Welcome to DATA 151

I’m so glad you’re here!

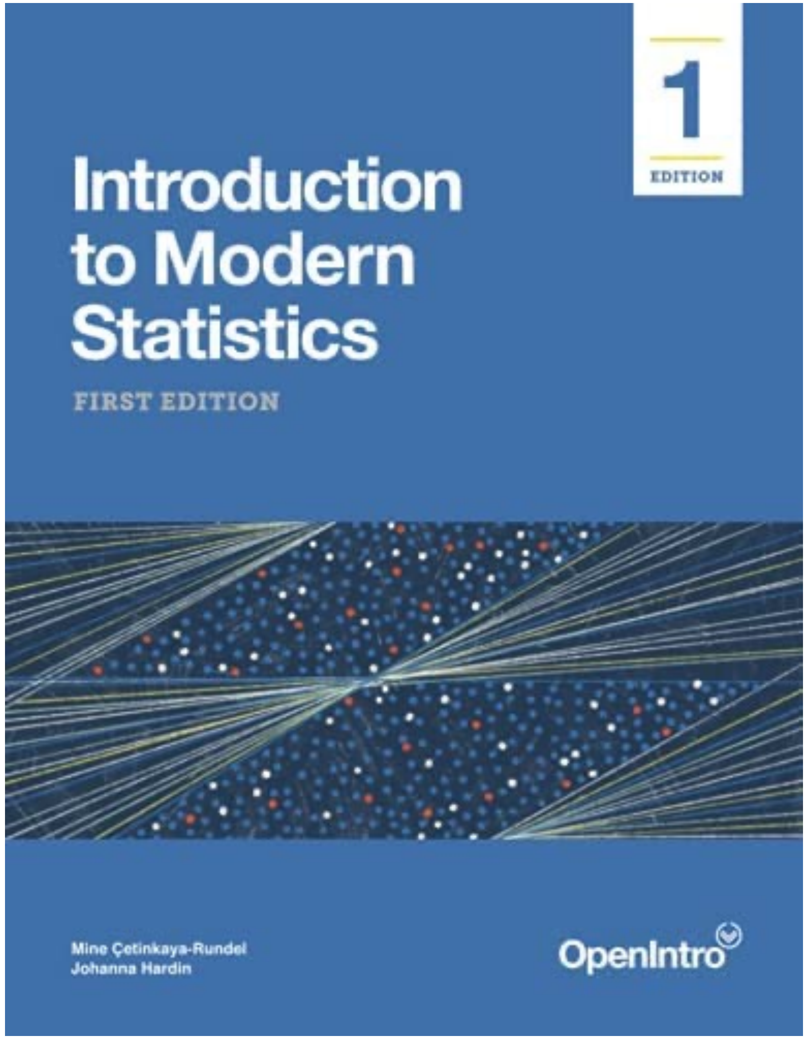
DATA 151: CLASS 11A
INTRODUCTION TO DATA SCIENCE (WITH R)

BIVARIATE RELATIONSHIPS

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

- Tuesday and Thursday:
  - Introduction to Modern Statistics
  - Ch 7: Relationships between two variables

HOMEWORK REMINDER

Due this week:

- ***NOTHING!***

HOMEWORK REMINDER

Due next week:

- ***DUE 11/17*** *Project Milestone #6*
- Relationships between two numeric
- ***DUE 11/17*** *HW #10: DC Correlation and Regression*

DESCRIBING RELATIONSHIPS BETWEEN TWO VARIABLES (BIVARIATE)

SCATTERPLOTS AND CORRELATION

- Have you ever thought to yourself
  - “Does age of a driver help explain accident deaths?”
  - “Does smoking influence life expectancy?”
  - “Can the number of hours a student studies for an exam help predict the score they receive on the exam?”

- In each of these questions there are two variables present. The relationship between them is being questioned. The variables each play a different role: One may help explain changes in the other

SCATTERPLOTS AND CORRELATION

- **Response variable (Y):** the variable one suspects is affected by the explanatory variable.
  - The variable that is of interest to study
  - In an experiment, this is the dependent variable

- **Explanatory variable (X)**: the variable whose effect one wants to study.
  - The explanatory variable (is thought to) explain or influence changes in a response variable.
  - In an experiment, this is the independent variable
  - "The explanatory variable HELPS EXPLAIN the response variable"

SCATTERPLOTS AND CORRELATION

Example 1: **"Does the interest rate help explain the number of loan applications?"**

- What two variables are under consideration in this question?

- Which of the two variables listed above is the *response* variable?

- Which of the two variables listed above is the *explanatory* variable?

SCATTERPLOTS AND CORRELATION

<u>Example 1:</u> **"Does the interest rate help explain the number of loan applications?"**

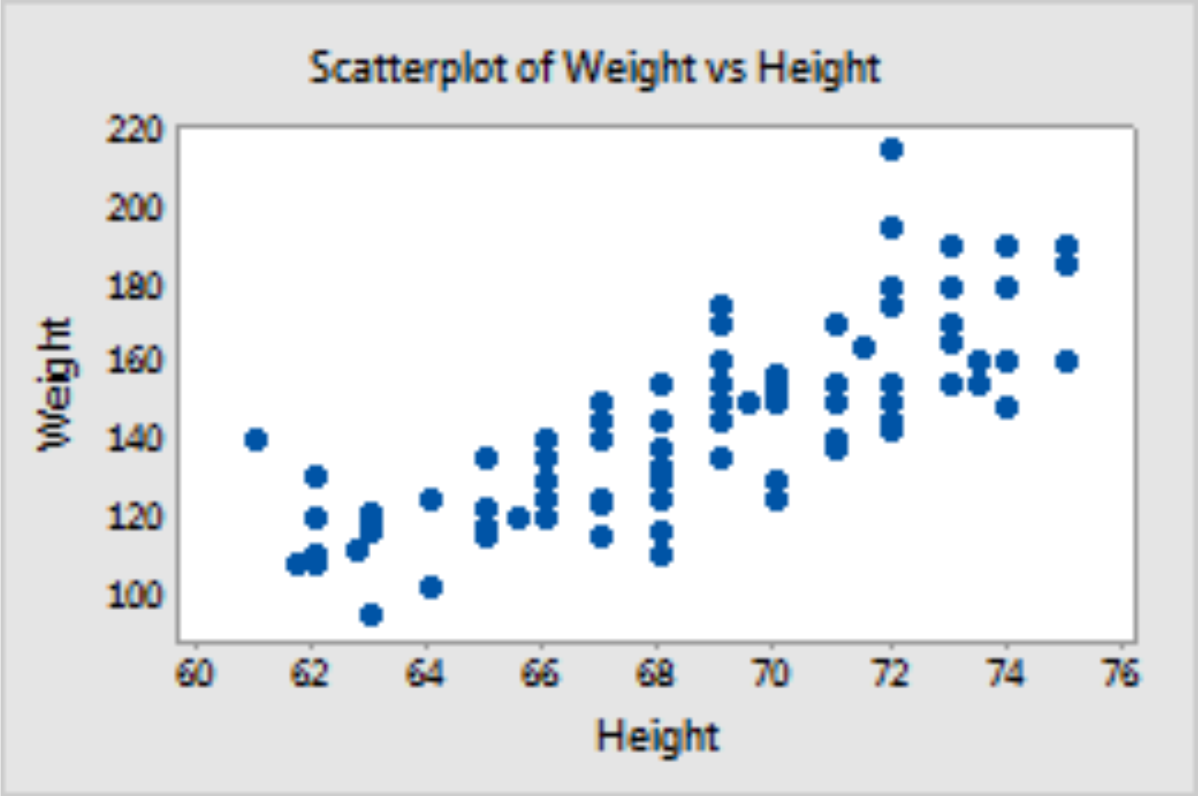
- What two variables are under consideration in this question?
  - Interest rate and number of loan applications
- Which of the two variables listed above is the *response* variable?
  - Number of loan applications
- Which of the two variables listed above is the *explanatory* variable?
  - Interest rate

SCATTERPLOTS AND CORRELATION

- Remember the FIRST step in data analysis is to explore the data.
- If both the response AND explanatory variable are quantitative then we can explore the relationship between these two variables using a **scatterplot**.
- A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals.

SCATTERPLOTS

Definition:
a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.



SCATTERPLOTS AND CORRELATION

How to make a Scatterplot

1. Decide which variable should go on each axis.
   - **explanatory variable → x-axis**
   - **response variable → y-axis.**

2. Label and scale your axes.

3. Plot individual data values. Each individual in the data appears as a point on the graph.

LET'S CONTINUE HIKING!

