CS 301 - Project Deliverable 2 Expectations

You are expected to continue writing the next section on the **same** document you used for Deliverable 1. Each Deliverable will cumulatively add on to the previous ones. This is factored in as a small portion of your project grade. This section should include the changes that were suggested after grading the deliverable. All code for this Deliverable should be included and updated in Section (1h). As a reminder, below are the included sections:

1) **Section 1: Overall Information**
    a) Title of Project
    b) Overview / Summary
    c) Problem Definition (Scope)
        i) Identify 3 questions which the group seeks to answer
    d) Period of Analysis
    e) Contact Information (Vendor of Data Set)
    f) Documentation (Given by Vendor)
    g) Sample Data
        i) URL Link to Set(s)
    h) Code - Link to the project repo on GitHub - DO NOT paste raw code here.
2) **Data Documentation Section 2: Data Exploration** (**Due 11/6 11:59PM EST**)
    a) Data Collection
        i) This subsection should include a full description of how the data was retrieved.
        ii) Links to ehe data sets must be made available in Section (1g), and screenshots of the data (first few rows of each data set) should be provided as well in the same section.
        iii) If the data is being created, then this section should also have a 1-2 paragraph description of how the data set was created, with code available in Section (1h).
    b) Column Descriptions
        i) Name and brief description of each column
            (1) Example: 'Dates' - Each date where the data was recorded, this includes the timestamp in YY-MM-DD-HH-MM-SS format.
        ii) What percentage of the data in the column is populated?
        iii) Data Type of the column (Categorical/Quantitative/Ordinal)
        iv) Describe the set of values in the column:
            (1) If it's a categorical column, you can say how many unique values are included in the set, and include *some* of the values. Example: "Celebrity Name' - There are 437 unique names in this column such as, {"Steve Jobs", "Mark Zuckerberg", "Bill Gates", …., etc}

(2) If it's a numerical column, you can simply provide the range of values, and mention the outliers that exist. Do not provide any more statistics other than range of values. Example: "Student Age" - Ages are integers ∃ [19, 25] inclusive, there are three outliers: {47, 48, 56}

(3) If it's a date column, you can simply provide the date range, and mention if it's daily, monthly, yearly, etc. Example: "Dates": 20020101-20180501 - Measurements are weekly

c) Data Processing

i) You should talk about the data as it was given to you. Remember that with any raw data set, there can potentially be six common problems:

(1) Outliers
(2) Outdated data
(3) Duplicate entries
(4) Incorrect values and/or column names
(5) Spelling/punctuation errors
(6) Missing entries

ii) Although you are not expected to have perfectly complete solutions for each, you should at least describe **if** the above problems exist in the data set, and what are your groups proposed solutions to solve those problems. When you submit, you should have the attempts made using code, clearly for the reader to understand, available in the GitHub link in Section (1h).

ADDITIONAL REQUIREMENTS:

1) Each group member must submit a Contribution Document. Maximum 1 double sided page, single spaced, size 12, Times New Roman. I expect that you discuss precisely the work you put into this stage of the project. You should include as much detail as you think is needed such that I can fully understand your work towards this deliverable. In addition, if you have any complaints / praises to specific teammates you would like to communicate, please do so. Please refrain from using any harsh language.
    a) This document should not be a group effort. I trust that each individual group member will work on this portion alone, and communicate to me their own honest opinions. This way, I know if I need to communicate to the group about sharing responsibilities, or if I need to adjust the points I give to any one particular group member.
2) Each group member is required to submit the Contribution Document on Canvas as a PDF.

3) Only one group member is required to submit the Group Deliverable Document on Canvas as a PDF, however if more than one person submits, you will not receive points off.
4) All code must be submitted in section (1h) of the Group Deliverable Document. It must also be zipped together into one folder and uploaded to Canvas.
5) For the Group Deliverable Document, you must produce at least one full page of text. The body of each section should be single spaced, Times New Roman font size 12. There is no specified format for the Document as a whole, however the Sample Data Documentation (on Canvas) is a good start.
6) If there are any screenshots you would like to include in the documents, please do so. However, you are required to cite where the image came from. If you took a screenshot of something you created, your citation can be a brief description of how you produced the image. If the screenshot is from Google or some other website, please provide the link to the screenshot. You do not have to use formal citations here.