# **Bytewise Data Science fellowship**

**Task 1 and 3:**

Summary:

* Data is everywhere
* Data science is extracting knowledge and insights from data.
* Types of data ,structured , unstructured , semi structured

| **Structured** | **Semi-structured** | **Unstructured** |
| --- | --- | --- |
| List of people with their phone numbers | Wikipedia pages with links | Text of Encyclopedia Britannica |

* Gathering data form different sources
* Use databases to store data
* Data processing
* Data utilization for analysis
* We live in a data world, where ethical considerations in data are crucial.
* Market trends predict a significant increase in data trading and usage, increased concerns about data privacy and potential harms from algorithms.
* By 2025, we will create over 180 zettabytes of data, necessitating strong data ethics to protect user rights and privacy.
* concepts like transparency, accountability, fairness, and reliability guide responsible data practices.
* Understanding and addressing ethical challenges through principles, case studies, and applied ethics helps establish a culture of ethics in data science and AI development.
* Data includes facts, information.
* It can be raw, quantitative, or qualitative
* organized as structured, unstructured, or semi-structured.
* Primary data is directly collected, while secondary data is shared.
* Common sources include databases, files, internet, APIs, and web scraping.

Probability measures event likelihood (0 to 1).

* Random variables can be discrete (countable outcomes) or continuous (range of real numbers)
* Mean: Average value.
* Variance: Data spread.
* Standard Deviation: Square root of variance
* Mode: Most frequent value.
* Median: Middle value.
* Quartiles: Divide data into four equal parts.
* Analyzing data using histograms and box plots
* Common real-world data pattern.
* Range where the true population parameter likely lies.
* Testing assumptions about data, such as comparing means.
* Larger samples approximate the population mean.
* Sample means' distribution approaches normal distribution.
* Covariance: Joint variability of two variables.
* Correlation: Normalized covariance, indicating relationship strength and direction

**Git Hub**