UNIT _ 1

Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Ben Fry's Model: Visualizing Data Process

1. Acquire
2. Parse (Analyze and put in proper format)
3. Filter
4. Mine (Discovering patterns or knowledge from datasets)
5. Represent
6. Refine
7. Interact

Jeff Hammerbacher's Model

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

7V's of Big Data

• Raw Data: Volume
• Change over time: Velocity
• Data types: Variety
• Data Quality: Veracity
• Information for Decision Making: Value
• Change in Data: Variability
• Presentation of Data: Visualization

Applications of Data Scienece

E Commerce	**Healthcare**
Identifying Consumers Recommending Products Analyzing Reviews	Medical Analysis frug Discovery Bioinformatics Virtual Assistants
Manufacturing	
Predicting Potential problems Monitoring Systems Automating Manufacturing Units Maintenance Scheduling Anomaly Detection	**Transport** Self-Driving Cars Enhanced Driving Experience Car Monitoring System Enhancing the safety of passengers
Banking	**Finance**
Fraud Detection Credit Risk Customer Lifetime Value	Customer Segmentation Strategic Decision Making Algorithmic Trading Risk Analytics

Data Science Life Cycle (Draw a Circle)

- Data Collection
- Data Preparation
- Exploratory Data Analysis
- Modelling
- Model Evaluation

Model Deployment

Area	**BI Analyst**	**Data Scientist**
Focus	Reports, KPIs, trends	Patterns, Correlations, models
Process	Static, comparative	Exploratory, experimentation, visuals
Data sources	Pre-planned, added slowly	On the fly, as needed
Transform	Upfront, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analytics	Retrospective, Descriptive	Predictive, Prescriptive, Preventative

Data Analyst Skills	**Data Scientist Skills**
Data Mining	Data Mining
Data Warehousing	Data Warehousing
Math, Statistics	Math, Statistics, Computer Science
Tableau and Data Visualization	Tableau and Data Visualization/ Storytelling
SQL	Python, R, Java, Scala, SQL, MATLAB, Pig
Business Intelligence	Economics
SAS	Big Data/Hadoop
Advanced Excel Skills	Machine Learning

NumPy/Python

NumPy is a Python library essential for working with arrays, providing a powerful array object called ND array. It is designed for high efficiency, with arrays stored in contiguous memory locations, allowing fast processing and manipulation compared to traditional Python lists. Besides array operations, NumPy includes functions for linear algebra, Fourier transforms, and matrix operations. Standing for Numerical Python, NumPy aims to offer an array object that is up to 50 times faster than standard Python lists, thanks to its optimized storage and extensive support functions that simplify array manipulations.

Pandas

Pandas, short for 'Python Data Analysis Library', is an open-source Python library widely used by data scientists for data exploration, manipulation, and visualization. It is renowned for its ease of use, speed, and powerful capabilities, making it comparable to Microsoft Excel in the Python ecosystem. Pandas integrates seamlessly with other visualization libraries and supports various data formats, including CSV, Excel, SQL databases, and even web pages. Its user-friendly interface and robust functionality make it an indispensable tool for efficient data analysis and processing.

Data Preprocessing: is a technique that is used to convert the raw data into a clean data set. Data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Tasks of Data Preprocessing

Data Cleaning: This is the first step which is implemented in Data Preprocessing. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers, minimizing duplication and computed biases within the data.

Data Integration: This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for analysis.

Data Transformation: This step is used to convert the raw data into a specified format according to the model.

Normalization – In this method, numerical data is converted into the specified range, i.e., between 0 and 1 so that scaling of data can be performed.

Aggregation – This method is used to combine the features into one. For example, combining two categories can be used to form a new group.

Generalization – In this case, lower-level attributes are converted to a higher standard (e.g. age 20, 40 – may be taken as Young, Old, etc.)

Data Reduction: After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data.

How we can deal with the Missing data

Ignoring the missing record: It is the simplest and efficient method for handling the missing data. But this method should not be performed at the time when the number of missing values are immense or when the pattern of data is related to the unrecognized primary root of the cause

Filling the missing values manually: This is one of the best-chosen methods. But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a time-consuming task.

Filling using computed values: The missing values can also be occupied by computing mean, mode or median of the observed given values. Another method could be the predictive values that are computed by using any Machine Learning or Deep Learning algorithm.

How we can deal with the Noisy data

Data Binning: In this approach sorting of data is performed concerning the values of the neighborhood. This method is also known as local smoothing.

Clustering: In the approach, the outliers may be detected by grouping the similar data in the same group, i.e., in the same cluster.

Machine Learning: A Machine Learning algorithm can be executed for smoothing of data. For example, Regression Algorithm can be used for smoothing of data using a specified linear function.

Removing manually: The noisy data can be deleted manually by the human being, but it is a time-consuming process, so mostly this method is not given priority.

Data Wrangling: is used in step during EDA and modeling to adjust data sets interactively while analyzing data and building a model. Data wrangling, also known as data munging, is the process of cleaning, restructuring, and transforming raw data from various sources into a more organized and usable format for analysis. This involves tasks like handling missing data, correcting inconsistencies, merging datasets, creating new variables, and ensuring data quality, ultimately preparing the data for further exploration and insights.

Tasks of Data Wrangling

• **Discovering:** Firstly, data should be understood thoroughly and examine which approach will best suit. For example: if have a weather data when we analyze the data it is observed that data is from one area and so primary focus is on determining patterns.

• **Structuring:** As the data is gathered from different sources, the data will be present in various shapes and sizes. Therefore, there is a need for structuring the data in proper format.

• **Cleaning:** Cleaning or removing of data should be performed that can degrade the performance of analysis.

• **Enrichment:** Extract new features or data from the given Validating: This approach is used for improving the quality of data and consistency rules so that transformations that are applied to the data could be verified.

• **Publishing:** After completing the steps of Data Wrangling, the steps can be documented so that similar steps can be performed for the same kind of data to save time.

