Abstract

This study was based on a NASA data set, which was used to obtain a series of aerodynamic and acoustic tests of multidimensional airfoil sections pertaining to the second and third degree. In terms of this report the data was put through several statistical analyses, which all were performed using Python. Visualizations were created to have a better understanding of the data that was presented and regression analysis was performed on some of the data categories that contained the strongest correlation values.

Introduction

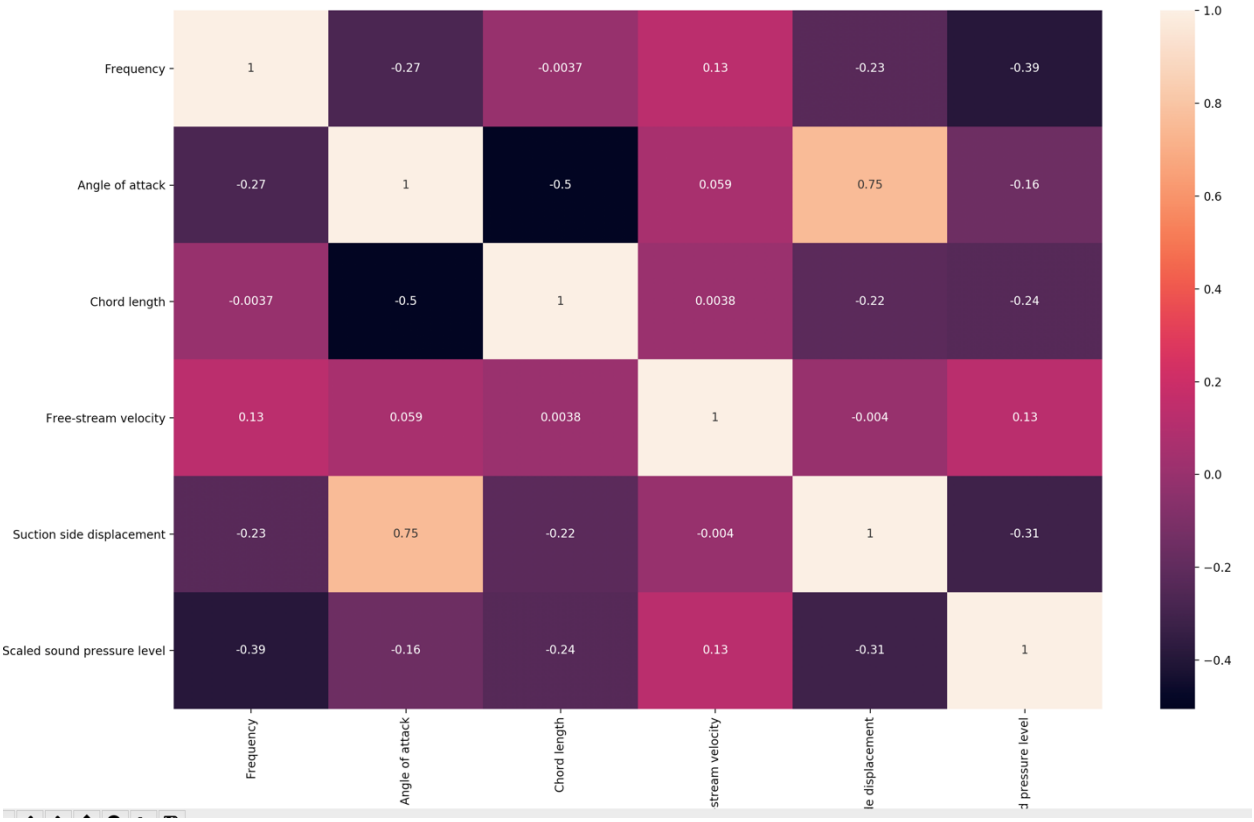
The purpose was to use Linear regression to predict future data points by creating a prediction line based on the correlation value. The higher the correlation the value, the better the results in having a prediction line in which it can be used to determine future data points between two variables. If a proper regression model is created it can be used to further understand and predict aerodynamic and acoustic results of airfoil in the second and third dimension.

Methodology

The data was provided through a CSV file. Python was used to create visualizations and statistical analysis on the data supplied in the file. For the assignment, visualizations (histograms and box plots) were required for the 6 data categories provided in this file. Another requirement was to provide a linear regression model for 2 different pairs of data. To do this, I created a correlation table and heat map to determine the data points that consisted of the highest correlation values, in order to determine which pairs would have the strongest prediction potential.

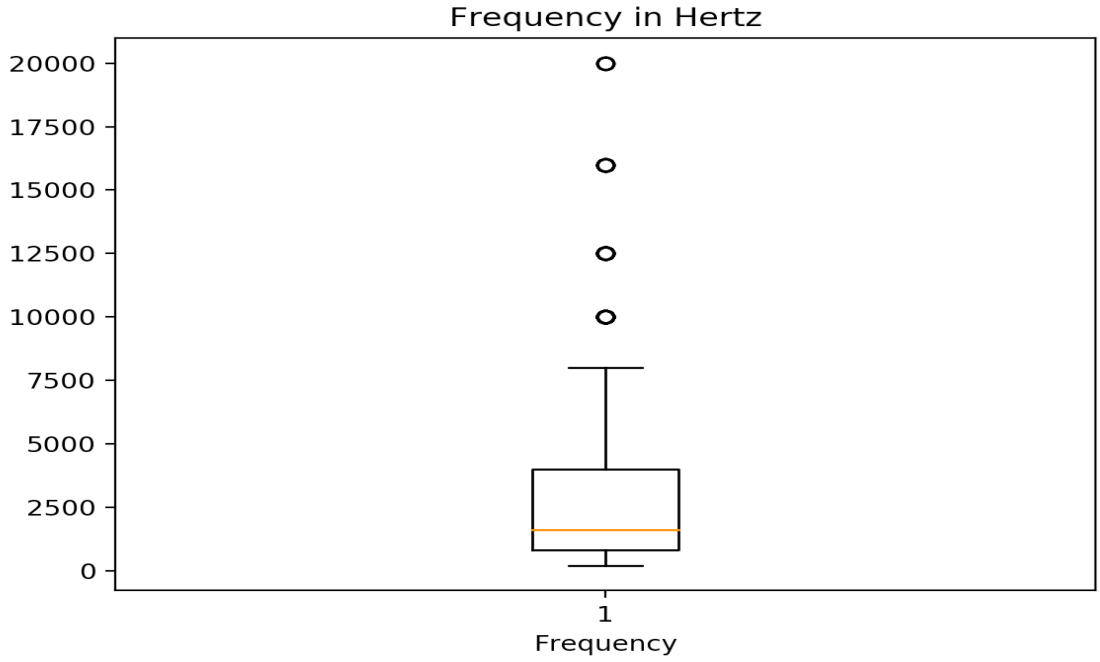
Correlation Table:

	Frequency	Angle of attack	Chord length	Free-stream velocity	Suction side displacement	Scaled sound pressure level
Frequency	1.000000	-0.272765	-0.003661	0.133664	-0.230107	-0.390711
Angle of attack	-0.272765	1.000000	-0.504868	0.058760	0.753394	-0.156108
Chord length	-0.003661	-0.504868	1.000000	0.003787	-0.220842	-0.236162
Free-stream velocity	0.133664	0.058760	0.003787	1.000000	-0.003974	0.125103
Suction side displacement	-0.230107	0.753394	-0.220842	-0.003974	1.000000	-0.312670
Scaled sound pressure level	-0.390711	-0.156108	-0.236162	0.125103	-0.312670	1.000000

