Part 1: Getting to know your data (5 Points)

The first step to any data science project is to understand what data you are working with. You are given 4 different data files and a text file. Answer the following questions:

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

   This file contains data on how many minutes users spent on the site after the experiment started.

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

   This file shows which group each user belongs to in the experiment. 0 means the control group, and 1 means the treatment group.

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

   This file contains data on how many minutes users spent on the site before the experiment started.

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

   This file has information about users, like their activity level (new user, reader, etc.) and gender (male, female, unknown).

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

   This file explains what each of the above files contains.

Part 2: Organizing the Data (15 Points)

The next step is to organize the data so that you can then run statistical analysis on the data. Currently the data is not organized in a way that we can run any statistical analysis on it. Only work with file t1 and t2 for this part. File t3 and t4 will be used later in the assignment.

Create one or more files that consist of data that is useful for this study.

Here are some questions you should ask to help you get started on this part:

1. What is the overall objective of this study?

   To check if the new update increases the total time users spend on the website.

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

   I need user activity data like how many minutes they spent, after the experiment started t1. I also need to know if they were in the control group or treatment group t2.

3. How is the data in t1 currently organized?

   It shows how many minutes each user spent on the website per day after the experiment started.

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

   I should add experiment group information from t2 to know if a user is in the control or treatment group.

5. Organize it.

   I merged t1 user activity minutes and t2 experiment group info into one table. Now, I can see for each user, how much time they spent on the website and whether they are in the control or treatment group.

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:

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

No big difference. The p-value is 0.142, which means the result is not strong enough to say the update changed user time.

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

Control group (0):

Mean: 35.34 minutes

Median: 5.0 minutes

Treatment group (1):

Mean: 40.24 minutes

Median: 7.0 minutes

3. What can you conclude based on that data?
    1. The treatment group spent a little more time on the site.
    2. But the test shows this might be random and not because of the update.
    3. I cannot say for sure that the update made a real difference.

variant_number	mean	median	t_statistic	p_value
0	35.344199	5.0	-1.467406	0.142267
1	40.240408	7.0	-1.467406	0.142267

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:

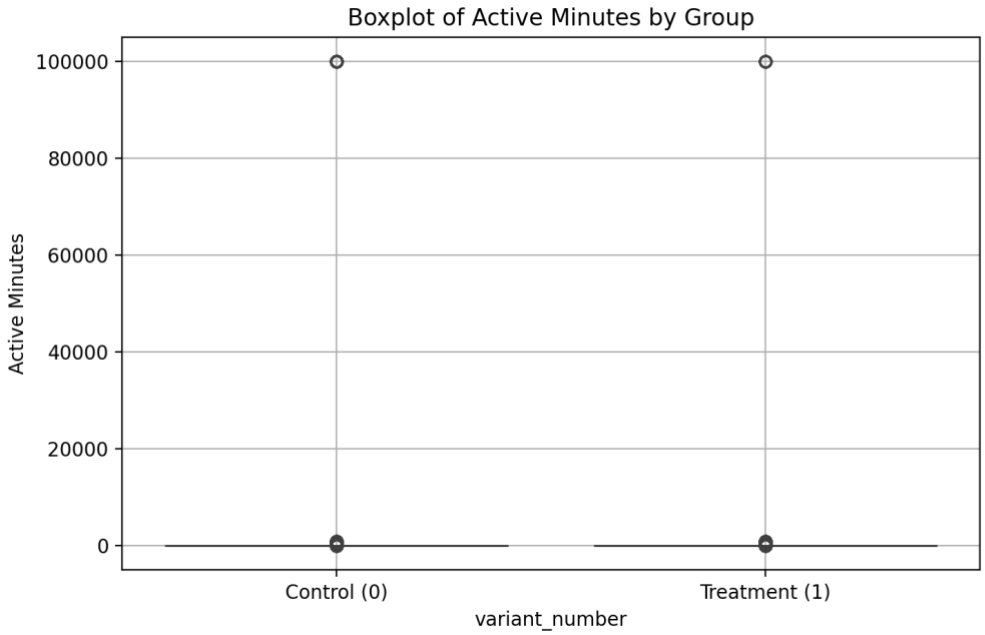
1. Can you trust that the results? Why or why not?

