

Assignment 1: R shiny App for Data Exploration

Introduction

This analysis report describes the findings of a comprehensive examination on the given dataset. This analysis aims to uncover hidden patterns, trends, and relationships within the dataset, in a hope to shed light on the context of the dataset, and to be able to extract meaningful insights from the data. Through using R shiny, it is possible to present findings in an informative manner through the use of data tables, charts and plots. Here discussed is an overview and summary of the dataset, followed by an analysis of the variables and observations therein.

Visualisations

Summary

Through the initial summary page in the R shiny app, we are able to see the structure of the dataset. The dataset indicates there are a total of 350 observations and 44 distinct variables, of which one is a 'date' data type, 12 are factors (categorical variables) and 31 are numeric variables. Through this summary we can also see there is one 'Y' variable which we can presume to act as the response variable data type. This conclusion can be reached, firstly through the standard naming convention of a Y-variable being a response variable, and secondly as this is the only numeric variable with no missing observations. The other 30 numeric variables are labelled 'sensor1' through 'sensor30', indicating these numeric variables are taking some sort of measurement of equipment, or the equipment itself is sensing something. This indicates the data could be coming from an engineering workplace or other type of factory.

From the summary output we can also observe the relative distribution of each sensor variable in the histogram chart (figure 1). These small charts indicate the general distribution of each variable. Through this, we see that the sensors 2, 9, 12, 15, 21, 22 and 24 all have a (relatively extreme) bimodal distribution, while the other sensors are indicating a more normal distribution of data. These sensors will be investigated further in the accompanying charts and plots. Finally, from the summary output we can observe the missingness of each relative variable. The variables with no missingness are indicated to be date, ID, Operator, Priority, Agreed and 'Y'. A variable with distinctly high missingness is 'sensor3', with 103 missing data points. 'Sensor3' is also indicating an interesting bi-modal distribution, but less so than those previously mentioned.
































Variable type: numeric											
	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	Y	0	1	18.9	8.17	-1.64	13.3	19.5	24.9	40.2	
2	sensor1	10	0.971	19.5	12.6	-7.39	9.38	19.0	30.1	44.9	
3	sensor2	7	0.98	77.6	125.	-12.5	12.1	24.6	36.7	370.	
4	sensor3	103	0.706	20.2	14.9	-9.24	6.89	26.1	33.5	54.4	
5	sensor4	12	0.966	19.5	12.5	-7.89	9.16	18.9	29.5	49.8	
6	sensor5	7	0.98	20.2	12.8	-10.4	9.74	20.2	30.3	52.0	
7	sensor6	9	0.974	20.3	13.2	-7.67	9.55	19.9	30.9	50.7	
8	sensor7	12	0.966	20.1	12.3	-9.83	9.73	20.6	30.4	49.0	
9	sensor8	10	0.971	19.9	12.4	-7.90	10.3	21.0	29.7	50.7	
10	sensor9	12	0.966	78.1	126.	-8.34	12.4	24.0	36.0	368.	
11	sensor10	11	0.969	20.0	12.5	-12.4	9.76	20.9	30.2	44.0	
12	sensor11	15	0.957	64.0	21.1	15.8	46.3	64.5	78.7	121.	
13	sensor12	14	0.96	114.	110.	11.1	48.6	71.4	92.4	370.	
14	sensor13	9	0.974	62.6	21.1	9.99	46.9	61.3	78.0	115.	
15	sensor14	4	0.989	63.6	20.5	6.43	49.2	62.9	77.6	112.	
16	sensor15	13	0.963	114.	111.	10.7	48.8	71.0	91.5	369.	
17	sensor16	11	0.969	62.0	20.8	17.7	44.7	61.9	78.7	112.	
18	sensor17	14	0.96	63.4	21.1	14.2	46.6	61.8	78.9	117.	
19	sensor18	7	0.98	62.8	21.1	14.2	46.2	62.9	80.8	108.	
20	sensor19	11	0.969	61.5	20.4	15.6	47.7	62.7	76.7	122.	
21	sensor20	7	0.98	62.1	21.4	17.0	45.7	62.0	78.5	116.	
22	sensor21	11	0.969	79.4	132.	-62.9	-9.67	34.3	88.2	370.	
23	sensor22	9	0.974	77.7	132.	-70.2	-9.19	36.2	84.9	370.	
24	sensor23	13	0.963	21.5	45.8	-70.1	-17.6	19.7	62.2	126.	
25	sensor24	12	0.966	79.6	130.	-62.3	-10.8	37.7	86.0	370.	
26	sensor25	10	0.971	21.9	44.2	-67.0	-15.8	20.1	60.4	120.	
27	sensor26	9	0.974	21.4	45.7	-67.6	-14.8	20.9	60.2	126.	
28	sensor27	10	0.971	20.0	44.9	-73.1	-18.9	16.2	58.2	126.	
29	sensor28	13	0.963	22.3	45.0	-76.3	-12.1	20.8	59.6	115.	
30	sensor29	13	0.963	22.2	45.7	-67.2	-13.4	18.9	60.9	114.	
31	sensor30	15	0.957	20.7	45.2	-68.5	-18.1	18.9	55.5	124.	

