Statistics with doodles

Thomas Levine
thomaslevine.com

Why we have statistics

Lots of numbers

95	16	74	56	83	49	34	85	38	54	91	37	65	18	90	91	96	98	16	81	14
81	14	57	99	1	14	94	52	45	86	41	24	92	11	67	41	22	55	94	7	2
7	2	58	77	5	42	10	50	49	76	70	82	19	17	65	2	31	53	85	80	26
80	26	65	9	41	7	10	22	60	10	98	52	78	8	32	45	42	19	80	93	10
93	10	37	50	5	47	37	36	82	79	20	42	16	59	36	89	42	6	41	47	83
47	83	57	59	3	23	18	7	75	75	94	14	46	72	79	83	73	45	23	12	15
12	15	38	69	90	13	83	16	22	24	47	98	77	49	44	55	79	69	69	65	70
65	70	13	35	35	32	79	25	33	10	50	86	51	55	43	86	55	63	23	33	24
33	24	65	31	68	95	17	35	67	99	95	2	63	90	36	20	44	45	51	81	64
81	64	6	60	93	8	40	98	23	39	19	22	62	28	29	63	78	8	70	80	67
80	67	0	3	69	33	93	85	16	96	63	76	25	3	31	33	33	24	42	54	33
54	33	4	92	35	89	36	76	80	24	76	39	53	74	93	56	36	22	33	24	57
24	57	24	76	61	58	44	32	17	17	29	46	98	74	8	27	71	90	0	66	83
66	83	46	76	27	15	37	90	19	68	89	96	75	66	54	7	43	32	72	22	59
22	59	42	78	52	32	79	17	56	74	89	72	94	19	88	83	0	41	14	83	63
83	63	41	28	31	32	70	7	24	2	1	76	26	30	69	36	11	88	51	67	66
67	66	80	28	49	51	35	36	40	0	59	21	58	47	69	40	54	77	12	40	48
40	48	28	66	53	65	8	40	78	52	33	62	61	78	76	15	78	88	19	12	45
12	45	24	8	13	94	68	62	61	54	85	91	47	22	32	51	65	79	91	12	48
12	48	54	35	73	100	38	23	82	79	67	53	8	80	69	44	31	25	23	40	79
40	79	40	53	99	89	35	11	17	76	2	23	54	99	78	78	96	6	17	59	95

It's hard to fit lots of numbers into our brains all at once.

69	2	58	55	71	37	11	45	15	85	50	79	3	78	56	42	70	78	22	0	52
0	52	62	85	16	40	84	42	69	88	83	16	85	35	89	75	15	64	14	31	60
31	60	97	5	0	25	90	50	40	37	66	48	50	30	100	4	14	21	67	67	12
67	12	77	61	10	62	80	7	54	39	79	67	41	74	46	95	83	64	34	36	70
36	70	66	8	72	16	60	25	3	14	18	36	34	94	34	91	54	99	12	11	4
11	4	78	93	11	79	68	70	71	64	40	98	1	61	15	90	16	87	26	43	62
43	62	10	57	89	28	78	79	72	5	78	53	74	25	66	84	84	67	19	19	19
19	19	14	16	64	36	61	22	64	85	12	58	33	64	73	18	89	94	74	80	7
80	7	32	4	67	45	94	26	17	65	10	41	32	13	41	4	51	88	19	9	96
9	96	79	34	91	76	63	71	24	44	41	64	37	72	65	79	44	8	7	58	23
58	23	41	54	10	74	43	87	37	73	32	67	56	68	100	37	33	50	87	29	97
29	97	43	19	41	97	66	12	28	16	74	74	93	39	83	68	39	52	81	58	68
58	68	79	57	23	65	77	96	12	50	73	68	58	73	98	8	96	46	34	79	57
79	57	66	89	75	25	48	58	79	95	1	84	39	60	43	79	1	60	12	32	32
32	32	57	34	40	80	44	16	81	24	28	21	4	21	6	40	45	29	52	29	57
29	57	76	35	27	70	74	88	52	59	68	37	87	63	54	49	67	49	26	53	58
53	58	65	82	70	42	31	90	64	31	24	56	49	83	71	88	45	98	40	89	74
89	74	4	76	70	80	32	47	86	18	7	28	49	18	65	31	17	97	30	39	92
39	92	77	52	21	36	12	47	29	12	73	42	81	8	96	90	63	29	5	85	45
85	45	74	55	66	22	84	5	31	25	94	39	86	12	94	87	96	11	72	100	97
100	97	55	27	10	96	98	90	31	53	51	86	19	41	91	17	45	81	24	51	69

So we invent numbers
that describe
lots of other numbers

So we invent numbers
that describe
lots of other numbers

(statistics)

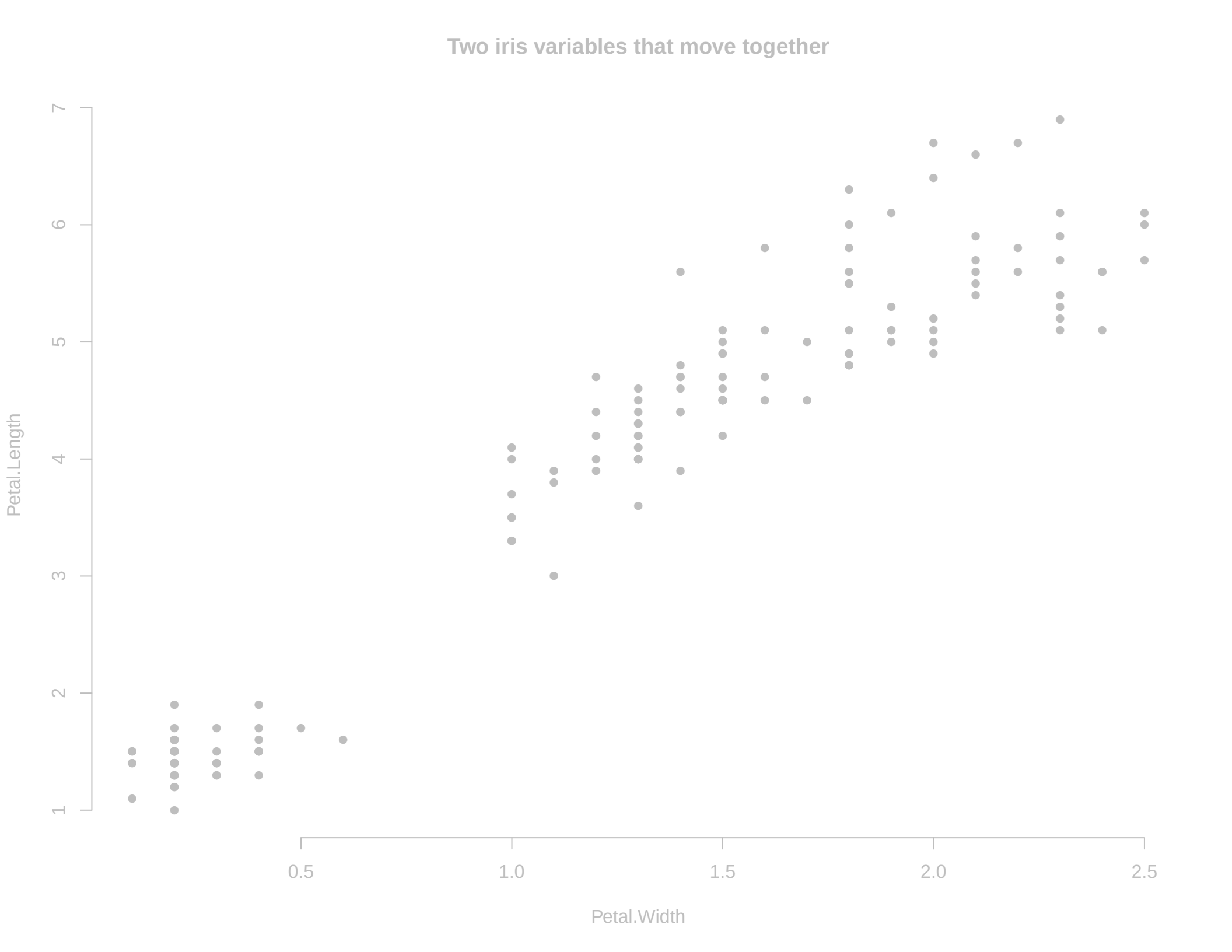
Here are some numbers:
1 2.2 pi 4 5 7 7

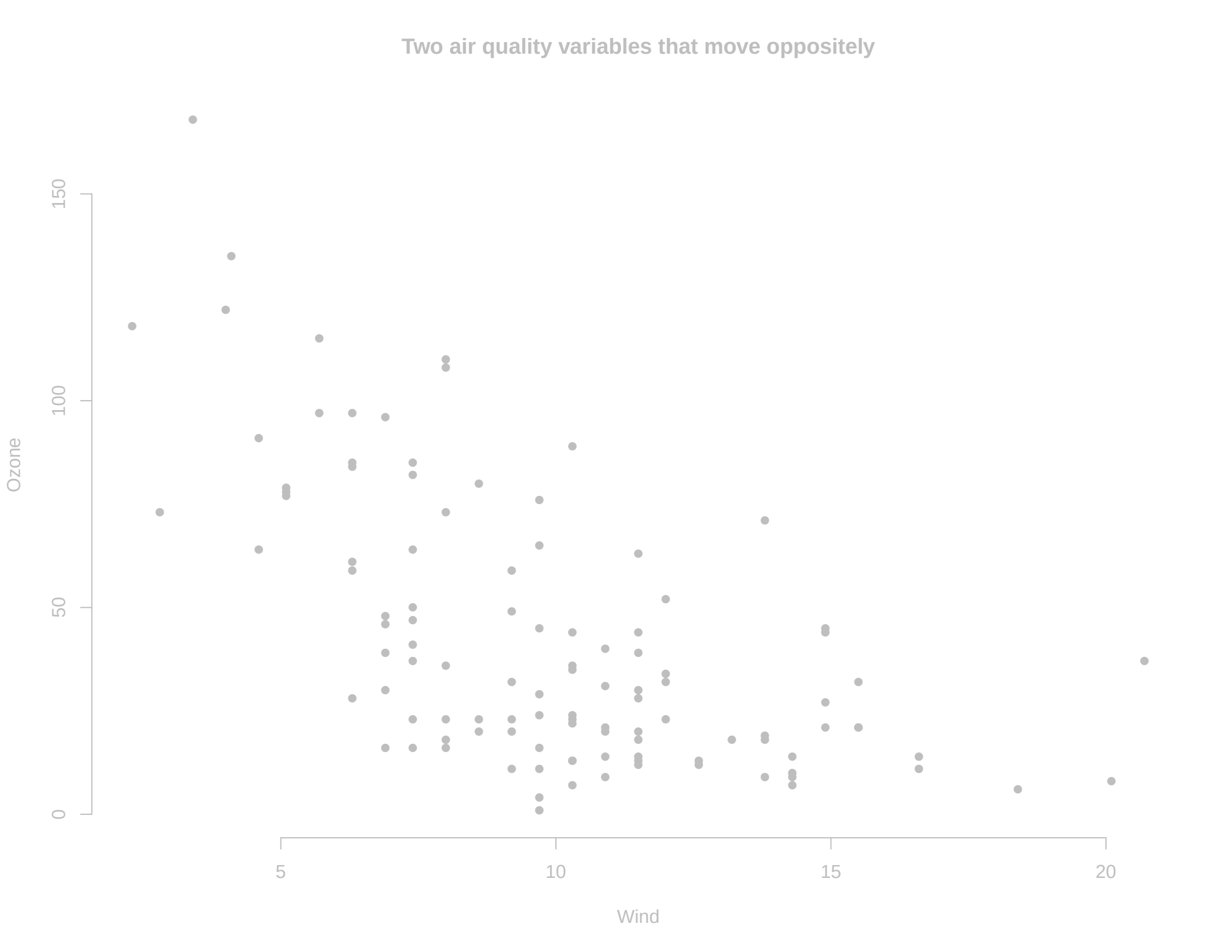
What are some statistics?

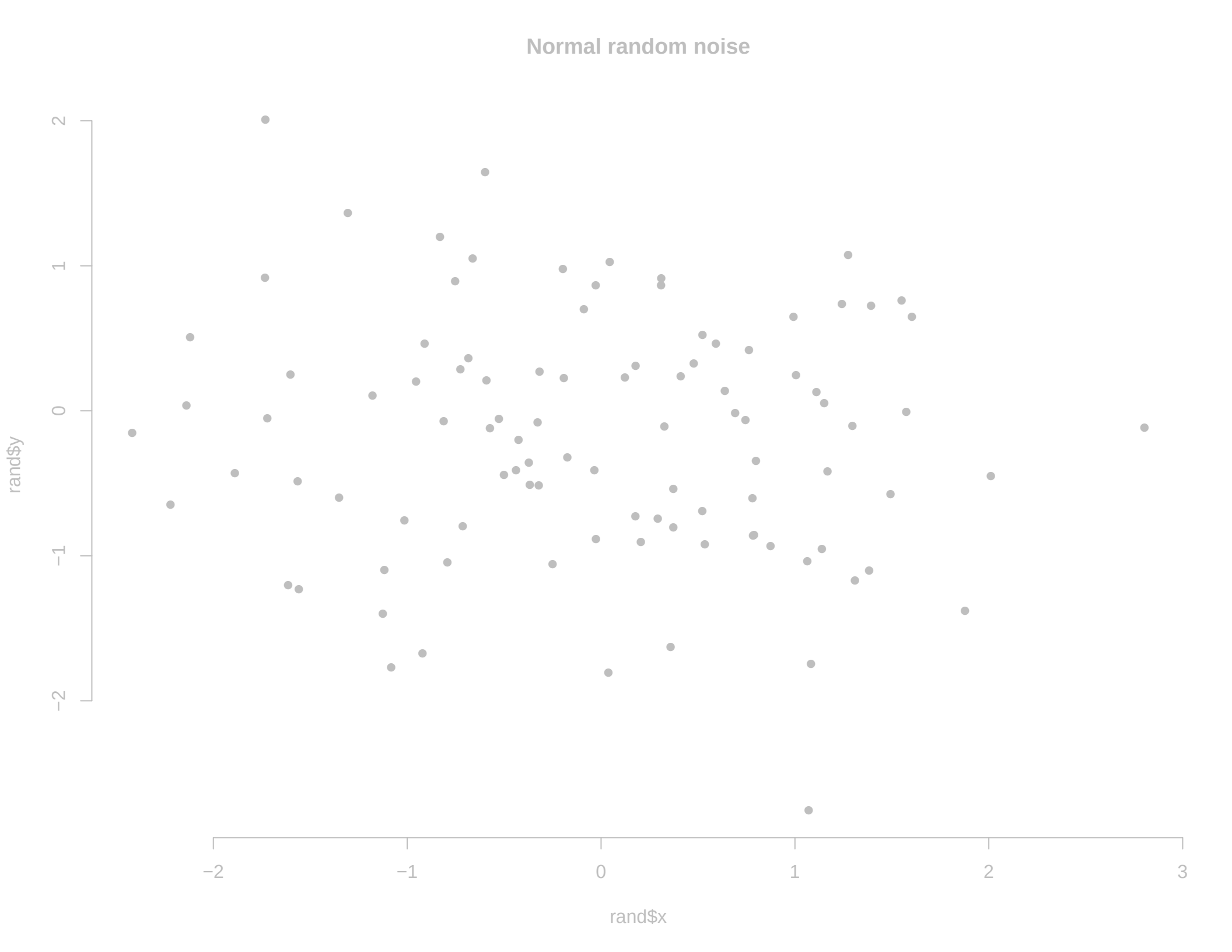
min, max,
mode, median, mean,
range, variance

how many integers,
whether the numbers are sorted
&c.

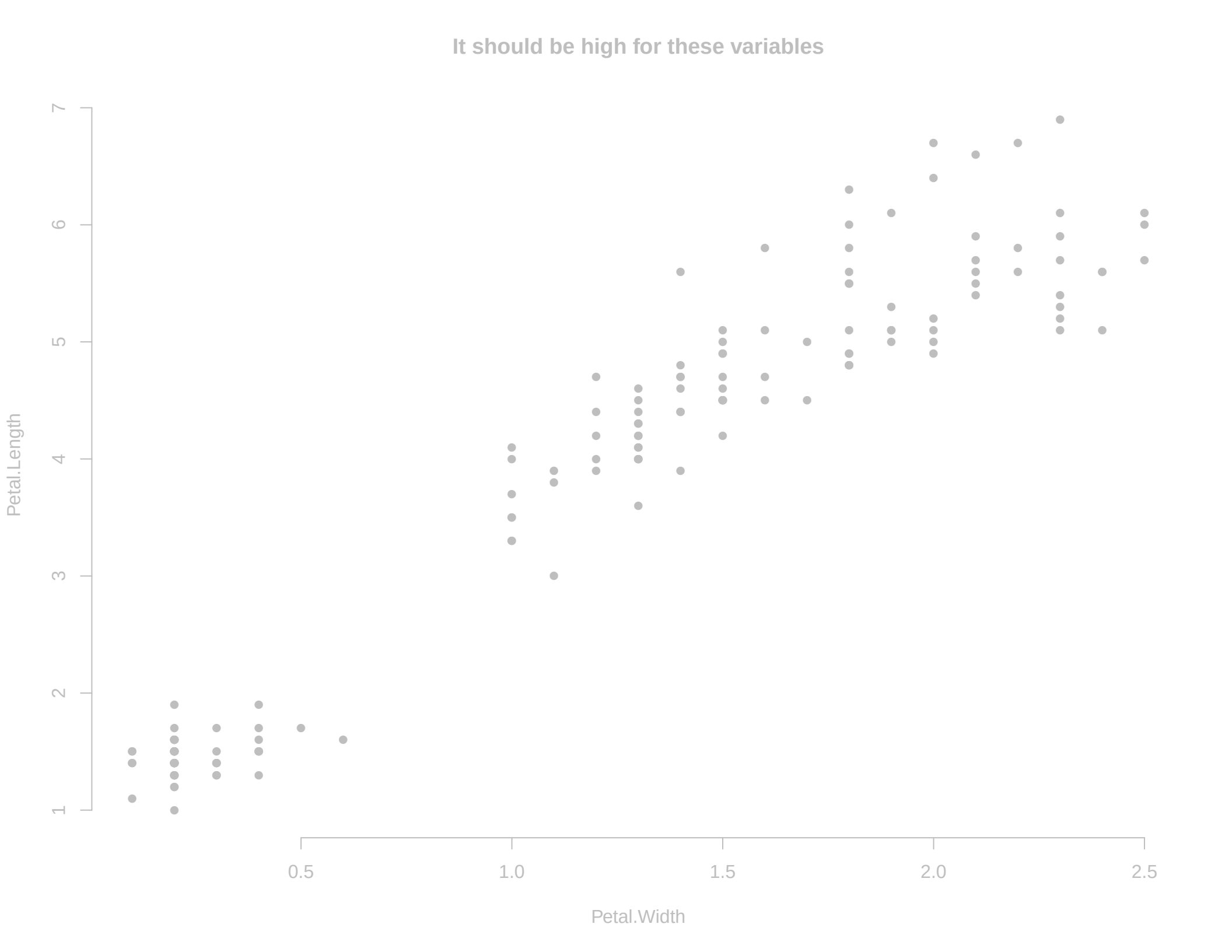
Measuring linear relationships

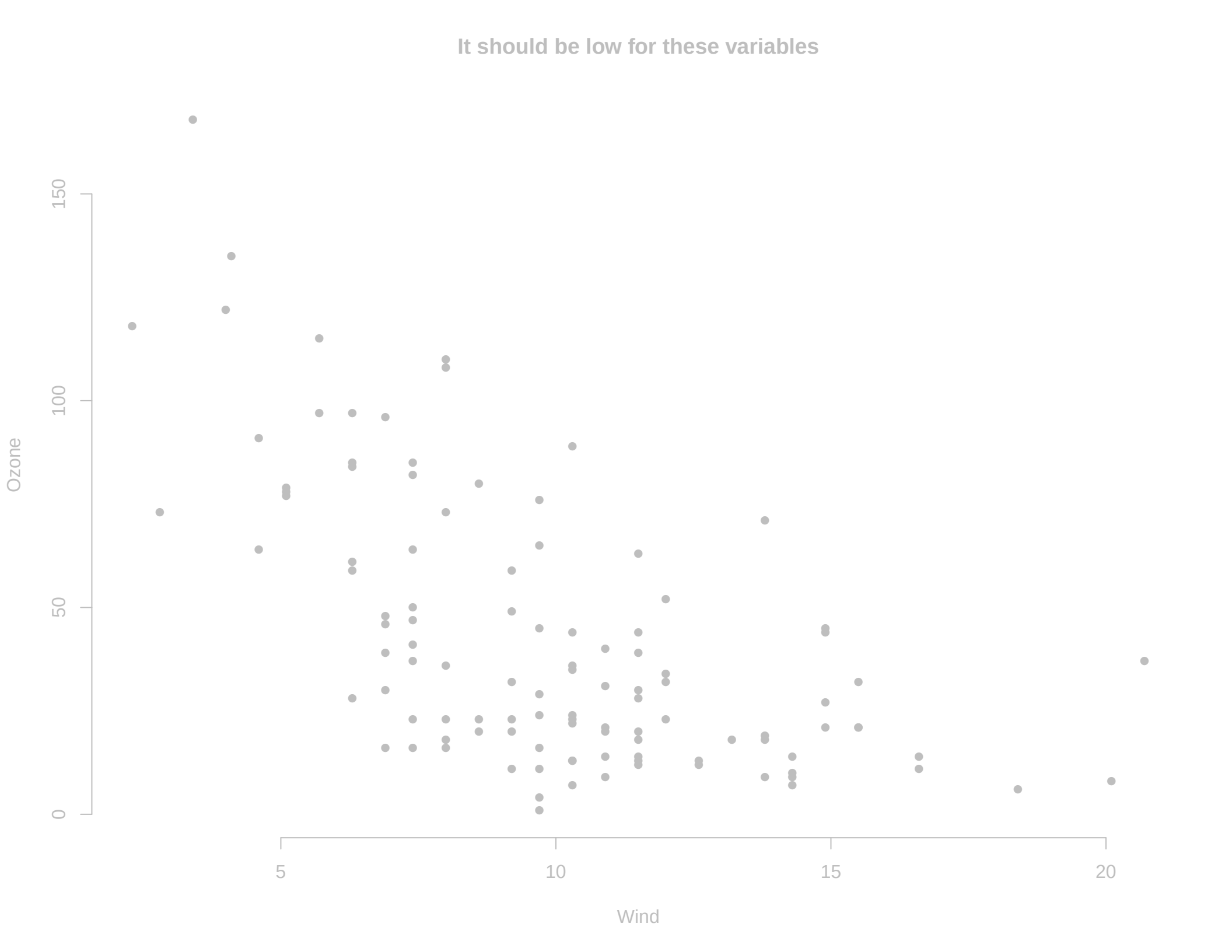

Two iris variables that move together
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

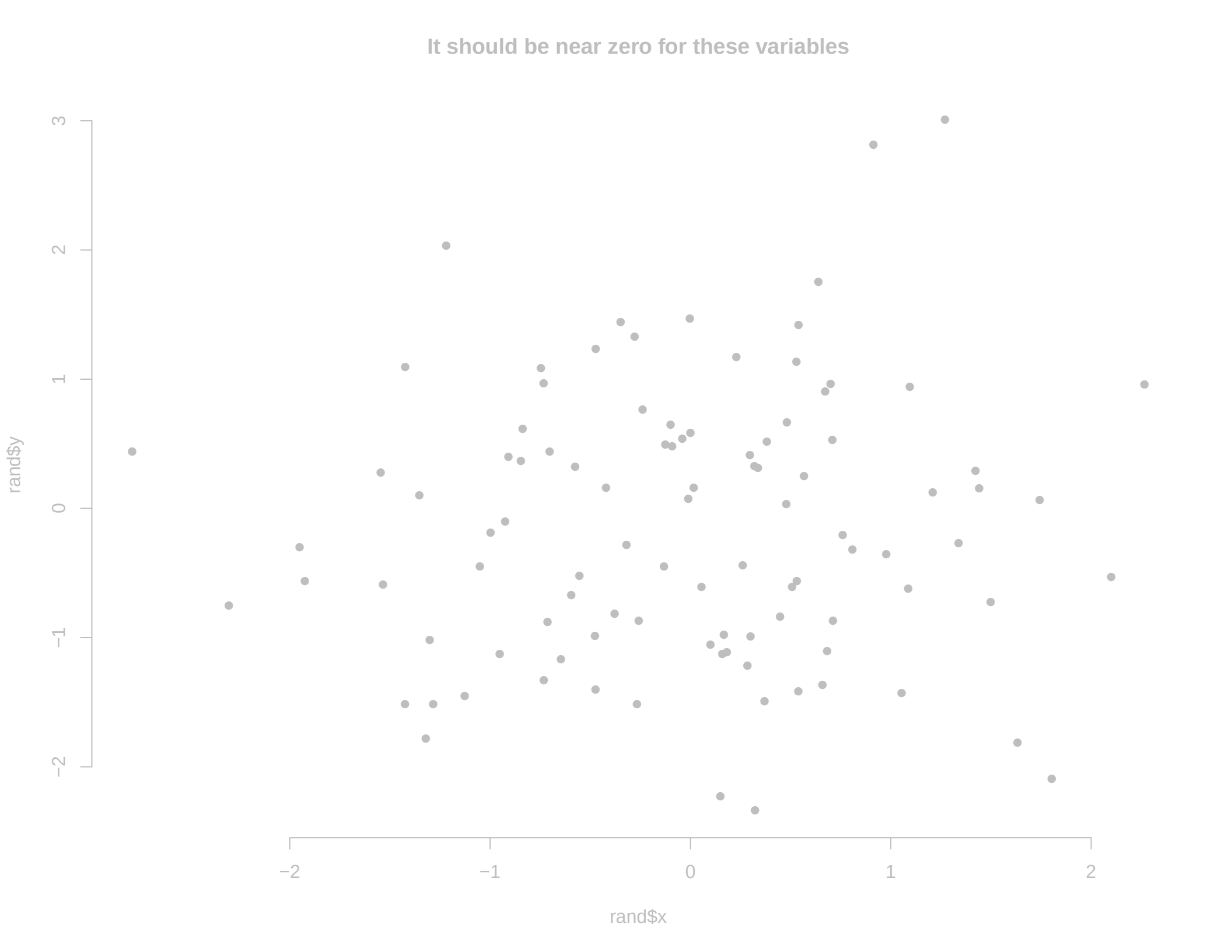

Two air quality variables that move oppositely
Ozone
0
50
100
150
5
10
15
20
Wind


Normal random noise
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

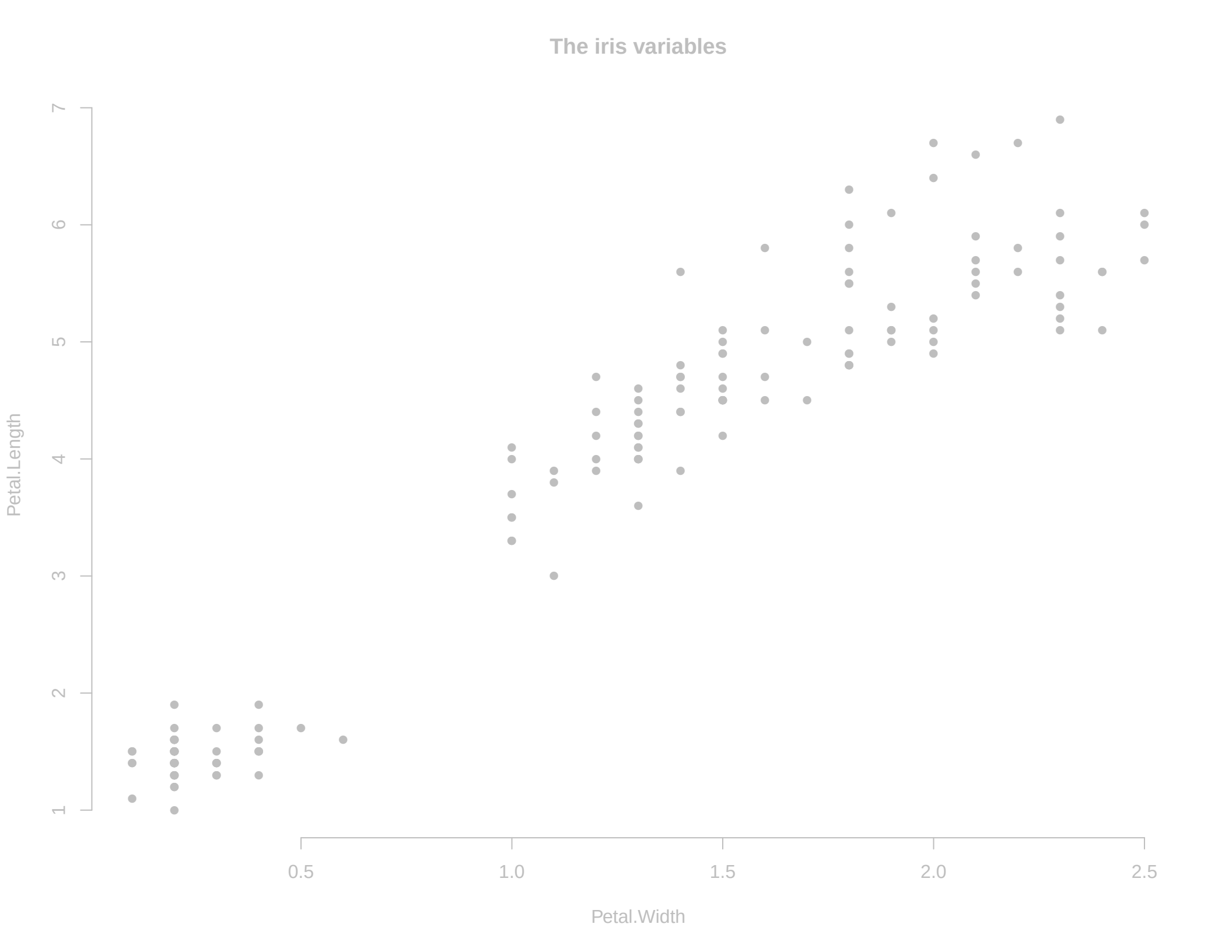
We want a number that describes whether two variables move together.

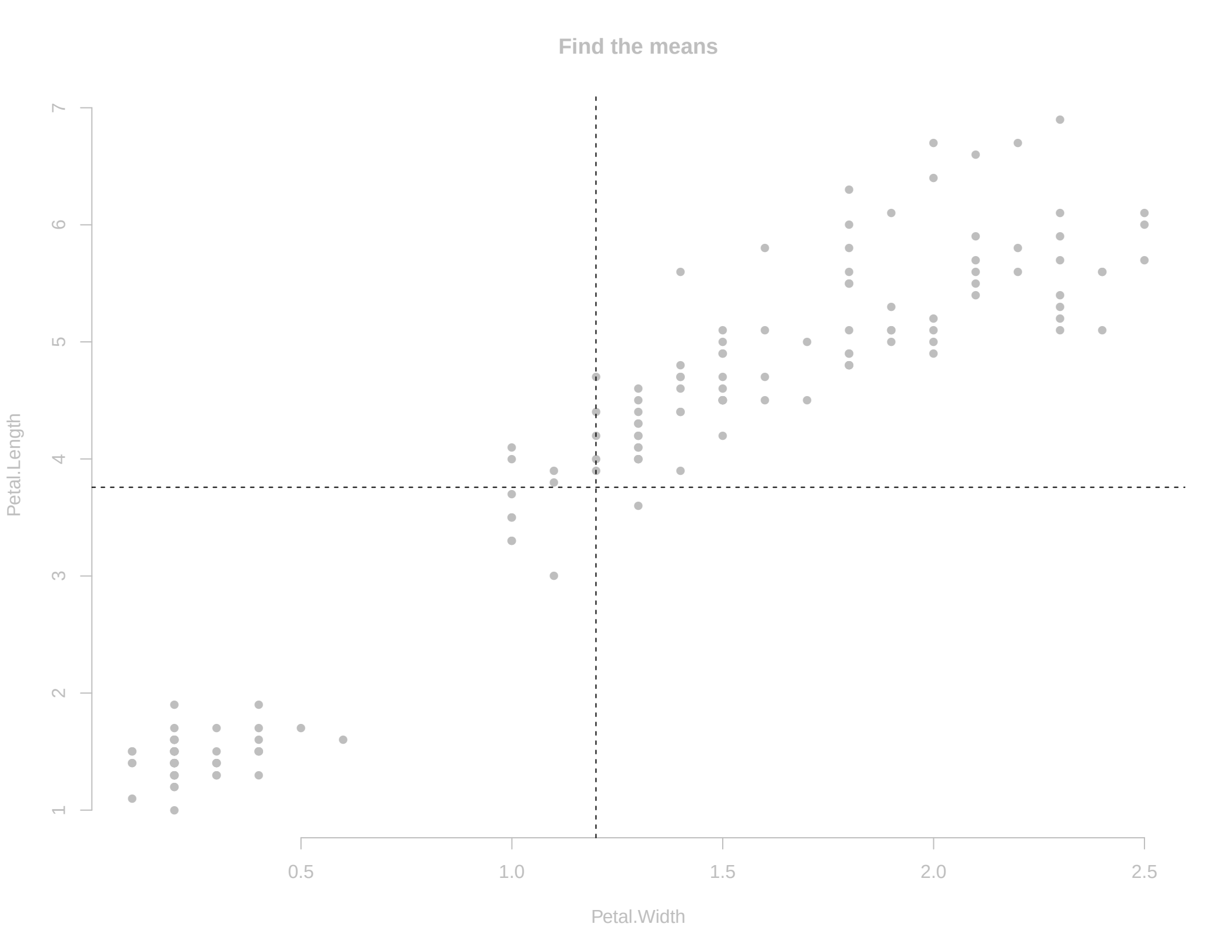

It should be high for these variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

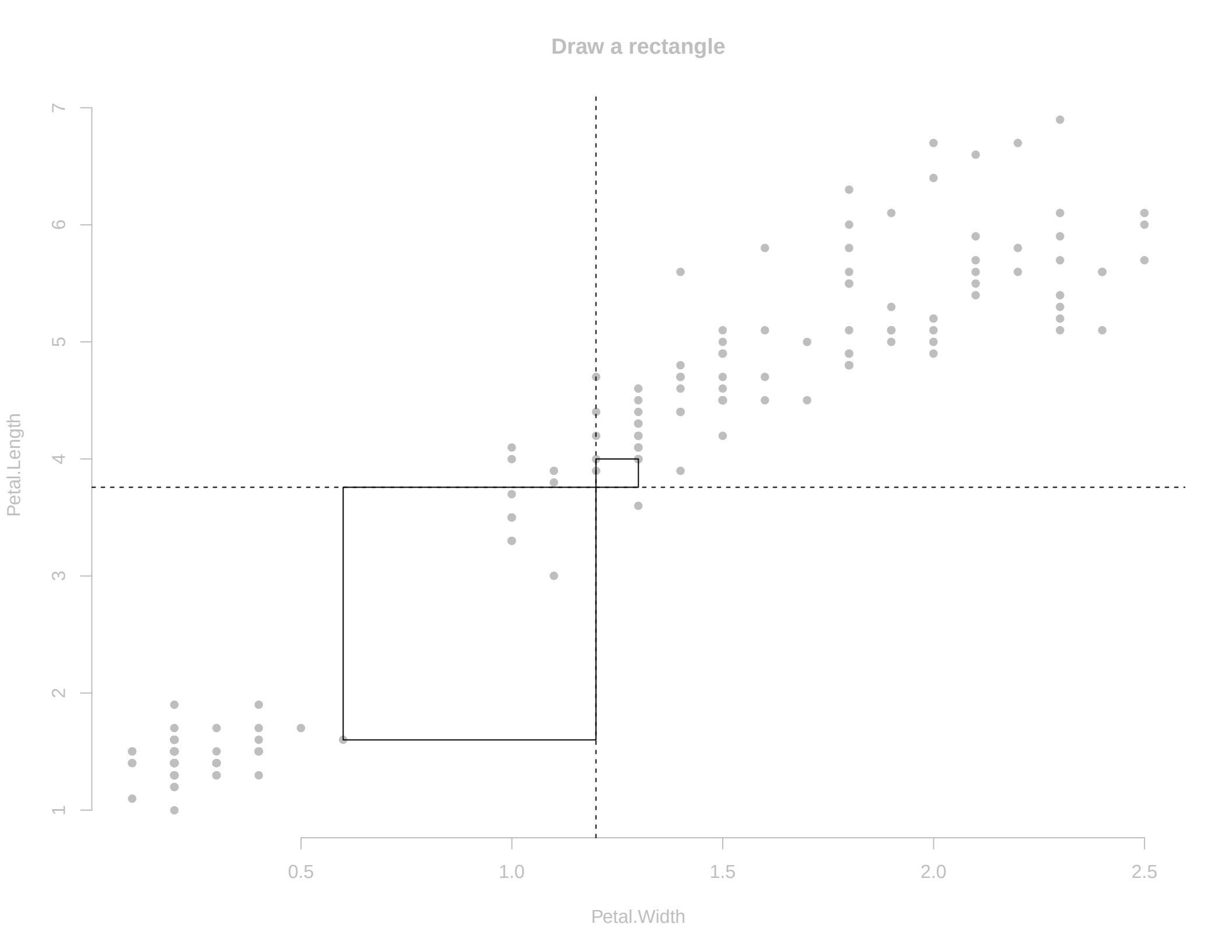

It should be low for these variables
Ozone
0
50
100
150
5
10
15
20
Wind

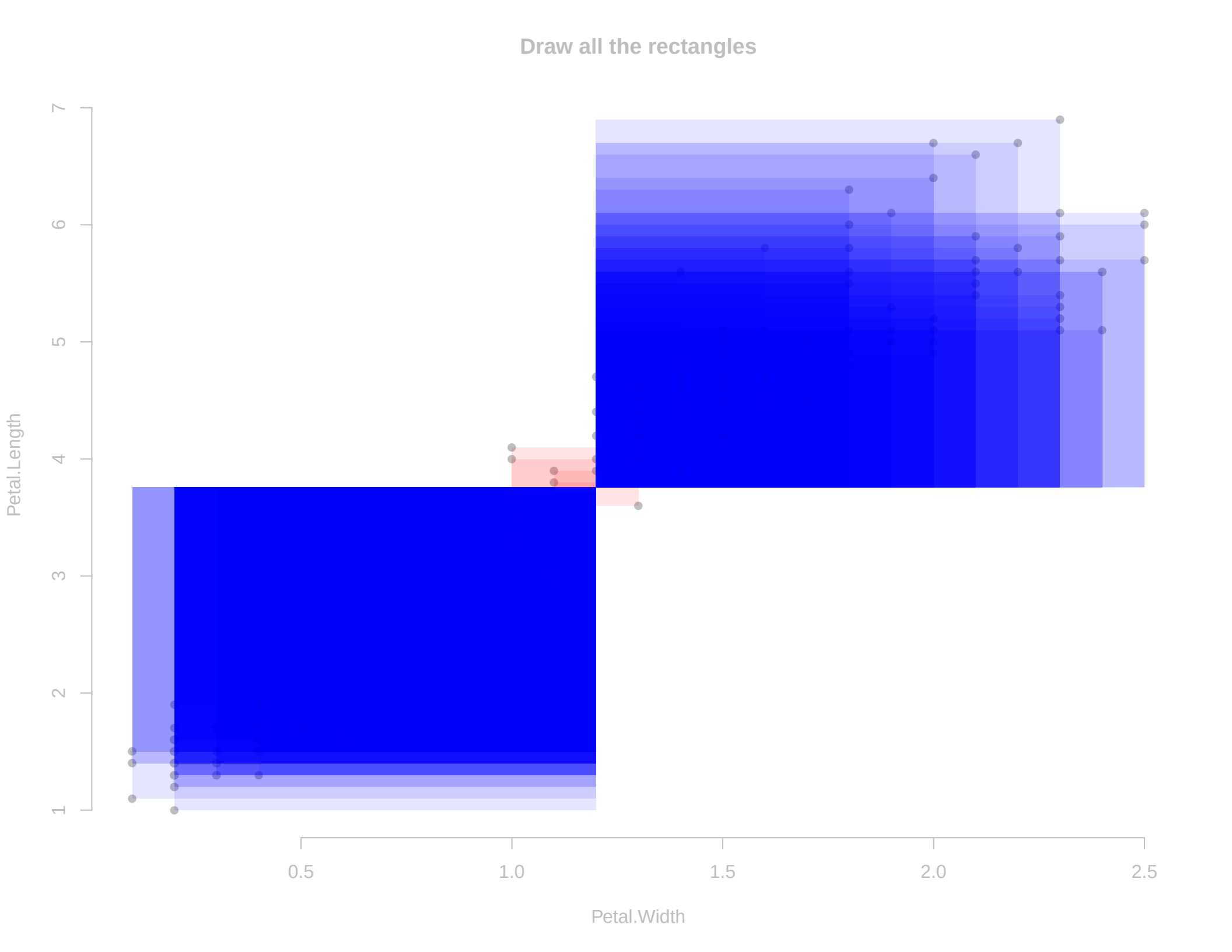

It should be near zero for these variables
rand$y
3
2
1
0
−1
−2
−2
−1
0
1
2
rand$x

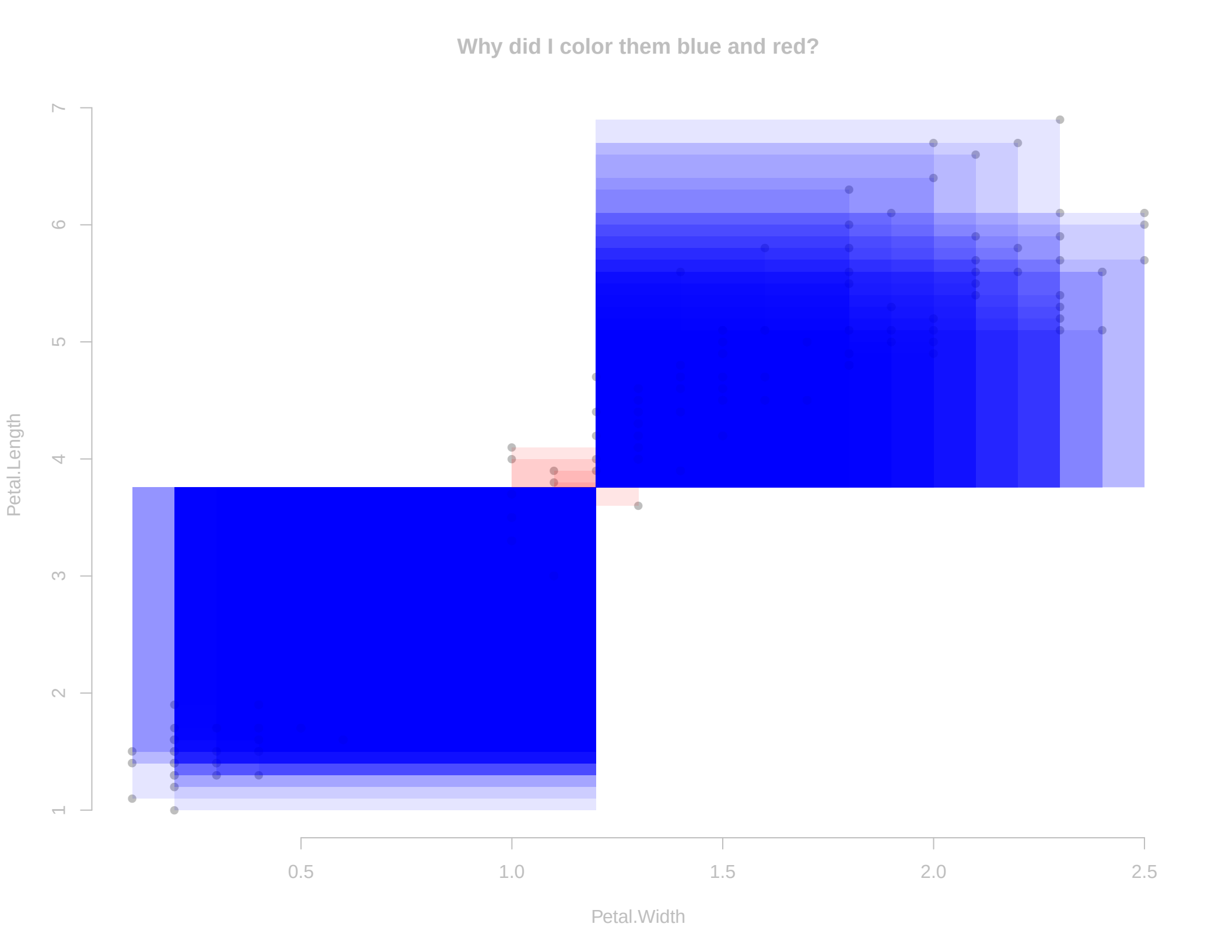
Covariance

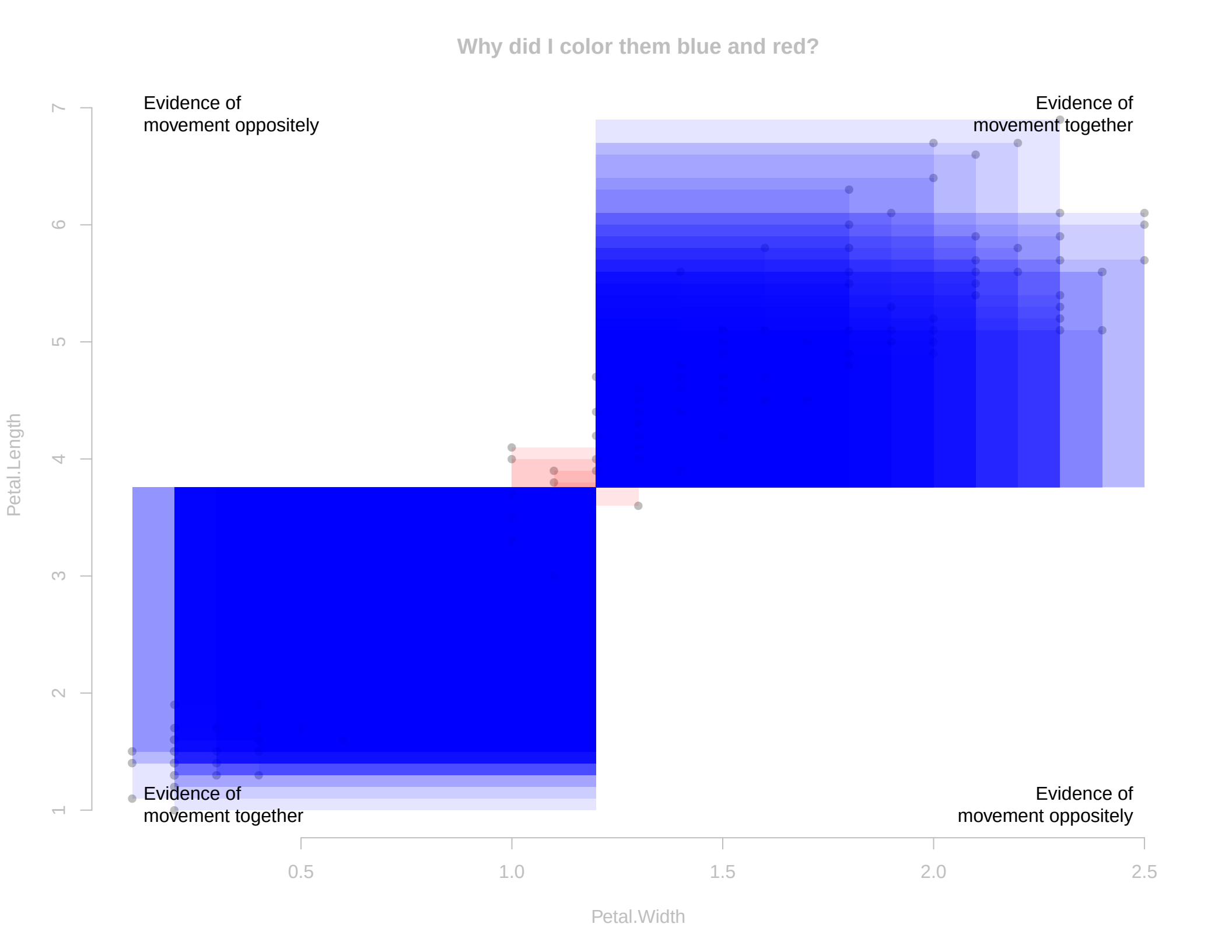

The iris variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Find the means
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

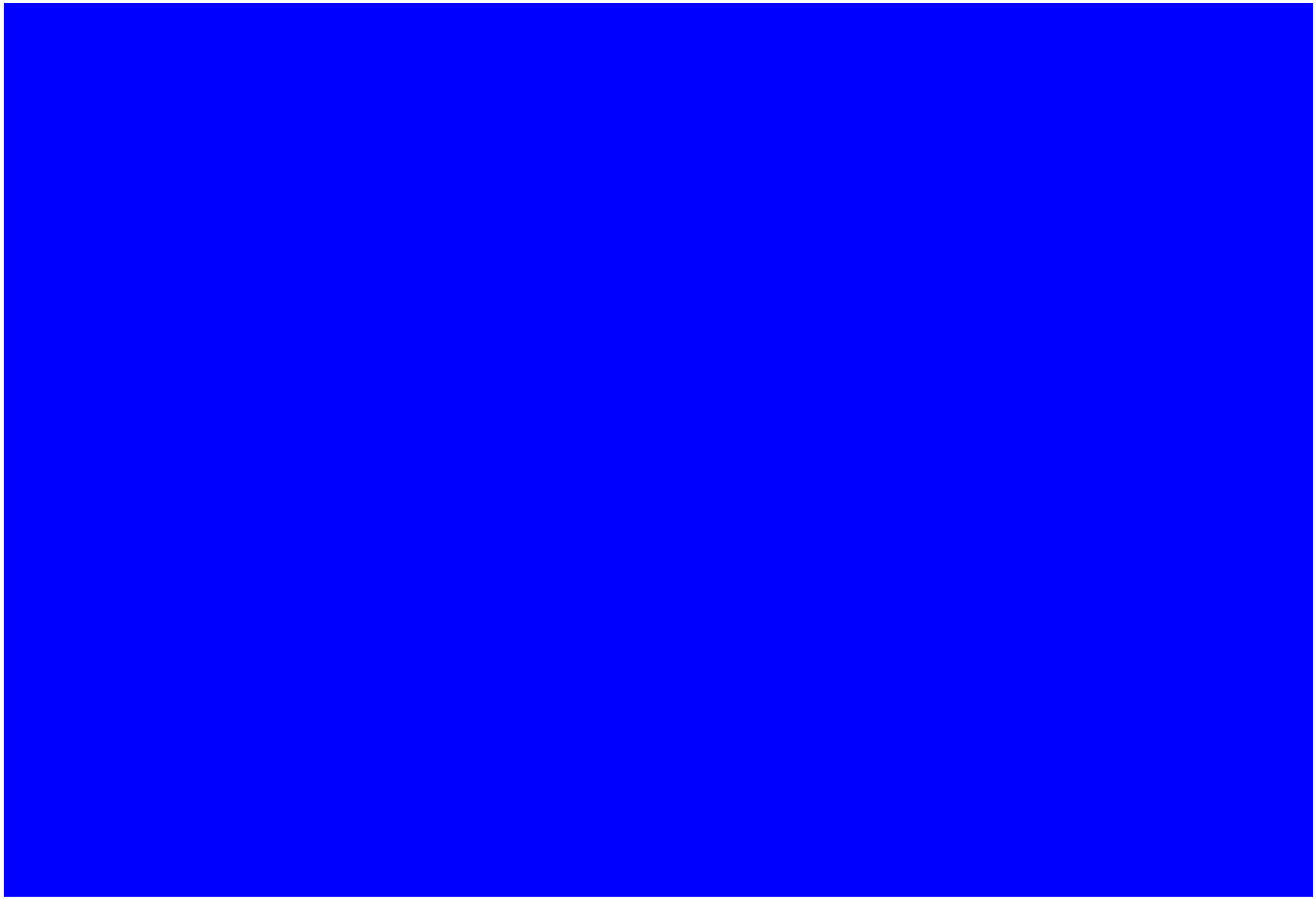

Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

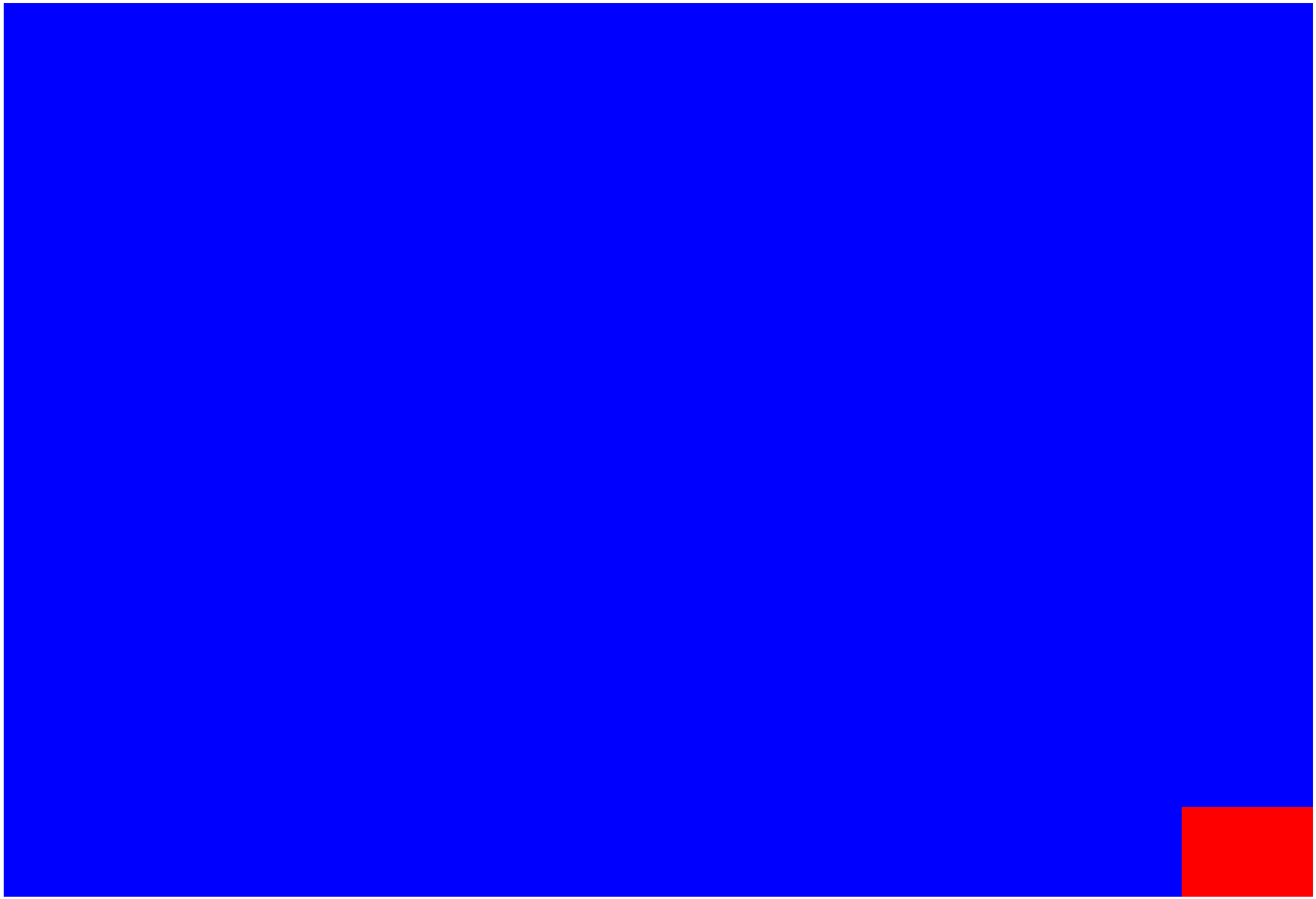

Why did I color them blue and red?
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


Why did I color them blue and red?
Evidence of
movement oppositely
Evidence of
movement together
Evidence of
movement together
Evidence of
movement oppositely
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

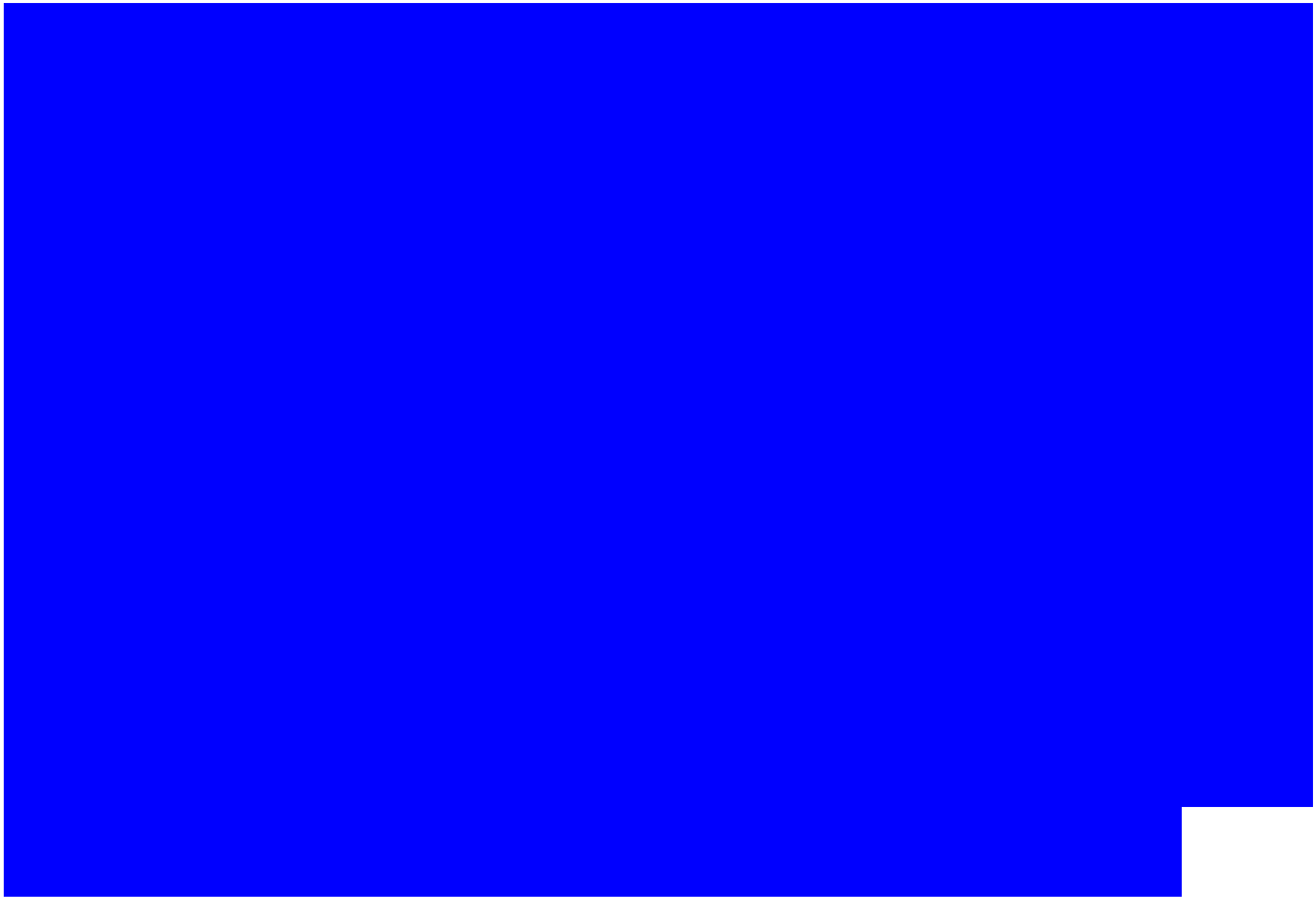
Add the blues together. (This is at a different scale.)



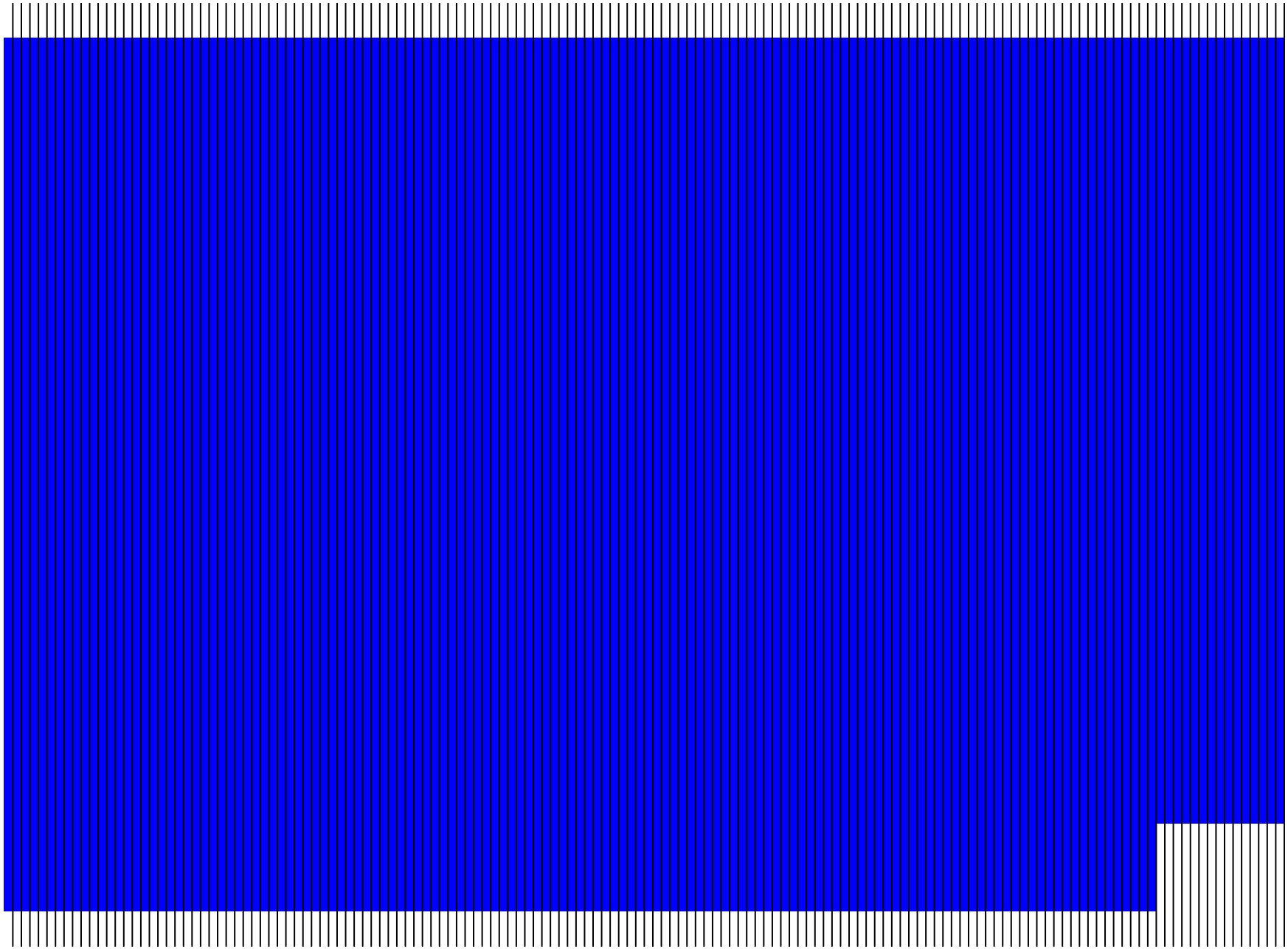
Add the reds together.



Subtract the reds.

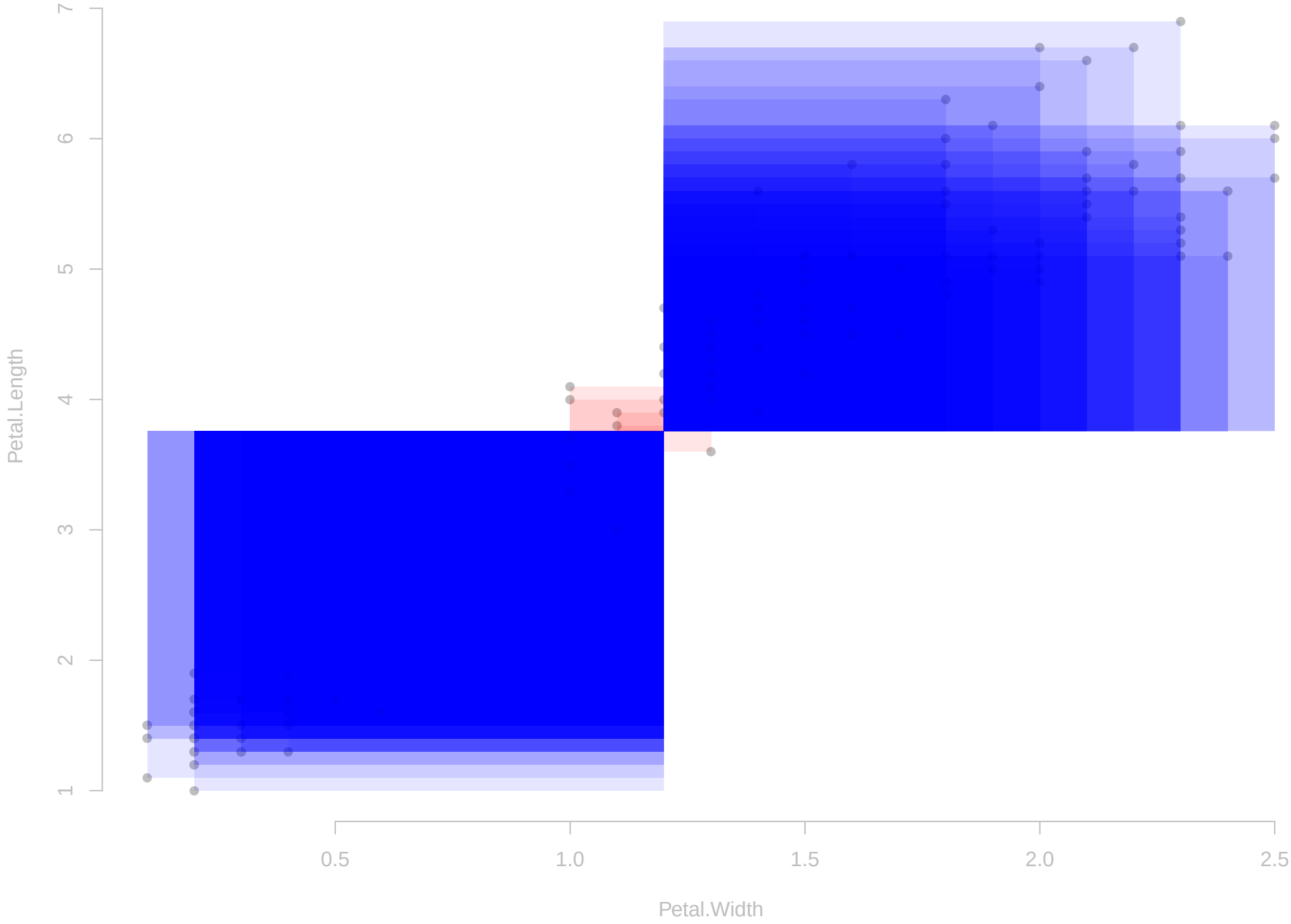


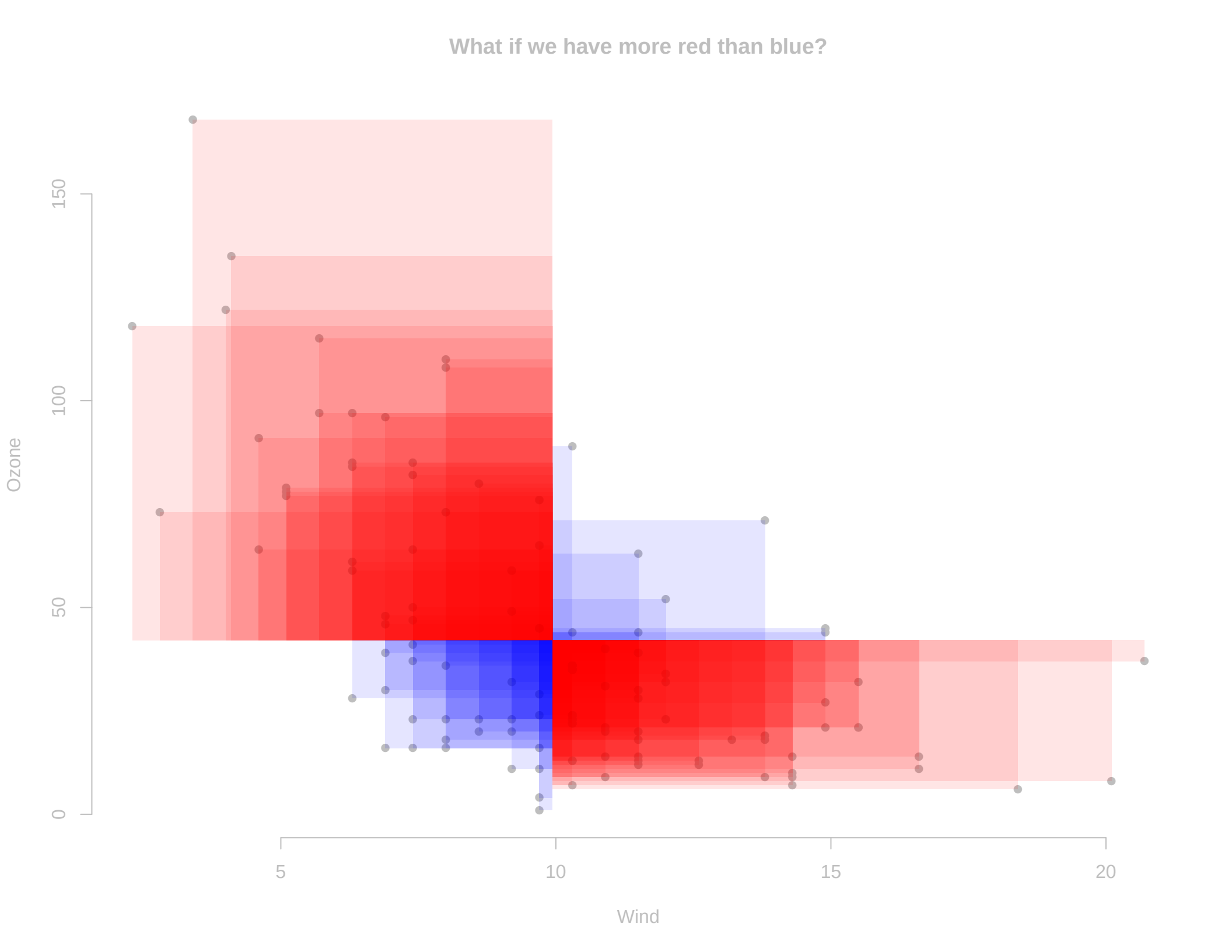
Divide into as many equal pieces as we have irises (n).



This blue sliver is the covariance.

