Section 1.2 Worksheet – Organizing and Displaying Categorical Data

MDM4U
Jensen

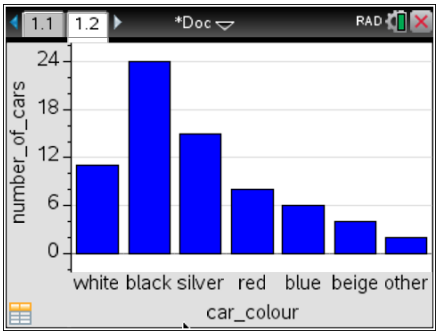
Refer to part 1, 2 & 3 from 1.2 lesson for help with the following questions

1) The colours of cars in the King's parking lot are recorded in the table below

Colour	Frequency
White	11
Black	24
Silver	15
Red	8
Blue	6
Beige	4
Other	2

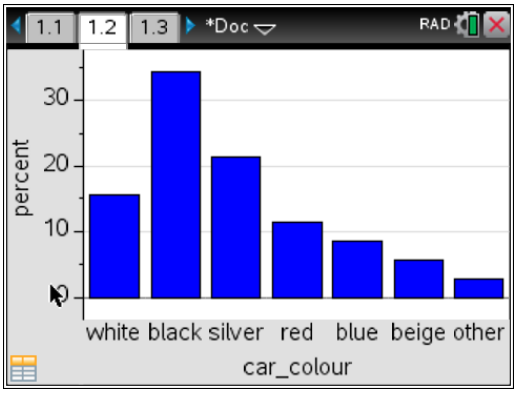
a) What type of variable is 'colour of car'? *It is a nominal, categorical variable*

b) Make a bar graph to display this data



c) Copy the table and add a relative frequency column
d) Make a relative frequency bar graph

Colour	Frequency	Relative Frequency
White	11	15.7%
Black	24	34.3%
Silver	15	21.4%
Red	8	11.4%
Blue	6	8.6%
Beige	4	5.7%
Other	2	2.9%

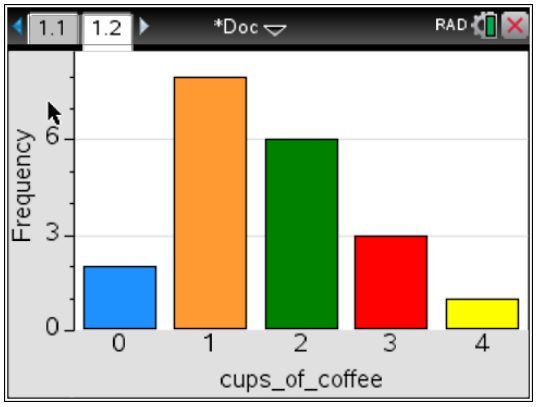


2) In order to set a reasonable price for a bottomless cup of coffee, a restaurant owner recorded the number of cups each customer ordered on a typical afternoon.

2	1	2	3	0	1	1	1	2	2
1	3	1	4	2	0	1	2	3	1

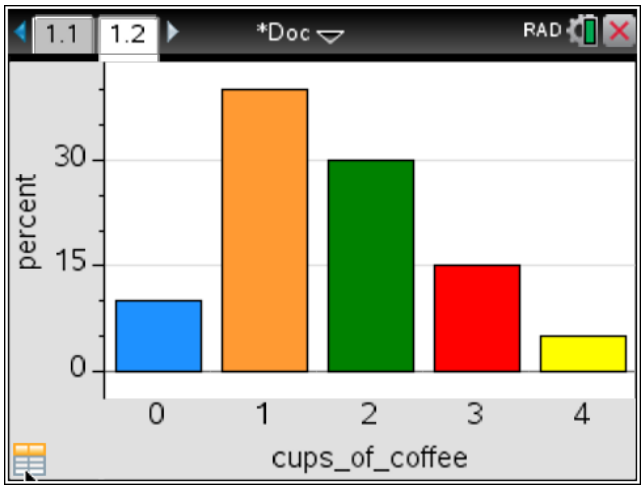
a) What type of variable is 'number of cups of coffee'? *Discrete, numeric variable*

b) Create a frequency table and bar graph to display the information



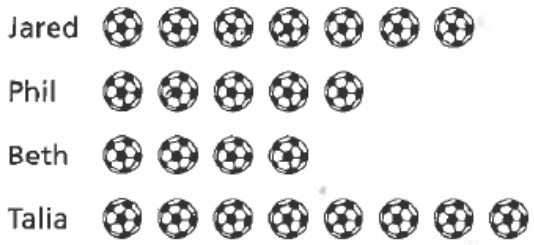
c) Create a relative frequency table and corresponding relative frequency bar graph

Cups of Coffee	Frequency	Relative Frequency
0	2	10%
1	8	40%
2	6	30%
3	3	15%
4	1	5%



Refer to part 5 from 1.2 lesson for help with the following question

3) The number of goals scored by the top four players on the school soccer team are displayed. Jared has 14 goals.



a) What information is missing from the graph? Provide it.

Legend is missing. Each soccer ball represents 2 goals.

b) How many goals does each player have?

