Looking at Data

Rob McCulloch

9/21/2017

Looking at the House Price Data

Mean and Variance

Covariance and Correlation

Looking at the House Price Data

House Price Data

Here is some *data*:

```
##   Home Nbhd Offers SqFt Brick Bedrooms Bathrooms  Price
## 1    1    2      2 1790    No        2         2 114300
## 2    2    2      3 2030    No        4         2 114200
## 3    3    2      1 1740    No        3         2 114800
## 4    4    2      3 1980    No        3         2  94700
## 5    5    2      3 2130    No        3         3 119800
## 6    6    1      2 1780    No        3         2 114600
```

- each observation corresponds to a recently sold house.
- each row is an *observation*.
- each column corresponds to a *variable*, which reports something about the house.

There are actually 128 rows: 128 observations.

The **goal** is to predict the price knowing the other variables.

Note

Nbhd is a *categorical variable*:

```
##
##  1  2  3
## 44 45 39
```

45 of the 128 houses are in neighborhood 2.

SqFt is a *numeric variable*:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1450    1880    2000    2001    2140    2590
```

Qu is for *Quartile*. The first quartile is the 25% *percentile* or *quantile*.

k% of the values are less than the k percentile.

e.g. 25% of the values are < 1880.

What percentile is the median? What percentile is the third quartile?

If **price** is *related* to the other variables we could use the relationship to predict the price of a house that has not yet sold.

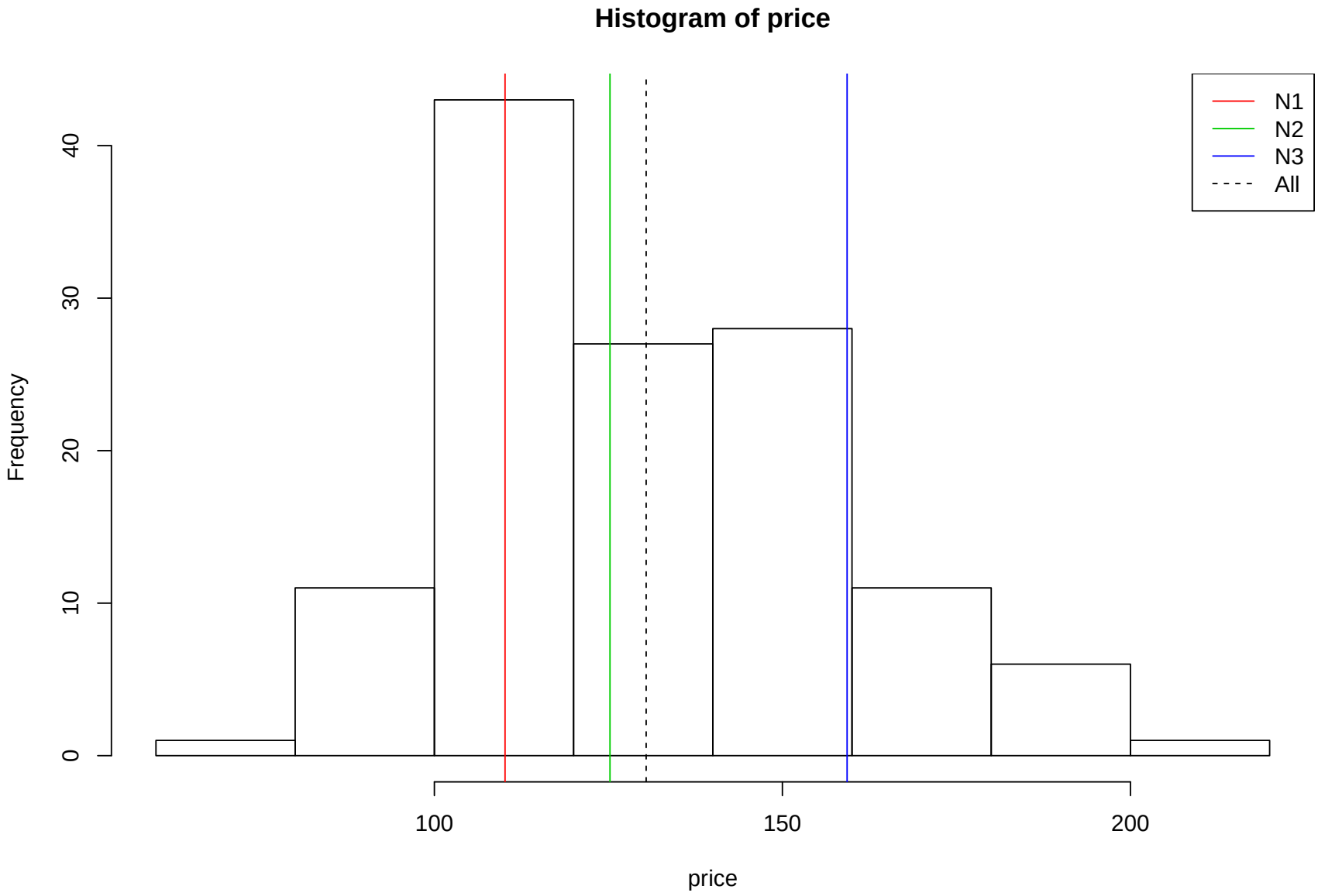
Is **price** related to **Nbhd** ??

Compute the average of the house price in each neighborhood.
I divided the prices by 1000 to make the numbers "nicer".

