June 15

- Use opendemand to run chapter_2_2021.ipynb
- Did it run without any errors?
- If not, send a message in Teams

Chapter 2

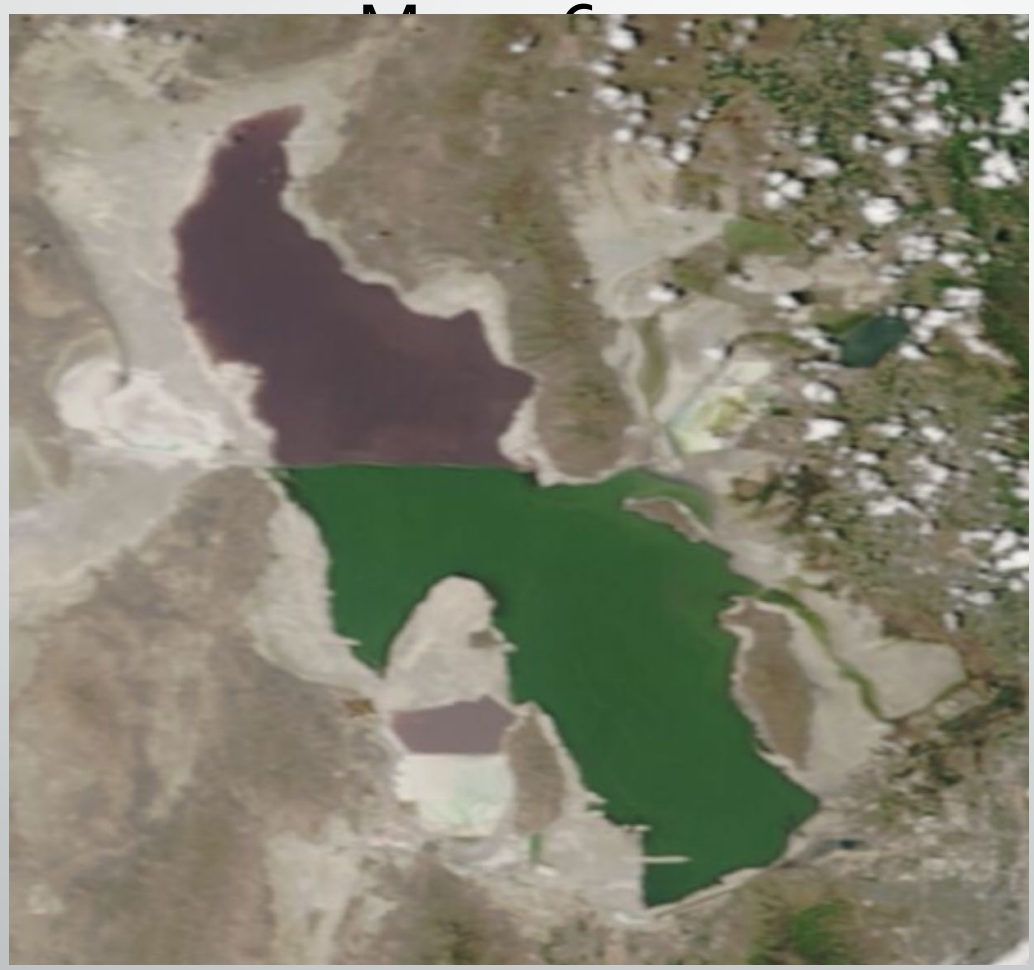
Exploratory Empirical Analyses

- Objective is to reduce the complexity (dimensionality) within a large data set
- What is a value commonly observed?
- How much variability is there among all the values?
- What are extreme cases that have been observed?

