**Principal Component Analysis (PCA) in R**

Aim of PCA

* To get a sense of the trends present within the data
* This can enable us to identify groups of samples that are similar and work out which variables make one group different from another.

Basics of PCA

* In a dataset with many variables, simplify that dataset by turning the original variables into a smaller number of "Principal Components"

Principal Components

* The underlying structure in the data
* The directions where there is the most variance
* The directions where the data is most spread out
* A linear combination with the original set of variables (Linear Algebra)
* Related to eigenvectors and eigenvalues

Example PCA

* mtcars dataset (built into R)
* 32 models of car, taken from an American motoring magazine (1974 Motor Trend magazine)
* For each car, there are 11 features, expressed in varying units (US units)

\*mpg: Fuel consumption (Miles per (US) gallon): more powerful and heavier cars tend to consume more fuel.

\*cyl: Number of cylinders: more powerful cars often have more cylinders

\*disp: Displacement (cu.in.): the combined volume of the engine's cylinders

\*hp: Gross horsepower: this is a measure of the power generated by the car

\*drat: Rear axle ratio: this describes how a turn of the drive shaft corresponds to a turn of the wheels. Higher values will decrease fuel efficiency.

\*wt: Weight (1000 lbs): pretty self-explanatory!

\*qsec: 1/4 mile time: the cars speed and acceleration

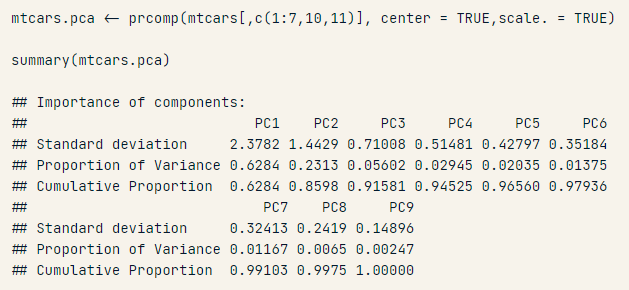
\*vs: Engine block: this denotes whether the vehicle's engine is shaped like a "V", or is a more common straight shape.

\*am: Transmission: this denotes whether the car's transmission is automatic (0) or manual (1).

\*gear: Number of forward gears: sports cars tend to have more gears.

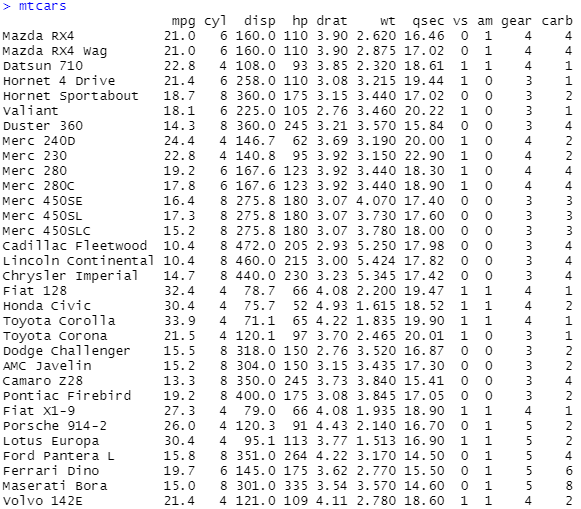
\*carb: Number of carburetors: associated with more powerful engines

* Exclude non-numeric data (vs, am)
* prcomp() function, center=TRUE, scale=TRUE
* ‘we are performing principle components analysis’
* center=TRUE: shift the data so that the average of the data becomes 0
* scale=TRUE: scale the data so that standard deviation and variance becomes 1
* standard way of preparing data before analysis, although we could do it differently
* summary()



* Then there are 9 principal components, PC1~PC9
* Each of these explains a percentage of the total variation in the dataset
* PC1: 63% of the total variance
* two-thirds of the information in the dataset (9 variables) can be encapsulated by PC1
* PC2 explains 23% of the variance
* So, by knowing the position of a sample in relation to just PC1 and PC2, you can get a very accurate view on where it stands in relation to other samples, as just PC1 and PC2 can explain 86% of the variance.
* For these reasons, it is good to choose PC1 and PC2 in our further analyses, since these two combined represents the data better than the other Principal Components.

