DS 5110

Homework 3

Due 2/7/25

Cody Snow

Part 1:

1. t1_user_active_min.csv contains active minutes data logged after experiment started. Each row represents the total number of minutes spent on site for each user on a date. If a user never visited the site for a given date, there wouldn't be data for that uid on that date.
2. t2_user_variant.csv contains users' treatment assignment. Each row represents the assignment information for a unique user.
3. t3_user_active_min_pre.csv contains active minutes data before the experiment started. It has a similar format as t1, except the dt range can extend before the experiment start date.
4. t4_user_attributes.csv contains data about some user attributes. Each row represents attributes of a unique user.
5. table_schema.txt contains the information stated above about each data file, as well as additional information about the fields found in each of those data files.

Part 2:

1. The objective of this study is to determine if the updates they made to their website have the desired effect—to increase usage time for their user base.
2. To reach this objective, we need a way to examine the impact the change has on user behavior. The data provide a before and after look at two groups of users—the treatment and the control groups.
3. Currently, t1 data consist of a user id, a date, and an "active minutes" value. Each row represents the total number of minutes spent on the site for each user on a date. Importantly, this dataset records time spent *after* the update.
4. To be useful, we need to group the users by treatment and control. For each user in both groups, we want to know the total time they spent before and after the update. Then we can demonstrate the effect the update had on the treatment vs. the control.
5. Data organization is done via Python code in Jupyter notebook/.py file.

Part 3:

1. I ran an independent t-test on the two dataframes I created (one for the control and another for the treatment group). The results are:

   ```
   TtestResult(statistic=-0.32346507126292273, pvalue=0.7463445065262613,
   df=46631.0)
   ```

2. Mean and median for group 1 and group 2:

   Treatment:
   Mean: 784.2028670721112
   Median: 71.0

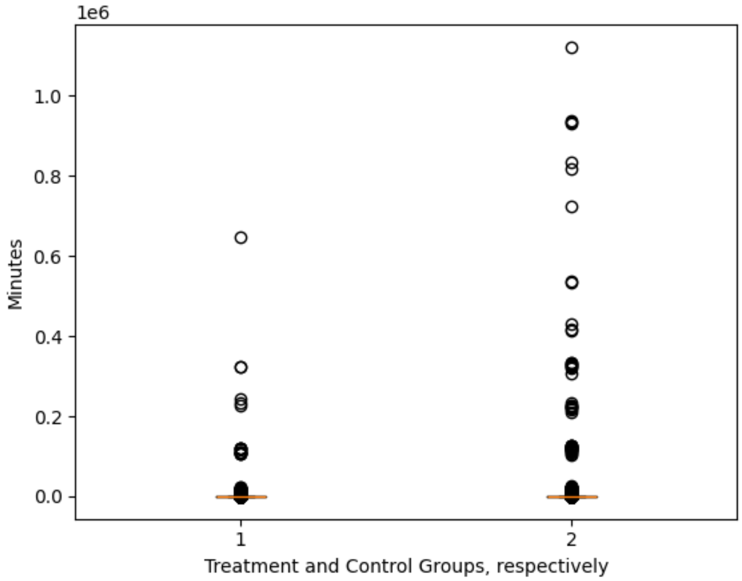
   Control:
   Mean: 837.6428857715431
   Median: 52.0

3. Based on the very large pvalue I got, there is no strong evidence of a difference between the two groups. However, the mean and median in both datasets are very far apart, but in each group they are far apart by a suspiciously similar amount. My gut tells me something is off with the data.

Part 4:

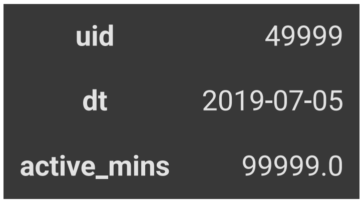
1. I do not trust these results. In analyzing the data, I noticed some very high values for user active minutes, so I suspect outliers are playing a role in skewing the data. Because I haven't done any analysis of this thus far, I certainly can't rule it out. The pvalue is very large—so I naturally ask why is it so large? I would expect at least some evidence of the website update having an impact on user behavior.
2. The data are not normally distributed, as evidenced by the large difference between the mean and median in each set.

3.



