This is a semester project devided into 4 steps and I need help with the step 1 below.

Project Step 1

Please note that being sloppy in these early steps will cause you to have compounding issues throughout the semester. Please go slowly, deliberately, and thoughtfully. Be mindful and have a reason for the things you do. MOST IMPORTANTLY, please ask for help or for clarification if there's something you're in doubt about.

The first assignment may be 'child's play' for some of you who are used to working with data. For those with less experience, pulling data together and making sense of it can be tedious and challenging. It is absolutely essential though, because very small errors in this phase of data analysis leads to total garbage down the line. One record being misaligned when it is imported can offset every other value, progressively throughout the file, and lead to computations that produce zero relations among data fields (variables), when they may actually exist.

Background: I've obtained data from a large survey done by the Pew Research Center. Using their words, "The Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. We conduct public opinion polling, demographic research, content analysis and other data-driven social science research. We do not take policy positions."

1. Download and open the data document in something like Notepad or Wordpad (any simple text editor that will not alter the file). Scroll through it in all directions. You will see it has a certain type of structure.

a. QUESTION 1a - What item divides each of the bits of data from each other?

Usually, the divider (aka "field delimiter") is put there by a database output program. That method can be very problematic. Imagine you have a large dataset that you're going to load, with variables divided from each other by that same item.

b. Question 1b - One common type of field delimiter is a file format called "Comma Delimited", or Comma Separated Values (a '.csv' file) What could go wrong when you are loading data, if the datafile uses a character such as a comma as a field separator? (If you are having trouble downloading, here's a video to walk you through: <https://youtu.be/9LJGsT946PY> ).

2. Download and open the "Metadata" documents. These tell how to interpret the information in the large data file. One tells you the questions that correspond to the variable labels.

a. How many variables are there?

b. How many data records are there?

c. Read through all the survey questions once.

d. Select and Thoroughly Consider a Set of Variables

d1. Go through the meta data showing the questions and anwers and flag 10 - 20 that you find interesting and list them for me. NO MORE THAN 20 should be listed, and no fewer than 10.

d2. Explain why EACH ONE looks interesting. "They are kinda cool" or any other group-level summary is insufficient.

d3. As you list EACH variable, show what the response options are. Show both the numeric code and the text corresponding to that code. (You will have to use both metadata files for this.)

d4. As you list EACH variable, identify what type of data it is. "They're mostly nominal" is an example of an insufficient answer.

(See the video AND the attachment "Exercises on Data Types" to understand these.)

e. Now that you have read the survey questions and answers, in a short paragraph (perhaps 3-8 sentences) describe what this lengthy survey is about.

3. Protect your original data and structure the data for analyses by saving the original data, and making a copy of that original file to use exclusively for analysis. It can be absolutely disastrous to miss this step, because tampering with the original data can cause permanent data loss when you do this on the job. Obtaining data is often very expensive, and, for example if you tamper with the original copy in an irrecoverable way, you cannot re-interview 1,000 customers to fix the data nearly as easily as the boss can fire you to solve the problem.

a. Open the data file using whatever software you have available for analysis. For most of you, that may be Excel. Excel will open a data import "wizard", which is allows you to specify how to read the data. The Wizard is not very smart, and should by no means be trusted! Carefully read through each wizardly step. First, you'll need to keep in mind your answer to Question 1a, above. A MUCH more powerful analytic software is available for free for students from TTU, called "SPSS". It can be located here: <https://www.depts.ttu.edu/itts/software/spss.php> . Some old computers may have trouble running this program, even though it is comparable in size to MS Office. You may choose any software available to you. Instructions for all assignments will be geared for Excel, since that is most prevalent.

b. Step 2 of the Wizard will show what happens if you mis-identify the delimiter (the divider between variables or fields). Select different options to see what effect it has on the data arrangement in the window at the bottom. The goal is to get fields (variables) into columns, such that there is a unitary bit of data in each column.

4. Clean and check for mis-imported data.

a. This largely involves graphically examining data ‘structure’, by scrolling through the data file and scanning it visually, and by producing plots of the frequencies of the values in each field or variable (column). You obviously do not have to produce graphs (bar chart, histogram, pie chart, line graph, etc.) for every variable. For data verification and validation, it is a good idea to examine the first and last variables in the dataset, because they're sensitive to data loading errors.

b. Choose 4 variables that you're interested in. (Recall the flagging of interesting variables you did in step 2d). Because this data is coded for verbal meaning, READ ALL of the metadata file explanations of each numeric value, and see how many of each response is being endorsed. Numeric values MIGHT NOT INDICATE A QUANTITY OF SOMETHING. You absolutely HAVE to understand whether your variables have numeric codes that are placeholders for categories, or if the numbers indicate 'more' or 'less' of something.

c. Watch these videos. I couldn't improve on these two videos for making graphics in Excel:

<https://www.youtube.com/watch?v=is14ehdy7jo> - shows how to produce histograms in Excel - which show the distribution shape for ordinal, interval, and ratio data.

