Shuiming Chen
01/27/2025

Part 1: Getting to know your data

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

Answer: file 1 contains user id, user login date, and the specific minutes of the user spent on that day. All data collected after the experiment begins.

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

Answer: file2 has user id, variant number, 1 for update version, 0 for control type which is the old version, and date which is the day the user was included in the experiment, all users were added on 2019-02-06, and also contains the user's initial sign up data.

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

Answer: file 3 contains user id, date and minutes spent on that day, all the data here came before the experiment started.

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

Answer: file 4 has users' attributes like user id, user type and gender.

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

Answer: this file has the brief introduction among table 1 to table 4, which can help us easily to know the basic information of those data.

Part 2: Organizing the data

1. What is the overall objective of this study?

Answer: The objective of this study is to determine whether the new layout and features of the social media platform can increase the users playing time or not. If it is added, the company will spend money on advertisements and launch it.

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

Answer:

- User ID
- Users daily active playing time, used to calculate the total playing time after the experiment started
- And users locating groups.

The first two data we can reach from table 1, t1_user_active_min.csv file. And the users locate groups can get from table 2, t2_user_variant.csv file.

3. How is the data in t1 currently organized?

Answer: t1 table contains user id, logged date and the minutes spent on that date. However, t1 neither handles the total time each person plays after the experiment, nor groups them into a control group or a treatment group. So the data in t1 only gives us the basic information, and needs to be reorganized.

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

Answer: combining t1 and t2 then reorganized a new file and let it contains:

- user id,
- user group(0 or 1), either control group or treatment group,
- Each user's total playing time after the experiment.

5. Organize it.

Answer:

- Create a python file name part2.py to handle table 1 and table 2
- Aggregate the table 1 by sum up playing time of each user
- Merge with table 2's group information, then generate a new csv file called part2.csv.
- All these new files will be stored in part 2 folder.

The new csv file data will look like:

part2

uid	total_act_mins	variant_number
0	43.0	0
1	15205.0	0
2	17.0	0
3	77.0	0
4	39.0	0
5	174.0	0
6	26.0	0
7	21.0	0
9	42.0	0
10	127.0	0
11	142.0	0
13	17.0	0

Part 3: Statistical Analysis

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

Answer:

To determine if there is a statistical difference between control group and treatment group, we can apply the t-test method. This method compares the means of the two different groups to determine whether the statistical difference is significant or not.

The code part can see the t_test.py file in part3 folder.

Print the results:

```
 TtestResult(statistic=np.float64(0.32346507126292273), pvalue=np.float64(0.7463445065262613), df=np.float64(46631.0))
========================================

              T           dof alternative     p-val              CI95%   cohen-d   BF10    power
T-test  0.405609  20207.387288   two-sided  0.685034  [-204.81, 311.69]  0.003763  0.014  0.06207
========================================

(np.float64(0.32346507126292273), np.float64(0.7463445065262613), np.float64(46631.0))
```

Conclusion:

- According to the result, after testing 3 methods of two sample t-test, the pvalue ranges from 0.685 to 0.746.
- Pvalue > 0.05, which means there is no statistically significant difference between these two groups.

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

Answer: see part 3 folder, mean_median_value.py file

- Group 1: control group:
    - Mean: 837.64
    - Median: 52.00

- Group 2: treatment group:
    - Mean: 784.20
    - Median: 71.00

3. What can you conclude based on that data?

Answer:

- Mean comparison
    - Group 1 and group 2 have a very close mean number of the users total playing time with the old version and updated version of platform, which means the update does not have had a significant effect on users.
- Median comparison
    - Control group has a smaller median total playing time compared with treatment, but considering there are 40, 000 users data in the control group and 10, 000 users data in the treatment group, the median data difference doesn't mean that the new version of the platform significantly adds users' playing time.

Conclusion, based on the previous data results, we cannot tell that the new version of platform has a significantly higher total playing time than the old version.

