Addressing
Data Quality Issues

After this video you will be able to..

- Define what ‘imputation’ means
- Illustrate three ways to handle missing values
- Describe the role of domain knowledge in addressing data quality issues

Data Quality Issues

Missing values

Duplicate data

Noise

Invalid data

Outliers

Removing Missing Data

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

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

Imputing Missing Data

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

- **Replace missing values with something reasonable**

Name	Age	Income
Angela	34	80
Sidney	***50***	***56***
Ratan	***10***	***50***
Kiril	***68***	***50***
Zhou	45	120

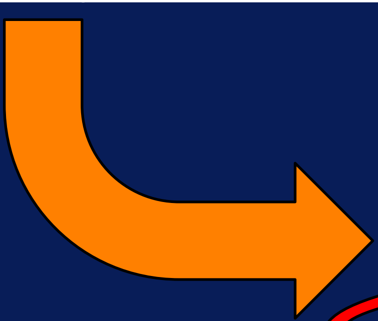
Ways to Impute Missing Data

- **Replace missing value with**
  - Mean
  - Median
  - Most frequent
  - Sensible value based on application

Duplicate Data

- **Delete older record.**
- **Merge duplicate records**

Name	Address
Sidney	7800 West View Street
Sid	7800 West View Street
Kiril	45 East 5th St
Kiril	1220 Mill Avenue



Name	Address
Sidney	7800 West View Street
~~***Sid***~~	~~***7800 West View Street***~~
~~***Kiril***~~	~~***45 East 5th St***~~
Kiril	1220 Mill Avenue

Invalid Data

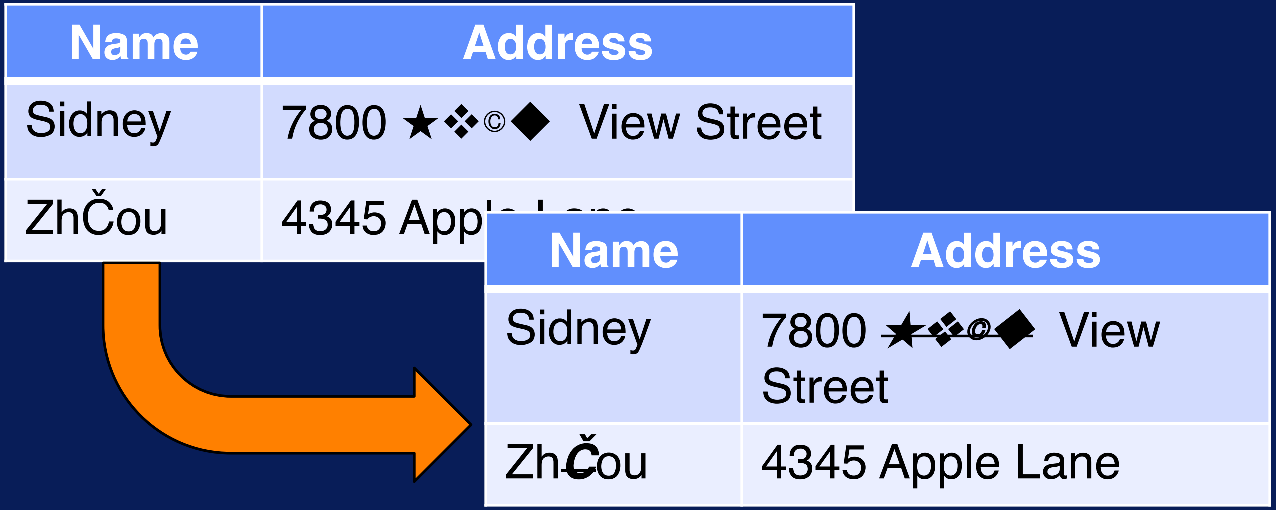
- Use external data source to get correct value
- Apply reasoning and domain knowledge to come up with reasonable value.

Name	Zip Code
Angela	346412
Ratan	8033A

Name	Zip Code
Angela	34641***2***
Ratan	8033***1***

Noise

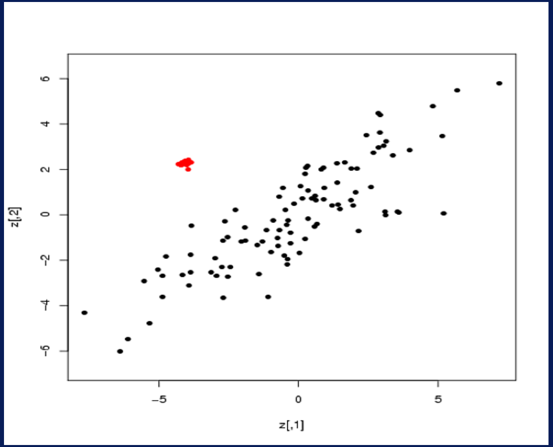
- Filter out noise component.
- May also filter out part of data, so care must be taken.



Name	Address
Sidney	7800 ★❖©◆ View Street
ZhČou	4345 App

Name	Address
Sidney	7800 ~~★❖©◆~~ View Street
Zh~~Č~~ou	4345 Apple Lane

Outliers



- Remove outliers if they’re not focus of analysis
- Analyze more closely if they are focus of analysis (e.g., fraud detection)

Domain Knowledge

- Required for addressing data quality issues effectively