Problem no=1

Problem name:

The information of 20 Person is given in the following table:

SN	Sex	level of education	Religion	SN	Sex	Religion	Level of education
1	Male	Primary	muslim	11	Female	Hindu	Primary
2	Female	Graduate	Hindu	12	male	christian	Graduate
3	male	Illitarate	muslim	13	male	Others	Secondary
4	male	Graduate	Hindu	14	Female	muslim	Secondary
5	Female	Primary	muslim	15	male	Hindu	Higher Secondary
6	Female	Graduate	muslim	16	male	christian	Others
7	male	Primary	Hindu	17	Female	muslim	Primary
8	Male	Illitarate	muslim	18	male	Others	Illiterate
9	Female	Others	Hinda	19	Female	muslim	Secondary
10	Male	Higher Secondary	Others	20	male	others	Secondary

i) Construct the frequency distribution for variables Religion and Level of education.

ii) Draw pie diagram for the variable "Religion and comment

iii) Draw bar diagram for the variable "level of Education and comment

Theory:

Frequency Distribution: A frequency distribution is a set of mutually exclusive classes on categories together with the frequency of occurrence of items, values on observation in each class on category in a given set of data Presented usually in a tabular form.

Pie Diagram: Pie diagram, also known as pie chart, is a useful device for Presenting Catagorical data. Data other than catagorical can also be employed for constructing pie diagram after suitable and meaningful classification on grouping of the data.

Bar Diagram:

A bar diagram also known as bar chart, is from of Presentation in which the frequency are represented by rectangles usually along the axis.

Procedure:

i) First we open the MS Excel and entry the data

ii) Then we open the Insert menu and we take the Pivot Table. Finally we get frequency distribution.

iii) Then we select the charts. The charts is two dimentional.

iv) In this chart option we select Pie chat

v) Last of all we select Bar chat in the chat option

Figure :

i) Frequency distribution for Relegion

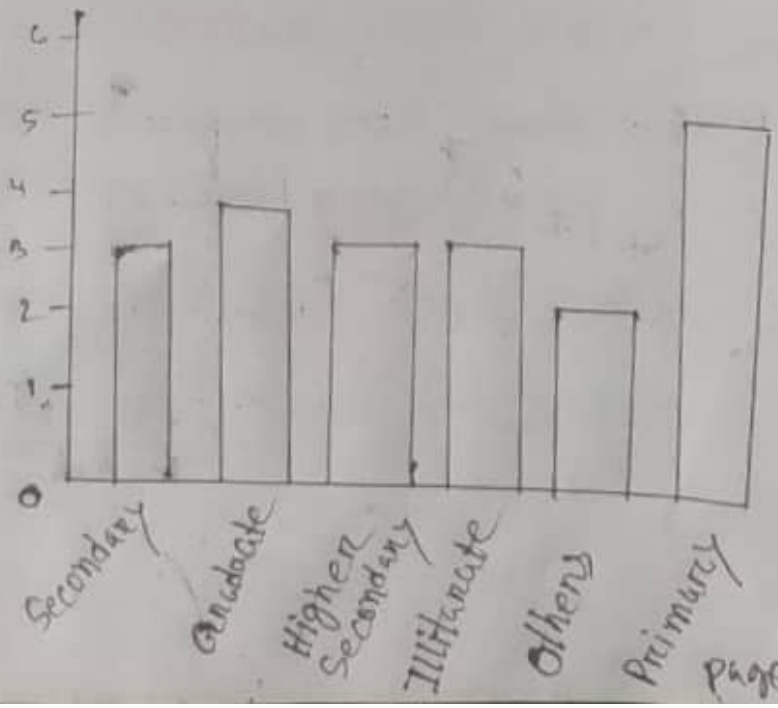
Relegion	Count of Relegion
christian	3
Hindu	4
Muslim	8
other	5
Grand total	20

ii) Frequency distribution for Level of education

Row labels	count
Secondary	3
Graduate	4
Higher Secondary	3
Illitarate	3
Others	2
Primary	5
total	20

ii) Pie diagram for Religion

iii) Bar diagram for level of education



Problem no=2

Problem name: Suppose that 100 students are enrolled in a statistics class and the following are the test scores received by them.

