Bivariate Data

Univariate (one variable) data

- Involves only a single variable
  - So cannot describe associations or relationships
- Descriptive Statistics
  - Central tendencies: mean, median, mode
  - Dispersion: range, max, min, quartiles, variance, standard deviation
- Visualizations
  - ~~Pie charts,~~ Bar charts, Line charts, Histograms, Box plots

Bivariate (two variables) data

- Involves two variables
  - So *can* describe associations or relationships
- Descriptive Statistics
  - Central tendencies: mean, median, mode
  - Dispersion: variance, standard deviation, covariance, correlation
- Visualizations
  - Scatter plots

Bivariate data and scatter plots

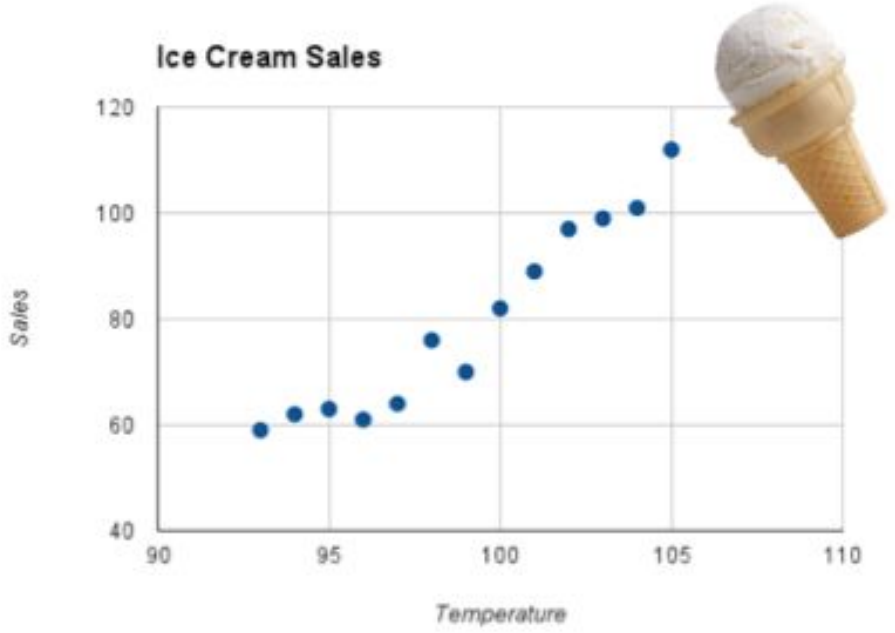


Image source

Bivariate data and scatter plots

- A scatter plot of bivariate data shows one variable vs. the other on a 2-dimensional graph
- If there is an explanatory variable, it is plotted it on the horizontal (*x*) axis, and the response variable is plotted on the vertical (*y*) axis
  - If there is no explanatory-response distinction either variable can be plotted on either axis
- *x* is also known as the independent variable, and *y* the dependent

Covariance & Correlation

Hybrid cars sold in the U.S. from 1997 to 2013

The variables:

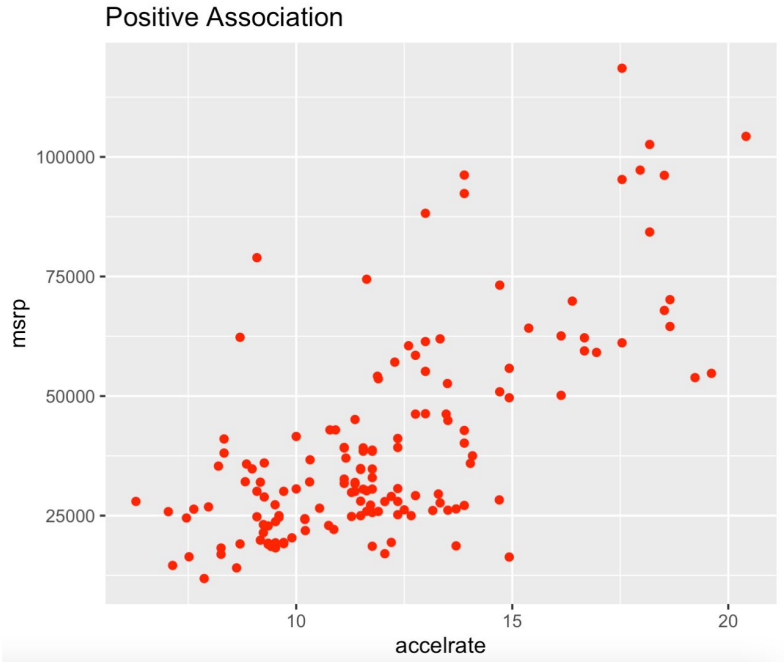
- Model of the car
- Year of manufacture
- MSRP (manufacturer's suggested retail price) in 2013 dollars
- Acceleration rate in km per second
- Fuel economy in miles per gallon
- Model’s class

