John Lorenz IV
CS 510: Data Engineering
Irvine
4/22/2022

In-Class Assignment3

Responses

Varun Jaisundar Raju:
Grabbed pokemon database from google, and there were missing fields like evolutions and types. The steps for solving it were to reorganize and independently validate the data

Myself (John Lorenz IV): Titanic survivors dataset off of Kaggle. I had to fill in missing information for many of the data points, or make assumptions for any relationships (if any existed), but ultimately with how old the information was it made more sense to simply re-represent the data and change the scope of the application.

Zhengmao Zhang: Sometimes people come to the wrong products when using NRP algorithm. Validated the result against some set of expected search results.

Chinmay Tawde: Import some records from a staging table, and two companies provided tables but sometimes it would be missing data. He would either say it was invalid and request correction, or it would be filled with default information

Assertions

Existence Assertion:

1. Every crash will have an ID associated with it.

Limit Assertion:

1. Every crash will have at least one participant.

Intra-Record Assertion:

1. If there are two vehicle IDs in a record, then there are at least two participant IDs in that record.

Inter-record Assertions:

1. If the vehicle seq# increments on addition of new vehicles to a record, then it must do that for all records.

2. A record containing a crash ID can’t have that same crash ID exist for a separate set of vehicle IDs.

Summary Assertions:

1. All records must be from the year 2019.
2. The largest vehicle seq# and largest participant seq# must add up to the sum of everyone involved in crashes on Oregon highway 26 for the year.

Statistical Distribution Assertions:

1. It is unlikely that crash data will be evenly distributed throughout the day. Bars closing, high commute times, and low commute times will influence this metric in a way that it won’t fit a normal distribution. However, that may change if we were to ‘zoom out’ on the data such as a decade-long ‘x crashes per year’.
2. Depending on the weather, some months may present more crashes than others.

Testing the Assertions

Limit

Using Pandas DF slicing magic in the function *limitAssertion* I found that each crash entry does in fact have **at least one participant.**

Intra-Record Assertion

For this one my assertion was wrong, and I likely just didn’t account for the presence of a tow truck or something which could contain more than one vehicle per participant, thus breaking the assertion.

```
Crash entry 1833678 has fewer than two participants for two or more vehicle
IDs.
     Crash ID  Record Type Vehicle ID Participant ID  ... Alcohol Use
Reported Drug Use Reported Participant Marijuana Use Reported Participant
Striker Flag\n
265   1833678            1                            ...
266   1833678            2    3453351                 ...
267   1833678            3    3453351        3934009  ...
268   1833678            2    3453352                 ...

[4 rows x 157 columns]
Assert FAILED
```