Understanding Data Attribute Types

Qualitative Attributes

1. **Nominal Attributes:** Values are names or symbols without numeric value, representing categories or states with no inherent order or rank among them.
2. **Binary Attributes:** Binary data has two states (e.g., yes/no, true/false). Symmetric binary attributes have equally important values with no preference in coding, while asymmetric attributes have one value more important than the other, coded as 1.
3. **Ordinal Attributes:** Values have a meaningful sequence or ranking but the magnitude of differences between them is not known.

Quantitative Attributes

1. **Interval-scaled Attributes**: Concerned with order and differences between values, allowing measurement of standard deviation and central tendency. Values can be added and subtracted but not multiplied or divided. They have no true zero, so ratios cannot be calculated. Example: temperature in Celsius.
2. **Ratio-scaled Attributes**: Have all properties of interval-scaled attributes, with the addition of a true zero indicating absence of the quantity. Values can be added, subtracted, multiplied, and divided. Example: weight in kilograms.

Major Tasks in Data Preprocessing
• **Data Cleaning:** Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
• **Data Integration:** Integration of multiple databases, or files
• **Data Transformation:** Normalization and aggregation
• **Data Reduction:** Obtains reduced representation in volume but produces the same or similar analytical results
• **Data Discretization** (for numerical data)

Data Cleaning: also known as **data cleansing** or **data preprocessing**, is a crucial step in the data science pipeline that involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability. Data cleaning is essential because raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it.

Python Collections to Store the Data
List: ordered and changeable. Allows duplicate members. e.g. [1, 2, 3, 4, 5]
Tuple: ordered and unchangeable Allows duplicate members. e.g. (1, 2, 3, 4, 5)
Set: unchangeable and unindexed. No duplicate members. e.g. {1, 2, 3, 4, 5}
Dictionary: ordered and changeable. No duplicate members. e.g. {1: “a”, 2: “b”, 3: “c”, 4: “d”, 5: “e”}

Data integration is the process of combining data from different sources to provide a unified view, making it accessible and useful for analysis and decision-making. This involves merging datasets from various databases, applications, and systems, ensuring consistency and coherence across the integrated data. Data integration techniques include ETL , data warehousing, and data virtualization, among others. Effective data integration addresses issues such as data redundancy, discrepancies, and fragmentation, enabling organizations to harness comprehensive and accurate insights from their data.
Redundancy Analysis
Redundancy analysis involves identifying and addressing redundant data that often arise when integrating multiple databases. Redundant data can occur when the same attribute or object has different names across databases, a situation known as object identification. Additionally, redundancy can appear when one attribute in a dataset is a derived attribute in another table, such as when annual revenue is calculated from monthly revenues.
Correlation Analysis
Correlation analysis is a statistical method used to measure and describe the association between random variables, helping to predict one quantity based on another. While correlation can suggest the presence of a causal relationship, it does not confirm causality. It serves as a foundational tool in many modeling techniques, providing a basic measure of association strength. Various methods exist to calculate the correlation coefficient, each capturing different types of relationships between variables.

Data Transformation
Data transformation is the process of converting data from one format or structure into another, facilitating its compatibility and usability across different systems and applications. This process often involves data cleaning, normalization, aggregation, and enrichment to ensure the data is accurate, consistent, and in a suitable format for analysis or integration. Transformation can include converting data types, standardizing values, and creating derived attributes. Effective data transformation enhances data quality and interoperability, enabling more efficient data analysis, improving insights, and supporting better decision-making. It is a critical step in data processing workflows, such as ETL (Extract, Transform, Load), which prepares data for storage in data warehouses or for use in advanced analytics and machine learning models.

1) Normalization by Decimal Scaling
Decimal scaling normalization is a data transformation method where the decimal point of attribute values is shifted based on the maximum value in the dataset. This shift ensures that all values are within a certain range, making comparisons and analyses more manageable. For example, if ages range from 25 to 70, dividing all ages by 10 yields values like 7.0 and 3.6, simplifying comparisons and analysis. This would result in age values like 7.0, 3.6, and 2.5, making them easier to work with and interpret, especially when combined with other normalized attributes or datasets.

2) Data Transformation by min-max Normalization
Min-max normalization is a technique in data transformation that scales numeric data to a specific range, typically between 0 and 1. For instance, if we have a dataset of exam scores ranging from 0 to 100, min-max normalization would convert these scores into values between 0 and 1 based on their relation to the minimum and maximum scores in the dataset. For example, if a student's score was 80 out of 100, after min-max normalization, it might become 0.8. This method is useful for standardizing data across different scales, ensuring that smaller and larger values contribute equally to analyses like machine learning models without losing the relationships between the original data points.

3) Data Transformation by z-score Normalization
Z-score normalization, also known as zero-mean normalization, is a technique in data transformation that standardizes numerical data based on the mean and standard deviation of the dataset. This method is particularly useful when dealing with data where the actual minimum and maximum values are unknown, or when outliers may significantly affect min-max normalization. For example, if we have a dataset of exam scores with a mean of 70 and a standard deviation of 10, a score of 80 would have a z-score of 1 (indicating one standard deviation above the mean), while a score of 60 would have a z-score of -1.d

Data Reduction : refers to the process of reducing the volume or complexity of a dataset while retaining its meaningful information. This can involve techniques such as dimensionality reduction, where the number of variables or features is decreased, or sampling methods, where a subset of data is selected to represent the whole. Data reduction aims to improve efficiency in data storage, processing, and analysis, making it easier to work with large datasets and extract relevant insights.

Data Reduction Strategies

1) Data Cube Aggregation

Data cube aggregation involves summarizing data at multiple levels of granularity within a data cube, effectively reducing the size of the dataset to be managed. By referencing appropriate levels of aggregation, it ensures that queries regarding summarized information are efficiently answered using the data cube. This technique leverages the smallest representation capable of solving the given task, optimizing storage and processing resources. Data cube aggregation is especially useful in scenarios where quick access to aggregated data is crucial, such as in OLAP (Online Analytical Processing) systems.

