IBM MACHINE LEARNING: EDA PROJECT

EDA of NASA Aerofoil Self Noise Dataset

1. Objective

In this project, we use the NASA Airfoil Self Noise dataset, where we are going to EDA on the dataset and make some hypothesis which will help us to do regression or classification as needed

2. Data:

The NASA data set comprises different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments. There are a total of 1503 observations. Following is the descriptions of the data features and labels.

- Input features:

    1. **f**: Frequency in Hertz [Hz].
    2. **alpha**: Angle of attack (AoA, α), in degrees [°].
    3. **c**: Chord length, in meters [m].
    4. **U_infinity**: Free-stream velocity, in meters per second [m/s].
    5. **delta**: Suction side displacement thickness (δ), in meters [m].

- Output:
    1. **SSPL**: Scaled sound pressure level, in decibels [dB].

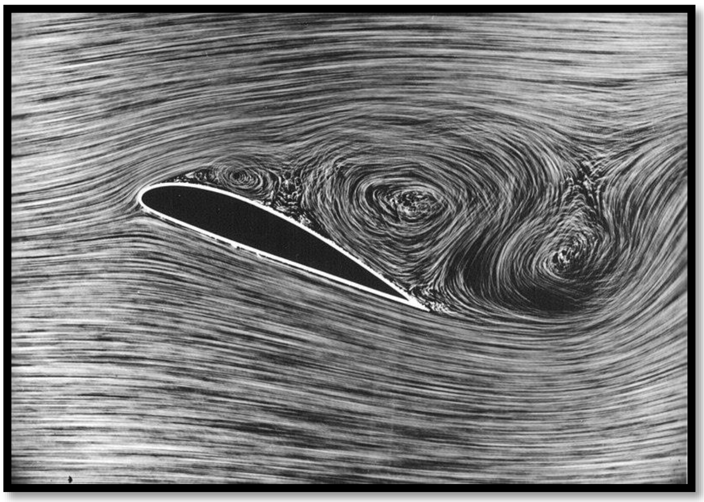
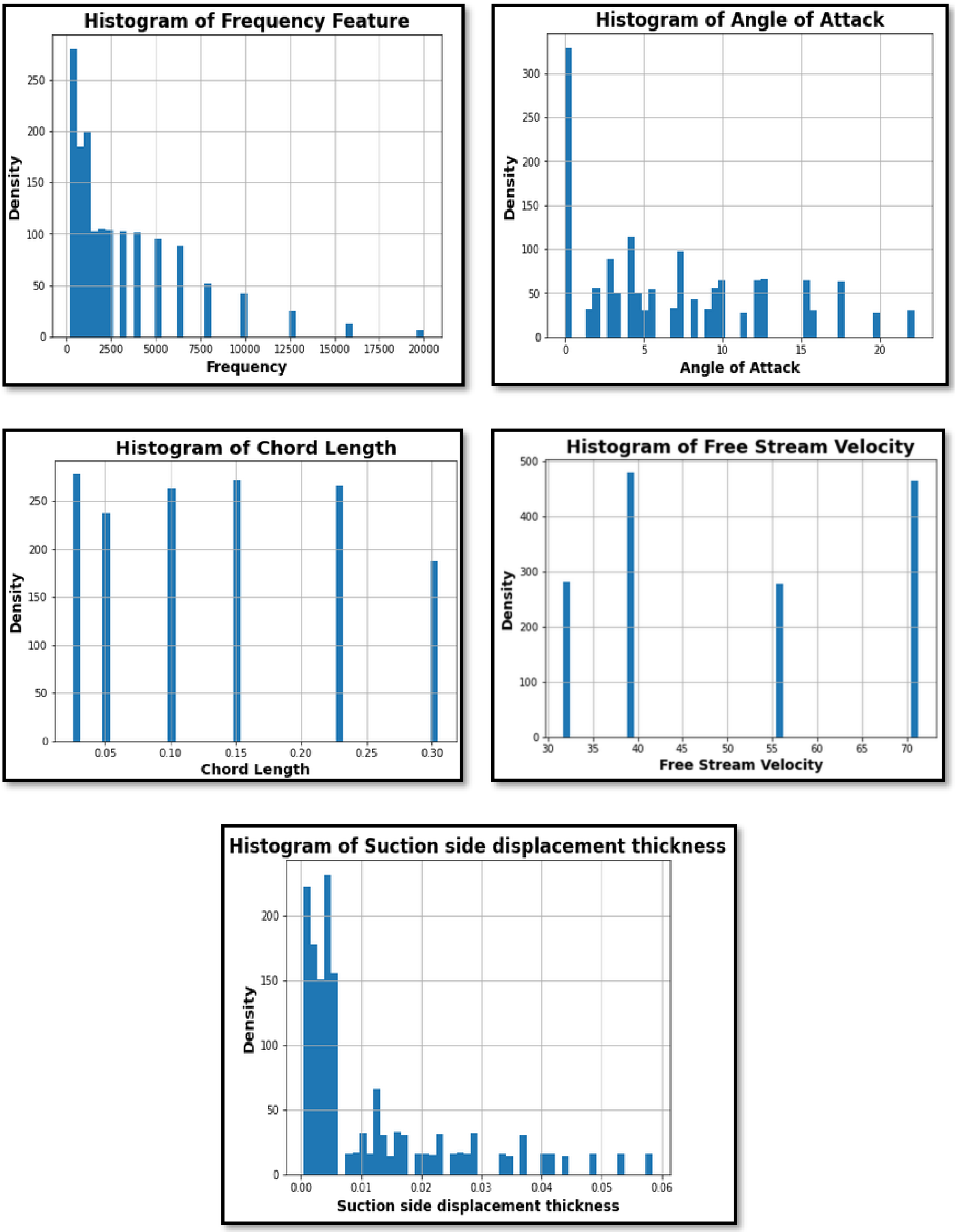


