Ques 1

x	40	10	100	110	120	150	20	90	80	130
y	56	62	195	240	170	270	48	196	214	286

a) $S_{xy} = \Sigma xy - n\bar{x}\bar{y}$

$= 34915$

$S_{xx} = \Sigma x^2 - n\bar{x}^2$

$= 20250$

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

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

$\bar{y} = 173.7$

$\bar{x} = 85$

$\hat{\alpha} = 27.14317$

$\hat{y} = 27.14317 + 1.72419x$

b) The regression line refers to its slope parameter β. Since B is significantly different from 0. So there exist a positive linear relationship b/w x and y.
Hence, there is a positive linear relationship b/w the amt. withdrawn and the number of hours until the next visit.

Date

Ques 2 =>

$X =$ Share price of Dell
$Y =$ Share price of IBM

(i) $Sxy = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$

$= 875.16$

$Syy = \Sigma y^2 - n\bar{y}^2$
$= 6409.7125$

$Sxx = \Sigma x^2 - n\bar{x}^2$
$= 301.66$

$\hat{\beta} = \frac{Sxy}{Sxx} = 2.90115$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
$= 3.74809$

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

(ii) Assumptions
-> The dist of x and y are all normal
-> The explonatory variables values are all independent of each other

95% CI interval for $\beta = \hat{\beta} \pm t_{10, 2.5\%} \sqrt{\frac{\hat{\sigma}^2}{Sxx}}$

$\hat{\sigma}^2 = \frac{1}{n-2}\left(Syy - \frac{S^2xy}{Sxx}\right)$

$= \frac{1}{10}\left(Syy - \frac{S^2xy}{Sxx}\right)$

$\hat{\sigma}^2 = 387.07447$

95% CI for $\beta = \hat{\beta} \pm t_{10/2.5\%} \sqrt{\frac{\hat{\sigma}^2}{Sxx}}$

$= 2.90115 \pm 2.52379$

95% CI for $\beta = (0.37736, 5.42494)$

#

Date

(iii) Mean IBM share price $= \mu_0$

$$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_{n-2}$$

$x_0 = 40$

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

$$\hat{\mu}_0 = 119.79409$$

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

$$= 10.59894$$

95% CI for $\mu_0 = \hat{\mu}_0 \pm t_{10;2.5\%} se(\hat{\mu}_0)$

$= 119.79409 \pm 23.61443$

95% CI for $\hat{\mu}_0 = (96.17966, 143.40852)$

Ques 3 ⇒ Comments on the plot

- The centers of the distribution differ for all the 4 cities. Thus, there is a prima facie case for suggesting that the underlying means are different.
- The difference b/w the mean time taken to commute to office in peak hours & non-peak hours are in the order city A (highest) to city D (lowest)
- The variation in the data for city C is lowest compared to city D which appears to be highest. However with only 7 observations for each city, we can't be sure that there is a real underlying difference in variance.

#

Date

(ii) Assumptions underlying analysis of variance:

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

(iii) Ho: The mean of difference is same for each city.
H1: The mean of difference is not the same.

$SSTOT = S_{yy}$
$= \Sigma y^2 - \frac{(\Sigma y)^2}{n}$

$= 1116.11$

$SSREG = 516.11$
$SSRES = SSTOT - SSREG = 600$

ANOVA Table

source	degree of freedom	sum of squares	mean sum of squares
Regression	3	516.11	172.04
Residuals	24	600	25
-			
Total	27	1116.11	

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

$T.S = 6.88$

$F_{3,24}\ 5\% = 3.009$ (Right tail test)

We have sufficient evidence to reject Ho at 5% LOS. Thus there are underlying diff b/w the cities.

