FACULTY OF COMPUTING
UTM Johor Bahru

Semester II 2024/2025

Subject	:	PROBABILITY & STATISTICAL DATA ANALYSIS (SECI 1143/SCST 1223)
Task	:	ASSIGNMENT 4 - Chapter 7 (60%) & Chapter 8 (40%)
Due Date	:	**JUNE 2025 (before 5 pm)** **1 week after release (please refer to the section's lecturer)**

INSTRUCTION:

1. This is a **GROUP** assignment. Please clearly write the group members **NAME G MATRIC NUMBER** in the front page of the submission.
2. This assignment contributes to 5% of overall course marks.
3. Only **HANDWRITTEN** submission is accepted:
    a. Submissions using any reporting or statistical tools (e.g.: MS Word, MS Excel, etc.,) will be **REJECTED**.
    b. Make sure the submission is neatly written. Any submission with handwriting that is unreadable, will be **REJECTED**.
    c. For answer that need to draw graphs, using graph paper is optional. You can use plain paper.
    d. Round your answers to **THREE** decimal places.
    e. Please scan/snapshot your work and save as a PDF file.
4. Submission via eLearning – only **ONE** group member needs to submit on behalf of the group.

NAME	MATIRC NO.
IZWAN AZIZ BIN ISMAIL @ ABD MALEK	SX241894ECJHF01
YUARAJ A/L PARTHIPAN	SX241919ECRHF01
SITI NURNAJIHAH BINTI MOHAMAD ANUAR	SX232351ECRHF04
FATIN SYAHIRAH BINTI NOR RASHID	SX241920ECRHF01
ASWINI A/P CHANDRASAGARAN	SX242452ECRHF01

PART 1 CHAPTER 7: CORRELATION AND REGRESSION (60%)

QUESTION 1 (10 MARKS)

A bakery production manager wants to investigate the linear relationship between the number of pastries produced and the production cost. To pursue his/her objective, the manager recorded the data on the number of pastries produced per day and the production cost per day (in thousands of Malaysian Ringgit) for 10 consecutive days as depicted in **Table 1**.

Table 1: Daily pastries produced and production cost for 10 consecutive days

Days	1	2	3	4	5	6	7	8	G	10
Number of pastries, x	35	50	45	60	70	55	40	65	75	80
Production cost, y	48	65	60	72	83	62	50	75	90	95

a) Calculate the correlation coefficient, r. (8 marks)

b) Based on the correlation coefficient, r obtained, make a conclusion on the linear relationship between the number of pastries produced and the production cost.

(2 marks)

$$r = \frac{\sum xy - \frac{(\sum x \sum y)}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})\,(\sum y^2 - \frac{(\sum y)^2}{n})}}$$

a) calculate the correlation coefficient r.

- data point $n = 10$

$x = [35, 50, 45, 60, 70, 55, 40, 65, 75, 80]$

$y = [48, 65, 60, 72, 83, 62, 50, 75, 90, 95]$

$\sum x = 575$

$\sum y = 700$

$\sum xy = 42125$

$\sum x^2 = 35125$

$\sum y^2 = 50576$

numerator $= \sum xy - (\sum x)(\sum y)/n$

$= 42125 - (575 * 700)/10$

$= 42125 - 40250$

$= 1875$

denominator $= \sqrt{(\sum x^2 - (\sum x)^2/n) * (\sum y^2 - (\sum y)^2/n}$

$= \sqrt{(35125 - 33065/10) * (50576 - 490000/10)}$

$= \sqrt{(35125 - 33062.5) * (50576 - 49000)}$

$= \sqrt{2052.5 * 1576}$

$= \sqrt{3236760}$

≈ 1798.54

$r = 1875 / 1798.54 = 0.986$

b) conclusion

Since the correlation $r \approx 0.986$ is very close to 1, we conclude that there is very strong positive linear relationship between the number of pastries produced and the production cost. As the number of patries increase, the production cost also increase in a nearly linear fashion.

