Principle Component Analysis

1. The first step is to make sure you data is your working directory (otherwise you will end up having to write out the entire file pathway) and is saved as a comma delimited file (CSV) file.

2. Assign your data to a variable (called data below) in “R”, and let it know that the first row is a header.
   data<-read.csv(“filename.csv”, header=T)

3. PCA can only work on numeric values so look at the data columns to determine if there are any categories that should be excluded
   names(data)

4. Notice that the first two categories (“Species” and Sex”, probably contain data that is not numeric. Assign a variable that excludes these groups.
   x<-data[,10:38]. This tells R that x contains rows 10-38 of the data-frame ”data”

5. Time for PCA! It is an easy “R” formula
   model<-prcomp(x, scale=T) runs a PCA on “x” and scales it to a uniform size.

6. To get a breakdown of you PCA use **summary(model).** In this case there are 29 components (because there were 29 variable in ”x”) in this case the first component accounts for almost 28 % of all variation.

7. To view the distribution of eigen values use **plot(model)**. This is known as a “scree plot”

8. To view a distribution of PC 1 plotted against PC 2 use **plot(model$x[c(1,2)])**. The numbers in the parentheses designate which components to plot. This same formula can be used to plot any two components against each other. Alternatively you can use **plot(model$x[,1],model$x[,2])**. It shows the same thing but the axis of the graph are labeled differently.
9. The graph from step 8 shows at least 3 groups and some outliers. It would be helpful to visualize how the groups of interest are distributed across the two principle components.

10. To do this** go back to the data table and create a column in which you designate a color to the groupings you would like (already done) and then set this column as a character **color<-as.character(data$COLOR).** Then, re-plot your graph using
    plot(model$x[,c(1:2)],col=color, pch=16)
    here:

    IC=red, IG=Blue, IT=orange, IGL=light blue, IGO=green, IGP=yellow, UNK=Black

    The letters are shorthand for Genus species subspecies. Some patterns are readily apparent. IGL seems to be composed of two distinct groups, one nested within IGG and one on its own.,You can now make an informed guess as to the identity of the three unknowns, one of the ICP appears to be morphologically more similar to IC etc…

11. To see which specimens correspond to which datapoints use the command **text(model$x[,1],model$x[,2],data$Specimen,cex=0.7,pos=4,col="black")** and zoom in.

12. A couple more potentially useful commands:
    model$x returns the data scores on each principle component.
    model$rotation will return the loadings (correlation coefficients) of each variable on each principle component.
    model$x[,1] will return the loading of each individual on the first principle component.
    (model$sdev)^2 will return the variance for each PC

*to see a full list of colors available in R type “colors()”

** I am convinced that there is an easier/better way