Unit - 5
VISUALIZING DATA: BAR CHARTS, LINE CHARTS, SCATTERPLOTS

```
pip install pandas
```

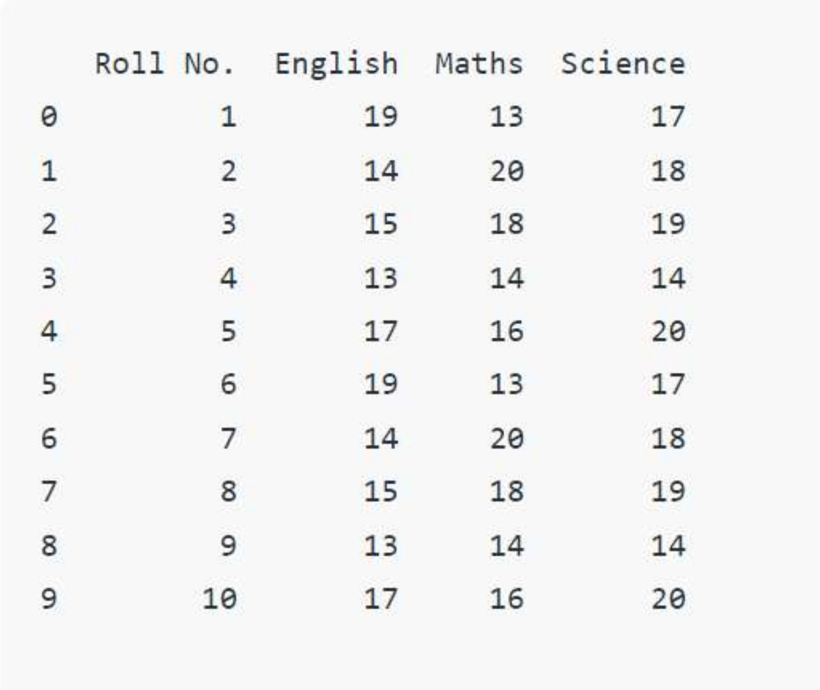
Working with data

- Reading Excel File using Pandas in Python
    - Installing and Importing Pandas
    - Reading multiple Excel sheets using Pandas
    - Application of different Pandas functions
    - **Installating Pandas**
    - To install Pandas in Python, we can use the following command in the command prompt:

```
pip install pandas
```

Reading Files using pandas

```
import pandas as pd
df = pd.read_excel('Example.xlsx')
print(df)
```



```
   Roll No.  English  Maths  Science
0         1       19     13       17
1         2       14     20       18
2         3       15     18       19
3         4       13     14       14
4         5       17     16       20
5         6       19     13       17
6         7       14     20       18
7         8       15     18       19
8         9       13     14       14
9        10       17     16       20
```

- Python provides inbuilt functions for creating, writing, and reading files.
- There are two types of files that can be handled in python, normal text files and binary files (written in binary language, 0s, and 1s).
- **Text files:** In this type of file, Each line of text is terminated with a special character called EOL (End of Line), which is the new line character (‘\n’) in python by default.
- **Binary files:** In this type of file, there is no terminator for a line, and the data is stored after converting it into machine-understandable binary language.

```
# Open function to open the file "MyFile1.txt"
# (same directory) in append mode and
file1 = open("MyFile.txt","a")

# store its reference in the variable file1
# and "MyFile2.txt" in D:\Text in file2
file2 = open(r"MyFile.txt","w+")
```

```
# a file named “myfile", will be opened with the reading mode.
file = open(‘myfile.txt', 'r')
# This will print every line one by one in the file
for each in file:
        print (each)
```

Data Cleaning

- **Data cleaning** is one of the important parts of machine learning.
- It plays a significant part in building a model.
- It surely isn’t the fanciest part of machine learning and at the same time, there aren’t any hidden tricks or secrets to uncover.
- However, the success or failure of a project relies on proper data cleaning.
- Professional data scientists usually invest a very large portion of their time in this step because of the belief that **“Better data beats fancier algorithms”**.

- What Is Data Munging?
- Data munging is the process of cleaning and transforming data prior to use or analysis.
- Without the right tools, this process can be manual, time-consuming, and error-prone.
- Many organizations use tools such as Excel for data munging.
- While Excel can be used for the data munging process, it lacks the sophistication and automation to make the process efficient.
- In most organizations, 80% of the time spent on data analytics is allocated to data munging, where IT manually cleans the data to pass over to business users who perform analytics.
- Data munging can be a time consuming and disjointed process that stands in the way of extracting true value and potential from data.

Why Is Data Munging Important?

- Data’s messy, and before it can be used for analysis and driving business objectives, it needs a little tidying up.
- Data munging helps remove errors and missing data so that data can be used for analysis.
- Here’s a look at some of the more important roles data munging plays in data management.
- Data Preparation, Integration, and Quality
- Data Enrichments and Transformation
- Data Analysis
- Time and Resource Efficiency

- **What is data manipulation?**
- Data manipulation is the process of organizing or arranging data in order to make it easier to interpret.
- Data manipulation typically requires the use of a type of database language called data manipulation language (DML).
- DML is a type of coding language that allows you to reorganize data by modifying it within its database program.

- Common operations used for data manipulation include:
- Aggregation
- Classification
- Mathematical formulas
- Regression analysis
- Row and column filtering
- String concatenation

- **Examples of data manipulation**
-

  Data manipulation can be used to:
- Arrange data alphabetically or by date to find individual entries
- Manage web server logs where website owners can monitor most-viewed web pages and traffic sources
- Create forecasts of stock market trends
- Assess the expense of products, pricing patterns or future tax obligations
- View online information in a more useful way to users based on code in a user-defined software program

- **Data Rescaling**
- Your preprocessed data may contain attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume.
- Many machine learning methods expect or are more effective if the data attributes have the same scale. Two popular data scaling methods are normalization and standardization.
- Normalization refers to rescaling real valued numeric attributes into the range 0 and 1.
- It is useful to scale the input attributes for a model that relies on the magnitude of values, such as distance measures used in k-nearest neighbors and in the preparation of coefficients in regression.