QUESTION 2 (25 MARKS)

A social media analytics company is investigating the relationships between various factors to understand how they influence the engagement on posts. The company has collected data on the following variables (**Table 2**) for 6 different social media posts:

- **Number of Likes**: The number of likes received by the post.
- **Number of Comments**: The number of comments received by the post.
- **Number of Shares**: The number of times the post was shared.
- **Post Length**: The length of the post in characters.
- **Engagement Score**: A score (from 1 to 100) representing the overall engagement on the post.
- **Sentiment Score**: An ordinal variable representing the sentiment of the post (1: Very Negative, 2: Negative, 3: Neutral, 4: Positive, 5: Very Positive).

Table 2: Factors influencing engagement on posts

Post ID	Likes	Comments	Shares	Post Length	Engagement Score	Sentiment Score
1	150	20	30	200	85	4
2	100	10	20	100	70	3
3	200	25	40	250	90	5
4	80	8	15	150	60	2
5	170	22	35	220	88	5
6	120	15	25	180	75	3

a) Compute the correlation coefficient for Engagement and Sentiment Score variables. Interpret the result. (9 marks)

b) Compute the correlation coefficient for Likes and Share. Interpret the result. (9 marks)

c) Test the null hypothesis that there is no correlation between 'Engagement Score' and 'Sentiment Score' against the alternative hypothesis that there is a significant correlation. Use a significance level of 0.05 and provide the test statistics and p- values. (7 marks)

a) Compute the correlation coefficient for Engagement & Sentiment Score variables. Interpret the result.

$$\text{correlation coefficient} = r = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\sqrt{[(\Sigma x^2) - (\Sigma x)^2/n][(\Sigma y^2) - (\Sigma y)^2/n]}}$$

x = engagement score
y = sentiment score

	Engagement Score (x)	Sentiment (y)	xy	x^2	y^2
	85	4	340	7225	16
	70	3	210	4900	9
	90	5	450	8100	25
	60	2	120	3600	4
	88	5	440	7744	25
	75	3	225	5625	9
Σ	468	22	1785	37,194	88

$$r = \frac{1785 - (468 \times 22)/6}{\sqrt{[(37194) - (468)^2/6][(88) - (22)^2/6]}} = 0.970$$

Correlation coefficient of 0.970 shows a very strong positive linear relationship between engagement score and sentiment score, which means as the sentiment becomes positive, the engagement tends to increase significantly.

b) Compute the correlation coefficient for ~~on~~ likes and shares. Interpret the result.

	Likes (x)	Shares (y)	xy	x^2	y^2
	150	30	4500	22500	900
	100	20	2000	10000	400
	200	40	8000	40000	1600
	80	15	1200	6400	225
	170	35	5950	28900	1225
	120	25	3000	14400	625
Σ	820	165	24,650	122,200	4,975

$$r = \frac{24,650 - (820 \times 165)/6}{\sqrt{[(122,200) - (820)^2/6][(4,975) - (165)^2/6}} \quad \begin{matrix} 2100 \\ 2105.548 \end{matrix} = 0.997$$

$10,133.33 \qquad 4375$

Correlation coefficient of 0.997 shows a very strong positive linear relationship between likes and share, which means that posts with more likes tends to be shared frequently.

c) Test the null hypothesis that there's no correlation between 'Engagement score' and 'Sentiment score' against the alternative hypothesis that there's a significant correlation. Use a significance level of 0.05 and provide the test statistics and p-values.

$H_0 : \rho = 0$ (no linear correlation)

$H_A : \rho \neq 0$ (linear correlation exists)

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad \frac{0.97}{\sqrt{\frac{1-0.97^2}{6-2}}} \quad 0.0591 = 7.980$$

significance level (α): 0.05 $\qquad t_{\alpha/2,4} = t_{0.025,4} = \pm 2.776$

degrees of freedom ($n-2$): 4

Since $t = 7.980 > t_{0.025,4} = 2.776$, and p-value < 0.005, reject H_0. There's sufficient evidence of a significant linear relationship between engagement score & sentiment score at the 5% level of significance.