#

Date

(iv) Analysis of mean diffrences

since $\bar{y}_{1*} = 9.14$

$\bar{y}_{2*} = 6.29$

$\bar{y}_{3*} = 1.14$

$\bar{y}_{4*} = -1.86$

$$\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$$

$$\hat{\sigma}^2 = \frac{SSRES}{n-2} = 25$$

The least significance diff b/w any part of means is:

$$t_{24, 0.025} \times f \sqrt{\frac{1}{7}+\frac{1}{7}} = 2.064 \times \sqrt{25} \times \sqrt{\frac{2}{7}} = 5.52$$

Now, we can examine the diffrences between each of the pairs of means. If the diff is less than the least significance diff & there is no significant diffrenes b/w the means

$\bar{y}_{1*} - \bar{y}_{2*} = 2.85$

$\bar{y}_{2*} - \bar{y}_{3*} = 5.15$

$\bar{y}_{3*} - \bar{y}_{4*} = 3$

All the three diffrence are less than 5.52, these pairs to show that they have no significant diffrence.

$$\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$$

Examining to see if the first 2 groups can be combined if

$\bar{y}_{1*} - \bar{y}_{2*} = 8$

There is a significant diffrence b/w means 1 & 3, so we cant combine the first 2 groups.

#

Date

ANOVA Tables

Source	degree of freedom	Sum of squares	mean sum of squares.
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

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

Test statistics = 6.88

$F_{3,24}$ 5% = 3.009 {Right tailed test}

Hence, we have sufficient evidence to reject Ho at 5% level of significance. Thus, there are underlying differences b/w the cities.

(iv) Analysis of mean difference

Since

$\bar{y}_{1*} = 9.14$

$\bar{y}_{2*} = 6.29$

$\bar{y}_{3*} = 1.14$

$\bar{y}_{4*} = -1.86$

$\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$

#

Date

Ques 4 a)

The mathematical model of one-way ANOVA is given by,

$$y_i = \mu + T_i + e_{ij} \quad , \quad i = 1, 2, ---, K$$
$$j = 1, 2, ---, n_i$$

where,

K = number of treatments
n_i = no. of responses from i^{th} treatment
- y_i, the j^{th} response from i^{th} treatment.
- T_i is the i^{th} treatment
- μ is the over mean response
- e_{ij} is the error term

Assumption : e_{ij} is iid $N(0, \sigma^2)$

b) H_0: mean claim amount of five companies are equal
H_1: mean ______ are not equal

$n_1 = 7$ $\quad$ $n_4 = 7$
$n_2 = 8$ $\quad$ $n_5 = 5$
$n_3 = 6$ $\quad$ $n = 33$

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_{ij} = 1856$$

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

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

$$SS_{TOT} = 174316 - CF$$
$$= 69930.06$$

#

Date

$$SSREG = \sum_{i=1}^{3} \left(\sum_{j=1}^{n_i} y_{ij}\right)^2 / n_i - CF$$

$$= 22325.69$$

$$SSRES = SSTOT - SSREG$$
$$= 47604.37$$

Source of variation	d.f.	SS	MSS
Regression	4	22326	5581.42
Residual	28	47604	1700.16
	32	69930	

$$T.S = \frac{MSSREG}{MSSRES}$$
$$= 3.283$$

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

We have sufficient evidence to reject H_0.

c) Observed frequency (O_i)

Paper Years	Salary per annum (in Rs lacs)				Total
	3-5	5-8	8-10	10-12	
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 (E_i)

Papers cleared	Salary per annum (in Rs lacs)			
	3 - 5	5 - 8	8 - 10	10 - 12
0 - 3	27.41772	23.08861	14.4303	11.06329
4 - 6	15.15189	12.75949	7.97468	6.11392
7 - 9	14.43038	12.15189	7.59484	5.82278

Cont. tables

H_0: Salaries are independent of the no. of arrear papers cleared

H_1: Salar are dependent ——

$$T.S. = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(r-1)(c-1)}$$

$$= 11.275 + 0.4137 + 4.92 \ldots$$
$$+ 6.55324$$

$$T.S = 49.91146$$

degree of freedom $= (r-1)(c-1)$
$= 2 \times 3$
$= 6$

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

Hence we have suff evide to reject H_0.
The salaries are dep on no. of arrears papers cleared

#

Ques 5 (i)

Assumptions under ANOVA

(i) • Populations are normal
• Populations have common variance
• observations are independent

The 4 sample variances are very diff from each other

Hence the common assumption var won't hold.