77	44	49	33	38	33	76	55	68	39	44	59	36	55	47	61	53	32
29	41	32	45	83	58	73	47	40	26	59	43	66	44	25	39	72	37
34	47	66	53	55	58	49	45	61	41	55	92	83	77	62	45	36	78
54	50	51	66	80	73	57	61	56	50	45	82	71	48	69	38	72	51
38	45	51	44	41	68	45	92	43	12	37	16	44	57	71	40	64	57

i) Compute Mean, Median, Mode, Variance and standard deviation of the above raw data and comment on your results.

ii) Find the five number Summaries.

iii) Select an appropriate class interval and organize the data set into a frequency distribution

iv) using the frequency distribution obtained in question (iii) construct a histogram and an ogive. Also approximate the median and mode with the help of ogive and histogram respectively.

v) Find the mean, median, and mode using the frequency distribution obtained in question (iii)

Theory:

Mean: There are three type of mean. There are
i) Arithmetic mean
ii) Geometric mean and
iii) Harmonic mean

Median: Median is the middle value of an array or a series, which divides the array into two equal parts half of the observations are above it and half of the observation are below it.

Arrange the series in ascending or descending order.

If the number of observation (n) is odd, then the formula of median is

median = Value of $\left(\frac{n+1}{2}\right)$th observation

If the number of observation (n) is even, then the formula of median is

median = Value of $\frac{1}{2}\left(\frac{n}{2}\right)$th observation + $\left(\frac{n}{2}+1\right)$th observation.

Mode: The value that occurs most often in a data set is called the mode.

A data set that has only one value that occurs with the greatest frequency is said to be unimodal.

When no data value occurs more than once, the data set is said to have no mode.

A data set can have more than one mode or no mode at all.

Variance: Variance is the arithmetic mean of the squared deviations from mean of the distribution. Mathematically, variance is

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{N}$$

It is an absolute measure of dispersion. Variance is not pure number.

Procedure:

i) At first we open the MS Excel file.

ii) Then we entry all data

iii) Then we use =AVERAGE(A1:T5) formula for calculate mean.

iv) we use =MEDIAN(A1:T5) formula for calculate median

v) we use @MODE.MULT(A1:T5) formula for calculate mode.

vi) Similarly we calculate variance, standard deviation and Quartile.

Figure :

i) Mean : 52.68
Median : 51
Mode : 45
Variance : 246.377
Standard deviation : 15.696

ii) Five number summerie
Minimum : 12
1st Quartile : 42.5
2nd Quartile : 51
3rd Quartile : 63.25
Maximum : 92

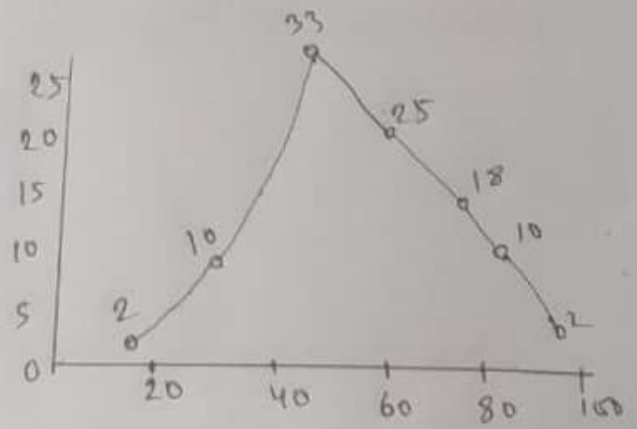
iii) class interval 12. We construct the frequency distribution table:

class limits	Bin	Frequency
12-24	24	2
24-36	36	10
36-48	48	33
48-60	60	25
60-72	72	18
72-84	84	10
84-96	96	2
Total		100

iv) Histogram and Ogive:

Bin	Frequency
24	2
36	10
48	33
60	25
72	18
84	10
96	2

Histogram



Problem no = 3

Problem name: The following data represents the ages of the 50 richest People in the world in 2009.

89, 89, 87, 86, 85, 83, 82, 81, 80, 78, 78, 77, 76, 73, 73, 73, 72, 69, 69, 68, 67, 66, 66, 65, 65, 64, 63, 61, 61, 60, 60, 59, 58, 57, 56, 54, 53, 53, 51, 49, 47, 46, 44, 43, 42, 36, 2000.

i) Find the mean, median, and mode of the ages of the 50 richest People. which measures of central tendency best describes a typical entry of this data set.

ii) Replace 35 instead of 2000 from the data set then rework (i). Compare these measures of central tendency with those found in (i)

iii) Construct a frequency distribution using the above data after replacing 35 instead of 2000.

iv) Construct a relative frequency histogram.

v) Find the mean, the median, the mode and the variance for grouped data. Comment on the results in the context of the data.

