

Stat 150 - Day 7
Scatterplots and Association

Example 1: Scrabble Names

(a) How many letters are in your first and last names? Record your answer in the table below:

ID	Name	Letters	Points	Ratio	ID	Name	Letters	Points	Ratio
1	Allison, Ryan				9	Pool, Benjamin			
2	Beemer, Joshua				10	Popoff, Zoya			
3	Ghirardo, Michael				11	Powers, Nicholas			
4	Klingmann, Maxwell				12	Said, Diana			
5	Maddalena, Julia				13	Samuels, Melody			
6	Maxwell, Alex				14	Shaffer, Matthew			
7	Moore, Christopher				15	Watts, Camille			
8	Nelson, Suzanne				16	Williams, Gregory			

(b) How many Scrabble points are in your names (again, combined)? Record your answer in the table above. (The point value of each letter is given in the table below.)

A	B	C	D	E	F	G	H	I	J	K	L	M
1	3	3	2	1	4	2	4	1	8	5	1	3
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	1	3	10	1	1	1	1	4	4	8	4	10

(c) Enter into Minitab the number of letters and number of Scrabble points in the last names of yourself and your classmates. (Do not bother to type in the students' names, but do name the columns appropriately. Do not worry about the ratio for now.)

A **scatterplot** displays the relationship between two quantitative variables. The *explanatory* variable (also called the *predictor*) appears on the horizontal axis and the response variable on the vertical axis. Each observational unit is represented by one point, fixed by the values of both variables for that individual.

(d) Use Minitab to create a scatterplot of height vs. foot length: use Graph> Scatterplot (clicking OK to create "Simple" scatterplot), then double click on the height column in the left box to specify it as the Y variable, then double click on the foot length column to specify it as the X variable.

(e) Does the scatterplot reveal a *tendency* for people with longer names to have more Scrabble points? Explain.

(f) Is it the case that *every* student with a longer name than another student has more Scrabble points? If not, identify a pair for which the student with the shorter name has more points.

- The overall pattern of a scatterplot includes the **form**, **direction**, and **strength** of the relationship.
- A common form is for the relationship to be **linear** (follow a straight line), although curved forms are also common, and another form is for the data to fall in clusters.
- Two variables are **positively associated** if above-average values of one variable tend to accompany above-average values of the other variable and below-average values of one tend to accompany below-average values of the other (reading left to right, the scatterplot slopes upward).
- Two variables are **negatively associated** if above-average values of one tend to accompany below-average values of the other, and vice (scatterplot slopes downward).
- The strength of the relationship concerns how closely the points follow a clear form.

(g) Describe the direction, strength, and form of the relationship between name length and Scrabble points.

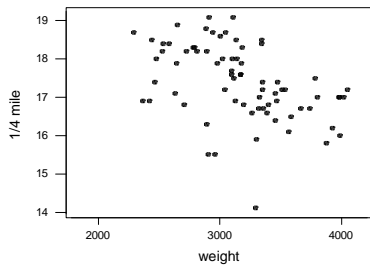
(h) Now use Minitab to calculate the *ratio* of points to letters for each person: At the MTB> prompt in the Session window, type: MTB> let c3=c2/c1. Name c3 appropriately. Examine a dotplot, histogram, and boxplot of the distribution of these ratios, and comment on what you observe.

(i) Produce a scatterplot to examine whether there is a relationship between number of letters and ratio. Then do the same for number of Scrabble points and ratio. Summarize what you find.

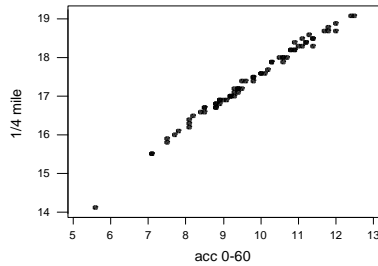
Example 2: New Car Data

The following nine scatterplots pertain to variables measured on models of new cars in 1999.