Figure 1: Numeric variable summary via Skim. The histograms on the right-hand side illustrate the general distribution of each variable.

Pairs Plot

The Pairs Plot constructed using ggpairs in R Shiny allows for a quick overview of various variables in a matrix format. Through this, identification of which variables have strong relationships, potential outliers, and other interesting distribution characteristics can be observed. Through the summary output as described above, the variables of most interest are sensors 2, 9, 12, 15, 21, 22 (hereafter referred to as sensors of interest). These have been selected as the initial display on the Pairs Plot in R shiny, and are colour coded by relevant factor variables (figure 2). The factor variable which has the most interesting effect is that of 'Operator'. By selecting the 'Operator' as the group-by variable, we can observe a distinction between the various operators and the relationship this seems to have with the distribution of the sensors. Specifically, the scatterplot in the Pairs Plot (bottom left) is able to show a clustering of observations by the 'HY' operator category (in red colouring), while the other operators have clustered together separately from the 'HY' operator. This pattern is seen across all sensors of interest. The histogram in the Pairs Plot also supports this finding, with the distribution of each variable indicating that the bi-modal distribution seen earlier is driven by the operator split between 'HY' and the other operators. Both the scatterplot and histogram show that the 'HY' operator is displaying sensor observations on the high end of the scale, that is to say, the 'HY' operator observations tend to cluster around the 300 values, while the other operators are clustering to the lower end of around 0-100. The correlation

values in the Pairs Plot also indicate that there is strong correlation between each of these sensors of interest, however this trend is being driven predominantly by the operators 'JA', 'MD' and 'WP', while 'HY' is consistently not showing strong a correlation with the other sensors.

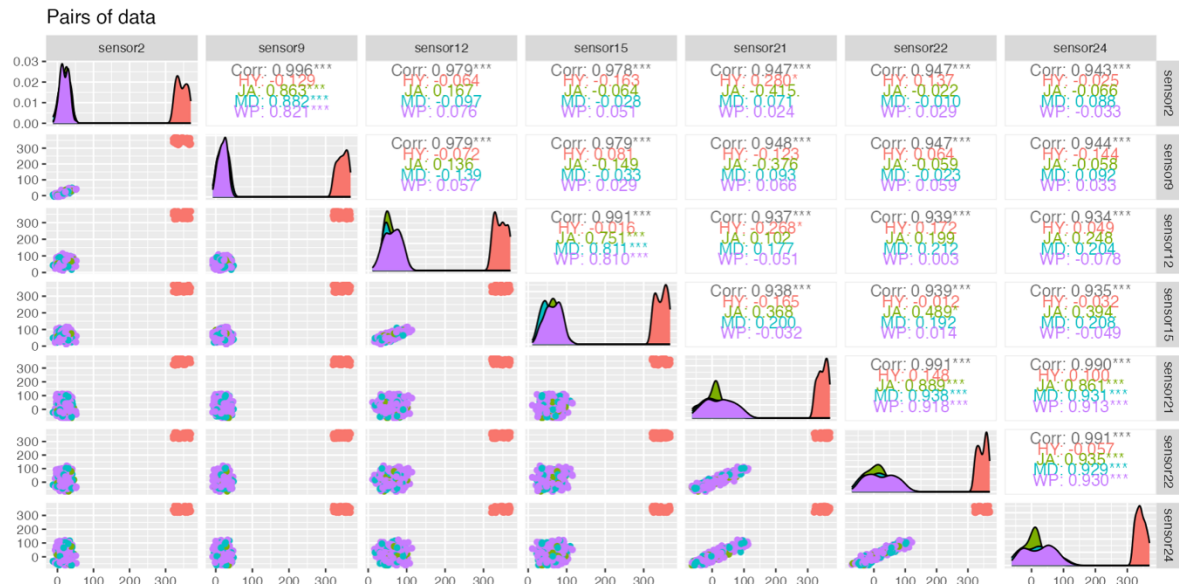


Figure 2: Pairs plot output of variables: Sensor2, sensor9, sensor12, sensor15, sensor21, sensor22, sensor24.