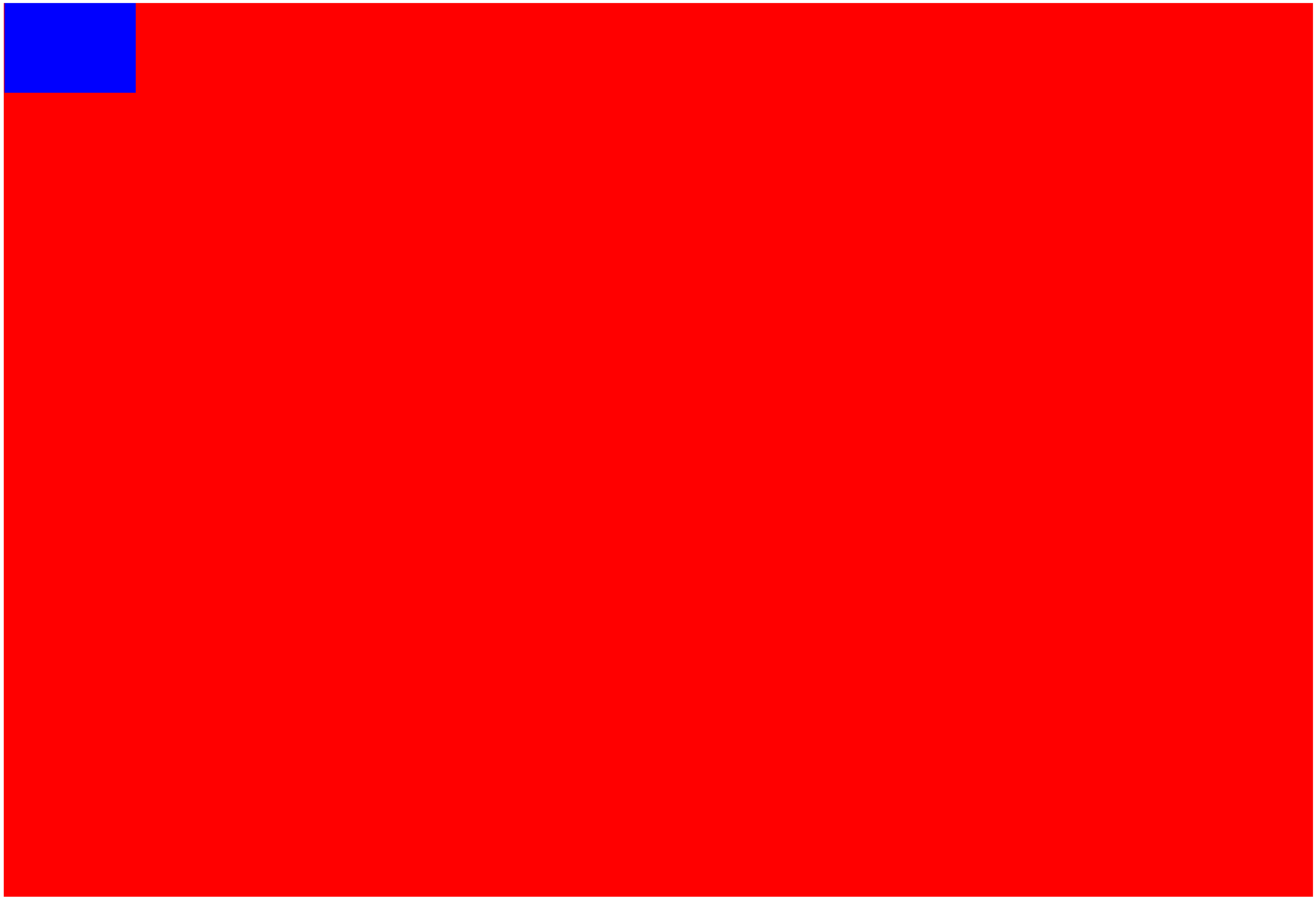
That was for this sort of relationship.



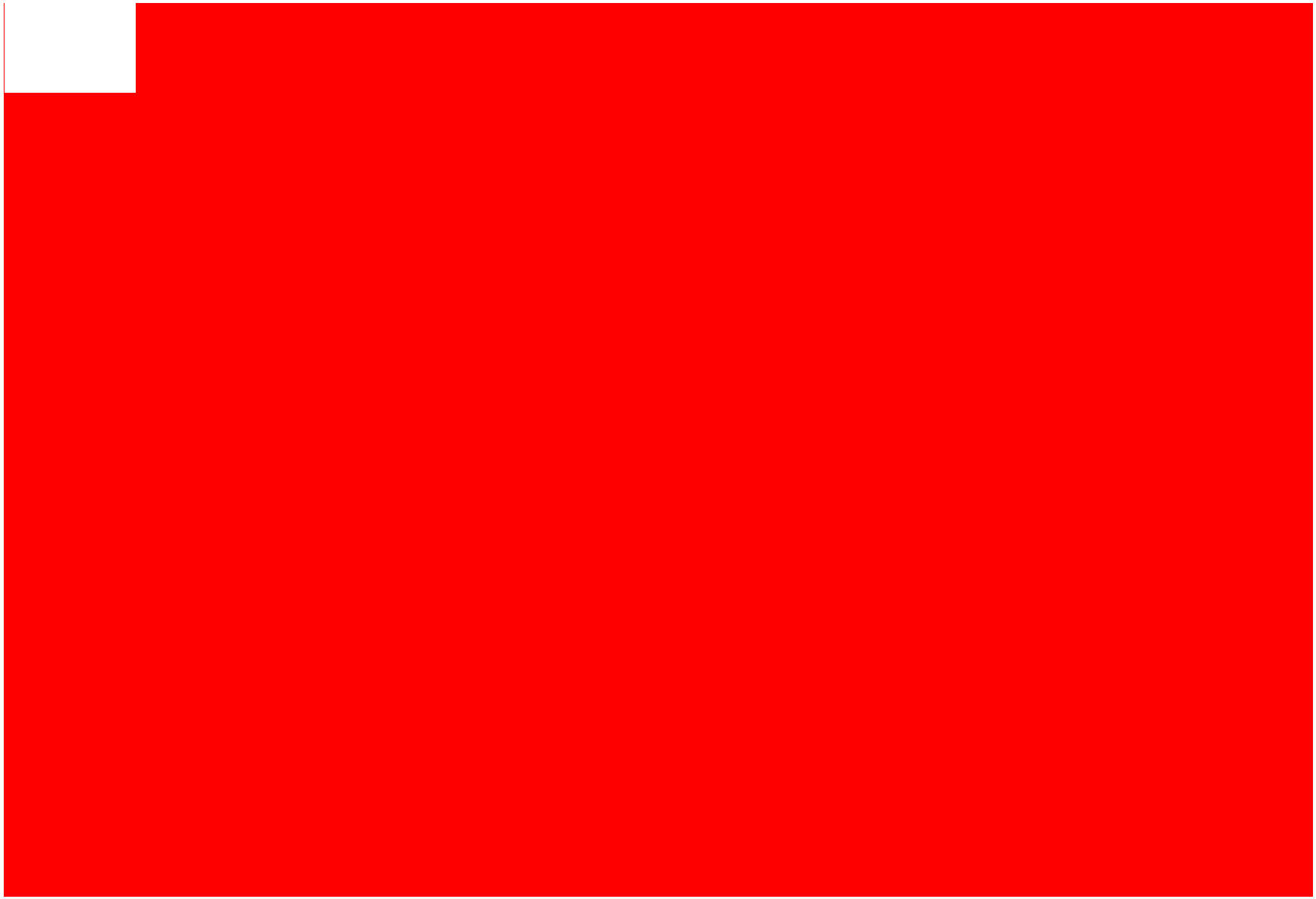

What if we have more red than blue?
Ozone
150
100
50
0
5
10
15
20
Wind

Add the blues together. (This is at a different scale.)

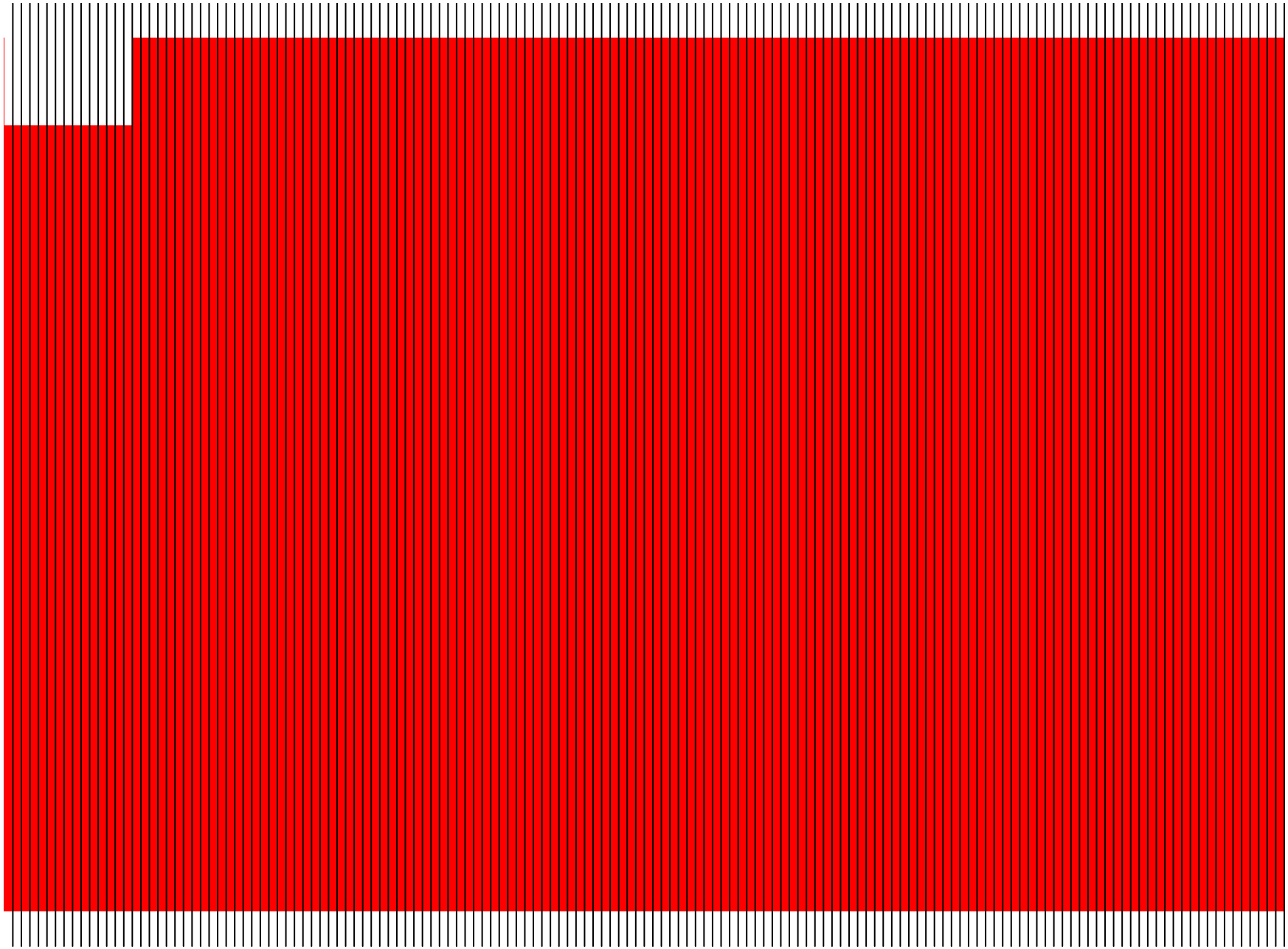
Add the reds together.



Subtract the reds.



Divide into as many equal pieces as we have irises (n).



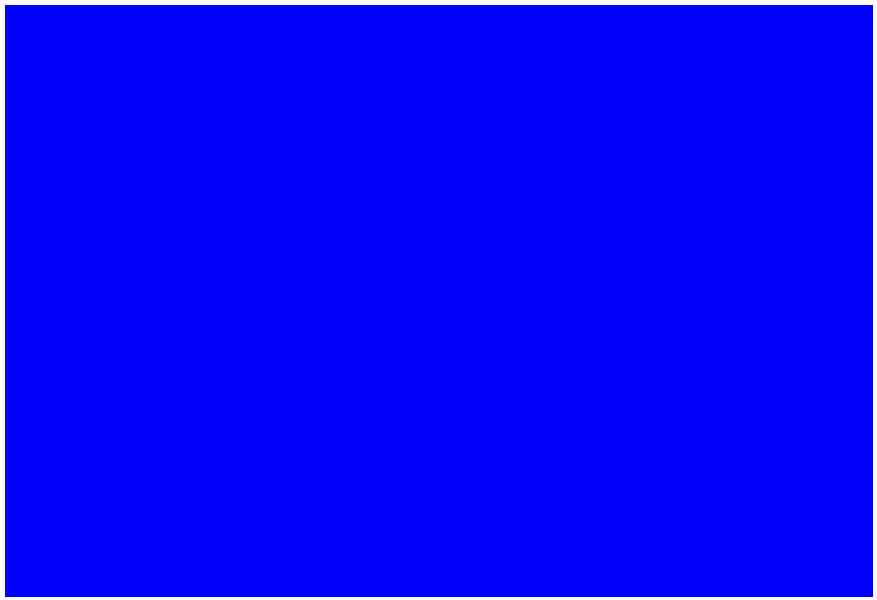
This red sliver is the covariance.

This red sliver is the covariance.

But it's negative!


What if we have as much red as blue?
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

Add the blues together. (This is at a different scale.)

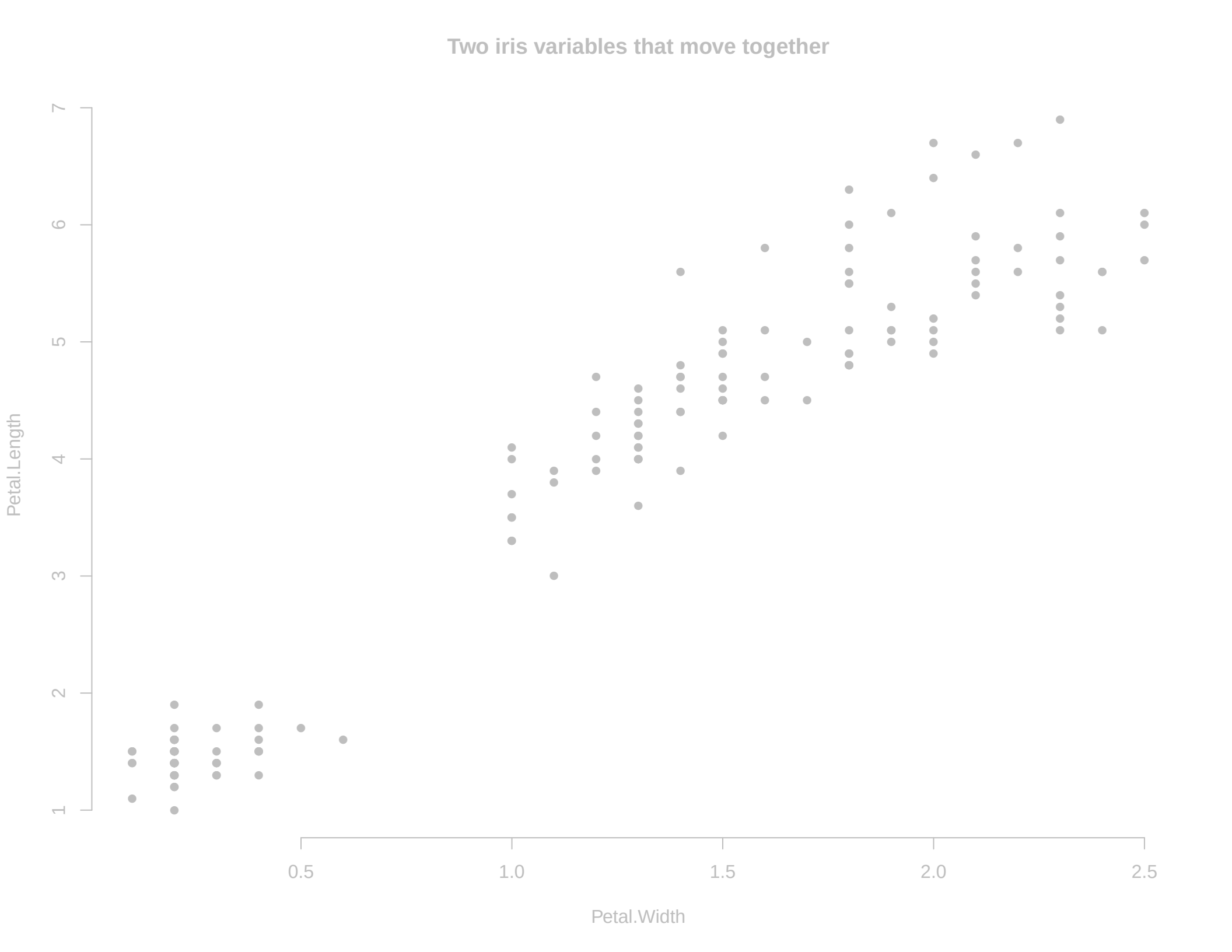


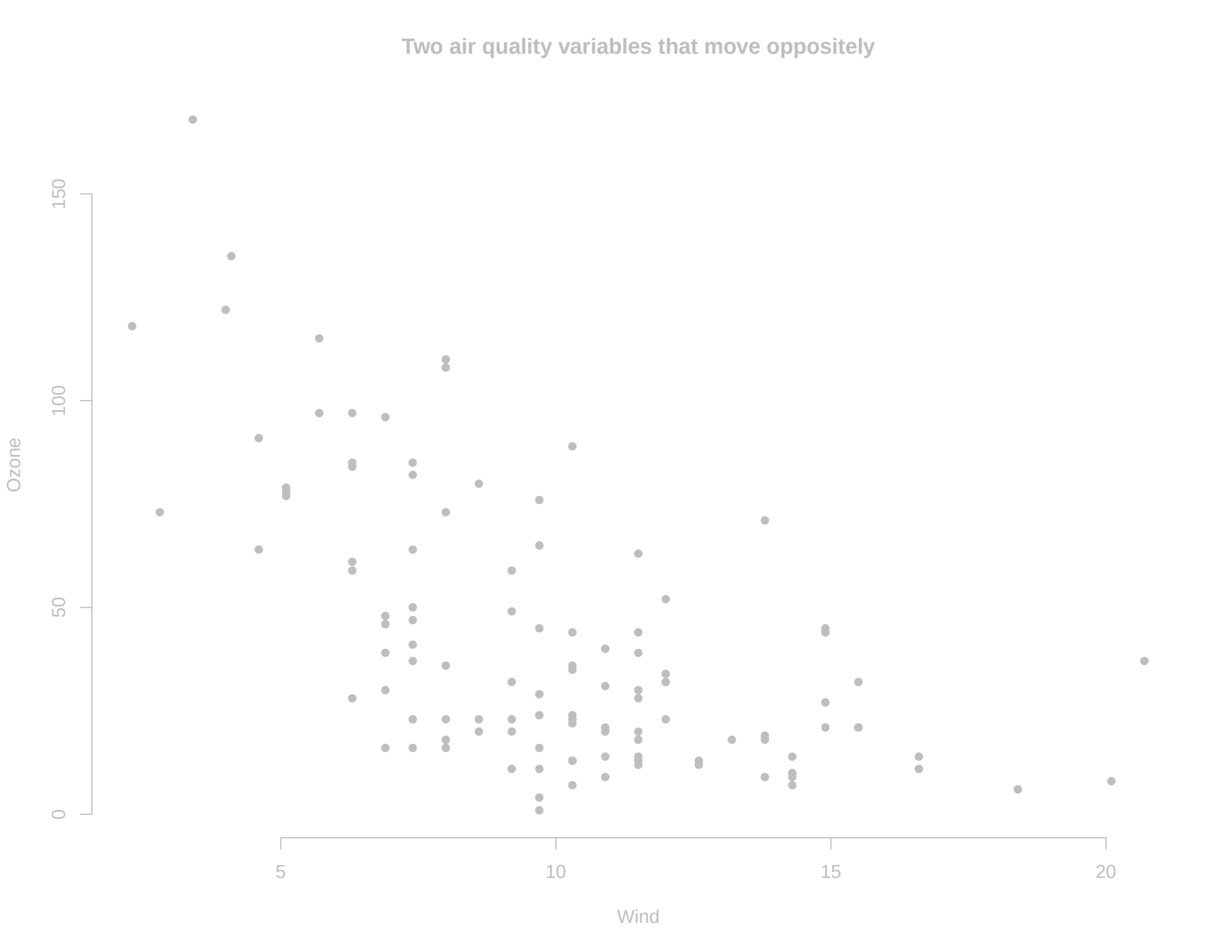
Add the reds together.

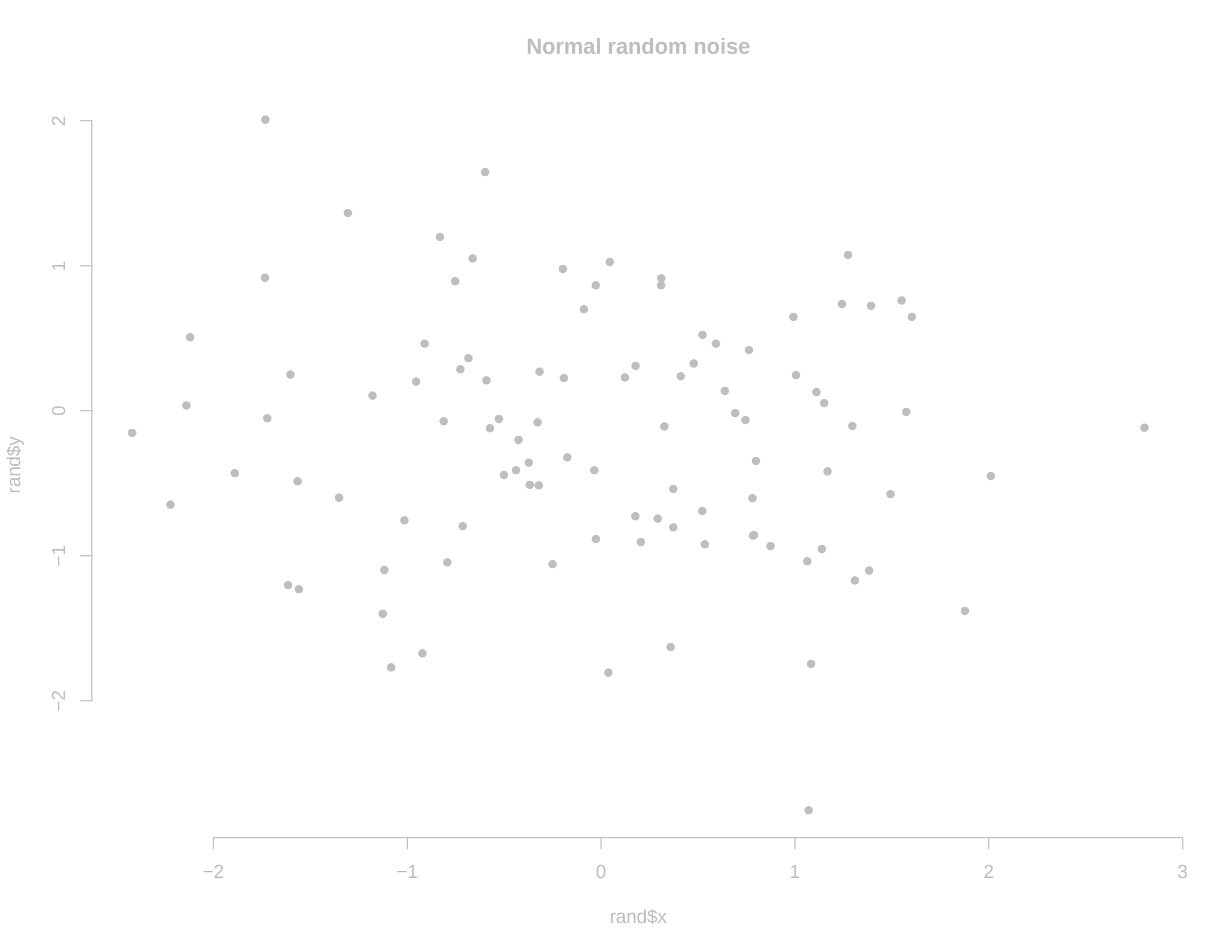
Subtract the reds.

(Covariance is zero.)

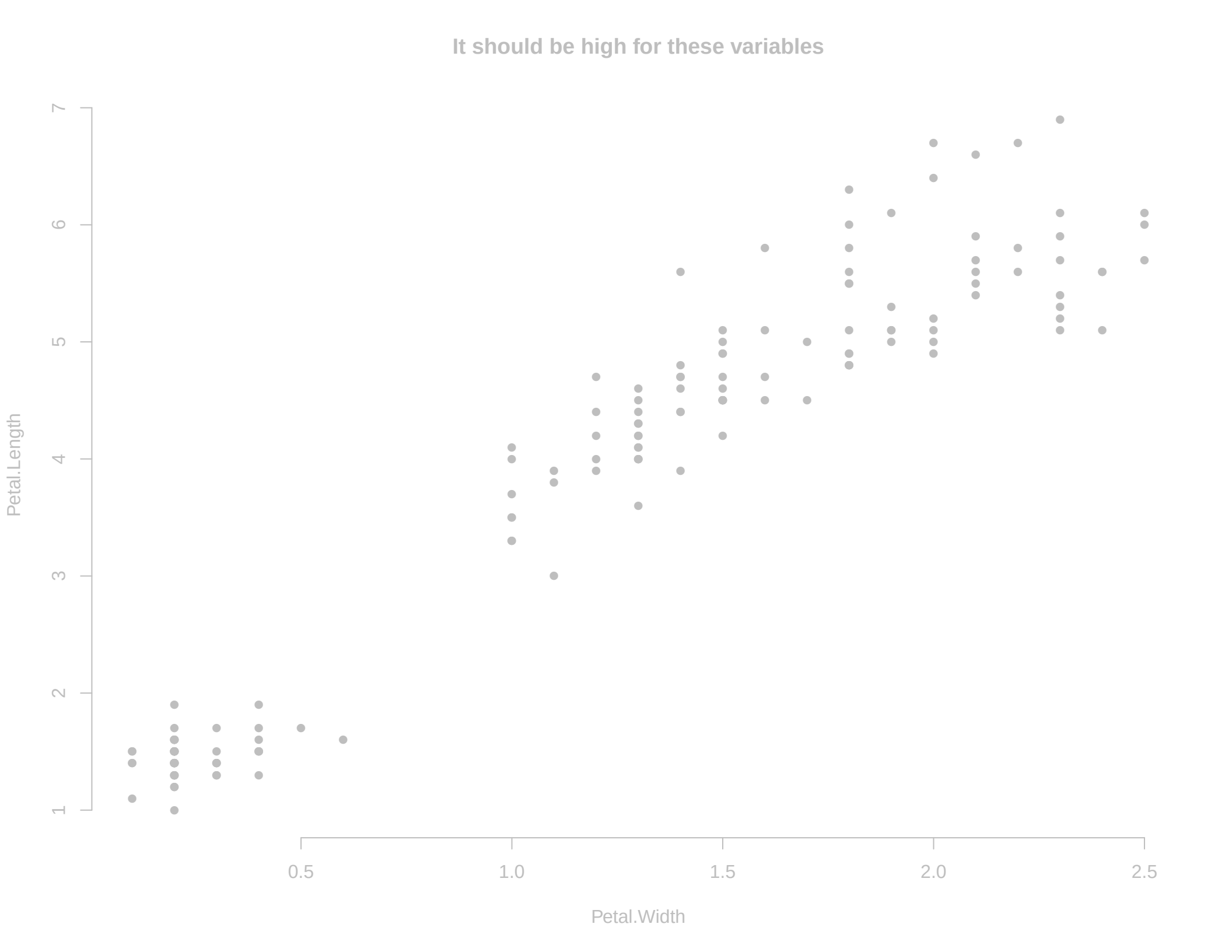
Let's review the previous slides quickly.

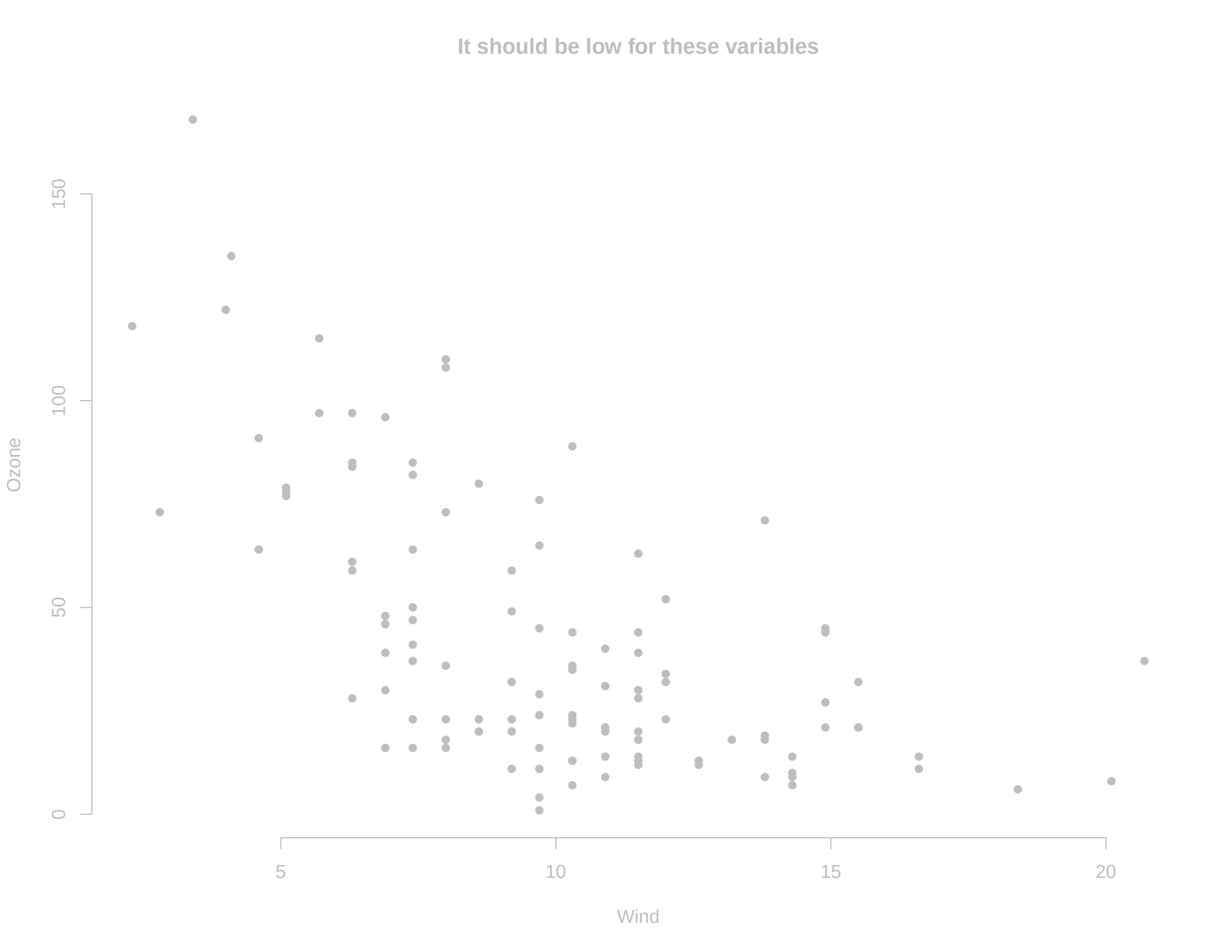

Two iris variables that move together
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

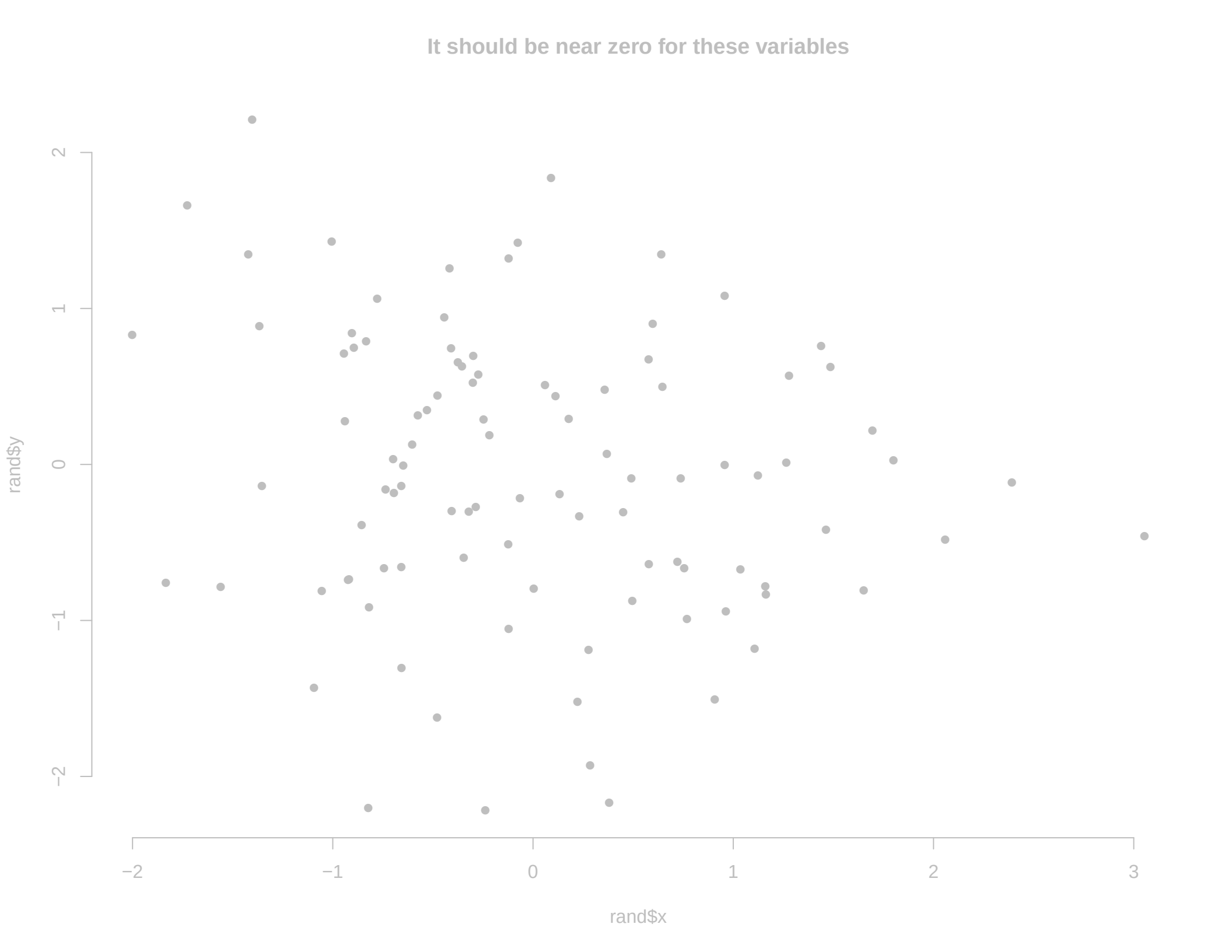

Two air quality variables that move oppositely
Ozone
0
50
100
150
5
10
15
20
Wind


Normal random noise
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

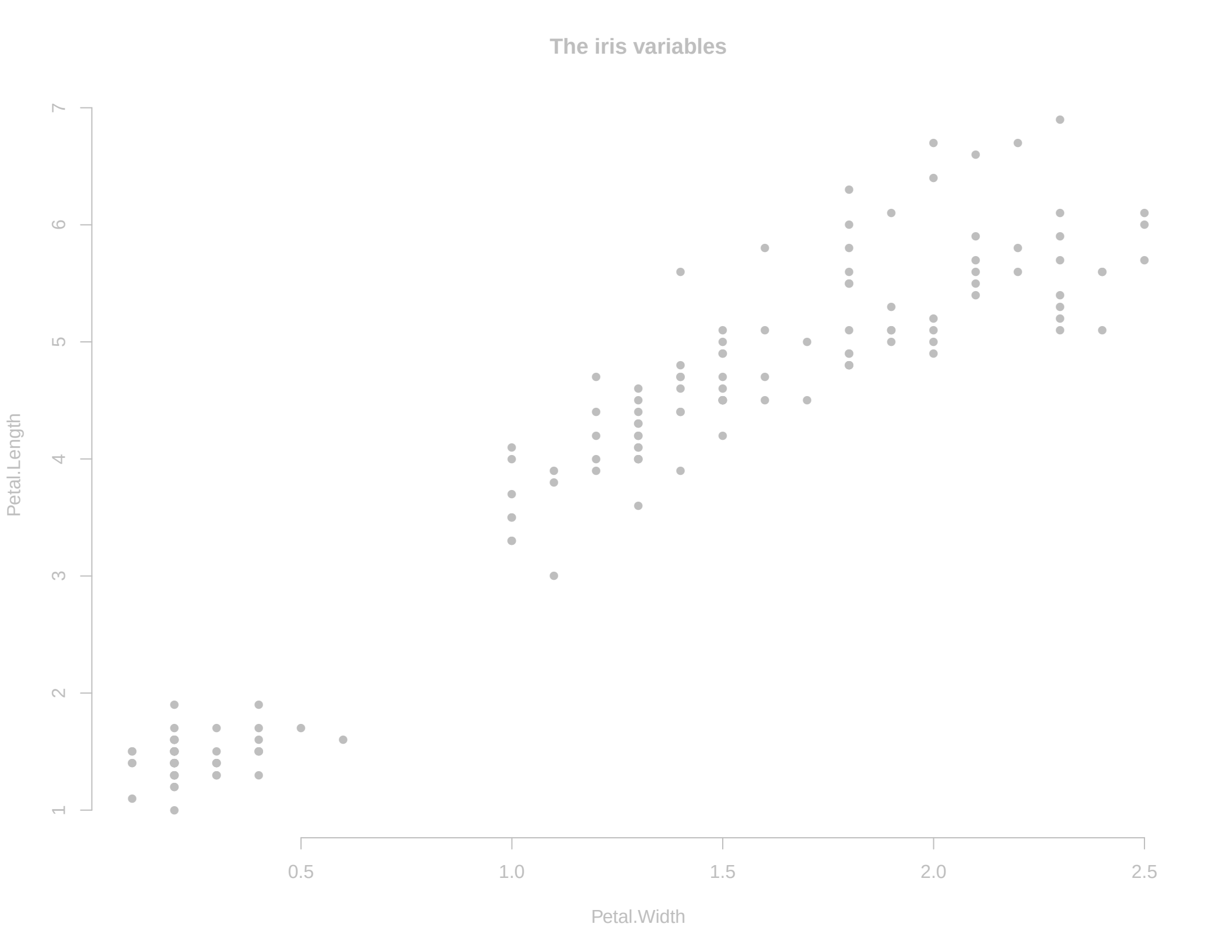
We want a number that describes whether two variables move together.

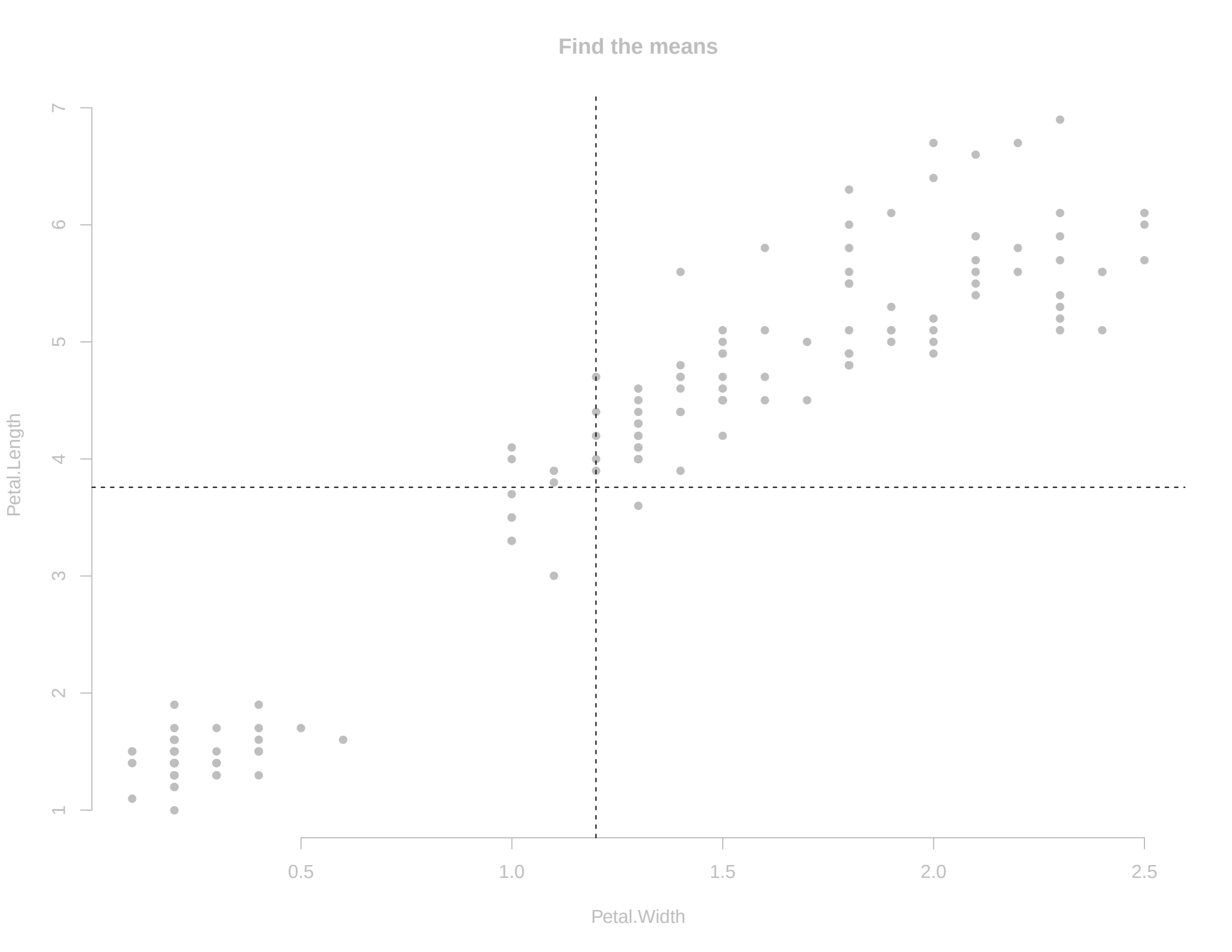

It should be high for these variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

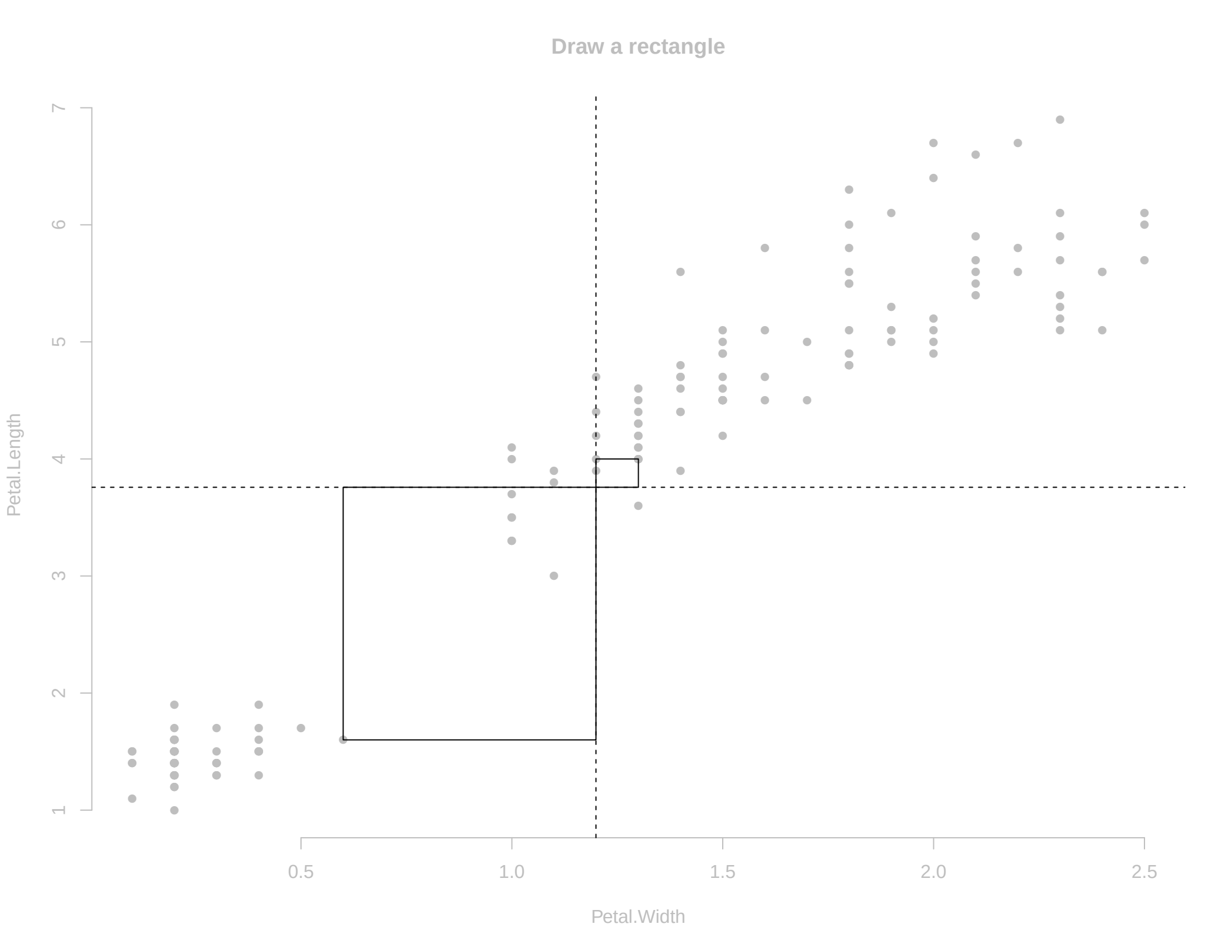

It should be low for these variables
Ozone
0
50
100
150
5
10
15
20
Wind

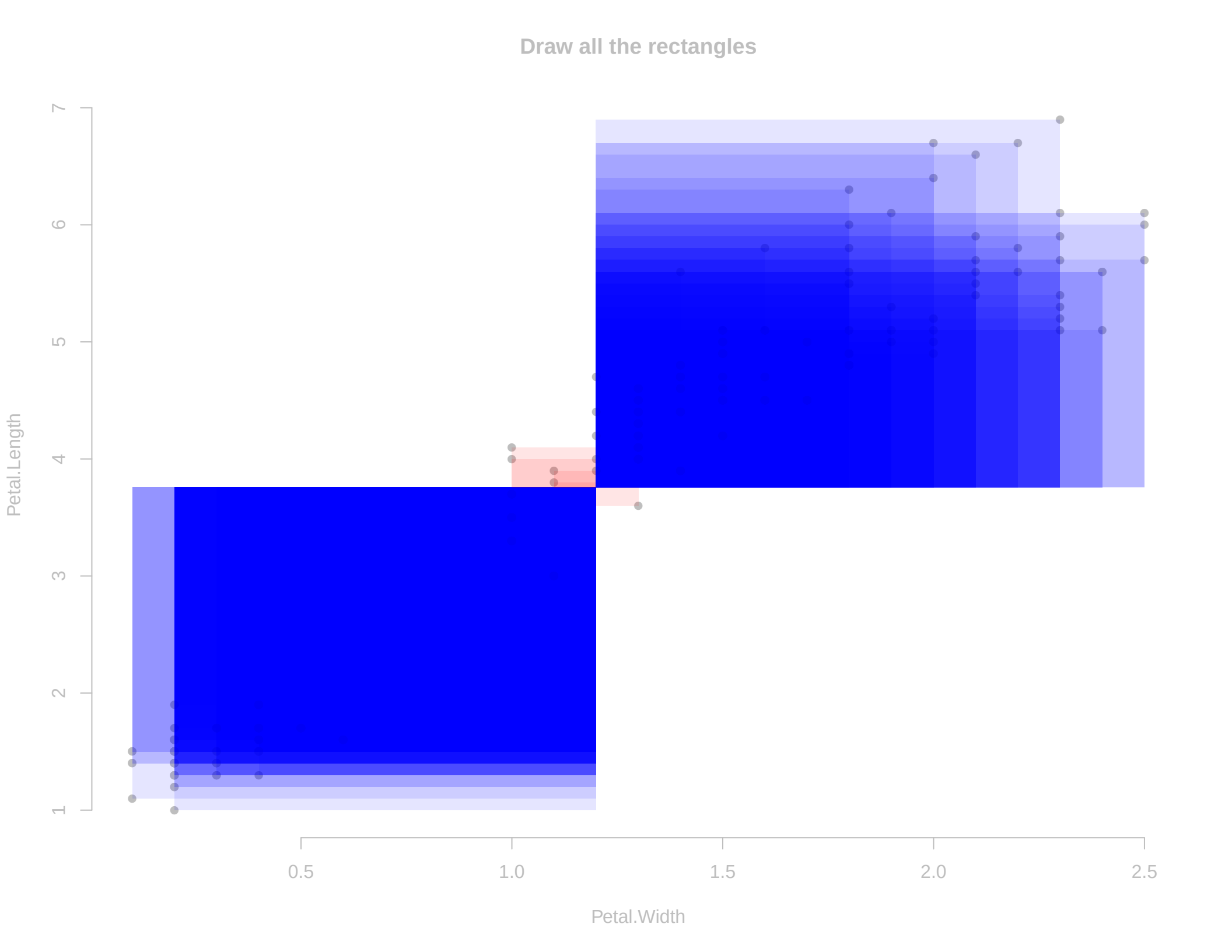

It should be near zero for these variables
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

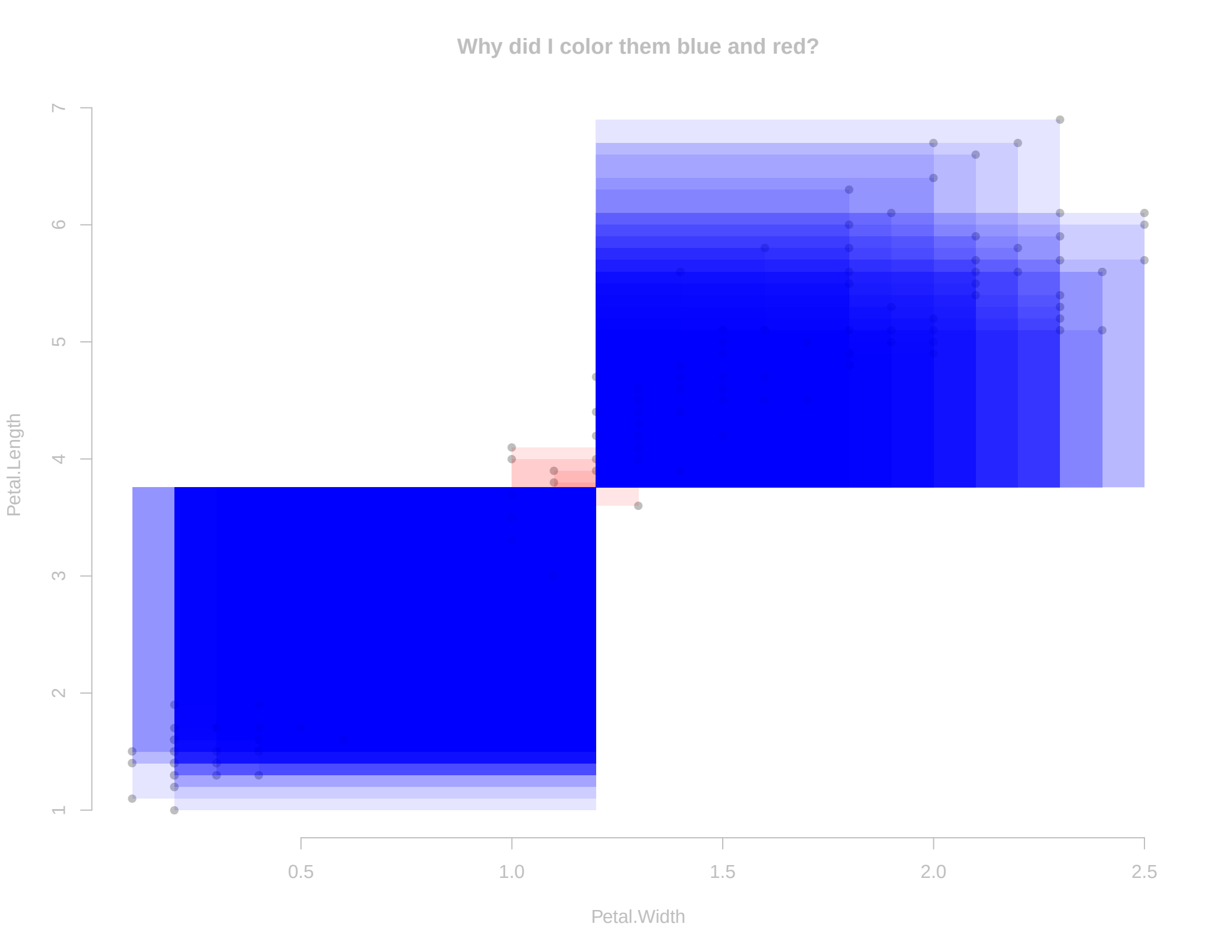
Covariance

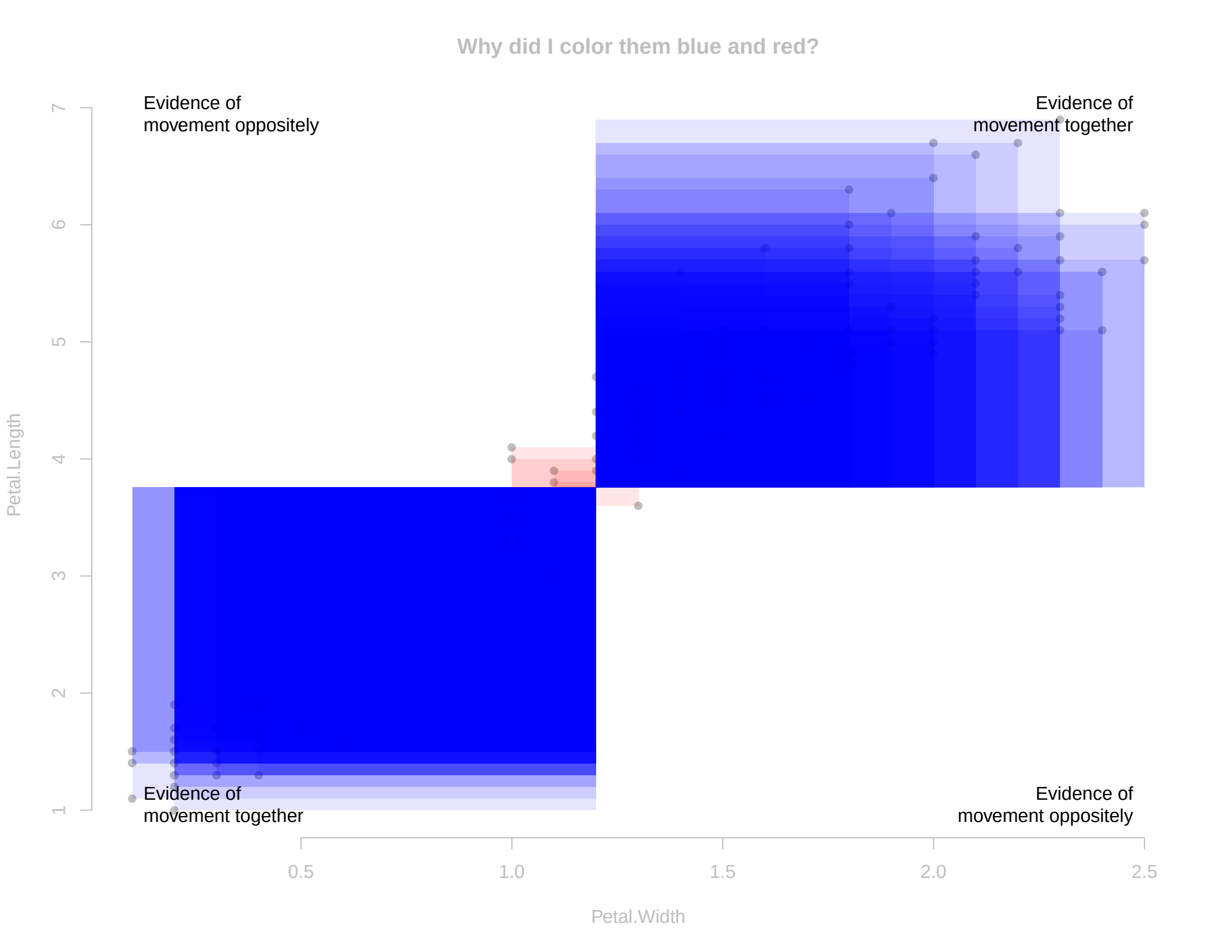

The iris variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Find the means
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

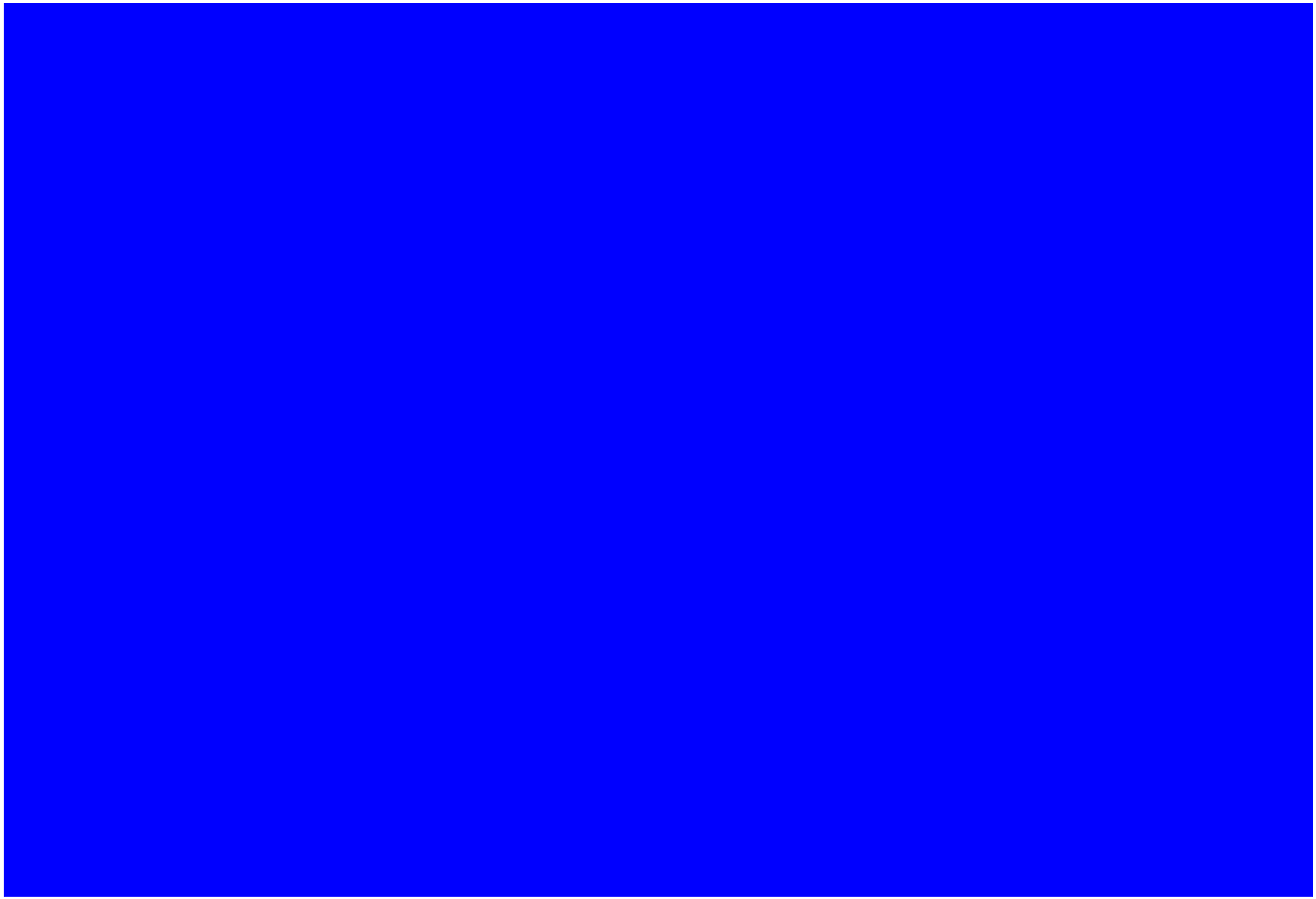

Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

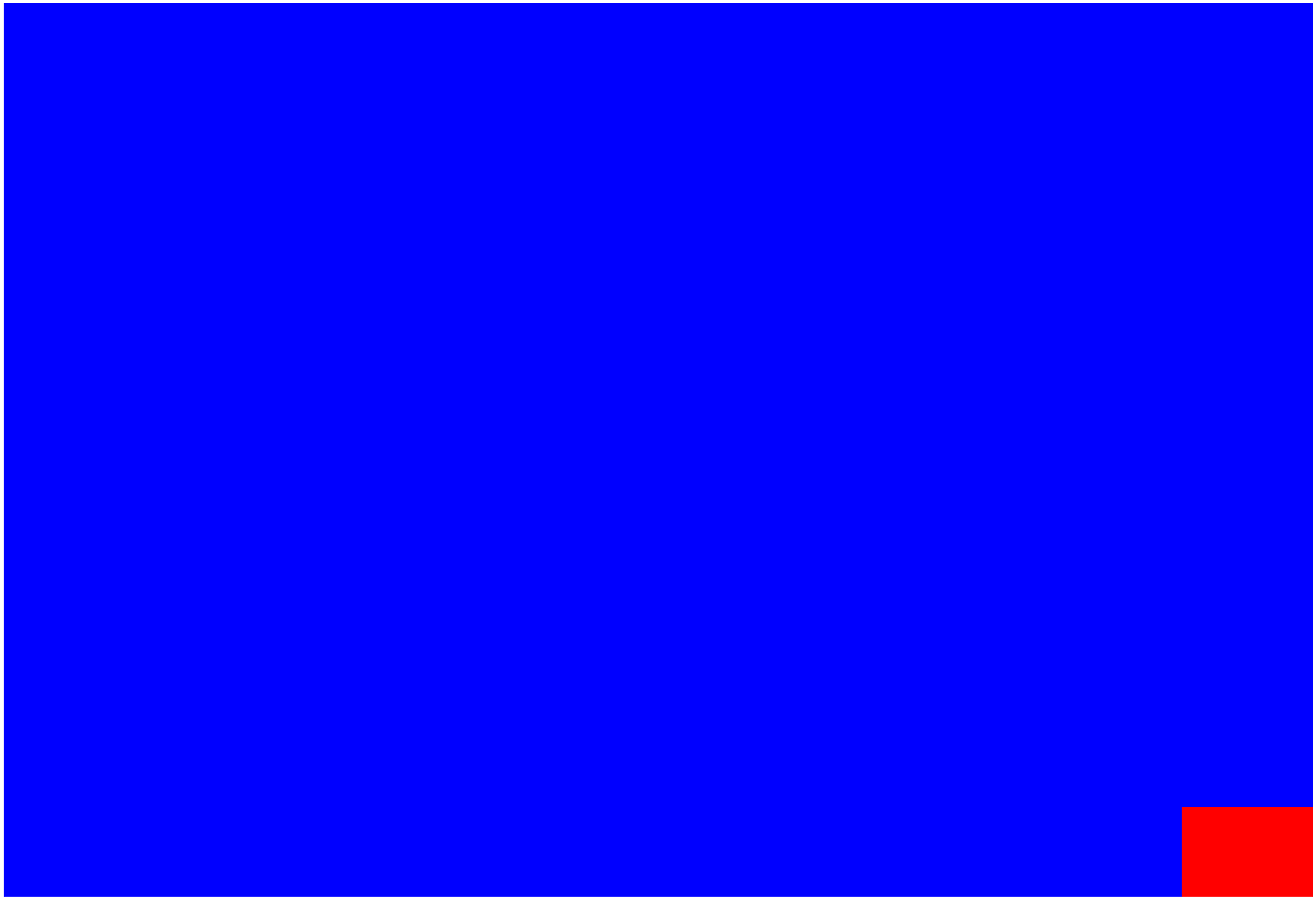

Why did I color them blue and red?
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


Why did I color them blue and red?
Evidence of
movement oppositely
Evidence of
movement together
Evidence of
movement together
Evidence of
movement oppositely
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

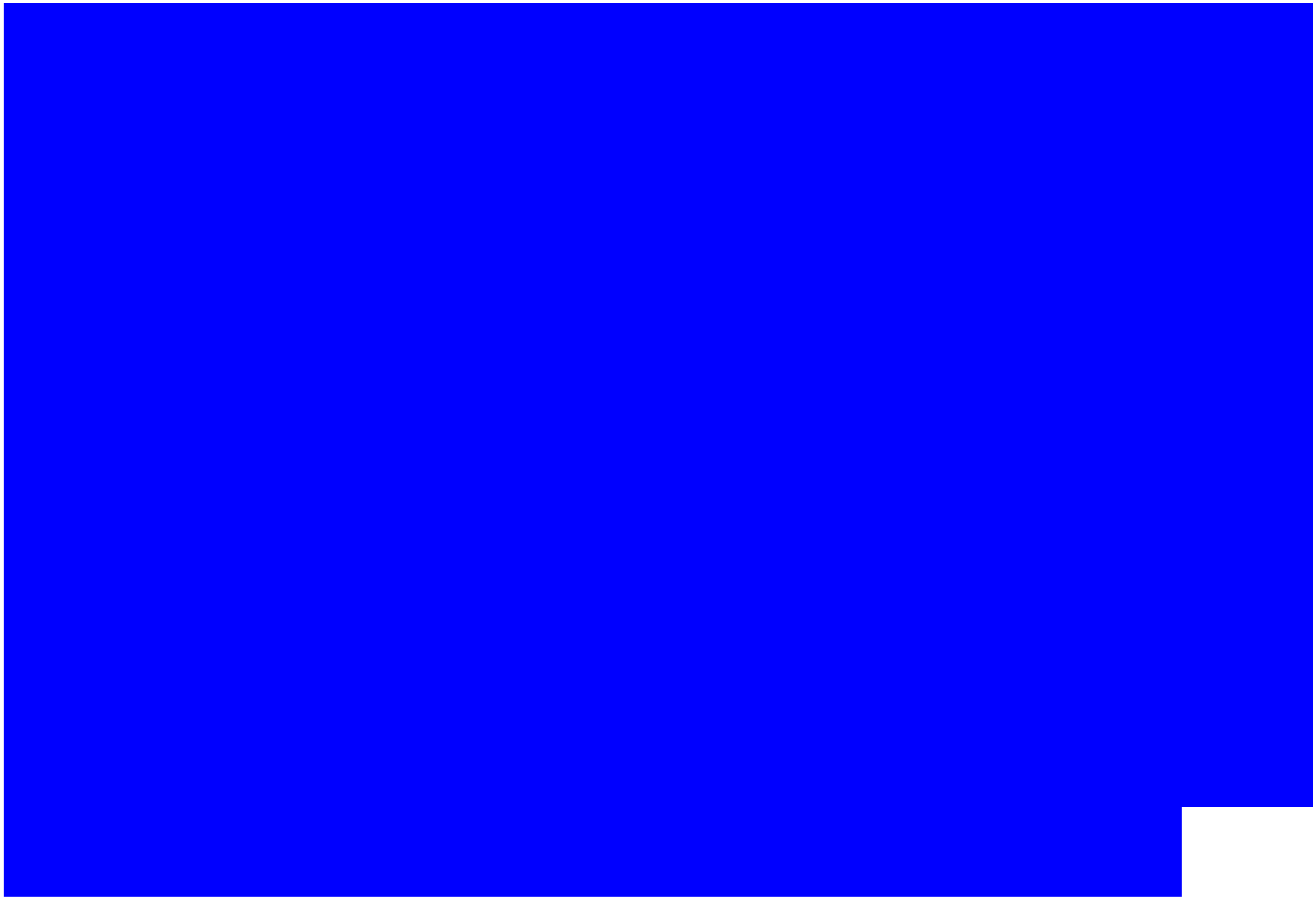
Add the blues together. (This is at a different scale.)



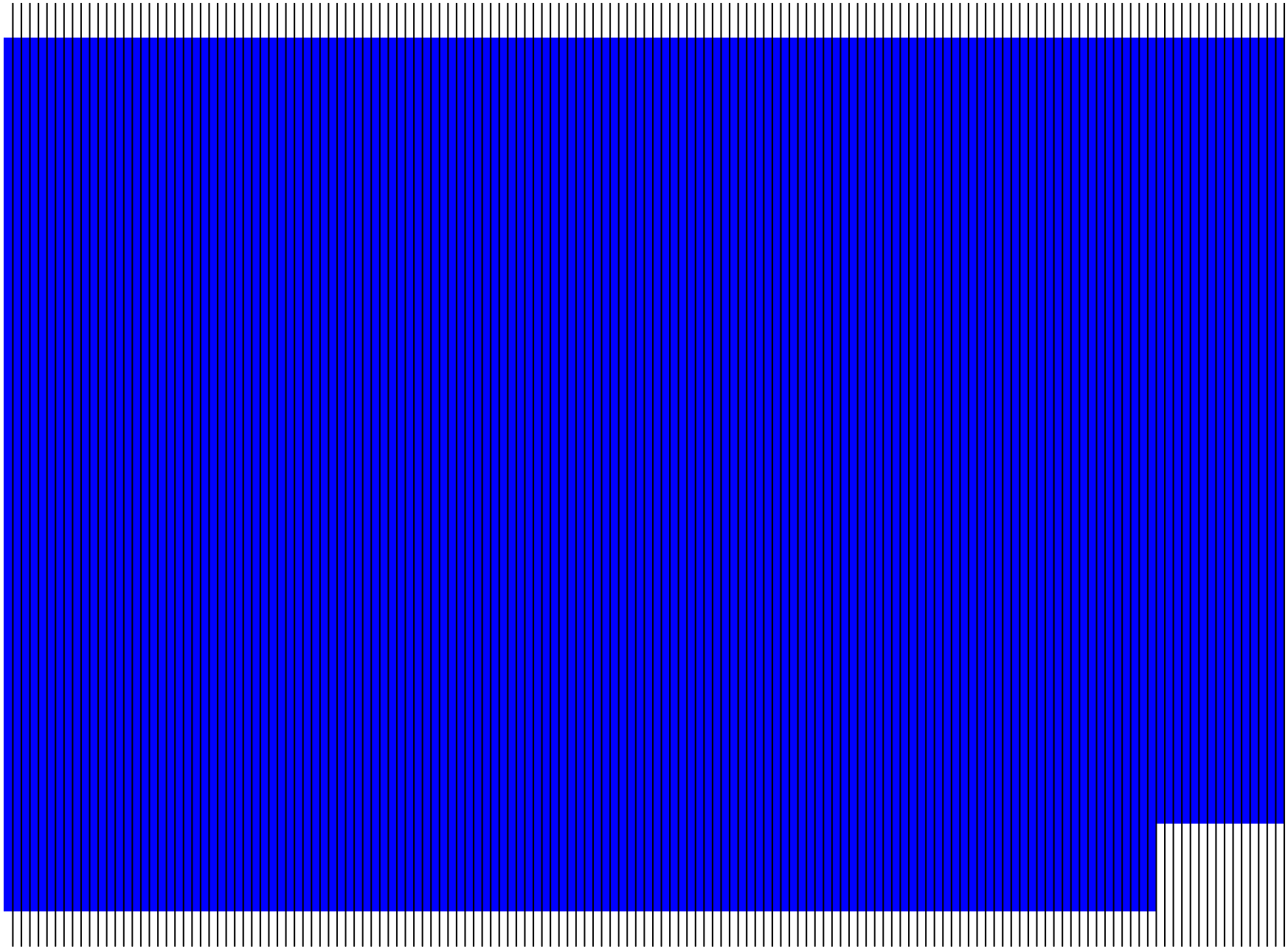
Add the reds together.



Subtract the reds.

