DS 09/21 Statistics Session-4

DS Statistics Session-4
Training Clarusway
Pear Deck - September 29, 2021 at 0:59PM

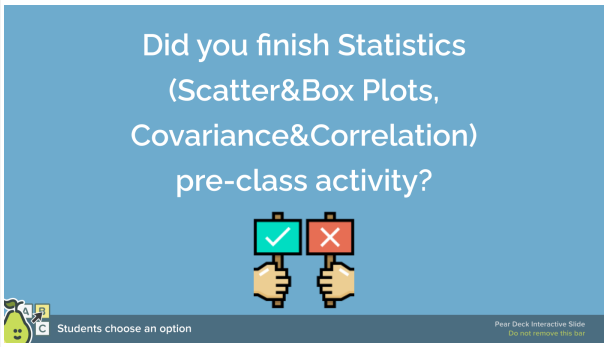
Part 1 - Summary

Use this space to summarize your thoughts on the lesson

Part 2 - Responses

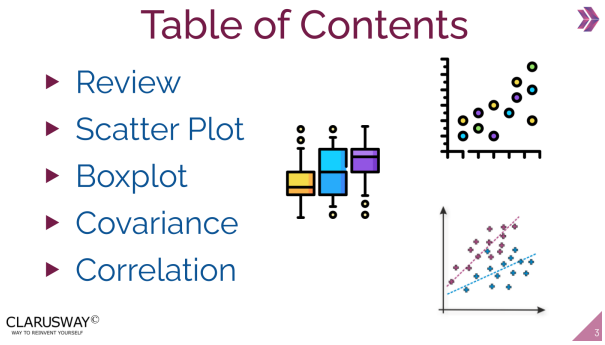
Slide 1

Use this space to take notes:

Slide 2	Your Response
	You Chose • **I finished partially.** Other Choices • I finished completely. • No, I didn't finish.

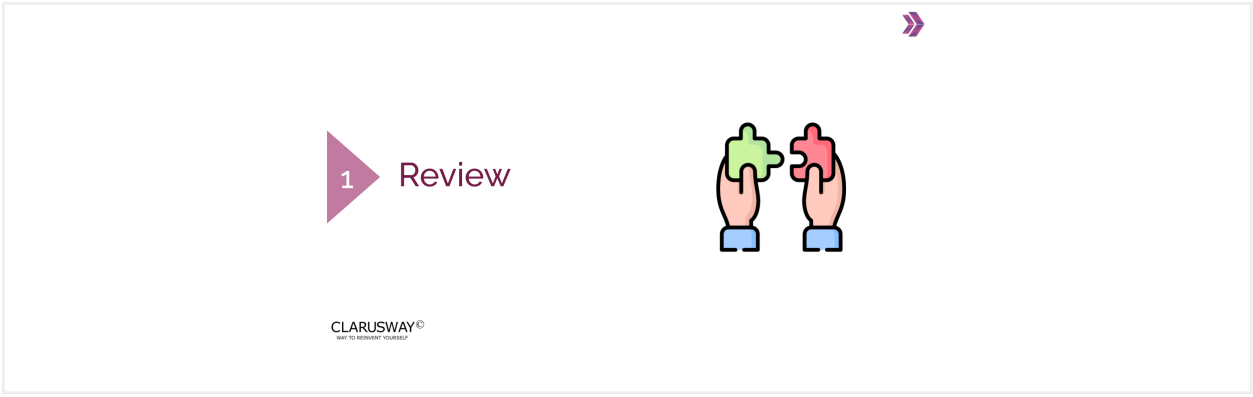
Use this space to take notes:

Slide 3



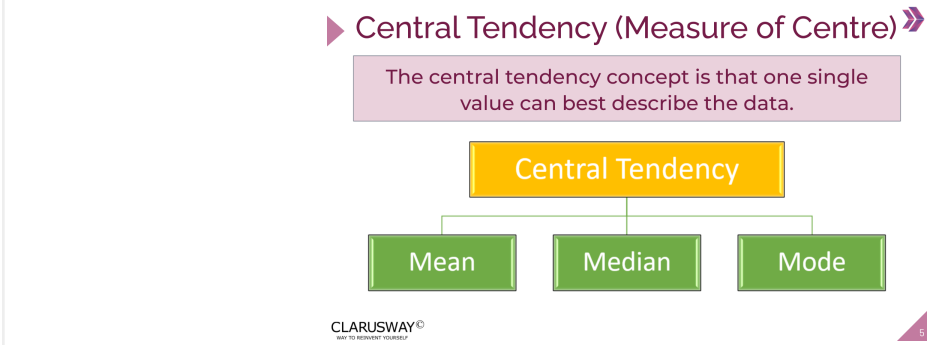
Use this space to take notes:

Slide 4



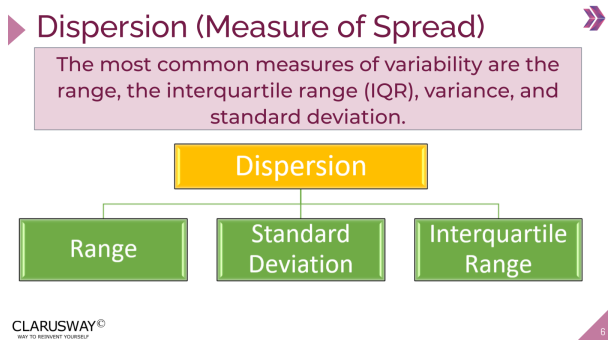
Use this space to take notes:

Slide 5



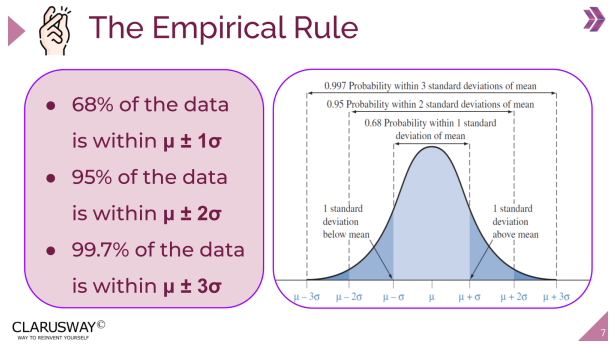
Use this space to take notes:

Slide 6



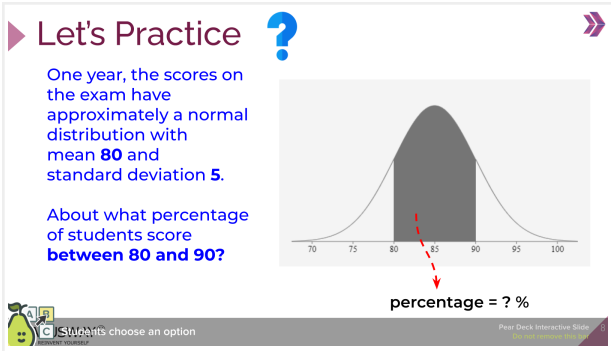
Use this space to take notes:

Slide 7



Use this space to take notes:

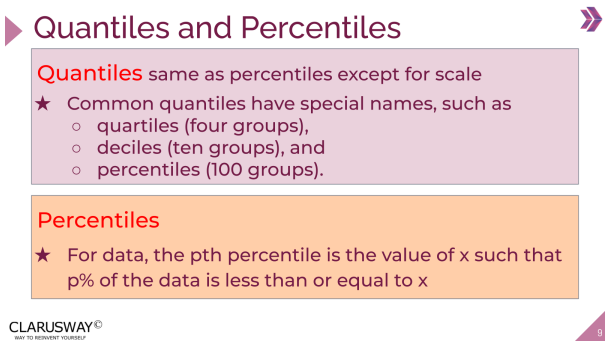
Slide 8	Your Response
	You Chose • **68** Other Choices



- 95
- 99.7
- Don't know

Use this space to take notes:

Slide 9

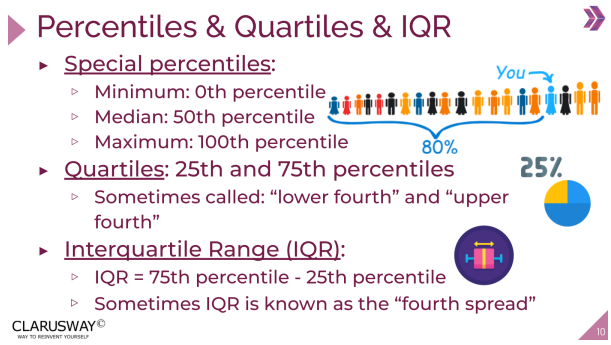


Link(s) on this slide:

- https://en.wikipedia.org/wiki/Quartile
- https://en.wikipedia.org/wiki/Decile
- https://en.wikipedia.org/wiki/Percentile

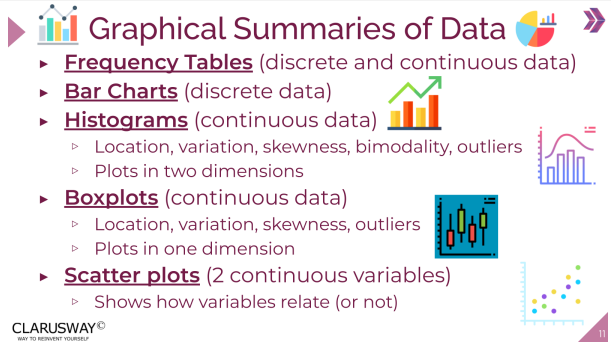
Use this space to take notes:

Slide 10



Use this space to take notes:

Slide 11

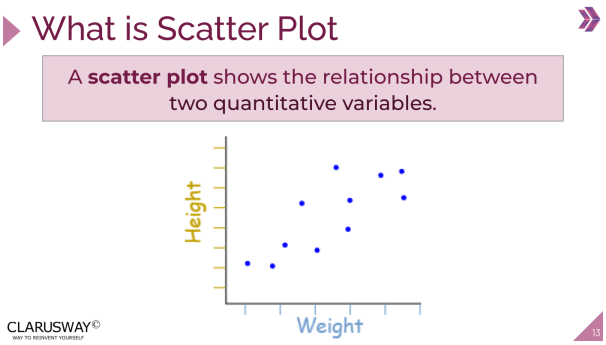


Use this space to take notes:

Slide 12

Use this space to take notes:

Slide 13

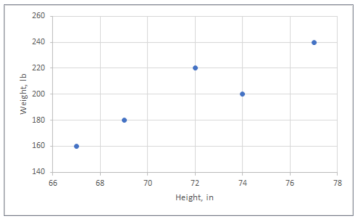


Use this space to take notes:

Slide 14

How to make a Scatter Plot

Height, in	Weight, lb
67	160
72	220
77	240
74	200
69	180



Scatter plots reveal patterns in bivariate data.

CLARUSWAY©
WAY TO REINVENT YOURSELF

14

Use this space to take notes:

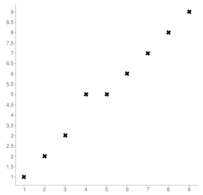
Slide 15

Patterns of Data in Scatter Plot

Linearity | Slope | Strength | Unusual Features

Linearity refers to whether a data pattern is linear (straight) or nonlinear (curved).

Linear

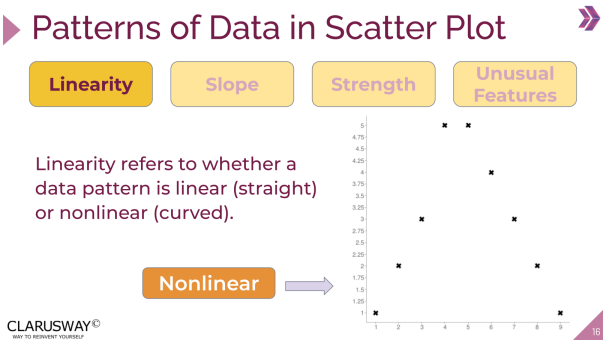


CLARUSWAY©
WAY TO REINVENT YOURSELF

15

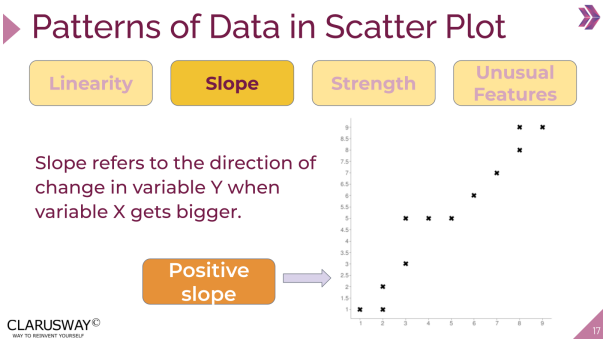
Use this space to take notes:

Slide 16



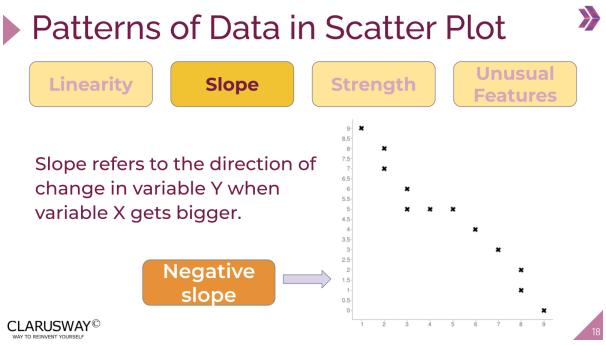
Use this space to take notes:

Slide 17



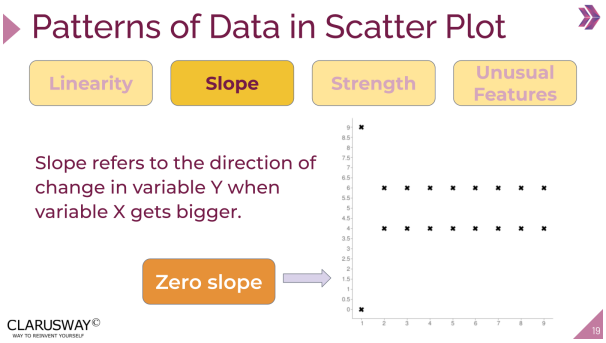
Use this space to take notes:

Slide 18



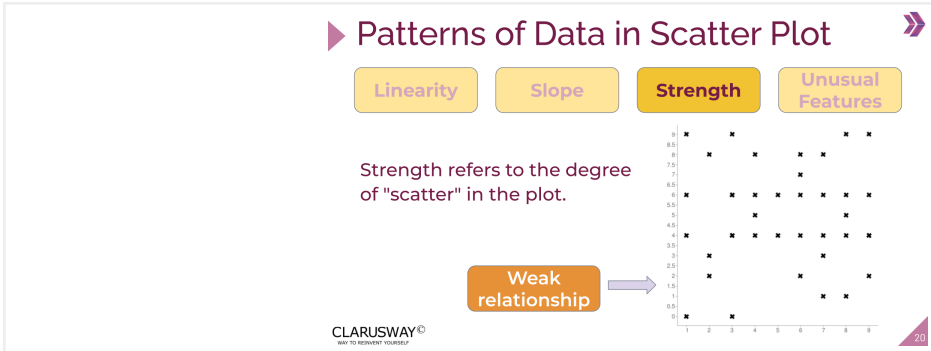
Use this space to take notes:

Slide 19



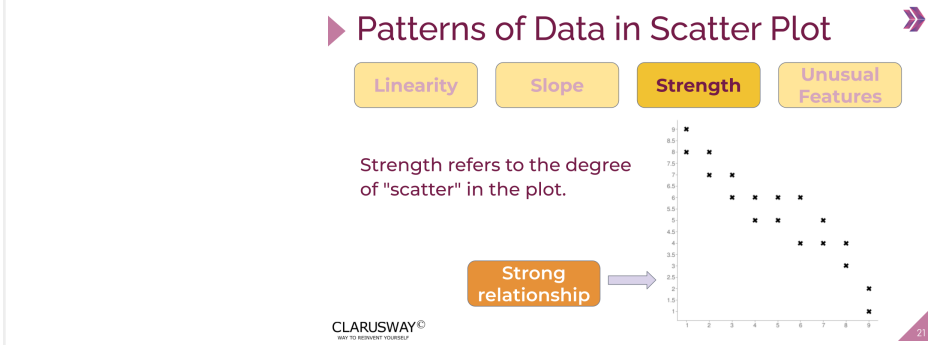
Use this space to take notes:

Slide 20



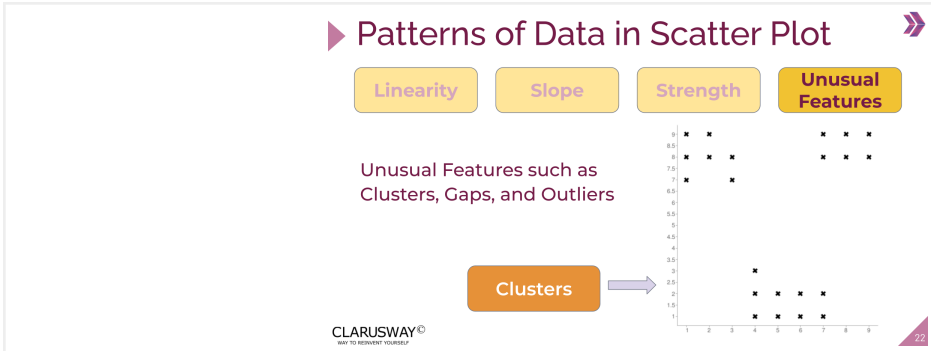
Use this space to take notes:

Slide 21



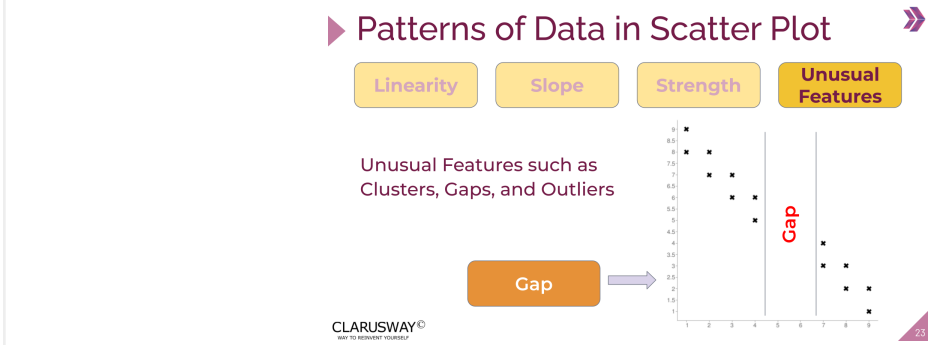
Use this space to take notes:

Slide 22



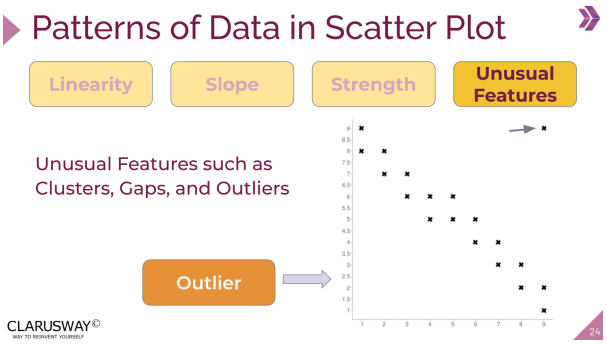
Use this space to take notes:

Slide 23

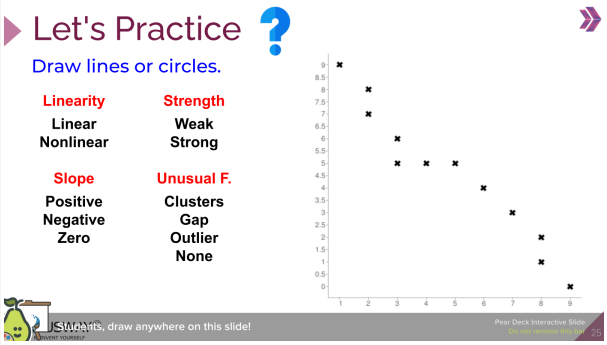
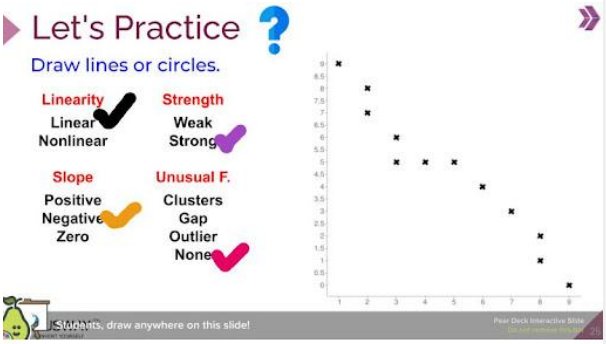


