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?

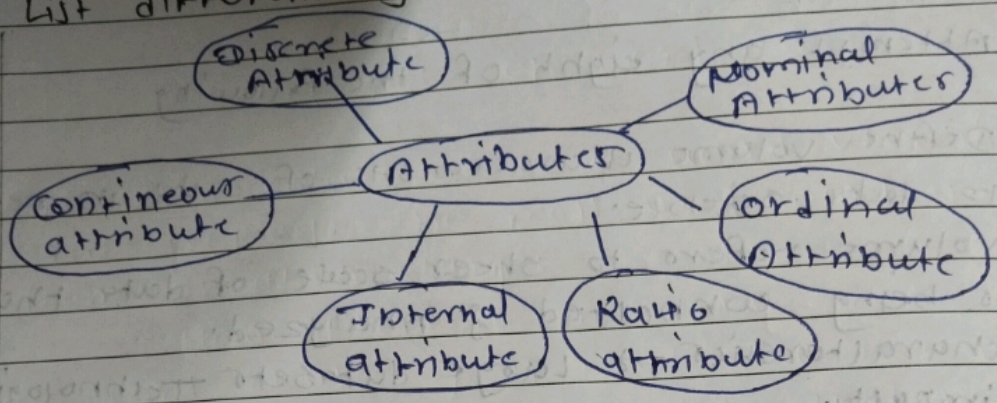
→ A quartile is statistical term that refers to division dataset into equal parts each containing 25% of data point when data arrange in ascending order.

f) Define data object.

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

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.

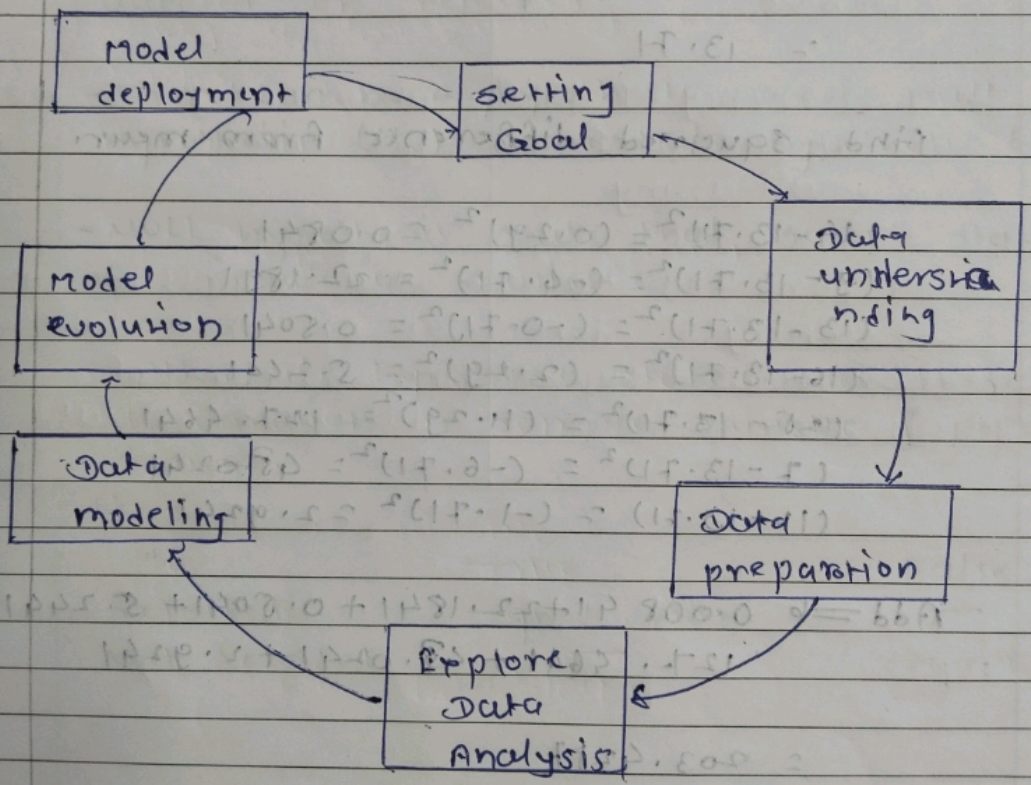
Media
social

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.

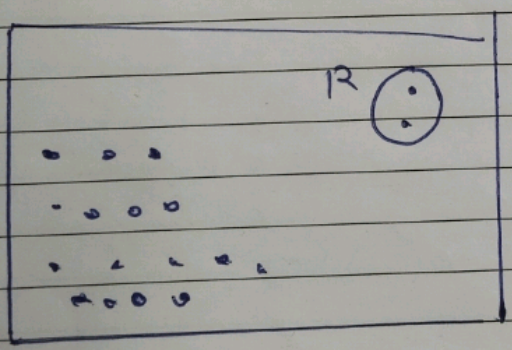
(i) Write any two application of data science.

→ (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.