Data Transformations

Agenda

- ETL
- Raw data
- Data Lake, Databases, Data Warehouse etc.

Raw Data

- Raw data, also known as primary data, are data (e.g., numbers, instrument readings, figures, etc.) collected from a source.
- Attributes of raw data are:
  - may possibly contain human, machine, or instrument errors,
  - may not be validated;
  - might be in different area (colloquial) formats; uncoded or unformatted;
  - some entries might be "suspect" (e.g., outliers), requiring confirmation or citation.
- An example from Kaggle
- Also mention the survey data

ETL: Extract-Transform-Load

- ETL is the name of overall process for organizations to combine data from multiple sources into a unified, single data store.
- In ETL, we clean, integrate and organize data.
- The goal in ETL is to store data in such a way that data is ready for the use of more advanced analytics.

Where can we extract data?

- SLQ and NoSql Servers
- Flat files
- Emails
- Webpages

Transform

- Cleaning data
- De-duplicating
- Validating
- Authenticating
- Filtering
- Putting data in a consistent format
- Removing, encrypting data

Load

- Loading is the process of inserting that formatted data into the target database, data store, data warehouse, or data lake.

Challenges of modern ETL

- Large volume of data
- Speed of data
- Diversity of data

To address these needs ETL tools and softwares are changing constantly.