What is data transform?

- is a process of converting data into a format that aids in building efficient ML models and deriving better insights.
- is an important step in the feature engineering and data processing stage of a Data Science project.
- is a technique used to transform raw data into a more appropriate format that enables efficient data mining and model building.
- Data Transformation process is also called an ETL process.

When to Transform Data?

- before Data Mining so that it can help in extracting meaningful patterns and insights.
- before training and developing an ML model.

Benefits of Data Transformation

1. **Maximize Value of Data:**
    - standardizes data from various data sources to increase its usability and accessibility.
        - This will ensure that maximum data is used in Data Mining and model building, resulting in extracting maximum value from data.
2. **Effective Data Management:**
    - remove inconsistencies in the data by applying various techniques so that it is easier to understand and retrieve data.
3. **Better Model Building and Insights:**
    - remove bias in the model by standardizing and normalizing features in the same range for highly skewed dataset.
4. **Improve Data Quality:**
    - improve data quality by handling missing values and other inconsistencies.

Challenges of Data Transformation

1. is an expensive and resource-intensive process. This cost depends upon many factors such as infrastructure, tools, company requirements, data size, etc.

2. requires professionals with appropriate subject matter expertise as faulty Data Transformation can lead to inaccurate business insights.

How does Data Transformation Works?

The entire process of Data Transformation is called ETL (Extract, Transform, and Load).

- ETL tools automate the entire transformation process and can enable efficient monitoring and management of Data Transformation.
- ETL tools can be `on-premise` (hosted on company servers) or `cloud-based` (hosted in the cloud).
- Most popular transformation techniques used in Data Science are
- ETL process can be defined as a six-step process as shown below :

1. **Data Discovery:** In this stage, Data Scientists and Analysts identify data sources that are relevant for required for further analysis. They also review its format and schema.
2. **Data Mapping:** During this phase, Data Scientists and Analysts determine how individual attributes across data sources are mapped, modified, and aggregated.
3. **Data Extraction:** ted from its primary source in this step. It could be in this step SQL database or data from the Internet using Web Scraping methods.
4. **Code Generation and Execution:** In this step, Data Scientists and Analysts prepare the code scripts for transformation and execute them.
5. **Review:** After the execution of code in the previous step, in this step, the output is reviewed to validate whether the transformation was accurate or not.
6. **Sending:** Once transformed data is reviewed and validated, it is sent for storage to the destination source, such as a database, data warehouse, data lake, etc.

Data Transformation Techniques

1. **Data Smoothing**
    - is used to remove noise in the data,
    - helps inherent patterns to stand out.
    - help in predicting trends or future events.
    - tools to perform Data Smoothing are `moving average`, `exponential average`, `random walk`, `regression`, `binning`, `clustering`, etc.

2. **Attribute Construction**
    - new attributes or features are created out of the existing features.
    - simplifies the data and makes data mining more efficient.
3. **Data Generalization**
    - process of transforming low-level attributes into high-level ones by using the concept of hierarchy.
    - applied to categorical data where they have a finite but large number of distinct values.
4. **Data Aggregation**
    - is the process of compiling large volumes of data and transforming it into an organized and summarized format that is more consumable and comprehensive.
    - enable the capability to forecast future trends and aid in predictive analysis.
5. **Data Normalization**
    - process of scaling the data to a much smaller range, without losing information to help minimize or exclude duplicated data and improve algorithm efficiency and data extraction performance.
    - The range of values for each attribute in a dataset can vary greatly. This introduces a bias in the model building. Therefore it is essential to normalize every feature in the dataset.
    - is a technique that is used to convert a numeric variable into a specified range such as [-1,1], [0,1], etc.
    - A few of the most common approaches to performing normalization include :

    1. *Min-Max Normalization*:
        - This is a linear transformation and will convert the data into the [0,1] range.
        - transforms features into a fixed range to ensure all values lie between the given min and max values.
        - doesn't work well with features having a lot of outliers.
    2. *Z-Score Normalization*:
        - utilizes the mean and standard deviation of the attribute to normalize it.
        - It will ensure that the attribute has a 0 mean and 1 standard deviation.
        - Z-Score Normalization is also called `Data Standardization` or `Data Scaling.`
        - most used transformation technique
        - transform all distributions into a normal distribution
    3. *Decimal Scaling:*

    - It normalizes the values of an attribute by changing the position of their decimal points
    - The number of points by which the decimal point is moved can be determined by the absolute maximum value of that attribute.
    - Normalized value of attribute $v'_i = \frac{v_i}{10^j}$, where j is the smallest integer such that $\max(|v'_i|) < 1$

6. **Data Discretization**
    - process of converting numerical or continuous variables into a set of intervals.
    - makes data easy to analyze and understand.
    - For example, the age features can be converted into intervals such as (0-10, 11-20, ..) or (child, young, ...).
7. **Log Transformation**
    - transform features with a heavily skewed distribution into a normal distribution.
8. **Reciprocal/Inverse Transformation**
    - an attribute x is replaced by its inverse i.e., $1/x$.
    - applied only to attributes having non-zero values.
9. **Square Transformation**
    - an attribute x is replaced by x^2.
    - applied to any feature having numeric values.
10. **Square Root Transformation**
    - x is replaced by its square root.
    - applied to features having only positive values.
    - has a moderate effect on input distribution.
11. **Box-Cox Transformation**
    - values in attribute x are replaced based on the formula as mentioned below :

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

    - converts non-normal data to approximate-normal distribution.

Data Integration

- it is not a data transformation technique but rather a critical step during the pre-processing phase.

- is the process of bringing together information from different sources to create a unified view of the data.
- The destinations for the data integrated from sources include `spreadsheets` (Google Sheets, Excel), `data warehouses` (BigQuery, Amazon Redshift, PostgreSQL), and even `BI tools` (Looker Studio, Power BI, Tableau, Qlik).

Data Blending

- is the process of combining data from multiple sources.
- also referred to as `data join`.
- In data integration, you can only extract data from `one source` or database, whereas data blending, extracts from `multiple sources`.

Data Manipulation

- process of making data more readable and organized.
- can be achieved by changing or altering raw datasets.