**Missing Value Plot:**

* Missing values are present in the dataset.
* The missing values are shown using missing distribution plot.
* We can see that the columns Date, Author, ID and Y does not have missing data whereas the column sensor3 has a lot of missing data. This tells that the data is not homogenous. Hence, we can say that the process used for collecting data might not be same.
* In the missing value plot, we can also see that at lower end that is till 45 there are no missing values for all the column except for sensor3.
* Missing values are represented as “NA”.

**About the Data:**

* The summary statistics tab gives us idea of the mean and median of the quantitative columns of the dataset.
* In the Data Visualization tab, we can see that the columns Y and sensor1 to sensor30 are all numeric data whereas the below fields are Factor.
* ID : Factor w/ 300 levels "D254","D255",..: 254 255 256 257 258 259 260 261 262 263 ...
* $ Author : Factor w/ 4 levels "KL","NW","PR",..: 2 2 2 2 2 2 2 3 3 2 ...
* $ Date : Factor w/ 300 levels "2001-01-07","2001-01-13",..: 1 2 3 4 5 6 7 8 9 10 ...
* $ Priority: Factor w/ 3 levels "High","Low","Medium": 3 1 1 2 2 2 2 3 2 2 ...
* $ Price : Factor w/ 3 levels "Cheap","Costly",..: 3 3 3 1 3 3 1 2 3 3 ...
* $ Speed : Factor w/ 3 levels "Fast","Medium",..: 2 1 2 2 2 1 1 2 1 1 ...
* $ Duration: Factor w/ 3 levels "Long","Short",..: 2 2 2 3 3 3 2 2 1 3 ...
* $ Scarcity: Factor w/ 3 levels "Cold","Hot","Warm": 2 3 3 3 2 2 2 2 2 2 ...
* $ Location: Factor w/ 3 levels "Boiler","Plant A",..: 3 2 2 1 2 3 3 2 1 3 ...
* $ Agreed : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 2 2 2 1 2 ...
* In general, the columns with more than 55 levels crash while modelling and hence in this case columns ID and Date needs to be handled with care.
* Few of the sensors are highly correlated such as sensor 8 and sensor 1 showing linear dependencies. Sensor2 and sensor5 are also highly correlated. Sensor8 and sensor5 are non- correlated.
* There are no columns in the dataset which are completely whole numbers. There are no predictors with integer value or counts.

**Data Roles:**

* In this dataset we have not been told about any of the columns hence we cannot assign the predictors or response variables.
* On looking at the plot data visualization tab, we can view the factor variables. The variables mentioned as factor has levels less than 10 except for the two variables that are ID and Date. The ID has 300 levels which can be considered as unique identifiers or can be assigned as observation identifiers. The column Date is shown as factor but it is not a factor as it has 300 levels.
* Column Y can be assigned the role of outcome that is the Y variable.
* Sensors can be assigned the data role of predictors that is X variable.
* Columns such as Priority, Price, Speed, Duration, Scarcity are qualitative data which are ordered.
* The column ID is not quantitative and are nominal data. It is 100% unique hence cannot be considered as predictor as it will overfit the training set and will behave inadequately when tested. The column ID is represented with some code at the beginning along with increasing number at the end.
* The column Date can be considered as grouping variable as its levels are not fixed and can keep increasing.
* For few categorical columns we might not be sure if the levels in the dataset are the only levels. In this dataset the columns such as State and Location might have more levels than mentioned for which we can add one more level that is “others”. We may have more levels for these columns if we test our model in some new set of data.
* None of the columns are constant or with zero variance. Constant columns need to be discarded when doing linear modelling.

**Collinearity:**

* Few columns are highly correlated which causes problem of multicollinearity when applying linear regression model to it. The collinearity of 0.99 has been seen between few data such as sensor2 and 13, sensor2 and 5 which needs to be taken into consideration when applying machine learning models to the data.

**Novelties:**

* Sensors 2, 5, 13, 15, 21, 26, 27 have outliers.
* In the continuity plot for the numeric data we can see that for the sensors 2,5,13,15,21,26 and 27 plots are not smooth, it is abrupt at the end with sudden rise in values.
* The column ID is a factor yet, it is continuous as it has as many levels as number of observations.
* From the correlation plot for different numeric values we can see that, few numeric columns are highly correlated with each other which can cause multicollinearity in regression analysis.

We can see univariate outliers in the Box plots for quantitative data 🡪 Some sensors such as 2, 5, 13, 15, 21, 26 and 27 shows outliers even when we increase the IQR range to 5. Thus, showing novelties in the dataset. The novelties need to be checked further and consulted with business to know more about the dataset. However, as on keeping the IQR value as 5, we are still not able to remove he outliers, this shows something unusual about the data.

**Continuity**:

The continuity plot shows that the sensors 2,3,5,13,15,21,26 and 27 are not continuous. There is a steep change in the continuity plot for sensor 3, thus we can say there is discontinuity in the plot for this column. Due to the sudden increase of the value for sensors 2,5,13,15,21,26 and 27 we can observe discontinuity in the dataset.

**Homogeneity**:

* On looking at the homogeneity graph we can see that the dataset is not homogenous in nature. The observations of sensor 2,5,13,15,21,26 and 27 does not follow the same pattern as rest of the columns when seen from left to right. Hence, we come across the unusual behaviour of the dataset.
* On plotting the line graph of numeric columns against the column Date we can see that for the sensors 2,5,13,15,21,26 and 27 we see abrupt rise in their values at the end of the year 2005. The line graph for these sensors are following different pattern from the rest of the numeric columns. The values of these sensors are close to 0 till the end of 2005 and there is abrupt rise after that. Observations of rest of the numeric columns are evenly distributed throughout the year 2001 to 2007.

**Mosaic Plot:**

* The mosaic plot for the categorical data shows that the red represents the rare set of data.
* In this plot we can see that the data with textured **surface**, Plant A **Location**, wooden **class** and medium **speed** are very rare as it is marked with red and is unusually rare.

**Centre and scaling do not work on non-numeric data. It works on numeric data.**

We cannot say that the data provided to us is structured as we can see lot of missing values in it. However, we can call **it semi-structured data**. It lies somewhere between the structured and unstructured data as it is not organized in a complex manner which makes sophisticated access and analysis possible; however, it provides us with information which allows elements contained to be addressed.