EP2300 Project - Network Management Estimating Conformance to Service Level Agreements (SLAs) using Machine Learning

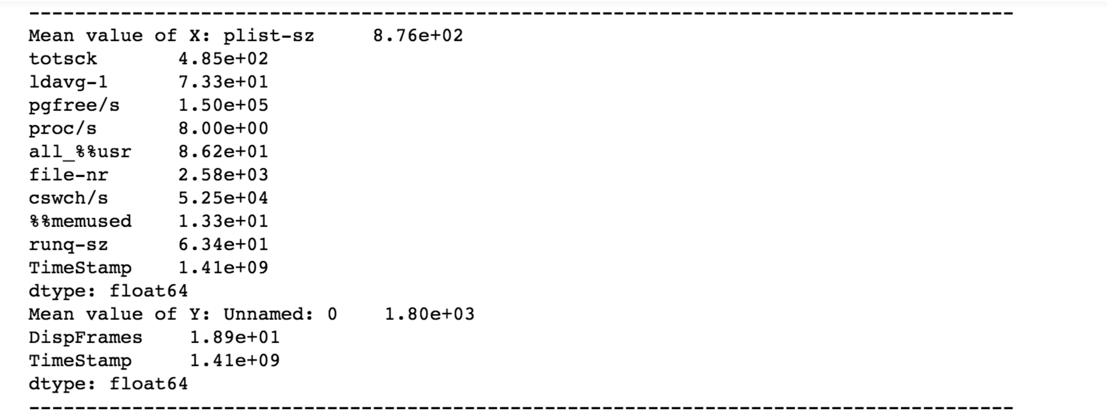
Jiani Jiang <jianij@kth.se>

Task I - Data Exploration:

1. Compute the following statistics for each feature of X and target of Y: mean, maximum, minimum, 25th percentile, 90th percentile, and standard deviation. Give no more than two digits after decimal point.

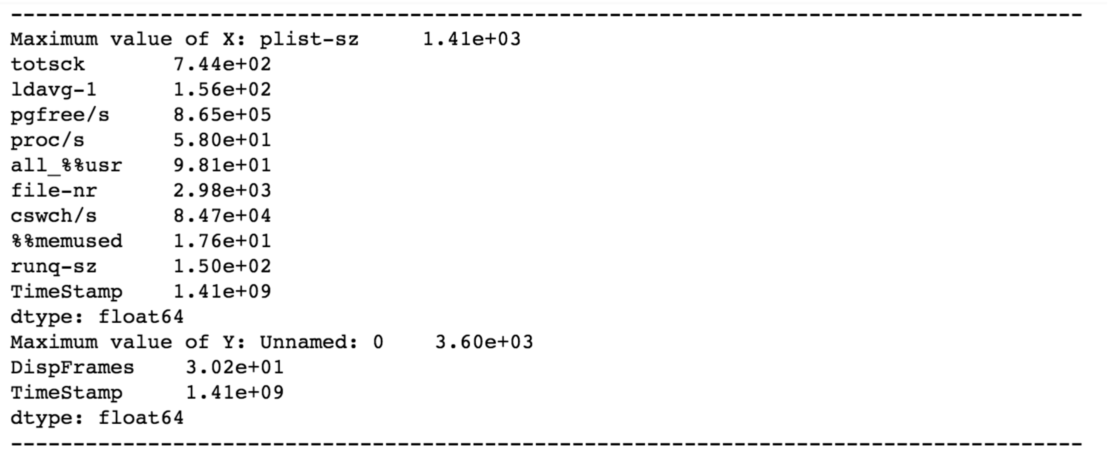
Answer:

Mean values of X and Y:



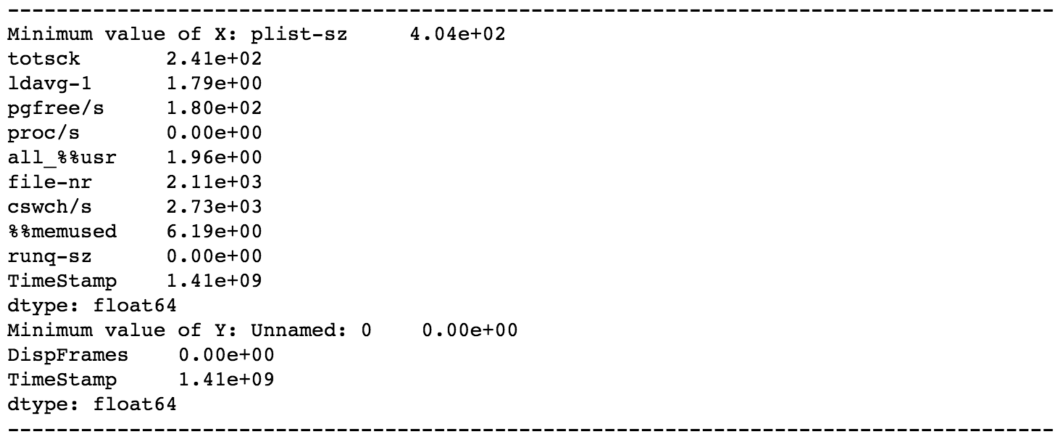
```
-------------------------------------------------------------------------------------
Mean value of X: plist-sz     8.76e+02
totsck       4.85e+02
ldavg-1      7.33e+01
pgfree/s     1.50e+05
proc/s       8.00e+00
all_%%usr    8.62e+01
file-nr      2.58e+03
cswch/s      5.25e+04
%%memused    1.33e+01
runq-sz      6.34e+01
TimeStamp    1.41e+09
dtype: float64
Mean value of Y: Unnamed: 0    1.80e+03
DispFrames    1.89e+01
TimeStamp     1.41e+09
dtype: float64
-------------------------------------------------------------------------------------
```

Maximum values of X and Y:



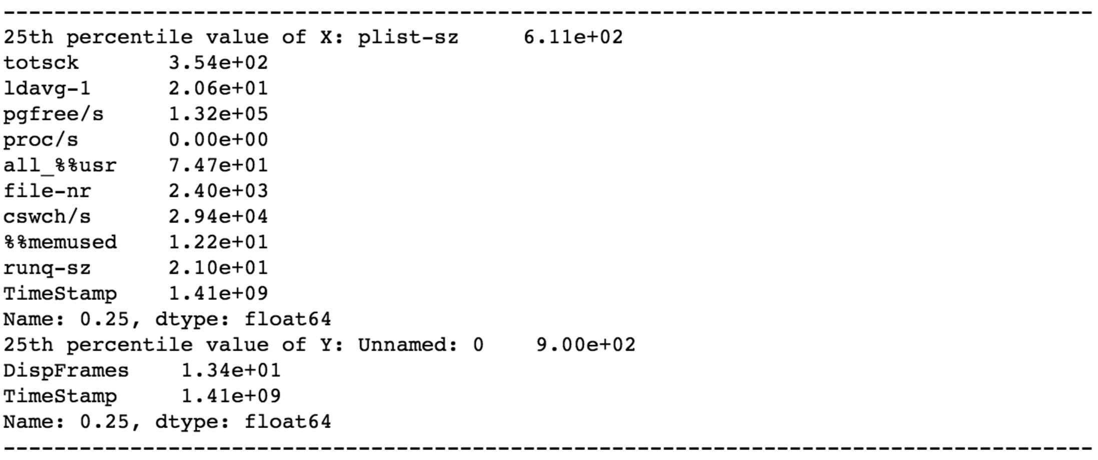
```
-------------------------------------------------------------------------------------
Maximum value of X: plist-sz     1.41e+03
totsck       7.44e+02
ldavg-1      1.56e+02
pgfree/s     8.65e+05
proc/s       5.80e+01
all_%%usr    9.81e+01
file-nr      2.98e+03
cswch/s      8.47e+04
%%memused    1.76e+01
runq-sz      1.50e+02
TimeStamp    1.41e+09
dtype: float64
Maximum value of Y: Unnamed: 0    3.60e+03
DispFrames    3.02e+01
TimeStamp     1.41e+09
dtype: float64
-------------------------------------------------------------------------------------
```

Minimum values of X and Y:



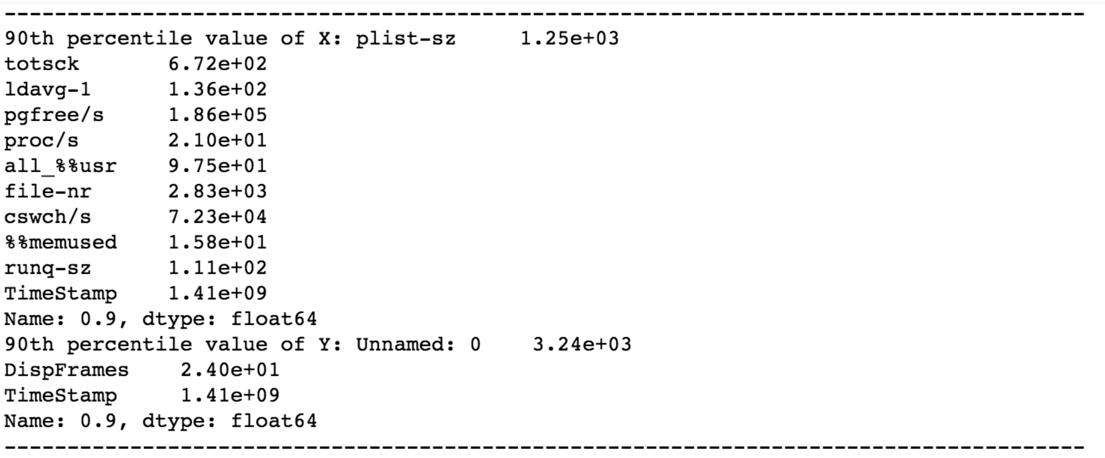
```
-------------------------------------------------------------------------------------
Minimum value of X: plist-sz     4.04e+02
totsck       2.41e+02
ldavg-1      1.79e+00
pgfree/s     1.80e+02
proc/s       0.00e+00
all_%%usr    1.96e+00
file-nr      2.11e+03
cswch/s      2.73e+03
%%memused    6.19e+00
runq-sz      0.00e+00
TimeStamp    1.41e+09
dtype: float64
Minimum value of Y: Unnamed: 0    0.00e+00
DispFrames    0.00e+00
TimeStamp     1.41e+09
dtype: float64
-------------------------------------------------------------------------------------
```

25th percentile value of X and Y:



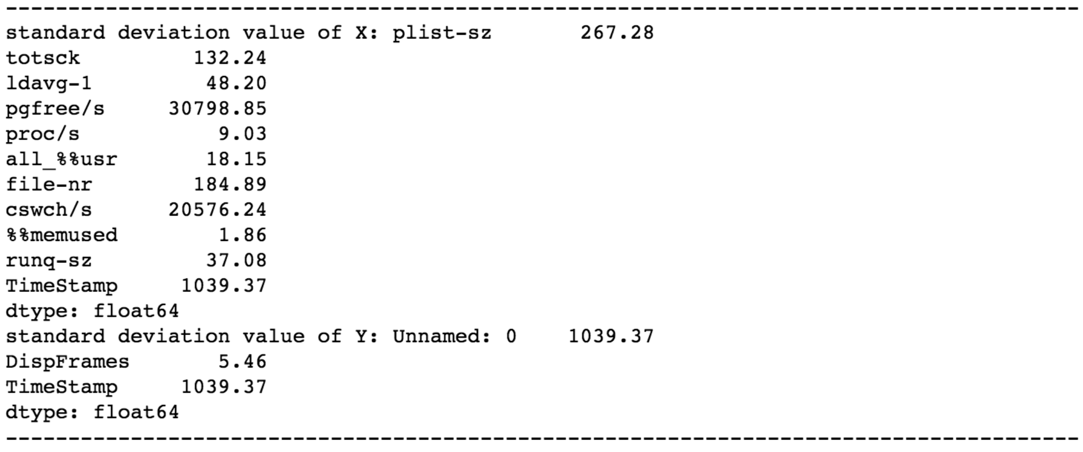
```
-----------------------------------------------------------------------------------
25th percentile value of X: plist-sz     6.11e+02
totsck       3.54e+02
ldavg-1      2.06e+01
pgfree/s     1.32e+05
proc/s       0.00e+00
all_%%usr    7.47e+01
file-nr      2.40e+03
cswch/s      2.94e+04
%%memused    1.22e+01
runq-sz      2.10e+01
TimeStamp    1.41e+09
Name: 0.25, dtype: float64
25th percentile value of Y: Unnamed: 0    9.00e+02
DispFrames    1.34e+01
TimeStamp     1.41e+09
Name: 0.25, dtype: float64
-----------------------------------------------------------------------------------
```

90th percentile value of X and Y:



```
-----------------------------------------------------------------------------------
90th percentile value of X: plist-sz     1.25e+03
totsck       6.72e+02
ldavg-1      1.36e+02
pgfree/s     1.86e+05
proc/s       2.10e+01
all_%%usr    9.75e+01
file-nr      2.83e+03
cswch/s      7.23e+04
%%memused    1.58e+01
runq-sz      1.11e+02
TimeStamp    1.41e+09
Name: 0.9, dtype: float64
90th percentile value of Y: Unnamed: 0    3.24e+03
DispFrames    2.40e+01
TimeStamp     1.41e+09
Name: 0.9, dtype: float64
-----------------------------------------------------------------------------------
```

standard deviation value of X and Y:



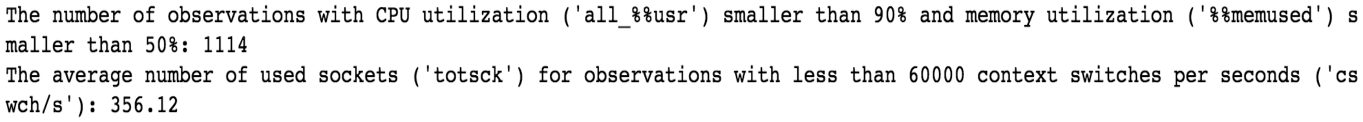
```
-----------------------------------------------------------------------------------
standard deviation value of X: plist-sz       267.28
totsck         132.24
ldavg-1         48.20
pgfree/s     30798.85
proc/s           9.03
all_%%usr       18.15
file-nr        184.89
cswch/s      20576.24
%%memused        1.86
runq-sz         37.08
TimeStamp     1039.37
dtype: float64
standard deviation value of Y: Unnamed: 0    1039.37
DispFrames       5.46
TimeStamp     1039.37
dtype: float64
-----------------------------------------------------------------------------------
```

2. Compute the following quantities of X:

(a) The number of observations with CPU utilization (“all %%usr”) smaller than 90% and memory utilization (“%%memused”) smaller than 50%;

(b) The average number of used sockets (“totsck”) for observations with less than 60000 context switches per seconds (“cswch/s”).

Answer:



```
The number of observations with CPU utilization ('all_%%usr') smaller than 90% and memory utilization ('%%memused') s
maller than 50%: 1114
The average number of used sockets ('totsck') for observations with less than 60000 context switches per seconds ('cs
wch/s'): 356.12
```

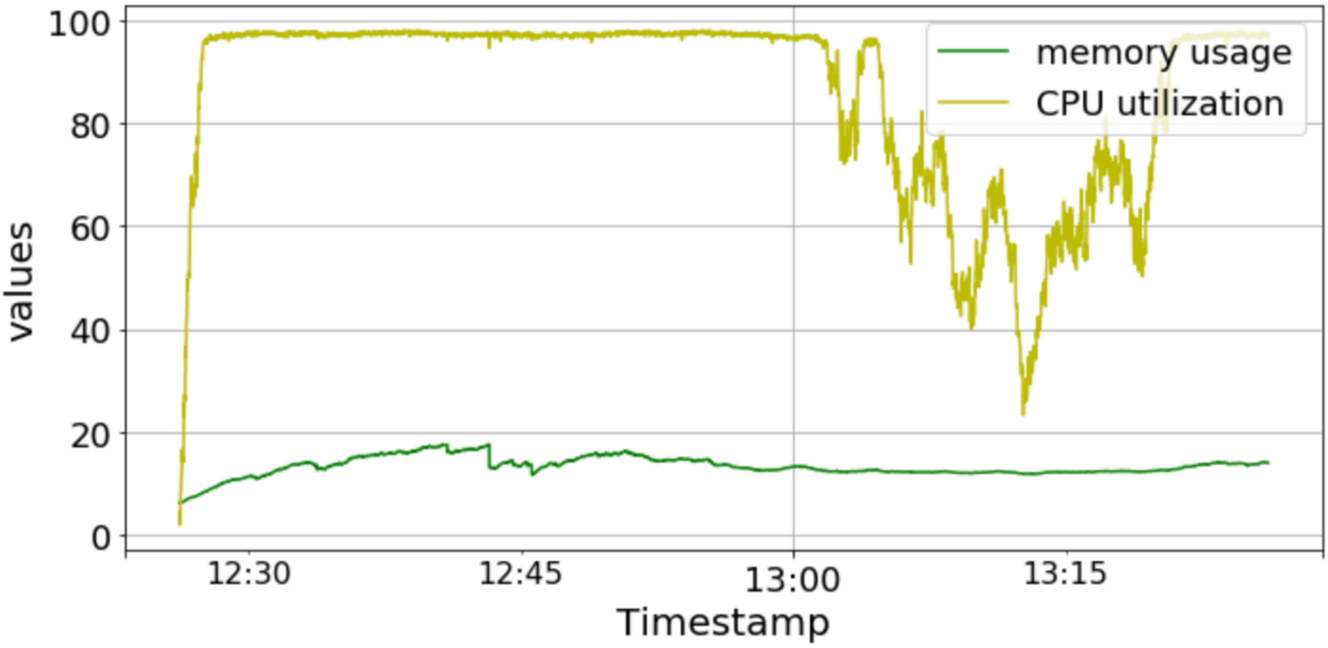
3. Produce the following plots:

(a) Time series of memory usage (“%%memused”) and CPU utilization (“all %%usr”), both curves in a single plot. Box plot of both features in a single plot.

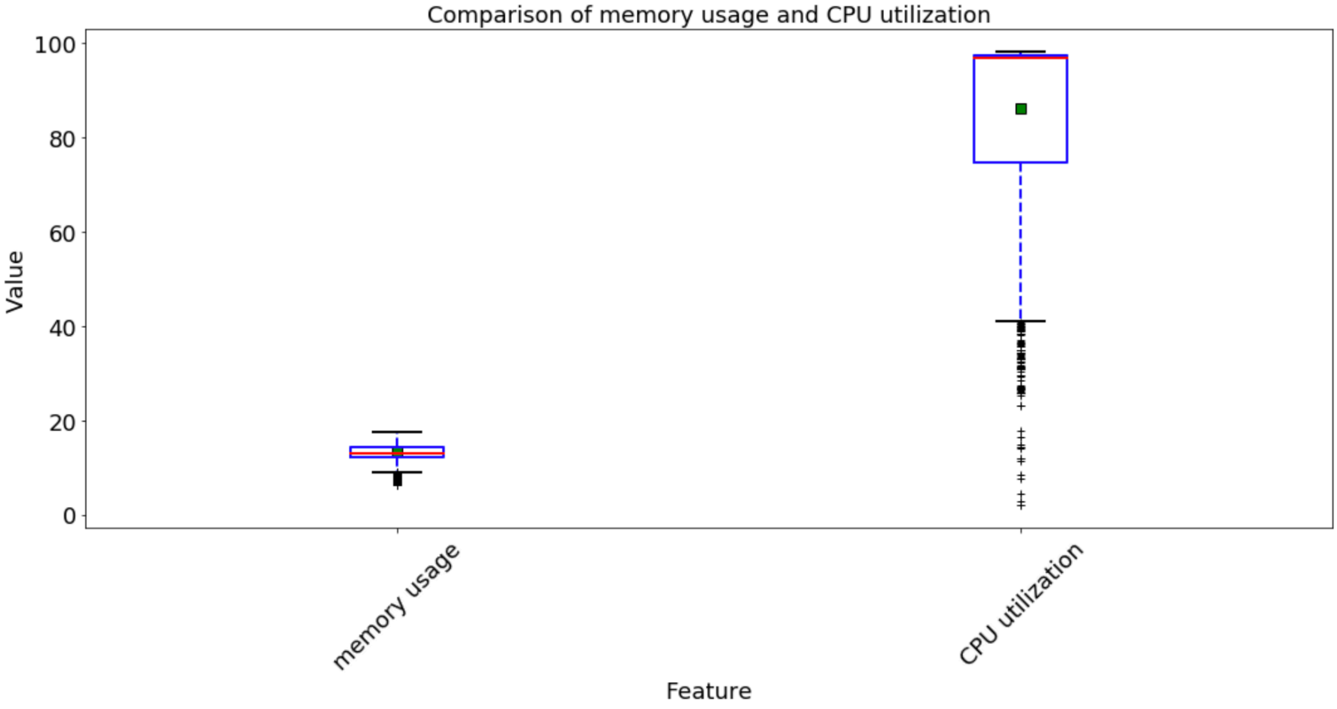
(b) Density plots of memory usage (“%%memused”) and CPU utilization (“all %%usr”), Histograms of both these features (choose a bin size of 1%), four plots in all.

