UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING

SEMESTER II 2020/2021

SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION: 07

LECTURER: DR. NIES HUI WEN

ASSIGNMENT 1

GROUP 6: SISTHREEHOOD

GROUP MEMBERS:

1. EUNICE LIM XIAN NI (A20EC0034)
2. SHABRINA SALSABILA SAKRONI (A20EC0334)
3. TEH JING LING (A20EC0228)

Question 1

1. a) Population : Football club in Malaysiana
   b) Sample : All the Malaysiana Super League
   c) Discrete variable : Total numbers of goals scored
   d) Continuous variable : Amount of time
   e) Primary data : Number of goals scored in each match
   f) Secondary data : Mean number of goals per match

2. a) Population : All students at a school in JayBee
   b) Sample : Second-year pupils
   c) Discrete variable : Gender of second-year pupils
   d) Continuous variable : Height of second-year pupils
   e) Primary data : Number of pupils weighed
   f) Secondary data : Weight of first-year pupils

Question 2

1.

Scores / Factors	Frequency				Total
	1	2	3	4	
Price	14	6	0	0	20
Condition of car	3	6	11	0	20
Fuel efficiency	1	11	7	1	20
Car depreciation	6	11	2	1	20

2. Price :

Scale	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
1	14	0.70	14	0.70
2	6	0.30	20	1.00
3	0	0.00	20	1.00
4	0	0.00	20	1.00
Total	20	1.00		

Condition of the car :

Scale	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
1	3	0.15	3	0.15
2	6	0.30	9	0.45
3	11	0.55	20	1.00
4	0	0.00	20	1.00
Total	20	1.00		

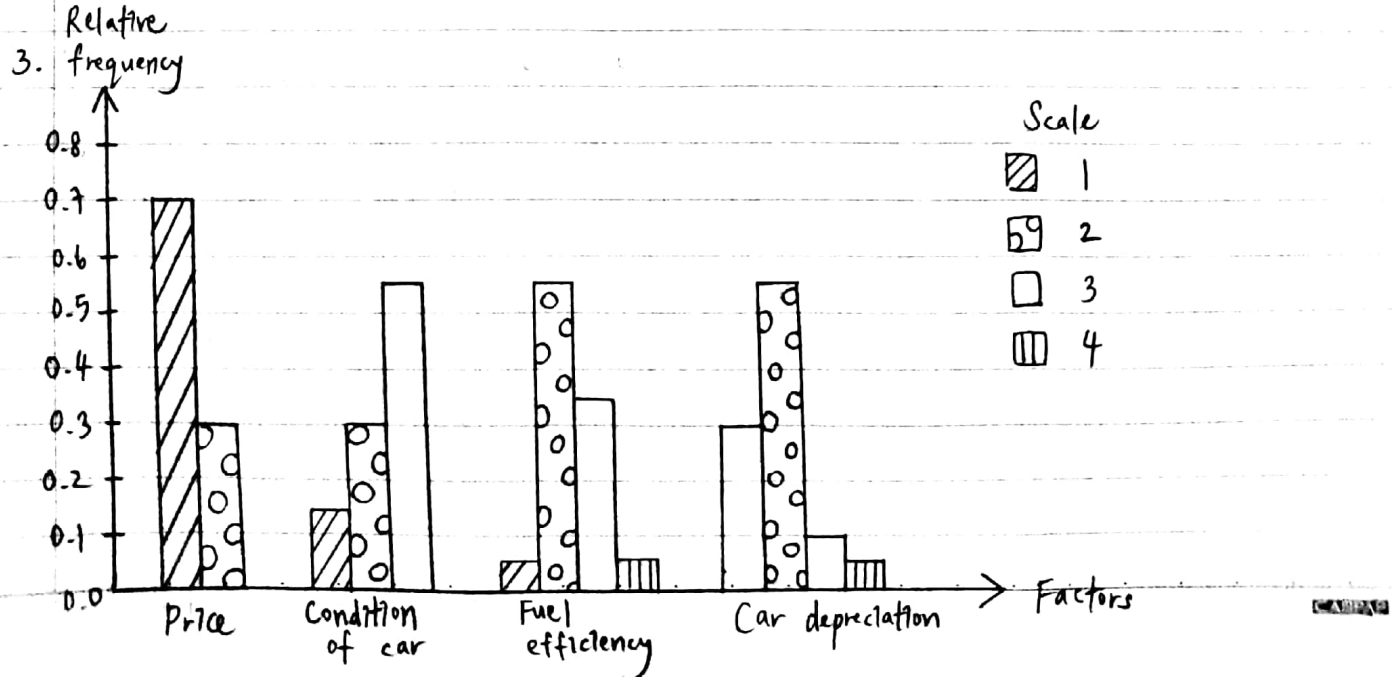
Fuel efficiency :

Scale	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
1	1	0.05	1	0.05
2	11	0.55	12	0.60
3	7	0.35	19	0.95
4	1	0.05	20	1.00
Total	20	1.00		

Car depreciation :

Scale	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
1	6	0.30	6	0.30
2	11	0.55	17	0.85
3	2	0.10	19	0.95
4	1	0.05	20	1.00
Total	20	1.00		

3.