Jared-14, Phil-10, Beth-8, Talia-16

c) What are the advantages and disadvantages of using a pictograph?

advantages-simple, visually appealing; disadvantages-hard to tell what fraction of the symbol has been drawn

Refer to part 6 from 1.2 lesson for help with the following questions

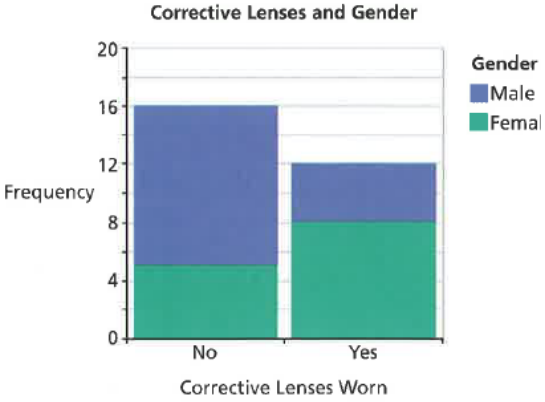
4) Here data from a survey conducted at eight high schools on smoking among students and their parents.

	Neither Parent Smokes	**One Parent Smokes**	**Both Parents Smoke**	**Total**
Student Does Not Smoke	1168 =86.1%	1823 =81.4%	1380 =77.5%	4371
Student Smokes	188 =13.9%	416 =18.6%	400 =22.5%	1004
Total	1356	2239	1780	5375

a) Copy the table and complete the totals
b) Calculate the conditional distribution of student smoking behavior based on parent smoking behaviour.
c) Use your conditional distribution to describe the relationship between the smoking behaviours of students and their parents. *When the number of parents smoking goes from 0 to 1 to 2, the percentage of students who smoke increases.*

5) Students were asked if they wear corrective lenses. The graph shows the responses. Do more females or more males wear corrective lenses? Explain.

	Male	Female	Total
Wears Corrective Lenses	4	8	12
Does Not Wear Corrective Lenses	11	5	16
Total	15	13	28



More females wear corrective lenses

6) Students were asked if they possess a valid driver's license. The results are shown below, broken down by gender. Does gender have any effect on whether a student has a license or not? Explain.

	Male	Female	Total
Has License	9 =64.3%	11 =73.3%	20
Does Not Have License	5 =35.7%	4 =26.7%	9
Total	14	15	29

To check if having a license depends on gender, I checked the conditional distribution for having a license based on gender (column percentages).

The proportion of females who have their license is higher than the proportion of males. This shows that gender may have an effect on whether a student has a license or not.

Section 1.3 Worksheet – Organizing and Displaying Quantitative Data

MDM4U
Jensen

Refer to Part 2 of 1.3 lesson for help with the following question

1) The number of hot dogs sold by a street vendor for each day in the month of June is recorded below

112	98	108	128	24	30	89
106	48	34	16	71	122	71
102	118	53	76	76	25	72
52	33	122	33	109	109	110
116	21					

a) Construct a stemplot to display the data

Stem	Leaf
1	6
2	1 4 5
3	0 3 3 4
4	8
5	2 3
6	
7	1 1 2 6 6
8	9
9	8
10	2 6 8 9 9
11	0 2 6 8
12	2 2 8

b) On what percent of days were more than 100 hotdogs sold?

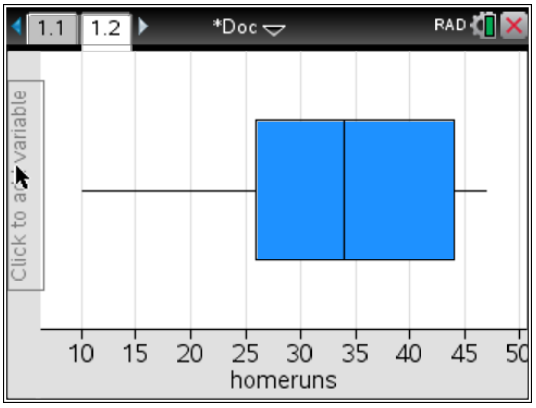
$\% > 100 = \frac{12}{30} = 0.4 = 40\%$

Refer to Part 3 of 1.3 lesson for help with the following question

2) Here are the number of homeruns that Hank Aaron hit in each of his 23 seasons. Make a boxplot for these data. Make sure to check for outliers.

13	27	26	44	30	39	40	34
45	44	24	32	44	39	29	44
38	47	34	40	20	12	10	

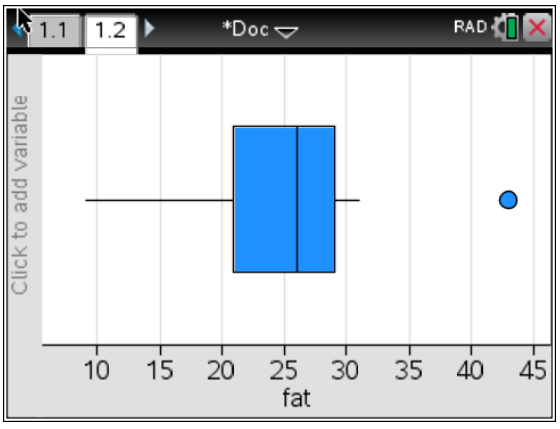
$Q_1 = 26$
$Q_2 = 34$
$Q_3 = 44$
$IQR = 18$
Lower Threshold $= 26 - 1.5(18) = -1$
Upper Threshold $= 44 + 1.5(18) = 71$
Therefore no outliers



3) McDonald's sells several different types of beef sandwiches. Below are the 12 amounts of fat in order. Make a boxplot for these data. Make sure to check for outliers.

9	12	19	23	24	26	26	27	29	29	31	43

$Q_1 = 21$
$Q_2 = 26$
$Q_3 = 29$
$IQR = 8$
Lower Threshold $= 21 - 1.5(8) = 9$
Upper Threshold $= 29 + 1.5(8) = 41$
Therefore 43 is an outlier



Refer to Part 4 of 1.3 lesson for help with the following question

4) The examination scores for a biology class are shown below.

