DATA VISUALIZATION ANALYSIS FOR DATASET-ASS1DATA USING SHINY

1.DATATABLE:

The given dataset Ass1Data.csv has 13 discrete variable columns, 40 continuous variable columns and 1 date column.

2.SUMMARY:

I have used summary, glimpse, descriptive statistics, skim, dfsummary to find different summary values for the dataset. For example, median, mean, skewness, IQR, standard deviation, minimum and maximum value.

3.VISUALISATIONS:

3.1 Mosaic Plot:

The mosaic plot is used to examine the relationship between two or more categorical data. For example, on observing the mosaic plot for our dataset if Priority is low, speed is low, duration is long or very long and state is checked then the event unusually common. If Priority is low, Speed is medium, duration is short, and state is checked then the event is unusually rare. I have used these variables as the default.

3.2 Missing Values:

The given dataset has missing values almost in all columns except ID,Y, Author and Priority. The sensor 6 has the least missing values and sensor 7 has the highest missing values. Therefore, the maximum and minimum missing values of the dataset stays next to each other. There is no interesting pattern in the dataset except this.

3.3 Rising Values:

On observing the rising value chart, I got three types of discontinuity in values. The first one is sensor 7 which is unique in discontinuity than other columns. This is because of its highest missingness in the data. The sensors 3,4,13,17,22,24,27 shows same level of discontinuity in their values. The other sensor columns and Y column show the same level of discontinuity in their values.

3.4 Correlation

I have used three types of correlation to analyse the pairs of numeric variables. Pearson type correlation shows 4 sets(7-10 variables in each set) that are highly correlated with each other. Spearman and Kendall type of correlation shows 3 sets in their correlation graph. This correlation is also used to plot gg pairs to draw more observations within that highly correlated sets.

3.5 Box Plot:

I have used a box plot for continuous variables to find patterns in the outliers. The sensors 3,4,13,17,22,24,27 has similar outliers in their data. These sensors are likely to be different

from other variables as its median line lies outside the box plot of other variables. The data in these sensors are less dispersed than the other variables since the interquartile ranges low. Among these sensors, sensor 3 and 4 are likely to be the same, sensors 13 and 17 are similar and sensors 22,24,27 are likely to be the same in their median and interquartile.

3.6 GG Pairs:

GG pairs plot shows the correlation between the pairs. I have used the various sets from the correlation to plot the GG pairs. If a set has more that 8 variables, I filtered it into less than or equal to 8 variables by their correlation. For example, Pearson correlation shows sensor 23 and 21 is highly correlated as its value is 0.982 and sensor 5 and sensor 22 is less correlated as its value is 0.16.

3.7 TabPlot:

For numeric values in the dataset, the tabplot shows the sensor 3,4,13,17,22,24,27 has similar patterns than the other variables. These variables are similar in all the plots and it forms a separate group.