Table of Contents

Introduction

This report examines the data quality of the dimensions that are present in the NYC311 service request dataset. NYC311 provides access to people of New York about various government programs and services such as homeless person assistance, complaints handling, repairs of roads or any other public issues in New York city (NYC311). People of New York can connect to NYC311 through various channels such as phone, social media, and mobile app etc. The service request dataset comprises of the issues reported and requests submitted through these channels.

According to (Gupta, 2021), data quality dimensions are used to measure the attributes of data which can be assessed, improved, and interpreted. The aggregated scores of these dimensions indicate the quality of data and how well it could be used for decision making. Though the service request dataset consists of 41 attributes, this report only investigates the data quality of six attributes which are Borough, Incident Zip, Complaint Type, Created Date, Closed Date, and Due Date. Also, this report makes use of the reference tables which consist of Zip code data developed by the postal service of the United States, complaint type data, and other summary tables computed from the main service request data set to help measure the quality of data.

Data Quality Analysis

To measure and analyse the data quality of the Service requests data, 8 metrics have been developed and the results are furnished below.

Metric-1: *Calculate the NULL value percentage in Closed Date column by Incident Zip code*

This metric explains the completeness of the data in the Closed Date attribute which represents the date a particular service request closed by the respective government agency.

Total	Incident Zip	count	null_percentage
2257	11426	45154	5
7986	11221	328570	2.43
1877	10037	66495	2.82
10307	10314	218093	4.73
3571	10035	131841	2.71
5989	10025	222344	2.69
6576	11211	238104	2.76
8432	11207	354983	2.38
3517	11365	95175	3.7

Table 1: Percentage of NULL values in Closed date by Zip code

As shown in table1, there is a minor percentage of closed dates are missing from the data for each zip code which impacts the ability to analyse the efficiency of NYC311 to close the service requests or complaints in a timely manner. This can lead to lack of clarity when the government wants to know if the public complaints are being resolved in the given timeframes for specific areas of the city.

Metric-2: *Percentage of service requests which contain Closed dates but no due dates*

This metric investigates both the consistency and completeness of Due date column which is the date the relevant government agency needs to action, update, or complete a particular service request. Table2 shows the percentage of data records that do not contain the due date but contains the closing date. It clearly shows the inconsistency of entering the due dates for the agencies to action upon a particular request which not only reduces the operational excellence but also impacts the analytical capabilities if the government wants to measure the efficiency of agencies actioning the requests by

due dates. Also, the high percentage of missing due dates shows that for most of the zip codes, more than 50% of the records are missing with due date values which lead to the incompleteness of the data.

Incident Zip	Totals	Count	missing_dates_percent
10803	101	199	50.75
10023	73737	145092	50.82
10280	3268	6366	51.34
10122	36	70	51.43
10021	43252	83387	51.87
10151	95	182	52.2
10172	80	153	52.29
11242	57	107	53.27
11040	3637	6826	53.28

Table 2: Percentage of service requests with Closed date but no due date by Zip code

Metric-3: *Calculate the percentage of wrong dates in Closing date column by complaint type*

This metric explains the validity of the Closed date column to see if it has the valid dates which are greater than the Created date. Because, in real time scenarios of NYC311, a complaint or request cannot be closed before it got created or existed. Table3 shows the four complaint types that contain more than 5% of invalid closed dates or a closed date which is smaller than the created date. The invalid dates in the Closed date column defeats the purpose of that attribute in the data as well as will mislead the results if for example NYC311 wants to know what the average time frame is taken to resolve complaints for each complaint type.