SCATTERPLOTS AND CORRELATION

Example : Making a scatterplot

Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

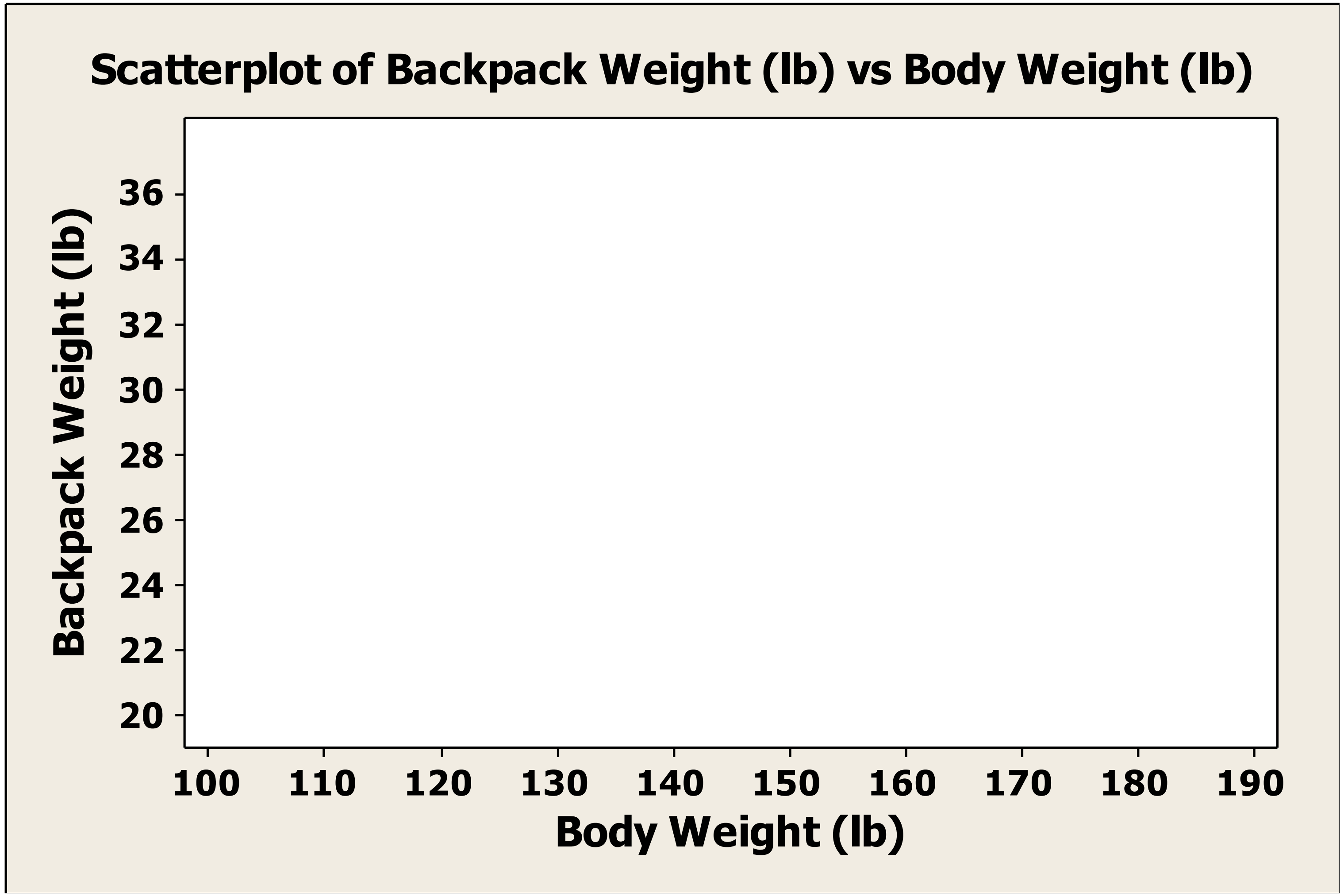
SCATTERPLOTS AND CORRELATION

Example : Making a scatterplot

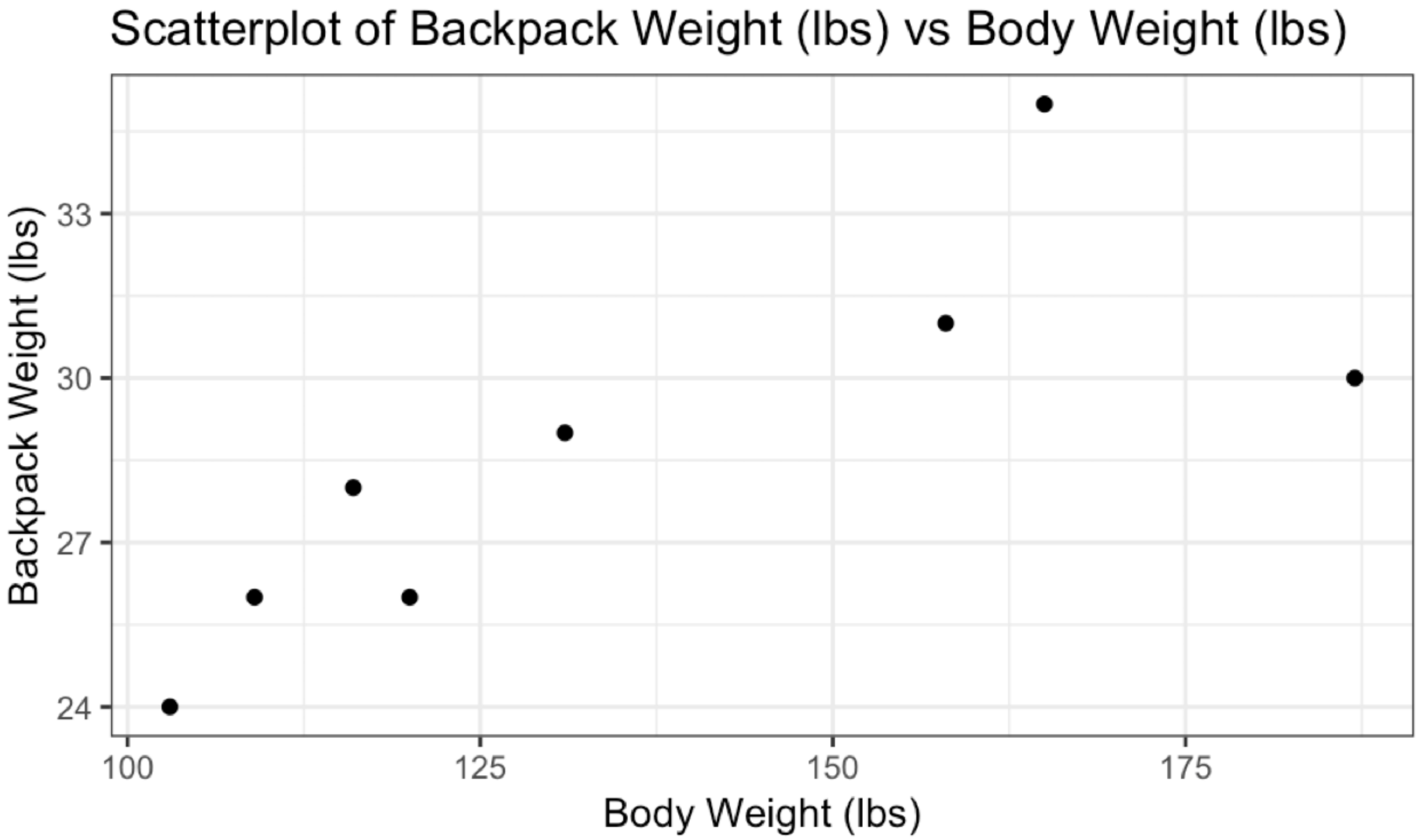
1. Decide which variable should go on each axis.
   - Explanatory variable (X) = **body weight**
   - Response variable (Y) = **backpack weight**
2. Label and scale your axes.
3. Plot individual data values.

Body weight (lb)	**120**	**187**	**109**	**103**	**131**	**165**	**158**	**116**
Backpack weight (lb)	**26**	**30**	**26**	**24**	**29**	**35**	**31**	**28**

SCATTERPLOTS AND CORRELATION



Example 2: Making a scatterplot



WHAT SHOULD I LOOK FOR IN A SCATTERPLOT?

SCATTERPLOTS AND CORRELATION

<u>Examining Scatterplots</u>

- As in any graph of data, look at and describe the *overall pattern* and for striking *departures* from that pattern.
  - You can describe the <u>overall pattern</u> of a scatterplot by the:
    - **<u>direction</u> – positive or negative**
    - **<u>form</u> – linear or non-linear**
    - **<u>strength</u> – strong (points close together) or weak (points spread out)**

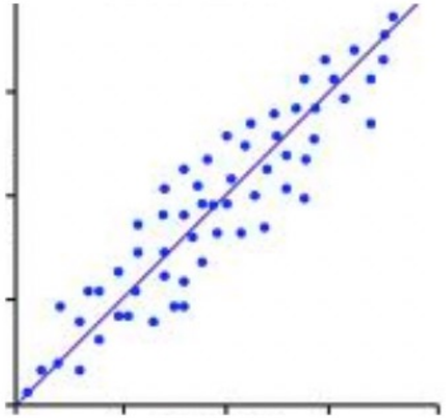
  - An important kind of departure is an outlier, an individual value that falls outside the overall pattern of the relationship

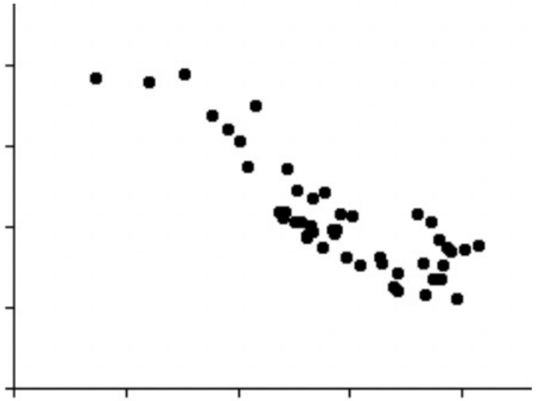
SCATTERPLOTS AND CORRELATION

DESCRIBING DIRECTION:

- **Positive association:** when increases in the explanatory variable are associated with increases in the response variable.

- **Negative (inverse) association:** when increases in the explanatory variable are associated with decreases in the response variable

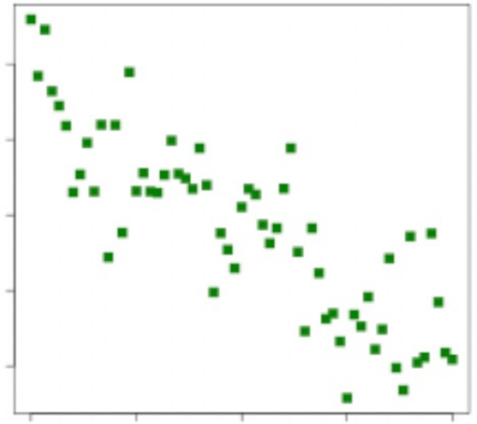




SCATTERPLOTS AND CORRELATION

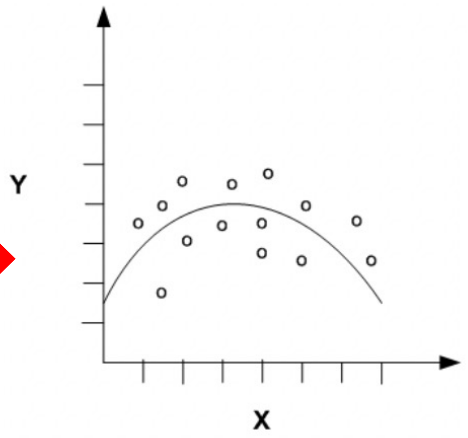
DESCRIBING FORM:

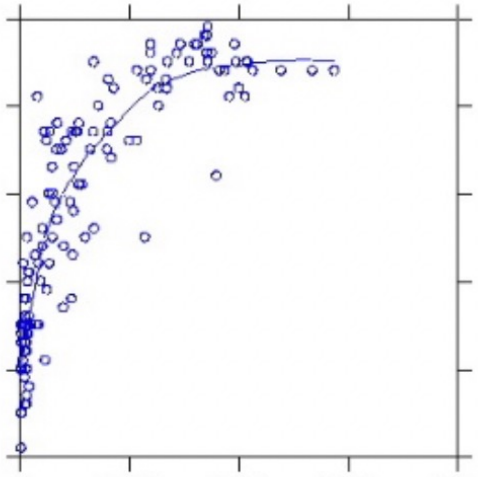
- **Linear**: If the overall trend follow a straight line



- **Non-linear**: If the overall trend has any kind of curvature

quadratic →



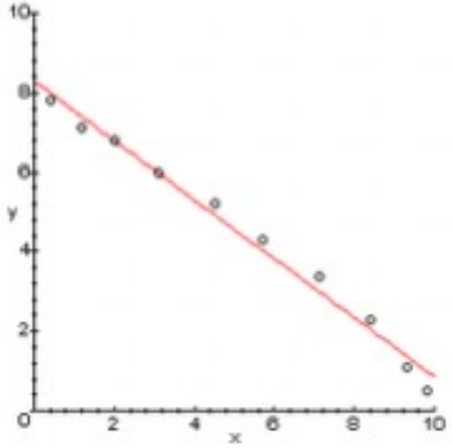


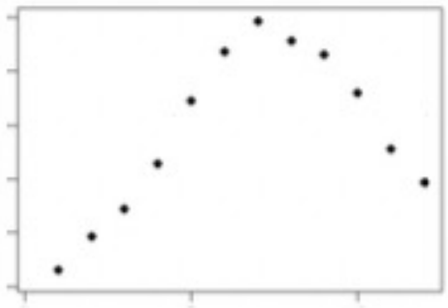
SCATTERPLOTS AND CORRELATION

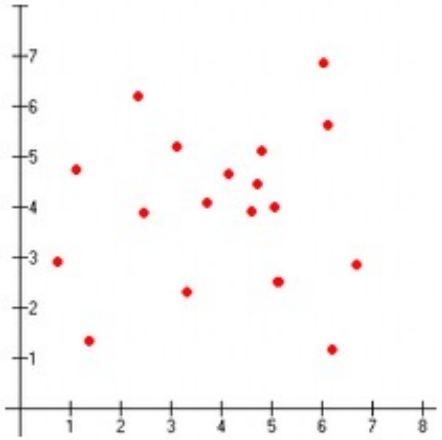
DESCRIBING STRENGTH:

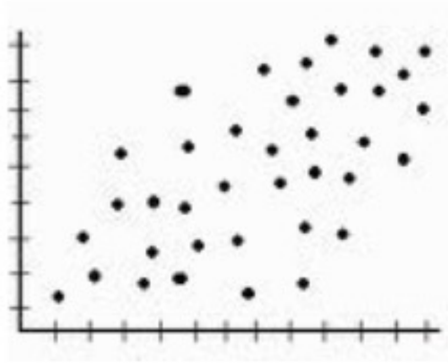
- **Strong:** When all points show a clear trend, with few departures from that trend.