Use this space to take notes:

Slide 24

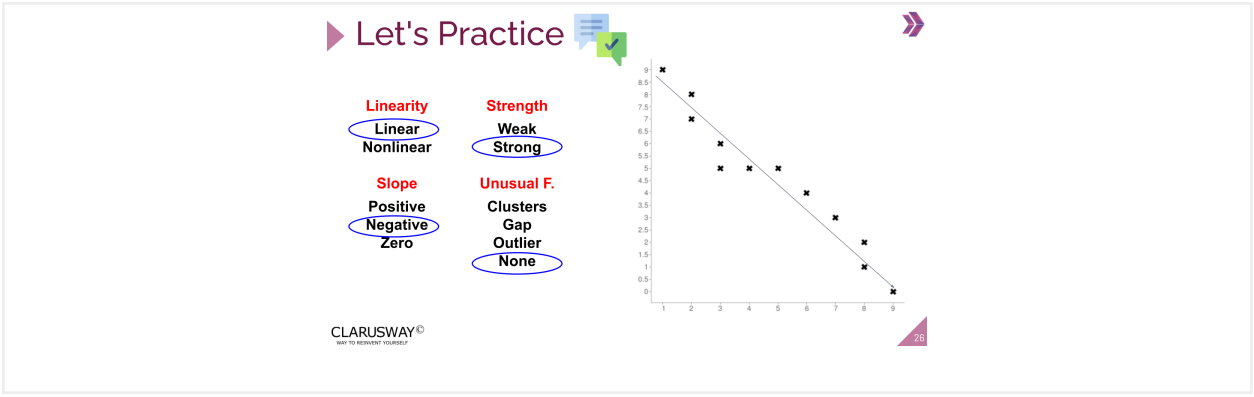


Use this space to take notes:

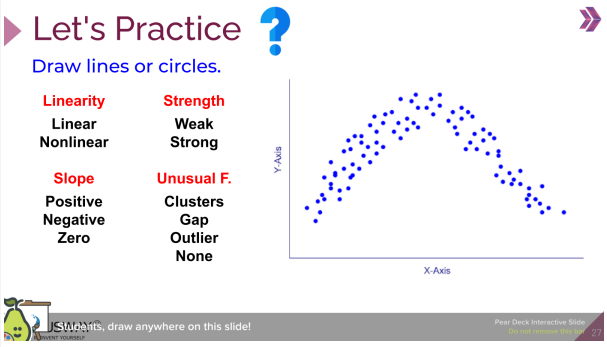
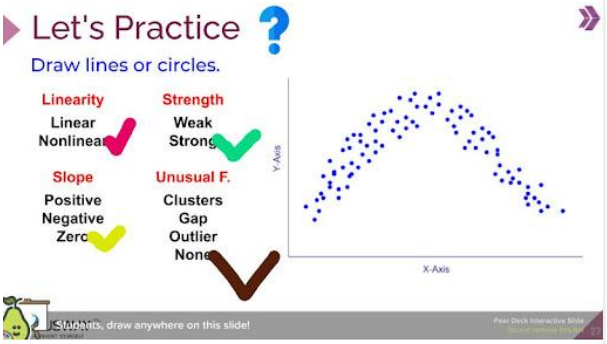
Slide 25	Your Response
	

Use this space to take notes:

Slide 26

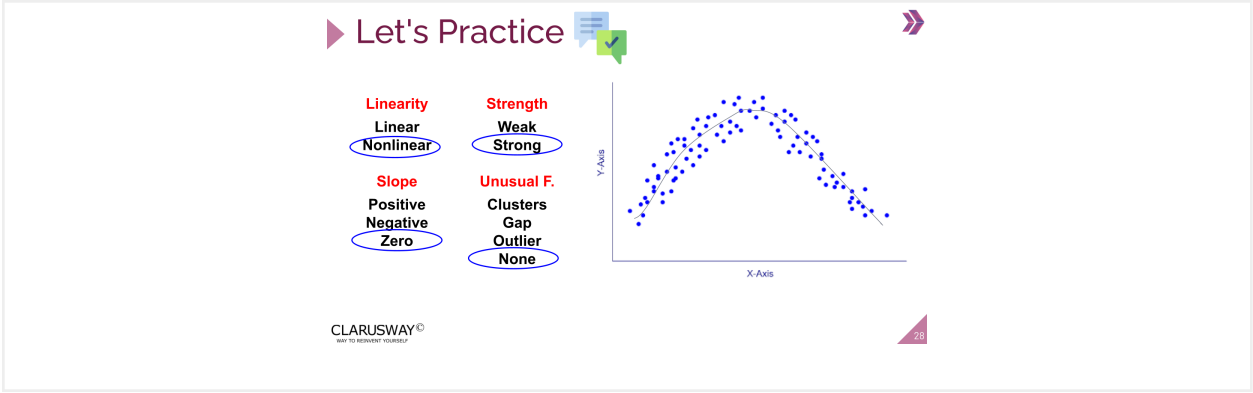


Use this space to take notes:

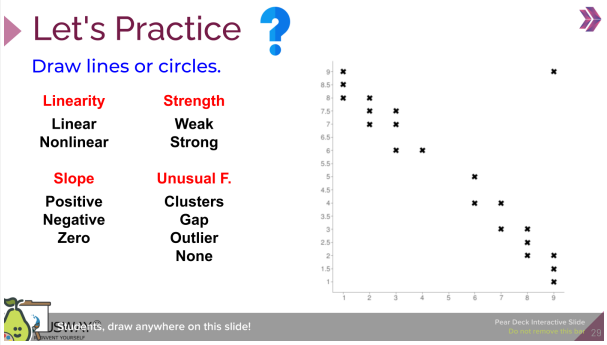
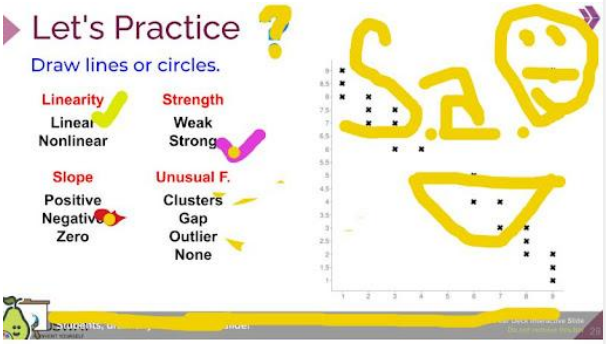
Slide 27	Your Response
	

Use this space to take notes:

Slide 28

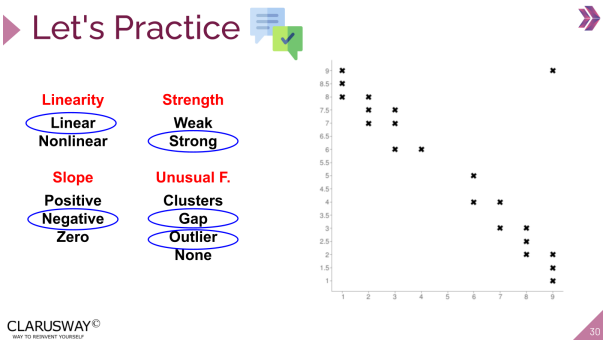


Use this space to take notes:

Slide 29	Your Response
	

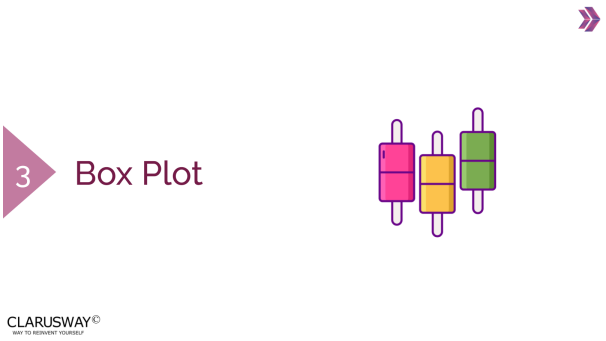
Use this space to take notes:

Slide 30



Use this space to take notes:

Slide 31



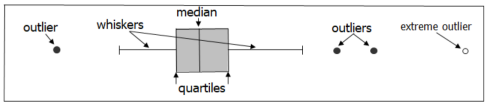
Use this space to take notes:

Slide 32

What is Box Plot

A **box plot** is a method for graphically depicting groups of numerical data through their *quartiles*.

A box plot generally shows
- ★ median
- ★ 1st quartile (Q1)
- ★ 3rd quartile (Q3)
- ★ outliers



CLARUSWAY©

32

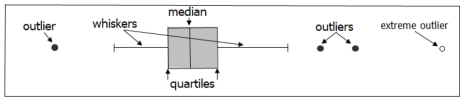
Use this space to take notes:

Slide 33

Box Plot

Boxplots show distribution in one dimension
- ★ Only useful for continuous variables
- ★ Good for comparing distributions of a continuous variable between categorical groups
- ★ Will not show multiple modes

Box plots are also known as **box and whisker plots**.



CLARUSWAY©

33

Use this space to take notes:

Slide 34

Box Plot

- Different variants exist, some do whiskers to min and max, others go to the lowest data that is at most 1.5*IQR from the edges of the rectangle.
- More outlying observations are identified as “outliers” or “extreme outliers” if greater than 3*IQR from the edges of the rectangle.

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 35

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 36

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

CLARUSWAY©
WAY TO REINVENT YOURSELF

36

Use this space to take notes:

Slide 37

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

Step 3: Find the quartiles.

25 28 29 29 30 34 35 35 37 38

Q1 = 29 *Q2 = 35*

CLARUSWAY©
WAY TO REINVENT YOURSELF

37

Use this space to take notes:

Slide 38

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

Step 3: Find the quartiles.

25 28 29 29 30 34 35 35 37 38

Q1 = 29 *Q2 = 35*

Step 4: Find the min and the max.

Min = 25 *Max = 38*

CLARUSWAY©

38

Use this space to take notes:

Slide 39

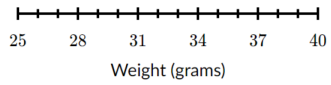
How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 1: Scale / label an axis that fits the five-number.

25 28 29 29 30 34 35 35 37 38

Min = 25 *Max = 38*



CLARUSWAY©

39

Use this space to take notes:

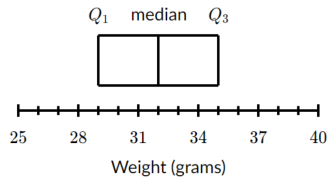
Slide 40

How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 2: Draw a box from *Q1* to *Q3* with a vertical line through the *median*.

25 28 29 29 30 34 35 35 37 38



CLARUSWAY©

Use this space to take notes:

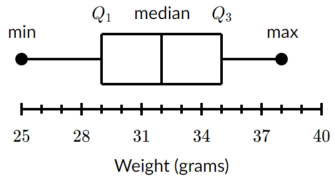
Slide 41

How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 3: Draw a whisker from *Q1* to the *min* and from *Q3* to the *max*.

25 28 29 29 30 34 35 35 37 38



CLARUSWAY©

Use this space to take notes:

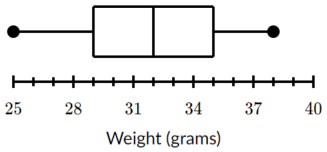
Slide 42

How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

We don't need the labels on the final product:

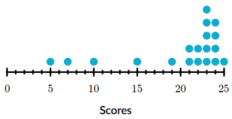
25 28 29 29 30 34 35 35 37 38



CLARUSWAY©

42

Use this space to take notes:

Slide 43	Your Response
## Identifying Outliers An **outlier** is a data point that lies outside the overall pattern in a distribution.  **Question:** How many outliers do you see? Students, enter a number! Pear Deck Interactive Slide Do not remove this bar 43	• 4

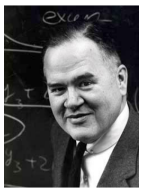
Use this space to take notes:

Slide 44

Identifying Outliers (1.5xIQR Rule)

1.5 IQR Rule: If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an outlier.

According to John Tukey,
1 IQR seemed like too little and
2 IQRs seemed like too much.



CLARUSWAY©
WAY TO REINVENT YOURSELF

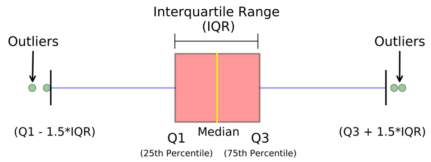
44

Use this space to take notes:

Slide 45

Boxplot with 1.5xIQR Rule

1.5 IQR Rule: If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an outlier.



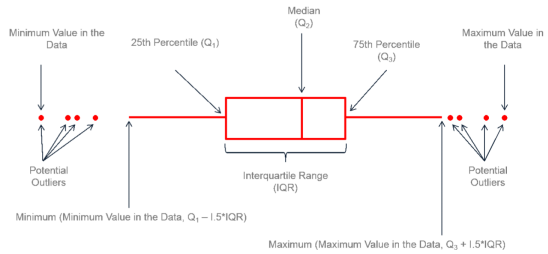
CLARUSWAY©
WAY TO REINVENT YOURSELF

45

Use this space to take notes:

Slide 46

Boxplot with 1.5xIQR Rule



CLARUSWAY©

Use this space to take notes:

Slide 47

How to make a Box Plot (1.5xIQR Rule)

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 1: Find the median, quartiles, and interquartile range.

Median = 23
Q1 = 19
Q3 = 24
IQR = Q3 - Q1 = 24 - 19 = 5

CLARUSWAY©

Use this space to take notes:

Slide 48

How to make a Box Plot (1.5xIQR Rule)

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 1: Find the median, quartiles, and interquartile range.

Step 2: Calculate 1.5 x IQR below the first quartile and check for low outliers.

