EDA-Data Transformation

Data transformation is the process of converting raw data into a suitable format for analysis. It involves modifying or cleaning up data to make it more meaningful and interpretable. The goal of data transformation is to ensure that data is consistent, accurate, and relevant for further processing.

Types of Data Transformation:

1. **Missing Value Imputation**: Replacing missing values with estimated values based on statistical methods.
2. **Data Normalization**: Scaling numerical data to a common range (e.g., 0-1) to prevent features with large ranges from dominating the analysis.
3. **Data Standardization**: Similar to normalization, but also considers the mean and standard deviation of each feature.
4. **Categorical Encoding**: Converting categorical variables into numerical representations (e.g., one-hot encoding).
5. **Time Series Transformation**: Transforming time series data into a suitable format for analysis (e.g., seasonal decomposition).

Example: Data Preprocessing with Pandas in Python

Suppose we have a dataset of student exam scores, which includes missing values and categorical variables.

Student ID	Name	Exam Score	City
1	John	85.0	New York
2	Emma	NaN	Los Angeles
3	Max	90.5	Chicago
...	...	...	...

```
 import pandas as pd

# Load the dataset
df = pd.read_csv('student_scores.csv')

# Print the first few rows of the dataset
print(df.head())

# Impute missing values with mean score (data transformation)
df['Exam Score'] = df['Exam Score'].fillna(df['Exam Score'].mean())

# One-hot encode categorical variable 'City' (data transformation)
df = pd.get_dummies(df, columns=['City'])

# Print the first few rows of the transformed dataset
print(df.head())
```

In this example, we use Pandas to load and manipulate a dataset. We impute missing values in the 'Exam Score' column with the mean score and apply one-hot encoding to the categorical variable 'City'. The resulting transformed dataset is then printed.

Data transformation is an essential step in data preprocessing, as it ensures that data is clean, consistent, and suitable for analysis.

Curated by Brajesh Kumar