Theory:

Mean: There are three type of mean. There are

i) Arithmetic mean
ii) Geometric mean
iii) Harmonic mean

Median: Median is the middle value of an array or a series which divides the array into two equal parts half of the observations are above it.

Mode: The value that occurs most often in a data set is called the mode. A data set that has only one value that occurs with the greatest frequency is said to be unimodal

Variance: Variance is the arithmetic mean of the squared deviations from mean of the distribution. Mathematically, variance is

$$\sigma^2 = \sum_{i=1}^{k} \frac{f_i(x_i - \bar{x})^2}{N}$$

It is an absolute measure of dispersion. Variance is not pure number.

Histogram: The most common form of graphical Presentation of a frequency distribution is the histogram. A histogram is constructed by placing the class boundaries on the horizontal axis of a graph and the frequencies on the vertical axis.

Procedure:

i) At first we open the MS Excel and entry all the data.

ii) we use =AVERAGE(A2:A51) formula for calculate mean.

iii) we use =MEDIAN(A2:A51) formula for calculate median.

iv) we use =@MODE.MULT(B2:B51)

Figure : i) Mean : 104.56
Median : 66
Mode : 73

ii) Mean : 65.26
Median : 65.5
Mode : 73

iii) Minimum : 35
Maximum : 89

iv) Relative frequency histogram

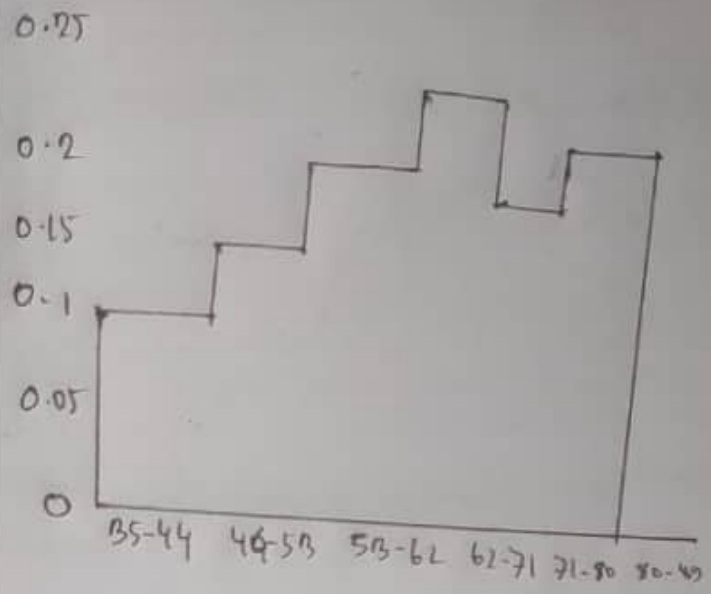
class Limit	Bin	Frequency
35-44	44	5
44-53	53	7
53-62	62	9
62-71	71	10
71-80	80	9
80-89	89	10

class Limit	Frequency	RF
35-44	5	0.1
44-53	7	0.14
53-62	9	0.18
62-71	10	0.2
71-80	9	0.18
80-89	10	0.2
Total	50	1

(v)

class Limit	Frequency	mid value	fx	cf
35-44	5	39.5	197.5	5
44-53	7	"	"	"
53-62	9	"	"	"
62-71	10	66.5	665	31
71-80	9	"	"	"
80-89	10	84.5	845	50
Total	50		3244	

class Limit	RF
35-44	0.1
44-53	0.14
53-62	0.18
62-71	0.2
71-80	0.18
80-89	0.2
Total	1



Mean: 64.88

Median: 65.6

l	62
N/2	25
f	10
F	21
i	9

Problem no=4

Problem name: The grade point average (GPA) in different Semesters of two students are shown below

Student	GPA in Semesters							
	1	2	3	4	5	6	7	8
A	2.5	2.5	3.0	3.5	3.5	4.0	3.5	3.5
B	2.5	3.0	4.0	4.0	4.0	2.0	2.5	4.0

which students would you consider better throughout the course of studies:

Theory:

Standard deviation: The standard deviation is the square root of the variance.

The symbol for the population standard deviation is σ.

The formula of Population standard deviation is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}}$$

The coeffecient of variation: A statistics that allows you to compare standard deviations when the units are differen as in this example is called the coeffeci ent of variation.