Q1 - 1.5 x IQR = 19 - 1.5 x 5
= 11.5

Low Outliers: 5 7 10

CLARUSWAY©
WAY TO REINVENT YOURSELF

48

Use this space to take notes:

Slide 49

How to make a Box Plot (1.5xIQR Rule)

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 1: Find the median, quartiles, and interquartile range.

Step 2: Calculate 1.5 x IQR below the first quartile and check for low outliers.

Step 3: Calculate 1.5 x IQR above the third quartile and check for high outliers.

Q3 + 1.5 x IQR = 24 + 1.5 x 5
= 31.5

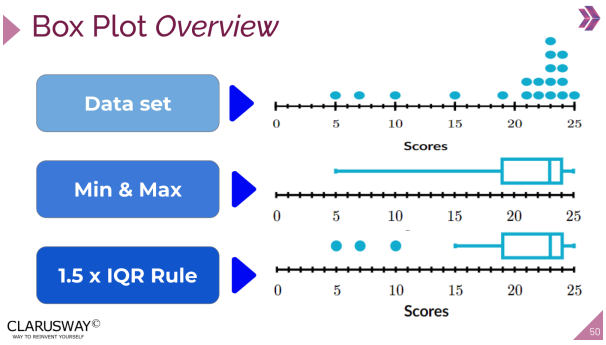
High Outliers: None

CLARUSWAY©
WAY TO REINVENT YOURSELF

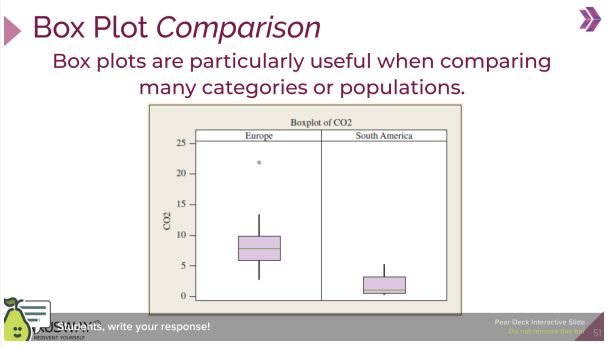
49

Use this space to take notes:

Slide 50



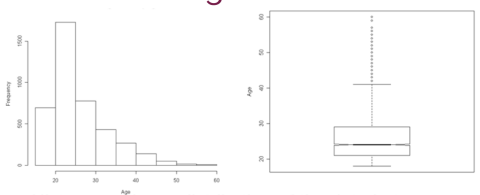
Use this space to take notes:

Slide 51	Your Response
	**Answer 1:** Yorumsuz :)

Use this space to take notes:

Slide 52

Box Plot vs. Histogram



- Histogram shows distribution of the data in two dimensions – the boxplot is in one dimension
  - Histogram shows frequency of observations within ranges
  - Boxplot only shows summary statistics

CLARUSWAY©

52

Use this space to take notes:

Slide 53	Your Response

Let's Practice *Draw Box Plot*

3	6	8	9	9	10	12	14	19

0 5 10 15 20

Students, draw anywhere on this slide!

Pear Deck Interactive Slide

Do not remove this bar

53

Let's Practice *Draw Box Plot*

3 6 8 9 9 10 12 14 19

0 5 10 15 20

Students, draw anywhere on this slide!

Use this space to take notes:

Slide 54

4 Covariance

CLARUSWAY©

Use this space to take notes:

Slide 55

Intro

- ★ We are interested in how variables behave together.
- ★ For example, suppose we want to investigate the relationship between smoking and lung capacity.
- ★ We might ask a group of people about their smoking habits, and measure their lung capacities.

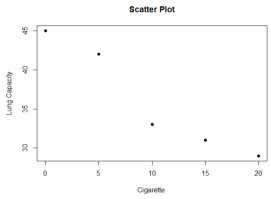
CLARUSWAY©

Use this space to take notes:

Slide 56

Scatter Plot

Cigarette (X)	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29



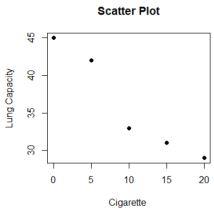
Scatter plots reveal patterns in bivariate data.

Use this space to take notes:

Slide 57

Scatter Plot

- We can see easily from the graph that as smoking goes up, lung capacity tends to go down.
- The two variables covary in opposite directions.



Use this space to take notes:

Slide 58

Covariance

- ★ **Covariance** is a measure of how much two random variables change together.
- ★ The covariance gives info on the strength of the linear relationship between two variables. Since it is not a scaled metric, it is not useful to compare two different values of covariances.

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

CLARUSWAY©
WAY TO REINVENT YOURSELF

58

Use this space to take notes:

Slide 59

Covariance Interpretation

x and y tend to move in the same direction	Cov(x,y) > 0
x and y tend to move in opposite directions	Cov(x,y) < 0
x and y are independent	Cov(x,y) = 0

CLARUSWAY©
WAY TO REINVENT YOURSELF

59

Use this space to take notes:

Slide 60

Covariance *Example*

Table shows years of education (S) and hourly earnings in dollars (Y).

Observation	*S*	*Y*	*Observation*	*S*	*Y*
1	15	17.24	11	17	15.38
2	16	15.00	12	12	12.70
3	8	14.91	13	12	26.00
4	6	4.50	14	9	7.50
5	15	18.00	15	15	5.00
6	12	6.29	16	12	21.63
7	12	19.23	17	16	12.10
8	18	18.69	18	12	5.55
9	12	7.21	19	12	7.50
10	20	42.06	20	14	8.00

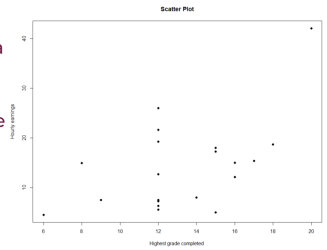
CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 61

Covariance *Example*

- Figure shows the data plotted as a scatter plot.
- You can see that there is a weak positive association between the two variables.



CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 62

Covariance *Example*

Observation	S	Y	$(S - \bar{S})$	$(Y - \bar{Y})$	$(S - \bar{S})(Y - \bar{Y})$
1	15	17.24	1.75	3.016	5.277
2	16	15.00	2.75	0.775	2.133
3	8	14.91	-5.25	0.685	-3.599
4	6	4.5	-7.25	-9.725	70.503
5	15	18.00	1.75	3.776	6.607
6	12	6.29	-1.25	-7.935	9.918
7	12	19.23	-1.25	5.006	-6.257
8	18	18.69	4.75	4.466	21.211
9	12	7.21	-1.25	-7.015	8.768
10	20	42.06	6.75	27.836	187.890
11	17	15.38	3.75	1.156	4.333
12	12	12.70	-1.25	-1.525	1.906
13	12	26.00	-1.25	11.776	-14.719
14	9	7.50	-4.25	-6.725	28.579
15	15	5.00	1.75	-9.225	-16.143
16	12	21.63	-1.25	7.406	-9.257
17	16	12.10	2.75	-2.125	-5.842
18	12	5.55	-1.25	-8.675	10.843
19	12	7.50	-1.25	-6.725	8.406
20	14	8.00	0.75	-6.225	-4.668

$$\text{Cov}(x, y) = \frac{1}{n}[(x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Therefore, covariance is calculated as

305.888 / 19 = 16.0993

CLARUSWAY©
WAY TO REINVENT YOURSELF

62

Use this space to take notes:

Slide 63

5 Correlation

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 64

Correlation

The **correlation** summarizes the direction of the association between two quantitative variables and the strength of its linear (straight-line) trend.

