ASSIGNMENT 2.

1) a) $\Sigma xy = 182560$ $\quad \Sigma x = 850$ $\quad S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 34,915$

$\Sigma x^2 = 92500$ $\quad \Sigma y = 1737$

$S_{xx} = 20,250$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 1.724198$ $\quad$ ~~S_{xx}~~ $S_{yy} =$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 27.14317.$

$\hat{y} = 27.14317 + 1.724194x$

b) The slope of the regress^n line refers to its slope parameter β. Since β is significantly different from 0, thus exist a positive linear relationship between x.

2) Let x = share price of Dell.
y = share price of IBM.

(i)

$S_{xy} = 875.16$ $\quad \hat{\beta} = \frac{S_{xy}}{S_{xx}} = 2.90115$

$S_{xx} = 301.66$

$S_{yy} = 6409.7125$ $\quad \hat{\alpha} = \bar{y} + \hat{\beta}\bar{x} = 3.74809$

$\hat{y} = 3.74809 + 2.90115x.$

(ii) Assumptions:

* the distribution of x and y are normally distributed.
* the explanatory variable are independent of each other.

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

Level of significance = 5%.

95% confidence interval of $\beta = \hat{\beta} \pm t_{10,2.5\%} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{10}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 387.07447.$$

95% confidence interval for β = $\hat{\beta} \pm t_{10, 2.5\%} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$= 2.90115 \pm 2.52379$

95% CI for β = $(0.37736, 5.42484)$

(iii) Mean IBM share price = μ_0

$$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_{n-2} \qquad \boxed{x_0 = 40}$$

$se(\hat{\mu}_0) = \sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)\hat{\sigma}^2} = 10.59894.$

$\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} x_0 = 119.78409$

95% confidence interval for μ_0: $\hat{\mu}_0 \pm t_{10; 2.5\%} \, se(\hat{\mu}_0)$

$= 119.78409 \pm 23.61443$

95% CI for μ_0. $= (96.17866, 143.40852)$

Q.3) (i) Comments

- The centres of the distribution differs for all the 4 cities. Thus, there is a ~~primer~~ case for suggestion that the underlying means are different.
- The difference between the mean time taken to commute to office in peak hours & non peak hr are in the order city A (heighest) to city D (lowest).
- The variation in the data for city C is the lowest compared to city D which appeals to ~~the~~ be highest. However, with only 7 observations for each city, we can't be sure that there is a real underlying difference in variance.

(ii) Assumptions underlying analysis of variance:

- The populations must be normal.
- The populations have a common variance.
- The observations are independent.

(iii) Ho: The mean of differences is same for each city

H1: The mean of differences is not the same.

$SS_{TOTAL} = S_{YY} = 1116.11$

$$SS_{REG} = \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right) - \frac{103^2}{28} = 516.11$$

$SS_{RES} = SS_T - SS_{REG} = 600$

ANOVA TABLE:

SOURCE.	DOF	SS	MSS
Regress^n	3	516.11	172.04
Residual	24	600	25
TOTAL.	27	1116.11	

Under H_0, $F = \frac{MSSREG}{MSSRES} = \frac{172.04}{25} = 6.88$

Test statistics = 6.88.

$F_{3,24,5\%} = 3.009$

Hence, we have significient ~~of~~ evidence to reject at 5% los. Thus, there are underlying differences b/w the cities.

(iv) Analysis of mean diff^r

Since, $\bar{y}_1 = 9.14$, $\bar{y}_2 = 6.29$, $\bar{y}_3 = 1.14$, $\bar{y}_4 = -1.86$.

$$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$$

$\hat{\sigma}^2 = \frac{SSRES}{n-k} = 25$ the least significance difference b/w any of means is:

$$t_{24,0.025} \, \hat{\sigma}^2 \sqrt{\frac{1}{7}+\frac{1}{7}} = 5.52.$$

Now, we can examine the difference b/w each of the pairs of means. If the difference is less than the level of significance difference then their is ~~not~~ no significant difference b/w the means.

We have,

$$\bar{y}_1 - \bar{y}_2 = 2.85$$
$$\bar{y}_2 - \bar{y}_3 = 5.15$$
$$\bar{y}_3 - \bar{y}_4 = 3$$

All the 3 differences are less than 5.52, we thus pairs to show that they have no significant ~~at~~ difference.

$$\overline{\bar{y}_1 > \bar{y}_2} > \bar{y}_3 > \bar{y}_4$$

Examining to see if the first 2 groups can be combined, $\bar{y}_1 - \bar{y}_3 = 8$. There is a significant difference b/w means 1&3, so we cant combine the first 2 groups. Examining to see if the last 2 groups can be combined $\bar{y}_2 - \bar{y}_4 = 8.15$, Hence we cant combine the last 2 grps. Therefore, the diagram remains as before.

Q.4) a) The mathematical model of one way ANOVA is given by,

$$Y_{ij} = \mu + \tau_i + e_{ij}, \quad i = 1,2,\cdots k. \quad j = 1,2,\cdots n$$

where

k = number of treatments.

n = no. of responses from i^{th} treatment.

