Statistics with doodles

Thomas Levine
thomaslevine.com

Why we have statistics

Lots of numbers

57	28	94	86	27	75	58	97	58	36	26	89	47	40	23	60	87	34	9	58	46
58	46	9	50	87	77	36	42	20	25	68	76	78	61	52	89	73	53	71	31	70
31	70	45	11	65	36	25	35	90	53	27	53	85	1	17	23	49	71	18	63	93
63	93	15	7	27	43	8	5	9	33	10	82	99	53	87	25	41	38	13	97	43
97	43	16	71	98	44	28	82	23	60	66	65	45	88	43	19	92	41	94	46	56
46	56	43	37	4	32	54	55	75	11	38	47	30	56	7	3	14	53	54	24	25
24	25	19	6	57	72	43	64	82	25	83	65	1	32	90	97	22	99	23	51	86
51	86	1	9	93	49	46	24	93	9	16	1	46	9	62	12	50	47	4	72	78
72	78	80	16	52	65	87	50	52	77	8	29	61	6	97	53	86	98	28	37	21
37	21	14	99	44	67	28	77	8	26	92	48	94	62	78	25	93	89	18	22	77
22	77	3	45	48	54	44	28	74	40	43	30	32	22	41	57	99	11	76	18	70
18	70	64	82	57	27	98	5	78	52	100	12	17	93	43	20	32	18	86	56	75
56	75	30	42	13	65	90	23	96	98	2	54	89	30	26	50	50	93	16	15	73
15	73	27	60	26	46	50	78	6	58	98	82	52	11	86	28	20	1	29	76	30
76	30	25	36	30	3	22	31	62	49	18	11	97	34	9	95	86	59	86	70	6
70	6	11	71	41	13	98	26	76	39	51	97	44	12	20	58	29	32	91	84	25
84	25	90	93	75	65	61	79	98	54	27	66	44	83	6	14	3	20	97	21	53
21	53	29	45	28	95	33	11	75	69	25	91	79	75	49	96	70	59	40	19	50
19	50	63	59	52	83	19	27	75	26	30	88	24	72	16	73	79	43	16	5	67
5	67	77	94	33	81	14	36	43	97	27	25	60	10	32	27	44	59	65	36	48
36	48	58	59	61	15	13	33	33	22	63	18	89	79	71	77	57	38	87	40	57

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

85	89	21	10	90	59	84	94	59	96	61	90	48	24	95	72	87	22	4	90	92
90	92	23	62	70	53	53	44	29	70	18	20	9	58	51	95	20	0	27	44	26
44	26	10	21	80	24	78	49	84	6	41	82	37	72	93	54	74	46	35	26	84
26	84	4	36	68	1	62	64	38	82	85	21	50	87	38	11	16	10	92	90	24
90	24	27	86	92	96	97	25	22	95	56	4	27	57	10	80	58	7	37	98	23
98	23	68	25	9	71	49	49	91	44	69	65	43	39	77	72	22	40	47	88	8
88	8	28	39	67	33	16	25	12	46	31	51	100	46	30	48	78	38	8	50	43
50	43	43	64	35	31	30	43	90	91	44	15	63	6	82	31	93	39	49	50	15
50	15	56	22	70	22	38	5	83	94	11	2	26	100	1	47	1	81	97	92	60
92	60	82	100	96	42	99	23	83	11	94	55	82	97	64	99	55	14	71	42	11
42	11	26	74	27	92	49	90	53	82	74	75	99	78	36	14	82	29	9	4	21
4	21	83	91	91	15	23	53	3	34	64	86	74	82	7	21	44	40	7	52	11
52	11	45	81	27	88	18	82	71	65	2	66	33	29	28	41	52	89	10	64	87
64	87	44	22	25	54	37	50	51	60	66	83	72	62	8	52	15	46	15	5	5
5	5	27	63	73	16	23	54	23	40	22	96	39	52	25	68	93	79	94	49	69
49	69	66	38	18	74	8	11	83	53	36	23	3	11	40	54	7	87	64	80	93
80	93	44	74	31	16	83	78	56	91	66	45	52	70	81	3	68	66	29	84	35
84	35	35	22	35	8	48	98	42	39	70	11	22	59	38	44	39	78	13	53	35
53	35	79	3	31	8	79	63	87	9	23	45	52	51	52	94	58	98	25	63	31
63	31	58	9	17	48	61	94	3	32	22	96	66	68	66	11	98	84	98	31	75
31	75	74	90	83	73	82	85	26	17	33	86	70	92	96	70	7	95	13	86	85

**So we invent numbers
that describe
lots of other numbers**

So we invent numbers that describe lots of other numbers

(statistics)

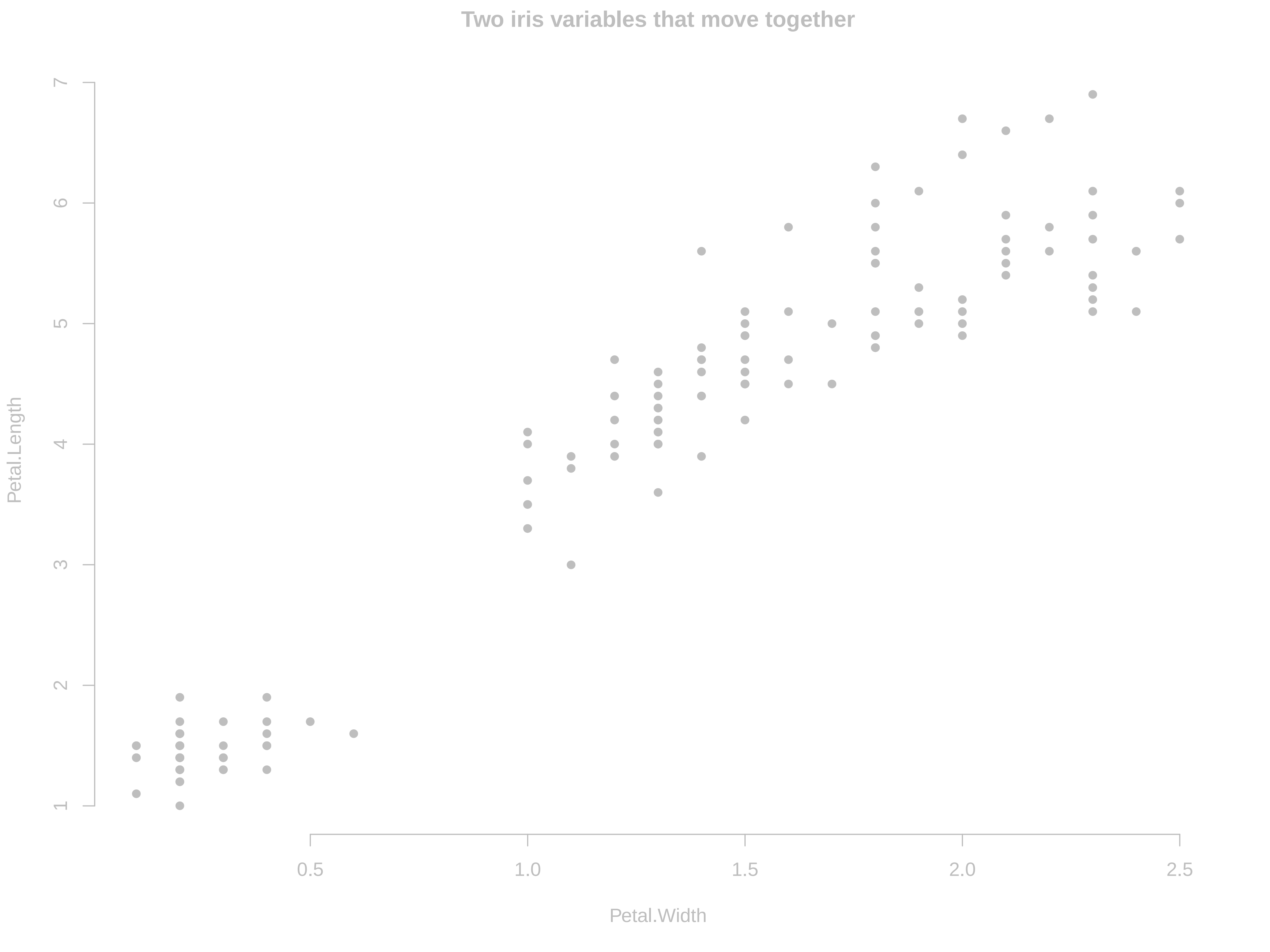
Here are some numbers:
1 2.2 pi 4 5 7 7

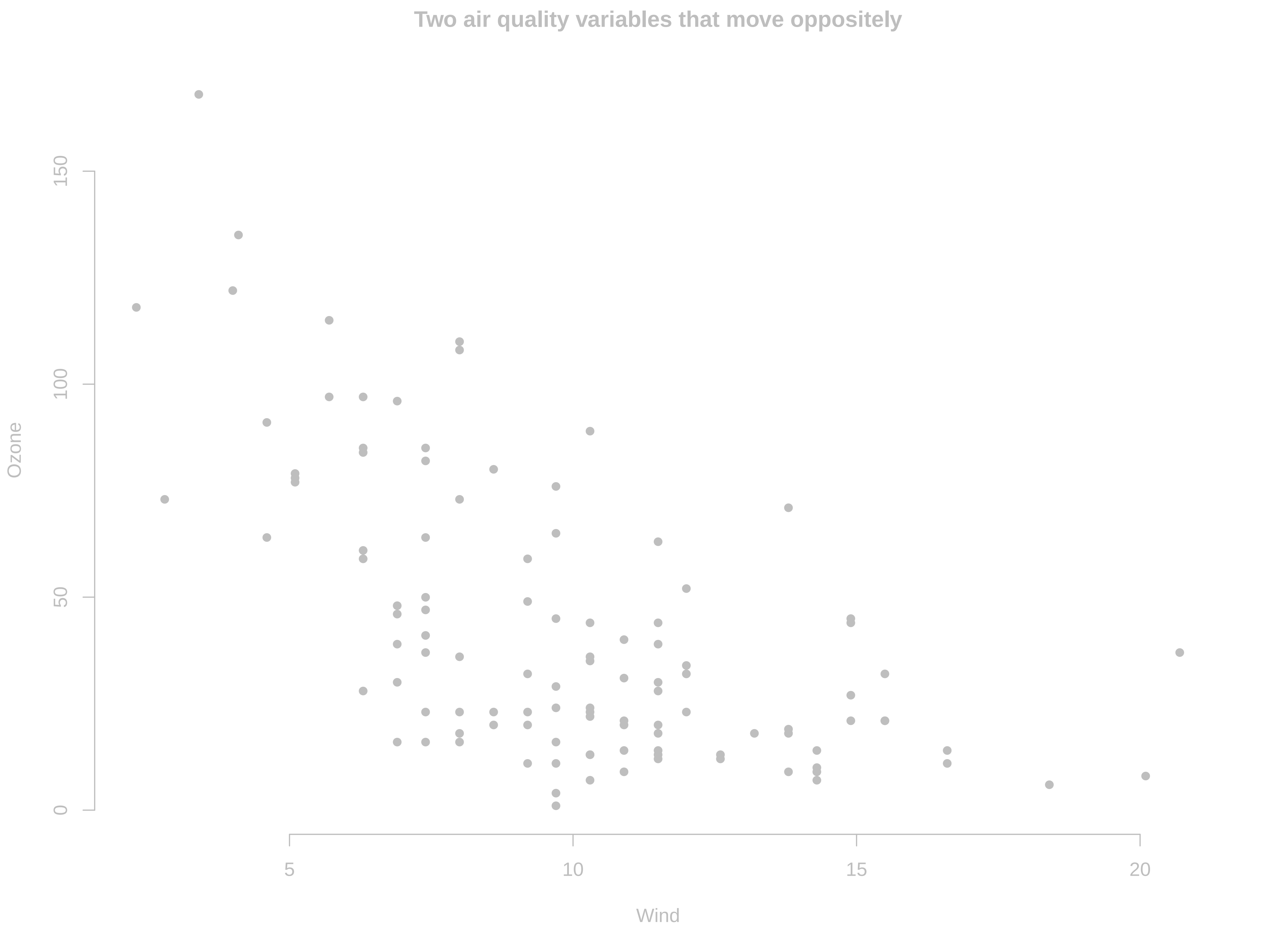
What are some statistics?

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

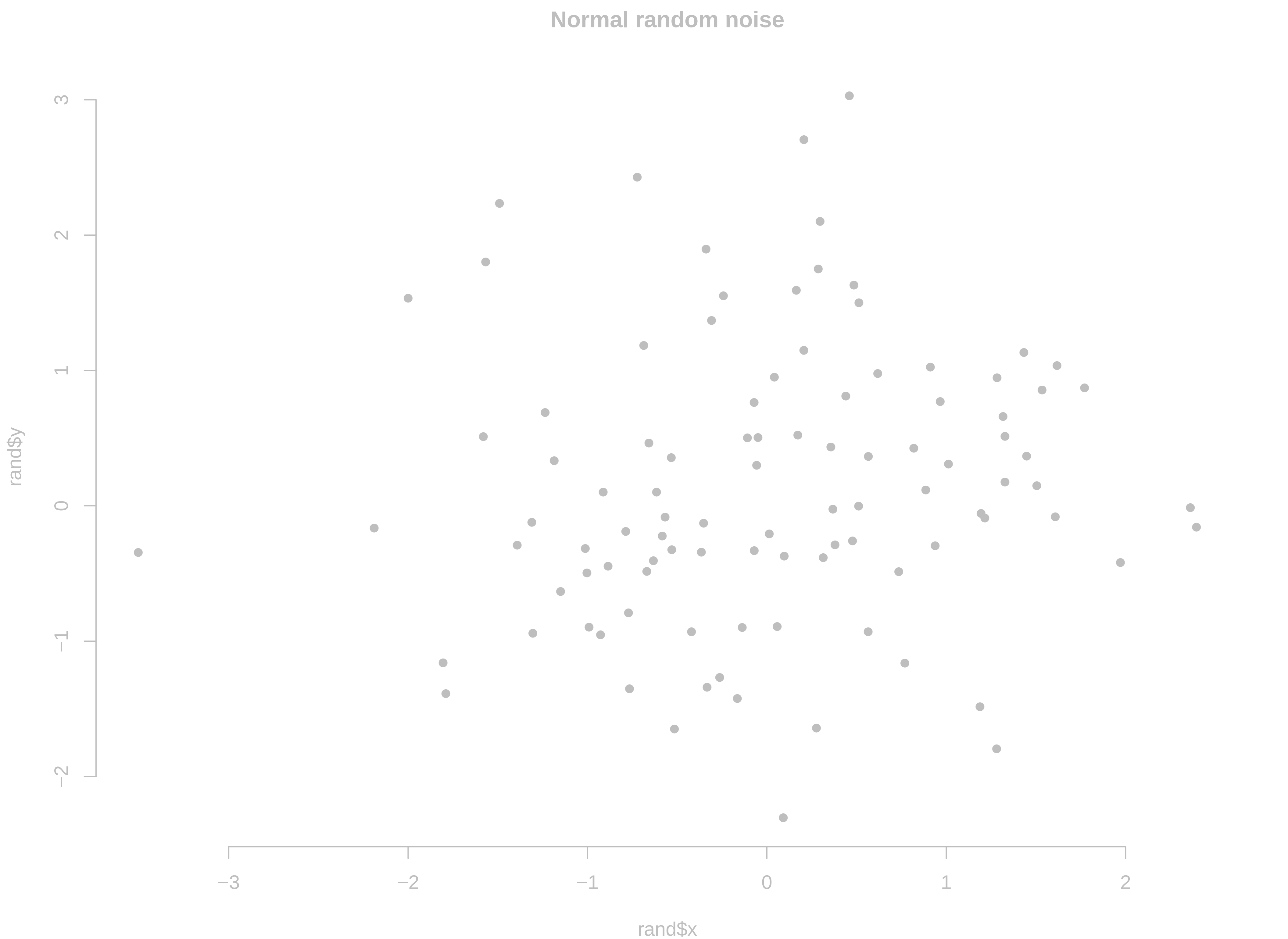
how many integers,
whether the numbers are sorted
&c.

Measuring linear relationships

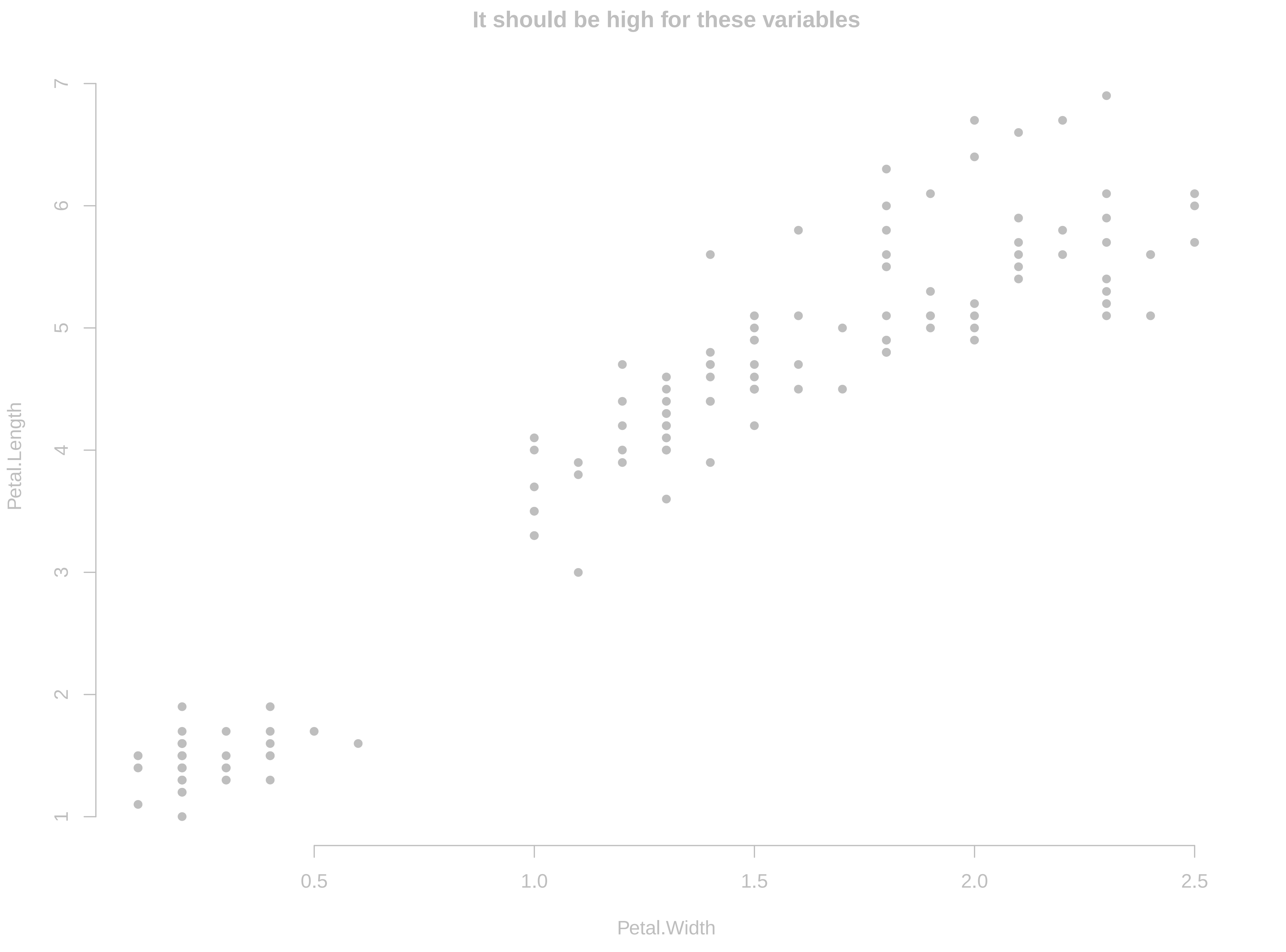

Two iris variables that move together
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

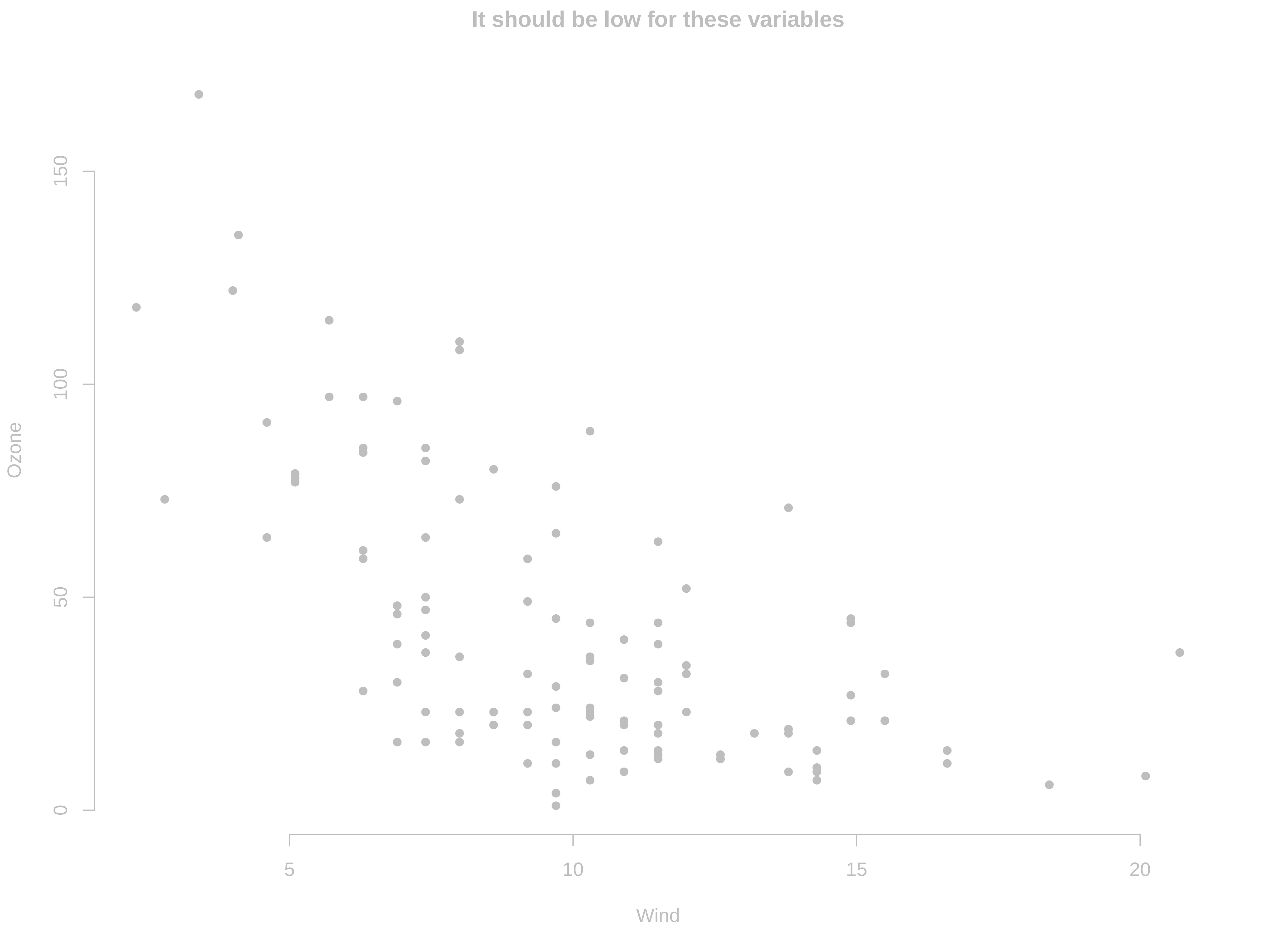


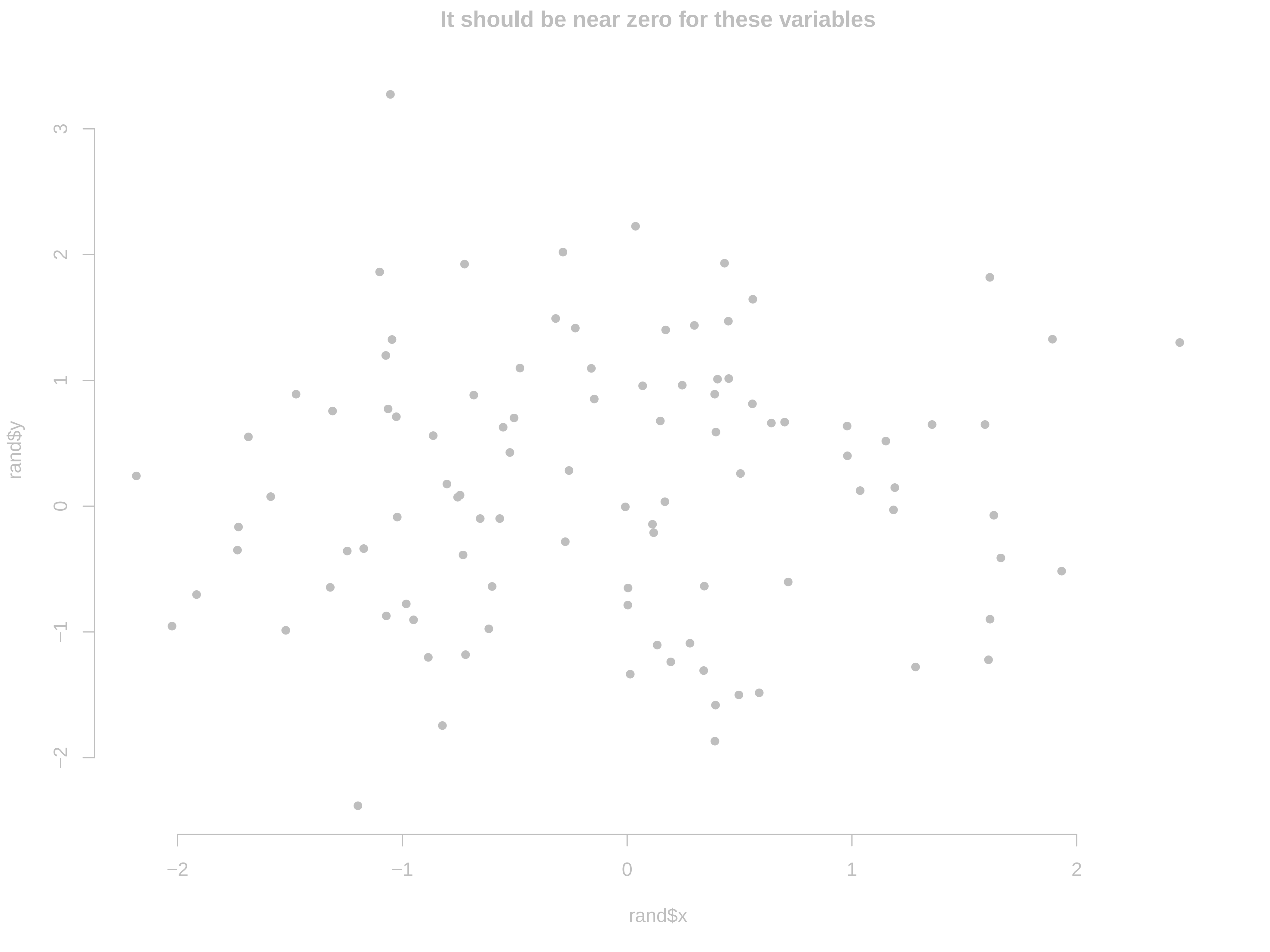
Two air quality variables that move oppositely
150
100
50
0
Ozone
5
10
15
20
Wind


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

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

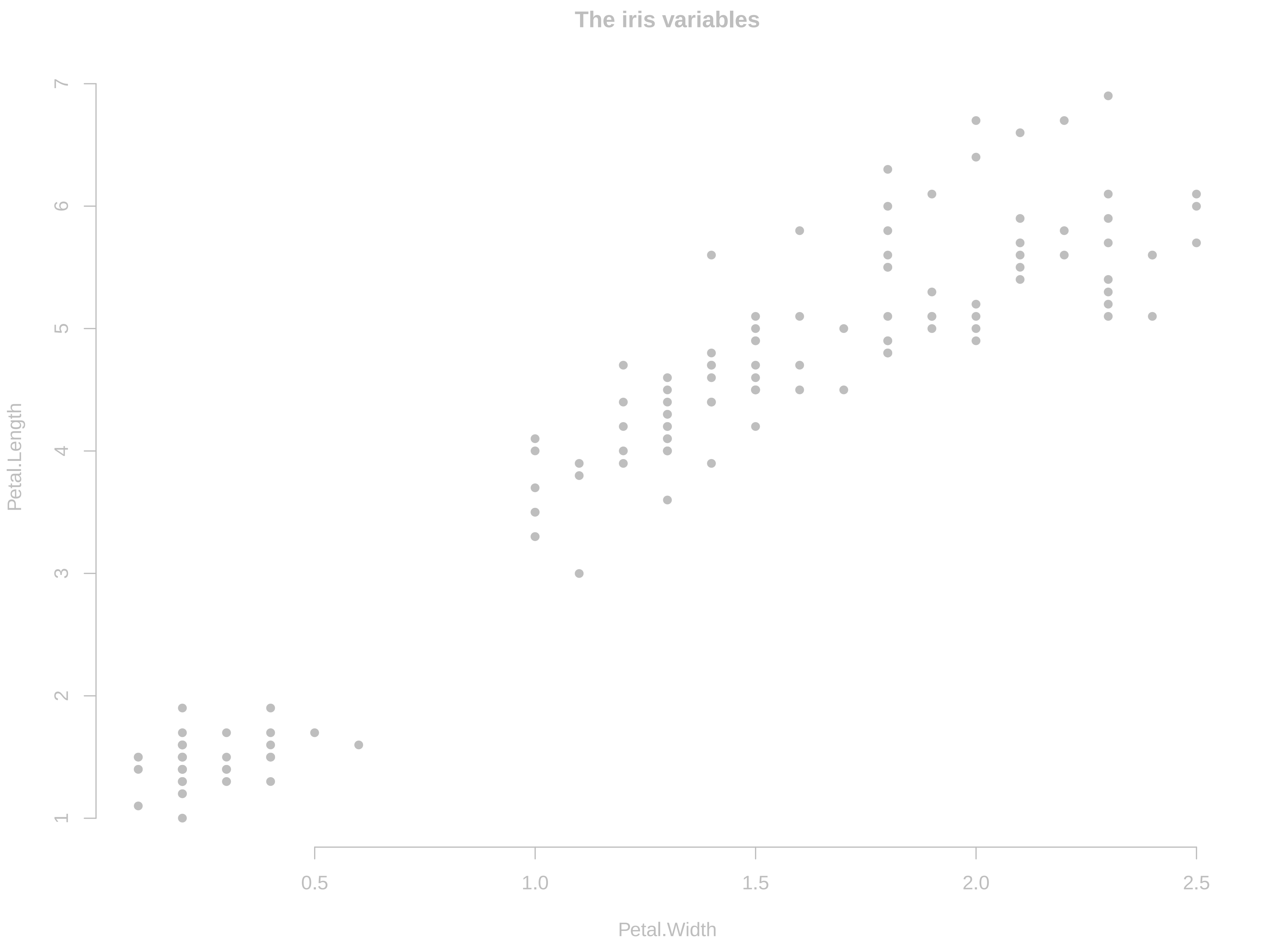

It should be high for these variables
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

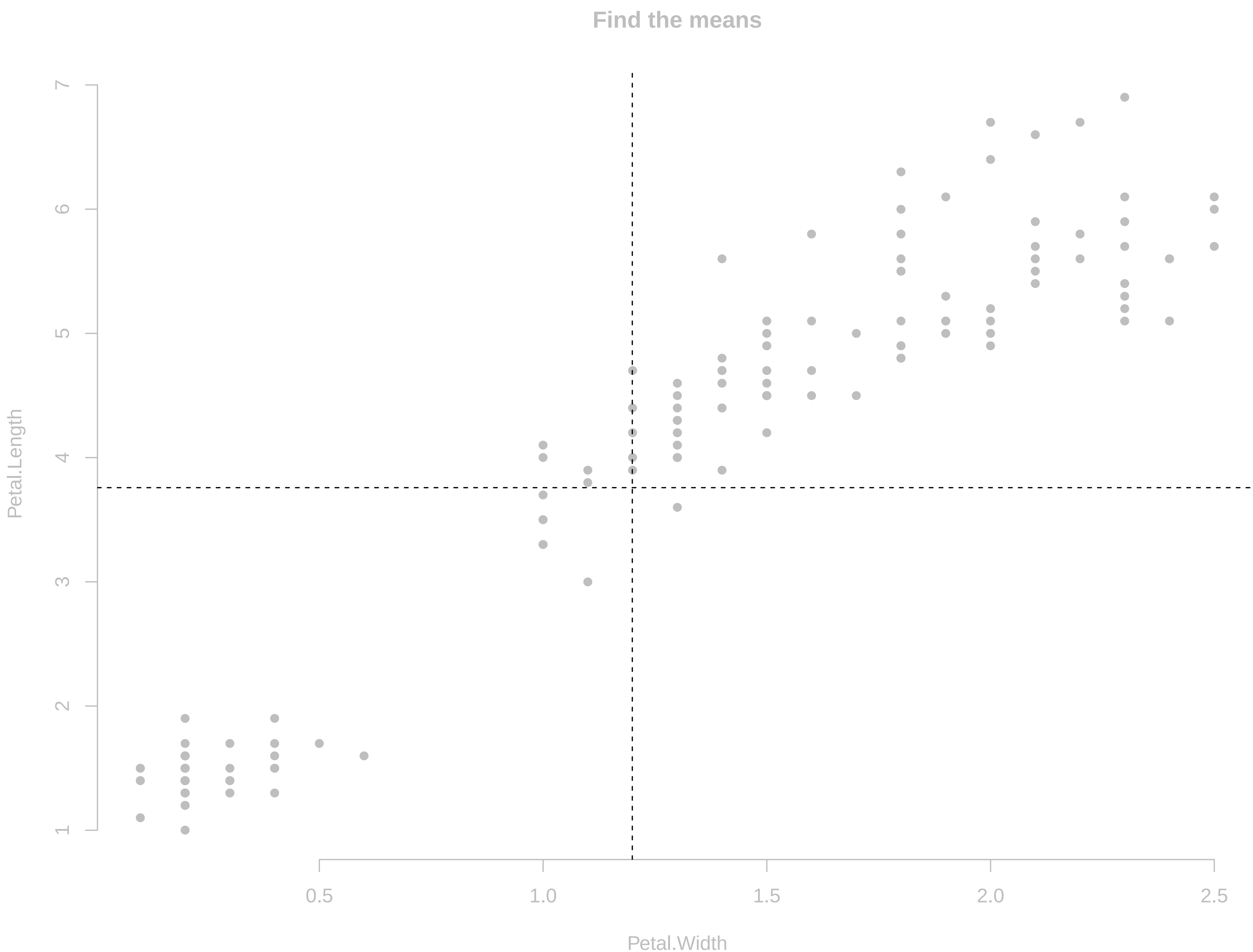

It should be low for these variables
Ozone
0
50
100
150
Wind
5
10
15
20

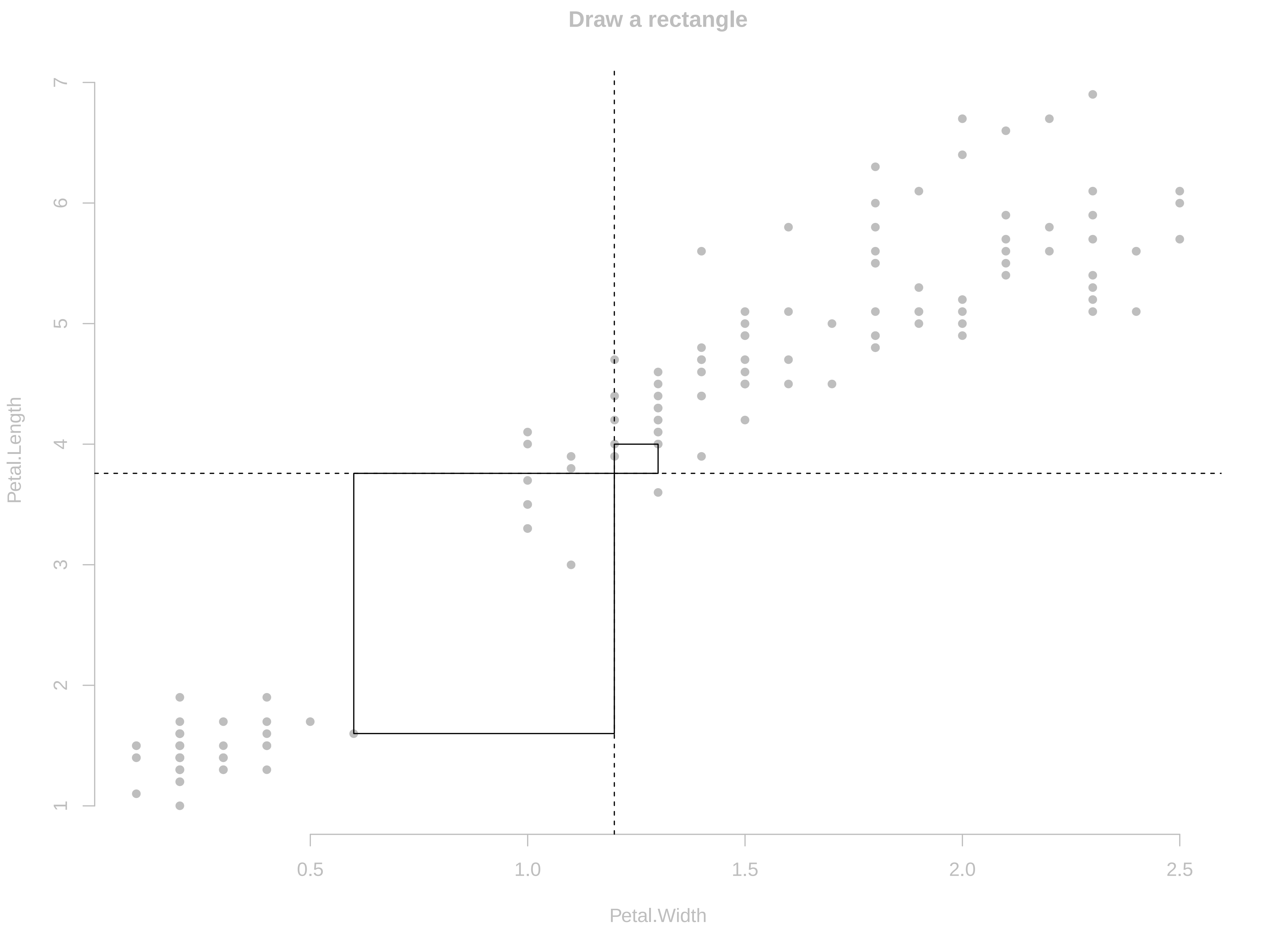


It should be near zero for these variables
rand$y
rand$x
−2
−1
0
1
2
3

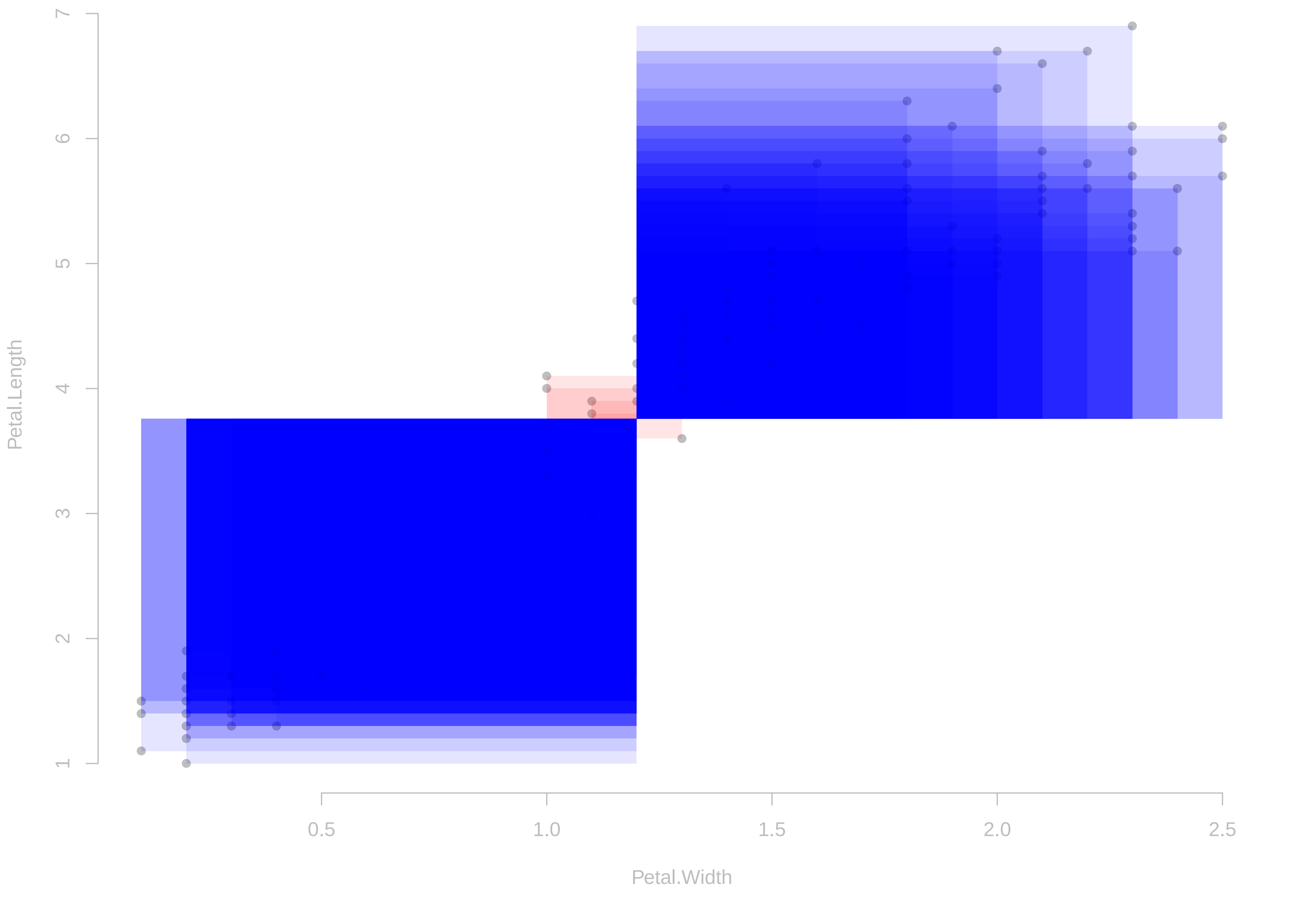
Covariance

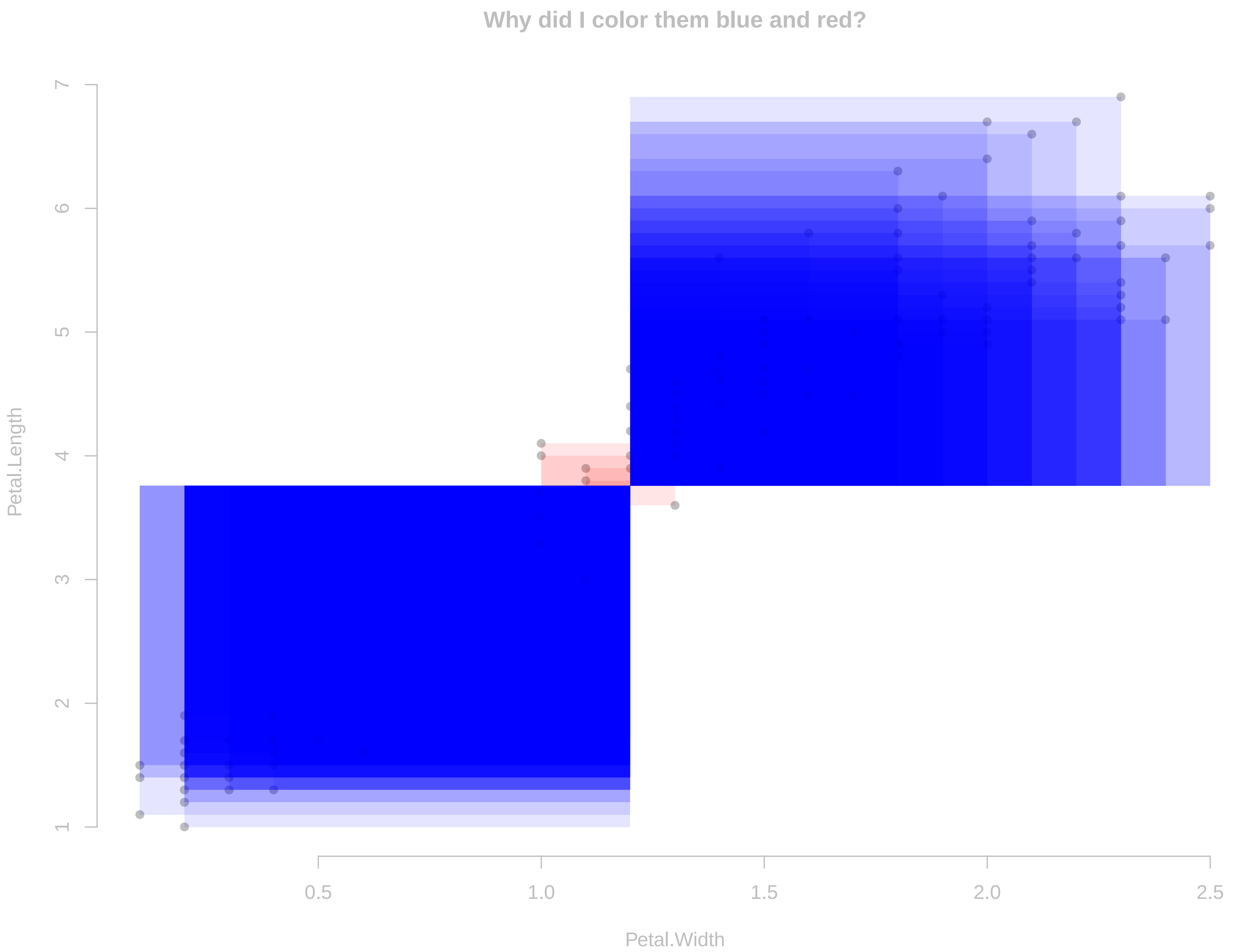

The iris variables
Petal.Length
1
2
3
4
5
6
7
Petal.Width
0.5
1.0
1.5
2.0
2.5


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

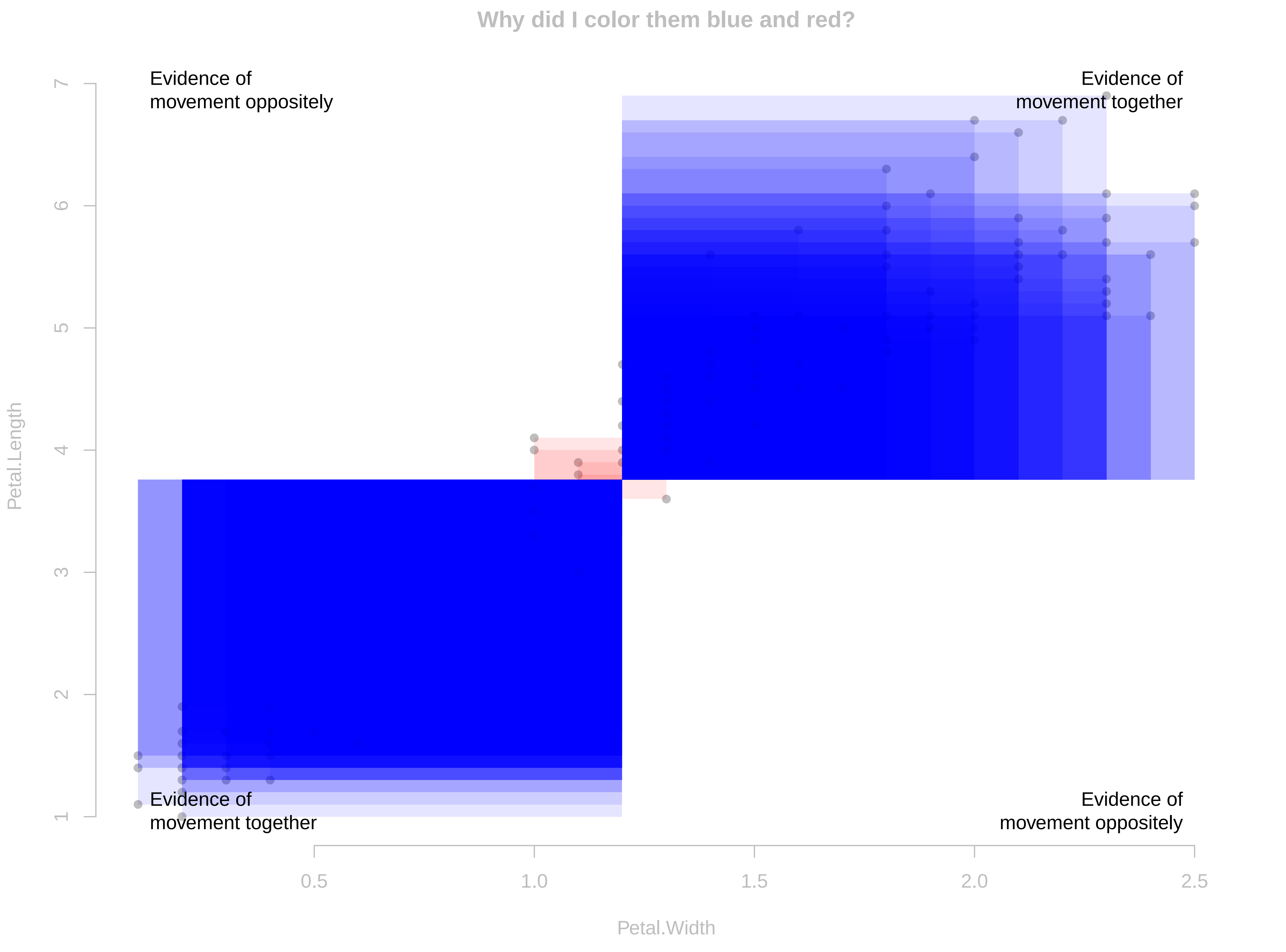

Draw a rectangle
Petal.Length
1
2
3
4
5
6
7
Petal.Width
0.5
1.0
1.5
2.0
2.5

Draw all the rectangles



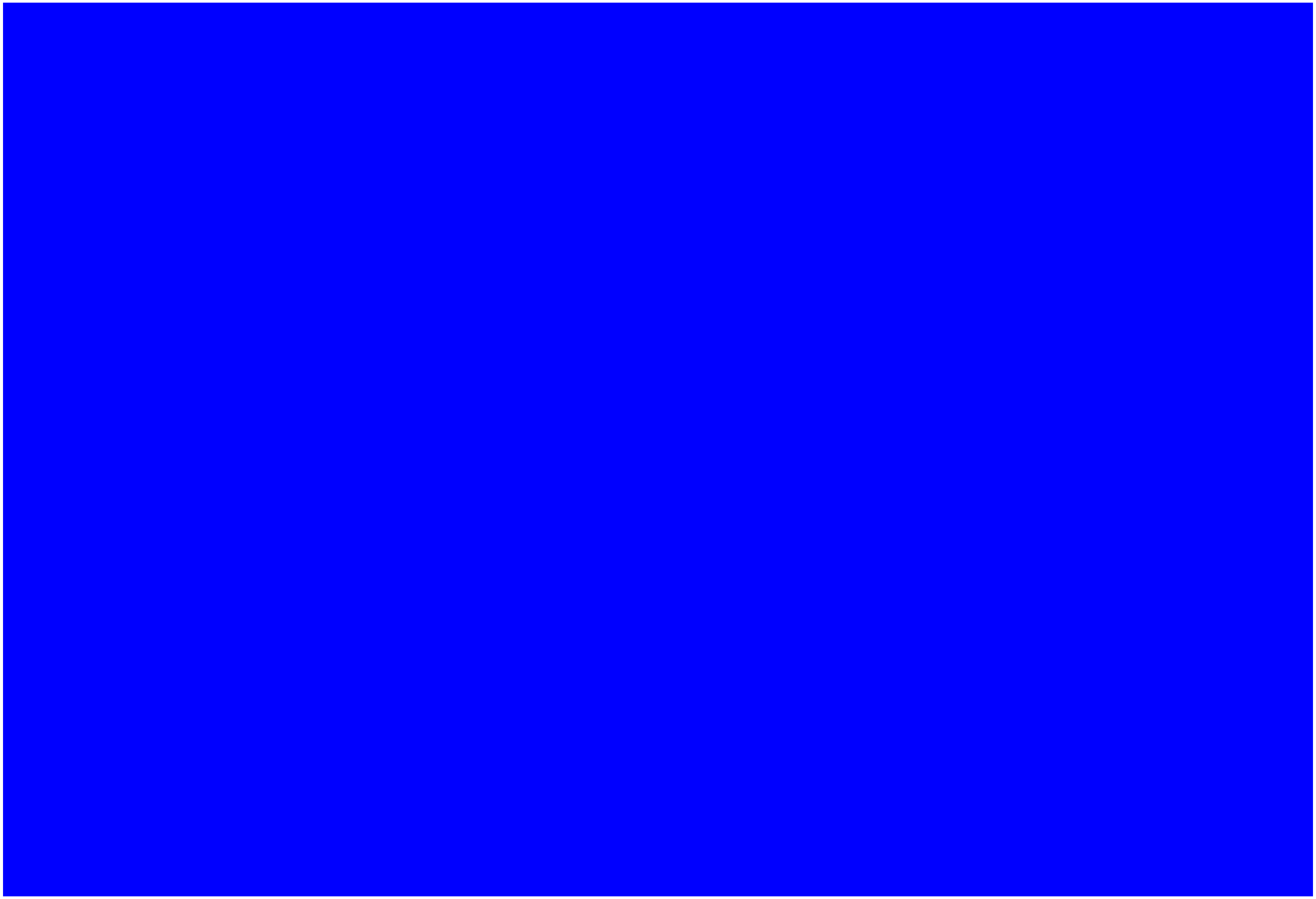


Why did I color them blue and red?
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

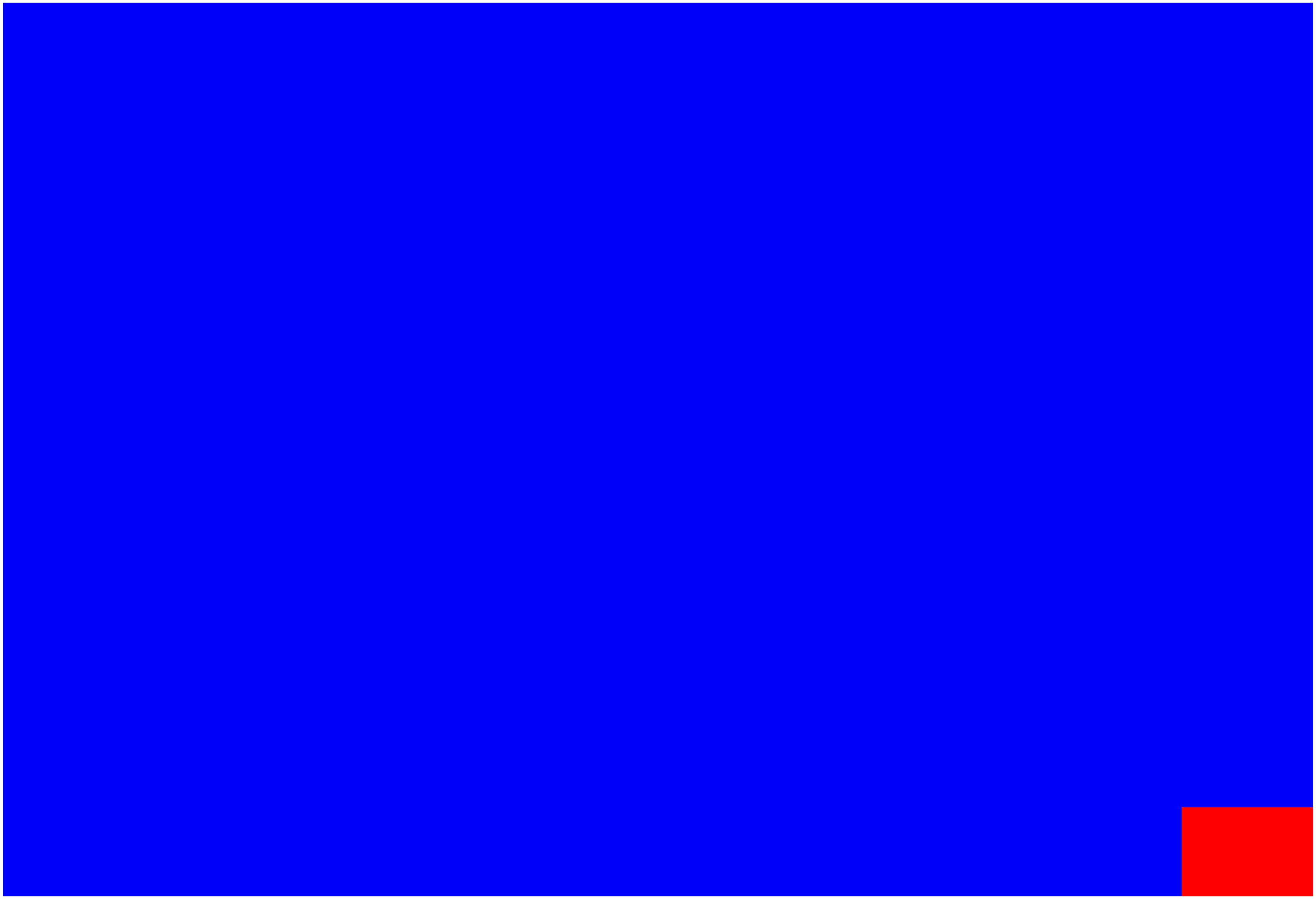


Why did I color them blue and red?
Evidence of
movement oppositely
Evidence of
movement together
Evidence of
movement together
Evidence of
movement oppositely
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

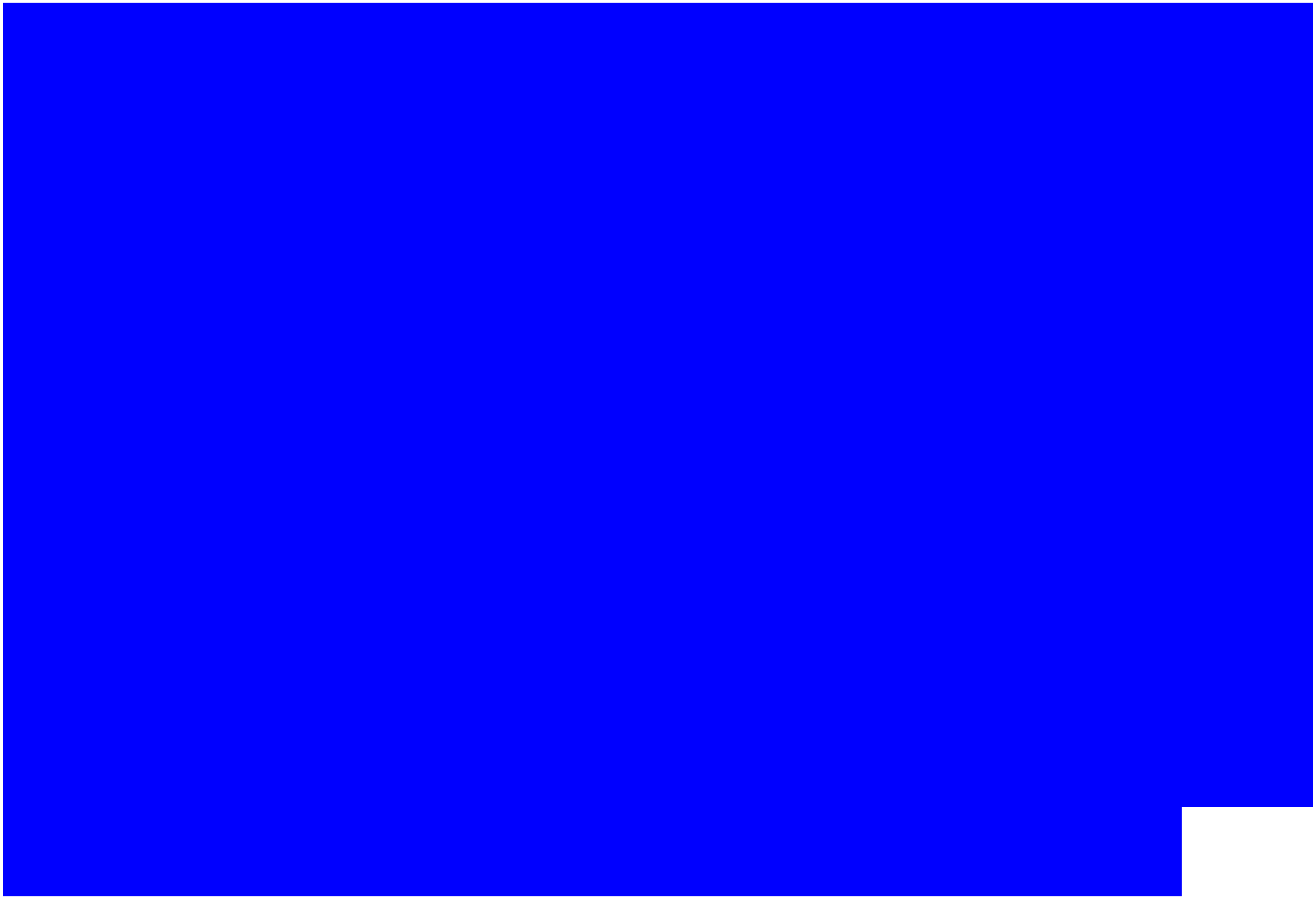
Add the blues together. (This is at a different scale.)



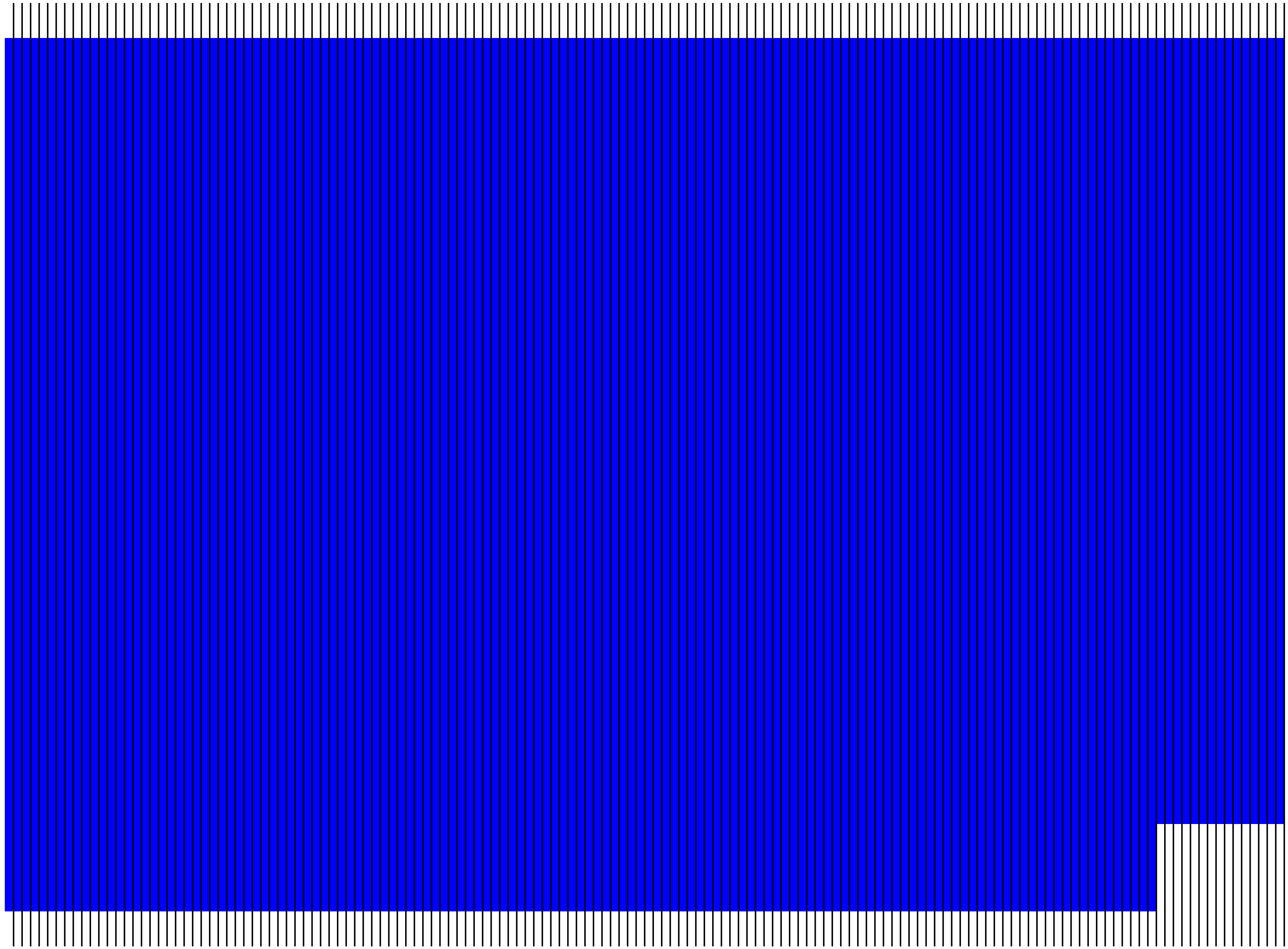
Add the reds together.



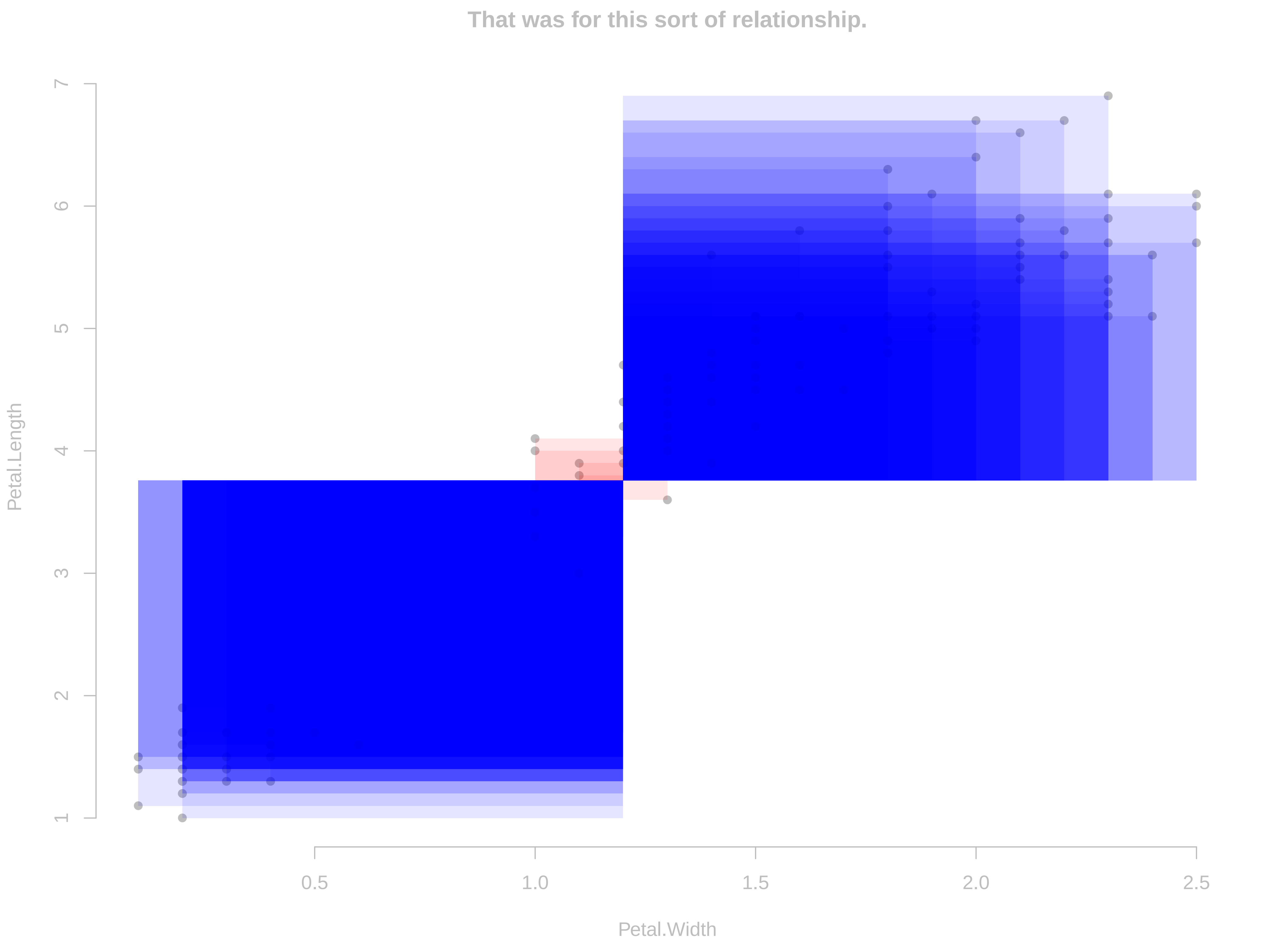
Subtract the reds.



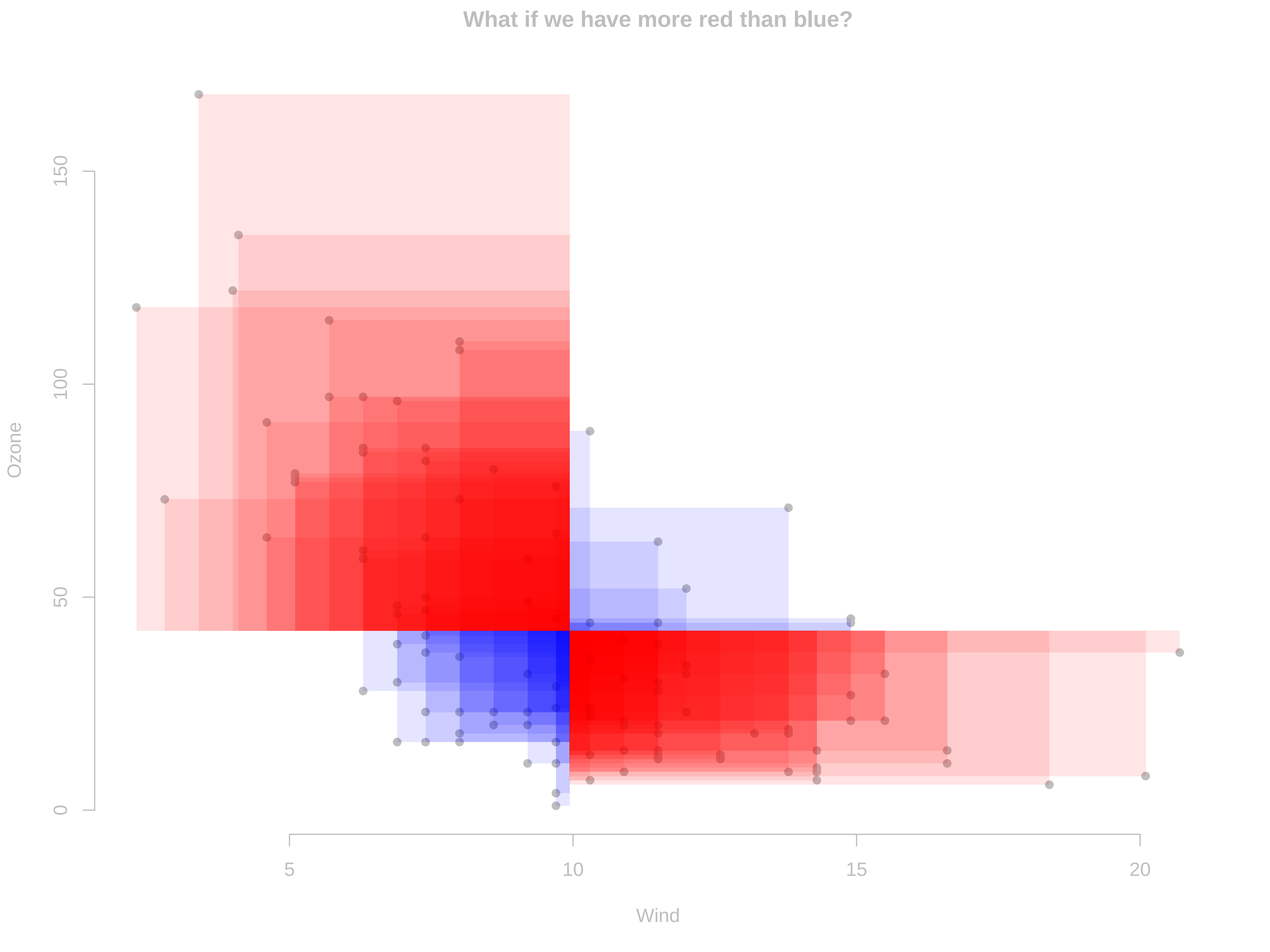
Divide into as many equal pieces as we have irises (n).



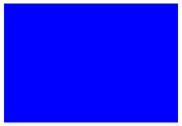
This blue sliver is the covariance.