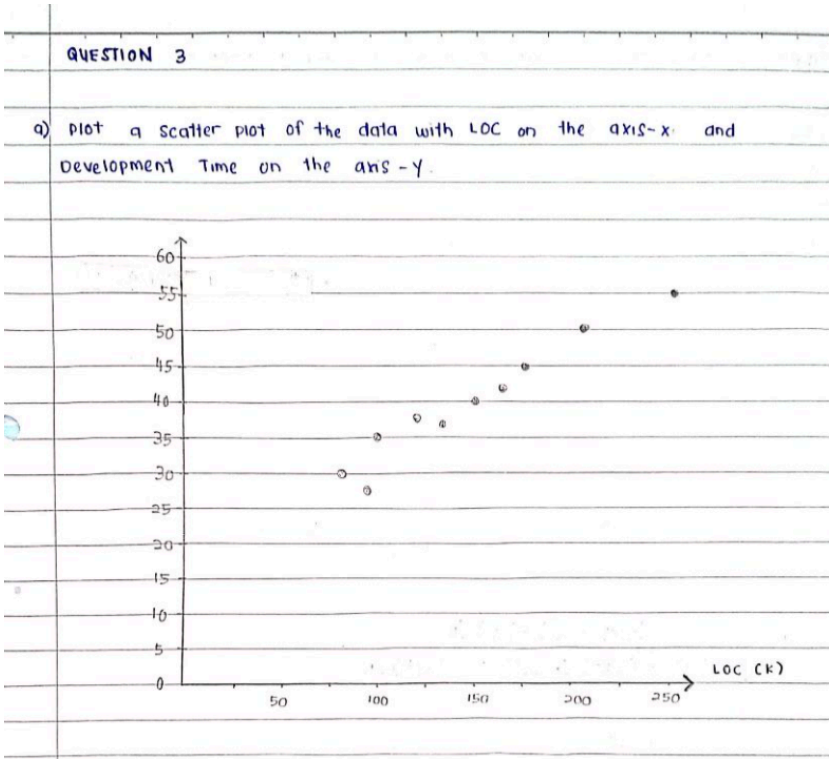
QUESTION 3 (25 MARKS)

A software development company wants to predict the time required to complete new software projects based on the estimated lines of code (LOC). The company has collected data on the development time (in weeks) and LOC for previous projects. Using this data presented in **Table 3,** answer the following questions in 2 decimal places:

Table 3: Length of code (LOC) and development time

Project ID	LOC (K)	Development Time (week)
1	150	40
2	100	35
3	200	50
4	80	30
5	170	45
6	120	38
7	160	42
8	90	28
9	250	55
10	130	37

a) Plot a scatter plot of the data with LOC on the x-axis and Development Time on the y-axis. (3marks)



b) Calculate the correlation coefficient between LOC and Development Time.

(8 marks)

a) calculate the correlation coefficient between LOC and Development Time.

x (LOC)	y (development time)	xy	x^2	y^2
150	40	6000	22500	1600
100	35	3500	10000	1225
200	50	10000	40000	2500
80	30	2400	6400	900
170	45	7650	28900	2025
120	38	4560	14400	1444
160	42	6720	25600	1764
90	28	2520	8100	784
250	55	13750	62500	3025
130	37	4810	16900	1369
1450	400	61910	235300	16636

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n}}$$

$$= \frac{10(61910) - (1450)(400)}{\sqrt{[10(235300) - (1450)^2/10][10(16636) - (400)^2}}$$

