Missing Data Lab

## Find snowfall data

I need to know how much snow fell *per day* in New York State in February 2017, on a county or more detailed level. I know some government agency measures and reports snowfall and puts the data online. Find and download the data. *Keep detailed records of every step of the process so you can reproduce your process. Take screenshots if necessary.*

## Find snowfall data

<https://www.ncdc.noaa.gov/snow-and-ice/daily-snow/>

## Study the data

1. What variables do we have?
2. What is GHCN ID? What is the meaning of the characters?
3. What do non-numeric codes in the day columns mean?

## Study the data

1. What variables do we have?
2. What is GHCN ID? What is the meaning of the characters?

<https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt>

1. What do non-numeric codes in the day columns mean?

## Mark missing data as NA

* Indicate NA string when reading the data

1. Use a heatmap style function to look for patterns in missing data. What do you observe?

* mi::image(missing\_data.frame())
* extracat::visna()
* ggplot2::geom\_tile()

1. Does the number of missing values vary by county?
2. Does the number of missing values vary by day of the month?
3. Does the number of missing values vary by type of day (weekday vs weekend)?
4. Does the number of missing values vary by average snowfall?
5. What is your overall conclusion regarding data quality?