<https://www.youtube.com/watch?v=IVhQTAF1guc> - A slightly longer video showing the distinction between a bar graph and a histogram.

5. After making appropriate graphics for your four variables, present the graphics, and describe your graphic results verbally, including AT LEAST:

a. The types of data evaluated for each (see Data Types video),

b. The shape of data on the graphs, as relevant. You can use (and look up the descriptions for "skewness vs symmetry, normalcy, etc.", but you are not limited to those. A normal English-language description is fine. Be absolutely sure to connect the shape you're describing to the responses and the meaning of the responses. We are trying to learn about the responses represented IN THE GRAPHIC, not about the graphic by itself.

c. The central tendency and spread of the data.

d. Provide an interpretation of the respondents' pattern of responses. An example of a complete answer, appropriate for a graduate business professional working in industry with an MBA would be, "Although a majority of respondents answered 'No' (Response code = "2", n=2,734, 64%), indicating that they do not have a smart device in the home, just under 2,000 indicated that they did (Response code = "1". Furthermore, 197 people were not presented this question (coded as "#N/A!"), and 3 people declined to answer (coded as "99"). Overall, XX% of people indicated they have a smart device in the home, which is (more/less) than I expected." An example of a terrible answer is "The mean answer was 'No'."

6. Using the data for two ordinal variables: use the following steps to compute the standard deviation for each variable.

a. Copy and paste both variables into a new tab (all the rows), in columns A and B.

b. Turn on the Filter, so that there are drop downs at the top of each column.

c. Use the meta-data files AND the filter drop down to identify the data that will not work. Typically, this is 99 and #NULL!, but there are stray invalid values in the data, so check what the metadata says the meaning of each coded data value is!

d. Sort the data, high to low, on column A. Delete the "bad" values (those which are not a meaningful response from a participant). Deleting the entire row from Excel is preferable. Re-sort the values so that the whole column is only valid data.

e. Repeat step D with the other variable, in Column B.

f. In Cell C1, put a heading "X1 - Xbar", and in cell C2, type '=A2 - average(A:A)'. Hit return key. If you cleaned the data successfully, then it will provide a value. Select cell C2 again, and double-click the little green nub in the lower right corner of the cell. If things went right, that will cause the formula to repeat for every single row, all the way down to the last row of data in Column A. You can grab and drag that little square green nub downward, to make it repeat the computation, if it does not fill automatically.

g. In cell D2, put a heading "(X1 - Xbar)^2", and in cell D2, type '=C2 \* C2'. Then double click the nub again (or drag it downward), and you'll have ALL the squared distances of every single person's response, from the average of that variable. Here's why we made that column: When added up, that column is the numerator of both "Variance" and "Standard Deviation". NEXT: Take the average of the squared column, and that single average value is called "Variance". Again, the average of the column of each person's response's squared distance from the mean is the VARIANCE of that variable. Go ahead and calculate the variance by taking the average of that column. When you have it, take the square root of that Variance value, and that square root is called "Standard Deviation" for the variable. It can be thought of as 'an average' distance of all the scores or values in the variable spread around the mean of that variable. It's the average spread of scores, and tells us how much the scores are spread out or distributed around the mean. Another way of thinking about it is it tells us how much the mean is crappy at summarizing all the data in that variable.

h. Repeat steps F and G, for whatever variable you pasted into Column B, to get the Variance and Standard Deviation of the 2nd variable.

i. Double click the Tab title, which is probably something like "Sheet3", and rename the tab "StandardDeviations".

7. Submit (upload) your Excel file, with your responses and actions for Items 1-6, above. Please paste the actual questions you're resonding to, from above, and insert your responses next to them.

Materials:

* Raw Data - Pew Research Survey ATTACHED
* Metadata File - Variable Labels and Survey Questions - Response ATTACHED
* "Dummy Codes"- Alternative Formats ATTACHED

video reference for the project

<https://www.youtube.com/watch?v=w2QCp2eJlrk>