• Attribute subset selection/Feature subset selection/feature creation: Irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

2) Dimensionality Reduction

Dimensionality reduction is the process of decreasing the number of random variables or attributes considered in a dataset, often through data encoding or transformations. Techniques like wavelet transforms and principal components analysis (PCA) project the original data onto a smaller space, yielding a compressed representation. This process includes attribute subset selection or feature creation, where irrelevant, weakly relevant, or redundant attributes are identified and removed. By simplifying the data structure, dimensionality reduction enhances computational efficiency and improves the performance of machine learning models.

3) Numerosity Reduction

Numerosity reduction techniques aim to decrease the original data volume by representing it in a more compact form. Both parametric and non-parametric methods are employed in this process. Parametric methods assume the data fits a particular model, estimate the model parameters, and store only these parameters, discarding the original data except for potential outliers. Non-parametric methods, which do not assume any specific model, include techniques such as histograms, clustering, and sampling. These methods effectively compress the data while preserving essential information, making it more manageable for analysis and storage

4) Data Compression

String compression • There are extensive theories and well-tuned algorithms • Typically lossless • But only limited manipulation is possible

Audio/video, image compression • Typically lossy compression, with progressive refinement • Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Clustering is a data reduction technique that involves grouping a dataset into clusters based on similarity and storing only the representative information for each cluster, such as centroids or cluster diameters. The quality of clusters is assessed by metrics like diameter (max distance within a cluster) or centroid distance (average distance from objects to cluster centers), making clustering effective when data is not widely dispersed.

Sampling is another data reduction method that simplifies large datasets by selecting representative subsets. Simple random sampling chooses items with equal probability, but it can perform poorly with skewed data distributions. Sampling without replacement removes selected items from consideration, while sampling with replacement keeps them in the pool. Stratified sampling partitions the dataset and draws samples proportionally from each partition, making it useful for skewed datasets where balanced representation is crucial.

5) Data Discretization

Discretization and Concept Hierarchies

1) Discretization

Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

Three types of attributes:

- Nominal — values from an unordered set
- Ordinal — values from an ordered set
- Continuous — real numbers

Discretization/Quantization:

divide the range of a continuous attribute into intervals Some classification algorithms only accept categorical attributes. Reduce data size by discretization Prepare for further analysis

2) Concept Hierarchies

Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

• Hierarchical and recursive decomposition using:
• Binning (data smoothing)
• Histogram analysis (numerosity reduction)
• Clustering analysis (numerosity reduction)
• Entropy-based discretization
• Segmentation by natural partitioning

UNIT – 2 Statistics

Type of Data

1) Raw or Primary data: when data collected having lot of unnecessary, irrelevant & un wanted information

2) Treated or Secondary data: when we treat & remove this unnecessary, irrelevant & un wanted information

3) Cooked data: when data collected not genuinely and is false and fictitious

4) Ungrouped data: when data presented or observed individually.

5) Grouped data: when we grouped the identical data by frequency.

Variable

A variable is something that can be changed, such as a characteristic or value. For example age, height, weight, blood pressure etc

Types of Variables

Independent variable: is typically the variable representing the value being manipulated or changed. • The independent variable is the cause. Its value is independent of other variables in study.

Dependent variable: is the observed result of the independent variable being manipulated. • The dependent variable is the effect. Its value depends on changes in the independent variable

Confounding variables: Are those that affect other variables in a way that produces spurious or distorted associations between two variables.

Types of measurement

1) Discrete: Quantitative data are called discrete if the sample space contains a finite or countably infinite number of values.

2) Continuous: Quantitative data are called continuous if the sample space contains an interval or continuous span of real numbers.

3) Nominal: Categorical variables. Numbers that are simply used as identifiers or names represent a nominal scale of measurement such as female vs. male.

4) Ordinal: An ordinal scale of measurement represents an ordered series of relationships or rank order. Likert-type scales (such as "On a scale of 1 to 10, with one being no pain and ten being high pain, how much pain are you in today?") represent ordinal data

5) Qualitative vs. Quantitative variables

–Qualitative variables: values are texts (e.g., Female, male), we also call them string variables.

–Quantitative variables: are numeric variables.

Population: any group of interest or any group that researchers want to learn more about. –Population parameters (unknown to us): characteristics of population

Sample: a group of individuals or data are drawn from population of interest. –Sample statistics: characteristics of sample

Descriptive & Inferential Statistics

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarize and describe data	Draw conclusions or predictions
Data Sample	Analyzes the entire dataset	Analyzes a sample of the data
Examples	Mean, Median, Range, Variance	Hypothesis testing, Regression
Scope	Focuses on data characteristics	Makes inferences about populations
Goal	Provides insights and simplifies data	Generalizes findings to a larger population
Assumptions	No assumptions about populations	Requires assumptions about populations
Common Use Cases	Data visualization, data exploration	Scientific research, hypothesis testing

Key Aspects of Descriptive Statistics:

• **Measures of Central Tendency:** Descriptive statistics include calculating the mean, median, and mode, which offer insights into the center of the data distribution.

• **Measures of Dispersion:** Variance, standard deviation, and range help us understand the spread or variability of the data.

• **Visualizations:** Creating graphs, histograms, bar charts, and pie charts visually represent the data's distribution and characteristics.

Key Aspects of Inferential Statistics:

• **Sampling Techniques:** Relies on carefully selecting representative samples from a population to make valid inferences.

• **Hypothesis Testing:** This process involves setting up hypotheses about population characteristics and using sample data to determine if these hypotheses are statistically significant.

• **Confidence Intervals:** These provide a range of values within which we're confident a population parameter lies based on sample data.

• **Regression Analysis:** Inferential statistics also encompass
techniques like regression analysis to model relationships between variables and predict outcomes.

