Let's talk about the PCA example.

Now, the packages we'll be using for

this are actually the ones we are already familiar with plus one more which is purrr,

which helps simplify certain operations.

I will explain this when it comes up.

The data we'll be dealing with is the crime rate in US States.

This is a pretty classic dataset.

These are rates of certain crimes in US States with a two-letter abbreviation,

which you can look up here if you're interested.

But basically you have a number of crime categories, murder,

rape, robbery, assault, burglary,

larceny, auto, meaning grand theft auto.

Then we have the states from Alabama to Alaska,

Arizona, Arkansas, California, Colorado,

Connecticut, and so on.

We'll do as we take this string,

this is a bit of a shortcut.

We take this long,

this is actually a long string of text.

Then we'll take that string,

we'll pipe it to pass it to text connection,

which will basically turn it into a fake file,

which can then be read using the read table function with the header equals true.

Then that gets assigned to head type of arrow,

this is to the crime variable.

Now again, what this does is basically it says it takes whatever's on

the left-hand side and plugs it into whatever's on the right-hand side.

Here's what the dataset looks like.

Well, let's do a quick exploratory data analysis.

Here are the summary statistics.

Here's the plot of different types of crime.

Note that they are generally positively correlated.

Also some of them like robbery seems to have a big outlier here and so

does grand theft auto possibly but and in fact,

these are two different outliers.

But otherwise they seem to follow a pretty centred normalish pattern.

Now the principal component analysis in R is pretty straightforward.

We just use the prcomp function and we give it a crime dataset and we tell it to scale

the variables just to put them on the same scale.

And it immediately gives us the standard deviation of each component,

which not ideal when we're scaling,

but also give us the rotation of this.

In fact, we can immediately see something interesting here,

which is that in the first principal component,

all values are negative.

In the second component,

some values are negative,

in particular, murder, rape, and assault,

but not robbery which is positive although it's only by a small amount and burglary,

larceny, and auto are pretty positive and

we can ask why that's the case and what does this represent.

Let's talk about selecting the number of principal components.

The way we can do that is we can look

at the proportion of variance explained by each of those components.

And in this case, we would see that we can look at

the cumulative proportion and the standard deviation of each component.

So we can be able to screeplot to say,

"Is there a rapid drop-off here?"

Another one we could do, though we see one,

the first component is much bigger than the others,

but the second is not much bigger than a third.

We can also represent them like this,

in terms of the variance,

reason because they're scaled.

Everything less above one here goes in according to the Kaiser's rule,

but everything below one doesn't.

The other is of course the cumulative variance explained,

which we can also see here.

This plot has been involved but here it is.

So first thing we'll do is we'll plot the proportions of the variance explained.

That is this component,

and we'll do that by setting up a plot.

On its horizontal axis we'll have a seq_along,

basically that says one through the length of the pcvars so

we get these one values one through seven.

On the vertical axis,

we have the variance explained here,

the eigenvalues divided by their total to get the proportion of

variance explained by each variable.

Then type "o" means this overplot where you have both the dashed line and a point and the rest is x label k,

number of components, y lab,

there's proportion of variance explained.

Now the other thing we're going to plot with

the same horizontal axis and now we're going to plot the cumulative variance

explained or proportion of variance explained.

We're going to use a different line type for

the cumulative variance explained here, dash lines.

Then we're going to print a legend on the right-hand side because that's where we have

blank space with the two line types

and caption it with the type of variance of explaining.

And then finally, we're going to plot just a line at

point nine and a line at a one over the total number of variables.

So the first is one I'm highlighting is for the Kaiser's rule,

this one is for the cumulative variance rule.

So we get this nice plot where we see for example

that according to the 90 percent variance explained rule,

we crossed the point nine barrier at the fourth component,

and for the Kaiser's rule,

we crossed it after the second one,

which we could also use this.

This is the minimal value for which it is true that

the variance explained exceeds point nine.

This one is the greatest value for which

the proportion of Variance Explained is relative to others is greater than one.

Now we can figure out what these components actually mean.

Now one useful tool for that,

and the way we looked at in

a little bit over here when I pointed out the signs on these variables.

But we can also have a useful tool called a biplot.

Here's what it looks like.

Now what we see here is each state is plotted on this for

its principal component value

So with rotation into the first and second principal component,

and we also have the variables which are plotted

according to which component do they represent or their component coordinates.

One thing to keep in mind is the principal component analysis,

if you multiply a particular component by minus

one or coefficients for a component by minus one,

it's not actually going to change the variance explained or anything else.

Highest crime states is actually going to end up with the lowest values of components.

Now, what do we see here?

Well, in the first principal component,

all of the variables are pointing in the same direction.

In fact here what we see is that the negative values of the components

correspond to the least crime,

so that means that the higher the value of the first principal component,

the lower the crime in the state.

This is the overall level of crime.

We could also look at the second principal component and here they fan out a little bit.

Some are positive, some are negative.

But notice that this forms a spectrum.

Murder is the ultimate crime against a person than assault and rape.

Robbery is something in between somebody who is the mode of his money although it does involve actually interacting with the victim.

Whereas grand theft auto,

that's stealing a car when nobody's paying attention,

larceny is just theft,

burglary also breaking in, but hopeful.

But we know for burglary ideally they're not going

to run into the person they're burglarising.

These are all crimes against property and these are all crimes against persons.

There's a spectrum, some states seem to

have high tendency towards criminality versus low tendency towards criminality.

But there's also a spectrum of some states

are really into crime against a person whereas others are against property.

In fact, some observations that are more specific to US geography,

Nevada is most known for Las Vegas,

which is the city which has a saying,

"What happens in Vegas stays in Vegas."

But it constitutes the Las Vegas crime statistics as gambling and sorts of things.

The states with the strongest crimes against a person values,

that is states, they're down here.

These tend to be,

the what's called the American southeast or the Deep South.

Whereas states with strongest crimes against

property tendency tend to be the northeast ones.

Deep South would be here, Alabama,

Louisiana, South Carolina, Mississippi,

and a bit to somewhat lesser extent,

New Mexico, Texas,

Missouri, Oklahoma, and Virginia, Kentucky, Arkansas.

Now, on the other hand,

the more crimes against property state,

the ultimate one is Massachusetts,

then followed by Rhode Island,

Connecticut, Delaware, New Jersey.

So all of these are New England or nearby states.

Hawaii is another very much property crime state,

that Hawaii is the state that's that island far into Pacific,

that is a US State.

We could also do other things.

For example, we could sort the states

by their component and see which states have the highest tendency towards criminality and

which states have the highest or lowest tendency towards property crime.

We can also make a combined plot of say,

the third component, although it is hard to tell exactly what it represents.

There's the auto robbery, murder,

and assault, larceny, rape, and burglary.

Again, not really clear what it actually means.

This concludes the demonstration of principal component analysis.

Try the challenge next.