DS5110 hw3
Name: Wenyi Ye
Date: Feb.7.2025

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

1. **t1_user_active_min.csv:**

This file contains active minutes users spent on the site from the start of the experiment. Each row reports one user's total time spent on the site for each date. If the user didn't visit the site on a given day, no record for the user on this date will be found.

2. **t2_user_variant.csv:**

This is the user treatment assignment. It tells us which users were placed in the control group (0) and which users were placed in the treatment group (1). It also has the date that each user joined the experiment (which is always '2019-02-06') and their signup date.

3. **t3_user_active_min_pre.csv:**

This table records users' active minutes before the experiment was started. It is in the same format as t1_user_active_min.csv, but the date range covers days before the experiment.

4. **t4_user_attributes.csv:**

This file contains user attributes like their user type ('new_user', 'non_reader', 'reader', or 'contributor') and gender ('male', 'female', or 'unknown').

5. **table_schema.txt:**

This file describes the format and interpretation of data in the other four CSV files, what the columns represent, and the intent of each dataset.

Part 2: Organizing the Data (15 Points)

1.The goal is to figure out if the new design and features on the social media platform have made users spend more time on the site. This will help the company decide whether to launch the update for everyone, with the aim of boosting user engagement and increasing ad revenue.

2.User Activity Data After the Experiment Started (t1_user_active_min.csv): This file tracks how much time users spent on the platform after the update was rolled out. It helps us understand whether the update influenced user engagement.

User Treatment Assignment (t2_user_variant.csv): This file tells us which users were in the control group (they didn’t get the update) and which were in the treatment group (they received the update). This distinction is key for comparing the two groups and seeing if the update had an impact.

3. The data in t1_user_active_min.csv is structured to show daily activity for each user. Each row represents a specific user (identified by uid) and a specific date (dt), along with the total number of minutes they spent on the platform that day (active_mins). This helps us track how much time users are engaging with the site after the experiment started. However, one thing to note is that this dataset doesn’t specify whether a user was part of the control group

or the treatment group. To figure that out, I need to combine this data with the information from t2_user_variant.csv, which tells us which group each user belonged to.

4. To make sense of the data, I combined t1_user_active_min.csv with t2_user_variant.csv. Right now, the activity data shows how much time each user spent on the platform daily, but it doesn't tell us whether they were in the control group or the treatment group By merging these two datasets, we can add a new column that specifies the experiment group for each user 0 for control and 1 for treatment. This way, each row in the final dataset will include the user ID, the date, the number of active minutes they spent on the platform that day, and their group assignment. With this structure, we can directly compare engagement levels between the two groups and see if the update actually made a difference.

5.

This is the code snippet for merging t1_user_active_min.csv and 2_user_variant.csv

```python
import pandas as pd

# Load the datasets
t1_path = "t1_user_active_min.csv"
t2_path = "t2_user_variant.csv"

t1_data = pd.read_csv(t1_path)
t2_data = pd.read_csv(t2_path)

# Merge the datasets on 'uid' to include the experiment group
merged_data = t1_data.merge(t2_data[['uid', 'variant_number']], on='uid', how='left')

# Save the merged data to a new CSV file
merged_data.to_csv("merged_user_activity.csv", index=False)

print(merged_data.head())
```

This is the organized data for t1_user_active_min.csv and 2_user_variant.csv

	A	B	C	D	E	F
1		uid	dt	active_mins	variant_number	
2	0	0	22/02/2019	5	0	
3	1	0	11/03/2019	5	0	
4	2	0	18/03/2019	3	0	
5	3	0	22/03/2019	4	0	
6	4	0	03/04/2019	9	0	
7	5	0	06/04/2019	1	0	
8	6	0	17/04/2019	1	0	
9	7	0	07/05/2019	3	0	
10	8	0	14/05/2019	1	0	
11	9	0	19/05/2019	1	0	
12	10	0	22/05/2019	3	0	
13	11	0	14/06/2019	5	0	
14	12	0	16/06/2019	2	0	
15	13	1	07/02/2019	79	0	
16	14	1	09/02/2019	211	0	
17	15	1	11/02/2019	23	0	
18	16	1	13/02/2019	49	0	
19	17	1	14/02/2019	143	0	
20	18	1	15/02/2019	50	0	
21	19	1	16/02/2019	13	0	
22	20	1	18/02/2019	82	0	
23	21	1	21/02/2019	192	0	
24	22	1	25/02/2019	203	0	
25	23	1	26/02/2019	216	0	
26	24	1	01/03/2019	74	0	
27	25	1	03/03/2019	203	0	
28	26	1	04/03/2019	144	0	
29	27	1	05/03/2019	118	0	
30	28	1	06/03/2019	471	0	
31	29	1	07/03/2019	203	0	

