Assignment 3 : data science

Ans (1) Region A: [10, 15, 12, 8, 14]

Region B: [18, 20, 16, 22, 15]

mean of each region

$\bar{x}$ of Region A = $\frac{59}{5}$ = 11.8

$\bar{x}$ of Region B = $\frac{91}{5}$ = 18.2

Ans (2)

customes satisfaction on scale 1 to 5

data : [4, 5, 2, 3, 5, 4, 3, 2, 4, 5]

Calculate mode of the survey

4 — 111

5 — 111

2 — 11

3 — 11

modes are [4, 5]

Q3

Ans ③

Salaries of two departments

A : [5000, 6000, 7000, 5500]

B : [4500, 5500, 5700, 6000, 5200]

median salary of each department

for A:

[5000, 5500, 6000, 7000]

median = $\frac{5500 + 6000}{2}$ = 5750

for B : [4500, 5200, (5500) 5800, 6000]

median = 5500

Q4
Ans4 Stock prices : [25.5, 24.8, 26.1, 25.3, 24.9]

Calculate the range

range = max - min

= 26.1 - 24.8

= [1.3]

Ans 5

Group A = [85, 90, 92, 88, 91]
Group B = [82, 88, 90, 86, 97]

perform t-test, if defference in mean score

H_0 : means are same $\mu_A = \mu_B$
H_1 : means are not same $\mu_A \neq \mu_B$

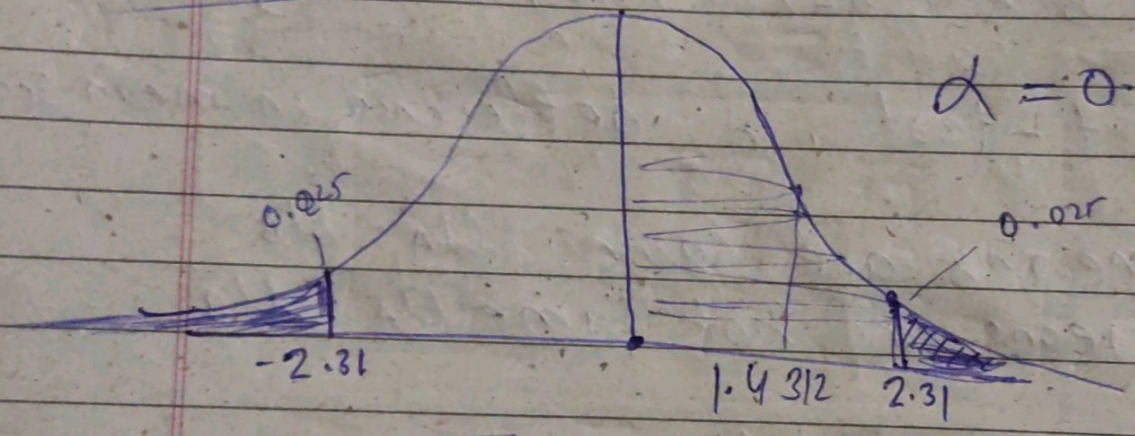
2-Sample t-test

	Group A	Group B
mean	89.2	86.6
Std (sample)	2.7749	2.9665
Sample Size	5	5
σ^2 (sample)	7.7	8.8

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

$$= \frac{89.2 - 86.6}{\sqrt{\frac{7.7}{5} + \frac{8.8}{5}}}$$

t-value = 1.4312



t-test Critical value = 2.31

So 1.4312 falls under the region where we don't reject H_0

So the test says:

Null hypo. to accept

$$\mu_A = \mu_B$$

Q6

Ans (6) Advertising: [10, 15, 12, 8, 14]
Sales : [25, 30, 28, 20, 26]

find the correlation

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1}$$

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \times \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$$

$$\boxed{\rho_{xy} = 0.876}$$

Q7

Ans 7

Height	μ	$(x-\mu)^2$
160	167.8	60.84
170	167.8	4.8
165	167.8	7.84
155	167.8	163.84
175	167.8	51.83
180	167.8	148.83
170	167.8	4.83

$$\sigma^2 = \frac{1}{N}\sum(x-\mu)^2$$

$$= \frac{442.87}{7}$$

$$\sigma^2 = 63.26$$

$$\text{St.d} = \sigma = \sqrt{63.26}$$

$$\boxed{\sigma = 7.95}$$

Q8

Ans 8

Tenure: [2, 3, 5, 4, 6, 2, 4]

