Data Quality Report

Dataset: Pune Air Quality Monitoring Sensors
Start Time: 2022-01-01 00:01:08
End Time: 2022-02-28 23:46:39
Number of Data Packets: 212475

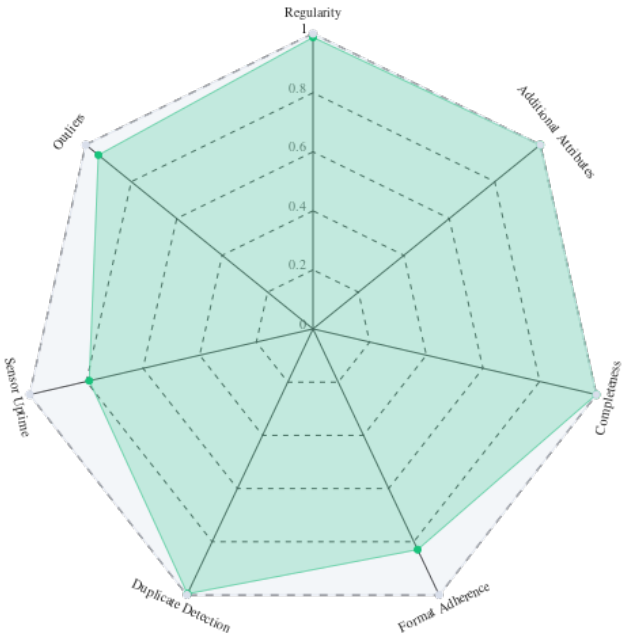
Overview

Metric	Score	Bar
Regularity of Inter-Arrival Time	0.988	0.988 / 0.012
Outlier Presence in Inter-Arrival Time	0.945	0.945 / 0.055
Sensor Uptime	0.79	0.79 / 0.21
Absence of Duplicate Values	0.996	0.996 / 0.004
Adherence to Attribute Format	1	1 / 0
Absence of Unknown Attributes	1	1 / 0
Adherence to Mandatory Attributes	0.83	0.83 / 0.17

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

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

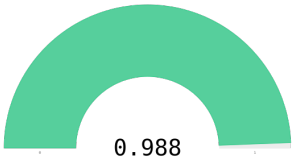
In the following pages you can find a detailed description and breakdown of each of these metrics.



Inter-Arrival Time (IAT)

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 this section, we will be analysing the regularity, outliers, and anomalous values of the inter-arrival times of this dataset.

IAT Regularity



The regularity metric of the inter-arrival time conveys how uniform this time interval is for a dataset in relation to the expected behaviour.
Considering the mode of the inter-arrival times to be the expected value (as per the specification), this metric measures the proximity of the spread of the normal distribution to the mode. The spread of the inter-arrival time values from the mode is computed using the formula:

$$Spread = \mu \pm \alpha \times \mu$$

where alpha is a constant from 0 to 1. In this case, 3 values have been considered: 0.25, 0.5, 0.75.

Considering the minimum and maximum values of this formula to be the lower and upper bounds, we compute the number of inter-arrival time values outside these bounds and divide by the total number of data packets using this formula:

$$IAT\ Regularity = 1 - \left(\frac{No.\ of\ data\ packets\ outside\ the\ bounds}{Total\ no.\ of\ packets}\right)$$

This value is computed for each alpha, and then averaged to give the overall metric score. The score is on a scale from 0 to 1, where 1 indicates the highest possible proximity to the mode and 0 indicates the opposite. The average of these three scores are taken to form the overall metric score.

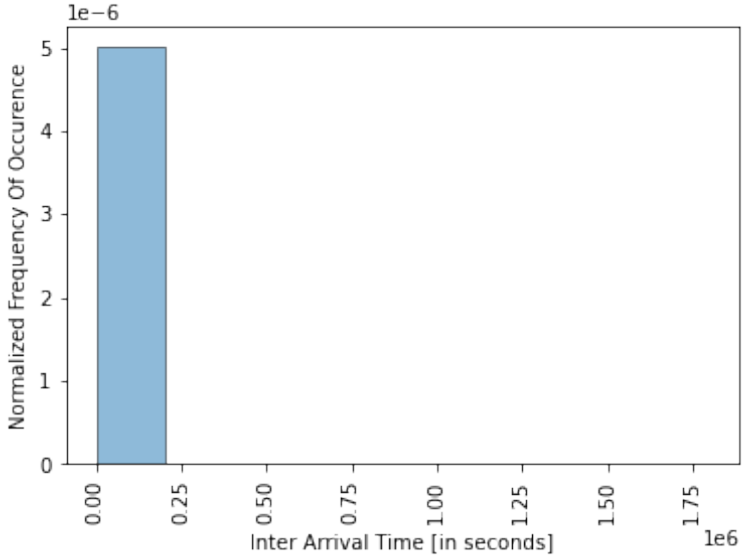
Alpha	$\mu - \alpha \times \mu$	$\mu + \alpha \times \mu$	**Regularity Score**
0.25	675.0	1125.0	0.987662
0.5	450.0	1350.0	0.987993
0.75	225.0	1575.0	0.988017

