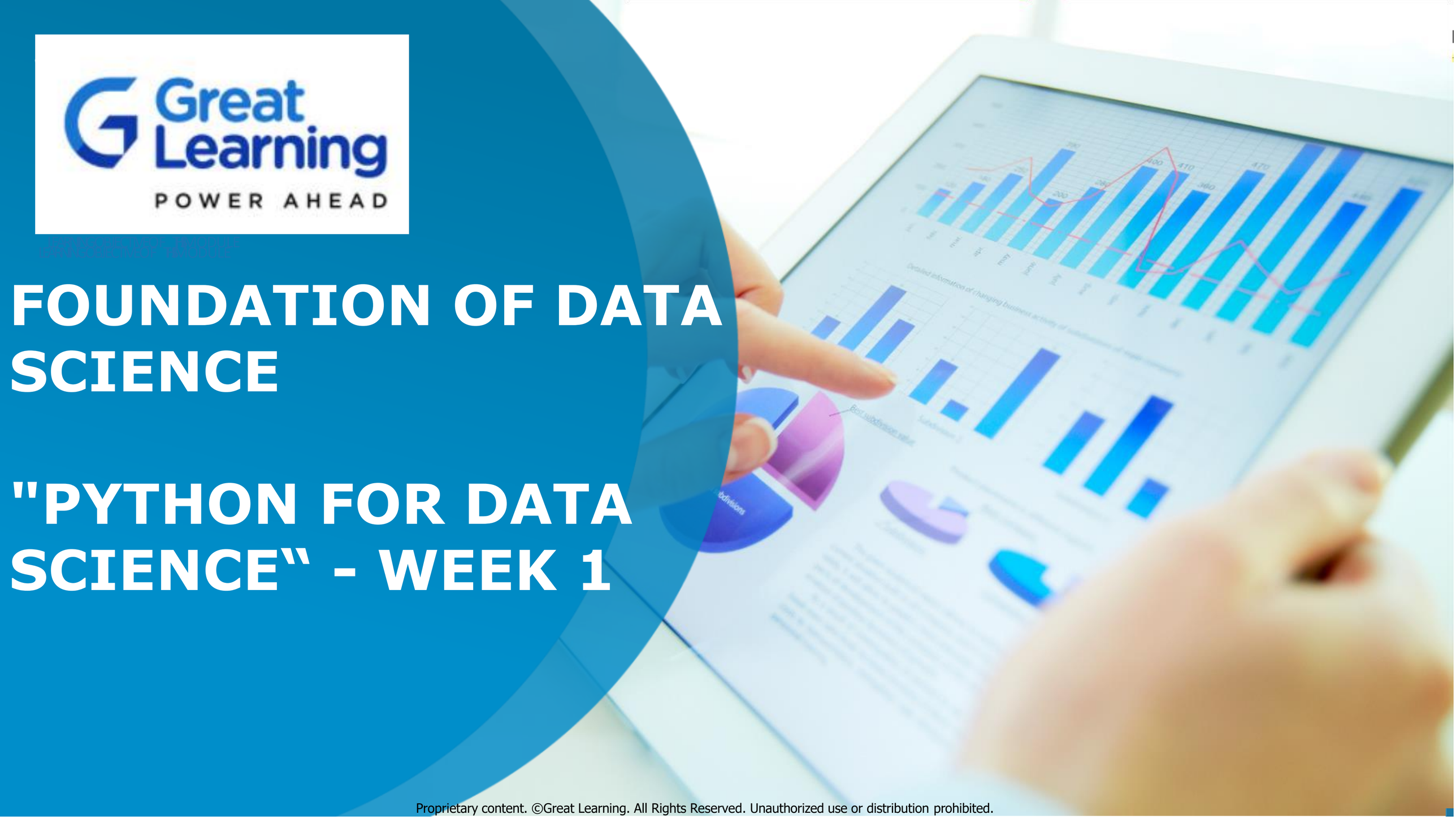
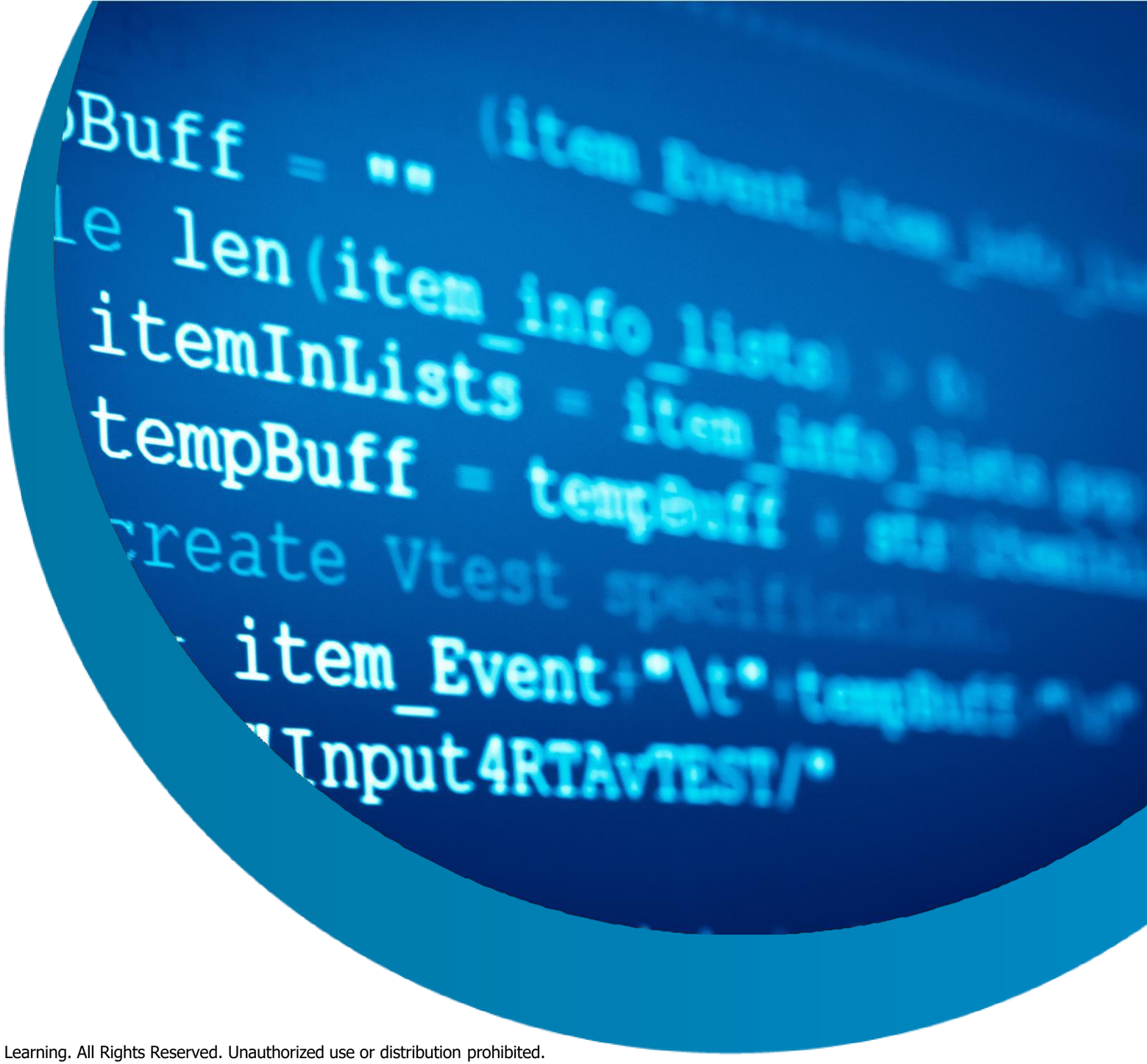


FOUNDATION OF DATA SCIENCE

"PYTHON FOR DATA SCIENCE" - WEEK 1

LEARNING OBJECTIVE OF THIS MODULE

- Basic Working proficiency in Python
- Basic Data-Manipulation using Python
- Basic Data-Visualization using Python
- Basics of Statistics



LET’S SET SOME GROUND RULES

- Come prepared for these sessions by watching the videos.
  - Concepts will be covered in the videos.

- Submit all assignments on time.

- Let’s be punctual & respect each other’s time.

A Few Analytics Application

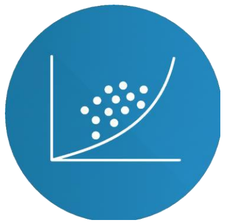
Case1: Can you predict which client will default the loan payment based on the client's spending?

- Why does the bank want to know who will default?
- What type of information I would need about the client to know the risk?
- Do you know what went wrong with ICICI bank and Yes bank

Case2: Can you predict when an employee will resign from his/her organization?

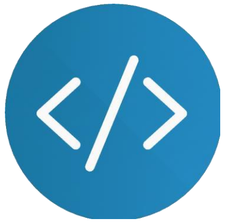
- Why is this important for a company?
- What type of information do we need to make an informed decision ?
- If my company is a 40-50 years old company, should one use all the available data to proceed with this analysis?

LEARNING OBJECTIVES OF THIS SESSION

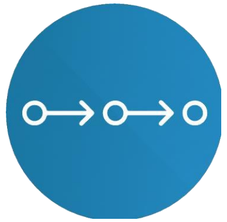


- Understand the big picture of Data Science & Machine Learning

- Introduction to Python



- Basic Operations in Python using a Case Study



- A journey of a thousand miles begins with a single step

Content Covered :

- Python
- Libraries Used
- NumPy
- Pandas
- Visualizations: matlplotlib, seaborn etc.
- Case Study

POP Up Questions:

- How do you define the data science life cycle ?
- What is data ?
- What are libraries and their uses in python ?

Python In Statistics

- One of the fastest growing programming languages
- Great functionality to deal with mathematics, statistics and data science applications.
- Easy to use, easy to debug language that also caters to people with non-programming backgrounds.
- Python libraries can be used as tools to assist you in working with data
- Quantitative analysis: Describes and summarizes data numerically
- Visual analysis: Illustrates data with charts, plots, graphs etc.

Frequently Used Libraries

Data Processing & Analysis

NumPy

Pandas

Data Visualization

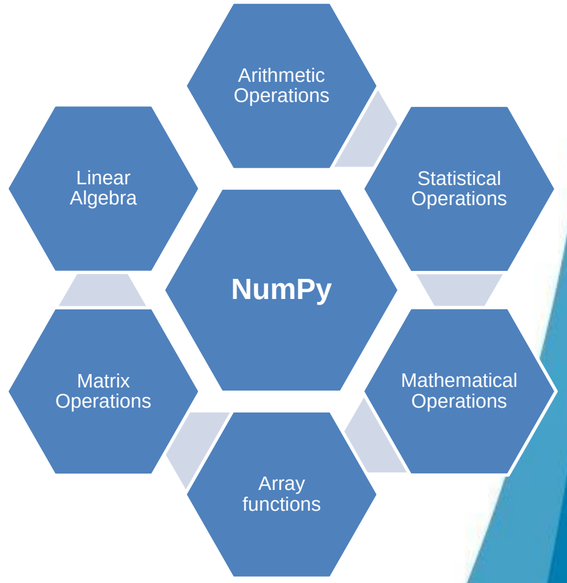
Matplotlib

Seaborn

Pandas

Numpy

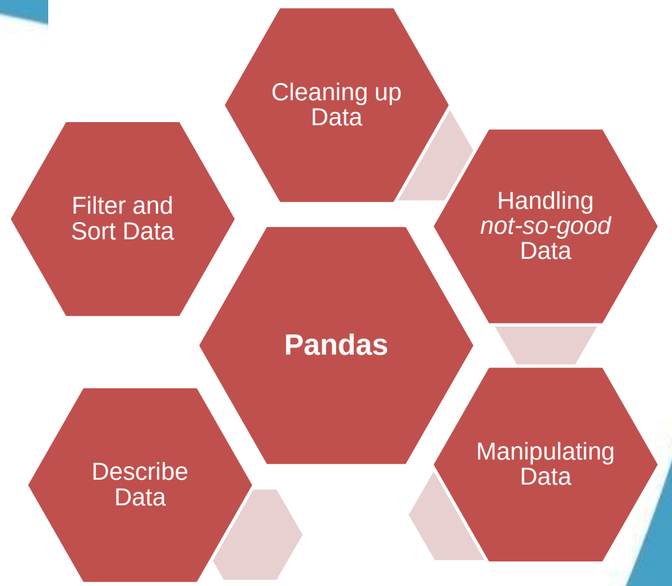
To use NumPy -> *import numpy as np*



- Stands for Numerical Python
- Library used for working with linear algebra and multi-dimensional arrays
- First step on the journey of a data scientist working with Python
- Pre-requisite for installing NumPy is Python itself.
- If you use the Anaconda distribution, you will automatically be able to use the common libraries, NumPy being one of them.

Pandas

To use Pandas -> *import pandas as pd*



- Built on top of NumPy
- Data processing, manipulation and analysis tool
- Used for typical data processing steps: load, prepare, manipulation and saving
- Additionally, it is also used for data merging and joining, data normalization, data modeling and analysis.
- If you use the Anaconda distribution, you will automatically be able to use the common libraries, pandas being one of them.

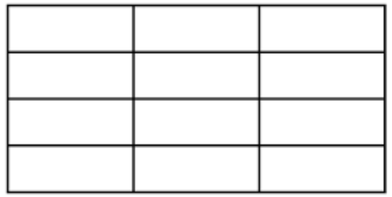
Data Visualization

- Visual analysis of data to determine patterns and/ or gather insights.
- Helps people to understand the summary of data by summarizing and presenting data in a simple, easy-to-understand format.
- Python offers some great libraries such as *matplotlib* and *seaborn* for creating graphs that are not only interactive but can also be customized.
- Some of the commonly used visualizations are bar plots, pie charts, line charts, scatter plots, boxplots, histograms, heatmaps etc.

Case Study -Introduction

Analyze the net worth of Forbes Top Billionaires 2020

Importing the Data



Loading the pre-requisite libraries:

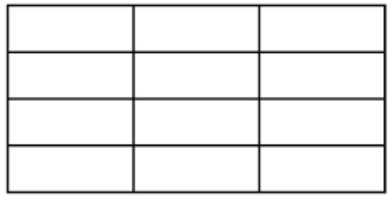
To use the functions available in a library, the libraries need to be imported as follows.

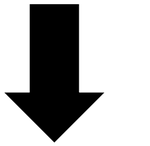
Instead of using the library names repeatedly, it is a good practice to give them shorter names.

import numpy as np

import pandas as pd

Importing the Data

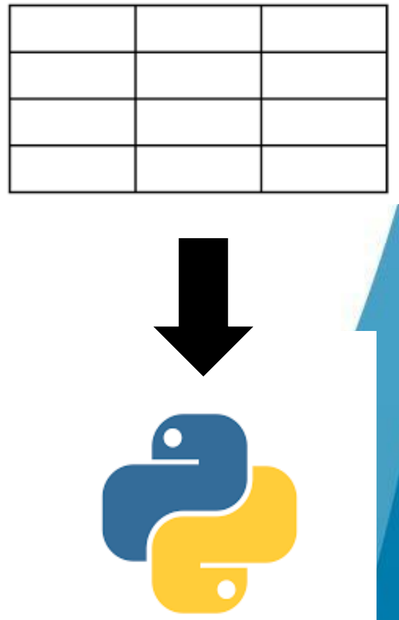




In case you need to use a library that doesn't pre-exist with the installation:

- Open Anaconda Command prompt as administrator
- Use /cd to come out of a particular path (if needed)
- Run pip install/uninstall command (for example, pip install seaborn)

Reading the Data



Reading the dataset in Python:

The data can be imported by giving any data frame name and

using the following command.

It is a good practice to keep the input file in the same location as

the python file.

df = pd.read_csv('Forbes_Billionaires.csv')

Knowing the Data

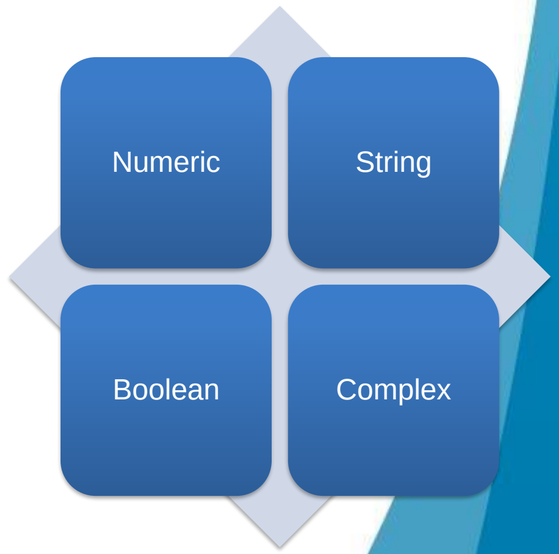
Knowing the data types:

To know the data type of each column in the dataset, use the following command.

The command lists all columns, the no. of values in them and their data types.

df.info()

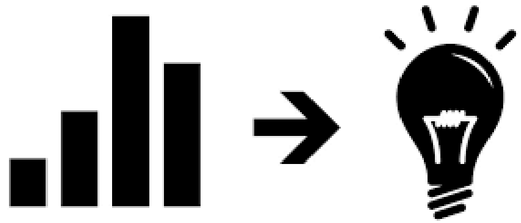
Knowing the Data



Data in python can belong to any of the following types:

- Numeric (integer or float)
- String (also c/a object)
- Boolean
- Complex

Understanding the Data



<u>Summary of the data</u>:

To understand each of the variable in the data, we may look at their mean, median and mode values.

These values can be calculated by appending the following to the column names:

.mean()

.median()

.mode()

We may also use the following command to get this information about all the columns:

df.summary()

Cleaning the Data

The data that we see is often “unclean” and not fit for use.

- It is a good practice to clean the data before it is analyzed.
- Following are few examples of unclean data:

–Improper column names (such as v1, v2)

–Lengthy column names

–Missing Values

–Outliers & others..

Missing Values

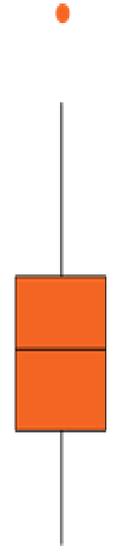
Finding and treating missing values in the data:

Often, the data contains missing values i.e., a few values of a column that are not available in the data.

To find if there are any missing values in the data, we may do the following:
df.isnull.sum()

If there are missing values in the data, they need to be treated before analysis.

Plots & Its Use



(Boxplot)

sns.boxplot()

A boxplot uses what is called as IQR to find outliers in the data:

- IQR stands for Inter-Quartile Range.
- It is defined as Q3-Q1, where Q3 is the 75th percentile of the data and Q1 is the 25th percentile
- Remember, median is the 50th percentile ☺

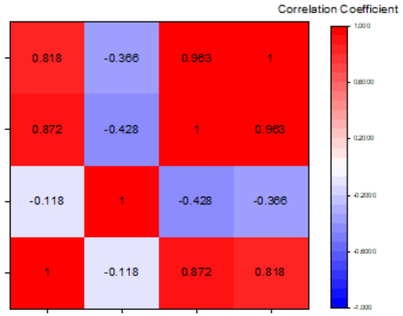
(Stripplot)

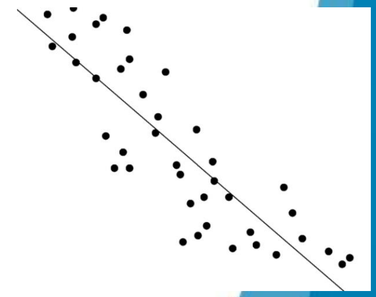
sns.stripplot()

(Countplot)

sns.countplot()

Heatmap & pairplot

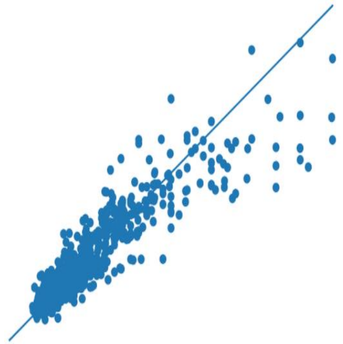


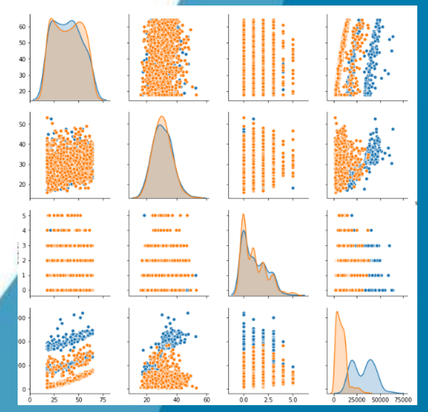


Correlation is the extent to which two variables are linearly related

Two variables (i.e., two numeric variables) may have one of the following relationship with each other:

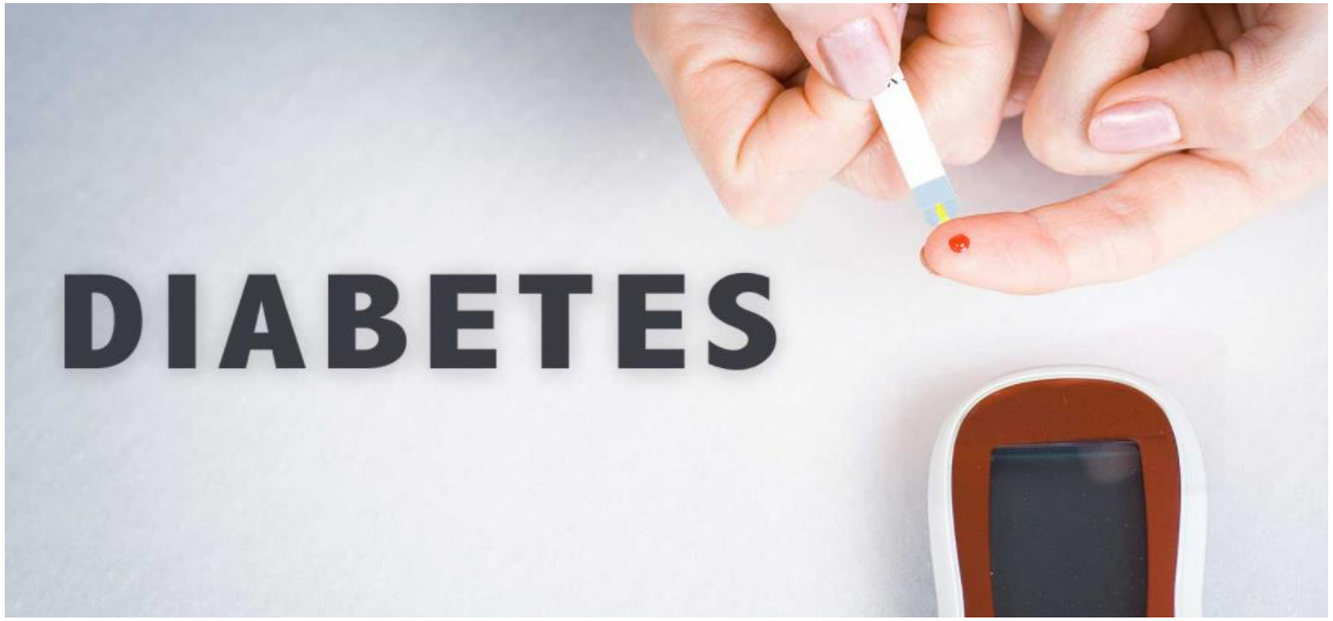
- Positively correlated
- Negatively correlated
- Not correlated





There are different ways to know the correlation in a data:
.corr(), .pairplot(), .heatmap()

PDS- Pima India Diabetes Project



Do you know?

In these 2 weeks, you will learn techniques to analyze the data and understand the patterns in the given data.

In the upcoming Project, you will be working on a real data which is based on the patients having diabetes. The objective of the project will be to analyze different aspects of having diabetes.

ANY QUESTIONS

HAPPY LEARNING