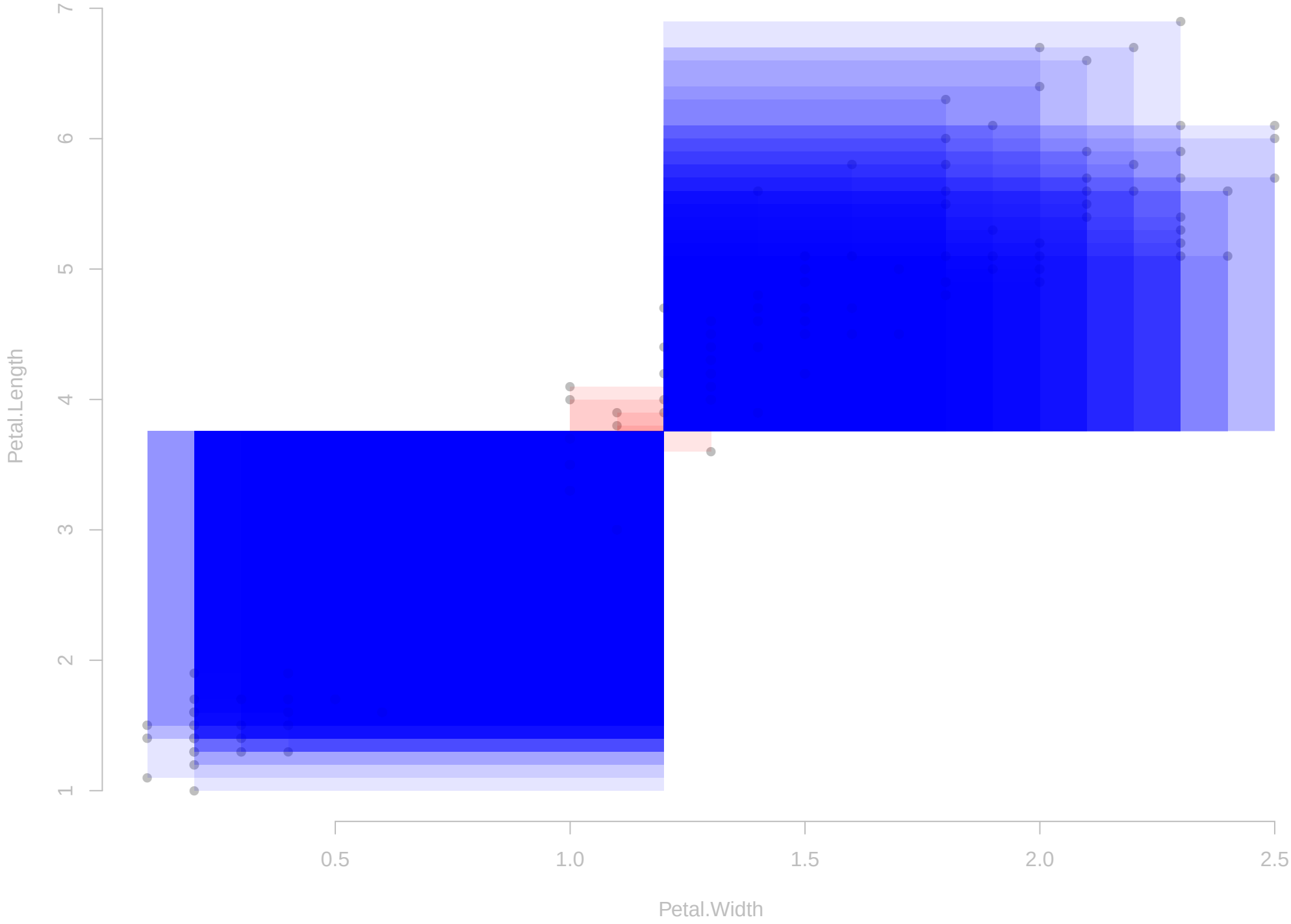


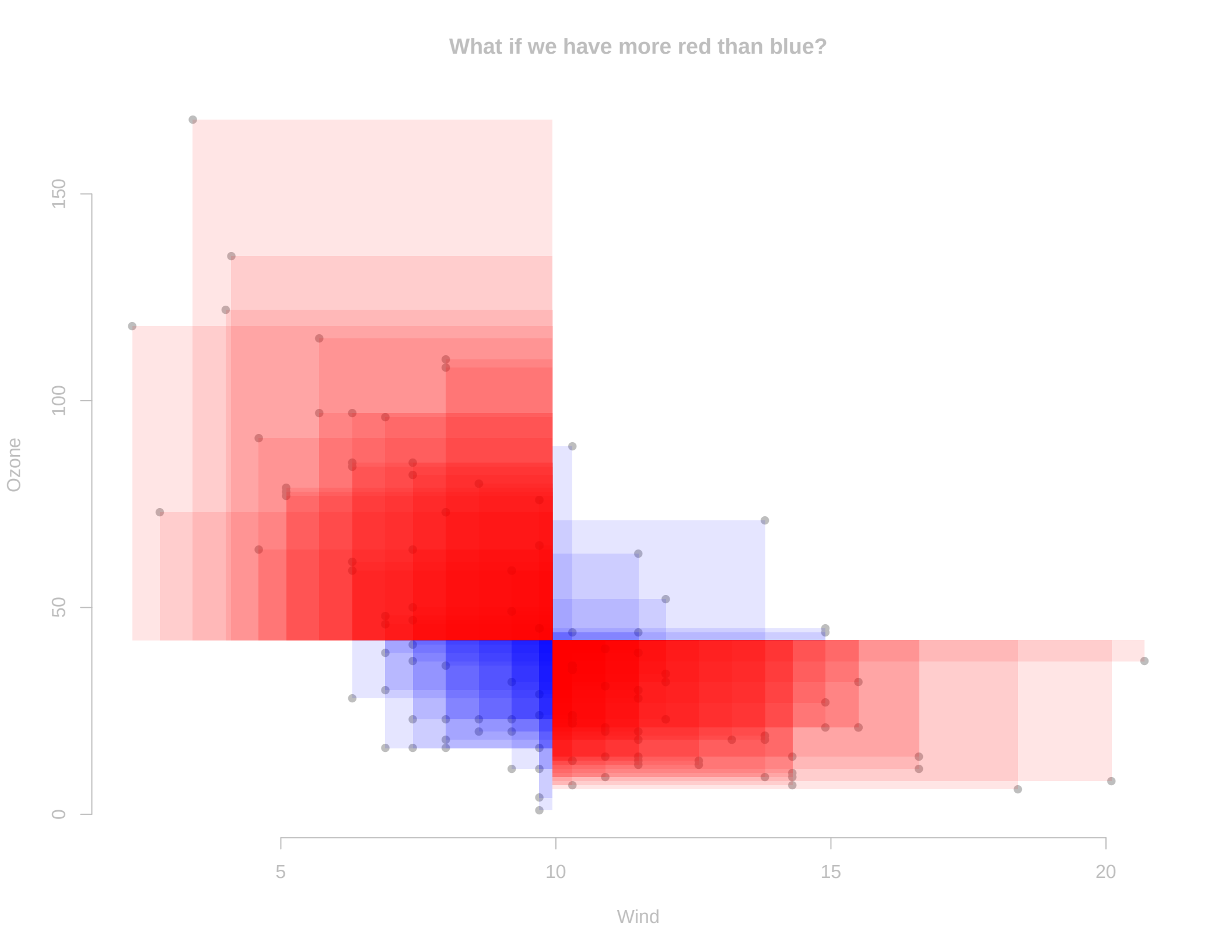
Divide into as many equal pieces as we have irises (n).



This blue sliver is the covariance.

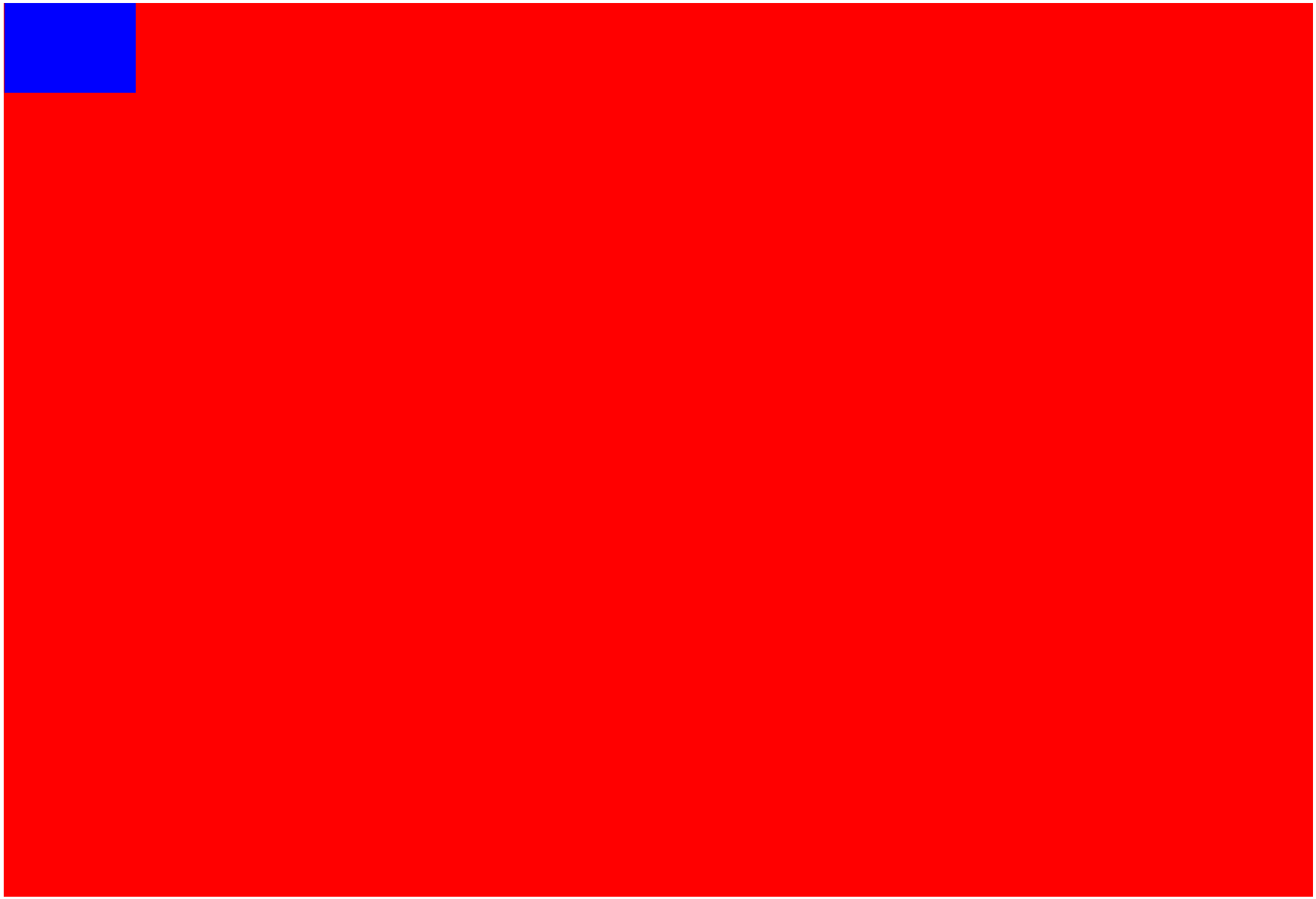
That was for this sort of relationship.



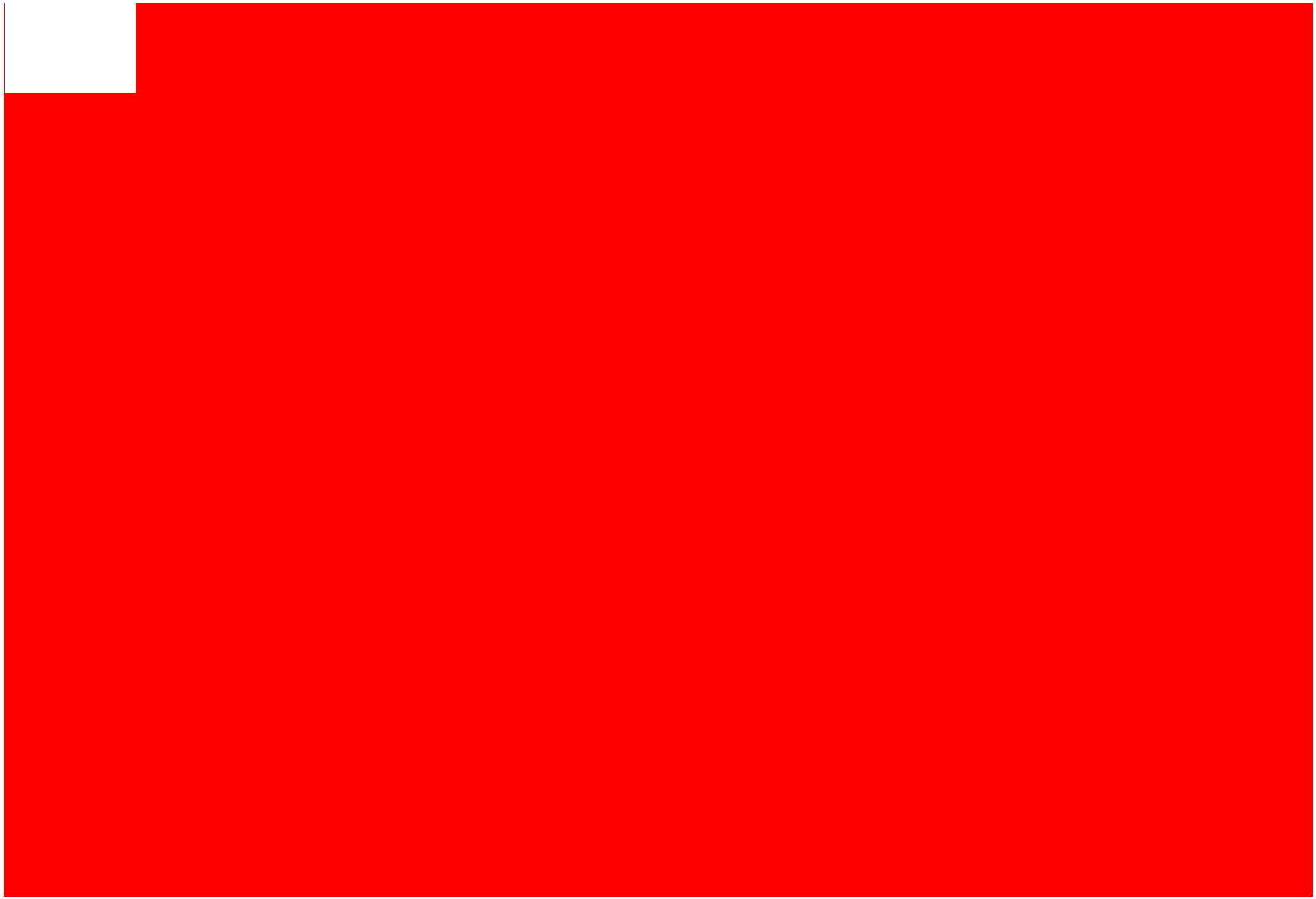

What if we have more red than blue?
Ozone
150
100
50
0
5
10
15
20
Wind

Add the blues together. (This is at a different scale.)

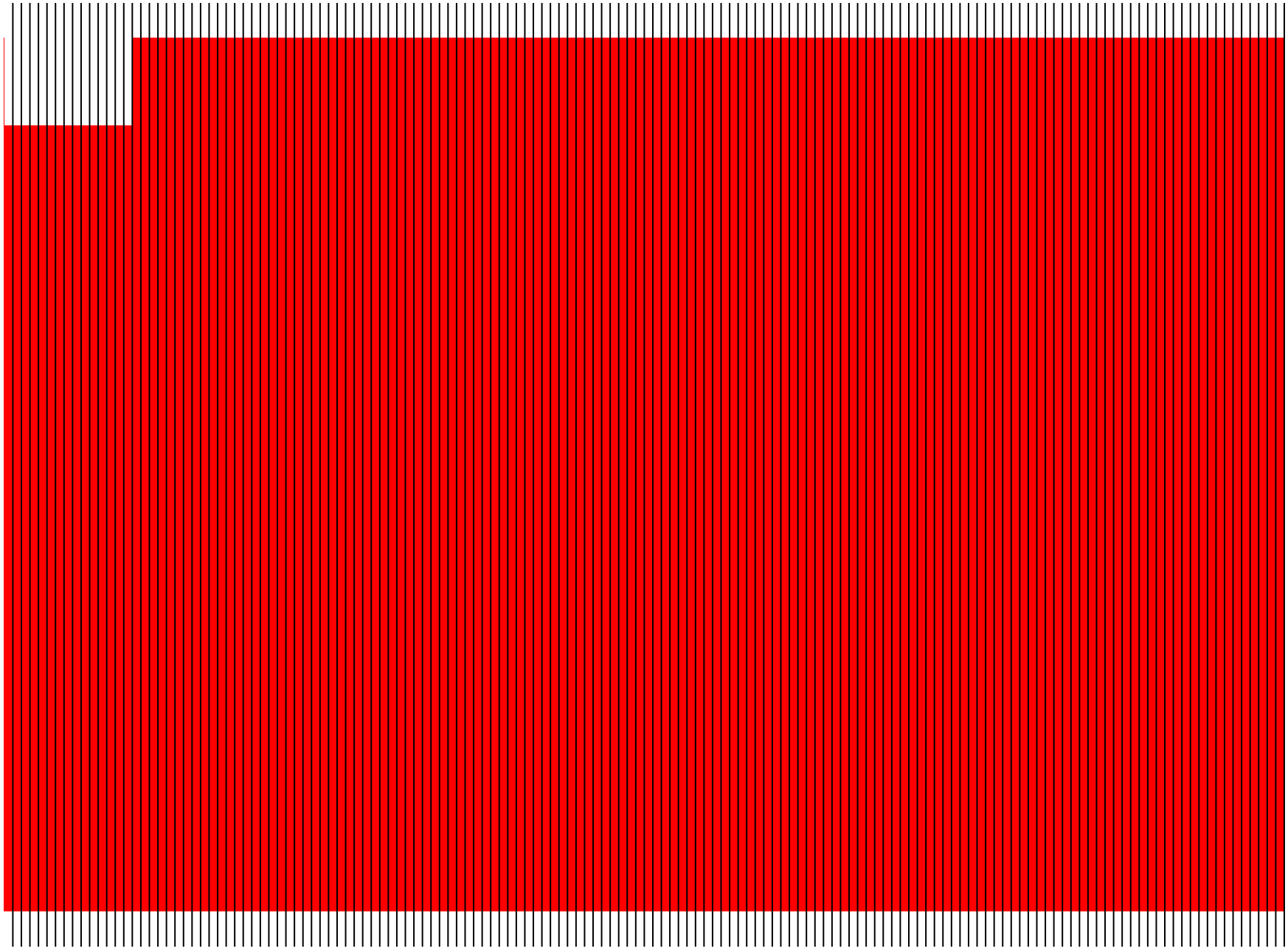
Add the reds together.



Subtract the reds.



Divide into as many equal pieces as we have irises (n).



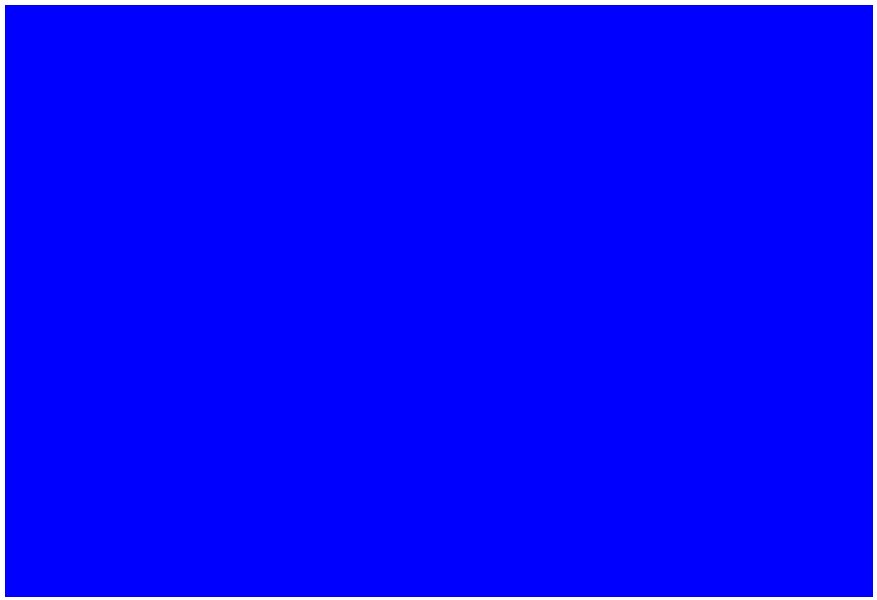
This red sliver is the covariance.

This red sliver is the covariance.

But it's negative!


What if we have as much red as blue?
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

Add the blues together. (This is at a different scale.)



Add the reds together.

Subtract the reds.

(Covariance is zero.)

Variance

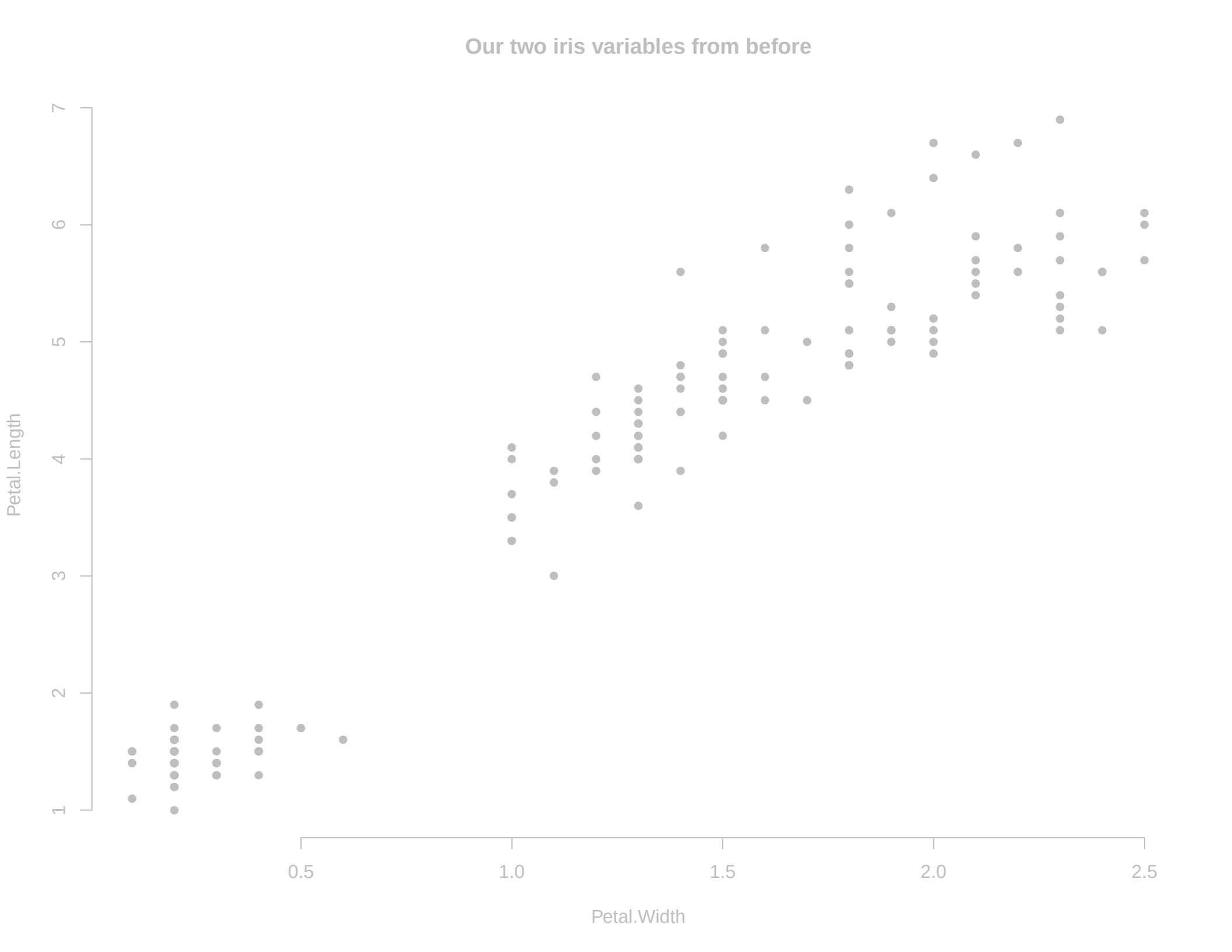
Variance tells us how spread out some numbers are.

1, 4, 8, 10

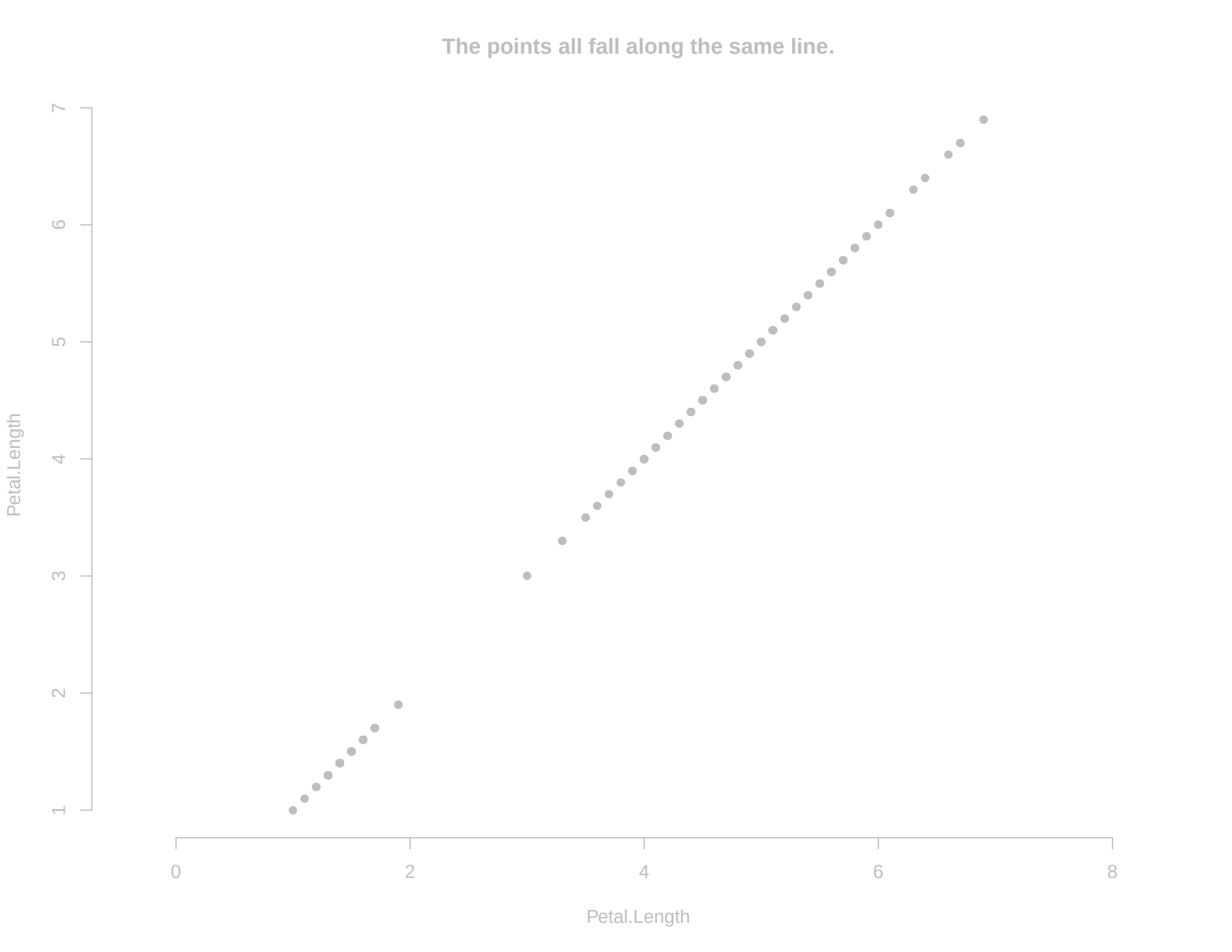
vs

4, 4, 5, 6

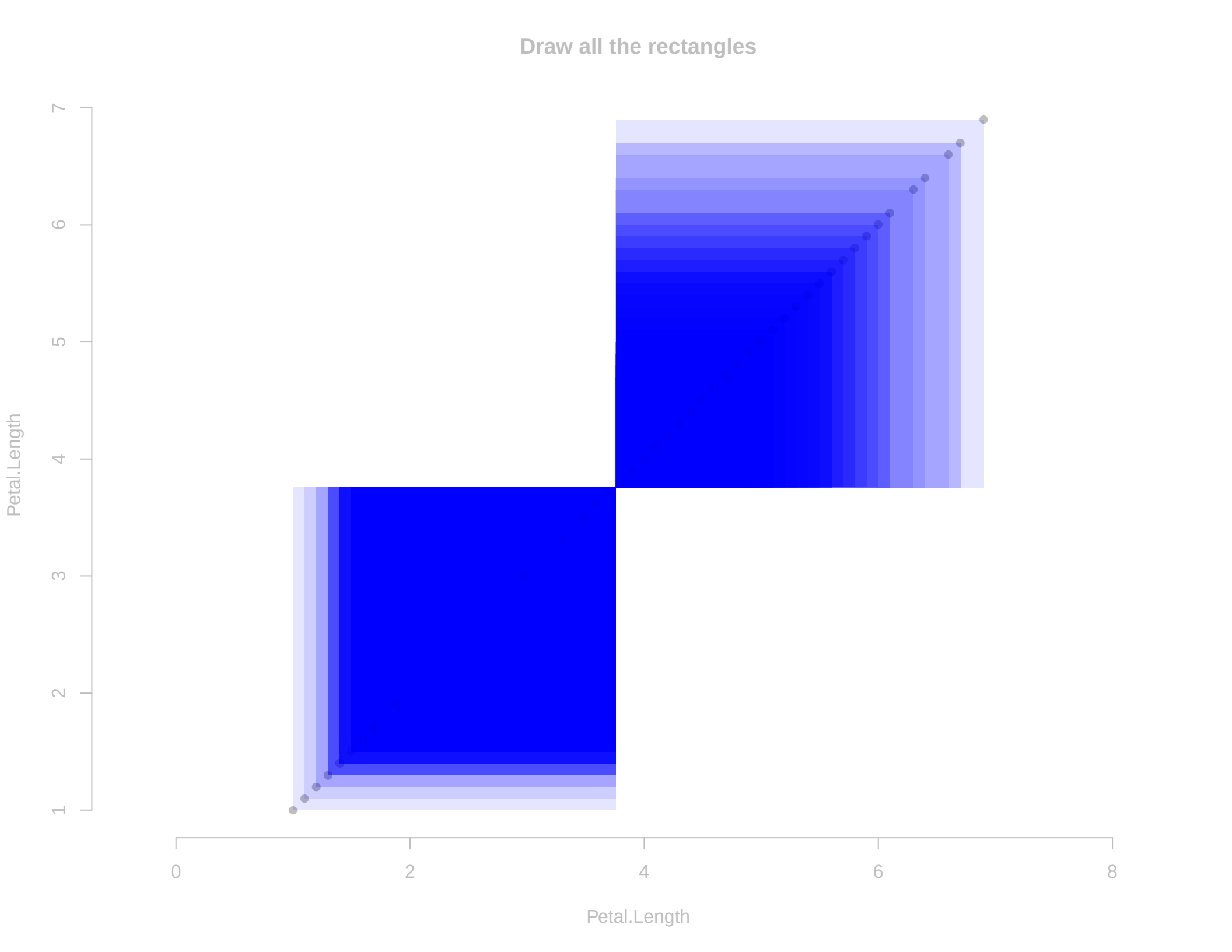
The variance of a variable is the covariance of the variable with itself.

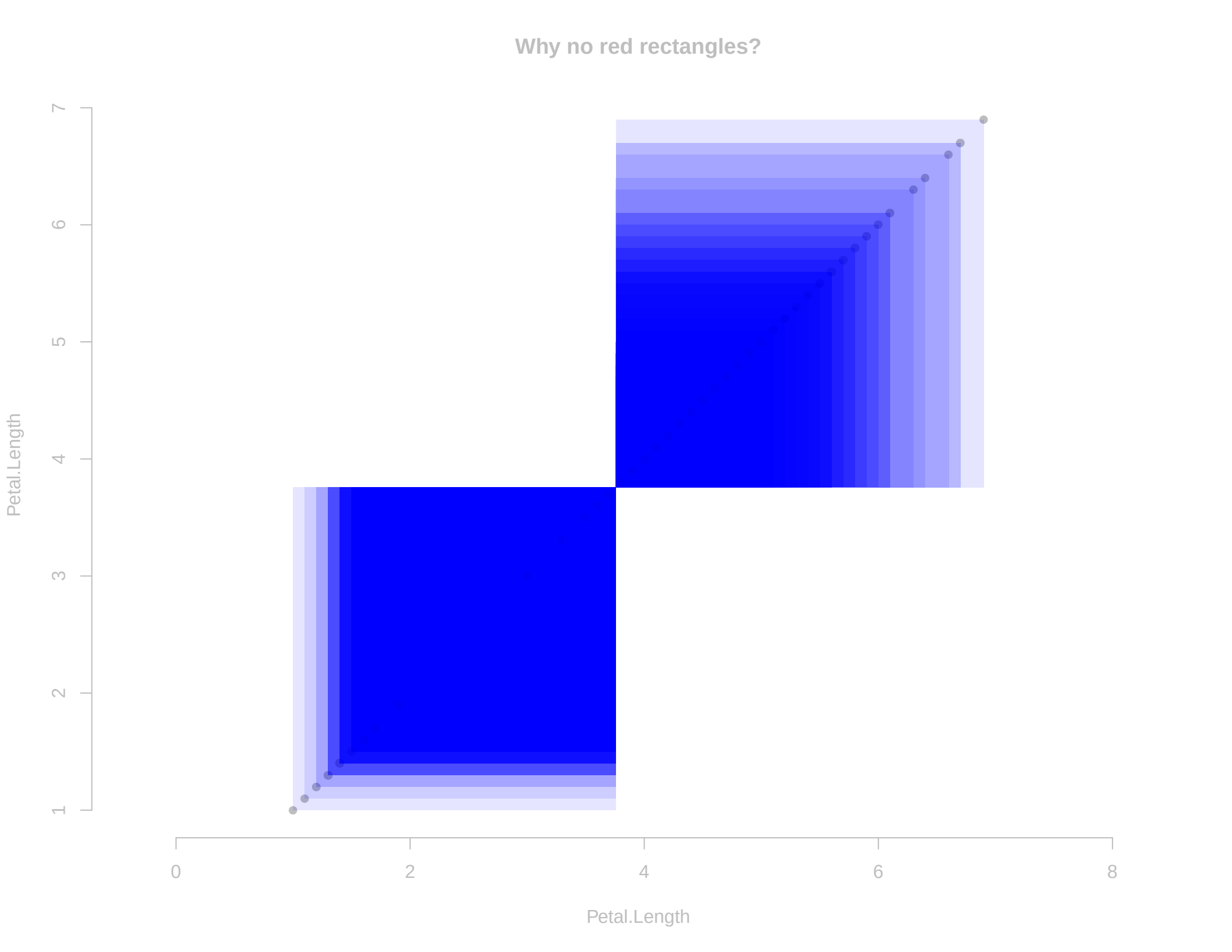

Our two iris variables from before
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

Let's look at just one of them.

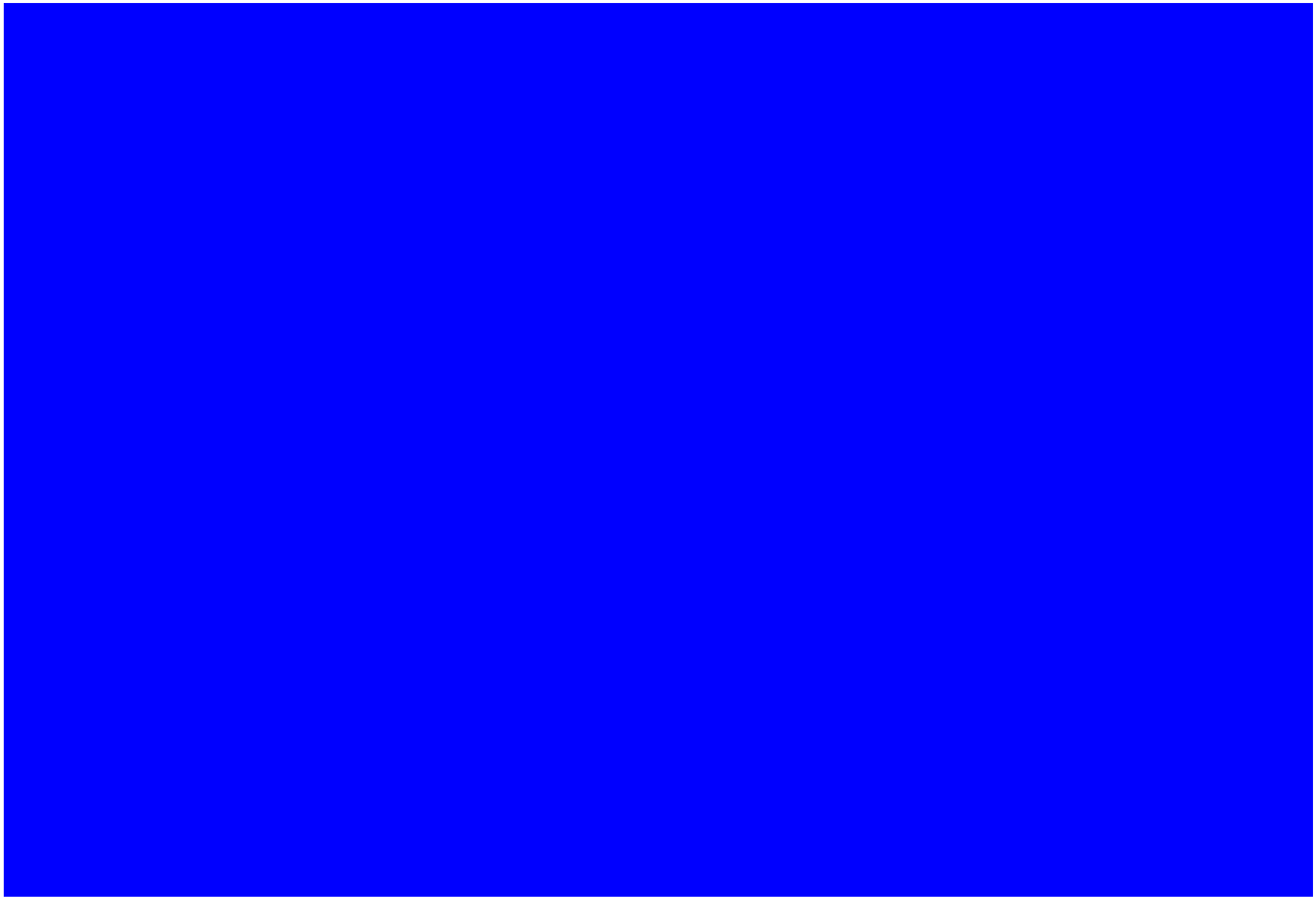

The points all fall along the same line.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

Let's find the variance of Petal.Length

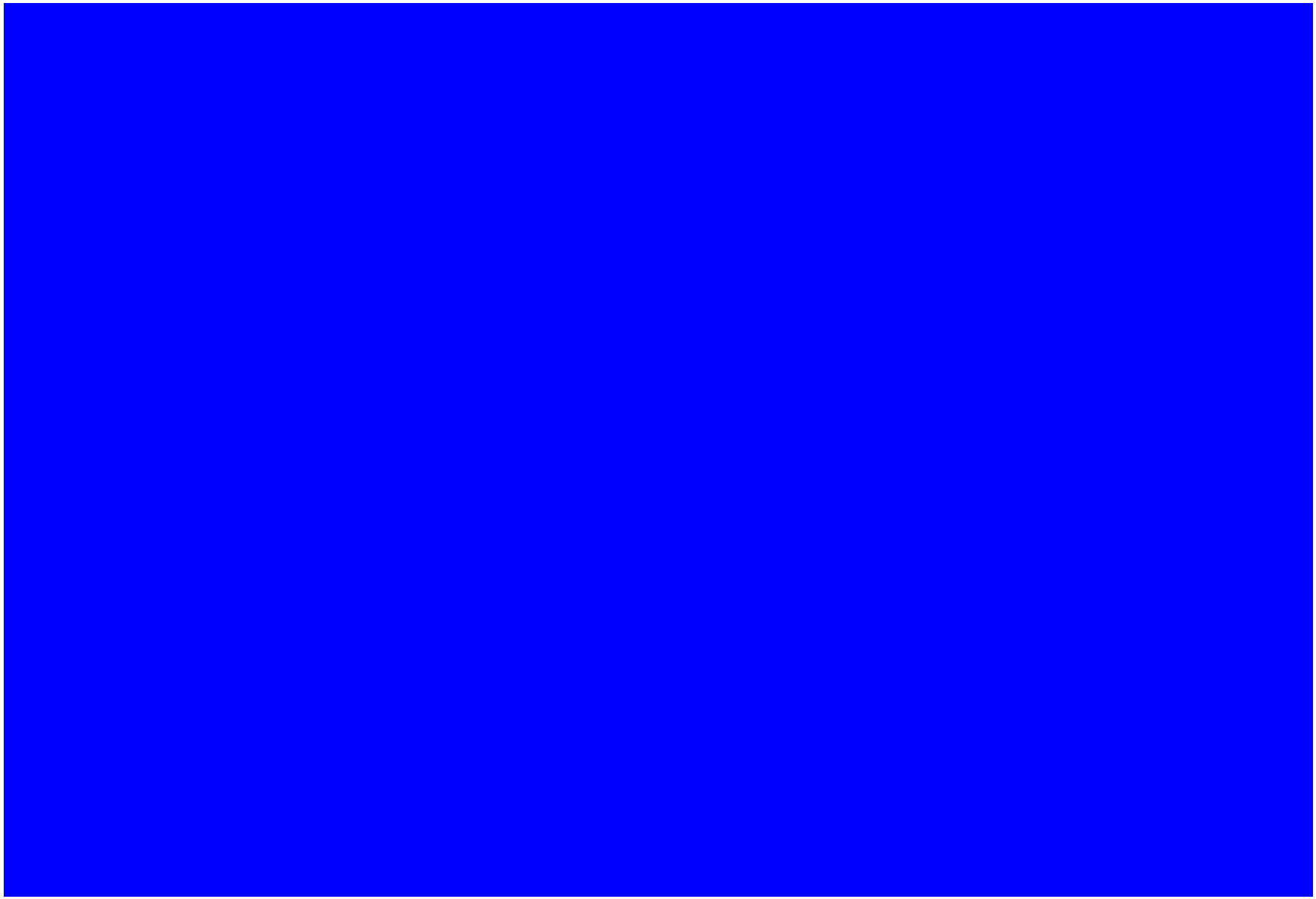

Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length


Why no red rectangles?
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

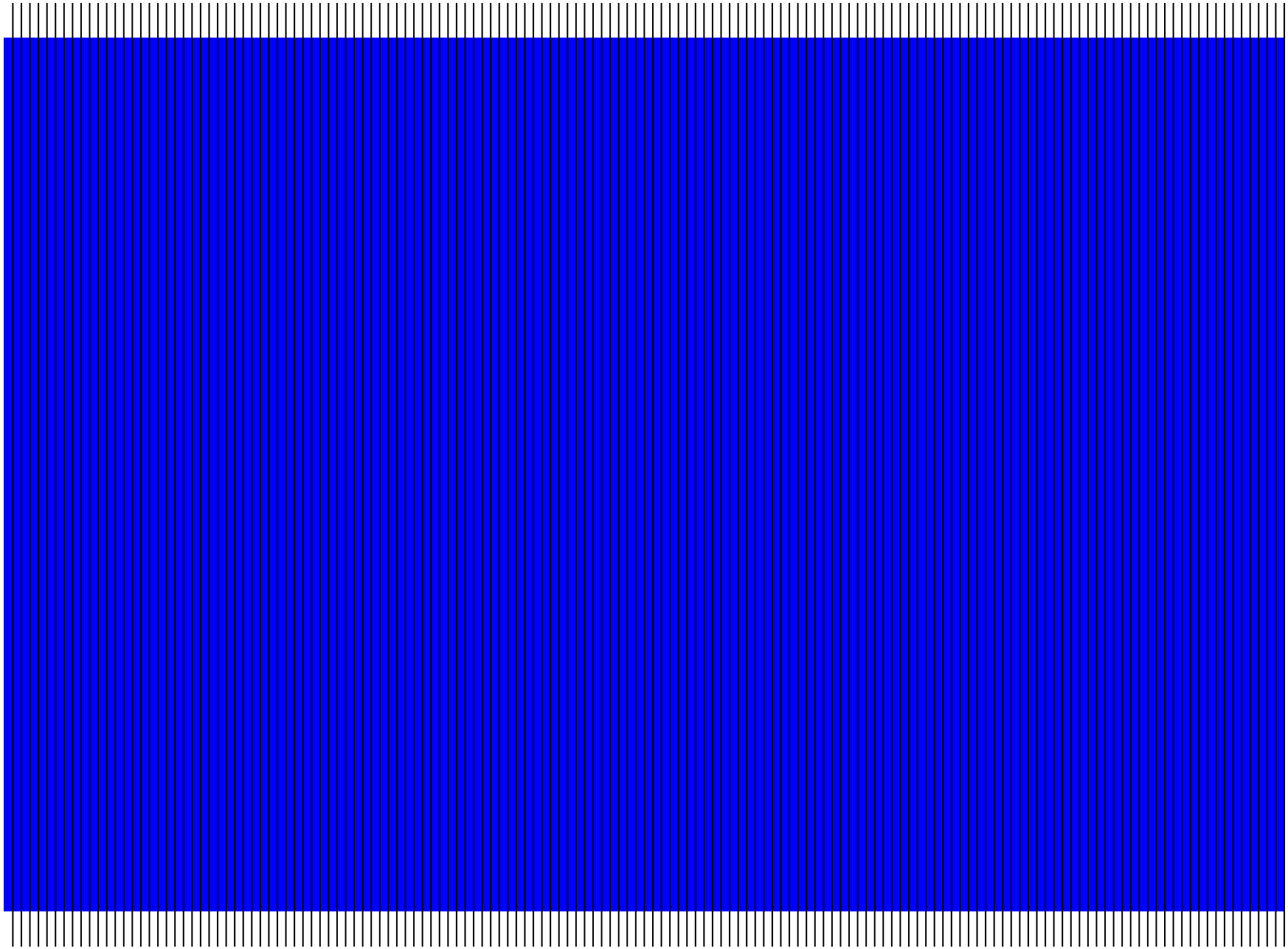
Add the blues together. (This is at a different scale.)



We have no reds to subtract.



Divide into as many equal pieces as we have irises (n).



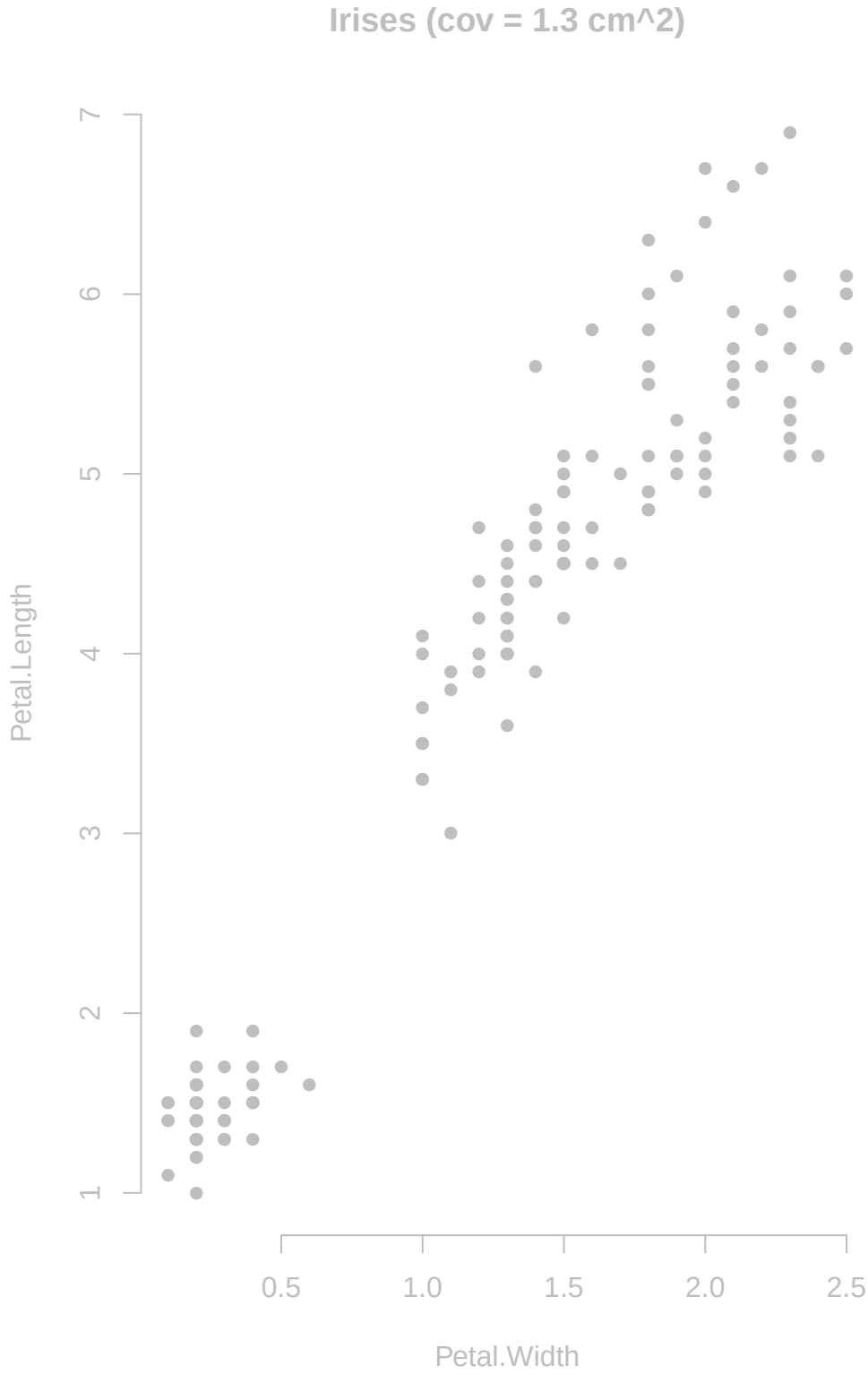
This blue sliver is the variance.

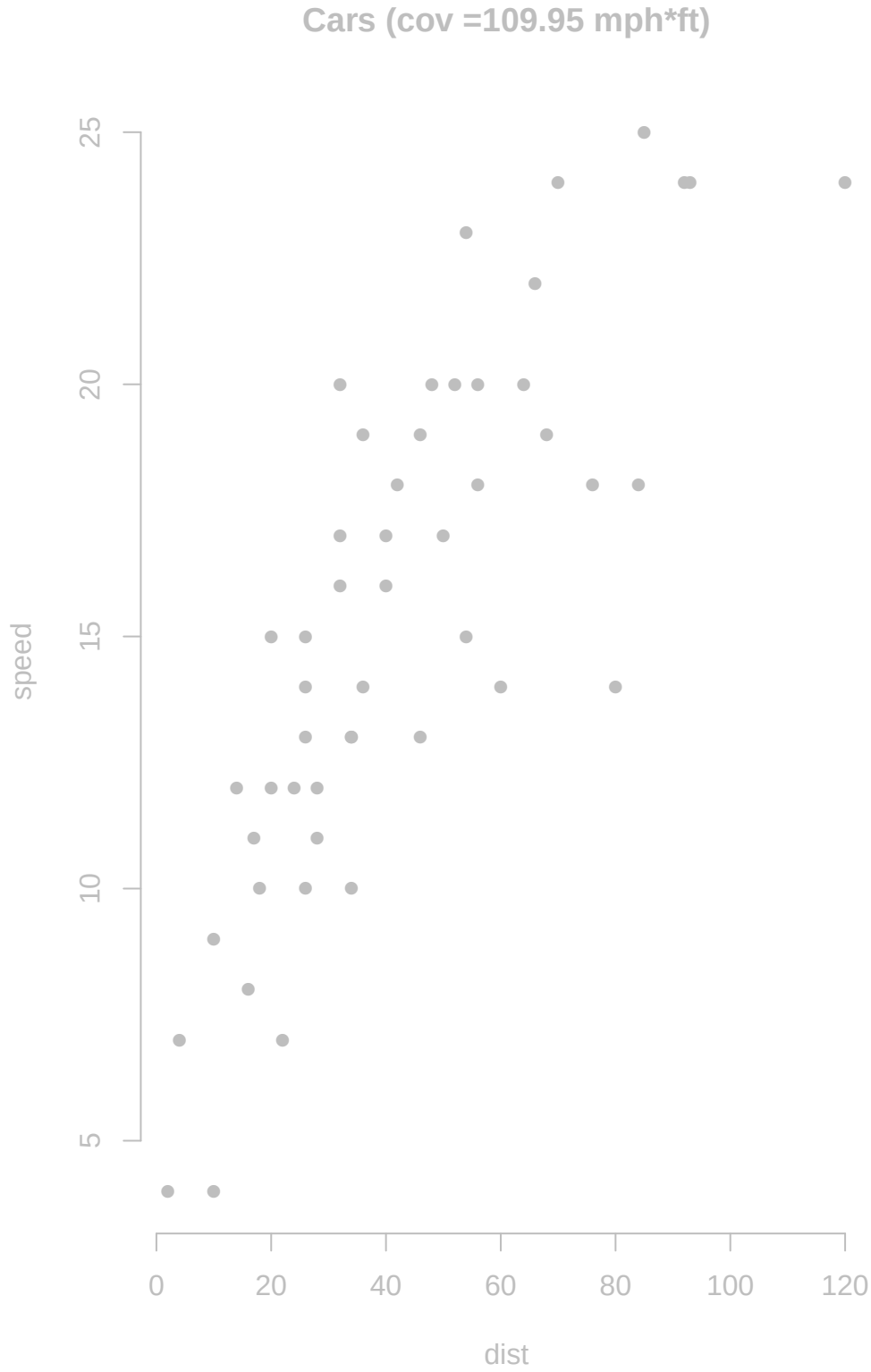
A problem with covariance

Covariance has units!

(x−unit times y−unit)

Which relationship is stronger (more linear)?

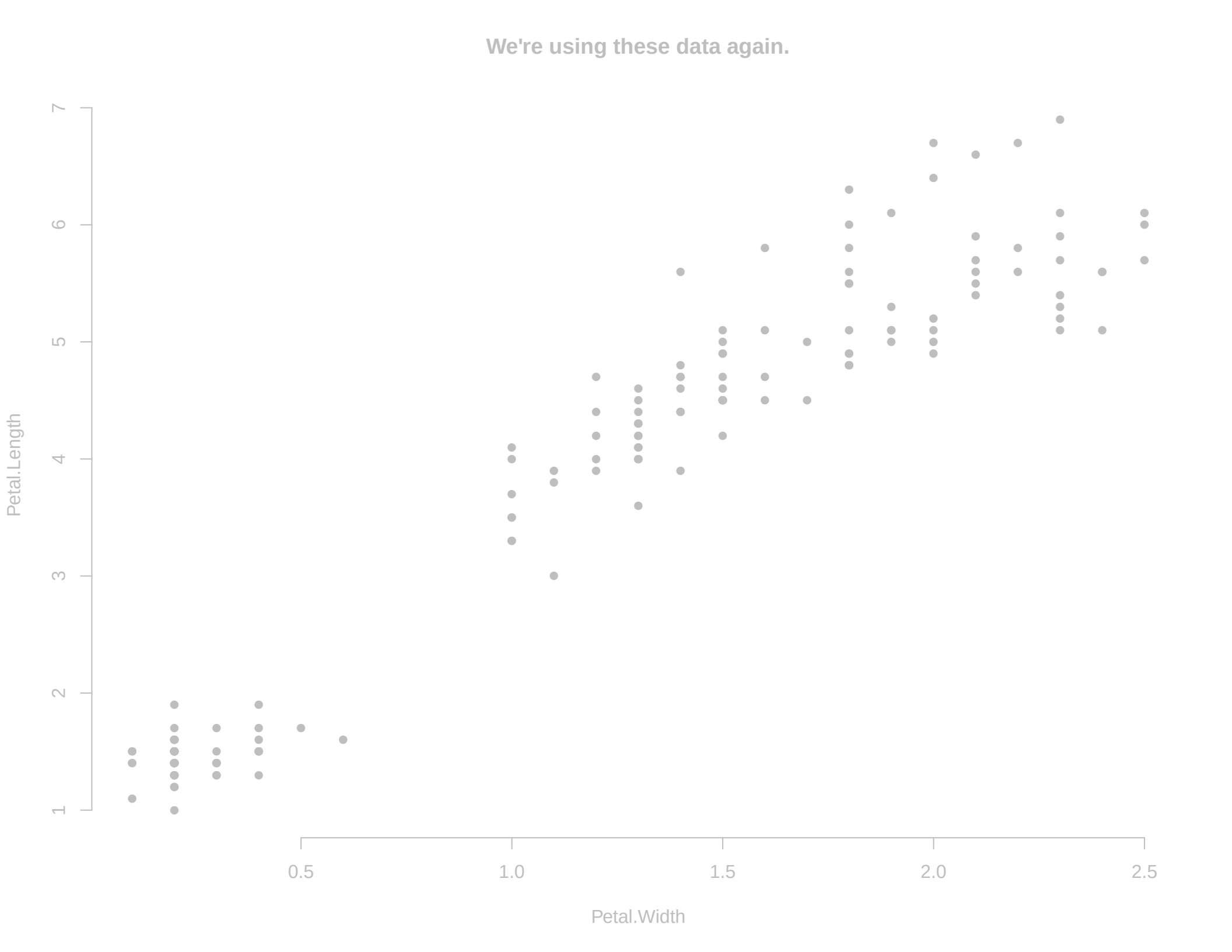

Irises (cov = 1.3 cm^2)
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

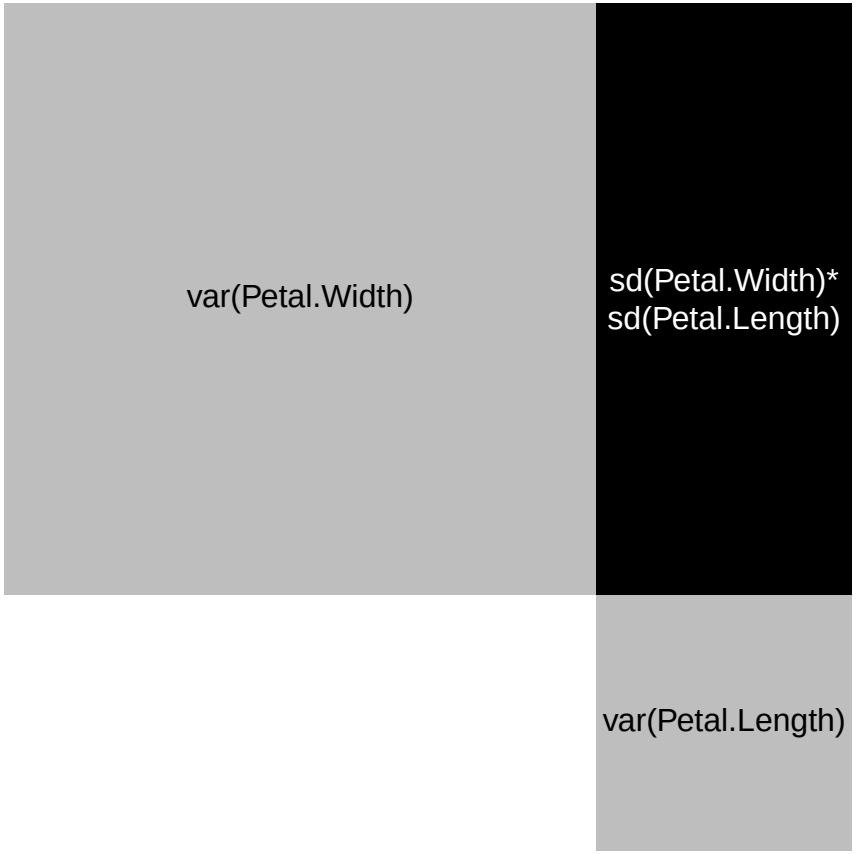

Cars (cov =109.95 mph*ft)
speed
dist
5
10
15
20
25
0
20
40
60
80
100
120

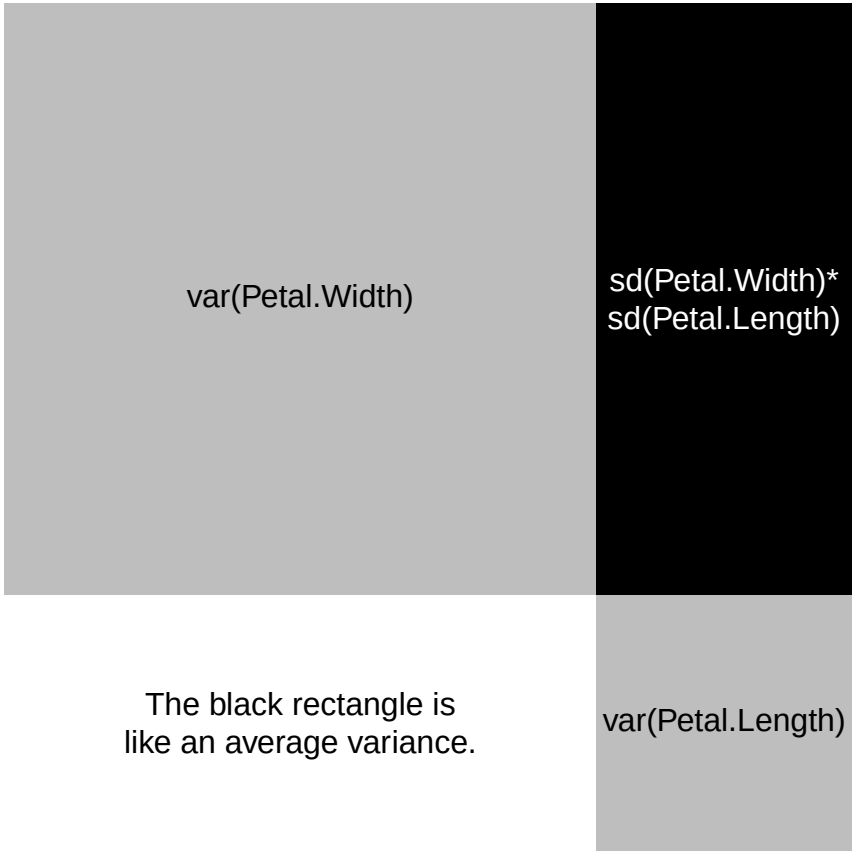
Oh noes!

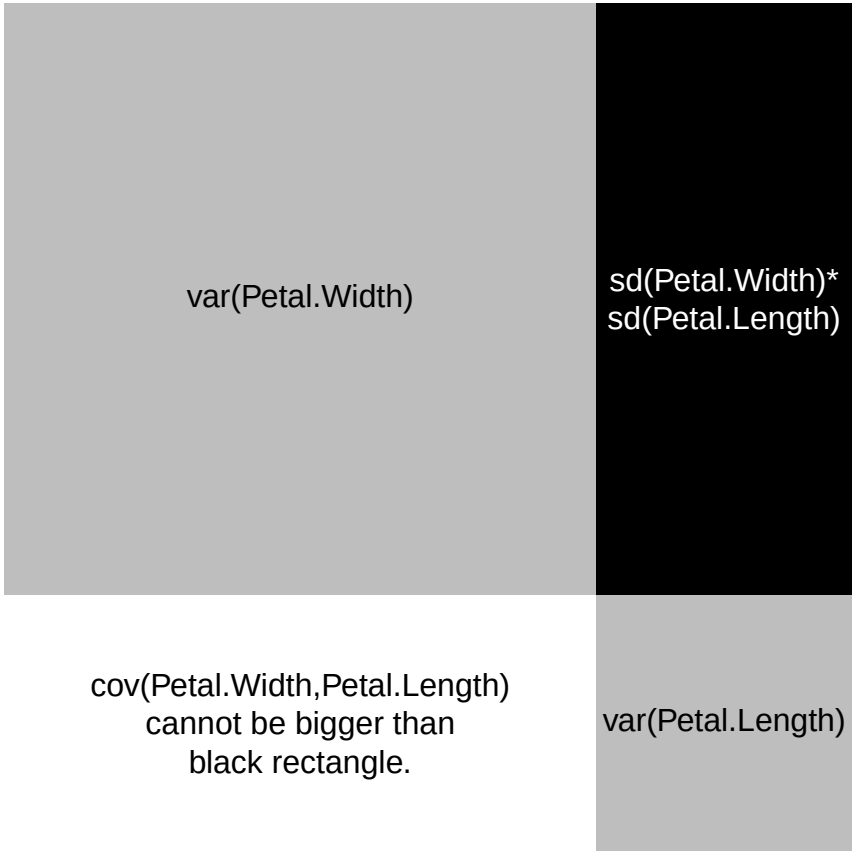
We can divide the covariance by the variances to standardize it.

We're using these data again.

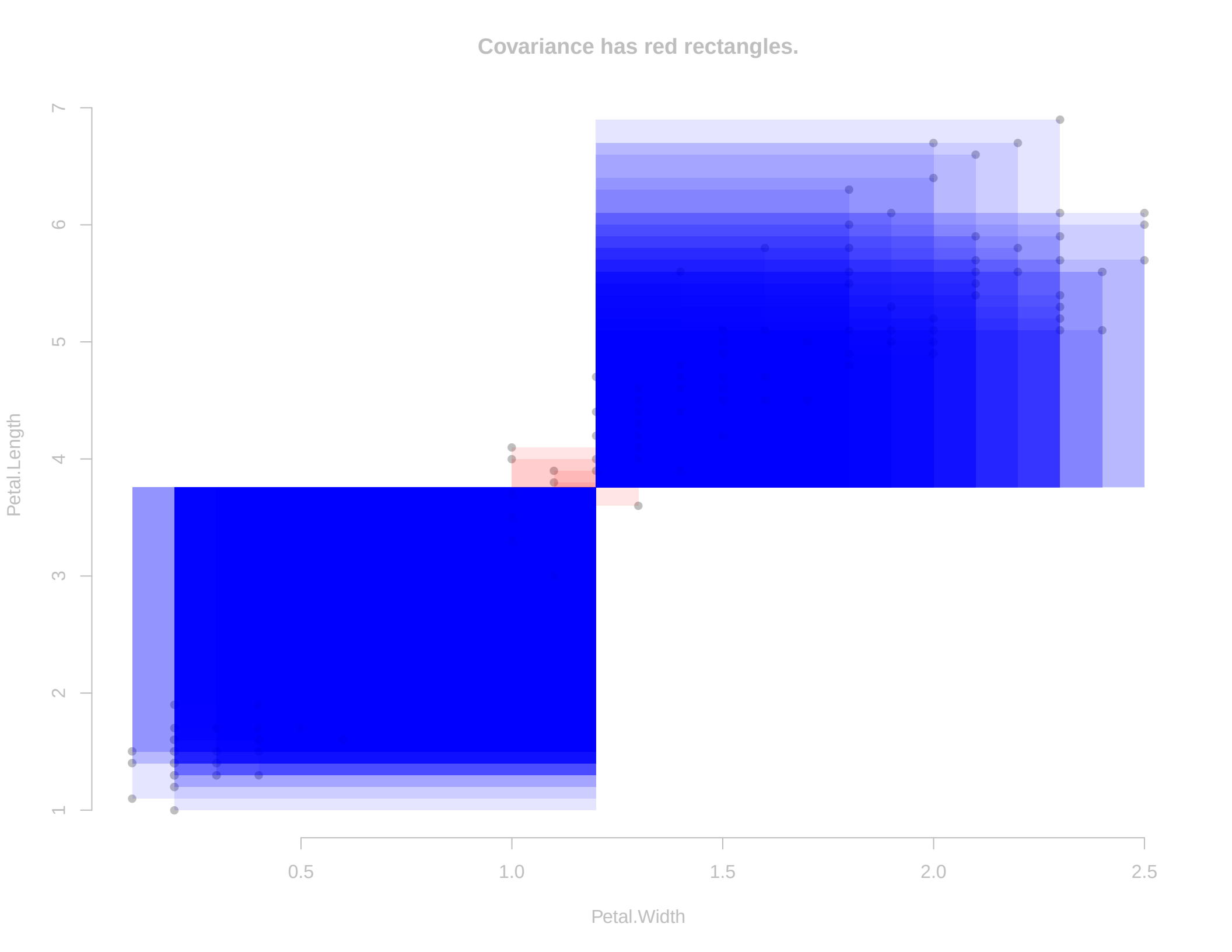


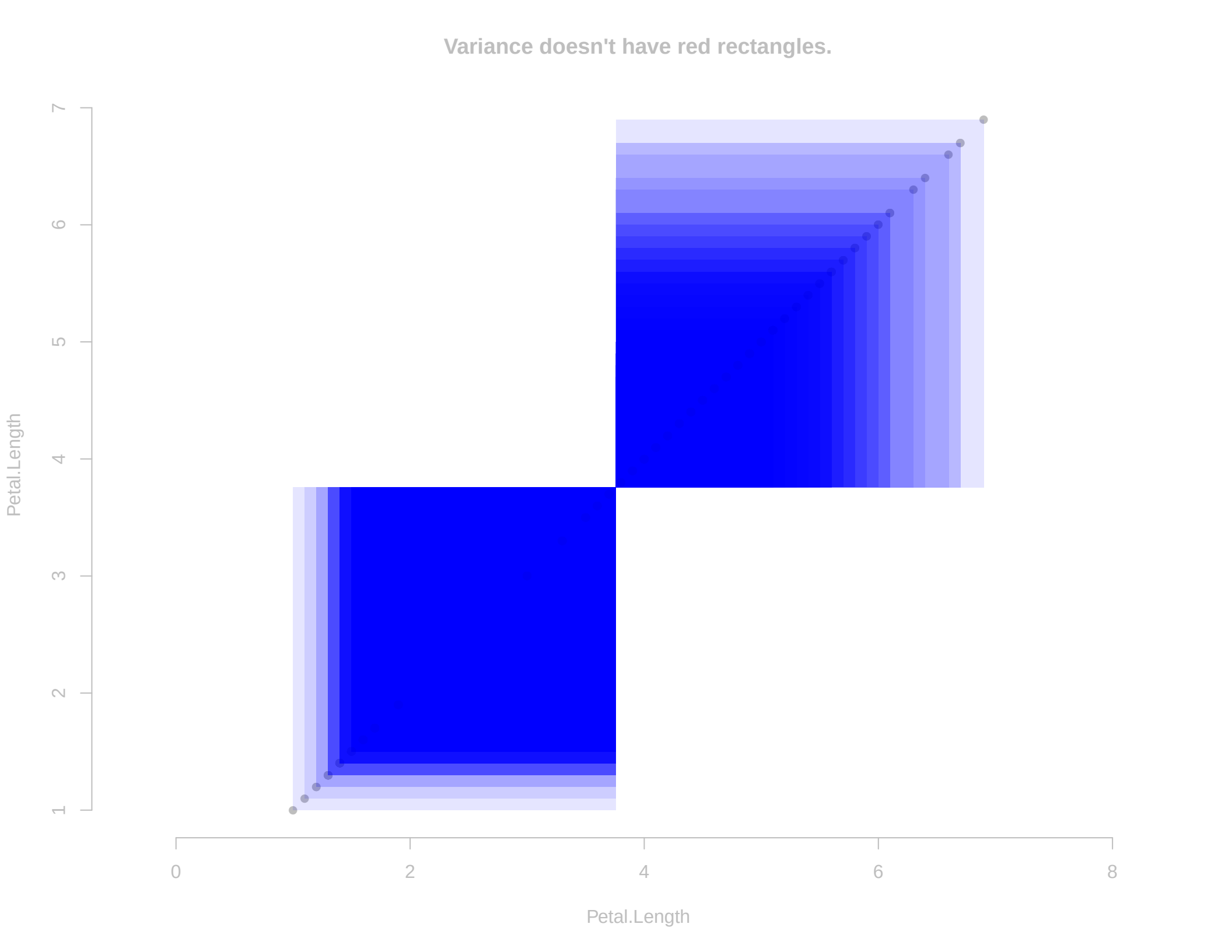

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
var(Petal.Length)

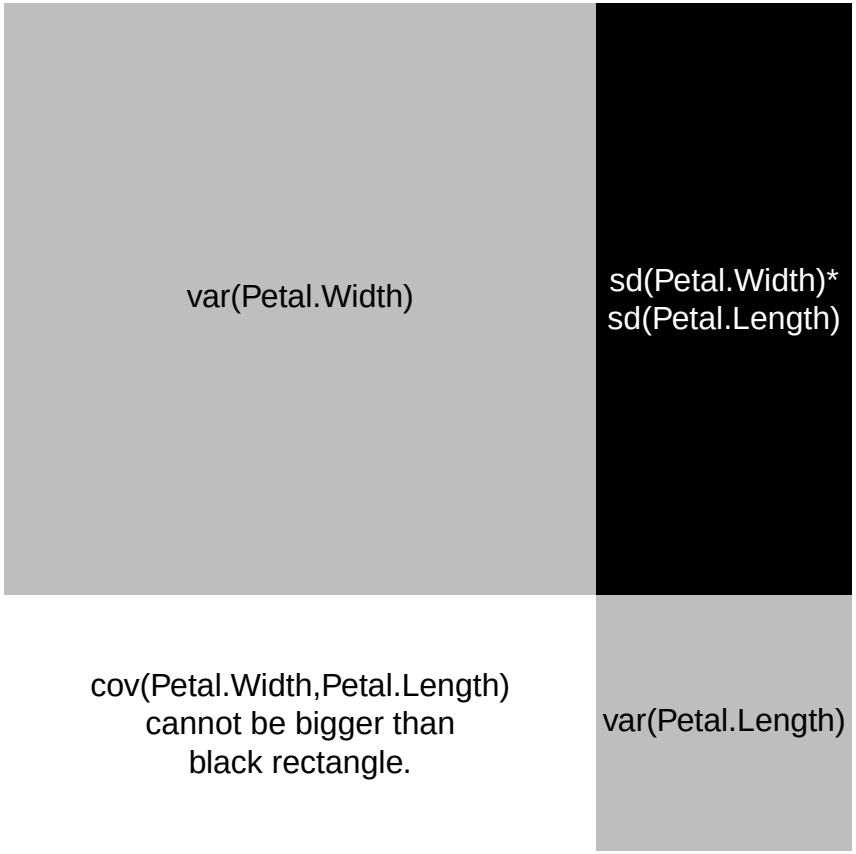

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
The black rectangle is
like an average variance.
var(Petal.Length)


var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
cov(Petal.Width,Petal.Length)
cannot be bigger than
black rectangle.
var(Petal.Length)

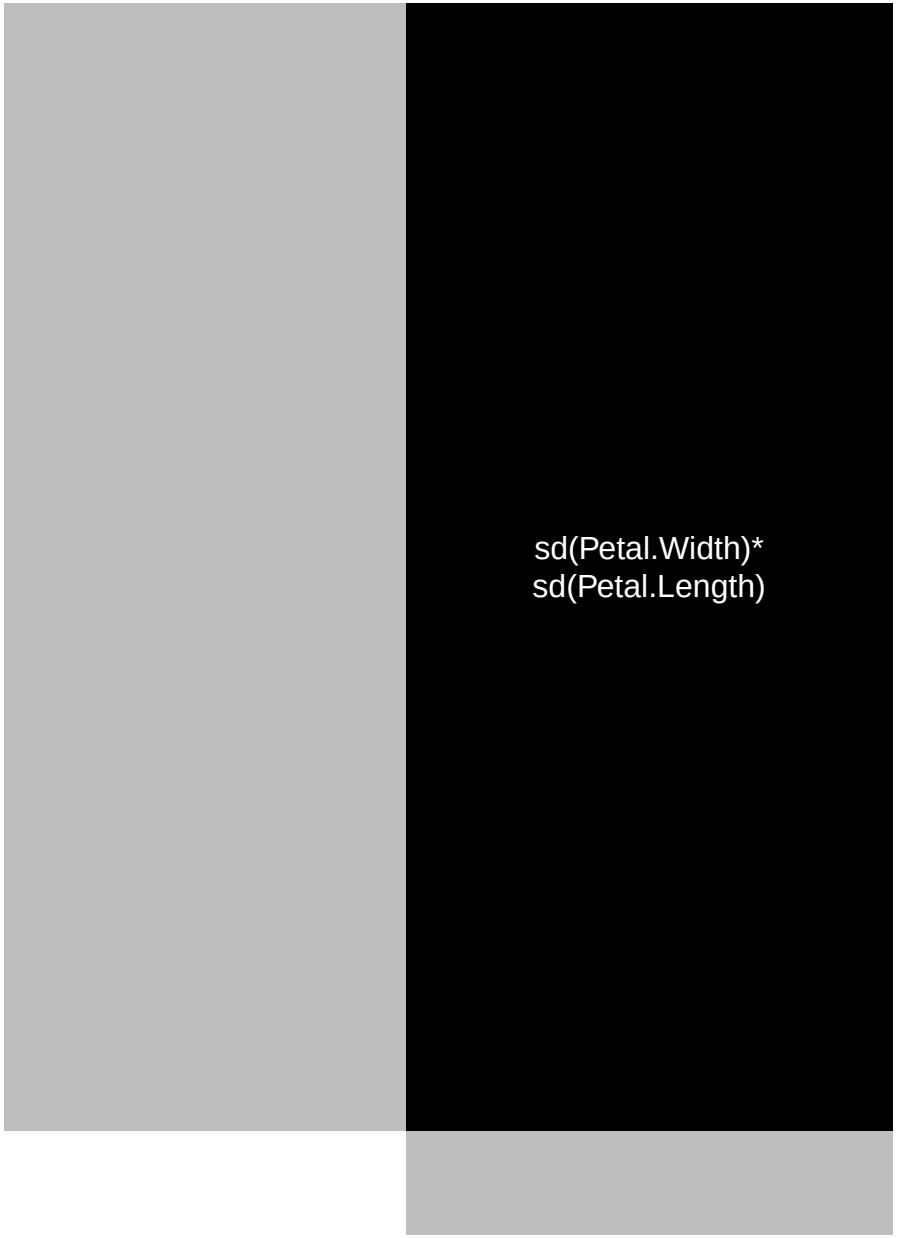
Why?