A high score for the inter-arrival time metric means that the data packets are received at regular intervals which is important for time-critical applications.

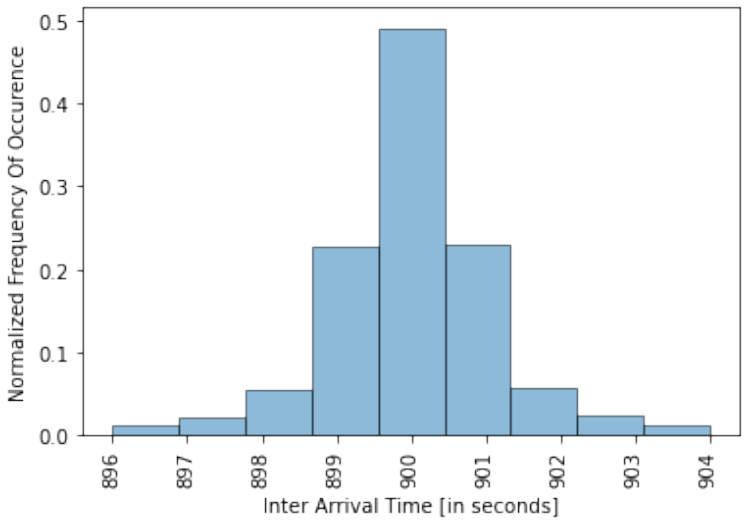
The table below shows a comparison of the statistics of the inter-arrival times of the dataset before and after outlier treatment using the Inter-Quartile Range method.

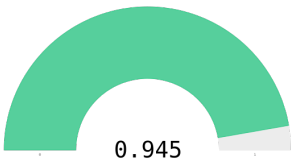
	Before Outlier Removal	After Removal of Outliers
Mean	1138	900
Median	900	900
Mode	900	900
Standard Deviation	8604	1
Variance	74035475	1
Skewness	112	0

The histogram of the inter-arrival times of the dataset prior to removal of outliers is on the right.



The histogram of the inter-arrival times of the dataset after removal of outliers using the inter-quartile range method is on the right.





IAT Outliers

The outliers of the inter-arrival time is defined as the number of data packets which are received outside the bounds specified by the inter-quartile method.

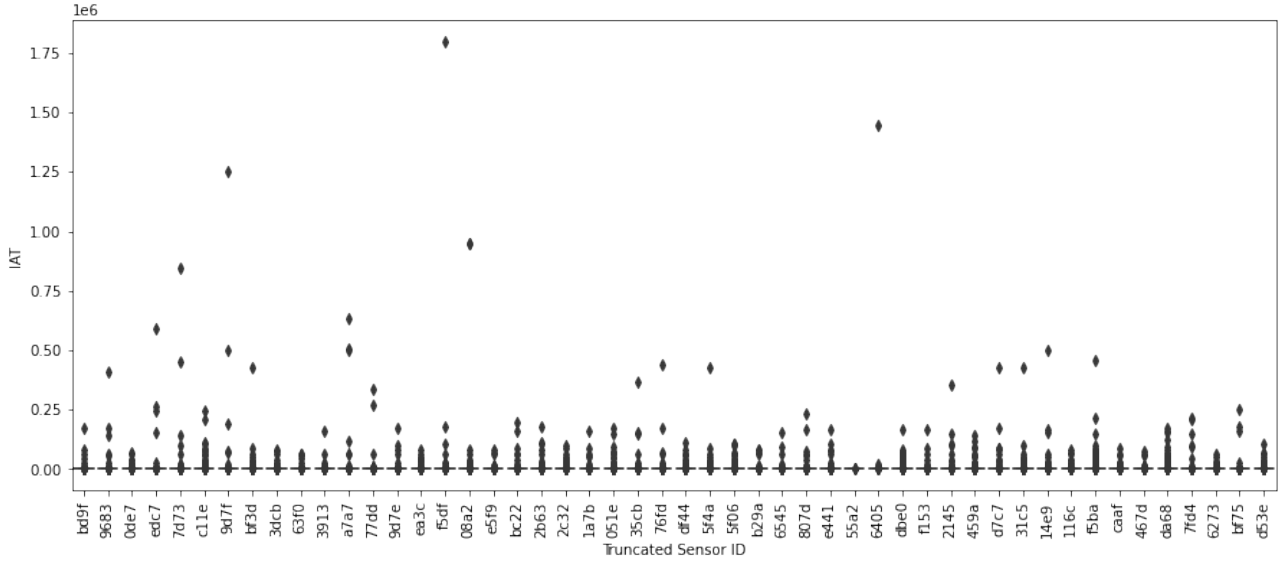
The Inter-Quartile Range of a dataset is defined by dividing the dataset into quartiles and selecting the middle two quartiles (50%) of values when ordered from lowest to highest.

