Introduction to Statistics I

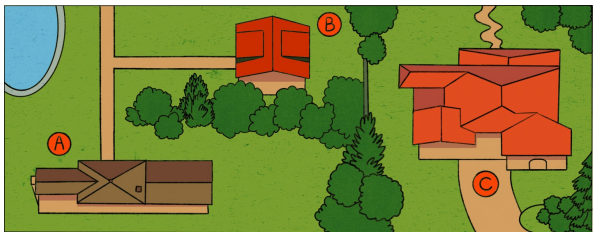
Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition by Rossman and Chance).

Previous Lecture

- z-score: $\frac{x-\bar{x}}{s}$
- Five Number Summary (FNS)
- Box plot 
- Outliers ●●● ●● ○
- Modified Box Plot

§4.1: Correlation

Correlation: describing how two variables are related to each other.



Example: How does square footage of a home affect its sale price? Explanatory/Response?

Explanatory: Square Footage (qual/quant?)

Response: Sale Price (qual/quant?)

Both vars are quantitative.

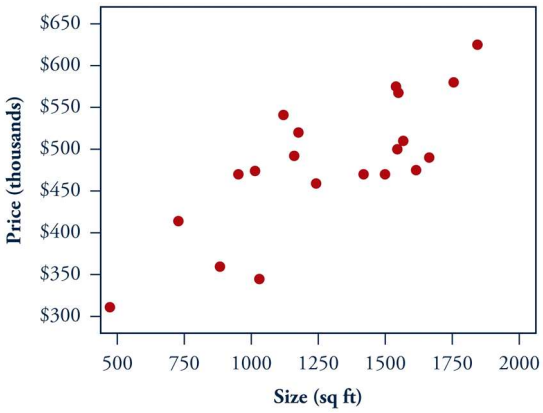
How should we describe the *relationship* between two quantitative vars?

House Price Data

Address	Price ($)	Size (sq ft)	Address	Price ($)	Size (sq ft)
2130 Beach St.	311,000	460	1030 Sycamore Dr.	490,000	1664
2545 Lancaster Dr.	344,720	1030	620 Eman Ct.	492,000	1160
415 Golden West Pl.	359,500	883	529 Adler St.	500,000	1545
990 Fair Oaks Ave.	414,000	728	646 Cerro Vista Cir.	510,000	1567
845 Pearl Dr.	459,000	1242	926 Sycamore Dr.	520,000	1176
1115 Rogers Ct.	470,000	1499	227 S. Alpine St.	541,000	1120
579 Halcyon Rd.	470,000	1419	654 Woodland Ct.	567,500	1549
1285 Poplar St.	470,000	952	2230 Paso Robles St.	575,000	1540
1080 Fair Oaks Ave.	474,000	1014	2461 Ocean St.	580,000	1755
690 Garfield Pl.	475,000	1615	833 Creekside Dr.	625,000	1844

table w/data from both vars

Construct a scatterplot!



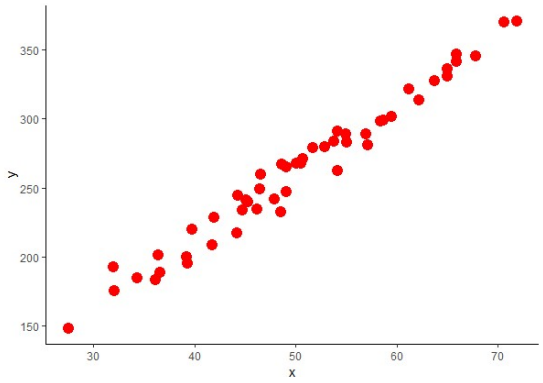
Scatterplots graphically display the relationship between two quantitative vars.

- Each pt is an obs-unit.
- Explanatory var is on horz-axis (*x*/bottom).
- Response var is on vertical-axis (*y*/side).
- Axes don't have to start at 0 for either var. Axes should cover values observed in data.

