Colm Mooney – ST204 - 20325583

CA4 Opt RainGrad.csv

#Question 1

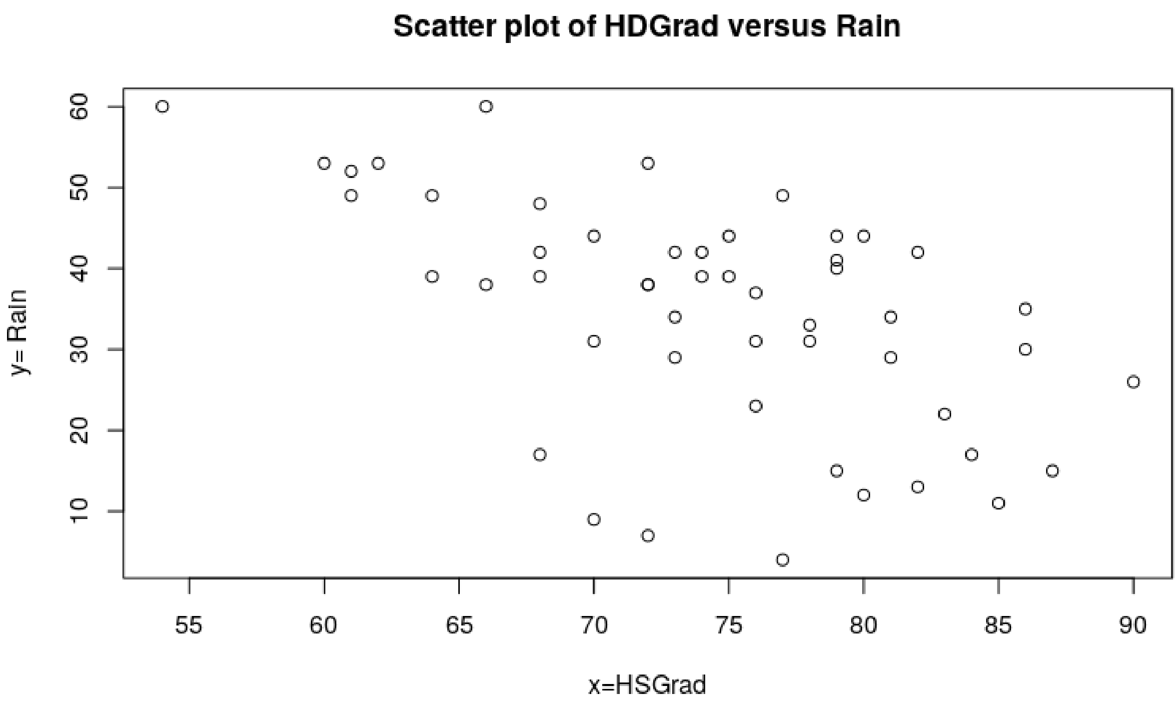
library(readr)

CA4_Opt_RainGrad <- read_csv("SharedFiles/ST204/Data/CA4_Opt_RainGrad.csv")

CA4_Opt_RainGrad

#(a)Generate a scatter plot of HDGrad versus Rain.

plot(CA4_Opt_RainGrad$HSGrad, CA4_Opt_RainGrad$Rain, ylab="y= Rain", xlab="x=HSGrad", main="Scatter plot of HDGrad versus Rain")



#(b) Estimate Pearson's correlation, Spearman's rank correlation, and Kendall's tau for these data

#and perform appropriate two-sided hypothesis tests for each.

#State clearly the null and alternative hypotheses as well as your conclusions in each case.

with(CA4_Opt_RainGrad, cor.test(HSGrad, Rain, method="pearson")) #We reject the hypothesis as p-value = 2.267e-05 is less than 0.05

with(CA4_Opt_RainGrad, cor.test(HSGrad, Rain, method="spearman")) #We reject the hypothesis as p-value = 2.863e-05 is less than 0.05

with(CA4_Opt_RainGrad, cor.test(HSGrad, Rain, method="kendall")) #We reject the hypothesis as p-value = 4.47e-05 is less than 0.05

#(c) Briefly discuss the statement ‘correlation does not imply causation’ in the context of these data.

#Just because it looks like people did better with less rainfall, doesn't imply that it is because of the rain.

#It could be better schooling, facilities or better textbooks.

#So just because it looks like people did better where there was less rainfall, doesn't mean it is BECAUSE there was less rainfall.

#(d) Let r and rs denote Pearson’s and Spearman’s correlation coefficients, respectively. Obtain

#these values in R and use them to verify by hand the values of the associated test statistics t and S. Show your workings.

#S = 34263, t = -4.6837

```
y<- c(CA4_Opt_RainGrad$Rain)

x<- c(CA4_Opt_RainGrad$HSGrad)

sort(x)

sort(y)

sum(y)

sum(x)

mean(x)

mean(y)

meany <- 34.63

meanx <- 74.24

meany

meanx

1:51

sum((y - meany)^2) * sum((x - meanx)^2)

sum((x - meanx) * (y - meany))

sum((x - meanx)^2) * sum((y - meany)^2)

with(CA4_Opt_RainGrad, cor.test(HSGrad, Rain, method="pearson"))  #0.56 = r

with(CA4_Opt_RainGrad, cor.test(HSGrad, Rain, method="spearman")) #0.55 = rho
```

	X = HSGrad		Y = Rain	
1	8.5	66	50.5	60
2	18.5	72	48	53
3	18.5	72	2	7
4	31.5	77	44	49
5	11.5	68	9.5	17
6	26.5	75	39.5	44
7	36.5	79	33	41
8	11.5	68	29.5	39
9	2	60	48	53
10	3.5	61	44	49
11	6.5	64	44	49
12	29	76	12	23
13	39.5	80	5	12
14	33.5	78	20	33
15	26.5	75	29.5	39
16	48.5	86	23	35
17	44.5	81	14.5	29
18	15	70	39.5	44
19	1	54	50.5	60
20	36.5	79	39.5	44
21	22	73	35.5	42
22	39.5	80	39.5	44
23	15	70	18	31
24	51	90	13	26
25	5	62	48	53
26	22	73	21.5	34
27	47	85	4	11
28	48.5	86	16	30
29	31.5	77	1	4
30	29	76	24	37
31	42.5	82	35.5	42
32	15	70	3	9

	X = HSGrad		Y = Rain	
33	6.5	64	29.5	39
34	11.5	68	35.5	42
35	50	87	7.5	15
36	18.5	72	26	38
37	29	76	18	31
38	18.5	72	26	38
39	36.5	79	32	40
40	24.5	74	35.5	42
41	3.5	61	46	52
42	46	84	9.5	17
43	11.5	68	42	48
44	8.5	66	26	38
45	36.5	79	7.5	15
46	44.5	81	21.5	34
47	24.5	74	29.5	39
48	22	73	14.5	29
49	33.5	78	18	31
50	45	83	11	22
51	42.5	82	6	13
	R(X)		R(Y)	

Pearson correlation coefficient (r)

$$r \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}}$$

$\bar{X} = 74.24$

$\bar{Y} = 34.63$

Sum of $X = 3786$

Sum of $Y = 1766$

Spearman

$$T = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6\sum 34263}{132651 - 51} = 1 - \frac{205578}{132600}$$

d = sum of $R(X) - R(Y)$ $= -0.55 = rho$

Number beside number = Rank

	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x}) \cdot (y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	
1	-8.24	25.37	-209.05	67.90	643.64	
2	-2.24	18.37	-41.15	5.02	337.46	
3	-2.24	-27.63	61.89	5.02	763.42	
4	2.76	14.37	39.66	7.62	206.50	
5	-6.24	-17.63	110.01	38.94	310.82	
6	0.76	9.37	7.12	0.58	87.80	
7	4.76	6.37	30.32	22.66	40.58	
8	-6.24	4.37	-27.27	38.94	19.10	
9	-14.24	18.37	-261.59	202.78	337.46	
10	-13.24	14.37	-190.26	175.30	206.50	
11	-10.24	14.37	-147.15	104.86	~~[illegible]~~	206.50
12	1.76	-11.63	-20.47	3.10	~~[illegible]~~ 135.26	
23	5.76	-22.63	-130.35	33.18	512.12	
14	3.76	-1.63	-6.13	14.14	2.66	
15	0.76	4.37	3.32	0.58	19.10	
16	11.76	0.37	4.35	138.30	0.14	
17	6.76	-5.63	-38.06	45.70	31.70	
18	-4.24	9.37	-39.73	17.98	87.80	
19	-20.24	25.37	-513.49	409.66	643.64	
20	4.76	9.37	44.60	22.66	87.80	
21	-1.24	7.37	-9.14	1.54	54.32	
22	5.76	9.37	53.97	33.18	87.80	
23	-4.24	-3.63	15.39	17.98	13.18	
24	15.76	-8.63	-136.01	248.38	74.48	
25	-12.24	18.37	-224.85	149.82	337.46	
26	-1.24	-0.63	0.78	1.54	0.40	
27	10.76	-23.63	-254.26	115.78	558.38	
28	11.76	-4.63	-54.45	138.30	21.44	
29	2.76	-30.63	-84.54	7.62	~~[illegible]~~ 938.20	
30	1.76	2.37	4.17	3.10	5.62	
31	7.76	7.37	57.19	60.22	54.32	
32	-4.24	-25.63	108.67	17.98	656.90	

	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X}) * (Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
33	-10.24	4.37	-44.75	104.86	19.10
34	-6.24	7.37	-45.99	38.94	54.32
35	12.76	-19.63	-250.48	162.82	385.34
36	-2.24	3.37	- 7.55	5.02	11.36
37	1.76	-3.63	- 6.39	3.10	13.18
38	-2.24	3.37	- 7.55	5.02	11.36
39	4.76	5.37	25.56	22.66	28.84
40	-0.24	7.37	- 1.77	0.06	54.32
41	-13.24	17.37	-229.98	175.30	301.72
42	9.76	-17.63	-172.07	95.26	310.82
43	-6.24	13.37	-83.43	38.94	178.76
44	-8.24	3.37	-27.77	67.90	11.36
45	4.76	-19.63	-93.44	22.66	385.34
46	6.76	-0.63	- 4.26	45.70	0.40
47	-0.24	4.37	-1.05	0.06	19.10
48	-1.24	-5.63	6.98	1.54	31.70
49	3.76	-3.63	-13.65	14.14	13.18
50	8.76	-12.63	-110.64	76.74	159.52
51	7.76	-21.63	-167.85	60.22	467.86
			$\Sigma = -3082.53$	$\Sigma = 3091.18$	$\Sigma = 9939.92$

$$r = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}, \quad r = \frac{-3082.53}{\sqrt{3091.18 \times 9939.92}} = \frac{-3082.53}{\sqrt{30726064}} = -0.556$$

	d	d^2
1	-42	1764
2	-29.5	870.25
3	16.5	272.25
4	-12.5	156.25
5	2	4
6	-13	169
7	3.5	12.25
8	-18	324
9	-46	2116
10	-40.5	1640.25
11	-37.5	1406.25
12	17	289
13	34.5	1190.25
14	13.5	182.25
15	-3	9
16	25.5	650.25
17	30	900
18	-24.5	600.25
19	-49.5	2450.25
20	-3	9
21	-13.5	182.25
22	0	0
23	-3	9
24	38	1444
25	-43	1849
26	0.5	.25
27	43	1849
28	32.5	1056.25
29	30.5	930.25
30	5	25

25 x .25 = 625

	d	d^2
31	8	64
32	12	144
33	-23	529
34	42.5	1806.25
35	-7.5	56.25
36	11	121
37	-7.5	56.25
38	4.5	20.25
39	-11	121
40	-11	121
41	36.5	1332.25
42	-30.5	930.25
43	-17.5	306.25
44	29	841
45	20	400
46	-5	25
47	7.5	56.25
48	15.5	240.25
49	34	1156
50	37.5	1406.25
51	-24	576

34263

34263

8.5 - 50.5 = -42
18.5 - 48 = -29.5
18.5 - 2 = 16.5
31.5 - 44 = -12.5
11.5 - 9.5 = 2
26.5 - 39.5 = -13
36.5 - 33 = 3.5
11.5 - 29.5 = -18
2 - 48 = -46
3.5 - 44 = -40.5
6.5 - 44 = -37.5
29 - 12 = 17
39.5 - 5 = 34.5
33.5 - 20 = 13.5
26.5 - 29.5 = -3
48.5 - 23 = 25.5
44.5 - 14.5 = 30
15 - 39.5 = -24.5
1 - 50.5 = -49.5
36.5 - 39.5 = -3
22 - 35.5 = -13.5
39.5 - 39.5 = 0
15 - 18 = -3
51 - 13 = 38
5 - 48 = -43
22 - 21.5 = 0.5
47 - 4 = 43
48.5 - 16 = 32.5
31.5 - 1 = 30.5
29 - 24 = 5
43.5 - 35.5 = 8
15 - 3 = 12

6.5 - 29.5 = -23
11.5 - 35.5 = -24
50 - 7.5 = 42.5
18.5 - 26 = -7.5
29 - 18 = 11
18.5 - 26 = -7.5
36.5 - 32 = 4.5
24.5 - 35.5 = -11
35 - 46 = -11
46 - 9.5 = 36.5
11.5 - 42 = -30.5
8.5 - 26 = -17.5
36.5 - 7.5 = 29
41.5 - 21.5 = 20
24.5 - 29.5 = -5
22 - 14.5 = 7.5
33.5 - 18 = 15.5
45 - 11 = 34
43.5 - 6 = 37.5

$d^2 = 34263$

#Question 2: The following data set was taken from a bivariate (X, Y) population.

A1 <- c(15,18,24,29)

A2 <- c(38,55,53,125)

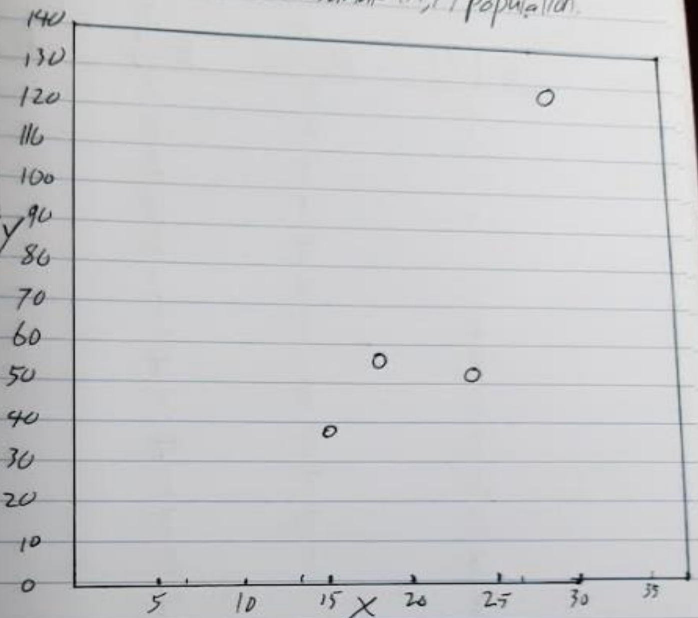
```
#For observed data, r* = 0.88129394, 24 permutations.
```

#(a) Sketch, by hand, a scatter plot of Y versus X.

#(b) Estimate Kendall's tau for these data by hand.

② The following data set was taken from a bivariate (X, Y) population.

	X	Y
1	15	38
2	18	55
3	24	53
4	29	125



T = 5, p-value = 0.1667

tau: = $\hat{\tau} = 0.67$

$$\hat{\tau} = \frac{4T}{n(n-1)} - 1$$

T = No. of pairs with a positive sign

(b)

Slopes	
1 vs 2	+
1 vs 3	+
1 vs 4	+
2 vs 3	−
2 vs 4	+
3 vs 4	+

$$\hat{\tau} = \frac{2}{n(n-1)} \times (\text{Pairs of } (+)\text{'s} - \text{Pairs of } (-)\text{'s})$$

$$\frac{2}{4(3)} \times (5-1) = \frac{2(4)}{4(3)} = \hat{\tau} = 0.6666667$$

(c) Here, we are doing a one-sided test against the alternative $H_A: \tau > 0$.
Observed $\hat{\tau}$ = 0.67 with p-value = 0.1667
We fail to reject the null hypothesis & conclude that we have no evidence of positive association between X & Y.

(15,18,24,29)

$R(X_1)$	$R(X_2)$	$R(X_3)$	$R(X_4)$		Association $\hat{r}^*$
1(15)	2(18)	3(24)	4(29)		Estimates
$R(Y_1)$	$R(Y_2)$	$R(Y_3)$	$R(Y_4)$		
1(38)	2(53)	3(55)	4(125)	*	0.88
1(38)	2(53)	4(125)	3(55)		0.40
1(38)	3(55)	2(53)	4(125)	*	0.86
1(38)	3(55)	4(125)	2(53)		0.37
1(38)	4(125)	2(53)	3(55)		-0.19
1(38)	4(125)	3(55)	2(53)		-0.20
2(53)	1(38)	3(55)	4(125)	*	0.82
2(53)	1(38)	4(125)	3(55)		0.34
2(53)	3(55)	1(38)	4(125)	-	0.66
2(53)	3(55)	4(125)	1(38)		0.08
2(53)	4(125)	1(38)	3(55)		-0.37
2(53)	4(125)	3(55)	1(38)		-0.44
3(55)	1(38)	2(53)	4(125)	*	0.79
3(55)	1(38)	4(125)	2(53)		0.30
3(55)	2(53)	1(38)	4(125)	-	0.66
3(55)	2(53)	4(125)	1(38)		0.07
3(55)	4(125)	1(38)	2(53)		-0.41
3(55)	4(125)	2(53)	1(38)		-0.51
4(125)	1(38)	2(53)	3(55)		-0.54
4(125)	1(38)	3(55)	2(53)		-0.56
4(125)	2(53)	1(38)	3(55)		-0.67
4(125)	2(53)	3(55)	1(38)		-0.78
4(125)	3(55)	1(38)	2(53)		-0.70
4(125)	3(55)	2(53)	1(38)		-0.81

1,2,3,4 = (38,53,55,128)

0.67

- For this upper-tailed test, count amount of permuted correlations $\geq \hat{r} =$ [illegible]
  There are 4 of 24 permuted $\hat{r}^*$ values as extreme or more extreme than our observed values.
- Thus, p-value = 4/24 = 0.166 ∴ Reject Ho, no evidence of positive association

#(c) Perform an appropriate hypothesis test. You may use the permutations function in R's gtools library.

```
library(gtools)

perms <- permutations(4, 4, A2)

N <- nrow(perms)

estimates <- numeric(N)


for(i in 1:N) {

  estimates[i] <- cor(A1, perms[i,], method="pearson")

}

output <- cbind(perms, estimates)

output


robs <- cor(A2, A1)

pval <- sum(estimates >= robs)/N # greater than

pval # fail to reject

robs

sum(estimates <= robs)/N # more than

sum(abs(estimates) >= abs(robs))/N # two-sided
```

#(d) Kendall's tau is an appropriate measure of association for these data as opposed to Pearson's Correlation because:

#With Pearson's Correlation Coefficient it goes with the assumptions that:

#Each observation should have a pair of values,

#Each variable should be continuous,

#It shouldn't have outliers

#Kendall correlation is best used when there are small samples or some outliers. Which this data is.

#3. The diameter and height of individuals of a particular type of plant were recorded in CA4 Opt Plant.csv. You may use R to aid doing this question.

CA4_Opt_Plant <- read_csv("SharedFiles/ST204/Data/CA4_Opt_Plant.csv")

CA4_Opt_Plant

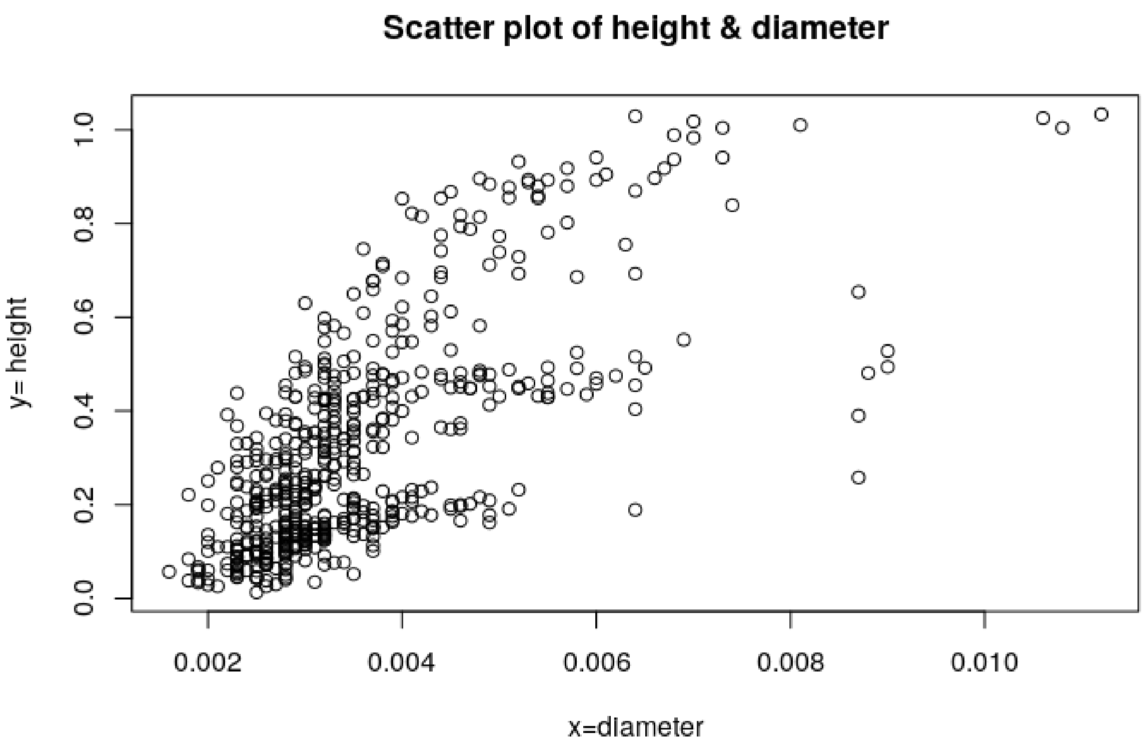
#(a) Create a scatter plot to illustrate the relationship between height & diameter, and estimate Pearson's correlation, Spearman's rank correlation, & Kendall's tau between the two variables.

plot(CA4_Opt_Plant$diameter, CA4_Opt_Plant$height, ylab="y= height", xlab="x=diameter",main ="Scatter plot of Height versus Diameter")

with(CA4_Opt_Plant, cor.test(diameter,height,method = "pearson")) #0.69

with(CA4_Opt_Plant, cor.test(diameter,height,method= "spearman")) #0.69

with(CA4_Opt_Plant, cor.test(diameter,height,method= "kendall")) #0.51



#(b) Construct 95% confidence intervals for each type of association using a bootstrap approach,

#with 5000 bootstrap samples.

with(CA4_Opt_Plant, cor.boot.ci(diameter,height, method="pearson", conf=0.95, nbs=5000)) #(0.64 - 0.74)

with(CA4_Opt_Plant, cor.boot.ci(diameter,height, method="spearman",conf=0.95, nbs=5000)) #(0.64 - 0.73)

with(CA4_Opt_Plant, cor.boot.ci(diameter,height, method="kendall",conf=0.95, nbs=5000)) #(0.47 - 0.55)

#(c) Need to Interpret?(Explain) The CIs.

#The Pearson & Spearman test are very similar in confidence intervals.

#The Kendall's Confidence interval considerable lower than the other two.

#Computing the Kendall association between R(A1) and R(A2) is exactly equivalent to computing the Kendall association between A1 and A2 as it implicitly uses ranks anyway

#(d) repeat (b), change .057 to 57.

#Which measure is the most sensitive to the influence of the outlier.

CA4_Opt_Plant[1,2] <- 57

with(CA4_Opt_Plant, cor.boot.ci(diameter,height, method="pearson", conf=0.95, nbs=5000)) #(-0.08 - 0.73)

with(CA4_Opt_Plant, cor.boot.ci(diameter,height, method="spearman",conf=0.95, nbs=5000)) #(0.62 - 0.72)

with(CA4_Opt_Plant, cor.boot.ci(diameter,height, method="kendall",conf=0.95, nbs=5000)) #(0.46 - 0.55)

#The one most sensitive to the influence of the outlier is the pearson test. Going from (0.64 - 0.74) to (-0.08 - 0.73), #A difference of (.72,.01)