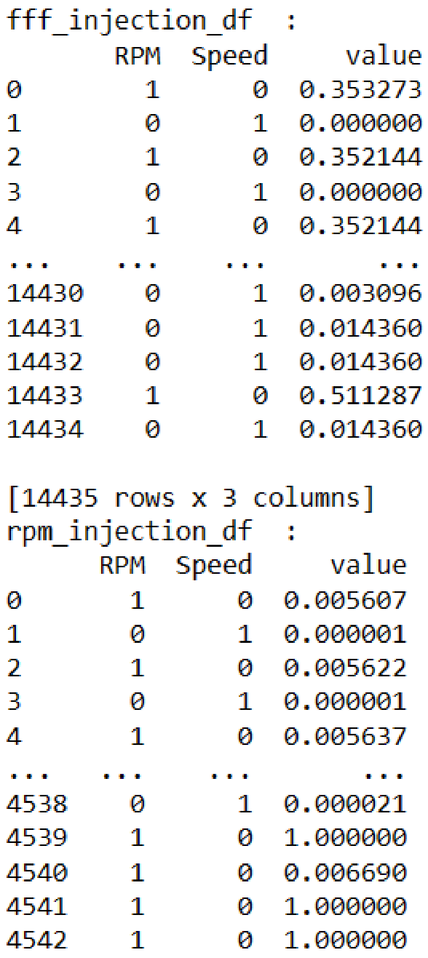
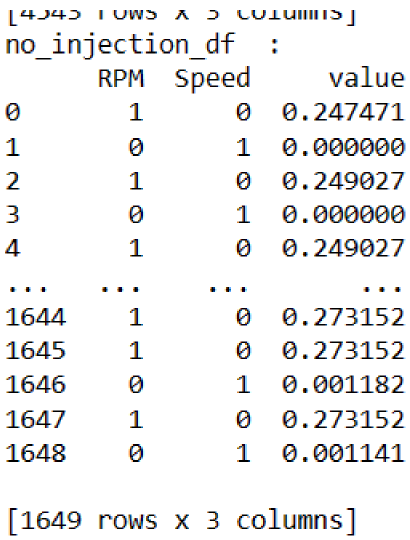
# Report

**Task1: -**



```
fff_injection_df  :
       RPM  Speed     value
0        1      0  0.353273
1        0      1  0.000000
2        1      0  0.352144
3        0      1  0.000000
4        1      0  0.352144
...    ...    ...       ...
14430    0      1  0.003096
14431    0      1  0.014360
14432    0      1  0.014360
14433    1      0  0.511287
14434    0      1  0.014360

[14435 rows x 3 columns]
rpm_injection_df  :
      RPM  Speed     value
0       1      0  0.005607
1       0      1  0.000001
2       1      0  0.005622
3       0      1  0.000001
4       1      0  0.005637
...   ...    ...       ...
4538    0      1  0.000021
4539    1      0  1.000000
4540    1      0  0.006690
4541    1      0  1.000000
4542    1      0  1.000000
```



```
[4543 rows x 3 columns]
no_injection_df  :
      RPM  Speed     value
0       1      0  0.247471
1       0      1  0.000000
2       1      0  0.249027
3       0      1  0.000000
4       1      0  0.249027
...   ...    ...       ...
1644    1      0  0.273152
1645    1      0  0.273152
1646    0      1  0.001182
1647    1      0  0.273152
1648    0      1  0.001141

[1649 rows x 3 columns]
```

The above are the three data frames contains three columns RPM, Speed and Value
1)FFM_injection_df=injection of FFF as the speed reading
2)rpm_injection_df= injection of RPM readings
3)no_injection_df=no injection of messages
As you observed, the three data-frames do not contain the attack column at all. Data collection is done. Task 1 completed.

**TASK 2: -**

**Speed Clusters comparisons: -**
**1. FFF Injection: - (Fig 1)**

- Observation: In this case, a distinct yellow line parallel to the Value (Y-axis) at 0.0 is visible. A unique cluster is shown by this line.

- Clusters: Two distinct clusters are present. While the other is centered around (1.0, 0.0), the first is centered about (0.0, 0.5).
- Discussion: There are two distinct clusters in the FFF injection scenario, indicating that the two states (Attack (0) and No Attack (1)) are clearly separated from one another.

**2. RPM Injection: - (Fig 5)**

- Observation: The red centroids in this scenario are roughly located at (0.0, 1.0) and (0.4, 0.0). Two violet spots are also present at (0.0, 0.0) and (1.0, 0.0).
- Clusters: Two clusters appear to exist based on the red centroids' presence. The violet dots, on the other hand, show which data points are not part of either of these clusters.
- Discussion: The RPM injection scenario appears to show two main clusters, but it also includes a collection of data points that don't fit well into either of these clusters, which could be a sign of noise in the data.

**3. No Injection:(Fig 3)**

- Observation: At 0.0, there is a yellow line parallel to the Y-axis (Value), just like in the FFF Injection scenario. This points to a distinct cluster.
- Clusters: There are two distinct clusters, same as the FFF Injection scenario. While the other is centered around (1.0, 0.0), the first is centered about (0.0, 0.5).
- Discussion: The two states, Attack (0) and No Attack (1), are clearly separated in the No Injection scenario, which displays two different clusters, much like the FFF Injection scenario.

**Comparison**

- With two distant clusters, the FFF Injection and No Injection scenarios display a very similar pattern in the Speed dataset, however the RPM Injection scenario displays a different pattern.
- RPM Injection appears to consist of two main clusters, but it also includes data points that don't fit into any of them, indicating some degree of uncertainty.
- In conclusion, the scatter plots demonstrate that the scenarios with FFF Injection and No Injection have more distinct clustering patterns than the RPM Injection scenario, which has a greater number of inconsistent and unclear data points. The discrepancies could be ascribed to the characteristics of the information and the influence of the injection measurements.