First 15 rows of data

	vehicle	year	msrp	acceleration	mpg	class
1	Prius (1st Gen)	1997	24509.74	7.46	41.26	C
2	Tino	2000	35354.97	8.20	54.10	C
3	Prius (2nd Gen)	2000	26832.25	7.97	45.23	C
4	Insight	2000	18936.41	9.52	53.00	TS
5	Civic (1st Gen)	2001	25833.38	7.04	47.04	C
6	Insight	2001	19036.71	9.52	53.00	TS
7	Insight	2002	19137.01	9.71	53.00	TS
8	Alphard	2003	38084.77	8.33	40.46	MV
9	Insight	2003	19137.01	9.52	53.00	TS
10	Civic	2003	14071.92	8.62	41.00	C
11	Escape	2004	36676.10	10.32	31.99	SUV
12	Insight	2004	19237.31	9.35	52.00	TS
13	Prius	2004	20355.64	9.90	46.00	M
14	Silverado 15 2WD	2004	30089.64	9.09	17.00	PT
15	Lexus RX400h	2005	58521.14	12.76	28.23	SUV

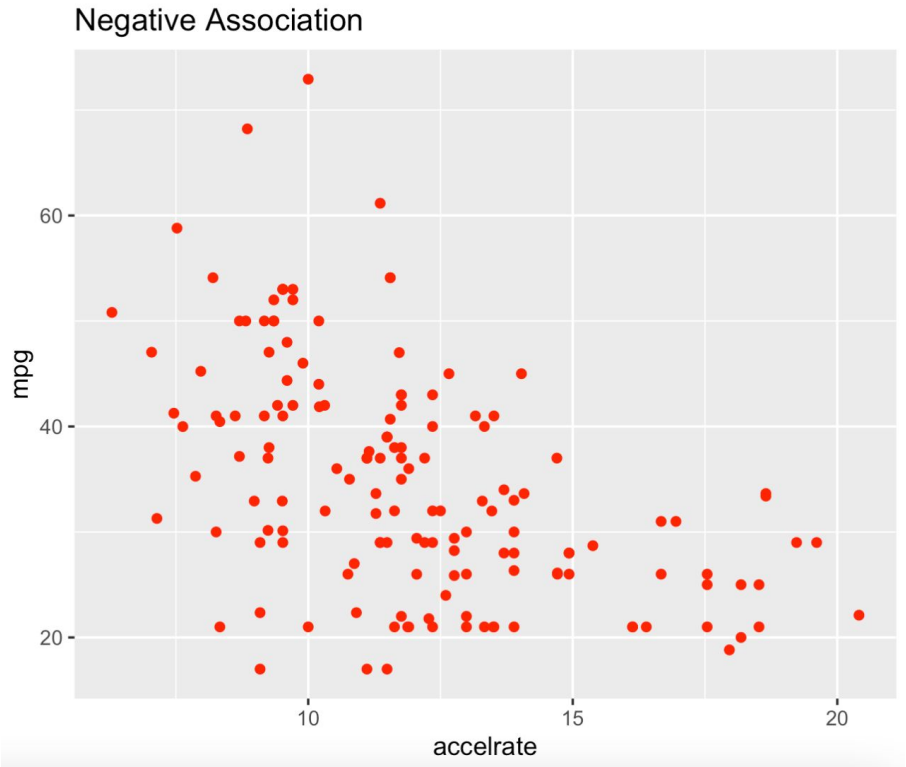
Positive association

- The points are scattered in an upward direction, indicating that cars with greater acceleration tend to cost more
- Conversely, cars that cost more tend to have greater acceleration
- This is an example of positive association: above average values of one variable tend to be associated with above average values of the other



