1.

a) Total number of customers = (31 + 78 + 49 + 81 + 117 + 13) = 369
So, a total of 369 customers were served on this day at this gas station.

b) Table for class width and class midpoints,

Class boundary	Class width	Class midpoint
0 to <4	4	2
4 to <8	4	6
8 to <12	4	10
12 to <16	4	14
16 to <20	4	18
20 to <24	4	22

c) Table for relative frequency and percentage,

Class Boundaries	f	Relative frequency	Percentage
0 to <4	31	0.08	8.40
4 to <8	78	0.21	21.14
8 to <12	49	0.13	13.28
12 to <16	81	0.22	21.95
16 to <20	117	0.32	31.71
20 to <24	13	0.04	3.52
	Σf = 369	Σ = 1	Σ = 100

d) Total number of customers purchased 12 gallons or more = (81 + 117 + 13) = 211
So, the percentage =

$$\frac{211}{\sum f} \times 100\% = \frac{211}{369} \times 100\% = 57.18\%$$

e) Since, 10 is not a boundary value, therefore, we cannot determine exactly how many customers purchased 10 gallons or less.

2.

a) Table for class limits,

Class limit
1 to 200
201 to 400
401 to 600
601 to 800
801 to 1000
1001 to 1200

b) Table for class boundary and class midpoints,

Class	Class boundary	Class midpoint
1 to 200	0.5 to <200.5	100.5
201 to 400	200.5 to <400.5	300.5
401 to 600	400.5 to <600.5	500.5
601 to 800	600.5 to <800.5	700.5
801 to 1000	800.5 to <1000.5	900.5
1001 to 1200	1000.5 to <1200.5	1100.5

3.

a) Minimum value of data = 2.2
Maximum value of data = 22.8
Let's consider, class width = 5

So, number of classes =

$$\frac{22.8 - 2.2}{5} = 4.12 \approx 5$$

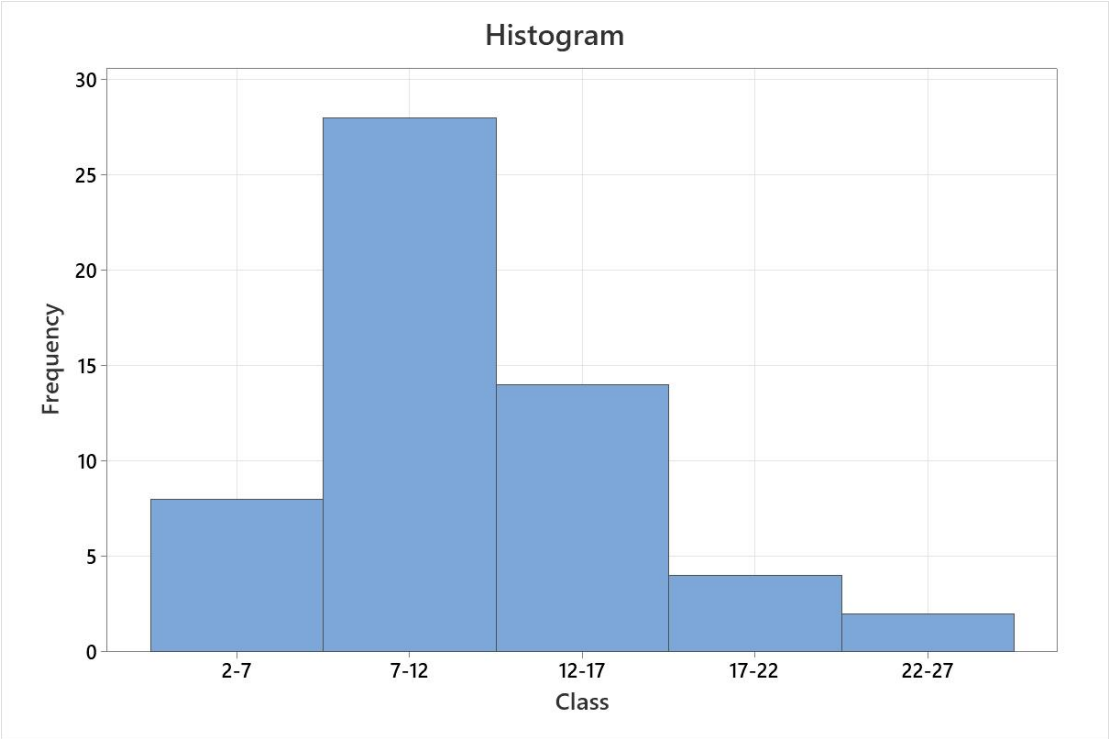
Table for frequency distribution,

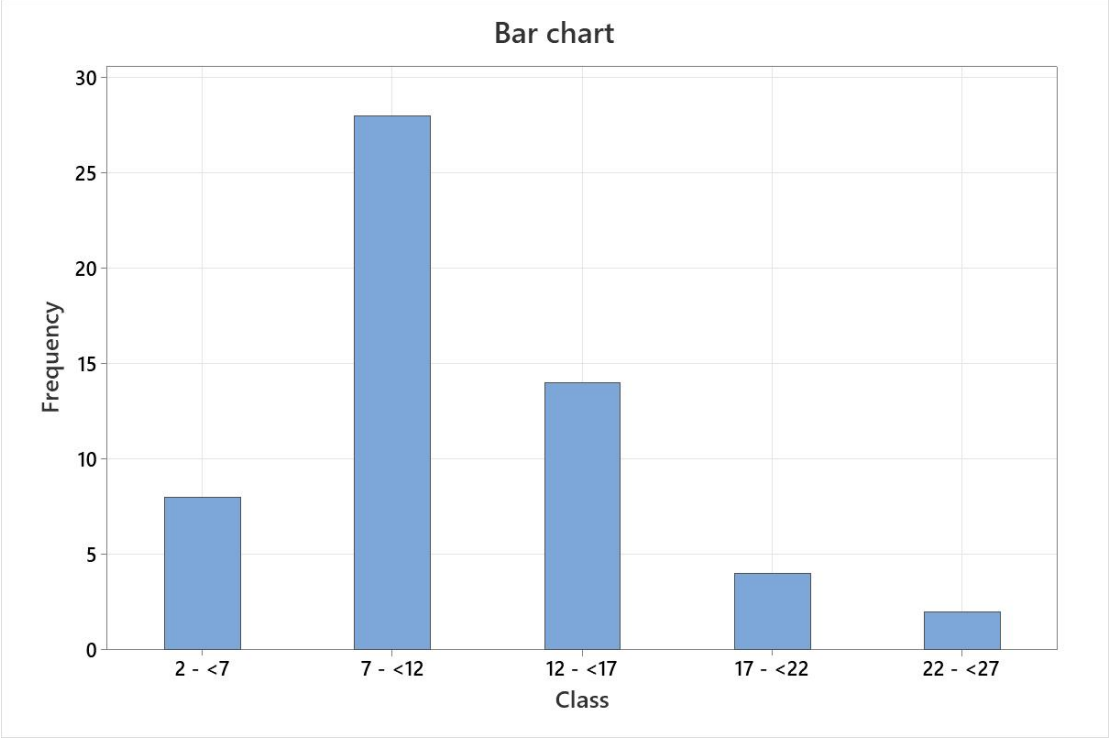
Class boundary	Tally	f
2 to <7	𝍸 |||	8
7 to <12	𝍸 𝍸 𝍸 𝍸 𝍸	28
12 to <17	|||	14
17 to <22	𝍸 𝍸 ||||	4
22 to <27	||||	2
	||	
		Σf = 56