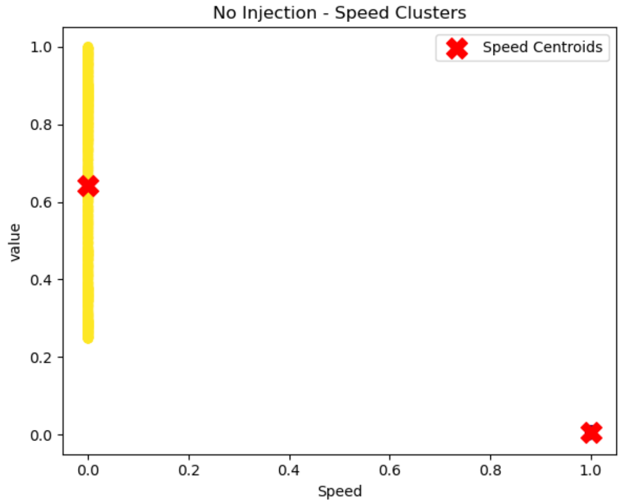


Fig 3

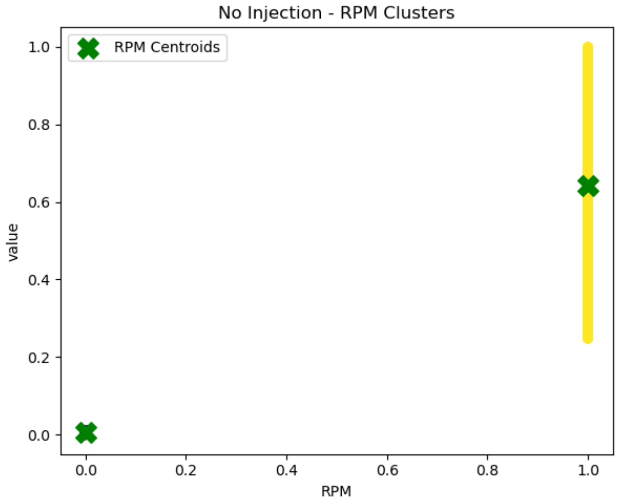


Fig 2

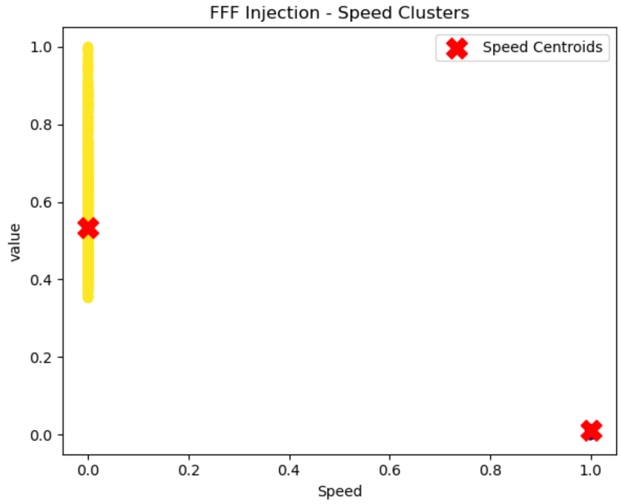


Fig 1

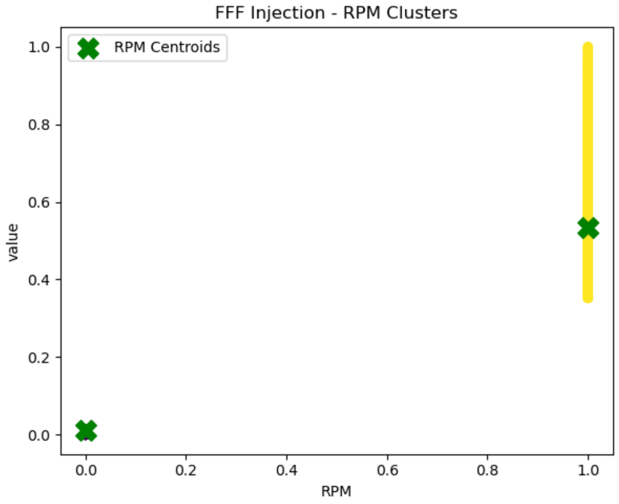


Fig 4

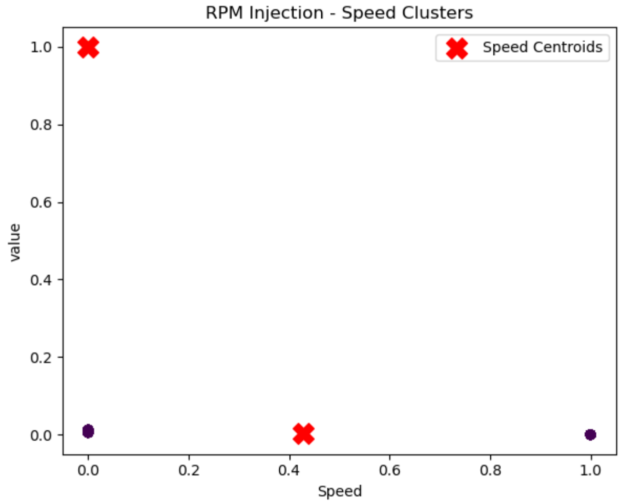


Fig 5

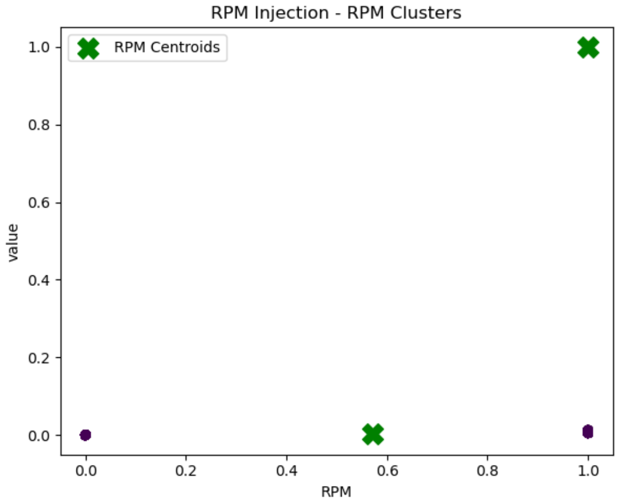


Fig 6

**RPM Clusters comparisons: -**

**1. FFF Injection: - (Fig 4)**

- Observation: In this case, a distinct yellow line parallel to the Value (Y-axis) at 1.0 is visible. A unique cluster is shown by this line.
- Clusters: Two distinct clusters are present. While the other is centered around (0.0, 0.0), the first is centered about (1.0, 0.6).
- Discussion: There are two distinct clusters in the FFF injection scenario, indicating that the two states (Attack (0) and No Attack (1)) are clearly separated from one another.

**2. RPM Injection: - (Fig 6)**

- Observation: The green centroids in this scenario are roughly located at (1.0, 1.0) and (0.6, 0.0). Two violet spots are also present at (0.0, 0.0) and (1.0, 0.0).
- Clusters: Two clusters appear to exist based on the green centroids' presence. The violet dots, on the other hand, show which data points are not part of either of these clusters.
- Discussion: The RPM injection scenario appears to show two main clusters, but it also includes a collection of data points that don't fit well into either of these clusters, which could be a sign of noise in the data.

**3. No Injection: -(Fig 2)**

- Observation: At 1.0, there is a yellow line parallel to the Y-axis (Value), just like in the FFF Injection scenario. This points to a distinct cluster.
- Clusters: There are two distinct clusters, same to the FFF Injection scenario. While the other is centered around (0.0, 0.0), the first is centered about (1.0, 0.5).
- Discussion: The two states, Attack (0) and No Attack (1), are clearly separated in the No Injection scenario, which displays two different clusters, much like the FFF Injection scenario.