Part 3: Statistical Analysis (10 Points)

1. After running an independent t-test, I found that the test statistic is around -1.47, with a p-value of 0.142. Because the p-value is higher than the typical threshold of 0.05, I don't have enough evidence to reject the null hypothesis. In simpler terms, this means there isn't a statistically significant difference in the amount of time users spent on the platform between the control group and the treatment group. So, based on this analysis, the update doesn't seem to have had a noticeable impact on user engagement.

This is the code snippet for independent t-test

```python
import pandas as pd
import scipy.stats as stats

t1_path = "Data/t1_user_active_min.csv"
t2_path = "Data/t2_user_variant.csv"

t1_data = pd.read_csv(t1_path)
t2_data = pd.read_csv(t2_path)

# Merge the datasets on 'uid' to include the experiment group
merged_data = t1_data.merge(t2_data[['uid', 'variant_number']], on='uid', how='left')

# Separate the two groups
control_group = merged_data[merged_data["variant_number"] == 0]["active_mins"].dropna()
treatment_group = merged_data[merged_data["variant_number"] == 1]["active_mins"].dropna()

# Perform an independent t-test
t_stat, p_value = stats.ttest_ind(control_group, treatment_group, equal_var=False)

# Compute mean and median for both groups
group_stats = merged_data.groupby("variant_number")["active_mins"].agg(["mean", "median"])

# Print results
print("Group Statistics (Mean and Median):")
print(group_stats)
print("\nIndependent t-Test Results:")
print(f"T-Statistic: {t_stat:.4f}")
print(f"P-Value: {p_value:.4f}")

# Interpretation
alpha = 0.05  # Significance level
if p_value <= alpha:
    print("\nResult: There is a statistically significant difference between the two groups.")
else:
    print("\nResult: There is no statistically significant difference between the two groups.")
```

2. This is the mean and median of both group, these values summarize the central tendency of user engagement for each group.

variant_number	mean	median	
0	35.3442	5	
1	40.24041	7	

3. Since the p-value is higher than the usual cutoff of 0.05, we don't have strong statistical evidence to say that the new platform layout made a significant difference in how much time users spent on the site. While there are some small differences in the average and median values between the groups, these differences aren't big enough to be considered meaningful from a statistical standpoint. This suggests that the new design might not have had the impact we were hoping for when it comes to boosting user engagement.

Part 4: Digging a Little Deeper (25 Points)

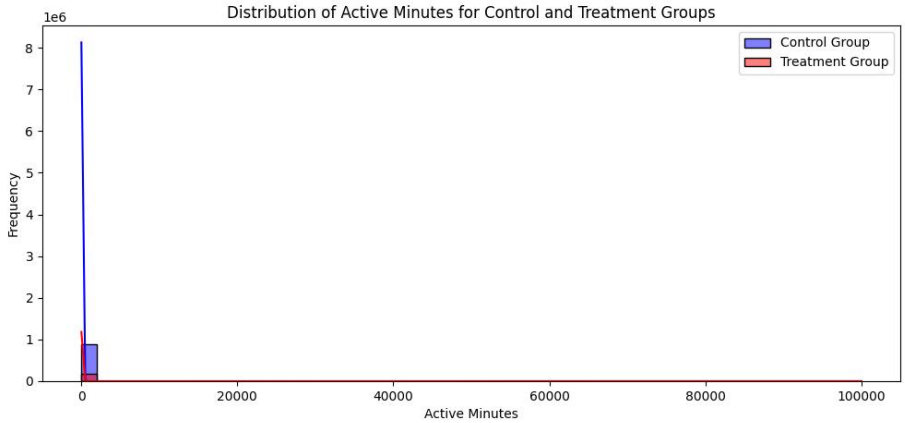
1. The results should be interpreted with caution. The Shapiro-Wilk normality test indicates that the data is not normally distributed