Figure 1. Wikipedia image of Flow over a Airfoil during testing.

3. Data Exploration:

3.1. Histogram of Features

A good way to explore the data is to use histograms and see the distributions of features. Following are the histograms of features.


Histogram of Frequency Feature

Histogram of Angle of Attack

Histogram of Chord Length

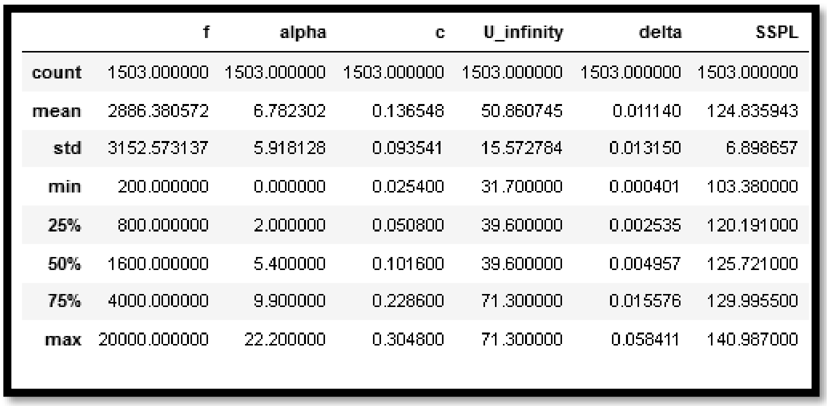
Histogram of Free Stream Velocity

Histogram of Suction side displacement thickness

Figure 2. Histograms of Features.

3.2 Data description:

A necessary step in EDA is to see the data description so that we get an idea about the data. Generally, we look at the percentiles and mean of the data. We also see wheather we have categorical or numerical data.