The coeffecient of variation formula is $CV = \frac{b}{\bar{x}} \times 100$

Procedure:

i) At first we open the MS Excel

ii) Then we enter the all data appropriately

iii) we use formula = AVERAGE (A7:A14)

iv) We use formula = STDEV.P (A7:A14) for standand deviation

v) Similarly we calculate CV

Figure :

XA	XB	$(XA-MA)^2$	$(XB-MB)^2$
2.5	2.5	0.5625	0.5625
2.5	3	0.5625	1
3	4	1	1
3.5	4	1	0.5625
3.5	4	0.0625	1
4	2	1	1
3.5	2.5	1	1
3.5	4	0.0625	0.5625
26	26	2	5

Mean, MA : 3.25

Mean, MB : 3.25

STD, SA : 0.5 STD, SB : 0.7905694

CV(A) : 15.385 CV(B) : 24.325213

Problem no = 5

Problem name: If X follows binomial distribution with $n = 50$ and $p = 0.6/0.5/0.2$.

i) Sketch the graph for binomial Probability distribution?

ii) Compute

a. $P(X = 35)$

b. $P(x \leq 20)$

c. $P(x \geq 15)$

d. $P(20 < x < 45)$

iii) Find first four central moment of the distribution.

iv) Find the skewness and kurtosis of the distribution.

Theory: A discreate random variable x is said to have a binomial distribution if it has a Probability function with

$$P(x; n, p) = {}^{n}c_{x} \cdot p^{x} \cdot (1-p)^{n-x}$$

$$= {}^{n}c_{x} \cdot p^{x} \cdot q^{n-x} \qquad x = 0, 1, 2 \cdots n$$

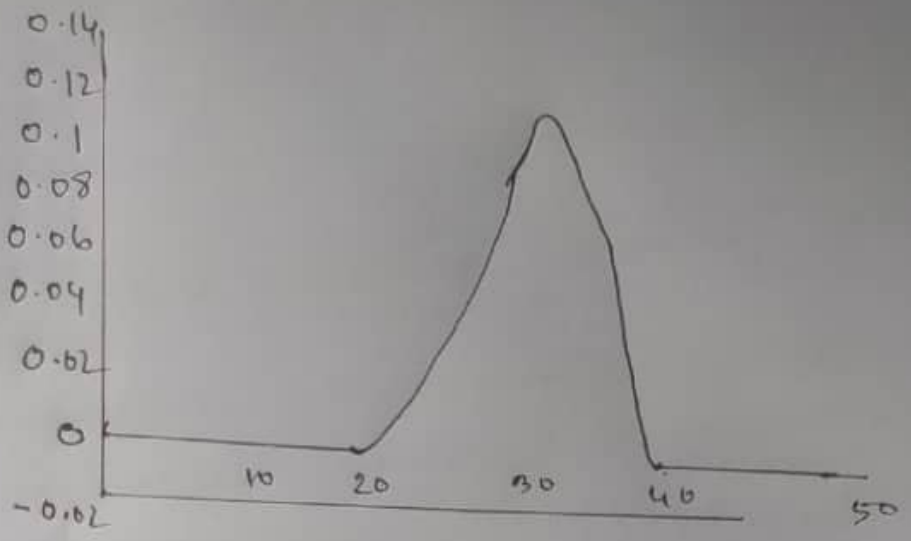
where $q = 1-p$ and ${}^{n}c_{x} = \binom{n}{x}$

n = number of trials

p = Probability of Success and

$0 \leq p \leq 1$

Figure: i)



ii) P(x=35) : 0.04154

P(x<=20) : 0.00336087

P(x>15) : 0.99998

P(20<x<45 : 0.99663

iii)

X	P(X)	X · P(X)	M1	M2	M3	M4
0	1.26765E-20	0	-3.80295E	1.1408	-3.4226	1.026
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
15	1.249435E-05	0.00782941	-0.0001842	0.000924	-0.00660	
1	1					
1	1					
1	1					
50	4.04263E-10	1.61652E-10	1.626345E-10	6.46626	1.2935	
Total		30	5.4632E-17	12	-2.4	

iv) skewness, gamma1 -0.057735

kuntosis, beta2 2.96333 platykurt

Problem no=7

Problem name: If $Z \sim N(0,1)$. For the following values of Z.

$-4, -3.9, -3.8 \ldots\ldots 3.8, 3.9, 4$.

i) Create pdf of Z. Draw standard normal curve and comment the shape characteristics of the distribution.

ii) Create pdf and cdf of $X \sim N(1000, 250000)$.

iii) Find

a) $P(X = 850)$

b) $P(X > 1200)$

c) $P(1000 < X < 2000)$

iv) Construct normal density curve and normal cumulative distribution curve. Comment on your results.

about the ordinates. The area under the normal curve representing proportion ate frequency is one.

