ANALYSIS OF CHARTS

MATRIX PLOT

- All sensors in the matrix plot have normal scattering apart from sensors 3, 4, 9, 22, 24, and 27, which have two distinct clusters within them. Sensor 9 has four clusters when charted with sensor 3 and 4, which is particularly unusual.
- When we group these by “Operator”, it is clear that there is something unusual happening with the KL operator, represented by the blue circles. In almost all cases, the KL scores above 300 while other operators score in the 0-100 range. It seems these high values are causing the multiple cluster formations.
- Another interesting observation surfaces when we group by “Duration” and see the scatter of the Y variable in relation to all the other sensors. The average “Very Long” value is above the “Long”, and the average “Short” value is above them both.

CORRELATION

- When we use Pearson’s correlation, four clusters emerge. All four clusters have a dark blue color, which indicates strong positive correlation.
- When we changed the setting to a non-linear correlation measure i.e. Kendall or Spearman. We only observe three distinct clusters having strong positive correlation.
- Whether its Peason’s, Kendall or Spearman’s correlation, the sensors within each cluster are related by their numbers. For example, sensors 1-10 are more likely to cluster. As are sensors 11-20 and 21-30.
- When we look at the categorical variables, expensive is correlated with fast, short, wooden and warm.

PAIRS

- The HH operator seems to have the most observations, and we can see this translate into other graphs.

MISSING VALUES

- Operator, Priority, Agreed and Y do not have any missing values.
- All sensors have some missing values in them and it is clear that sensor 9 has the most missing.
- The missingness of data becomes more frequent near the 90 observation mark. Before, it was quite rare to see a missing value. This may indicate that the sensors need more maintenance after 90 observations pass.

BOXPLOT

- When not standardized, the boxplot reinforces what we observed in the correlation matrix of sensors grouping by intervals 1-10, 11-20 and 21-30.
- The variance of measurements also tends to increase as the sensor number increases. Sensors 1-10 have the smallest variance, sensor 11-20 have bigger variances and sensors 21-30 have the biggest variances.

- When we standardize, most sensors actually look similar apart from sensors 3, 4, 13, 17, 22, 24 and 27. Note, five of the sensors just mentioned were the same ones behaving oddly in the matrix plots with the outlier observations of 300.
- And when we tick show outliers, these same sensors have the most outliers unsurprisingly.
- When we adjust the IQR multiplier past 3.5, sensors 22 and 24 stop having outliers. Whereas sensors 3, 4 13 and 17 have outliers even with a multiplier of 5.

MOSAIC

- When Duration is split with Agreed they are evenly balanced.
- It seems most pairings of categorical variables produce gray coloured shapes, which indicate there is nothing unusually uncommon or common about them.

RISING VALUE

- It is clear that a lot of the discontinuity happens at around the 80th percentile. The sensors associated with a gap around this mark are 3, 4, 13, 17, 22, 24, and 27.
- Again sensor 9 seems to be particularly unusual as it is the only sensor that has missing values before the 80th percentile. Gapping before the 40th percentile. It also falls short and does not reach the end due to the missingness. However, sensor 9's gap is smaller compared to the rest.
- All the other sensors behave similarly and are pretty much copies of each other.

HISTOGRAM

- Most sensors are distributed approximately normally apart from sensors 3, 4, 13 , 17, 22, 24 and 27, which have bimodal distributions.
- Sensor 9 looks like it has a bimodal distribution but it may just have missing values between 15-25. This is because in all other cases the values of the distribution range from 0-300. Sensor 9's only range from 0-50. Therefore, the gap in the middle is most likely due to missing data. The missing values chart also supports this fact.
- For categorical data, priority, agreed, location, state and duration has a fairly uniform distribution of factors.
- From operators, KL has the lowest count. Therefore what we observed in the matrix plot may be outlier events.
- From price, expensive has the highest count. And from class, wooden and ceramic have the highest count