Executing PCA and MANOVAs

Module 7 & 8

Victor Valdez

Getting started

- RStudio
  - Makes R easier to use
  - Start by keeping your work organized by:

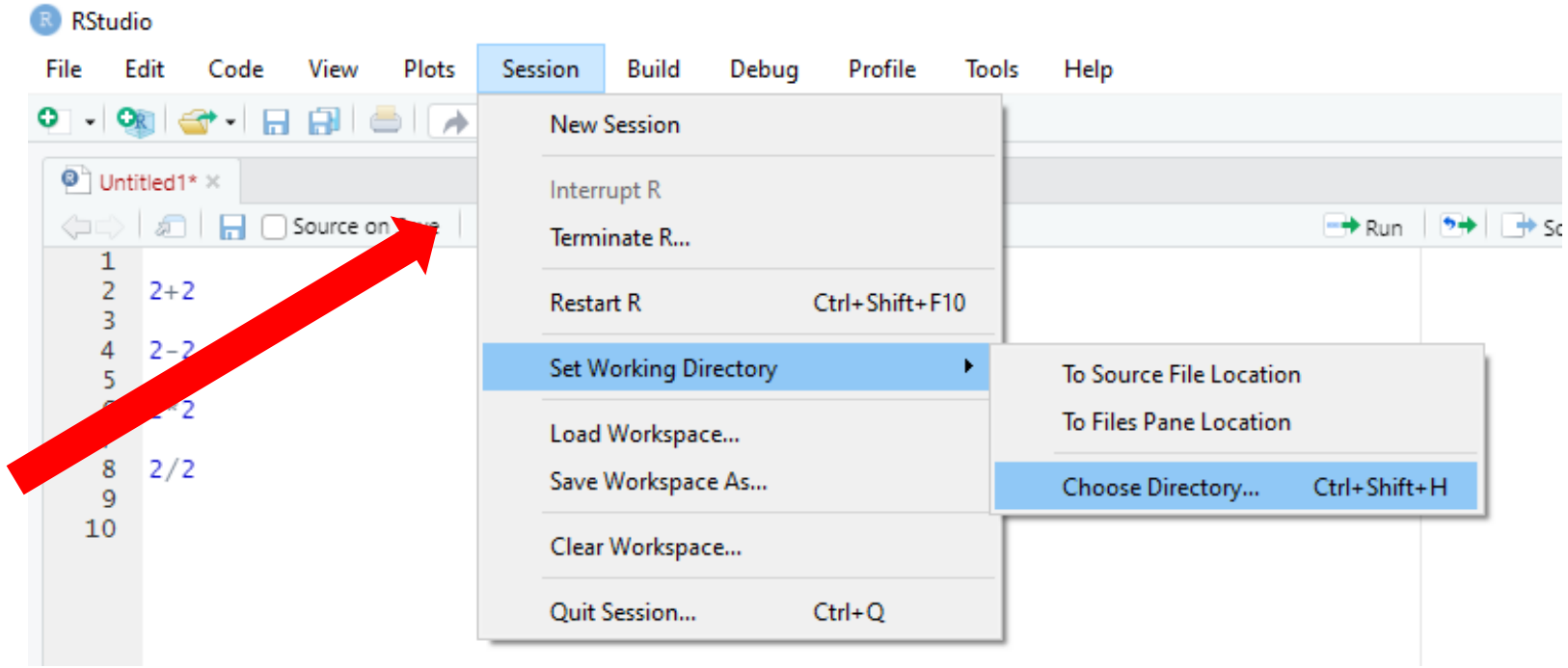
    Set your *<u>"Working Directory"</u>*

    OR

    Ctrl+shift+H

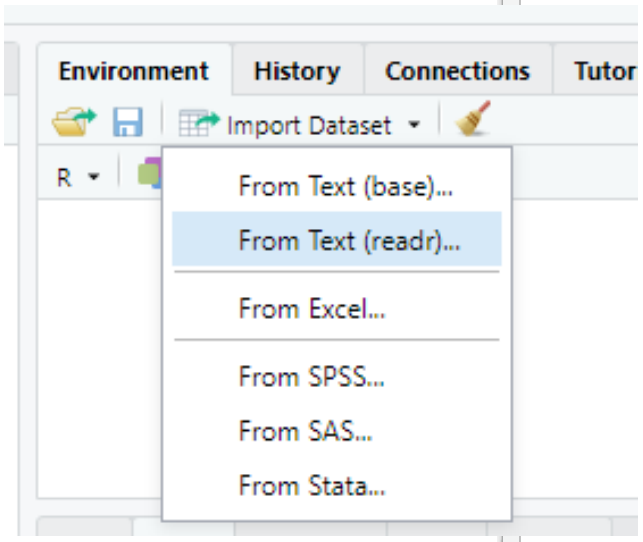
    OR

    setwd("C:/Users/Home/Downloads/Workshop_data")

