FDS(SEM V)

a) List any two application of Data Science.

--> 1.*Healthcare: Data science is used to analyze medical records, identify disease patterns, and develop personalized treatment plans.*

2.Finance: Data science is used to detect fraud, assess credit risk, and optimize investment strategies.

b) What is outlier?

--> *An outlier is a data point that differs significantly from other observations in a dataset. It can be either extremely high or low compared to the rest of the data. Outliers can be caused by measurement errors, data entry errors, or genuine anomalies in the data.*

c) What is missing values?

--> *Missing values refer to data points that are absent or incomplete in a dataset. They can occur due to various reasons like data entry errors, equipment failures, or missing responses in surveys. Missing values can significantly impact data analysis and modeling, so it's crucial to handle them appropriately.*

d) Define variance.

--> Variance is a statistical measure that quantifies the dispersion or spread of data points from their mean. It calculates the average squared difference between each data point and the mean. A higher variance indicates greater variability in the data,while a lower variance indicates less variability.

e) What is nominal attribute?

--> *Nominal attributes are categorical data where the values represent different categories or labels without any inherent order or ranking. Examples include gender, color, or country.*

f) What is data transformation?

--> *Data transformation involves converting raw data into a suitable format for analysis. This includes techniques like normalization, standardization, discretization, and feature engineering to improve data quality, handle missing values, and extract meaningful insights.*

g) What is one hot coding?

--> One-hot encoding is a technique used to convert categorical data into numerical data. It creates a new binary feature for each category, assigning a value of 1 to the corresponding category and 0 to others. This allows machine learning algorithms to process categorical data effectively.

h) What is the use of Bubble plot?

--> A bubble plot is a type of chart used to visualize data points as bubbles. The size of each bubble represents the magnitude of a third variable, while the x and y axes represent two other variables. Bubble plots are useful for visualizing relationships between three variables simultaneously, making it easier to identify patterns and trends.

i) Define Data visualisation.

--> *Data visualization is the process of representing data graphically to make it easier to understand and interpret. It involves creating visual representations of data, such as charts, graphs, and maps, to highlight patterns, trends, and anomalies. Effective data visualization can help people make better decisions, identify opportunities, and solve problems.*

j) Define Standard deviation?

--> Standard deviation is a statistical measure that quantifies the dispersion or spread of data points from their mean. It measures how much the data points deviate from the average value. A higher standard deviation indicates greater variability in the data, while a lower standard deviation indicates that the data points are clustered closer to the mean.

a) Define volume characteristic of data in reference to data science.

--> Volume in the context of data science refers to the sheer size and quantity of data being generated and stored.As technology advances, the volume of data generated by various sources (e.g., social media, IoT devices, scientific experiments) is rapidly increasing. This massive volume of data presents both challenges and opportunities for data scientists, requiring specialized tools and techniques to store, process, and analyze it effectively.

b) Give examples of semistructured data.

--> *<book>*

<title>The Lord of the Rings</title>

<author>J.R.R. Tolkien</author>

<genre>Fantasy</genre>

</book>

Semistructured Data is data that doesn't conform to a rigid, predefined data model. It has a partial structure, often using tags or markers to delimit data elements. Examples include XML, JSON, and HTML. While it lacks the strict structure of relational databases, it offers flexibility for representing complex information.

c) Define Data Discretization.

--> Data discretization, also known as binning or quantization, is the process of converting continuous numerical data into discrete intervals or bins. This technique is often used to simplify data analysis, improve data quality, and reduce the dimensionality of data. By grouping similar values together, discretization can help in identifying patterns, trends, and outliers in the data.

d) What is a quartile?

--> A quartile is a statistical measure that divides a dataset into four equal parts. There are three quartiles:

1.First Quartile (Q1): Divides the lowest 25% of the data.

2.Second Quartile (Q2): Also known as the median, divides the lowest 50% of the data.

3.Third Quartile (Q3): Divides the lowest 75% of the data.

Quartiles are used to understand the distribution and variability of data.

e) List different types of attributes.

--> Attributes can be broadly categorized into two types:

Categorical Attributes:

Nominal: No inherent order (e.g., color, gender).

Ordinal: Has a natural order (e.g., low, medium, high).

Numerical Attributes:

Discrete: Countable values (e.g., number of children).

Continuous: Infinitely many possible values (e.g., height, weight).

f) Define Data object.

--> A data object is a structured collection of data elements, often represented as key-value pairs. It can be used to store and organize information about a person, product, event, or any other entity. Data objects are commonly used in various data formats like JSON, XML, and CSV.

g) What is Data Transformation?

--> Data transformation is the process of converting raw data into a suitable format for analysis. It involves techniques like normalization, standardization, discretization, and feature engineering to improve data quality, handle missing values, and extract meaningful insights.

h) Write the tools used for geospatial data.

***-->** Some popular tools for geospatial data analysis and visualization include:*

1.QGIS: A free and open-source GIS software.

2.ArcGIS: A powerful commercial GIS software.

3.Google Earth Engine: A cloud-based platform for planetary-scale geospatial analysis.

4.Python libraries: Geopandas, Shapely, Folium, and others.

i) State the methods of feature selection.