Measures of Central Tendency

1) Mean
The mean, also known as the average, is a measure of central tendency in a dataset. It is calculated by summing up all the values in the dataset and then dividing by the total number of values. The formula for the mean
Mean = sum of all observations / number of all observations

2) Median:
The median is the middle value in a dataset when the values are arranged in ascending or descending order. If there is an odd number of values, the median is the middle value itself. If there is an even number of values, the median is the average of the two middle values. The median is useful because it is not influenced by extreme values (outliers) in the dataset.

3) Mode
The mode is the value or values that appear most frequently in a dataset. A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal). If no value repeats, the dataset is said to have no mode.

Dispersion (Variability): a measure of the spread of scores in a distribution
Variability commonly measured with the following:
– Range Diff. between highest and lowest values
– Inter Quartile Range (IQR): Range of the middle half of a distribution
– Standard deviation: Average distance from the mean
– Variance: Average of squared distances from the mean

4) Range
Range in statistics refers to the difference between the highest and lowest values in a dataset. It provides a measure of the spread or variability of the data. It is informative for data without outliers To calculate the range:
Range = Highest Value − Lowest Value

Quartiles: Quartiles are values that divide a dataset into four equal parts, each representing 25% of the data.
There are three quartiles:
First Quartile (Q1): The value that separates the lowest 25% of the data from the rest. It is also the 25th percentile.
Second Quartile (Q2): The median, which divides the data into two equal halves. It is also the 50th percentile.
Third Quartile (Q3): The value that separates the lowest 75% of the data from the highest 25%. It is also the 75th percentile.
5) Inter Quartile Range (IQR)
Interquartile range (IQR) = Value of third quartile (Q3) – Value of first quartile (Q1)

Outliers
Outliers are data points that significantly differ from the rest of the dataset, often indicating potential errors, anomalies, or rare occurrences. They can skew statistical analyses and affect the accuracy of models if not properly addressed.

When there are no outliers in a sample, the mean and standard deviation are used to summarize a typical value and the variability in the sample, respectively.
When there are outliers in a sample, the median and interquartile range are used to summarize a typical value and the variability in the sample, respectively.

Tukey's fences offer a systematic approach to identifying outliers based on the distribution of the data, helping to ensure robust and accurate statistical analyses
Tukey's fences are based on the Interquartile Range (IQR) and are calculated as follows:
Calculate the IQR: IQR = Q3 - Q1 (where Q3 is the third quartile and Q1 is the first quartile).
Lower Fence: Q1 - 1.5 * IQR
Any data point below this lower fence is considered an outlier.
Upper Fence: Q3 + 1.5 * IQR
Any data point above this upper fence is considered an outlier.

BOX – PLOT – refer (50 -61)
A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It displays key statistical measures such as the median, quartiles (Q1 and Q3), and potential outliers in a compact and informative way. Here's how a box plot is constructed and calculated:

1. **Median (Q2)**: The middle value of the dataset, represented by a line inside the box.
2. **First Quartile (Q1)**: The median of the lower half of the dataset, forming the lower boundary of the box.
3. **Third Quartile (Q3)**: The median of the upper half of the dataset, forming the upper boundary of the box.
4. **Interquartile Range (IQR)**: The range between Q1 and Q3, represented by the height of the box.
5. **Whiskers**: Lines extending from the top and bottom of the box to indicate the range of non-outlier data points.

- **Lower Whisker:** Typically extends to the lowest data point within 1.5 times the IQR below Q1, or to the minimum data point if no outliers are present.
- **Upper Whisker:** Extends to the highest data point within 1.5 times the IQR above Q3, or to the maximum data point if no outliers are present.

Outliers: Individual data points beyond the whiskers are considered potential outliers and are often marked separately on the plot.

6) Percentiles:
Values below which a given percentage of observations fall. Percentiles are statistical measures that divide a dataset into hundred equal parts, each representing 1% of the data. They are useful for understanding the distribution of values within a dataset and identifying specific points relative to the entire dataset
Formula: p th percentile:
- p percent of observations below it
- (100 - p)% above it.
Example: Like 95% of CAT percentile means 95% are below it and 5% are above it

7) Standard Deviation
A measure of the amount of variation or dispersion of a set of values. Standard deviation is a measure of the dispersion or variability of a set of data points around their mean (average) value. It indicates how spread out the values in a dataset are from the mean, providing insight into the degree of consistency or deviation within the dataset

8) Variance
A measure of how spread out the values in a dataset are. Variance is a statistical measure that quantifies the dispersion or spread of a set of data points around their mean (average) value. It is the average of the squared differences between each data point and the mean.

Standard normal distribution or Z distribution
The standard normal distribution, also known as the Z distribution, is a specific type of normal distribution with a mean (average) of 0 and a standard deviation of 1. It is a fundamental concept in statistics and probability theory and serves as a reference for many statistical analyses.
Z Score
$z = (x – \mu) / \sigma$
– x observation
– μ mean
– σ standard deviation

Characteristics of normal distribution
▪ The normal distribution is mathematically defined
▪ The normal distribution is theoretical.
▪ The mean, median, and mode are all the same values at the center of the distribution.
▪ The normal distribution is symmetrical.
▪ The form of a normal distribution is determined by its mean and standard deviation.
▪ Standard deviation can be any positive value.
▪ The total area under the curve is equal to 1.
▪ The tails of normal distribution are always approaching to x-axis, but never touch it.

Histograms
Histograms are graphical representations of the distribution of a dataset. They display the frequencies or counts of data points falling within certain intervals or bins. Histograms are widely used in statistics and data analysis to visualize the distribution of continuous or discrete numerical data.
Bins or Intervals: The range of values is divided into intervals or bins along the horizontal axis (x-axis).
Frequency or Count: The vertical axis (y-axis) represents the frequency or count of data points falling within each bin.
Bars: Each bin is represented by a bar whose height corresponds to the frequency or count of data points in that
No Gaps: There are no gaps between the bars in a histogram, as it represents continuous data.
Area Under the Curve: The area of each bar is proportional to the frequency or count of data points in the corresponding interval.