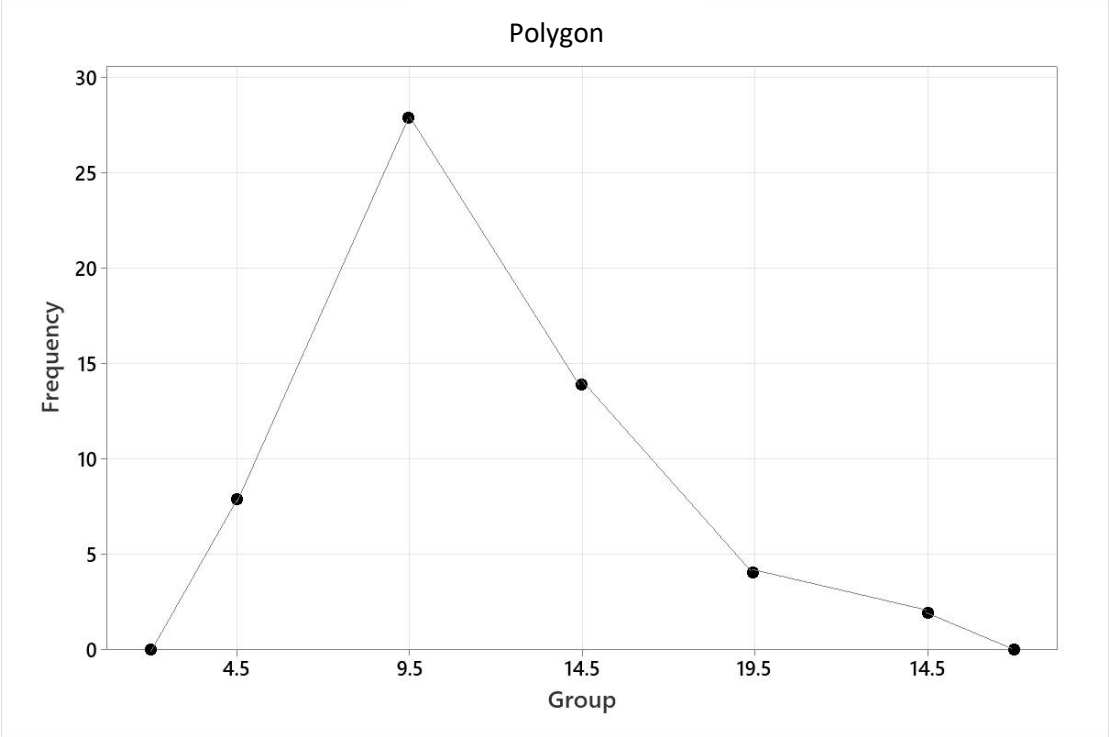
b) Table for relative frequency and percentage,

Class boundary	f	Relative frequency	Percentage
2 to <7	8	0.14	14.29
7 to <12	28	0.50	50.00
12 to <17	14	0.25	25.00
17 to <22	4	0.07	7.14
22 to <27	2	0.04	3.57
	Σf = 56	Σ = 1	Σ = 100

c) Histogram, bar diagram and a polygon for the birth rate percentage distribution,







d) Data less than 11 =
2.2 3.1 4.6 5.4 5.6 5.6 6.5 6.6 7.4 7.7 7.8 8.1 8.2 8.5 8.6 8.8 8.9 9.3 9.4 9.6 9.7 9.7 9.8 9.8 9.9 10.1 10.2 10.2 10.3 10.5 10.5 10.8 10.9 10.9

Number of data less than 11 = 34
Total frequency, Σf = 56

percentage of the countries had a birth rate of less than 11 births per 1000 people =

$$\frac{34}{56} \times 100\% = 60.71\%$$

4.
a) Smallest data = 1039
Largest data = 5490
Let's consider, class width = 1000
In one possible table,
Then 5 classes will hold all the data values if started from 1001.

Table for frequency (f) and cumulative frequency (F),

Class	f	F
1001-2000	8	8
2001-3000	3	11
3001-4000	0	11
4001-5000	0	11
5001-6000	1	12
	Σf = 12	

b)

$$mean = \frac{\sum f}{number\ of\ states} = \frac{21636}{12} = 1803$$

Data in ascending order =
1039 1086 1113 1137 1139 1166 1374 1673 2009 2110 2300 5490

Number of states, n = 12 , an even number

$$median = \frac{(\frac{n}{2})^{th}\ term\ +\ (\frac{n}{2}\ +1)^{th}\ term}{2}$$

$$= \frac{(\frac{12}{2})^{th}\ term\ +\ (\frac{12}{2}\ +1)^{th}\ term}{2}$$

$$= \frac{(6)^{th}\ term\ +\ (7)^{th}\ term}{2}$$

$$= \frac{1166}{1374}$$

$$= 1270$$

$$mode\ =\ None$$

c) From (b),
mean = 1803
median = 1270
The outlier in the data set = 5490

Data in ascending order after dropping the outlier =
1039 1086 1113 1137 1139 1166 1374 1673 2009 2110 2300

mean and median without the outlier;

$$mean = \frac{\sum f}{number\ of\ states} = \frac{16146}{11} = 1467.818$$

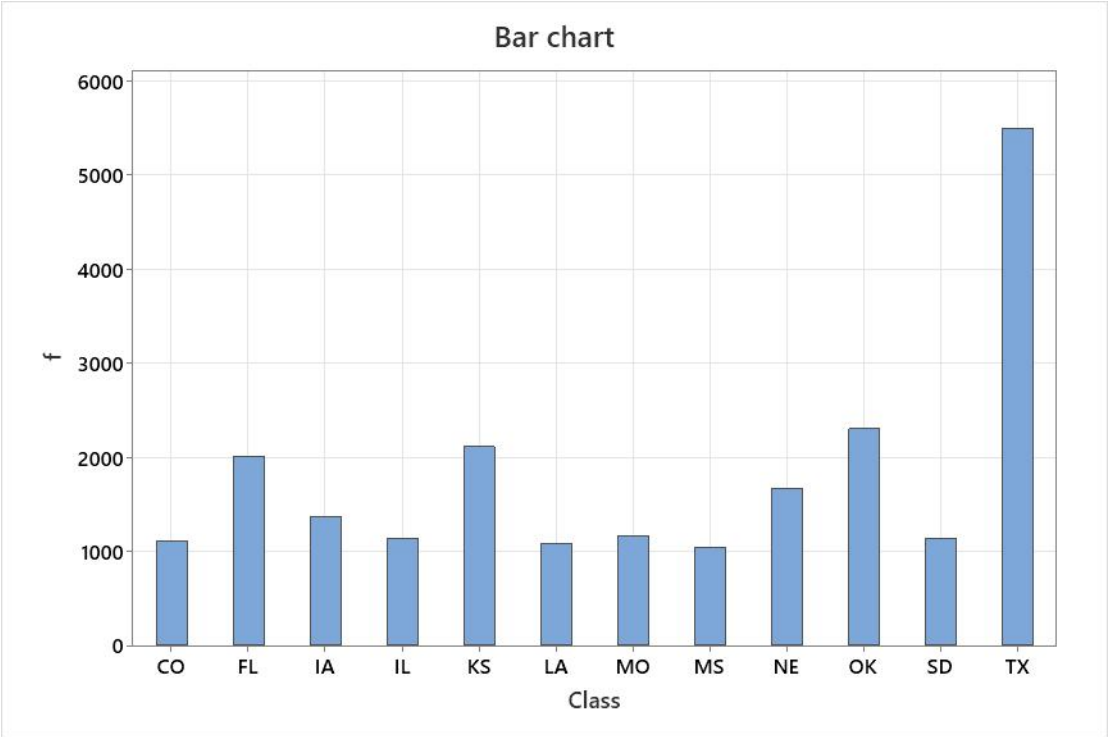
$$\begin{aligned} median &= (\frac{n+1}{2})^{th}\ term \\ &= (\frac{11+1}{2})^{th}\ term \\ &= (6)^{th}\ term \\ &= 1166 \end{aligned}$$

