Probability & Statistic II
Assignment - 2

1)

	x	y	xy	x^2
1	40	56	2,240	1600
2	10	62	620	100
3	100	195	19,500	10,000
4	110	240	26,400	12,100
5	120	170	20,400	14,400
6	150	270	40,500	22,500
7	20	48	960	400
8	90	196	17,640	8,100
9	80	214	17,120	6,400
10	130	286	37,180	16,900
Total	850	1,737	182,560	92,500

(a)

$$S_{xy} = \Sigma x_i y_i - \Sigma x_i \Sigma y_i / n = 182560 - (850)(1737)/10$$
$$= 34915$$

$$S_{xx} = \Sigma x_i^2 - (\Sigma x_i)^2/n = 92500 - 850^2/10 = 20250$$

$$\hat{b} = S_{xy} / S_{xx} = 1.72$$

$$\hat{a} = \bar{y} - b\bar{x} = (1737/10) - 1.72 \times (850/10) = 27.14$$

$$y = 27.14 + 1.72x$$

(b)

Gradient represents the amount of hours per rupee spent.

2) (i) $\Sigma x = 3852$, $\Sigma y = 1162.5$, $\Sigma xy = 38191.41$

$n = 12$, $\Sigma x^2 = 12666.58$, $\Sigma y^2 = 119026.9$

$S_{xx} = \Sigma x^2 - n\bar{x}^2 = 12666.58 - 12(3852/12)^2$

$= 301.66$

$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 119026.90 - 12(1162.5/12)^2$

$= 6409.71$

$S_{xy} = \Sigma xy - \cancel{\Sigma} n\bar{x}\bar{y} = 38191.41 - 12(385.2/12)(1162.5/1$

$= 875.16$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = (1162.5/12) - 2.90(385.2/12) = 3.78$

$y = \hat{\alpha} + \hat{\beta}x = 3.78 + 2.90x$

(ii) 95% C.I. for $\beta \pm t_{n-2}(2.50\%)\ s.e.(\hat{\beta})$

$s.e.(\hat{\beta}) = \sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}$

$\hat{\sigma}^2 = \dfrac{1}{n-2}\left[S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}\right] = 387.07$

$\therefore\ s.e.(\hat{\beta}^2) = \sqrt{\dfrac{387.07}{301.66}} = 1.13$

$\therefore$ 95% C.I. for β : $2.90 \pm 2.228(1.13)$

$= (0.38, 5.42)$

(iii) $\hat{y}_{x_0} \pm t_{n-2(2.5\%)}\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{(x_0-\bar{x})^2}{S_{xx}}\right)}$

$x_0 = 40$ (Dell share price in U.S.)

$\hat{y}_{x_0} = 3.48 + 2.90(40) = 119.78$

$\therefore$ 95 C.I. $= 119.78 \pm 2.228 \times \sqrt{387.07\left[1/2 + \dfrac{(40-32.1)^2}{301.66}\right]}$

$= (96.17, 143.89)$

3) (i) The centers of the distribution differs for all the four cities
The diffrence between the mean time taken to commute to office in peak hours and non-peak hours are in the order city A to city D (highest to lowest)

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

(iii) H_0 : The mean of diffrences is greater for each city

v/s

H_1 : The mean of diffrences are not the same for all cities.

$$SS_T = 1495 - \frac{103^2}{28} = 1116.11$$

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

ANOVA Table

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	26	600	25	
Total	27	116.11		

(iv) Under H_0, $F = 172.04/25.00 = 6.88$ $(\because F_{3,24})$
The 5% critical point ~~to reject~~ is 3.009, so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is resonable to conclude that there are underlying diffrences between cities

5) (i) The following assumptions are required for one-way analysis of variance (ANOVA):

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

(ii) For the transformation $x \to \sqrt{x}$, the sample mean for rate 1 will be:-

$$\frac{1}{3}\left(\sqrt{29}+\sqrt{13}+\sqrt{21}\right) = 4.52$$

For the transformation $x \to \log x$, the value of sample varience for ~~value~~ rate 2 will be:-

$$\frac{1}{2}\left[(\log_e 180 - 4.827)^2 + (\log_e 90 - 4.827)^2 + (\log_e 120 - 4.827)^2\right]$$

$$= 0.1213$$

(iii) The scientist was correct in asserting that the $\log_e$ transformation must be done before carrying out a one-way ANOVA as for this transformation it can be claimed that the assumption of common variance for the ~~as~~ underlying population holds.