Procedure: We create open the excel file and create a new excel document

ii) The code we provided excel file to be already created with matching column headers

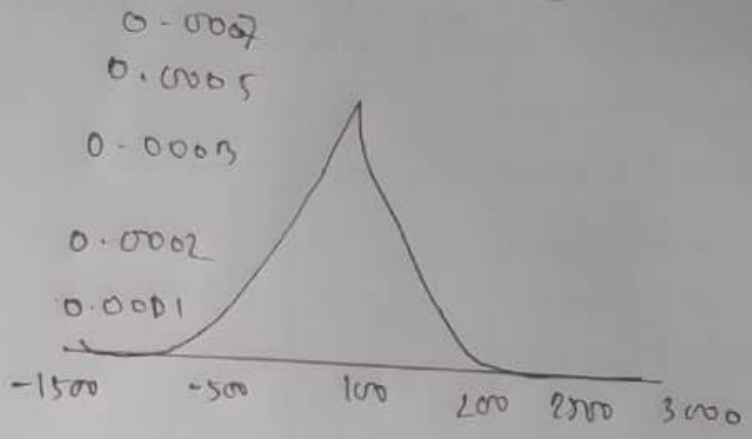
iii) We create the file and include the value and finally calculate the total value.

iii) $P(x = 850) = 0.000762$

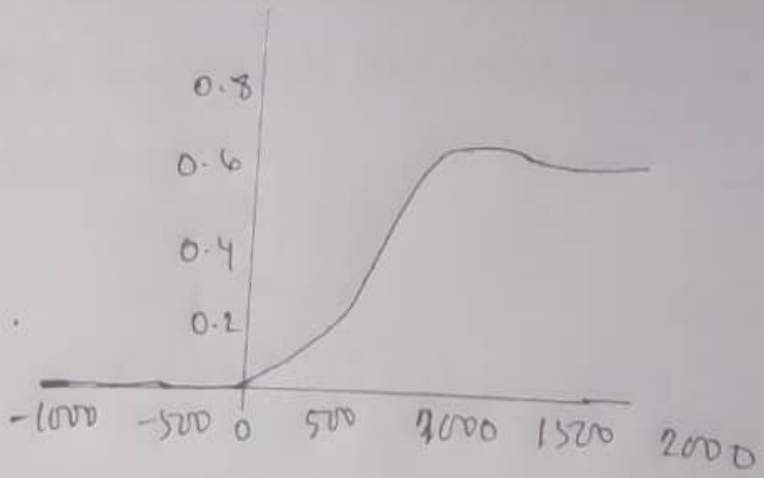
$P(x > 1200) = 0.006525$

$P(1000 < x < 1200) = 0.009090$

iv) Normal Density curve



Normal Cumulative distribution curve



Problem no = 8

Problem name: The following data gives the number of printing mistakes in a book of five hundred pages.

No. of Printing mistake	No of page	No of Printing mistake	No. of page
0	150	9	18
1	37	10	17
2	34	11	16
3	30	12	14
4	28	13	11
5	25	14	9
6	24	15	8
7	22	16	7
8	21	17 and above	5

i. Finst position distribution to the above data.

ii. sketch the graph for Posisson distribution.

iii. Compute

a. $P(x = 15)$

b. $P(x < 10)$

c. $P(x > 15)$

d. $P(4 < x < 17)$

iv) Find the skewness and kuntosis of the distribution.

v) Find the Probability of position distribution using recurrence relation

Theory: Poission distribution is one of the important families of Probability distributions named atten a french mathematician Simeon Renis Poisson, who discovered in 1837. It is also known as the law of small numbers. There are many appnoaches to study the poisson Process and poisson distribution The simplest way to obtain the poisson distribution is an approximation to the binomial distribution in whene

$$P(x) = \frac{e^{-m} \cdot m^x}{x!}$$

A poisson distribution is assumed to the following the occurance of the events in an interval of space on time is independent on the Probability of a Second occurances of the event in any othen disjoint time interval.