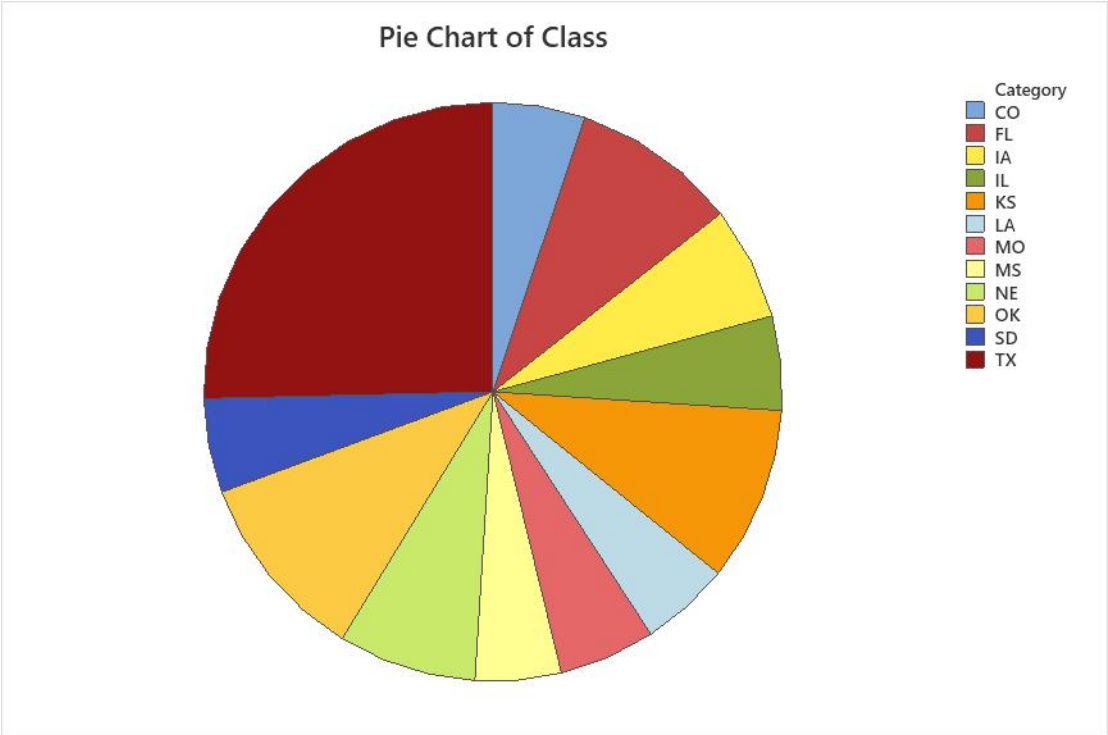
Difference in mean = 1803 - 1467.818 = 335.182
Difference in median = 1270 - 1166 = 104

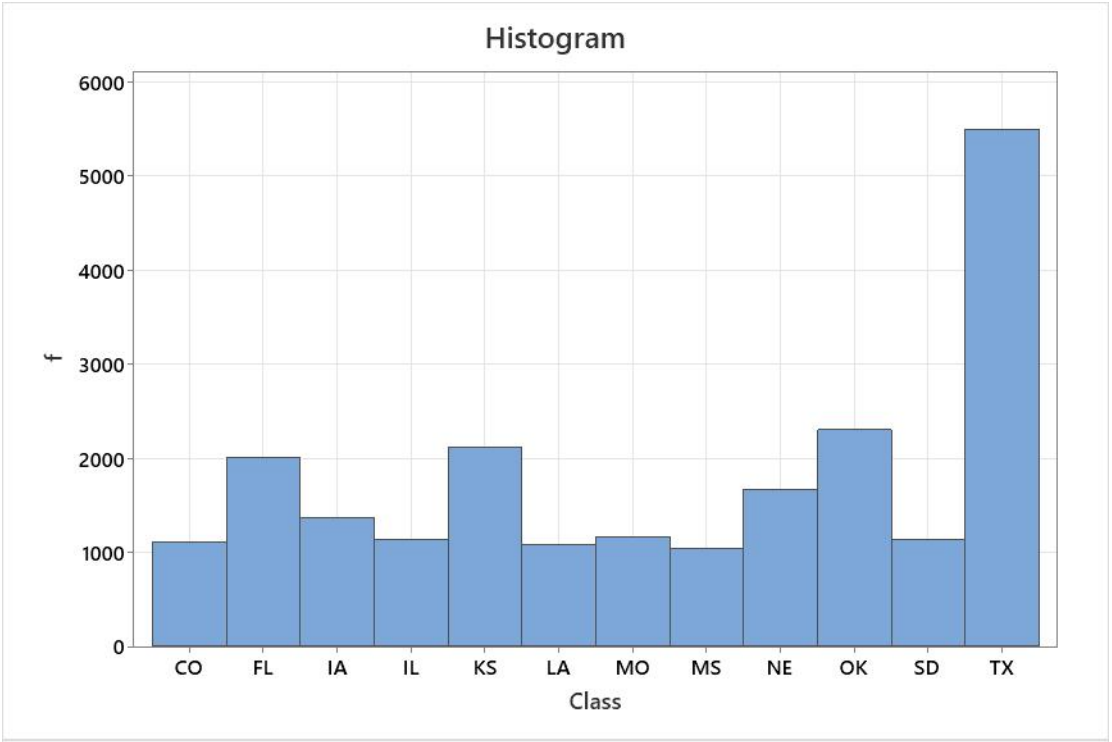
Therefore, mean is the summary measures changes by a larger amount when dropped the outlier.

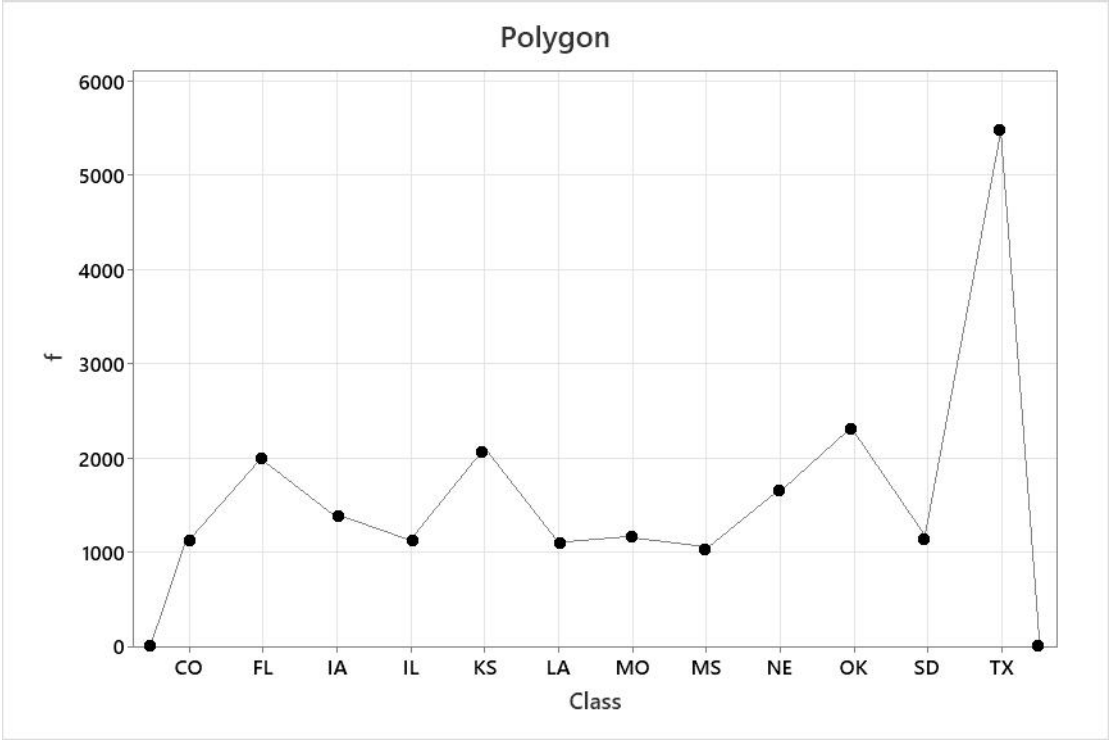
d) Since there is a significant outlier in the data set, therefore median is the better summary measure for these data than the mean and the mode.