68	77	91	66	52	58	79	94	81
60	73	57	44	58	71	78	80	54
87	43	61	90	41	76	55	75	49

a) Determine the range of the data.

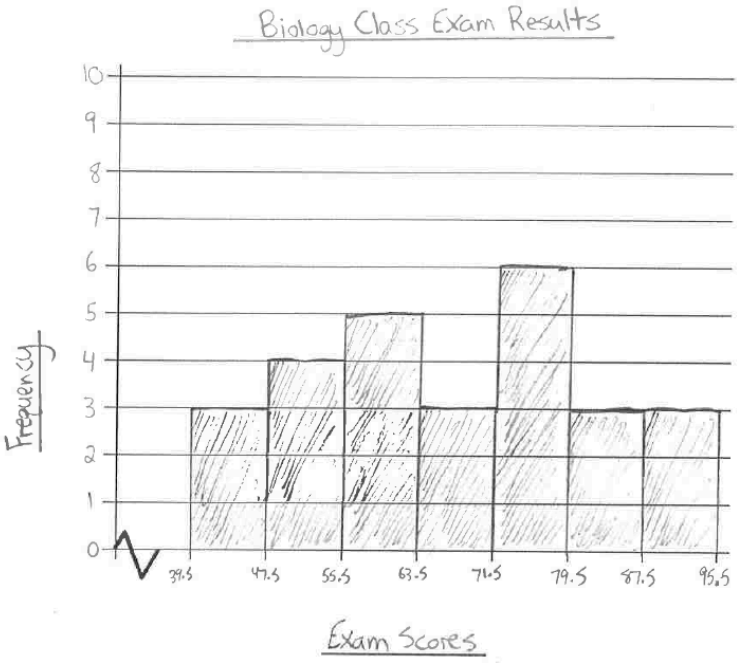
$Range = 94 - 41 = 53$

b) Determine an appropriate bin width that will divide the data into 7 intervals.

$$Bin\ Width = \frac{rounded\ range}{\#\ of\ intervals} = \frac{56}{7} = 8$$

c) Create a frequency table for the data
d) Create a histogram of the data

Grade Interval	Frequency
39.5 – 47.5	3
47.5 – 55.5	4
55.5 – 63.5	5
63.5 – 71.5	3
71.5 – 79.5	6
79.5 – 87.5	3
87.5 – 95.5	3



5) The bowling scores for a sample of league members are shown below.

154	257	195	220	182	240	177	228	235
146	174	192	165	207	185	180	264	169
225	239	148	190	182	205	148	188	

a) Determine the range of the data.

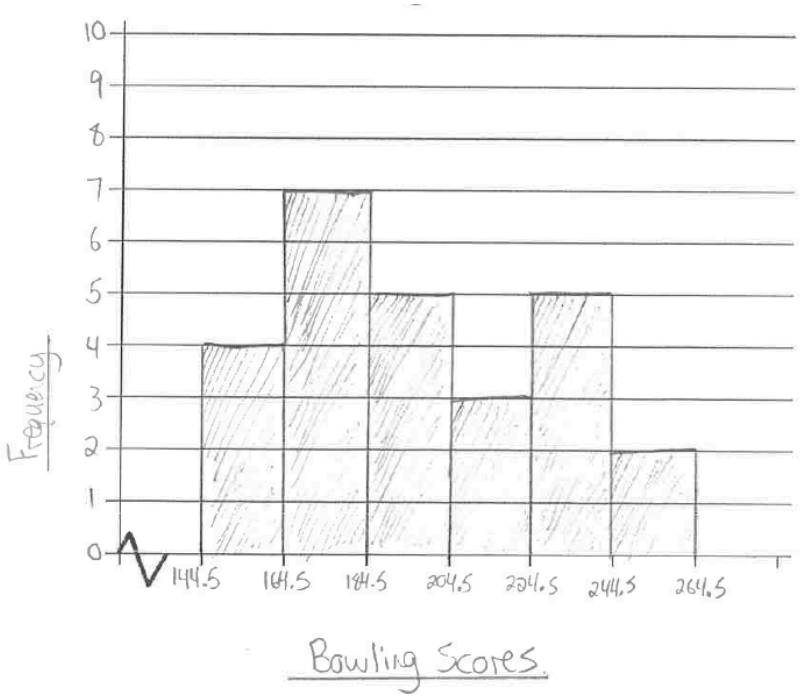
$Range = 264 - 146 = 118$

b) Determine an appropriate bin width that will divide the data into 6 intervals.

$$Bin\ Width = \frac{rounded\ range}{\#\ of\ intervals} = \frac{120}{6} = 20$$

c) Create a frequency table for the data
d) Create a histogram of the data

Bowling Score	Frequency
144.5 – 164.5	4
164.5 – 184.5	7
184.5 – 204.5	5
204.5 – 224.5	3
224.5 – 244.5	5
244.5 – 264.5	2