***-->** There are several methods for feature selection in machine learning:*

1.Filter Methods: Statistical measures like correlation, chi-square test, and information gain are used to rank features.

2.Wrapper Methods: Algorithms like forward selection, backward elimination, and recursive feature elimination evaluate subsets of features.

3.Embedded Methods: Feature selection is integrated into the model building process, such as regularization techniques like L1 and L2 regularization.

j) List any two libraries used in Python for data analysis.

***-->** Two popular Python libraries for data analysis are:*

1.Pandas:

Provides high-performance, easy-to-use data structures and data analysis tools.

Used for data manipulation, cleaning, analysis, and visualization.

2.NumPy:

Offers efficient numerical computations and array operations.

Provides fundamental building blocks for scientific computing and data science.

a) What is Data science?

--> Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

b) Define Data source?

--> A data source is the origin of data, such as databases, files, APIs, or real-time streams. It provides the raw material for data analysis and processing. Data sources can be structured (e.g., relational databases), semi-structured (e.g., JSON, XML), or unstructured (e.g., text, images).

d) List the visualization libraries in python.

--> *Here are some popular Python libraries for data visualization:*

1.Matplotlib: Versatile library for creating static, animated, and interactive visualizations.

2.Seaborn: High-level data visualization library built on top of Matplotlib, providing a more attractive and informative style.

3.Plotly: Interactive visualization library for creating dynamic and shareable plots.

4.Bokeh: Interactive visualization library for creating web-based visualizations.

5.Altair: Declarative statistical visualization library based on Python.

e) List applications of data science.

--> *Data science has applications across various industries:*

Healthcare: Disease prediction, drug discovery, personalized medicine

Finance: Fraud detection, risk assessment, algorithmic trading

Marketing: Customer segmentation, targeted advertising, sentiment analysis

Retail: Recommendation systems, inventory management, demand forecasting

E-commerce: Personalized product recommendations, customer behavior analysis.

g) Define Hypothesis Testing?

--> Hypothesis testing is a statistical method used to determine whether a hypothesis about a population parameter is likely to be true or false. It involves collecting sample data, calculating test statistics, and comparing them to a critical value or p-value to make a decision.

h) What is use of Bubble plot?

--> *A bubble plot is a type of chart used to visualize data points as bubbles. The size of each bubble represents the magnitude of a third variable, while the x and y axes represent two other variables. Bubble plots are useful for visualizing relationships between three variables simultaneously, making it easier to identify patterns and trends. For example, you could use a bubble plot to visualize the relationship between GDP, population, and life expectancy for different countries.*

i) Define Data cleaning?

--> *Data cleaning is the process of detecting and correcting errors and inconsistencies in data. It involves tasks like handling missing values, removing duplicates, formatting data, and identifying outliers. Clean data is essential for accurate and reliable data analysis.*

a) List the tools for data scientist.

--> *Data scientists utilize a variety of tools throughout their workflow. Here are some key categories:*

1.Programming Languages: Python (dominant), R (statistics), SQL (databases)

2.Data Analysis Libraries: Pandas (manipulation), NumPy (numerical computing)

3.Machine Learning Libraries: Scikit-learn (popular algorithms), TensorFlow/PyTorch (deep learning)

4.Visualization Libraries: Matplotlib, Seaborn (creation of charts)

5.Version Control Systems: Git (collaboration and tracking changes)

6.Cloud Platforms: AWS SageMaker, Google Cloud AI Platform (scalable computing)

7.Data Wrangling Tools: OpenRefine (interactive data cleaning)

b) Define statistical data analysis?

--> Statistical data analysis involves applying statistical methods to collect, organize, analyze, interpret, and present data. It helps in understanding data patterns, making inferences, and drawing conclusions. Statistical techniques include descriptive statistics (mean, median, mode, standard deviation), inferential statistics (hypothesis testing, confidence intervals), and exploratory data analysis (visualization, summary statistics).

c) What is data cube?

--> A data cube is a multidimensional data structure that organizes data along multiple dimensions, such as time, location, and product category. It allows for efficient data analysis and

reporting by enabling users to slice and dice the data along different dimensions to answer specific questions. Data cubes are commonly used in business intelligence and data warehousing applications.

d) Give the purpose of data preprocessing?

--> *Data preprocessing is a crucial step in data mining and machine learning. It involves cleaning, transforming, and preparing raw data to improve its quality and suitability for analysis. The main purposes of data preprocessing include:*

1.Handling missing values: Imputing missing values or removing records with missing data.

2.Noise reduction: Identifying and removing noise or outliers.

3,Data integration: Combining data from multiple sources.

4,Data transformation: Normalization, standardization, and feature engineering.

5,Data reduction: Dimensionality reduction and feature selection.

a) Explain any two ways in which data is stored in files.

--> 1. Text-based Files:

Data is stored as plain text characters.

Common formats: CSV, TSV, JSON, XML

Simple to read and write but less efficient for large datasets.

2. Binary Files:

Data is stored in binary format, which is more efficient for storing large amounts of data.

Common formats: Databases, images, audio, video

Requires specific software or libraries to read and write.

b) Explain role of statistics in data science.