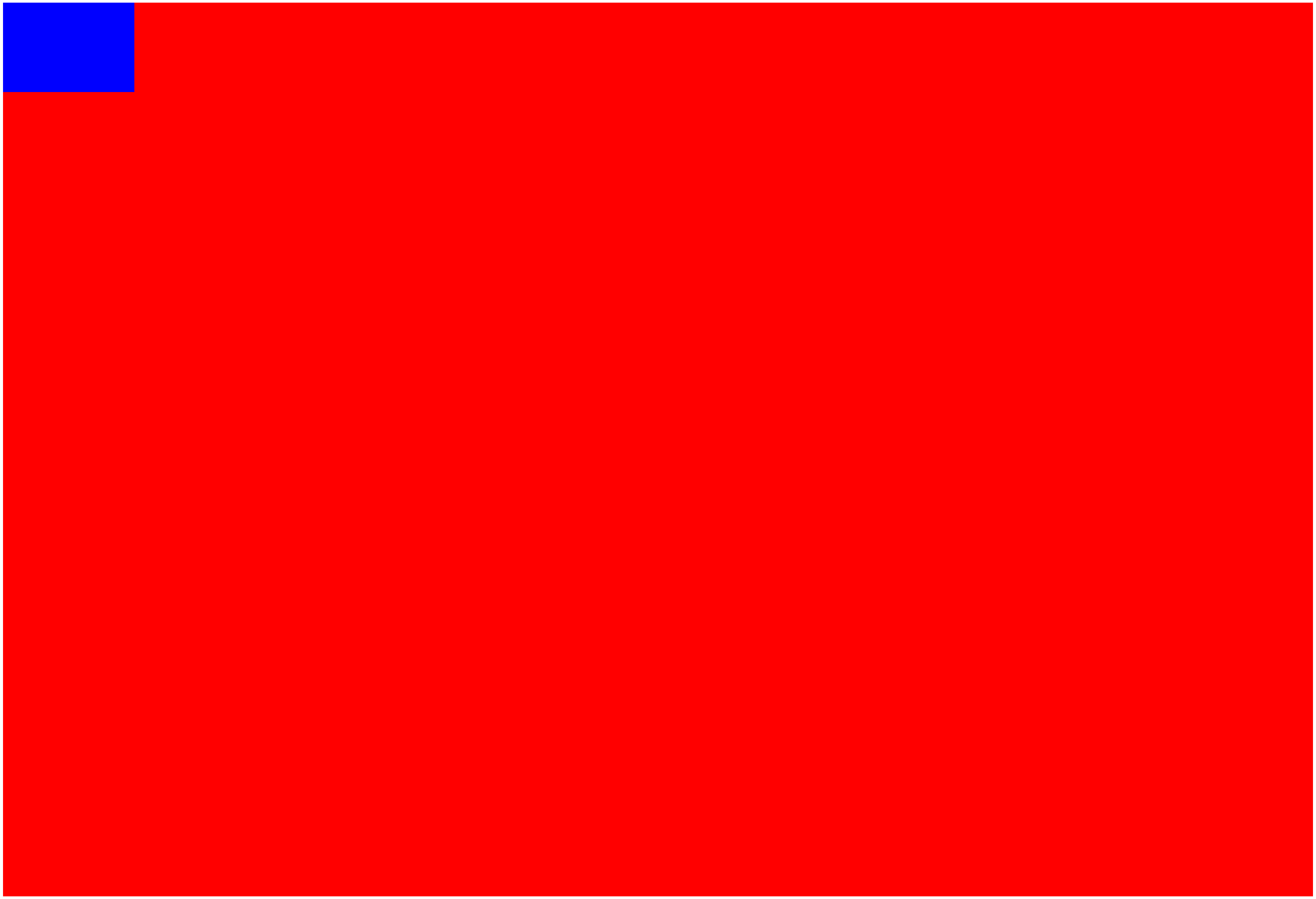
That was for this sort of relationship.
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


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

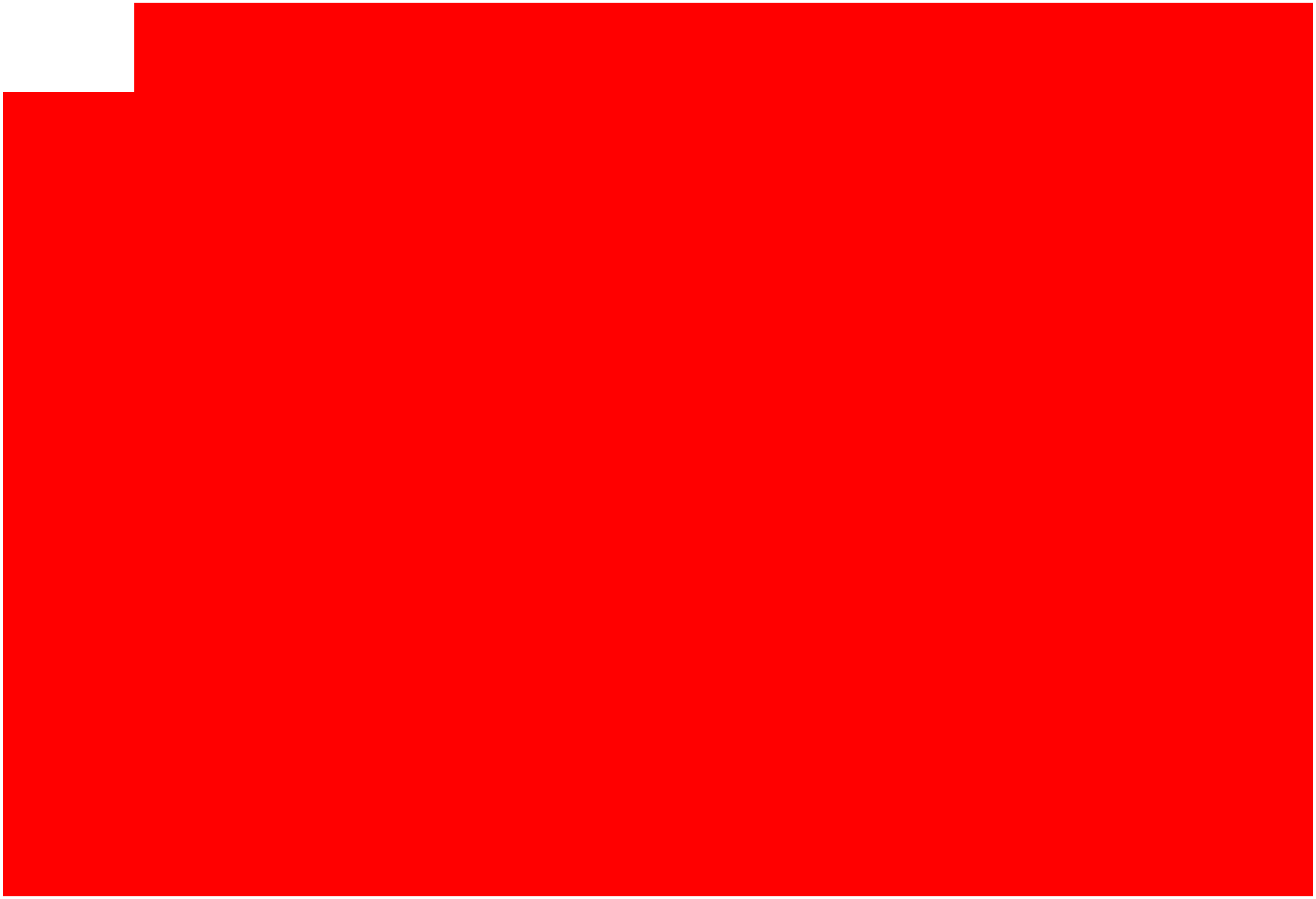
Add the blues together. (This is at a different scale.)



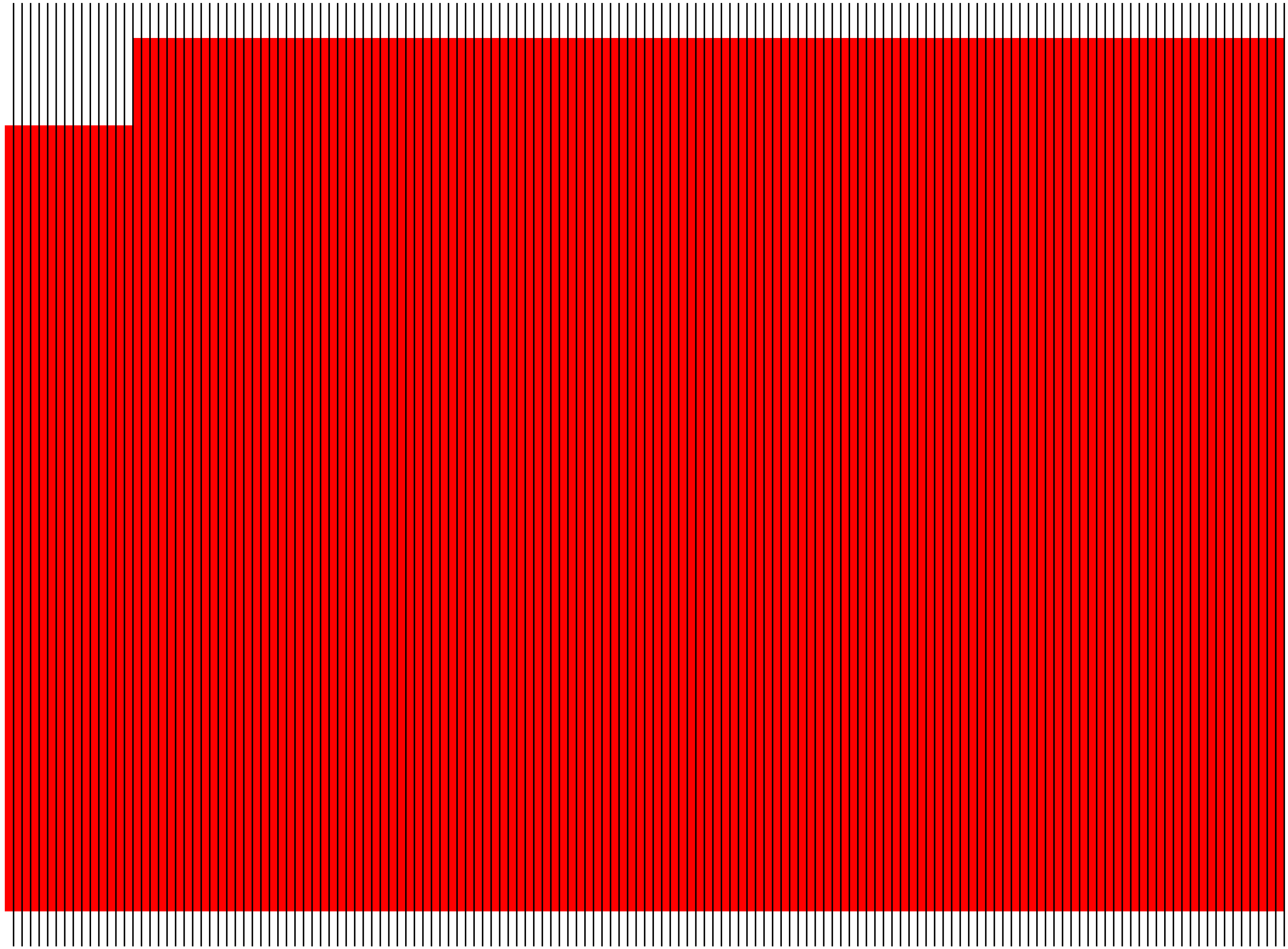
Add the reds together.



Subtract the reds.



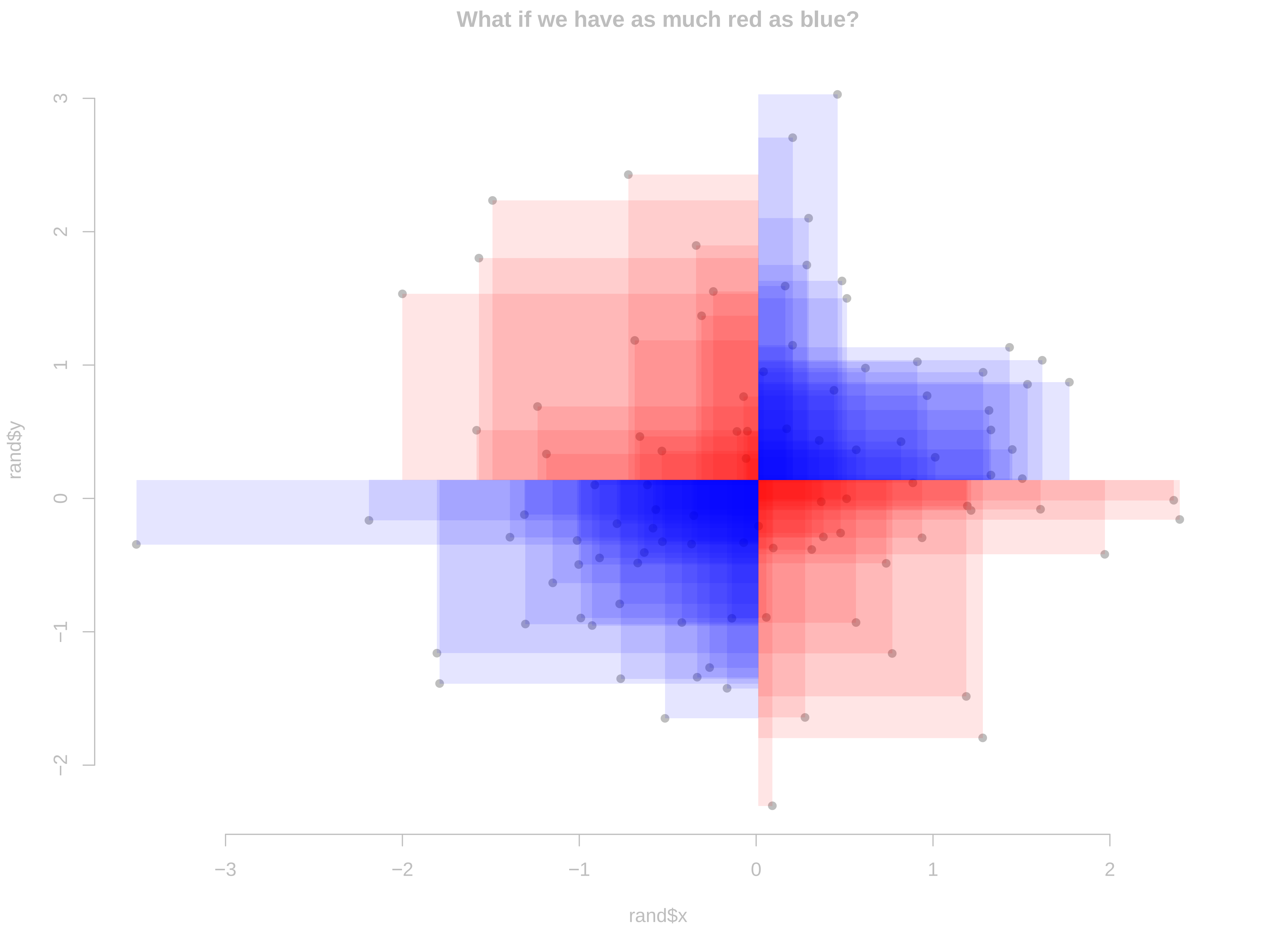
Divide into as many equal pieces as we have irises (n).



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$y
rand$x
3
2
1
0
−1
−2
−3
−2
−1
0
1
2

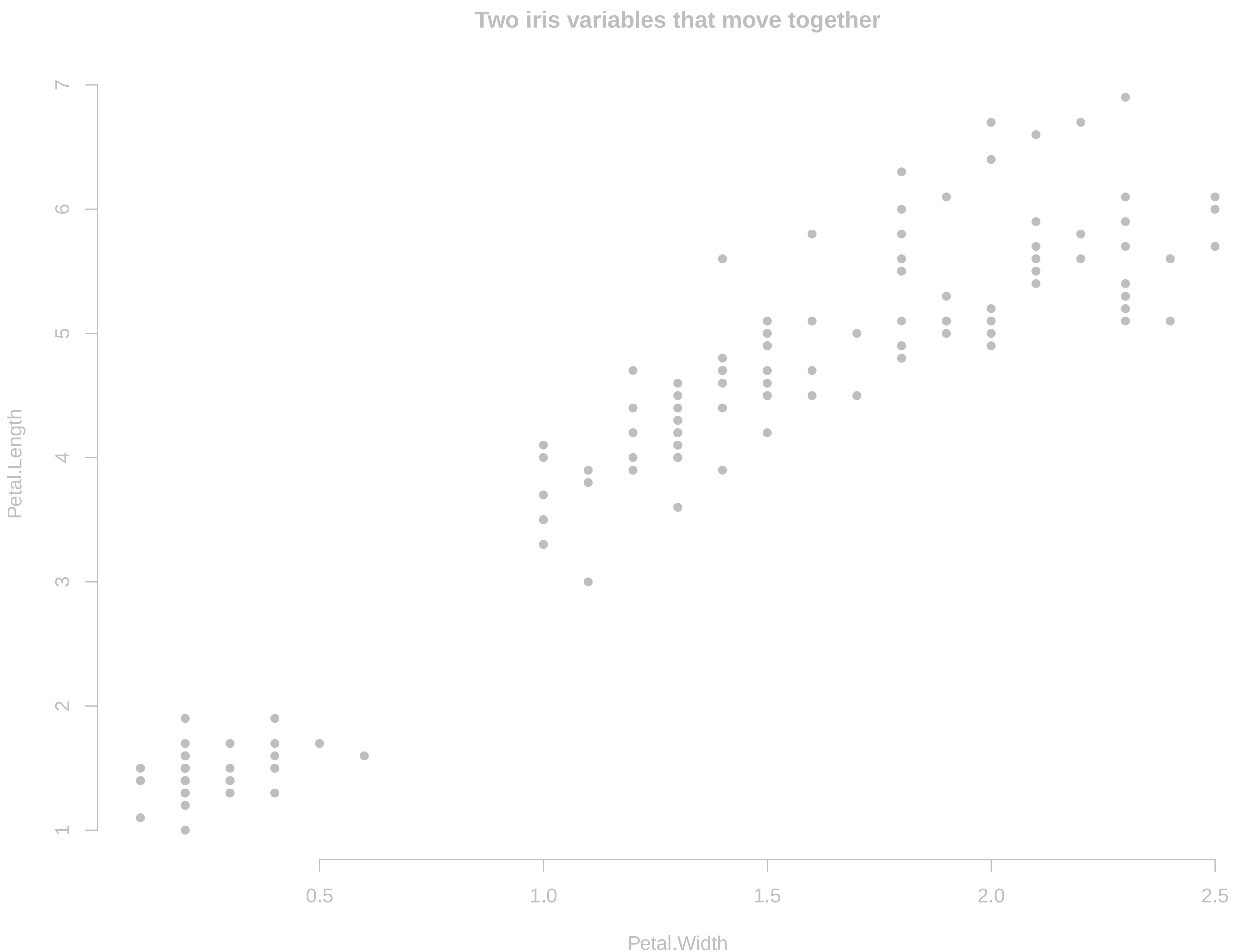
Add the blues together. (This is at a different scale.)

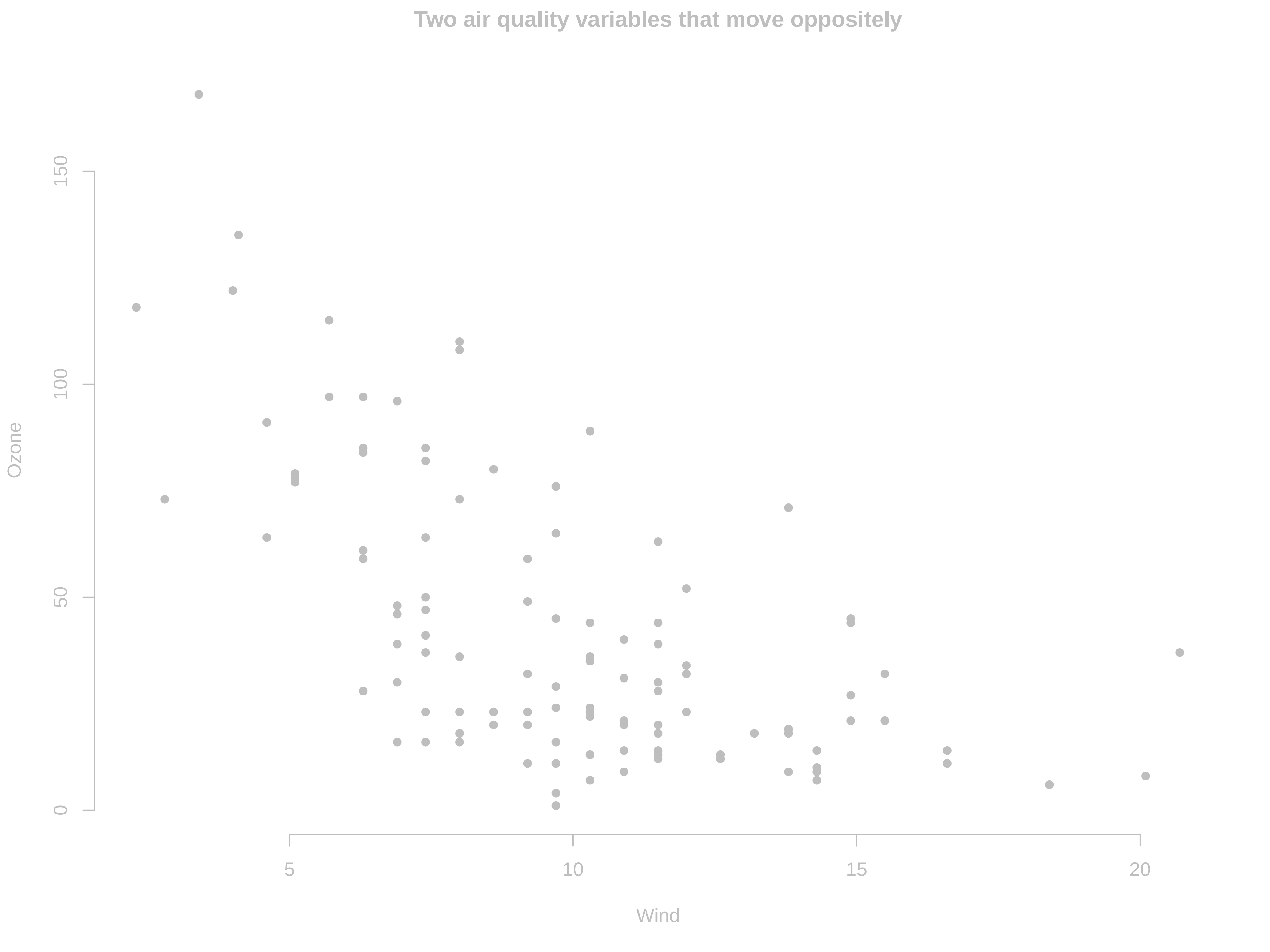
Add the reds together.

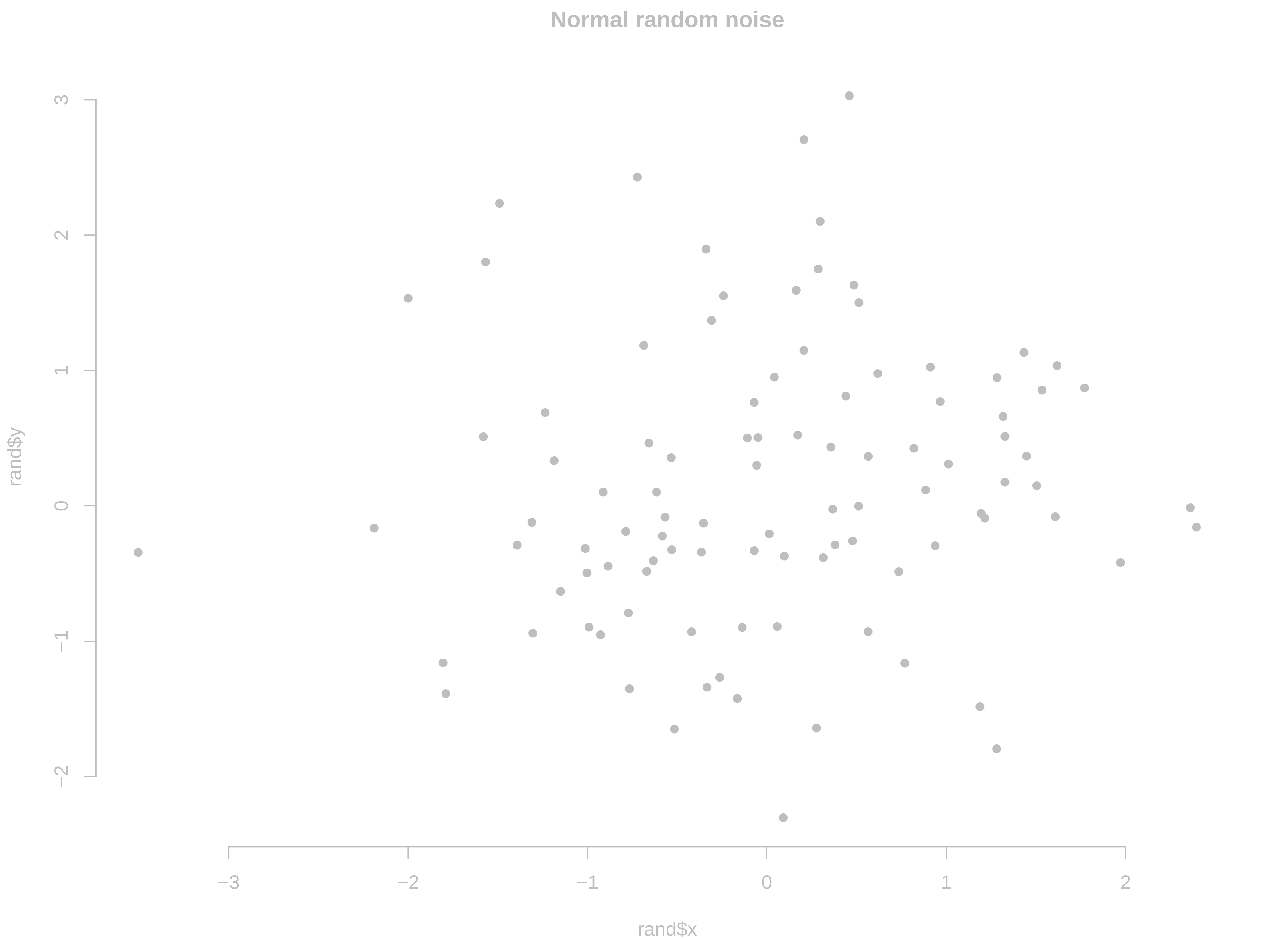
Subtract the reds.

(Covariance is zero.)

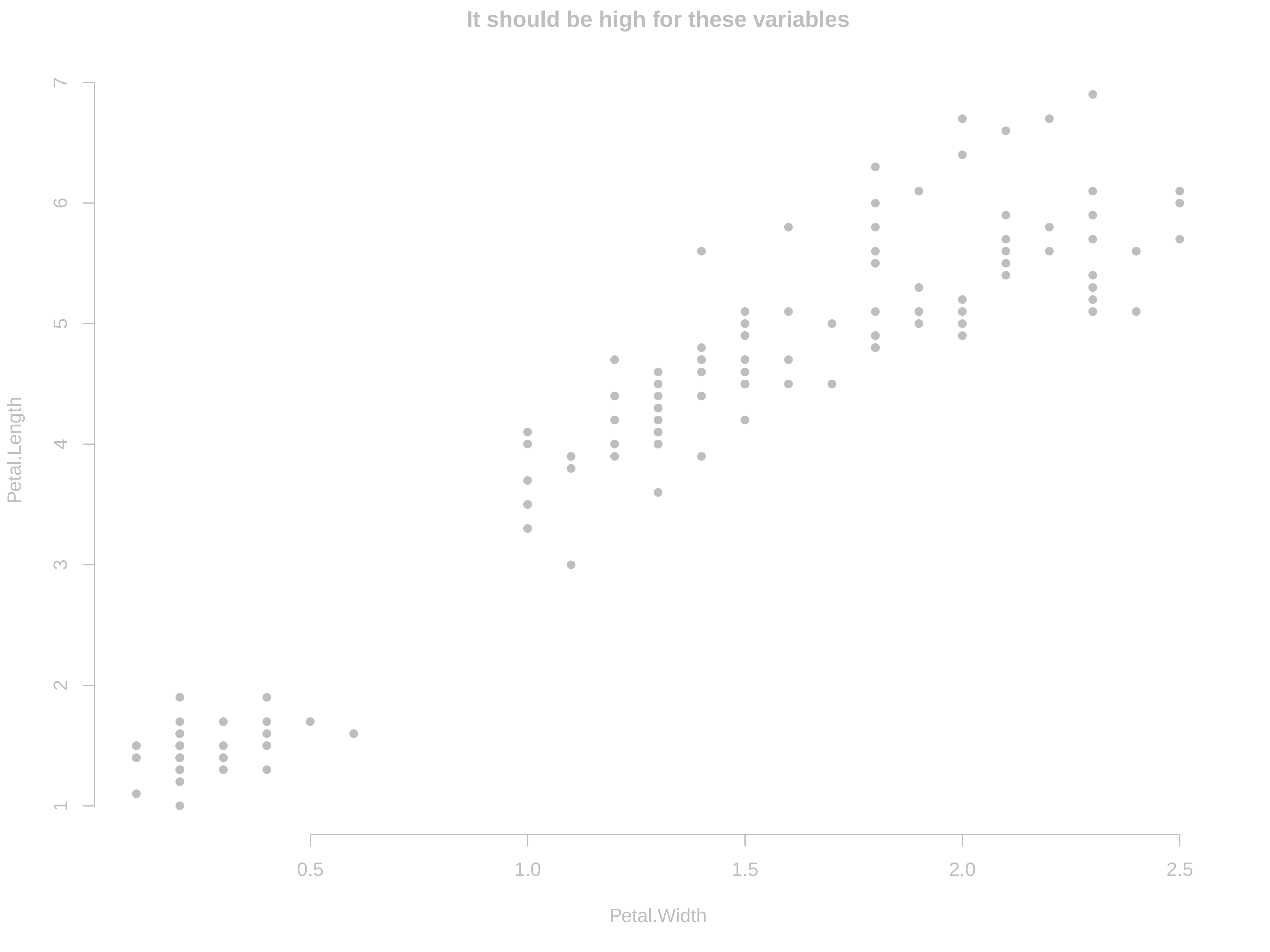
Let's review the previous slides quickly.

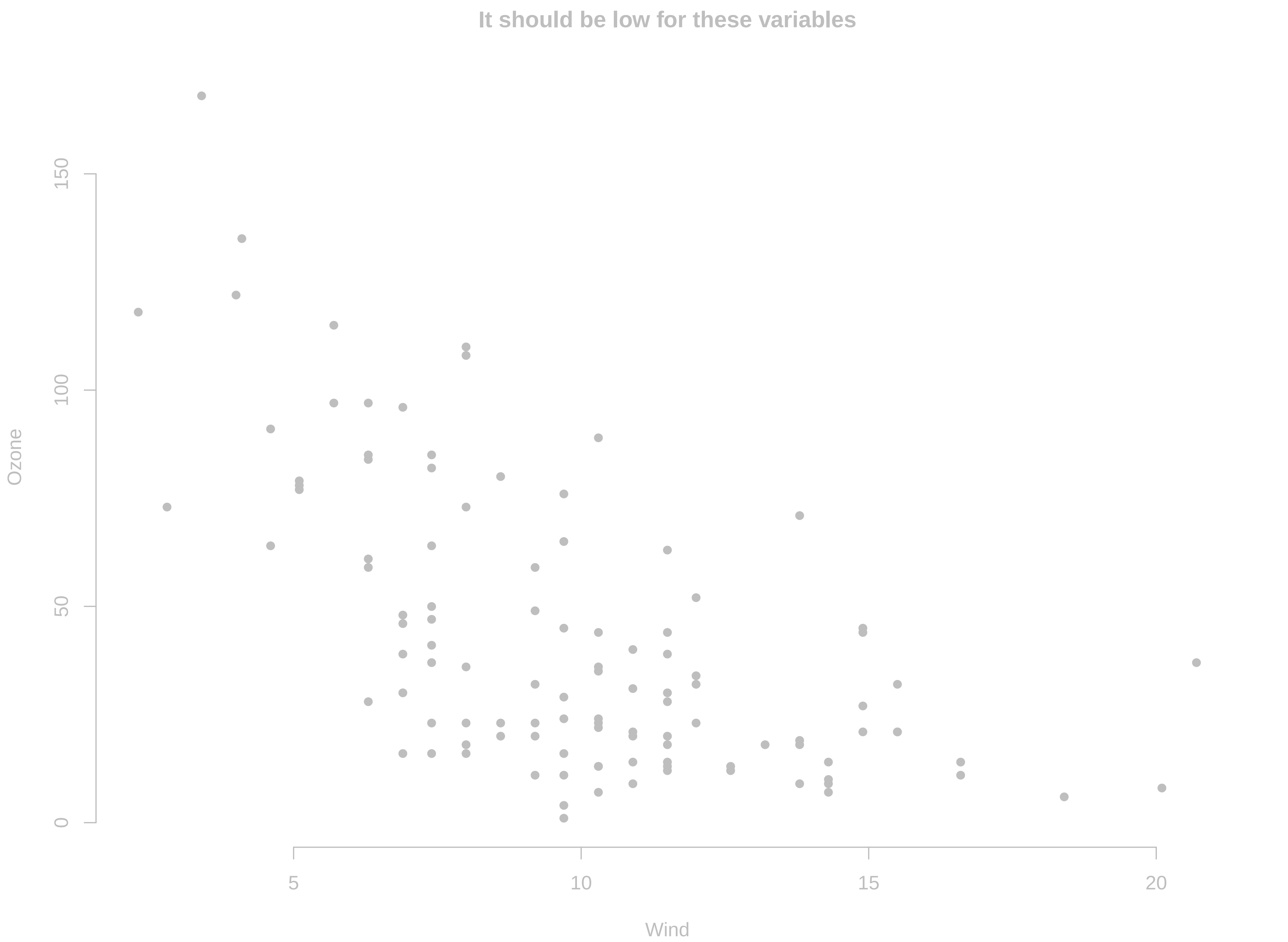

Two iris variables that move together
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width

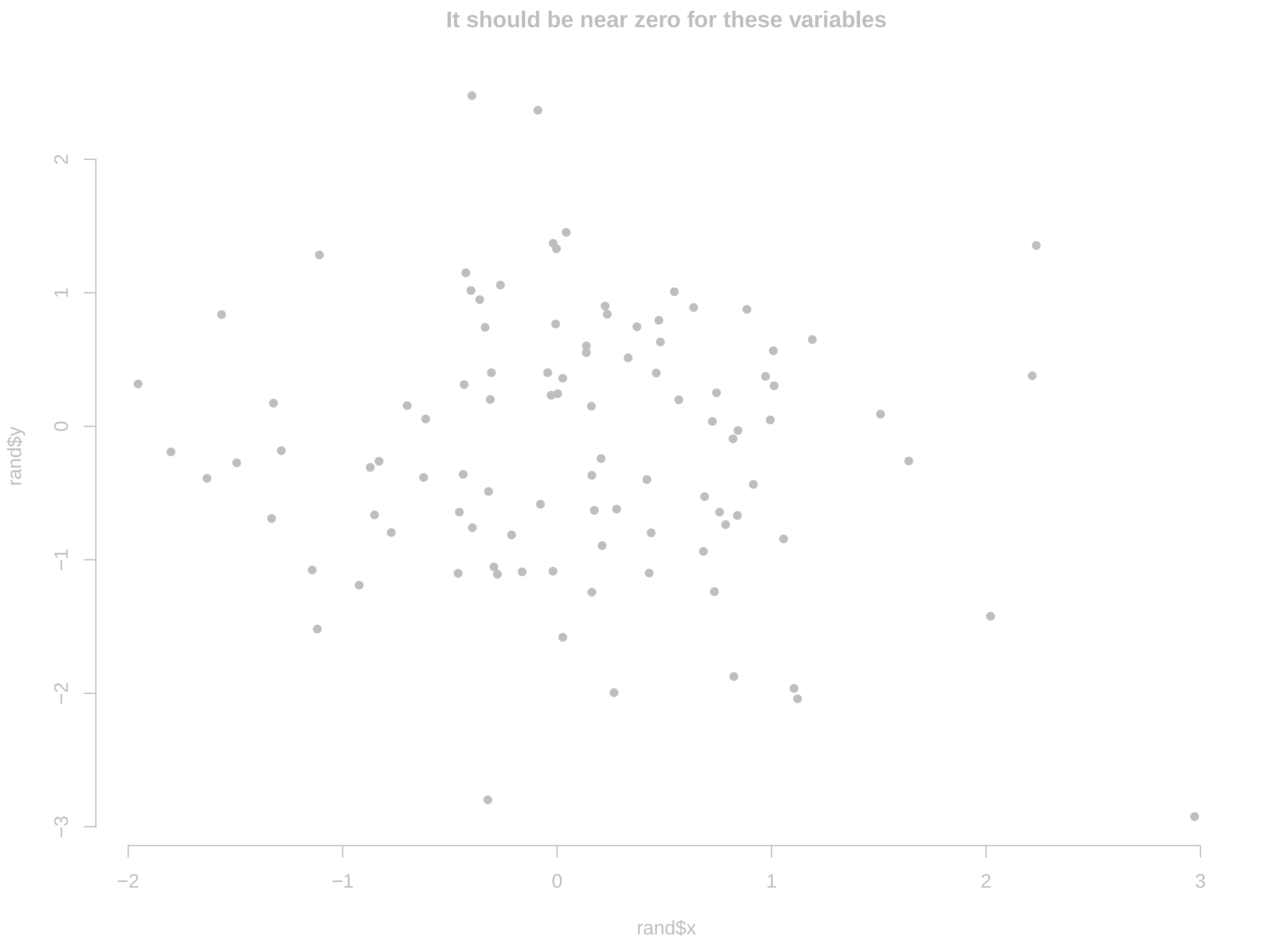

Two air quality variables that move oppositely
Ozone
0
50
100
150
Wind
5
10
15
20


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

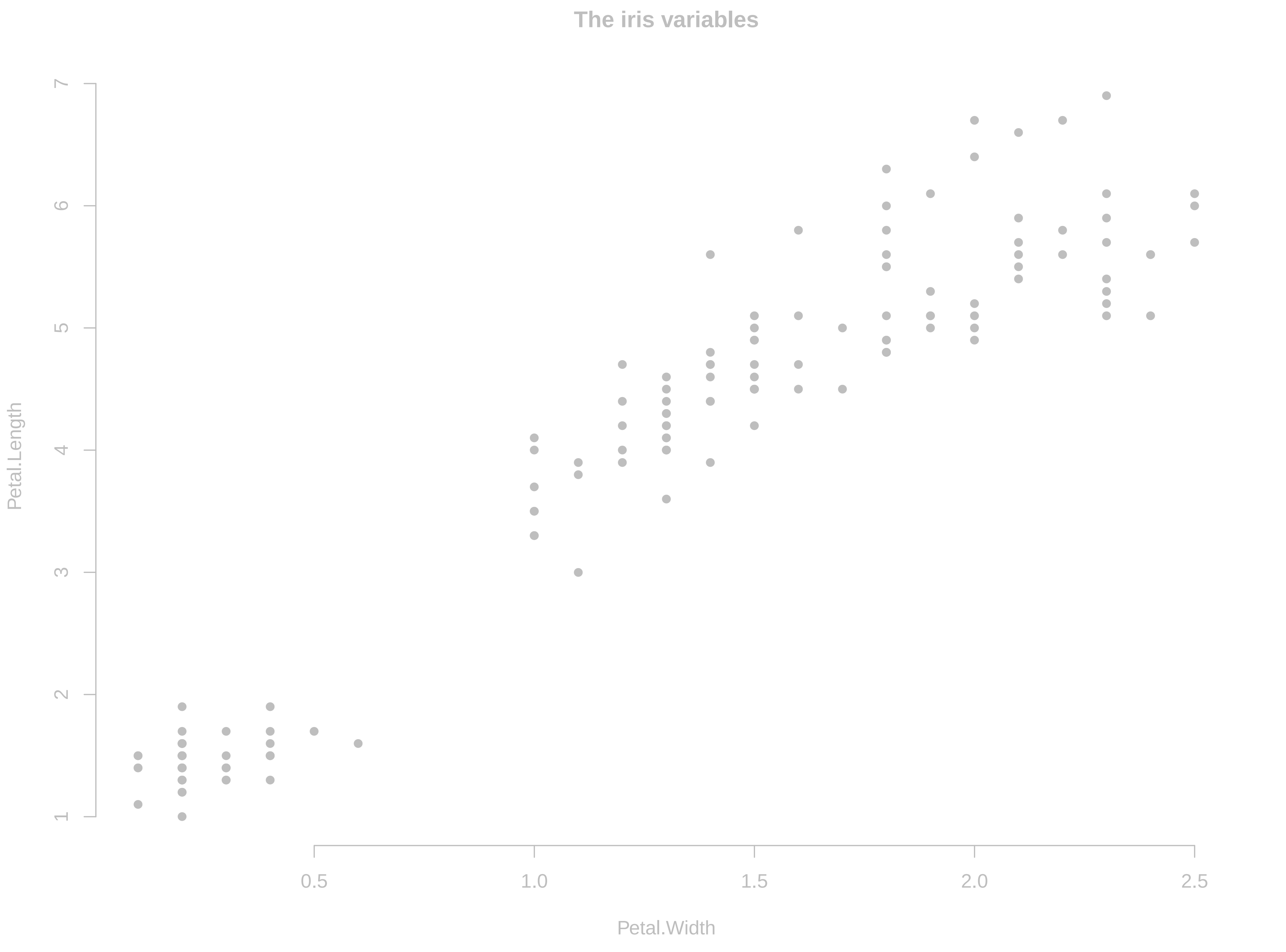
We want a number that describes whether two variables move together.


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

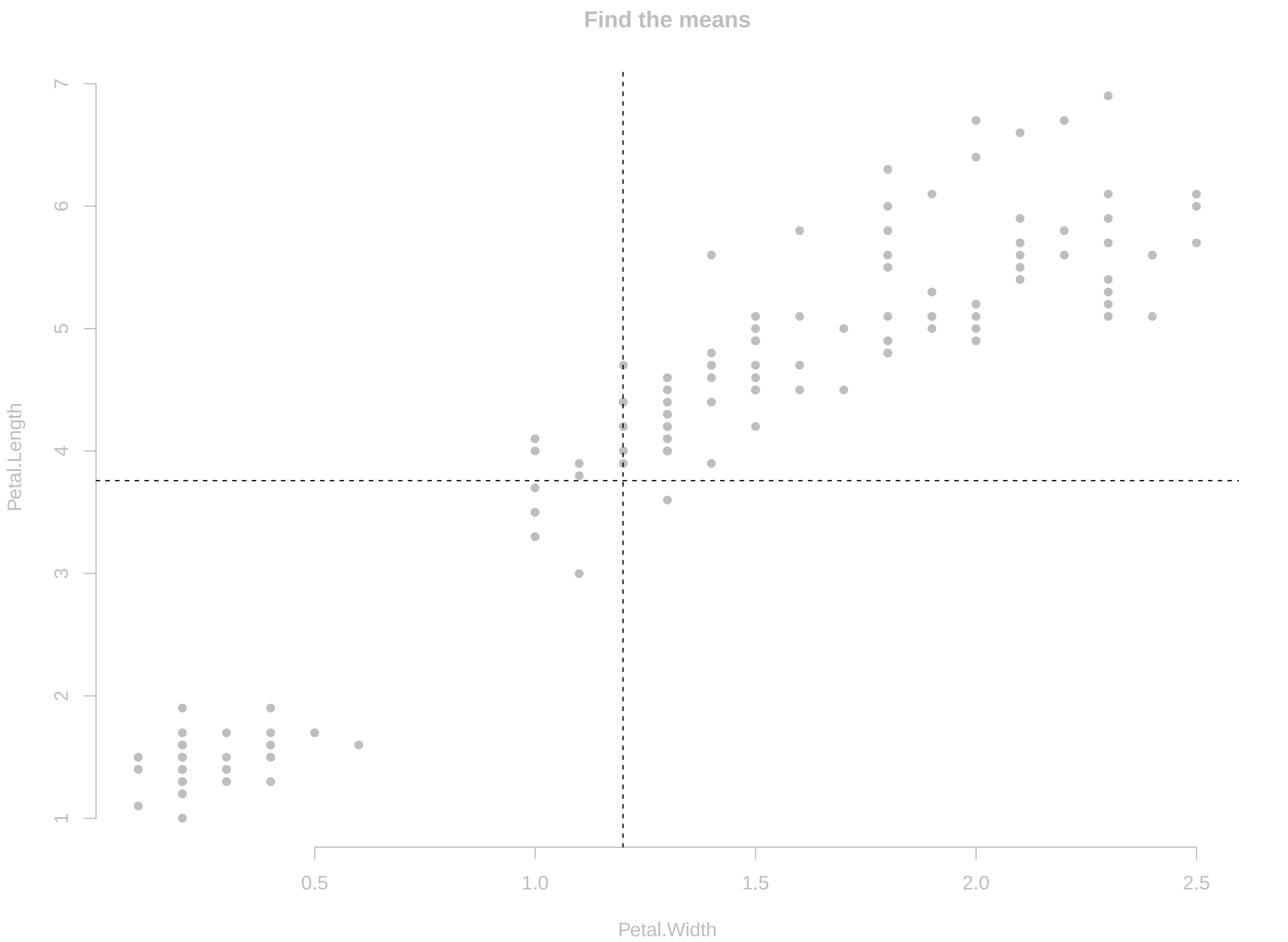

It should be low for these variables
Ozone
0
50
100
150
Wind
5
10
15
20


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

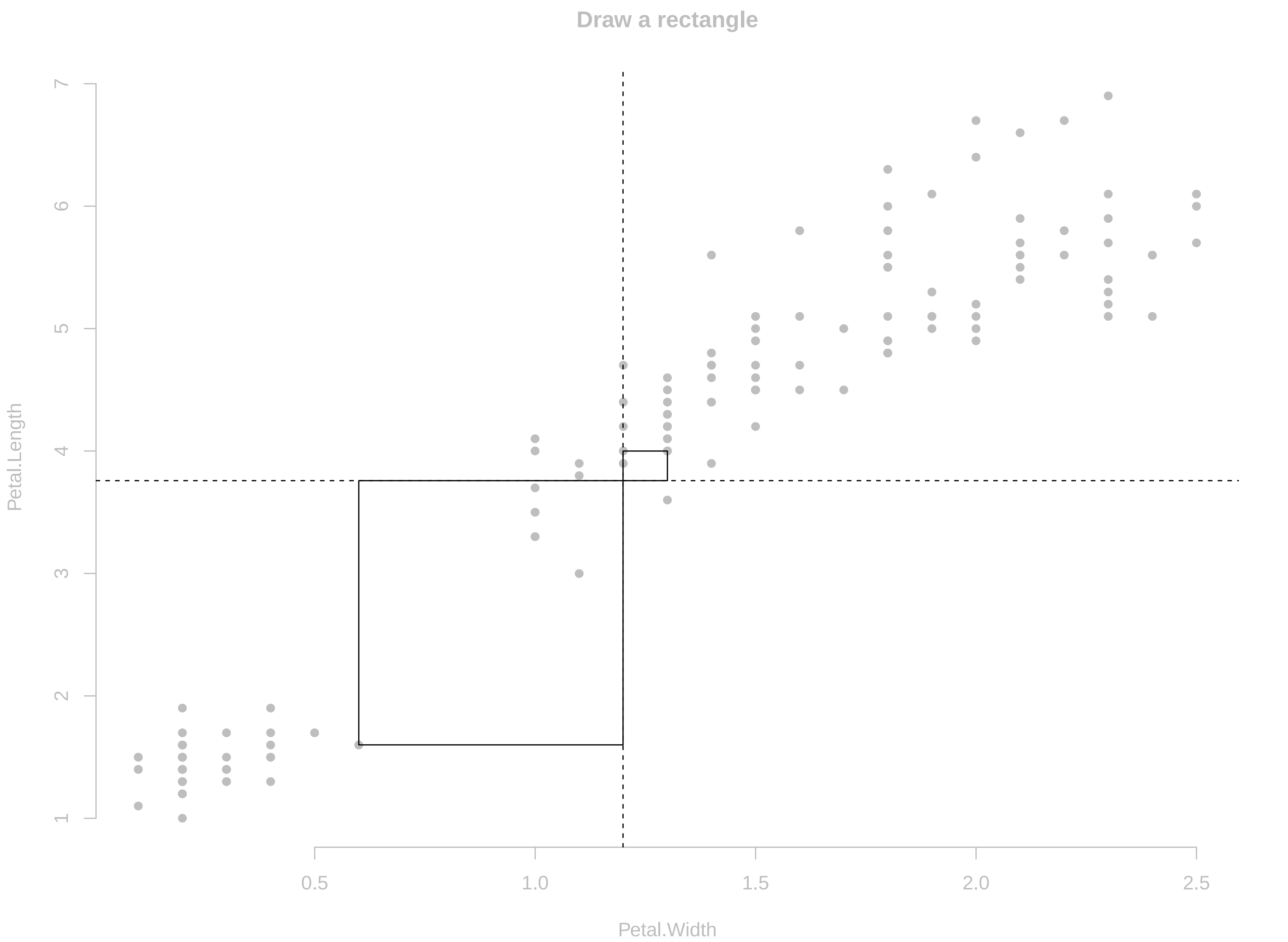
Covariance



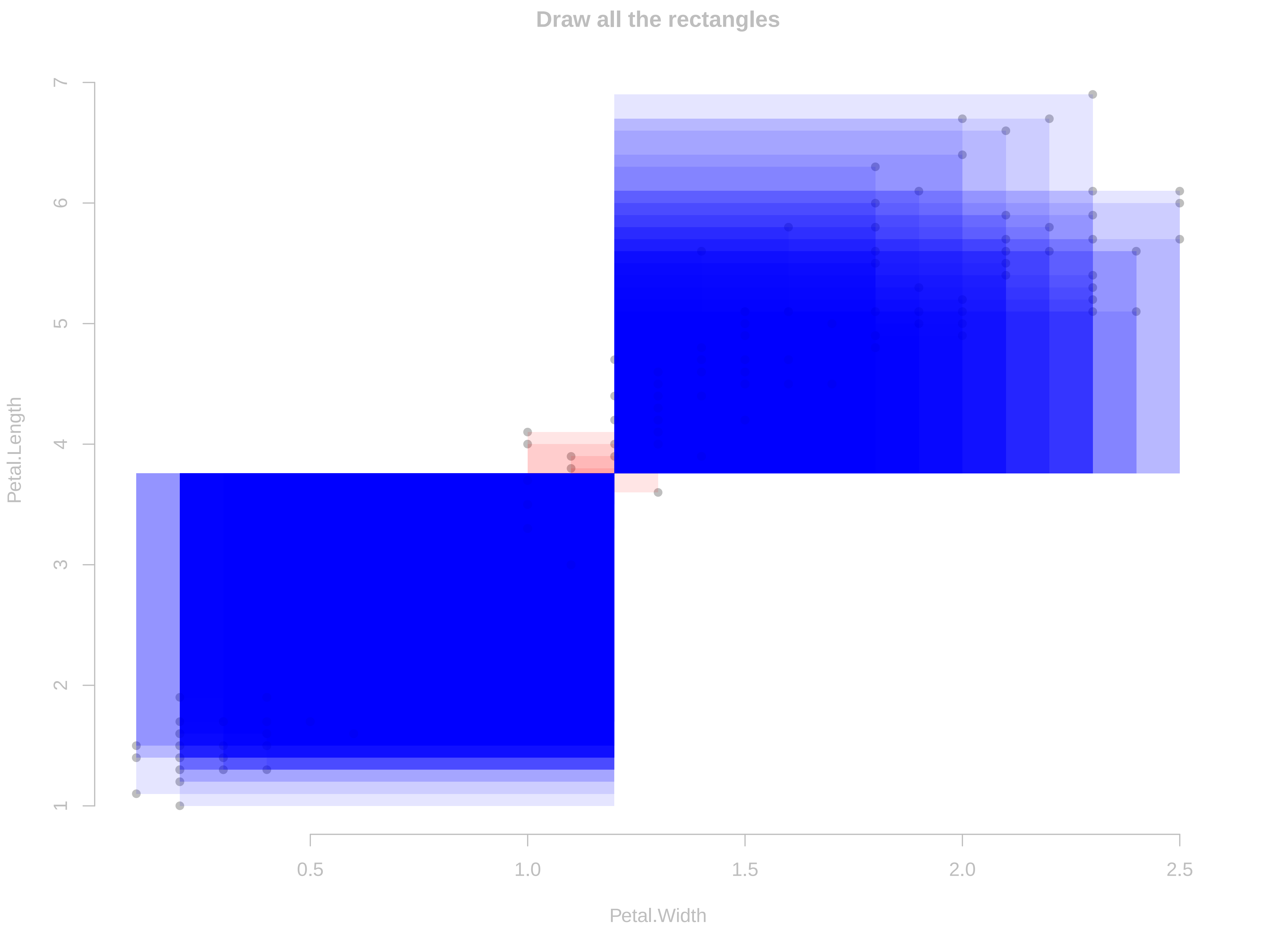
The iris variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5



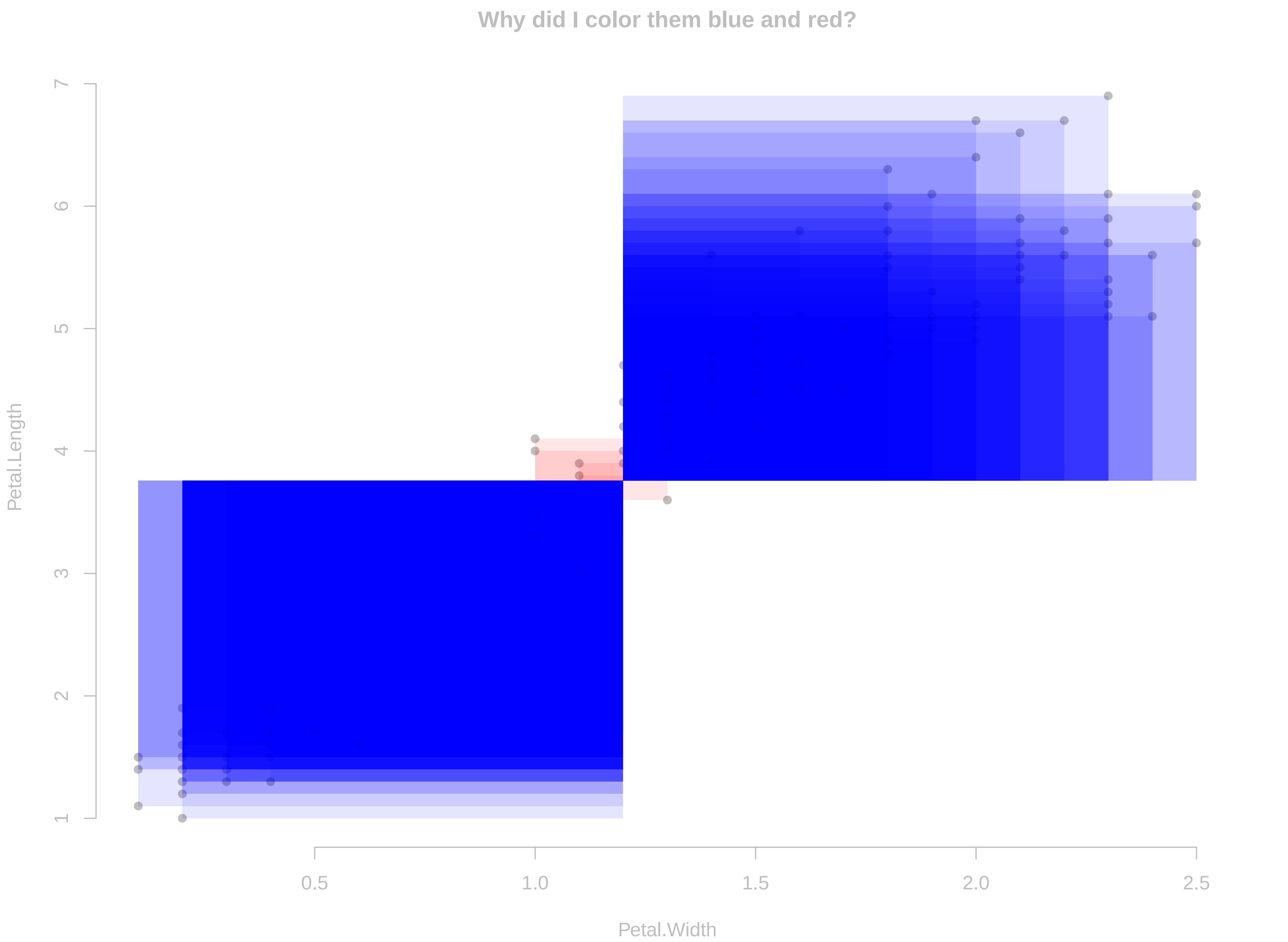
Find the means
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

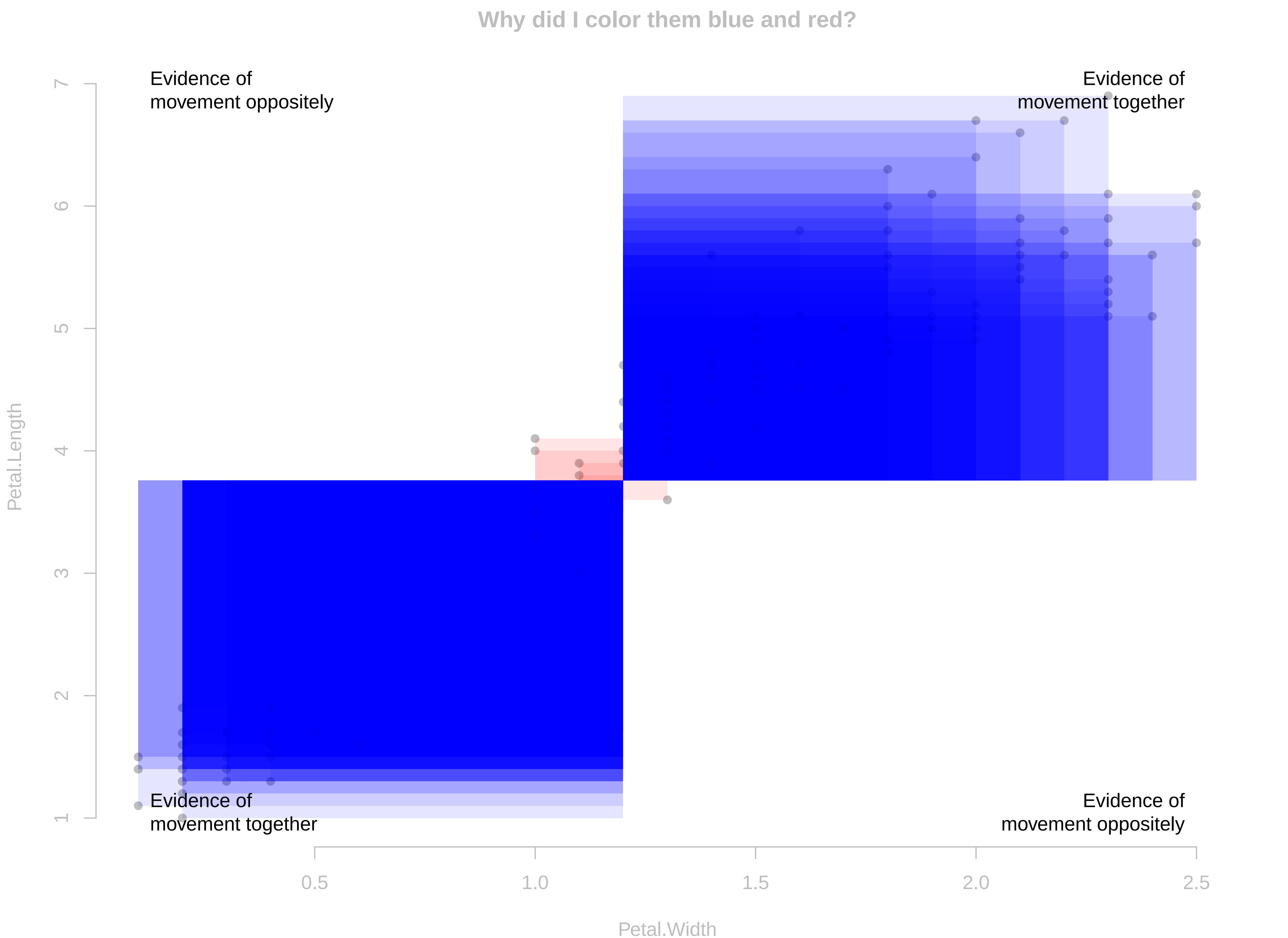


Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

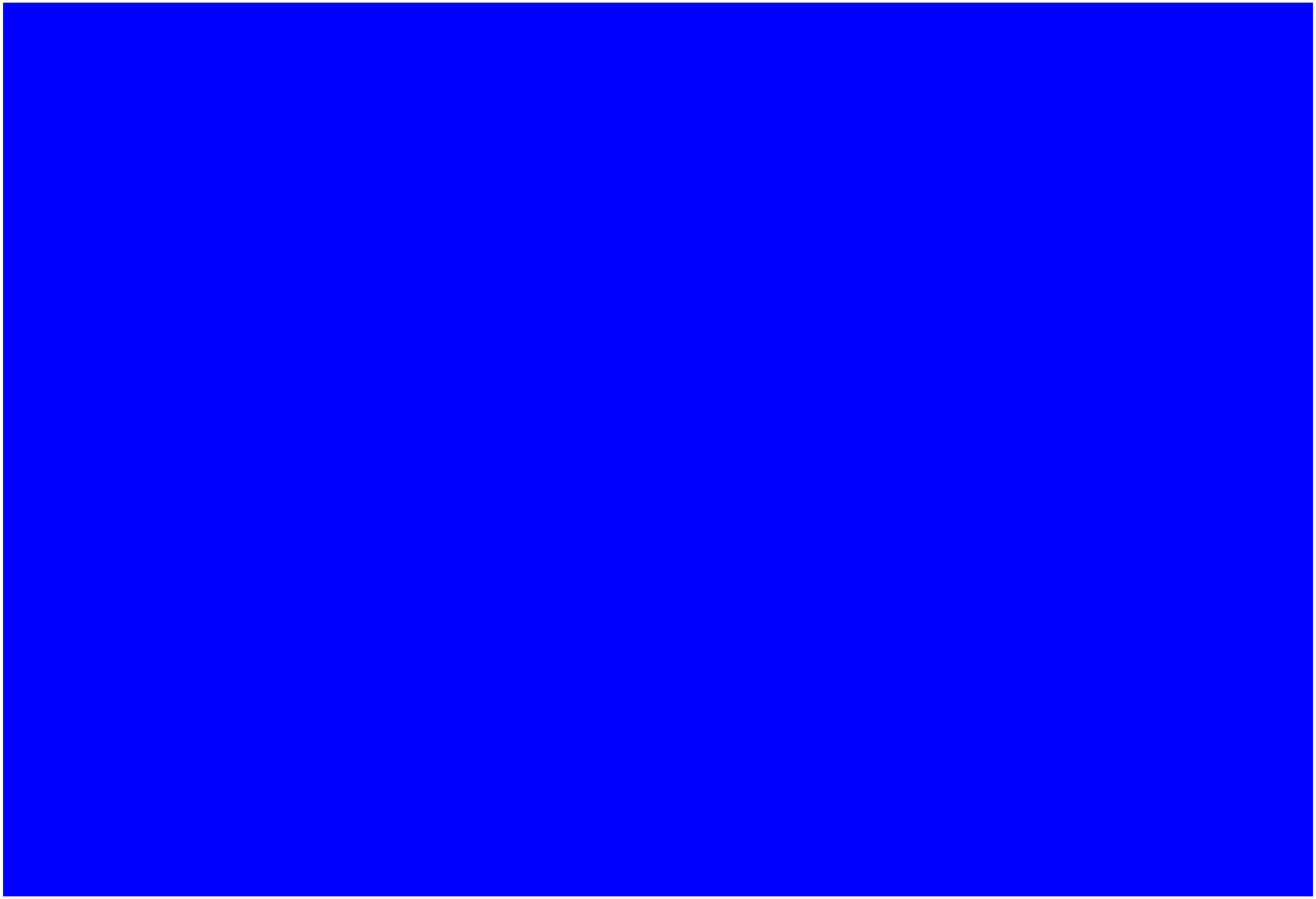


Draw all the rectangles
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

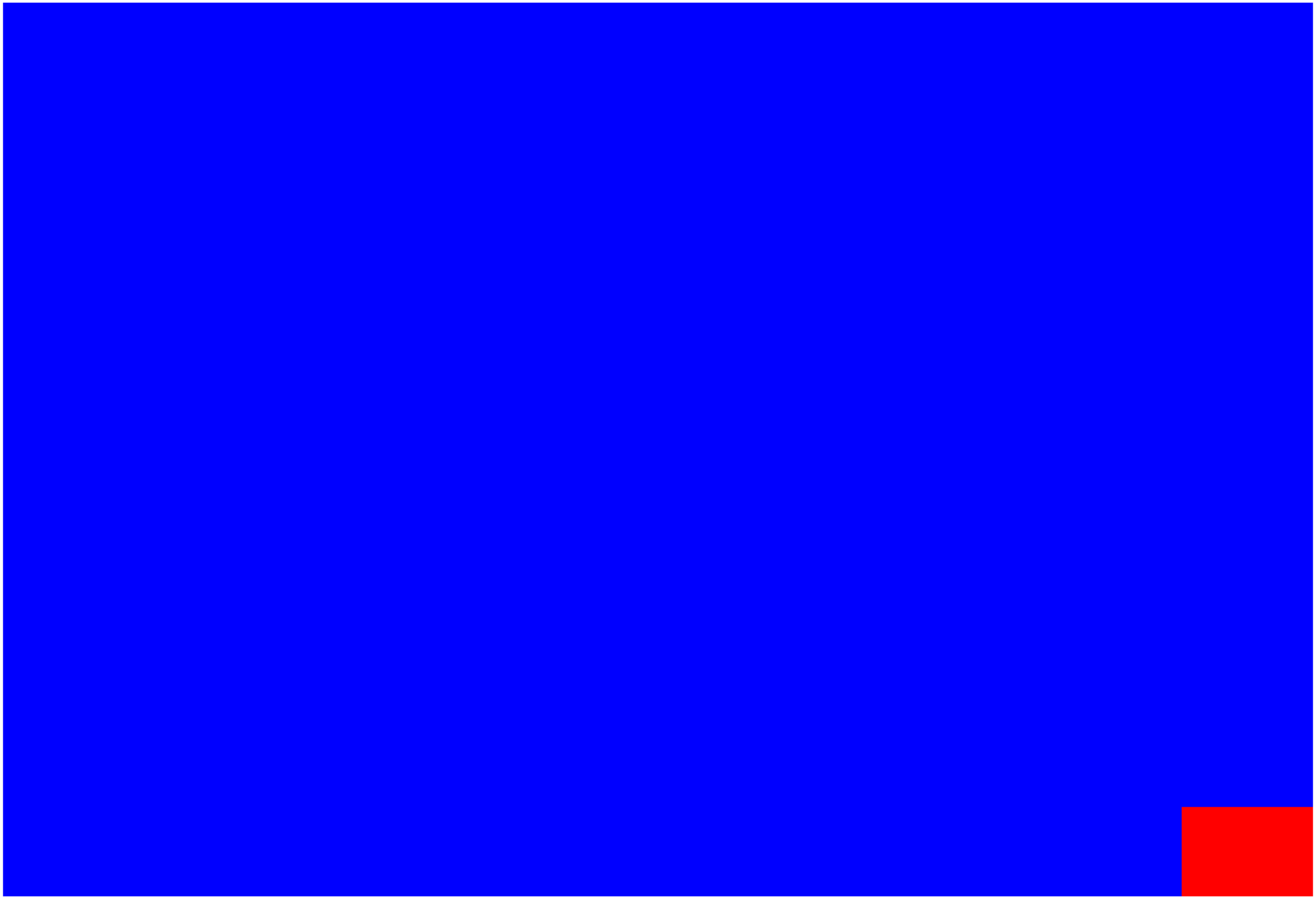

Why did I color them blue and red?
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


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
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

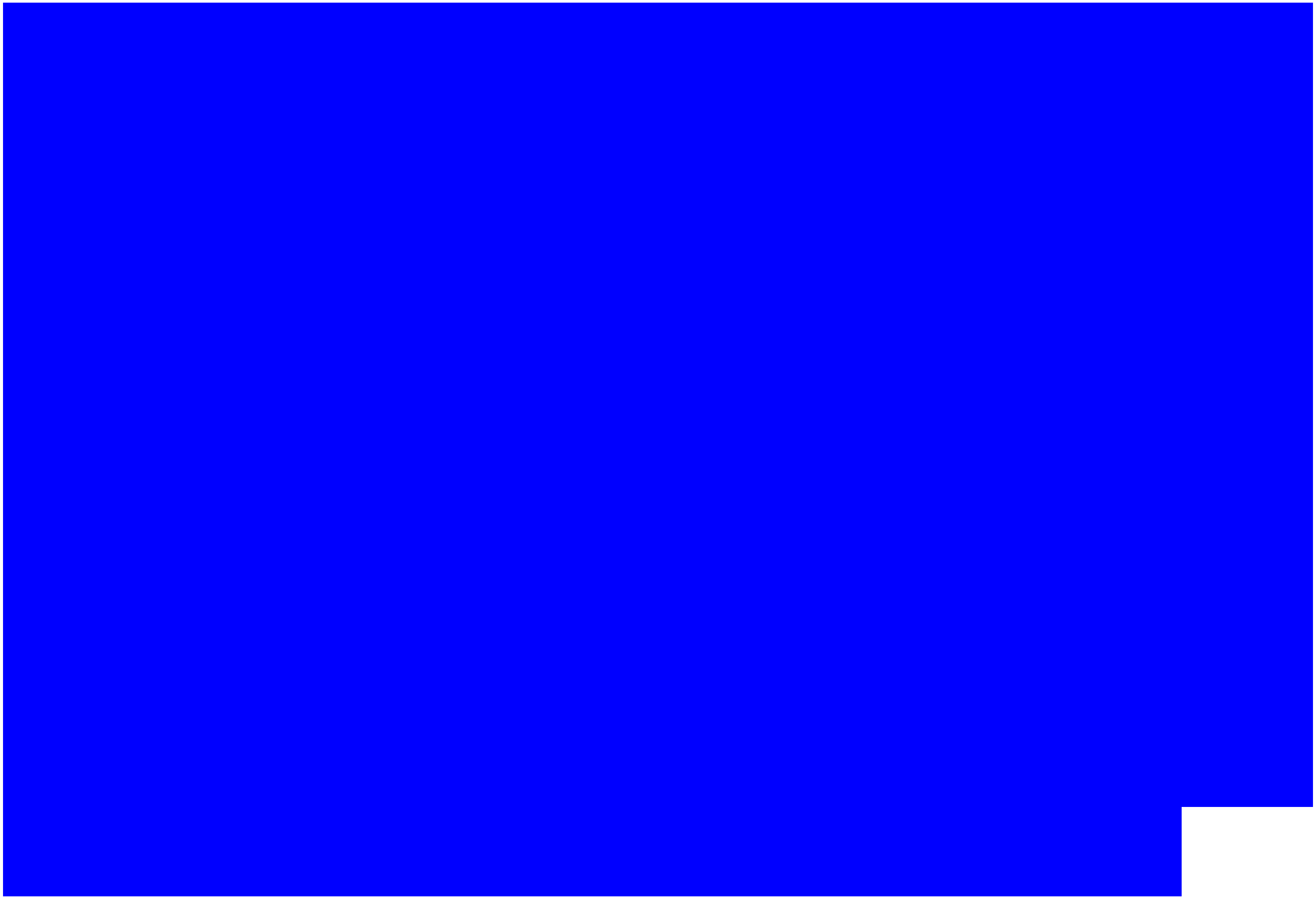
Add the blues together. (This is at a different scale.)



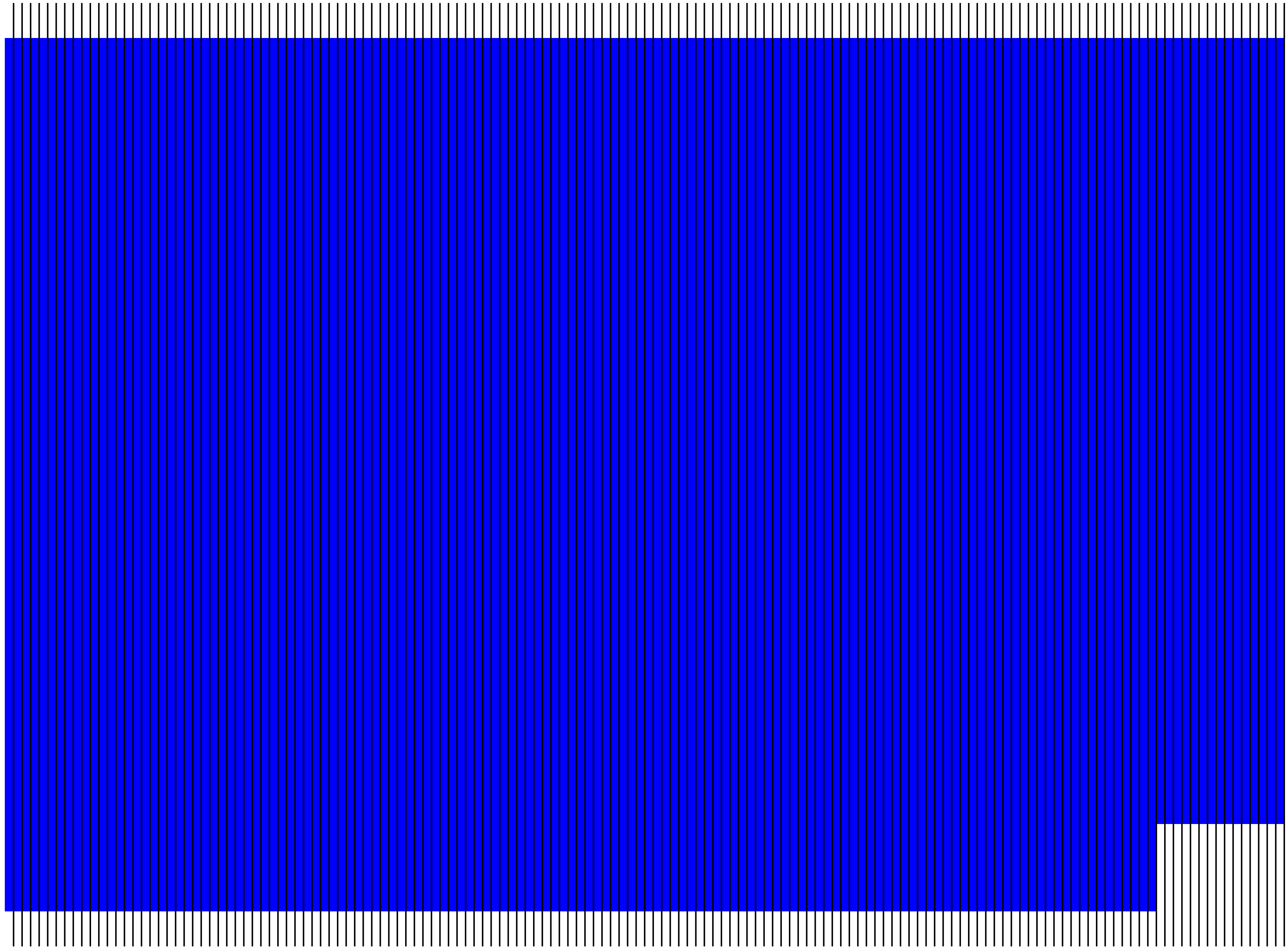
Add the reds together.



Subtract the reds.

