Data Science

Unit I

- **Need for data science:**

- Making better business decisions. ...
- Measuring performance. ...
- Providing information to internal finances. ...
- Developing better products. ...
- Increasing efficiency. ...
- Mitigating risk and fraud. ...
- Predicting outcomes and trends. ...
- Improving customer experiences.

There are many facets of data science, including:

- Identifying the structure of data.
- Cleaning, filtering, reorganizing, augmenting, and aggregating data.
- Visualizing data.
- Data analysis, statistics, and modeling.
- Machine Learning.
- Assembling data processing pipelines to link these steps.

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

1) 5 characteristics of quality data

1. ***Validity.*** *The degree to which your data conforms to defined business rules or constraints.*

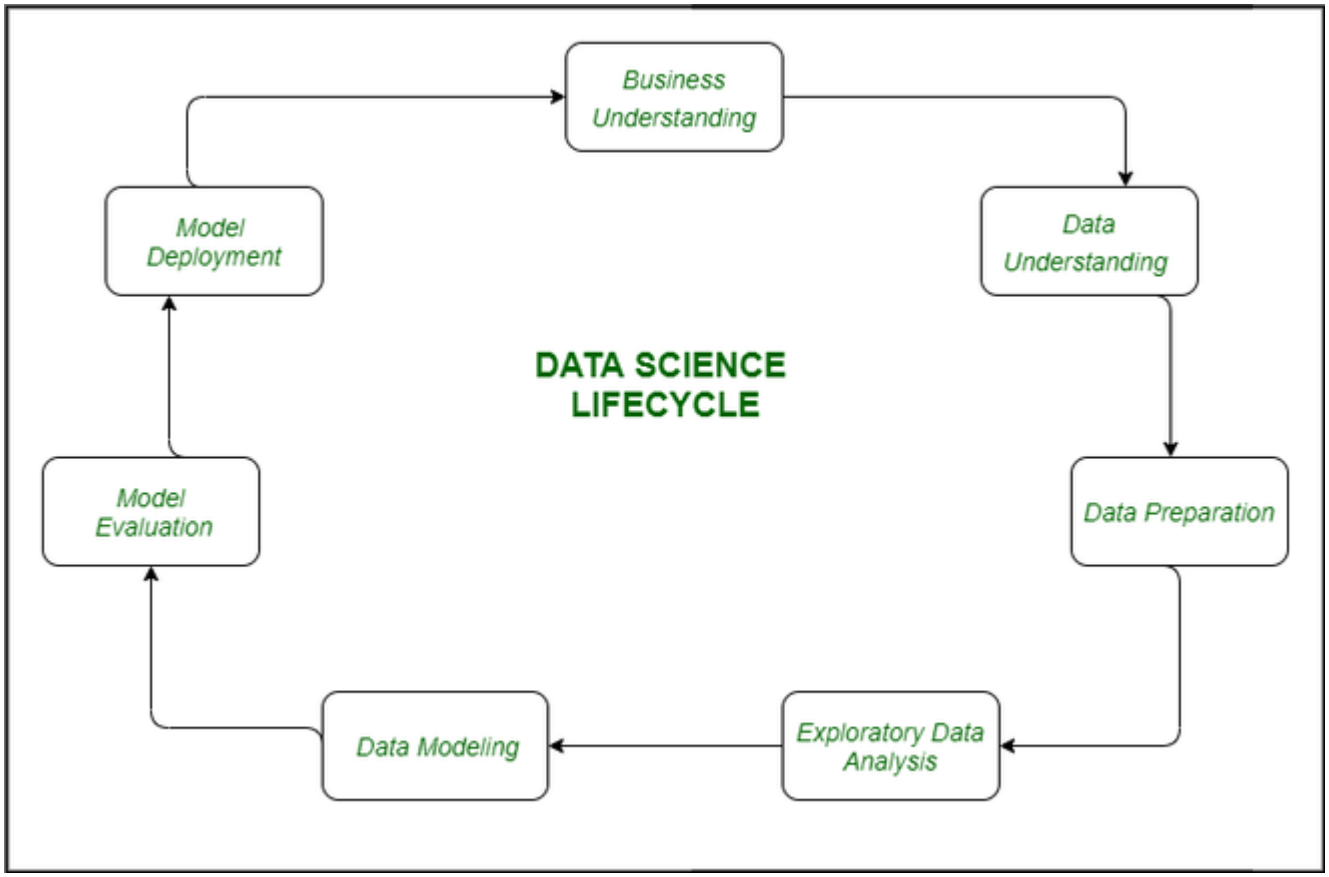
2. **Accuracy.** Ensure your data is close to the true values.
3. **Completeness.** The degree to which all required data is known.
4. **Consistency.** Ensure your data is consistent within the same dataset and/or across multiple data sets.
5. **Uniformity.** The degree to which the data is specified using the same unit of measure.

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

What is exploratory data analysis example?

There are dress shoes, hiking boots, sandals, etc. Using EDA, you are open to the fact that any number of people might buy any number of different types of shoes.
You visualize the data using exploratory data analysis to find that most customers buy 1-3 different types of shoes.

Data Science Lifecycle revolves around the use of machine learning and different analytical strategies to produce insights and predictions from information in order to acquire a commercial enterprise objective. The complete method includes a number of steps like data cleaning, preparation, modelling, model evaluation, etc. It is a lengthy procedure and may additionally take quite a few months to complete. So, it is very essential to have a generic structure to observe for each and every hassle at hand. The globally mentioned structure in fixing any analytical problem is referred to as a Cross Industry Standard Process for Data Mining or CRISP-DM framework.



1. Business Understanding: The complete cycle revolves around the enterprise goal. What will you resolve if you do no longer have a specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

2. Data Understanding: After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

3. Preparation of Data: Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them. Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate

undesirable columns and features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

4. Exploratory Data Analysis: This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

5. Data Modeling: Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

6. Model Evaluation: Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

7. Model Deployment: The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

II. Difference Between Data Science and Data Mining

- **Difficulty Level :** Basic
- **Last Updated :** 01 Feb, 2023

- Read
- Discuss

Data Science: Data Science is a field or domain which includes and involves working with a huge amount of data and uses it for building predictive, prescriptive and prescriptive analytical models. It's about digging, capturing, (building the model) analyzing(validating the model) and utilizing the data(deploying the best model). It is an intersection of Data and computing. It is a blend of the field of Computer Science, Business and Statistics together. **Data Mining**: Data Mining is a technique to extract important and vital information and knowledge from a huge set/libraries of data. It derives insight by carefully extracting, reviewing, and processing the huge data to find out pattern and co-relations which can be important for the business. It is analogous to the gold mining where golds are extracted from rocks and sands.

Data science and data mining are related but distinct fields that involve the extraction of useful information from large data sets.

Data science is a broad field that encompasses various techniques and tools for analyzing and interpreting data. It involves using statistical, machine learning, and programming techniques to extract insights and knowledge from data. Data scientists use data to build predictive models, visualize data, and communicate findings to stakeholders.

Data mining, on the other hand, is a specific technique used within data science to extract patterns and knowledge from large data sets. It typically involves using algorithms and statistical methods to discover hidden patterns and relationships in data. The goal of data mining is to identify useful information from data, such as customer behavior, product preferences, and market trends, that can be used to make better decisions.