--> *Statistics plays a crucial role in data science by providing the tools and techniques to analyze, interpret, and draw meaningful insights from data. It helps in:*

Data exploration and cleaning: *Identifying patterns, anomalies, and missing values.*

Feature engineering: *Creating new features from existing ones.*

Model building and evaluation: *Selecting appropriate models, training, and evaluating their performance.*

Hypothesis testing: *Making inferences about the population based on sample data.*

Data visualization: *Creating informative visualizations to communicate findings.*

c) Explain two methods of data cleaning for missing values.

-->

1. Deletion:

Listwise deletion: Removes entire records with missing values.

Pairwise deletion: Excludes cases with missing values only for specific analyses.

Simple to implement but can lead to loss of information.

2. Imputation:

Mean/Median/Mode Imputation: Replaces missing values with the mean, median, or mode of the respective variable.

Regression Imputation: Predicts missing values using regression models.

Hot Deck Imputation: Replaces missing values with values from similar records.

d) Explain any two tools in data scientist tool box.

--> 1. Python:

Versatile programming language for data analysis, machine learning, and data visualization.

Popular libraries: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn

2. SQL:

Essential for working with relational databases.

Used for data extraction, transformation, and loading (ETL) processes.

Enables querying and manipulating large datasets.

e) Write a short note on wordclouds.

--> A word cloud is a visual representation of text data where words are displayed in different sizes, with larger words representing more frequent terms. It's a useful tool for quickly identifying the most important keywords or themes within a text document or corpus. Word clouds are often used

in text analysis, natural language processing, and information visualization. By visually highlighting the most prominent words, word clouds can help users gain insights into the underlying topics and sentiments of the text.

a) Differentiate structured and Unstructured Data.

-->

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF(Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than structured data	It is more scalable.

Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

b) What is inferential statistics?

--> JSON code,

{

"name": "Alice",

"age": 30,

"city": "New York"

}

Inferential statistics is a branch of statistics that involves drawing conclusions about a population based on a sample of data. It uses statistical tests and models to make inferences about the population parameters, such as the mean, standard deviation, or proportions. Common techniques include hypothesis testing, confidence intervals, and regression analysis.

e) What is visual encoding?

--> Visual encoding is the process of representing data visually using different visual elements, such as color, size, shape, and position. It helps in conveying information effectively and efficiently. For example, a bar chart uses the length of bars to represent numerical values, while a scatter plot uses the position of points to represent two numerical variables. By using appropriate visual encodings, data visualization can help uncover patterns, trends, and insights that might not be apparent from raw data alone.

a) Explain outlier detection methods in brief.

--> Outlier detection is the process of identifying data points that deviate significantly from the rest of the data. Here are some common methods:

1.Statistical Methods:

Z-score: Measures how many standard deviations a data point is from the mean.

Interquartile Range (IQR): Identifies outliers based on quartiles and the IQR.

2.Machine Learning Methods:

Isolation Forest: Isolates anomalies by randomly selecting features and splitting data.

Local Outlier Factor (LOF): Compares the density of a data point to its neighbors.

3.Visualization Techniques:

Box Plots: Visually identify outliers based on quartiles.

Scatter Plots: Spot outliers visually.

By detecting and handling outliers, you can improve the accuracy and reliability of your data analysis.

b) Write different data visualization libraries in python.

--> 1.Matplotlib: Versatile library for creating static, animated, and interactive visualizations.

2.Seaborn: High-level data visualization library built on top of Matplotlib, providing a more attractive and informative style.

3.Plotly: Interactive visualization library for creating dynamic and shareable plots.

4.Bokeh: Interactive visualization library for creating web-based visualizations.

5.Altair: Declarative statistical visualization library based on Python.

c) What is data cleaning? Explain any two data cleaning methods.

--> Data cleaning is the process of detecting and correcting errors and inconsistencies in data. It involves tasks like handling missing values, removing duplicates, formatting data, and identifying outliers.

Two common data cleaning methods:

1.Handling Missing Values:

Imputation: Replacing missing values with estimated values (e.g., mean, median, mode, or predicted values).

Deletion: Removing records with missing values, but this can lead to information loss.

2.Outlier Detection and Handling:

Statistical Methods: Using techniques like Z-score or IQR to identify outliers.

Visualization: Using box plots or scatter plots to visually identify outliers.

Handling Outliers: Removing, capping, or imputing outliers.

a) Explain data science life cycle with suitable diagram.

--> JSON code,

```
{
  "name": "Alice",
  "age": 30,
  "city": "New York"
}
```

The data science life cycle is a structured approach to solving data-driven problems. It typically involves the following steps:

1.Problem Definition: Clearly define the problem to be solved.

2.Data Acquisition: Gather relevant data from various sources.

3.Data Cleaning and Preparation: Clean and preprocess the data to remove errors and inconsistencies.

4.Exploratory Data Analysis (EDA): Explore the data to understand its characteristics and identify patterns.

5.Feature Engineering: Create new features or transform existing ones to improve model performance.

6.Model Building: Select and train appropriate machine learning models.

7.Model Evaluation: Assess the performance of the model using evaluation metrics.

8.Deployment: Deploy the model to a production environment.

9.Monitoring and Maintenance: Monitor the model's performance and retrain as needed.

b) Explain concept and use of data visualisation.

--> Data visualization is the process of representing data graphically to make it easier to understand and interpret. It involves creating visual representations of data, such as charts, graphs, and maps, to highlight patterns, trends, and anomalies.

