P P Savani University
School of Engineering

Institute of Computer Science and Application

SSCA3021: Data Science

By Misha Patel

PPSU
P P SAVANI UNIVERSITY

Module 1

What Is Data Science?

What is Data Science?

"Torture the data, and it will confess to anything."
~ Ronald Coase, Economics, Nobel Prize

Data Science is the process of extracting knowledge and insights from data by using *scientific methods*.

Scientific methods:

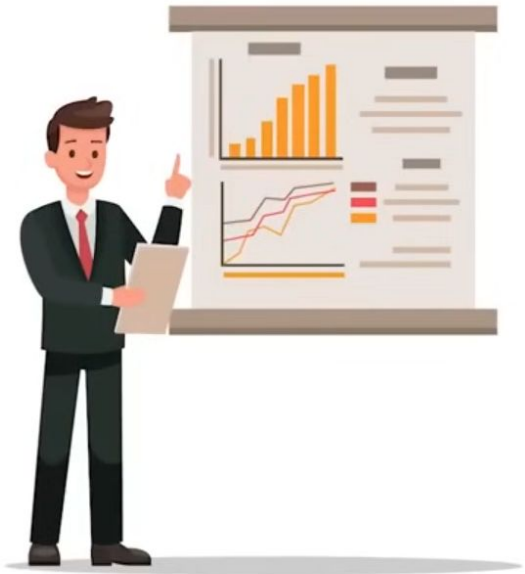
Programming + **Statistics** + **Business**

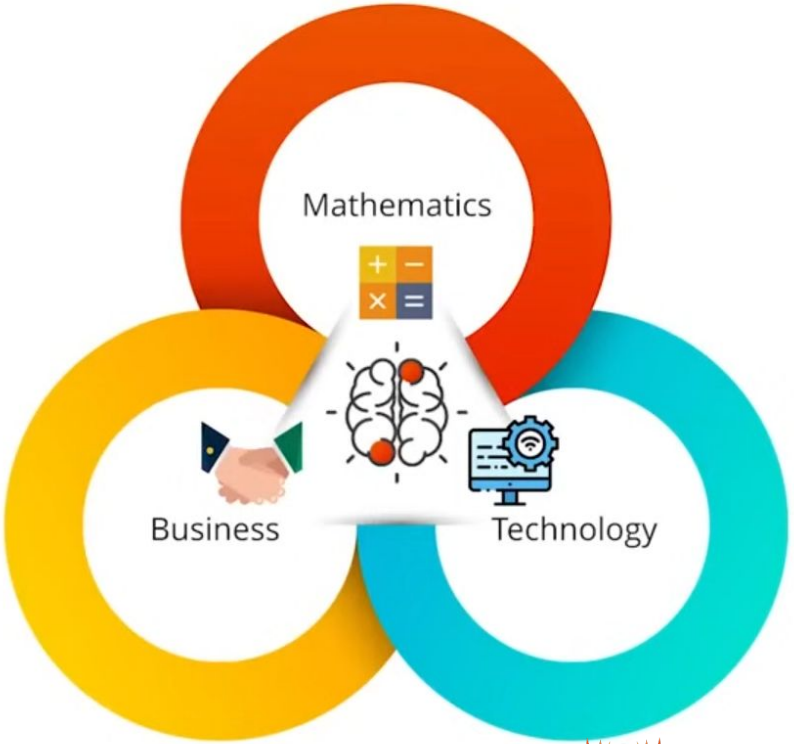
PPSU
P P SAVANI UNIVERSITY
RECOGNISED BY UGC

What is Data Science?

- Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms, and systems to extract or extrapolate knowledge and insights from data.
- Data science is collecting, analyzing, and interpreting data to gather insights into the data that can help decision-makers make informed decisions.
- In simple terms, data science helps to analyze data and extract meaningful insights from it by combining statistics & mathematics, programming skills, and subject expertise.
- Data can be proved to be very fruitful if we know how to manipulate it to get hidden patterns from it. This logic behind the data or the process behind the manipulation is what is known as data science.

Who Is A Data Scientist?





