Report

Task 5:

1. What were the shortcomings/limitations of using unsupervised ML on this problem?

Unsupervised learning(K-means and Isolation Forest) has several limitations.

a. In unsupervised learning, the labeled data is not present which makes model peformance evaluation challenging.

b. The assumption that similar data points should belong to the same cluster may not be true in the complex datasets.

c. In K-means, K value should be chosen very correctly. Else, if the k value is not chosen correctly the entire model may not give accurate results.

2. What could be done to make the models perform better?

Dimensionality reduction techniques like PCA can be applied before clustering to improve the accuracy of the model. Using multiple clustering algorithms and ensemble methods can also improve the results. Also, employing feature engineering techniques helps in creating more meaningful features.

3. What do you suggest doing to solve the problem?

I would suggest that refining data pre-processing methods are helpful, to ensure that the data is properly cleaned and prepared. Trying out various clustering algorithms and optimizing their parameters can help in achieving more accurate results. For unsupervised learning, visualizations and exploratory data analysis are important to understand the nature of the data. Finally, trying out the different models, tuning parameters, and comparing results with the other models can help in increasing the accuracy.

Task 2:

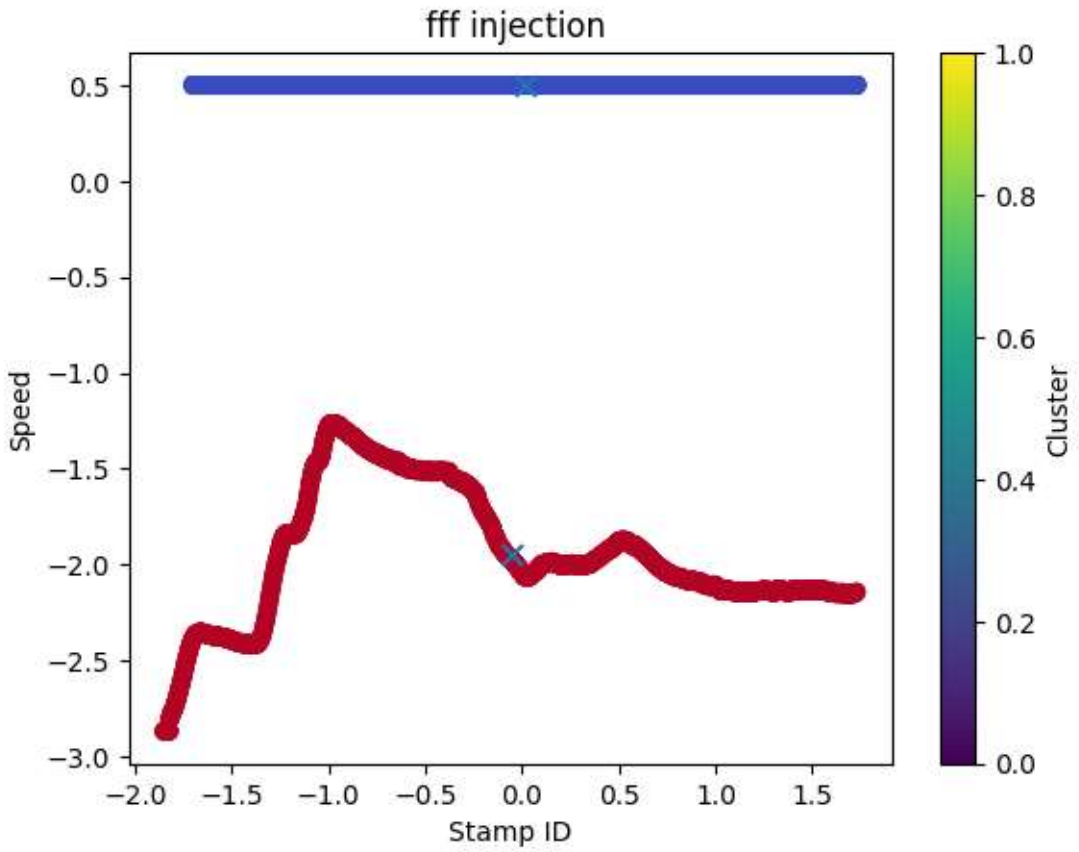
Outputs:

```
[280] # printing the centroids
      print("f_inj_Speed_centroid:",f_inj_Speed_centroid)
      print("f_inj_RPM_centroid:",f_inj_RPM_centroid)
      print("r_inj_Speed_centroid:",r_inj_Speed_centroid)
      print("r_inj_RPM_centroid:",r_inj_RPM_centroid)
      print("no_inj_Speed_centroid:",no_inj_Speed_centroid)
      print("no_inj_RPM_centroid:",no_inj_RPM_centroid)
```