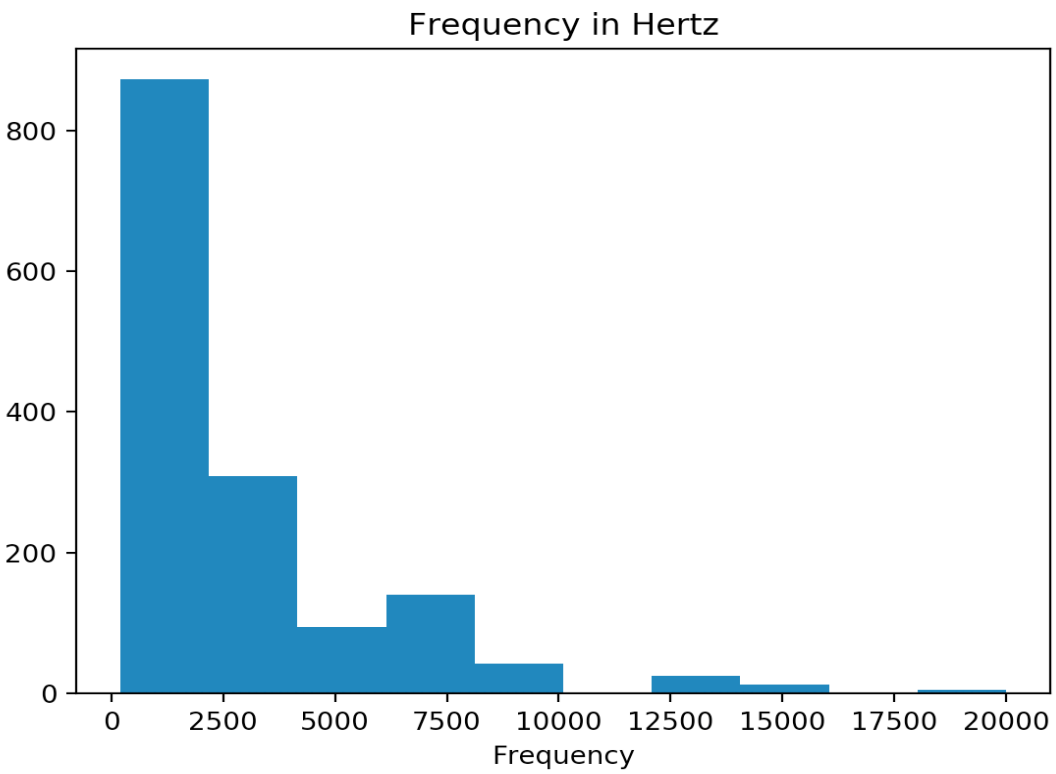


From the correlation table or the heat map provided above, it can be determined that the categories that have the highest correlation values is Angle of Attack vs Suction Side Displacement with 0.75, this shows that there is a stronger positive correlation between these two data sets. The second highest pair is Angle of Attack vs Chord length with -0.50, which shows there is a negative and a weaker correlation between these two data sets. We should be able to create a regression analysis on these two pairs of data.