Control Group: Statistic = 0.4242, p-value = 3.14e-83
Treatment Group: Statistic = 0.0120, p-value = 1.45e-95

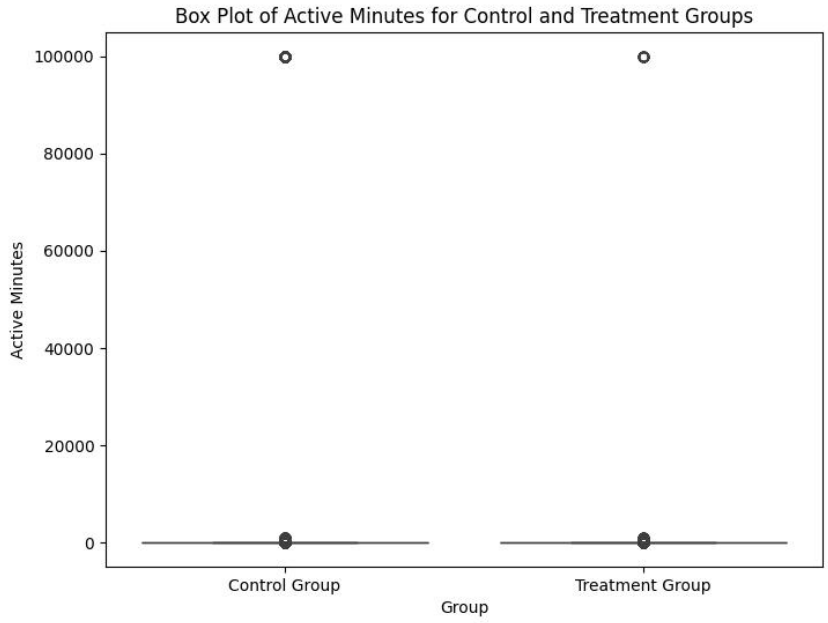
Since the p-values are extremely small, This violates the assumption required for an independent t-test, meaning the results could be misleading. Additionally, there are extreme

outliers, some users had 99,999 active minutes, which is unrealistic, which further affects the reliability of the conclusions.

2. No, the Shapiro-Wilk normality test produced an extremely small p-value of for both the control and treatment groups, meaning we reject the null hypothesis and conclude that the data is not normally distributed. The histogram also confirms that the data is highly skewed.



3. The box plot shows a wide range of values, with some extreme outliers present in both groups.



4. Yes, the maximum recorded value in t1_user_active_min.csv is 99,999 minutes, which is clearly unrealistic, as there are only 1,440 minutes in a day. This indicates major outliers in the dataset.

5. Since the maximum possible minutes per day is 1,440, any values significantly above this are likely errors or unrealistic user behavior.

6. All records where active_mins greater than 1,440 we're all removed.

7. The cleaned mean and median values for each group after removing extreme outliers:

Control Group (variant_number = 0)
Mean Active Minutes: 19.34
Median Active Minutes: 5.0

Treatment Group (variant_number = 1)

Mean Active Minutes: 23.53
Median Active Minutes: 7.0

T-Test Results After Removing Outliers
T-Statistic: -30.686846737487123
P-Value: 2.219758340477041e-206

8. The new t-statistic is -30.69, and the p-value is nearly 0 (2.22e-206), Since $p < 0.05$, we reject the null hypothesis, meaning there is a statistically significant difference between the control and treatment groups. This suggests that the new platform layout did have a measurable impact on user engagement, which was previously masked by extreme outliers.

