Part 1:

This paper compiles data from multiple different studies that look at the relationship between temperature and work performance. This study analyzes all that data to see if there is any correlation between temperature of the workplace and actual work performance.

Part 2:

Section 1: Data & Research Questions

Data:

This paper is https://www.sciencedirect.com/science/article/pii/S036013232100439X?via%3Dihub titled, "Meta-analysis of 35 studies examining the effect of indoor temperature on office work performance." Using data taken from: https://datadryad.org/stash/dataset/doi:10.6078/D1G42R#citations

In this paper, the authors looked at 35 different studies to see if there is a correlation between the temperature of the workplace and the productivity of work done. The authors gathered:

- The temperature of different workplaces on a given day or period of time
- The relative productivity at the workplace on a day or period of time.

The observations are the different participants in each of the studies compiled into this one. The features are the year, region context, climate, participant number, temperature and different behaviors in the workplace.

Different calculations are made with this data to due further analysis with the data provided in the studies and compiled into this one

This is the data set: https://datadryad.org/stash/downloads/file_stream/791518

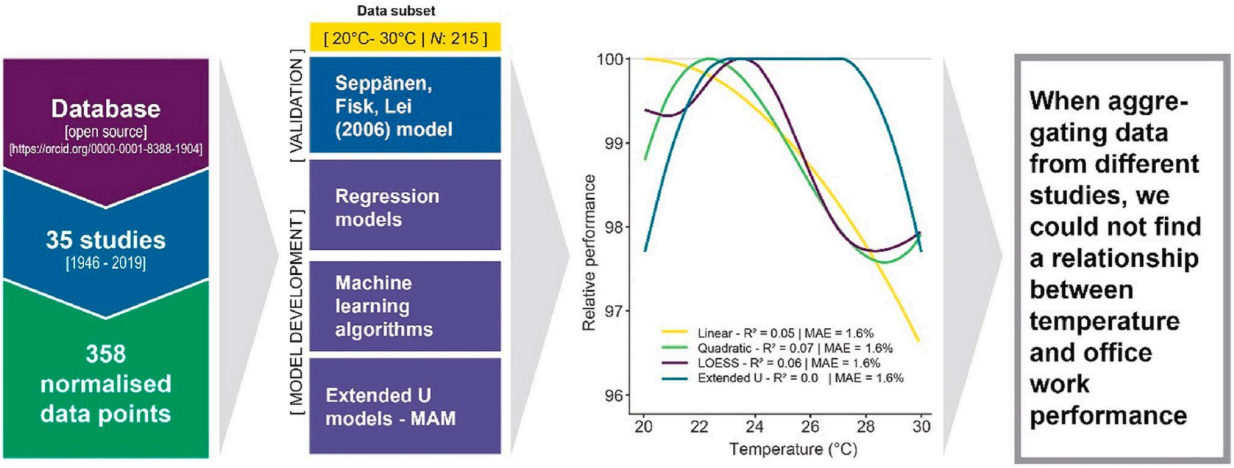
Research Questions:

The paper posits the question whether you can accurately predict work performance based on the temperature of the workplace. The hypothesis being that a comfortable temperature that is not too hot or too cold will lead to a better work environment and therefore, more productivity.

You could look to see whether the temperature of the workplace will impact the productivity of women more than the men. This question can be looked at and checked based on the raw data collected, since one feature is the gender.

Section 2: Visualizations & Summaries

Figure 1:



This figure compares the database, the 35 studies, the normalized data, and the linear control line.

