14.2) Correlation

Vitor Kamada

December 2019

Reference

Tables, Graphics, and Figures from

Computational and Inferential Thinking: The Foundations of Data Science

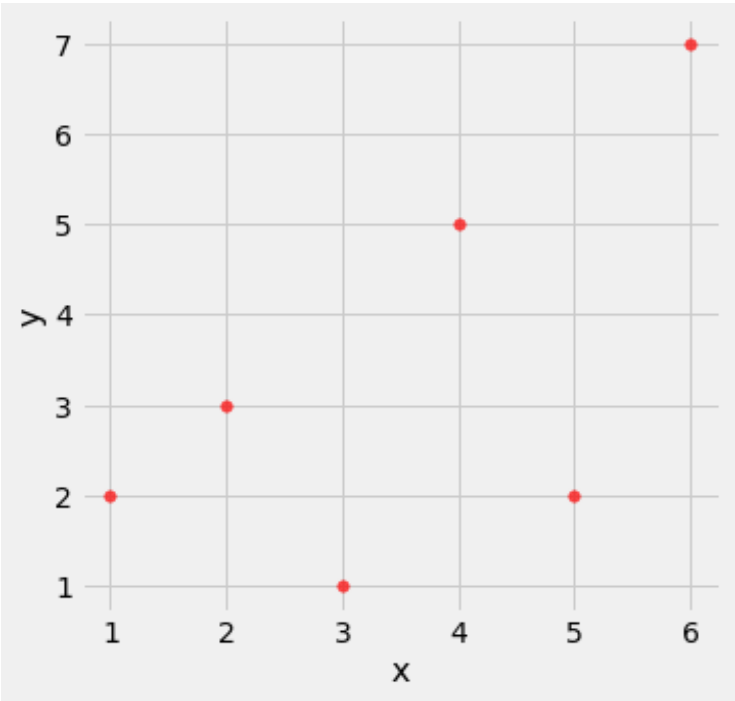
Adhikari & DeNero (2019): Ch 15.1 Correlation

https://www.inferentialthinking.com/

```
from datascience import *
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
```

```
x = np.arange(1, 7, 1)
y = make_array(2, 3, 1, 5, 2, 7)
t = Table().with_columns(
        'x', x,
        'y', y)
```

```
t.scatter(0, 1, s=30, color='red')
```



```
def standard_units(any_numbers):
    "Convert any array of numbers to standard units."
    return (any_numbers - np.mean(any_numbers))/np.std(any_numbers)
```

```
t_su = t.with_columns(
        'x (standard units)', standard_units(x),
        'y (standard units)', standard_units(y))
```

x	y	x (standard units)	y (standard units)
1	2	-1.46385	-0.648886
2	3	-0.87831	-0.162221
3	1	-0.29277	-1.13555
4	5	0.29277	0.811107
5	2	0.87831	-0.648886
6	7	1.46385	1.78444

```
t_product = t_su.with_column('product of standard units',
                             t_su.column(2) * t_su.column(3))
```

x	y	x (standard units)	y (standard units)	product of standard units
1	2	-1.46385	-0.648886	0.949871
2	3	-0.87831	-0.162221	0.142481
3	1	-0.29277	-1.13555	0.332455
4	5	0.29277	0.811107	0.237468
5	2	0.87831	-0.648886	-0.569923
6	7	1.46385	1.78444	2.61215

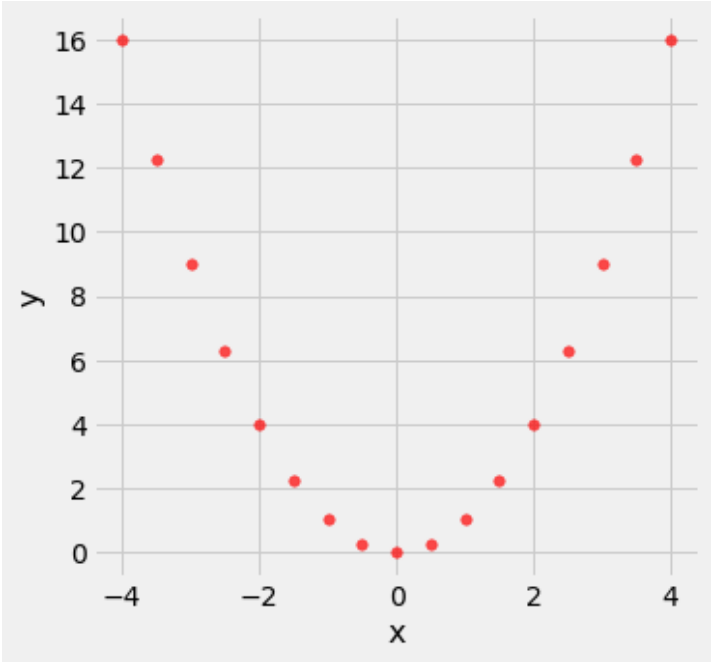
```
r = np.mean(t_product.column(4))
```

```
0.6174163971897709
```

