Task 1

1. Defining Data Science

In this first topic I got knowledge about data and data science. Anything surrounded by us is data and Data science is a scientific field that uses scientific methods to extract knowledge and insights from structured and unstructured data. The related fields of data science are such as:

- Databases: It organizes data for efficient processing.
- Big data: It manages large volumes of simple data
- Machine Learning: It builds predictive models.
- Artificial Intelligence: It creates complex models emulating human thoughts.
- Visualization: It helps to make sense of data through visual representation.

Data can be Structured, Semi Structured or Unstructured and the key data operations include: Data Acquisition, Data Storage, Data Preprocessing, Visualization and Training a predictive model.

2. Data Science Ethics

Ethics is about the shared values and moral principles that govern our behavior in society. Ethics is not based on laws but on widely accepted norms of what is right vs wrong. Data ethics is a branch of ethics that evaluates moral issues related to data and algorithms. Applied ethics involves investigating and correction ethical issues in real world contexts to align with ethical values while ethics culture ensures consistent and scalable adoption of ethical principles across organizations. Ethical Challenges include handling personal data, ownership, informed consent, intellectual property, data privacy, the right to be forgotten, dataset bias, data quality algorithm fairness, misrepresentation and the illusion of free choice. A case study on data misrepresentation highlights misleading data reporting during the COVID-19 pandemic, which hindered informed public health decisions, particularly in states like Georgia and Virginia, and at the federal level with discrepancies in CDC data. This misrepresentation, whether due to sloppiness or poor science, underscores the importance of accurate data for effective policy-making.

3. Defining Data

Data is of several types such as:

- Raw data: Data is unprocessed and needs organization for analysis. In order to make sense of what is happening with a dataset, it needs to be organized into a format that can be understood by humans as well as the technology they may use to analyze it further.
- Quantitative data: Data consists of numerical observations suitable for mathematical analysis.
- Qualitative data: Data captures subjective qualities and cannot be measured objectively.
- Structured data: Data is organized into rows and columns with a consistent format.

- Semi Structured data: Data combines elements of both Structured and unstructured, being organized but not strictly in rows and columns often following a specific or set of rules.
- Unstructured data: Data lacks a defined format making it flexible for adding new information.

4. Introduction to Statistics

Probability is a measure between 0 and 1 that expresses the likelihood of an event, such as the probability of rolling an even number on a die being 0.5. Probability distribution for discrete random variables assigns probabilities to each event in a sample space, summing to 1. The average indicates central tendency. Average has several types i.e.

- Arithmetic Mean: Sum of values divided by their count, e.g., 4, 3, 1, 6, 1, 7 the mean is (4+3+1+6+1+7)/6= 22/6 = 3.6.
- Median: Middle value in an ordered set, e.g., 1, 1, 3, 4, 6, 7 the median is 3.5.
- Mode: Most common value in number set e.g., 4,3,1,6,1,7 the mode is 1.