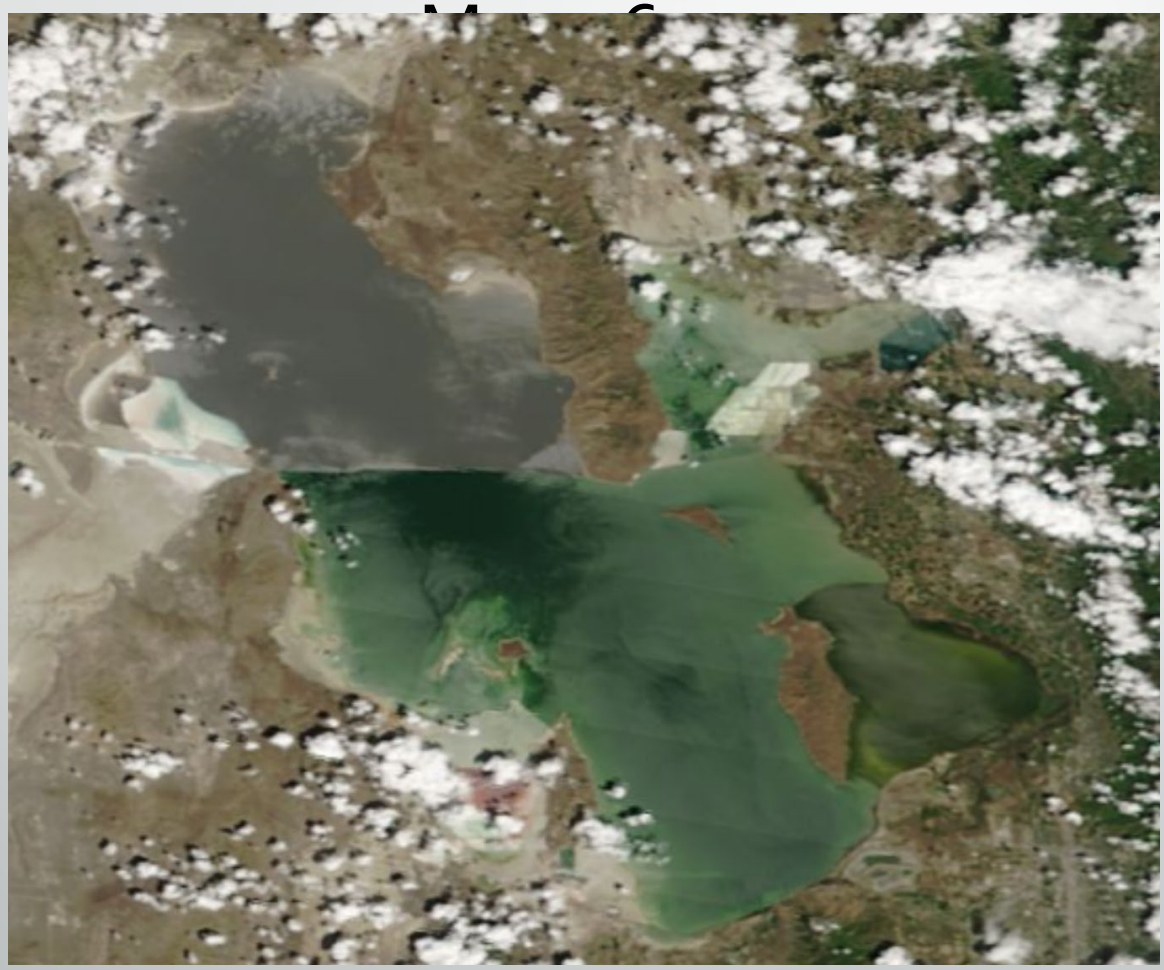
Exploring Data: What is the Objective?

- Summarizing some of the typical characteristics of the data
- How often are critical thresholds for specific applications reached?
  - Road temperature below freezing point
  - Hot, dry, windy conditions potentially leading to wildfires
- Approach to be used will depend on what is considered important to know to address the objective

