Write-up Question 0:

- Collaborated with Marie-Liesse Gouilliard only on Task 2 to understand the iterative process at a high level

Write-up Question 1 (4 points): In the write-up, present (and defend) your characterization of the missing data using the terminology (e.g., MCAR) introduced in lecture.

Comparing the NaN values between the data.csv and data-full.csv files, I broke down the columns and values in the columns for author_type, title, and score.

For the scores, the most correlation I saw between score NaN values and another column was with scores and authors, with scores showing the most relation with NaN values. This means that the scores that are missing could be considered to be MNAR as there are certain values (in this case it seems like the scores that are missing are 6.6 with 343 of these scores missing).

For author_type, the missing data looks like it could be correlated with scores, particular author names, or publication dates (pub_dates). These correlations seem to not have one particular column type/data type that relates back to the NaN values, which indicates that the missing data might be MCAR, meaning that the values that are missing from the dataset are randomized and not totally correlated to the values in the data set.

Lastly, for title, there is a strong correlation between the NaN values for the title and the authors of the music, with the Rob Mitchum entries under author correlating to 38 of the 58 NaN entries relating to the titles. This would indicate that the missing values are in fact correlated to the dataset but are not correlated to the dataset in the particular way we might initially think (certain title names being NaN in data.csv) which leads me to believe the missing data is MAR.

(4 points) Now pretend that you've never seen the full dataset. Characterize the data missing from the **title**, **score**, and **author_type** columns using the terminology introduced in lecture based **only** on the information contained in data.csv.

Using data only from data.csv, the missing score values seem to have a slight correlation with the author value in the data.csv file. While there is not many, I would still think this could be placed under the classification of either MCAR or MAR data deletion types.

For the NaN values for the title, just with data.csv information, we can still see the MAR correlation as we can see how the NaN titles still correlate to Rob Mitchum as the clearly largest value correlated to the missing NaN values (refer to the first part of Write-Up Question 1).

Finally, for the NaN values for the author_type, we can still observe strong correlations between the scores, authors, and some publication dates (pub_dates). In general, however, this would mean that the dataset would be classified as MCAR as the values missing from the dataset are seemingly random and not correlated to any only column type of the values in the data frame.

Write-up Question 2 (4 points): In the write-up, again present (and defend) your characterization of the missing data using the terminology introduced in lecture.

Overall, considering the introduction of the data from data-full.csv and lack of data knowledge from data-full.csv, it seems that the data deletions are a wide array of types.

For the scores, it is most likely that the deletion is MNAR which would not be known without the use of data-full.csv since the data that was deleted was, at a surface level, deleted according to a certain score (or data of that same, given type).

Author_type seems to have been deleted according to MCAR since there was no strong correlation between the deleted author types and any column present in the data. The deletion could be considered MCAR since it appears to be random deletions that aren't presently apparent.

Finally, the title deletions look to be MAR as the deletion of data is correlated to the data set in a way that is not confined in the title column but another observable and recorded set of data relating to another column in the dataframe.

Task 2: Synthesis: Multiple Testing (28 points)

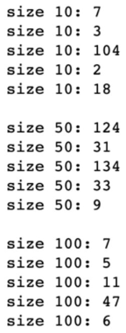
(3 points) First, create two normal distributions both **with a mean of 8 and standard deviation of 2**. Randomly sample *n=30* points from each, and run a t-test to compare these two samples. Repeat this process automatically until you get a false positive **(t-test result indicating that there is a significant difference between the two samples even though you know they came from the same distribution**). Repeating this whole process until you are comfortable, about how many times on average do you need to run your simulation before getting a false positive?

Write-up Question 3 (2 points): In your write-up, report this *average number of trials* you determined.

I ran the test 5 time, receiving the number of iterations to be 22, 45, 5, 50 and 56. This means that the average number of trials to be run to receive a false positive is 35.6 trials to receive a false positive.

(3 points) Now try varying *n* and graph the average number of trials required to get a false positive for different values of *n*.

When the size is 10, the average iterations required to receive a false positive is 26.8. For a random sample size of 50, the average number of iterations is 66.2. Finally, for a random sample size of 100, the average number of iterations to receive a false positive is 15.2 (please refer to Figure 1).



```
size 10: 7
size 10: 3
size 10: 104
size 10: 2
size 10: 18

size 50: 124
size 50: 31
size 50: 134
size 50: 33
size 50: 9

size 100: 7
size 100: 5
size 100: 11
size 100: 47
size 100: 6
```

Figure 1

Write-up Question 4 (2 points): In your write-up, include your graph and describe what you learned.

I learned about the use of p-values for false positives and the appropriate values to look for when trying to assess data for false positives.

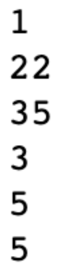
Write-up Question 5 (2 points): In your write-up, describe in prose how each of these three methods seems to impact type I errors (false positives).

It seems that Bonferroni takes more iterations to achieve a Type I error, followed by Benjamini/Hochber, and Holm taking the fewest iterations to achieve a False Positive.