4. There are definitely outliers present in the data.

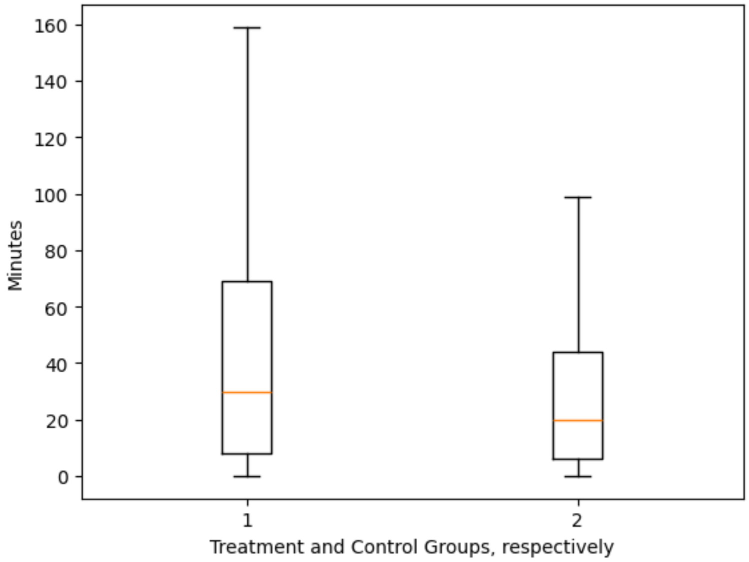
5. df_t1['active_mins'].max() gives us 99,999.0 minutes for , which is almost 70 days, for uid 49999:



uid	49999
dt	2019-07-05
active_mins	99999.0

This is obviously an erroneous data point and must be removed.

6. Outliers are removed:



7. I am assuming the intent is to redo parts 3 and 4 (not 2 and 3). The result of the t-test is dramatically different:

```
TtestResult(statistic=41.75965577212747, pvalue=0.0, df=30265.0)
```

```
Treatment mean: 43.5044994375703
Treatment median: 30.0
Control mean: 28.201643918256543
Control median: 20.0
```

8. With a pvalue of 0, there can be little doubt there is a statistical difference between the two groups. Our data is still not perfectly normal, as the means and medians for each set are not the same. However, they are much closer than they previously were. The Shapiro-wilk test tells us we are somewhere in the 87-88% fitment into a normal range for each set:

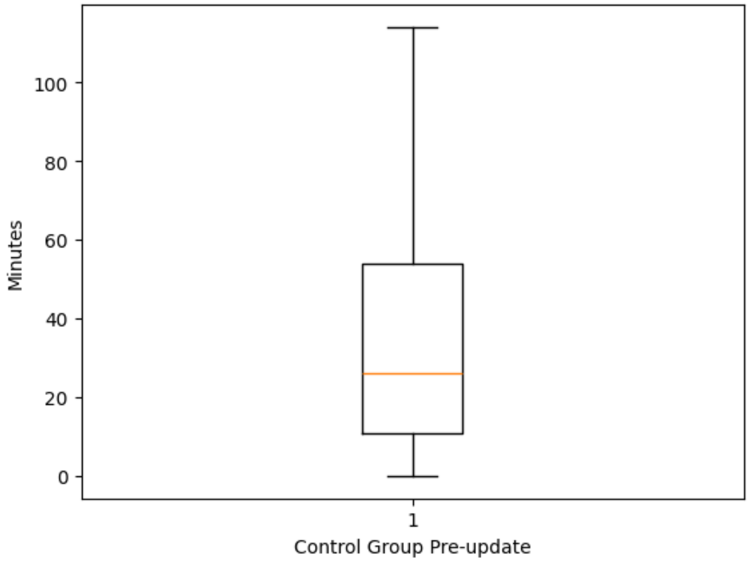
```
ShapiroResult(statistic=0.8758864369089674, pvalue=1.0426198327594599e-59)
ShapiroResult(statistic=0.8847444331415818, pvalue=3.3708985625458386e-86)
```

   I am not sure what to make of the pvalues here, though. Overall, my conclusion is that we have brought the data much closer to normality and have a statistically significant difference between the control and the treatment groups.

Part 5:

1. This dataset allows us to observe any changes in user behavior in the control group. We can look for statistically significant changes between the control's before and after usage numbers, which might suggest something else impacted the difference between the control and treatment groups.

2.



With outliers removed and controls separated from treatment in t3, we have the boxplot above and mean/median below.

