(06/08/2020) (Pre-Processing Data)

Missing Values⇒

missing value occur when no data value is stored for a variable (feature) in an observations.

Reperesen'ted as → ?, N/A, "0".

Dealing with missing Values ⇒ Check with data collection Source :⇒

* Drop the missing values:-
  - drop the variable
  - drop the data entry

* Replace the missing value -
  - replace it with an average (of similar datapoin
  - replace it by frequency (if data is categorical)

* Leave it as missing data.

How to drop missing values in Python ⇒

use →'

dataframes.dropna( ):

axis=0 drop the entire row

axis=1 drop the entire cloumn

Data Formatting ⇒

non formatted:
- Confusing
- hard to aggregate
- hard to compare

formatted:
- more clear
- easy to aggregate
- easy to compare

Correcting Data types ⇒

- To identify data types:—
  (1) dataframe.dtypes( )
- To convert data types:—
  dataframe.astype( )

for e.g df["Price"] = df[Price].astype("int")

Data Normalization ⇒

age	income
20	100000
30	20000
40	500000

⇒

age	income
0·2	0·2
0·3	0·04
0·4	1

method of Normalization data ⇒

(1.) Simple feature Scaling

$$\boxed{X_{new} = \frac{X_{old}}{X_{max}}}$$

(ii) Min-Max ⇒

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$

(iii) Z-Score ⇒

$$x_{new} = \frac{x_{old} - \mu}{\sigma}$$

Binning In Python ⇒

- Binning ⇒ Grouping of values into "bins"
- Convert numeric into categorical variables
- Group a set of numerical values into a set of "bins",

Turning categorical variable into Quantertive Variable in Python ⇒

One-hot-encoding ⇒

Using Pandas Library ⇒

(I) Pd.get dummies ( ) :
convert categorical variable to dummy variable (0 or 1).

e.g - Pd.get-dummies(df('fuel')).