Intro to Data Science

Software's:

- R Programming
- Data Mining
- Power BI
- Tableaeu
- Machine Learning

Python: high level programming language, made from C programming language.

- Games
- Data Analysis
- Machine Learning
- Natural Language Processing
- Deep Learning
- Web Development
- Data Scraping

Big Data = 1 TB or more

1024 MegaBytes = 1 Gigabyte
1024 Gigabytes = 1 Terabyte
1024 Terabytes = 1 Petabyte*
1024 Petabytes = 1 Exabyte*
1024 Exabytes = 1 Zettabyte
1024 Zettabytes = 1 Yottabyte

R – S Programming Language (statistical)

- Specialty Language
- Used in Finance & Healthcare Industry

Types of Data:

1. Digital – websites, apps, instant messages, email, voicemail transcripts.
2. Physical – geolocations, sales transactions, traffic monitors.

Data Science – expert study of data. Sees raw data.

Data Analysis – expert study of data using statistics. Only sees structured data.

- ➢ Rows – Attributes
- ➢ Columns – Features

- Unstructured – images, audio, video, e-mails, files, xml.
- Structured – numbers & text.

Line Chart – used to plot change over time and draw attention to the total value across a trend.

1. Continuous data is numerical.
2. Category data = Qualitative = Non-Parametric

- ➢ Quantitative Parametric = Numeric
- ➢ Qualitative = Non-Parametric = Categorical

Mean – Average

Median – Middle Number

Histogram – distribution of variables. Plot quantitative data.

Credible sources for learning:

- **ResearchGate**
- **Scikit – Learn**
- **Khan Academy**

Bias

1. Selection Bias – occurs when a sample population does not reflect the true population.
2. Non – Response Bias
3. Social Desirability Bias

Velocity - Speed data is Transformed
Volume - Infrastructure and amount of data
Variety - Data Sources the data was collected from
Veracity - Quality and usefulness data
Value - Profits it can make a company

Variance

- Measures how far the set of (random) numbers are spread out from mean (average value).
- Tells us the numbers of our mean

Two Types of Variances:

Population – variance from all data.

Sample – variance from a sample data.

Probability Distribution – function used in data science to describe all possible values or outcomes with random variables.

Sampling Distribution:

- Simple Random Sampling
- Systematic Sample (Sample Interval) – with a system (every 3rd person)
- Stratified Random Sample
- Cluster Sample