- **Weak:** If the points are spread all over the graph, showing no discernable pattern and high variability









OUTLIERS AND INFLUENTIAL POINTS

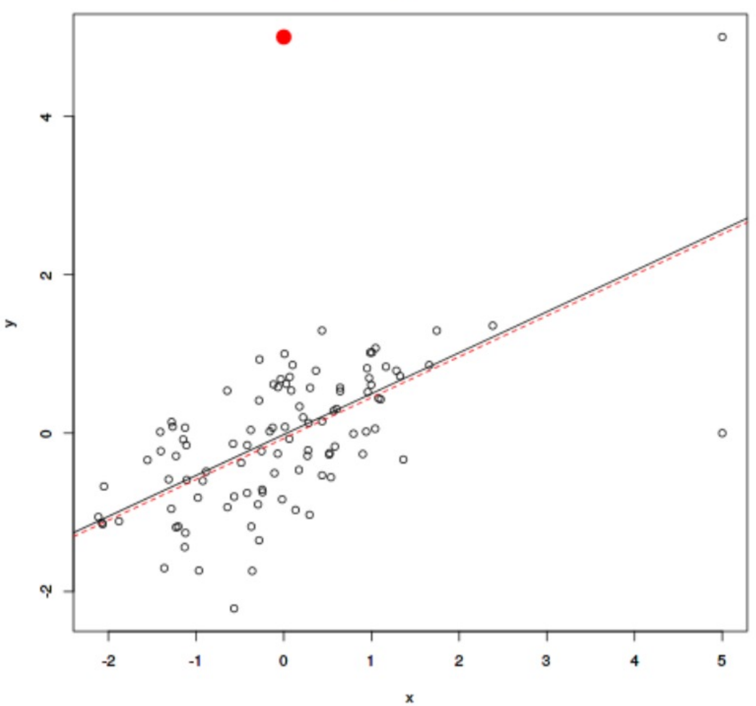
OUTLIERS/INFLUENTIAL POINTS

- **Outliers and Influential Points**
- An ***outlier*** is an observation that lies far away from the other observations.
  - **Outliers in the y direction have large residuals.**
  - **Outliers in the *x* direction are often (but not always) *influential* for the least-squares regression line.**
- If an outlier is ***influential,*** it simply means that the removal of such points would noticeably change the regression equation of the line.

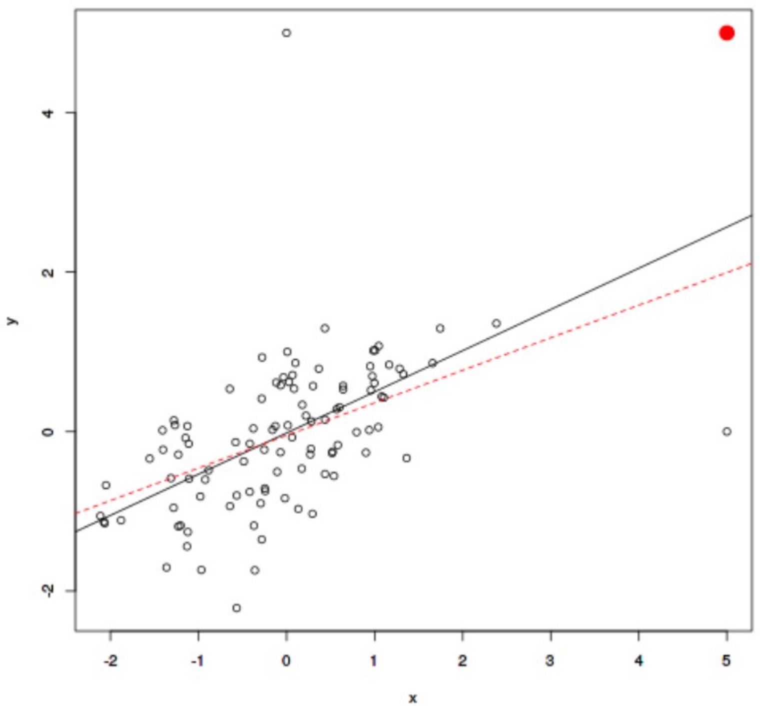
OUTLIERS/INFLUENTIAL POINTS

- Here are three possible scenarios for outliers:

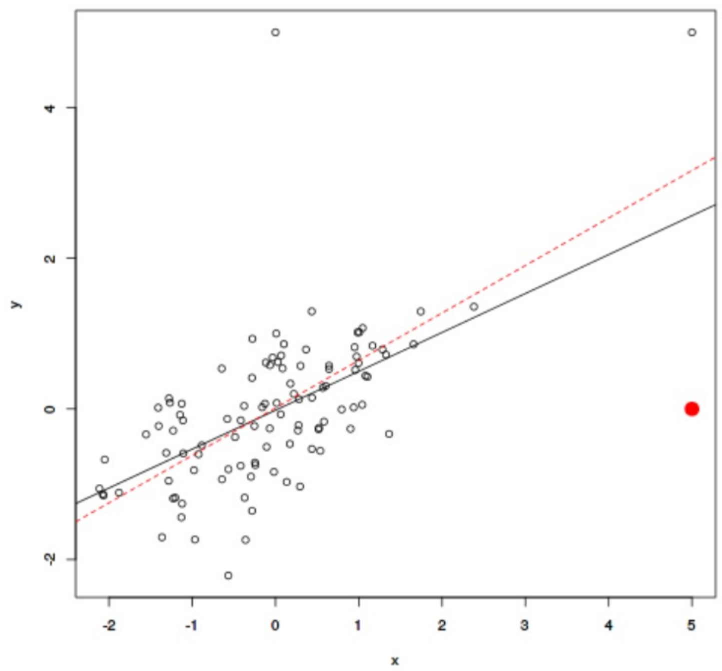
Before removing the red point: r = 0.58, after r = 0.66



Before removing the red point: r = 0.58, after r = 0.47



Before removing the red point: r = 0.58, after r = 0.64



The **black** lines are regression lines using the **entire** data set.

The **red** dotted lines are regression lines with the red, bold **outlier removed**.

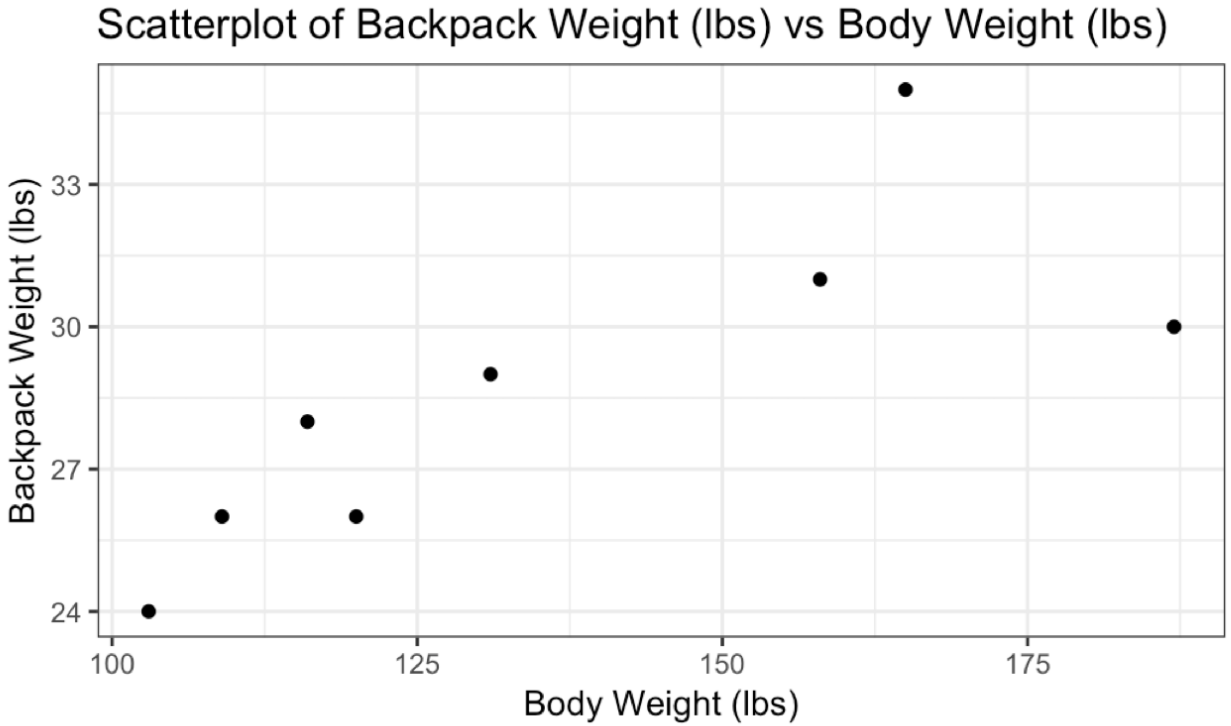
Notice how the regression lines change when each of the outliers are removed.

LET'S TRY IT!

SCATTERPLOTS AND CORRELATION

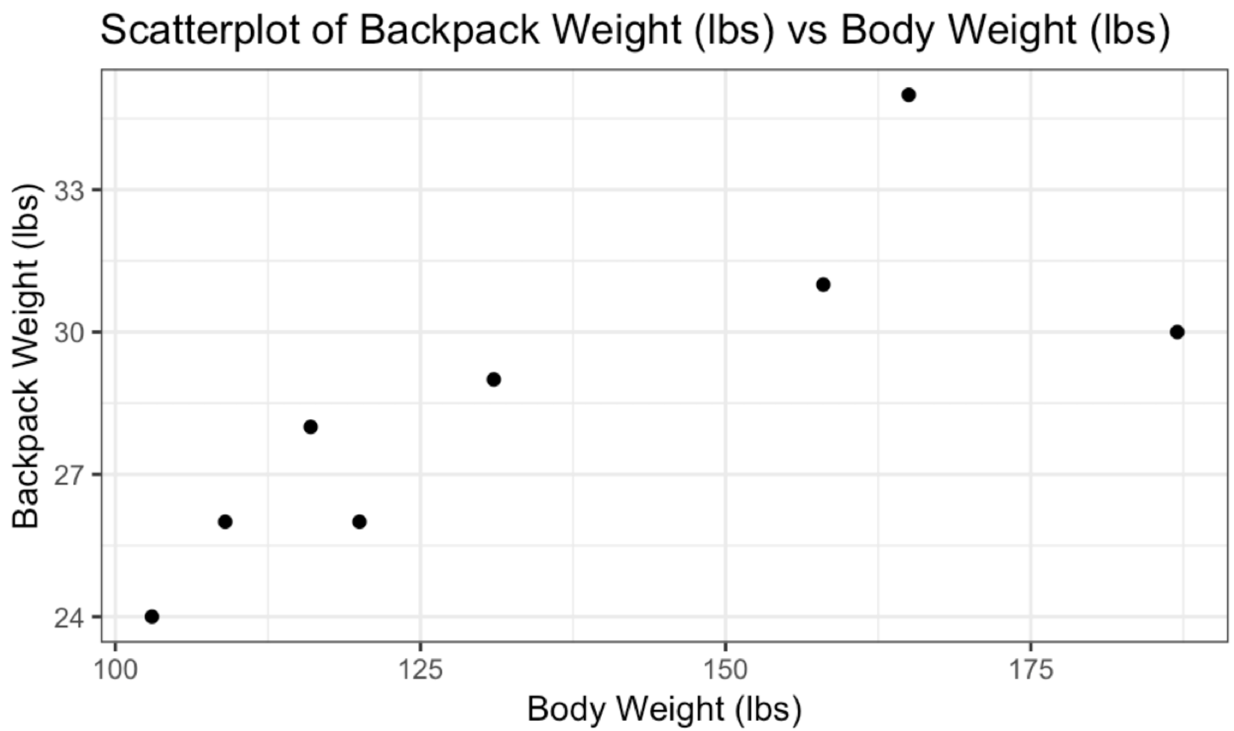
***Example 3*:** Putting it all together

Earlier we constructed a scatterplot of hiker's pack weight against their body weight. How would you describe the following scatterplot?



SCATTERPLOTS AND CORRELATION

SOLUTION: There is a moderately strong, positive, linear relationship between body weight and pack weight with one outlier who has a body weight of 187.



