Weekly Progress Report

Submitted by:

Sandesh Pabitwar

Modern Education Society’s College of Engineering Pune

Project Title: INTP22-ML-2 Remaining Usable Life Estimation (NASA Turbine dataset)

Objectives:

1. Understand the components and working of Turbo Engine.
2. Understand the dataset and analyze dataset
3. To learn about different Python libraries and their implementation.
4. To develop model for remaining usable life estimation

Task done in this week:

1. Downloaded and started exploring dataset.
2. Converted txt files into csv files.
3. Understood different functions of panda’s library.
4. Completed Exploratory data analysis on dataset 1.
5. Reading research papers on remaining usable life.

Downloaded and started exploring dataset:

At the starting of the week I strated reading the guidelines about the dataset and project then dowloaded the dataset folder and starting exploring about the dataset, gone through diiferent articles explaning about the dataset and after doing all these activities I got the clear image of the dataset. Some observations about the dataset are listed in below table.

Dataset	Operating condition	Fault mode	Train trajectories	Test trajectories
FD001	1	1	100	100
FD002	6	1	260	259
FD003	1	2	100	100
FD004	6	2	248	249

Converted txt files into csv files:

Link for converting txt file to csv file was given in the guideline section from there I had converted txt file into csv file.

Understood different functions of panda's library:

Before staring the analysis of data I gone through one tutorial of the pandas and understood the pandas functions such as read_csv() ,head() ,describe() , memory_usage() , astype() , value_counts() , drop_duplicates() , groupby() any many more.

These functions are briefly explained below.

- read_csv().
  read_csv() function helps read a comma-separated values (csv) file into a Pandas DataFrame

- head ()
  head(n) is used to return the first n rows of a dataset

- describe ()
  describe () is used to generate descriptive statistics of the data in a Pandas DataFrame

- memory_usage()
  memory_usage() returns a Pandas Series having the memory usage of each column (in bytes)

- astype()
  astype() is used to cast a Python object to a particular data type

Completed Exploratory data analysis on dataset 1:

Data of the csv file was not labeled so first step was to label the data as mentioned in the guidline I labeled the data.

Coloum1: Unit number

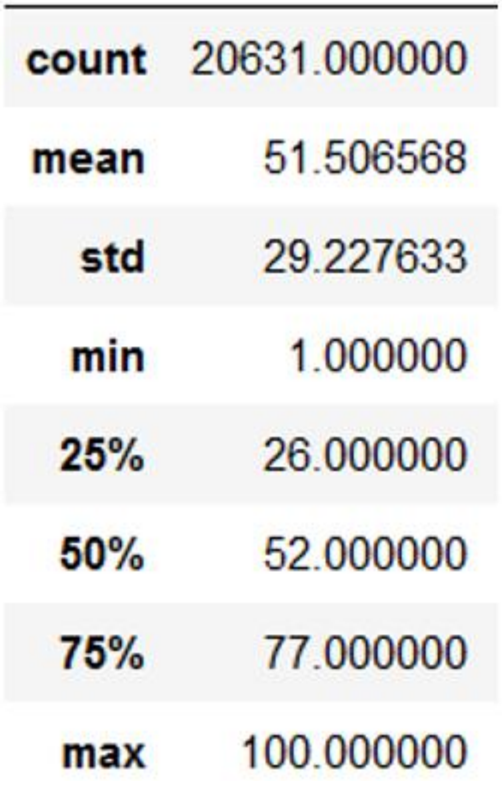
Coloum2: Time (in cycles)

Coloum 3,4,5: Operational Setting 1,2,3 respectively.