Quartiles are three percentiles that separate an ordered dataset into four parts. In this case, our quartiles are: 25, 50, 75

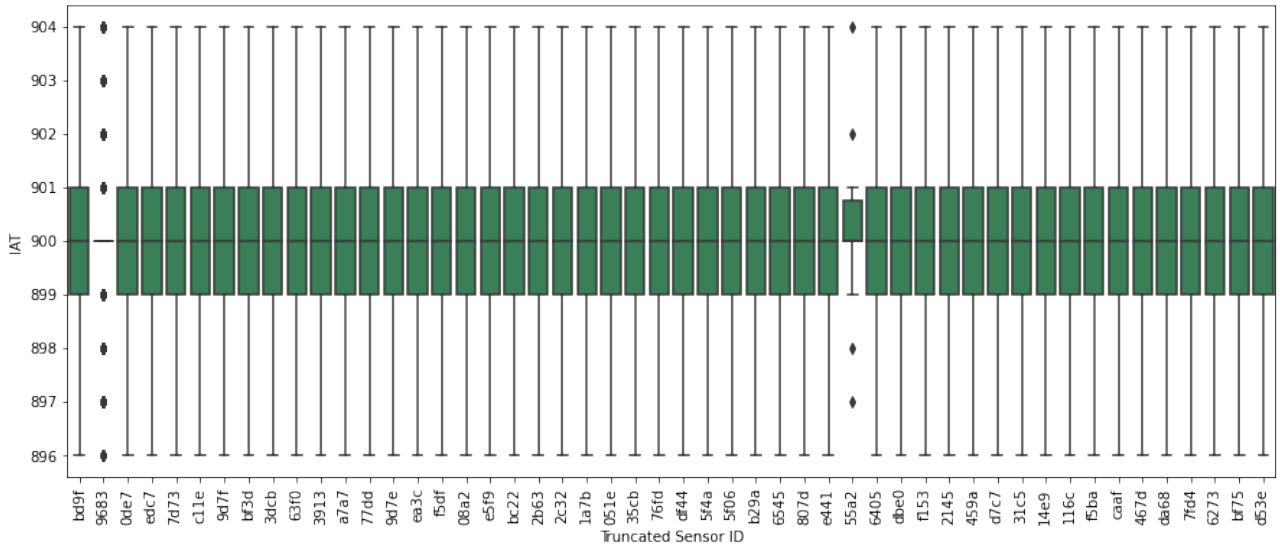
The metric score is computed as below:

$$IAT\ Outliers\ =\ 1\ -\ (\frac{No.\ of\ outliers}{No.\ of\ data\ packets})$$

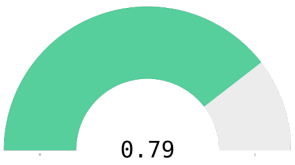
This score is computed on a scale from 0 to 1, with 0 being the lowest possible score, indicating that there are no data packets within the inter-quartile range, and 1 being the highest possible score indicating that all the data packets are within the inter-quartile range.
Before applying the interquartile method, a boxplot of the dataset is given below:



After the interquartile method is applied to remove the outliers, a boxplot of the dataset is given below:
The value of the lower bound is: 896.0
The value of the upper bound is: 904.0



Sensor Uptime



Sensor uptime is defined as the duration in which the sensor is actively sending data packets at the expected time intervals.

