DataEng S22: Data Validation Activity

Name: Deepa Subray Hegde

PSU ID: 954680127

High quality data is crucial for any data project. This week you’ll gain experience with validating a real data set.

**Submit**: Make a copy of this document and use it to record your results. Store a PDF copy of the document in your git repository along with any needed code before submitting using the in-class activity submission form.

**Initial Discussion Question** - Discuss the following question among your working group members at the beginning of the week and place your responses in this space. Or, if you have no such experience with invalid data then indicate this in the space below.

*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 1: No experience with data outside of Data Engineering

Response 2: While working on project during Database course, whenever there was a data missing in the field, scanned it manually and entered dummy data of the same type

Response 3: While working on the Database project, validating the currency field when the amount was in different currencies. Scanning and conversion was done using manual naive approach

Response 4: While working with machine learning data. Manually checked whether the data belonged to a particular class

The data set for this week is [a listing of all Oregon automobile crashes on the Mt. Hood Hwy (Highway 26) during 2019](https://docs.google.com/spreadsheets/d/1GFBcW39v218FOjVZhsb5u-3LB7ok839BRuJA_2KwfT4). This data is provided by the [Oregon Department of Transportation](https://www.oregon.gov/odot) and is part of a [larger data set](https://tvc.odot.state.or.us/tvc/) that is often utilized for studies of roads, traffic and safety.

Here is the available documentation for this data: [description of columns](https://docs.google.com/spreadsheets/d/1AFa1p74257uTbb_cN8l2ztMPuFh2WGLq), [Oregon Crash Data Coding Manual](https://www.oregon.gov/ODOT/Data/documents/CDS_Code_Manual.pdf)

Data validation is usually an iterative three-step process.

1. Create assertions about the data
2. Write code to evaluate your assertions.
3. Run the code, analyze the results and resolve any validation errors

Repeat this ABC loop as many times as needed to fully validate your data.

## A. Create Assertions

Access the crash data, review the associated documentation of the data (ignore the data itself for now). Based on the documentation, create English language assertions for various properties of the data. No need to be exhaustive. Develop one or two assertions in each of the following categories during your first iteration through the ABC process.

1. *existence* assertions. Example: “Every crash occurred on a date”

* Every crash has vehicle ID
* Every crash has crash ID

1. *limit* assertions. Example: “Every crash occurred during year 2019”

* Every crash occurred in 2019 (Crash Year = 2019)
* Every crash month should be in the range 1 to 12

1. *intra-record* assertions. Example: “Every crash has a unique ID”

* When latitude degree field has a value, latitude minutes field should also have a value
* The sum of each record latitude degree,latitude minute and latitude second must be in the range of (41+0+0.00) and (47+59+59.99)

1. Create 2+ *inter-record check* assertions. Example: “Every vehicle listed in the crash data was part of a known crash”

* Every crash with record type 1 will have both record type 2 and 3
* Every crash ID has at least one vehicle ID

1. Create 2+ *summary* assertions. Example: “There were thousands of crashes but not millions”

* Crash ID is always unique
* Latitude degree is never “null” (it always has a value in the range 41 to 46)

1. Create 2+ *statistical distribution assertions*. Example: “crashes are evenly/uniformly distributed throughout the months of the year.”

* Crashes are evenly distributed throughout the day
* Crashes are evenly distributed throughout the months of the year

These are just examples. You may use these examples, but you should also create new ones of your own.

## B. Validate the Assertions

1. Study the data in an editor or browser. Study it carefully, this data set is non-intuitive!.
2. Write python code to read in the test data. You are free to write your code any way you like, but we suggest that you use pandas’ methods for reading csv files into a pandas Dataframe.
3. Write python code to validate each of the assertions that you created in part A. The pandas package eases the task of creating data validation code.
4. If needed, update your assertions or create new assertions based on your analysis of the data.

## C. Run Your Code and Analyze the Results

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

* Statistical Violation - From statistical assertion validation, it is evident that the crashes are mostly during the peak hours during day

For each assertion violation, describe how to resolve the violation. Options might include:

* revise assumptions/assertions
* discard the violating row(s)
* Ignore
* add missing values
* Interpolate
* use defaults
* abandon the project because the data has too many problems and is unusable

No need to write code to resolve the violations at this point, you will do that in step E.

## D. 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?

* I had few violations at the beginning when I started. However, that was due to my mistake of validating it the wrong way. This way I learned how to correctly validate the field in the dataframe. In the statistical assertion, I assumed that the crashes are more during the midnight and early morning hours than during the day. But the graph revealed that the crashes are more during the day than in the midnight early morning hours.

Next, iterate through the process again by going back through steps A, B and C at least one more time.

## E. Resolve the Violations and Transform the Data

For each assertion violation write python code to resolve the violation according to your entry in the “how to resolve” section above.

For statistical assertion violation - Since this was because of my wrong assumption, I would revise the assumption and change the assertion

Output the validated/transformed data to new files. There is no need to keep the same, awkward, single file format for the data. Consider outputting three files containing information about (respectively) crashes, vehicles and participants.

I have added the cleaned CrashesDF, VehiclesDF, ParticipantsDF into CrashesInfo.csv, VehiclesInfo.csv and ParticipantsInfo.csv respectively.