Y_{ij} is the j^{th} response for i^{th} treatment.

τ_i is the i^{th} treatment.

μ is the over mean response.

e_{ij} is the error term. (We are assuming that e_{ij} is i.i.d normal with mean = 0 and variance = σ^2)

b) H_0: mean claim amounts of five companies are equal.

H_1: mean claim amounts of five companies are not equal.

$n_1 = 7$ $\quad n_4 = 7$

$n_2 = 8$ $\quad n_5 = 5$

$n_3 = 6$ $\quad n = 33$

$$\sum_{i=1}^{5}\sum_{j=1}^{n_i} y_{ij} = 1856 \qquad \sum_{i=1}^{5}\sum_{j=1}^{n_i} y^2_{ij} = 174,316.$$

$$CF = \left(\sum_{i=1}^{5}\sum_{j=1}^{n_i} y_{ij}\right)^2 / n = 104385.94$$

$SS_{TOT} = 174316 - CF = 69,930.06.$

$$SS_{REG} = \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{384^2}{5} - 104385.84$$

$= 22,325.69$

$SS_{RES} = SS_{TOT} - SS_{REG}$

$= 47,604.37.$

Source of variation	DOF	SS	MSS
Regression	4	22326	5581·42
Residual	28	47604	1700·16
TOTAL	32	69930	

Test statistics $= \frac{MSS_{REG}}{MSS_{RES}} = 3.283$

$F_{4,28,0.05} = 2.714$.

We have sufficient evidence to reject Ho.

c) (ii)

Papers cleared	3-5	5-8	8-10	10-12	TOTAL
0-3	45	20	6	5	76
4-6	7	20	9	6	42
7-9	5	8	15	12	40
TOTAL	57	48	30	23	158

Expected frequency.

P.C.	3-5	5-8	8-10	10-12	TOTAL.
0-3	27·417	23·08861	14·43	11·063	
4-6	15·1518	12·759	7·97	6·11	
7-9	14·430	12·1518	7·59	5·80	

No clubbing for Ei's required as all Ei's are greater than 5.

Contingency table:

H_0: Sales are independent of the no of acturial papers.
H_1: Sales are dependent of the no of actuarial papers cleared.

(one sided test) because of χ^2 test.

Test statistics = $\Sigma \frac{(O_i - E_i)^2}{E_i} = 49.91146$

$\chi^2_{6, 5\%} = 12.59$

Hence, we have sufficient evidences to reject H_0.
The salaries are dependend on ~~not~~ number of acturial paper cleared.

Q.5) (i) Assumptions:
- Populations are normal
- Populations have common variance
- Operations are independent

(ii) Mean at Rate 1 for $\sqrt{x}$ = $\frac{\sqrt{28} + \sqrt{13} + \sqrt{21}}{3} = 4.52443$.

Variance at rate 2 for $ln(x)$ = $\frac{\Sigma (x - \bar{x})^2}{n-1} = 0.12127$.

(iii) The scientist was correct is asserting that the ln transformation must be close before carrying one way ANNOVA as for the transformation it can be claimed that the assumption of common var for the underlying population holds.

(iv) a) $Y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1, 2, 3, 4$; $\forall j = 1, 2, 3$

- Y_{ij} is the ln transformational value of the j^{th} observation of the j^{th} observation of the number of food observed when ~~th~~ the ~~th~~ i^{th} ~~result~~ rate was applied.

- μ is the overall population mean.
- τ is the deviation of the i^{th} rate mean such as $\Sigma \tau_i = 0$
- e_{ij} are the independent error. and $e_{ij} \sim N(0, \sigma^2)$

b) FOR ANOVA.

$H_0: \tau_i = 0$

$H_1: \tau_i \neq 0$ for atleast one i

Rate	Mean	Variance	y_i^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.716	0.186	294.053
4	6.575	0.148	389.060
TOTAL	20.110	0.618	973.373

Here, $y_i^2 = (\Sigma y_{ij})^2 = (3 \times \text{Mean}_i)^2$

$$SS_{RES} = \sum_{i=1}^{4} \left\{ \sum_{j=1}^{3} (y_{ij} - \bar{y}_i)^2 \right\}$$

$$= \Sigma (2 \times \text{variance})$$

$$= 2 \times 0.618$$

$$= 1.236.$$

Sources of variance	DOF	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
TOTAL	11	22.38	

Test statistics $= \frac{MSSRES}{MSSREG} = 45.638$

$F_{3,8,H} = 7.951$

(c) Given the observed F statistic value as much larger than this, we can state the p value for this test is almost near to 0 or is in other words there is overwhelming evidence against H_0. Thus it can be conducted that the underlying means arenot equal.

