Avinash V Cherukuri
avin2@pdx.edu

CS 510 - Data Engineering
7 February 2021

Answers for Data Validation Activity

Part A - Create Assertions

1) Create 2+ existence assertions. Example, "Every record has a date field".
   - Every Record has a Crash ID and a Record Type

2) Create 2+ limit assertions. The values of most numeric fields should fall within a valid range. Example: "the date field should be between 1/1/2019 and 12/31/2019 inclusive"
   - Week Day Code and Age

3) Create 2+ intra-record check assertions.
   - Total Count of Persons Involved
   - Total Persons using Safety Equipment

4) Create 2+ inter-record check assertions.
   - Latitude is always 45 Degrees and has minutes ranging from 17 to 41

5) Create 2+ summary assertions. Example, "Every crash has a unique ID".
   - Every Record has a Vehicle ID and Participant ID

6) Create 2+ referential integrity insertions. Example "every crash participant has a Crash ID of a known crash"
   - Every crash record has a County Code and Vehicle ID

7) Create 2+ statistical distribution assertions. Example: "crashes are evenly/uniformly distributed throughout the year."
   - Crashes are discretely distributed throughout the year.
   - Crashes are kind of uniformly distributed across Crash day.

Part B - Validate the Assertions

5) Run your code and note any assertion violations. List the violations here.
   - Age Column Violation

Part C - Evaluate the Violations

- Found Age column has Empty cells. I dropped these rows to overcome the violation.

Part D - Learn and Iterate

- The dataset is huge and it gets complicated when information from various tables are related. There are no clear indications of why somethings are the way they are. For example, empty cells in age column can sometimes indicate no data and sometimes they can indicate that the data is lost during any operations.

Part E - Resolve the Violations

- I segregated the data into multiple data frames to better understand what the data is trying to tell. I also dropped the rows with empty cells (I am not 100% sure on this but I did it)