Data Quality Report

Dataset: Surat Public Transit Bus Realtime Info

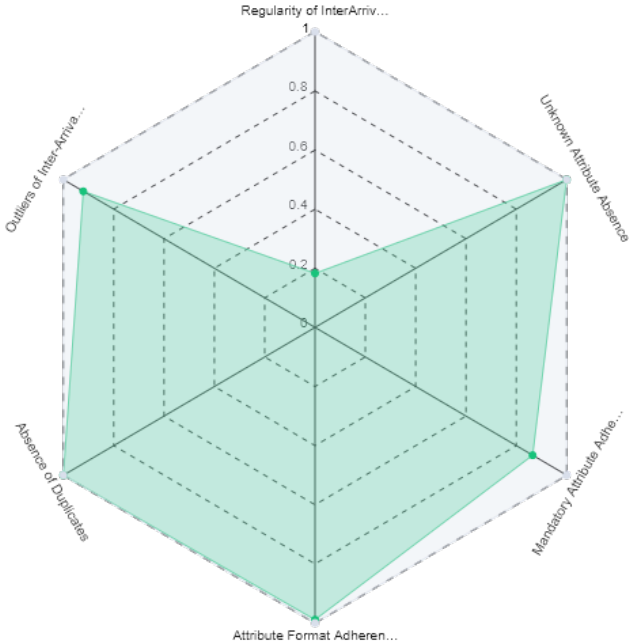
Number of Data Packets: 5000 | Start Time: 2022-01-01 10:10:35 | End Time: 2022-01-01 10:41:03

Overview

Metric	Score	Bar
Inter-Arrival Time Regularity	0.183	0.183 / 0.817
Inter-Arrival Time Outliers	0.921	0.921 / 0.079
Duplicate Presence	1.0	1 / 0
Adherence to Attribute Format	0.865	0.865 / 0.135
Absence of Unknown Attributes	1.0	1 / 0
Adherence to Mandatory Attributes	0.989	0.989 / 0.011

The Overall Data Quality Score of the dataset, computed by calculating an average of the above scores is:

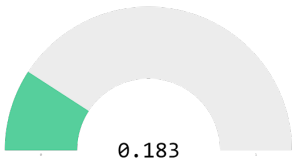
0.826/1.00 or 82.6%



This data quality assessment report shows the score for six metrics that contribute to data quality.

The chart on the right shows an overview of the data quality of the dataset.

In the following pages you can find a detailed description and breakdown of each of these metrics.



Inter-Arrival Time Regularity

Inter-arrival time is defined as the time elapsed after the receipt of a data packet and until the receipt of the next packet. For sensor data, this is an important factor to evaluate as sensors are often configured to send data at specific time intervals.

In order to compute this metric we analyse the deviation of each inter-arrival time from the mode. The assumption here is if most of the sensors are operating nominally most of the time, then the mode of the inter-arrival times will represent the expected nominal behaviour of the sensors. To compute this deviation, we define:

$$Relative\ Absolute\ Error\ (RAE) = \frac{\left|x_i - \bar{x}\right|}{\bar{x}}$$

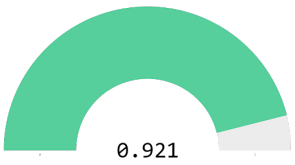
Here, xi is the inter-arrival time, and x is the mode of the inter-arrival time. We consider an RAE value of 0.5 to be the crossover point between good and poor values of inter-arrival time, i.e. RAE > 0.5 is poor. We also want to penalise the score proportionately to the RAE value, meaning the greater the RAE value, the greater the penalty. RAE is thus bound as RAE belongs to [0, inf).
The metric computation can also be represented as an equation:

$$M(x) = \frac{\sum_{x_i,\, RAE_i \leq 0.5} (1 - 2RAE_i)}{\sum_{x_i,\, RAE_i \leq 0.5} 1 + \sum_{x_i,\, RAE_i > 0.5} (2RAE_i)}$$