Covariance has red rectangles.
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

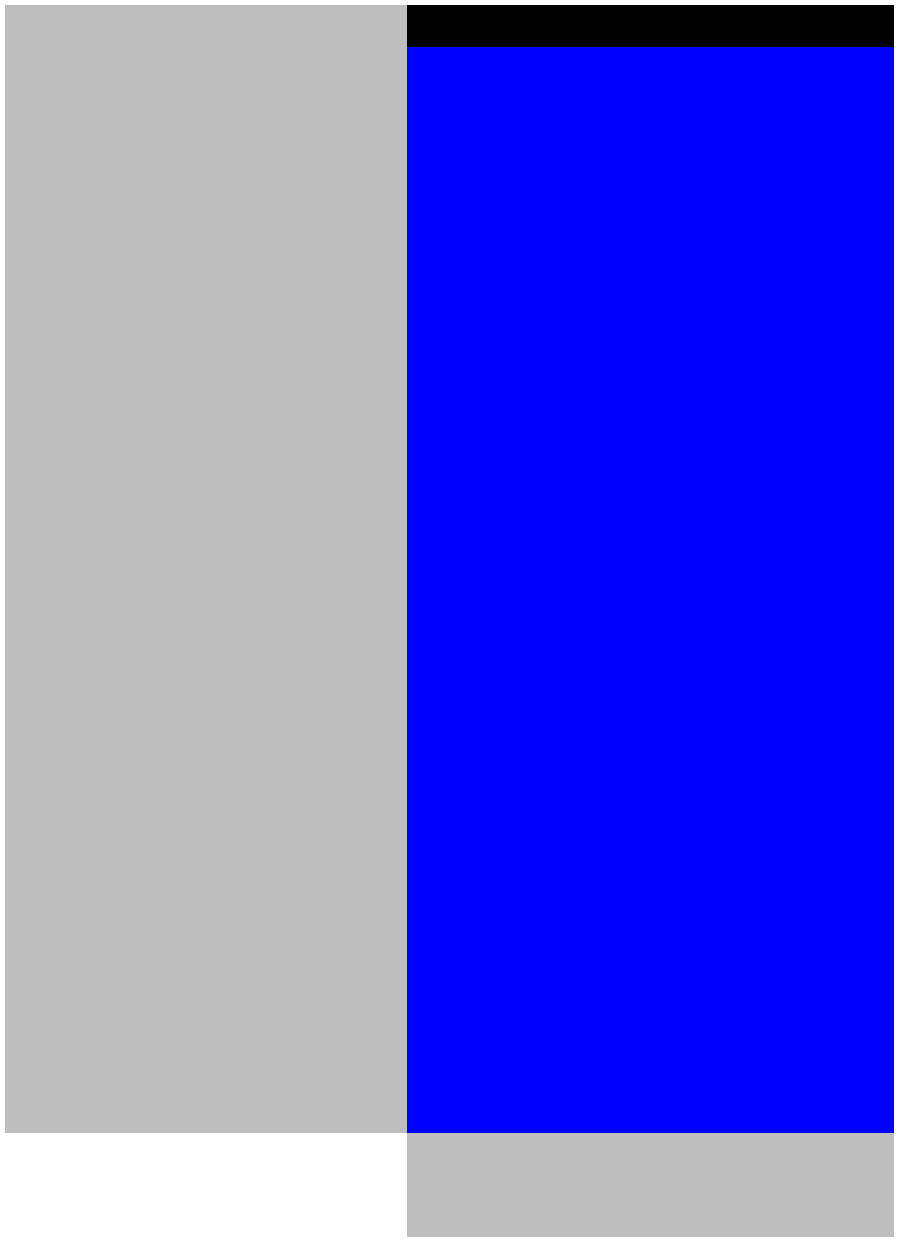

Variance doesn't have red rectangles.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length


var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
cov(Petal.Width,Petal.Length)
cannot be bigger than
black rectangle.
var(Petal.Length)

Let's zoom in.

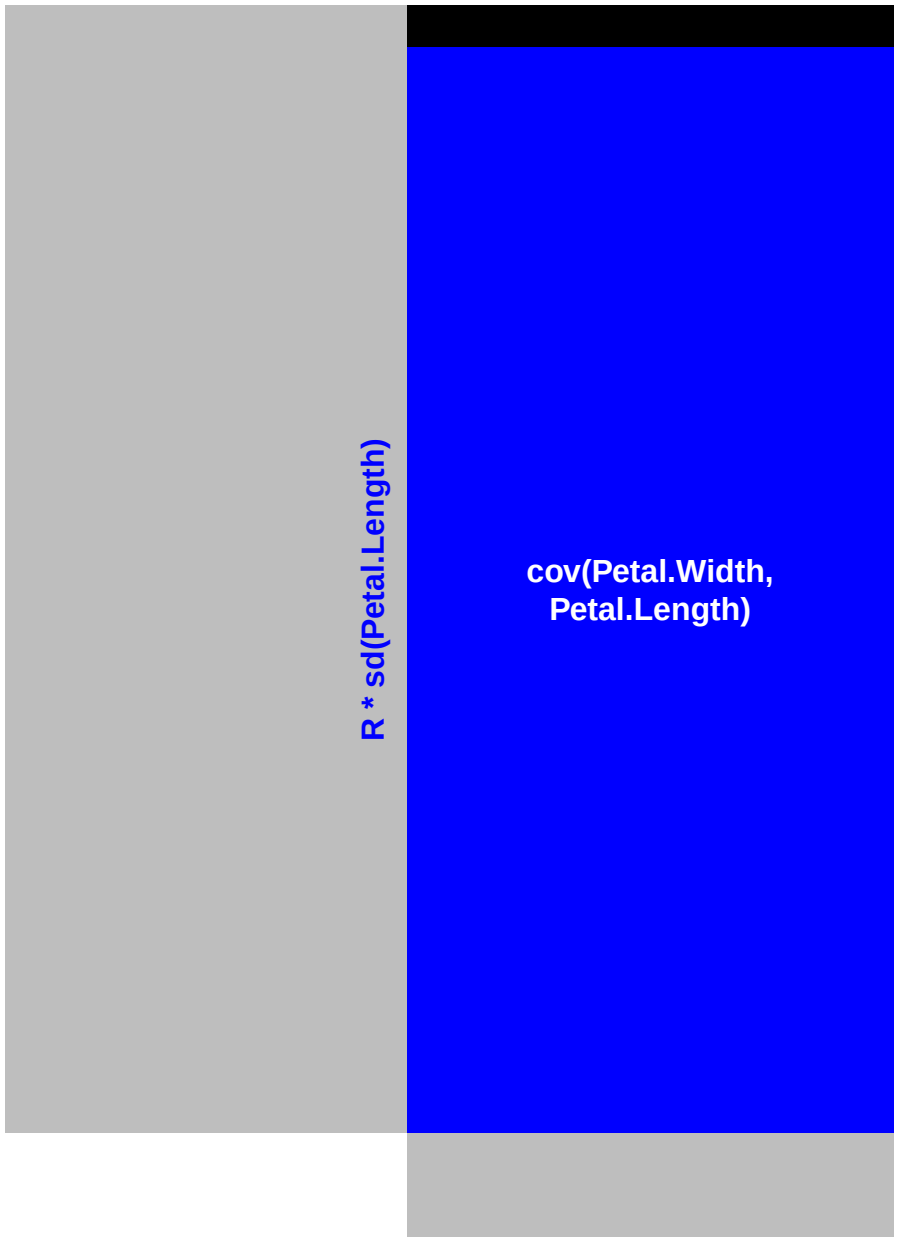


Squish covariance vertically into the rectangle.

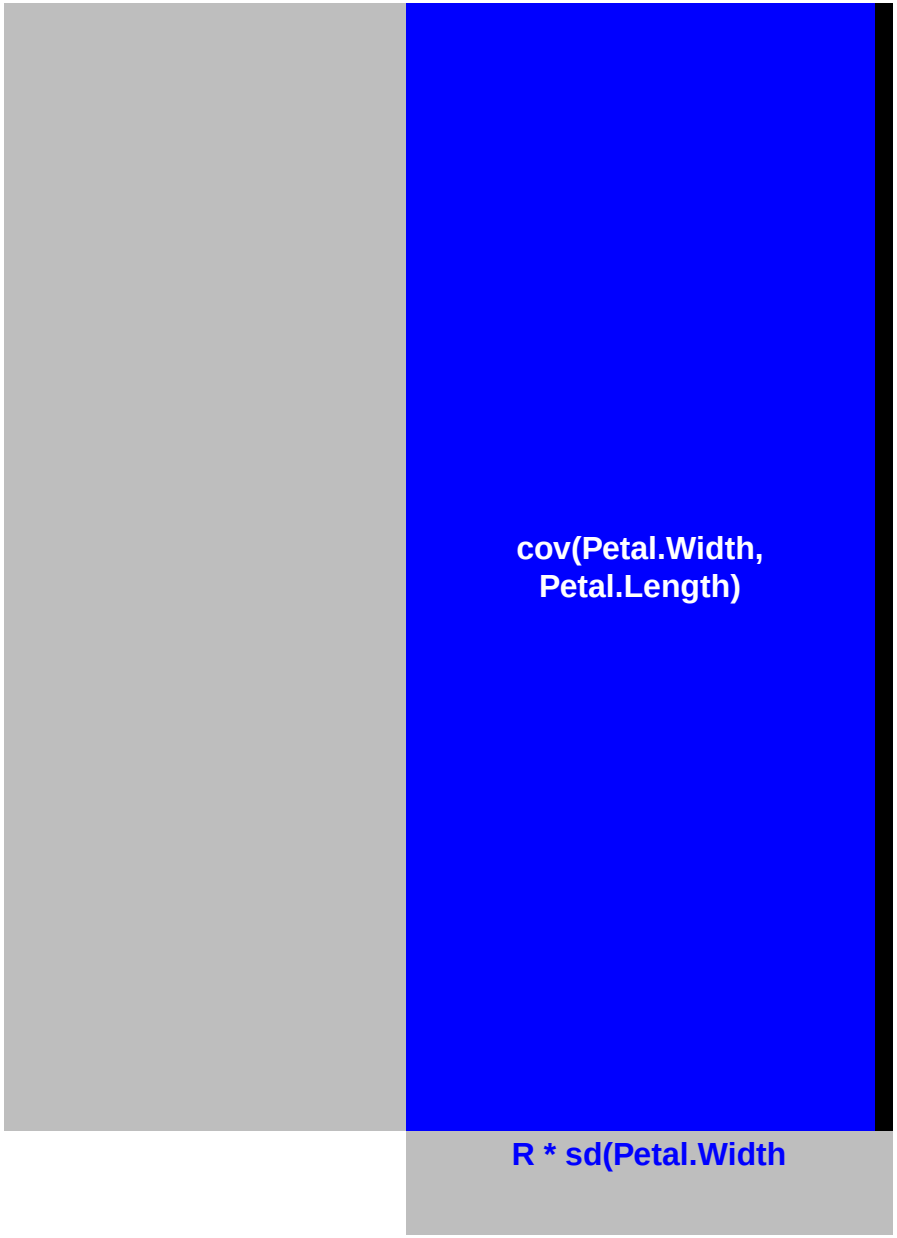


Correlation (R)
is the ratio of
the small rectangle
to the big rectangle.

Squish covariance vertically into the rectangle.

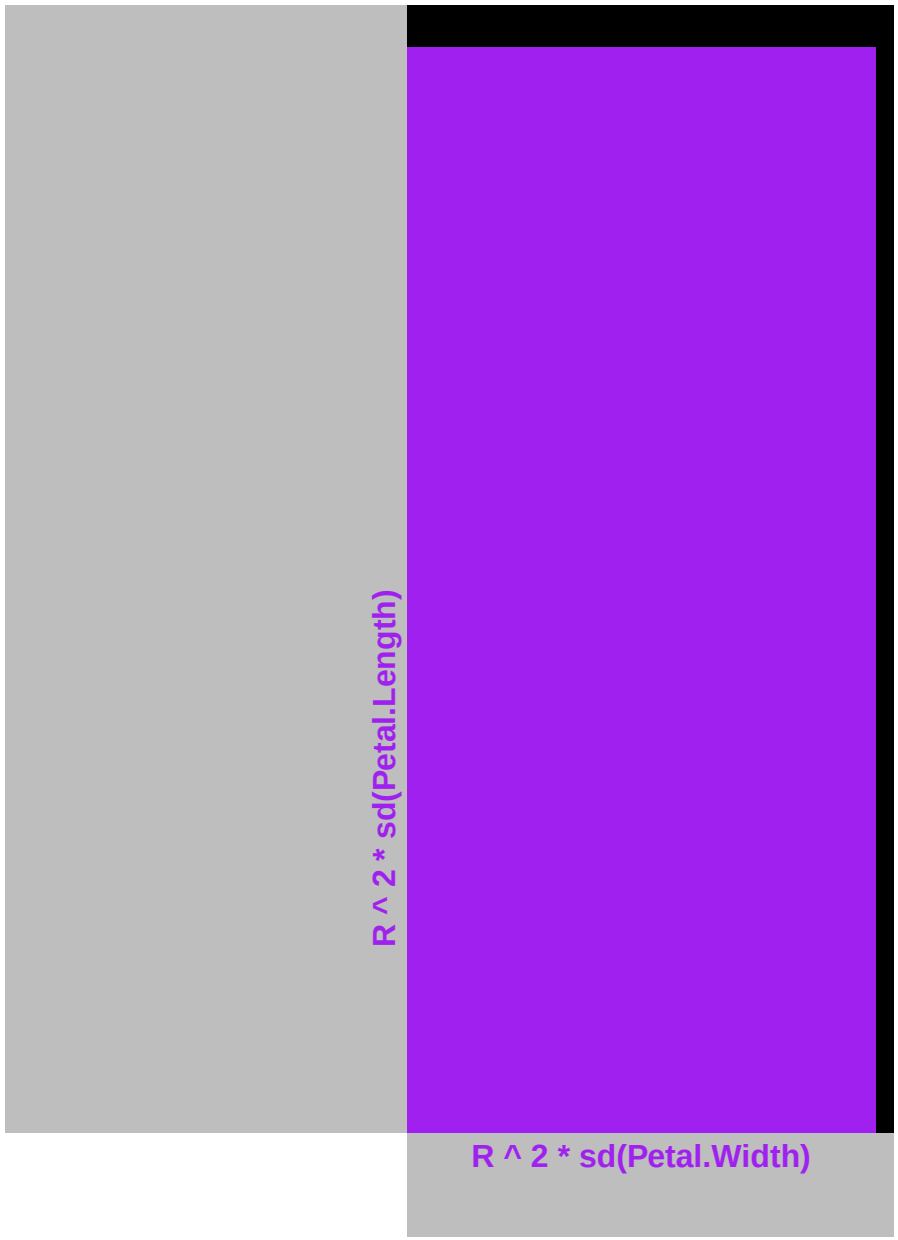


Squish covariance horizontally into the rectangle.

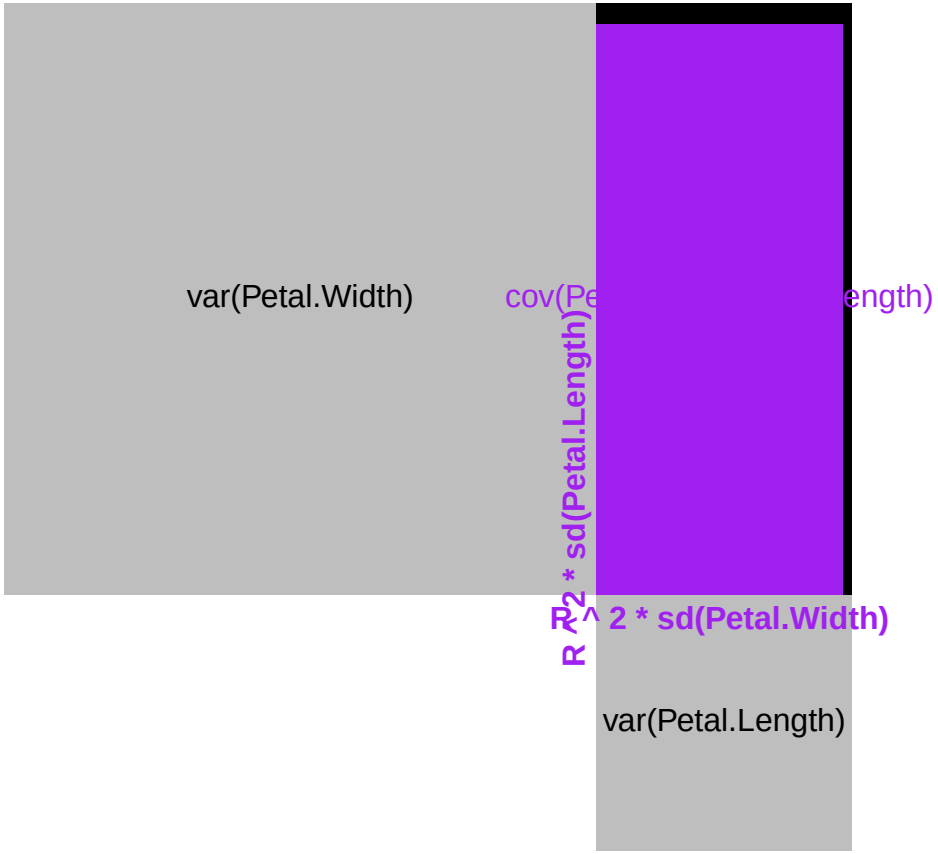


People like to
talk about R−squared.

Intersect the two squished covariance rectangles.



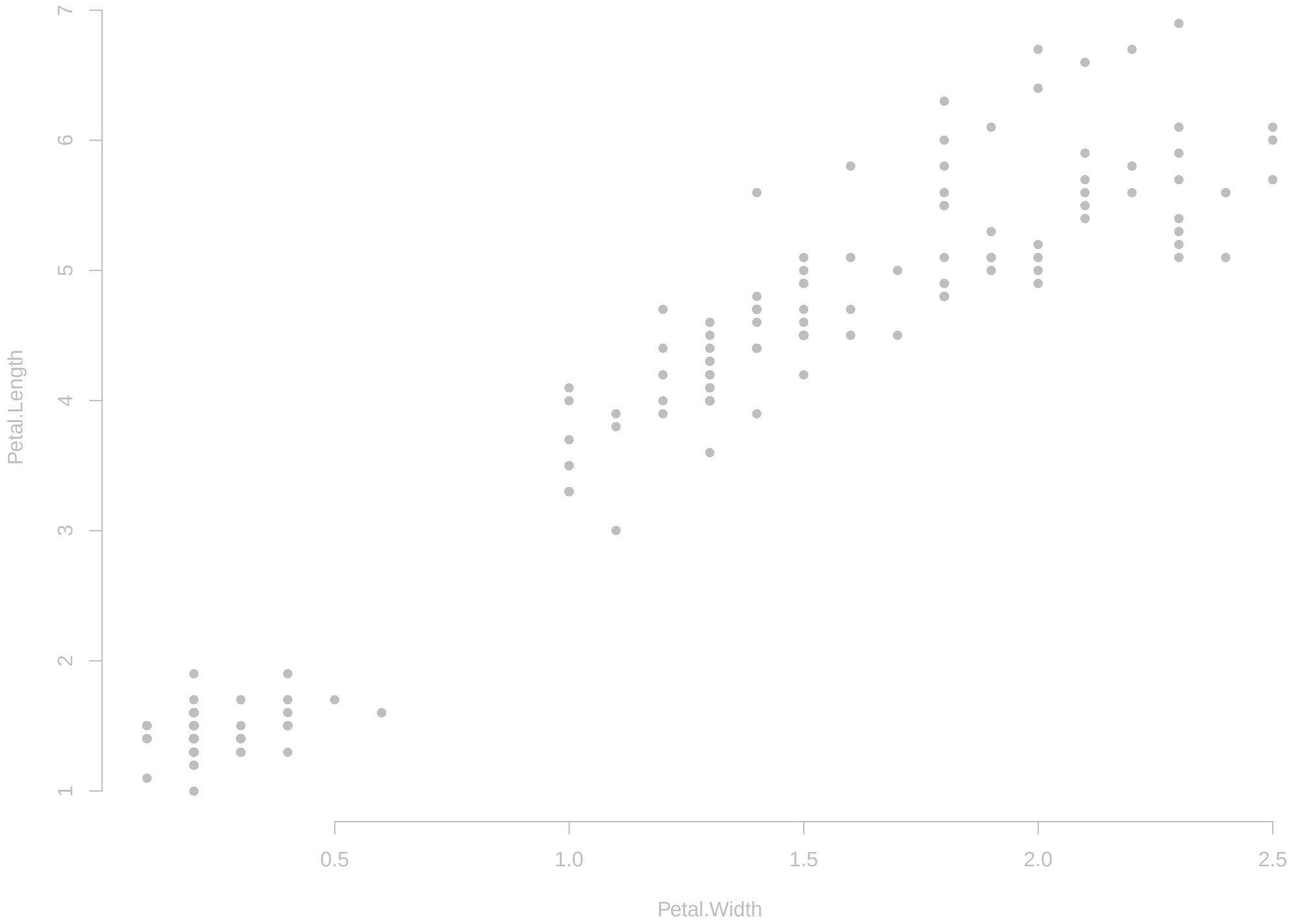
Intersect the two squished covariance rectangles.

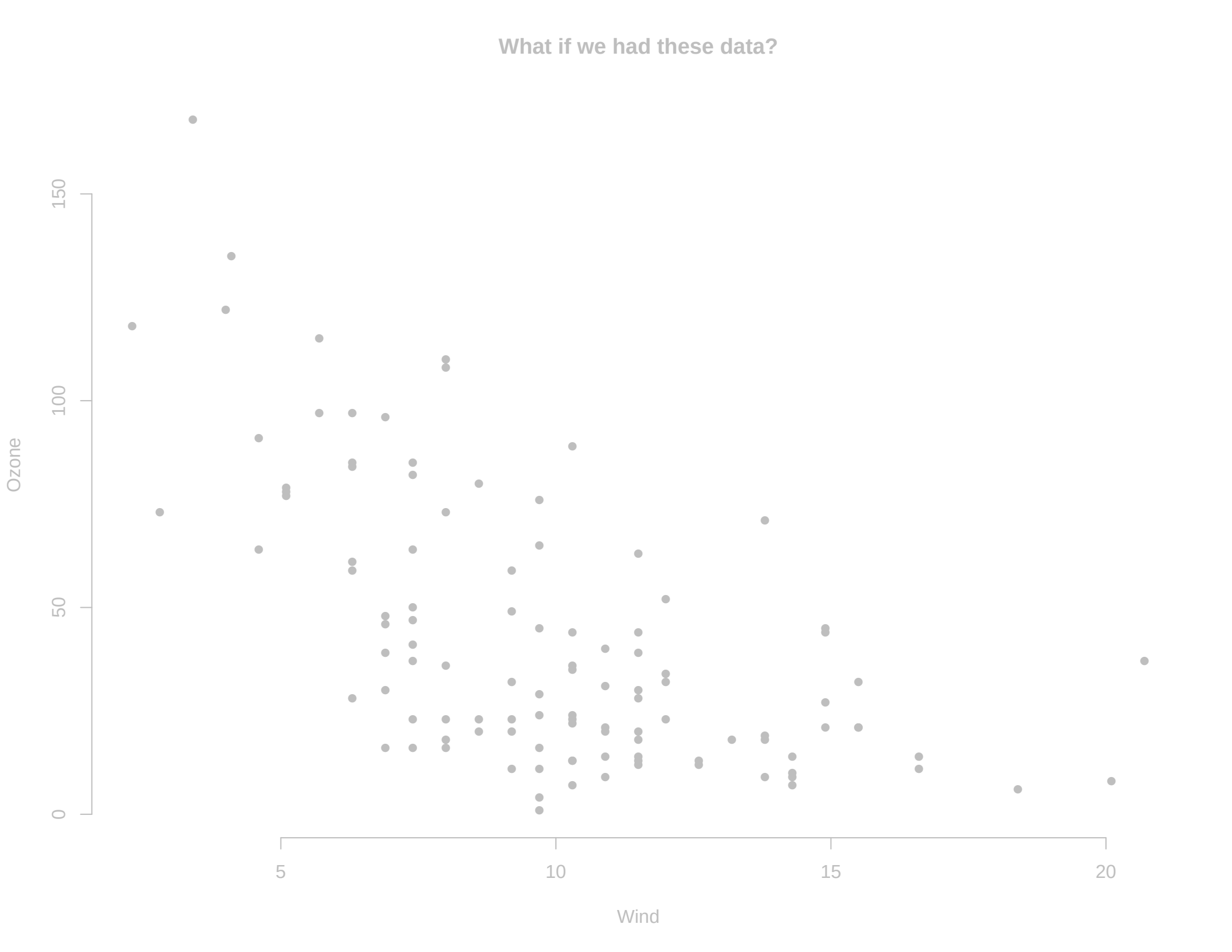


That was for very positive (blue) covariances.

What if covariance is negative (red)?

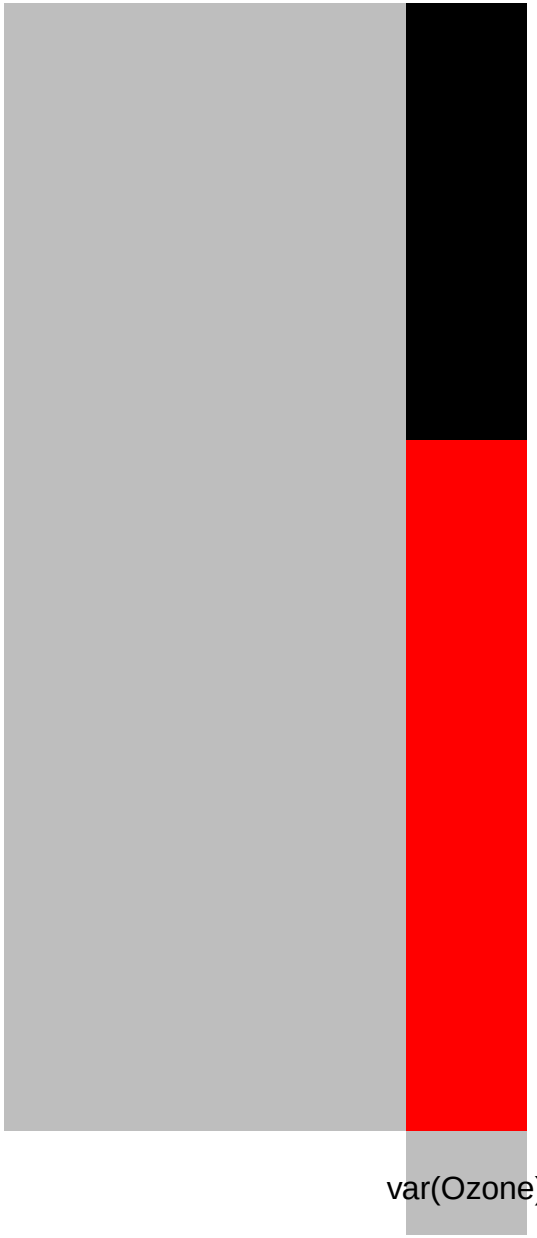
We were just using these data.



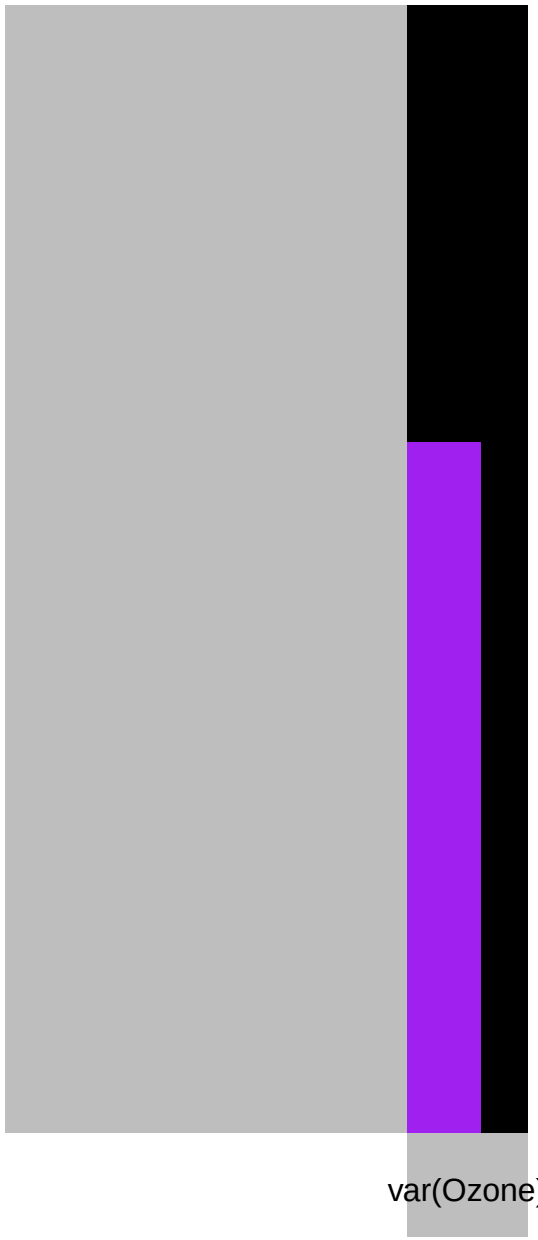

What if we had these data?
Ozone
0
50
100
150
5
10
15
20
Wind


What if we had these data?
Ozone
0
50
100
150
5
10
15
20
Wind

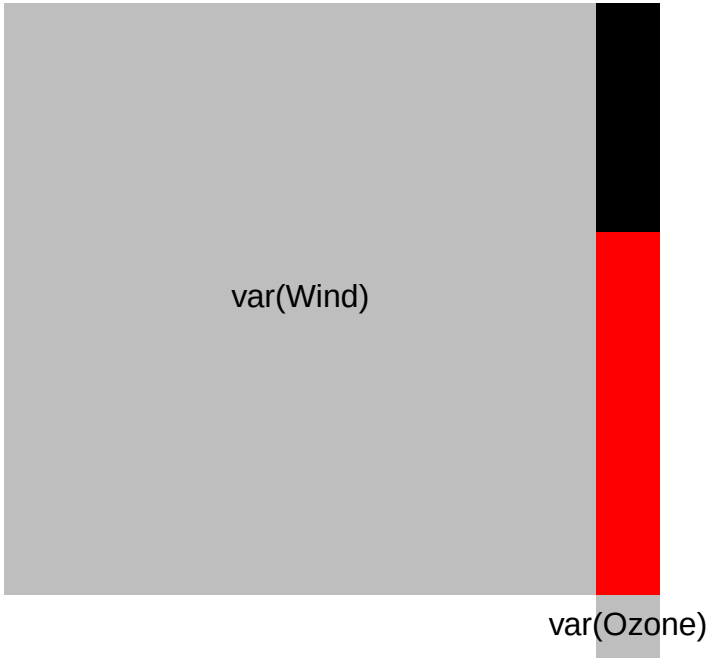
R is the same, just negative.



R−squared is the same, and it is always positive.

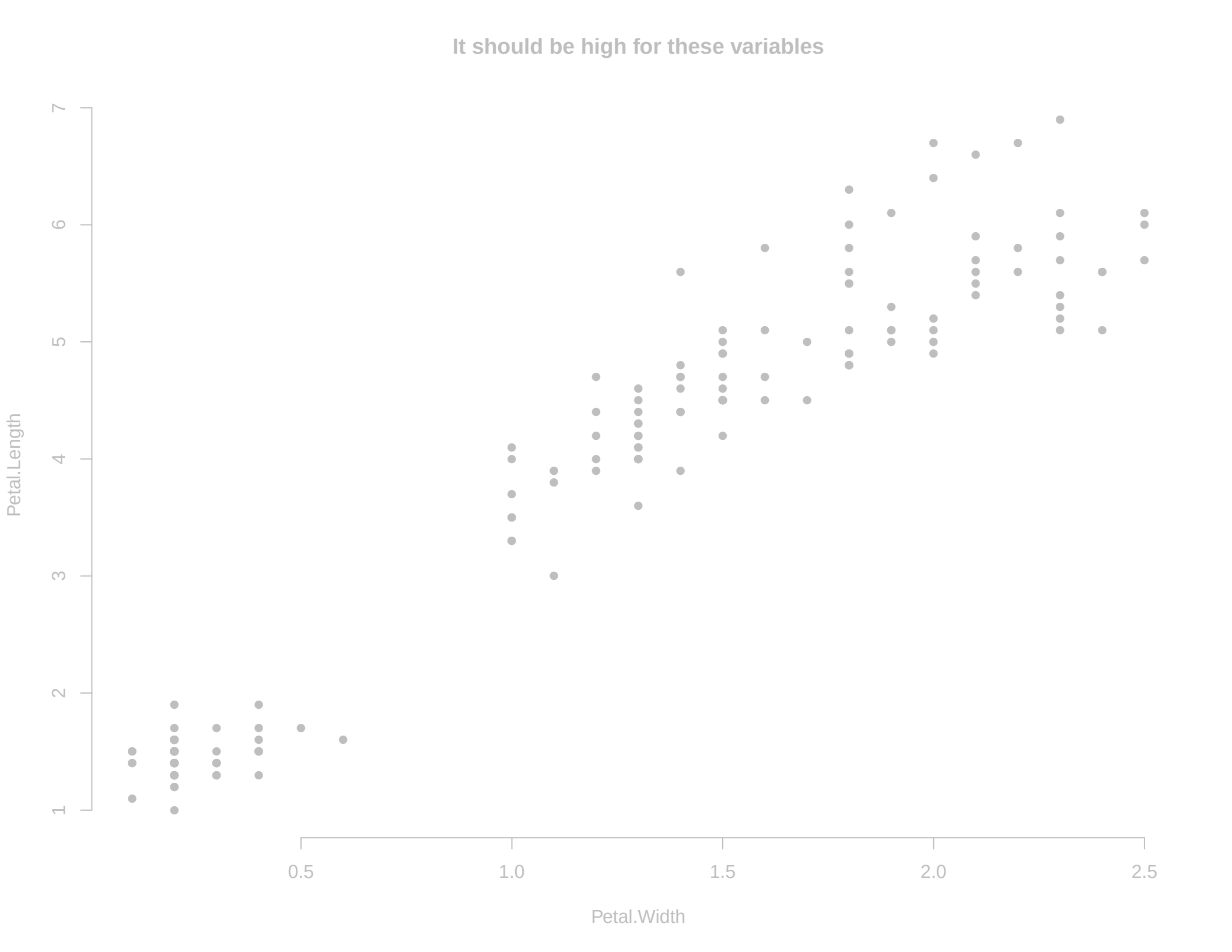


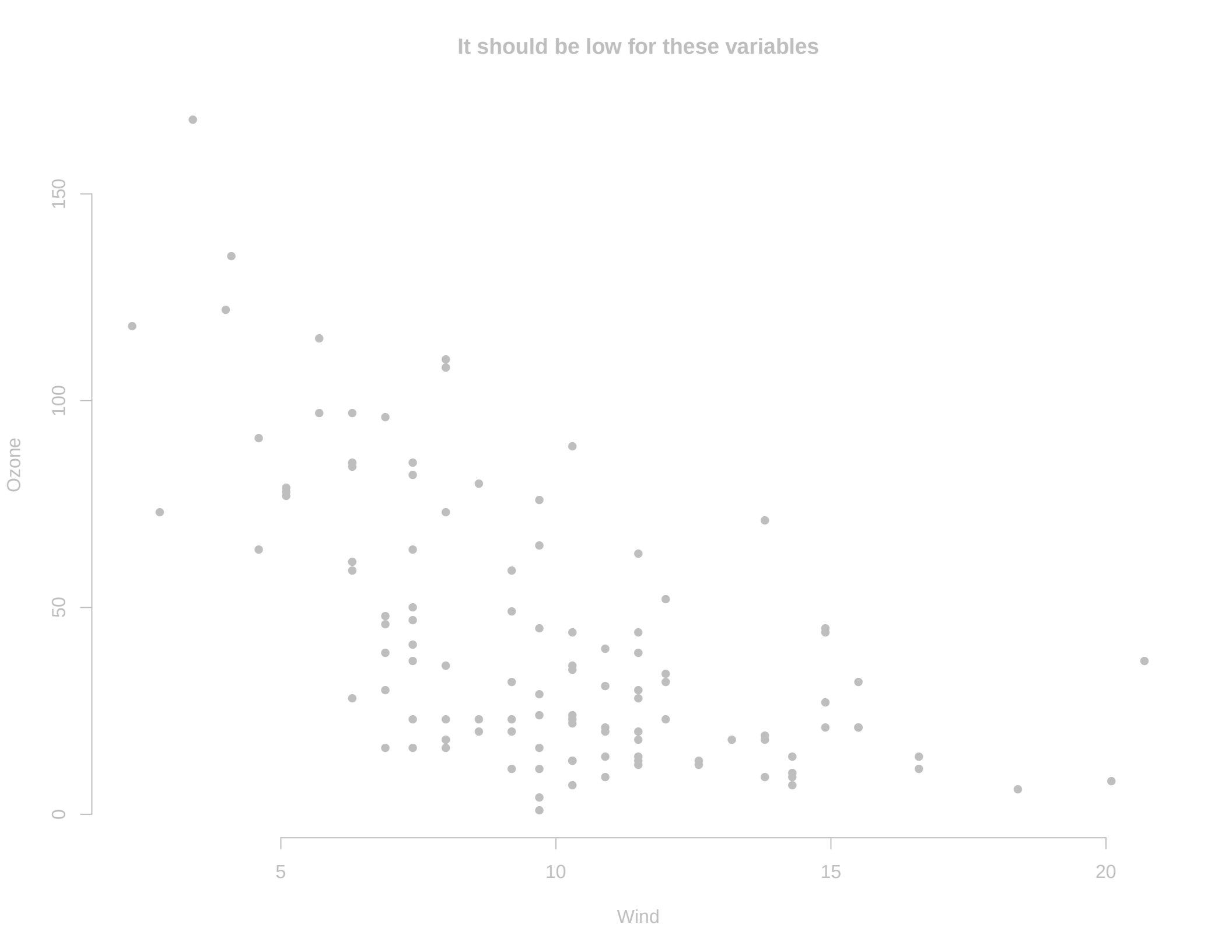
Zoom back out.

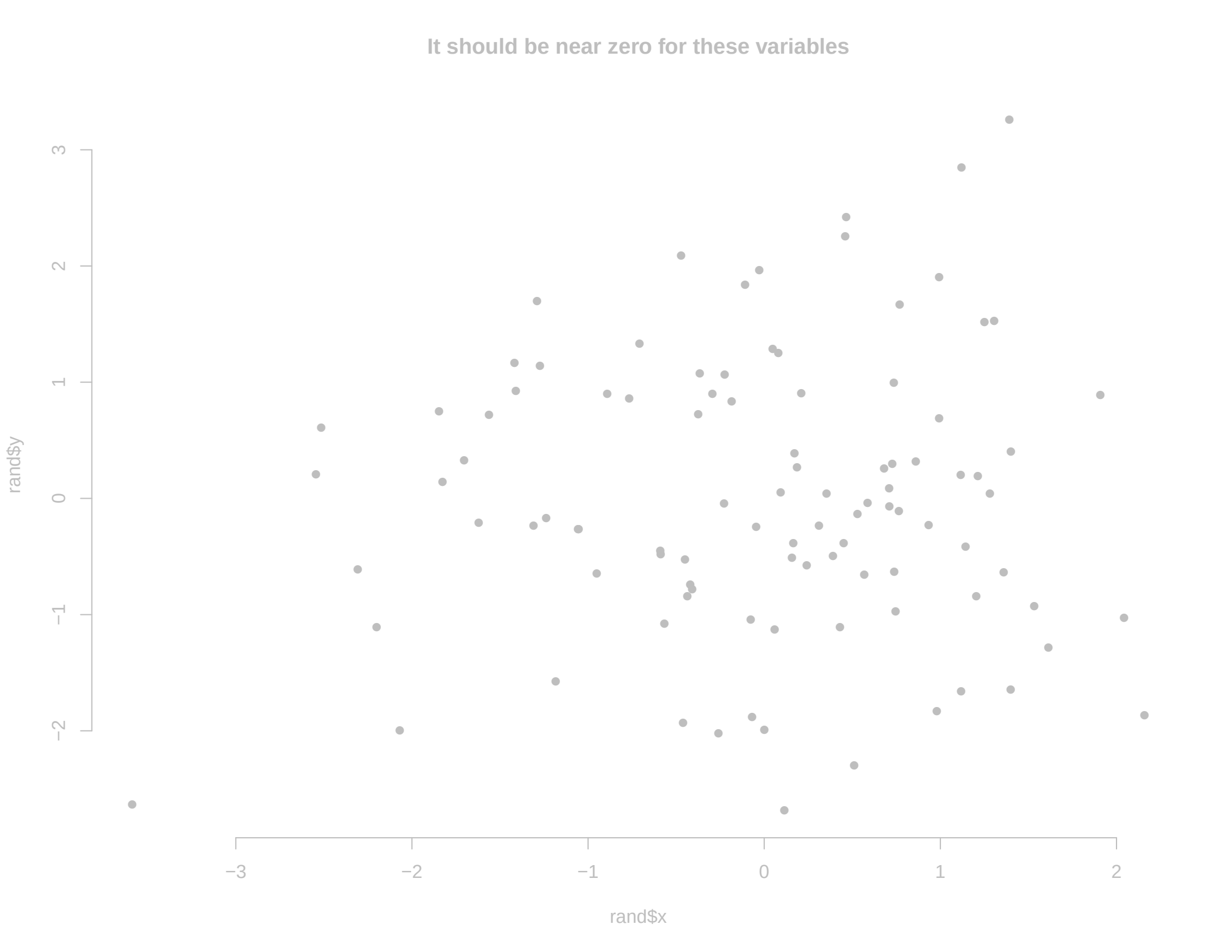


Remember how this fits in.

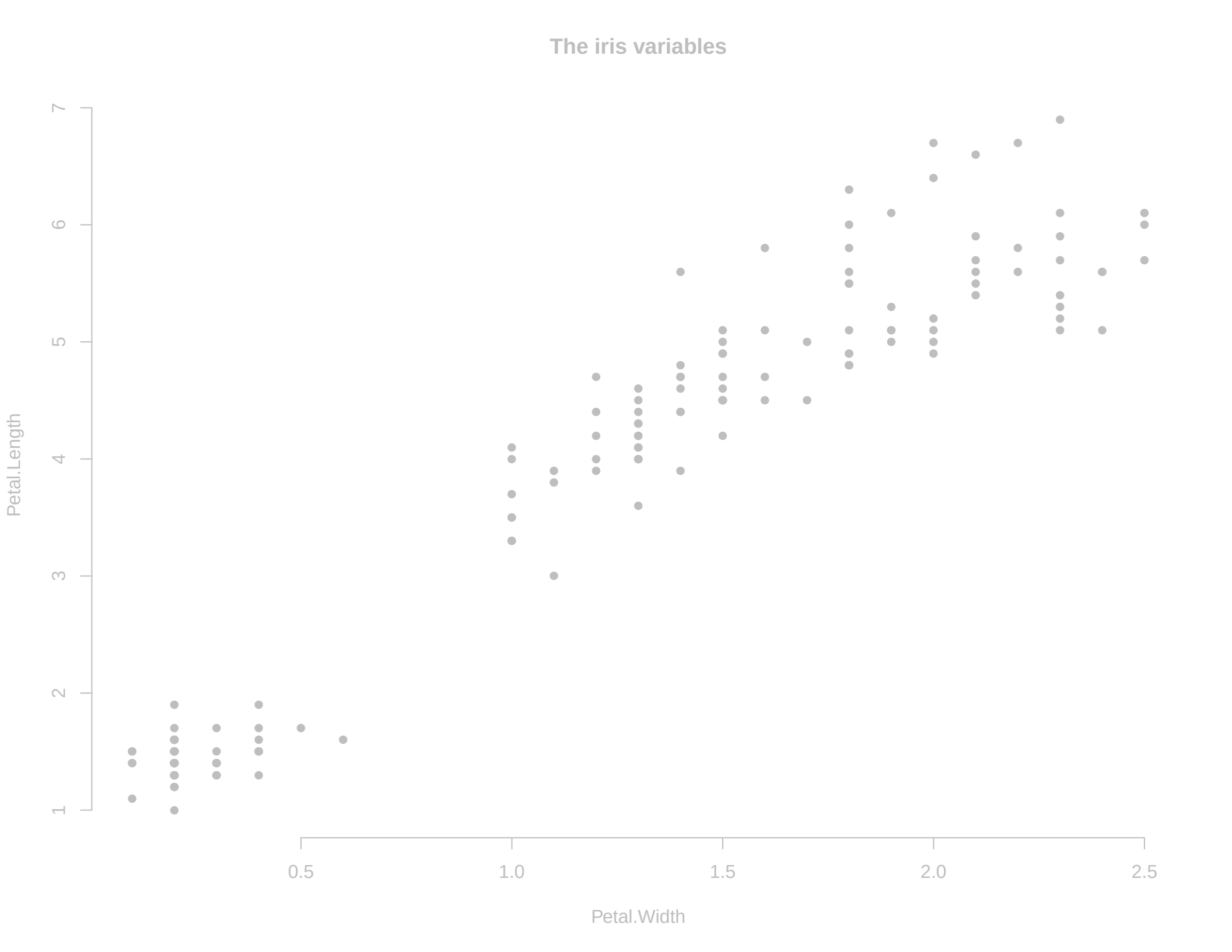
We want a number that describes whether two variables move together.

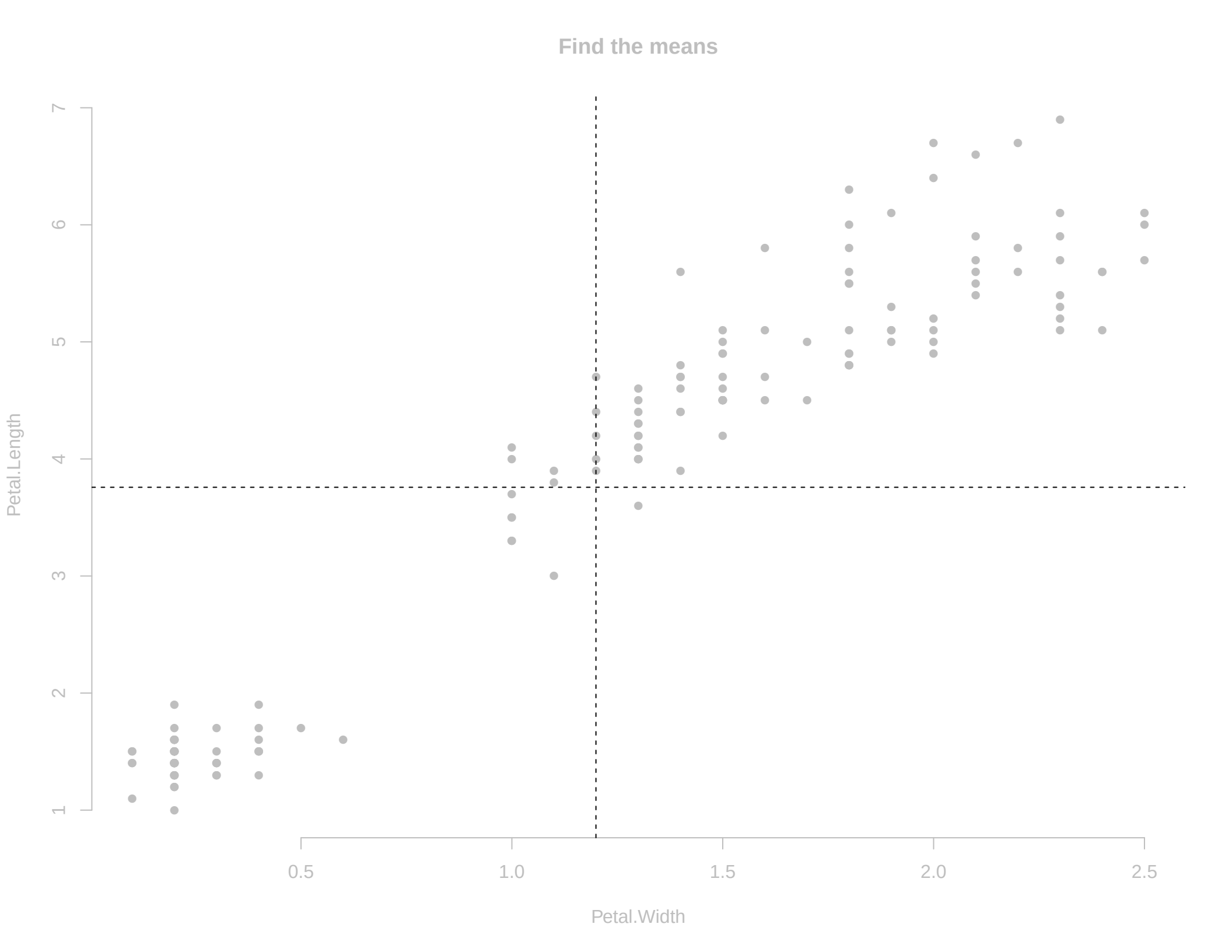

It should be high for these variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

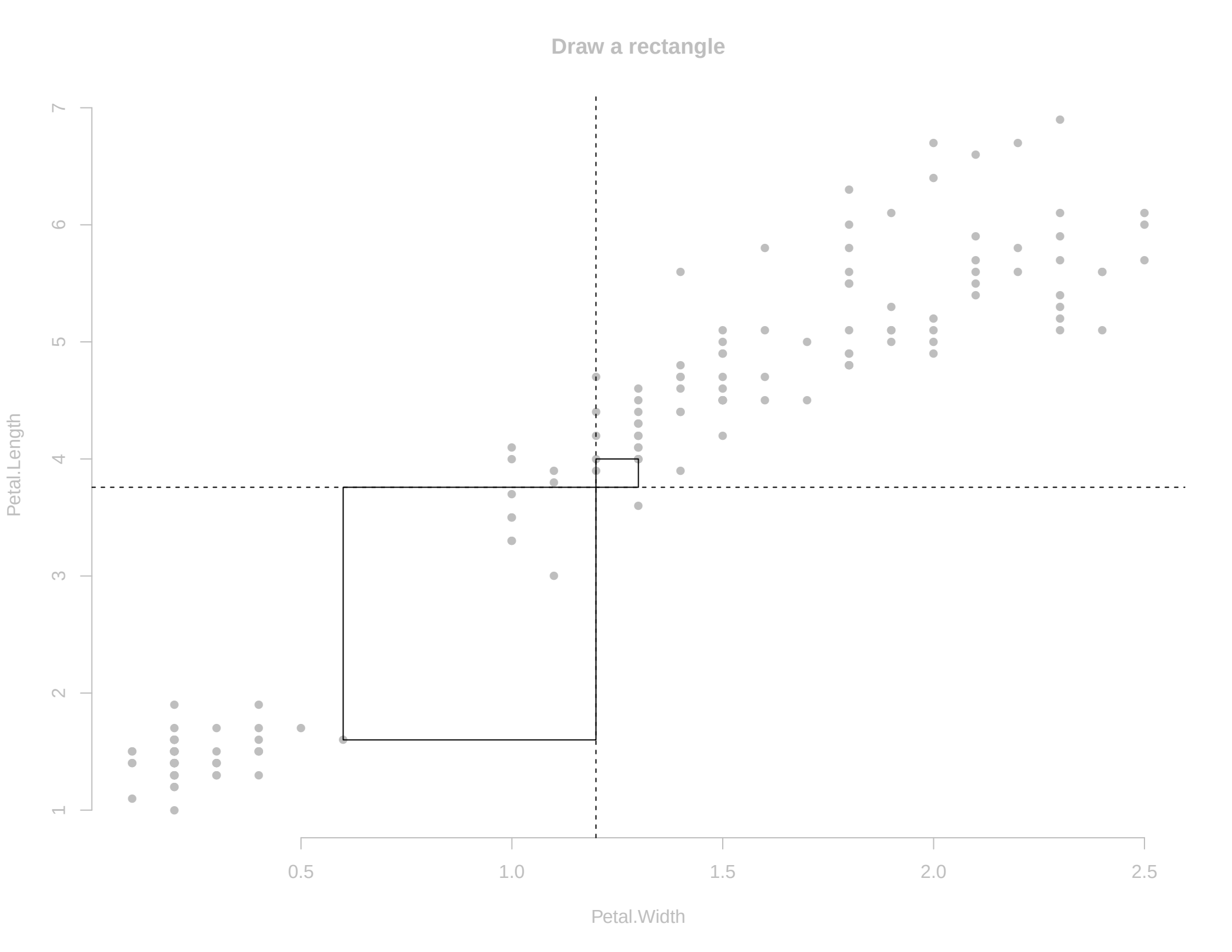

It should be low for these variables
Ozone
0
50
100
150
5
10
15
20
Wind

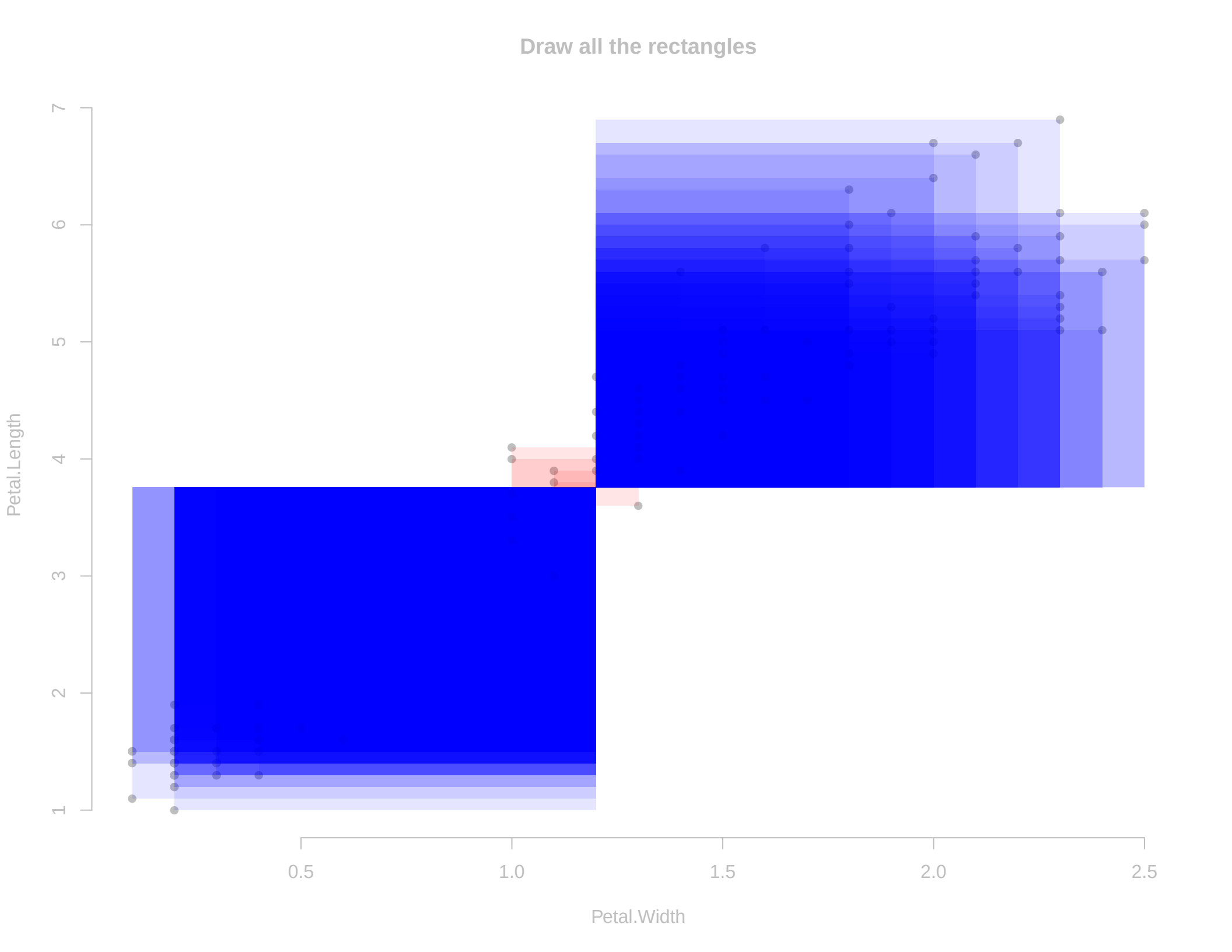

It should be near zero for these variables
rand$y
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
rand$x

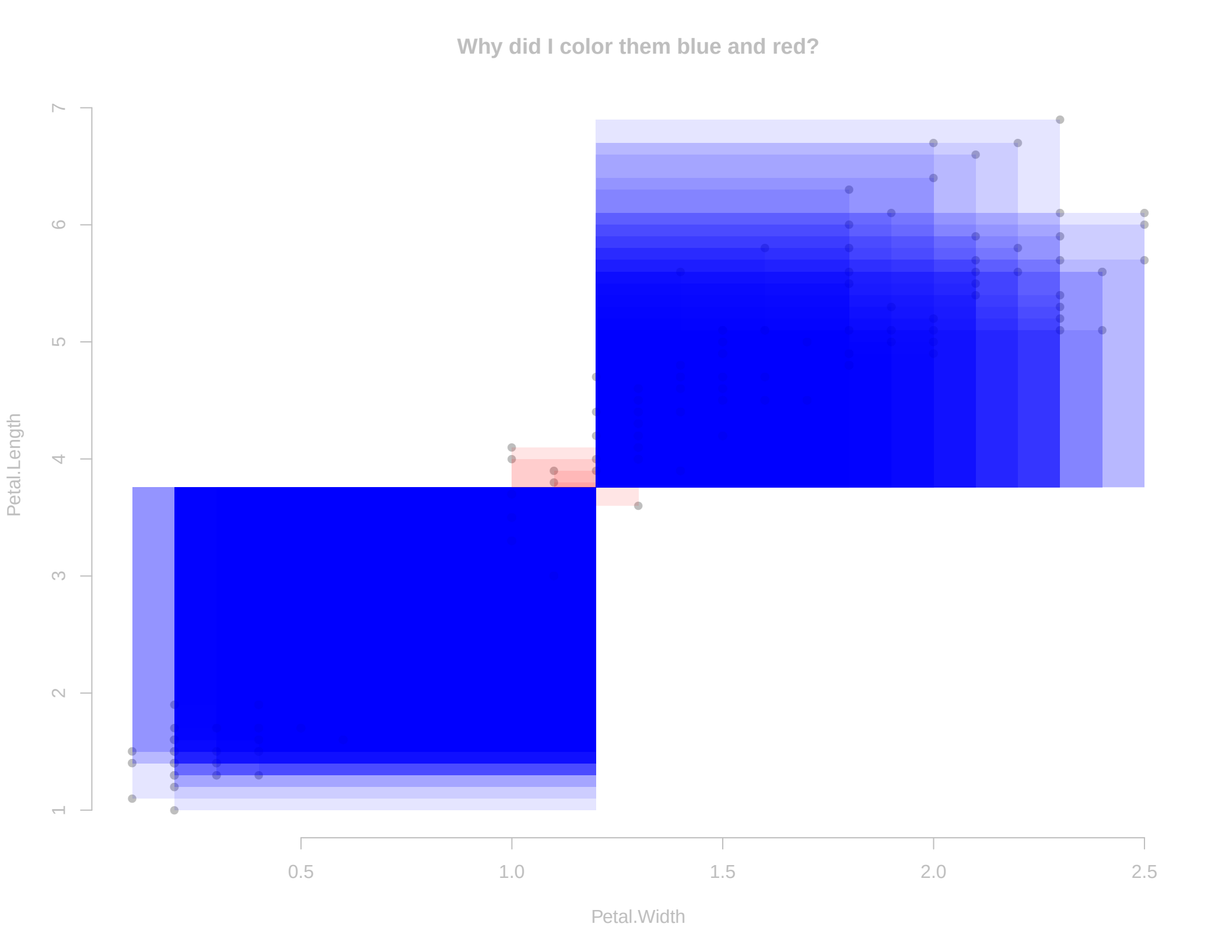
Covariance

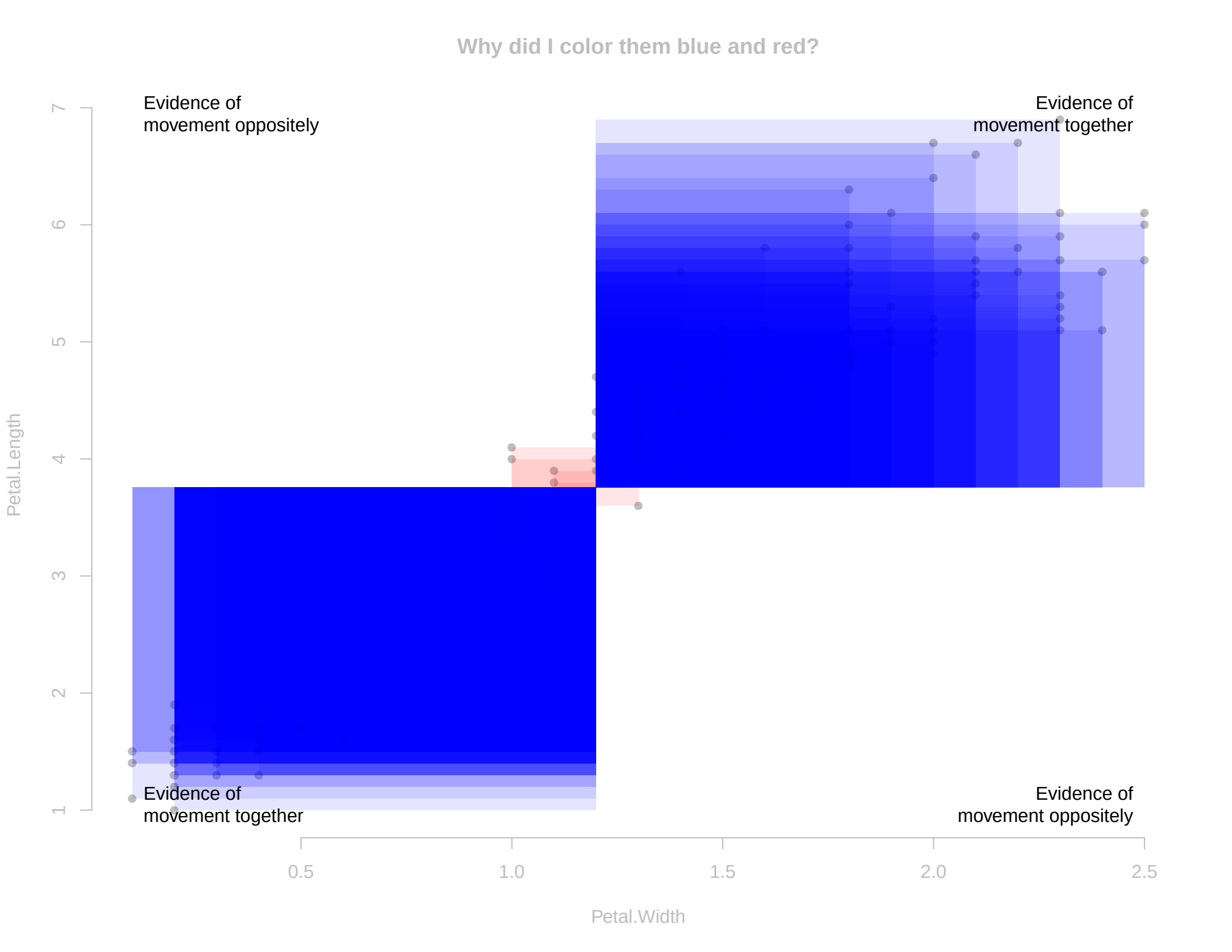

The iris variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Find the means
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

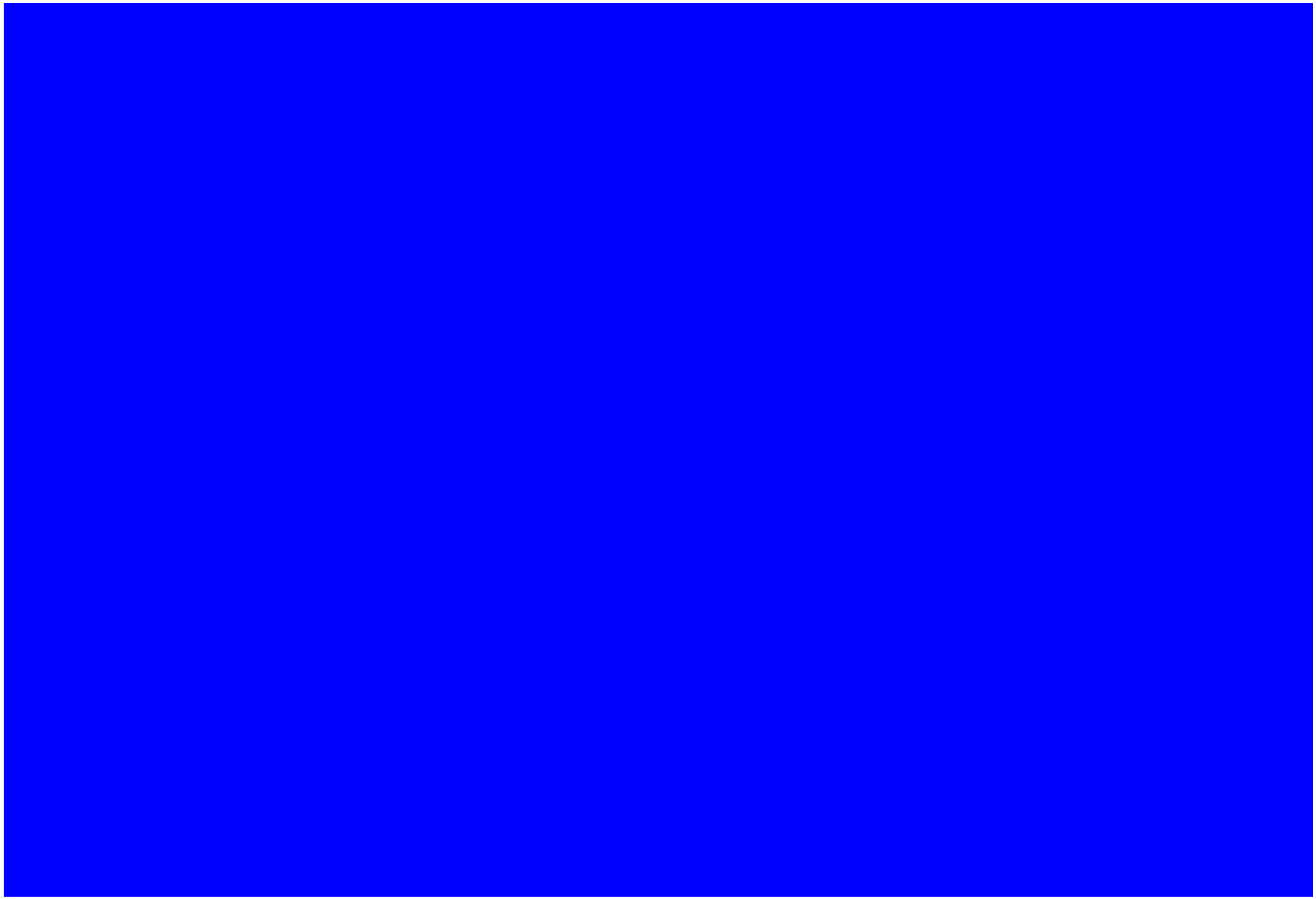

Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

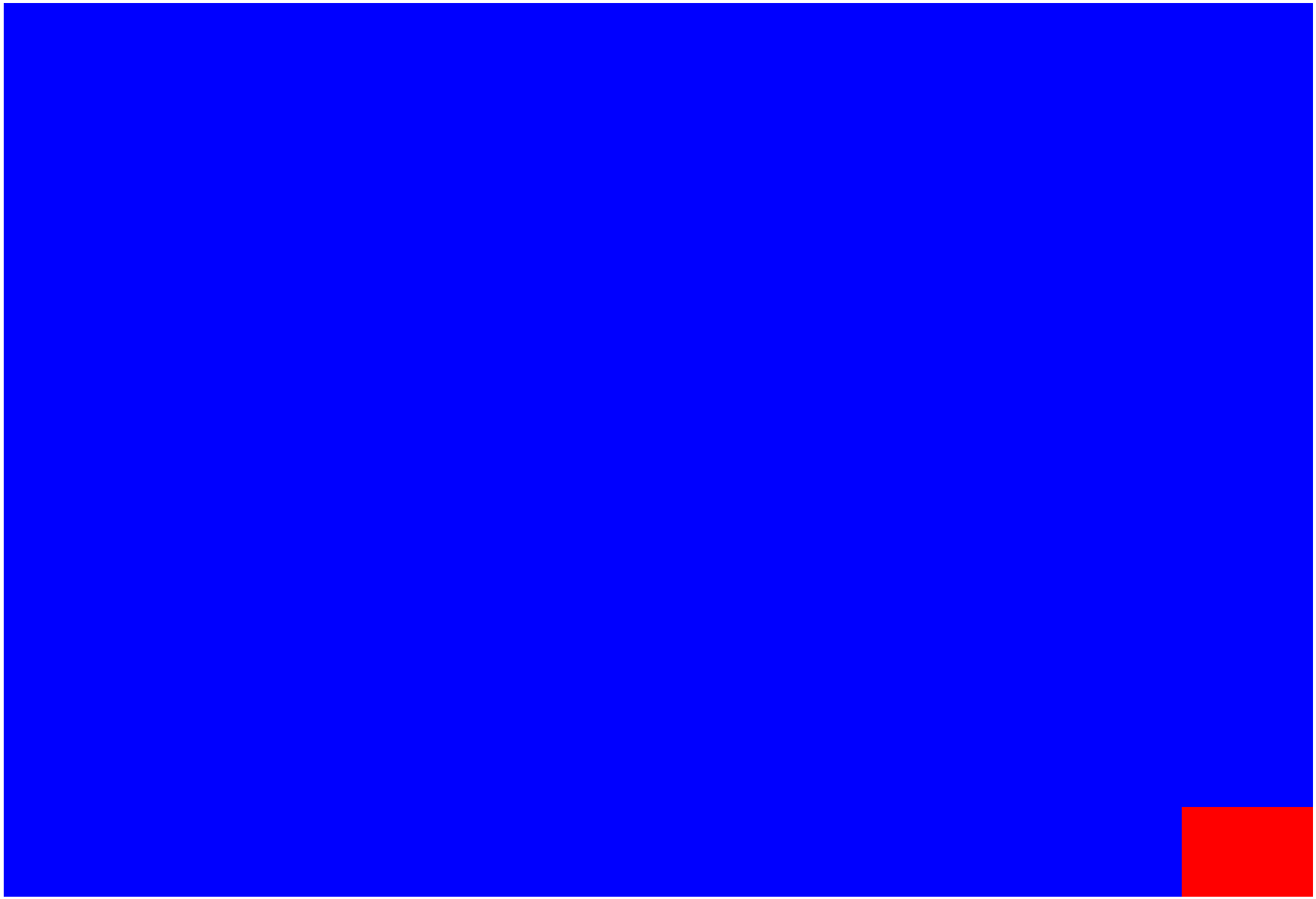

Why did I color them blue and red?
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


Why did I color them blue and red?
Evidence of
movement oppositely
Evidence of
movement together
Evidence of
movement together
Evidence of
movement oppositely
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

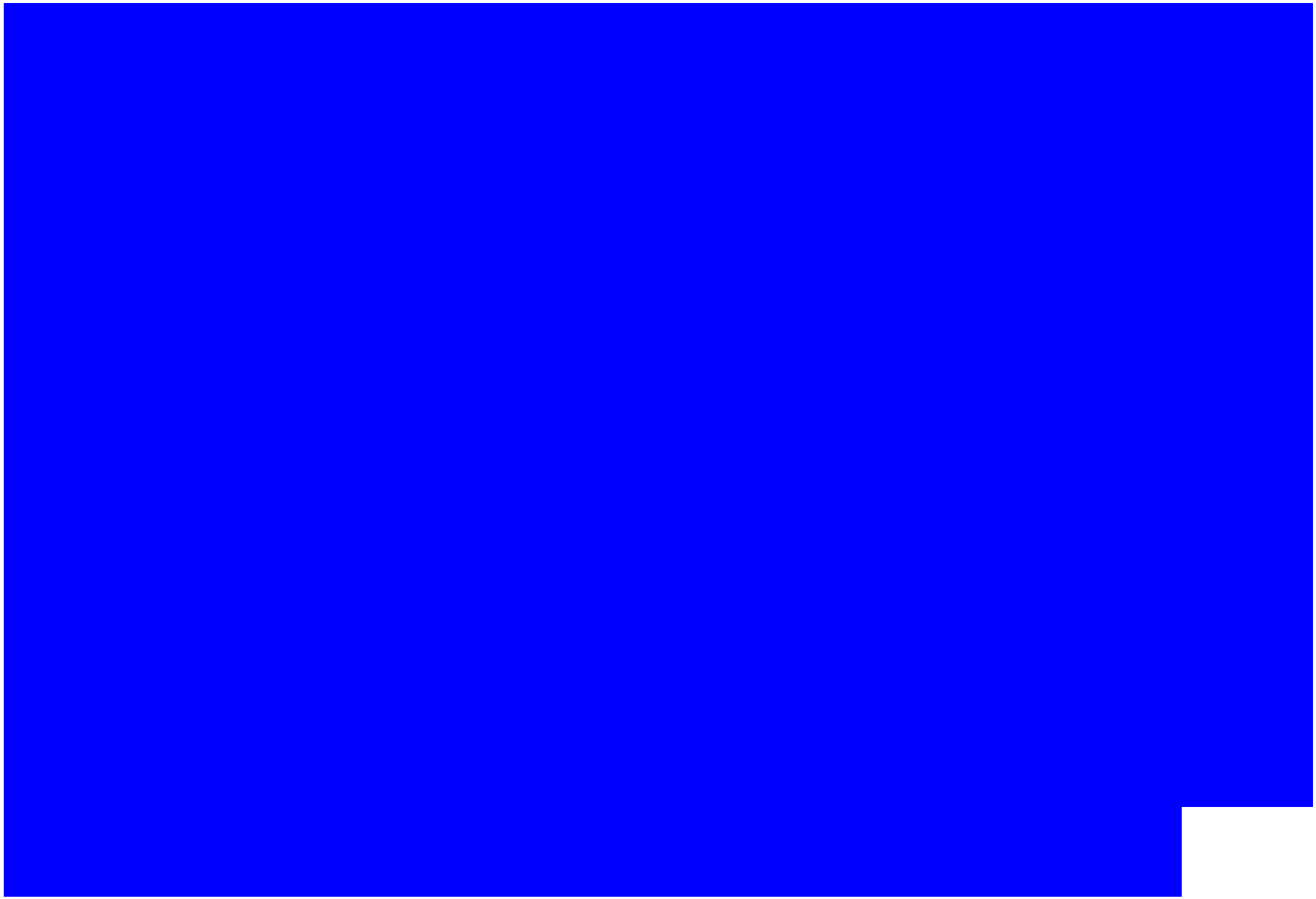
Add the blues together. (This is at a different scale.)