NASA Worldview

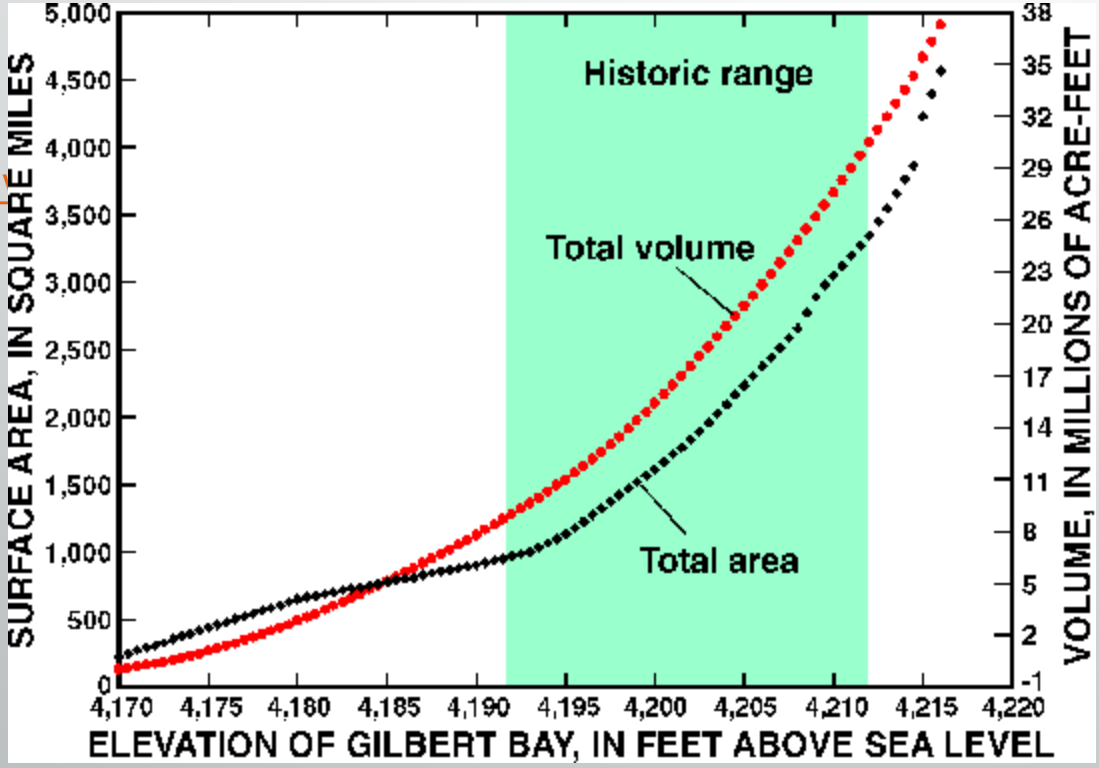


Great Salt Lake

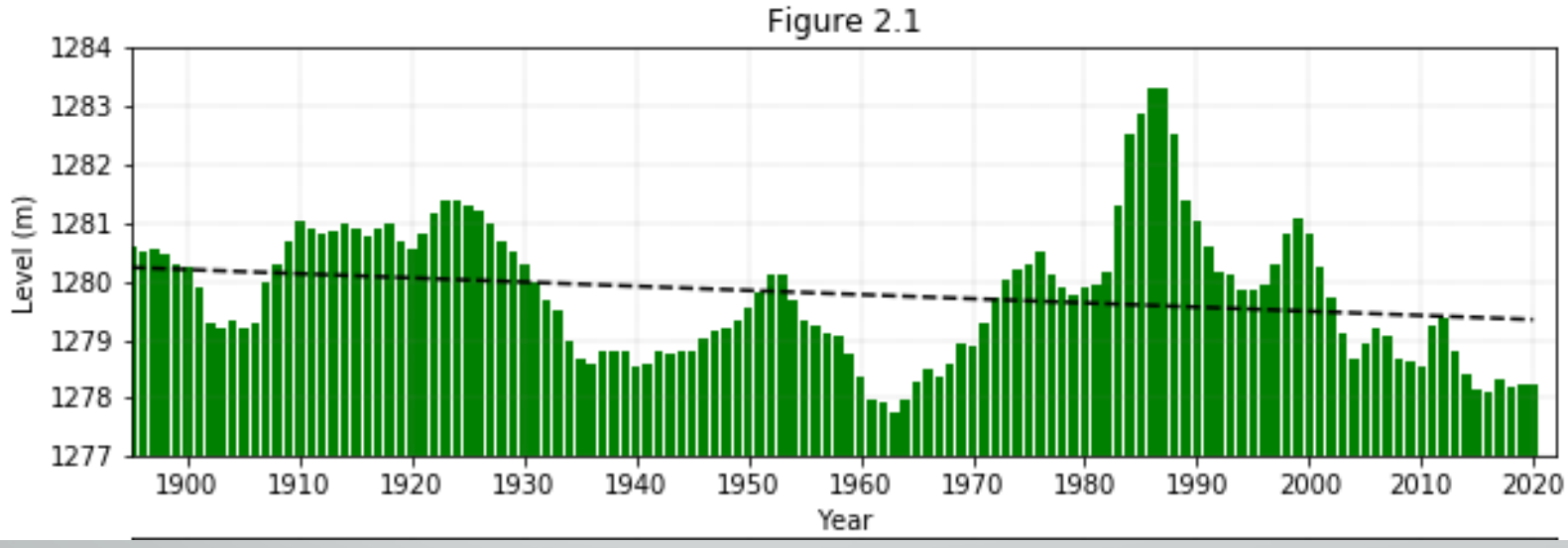


Great Salt Lake

- http://ut.v



Great Salt Lake level