Question 3

①

stem	Leaf
0	4 8 9
1	6 8
2	1 3 4
3	0 2 3 8
4	2 3 4
5	5
6	5
8	1.

key
0|4 4
2|1 21

② a. mean

↝ $\frac{586}{180} = 32,56$ //

b. mode

↝ there is no mode //

c. median

↝ $\frac{32+30}{2} = \frac{62}{2} = 31$ //

4 8 8 16 18 21 25 24 30 32 33 38 42 43
49 55 65 81

③ a. $Q_1 = i = \frac{25}{100} \cdot 18$

$= \frac{1}{4} \cdot 18 = 4,5$

$= k = 5$ //

$Y[k] = Q$

$Y[5] = 18$

↝ $Q_1 = 18$ //

• $Q_2 = i = \frac{50}{100} \cdot 18$

$= \frac{1}{2} \cdot 18 = 9 = k$

↝ $k+1 = 10$

$Y[9] = 30$

$Y[10] = 32$ ↝ $\frac{32+30}{2} = \frac{62}{2}$

$= 31 = Q_2 = \text{median}$ //

• $Q_3 = i = \frac{75}{100} \cdot 18$

$= \frac{3}{4} \cdot 18$

$= 13.5 = 14$

$k = 14$

$Y[14] = 43$

↝ $Q_3 = 43$ //

b. iqr $= Q_3 - Q_1$

$= 43 - 18$

$= 25$ //

c. mwl $=$ iqr $\times 1,5$

$= 25 \times 1,5$

$= 37,5$ //

• Lower limit

↝ $18 - 37,5 = -19,5$ //

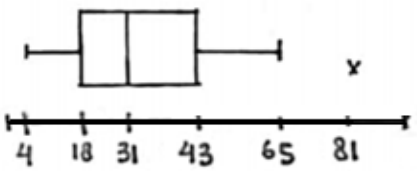
• Upper limit

↝ $43 + 37,5 = 80,5$ //

No things is below lower limit.

81 is above upper limit. thus 81 is a mild outliens.

65 is new maximum



Question 4

① mean

$$\frac{(175000 \times 1) + (250000 \times 5) + (500000 \times 4) + (700.000 \times 1)}{11}$$

$$= \frac{4125000}{11} = 375000 \text{ "}$$

② Median

$$i = \frac{50}{100} \cdot 11$$

$$= \frac{1}{2} \cdot 11 = 5,5$$

$$k = 6$$

$$\rightsquigarrow Y[6] = Q_2$$

$$= 250000 \text{ "}$$

③ median is better than mean because it has the best measurement for measuring center of dataset and does not effected by extremly large values of the data. Thus median is the "best" represent the value

Question 5

1. Age = ratio
   Current Smoker = nominal
   Body Mass Index (BMI) = Interval
   Hypertension = nominal

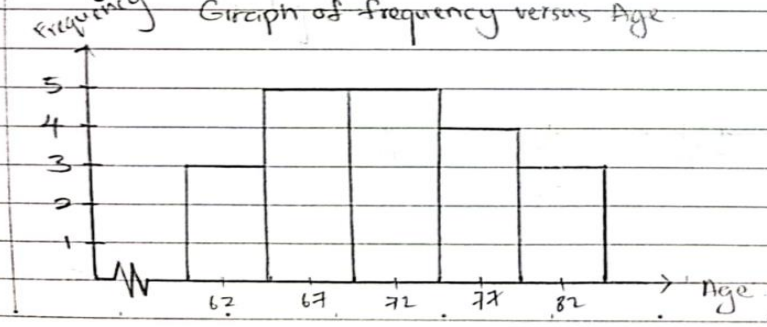
2.

CLASS INTERVAL	CATEGORY	FREQUENCY
$0 < BMI < 18.5$	Underweight	0
$18.5 <= BMI < 25.0$	Normal	12
$25.0 <= BMI < 80.0$	Overweight	7
$80.0 <= BMI$	Obesity	1
Total		20

3.

CLASS INTERVAL	CLASS BOUNDARIES	CLASS MIDPOINT	FREQUENCY	CUMULATIVE FREQUENCY
60-64	59.5 - 64.5	62	3	3
65-69	64.5 - 69.5	67	5	8
70-74	69.5 - 74.5	72	5	13
75-79	74.5 - 79.5	77	4	17
80-84	79.5 - 84.5	82.	3	20
Total			20	

4 Refer graph.



DATE: .....

QUESTION 6

1.