Answer:

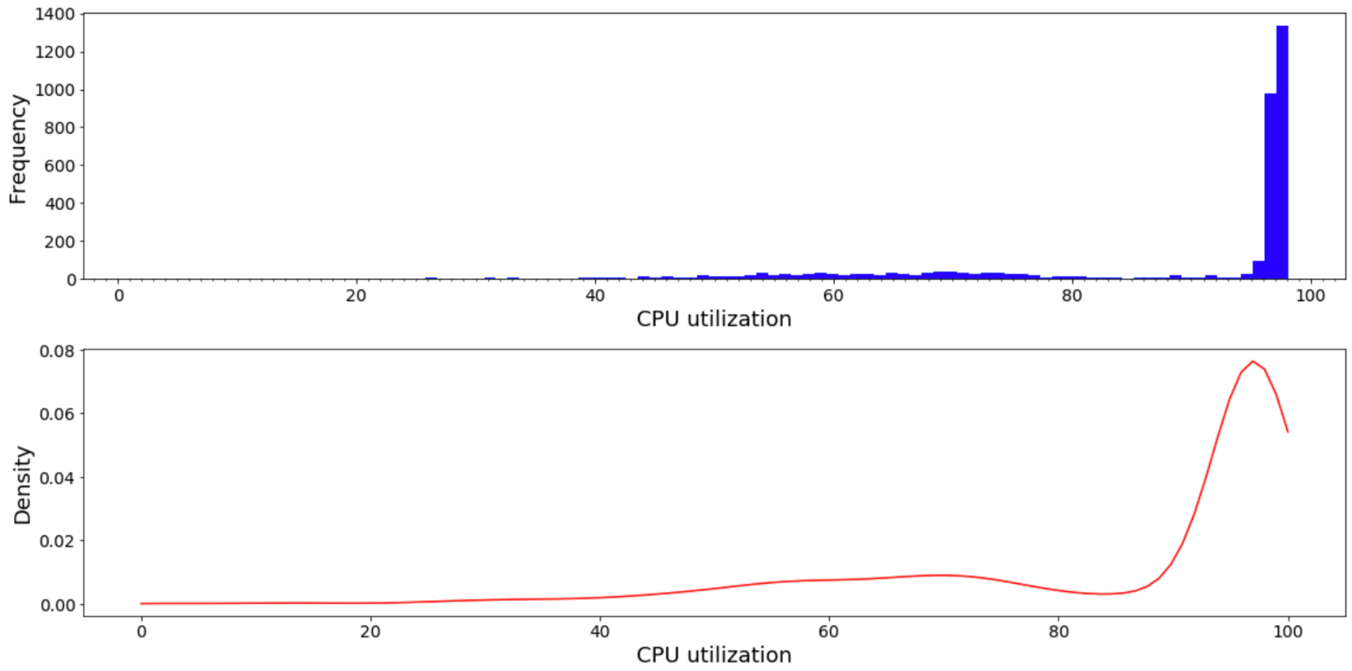
(a) Time series of memory usage ("%%memused") and CPU utilization ("all_%%usr"):



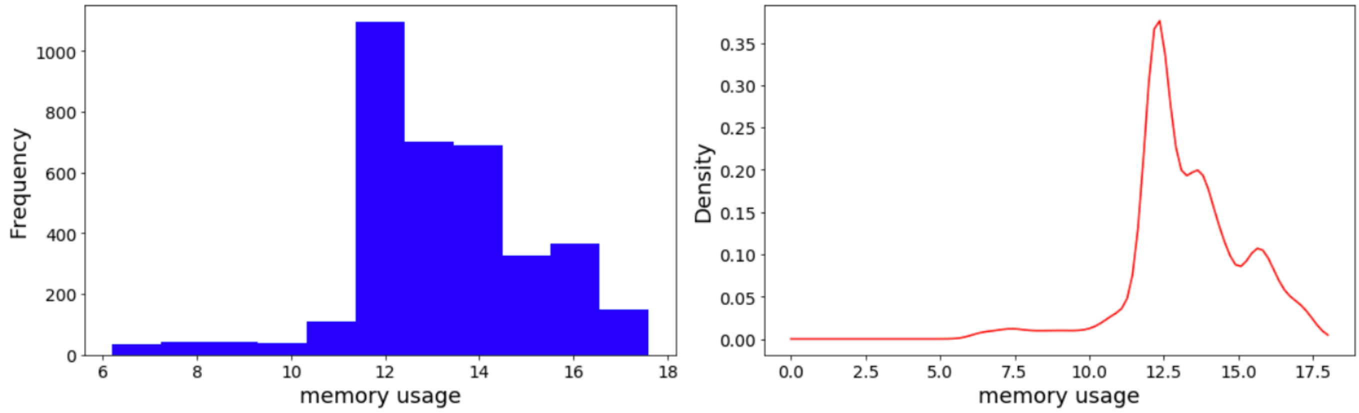
Box plot of both features in a single plot:



(b) Density plot and histogram plot of CPU utilization



Density plot and histogram plot of memory usage

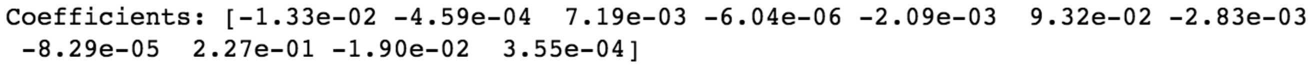


Task II - Estimating Service Metrics from Device Statistics:

1. Evaluate the Accuracy of Service Metric Estimation.

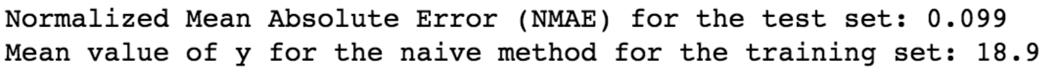
Answer:

(a)