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

```
def correlation(t, x, y):
    return np.mean(standard_units(t.column(x))*standard_units(t.column(y)))
```

```
new_x = np.arange(-4, 4.1, 0.5)
nonlinear = Table().with_columns(
        'x', new_x,
        'y', new_x**2)
nonlinear.scatter('x', 'y', s=30, color='r')
```




```
correlation(nonlinear, 'x', 'y')
```

0.0

Ecological Correlations Should be Interpreted with Care

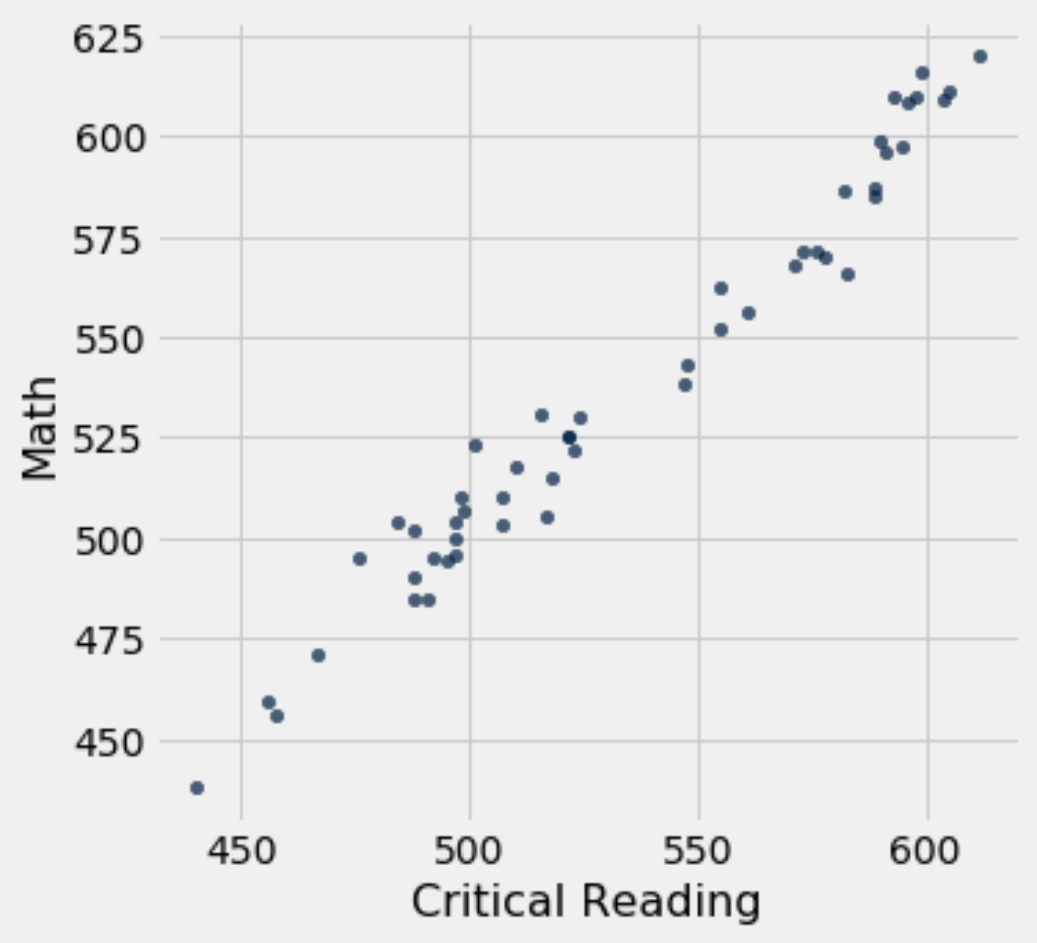
Correlations based on aggregated data can be misleading