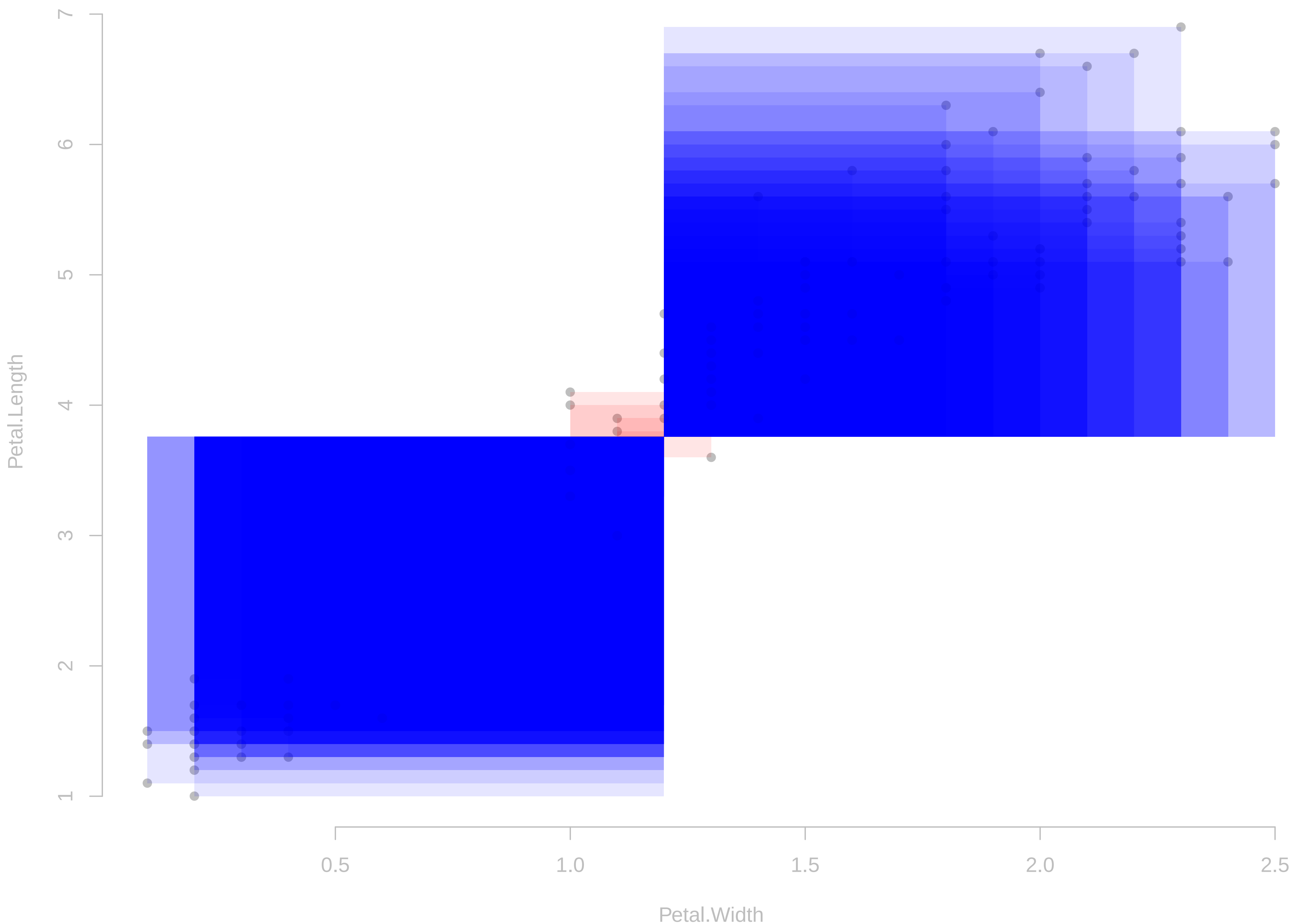


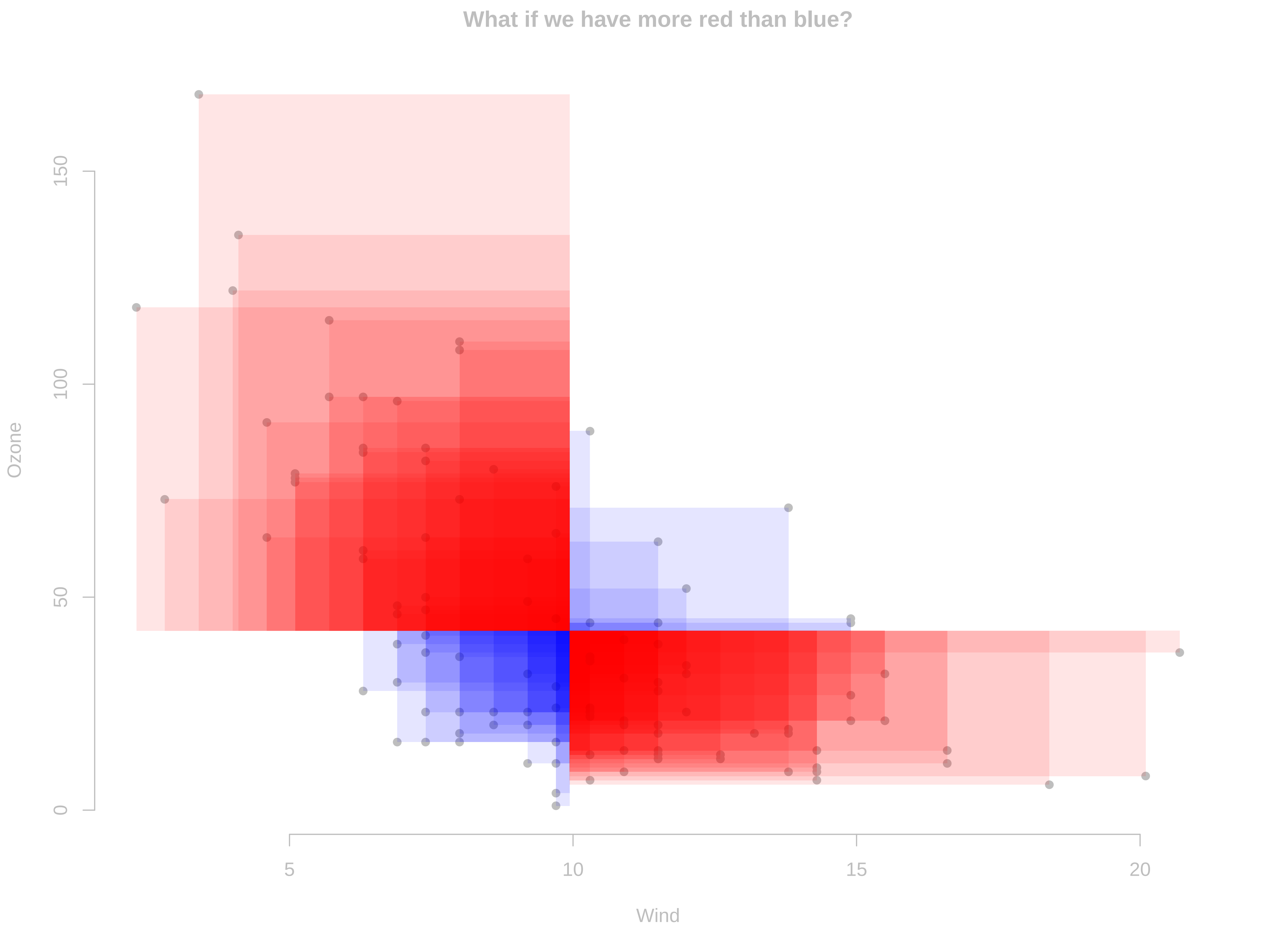
Divide into as many equal pieces as we have irises (n).



This blue sliver is the covariance.

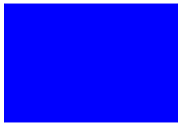
That was for this sort of relationship.



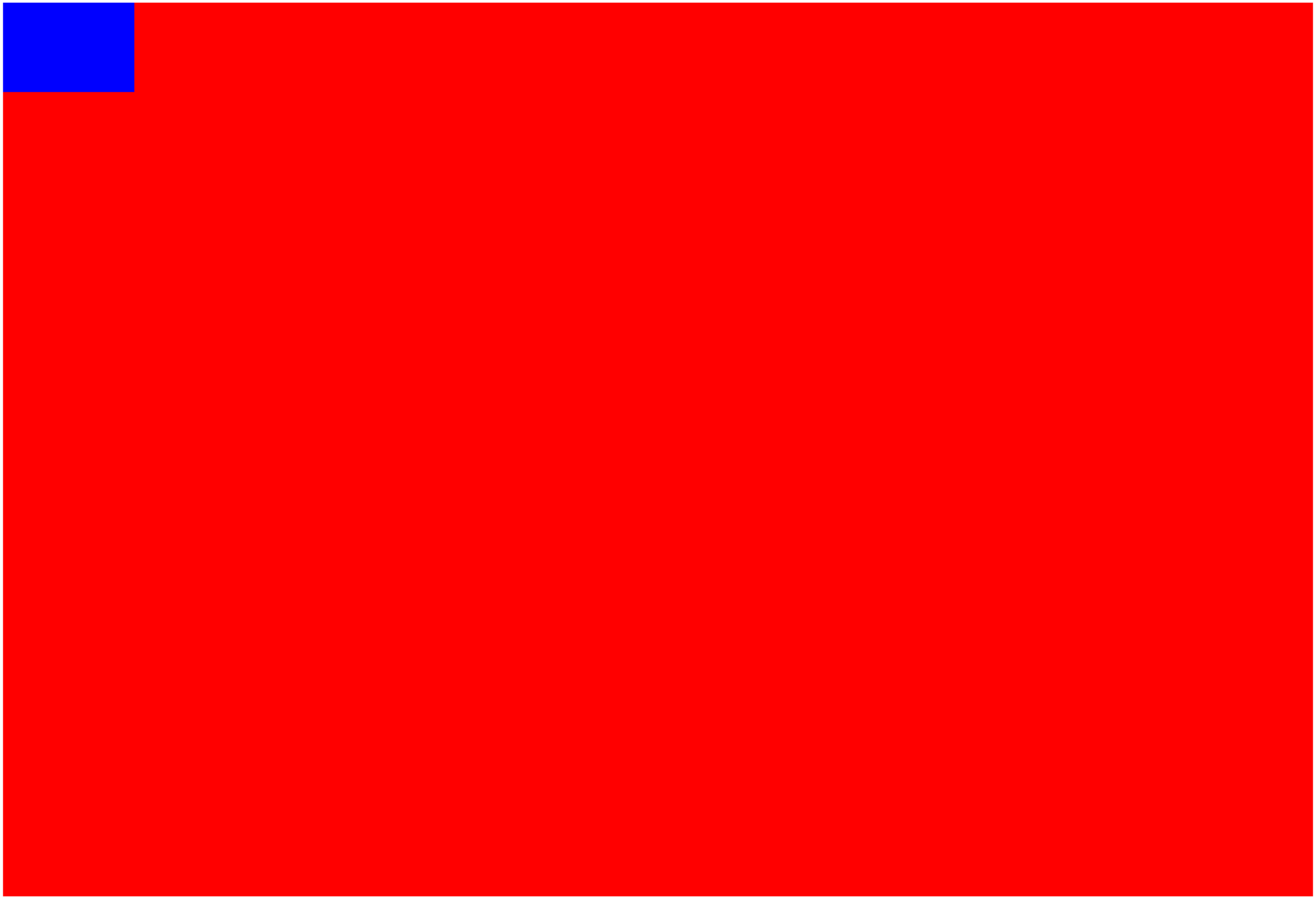


What if we have more red than blue?
Ozone
150
100
50
0
5
10
15
20
Wind

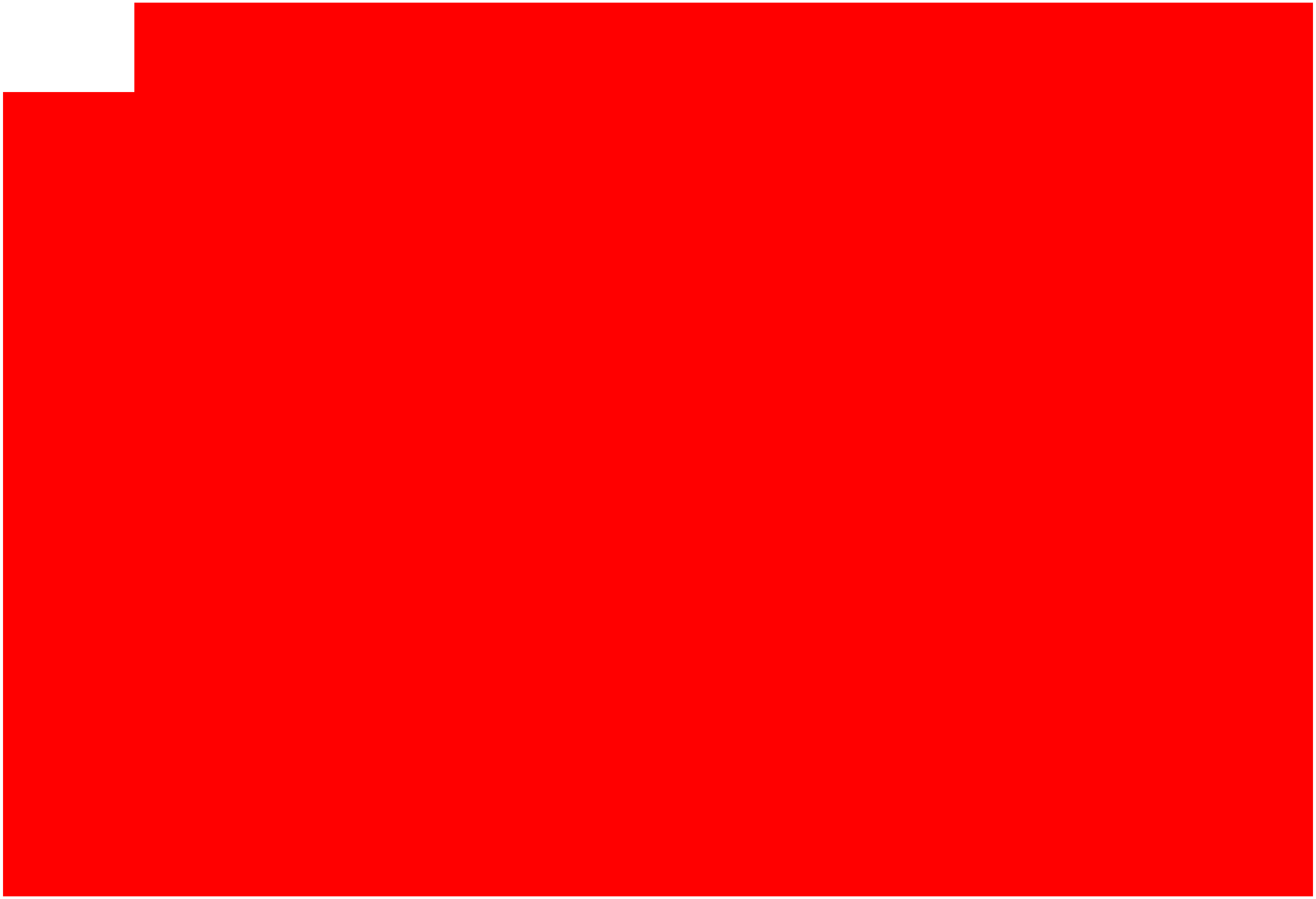
Add the blues together. (This is at a different scale.)



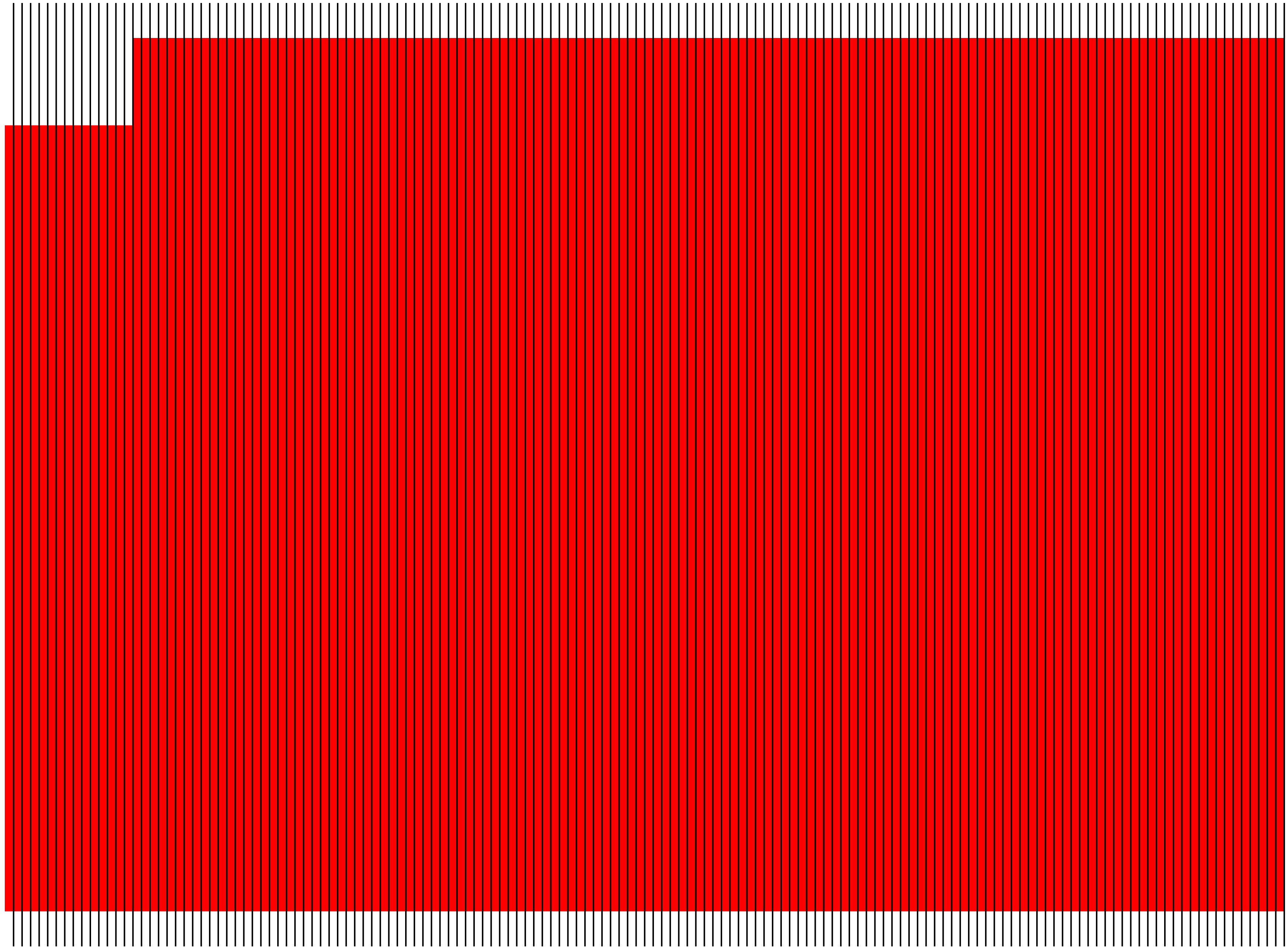
Add the reds together.



Subtract the reds.



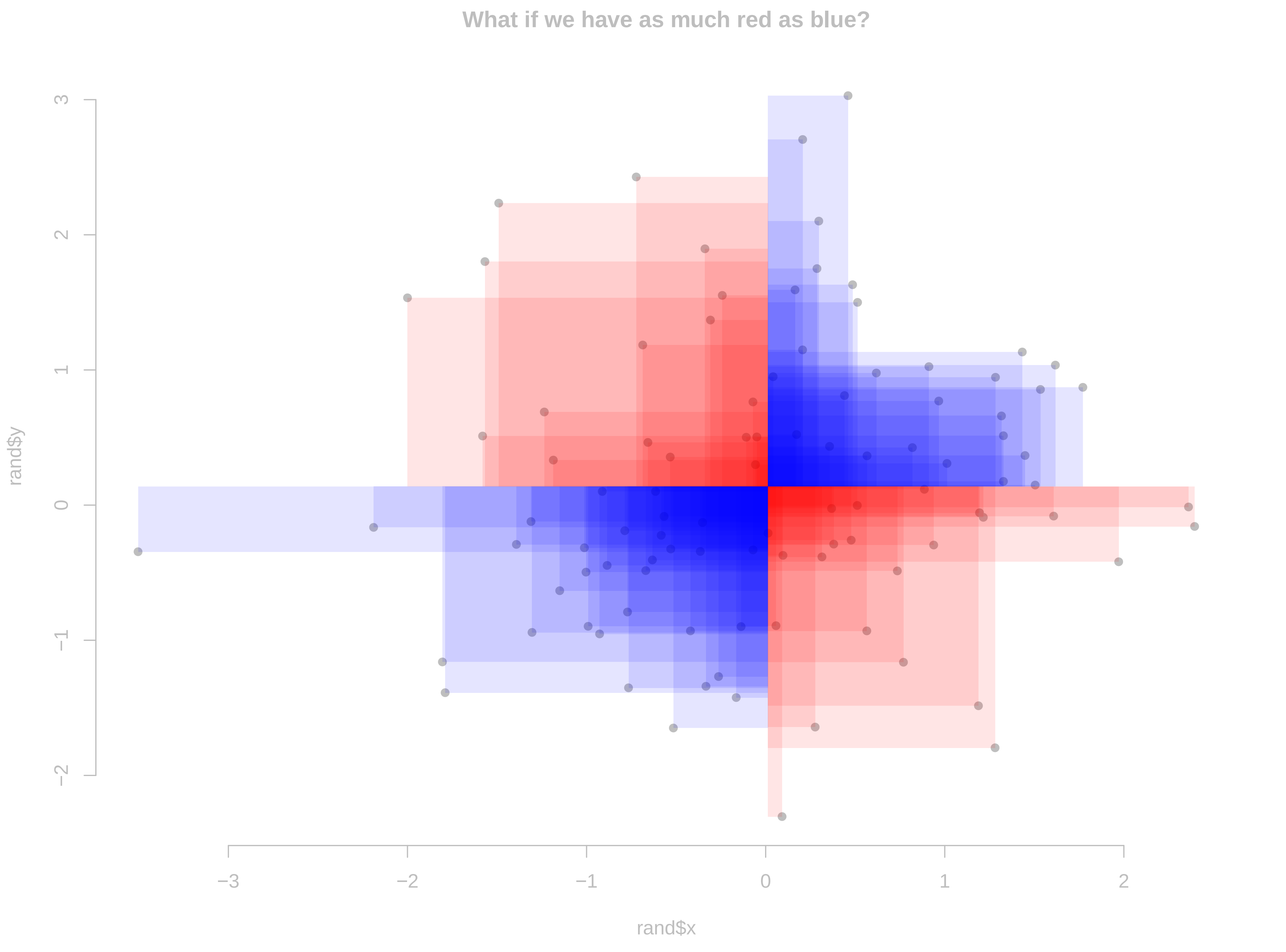
Divide into as many equal pieces as we have irises (n).



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$y
3
2
1
0
−1
−2
rand$x
−3
−2
−1
0
1
2

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

Add the reds together.

Subtract the reds.

(Covariance is zero.)

Variance

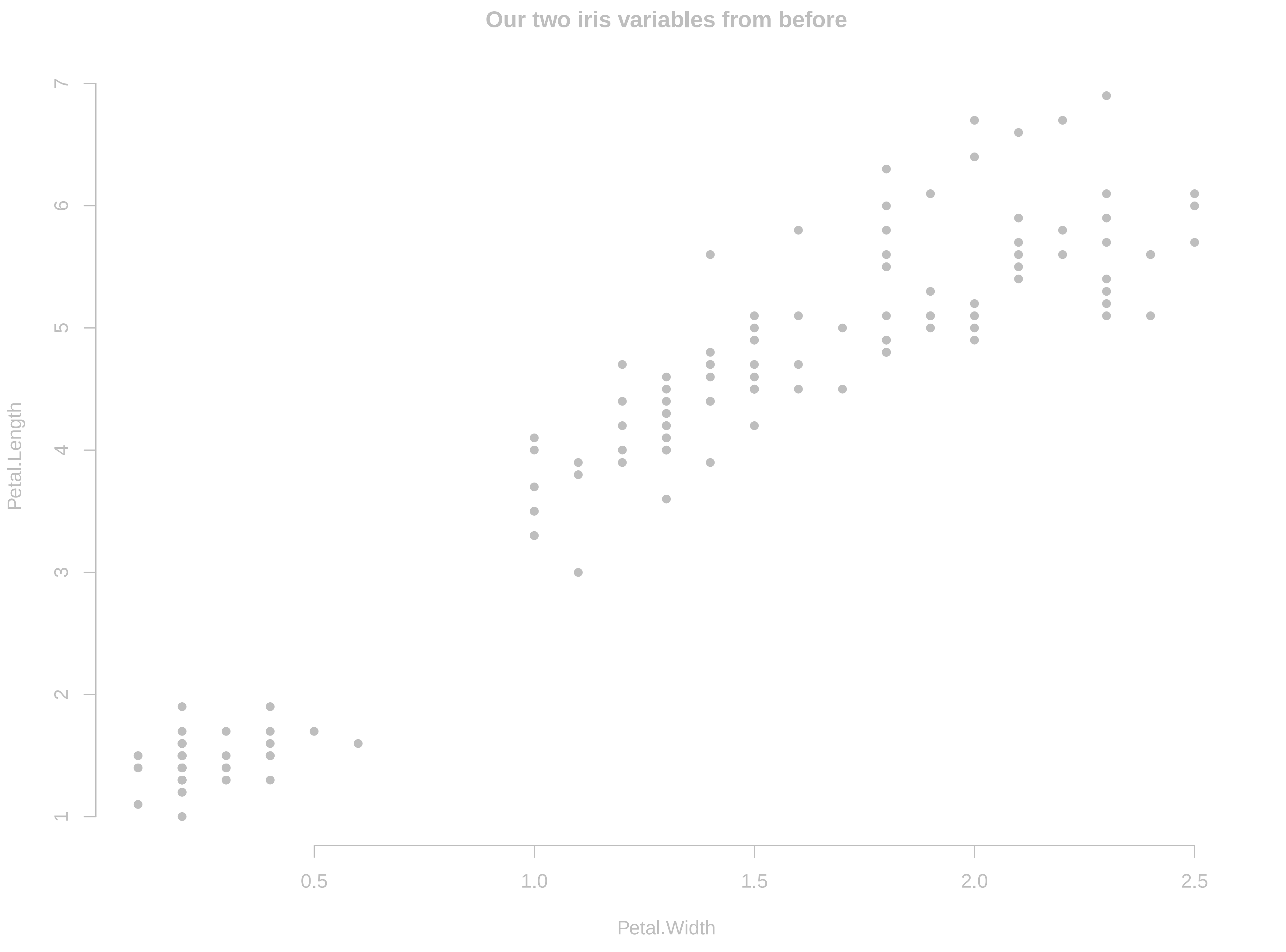
Variance tells us how spread out some numbers are.

1, 4, 8, 10

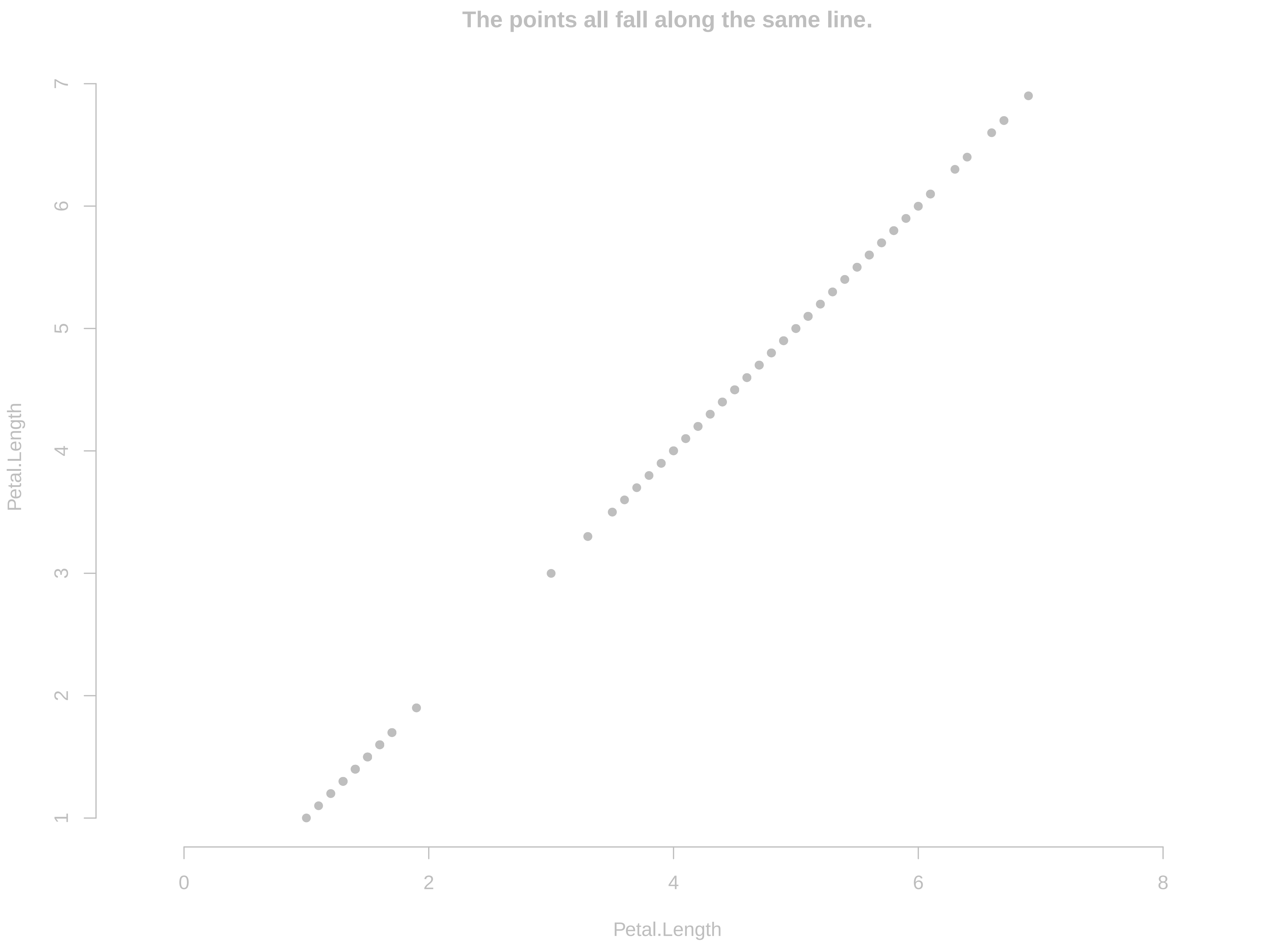
vs

4, 4, 5, 6

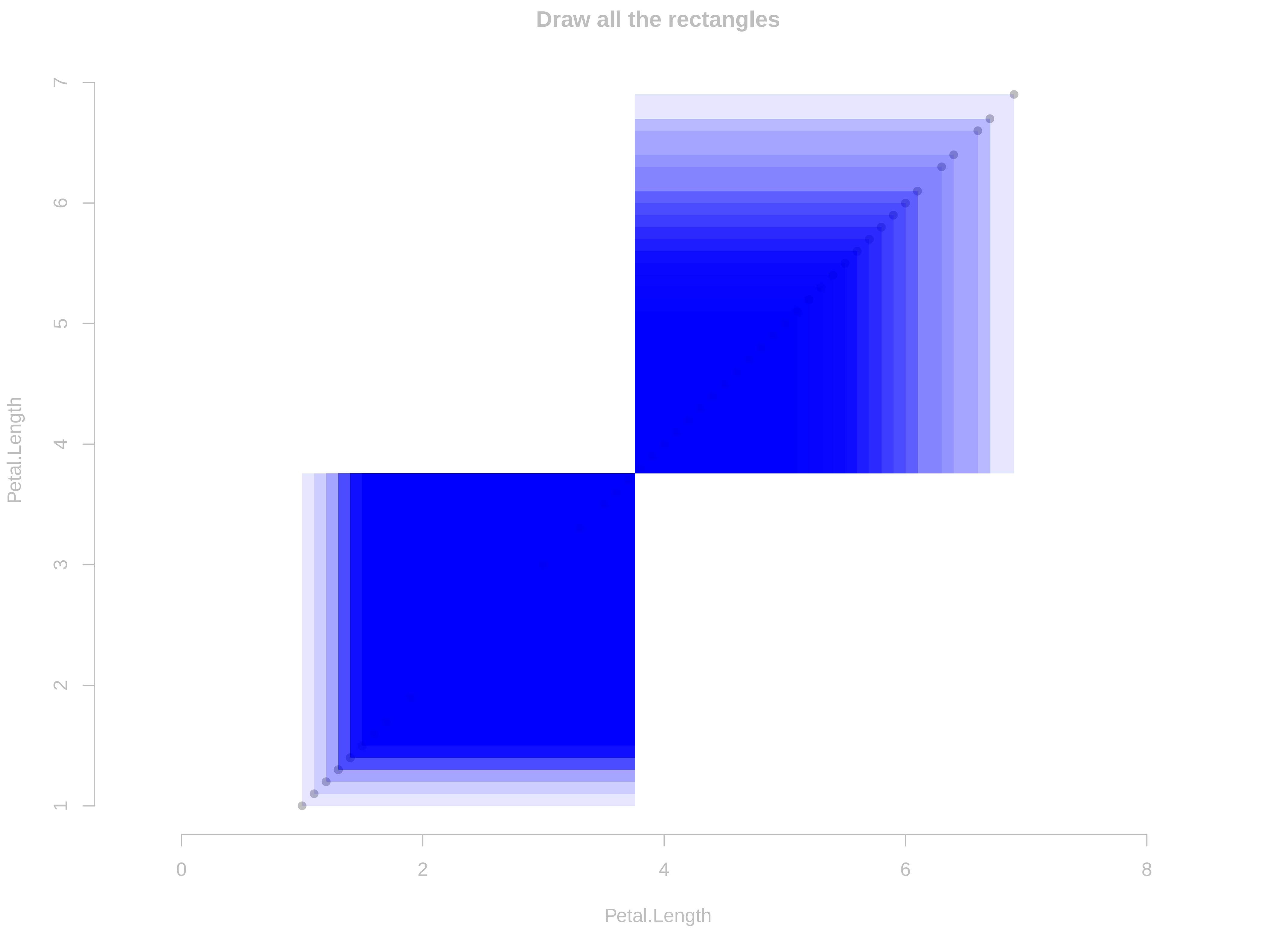
The variance of a variable is the covariance of the variable with itself.

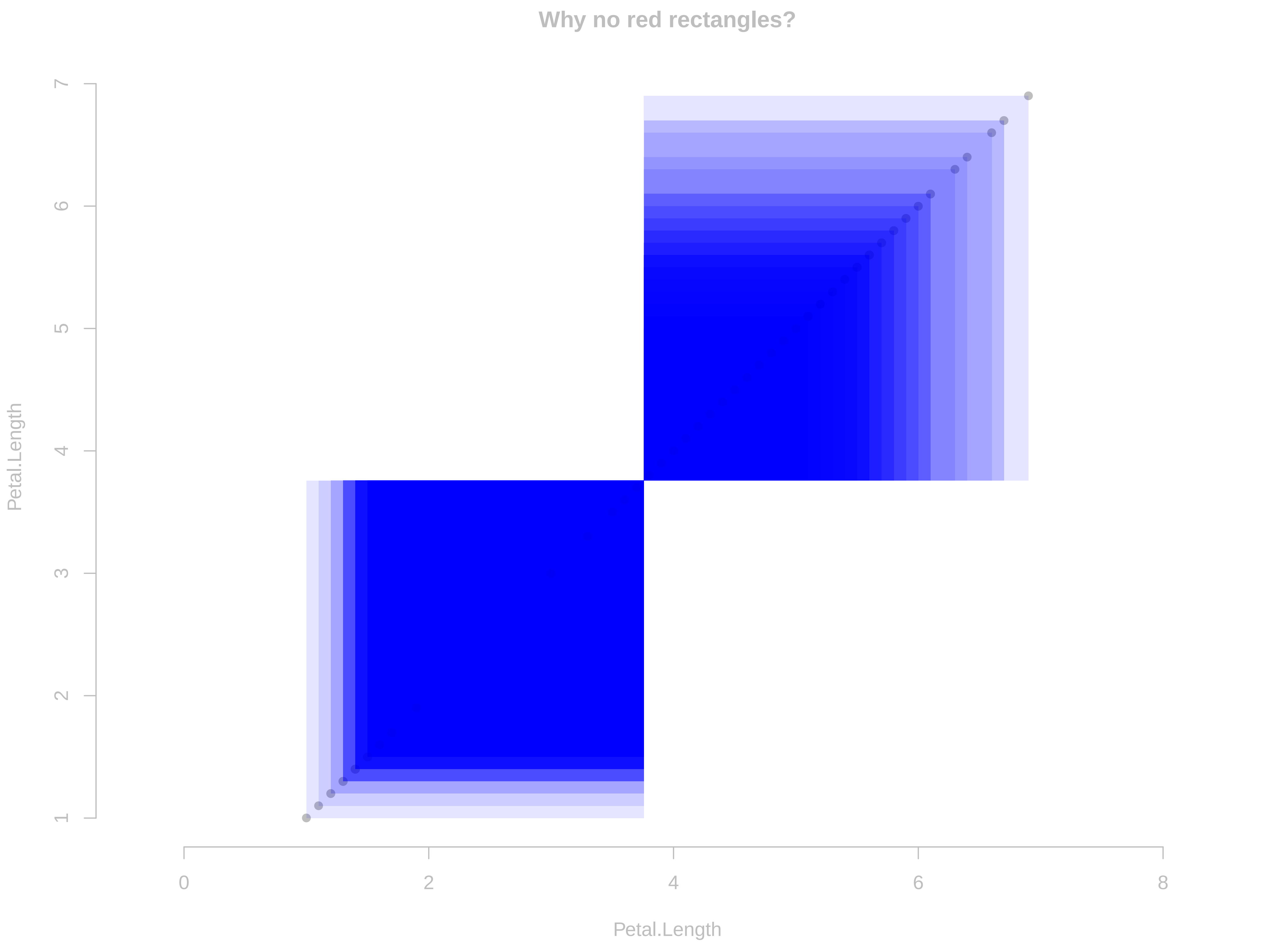

Our two iris variables from before
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

Let's look at just one of them.


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

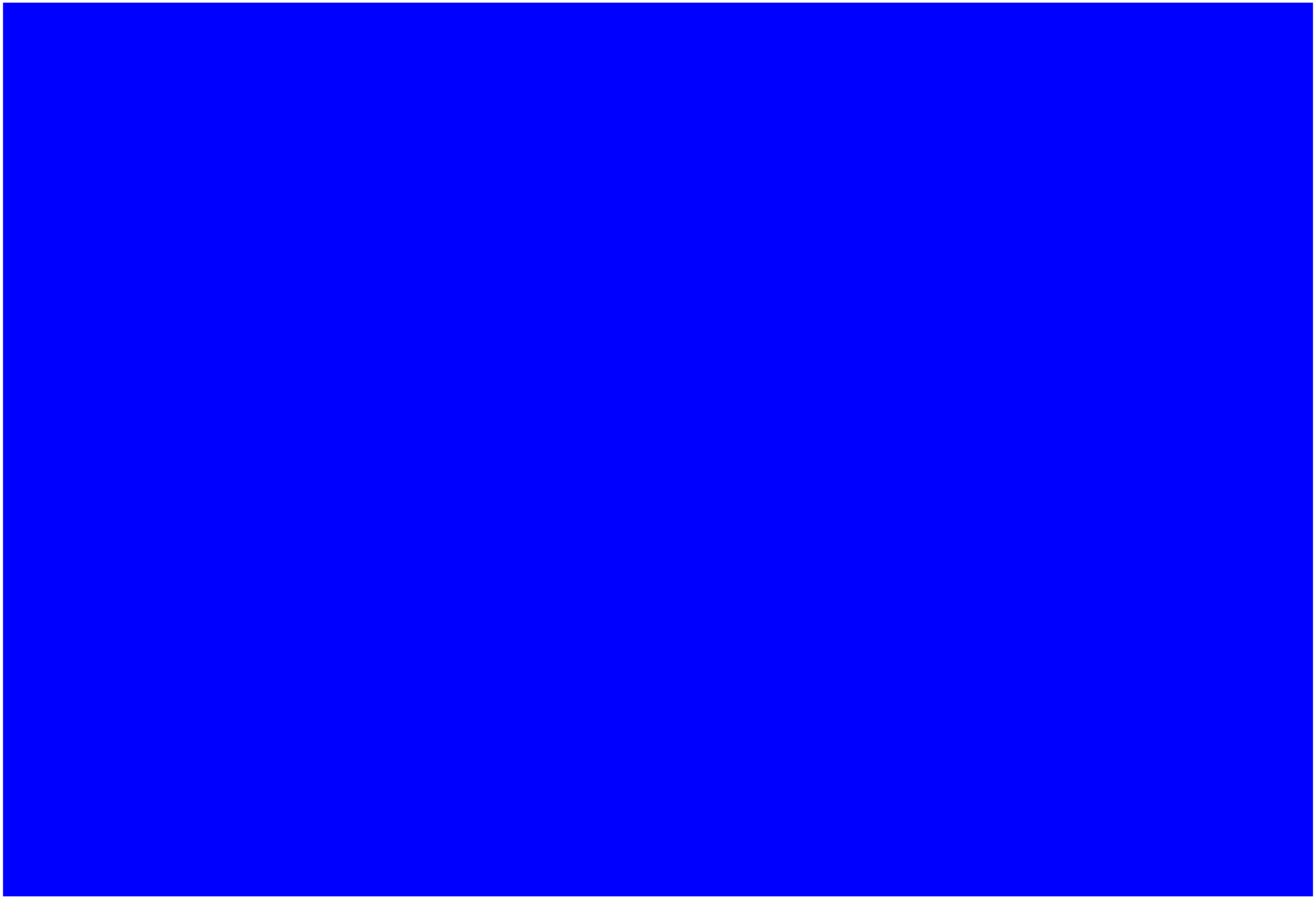
Let's find the variance of Petal.Length


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

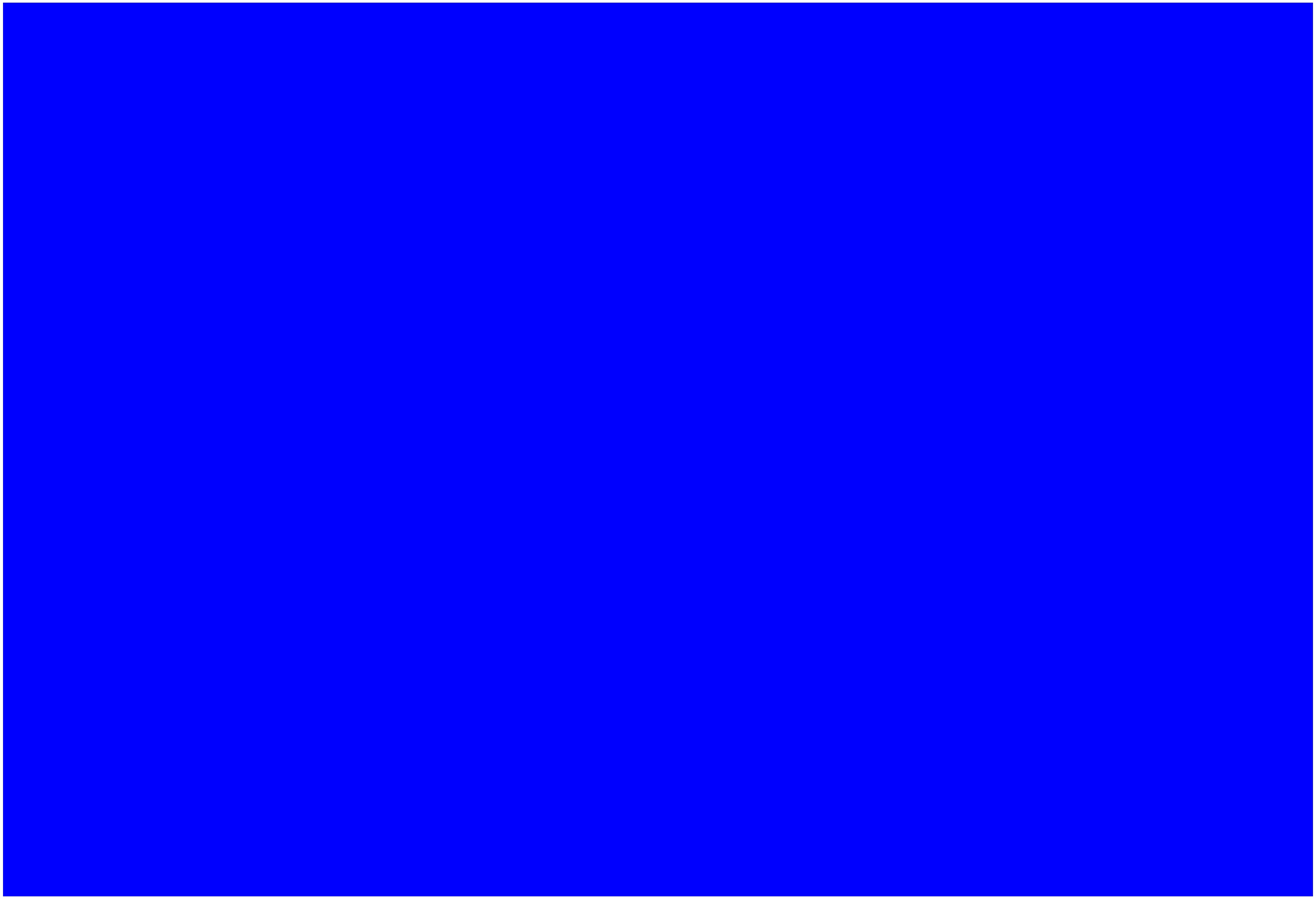


Why no red rectangles?
Petal.Length
Petal.Length
0
2
4
6
8
1
2
3
4
5
6
7

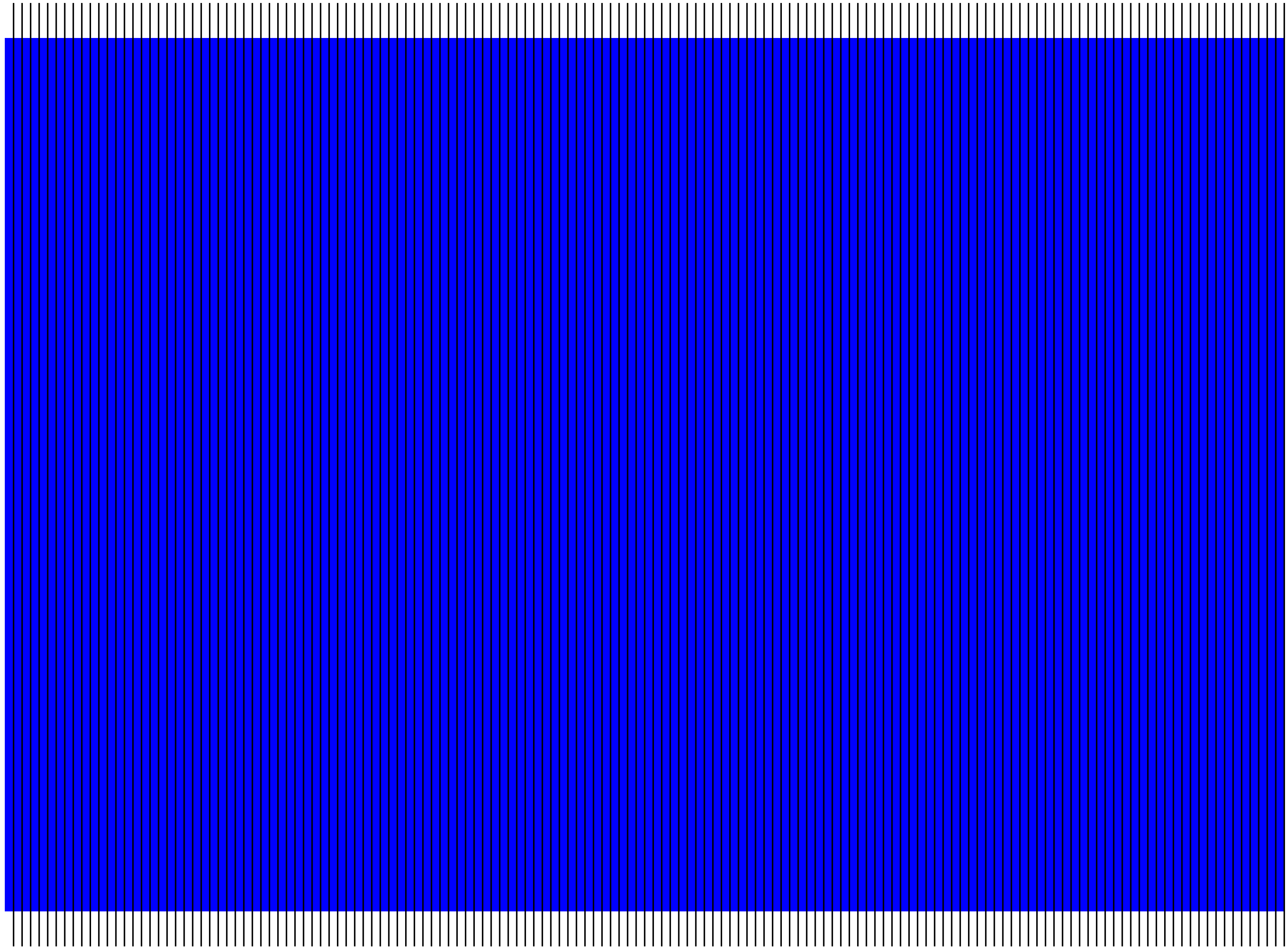
Add the blues together. (This is at a different scale.)



We have no reds to subtract.



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



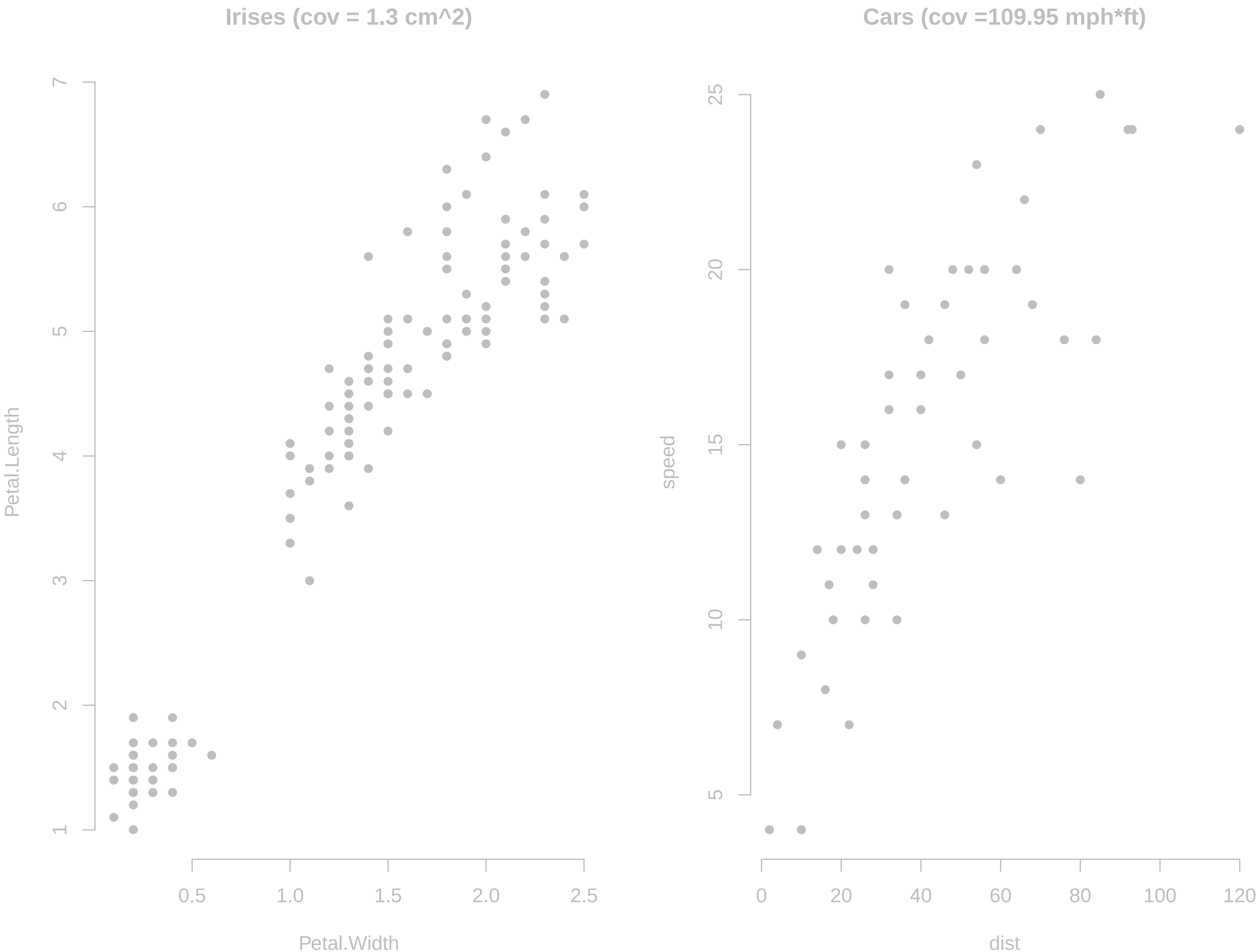
This blue sliver is the variance.

A problem with covariance

Covariance has units!

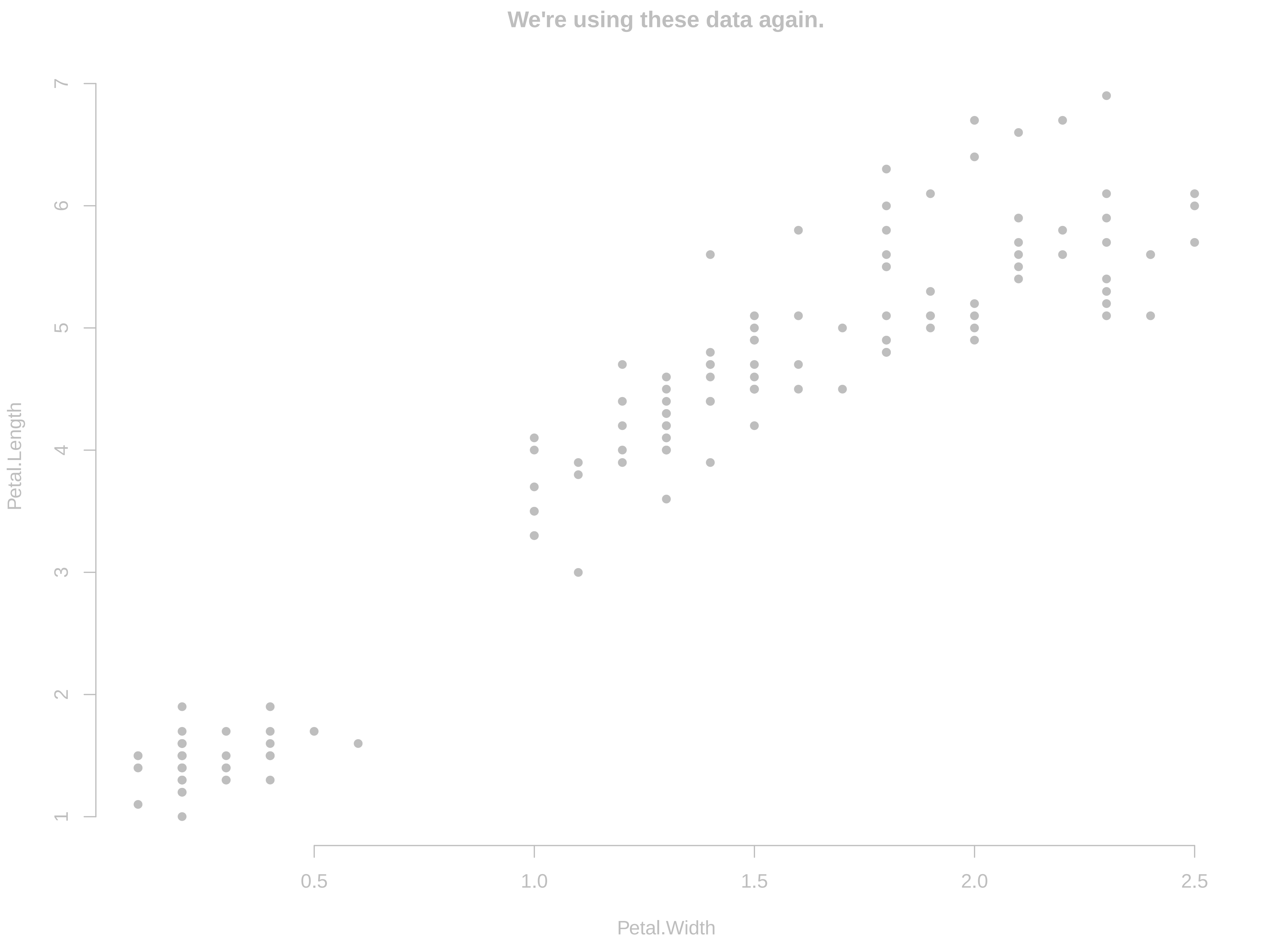
(x–unit times y–unit)

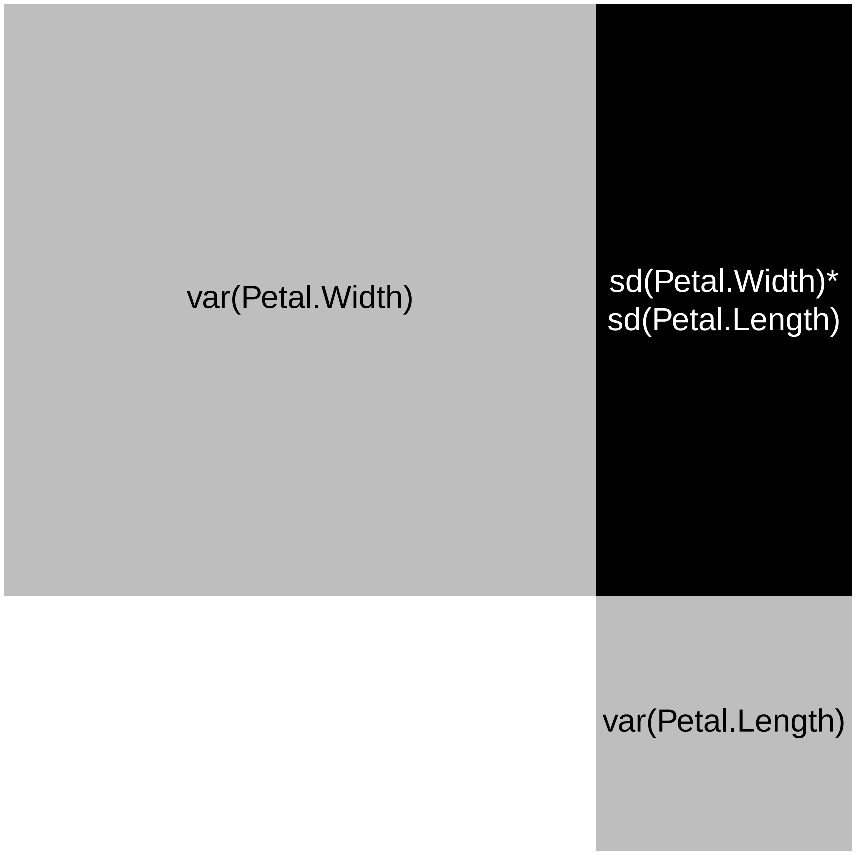
Which relationship is stronger (more linear)?

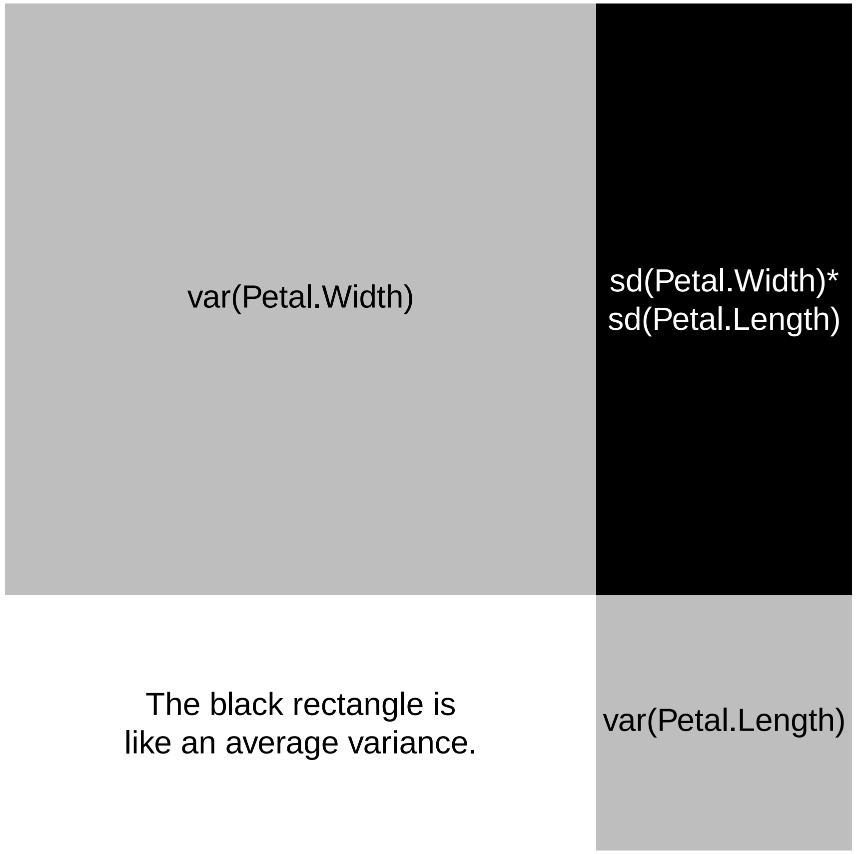

Irises (cov = 1.3 cm^2)
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Cars (cov =109.95 mph*ft)
speed
dist
5
10
15
20
25
0
20
40
60
80
100
120

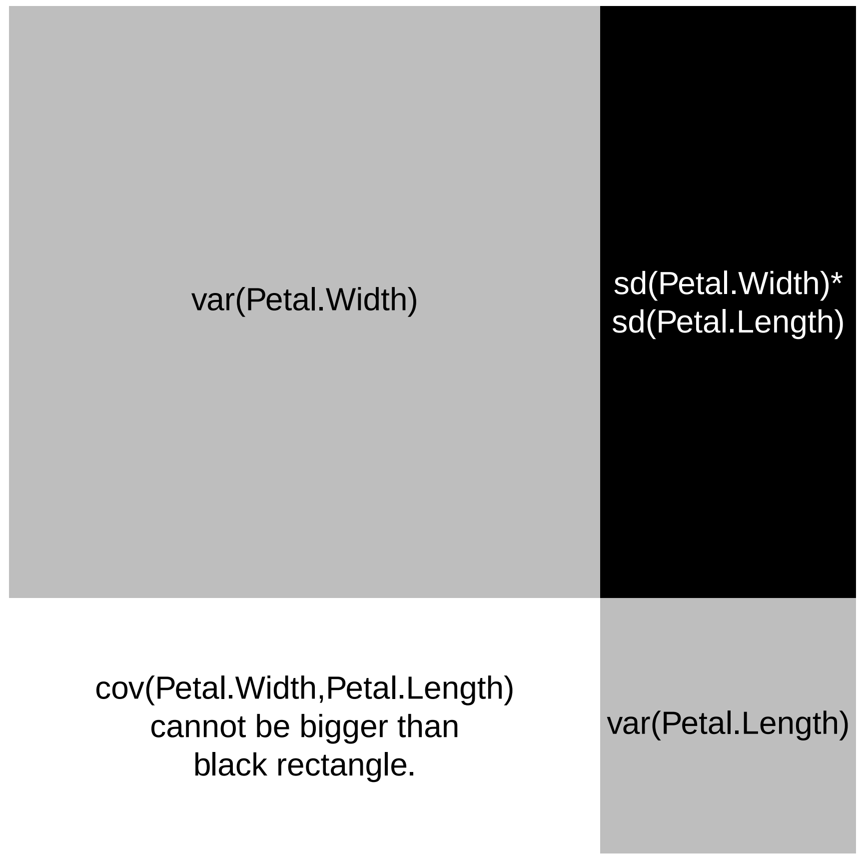
Oh noes!

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

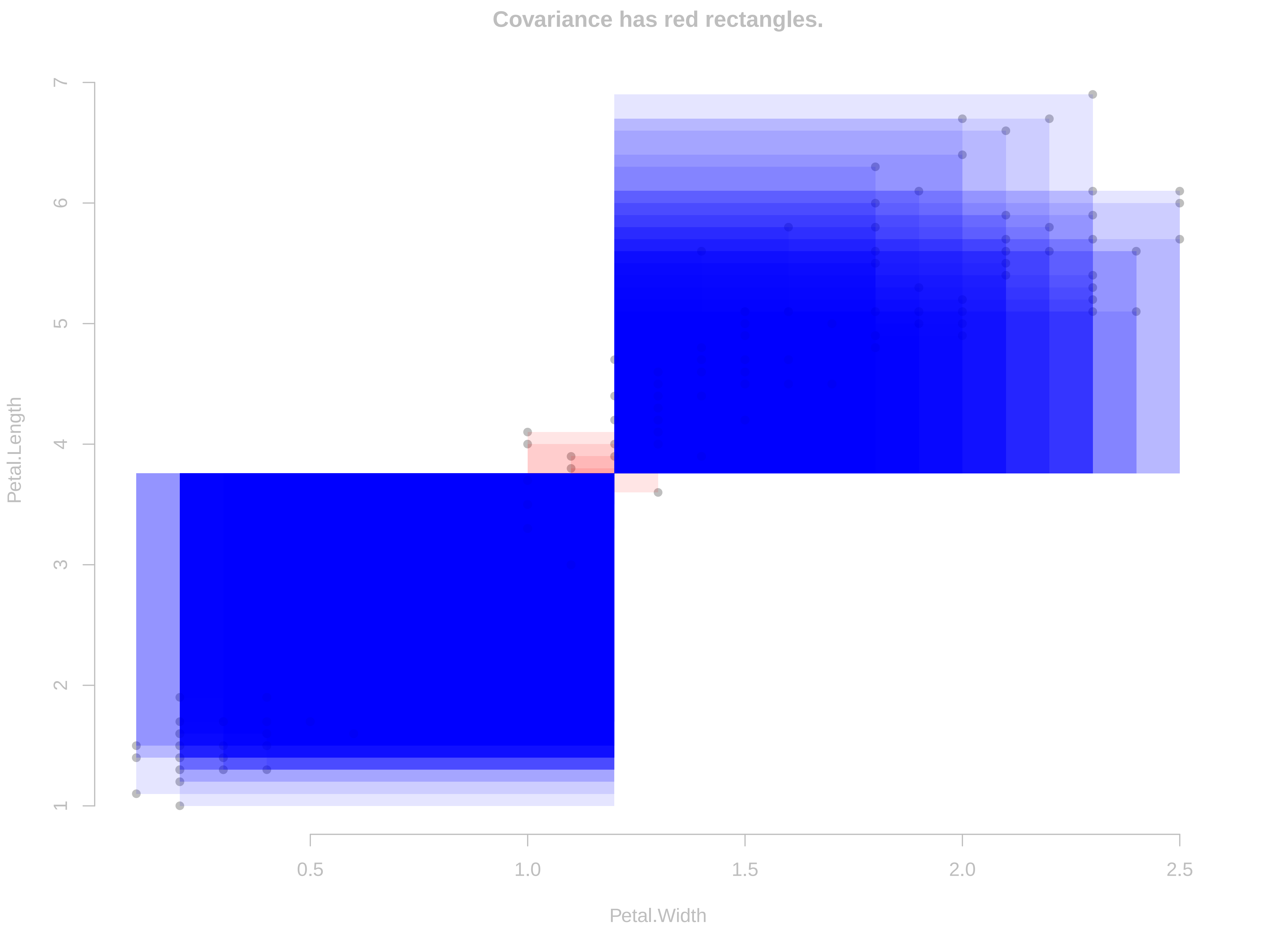

We're using these data again.
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

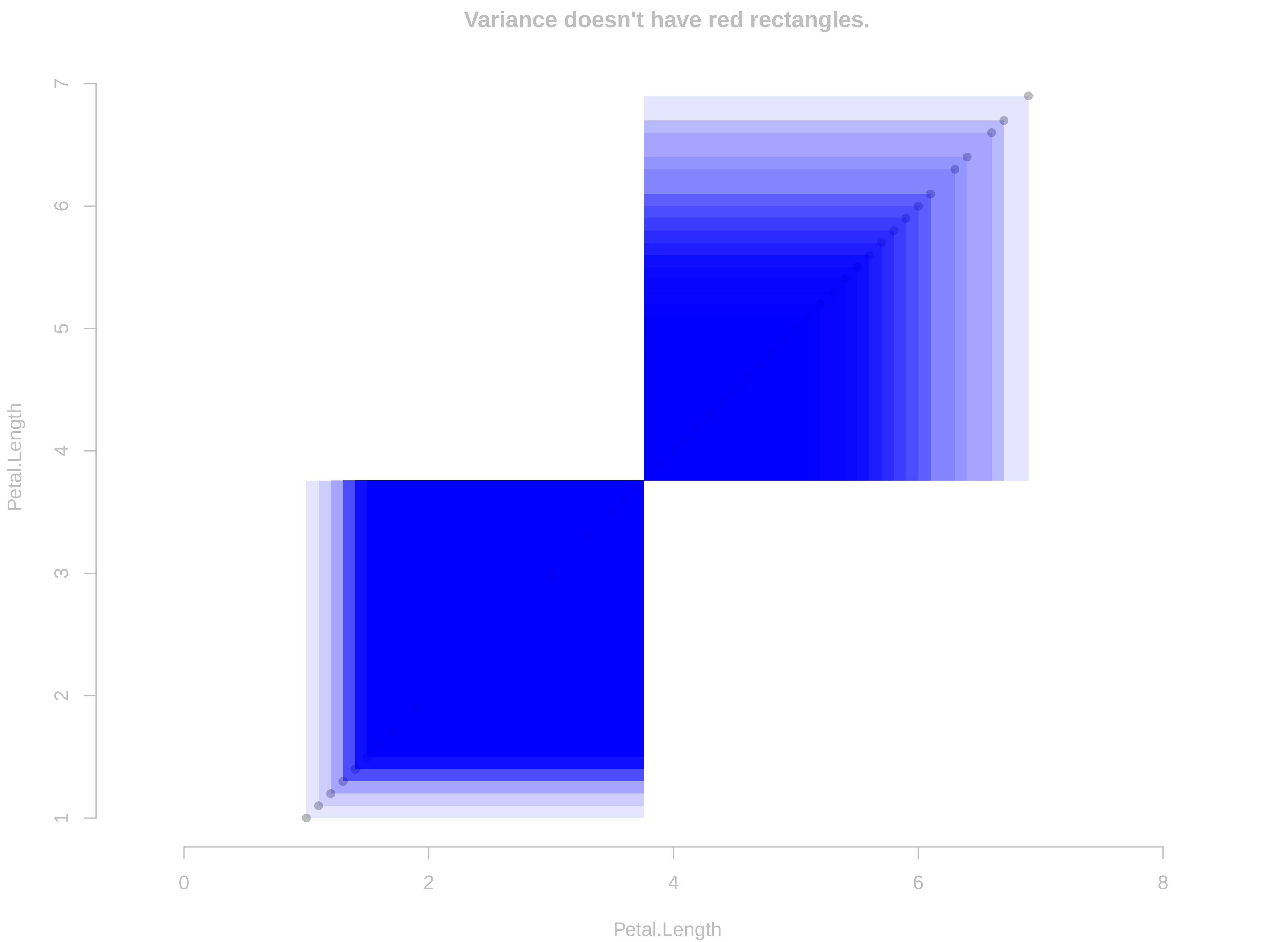

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
var(Petal.Length)

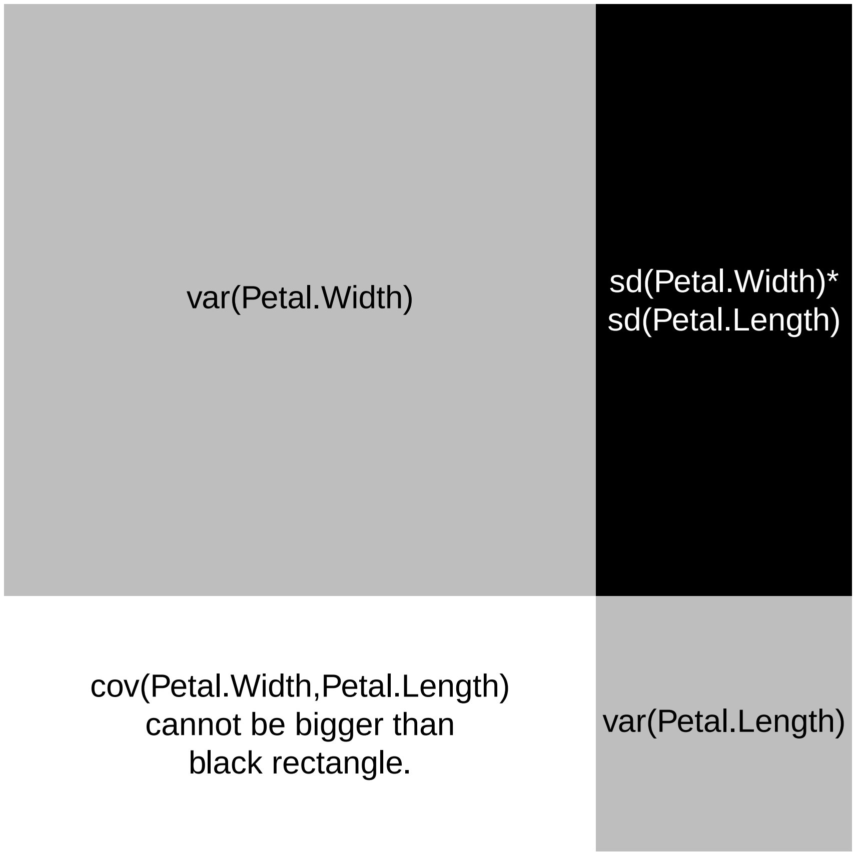

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
var(Petal.Length)
The black rectangle is
like an average variance.