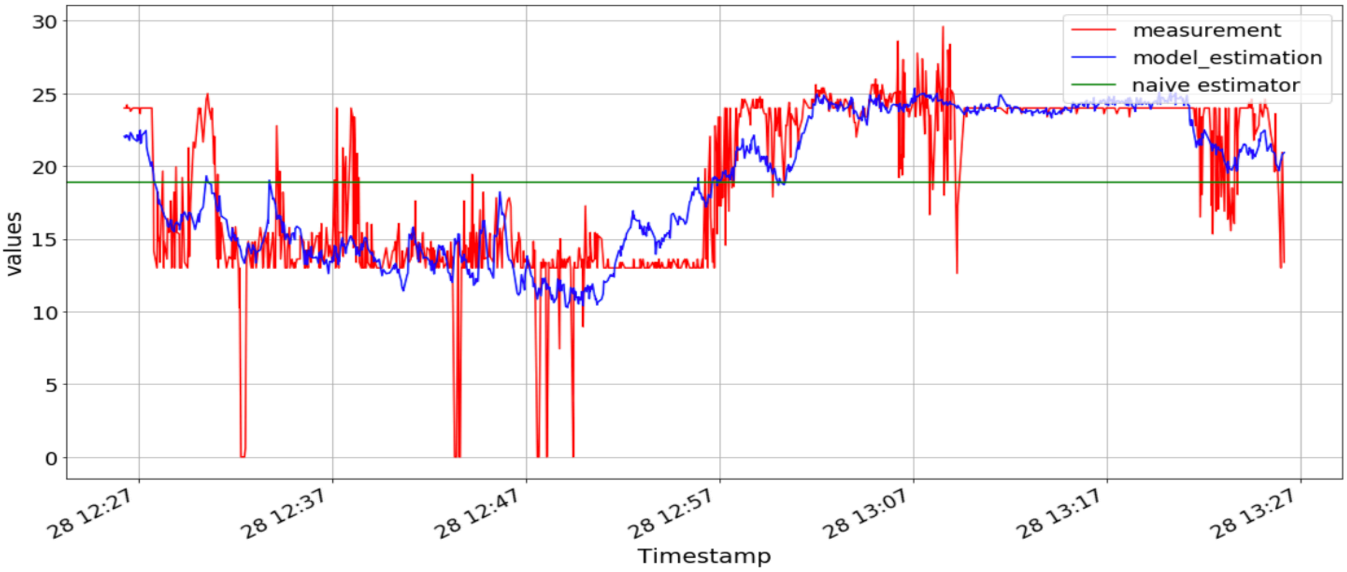
```
Coefficients: [-1.33e-02 -4.59e-04  7.19e-03 -6.04e-06 -2.09e-03  9.32e-02 -2.83e-03
 -8.29e-05  2.27e-01 -1.90e-02  3.55e-04]
```

(b)

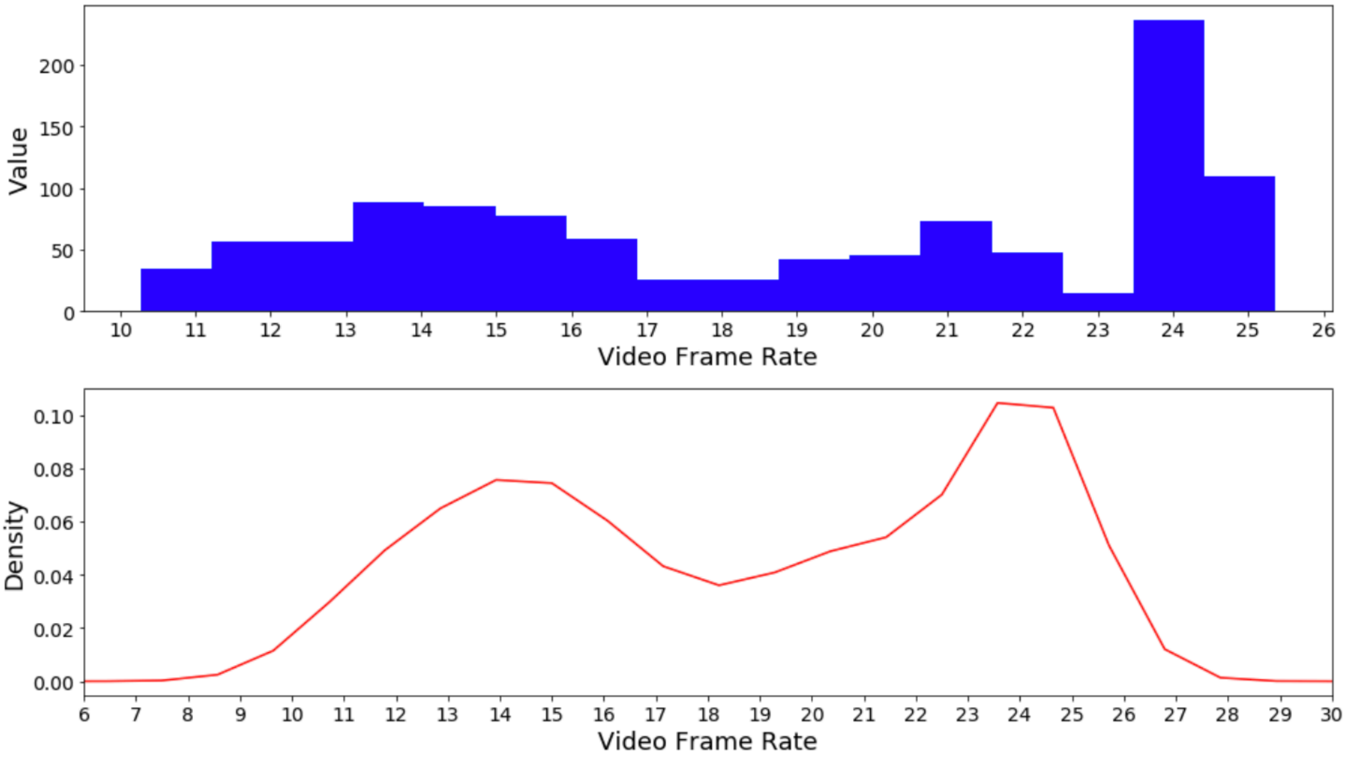


```
Normalized Mean Absolute Error (NMAE) for the test set: 0.099
Mean value of y for the naive method for the training set: 18.9
```

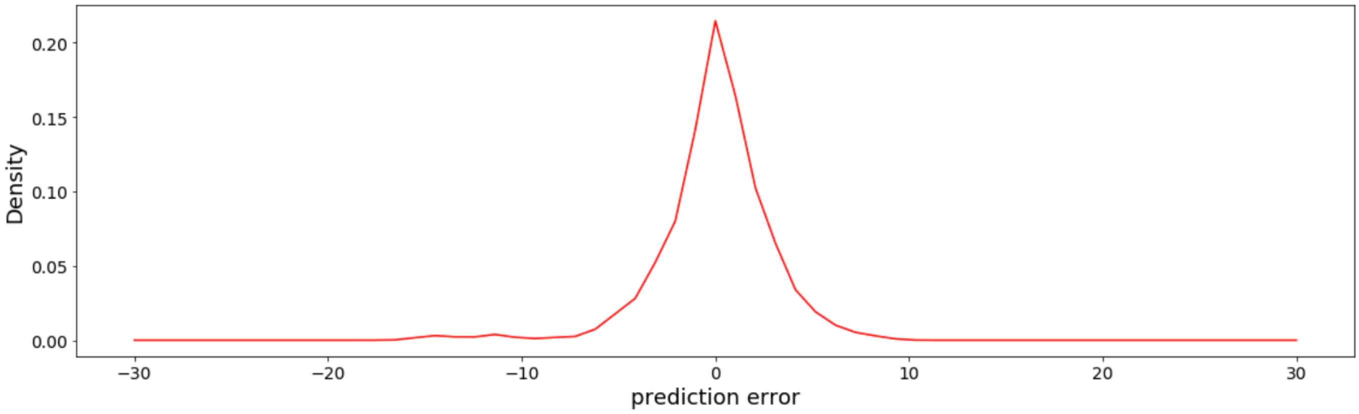
(c) The time series plot of measurements and model estimations for M for Video Frame Rate:



(d) The density and histogram plots for Video Frame Rate in test set:



(e) The density plot for the prediction error:



(f) In this module, I use ten features to estimate the Video Frame Rate, and get 10 coefficients and one bias using linear regression to train a model M with the training set. The figure of NMAE directly shows that for the test set, the difference between continued measured and estimated values is around 9.9%. A lower NMAE value means a better model. So the result is just barely satisfactory. I think the reason is insufficient number of data reduce the accuracy. Also, I use naïve method to simply predict target variables in test set of Y that only rely on the train set of Y. The naïve estimator here is 18.9 and the first plot shows that the red curve (the trend of Y_test) is basically around the green line (the naïve estimator). But the error of naïve method is obviously larger than NMAE. So I can safely draw a conclusion that NMAE is more accurate that naïve method.

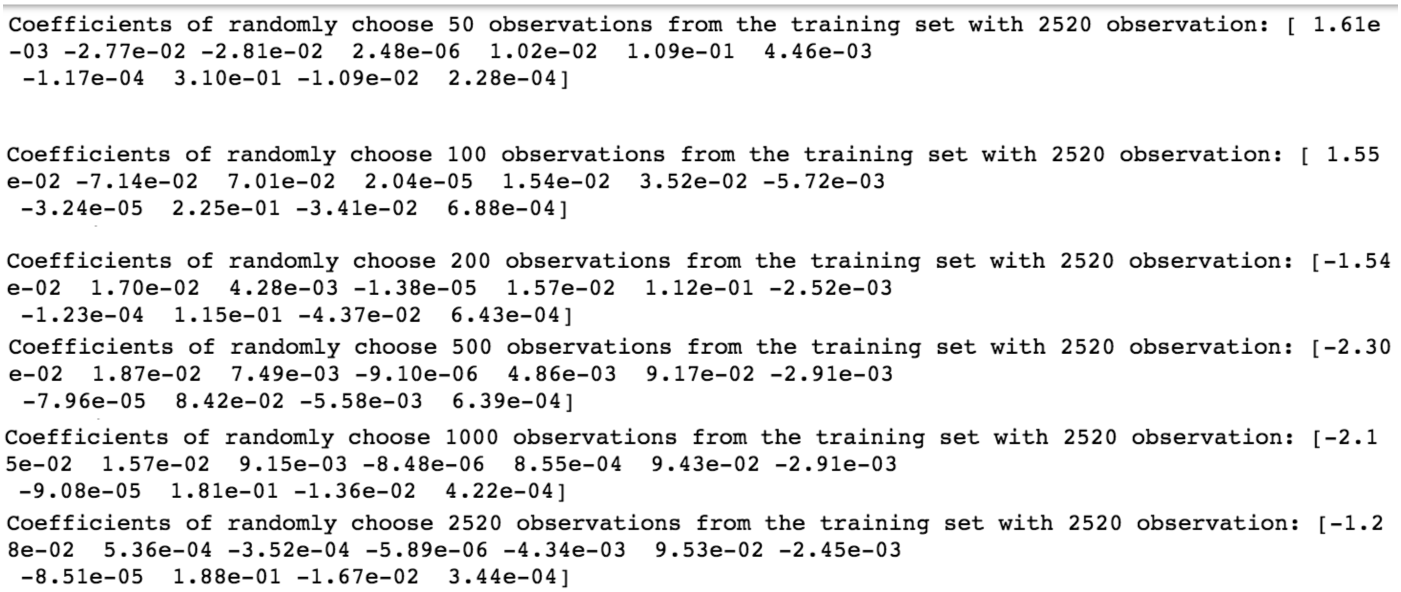
The density and histogram plots both reflect the data distribution of video frame rate, and they are mostly consistent here. They show that video frame rate has two peaks at 14 frame/s and 24 frame/s, so, more predicted video frame rates are distributed at two intervals of 14 frame/s and 24 frames/s. And it reaches the maximum number of 220 (around 0.105 of the density) when the rate is 24 frame/s. Meanwhile, the value turns to 0 when a frame rate is bigger than 26 or smaller than 10, and the curve of density plot tend to be smooth in this two intervals as well.

The density plot for prediction error illustrates that there is only one peak appears at interval [-8,8] and reach the maximum value of 0.22 at point 0, which illustrates that almost all predicted rates have the error that less than 8 frame/s compared to measured rates, and most of them have extremely small errors that close to 0.

2. Study the Relationship between Estimation Accuracy and the Size of the Training Set.

Answer:

(a)