```
##    N1    N2    N3
## 110.2 125.2 159.3
```

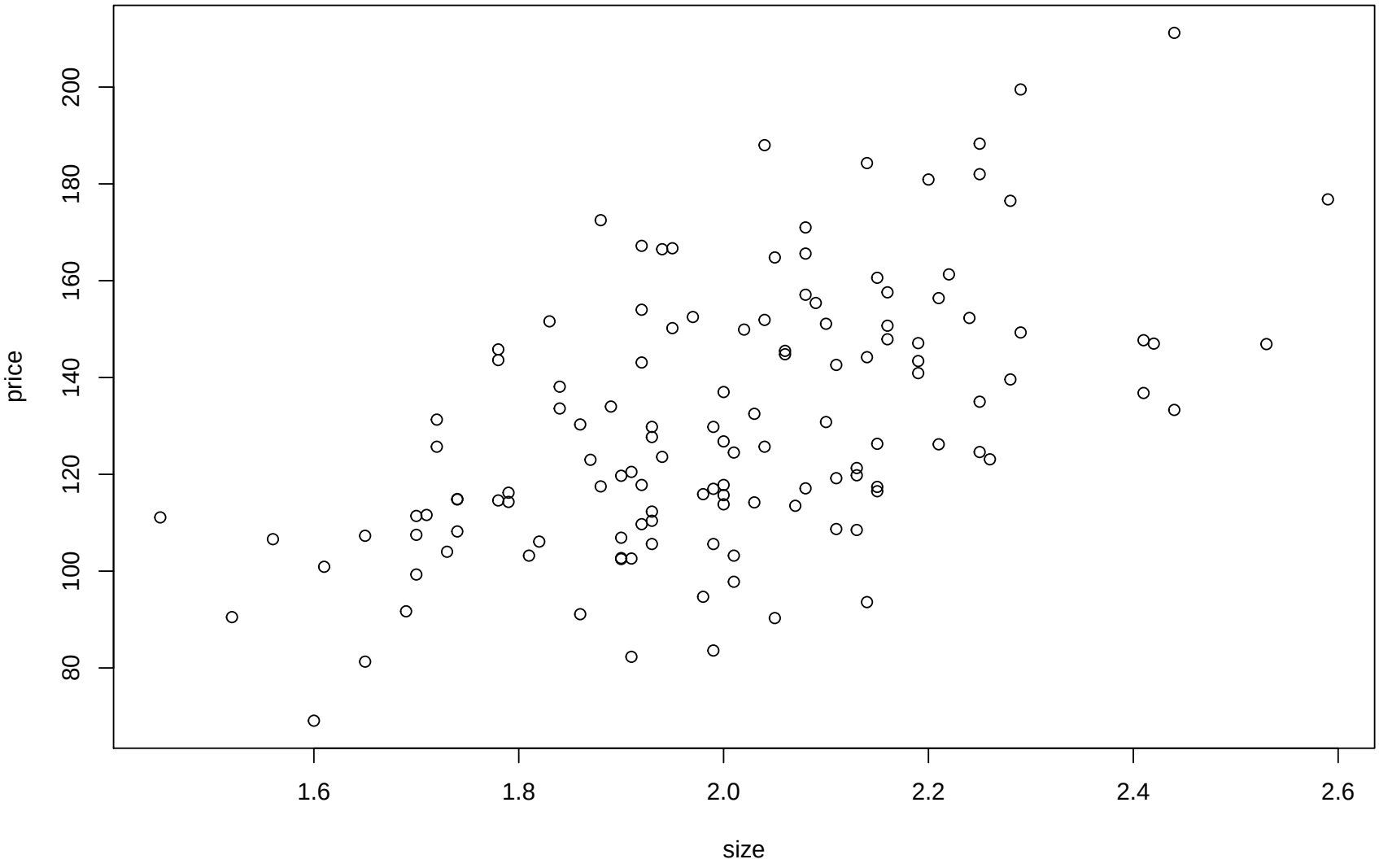
The average house price is 130.4, whilst the average of the houses in N3 is 159.3.

Histogram of `price`. The height of each bar tells us how many of the observations are in the interval.



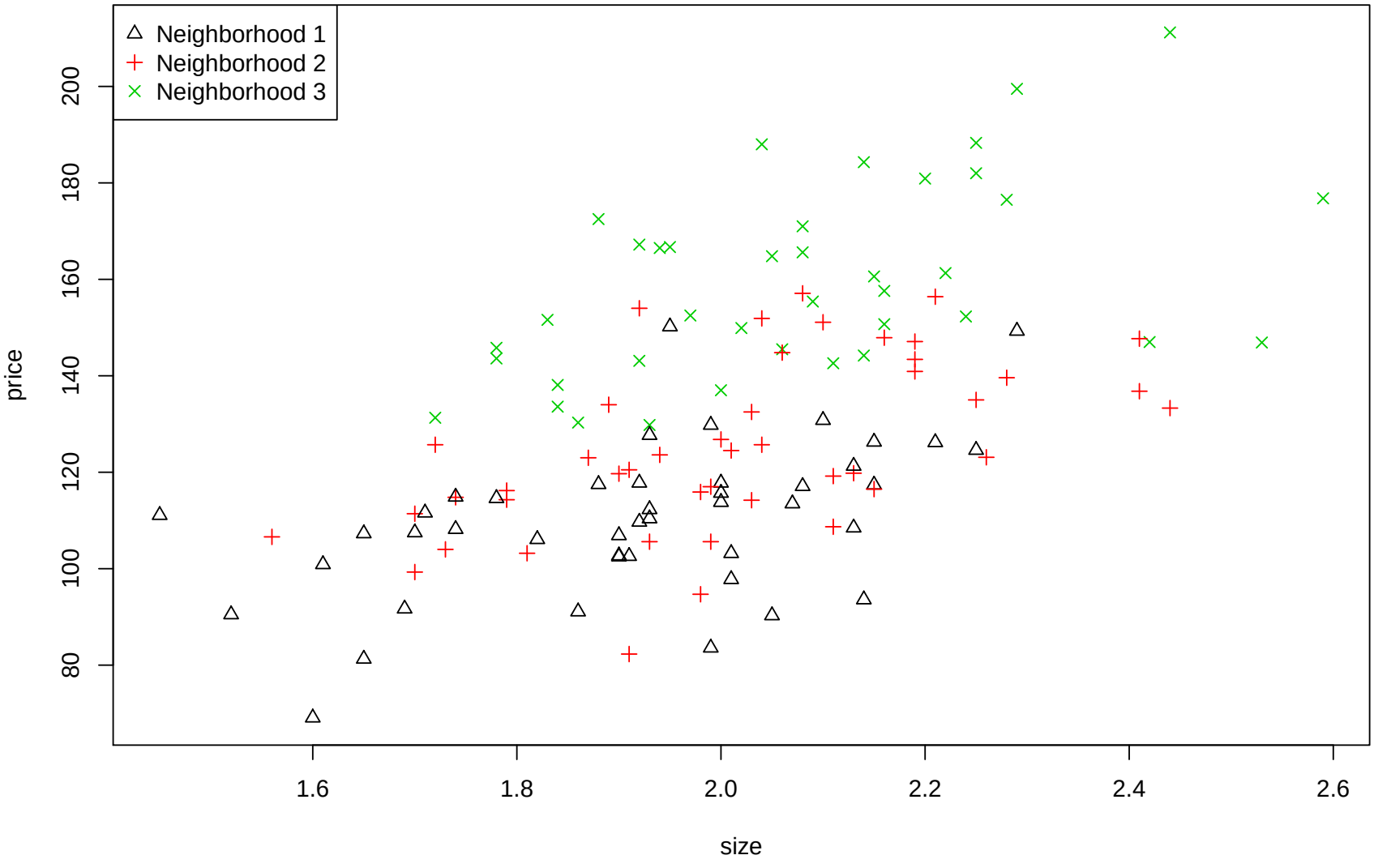
Vertical lines at the three neighborhood means and the overall mean ("All").

Is price related to size?



How would you predict the price of a house if you knew its size?

Is price related to size and neighborhood?



How would you predict the price of a house if you knew its size and neighborhood?

Mean and Variance

The Mean and Summation Reviewed

We summarized the `price` variable with its mean.

We compared house prices in different neighborhoods by comparing the average house price for each neighborhood.

Given numbers $y_1, y_2, \ldots, y_n$ the mean is