Part 4: Digging a Little Deeper

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

Answer:

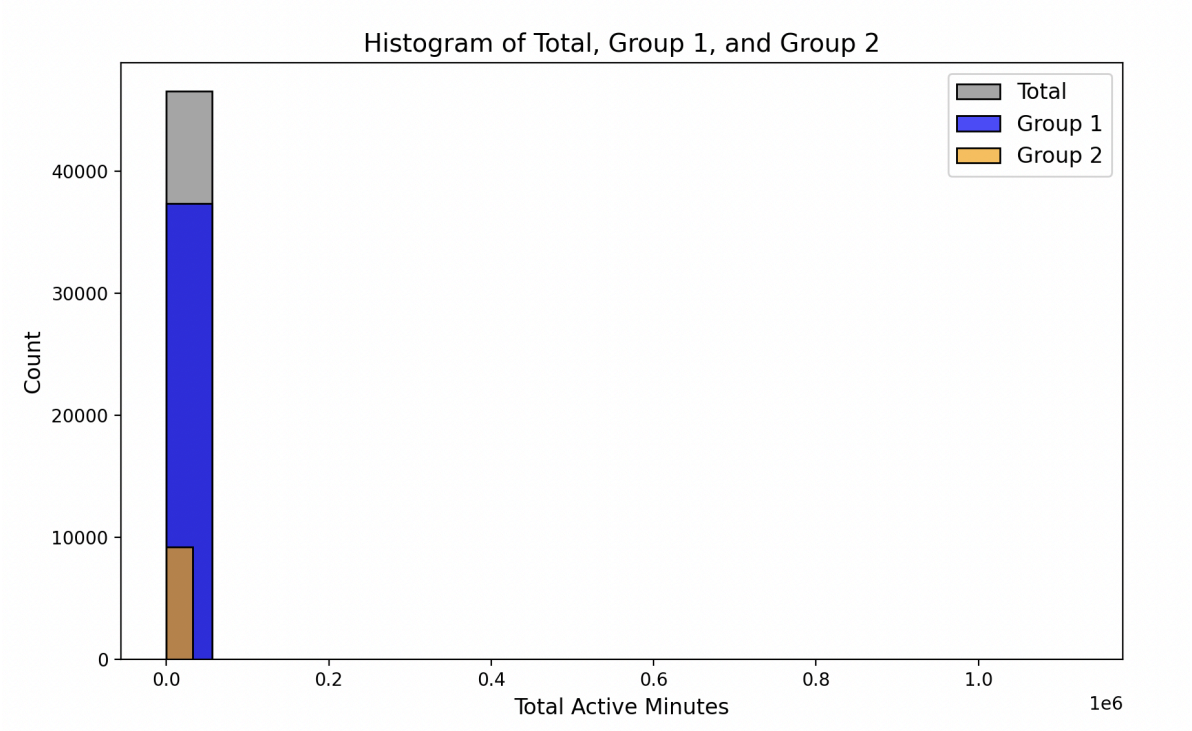
The result can not be trustful.

- The dataset's independence varies, and not handle incorrect value
- The dataset has extremely high values/outliers
- The dataset is not normally distributed, see question 2 explanation

2. Is the data normally distributed?

Answer:

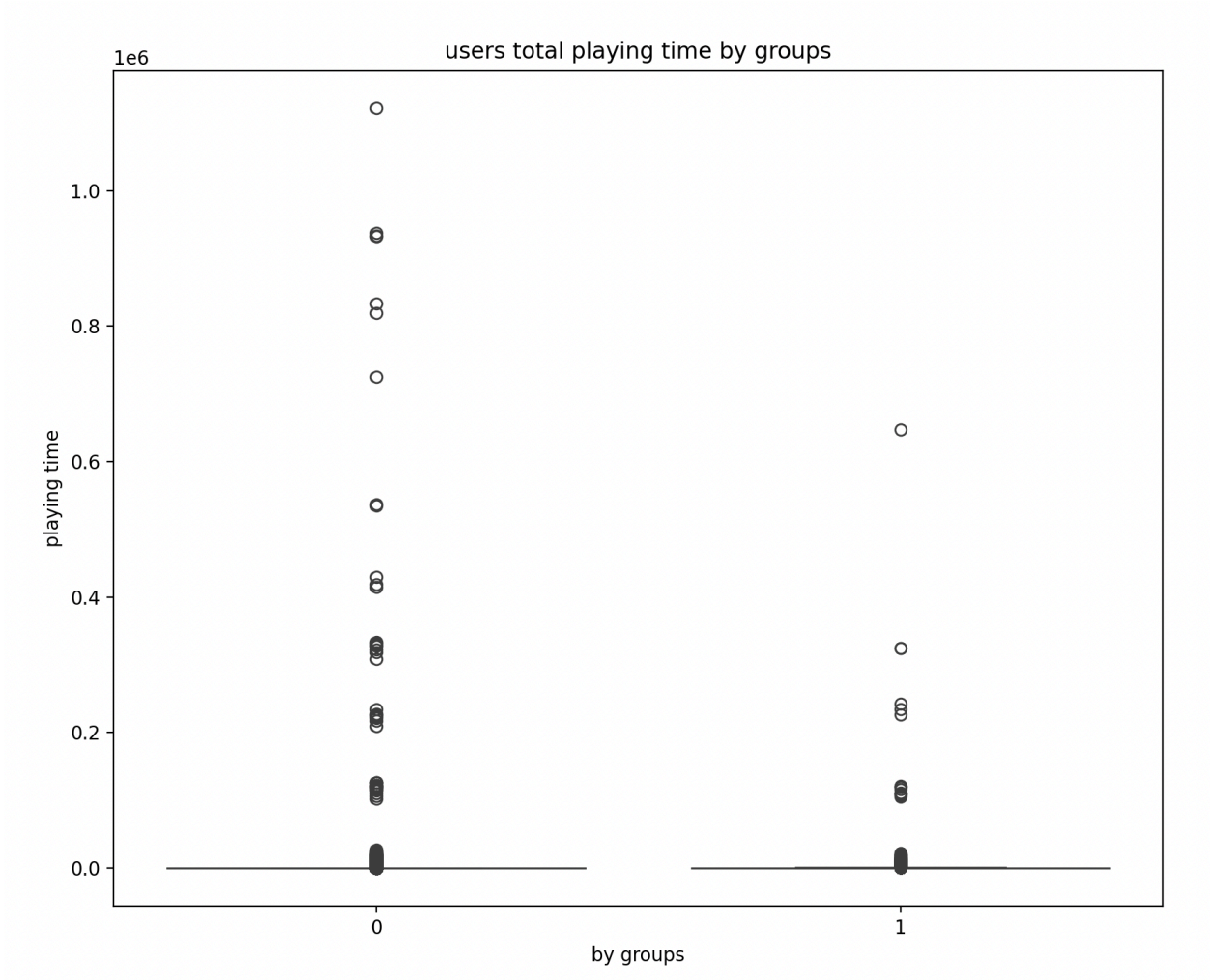
The data is not normally distributed, we create a histogram graph, it doesn't have a symmetric "bell curve" shape, which means most of the data are clustered in the low number area, and the dataset has extremely high value. See the figure below. And histogram.py file in part4 folder.



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

Answer:

See boxplot.py file in part4 folder



4. Are there any outliers?

Answer:

Yes, both group1 and group2 have outliers, see the figure above.

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

Answer:

- After checking table 1, we find out that the highest value in table 1 is “99999” minutes in one day, which is obviously incorrect.
- and the total number of rows in table 1 with incorrect value is 172. See the extream_value.py file in the part4 folder.

All these incorrect values recorded in table 1 causing the outliers.

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

Answer:

Remove the rows in t1_user_active_min.csv file when active_mins is more than 12*60.The reason why I choose 12 hours instead of 24 hours is because in most cases most of the users won’t spend more than 12 hours a day. This is not a gaming platform, it’s a social media platform.
See part4_redo_part2_3 folder.

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

Answer: See part4_redo_part2_3 folder.