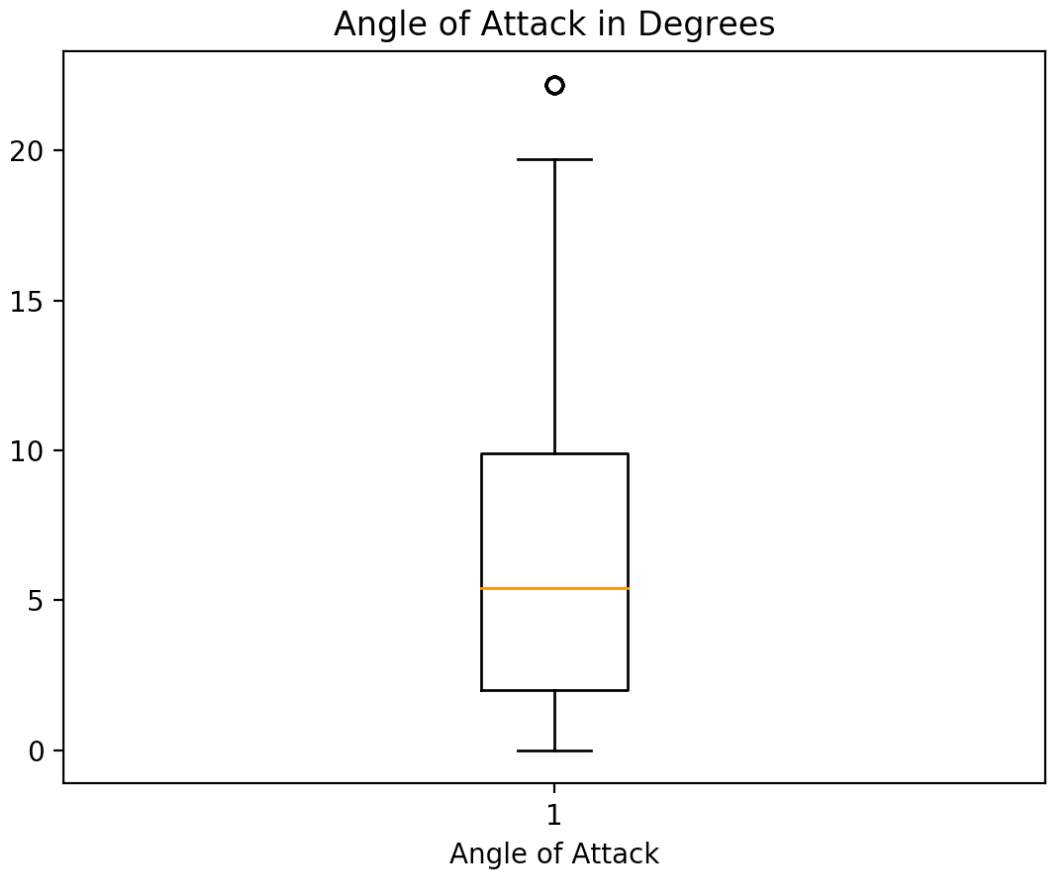
Data Categories



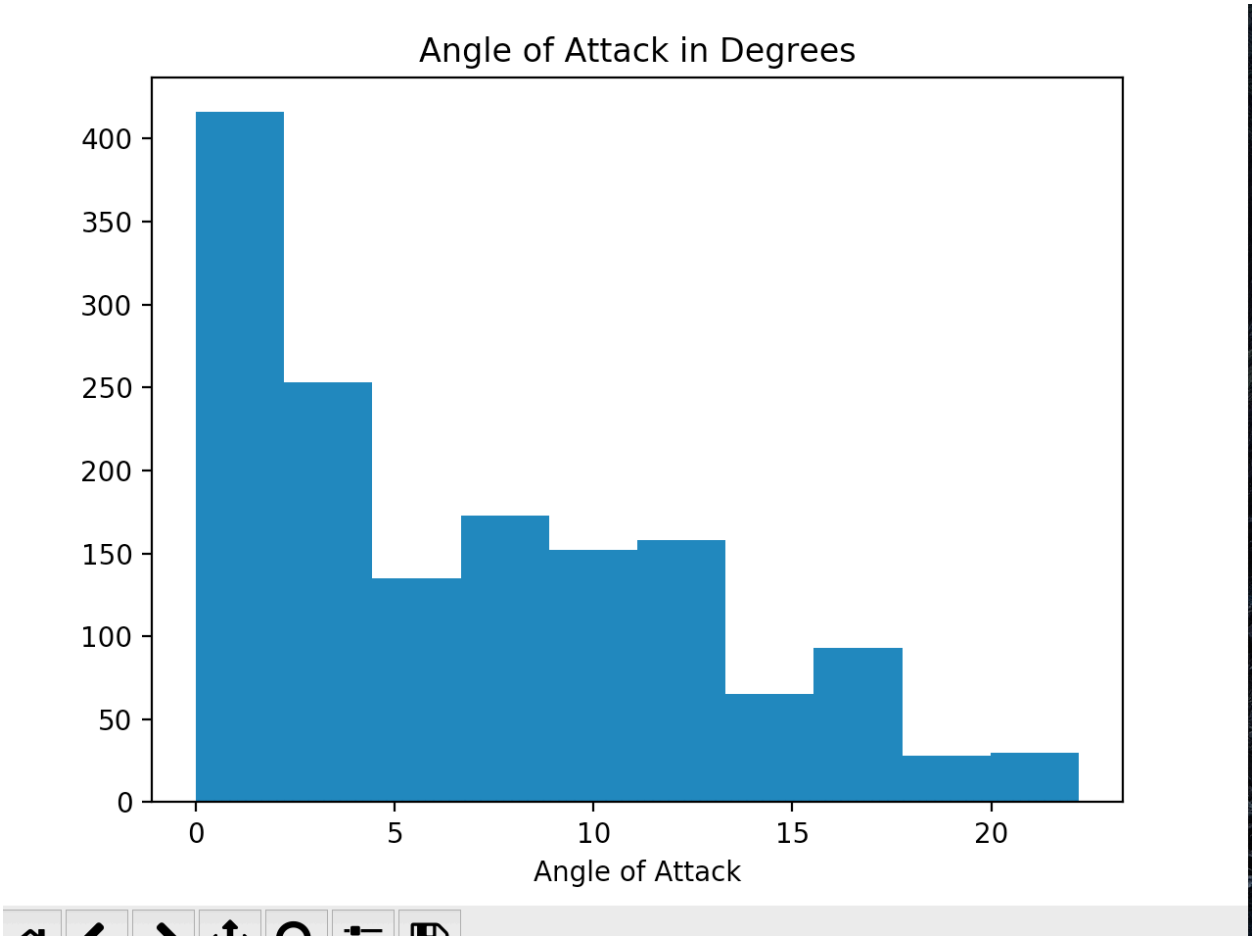
One of the first data categories for our data set is Frequency in hertz. Frequency is described to be the number of occurrences of repeating a certain event or task per unit of time. As we can see from the above box plot, frequency has a median value of 1600 Hz. The 25% quartile value is 800 Hz and the 75% quartile value is 4,000 Hz. We are able to see from this box plot that this data set contains several outliers which range between 10-20 thousand hertz. The maximum value for this data is 20,000 hertz while the minimum value is 200 Hz. The mean is 2886.38 hertz and the standard deviation is 3152.57 hertz.



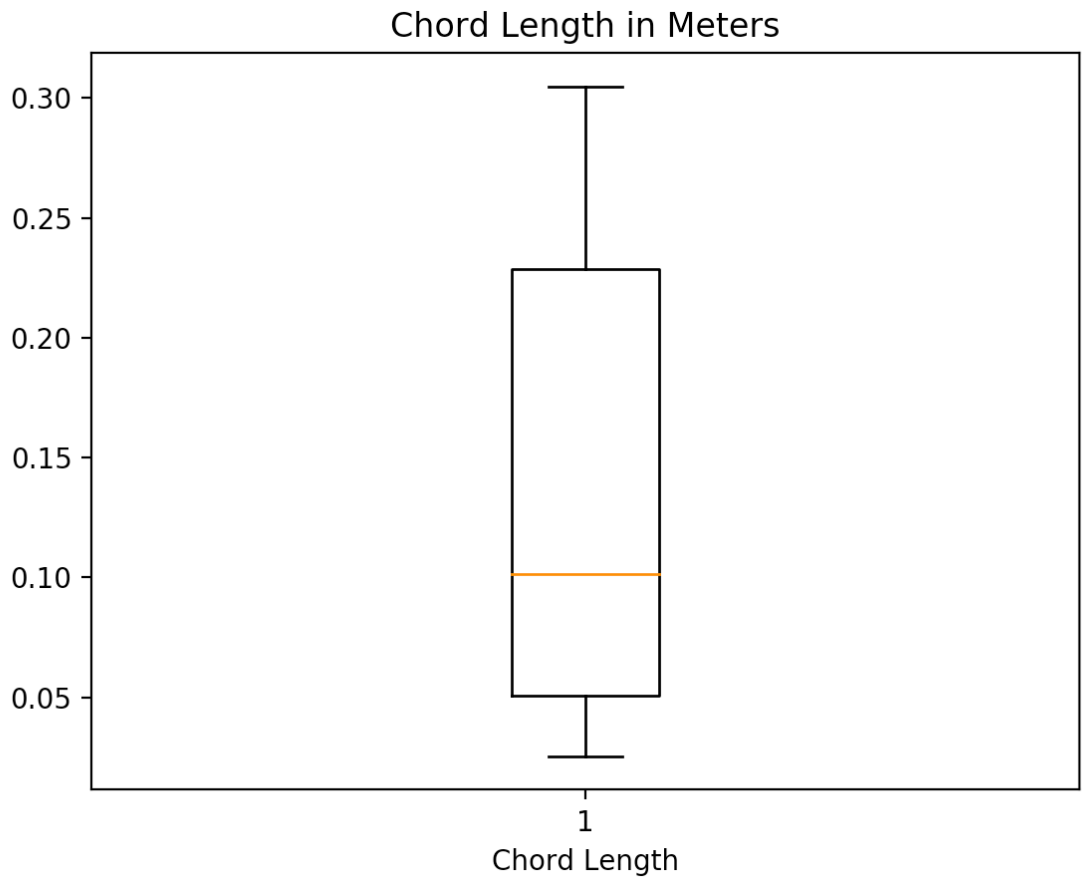
The histogram shows that the frequency data set has a strong positive skewness and we are able to see that most of the data clusters between 0 and 5,000 hertz. This data set is far from a normal distribution.