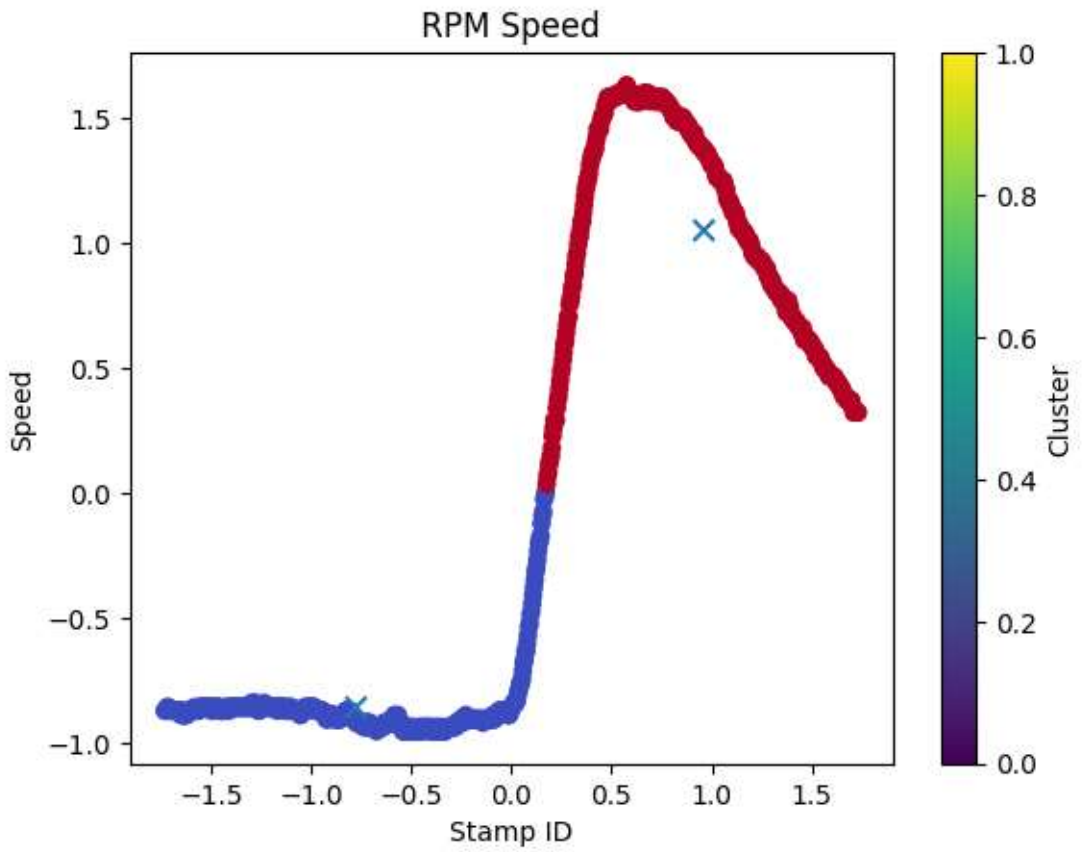
```
f_inj_Speed_centroid: [[ 0.0147652   0.50089906]
 [-0.05743875 -1.94856666]]
f_inj_RPM_centroid: [[-0.86156837  0.61731336]
 [ 0.8489947  -0.60830435]]
r_inj_Speed_centroid: [[-0.77899649 -0.85721514]
 [ 0.95305006  1.0487455 ]]
r_inj_RPM_centroid: [[-0.26694906 -1.55842462]
 [ 0.10991432  0.64166953]]
no_inj_Speed_centroid: [[-1.185739   -0.93419799]
 [ 0.54632192  0.43042593]]
no_inj_RPM_centroid: [[-0.62515706  0.53833345]
 [ 1.10689583 -0.95316687]]
```

0s completed at 11:04 PM

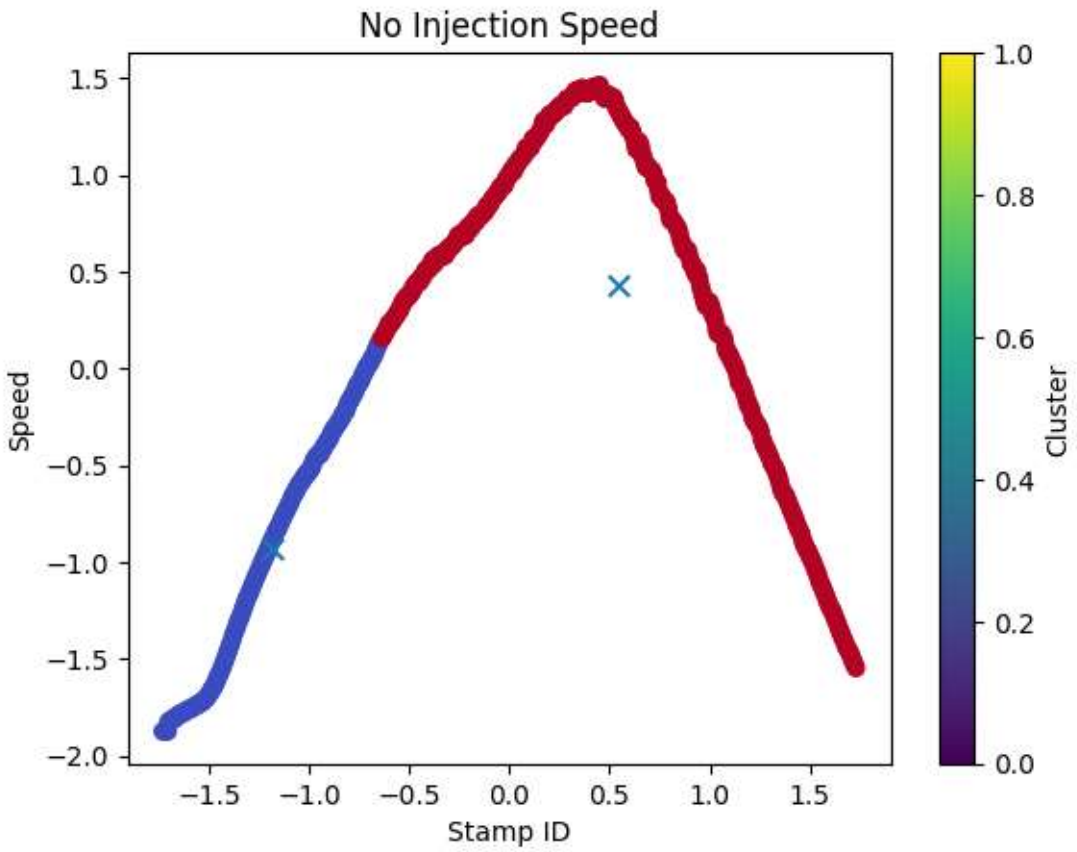
1. FFF Injection (Stamp ID vs Speed)