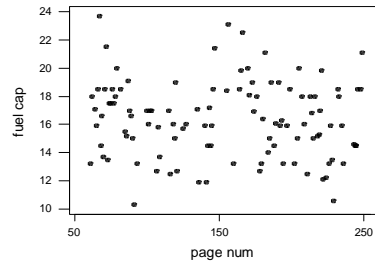
A:



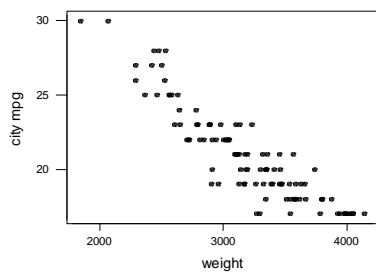
B:



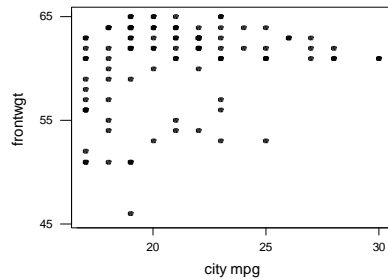
C:



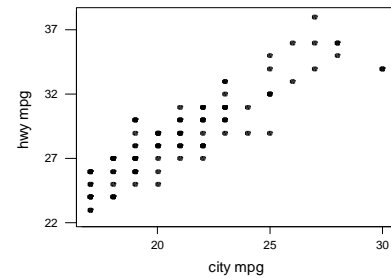
D:



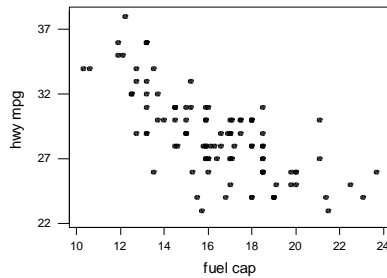
E:



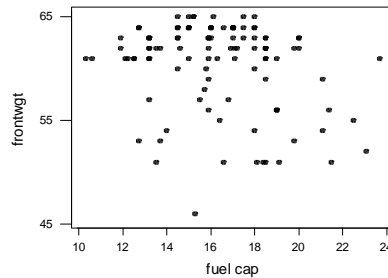
F:



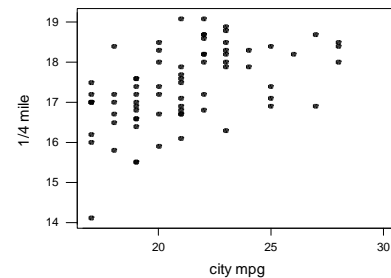
G:



H:



I:



Arrange these plots from the most strongly negative to the most strongly positive association, by writing the appropriate letters in the middle row of the table:

Strong negative		Moderate negative		Virtually none		Moderate positive		Strong positive

Example 3: Televisions and Life Expectancy

The data in `tvlife.mtw` provides information on life expectancy and number of televisions per thousand people in a sample of 22 countries, as reported by the *2006 World Almanac and Book of Facts*.

(a) Which of the countries has the fewest televisions per thousand people? Which country has the most televisions? Record those numbers.

Fewest:

Most:

(b) Use Minitab to produce a scatterplot of *life expectancy vs. televisions per thousand people*. Does there appear to be an association between the two variables? If so, describe its direction, strength, and form.

(c) One way to explore the form of the relationship is using a *smoother*. Right click on the scatterplot and choose Add > Smoother. Click OK in the next window. Using this smoothing function as a guide, do you think the relationship is best described as linear?

(d) Right click on the scatterplot again and select Add > Data Labels. Select the “Use labels from columns” option and specify C1 (country) in the box. Click OK. Which countries stand out as not following the same pattern as the rest of the data?

(e) Because the association between the variables is so strong, you might conclude that simply sending televisions to the countries with lower life expectancies would cause their inhabitants to live longer. Comment on this argument.