INTRODUCTION TO PYTHON PROGRAMMING FOR DATA SCIENCE

Rahama Sani

What Is Data Science

- Data Science is about data gathering, analysis and decision making
- Data Science is about finding patterns in data, through analysis, and making future predictions,
- Data Science is a process, not an event. It is the process of using data to understand different things.
- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. In simpler terms, data science is about obtaining, processing, and analyzing data to gain insights for many purposes.

Why Data Science?

- Data Science is used in many industries in the world today such as banking, consultancy, healthcare and manufacturing
- Data Science has witnessed recent remarkable growth due to the abundance of electronic data, computing power, advancements in AI and demonstrated business value.
- There is an increased adoption of data science across industries. This has fueled demand for skilled data scientists.

What Is Data?

- Data is the foundation of Data Science. It is the material on which all the analyses are based.
- There are two classes of data: traditional and big data.
- Traditional data is data that is structured and stored in databases which analysts can manage from one computer. It is usually in table format, containing numeric or text values.
- Big data not only contains numeric or text values but could also contain image data, audio data e.t.c. It also has a larger volume and is usually distributed across a network of computers

Data types

- There are two main data types: numerical and categorical. Numerical data is quantitative and can be represented by numbers. Categorical data is qualitative and can be represented by labels or names.

Nominal and ordinal data

- Nominal data is a type of data that consists of categories that cannot be ordered. For example favorite colours.
- Ordinal data is a type of data that can be ordered but not necessarily measured. For example, product rating from 1-5

Discrete and continuous data

- Discrete data can only take discrete values. It includes whole, concrete numbers with specific and fixed data values determined by counting.
- Continuous data is data that can be calculated. It has an infinite number of possible values. It is usually measured over a particular time interval.

Examples of where Data Science is needed

- To foresee delays for flight/ship/train e.t.c

  (through predictive analysis)

- To find the best suited time to deliver goods

- To forecast the next years revenue for a company

How does a Data Scientist work?

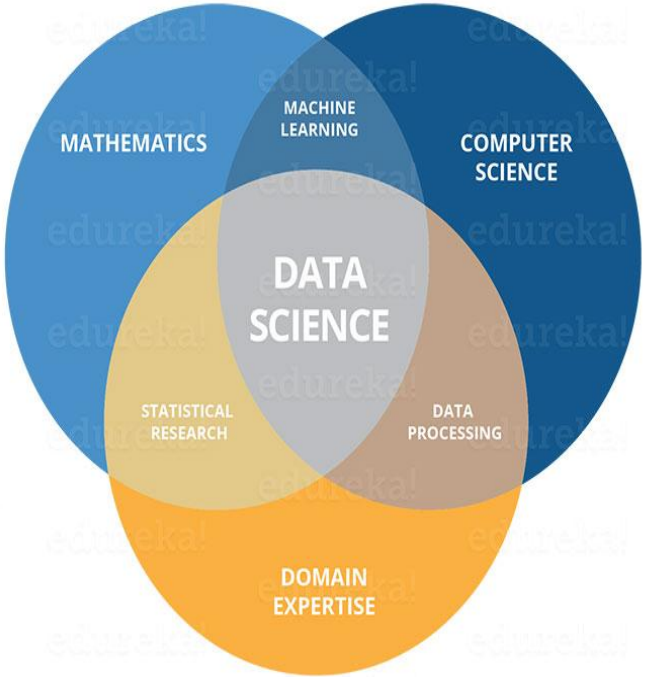
A Data Scientist requires expertise in several backgrounds which include

- **Mathematics**: Mathematics is an important part of data science. Make sure you know the basics of university maths from calculus to algebra. The more you know, the better.
- **Statistics**: You must know statistics in order to infer insights from smaller datasets onto a larger population. You need to know statistics in order to play with data. Statistics allow you to slice and

dice through data, extracting the insights you need to make reasonable conclusions. To understand data science, you need to know the basics of hypothesis testing and experiment design in order to understand the meaning of your data.

- **Python(or R) programming**: Python is a programming language widely used by Data Scientists. It has inbuilt mathematical libraries and functions making it easier to calculate mathematical problems and perform data analysis.

- **Databases**: A database is an organised collection of structured information or data typically stored electronically in a computer system. A database management system is a software application that interact with the user, other applications and the database to capture and analyze data.
- **Machine Learning**: Machine learning is used to automate data analysis process and make predictions in real time without any human involvement.


MATHEMATICS
MACHINE LEARNING
COMPUTER SCIENCE
DATA SCIENCE
STATISTICAL RESEARCH
DATA PROCESSING
DOMAIN EXPERTISE

Tools used in Data Science

With your skillset developed, you will need to understand how to use modern data science tools. Each tool has their strengths and weaknesses and plays a different role in the data science process. You can choose one of them to use or use all of them. We are going to talk about the most popular data science tools here. But before then let us first talk about file formats.

File Formats

Data can be stored in different file formats but some of the most popular are CSV, JSON, XLSX, HTML, TXT.

- CSV: CSV stands for Comma-separated values. as-well-as this name CSV file is use comma to separated values. In CSV file each line is a data record and Each record consists of one or more than one data fields, the field is separated by commas.

- JSON: JSON is stand for JavaScript Object Notation. JSON is a standard text-based format for representing structured data based on JavaScript object syntax.
- XLSX: The XLSX file is Microsoft Excel Open XML Format Spreadsheet file. This is used to store any type of data but it's mainly used to store financial data and to create mathematical models etc.
- HTML: HTML is stand for stands for Hyper Text Markup Language and it is used for creating web pages.

- Excel: Excel allows you to easily manipulate data with what is essentially a what you see is what you get editor. It allows you manipulate data without working on code at all. It is easy to start with and is very useful in communicating data to people who may not have programming skills.

- SQL: Data Scientists need SQL. SQL is a programming language specifically designed to extract data from databases.

- R: R is a staple in the data science community because it is designed explicitly for data science needs. It is one of the most popular programming environment in data science. R shines when it comes to building statistical models and displaying the results.

- Python: Python is the most powerful, versatile programming language in data science. It also has a simple syntax that is easy to learn.

Bringing tools into the data science process

Each of the tools listed above can be used at different stages of the data science process. This is summarized in the table below.

	Excel	SQL	Python	R
Collect data		X	X	X
Process data	X	X	X	X
Explore data	X		X	X
Analyse data	X		X	X
Communicate data	X		X	X

A data scientist must find patterns within the data. Before he/she can find patterns, he/she must organise the data in a standard format. Here is how a data scientist works:

- Ask the right questions: To understand the business problem. If you do not understand the problem correctly, you might end up with a wrong solution.
- Explore and collect data: Sometimes, finding the data that you need is hard. Thankfully, there are many helpful resources. You could either source data yourself e.g., via customer feedback, weblogs,

database e.t.c. or you could download public datasets from various websites e.g., kaggle.

- Process the data: Transform the data into standardized format. Remove duplicate rows/columns, filter the data you need, take care of erroneous missing or inconsistent values. Then you normalize the data by scaling values in a practical range. It is more quicker to process large datasets with python than with excel. It is also more functional.
- Explore data: Add columns, get averages, do basic statistical and numerical analysis.

- Find patterns and make future predictions.
- Represent the result: Present the result with useful insights in a way the “company” can understand.