$$\bar{y} \;=\; \frac{y_1 + y_2 + \ldots + y_n}{n} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i$$

Recall: $\sum_{i=1}^{n} y_i$ means for each i from 1 to n add in y_i.

Example:

$$y_1 = 1, y_2 = 2, y_3 = 3, y_4 = 4, y_5 = 5$$

$$\sum_{i=1}^{n} y_i = 1 + 2 + 3 + 4 + 5$$

$$\sum_{i=1}^{n} (y_i - \bar{y}) = ???$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = ???$$

The Canadian and Japanese Returns

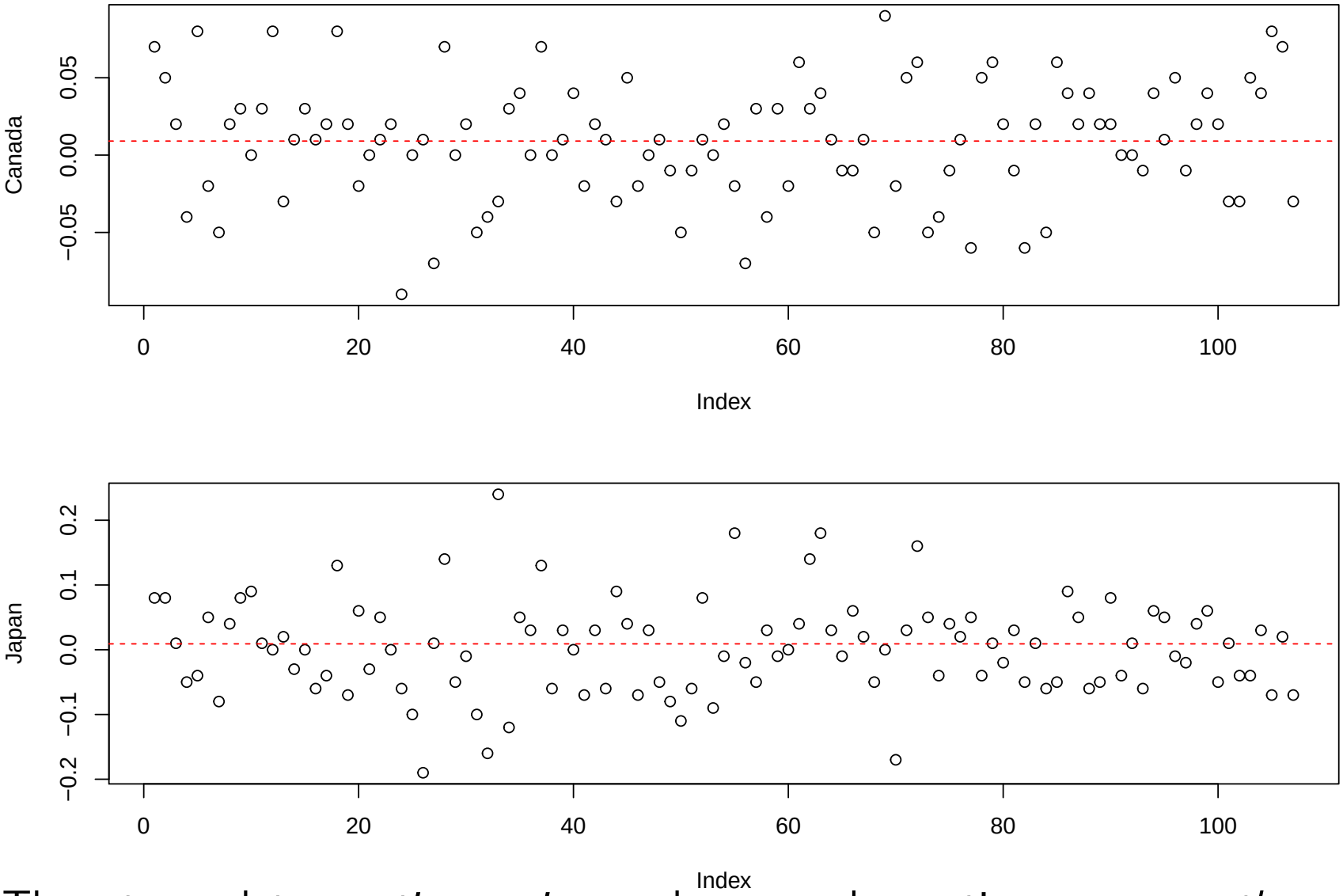
A *return* is the percentage increase your wealth goes up by if you invest over a given period.

So, if the return $r = .1$ then if you put put $W = 100$ in at the beginning of the period and cash out at the end of the period, you get

$$W(1 + r) = 100(1 + .1) = 110$$

at the end of the period.

Monthly returns on a portfolio of Canadian equities and a portfolio of Japanese equities over the same months.



The returns data are *time series*, we have an observation *every month*.
For time series data, obviously the order matters.