~~Q#)~~

Q.6) $S_{xy} = 661.2$

$S_{xx} = 54.9$

$S_{yy} = 8923.6$

(i) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.94466$, $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = 12.04371$

(ii) Fitted regress^n equation

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 9.2295$

$\hat{y} = 9.2295 + 12.04371x$

(iii) $SS_{TOTAL} = S_{yy} = 8923.6$

$SS_{REG} = \dfrac{S^2_{xy}}{S_{xx}} = 7963.3048$

$SS_{RES} = SS_{TOTAL} - SS_{REG} = 960.29$

$R^2 = \dfrac{S^2_{xy}}{S_{xx} S_{yy}} = 89.238703\%$

model is a good fit as model explains 89.238703% of the variability ~~is~~ in the output.

Q.7) (i) $S_{xx} = 4788.4221$

$S_{xy} = 2176.84$

$S_{yy} = 1189.20769$

$\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = 0.4545$,

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 10.78056$.

$\hat{y} = 10.78056 + 0.45451x$.

(ii) $H_0: \beta = 0$

$H_1: \beta \neq 0$

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right) = 22.20136$$

$$\text{test statistics} = \frac{\hat{\beta} - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = 6.67568$$

$t_{9, 2.5\%} = 2.262$

as test statistics is greater than critical value, we will reject H_0.

Assumpt^n made → populations have a common variance.
population follow the normal distribution

(iii) $$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.91213.$$

(iv) For mean response (μ_0)

$$se(\hat{\mu}_0) = \sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)} \hat{\sigma} = 1.55181$$

$$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_9$$

95% CI of μ_0: $\hat{\mu}_0 \pm t_{9, 2.5\%} \cdot se(\hat{\mu}_0) = (22.15331 \pm 3.51018)$
$= (18.64313, 25.66349)$

individual ~~resp~~ response (y_0)

$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0 = 22.153$

$$se(\hat{y}_0) = \sqrt{\left(\frac{1}{n} + 1 + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \hat{\sigma}^2} = 4.9607$$

$$\frac{\hat{\mu_0} - \mu_0}{se(\hat{\mu_0})} \sim t_{n-2}$$

95% CI for $\hat{\mu_0}$: $\hat{\mu_0} \pm t_{9,.25}\, se(\hat{\mu_0})$ = $22.15331 \pm \cancel{3.51018}\ 11.22131$

$= \cancel{(18.64313, 25.6634)}$

$= (10.932, 33.37462)$

(v) There is clearly some relationship between the residuals and the explanatory variable values. Hence our assumption of normally might be wrong. The observation are not independent as well. The model is not a good fit to the data.

Q.8) (i) The main purpose of factor analysis is to reduce many individual item into a fewer number of dimension.

- Factors analysis can be used to simplify data, such as reducing the number of variables in regression model.
- Principle component analysis is a technique for reducing the dimensionality of such data sets increasing interpretability but at the same time minimising information loss.

While newly identified principle components are chosen to be:

- Uncorrelated.
- Linear combination of the original variable of the data which maximise the variance

(ii)

PC	Diagonal entry (PCi)	~~PCi / Sum~~ PCi / (sum(PCi) over 1to5)
PC1	0.456	65%
PC2	0.137	18.5%
PC3	0.08	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.1015	100%

(iii) As 1^{st} 3 prinicipal components explain over 95% of total variance, dimensionality can be reduced to 3 for this dataset. The 1^{st} 3 principal components can then be used for building further classification or regression modelling purpose.

Q.9) (i) There appears to be a negative correlation b/w marks scored & hours spent per day on social media.

(ii) $S_{xx} = 75$

$S_{yy} = 2482.4$

$S_{xy} = -296$

$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -0.686002$ Hence, they have a weak (-)ve correlation.

(iii) $H_0: \rho = 0$

$H_1: \rho < 0$

$\tanh^{-1}(r) - \tanh^{-1}(\rho) \sim N(0, \frac{1}{n-3})$

$\frac{\tanh^{-1}(r) - \tanh^{-1}(\rho)}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$

Z calculated $= -2.07893$

$Z_{5\%} = 1.6449$

$|Z_{cal}| > Z_{tab}$ Hence, we will reject H_0.

$\therefore \rho < 0$.

(iv) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -3.94667$.

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 82.16$

$\hat{y} = 82.16 - 3.94667x$

(v) $R^2 = \frac{S_{xy}}{S_{xx} S_{yy}} = 47.05983\%$. Hence, the model explains only 47% of the variability. The model is not a good fit of data.

(vi) $x_0 = 1$

$\hat{\mu}_0 = \hat{\alpha} - \hat{\beta} x_0$, ~~ta~~

$= 78.21333$

10) $S_{xx} = 0.0042$

$S_{yy} = 0.00126$ $SS_{TOT} = S_{yy} = 0.00126$

$S_{xy} = 0.0022$ $SS_{REG} = \frac{S^2_{xy}}{S_{xx}} = 0.00115$

$SS_{RES} = 0.00011$

ANOVA TABLE

Sources of variation	dOF	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011
Total	2	0.00126	

$H_0: \beta = 0$

$H_1: \beta \neq 0$ (one sided F test)

Test statistics $= \frac{MSS_{REG}}{MSS_{RES}} = 10.45455$.

$F_{1,1,5\%} = 161.4$

$F_{cal} < F_{tab}$,

We will not reject H_0;