Uses of data visualization:

Exploratory Data Analysis (EDA): To understand data distribution and relationships.

Communication: To convey insights and findings effectively to stakeholders.

Decision Making: To support informed decision-making.

Storytelling: To create compelling narratives from data.

c) Calculate the variance and standard deviation for the following data.

X : 14 9 13 16 25 7 12

--> *import numpy as np*

```
data = [14, 9, 13, 16, 25, 7, 12]

# Calculate the mean
mean = np.mean(data)

# Calculate the variance
variance = np.var(data)

# Calculate the standard deviation
std_dev = np.std(data)

print("Mean:", mean)
```

print("Variance:", variance)

print("Standard Deviation:", std_dev)

a) What are the measures of central tendency? Explain any two of them in

brief.

--> Measures of central tendency describe the central or typical value of a dataset. The three main measures are:

Mean: *The average of all data points.*

Calculated by summing all values and dividing by the number of values.

Sensitive to outliers.

Median: *The middle value when data is sorted.*

Not affected by outliers.

Better for skewed data.

Mode: *The most frequently occurring value.*

Can be used for both numerical and categorical data.

b) What are the various types of data available? Give example of each?

--> Data can be categorized into various types:

Numerical Data:

Discrete: Countable values (e.g., number of students, number of cars).

Continuous: Infinitely many possible values (e.g., height, weight, temperature).

Categorical Data:

1.Nominal: *No inherent order (e.g., color, gender).*

2.Ordinal: *Has a natural order (e.g., low, medium, high).*

3.Text Data: *Unstructured text (e.g., emails, social media posts, news articles).*

Image Data: *Visual information (e.g., photos, videos).*

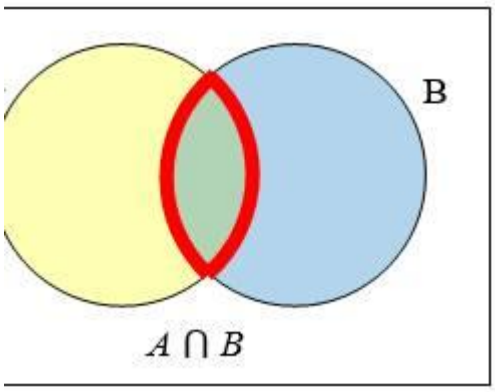
Audio Data: *Sound recordings (e.g., speech, music).*

c) What is venn diagram? How to create it? Explain with example.

--> A Venn diagram is a visual representation of sets and their relationships. It consists of overlapping circles, where each circle represents a set and the overlapping regions represent the intersection of sets.

Example:

Consider two sets: A = {1, 2, 3, 4} and B = {3, 4, 5, 6}.



The overlapping region represents the intersection of the two sets, which is {3, 4}. Venn diagrams are useful for understanding and visualizing set operations like union, intersection, and difference.

a) Explain different data formats in brief.

--> Data can be stored in various formats, each with its own advantages and disadvantages:

1. Text-based Formats:

CSV (Comma-Separated Values): Simple tabular format, easy to read and write.

JSON (JavaScript Object Notation): Flexible format for structured data, commonly used for web APIs.

XML (eXtensible Markup Language): Hierarchical format for structured data, often used for configuration files and data exchange.

2. Binary Formats:

Database Files: Efficiently store and retrieve large amounts of structured data.

Image Formats: Store visual information (e.g., JPEG, PNG, GIF).

Audio and Video Formats: Store sound and video data (e.g., MP3, WAV, MP4).

3. Specialized Formats:

Excel: *Spreadsheet format for data analysis and visualization.*

PDF: *Portable document format for sharing documents***.**

Word: *Document format for text and formatting.*

The choice of data format depends on factors like the type of data, the intended use, and the desired level of structure and flexibility.

b) What is data quality? Which factors are affected data qualities?

--> *Data quality refers to the accuracy, completeness, consistency, and timeliness of data. High-quality data is essential for reliable data analysis and decision-making.*

Factors affecting data quality include:

Accuracy: Data should be free from errors and inconsistencies.

Completeness: Data should be complete and not contain missing values.

Consistency: Data should be consistent across different sources and formats.

Timeliness: Data should be up-to-date and relevant.

Relevance: Data should be relevant to the specific analysis or task.

Validity: Data should be accurate and meaningful.

Poor data quality can lead to incorrect insights, biased models, and bad decisions. Therefore, data cleaning and preprocessing are crucial steps in any data analysis project.

a) Write a short note on hypothesis testing.

--> Hypothesis testing is a statistical method used to determine whether a claim or hypothesis about a population parameter is true or false. It involves formulating a null hypothesis (H_0) and an alternative hypothesis (H_1).

Steps in Hypothesis Testing:

State the Hypotheses: Define the null and alternative hypotheses.

Set the Significance Level: Choose a significance level (α) to determine the risk of rejecting a true null hypothesis.

Collect Data: Gather a sample of data.

Calculate the Test Statistic: Calculate a test statistic based on the sample data.

Determine the P-value: Calculate the probability of obtaining the observed test statistic or a more extreme value under the null hypothesis.

a) Explain 3V's of Data Science.

