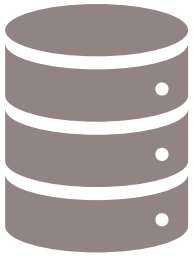
Data Collection and Cleaning

Agenda



Data Collection

How do we collect data?

What are key considerations when collecting?

Data Cleaning

Goals of Cleaning Data

Why is it important?

Cleaning SS Data Sources

Examples of Rogue/Dirty Data

- Duplicated rows or columns
- Inaccurate data
- Incorrectly formatted
- Inconsistent
  - Same data exists in different formats
- Missing data
  - Incomplete observations
  - Blank columns

Goals of Data Cleaning

Maintain	Remove	Keep
Maintain high data quality • Valid • Accurate • Complete • Consistent • Uniform • Relevant	Outliers Irrelevant observations Dupes	Keep as much of the data intact as possible

Considerations

General

- Become familiar with dataset
- Outline the ultimate goal of the analysis
- Figure out which columns are most important

Analytics

- Outdated data
- Data Sparsity
- Skewed Data
- What columns contain the most information (variance)?

About **25-30%** of data is **inaccurate**

Analysts spend **60-80% of their time** cleaning data

Why is it important?

- GIGO
  - Impact the results of your analysis
  - Flawed results = incorrect actions taken
- Ability to combine data sources

Analyses are only as good as the data they are based on

Cleaning Sport Science Data

1. Become familiar with the data that you are working with
    1. Is there anything about the structure that needs to be changed?
2. Look for values that are very unlikely to be true and any rows that may be duplicated
3. Remove duplicates and decide how to handle outliers
    1. Is just one of the columns messed up for the observation or is the entire row inaccurate?
    2. Can require looking at basic statistics such as min, max, mean, median, range, etc
4. Determine the most important columns
    1. Ask yourself what columns contain the information most related to the goal of your analysis
5. Remove columns that don't contain relevant information
6. Add columns with additional information if needed for analysis
7. Update formatting of columns if needed
8. Combine multiple files or data sources
    1. Look for unique identifiers that exist in every data source

Sources

- https://careerfoundry.com/en/blog/data-analytics/what-is-data-cleaning/
- https://www.alteryx.com/glossary/data-cleansing
- https://www.sigmacomputing.com/resources/learn/what-is-data-cleaning#:~:text=The%20main%20objective%20of%20data,of%20refining%20and%20optimizing%20datasets

Catapult Appendix

R Studio

- Set working directory: setwd(“*Folder path*")
- Get file names: csv_files <- **list.files**(pattern = "\\.csv$")
- Apply cleaning to multiple files: **lapply**(csv_files, function(file) { read.csv(file, skip = 9) })
  - Applies function to list or vector
- Find columns containing specific word: columns_with_target <- **grepl**(“Load", **colnames**(*dataset_name*))
  - print(colnames(*dataset_name*)[columns_with_target])
- Get summary stats of columns: **summary**(*df*)
- Clean text strings in column names: **gsub**(“*what to replace”, “replace with”,* colnames(*df))*
- Find columns with all 0’s: zero_columns <- sapply(*df*, **function**(x) **all**(x == 0))
  - Get names of cols: **names**(zero_columns[zero_columns])

Python

- Folder_path = “*Folder path*”
  - Os.chdir(Folder_path)
- Pattern = “*.csv”
  - Csv_files = glob.glob(pattern)
- [**pd.read_csv**(file, skiprows = 9) for file in Csv_files]
- [col for col in *dataset_name*.**columns** if ‘Load’ in col]
- *dataset_name*.describe()
- *dataset_name*.columns.**str.replace**(“*what to replace”, “replace with”)*
- zero_columns = *df*.columns[(*df* == 0).**all**()]