Name: Shyed Shahriar Housaini
ID: MCE 079 05536

Ans to the q.no- 1 (a)

(i) Entry for students ; C, D, F, G, B should be should be cleared.

Entry C has noisy and missing data. (eg 3,39 and .0 - CGPA & IELTS) and a NULL Entry.
Entry D has noisy data (eg 12.0 IELTS score)

Entry F and G are probable duplicated entry.
Entry B has a NULL Entry.
~~Cle~~ Cleaned entries will be

Student	CGPA	IELTS	Publication	Job Exp	Scholarship
B	2.8	5.5	0	1	NO
C	3.9	7.0	2	3	YES
F	3.6	7.0	3	1	YES

Ans to the q.no 1 (b)

Samples: -34, 21, 13, 46, 999

Sample Size = 5

$v' = \frac{v - \mu}{6}$;

Mean $\mu = \frac{1045}{5}$

$= 209$

Standard deviation, $\sigma =$

$$= \sqrt{\frac{(-34-209)^2 + (21-209)^2 + (13-209)^2 + (46-209)^2 + (999-209)^2}{5}}$$

$= 395.85$

for, -34

$v' = \frac{-34 - 209}{395.85}$

$= -.613$

for, 21

$v' = \frac{21 - 209}{395.85} = -0.474$

for, 13

$v' = \frac{13 - 209}{395.85} = -0.495$

for 46,

$v' = \frac{46 - 209}{395.85}$

$= -0.412$

for, 999

$v' = \frac{999 - 209}{395.85}$

$= 1.996$

Ans

Ans to the q. no 2 (a)

A	B	$a=(A-\bar{A})$	$b=(B-\bar{B})$	axb	a^2	b^2
121.1	4.9	28.5667	-3.866	-110.45	816.05	14.95
90.8	8.9	-1.733	0.133	-0.231	3.0043	0.0178
65.7	12.5	-26.833	3.733	-100.167	720.026	13.937
$\bar{A}=92.53$	$\bar{B}=8.766$			-210.8667	1539.0867	28.9067

$\therefore$ Corelation Coefficient =

$$\frac{\sum_{i=1}^{n}(A_i-\bar{A})(B_i-\bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i-\bar{B})^2}}$$

$$= \frac{216.8667}{\sqrt{1539.0867}\times\sqrt{28.9067}} = \frac{-210.866}{210.9262}$$

$$= -0.9997$$

The data sets have strong negative correlation.
The variable moves in opposite direction.

Ans to the q. no. 2(b)

Sorted data is: 12, 23, 23, 43, 45, 54, 65, 67, 90, 100, 234, 1090

With intervals will be $w = \frac{b-a}{n}$

$= \frac{1090-12}{4} = 269.5$

≈ 270

Output:

for range [12 ~ 270]:-

[12, 23, 23, 43, 45, 54, 65, 67, 90, 100, 234]

for range [271 ~ 540]:- No data.

for range [541 ~ 810]:- No data.

for range [811 ~ 1080]:- No data

for Range [1081 ~ 1090]:-

[1090]

Smoothing by Bin Boundaries

Bin 1: [12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 234]

Bin 2: [1090]

3(a)(i) From Sorted data we can see position of 49 is 5th = Pt ;
total sample is $n = 12$

so ; $\frac{5}{12} \times 100\% = 42$

So, 49 would fall in 42 percentile

Ans: 42 Percentile

3(a)(ii)

Parametric

Non-P

Ans to the q. no- 3 a(ii)

Numerocity Reduction: It is a data reduction technique which replaces the original data with alternative small data; this is a brief form of data representation. Regression, Log-Linear Models, Histograms, clustering, sampling are examples.

Parametric Reduction	Non-Parametric Reduction
1. Uses fixed number and test group means	1. Uses flexible numbers and test medians.
2. Applicable for variables, consider strong assumptions and less data.	2. Used for variables & attributes, fewer assumptions, and more data.
3. Normal distribution	3. No assumed distribution.
4. Handles interval data, ration data	4. Handles original data

Ans to the q.no- 3 (b) : Observation

From figure(Fig-1) boxplot, we can see that

(a) Only Method-2 has and outlier.

(b) Median score is diffirent for different methods.

(c) Median for method 1 and 2 is closer together (25); But with different distribution

(d) Method 3 has the widest rang of data, Method 2 has the most concentrated score data.

(e) Method 1 and 4 has almost similar IQR but ~~therir~~ most of the score in method 1 is lower than Method 4.

(f) 50% of all data in method 4 has higher score than 75% of method 3.

(g) Method 3 & 4 has similar upper data/score ~~range~~ limit.