--> The 3Vs of Big Data refer to the characteristics that define large and complex datasets:

Volume: The sheer amount of data generated and stored. As technology advances, the volume of data continues to grow exponentially.

Velocity: The speed at which data is generated and processed. Real-time data streams from IoT devices, social media, and other sources require rapid analysis.

Variety: The diverse types and formats of data, including structured, semi-structured, and unstructured data. This diversity presents challenges in data integration and analysis.

Understanding and addressing the 3Vs is crucial for effectively handling and extracting insights from big data.

b) Explain data cube aggregation method in detail.

--> Data cube aggregation involves summarizing data across multiple dimensions to create a more concise and informative representation. This is often used in data warehousing and business intelligence applications.

Methods of Aggregation:

Roll-up: Aggregating data from a lower level of detail to a higher level. For example, summing up sales data for individual products to get total sales by product category.

Drill-down: Navigating from a higher level of detail to a lower level. For example, drilling down from total sales by region to sales by individual store within a region.

Slice and Dice: Selecting specific subsets of data by applying filters to one or more dimensions. For example, filtering data by a specific time period or product category.

c) Explain any two data transformation technique in detail.

--> Data transformation is the process of converting raw data into a suitable format for analysis. Two common techniques are:

<u>Normalization:</u>

Scales numerical data to a specific range (e.g., 0 to 1 or -1 to 1).

Helps in improving the performance of machine learning algorithms.

Techniques include min-max scaling and z-score normalization.

<u>Discretization:</u>

Converts continuous numerical data into discrete intervals or bins.

Reduces the number of values, simplifies analysis, and can improve model performance.

Methods include equal-width binning, equal-frequency binning, and clustering-based binning.

a) Write a short note on feature extraction.

-> {

"name": "Alice",

"age": 30,

"city": "New York"

}

Feature extraction is the process of selecting and transforming relevant features from raw data to improve the performance of machine learning models. It involves identifying the most informative characteristics of the data that contribute to the prediction or classification task.

Key Techniques:

Feature Selection: *Choosing a subset of the most relevant features.*

Feature Engineering: *Creating new features from existing ones.*

Dimensionality Reduction: *Reducing the number of features using techniques like Principal Component Analysis (PCA) or t-SNE.*

By effectively extracting and engineering features, you can enhance the accuracy and efficiency of machine learning models.

b) Explain Exploratory Data Analysis (EDA) in detail.

--> JSON code,

```
{
  "name": "Alice",
  "age": 30,
  "city": "New York"
}
```

Exploratory Data Analysis (EDA) is an essential step in the data science pipeline. It involves understanding the data through statistical summaries, visualizations, and other techniques. The goals of EDA are to:

1.Understand the Data:

Identify the data types (numerical, categorical)

Check for missing values and outliers

Examine the distribution of variables

2.Discover Patterns:

Identify trends, correlations, and relationships between variables

Find clusters and anomalies

3.Prepare for Modeling:

Transform and clean the data

Select relevant features for modeling

Key Techniques:

Univariate Analysis: Analyzing individual variables.

Summary statistics (mean, median, mode, standard deviation)

Histograms, box plots, and density plots

Bivariate Analysis: Analyzing the relationship between two variables.

Scatter plots, correlation matrices, cross-tabulations

Multivariate Analysis: Analyzing the relationship between multiple variables.

Principal Component Analysis (PCA), Factor Analysis

By performing EDA, data scientists can gain valuable insights into the data, make informed decisions about data cleaning and preprocessing, and select appropriate modeling techniques.

Date: / /

Q.1) Attempt any eight of the following. 4

a) List any 2 application of Data science.
→ i) Gamming world
ii) Helth core sector.

b) What is outlier? 1M
→ The outlier is an observation point that is distant from other ~~outlier~~ observation.

c) What is missing value?
→ some values in the data may not be filled up for various reasons & hence are considered missing value

d) Define variance
→ variance is the measure of dispersion that is related to the standard deviation. it is calculated by finding the square of the standard deviation of given data distribution.

e) What is Data transformation?
→ Data transformation the process of converting cleaning, structuring data into a usable format that can be analyzed support decision making process to the growth an organization.

e) What is Nominal attribute?
→ Nominal means relating to name's the value of nominal attribute are symbol or names of things.

g) what is one hot coding.
→ one hot coding is one method of converting data to prepare it for an algorithm & get a better predication.

h) What is the use of Bubble plot?
→ A bubble plot is scatter plot where a third dimension is added the value of an additional numeric variable is represented through the size of the dots.

i) Define Data visulation
→ Data visulization is the presentation of data ingraphical format. Data visulatation is generic term used which describes any atempt to help understanding of data by providing visual representation.

j) Define standard deviation.
→ A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean

Q.2) Attempt any four the following
a) Differentiate structure & unstructured data
→ i) Structure Data :- i) structure data as name suggest type data is well organized. ii) structured data is data the depends on a data model & resides in a fixed field within a record.
ii) Unstructure Data :-
i) unstructure data is data that is not organized in a pre-defined manner does not have pre-defined data model.
ii) unstructured data has internal structure but not structured via pre-defined data model or schema.

b)
→

Q.4)
a)
→

b) What is inferential statistics.