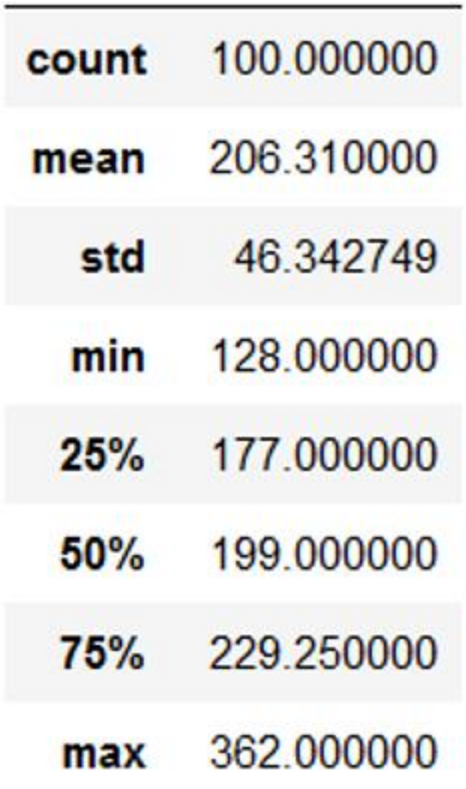
Column 6-26: sensor mesurment.

After doing this a using **raw_data[raw_data[“Unit_Number”] == 1]** this command I was exploring the data of each unit number. While doing this a got the end-of-life cycles value for each unit.

Now the actual data analysis was started using the describe function of pandas. Using the describe function on unit and cycles column I got the following output.

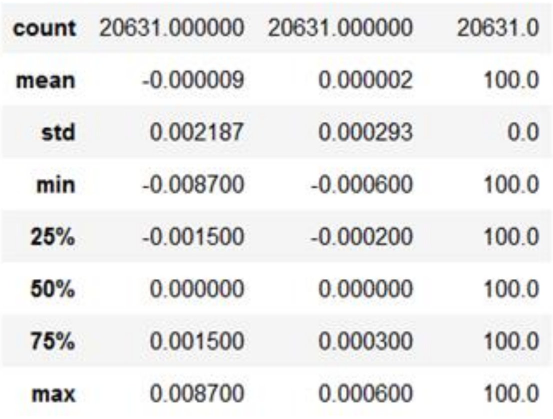


```
count  20631.000000
 mean     51.506568
  std     29.227633
  min      1.000000
  25%     26.000000
  50%     52.000000
  75%     77.000000
  max    100.000000
```



```
count  100.000000
 mean  206.310000
  std   46.342749
  min  128.000000
  25%  177.000000
  50%  199.000000
  75%  229.250000
  max  362.000000
```

After applying describe function on the operational setting I got following output.



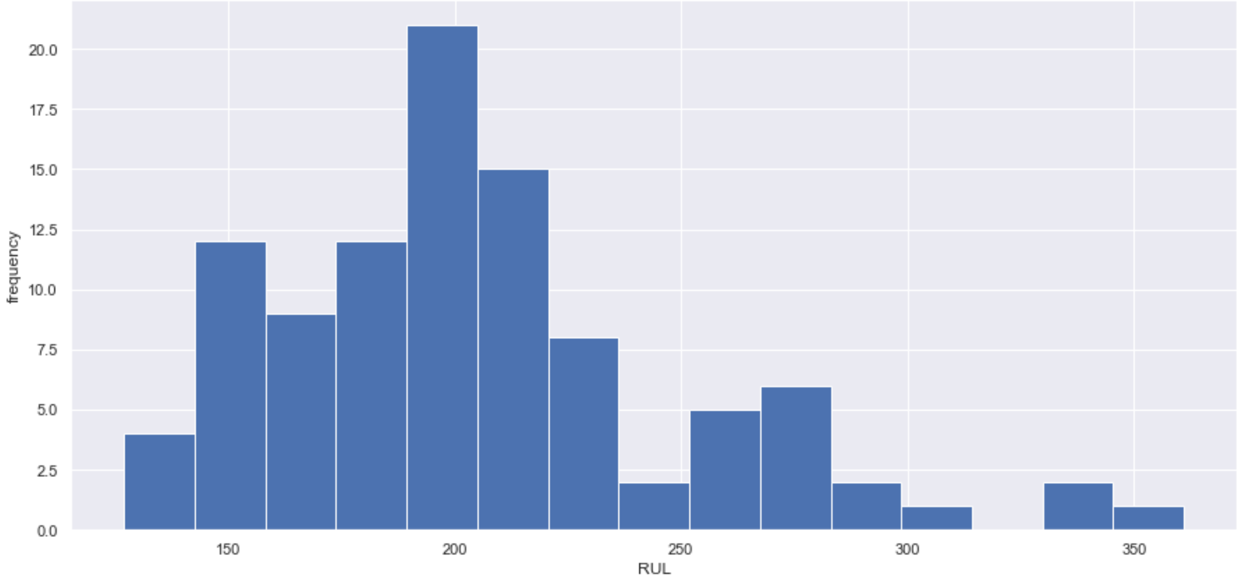
```
count  20631.000000  20631.000000  20631.0
 mean     -0.000009      0.000002    100.0
  std      0.002187      0.000293      0.0
  min     -0.008700     -0.000600    100.0
  25%     -0.001500     -0.000200    100.0
  50%      0.000000      0.000000    100.0
  75%      0.001500      0.000300    100.0
  max      0.008700      0.000600    100.0
```