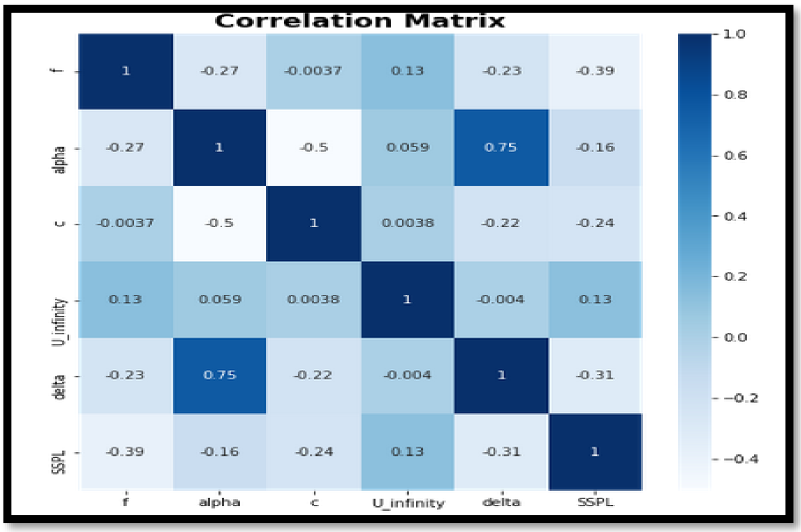
	f	alpha	c	U_infinity	delta	SSPL
count	1503.000000	1503.000000	1503.000000	1503.000000	1503.000000	1503.000000
mean	2886.380572	6.782302	0.136548	50.860745	0.011140	124.835943
std	3152.573137	5.918128	0.093541	15.572784	0.013150	6.898657
min	200.000000	0.000000	0.025400	31.700000	0.000401	103.380000
25%	800.000000	2.000000	0.050800	39.600000	0.002535	120.191000
50%	1600.000000	5.400000	0.101600	39.600000	0.004957	125.721000
75%	4000.000000	9.900000	0.228600	71.300000	0.015576	129.995500
max	20000.000000	22.200000	0.304800	71.300000	0.058411	140.987000

Figure 3. Data description.

We see that there are a total of 1503 observations. We also see that the features and target is distributed over various scales of numbers, which tell us that the data has to be normalised.

3.3 Correlation Matrix:

The correlation matrix tells us how the features and target are related with each other. A correlation of 1 tells us that the variables are positively corelated and a correlation of -1 tells us that they are negatively correlated. If the correlation is 0 then there is no correlation between the variables. Such variables need not be used for further Machine Learning Task.



	f	alpha	c	U_infinity	delta	SSPL
f	1	-0.27	-0.0037	0.13	-0.23	-0.39
alpha	-0.27	1	-0.5	0.059	0.75	-0.16
c	-0.0037	-0.5	1	0.0038	-0.22	-0.24
U_infinity	0.13	0.059	0.0038	1	-0.004	0.13
delta	-0.23	0.75	-0.22	-0.004	1	-0.31
SSPL	-0.39	-0.16	-0.24	0.13	-0.31	1

Figure 4. Correlation Matrix.

We see that all features have some correlation with the Target feature ‘SSPL’. Therefore, all features can be included in the Machine Learning Model.

3.4 Box Plots:

Box Plots are good way to describe the distribution of the dataset. It tells us about the percentiles distribution and also shows wheather the data has certain outliers which play a huge role in Machine Learning Model evaluation.