Part 5: Digging Even Deeper (25 Points)

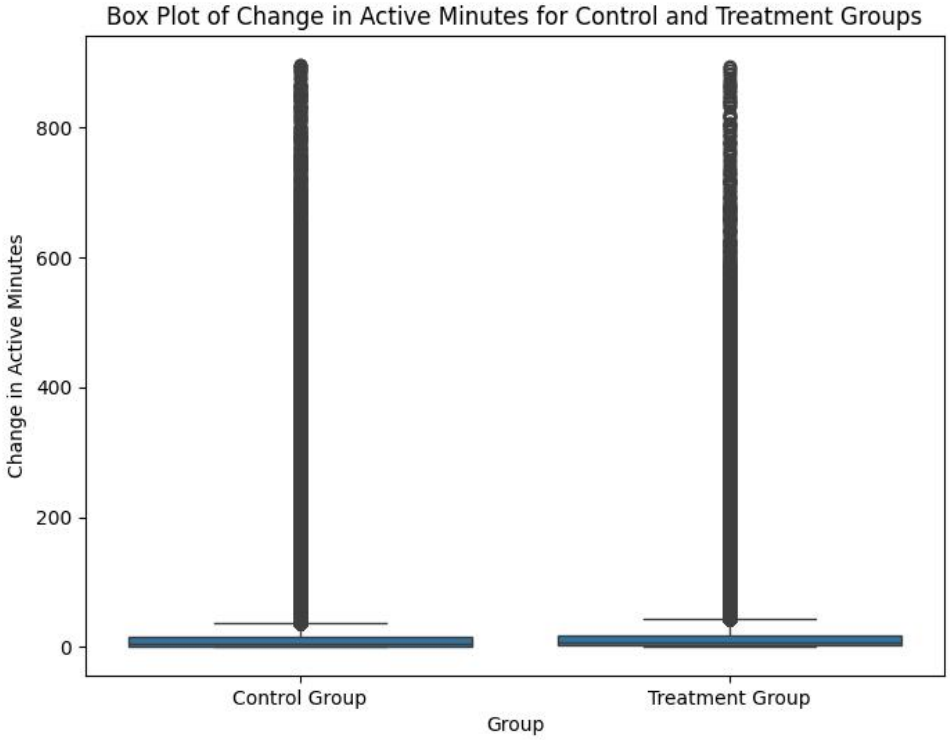
1. The t3_user_active_min_pre.csv dataset contains pre-experiment active minutes, which allows us to measure the true impact of the experiment by considering changes in user behavior. Instead of just comparing post-experiment engagement, we can now analyze, How much time users spent before the update and after the update, whether the treatment group truly increased engagement compared to the control group, independent of prior activity and If pre-existing differences in user behavior explain the observed effects. By accounting for pre-experiment activity, we get a more reliable measure of the causal impact of the new platform update.

2.

Mean, Median, T-Test, T-Statistics and P-Value when data from t3 was taken into account

```
Mean and Median Change in Active Minutes:
                    mean  median
variant_number
0              19.337660     5.0
1              23.526294     7.0

T-Test Results on Change in Active Minutes:
T-Statistic: -30.686846737487123
P-Value: 2.219758340477041e-206
```



3. The p-value (2.22e-206 very close to 0) is still extremely significant, this means that the change in engagement between the control and treatment groups is not due to random chance. The treatment group still showed a significant increase in active minutes compared to the control group.This suggests that the new platform update had a real effect on increasing engagement, even after accounting for prior user behavior.

Part 6: Exploring other conclusions (10 Points)

After merging user attributes (t4) with our cleaned dataset, we can analyze how different user types and genders responded to the platform update.