mtcars ### so this is the data we’re going to use (before exclusion of vs and am)



library(devtools) ### already installed in computer (comes pre-installed)

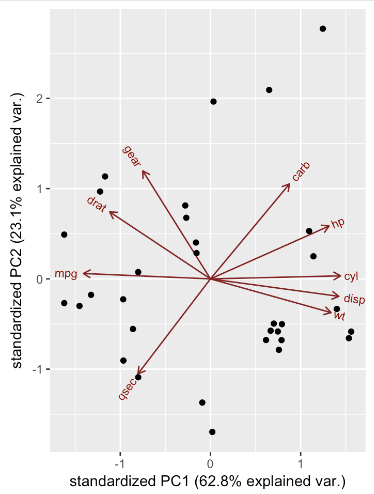
install\_github("vqv/ggbiplot")

### from github repository vqv, install ggbiplot, a package that creates biplots

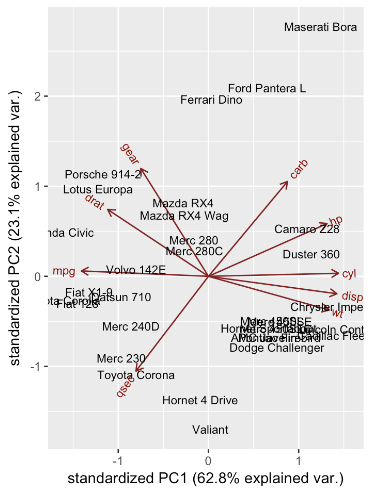
### biplot is a two-variable scatterplot

library(ggbiplot) ### declare we use ggbiplot

ggbiplot(mtcars.pca) ### makes it easy to see the variance in the data (how similar/different)



ggbiplot(mtcars.pca, labels=rownames(mtcars)) ### denote what the ‘points’ mean



mtcars.country <- c(rep("Japan", 3), rep("US",4), rep("Europe", 7),rep("US",3), "Europe", rep("Japan", 3), rep("US",4), rep("Europe", 3), "US", rep("Europe", 3))

### in the dataset, there are, in order, 3 Japanese cars, 4 US cars, 7 European cars, 3 US cars, 1 European car, 3 Japanese cars, 4 US cars, 3 European cars, 1 US car, and 3 European cars

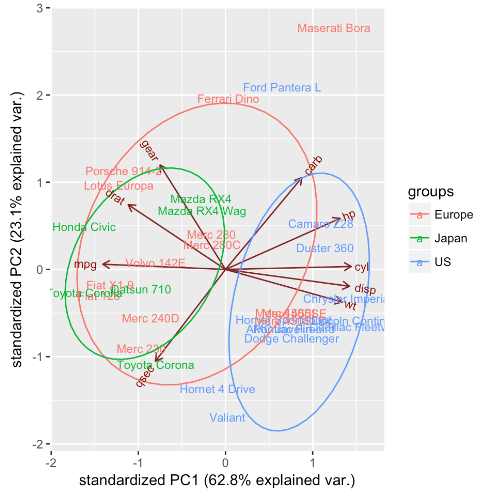
### rep stands for replicate—replicate elements of vectors and lists

### here, one long vector goes into mtcars.country (ordered group of strings)

ggbiplot(mtcars.pca,ellipse=TRUE, labels=rownames(mtcars), groups=mtcars.country)

### make ellipses that show where each group is (Europe, Japan, US)

### show similarities and differences between groups, and the role of each variable



**References & Further reading**

Datacamp - Principal Component Analysis in R Tutorial

* <https://www.datacamp.com/tutorial/pca-analysis-r>

R Documentation - prcomp: Principal Components Analysis

* <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp>

R Documentation - rep: Replicate Elements of Vectors and Lists

* <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/rep>

If you are interested in the math behind it

* <https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643>
* <http://rstudio-pubs-static.s3.amazonaws.com/170696_a27b54d301c3412fa70e9236fc6a8762.html>
* <https://www.statology.org/principal-components-analysis-in-r/>
* <https://uc-r.github.io/pca>

**Exercise**

Use UsArrests dataset instead of mtcars (built into R)

[5 US Regions Map and Facts | Mappr](https://www.mappr.co/political-maps/us-regions-map/)