→ In inferential statistics, we make an inference from a sample about the population. The main aim of inferential statistics is to draw some conclusions from the sample generalize them for the population data. Statistical inference mainly deals with two different kind of problems of hypothesis testing & estimation of parameter values.

c) What do you mean by data preprocessing?

→ It is the task of transforming raw data to be ready to be fed into an algorithm it is a time consuming yet important step that cannot be avoided for the accuracy of result in data analysis.

d) Define Data discretization?

→ i) Data discretization is the process of converting continuous data into an a set of discrete inference or categorious. ii) this technique is used to for data reduction data simplification, or to make the data suitable format. for analysis & it is typically used for very large datasets.

e) What is visual encoding.

→ i) The visual encoding is the way in which data is mapped into visual structure, upon which we build the images on the screen.

ii) encoding in data visulaization means transforming the data into a visual element on a chart or map through position, shape, size, symbol & colour.

Q.3) Attempt any two of the following

a) Explain outlier detection methods in brief.

→ The outlier detection methods can be divided into supervised method, supervised method & unsupervised method.

i) Supervised method :-

a) Supervised method model data normality & abnormality.

b) Domain ~~expt~~ experts examine & label a sample of underlying data.

2] Unsupervised Data :-

a) In same application scenarios labelled as "normal" or "outlier" are not available. Thus an ~~Un~~unsupervised learning method has to be used.

b) Unsupervised outlier detection methods like an implicit assumption, the normal object are some what "clustered".

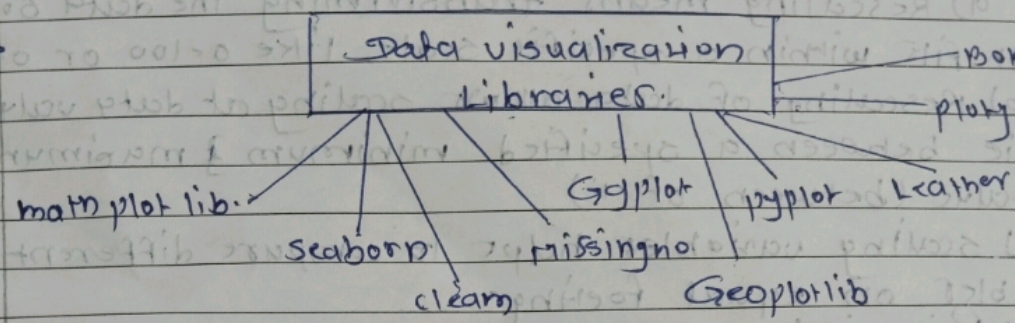
3] Semi supervised method :-

a) Semi-supervised outlier detection, method were developed to take such scenarios.

b) Building a model for outlier based on only a few labbeled outlier is unlikely to be effective.

b) Write different data visualization libraries in python.

→



Q.4) Attempt any two of the following

a) Explain 3V's of Data science.

→ i) The 3V's are (volume, velocity, variety)

ii) Due to the expansion of data of the turn of the 21st century evolved by the so-called 3V's of data science which are valuror, velocity variety.

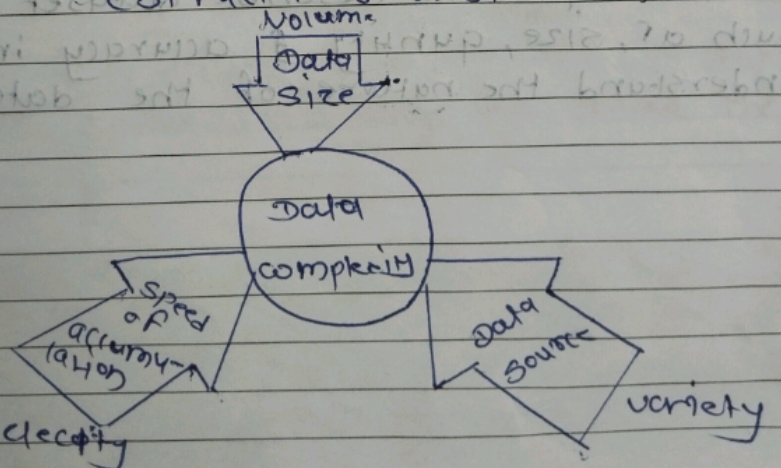
iii) Volume refers to the increasing size of data, velocity the speed of at which data is acquired & velocity the diverse types of data that the are available.

iv) The 3V's are explained below -

a) velocity :- The speed at which data is accumulated.

b) volume :- The size of the scope of the data.

c) variety :- The massive array of data & types (structured unstructured).



Page No.:
Date: / /

Q.1) Attempt any eight of the following

a) Define volume characteristic of data in reference to data science.

→ volume refers to sheer scale of data that is being connected & analyzed.
characteristics → Large datasets Technological impact.

b) Give example of semistructured data.

→ ① markup language XML
② open standard JSON (Javascript object notation)
③ NO SQL

c) Define data discretization.

→ Data discretization is characterized as a method of translating attribute values of continuous data into a finite set of intervals with nominal information loss.

d) What is quartile?

→ A quartile is statistical term that refers to division dataset into equal parts each containing 25% of data point when data arrange in ascending order.

f) Define data object.

→ A data object is a collection of attributes that together describe an entity or instance in dataset.

