HW3
Zichong Meng
DS5110
Febuarary 6 2025

Part 1

1. In the t1_user_active_mis.csv file, the data of each row is about each user's activate minutes on the website on each specific date after the experiment starts. However, there will be no data if the user is never activated.
   uid is the user ID
   dt is the specific date
   active_mins is the total activate duration in minutes in that specific dates of that user
2. In the t2_user_variant.csv file, the data are about which group the user is in, if it is 0 the user is in the control group, if it is 1 the user is in the treatment group. The data also record the date the experiment starts which is all the same dates, and when the user signed up.
   uid is the user ID
   variant_number is the group users are separated into (0 for controlled and 1 for treatment)
   dt is the date the user starts the experiment which is 2019-02-06 for all users
   signup_date is the date users signed up
3. In the t3_user_active_min_pre.csv file, the data of each row is about each user's activate minutes on the website on each specific date before the experiment starts. If the user is never activated, there will be no data.
   uid is the user ID
   dt is the specific date
   active_mins is the total activate duration in minutes on the specific dates of that user
4. In the t4_user_attributes.csv file, the data are about user's attributes like gender, and what user group each user belongs to based on their activity level
   uid is the user ID
   gender is the user's gender
   user_type is the user group based on their activity level
5. table_schema.txt file contains the description of all the schemas what's in each CSV file and what each variable in each CSV file means.

Part 2

1. The overall object of this study is to study how the new updated design of the website can lead to much increase in the users' active total minutes. In the current setting, this is done with A/B testing where one group is the controlled group with no update and one group is the treatment group where new updates are available.
2. So in this setting, user ID, the total time regardless of the specific date (in all dates) they are active on the website, and what group they are in are very helpful to compare the impact of the new update in A/B testing.

3. The data in t1 only have user ID, the data they access the website, the total time on the specific date they are active on the website. There is no data about what group they are in and no sum of the total time regardless of the specific date.
4. We should merge the user's group information from t2 to t1 to make it helpful as indicated in 2 and also we should calculate the sum so we can get the total time user spend during the entire experiment.
5. The organized file in the data folder and named 'Data/new_t1.csv' and the merging script is part2.py.

Part 3

1. The python file to get p-value and t-test stats are in part3.py and the results are in Data/part3.csv. P-value is 0.6850342487187623 and the t-stats is 0.4056089463388251. Since p-value is realy really large so null hypothesis has larger probability to be true and t-stats is also not very large, therefore there is no statistical difference between group 1 and group 2.
2. The python file to get mean and medium are in part3.py and the results are in Data/part3.csv. For controlled group , it is 837.6428857715431 for mean and 52.0 for medium, for treatment group it is 784.2028670721112 for mean and 71.0 for medium.
3. We can not conclude based on the data because there may be outliers since mean and medium tell us very different things, in mean level group 1 is not as good as group 2, but in medium level it is completely different.

Part 4

1. No, we can not fully trust the result. People's willingness and their own preference can make it an impact on the website activity, so there will be outliers in each group that impact the results, for example some one who stay just longer than all other people no matter what design the website is.
2. The data is not normally distributed this is because p-value from part 3 indicates that it is much much larger than 0.05 which fails to reject null hypothesis.
3. The python file to get boxplot is in part4.py and the results are in figure_part4.png.
4. There are outlier as indicated in the boxplot, in both groups there are some user tends to stay very very long than other people.
5. In t1, we can see most of people in each day theta re active are usually no more than 10 minutes, but some users tends to be active for more than hundred of minutes causing them to be outliers.
6. The python file to remove outlier is in part4.py and the new data is in 'Data/new_new_t1.csv', I use IQR to remove outliers.
7. The python file to redo part2 and part 3 is in part4.py and the new data is in 'Data/part4.csv'.
8. From the new results, we can see that p-value is 3.5034641363021703e-28 which is lower than 0.05 so there is statistically difference between group 1 and group 2. Also because mean for controlled group is 77.7463586922548 and medium is 38.0, and for treatment group it is 91.88035280582896 and medium is 52.0, both mean and medium tells that the treatment group have a longer average total active time on the website, but we still can not conclude the impact of new design as we still need to know the pre active time before experiment to be more accurate.

Part 5

1. We care about the data in t3 because the difference of increase before and after the experiment can help us reduce the impact of personal behavior of the results. For example, if someone actually already like the website before the experiment, and he is in treatment group, he might still have high active time but the increase or decrease can tell whether he really like the new look or not.
2. The python file to redo part2 and part 3 is in part5.py and the new data is in 'Data/part5.csv'.
3. From the new results, we can see that p-value is 2.0408012812569588e-137 which is lower than 0.05 so there is statistically difference between group 1 and group 2. Mean for controlled group is -11.810646490576234 and medium is -5.0 and for treatment group it is 6.285897435897436 and medium is 4.0, both mean and medium tells that the treatment group have a longer increased in average total active time on the website, and I think we can obtained the information that new design tends to increase people active time on website.

Part 6

1. I merged t4 into the new table from part 5, and save the csv file is in 'Data/part6_attr.csv' and 'Data/part6_genders.csv', it seams like new user, non_reader tends to increase active time with new design, and all genders tends to increase active time.

Part 7

1. **Part 1**
    - In part 1, I find out what data in each table is and which helps me to do the further studies.
2. **Part 2**
    - In part 2, I rethink about what is our primary objective for this experiment and what data we need which helps me to merge and process the tables.
3. **Part 3**
    - I analyzed the result we obtained from the processed table, this helped us to see that is wrong in data and why it can not conclude.
4. **Part 4**
    - Removing the outlier and reprocessing again helps to remove the flaws in data and build a better connection, but still can not conclude.
5. **Part 5**
    - Calculating the difference rather than actuall time helps more on seeing the impact of new design on people's active time on website.
6. **Part 6**
    - User attribute and gender analysis helps to break further down on which group people will increase active time and which group of people will not which helps us to lead a more detailed conclusion.