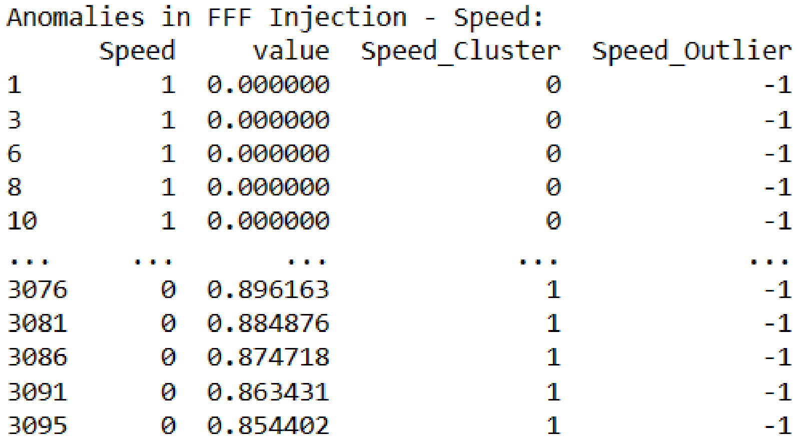
**Comparison**

- With two distant clusters, the FFF Injection and No Injection scenarios display a very similar pattern in the RPM dataset, however the RPM Injection scenario displays a different pattern.
- RPM Injection appears to consist of two main clusters, but it also includes data points that don't fit into any of them, indicating some degree of uncertainty.
- In conclusion, the scatter plots demonstrate that the scenarios with FFF Injection and No Injection have more distinct clustering patterns than the RPM Injection scenario, which has a greater number of inconsistent and unclear data points. The discrepancies could be ascribed to the characteristics of the information and the influence of the injection measurements.

**Task 3: -**

**FFF Injection:(Fig 1.1): -**

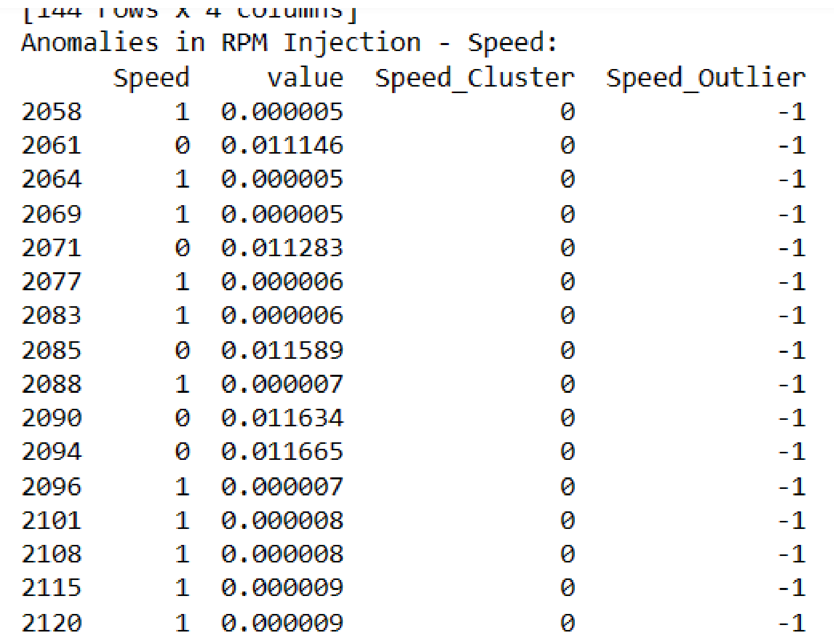
The result that is given displays anomalies that were found in the speed dataset under three scenarios (FFF Injection, RPM Injection, and No Injection).



```
Anomalies in FFF Injection - Speed:
      Speed     value  Speed_Cluster  Speed_Outlier
1         1  0.000000              0             -1
3         1  0.000000              0             -1
6         1  0.000000              0             -1
8         1  0.000000              0             -1
10        1  0.000000              0             -1
...     ...       ...            ...            ...
3076      0  0.896163              1             -1
3081      0  0.884876              1             -1
3086      0  0.874718              1             -1
3091      0  0.863431              1             -1
3095      0  0.854402              1             -1
```

When the speed is zero and the value is approximately greater than 0.8 in the FFF Injection scenario, most anomalies are found. In the dataset, these anomalies are identified as Speed Outlier = -1. This implies that there are anomalies in the FFF Injection scenario when the speed is low, and the value is high. This is correctly shown in the scatter plot in fig 1 that purple approximately greater 0.8 indicates anomalies. Anomaly 1 at (0.0, 0.8) denotes the existence of situations in which the value is comparatively high (0.8) while the speed is low (0.0). Cases where the speed is at its highest (1.0), but the corresponding value is extremely low (0.0) are represented by Anomaly 2 at (1.0, 0.0). The scatter plot of the anomalies data frame is above.

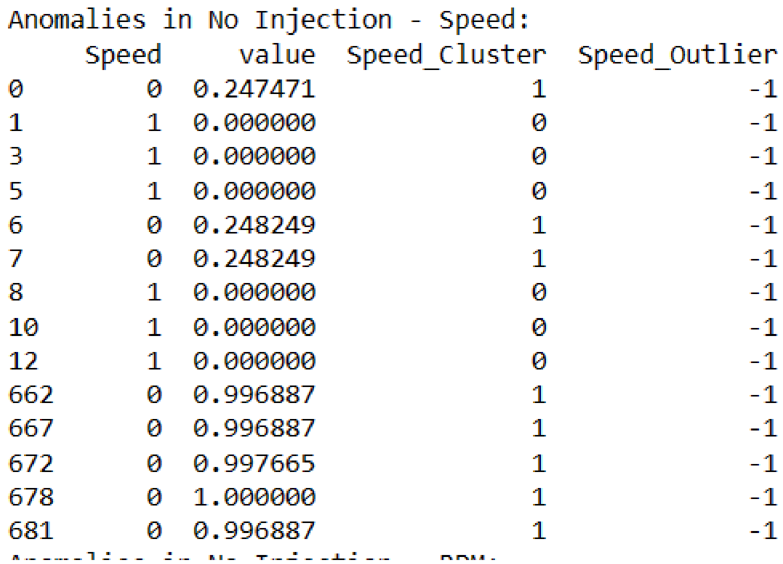
**RPM Injection:(Fig 2.1)**



```
[144 rows x 4 columns]
Anomalies in RPM Injection - Speed:
      Speed     value  Speed_Cluster  Speed_Outlier
2058      1  0.000005              0             -1
2061      0  0.011146              0             -1
2064      1  0.000005              0             -1
2069      1  0.000005              0             -1
2071      0  0.011283              0             -1
2077      1  0.000006              0             -1
2083      1  0.000006              0             -1
2085      0  0.011589              0             -1
2088      1  0.000007              0             -1
2090      0  0.011634              0             -1
2094      0  0.011665              0             -1
2096      1  0.000007              0             -1
2101      1  0.000008              0             -1
2108      1  0.000008              0             -1
2115      1  0.000009              0             -1
2120      1  0.000009              0             -1
```

When the speed is 1 and the value is extremely low (almost 0), anomalies in the RPM Injection scenario are mostly seen. Speed Outlier = -1 is used to identify these anomalies. This suggests that when the speed is high, but the value is very low, anomalies in the RPM Injection scenario happen. The scatterplot is green. It does not contain more anomalies; it contains very little shown in the scatter plot in fig 2. Indicating in the RPM injection it does not affect the speed. The graph is similar to a normal graph which was drawn in task 2. Very Less anomalies. The whole scatter plot is green in color and anomalies are very less.