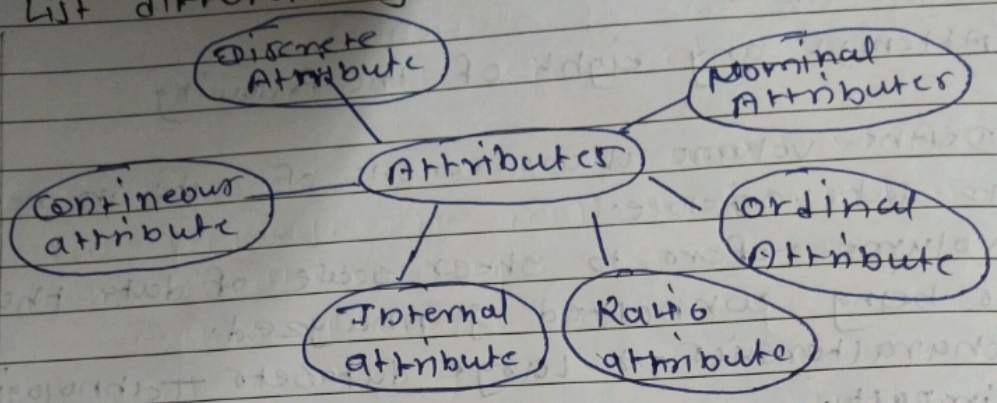
Page No.:
Date: / /

5) Explain any two data transformation technique in detail
→ i) Rescaling :-
a) Rescaling means transforming the data so that it fits within a specific scale, like 0-100 or 0-1
b) Rescaling of data allows scaling at data value to lie between a specified minimum & maximum value (between of 1)
c) scaling variables helps to compare different variables on equal footing.
ii) Binarizing :-
a) It is the process of converting data to either 0 or 1 based on a ther threshold value.
b) All data value above the threshold value of are marked where as all the data value equal to or below the threshold value are marked as 0.

Q.5: Attempt any one of the following
a) Explain Exploratory Data analysis (EDA) in details
→ i) Exploratory Data Analysis :- an analysis approach that identifies general pattern in the data.
ii) These patterns include outliers & feature of the data that might be unexpected.
iii) EDA is an import first step in any data.
iv) EDA is a crucial initial step in data science project
v) Data Analysis use data visulization & statistical techniques to describe dataset characterization such as, size, quality & accuracy in order to better understand the nature of the data.

e) List different type of attributes.

→



g) What is data transformation.?

→ Data transformation is the process of converting row data into a format or structure that would be more suitable.

h) Write the tools used for geospatial data.

→ ① GIS software.
② programming libraries & framework
③ spatial dataset
④ Remote sending tools.
⑤ visualization & mapping tools.
⑥ Geospatial data format and tools
⑦ web-based geospatial formats.

Q.4) Attempt any four of the following.

a) Explain any two ways in which data is stored in files.

→ ① Flat files
- format - Data is stored in next format usually structured or semi-structured such as CSV, TSV, JSON

② Relational Databases -
Data is stored in a more structured format typically in relational database management.

Date: / /

b) Explain role of statistical in data science.

→ - some roles

① Predictions & classification
statistical help in prediction & classification of data whether it would be correct or client by either previous usage.

② Help to create probability distribution and estimation is crucial in understanding basics of machine learning & logistic regressions.

③ cross validation & other techniques
they are also in heavy statistical tools that have been brought into machine learning data analytics world.

d) Explain any tools in the data science.

→ ① Python programming.

- python refers various libraries designed explicity for data science operations.
- It was found in 1990 by Guido van Rossum
- The rich set of libraries are core strengths.

② Tableau public
- Tableau is data visualization software public which has it free version named as, Tableau
- It data visualization software packet with powerfull graphics to make interactive visualization

e) Write a short note on word clouds.

→ A word ~~colu~~ cloud is a word visualization that displays the most used word in a text from small to large.

-How often it appears.

- a word cloud tag or cloud that is a visual representation of textual data that present a list of words.

- most often used for aesthetic purpose at depict categorical data.