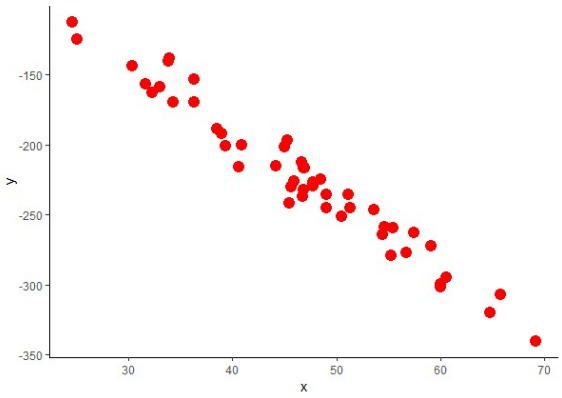
Describing an Relationship

When describing a relationship between two quantitative vars, we want to focus on:

Direction **Strength** **Form**

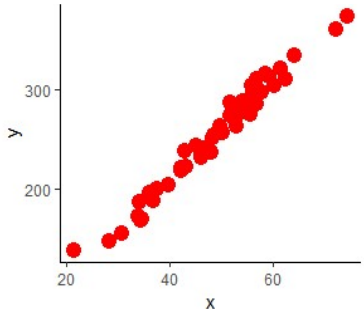


Positive Direction: As one var increases, so does the other.

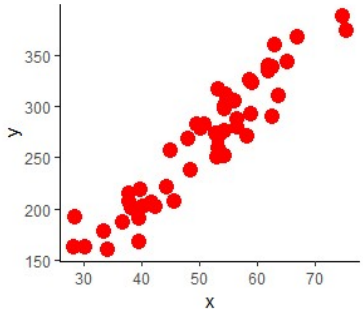


Negative Direction: As one var increases, the other var decreases.

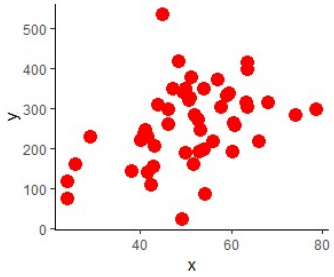
Strength: The closer the pts are to being a line, the stronger the relationship.



Strong

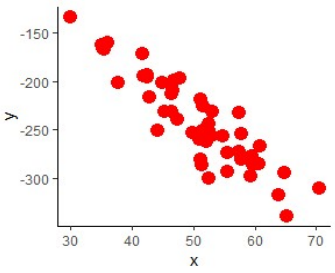


Moderate

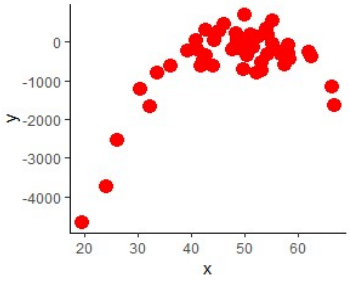


Very Weak/ None

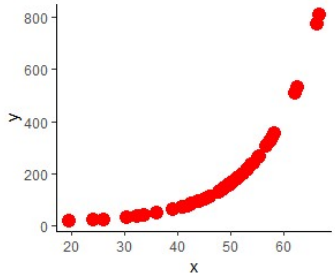
Form: can either be *linear* or *non-linear*.



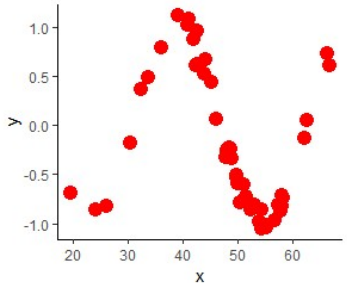
Linear



Non-Linear

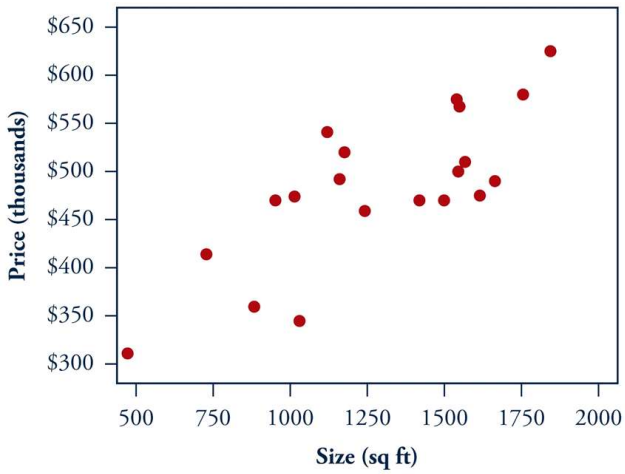


Non-Linear



Non-Linear

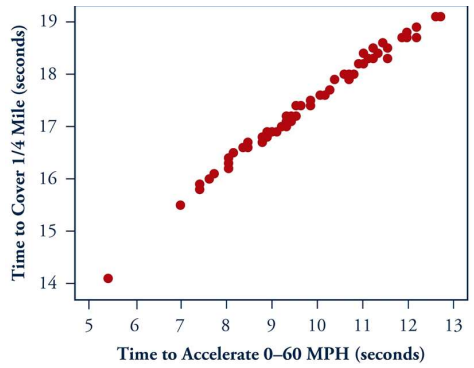
Recall: **House Price Data** - Assess this relationship.