This metric is calculated by performing an analysis of the inter-arrival time of the sensors. Each value of the inter-arrival time that is greater than twice the mean is selected and sorted by sensor. These values are then summed for each sensor and an overall average is taken. This overall average value is then divided by the total query time of the dataset.
Total query time is the time for which the dataset is queried, i.e. the difference between the timestamps of the first and last data packets in the dataset.

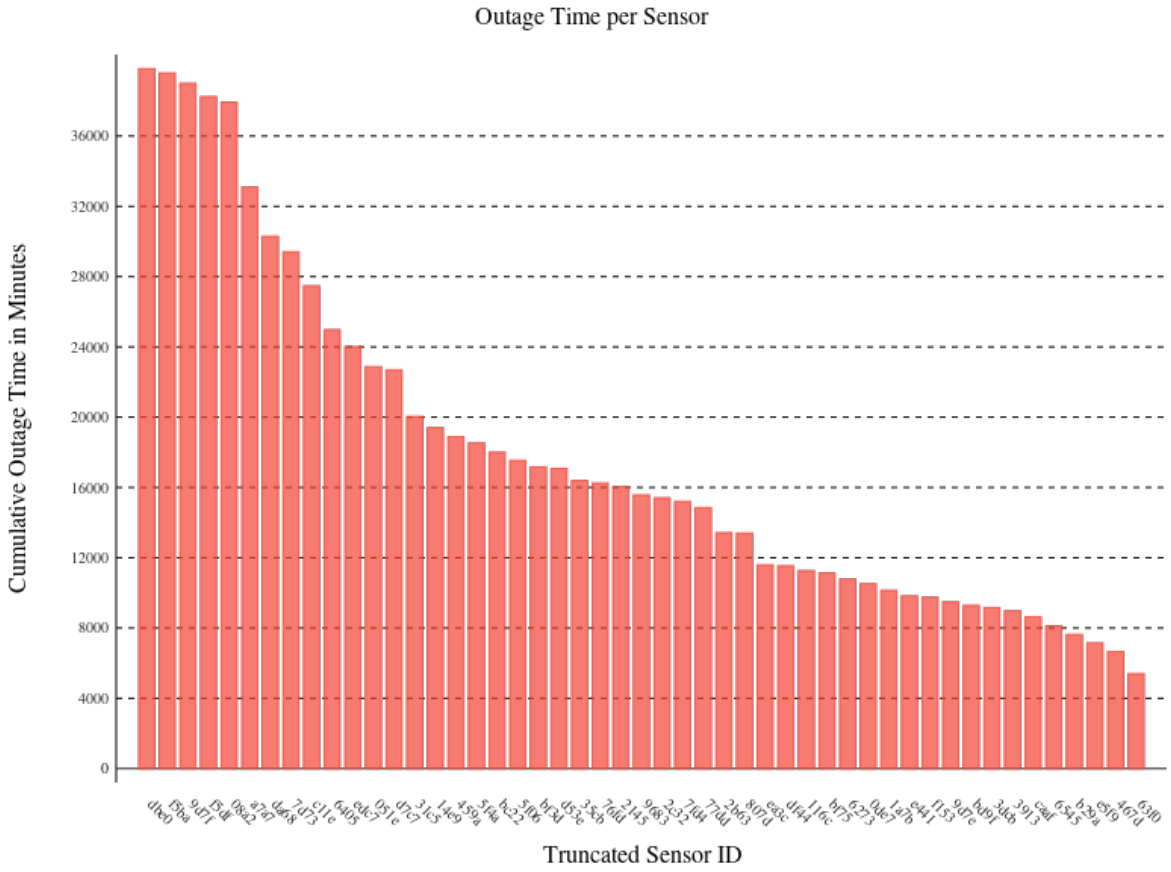
The metric score is computed as below:

$$Sensor\ Uptime\ =\ 1\ -\ \left(\frac{avg.\ Outage\ Time\ per\ Sensor}{Total\ Query\ Time}\right)$$

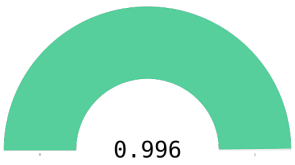
Assuming that a high value for the inter-arrival time means that the sensor is not sending data packets at the expected intervals and is assumed to be "down". Sensor uptime can be understood as the time during which the sensor is not undergoing an outage and is functioning as expected.

The metric is calculated on a scale from 0 to 1, with 0 being the lowest score indicating that there is a high degree of sensor outage in the dataset, and 1 being the highest score indicating that there are no inter-arrival times greater than twice the mean.

The chart below shows the outages in the dataset on a "per sensor" basis.