The table shows the mean and median change in active minutes for different user types and genders:

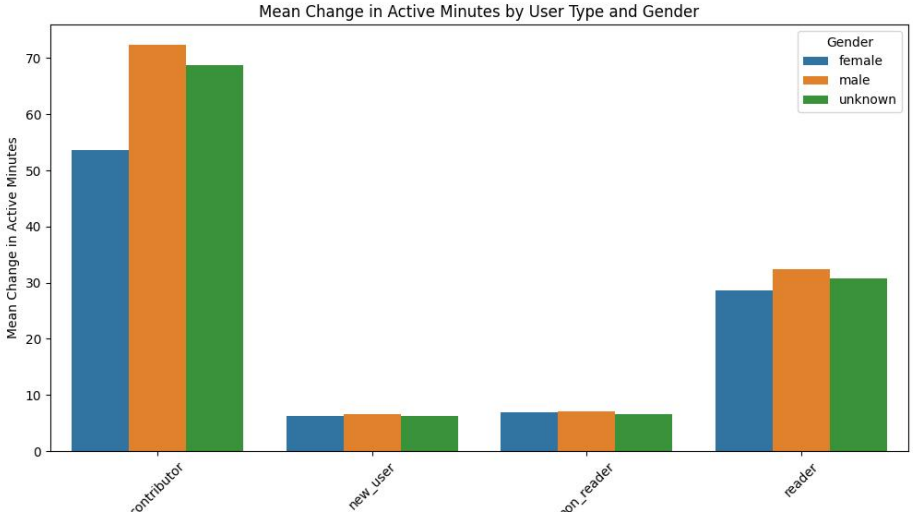
	A	B	C	D
1	user_type	gender	mean	median
2	contributor	female	53.62801369	21
3	contributor	male	72.37966676	32
4	contributor	unknown	68.71606994	27
5	new_user	female	6.198917944	3
6	new_user	male	6.57788587	3
7	new_user	unknown	6.1827799	3
8	non_reader	female	6.922043365	3
9	non_reader	male	7.139087693	3
10	non_reader	unknown	6.550591653	3
11	reader	female	28.58532818	11
12	reader	male	32.35431967	13
13	reader	unknown	30.82066814	12

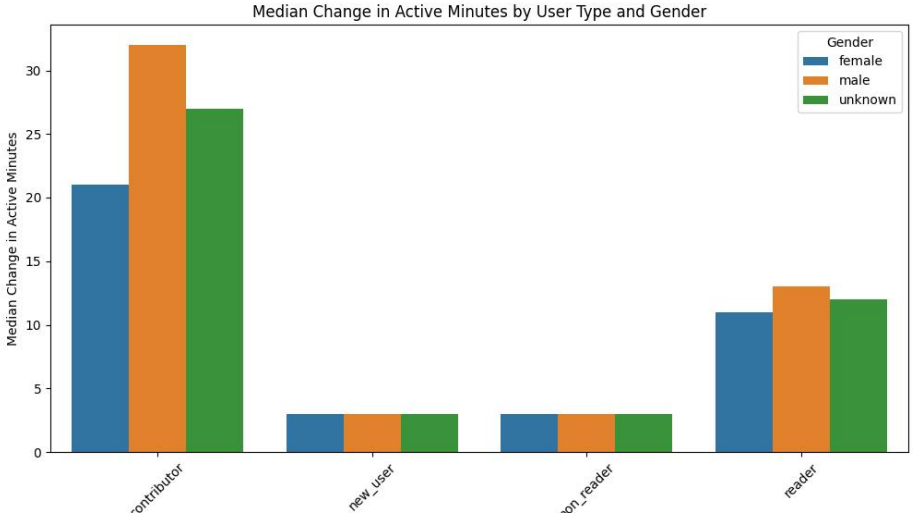
Contributors had the highest increase in active minutes, with males gaining +72.38 minutes on average. Contributors engage more actively than other user types, and the new update may have enhanced their experience the most.