**No Injection :(Fig 3.1)**

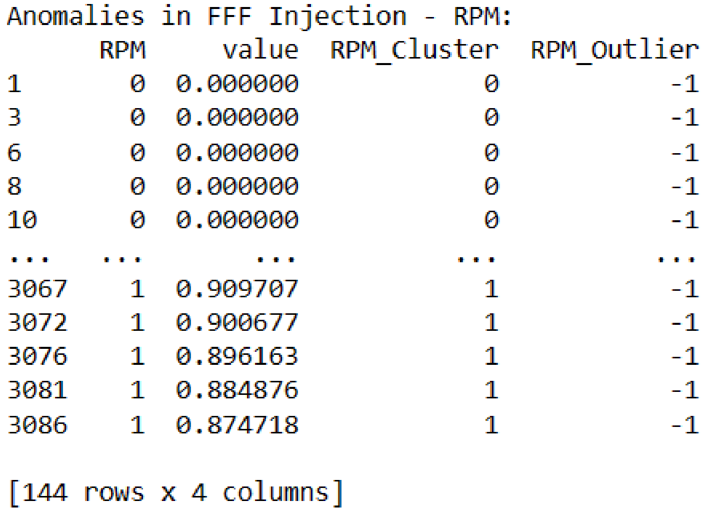


```
Anomalies in No Injection - Speed:
     Speed     value  Speed_Cluster  Speed_Outlier
0        0  0.247471              1             -1
1        1  0.000000              0             -1
3        1  0.000000              0             -1
5        1  0.000000              0             -1
6        0  0.248249              1             -1
7        0  0.248249              1             -1
8        1  0.000000              0             -1
10       1  0.000000              0             -1
12       1  0.000000              0             -1
662      0  0.996887              1             -1
667      0  0.996887              1             -1
672      0  0.997665              1             -1
678      0  1.000000              1             -1
681      0  0.996887              1             -1
Anomalies in No Injection - RPM:
```

Anomalies are found in the No Injection scenario in several scenarios where the speed is between approximately 0 and value greater than 0.9. Speed Outlier = -1 is used to identify these anomalies. Compared to the FFM-injection it is almost similar to it. But the anomalies are very less compared to FFM almost equal to 10 anomalies. The scatter plot Fig 3.1 shows purple at top of the straight line at speed ==0.0 and value greater than 0.9 and remaining all are green.

**RPM dataset: -**

**FFF Injection - RPM:(4.1)**

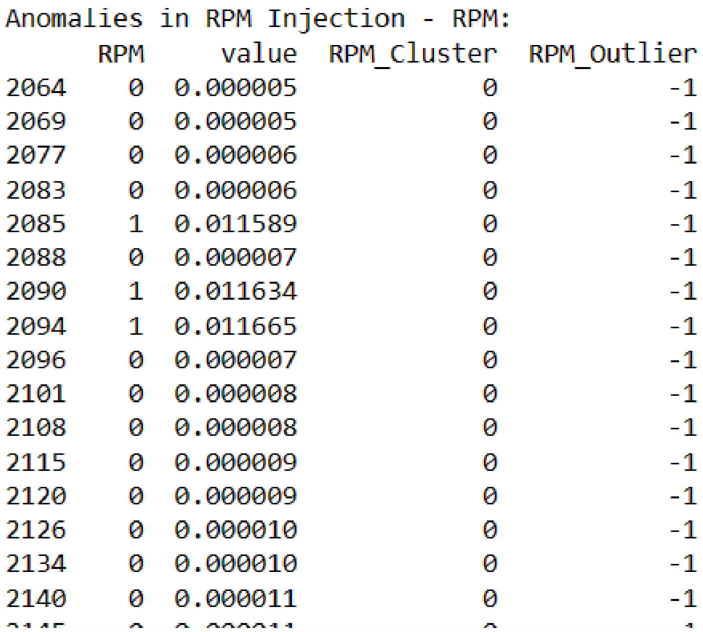


```
Anomalies in FFF Injection - RPM:
      RPM     value  RPM_Cluster  RPM_Outlier
1       0  0.000000            0           -1
3       0  0.000000            0           -1
6       0  0.000000            0           -1
8       0  0.000000            0           -1
10      0  0.000000            0           -1
...   ...       ...          ...          ...
3067    1  0.909707            1           -1
3072    1  0.900677            1           -1
3076    1  0.896163            1           -1
3081    1  0.884876            1           -1
3086    1  0.874718            1           -1

[144 rows x 4 columns]
```

When the RPM is zero and the value is approximately greater than 0.8 in the FFF Injection scenario, the majority of anomalies are found. In the dataset, these anomalies are identified as Speed Outlier = -1. This implies that there are anomalies in the FFF Injection scenario when the rpm is low, and the value is high. This is correctly shown in the scatter plot in fig 4 that purple approximately greater 0.8 indicates anomalies. Anomaly 1 at (0.0, 0.8) denotes the existence of situations in which the value is comparatively high (0.8) while the speed is low (0.0). Cases where the rpm is at its highest (1.0) but the corresponding value is extremely low (0.0) are represented by Anomaly 2 at (1.0, 0.0). The scatter plot of the anomalies data frame is above. This is similar to FFF-injection with speed data frame.

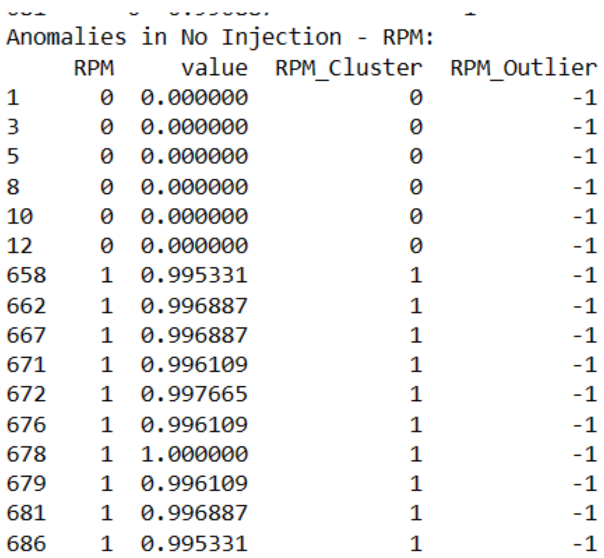
**RPM Injection- RPM:**



```
Anomalies in RPM Injection - RPM:
      RPM     value  RPM_Cluster  RPM_Outlier
2064    0  0.000005            0           -1
2069    0  0.000005            0           -1
2077    0  0.000006            0           -1
2083    0  0.000006            0           -1
2085    1  0.011589            0           -1
2088    0  0.000007            0           -1
2090    1  0.011634            0           -1
2094    1  0.011665            0           -1
2096    0  0.000007            0           -1
2101    0  0.000008            0           -1
2108    0  0.000008            0           -1
2115    0  0.000009            0           -1
2120    0  0.000009            0           -1
2126    0  0.000010            0           -1
2134    0  0.000010            0           -1
2140    0  0.000011            0           -1
2145    0  0.000011            0           -1
```

