Data Statistics: Teacher Guide

Level: Advanced

Subject: Mathematics

Duration: 1 hour

Type: Guided Classroom Activity

Learning Goals:

- Reinforce prior knowledge of statistics
- Calculate statistics on large data sets in Excel

Materials:

- Internet access to download a dataset at School2School.net
- Computer access with Excel or comparable spreadsheet program

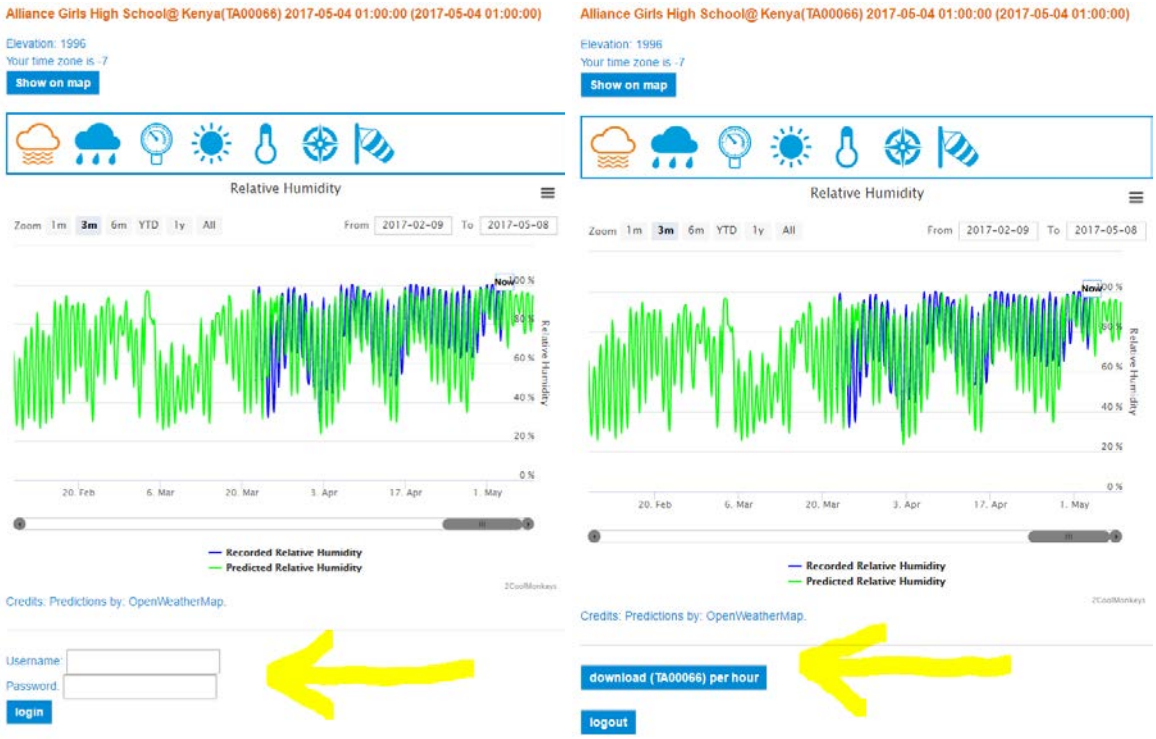
Background:

Discuss with the class why it is useful to mathematically describe our datasets. Descriptive statistics are used to quantitatively describe and summarize and identify trends in the data, particularly for very large datasets where hand calculations are difficult. Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation, the minimum and maximum values, and skewness of the variables. Students should already be familiar with these measures of statistics (minimum, maximum, average, median, mode, and standard deviation) as this lesson plan does not include definitions and hand calculations.

- minimum- the lowest value in the dataset
- maximum- the largest value in the dataset
- mean- the average value of the dataset
- median- the middle number in the dataset
- mode- the value that is repeated the most times in the dataset
- standard deviation- the measures of the spread of the data
- skewness - the measure of the asymmetry of a probability distribution

Methods:

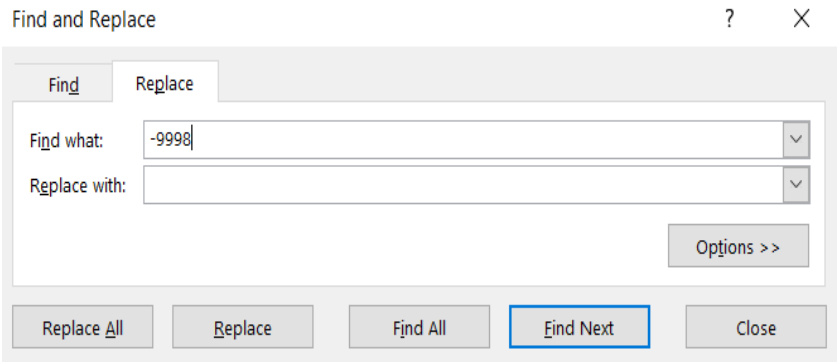
Navigate to the School2School.net website. In the stations tab choose the school that you would like to analyze. Directly below the plots there is a section to fill out the username and password associated with that site. The login information is site specific, so if you were given login information it will only work for your one station. After typing your username and password and clicking the "login" button, the page will reload and a new button to download the hourly data will appear. Choose that button and the hourly data for the station will download in a CSV format. This format is compatible with spreadsheet programs like Excel. Open the downloaded data in your spreadsheet program. [Note: There is an instructional video with step-by-step directions on how to download the TAHMO station data on the S2S website].



Look through the data, are there any values that seems wrong? Try plotting a histogram of one of the variables, do you notice any values that aren't realistic? Take relative humidity for example, was are the bounds for this variable? Can there be negative relative humidity? [Answer: No, relative humidity is expressed as a percent as is bound from 0 to 100, therefore any values outside of this range are suspicious.]

For TAHMO stations, any measurement with a value of -9998 is an error message. It is common for instruments to only be able to log numerical values, so when there is an error in the sensor it assigns a default error value. Any cell containing a value of -9998 should not be considered in analysis. Ask the students why do you think that the TAHMO engineers chose a number like -9998 to be the default error value? [Answer: A very large negative number like -9998 is easy to find in this dataset, and will not be mistaken for a measured value because it is outside of the range of possible measured values; that is why it was chosen]

In order to exclude -9998 error value from the analysis, we can delete all cells with this value and replace them with blank cells, signifying no data (replacing the value with a 0 would indicate the measured value was actually 0 and we don't want that). To do this, type Control + H while your cursor is on your Excel sheet. This will bring up the replace command in Excel, type -9998 into the find bar and leave the replace bar blank, click replace all.



Now our data is ready to use. Start by creating a table in your spreadsheet with the following headings, students will have a similar table on their student worksheet to fill in as you do the spreadsheet calculations.

	Relative Humidity (%)	**Precipitation (mm/hr)**	**Pressure (kPa)**	**Solar Radiation ($W/m^2/hr$)**	**Temperature (^oC)**	**Wind Speed (m/s)**
min						
max						
mean						
median						
mode						
standard deviation						

While descriptive statistics can all be computed by hand, but since our dataset is so large it is convenient to use Excel's built in functions instead. The minimum function in Excel is "min"; to implement this in Excel click on an empty cell and type "=min(datarange)" but instead of datarange choose all of the data for one variable. Below is an example of each of the statistical Excel functions calculated for humidity. Calculate the statistics for each variable. Note: Wind direction entries are not numeric, therefore the statistics don't work for this variable. Repeat the process to calculate the statistics for each variable and fill in the table shown above.