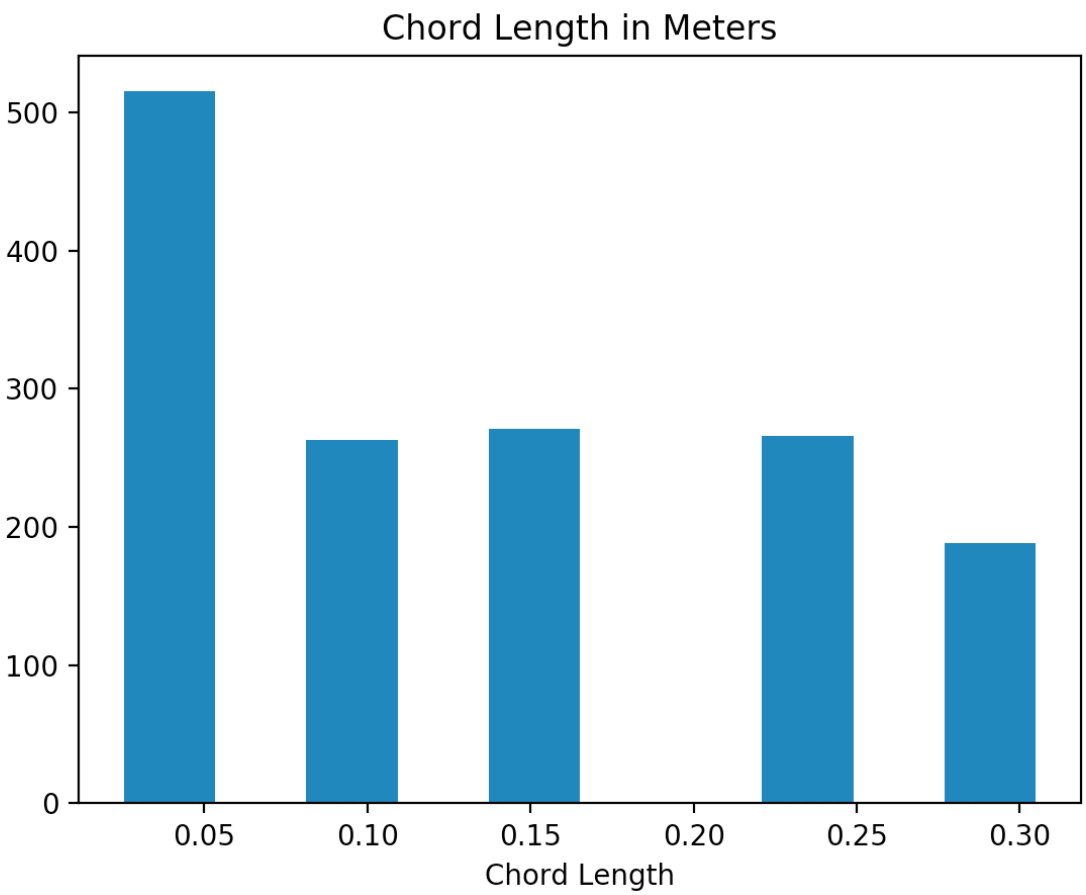
The angle of attack data set has a median value of 5.40 degrees. The 25% quartile is 2 degrees and the 75% quartile is 9.90 degrees. The mean is 6.78 degrees. The standard deviation is 5.92 degrees. The minimum value of the data set is 0 degrees and the maximum value is 22.20 degrees. From the box plot we can also see that the maximum value of 22.20 degrees is an outlier for this data set.



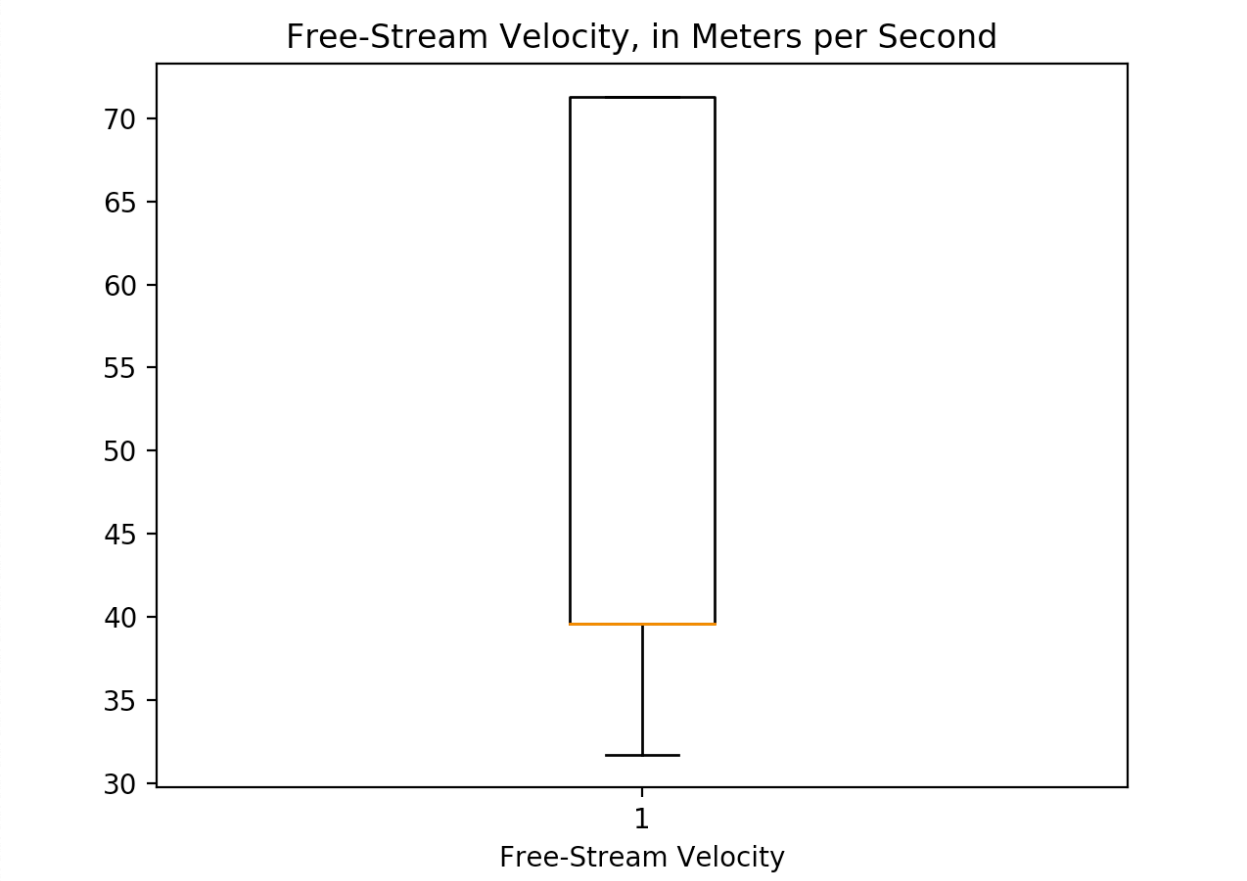
The angle of attack is positively skewed. We can see that the major cluster of data falls between 0 and 5 degrees, this is where the highest frequency of the distribution is located. Then between 5 and 13 degrees the distribution has a frequency between 100 and 175.