When the rpm is 1 and the value is extremely low (almost 0), anomalies in the RPM Injection scenario are mostly seen. Speed Outlier = -1 is used to identify these anomalies. This suggests that when the rpm is high, but the value is very low, anomalies in the RPM Injection scenario happen. The scatterplot is green. It does not contain more anomalies; it contains very less   shown in the scatter plot in fig 5.1. Indicating in the RPM injection it does not affect the RPM.The graph is similar to normal graph which was drawn in task 2. Very Less anomalies. The whole scatter plot is green in color and anomalies are very less. This graph is similar to RPM injection with speed data frame.

**No Injection:**



```
Anomalies in No Injection - RPM:
     RPM     value  RPM_Cluster  RPM_Outlier
1      0  0.000000            0           -1
3      0  0.000000            0           -1
5      0  0.000000            0           -1
8      0  0.000000            0           -1
10     0  0.000000            0           -1
12     0  0.000000            0           -1
658    1  0.995331            1           -1
662    1  0.996887            1           -1
667    1  0.996887            1           -1
671    1  0.996109            1           -1
672    1  0.997665            1           -1
676    1  0.996109            1           -1
678    1  1.000000            1           -1
679    1  0.996109            1           -1
681    1  0.996887            1           -1
686    1  0.995331            1           -1
```

Anomalies are found in the No Injection scenario in a number of scenarios where the RPM is between approximately 0 and value greater than 0.9. Speed Outlier = -1 is used to identify these anomalies. Compared to the FFM-injection it is almost similar to it. But the anomalies are very less compared to FFM almost equal to 10 anomalies. The scatter plot Fig 6.1 shows purple at top of the straight line at speed ==0.0 and value greater than 0.9 and remaining all are green. This graph is almost similar to the no injection drawn with speed data frame.