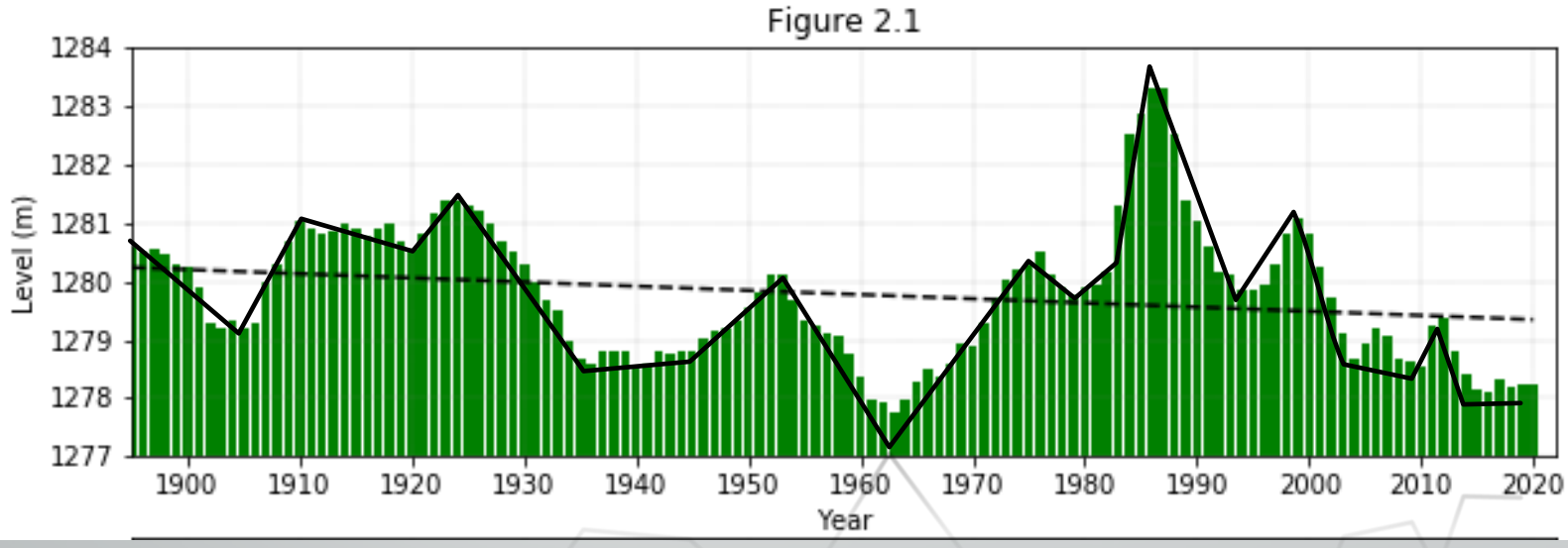


Great Salt Lake Level

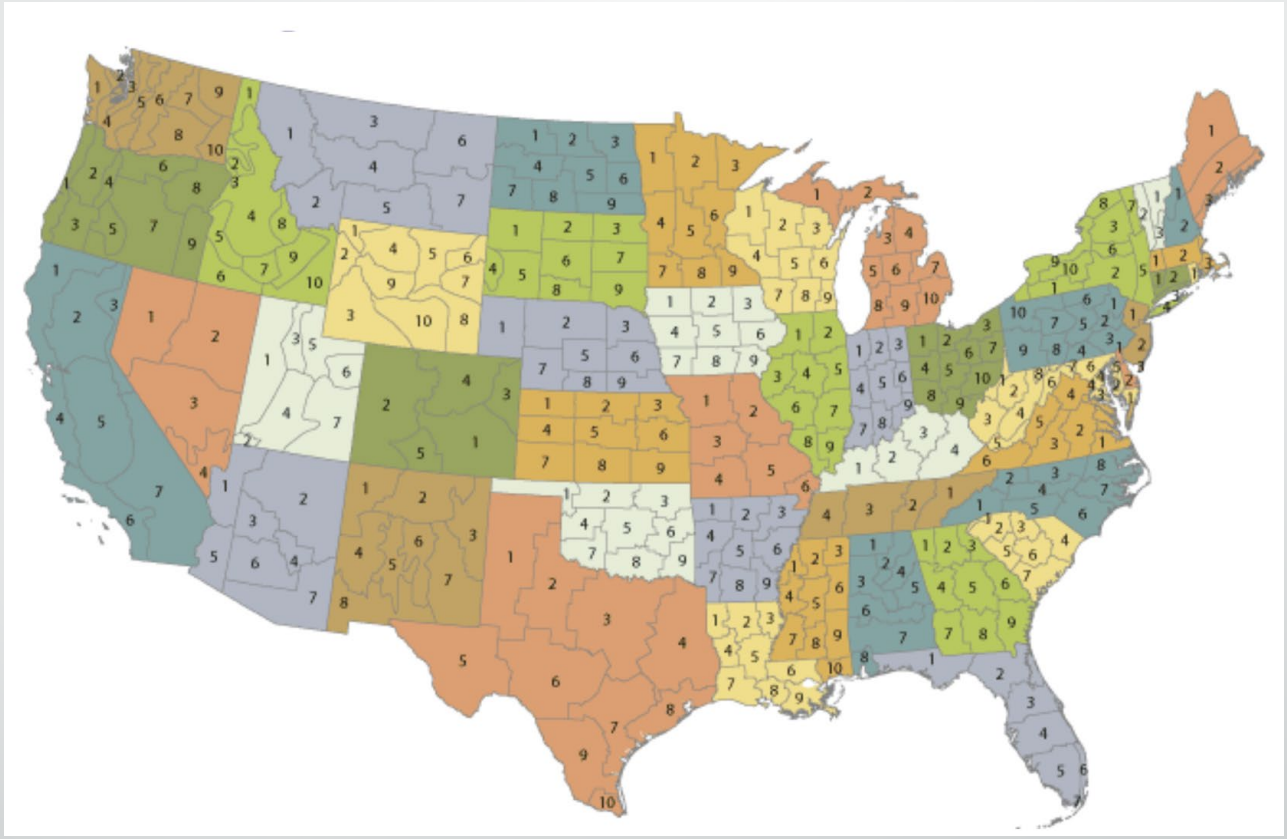
- Where did this data come from?
- What format?
- Understanding IO is the most critical step