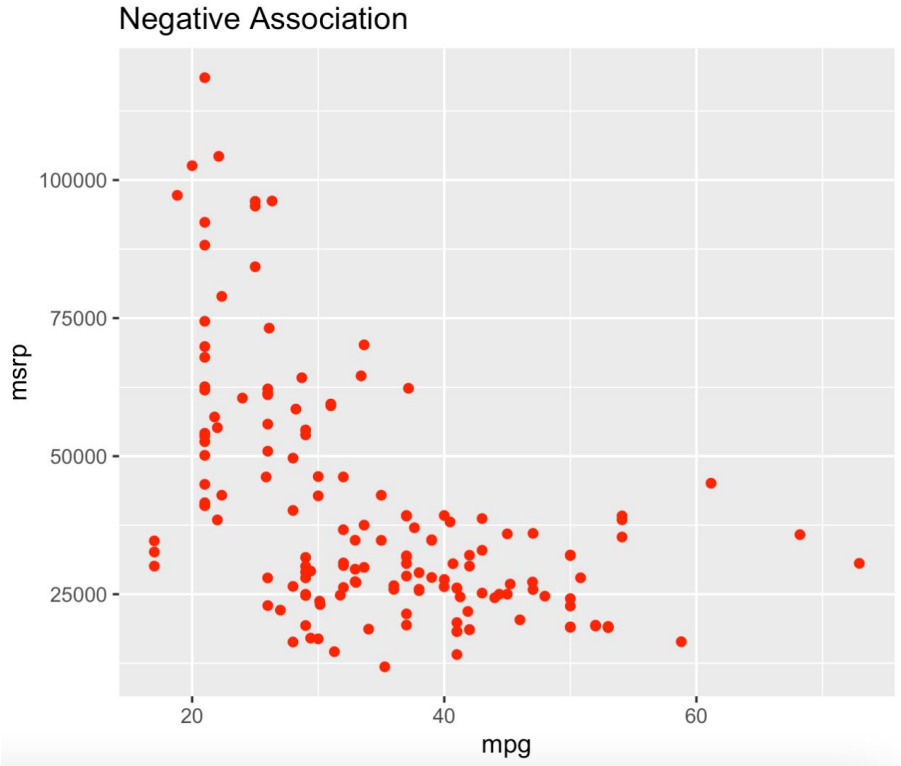
Negative association

- There is a clear downward trend: i.e., a negative association
- Hybrid (or perhaps all) cars that accelerate faster tend to get fewer miles per gallon; conversely, cars that get fewer miles per gallon tend to accelerate more slowly



Negative association

- There is a clear downward trend: i.e., a negative association
- Hybrid cars with higher mpg tend to cost less; conversely, cars that cost more tend to have lower mpg
- This might seem confusing at first, until we consider that cars that accelerate faster tend to be less fuel efficient and have lower mpg



Sample Covariance

c is the average product, across all observations, of deviations (i.e., the differences between the measurements and their sample means)

$$s_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x}) (y_i - \bar{y})$$

Standard Units

Problem: scales differ (e.g., Fahrenheit degrees *x* dollars vs. Fahrenheit degrees *x* euros vs. Celsius degrees *x* euros)

Solution: normalize the deviations
I.e., divide by a measure of spread

$$\frac{x_i - \mu_x}{\sigma_x}$$

Interpretation:
The measurement is $\frac{x_i - \mu_x}{\sigma_x}$ many standard deviations from the mean

Correlation = normalized covariance

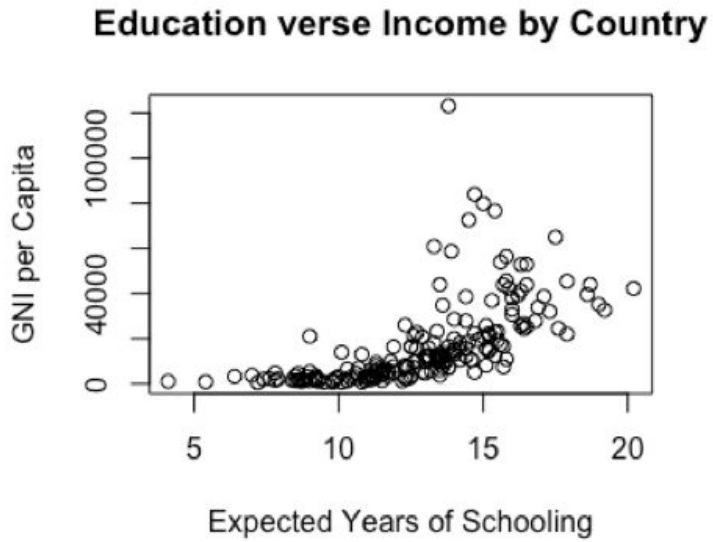
Sample Correlation

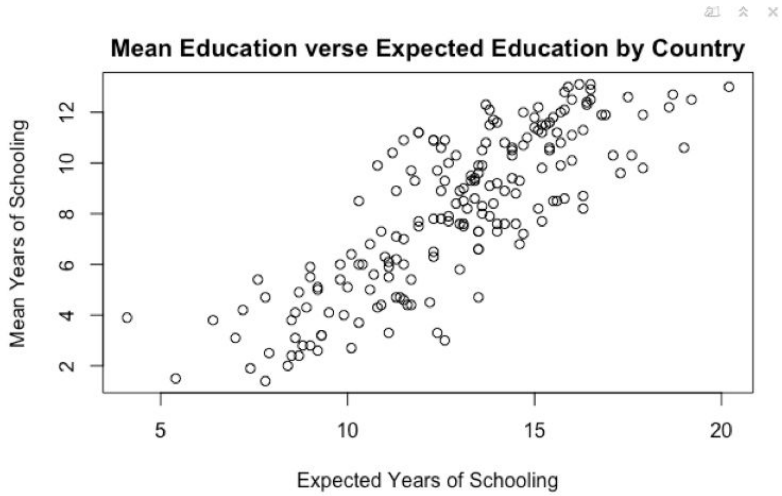
r is the average product, across all observations, of deviations, measured in standard units

$$r_{XY} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_{XX}} \right) \left(\frac{y_i - \bar{y}}{s_{YY}} \right)$$

