Brian West HW3

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 txt file. Answer the following questions:

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

A unique identifier for users, a date, and the minutes active during the date (session?)

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

A unique identifier for users, which variant the user received, the date the user entered the experiment, and the date the user signed up.

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

A unique identifier for users, a date, the number of minutes the user spent on the site on that date.

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

A unique identifier for users, the type of user the user is, and the gender of the user.

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

An explanation for the data in the tables t1,t2,t3, and t4.

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?

The objective is to understand how the different variants influence user time on the site.

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

We need to aggregate the user's time into total time spent on the site and which variant the user was given.

3. How is the data in t1 currently organized?

The data in T1 is currently organized by the total time the user spent on the site on a given day.

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

It should include the variant the user was in as well and the total time the user spent on the site during the trial.

5. Organize it.

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?

Tvalue: 0.32346507126292273

Pvalue: 0.7463445065262613

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

Variant 0 Mean: 837.6428857715431

Variant 0 Median: 52.0

Variant 1 Mean: 784.2028670721112

Variant 1 Median: 71.0

3. What can you conclude based on that data?

Just looking at the T-value I would assume that the data sets are different. The difference between mean and median between both groups suggest a large skew.

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:

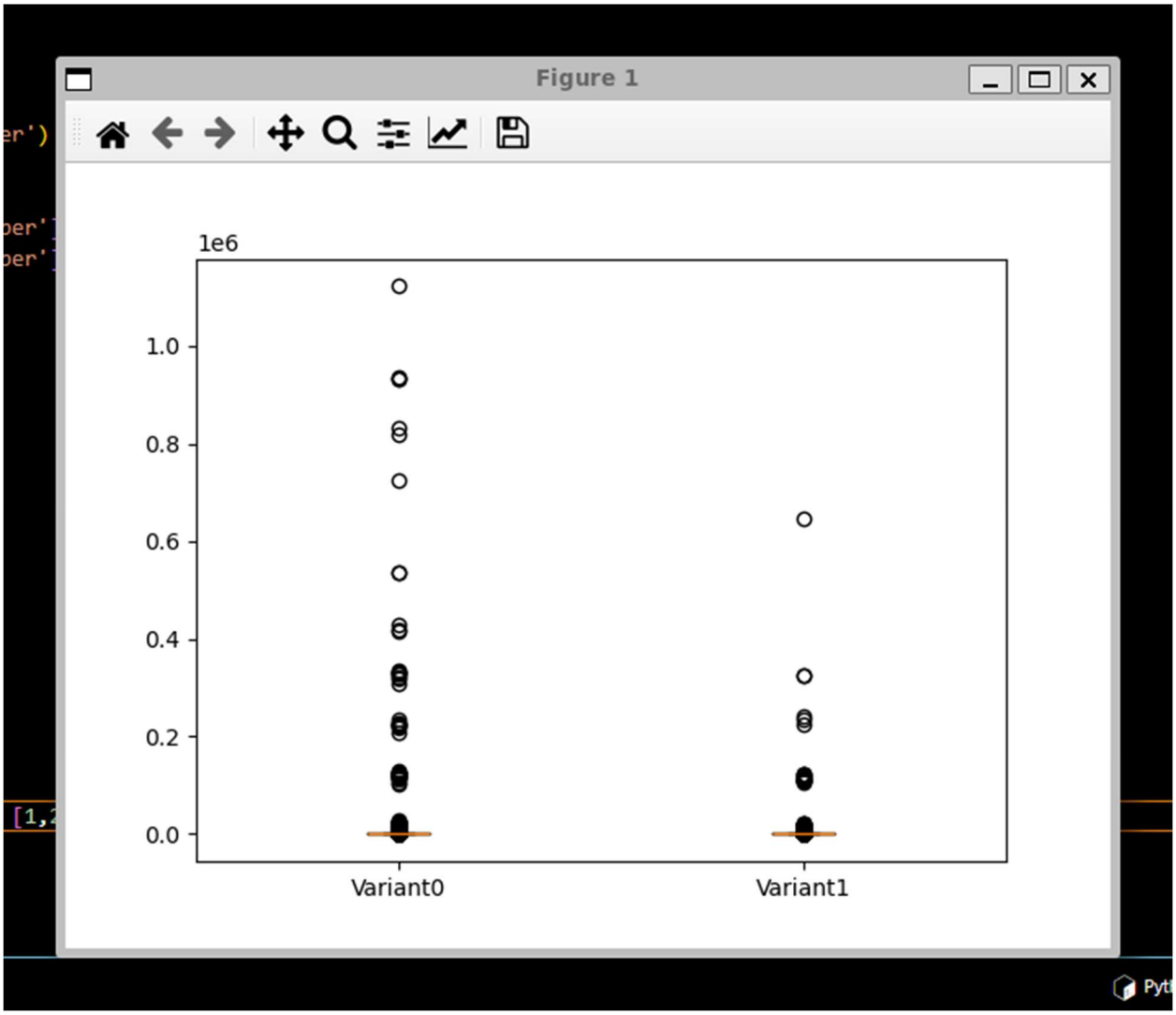
1. Can you trust that the results? Why or why not?

