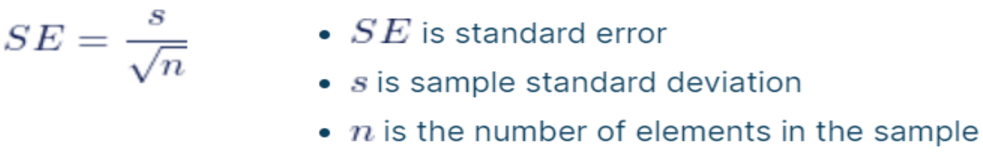
1	Answer all Questions

a)More data analysis will result into greater analytical accuracy explain in brief.

Ans:
More data for analysis will result into greater analytical accuracy and greater confidence in the decisions based on the analytical findings. This would entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time,and innovating on new products, new services and optimizing existing services

b)In a random sample of 200 students, the mean Data Analytic score is 550. The Standard deviation of the Data Analytic score is 180. What is the standard error for Data Analytic score.



$$SE = \frac{s}{\sqrt{n}}$$

- SE is standard error
- s is sample standard deviation
- n is the number of elements in the sample

Ans:12.78 (1 mark)

c)A website captures information about each customer's order. The total dollar amounts of the last 8 orders are listed below. Calculate average distance between each data value and the mean.12, 23, 31, 15, 26, 24, 16, 23

Ans.Find the mean absolute deviation (MAD).
Step 1: Calculate the mean.
Mean = 21.25
Step 2: Calculate the distance between each data point and the mean.

Data Point	Distance from mean
12	12-21 = 9
23	23-21 = 2
31	31-21 = 10
15	15-21 = 6
26	26-21 = 5
24	24-21 = 3
16	16-21 = 5
23	23-21 = 2

Step 3: Add the distances together.

9+2+10+6+5+3+5+2 = 42

Step 4: Divide the sum by the number of data points.

42/8 = 5.25(if process correct 0.5 mark and if both process and result correct 1 mark)

d)State the purpose of Map-reduce in Hadoop, with a suitable real life example.

Ans:MapReduce is a programming model used to perform distributed processing in parallel in a Hadoop cluster, which Makes Hadoop working so fast.When you are dealing with Big Data, serial processing is no more of any use.

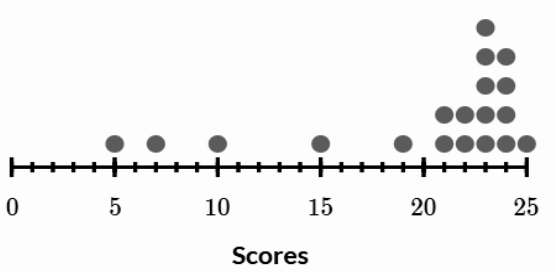
On a daily basis the micro-blogging site Twitter receives nearly 500 million tweets, i.e., 3000 tweets per second.If Twitter data is an input, and MapReduce performs the actions like **Tokenize, filter, count and aggregate counters**.(only purpose is explained 0.5 mark and example also given 1 mark)

e)What is the role of marginal's in descriptive measures for categorical variables?

Ans:Marginals show the total counts or percentages across columns or rows in a contingency table.

Q2.2.

(i)The distribution below shows the scores on a player's test for 19 applicants. Identify the range of both low & high outliers in the given scatter plot. Based on Q1, Q2, Q3 and IQR



(ii)Draw a box-plot for the above data with proper labelling.

Ans:

(i)For the data set {5,7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25}

The first quartile (Q1) is 19
The second quartile (Q2) is 23
The third quartile (Q3) is 24
IQR is (Q3-Q1)= 5

Low outliers: Q1−1.5*IQR=19−1.5(5)=19−7.5=11.5
High outliers: Q3+1.5*IQR=24+1.5(5)=24+7.5=31.5
(Process only correct 2 marks,process and answer both correct 3 mark)

(ii)Student will draw the box plot.(2 marks)

3.

(i) Enlist the different types of analytics present in data analytics with suitable example

(ii) State the difference between ETL and ELT.

Ans:All the four analytics:
(i) Descriptive
(ii) Diagnostic
(iii) Predictive
(iv) Prescriptive (definition with example 4 marks)

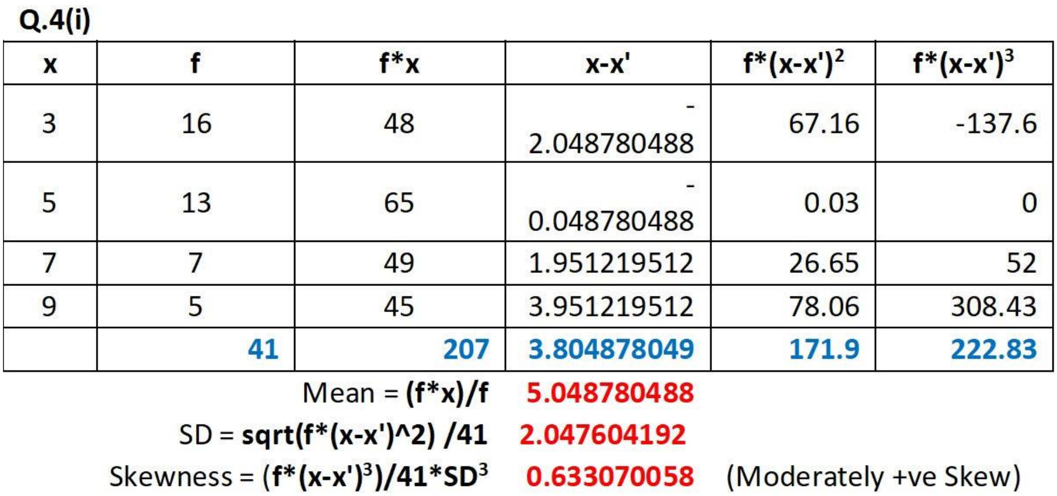
(ii)difference (1 marks)

4.(i) Design a group frequency table for a paediatric data having the following distribution of data, and calculate the Skewness from the data.

The age group of children are, (i). 2 to 4 age are 16 nos of children.
(ii). 4 to 6 age are 13 nos of children.
(iii). 6 to 8 age are 7 nos of children.
(iv). 8 to 10 age are 5 nos of children.

(ii)List down at least two differences between stratified and cluster sampling.

Ans:



Q.4(i)

x	f	f*x	x-x'	f*(x-x')^2	f*(x-x')^3
3	16	48	- 2.048780488	67.16	-137.6
5	13	65	- 0.048780488	0.03	0
7	7	49	1.951219512	26.65	52
9	5	45	3.951219512	78.06	308.43
	41	207	3.804878049	171.9	222.83

Mean = **(f*x)/f** 5.048780488

SD = **sqrt(f*(x-x')^2) /41** 2.047604192

Skewness = **(f*(x-x')^3)/41*SD^3** 0.633070058 (Moderately +ve Skew)

(Process and ans correct-3 marks)

(ii) Two differences-2 marks

5**. Refer to Table 1, which illustrates the fish data-set. With this data-set, establish a simple linear regression model for each species by estimating the weight of the fish from its length. By demonstrating a step-by-step procedure, the regression model has to capture the values of the slope and intercept for each species. In the data-set, the weight of a fish is expressed in grams, and its length and height are in centimeters.**

Species	**Weight**	**Length**	**Height**
Bream	242	25.4	11.52
Bream	290	26.3	12.48
Bream	340	26.5	12.73
Bream	363	29	12.75
Bream	500	29.7	13.65
Bream	1000	37	18.9
Roach	200	23.5	7.3
Roach	180	25.2	7.10
Roach	290	26	8.88
Roach	390	31.7	9.5
Roach	160	22.5	6.5
Roach	140	20.8	6.4
Roach	40	14.5	4.15

Ans:

Solution:

(i)The objective is to establish a simple linear regression model for each species by estimating the weight of the fish from its length. So, the dataset to be divided into Bream and Roach species respectively with dependent variable (DV) as weight and independent variable (IV) as length. Post to the split, two regression model to be established i.e., one of specie and the other for roach species.

Both models are presented in the following form:

Weight$_{\text{bream}}$ = a$_{\text{bream}}$ + b$_{\text{bream}}$ $*$ Length$_{\text{bream}}$ and Weight$_{\text{roach}}$ = a$_{\text{roach}}$ + b$_{\text{roach}}$ $*$ Length$_{\text{roach}}$

After splitting, the dataset looks as follows:

Weight	**Length**
242	25.4
290	26.3
340	26.5
363	29
500	29.7
1000	37

Weight	**Length**
200	23.5
180	25.2
290	26
390	31.7
160	22.5
140	20.8
40	14.5

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - \left(\Sigma X\right)^2} \qquad a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n}$$

Calculation of slope and intercept for Bream spices:

n = 6

b$_{\text{bream}}$ = (510964.8 - 475616.5) / (30787.14 - 30241.21) = 35348.3 / 545.93 = 64.74

a$_{\text{bream}}$ = 455.83 – (64.74 * 28.98) = -1420.8

Therefore, the linear regression model is Weight$_{\text{bream}}$ = **-1420.8** + **64.74** $*$ Length$_{\text{bream}}$

Calculation of slope and intercept for Roach spices:

n = 7

b$_{\text{roach}}$ = (253617 - 229880) / (28121.24 - 26961.64) = 23737 / 1159.6 = 20.46

a$_{\text{roach}}$ = 200 – (20.46 * 23.457) = -280.16

Therefore, the linear regression model is Weight$_{\text{roach}}$ = **-280.16** + **20.46** $*$ Length$_{\text{roach}}$

(Correct process only(2 marks) correct process with correct answer (3 marks))

(ii)Proper Definition 2 marks

-----End------