Final Project

Grace Curtis

Intro to Data Science (DS 210)

Introduction / Stating question

The provided data set is on crickets chirps and the temperature. With this we will try to find out “Can the outside temperature be estimated by the frequency of cricket chirps?”. My personal guess is that the warmer the temperature the more chirps there will be, it being warm and more crickets are active. In the data set the first column is how many chirps happen every 15 seconds (Chips15s) and the second row is the temperature in Fahrenheit (TempFarenheit).

Exploratory Data Analysis

Now that I have the data I need to clean it up. First I would need to remove the missing values. Followed by removing the outliers, which are the values that might not make sense whether it be an error in the system or an exception. Both will be done using Python and Pandas. Using df = df.dropna() will remove all the rows with missing values. Then we get rid of the outliers by first figuring out what the outliers are by finding the median for q2 ,then the median of the lowest value and q2 for q1, and the IQR, which is Q3 - Q1.

First sort the values , then use the following equations to find the values,

Q2 = df.median()

Q1= df[ df <= df.median()].median()

Q3= df[ df <= df.median()].median()

IQR= Q3 - Q1

Low Outliers = Q1 - 1.5(IQR)

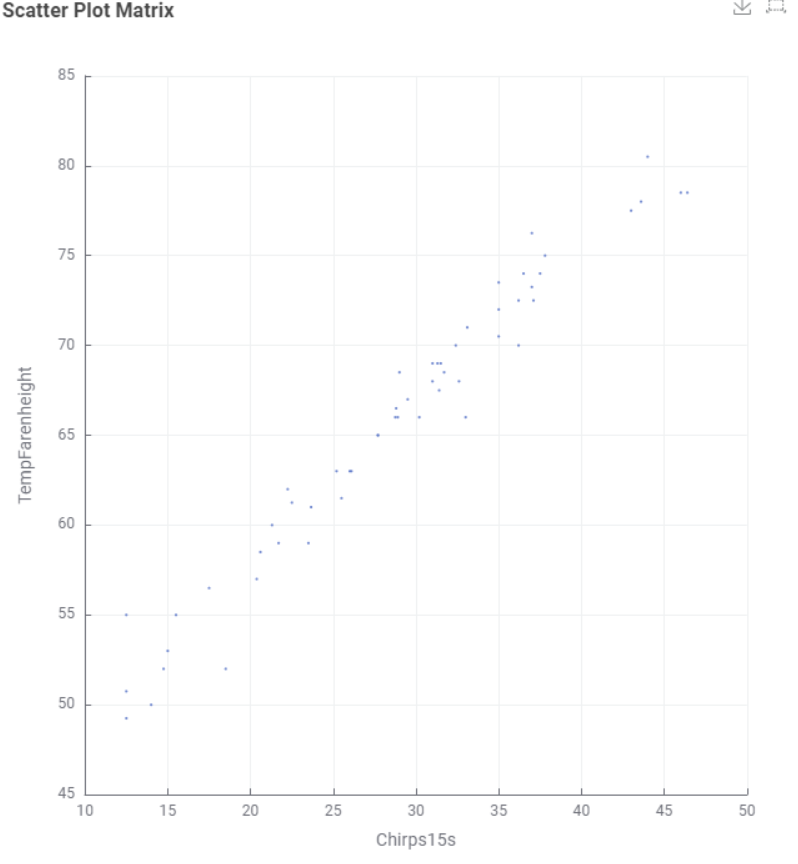
High Outliers = Q3 +1.5(IQR)

,then repeat for the other category.

Once the outliers are found remove them from the list (ex. df = df[df.value < x ] ).

| Value | Q1 | Q2 | Q3 | IQR | Too Low | Too High |
| --- | --- | --- | --- | --- | --- | --- |
| Chirps | 22.5 | 29.5 | 35 | 12.5 | 3.75 | 53.75 |
| Temp. | 60 | 66.5 | 72 | 12 | 42 | 90 |

The mean of the final result is 28.8 for Chirps15s and 65.7 for TempFarenheight which is found with df.mean().



