Page No.:
Date: / /

Q.1.) Attempt any eight of following.

(a) What do you mean by primary data?
→ a primary data is data that never collected ~~wayo~~ before & can be generated in variety of ways such as conducting interviewo schedules & so on.

(b) What do you mean by data quality?
→ Data quality is mean by 'the ability of a given data set to serve an intendd purpose

c) Define outlier.
→ An outlier is an observation point that is distandt from other observation

d) Define interquartile range.
→ - It is a measure of statistical dispersion which represents range within which central 50% of data value lie.
IQR = Q3 - Q1
Q1 → value below 25% of data falls
Q3 → The value below 75% of data falls

e) What do you mean by missing value?
→ values in data may not be filled up for various reasons & hence are considere missing.

f) What are uses of zip files?
→ - commonly used for emailing & sharing files over the internet.

Page No.:
Date: / /

g) What do you mean by XML files data format?
→ An XML is extensible markup language file designed to be human & machine readable & be used to store & transport data.

i) What is tag cloud?
→ A tag cloud is word visualization that displays most used words in texts from small to large according to how often each appears.

Q.2) Attempt any four of the following

a) Explain different application of data science.
→ (i) Finance → (1) Fraud detection → Data science is used to detect pattern in transactions.

(ii) sales & marketing → (1) Analyse customer data to segment the market for targeted marketing
(iii) E-commerce → (1) Dynamic pricing models adjust prices in real time based demand competition
(iv) predictive maintence → Analysis equipment data to predict failures & schedule maintenance.

(b) Explain NULL & alternate hypothesis.

→ NULL Hypothesis → NULL hypothesis state hypothesis that sample of statistic is equal to population statistic.
eg → there has been no diff found in ~~error~~ exam grade of students after coaching classes
Alternate Hypothesis → Alternate hypothesis states hypothesis there is variation in sample statistic & population.

eg → there has been significant grade improvement found in exam grade of students after conducting coaching classes.

c) What do you mean by noisy data? Explain any two causes of noisy data.

→ - Noisy data contains error or outliers.
eg → value of age attribute are within range 22-45 years where as are record, reflect the age attribute value as 80.

causes → ① Facility data collection instruction
② Data entry problems.

d) What do you mean by data visualization? Give example of any two data visualization libraries.

→ There data visualization is generic term used they which describes any attempt, to help understanding of data by providing visual representation.

- Libraries.

① matplotlib libraries.
- It is common standard python library used for plotting 2D data visualization
- mainly used for creating plots that zoomed in a section of plots & pan around plot wing toolbar.

(ii) seaborn libraries.
→ - seaborn library to create artistic charts with very few lines of code.
- User creative styles, rich color palettes which use violin plots Heatmaps time series plots.

Page No.:
Date: / /

Q.3 Attempt any two of the following

(a) Explain data cube aggregation method in context of data reduction.

→ - A data cube is multi-dimensional array of value. A data cube is generally used to easily interpret data.
- useful when representing data together with dimensions as certain measures of business requirment.

- Data cube aggregation is multi-dimensional aggregation which eases multidimensional analysis.

- Data cube present pre-computed & summarized data which eases the data mining into fast access.

b) What is mean, mode, median & range of the following list of values.

24, 29, 24, 25, 27, 25, 32, 24.

→ mean → $\frac{24+29+24+25+24+27+25+32+24}{9}$

$= \frac{234}{9}$

$= 26$

Median → The median is the middle value

since the are 9 number (odd number) the median is 5th number

Median = 25

Page No.:
Date: / /

Mode → value that appears most frequently in data

24 appears five times

mode = 24

① Histogram

- A histogram is graphical display of data using bars of diff height.
- To create histogram we need divide the range of value into a series of intervalues. & second count how many values fall into interval.

② Bar charts

- Bar charts are used for comparing quantities of diff categories or groups.
- value of category are represented by bars with vertical or horizontal bars
- Bar charts has gaps bet bars compare to histograms

③ Line plot

- It is 2-dimensional plotting of value
- value are displayed in scattered manner & connected.

line plot displays information as series of data points called "markers" connected by straight lines.

Q.4) Attempt any two of the following

a) Difference betn structure & unstructured

structured	unstructured
- This data is typically well organised.	- this data is typically non-organised.
- organised by mean of relational database	- organised by simple character & binary data.
- managment & concurrency of data are present	- no transaction managment & no concurrency are present.
- less flexible as well as less scalable	- more flexible as well as more scalable.

(b) What do you mean by data attributes? Explain types of attribute with example.

→ - An attribute is property or charaltersHic of an object. A data attribute is a single value descriptor for data object. ever color of person, name of student etc.

Page No.:
Date: / /

Types ① Nominal means "relating to names" Nominal attribute are symbols or names.

eg → Branch of relation of ATTRONICS company has attributes like BID & name which are Nominal type.

② Binary attribute → A binary attribute is a nominal attribute with only two categories 0 or 1 where 0 is absent, 1 is present

③ ordinal Attribute → an ordinal Attribute is attribute with possible value that have meaningful order successive value is not known.
e.g. → rankings (eg → taste of potato chips on scale from 1-10).
grades, height in (tall, medium, short).

④ Numeric Attributes → A numeric attribute is quantitative i.e. → is a measurable quantity represented in integer or real values.
- ① Interval scaled

⑪ Ratio scaled

Interval eg → temperature in celcius or fahrenheit etc.

Ratio eg → temperature in kelvin, length time, counts, etc.

Page No.:
Date: / /

Q.5.) Attempt any one of the following

(a) What do you mean by data transformation explain strategies of data transformation

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

① Rescaling

- Rescaling means transforming the data so that it fits within specific scale like 0-100 or 0-1
- Rescaling data allows scaling all data values to lie betn a specified min & max value.

	Col1	Col2	Col3
0	2	14	24
1	4	4	36
2	8	9	-13

scaled data

[[0.1 0.78
[0.33 0.09 1.]
[1. 0.55 0.]]

② Normalizing

- Measurment unit used can affect data analysis.
Normalizing scaled attribute data so as to fall within smaller range 0.0 to 1.0 or -1.0 to 1.0.

- Normalizing data attempt to give all attribute an equal weight.

(3) Binarizing → It is process of converting data to either 0 or 1 based on threshold value.

(4) Standarizing →
- It is also called as removal - it is process of transforming attributes having gaussion distribution with differing mean & standard deviation values into a standard Gaussion distribution with mean 0 & standard deviation 1

(5) Label encoding → It is process used to convert textual labels into numeric form in order to prepare it to be used readable form
eg → Gender → male → & female.

(6) One Hot coding →
- used to transform categorical variables into a format that can be binary.
- one hot coding refers to splitting the column which contains numerical categorical data to ~~any~~ many column.
depending on no. of categories present in column