ROLL UP YOUR SLEEVES FOR SOME DATA CLEANING!

Whether you have collected your own data or will be reusing existing datasets, you probably need to clean them up before you move forward with data analysis. This process includes fixing or removing incorrect, corrupted, unformatted, duplicate, or incomplete data. While the cleaning-up process may look different depending on the dataset you have at hand, this handout covers some essential tips to complete this task more efficiently while making your data more consistent, accurate, and high quality.

PLAN & ORGANIZE YOUR "CHORE"

MAKE A COPY OF THE DATA

Start working on a separate copy and keep a backup and untouched copy of the original data.

CHOOSE YOUR TOOLS

Select an open source tool to help you automate the process and optimize your time (see some options below).

KNOW THE DATA

Explore and inspect the dataset and documentation to get a sense of its structure, types and contents.

DOCUMENT YOUR WORK

Keep a record of all changes and transformations performed for transparency and reproducibility.

DETECT & REMOVE THE DIRT

DUPLICATED ROWS

Exclude duplicate cases/records from the dataset to avoid bias and a skewed analysis.

EXTRA SPACES

Eliminate unnecessary leading, trailing, or multiple embedded space characters or nonprinting characters.

IRRELEVANT COLUMNS

Inspect and delete any columns that are not important for the project or that should not be included in the analysis.

TYPOS & ODD CHARACTERS

Look for any typos or odd characters wrongly added during data entry, exporting or as a result of encoding.

SCRUB OFF & TOUCH UP

INCONSISTENCIES

Reconcile any inconsistencies in the data (e.g., different date formats, capitalization, out of range values, uncontrolled vocabularies).

NULL VALUES

Watch out for values like, blanks,"0", "Not Applicable", "NA", "None", "Null" and make sure they are uniformly encoded.

STANDARDS

Check if units of measurements conform to desired standards and transformed accordingly (e.g., temperature, distance, weight).

RELATIONSHIPS

If working with relational data, ensure the dataset does not contain any "dangling" foreign keys linking to nonexistent tables or columns.

RECOMMENDED TOOLS

OpenRefine

A tool that does not require programming skills and supports effortless cleaning, transformation, and error identification in large datasets.

pandas

A powerful data processing and manipulation library for the Python programming language to deal with messy data.

tidyverse

A collection of R packages designed to help researchers "tidying" and preparing the structure of datasets to facilitate analysis.

Want to learn more? rds@library.ucsb.edu

UC **SANTA BARBARA**
Library

www.library.ucsb.edu