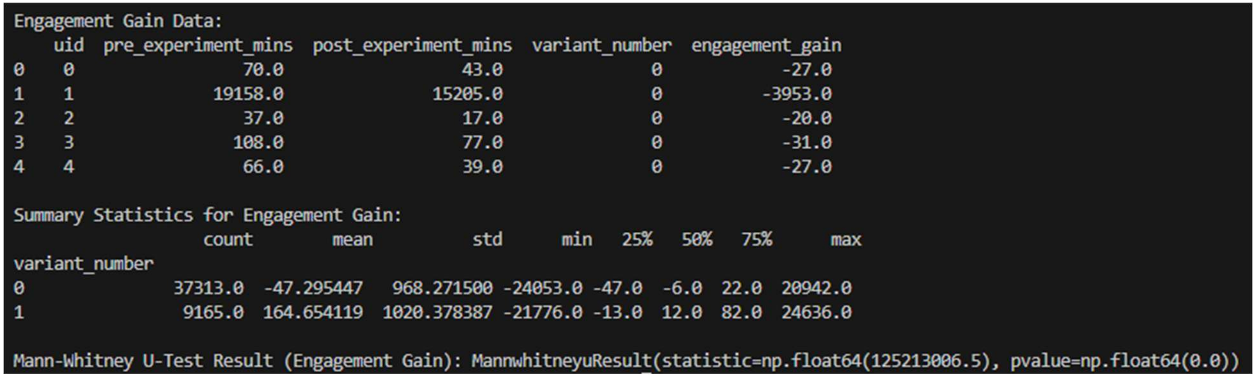
- Added up total pre-experiment active minutes per user
  Instead of looking at daily data, we calculated total minutes spent by each user before the experiment.
- Compared the control and treatment groups before the update
  -We calculated average (mean) and median active minutes for both groups.
  -We performed a Mann-Whitney U-Test to check if the groups were different before the update.
  -We calculated engagement gain (post-experiment minutes - pre-experiment minutes) to measure how much each group's engagement changed after the update.



```
Summary Statistics for Pre-Experiment Engagement:
                  count        mean          std  min   25%   50%    75%      max
variant_number
0               39776.0  477.659946  1820.584432  1.0  18.0  55.0  191.0  47307.0
1                9921.0  274.721097  1078.118840  1.0  15.0  45.0  140.0  39077.0
```