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

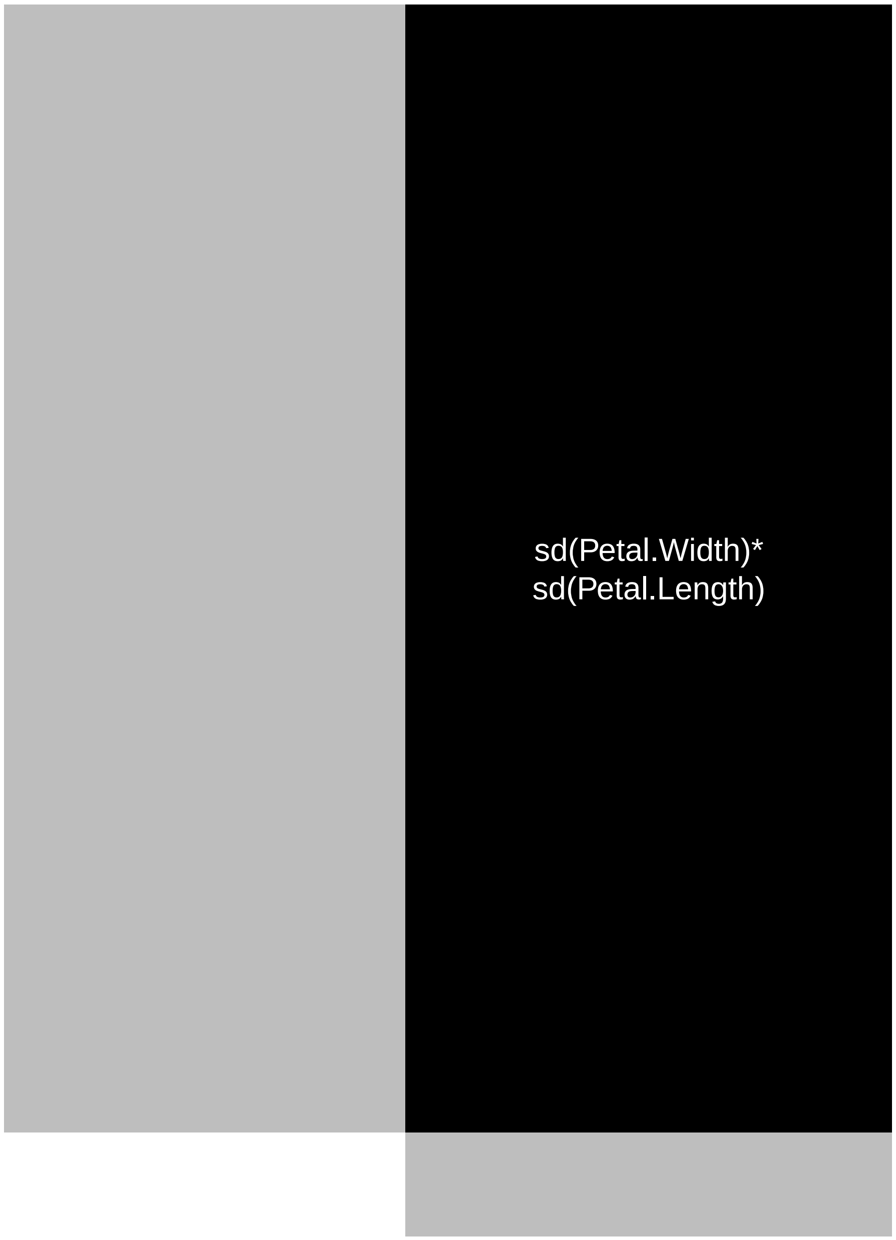
Why?


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

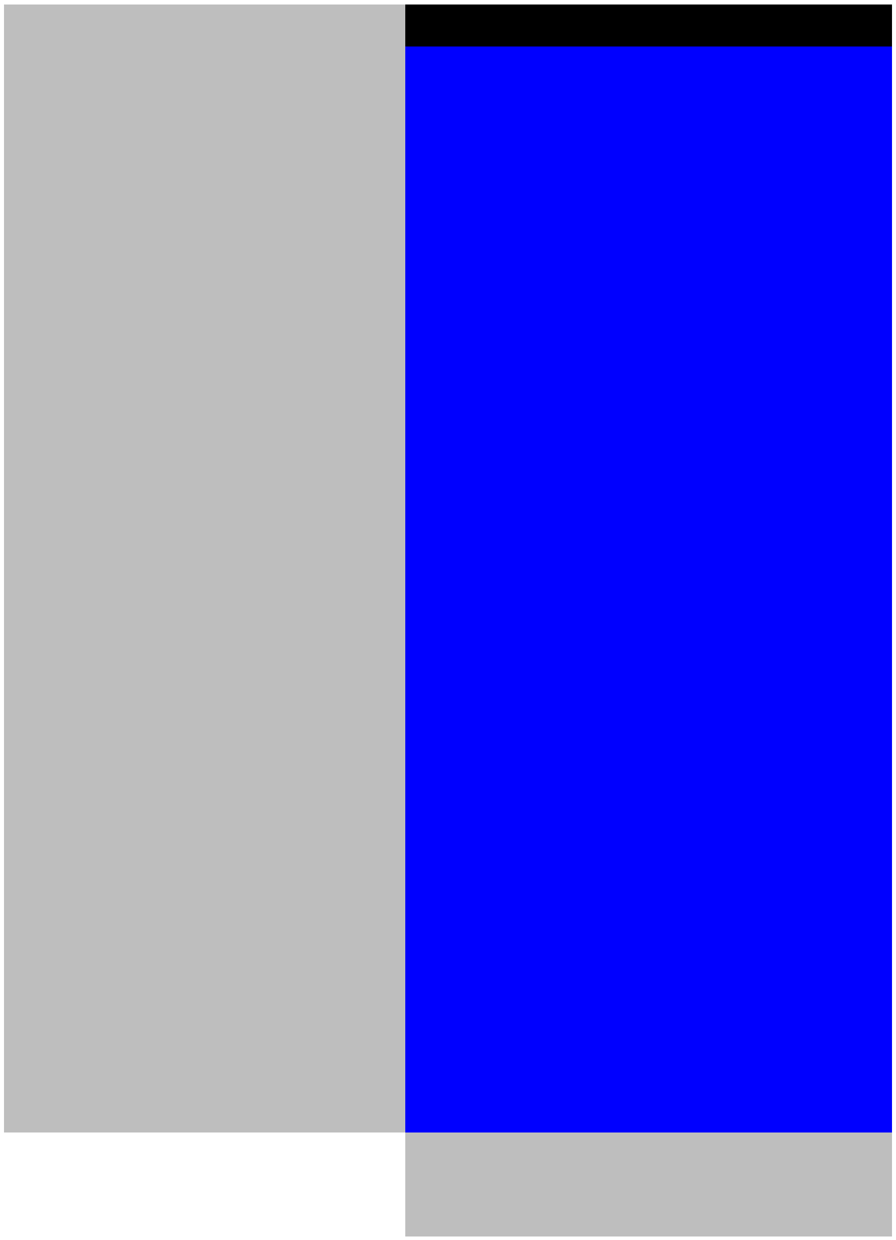

Variance doesn't have red rectangles.
Petal.Length
Petal.Length
0
2
4
6
8
1
2
3
4
5
6
7


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.

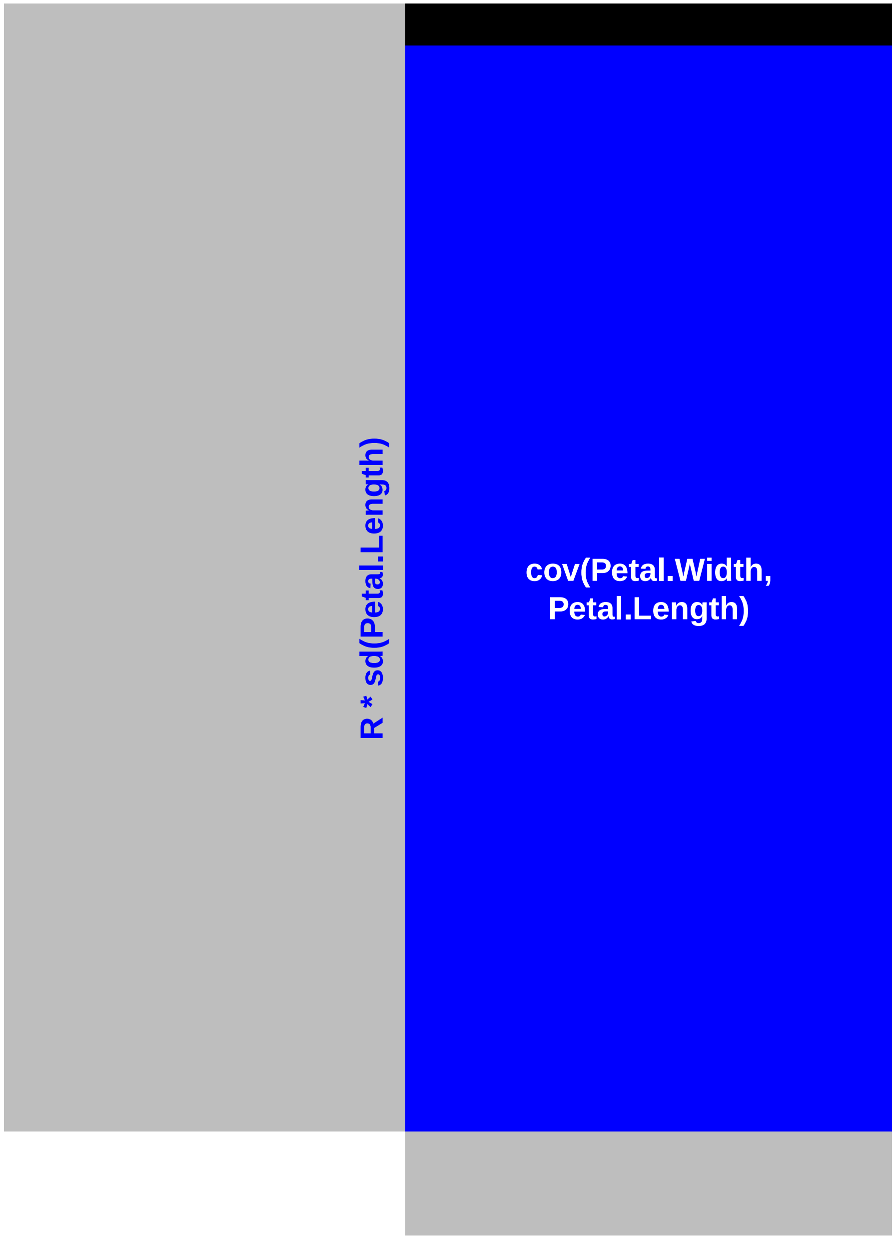


Squish covariance vertically into the rectangle.

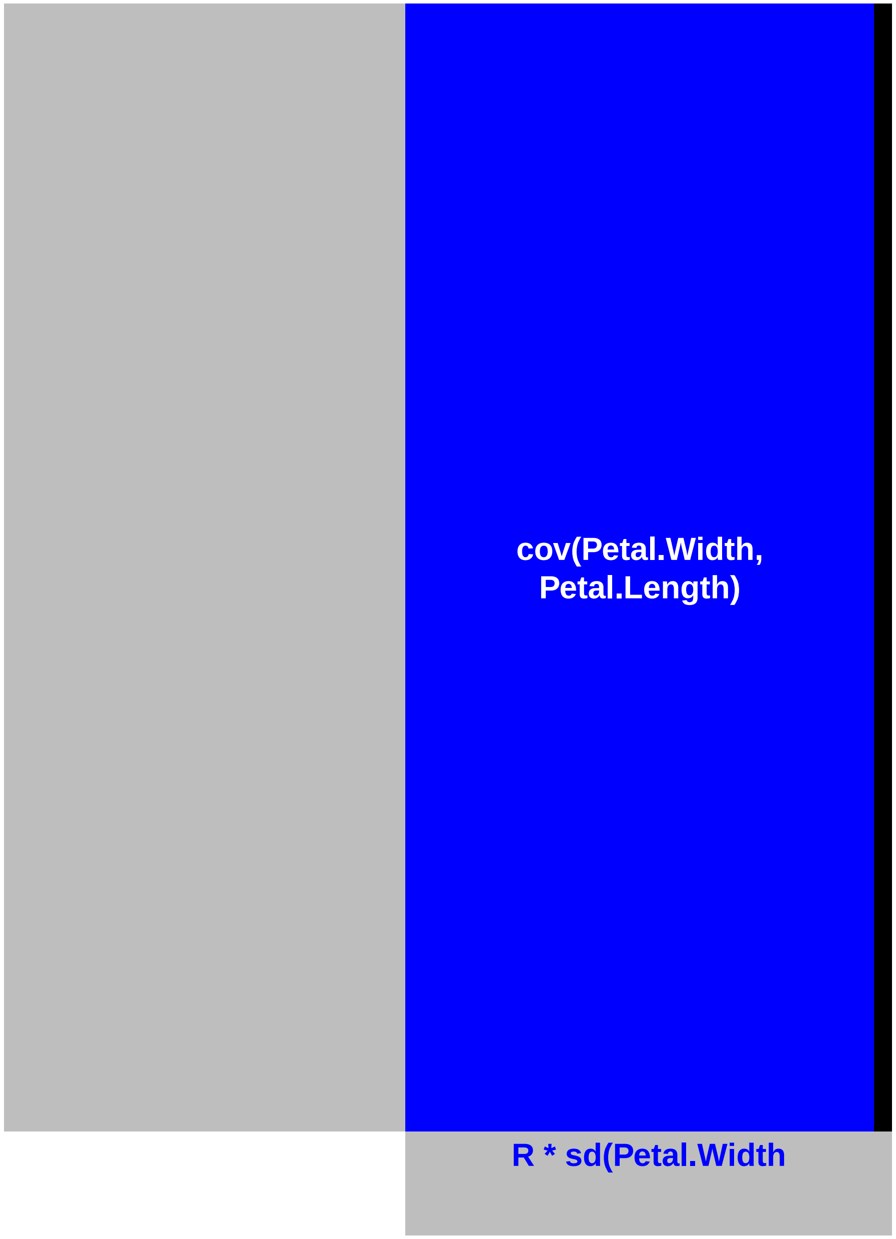


Correlation (R)
is the ratio of
the small rectangle
to the big rectangle.

Squish covariance vertically into the rectangle.

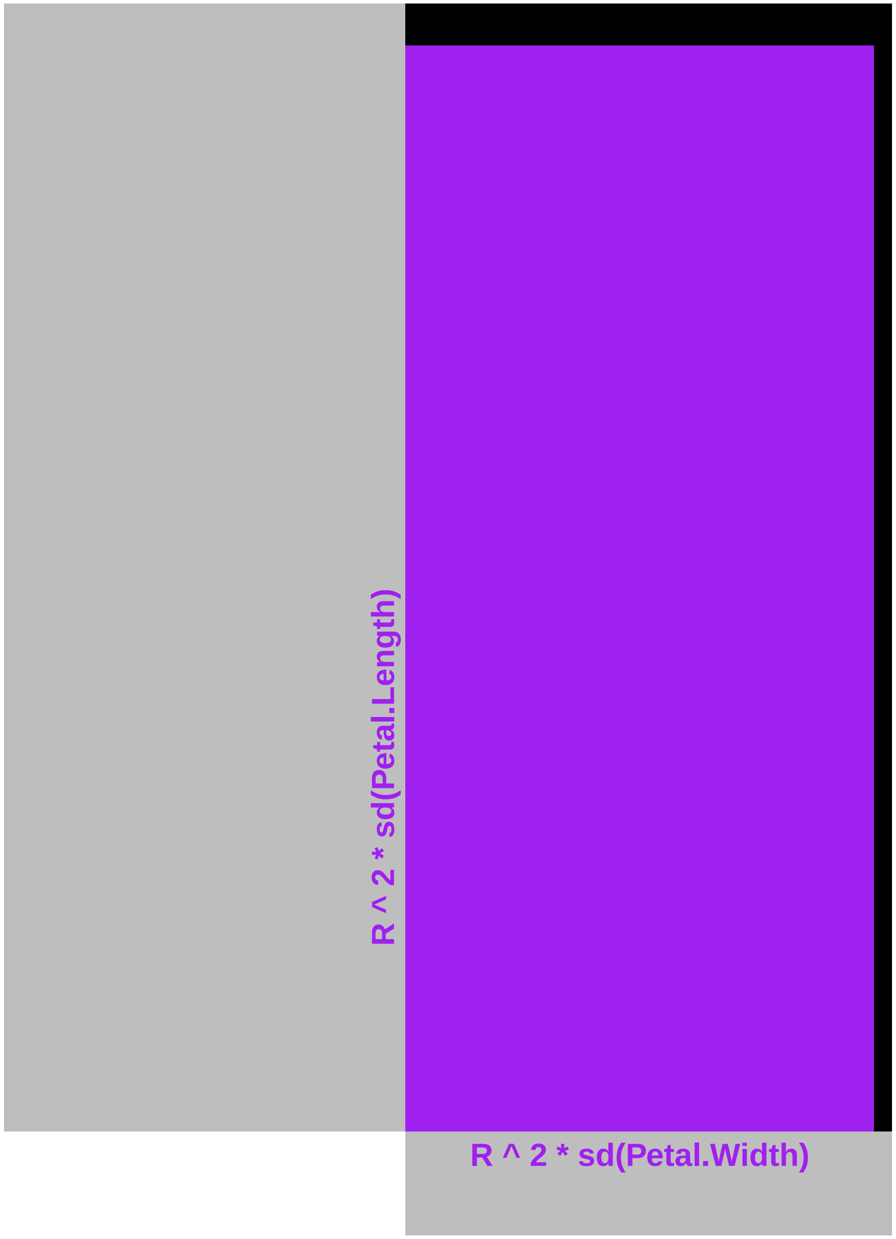


Squish covariance horizontally into the rectangle.

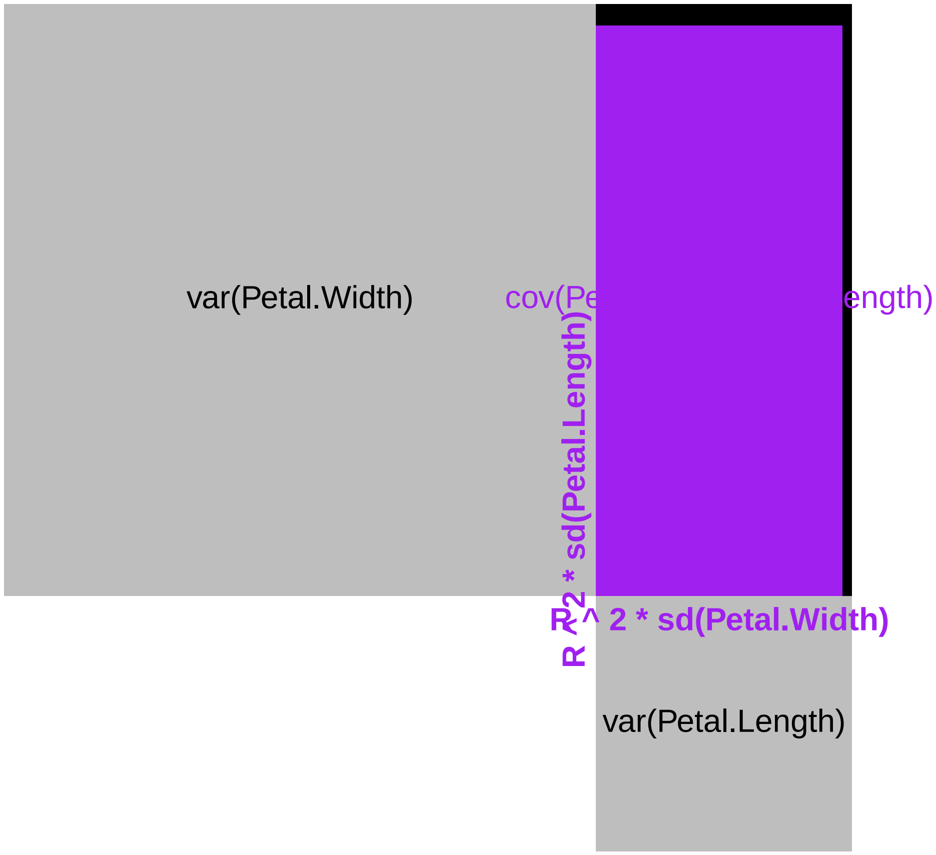


People like to talk about R–squared.

Intersect the two squished covariance rectangles.

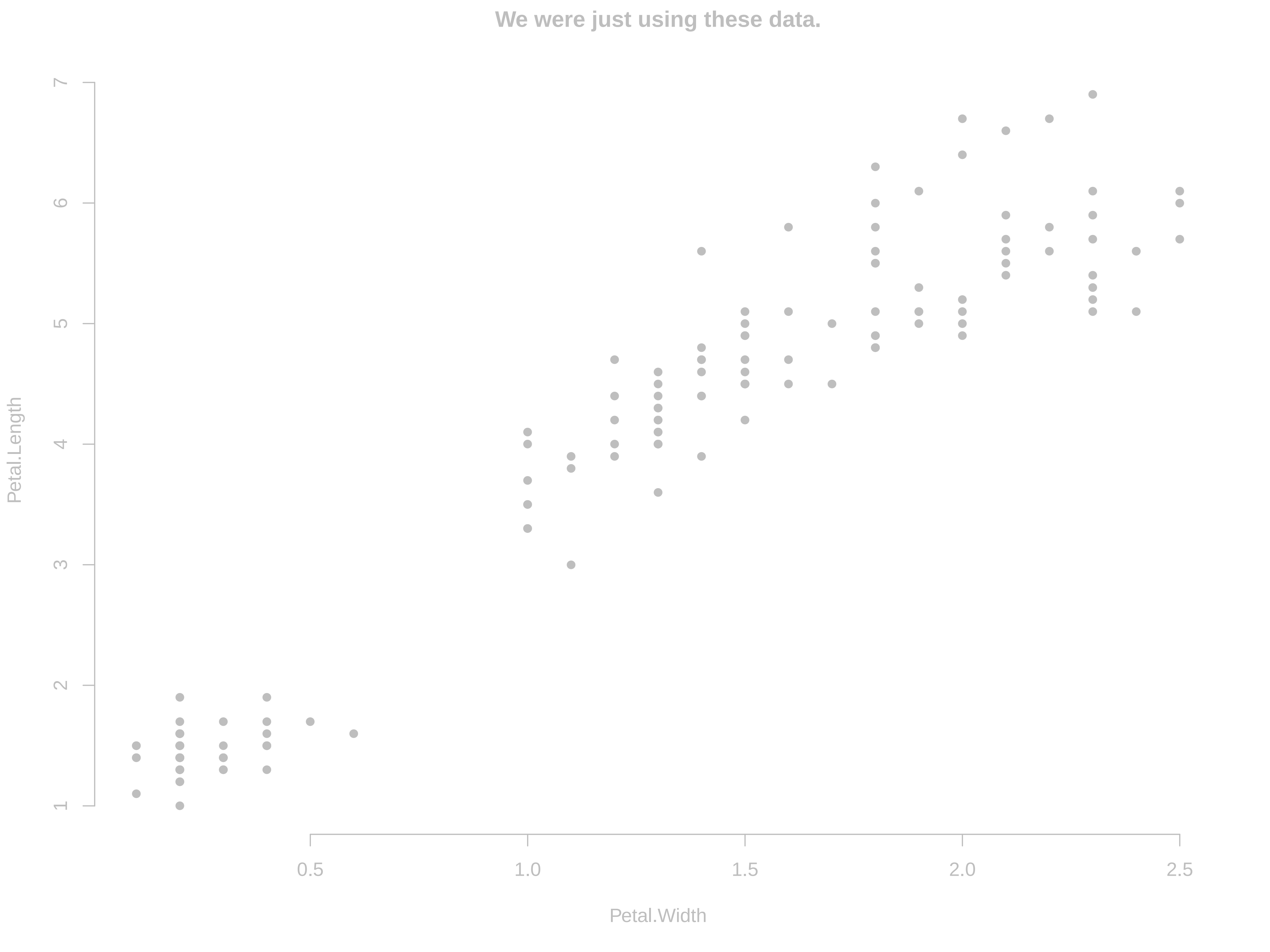


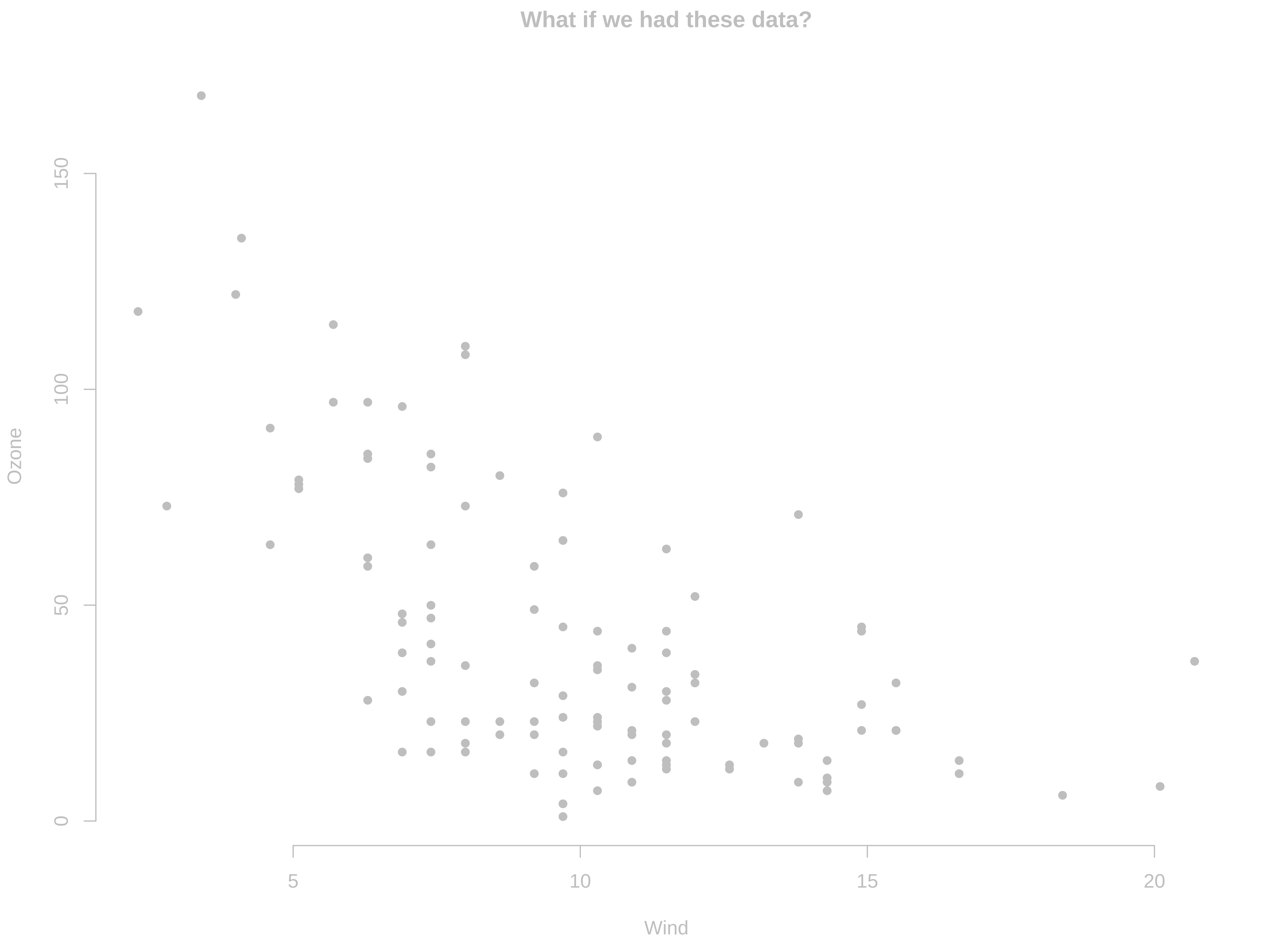
Intersect the two squished covariance rectangles.

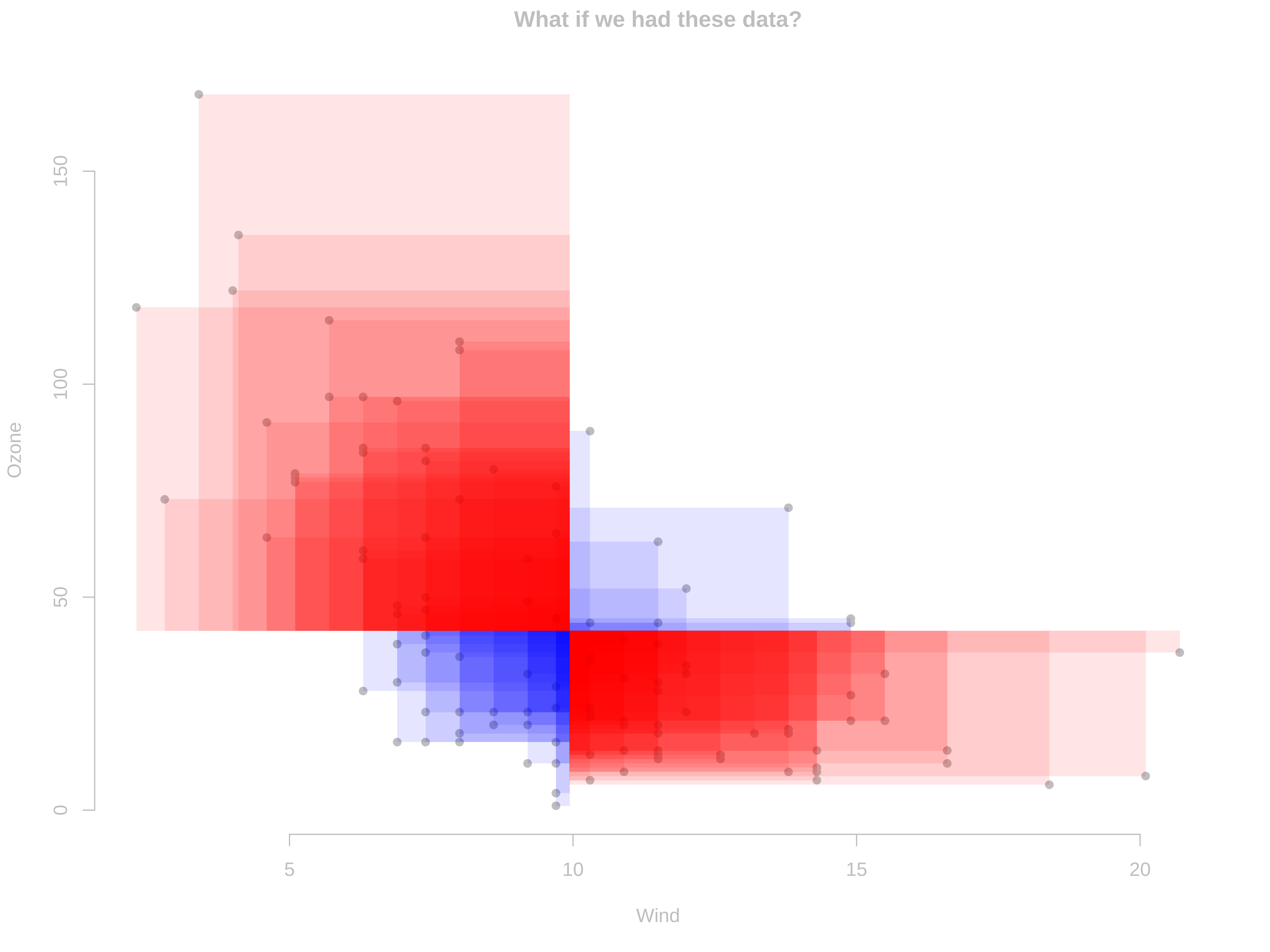


That was for very positive (blue) covariances.

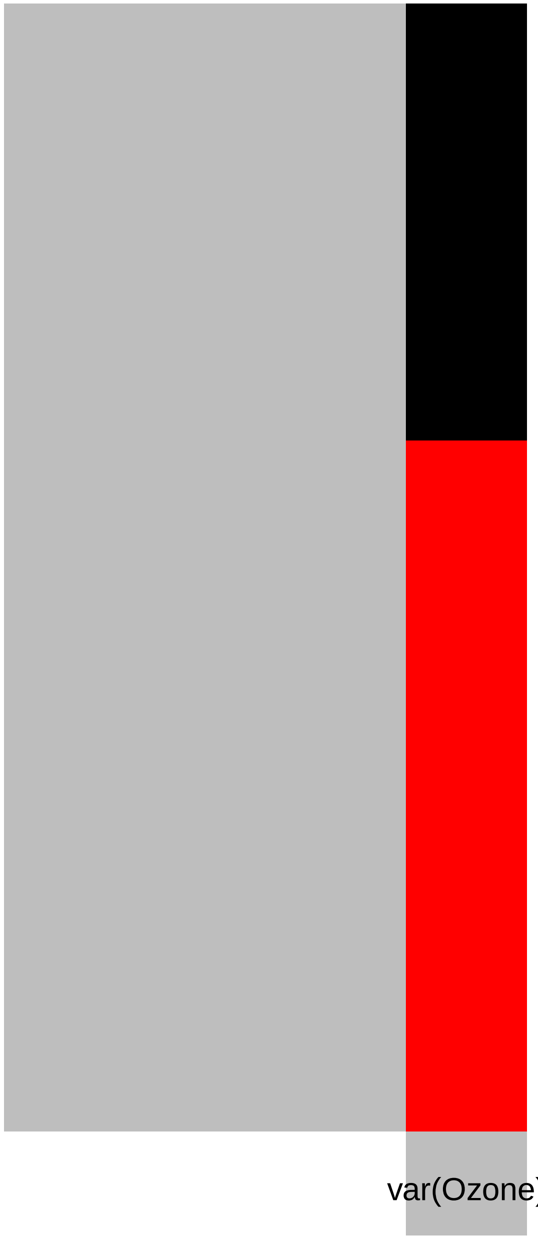
What if covariance is negative (red)?


We were just using these data.
Petal.Length
7
6
5
4
3
2
1
0.5
1.0
1.5
2.0
2.5
Petal.Width

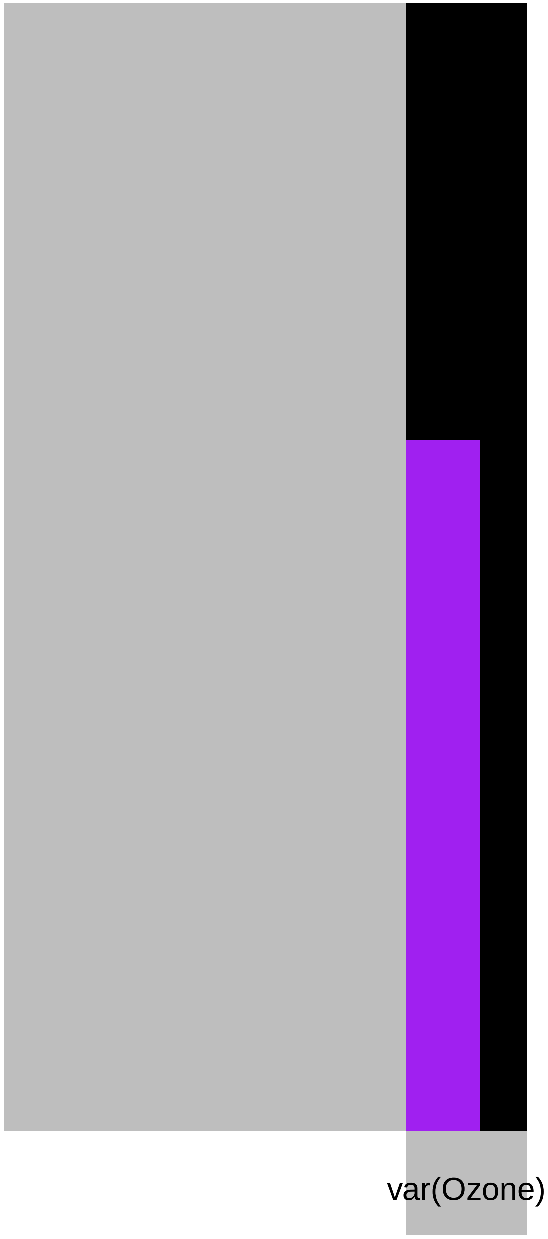

What if we had these data?
Ozone
0
50
100
150
Wind
5
10
15
20


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

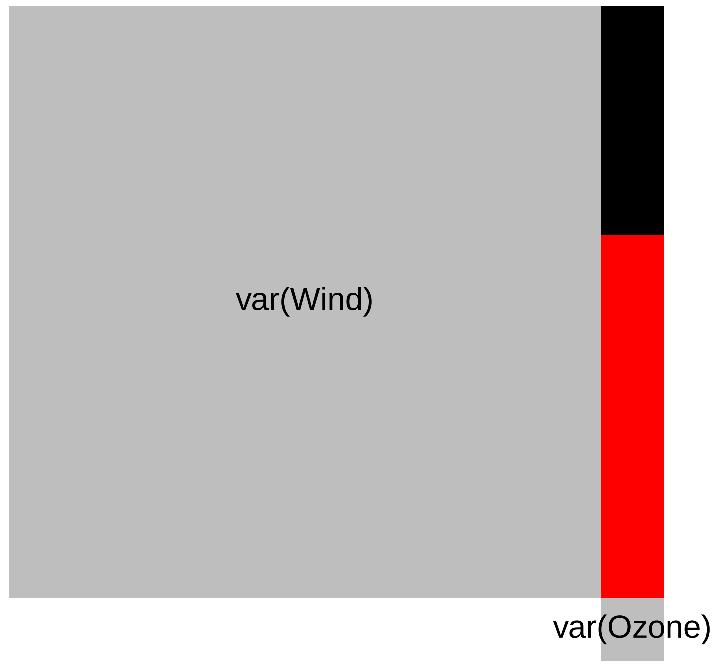
R is the same, just negative.



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

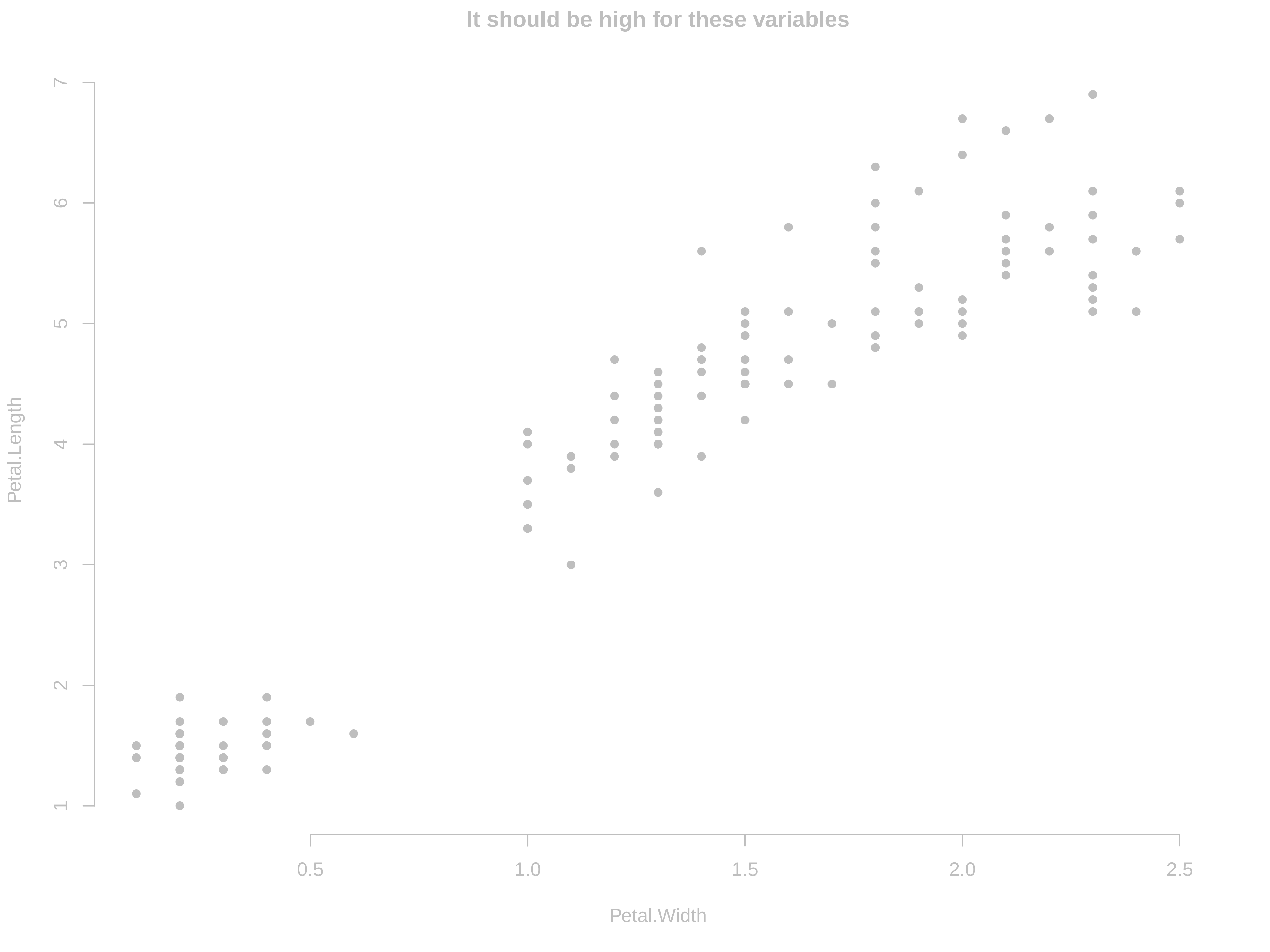


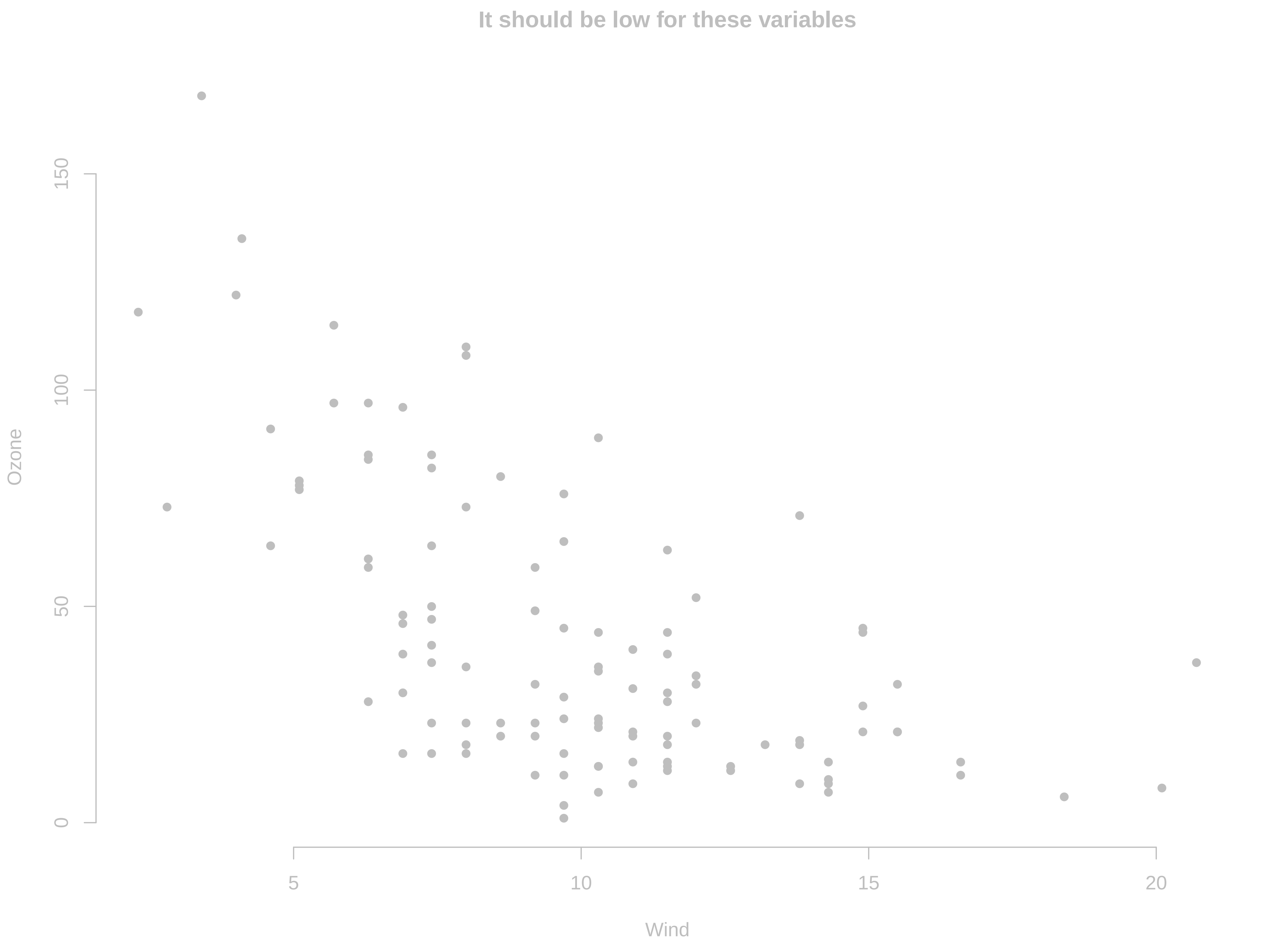
Zoom back out.

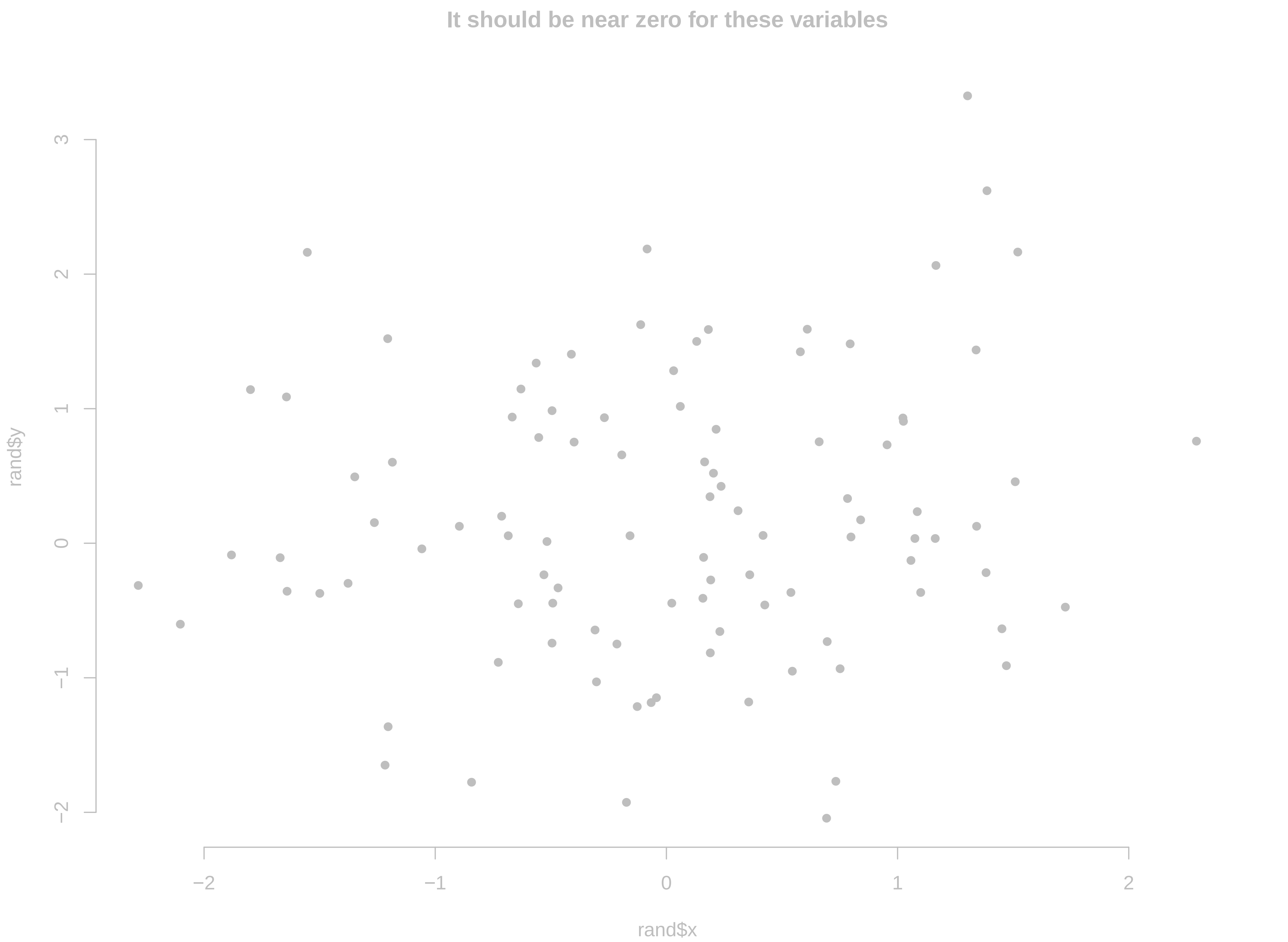


Remember how this fits in.

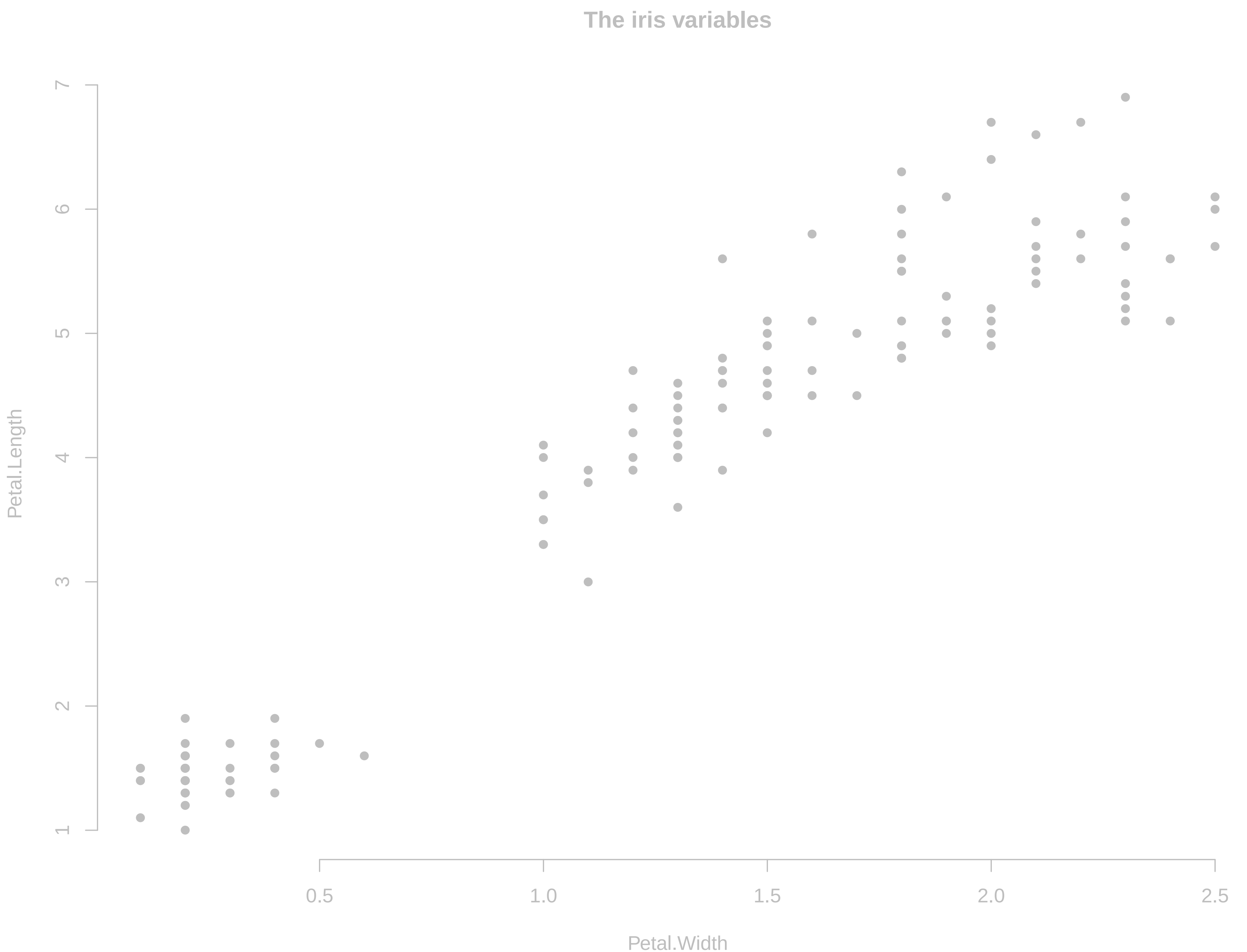
We want a number that describes whether two variables move together.

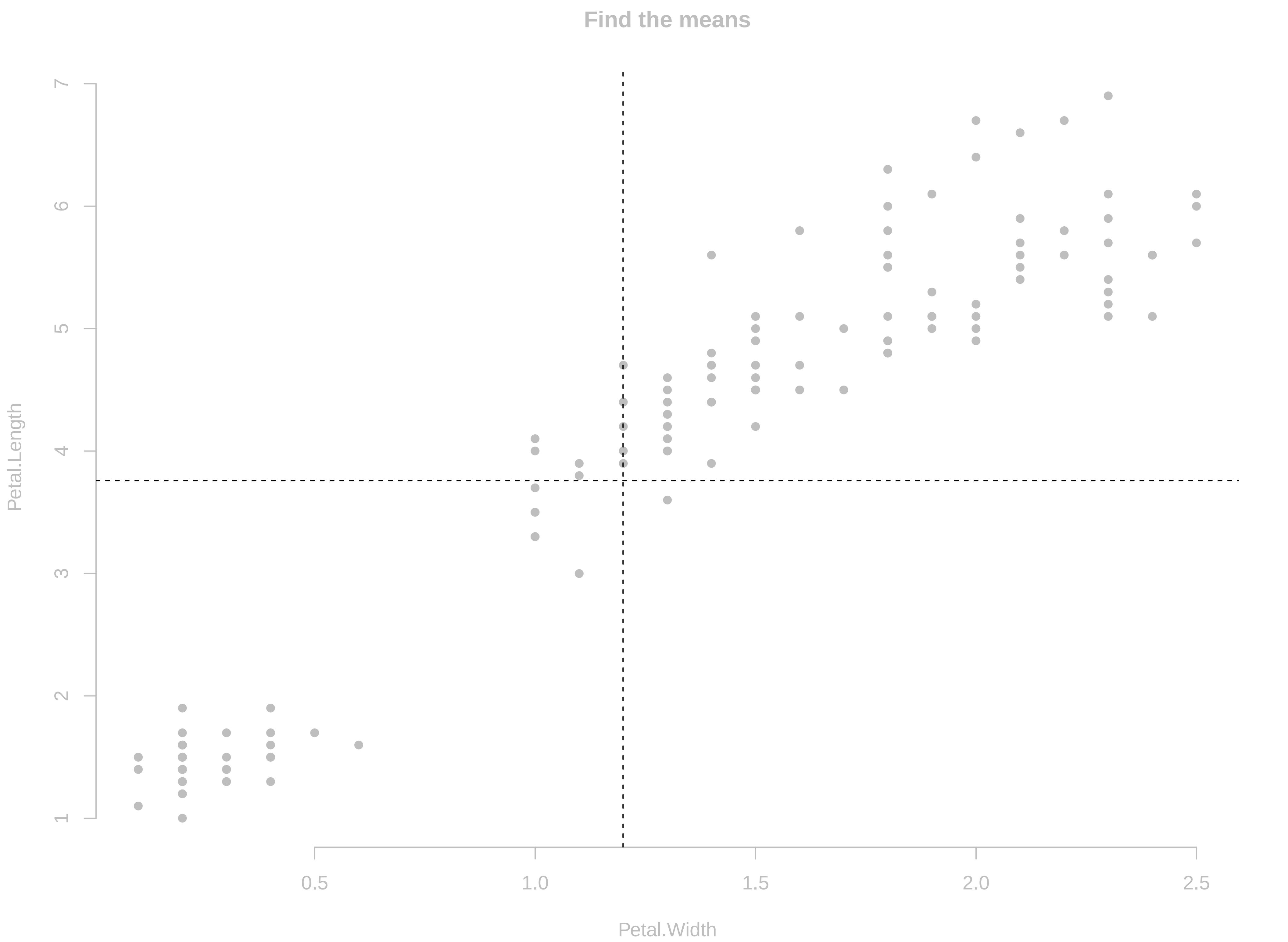

It should be high for these variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

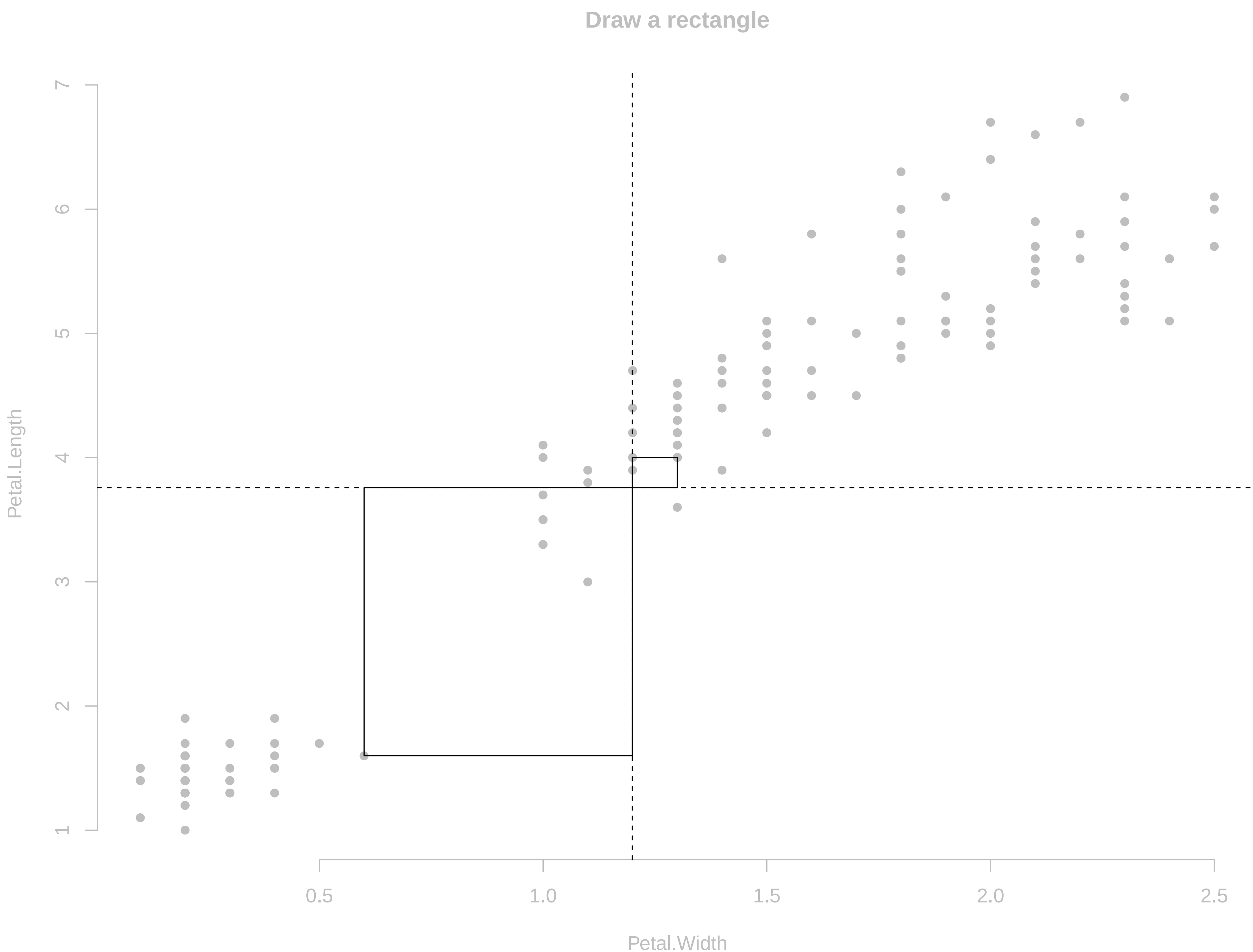

It should be low for these variables
Ozone
0
50
100
150
Wind
5
10
15
20


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

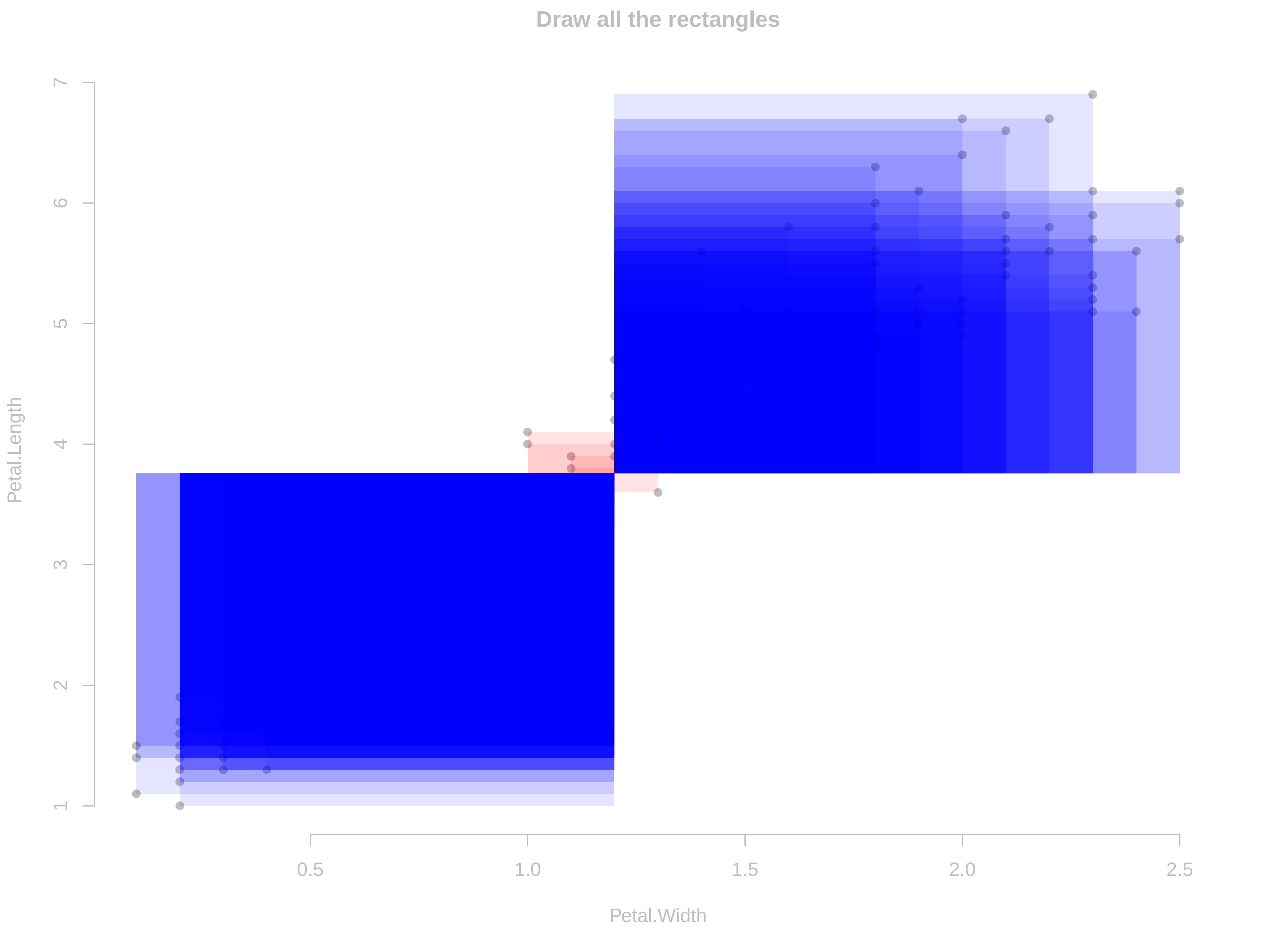
Covariance

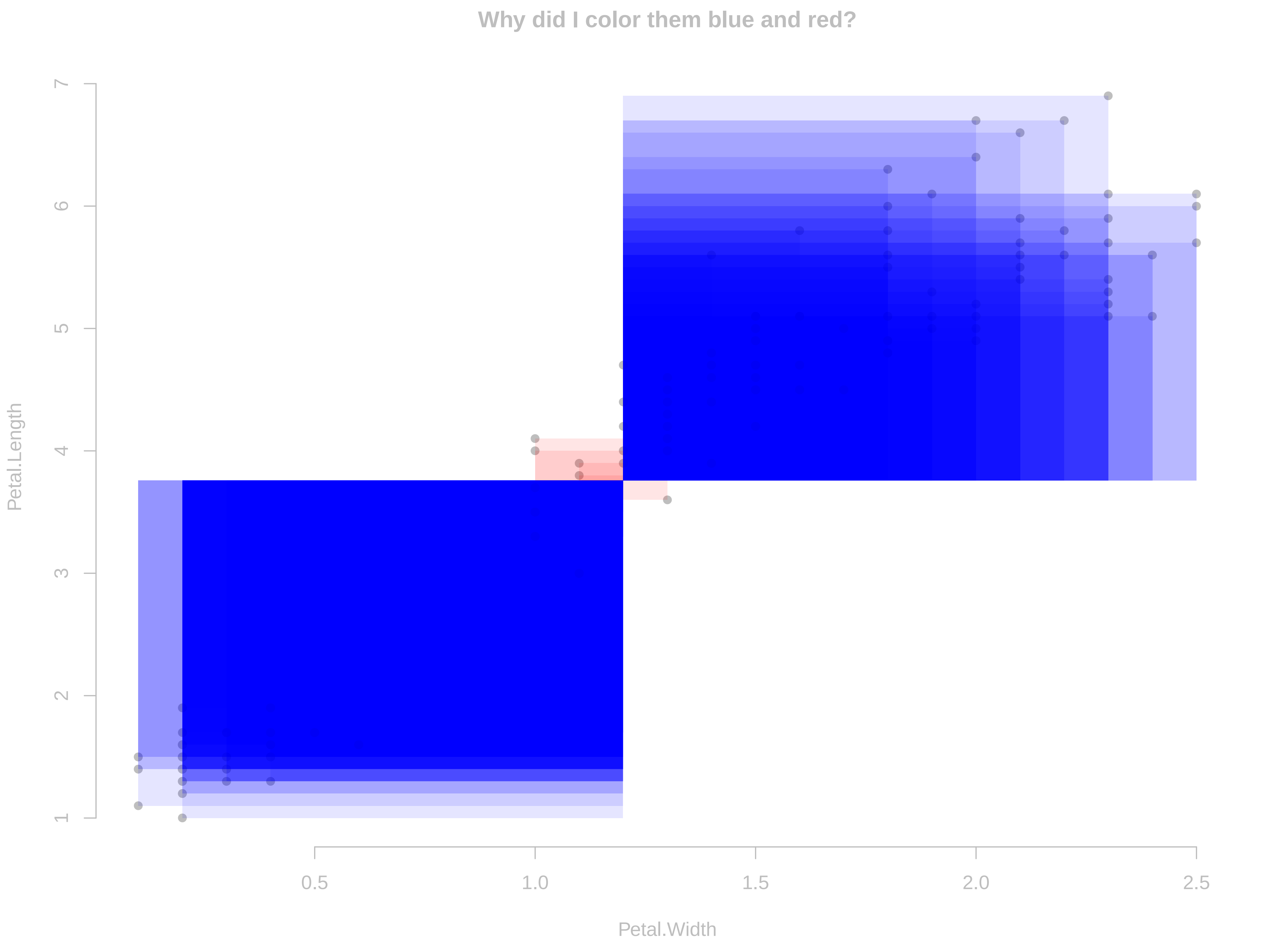

The iris variables
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


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

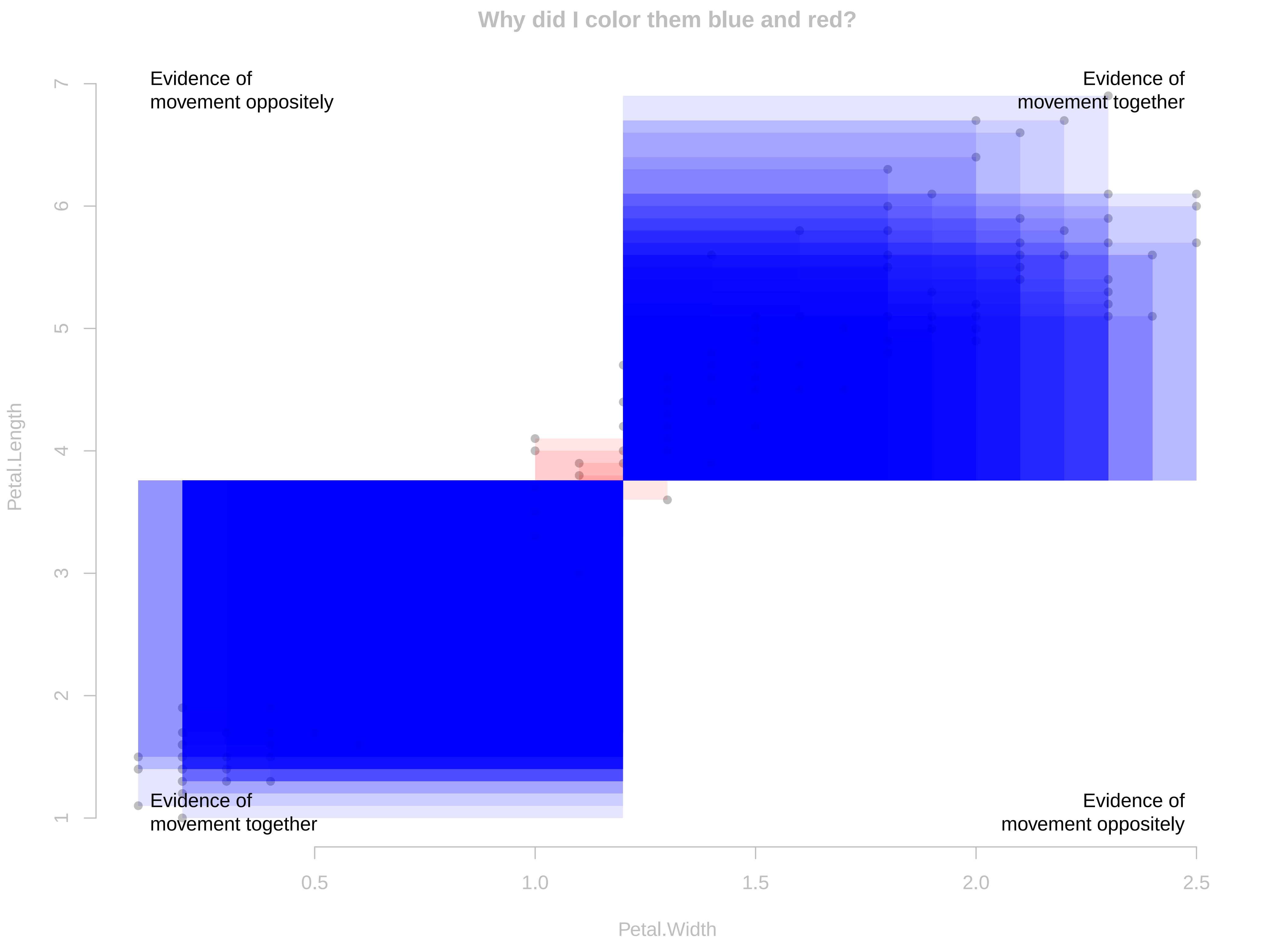


Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


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

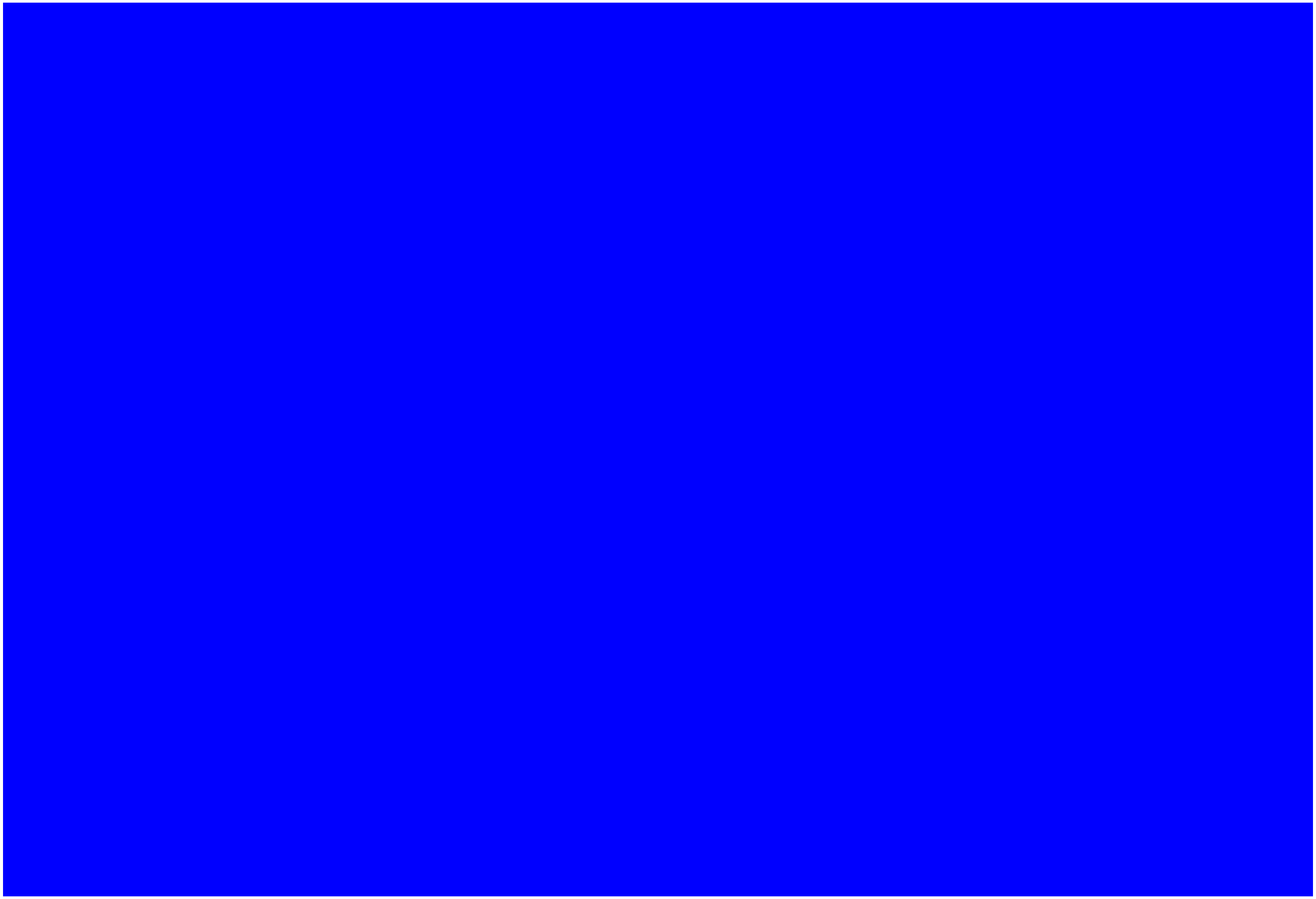


Why did I color them blue and red?
Petal.Length
1
2
3
4
5
6
7
Petal.Width
0.5
1.0
1.5
2.0
2.5

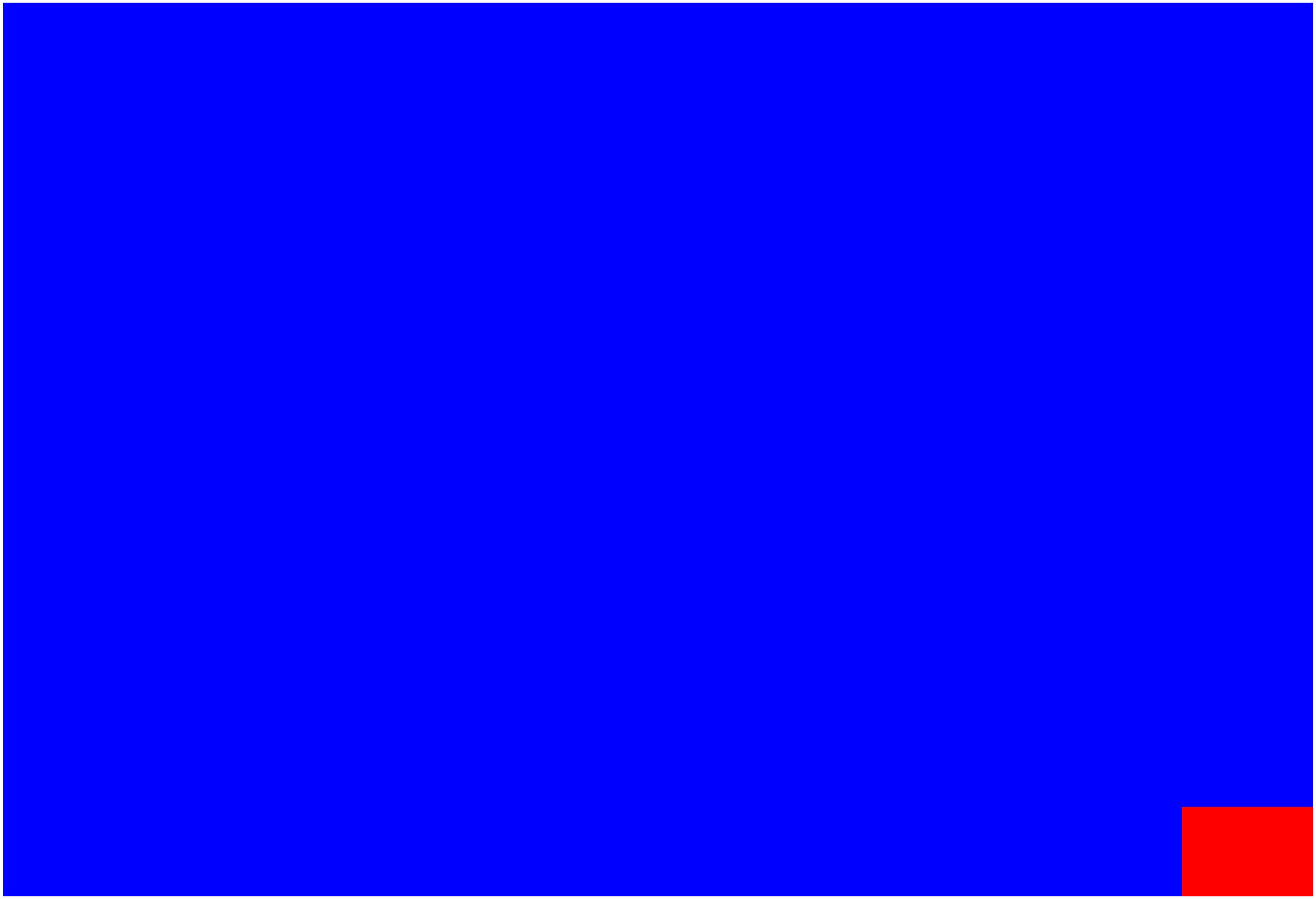


Why did I color them blue and red?
Evidence of
movement oppositely
Evidence of
movement together
Evidence of
movement together
Evidence of
movement oppositely
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

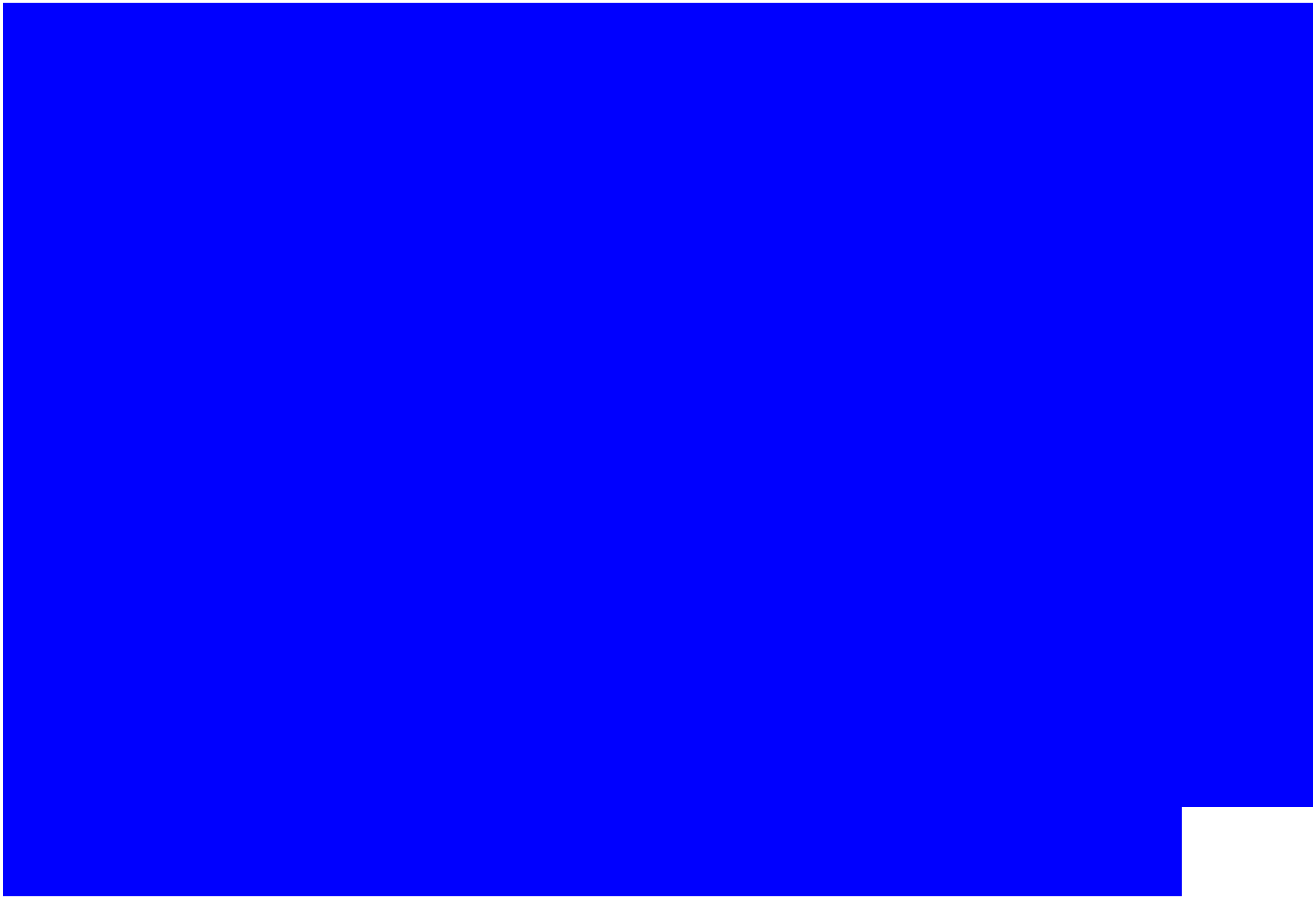
Add the blues together. (This is at a different scale.)