Control group before update mean: 35.1200829719027
Control group before update median: 26.0

As we recall, after the update, it's:
Control mean: 28.201643918256543
Control median: 20.0

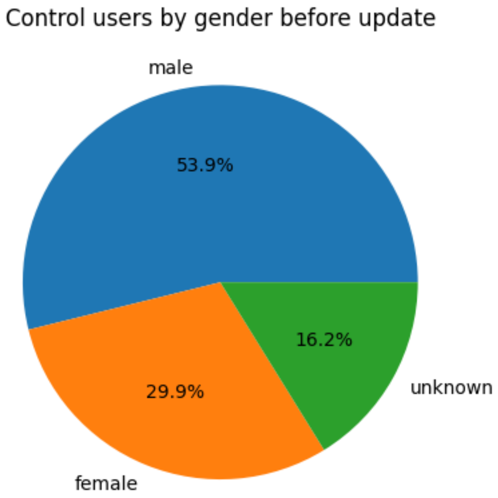
Control pre-update ShapiroResult(statistic=0.8951517367103534, pvalue=5.456329239401593e-84)

Another t-test shows us the following:
TtestResult(statistic=16.702107010397665, pvalue=1.8660238875990016e-62, df=50236.0)

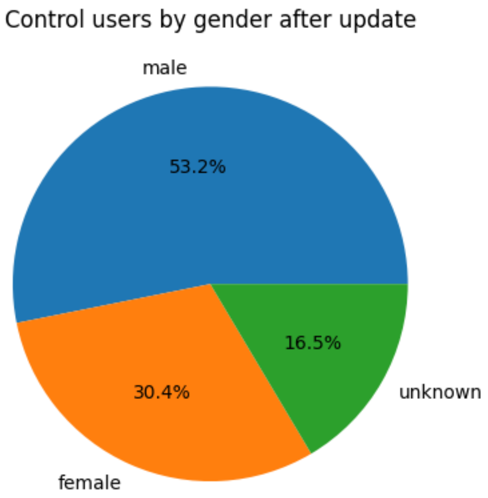
3. According to the results above, there is a statistical difference between the control groups before and after the update. I am not sure what could be causing this, but it suggests that something changed in our control group after the update was applied. Perhaps a controversy on the platform put some users off? Was there a policy change that occurred that might account for the shift? Regardless of the cause, this casts some doubt on the conclusions we have drawn about the treatment vs. control groups, as it calls into question whether our control is truly a control.

Part 6:

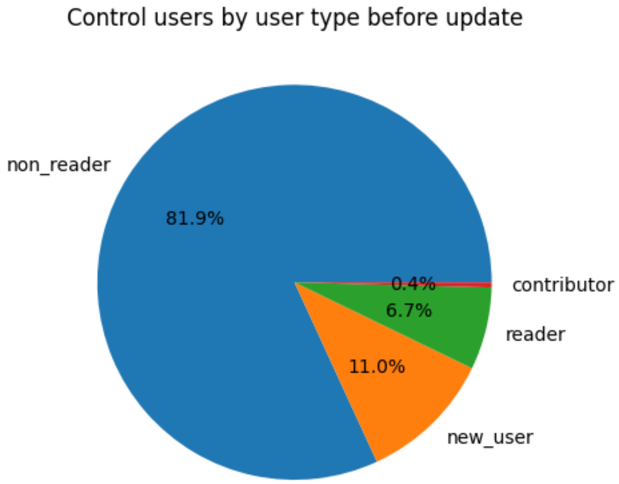
I created four new dataframes by combining the user data from t4 with the pre- and post-update datasets. They are pre- and post-update treatment, and pre- and post-update control groups. After grouping, I removed the outliers using their respective interquartile ranges. I retained the user's gender and user type fields in each set so I could plot them and examine the data a bit more.

