Response 1: Mahshid: During my NLP course, I encountered a dataset with errors while working with Copra, such as mislabeled categories and inconsistent values. I discovered these errors during the exploratory data analysis phase and addressed them using appropriate data cleaning techniques “standardizing labels” & “ filling in missing values”

Response 2: Yes, I (Max) had to manually fix a CSV in a text file for work, about 20 lines changing one value. It would have needed a script if there were more data errors

Response 3: Timofey: I've run into (sort of an error) the Trimet data we are using for the project. On the days when no data is available, the API returns [{"data not available for this day"}], which is a problem because it's not a valid JSON object. To avoid attempting to dump such data into a Python dictionary, I would first check the size of the string. If it was relatively short, I would assume that it's an error message and would not parse it.

Response 4: Dan: I personally haven't. But I feel like we should count the dataset that we worked on the previous week. Two datasets that came from the same website (for BeautifulSoup), created problems with our code being robust. I ended up giving up on trying to make the code robust but if I had more time I would've definitely spent it on fixing the code to be able to include both the datasets.

Existence assertions:

1. Every crash has a unique Crash ID.
2. Each crash has a Crash Date.

Limit assertions:

1. All Crash Hour Codes are within the range of 0 to 23.
2. All Posted Speed Limit Values are non-negative integers.

Intra-record assertions:

1. If there is a value for 'Latitude (Decimal Degrees)', there should also be a value for 'Longitude (Decimal Degrees)'.
2. If the 'Crash Year' is 2019, then the 'Crash Month' should be between 1 and 12.

Inter-record checks assertions:

1. The total number of 'Crash ID' should match the total number of unique 'Crash ID' in the dataset.

Summary assertions:

1. The total number of crashes is greater than zero.

Statistical distribution assertions:

if the crashes are evenly distributed throughout the months of the year by calculating the coefficient of variation, which is the ratio of the standard deviation to the mean. A lower coefficient of variation indicates that the crashes are more uniformly distributed. The threshold of 0.1 is just an example, and you can adjust it based on your requirements and analysis.

AssertionError: Participant ID column contains a null value
remove the rows with missing values in the 'Participant ID' column.

AssertionError: Vehicle ID column contains null values
remove the rows with missing values in the 'Participant ID' column.

AssertionError: Not all Crash Hour Codes are within the range of 0 to 23
remove the rows with missing values in the 'Participant ID' column.

AssertionError: Crashes are evenly/uniformly distributed throughout the months of the year
I provided an alternative approach to handling the AssertionError by printing a message instead of raising an error. This falls under the "Ignore" method from the options you provided.

I have learned that data validation requires a high level of accuracy and that a significant amount of effort is needed to properly prepare data before it can be analyzed in an official capacity.