SAT scores in 2014

```
path_data = 'https://github.com/data-8/textbook/raw/gh-pages/data/'
sat2014 = Table.read_table(path_data + 'sat2014.csv').sort('State')
```

State	Participation Rate	Critical Reading	Math	Writing	Combined
Alabama	6.7	547	538	532	1617
Alaska	54.2	507	503	475	1485
Arizona	36.4	522	525	500	1547

```
sat2014.scatter('Critical Reading', 'Math')
```




```
correlation(sat2014, 'Critical Reading', 'Math')
```

```
0.9847558411067434
```

New England Journal of Medicine 2012

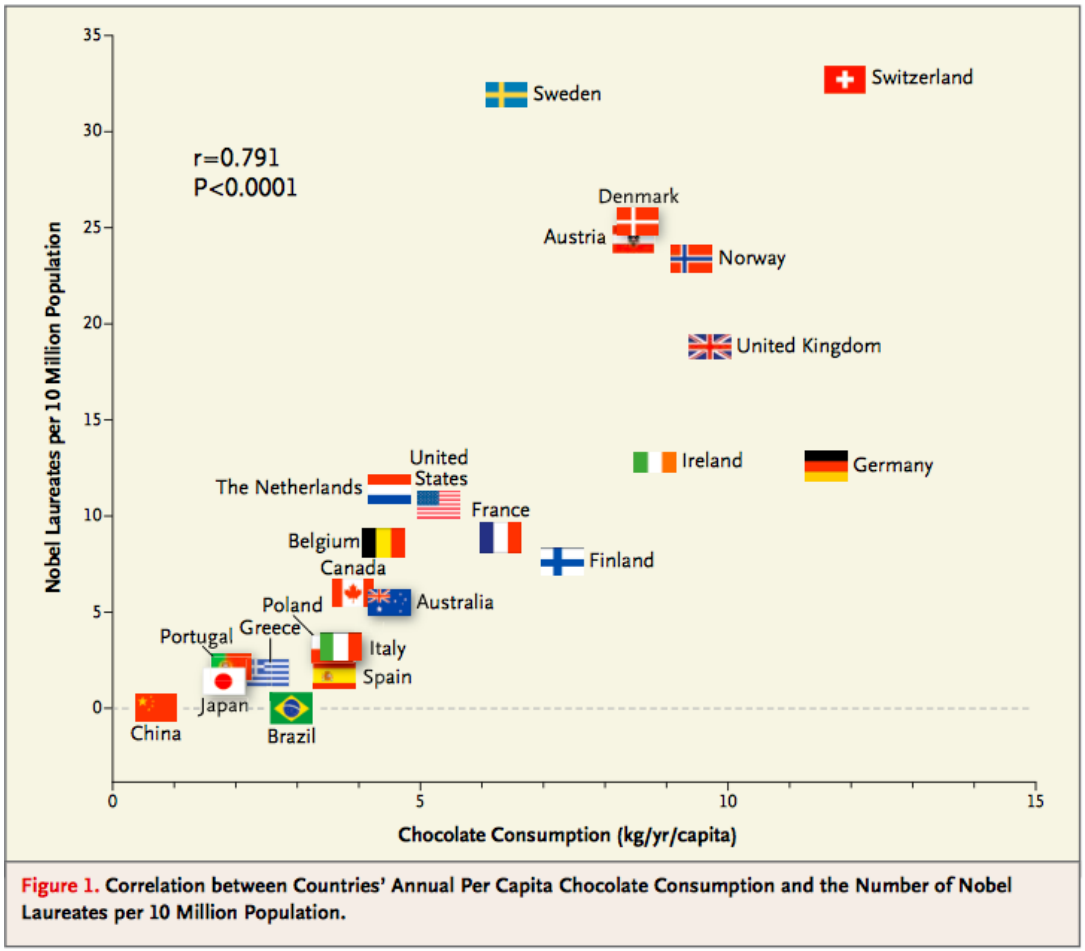


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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