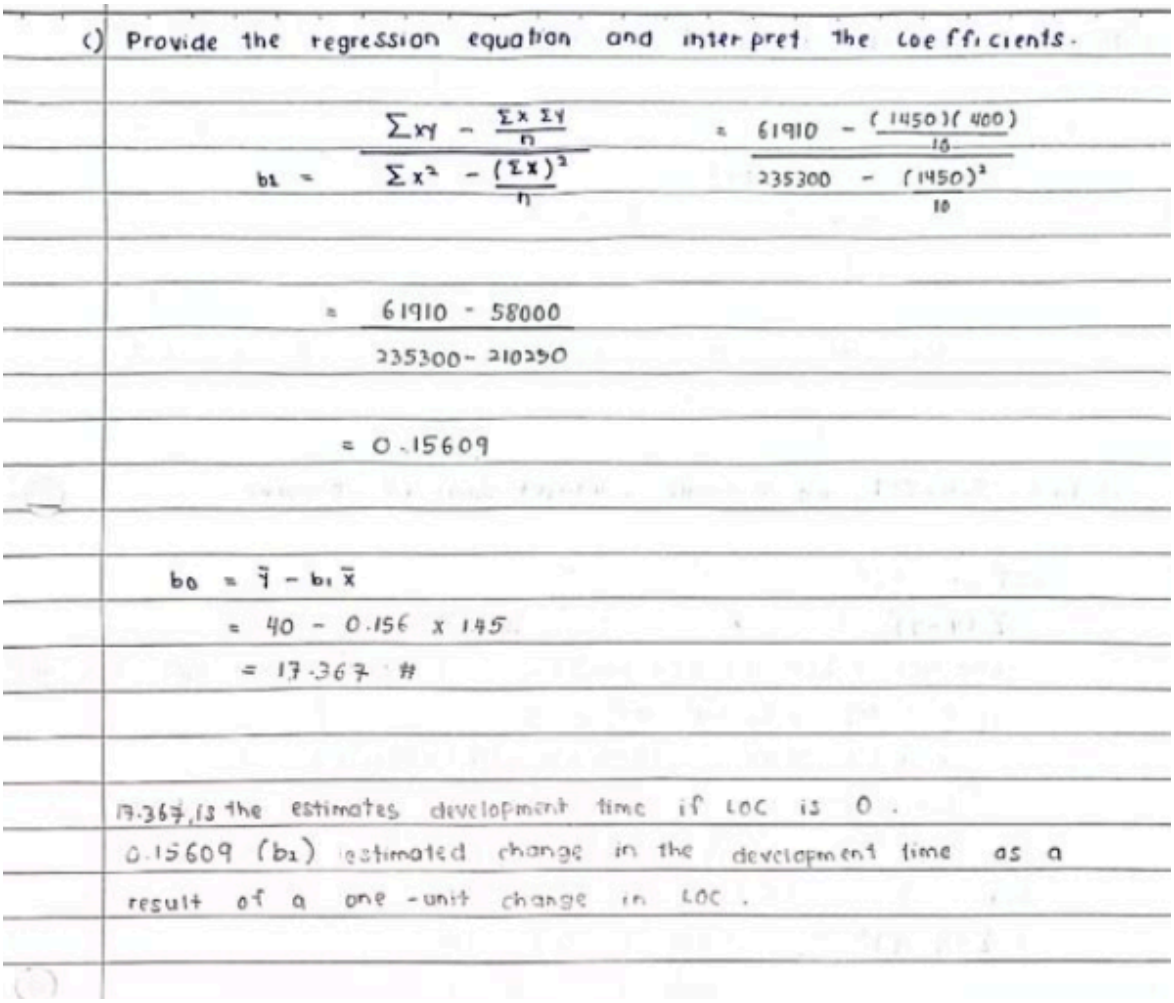
$$= \frac{619100 - 580000}{\sqrt{(250500)(6360)}}$$

