1 Question 1

Exercise A

a. The goal was to understand how fast my cat ate his dry food at each meal. The modes of collection were observation and measurement as I observed the time of day and measured how long it took him to eat his food.

b. The data was acquired by measuring the time it took my cat to eat at each meal using the stopwatch function on my phone. The date, time of day, time taken, and any notes I had were then recorded in to a note on my phone. Approximately every 30 entries, the note was then uploaded to my computer as a text file and then imported into excel using linear alignment. The time was stopped roughly when I saw him eat the last piece of kibble which was problematic when he chose to leave three pieces and then walk away to demand attention.

c. Initially I was using the AM/PM format for time, and I decided that 24-hour would be more recognizable and simpler for data purposes because it was only numbers. The use of year/month/day format for the date was very easy to use and much more readable. All the timing data was recorded to the 10th of a second, but that is definitely too precise for the actual measurement technique, so I might drop any decimal places.

d. The only metadata stored was what each column of data was for and the format of the data which was included in the headers of the columns of the excel sheet. Provenance about how the data was recorded, such as one tablespoon of kibble in the morning and at night, were recorded into a text document that accompanies the data.

e. The csv file can be found at the GitHub Repository for this project.

Exercise B

a. The goal of this exercise was to understand the very general sentiment people had towards the state o the world and how it was in the past. The question I asked myself for this exercise was "Do people think the world is getting worse?" The mode of collection was generation as I sent out a survey to obtain these results.

b. The survey was performed using the service Google Forms. The survey link was then sent out to r/SampleSize to get data results. Initially, I said I was going to send it to friends and family as well, but I didn't feel that was necessary. Google Forms has an automated system that automatically translates the data from the form format to tabled form. I then downloaded this file as a csv.

c. Because the data was only collected from a Reddit subreddit the data was biased towards the 18-24 demographic. Also because this was posted in a public place without regulation of who was taking the survey some answers were done only to be problematic. These sorts of answers will have to be cleaned from the data manually. In the plan it was stated that this data would be stored in JSON, currently the data is stored as a csv. If the need is warranted the data will be converted to a JSON.

d. Metadata for the answer fields is recorded in the header columns of the csv file. Some metadata will need to be recorded about the ranking process used for the survey questions. Provenance about the population from which the data was obtained and survey process was recorded to a separate text file.

e. The csv file can be found at the GitHub Repository for this project.

2 Question 2

Exercise A

a. The biggest success of the data management plan was the use of GitHub. As it is a revision tracking software whenever I make changes to the data now, it will be tracked. The method of

handling the data was altered. I did not think about the fact that when I feed him right after I get up I won't wait to turn on my computer to record the data into the excel sheet. Instead I just wrote it on my phone in an organized fashion. Most parts of my data management plan did come out as expected, this is due to the simplicity of the data collection process.

b. I don't believe I would need to change anything for another iteration of this data collection exercise. A possible change would be feeding him at exactly the same time everyday to see if that would normalize how long it took him to eat.

c. There was some variation in the quality of the data because I just assumed that he would finish his dry food every time. Fortunately, most of the time he did but occasionally it was a problem. The metadata collection had no problems.

Exercise B

a. As stated for Exercise the Data ownership, Persistence, and Discovery categories were handled collectively by using GitHub to store my data. GitHub also has no file standards so my data can be in any format while held there. For a short time I was unsure if I'd continue with the rank your agreement with the statement on a scale of 1 to 5, but I did and I think it worked quite well. It is allowing me to see the trends in the data appropriately. I had planned to collect data from more than just r/SampleSize, but I received nearly 200 responses so I stopped there.

b. As most of the data was collected from Reddit the groups tend to be 18-24 year old students. I would have liked to have more data from the 50+ categories. In order to get that data I would have had to specifically seek out those groups which is certainly more difficult.

c. The data will definitely be skewed heavily towards a liberal perspective. Except for a select few subreddits, Reddit is fairly liberal which could potentially skew answers to questions about human rights and environmental protections. The questions also asked respondent to compare to when they were children. Since a plurality of the respondents were in the 18-24 age range, they were "children" roughly six years ago which isn't a huge time difference. Overall I am happy with the quality of the data.

3 Question 3

Exercise A

a. I mostly used excel, but I also used the data format of putting the results into a text document on my phone which is abysmal. I did it out of convenience of it being early in the morning and doing things quickly.

b. Conventions on the identification of a type of cat would be helpful just for clarity on the provenance.

c. I was not aware of a best practice for the type of collection I carried out. I took assumptions on the data fields that I would need for this process as it was a relatively basic one.

Exercise B

a. I didn't use any officially recognized standards for my data which could potentially lead to some problems for the collection. Having a metadata standard would be helpful considering how many words are in the some of the survey questions.

b. An international standard on how to collect race and ethnicity data would have been helpful. I was aware that Reddit was global platform and so I asked for people's country of residence but didn't think that the way I was asking ethnicity questions would then not be applicable.

c. I used the best practice for race and ethnicity as specified by the U.S. Government. I also used a 10 year age range standard in order to increase respondent privacy.