Data Source: 2014 HDI Index

Standardized Units



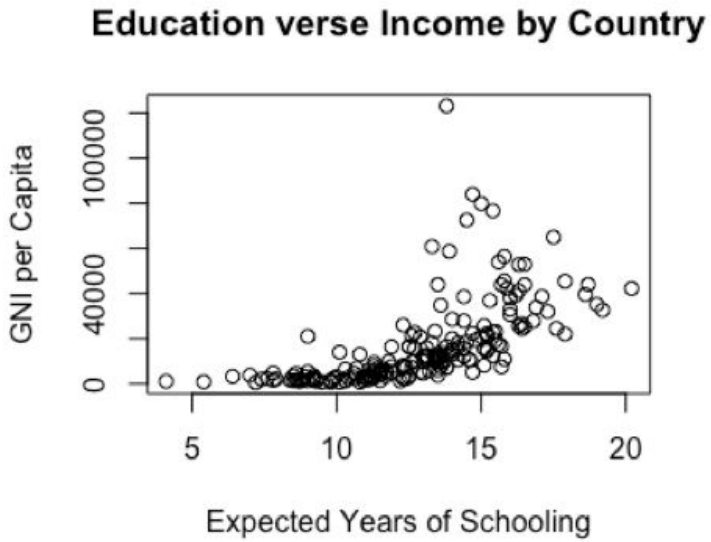


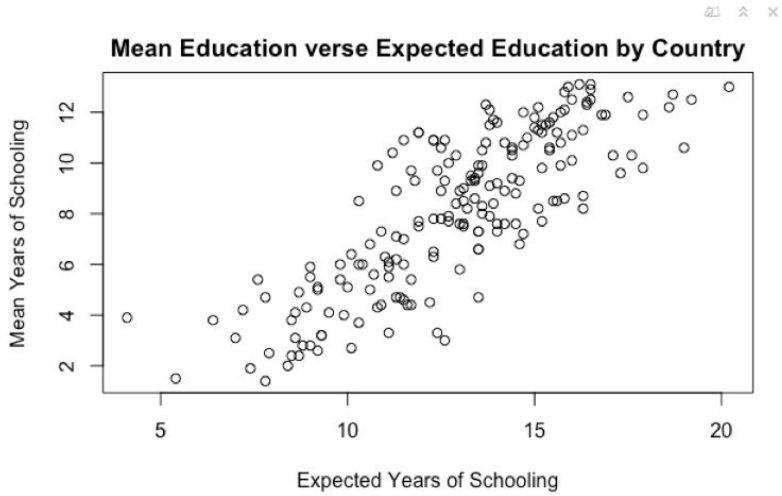
- Cov(GNI per Capita, Expected Years of Schooling) ∽ 32898.63
- Cov(Mean Years of Schooling, Expected Years of Schooling) ∽ 7.232066
- This is *not* intuitive: Covariance with Mean Years of Schooling should be higher!

Expected Years of Schooling: number of years of schooling a child should expect, given current enrollment rates

Data Source: 2014 HDI Index

Standardized Units



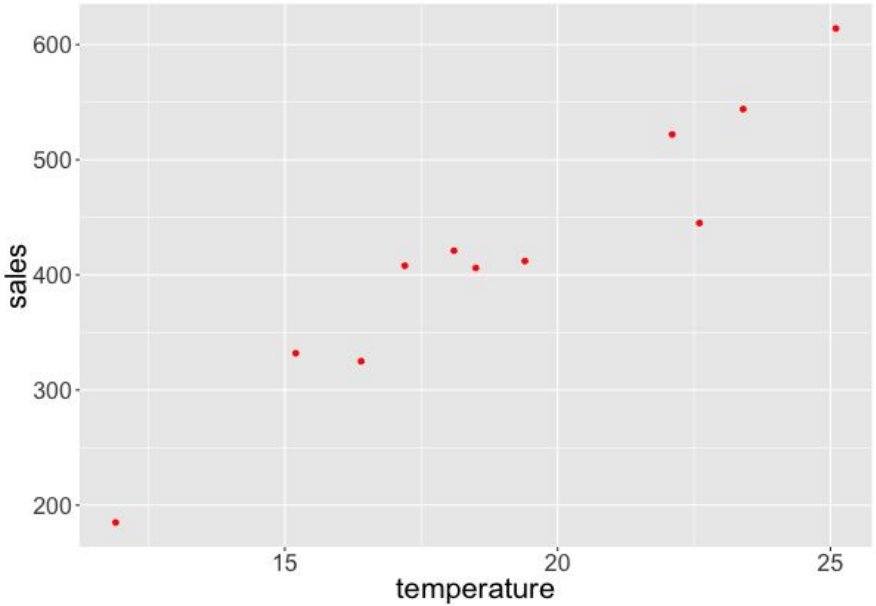


- Cor(GNI per Capita, Expected Years of Schooling) ∽ 0.610305
- Cor(Mean Years of Schooling, Expected Years of Schooling) ∽ 0.8152542
- This *is* intuitive: Covariance with Mean Years of Schooling is now higher!