$$= \frac{39100}{\sqrt{1593180000}}$$

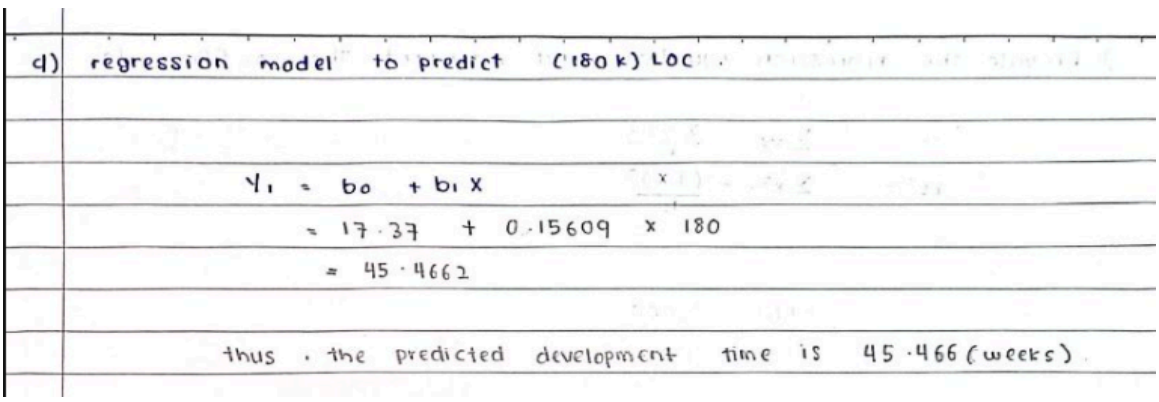
$$= \frac{39100}{39914.659}$$

$$= 0.980 \#$$

c) Fit a simple linear regression model using LOC as the independent variable and Development Time as the dependent variable. Provide the regression equation and interpret the coefficients. (5marks)



d) Use the regression model to predict the development time for a new project with an estimated 180K LOC. (2marks)



e) Find value of SSR, SST and R-Squared. Interpret value of R-Squared. (6 marks)

e) Find SSR, SST and R-square, interpret value of R-square.

SST = 636

$= \sum (y - \bar{y})^2$

$= (40-40)^2 + (35-40)^2 + (50-40)^2 + (30-40)^2 + (45-40)^2 + (38-40)^2 + (42-40)^2$
$+ (28-40)^2 + (55-40)^2 + (37-40)^2$

$= 0 + 25 + 100 + 100 + 25 + 4 + 4 + 144 + 225 + 9$

$= 636$ #

SSR

$= \sum (\hat{y} - \bar{y})^2$

$\hat{y} = 17.37 + 0.15609 \times (X)$

$SSR = (40.78 - 40)^2 + (32.98 - 40)^2 + (48.59 - 40)^2 + (29.85 - 40)^2 +$
$(43.91 - 40)^2 + (36.10 - 40)^2 + (42.34 - 40)^2 + (31.42 - 40)^2 +$
$(56.39 - 40)^2 + (37.66 - 40)^2$

$= 0.61 + 49.28 + 73.79 + 103.02 + 15.29 + 15.21 + 5.48 +$
$73.62 + 268.63 + 5.48$

$= 610.41$ #

R - squared.

$$R^2 = \frac{SSR}{SST}$$

$$= \frac{610.41}{636}$$

$= 0.960$ #

PART 2 CHAPTER 8: ANOVA (40%)

QUESTION 4 (20 MARKS)

A local agricultural researcher is conducting a study to determine whether different types of fertilizers affect plant height after 30 days. Three fertilizers (A, B, and C) are applied to separate groups of plants. Each group contains 5 plants, and all other growing conditions (light, water, soil) are kept constant. **Table 4** below shows the height (in cm) of plants after 30 days.