No, I have not interrogated the integrity of the data. There could be outliers or bad data that are skewing the results of the previous evaluation.

2. Is the data normally distributed?

No, the mean and median of both groups have significant differences between them.

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



4. Are there any outliers?

Yes, there are a number of outliers in both groups.

5. What might be causing those outliers? (Hint, look at the data in t1.

What is the maximum time a user should possibly have?).

There are 172 rows with more minutes logged than minutes in a day for a given date.

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

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

Tvalue: -0.009396017075725334

Pvalue: 0.9925032135466763

Variant 0 Mean: 458.2211623246493

Variant 0 Median: 52.0

Variant 1 Mean: 458.4024761077324

Variant 1 Median: 71.0

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

The difference between the means and medians still suggests a heavy skew. The T value suggests that the two groups are similar. The p value is extremely high, which means we cannot make a conclusion about the two groups from the t value.

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?

The data from t3 describes user behavior before the experiment started. We can use this to the changes in user behavior during the experiment.

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

Tvalue: -18.574902340661215

Pvalue: 9.737307462134446e-77

Variant 0 Mean: -47.29544662718087

Variant 0 Median: -6.0

Variant 1 Mean: 164.65411893071467

Variant 1 Median: 12.0

3. Are there any new conclusions?

The p value and t value imply that there is a difference between the two groups data set. The means and medians still having large gaps still indicate that our data is probably not normal.

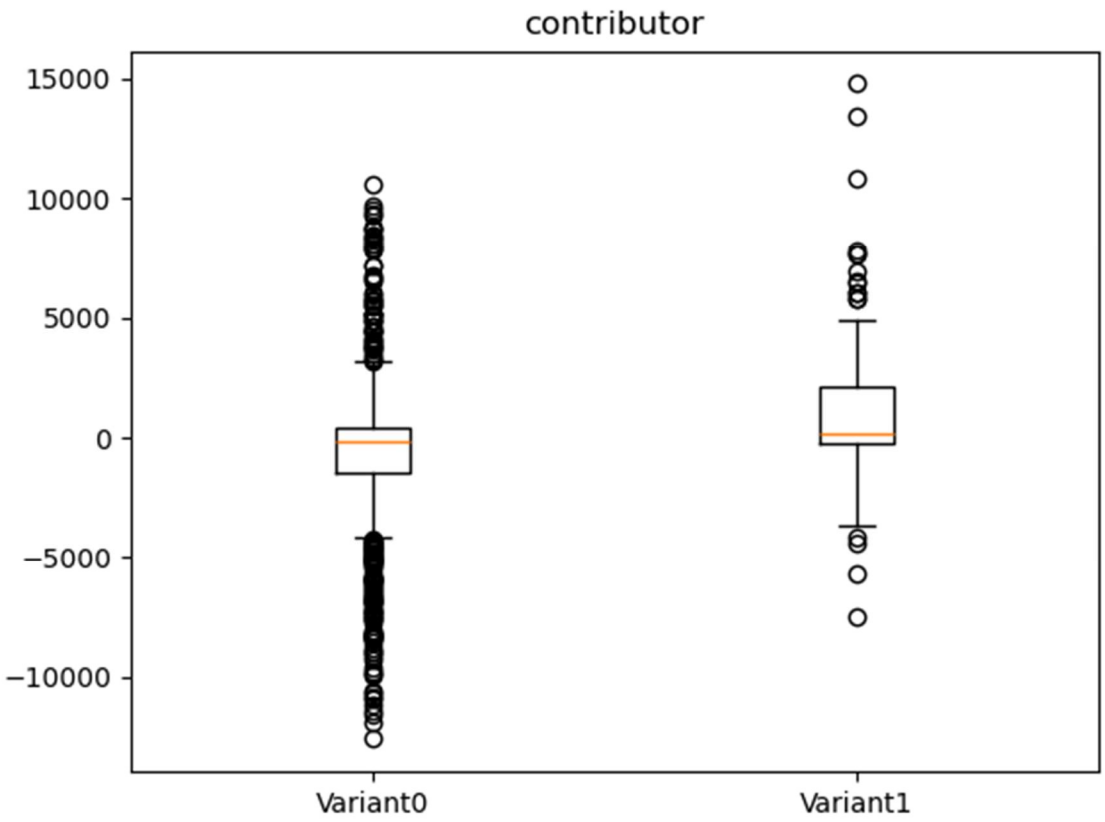
Part 6: Exploring other conclusions (10 Points)