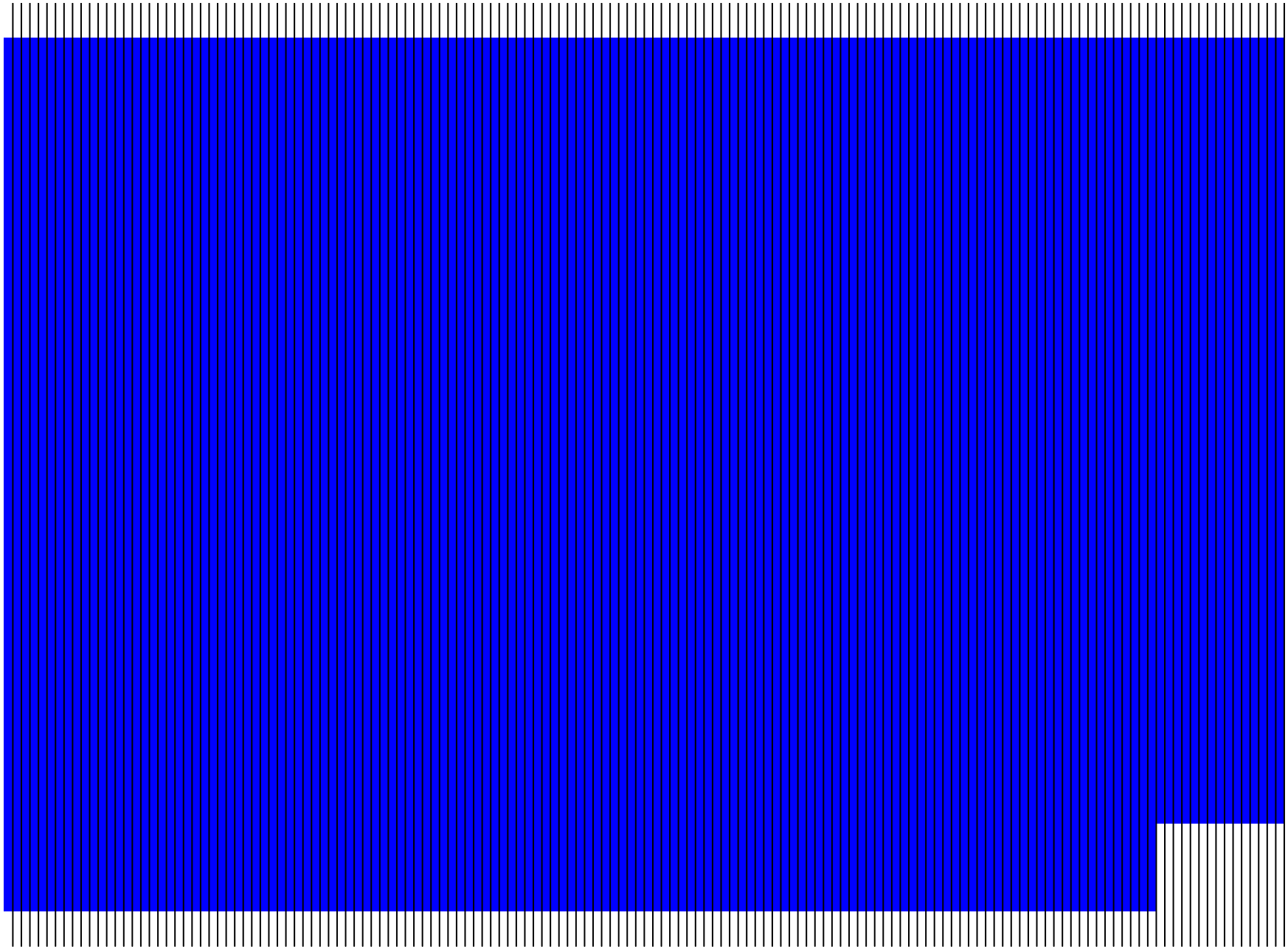
Add the reds together.



Subtract the reds.



Divide into as many equal pieces as we have irises (n).



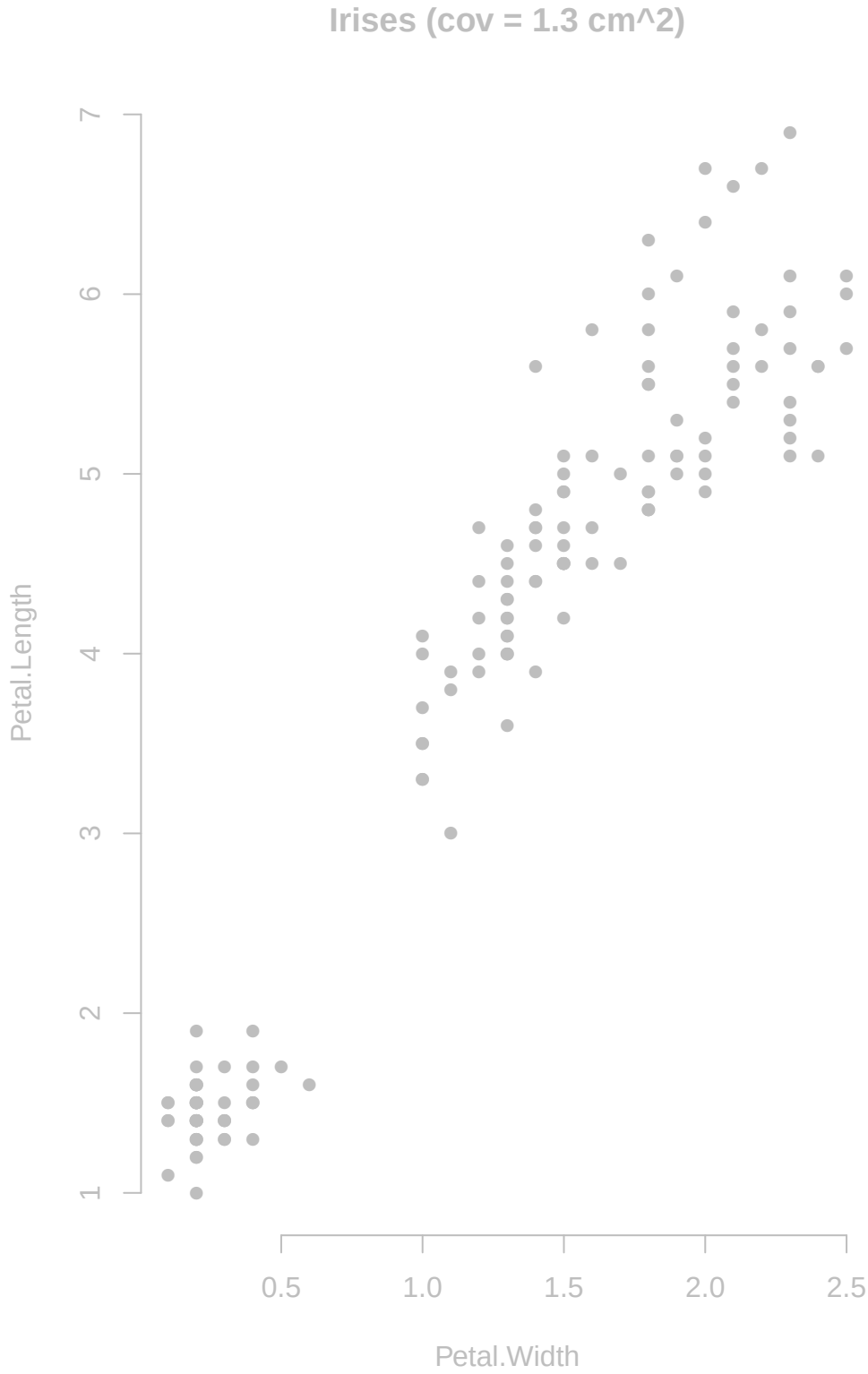
This blue sliver is the covariance.

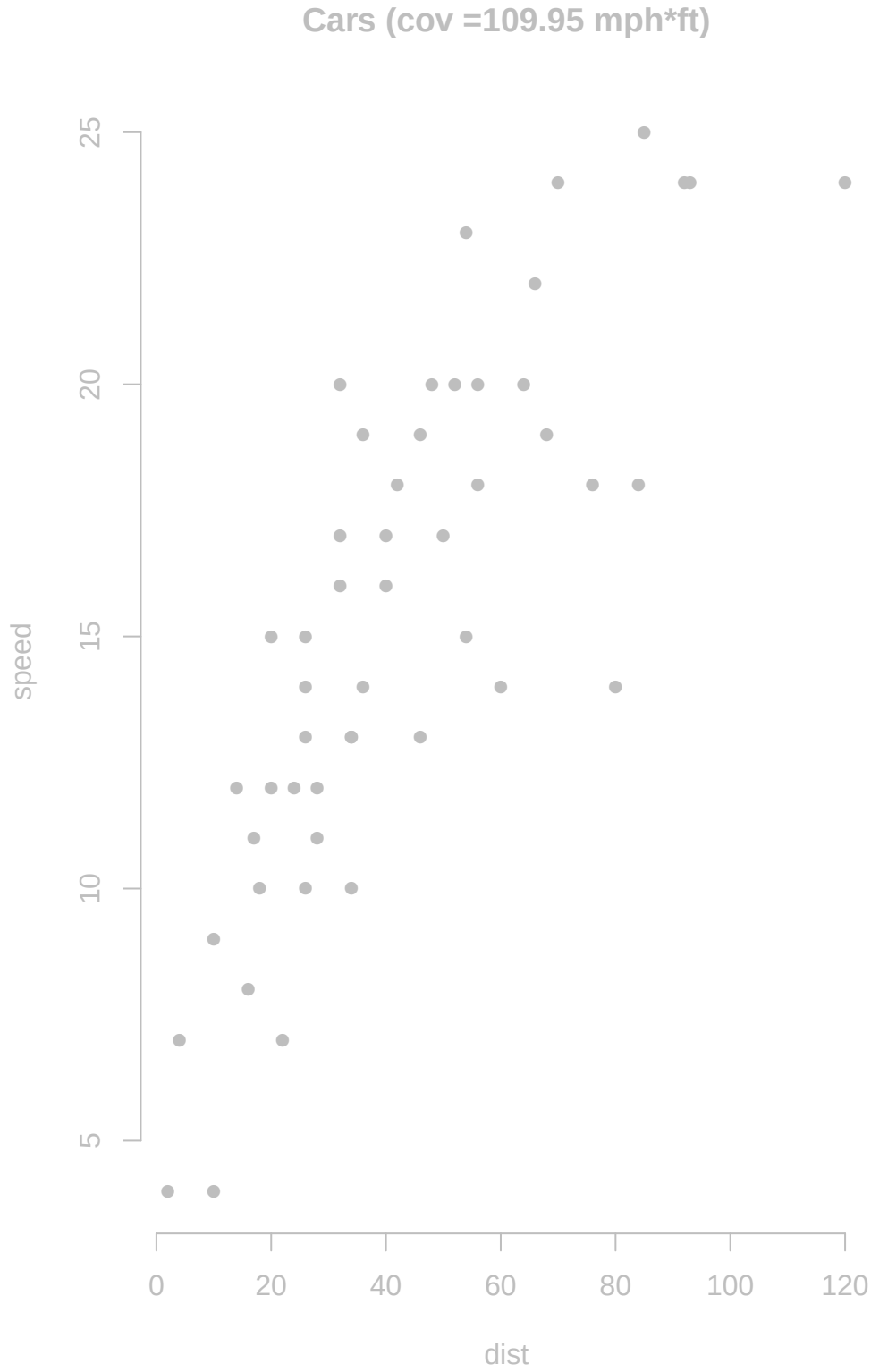
A problem with covariance

Covariance has units!

(x−unit times y−unit)

Which relationship is stronger (more linear)?

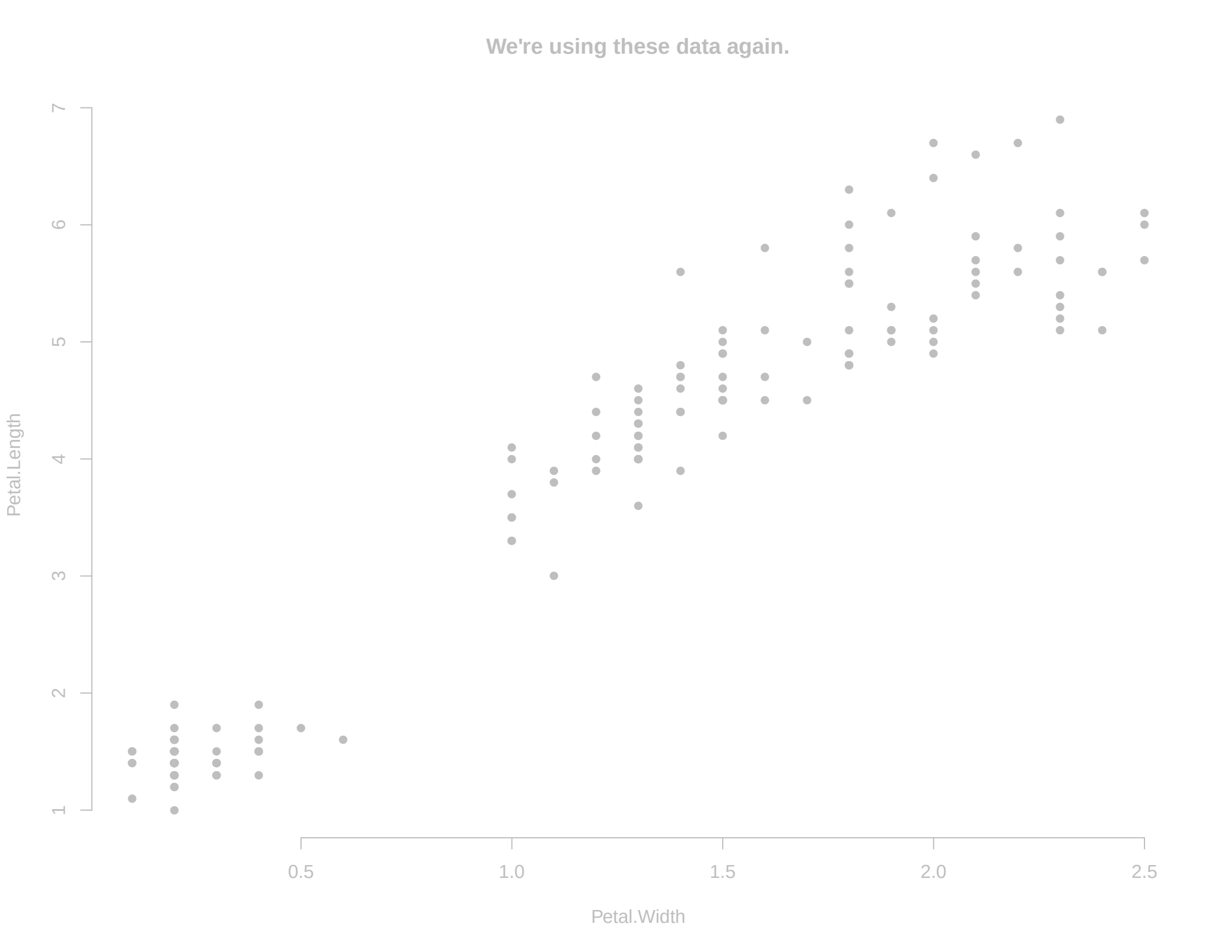

Irises (cov = 1.3 cm^2)
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

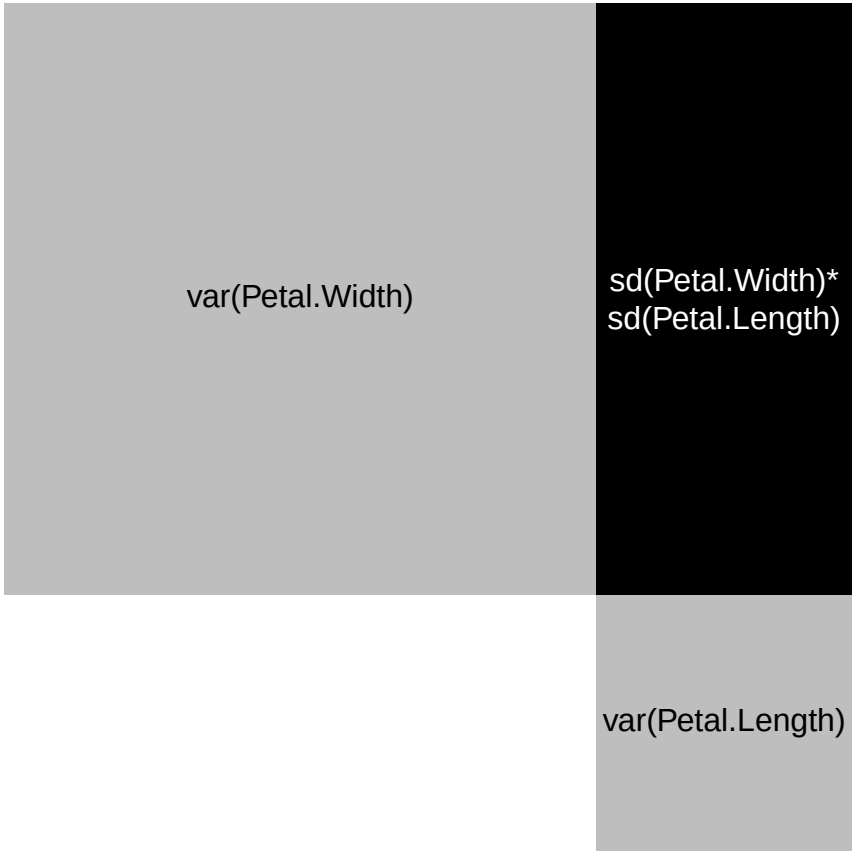

Cars (cov =109.95 mph*ft)
speed
dist
5
10
15
20
25
0
20
40
60
80
100
120

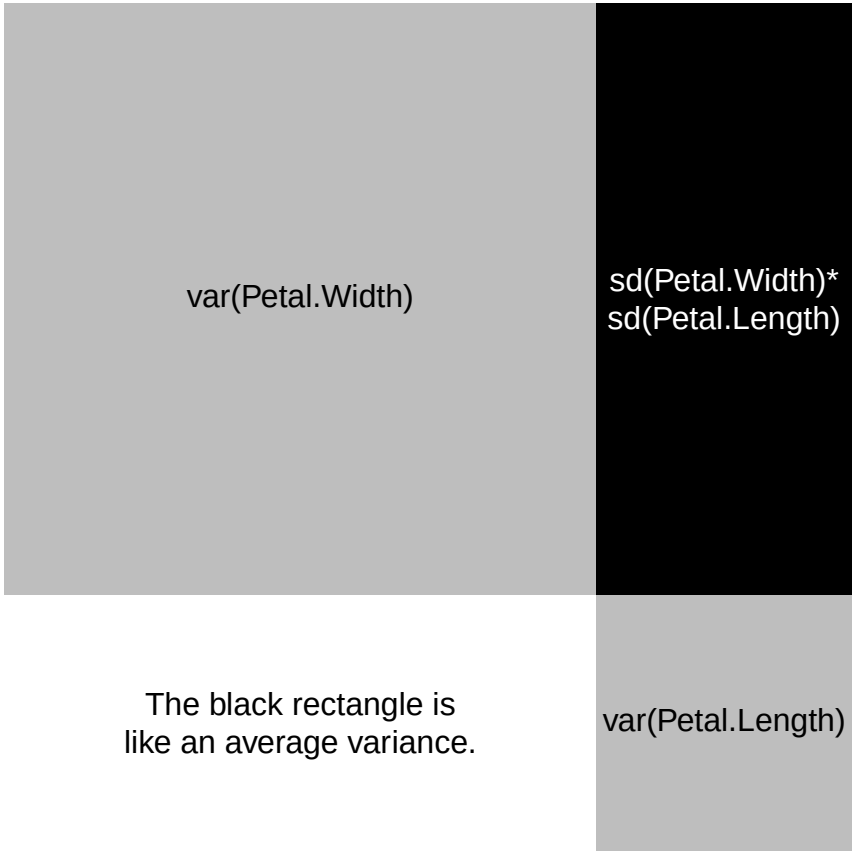
Oh noes!

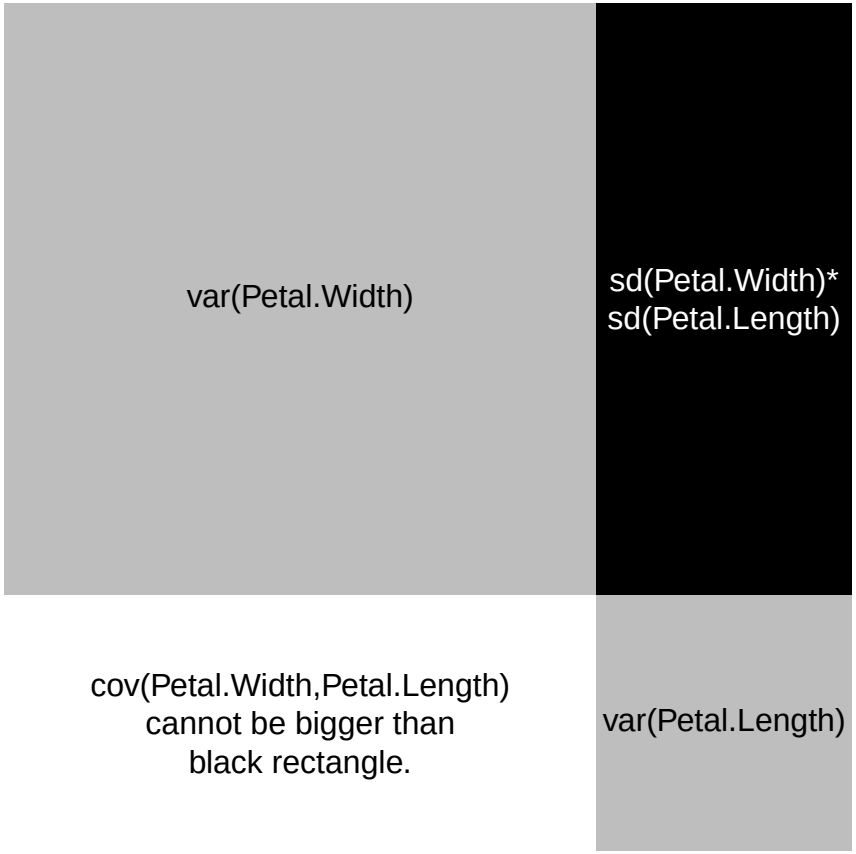
We can divide the covariance by the variances to standardize it.

We're using these data again.

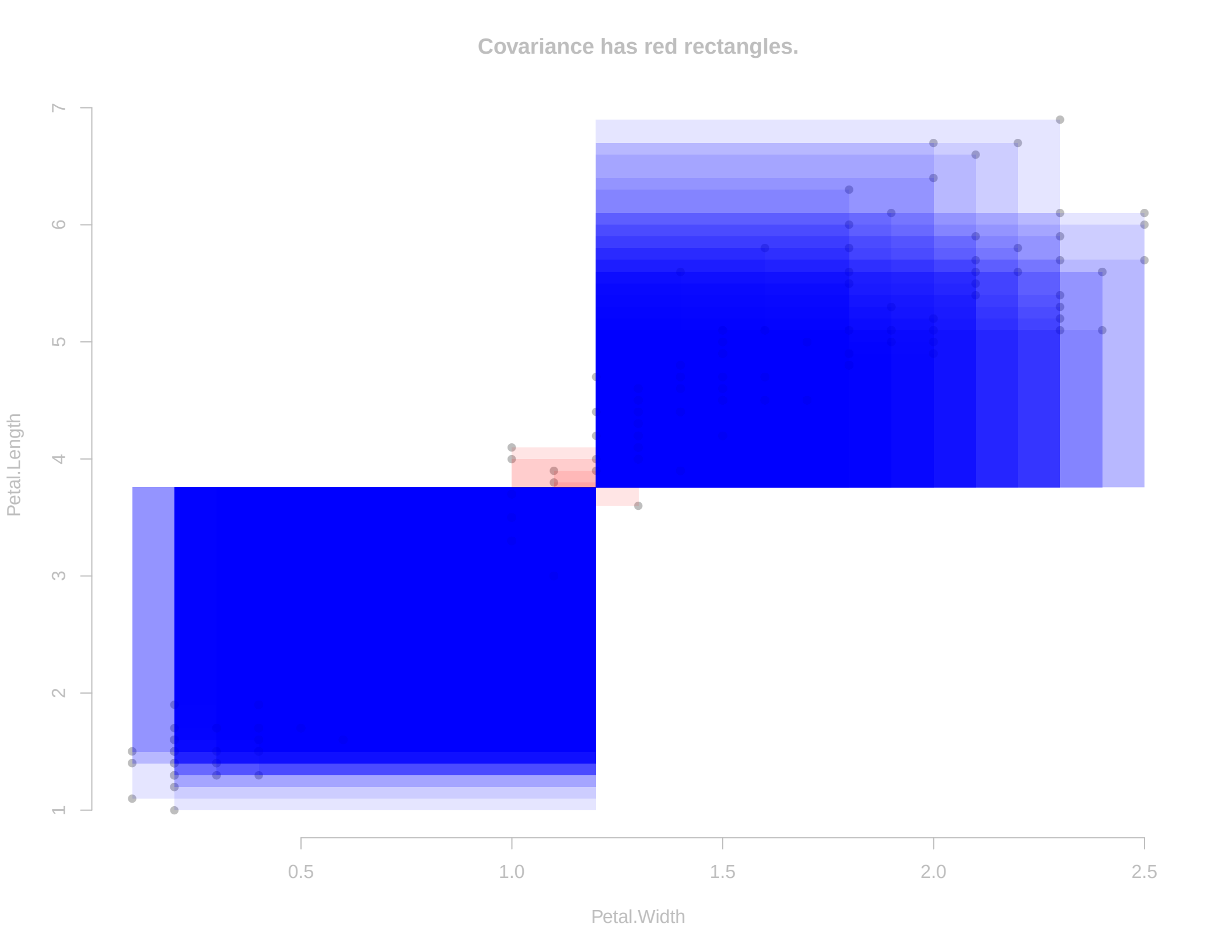


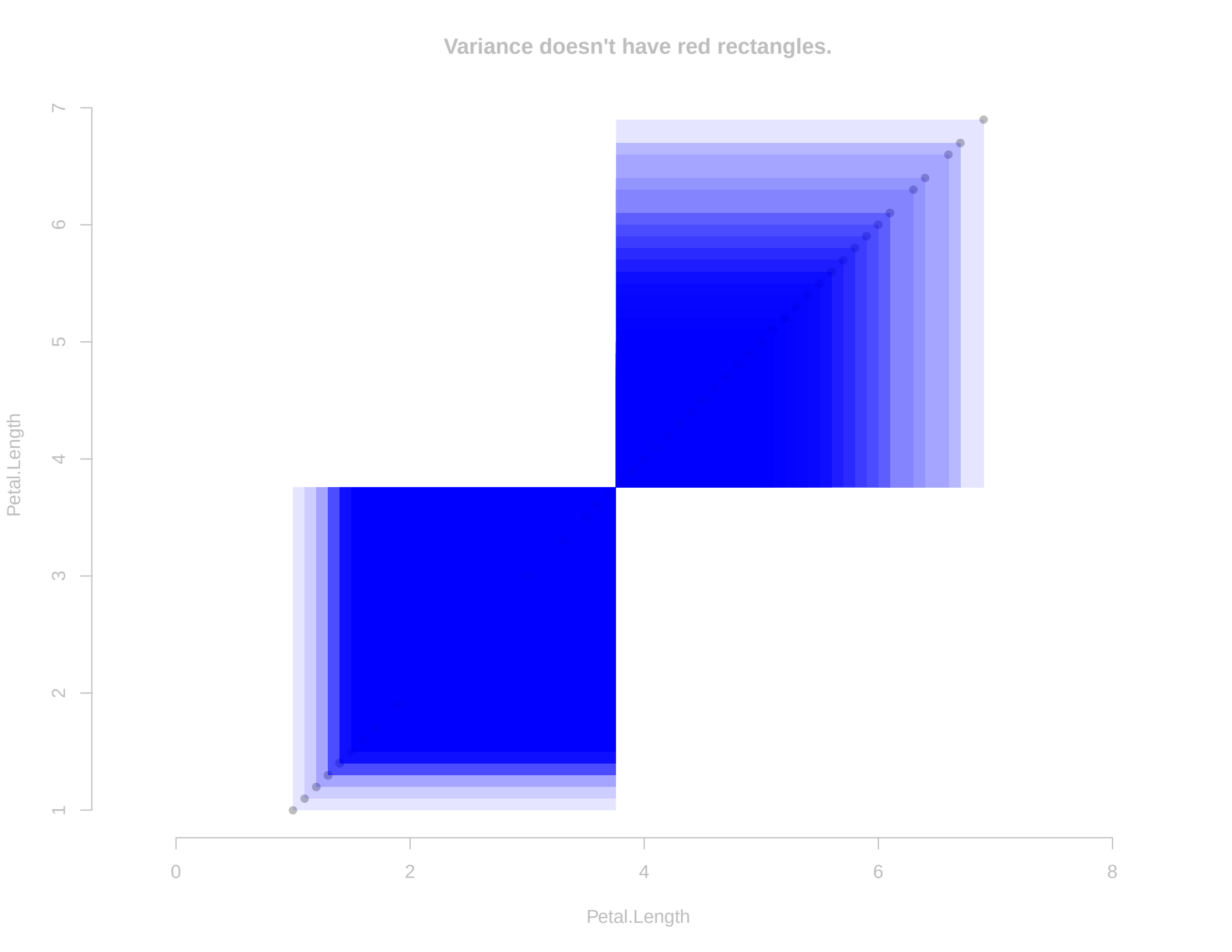

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
var(Petal.Length)

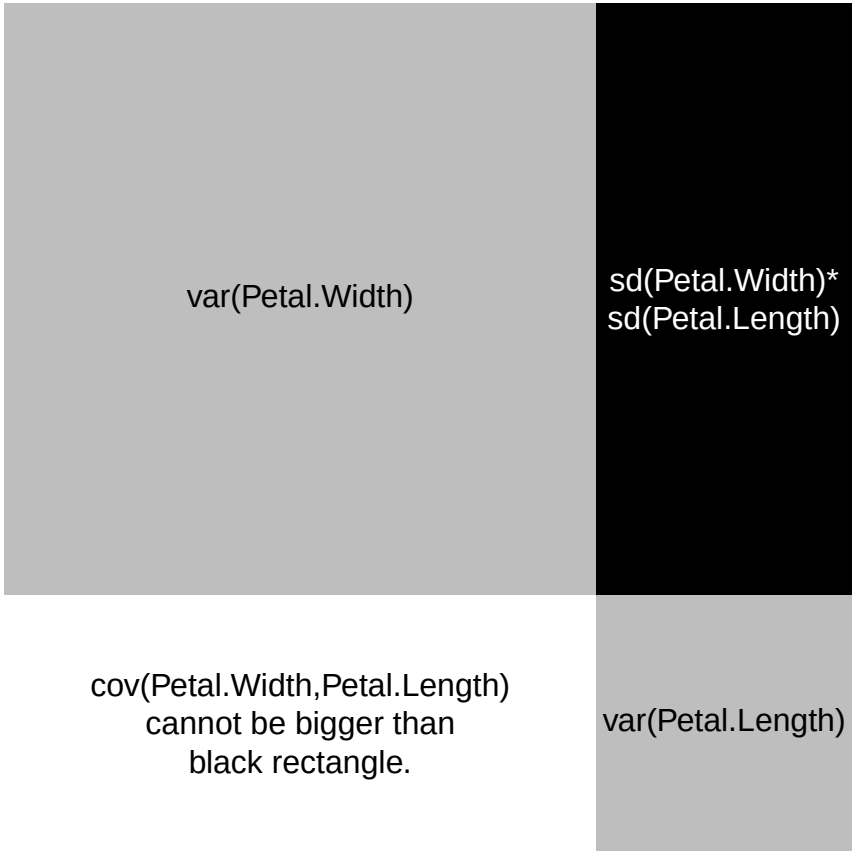

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
The black rectangle is
like an average variance.
var(Petal.Length)


var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
cov(Petal.Width,Petal.Length)
cannot be bigger than
black rectangle.
var(Petal.Length)

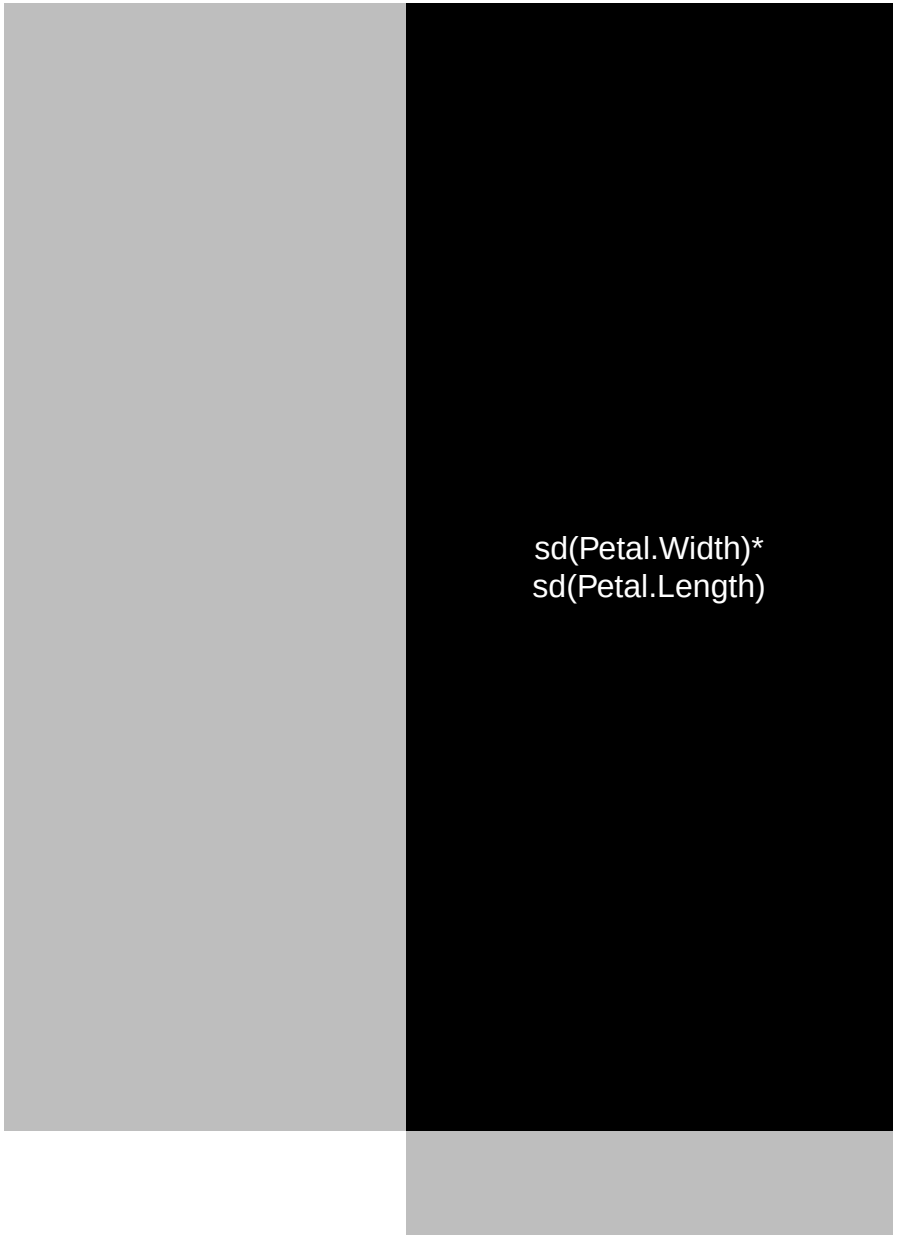
Why?


Covariance has red rectangles.
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

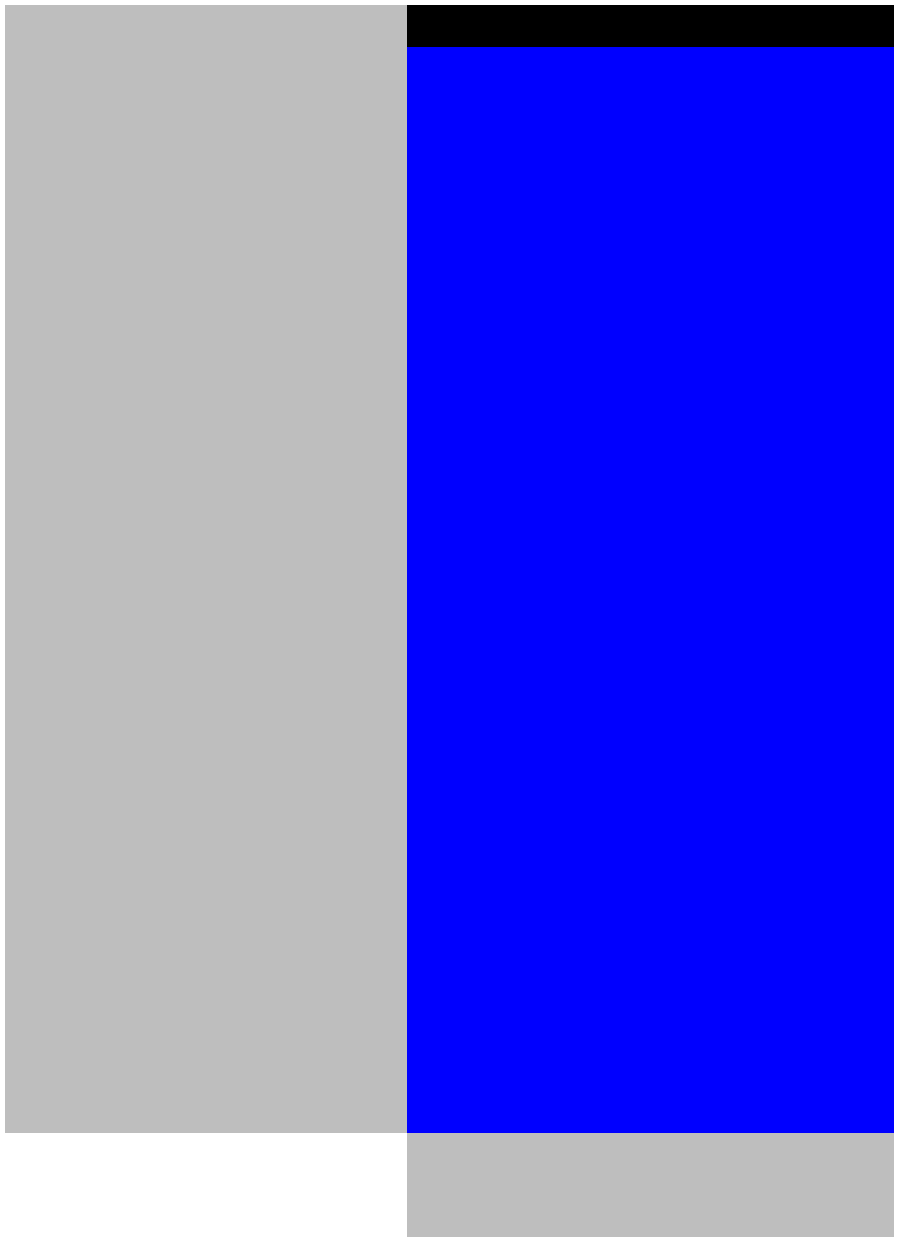

Variance doesn't have red rectangles.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length


var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
cov(Petal.Width,Petal.Length)
cannot be bigger than
black rectangle.
var(Petal.Length)

Let's zoom in.

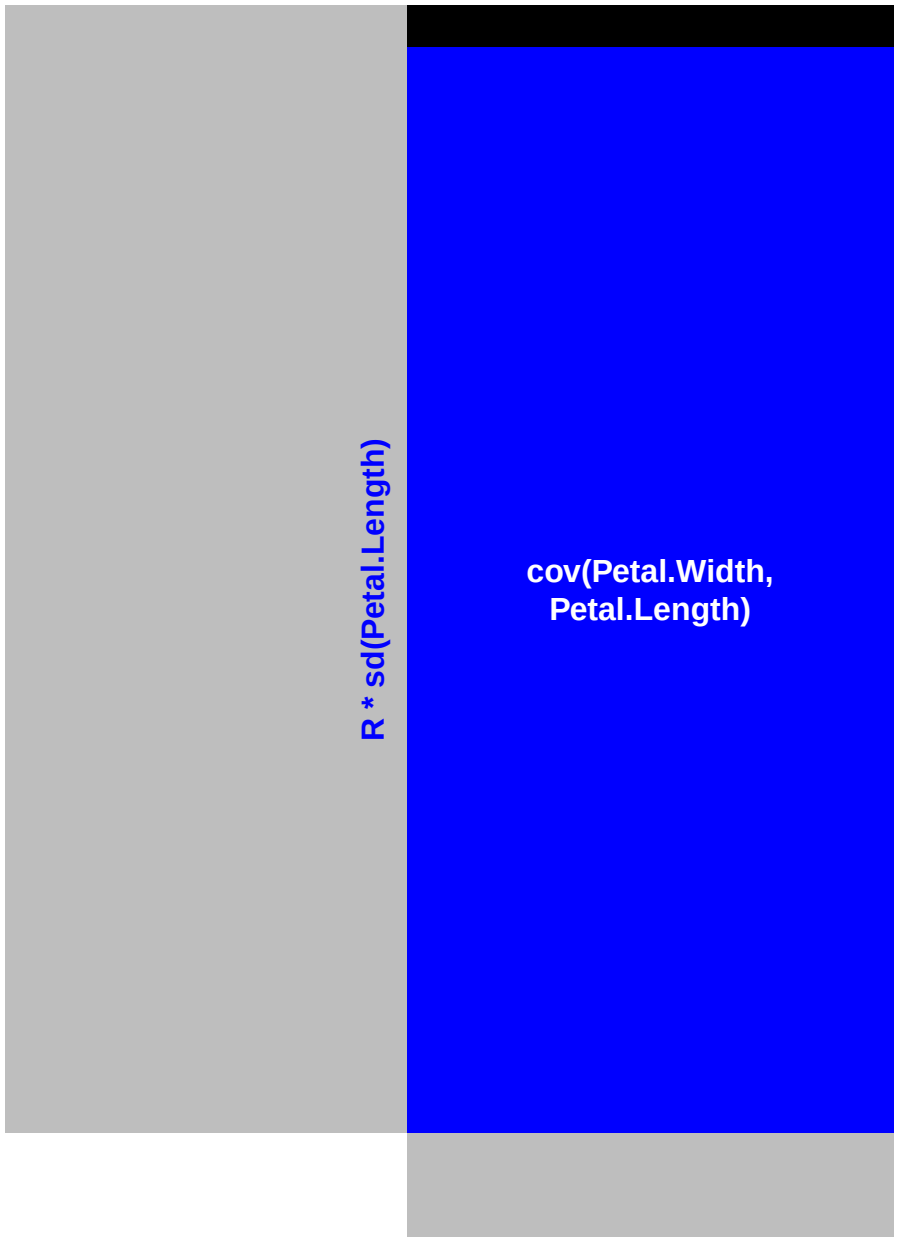


Squish covariance vertically into the rectangle.

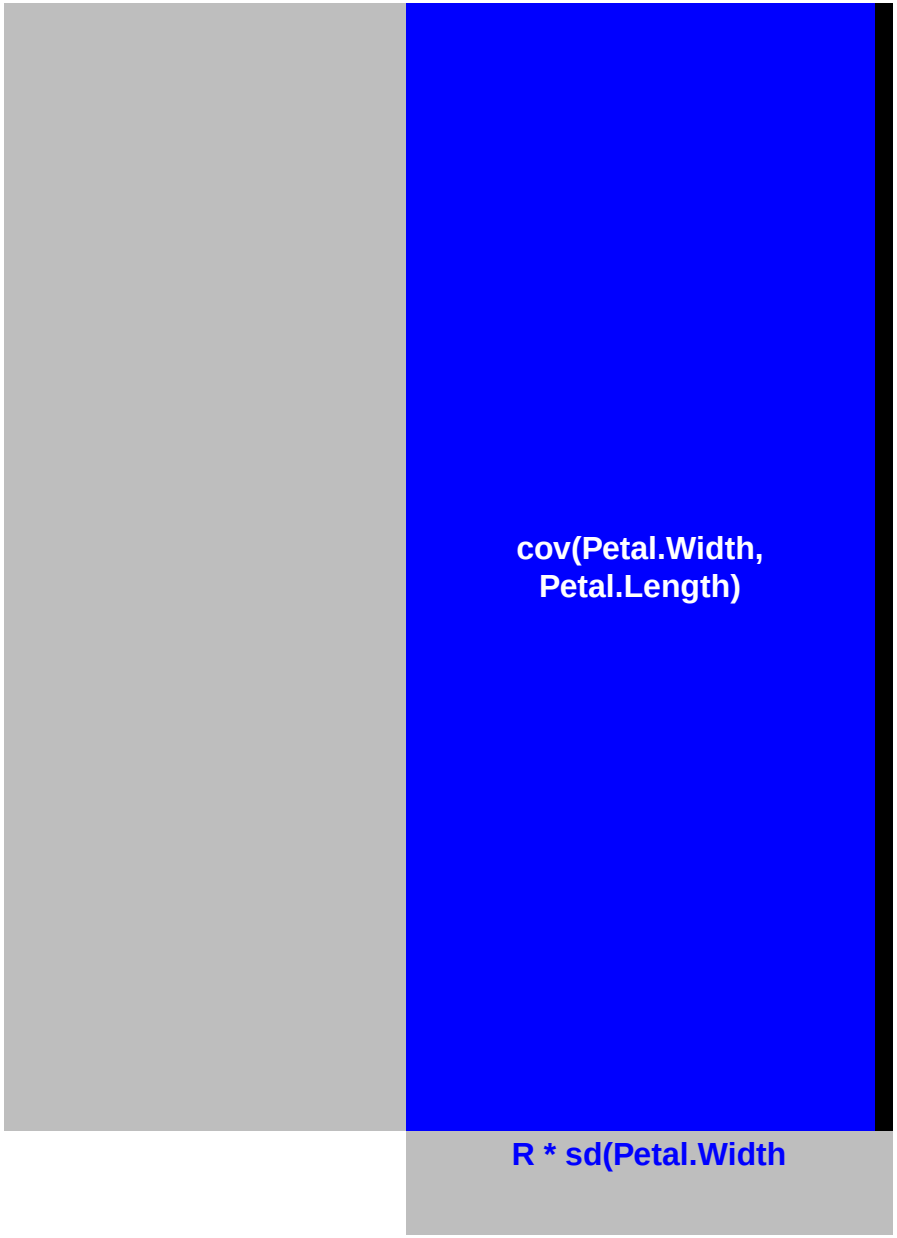


Correlation (R)
is the ratio of
the small rectangle
to the big rectangle.

Squish covariance vertically into the rectangle.

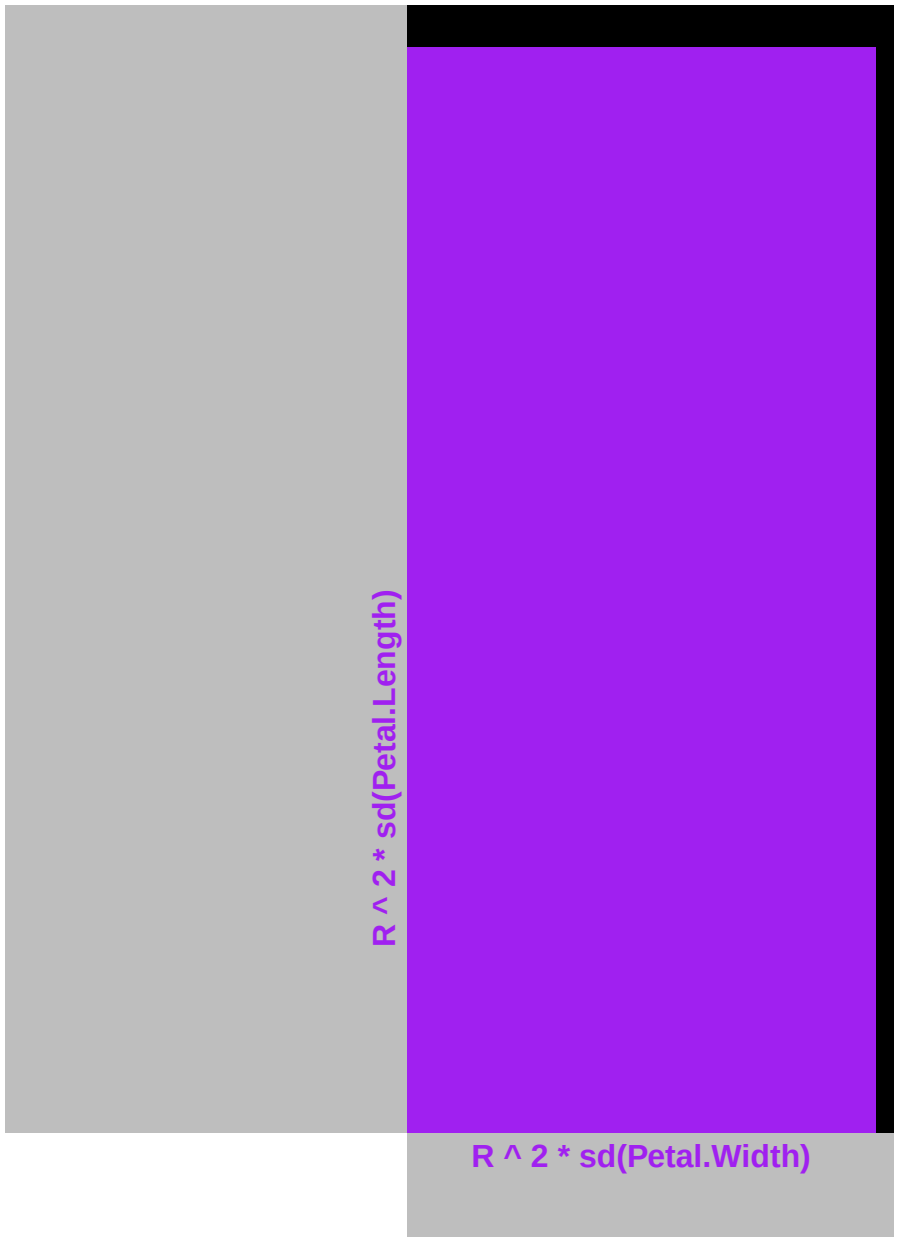


Squish covariance horizontally into the rectangle.

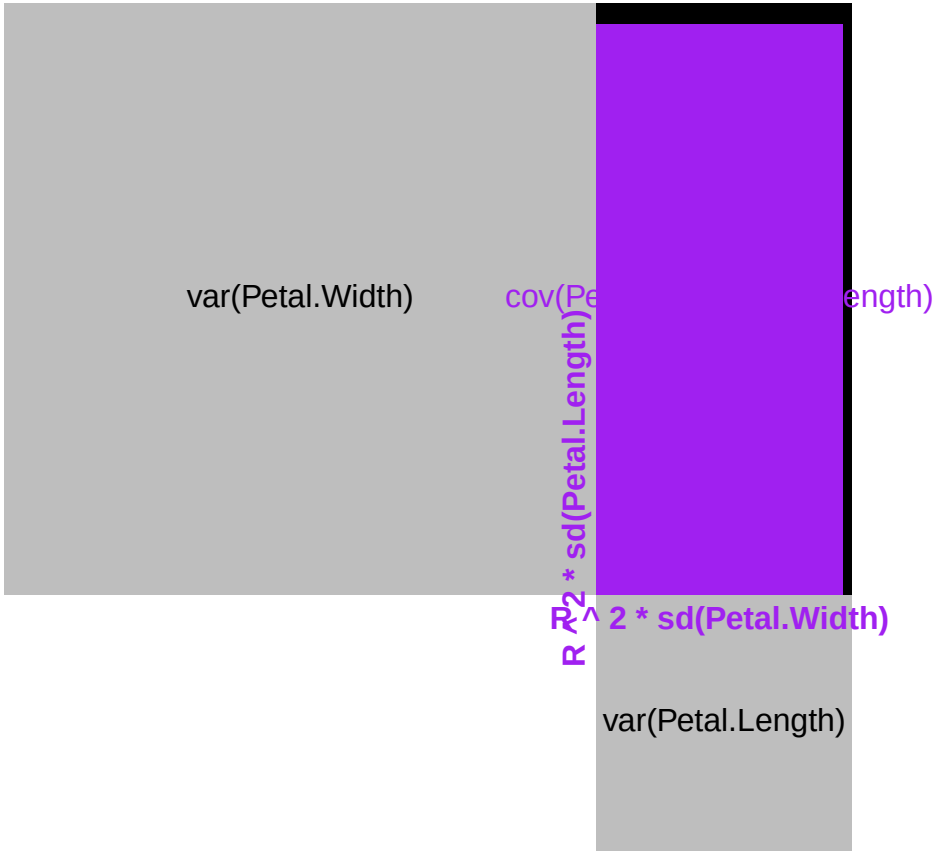


People like to
talk about R−squared.

Intersect the two squished covariance rectangles.



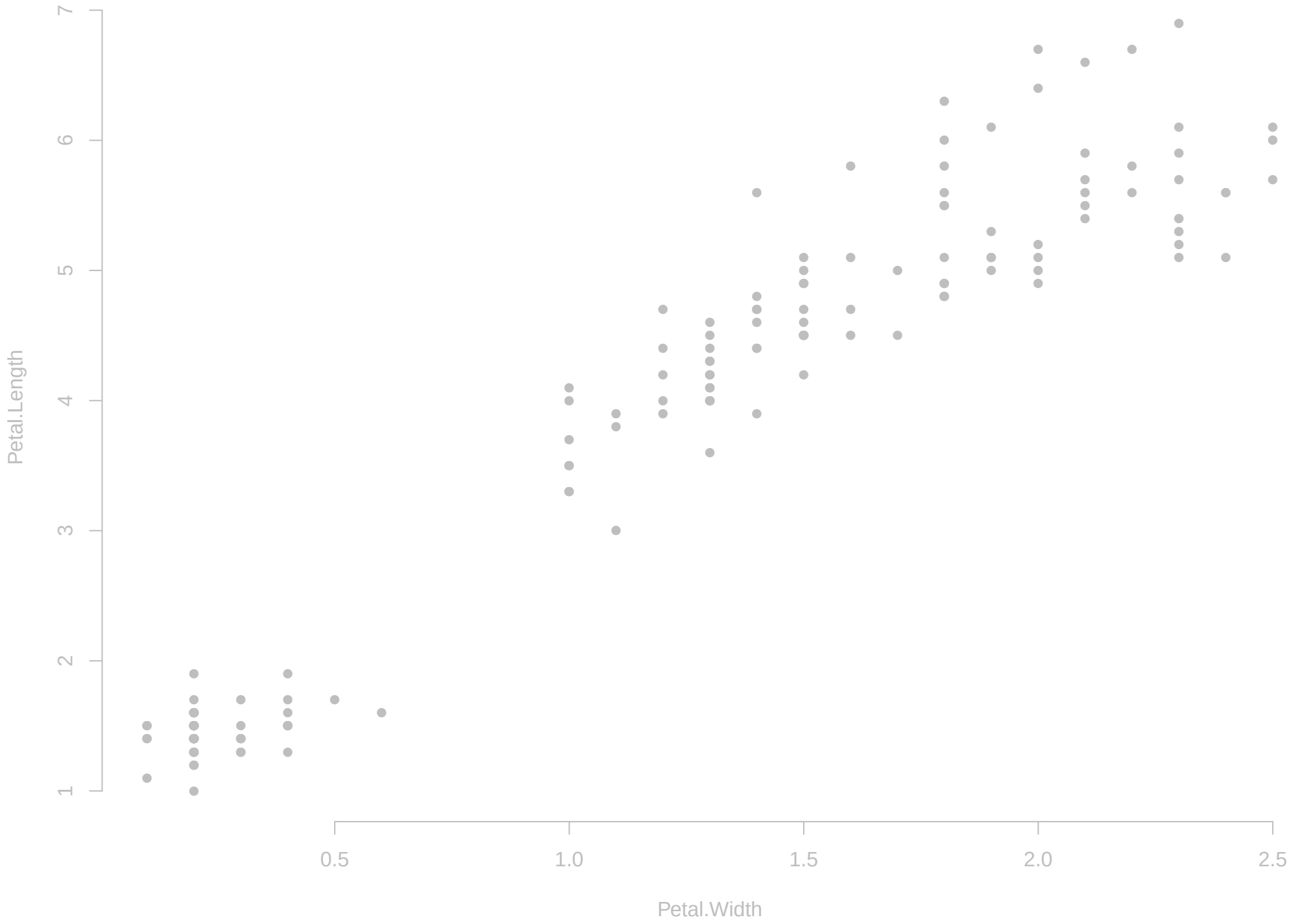
Intersect the two squished covariance rectangles.

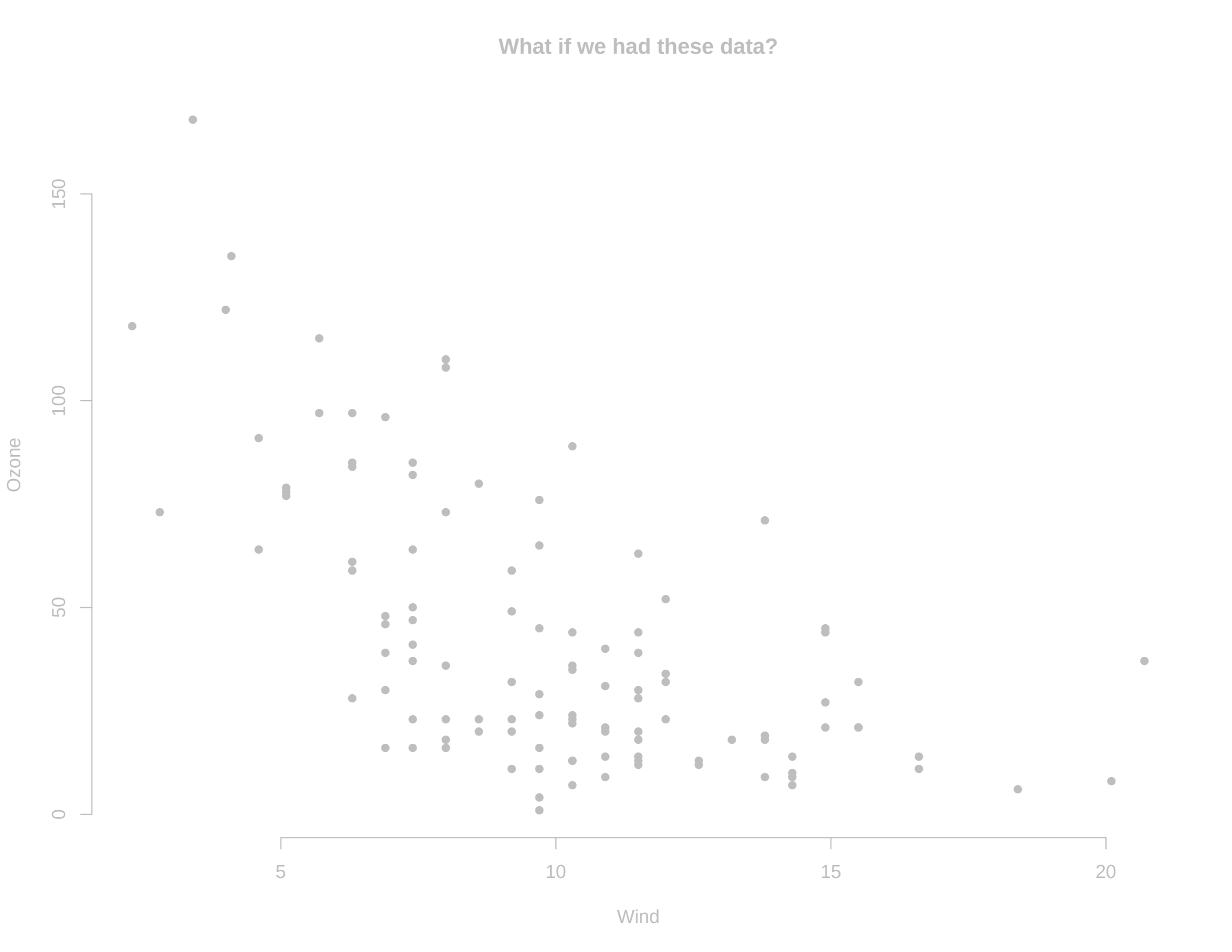


That was for very positive (blue) covariances.

What if covariance is negative (red)?

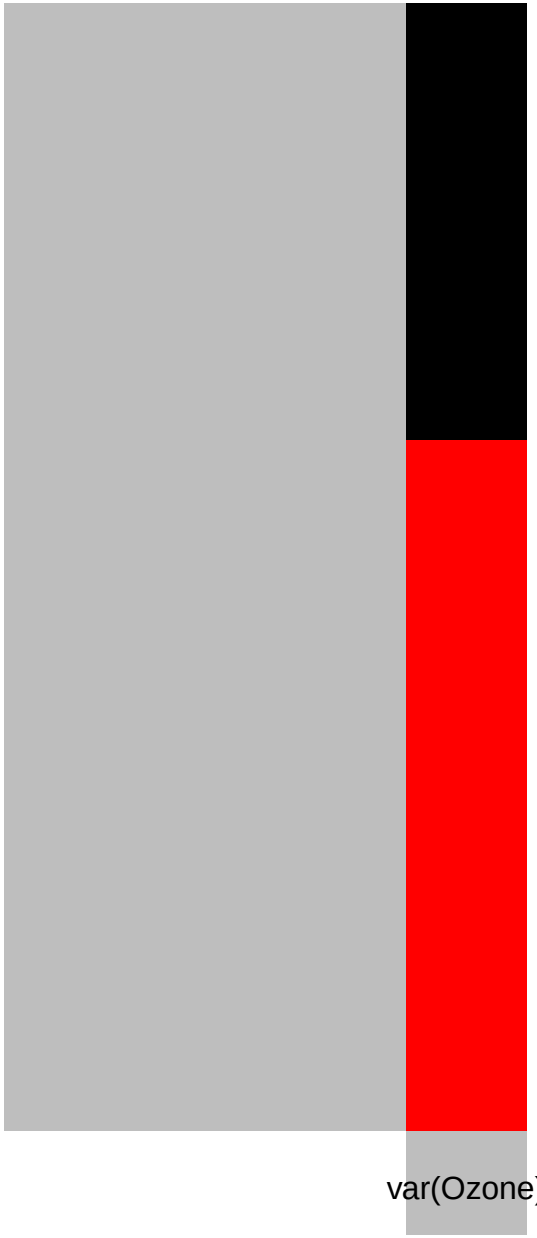
We were just using these data.



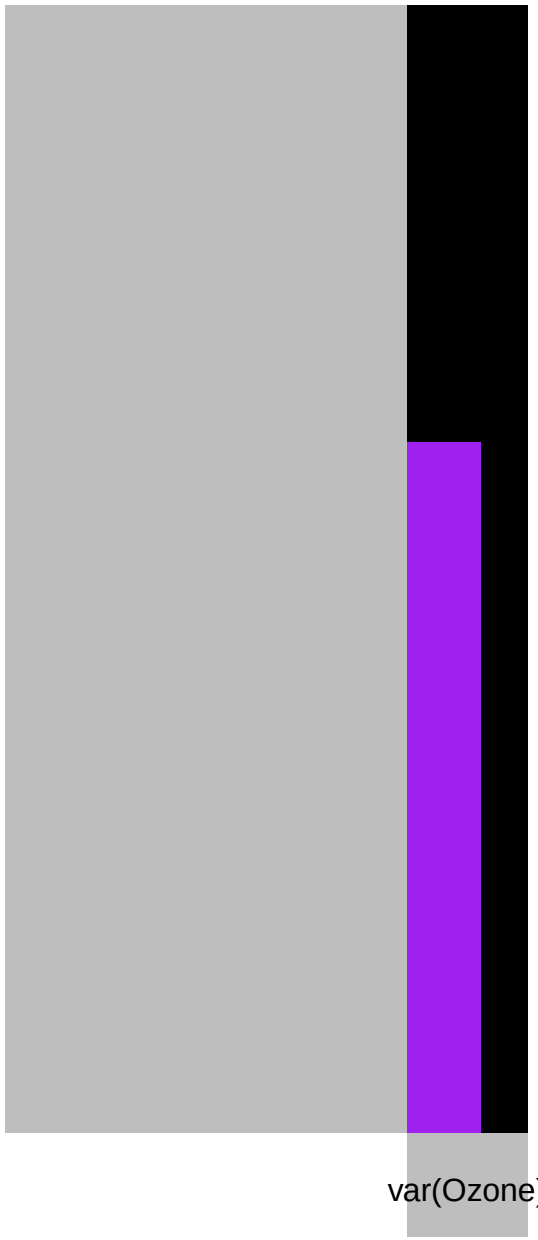

What if we had these data?
Ozone
0
50
100
150
5
10
15
20
Wind


What if we had these data?
Ozone
0
50
100
150
5
10
15
20
Wind

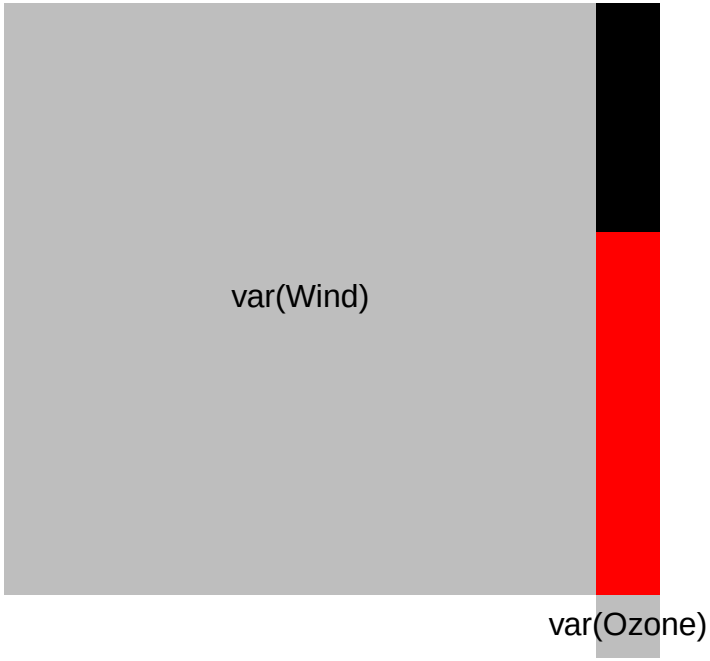
R is the same, just negative.