- https://waterdata.usgs.gov/nwis/uv?site_no=10010000

Great Salt Lake level

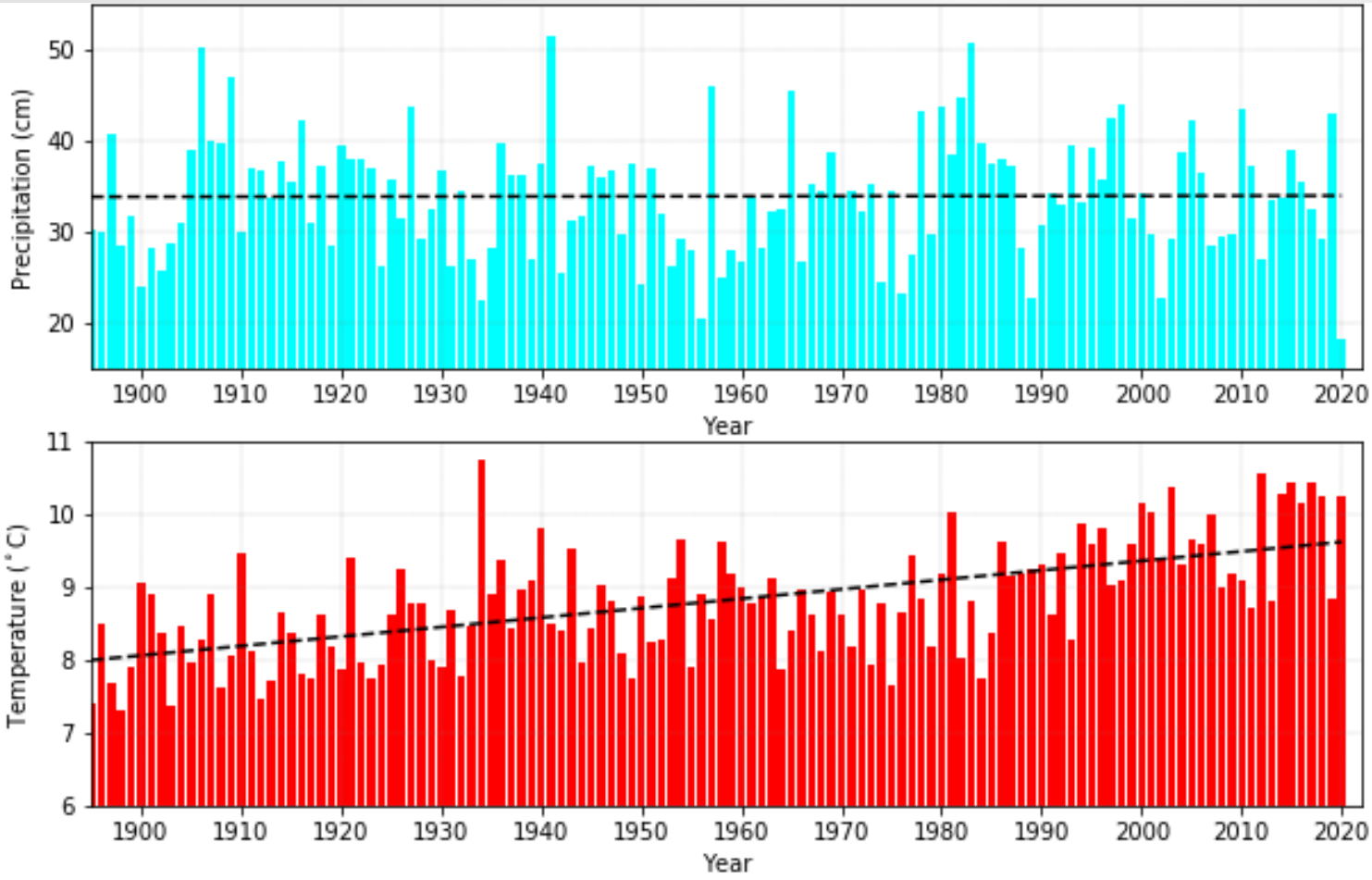


Climate Division Data



- https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-divisions.php

Utah Precipitation and Temperature



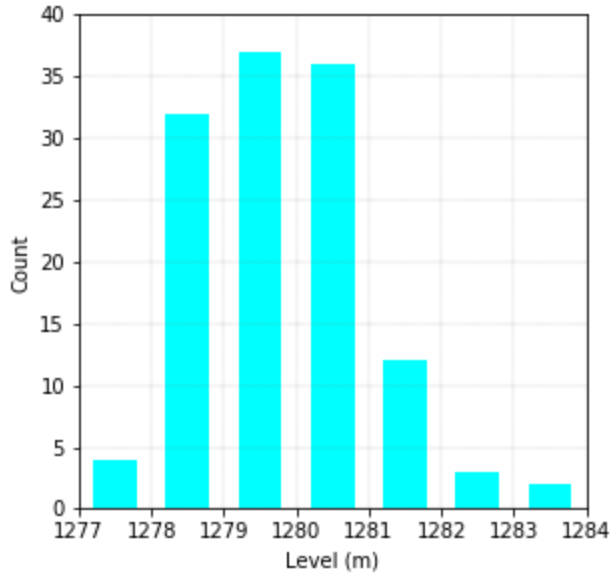
Utah Temperature and Precipitation

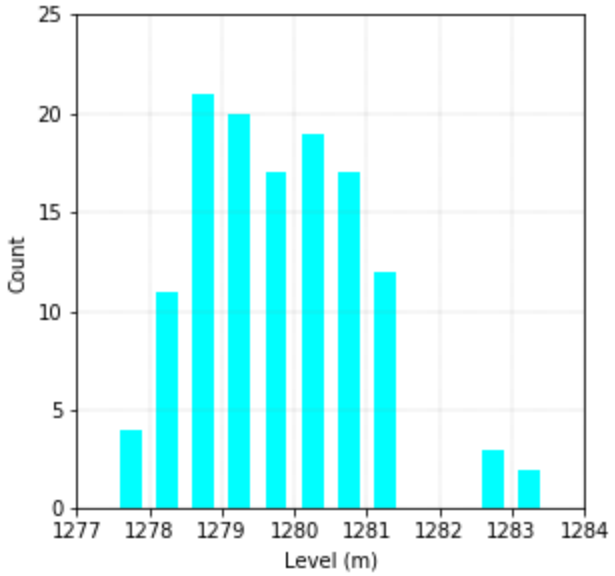
- Where did this data come from?
- What format?
- IO is the most critical step
- http://www.wrcc.dri.edu/cgi-bin/divplot1_form.pl?4203

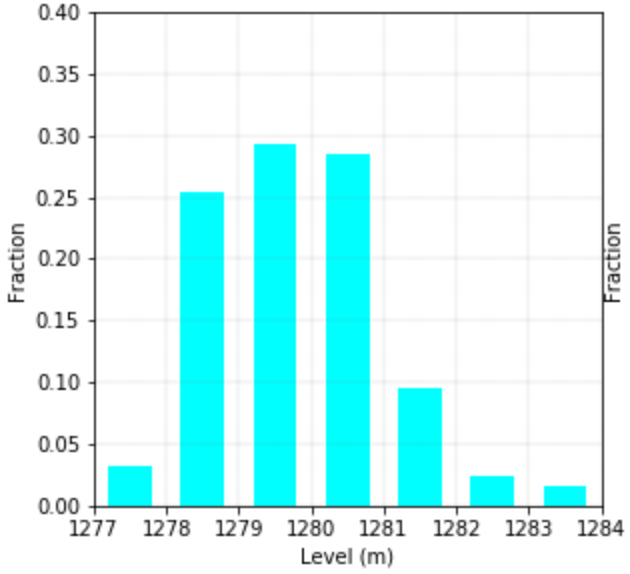
b. Data Distributions and Histograms

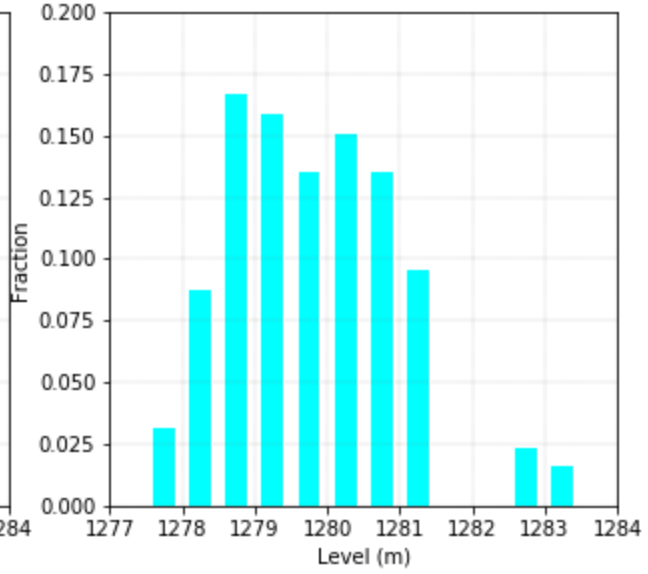
- Be careful with time series data- using histograms (frequency distributions) and cumulative frequency distributions loses the sense of time dependence of the underlying data
- Every observation in a time series is not likely independent of the others, yet that is a common misinterpretation of data distributions