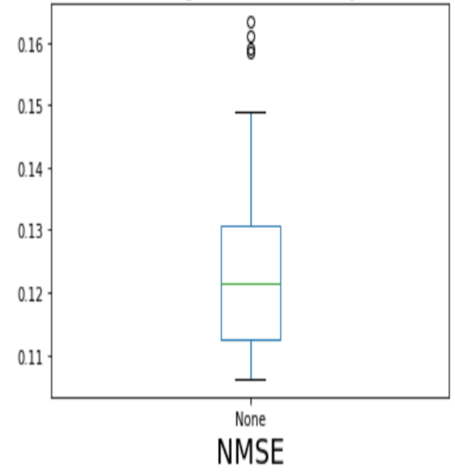


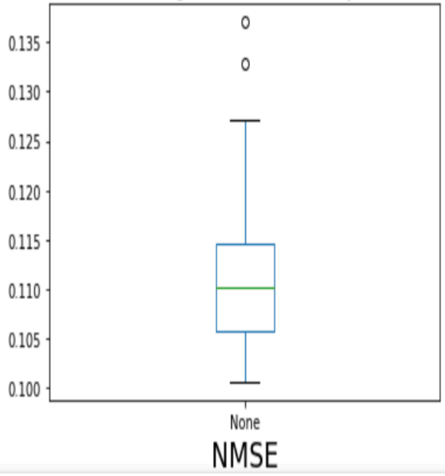
```
Coefficients of randomly choose 50 observations from the training set with 2520 observation: [ 1.61e
-03 -2.77e-02 -2.81e-02  2.48e-06  1.02e-02  1.09e-01  4.46e-03
 -1.17e-04  3.10e-01 -1.09e-02  2.28e-04]


Coefficients of randomly choose 100 observations from the training set with 2520 observation: [ 1.55
e-02 -7.14e-02  7.01e-02  2.04e-05  1.54e-02  3.52e-02 -5.72e-03
 -3.24e-05  2.25e-01 -3.41e-02  6.88e-04]

Coefficients of randomly choose 200 observations from the training set with 2520 observation: [-1.54
e-02  1.70e-02  4.28e-03 -1.38e-05  1.57e-02  1.12e-01 -2.52e-03
 -1.23e-04  1.15e-01 -4.37e-02  6.43e-04]
Coefficients of randomly choose 500 observations from the training set with 2520 observation: [-2.30
e-02  1.87e-02  7.49e-03 -9.10e-06  4.86e-03  9.17e-02 -2.91e-03
 -7.96e-05  8.42e-02 -5.58e-03  6.39e-04]
Coefficients of randomly choose 1000 observations from the training set with 2520 observation: [-2.1
5e-02  1.57e-02  9.15e-03 -8.48e-06  8.55e-04  9.43e-02 -2.91e-03
 -9.08e-05  1.81e-01 -1.36e-02  4.22e-04]
Coefficients of randomly choose 2520 observations from the training set with 2520 observation: [-1.2
8e-02  5.36e-04 -3.52e-04 -5.89e-06 -4.34e-03  9.53e-02 -2.45e-03
 -8.51e-05  1.88e-01 -1.67e-02  3.44e-04]
```

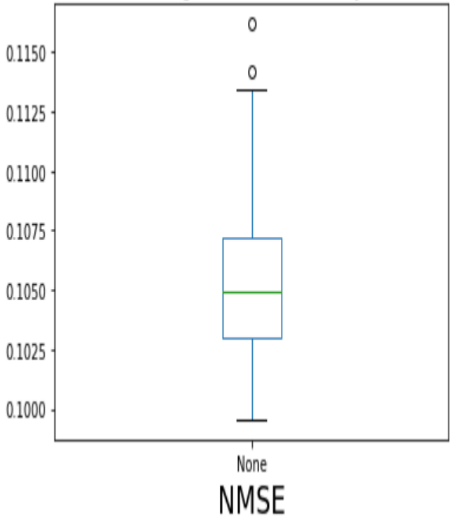
(b)

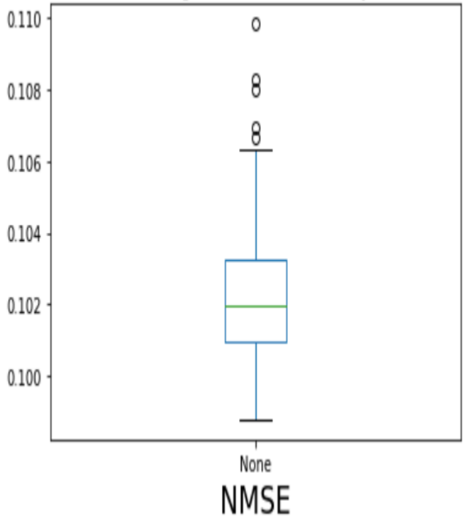
Normalized Mean Absolute Error (NMAE) of randomly choose 50 observations: 0.12488
Normalized Mean Absolute Error (NMAE) of randomly choose 100 observations: 0.11096
Normalized Mean Absolute Error (NMAE) of randomly choose 200 observations: 0.10527
Normalized Mean Absolute Error (NMAE) of randomly choose 500 observations: 0.10217
Normalized Mean Absolute Error (NMAE) of randomly choose 1000 observations: 0.10126
Normalized Mean Absolute Error (NMAE) of randomly choose 2520 observations: 0.10126

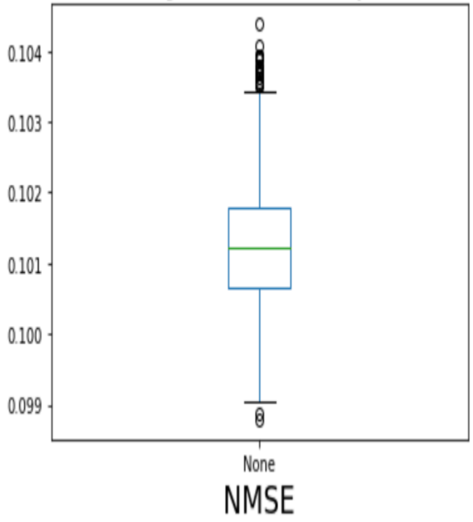
(d)

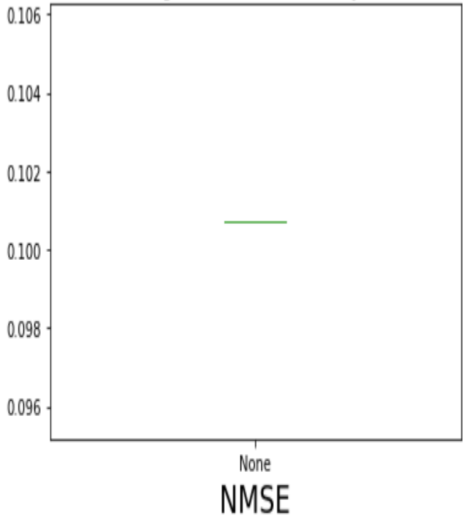

The Range of NMSE Values for training sets with 50 randomly selected observations; 50 runs