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.

The first thing I noticed with T4 was the different categories of users based on "activity level". I would love to see what determines these activity levels. I also removed any outliers that were over 3 standard deviations away using Z scores. Below are box plots and the associated statistics.

I did try further filtering with gender, but this really didn't provide any additional clarity to the results of the experiment. Additionally, I also tried taking into consideration user's sign-up dates, since this could skew my aggregate metric of "active_mins_t3", but that also did not provide any additional clarity.

The Tvalues and Pvalues imply that there is a difference between the two groups for each user category. The descriptive statistics also imply that users in variant 1 spend more time on the site than those in variant 0. Viewing histograms (not shown) for the data also shows that the data is fairly normally distributed for the user categories.



Statistics for contributor

Tvalue: -5.927299007088535

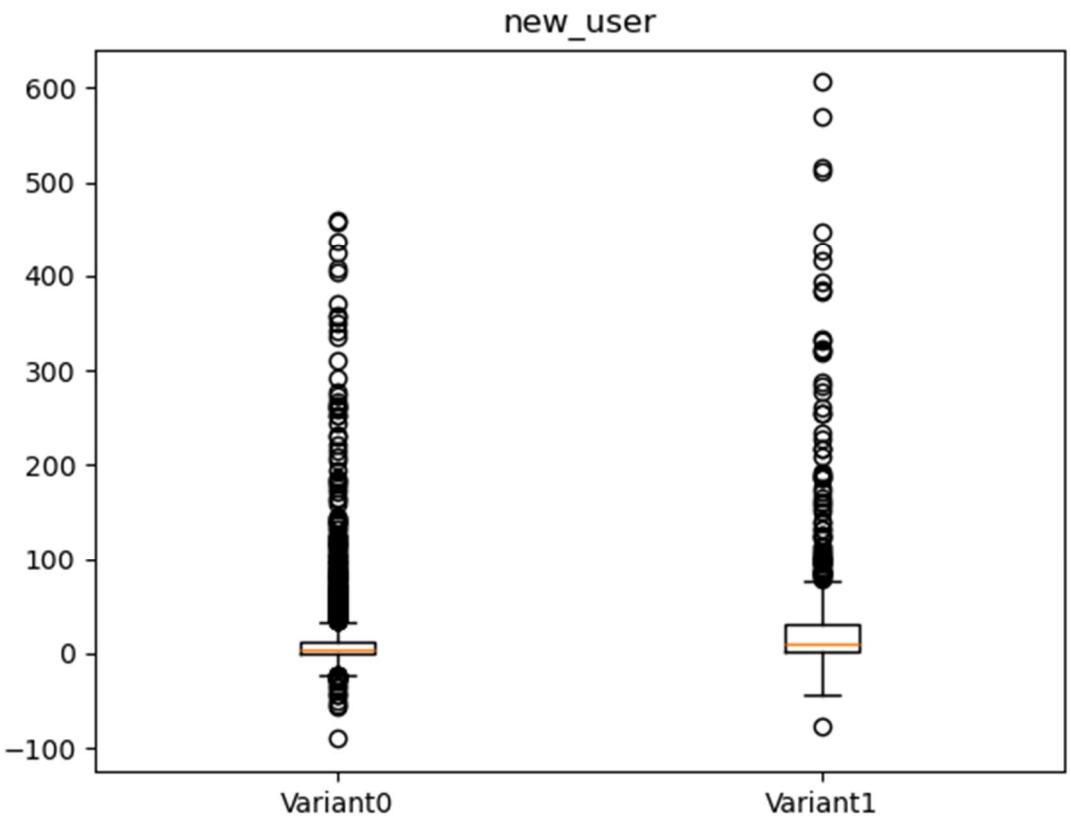
Pvalue: 4.2271761664181e-09

Variant 0 Mean: -600.3115124153499

Variant 0 Median: -176.5

Variant 1 Mean: 1164.655737704918

Variant 1 Median: 166.5



Statistics for new_user

Tvalue: -7.927475527313491