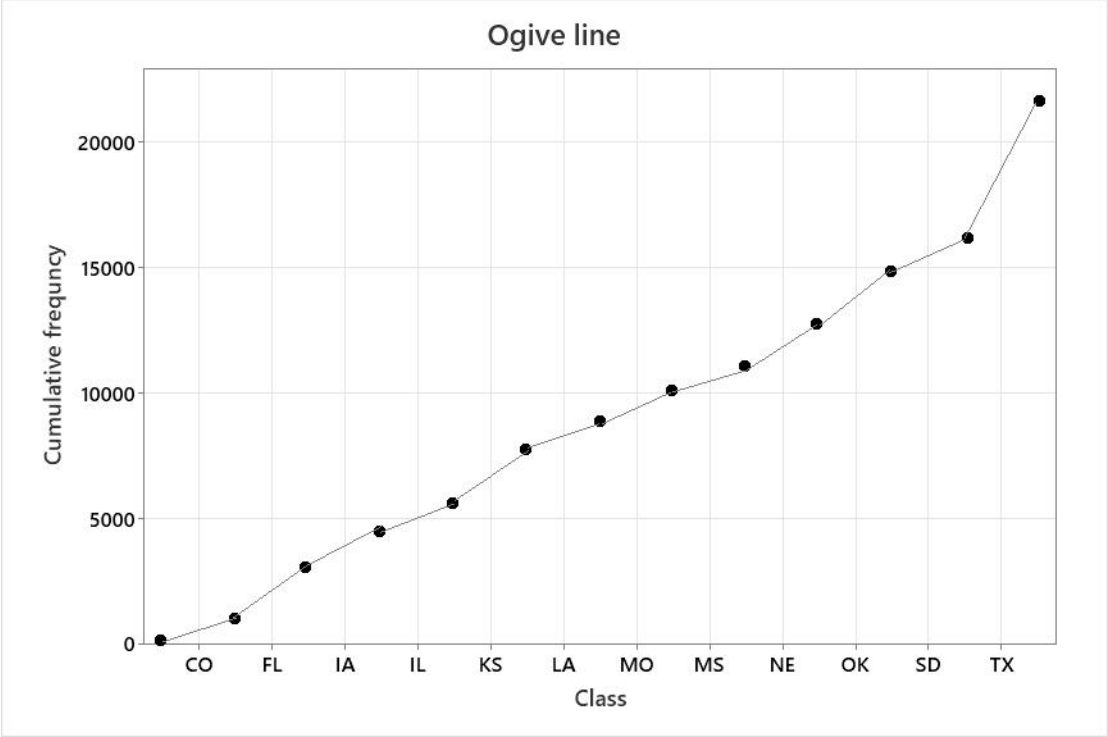
e) bar diagram, pie chart, histogram, and polygon, ogive line






Histogram
6000
5000
4000
3000
2000
1000
0
f
CO
FL
IA
IL
KS
LA
MO
MS
NE
OK
SD
TX
Class


Polygon
6000
5000
4000
3000
2000
1000
0
f
CO
FL
IA
IL
KS
LA
MO
MS
NE
OK
SD
TX
Class


Ogive line
Cumulative frequncy
0
5000
10000
15000
20000
CO
FL
IA
IL
KS
LA
MO
MS
NE
OK
SD
TX
Class

5.

a)

$$mean,\ \overline{x} = \frac{\sum x_i}{n} = \frac{7+10+8+3+15+12+6+11}{8} = 9$$

Table for deviations of the data values from the mean ($x_i - \overline{x}$),

x	$x - \overline{x}$
7	-2
10	1
8	-1
3	-6
15	6
12	3
6	-3
11	2
	Σ = 0

Yes, the sum of these deviations is zero.

b) Smallest data value = 3
Largest data value = 15

Range = Largest - Smallest = 15 - 3 = 12

Table for calculating variance,

x	x^2
7	49
10	100
8	64
3	9
15	225
12	144
6	36
11	121
Σx = 72	Σx^2 = 748

Number of sample, n = 8

variance, s^2 = $\dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$

$= \dfrac{748 - \frac{(72)^2}{8}}{8-1}$

$= 14.29$

standard deviation, s = $\sqrt{s^2}$

$= \sqrt{14.29}$

= 3.78

Data in ascending order,

3 6 7 8 10 11 12 15

Since, n = 8, which is divisible by 4

First quartile, Q_1 = $\dfrac{(n/4)^{th}\ term + ((n/4)+1)^{th}\ term}{2}$

= $\dfrac{(8/4)^{th}\ term + ((8/4)+1)^{th}\ term}{2}$

= $\dfrac{(2)^{th}\ term + (3)^{th}\ term}{2}$

= $\dfrac{6+7}{2}$

= 6.5

Second quartile, Q_2 = $\dfrac{(2n/4)^{th}\ term + ((2n/4)+1)^{th}\ term}{2}$

= $\dfrac{(2\times 8/4)^{th}\ term + ((2\times 8/4)+1)^{th}\ term}{2}$

= $\dfrac{(4)^{th}\ term + (5)^{th}\ term}{2}$

$$= \frac{8+10}{2}$$

= 9

Third quartile, Q_2 = $\frac{(3n/4)^{th}\ term + ((3n/4)+1)^{th}\ term}{2}$

$$= \frac{(3 \times 8/4)^{th}\ term + ((3 \times 8/4)+1)^{th}\ term}{2}$$

$$= \frac{(6)^{th}\ term + (7)^{th}\ term}{2}$$

$$= \frac{11+12}{2}$$

= 11.5

Since 50 × n = 50 × 8 = 400 ; is divisible by 100

50^{th} percentile, P_{50} = $\frac{(50n/100)^{th}\ term + ((50n/100)+1)^{th}\ term}{2}$

$$= \frac{(50 \times 8/100)^{th}\ term + ((50 \times 8/100)+1)^{th}\ term}{2}$$

$$= \frac{(4)^{th}\ term + (5)^{th}\ term}{2}$$

$$= \frac{8+10}{2}$$

= 9

6.