(ii) mean at Rate for $\sqrt{x}$ = $\frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3}$

= 4.52443

Var at Rate 2 for $\log_e (x)$ = $\frac{\Sigma (x-\bar{x})^2}{n-1}$ = 0.12127

(iii) a) $X_{ij} = \mu + \tau_i + e_{ij}$., $i = 1,2,3,4$, $j = 1,2,3$

→ X_{ij} is the $\log_e$ transformed value of the j^{th} observation of the no. of generation per square foot observed where the i^{th} rate was applied.
→ μ is the overall population mean.
→ τ_i is the deviation of the i^{th} mean such that $\Sigma \tau_i = 0$
→ e_{ij} are the independent error terms.
$e_{ij} \sim N(0, \sigma^2)$

b) For ANOVA

$H_0: \tau_i = 0$, $i = 1,2,3,4$
$H_1: \tau_i \neq 0$ for atleast one i.

#

Date

Rate	$\ln(x)$ 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	20.11	0.618	973.373

Here $y_i^2 = \left(\sum_{j=1}^{3} x_{ij}\right)^2 = (3 \times \text{Mean } i)^2$

$$SS\,Rate = \sum_{i=1}^{4} \frac{y_i^2}{3} - \frac{y^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.11)^2}{12}$$

$$SS_{Rate} = 21.153$$

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

$$= \sum_{i=1}^{4} [2 \times \text{variance}_i]$$

$$= 2 \times 0.618$$
$$= 1.236$$

Source of variation	d.f	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
	11	22.38	

Test statistics $= \dfrac{MSS_{Reg}}{MSS_{Res}}$

$= 45.638$

$F_{3,8,1\%} = 7.591$

#

Date

Given the observed F statistics value as much larger than this, we can state the p value for this test is almost near to 0 or in other words there is overwhelming evidence against H0. Thus it can be concluded that the underlying means are not equal.

Name = Rittle Jain

Prob. & Statistics Assignment 2

Date

Ques 6 ⇒

X	5 9 2 3 3 11 6 5 4
y	73 120 34 46 35 24 86 93 45 66

$n = 10$

(i) correlation coefficient $= \dfrac{Sxy}{\sqrt{Sxx\ Sxy}}$

$Sxy = \Sigma xy - n\bar{x}\bar{y}$

$= 2853 - 10 \times \frac{39}{10} \times \frac{562}{10} = 661.2$

$Sxx = \Sigma x^2 - n(\bar{x})^2 = 54.9$

$Syy = \Sigma y^2 - n(\bar{y})^2 = 39946$

Coefficient correlation $= 0.9446623$

(ii) Fitted regression equation

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

$\hat{\beta} = \dfrac{Sxy}{Sxx}$

$\hat{\alpha} = 9.2295$

$\hat{\beta} = 12.04371$

$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$

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

(iii) $SS_{TOT} = Syy = 39946$

$SS_{REG} = \dfrac{S^2xy}{Sxx} = \dfrac{(661.2)^2}{54.9} = 7963.304918$

$SS_{RES} = SS_{TOT} - SS_{REG} = 31982.69508$

Date

Coefficient of determination $= \frac{SSREG}{SSTOT} = 0.1993517$

$= 19.935\%$

Comment:

→ model is not a good fit because it only considers about 20% variation and does not account for remaining 80% variation

→ we need to alter the model to make it good fit.

Ques 7 ⇒

(i) Linear regression equation

$$\hat{y_i} = \hat{\alpha} + \hat{\beta} x$$

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

$$\hat{\alpha} = 10.7896$$

$$\hat{\beta} = \frac{Sxy}{Sxx} = \frac{2176.7393}{4789.5042} = 0.4544$$

$$\hat{y_i} = 10.7896 + 0.4544x$$

(ii)

$$TS = \frac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t_{n-2}$$

$$H0: \beta = 0$$

$$H1: \beta > 0$$

$$TS = \frac{0.4544 - 0}{se(\hat{\beta})}$$

#

Date

$$se(\hat{\beta}) = \sqrt{Var(\hat{\beta})} = \sqrt{\frac{\hat{\sigma}^2}{Sxx}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[Syy - \frac{S^2xy}{Sxx}\right]$$