These are *time series plots*, the value on the y-axis and time on the x-axis.

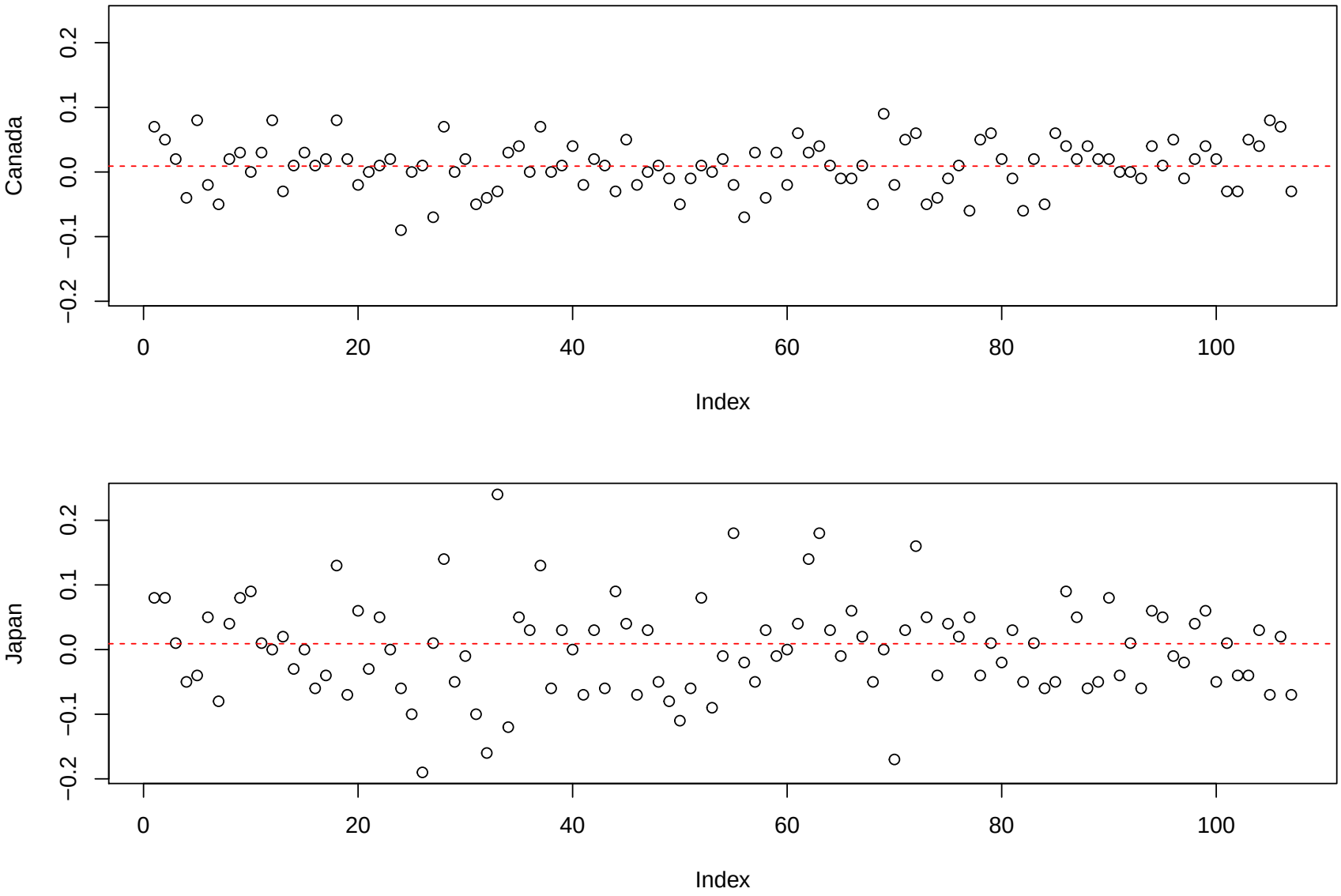
How do the Canada and Japan returns compare for these months?

Red lines drawn at the mean return.

Canadian mean return is 0.0090654.

Japanese mean return is 0.0023364.

Let's plot them on the same scale.


Canada
0.2
0.1
0.0
−0.1
−0.2
0
20
40
60
80
100
Index
Japan
0.2
0.1
0.0
−0.1
−0.2
0
20
40
60
80
100
Index

The Sample Variance and Standard Deviation

Sometimes we want to summarize how *variable* a bunch of numbers are.

The *sample variance* is the average squared distance to the mean:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

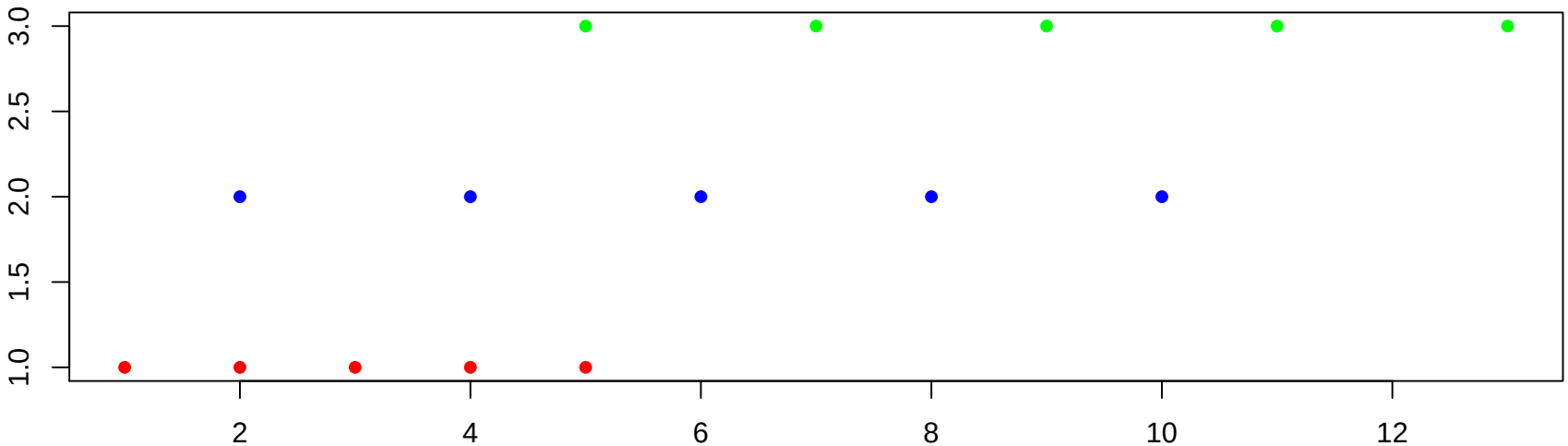
Just think of the $n-1$ as n for now, we'll explain later.

The *sample standard deviation* is the square root of the variance.

$$s_y = \sqrt{s_y^2}$$

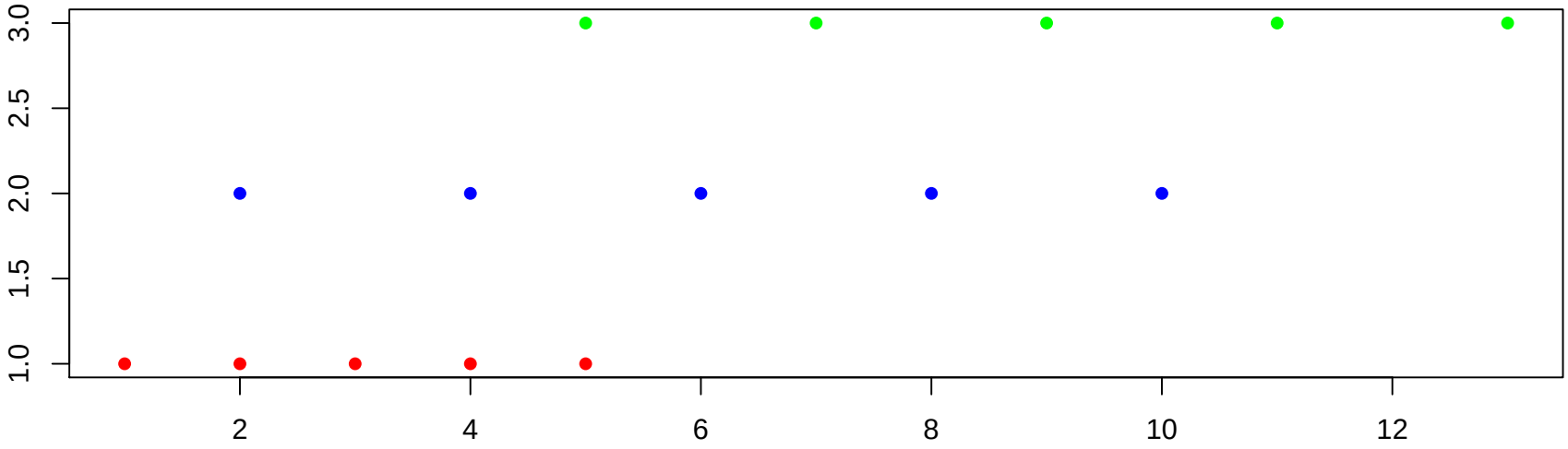
Example:

```
## x:
## [1] 1 2 3 4 5
## y:
## [1]  2  4  6  8 10
## z:
## [1]  5  7  9 11 13
```



$$
\begin{aligned}
s_x^2 &= ((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2)/4 \\
&= (2^2 + 1^2 + 0^2 + 1^2 + 2^2)/4 \\
&= 10/4 = 2.5
\end{aligned}
$$

$s_x = \sqrt{2.5} = 1.5811388.$



$$s_y^2 = (4^2 + 2^2 + 0^2 + 2^2 + 4^2)/4$$
$$= 40/4 = 10$$

$s_y = \sqrt{10} = 3.16.$

y is more *spread out* than x, so it has a bigger standard deviation!!!

What is s_z?

The mean and standard deviation of the Canada returns are:

Mean: 0.0090654
SD: 0.0383266

The mean and standard deviation of the Japan returns are:

Mean: 0.0023364
SD: 0.0736844

The Japan returns have a lower mean *and* they are more spread out.

Interpreting the Standard Deviation

The mean is easy to interpret.
E.g. "the average price is 130.43 (thousands of dollars)".

The standard deviation and variance are trickier.

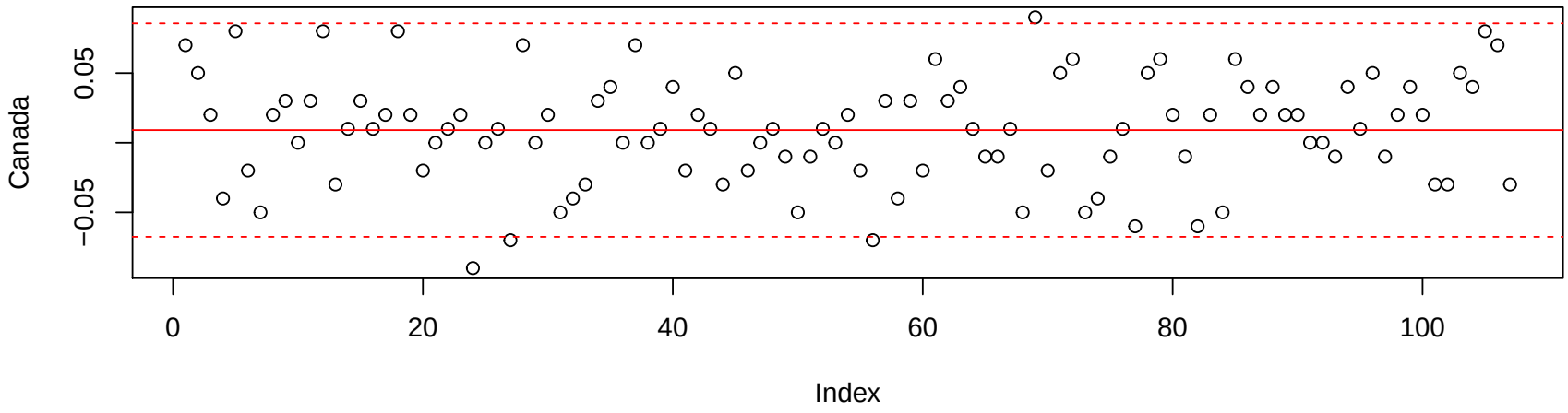
If y is in feet, what are the units of s_y^2?

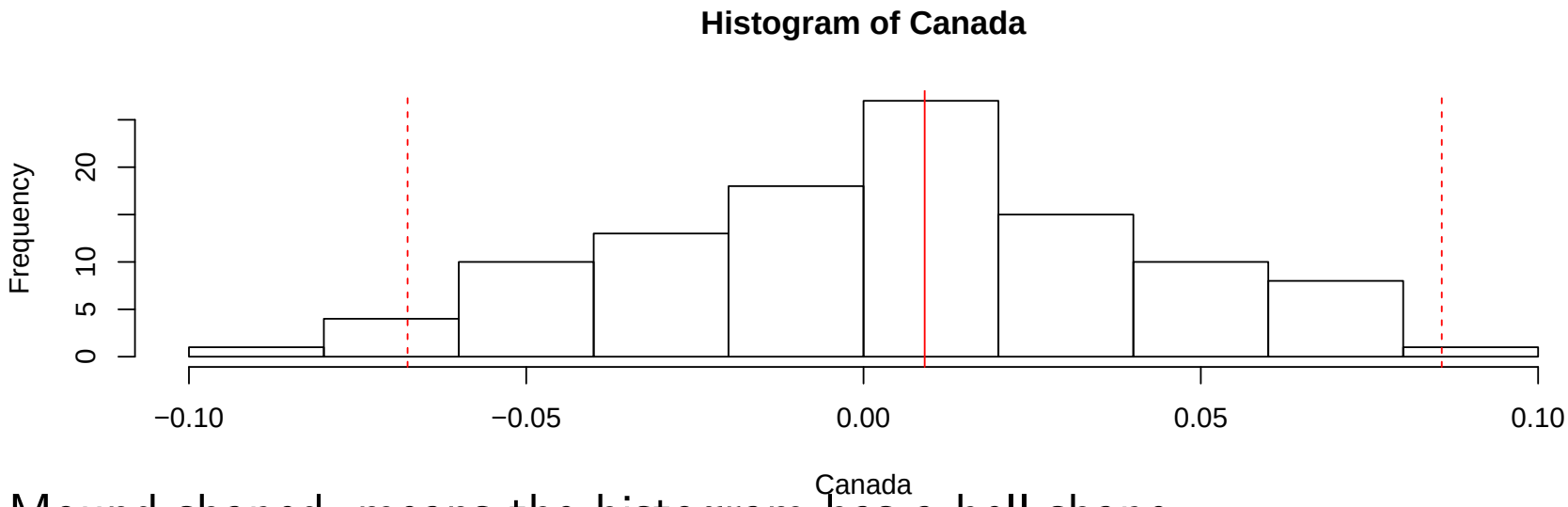
If y is in feet, what are the units of s_y?

In general, the standard deviation is easier to interpret because at least the units are comprehensible.

For "mound shaped data" one way to think about the standard deviation is