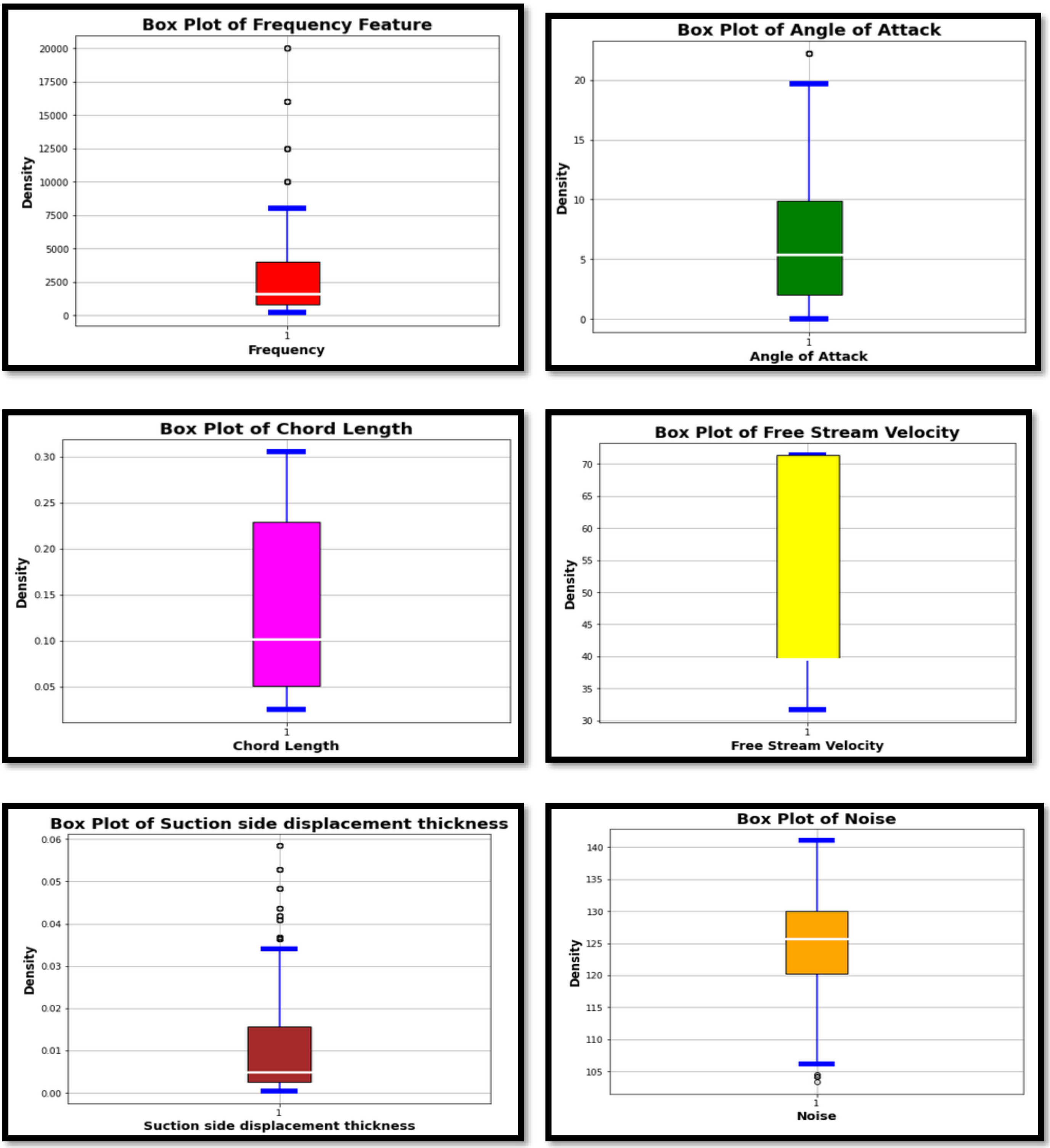


Figure 5. Box Plot of all the features and target variables.

We see that certain features have outliers and certain do not. Outliers are something which has to be kept in mind for further Machine Learning Modelling.

4. Hypothesis Testing

One of the assumptions in regression is that the residuals obtained are assumed to be normally distributed. Statistically this means that the mean is centred around zero and standard deviation is one (unit variance). The distribution is in the shape of a bell curve.

4.1 Hypothesis:

Null Hypothesis: The target values are normally distributed.

Alternate Hypothesis: The target values are not normally distributed.

Statisticians say if $p > 0.05$, then the dataset is normally distributed. If observation is not normally distributed, so we reject the Null Hypothesis.