A toy example: Ice cream sales

Based on the scatter plot of temperature vs. ice cream sales, we expect *r* to be positive and close to 1, as there seems to be a strong linear relationship

	temperature	sales
1	16.4	325
2	11.9	185
3	15.2	332
4	18.5	406
5	22.1	522
6	19.4	412
7	25.1	614
8	23.4	544
9	18.1	421
10	22.6	445
11	17.2	408



Step 1: Convert values to standard units

- Let's calculate temperature in standard units, starting with 16.4
- mean of temperature: 19.08
- standard deviation of temperature: 3.755

Value in standard units =
(value - average) / SD

Temperature 16.4 in standard units =
(16.4 - 19.08) / 3.755 =
-0.714116

All the values in the columns of temperatures and sales in standard units were calculated this way.

	temperature	sales	temperature_standard_units	sales_standard_units
1	16.4	325	-0.71411617	-0.84811268
2	11.9	185	-1.91237891	-2.10518057
3	15.2	332	-1.03365290	-0.78525929
4	18.5	406	-0.15492690	-0.12080912
5	22.1	522	0.80368329	0.92076141
6	19.4	412	0.08472565	-0.06693478
7	25.1	614	1.60252511	1.74683459
8	23.4	544	1.14984808	1.11830065
9	18.1	421	-0.26143914	0.01387672
10	22.6	445	0.93682359	0.22937408
11	17.2	408	-0.50109169	-0.10285101

Step 2: Multiply corresponding pairs of values in standard units

	temperature	sales	temperature_standard_units	sales_standard_units	product_of_standard_units
1	16.4	325	-0.71411617	-0.84811268	0.605650985
2	11.9	185	-1.91237891	-2.10518057	4.025902917
3	15.2	332	-1.03365290	-0.78525929	0.811685544
4	18.5	406	-0.15492690	-0.12080912	0.018716583
5	22.1	522	0.80368329	0.92076141	0.740000559
6	19.4	412	0.08472565	-0.06693478	-0.005671093
7	25.1	614	1.60252511	1.74683459	2.799346300
8	23.4	544	1.14984808	1.11830065	1.285875854
9	18.1	421	-0.26143914	0.01387672	-0.003627919
10	22.6	445	0.93682359	0.22937408	0.214883045
11	17.2	408	-0.50109169	-0.10285101	0.051537786

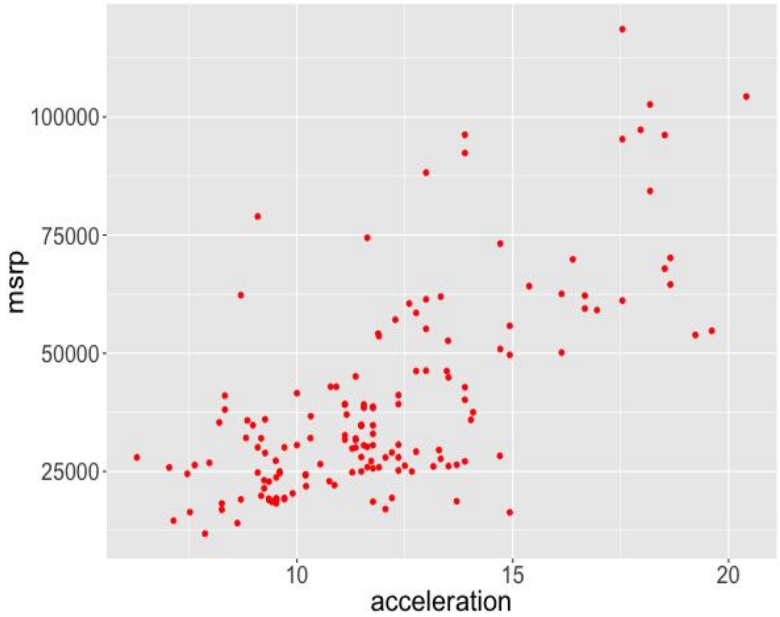
Step 3: *r* is the average of the products of the values in standard units

- We compute the mean of the products in standard units
- In this case *r* is 0.9585728
- This confirms our conjecture that *r* is positive and close to 1

product_of_standard_units
0.605650985
4.025902917
0.811685544
0.018716583
0.740000559
-0.005671093
2.799346300
1.285875854
-0.003627919
0.214883045
0.051537786