a) Table for determining mean,

Class	m	f	mf
10-20	15	5	75
20-30	25	8	200
30-40	35	12	420
40-50	45	7	315
50-60	55	4	220
		Σf = 36	Σmf = 1230

mean = $\frac{\sum mf}{\sum f}$

= $\frac{1230}{36}$

= 34.1667

b) Table for determining median,

Class	f	F
10-20	5	5
20-30	8	13
30-40	12	25
40-50	7	32
50-60	4	36
	Σf = 36	

median = $l + \frac{(n/2) - F}{f} \times h$

= $30 + \frac{(36/2) - 13}{12} \times 10$

= 34.1667

Here,
l = 30
n = Σf = 36
n/2 = 36/2 = 18
F = 13
f = 12
h = 20-10 = 10

c) Table for determining mode,

Class	f
10-20	5
20-30	8
30-40	12
40-50	7
50-60	4
	Σf = 36

mode = $l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

$= 30 + \frac{12 - 8}{2 \times 12 - 8 - 7} \times 10$

= 34.4444

Here,
l = 30
n = Σf = 36
n/2 = 36/2 = 18
F = 13
f = 12
h = 20-10 = 10

d) First quartile, Q_1 = $l_1 + \frac{h}{f_1}(\frac{n}{4} - F_{(1)-1})$

$= 20 + \frac{10}{8}(\frac{36}{4} - 5)$

= 25

Second quartile, Q_2 = $l_2 + \frac{h}{f_2}(\frac{2n}{4} - F_{(2)-1})$

$= 30 + \frac{10}{12}(\frac{2 \times 36}{4} - 13)$

= 34.1667

Third quartile, Q_3 = $l_3 + \frac{h}{f_3}(\frac{3n}{4} - F_{(3)-1})$

= $40 + \frac{10}{7}(\frac{3 \times 36}{4} - 25)$

= 42.8571

25th percentile, P_{25} = $l_{25} + \frac{h}{f_{25}}(\frac{25n}{100} - F_{(25)-1})$

= $20 + \frac{10}{8}(\frac{25 \times 36}{100} - 5)$

= 25

75th percentile, P_{75} = $l_{75} + \frac{h}{f_{75}}(\frac{75n}{100} - F_{(75)-1})$

= $40 + \frac{10}{7}(\frac{75 \times 36}{100} - 25)$

= 42.8571

e) Table for mean deviation and coefficient of mean deviation, range coefficient, quartiles deviation, variance, standard deviation,

Class	x	|x-mean|	f	f|x-mean|	f(x-mean)^2
10-20	15	19.17	5	95.85	1837.44
20-30	25	9.17	8	73.36	672.71
30-40	35	0.83	12	9.96	8.27
40-50	45	10.83	7	75.81	821.02
50-60	55	20.83	4	83.32	1735.56
			Σf = 36	Σ = 338.3	Σ = 5075

Here, mean = 34.17

mean deviation = $\frac{\sum f|x - mean|}{\sum f}$

= $\frac{338.3}{36}$

= 9.3972

Coefficient of mean deviation = (mean deviation) / mean

= 9.3972 / 34.17

=0.28

Range coefficient = $\frac{Mid-points\ of\ the\ Highest\ Class - Mid-points\ of\ the\ Lowest\ Class}{Mid-points\ of\ the\ Highest\ Class + Mid-points\ of\ the\ Lowest\ Class}$