CORRELATION COEFFICIENT

SCATTERPLOTS AND CORRELATION

Measuring Linear Association

- Our eyes are not always a good judge of how strong a linear relationship is.
- Our eyes can be fooled by changing the scale of the scatterplot or the amount of space around each point.
- That is why we calculate *r*. The **correlation *r*** measures the direction and strength of the ***linear*** relationship between two quantitative variables.
- In practice we will use R to calculate r, but here is the formula :

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

SCATTERPLOTS AND CORRELATION

<u>Some notes about correlation:</u>

To ensure units don't matter when measuring the strength, we can remove them by standardizing each variable. Now, for each point, instead of values (x,y) we'll have the standardized coordinates (z_x, z_y). Remember that to standardize values, we subtract the mean of each variable and then divide by its standard deviation:

$$(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$$

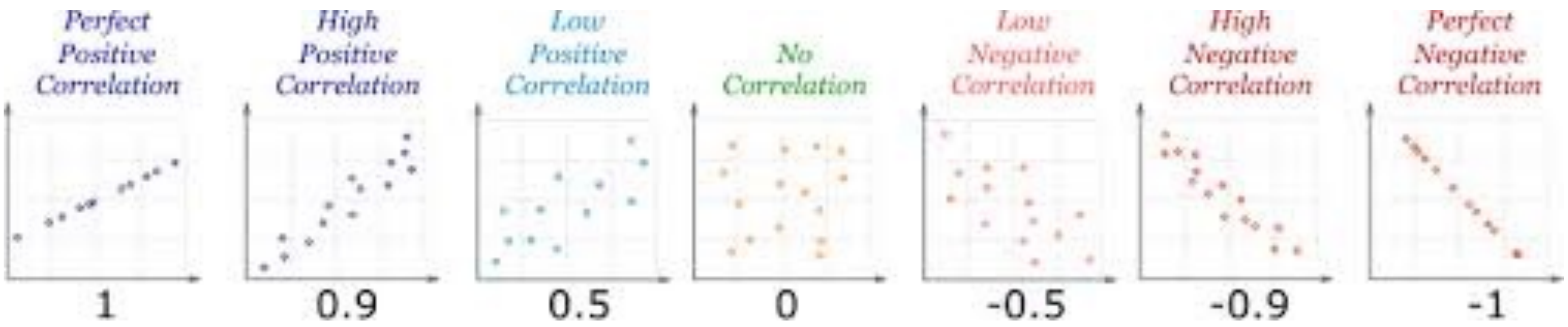
Because standardizing makes the means of both variables 0, the center of the new scatterplot is at the origin. The scales on both axes are now standard deviation units, making the scaling consistent and providing a fairer impression of the strength of the association.

CORRELATION

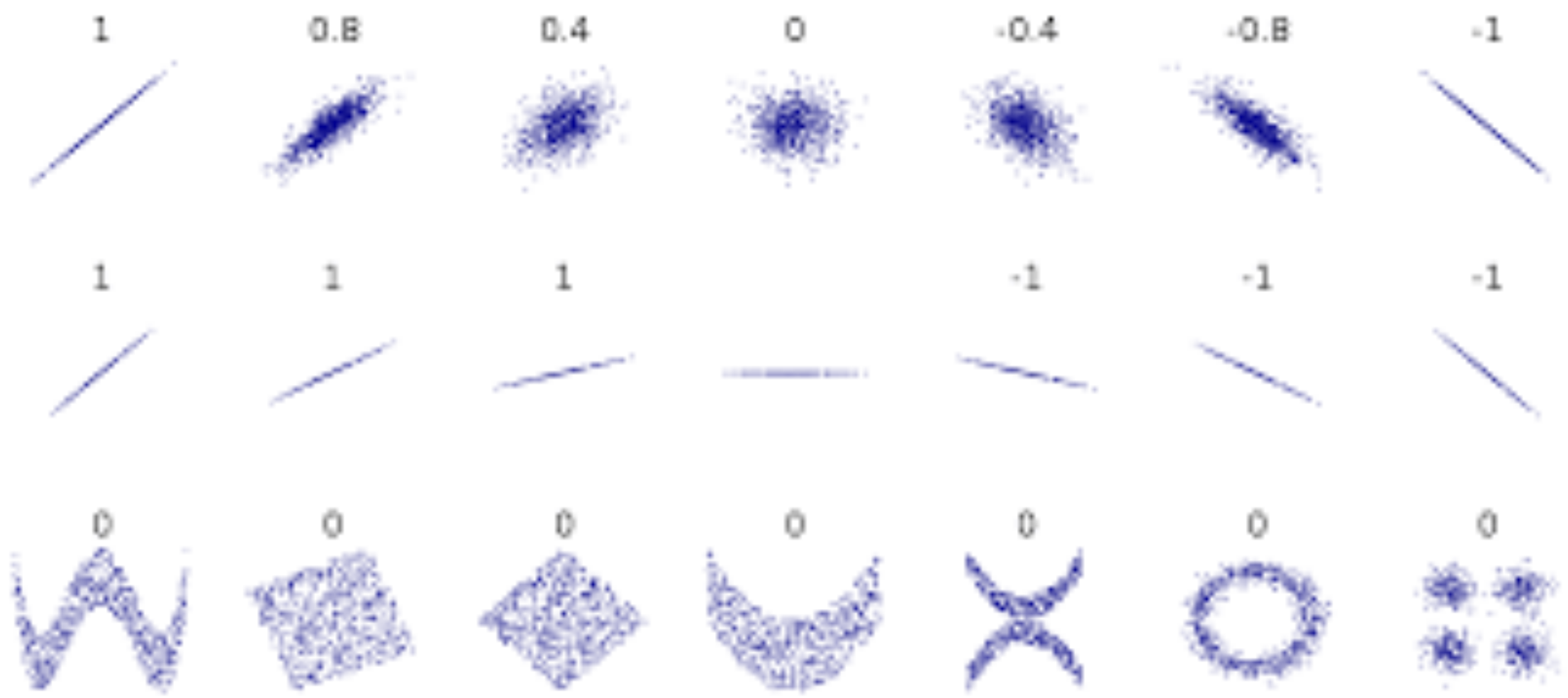
Definition:

a metric for the strength of **<u>linear</u>** relationship between two numeric variables

MORE on this when we cover regression!

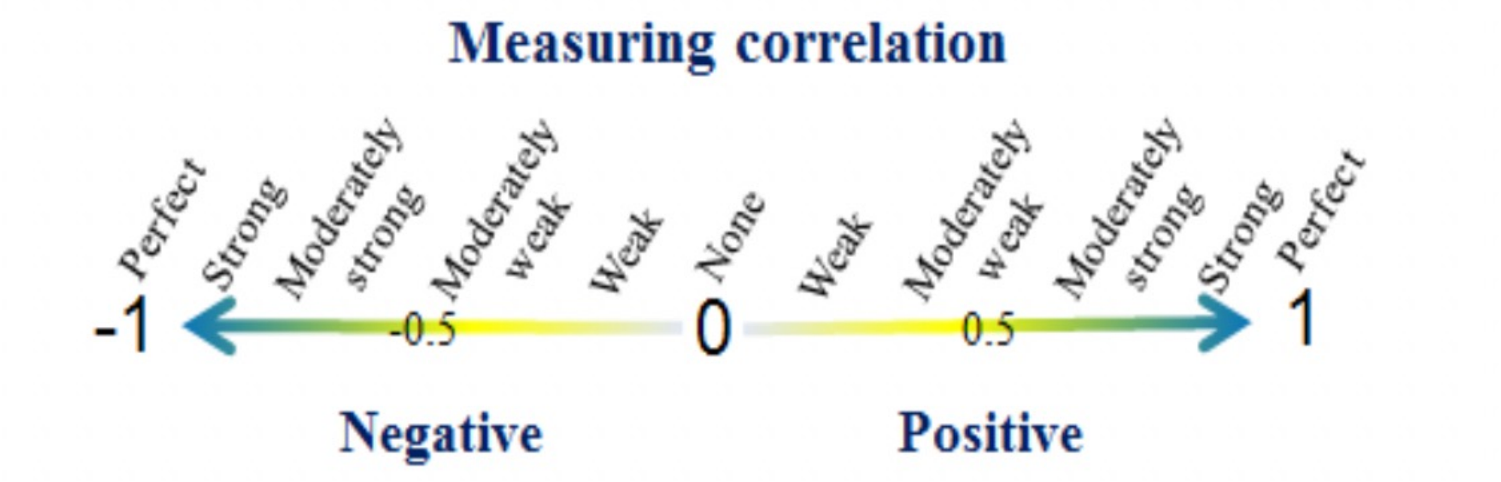


CORRELATION



SCATTERPLOTS AND CORRELATION

Use this continuum to describe the strength of relationship during r (the correlation coefficient):

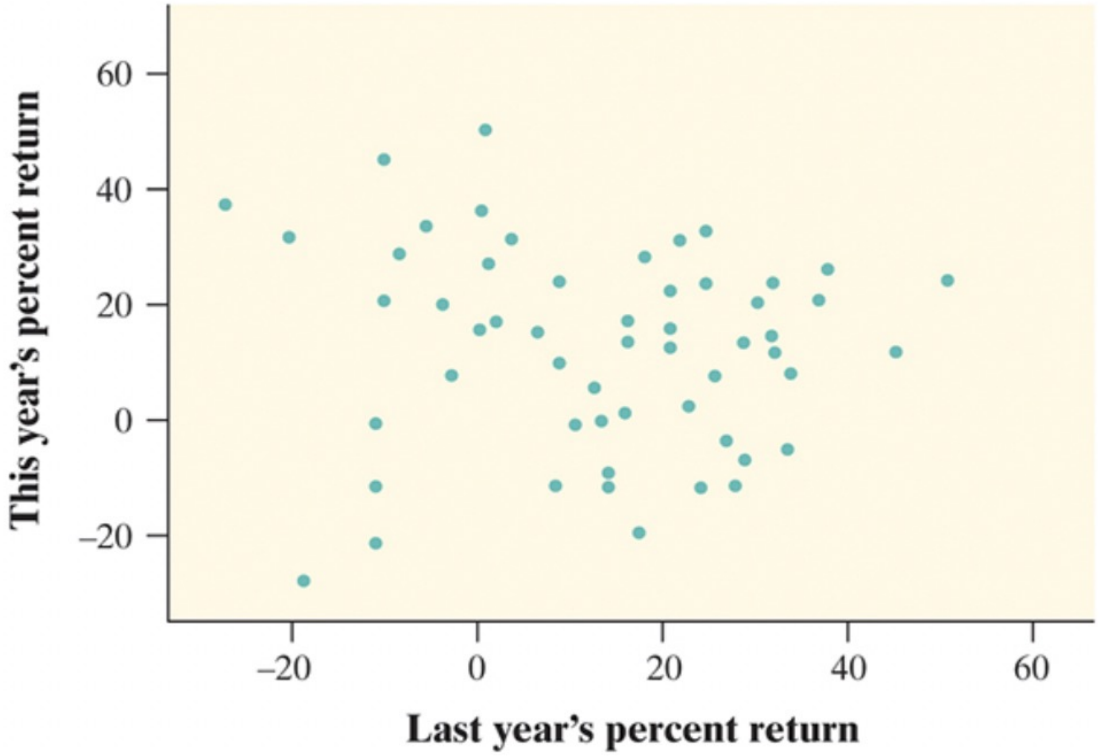


SCATTERPLOTS AND CORRELATION

Example 4:

What is the best estimate for the correlation for this graph?