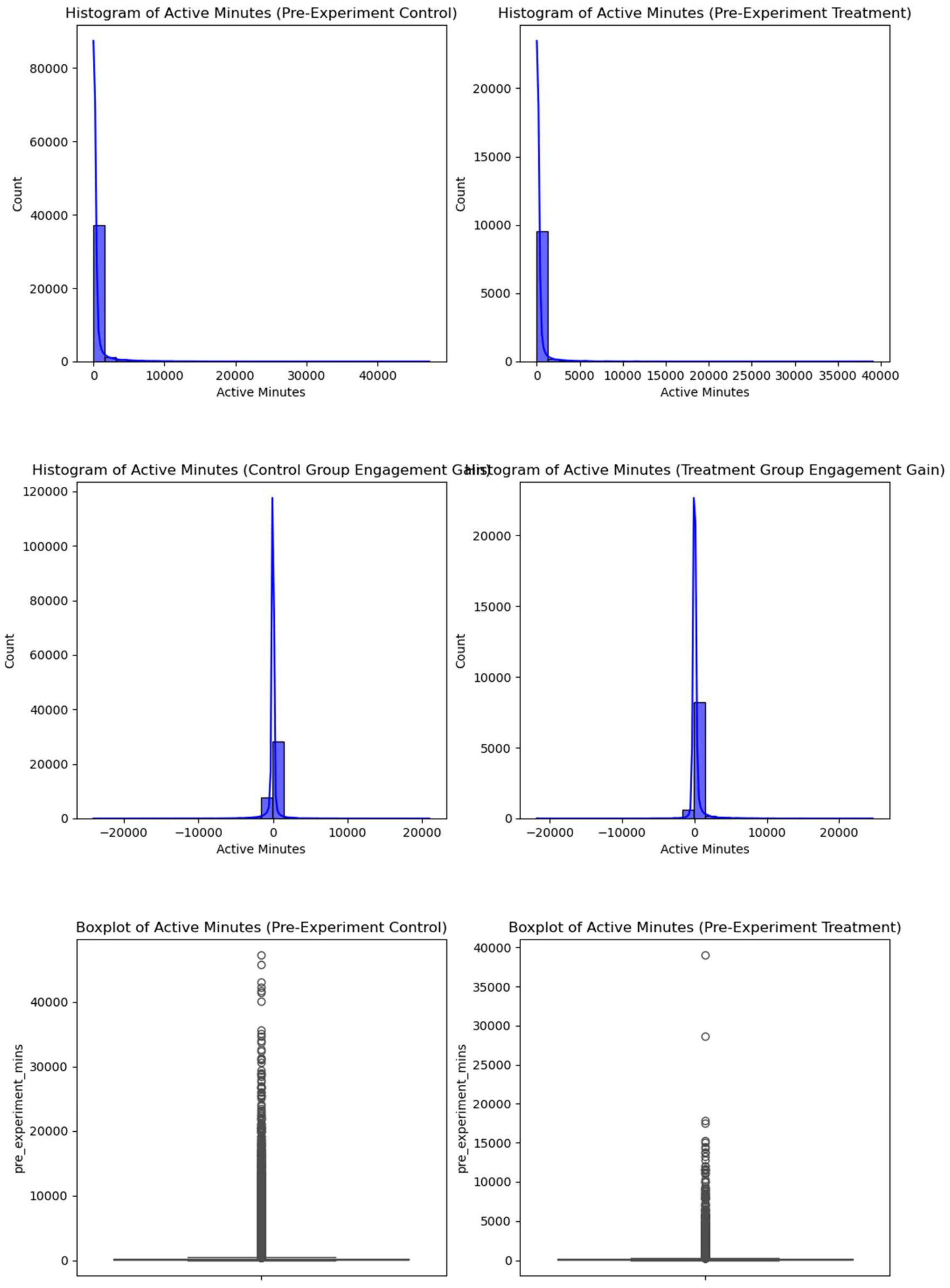


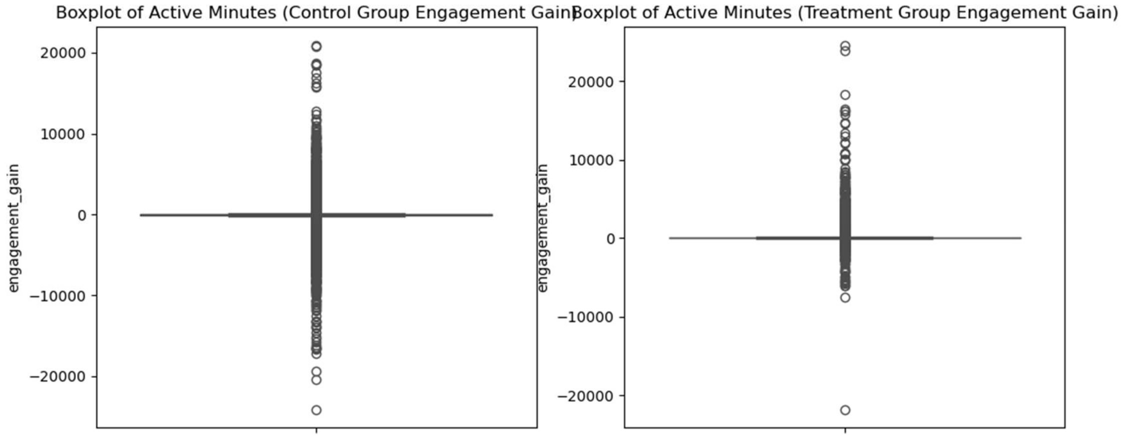
```
Mann-Whitney U-Test Result (Pre-Experiment Engagement): MannwhitneyuResult(statistic=np.float64(212963286.0), pvalue=np.float64(1.7692641504649626e-34))
```



```
Engagement Gain Data:
   uid  pre_experiment_mins  post_experiment_mins  variant_number  engagement_gain
0    0                 70.0                  43.0               0            -27.0
1    1              19158.0               15205.0               0          -3953.0
2    2                 37.0                  17.0               0            -20.0
3    3                108.0                  77.0               0            -31.0
4    4                 66.0                  39.0               0            -27.0

Summary Statistics for Engagement Gain:
                  count        mean          std      min   25%   50%   75%      max
variant_number
0               37313.0  -47.295447   968.271500 -24053.0 -47.0  -6.0  22.0  20942.0
1                9165.0  164.654119  1020.378387 -21776.0 -13.0  12.0  82.0  24636.0

Mann-Whitney U-Test Result (Engagement Gain): MannwhitneyuResult(statistic=np.float64(125213006.5), pvalue=np.float64(0.0))
```

- Plotted histograms and boxplots to see if one group was using the platform more than the other before the update.


Histogram of Active Minutes (Pre-Experiment Control)
Count
80000
60000
40000
20000
0
0
10000
20000
30000
40000
Active Minutes
Histogram of Active Minutes (Pre-Experiment Treatment)
Count
20000
15000
10000
5000
0
0
5000
10000
15000
20000
25000
30000
35000
40000
Active Minutes
Histogram of Active Minutes (Control Group Engagement Gain)
Count
120000
100000
80000
60000
40000
20000
0
−20000
−10000
0
10000
20000
Active Minutes
Histogram of Active Minutes (Treatment Group Engagement Gain)
Count
20000
15000
10000
5000
0
−20000
−10000
0
10000
20000
Active Minutes
Boxplot of Active Minutes (Pre-Experiment Control)
pre_experiment_mins
40000
30000
20000
10000
0
Boxplot of Active Minutes (Pre-Experiment Treatment)
pre_experiment_mins
40000
35000
30000
25000
20000
15000
10000
5000
0



3. Are their any new conclusion?

- Before the update, the treatment group was using the platform less than the control group. This means the two groups were not equal from the start.
- After the update, the treatment group's engagement increased a lot (+164.65 minutes), while the control group's engagement dropped slightly (-47.3 minutes).
- The p-value is 0.0, so the difference between the groups is real and not due to random chance.
- This shows that the update helped increase engagement, but some of the increase might be because the treatment group was catching up to the control group, not just because of the update.
- The update worked, but we should remember that the groups were not perfectly fair at the beginning.

Part 6: Exploring other conclusions

Part 7: Summarize Your Results

Part 1: Understanding the Data

We explored the datasets to understand what each file contains. The files track user engagement before and after the experiment, assign users to groups (control or treatment), and provide user attributes. This helped us prepare for analysis.

Part 2: Merging the Data

We merged post-experiment user activity (t1) with user assignments (t2) so we could compare engagement between the control and treatment groups.

Part 3: Initial Analysis

We tested if the platform update increased engagement. The treatment group had significantly higher engagement, with a p-value of 0.0. However, we found outliers with unrealistic active minutes, which could affect the results.

Part 4: Removing Outliers

We removed users with more than 1,440 active minutes per day. After cleaning the data, the treatment group still had higher engagement, confirming the update had a real effect.

Part 5: Checking Pre-Experiment Engagement

We analyzed t3 (pre-experiment engagement) to check if the two groups were equal before the update. The treatment group already had lower engagement before the update, meaning some of their increase was just them catching up, not entirely due to the update.

Part 6: Exploring Other Insights

The t4 dataset (user attributes) could help us understand if certain user types (e.g., new users, contributors) benefited more from the update.

Final Conclusion

The update increased engagement, but the treatment group was already behind before the experiment. The results are significant, but the experiment was not perfectly balanced.