

# Final Project

Camilla Sisemore

## Intro to Data Science (DS 210)

### Introduction and Stating the Question

A scientist, Dr. Margaret LeMone, once wondered if it was true that there was a correlation between the frequency of cricket chirps and the outside temperature. She looked up some formulas on the internet, enlisted the help of her husband and began collecting data so she could begin her experiment and analysis<sup>1</sup>. The data that I was provided and have been working with consists of two columns of data – the number of chirps counted in a 15-second period, and the temperature in Fahrenheit at the time – and 58 rows, determined by using the `df.count()` command in Pandas. I will be attempting to answer the same question with the data I was provided using Python with the Pandas library, and the KNIME program.

### Exploratory Data Analysis

Before embarking on the actual analysis, I used Python with the Pandas library to clean the data. First, I sorted the dataset on the Chirps15s column in ascending order, using the `df.sort_values('Chirps15s', ascending=True)` command. This enabled me to quickly glance at the data and see that there was an obvious outlier, with a value of 361 – if the dataset was larger, I might have used the mean instead. I also noticed one null value, which I validated using `sum(df['Chirps15s'].isnull())`. I looked at the corresponding temperature for the 361 chirps per second value, and it was 73.0° F. Since the subset was so small, I looked at the chirps for other temperatures in the low 70s°F, and decided to replace the 361 with 36.1, making the assumption that it would be a much closer value than using the mean. To do this, I used `df['Chirps15s'] = df['Chirps15s'].replace([361], 36.1)`. I thought about doing the same with the null value, but decided to use the mean of the chirps instead, using the `df.Chirps15s = df.Chirps15s.fillna(df.Chirps15s.mean())` command.

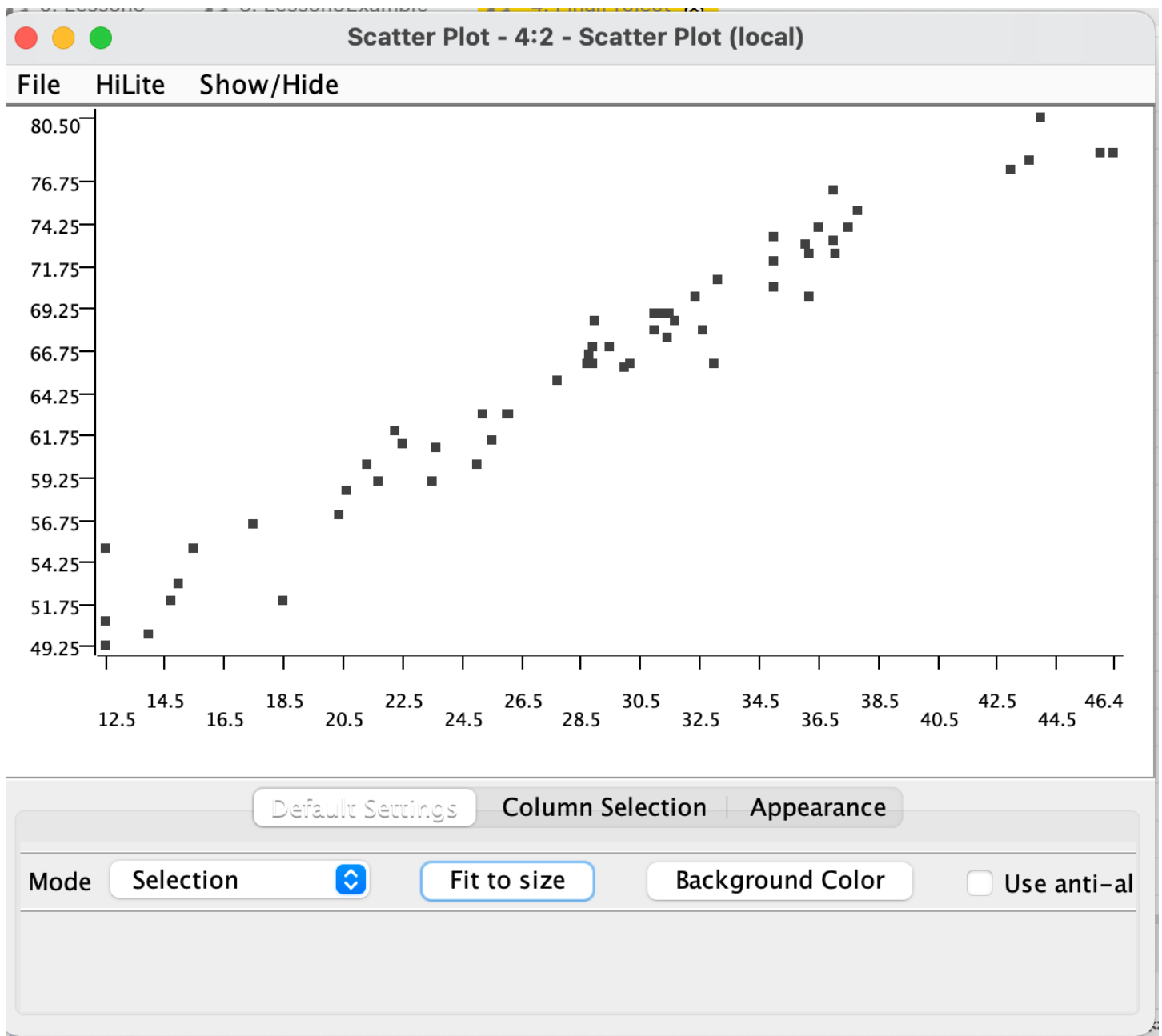
I then tackled the data in the TempFarenheight column using similar methodology. I sorted the data using `df.sort_values('TempFarenheight', ascending=True)`. I found one outlier, where the temp is 6°F, so I looked at the data sorted by Chirps15s, and saw that it fell into where the chirps were in the 25 chirps/15 second range, so I assigned it a value of 60°F (again, if the dataset was larger, I likely would not use this method) using the `df['TempFarenheight'] = df['TempFarenheight'].replace([6], 60.0)` command. There was one null value, which I found by using `sum(df['TempFarenheight'].isnull())`. I replaced the null temperature with the mean by using `df.TempFarenheight = df.TempFarenheight.fillna(df.TempFarenheight.mean())`.

I exported the cleaned data to a new csv file using `df.to_csv('Crickets.csv')`, so that I could import the dataset into KNIME so I could do the remainder of the analysis.

I imported the Crickets.csv file to KNIME, and then added a scatterplot module.

---

<sup>1</sup> [https://www.globe.gov/explore-science/scientists-blog/archived-posts/sciblog/index.html\\_p=45.html](https://www.globe.gov/explore-science/scientists-blog/archived-posts/sciblog/index.html_p=45.html)



### Refining the Question

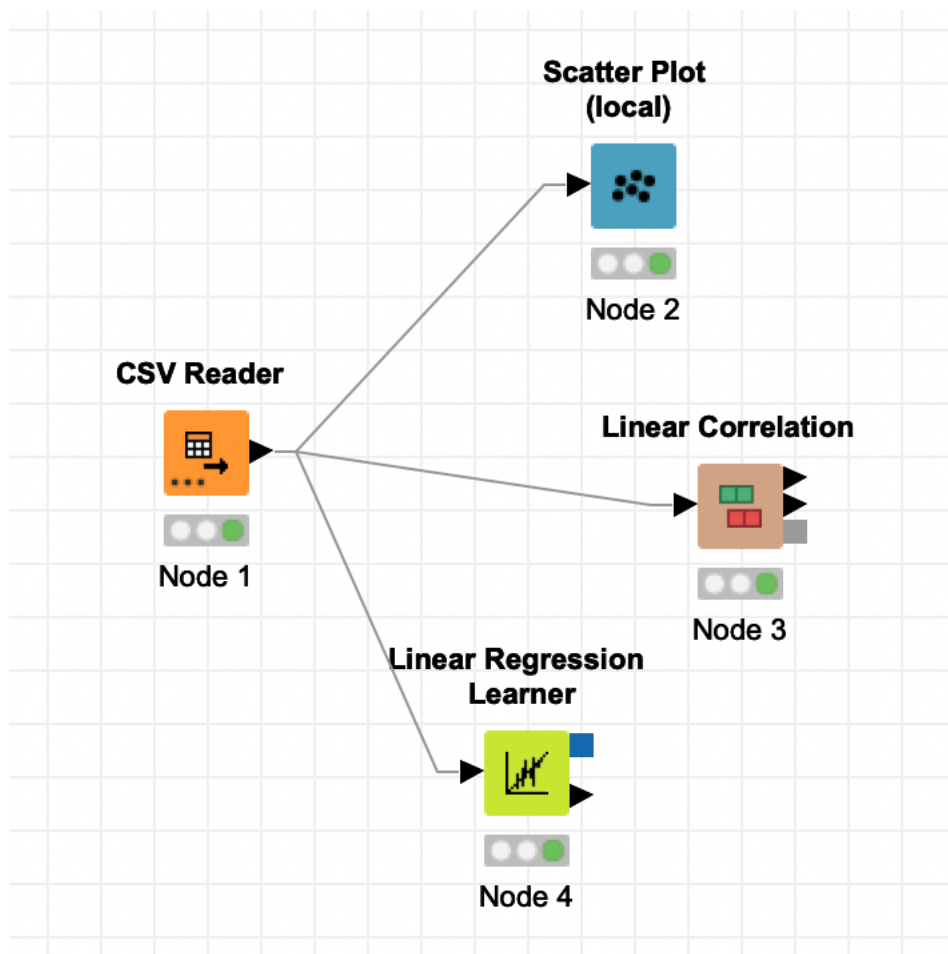
10 pts

Now that I have the scatter plot, it appears that there is a fairly strong correlation between the frequency of cricket chirps and the outdoor temperature. I think the original question can remain as is.

## Model Building

30 pts

Next came the model building. I built a linear regression model using KNIME, so that I would be able to use it to predict the temperature based on the number of cricket chirps I heard in 15 seconds. By using one feature to predict another, this falls under the supervised learning category. Below is my KNIME workflow.

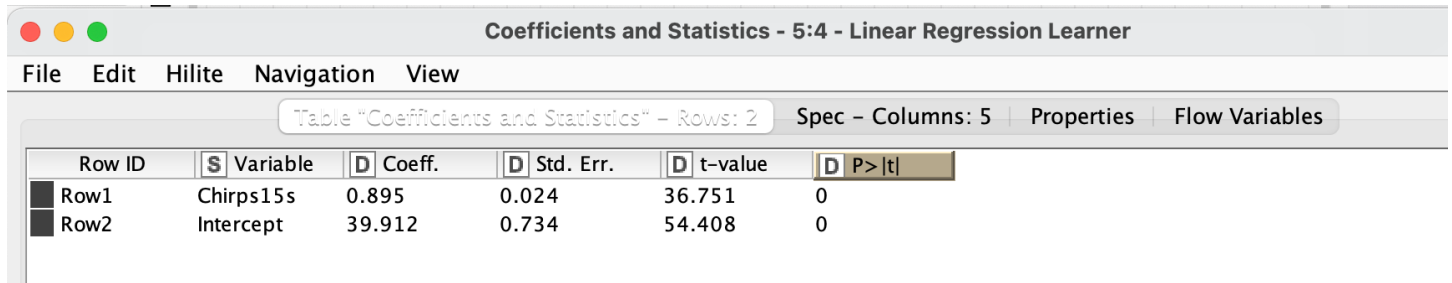


The Correlation measure screen, in particular the Correlation value, tells me that there is a strong correlation coefficient between the Chirps15s and the TempFarenheight values – the closer to 1 (or -1) that it is, the stronger the relationship.

Correlation measure - 4:3 - Linear Correlation					
File Edit Hilite Navigation View					
Table "default" - Rows: 1 Spec - Columns: 5 Properties Flow Variables					
Row ID	First column name	Second column name	Correlation value	p value	Degrees of freedom
Row0	Chirps15s	TempFarenheight	0.97954444009069...	0.0	57

The Coefficients and Statistics window from the Linear Regression Learner module shown below gives the values that I will need to put into my formula so that I can manually calculate the temperature when given a certain number of chirps. This would also be used if I were to create a regression predictor module in my KNIME workflow, along with a

data file containing just Chirps15s values – I could connect the data file to the regression predictor, and connect it to the Linear Regression Learner, and it would do all the calculations for me, using these values.



Row ID	Variable	Coeff.	Std. Err.	t-value	P> t
Row1	Chirps15s	0.895	0.024	36.751	0
Row2	Intercept	39.912	0.734	54.408	0

A regression line is the line that signifies the relationship of the data, usually in a scatter plot. It is only useful if the data appears in a linear fashion. The formula for manually calculating the values is  $y' = mx + b$  where m is the Chirps15s coefficient, b is the Intercept coefficient, and x is the number of chirps that I will be calculating for.  $y'$  will be the temperature.

For example, if the crickets chirp 40 times in 15 seconds one night, I can plug these values into my formula and get an estimate of the temperature outside – which, after rounding, appears to be 75.7° Fahrenheit.

$$\begin{aligned}y' &= mx + b \\y' &= .895x + 39.912 \\y' &= .895(40) + 39.912 \\y' &= .895(40) + 39.912 \\y' &= 35.8 + 39.912 \\y' &= 75.712 \\y' &= 75.7^\circ \text{ F}\end{aligned}$$

## Interpretation/Summary

10 pts

In this project, I have cleaned a small set of temperature and cricket chirps per 15 seconds data using Python and Pandas. I then imported the cleaned data into KNIME and created a workflow that included a scatter plot, a linear correlation node, and a linear regression node that could easily be expanded to be used for modeling future data. I used what I learned from the linear regression node to do the math via equation on what the temperature would be if I heard a cricket chirp 40 times in a 15-second period. Finally, I learned that if an EMP takes out all my electronics, I will be able to know that yes, cricket chirps can tell you the outside temperature (as long as it is over 50°F).