Satisfaction: [7, 8, 6, 9, 5, 7, 6]

$$SE_{line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots + (y_n - (mx_n + b))^2$$

we have to find m & b that minimize ↓ SE

$$SE = y_1^2 - 2y_1(mx_1+b) + (mx_1+b)^2$$
$$+ y_2^2 - 2y_2(mx_2+b) + (mx_2+b)^2$$
$$\vdots$$
$$+ y_n^2 - 2y_n(mx_n+b) + (mx_n+b)^2$$

$$= y_1^2 - 2y_1mx_1 - 2y_1b + m^2x_1^2 + 2mx_1b + b^2$$
$$+ y_2^2 - 2y_2mx_2 - 2y_2b + m^2x_2^2 + 2mx_2b + b^2$$
$$\vdots$$
$$+ y_n^2 - 2y_nmx_n - 2y_nb + m^2x_n^2 + 2mx_nb + b^2$$

$$= (y_1^2 + y_2^2 + y_n^2) - 2m(x_1y_2 + x_2y_2 + \cdots x_ny_n)$$
$$- 2b(y_1 + y_2 + \cdots y_n) + m^2(x^2 + x_2^2 - \cdots x_n^2)$$
$$+ 2mb(x_1 + x_2 + \cdots x_n) + nb^2$$

$$SE_{Line} = n\bar{y}^2 - 2mn\bar{x}\bar{y} - 2bn\bar{y} + m^2n\bar{x}^2 + 2mbn\bar{x} + nb^2$$

In order to minimize SE_{line}

$$\frac{\partial SE}{\partial m} = 0 \quad + \quad \frac{\partial SE}{\partial b} = 0$$

$$-\bar{xy} + m\bar{x}^2 + b\bar{x} = 0$$

$$-\bar{y} + m\bar{x} + b = 0$$

$$m\bar{x}^2 + b\bar{x} = \overline{xy}$$

$$m\bar{x} + b = \bar{y}$$

Note:

$y = mx + b$

point $(\bar{y}, \bar{x})$ lies on this eqn

by solving the eqn

$$m = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}} = \frac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}$$