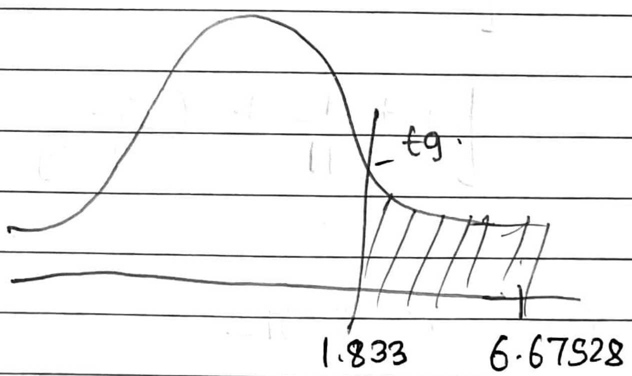
$$= \frac{1}{11-2}\left[\cancel{229684}\ 1189.029 - \frac{(2176.7393)^2}{4789.5042}\right]$$

$$\hat{\sigma}^2 = 22.193561$$

$$se(\hat{\beta}) = \frac{4.711004311}{\sqrt{4789.5042}} = 0.06807195$$

$$TS = \frac{0.4544}{0.06807195} \sim t_9$$

$$= 6.67528 \sim t_9$$



Since the test statistic lies in rejection region we have sufficient evidence to reject Ho at CS 5%. So, we conclude that $\beta > 0$, which means there exists some linear (positive) relationship with positive correlation.

Assumptions ⇒ → All e_i follow normal dist with mean 0 & var σ^2.
→ All variables are iid's.
→ All explanatory & response variables follow normal dist

Date

iii) Calculate sample correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.91214$$

(iv) $\hat{x}^3 = 25$

Individual response

$$\hat{y} = \hat{\alpha} + \hat{\beta}x_0$$
$$\hat{y} = 10.7896 + 0.4544(25)$$
$$\underline{\hat{y} = 22.1496}$$

T.S ⇒ $P:Q \cdot = \dfrac{\hat{y}_0 - y_0}{S_c(\hat{y}_0)} \sim t_{n-2}$

$$S_c(\hat{y}_0) = \sqrt{Var(\hat{y}_0)}$$

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

$$= \left[1 + \frac{1}{11} + \frac{(25-15.83090)^2}{4789.5042}\right] \times 22.193561$$

$$= 24.60073$$

$$S_c(\hat{y}_0) = 4.959912$$

95% CI

$$P(-2.262 < t_9 < 2.262) = 0.95$$

CI:

$$\hat{y}_0 \pm 2.262\, S_c(\hat{y}_0)$$

CI ⇒ $(10.93027, 33.3689)$

#

Mean response

$\hat{\mu}_0 = 22.1496$

$$Var(\hat{\mu}_0) = \left[\frac{1}{11} + \frac{(25-15.83090)^2}{4789.5042}\right] \times 22.193561$$

$= 24.60073 - 22.193561$

$= 2.407169$

$Se(\hat{\mu}_0) = 1.551505$

95% CI

$P(-2.262 < t_9 < 2.262) = 0.95$

CI

$\hat{\mu}_0 \pm 2.262\, Se(\hat{\mu}_0)$

CI = $(18.64009, 25.6591)$

v) After seeing the above residual plot we see that errors are not iid random variables and are not completely random in nature and there is a pattern in errors.

So, ϵ_i does not follow $N(0, \sigma^2)$.

We need to alter our model to fit all our assumptions right.

Date

Ques 8 :-

(i) -> The main purpose of factor analysis is to reduce many individuals item into a fewer number of dimensions.

-> Factor 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 datasets increasing interpretability but at the same time minimizing information loss.

While newly indentified principle components are chosen to be

- Uncorrelated
- linear combinations of the original variables of the data
- which maximizes the variance.

(ii)

Principle Component	Diagonal entry (PCi)	PCi / (Sum (PCi) over 1 to 5)
PC1	0.456	65%
PC2	0.137	19.5%
PC3	0.08	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	100%

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

#

Ques 9 :-

$$S_{XX} = \Sigma x^2 - n\bar{x}^2 = 75$$
$$S_{YY} = \Sigma y^2 - n\bar{y}^2 = 2482.4$$
$$S_{XY} = \Sigma xy - n\bar{x}\bar{y} = -296$$

(i) Based on above scatterplot we can clearly see that there is a negative correlation between marks scored and hours spent per day on social media.

