**Assignment 2: Data Scavenger Hunt**

**Niranjana Sathish Avilery**

The main goal of this assignment is to utilize visualizations to observe outliers in a dataset. Two datasets “waveform.csv” and “times.csv” were provided which had five matrices (LinearAcceleration, RotationalVelocity, RotationalAcceleration, TimeLinear, TimeRotational).

There are 65 instances of data, of which 5 are statistically "different" than the other 60. To find the five data instances, new features were generated by iterating across the three waveform matrices (LA, RV, and RA). The minimum, maximum, and mean values of each individual waveform were computed. On finding the summary statistics for each of these features, three features appeared to be the most useful for retrieving the “different” data instances from the rest of them. The feature selection was done by choosing the features with the highest range (computed by the difference between maximum and minimum) values and the selected features were PLA, PRV, and PRA.

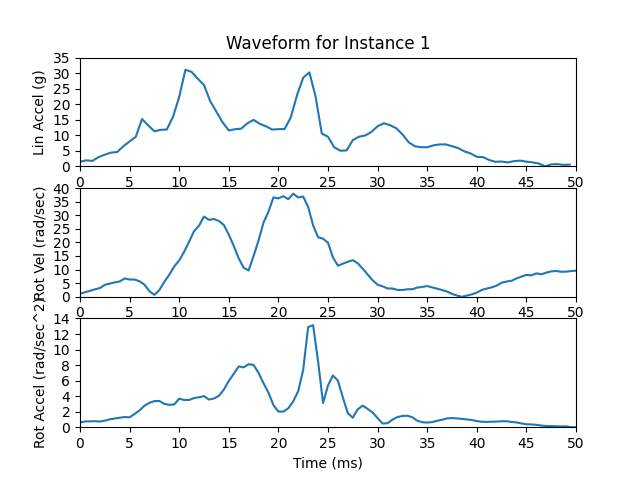
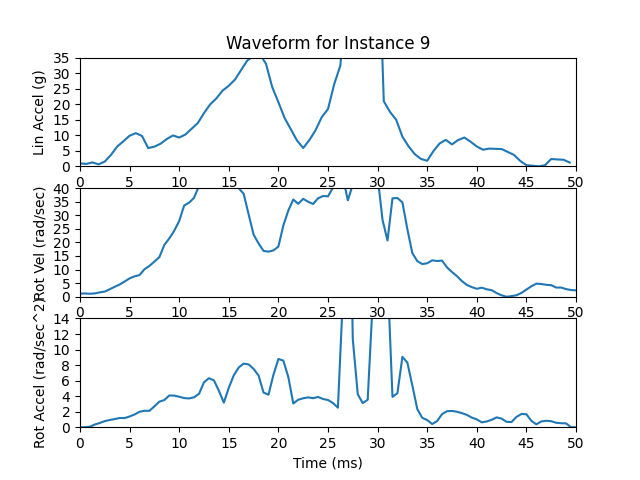


Figure 2 – Waveform of Instance 1 (Example for “normal” data instance

Figure 1 – Waveform of Instance 9 (Example for “different” data instance

The waveforms in Fig(1) and Fig(2) assist in visualizing how contrasting the “different” data instance is from the “normal” instance. The plots for instance 9 have values that are outside of the range of the y-axis (magnitude) which is not the case in instance 1.

Further investigation of these features through visualization (Fig(3), Fig(4), and Fig(5)) shows the “Top 5” largest feature values for each of the features. The “normal” instances are closely grouped together and far away from the outliers in Fig(4) than the other two scatter plots. This helps us infer that ‘PRA vs PLA’ plot is more useful in determining the anomaly. The data instances in Table 1 were found by sorting the NumPy arrays (PLA, PRA, and PRV) and by slicing the first five values.

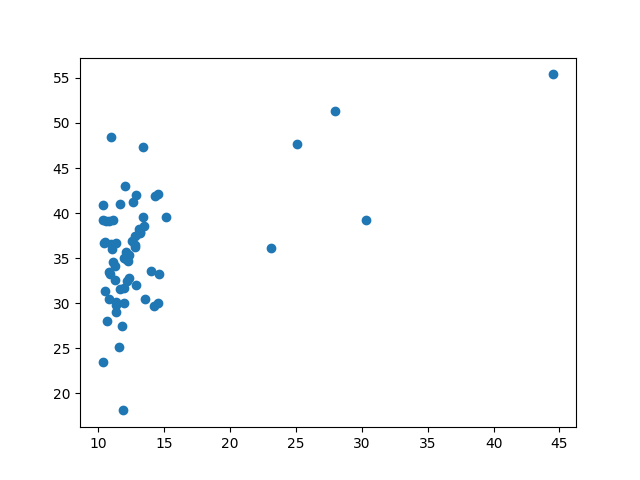
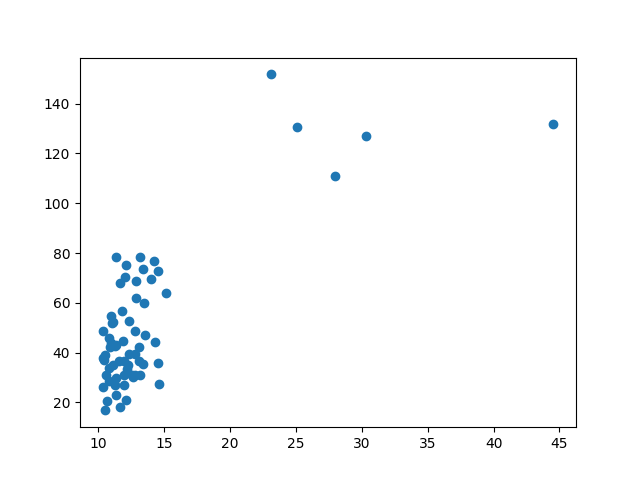
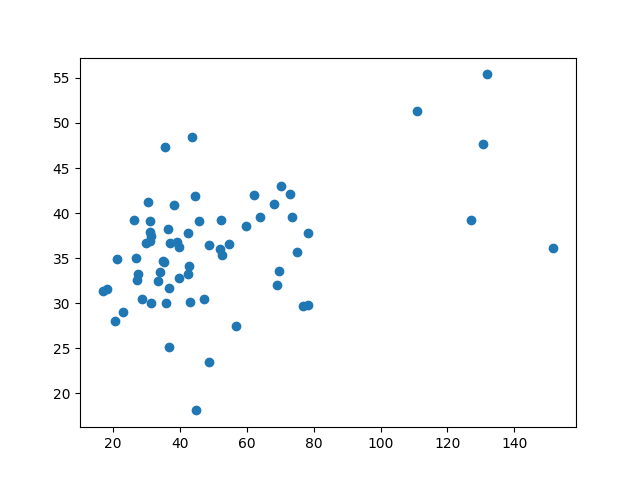


Figure 5 – Scatter plot of PRA vs PRV

Figure 4 – Scatter plot of PRA vs PLA

Figure 3 – Scatter plot of PLA vs PRV

|  |  |
| --- | --- |
| **Feature** | **Instance Numbers of Top 5 largest feature values** |
| Peak Linear Acceleration (PLA) | 20, 9, 11, 19, 8 |
| Peak Rotational Acceleration (PRA) | 9, 19, 8, 11, 20 |
| Peak Rotational Velocity (PRV) | 9, 8, 60, 11, 34 |

Table 1 – Features with Instance Numbers of Top 5 largest feature values

The instance numbers for two of the features (PLA, PRA) in the table above were an exact match, indicating that these are the five "different" data instances.