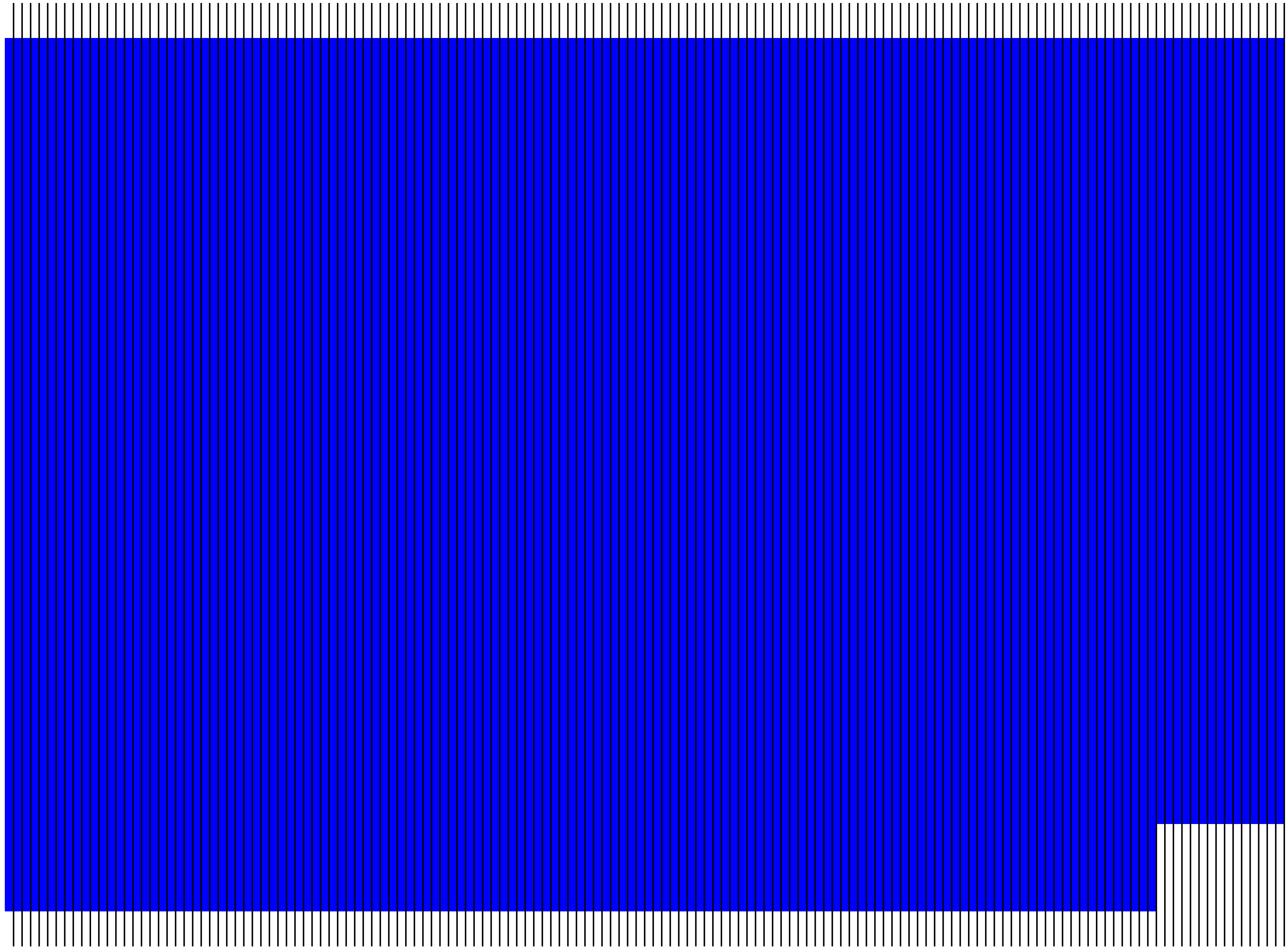
Add the reds together.



Subtract the reds.



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



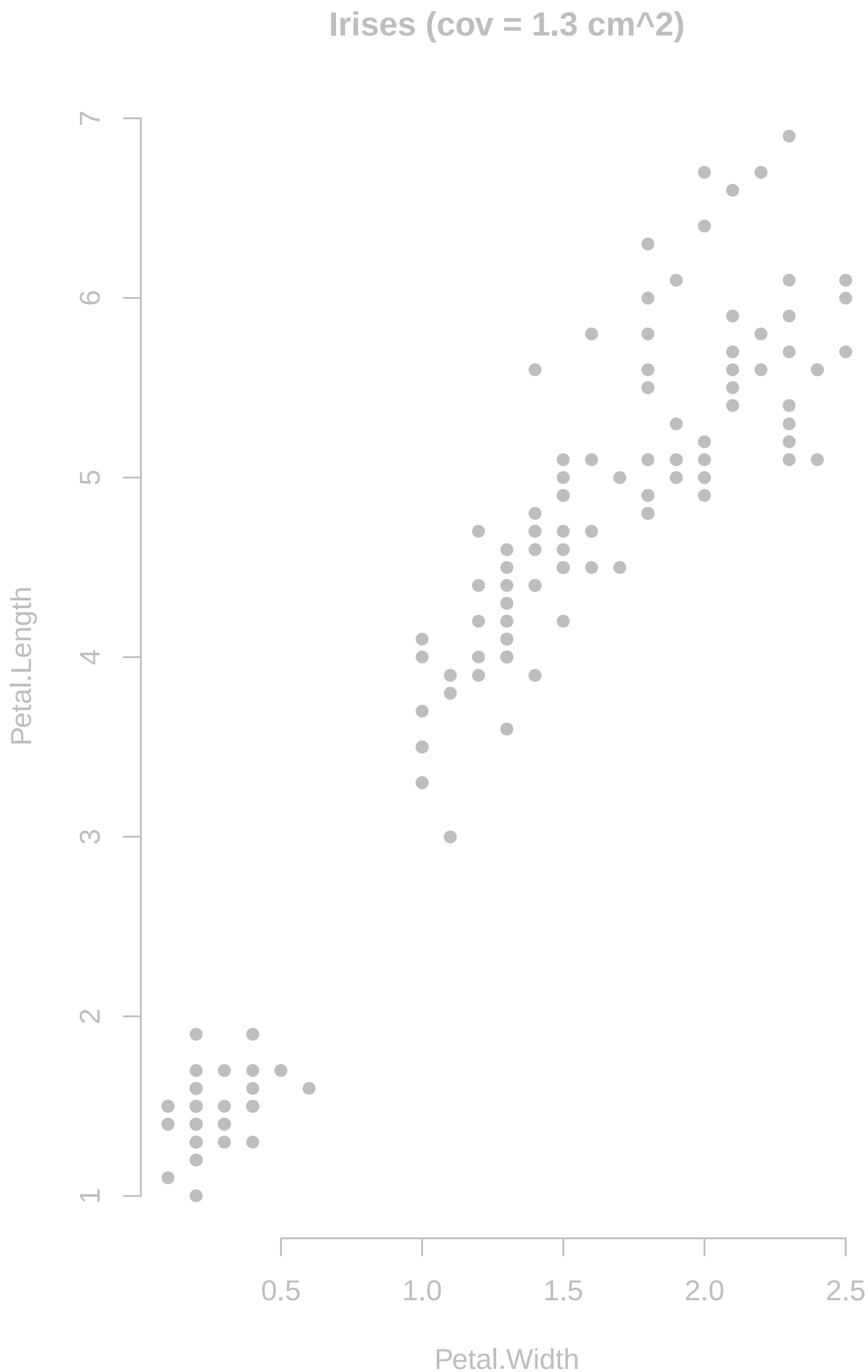
This blue sliver is the covariance.

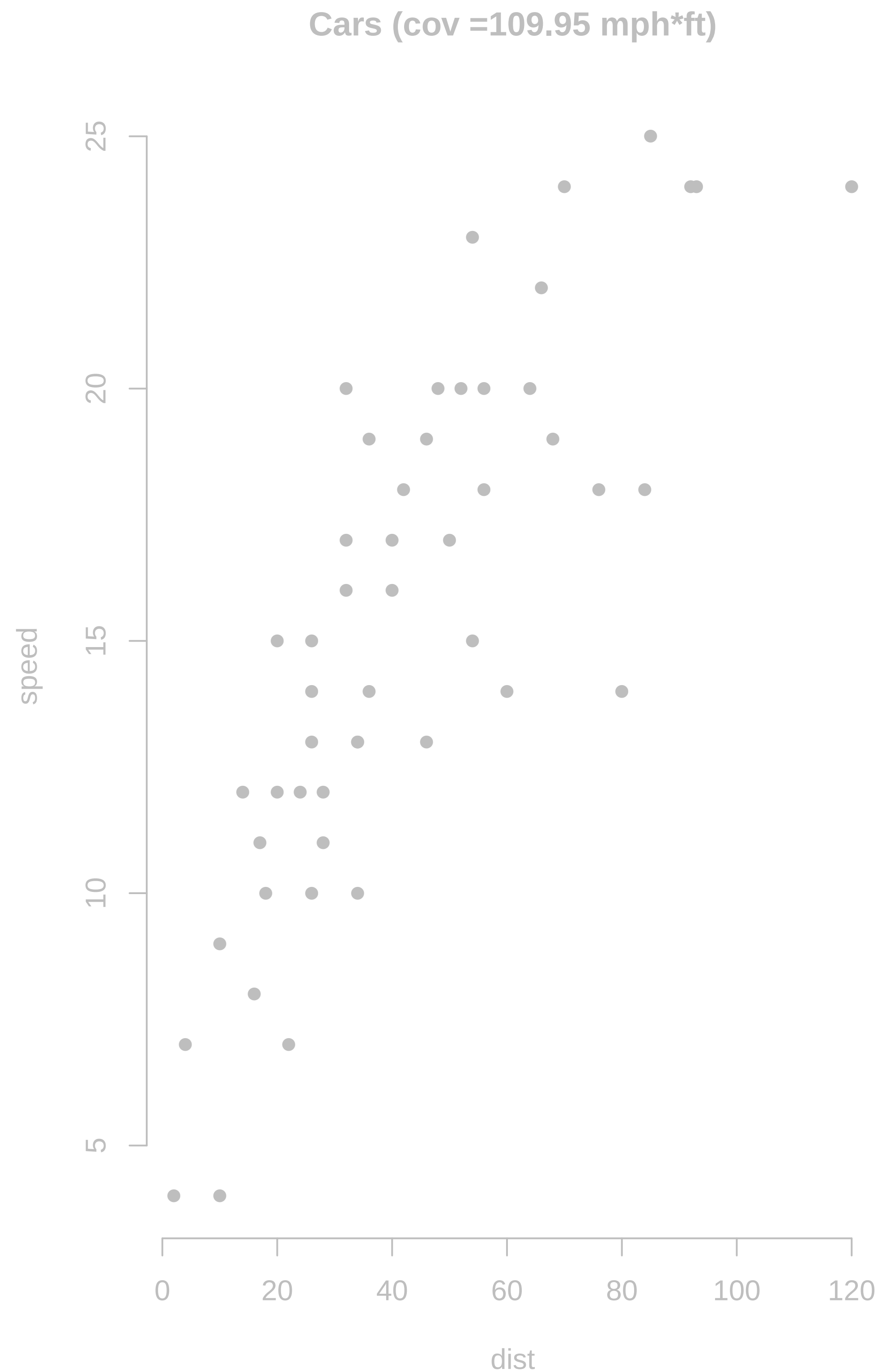
A problem with covariance

Covariance has units!

(x−unit times y−unit)

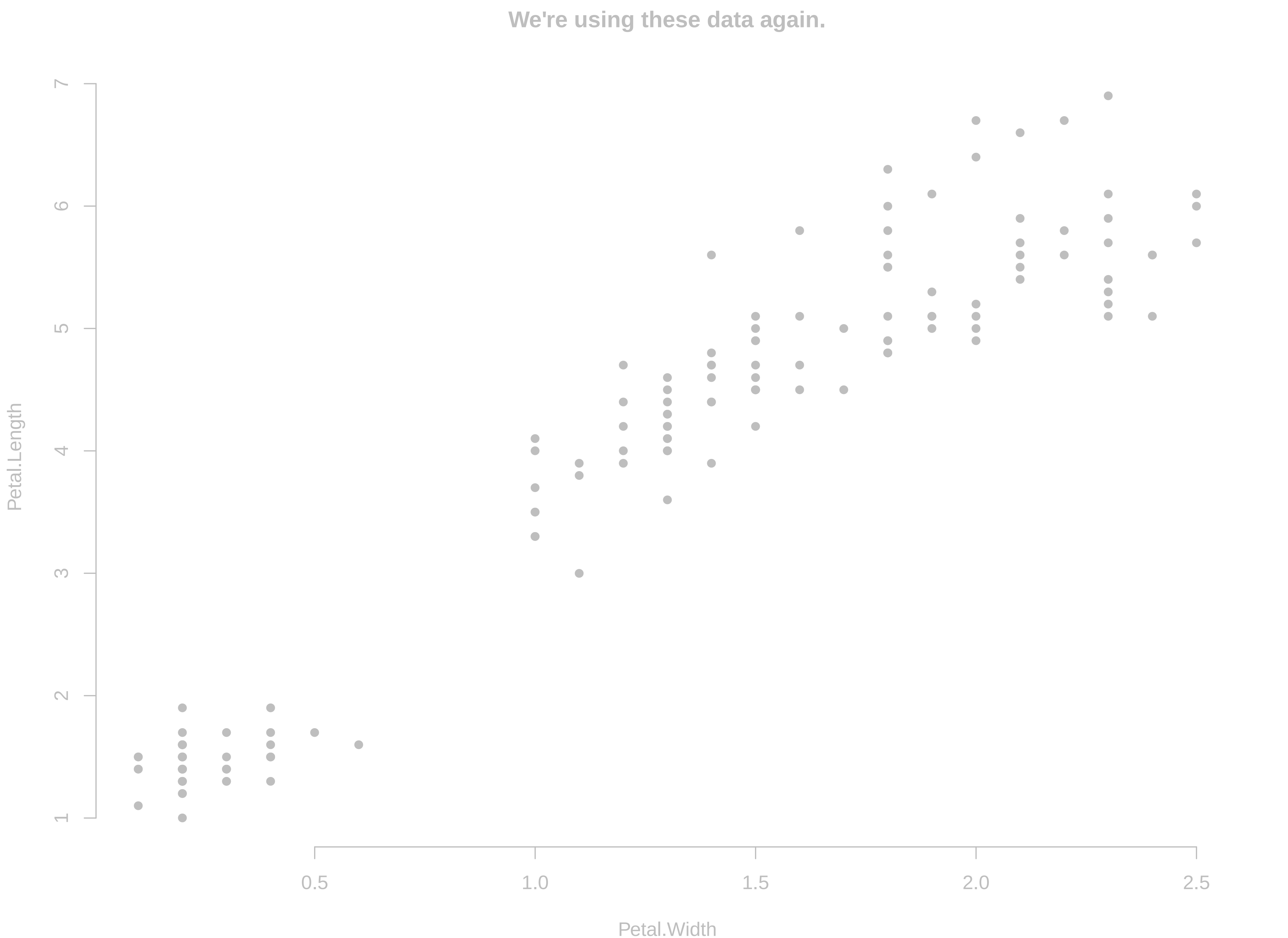
Which relationship is stronger (more linear)?

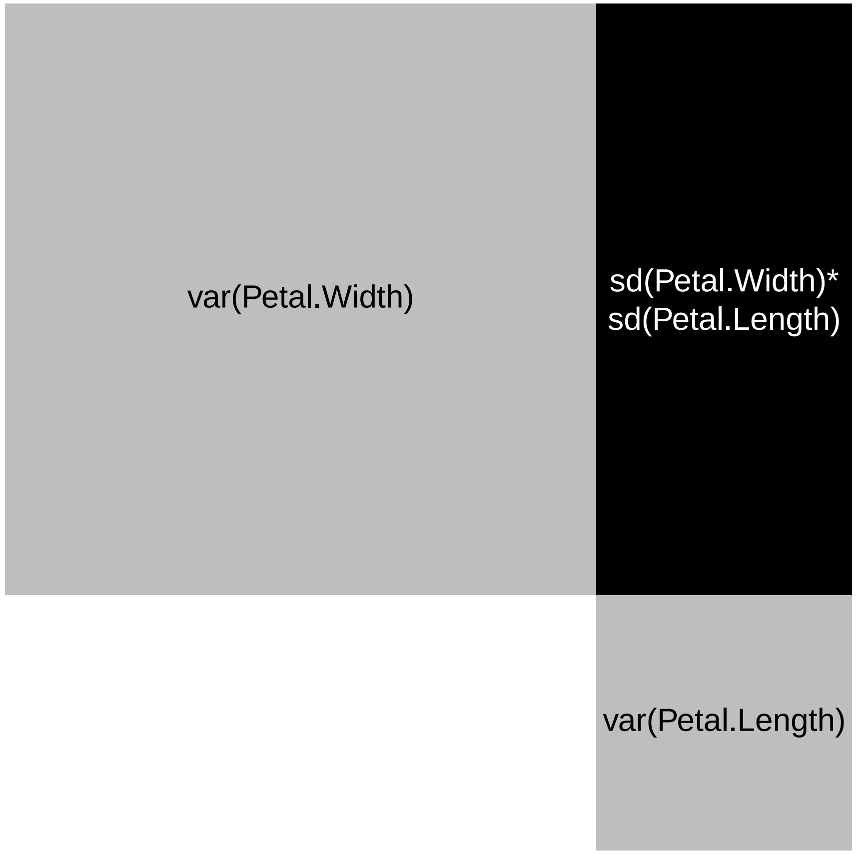

Irises (cov = 1.3 cm^2)
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

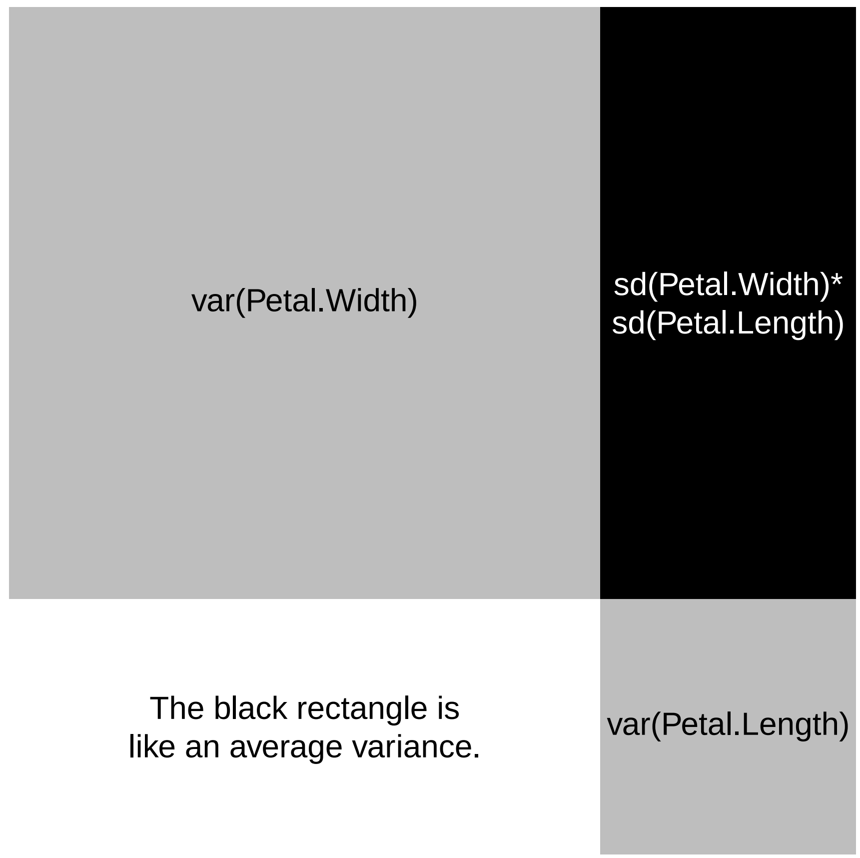

Cars (cov =109.95 mph*ft)
speed
dist
5
10
15
20
25
0
20
40
60
80
100
120

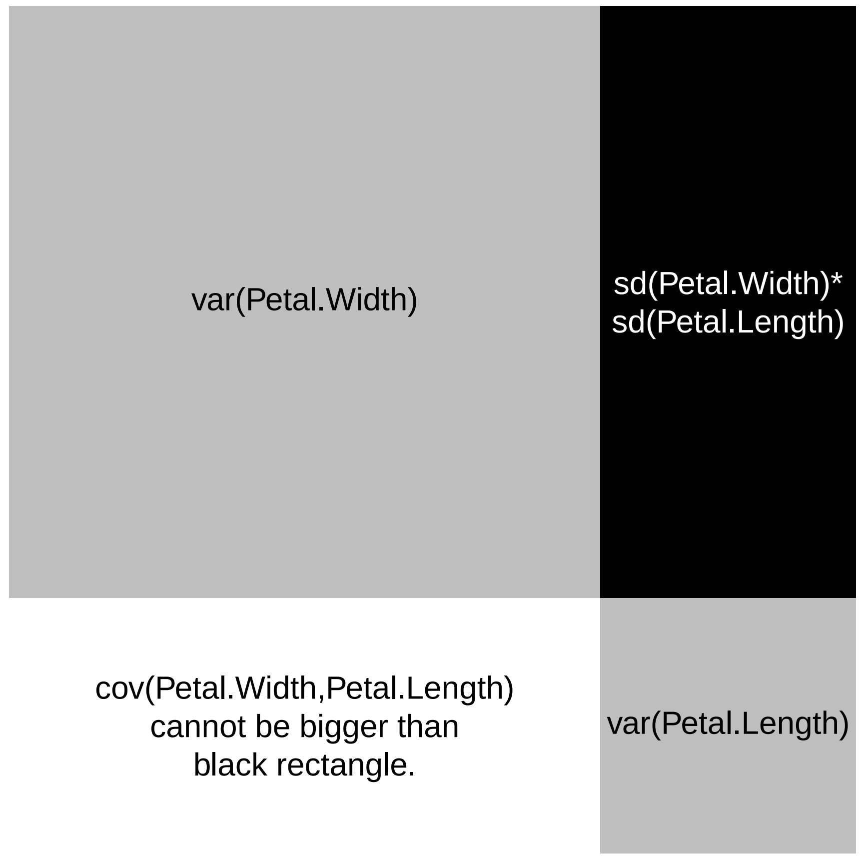
Oh noes!

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

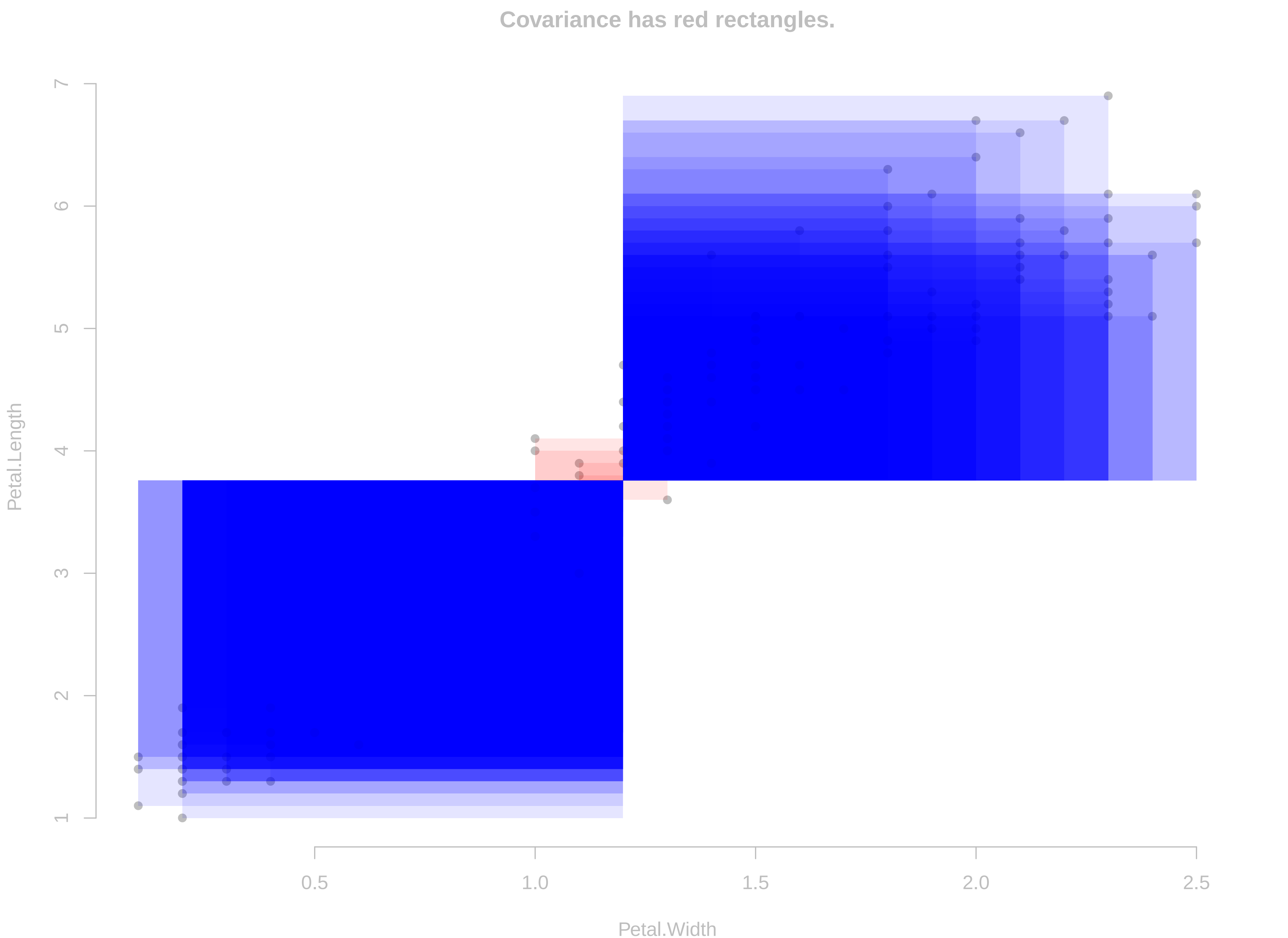

We're using these data again.
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

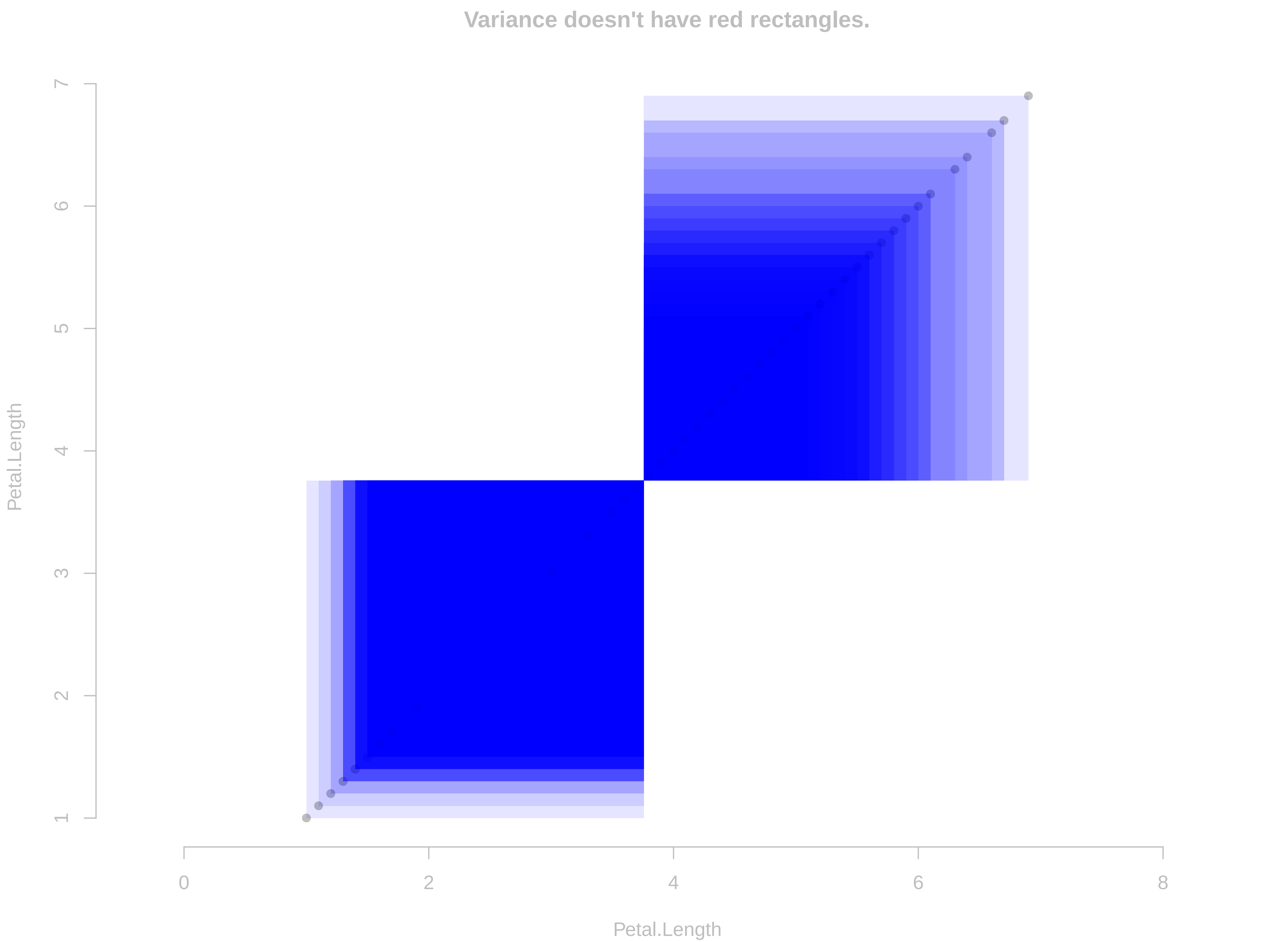

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
var(Petal.Length)

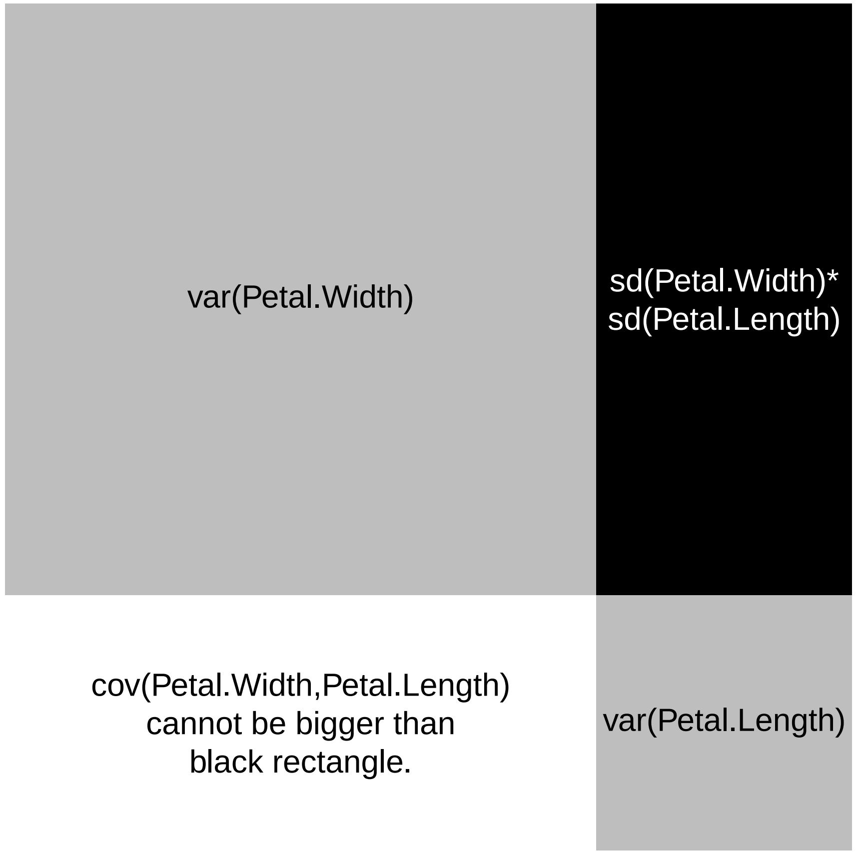

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
The black rectangle is
like an average variance.
var(Petal.Length)


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

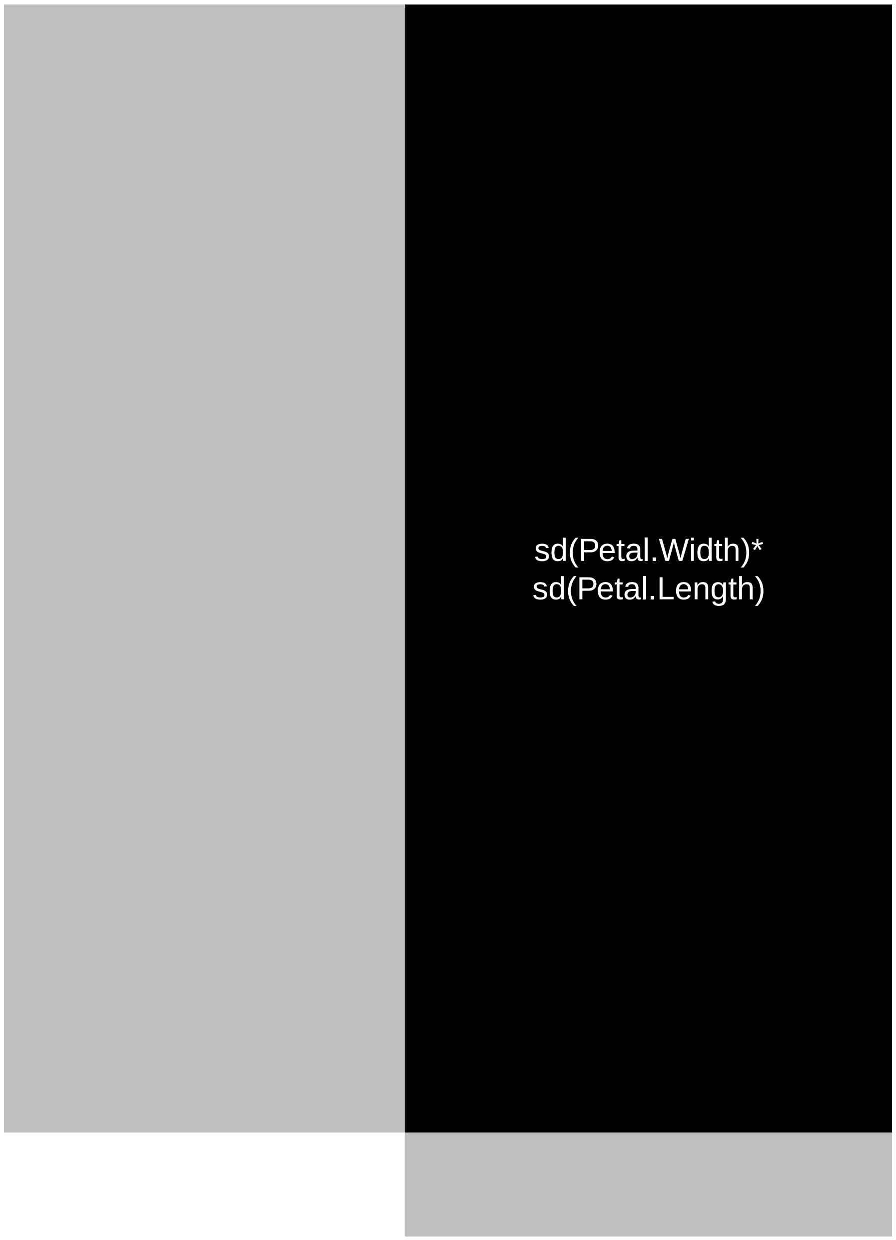
Why?


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

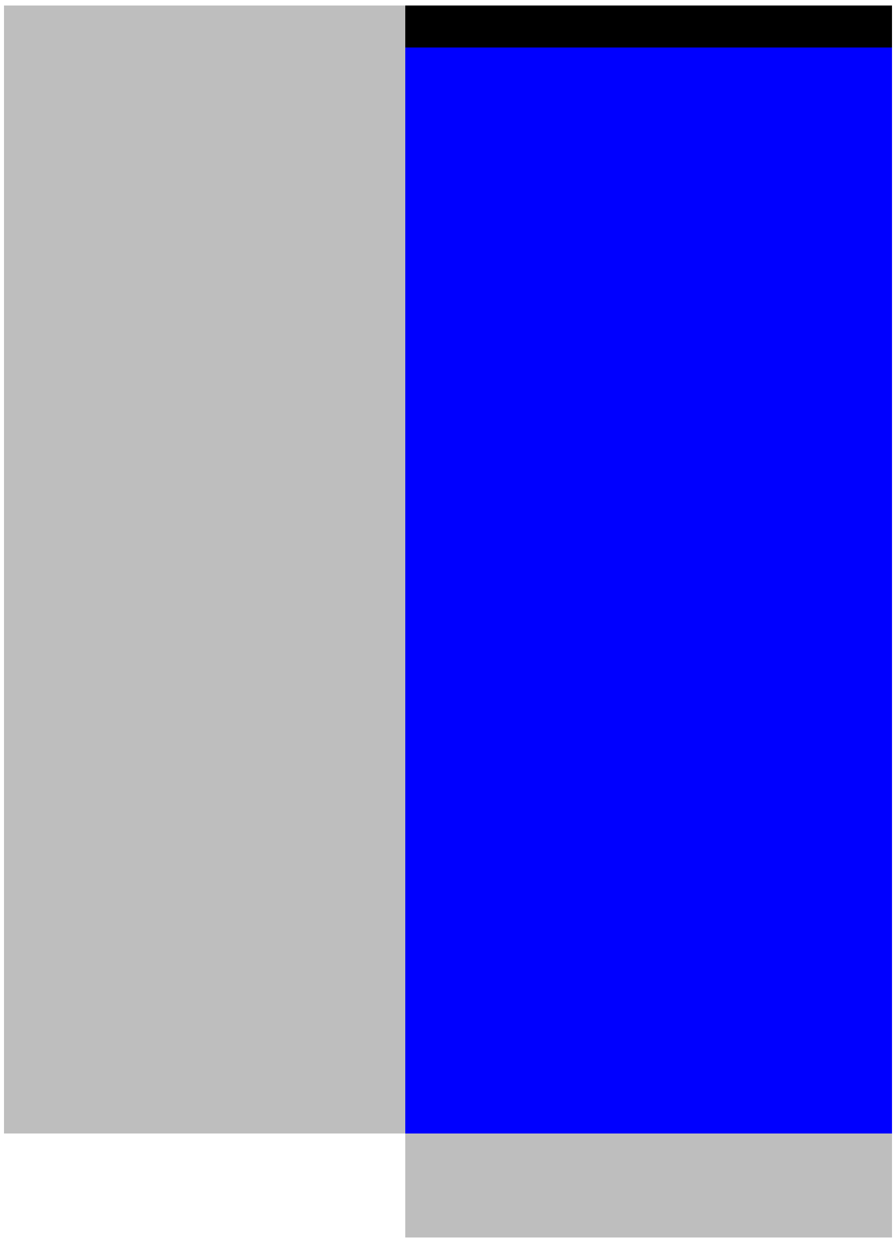

Variance doesn't have red rectangles.
Petal.Length
Petal.Length
0
2
4
6
8
1
2
3
4
5
6
7


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.

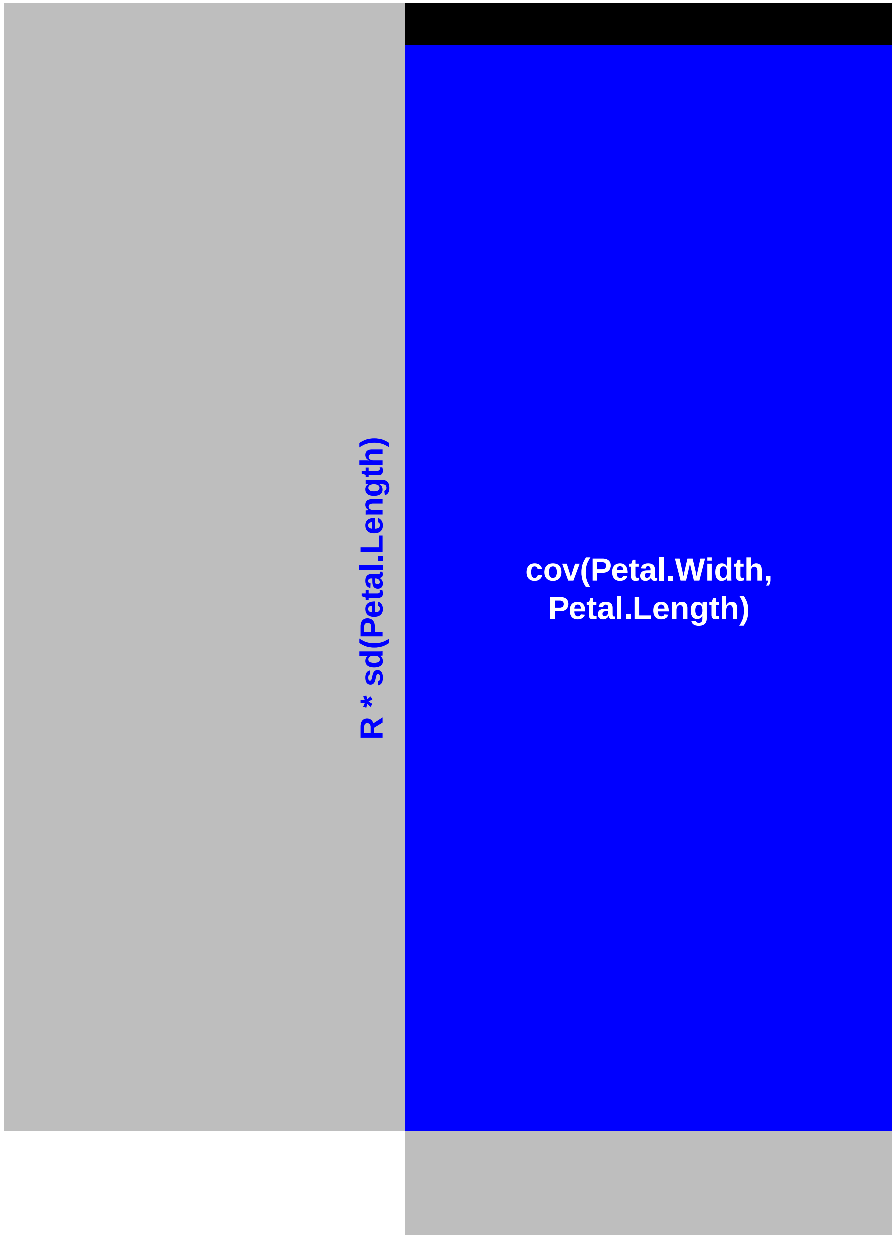


Squish covariance vertically into the rectangle.

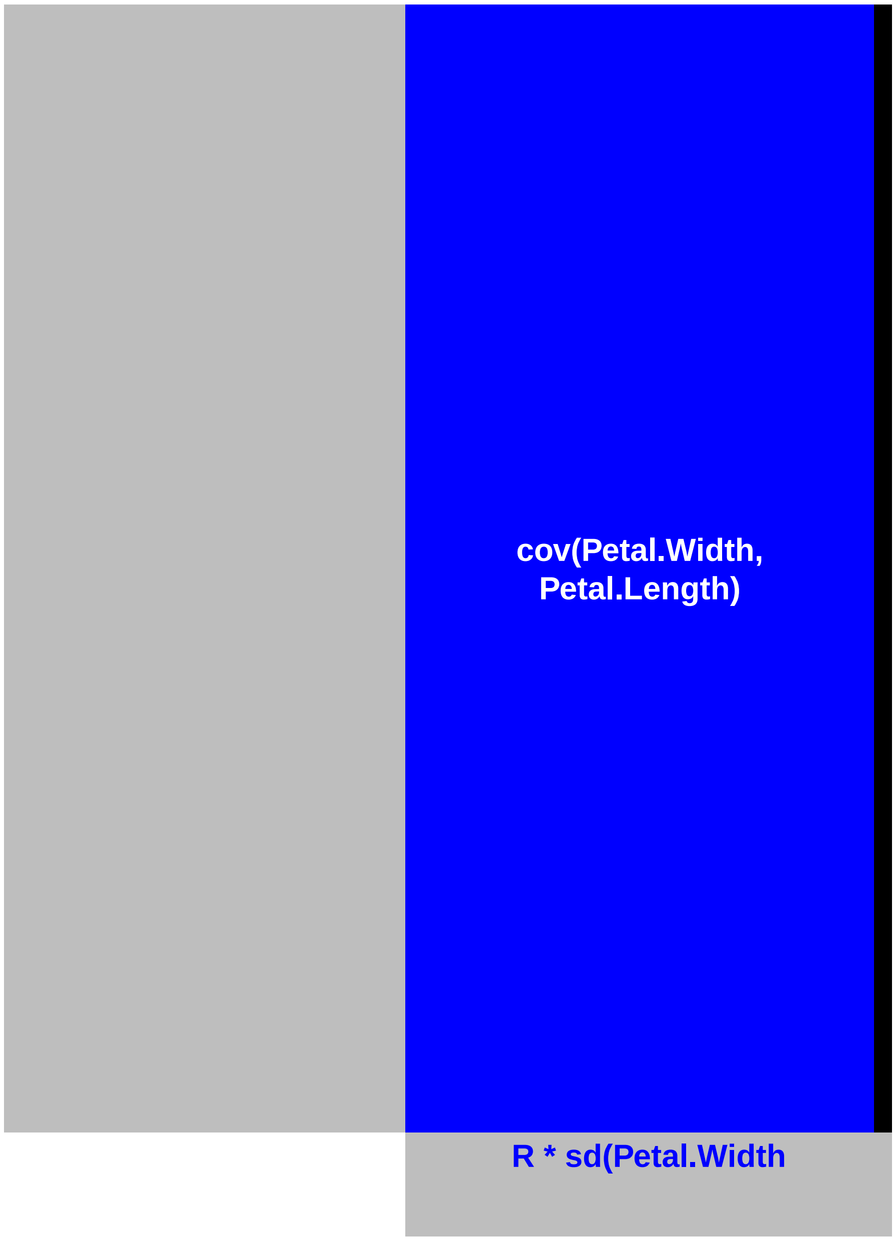


Correlation (R)
is the ratio of
the small rectangle
to the big rectangle.

Squish covariance vertically into the rectangle.

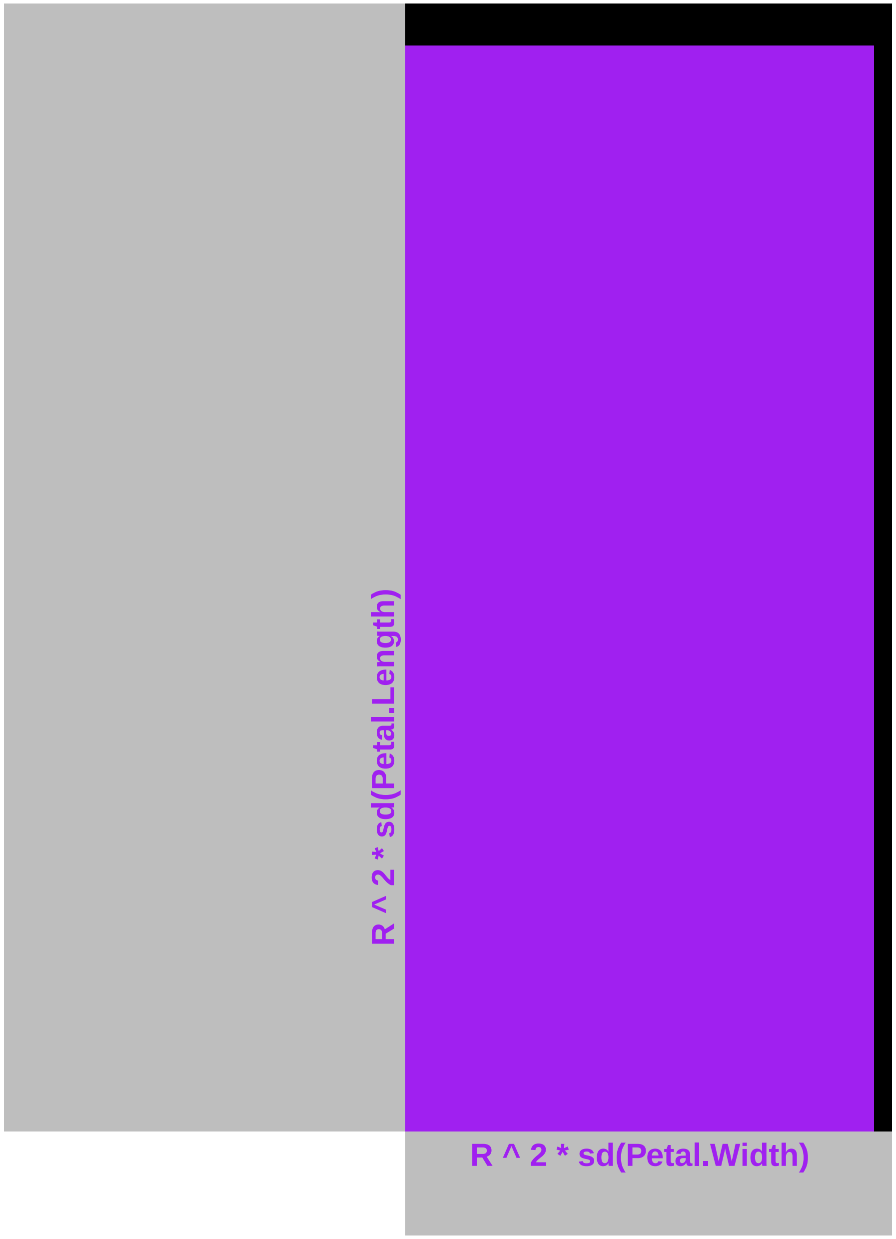


Squish covariance horizontally into the rectangle.

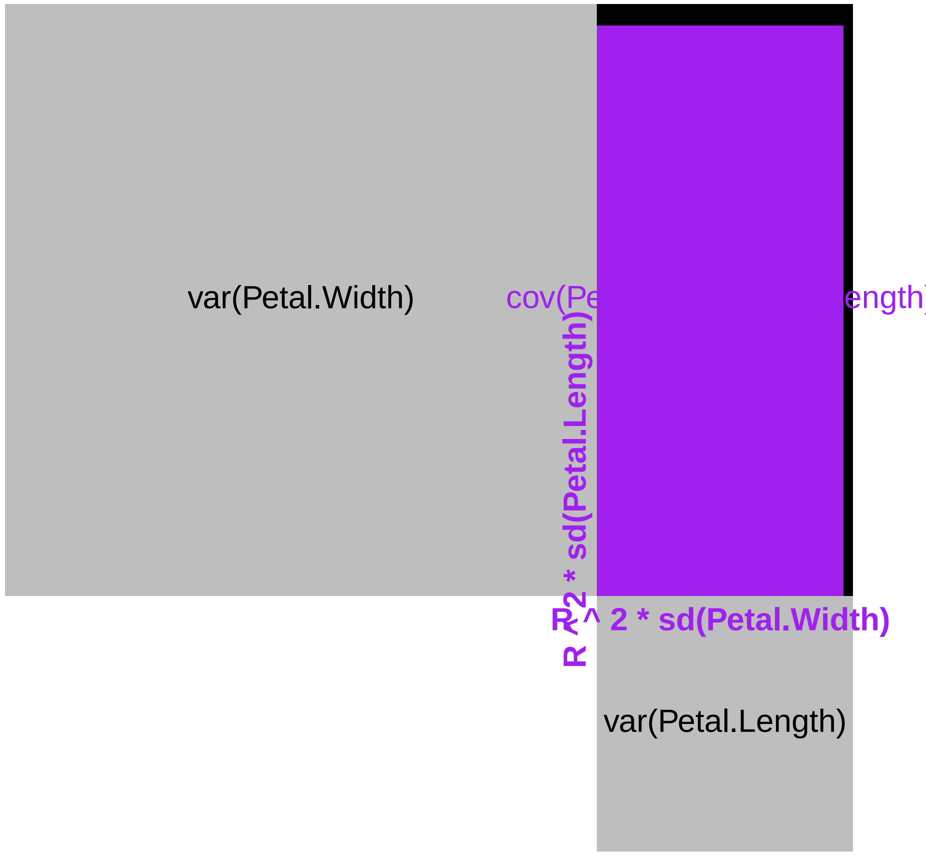


People like to talk about R–squared.

Intersect the two squished covariance rectangles.

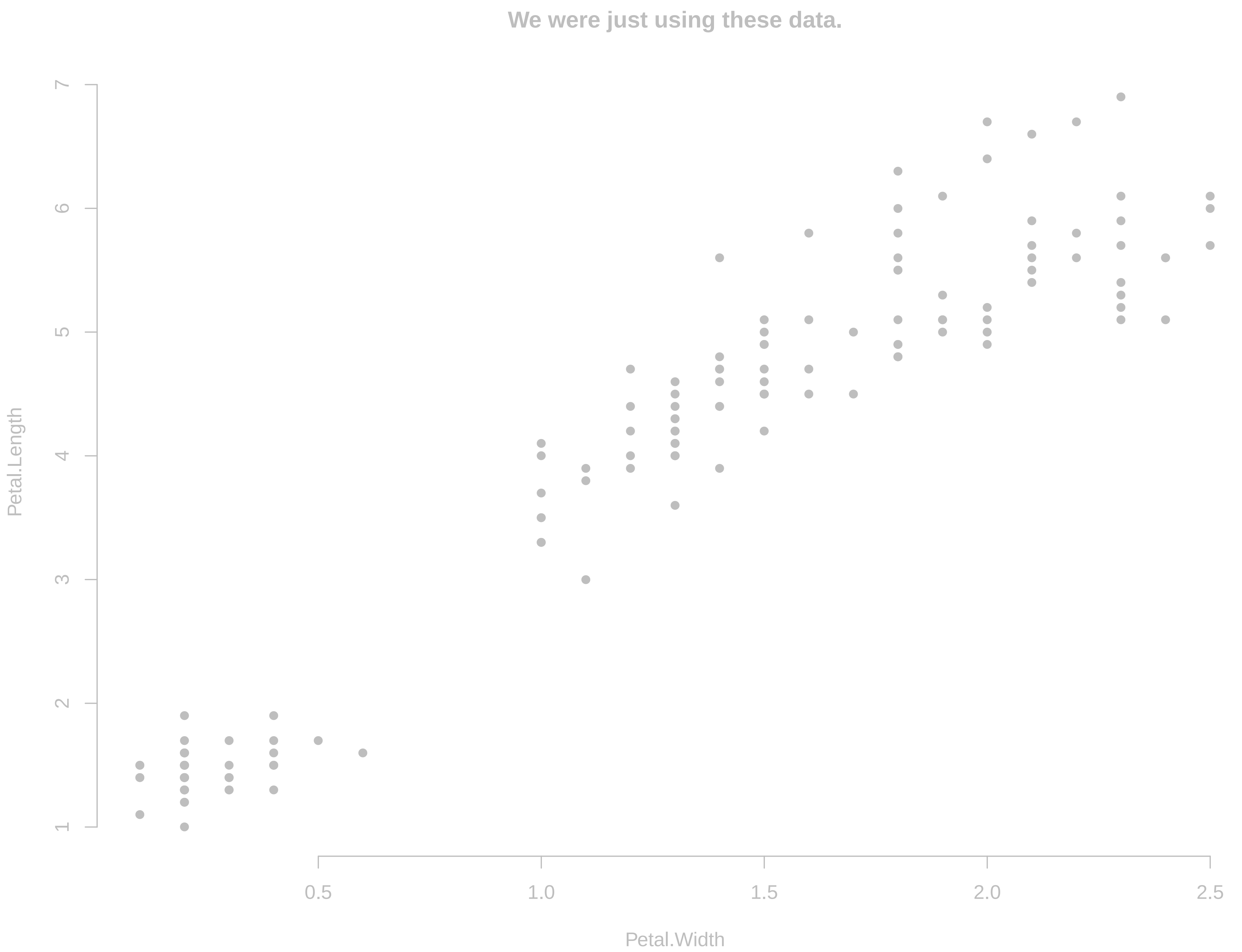


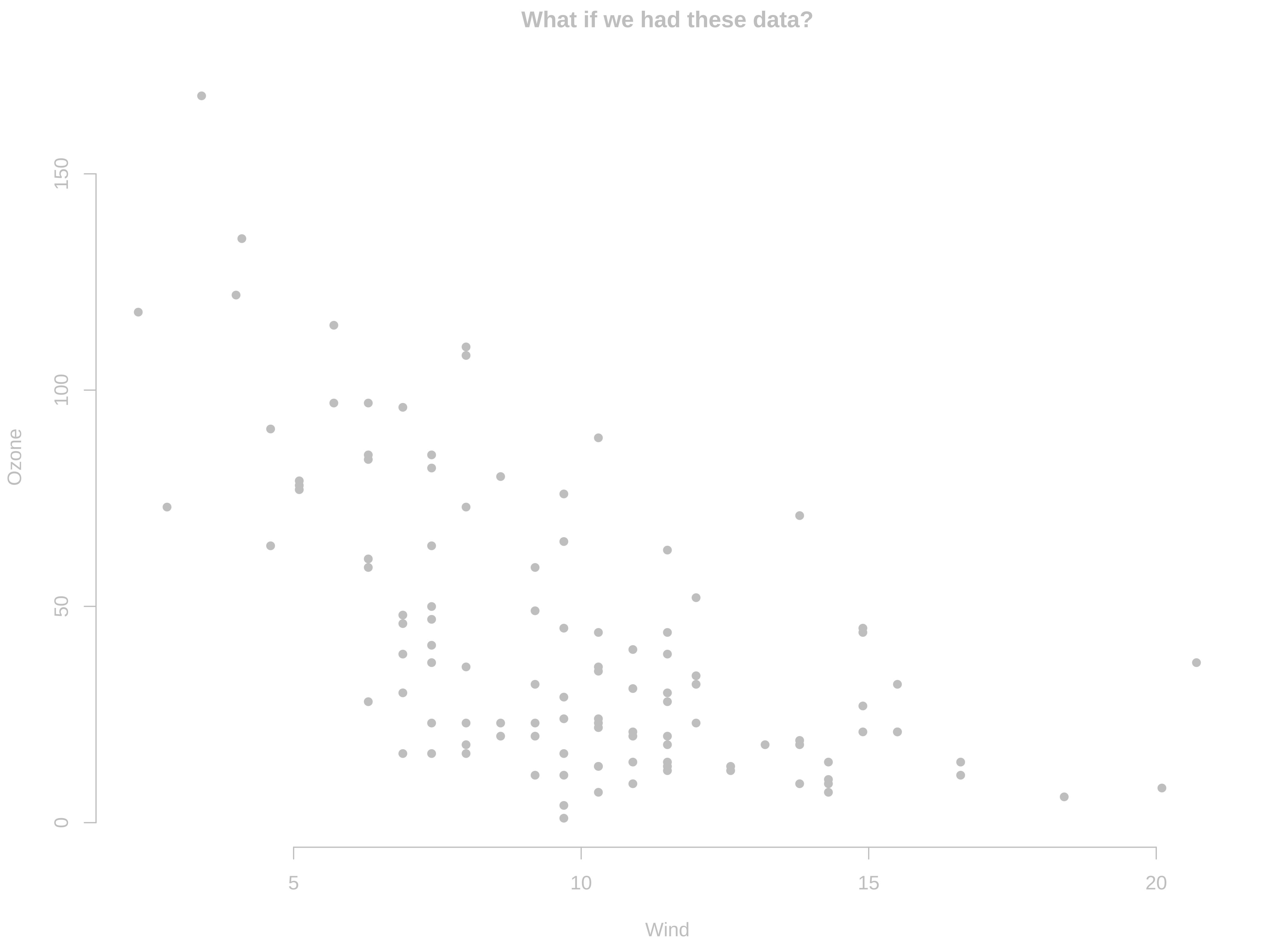
Intersect the two squished covariance rectangles.

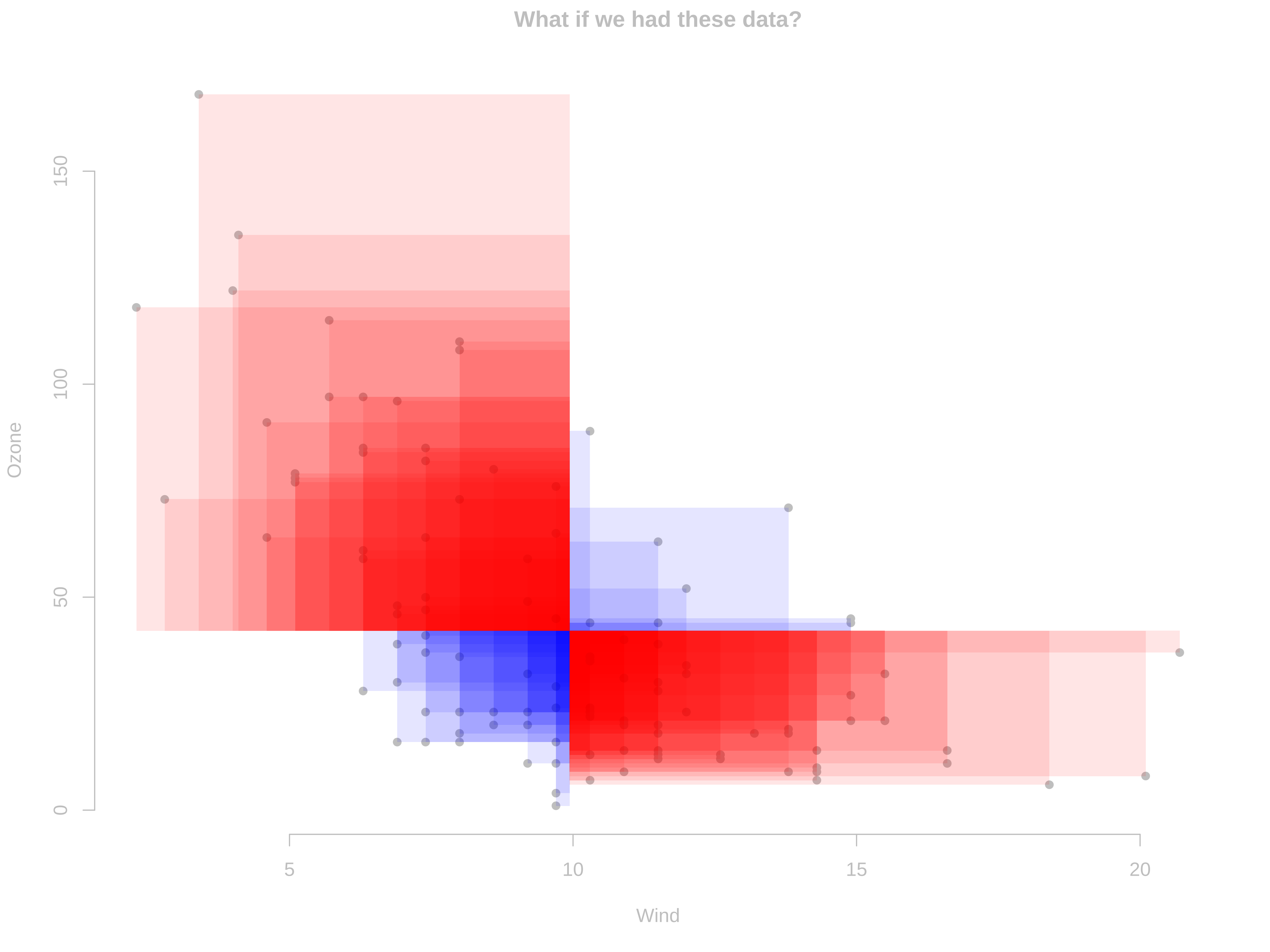


That was for very positive (blue) covariances.

What if covariance is negative (red)?