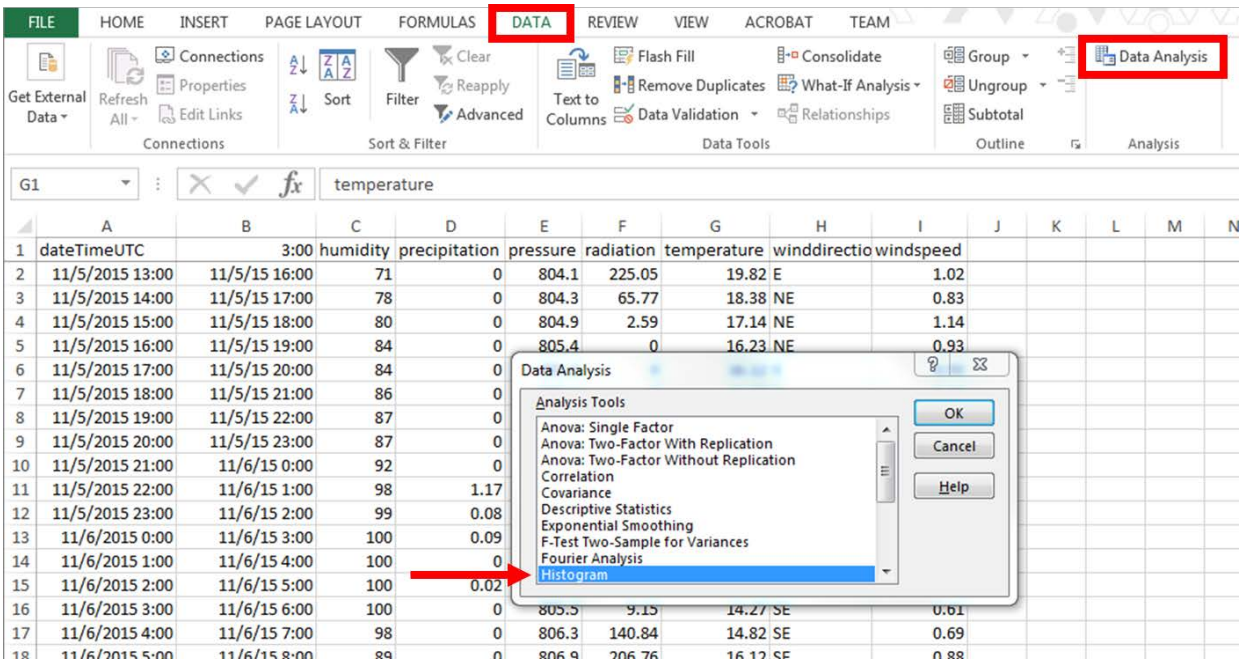
	A	B	C	D	E	F	G	H	I	J	K
1	dateTimeUTC	humidity	precipitation	pressure	radiation	temperature	winddirection	windspeed			humidity
2	11/5/2015 13:00	71	0	804.1	225.05	19.82	E	1.02		min	=MIN(B2:B13073)
3	11/5/2015 14:00	78	0	804.3	65.77	18.38	NE	0.83		max	=MAX(B2:B13073)
4	11/5/2015 15:00	80	0	804.9	2.59	17.14	NE	1.14		mean	=AVERAGE(B2:B13073)
5	11/5/2015 16:00	84	0	805.4	0	16.23	NE	0.93		median	=MEDIAN((B2:B13073))
6	11/5/2015 17:00	84	0	805.8	0	16.12	E	0.93		mode	=MODE.SNGL(B2:B13073)
7	11/5/2015 18:00	86	0	806.3	0	15.69	E	0.63		std	=STDEV.P(B2:B13073)
8	11/5/2015 19:00	87	0	806.5	0	15.42	SE	0.62		skew	=SKEW(B2:B13073)
9	11/5/2015 20:00	87	0	806.5	0	15.83	E	0.66			

Once the students have filled in their table, take a look at the different values and discuss with the class what they notice. Which measure of statistics is used to describe variability? [Answer: standard deviation is the measure that describes the amount of variation in the data, with low standard deviations indicating data points that are close to the mean while high standard deviation values indicate the data points are spread out]. Which parameter had the most variability?)? [Answer: Solar Radiation has a large range and the most variability]. Discuss with the class why they think solar radiation has such a large range of values. Ask them to think about the nature of the variable, what values can be expect on different timescales (hourly, daily, monthly, yearly, or other). [Answer: Solar Radiation has diurnal (daily) variation, with values of zero at night and large values during the day]

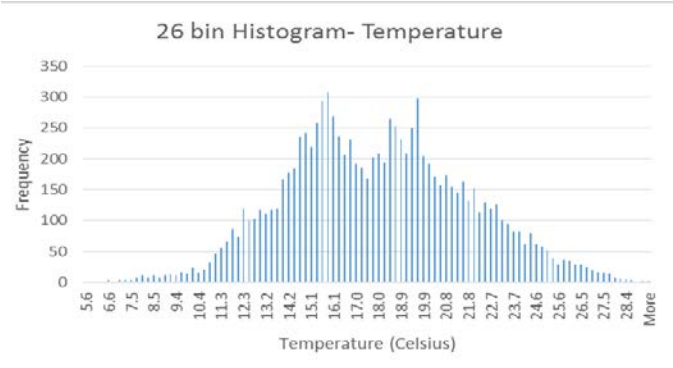
Brainstorm as a class which variable they think would have the least variability, be sure to explain this in context. For example, a student might say that they expect precipitation to have the least variability because most days in the dry season it doesn't rain and most days in the wet season it does rain OR they expect pressure to have the least variability because the altitude stays constant. This is only a brainstorming question, there is not correct answer. Compare the student predictions with the tables, do the numbers support your hypothesis or is it hard to tell? Many students might find it difficult to make find information to prove or disprove their hypothesis based on a table. What other ways of visualizing data might be more helpful than a table? Why is it useful to graphically describe our dataset? [Answer: Figures like a histograms are useful to represent data if you are trying to summarize information of display patterns in the data]. Compare and contrast different reasons to describe your data in a table to a graph. Identify and describe one data statistic that is easier to identify in the table than in the histogram. [Answers may vary: example could be the shape of the distribution is easier to see in a graph than in a table.] Identify and describe one data statistic that is easier to identify in the histogram than in the table. [Answers vary: example could be tables show an exact number as the mean but with the histogram you can only estimate.]