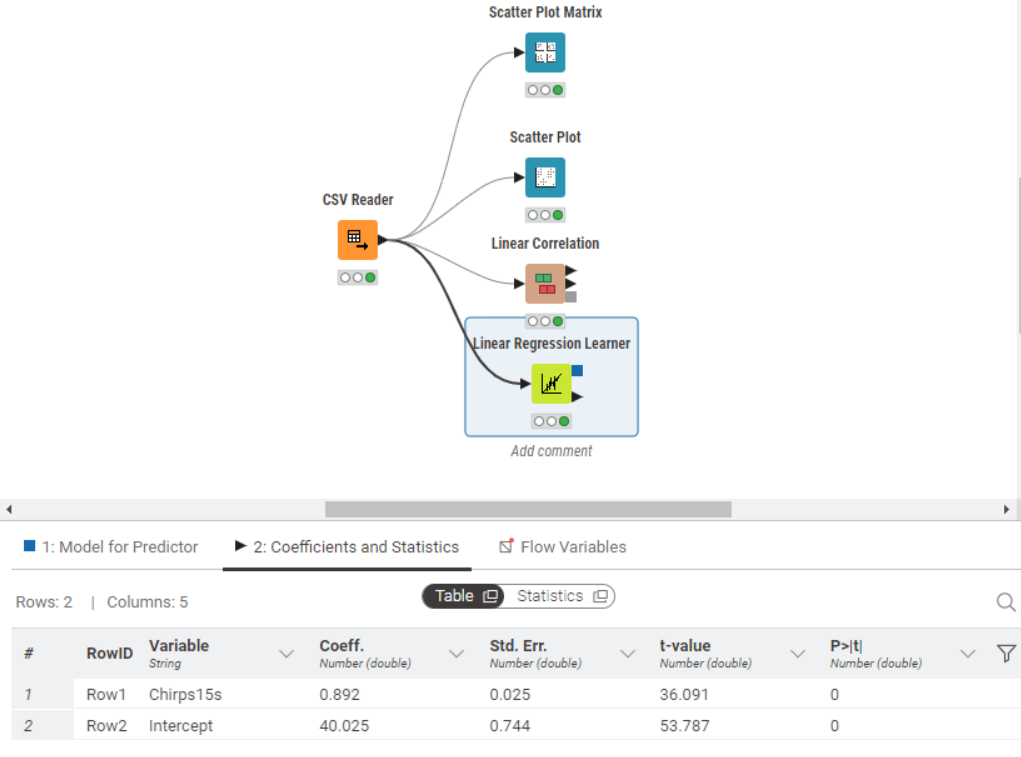
After the data is cleaned, we put the cleaned up file into Kinme. We use the CVS Reader node to put it in and either the Scatter Plot or Scatter Plot Matrix node to create the scatter plot as shown above.The plot showed that when the temperature goes up so does the amount of chirps every 15 seconds.

Refining the Question

Next step is to refine the question if needed. We would also need to consider if there were any unconsidered variables. In this case there was a high correlation and we believe there are no other factors to consider. In this case there is no need to change the question.

Model Building

Next step is to find the Regression line aka the line of best fit. The equation is the same as slope intercept, y = mx + b. The regression line will find a number variable making this supervised learning.



With this we can find the coefficient for chrips ,m, which is 0.892 and the coefficient for the intercept ,b, which is 40.025.

This would make the equation for the line of regression for chirps and temp be y = 0.892x + 40.025. Now we can use this to predict the weather, y , for the amount of chirps (x). For our example we will find the temp for 40 chirps in 15 sec.

y=0.892(40) + 40.025 = 75.705 degrees fahrenheit

Summery

To summarize, we asked the question “Can the outside temperature be estimated by the frequency of cricket chirps?”and determined that there was a correlation. After determining we did not need to change the question ,we then cleaned up the data. Using the cleaned up data we were able to make an equation to help us make a 15 seconds, the temperature would be 75.705 degrees Fahrenheit.