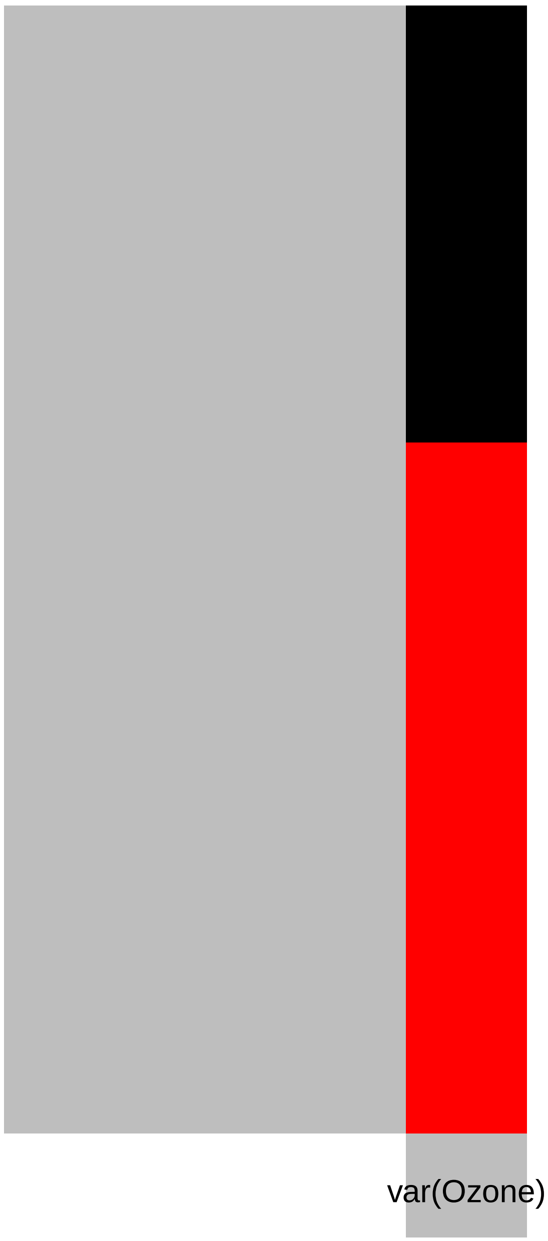

We were just using these data.
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


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

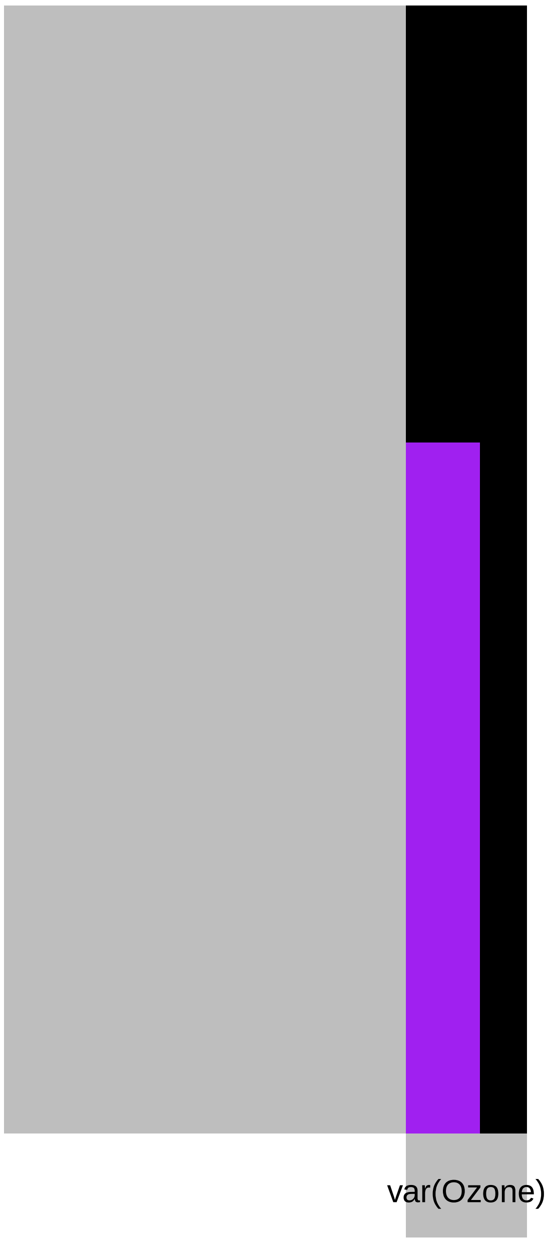


What if we had these data?
Ozone
0
50
100
150
Wind
5
10
15
20

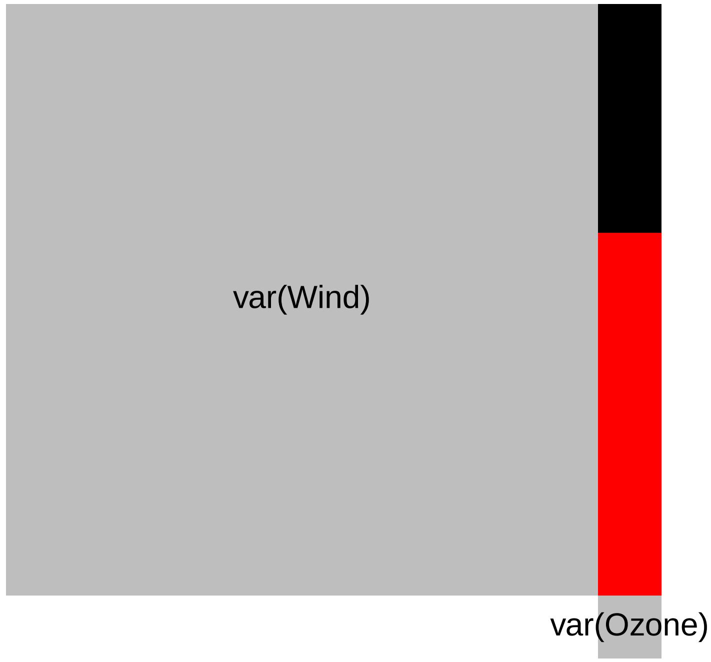
R is the same, just negative.



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 high
and x is high,
y should be high.

(b1 is very positive.)

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

(b1 is very negative.)

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

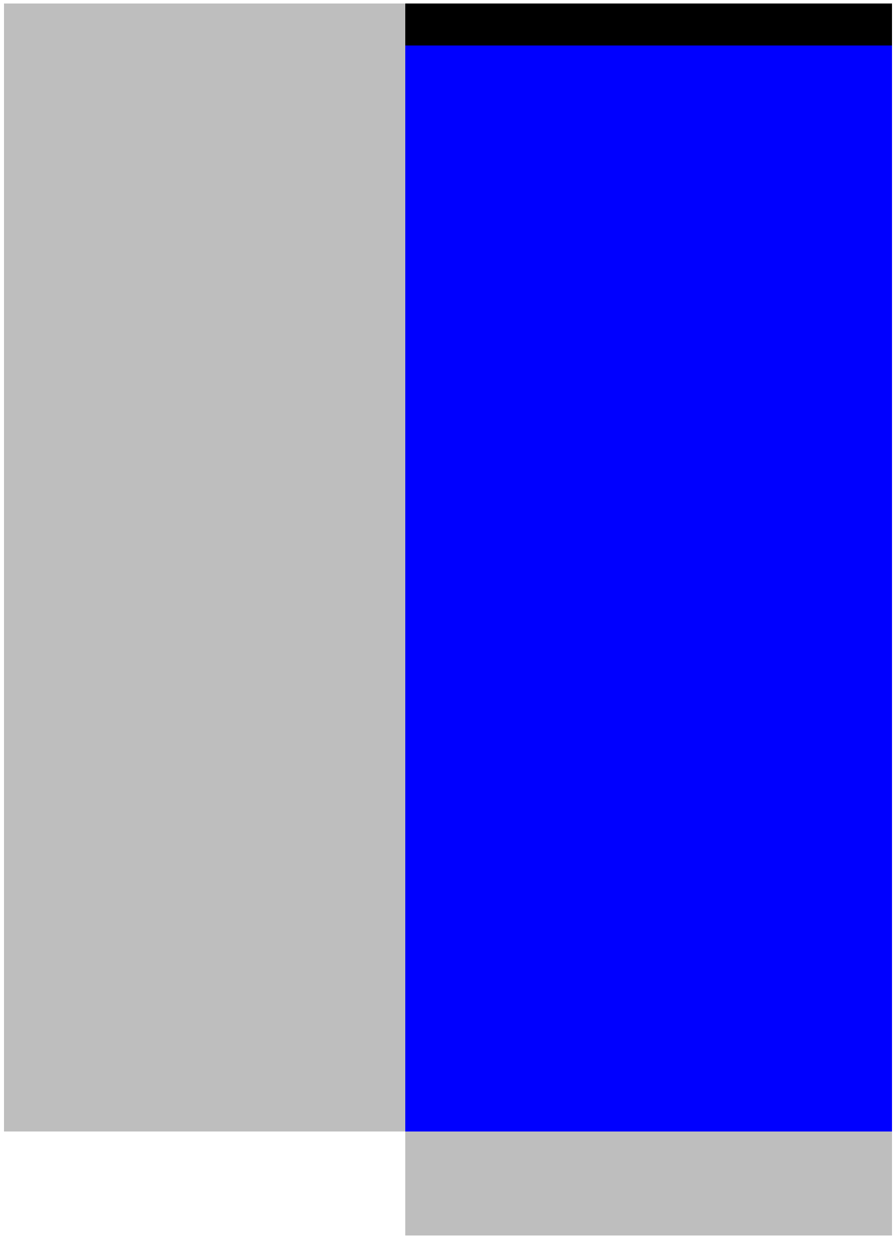
(b1 is around zero.)

Let's think about units again.

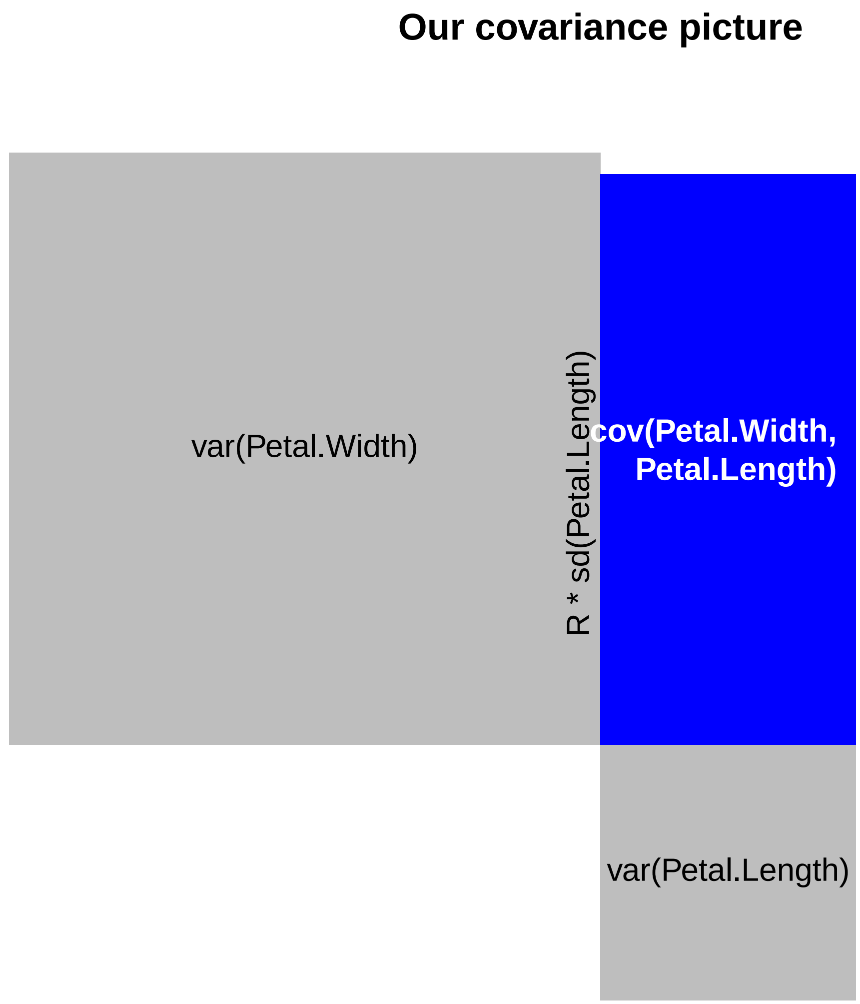
Covariance is an area; its unit is the product of the x and y units.

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

Correlation is a ratio of areas with the same units.

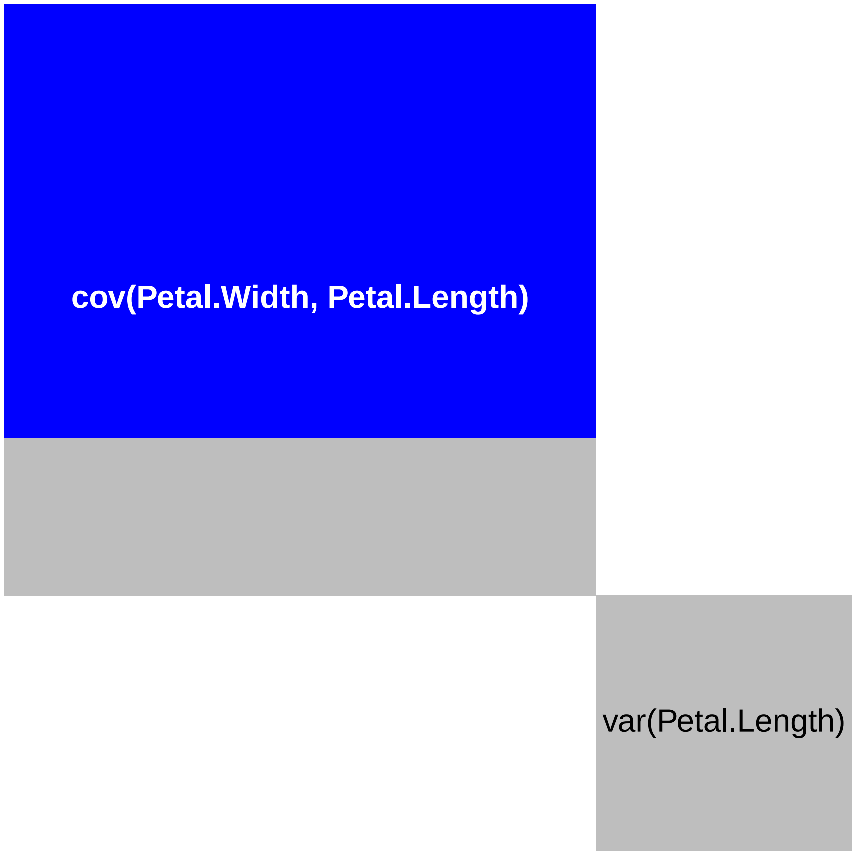


The unit of b1 must be y–unit/x–unit.

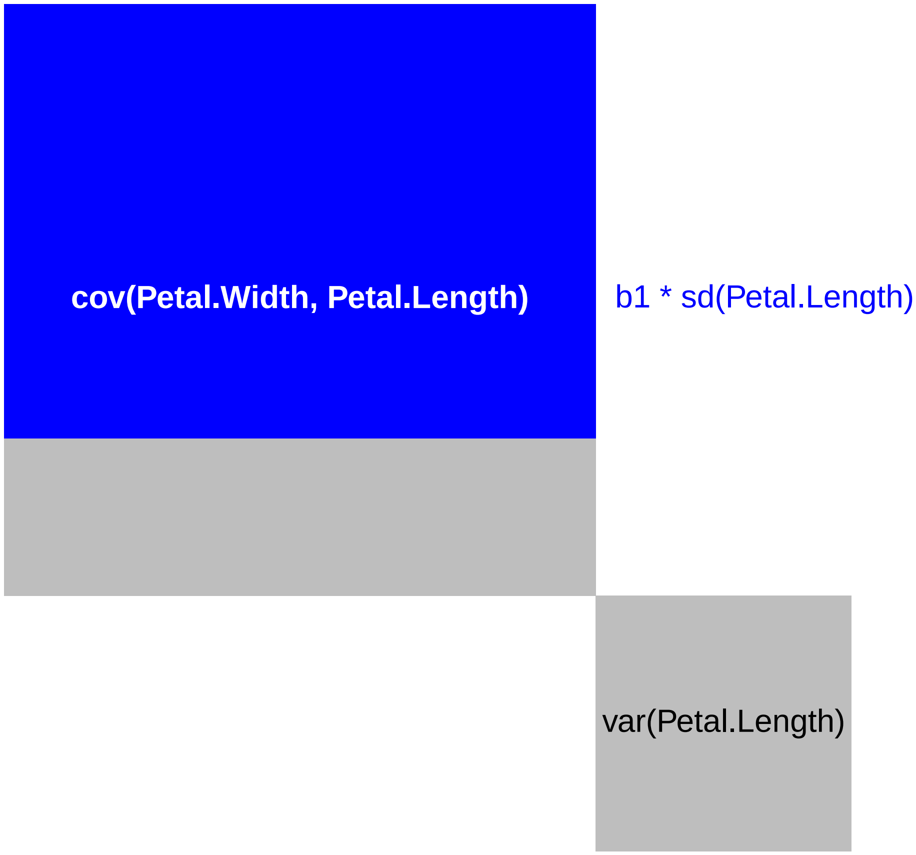


Our covariance picture
var(Petal.Width)
R * sd(Petal.Length)
cov(Petal.Width,
Petal.Length)
var(Petal.Length)

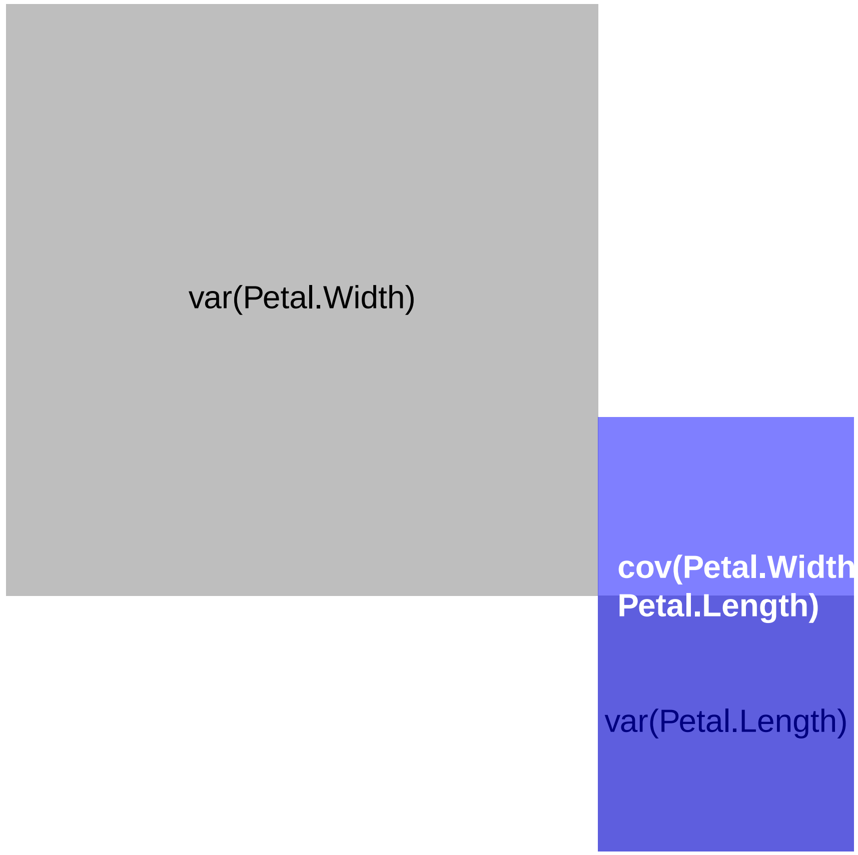
Lay the covariance over one of the variances instead.



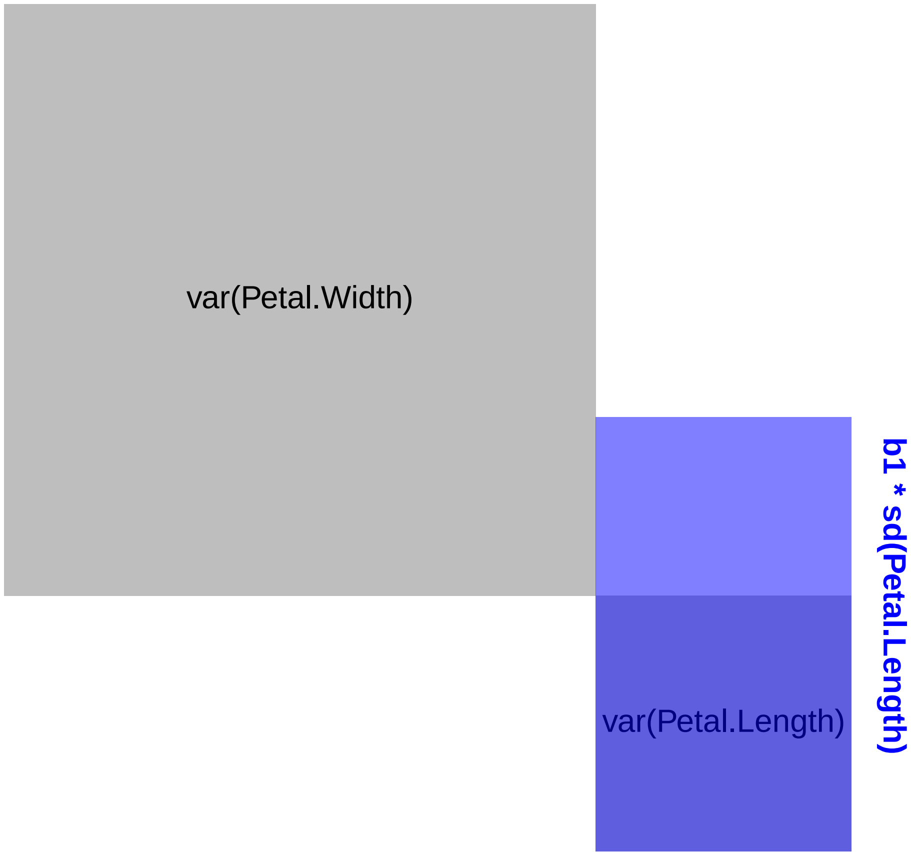
Petal.Width = b0 + b1 * Petal.Length



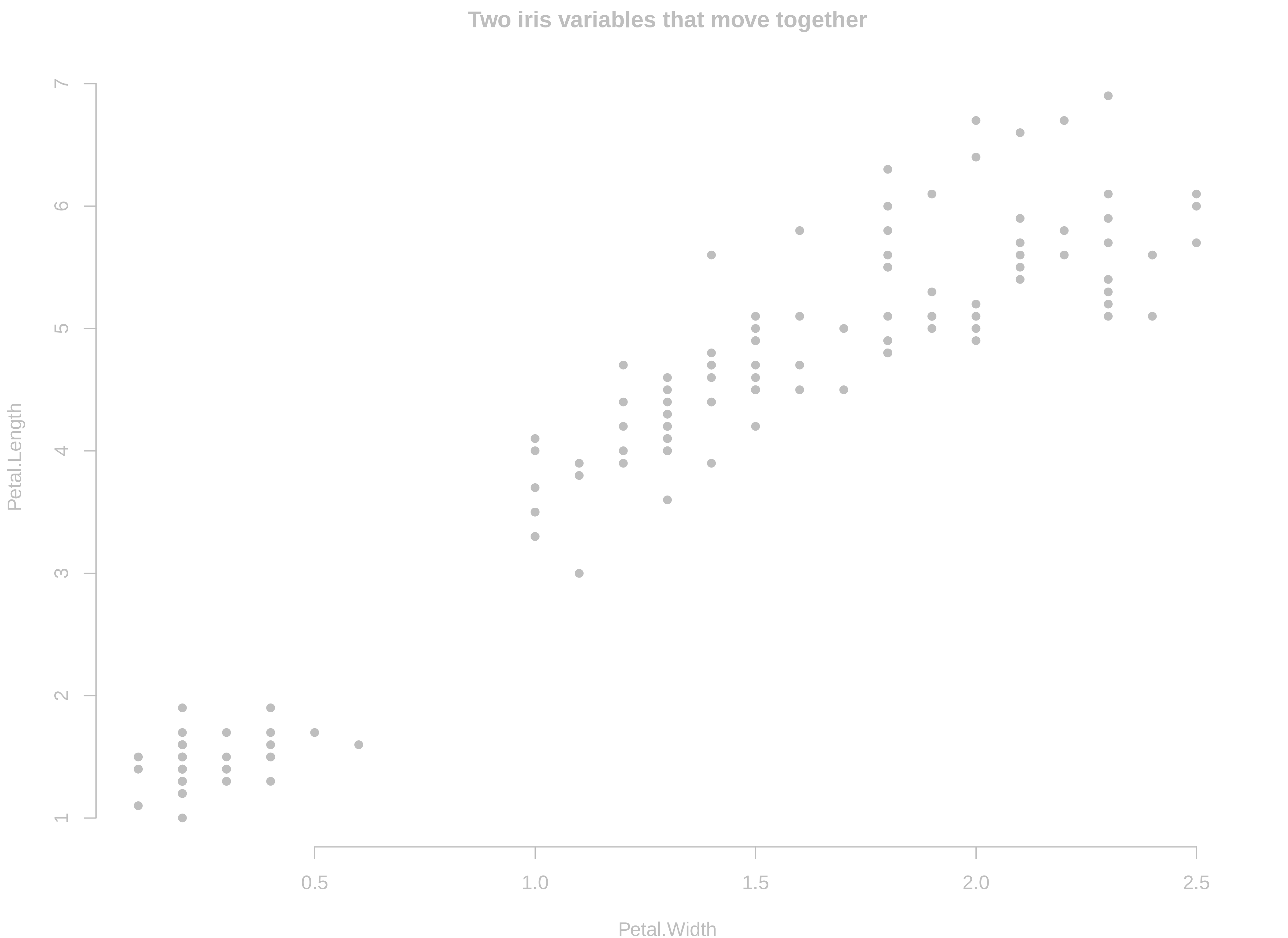
Lay the covariance over the other variance.

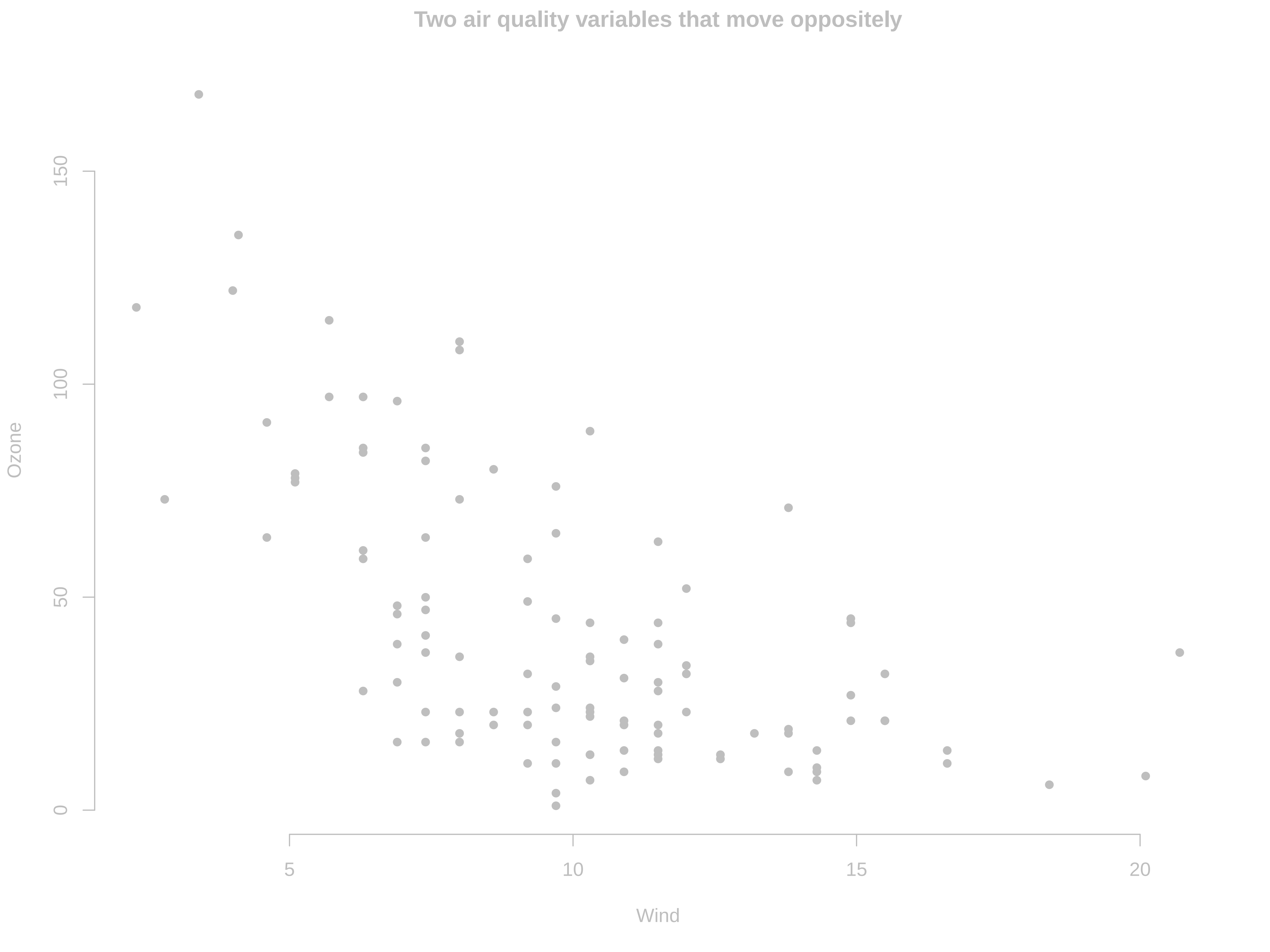


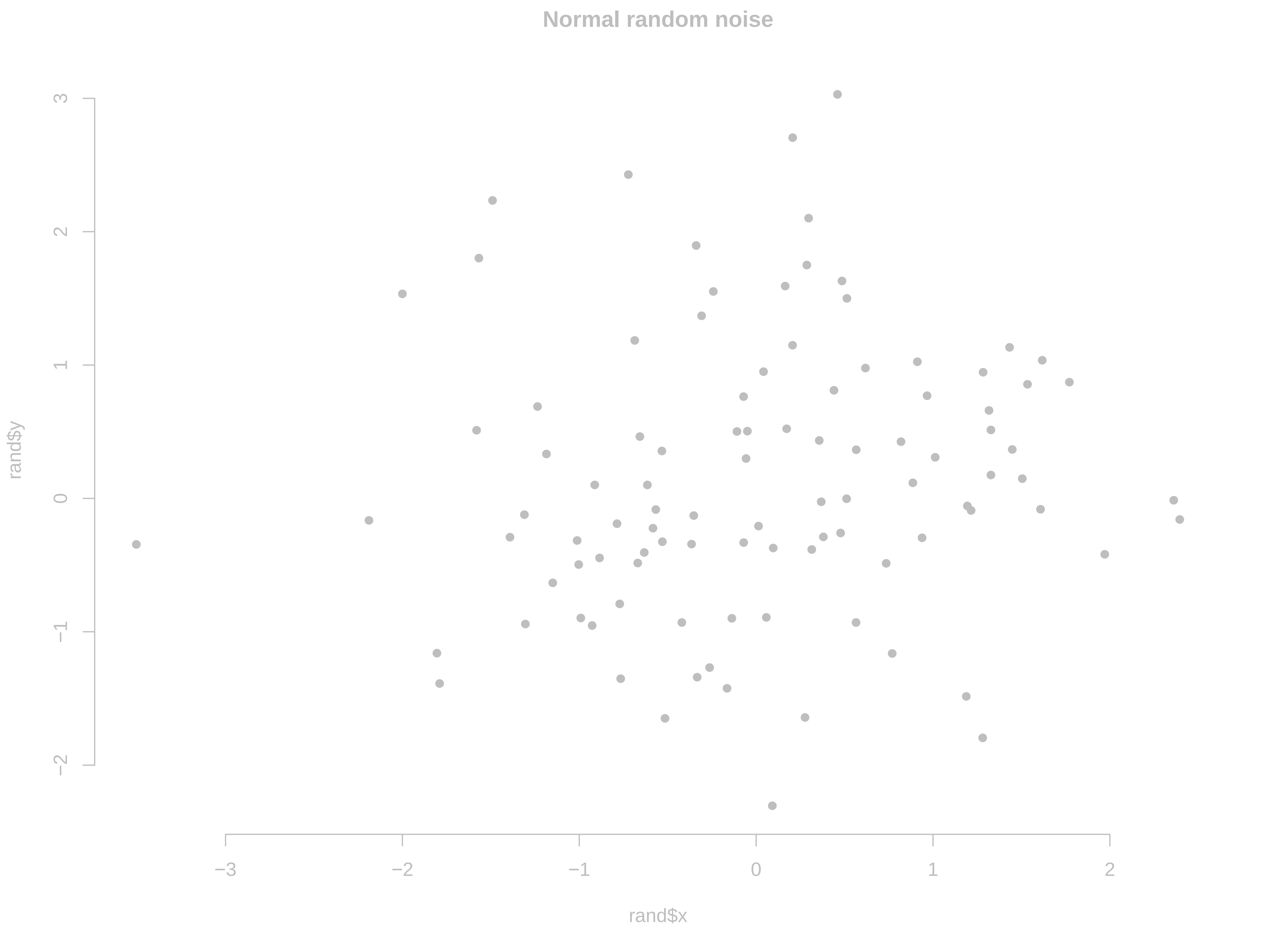
Petal.Length = b0 + b1 * Petal.Width



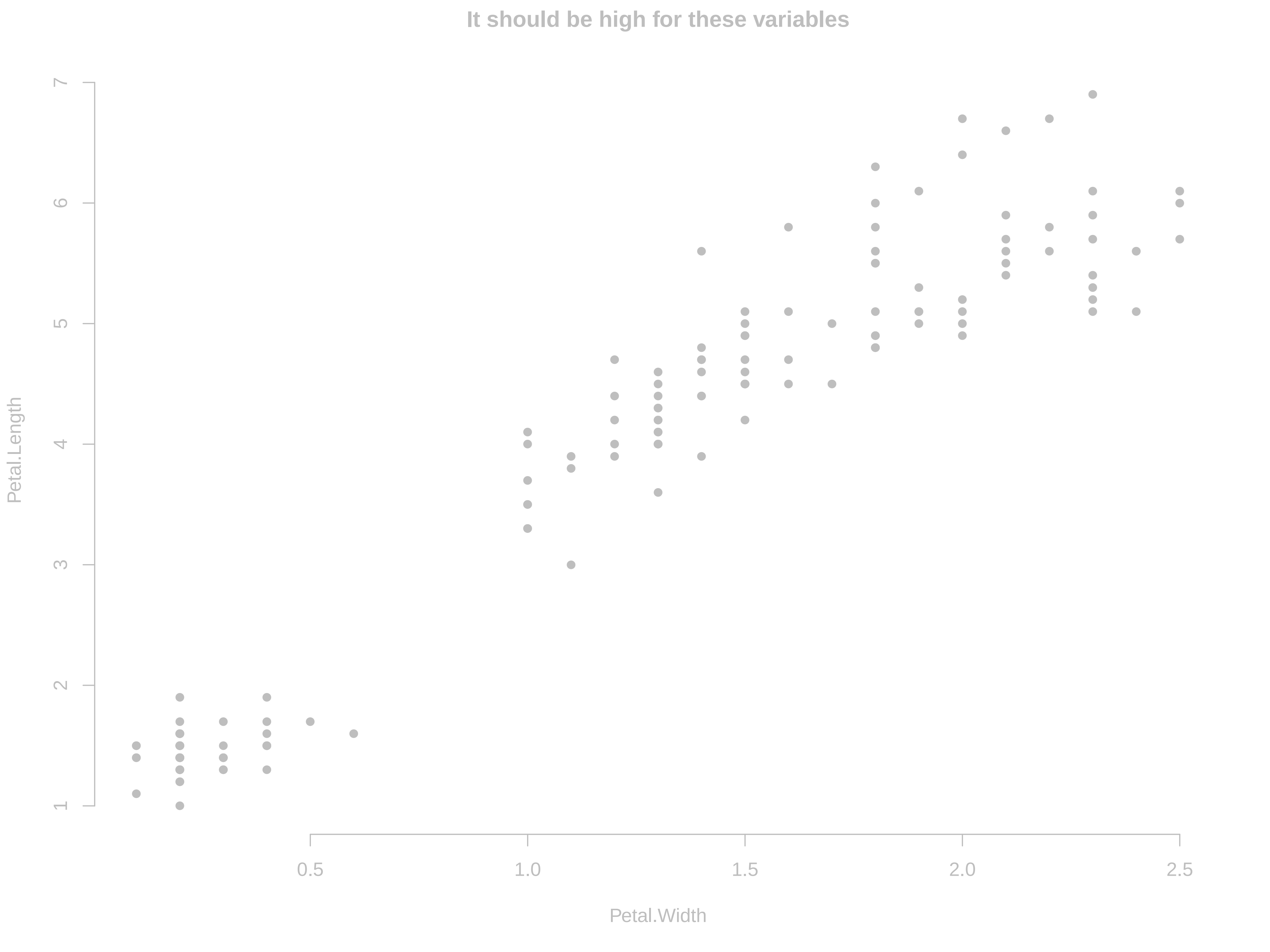
Let's go over that again.

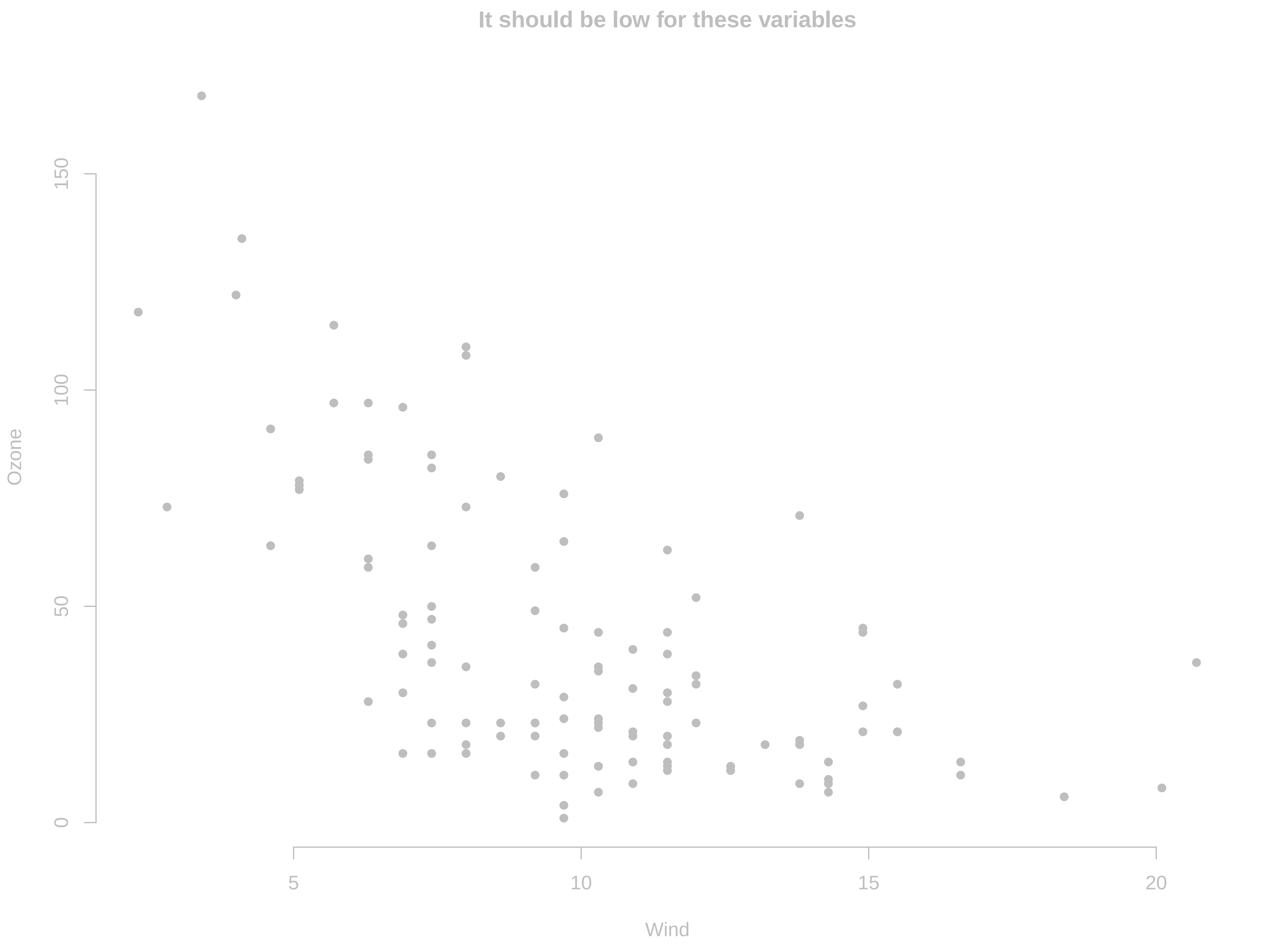

Two iris variables that move together
Petal.Length
1
2
3
4
5
6
7
Petal.Width
0.5
1.0
1.5
2.0
2.5


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

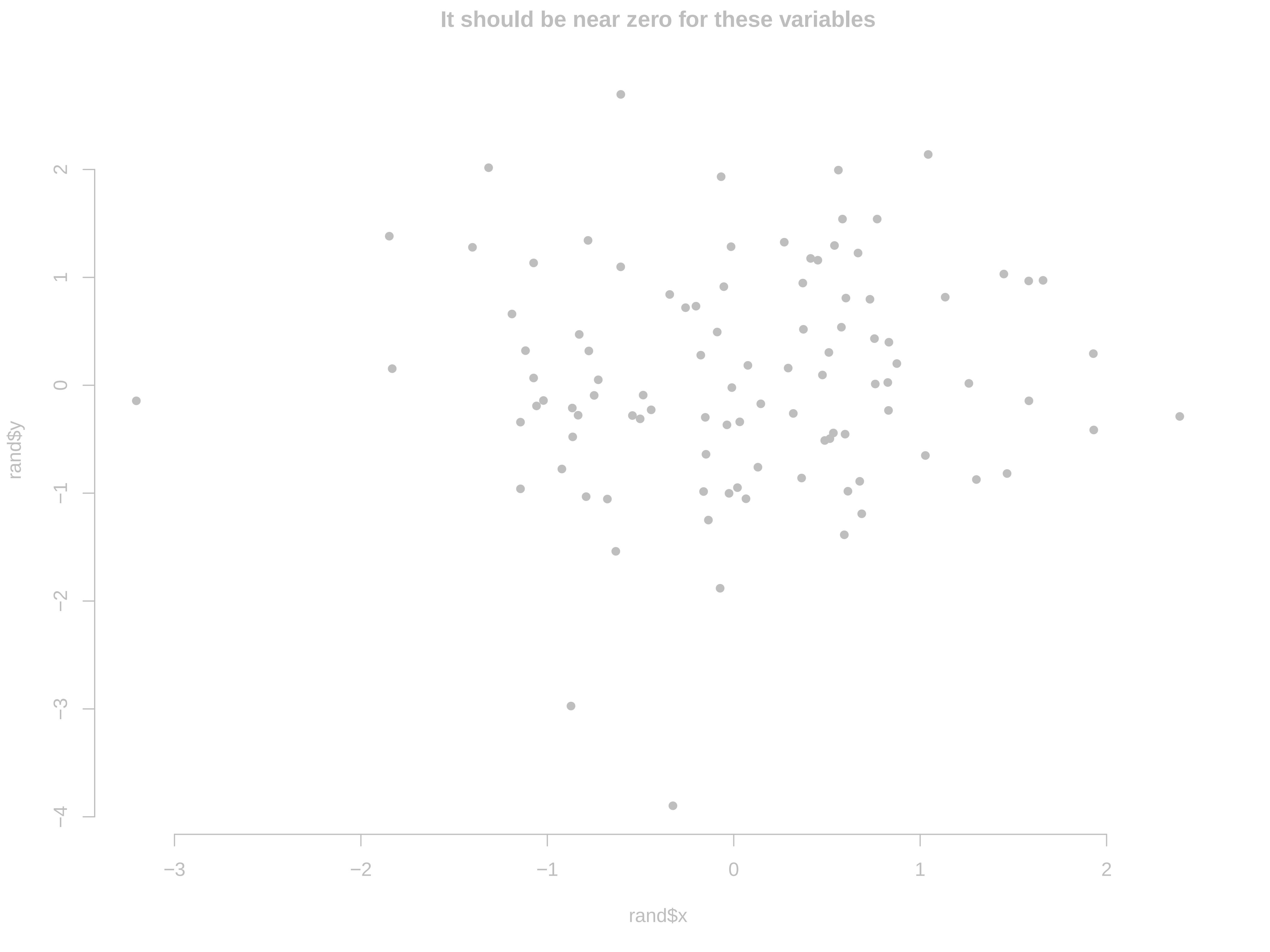

Normal random noise
3
2
1
0
−1
−2
rand$y
−3
−2
−1
0
1
2
rand$x

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

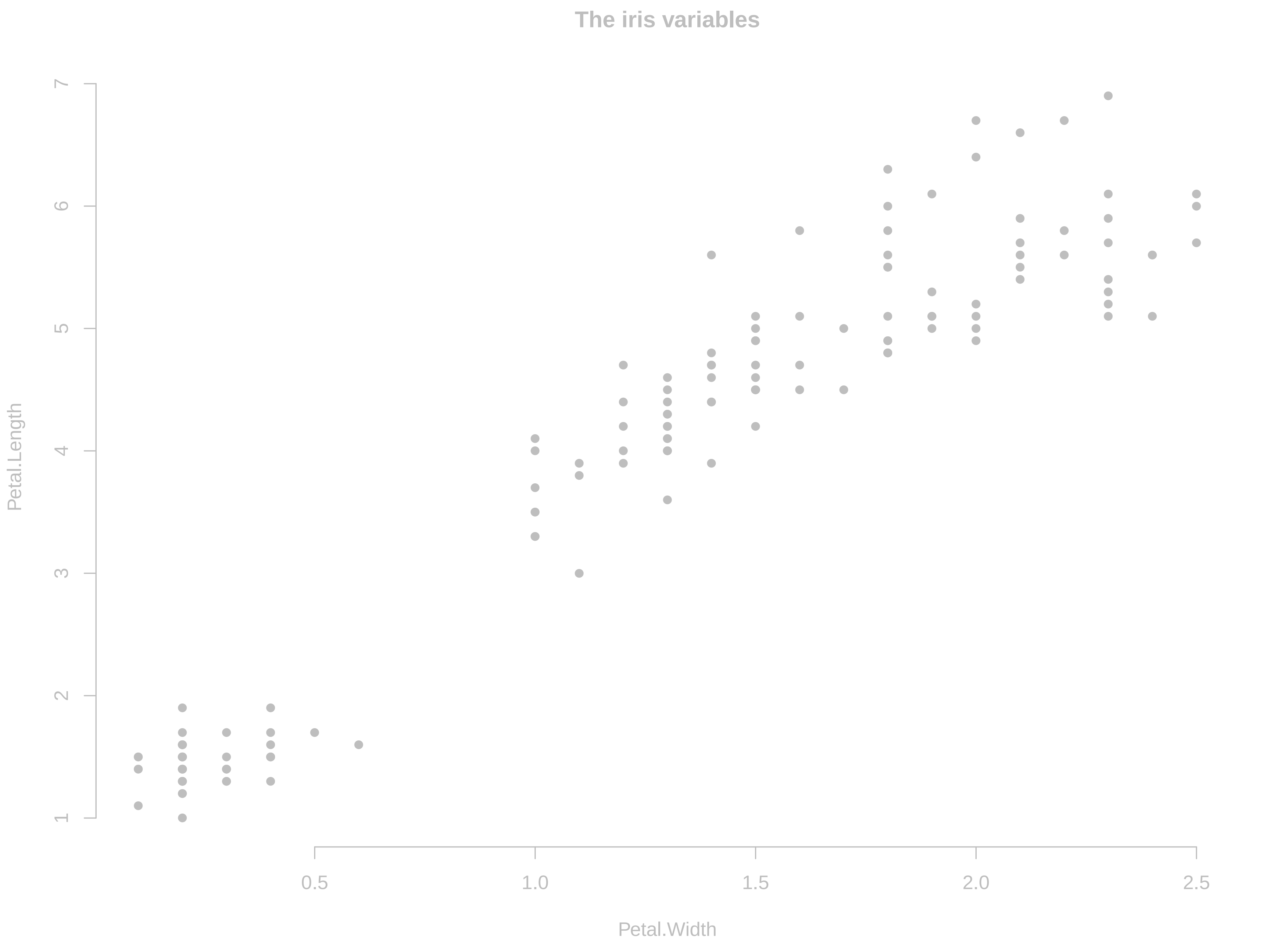

It should be high for these variables
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

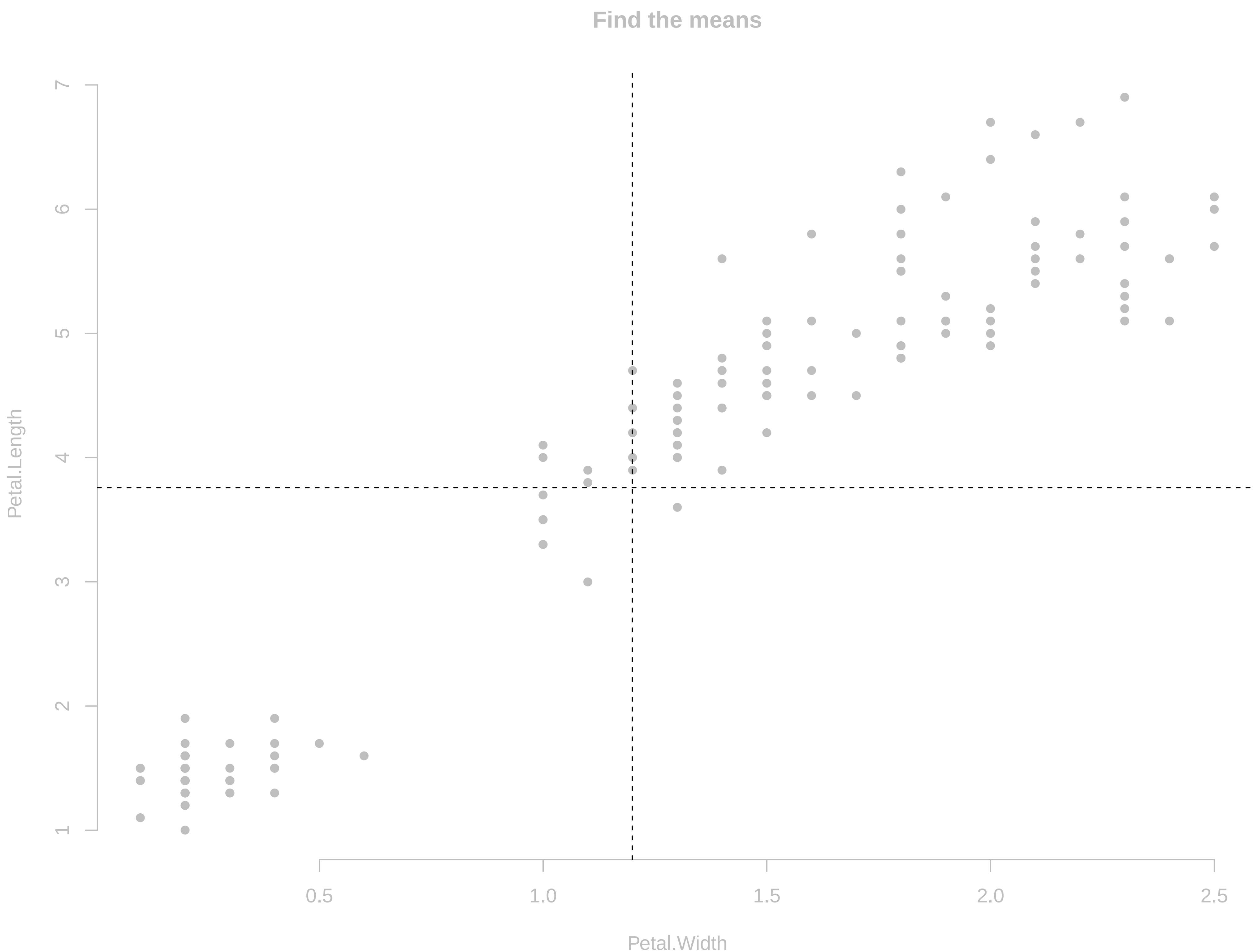


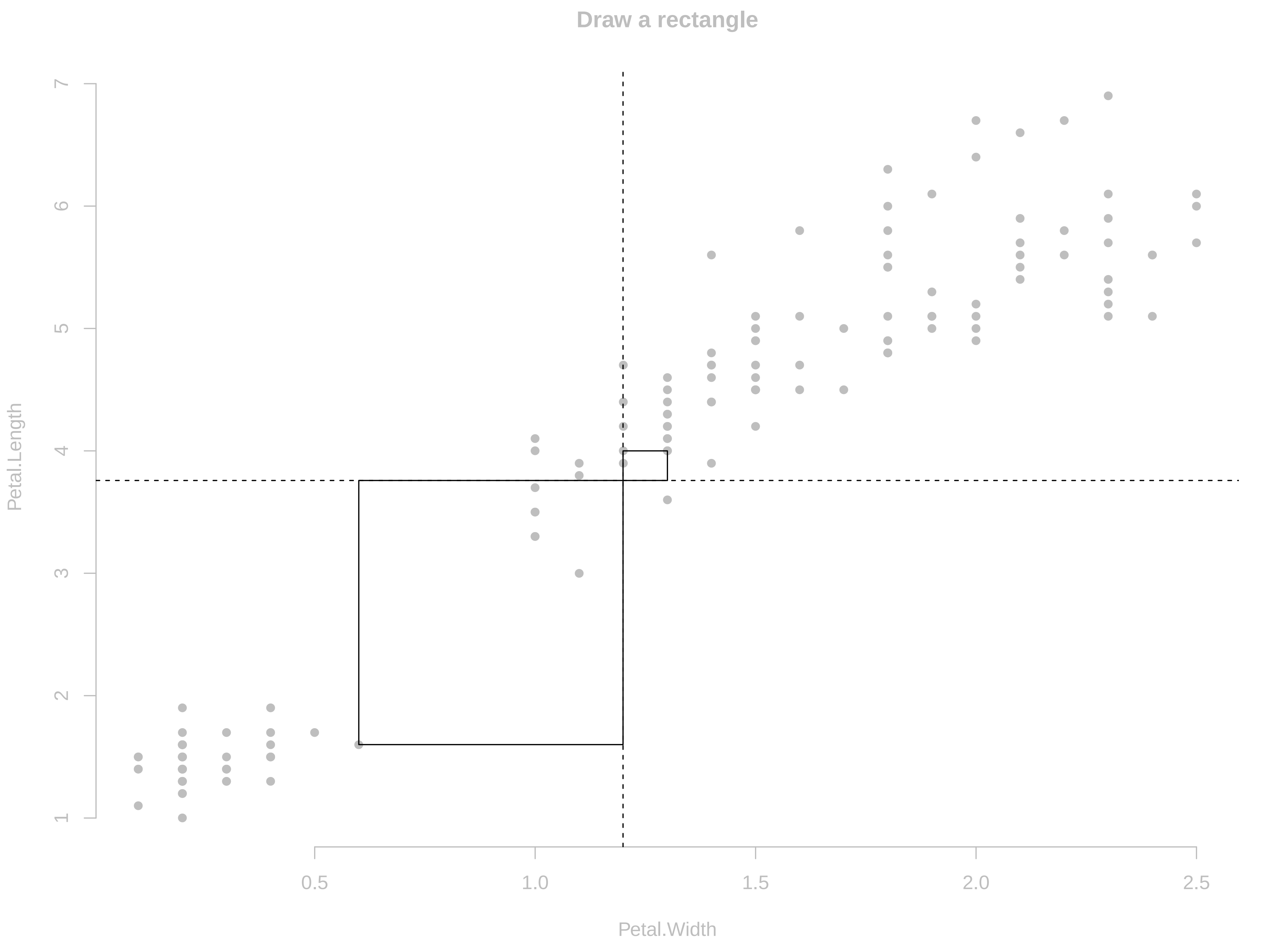
It should be low for these variables
Ozone
0
50
100
150
5
10
15
20
Wind


It should be near zero for these variables
rand$y
2
1
0
−1
−2
−3
−4
rand$x
−3
−2
−1
0
1
2

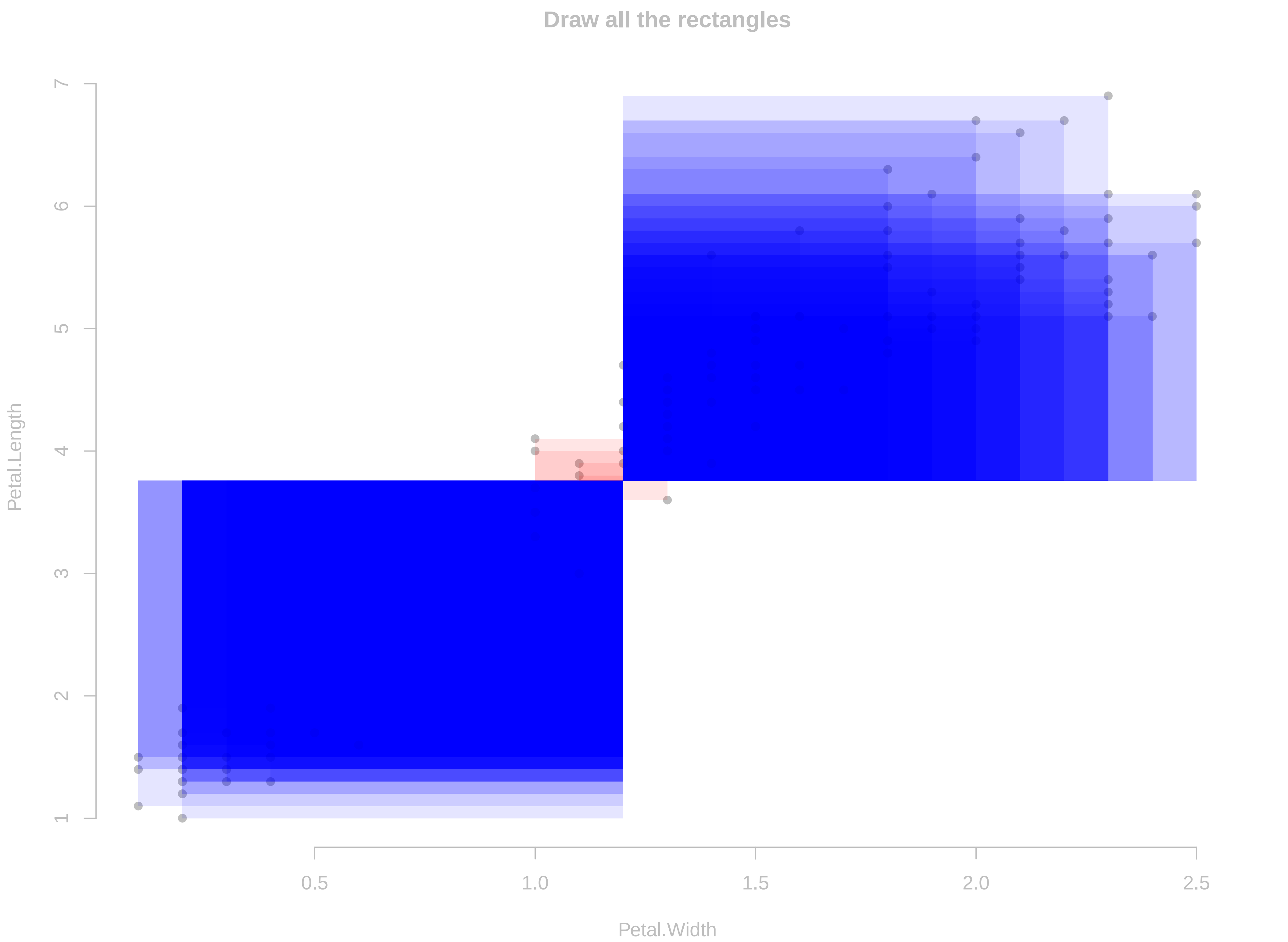
Covariance

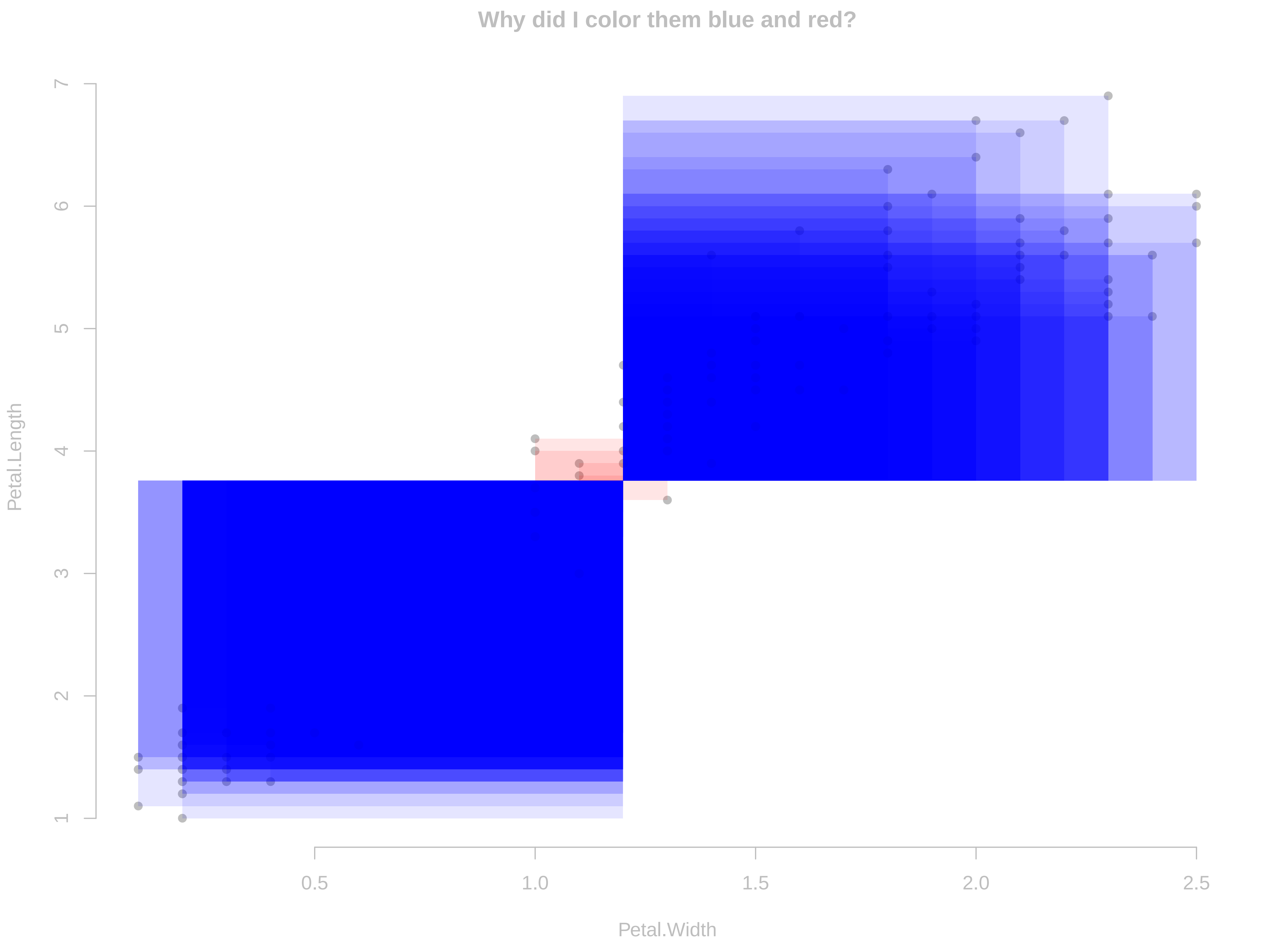

The iris variables
Petal.Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal.Width


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

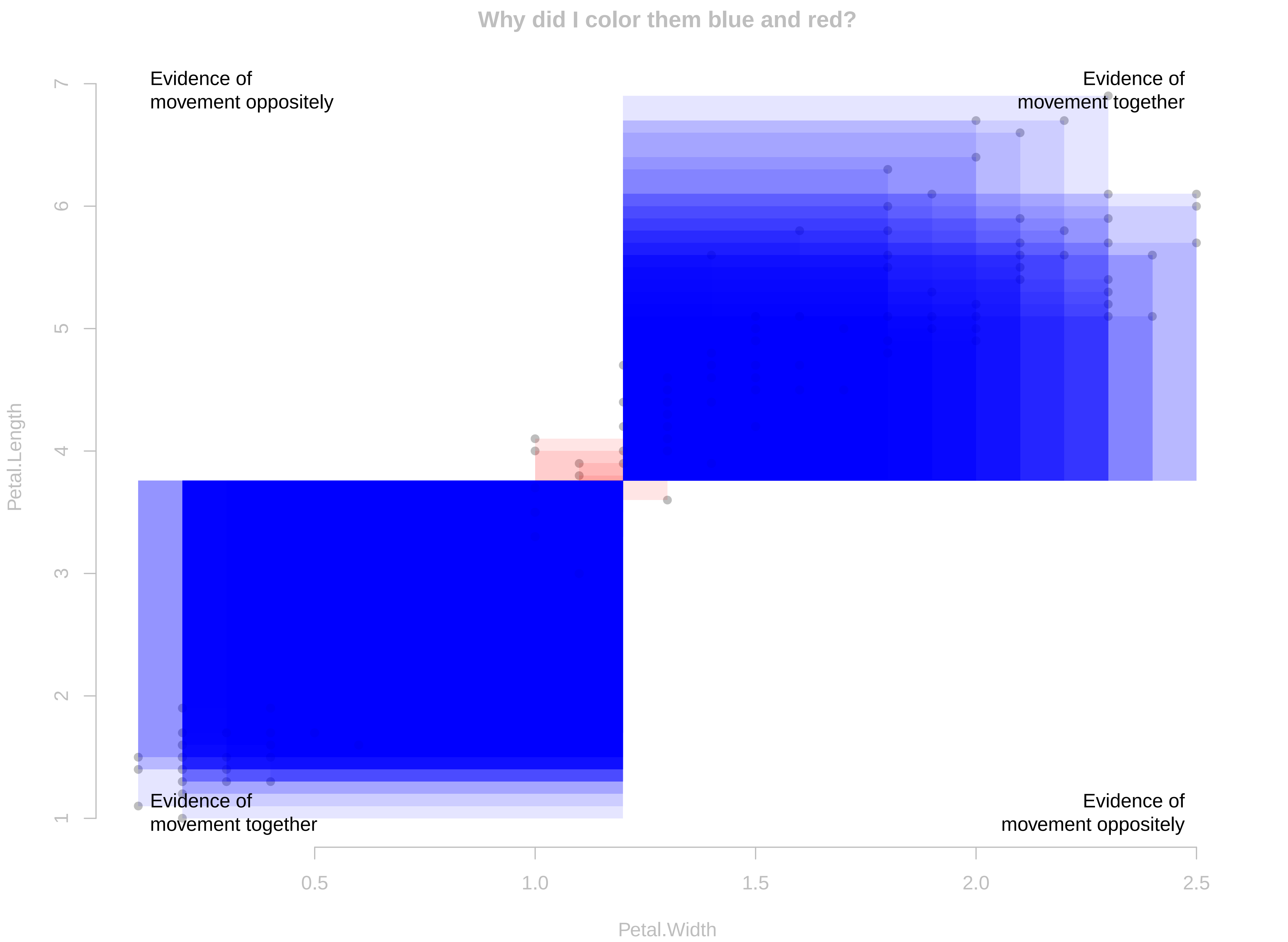


Draw a rectangle
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

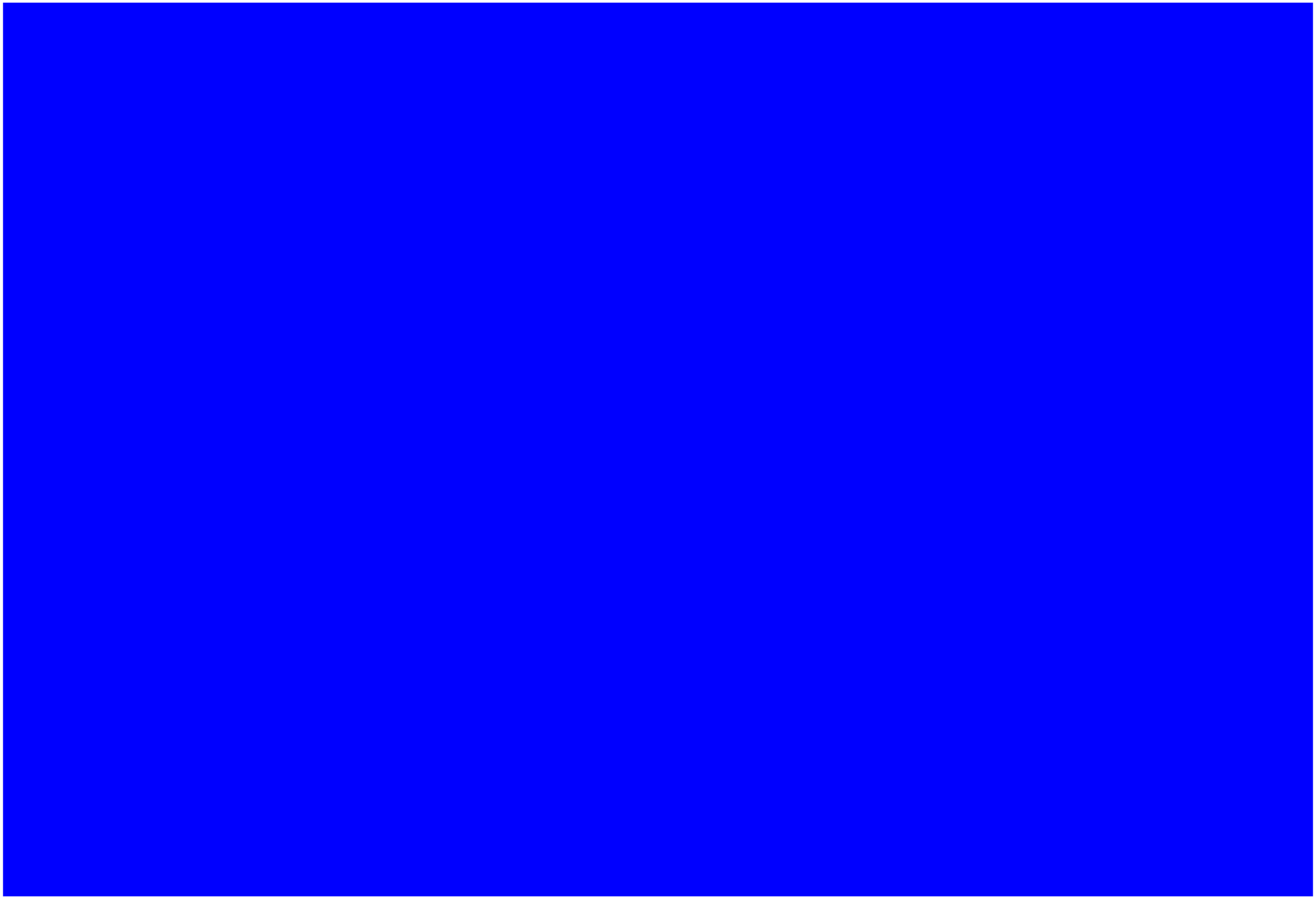

Draw all the rectangles
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5



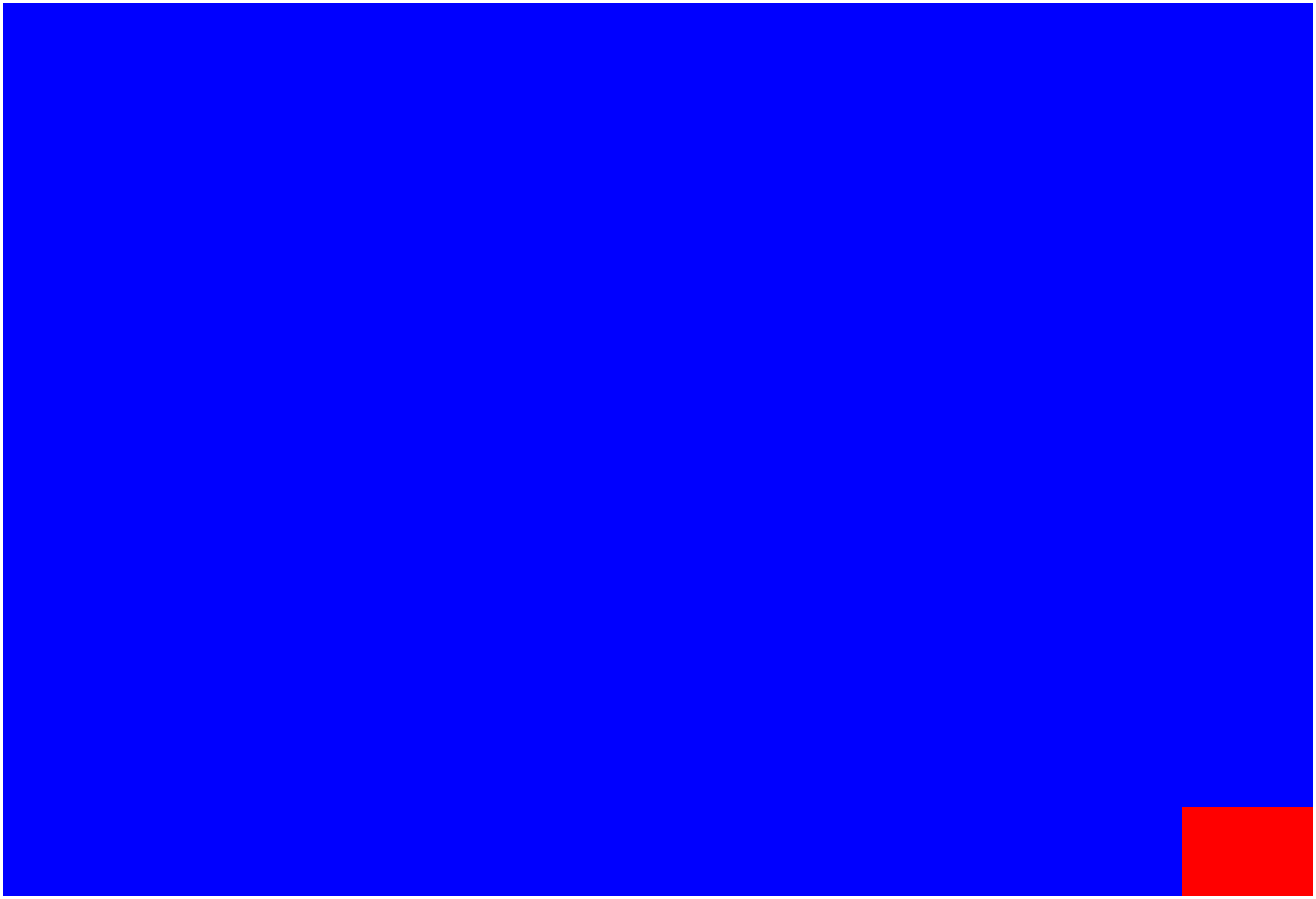
Why did I color them blue and red?
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


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
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

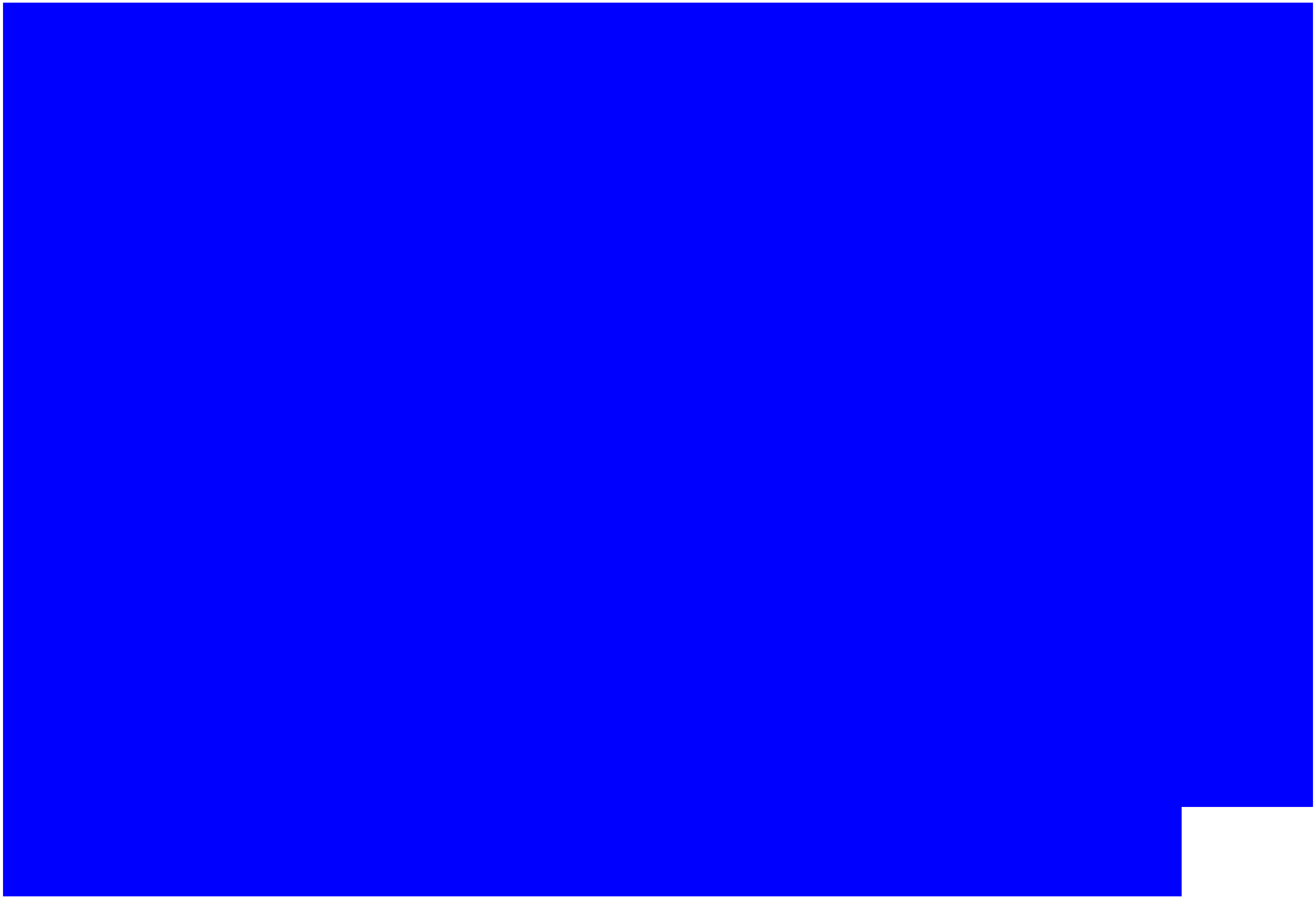
Add the blues together. (This is at a different scale.)