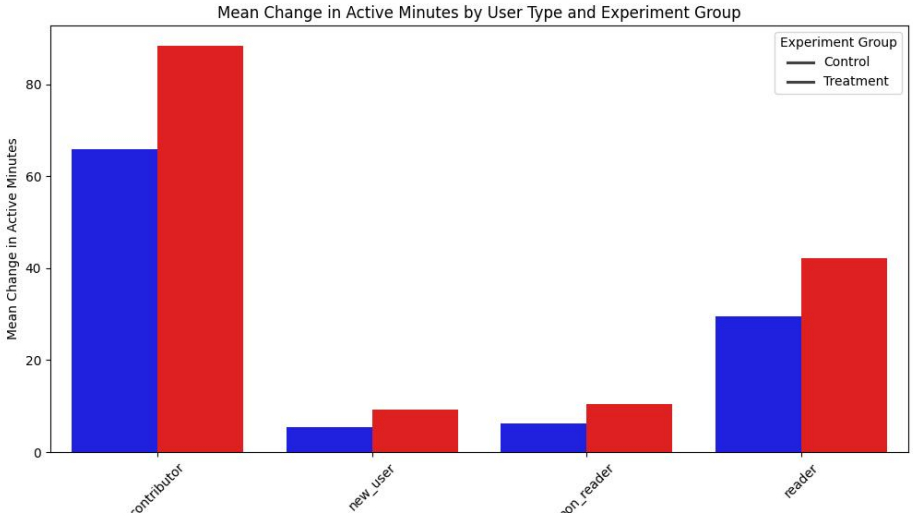
Readers showed a moderate increase in active minutes (+28 to +32 minutes on average). They engaged more than non-readers or new users, suggesting the new platform design benefited frequent users.

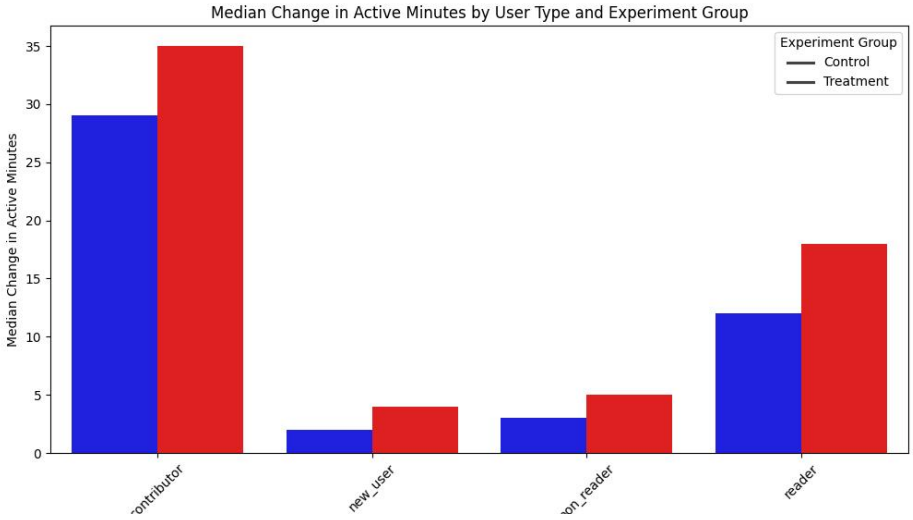
New users and non-readers only gained ~6 minutes on average, with a median of +3 minutes. This suggests that the new platform features did not strongly influence less engaged users. The company may need to optimize onboarding strategies to retain new users.

Across all user types, males increased their active minutes more than females. This may indicate differences in content preferences or how different demographics interact with the platform.









Part 7: Summarize Your Results(10 Points)

Part 1:

We examined the datasets to understand their structure and identified key information such as user activity, experiment groups, pre-experiment behavior, and user attributes. This helped us determine which data was essential for analysis and potential issues like missing values or outliers.

Part 2:

We merged user engagement data with experiment group labels (control vs. treatment) to create a structured dataset. Cleaning the data ensured it was ready for statistical testing and meaningful comparisons.

Part 3:

An initial t-test showed that the treatment group had significantly higher engagement. However, a normality test revealed that the data was not normally distributed, suggesting that a t-test alone might not be fully reliable.
Part 4:

Investigating outliers revealed extreme values, such as users logging impossible activity times (99,999 minutes). After removing outliers, the results remained significant, confirming the update's impact on engagement.

Part 5:

We analyzed engagement changes before and after the update rather than just post-experiment behavior. Even after accounting for pre-existing activity levels, the treatment group still showed a significant increase, reinforcing that the update was effective.

Part 6:

We examined how different user types (contributors, readers, new users, etc.) and genders responded to the update. Contributors and readers showed the highest engagement increase, while new users had minimal change. Males engaged more than females across most user types.