Data Science – Skill Set

Statistics

Programming languages

Data extraction & processing

Data wrangling & exploration

Machine Learning

Big Data processing frameworks

Data visualisation

Data science Terminology

- **Algorithm:** A set of rules or steps used to solve a problem or perform a computation.
- **Big Data:** Extremely large datasets that can be analyzed computationally to reveal patterns, trends, and associations.
- **Classification:** A machine learning task of predicting a discrete label or category for an input.
- **Clustering:** The task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- **Data Cleaning:** The process of correcting or removing inaccurate records from a dataset.
- **Data Mining:** The practice of examining large pre-existing databases to generate new information.
- **Supervised Learning:** A type of machine learning where the model is trained on labeled data.
- **Unsupervised Learning:** A type of machine learning where the model is trained on unlabeled data and must find structure in the input data.

Where we Find DATA?

Data Sources

Evolution of Technology

IOT

Social Media

Other factors

Telephone

Desktop

Car

Mobile

Cloud

Smart Car

Data Sources

- Evolution of Technology
- IOT
- Social Media
- Other factors

Data Sources

- Evolution of Technology
- IOT
- Social Media
- Other factors

1,736,111 pictures

347,222 tweets

204,000,000 emails

4,166,667 likes & 200,000 photos

300 hours of video uploaded

Data Sources

- Evolution of Technology
- IOT
- Social Media
- Other factors

Data Analysis At Walmart

Halloween and cookie sales

Data scientist at Walmart found a connection between Halloween and the sales of cookies.

Data Analysis At Walmart

Social media and cake pops

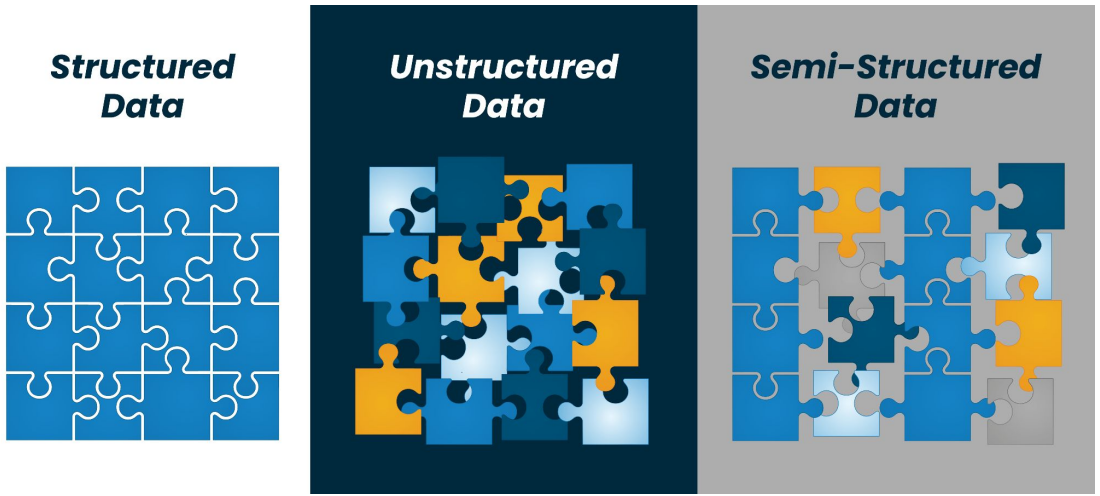
Walmart is leveraging social media data to find about the trending products so that they can be introduced to the Walmart stores across the world

WHAT IS DATA?

Data refers to facts and statistics collected together for reference or analysis.

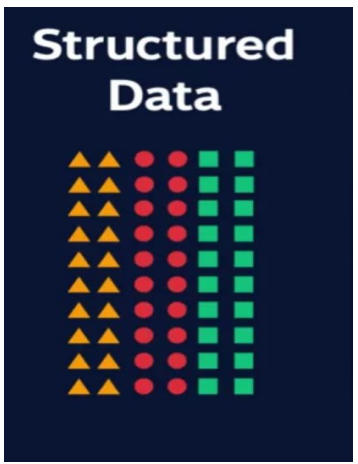
Types of DATA.