4.2 Determine Normality

1. If the target variable is Normally distributed, it obtains better results.

2. If our target is not normally distributed, we can apply a transformation to it and then fit our regression to predict the transformed values

3. To see if target values are normally distributed, check:

   a. Visually

   b. Statistically

1. Visual Test using Histogram of the Label Feature of SSPL Noise:

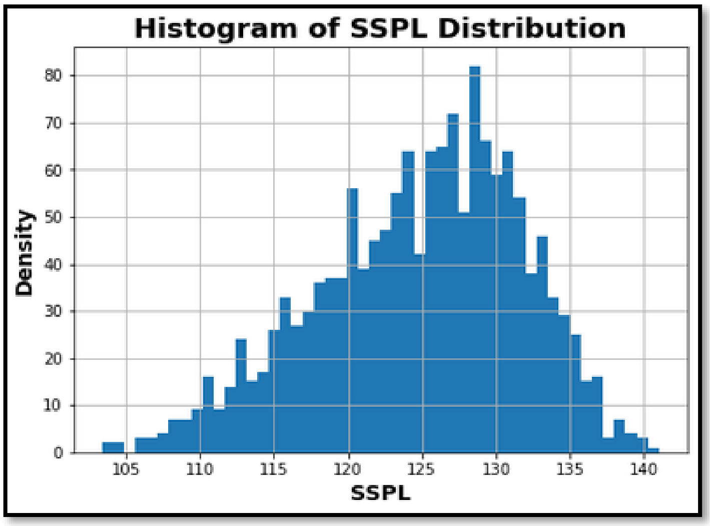
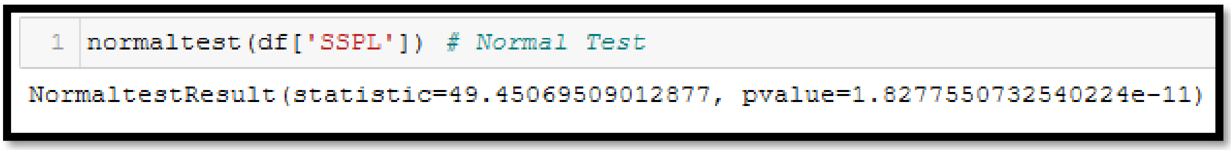


Figure 6. Histogram of SSPL Output Label.

Visually inspecting we can say that the histogram is not normally distributed and is skewed towards the left.

2. Statistic Test using p-value:

A normal test is conducted on the 'SSPL' label and the p-value obtained is less than 0.05.



```
normaltest(df['SSPL']) # Normal Test
NormaltestResult(statistic=49.45069509012877, pvalue=1.8277550732540224e-11)
```

Figure 7. P-value of 'SSPL'.

Thus, the desired output 'SSPL' is not normally distributed.

4.3 Transformations

Transformation of the output label can make the it normally distributed. Some transformation which can make a non-normally distributed to a normally distributed are:

a. Log Transformation

b. Square Root Transformation

c. Box-Cox Transformation

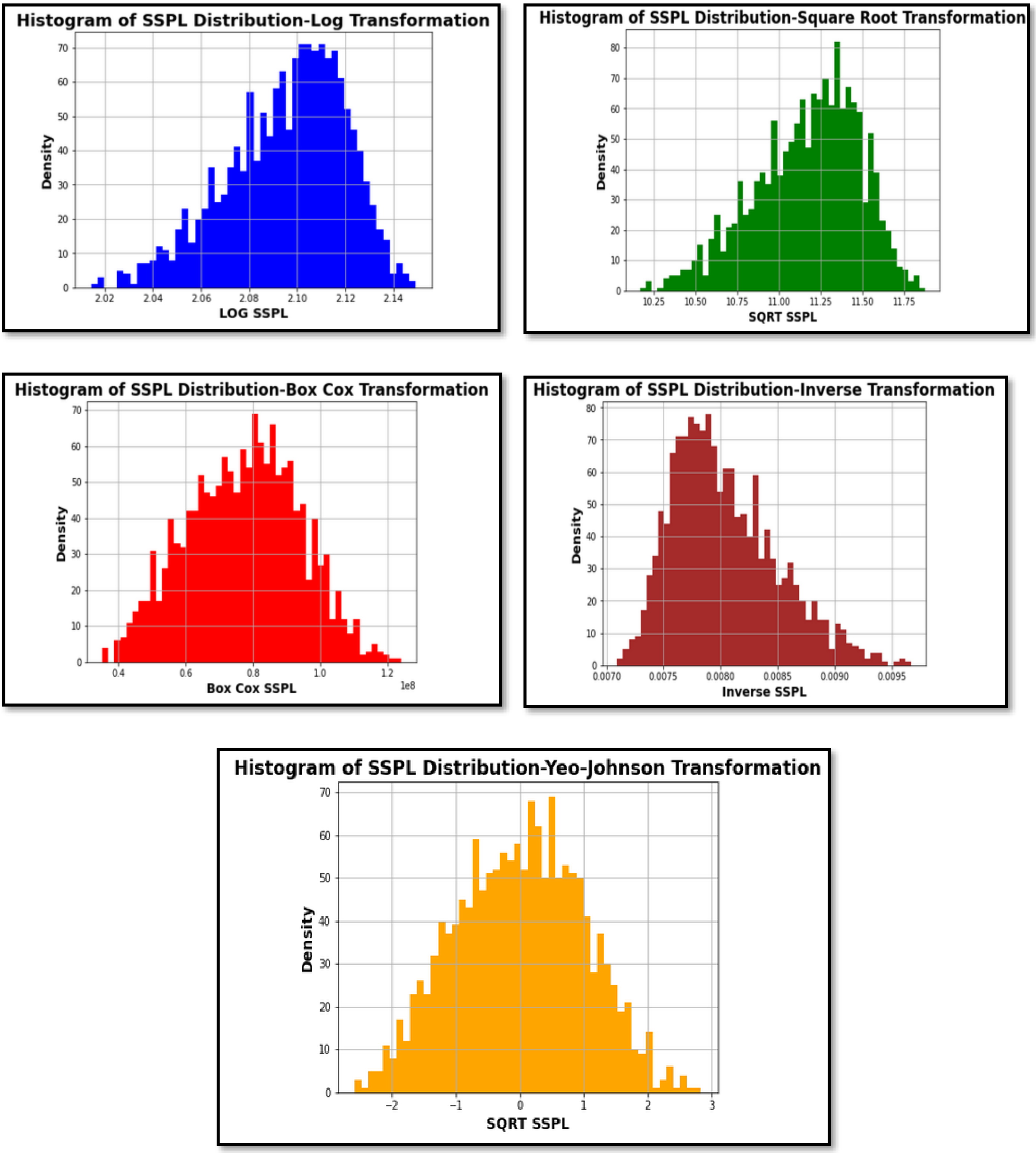
d. Inverse Transformation

e. Yeo-Johnson Transformation

We try all the above transformation and conduct a normal test after transformation. Following table describes the p-value obtained during different transformations

Serial No	Transformation	p-value
1	Log	1.7218157371855 87e-15
2	Square Root	3.5078566279140157e-13
3	Box-Cox	1.7951647448346226e-08
4	Inverse	5.301185393463406e-22
5	Yeo-Johnson	1.63207892e-07

It looks like all the transformation were not able to obtain a normal distribution, with Yeo-Johnson Transformation performing the best since p-value is close to 0.05 when compared with others. The transformations are also visually inspected using histograms.



Histogram of SSPL Distribution-Log Transformation

Histogram of SSPL Distribution-Square Root Transformation

Histogram of SSPL Distribution-Box Cox Transformation

Histogram of SSPL Distribution-Inverse Transformation

Histogram of SSPL Distribution-Yeo-Johnson Transformation

Figure 8. Histograms of "SSPL" after various transformations.

Since none of the transformation were able to make the output label normally distributed, we reject the null hypothesis that the targe values are normally distributed and go ahead with the alternate hypothesis without any transformations.

5. Conclusion

The EDA of NASA Airfoil Dataset was done in the project. We found that all the features are correlated with the Target variable and they have an impact in further machine learning task. We also did the hypothesis testing for regression task where we had to reject the null hypothesis that the target feature is normally distributed that means we have a Type 1 error.

6. Reference

1. Data Reference:

- https://archive.ics.uci.edu/ml/datasets/airfoil+self-noise
- https://www.kaggle.com/fedesoriano/airfoil-selfnoise-dataset