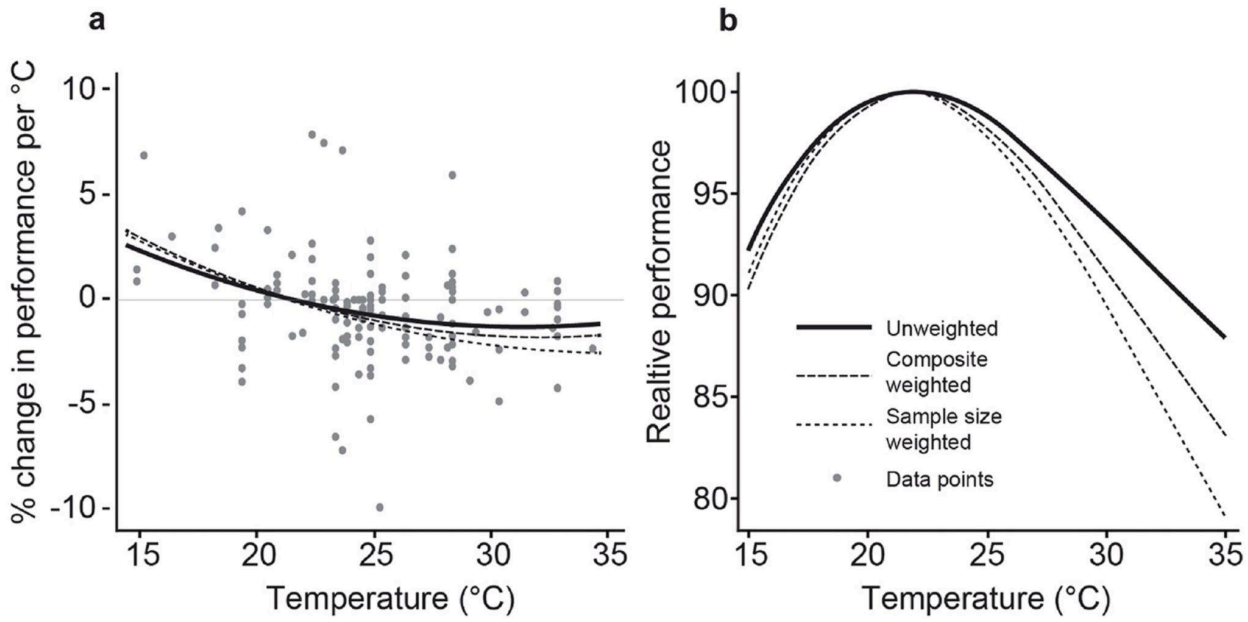
The left margin shows the data subsets that are being shown on the graph. This shows what the control is, and what actual data is being used. It shows the variation between the studies and lets you see clearly the difference between the points of data.

The x-axis shows the temperature in celsius and the y-axis shows the relative performance in the workplace. The data points show how high the work performance was at a given temperature in each of the different data subsets.

The author includes this because it shows you the difference between all the subsets before they generalize it all in the rest of the figures to make a more conclusive data point. This is characterizing the data to add validity and the needed skepticism for the rest of the claims made in the paper.

From this figure alone, I can see that the majority of the graphs seem to have a work productivity peak between 21C-23C, which is right around room temperature. It shows an interesting, very small, dip back up near 30C.

Figure 2:



This figure compares the percentage change in performance based on the unweighted data, composite data, sample size data, and the actual data points. The second graph shows the comparison between unweighted data, composite data, and sample size data based on temperature and work productivity.

Both figures have an x-axis of temperature in celsius. The first graph shows it compared to the y-axis % change in performance per degrees of celsius. The second graph shows it compared to the relative performance in the workplace productivity. The data points show the specific relative productivity found in each of the studies and where they fell on the graph. The lines show how the different data subsets general trend lines are.

The authors of this paper would want to sum up the data in trends to make it more digestible and easy to understand where the majority of each subset show the productivity of work to temperature would be. Showing the percentage that it changes further characterizes the data to show how actually dramatic the changes are, 0 being absolutely no change. It grounds the data to not make it seem so extreme when looking at the graphs.

These graphs show that room temperature seems to be where work productivity is relatively the highest in all data subsets. But, the percentage graph shows that by percentage, the actual improvement to productivity is not actually that big of a difference, the highest difference in regular performance only increasing by less than 5% or decreasing by very close to 1%.

Another Figure:

I think another figure that could be useful for this is showing the comparison of the different data points obtained in each of the individual studies and comparing them on a graph. I think having another table that shows if the changes in workplace productivity changes were generally the same in every workplace would go a long way to prove the hypothesis true or

false. These graphs would either look very similar to the graphs already shown in the paper or it will have a lot more outliers. The x-axis would stay the temperature and the y-axis would stay the relative productivity. The difference would be 35 differently colorful lines that will show a different trendline of the rate in change, on a scatter plot.

Section 3: The Data Sets

The dataset I loaded looks at the study, authors, year, source, country climate, information of participants, and information about temperature.

- There are 358 rows of data in the table and 357 of them are the observations taken from every paper used in the study. They contain the data collected in each study and neatly organize it into one table. There are 28 columns and 24 of them are actual features. The other 4 columns, i.e. study, authors, metric, and SA are descriptors of the data added after the fact, not actual data collected in the study.
- The format is a csv file made into a nice data table. This means the delimiter of this file are commas, as that is the standard delimiter in csv files. It is formatted to show the different data collected from each study and to organize it all into one table. It is essentially a compilation of a lot of other raw data.