Section 1.4 Worksheet – Scatterplots and Correlation vs. Causation

MDM4U
Jensen

Refer to Part 2 of the 1.4 lesson for help with the following question

1) Identify the explanatory and the response variable in a correlation study of

a) heart disease AND cholesterol level

explanatory: cholesterol level *response: heart disease*

b) hours of basketball practice AND free-throw success

explanatory: hours of basketball practice *response: free-throw success*

c) amount of fertilizer used AND height of plant

explanatory: amount of fertilizer used *response: height of plant*

d) income AND level of education

explanatory: education *response: income*

e) running speed AND pulse rate

explanatory: running speed *response: pulse rate*

Refer to Part 3 of the 1.4 lesson for help with the following question

2) Classify the direction of linear correlation that you would expect with the following pairs of variables

a) hours of study, examination score

positive

b) speed in excess of the speed limit, amount charged on a traffic fine

positive

c) hours of television watched per week, final mark in calculus

negative

d) a person's height, sum of the digits in the person's phone number

no correlation

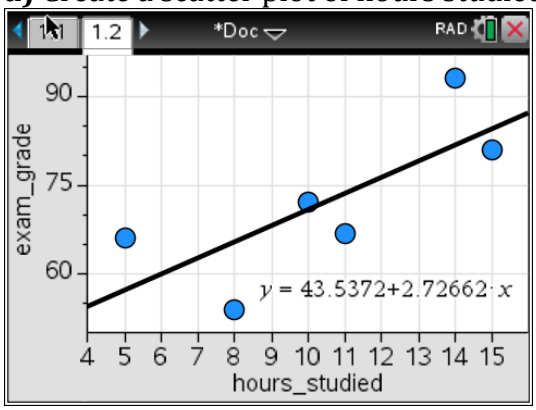
e) a person's height, the person's strength

positive

3) For a week prior to their final physics exam, a group of friends collect data to see whether time spent studying or time spent watching TV had a stronger correlation with their marks on the exam.

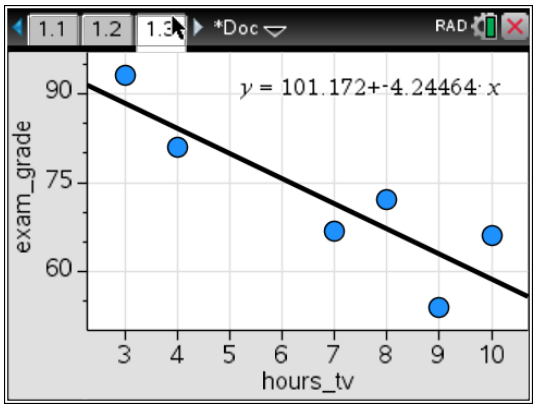
Hours Studied	Hours Watching TV	Exam Score
10	8	72
11	7	67
15	4	81
14	3	93
8	9	54
5	10	66

a) Create a scatter plot of hours studied versus exam score. Classify the linear correlation.



There appears to be a moderate to strong positive linear correlation between hours studied and exam grade. This means that the more a student studied, the higher their exam score was.

b) Create a scatter plot of hours watching TV versus exam score. Classify the linear correlation.



There appears to be a strong negative linear correlation between hours spent watching TV and exam grade. This means that the more a student watched TV, the lower their exam score was.

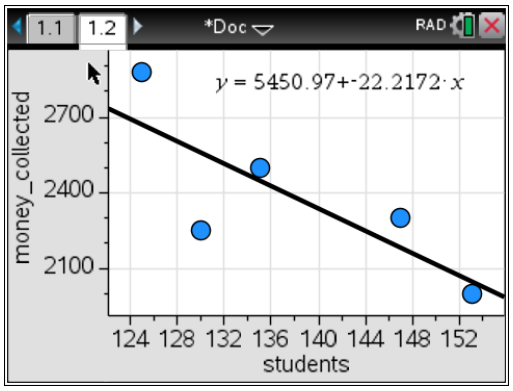
c) Which explanatory variable has a stronger correlation with exam scores? Explain.

It appears that hours spent watching TV has a stronger correlation with exam grade. The points seem less spread out indicating a stronger correlation.

4) Every year, students at a local high school collect money for a local charity. They keep track of the number of students who participate, as well as the amount of money that is collected. The information for the past five years is listed in the table below.

Year	Number of Students	Amount Collected ($)
1	130	2250
2	125	2875
3	135	2500
4	147	2300
5	153	2000

a) Create a scatterplot of the data



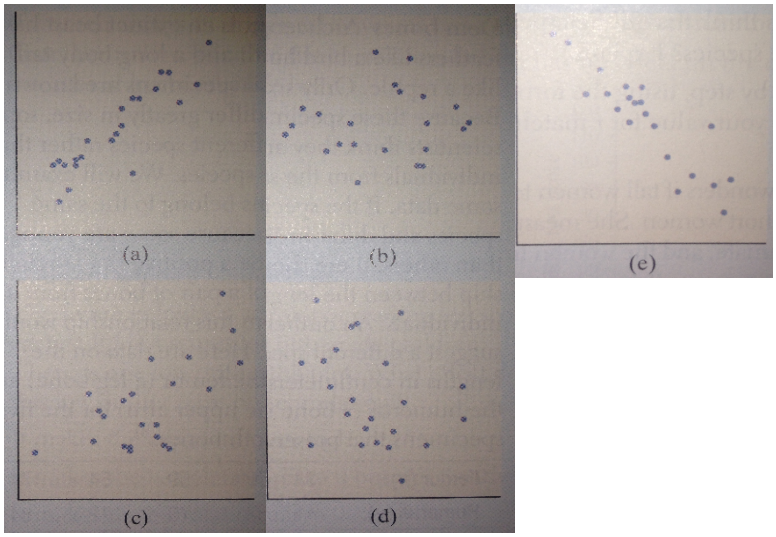
b) Describe the correlation that is observed in the data

There appears to be a moderate negative linear correlation between number of students participating and money collected. This indicates that the more students that participate, the less money that is collected.