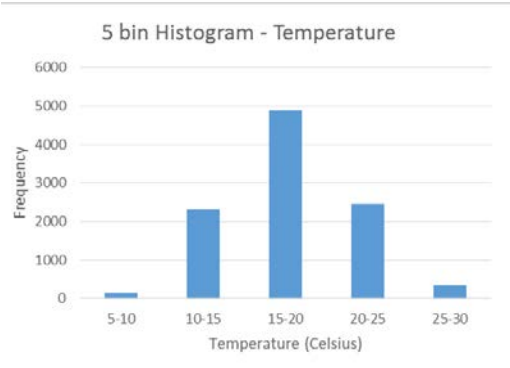
Next we are going to create a histogram for a weather variable of our choosing. You can repeat this process for as many weather variables as your class has time for. A histogram is a column chart that displays the frequency of different values. If you are using Microsoft Excel, you will need to install the Analysis ToolPak. To do this, on the File menu, click Options. When the Excel Options window pops up, select Add-Ins from the menu on the left hand side. Next click on the Manage list, select Excel Add-ins, and then click Go. In the Add-ins dialog box, click the box next to the row that says Analysis ToolPak and then click OK. Once you have installed the Analysis ToolPak, you will be ready to plot histograms. In other program such as Google Sheets, you have the ability to directly create histograms.

After you have the Analysis ToolPak installed, under the Data tab, choose the Data Analysis button. When the Data Analysis dialog box opens, choose the row Histogram and click OK. In the Histogram dialog box, select the data range of the variable you want to look at and click OK (you can also choose how many columns (bins) you want and where you would like the output results). A new sheet will open up with the Bin and Frequency headings.



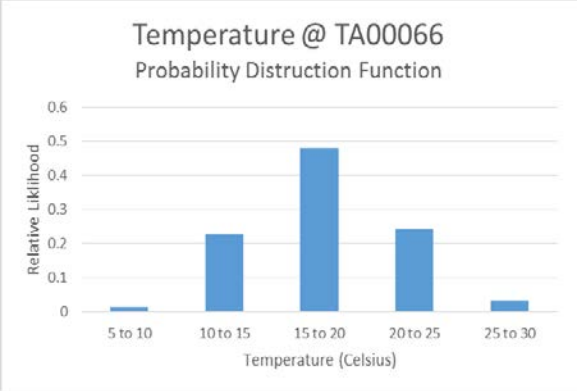
Open the new worksheet and select both rows, in the Insert tab select a column chart. In the new histogram created, label the axis appropriately. Below are two examples of a histogram for Temperature with different bin values (one with 26 bins and one with 7 bins) [Please note that these graphs may different for different weather stations based on local climate and weather- the graphs below are for reference only and use TA00066]. The number of bins dictates how detailed the distribution is. The advantage of having many bins is very details distributions, while the advantage of having only a few bins is that you can easily identify the overall trend and expected value.

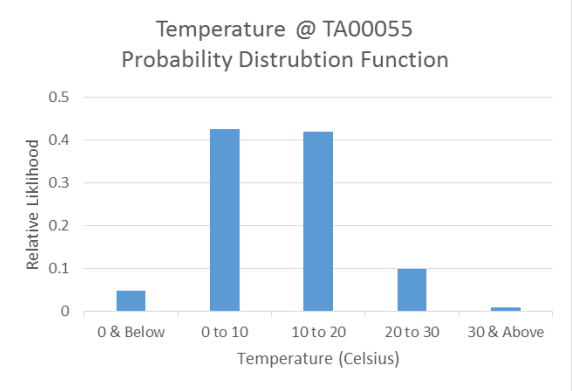




A histogram is a convenient way to summarize the data to identify statistical descriptors like minimum and maximum, while also allowing you to see the type of distribution of the data (uniform, binary, single peak, two peaks, normally distributed, etc). The histogram gives you an idea of the