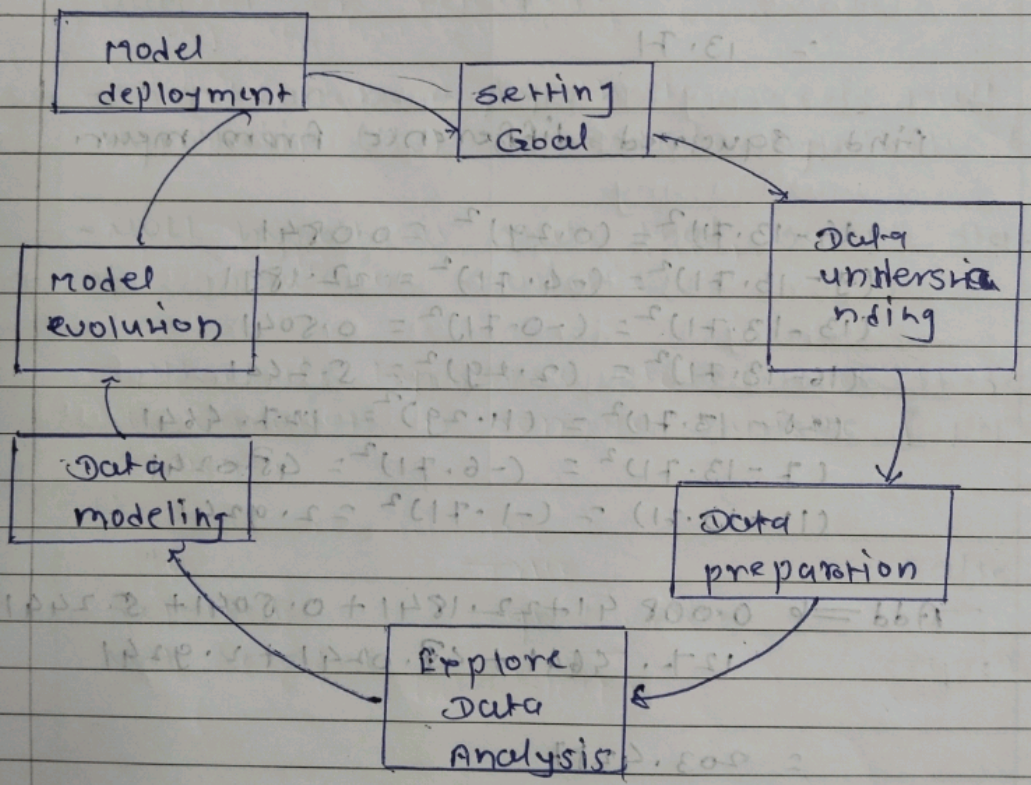
Media
social

Page No.:
Date:

Q.3) Attempt any two of the following

a) Explain data science life cycle with suitable diagram.

→ - Life ~~science~~ cycle of data science outlierr phase from start to finish
- provides framework for best performancer of each phase from creation to irs. completion



- Setting Goal :- Entire circle revolver around the business or research goal
- Data understanding :- Involves collection of available data.
- Data prep → include selecting of relavan data.
- Exploratory data analysis :- Involver getting idea.

about solution of factor affecting building model

- Data modeling :- It heart of data analysis.

c) calculated variance & standard deviation. for the following data.

x : 14 9 13 16 25 7 12

→ mean = $\frac{14+9+13+16+25+7+12}{7}$

= 13.71

find squared difference from mean

$(14-13.71)^2 = (0.29)^2 = 0.0841$

$(9-13.71)^2 = (-4.71)^2 = 22.1891$

$(13-13.71)^2 = (-0.71)^2 = 0.5041$

$(16-13.71)^2 = (2.29)^2 = 5.2441$

$(25-13.71)^2 = (11.29)^2 = 127.4641$

$(7-13.71)^2 = (-6.71)^2 = 45.0241$

$(12-13.71) = (-1.71)^2 = 2.9241$

Add ⇒ 0.0841 + 22.1841 + 0.5041 + 5.2441 + 127.4641 + 45.0241 + 2.9241

= 203.4287

= $\frac{203.4287}{7}$ = 29.0612

variance = 29.0612

standard deviation = $\sqrt{29.0612}$

= 5.39

Q.4. A

a) W

→ =T

h

s

-

4

a

-

-

Page No.:
Date: / /

Q.4. Attempt any two of the following.

a) Write a short note on hypothesis testing.

→ The process of determine whether stated hypothesis is accepted or rejected from sample data is called hypothesis testing.

- It is most inferential statistical techniques used to check whether a hypothesis is accepted or rejected.

- There can be 2 hypothesis namely null hypothesis (H_0) alternate hypothesis (H_a)

- NULL hypothesis state that sample statistic & population statistic.

- Alternate hypothesis states that there is variation in sample statistic & population statistic.

H_0	True	False
Rejected	Type I error	✗
Not Rejected	✓	Type II error

b) Difference between structure & unstructured data.

structured	unstructured.
- Data it well be organized	Data is not well organized
- Get organised by mean of relational Database	- Baled on simple character & binary data.
- Concarrency of data is present preferred in multitasking	- concurrency is not present.
- support RDB so versioning is done over rows & tuples table.	- data is possible only at on whole data as no support of database at all.

Q.5) Attempt any one of the following.

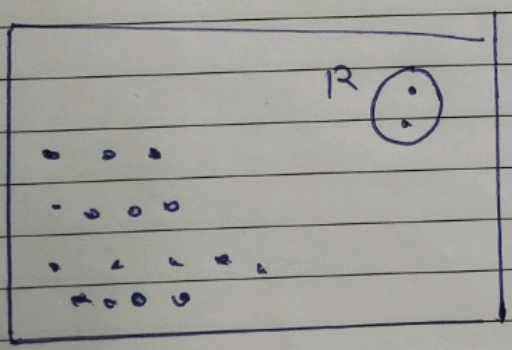
(i) Write any two application of data science.

→ (1) Image recognition & speech Recognition
(2) Gaming world.
(3) Internet search.
(4) Transport.

(ii) Explain any type of outlier in detail.

→ (1) Global Outlier

- If an individual data point can be considered ~~as~~ anomalous with respect the rest of data, then data is termed as point outlier.



- To detect global outlier critical issue is to find ~~the~~ on appropriate measurment of deviation with ~~ref~~ respect to application in question.