Section 1.5 Worksheet – Linear Regression Using Technology

MDM4U
Jensen

1) Match each of the following scatterplots to the r below that describes it. Then describe the direction and strength of the correlation. (Some r's will be left over)



$r = -0.9$ $r = -0.7$ $r = -0.3$

$r = 0$ $r = 0.3$ $r = 0.7$

$r = 0.9$

a) $r = 0.9$ **b)** $r = 0$ **c)** $r = 0.7$ **d)** $r = -0.3$ **e)** $r = -0.9$

2) Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The researchers observed a linear pattern over time. They reported that the regression line $\widehat{pH} = 5.43 - 0.0053(weeks)$ fit the data well.

a) Identify the slope of the line and explain what it means in this setting.

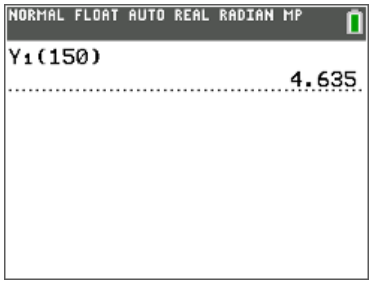
For every 1 unit increase in weeks, our model predicts a 0.0053 unit decrease in pH.

b) Identify the y-intercept of the line and explain what it means in this setting.

At 0 weeks, our model predicts an pH of 5.43

c) According to the regression line, what was the pH at the end of this study?

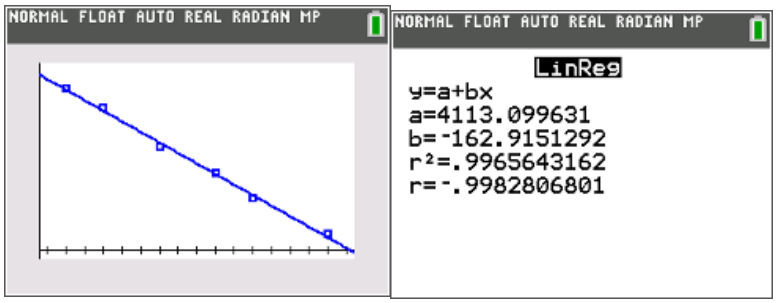
$\widehat{pH} = 5.43 - 0.0053(150) = 4.635$ *At the end of the study, the model predicts a pH of 4.635*



3) Market research has provided the following data on the monthly sales of a licensed T-shirt for a popular rock band.

Price ($)	Number of Shirts Sold
10	2500
12	2200
15	1600
18	1200
20	800
24	250

a) Make a scatterplot of the data.



b) Find the equation of the regression line and interpret the slope and y-intercept in context.

$\widehat{shirts\ sold} = 4113.1 - 162.9(price)$

The slope of -162.9 tells us that for every $1 increase in price, our model predicts a decrease of 162.9 shirts sold.

The y-intercept of 4113.1 tells us that at a price of $0, our model predicts 4113.1 shirts would be sold.

c) Find and interpret correlation coefficient, *r*.

$r = -0.998$*; this tells us there is a strong negative linear correlation between price and shirt sales.*

d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

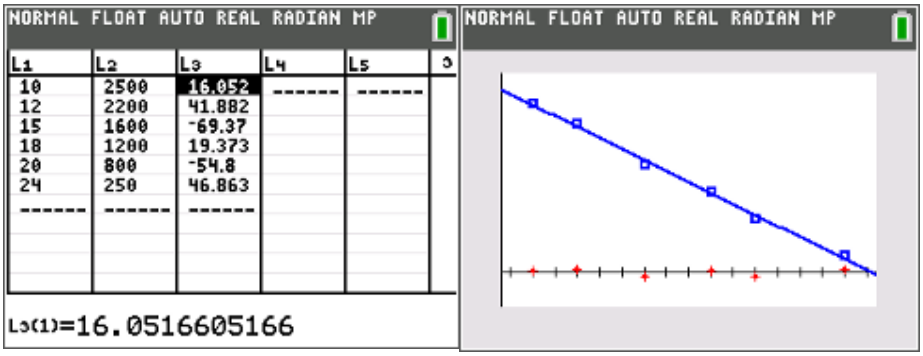
$r^2 = 0.9966$*; this tells us that about 99.66% of the variation in in shirts sold can be explained by the approximate linear relationship with price.*

e) Predict the sales if the shirts are priced at $19.

$\widehat{shirts\ sold} = 4113.1 - 162.9(19) = 1018$*; our model predicts that 1018 shirts would be sold*

f) Calculate the residual values, record them and analyze them using the residual plot to help. Is a linear model a good fit?

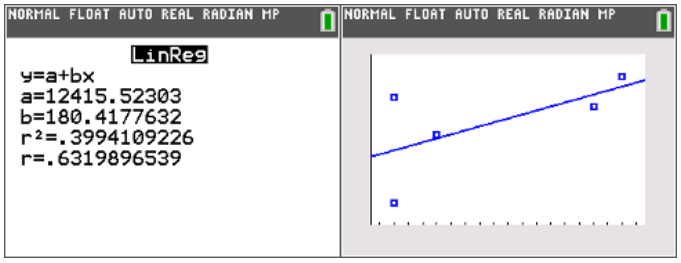
There is no distinguishable pattern in the residual plot and the residual values are relatively small. This indicates that a linear model is an appropriate model for the variables and there is a strong correlation between them.



4) Average home attendance and number of home wins for the 2009 – 2010 NBA Pacific Division teams were as follows:

	Lakers	Suns	Clippers	Warriors	Kings
Home Wins, x	34	32	21	18	18
Average Attendance, y	18 997	17 648	16 343	18 027	13 254

a) Make a scatterplot of the data.



b) Find the equation of the regression line and interpret the slope and y-intercept in context.

$\widehat{attendance} = 12415.5 + 180.4(home\ wins)$

The slope of 180.4 tells us that for every 1 more win, our model predicts a 180.4 person increase in attendance.

The y-intercept of 12 415.5 tells us that with 0 home wins, our model predicts an attendance of 12 415.5 people.

c) Find and interpret correlation coefficient, *r*.

$r = 0.63$*; this tells us there is a moderate, positive, linear correlation between home wins and attendance.*

d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

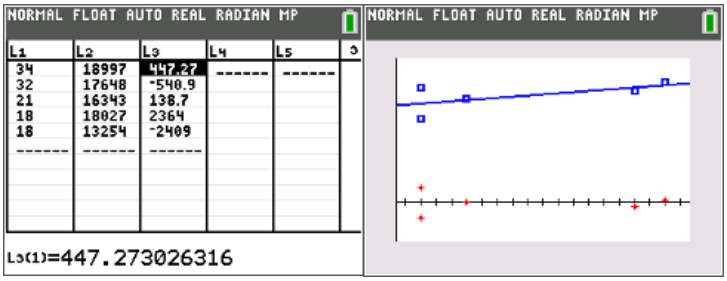
$r^2 = 0.3994$*; this tells us that about 39.94% of the variation in attendance can be explained by the approximate linear relationship with home wins.*

e) Predict the average attendance for a team with 25 home wins.

$\widehat{attendance} = 12415.5 + 180.4(25) = 16\ 925.5$*; Our model predicts an attendance of about 16 926.*

f) Calculate the residual values, record them and analyze them using the residual plot to help. Is a linear model a good fit?

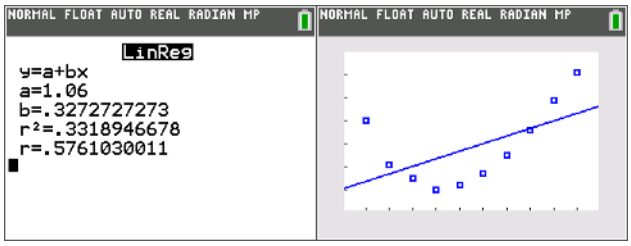
There is no distinguishable pattern in the residual plot. This tells us that the linear regression is a good model for the data.



5) Suppose the drying time of a paint product varies depending on the amount of a certain additive it contains.

Additive (oz), x	1	2	3	4	5	6	7	8	9	10
Drying Time (hr), y	4	2.1	1.5	1	1.2	1.7	2.5	3.6	4.9	6.1

a) Make a scatterplot of the data.



b) Find the equation of the regression line and interpret the slope and y-intercept in context.

$\widehat{drying\ time} = 1.06 + 0.327(additive)$

The slope of 0.327 tells us that for every 1 ounce increase in additive, our model predicts a 0.327 hour increase in drying time.

The y-intercept of 1.06 tells us that with 0 ounces of additive, our model predicts a drying time of 1.06 hours.

c) Find and interpret correlation coefficient, *r*.

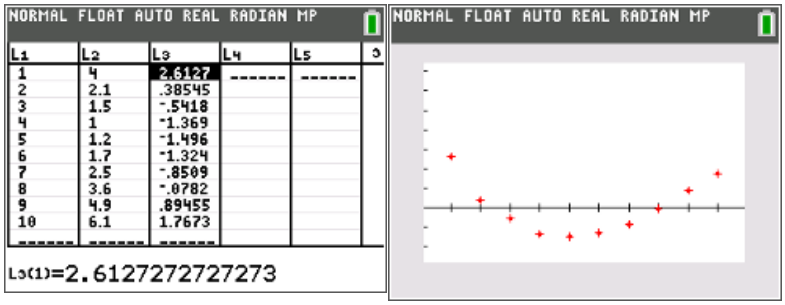
$r = 0.576$; this tells us there is a moderate, positive, linear correlation between drying time and amount of additive.

d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

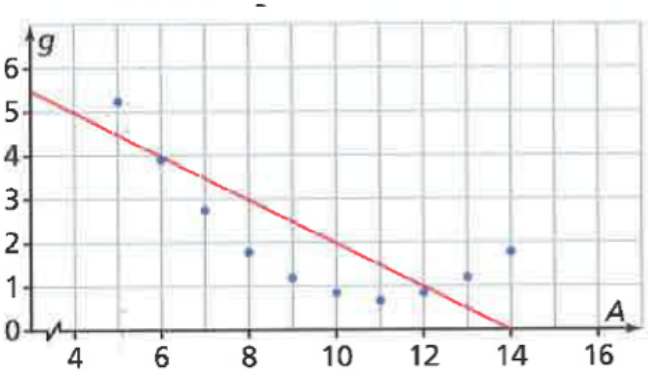
$r^2 = 0.3319$; this tells us that about 33.19% of the variation in drying time can be explained by the approximate linear relationship with amount of additive.

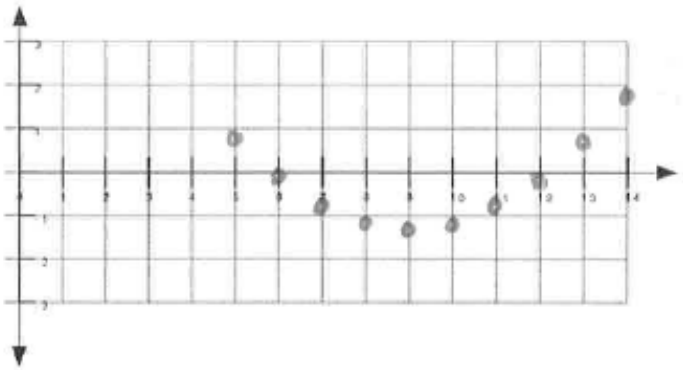
e) Calculate the residual values, record them and analyze them using the residual plot to help.

