**Introduction to data processing**

The second segment is “Research data processing”.

In research, We need data to work with but we can’t work with raw data, let it be a human or computer or a code snippet. In order to get accurate results, the data must be filtered to get the desired viable data and to make a meaning out of it. This process of making the datas viable is called data processing.

If our model does not give expected output, We recollect data and go through all the steps again and there’s still no guarantee of the new collected data giving the expected outputs, hence, there’s no end to data processing.

Two popular libraries are mostly used in python to process data. Those are,

1. Numpy
2. Pandas

**Numpy:** *Numpy* is basically used to work with or manipulate large multidimensional array and matrix data. A large collection of mathematical functional operations like linear algebra routines, or fourier transforms can be done on these types of data using the *Numpy* library.

Other operations such as array creation or indexing or slicing these arrays/matrices or other numpy operations such as searching or sorting in an array can be done using the *numpy* library.

**Pandas:** *Pandas* is basically used for manipulating data and doing data analysis. It inherits some data structures & operations that are able to manipulate numerical tables and time series.

Pandas.

Tasks like Data cleaning, Data transformation, Future election, Encoding categorical data, Correlation, Imbalance data handling, Outlier detection and all sorts of other tasks can be done using *Pandas*.

Pandas library’s objects are heavily dependent on *Numpy* library’s objects. This enables us to convert data to dataframe and vice versa.

The third segment is “Research Data visualization”.

In research, We sometimes need to see the data or need to visualize it. That segment will be covered here.

Two popular libraries are mostly used in python to visualize data. Those are,

1. Matplotlib
2. Seaborn

**Matplotlib:** *Matplotlib* is basically used to visualize data in multiple ways in all sorts of representation.

**Seaborn:** *Seaborn* is also used to visualize data but it enhances the representation of *Matplotlib*. Basically, It works like an extension of the *Matplotlib* library.

The fourth segment is “Research Data Analysis”.   
Here, we will work with all the datas we’ve cleaned so far. It has 2 modules. Those are,

1. Model development
2. Evaluation

**Model development:** Here we develop a model and feed the data we far cleared up.

**Evaluation:** According to the given output as result, We will evaluate it how accurately is it working.

**Data:** Data is basically a collection of information. There could be various forms of data.

1. Numerical
2. Word
3. Sound
4. Image

**Dataset:** Dataset is basically a collection of data.

Here, We will talk about the second segment “Research Data Processing”.

Data processing is done within a cycle known as the data processing cycle.

**Data processing cycle:** The data processing cycle is the set of operations used to transform data into useful information. The intent of this processing is to create actionable information that can be used to enhance a business.

There are multiple types of data processing which are assorted by the source of the data and steps taken to process it by the processing unit.