(3 points) Now, create two normal distributions with means 7 and 9, respectively, and standard deviation 1.5 (for both). Note that, per the assumptions of the t-test, both distributions should have the same standard deviation, though the means should differ. Randomly sample *n=30* points from each (again choosing n), and run a t-test to compare these two samples. Repeat this process automatically until you get a false negative (t-test result indicating that there is not a significant difference between the two samples even though you know they came from different distributions). Repeating this process until you are comfortable, about how many times on average do you need to run your simulation before getting a false negative?

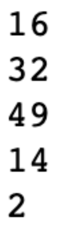
Write-up Question 6 (2 points): In your write-up, report this *average number of trials*.

After running a standard deviation of 1.5 with a mean of 7 on a random sample of 30, I retrieved the values seen in Figure 2. The average of the values indicating the number of times the trails needed to be run to achieve a false negative is 14.2. In Figure 3, we see the same trial run but, instead of the mean being 7 it is set to 9, where the average number of times the trial was run to achieve a false negative is 22.6.



```
1
22
35
3
5
5
```

Figure 2



```
16
32
49
14
2
```

Figure 3

(3 points) Now try varying *n*, and graph the average number of trials required to get a false negative for different values of *n*.

When varying the sized of the random samples, when the mean is 7, I vary the size using the values 10, 50, and 100. For a size of 10, the average number of trials to achieve a false negative is 25.2, for a size of 50 the average number of trials is 14.8, and for a size of 100, the average number of trials is 16.2.

When the mean is 9, the size is varied in the same way as previously stated with a mean of 7 (10, 50, 100). When the random sample size is 10, the average number of trials to retrieve a false positive is 13.0, when the size is 50, the average number of trials is 11.8, and when the size is 100, the average number of trials is 14.2.

Write-up Question 7 (2 points): In your write-up, include your graph and describe what you learned.

I learned about the various types of methods and how to analyze their trends in the data based on the output of the code written.

(2 points) Now repeat the previous simulation three times, using three different methods for correcting for multiple testing: Bonferroni, Holm, and Benjamini-Hochberg.

When the method type is Bonferroni, the average trials for a mean of 7 is 14.4 for a size of 10, 8.6 for a size of 50, and 17.6 for a size of 100. When the mean is 9 for Bonferroni, the average number of trials for a false negative is 15.6 for a size of 10, 9.8 for a size of 50, and 42.8 for a size of 100.

For Holms, when the mean is 7, and the size is 10 the average number of trials is 19.2, a size of 50 is 51.2 times, and a size of 100 is 14. With a mean of 9, when the size is 10 the average number of trials to receive a false negative is 19, 25.9 trials for size 50, and 23.2 when the size is 100.

Finally, for Benjamin/Hochberg, when the mean is 7, we see average number of trials to receive a false negative averaging at 12.2 trials for a size of 10, 12.4 for a size of 50, and 16 ties for a size of 100. When the mean is 9, we see the average number of trails clocking in at 14.6 for size 10, 47.2 for a size of 50, and 32.4 when the random sample size is 100.

Write-up Question 8 (3 points): In your write-up, describe in prose how each of these three methods seems to impact type II errors (false negatives).

In short, Bonferroni and Holms seem to have the most similarities between the means of 7 and 9 for the Type II errors, where as Benjamin/Hochberg has lower average trial values compared to Bonferroni and Holms.

Write-up Question 9: **Research questions (4 points).** Please specify what your research questions are. These research questions must be precisely stated, unambiguous, falsifiable, and concrete.

1. How might the incorporations of dark patterns (such as the ones in the article) increase the user time spent interacting with the ads?
2. What is the average age of users that interact with the dark patterned ads in a way that aligns with the expected/desired outcome of the manipulative presentation?

Write-up Question 10: **Interface Design (4 points).** Please describe the design decisions you made in creating your interfaces and the reasons behind your design decisions. If you changed your design after piloting, please also include what suggestions your participants provided and what changes you made based on them.

For the pilot, I did not include the Current Time questions, but, after reviewing the research questions with friends who partook in the pilot testing, I added the ability for the research participants to input their time at the times in which they answered each specified section of the survey. Additionally, my design decisions include three main components, ads that do and do not contain dark patters, time related questions, and age specifications. For the creation of either type of ad in Figma, I kept the word count around the same so that users will have to read the same amount for either ad to try and keep consistency (so that time spent reading the ads would not pose as a confounding variable in the experiment) and incorporated common types of phrasings that would guide the user to pick a desired outcome. I had some issues initially with the layout of my questionnaire regarding the phrasing of questions (the times inputs were confusing for those taking the quiz and some questions were ambiguous as to which question corresponded to which image). From there, I changed the way I worded particular questions pertaining to the time inputs and how I referenced the images in my question to be clearer about what was being asked of the research questionnaire participant.