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

Answer:

(1) Mean and median

Group 1: control group:

- Mean: 453.94
- Median: 52.00

Group 2: treatment group:

- Mean: 450.68
- Median: 71.00

- Mean comparison
    - Group 1 and group 2 have a very close mean number of total playing time with old version and updated version of platform, which means the update does not have had a significant effect.
- Median comparison
    - Control group has a smaller median total playing time compared with treatment, but considering there are 40, 000 users data in the control group and 10, 000 users data in the treatment group, the median data difference doesn't mean that the new version of the platform has had a significant effect.

(2) t-test

After running a T-test file with 3 methods, the pvalue is 0.86 among all of them.

This means:

- Pvalue > 0.05, there is no statistically significant difference between these two groups.

Conclusion, based on the previous data results, we cannot tell that the new version of platform has a significantly higher total playing time than the old version.

Part 5: Digging even Deeper

1. Why do we care about the data from t3?

Answer:

T3 table provides the user id, user logged date and the minutes spent on that day, all the data came before the experiment started.

- Comparison with previous data: with table 3 data, we can build the user's previous behavior, and can help to observe changes compared with the data from table 1.
- Help us better to detect if there exists any significant playing time changes after the experiment, since we have the data of table3, users can be divided into two groups, control group and treatment group, each group has its before and after experiment data. Unlike before, each data only has after experiment data.
- Help us to know users' changing: like before the experiment started, user's would spend more time or less than nowadays.
- Differentiate of new users and old users active time on the platform.

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

Answer:

- Handle table 3 and combine with table 1, see process_t3.py file
- Calculated table 3 mean = 432.80, median = 52.00 of total active time
- reRun t-test with new combined csv file, details see t-test.py in part5 folder, divided data into control group and treatment group, in each group:
    - Control group:
        - total active time after experiment
        - Total active time before experiment
    - Treatment group
        - total active time after experiment
        - Total active time before experiment

- After running the t-test, all the results of pvalue < 0.05.

3. Are there any new conclusions?

Answer:

Yes.

After organizing data from table 3, removing extreme value and combining with already handled data from table 1, we have a new csv file, see part5 folder, name combine_t1_t3.csv.

And then after running t-test, see t_test.py file in part5 folder, the pvalue is much less than 0.05, which is typically considered to be statistically significant.

Conclusion:

After all these data tests, we can conclude that:

The new layout and features of the social media platform can increase the users total playing time. the company can spend money on advertisements and launch it.

Part 6: Exploring other conclusions

Answer:

(1) There are different user types, we can count how many of them in total, see analysis.py file in part6 folder.

	user_type	count
0	non_reader	36066
1	reader	8002
2	new_user	4888
3	contributor	1044

(2) Combining table 4 with previous table 1 and table 3, we can conclude which type of user play the social media platform most the time(mean value), also see analysis.py file in part6 folder

user_type	total_act_mins	total_act_mins_pre
contributor	4212.785164	4594.983622
new_user	34.850047	6.201769
non_reader	116.594128	106.011033
reader	1590.443172	1632.224670

Conclusion:

Contributor users generally contribute most of the time on playing social media, then the reader type of user, the new users have the least playing time.

Part 7: Summarize Your Results

Answer:

Part1:

explored the datasets from table 1 to table 4, and understood the information those tables contained, including user id, active time, experimental groups, and demographic details.

Part2:

merged and organized the data by dividing into 2 groups, aggregating the users' active time and structuring the data for analysis, also keeping the new organized data into a corresponding folder.

Part3:

conducted initial statistical tests to examine differences in total active time levels before and after the platform version updated. The T-test concluded some statistical analysis, but the conclusion doesn't show a good result.

Part4:

further investigated potential factors impacting previous t-test results, after applying histogram, boxplot and analyse these, there are some extreme values didn't removed from table 1. Even after modifying the previous table data, the result is still not the right one.

Part5:

combining with previous data of table 3 before experiment started, I have old data and current datasets of the same group of people, I finally can track user behavior over time and refine the analysis

Part6:

explored additional insights from table 4 with users' types, finding that non-readers were the most user type and contributors are the main users who spend most of the time active daily on the platform.