Data Preprocessing

Data Cleaning

- Data Cleaning is the process of ensuring that your data is correct, consistent and useable
- Old and inaccurate data can have an impact on results
- Data Cleaning consists of,
  - *identifying any errors or corruptions in the data,*
  - *correcting or deleting them,*
  - *manually processing them as needed to prevent the error from happening again.*

Benefits of Data Cleaning

- It removes major errors and inconsistencies that are inevitable when multiple sources of data are getting pulled into one dataset.
- Using tools to cleanup data will make everyone more efficient since they'll be able to quickly get what they need from the data.
- The ability to map the different functions and what your data is intended to do and where it is coming from your data.

Data Quality

- Validity
- Accuracy
- Completeness
- Consistency
- Uniformity

Validity (1/2)

The degree to which the data conform to defined business rules or constraints.

- Data-Type Constraints
  - Particular column must of a particular datatype e.g numeric
- Range Constraints
  - Numbers or Dates should fall within a certain range
- Mandatory Constraints
  - Certain columns cannot be empty.

Validity (2/2)

- Unique Constraints
  - A field/(s) must be unique across a dataset
- Set-Membership Constraints
  - Values of a column come from a set of discrete values e.g. Male /Female from enum values
- Regular expression patterns
  - Text fields that have to be in a certain pattern e.g (999) 999–9999
- Cross-field Validation
  - Date birth of new students cannot be above current date

Accuracy

- The degree to which the data is close to the true values.
- While defining all possible valid values allows invalid values to be easily spotted, it does not mean that they are accurate.
- Outliers

Completeness

- The degree to which all required data is known.

- Missing data is going to happen for various reasons. One can mitigate this problem by questioning the original source if possible, say re-interviewing the subject.

Consistency

- The degree to which the data is consistent, within the same data set or across multiple data sets.
- Inconsistency occurs when two values in the data set contradict each other.
  - *E.g A customer is recorded in two different tables with two different addresses.*

Uniformity

- The degree to which the data is specified using the same unit of measure.

- If not, data must be converted to a single measure unit.

Workflow

Steps

- Inspection
  - Detect unexpected, incorrect, and inconsistent data.
- Cleaning
  - Fix or remove the anomalies discovered.
- Verifying
  - After cleaning, the results are inspected to verify correctness.
- Reporting
  - A report about the changes made and the quality of the currently stored data is recorded.