**Scatter Plots: -**

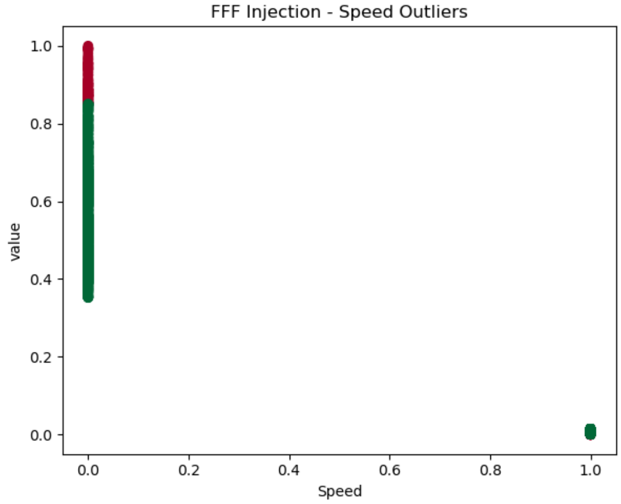


**Fig 1.1**

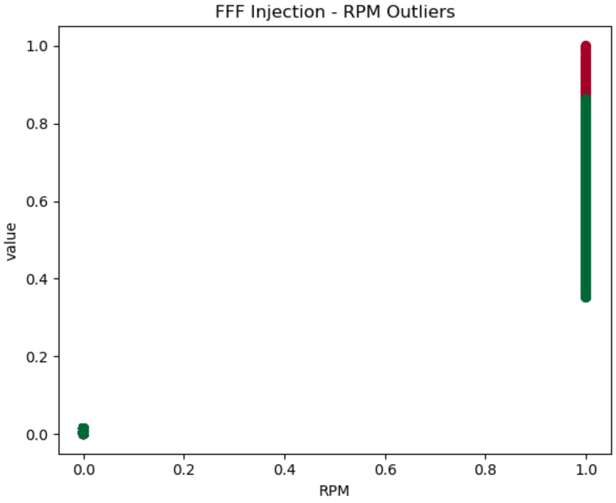


**Fig 4.1**

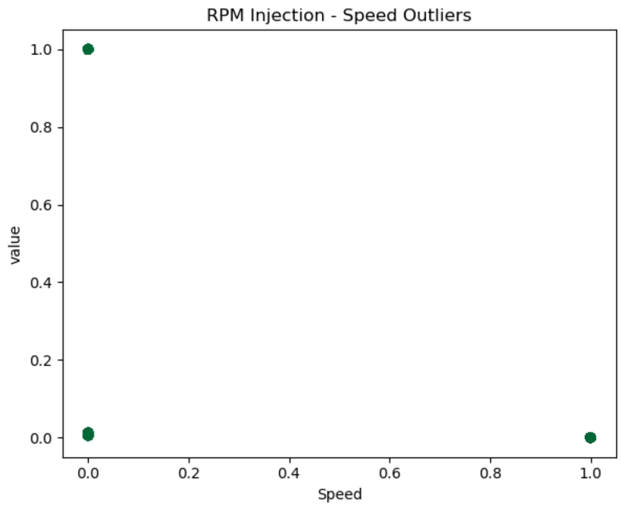


Fig 2.1

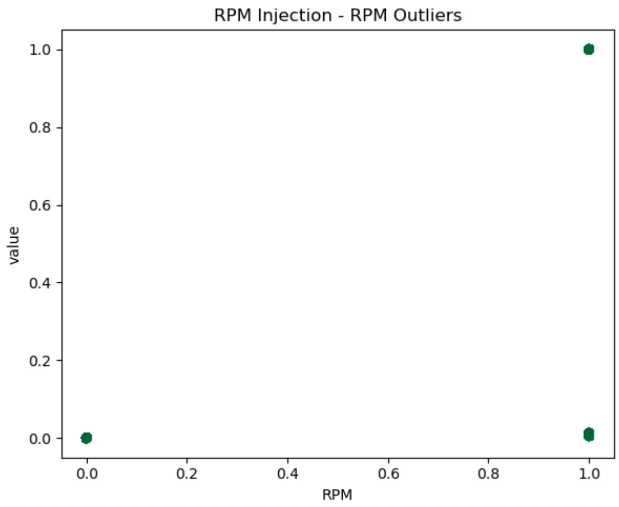


Fig 5.1

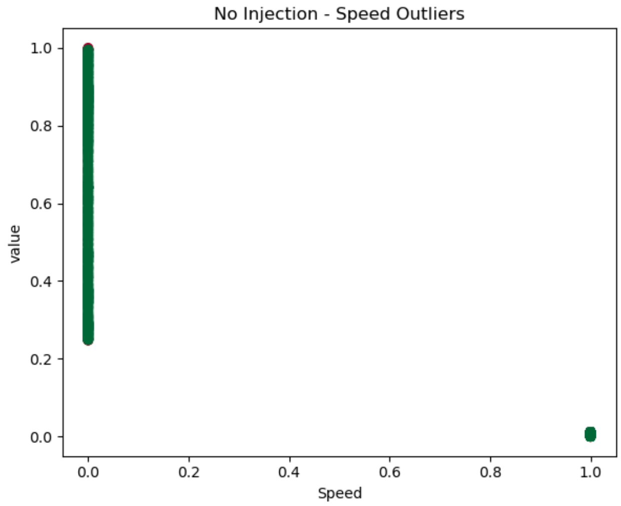


Fig 3.1

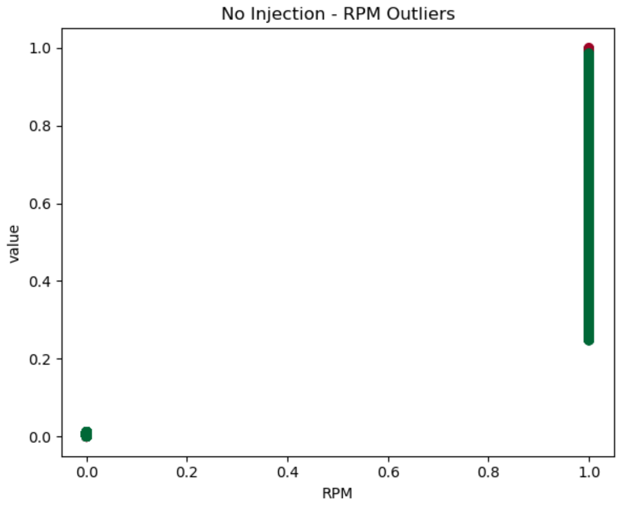


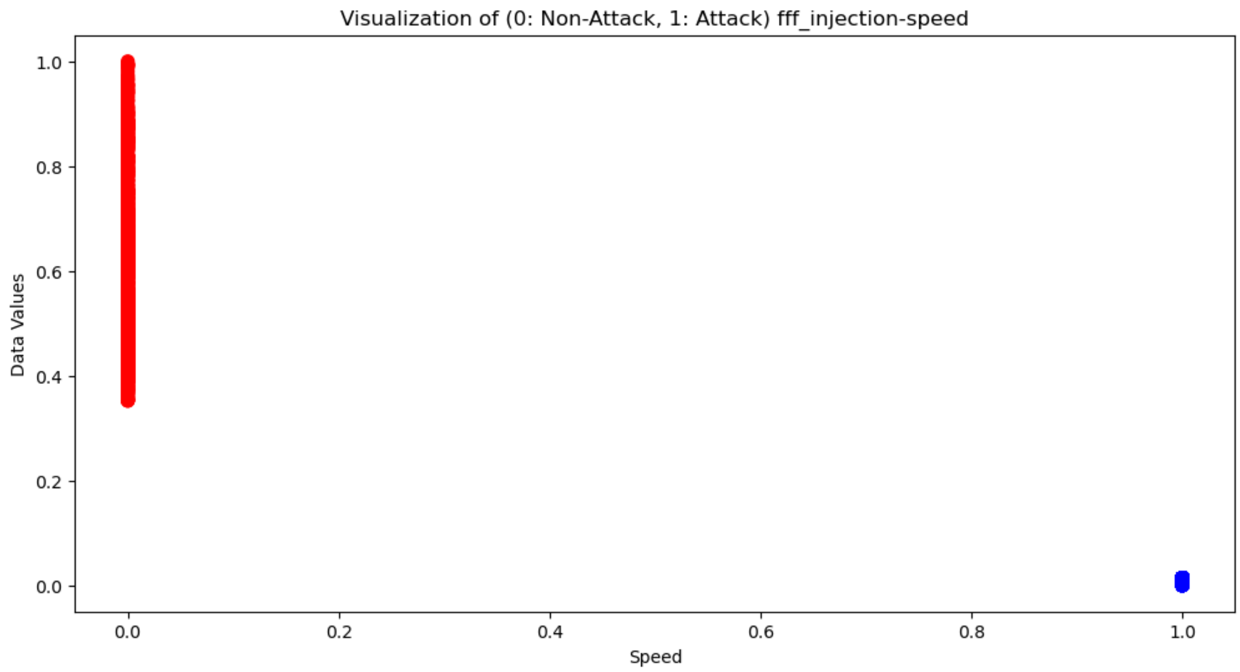
Fig 6.1

**Green -In liners**

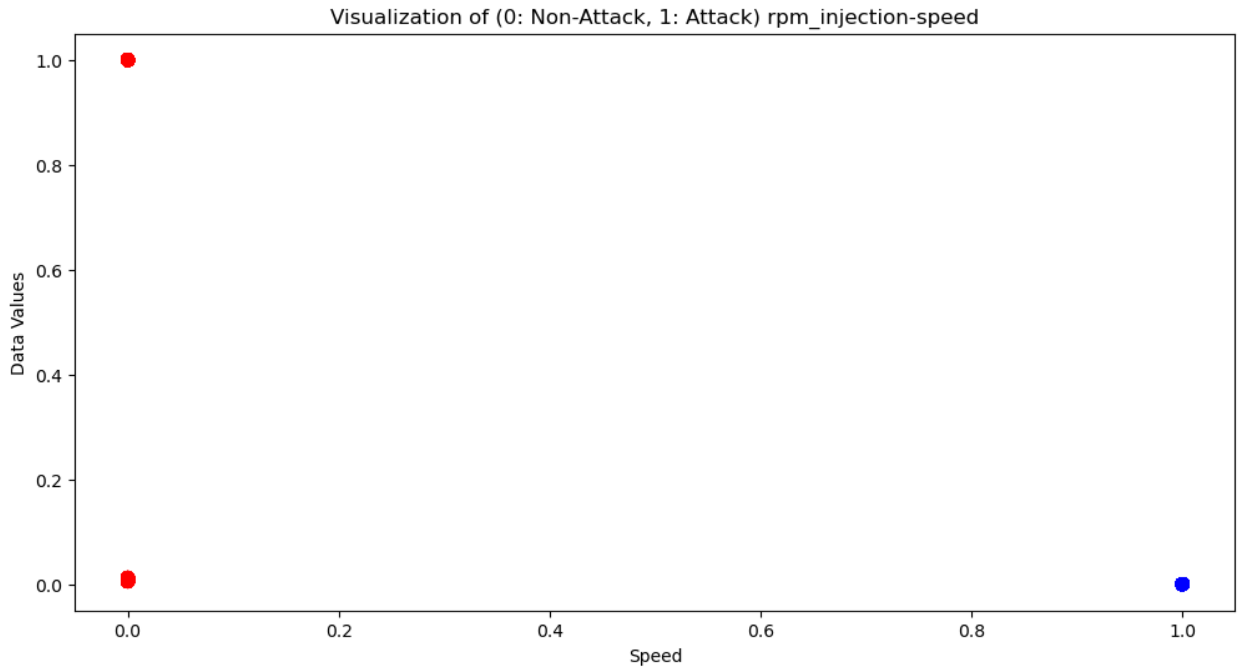
**Purple-Outliers**

**Task 4: -**

**Speed**
**Ffm_injection:-**
Actually it contains two clusters ,two centriods .One cluster contains no_attack ,most of the cluster is blue as youn can see below the centriod is almost approximetly equal to 0.0,0.0 and one cluster is red in color contains most of the data is attack