(iv) We will perform an ANOVA on $\log_e x$ data. We would assume the following model:-

$$Y_{ij} = \mu + \tau_i + e_{ij}, \quad i = 1,2,3,4,$$
$$j = 1,2,3$$

Here :-

Y_{ij} is $\log_e$ transformed value of the j^{th} observation of the no. of germinators per square foot observed when the i^{th} rate was applied

- μ is the overall population mean.
- τ_i is the deviation of the i^{th} rate mean such that $\sum \tau_i = 0$
- e_{ij} = are the independent error terms which follows Normal distribution with mean 0 and common unknown variance σ^2

For ANOVA, the null hypothesis being tested here is:-

$H_0 : \tau_i = 0,\ i = 1,2,3,4$ against $H_1 : \tau_i \neq 0$ for at least one i

Sum of squares table: ($\log_e(x)$)

Rate	Mean	Variance	y_i^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.719	0.186	294.053
4	6.575	0.148	389.060
	20.110	0.618	973.373

Here : $y_i = \left(\sum_{j=1}^{3} y_{ij}\right)^2 = (3 \times \text{mean}_i)^2$

Now :-

$$SS(\text{Rate}) = \sum_i^4 = -\frac{y^2}{1^2} = \frac{973.373}{3} - \frac{(3 \times 20.110)^2}{1^2}$$

$$= 21.153$$

$$SS(\text{Residuals}) = \sum_i^4 = 1 \left\{ \sum_j^3 = 1 \left(y_{ij} - \bar{y}_i\right)^2 \right\} = 2 \times 0.618 = 1.236$$

$$= \sum_{i=1}^4 \{2 \times \text{var}_i\}$$

The ANOVA table is as follows

Source of variation	d.f	Sum of Squares	Mean squares	F
Rates	3	21.153	7.051	45.638
Residuals	8	1.236	0.155	
Totals	11	22.381		

The 1% critical value for $F(3,8)$ distribution is 7.951.

Considering the F statistic value is much larger than this, we can observe that the p-value for this, we can observe that the test is almost near to zero. This is an evidense against the null hypothesis H_0. Thus it can be concluded that the underlying means are not equal.

(v) A 95% confidence interval for the i^{th} rate mean is given by:-

$$\bar{y}_i \pm t_{8;0.975} \times \frac{\hat{\sigma}}{\sqrt{3}} \quad \text{for } i = 1,2,3,4$$

Here: $t_{8;0.975}$ is the 97.5% critical point of a t-distibution with 8 degrees of freedom and equals 2.306.

The common error variance σ^2 is estimated as

$$\hat{\sigma}^2 = \frac{\text{Residuals sum of squares}}{8} = 0.155$$

The 95% confidence intervals are tabulated in the following table.

Rate	Observed mean	95% C.I. (Transformed scale) LCI	95% C.I. (Transformed scale) UCI	95% C.I. (original scale) LCI	95% C.I. (original scale) UCI
1	2.992	2.469	3.516	11.80	33.635
2	4.827	4.303	5.350	73.954	210.623
3	5.716	5.193	6.239	179.950	512.505
4	6.575	6.052	7.098	425.770	1209.76

6) (i) $S_{xx} = \Sigma x^2 - n\bar{x}^2 = 207 - 10(39/10)^2 = 54.90$

$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 2853 - 10(39/10)(563/10)$

$= 661.20$

$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 40508 - 10(562/10)^2 = 8923.66$

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

$= \frac{661.20}{\sqrt{54.90}\sqrt{8923.6}} = 0.945$

(ii) Fitted linear regression equation

The coefficients of the regression eqⁿ are:-

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04$

$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \hat{x} = (562/10) - 1204(39/10) = 9.23$

$\therefore y = \hat{\alpha} + \hat{\beta}x = 9.23 + 12.04x$

(iii) Relation : $SS_{TOT} = SS_{REG} + SS_{RES}$

$SS_{TOT} = S_{yy} = 8923.60$

$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 8923.60 - \frac{(661.20)^2}{54.90}$

$= 960.30$

$SS_{REG} = S_{TOT} - S_{RES} = 8923.60 - 960.30$

$= 7963.30$

(iv) Coefficient of Determination :-

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{SS_{Reg}}{SS_{TOT}} = \frac{7963.30}{8923.60} = 0.8924$$

For the simple linear regression model, the value of the coefficient of determination is the square of the correlation coefficient for the data, since,

$$0.945 = r = \frac{S_{xy}}{(S_{xx} . S_{yy})^{0.5}} = \sqrt{R^2} = \sqrt{0.8924}$$

8) (i) Factor analysis / Principal component analysis is. A method for reducing the dimensionality of data. It seeks to identify key components nessesary to model and understand data original data variables may be.

- Correlated with each other

while newly identified principles components are chosen to be

- uncorrelated
- linear combinations of the original variables of the data
- which maximise the variance

(ii)

Principal component	Diagonal entry (PC_i)	PC_i / (Sum(PC_i) over 1 to 5)
PC_1	0.456	65.0%
PC2	0.137	19.5%
PC3	0.08	11.6%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	100.0%

(iii) As 1st 3 principal components explain over 95% of total variance, dimentionality can be reduced to 3 for this dataset. The 1st 3 principal components can then be used for building further classifaction or regression modeling purpose.