$$\boxed{m = \frac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}}$$

$$\boxed{b = \bar{y} - m\bar{x}}$$

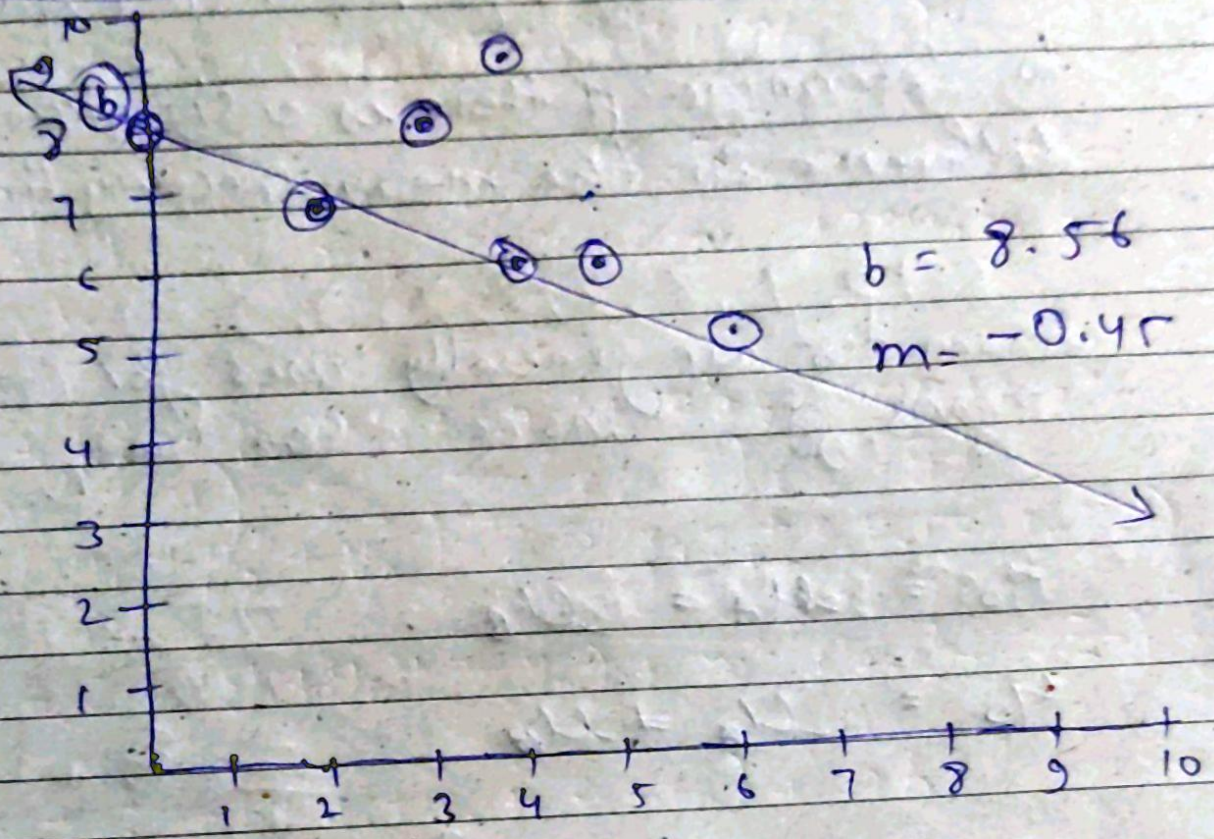
Now predict the Job Satisfa

$$m = \frac{25.4 - 24.57}{13.79 - 15.71}$$

$$\boxed{m = -0.46}$$

$b = \bar{y} - (-0.46)\bar{x}$

$\boxed{b = 8.56}$



Ans 9. medication A : [10, 12, 14, 11, 13]
B : [15, 17, 16, 14, 18]

perform analysis of variance (ANOVA) to determine if there is a significant difference in the mean recovery time.

Anova test are generally used for comparing means of more than two groups.

$H_0 = \mu_A = \mu_B$

$H_0 = \mu_B \neq \mu_B$

$SSW = \Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2 = 20$

$n(n-1) = 8$

SSB = | Grand mean = 14

d.f = 9

$SSB = 5(12 - 14)^2 + 5(16 - 14)$

$= 40$

F stat = $\frac{SSB}{n-1}$ / $\frac{SSW}{n(n-1)}$

$\frac{2}{20} \times 8$

F stat = 16

$\alpha = 0.10$

Fcritical = 3.46

Reject Null Hypothesis

$\mu_A \neq \mu_0$

Q 10
Ans [8, 9, 7, 6, 8, 10, 9, 8, 7, 8]

In order to calculate 75% percentile first we have to arrange in ascending order

6, 7, 7, 8, 8, 8, 8, 9, 9, 10

$n = 10$

75^{th} percentile index $= \frac{75}{100} \times n + 1$

$\frac{3}{4} \times 10 + 1$

$\frac{3}{4_2} \times \overset{5}{\cancel{10}} + 1$

$\text{lower}(7.5) + 1$

$= 8$ index

75^{th} percentile $= 9$

Q 11
Ans [10.2, 9.8, 10.0, 10.5, 10.3, 10.1]

Hypothesis test if mean are differ from 10

one Sample t-test

mean of sample $\bar{x} = 10.15$

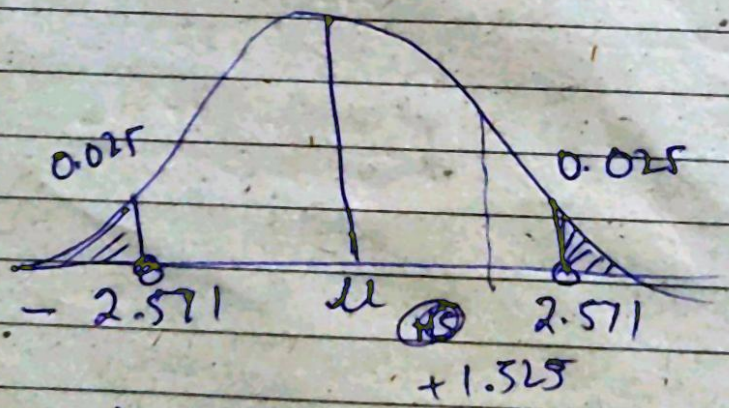
$$t_{stat} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{10.15 - 10}{\frac{0.24}{\sqrt{6}}}$$

$$= \frac{0.15 \times \sqrt{6}}{0.24}$$

$$\boxed{t_{stats} = 1.525}$$

$H_0 \Rightarrow \bar{x} = \mu = 10$ degree of freedom = 5
$H_1 \Rightarrow \bar{x} \neq \mu$ or $\neq 10$

$\alpha = 0.05$



Since t stats falls under the acceptance region

we accept Null Hypothesis

$$\boxed{H_0 = \mu = 10}$$