This represents the "badness" of the inter-arrival time when compared to the modal value. The further the inter-arrival time is from the mode, the greater the penalty contribution to the regularity score for that inter-arrival time. A value of 0.5 for RAE is chosen as the crossover point between "goodness" and "badness" of inter-arrival time as it represents a window of values corresponding to:

$$\bar{x} \pm \bar{x}/2$$

A higher IAT Regularity score indicates lower dispersion of IAT values around the mode, and vice versa. A higher score indicates that there is a higher clustering of IAT values close to the mode of the sensor. This regularity is particularly important for time-critical applications where a consistent and predictable arrival pattern is desired. By evaluating the IAT Regularity metric, researchers can gain insights into the reliability and efficiency of the data transmission process in IoT networks, contributing to the optimization of various IoT applications and services.



Inter-Arrival Time Outliers

The outlier metric of the inter-arrival time is an evaluation of the number of IAT values that show a significant deviation from the expected behaviour.

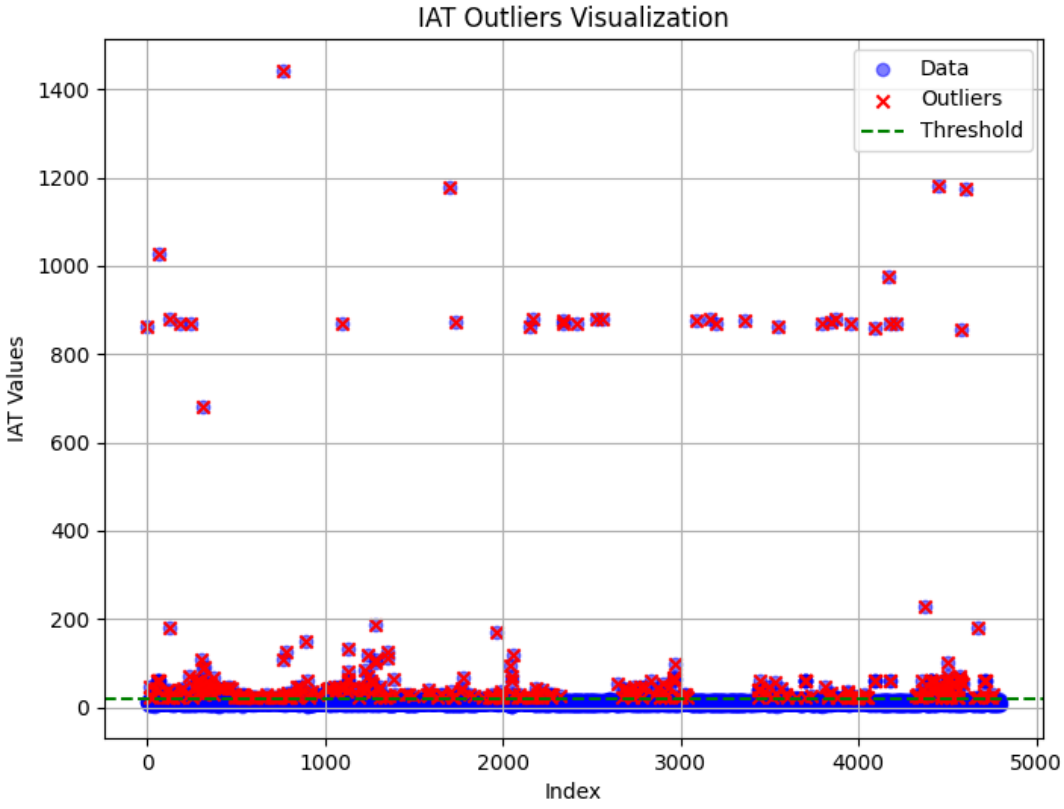
There are multiple ways to identify outliers in a dataset, and the choice of method is dependent on the independent characteristics of the dataset. In our case, we apply the modified z-score method proposed by Iglewiscz and Hoaglin.
Let the Median Absolute Deviation of the data be defined as:

$$m(x) = median_i\{|x_i - \bar{x}|\}$$

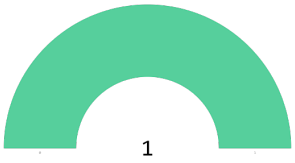
where xi is the observation for which the MAD is being computed and x is the mode of the data. We use the mode in place of the median as used by Iglewiscz and Hoaglin because we want to evaluate the deviation of the inter-arrival times from the mode, and we consider the mode to represent the expected behaviour of the dataset. Then the modified Z-score Mi is:

$$M_i = \frac{0.6745(x_i - \bar{x})}{m(x)}$$

Here, Iglewiscz and Hoaglin suggest that observations with |Mi| > 3.5 be classified as outliers, with variations to this cut-off value depending on the distribution of x. For our purposes, we will use this value to label inter-arrival time values as outliers. The outliers for this dataset are shown in the plot below.



Duplicate Detection



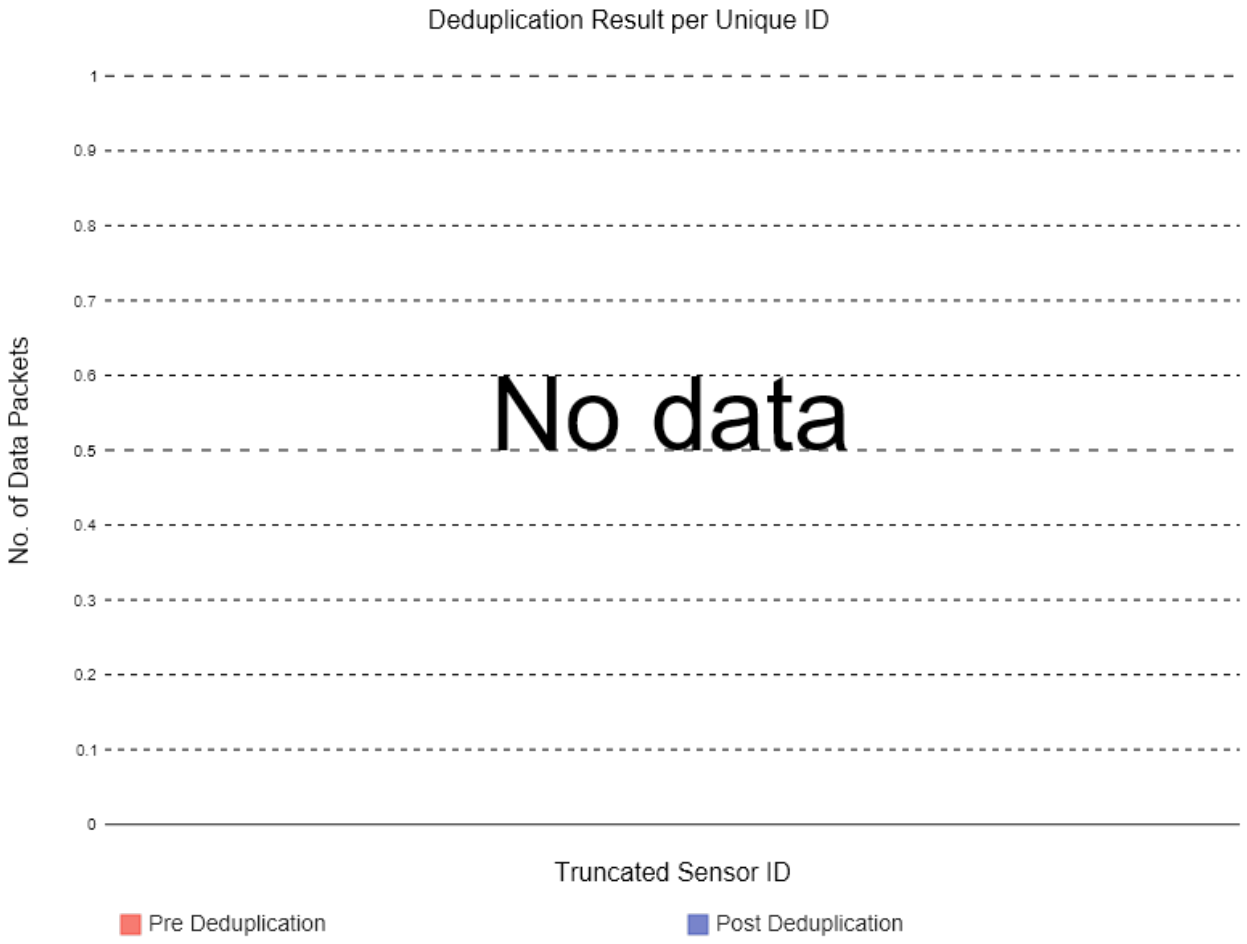
This metric conveys how many duplicate data points are present in the dataset.