A. -1.25
B. -1
C. -0.08
D. 0.584
E. 0.95

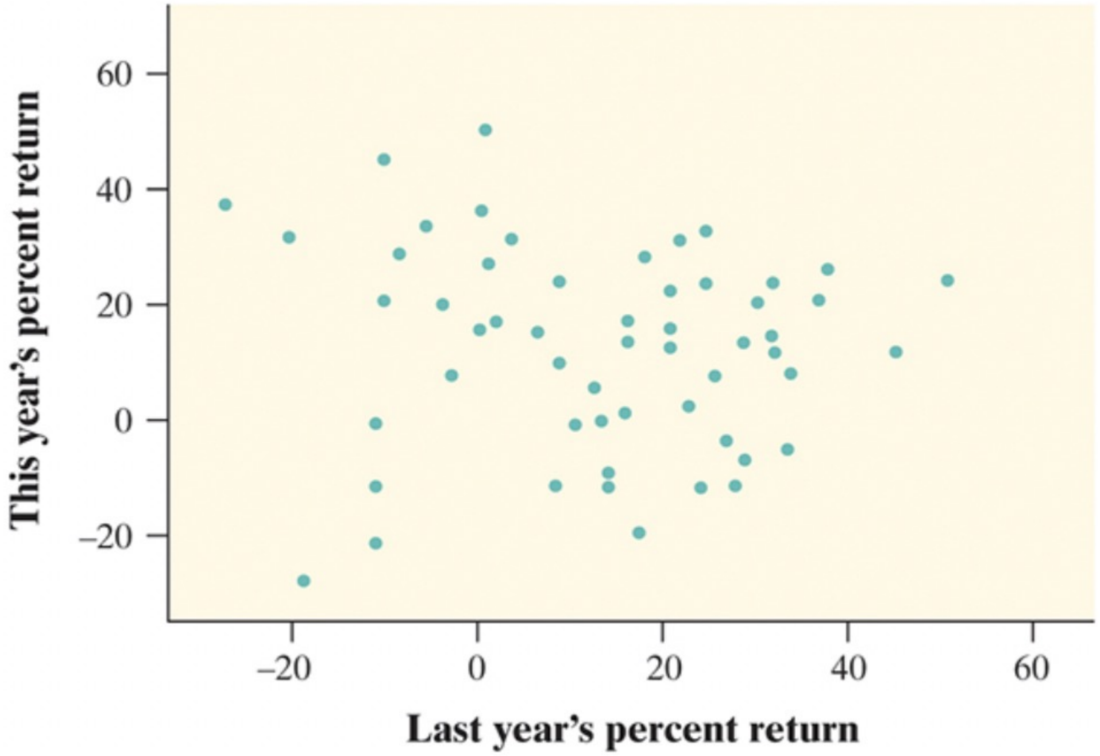


SCATTERPLOTS AND CORRELATION

<u>Example 4</u>:

What is the best estimate for the correlation for this graph?

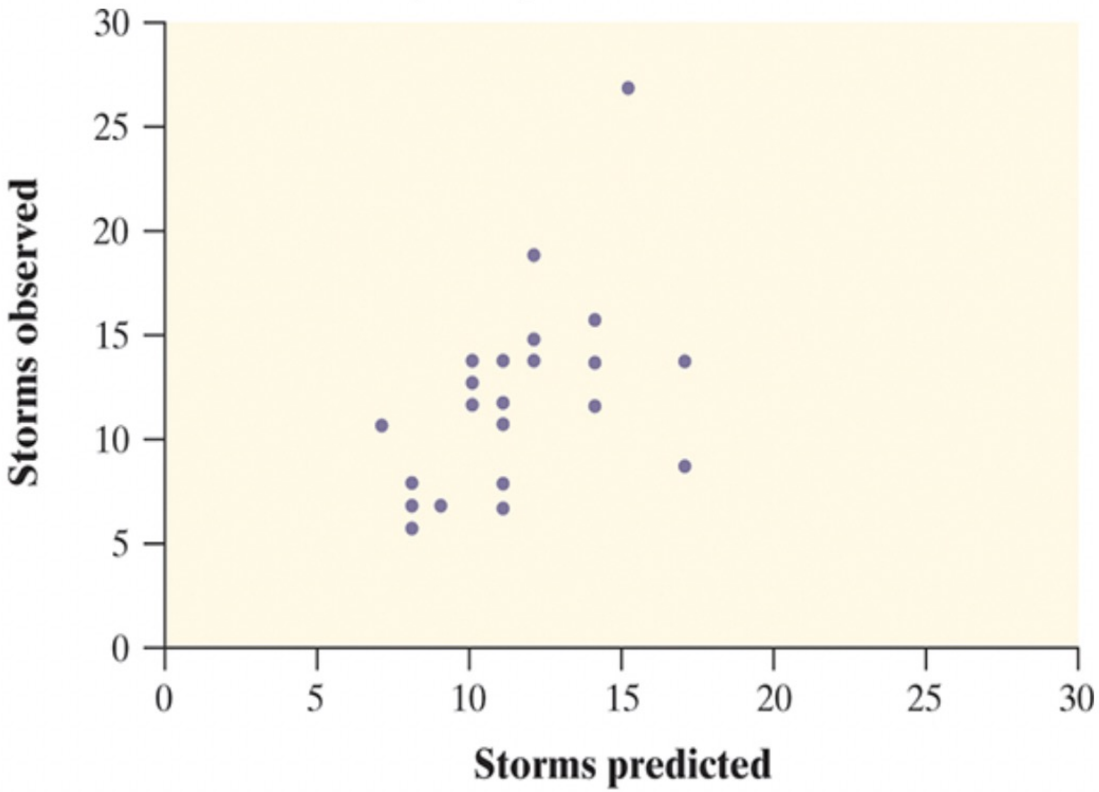
A. -1.25
B. -1
C. -0.08
D. 0.584
E. 0.95



SCATTERPLOTS AND CORRELATION

***Example 5*:**
What is the best estimate for the correlation for this graph?

A. -1
B. -0.9
C. -0.081
D. 0.584
E. 0.95
F. 1



SCATTERPLOTS AND CORRELATION

***Example 5*:**

What is the best estimate for the correlation for this graph?

A. -1

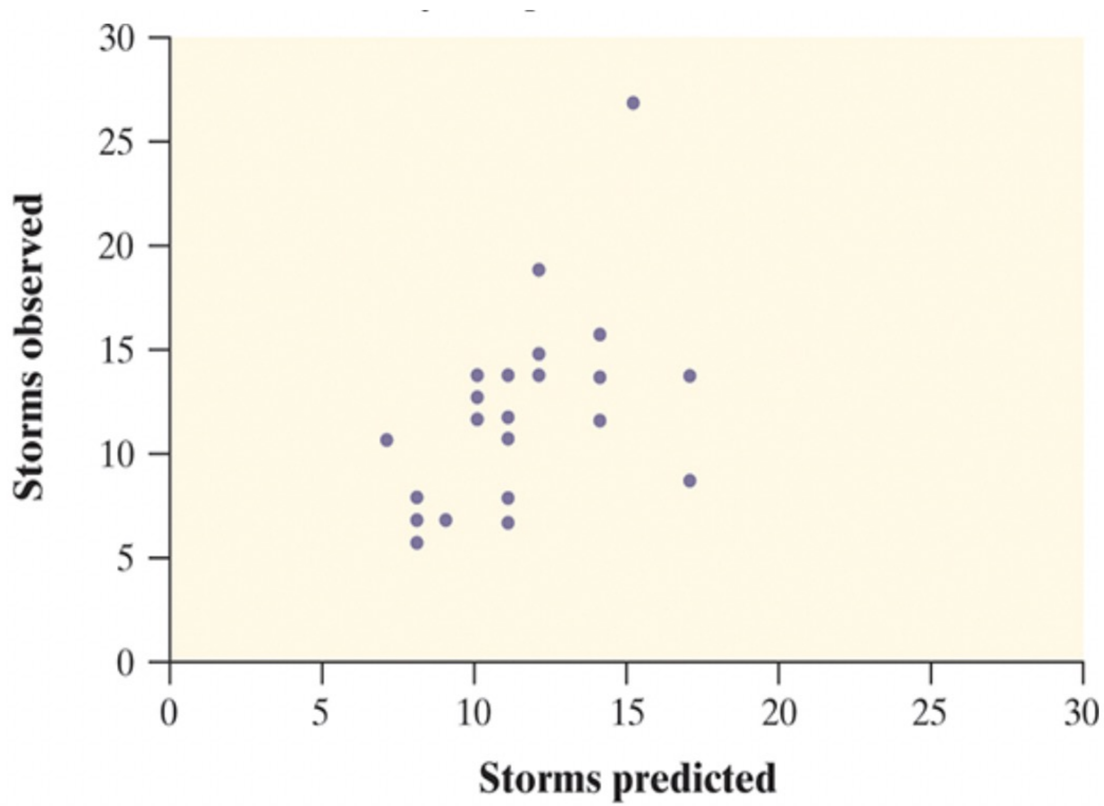
B. -0.9

C. -0.081

D. 0.584

E. 0.95

F. 1



SCATTERPLOTS AND CORRELATION

Fun Facts about the correlation coefficient:

- r is always a number between -1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of r near 0 indicate a very weak linear relationship.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect ***linear*** relationship.
- Correlation makes no distinction between explanatory and response variables.
    - Meaning if you switch around the explanatory and the response variable (switch the x and y), you will get the same value for the correlation
- r has no units and does not change when we change the units of measurement of x, y, or both.

ACTIVITY

STEP 5: Activity

First: Load in the data

```
data("anscombe")
str(anscombe)
```

```
## 'data.frame':    11 obs. of  8 variables:
##  $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ x4: num  8 8 8 8 8 8 8 19 8 8 ...
##  $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
##  $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
##  $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
##  $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
```

Directions:

If your birthday is:

- January - March: Use variables `x1` and `y1`
- April - June: Use variables `x2` and `y2`
- July - September: Use variables `x3` and `y3`
- October - December: Use variables `x4` and `y4`

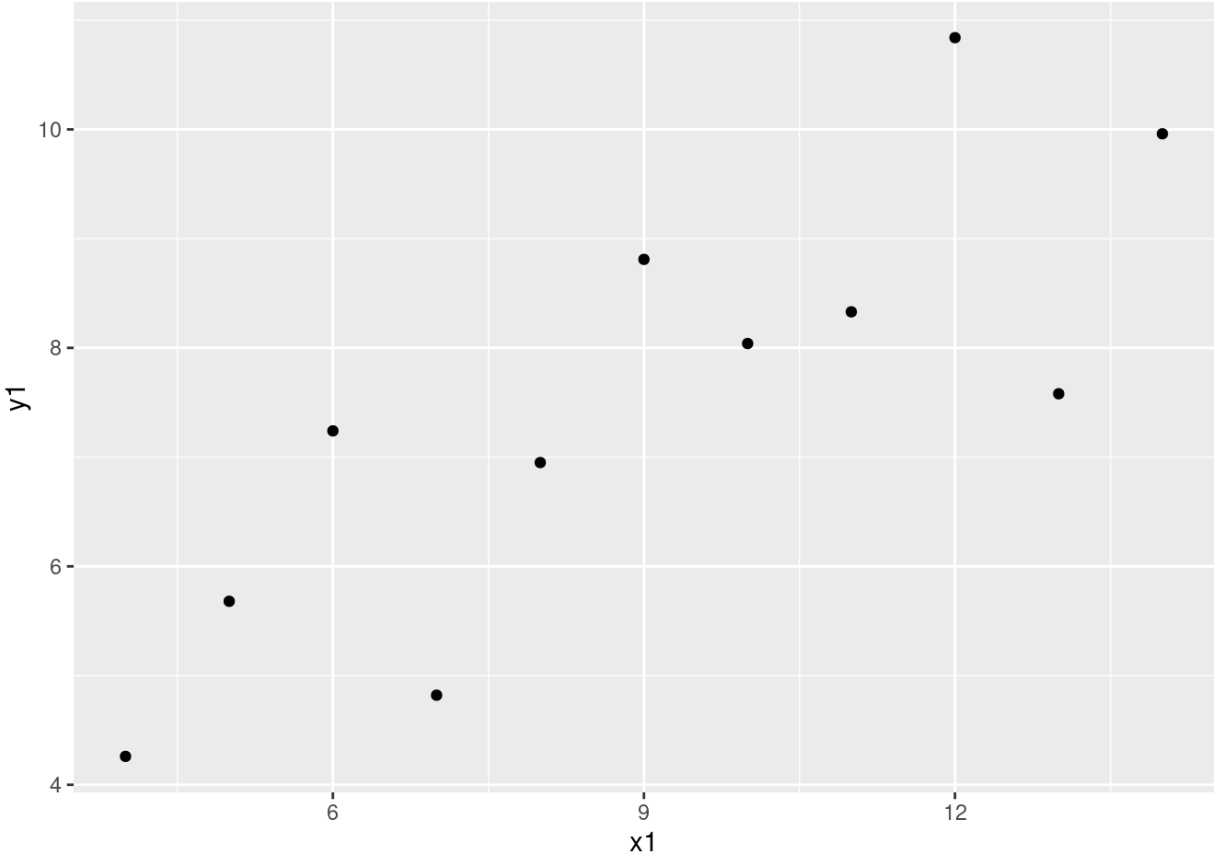
Complete the following tasks:

- Create a scatterplot and describe it
- Calculate the mean and standard deviation for both your `x` and `y` variables
- Calculate the correlation coefficient
- Compare the information you have obtained with your neighbor

SOLUTIONS

GROUP 1

```
## SPACE FOR YOUR WORK ##
ggplot(anscombe, aes(x1, y1))+
  geom_point()
```




```
mean(anscombe$x1)
```

```
## [1] 9
```

```
sd(anscombe$x1)
```

```
## [1] 3.316625
```

```
mean(anscombe$y1)
```

```
## [1] 7.500909
```

```
sd(anscombe$y1)
```

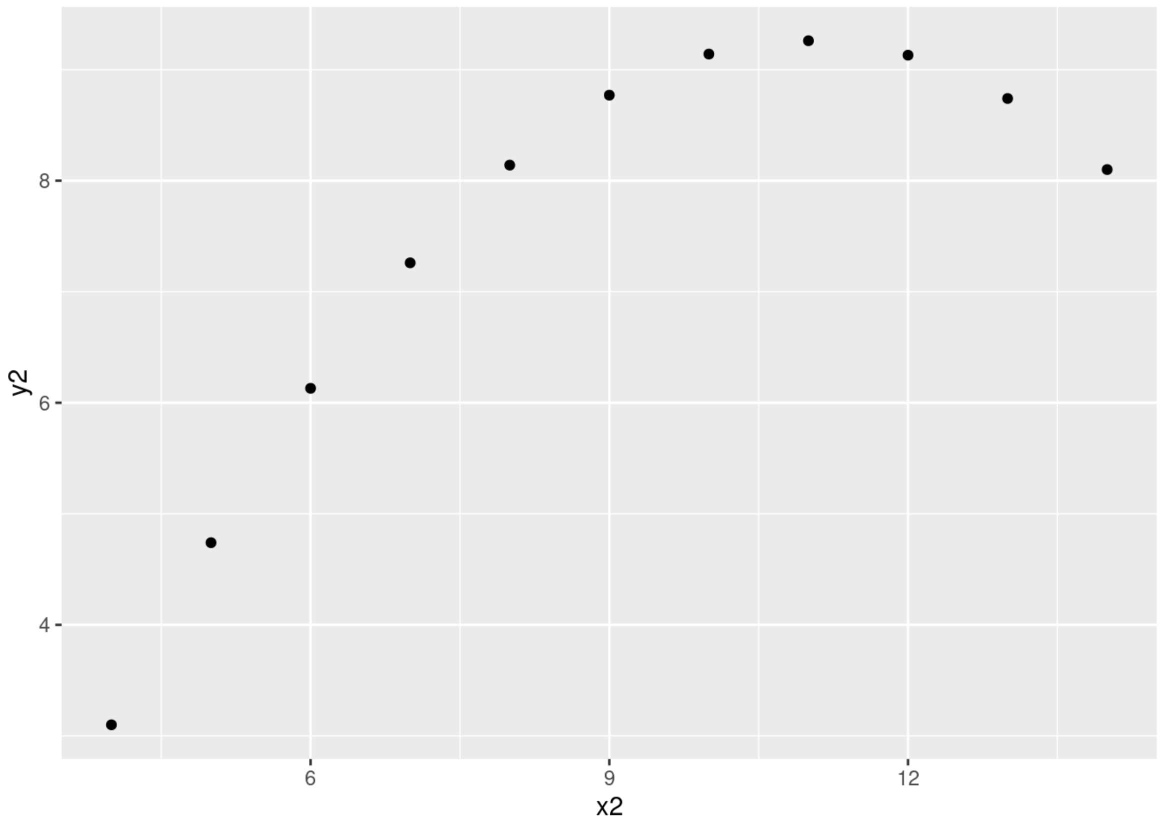
```
## [1] 2.031568
```

```
cor(anscombe$x1, anscombe$y1)
```

```
## [1] 0.8164205
```

GROUP 2

```
## SPACE FOR YOUR WORK ##
ggplot(anscombe, aes(x2, y2))+
  geom_point()
```




```
mean(anscombe$x2)
```

```
## [1] 9
```

```
sd(anscombe$x2)
```

```
## [1] 3.316625
```

```
mean(anscombe$y2)
```

```
## [1] 7.500909
```

```
sd(anscombe$y2)
```

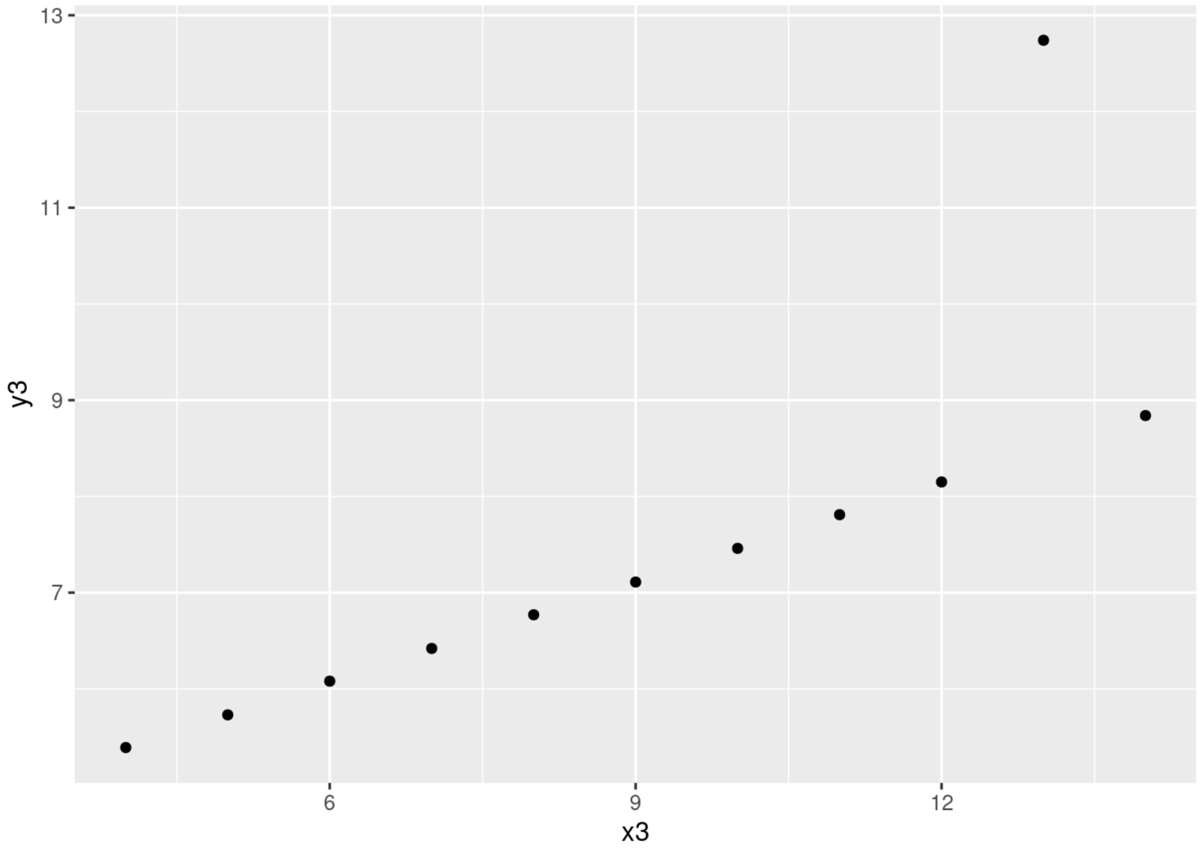
```
## [1] 2.031657
```

```
cor(anscombe$x2, anscombe$y2)
```

```
## [1] 0.8162365
```

GROUP 3

```r
## SPACE FOR YOUR WORK ##
ggplot(anscombe, aes(x3, y3))+
  geom_point()
```




```r
mean(anscombe$x3)
```

```
## [1] 9
```

```r
sd(anscombe$x3)
```

```
## [1] 3.316625
```

```r
mean(anscombe$y3)
```

```
## [1] 7.5
```

```r
sd(anscombe$y3)
```

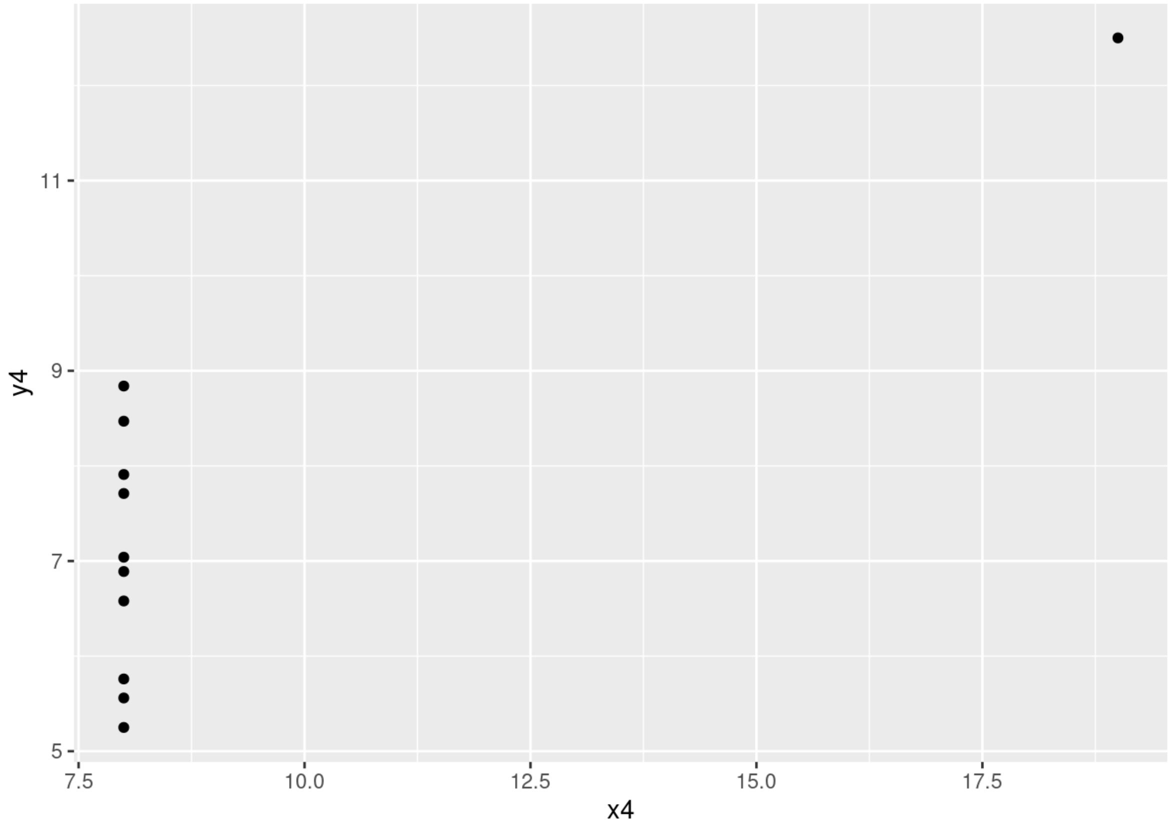
```
## [1] 2.030424
```

```r
cor(anscombe$x3, anscombe$y3)
```

```
## [1] 0.8162867
```

GROUP 4

```
## SPACE FOR YOUR WORK ##
ggplot(anscombe, aes(x4, y4))+
  geom_point()
```




```
mean(anscombe$x4)
```

```
## [1] 9
```

```
sd(anscombe$x4)
```

```
## [1] 3.316625
```

```
mean(anscombe$y4)
```

```
## [1] 7.500909
```

```
sd(anscombe$y4)
```

```
## [1] 2.030579
```

```
cor(anscombe$x4, anscombe$y4)
```

```
## [1] 0.8165214
```

COMPARE

```
mean(anscombe$x1)
```

```
## [1] 9
```

```
sd(anscombe$x1)
```

```
## [1] 3.316625
```

```
mean(anscombe$y1)
```

```
## [1] 7.500909
```

```
sd(anscombe$y1)
```

```
## [1] 2.031568
```

```
cor(anscombe$x1, anscombe$y1)
```

```
## [1] 0.8164205
```

```
mean(anscombe$x2)
```

```
## [1] 9
```

```
sd(anscombe$x2)
```

```
## [1] 3.316625
```

```
mean(anscombe$y2)
```

```
## [1] 7.500909
```

```
sd(anscombe$y2)
```

```
## [1] 2.031657
```

```
cor(anscombe$x2, anscombe$y2)
```

```
## [1] 0.8162365
```

```
mean(anscombe$x4)
```

```
## [1] 9
```

```
sd(anscombe$x4)
```

```
## [1] 3.316625
```

```
mean(anscombe$y4)
```

```
## [1] 7.500909
```

```
sd(anscombe$y4)
```

```
## [1] 2.030579
```

```
cor(anscombe$x4, anscombe$y4)
```

```
## [1] 0.8165214
```

```
mean(anscombe$x4)
```

```
## [1] 9
```

```
sd(anscombe$x4)
```

```
## [1] 3.316625
```

```
mean(anscombe$y4)
```

```
## [1] 7.500909
```

```
sd(anscombe$y4)
```

```
## [1] 2.030579
```

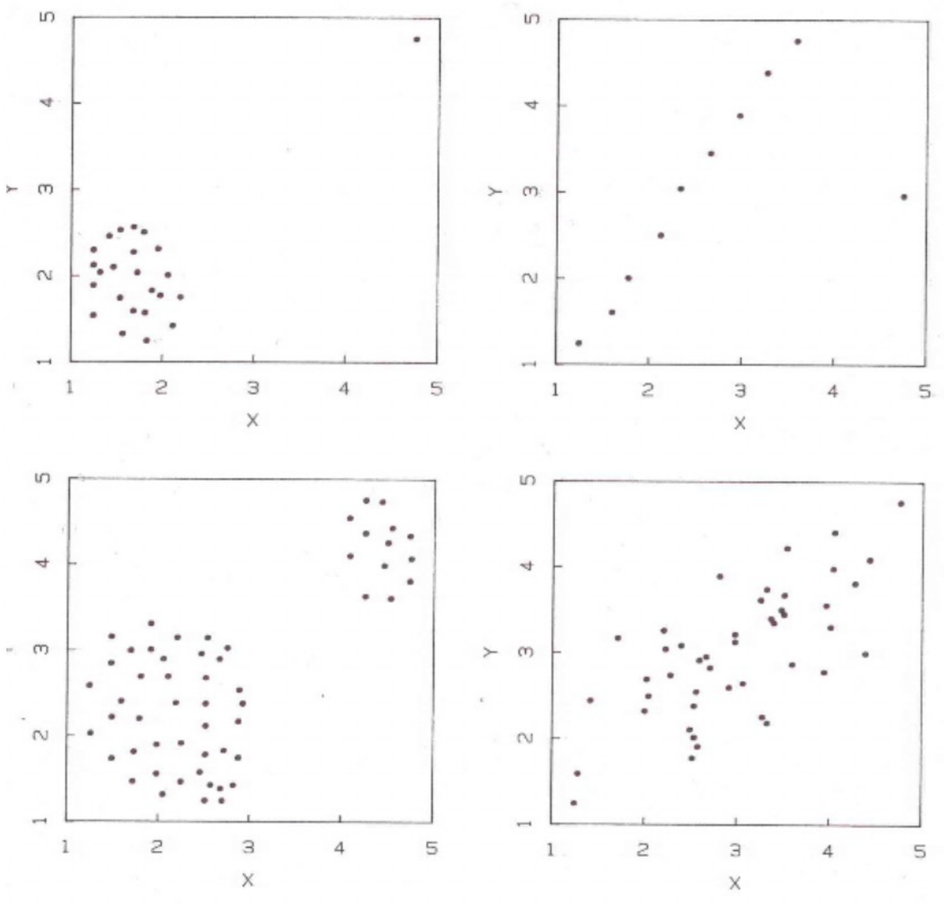
```
cor(anscombe$x4, anscombe$y4)
```

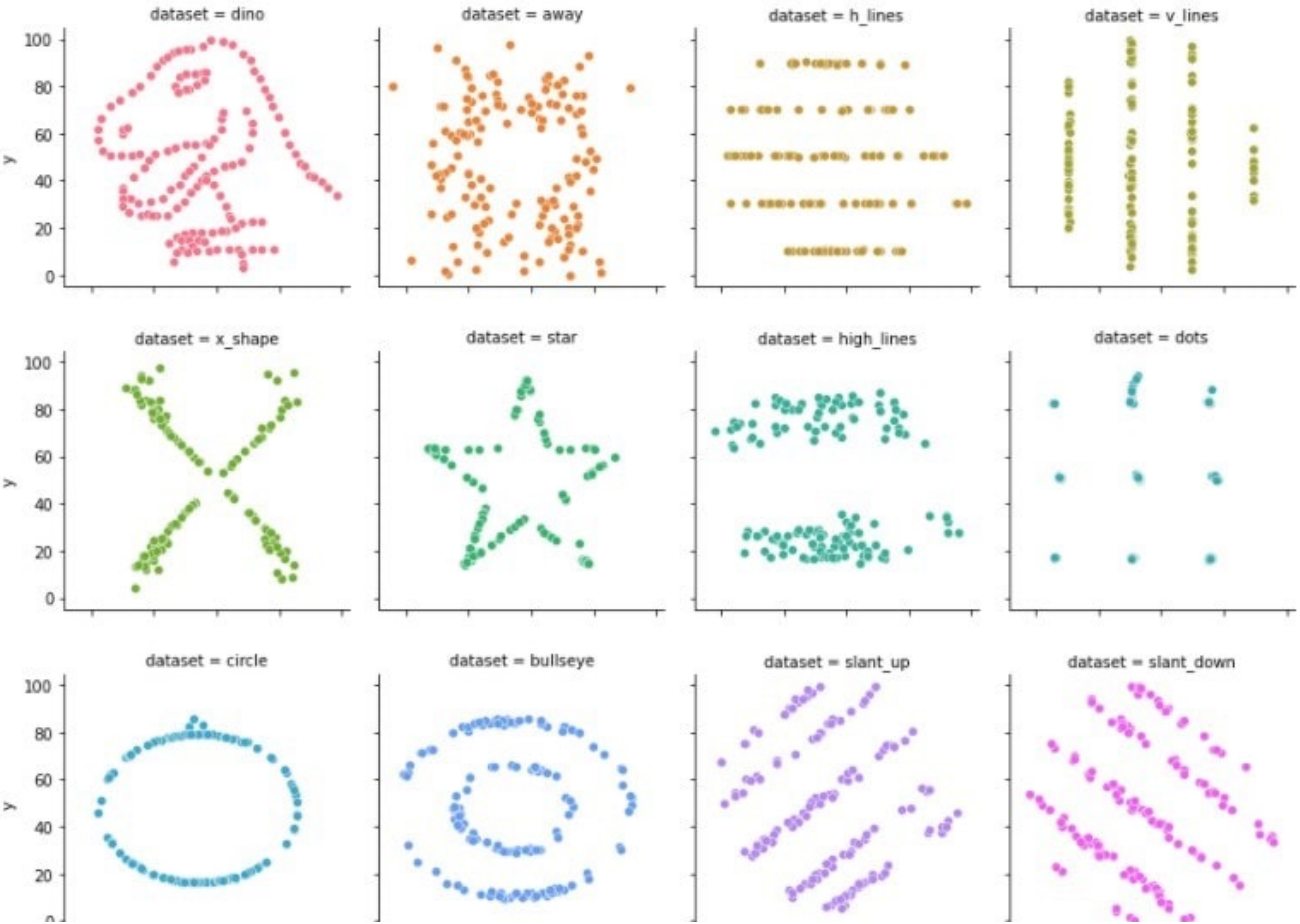
```
## [1] 0.8165214
```

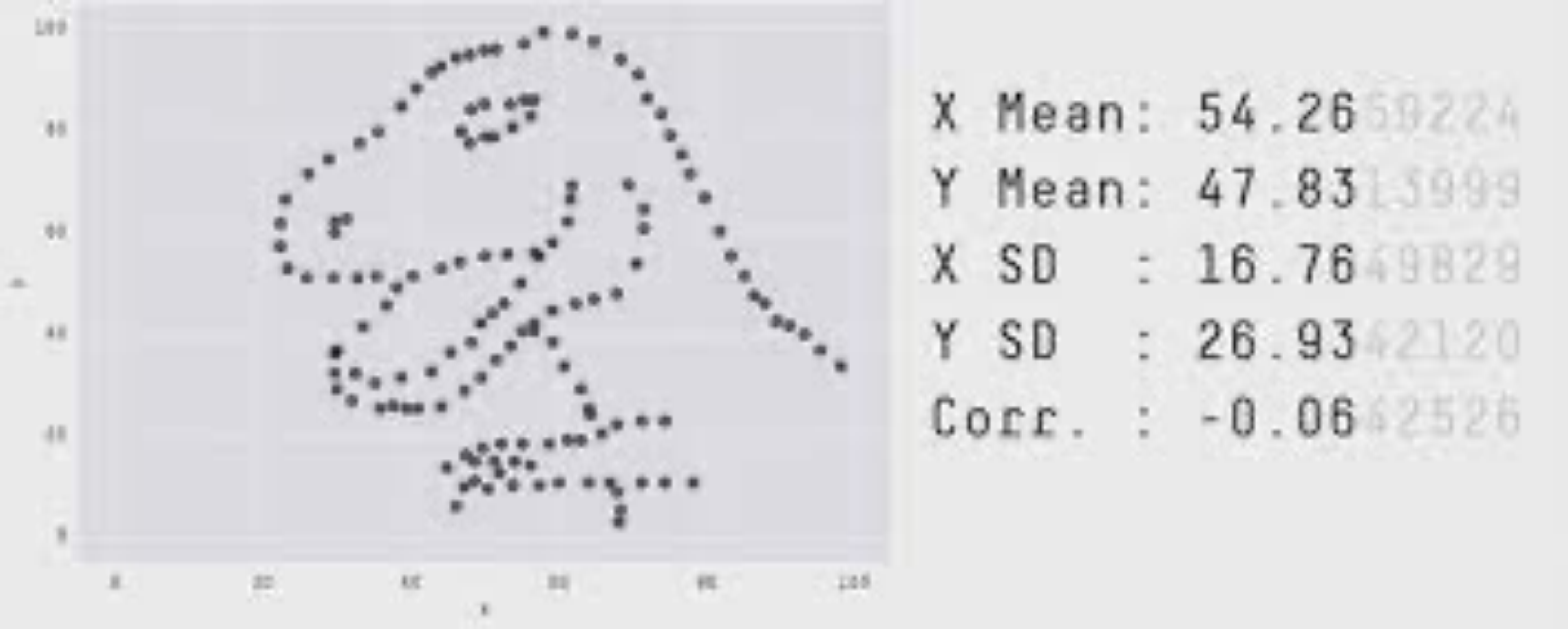
SCATTERPLOTS AND CORRELATION

Cautions about correlation:

- Correlation requires that both variables be quantitative.
- Correlation DOES NOT describe curved relationships between variables, no matter how strong the relationship is.
- Correlation is not resistant to outliers. *r* is strongly affected by a few outlying observations.
- Correlation is not a complete summary of two-variable data. Consider these 4 graphs. All four graphs have r = 0.7, but you can see all four graphs look very different.