long term expectation of the expected values for that variable. We can redraw a histogram by changing the y-axis from total frequency to relative frequency by dividing the frequency for each bar by the total number of samples. The benefit of using a histogram with the relative frequency is that we can compare graphs quickly even if they have different sample sizes. A histogram with the relative frequencies, also called relative likelihood, is commonly known as a probability distribution function (PDF). The area under the PDF curve is equal to one, representing that the graph shows all the given values for that variable. Create a PDF for one of the weather variables that you already have a histogram for. As a class think about how you interpret a PDF. Using the example PDF for TA00066 of temperature below, a correct interpretation could be that about 48% of the time the temperature is between 15 and 20 degrees Celsius OR less than 5% of the time the temperature is below 10 degrees Celsius or above 25 degrees Celsius. Ask the class, if they didn't know the weather forecast for tomorrow but they did have a PDF of past values, how would they make their predictions for tomorrow's temperature? [Answer: for station TA00066 students might guess that the temperature will likely be between 15 and 20 degrees Celsius but it could be as low as 5 or as high as 30, for station TA00055 stations might guess that the temperature is 80% likely to be between 0 and 20 degree Celsius]





What are other ways we might use to describe the data that we didn't already use? [Answer: There are many answers to this question. Some examples are: using plots like histograms and boxplots, variance, range, number of samples, skewness, five number summary, etc]

We looked at the entire dataset for our analysis, but we could have only looked at a subset of the data. What would be one useful reason to only look at a subset of the data? [Answers may vary: you could look at only day/night/seasonal temperatures, pressure, wind speed, or humidity to understand variations on a smaller scale like: what is the average temperature for a summer/winter, day/night?]

Data Statistics: Student Worksheet

Match each statistical measure with its definition by drawing a line between them:

Minimum	The middle number in the dataset
Maximum	The largest value in the dataset
Mean	The lowest value in the dataset
Median	The average value of the dataset
Mode	The measures of the spread/variability of the data
Standard deviation	The value that is repeated the most times in the dataset

Download hourly data for a TAHMO weather station. Using Excel functions, fill out the table below.

	Relative Humidity (%)	**Precipitation (mm/hr)**	**Pressure (kPa)**	**Solar Radiation ($W/m^2/hr$)**	**Temperature (^oC)**	**Wind Speed (m/s)**
min						
max						
mean						
median						
mode						
standard deviation						

Which parameter had the most variability? What is the timescale of the variability (hourly, daily, monthly, yearly, or other)?

_____

_____

What are other ways might we use to describe the data that we didn't already use?

_____

_____

Why is it useful to mathematically describe our datasets?

_____

_____

Sketch a histogram of one of the parameters that you looked at. Be sure to create a title for your graph and label you axis. Label features on your graph including minimum, maximum, and mode.

Why is it useful to graphically describe our dataset?

_____

_____

Identify and describe one data statistic that is easier to identify in the table than in the histogram.

_____

_____

_____

Identify and describe one data statistic that is easier to identify in the histogram than in the table.

_____

_____

_____

We looked at the entire dataset for our analysis, but we could have only looked at a subset of the data. What would be one useful reason to only look at a subset of the data?