Normalization:

- **Purpose**: The goal of normalization is to rescale the data so that it falls within a specific range, typically between 0 and 1. This process is helpful when the features (variables) in the dataset have different scales or units.
- **Method**: The most common normalization technique is Min-Max normalization, which transforms each data point xx to a new value $x'x'$ using the formula

- Effect: Normalization preserves the relative relationships and proportions between data points, but it may be sensitive to outliers, especially when using the Min-Max technique.

Standardization:

- **Purpose**: Standardization aims to center the data around a mean of 0 and a standard deviation of 1. It is particularly useful when the features in the dataset have different scales and follow a normal distribution or when performing certain statistical analyses like regression.
- **Method**: The standardization process involves subtracting the mean ($\mu\mu$) from each data point and then dividing by the standard deviation ($\sigma\sigma$). The formula

- Effect: Standardization results in a distribution with a mean of 0 and a standard deviation of 1, making it easier to compare and interpret the relative importance of different features in the dataset. It is less affected by outliers compared to normalization.

9) Skewness

A measure of the asymmetry of the distribution of values.
• Skewness is a number that indicates to what extent a variable is asymmetrically distributed.
• It is the degree of distortion from the symmetrical bell curve or the normal distribution.
• A symmetrical distribution will have a skewness of 0.

Symmetric Distribution (No Skew):
In a symmetric distribution, the data is evenly distributed around the mean, and the skewness value is close to 0. The tails on the left and right sides of the distribution are balanced.

Negative Skew (Left Skew): Mode >= Median >=Mean
In a negatively skewed distribution, the tail on the left side of the distribution is longer or stretched out, indicating more low values. The mean is less than the median, and the skewness value is negative.
Example: Income distribution in a country where a few people have very high incomes (long left tail).

Positive Skew (Right Skew): Mean >= Median >= Mode
In a positively skewed distribution, the tail on the right side of the distribution is longer or stretched out, indicating more high values. The mean is greater than the median, and the skewness value is positive.
Example: Test scores distribution where most students score well but a few score very low (long right tail).

10) Kurtosis

Kurtosis is a statistical measure that quantifies the degree of peakedness or flatness of the distribution of data points in a dataset. It assesses whether the data is more or less peaked than a normal distribution and can provide insights into the presence of outliers or extreme values in the dataset.

• **High kurtosis** in a data set is an indicator that data has heavy tails or outliers. investigate why do we have so many outliers.
• **Low kurtosis** in a data set is an indicator that data has light tails or lack of outliers need to investigate and trim the dataset of unwanted results
• **Mesokurtic:** This distribution has kurtosis statistic similar to that of the normal distribution (Kurtosis = 0).
• **Leptokurtic (Kurtosis > 3):** Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or more outliers.
• **Platykurtic: (Kurtosis < 3):** Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are lighttailed or lack of outliers.

11) Correlation

Correlation is a statistical measure that quantifies the relationship or association between two variables in a dataset. It indicates the extent to which changes in one variable are associated with changes in another variable. Correlation describes strength of association between two variables

• Falls between -1 and +1, with sign indicating direction of association (formula & other details later)

• The larger the correlation in absolute value, the stronger the association (in terms of a straight line trend)

Formula:

12) Covariance

13) Regression

Regression analysis is a statistical method used to model the relationship between a dependent variable (response variable) and one or more independent variables (predictor variables).

Types of Regression:

Simple Linear Regression: Involves one independent variable to predict the dependent variable.

Multiple Linear Regression: Includes multiple independent variables to predict the dependent variable.

Polynomial Regression: Fits a curve to the data by using polynomial functions.

Logistic Regression: Used for binary classification problems where the dependent variable is categorical.

Covariance and Correlation

Aspect	Covariance	Correlation
Definition	Measures directional association.	Standardized linear relationship.
Range of Values	-∞ to +∞ (unbounded)	-1 to +1 (bounded)
Standardization	Not standardized	Standardized
Strength Indicator	Directional association only	Indicates strength of relationship
Application	Used in analysis, scale dependent	Widely used, scale independent

Quartile Vs Quantile

Quartiles

- **First quartile:** Also known as Q1 (the number halfway between the lowest number and the middle number).
- **Second quartile**: Also known as Q2 or the median (the middle number halfway between the lowest number and the highest number).
- **Third quartile:** Also known as Q3, or the upper quartile (the number halfway between the middle number and the highest number).

Quantile

- Are values that split sorted data or a probability distribution into equal parts (In general, q-quantile divides sorted data into q parts).
- A quartile is a type of quantile.

▪ Quartiles (4-quantiles): Three quartiles split the data into four parts.

▪ Deciles (10-quantiles): Nine deciles split the data into 10

▪ Percentiles (100-quantiles): 99 percentiles split the data into 100 parts.

- There is always one fewer quantile than there are parts created by the quantiles

Quantile-Quantile Plots (QQ Plots)

The quantile-quantile (q-q plot) plot is a **graphical method for determining if a dataset follows a certain probability distribution** or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

Quantile-Quantile Plots (QQ Plots)

1.Collect the Data: Gather the dataset for which you want to create the Q-Q plot. Ensure that the data are numerical and represent a random sample from the population of interest.

2.Sort the Data: Arrange the data in either ascending or descending order. This step is essential for computing quantiles accurately.

3.Choose a Theoretical Distribution: Determine the theoretical distribution against which you want to compare your dataset. Common choices include the normal distribution, exponential distribution, or any other distribution that fits your data well.

4.Calculate Theoretical Quantiles: Compute the quantiles for the chosen theoretical distribution. For example, if you're comparing against a normal distribution, you would use the inverse cumulative distribution function (CDF) of the normal distribution to find the expected quantiles.