Part 3:

- 3.1: Load and clean data
    - Describe steps someone can take to obtain the data files you are working with. For example, provide the link to a the paper's supplementary Excel file, and explain that you opened it in excel and then saved it as a CSV file called "frogs.csv" in the same directory as the notebook (or whatever, as long as what you instruct them to do matches what your code is so your code works). **ONLY IN RARE CASES:** Only if you check with Dr. Melamed first, and you have some data that requires special permission to access, or some other big hassle to access, you can do this some alternate way. Your data should be directly downloadable from the source and not your personal shared folder.
    - Provide code to read in all relevant data files into data frames. Explain your code and why you did it that way. Show the "head" (first few lines/rows/columns) of each data frame.
    - If any cleaning steps were needed at this point, explain these cleaning steps. Otherwise, explain how you checked that the data frames were suitable for the further analyses.
- 3.2: Describe data numerically
    - Provide code to obtain the shape of the data files. Describe how this shape relates to the number of observations and the number of features. Be precise, such as "This data frame has 6000 rows which is the number 500 mice times the 2 treatments times the 6 time points per treatment".

    - Pick **two** features to investigate. For each feature.
        - Explain what you expect the "describe" function would output, based on your understanding of that features. How many observations have a recorded value of that feature and what is the average across observations?
        - Run the "describe" function and compare the results to what you predicted.
- 3.3: Visualizations. Make two visualizations, one for each of the features from in 3.2. They could be univariate visualizations (just the one feature) or more than one feature visualizations. For each one.
    - Describe what kind of visualization you want to make, why this is appropriate for this feature and data set, and how the visualization will provide insight into the data.
    - Provide code and explain your code to make the visualization.
    - Interpret the visualization: compare it to the "describe" function output from 3.2, and explain what insight into the data you can make with the visualization.
    - Describe how your visualization relates to one of the hypotheses or figures from the paper.

3.1

The paper's data is saved as a csv: DRYAD_TOWP_DB01.csv. I downloaded this file into my downloads folder which is the directory I use for all my python notebooks. These files were then opened in the Apple Numbers application to read the csv file.

- The line of code to read in this file was: dryad = pd.read_csv("DRYAD_TOWP_DB01.csv",index_col=0, encoding ="latin1"). I set the data into a variable "dryad". I read in the file using pd.read_csv and write in the csv file. I indexed it with index_col=0 to start at column 0. I originally ran this and got an error saying python could not read in a utf-8 file. So, with a little research I found out that python by default uses utf-8 for encoding and if your file uses something else you need to specify which encoding it is. After some trial and error, I found that setting it to encoding="latin1" worked. Latin1 is a single-byte character encoding which will cover a lot of special characters.

3.2

I used dryad.shape to get (358, 27). This dataframe has 358 rows which are the observations of the different studies and 27 columns of the features that were observed in the participants of the studies.

I'm going to look at the features of Mean Low Performance (PLow) and Low Temperature in Celsius (TLow_C). PLow is the mean performance of the participants at low thermal temperatures. TLow_C Low temperature is the mean low temperature at low thermal conditions in celsius.

- Both these features have 358 observations for them. I expect the mean of the TLow_C to be around 20 degrees celsius, a minimum of 17.9 and the max should be at 89 celsius. I would expect the PLow to have a minimum of 0.14, a maximum of 4487, and I

would expect the mean to be around 220. Which I would expect to be seen in the describe function

- Running the function, observations, the minimums, and the maximums were accurate. My means were close but a little off. TLow_C has a mean of 23 and PLow was 181.6.

3.3

I want to make a histogram and a scatter plot from this data. The histogram I made takes a look at how the temperatures at low thermal conditions to see where most office temperatures at low thermal conditions were. The scatter plot looks at the comparison between PLow and TLow_C. Comparing work performance at low thermal conditions against temperatures at low thermal conditions. This graph looking to see if there is a cluster that shows any correlations between the performance and temperatures.

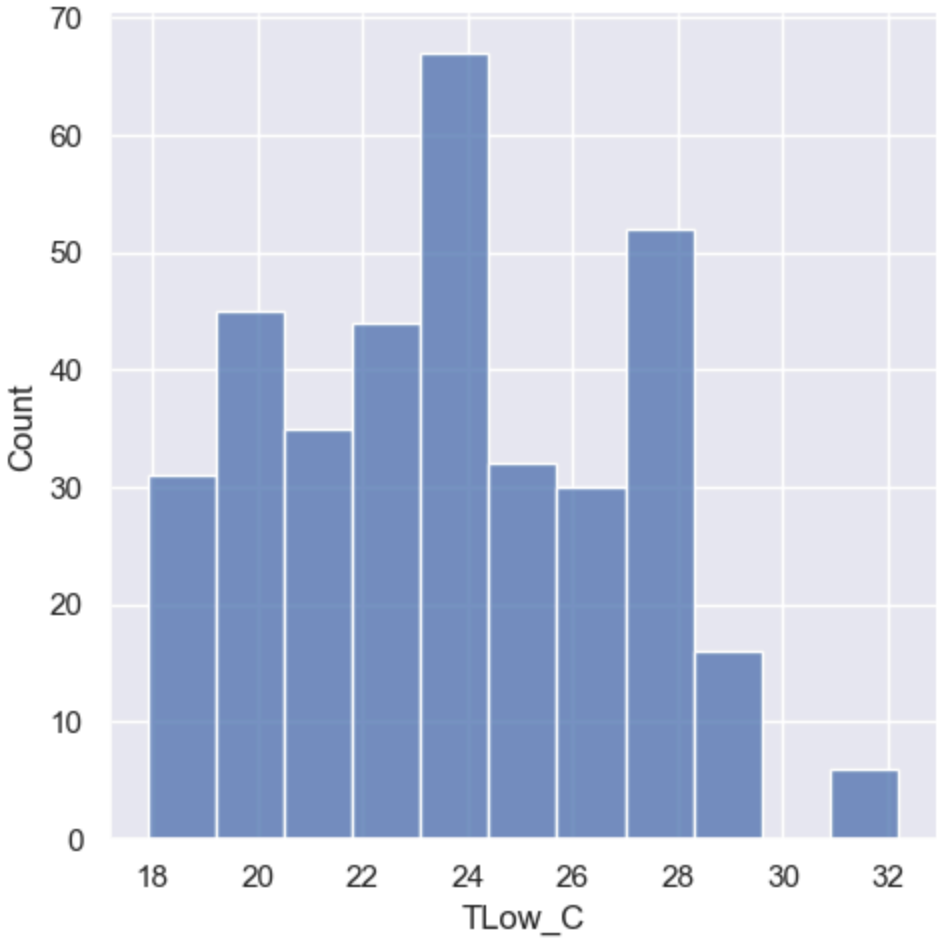
```
sns.displot(dryad['TLow_C'])
```

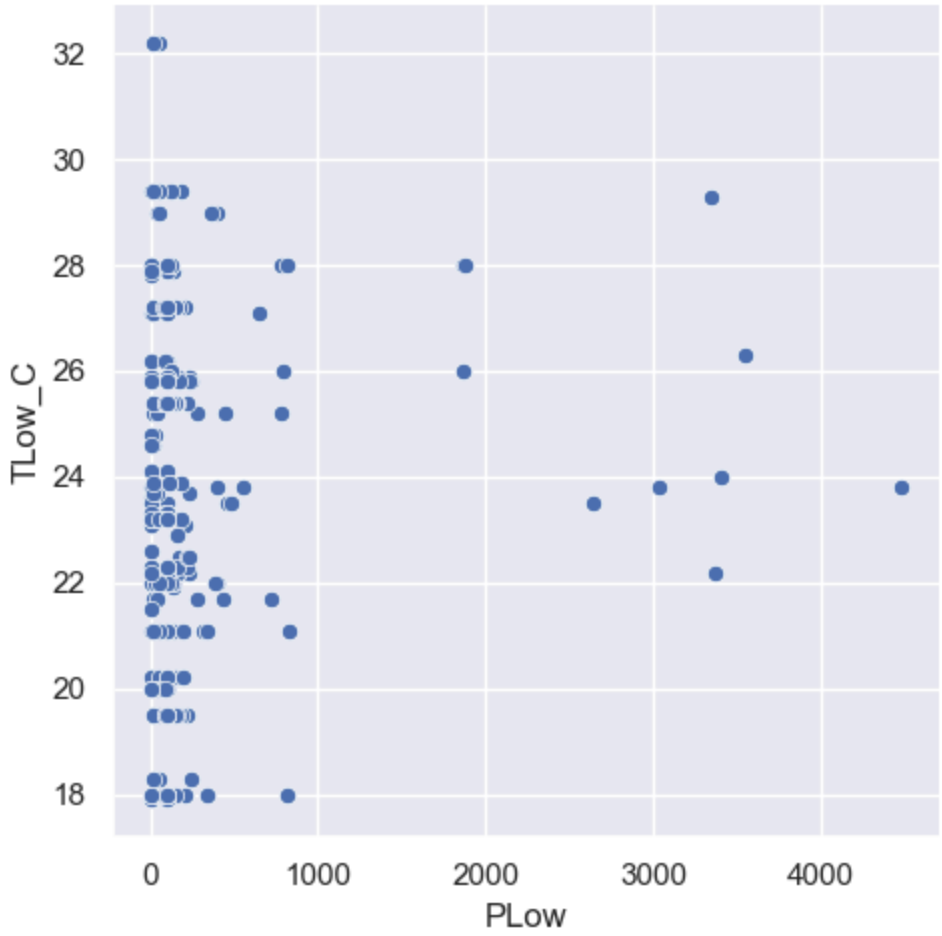
This code creates a histogram pulling from the TLow_C in the dryad data set. This will create evenly spaced bins based on low thermal temperatures.

```
sns.relplot(
    data=dryad, x='PLow', y='TLow_C', kind='scatter'
)
```

This code creates the scatter plot. It uses my data frame dyad pulling the column PLow for x and TLow_C for y. I specified that I wanted this graph to be a scatter plot.

For both graphs the minimum and maximums were expected. For TLow_C histogram, it was described to be a minimum of 17.9 and the max at 89 celsius. Which based on the graph (graph below) is accurate. With PLow (graph below), I expected the PLow to have a minimum of 0.14, a maximum of 448. The mean for the first histogram graph, most of the values are clearly in the bin between 22-24. The mean for TLow_C is 23.5 which is expected. With the second graph you can see the biggest clusters are around 22-24 in track with that mean value. For PLow, the mean expected is 181.6 and a lot of values seem to be around that number on that graph.


70
60
50
40
30
20
10
0
Count
18
20
22
24
26
28
30
32
TLow_C



This data puts into perspective where the normal temperature ranges were in the office and puts in perspective where temperature ranges and productivity intersect. This will allow the researchers to analyze where the biggest impacts in temperature and performance actually are.

The scatter plot demonstrates that most temperatures even higher or lower than room temperature often have the same level of work performance. This goes to proving the theory put out in the paper that temperature is not that important for work performance and not economical to focus on.

CODE:

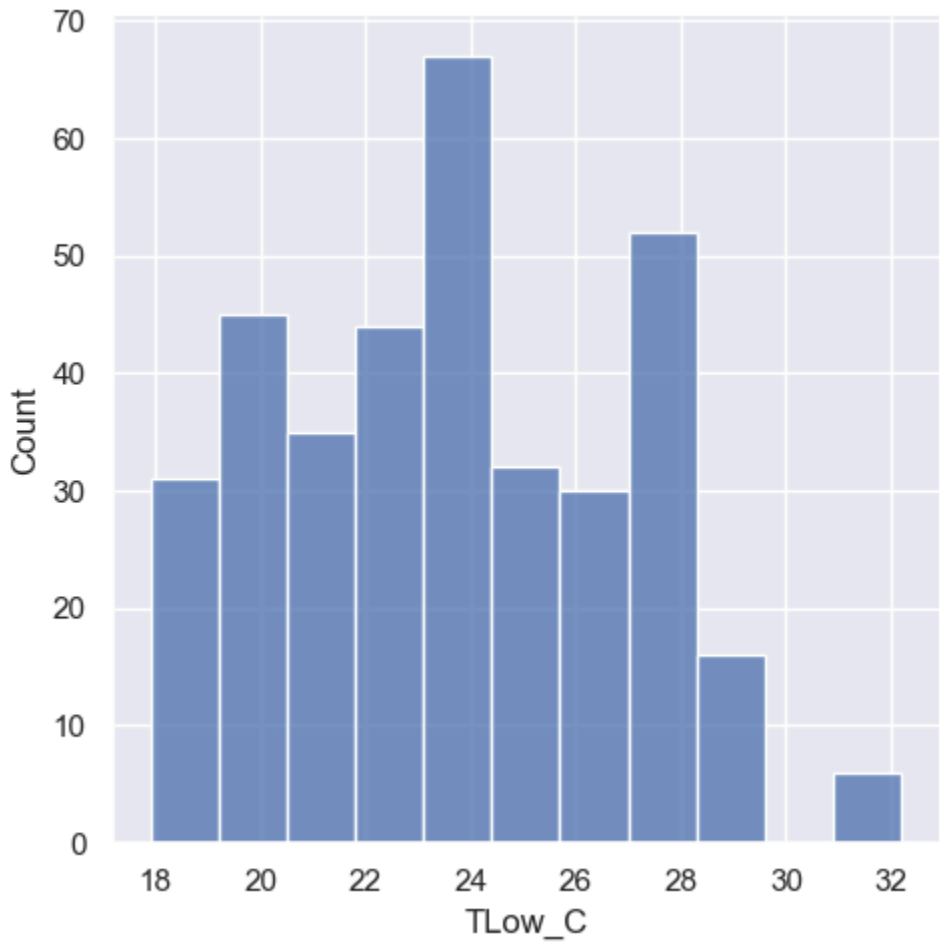
```
import pandas as pd

import seaborn as sns
sns.set_theme()

from scipy.stats import norm

import numpy as np
```

```
dryad = pd.read_csv("DRYAD_TOWP_DB01.csv",index_col=0, encoding ="latin1")
dryad
```