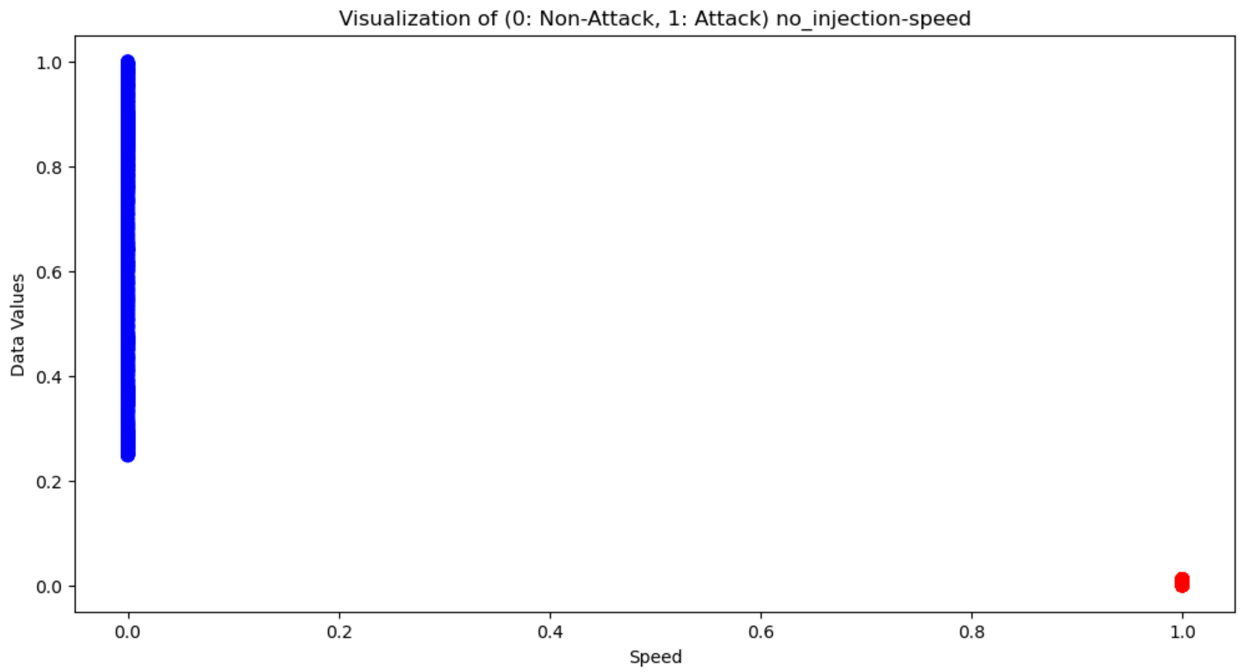


**Rpm_injection:-**



Most of the data is red as you can see.The hidden markov model formed clusters with three ,two red and ine blue.As blue indicates the no_attack and red indicates the attack.When thr value of speed is two low and data value is high .It is indicating red where attack==1 but it is an outlier .

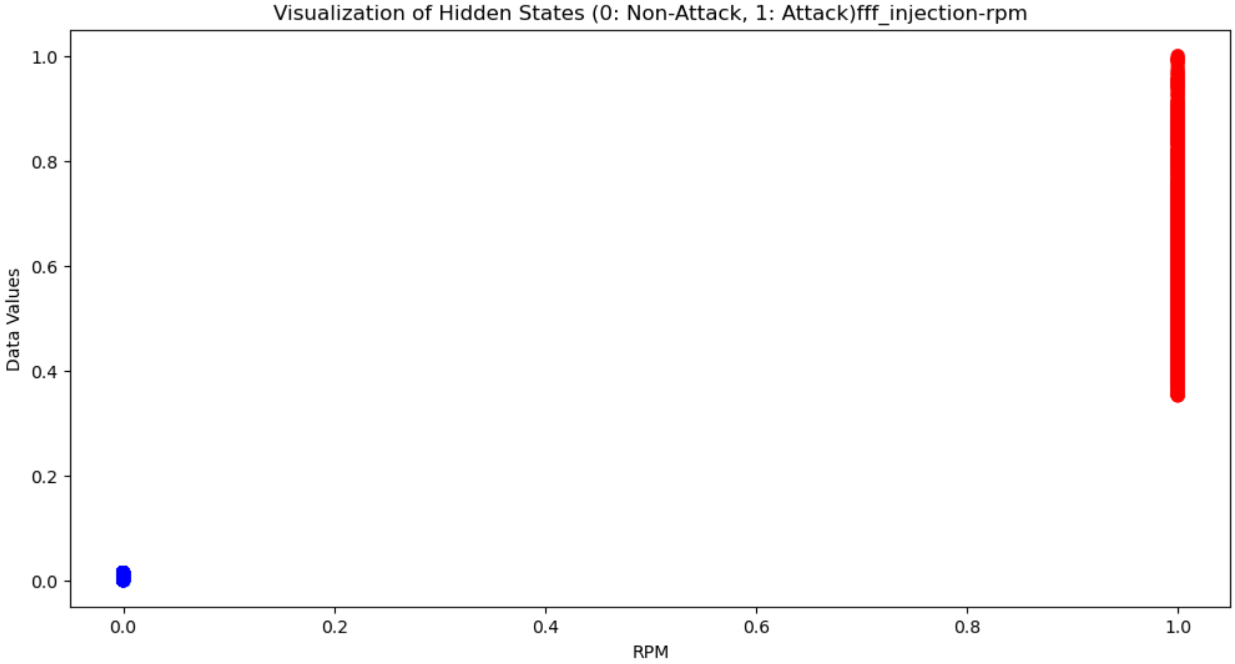
**No-injection: -**



As there is no injection of speed or RPM reading, the red dominance is very less in it and blue is high. The red also includes the outlier as speed is high and data value is low. The attack probability is too low in this case

