

9

In this StatQuest, we will walk through Principal Component Analysis (PCA) one step at a time. We'll start with the basic idea of Principal Component Analysis (PCA), then move on to how PCA decomposes data.

You'll learn what PCA does, and how to use it to gain insight into your data.

Mouse	Mouse	Mouse	Mouse
1	2	3	4

In this StatQuest, we will walk through Principal Component Analysis (PCA) one step at a time. We'll start with the basic idea of Principal Component Analysis (PCA), then move on to how PCA finds the components, and finally how to use PCA to gain insight into your data.

Value Decomposition

You'll learn what PCA does, and how to use it to gain insight into your data.



In this StatQuest, we
through Principal Com
(PCA) one step at a tim
Value Decomposit

You'll learn what PCA
does it, and how to use
insight into yo



In this StatQuest, we
through Principal Com
(PCA) one step at a tim
Value Decomposit

You'll learn what PCA
does it, and how to use
insight into yo

In this StatQuest, we will walk through Principal Component Analysis (PCA) one step at a time. We'll start with the basic idea of Principal Component Analysis (PCA), then move on to how PCA finds the components, and finally we'll learn how to use PCA to gain insight into your data.

Value Decomposition

You'll learn what PCA does, and how to use it to gain insight into your data.

In this StatQuest, we will walk through Principal Component Analysis (PCA) one step at a time. We'll start with the basic idea of Principal Component Analysis (PCA), then move on to how PCA finds the components, and finally we'll learn how to use PCA to gain insight into your data.

Value Decomposition

You'll learn what PCA does, and how to use it to gain insight into your data.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

We've measured transcript levels for two genes, Gene 1 and Gene 2.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Sample 1	Sample 2	Sample 3
Variable 1	10	11	8
Variable 2	6	4	5

	Sample 1	Sample 2	Sample 3
Variable 1	10	11	8
Variable 2	6	4	5

...and the genes as variables we measure for each sample

	Student 1	Student 2	Student 3
Math	95	88	93
Reading	96	79	98

	Business 1	Business 2	Business 3
Market Cap	9.5 million	88 million	93 million
# Employed	960	79	93

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

0

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3



	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3



	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5



	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	7	6	2

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

We'll start by plotting

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3

From this point on, we
what happens in the gr
longer need the origin

Now we'll shift to
that the center is
the origin $(0,0)$ in



Now we'll shift to
that the center is
the origin $(0,0)$ in



NOTE: Shifting the
change how the data
positioned *relative*
other.





Now that the data
centered on the origin,
can try to fit a line

To do this...



...we start by
random lines
through the



...then we rotated
it fits the data
can, given that
through the



Ultimately
be



Ultimately
be

...but I'm getting
myself. First, we
about how PCA
fit is good



So let's go back
original “random”
goes through the



To quantify how good a line fits the data, we project the data onto the line.





...or it can try to find a line that ***maximizes*** the distances from projected points to the origin.



If those options do not
equivalent to





**...while these
get larger when
fits better**



**...while these
get larger when
fits better**

















...then we can use
Pythagorean theorem
to find out
how **b** and **c** are
related.

$$a^2 = b^2 + c^2$$



Since a (and thus a^2)
doesn't change...

$$a^2 = b^2 + c^2$$



If b gets big

$$a^2 = b^2 + c$$



....then c must ge

$$a^2 = b^2 + c$$

Likewise, if c gets

$$a^2 = b^2 + c^2$$

...then b must ge

$$a^2 = b^2 + c^2$$



Thus, PCA can either
the distance to t

$$a^2 = b^2 + c^2$$



...or ***maximize*** the
the projected point

$$a^2 = b^2 + c^2$$



The reason I'm making
about this is that, it
makes sense to minimize
distance from the point

$$a^2 = b^2 + c^2$$



The reason I'm making
about this is that, it
makes sense to minimize
distance from the point

$$a^2 = b^2 + c^2$$



...but it's actually easier
the distance from the p
to the origin, so PCA t
fitting line by maximizi
**the squared distanc
projected points to**

$$a^2 = b^2 + c^2$$



So for this



....PCA projects t



...and then measures the distance from this point to the origin (and call it d_1)...



d_1



NOTE: I'm going to
keep track of the
distances we measure
up here...



d_1 d_2

...and then PCA measures
distance from this point
to the origin...



d_1 d_2 d_3

...and then PCA measures
distance from this point
to the origin...



d_1 d_2 d_3

...and then PCA measures
distance from this point
to the origin...



d_1 d_2 d_3

...and then PCA measures
distance from this point
to the origin...



d_1 d_2 d_3

...and then PCA measures
distance from this point
to the origin...



d_1

d_2

d_3

Here are all
that we have



d_1^2

d_2^2

d_3^2

The next three
square a



$$d_1^2 \quad d_2^2 \quad d_3^2$$

The distances are so
that negative val

$$d_1^2 \quad d_2^2 \quad d_3^2$$

The distances are so
that negative val

...don't cancel out
values.



$$d_1^2 + d_2^2 + d_3^2$$

Then we
these
dista



$$d_1^2+d_2^2+d_3^2$$



$$d_1^2+d_2^2+d_3^2$$



Now we rotate



...project
onto the



$$d_1^2 + d_2^2 + d_3^2$$

...and then
squared
from the
points to t



$$d_1^2 + d_2^2 + d_3^2$$

...and then
squared
from the
points to t



$$d_1^2 + d_2^2 + d_3^2$$

...and then
squared
from the
points to t



$$d_1^2 + d_2^2 + d_3^2$$

...and then
squared
from the
points to t



$$d_1^2 + d_2^2 + d_3^2$$

...and then
squared
from the
points to t



$$d_1^2 + d_2^2 + d_3^2$$

...and then
squared
from the
points to t



$$d_1^2 + d_2^2 + d_3^2$$

...and we repeat until we get a minimum. We do this by calculating the gradient of squared distances between each point and the line. We then move the line with the line with the largest negative gradient. We repeat this process until the gradient is zero or very small.



$$d_1^2 + d_2^2 + d_3^2$$

Ultimately, we
this line. It has
SS(distan



This line is called
Component 1. (PC)



PC1 ha









...and
spread
Ge



One way to think about
terms of a cocktail recipe



One way to think about
terms of a cocktail recipe

To make PC-



One way to think about
terms of a cocktail recipe

To make PC-
Mix 4 parts Gen-



One way to think about
terms of a cocktail recipe

To make PC-
Mix 4 parts Genet
with 1 part Genet



One way to think about
terms of a cocktail recipe

To make PC-
Mix 4 parts Genet
with 1 part Genet

Pour over ice and



One way to think about terms of a cocktail recipe

To make PCP
Mix 4 parts Gene 1
with 1 part Gene 2

The ratio of Gene 1 to Gene 2 tells you that Gene 1 is more important when it comes to describing

To make PC-
Mix 4 parts Gen-
with 1 part Gen-

Terminology Alert

Mathematicians call this
recipe a “linear combination.”

**To make PC-
Mix 4 parts Gen-
with 1 part Gen-**

I mention this because
someone says, “PC1 is
a combination of various

To make PC-
Mix 4 parts Gen-
with 1 part Gen-



**...this is what they
are talking about**

To make PC-
Mix 4 parts Gen-
with 1 part Gen-



...this is what they
are talking about

To make PC-
Mix 4 parts Gen
with 1 part Gen























$$\frac{4.12}{4.12} = \frac{\sqrt{4^2 + 1^2}}{4.12} = \sqrt{\left(\frac{4^2 + 1^2}{4.12}\right)}$$

$$= \sqrt{\left(\frac{4^2 + 1^2}{4.12}\right)}$$

For those of you keeping score at home,
here's the math worked out. It shows that all we need to do is divide all 3 sides by 4.12.



The new values change
the recipe...

