DataEng S24: Data Validation Activity

A. Initial Discussion Question

Have you ever worked with a set of data that included errors? Describe the situation, including how you discovered the errors and what you did about them.

Response:
I found an issue during a data migration where the date columns had different formats between the old and new databases. To align the formats, we needed to calculate the century part of the year based on the current year's last two digits. However, during the migration, the year changed to 24, leading to incorrect updates (e.g., 1923 instead of 2023). Upon validation, I identified and reported this error to the database administrator. We updated the code responsible for date conversion and checked other code segments to prevent similar database update errors. After correcting these issues, we successfully loaded the new database with the correct dates.

B. Create Assertions

1. *existence* assertions : Every crash record has a crash date
2. *limit* assertions : The crash day must be between 1 and 31,The crash month must be between 1 and 12
3. *intra-record* assertions : Each participant has a age
4. *inter-record check* assertions : Every participant was involved in a known crash, every participant is associated with a vehicle ID
5. *summary* assertions : All of the crashes occurred in the year 2019, All of the crashes occurred in the highway number 26
6. *statistical distribution assertions* : The crashes were evenly distributed in all days of the week, The crashes are evenly distributed on all hours

D. Run Your Code and Analyze the Results

In this space, list any assertion violations that you encountered:

- Intra record assertion - Every participant has an age
  Violation : Not every participant has an age associated.
  To resolve this violation we can fill in the missing values with a particular age code that would correspond to an unknown age.

Srivashinie Dhamodharasamy (954814608)

F. Learn and Iterate

The process of validating data usually gives us a better understanding of any data set. What have you learned about the data set that you did not know at the beginning of the current ABC iteration?

Response:

From validating the dataset I have understood that there is a lot of data that is decoded to protect the identity of a person.The data only includes crashes from 2019 on Highway 26. It's interesting that crashes happened at similar rates on all days and hours. Also, the dates of the crashes have been recorded accurately. Each crash is associated with vehicle and participant IDs.For every recorded crash, there is corresponding vehicle and participant information.

New Assertion:

1. *existence* assertions : Every crash has a highway number and roadway number
2. *limit* assertions : The latitude values are between -90 and 90, and longitude values are between -180 and 180
3. *intra-record* assertions : If the crash month is January, March, May, July, August, October, and December the days must be within 1 and 31, if the crash month is April, June, September, and November then crash day must be 1 and 30. If the crash month is february then the crash day must be within 1 and 28.
4. *inter-record check* assertions : Crash ID must be the same in the crash, vehicle and the participant dataset
5. *summary* assertions : The total number of crashes is less than 1000, The total number of participants in the crashes were less than 1300, The total vehicles involved in the crashes were less than 1100
6. *statistical distribution assertions* : All types of vehicles were involved in the crashes