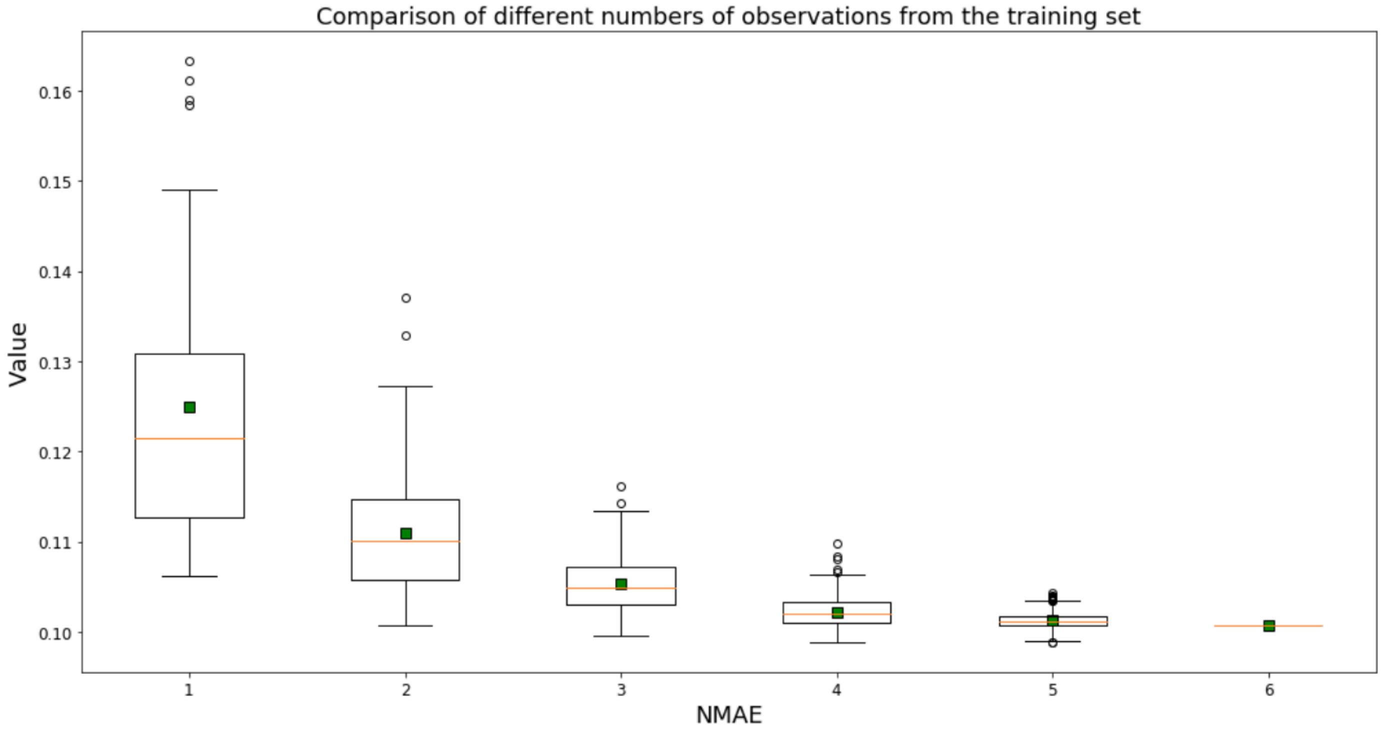

The Range of NMSE Values for training sets with 100 randomly selected observations; 50 runs


The Range of NMSE Values for training sets with 200 randomly selected observations; 50 runs


The range of NMSE Values for training sets with 500 randomly selected observations; 50 runs


Range of NMSE Values for training sets with 1000 randomly selected observations; 50 runs

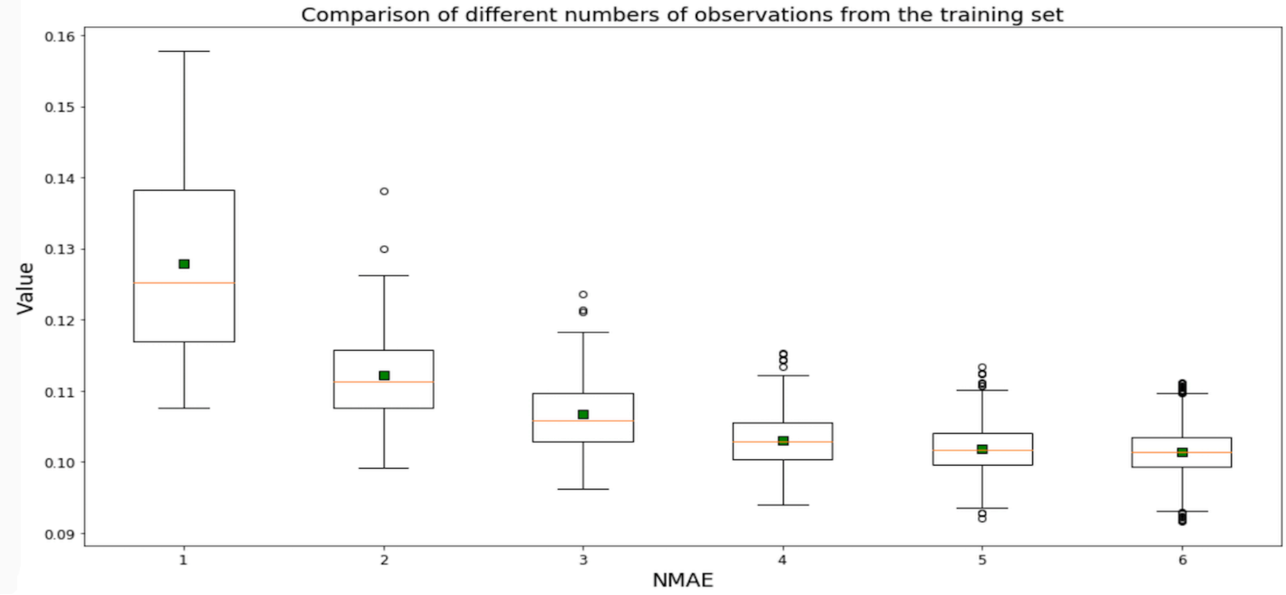

Range of NMSE Values for training sets with 2520 randomly selected observations; 50 runs



(e) In this task, I creat another six training sets from the original one to test the relationship between the size of the training set and estimation accuracy. The figures of NMAE shows that, as the size of training set grows, the value of NMAE decreases and accurancy increased.

To avoid contigency, I train linear models and compute the NMAE for each model above 50 times, and use box plots to show the range of the NMAE values for each given set size. As the size grows, the green point which represents the mean value of MNAE shows a downward trend and it means the accuracy is improving. Meanwhile, the distance between upper quartile and lower quartile is smaller with the training size increasing. The smaller the distance, the more concentrated it is. And it’s the same for upper limb and lower limb. The statistics lead us to the conclusion that, as the size of training set increases, the results of NMAE are more concentrated and the estimation accuracy is better.

PS: I considered another way to train models and perform 50 times: Every time before I select the new set from the training set(e.g.100), I first re-split the whole set (to 70% and 30%) to training set and test set again. So this means I re-split the original set to 70% and 30% 50 times as well. The result is :



It has the same pattern that, when we increase the size of train set, the values of MNAE are more concentrated and the mean value of MNAE becomes smaller. But I find that, with re-spliting the data set, the errors are more random. Because when I split the data set, the test set will be different. Some test sets will gain good results for some models and bad for others. So, it’s not fair to use different test sets for comparison. To evaluate the models, it’s better to compare the results using same data for evaluation.

Also, I observed that, it has a slower convergence than without re-spliting data set method. I think the reason is randomness of new selected set and test set largely increased. And the uncertainty of errors also contributes to this result. Due to the reason of re-spliting data set, it can never converge to a straight line as previous one.