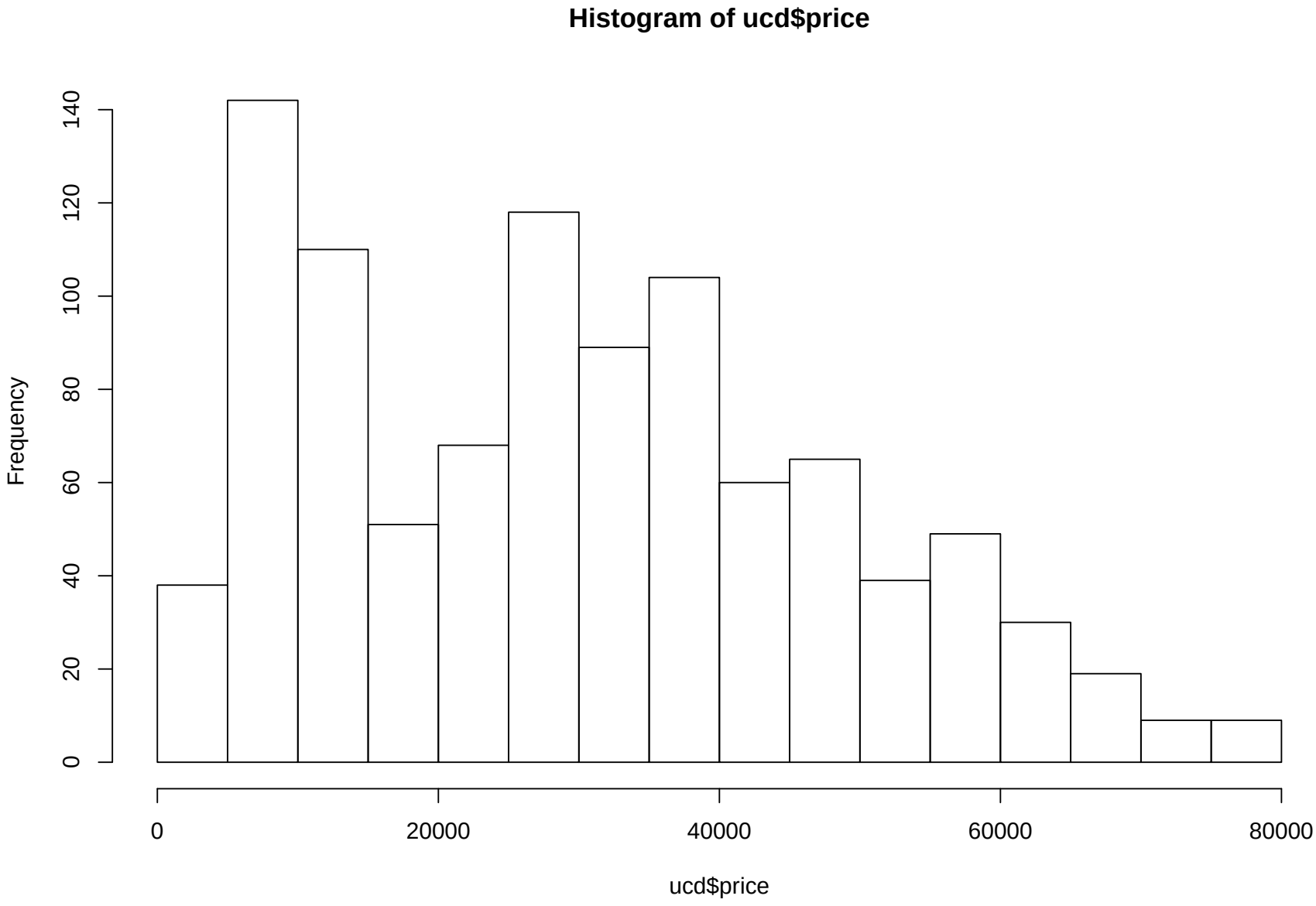
- about 95% of the data is in the interval $\bar{y} \pm 2s_y$.
- about 68% of the data is in the interval $\bar{y} \pm s_y$.





Mound shaped, means the histogram has a bell shape.
Later, we will relate this to the normal distribution.

Prices of used cars.



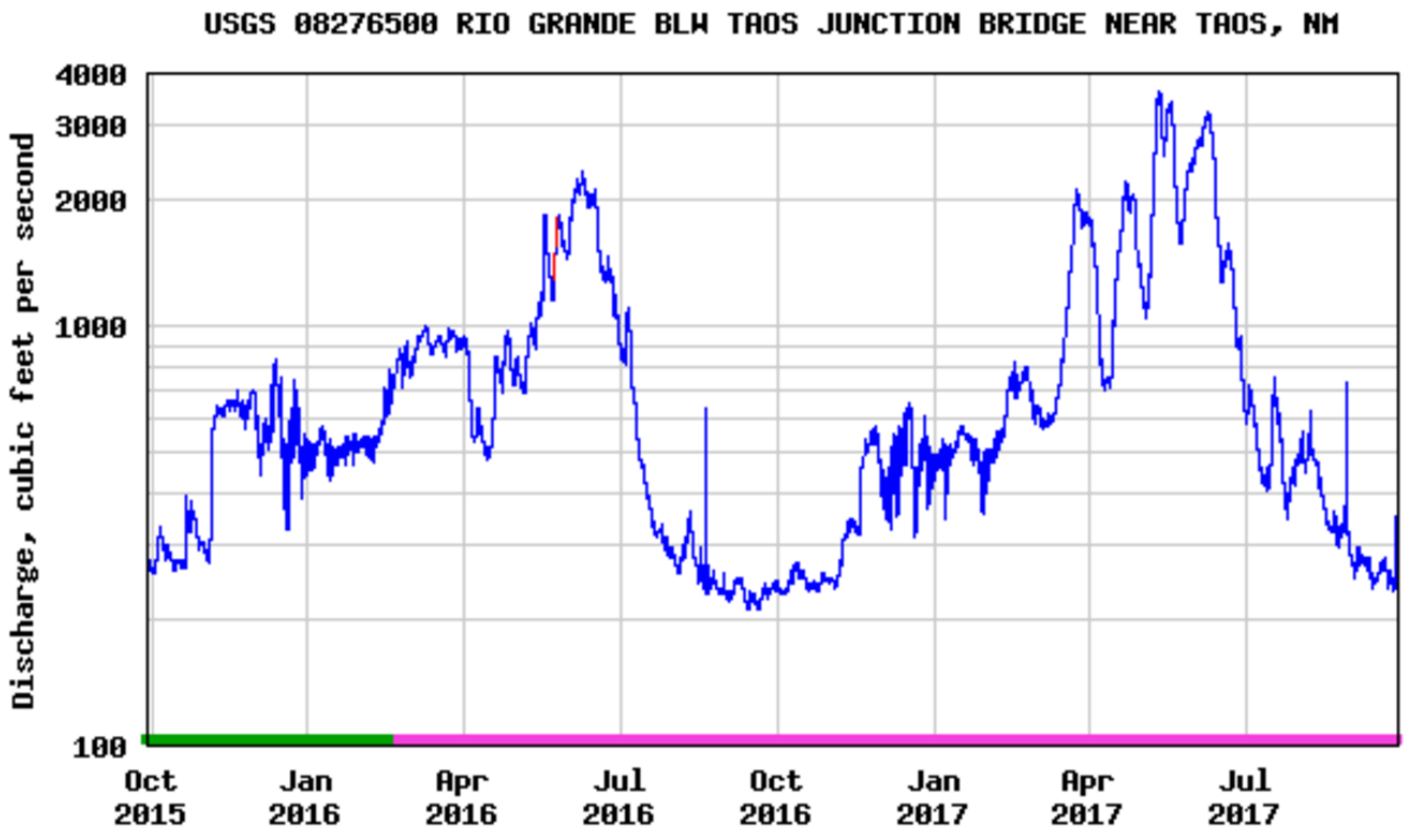
not mound shaped, skewed right !!

Note:

The time series plots of the Canada and Japan returns did not exhibit a temporal pattern.

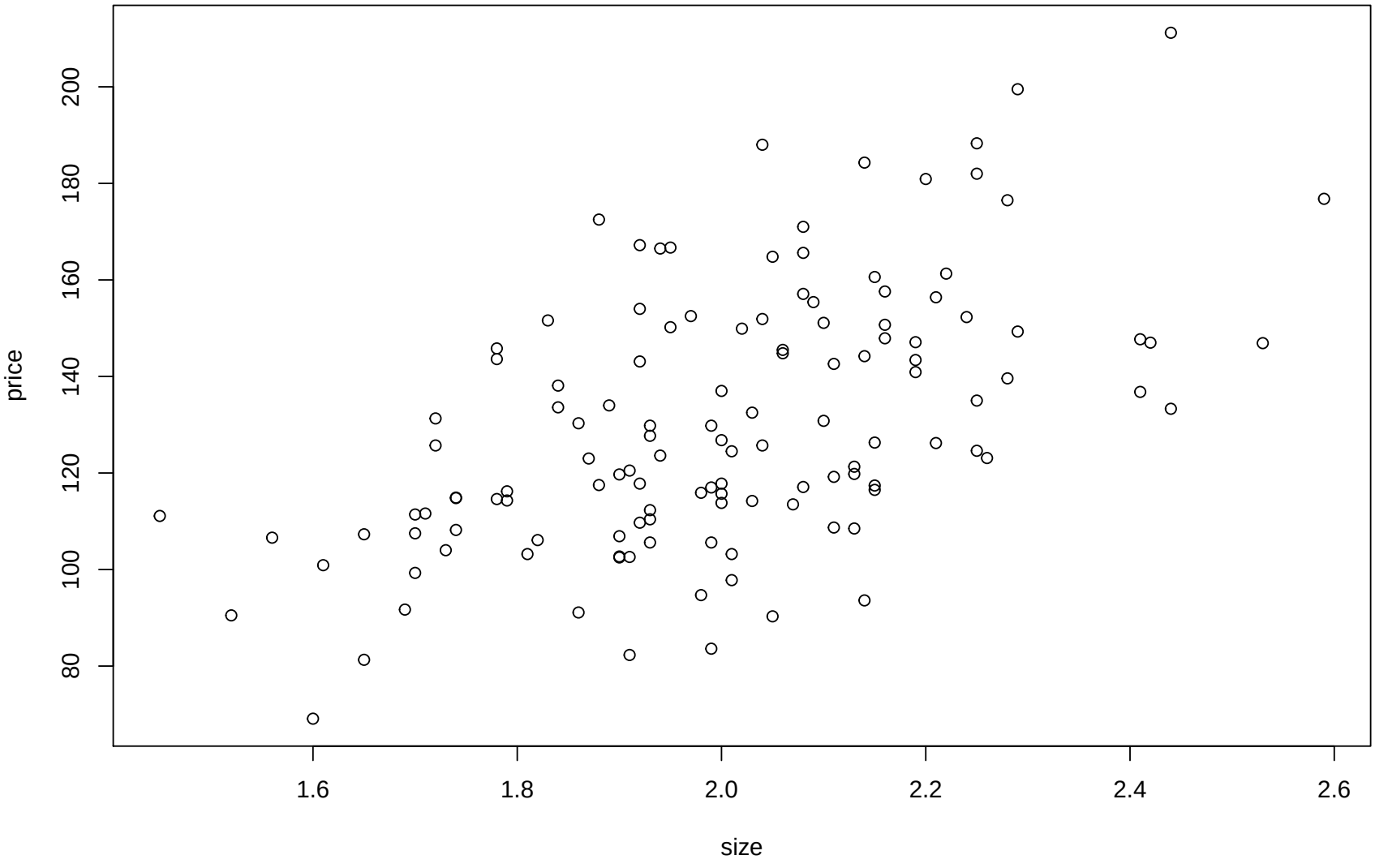
This is an interesting feature of (some) returns data!!

Here is a time series with a clear temporal pattern, daily stream flow numbers for the Rio Grande near Taos, New Mexico.



Covariance and Correlation

Back to the housing data....



From the plot, we see a *linear relationship* between the two variables.

The covariance and correlation are used to summarize the strength of the linear relationship between two numeric variables.

The *sample covariance* between x and y is

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

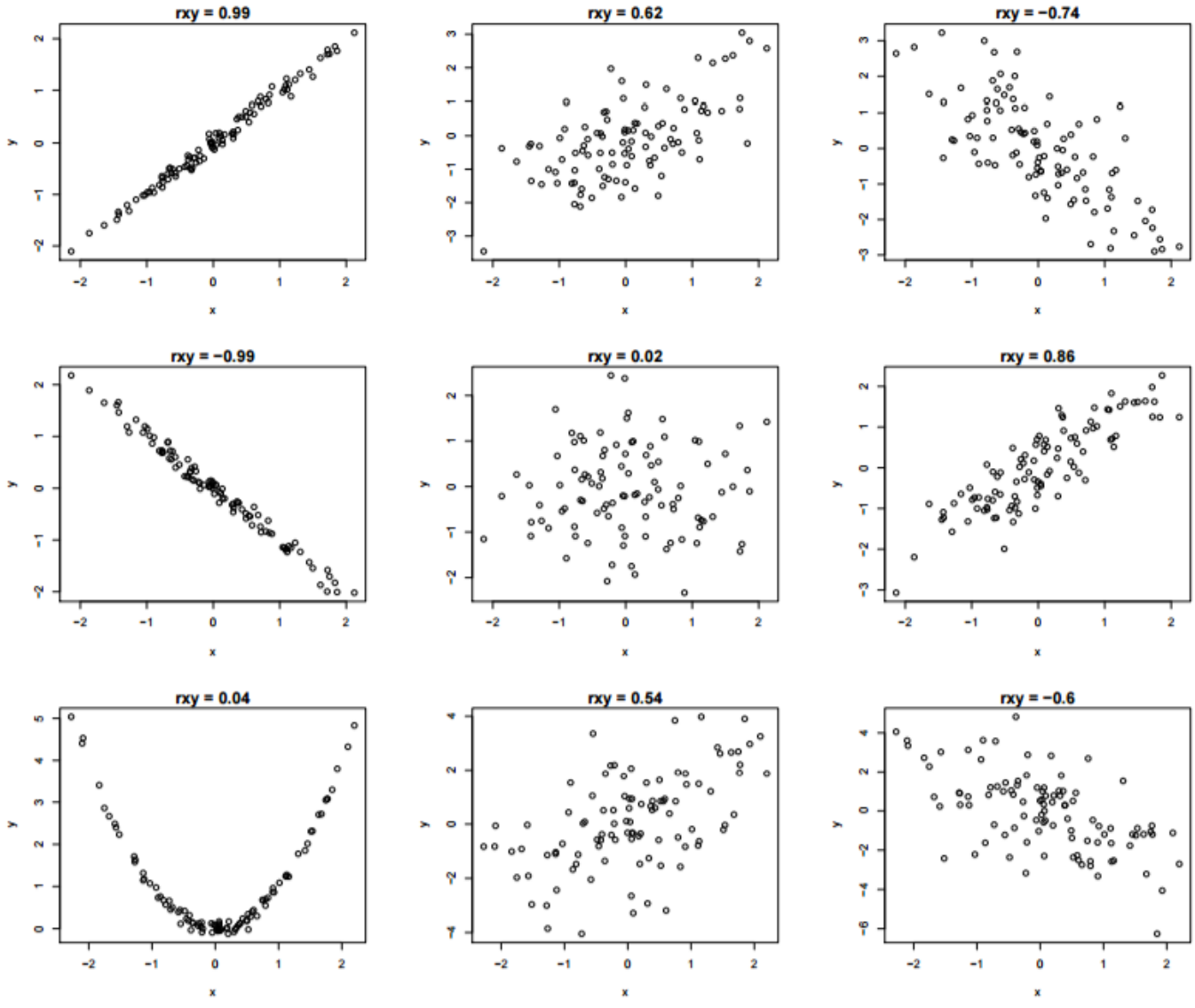
The *sample correlation* between x and y is

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y}$$