Duplicate Detection



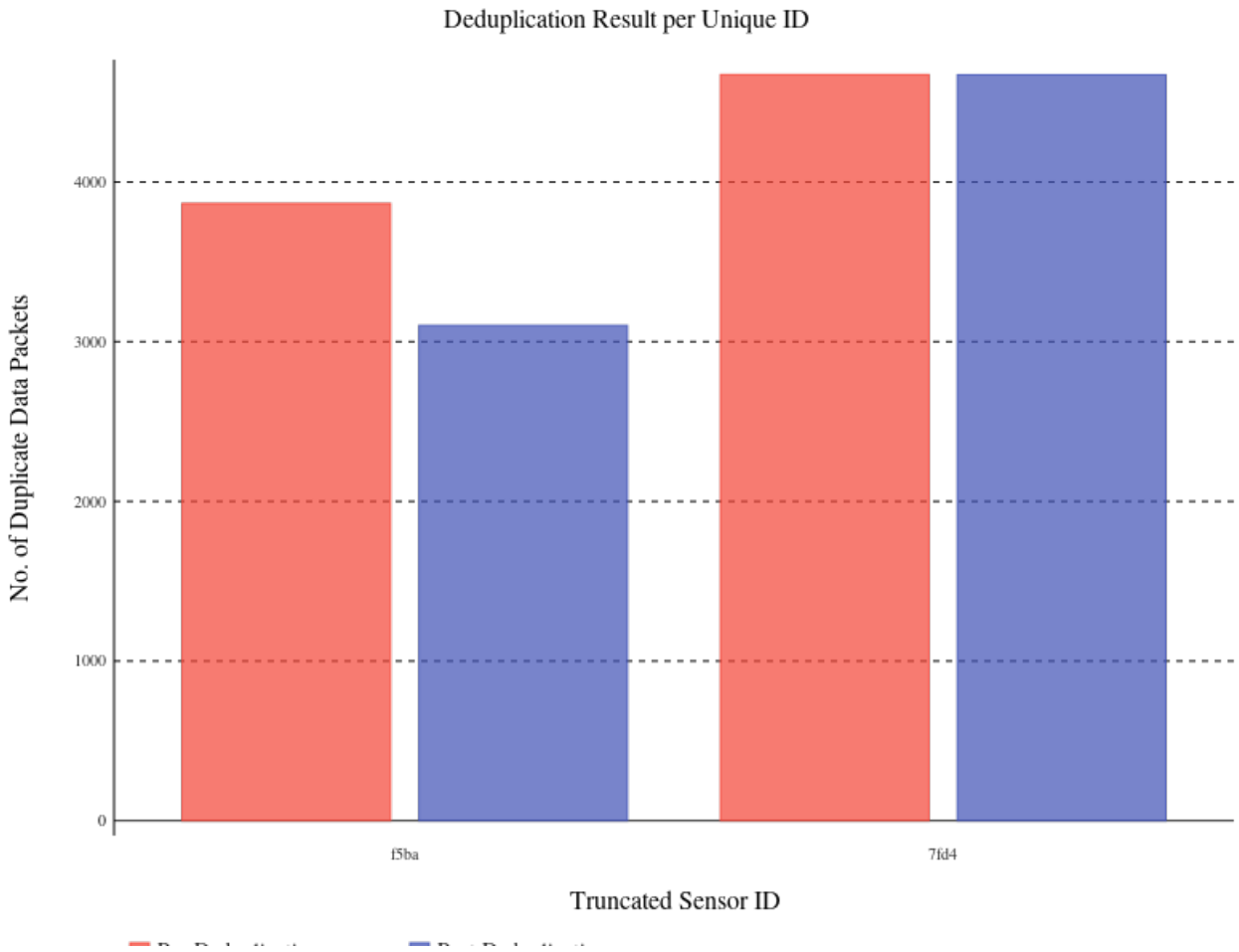
This metric conveys how many duplicate data points are present in the dataset.

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 sensor may not send two data packets with a single timestamp.

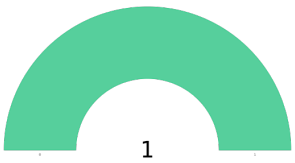
For this dataset, the attributes chosen for deduplication are:

id
observationDateTime

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.



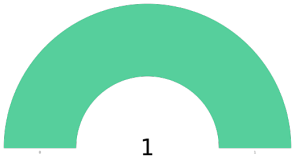
Attribute Format Adherence



This metric assesses the level of adherence of the data to its expected format as defined in the data schema. It is quantified by taking the ratio of packets adhering to the expected schemas to the total number of data packets.