TIME (MINUTES)	Midpoint, x	FREQUENCIES	fx	CUMULATIVE FREQUECY
15.5 - 30.5	23	3	69	3
30.5 - 45.5	38	13	494	16
45.5 - 60.5	53	30	1590	46
60.5 - 75.5	68	25	1700	71
75.5 - 90.5	83	14	1162	85
90.5 - 105.5	98	8	784	93
105.5 - 120.5	113	4	452	97
120.5 - 135.5	128	2	256	99
135.5 - 150.5	143	1	143	100
Total		100	6650	

$$\text{Mean} = \frac{6650}{100} = 66.5$$

$$\text{Mode} = 45.5 + \left(\frac{30-13}{2(30)-13-25}\right)15$$

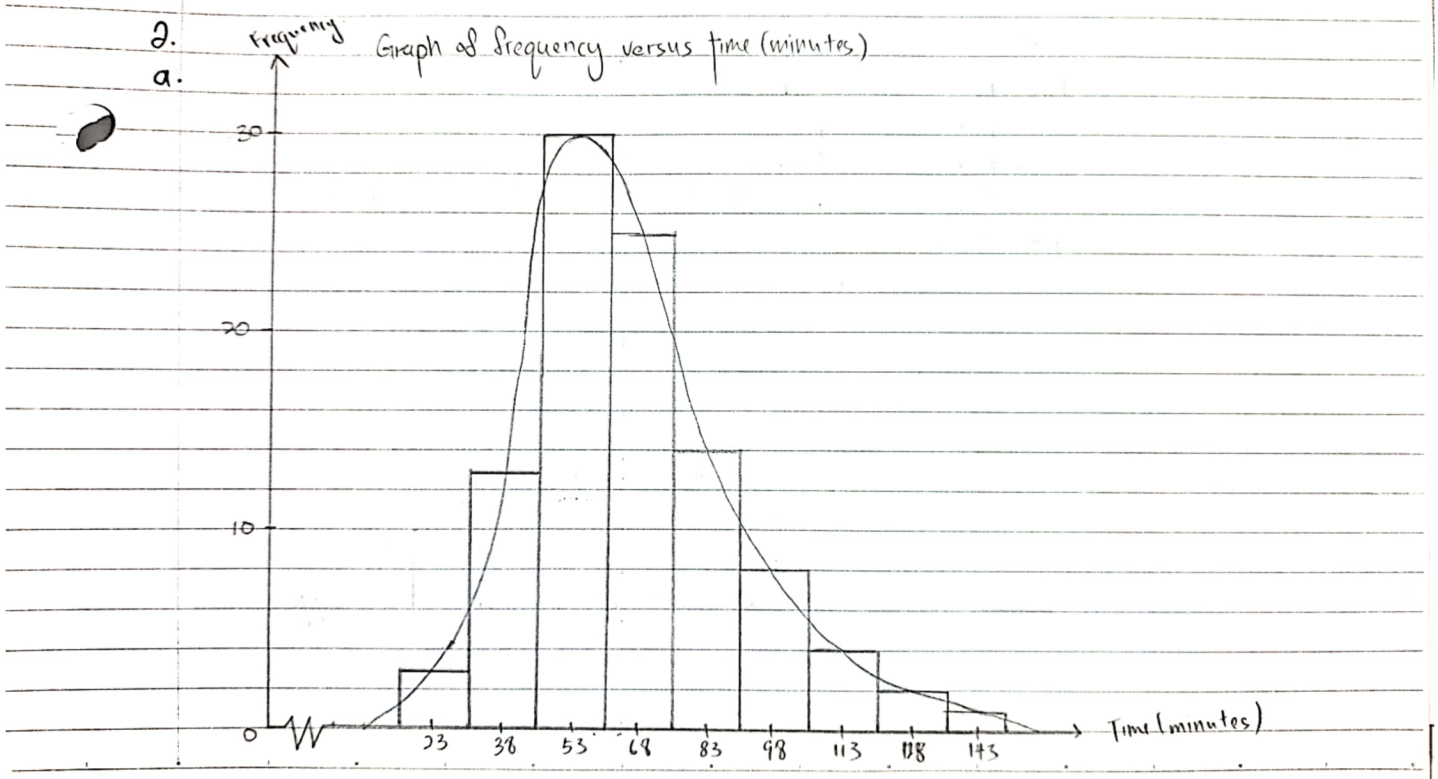
$$= 57.09$$

$$\text{Median} = 60.5 + \left(\frac{\frac{100}{2} - 46}{25}\right)15$$

$$= 62.9$$

2.

a.



Subject : Date:

b) The shape of the distribution is right skewed because Mean > Median > Mode.
The shape is also leptokurtic because the graph has heavier and fatter tail as compare to normal distribution.

c) The graph is positively skewed shows that most data are centered around mode, 57.09 min while right tail has longer distribution. This indicates that flight to be delayed for 57.09 min will often occur while the delay time more than 66.5 min will rarely occur.

Leptokurtic graph which has heavier and fatter tail means there are outliers in the data. These outliers indicates that the flight to be delayed will concentrate around mean, which is 66.5 min.