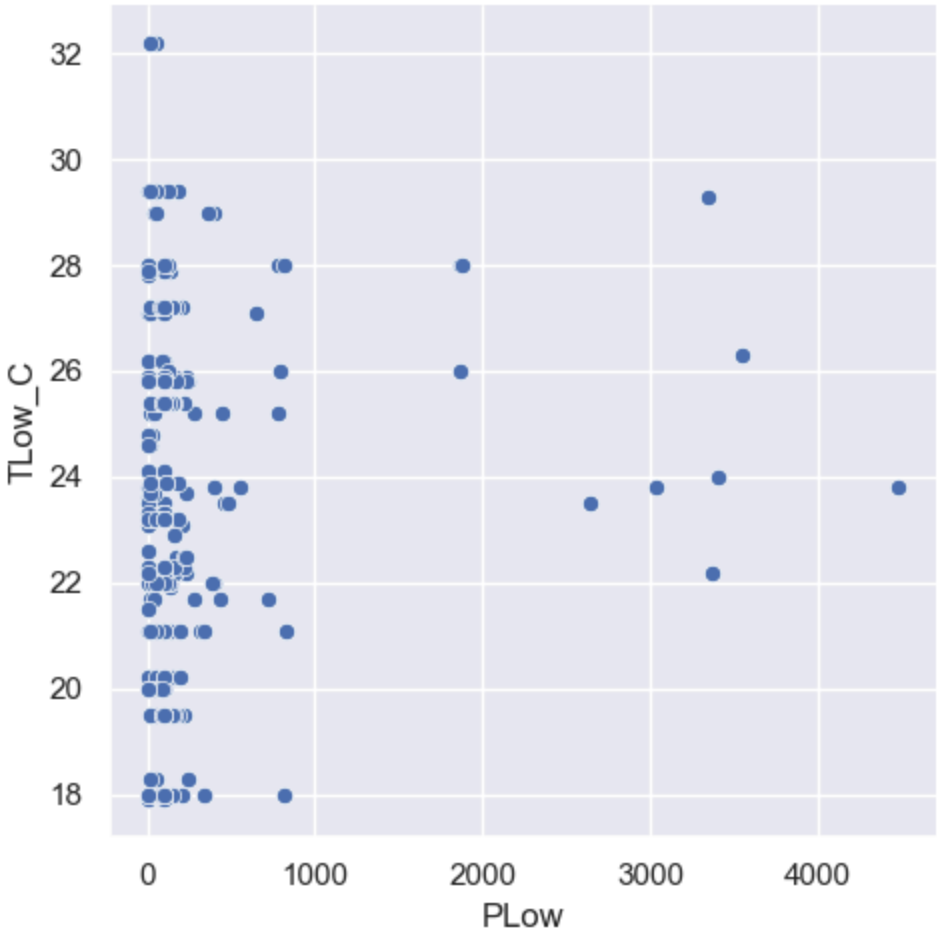
```
dryad.shape
```

```
dryad.describe()
```

```
sns.displot(dryad['TLow_C'])
```




```
sns.relplot(
    data=dryad, x='PLow', y='TLow_C', kind='scatter'
)
```


32
30
28
26
24
22
20
18
TLow_C
0
1000
2000
3000
4000
PLow