MBIS623 Case Study Report

The NYC311 Service Request Dataset Quality Metrics

Due: Sunday, 8 May 2022, 9:00 p.m.

Grade contribution: 30%

1. Overview

The MBIS623 Data Management course material examines the dimensions of data quality listed as **accuracy, timeliness, relevance, completeness, understandability, validity, uniqueness, integrity, and trustworthiness**. The NYC311 case study materials include the NYC311 service request dataset with 27,736,190 records. The course project is focused on the examination of the NYC311 service request dataset from **the point of view of data quality**. While only a subset of the quality dimensions listed above can be examined, you are required to determine the relevant dimensions and **develop a set of automated data quality metrics** to be used for data quality assessment.

The automated data quality metrics developed for the NYC311 dataset would have a lot of applications of determining the trustworthiness of data analytics which can be built into the data warehouse for the estimation of service quality level, strategic planning, resource allocation, operations support, publicity and marketing activities. **Metadata such as data quality definitions as well as reference data provide the foundation** for the ability of data governance activities **to qualify and enhance the value of the data as the key asset** of the NYC311 call centre. One of the key reference data sources in this case study are the data for the Zone Improvement Plan (ZIP) codes which were developed by the Postal Service in the United State and used by a wide range of government and corporate entities.

2. Project Tasks

You are required to **focus on the following fields/columns** of the NYC311 service request table:

- Borough
- Incident Zip
- Complaint Type
- The three dates fields: Created Date, Closed Date, and Due Date,

while evaluating the quality of the relevant quality dimensions which you are required to select from the set of the **listed data quality dimensions**. You are invited to examine the Collibra Data Quality White Paper available alongside the course project specifications.

Date Fields Data Quality

When examining the date fields for the NYC311 dataset, one of the data quality metrics could examine the properties of the data as defined by the data domain. For example, **how well do the date values fit in the expected domains for each of the attributes?** To answer this question each of the **date fields domains have to be established as part of the metadata** guiding the quantification of the data quality with the automated data quality metrics.

Incident Zip Field Data Quality

Similarly, reference data can be used to evaluate the quality of other fields/attributes. For example, what can you conclude about the quality of the Incident Zip and Complaint Type fields by **examining the summary tables** which include the aggregated counts of the values encountered in the dataset? And so, how well do the Incident Zip field values correspond to the **values listed in the reference Zip code dataset**?

Complaint Type Data Quality

As for the Complaint Type data, the picture is even more complex. For instance, in absence of a reference dataset for Complaint Type values, what can be done to estimate the domain requirements? What may come to the rescue here is the **definition of a map/association between the available values and either of the two reference datasets listing the problem areas or service request types** which can be used for the analysis and classification of the service request records. If you are choosing to work with the Complaint Type field, you are invited to have a go at developing a table for this kind of map/association.

NY Borough Field Data Quality

With the Borough attribute the picture is a bit more straightforward as it is supported by a compact and easily understood reference dataset which maps the zip codes to NY boroughs.

Submission Components/Requirements

Develop eight-to-twelve data quality metrics to be automated as database queries in SQL format. Please note: you don't need to cover all four fields examined in this section, but it is suggested that you use at least three of those to focus your attention on.

The submission must include two components submitted via UC LEARN, the Course Project drop box as two separate files, as follows:

1. A document (PDF or MS Word file) **three-to-five pages** detailing the following:
    - your very-well researched and supported **assumptions for the domains and data properties** supporting the dataset fields you are choosing to focus on,
    - the **list and the descriptions for the automated data quality metrics** you are choosing to work with. Be sure to state the quality dimensions you are attributing each metric to, supported by a brief explanation of your thinking,
    - a table of the **calculated values for each of the metrics**,
    - the **overall analysis and conclusions** of the data quality for the attributes you've examined.
2. A **set of queries** (SQL/MySQL) to be run against the NYC311 service request dataset alongside with the other reference and summary datasets.
    - Every query should test an aspect of the target data quality definition for a column and **express the result as a percentage or proportion** of the overall number of available records/rows.
    - Please make sure each query can be executed against the **service request sample datasets (either 1k or 10k)** because the queries run against the full dataset may be too costly to complete in terms of the time and memory resources required.
    - If you are developing any **additional tables** by extracting the data from the service request dataset or other provided datasets, please be sure to **include those queries** as this will be essential for validating your code.

3. Marking Schedule

Please note the following allocation of marks for your project submissions:

Report presentation, structure and references (if applicable), communication style, spelling and grammatical correctness:	20%
Appropriate level of the concepts' introduction, soundness of metadata descriptions, domain value assumptions, and data quality metrics descriptions:	20%
The coherence and clarity of the analysis and conclusions:	30%
Correctness and clarity/readability of the SQL code:	30%
Total:	100%

4. Getting Started and Seeking Clarifications

Some of the **material required for the completion of this assignment is still to be covered** in the course, including both lecture and tutorials, and so, it is essential you keep up with the theoretical and practical data management matters to be covered before the due date of the assignment.

Then, it is natural you will have questions to clarify some of the **requirements and various technical matters related to the dataset**, MySQL Workbench and the rest of it all. Please note: don't email these questions to your course lecturer/coordinator/tutor; instead, please make sure to **post these questions via the Discussion Forum**.

Given the current context we are working in, all things around distance learning, COVID restrictions/regulation and the like, there's no resources to maintain one-on-one communication about course-related matters when it comes to the questions that may be of relevance to the rest of the class. The one-on-one emails to the course teaching team will not likely to merit a response if they are suitable to be dealt with via the Discussion Forum.