Write-up Question 11: **Detailed Survey Design (4 points).** Please briefly describe your survey protocol, such as what questions you have designed and what techniques you have used to make the survey more valid (and avoid confounds). For each question, please provide information such as why you put this question here, what types of data you hope to collect and how that can help you answer your research questions. If you thought about including other questions and chose not to include them, or eliminated them through piloting, include them here and explain why.

Since my questionnaire was based around asking questions to a wide range of ages, I wanted to keep it as straight forward as possible as to not confuse the participants that might not have as much experience taking electronic quizzes. My research questions are based on how dark patterns affect people of different ages as well as the times that users spend on each online add that contain dark patterns. For this reason, I start off the quiz by taking the age of the quiz taker. For more concise data, I use age ranges and grouping as discussed earlier in the course (as noted from in class discussions after the completion of Assignment 1) for the ages. The time is taken at the beginning of the question pertaining to the dark pattern image. This is to act as a makeshift "stopwatch" to take the times of the users to gauge how long they are spending on either the dark pattern or non-dark pattern ads. Image one is based on regulations placed on baggage protection, flight protection, and seat upgrades that are already in Delta's flight policies but are worded to "trick" the user into thinking they are paying $30 for added benefits that are perhaps better or more impactful than they are, in reality. Additionally, the wording of the Flight Protection purchase, and UI clearly indicate that there is a perceived right and wrong choice. By using selective word choice, red and green coloring, color contrast, and check vs x symbols, there is a clear choice as to what the company would want the user to select. The user then uses the UI of the questionnaire to select their choice on the drop-down menu. This process is repeated for Image 2 (the ad without intentional dark patterns) with a similar word count to prevent confounding variables of reading time (as opposed to comprehension time). Finally, a change with the ad lacking in dark patterns is the box to either check or leave blank the indication of the purchasing of flight protection. Again, I used colors, contrast, and word choice to eliminate any bias that the company may have to push any agenda.

Write-up Question 12: **Piloting (4 points).** Describe your process of running pilots and describe in detail what you changed about the interface and the survey after piloting, and why.

Initially, I ran into interface issues using Gorilla for my questionnaire, especially the addition of options for the drop down menus, as some of the options included commas, but the delimiter for the choices was commas (on the set up side of the questionnaire). This, along with wordings of the questions needed to be changed in particular. For some questions, I had friends/experiment participants confused as to what time they were to input in the three different time categories (whether they should put the time they see that question, put all times at the beginning, or put all times at the end of the survey). For this, I attempted to indicate that they should select the current time at which they arrive at that given question/input. Additionally, for Image 2, I had pilot testers confused as to the choices present in the drop down since it was not quite the same as Image 1 (image 1 having two distinct options, whereas image 2 has one distinct option that you can either comply with or reject as a choice). I then changed the wording for questions relating to both Images 1 and 2 as to keep consistency between the questions.

Write-up Question 13: **Analysis Plan (4 points).** Please describe how you would analyze the data if you were to collect enough data in the future to make analysis meaningful. We acknowledge that your sample size is too small for analysis. Therefore, for this part, please imagine that you have already collected enough data for any analysis. What would you do to analyze them?

In relation to my research questions, I would first explore the relation between the age groupings and the number of people who indicated that they would buy the flight protection with the dark pattern as well as a second analysis looking for the correlation between age and buying the flight protection with the ad that did not consist of dark patterns. Additionally, I would look to analyze the hypothesized correlation between the users who did not buy flight protection with the ad containing dark patterns and their age.

Pertaining to my second research question regarding the time users spent on either dark pattern or no dark pattern ad, I would compare the average times spent for each age group on either ad and compare the times. I believe that the users would spend more time reading and deciphering the ad containing dark patterns than the ad without, which may be reflected in the values collected from the survey.

Write-up Question 14: **Results (4 points).** Please provide and discuss your results. Because of the small sample size, you don't have to follow the analysis plan you proposed in the last part but describe any interesting initial trends that could only be generalized with much more data. Your report will be graded based on the quality of your study design, *not* on the quality of your data. Therefore, don't worry too much if the data says your design doesn't work. Producing a valid and fair experiment and discussing your result objectively are what we value.

I found a surprisingly large number of participants that were of the age range 18-24 years old that also selected to protect their trip in the dark pattern ads. I also saw a trend that, in this age range, those who selected to protect their trip with the ad that contained dark patterns also selected to protect their trip with the ad that did not contain any intentional dark patterns.

Something I also found surprising is that, with the participants who were older, they selected *not* to protect their trip, even with the dark pattern ads. This is something that, if I were conducting a large-scale experiment, may be avoided as the older participants I shared the link with were my parents, grandparents, and friend’s parents, which, now that I am thinking about it could be an issue of their respective lines of work requiring them to travel a lot! I also noticed on participant that was of the 55–64-year-old range that opted to *not* purchase the flight protection with the dark patterned ad, but *did* purchase the flight protection with the ad that did not contain intentional dark patterns. The additional information they left regarding their decision for both Images 1 and 2 related questions was simply “plane ticket”. In the future, I would hope to add more detailed questions as to resolve this issue, or rather learn more about the thought process behind the selection of this option.