- One purpose of Data Science is to structure data, making it interpretable and easy to work with.
- Data can be categorized into three groups:
  - Structured data
  - Semi-structured data
  - Unstructured data



Structured Data

- Structured data is the data that conforms to a data model, has a well-defined structure, follows a consistent order, and can be easily accessed and used by a person or a computer program.
- Structured data is usually stored in well-defined schemas such as Databases. It is generally tabular with columns and rows that clearly define its attributes.
- SQL (Structured Query Language) is often used to manage structured data stored in databases.
- **Sources of Structured Data:**
  - SQL Databases
  - Spreadsheets such as Excel
  - Online Transaction Processing (OLTP) systems
  - Online forms
  - Sensors such as GPS or RFID tags
  - Network and Web server logs
  - Medical devices

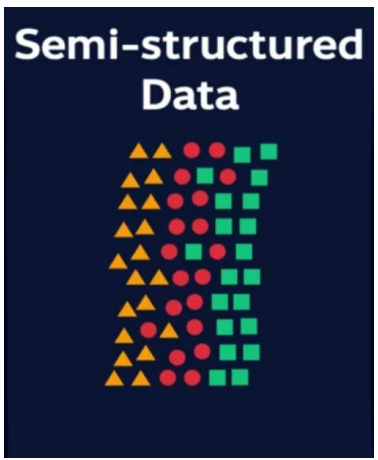


Characteristics of Structured Data

- Data conforms to a data model and has easily identifiable structure
- Data is stored in the form of rows and columns
- **Example:** Database
- Data is well organized so, Definition, Format and Meaning of data is explicitly known
- Data resides in fixed fields within a record or file.
- Similar entities are grouped together to form relations or classes
- Entities in the same group have same attributes
- Easy to access and query, So data can be easily used by other programs
- Data elements are addressable, so efficient to analyse and process

Semi-Structured Data

- Semi-structured data is data that does not conform to a data model but has some structure.
- It lacks a fixed or rigid schema.
- It is the data that does not reside in a rational database but that has some organizational properties that make it easier to analyze.
- With some processes, we can store them in the relational database.
- **Sources of Semi-Structured Data:**
  - E-mails
  - XML and other markup languages
  - Binary executables
  - TCP/IP packets
  - Zipped files
  - Integration of data from different sources
  - Web pages

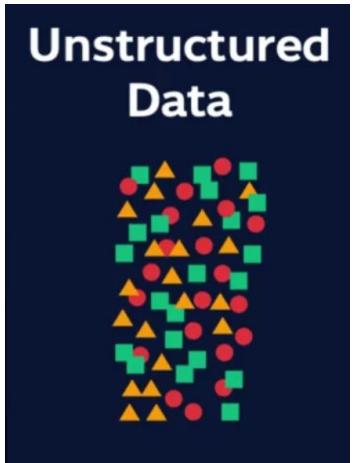


Characteristics of Semi-Structured Data

- Data does not conform to a data model but has some structure.
- Data can not be stored in the form of rows and columns as in Databases
- Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored
- Similar entities are grouped together and organized in a hierarchy
- Entities in the same group may or may not have the same attributes or properties
- Does not contain sufficient metadata, which makes automation and management of data difficult
- Size and type of the same attributes in a group may differ
- Due to lack of a well-defined structure, it cannot be used by computer programs easily

Unstructured Data

- Unstructured data is not organized. We must organize the data for analysis
- purposes.
- Unstructured data is the data that does not conform to a data model and has no easily identifiable structure such that it cannot be used by a computer program easily.
- Unstructured data is not organized in a pre-defined manner or does not have a pre-defined data model; thus, it is not a good fit for a mainstream relational database.
- **Sources of Semi-Structured Data:**
  - Web pages
  - Images (JPEG, GIF, PNG, etc.)
  - Videos
  - Memos
  - Reports
  - Word documents
  - PowerPoint presentations
  - Surveys