(ii) Correlation coefficient $= \dfrac{S_{XY}}{\sqrt{S_{XX} S_{XY}}} = -0.686001$

(iii)

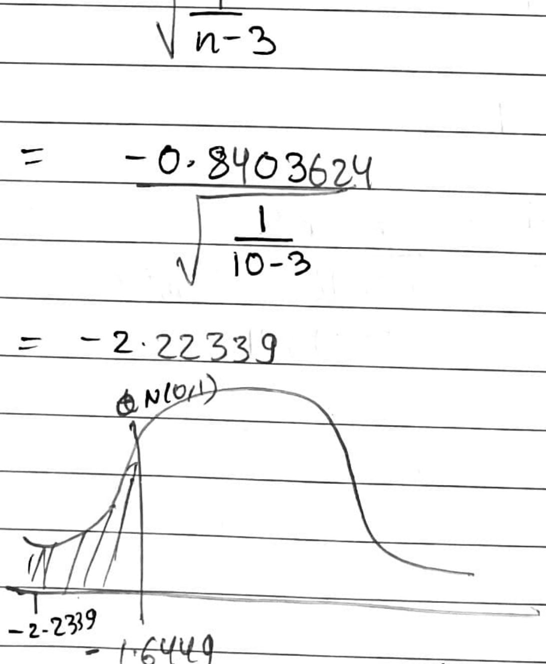
$H_0: \rho = 0$

$H_1: \rho < 0$ $\qquad r = -0.686001$

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

$$= \frac{-0.8403624}{\sqrt{\frac{1}{10-3}}}$$

$$= -2.22339$$



since the TS lies in rejection region

We have sufficient evidence to reject H_0 at CS 5%.
So we can conclude that there is a negative correlation.

Date

(iv)

Response variable = Marks
Hours spent = Exploratory variable

$\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i$

$\hat{\beta} = \frac{S_{XY}}{S_{XX}}$

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

$\hat{\alpha} = 82.16$

$= -3.946666$

Fitted linear regression model

$\hat{y_i} = 82.16 - 3.94666x$

v) Coefficient of Determina = r^2 = 0.470598
or 47.059%.

It means that our model only explains about 47% of variance i.e model is not a good fit.
We have to tweak our model to make it good fit.

(vi)

$x_0 = 1$

$\hat{y_i} = \hat{\alpha} + \hat{\beta} x_1$

$= 82.16 - 3.94666 \times 1$

$= 78.21334$

Ques 10 ⇒

ANOVA Table

Source of variation	Degree of freedom	Sum of squares	Mean sum of squares
Regression	1	0.00115238	0.00115238
Residual	3 - 2	0.00011422	0.00011422
TOTAL	1	0.0012666	

$SS_{TOT} = S_{YY} = 0.0012666$

$SS_{REG} = \dfrac{S^2_{XY}}{S_{XX}} = \dfrac{(0.0022)^2}{0.0042} = 0.00115238$

$SS_{RES} = SS_{TOT} - SS_{REG} = 0.00011422$

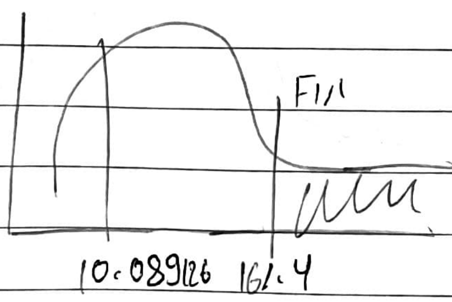
H0 : Industry has no impact on ~~regression~~ resignation rate

H1 : Industry has impact on resignation rate

$T.S = \dfrac{SS_{REG}}{SS_{RES}/n-2} \sim F_{1,1}$

$= \dfrac{0.00115238}{0.00011422} \sim F_{1,1}$

$= 10.089126 \sim F_{1,1}$



Sin TS lies in AR.

Do not reject H0. So, therefore Industry has no impact on resignation ~~rate~~