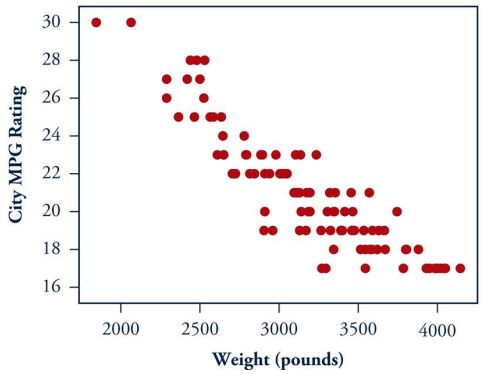


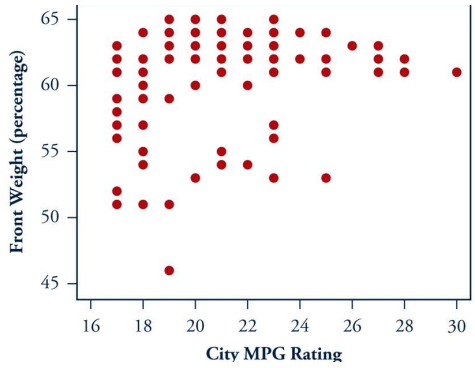
Housing Prices and Sqr Footage have **moderate**, **positive**, **linear** relationship.

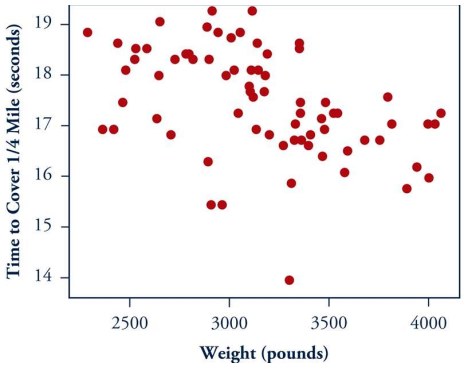
Automobile Stats

Assess the following relationships.





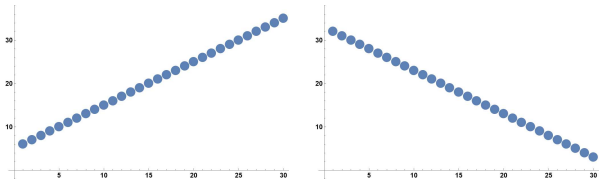




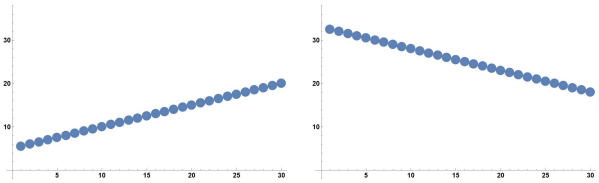
Correlation Coefficient (r)

Correlation coefficient describes direction & strength of a **linear** relationship.

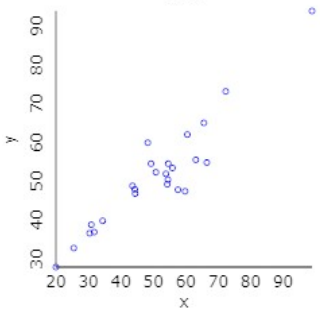
Correlation r can take values from -1 to 1. It measures how close the data follows a straight line.