Characteristics of Unstructured Data

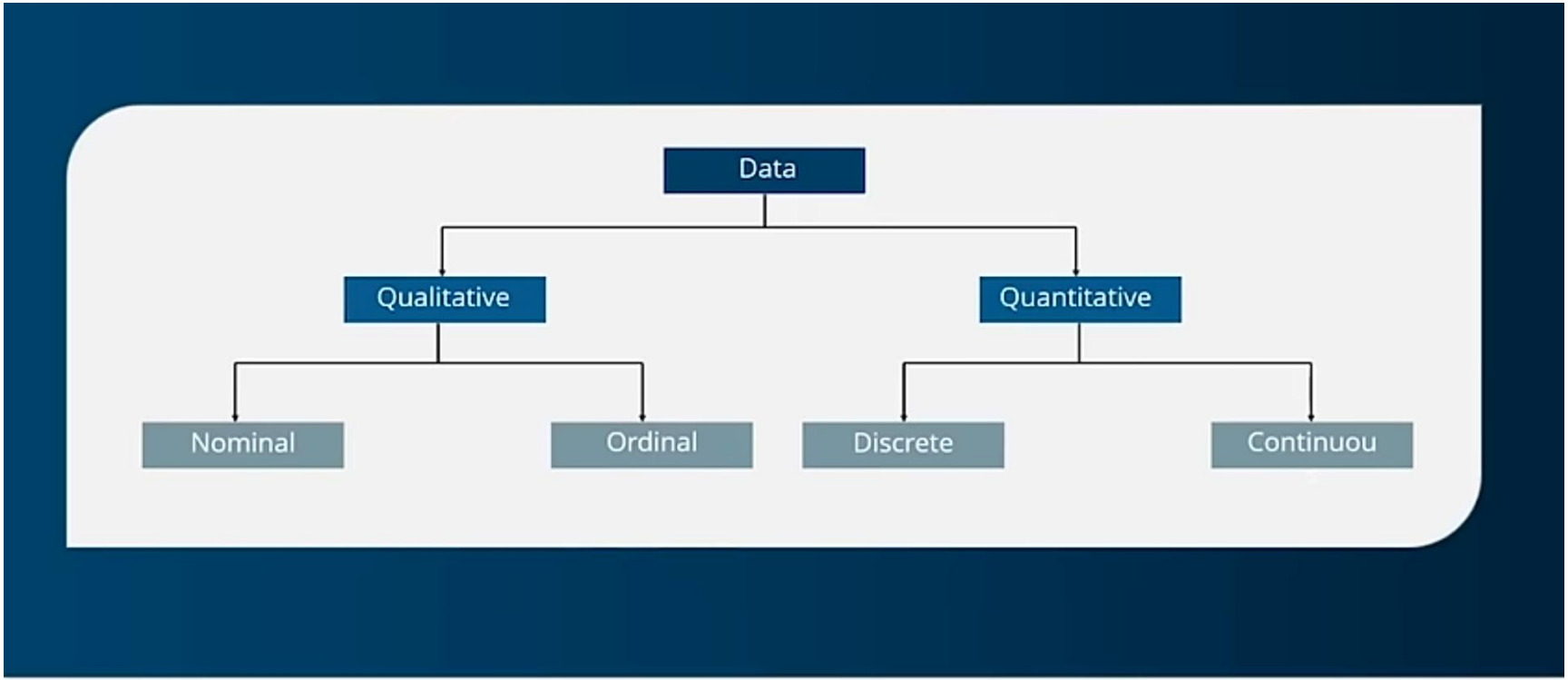
- Data neither conforms to a data model nor has any structure.
- Data can not be stored in the form of rows and columns as in Databases
- Data does not follow any semantic or rules
- Data lacks any particular format or sequence
- Data has no easily identifiable structure
- Due to lack of identifiable structure, it cannot be used by computer programs easily

Categorize the following data

- Customer database table: ...?
- Chat messages from a messaging app: ...?
- Amazon product reviews in JSON format: ...?
- Digital photographs: ...?
- Employee records in an Excel file: ...?
- HTML source code of a webpage: ...?
- System error logs: ...?
- Research papers in PDF format: ...?
- Sales data in a CSV file: ...?
- Emails with attachments: ...?