R−squared is the same, and it is always positive.



Zoom back out.



If we transform the covariance a bit, we can also make predictions.

Let's use x to predict y.

$$y = b0 + b1 * x$$

Let's invent b1.

What values should it have?

If covariance is very positive and x is high, y should be high.

(We want b1 to be positive.)

If covariance is very negative and x is high, y should be low.

(We want b1 to be negative.)

If covariance is low,
we have no idea what y is.

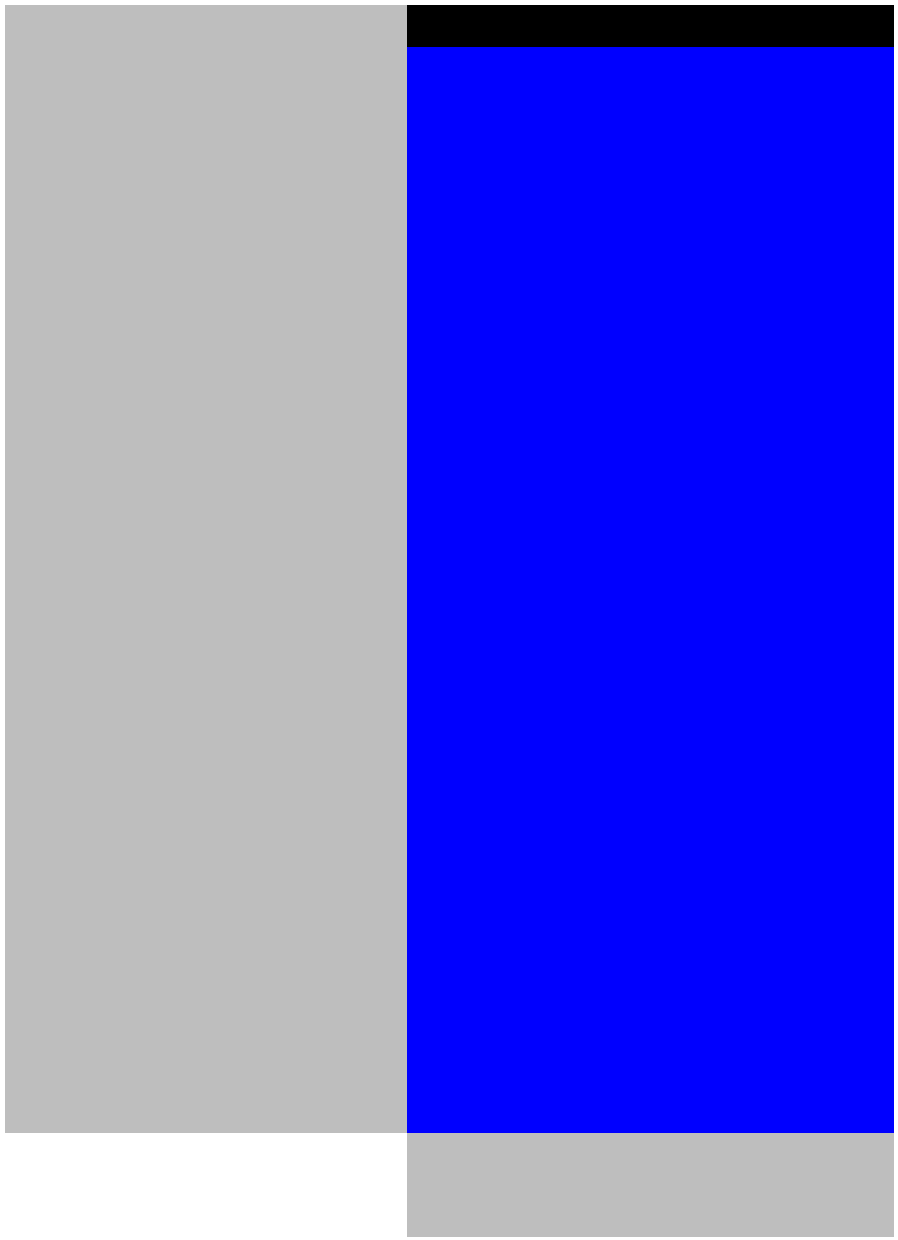
(b1 is around zero.)

Let's think about units again.

Covariance is an area; its unit is the product of the x and y units.

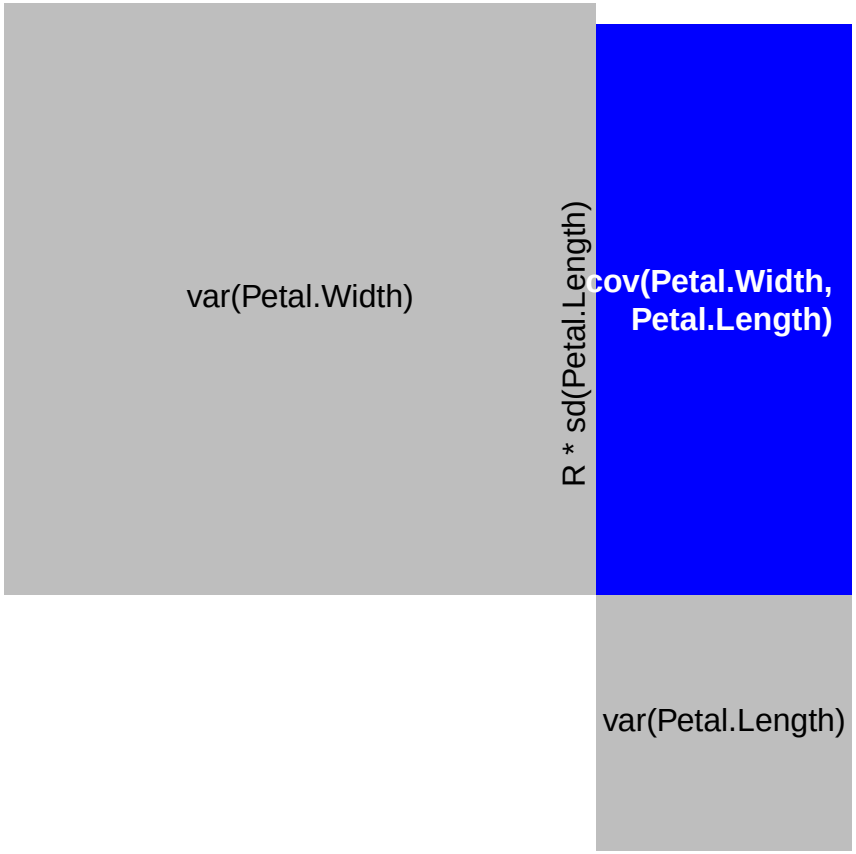
Variance is a special covariance; its unit is the square of the x unit.

Correlation is a ratio of areas with the same units.

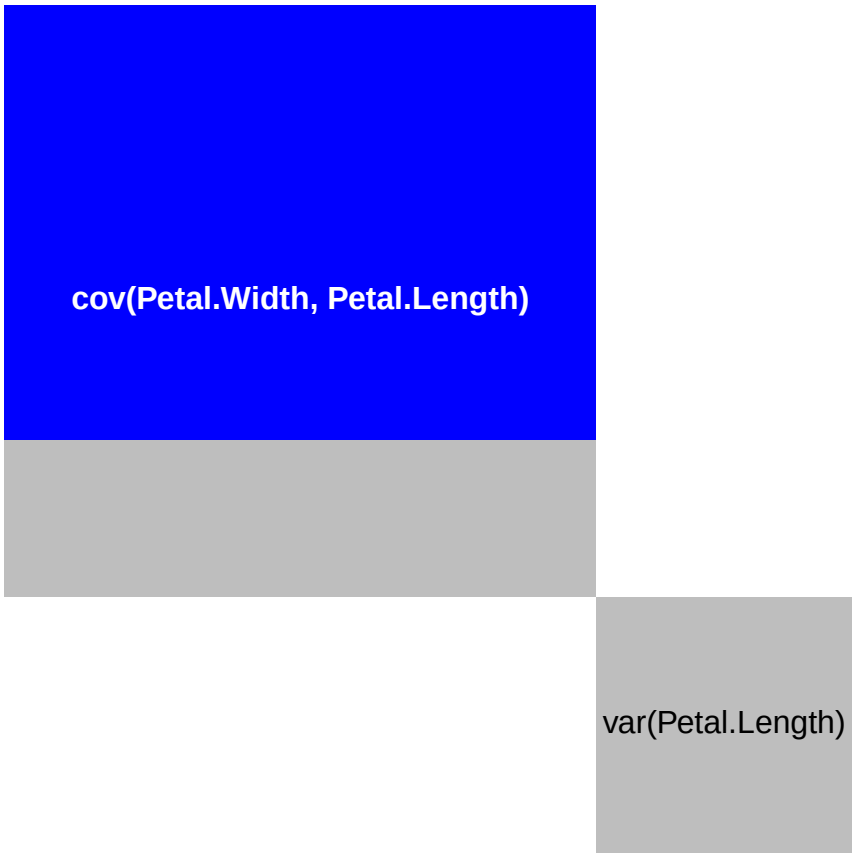


The unit of b1 must be y−unit/x−unit.

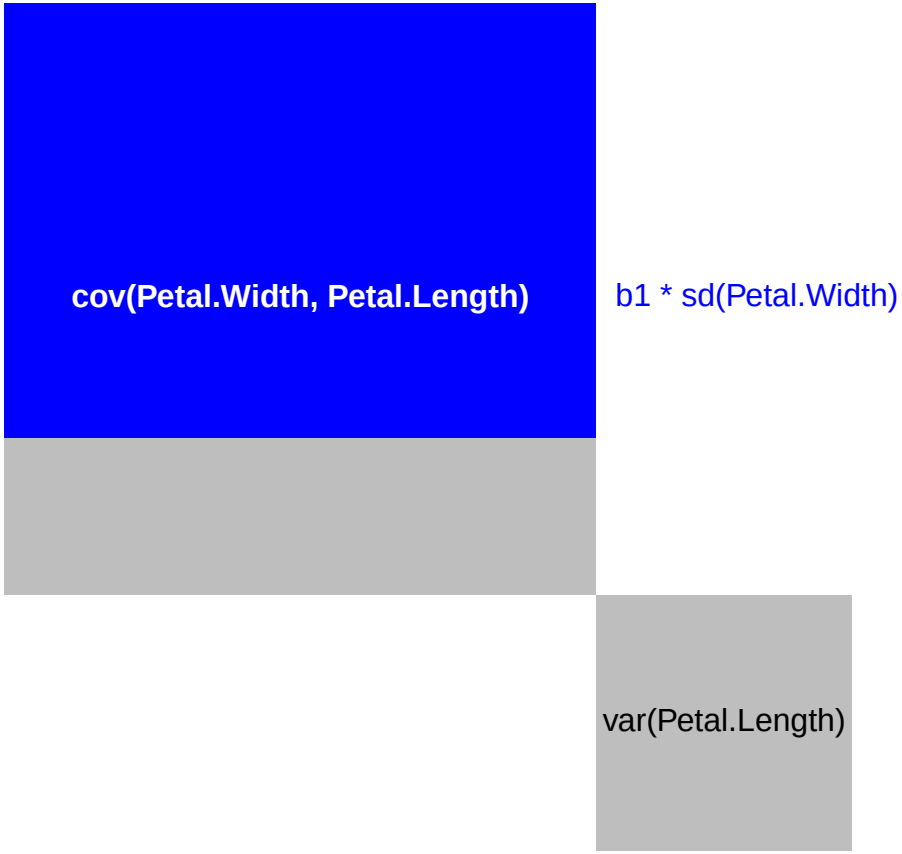
Our covariance picture



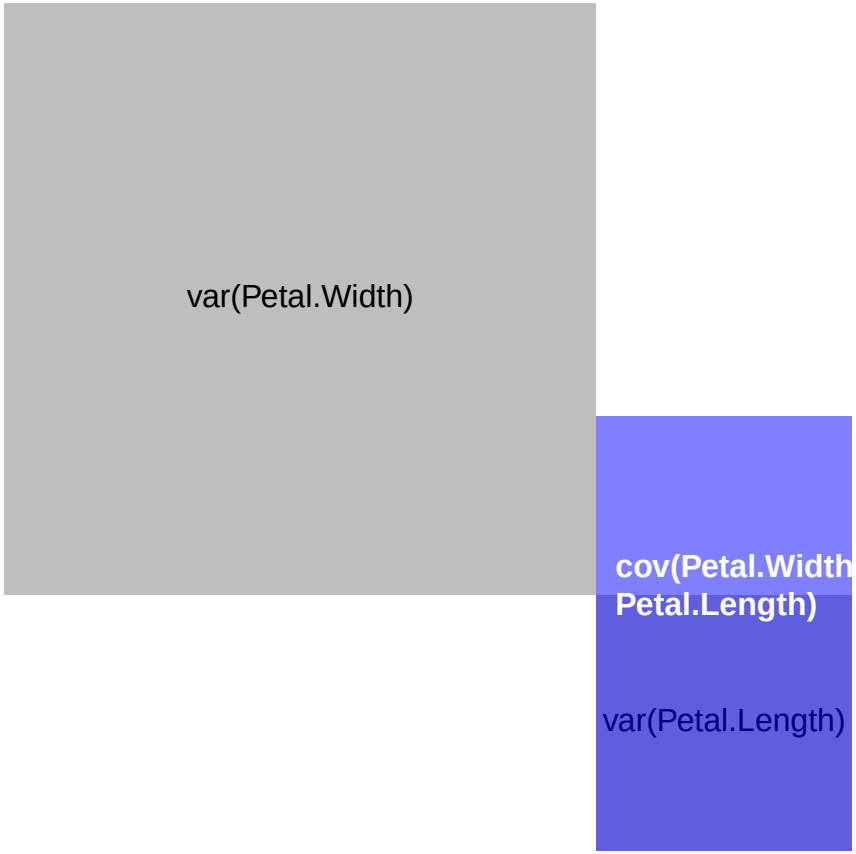
Lay the covariance over one of the variances instead.



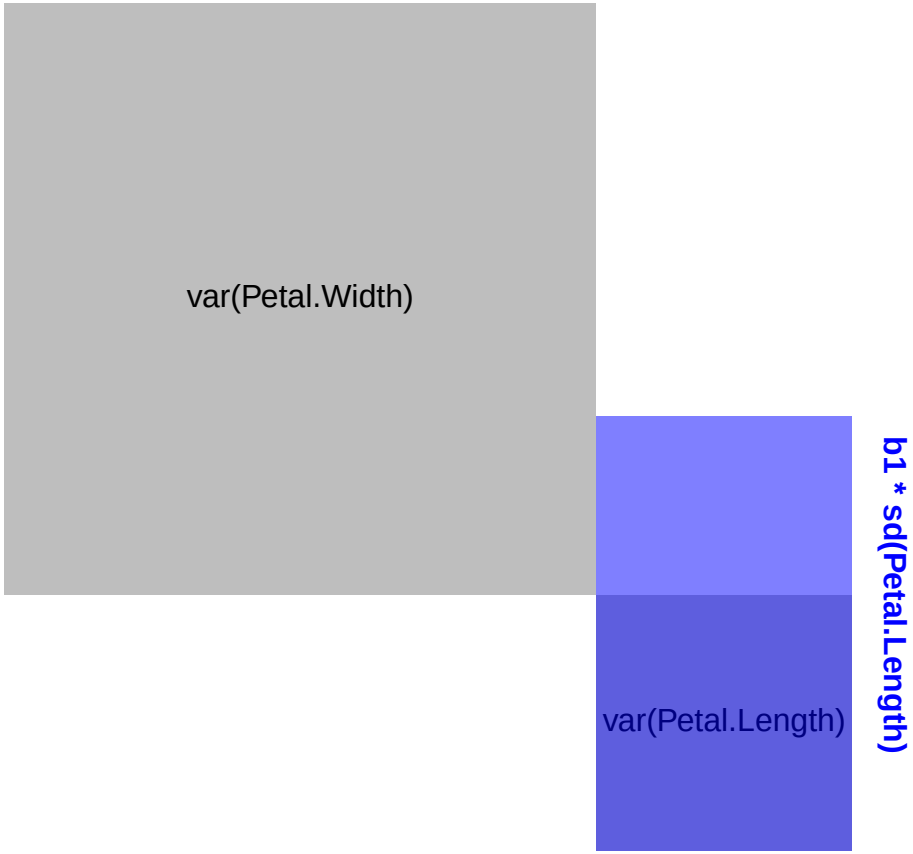
Petal.Length = b0 + b1 * Petal.Width



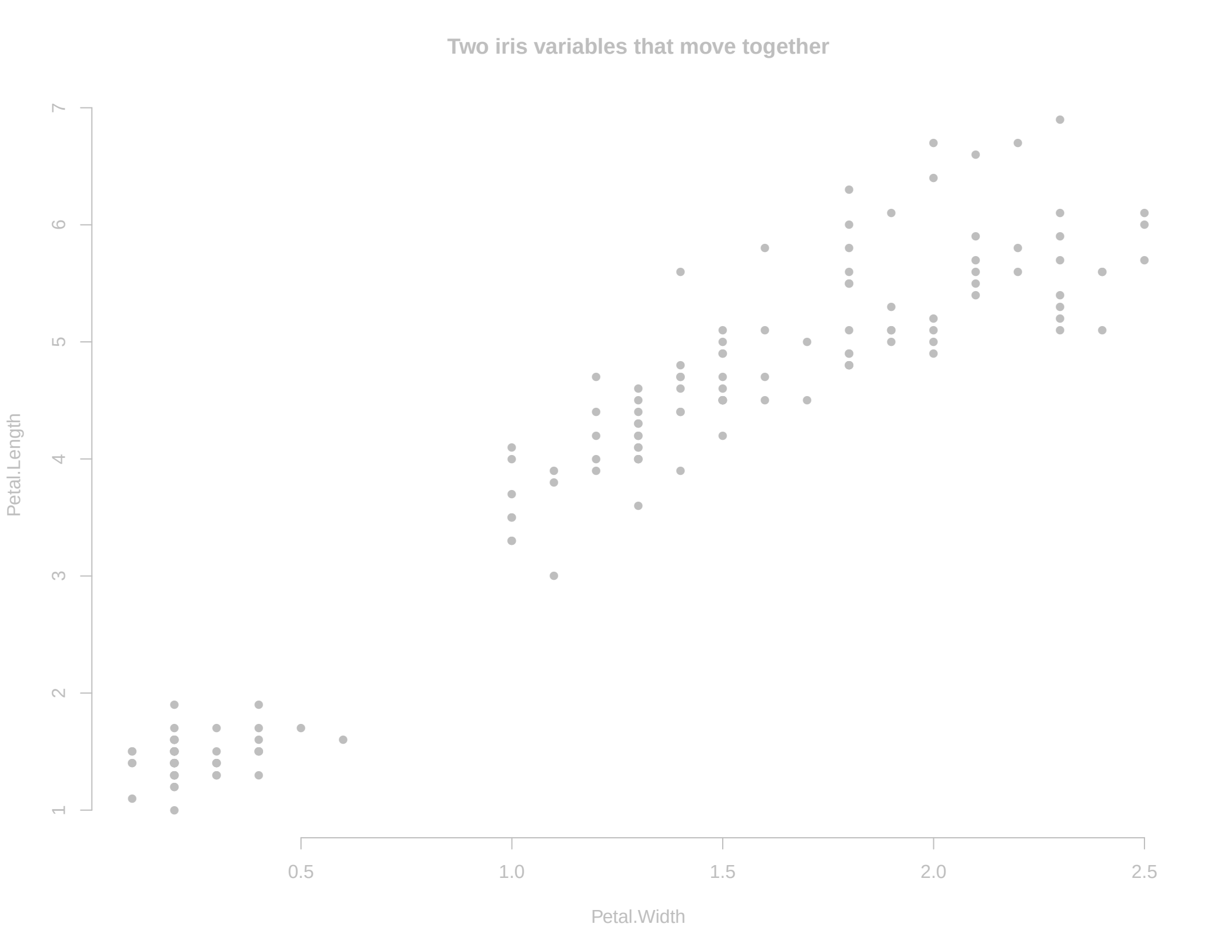
Lay the covariance over the other variance.

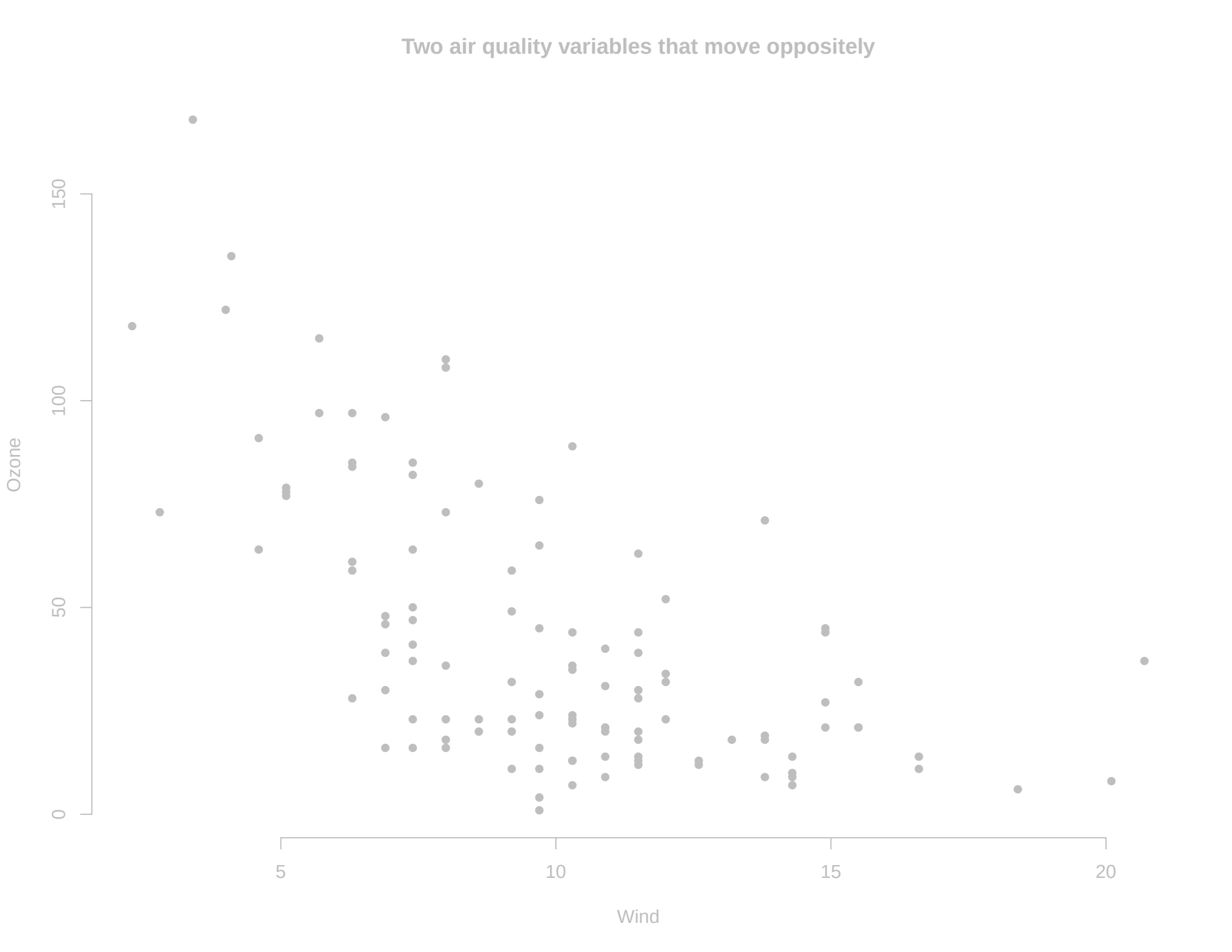


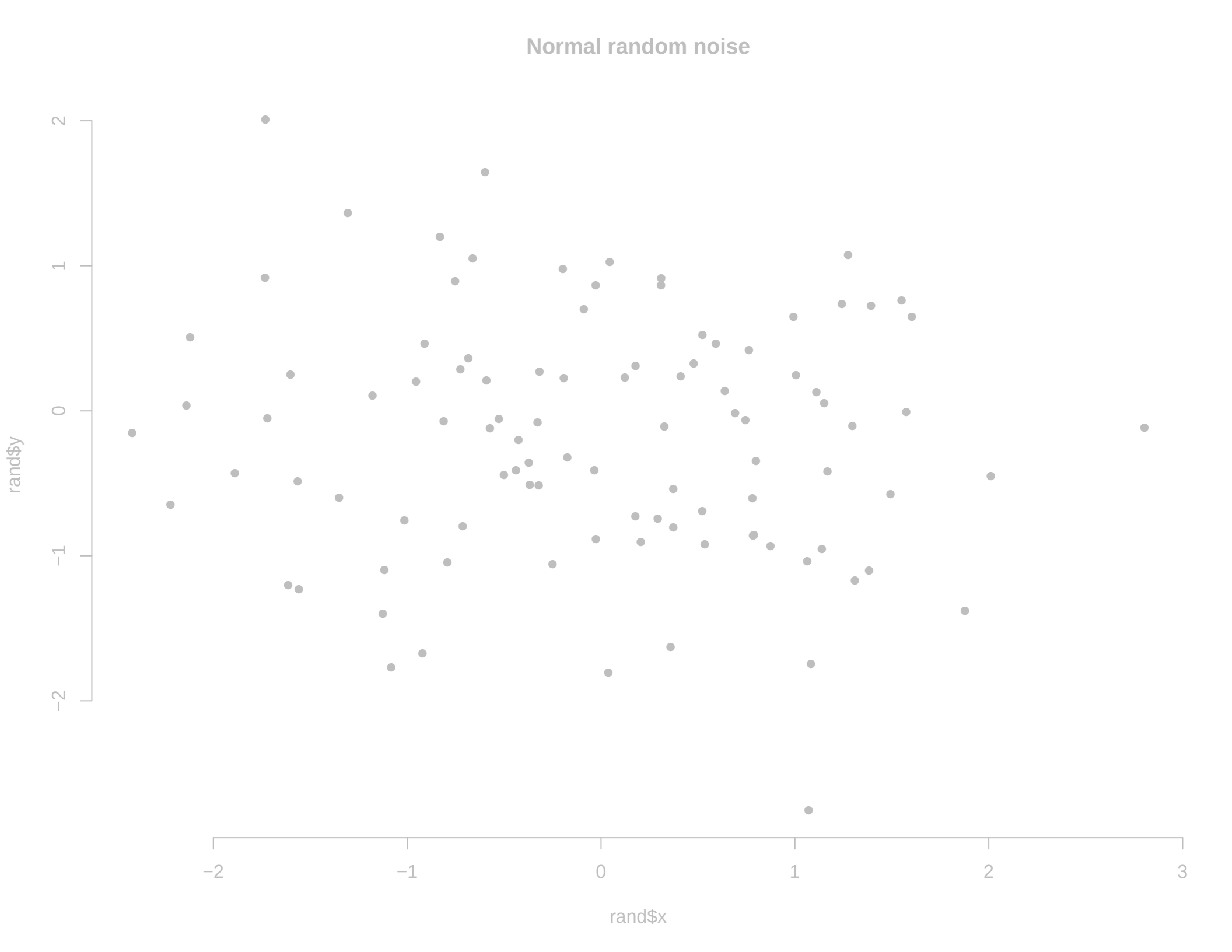
Petal.Width = b0 + b1 * Petal.Length



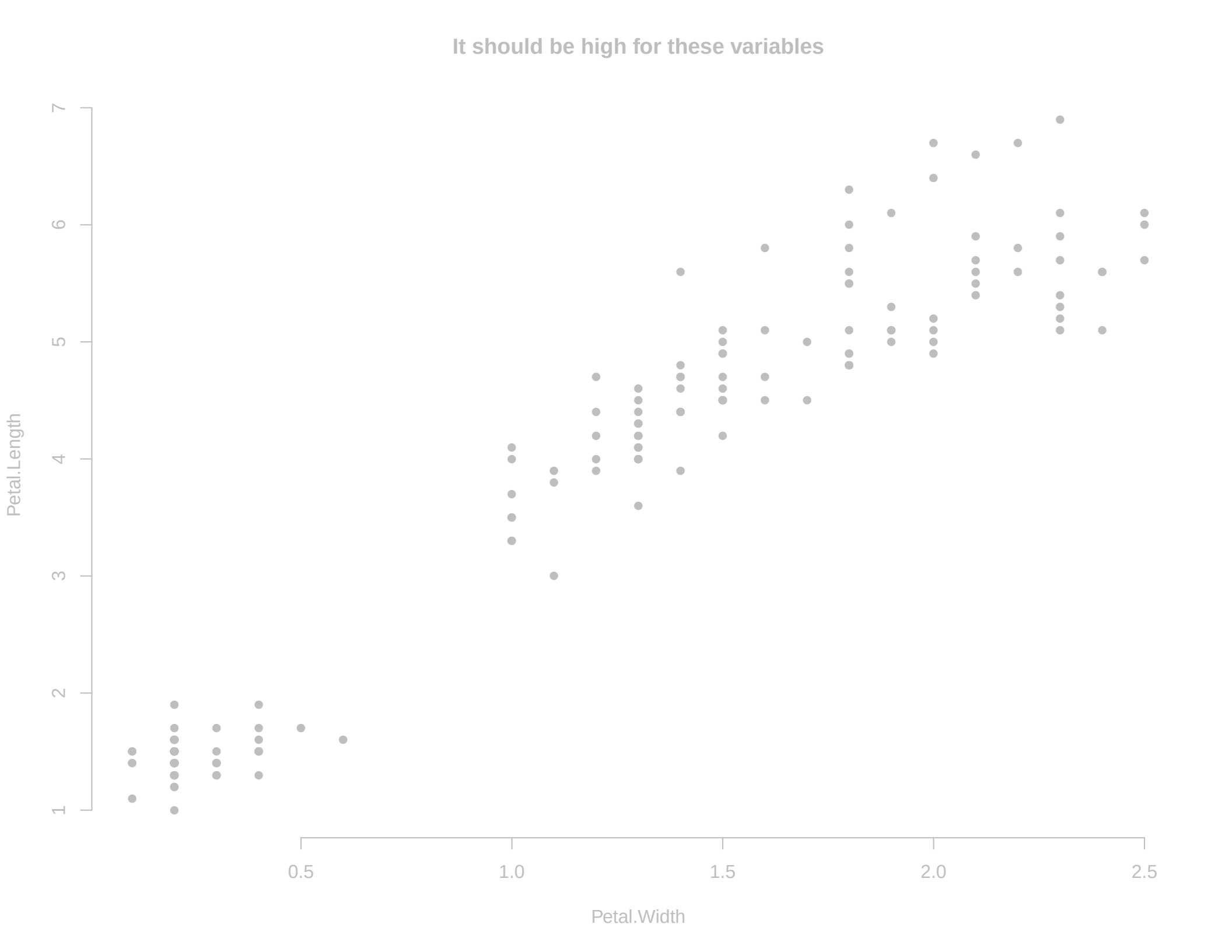
Let's go over everything one last time.

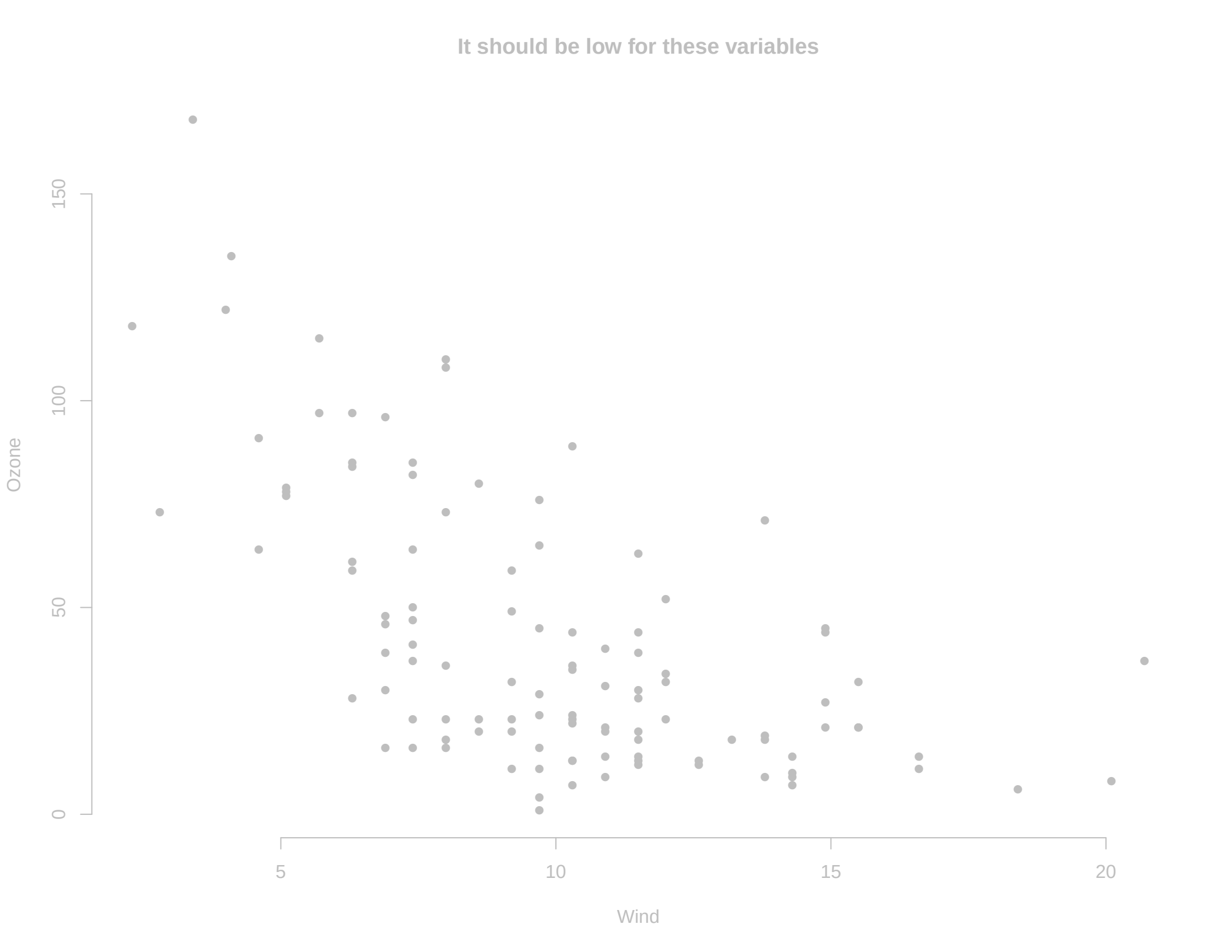

Two iris variables that move together
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

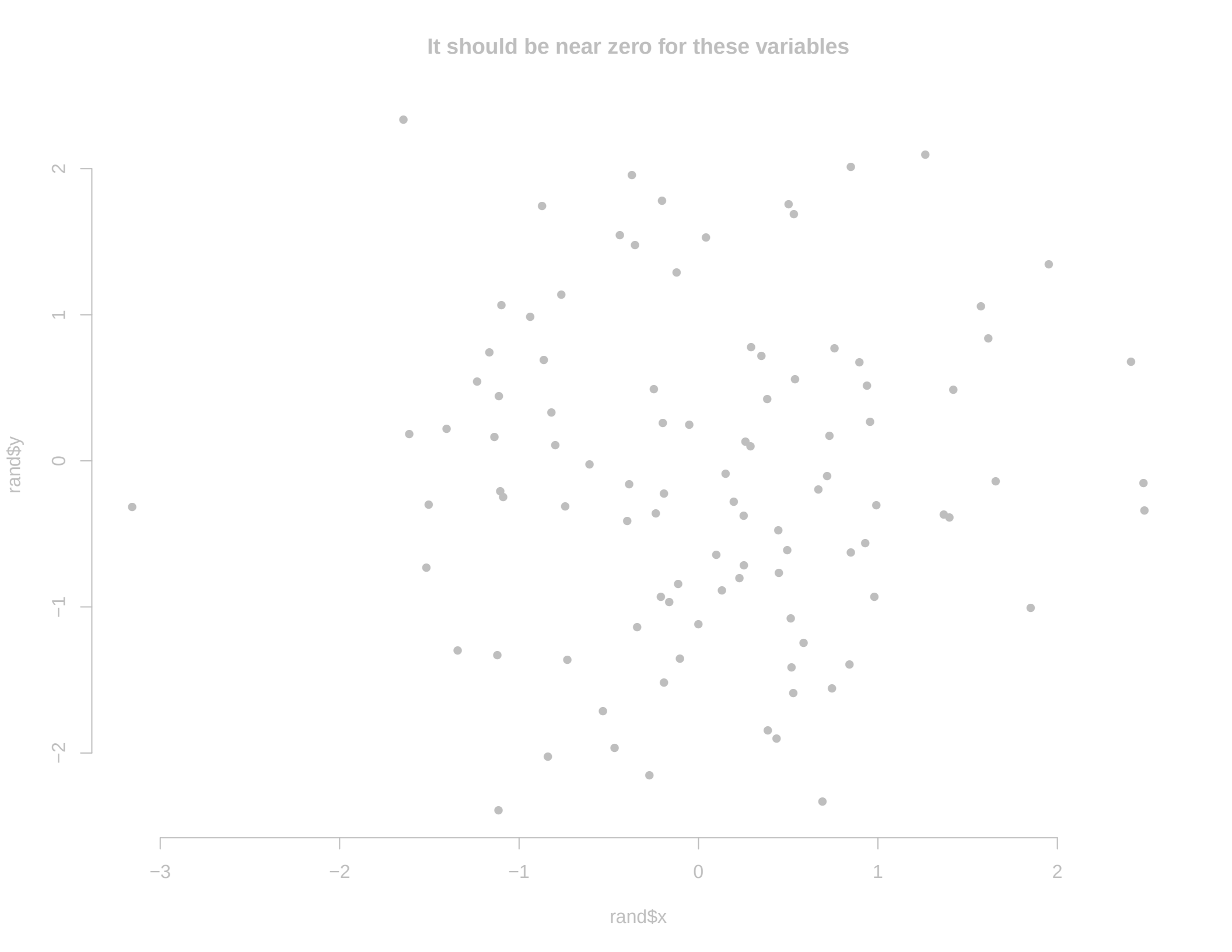

Two air quality variables that move oppositely
Ozone
0
50
100
150
5
10
15
20
Wind


Normal random noise
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

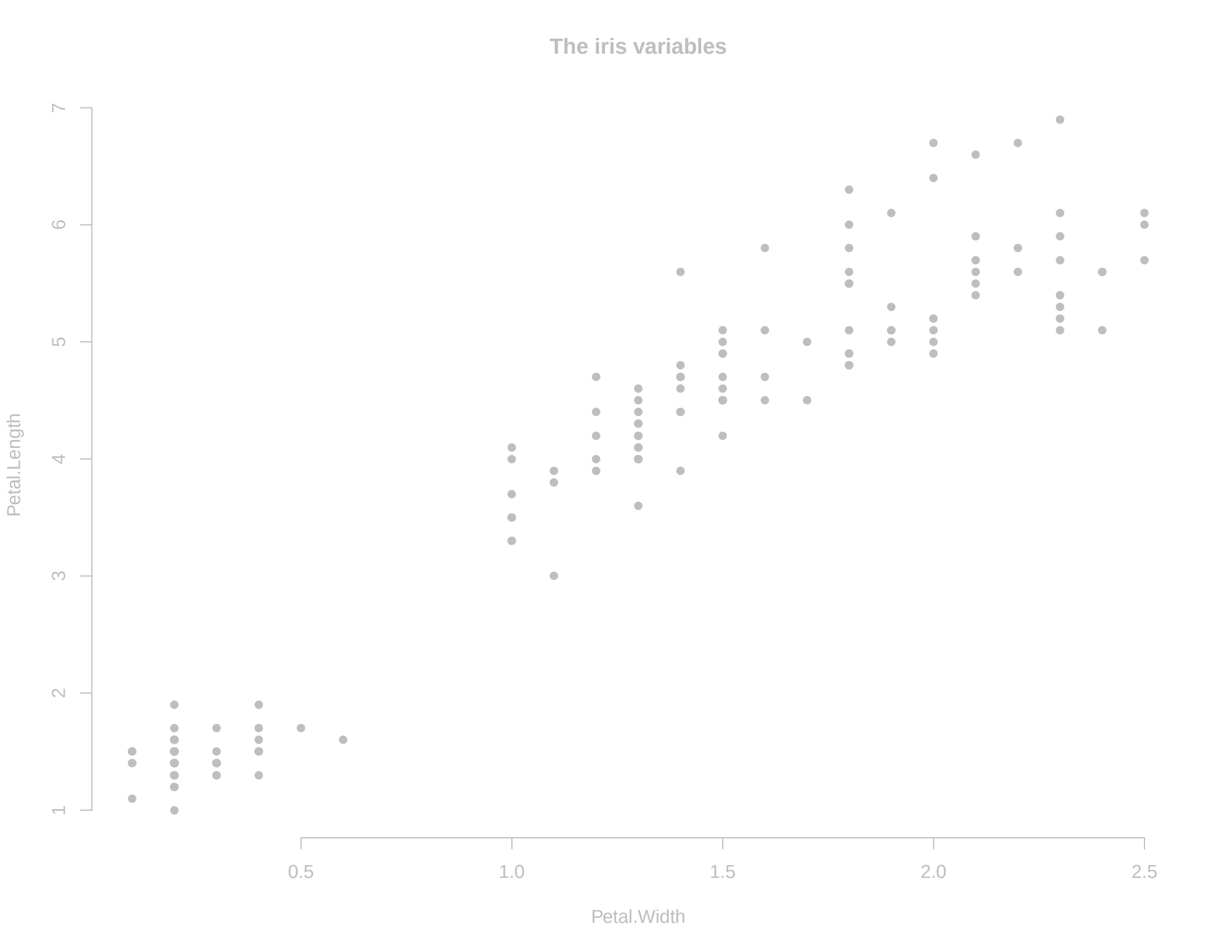
We want a number that describes whether two variables move together.

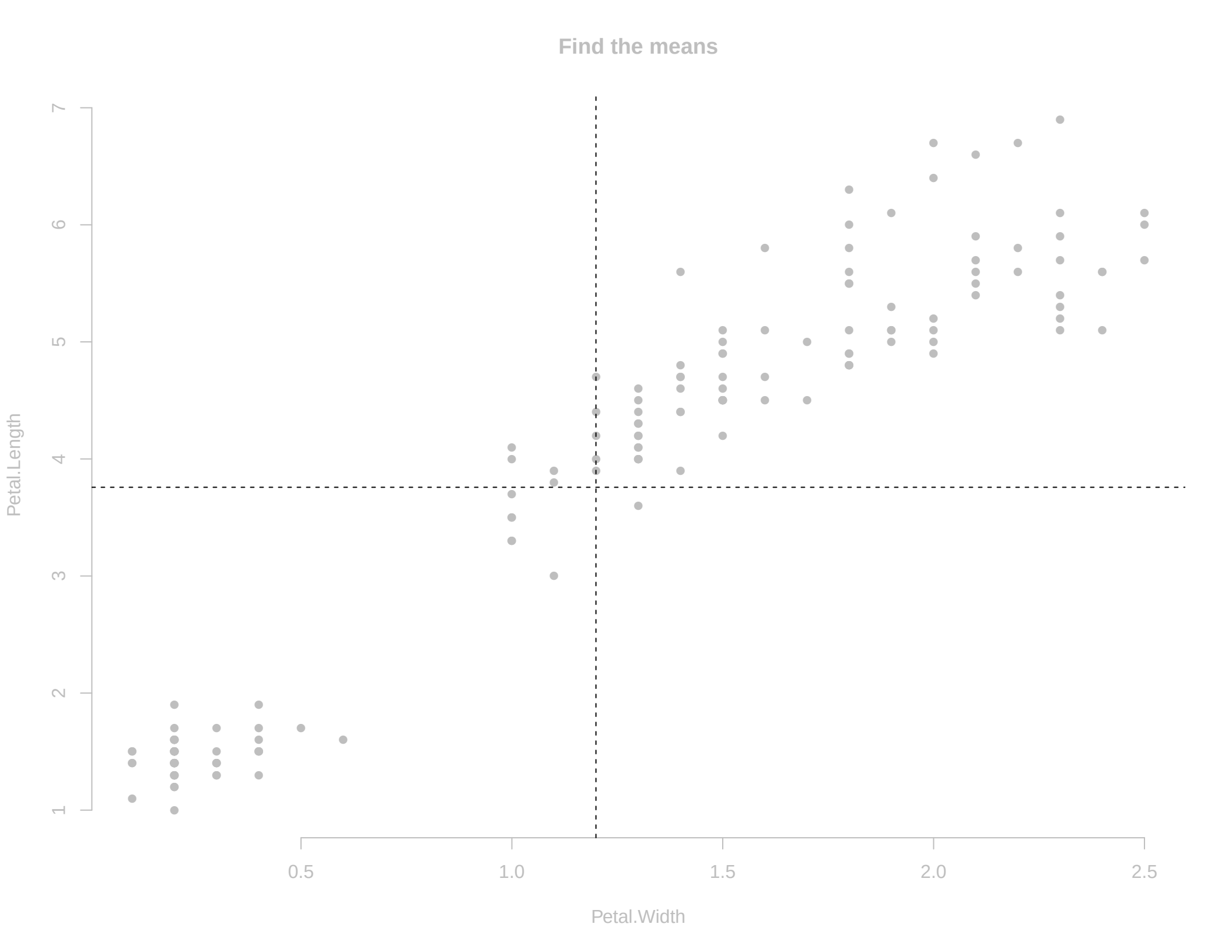

It should be high for these variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

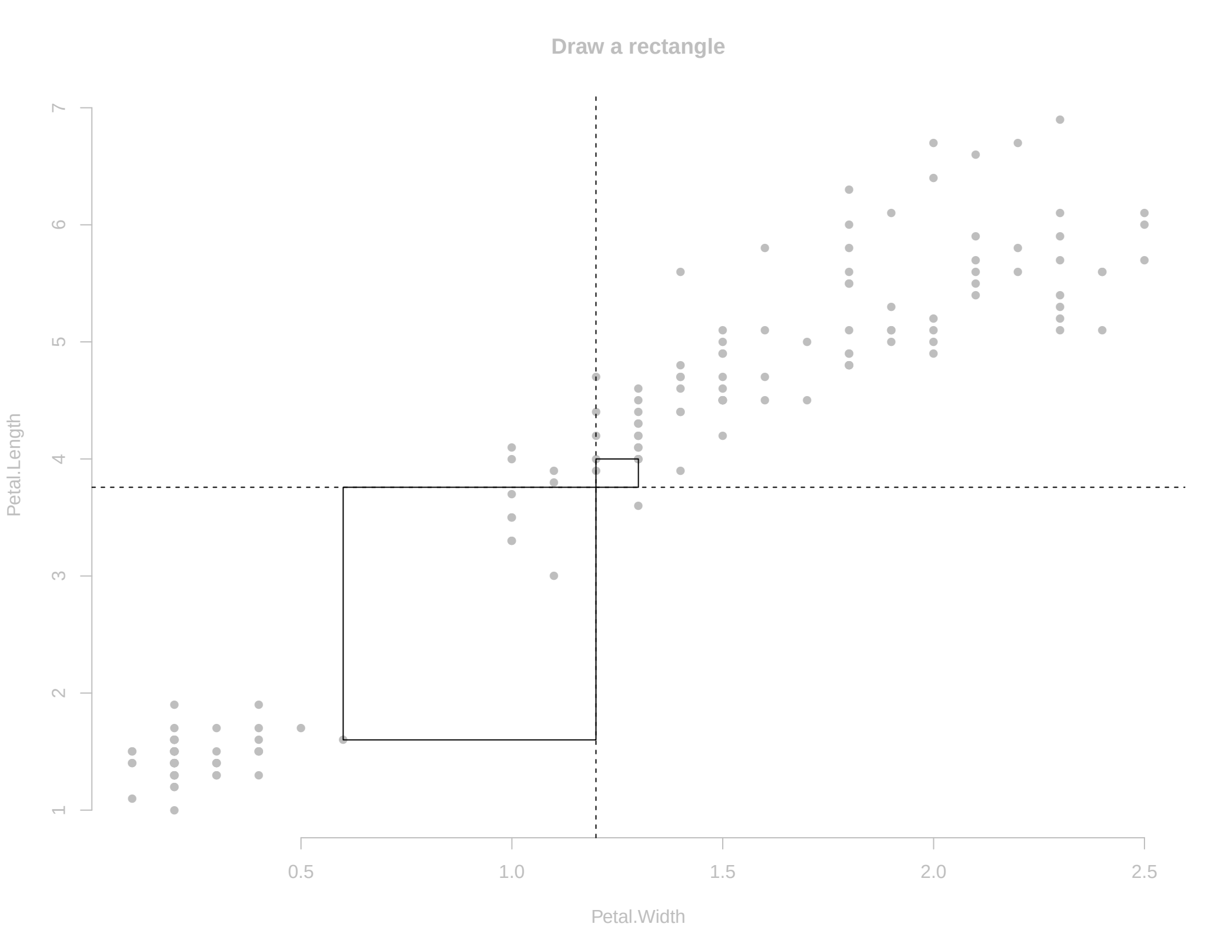

It should be low for these variables
Ozone
0
50
100
150
5
10
15
20
Wind

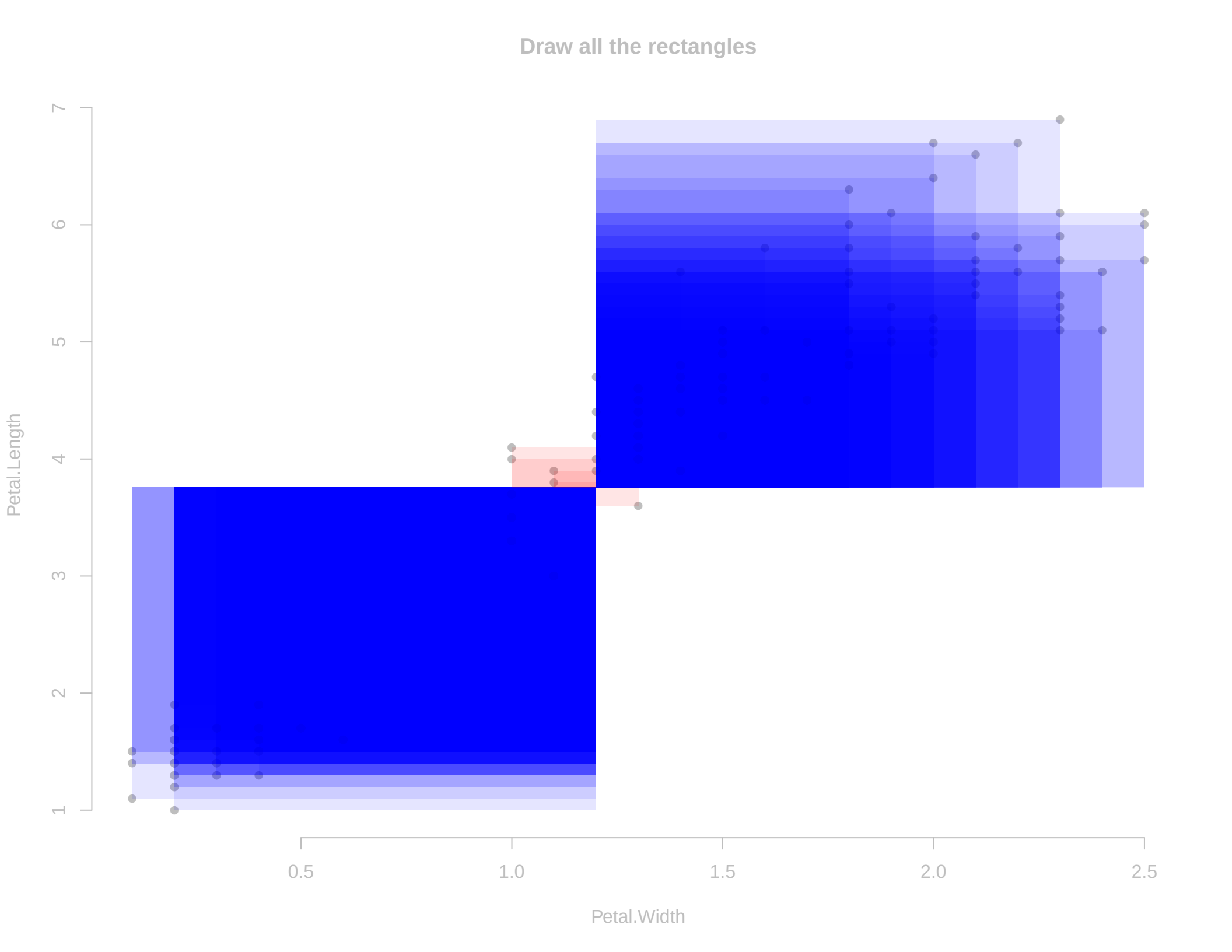

It should be near zero for these variables
rand$y
2
1
0
−1
−2
−3
−2
−1
0
1
2
rand$x

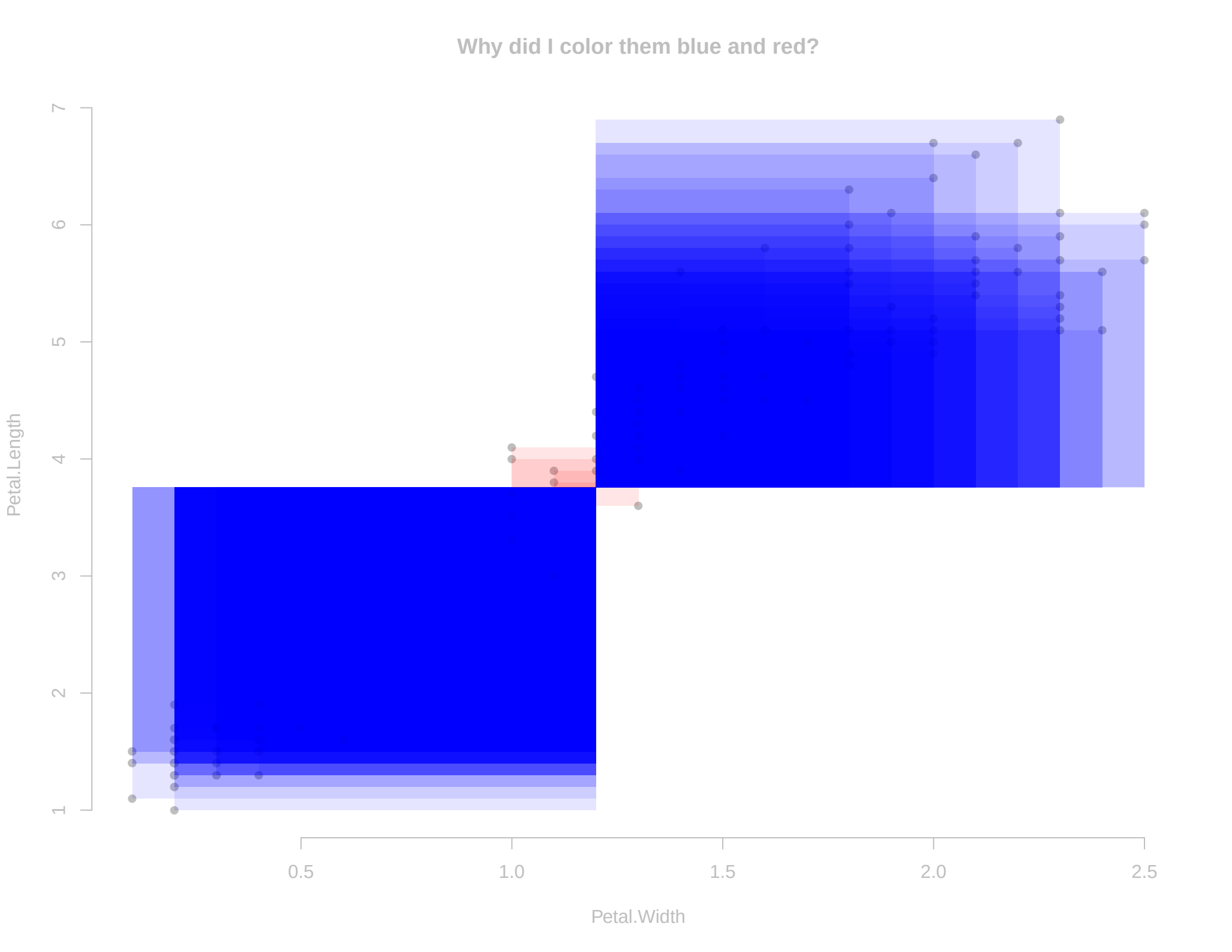
Covariance

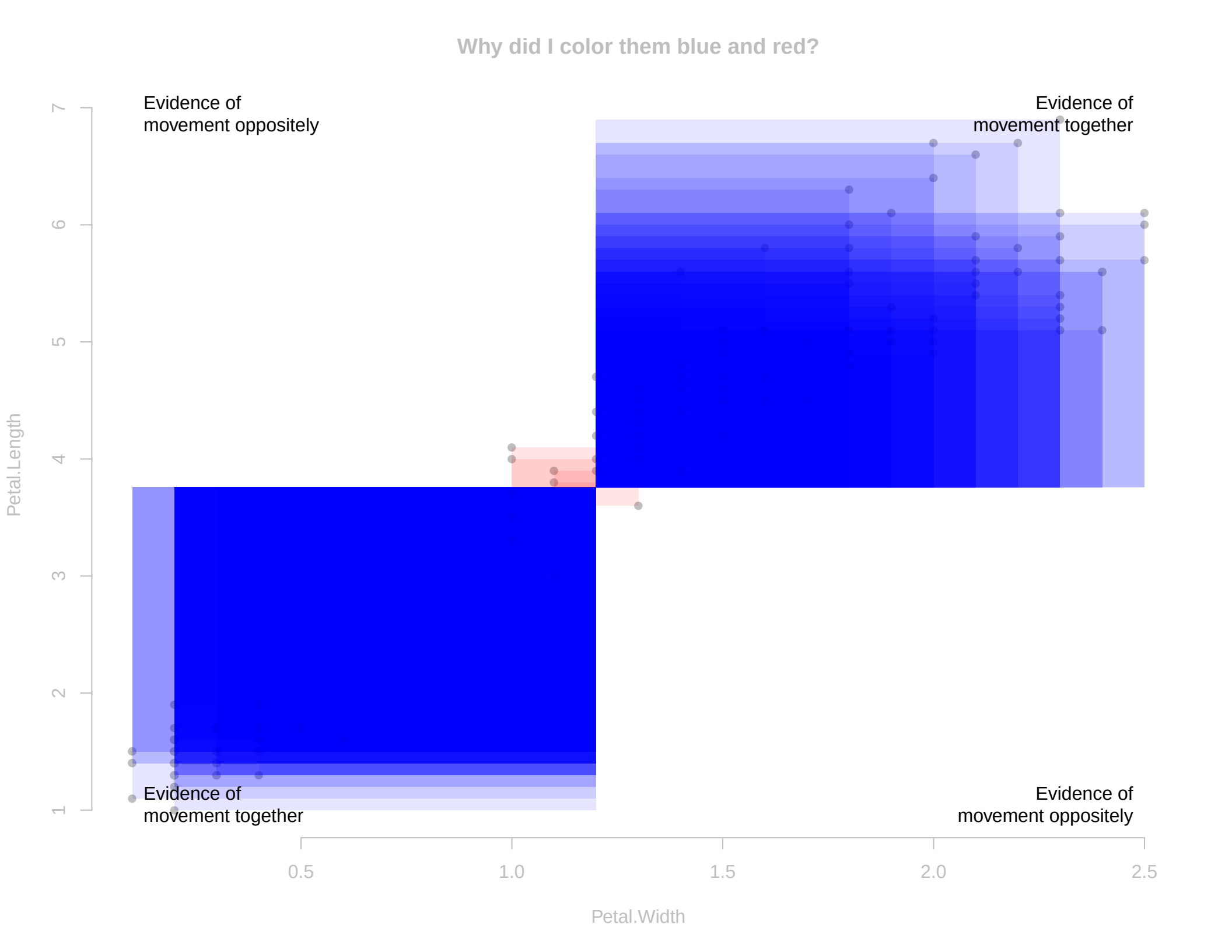

The iris variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Find the means
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

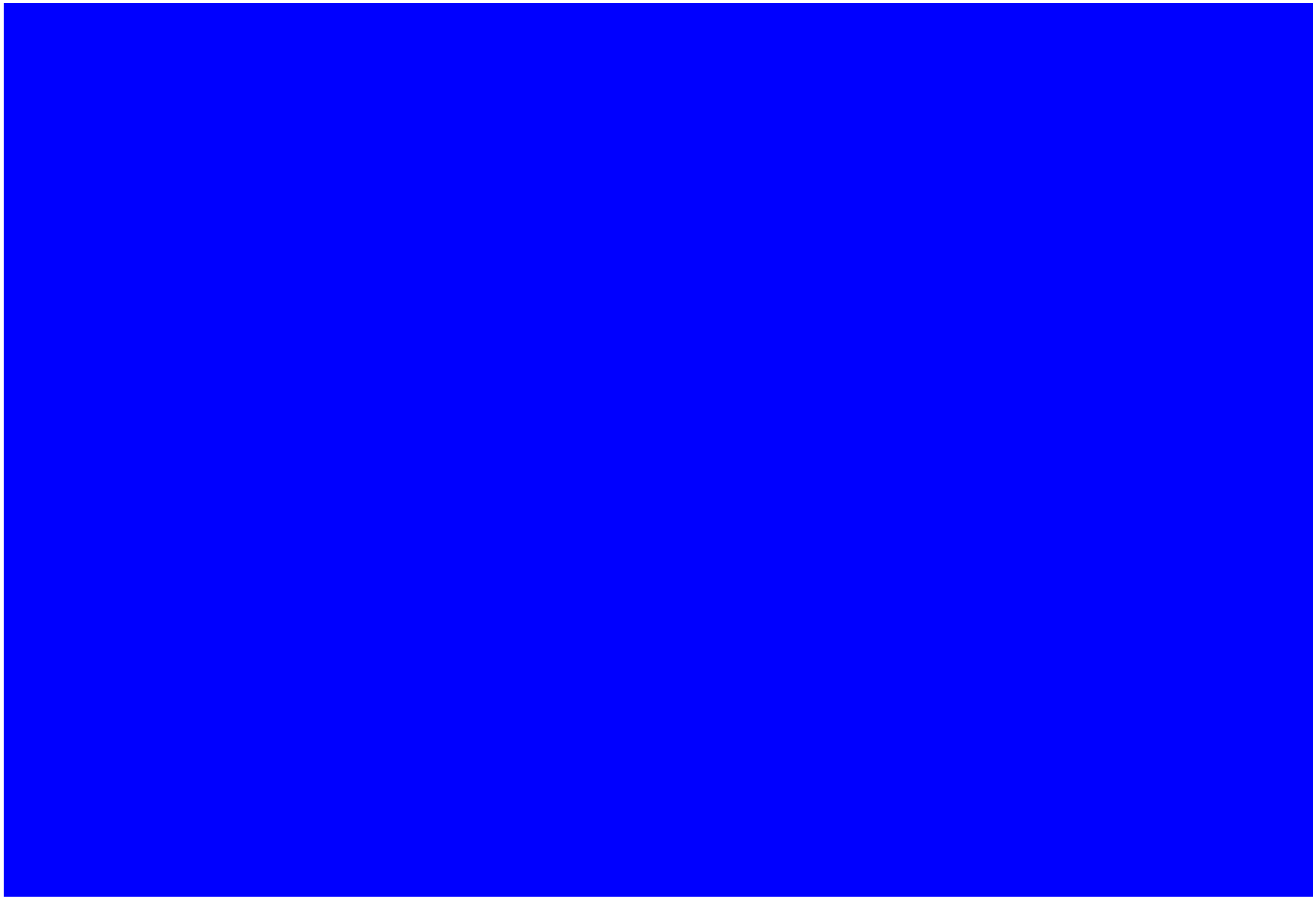

Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

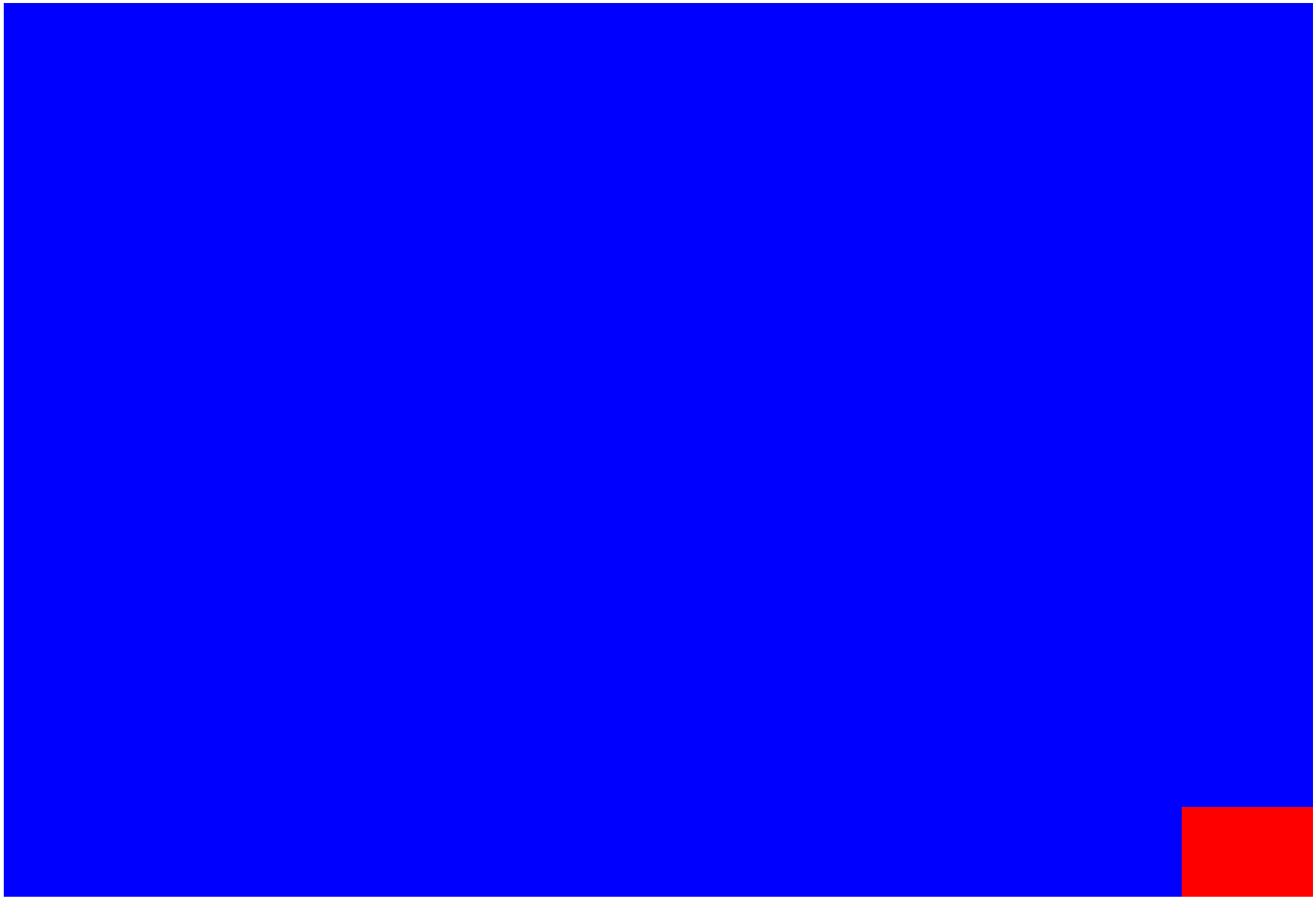

Why did I color them blue and red?
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


Why did I color them blue and red?
Evidence of
movement oppositely
Evidence of
movement together
Evidence of
movement together
Evidence of
movement oppositely
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

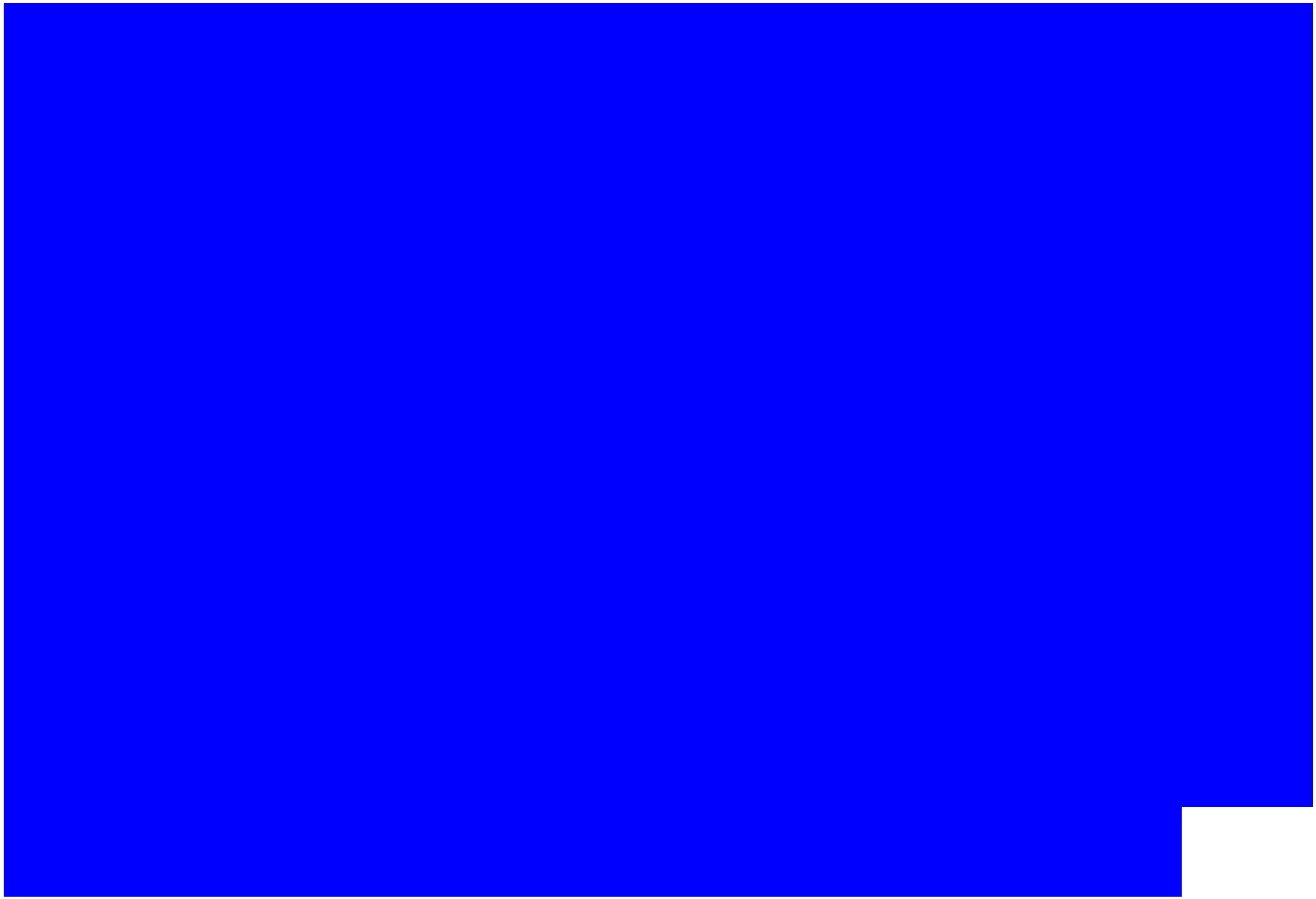
Add the blues together. (This is at a different scale.)