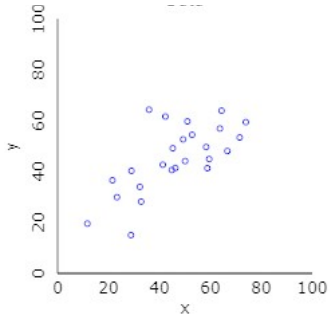
$r = 1$ $r = -1$



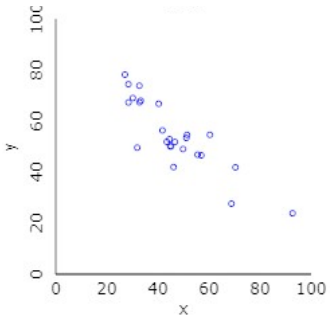
$r = 1$ $r = -1$



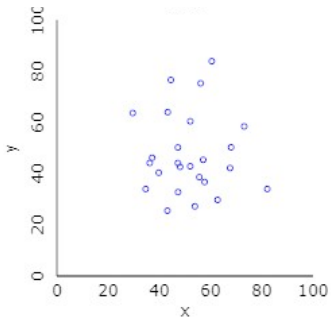
$r = 0.934$



$r = 0.684$

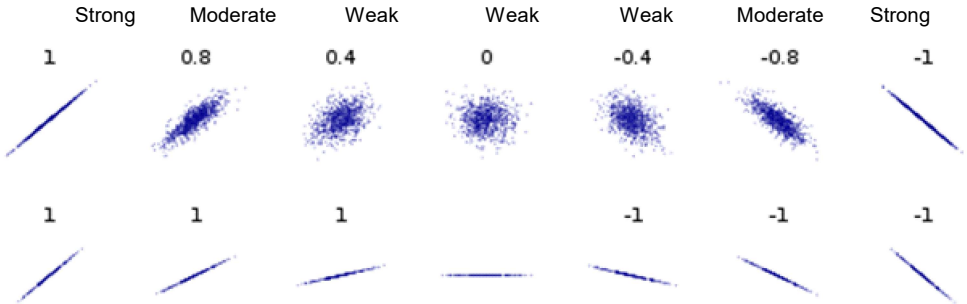


$r = -0.867$



$r = -0.035$

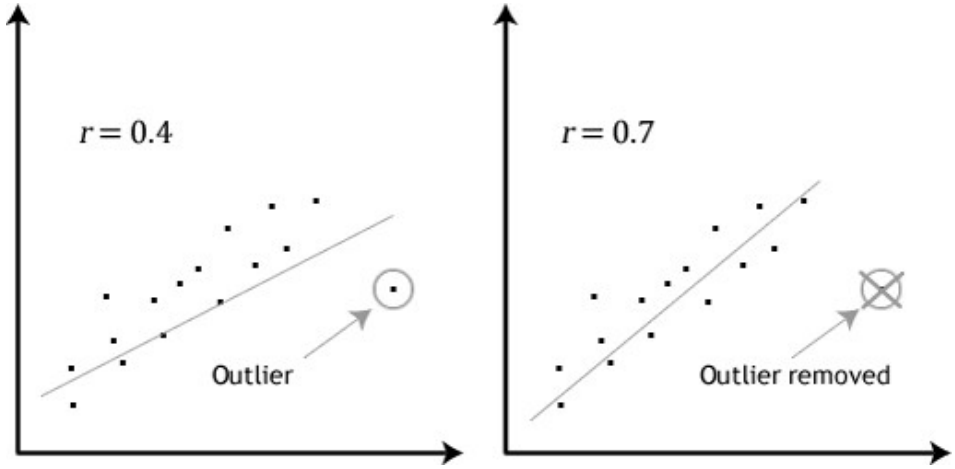
In case you wondered: $r := \frac{1}{n-1}\sum_{i-1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_i}\right)$. (!!!)



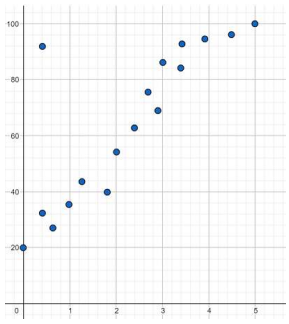
r values

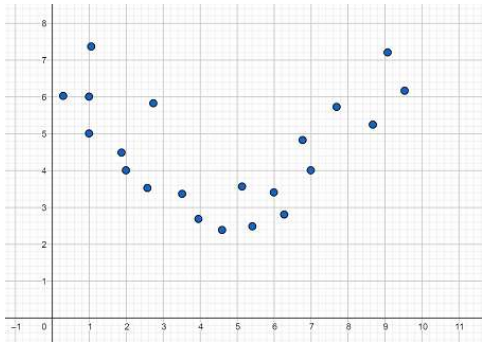
Note that the correlation coefficient does not assist us with nonlinear scatterplots.

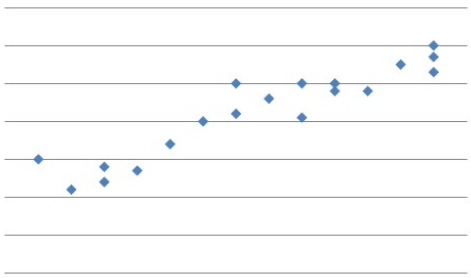
Correlation coefficients are sensitive to outliers. In other words, they are not resistant.



Which of the following scatterplots has the most trustworthy correlation coefficient?







Activity: 4.1a

What did we learn?

- Graphically display relationships between two quantitative vars
- Scatterplots: Direction, Strength, Form
- Correlation Coefficient (r)