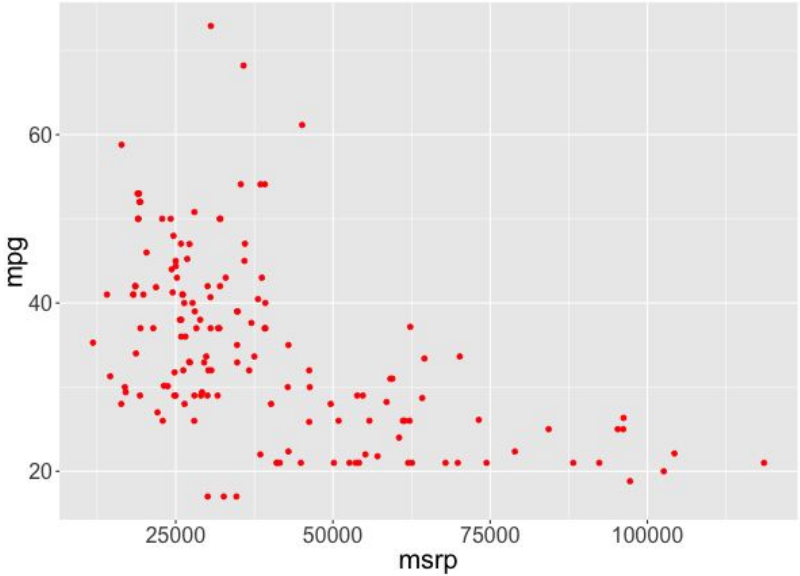
Computing *r* in R

```
> hybrid <- read.csv("hybrid.csv")
> cor(hybrid$msrp, hybrid$accelrate)
[1] 0.6955779
```



Computing *r* in R

```
> hybrid <- read.csv("hybrid.csv")
> cor(hybrid$msrp, hybrid$mpg)
[1] -0.5318264
```



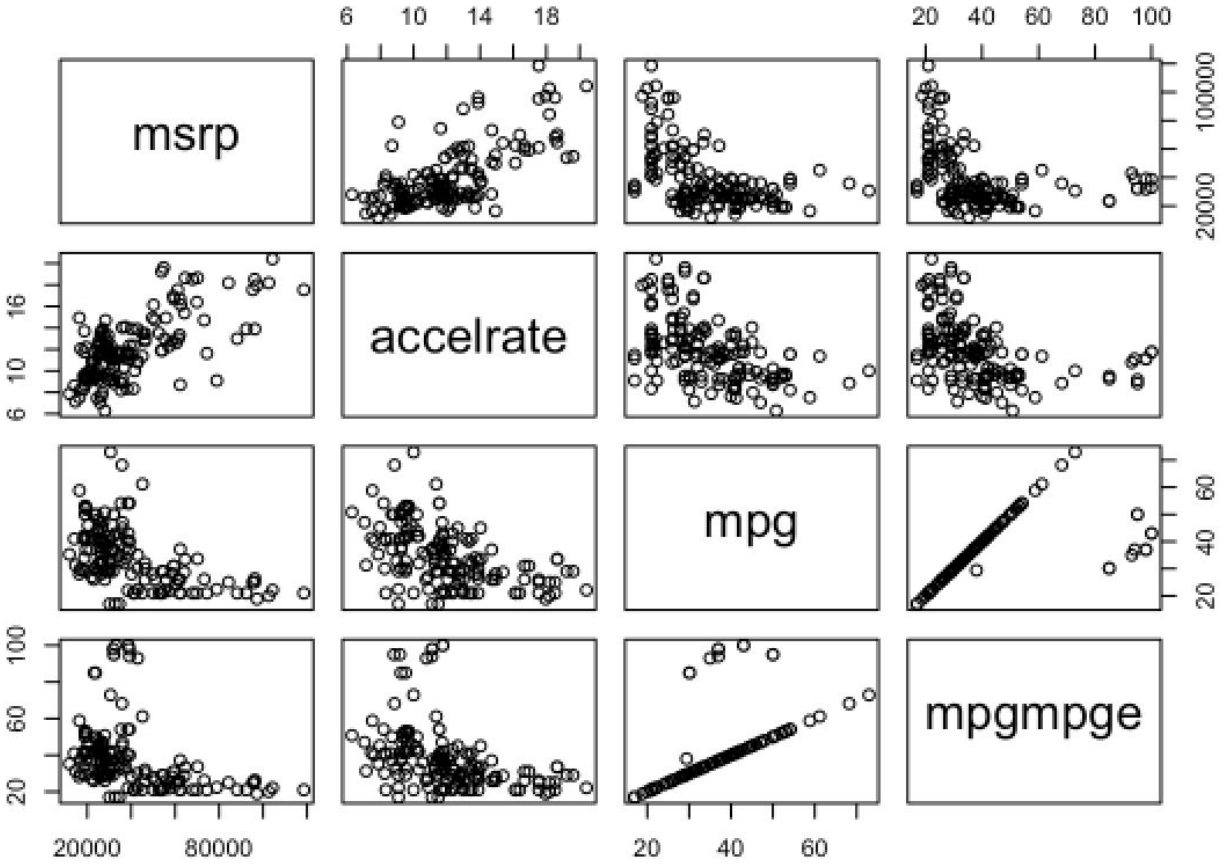
Correlation Matrix

```
> hybrid <- read.csv("hybrid.csv")
> x <- select(hybrid, msrp:mpgmpge)
> corr(x, x)
```

	msrp	accelrate	mpg	mpgmpge
msrp	1.0000000	0.6955779	-0.5318264	-0.3722185
accelrate	0.6955779	1.0000000	-0.5060704	-0.3988673
mpg	-0.5318264	-0.5060704	1.0000000	0.6677531
mpgmpge	-0.3722185	-0.3988673	0.6677531	1.0000000