Historgrams: Lake level

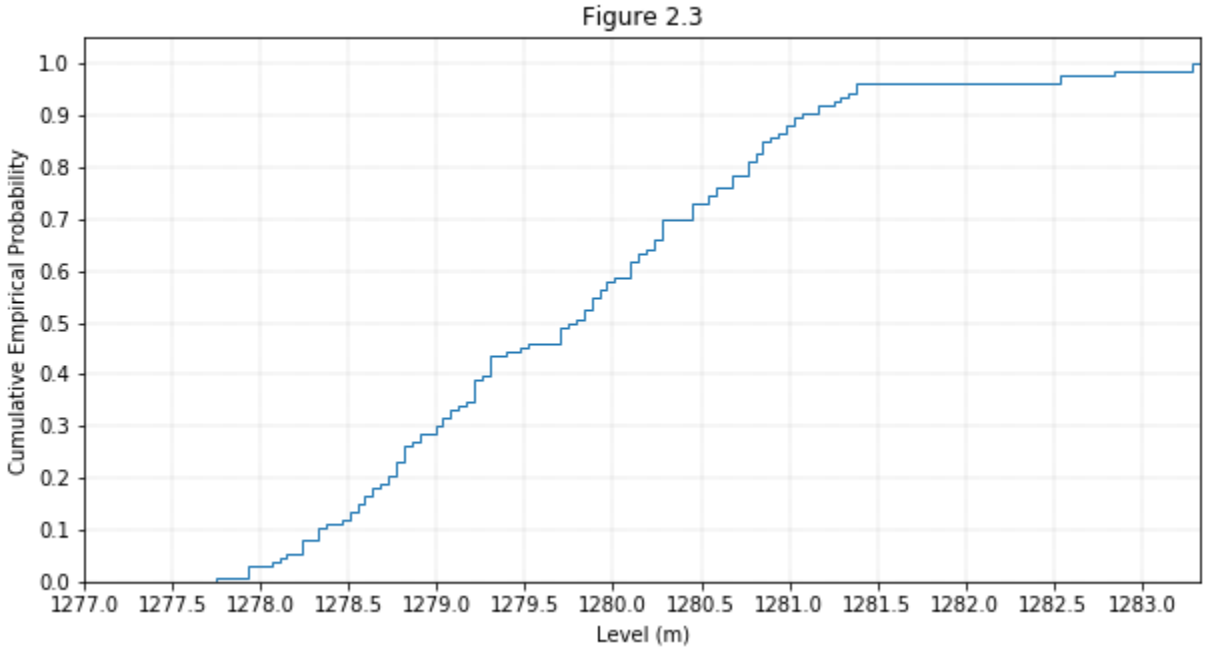








Empirical Cumulative Frequency Distribution Lake Level



c. Central Value, Spread, and Symmetry

- **central value** (central tendency or typical value)
- **spread** (variation or dispersion about the central or typical value)
- **symmetry** (degree to which the values tend to be larger or smaller than the central value)

What we want…

Metrics that are:

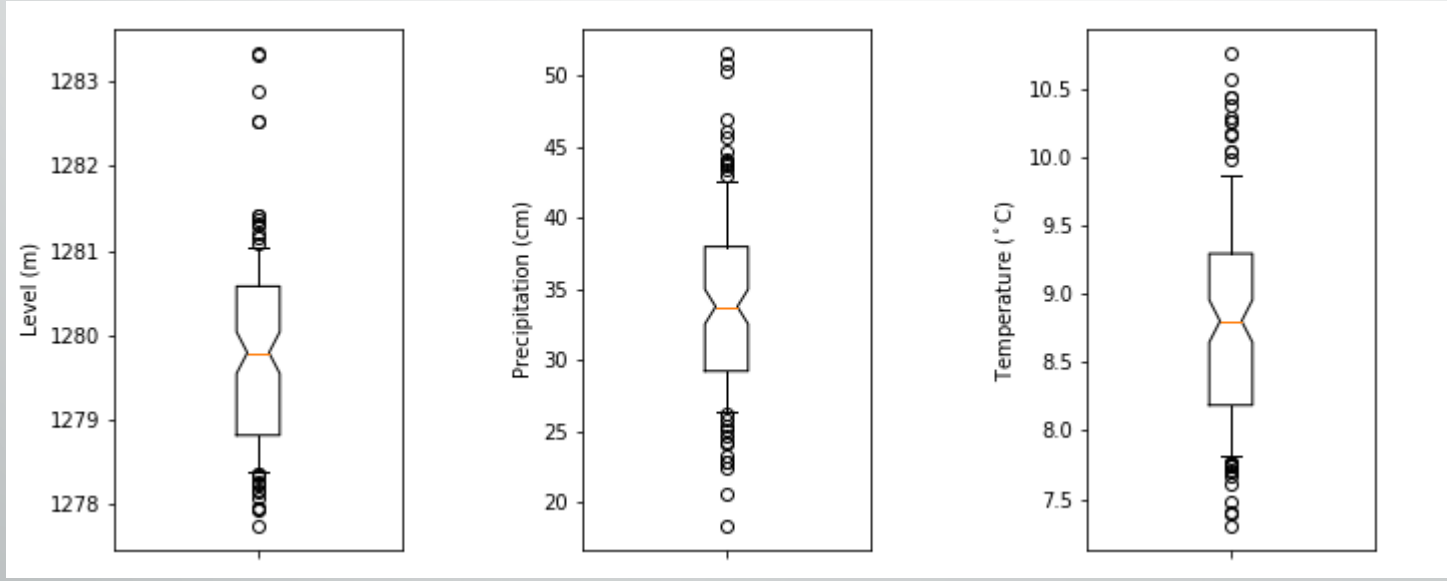
- ***robust***- which means not overly sensitive to the characteristics of the entire sample of data values. In other words, we want the measure to perform reasonably well no matter how the data values are distributed.
- ***resistant-*** not unduly influenced by outliers in the sample. For example, the range is not resistant to outliers.

Quantiles

- **Quantiles**: percentage thresholds of the data
- q_{25} – lower quartile- 25% of the sample lies below that value and 75% lies above
- q_{50} – median- 50% of the sample lies below that value and 50% lies above
- q_{75} – upper quartile- 75% of the sample lies below that value and 25% lies above

- **Median** is a very good measure of the central tendency of the data, i.e., the typical value.

Box and whiskers plots
Box: 75^{th}, 50^{th}, 25^{th} percentiles
Whiskers: 90^{th} and 10^{th} percentiles
Circles: outliers



Metrics

Central tendency

- Median
- Mode
- Mean
- Trimmed mean

Spread

- Range
- Interquartile Range
- Median absolute deviation
- Standard deviation
- Variance

Mode, Mean, Trimmed Mean

- Mode- most common values
- Mean- not robust or reliant- sensitive to outliers

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Trimmed mean- robust and reliant but throwing away data: k values on either end of the distribution

$$\overline{X_\alpha} = \frac{1}{n-2k}\sum_{i=k+n}^{n-k} x_i$$

Spread

- Range- very sensitive to outliers
- IQR- $75^{th} - 25^{th}$ percentiles (ignores 50% of the data)
- MAD = median $| x_i - q_{.5} |$- robust and reliant
- **standard deviation**, s, is the most commonly used measure of spread, but not resistant to outliers or robust

(2.c.3)

Standard Deviation/Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Standard deviation, s, is not resistant to outliers or robust

Variance, s^2, an **unbiased estimate of the population variance**

Why n-1 rather than n???

Estimating Population Variance

- Assume population has a mean of zero and variance σ^2
- Sample variance is s_x^2 (note: dividing by n)

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \overline{x^2}$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\sum_{i=1}^{n} (x_i \bar{x}) + \frac{1}{n}\sum_{i=1}^{n} (\bar{x})^2$$

$$\sum_{i=1}^{n} a = na \qquad \sum_{i=1}^{n} ax_i = na\bar{x}$$

$$s_x^2 = \overline{x^2} - \frac{2}{n}\bar{x}\sum_{i=1}^{n} (x_i) + \frac{1}{n}\bar{x}^2 n = \overline{x^2} - 2\overline{\bar{x}\bar{x}} + \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

For a large sample

$$\overline{x^2} = \frac{1}{n}\overline{x}^2 \quad from\ Central\ Limit\ theorem$$

$$s_x{}^2 = \overline{x^2} - \frac{1}{n}\overline{x^2} = \frac{n-1}{n}\overline{x^2} = \frac{n-1}{n}\sigma^2$$

$$s^2 = \frac{n}{n-1}s_x{}^2 = \sigma^2$$

- Much ado about nothing? Yes if n is large!!
- Sample variance: what we measure- use it
- Population estimate: use when assessing statistical significance

Symmetry

- Skewness- a measure of how the values are symmetric or asymmetric

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

What you should be doing

- Make sure you can run the chapter 2 python notebook
- Read all of Chapter 2 notes
- Begin Check Your Understanding #2