DATA CLEANING

Data quality

Validity

- *Data-Type Constraints (numeric vs character)*
- *Range Constraints (numbers, dates within range)*
- *Mandatory Constraints (no missing ids)*
- *Unique Constraints (unique combinations)*
- *Set-Membership constraints (predefined values for gender etc)*
- *Foreign-key constraints (foreign key may not include something that’s not in primary key)*
- *Regular expression patterns (ssn, phone number, zipcode rules)*
- *Cross-field validation (may not depart before arriving)*

Accuracy

- Validity does not mean accuracy!
- The entered zipcode, phone number, email may be in a valid format but not accurate

Completeness – missing data?

Consistency

- Inconsistency occurs when two values in the data set contradict each other.
- Age is 10 and the person is married; or zipcode is entered as 78249 but the person’s address is in California

Uniformity

- Convert data into a single measure unit.
- Pounds or kilos; miles or kilometers; be clear: days, months, years

The workflow

The workflow is a sequence of three steps aiming at producing high-quality data and taking into account all the criteria we've talked about.

1. Inspection: Detect unexpected, incorrect, and inconsistent data.
    - Data profiling: Utilize summary statistics, describe data, what are the data characteristics, etc.
    - Do visualization: tables, figures, graphs etc. Summarize data.
2. Cleaning: Fix or remove the anomalies discovered.
    - Clean irrelevant data
    - Remove duplicates
    - Make necessary type conversions
    - Correct syntax errors: extra white space, typos
    - Standardize: uppercase, lowercase, same measure unit
    - Scale if necessary
    - Drop missing values if you have enough remaining data
    - Take into account the outliers, make a note if you remove anything and why you did it.
3. Verifying: After cleaning, the results are inspected to verify correctness.
4. Reporting: A report about the changes made and the quality of the currently stored data is recorded.

References

Elgabry, O. (2019, March 2). The Ultimate Guide to data cleaning. Medium. Retrieved October 19, 2021, from https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.