Basic # Data cleaning
Data Preparation.

DOMS | Page No.
Date / /

Q Some important concepts to keep in mind while doing Data cleaning and Preparation.

① Handling Missing Data

Why?
- To maintain data integrity
- To avoid Bias and distortion
- To maintain data completeness

How?
- Completely filter out rows or values using pandas.
- Imputation i.e, filling missing values either by mean median or by any other value.
  - either 1. domain specific imputation
    2. Time series imputation. or
    3. statistical imputation.

② Data Transformation.

Why?
- Irregular scales of data and different range makes analysis difficult.
- Duplicacy of data which will be of no use
- To handle data skewness
- For Further processing and analysis DT is must.

Basic # Data cleaning
Data preparation

2015 | Page No.
Date / /

- For make data in form which is can be used for further computation like handling categorical data.
- To Extract out new and better features of requred
- To make data in form suitable for other libraries like numpy.

How?

1. Removing Duplicates:

   Pandas has inbuilt tools to remove duplicates

2. Transforming Data using Function mapping or dictionary mapping.

   using dictionary or functions to map with data by using map() function and manipulation existing columns or creating new one.

3. Replacing values:

   using replace function to replace existing values by null or any other value as required.

4. using 2. we can also rename indec and manipulate them as per requirement without creating new data structure.

   we can also use rename() feul in pandas.

Basic # Data cleaning
Data preparation.

DOMS | Page No.
Date / /

5. Discretization and Binning:

Sometimes it is hard to analyse continuous data so it is discretized into "bins" for analysis.

You can convert

continuous data ⟶ categorical data

by creating bins. (use pandas) ✓

6. Detecting and Filtering outliers:

- you can use matplotlib to detect outliers and anomaly and then use pandas to filter out the outliers
- We can use various statistical tests and methods for detecting outliers
- use imputation or deletion for filtering or flag them.
- For detection there are various methods available but used according to problem statement and requirement

1. Visualization.
2. Z-score
3. IQR method
4. Mahalanbis distance
5. Isolation forest
6. Local outlier factor (LOF)

...

Basic
Data Preparation.

7. Random sampling & recording for avoiding bias and noise learning.

✓ 8. Converting categorical variables or data into column form. where each column represents a categorical variable. This is required for modelling and consider as a important step.

NOTE: For string manipulation we can use regix by importing re library. i.e. regular expression.

(3.) More on Categorical Data.

categorical data into numerical representation.

Why?

- Algorithm compatibility:
  many ML algo. are compatible into numerical inputs.
- Preventing misinterpretation of data and handling multible categories.

How?

- There are various techniques to do so
  - using dummy variables or indexing
  - one hot encoding
  - label encoding
  - target encoding and various other.