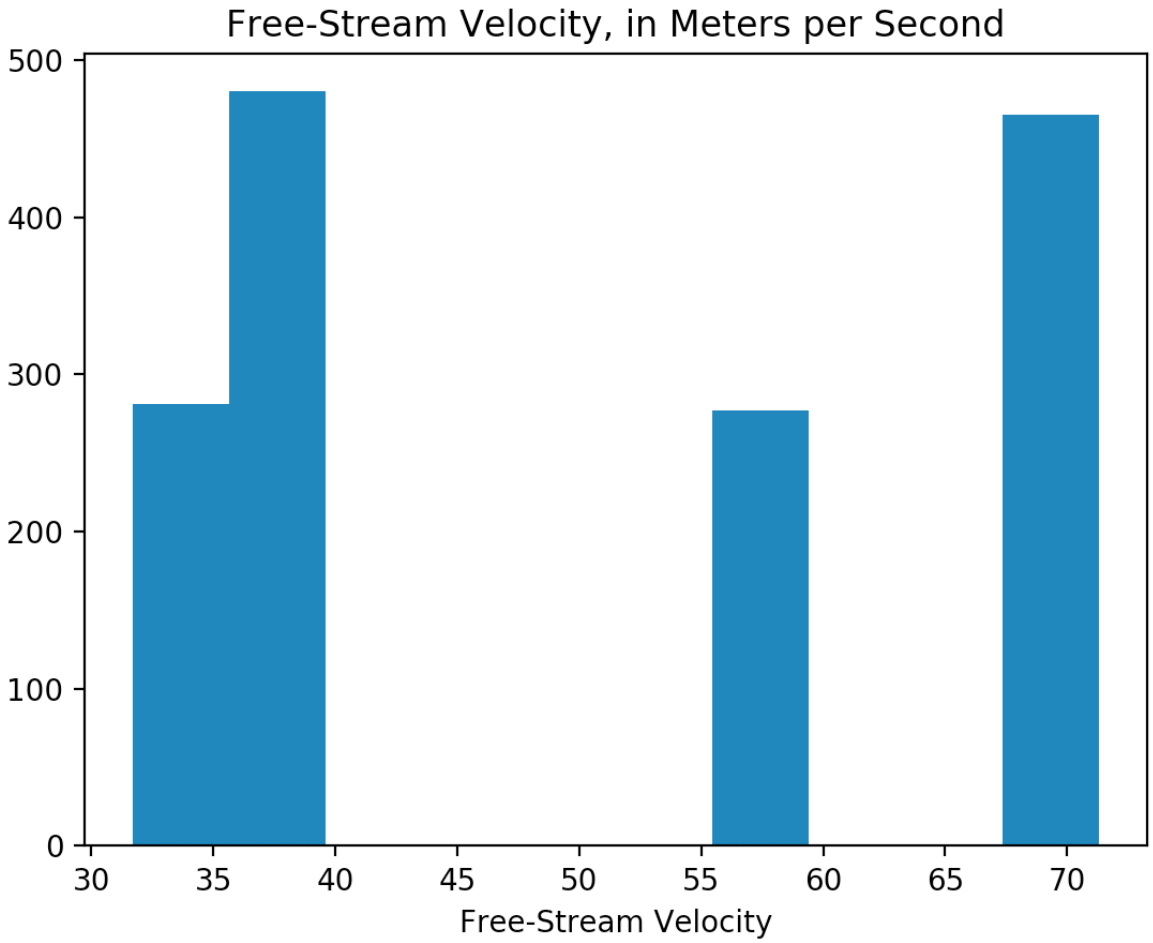
The Chord length data category has a median value of 0.10 meters. The mean is 0.14 meters. Chord length data set has a 25% quartile value of 0.05 and a 75% quartile value of 0.23 meters. The minimum value was 0.03 meters and the maximum value was 0.30 meters. This data set has a standard deviation of 0.09 meters. From the visualization above you can also see that this data set has no outliers.



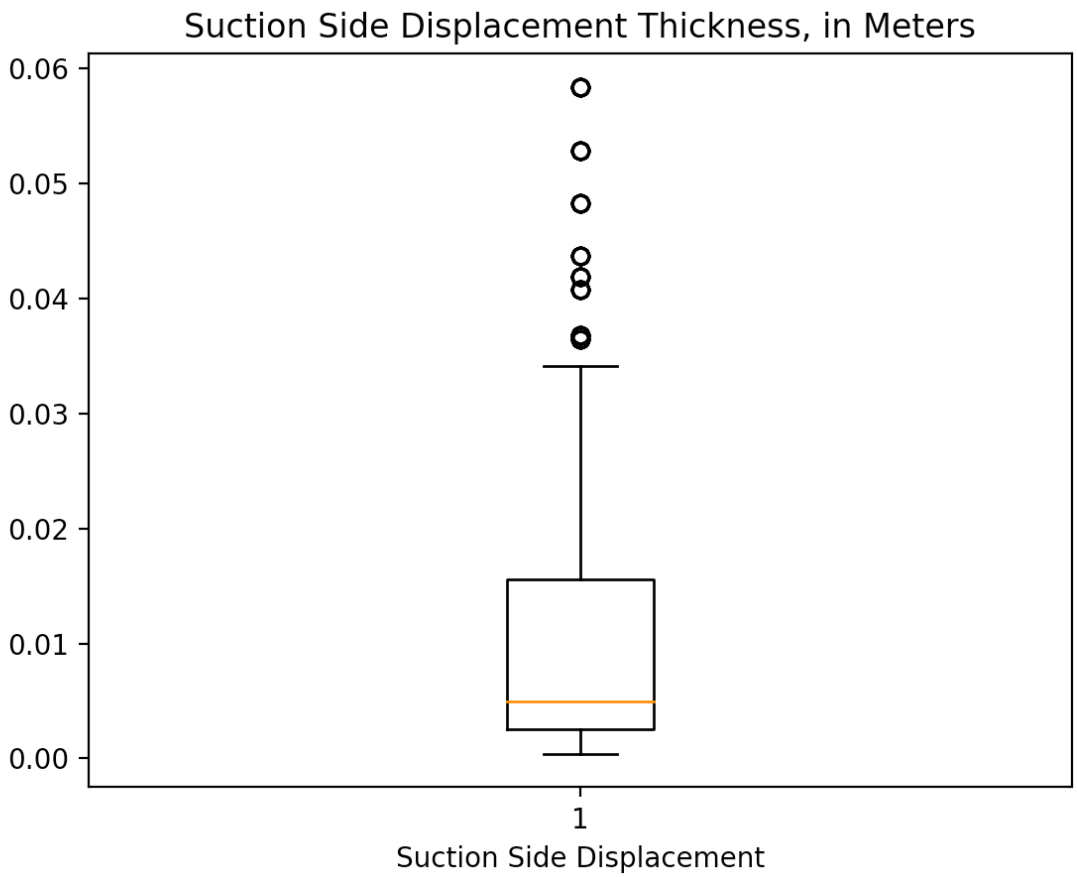
The data set for Chord length follows a positively skewed distribution. The highest frequency for the distribution falls around 0.05 meters with a count of close to 500. The remainder of 0.10, 0.15 and 0.25 all have close to the same frequency count.