Represents the completeness of the attributes of the dataset.

Absence of Unknown Attributes

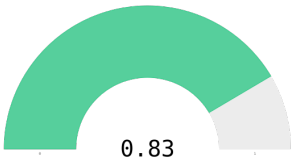


This metric checks whether there are any additional attributes present in the dataset apart from the list of required attributes.

This metric is computed as (1 - r) where r is the ratio of packets with unknown fields (fields that are not present in the list of mandatory attributes) to the total number of packets.

This metric represents the total number of unknown attributes in the dataset.

Adherence to Mandatory Attributes



This metric checks whether all the required attributes defined in the schema are present in the dataset.

It is computed as follows: For each mandatory attribute, i, compute r(i) as the ratio of packets in which attribute i is missing. Then output 1 - average(r(i)) where the average is taken over all mandatory attributes.

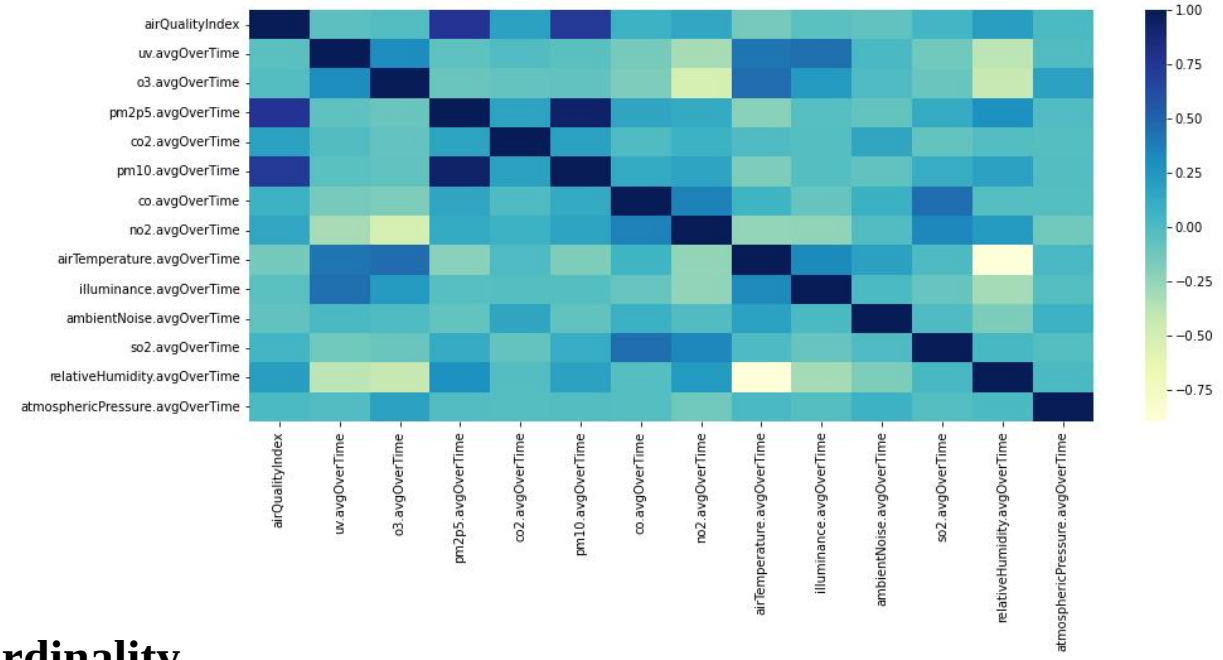
The metric is computed on a scale from 0 to 1, where a score of 0 indicates that all the data packets in the dataset are missing the required attributes, and 1 indicating that all the data packets are adherent to the list of required attributes. The metric represents the completeness of the attributes of the dataset.

Additional Information about the Data

In this section are some useful visualizations that describe certain data statistics that can be used by the end user to determine the usability of the data. These subheadings may not explicitly fall under the umbrella of data quality and so are not counted as part of the overall score.

Correlation

Correlation here refers to a causal relationship between different attributes found in the dataset. This relationship might be either directly or inversely proportional.
This relationship is shown in the heat map below, with darker colors referring to a stronger direct relationship, and lighter colors referring to a stronger inverse relationship.



Cardinality

Cardinality of a dataset is defined here as the number of unique values of in that dataset. A higher value of cardinality indicates a higher proportion of unique values.