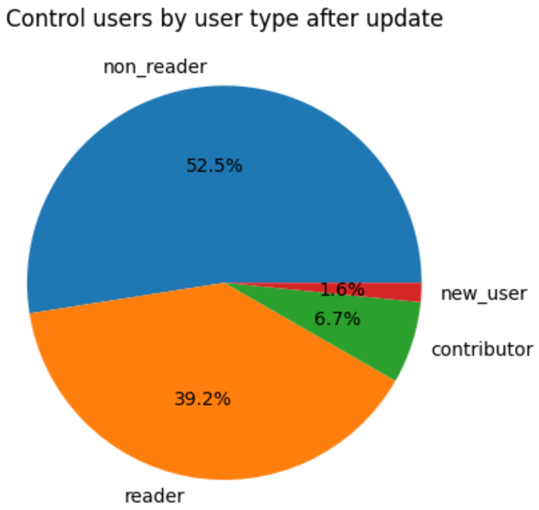


The user base in the control group skews heavily male



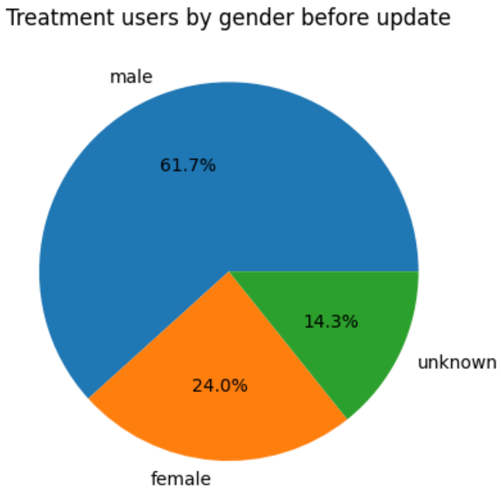
Similar for user type:



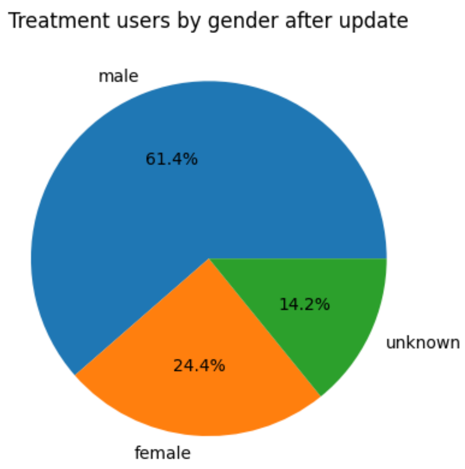


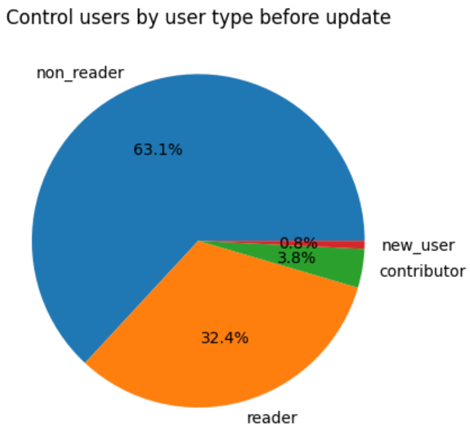
There is a remarkable increase in the number of readers after the update in the control group. Why would the control group change their user type so dramatically?

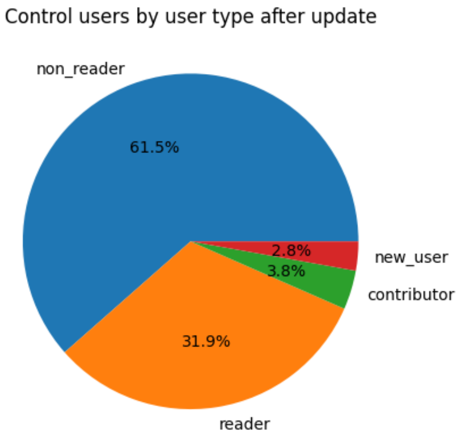
Now for the treatment group:



Here, the skew is even more heavily male.







Here we do *not* see the same large shift in user type. Could this account for the difference we saw in the before/after groups of control users?

Part 7:

Part 1 is simply a data overview, describing what each dataset contains.

Part 2 describes the objective: to determine if website updates increased user activity, comparing the treatment and control groups before and after the update; and data organization: splitting users into treatment and control groups based on the data files, and aggregating their active minutes before and after the update.

Part 3 is the initial analysis, where I use independent T-Tests to compare the treatment and control groups post-update. The p-value (0.746) suggests no significant difference, though there are large differences in means and medians, which raises suspicion.

Part 4 reveals the outlier impact, where the outliers' high values (e.g., 99,999 minutes) are identified, skewing results. After removing outliers, the new t-test results show a significant difference between the groups (p-value = 0.0). A Shapiro-Wilk Test shows the data is closer to normal but not perfectly normal.

Part 5 is a control group analysis, performing a before-and-after analysis of the control group, which suggests a significant change, but this raises concerns about whether the control group is truly unaffected by external factors.

Part 6 is about combining data and investigating user attributes. Merging user data with pre- and post-update datasets and removing outliers, I was able to analyze gender and user type. In the control group, I observed that the user base skews male, and there is a notable increase in readers after the update. In the treatment group, I saw a similar male-skewed user base, but no dramatic shift in user type. This calls into question the selection of the users for the study and raises concerns about selection bias and the adequacy of population size within the datasets.