The pattern in the residual plot indicates that the linear regression is not a good model for the relationship between drying time and amount of additive.



6) Sketch the residual plot for the following scatterplot. Explain what it shows about the linear model.





Section 1.6 Worksheet – Linear Regression by Hand

MDM4U
Jensen

1) Sand driven by wind creates large dunes at the Great Sand Dunes National Monument in Colorado. Is there a linear relationship correlation between wind velocity and sand drift rate? A test site at the Great Sand Dunes National Monument gave the following information about x, wind velocity in cm/sec, and y, drift rate of sand in g/cm/sec.

a) Complete the chart

Wind Speed [x]	**Drift Rate [y]**	x^2	y^2	xy
70	3	4 900	9	210
115	45	13 225	2 025	5 175
105	21	11 025	441	2 205
82	7	6 724	49	574
93	16	8 649	256	1 488
125	62	15 625	3 844	7 750
88	12	7 744	144	1 056
$\sum x = 678$	$\sum y = 166$	$\sum x^2 = 67\,892$	$\sum y^2 = 6\,768$	$\sum xy = 18\,458$

b) Determine the equation of the least squares regression line ($\hat{y} = a + bx$). Interpret the slope and y-intercept in context.

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

Slope:

Slope = $b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{7(18\,458) - (678)(166)}{7(67\,892) - (678)^2} = \frac{16\,658}{15\,560} = 1.07$

This indicates that for every 1 cm/sec increase in wind velocity, the model predicts a 1.07 g/cm/sec increase in drift rate of sand.

y-intercept:

$\bar{x} = \frac{\sum x}{n} = \frac{678}{7} = 96.857$

$\bar{y} = \frac{\sum y}{n} = \frac{166}{7} = 23.714$

y-intercept = $a = \bar{y} - b\bar{x}$

$y - intercept = a = \bar{y} - b\bar{x} = 23.714 - 1.07(96.857) = -79.92$

This tells us that at a wind speed of 0, the model predicts a sand drift rate of -79.92 g/cm/sec.

Linear Regression Equation:

$\hat{y} = a + bx \rightarrow predicted\ drift\ rate = -79.92 + 1.07(wind\ velocity)$

c) Compute the correlation coefficient using the formula. Interpret r and r^2 in context.

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

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{7(18\,458) - (678)(166)}{\sqrt{[7(67\,892) - (678)^2][7(6\,768) - (166)^2]}} = \frac{16\,658}{17\,561.29836} = 0.94856$$

$r = 0.94856$; This indicates that there is a strong, positive, linear correlation between wind speed and drift rate.

$r^2 = 0.8998$; This tells us that about 89.98% of the variation in drift rate can be explained by the approximate linear correlation with wind speed.

2) A study was conducted to determine if larger universities tend to have more property crime. Let x represent student enrollment (in thousands) and let y represent the number of burglaries in a year on the campus. A random sample of 8 universities in California gave the following information:

a) Complete the chart

Student Enrollment [x]	**Burglaries [y]**	x^2	y^2	xy
12.5	26	156.25	676	325
30	73	900	5 329	2 190
24.5	39	600.25	1 521	955.5
14.3	23	204.49	529	328.9
7.5	15	56.25	225	112.5
27.7	30	767.29	900	831
16.2	15	262.44	225	243
20.1	25	404.01	625	502.5
$\sum x = 152.8$	$\sum y = 246$	$\sum x^2 = 3\,350.98$	$\sum y^2 = 10\,030$	$\sum xy = 5\,488.4$

b) Determine the equation of the least squares regression line ($\hat{y} = a + bx$) by hand. Interpret the slope and y-intercept in context.

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

Slope:

$Slope = b = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} = \frac{8(5488.4)-(152.8)(246)}{8(3350.98)-(152.8)^2} = \frac{6318.4}{3460} = 1.826$

The slope tells us that for every 1000 more students enrolled, the model predicts 1.826 more burglaries a year.

y-intercept:

y-intercept = $a = \bar{y} - b\bar{x}$

$\bar{x} = \frac{\sum x}{n} = \frac{152.8}{8} = 19.1$

$\bar{y} = \frac{\sum y}{n} = \frac{246}{8} = 30.75$

$y - intercept = a = \bar{y} - b\bar{x} = 30.75 - 1.826(19.1) = -4.1266$

The y-intercept tells us that if 0 students were enrolled, the model predicts -4.1266 crimes a year.

Linear Regression Equation:

$\hat{y} = a + bx \rightarrow predicted\ burglaries = -4.1266 + 1.826(student\ enrollment)$

c) Compute the correlation coefficient using the formula. Interpret r and r^2 in context.

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

$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{8(5488.4) - (152.8)(246)}{\sqrt{[8(3350.98) - (152.8)^2][8(10030) - (246)^2]}} = \frac{6318.4}{8261.055623} = 0.7648$

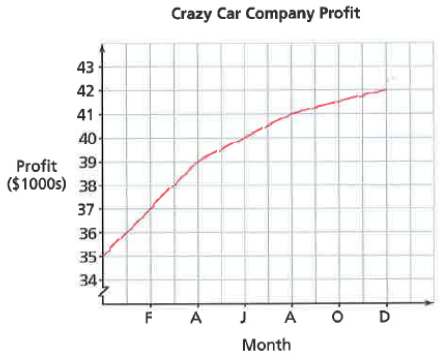
$r = 0.7648$; this tells us there is a moderate, positive, linear correlation between student enrollment and burglaries.