The sample correlation:

- $-1 \leq r_{xy} \leq 1$.
- r_{xy} close to 1 means there is a strong linear relationship, with a positive slope.
- r_{xy} close to -1 means there is a strong linear relationship, with a negative slope.
- r_{xy} close to 0 means no linear relationship.

The correlation answers the question *do you see a line* .



Note:

What are the units of the covariance (in terms of the units of x and y) ?

What are the units of the correlation (in terms of the units of x and y) ?

Think of the correlation as a version of the covariance *scaled* so that we it is more interpretable.

Since $-1 \leq r_{xy} \leq 1$, we have some sense of what a "big one" is.

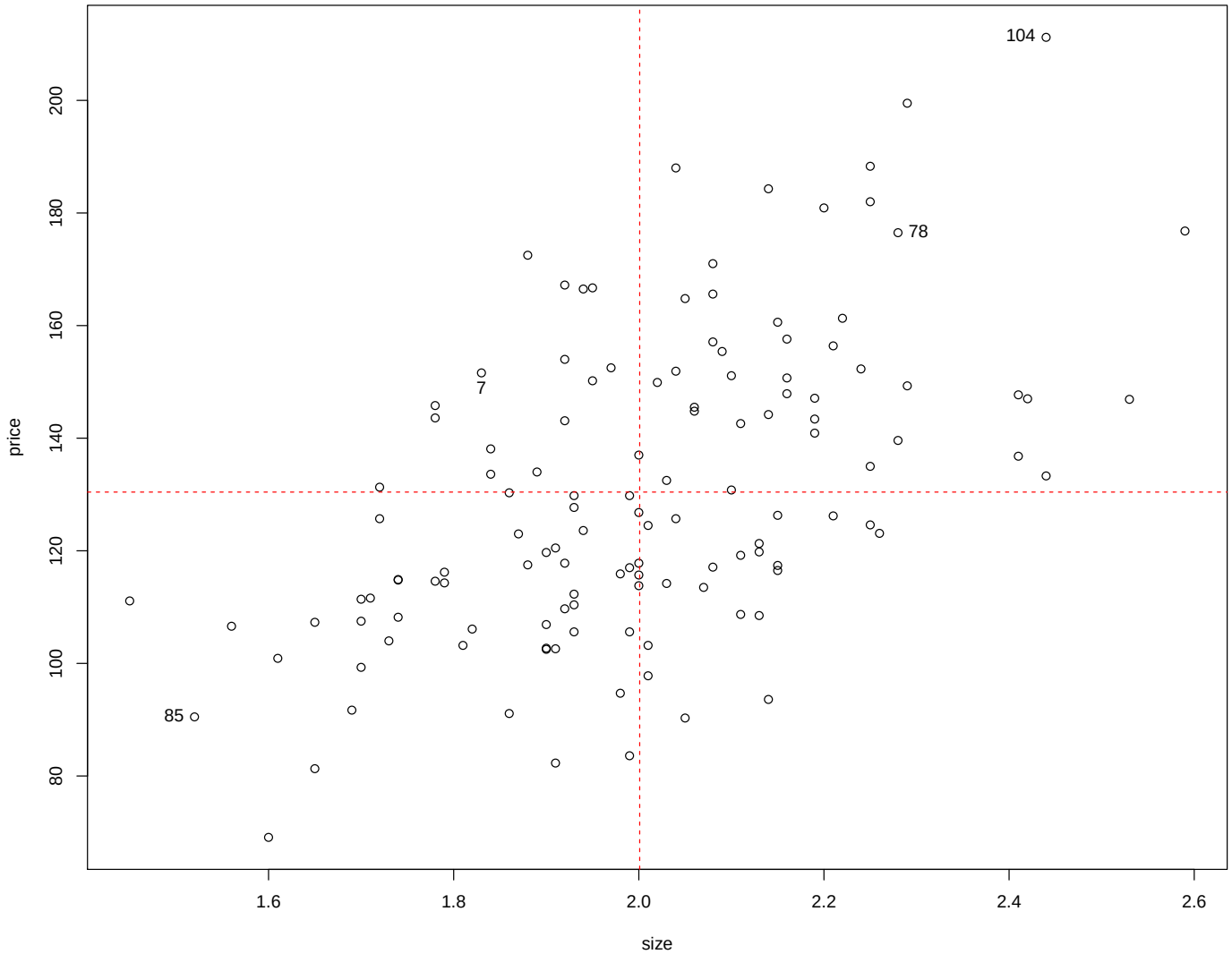
We have a similar situation with s_y^2 and s_y, obviously, they are about the same thing, but s_y tends to be the more interpretable version.

Why does this wacky formula work?

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

Well, of course, it is the inner product of the demeaned vectors, but . . . .

To get a feeling for how things work let see what happens at the observations: 104, 78, 85, and 7.



Dashed red lines at $\bar{x}$ and $\bar{y}$.

Here are the four points.

