Data Structures

Introduction to Data Visualization

The Graduate Center at CUNY | Summer 2019

Video Session

Data Cleaning Best Practices

- NEVER DELETE the originals
- Document how values were altered

Data Shapes

- Most data you will deal with will be rectangular (rows/columns, observations/attributes, records/categories, instances/variables).
- These usually come as Excel or CSV files
  - Excel is a proprietary format owned by Microsoft.
  - CSV are “comma separated values” and can be read in any program so long as the character encodings are readable
  - TSV are ”tab separated values” similar to CSVs

- Some is dictionary-like (json, avro, headless databases)
- Some is relational (many linked rectangles, i.e., SQL)

Inspection: Headers

- Headers
  - Data in the headers: columns that could be grouped into a larger category such as “day of the week”, “language”, etc.
    - Pivot the data
  - Codes in the headers: A000TF15
    - This is not a problem, but you may want to change these to make your table easier to work with.

Tragic Problems (& some solutions)

- No headers
  - Can categorize by location, but then you need to know location
- Character Encoding Errors
  - Trial and error decoding in potential formats (ASCII, big 5, Windows, etc.)
  - Always save in Unicode!
- Incorrect Category Labels
  - Manual manipulation
  - Re-align labels
- Disorganized data (different data types in one variable)
  - Manual manipulation

Problems (& some solutions)

- Wrong character encodings
  - Change the encoding
- Dates as strings
  - Lubridate
- Stray characters
  - Strip, replace, remove
- Data in the column names
  - Pivot/Gather
- Inconsistent formatting
  - Transformations

Tidy Data

- All observations are rows
- All variables are columns
- Only one type of observation unit per table

First Name	Child	Age
Wanda	Maria	12
Wanda	Lacey	15
Floyd	Marcus	5
Clyde	Luella	7

First Name	Age	Education	Occupation
Wanda	42	Chemistry	Chemist
Floyd	51	Physics	Scientist
Clyde	28	Mathematics	Engineer
Clyde	28	Mathematics	Data Analyst

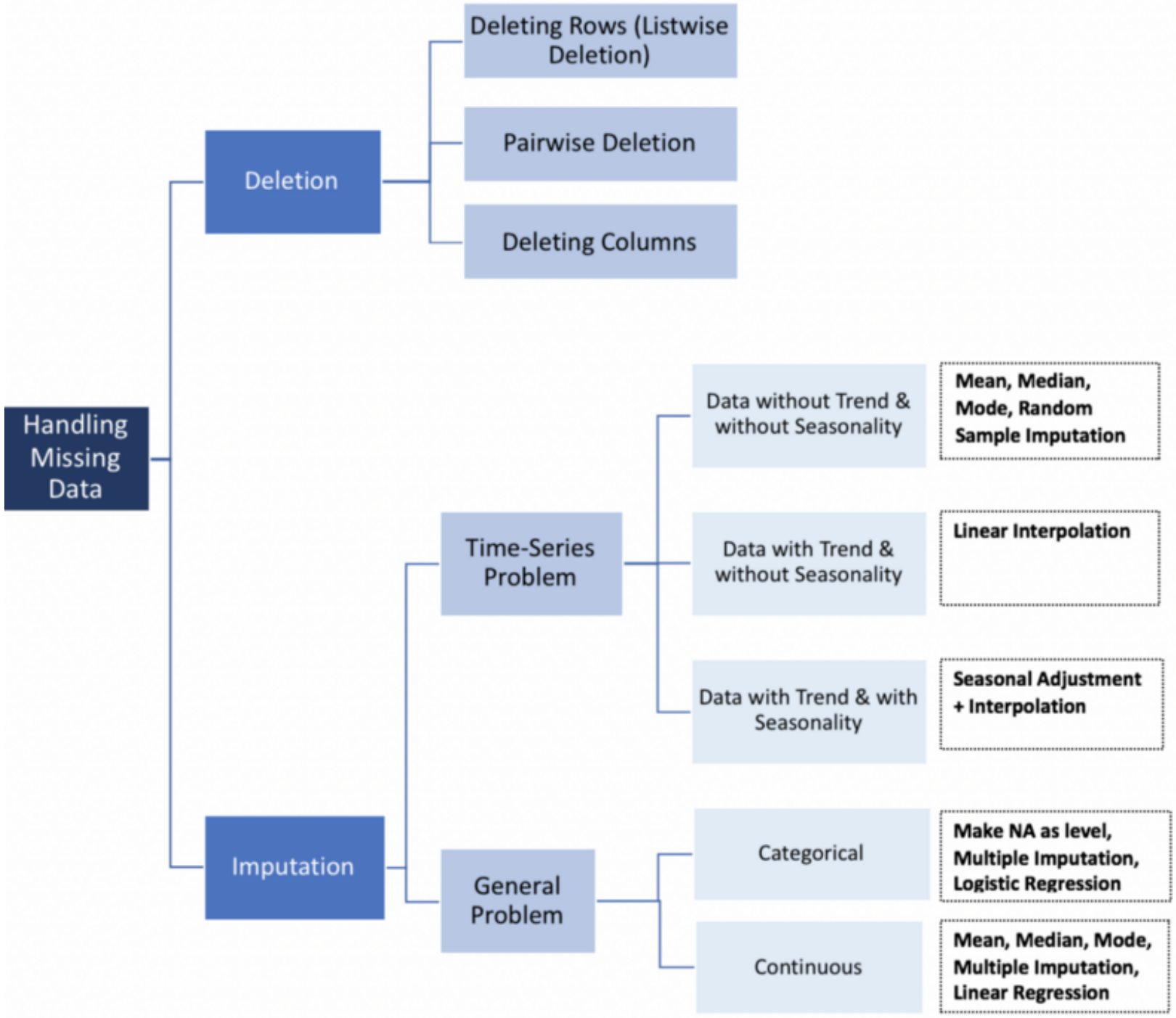
First Name	Age	Education	Occupation	Child	Age
Wanda	42	Chemistry	Chemist	Maria	12
Wanda	42	Chemistry	Chemist	Lacey	15
Floyd	51	Physics	Scientist	Marcus	5
Clyde	28	Mathematics	Engineer	Luella	7

Missing Data

- Missing not at Random
  - The missing data tells a story
  - Missing data is a data point
  - Imputation depends on context
- Missing at Random
  - Due to the measurement or collection system
- Missing Completely at Random
  - Simple human error
  - Messiness

Missing Data

- Remove
- Fill with
  - Mean
  - Median
  - Maximum
  - Minimum
  - 0
  - 1



https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4