$r^2 = 0.5849$; this tells us that about 58.49% of the variation in burglaries can be explained by the approximate linear correlation with student enrollment.

Section 1.7 Worksheet – Misrepresentations of Data

MDM4U
Jensen

1) The two graphs below show the profits of the Crazy Car Company.



a) How are the graphs similar? How are they different?

The two graphs show the same set of data, but using different scales.

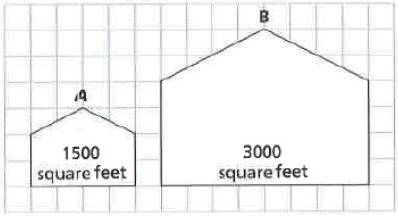
b) How much has the profit increased on each graph?

Both graphs show the same profit from $35 000 to $42 000 over 12 months.

c) What false impressions are conveyed by the two graphs?

The first graph uses a truncated y-axis to try and show a large profit growth over the year. The second graph shows very little profit over the year by using a large scale.

2) The increase in the size of homes purchased is shown in the graph below.



a) What is similar about the homes?

The homes are identical in shape except in size.

b) Using the tiles of the graph, how many times bigger is the area of the shape of house B than the area of house A?

The area of house B is 4 times larger than the area of house A.

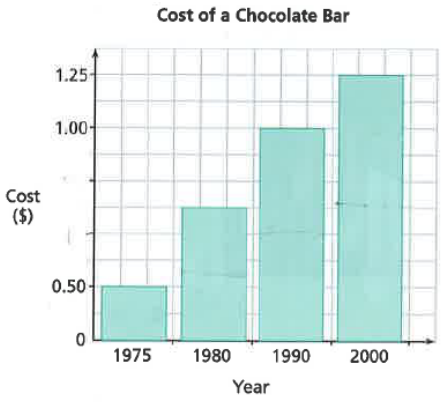
c) By how much has the actual size of the home increased.

Home size increased from 1500 square feet to 3000 square feet. It has increased by a factor 2.

d) List any false impressions conveyed by the graph.

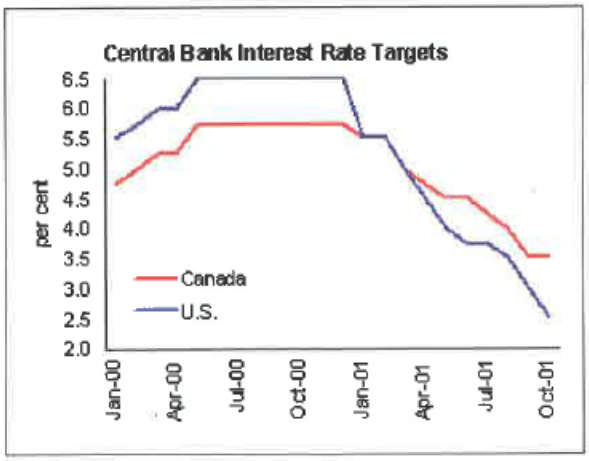
Because the area of the second house is so much larger than the first house, it appears that the size of homes has increased tremendously.

3) List the false impressions conveyed by this graph. How could you change the graph to correct the false impressions?



It seems that the price in 1980 is 2.5 times the price in 1975, but really it is only 1.5 times. The price in 1990 seems 4 times the price in 1975, but really it is only 2 times. Also, it does not look like a steady increase in price, although it is a steady increase. We can change the graph to correct the false impressions by changing the vertical scale (cost) so that is has regular intervals.

4) Examine the graph below.



Source: Phillips, Hager & North Investment Management Ltd.

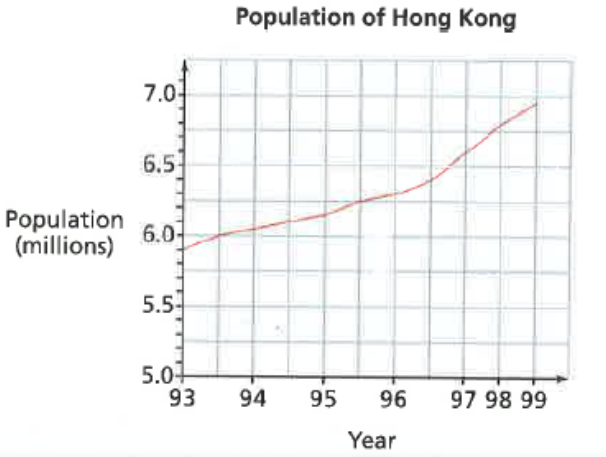
a) Has the data been misrepresented to bias the reader? Give reasons.

Yes; the vertical scale (percent) is truncated. The differences look larger.

b) How could you modify the graph to display the data accurately?

Can modify the graph by using correct scales (start at 0 on the y-axis and increase at equal and evenly spaced intervals)

5) The graph below shows the population of Hong Kong from 1993 to 1999. Explain why this graph would cause incorrect interpretations of the data.



The graph would cause incorrect interpretations of the data because the horizontal scale (year) does not have regular intervals AND the y-axis is truncated

6) Suppose that in a recent magazine article, the graphic below was used to show how the use of cell phones changed between 1994 and 1998. Explain why this picture is misleading.

420/71= 5.9 In 1998, the use of cell phones has increased approximately by a factor of 6 compare to 1994. The graphic in the margin shows the 1998 cell phone to be much larger than the 1994 cell phone. The 1998 cell phone size appears to be 25 times larger than the 1994 cell phone.