= $\frac{55 - 15}{55 + 15}$

= 0.57

Quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

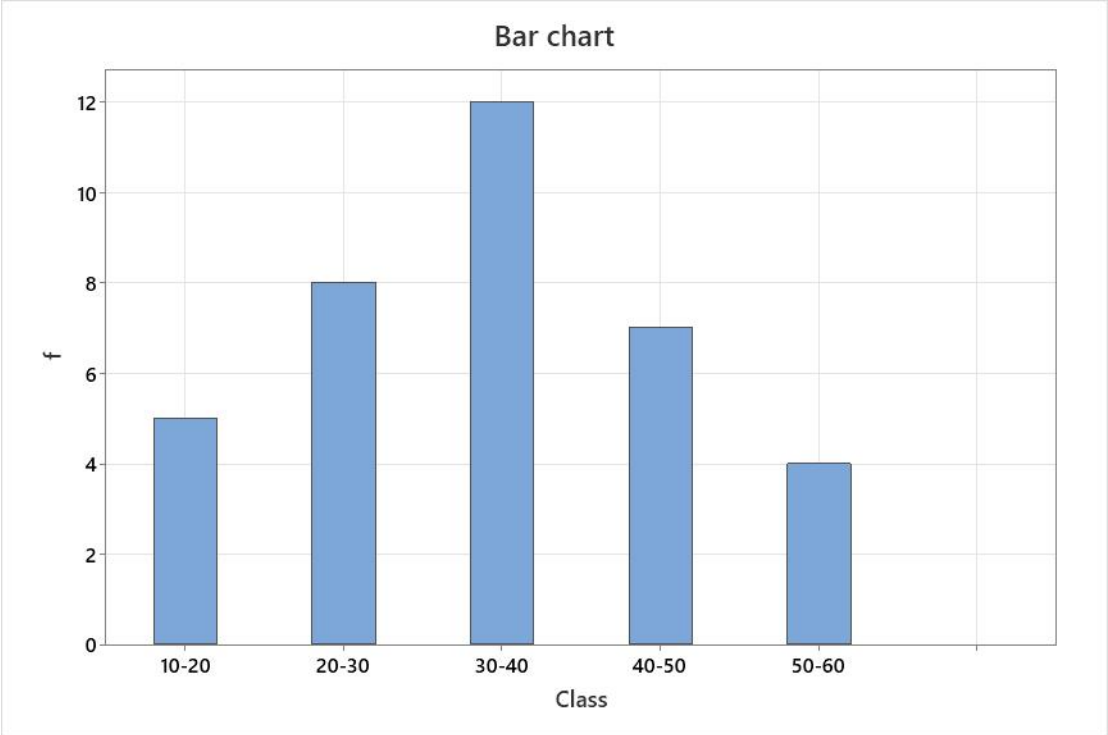
= $\frac{42.86 - 25}{42.86 + 25}$

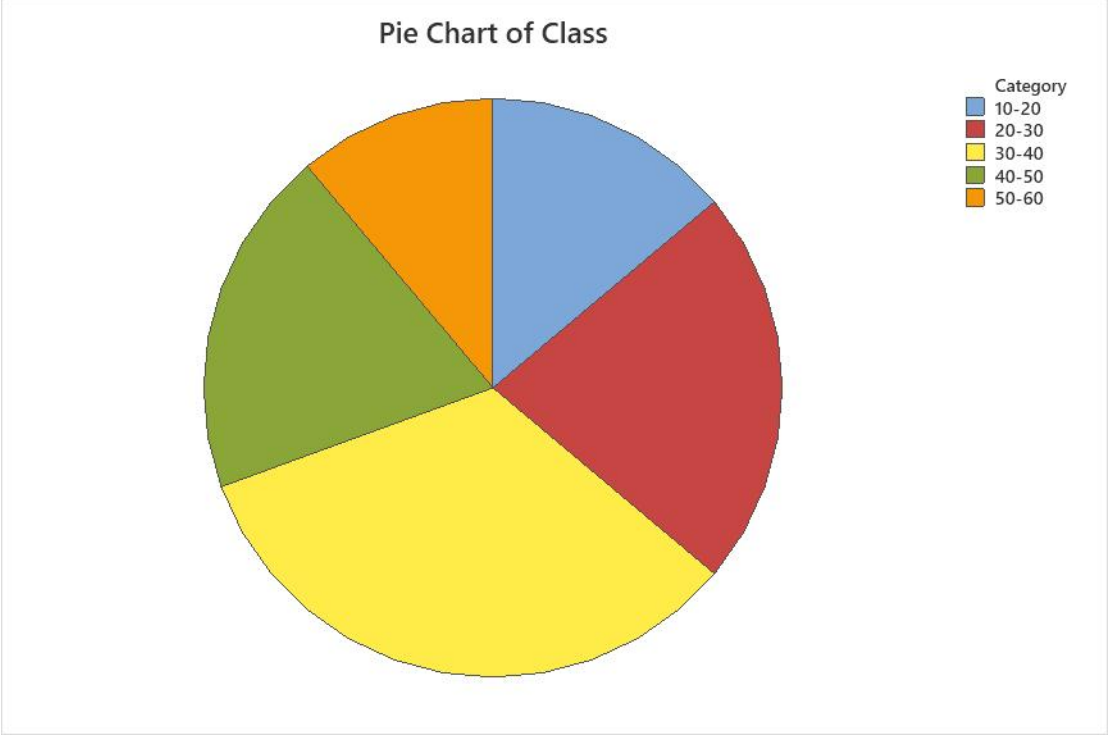
= 0.26

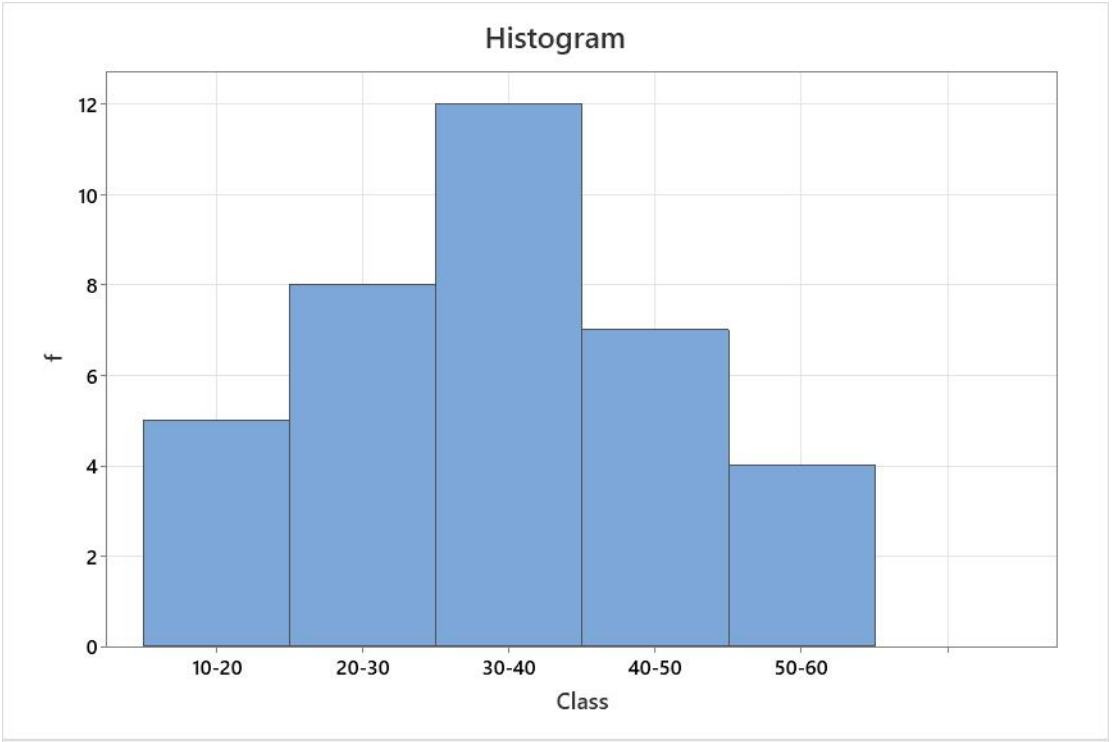
variance = $(\sum f(x - \text{mean})^{2)}/n$
= (5075)/36
= 140.97

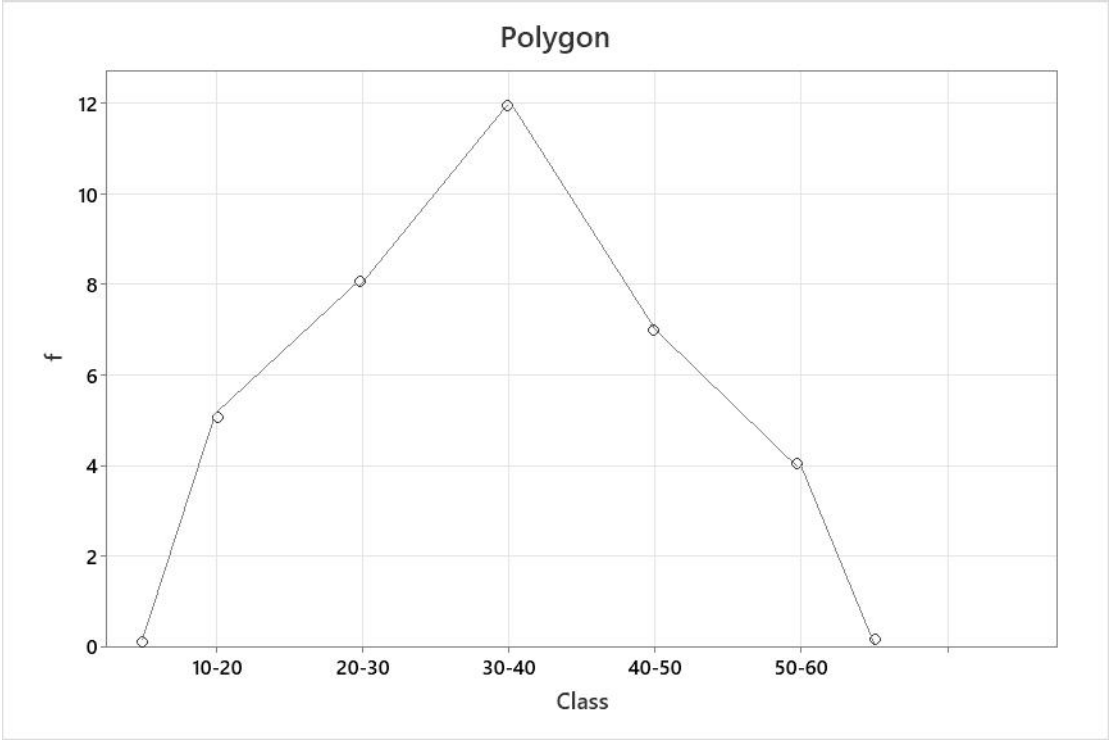
standard deviation = $\sqrt{variance}$
= $\sqrt{140.97}$
=11.87

