APSSDC

Andhra Pradesh State Skill Development Corporation

Data Cleaning using Pandas

Content

- Working with Duplicates and Missing Values
- Which values should be replace with missing values based on data
- Identifying and Eliminating Outliers
- Dropping duplicate data
- Filling missing data

Working Missing Values

- Missing Data can occur when no information is provided for one or more items or for a whole unit.
- In Pandas missing data is represented by two value:
  - None: None is a Python singleton object
  - NaN : NaN (an acronym for Not a Number),
- There are useful functions for finding missing values in Pandas DataFrame :
  - isnull()
  - notnull()
  - Dropping missing values by using following methods
    - dropna(),dropna(how='all'),dropna(axis=0),dropna(axis=1)

Working with duplicates in DataFrame

- An important part of Data analysis is analyzing *Duplicate Values* and removing them.
- Pandas **duplicated()** method helps in analyzing duplicate values only. It returns a boolean series which is True only for Unique elements.
- the duplicated() method returns False for Duplicates, the NOT of the series is taken to see unique value in Data Frame.
    - Step1 : Returning a boolean series
    - Step2: Removing duplicates

Replacing data

- Pandas dataframe.replace() function is used to replace a string, list, dictionary, series, number etc. from a dataframe
- Values of the DataFrame are replaced with other values dynamically
  - df.replace(to_replace=value, value)
  - df.replace(to_place=[value1,value2,value3],value)
  - df.replace(to_place=np.nan,value)

Filling missing data

- If only a few of the values are missing, we can perform data imputation to substitute the missing data with some other value(s).
- There are many different methods for to replace missing values
  - Using the mean,median value and the most frequent value
  - Using Filling in missing values with a constant by **fillna(constant)**
  - Filling null values with the previous ones by
    - df[‘column_name’].fillna(method = "pad")
  - Filling null values with the before ones by
    - df[‘column_name’].fillna(method = "bfill")

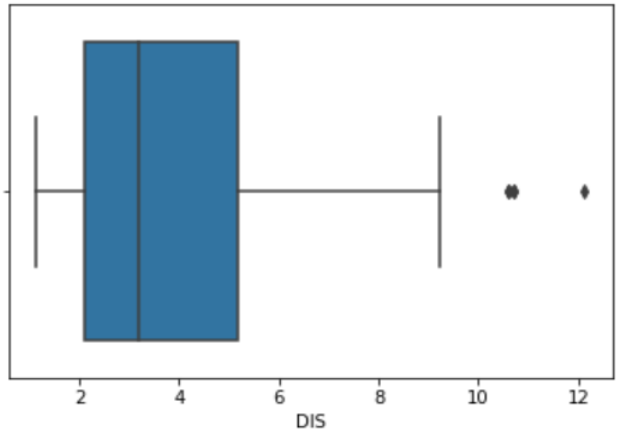
Dropping duplicate data

An important part of Data analysis is analyzing Duplicate Values and removing them

- drops all rows having null values using df.dropna() method
- drop_duplicates() - Return DataFrame with duplicate rows removed.
- df.drop_duplicates(subset = "column_name")
- df.drop_duplicates(subset = ["column_name1",’column_name2’])

Identifying and Eliminating Outliers

- outliers are observations that are significantly different from other data points
- outliers can adversely affect the training process of a machine learning algorithm, resulting in a loss of accuracy.
- need to use the mathematical formula and retrieve the outlier data.
- **interquartile range(IQR) =**
  Q3(quantile(0.75)) – Q1(quantile(0.25))



Skill AP
APSSDC

pandas