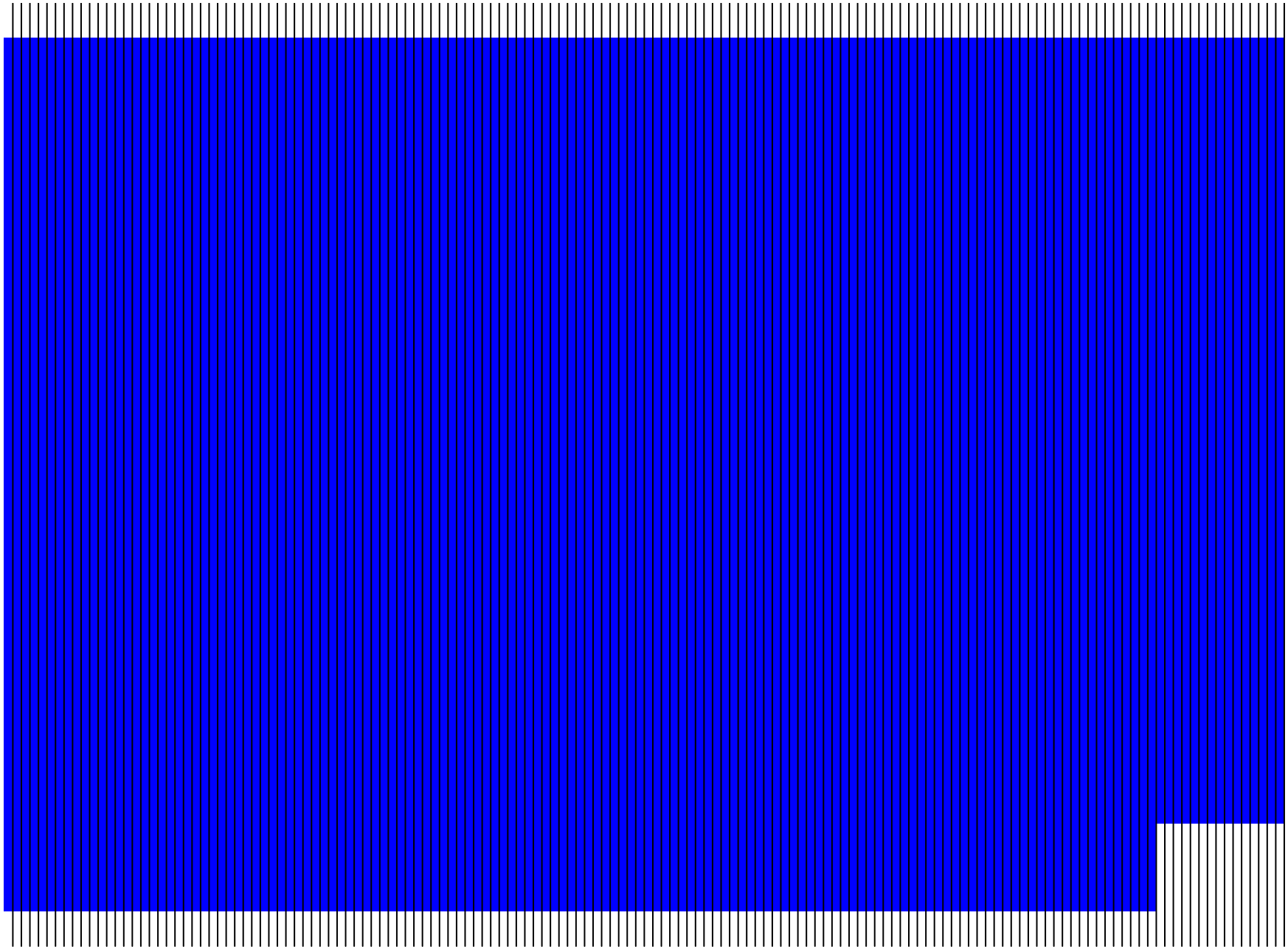
Add the reds together.



Subtract the reds.

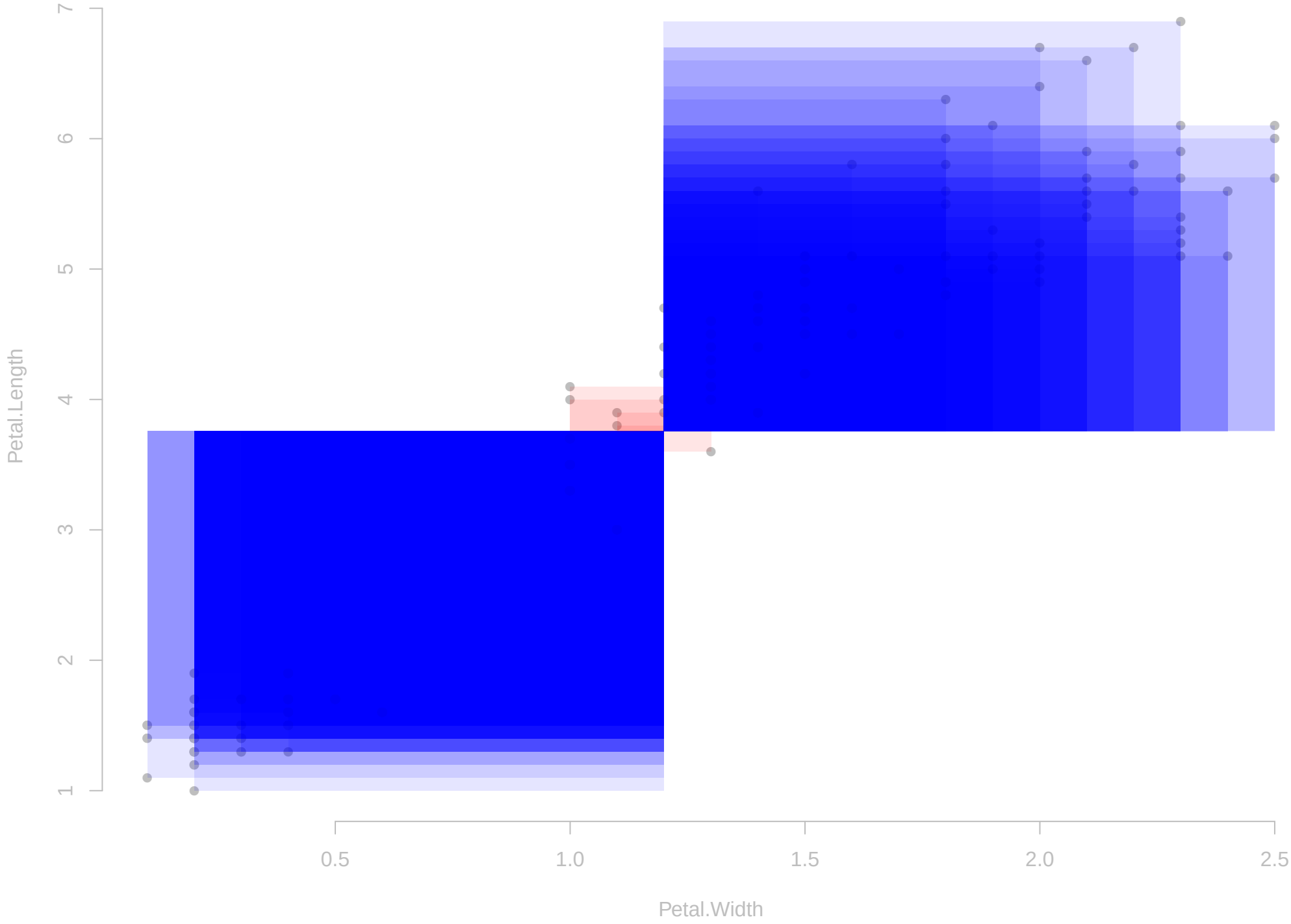


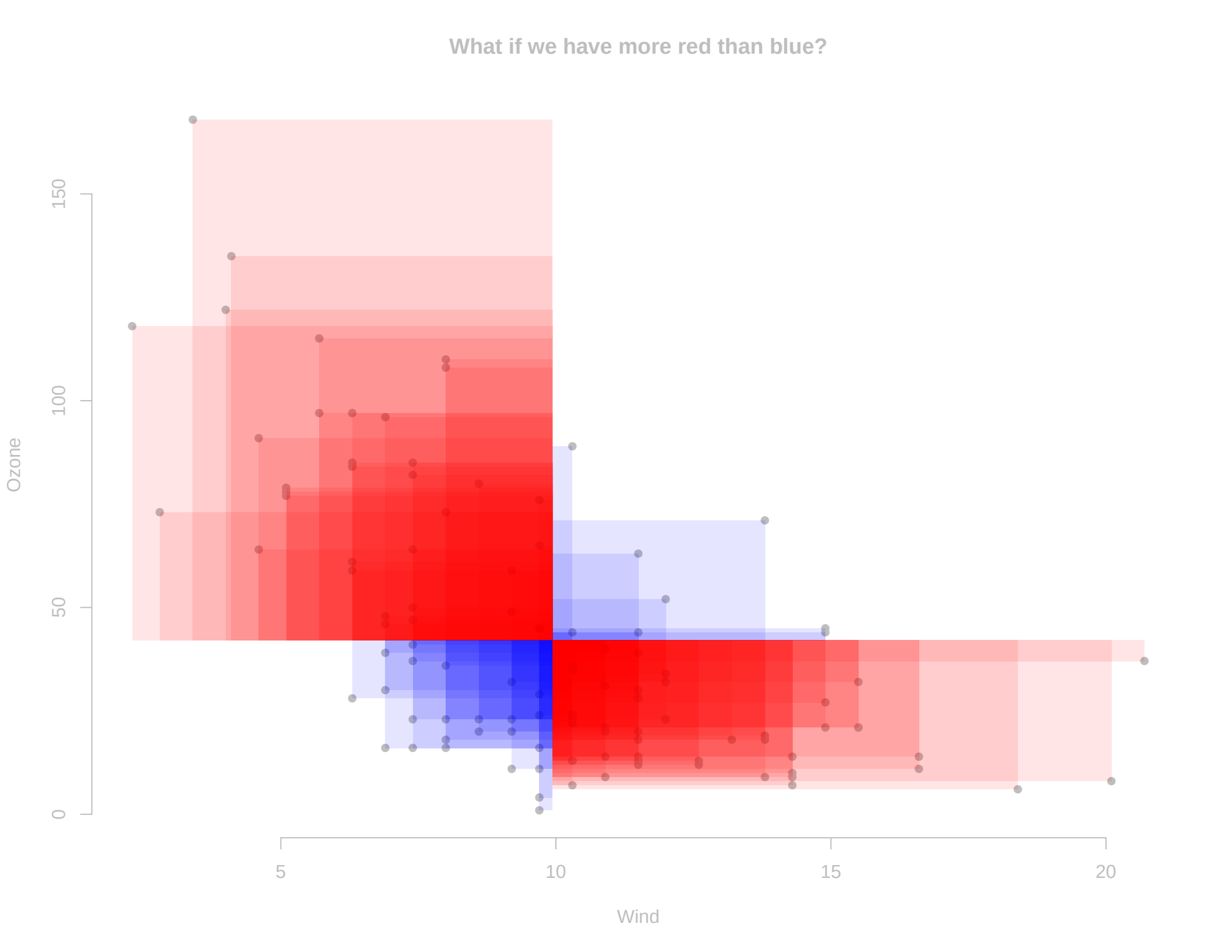
Divide into as many equal pieces as we have irises (n).



This blue sliver is the covariance.

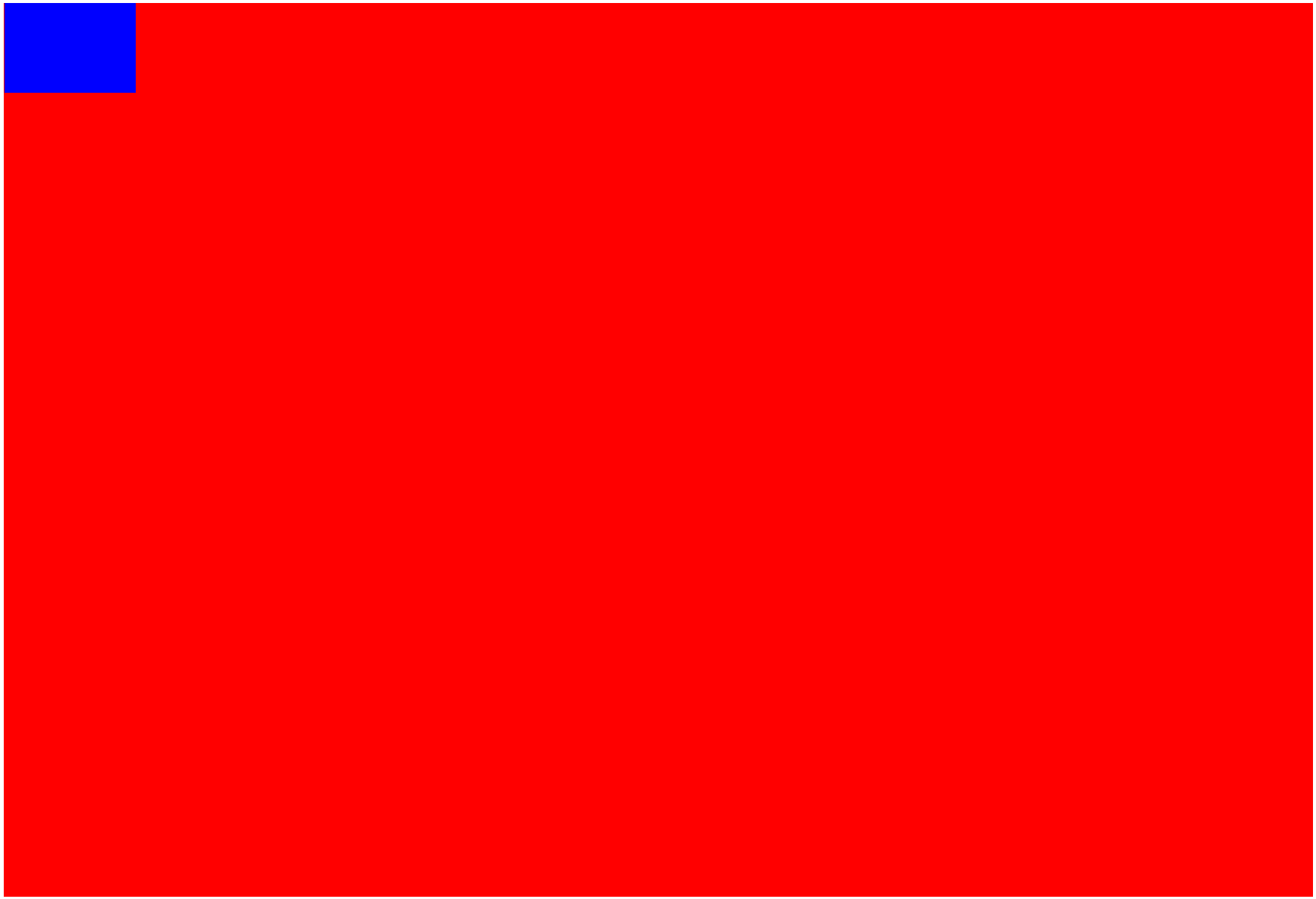
That was for this sort of relationship.



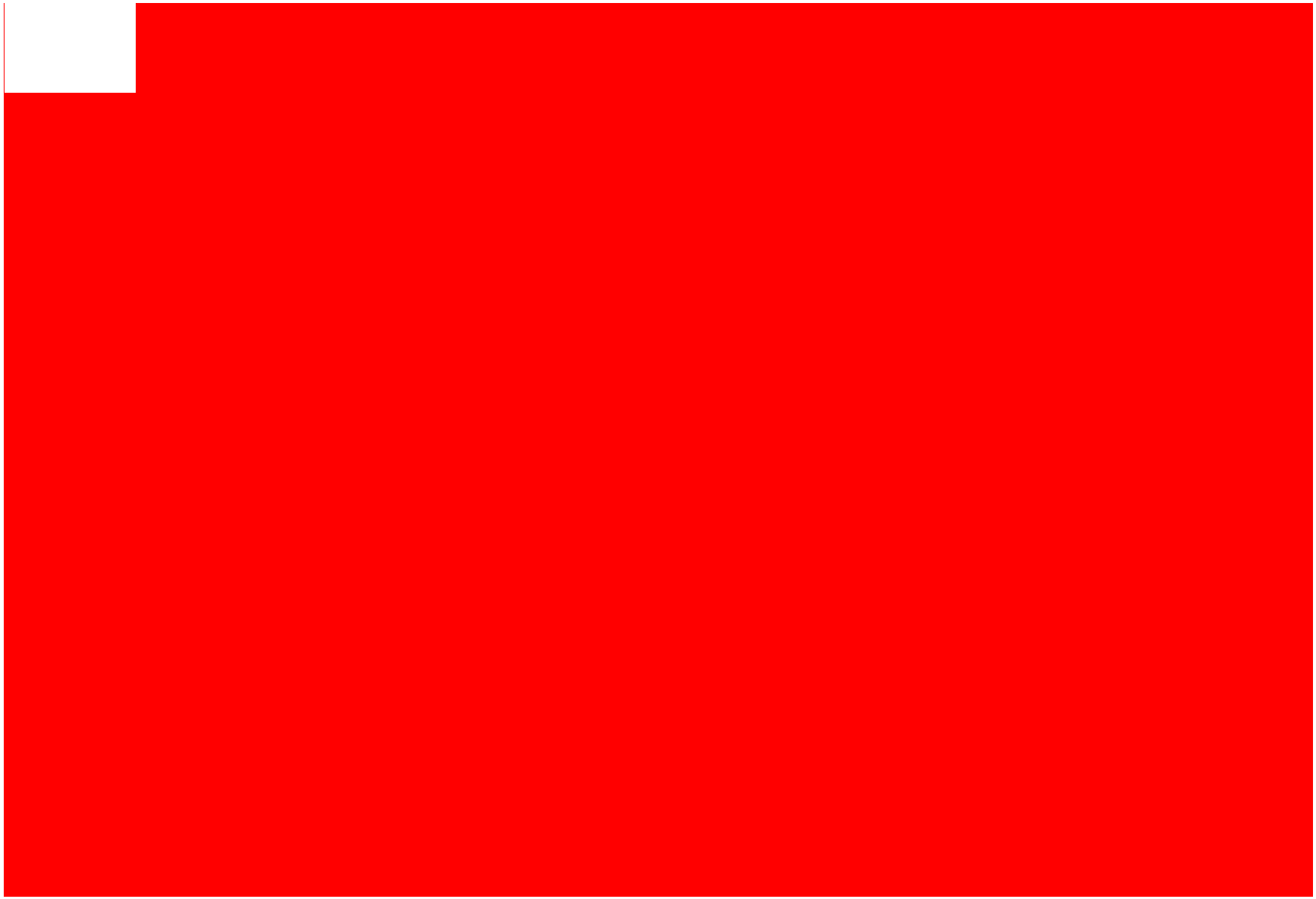

What if we have more red than blue?
Ozone
150
100
50
0
5
10
15
20
Wind

Add the blues together. (This is at a different scale.)

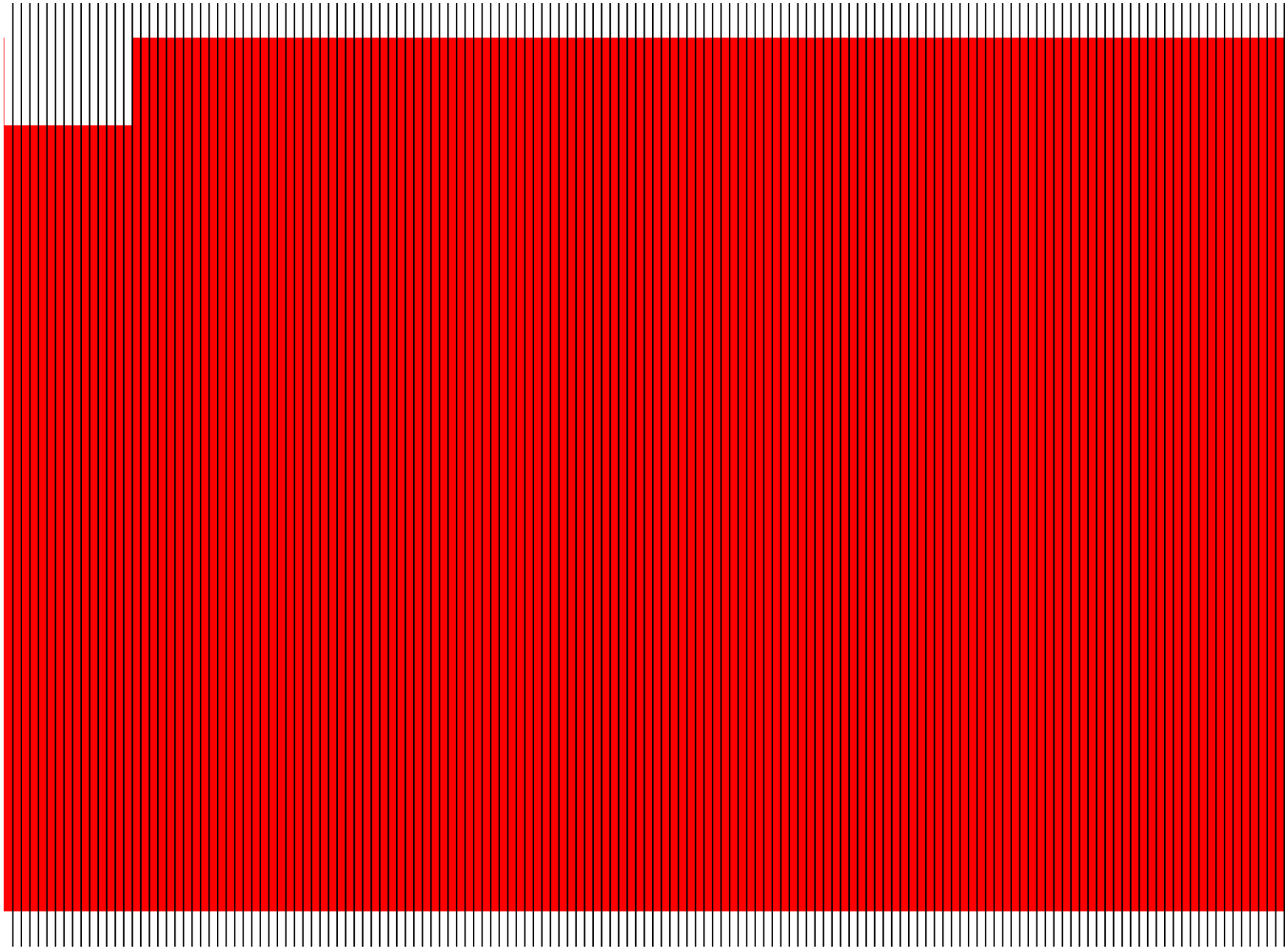
Add the reds together.



Subtract the reds.



Divide into as many equal pieces as we have irises (n).



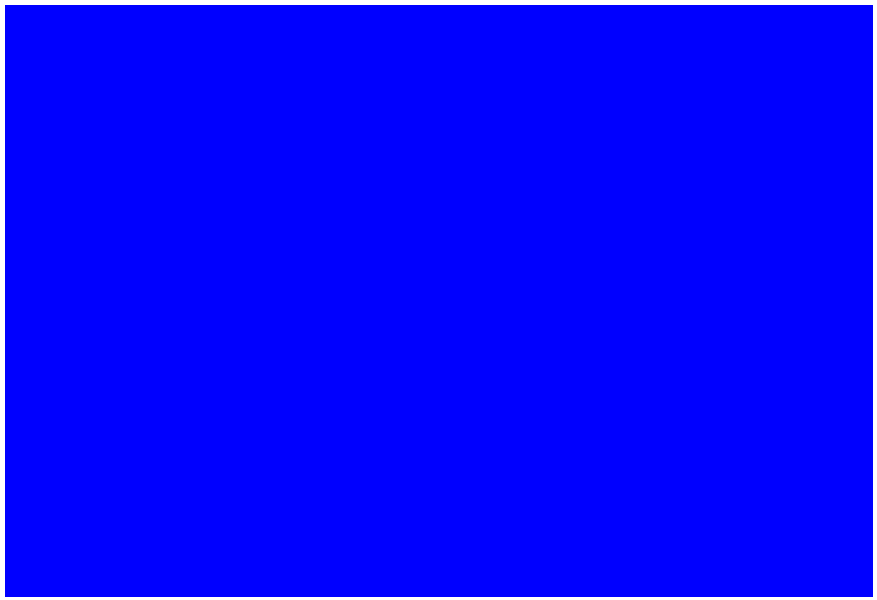
This red sliver is the covariance.

This red sliver is the covariance.

But it's negative!


What if we have as much red as blue?
rand$y
2
1
0
−1
−2
rand$x
−2
−1
0
1
2
3

Add the blues together. (This is at a different scale.)



Add the reds together.

Subtract the reds.

(Covariance is zero.)

Variance

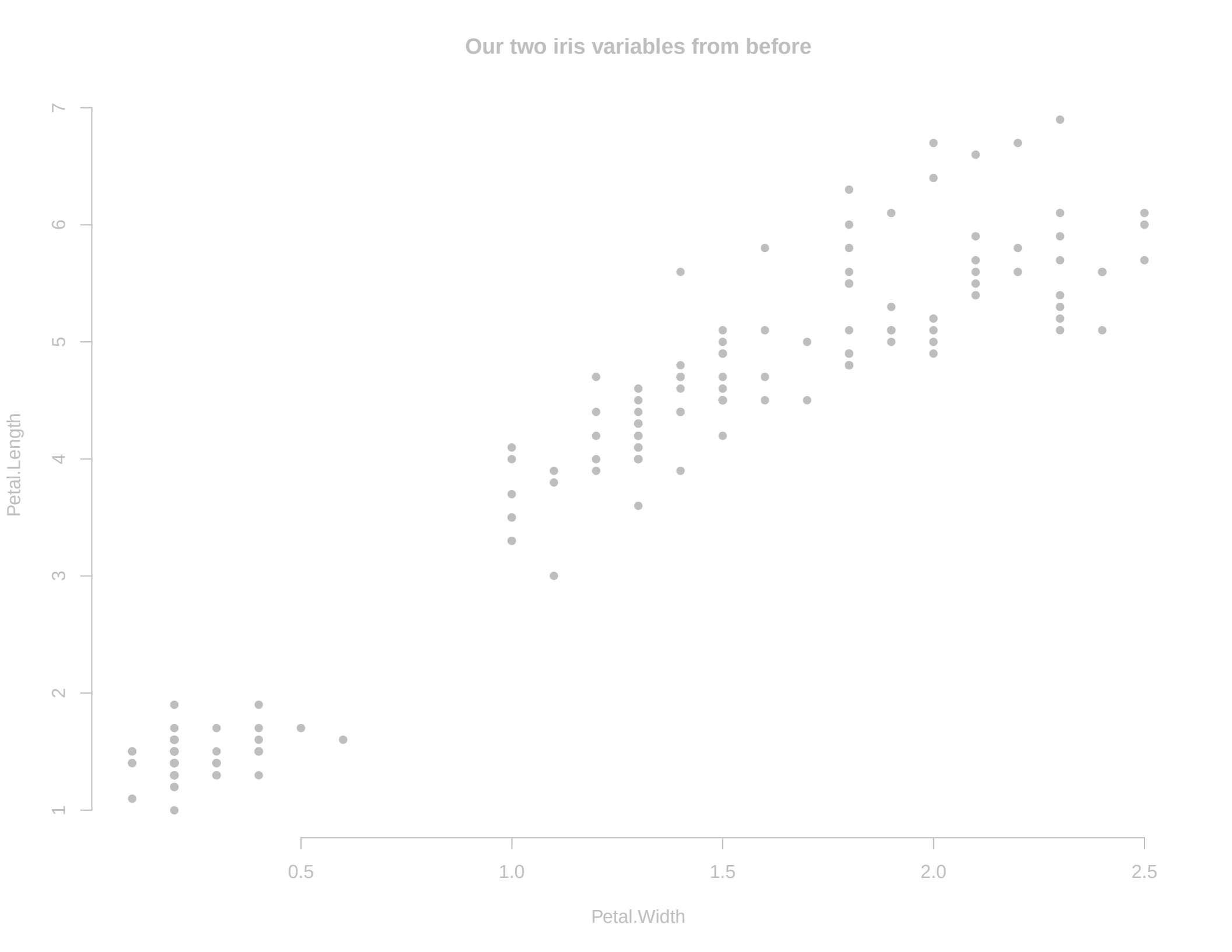
Variance tells us how spread out some numbers are.

1, 4, 8, 10

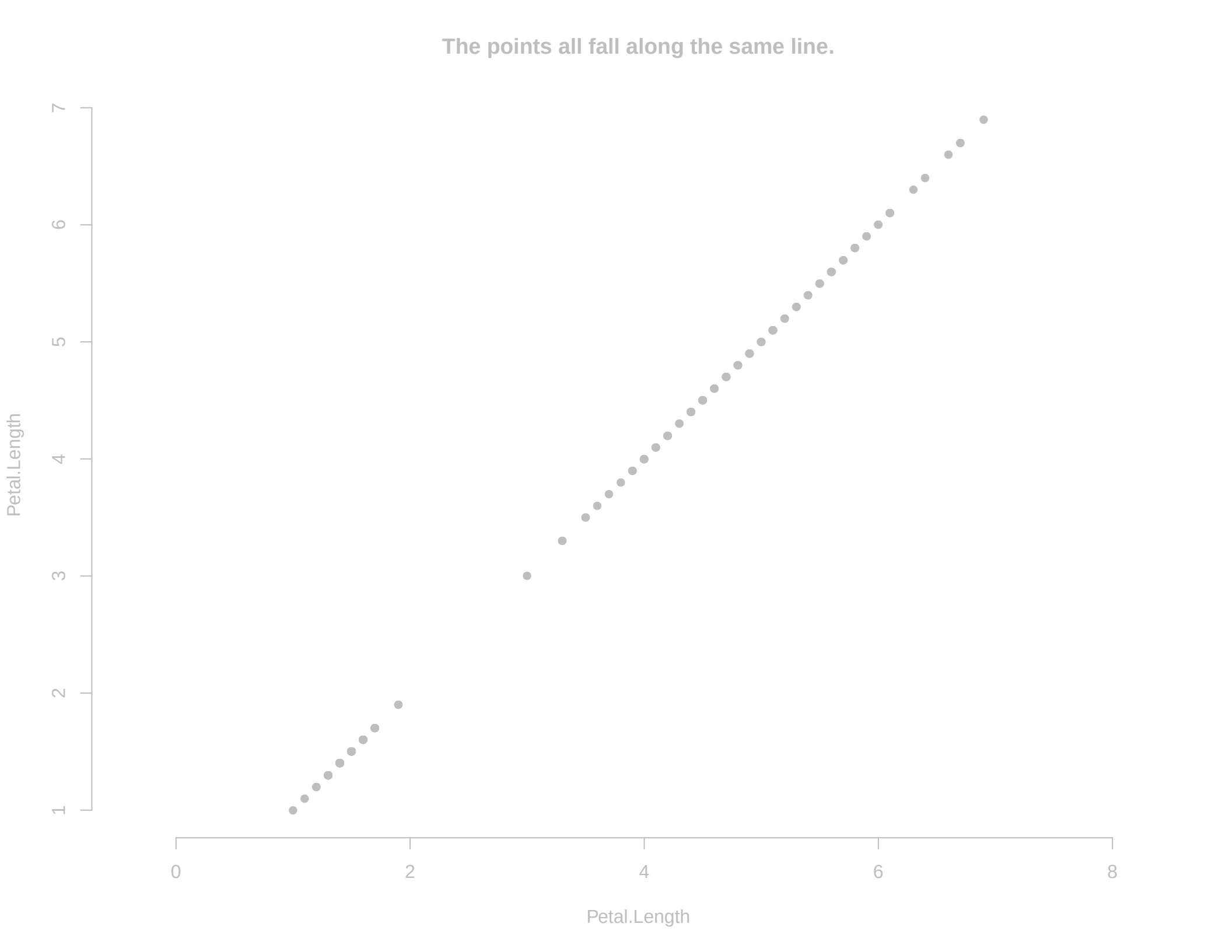
vs

4, 4, 5, 6

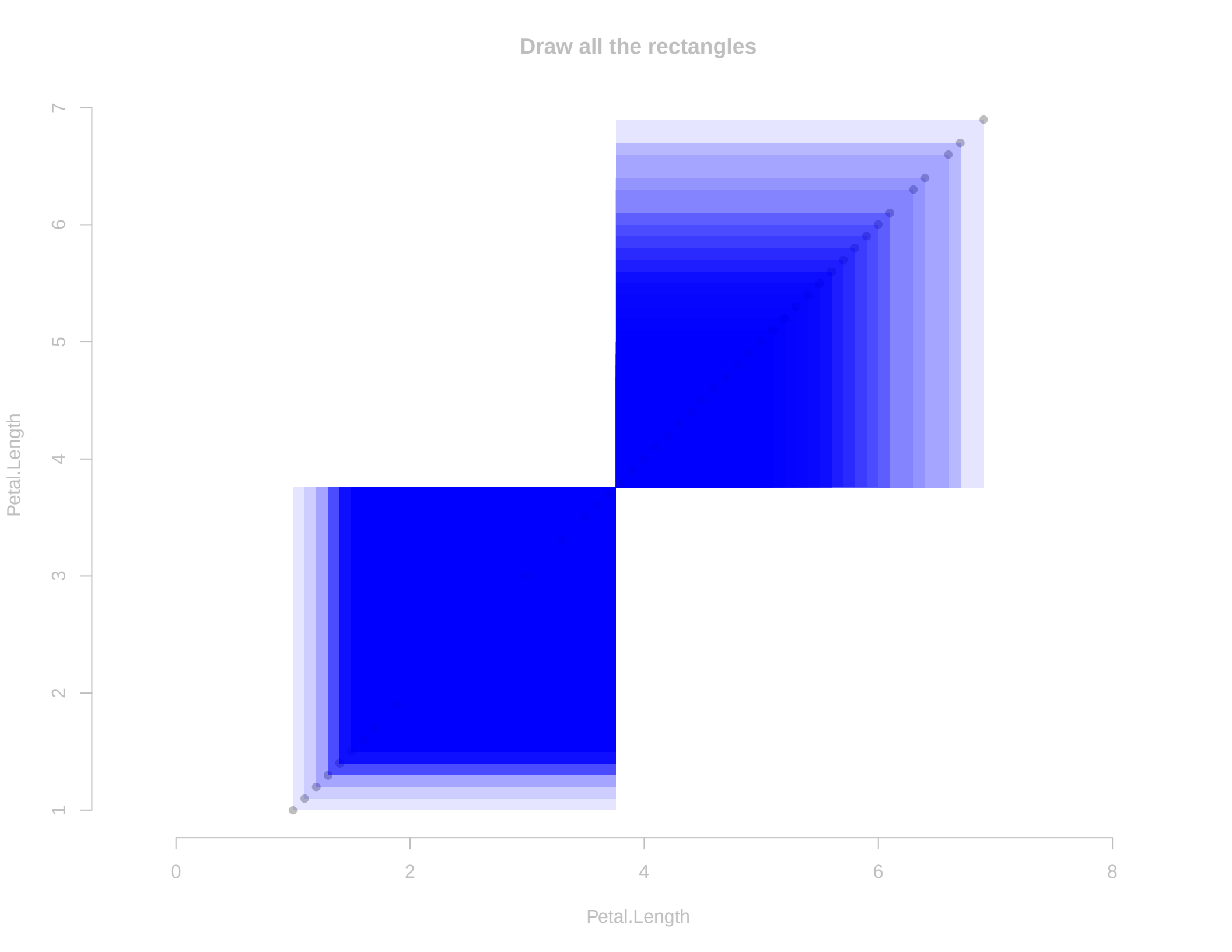
The variance of a variable is the covariance of the variable with itself.

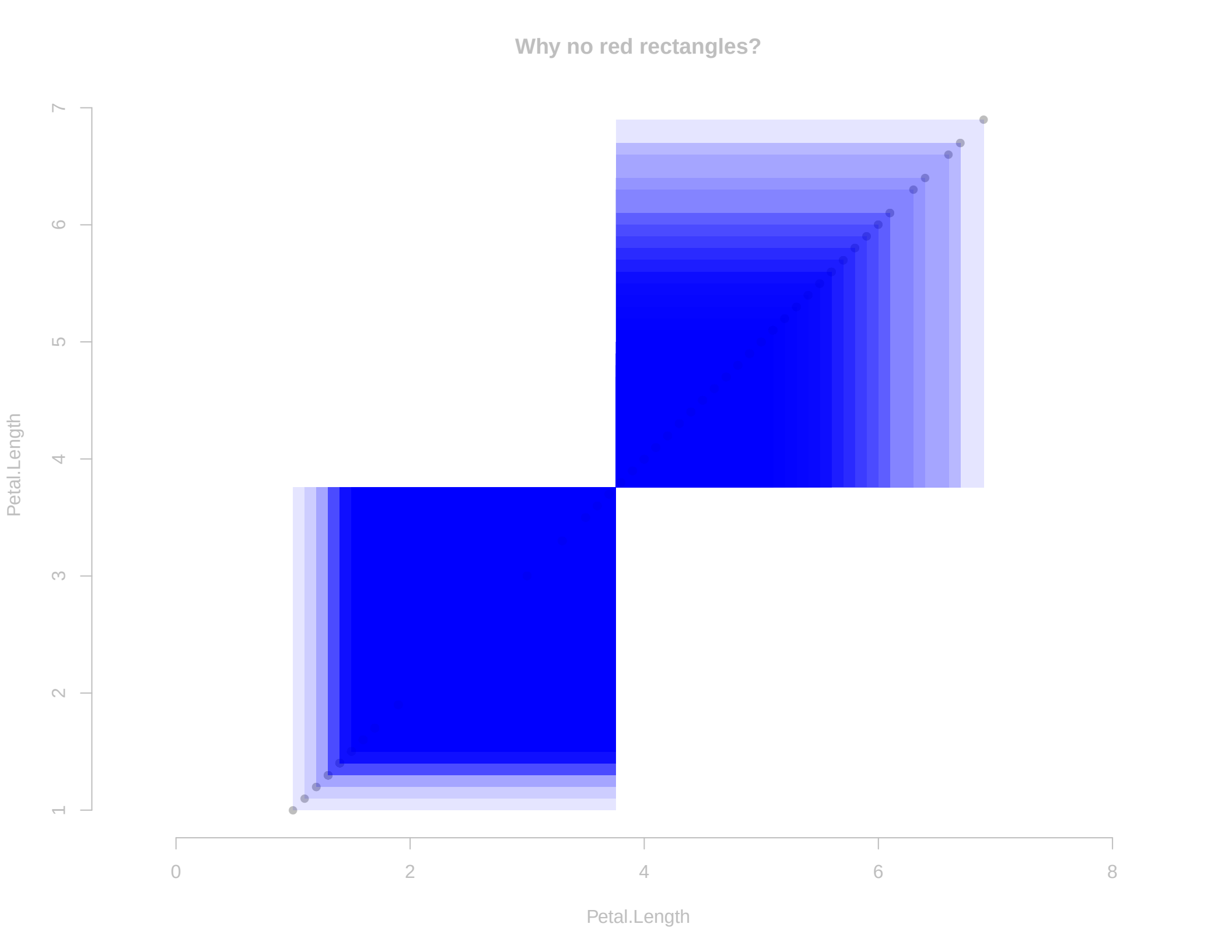

Our two iris variables from before
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

Let's look at just one of them.

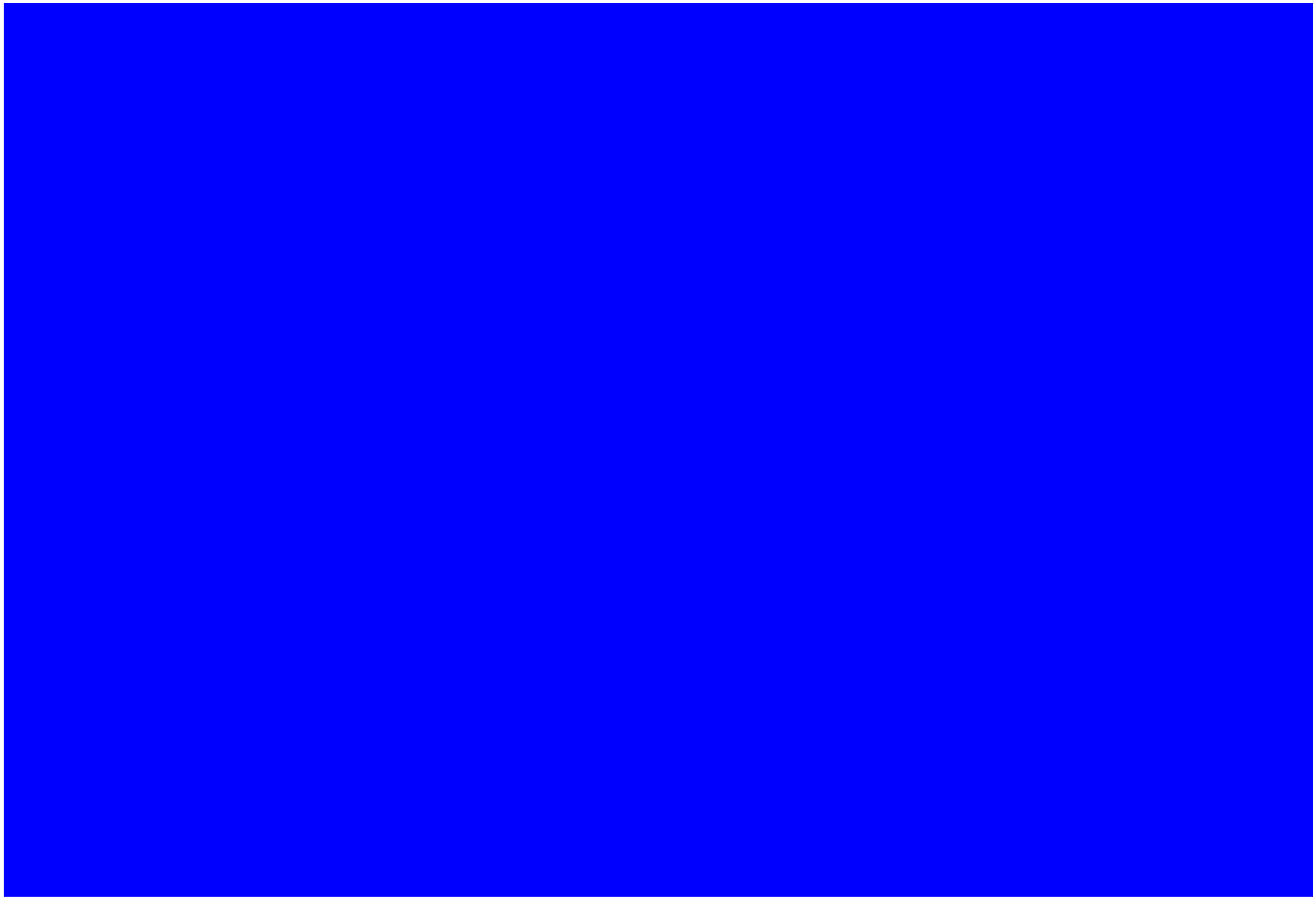

The points all fall along the same line.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

Let's find the variance of Petal.Length

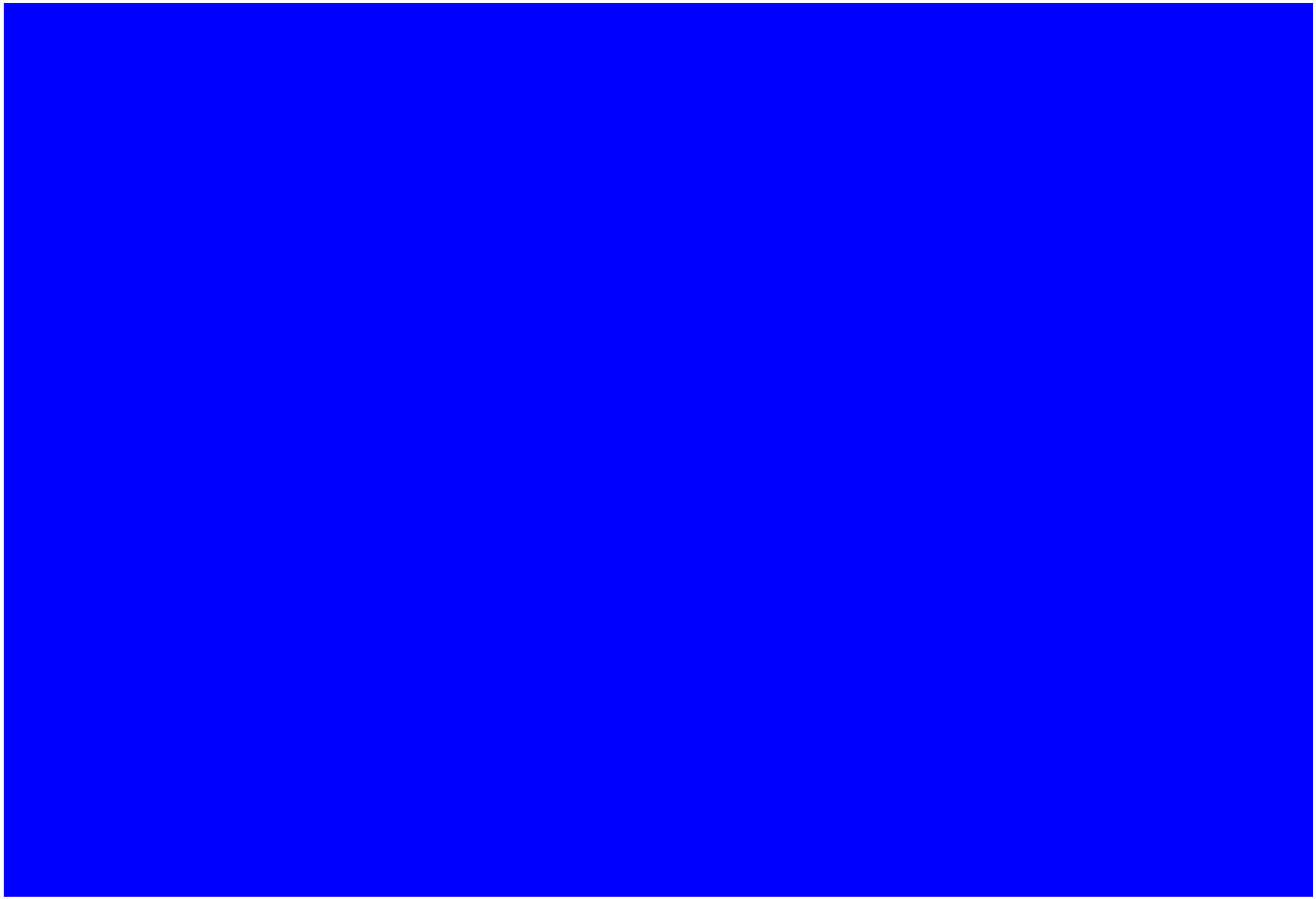

Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length


Why no red rectangles?
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

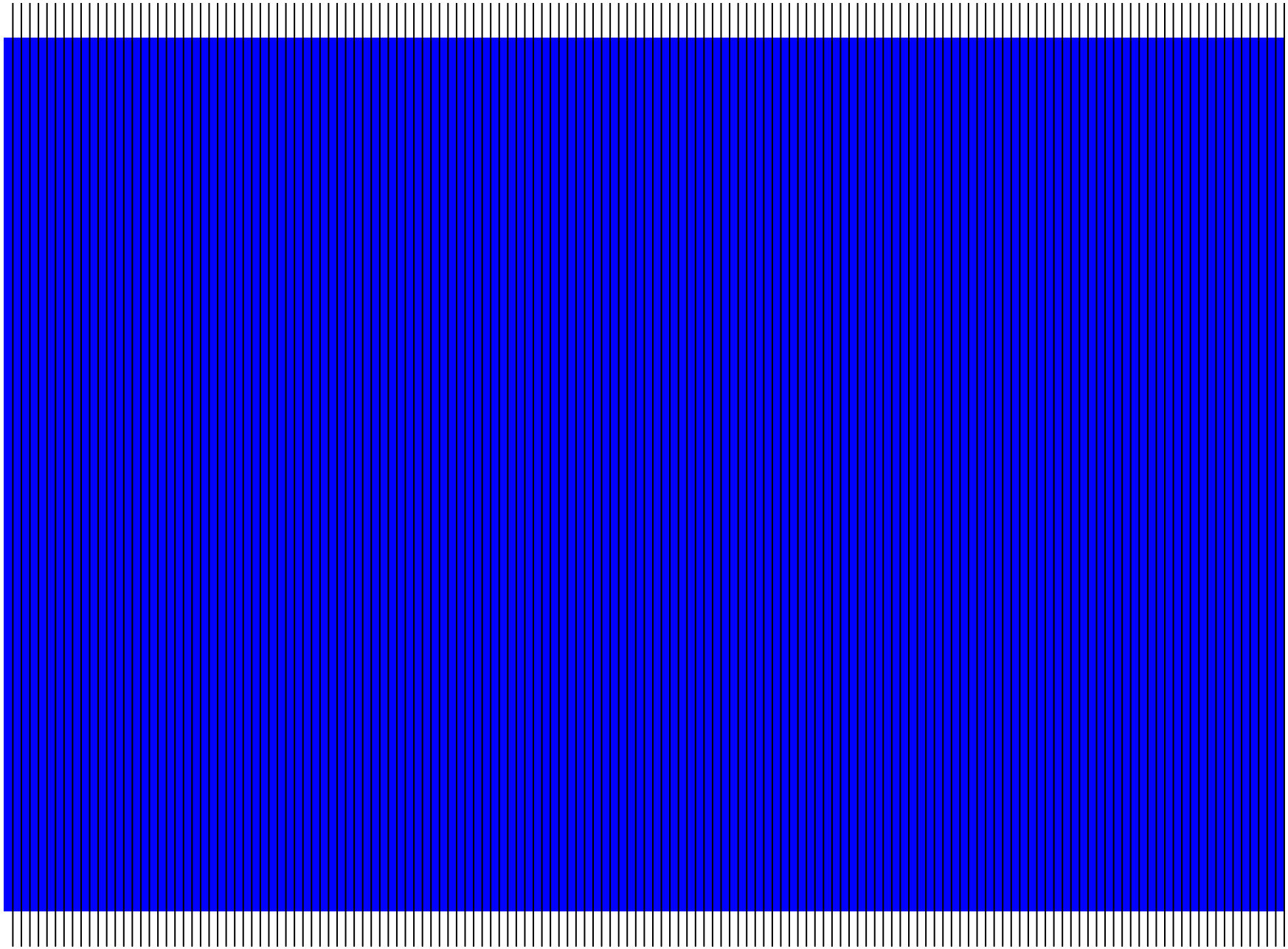
Add the blues together. (This is at a different scale.)



We have no reds to subtract.



Divide into as many equal pieces as we have irises (n).



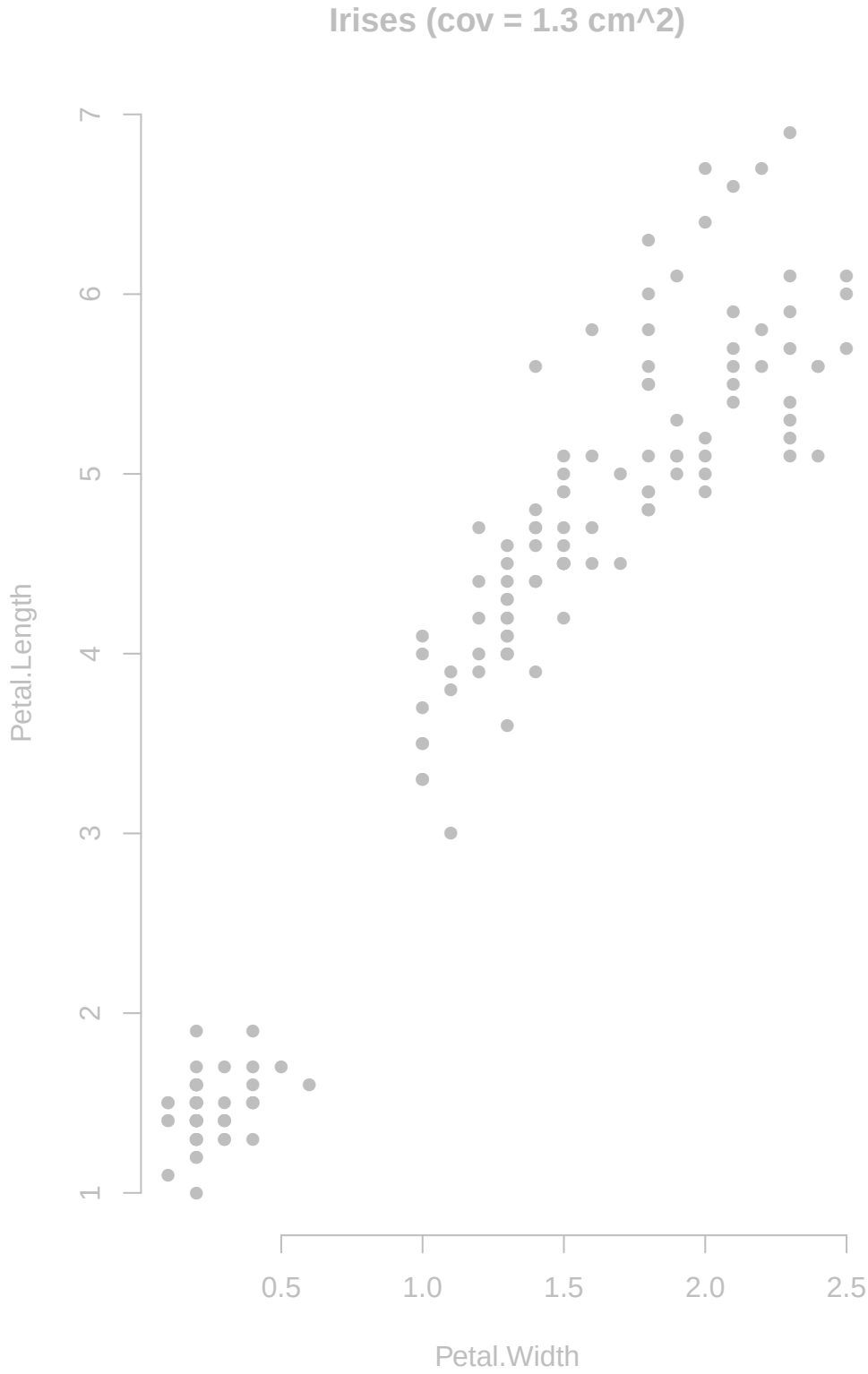
This blue sliver is the variance.

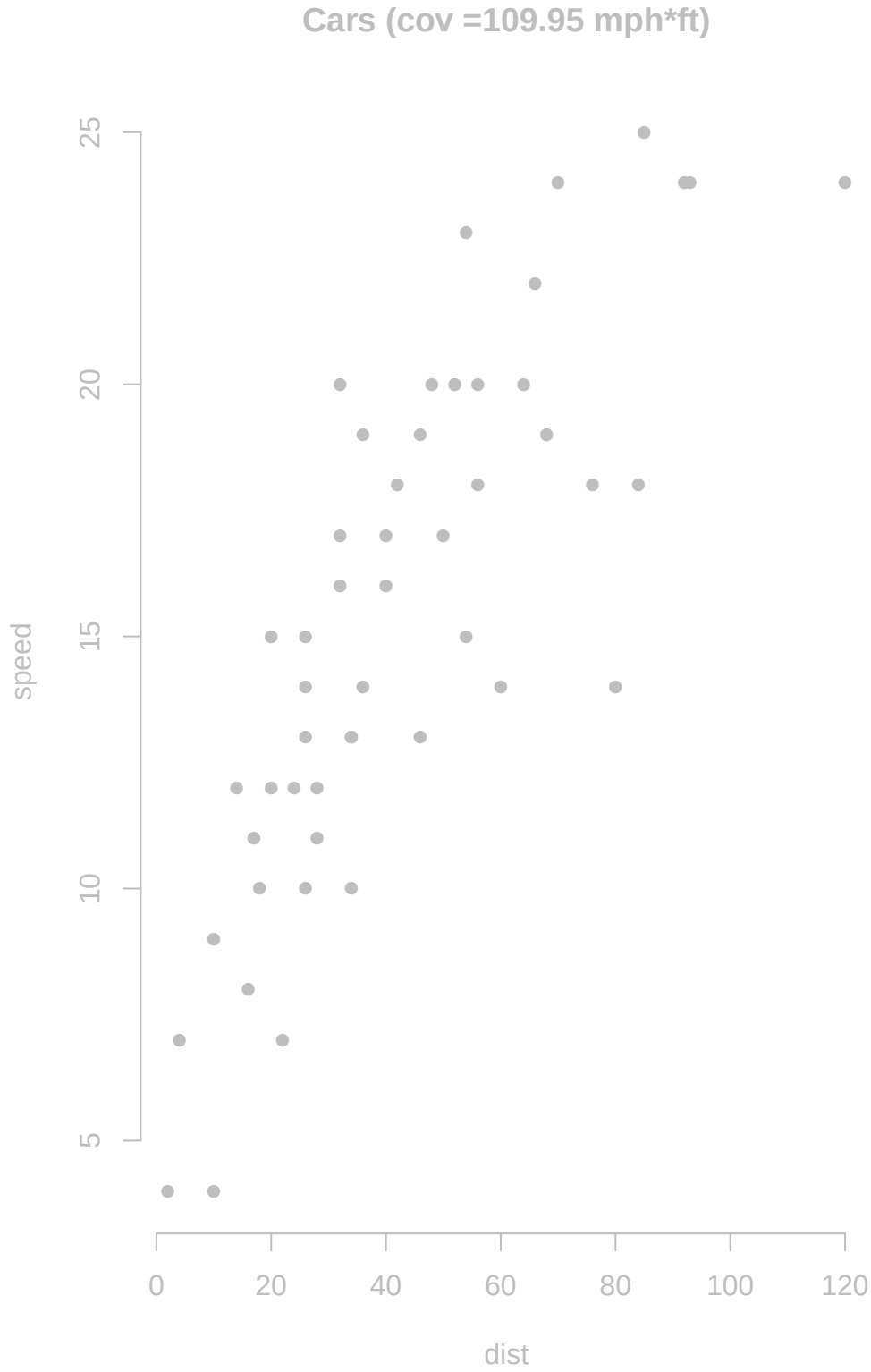
A problem with covariance

Covariance has units!

(x−unit times y−unit)

Which relationship is stronger (more linear)?

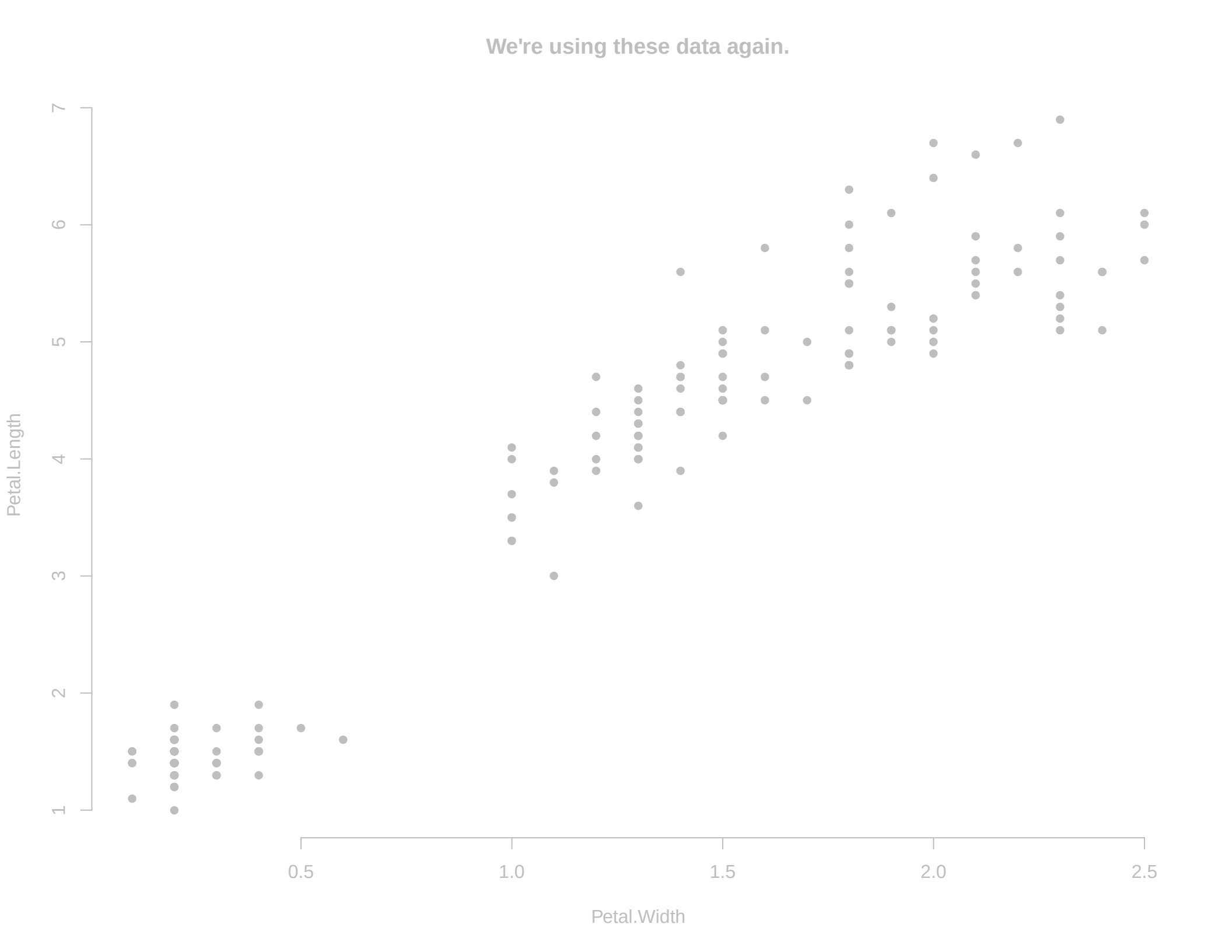

Irises (cov = 1.3 cm^2)
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

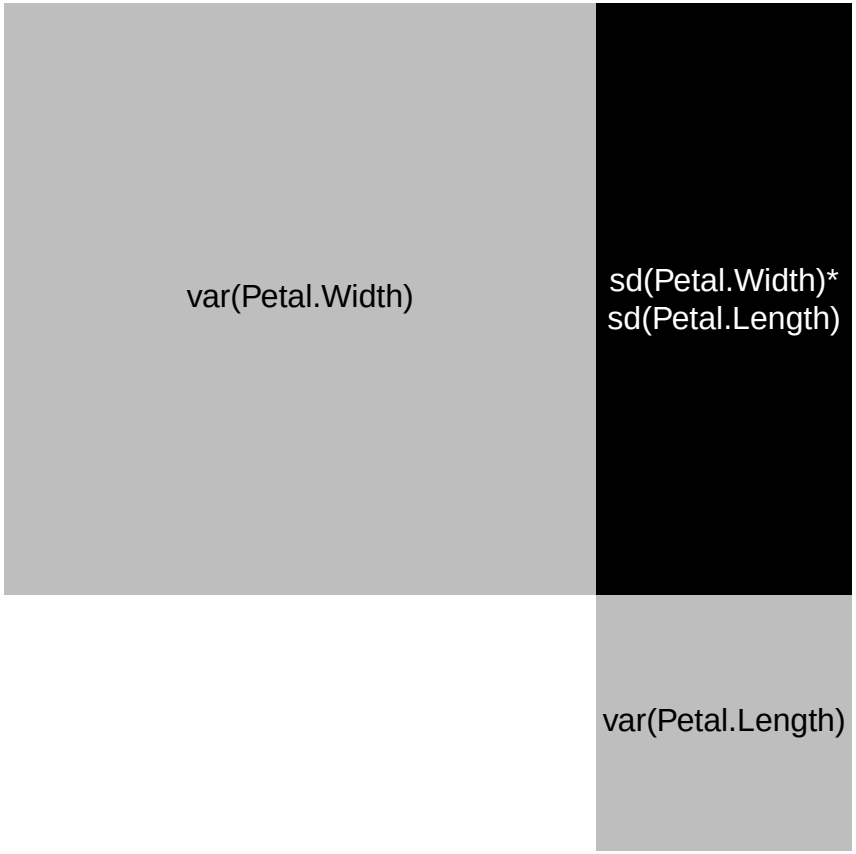

Cars (cov =109.95 mph*ft)
speed
dist
5
10
15
20
25
0
20
40
60
80
100
120

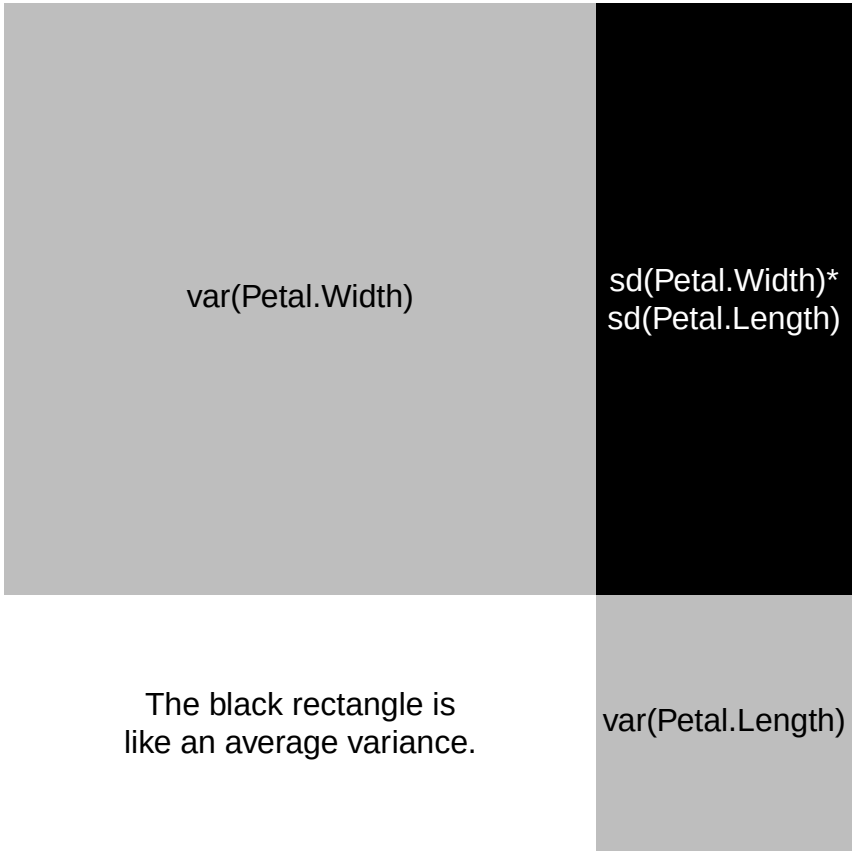
Oh noes!

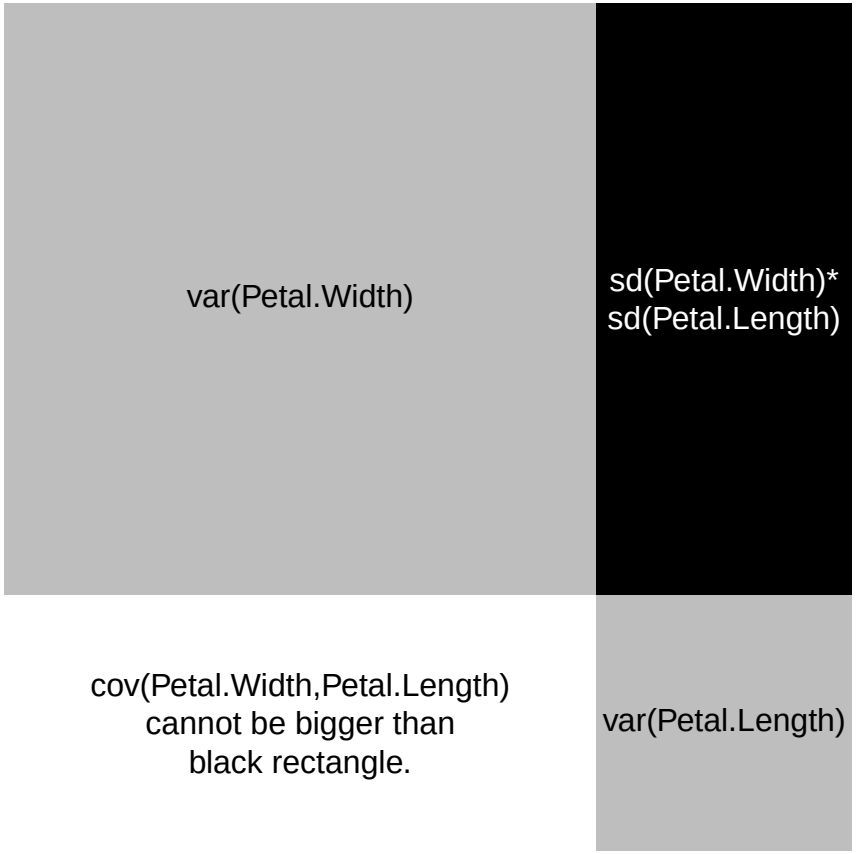
We can divide the covariance by the variances to standardize it.

We're using these data again.

