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".

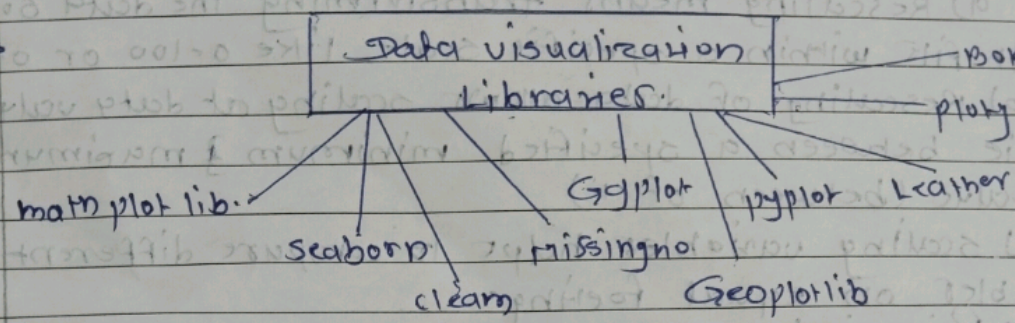
3] Semi supervised method :-

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.

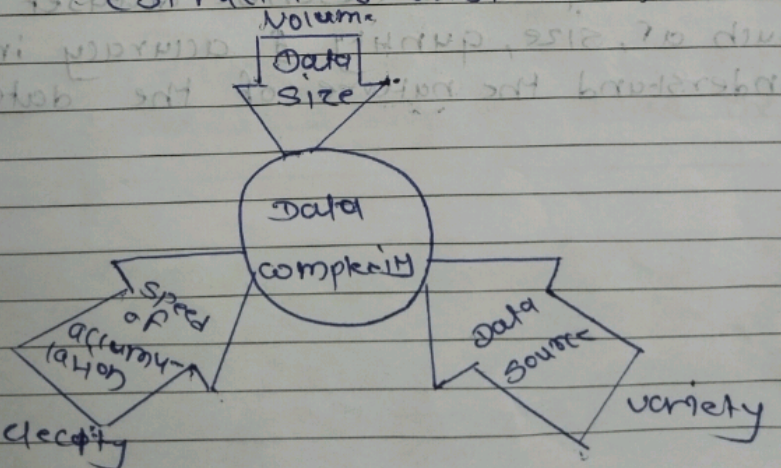
iii) Volume refers to the increasing size of data, velocity the speed of at which data is acquired & velocity the diverse types of data that the are available.

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: / /

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.