Procedure:

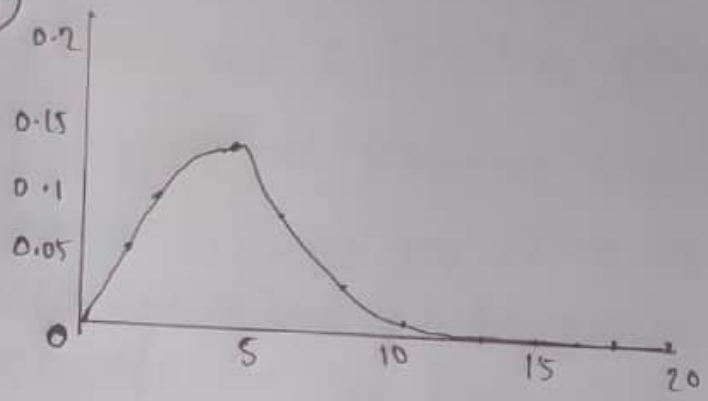
i) We create open the excel file and create a new excel document.

ii) The code we provided excel file to be already created with matching column headers

iii) We create the file include the value and finding calculate the total value.

NO. of Printing mistke	NO. of Page	fx	P(x)	P(x)R
0	150	0	0.066461	0.006461
1	1	1	1	1
1	1	1	1	1
15	9	135	0.001412	0.018665
1	1	1	1	1
20	4	76	1.19E-06	1.19E-06

ii)



iii) $P(X=15) = 0.000170935$

$P(X<10) = 45$

$P(X>15) = 7.55043E-05$

$P(5<X<17) = 0.56682$

Problem no=9

Problem name: Hypothetical data on weekly family Consumption expenditure (Y) and weekly family income (X) are given below:

X	70	65	90	95	110	115	120	140	155	150
Y	80	100	120	140	160	180	200	220	240	260

i) Construct a Scatter plot of the weekly family Consumption expenditure Y and weekly family income (X). Do you think are any relationship between X and Y? Do you think a linear model is appropriate for this data?

ii) Find the coefficient of correlation between weekly family Consumption expenditure (Y) and weekly family income (X) and Comment on your result.

iii) obtain the line of best fit for Y on X.

iv) How do you interpret the intercept and the slope of the regression line?

vi) check the goodness of fit of the least squares fit.

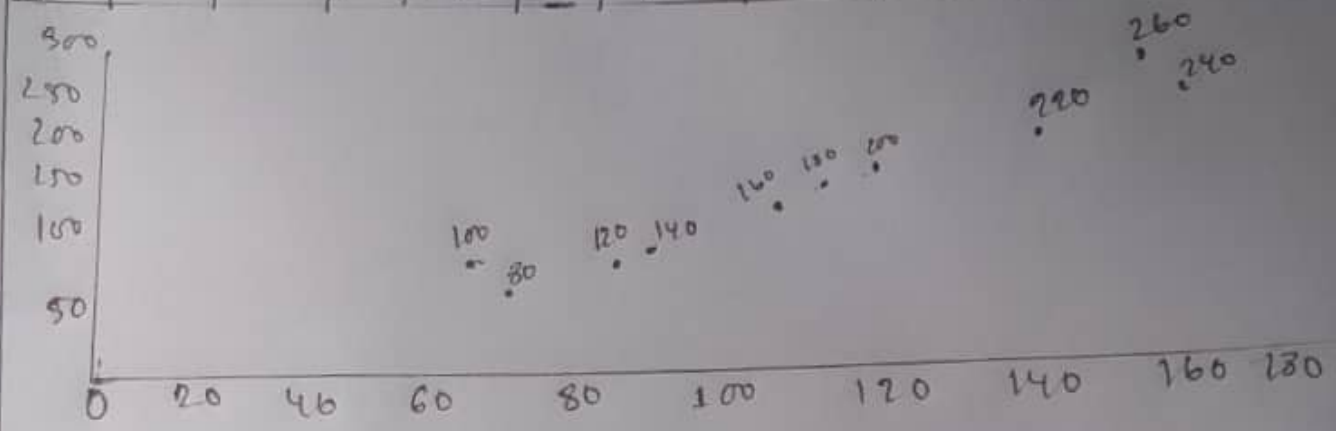
Theory:

Linear Regression: The inherent complexity of most real word problems suggests that we can more accurately describe Predict and control an outcome variable by using a regression model that employs more than one independent variable. Such a model is called a multiple regression model in constrast to linear regression model.

Correlation: The connelation coetficient is symbolized by R which shows the connelation among more than two variables. As with r^2, R^2 indicates the Proportion of variance in the connelation.

Figure :

Y	70	65	90	95	110	115	120	140	155	150
X	80	100	120	140	160	180	200	220	240	260



ii) Co. coo, R = 0.9808473

iii) Summary output

Multiple R	0.98087
R Square	0.9620
Adjusted	0.957314

iv) Y175 113.5454545