The duplicates in a dataset are identified using the timestamp and any one unique identifier for each data packet. For example: AQM Sensor ID, Vehicle ID, etc. may be used as unique identifiers for a dataset.
If any unique identifier sends two data packets with the same timestamp, then one of the two data packets is counted as a duplicate. This is because it is assumed that any one device or sensor may not send two data packets with a single timestamp.

For this dataset, the attributes chosen for deduplication are:

trip_id
observationDateTime

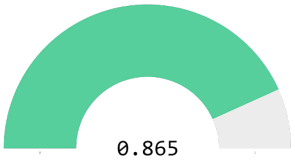
Using these attributes, 0 duplicate data packets have been identified in the dataset.
This metric is calculated on a score from 0 to 1, where a score of 0 indicates that all the data packets are duplicates and a score of 1 indicates that none of the data packets are duplicates.
The chart below shows the number of data packets before and after deduplication on a per unique ID basis. If a unique ID is not represented in the chart, it means that there were no duplicate values received from that unique ID.



Metrics for Schema Analysis

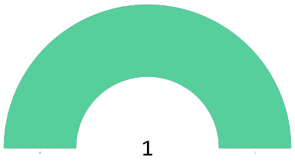
The remaining three metrics are an analysis of the metadata that is provided along with the dataset. This metadata is provided in the form of a schema, a document that delineates the different types of attributes, the data types of each attribute (integer, float, string, etc.) as well as the range of the observations under each attribute. This document also provides the mandatory attributes that the dataset must contain, as well as a list of all the expected attributes in the dataset.

Attribute Format Adherence



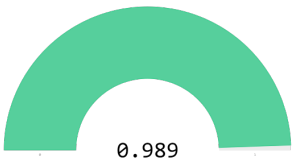
The attribute format metric checks whether the format of the data packets being evaluated matches the format defined in the data schema. The various possible formats include number, string, float, and object.
The format adherence metric is computed using the json schema validation method. The count of errors is incremented when the data type of an evaluated data packet does not match the data type specified in the data schema.
A higher score for the attribute format metric indicates a relatively lower proportion of data packets that contain attributes that do not adhere to the format defined in the schema, and a lower score for the attribute format metric indicates a relatively greater proportion of data packets with incorrect attribute formats.

Absence of Unknown Attributes



The unknown attributes A higher score for the attribute format metric indicates a relatively lower proportion of data packets that contain attributes that do not adhere to the format defined in the schema, and a lower score for the attribute format metric indicates a relatively greater proportion of data packets with incorrect attribute formats.metric computes the number of data packets with attributes that are present in the dataset but are not specified in the schema in any capacity.
This metric is computed by validating the data against the schema. A higher score for this metric indicates a relatively lower proportion of data packets that contain attributes that are not present in the data schema and a lower score indicates a relatively greater proportion of data packets with unknown attributes.
This metric represents the total number of unknown attributes in the dataset.

Adherence to Mandatory Attributes



The mandatory attributes metric checks whether the list of mandatory attributes defined in the data schema are all present in the dataset. This validation is performed for each data packet in the dataset.
A higher score for the mandatory attributes metric indicates that there is a relatively greater proportion of data packets with values present for all the mandatory attributes, and a lower score for the mandatory attributes metric indicates that there is a relatively lower proportion.
This metric is an indicator of the completeness of the dataset. Null values received under mandatory attributes are also included in the count of the number of missing attributes.