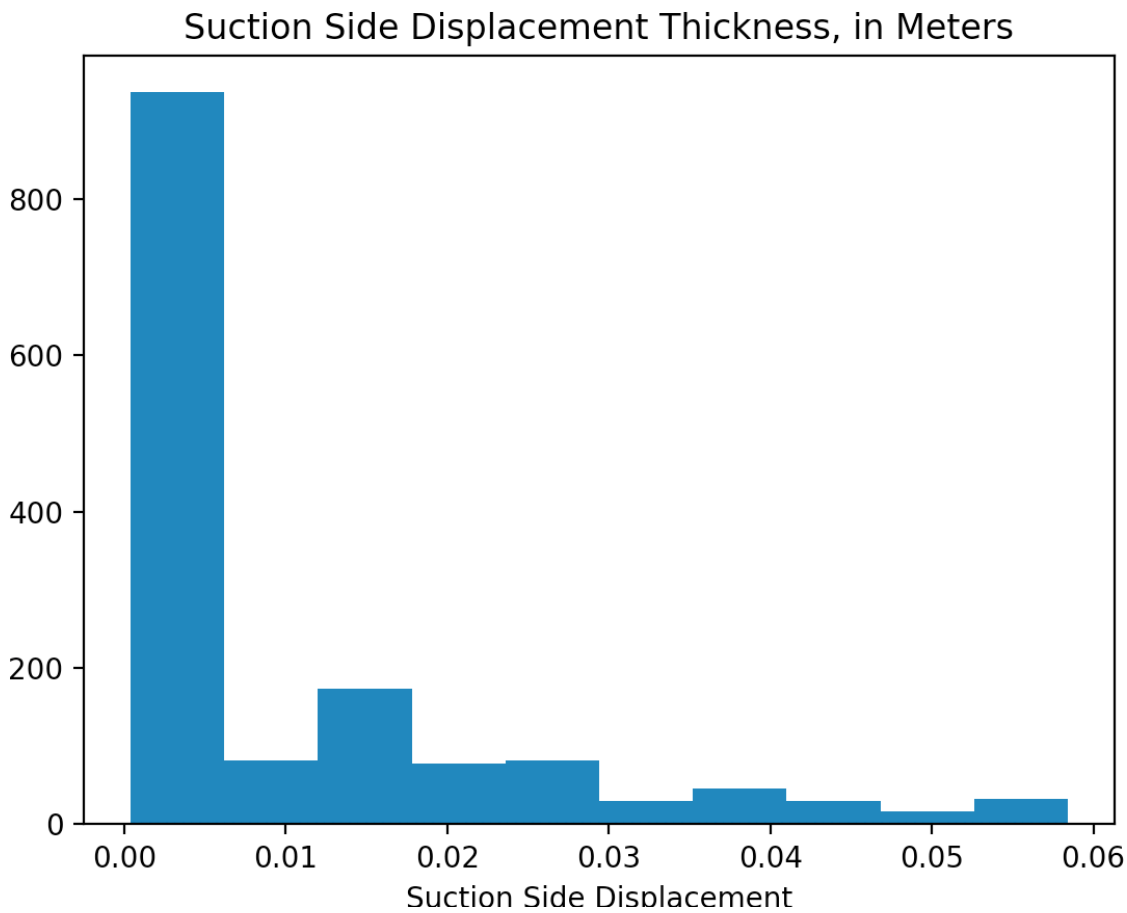
Free-stream velocity is known as the air far upstream of an aerodynamic body, this is before the body has a chance to deflect, slow down or compress any of the air. The median is 39.60 meters per second. The mean value for this data set is 50.86 meters per second. The 25% quartile for this data set is 39.60 meters per second and the 75% quartile is 71.30 meters per second. The standard deviation for this data set is 15.57 meters per second. The minimum value is 31.70 meters per second and the maximum value is 71.30 meters per second. From this box plot you can see that there were no outliers in this data collection for Free-stream velocity.



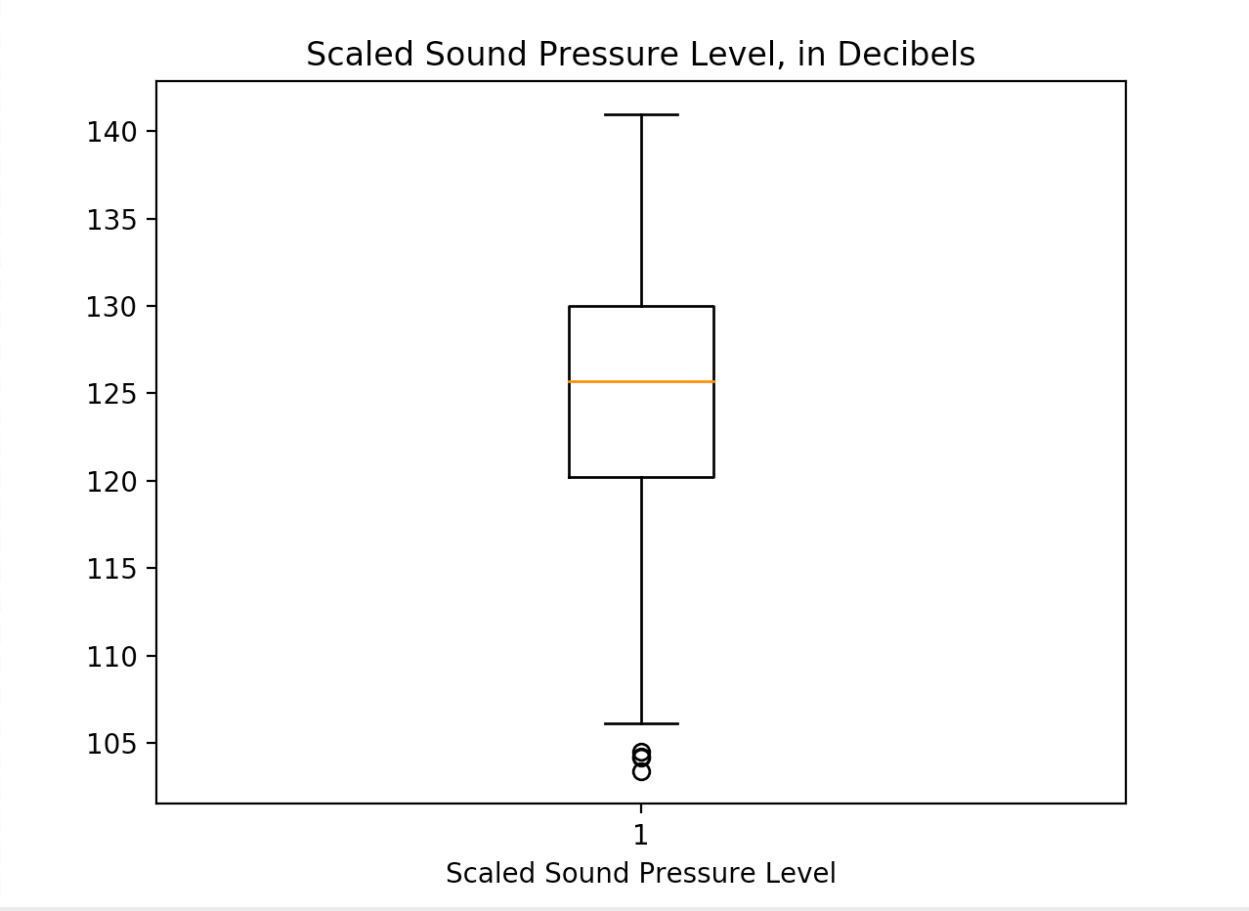
From looking at this histogram we are able to see that this data set looks similar to a bi-model distribution, the reason being is because it contains 2 major data clusters that have roughly the same peak in frequency. It is safe to say that we can infer that these two values (39.60 and 71.30) with the high frequency would be considered our mode for this data set.



