The data used in Assignment 1 consists of 350 observations of 44 variables, the variables are a mix of numerical and categorical data. Some manual intervention was required to interpret the Date value into a recognisable date and ordering was required for the ordinal values in the dataset to properly represent their natural order. The summary gives minor insights to the context behind the data, as information is being recorded from 30 sensors with different categorical variables assigned to each day's recording.

The box plots used start off by showing the first 5 numerical variables in the dataset, this first look shows that Sensor's 1-4 and variable Y appear similar. Sensors 3 and 4 have similar outlier's with and IQR multiplier of 1.5. When expanded to include all numerical variables some clearer patterns start to appear within the uncentred and unscaled data. The sensors can be grouped in tens with the first 10 and variable Y having high similarity, Sensors 11 to 20 are the next group to have similarities and finally Sensors 21 to 30 also have similarities. Sensors 3, 4, 13, 17, 22, 24 and 27 have high outliers only and when the IQR multiplier is reduced, these outliers still show a large gap between them and the main data. When the IQR multiplier is set to 3.4 the aforementioned Sensors greater than 20 lose their outliers, none of the other sensors lose their outliers even when the IQR multiplier is set to 5.

In the Missing Value chart, a down sampled selection of the data has been selected, primarily to display the large scale missingness of Sensor 9 when in comparison to other similar Sensors. Overall, only 3.4% of the overall data is missing from the dataset and we can also clearly see that variables Y, ID, Operator, Date, Priority and Agreed have no missing values. When the missingness is clustered, it is clear there is relatively minor missingness in earlier observations and this subsides until later observations where the minor missingness appears again. As mentioned earlier Sensor 9 has relatively high missingness with 30.29% of the data missing, other variables hover around 4% for missingness. For the other variables there aren't any obvious patterns in the missingness, and it seems fairly random.

The Rising Value chart shows 3 very distinct patterns for data with no discontinuity. This seems to match the similarity observed in the box plots where the data can be separated into their groups of 10. The relatively low missingness of the majority of the variables means that most of the lines are fairly smooth with minimal observable stepping. The variables that fall outside of this observation are as follows Sensors 3, 4, 9, 13, 17, 22, 24 and 27. Variable 9 falls very short of the end of the chart due to all its missing values, it also shows a discontinuity in which the values jump from 15 to 25 around 0.38 percentile. The other 7 variables all show a discontinuity around the 0.8 percentile mark, they also all start similarly to their continuous counterparts with the aforementioned pattern being observable. Sensors 3 and 4 have values that jump from approximately 50 to 300, whereas Sensors 13, 17, 22, 24 and 27 have a jump in value from around 100 to 300. These other variables only fall minimally short in comparison to the rest. When centred and scaled the data with no discontinuity loses the observable pattern and they all follow a very similar path.

There are no overly dark zones in the Mosaic chart, and only a couple of variables get coloured when being compared, for example Price and Class get one light red for Plastic and Cheap, and when being split by Speed three light blue areas appear. This shows that the data is fairly evenly balanced throughout. When increasing the amount of variables it is easy for this data to get muddled as there are many options for comparison, at 4 variables the chart is already quite confusing. Location has too high a cardinality to be useful in the mosaic chart even though it shows some difference between the combinations of factors. The combination of Speed, State, Operator and Price increases the residuals for 4.6 for two of the combinations, including Temp increases this residual to 6.0 but this makes the graph nearly completely illegible. Overall, more light blue appears when comparing variables which shows that there are a fair number of combinations of factors that are more common than usual.

The correlation charts have very interesting grouping for each method and helps to display the patterns throughout the data. Spearman correlation also highlights the previously noticed pattern in the data set, whereas the Pearson method splits the data into four distinct groups with strong correlation. An example of each of these groups are how Sensors 1-10 have high correlation with the Spearman method under OLO grouping whereas Pearson shows negative correlation with Sensor 3 in combination with all other sensors less than 10 other than 4 which still has strong correlation. The Kendall method shows a lot more correlation across the board than the other two methods but sticks with similar groups to the Spearman method. When absolute correlation is turned off, more light blue is shown to display that there is a majority of positive correlation throughout the dataset.

The pairs chart has a lot of information on display and has been given the options to adjust the variables shown and which categorical variable to colour the data by. Variables such as ID and Date have been excluded in this graph as their cardinality is too high for any cohesive information to be shown. Any numerical data in combination with a categorical data point does not scale as there are too many variables to display, but the option is still there for a smaller set to be observed. Sensor 1 and Sensor 2 show a strong positive linear relationship, whereas Sensor 11 and Sensor 29 show close to no relationship. The scatter plot has no discernible trend. Sensor 13 and Sensor 21 show two distinct sets of points, when coloured by Operator, one set of points is only created by a specific operator whereas the other set is a mix. Variable Y shows an interesting grouping when coloured by Duration which may indicate some correlation between the two variables.

Overall, exploring the data found many different groupings and sets throughout which has created many questions to further discover the data. The many different charts and graphs enabled the chance to link a clear pattern through many different methods rather than just one. The large amount of data shows how easy it is to get lost in the quantity and a clear question to cut out the extra data would benefit further data exploration for a user.