Corrgram

The Corrgram chart is able to display the correlation of the relevant numeric variables in the dataset. The Corrgram employs the use of Pearson correlation, Spearman correlation or Kendall correlation, with a choice of grouping via OLO, HC, GW, or no grouping. The choice of absolute correlation also allows for the user to visualise the strength of the correlation between the numeric variables regardless of whether is correlation is positive or negative. From the initial Corrgram with absolute Pearson's correlation and OLO grouping, we observe how the sensors of interest have all clustered together, indicating that these sensors are more closely correlated with each other than the other variables (figure 3). This strong grouping remains when HC or GW grouping is used. This same grouping is not as strongly observed when using Spearman's or Kendall's correlation method.

Correlation of data using pearson correlation method

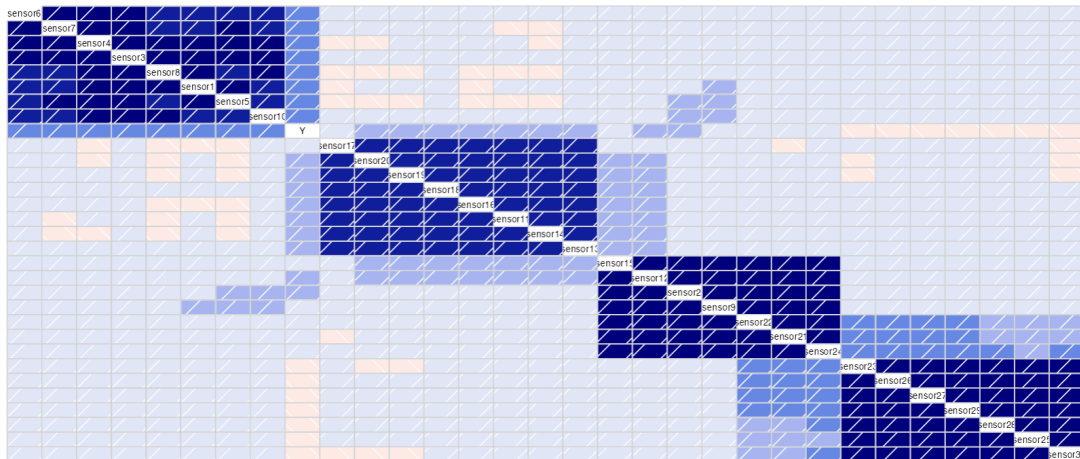


Figure 3: Corrgram of Data using Pearson's correlation method, OLO grouping (absolute correlation). Of particular interest is the group 3rd from the top.

Mosaic Chart

The Mosaic Chart is able to show relevant categorical variables of low cardinality. Consequently, variables 'ID' (and 'Date') have been excluded from this chart. The mosaic chart with the variables 'Operator', 'Priority' and 'Duration' show that observations in the 'WP' operator group with 'low' priority and 'very long' duration are moderately over-represented (moderately unexpectedly common). However, the mosaic chart is in this case not able to inform greatly on the categorical variables in this dataset beyond this slight observation, and does not provide great interesting insights.

BoxPlot

The BoxPlot is able to show the relative distributions of numeric variables in the dataset, with each variable's summary statistics and outliers able to be shown and compared across one plot. Initially, the plot shows the boxplots of each variable with an IQR multiplier of 1.5 (no outliers and no centring/scaling). It's possible to observe a general trend where by sensors 1 through 10 seem to group in similarity of distribution, sensors 11 through 20 group together, and sensors 21 through 30 group together through being more widely distributed than the other sensor groups. All these groupings seem to group on similarity of medians. Upon employing standardization and showing outliers in the dataset, some interesting distributions can be observed, where some sensors are indicating a high number of outliers in the dataset which is driving the distribution of these variables without standardisation. Upon further investigation, we can see that there are 7 of these variables which are being affected by strong outlier influence; these 7 variables are the same sensors of interest that have shown unusual distribution patterns in previous charts. This is illustrated in the image below (figure 4). Sensors 21, 22 and 24 lose their outliers when the IQR multiplier is ≥ 3.5 .