_____

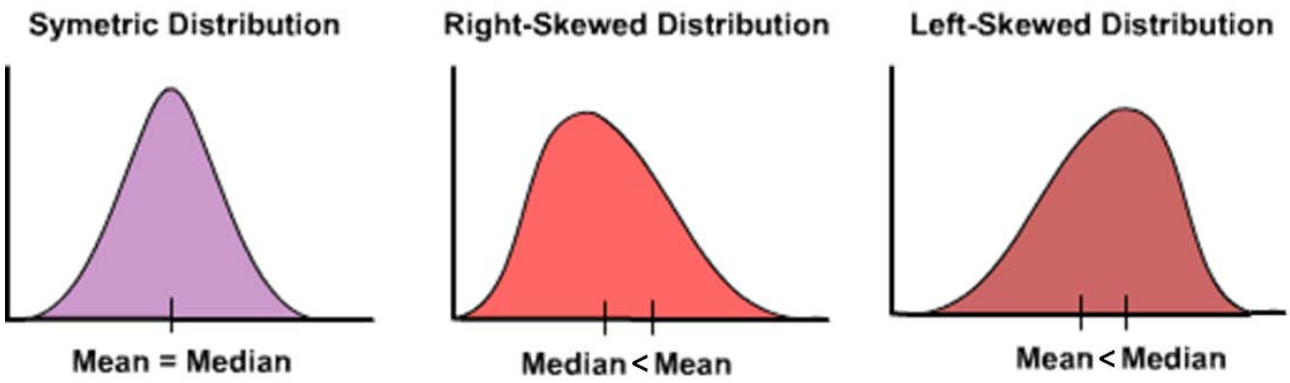
_____

_____

Advanced Topics: Skewness of Data

This advanced topic is intended for classes that want to continue to explore statistical descriptions of datasets. This advanced topic is option and should be determined based on the skill level of the classroom. This activity is designed to be done after the main lesson plan is completed, and does use values obtained earlier. Allow an additional 20 minutes for this activity.

How can you use the mean and median to predict the shape of the distribution of the dataset? If the mean is smaller than the median, the data is said to be skewed to the left. Skewness is a measure of the symmetry of a probability distribution. If the median is smaller than the mean, the data is said to be skewed to the right. If the mean and the median are equal, the data is said to be symmetrical.



Based on the mean and the median, describe the shape of the distribution for each variable. Creation of a histogram plot (value is the independent variable, frequency is the dependent variable) to verify your predictions.

Variable	**Mean (=,<,>) Median**	**Predicted Skewness**	**Excel's Skew (+,-)**	**Plot**
Relative Humidity				
Precipitation				
Pressure				
Solar Radiation				
Temperature				
Wind Speed				

Example Answer Key: Alliance Girls School TA00066

The plots and tables represented in this example answer key are only valid for the TAHMO station TA0006. The data used was downloaded on 2 May 2017, and can be accessed here: https://drive.google.com/file/d/0BxNcyc_sE4pPUzlOdVRVTl9uYnc/view?usp=sharing. Analysis for other stations may not display the same trends, so this example answer key should be used as a guide.

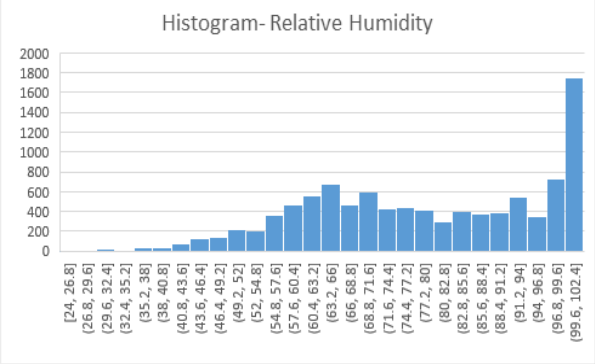
Answer Key: Data Statistics

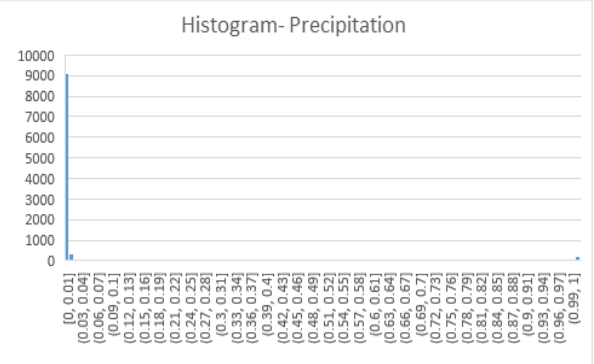
	Relative Humidity (%)	**Precipitation (mm/hr)**	**Pressure (kPa)**	**Solar Radiation ($W/m^2/hr$)**	**Temperature (^oC)**	**Wind Speed (m/s)**
min	24	0	781	0	5.61	0
max	100	35.5	809.3	1031	29.4	2.5
mean	78.18	0.12	804.99	132.50	17.91	0.51
median	79	0	805	0	17.73	0.49
mode	100	0	804.5	0	19.4	0
std	17.65	1.13	1.52	243.96	3.80	0.39