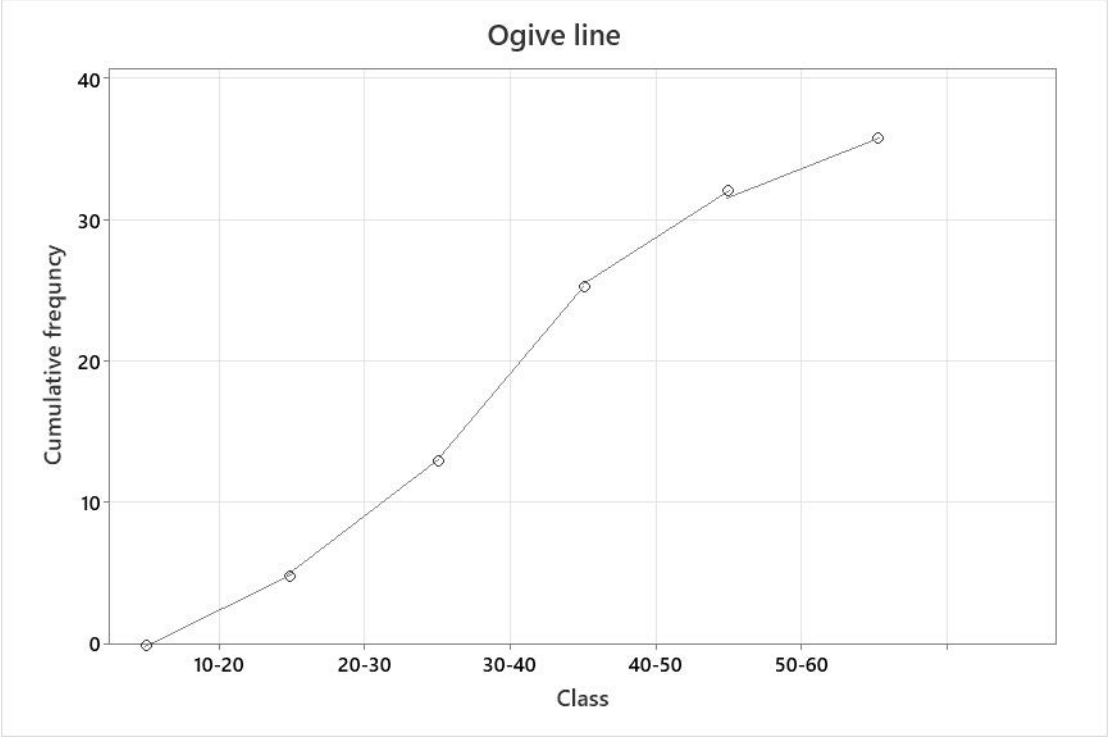
f) bar diagram, pie chart, histogram, and polygon, ogive line






Histogram
f
0
2
4
6
8
10
12
10-20
20-30
30-40
40-50
50-60
Class


Polygon
f
0
2
4
6
8
10
12
10-20
20-30
30-40
40-50
50-60
Class

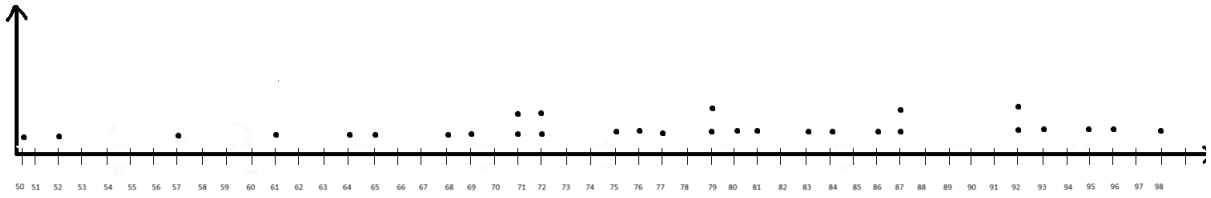

Ogive line
Cumulative frequncy
40
30
20
10
0
10-20
20-30
30-40
40-50
50-60
Class

7.

a)
Stem-and leaf display:

5	2 0 7
6	5 9 1 8 4
7	5 9 1 2 6 9 7 1 2
8	0 7 1 6 3 4 7
9	6 3 5 2 2 8

b)
Dot plot:



c)
The dot plot is more efficient for drawing conclusion about scattering, as we can easily visualize outliers, modes, minimum and maximum etc.

8.

Kurtosis is the degree of peakedness of a distribution, ususally taken in relation to a normal distribution. It quantifies how much the tails of a distribution differ from those of a normal distribution.
Here's how kurtosis relates to understanding a frequency ditribution:
i) Hight kurtosis indicates that the distribution has heavy tails and a sharp peak, suggesting that the data has more outliers compared to a normal distribution.

ii) Low kurtosis indicates that the distribution has lighter tails and is more flat-topped, suggesting that the data has fewer outliers compared to a normal distribution.

There are different measures of kurtosis, but the most commonly used is pearson's moment coefficient of kurtosis, denoted by β_2 .

a)
Given the first four moments of distribution,
$\mu'_1 = -1.5$
$\mu'_2 = 17$
$\mu'_3 = -30$
$\mu'_4 = 108$

So,
central moments,
$\mu_1 = 0$
$\mu_2 = \mu'_2 - \mu'^2_1 = 17-(-1.5)^2$
$= 14.75$
$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'^3_1$
$= -30 - 3(-1.5)(17) + 2(-1.5)^3$
$= 39.75$
$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'^2_1\mu'_2 + 3\mu'^4_1$
$= 108 - 4(-1.5)(-30) + 6(-1.5)^2(17) + 3(-1.5)^4$
$= 172.6875$

b)

$\beta_1 = \mu_3^2 / \mu_2^3$

$= (39.75)^2 / (14.75)^3$

$= 0.4924 > 0$, so, positively skewed

$\beta_2 = \mu_4 / \mu_2^2$

$= (172.6875)^2 / (14.75)^2$

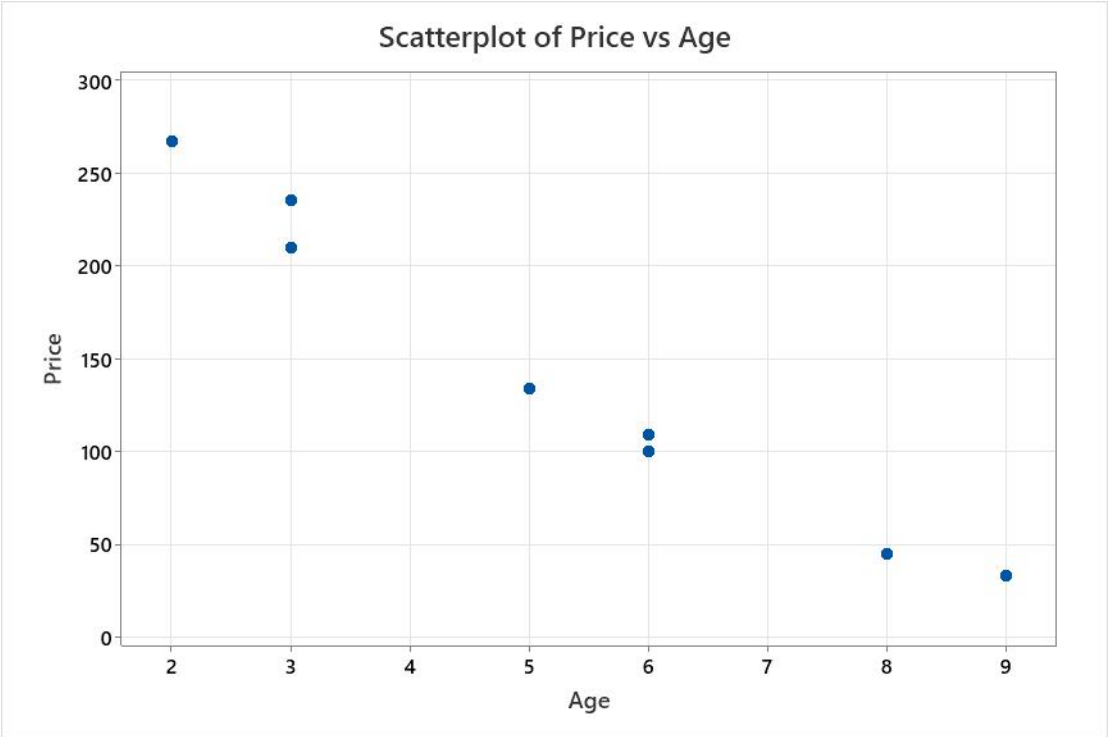
$= 0.7937 < 3$, so, platykurtic

c)

While skewness tells us about the asymmetry of the distribution kurtosis provides additional information about the shape of the distribution's tails. In this case, both measures are valuable for understandng the distribution's charecteristics with kurtosis indicating a distribution with heavier tails and positive skewness indicating that the mode > median > mean.