Table 4: Plant Heights after 30 Days (cm)

Fertilizer A	Fertilizer B	Fertilizer C
262	235	223
246	271	223
266	255	233
288	230	201
244	250	204

Conduct the ANOVA test for the above data by;

a) Define the hypothesis statement. (2m)

b) Calculate mean and variance. (3m)

c) Calculate the test statistics. (10m)

d) Calculate numerator and denominator degree of freedom. Use α = 0.05 (2m)

e) State the critical value. (1m)

f) Test the claim and state the conclusion. (2m)

Subject: Q4

a)

- Null Hypothesis (H_0):
  $\mu_A = \mu_B = \mu_C$
  (There are no difference in plant height)

- Alternative Hypothesis (H_1):
  At least one group mean is significantly different

b) Group means

$$\text{Mean A} = \frac{262 + 246 + 266 + 288 + 244}{5} = \frac{1306}{5} = 260.8$$

$$\text{Mean b} = \frac{235 + 271 + 255 + 230 + 250}{5} = \frac{1241}{5} = 248.2$$

$$\text{mean c} = \frac{223 + 223 + 233 + 201 + 204}{5} = \frac{1084}{5} = 216.8$$

Group variances.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Fertilizer A:

$$s^2_A = \frac{(262 - 260.8)^2 + (246 - 260.8)^2 + (266 - 260.8)^2 + (288 - 260.8)^2 + (244 - 260.8)^2}{4}$$

$$= \frac{1.44 + 220.90 + 27.04 + 739.84 + 282.24}{4} = \frac{1271.46}{4}$$

$$= 269.42$$

Fertilizer B:

$$s^2_B = \frac{(235 - 248.2)^2 + (271 - 248.2)^2 + (255 - 248.2)^2 + (230 - 248.2)^2 + (250 - 248.2)^2}{4}$$

$$= \frac{174.24 + 522.72 + 46.24 + 331.24 + 3.24}{4} = \frac{1077.68}{4} = 269.42$$

Subject: Date:

~~Exact~~ Fertilizer C :

$$S^2_C = \frac{(223-216.8)^2 + (223-216.8)^2 + (233-216.8)^2 + (201-216.8)^2 + (204-216.8)^2}{4}$$

$$= \frac{38.44 + 38.44 + 262.44 + 253.44 + 1638.44}{4} = \frac{2231.2}{4}$$

$$557.80 \text{ N}$$

c) ~~Cal~~ Grand Mean

$$\bar{x}_{grand} = \frac{1306 + 1241 + 1084}{15} = \frac{3631}{15} = 242.07$$

Sum of squares between groups

$$SSB = n \cdot \left[(\bar{x}_A - \bar{x}_{grand})^2 + (\bar{x}_B - \bar{x}_{grand})^2 + (\bar{x}_C - \bar{x}_{grand})^2 \right]$$

$$= 5 \cdot \left[(260.8 - 242.07)^2 + (248.2 - 242.07)^2 + (216.8 - 242.07)^2 \right]$$

$$= 5 \cdot [355.51 + 37.73 + 638.73]$$

$$= 5 \cdot 1031.97$$

$$= 5159.85$$

Sum of squares within groups

$$SSW = (nA-1)s^2_A + (nB-1)s^2_B + (nC-1)s^2_C$$

$$= 4 \cdot 317.87 + 4 \cdot 269.42 + 4 \cdot 557.80$$

$$= 1271.48 + 1077.68 + 2231.20$$

$$= 4580.36$$

Degrees of freedom.

Between groups : $df_1 = k - 1 = 3 - 1 = 2$

Within groups : $df_2 = N - k = 15 - 3 = 12$

Mean squares

$$MSB = \frac{SSB}{df_1} = \frac{5159.85}{2} = 2579.93$$

$$MSW = \frac{SSW}{df_2} = \frac{4580.36}{12} = 381.70$$

F-Ratio

$$F = \frac{MSB}{MSW} = \frac{2579}{381.70} = 6.76$$

d) Numerator (between groups) : $df_1 = 2$
Denominator (within groups) $= df_2 = 12$

e) From F-distribution table with $\alpha = 0.05$, $df_1 = 2$, $df_2 = 12$:

$F_{critical} = 3.89$

f) Since

$F_{calculated} = 6.76 > F_{critical} = 3.89$

We reject the null hypothesis.

Conclusion : There is significant evidence to conclude that at least one fertilizer produces a different effect on plant height after 30 days.

