DS5110
Homework 2
Due 2/7/25

Hypothesis Testing / AB Testing

This homework is designed to simulate what type of technical questions you might encounter while interviewing for a data science internship. This is done by having you answer questions closely related to a real interview that a former student had while searching for a data science internship.

A company runs a social media platform and wants to increase the total time users spend on their website. The company decides to create a new layout of their platform with new features. Before they launch, they want to test the new features on a small group of subjects to see if the new version of their platform increases total time users spent on the website so that they can cash in on that ad money.

They split their users into two groups. Group 1 is the control group. They did not receive the new update. Group 2 is the treatment group. They received the new update. Data was collected to see how long they spent on their website both before and after the update was released. Your job is to see if the update increased the total time spent by users on the platform.

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"?

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

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

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

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

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?
2. What data do we need to reach that objective?
3. How is the data in t1 currently organized?
4. How should the data in t1 be organized to be useful?
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?
2. What is the mean and median for group 1 and group 2?
3. What can you conclude based on that data?

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?
2. Is the data normally distributed?

3. Plot a box plot of group 1 and group 2.
4. Are there any outliers?
5. What might be causing those outliers? (Hint, look at the data in t1. What is the maximum time a user should possibly have?).
6. Remove any data point that might be causing outliers.
7. Redo part 2 and 3 with the new data without those data points.
8. What is the new conclusion based on the new data?

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?
2. Accounting for the data from t3 rerun part 2 and 3.
3. Are their any new conclusion?

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.

Part 7: Summarize Your Results(10 Points)

Write a summary for each part of this assignment and how it impacted your results.

How to turn in

Turn in your a .pdf copy of you answers to each question along with any graphs that you create to help visualize your results into Canvas. Also turn in any python/data file that you created for this assignment.

Good Luck Cat