Answer Key: Advanced Topics and Skewness

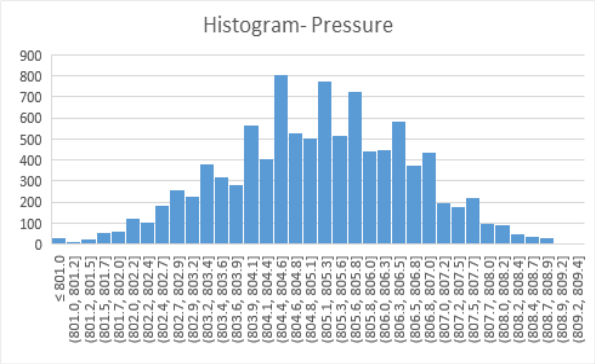
Variable	**Mean (=,<,>) Median**	**Predicted Skewness**	**Excel's Skew (+,-)**	**Plot**
Relative Humidity	<	left	- 0.31	left/-
Precipitation	>	right	+ 17.21	right/+
Pressure	=	symmetrical	- 0.45	left/-
Solar Radiation	>	right	+ 1.83	right/+
Temperature	=	symmetrical	+ 0.18	symmetrical
Wind Speed	>	right	+ 0.55	right/+

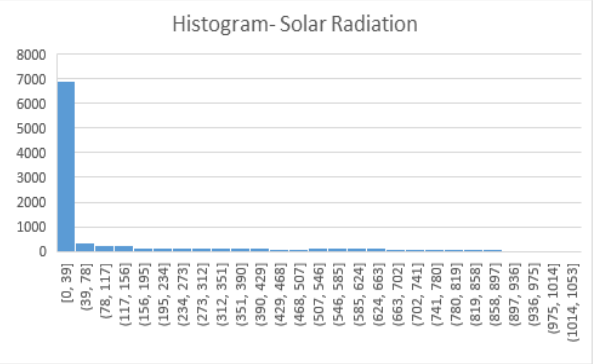
Predictions of skewness based on the mean and median are given as left, right, or symmetrical. When talking about skewness, we describe the direction of the tail: If a probability function is described as skewed right then you can imagine that the hump of the data is on the left and the tail end of the data is on the right. In Excel, the skewness function gives you a quantitative degree of skewness. A value greater than +1 means that there is a high degree of skewness to the right. A value less than -1 means that the data is very skewed to the left. A value between -0.25 and +0.25 is approximately symmetrical. A value between +0.25 and 1 is slightly skewed to the right. A value between -1 and -0.25 is slightly skewed left. Plotting a histogram (data values on the x-axis and frequency on the y-axis) is visually helpful to confirm the skewness trends.



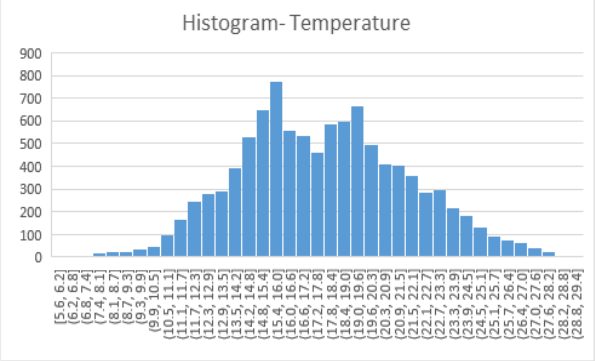


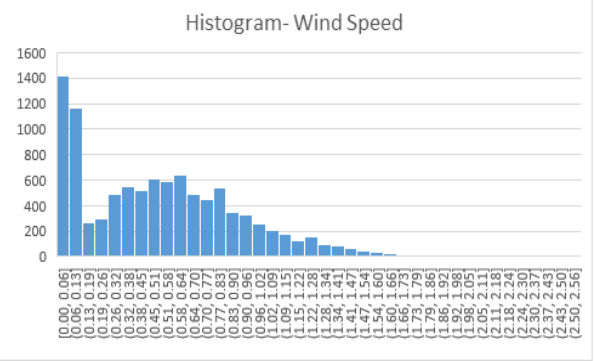
Relative Humidity looks to be skewed left, with the mode of the data on the right. Most precipitation values are 0 when it is not raining but there are values with heavy rain, thus the graph looks to be skewed right.





Pressure appears to be approximately symmetrical but it is hard to determine. Solar Radiation is strongly skewed right: half of the data is 0 and this represents the nighttime value.





Temperature looks to be approximately symmetrical, there are two peaks in this graph so we would call is a bimodal symmetrical distribution. Wind Speed is strongly skewed right, with a lot of data points at low wind speeds.