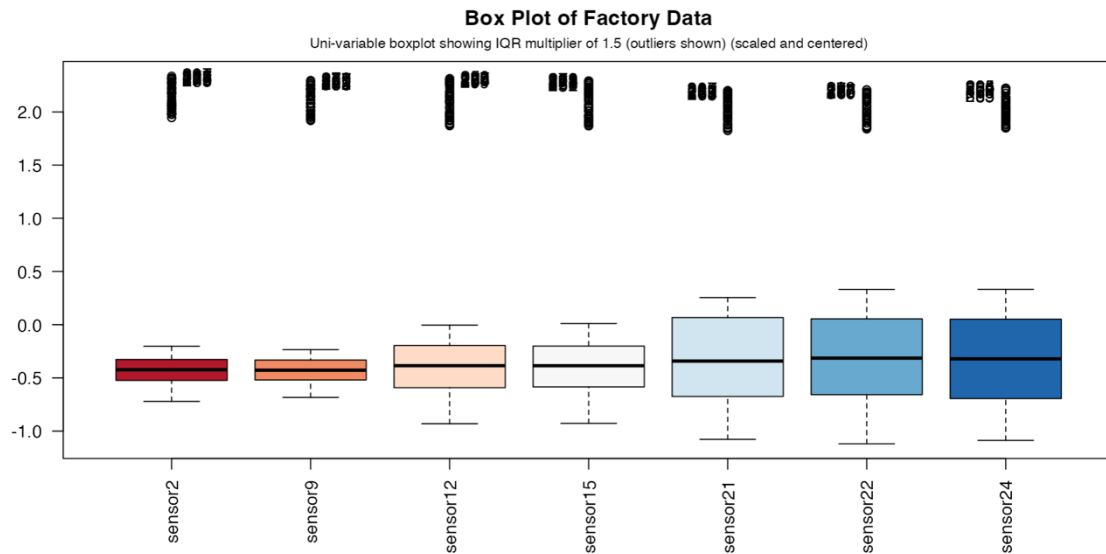


Figure 4: BoxPlot of Data. Sensors of interest shown with IQR multiplier of 1.5, Scaled/Centred with outliers showing.

TablePlot

The Tabplot::Tableplot plot in R Shiny helps with visualisation of the homogeneity of a dataset. From our previous graphs we have seen an interesting distribution in the sensors of interest, and there seems to be a relationship between these sensors of interest and the 'HY' operator group. To further investigate this, we can look at the Tableplot to see this relationship more clearly. As shown in the Tableplot when grouped by Date (figure 5), a trend is appearing between the operator HY (which is clearly grouped in blue), the IDs of some of the observations and the sensors of interest. Grouping by 'Date' variable has the same effect as no grouping, as the 'Date' has the same order as observation order, which is why no grouping is not available. When selecting more sensors, it's clear to see this trend is specific to those previously identified sensors. A similar trend is visible when selecting on ID. From this plot, it is clear there is some sort of relationship between the operator, the date the observations were made, and the observations impacted (ID). In order to investigate this further, we can look at the DataTable.

Dataset - Tabplot

Show variables:

Y ID Operator Date Priority sensor2
 sensor9 sensor12 sensor15 sensor21
 sensor22 sensor24

Sort on:

Date

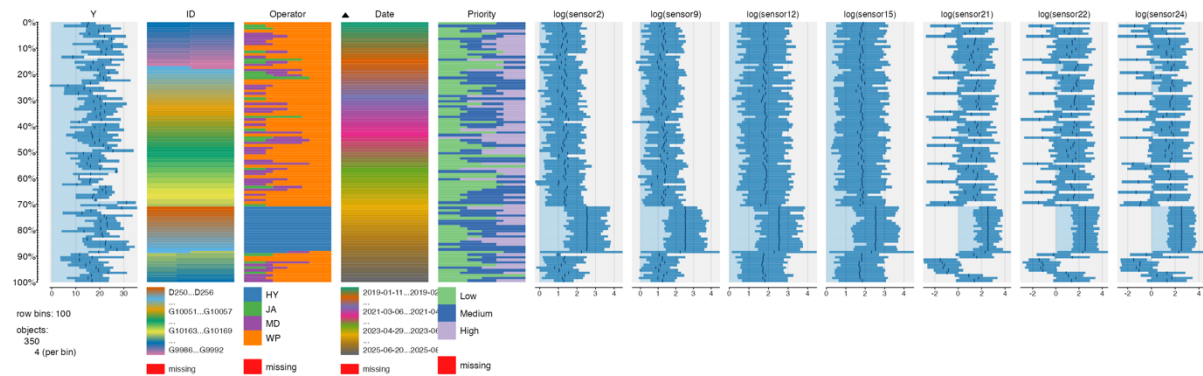


Figure 5: TablePlot of variables Y, ID, Date, Priority, sensor2, sensor9, sensor12, sensor15, sensor21, sensor22, sensor24, grouped by date.

MatPlot

The MatPlot is also able to show the relative homogeneity of a dataset. In this case, the homogeneity of the data is not indicated to be homogenous across all variables. There is a distinct jump in a select number of variables from the rest of the variables (figure 6), indicating these variables are not homogenous with the rest of the dataset, and do not share similar measures of mean and spread.

Dataset Homogeneity

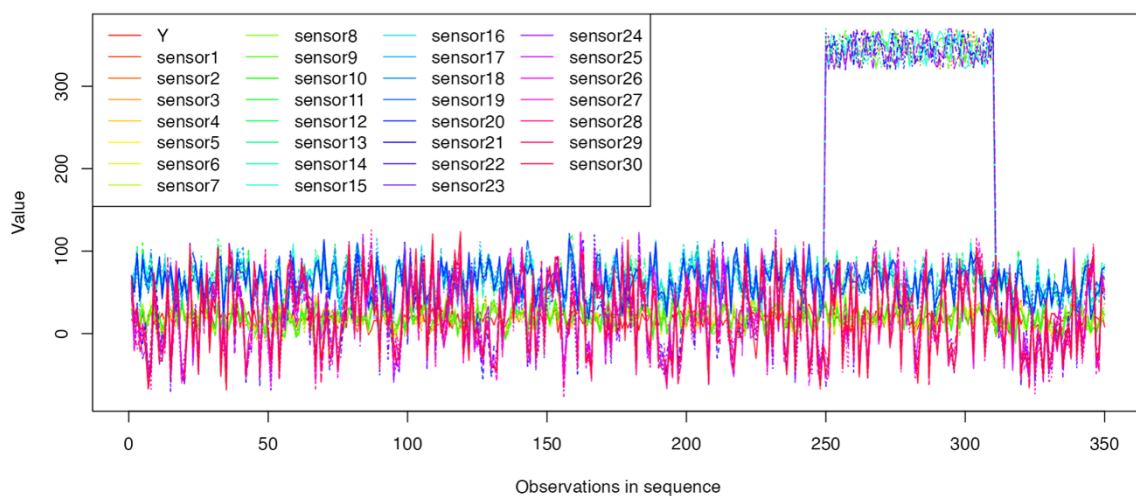


Figure 6: Dataset homogeneity via Matplot, indicating non-homogenous dataset.

Missingness with Vis_Miss

The missingness chart is a clear way to visualise the missingness across the dataset. Variables 'Y', 'ID', 'Operator', 'priority' and 'agreed' are all variables with no missingness ('Date' also has no missingness but has been left out of this graph for clearer visualisation purposes). Immediately noticeable is the abundant missingness of the 'sensor3' variable, which was commented on previously. This variable has the most missingness at 29% of the data missing (figure 7). It might be worthwhile managing sensor 3 more closely to improve this data quality. The rest of the sensor variables have relatively consistent missingness, only having 1%-4% missing data across the sensor variables. The spread of the missingness is also relatively consistent, save for a gap just below the 300 mark where there is distinctly no missingness (aside from sensor3).

When clustering the missingness, we can see most of the missingness is concentrated in the earlier observations, and that most variables don't have too much missing data in general. There does not seem to be a connection between missingness and the sensors of interest.

Visualising Missing Data



Figure 7: Visualisation of the missingness of dataset (excluding Date), where FALSE is not missing, TRUE indicates missing datapoints.

Rising Values Chart

The rising values chart is able to show the continuity of values for the data, specifically for numeric variables. The rising values chart on our dataset shows that some variables have steps in the chart. As we would assume these numeric values to be continuous, these variables with a step are unexpected and suspicious. Investigation into these variables with a significant step in the rising values chart showed that these variables in question are once again the sensors of interest, as illustrated below (figure 8). All sensors of interest seem to step at around the 80th percentile. Sensor3, the sensor with the previously described abundant missingness, also did not follow a regular trend much shorter than the other variables, again indicating missingness (not pictured). This finding in the rising values chart is consistent with findings in other charts, and further supports the evidence that these sensors do not follow usual trend patterns.

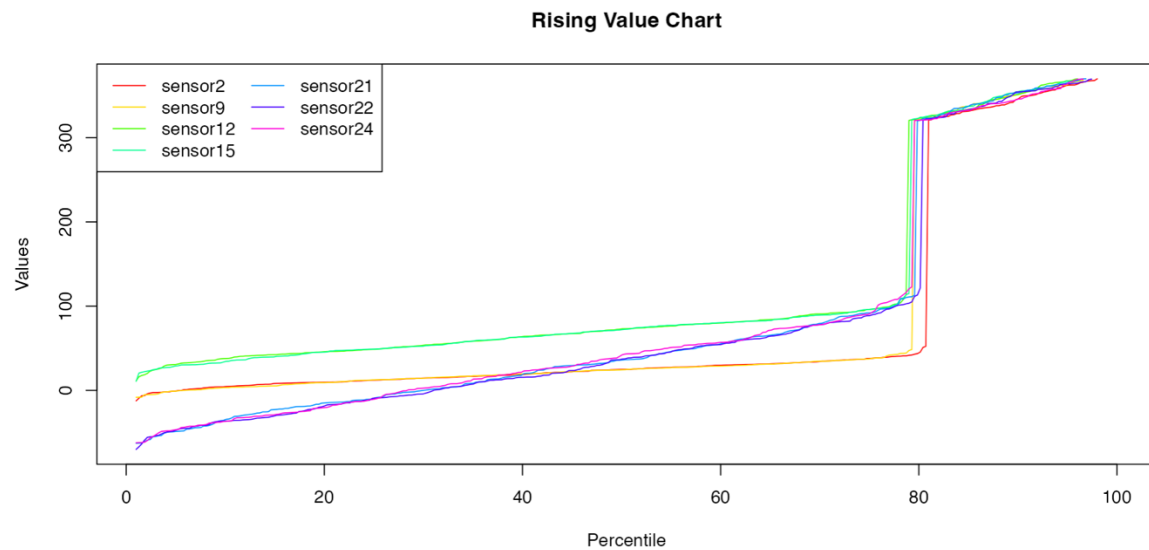


Figure 8: Rising Value Chart, only picturing sensor2, sensor9, sensor12, sensor15, sensor21, sensor22, sensor24.

Data Table

Leading on from the TablePlot investigations, when looking at the Data Table, it is possible to see the raw observations, and to group the observations based on each variable. When grouping by operator, the operator of interest, 'HY' looks to have recorded observations sequentially from the date 2023-10-20 until 2024-12-13. Observing the values for the sensor data, we can also see that sensor2, sensor9, sensor12, sensor15, sensor21, sensor22 and sensor24 do consistently report values of around 300 for the 'HY' operator. As we have seen in our previous data exploration efforts, these particular sensor variables do not follow the same trend as the other sensor variables.

Final comments

Across the various visualisations in the R Shiny app, we have been able to visualise interesting trends and patterns across the dataset. We have found 1 variable of high missingness (sensor3), and it would be recommended to investigate this sensor further for faults and further management. We have also found there to be a group of sensors (sensor2, sensor9, sensor12, sensor15, sensor21, sensor22, sensor24) which do not follow the same trend as the other sensors, and instead tend to group together across different metrics. The correlation between these variables is stronger among these variables than the other variables, as seen in the Corrgram. All these variables of interest also appear to have a bimodal distribution, as seen in the initial summary statistics of the dataset. Upon further investigation, we found that these variables are suspiciously linked to the 'operator' factor variable, specifically relating to the 'HY' operator as indicated in the pairs plot. The TablePlot further supports this link between the 'HY' operator and these sensors of interest. Further investigations are recommended to identify the cause of this link, either through investigation of metadata surrounding this dataset, or further investigation via scatterplots and density plots.