The Basic Practice of Statistics
Ninth Edition

David S. Moore William I. Notz

Chapter 4
Scatterplots and Correlation

Lecture Slides

In Chapter 4, we cover …

- Explanatory and response variables
- Displaying relationships: Scatterplots
- Interpreting scatterplots
- Adding categorical variables to scatterplots
- Measuring linear association: Correlation
- Facts about correlation

Response Variables and Explanatory Variables

A **response variable** measures an outcome of a study.

An **explanatory variable** may explain or influence changes in a response variable.

Scatterplot (1 of 3)

- The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.

- A **scatterplot** shows the relationship between two quantitative variables that are measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

- Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Scatterplot (2 of 3)

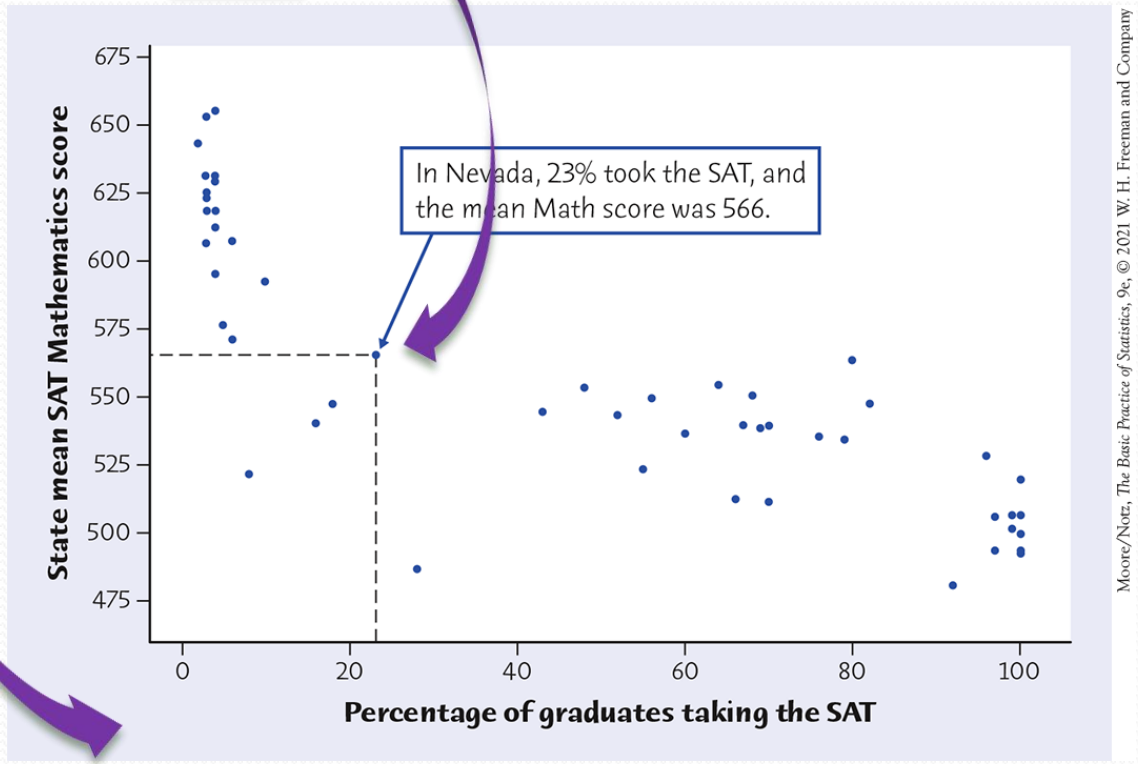
Making the scatterplot in the context of our four-step process:

- **STATE**: The research question of interest is stated as a statement (or a query about a statement) of the association between two variables in your data.
- **PLAN**: The solution of your problem is planned by plotting the variables according to the guidelines on the previous slide.
- **SOLVE**: Examine the scatterplot, taking note of any relationship present.
- **CONCLUDE**: We will explore this step later …

Scatterplot (3 of 3)

Example 4.3: Make a scatterplot of the relationship between the percent taking the SAT and the state's mean SAT Mathematics score.

% of state's grads taking SAT	23	55	80	8	18	10	45	100
State mean SAT Math score	566	551	538	571	548	592	547	490



Example 4.3 in Python: Make a scatterplot of the relationship between the percent taking the SAT and the state's mean SAT Math score.

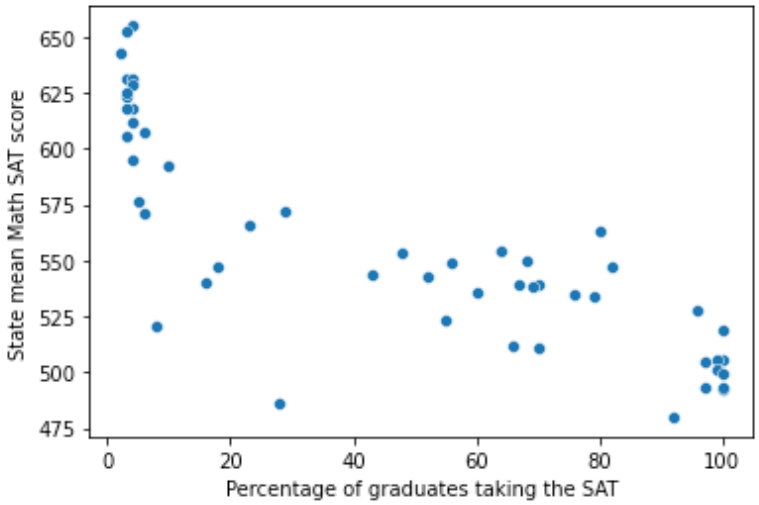
In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
```

In [2]:
```
mathsat = pd.read_csv("eg04-03mathsat.csv")
mathsat.head()
```

Out[2]:

	State	PctSAT	MathSAT2018
0	Alabama	6	571
1	Alaska	43	544
2	Arizona	29	572
3	Arkansas	5	576
4	California	60	536

In [3]:
```
sns.scatterplot(x = "PctSAT", y = "MathSAT2018", data = mathsat)
plt.xlabel("Percentage of graduates taking the SAT")
plt.ylabel("State mean Math SAT score")
plt.show()
```



Interpreting Scatterplots

To interpret a scatterplot, follow the basic strategy of data analysis from Chapters 1 and 2. Look for patterns and important departures from those patterns.

EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **direction, form,** and **strength** of the relationship.

An important kind of departure is an **outlier—**an individual value that falls outside the overall pattern of the relationship.

Direction of Association

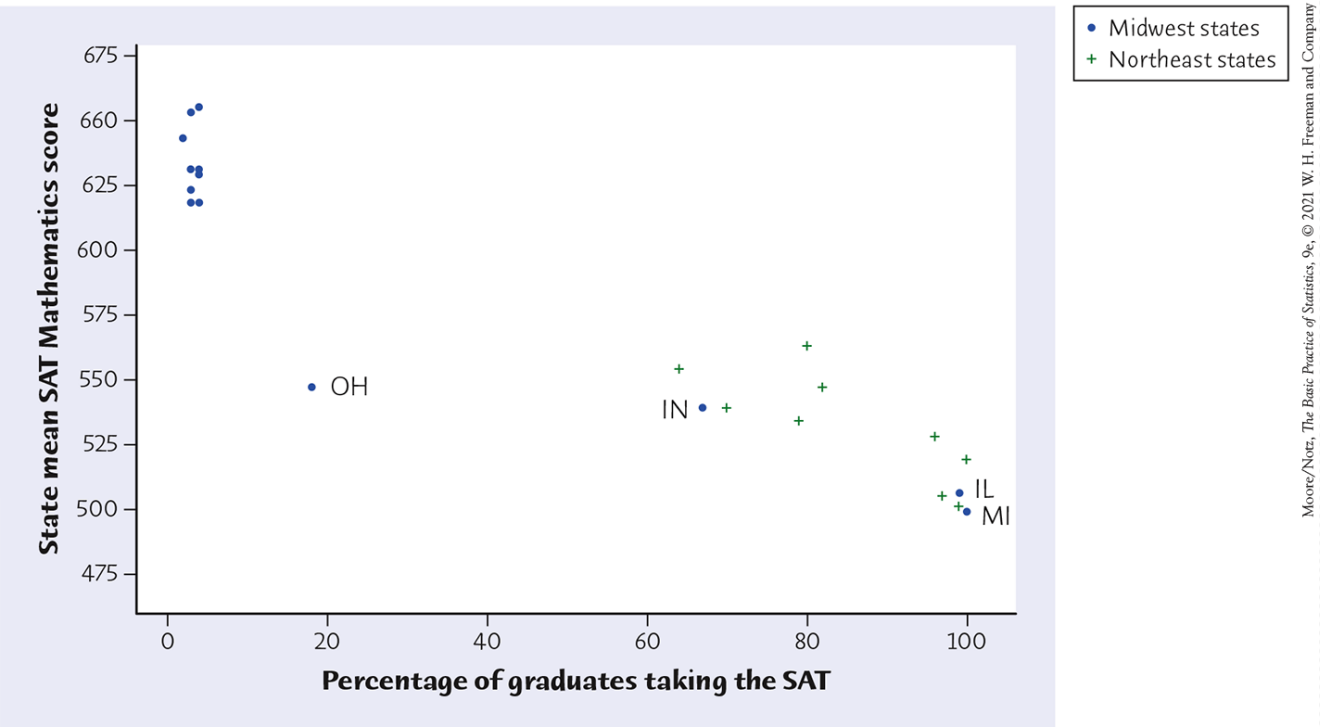
POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

- Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and below-average values also tend to occur together.
- Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa.

Adding Categorical Variables

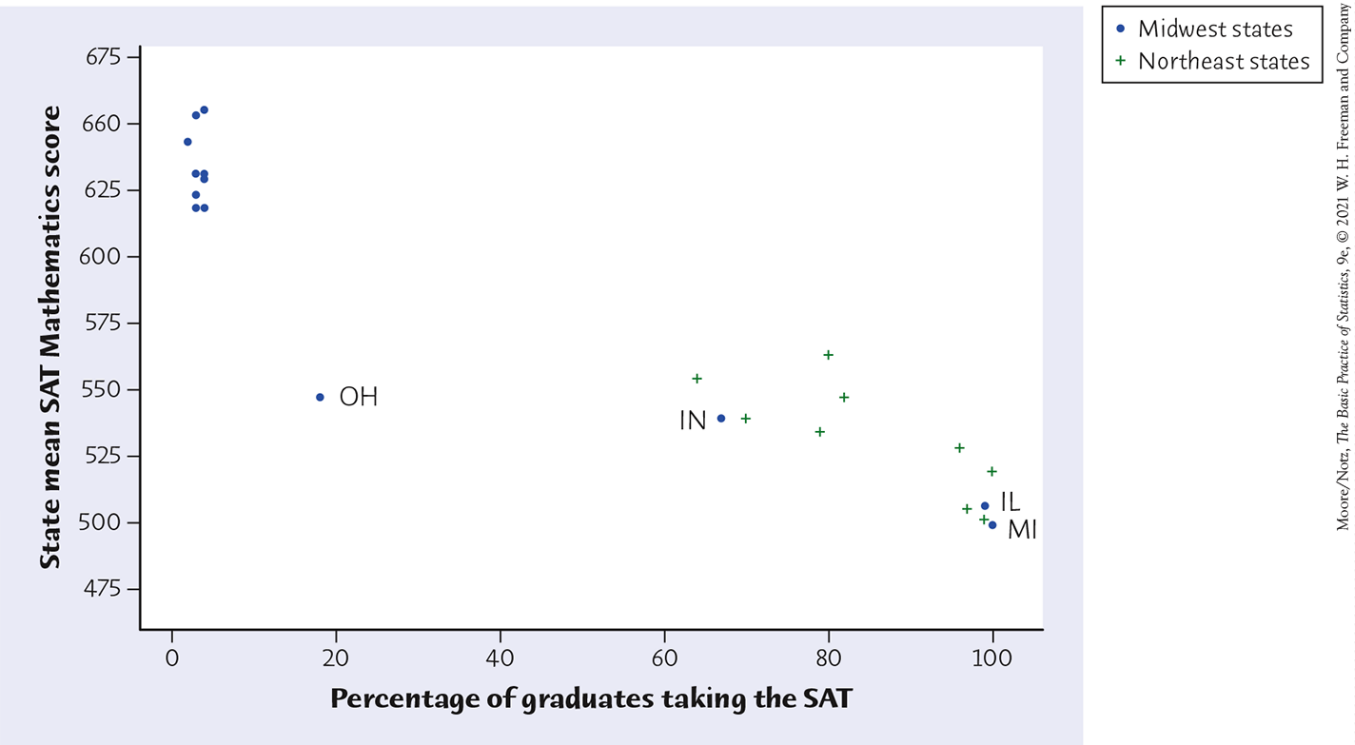
Consider the relationship between mean SAT Mathematics score and percent of high-school grads taking the SAT for each state.

Mean SAT Mathematics score and percent of high school graduates who take the test for only the Midwest (•) and Northeast (+) states.



Categorical Variables in Scatterplots

To add a categorical variable, use a different plot color or symbol for each category.



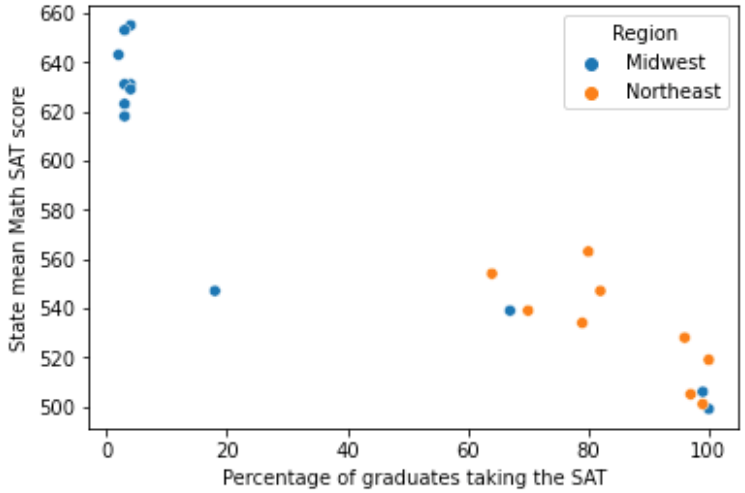
Categorical Variables in Scatterplots

```
In [4]: sat = pd.read_csv("eg04-04Midwest_Northeast.csv")
        sat.head()
```

Out[4]:

	State	PctSAT	MathSAT2018	Region
0	Minnesota	4	655	Midwest
1	Wisconsin	3	653	Midwest
2	NorthDakota	2	643	Midwest
3	Kansas	4	631	Midwest
4	Iowa	3	631	Midwest

```
In [5]: sns.scatterplot(x = "PctSAT", y = "MathSAT2018", hue="Region", data = sat)
        plt.xlabel("Percentage of graduates taking the SAT")
        plt.ylabel("State mean Math SAT score")
        plt.show()
```



- Mean SAT Mathematics score and percent of high school graduates who take the test for only the Midwest (•) and Northeast (•) states.

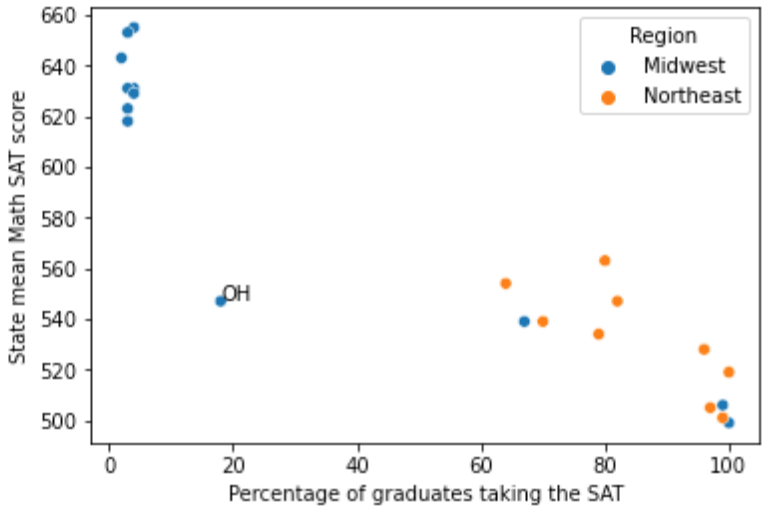
Categorical Variables in Scatterplots

```
In [6]: sat[sat.State=="Ohio"]
```

Out[6]:

	State	PctSAT	MathSAT2018	Region
8	Ohio	18	547	Midwest

```
In [7]: sns.scatterplot(x = "PctSAT", y = "MathSAT2018", hue="Region", data = sat)
        plt.xlabel("Percentage of graduates taking the SAT")
        plt.ylabel("State mean Math SAT score")
        ## adding text into the plot to indicate the point corresponding to Ohio
        plt.text(x=18, y=547, s='OH')
        plt.show()
```



Adding text to the plot to indicate OH

- Mean SAT Math score and percent of high school graduates who take the test for only the Midwest (•) and Northeast (•) states.

Measuring Linear Association

A scatterplot *displays* the strength, direction, and form of the relationship between two quantitative variables.

- The **correlation (or Pearson correlation coefficient)**, r, *measures* the direction and strength of the linear relationship between two quantitative variables.
- Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_y}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\cdots+\left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

Shorter form:

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

Facts about Correlation (1 of 2)

- Correlation makes no distinction between explanatory variables and response variables.
- r has no units and does not change when we change the units of measurement of x, y, or both.
- Positive r indicates positive association between the variables, and negative r indicates negative association.
- The correlation r is always a number between –1 and 1.

Facts about Correlation (2 of 2)

Cautions:

- Correlation requires that both variables be quantitative, so it makes sense to do the arithmetic indicated by the formula for r.
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is between them.
- Correlation is not resistant; *r* is strongly affected by a few outlying observations.
- Correlation is ***not*** a complete summary of two-variable data.

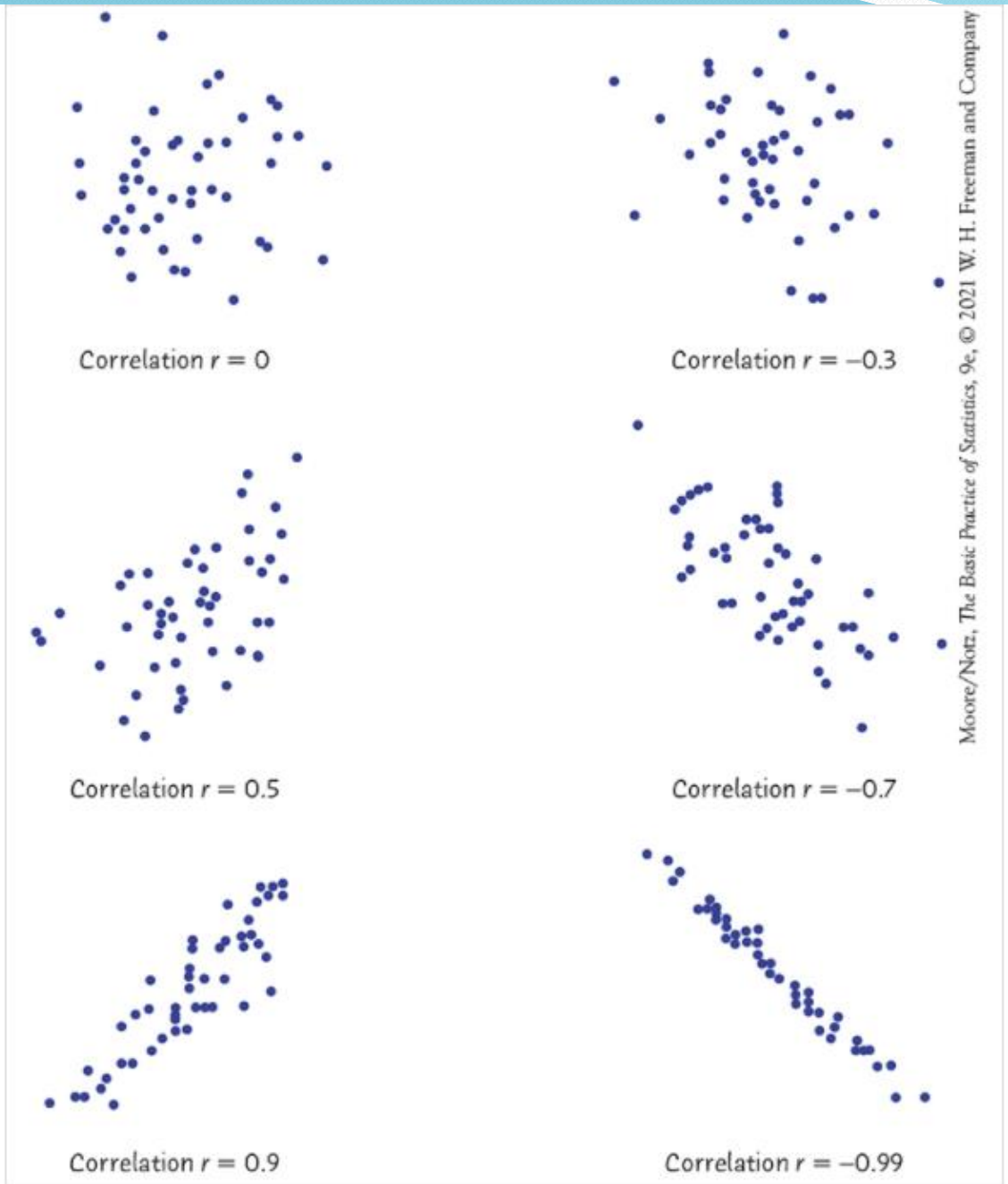


FIGURE 4.6

The scatterplots in Figure 4.6 illustrate how values of *r* closer to 1 or -1 correspond to stronger linear relationships.

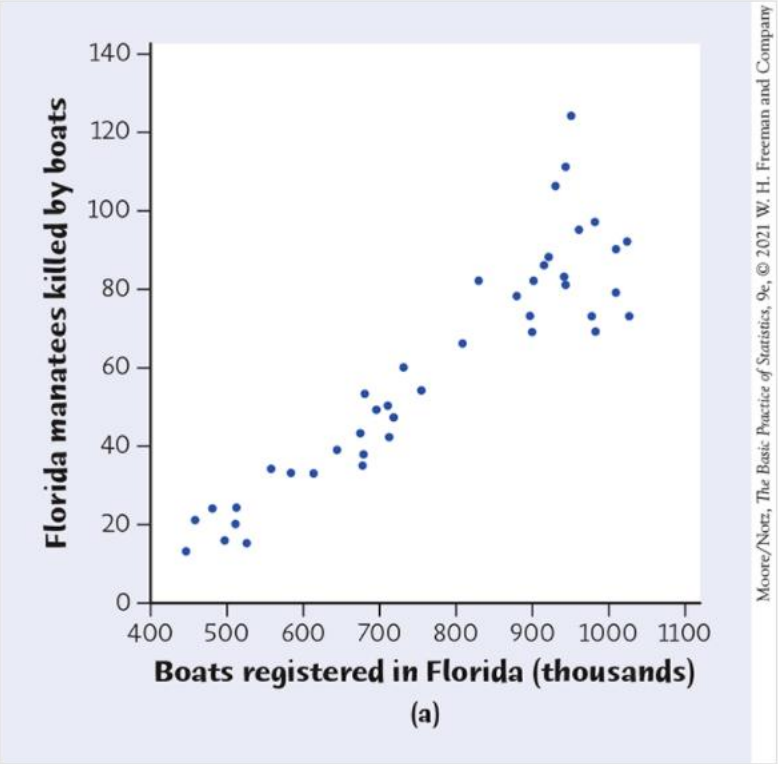


Figure 4.7 (a)

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

$r = 0.919$

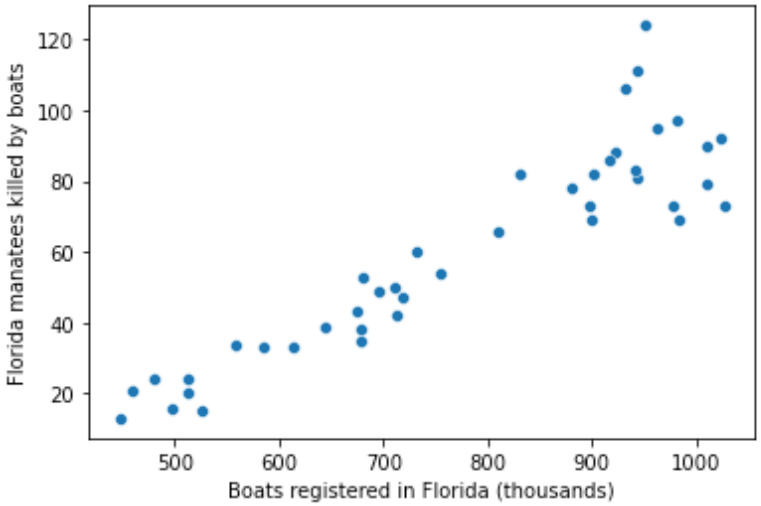
```
In [8]: florida = pd.read_csv("eg04-05manatee.csv")
        florida.head()
```

Out[8]:

click to scroll output; double click to hide

	Year	Boats	Kills
0	1977	447	13
1	1978	460	21
2	1979	481	24
3	1980	498	16
4	1981	513	24

```
In [9]: sns.scatterplot(x = "Boats", y = "Kills", data = florida)
        plt.xlabel("Boats registered in Florida (thousands)")
        plt.ylabel("Florida manatees killed by boats")
        plt.show()
```




```
In [10]: stats.pearsonr(x=florida['Boats'], y=florida['Kills'])[0]
```

Out[10]: `0.9189057628743729`

Example from Figure 4.7 (a) in Python

$r = 0.919$