Direction

- Positive : Move in same direction
- Negative : Move in opposite directions

Strength

- Weak : Widely spread
- Strong : Concentrated around a line

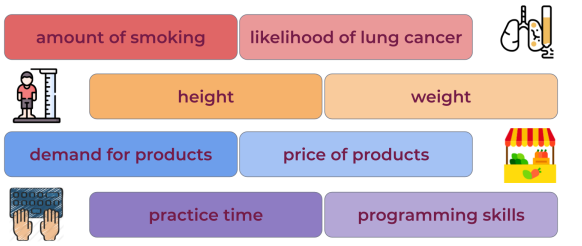
CLARUSWAY©
WAY TO REINVENT YOURSELF

64

Use this space to take notes:

Slide 65

Examples



CLARUSWAY©
WAY TO REINVENT YOURSELF

65

Use this space to take notes:

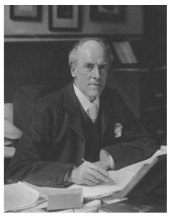
Slide 66

Correlation *Coefficient*

The correlation coefficient, ***r***, developed by *Karl Pearson* in the early 1900s.

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

n = the number of data points.



CLARUSWAY©
WAY TO REINVENT YOURSELF

66

Use this space to take notes:

Slide 67

Correlation *Coefficient*

Denoted by *r*, it takes values between *-1* and *+1*.

Strength

- Greater absolute value, stronger relationship
- Strongest relationship is correlation of -1 or 1
- Weakest relationship is correlation of zero.

Direction

- Sign of coefficient describes direction
- Positive sign: Variables move in same direction
- Negative sign: Move in opposite directions

CLARUSWAY©
WAY TO REINVENT YOURSELF

67

Use this space to take notes:

Slide 68

Correlation *Notation*

Sample Correlation	r
Population Correlation	ρ or R

CLARUSWAY©
WAY TO REINVENT YOURSELF

68

Use this space to take notes:

Slide 69

Correlation *Coefficient Formula*

Sample Correlation	$r = \frac{Cov(x,y)}{s_X s_Y}$
Population Correlation	$\rho = \frac{Cov(x,y)}{\sigma_X \sigma_Y}$

CLARUSWAY©
WAY TO REINVENT YOURSELF

69

Use this space to take notes:

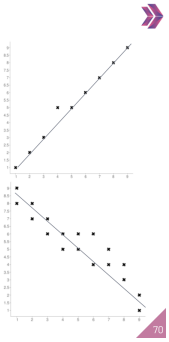
Slide 70

Linear Relationship

When the data points follow a roughly straight-line trend, the variables are said to have an approximately **linear relationship**.

On the line

Around the line



CLARUSWAY©
WAY TO REINVENT YOURSELF

70

Use this space to take notes:

Slide 71

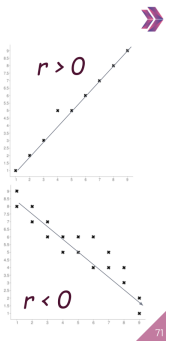
Direction

Slope determines the sign of the correlation coefficient.

When the slope of the line in the plot is negative, the correlation is negative; and vice versa.

Positive

Negative

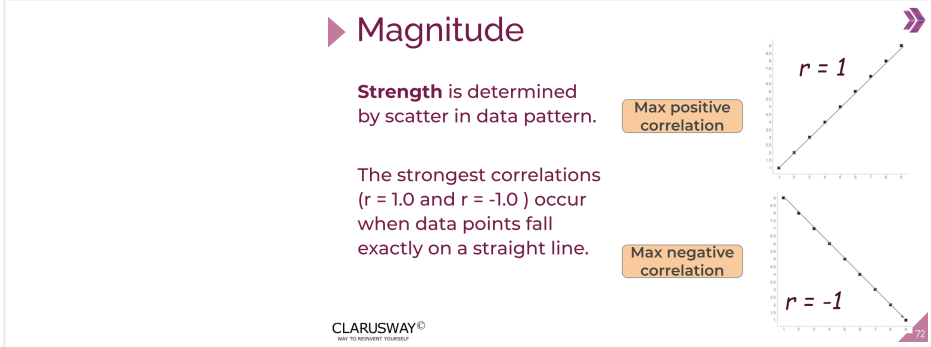


CLARUSWAY©
WAY TO REINVENT YOURSELF

71

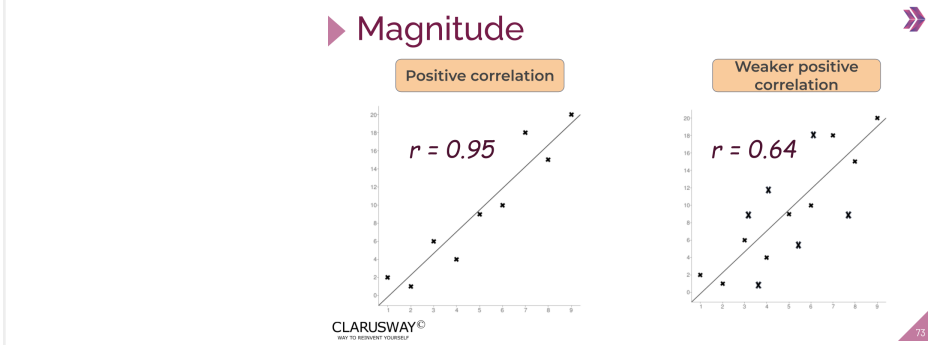
Use this space to take notes:

Slide 72



Use this space to take notes:

Slide 73

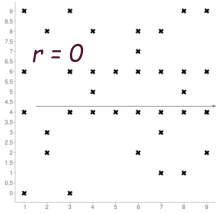


Use this space to take notes:

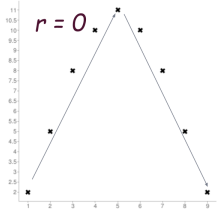
Slide 74

Zero Correlation

Random pattern



Curvilinear pattern



CLARUSWAY©

74

Use this space to take notes:

Slide 75	Your Response

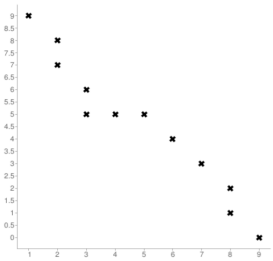
Patterns of Data in Scatter Plot

Draw lines or circles.

Linearity
Linear
Nonlinear

Strength
Weak
Strong

Slope
Positive
Negative
Zero



Students, draw anywhere on this slide!

Pear Deck Interactive Slide
Do not remove this bar

75

Use this space to take notes:

Slide 76

Examples of Approximate r Values

Graph A (r = 1.0): perfect positive correlation between x and y

Graph B (r = -1.0): perfect negative correlation between x and y

Graph C (r = 0.6): a moderately positive relationship: y tends to increase as x increases, but not necessarily at the steady rate we observed in Graph A

Graph D (r = -0.4): a relatively weak negative relationship: the correlation coefficient is closer to zero, negative r value so y tends to decrease as x increases

Graph E (r = 0): no relationship between x and y

CLARUSWAY©

76

Use this space to take notes:

Slide 77	Your Response
## Let's Practice Which pair of variables exhibits the strongest linear relationship? Which pair of variables exhibits the weakest linear relationship?	You Chose ● **Strongest: B | Weakest: E** Other Choices ● Strongest: C | Weakest: E ● Strongest: E | Weakest: B

Pair	Variable 1	Variable 2	Correlation
A	Internet Users	Facebook Users	0.507
B	Internet Users	Broadband Subscribers	0.949
C	Internet Users	Population	0.744
D	Facebook Users	Broadband Subscribers	0.619
E	Facebook Users	Population	0.097
F	Broadband Subscribers	Population	0.533

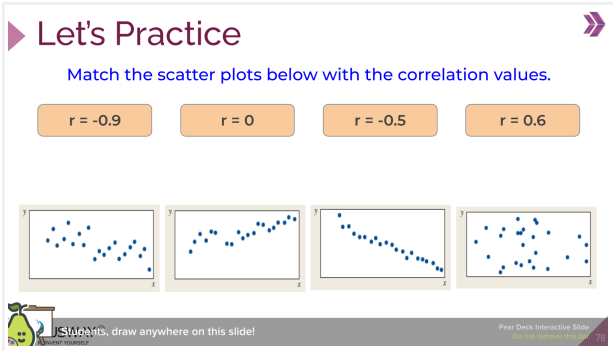
Students choose an option

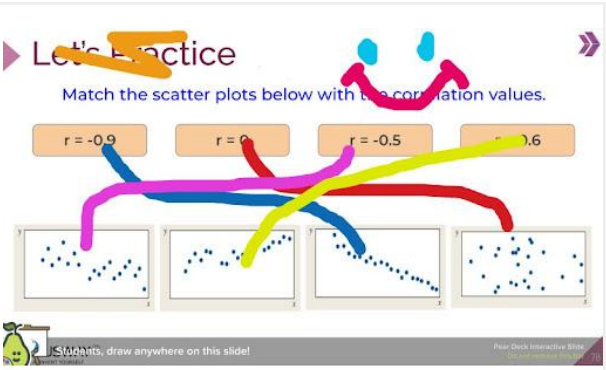
Pear Deck Interactive Slide
Do not remove this bar

77

Use this space to take notes:

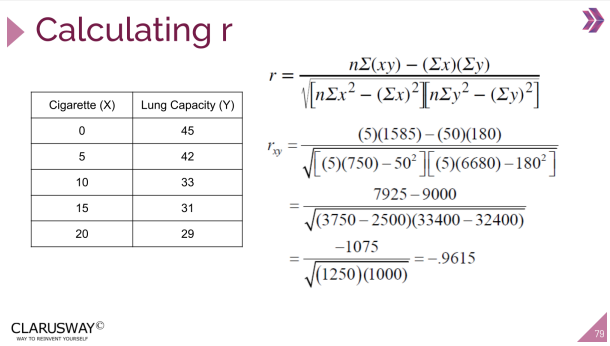
Slide 78	Your Response





Use this space to take notes:

Slide 79

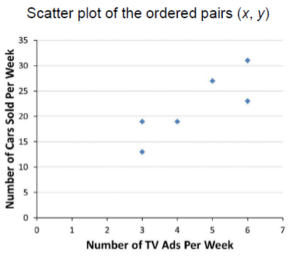


Use this space to take notes:

Slide 80

Calculating r

Week	Number of TV Ads x	Number of Cars Sold y
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19



CLARUSWAY©

80

Use this space to take notes:

Slide 81

Calculating r

Week	Number of TV Ads x	Number of Cars Sold y	xy	x^2	y^2
1	3	13	39	9	169
2	6	31	186	36	961
3	4	19	76	16	361
4	5	27	135	25	729
5	6	23	138	36	529
6	3	19	57	9	361
	$\Sigma x = 27$	$\Sigma y = 132$	$\Sigma xy = 631$	$\Sigma x^2 = 131$	$\Sigma y^2 = 3110$

CLARUSWAY©

81

Use this space to take notes:

Slide 82

Calculating r

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

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{(6)(631) - (27)(132)}{\sqrt{\left[(6)(131) - (27)^2\right]\left[(6)(3110) - (132)^2\right]}}$$

$$= \frac{222}{\sqrt{[57]\,[1236]}} = \frac{222}{265.43} = 0.836$$

CLARUSWAY©
WAY TO REINVENT YOURSELF

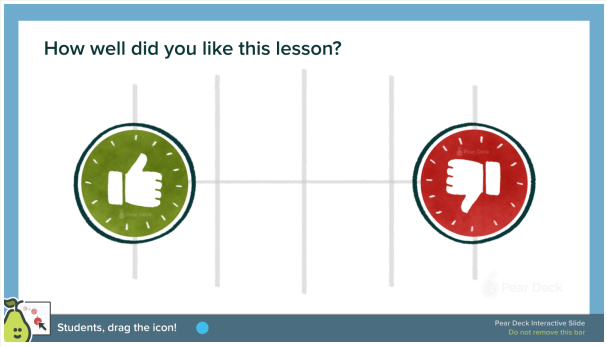
82

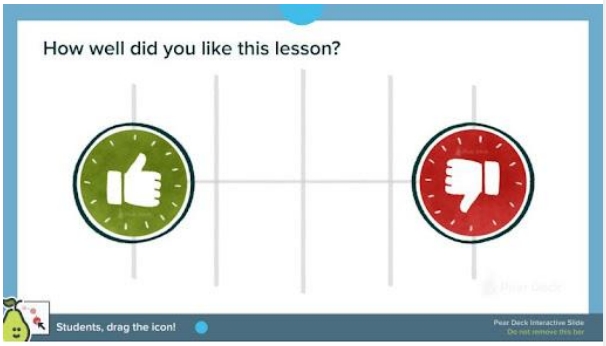
Use this space to take notes:

Slide 83	Your Response
## Let's Practice If one y value is changed, the correlation for the five pairs equals 1. Identify the y value and how it must be changed for this to happen. **Variable 1** / **Variable 2** 3 / 8 4 / 13 5 / 12 6 / 14 7 / 16 Students, write your response! Pear Deck Interactive Slide Do not remove this bar 83	

Use this space to take notes:

Slide 84	Your Response





Use this space to take notes:

Slide 85

Use this space to take notes: