Data Quality

After this video you will be able to..

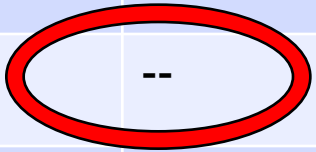
- Describe three data quality issues
- Name three reasons for poor data quality
- Explain why data quality issues need to be addressed

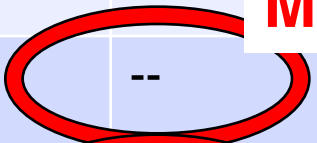
Data Quality Issues

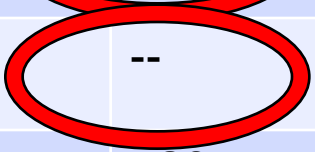
- Real-world data is messy!

Missing Data

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120







Missing Values

Duplicate Data

Name	Address
Angela	430 Park Drive
Sidney	7800 West View Street
Sid	7800 West View Street
Ratan	12443 Mountain Avenue
Kiril	45 East 5th St
Kiril	1220 Mill Avenue
Zhou	4345 Apple Lane

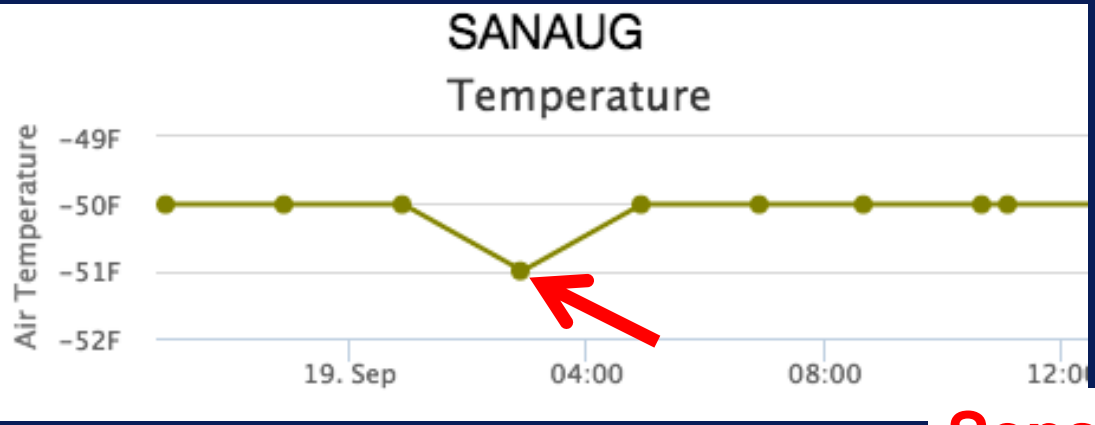
Invalid Data

Name	Zip Code
Angela	346412
Sidney	92618
Ratan	8033A
Kiril	11012
Zhou	59285

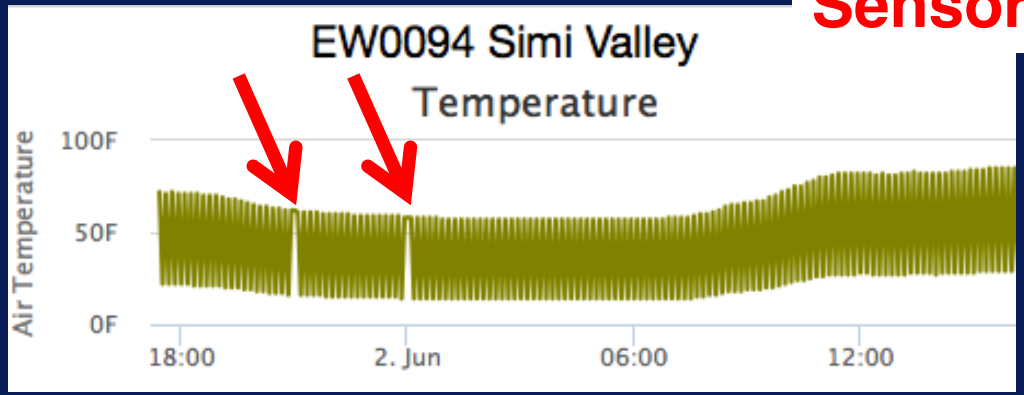
Noise

Name	Address
Angela	430 Park Drive
Sidney	780 ★❖©◆ ew Street
Ratan	12443 Mountain Avenue
Kiril	1220 Mill Avenue
ZhČou	4345 Apple Lane

Outliers



Sensor Failure



Why Address Data Quality Issues?

Poor Data Quality

Poor Analysis Results