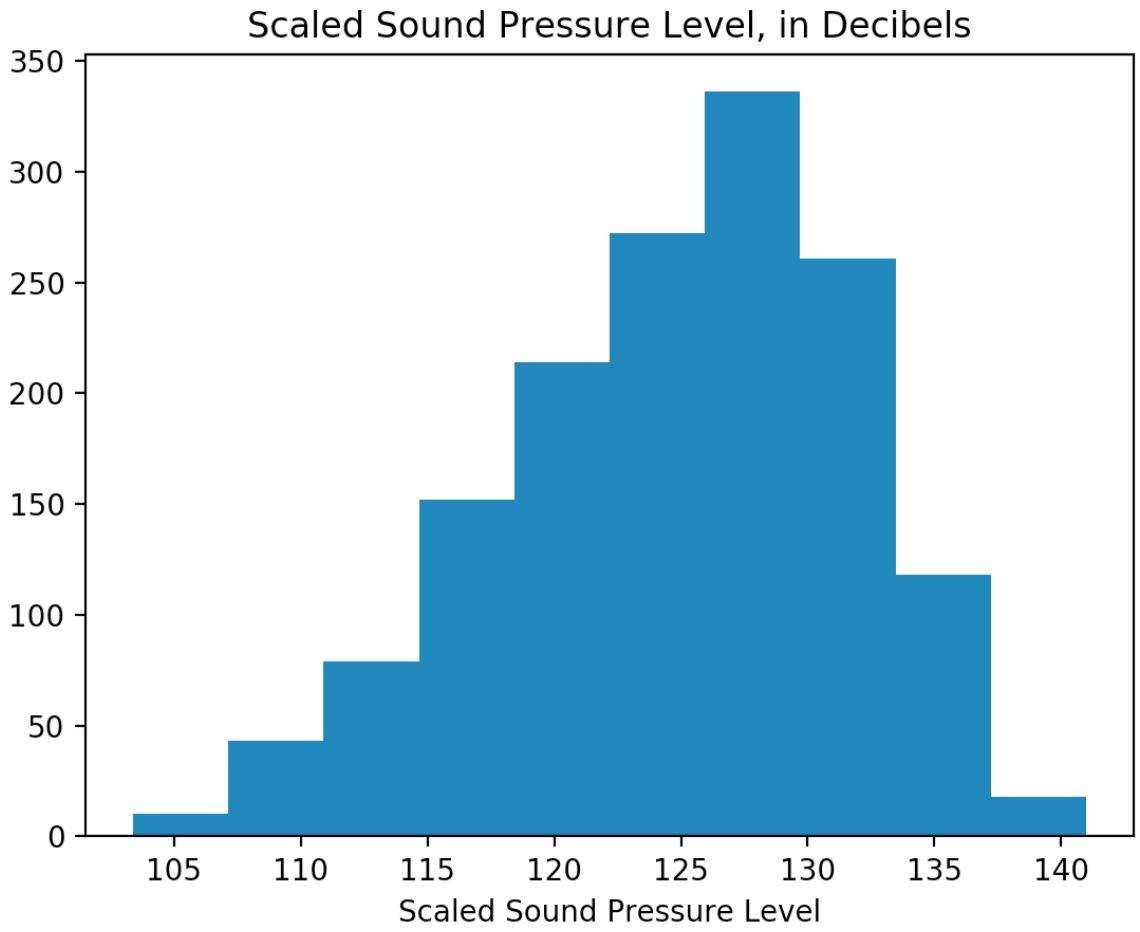
The mean value for suction side displacement thickness 0.011 meters. The median value 0.005 meters. The 25% quartile is 0.003 meters and the 75% quartile is 0.016 meters. The standard deviation for this data set is 0.013 meters. The minimum value for this data set is 0.0004 meters and the maximum value is 0.058 meters. From the box plot we can see that the data set has many outliers as it moves past 0.035 meters.



The data set of Suction side displacement thickness follows a heavy positively skewed distribution. We can see that a major cluster of the data falls 0.002 and 0.004 meters. This is where the mode of the data is located numerically.



Scaled sound pressure level has a mean value of 124.84 decibels. The standard deviation is 6.90 decibels. The median for the data set is 125.72. The 25% quartile for the data set is 120.19 decibels and the 75% quartile is 130.00 decibels. The minimum value is 103.38 decibels and maximum value is 140.99 decibels. We are able to tell from the box plot that there are outliers in this data set. The outliers fall below 105 decibels.



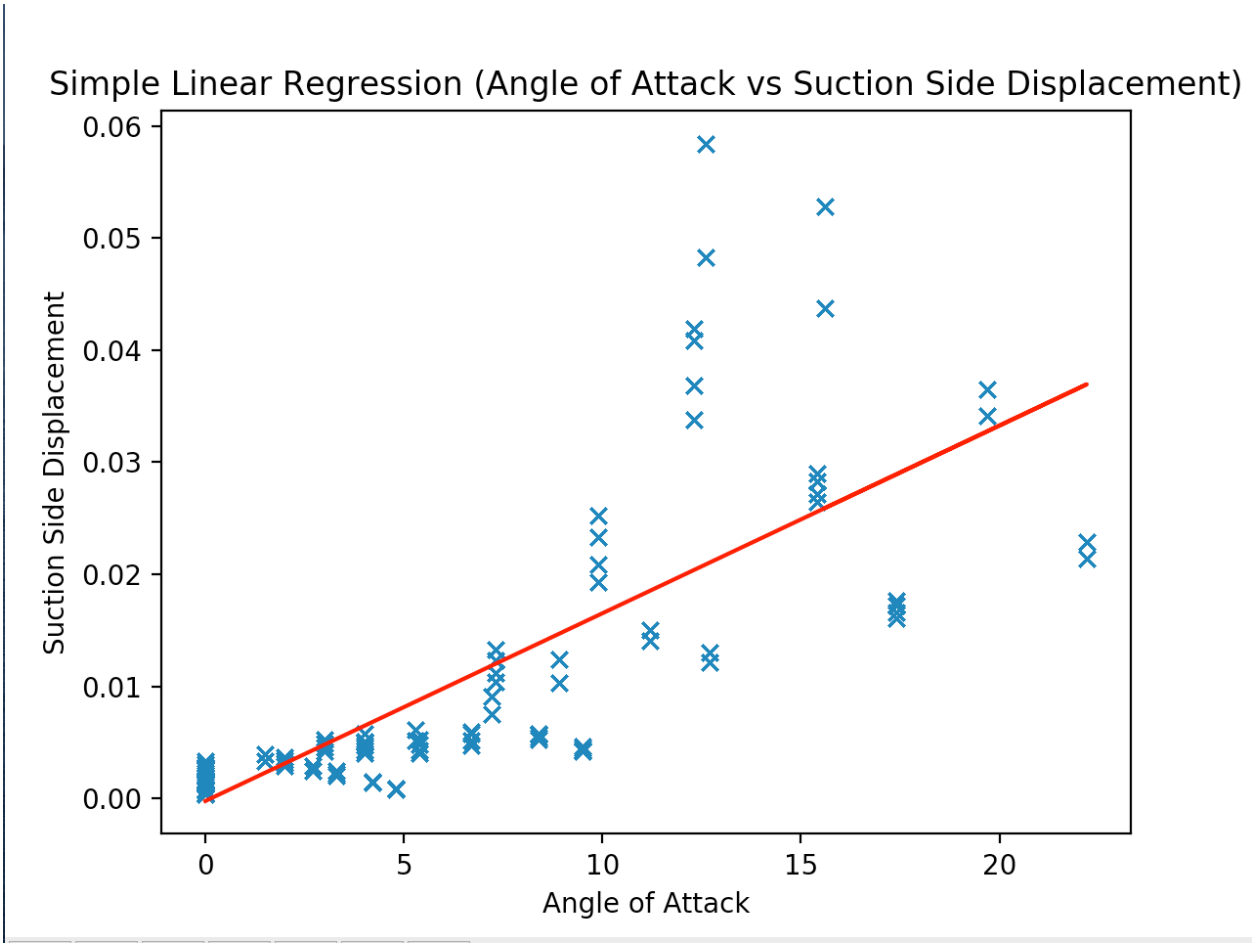
From the histogram above, we can clearly see that the data is slightly negatively skewed, however, the data is very close to a normal distribution and Scaled sound pressure level is the only data category that follows the structure of a normal distribution.