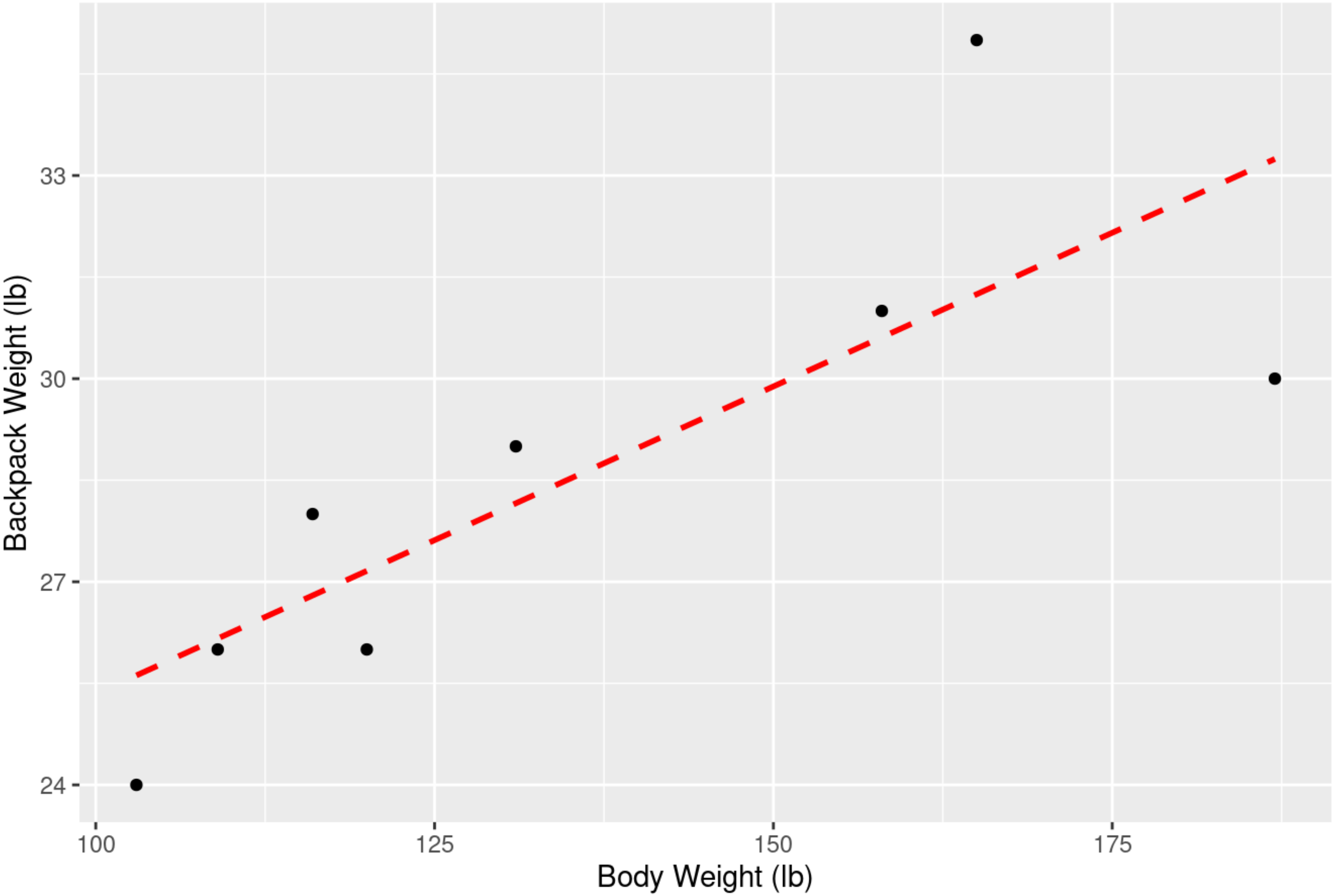



dataset = dino
dataset = away
dataset = h_lines
dataset = v_lines
dataset = x_shape
dataset = star
dataset = high_lines
dataset = dots
dataset = circle
dataset = bullseye
dataset = slant_up
dataset = slant_down
y
0
20
40
60
80
100


X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

LINE OF BEST FIT

Scatterplot of Backpack Weight vs Body Weight



LINEAR REGRESSION

<u>Goals of Simple Linear Regression</u>

1. Describe a relationship with a mathematical **model.**
   - This model is called the Least-squares regression equation (LSRE).
   - $\hat{y} = \boldsymbol{b}_0 + \boldsymbol{b}_1 x$

2. ***Predict or estimate*** a mean response value with a mathematical model.
   - We will use the LSRE to do this prediction

3. From a sample, estimate the change that one variable explains in another in a population.
   - Hypothesis testing and confidence intervals

LINEAR REGRESSION

Least-Squares Regression Line:

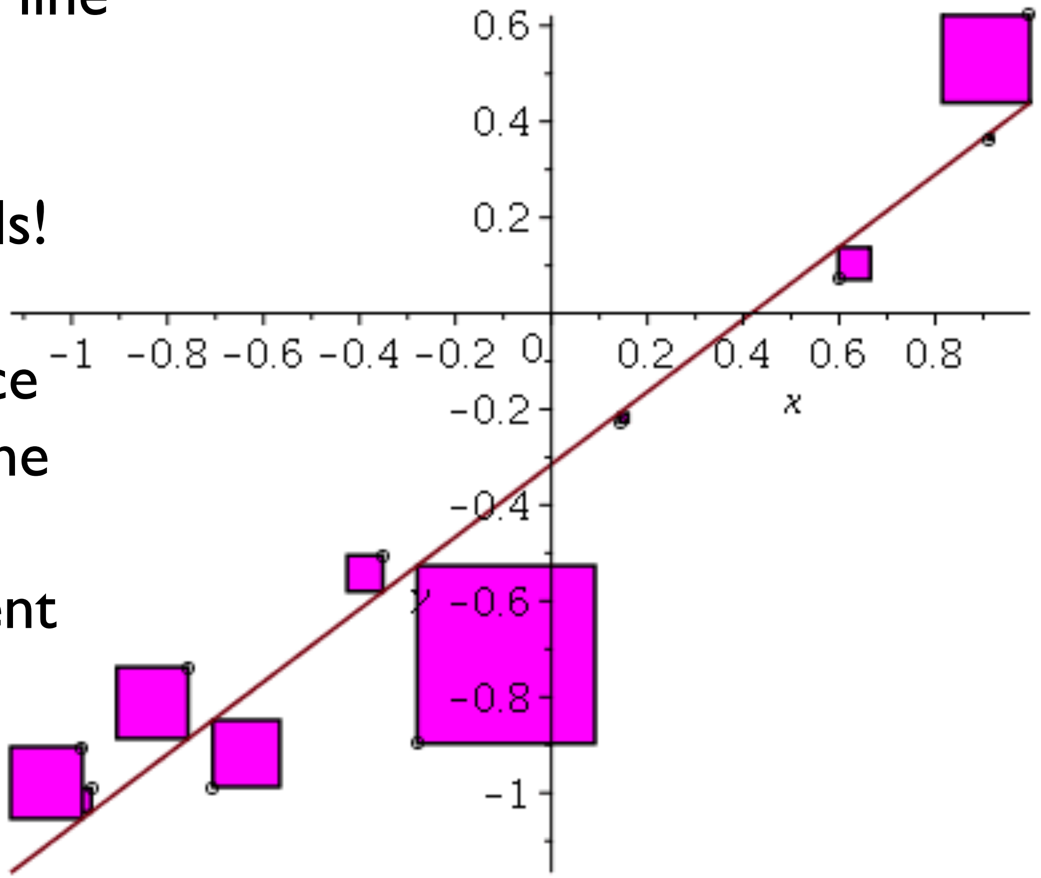
- The estimated simple regression line is $\hat{y} = \boldsymbol{b_0} + \boldsymbol{b_1}x$
- For a given data set there are many regression lines that *could* be drawn to fit the data. If we are going to make predictions from this regression line, we want the **BEST** regression line we can draw! What I mean is that we want a regression line that fits the data best.
- The straight line that ***minimizes*** the sum of the squares of the vertical distances of the data points from the line.
- This line is calculated in such a way that the distance from the point to the line is minimized for all points in the data set.

LINEAR REGRESSION

The least squares regression line finds the line that:

minimizes the sum of the squared residuals!

- Residuals are computed as the difference between the observed data point and the predicted value, given by the line.
- In this graphic, the pink squares represent the squared residuals.



LINEAR REGRESSION

The Equation of the Regression Line

The equation of a line you may have seen in a previous math class is the same equation we will use this this class, with minor differences in notation:

What you saw in previous math classes:

- **y = mx + b**
  - **m = slope**
  - **b = y-intercept**

The Equation of the Regression Line

What you will see in statistics classes:

$$\hat{y} = \boldsymbol{b}_0 + \boldsymbol{b}_1 x$$

- $\hat{y}$ = ***"y-hat"*** is the predicted value of the response variable for a given value of *x*.
- $\boldsymbol{b}_0$ is the **intercept,** the value of *y* when *x* = *0*.
- $\boldsymbol{b}_1$ is the **slope,** the amount by which *y* changes for each one-unit increase in *x*.
  - *I will ask you to interpret the slope of a regression equation in the context of the problem.*
  - *It should sound something like, "For every one-unit increase in x, the mean of our response variable (y) is PREDICTED to increase/decrease by the value of the slope."*
- x is the value of the explanatory variable.

LINEAR REGRESSION

Equation of the LSRL: Calculating by hand

Below are the formulas you may use to find the slope and intercept of a regression equation.

$$slope = b_1 = r \times \frac{s_y}{s_x}$$

$$intercept = b_0 = \bar{y} - b_1 \bar{x}$$

where s_x and s_y are the standard deviations of the two variables, and *r* is their correlation.

LINEAR REGRESSION

<u>**Facts about Least-Squares Regression and Correlation**</u>

The distinction between explanatory and response variables is essential.

The slope b_1 and correlation ***r*** always have the same sign.

Both regression and correlation describe <u>linear</u> relationships.

It does not make sense to use correlation coefficient, ***r***, to describe the strength of the relationship between X and Y when the scatterplot <u>does not</u> show a LINEAR trend!

Both LSRL and correlation ***r*** are influenced by outliers.

REMEMBER - Always plot the data before interpreting!

STATISTICAL OUTPUT

READING STATISTICAL SOFTWARE OUTPUT

Below is a general table regression output. Statistical software will provide you with a table with information on the slope and intercept of the regression equation.

	Estimate	***Std. Error***	***T- value***	***P-value***
Intercept	$\hat{\beta}_0$	$SE_{\hat{\beta}_0}$	$\frac{\hat{\beta}_0}{SE_{\hat{\beta}_0}}$	P-value for t-test on intercept
Explanatory Variable	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	$\frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$	P-value for test on slope

LINEAR REGRESSION

```
Call:
lm(formula = backpack_wgt ~ body_wgt, data = backpack_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2444 -1.2750  0.1133  0.9308  3.7532

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.26493    3.93692   4.131  0.00614 **
body_wgt     0.09080    0.02831   3.207  0.01844 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LINEAR REGRESSION

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	16.26493	3.93692	4.131	0.00614	**
body_wgt	0.09080	0.02831	3.207	0.01844	*

- What is the value of the slope?
- What is the value of the intercept?
- What is the regression equation?

LINEAR REGRESSION

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.26493    3.93692   4.131  0.00614 **
body_wgt     0.09080    0.02831   3.207  0.01844 *
```

- What is the value of the slope?
  0.09
- What is the value of the intercept?
  16.26
- What is the regression equation?

$\hat{y}$ **= 16.26 + 0.91x**

SOLUTIONS

INTERPRETING THE SLOPE

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.26493    3.93692   4.131  0.00614 **
body_wgt     0.09080    0.02831   3.207  0.01844 *
```

A backpacker will tend to carry 0.09 pounds more *on average*, for every pound that they weight.

WHAT DOES THE INTERCEPT MEAN?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.26493    3.93692   4.131  0.00614 **
body_wgt     0.09080    0.02831   3.207  0.01844 *
```

A backpacker who weights zero pounds will carry a 16.26 pound backpack???

ANOTHER EXAMPLE

LINEAR REGRESSION

Example : Nutrition

In one study, the sodium was used to predict the number of calories for a number of food items. Use this information to write the least squares regression equation.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**103.7587**	18.8678	5.499	0.000
sodium	**0.1366**	0.0810	1.686	0.1028

LINEAR REGRESSION

Example : Nutrition

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**103.7587**	18.8678	5.499	0.000
sodium	**0.1366**	0.0810	1.686	0.1028

- What is the value of the slope?
  - **0.1366**
- What is the value of the intercept?
  - **103.7587**
- What is the regression equation?
  - $\hat{y}$ **= 103.7587 + 0.1366x**