Categorize the following data

- Customer database table: **Structured**
- Chat messages from a messaging app: **Unstructured**
- Amazon product reviews in JSON format: **Semi-structured**
- Digital photographs: **Unstructured**
- Employee records in an Excel file: **Structured**
- HTML source code of a webpage: **Semi-structured**
- System error logs: **Semi-structured**
- Research papers in PDF format: **Unstructured**
- Sales data in a CSV file: **Structured**
- Emails with attachments: **Semi-structured**



Types Of Data

Qualitative or Categorical Data

- Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers.
- These types of data are sorted by category, not by number. That's why it is also known as Categorical Data.
- These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or other, is qualitative data.
- Qualitative data tells about the perception of people.
- This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.
- **The other examples of qualitative data are,**
    - What language do you speak
    - Favorite holiday destination
    - Opinion on something (agree, disagree, or neutral)
    - Colors

QUALITATIVE DATA

Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively.

Nominal Data

Data with no inherent order or ranking such as gender or race, such kind of data is called Nominal data

Gender
Male
Female
Male
Male

Ordinal Data

Data with an ordered series, such as shown in the table, such kind of data is called Ordinal data

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

Qualitative or Categorical Data-Nominal Data

- Nominal Data is used to label variables without any order or quantitative value.
- The color of hair can be considered nominal data, as one color can't be compared with another color.
- The name "**nominal**" comes from the Latin name "**nomen**," which means "**name**."
- With the help of nominal data, we can't do any numerical tasks or give any order to sort the data.
- These data don't have any meaningful order; their values are distributed into distinct categories.
- **The other examples of Nominal Data are,**
  - Colour of hair (Blonde, red, Brown, Black, etc.)
  - Marital status (Single, Widowed, Married)
  - Nationality (Indian, German, American)
  - Gender (Male, Female, Others)
  - Eye Color (Black, Brown, etc.)

Qualitative or Categorical Data-Ordinal Data

- Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale.
- These data are used for observation, like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.
- Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered "in-between" qualitative and quantitative data.
- The ordinal data only shows the sequences and cannot be used for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.
- **The other examples of Ordinal Data are,**
    - When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
    - Letter grades in the exam (A, B, C, D, etc.)
    - Ranking of people in a competition (First, Second, Third, etc.)
    - Economic Status (High, Medium, and Low)
    - Education Level (Higher, Secondary, Primary)

Quantitative or Numerical Data

- Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis.
- These kinds of data are also known as Numerical data. It answers questions like “how much,” “how many,” and “how often.”
- For example, the price of a phone, the computer’s RAM, the height or weight of a person, etc., falls under quantitative data.
- Quantitative data can be used for statistical manipulation.
- These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.
- **The other examples of Quantitative Data are,**
    - Height or weight of a person or object
    - Room Temperature
    - Scores and Marks (Ex: 59, 80, 60, etc.)
    - Time

QUANTITATIVE DATA

Quantitative data deals with numbers and things you can measure objectively.

Discrete Data

Also known as categorical data, it can hold finite number of possible values.

Example: Number of students in a class

Continuous Data

Data that can hold infinite number of possible values.

Example: Weight of a person

Quantitative or Numerical Data—Discrete Data

- The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers.
- The total number of students in a class is an example of discrete data.
- These data can't be broken into decimal or fraction values.
- The discrete data are countable and have finite values; their subdivision is not possible.
- These data are represented mainly by a bar graph, number line, or frequency table.
- **The other examples of Discrete Data are,**
    - Total number of students present in a class
    - Cost of a cell phone
    - Numbers of employees in a company
    - The total number of players who participated in a competition
    - Days in a week

Quantitative or Numerical Data—Continuous Data

- Continuous data are in the form of fractional numbers. It can be the version of an Android phone, the height of a person, the length of an object, etc.
- Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.
- The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers.
- **The other examples of Continuous Data are,**
    - Temperature
    - height
    - Width
    - Time
    - Speed

Categorize the following data items...

- The color of a car...?
- The number of students in a classroom...?
- The height of a person...?
- The type of cuisine (e.g., Italian, Chinese, Mexican)...?
- The temperature in Celsius...?
- The brand of a smartphone...?
- The age of a tree...?
- The rating of a movie (e.g., stars out of 5)...?
- The type of pet (e.g., dog, cat, bird)...?
- The number of pages in a book...?

Categorize the following data items...