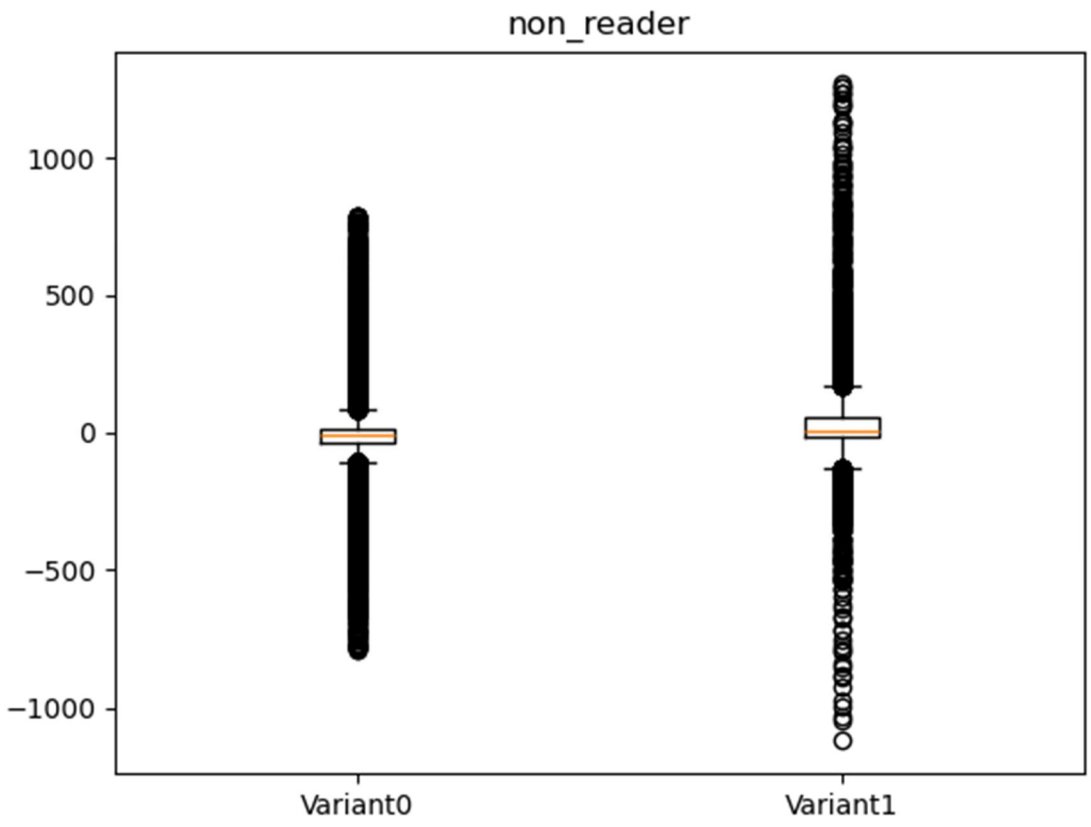
Pvalue: 3.1162150674538348e-15

Variant 0 Mean: 14.800793301013663

Variant 0 Median: 4.0

Variant 1 Mean: 31.868766404199476

Variant 1 Median: 10.0



Statistics for non_reader

Tvalue: -28.124678801081636

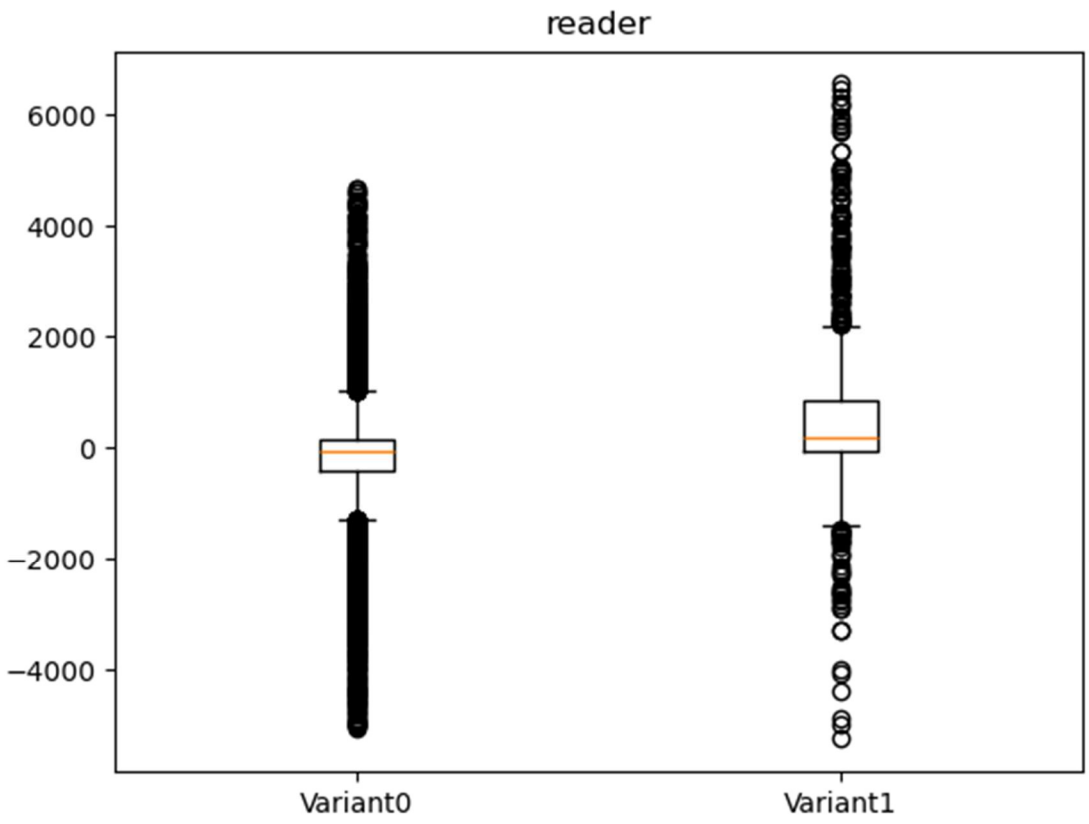
Pvalue: 4.578887734231996e-172

Variant 0 Mean: -8.4955530132487

Variant 0 Median: -6.0

Variant 1 Mean: 38.6529963898917

Variant 1 Median: 9.0



Statistics for reader

Tvalue: -18.85151619080644

Pvalue: 1.5159749764306695e-77

Variant 0 Mean: -172.1532332209623

Variant 0 Median: -65.0

Variant 1 Mean: 501.0615883306321

Variant 1 Median: 191.0