Complaint Type	Total_entries	Percentage_of_wrongEntries
Smoking	3009	11.35
Non-Residential Heat	1929	13.86
Street Light Condition	257581	24.69
FATF	157	31.72

Table 3: Percentage of wrong dates entered in closing Date by complaint type

Metric-4: *Number of Zip codes that are out of range compared to Zip codes in reference data*

This metric shows how many Zip codes are out of range in the service requests data when compared to the Zip codes developed by US postal services for New York city. The valid Zip codes for New York city ranges from 10001 till 11697 but the service request dataset contains numerous records with an invalid or incorrect zip code which again reduces the dependability of the data when producing results based on Zip codes and makes the data inaccurate. Table 4 shows the total of records present that are either below or above the valid Zip code range. It seems out of 27,736,190 records in the dataset, 26289624 records contain valid zip code whereas 1440952 records are below the range and 4703 records are above the valid zip code range.

below_ZipCode_range	valid_ZipCode	above_ZipCode_range
1440952	26289624	4703

Table 4: Out of range Zip codes compared to reference data Zip codes

Metric-5: *Percentage of Zip codes present in the data that are outside the New York region.*

This metric evaluates the accuracy of the NYC311 service requests data in terms of the Incident ZIP attribute containing correct zip codes that belongs to New York city. It has been found that 94.29% of the data contains valid Zip codes for each service request whereas only 5.71% of the records have Zip codes that are not of New York when compared to the reference data. Though this shows the service request table has 94.29% accuracy in terms of Zip code data, it will still impact on any analysis carried out for costs associated decision making.

Metric-6: *Borough names that are incorrectly entered/geo-validated against the Zip codes*

This metric examines the accuracy of the Borough attribute to see if a borough name is entered or otherwise geo-validated correctly against a particular Zip code that are present in the reference borough names data. It seems Bronx, Queens, Manhattan, and Brooklyn boroughs are being confused with each other and being incorrectly entered. Also, some of the records contain value as 'Unspecified' which indicates either the users did not report the borough name, or it was not validated correctly when entering the system. A total of 8.87% of records of wrong borough names attached against to their respective Zip codes.

Incident Zip	Sr_borough	Zip	ref_borough
11210	bronx	11210	brooklyn
11106	bronx	11106	queens
10466	bronx	10466	bronx
11226	bronx	11226	brooklyn
11212	bronx	11212	brooklyn
11213	bronx	11213	brooklyn
10019	bronx	10019	manhattan
10463	bronx	10463	bronx
10456	bronx	10456	bronx

Table 5: Wrong borough names entered against the Zip codes compared to reference data

Metric-7: *Percentage of complaint types that are not present in the 26 complaint types mentioned in reference data*

This metric evaluates what percentage of complaint types are present in the service requests data that are not part of the 26 complaint types that exist in reference data. It is surprising to see 40% of the complaint types entered in service request data are not part of the reference data. This leads to inaccuracy in results produced by complaints type data. Also, if the government wants to analyse the data using the predefined 26 complaint categories, they would only be able to use 60% of the correct complaint type data or otherwise the analysis will make no sense and will result in wrong decision making.

Metric-8: *The percentage of data that has no NULL values in the selected columns (Created Date, Closed Date, Due date, Complaint Type, Borough)*

This metric evaluates the completeness of the records present in the data in terms of the chosen six columns for this report. Only 30% of the records are found to be complete with no null values whereas the rest of the records have at least one null value in one of these six columns. This incompleteness of data reduces the statistical power and representativeness of data sample taken for analysis.

Conclusion

Finally, the eight metrics developed in this report indicates that the service requests data contains lot of inaccurate data in the attributes representing the Zip codes and complaint types. It has been found that the data lacks completeness for most of the The Created date and closed dates are inconsistent as some of the closed dates do not have a respective due date. Some closing dates are inaccurate as they are smaller than the Created dates. There are numerous records that do not follow the metadata rules, for example out of range zip codes and self-created complaint types. When automated these metrics in the database, NYC311 can consistently monitor their data quality and take appropriate actions to enhance it.

References

Gupta, A. (2021, April 06). *The 6 dimensions of data quality.* Retrieved from Collibra: https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality

NYC311. (n.d.). Retrieved from NYC311: https://portal.311.nyc.gov/article/?kanumber=KA-02498