```
##     size price
## 104 2.44 211.2
## 78  2.28 176.5
## 85  1.52  90.5
## 7   1.83 151.6
```

The mean of size is 2.0009375 and the mean of price is 130.4273438.

Here are the four points with the means subtracted off.

```
##           size     price
## 104  0.4390625  80.77266
## 78   0.2790625  46.07266
## 85  -0.4809375 -39.92734
## 7   -0.1709375  21.17266
```

e.g.
2.44-2.000938 = 0.439062

211.2-130.4273 = 80.7727

Here are the four products $(x_i - \bar{x})(y_i - \bar{y})$:

```
## [1] 35.464244 12.857151 19.202557 -3.619201
```

e.g.

0.4390625* 80.77266 = 35.46425

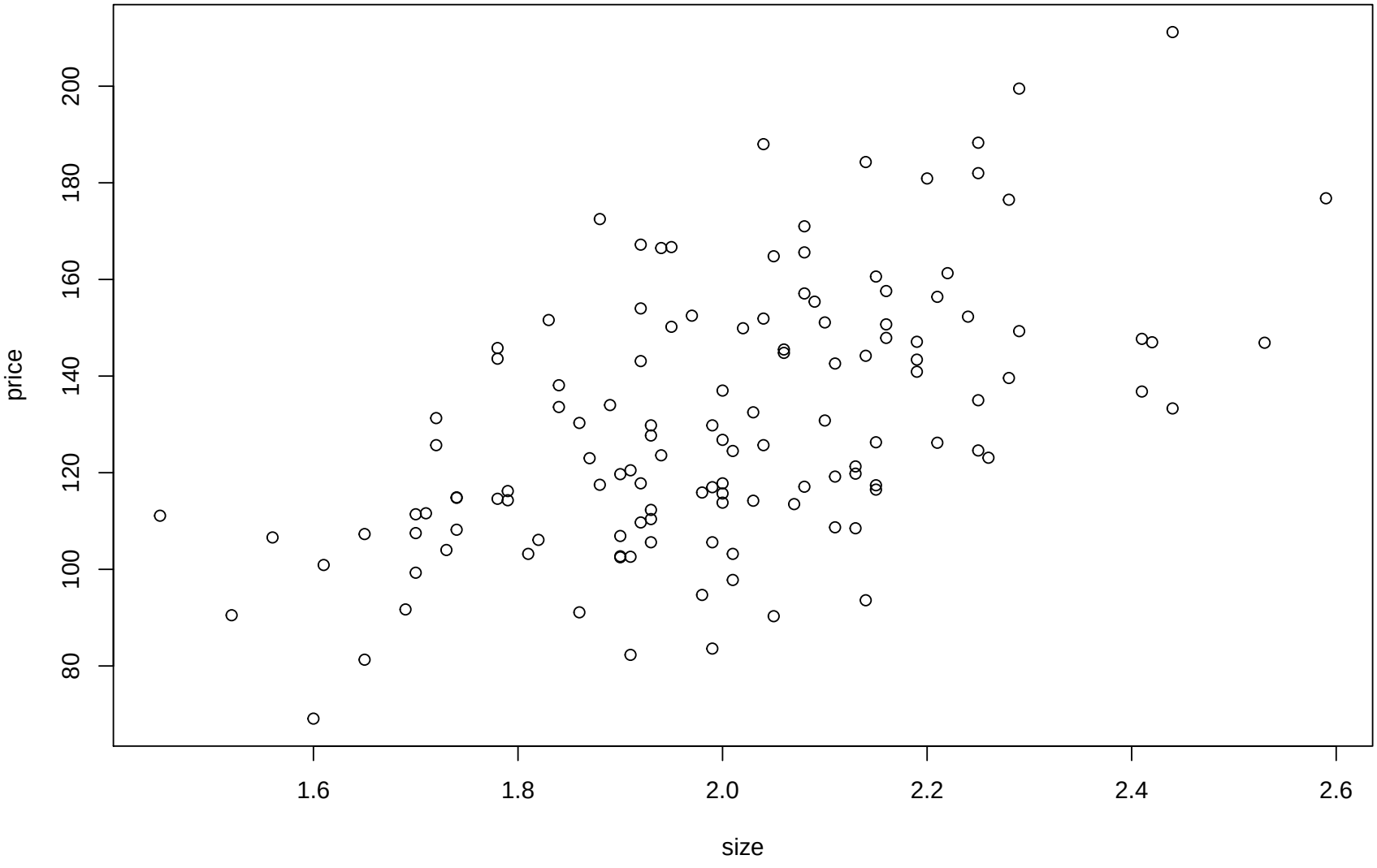
$(x_i - \bar{x})(y_i - \bar{y})$:

So, if both x and x are above (or below) the mean, you get a positive contribution.

If one is up and the other is down, you get a negative contribution.

The further out you are from $(\bar{x}, \bar{y})$, the larger (in absolute value) the contribution is.

So, you get an overall summary of how much they move together.



Summarize:

The mean and standard deviation of price is 130.43 and 26.87.

The mean and standard deviation of size is 2 and 0.21.

The correlation between size and price is 0.55.