2. RPM Speed (Stamp Id Vs Speed)

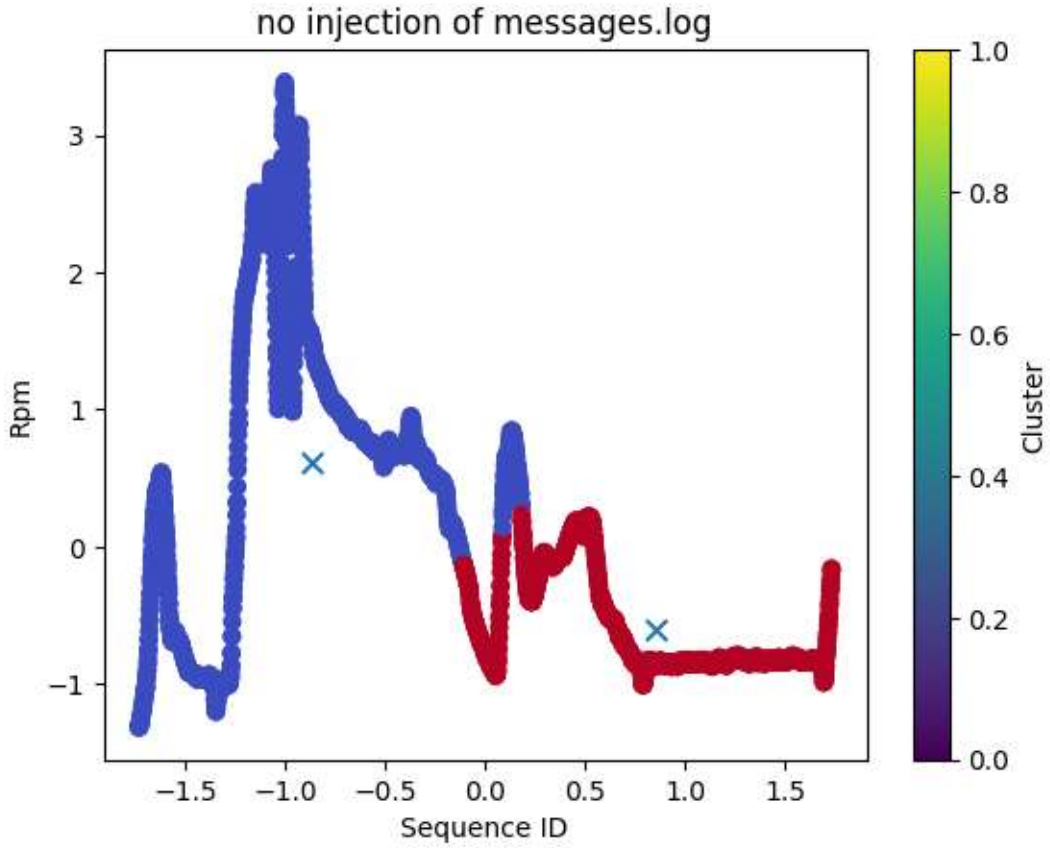


3. No Injection (Stamp ID vs Speed)

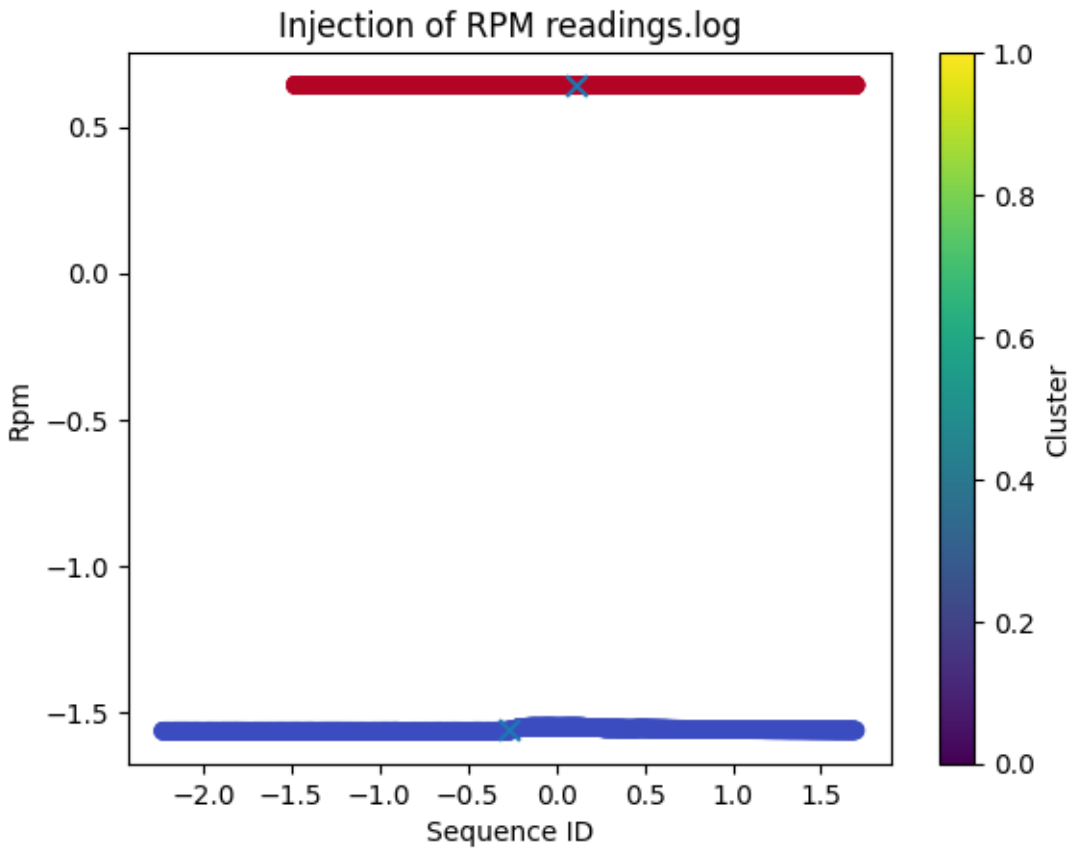


Plots for RPM:

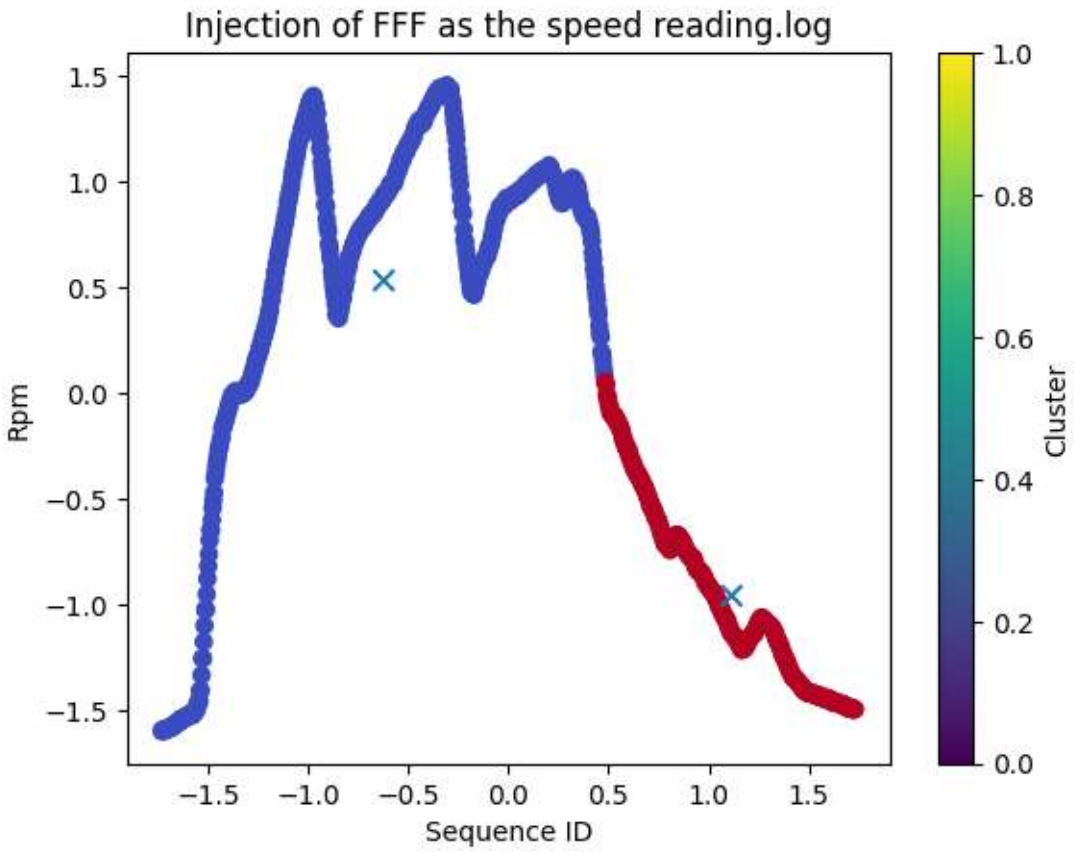
1. No Injection of messages



2. Injection of RPM readings.log



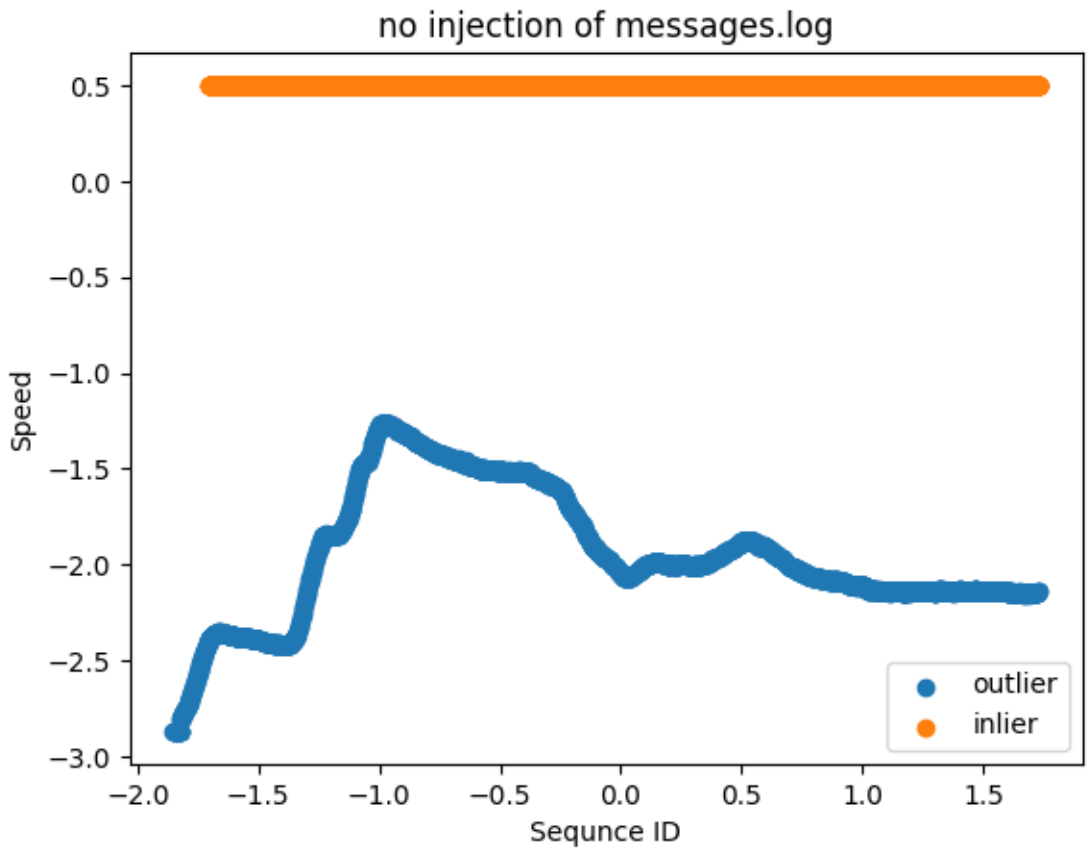
3. Injection of FFF as the speed reading.log



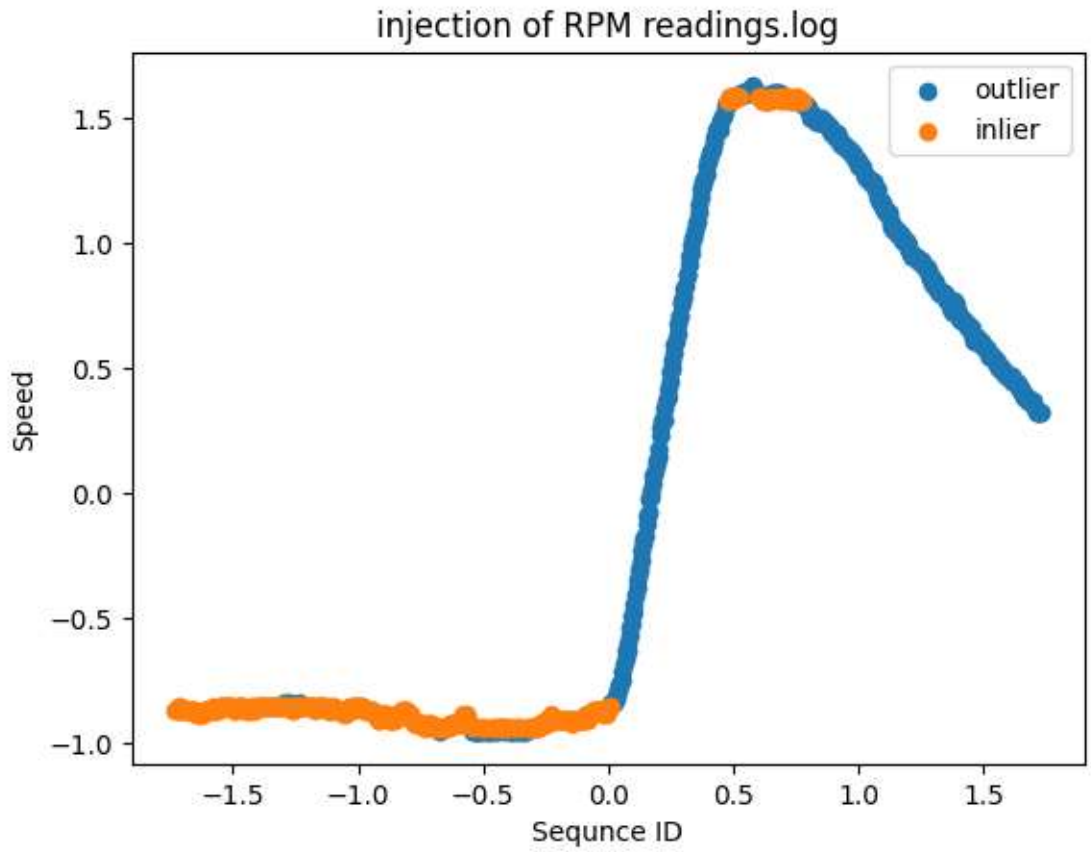
4.

Task 3:

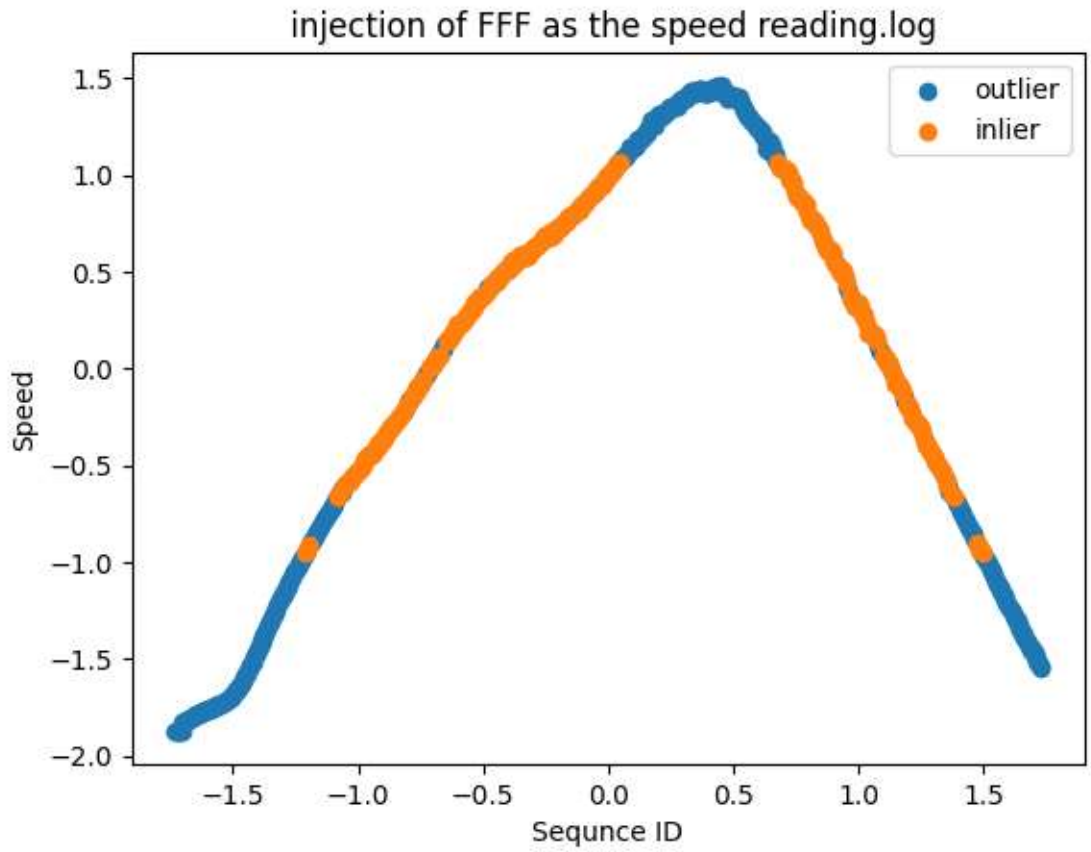
1. no injection of messages.log



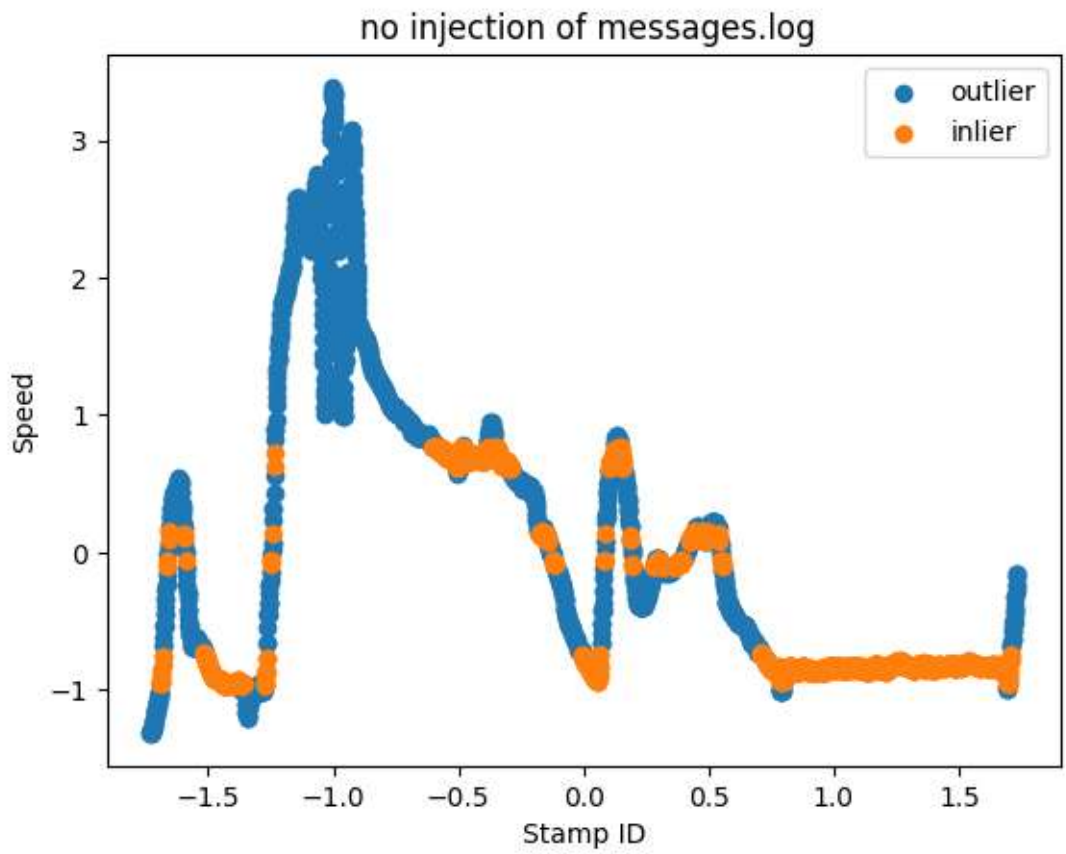
2. injection of RPM readings.log



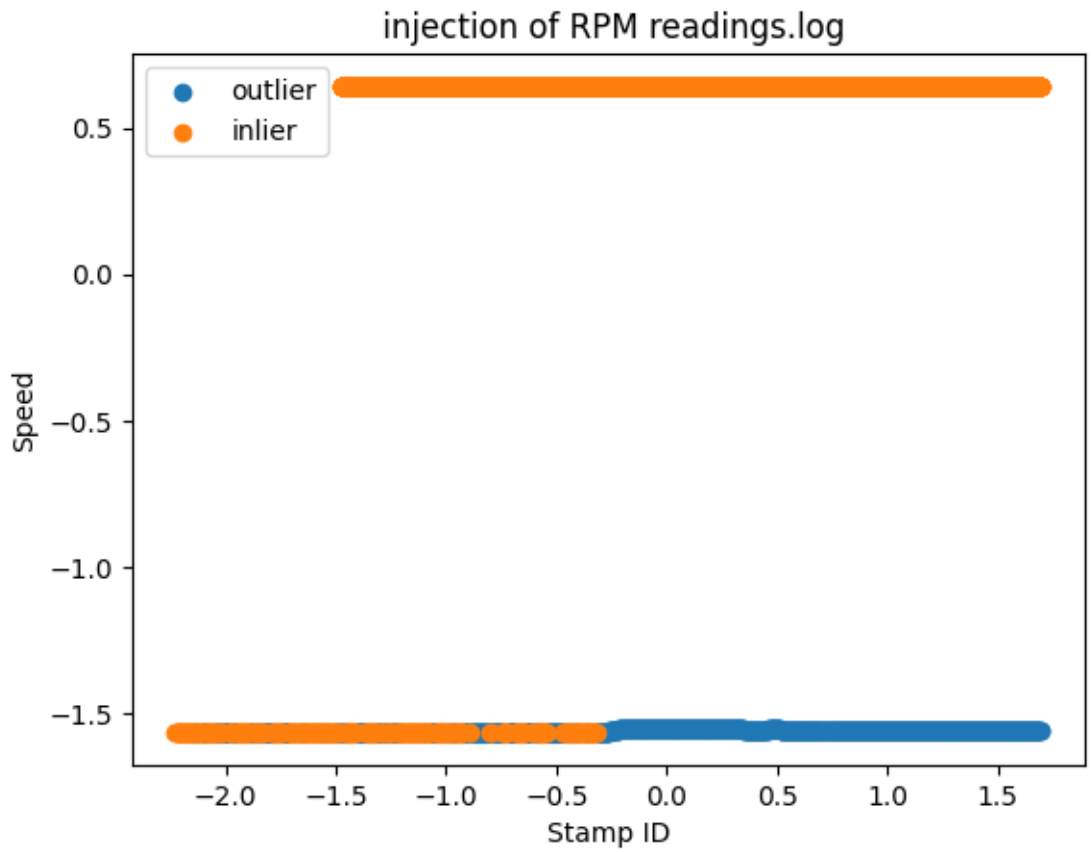
FFF injection in the speed reading.log



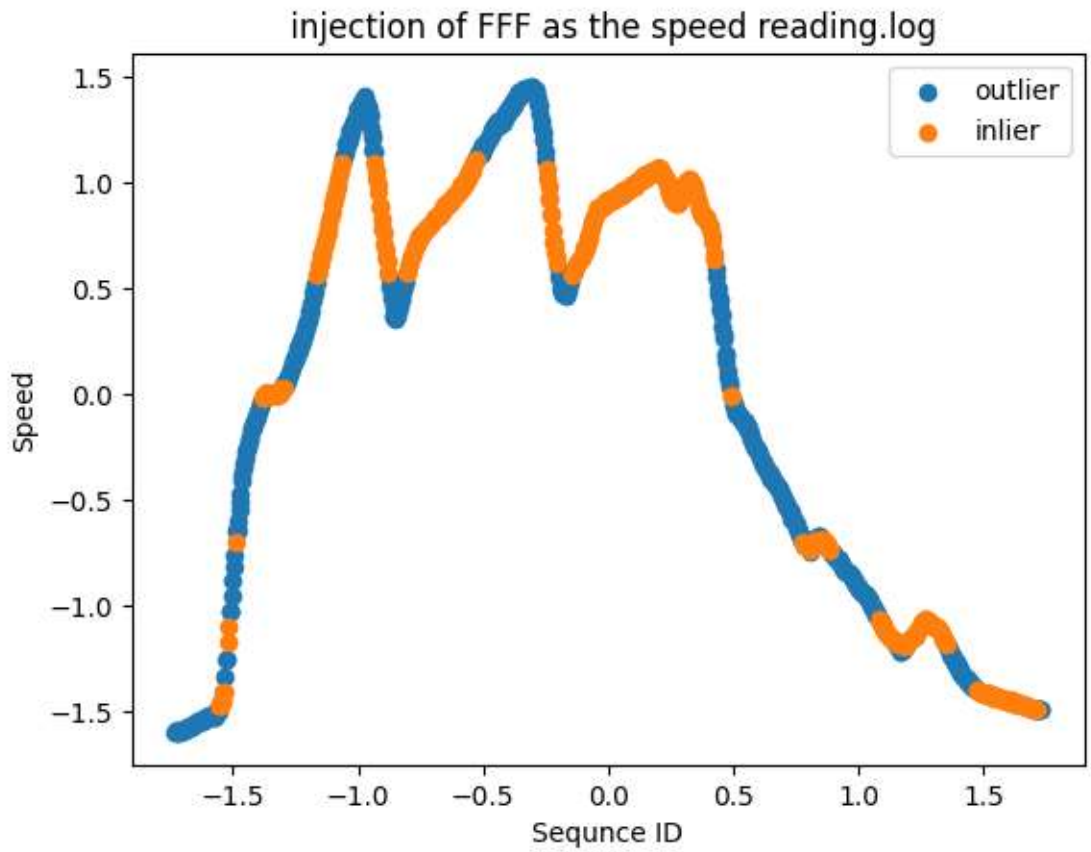
No injection in messages.log


no injection of messages.log
outlier
inlier
Speed
3
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Stamp ID

RPM Injection readings.log:



FFF injection of speed in reading.log:



TASK 5 :
Hidden Markov Model:

```
[277] # Hidden Markov Model for fff Injection, RPM
      HMModel(f_inj_RPM)

      Actual Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Predicted Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,
      Accuracy: 0.5797379354426334

[272] # Hidden Markov Model for fff Injection, Speed
      HMModel(f_inj_Speed)

      Actual Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Predicted Value: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
      Accuracy: 0.0

[273] # Hidden Markov Model for RPM Injection, RPM
      HMModel(r_inj_RPM)

      Actual Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Predicted Value: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
      Accuracy: 0.0
```

```
[274] # Hidden Markov Model for RPM Injection, Speed
      HMModel(r_inj_Speed)

      Actual Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Predicted Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Accuracy: 0.5367647058823529

[275] # Hidden Markov Model for NO Injection RPM
      HMModel(no_inj_RPM)

      Actual Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Predicted Value: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
      Accuracy: 0.6093418259023354

[276] # Hidden Markov Model for NO Injection Speed
      HMModel(no_inj_Speed)

      Actual Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Predicted Value: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      Accuracy: 0.7906647807637907

✓ 0s completed at 11:04 PM
```

What I did in 5 Tasks:

Task 1: Data Preparation

1. Loaded three text files containing bus logs for different scenarios and I dropped the 'Attack' column as it is not used.

Task 2: K-means Clustering

- Implemented K-means clustering separately for six datasets (Speed and RPM, three scenarios each), plotted fitted clusters using scatter plots, and provided centroids for each.

Task 3: Isolation Forest Algorithm

- Implemented Isolation Forest separately for the six datasets, created scatter plots displaying the fitted isolation forest clusters, and recorded anomalous data points identified by the algorithm.

Task 4: Hidden Markov Models

- Constructed Hidden Markov Models to predict Attack vs. no Attack.

Task 5: Discussion

- Reflected on the limitations of unsupervised ML in this context, proposed methods for better model performance, and suggested refining data pre-processing and exploring alternative clustering approaches.