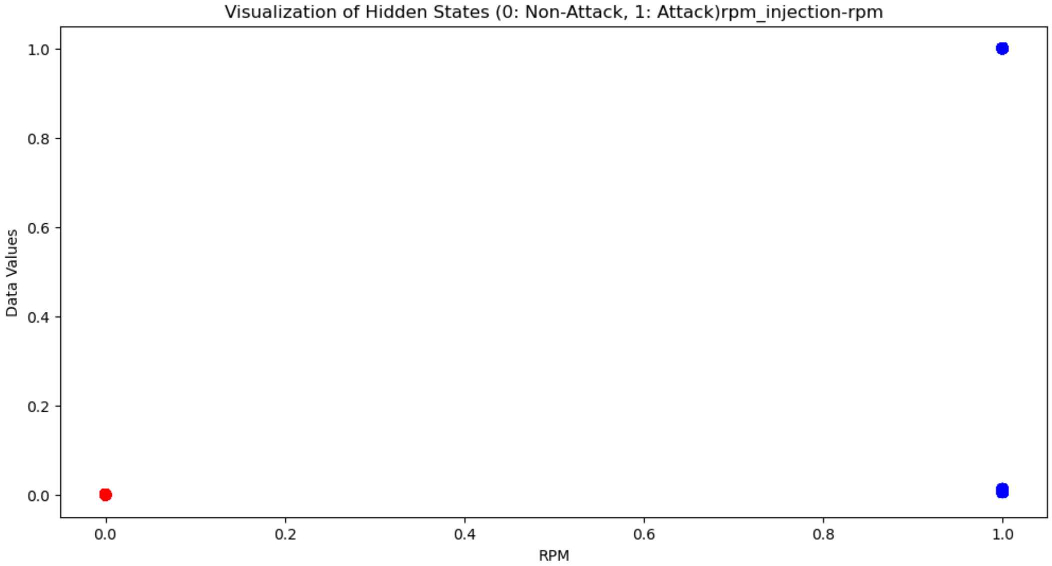
**RPM: -**

**FFF_injection: -**



It is similar to the fff_injection in speed scatter plot and when the RPM is too high. It is showing attack Probability is too high and blue is very less in this case. Indicating the attack is more when RPM is high and it contains two clusters, one of the clusters is pre dominant to attack in this case.

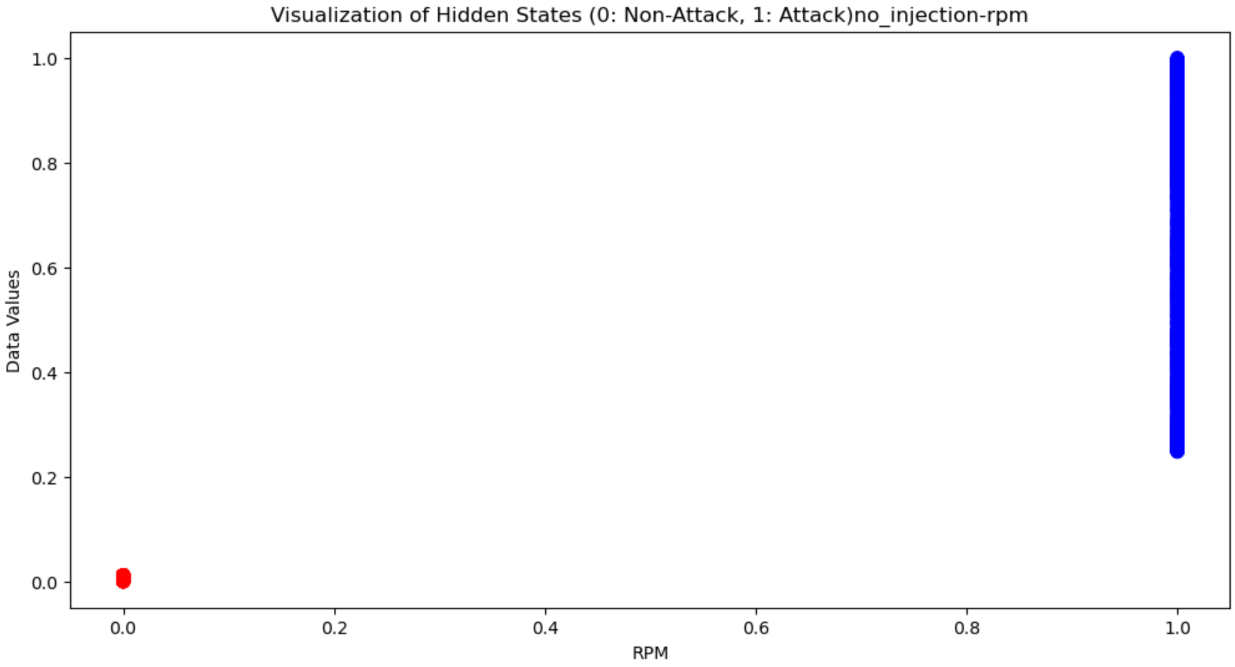
**RPM-injection: -**



The scatter plot indicates the attack is very less in this case and most of the clusters are blue in color indicating the nonattack and attack is possible when the data values are less, and rpm is also less.

**No_injection: -**

_



By using HMM, we predict the attack and non-attack. In this case the non-attack is more dominant as there is no injection in this case like speed RPM reading. The attack is when the data value are less and RPM is also less.

**Task 5: -**

We explored the field of unsupervised machine learning in this project, concentrating on using Hidden Markov Models (HMM) and clustering methods like K-Means and Isolation Forest to identify assaults on a Controller Area Network (CAN) bus. Unsupervised techniques have certain drawbacks even if they can be useful in situations involving unlabeled data.

The absence of ground truth labels in this situation is a significant drawback of unsupervised machine learning. It is difficult to assess clustering algorithms' performance properly in the absence of labeled data. Furthermore, unsupervised models could miss minor trends in the data, which could result in false positives or negatives when detecting attacks. Additionally, choosing the appropriate parameters, like the quantity of clusters or the threshold for anomaly detection, can be subjective and challenging. One cannot figure out the accuracy, True positive etc. unlike the supervised. The n_cluster variable in building the k-means is 2 as attack is 0 or 1 but if it not binary, it would be difficult to build the model. In the Isolation Forest algorithm, the biggest trouble is finding the Contamination value in it, first I gave the variable as 1, most of the data is outlier, then I gradually decreased and made to 0.01 and now most of the data is in linear. The major headache is Hidden Markov Model, I have used various methods to link between the features (speed or rpm) with the attack. The hidden states how they are linked to the attack output. Only by the visualization of I got the

Visualization is the only saviour in unsupervised algorithms. In order to optimize the performance of the unsupervised models, at least a portion of the dataset must have labeled data in order to be evaluated and adjusted. To improve the quality of data used in unsupervised machine learning, feature engineering and data pretreatment are essential. Maintaining an edge over changing attack patterns also requires constant observation and model upgrades.  I performed the data normalization also, normalized the value column in the range of [0,1] for better analysis. Contamination parameters also improved the model for me. Understanding the anomalies also helped to improve the model by looking at the scatter plot.

In conclusion, the issue of successfully identifying attacks on a CAN bus network requires a combination of supervised and unsupervised techniques, as well as a focus on feature engineering and data quality. Although this is a difficult to use unsupervised algorithm. If it is supervised, we can easily understand the model efficiency, whether its True positive, False Negative or Heatmap to get clarity on model. I understood the importance of label data and importance of clustering the similarity in data points and Isolated Forest algorithm.