5.Plotting:

Plot the sorted dataset values on the x-axis.

Plot the corresponding theoretical quantiles on the y-axis.

Each data point (x, y) represents a pair of observed and expected values.

Connect the data points to visually inspect the relationship between the dataset and the theoretical distribution

Comparing the three visual techniques
Histograms
• Advantages: With properly-sized bins, histograms can summarize any shape of the data (modes, skew, quantiles, outliers)
• Disadvantages: Difficult to compare side-byside (takes up too much space in a plot) Depending on the size of the bins, interpretation may be different

Boxplots
Advantages:
Can identify whether the data came from a certain distribution.
Don't have to tweak with "graphical" parameters (i.e. bin size in histograms)
Summarize quantiles
Disadvantages:
Difficult to compare side by-side
Difficult to distinguish skews, modes, and outliers

QQ Plots
Advantages:
– Don't have to tweak with "graphical" parameters (i.e. bin size in histograms)
– Summarize skew, quantiles, and outliers
– Can compare several measurements side-by-side
Disadvantages:
– Cannot distinguish modes

Binomial distribution
• Binomial distribution is a type of discrete probability distribution representing probabilities of different values of the binomial random variable (X) in repeated independent N trials in an experiment.
• Thus, in an experiment comprising of tossing a coin 10 times (n), the binomial random variable (number of heads represented as successes) could take the value of 0-10.
• The binomial probability distribution is the probability distribution representing the probabilities of a random variable taking the value of 0-10
• **The necessary conditions and criteria to use binomial distributions:**
• **Rule 1:** Situation where there are only two possible mutually exclusive outcomes (for example, yes/no survey questions).
• **Rule2:** A fixed number of repeated experiments and trials are conducted (the process must have a clearly defined number of trials).
• **Rule 3:** All trials are identical and independent (identical means every trial must be performed the same way as the others; independent means that the result of one trial does not affect the results of the other subsequent trials).
• **Rule: 4:** The probability of success is the same in every one of the trials.

Poisson distribution
- The Poisson distribution is a type of discrete probability distribution used to model the number of events occurring within a fixed interval of time or space, given the average rate of occurrence ($\lambda\lambda$).
- For example, in a Poisson process where events occur randomly but at a constant average rate of 5 events per hour (λ=5λ=5), the Poisson random variable (X) could take values 0, 1, 2, 3, and so on, representing the number of events in that interval.
- The Poisson probability distribution represents the probabilities of X taking different values based on the Poisson parameter ($\lambda\lambda$).
- **Conditions for Poisson Distribution**:
    - Rule 1: Events occur randomly and independently of each other.
    - Rule 2: The average rate of occurrence ($\lambda\lambda$) is constant throughout the interval.

Rule 3: The probability of more than one event occurring in an infinitesimally small interval is negligible

Confidence Interval
Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval. • Your desired confidence level is usually one minus the alpha (a) value you used in your statistical test: • Confidence level = 1 − a

Central Limit Theorem
• The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger, no matter what the shape of the data distribution. • An essential component of the Central Limit Theorem is the average of sample means will be the population mean. • Similarly, if you find the average of all of the standard deviations in your sample, you will find the actual standard deviation for your population. • Mean of sample is same as the mean of the population. • The standard deviation of the sample is equal to the standard deviation of the population divided by the square root of the sample size. • Central limit theorem is applicable for sufficiently large sample sizes (n ≥ 30). The formula for central limit theorem can be stated as follows:

UNIT- 3

Machine Learning

Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. The goal of machine learning is to create systems that can automatically learn and improve from experience without being explicitly programmed for every task

Machine Learning Types:

1) Supervised Learning
   - Housing Price Prediction
   - Medical Imaging
2) Unsupervised Learning
   - Customer Segemntation
   - Market Basket Analysis
3) Semi-Supervised Learning
   - Text Classification
   - Lane-finding on GPS data
4) Reinforcement Learning
   - Optimized Marketing
   - Driverless Cars

Regression:

• Simple Linear Regression Algorithm
• Multivariate Regression Algorithm
• Decision Tree Algorithm
• Lasso Regression

Classification:

• Naïve Bayes Classifier Algorithm
• Random Forest Algorithm
• Decision Tree Algorithm
• Logistic Regression Algorithm
• Support Vector Machine Algorithm

Clustering

•K-Means Clustering algorithm
•Mean-shift algorithm
•DBSCAN Algorithm
•Principal Component Analysis
•Independent Component Analysis

Supervised Learning:

Supervised learning is a type of machine learning where the algorithm learns from labeled training data. This means each input data point is paired with a corresponding correct output or target variable. The objective is to train the algorithm to learn a mapping function from inputs to outputs, which can then be used to predict or classify new, unseen data accurately. Common tasks in supervised learning include regression, where the output is a continuous value (e.g., predicting house prices), and classification, where the output is a discrete label (e.g., classifying emails as spam or not spam). By iterating through the training data and adjusting its parameters, the algorithm minimizes the error between its predictions and the actual outcomes, effectively learning from the examples provided.

Example: Suppose you have a dataset of housing prices with features like size, location, number of bedrooms, and the corresponding sale prices. In supervised learning, you would use this labeled data to train a model to predict the price of a house given its features.

Types of Supervised Learning Algorithms:

Classification: In classification tasks, the algorithm predicts a discrete label or category. For example, predicting whether an email is spam or not spam based on its content.

Regression: Regression tasks involve predicting a continuous value. For instance, predicting the price of a house based on its features.

Unsupervised Learning:

Unsupervised learning deals with unlabeled data, where the algorithm attempts to identify patterns, structures, or relationships within the data without explicit guidance or pre-defined categories. The primary goal is to explore the data and extract meaningful insights or groupings that were not previously known. Common tasks in unsupervised learning include clustering, where the data points are grouped into clusters based on similarity (e.g., customer segmentation in marketing), and dimensionality reduction, where the data is simplified while preserving its essential features (e.g., reducing the number of variables in a dataset while retaining significant information). Unlike supervised learning, unsupervised learning does not rely on labeled training data, making it particularly useful for exploratory data analysis and discovering hidden structures within large datasets.

Example: Consider a dataset containing customer purchase histories without any labels. In unsupervised learning, you might use clustering algorithms to group similar customers together based on their purchasing behavior.

Types of Unsupervised Learning Algorithms:

Clustering: Clustering algorithms group similar data points together into clusters. K-means clustering is a popular technique used for this purpose.

Dimensionality Reduction: These algorithms aim to reduce the number of features in a dataset while preserving important information. Principal Component Analysis (PCA) is a widely used dimensionality reduction technique.

Linear Regression
Linear regression is a statistical method used in machine learning and statistics to model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the dependent variable, represented by a straight line. The basic concept of linear regression involves fitting a line to a set of data points to predict the value of the dependent variable based on the values of the independent variables. The equation of a simple linear regression model can be written as:
y=mx+b
Where:

- y is the dependent variable (target).
- x is the independent variable (feature).
- m is the slope of the line, representing the relationship between xx and yy.
- b is the y-intercept, indicating the value of yy when xx is 0.

Types of Linear Regression:
Simple Linear Regression: Involves one independent variable (feature) and one dependent variable (target).

Multiple Linear Regression: Deals with multiple independent variables to predict a single dependent variable.

Classification
In classification, the goal is to build a model that can accurately assign class labels to instances based on their features. The model learns the patterns and relationships in the training data and uses this knowledge to make predictions on new data.
Types of Classification Algorithms:
Logistic Regression: Although named "regression," logistic regression is a classification algorithm used for binary classification tasks. It predicts the probability that an instance belongs to a particular class.
Decision Trees: Decision trees partition the feature space into regions and assign a class label to each region based on the majority class of training instances within that region
Support Vector Machines (SVM): SVMs find the optimal hyperplane that separates classes in the feature space, maximizing the margin between classes
Random Forest: A random forest is an ensemble technique that combines multiple decision trees to improve classification accuracy and robustness.
Naive Bayes: Naive Bayes classifiers are based on Bayes' theorem and assume that features are independent given the class. They are particularly effective for text classification tasks.
Neural Networks: Deep learning models such as neural networks can also be used for classification tasks, especially when dealing with complex, high-dimensional data.

Bayes' theorem
Bayes' theorem is a fundamental concept in probability theory and statistics, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the probability of an event based on new evidence or information. The theorem is often used in machine learning, especially in Bayesian inference and Bayesian networks, where it plays a key role in making predictions and decisions under uncertainty.

Where:

- $P(A|B)P(A|B)$ is the probability of event AA occurring given that event BB has occurred (posterior probability).
- $P(B|A)P(B|A)$ is the probability of event BB occurring given that event AA has occurred (likelihood).
- $P(A)P(A)$ is the prior probability of event AA occurring.
- $P(B)P(B)$ is the prior probability of event BB occurring.

Naive Bayes Classification
Naive Bayes is a popular classification algorithm in machine learning, particularly suited for text classification tasks such as spam detection, sentiment analysis, and document categorization. It's based on Bayes' theorem, assuming independence among features, hence the term "naive." The Naive Bayes classifier assumes that the presence of a particular feature in a class is independent of the presence of other features. Despite this simplifying assumption, Naive Bayes often performs well, especially with large datasets and when the independence assumption approximately holds.
Example:
Naive Bayes for email classification works like this: Imagine you have a few emails, some are spam, and some are not. You look at these emails to see how often certain words like "free" or "buy" appear and how long the emails are. Then, using this information, when a new email arrives, you check if it has words like "free" or "buy" and how long it is. Based on how often these features occur in spam or non-spam emails in your data, you make a guess about whether the new email is likely to be spam or not. It's like making an educated guess based on patterns

Classification in Machine Learning

• Types of predictive models in machine learning are:

Binary Classification for Machine Learning:

• The most popular algorithms which are used for binary classification are :

– K-Nearest Neighbours (Also supports multiple labels)
– Logistic Regression (Only for two labels)
– Support Vector Machine (Only for two labels)
– Decision Trees (Also supports multiple labels)
– Naive Bayes (Also supports multiple labels)

Multi-Class Classification

• These types of classification problems have no fixed two labels but can have any number of labels.

Multi-Label Classification for Machine Learning

• Here, we refer to those specific classification tasks where we need to assign two or more specific class labels that could be predicted for each example.

• The main difference with multi-class is the ability to predict multiple labels and not just one

Imbalanced Classification

• An Imbalanced Classification refers to those tasks where the number of examples in each of the classes are unequally distributed.

Generally, imbalanced classification tasks are binary classification jobs where a major portion of the training dataset is of the normal class type and a minority of them belong to the abnormal class.

Clustering

Clustering is an unsupervised machine learning technique that groups similar data points into clusters based on their characteristics, aiming to uncover inherent patterns and structures within the data without predefined labels. This method is useful for exploratory data analysis, helping to identify natural groupings that can inform decision-making across various fields. Popular algorithms like k-means, hierarchical clustering, and DBSCAN analyze datasets to form these clusters by measuring the similarity between data points. Clustering has diverse applications, such as customer segmentation for targeted marketing, image segmentation in computer vision, and pattern recognition in bioinformatics, making it a versatile tool for revealing hidden insights and organizing complex datasets.

For example, in healthcare, clustering can be used to group patients with similar medical histories or symptoms, enabling healthcare providers to identify patterns in disease progression or response to treatments. This can lead to more personalized and effective treatment plans, improve patient outcomes, and assist in early detection of health issues. By analyzing patient data through clustering, medical researchers can also uncover new insights into the relationships between different health conditions, facilitating advancements in medical research and public health strategies.