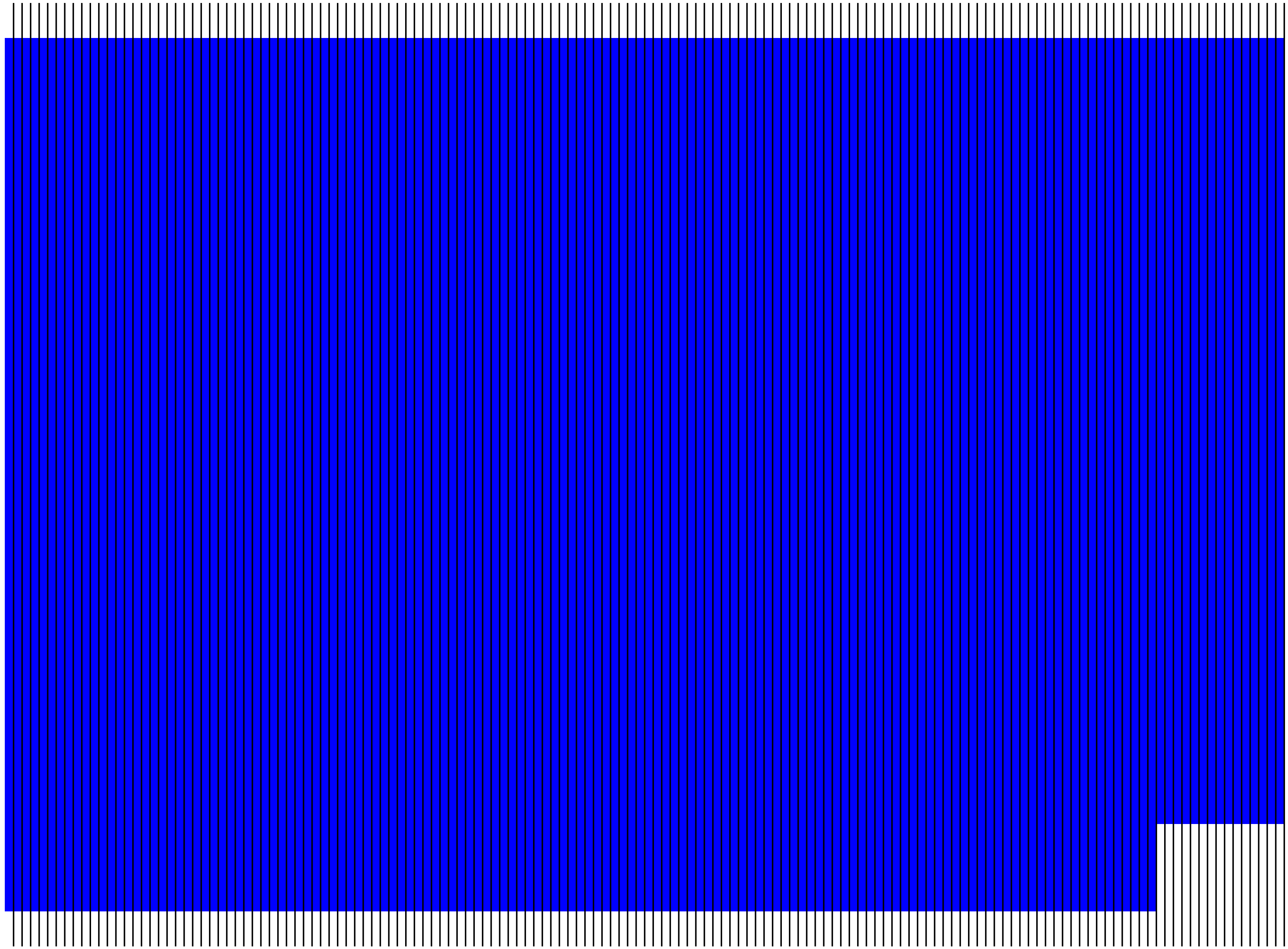
Add the reds together.



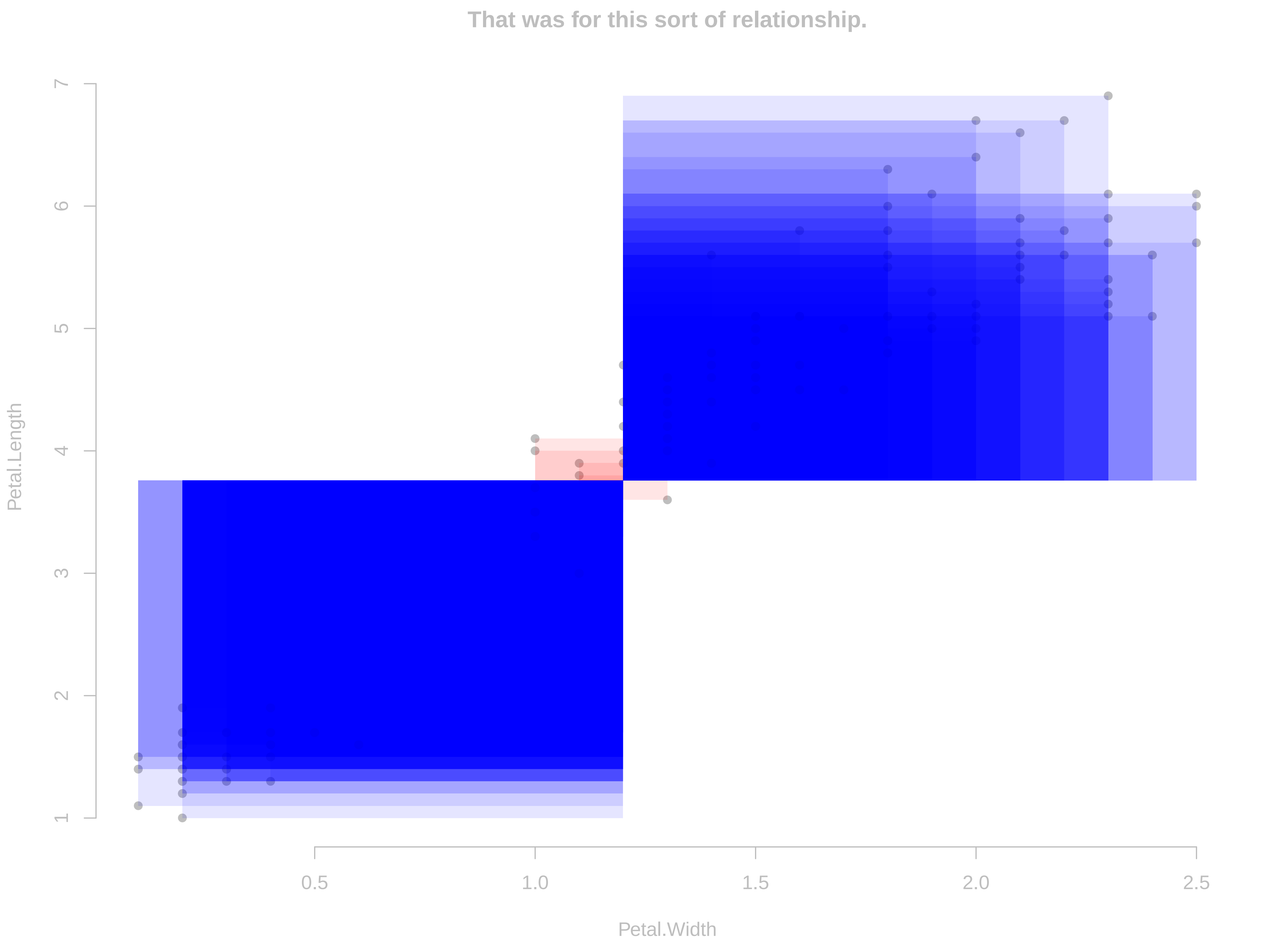
Subtract the reds.

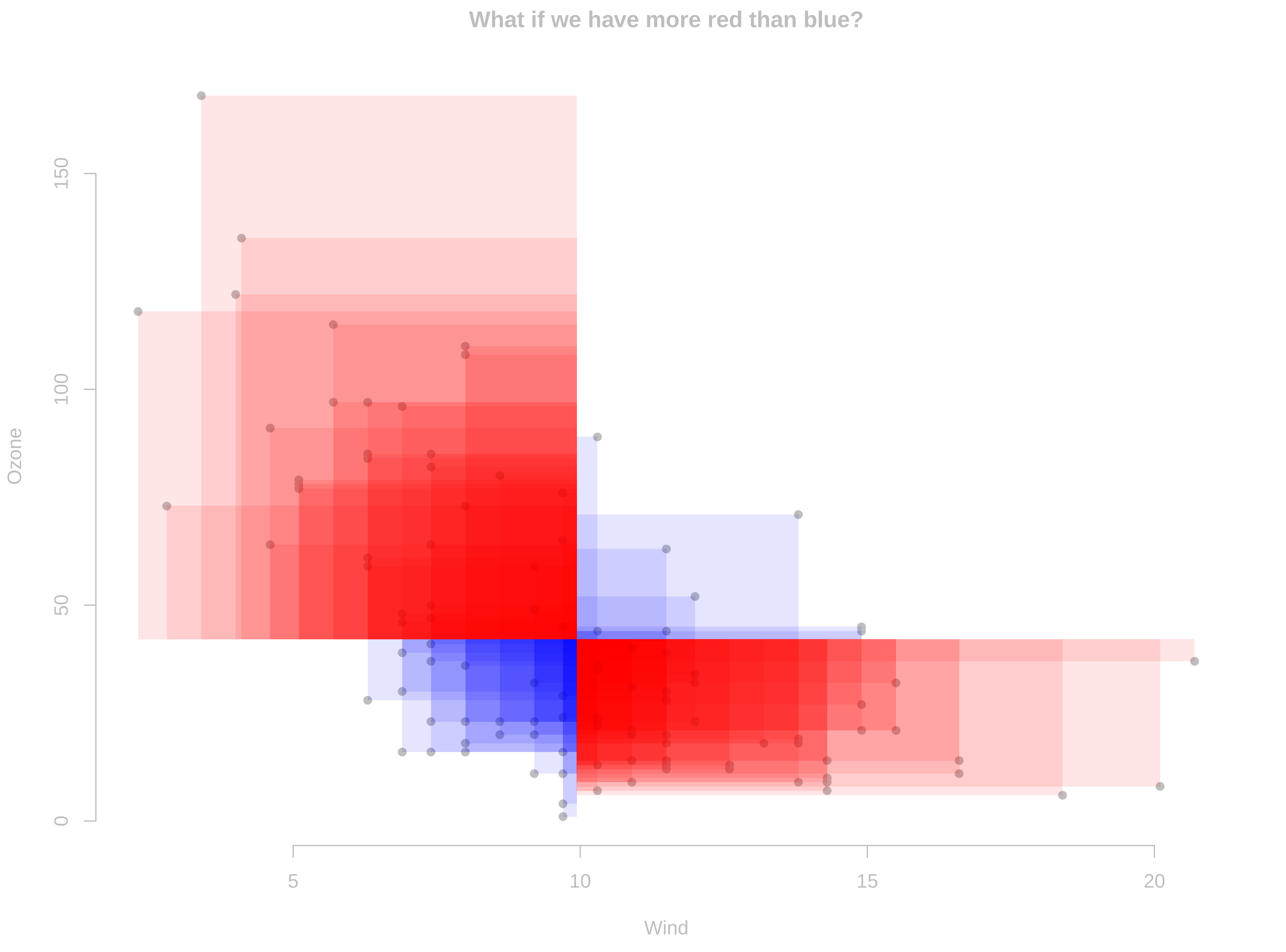


Divide into as many equal pieces as we have irises (n).

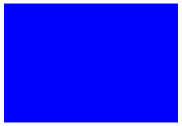


This blue sliver is the covariance.

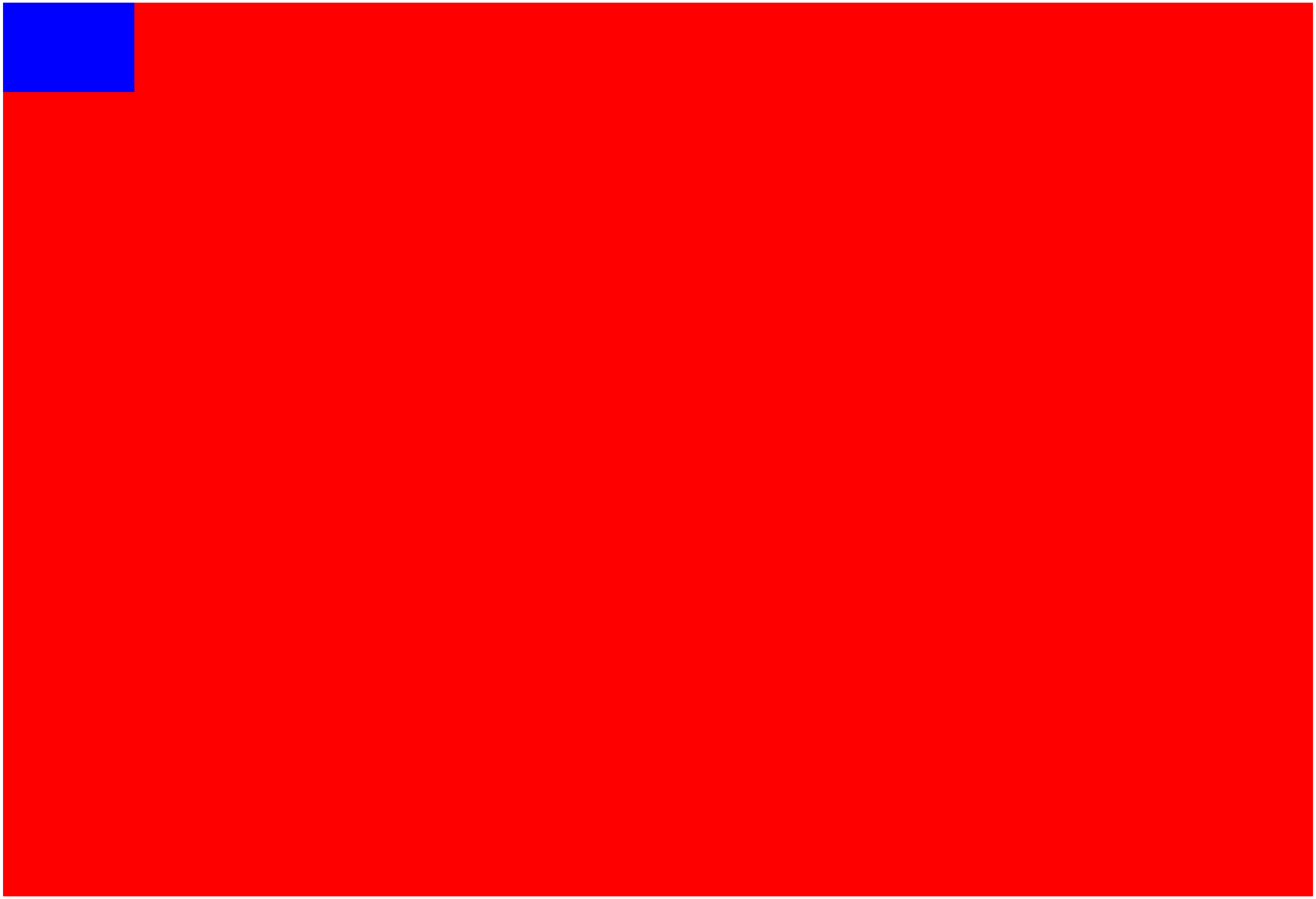

That was for this sort of relationship.
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


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

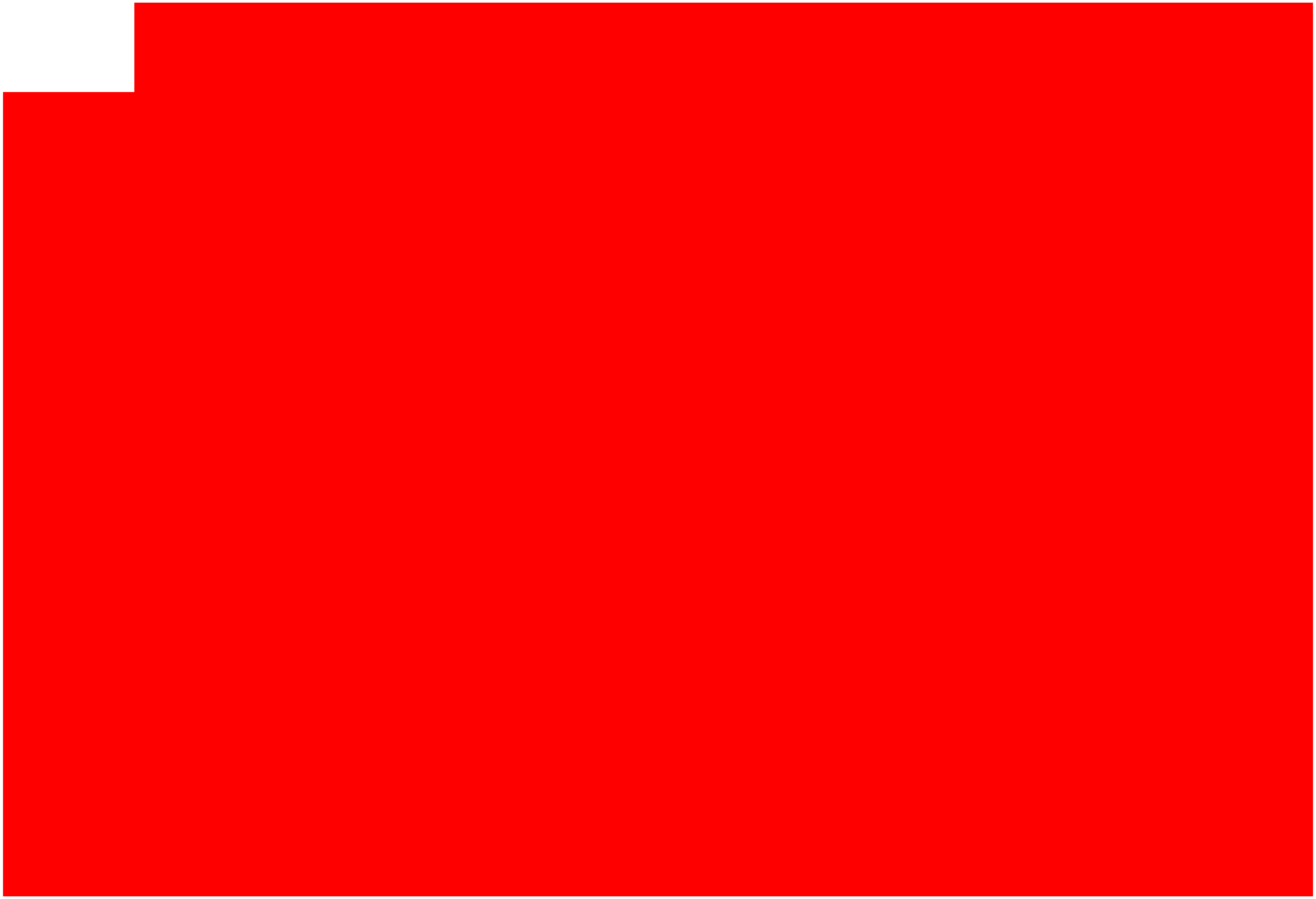
Add the blues together. (This is at a different scale.)



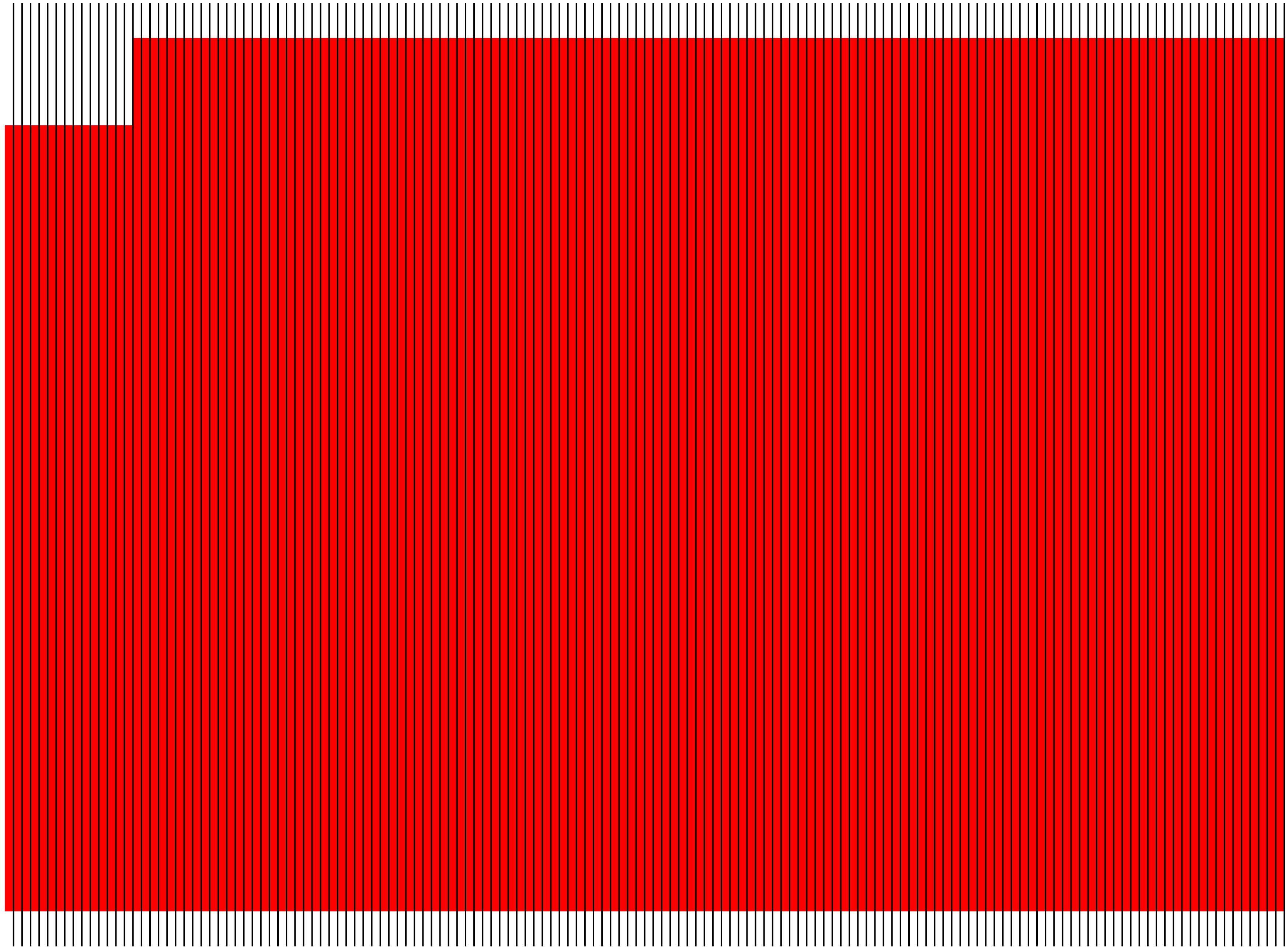
Add the reds together.



Subtract the reds.



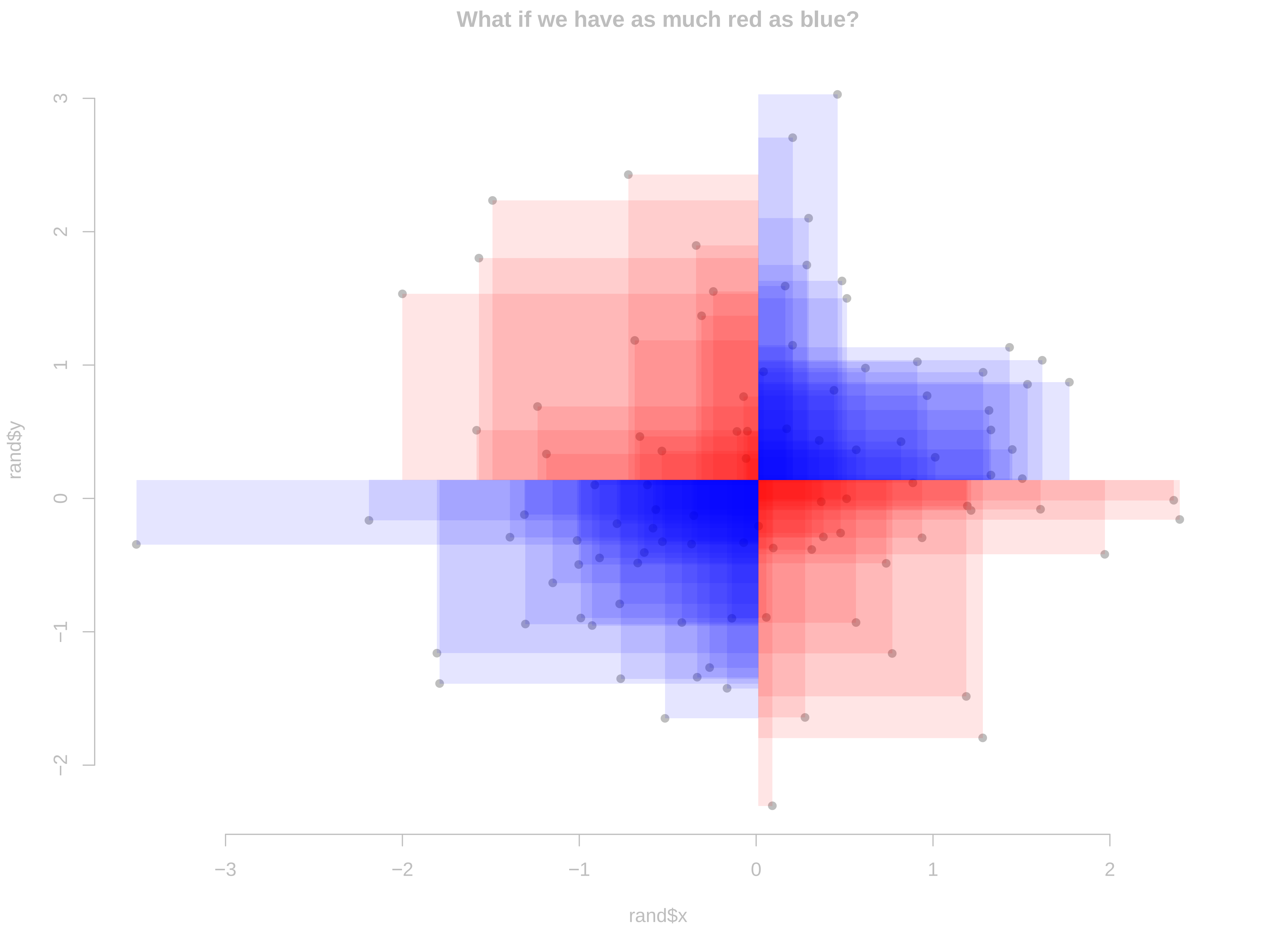
Divide into as many equal pieces as we have irises (n).



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?
3
2
1
0
−1
−2
rand$y
−3
−2
−1
0
1
2
rand$x

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

Add the reds together.

Subtract the reds.

(Covariance is zero.)

Variance

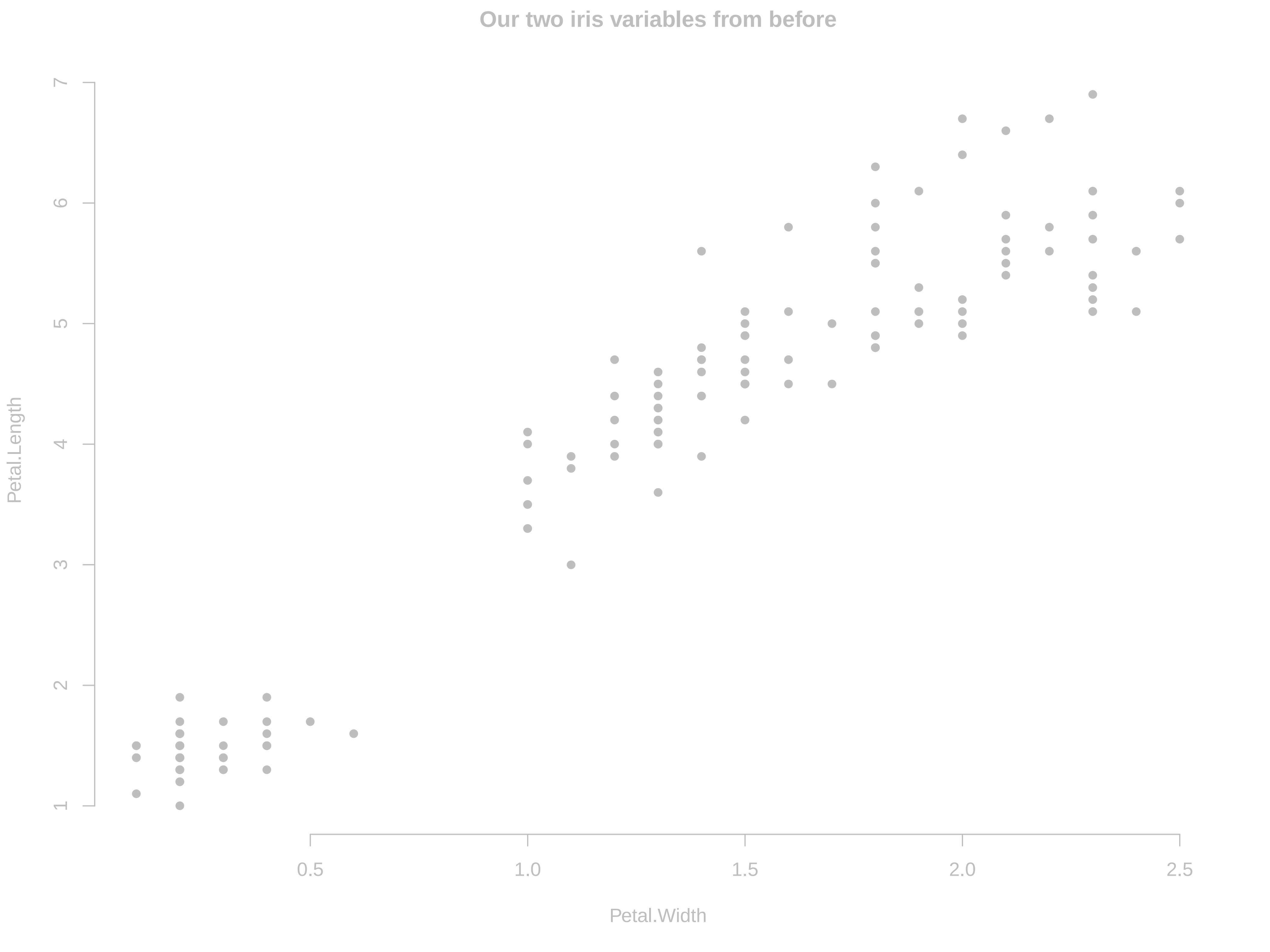
Variance tells us how spread out some numbers are.

1, 4, 8, 10

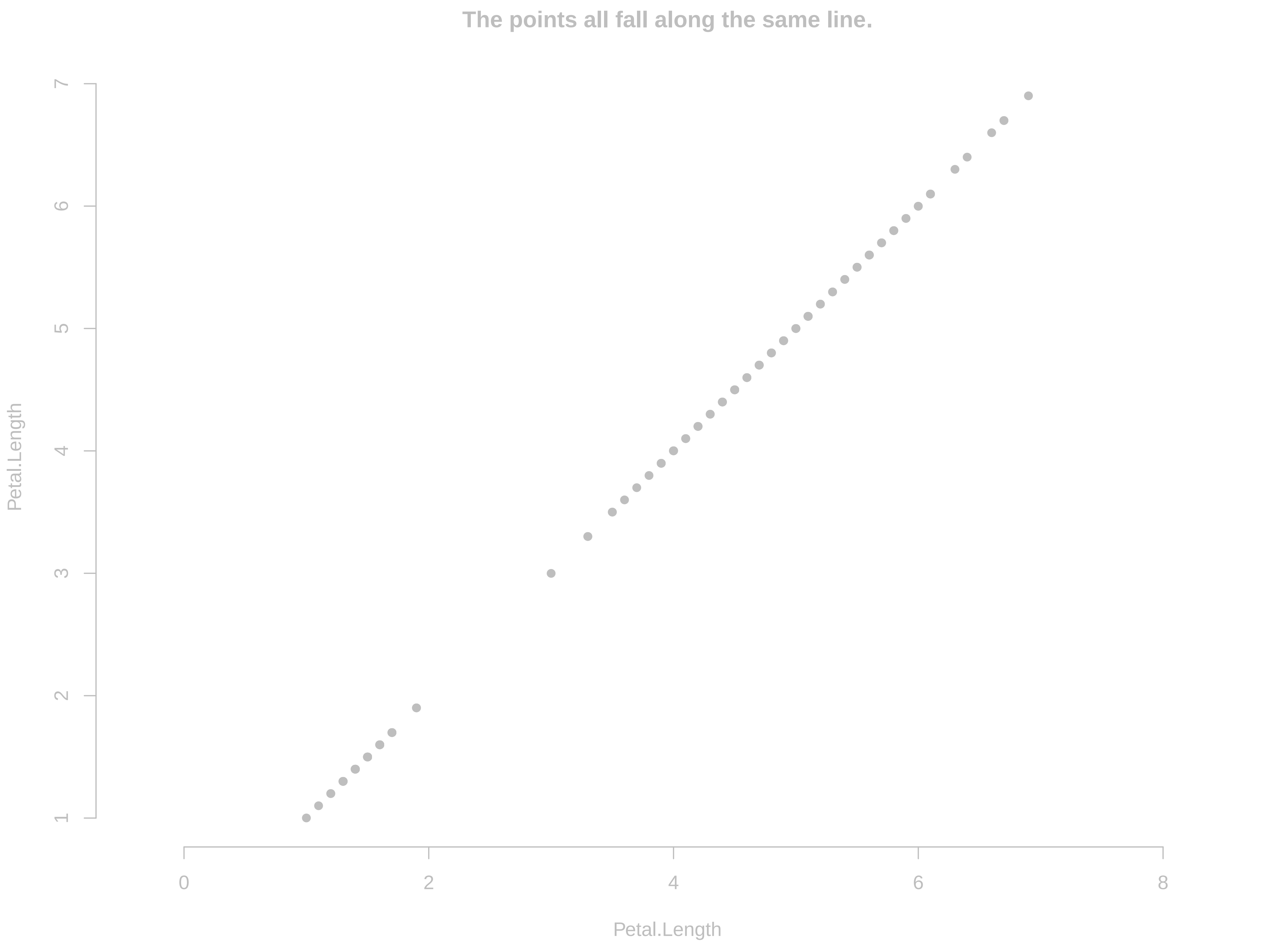
vs

4, 4, 5, 6

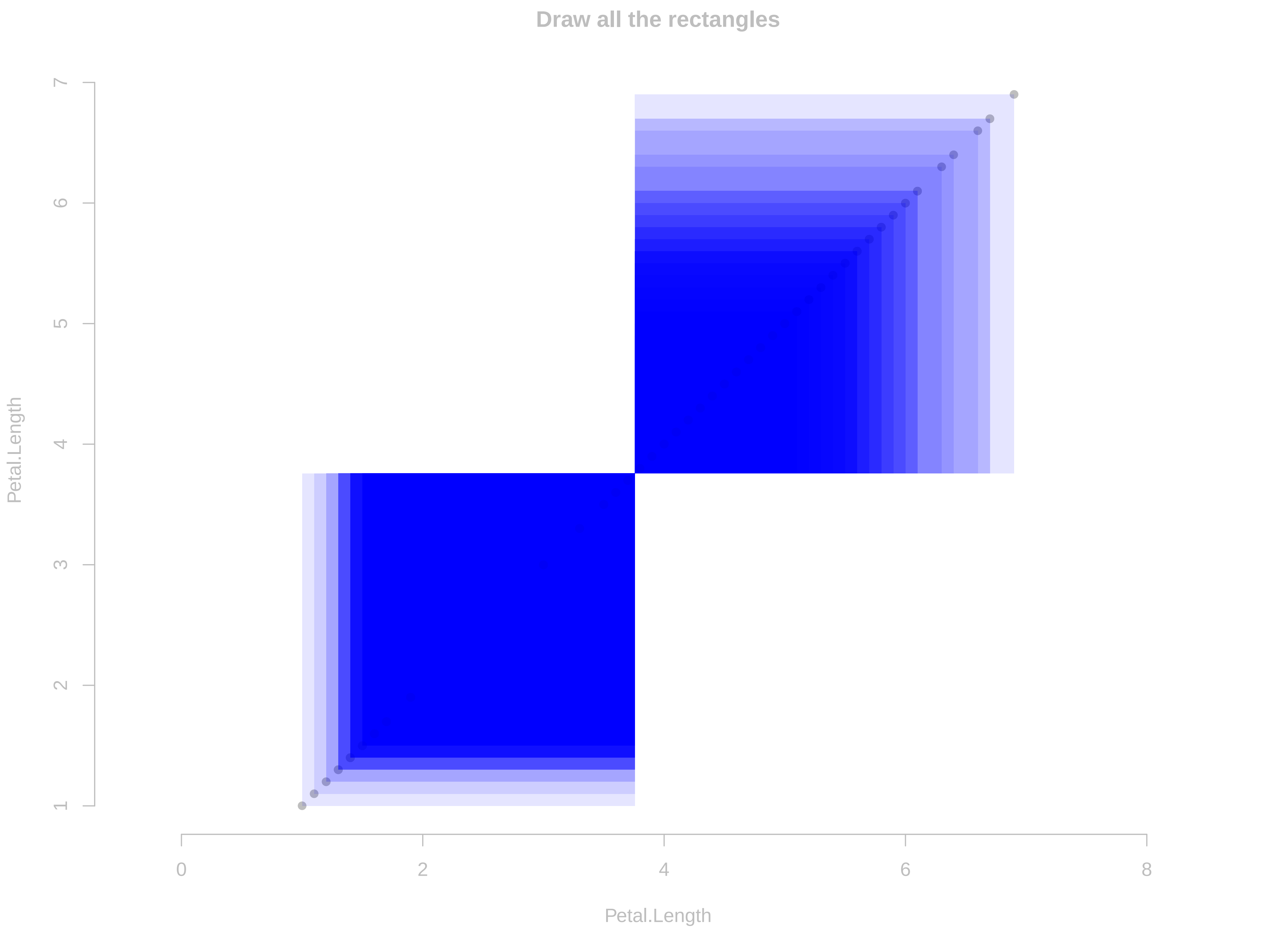
The variance of a variable is the covariance of the variable with itself.


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

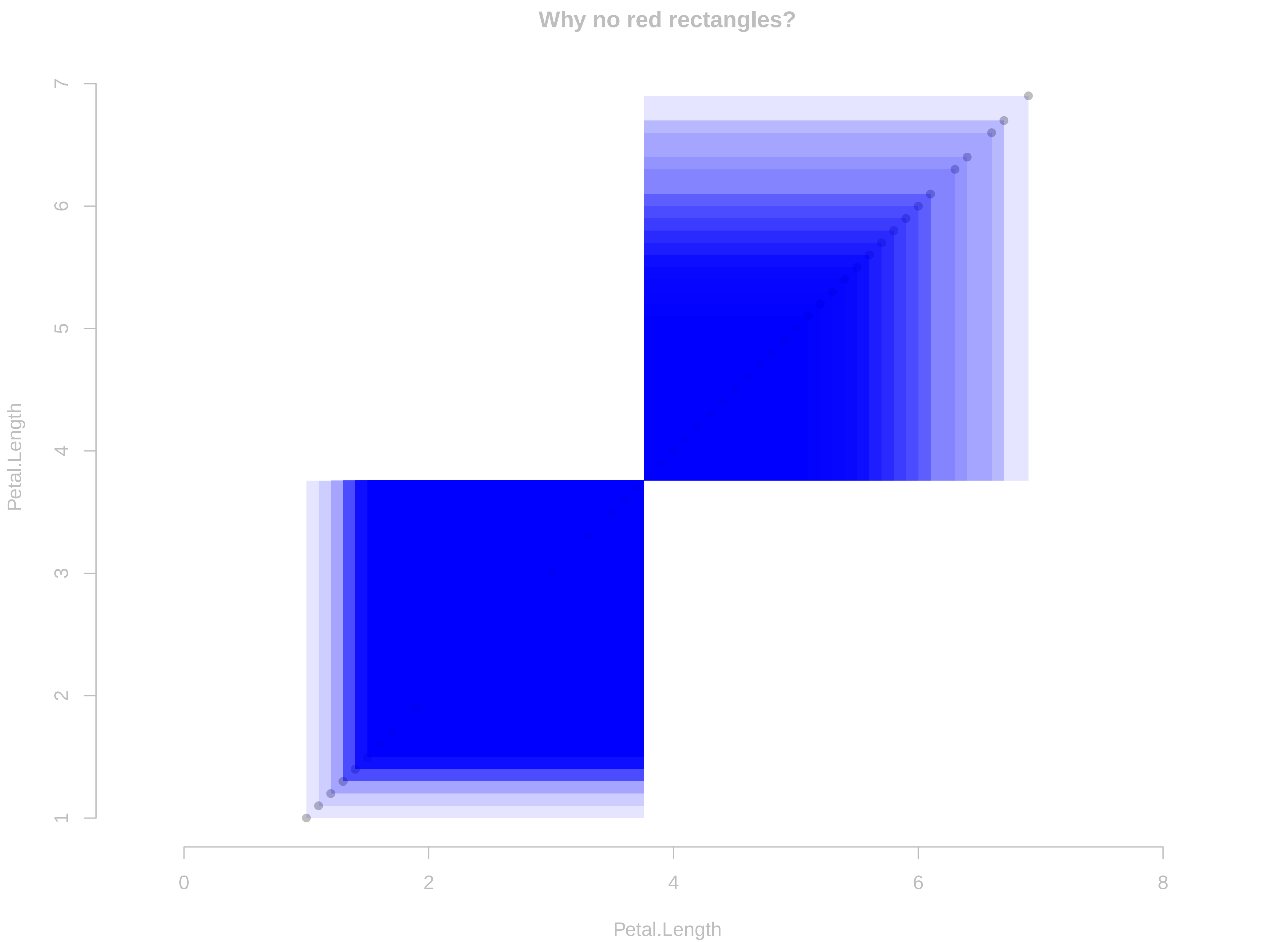
Let's look at just one of them.


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

Let's find the variance of Petal.Length

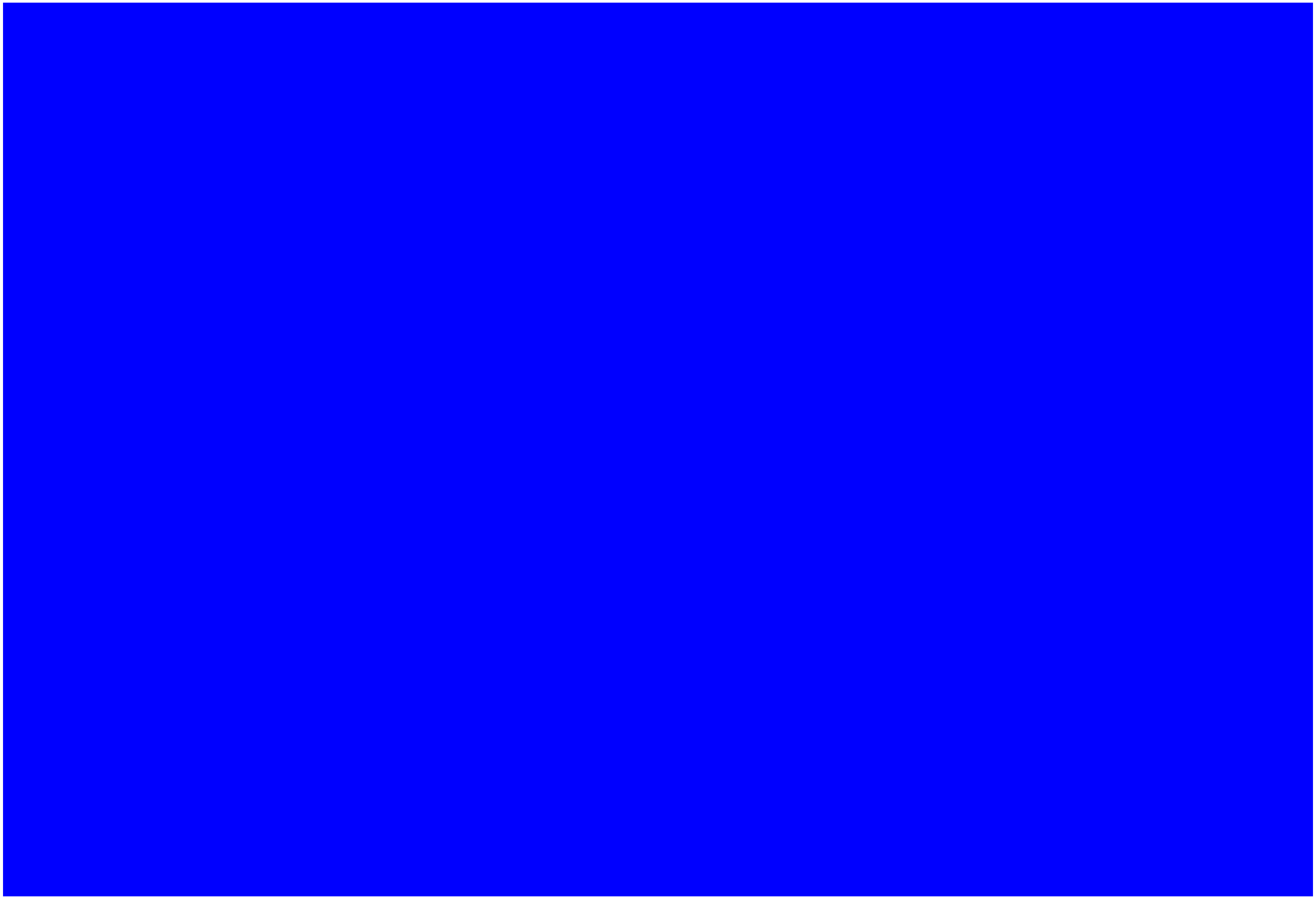


Draw all the rectangles
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

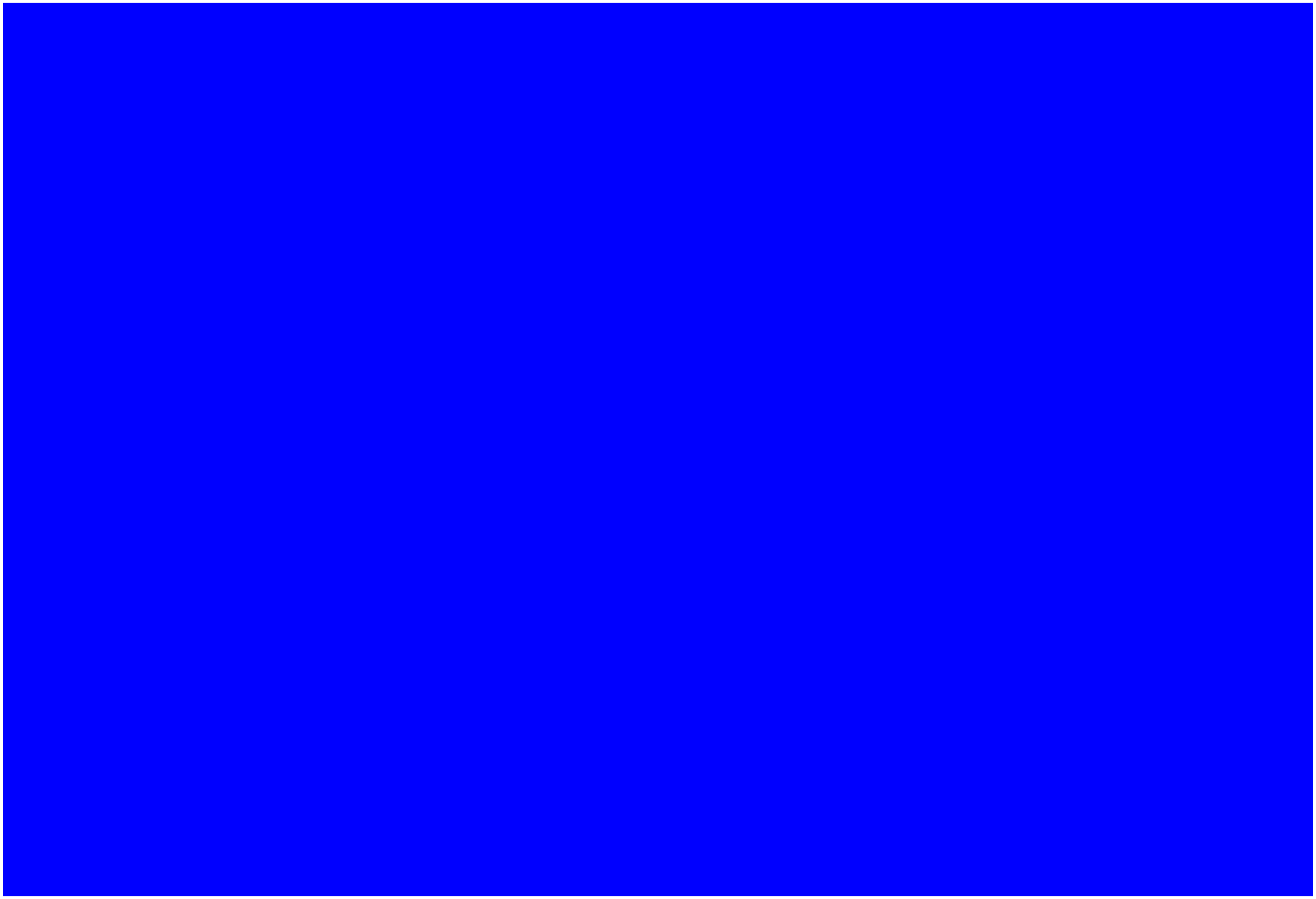


Why no red rectangles?
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length

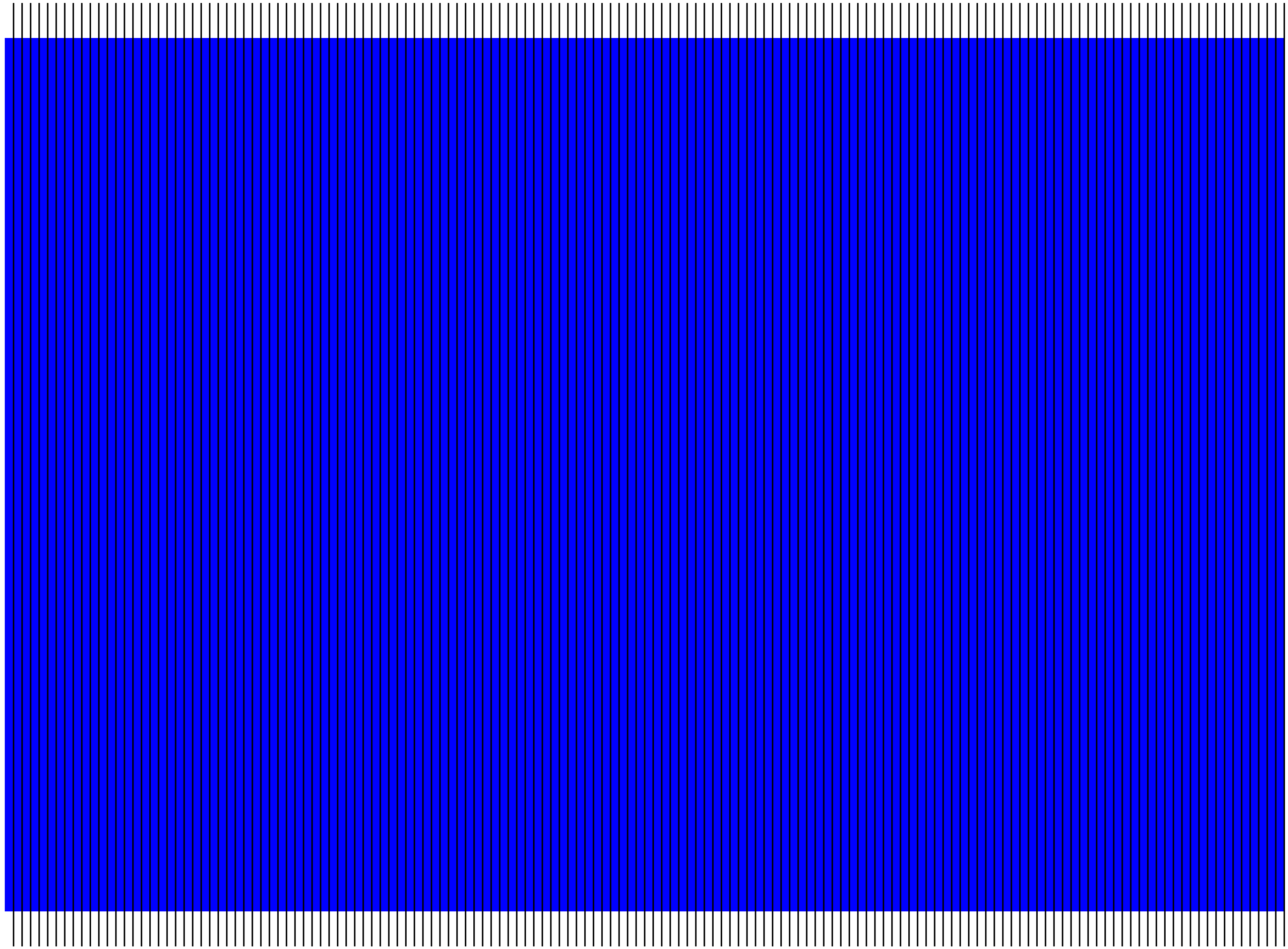
Add the blues together. (This is at a different scale.)



We have no reds to subtract.



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



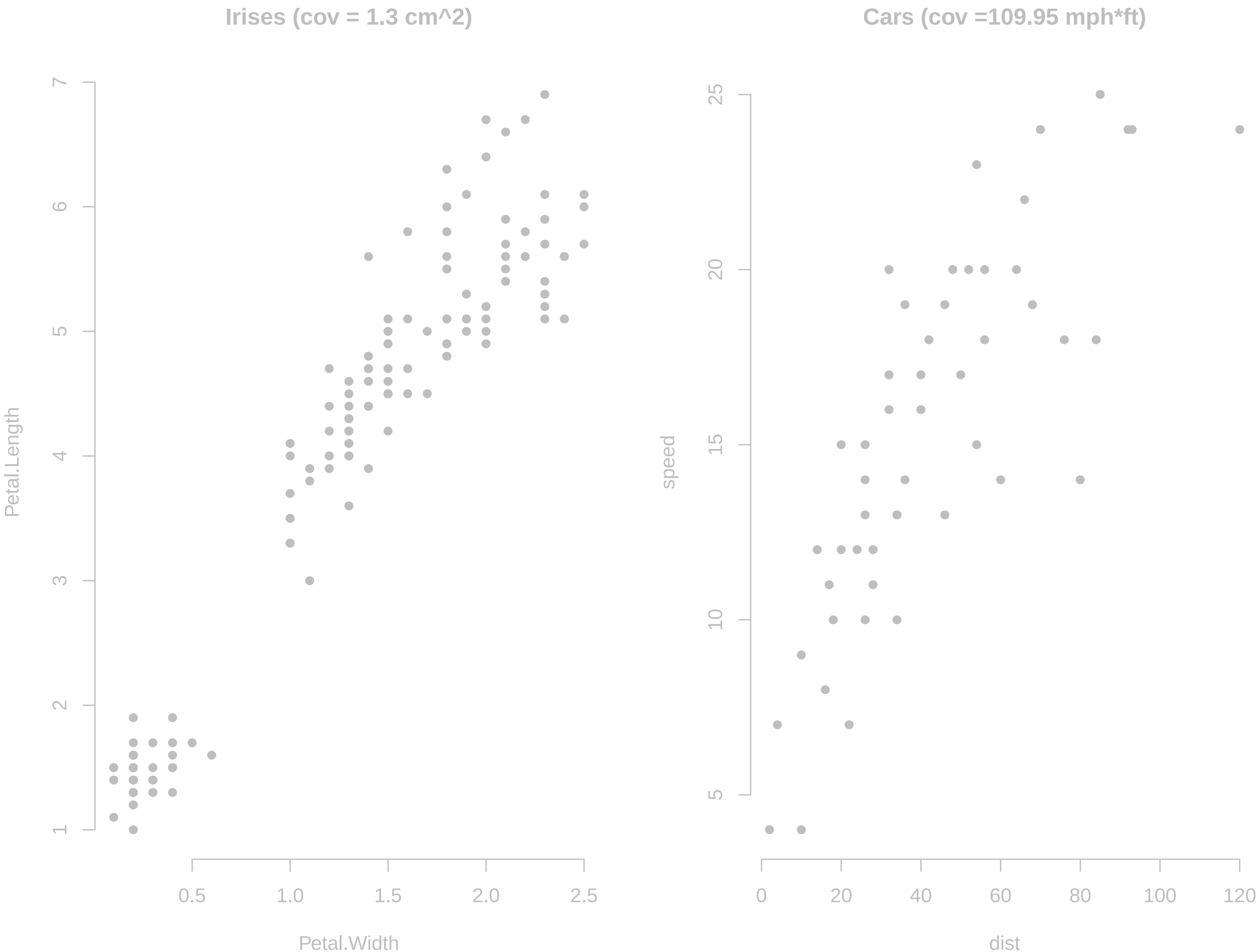
This blue sliver is the variance.

A problem with covariance

Covariance has units!

(x–unit times y–unit)

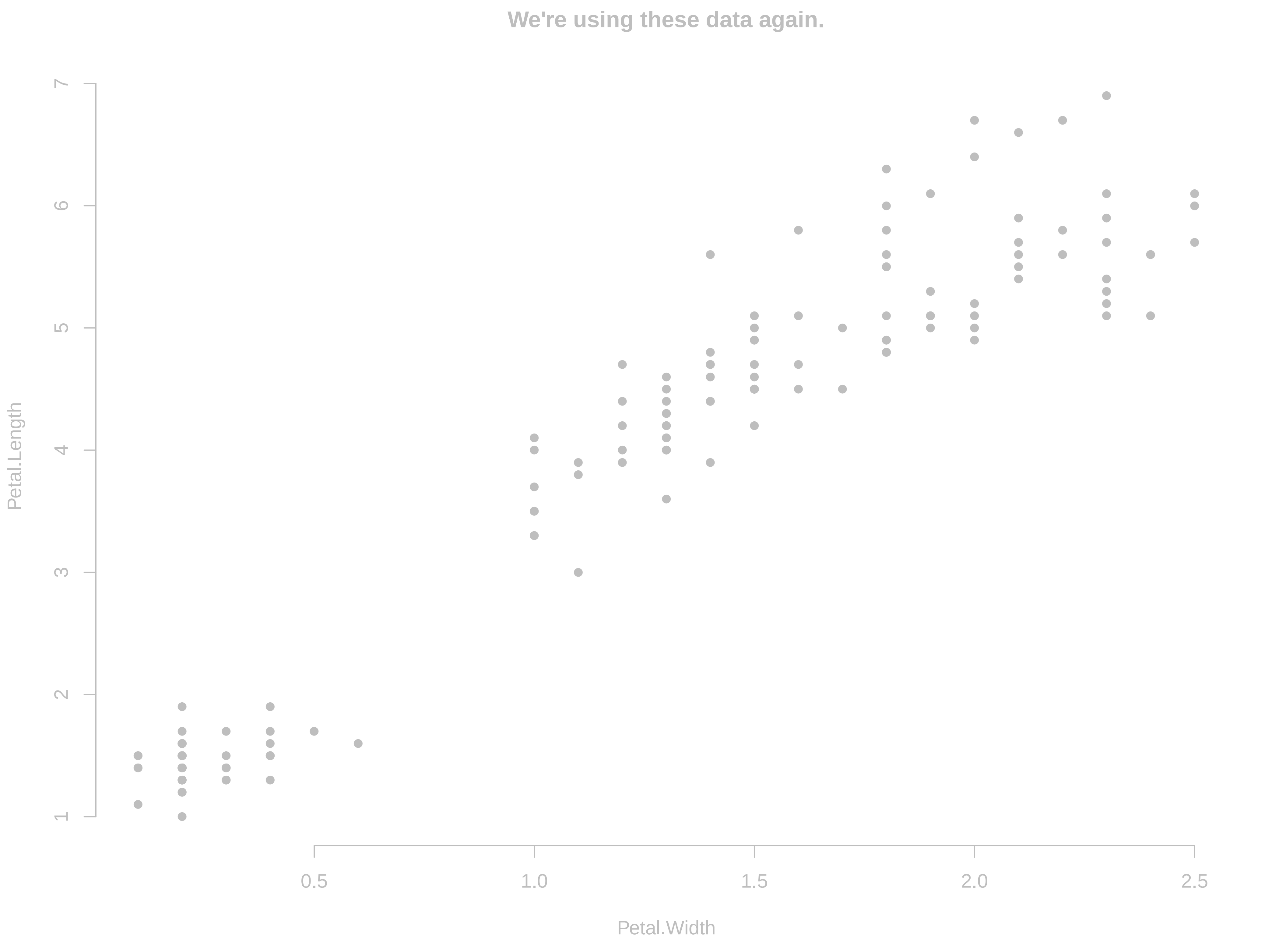
Which relationship is stronger (more linear)?

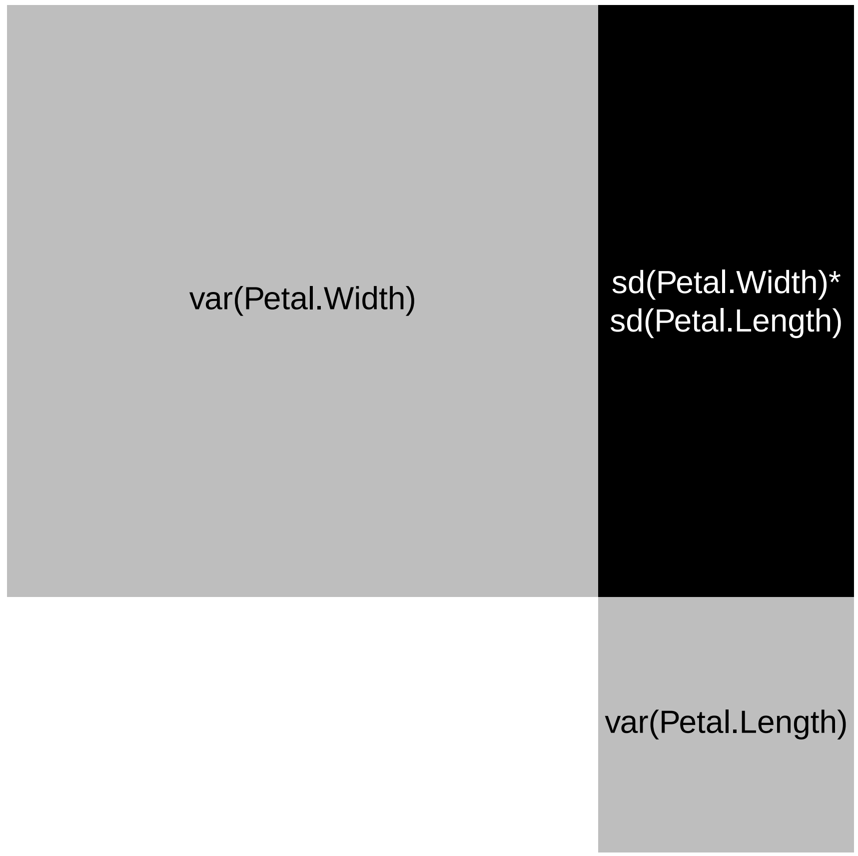


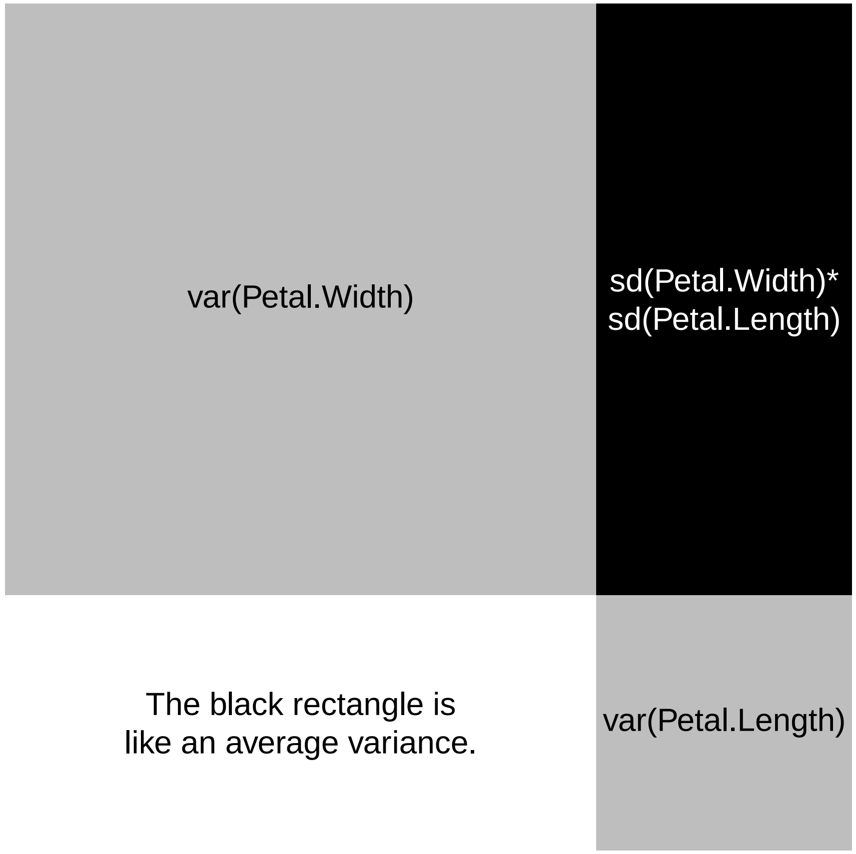
Irises (cov = 1.3 cm^2)
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Cars (cov =109.95 mph*ft)
speed
dist
5
10
15
20
25
0
20
40
60
80
100
120

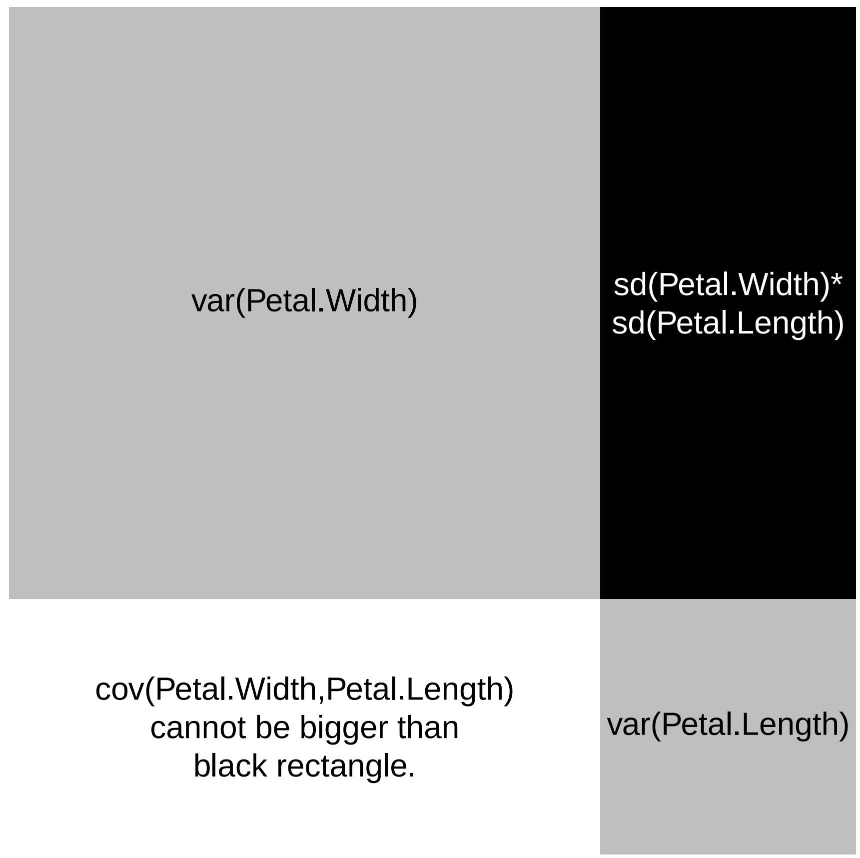
Oh noes!

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

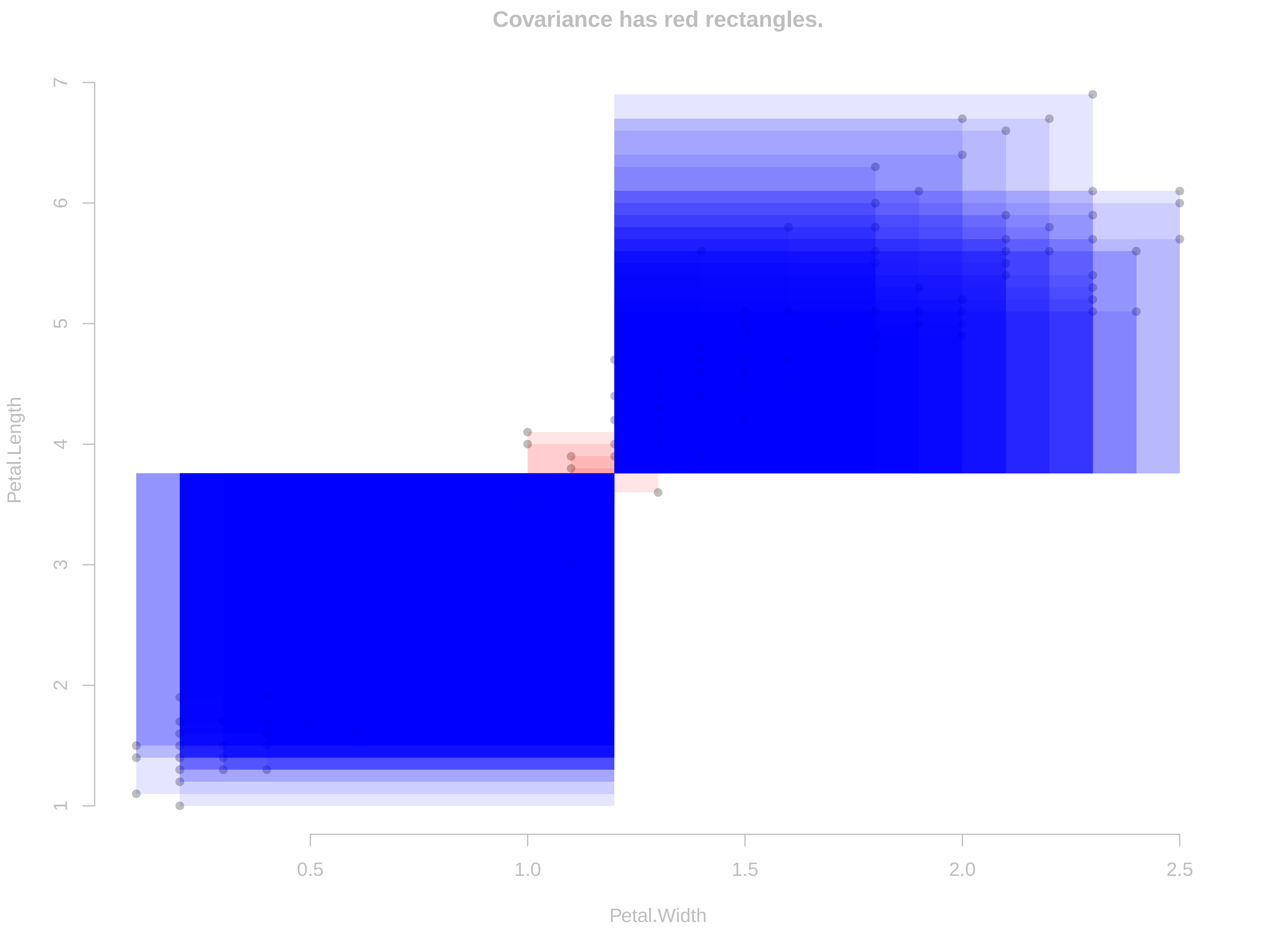

We're using these data again.
Petal.Length
1
2
3
4
5
6
7
Petal.Width
0.5
1.0
1.5
2.0
2.5


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

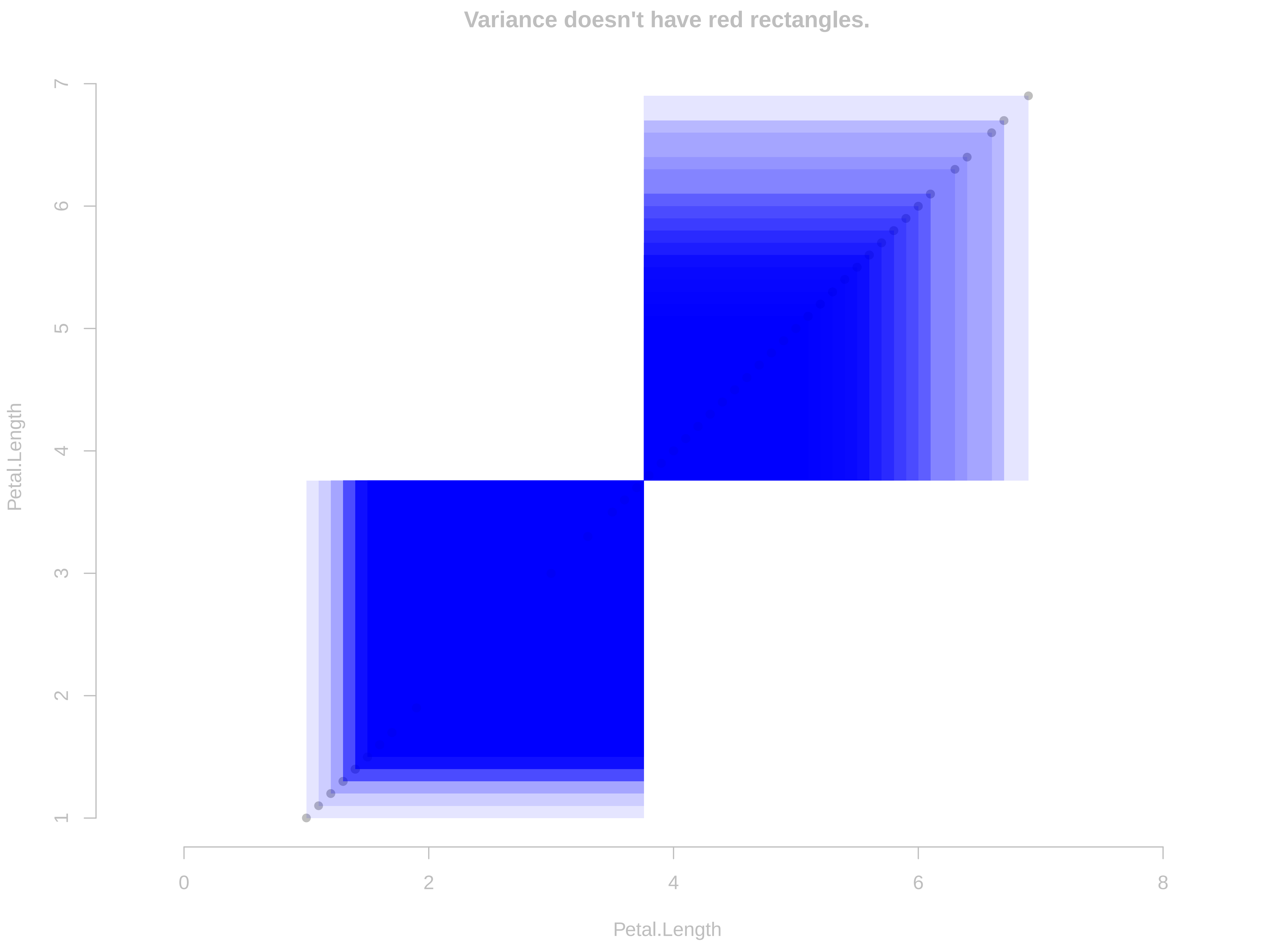

var(Petal.Width)
sd(Petal.Width)*
sd(Petal.Length)
The black rectangle is
like an average variance.
var(Petal.Length)


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

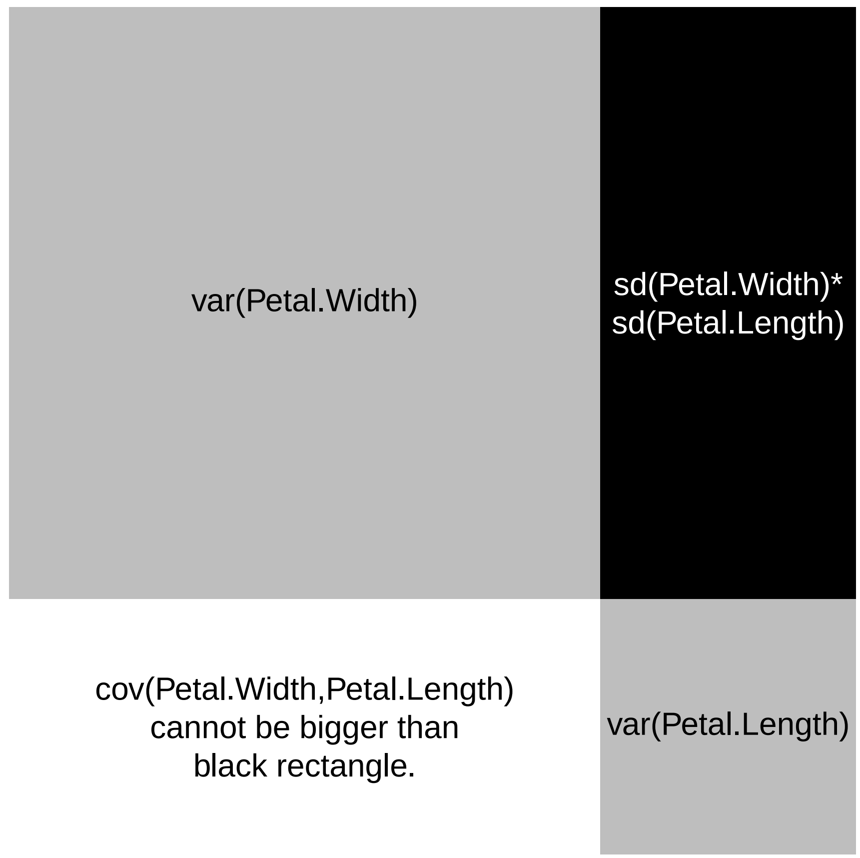
Why?



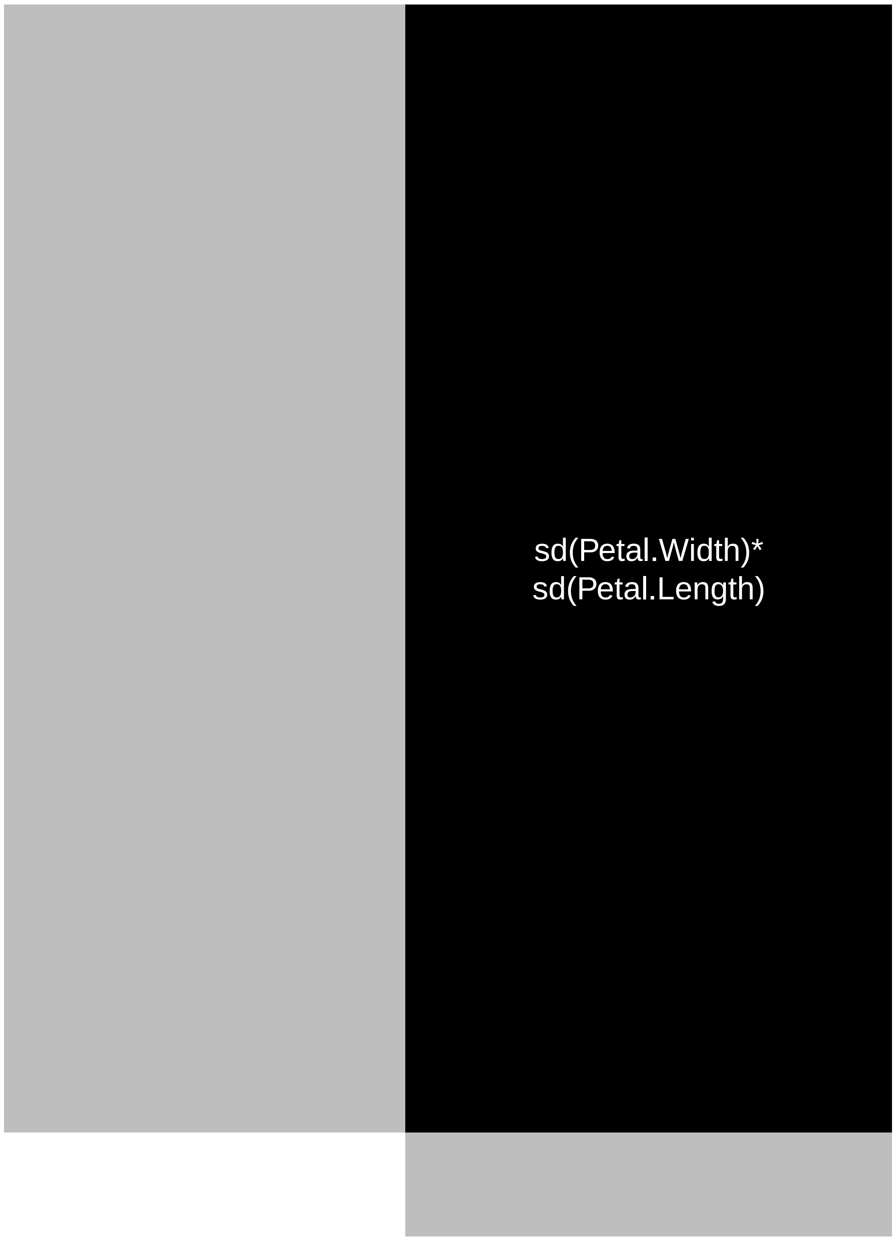
Covariance has red rectangles.
Petal.Length
1
2
3
4
5
6
7
Petal.Width
0.5
1.0
1.5
2.0
2.5



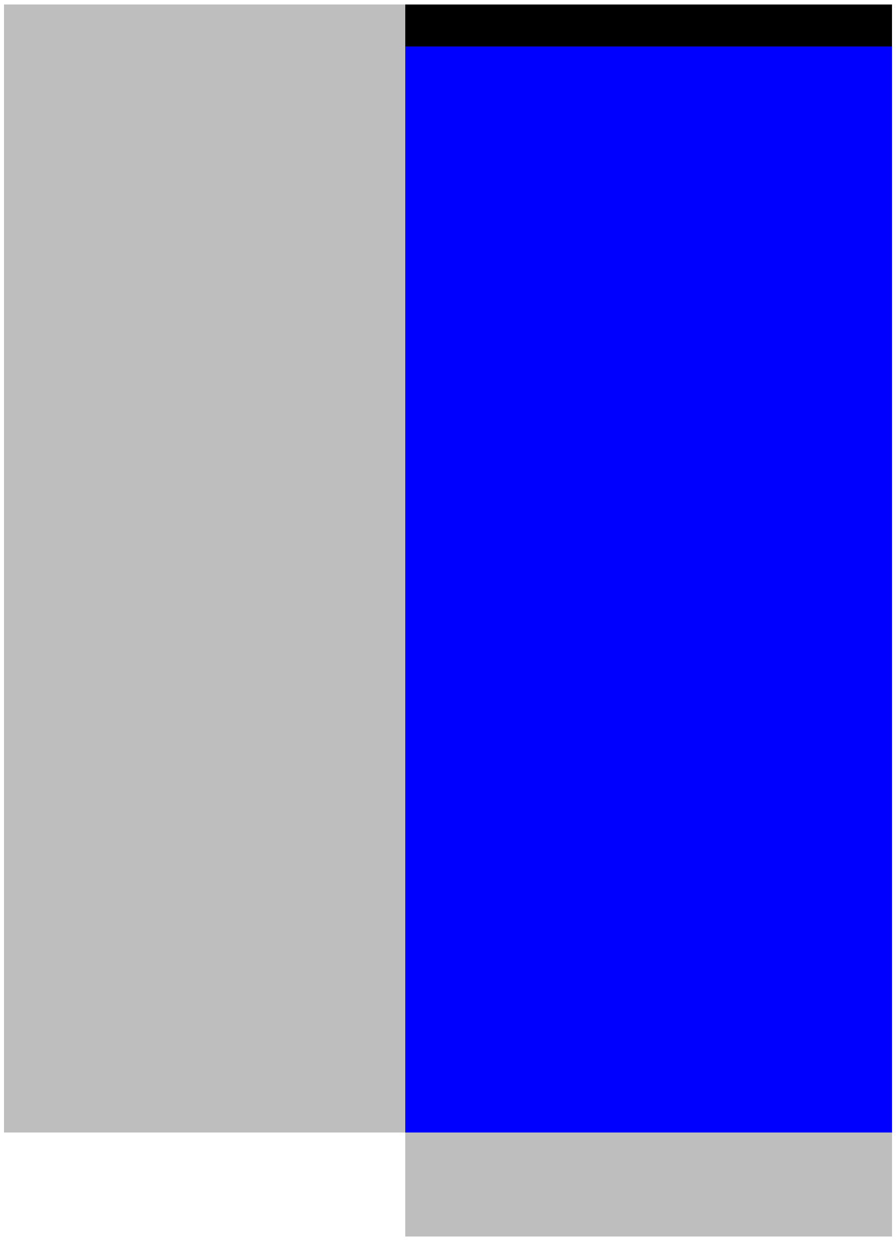
Variance doesn't have red rectangles.
Petal.Length
1
2
3
4
5
6
7
0
2
4
6
8
Petal.Length


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.

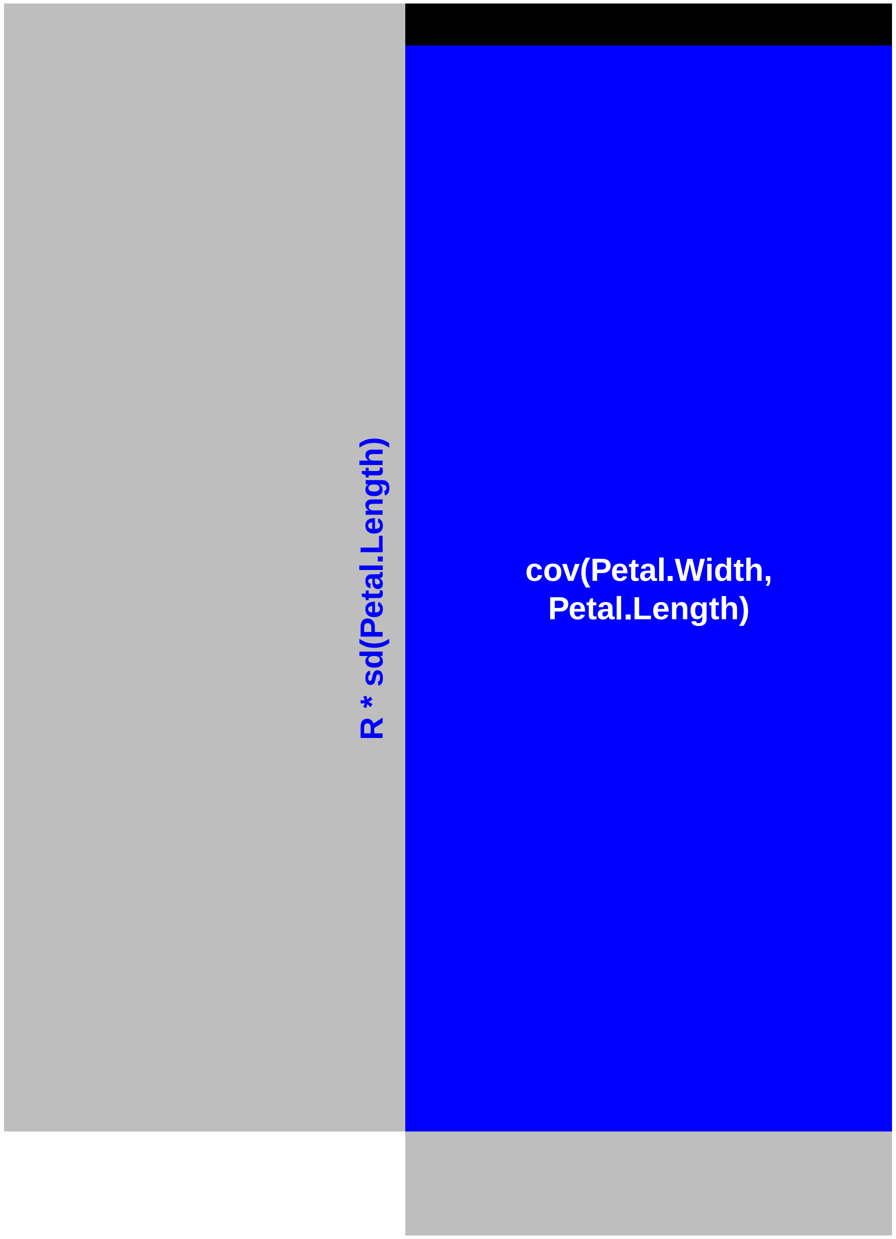


Squish covariance vertically into the rectangle.

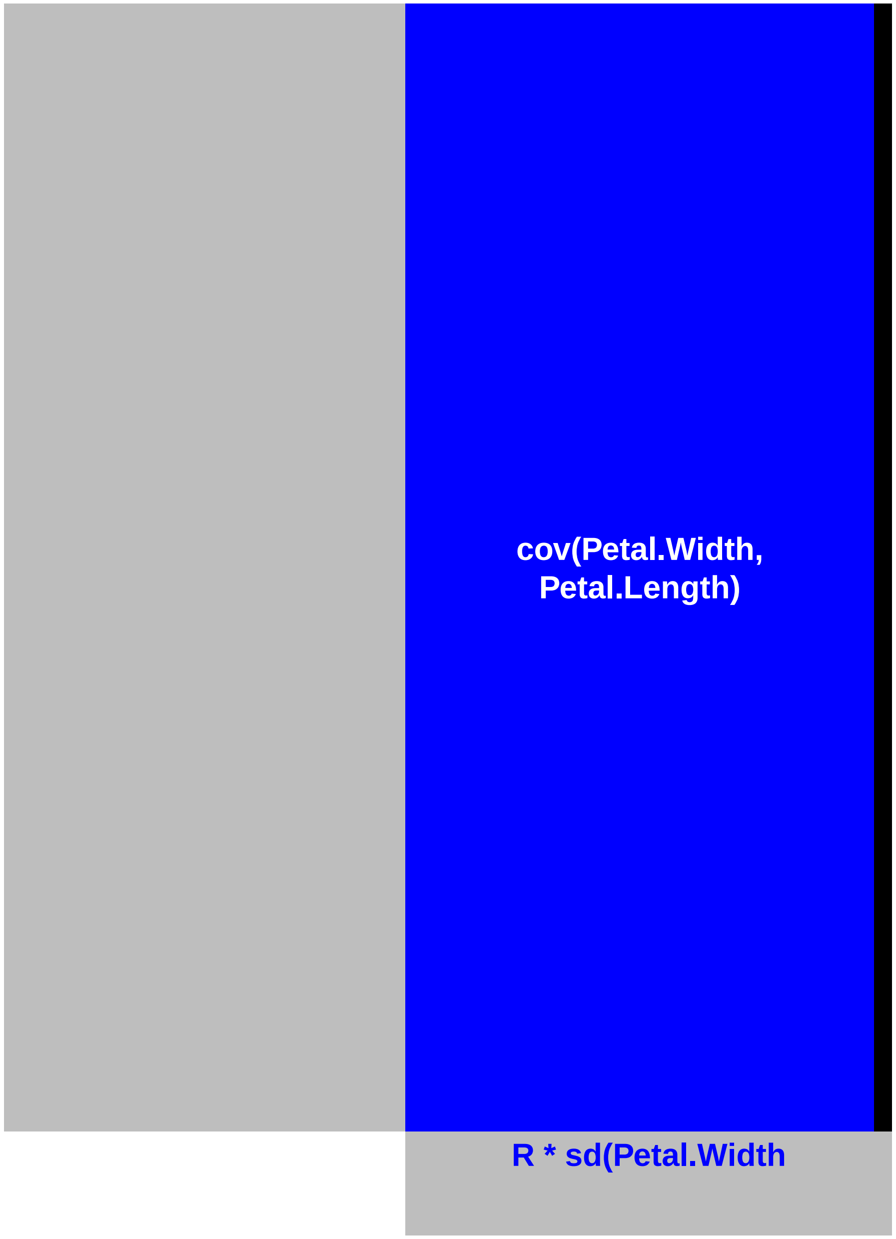


Correlation (R)
is the ratio of
the small rectangle
to the big rectangle.

Squish covariance vertically into the rectangle.

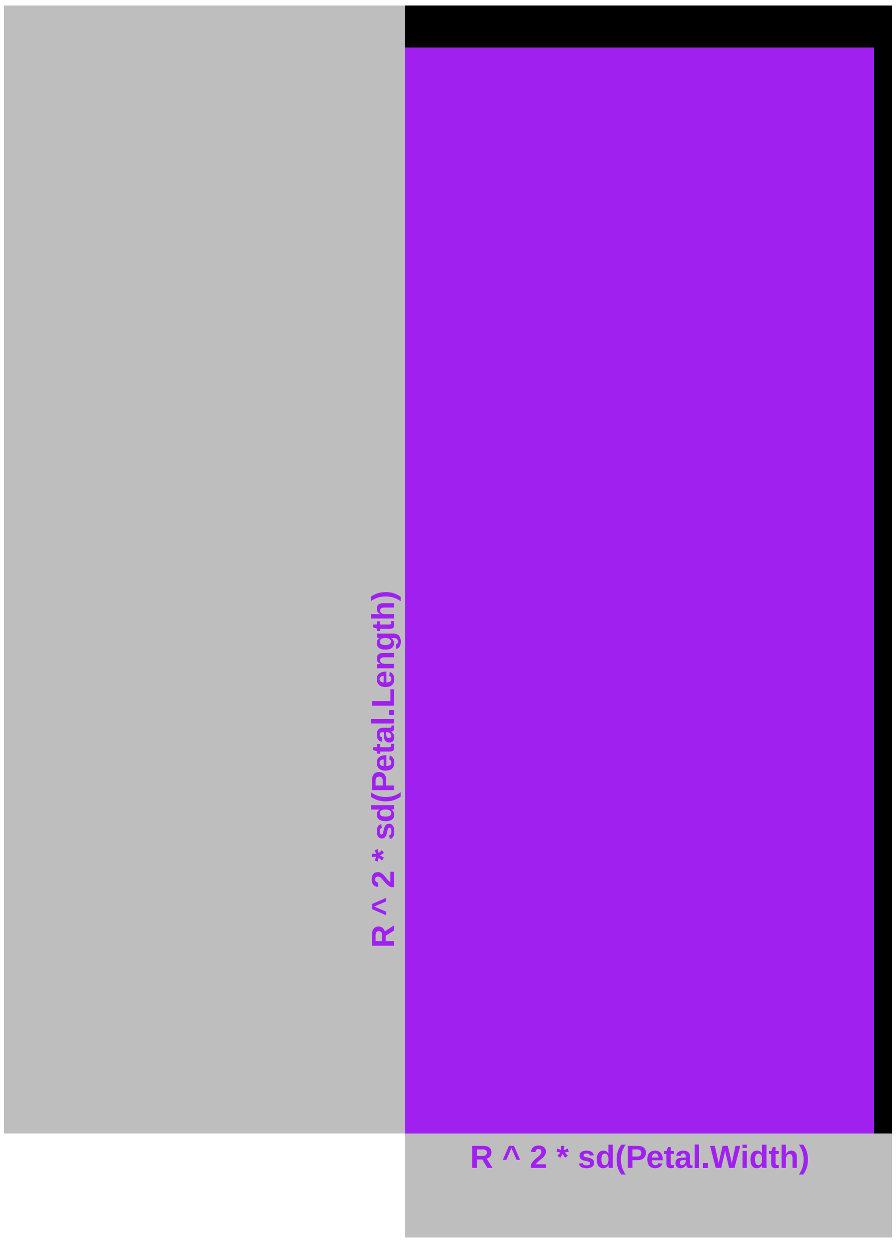


Squish covariance horizontally into the rectangle.

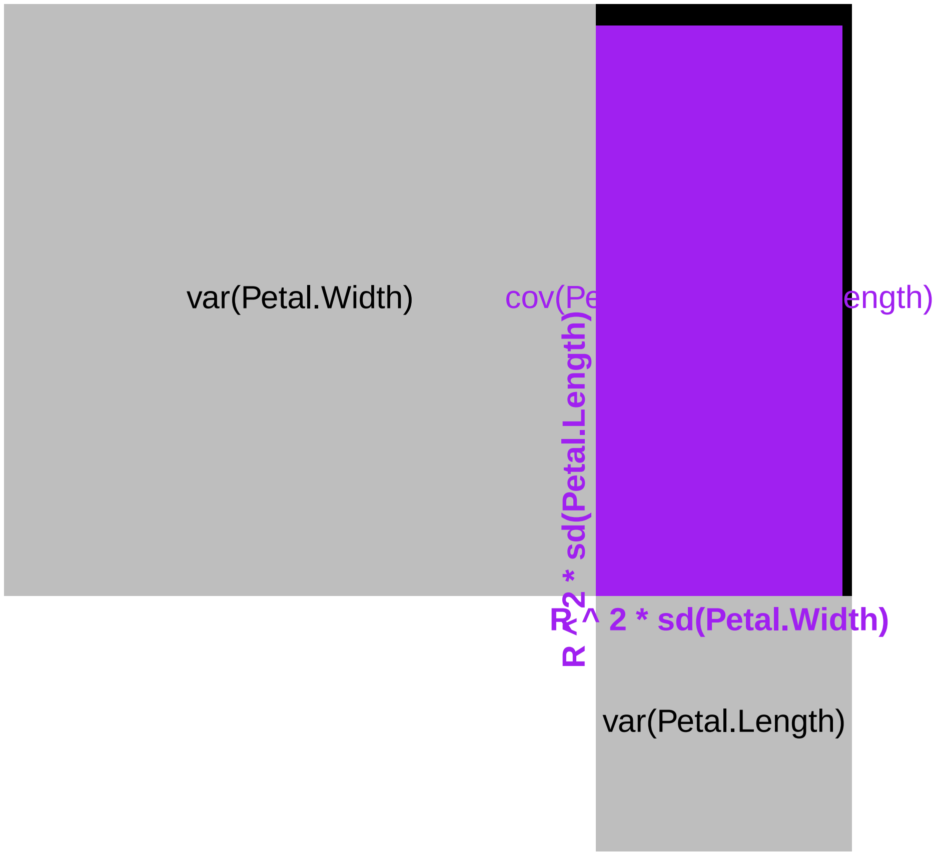


People like to
talk about R–squared.

Intersect the two squished covariance rectangles.

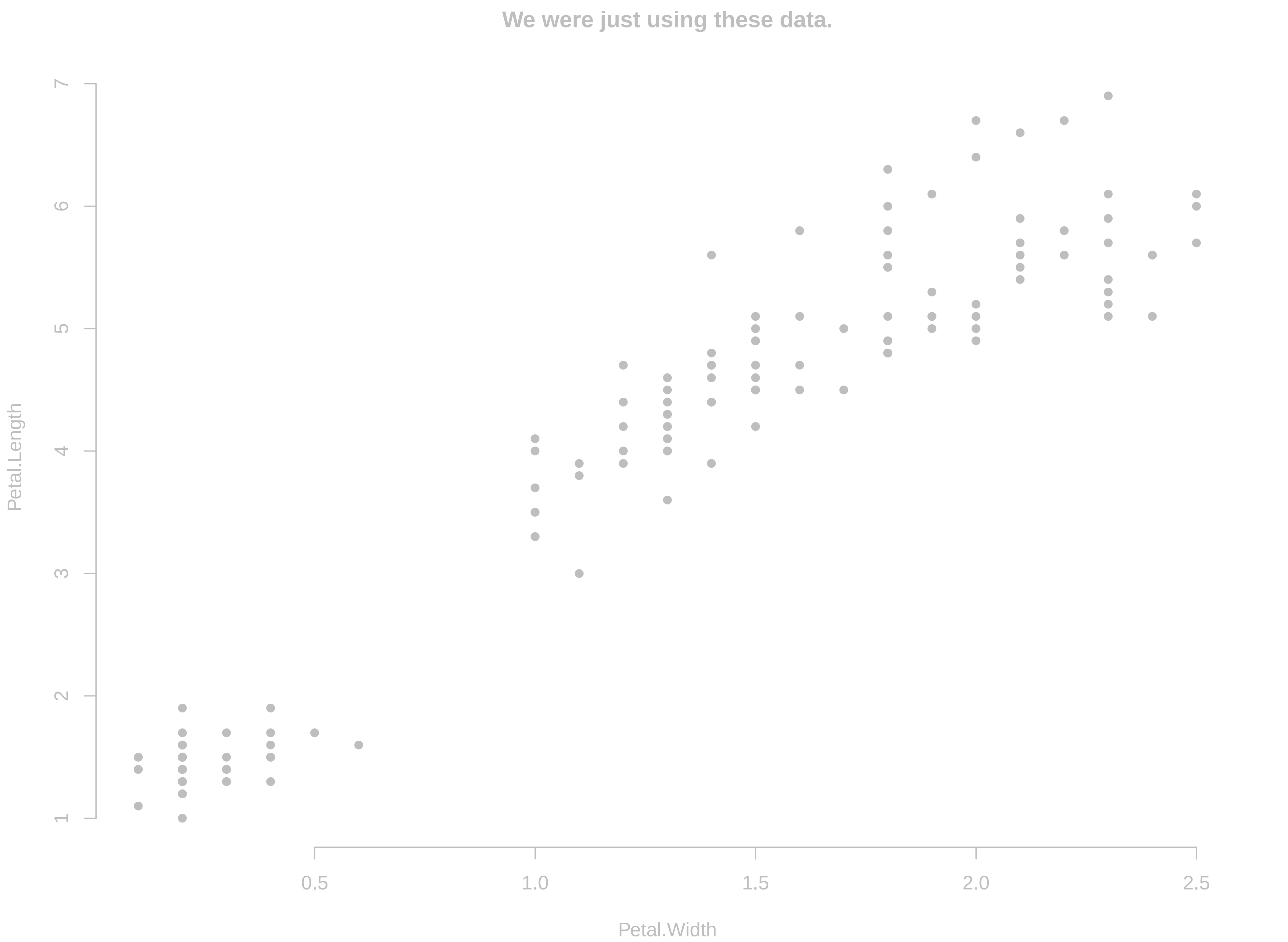


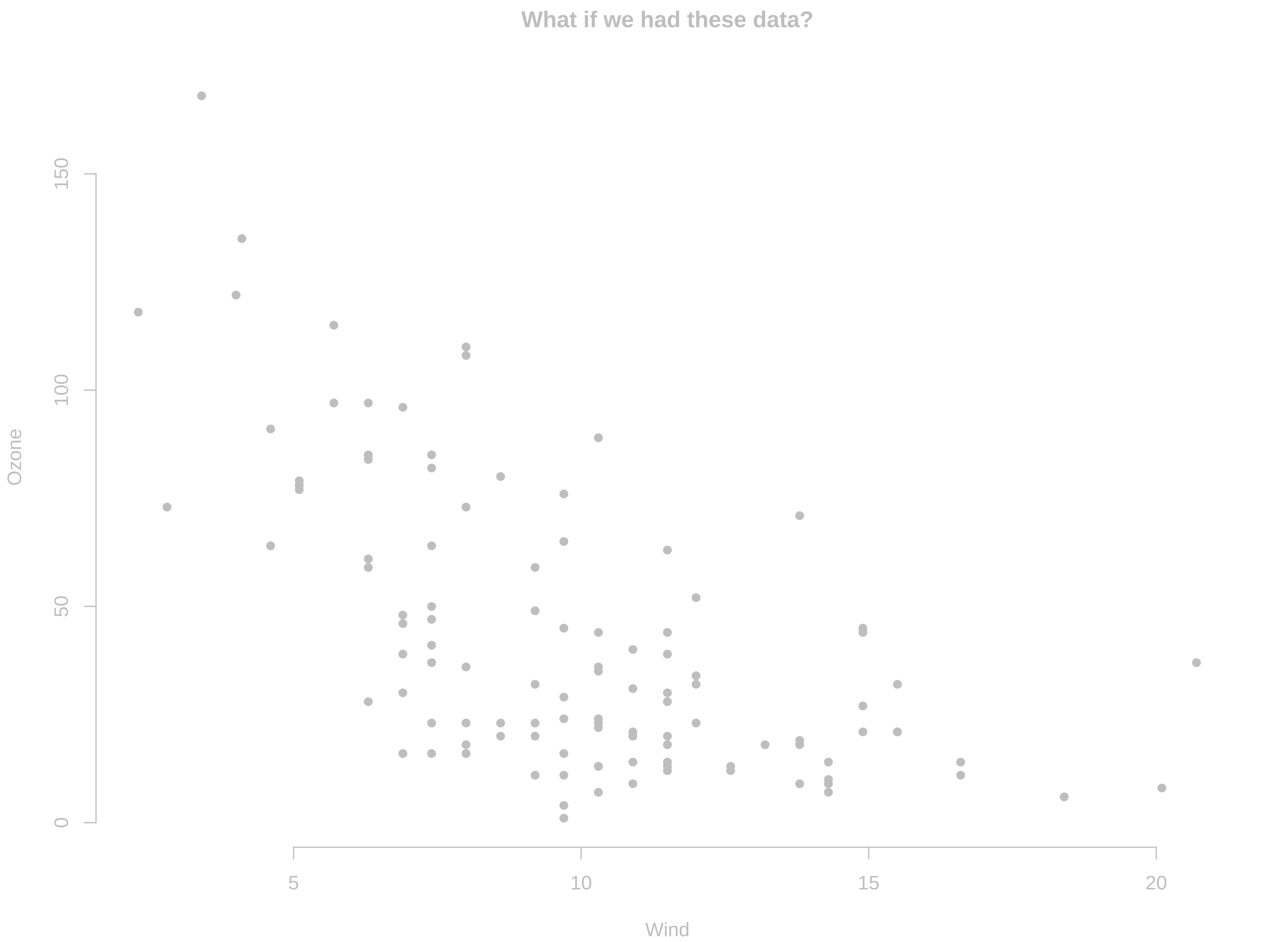
Intersect the two squished covariance rectangles.

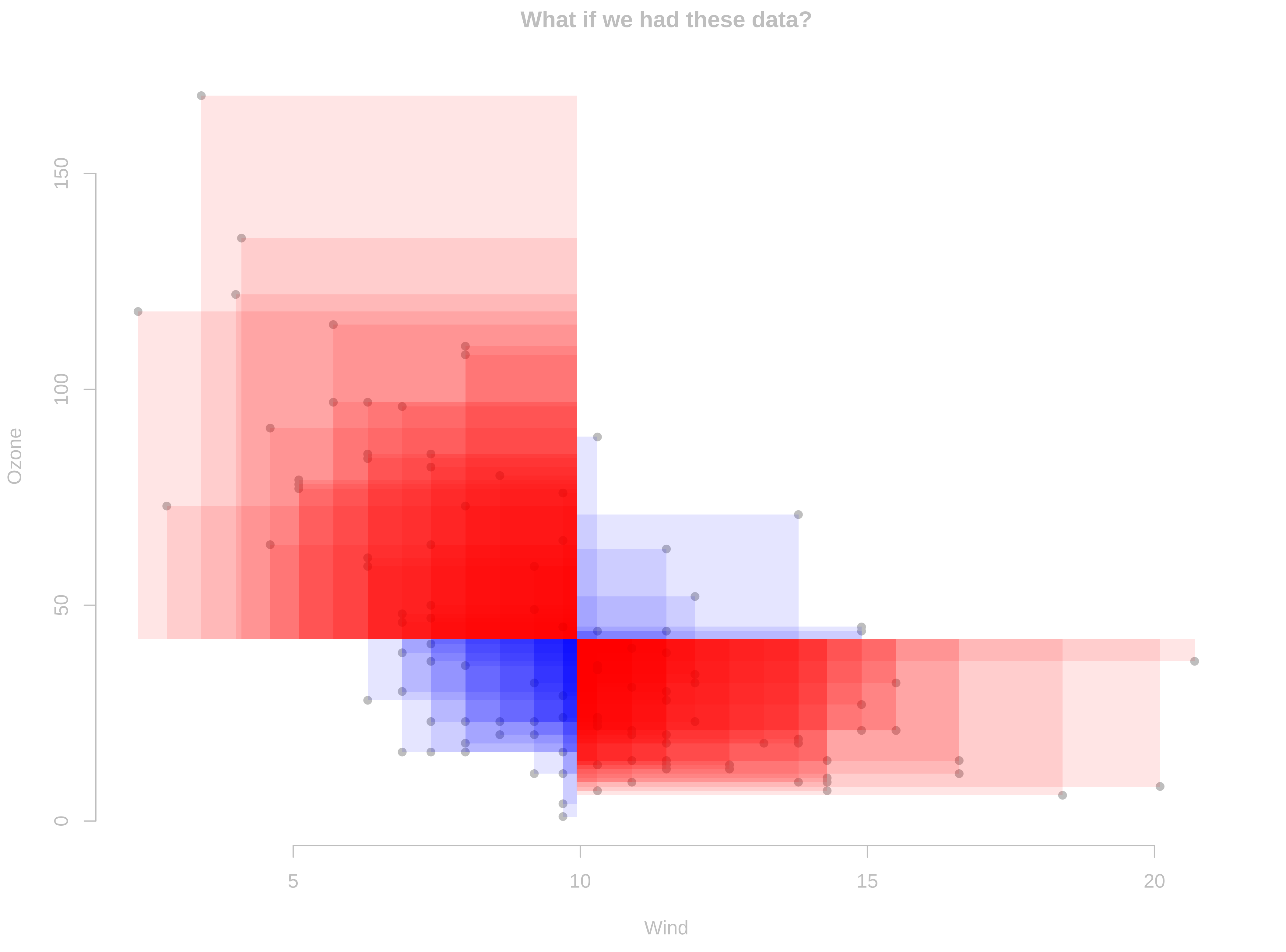


That was for very positive (blue) covariances.

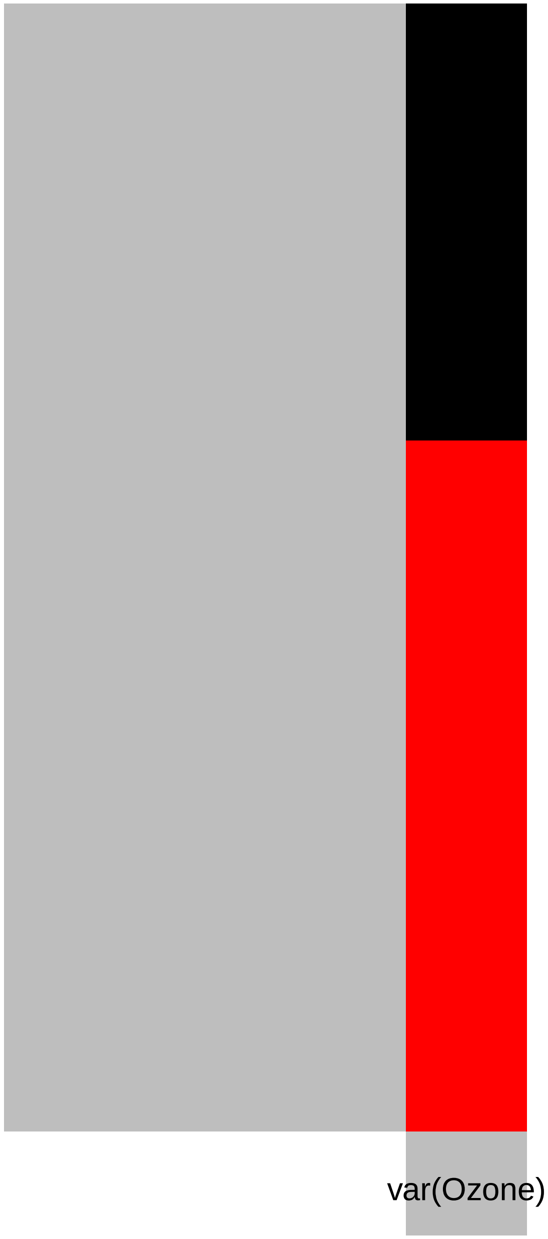
What if covariance is negative (red)?


We were just using these data.
Petal.Length
Petal.Width
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7

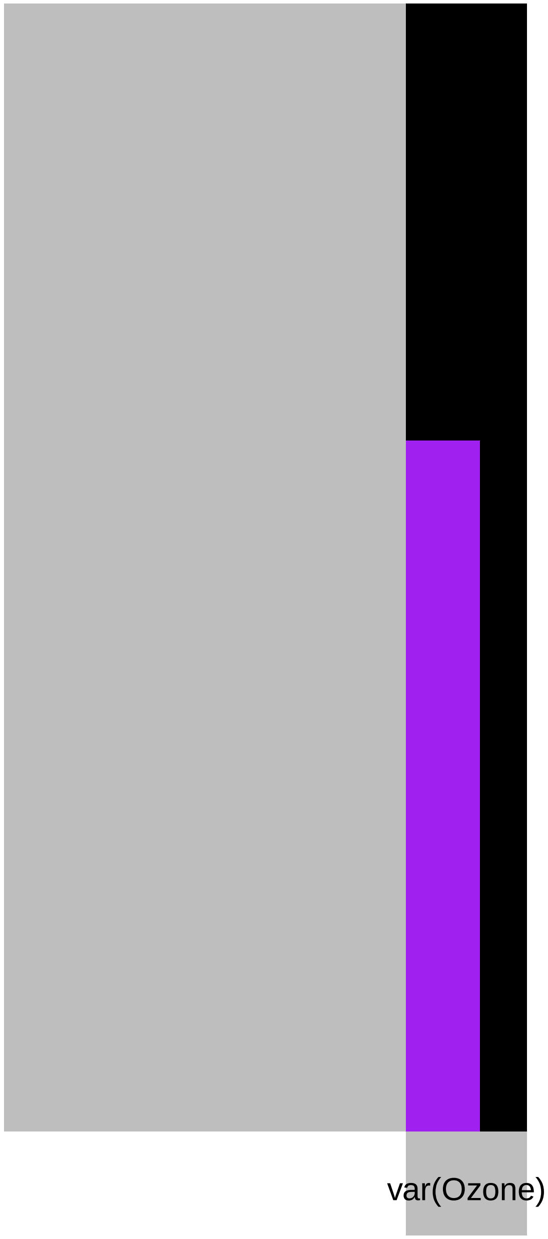

What if we had these data?
Ozone
0
50
100
150
Wind
5
10
15
20


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

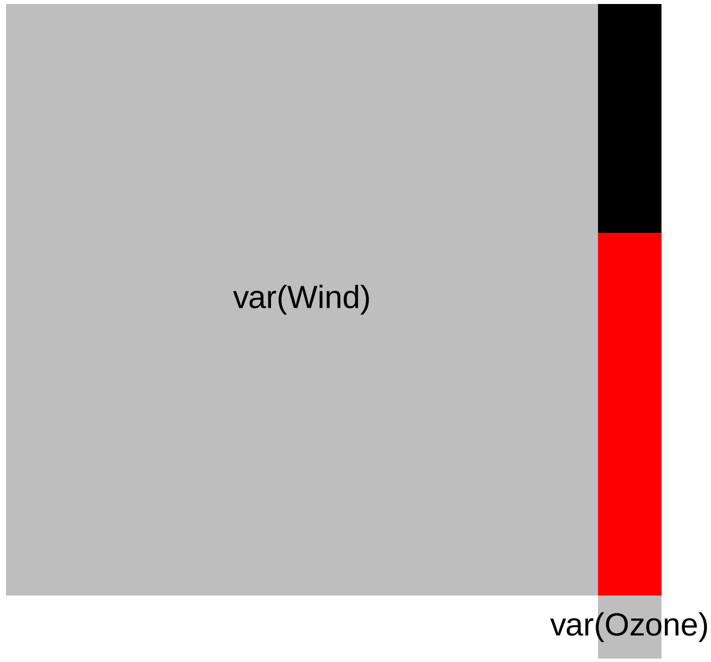
R is the same, just negative.



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 high
and x is high,
y should be high.

(b1 is very positive.)

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

(b1 is very negative.)

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

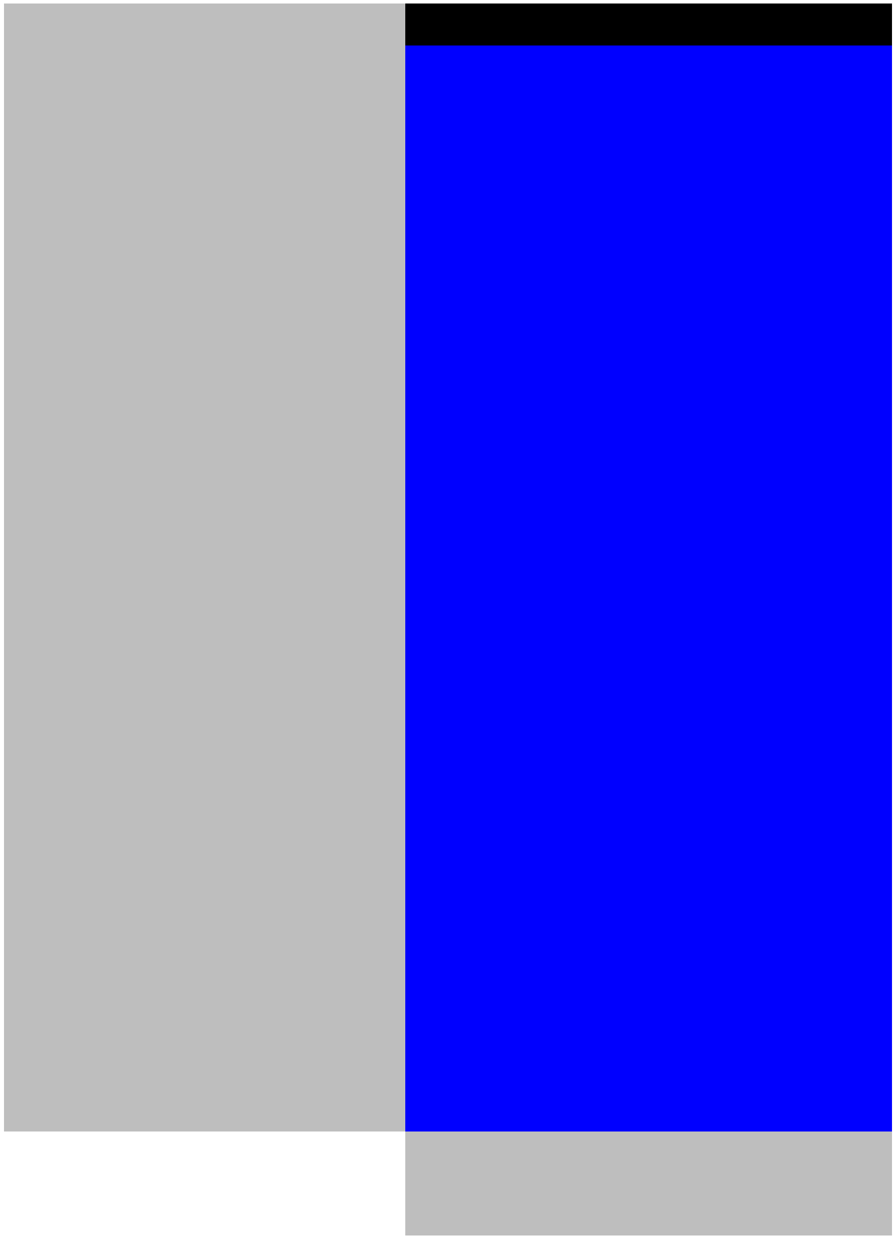
(b1 is around zero.)

Let's think about units again.

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

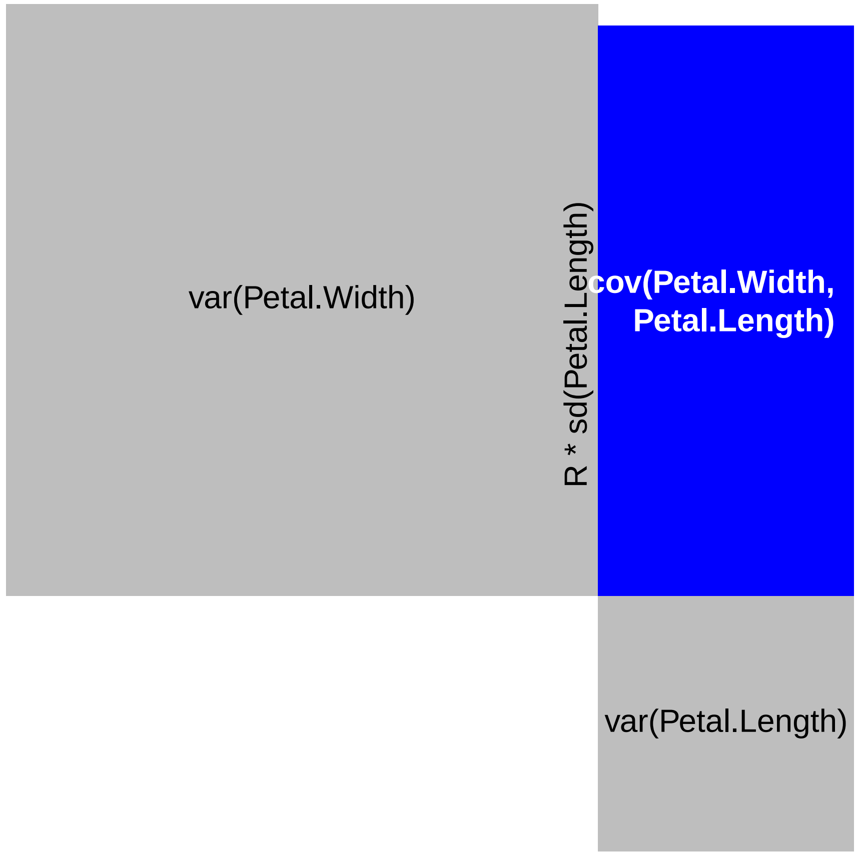
Variance is a special covariance; its unit is the square of the x unit.

Correlation is a ratio of areas with the same units.

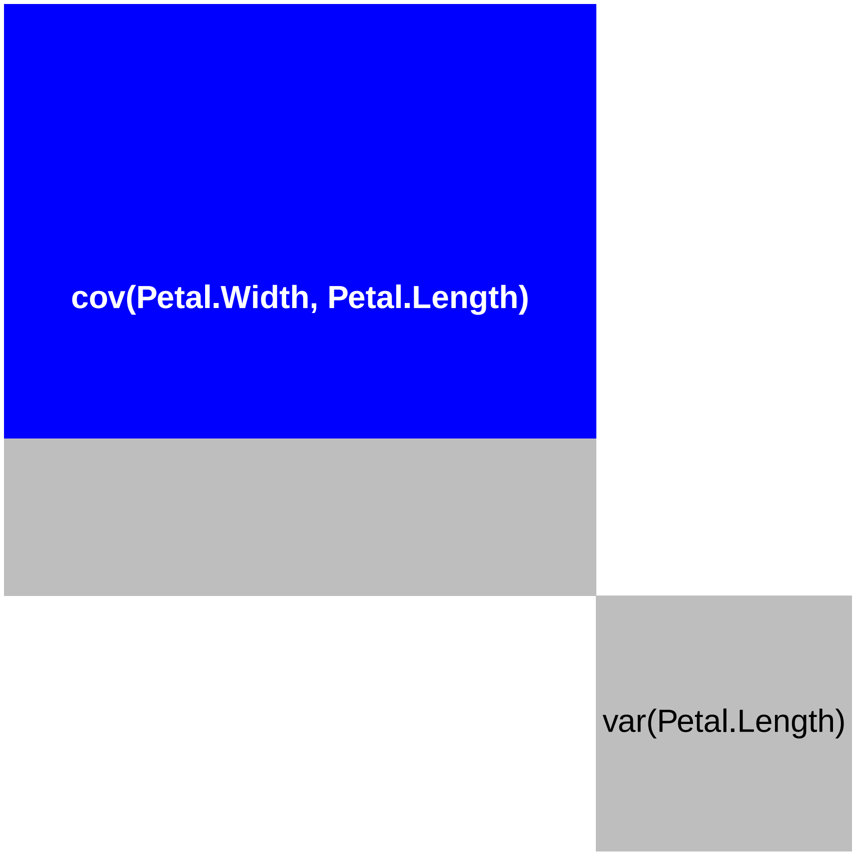


The unit of b1 must be y−unit/x−unit.

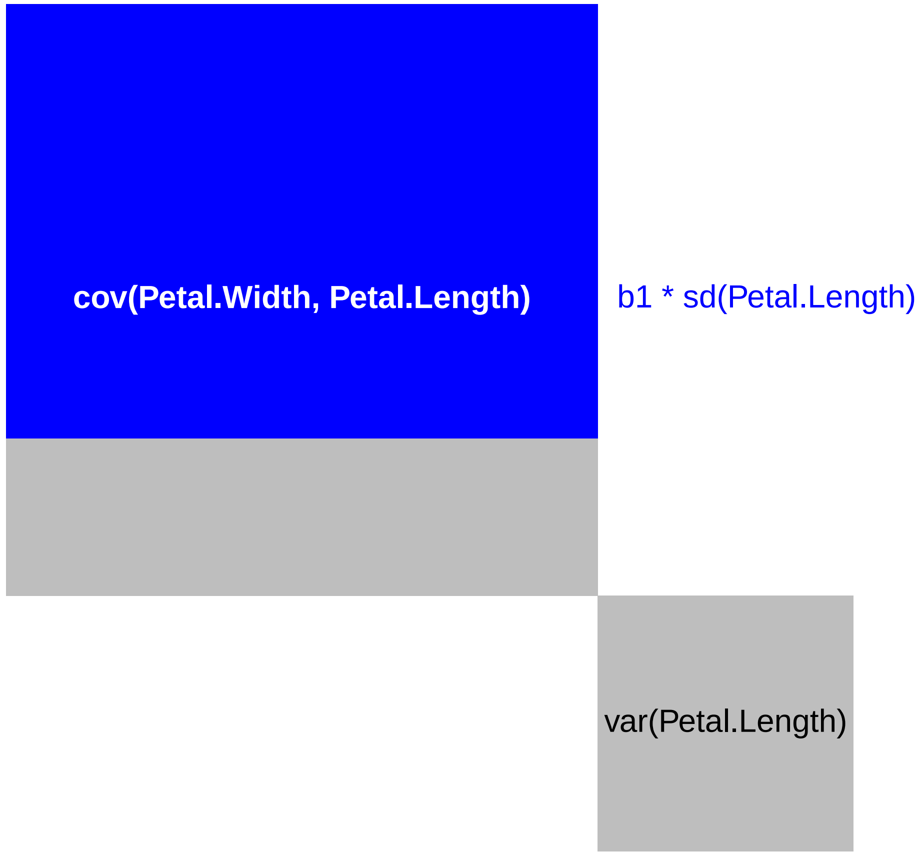
Our covariance picture



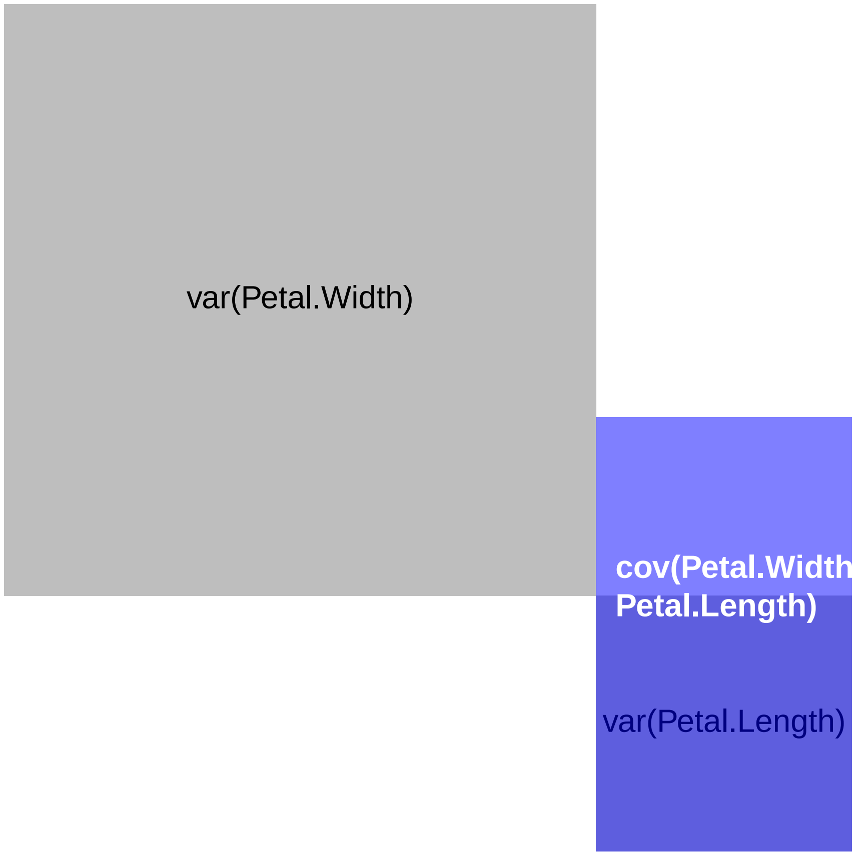
Lay the covariance over one of the variances instead.



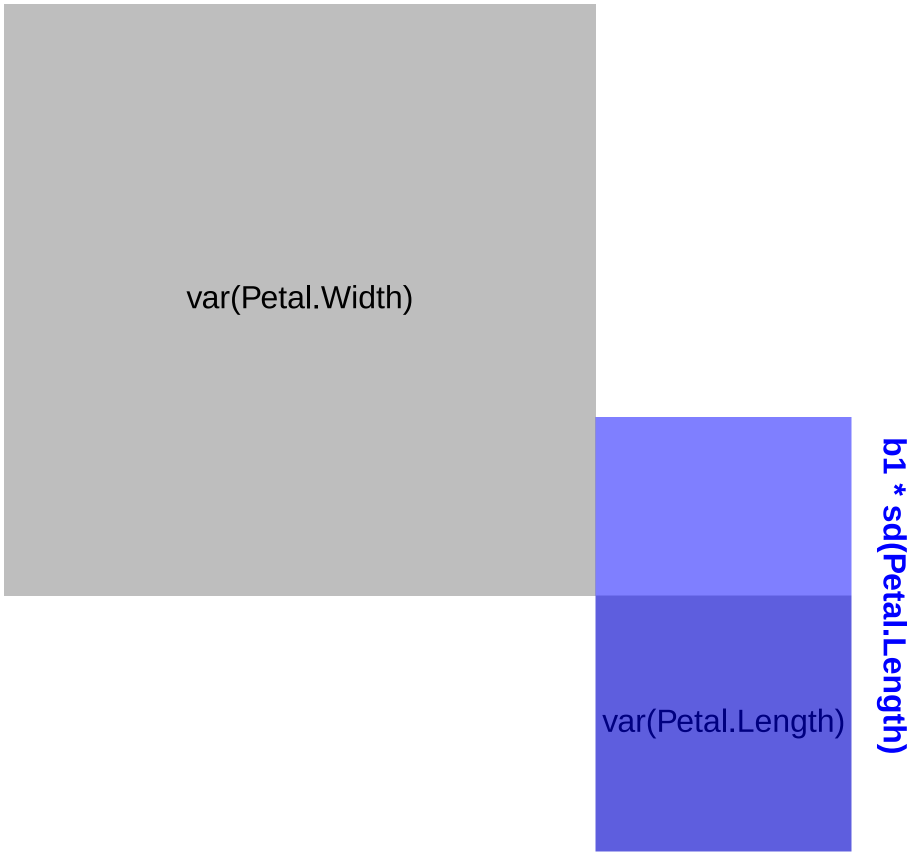
Petal.Width = b0 + b1 * Petal.Length



Lay the covariance over the other variance.



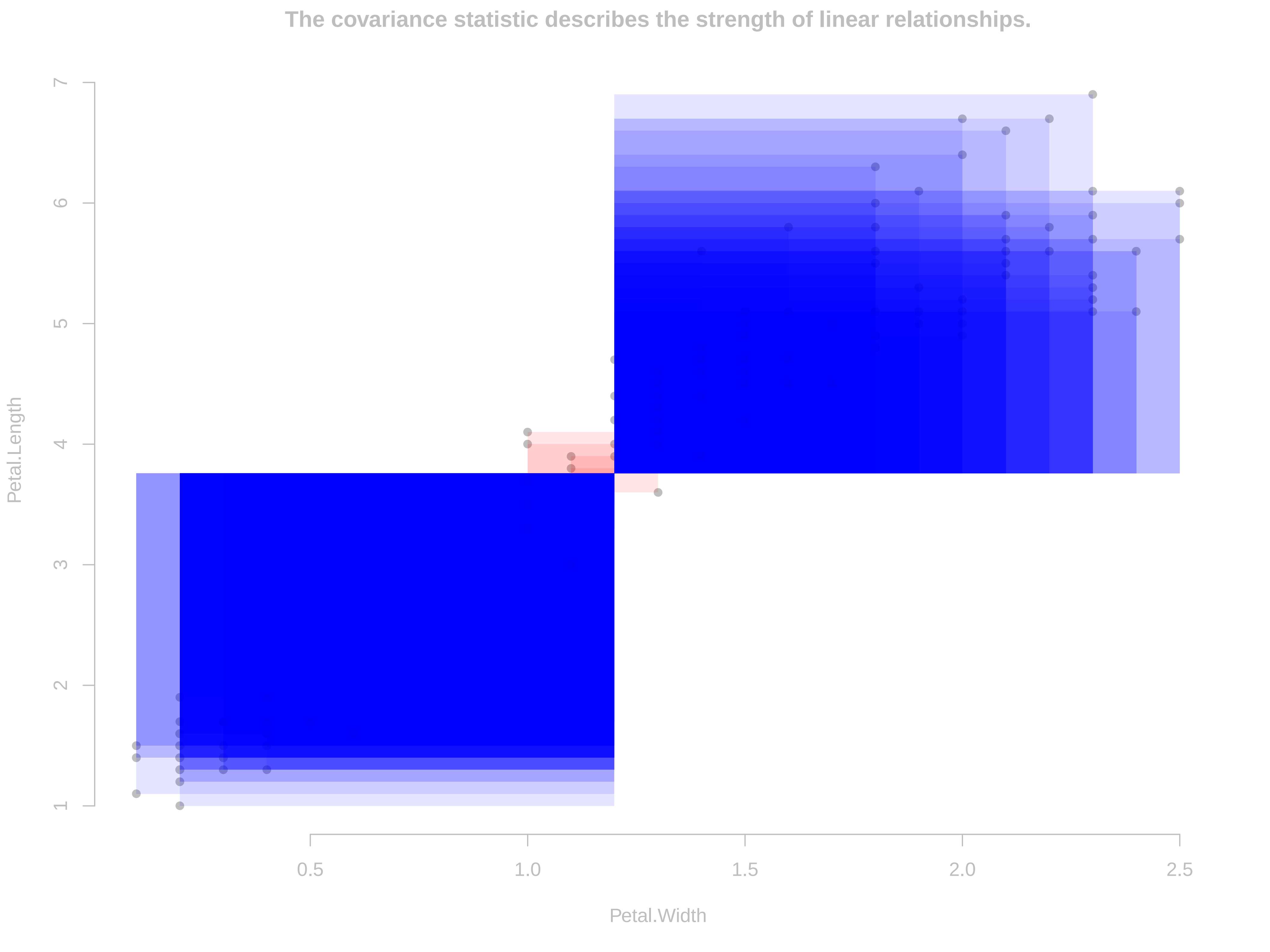
Petal.Length = b0 + b1 * Petal.Width

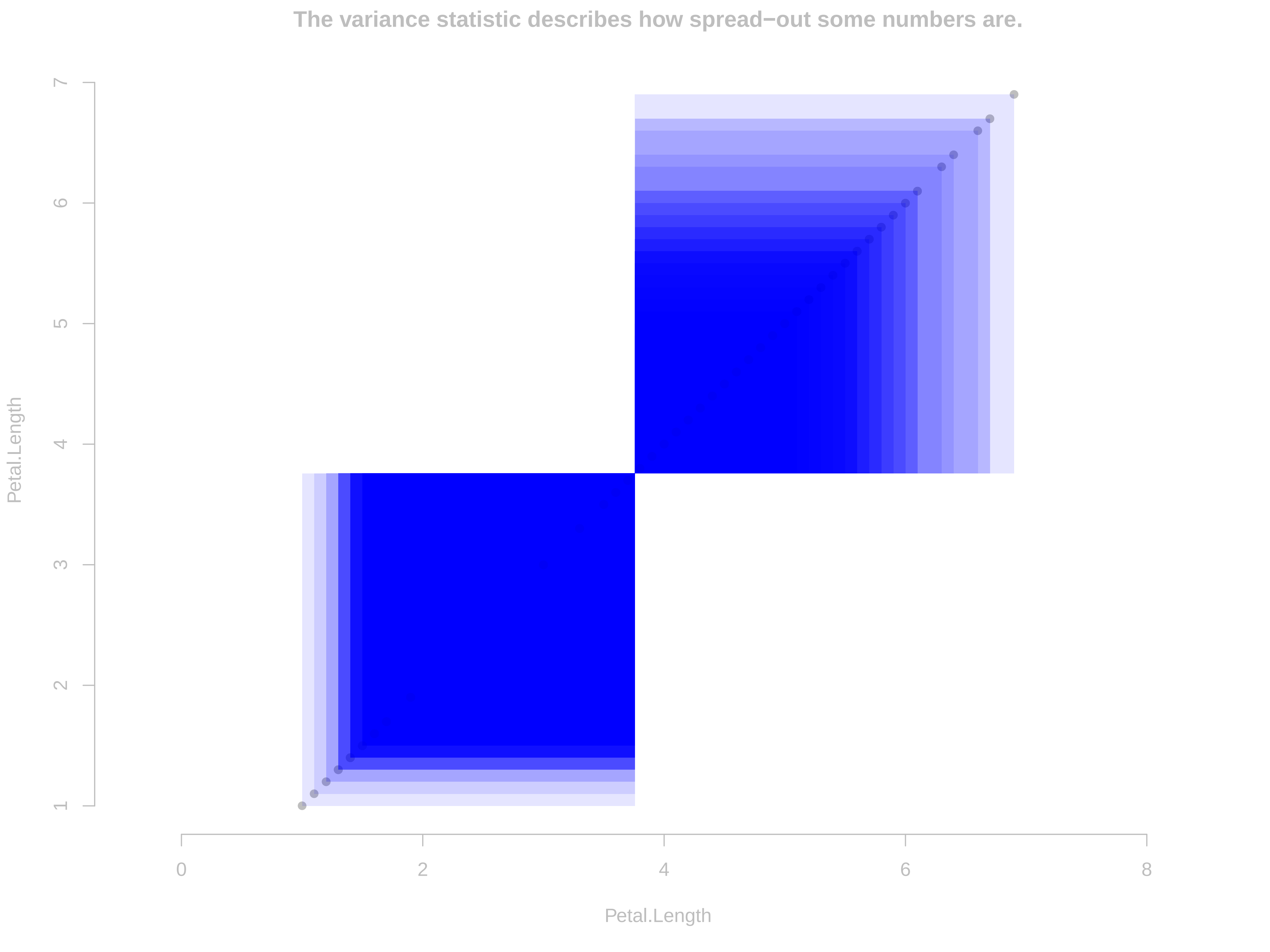


Some things to remember

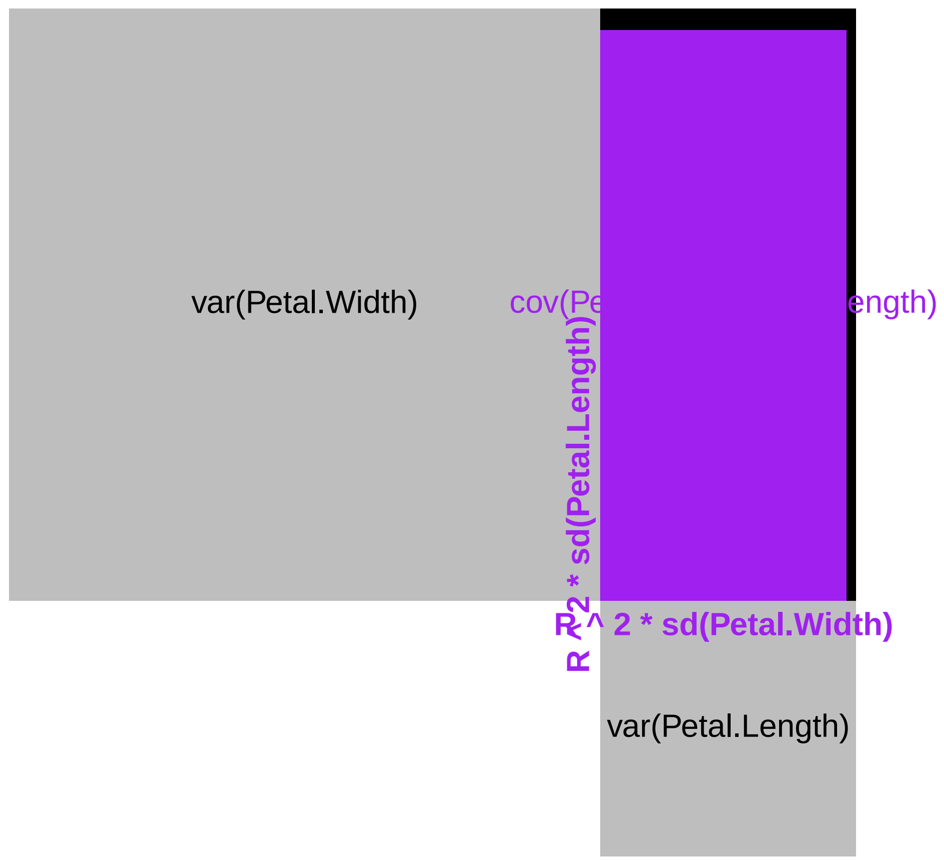
A statistic is a number that describes a lot of other numbers.

13	15	20	80	50	52	52	12	42	2	74	56	87	70	89	3	87	12	26	26	1
26	1	42	11	100	68	22	6	94	31	17	61	59	53	84	84	20	32	0	66	31
66	31	53	79	93	0	40	67	93	51	13	10	84	98	71	95	61	97	33	70	58
70	58	46	10	6	63	83	37	19	72	62	92	53	42	70	54	34	53	22	22	53
22	53	71	24	16	34	5	61	95	39	91	14	36	92	34	11	86	12	53	24	21
24	21	37	27	32	27	94	36	2	30	63	23	7	6	31	29	98	88	41	85	11
85	11	48	61	16	32	24	51	32	48	20	12	56	36	97	63	36	87	26	5	56
5	56	95	36	88	42	73	49	63	99	98	64	57	54	95	20	67	23	80	69	60
69	60	54	76	92	83	0	46	72	23	2	96	38	76	60	10	87	50	69	58	23
58	23	78	41	87	11	58	18	11	2	34	39	56	85	64	34	35	65	40	9	95
9	95	62	36	49	65	94	72	74	8	25	28	49	92	34	25	35	14	44	13	32
13	32	85	36	52	57	99	32	43	74	48	79	59	12	23	67	91	5	59	39	74
39	74	87	28	58	92	94	46	88	63	47	37	40	60	4	16	4	77	5	41	76
41	76	25	37	73	98	7	29	52	28	47	97	70	90	75	94	87	46	32	27	46
27	46	49	14	8	52	43	29	54	12	82	20	26	70	53	84	28	5	94	16	31
16	31	33	98	56	40	45	51	83	18	77	34	97	47	61	95	24	57	87	1	36
1	36	67	8	70	82	1	6	21	45	43	99	22	15	37	14	86	29	45	47	73
47	73	69	94	5	20	35	45	67	29	38	94	33	69	95	6	89	62	75	14	9
14	9	36	99	16	27	83	57	93	60	52	43	14	54	65	31	91	99	64	70	89
70	89	18	66	13	97	91	16	3	73	95	21	60	9	96	74	21	11	62	64	18
64	18	63	49	63	18	52	6	95	4	51	5	79	15	32	70	60	80	52	70	13

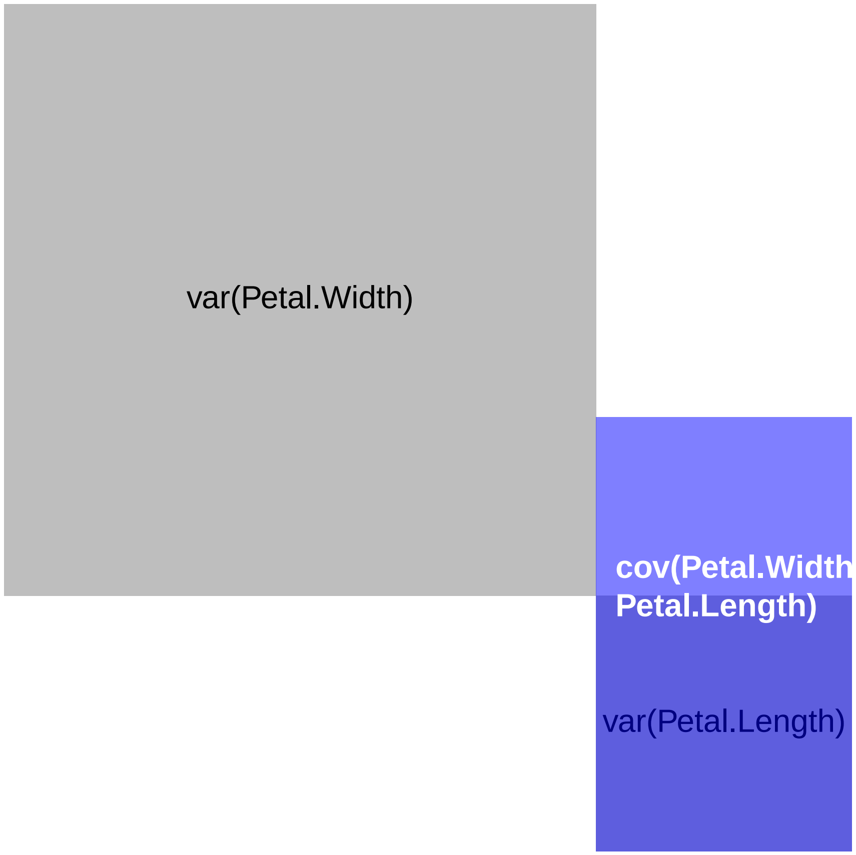

The covariance statistic describes the strength of linear relationships.
Petal.Length
Petal.Width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5


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

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.