Looking at the standard deviations of settings 1 and 2, they aren’t completely stable.

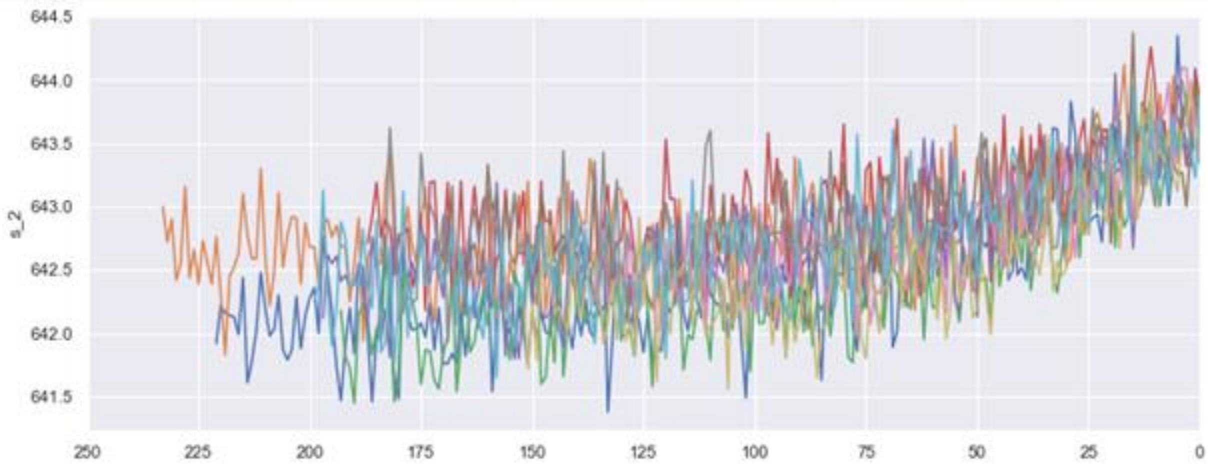
Same function I applied on sensors data and got to know that by looking at the standard deviation it's clear sensors 1, 10, 18 and 19 do not fluctuate at all, these can be safely discarded as they hold no useful information. Inspecting the quantiles indicates sensors 5, 6 and 16 have little fluctuation and require further inspection. Sensors 9 and 14 have the highest fluctuation, however this does not mean the other sensors can't hold valuable information.

After this I calculated the remaining useful life for each unit then plotted the histogram of the data.

Histogram looks like.



Then plotted the graph of RUL against the sensors data for all the sensors. Below graph is of sensor2 vs RUL.



Reading research papers on remaining usable life.

I gone through research paper titled as "Prediction of Remaining Useful Lifetime (RUL) of Turbofan Engine using Machine Learning By Vimala Mathew, Tom Toby, Vikram Singh, B Maheswara Rao, M Goutham Kumar. From this research paper I got idea about what all the activities I must do after doing exploratory data analysis.