Likhitha Vanga
977673912

DataEng: Data Validation Activity

1. Create 2+ *existence* assertions. Example, "Every record has a date field".
    a. Every crash ID has Record Type. (Validated)
    b. For record type 3, with vehicle ID 0, Do not generate a vehicle record for pedestrians, pedal-cyclists, or another non-motorist. (Validated)
    c. 90% of the records have Vehicle ID. (Validated)
    d. For record type 3, with vehicle ID 0, and injury severity 7. The entry is invalid. (Validated)
2. Create 2+ *limit* assertions. The values of most numeric fields should fall within a valid range.
    a. Highway Number – 26
    b. Year – 2019 (validated)
3. Create 2+ *intra-record check* assertions.
    a. Total Non-Fatal Injury Count = Total Suspected Serious Injury (A) Count + Total Suspected Minor Injury (B) Count + Total Possible Injury (C) Count - Total Fatality Count
    b. Total Persons Not Using Safety Equipment = Total Count of Persons Involved - Total Persons Using Safety Equipment
4. Create 2+ *inter-record check* assertions.
    a. For record type 3, with vehicle ID 0, the Vehicle Coded Seq# is null and injury severity is died prior to crash. (Validated)
    b. Special Jurisdiction is for roads that belong to an agency other than the State, County, or City transportation department.
5. Create 2+ *summary* assertions
    a. Every crash-id has a unique serial number (Validated)
    b. Participant ID is unique. (Validated)
6. Create 2+ referential integrity insertions.
    a. Every Participant ID has a crash ID, vehicle ID for a crash (Validated)
    b. Every Participant ID has a Participant Display Seq# for a given crash
7. Create 2+ *statistical distribution assertions*
    a. crashes are evenly/uniformly distributed throughout the year
    b. Most of the crash types are 4. (Validated)
    c. Most of the crashes occurred in the beginning of the year (Validated)
    d. Crashes occur mostly on Saturday and Friday (Validated)