Data

The linear regression model for Angle of Attack vs Suction Side Displacement Thickness

Suction Side Displacement	Angle of Attack
Slope	0.0016740605398028673
Y-intercept	-0.00021410386070689832
R value	0.7533937846545843
R-squared	0.5676021947561581
P value	1.4688854244811825e-275
Std err	3.7713819728467515e-05

Regression Model: Suction Side Displacement = -0.00021410386070689832 + 0.0016740605398028673 * Angle_of_Attack(in degrees)

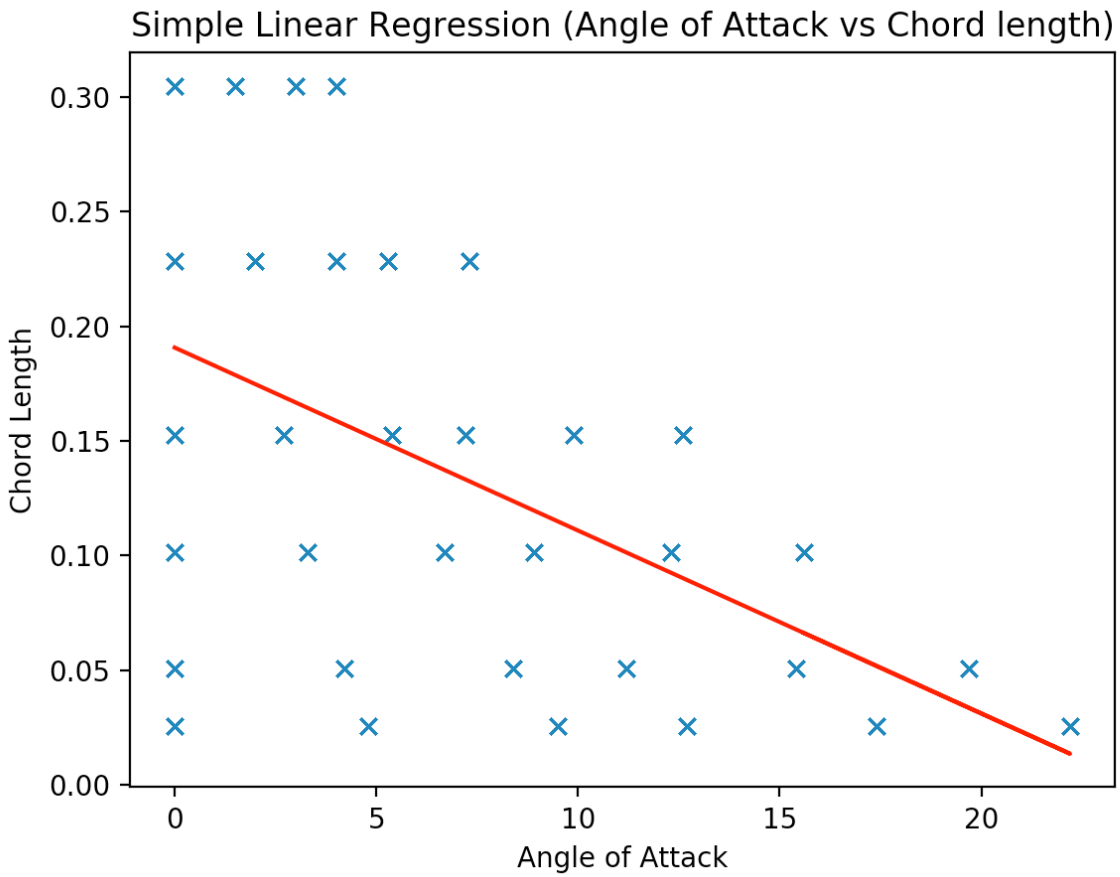


The blue X's represent data points between the two data categories. The red line is used to represent a prediction line for this model. As you can see it follows the data quite very closely when both the X and Y values are low. However, when it increases the value started to drift away from the line.

The linear regression model for Angle of Attack vs Chord Length

Chord Length	Angle of Attack
Slope	-0.007979843197914512
Y-intercept	0.19066994383958813
R value	-0.5048681497642213
R-squared	0.25489184864634823
P value	5.1316183664901356e-98
Std err	0.00035215648893861513

Regression Model: Chord Length = 0.19066994383958813 + -0.007979843197914512 * Angle_of_Attack(in degrees)



The blue X's represent data points between the two data categories. The red line is used to represent a prediction line for this model. The data pairs weaker correlation value and negative correlation value can be easily seen through this data visualization.

Results

The first regression model displayed (Angle of Attack vs Suction Side Displacement Thickness) provided a stronger model compared to the second model. The R^2 value, which is better known as the correlation coefficient squared, had a value of 0.5676021947561581. Now, from the value it can be seen that it falls closer to 1 than 0, which tells us that the regression line falls in a closer proximity to a large number of true data points. However, this R^2 value is still quite low, and it is unlikely that it would be a good form in predicting future data points. The standard error value is 3.7713819728467515e-05, this describes that the average distance of the true value from the line of prediction is quite small. The p-value is 1.4688854244811825e-275 is significantly less than 0.05, this determines that Angle of Attack is a good predictor for Suction Side Displacement. The regression model is formed through this equation.

Suction Side Displacement = -0.00021410386070689832 + 0.0016740605398028673 * Angle_of_Attack(in degrees)

The second regression model displayed (Angle of Attack vs Chord Length) provided a weaker model compared to the first model. The R^2 value, which is better known as the correlation coefficient squared, had a value of 0.25489184864634823. Now, from the value it can be seen that it falls closer to 0 than 1, which tells us that the regression line does not fall in a close proximity to a large number of true data points. This R^2 value is quite low, and it is unlikely that it would be a good form in predicting future data points. The standard error value is 0.00035215648893861513, this describes that the average distance of the true value from the line of prediction is quite small. The p-value is 5.1316183664901356e-98 is significantly less than 0.05, this determines that Angle of Attack is a good predictor for Chord Length. The prediction line, however, does not show a clear, general path for the true data points. The regression model is formed through this equation.

Chord Length = 0.19066994383958813 + -0.007979843197914512 * Angle_of_Attack(in degrees)

Conclusion

Regression analysis has been performed to understand and help predict future aerodynamic and acoustics of airfoil in the second and third dimensions. The use of regression could be quite beneficial for NASA to use in order to obtain a better understanding of the data. Even though there were many data pairs that had very low correlation values. If NASA was able to get strong correlations between variables, they would be able to formulate better regression analysis and predictions.

Resources

https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise#