

Final Project

Amanda Spake

Intro to Data Science (DS 210)

Introduction and Stating the Question

10 pts

This dataset from the Globe Program's Science blog post from October 2007 provides data to help answer the question "Can we estimate outside temperature by the frequency of cricket chirps"? This question is hypothesized by using the following formulas:

Temp in degrees F = # of cricket chirps in 15 seconds + 37

This data set has 58 records, starting from August 1st of the year. The temperature value was measured by taking the average temperature from two separate thermometers. The value for cricket chirps was measured by counting chirps for five 30-second periods, averaging the numbers, and then dividing that average by 2.

Exploratory Data Analysis

40 pts

Prior to completing a full data analysis on this data set, I needed to ensure the data set was clean, so I pulled the data into Python/Pandas and created a data set for ProjectDataRevised.csv.

TempFarenheight feature

First I wanted to determine if there were any outliers in the data. For the TempFarenheight feature, I ran the `df.describe()` command to see the statistics for that column. From this calculation I saw that the minimum value was 6.0, whereas the max was 80, 75% quartile was 71.7, 50% quartile was 66, and 25% quartile was 60, so 6.0 was definitely an outlier based on the rest of the data in this feature.

I then ran the `df['TempFarenheight']` command to view all of the values listed in that column and saw there were two values that were outliers, 6.0 and a null value.

To drop the row with the null value, I ran the command `df=df.dropna(subset=['TempFarenheight'])`.

To drop the row with the outlier value, I ran the command `df=df[df.TempFarenheight>6.0]`

Chirps15s feature

For the Chirps15s feature, I ran the `df.describe()` command to see the statistics for that column. From this calculation I saw that the maximum value was 361, whereas the min was 12.5, 25% quartile was 22.4, 50% quartile was 29.8, and 75% quartile was 35.3, so a value of 361 was definitely an outlier for this feature.

As with the other feature, I then ran the `df['Chirps15s']` command to view all of the values listed in that column and saw there were two values that were outliers, 361 and a null value.

To drop the row with the null value, I ran the command `df=df.dropna(subset=['Chirps15s'])`.

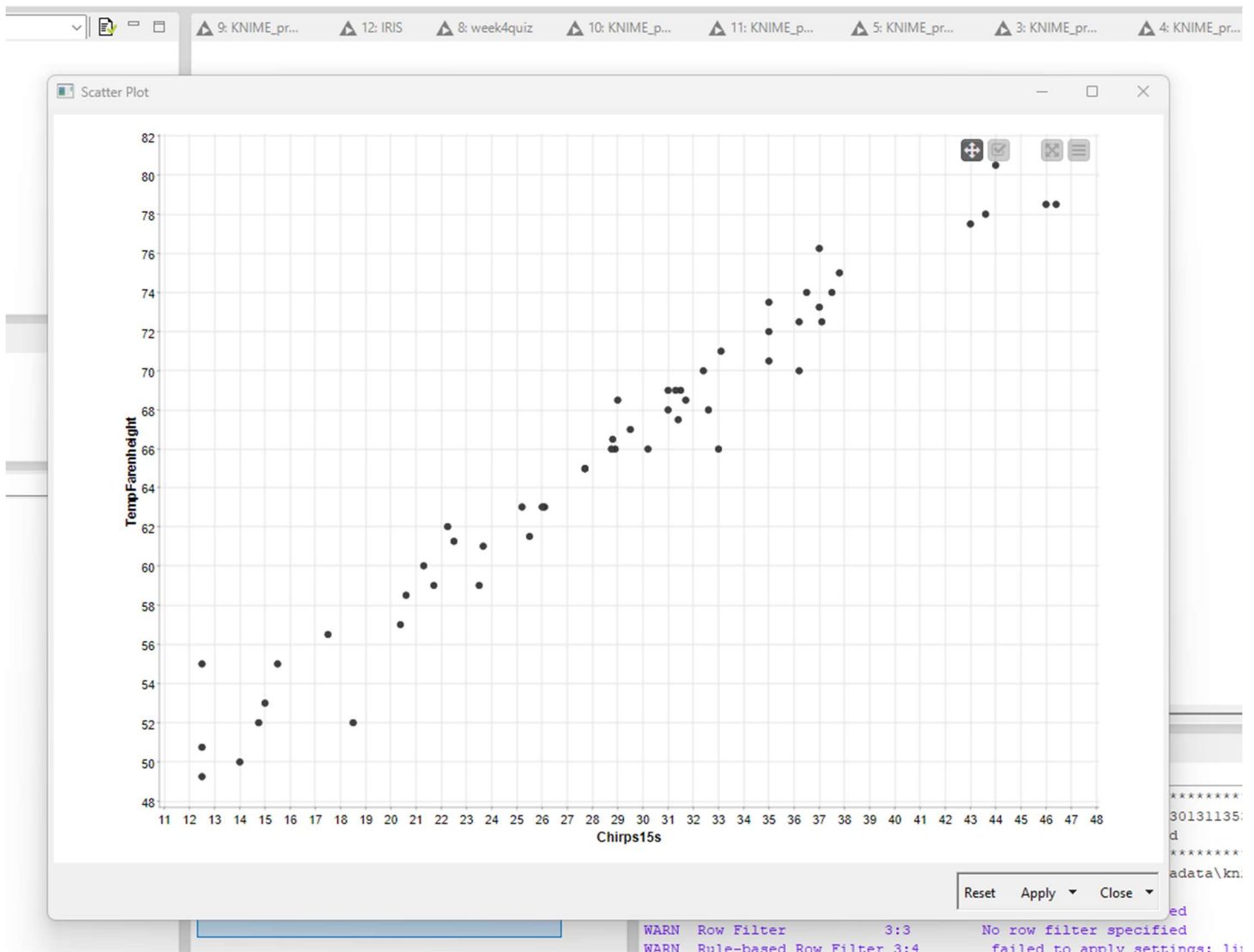
To drop the row with the outlier value, I ran the command `df=df[df.Chirps15s<100]`

Once the nulls and outliers were removed, 55 records remained for the dataset.

Once the data was cleaned, I ran the `df.describe()` command to get the following values:

Chirps15s	TempFarenheight
Mean=28.81	Mean=65.71
Median=29.5	Media=66.5

Scatterplot from KNIME:



Refining the Question

10 pts

In this step, we review our results of the exploratory data analysis. As we can see from the scatter plot, there seems to be a linear correlation between the outside temperature estimates and the frequency of cricket chirps, so there is no reason to change the original question, as this data set provides enough data to answer our original question. However, if that were not the case here, this is where we could review and revise our original question.

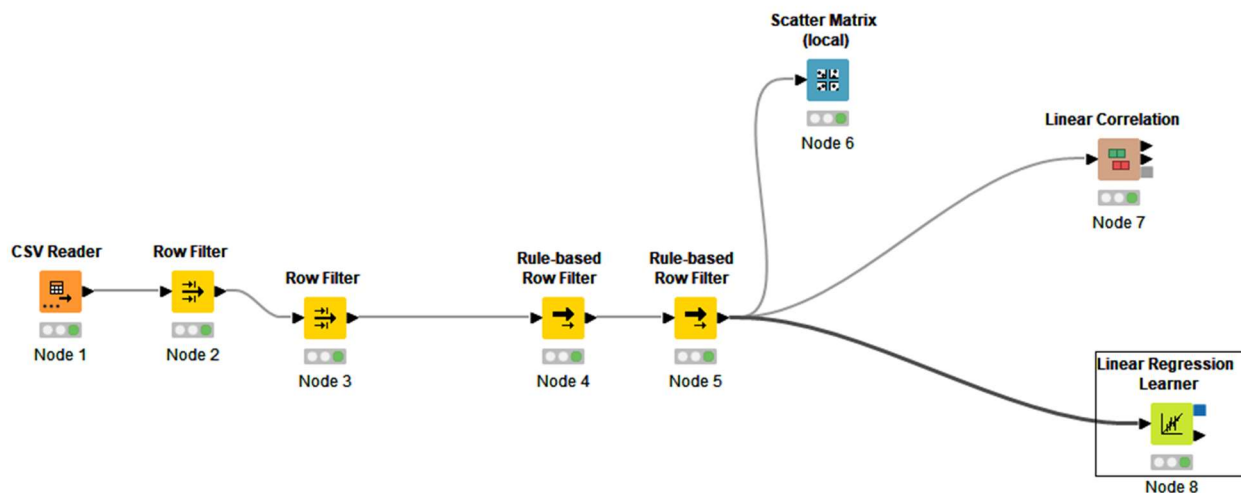
Model Building

30 pts

Now we'll build a linear regression model using KNIME so that we can predict the temperature based upon the number of cricket chirps per 15 seconds. Since we are using one feature (chirps) to predict the value of another feature (temp), this will be an example of supervised learning.

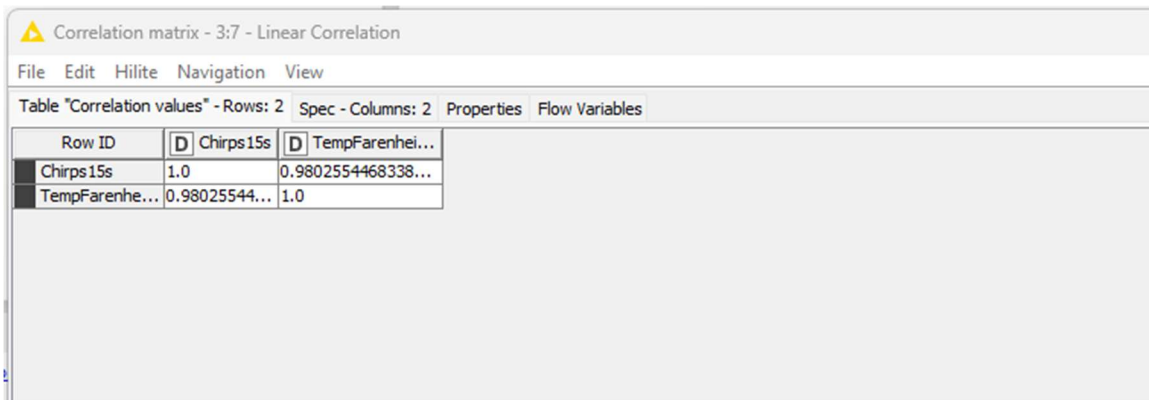
Knime Workflow

▲ 9: KNIME_pr... ▲ 12: IRIS ▲ 8: week4quiz ▲ 10: KNIME_p... ▲ 11: KNIME_p... ▲ 5: KNIME_pr... ▲ 3: KNIME_pr... ▲ 4: KNIME_pr... Welcome to K...



Correlation Coefficient

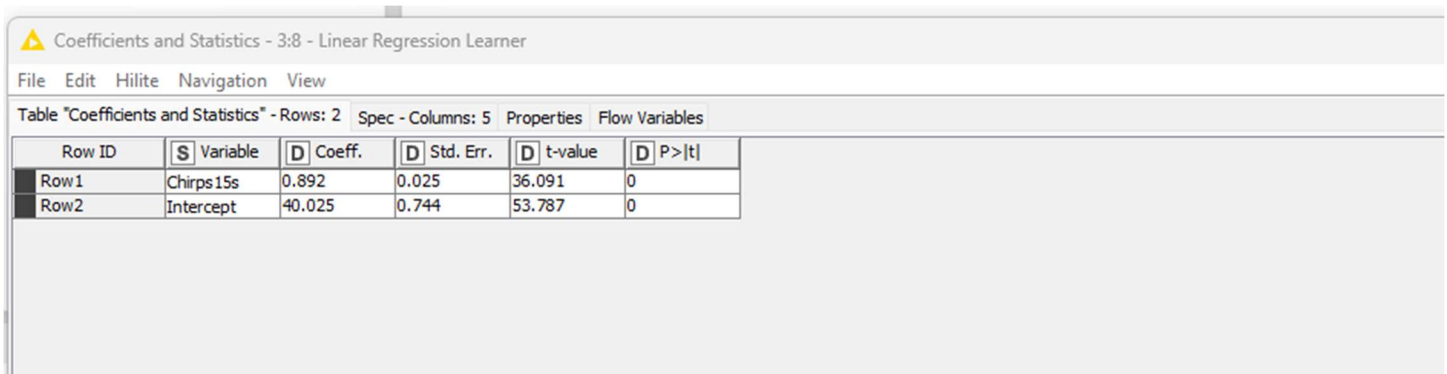
From my KNIME calculations, and the Linear Correlation node, we can see the correlation coefficient for our data is 0.98, which aligns with our scatter plot findings that there is a very close linear correlation between temperature and chirps.



Row ID	D Chirps15s	D TempFarenhe...
Chirps15s	1.0	0.9802554468338...
TempFarenhe...	0.98025544...	1.0

Coefficients and Statistics

From the Linear Regression Learner node, we can see the correlation coefficient, as well as the intercept value. These values will help us to calculate the regression line of $y=mx+b$, where $m=.892$ and $b=40.02$. This regression line will help us to predict values for temperature based on the number of chirps, based on our current data set, and provides us the value where the regression line will intercept the Y axis, based on the value of x.



Row ID	S Variable	D Coeff.	D Std. Err.	D t-value	D P> t
Row1	Chirps15s	0.892	0.025	36.091	0
Row2	Intercept	40.025	0.744	53.787	0

Regression Line equation

$$Y = .892x + 40.02$$

$$Y = .892(\text{chirps15s}) + 40.02$$

For a night when we hear 40 cricket chirps in 15 seconds, we can use this equation to estimate what the outside temperature is, based on our data.

$$Y = .892(40) + 40.02$$

$$Y = 35.68 + 40.02$$

$$Y = 75.7 \text{ degrees for 40 chirps per 15 seconds}$$

Interpretation/Summary

10 pts

From our analysis and regression line, we can estimate outside temperature by the frequency of cricket chirps based on our linear regression model of $y=mx+b$. I have done this by first examining the data set to make sure the data is clean and valid, and then by determining whether there was any sort of relationship between the temperature and chirps features. Once I determined the linear regression equation, it was clear that this equation provided a solid answer to our question, and that we definitely can estimate outside temp by chirp frequency.