In all data obtained in the real world, there are bound to be some missing data or even data that do not fit in with the rest. What if a sensor was malfunctioning or inactive? What if some time series yielded radically different results than previous times? What if a mistake was made in data entry? Data that are extremely different from norms and averages, whether it be higher or lower, are known as *outliers*. I will be explaining how both outliers and missing data are dealt with -- that is, what we can do to remedy them.

Take, for example, a time series displaying annual temperature data since the Industrial Revolution began (shown below). We can see that during most years, the global temperature anomaly remained at or slightly below average temperature. However, after 1980, the temperature began rising sharply. If thermometers read a temperature much higher or much lower than the observed trend, those values would be considered outliers (e.g., the value for 1878 is much higher than its surrounding values). Similarly, if no temperature sensor could collect temperature data for that year, or if the data was never recorded, then missing data would arise.

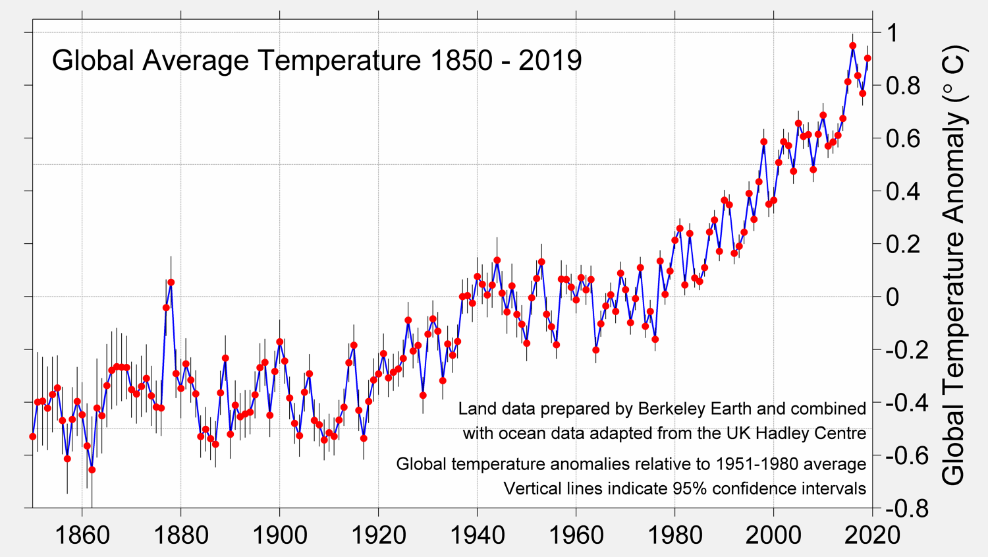


Image: <http://berkeleyearth.org/archive/2019-temperatures/>

Before dealing with outliers, we have to confirm a number is an outlier. Missing data is obviously to known to be missing, but how can we know if a present value is sufficiently different to constitute an outlier? We can use Z-scores, or the number of standard deviations from the mean, to find out. Usually, if a value is more than 3 standard deviations away from the mean, it is considered an outlier. This comes from the fact that 99.7% of data in a normal distribution is within 3 standard deviations of the mean.

Once a value is identified as missing or as an outlier, we must decide how to deal with it. There are three actions we can take with such values: we can remove them entirely, replace them with another value, or keep them. When would we choose to do each action? Here is a guide.

**When to Drop Outliers/Missing Values:**

1. If you know for a fact a point is way off or should not be missing, then you can safely delete it.

2. If you have lots of data that wouldn't be affected by dropping them, then you can delete the blanks and outliers. Because of this, I make it a point to collect as much time series data as I can at my job.

**When to Replace Outliers/Missing Values:**

1. If you can go back and find the correct data, then you can replace the points with that data.

2. If noise needs to be eliminated from the data, such as with moving averages, then the erroneous values can be replaced with the moving average.

**When to Keep Outliers/Missing Values:**

1. If you're dealing with critical data, such as structural stress testing, then it is important to keep all data as it is, even the vacancies.

2. If you are certain that the outlier or vacancy is not an error and is supposed to occur, such as with regular voltage spikes or regular power outages (as in some countries), then you can keep it.

Usually, missing data and outliers should be few and far in between, but if there are a lot of them where dropping them would cause radically different results to occur, that could potentially reveal something about the data or the collection method. In these cases, two analyses (with and without outliers/missing data) might be preferable in order to be as impartial in the analysis as possible.

References:

Sarkar, T. & Roychowdhury, S. (2019). *Data Wrangling with Python.*Packt Publishing.

Ferguson, K.  (2018). *When Should You Delete Outliers from a Dataset?* Humans of Data. Retrieved on January 8, 2021 from https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/.