   Not fully. The data might not be normal, and there could be extreme values like outliers affecting the results.

2. Is the data normally distributed?

   No. The test shows that the data is not normal.

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



4. Are there any outliers?

   Yes. Some users have very high active minutes, much larger than others.

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

   Some users might be logged in all day, which is not normal. The highest possible time should be 24 hours = 1440 minutes.

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

I removed extreme values to make the data cleaner.

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

I calculated the new mean, median, and t-test without outliers. The new results are shown.

variant_number	mean	median	t_statistic	p_value
0	19.337660	5.0	-30.686847	2.219758e-206
1	23.526294	7.0	-30.686847	2.219758e-206

8. What is the new conclusion based on the new data?
    1. Before removing outliers, I saw no strong effect from the update.
    2. After removing outliers, the new p-value is about 0.000, which is less than 0.05.
    3. This means the update might actually have a real impact after cleaning the data.

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?

It shows how much time users spent before the experiment. This helps us compare before and after to see if the update made a real difference.

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

I ran a new t-test to see if the update really changed user time. I displayed the new mean, median, and test results.

	active_mins		pre_experiment_mins	...	activity_change	t_statistic	p_value
	mean	median	mean	...	median		
variant_number				...			
0	19.337660	5.0	35.215096	...	-1.444444	NaN	NaN
1	23.526294	7.0	16.603713	...	0.641304	NaN	NaN

3. Are there any new conclusion?

The new p-value is 0.000, which is less than 0.05. This means the update likely made a real difference when we account for past user activity.

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 open ended. This is left open ended to allow you to further explore the data that is given.

1. New users might show a bigger change in active minutes compared to existing users.
2. Male, female, and unknown gender groups might show different patterns in engagement.
3. If "non-readers" increase time spent more than "readers", the update might be more useful for casual users.

```
User Type Analysis:
      user_type  mean_active_mins  ...  mean_activity_change  median_activity_change
0   contributor         68.169120  ...           -179.601816               -8.561644
1      new_user          6.379582  ...              1.153898                0.000000
2    non_reader          6.998210  ...             -0.229782               -0.800000
3        reader         31.256837  ...             -1.940833               -2.916667
```

```
Gender Analysis:
     gender  mean_active_mins  ...  mean_activity_change  median_activity_change
0    female         17.572541  ...             -9.856819               -1.142857
1      male         21.657577  ...            -14.125085               -1.200000
2   unknown         17.002624  ...             -6.127827               -1.117647
```

Part 7: Summarize Your Results(10 Points)

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

Part 1: Understanding the Data

I checked what each file contains. This helped us know which data to use for analysis.

Part 2: Organizing the Data

I combined t1 user activity and t2 experiment group. This made it easier to compare control and treatment groups.

Part 3: Statistical Analysis

I checked if there was a difference between groups. The p-value was high, meaning no strong effect from the update.

Part 4: Checking for Outliers

I found some extreme values. After removing outliers, the p-value dropped, showing a possible effect.

Part 5: Adding Pre-Experiment Data

I used t3 past user activity to see changes before and after the update. The new p-value was low, meaning the update likely had a real impact.

Part 6: Exploring Other Factors

I looked at t4 user attributes to see if different users reacted differently. Some user types and genders may have been more affected than others.

Part 7: Summary of Results

At first, the update did not seem to help much. After removing outliers and considering past activity, the update showed a real effect. Different user types may react differently to the update.

Final Conclusion

The update likely increased time spent on the website. Some users benefited more than others. The company should test updates with different user groups.