- The color of a car - **Qualitative**
- The number of students in a classroom - **Quantitative**
- The height of a person - **Quantitative**
- The type of cuisine - **Qualitative**
- The temperature in Celsius - **Quantitative**
- The brand of a smartphone - **Qualitative**
- The age of a tree - **Quantitative**
- The rating of a movie - **Quantitative**
- The type of pet - **Qualitative**
- The number of pages in a book - **Quantitative**

Data Science Life Cycle

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Understand the problem

Identify central objectives

Identify variables that need to be predicted

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

What data do I need for my project?

What are the data sources?

How can I obtain the data?

What is the most efficient way to store and access all of it?

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Transform data into desired format

Data cleaning

- Missing values
- Corrupted data
- Remove unnecessary data

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

understand the patterns in the data

Retrieve useful insight

form hypotheses

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Determine optimal data features for the machine-learning model

Create a model that predicts the target most accurately

Evaluate & test the efficiency of the model

Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Check the deployment environment for dependency issues

Deploy the model in a pre-production/ test environment

Monitor the performance

Data Science Process Life Cycle

- **Understanding the Business Problem:** This is the first step, where you identify and understand the problem that needs to be solved. It involves discussing with stakeholders to gather requirements and define the objectives of the project.
- **Data collection:** The process of systematically gathering and measuring information from various sources to answer research questions, solve problems, and inform decision-making. It's the foundational step in any data science project, providing the raw material for analysis and insights.
- **Preparing the Data:** In this step, you collect and organize the data needed for analysis. This may involve cleaning the data to remove any inconsistencies or errors and transforming it into a suitable format for analysis.
- **Exploratory Data Analysis (EDA):** Here, you analyze the data to discover patterns, relationships, or insights. This helps in understanding the data better and identifying which variables are important for modeling.

Data Science Process Life Cycle: Continue

- **Modeling the Data:** In this step, you use statistical and machine learning techniques to build models that can predict outcomes or classify data based on the patterns identified in the EDA step.
- **Evaluating the Model:** Once the model is built, you need to evaluate its performance. This involves testing the model with a separate dataset to see how well it predicts or classifies the data and making adjustments as necessary.
- **Deploying the Model:** After evaluating and refining the model, the final step is to deploy it into a production environment where it can be used to make predictions or provide insights on new data.

Data science Toolkit

Data science Toolkit

- **Programming Languages**
  - **Python:** Widely used for its simplicity and extensive libraries for data analysis and machine learning.
  - **R:** Popular for statistical analysis and data visualization.
- **Libraries and Frameworks**
  - **Pandas:** A Python library for data manipulation and analysis.
  - **NumPy:** A Python library for numerical computations.
  - **Scikit-learn:** A Python library for machine learning.
  - **TensorFlow:** An open-source framework for machine learning and deep learning.
  - **Keras:** A high-level neural networks API, running on top of TensorFlow.
  - **PyTorch:** An open-source machine learning library based on the Torch library.
  - **Matplotlib:** A plotting library for creating static, interactive, and animated visualizations in Python.
  - **Seaborn:** A Python visualization library based on Matplotlib, providing a high-level interface for drawing attractive statistical graphics.
  - **NLTK:** A library for natural language processing in Python.
  - **SpaCy:** An open-source library for advanced natural language processing in Python.

Data science Toolkit: Continue

- **Data Visualization Tools**
  - **Tableau:** A powerful data visualization tool used for creating interactive and shareable dashboards.
  - **Power BI:** A business analytics tool by Microsoft for visualizing data and sharing insights.
  - **D3.js:** A JavaScript library for producing dynamic, interactive data visualizations in web browsers.
- **Data Storage and Management**
  - **SQL:** A language used for managing and querying relational databases.
  - **Hadoop:** A framework for distributed storage and processing of large datasets.
  - **Spark:** An open-source distributed computing system for big data processing.

Data science Toolkit: Continue

- **Data Collection and Scraping**
  - **Beautiful Soup:** A Python library for web scraping.
  - **Scrapy:** An open-source web crawling framework for Python.
  - **Selenium:** A tool for automating web browsers, often used for web scraping.
- **Cloud Platforms**
  - **AWS:** Amazon Web Services, offering a suite of cloud computing services.
  - **Google Cloud Platform (GCP):** A suite of cloud computing services by Google.
  - **Microsoft Azure:** A cloud computing service created by Microsoft.

Examples and Applications of Data Science

You need to prepare a case study.