Pairs

```
> pairs(x)
```



Summary

Correlation measures the direction and strength of a linear relationship between two quantitative variables

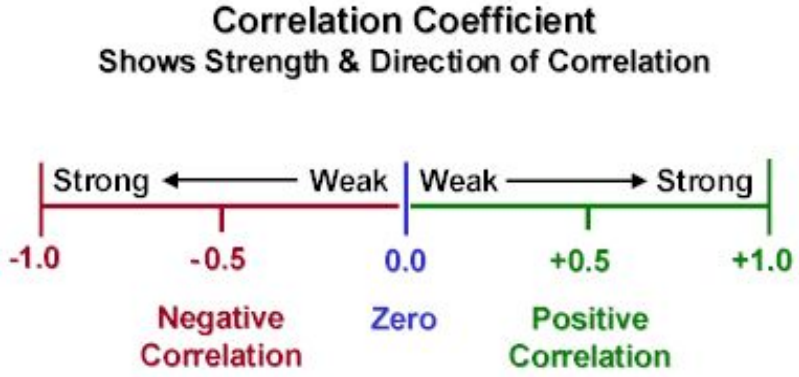


Image source


CAUTION

⚠ CAUTION

Correlation is powerful and simple but easy to misinterpret:

- *Correlation does not imply causation!*
  - Correlation only measures association.
- Correlation only measures linear association.
- Outliers can have a significant effect on correlation.
- Correlation can be misleading when data are aggregated.

Correlation does not imply causation

Intuitive Example:

- Imagine a positive correlation between math abilities and kids' weight.
- Does this imply that students who are better at math gain weight easily?
- Or that gaining weight can improve a student's math abilities?
- No! Of course not!
- Age is a confounding variable, which explains the correlation.

Older children both weigh more and are better at math than younger children.

A 2012 paper in the New England Journal of Medicine

Some responded harshly:
http://blogs.scientificamerican.com/the-curious-wavefunction/chocolate-consumption-and-nobel-prizes-a-bizarre-juxtaposition-if-there-ever-was-one/

Other responses were more nuanced:
http://www.reuters.com/article/us-eat-chocolate-win-the-nobel-prize-idUSBRE8991MS20121010#vFdfFkbPVlilSjsB.97

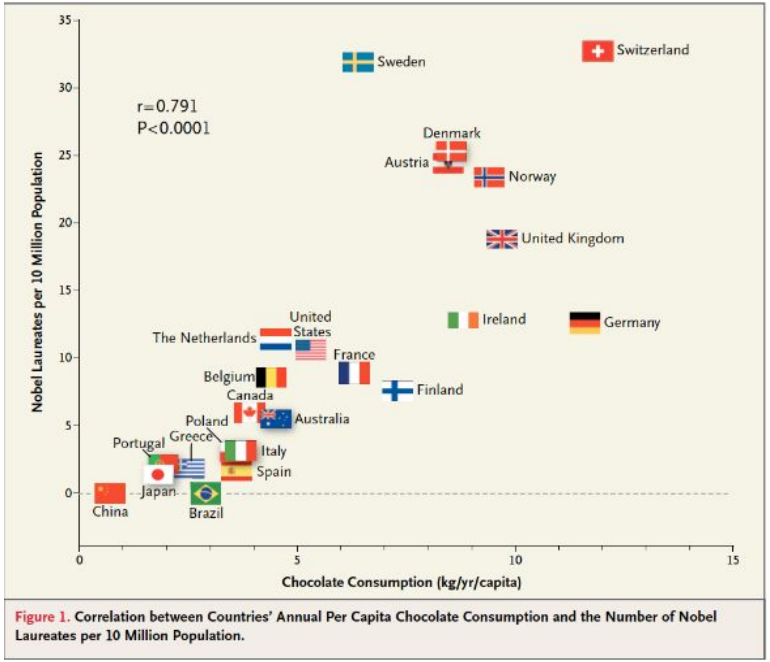
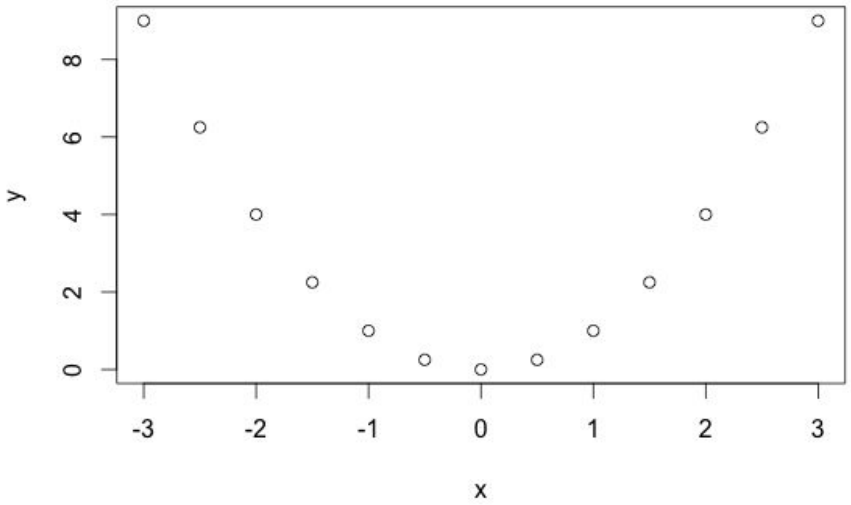


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Image source

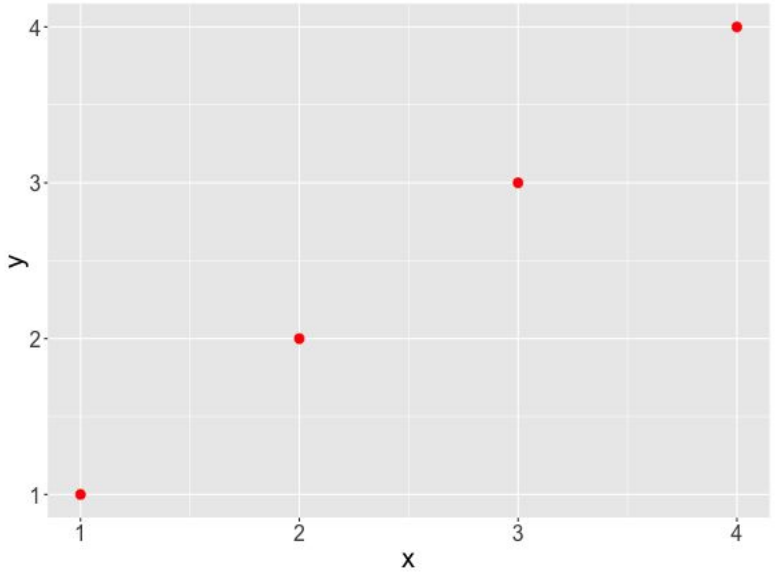
Correlation measures linear associations

```
> x <- seq(-3, 3, by = 0.5)
> y <- x ** 2
> cor(x, y)
[1] 0
> plot(x, y, xlab = "x", ylab = "y")
```

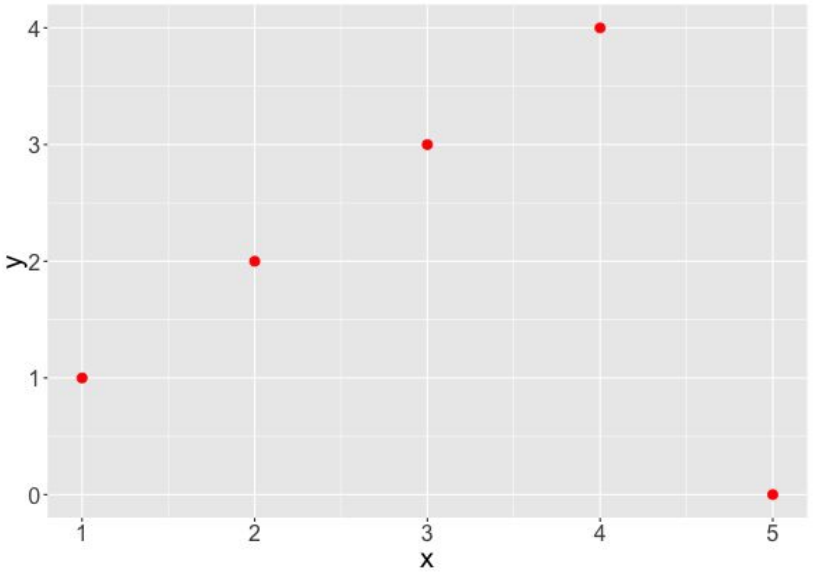


Outliers can gravely impact correlation

```
> cor(line$x, line$y)
[1] 1
```




```
> cor(outlier$x, outlier$y)
[1] 0
```



Properties of Covariance

- Symmetric measure: does not distinguish between the explanatory and response variables
- Both variables must be quantitative
- If two variables are independent, then their covariance is 0.
- But if the covariance of two variables is zero, they are not necessarily independent, because covariance captures only linear associations.
    - The following data sets exhibit zero or near-zero covariance:

Image Source

Properties of Correlation

- Symmetric measure: does not distinguish between the explanatory and response variables
- Both variables must be quantitative
- If two variables are independent, then their correlation is 0.
- But if the correlation of two variables is zero, they are not necessarily independent, because correlation captures only linear associations.
- Is standardized, so is invariant to change of units
- Is a number between -1 and 1