QUESTION 5 (20 MARKS)

A car manufacturer wants to test the effectiveness of four different brands of brake tires. The stopping distances (in meters) under identical conditions were measured using 5 test runs for each tire brand:

Table 4: Plant Heights after 30 Days (cm)

Tire Brand	Stopping Distance (meters)
L	35.5, 34.8, 36.1, 35.2, 34.9
M	38.2, 37.7, 38.5, 37.9, 38.0
N	33.9, 34.2, 33.7, 34.0, 33.5
O	36.8, 37.0, 36.5, 36.9, 37.1

At the 0.05 significance level, conduct the ANOVA test whether the mean stopping distances are the same for all four tire brands.

a) Define the hypothesis statement. (2m)

b) Calculate mean and variance. (3m)

c) Calculate the test statistics. (10m)

d) Calculate numerator and denominator degree of freedom. Use α = 0.05 (2m)

e) State the critical value. (1m)

f) Test the claim and state the conclusion. (2m)

Aswini A/p Chandrasagoran No. Date

Assignment 4

Question 5

a) Null Hypothesis (H_0):
The mean stopping distances for all four tire brands are equal.
$H_0: \mu L = \mu M = \mu N = \mu O$

Alternative Hypothesis (H_1):
At least one of the mean stopping distances is different.
H_1: Not all means are equal.

b) $L = \frac{35.5 + 34.8 + 36.1 + 35.2 + 34.9}{5} = \frac{176.5}{5} = 35.3$

$M = \frac{38.2 + 37.7 + 38.5 + 37.9 + 38.0}{5} = \frac{190.3}{5} = 38.060$

$N = \frac{33.9 + 34.2 + 33.7 + 34.0 + 33.5}{5} = \frac{169.3}{5} = 33.860$

$O = \frac{36.8 + 37.0 + 36.5 + 36.9 + 37.1}{5} = \frac{184.3}{5} = 36.860$

L $\frac{(Lx - 35.3)^2}{5-1}$

$= \frac{(35.5-35.3)^2 + (34.8-35.3)^2 + (36.1-35.3)^2 + (35.2-35.3)^2 + (34.9-35.3)^2}{4}$

$= 0.275$

M $\frac{(Mx - 38.06)^2}{5-1}$

$= \frac{(38.2-38.06)^2 + (37.7-38.06)^2 + (38.5-38.06)^2 + (37.9-38.06)^2 + (38.0+38.06)^2}{4}$

$= 0.093$

N $\frac{(Nx - 33.86)^2}{5-1}$

$= \frac{(33.9-33.86)^2 + (34.2-33.86)^2 + (33.7-33.86)^2 + (34.0-33.86)^2 + (33.5-33.86)^2}{4}$

$= 0.073$

O $\frac{(Ox - 33.86)^2}{5-1}$

$= \frac{(36.8-33.86^2) + (37.0-33.86)^2 + (36.5-33.86)^2 + (36.9-33.86)^2 + (37.1-33.86)^2}{4}$

$= 0.053$

No. Date

c) $\frac{176.5+190.3+169.3+184.3}{20} = \frac{720.4}{20} = 36.02$

$= 5(35.3-36.02)^2 + 5(38.06-36.02)^2 + 5(33.86-36.02)^2 + 5(36.86-36.02)^2$

$= 5(0.5184) + 5(4.1604) + 5(4.6656) + 5(0.7056)$

$= 50.250$

SSW

$= 4(0.275) + 4(0.093) + 4(0.073) + 4(0.053)$

$= 1.100 + 0.372 + 0.292 + 0.212 = 1.976$

between $= k-1 = 4-1 = 3$; within $N-k = 20-4 = 16$.

$\frac{50.250}{3} = 16.750$ $\frac{1.976}{16} = 0.12475$.

$F = \frac{16.750}{0.12475} = 134.256$ ✱

d) Numerator (between groups) df = 3
Denominator (within groups) df = 16

e) 3.24

f) F calculated = 134.256 > F critical = 3.24.

There is a significant difference in mean stopping distances among four tire brands.