9.

a)



by looking at the scatter diagram , we can observe that there exists a strong linear relationship between car age and price. If a straight line is drawn through the points, the points will be scattered closely around the line.

b)

Age x	Price y	xy	x^2
8	45	360	64
3	210	630	9
6	100	600	36
9	33	297	81
2	267	534	4
5	134	670	25
6	109	654	36
3	235	705	9
Σx = 42	Σy = 1133	Σxy = 4450	Σx^2 = 264

So,

$\overline{x}$ = Σx / n = 42 / 8 = 5.25

$\overline{y}$ = Σy / n = 1133 / 8 = 141.625

Now,

ss_{xy} = Σxy - (Σx)(Σy) / n
= 4450 - (42)(1133) / 8
= -1498.25

ss_{xx} = Σx^2 - $(\Sigma x)^2$ / n
= 264 - $(42)^2$ / n
= 43.5

So,

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

$$= \frac{ss_{xy}}{ss_{xx}}$$

$$= \frac{-1498.25}{43.5}$$

$$= -34.4425287356$$

$$a = \overline{y} - b\overline{x}$$

= 141.625 - (-34.44)(5.25)
= 322.435

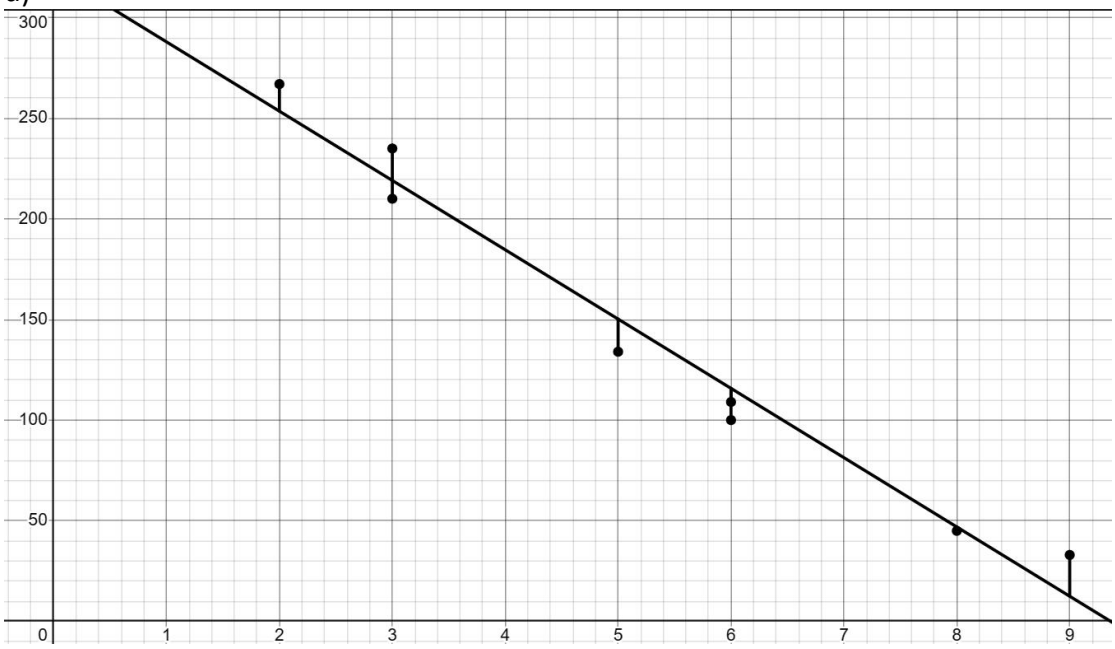
Thus the regression line estimated by employing least-square method is,

$$\hat{y} = 322.435 - 34.44x$$

c)
'b' represents an estimate of the average change in the value of the dependant variable y for each unit change in the independent variable x. In this particular case, the slope 'b' is negative which implies that as the age of cars (x) increases, the price decreases. So, the value b=-34.44 means that for an average increase of one year age of a car, the price would decrease on average by 34.44 hundred dollars.

The estimate of 'a' locates the regression line at point when x=0. Thus, if the age of a car is 0 years aka brand new, the average increase in price will be almost 322.435 hundred dollar.

d)



Graph of scatterplot and regression line and error shown with vertical lines.

Table for correlation coefficient,

Age x	Price y	xy	x^2	y^2
8	45	360	64	2025
3	210	630	9	44100
6	100	600	36	10000
9	33	297	81	1089
2	267	534	4	71289
5	134	670	25	17956
6	109	654	36	11881
3	235	705	9	55225
Σx = 42	Σy = 1133	Σxy = 4450	Σx^2 = 264	Σy^2 = 213565

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

$$= \frac{4450 - \frac{(42)(1133)}{8}}{\sqrt{264 - \frac{(42^2)}{8}}\sqrt{213565 - \frac{(1133^2)}{8}}}$$

= - 0.986 , value close to -1 indicates strong linear relationship with negative slope.

Since the regression model, $\hat{y} = 322.435 - 34.44x$
Regression coefficient = -34.44
Which indicates negative slope and also means that for an average increase of one unit in the independent variable, the value of dependant variable would decrease on average by 34.44 unit.

Therefore, data is not perfectly scattered.

e)
The best prediction of the price of a 7-year-old car of this model is,

$\hat{y}_{(7)} = 322.435 - 34.44(7)$
= 81.355

So, 81.355 hundreds of dollars or $8135.5

f)
The best prediction of the price of a 18-year-old car of this model is,

$\hat{y}_{(18)} = 322.435 - 34.44(18)$
= -297.485

But, price can’t be negative.

10.

a)

Figure 1: obtained from MINITAB

b)

The regression coefficients are 15.07, 0.1669 and − 0.1323

The value of a = 15.07 gives the value for ˆ y when x1 = 0 and x2 = 0. However, since x1 = 0 and x2 = 0 do not occur together in the sample data, the estimate is invalid.
The value b1 = 0.1669 gives the change in ˆ y for a one-unit change in x1 when x2 is held constant.
The value b2 = −0.1323gives the change in ˆ y for a one-unit change in x2 when x1 is held constant.

c)

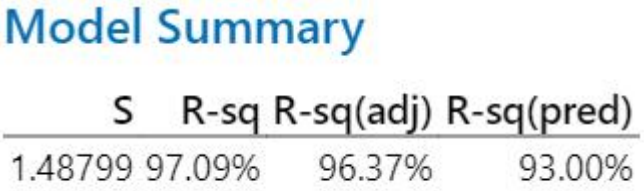


Figure 2: obtained from MINITAB

standard deviation of errors, s_e = 1.488
the coefficient of multiple determination, R^2 = 0.9709
the adjusted coefficient of multiple determination, $\overline{R}^2$ = 0.9637

d)

y^ = 15.07 + (0.1669)x1 - (0.1323)x2 = 15.07 + .1669(87) − .1323(54) = 22.4461

e)

y^ = 15.07 + (0.1669)x1 - (0.1323)x2 = 15.07 + .1669(95) − .1323(49) = 24.4428

f)
df = n − k − 1 = 11 − 2 − 1 = 8
The 99% confidence interval for B1 is
$b_1 \pm t\, s_{b1}$ = .167 ± (3.355)(.034) = .167 ± .114 = .053 to .281

g)
H_0: B2 = 0, H_1: $B_2 < 0$
Since $\sigma_{\in}$ is unknown, use the t distribution.
For α = .01 with df = 8, the critical value of t is −2.896.
$t = (b_2 - B_2)/ s_{b2}$ = −1.919
Do not reject H_0 since −1.919 > −2.896.
So, B_2 is not negative.