Understanding your Data

How can we properly analyze our data in beggining

We basically ask few questions related to our data.

Question 1: How big is the Data?

⇒ using df.shape. Our before import we should know how big is our data

Question 2: How does the data look like?

⇒ Df.head(). we need to see the overview of the data.

OR ⇒ better for Overview of Data.

df.sample() ⇒ gets a random rows, so that we need to find any bias in the data.

Question 3: What type of Columns do I have?

⇒ Df.info() can be used to understand what type of data do we passes.

Imp Question ⇒ Question 4: Do we have missing values?

⇒ Dealing with missing values, is great headache.

df.isnull().sum() ⇒ gives number of missing values as sum for each columns.

So this helps us decide which column to remove or which column do we have put some values in,

Question 5: How does the data look mathematically?

Df.describe() ⇒ help us get the math value like mean, min, max, std, 25% value, count for each column,

Question 6: Are there Duplicate values?

⇒ This causes most trouble for our algo.

df.duplicated.sum() ⇒ get number of duplicated rows.

Question 7: How is the correlation btw cols?

which input column has no impact on output, so we generally remove those cols.

df.corr() ⇒ gives correlation btw columns.

This is pearson correlation of each column with every other column.