CSCE 5310 - Empirical Analysis
Assignment 2

1. Try these multiple-choice questions

   1. A correlation coefficient of -0.95 means there is a _____________ between the two variables.
   A. Strong positive correlation

   B. Weak negative correlation

   C. Strong negative correlation

   D. No Correlation

2. According to the data reported by the New York State Department of Health regarding West Nile Virus for the years 2000-2004, the least squares line equation for the number of reported dead birds (x) versus the number of human West Nile virus cases (y) is ^y= -10.2638+0.0491x. If the number of dead birds reported in a year is 732, how many human cases of West Nile virus can be expected?

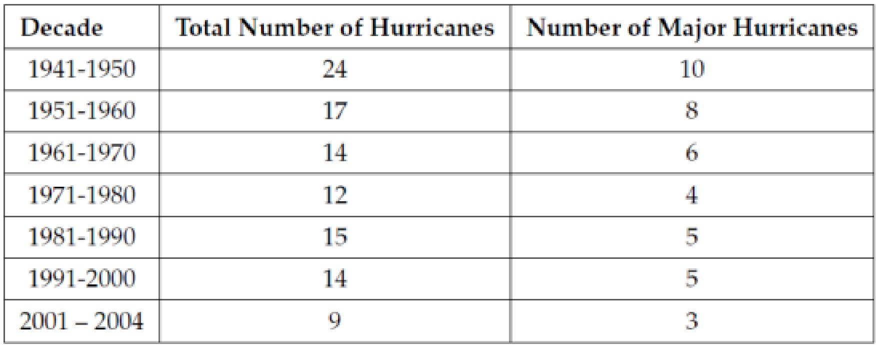
   A. 25.7

   B. 46.2

   C. -25.7

   D. 7513

3. The next two questions refer to the following data: (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from www.nhc.noaa.gov/gifs/table6.gif 13 A major hurricane is one with a strength rating of 3, 4 or 5.



Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	9	3

Using only completed decades (1941 – 2000), calculate the least squares line for the number of major hurricanes expected based on the total number of hurricanes.

Total number of hurricanes (x)	Number of major Hurricanes (y)	x^2	y^2	xy
24	10	576	100	240
17	8	289	64	136
14	6	196	36	84
12	4	144	16	48
15	5	225	25	75

	14	5	196	25	70
	9	3	81	9	27
Total	105	41	1707	275	680

Slope = $n\sum xy - \sum x\sum y / n\sum x^2 – (\sum x)^2$

= $7(680)-105(41) / 7(1707)-(105)^2$

= 0.4924 = 0.5

y-intercept = $\sum y - m\sum x / n$

= (41 – (0.6*105)) / 7

= -1.5292

A. ^y = -1.67x + 0.5

B. ^y = 0.5x - 1.67

C. ^y = 0.94x - 1.67

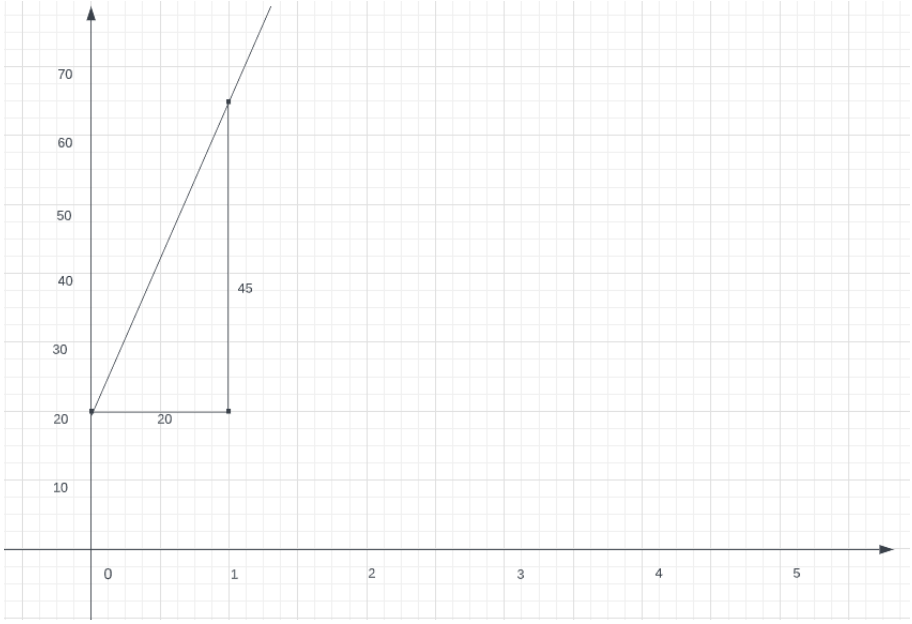
D. ^y = -2x + 1

4. The data for 2001-2004 show 9 hurricanes have hit the mainland United States. The line of best fit predicts 2.83 major hurricanes to hit the mainland U.S. Can the least squares line be used to make this prediction?
A. No, because 9 lies outside the independent variable values
B. Yes, because, in fact, there have been 3 major hurricanes this decade
C. No, because 2.83 lies outside the dependent variable values
D. Yes, because how else could we predict what is going to happen this decade.

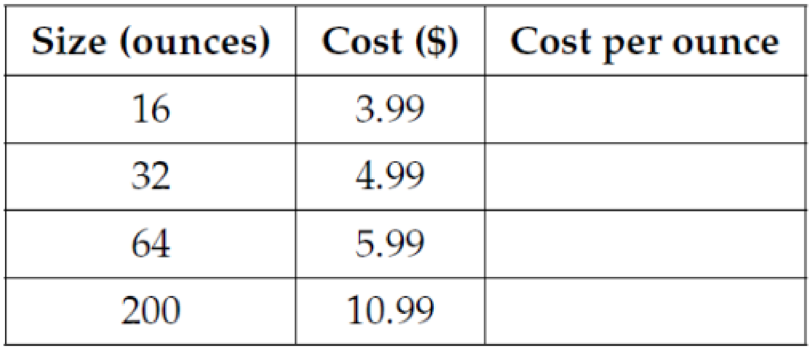
5. Coach Jack trains kids in soccer skills to make extra money. For each session, he charges a one-time fee of $20 plus $45 per hour of training. A linear equation that expresses the total amount of money Jack earns for each session she trains is y = 20 + 45x. What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution: y = 45x + 20 (y = a + bx, b = slope, and a = y-intercept)
Independent variables: x
Dependent variables: y
y-intercept: 20
slope: 45

As the per-hour training rate increases the total amount also increases linearly. As the per-hour training increases by 1 hour, the total amount increases by 45 times.
From algebra recall that a slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point (0, a) i.e. (0, 20) where the line crosses the y-axis. If b > 0, the line slopes upward to the right.



6. The cost of a leading liquid laundry detergent in different sizes is given below

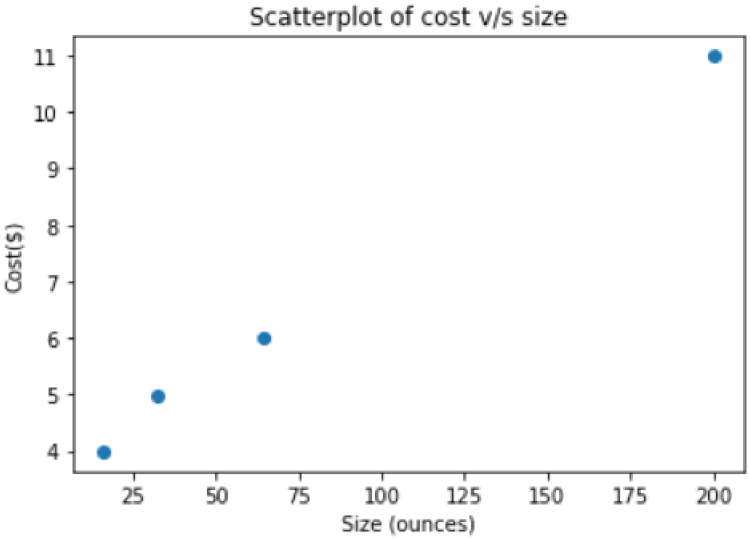


Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Part 1:

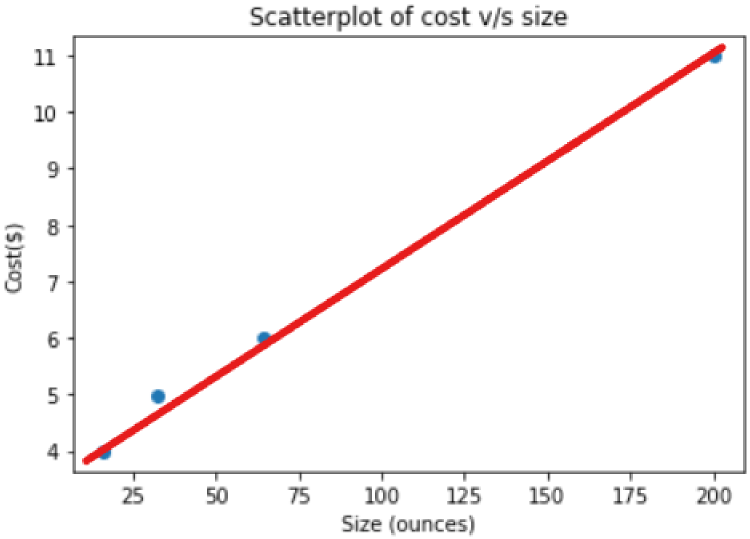
a) Using “size” as the independent variable and “cost” as the dependent variable, make a scatter plot.

CSCE 5310 - Empirical Analysis
Assignment 2



b) Does it appear from inspection that there is a relationship between the variables? Why or why not?

Yes, from the scatterplot, we can notice that as the size (x) increases the cost (y) also increases. There is a strong positive relationship between size and cost because the points lie around the straight line that can be drawn joining the given points.



c) Calculate the least squares line. Put the equation in the form of: ^y = a + bx

	Size(x)	Cost(y)	x^2	y^2	xy
	16	3.99	256	15.9210	63.84
	32	4.99	1024	24.9001	159.68
	64	5.99	4096	35.8801	383.36
	200	10.99	40000	120.7801	2198
Total	312	25.96	45376	197.4813	2804.88

CSCE 5310 - Empirical Analysis
Assignment 2

Slope = $n\sum xy - \sum x\sum y \,/\, n\sum x^2 - (\sum x)^2$

$= 4(28.88)-321(25.96) \,/\, 4(45376)-(312)^2$

$= 3120/84160 = 0.03707$

y-intercept = $\sum y - m\sum x \,/\, n$

$= (25.96 - (0.03707*312)) \,/\, 4$

$= 25.96 - 11.56584 \,/\, 4$

$= 3.5985$

Equation: $^{\wedge}y=3.5985 + 0.03707x$

d) Find the correlation coefficient.

Correlation coefficient(r)= $n\sum xy - \sum x\sum y \,/\, \sqrt{}n\sum x^2 - (\sum x)^2 \;\; \sqrt{}n\sum y^2 - (\sum y)^2$

$= 4(28.88)-321(25.96) \,/\, \sqrt{}4(45376)-(312)^2 \;\; \sqrt{}4(197.4813) - (25.96)^2$

$= 3120/\sqrt{}84160\sqrt{}116$

$= 3120/290.1*10.77 = 3210/3124.377$

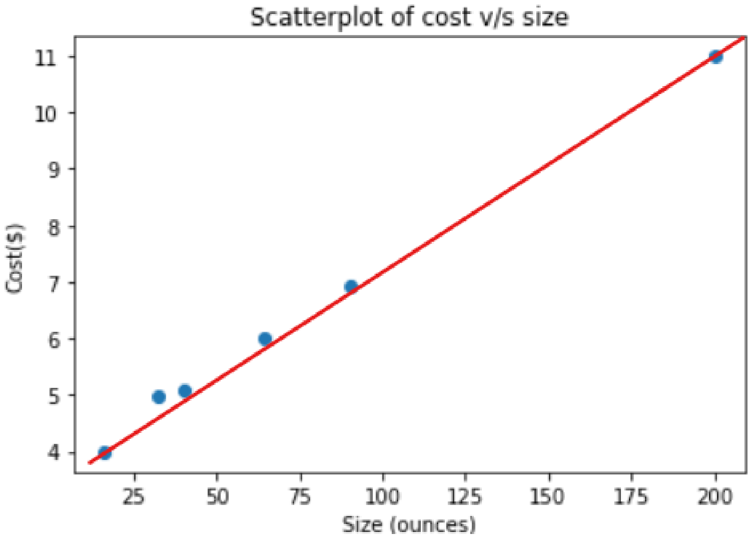
$= 0.99859908$

e) If the laundry detergent was sold in a 40-ounce size, find the estimated cost.
We have the equation $^{\wedge}y=3.5985 + 0.03707x$
Given the size (x-intercept) = 40
Estimated cost (y-intercept) = $^{\wedge}y=3.5985 + 0.03707\,(40)$
$= 3.5985 + 1.4828$
$= 5.0813$

f) If the laundry detergent was sold in a 90-ounce size, find the estimated cost.
6.93
We have the equation $^{\wedge}y=3.5985 + 0.03707x$
Given the size (x-intercept) = 90
Estimated cost (y-intercept) = $^{\wedge}y=3.5985 + 0.03707\,(90)$
$= 3.5985 + 3.3363$
$= 6.9348$

g) Use the two points in (e) and (f) to plot the least squares line on your graph from (a).

CSCE 5310 - Empirical Analysis
Assignment 2



h) Does it appear that a line is the best way to fit the data? Why or why not?
Yes, it is the best way to fit the data because the data points in the graph lie closer to the line, and the correlation coefficient value is 0.9986 significant. It shows a clear relationship between size and cost.

i) Are there any outliers in the above data?
There are no specific outlies but (200,10.99) would be an outlier because it lies away from the other data points.

j) Is the least squares line valid for predicting what a 300-ounce size of laundry detergent would cost? Why or why not?
It is not valid for predicting what a 300-ounce size of laundry detergent would cost because 300 ounces does not outside the range of x and would be an outlier. It might not be possible to determine the cost.

k) What is the slope of the least squares (best-fit) line? Interpret the slope.

The slope of the least squares (best-fit) line is 0.03707. We interpret that the change in size changes the cost i.e., for a unit change in size x there will be 0.03707 times increase in cost y.

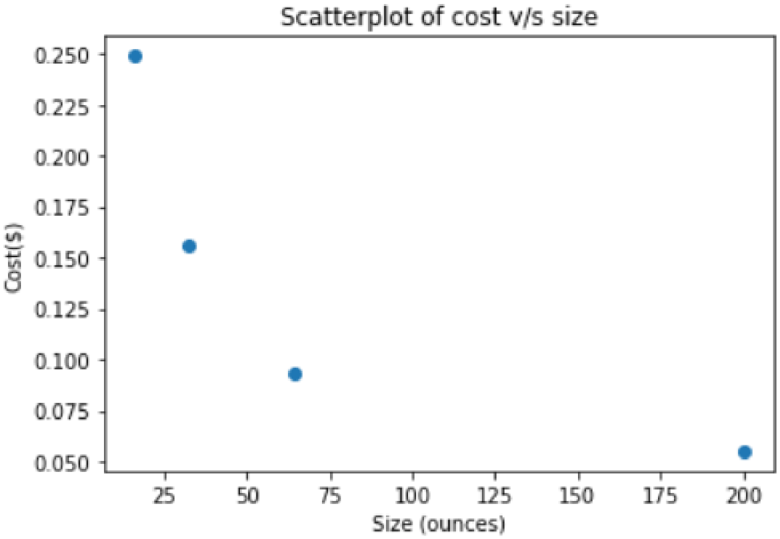
Part 2:

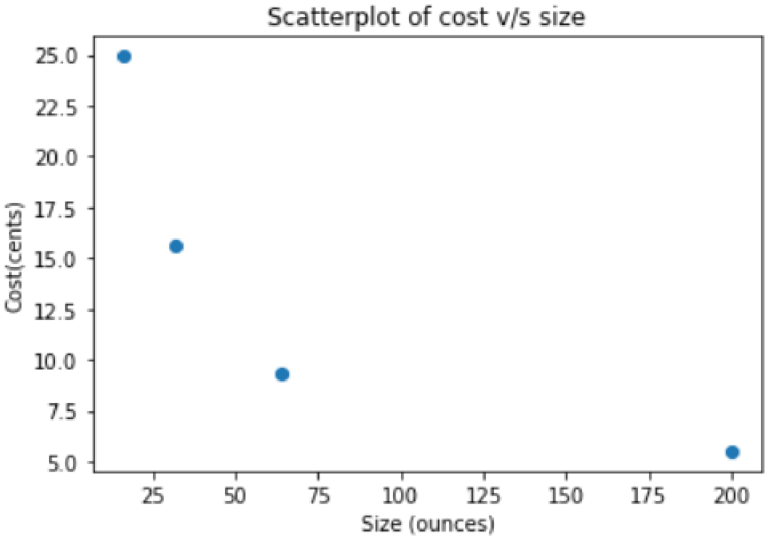
a) Complete the above table for the cost per ounce of the different sizes.

Size(ounces)	Cost ($)	Cost per ounce ($)	Cost per ounce (cents)
16	3.99	0.249	24.937
32	4.99	0.156	15.593
64	5.99	0.093	9.359

200	10.99	0.055	5.495
			55.384

b) Using “Size” as the independent variable and “Cost per ounce” as the dependent variable, make a scatter plot of the data.





c) Does it appear from inspection that there is a relationship between the variables? Why or why not?

Yes, from the scatterplot, we can notice that as the size (x) increases the cost (y) decreases, and there is a linear relationship for sizes 16,32, and 64 but it does not hold true for 200.

CSCE 5310 - Empirical Analysis
Assignment 2

d) Calculate the least squares line. Put the equation in the form of ^y = a + bx

	Size(x)	Cost per ounce ($)	Cost per ounce (cents)	x^2	y^2	xy
	16	0.249	24.937	256	621.853	398.992
	32	0.156	15.593	1024	243.141	498.976
	64	0.093	9.359	4096	87.590	598.976
	200	0.055	5.495	40000	30.195	1099
Total	312	0.553	55.384	45376	982.689	2595.944

Slope = $n\sum xy - \sum x\sum y \,/\, n\sum x^2 - (\sum x)^2$

= $4(2595.944)-321(55.384) \,/\, 4(45376)-(312)^2$

= 10383.776-17778.264 /84160 = -0.0819

y-intercept = $\sum y - m\sum x \,/\, n$

= (55.384 – (-0.08786226*312)) / 4

= 20.237

Equation: $^\wedge$y=20.237-0.0819x

e) Find the correlation coefficient.

Correlation coefficient(r)= $n\sum xy - \sum x\sum y \,/\, \surd n\sum x^2 - (\sum x)^2 \;\; \surd n\sum y^2 - (\sum y)^2$

= $4(2595.944)-321(55.384) \,/\, \surd 4(45376)-(312)^2 \;\; \surd 4(982.689) - (55.384)^2$

= -7394.488/√84160√ 863.369

= -7394/290.1*29.383 = -7394/8524.

= -0.8088262

f) If the laundry detergent was sold in a 40-ounce size, find the estimated cost per ounce.

We have the equation $^\wedge$y=20.237-0.0819x

Given the size (x-intercept) = 40

Estimated cost (y-intercept) = $^\wedge$y=20.237-0.0819 (40)

= 16.961

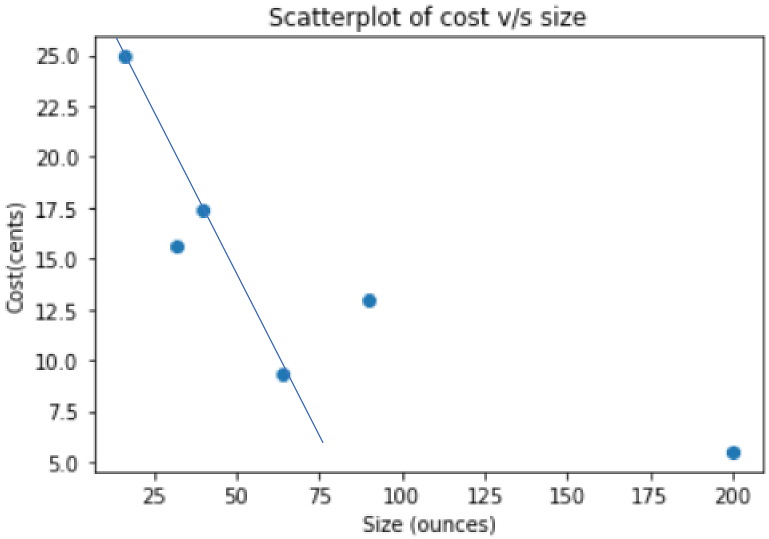
g) If the laundry detergent was sold in a 90-ounce size, find the estimated cost per ounce.

We have the equation $^\wedge$y=20.237-0.0819x

Given the size (x-intercept) = 90

Estimated cost (y-intercept) = $^{\wedge}$y=20.237-0.0819 (90)
= 12.865

h) Use the two points in (f) and (g) to plot the least squares line on your graph from (b).



i) Does it appear that a line is the best way to fit the data? Why or why not?

The line is not the best way to fit the data, it does not cover all the data points. The relation is also not linear for all the data points.

j) Are there any outliers in the above data?
There is no proper trend between the data points to indicate the outliers. So, there are no outliers, but the size (200,5.49) lies away from the other data points.

k) Is the least squares line valid for predicting what a 300-ounce size of laundry detergent would cost per ounce? Why or why not?
It is not valid for predicting what a 300-ounce size of laundry detergent would cost because 300 ounces is outside the range of x. It might not be possible to determine the cost.

l) What is the slope of the least squares (best-fit) line? Interpret the slope.
As the size in ounces increases the cost per ounce decreases by 0.8088262. Because the slope is -0.8088262.