K-Means Clustering Algorithm:

K-means is a clustering algorithm in machine learning that aims to partition a set of data points into K clusters. It works by iteratively assigning each data point to the nearest centroid (center) of a cluster and then recalculating the centroids based on the new assignments. The goal is to minimize the sum of squared distances between data points and their respective cluster centroids, creating clusters where data points within the same cluster are similar to each other and dissimilar to those in other clusters. K-means is an unsupervised learning algorithm commonly used for data exploration, pattern recognition, and segmentation tasks.

K-Means Steps:

1. Choose the number of clusters (K) and initialize centroids randomly or using K-means++.
2. Assign data points to the nearest centroid based on distance, updating assignments until convergence.
3. Update centroids by computing the mean of data points in each cluster, repeating until convergence.
4. Convergence occurs when assignments and centroids stabilize or change minimally.
5. Output final cluster assignments and centroids as groups of similar data points.

Elbow Method:

The Elbow method is a technique used to determine the optimal number of clusters KK for a dataset in K-means clustering. It helps find the "elbow point" in a plot of the within-cluster sum of squares (WCSS) versus the number of clusters KK. Here's how it works:

1. WCSS Calculation:

Run the K-means algorithm for a range of KK values, typically from 1 to a maximum number based on domain knowledge or trial and error.

For each KK, calculate the within-cluster sum of squares (WCSS), which measures the sum of squared distances of data points to their cluster centroids.

2. Elbow Point Identification:

Plot a graph of KK versus WCSS, where KK is on the x-axis and WCSS is on the y-axis.

Look for the "elbow point" on the graph, which is the point where the rate of decrease in WCSS slows down significantly (forming an elbow-like shape).

3. Optimal KK Selection:

The optimal number of clusters KK is typically chosen at the elbow point, as it represents a balance between minimizing WCSS (better clustering) and avoiding overfitting (too many clusters).

UNIT - 4

Data Visualization

Data Visualization techniques involve the generation of graphical or pictorial representation of DATA, form which leads you to understand the insight of a given data set. This visualization technique aims to identify the Patterns, Trends, Correlations, and Outliers of data sets.

Benefits of Data Visualization

Patterns in business operations: Data visualization techniques help us to determine the patterns of business operations. By understanding the problem statement and identifying the solutions in terms of pattering and applied to eliminate one or more of the inherent problems.

Identify business trends and relate to data: These techniques help us identify market trends by collecting the data on Day-To-Day business activities and preparing trend reports, which helps track the business how influences the market. So that we could understand the competitors and customers. Certainly, this helps to long-term perspective.

Storytelling and Decision making: Knowledge of storytelling from available data is one of the niche skills for business communication, specifically for the Data Science domain which is playing a vital role. Using best visualization this role can be enhanced much better way and reaching the objectives of business problems.

1) Line Chart

A line chart is a type of graph that displays data points connected by straight line segments. It is commonly used to visualize trends over time, such as stock prices, temperature changes, or sales figures. The x-axis typically represents time or some other continuous variable, while the y-axis represents the value being measured. Line charts are useful for identifying patterns, trends, and fluctuations in data over time.

2) Histogram

A histogram is a graphical representation of the distribution of numerical data. It consists of a series of bars, where each bar represents the frequency or count of data points falling within a specific range, known as a bin. Histograms are particularly useful for visualizing the distribution of continuous data, such as exam scores, ages of people, or product prices. They help to understand the central tendency, variability, and shape of the data distribution.

3) Pie Chart

A pie chart is a circular graph divided into sectors, where each sector represents a proportion or percentage of the whole dataset. Pie charts are commonly used to show the composition or distribution of categorical data, such as market share, budget allocation, or survey responses. Each sector's size is proportional to the quantity it represents, making it easy to compare relative sizes of categories within the dataset.

4) Scatter plots

A scatter plot is a two-dimensional graph that uses dots to represent individual data points. It is used to visualize the relationship between two variables, typically one plotted on the x-axis and the other on the y-axis. Scatter plots are useful for identifying patterns, correlations, clusters, or outliers in the data. They are especially effective for exploring relationships between continuous variables, such as height versus weight or temperature versus humidity.

5) Hexbins plots

A hexbin plot is a variation of a scatter plot that uses hexagonal bins to represent the density of data points in a two-dimensional space. Each hexagon's color or intensity indicates the number of data points it contains, helping to visualize areas of high or low density. Hexbin plots are useful for handling large datasets and identifying patterns or clusters that may not be obvious in a traditional scatter plot.

6) Heatmap

A heatmap is a graphical representation of data where values in a matrix are represented as colors. It is commonly used to visualize correlations, distributions, or relationships in a dataset. Heatmaps are particularly effective for highlighting patterns, trends, or anomalies in large datasets. They are commonly used in fields such as biology (gene expression), finance (stock correlations), and social sciences (survey responses).

7) Boxplot

A boxplot, also known as a box-and-whisker plot, is a graphical representation of the distribution of numerical data through quartiles. It consists of a box that represents the interquartile range (IQR) of the data, with a line inside the box indicating the median. Whiskers extend from the box to show the range of the data, excluding outliers. Boxplots are useful for identifying central tendency, variability, and outliers in the data distribution.

8) Pairplot

A pairplot is a grid of scatter plots and histograms that visualizes pairwise relationships between variables in a dataset. It displays scatter plots of each pair of variables along the diagonal and histograms on the off-diagonal cells. Pairplots are useful for exploring correlations, distributions, and patterns between multiple variables simultaneously. They are commonly used in exploratory data analysis (EDA) to gain insights into data relationships.

9) Bar Chart

A bar chart is a graphical representation of data using rectangular bars of varying lengths. It is used to compare categories or show changes over time. The length of each bar represents the value it represents, making it easy to compare quantities across different categories. Bar charts are effective for visualizing categorical data, such as sales by product category, population by country, or survey responses by option.