Data Wrangling

What is Data Wrangling?

- Data wrangling, also known as **Data Munging**, is an iterative process that involves data exploration, transformation, validation and making it available for a credible and meaningful analysis.
- It is a 4-step process that involves, -
  - (i) **Discovery-** about understanding our data better with respect to our use case. The objective is to figure out specifically how best we can clean, structure, organize and map our data for our use case.
  - (ii) **Transformation-** involves the tasks we undertake to transform the data such as,
    - **(a) Structuring-** includes actions that can change the form of our data.
    - **(b) Normalization-** focuses on cleaning the database of unused data and reducing redundancy and inconsistency.
    - **(c) Denormalization-** used to combine data from multiple tables into a single table so that it can be queried faster.
    - **(d) Cleaning-** actions that fix irregularities in data in order to produce a credible and accurate analysis.
    - **(e) Enriching-** to look at additional data points that could make analysis more meaningful.
  - (iii) **Validation-** checking the quality of data after transformation. It refers to repetitive programming steps to verify consistency, quality and security of the data we have.
  - (iv) **Publication-** involves delivering the output of the wrangled data for downstream project needs.

Tools for Data Wrangling

- Spreadsheets (Microsoft Power Query for Excel)
- Python
- R

Data Cleaning

- Data cleaning workflow includes, -

(i) **Inspection-** detect the different types of issues and errors that dataset may have. We can use scripts and tools that allow to define specific rules and constraints and validate data against these rules and constraints. We can also use data profiling and data visualization tools for inspection. Data profiling helps to inspect the source data to understand the structure, content, and interrelationships in data. It uncovers anomalies and data quality issues.

(ii) **Cleaning-** dealing with missing values, removal of duplicate data, irrelevant data.

(iii) **Verification-** inspect the results to establish effectiveness and accuracy achieved as a result of the data cleaning operation. We need to re-inspect the data to make sure the rules and constraints applicable on the data still hold after the corrections we made.

At the end, it is important to note that all changes undertaken as part of the data cleaning operation need to be documented. Not just the changes, but also the reasons behind making those changes, and the quality of the currently stored data. Reporting how healthy the data is, is a very crucial step.