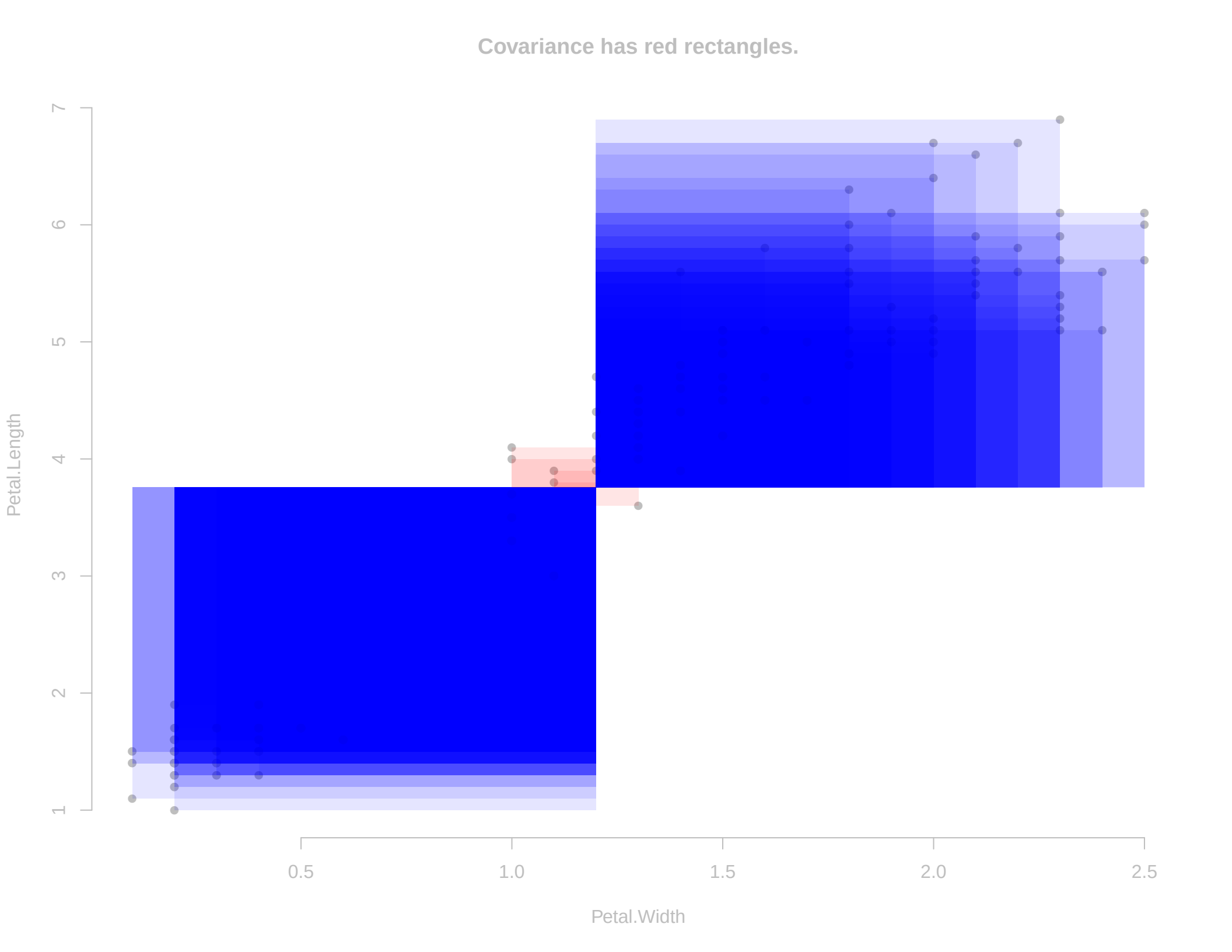


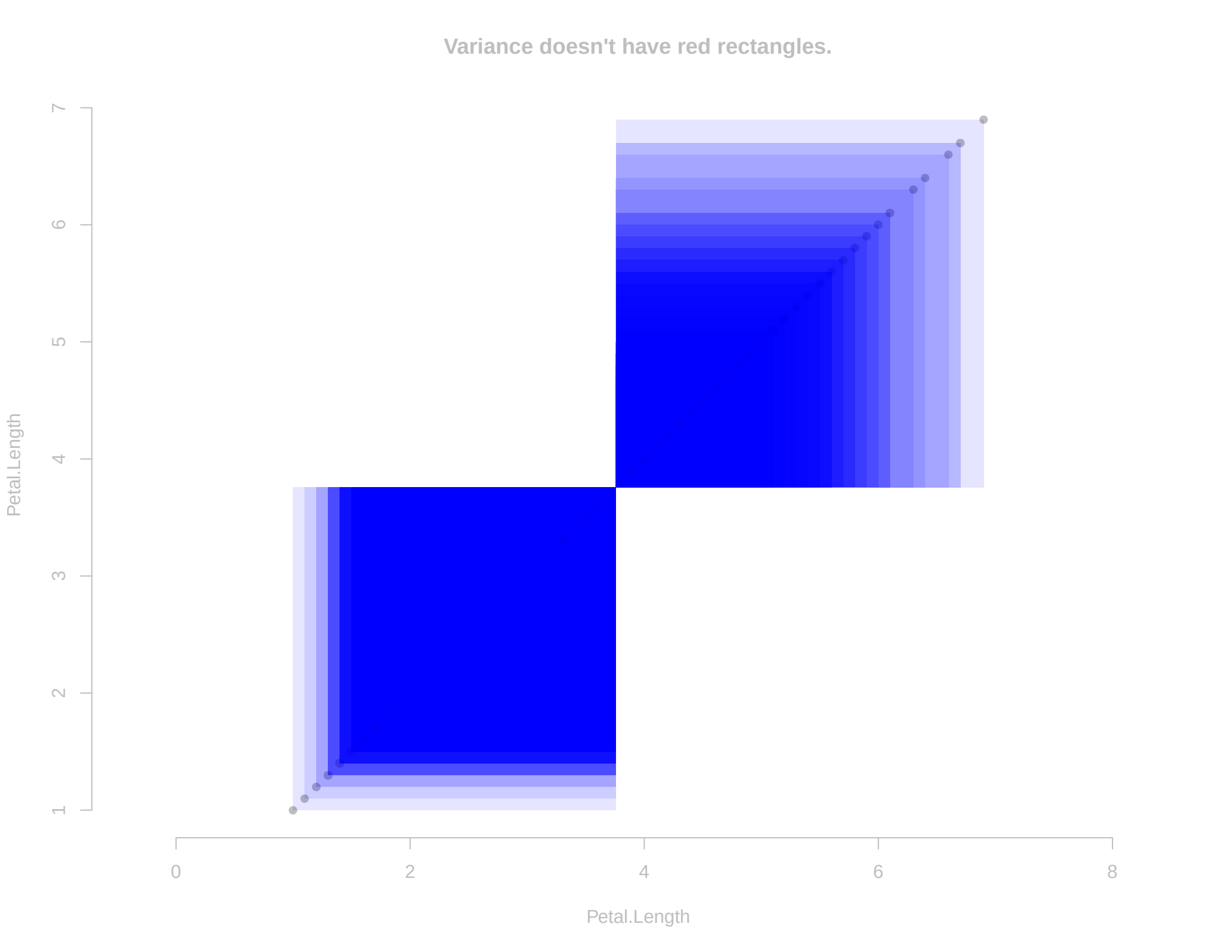

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
var(Petal.Length)

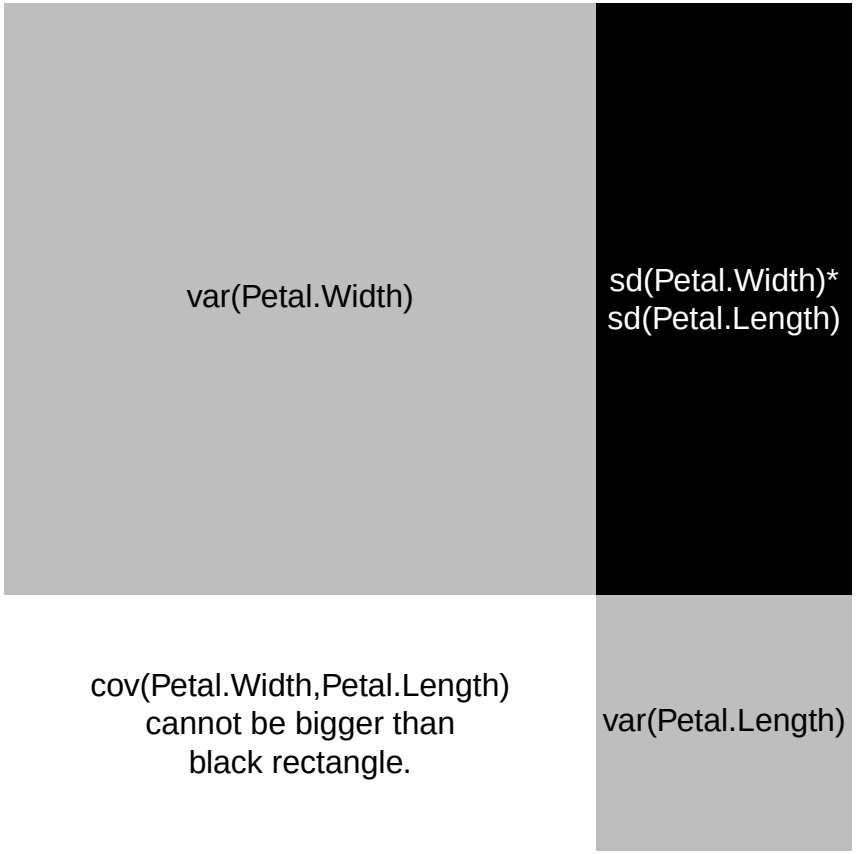

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
The black rectangle is
like an average variance.
var(Petal.Length)


var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
cov(Petal.Width,Petal.Length)
cannot be bigger than
black rectangle.
var(Petal.Length)

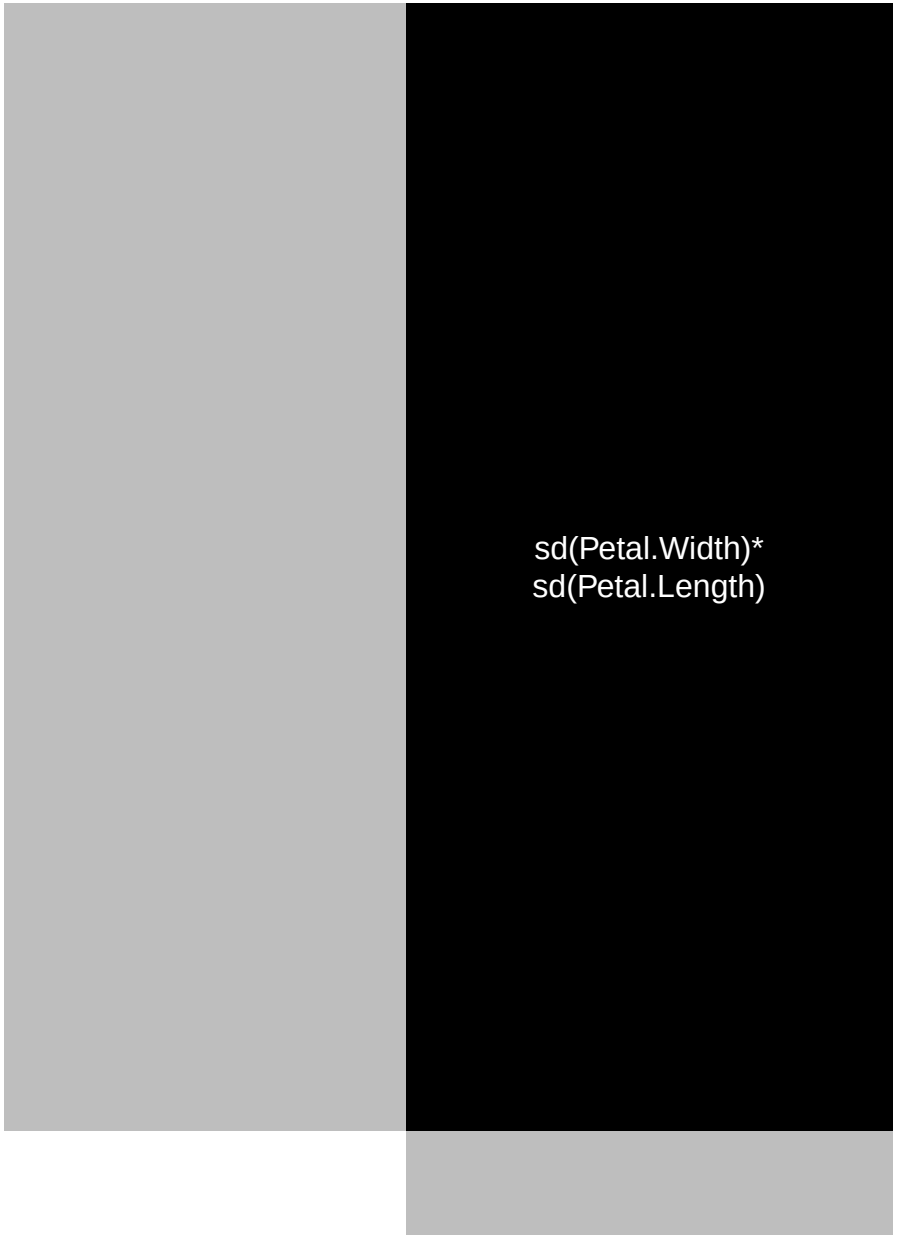
Why?


Covariance has red rectangles.
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

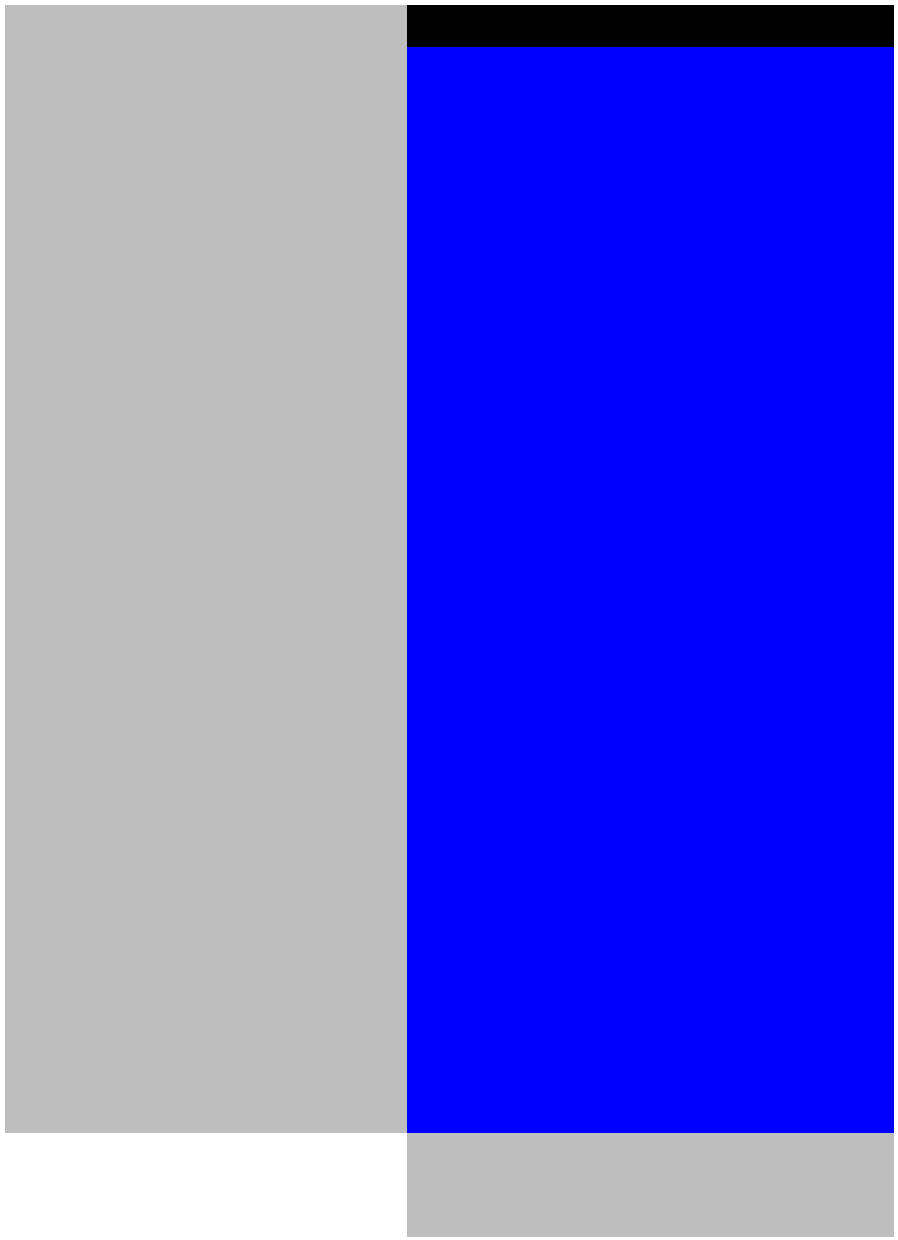

Variance doesn't have red rectangles.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length


var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
cov(Petal.Width,Petal.Length)
cannot be bigger than
black rectangle.
var(Petal.Length)

Let's zoom in.

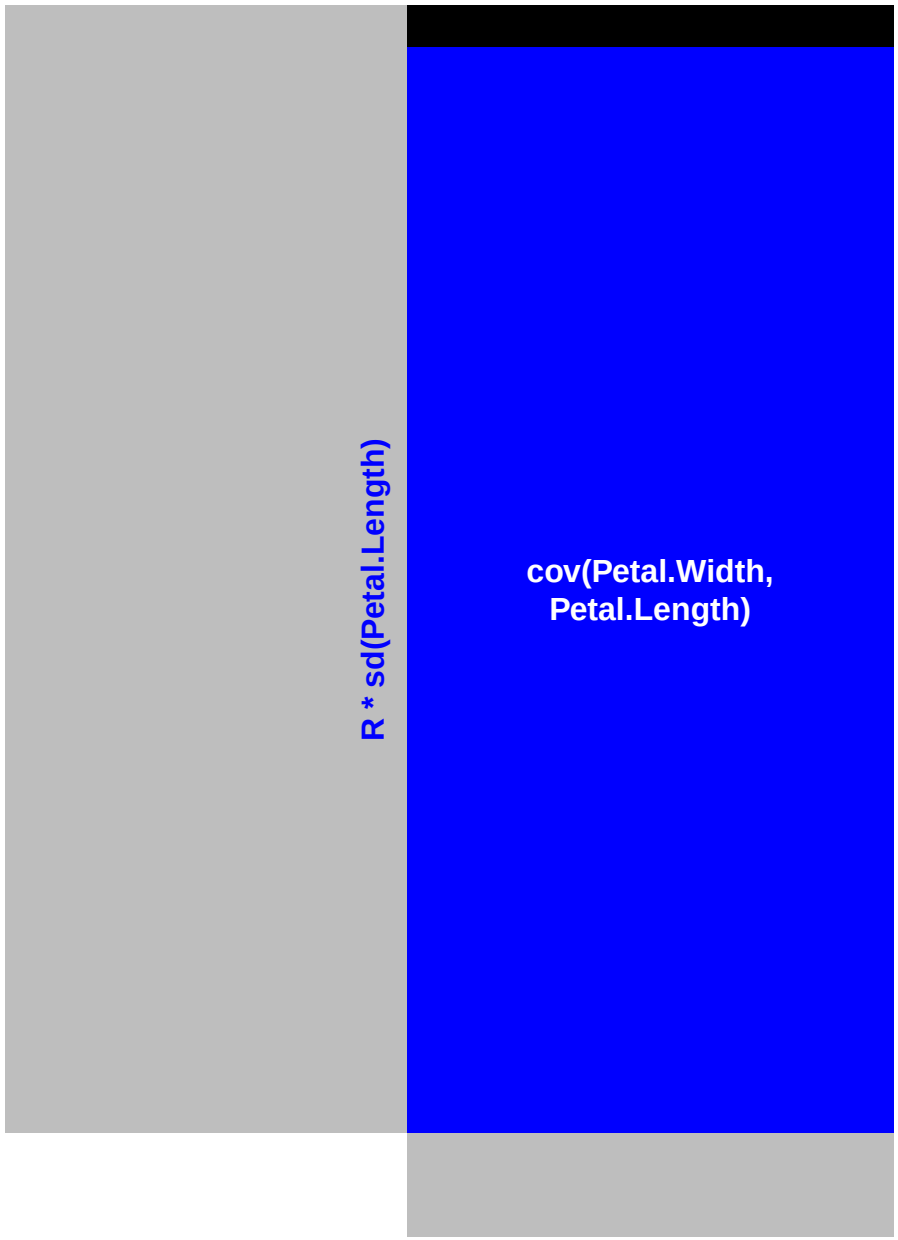


Squish covariance vertically into the rectangle.

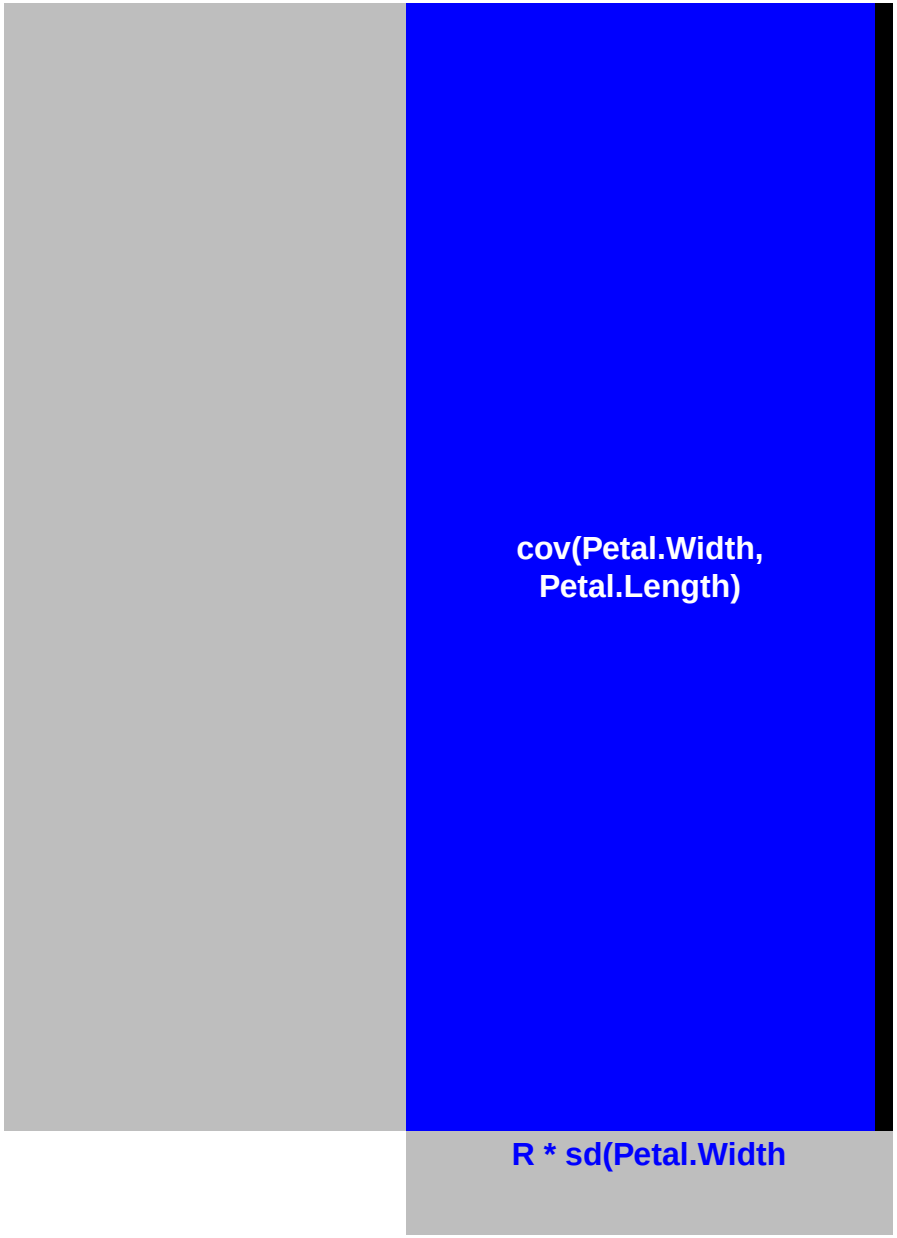


Correlation (R)
is the ratio of
the small rectangle
to the big rectangle.

Squish covariance vertically into the rectangle.

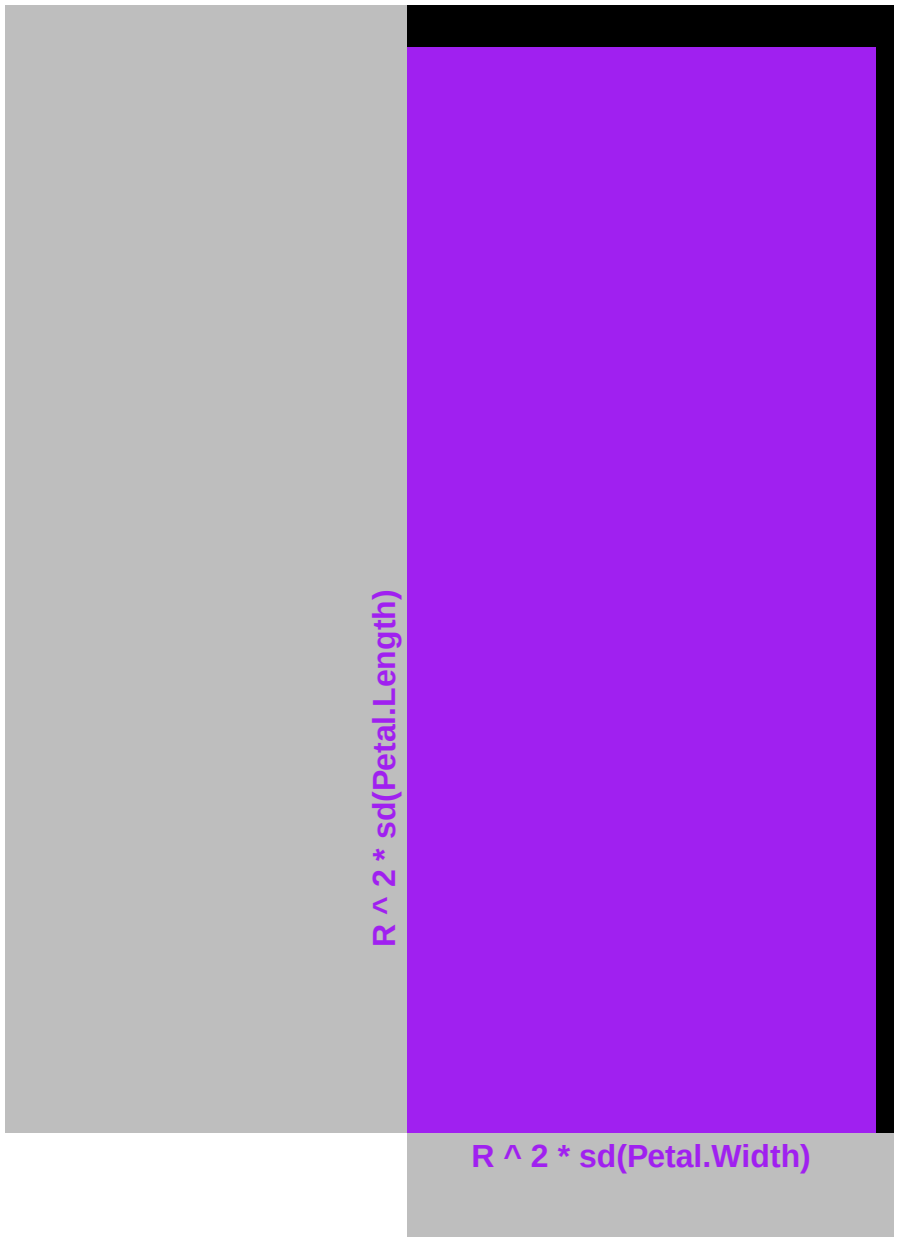


Squish covariance horizontally into the rectangle.

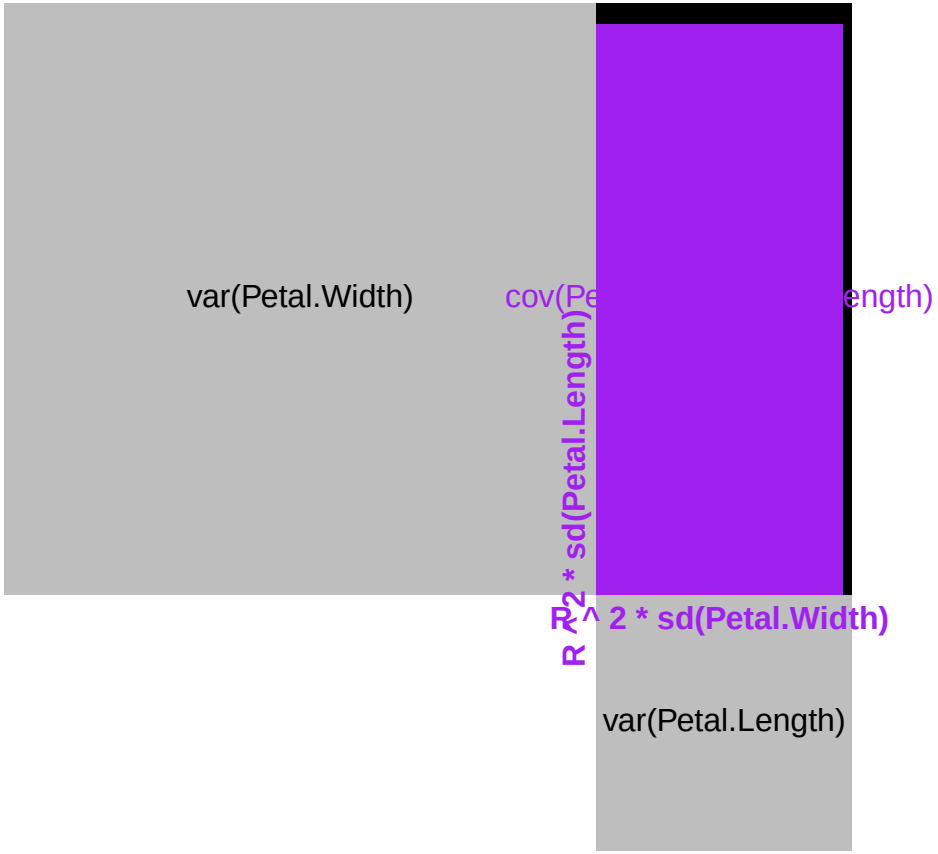


People like to
talk about R−squared.

Intersect the two squished covariance rectangles.



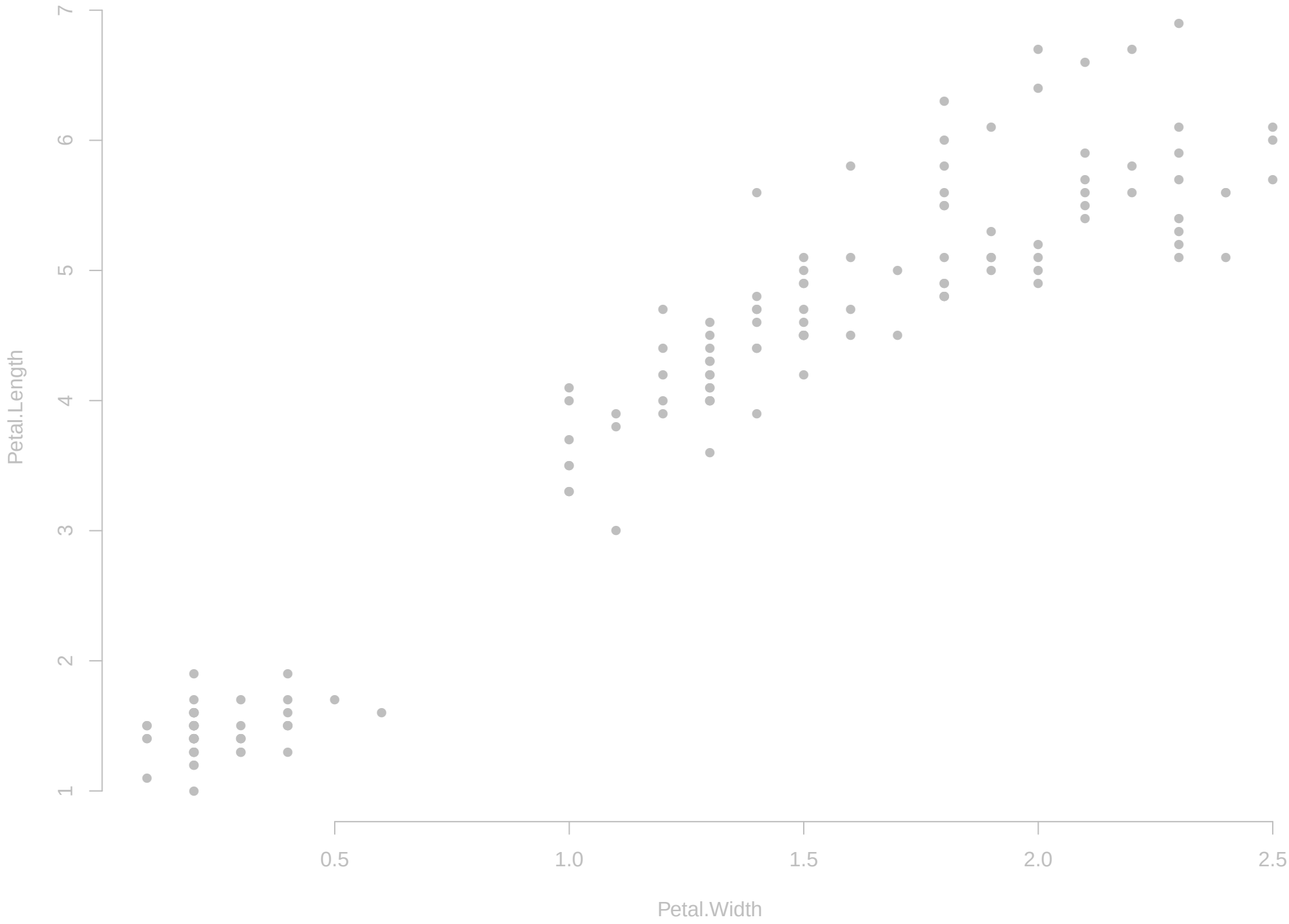
Intersect the two squished covariance rectangles.

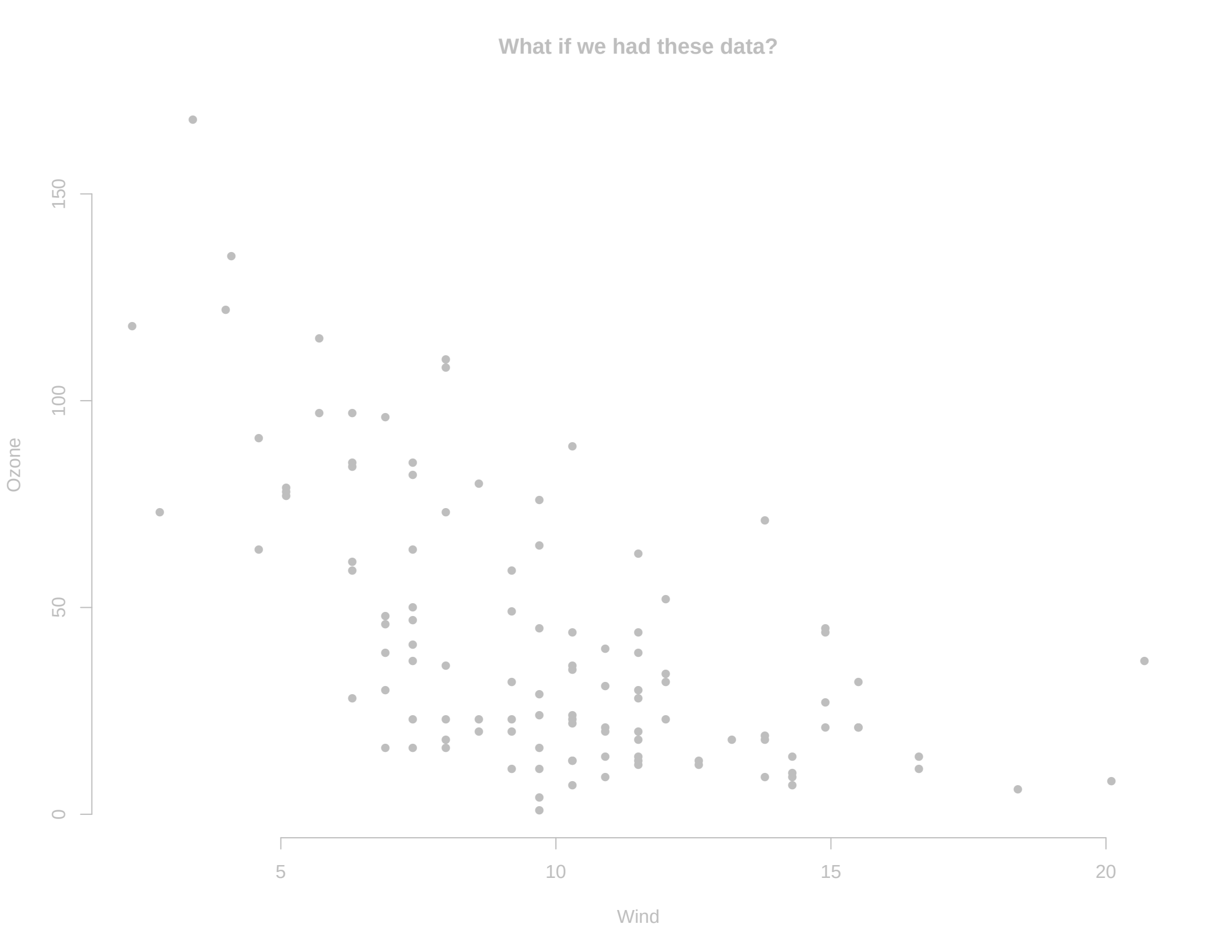


That was for very positive (blue) covariances.

What if covariance is negative (red)?

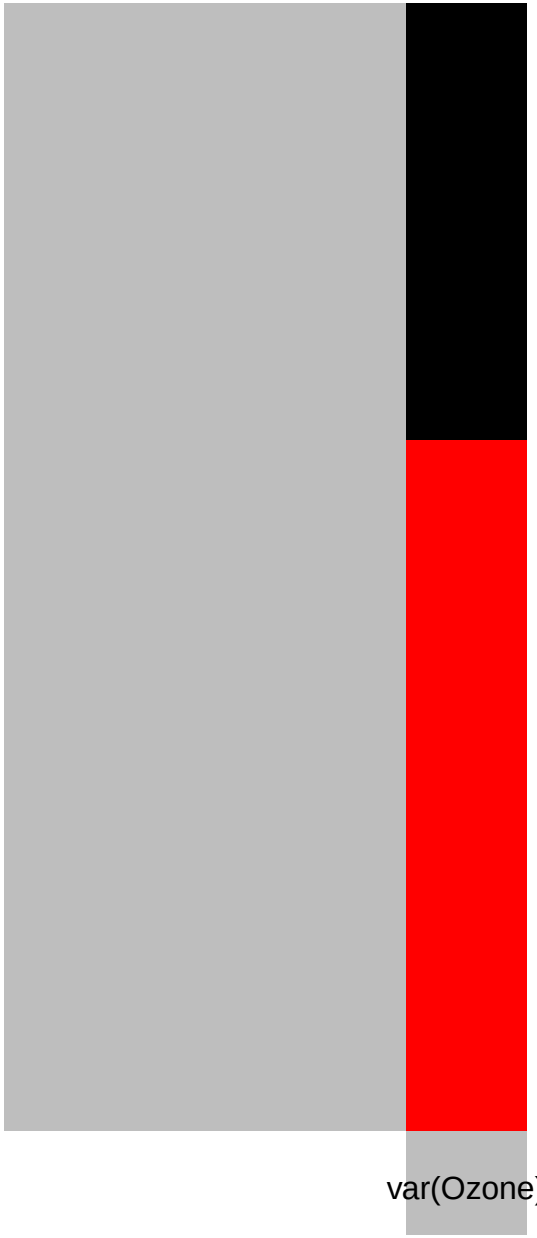
We were just using these data.



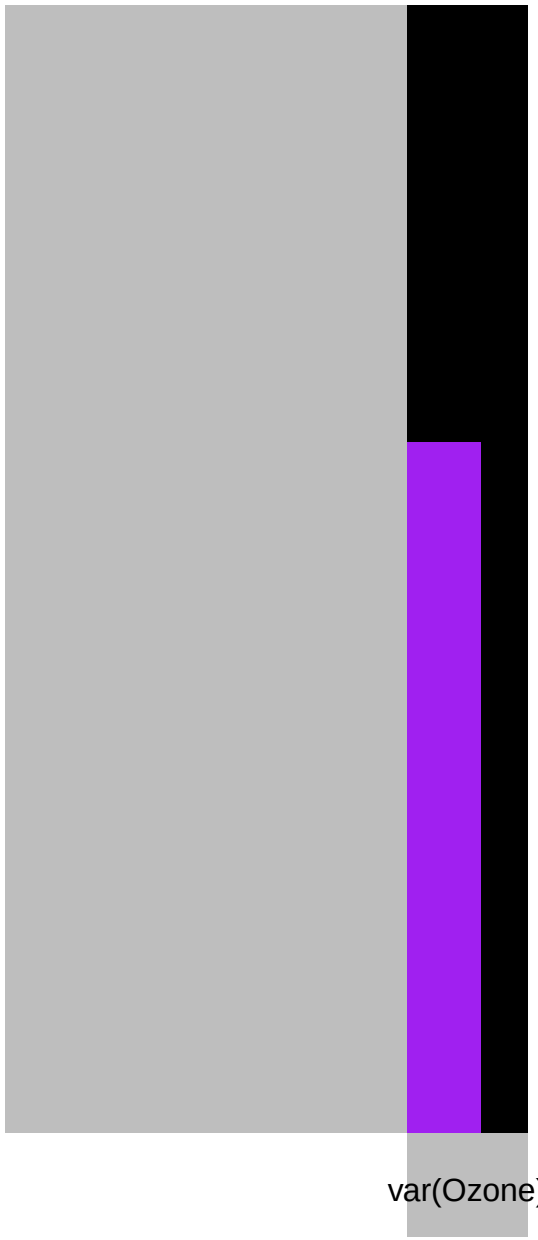

What if we had these data?
Ozone
0
50
100
150
5
10
15
20
Wind


What if we had these data?
Ozone
0
50
100
150
5
10
15
20
Wind

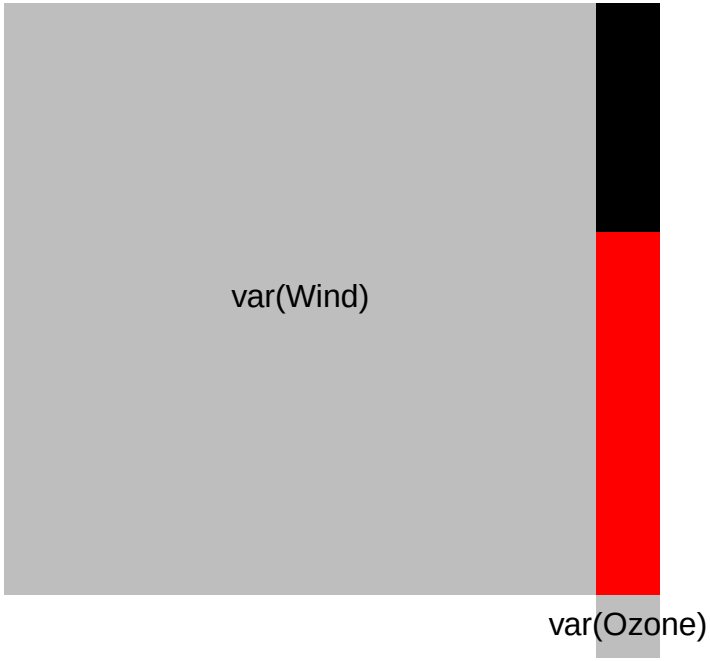
R is the same, just negative.



R−squared is the same, and it is always positive.



Zoom back out.



If we transform the covariance a bit, we can also make predictions.

Let's use x to predict y.

$$y = b0 + b1 * x$$

Let's invent b1.

What values should it have?

If covariance is very positive and x is high, y should be high.

(We want b1 to be positive.)

If covariance is very negative and x is high, y should be low.

(We want b1 to be negative.)

If covariance is low,
we have no idea what y is.

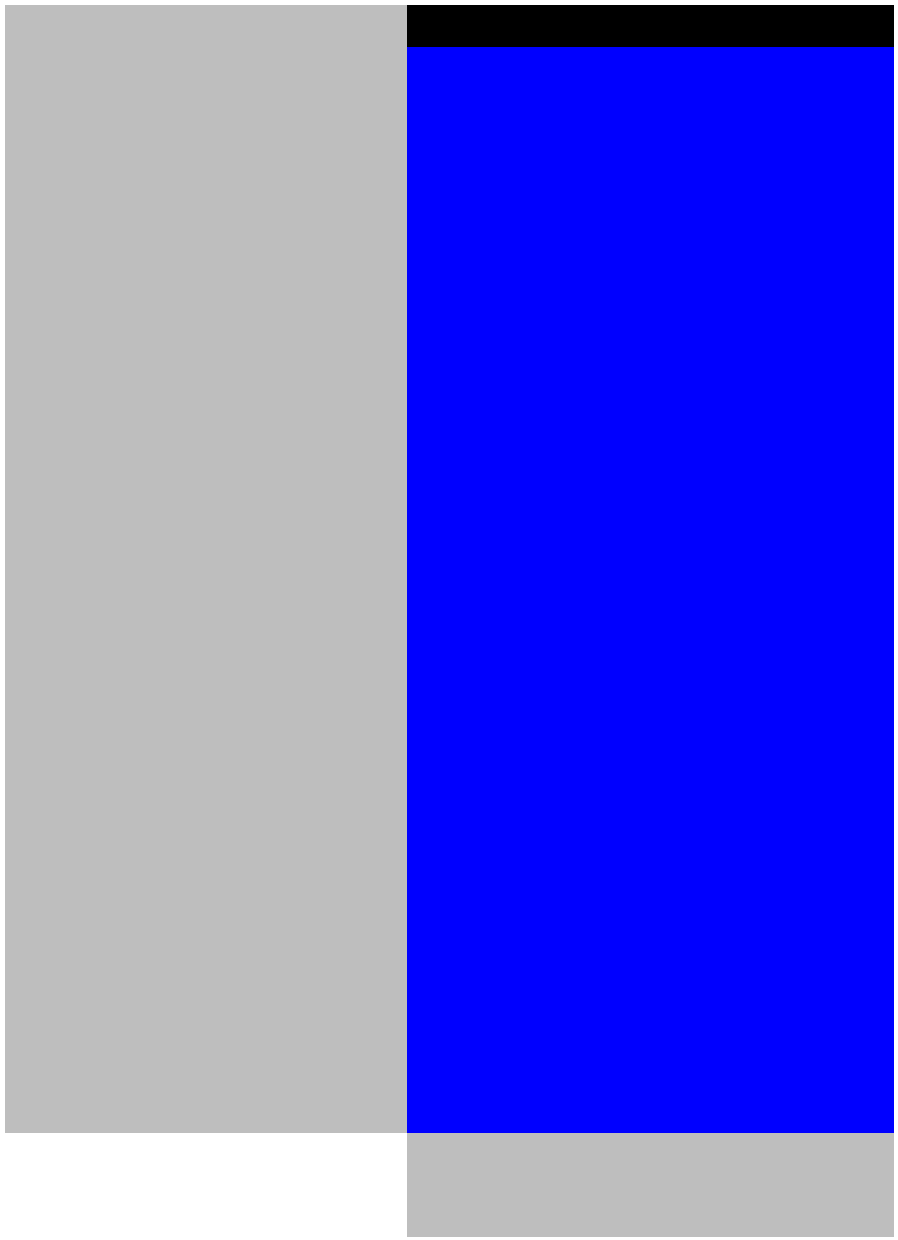
(b1 is around zero.)

Let's think about units again.

Covariance is an area; its unit is the product of the x and y units.

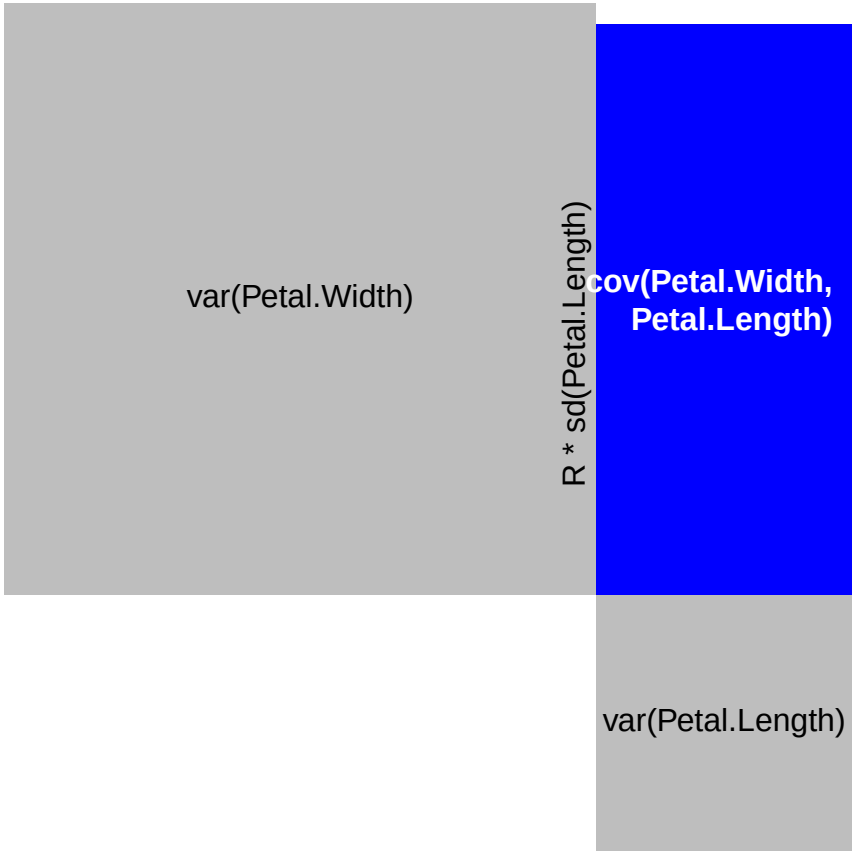
Variance is a special covariance; its unit is the square of the x unit.

Correlation is a ratio of areas with the same units.

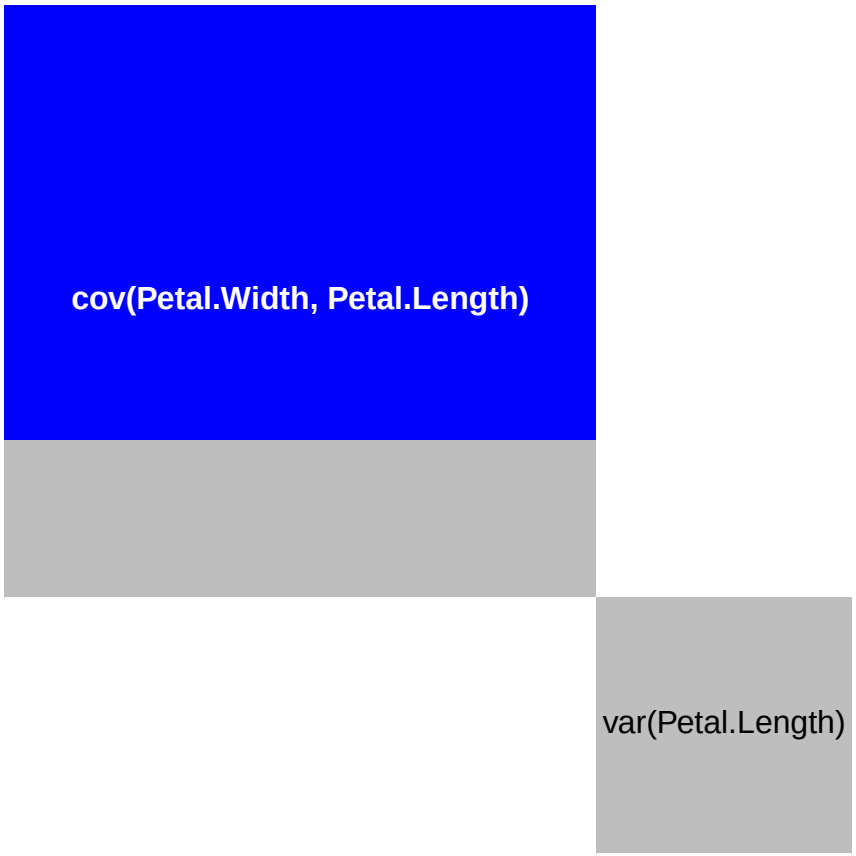


The unit of b1 must be y−unit/x−unit.

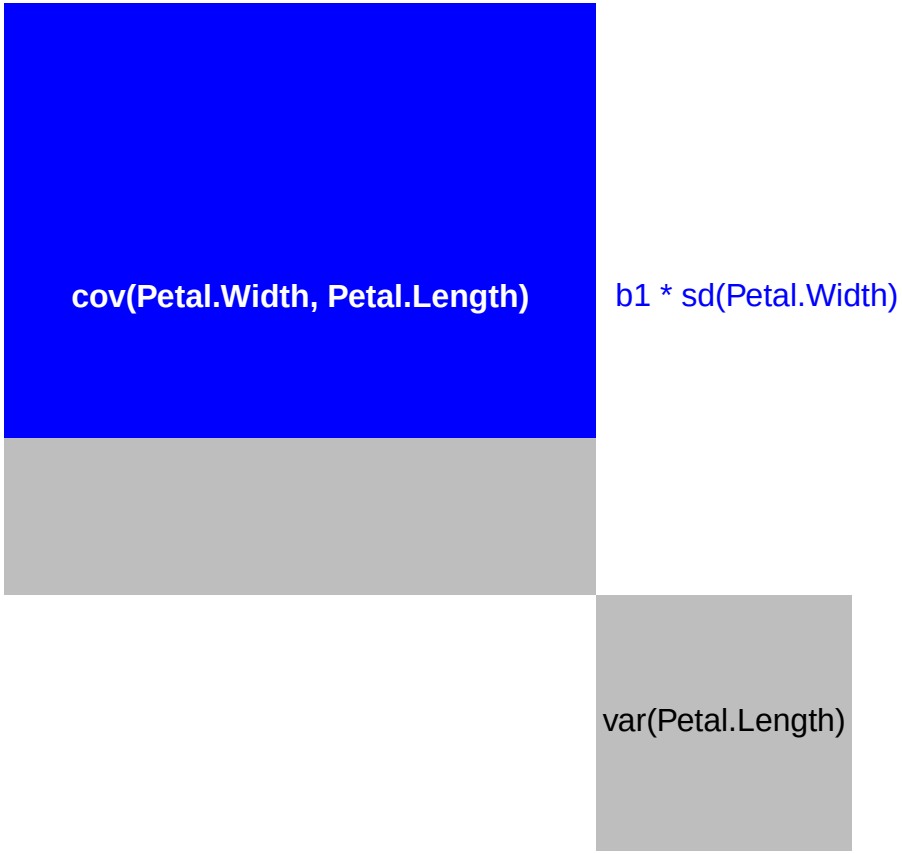
Our covariance picture



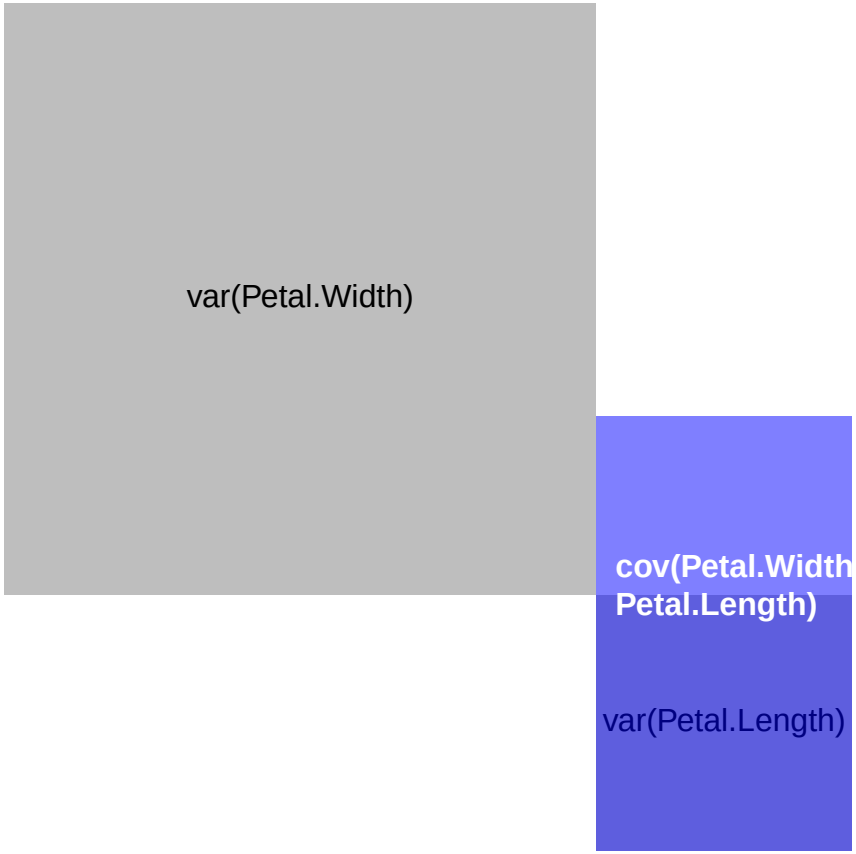
Lay the covariance over one of the variances instead.



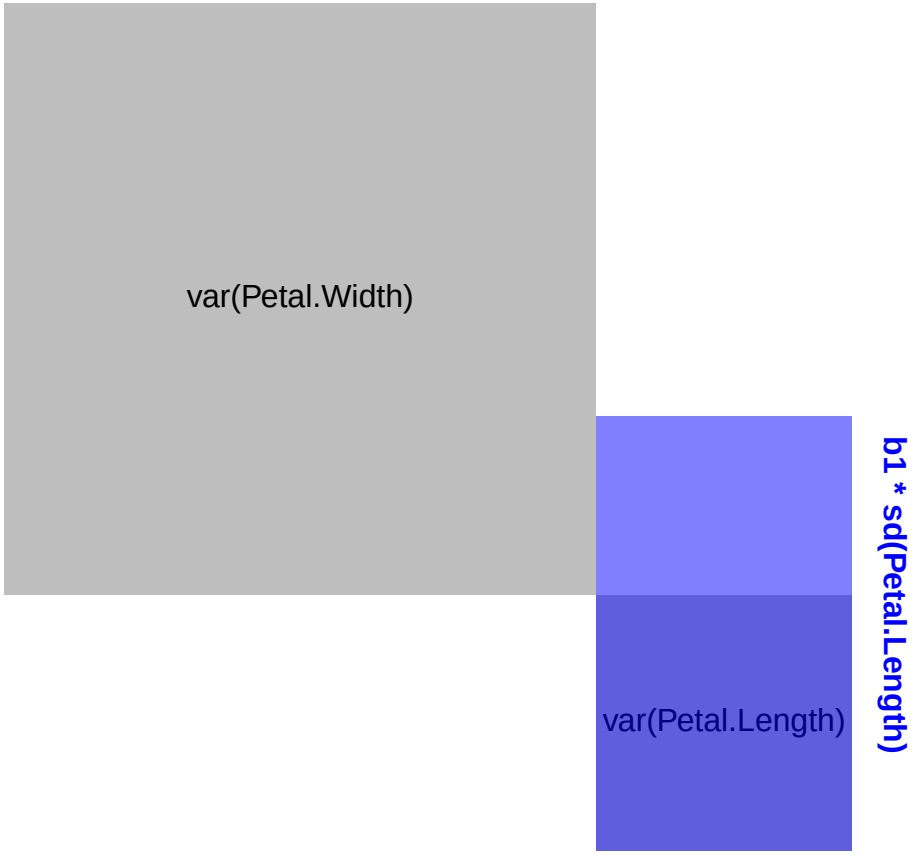
Petal.Length = b0 + b1 * Petal.Width



Lay the covariance over the other variance.



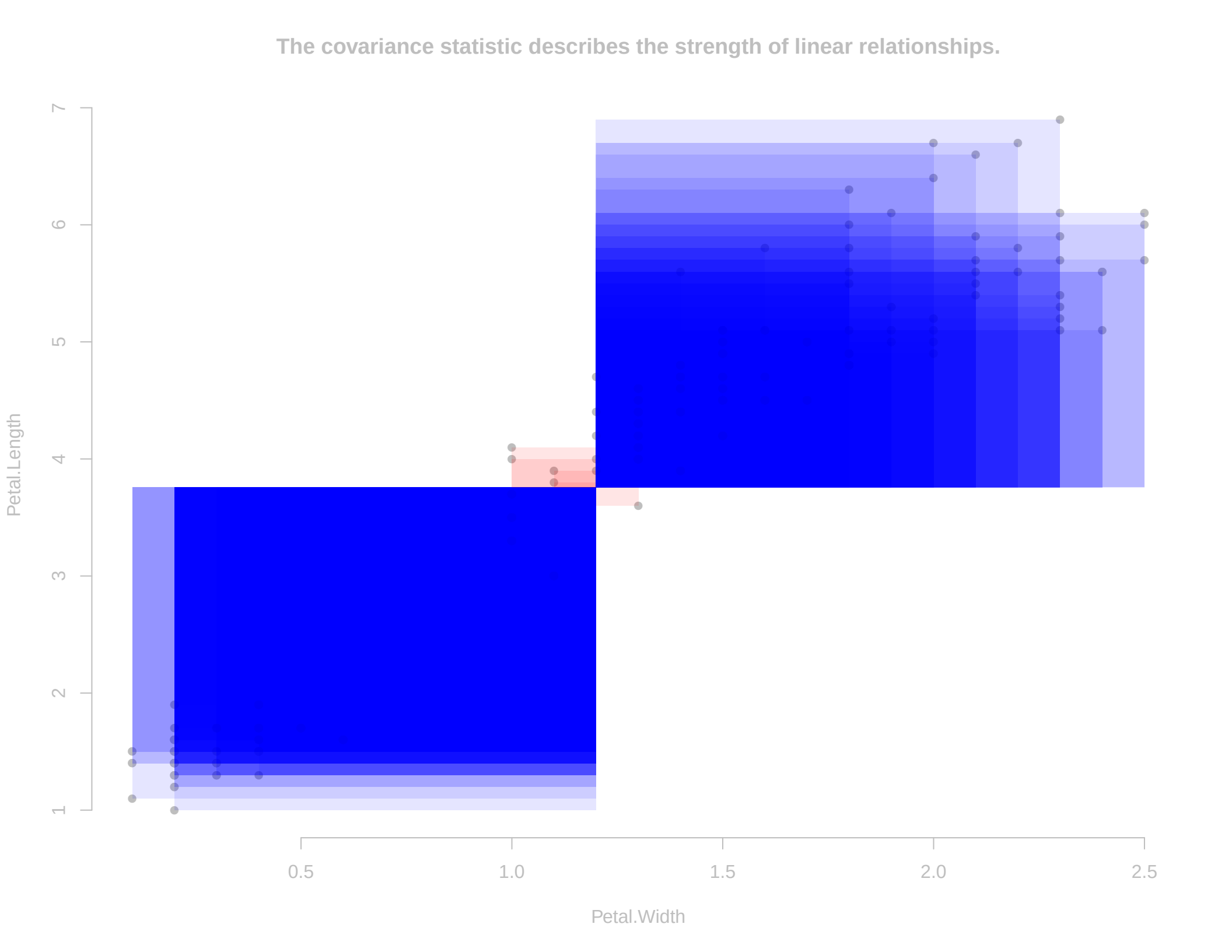
Petal.Width = b0 + b1 * Petal.Length

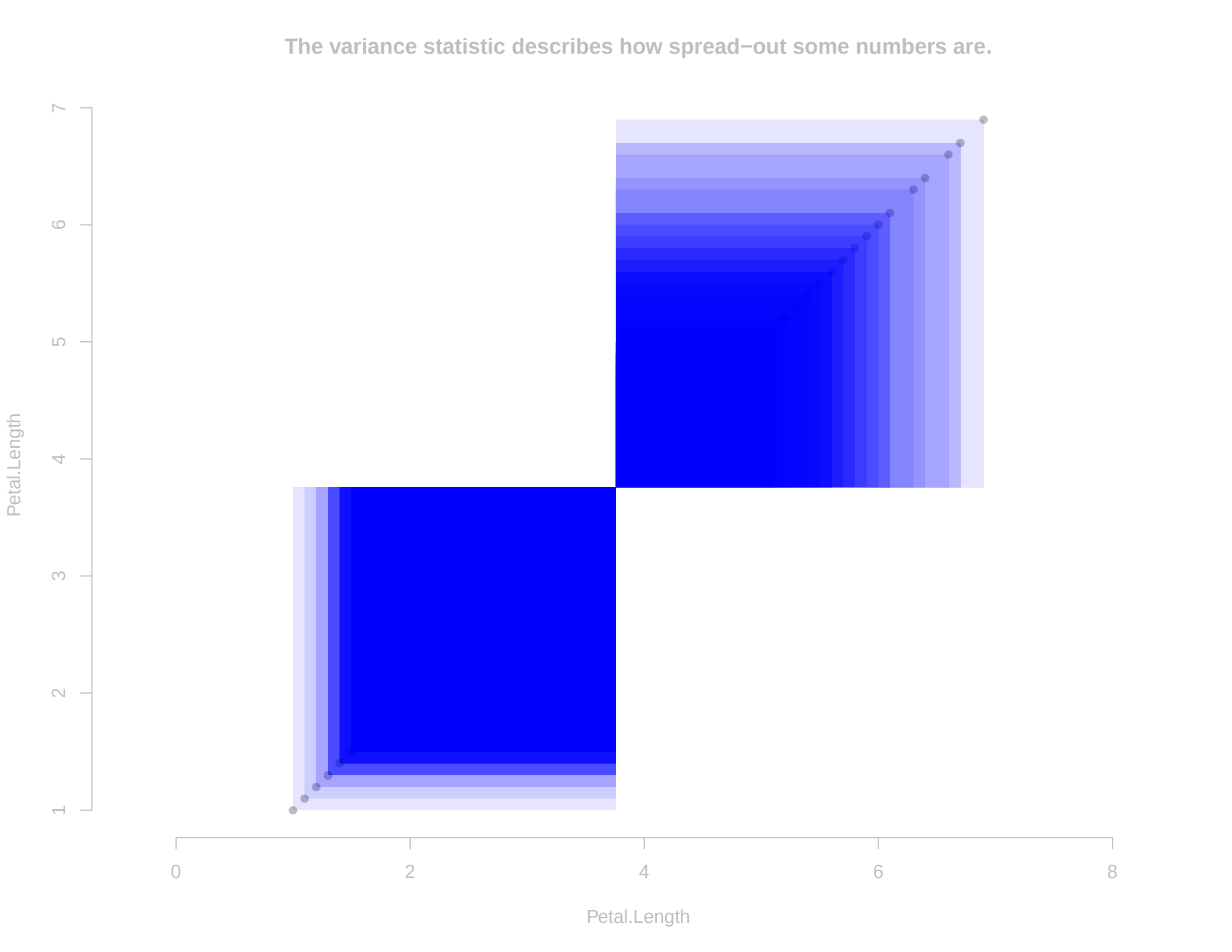


Some things to remember

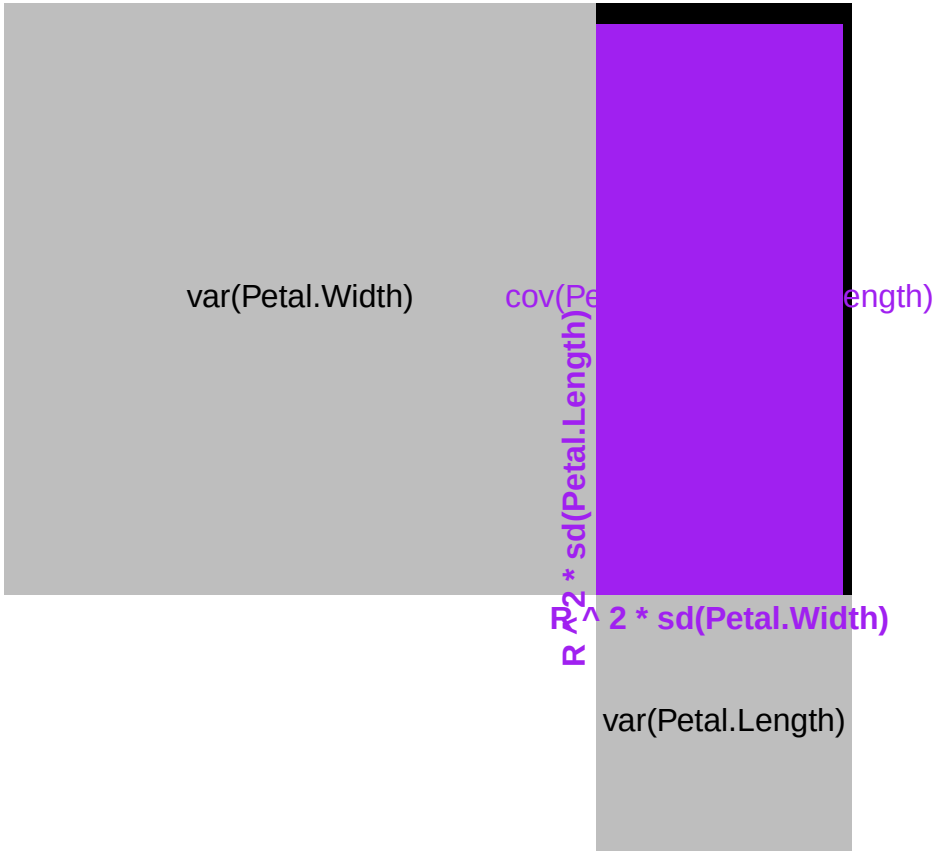
A statistic is a number that describes a lot of other numbers.

28	93	82	69	26	33	13	15	61	39	16	93	1	26	33	97	96	63	71	67	94
67	94	91	77	72	59	46	39	88	68	28	17	63	100	10	73	75	65	1	73	52
73	52	78	97	53	55	42	78	58	41	38	79	87	33	90	69	97	22	40	83	9
83	9	18	89	25	80	74	89	70	6	65	61	10	35	54	54	6	13	91	22	67
22	67	49	61	45	72	77	16	74	49	1	3	54	14	61	86	58	54	14	89	62
89	62	95	37	67	28	54	97	40	37	17	33	18	97	30	39	35	61	83	96	86
96	86	28	49	74	39	16	1	88	13	19	76	53	12	66	65	73	85	1	97	97
97	97	69	5	61	89	30	23	64	33	65	26	30	25	71	60	52	56	56	81	78
81	78	92	40	37	80	17	68	73	42	43	86	9	55	64	87	86	74	35	68	8
68	8	88	67	62	54	34	58	65	64	78	46	92	12	17	67	96	35	44	76	95
76	95	30	58	29	59	17	53	6	80	57	42	23	98	47	55	94	50	82	54	97
54	97	22	82	8	64	64	6	72	23	15	52	53	50	21	56	23	62	11	50	52
50	52	37	39	76	75	64	24	5	27	10	3	91	10	43	41	46	51	70	51	77
51	77	47	96	53	6	89	6	59	45	22	27	51	69	66	0	6	63	80	27	20
27	20	5	70	39	42	77	95	20	71	74	14	23	47	84	33	3	55	33	16	61
16	61	22	10	90	45	13	99	6	60	24	41	80	48	44	91	57	72	58	5	61
5	61	62	72	17	90	28	15	64	3	64	59	2	26	19	81	59	38	35	22	57
22	57	79	85	90	21	16	61	37	60	6	36	82	40	8	82	24	78	94	59	68
59	68	61	45	96	64	48	41	59	76	40	68	4	86	51	91	62	5	61	43	57
43	57	26	26	46	8	71	94	88	15	7	91	55	39	31	45	56	50	14	95	58
95	58	44	54	8	83	93	43	84	22	83	76	68	18	34	67	2	73	97	11	28

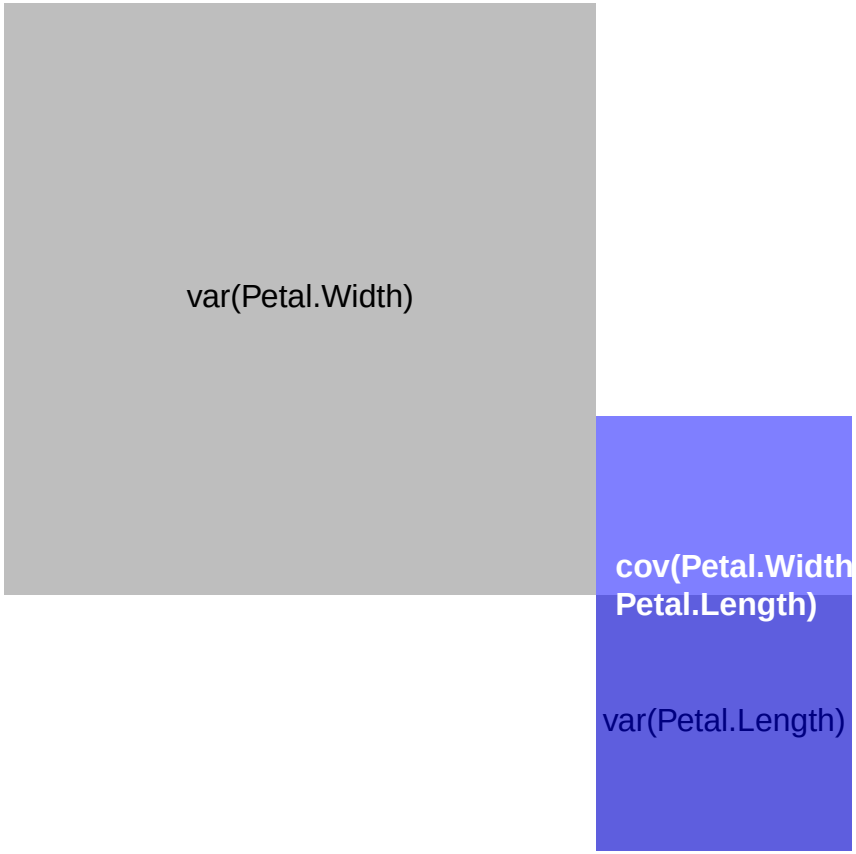

The covariance statistic describes the strength of linear relationships.
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


The variance statistic describes how spread−out some numbers are.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

The correlation statistic is a standardized version of covariance.



(Beta coefficients for) least−squares regression predict one variable based on another.



You can pretty much always draw math.