To make PC-
Mix 0.97 parts GE
with 0.242 parts G

The new values change
the recipe...

To make PC-
Mix 0.97 parts Gene 1
with 0.242 parts Gene 2

...but the ratio is the same.
use 4 times as much Gene 1
as Gene 2.



So now we are back



So now we are back

- The data



So now we are back

- The data
- The best fitting line

So now we are back

- The data
- The best fitting line
- The unit vector that calculated.



To make PC-
Mix 0.97 parts GE
with 0.242 parts G

...and the proportions of
are called “Loading S



$$d_1^2+d_2^2+d_3^2$$



$$d_1^2 + d_2^2 + d_3^2$$

$$\frac{\text{SS(distances for PC1)}}{n - 1} =$$



$$d_1^2 + d_2^2 + d_3^2$$

$$\frac{\text{SS(distances for PC1)}}{n - 1} =$$

$$\sqrt{\text{SS(distances for PC1)}}$$

That's a



Now that we've got
figured out let's work



Because this is a graph, **PC2** is simply through the origin perpendicular to **P**. Any further optimization to be done



....and this mean
recipe for PC

-1 Parts Ge
4 Parts Ge



If we scale everything
we get a unit vector
that is...

-0.242 Parts C
0.97 Parts G



**-0.242 Parts C
0.97 Parts G**



These are the L
Scores for

-0.242 Parts C
0.97 Parts G



These are the
Scores for

-0.242 Parts C
0.97 Parts G

They tell us that,
of how the value
projected onto PC
2 is 4 times as im-

$$d_1^2 + d_2^2 + d_3^2$$

$$\frac{\text{SS(distances for PC2)}}{n - 1} =$$



Hooray!

We've worked out



To draw the fir

We simply rotate
that PC1 is l

We simply rotate
that PC1 is 1

...then we use the
to find where the
the PC

For example,
projected
correspond to S

...so Sample 6 goes h

Double

That's how
using Si
Decompo

OK

s

Remember the equation

$$\frac{\text{SS(distances for PC1)}}{n - 1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1}$$

Remember the equation

$$\frac{\text{SS(distances for PC1)}}{n - 1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1}$$

Remember the equation:

$$\frac{\text{SS(distances for PC1)}}{n - 1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1}$$

Remember the e



$$\frac{\text{SS(distances for PC1)}}{n - 1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1}$$

Well, if you are familiar
for variation, you will note
are just measures

$$\frac{\text{SS(distances for PC1)}}{n - 1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1}$$

For the sake of the example, let's assume that the Variation for the variation for

$$\frac{\text{SS}(\text{distances for PC1})}{n - 1}$$

$$\frac{\text{SS}(\text{distances for PC2})}{n - 1}$$

For the sake of the example, let's assume that the Variation for the variation for

$$\frac{\text{SS}(\text{distances for PC1})}{n - 1}$$

$$\frac{\text{SS}(\text{distances for PC2})}{n - 1}$$

For the sake of the example, let's assume that the Variation for PC1 is 100% and the variation for PC2 is 90%.

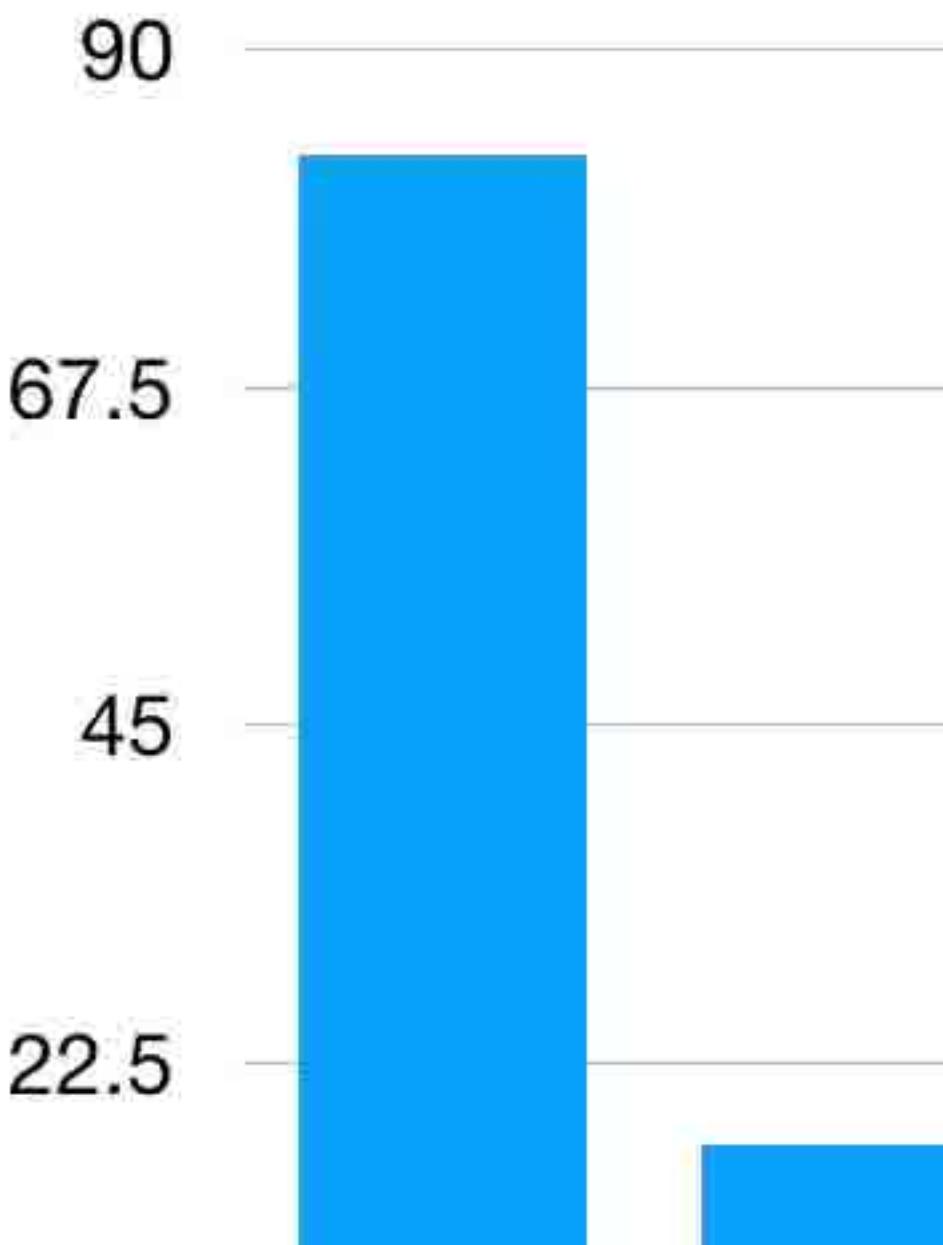
$$\frac{\text{SS(distances for PC1)}}{n - 1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1}$$

PC2 accounts for
17% of the total
the P

TERMINOLOGY ALL

Plot is a graphical representation showing the percentages of various PC accounts.



We'll talk more about
later..

90

67.5

45

22.5



•
•

	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5

You c



You then find
line that goes
through the
origin.



Just
best f



But the recipe
now has 3 ingre

0.62 Parts G

0.15 Parts G

0.77 Parts G

0.62 Parts G

0.15 Parts G

0.77 Parts G

In this case, Gene C
is an important ingredient.



You then find PC2,
line given that it
origin and it is pe



You then find PC2,
line given that it
origin and it is pe



You then find PC2,
line given that it
origin and it is pe



Here's the record
PC2...

0.77 Parts G

0.62 Parts G

0.15 Parts G



0.77 Parts G

0.62 Parts G

0.15 Parts G

In this case, Gene -
important ingredient



Lastly, we find the best fitting line through the origin perpendicular to PC2



If we had more genes
finding more and
components by adding
and rotating



If we had more genes
finding more and
components by adding
and rotating

In theory there is one p
but in practice, the num
number of variables or
whichever i



If this is confusing,
not super important
make a separate vid
the next

In theory there is one p
but in practice, the num
number of variables or
whichever i



Once you have all components figured out, calculate eigenvalues (i.e. $SS(\text{diagonal}) / \text{number of variables}$) to determine the proportion of variance each component accounts for.

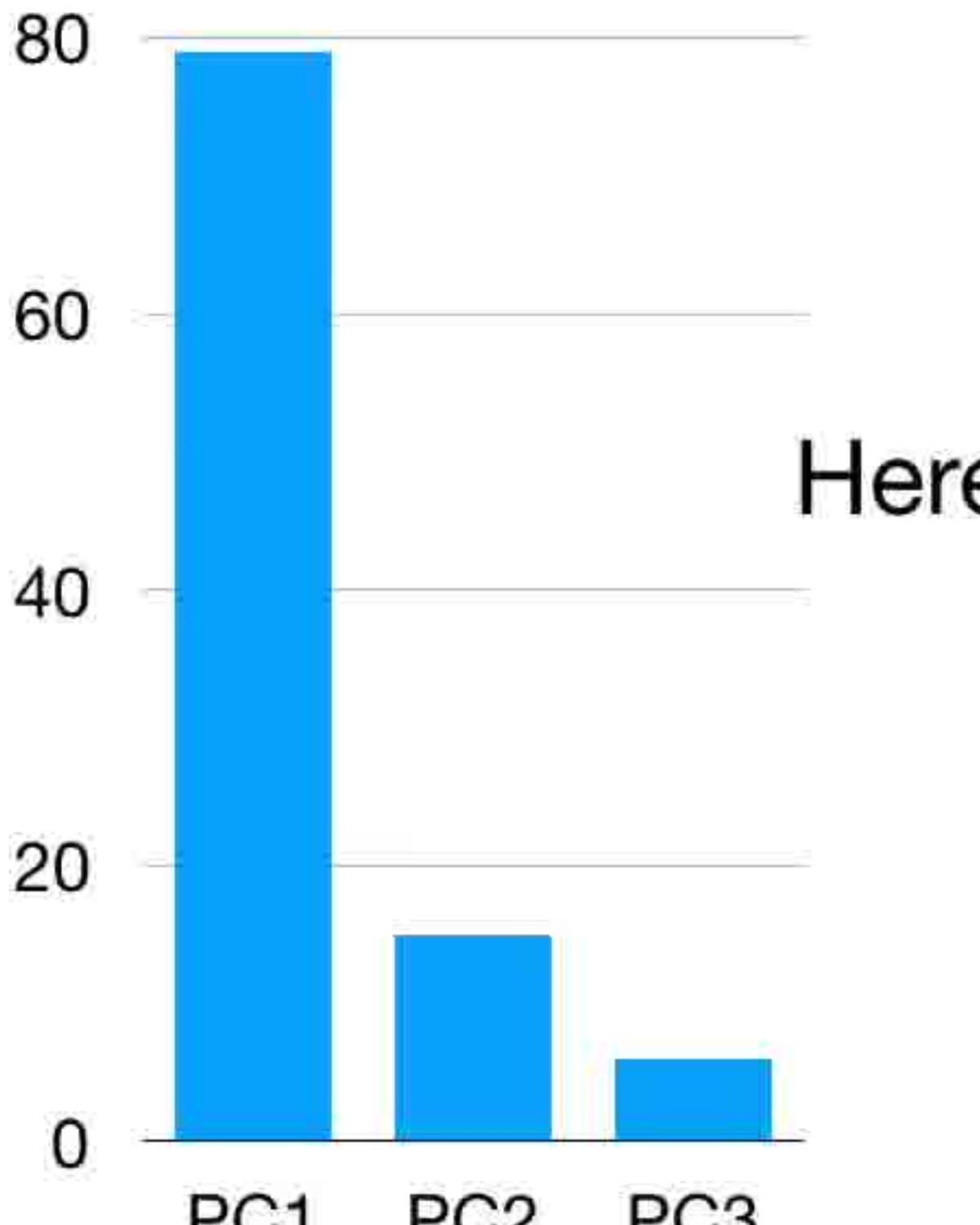






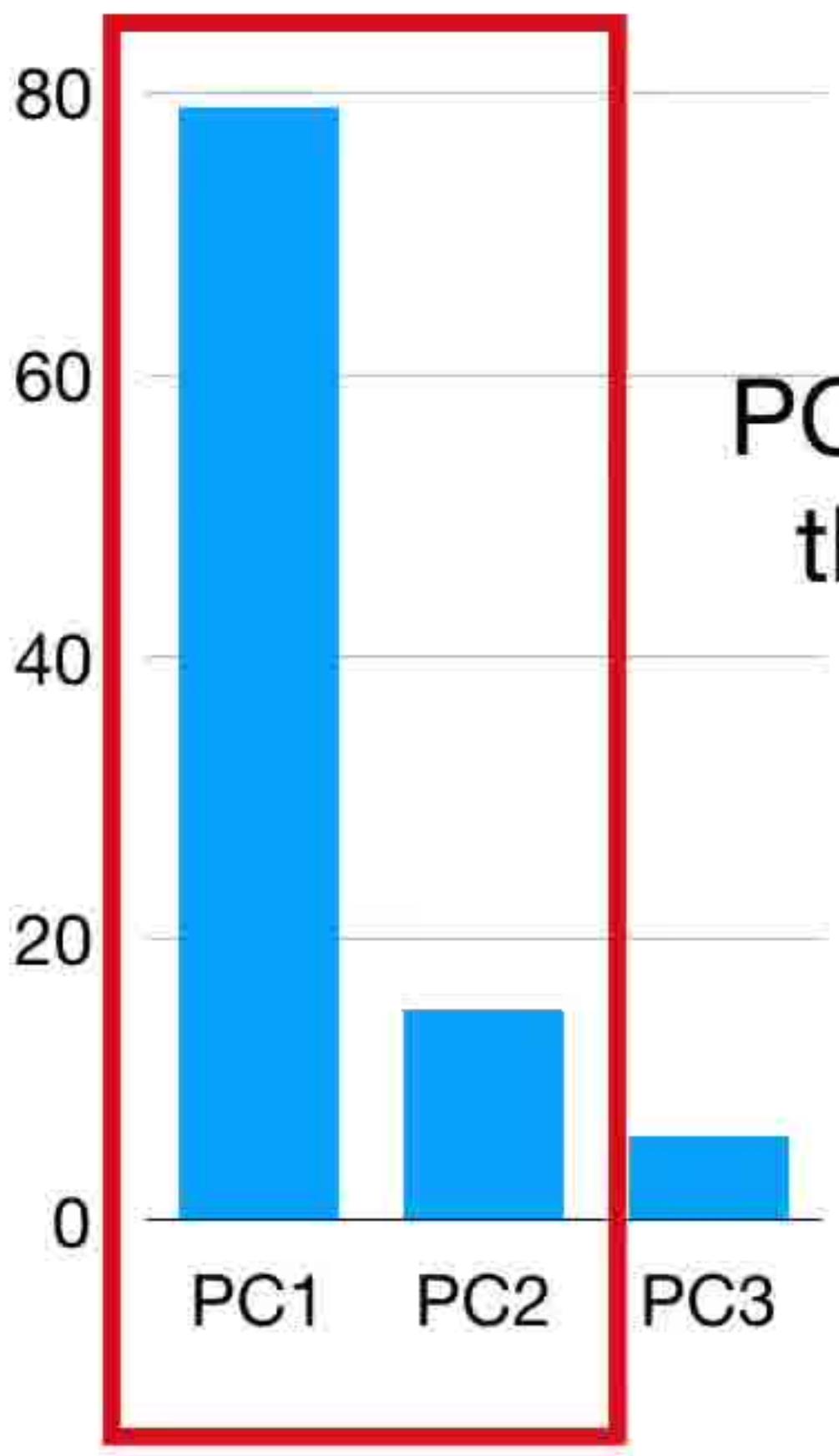
... ■ ■ ■





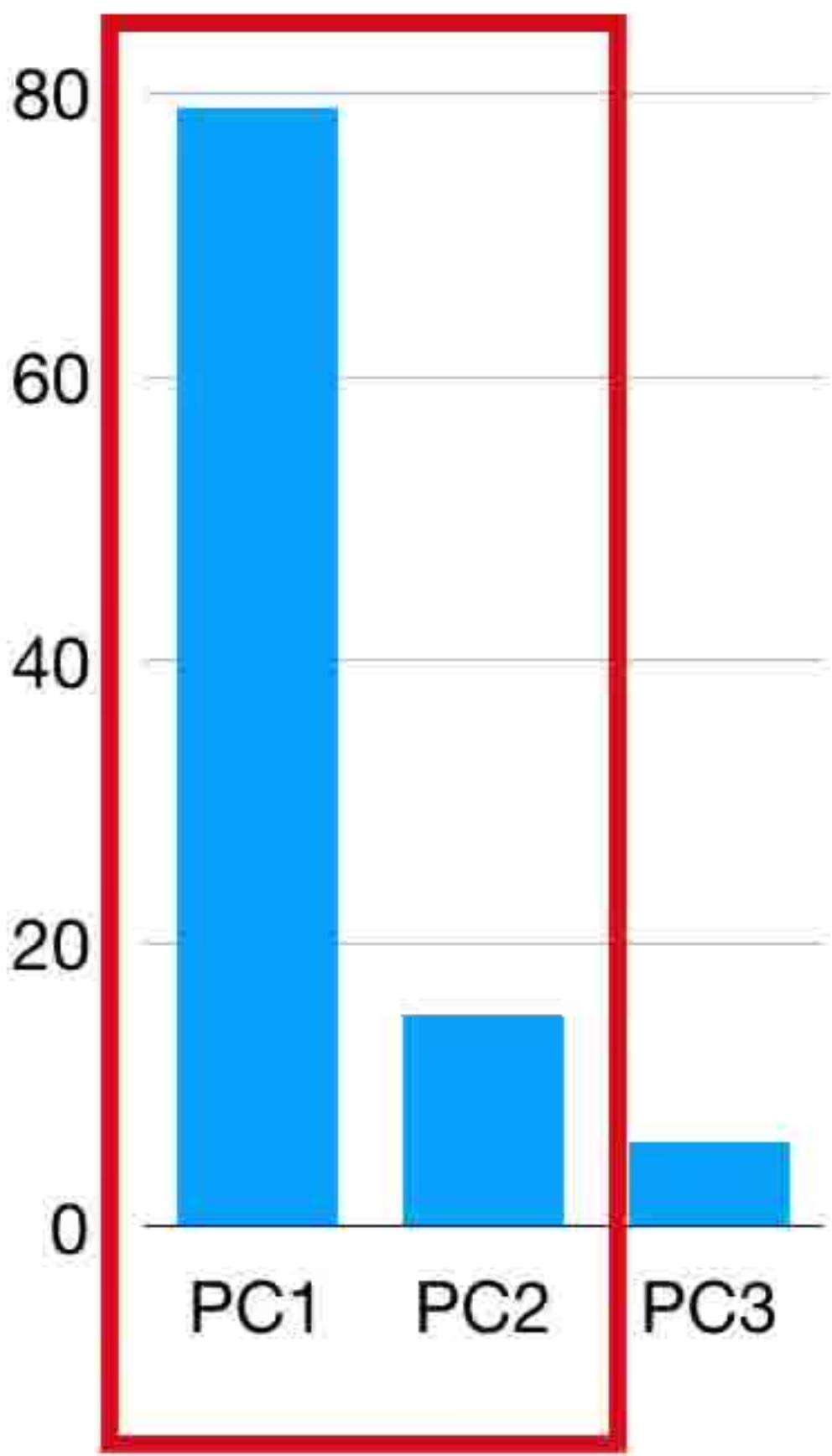
Here





PC1 at
the w





Than
and
the



To convert the 3-D graph, we just strip away
data and PC1

To convert the 3-D graph, we just strip away data and PC1

Then project the sam

...and □

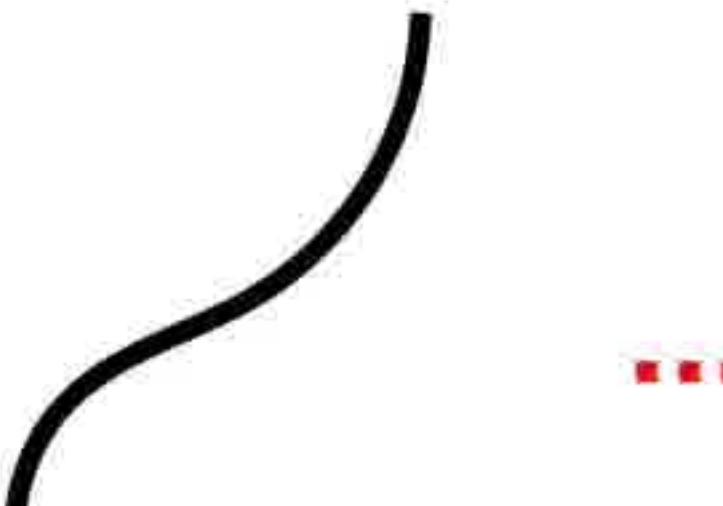
Then we rotate so that
PC2 is vertical (this just
look a

Then we rotate so that
PC2 is vertical (this just
look a

Since these projects correspond to Sa



This is where Sam
in our new PCA

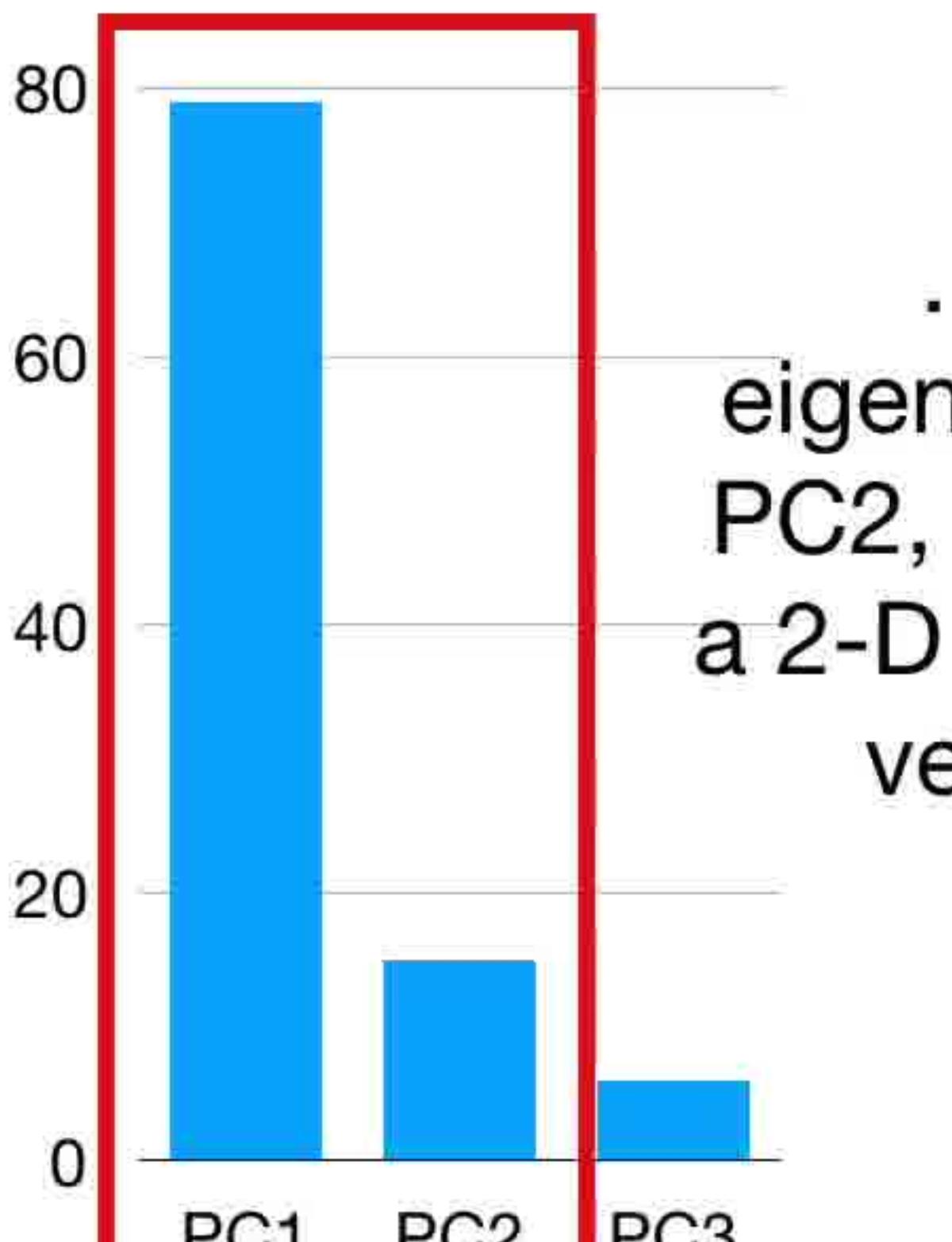


...

To review, we started with an awkward graph that was kind of hard to read.

...then we calculate
principal components





•
eigen
PC2,
a 2-D
ve

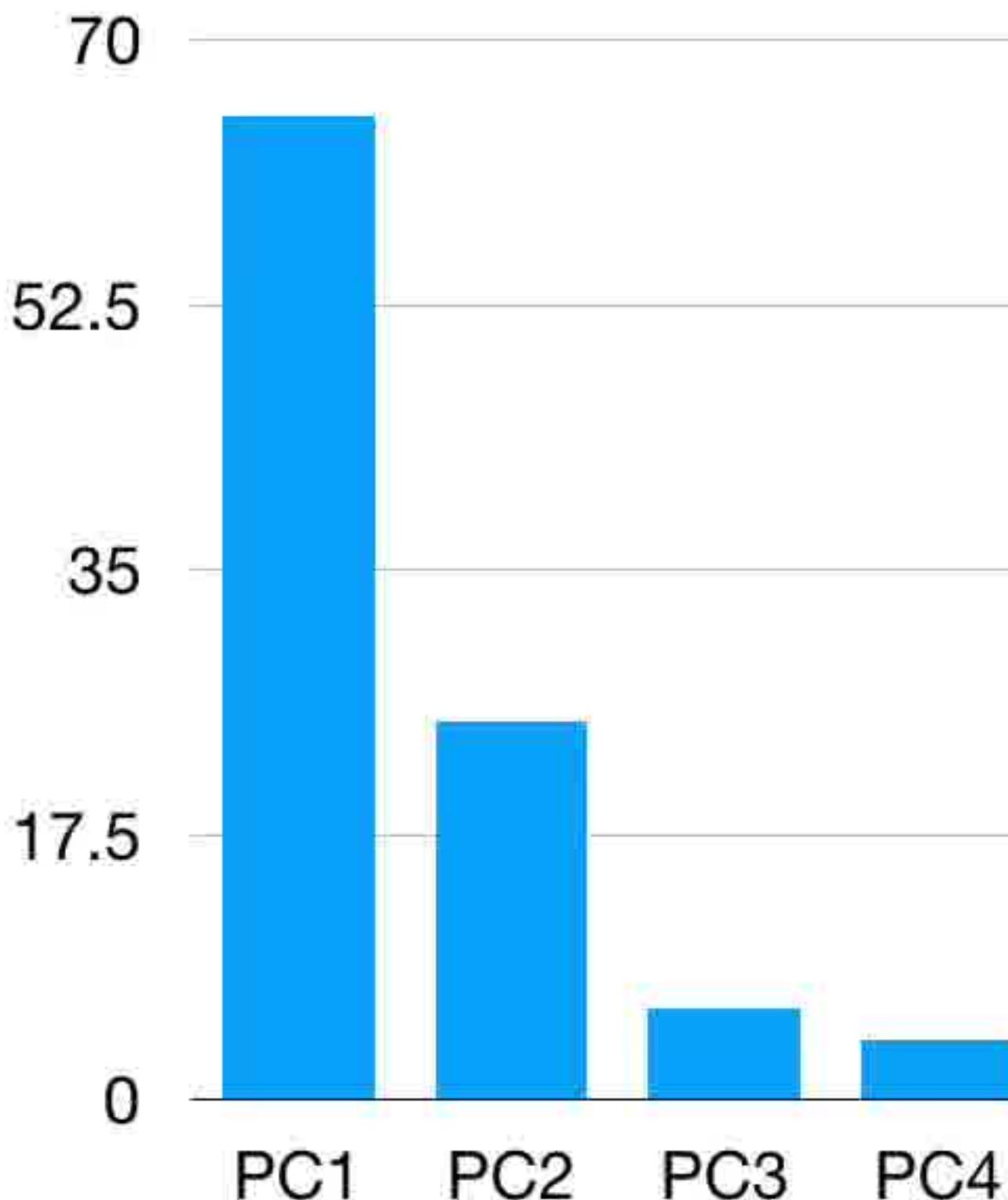
...lastly,
and PC2

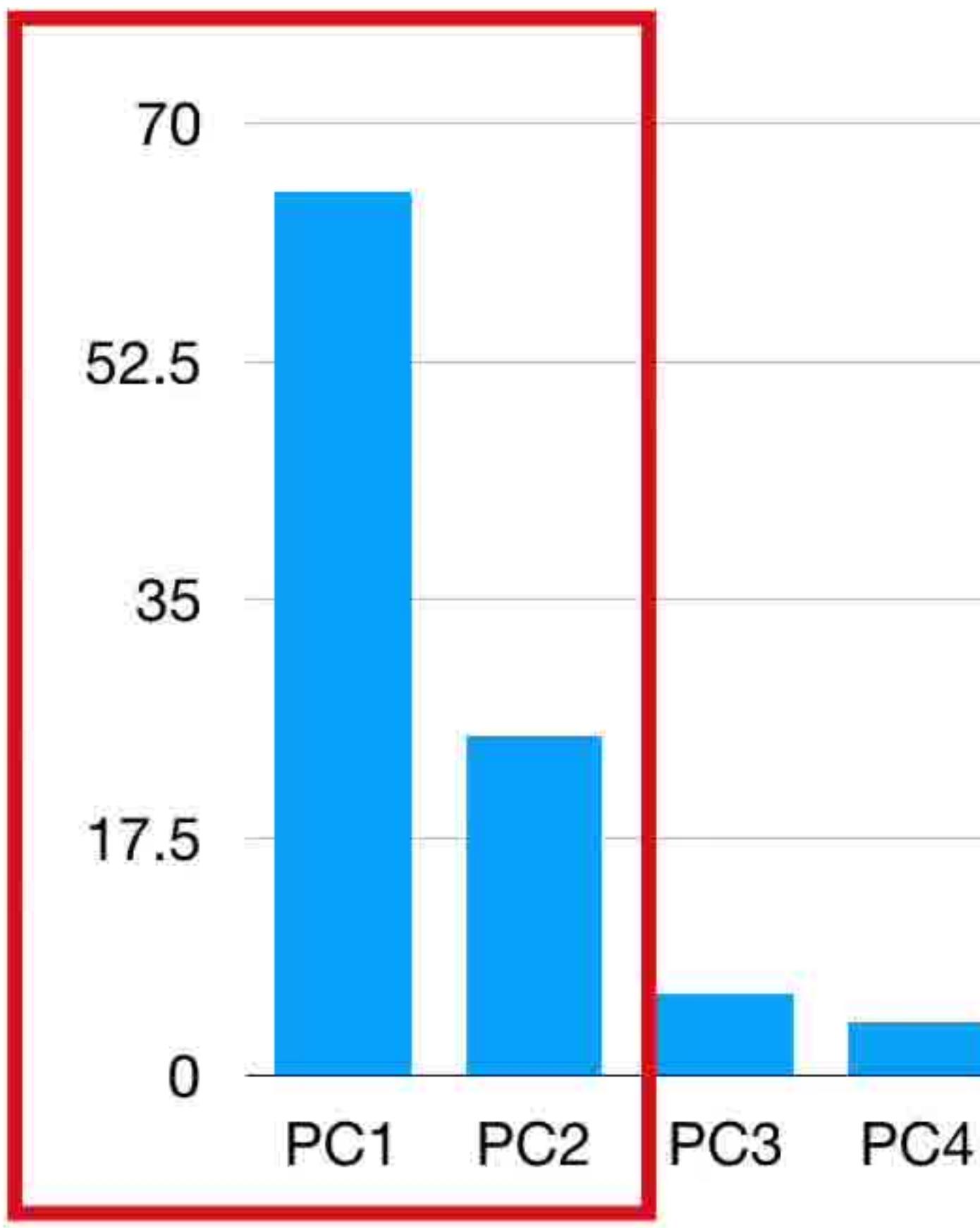
Dimensions

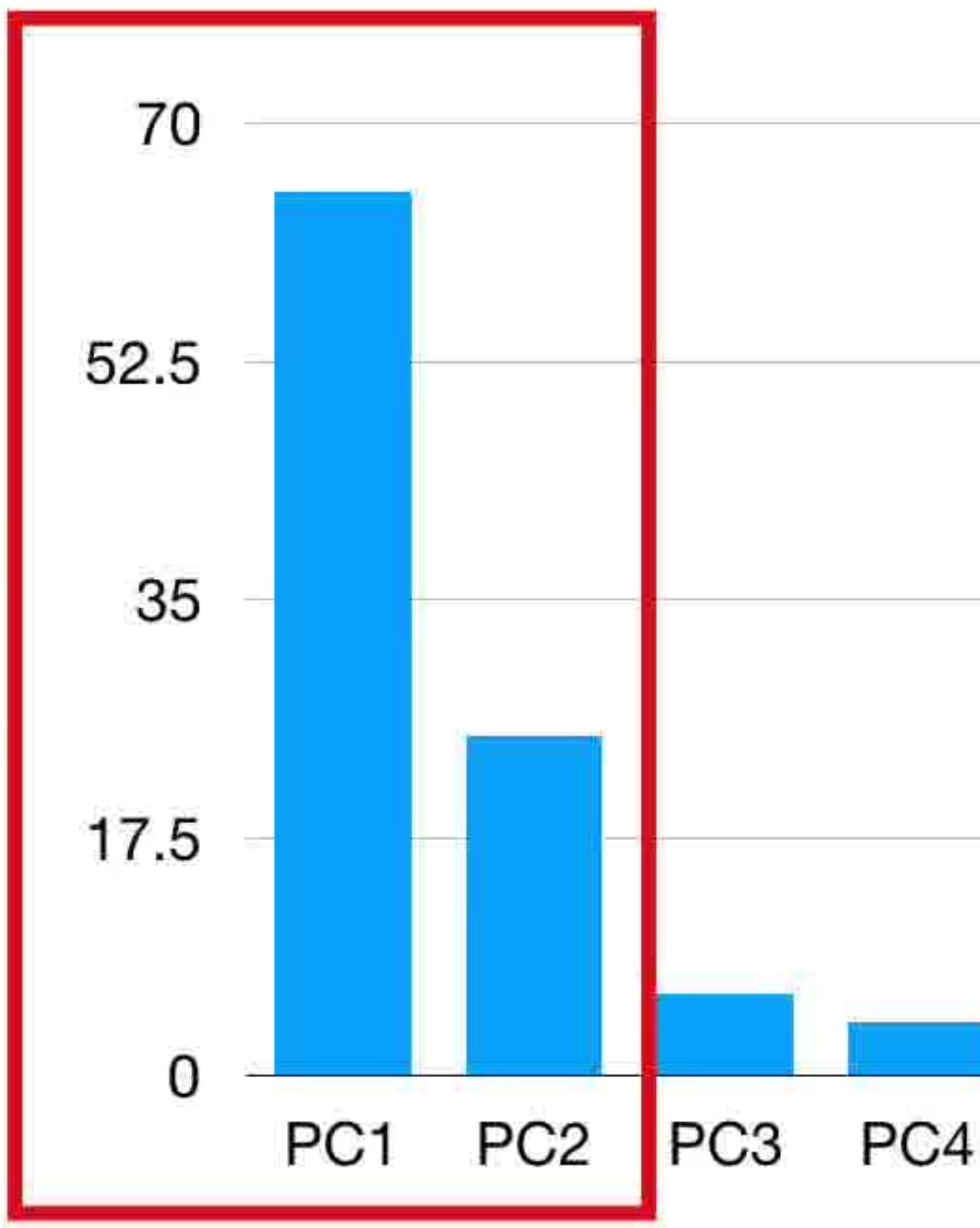
the

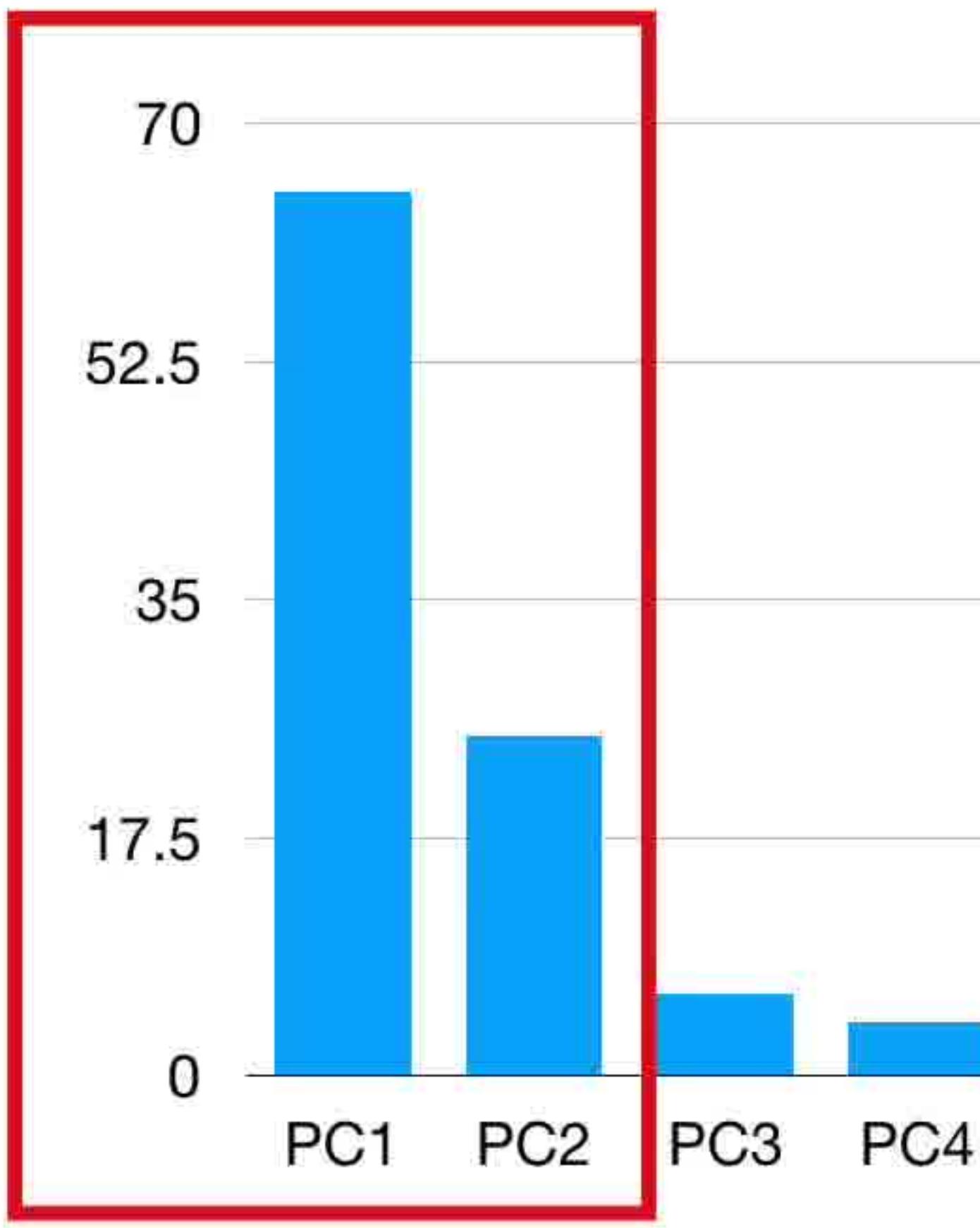


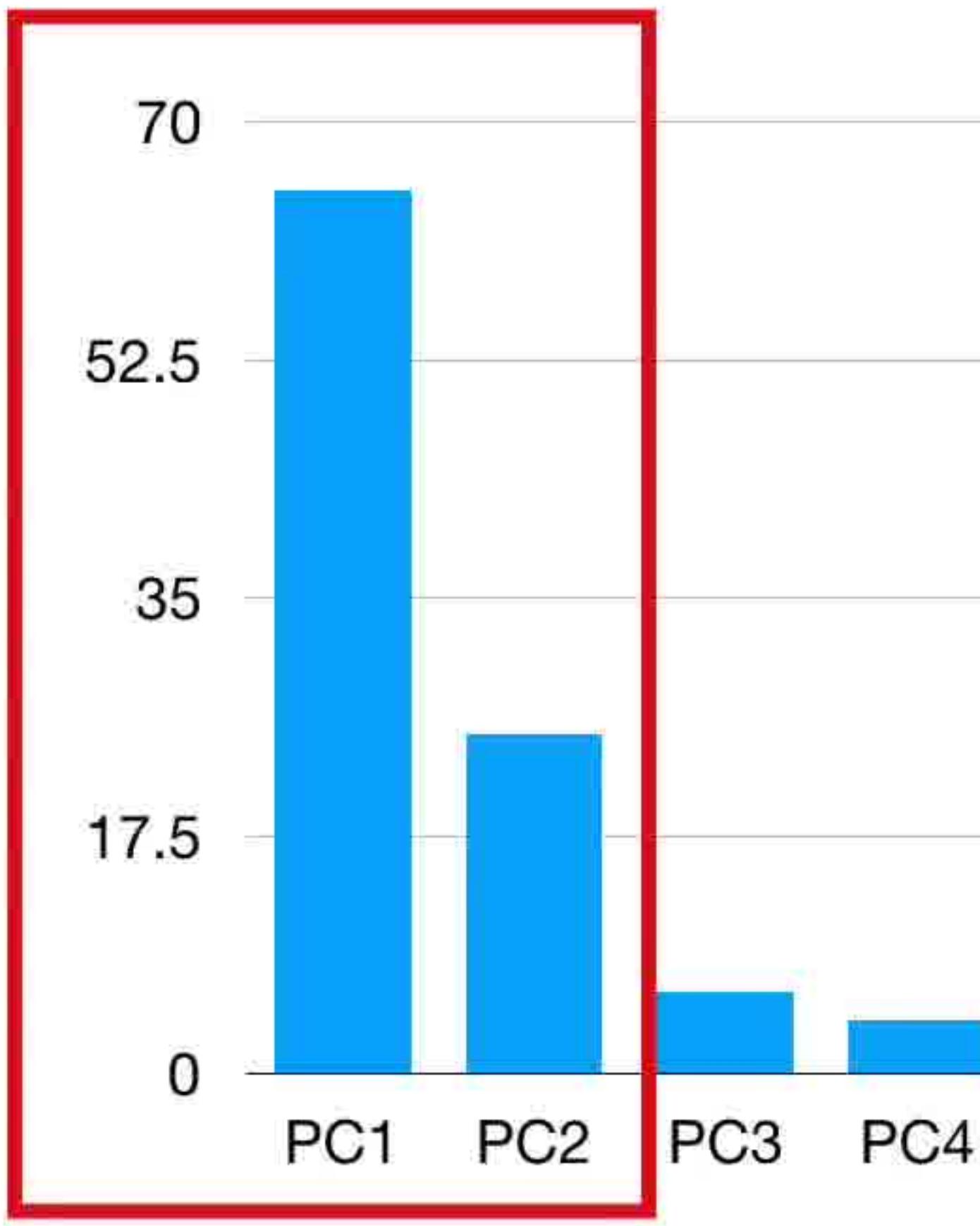
	Mouse 1	Mouse 2	Mouse 3	Mouse 4
Gene 1	10	11	8	3
Gene 2	6	4	5	3
Gene 3	12	9	10	2.5
Gene 4	5	20	6	2

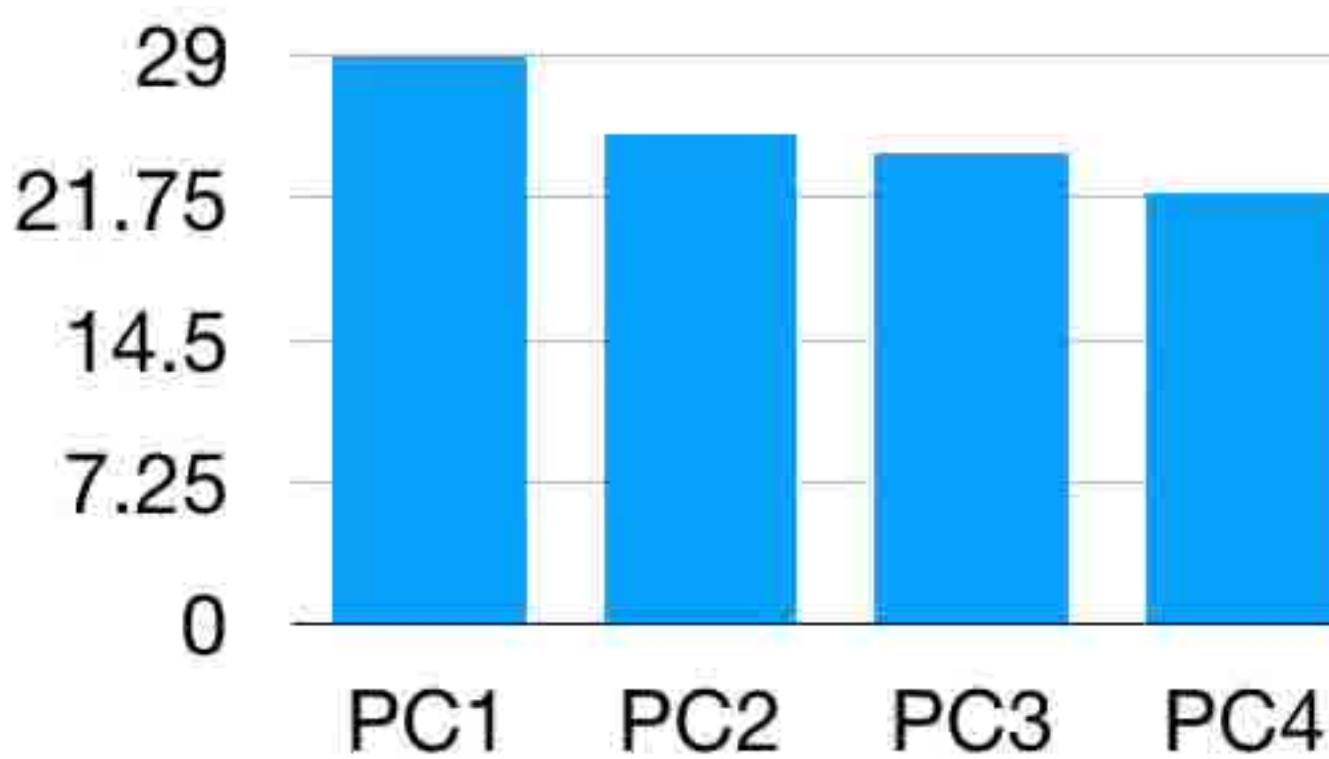




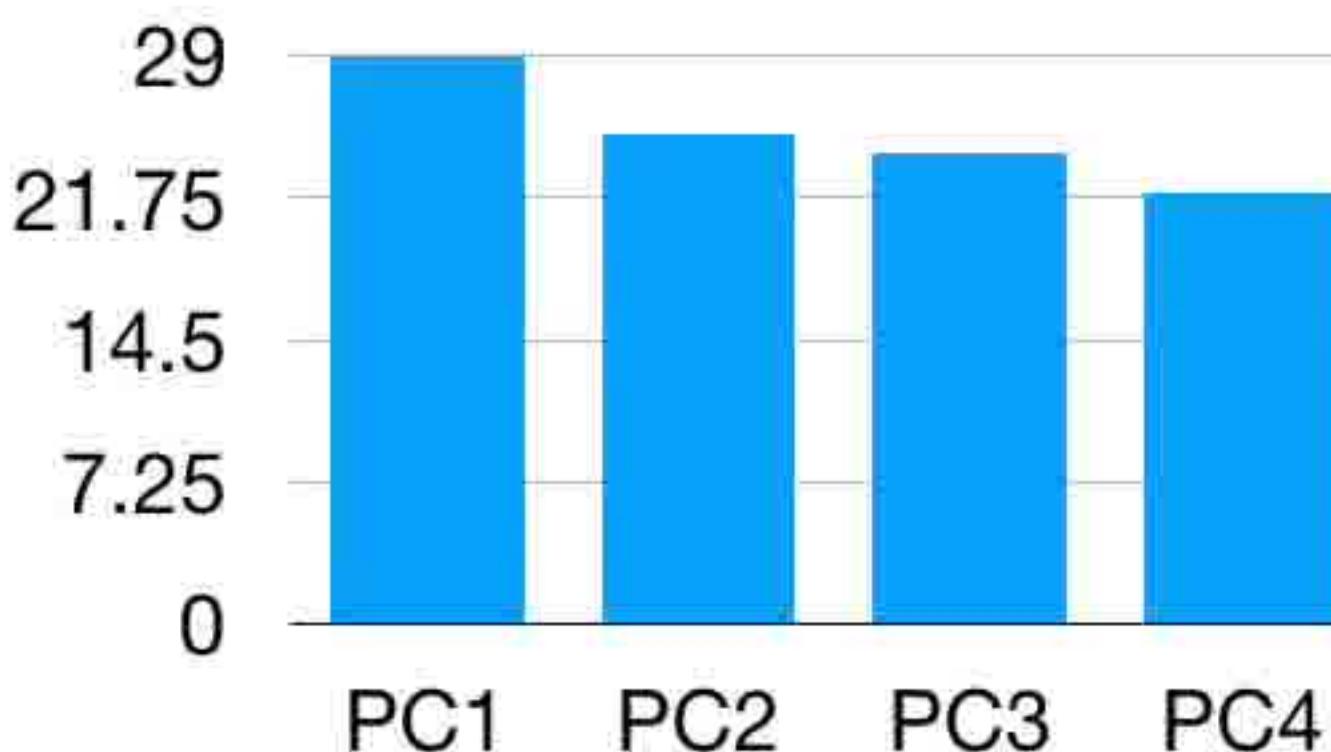




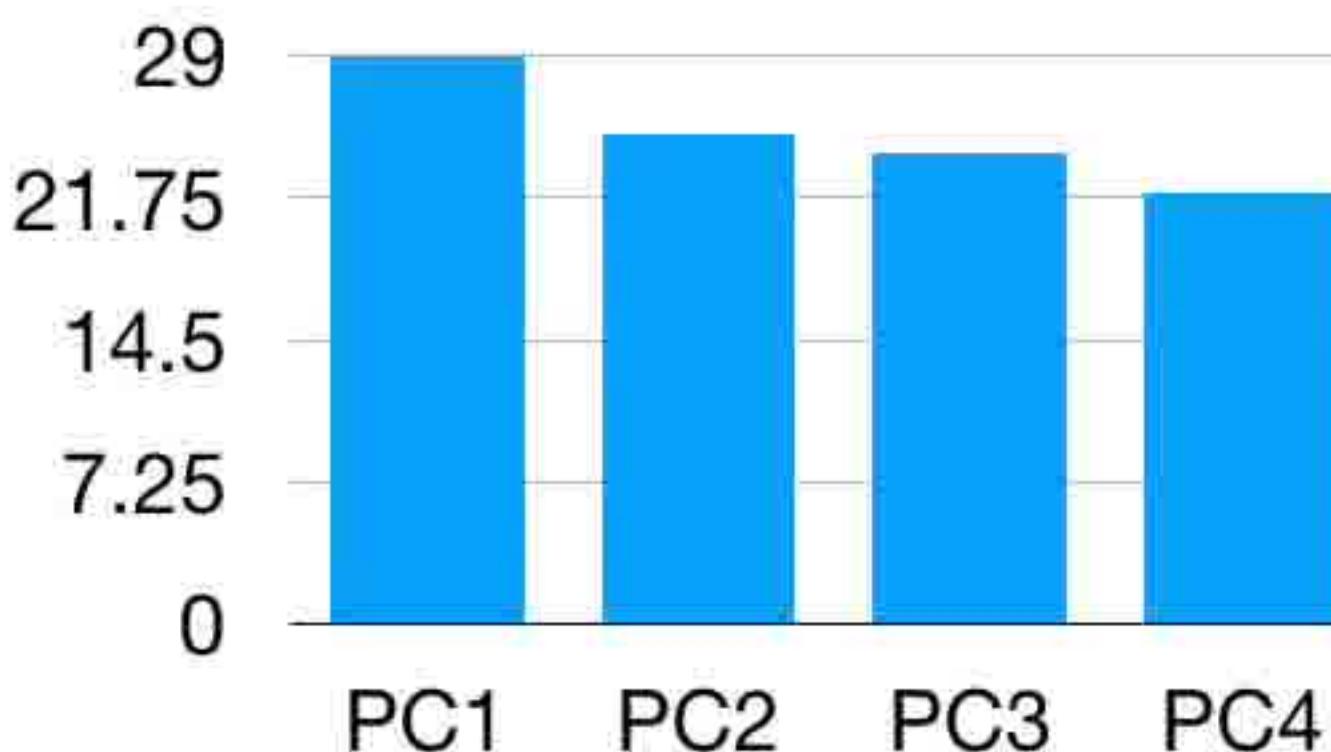




However, even a noisy
can be used to identify



These samples
each other the



Th