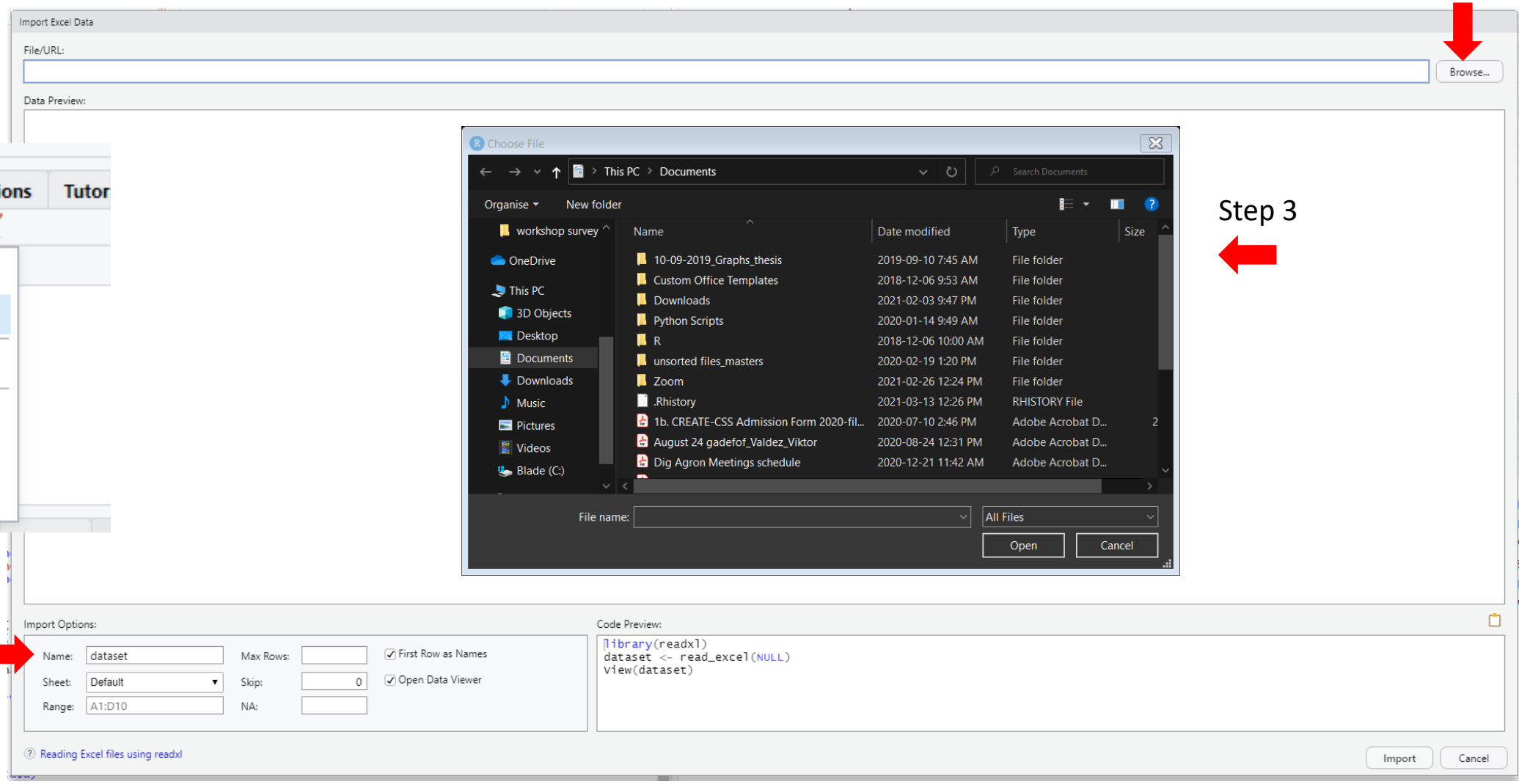


Import data in ***text format*** to R

Step 1



Step 2



Step 3

Step 4

- You can rename your data here or use another function

- You should now see the following; ***change*** delimiter to <u>whitespace</u>; ***click*** import

Import Text Data

File/URL:

C:/Users/Victor/Downloads/PCA analysis/BumpusSparrowsJanzenStern.txt

Browse...

Data Preview:

sex (double)	Fit (double)	len (double)	ext (double)	wt (double)	head (double)	humer (double)	femur (double)	tibio (double)	skull (double)	stern (double)
1	1	154	241	24.5	31.2	0.687	0.668	1.022	0.587	0.830
1	1	160	252	26.9	30.8	0.736	0.709	1.180	0.602	0.841
1	1	155	243	26.9	30.6	0.733	0.704	1.151	0.602	0.846
1	1	154	245	24.3	31.7	0.741	0.688	1.146	0.584	0.839
1	1	156	247	24.1	31.5	0.715	0.706	1.129	0.575	0.821
1	1	161	253	26.5	31.8	0.780	0.743	1.144	0.607	0.893
1	1	157	251	24.6	31.1	0.741	0.736	1.153	0.610	0.862
1	1	159	247	24.2	31.4	0.728	0.718	1.126	0.609	0.793
1	1	158	247	23.6	29.8	0.703	0.673	1.079	0.602	0.820
1	1	158	252	26.2	32.0	0.749	0.739	1.153	0.614	0.857
1	1	160	252	26.2	32.0	0.741	0.723	1.129	0.624	0.892
1	1	162	253	24.8	32.3	0.766	0.752	1.134	0.633	0.923
1	1	161	243	25.4	31.8	0.721	0.722	1.126	0.597	0.891
1	1	160	250	23.7	29.8	0.730	0.703	1.103	0.590	0.820
1	1	159	247	25.7	31.4	0.729	0.717	1.141	0.592	0.927
1	1	158	253	25.7	31.9	0.743	0.699	1.150	0.600	0.860
1	1	159	247	26.5	31.6	0.733	0.714	1.155	0.611	0.923
1	1	166	253	26.7	32.5	0.767	0.765	1.230	0.600	0.878
1	1	159	247	23.9	31.4	0.752	0.723	1.113	0.602	0.825
1	1	160	248	24.7	31.3	0.752	0.737	1.176	0.603	0.803
1	1	161	252	28.0	31.8	0.770	0.731	1.190	0.590	0.885

Previewing first 50 entries.

- Our excel sheet with columns and data

Import Options:

Name: BumpusSparrowsJanzenS

Skip: 0

First Row as Names

Trim Spaces

Open Data Viewer

Delimiter: Whitespace

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

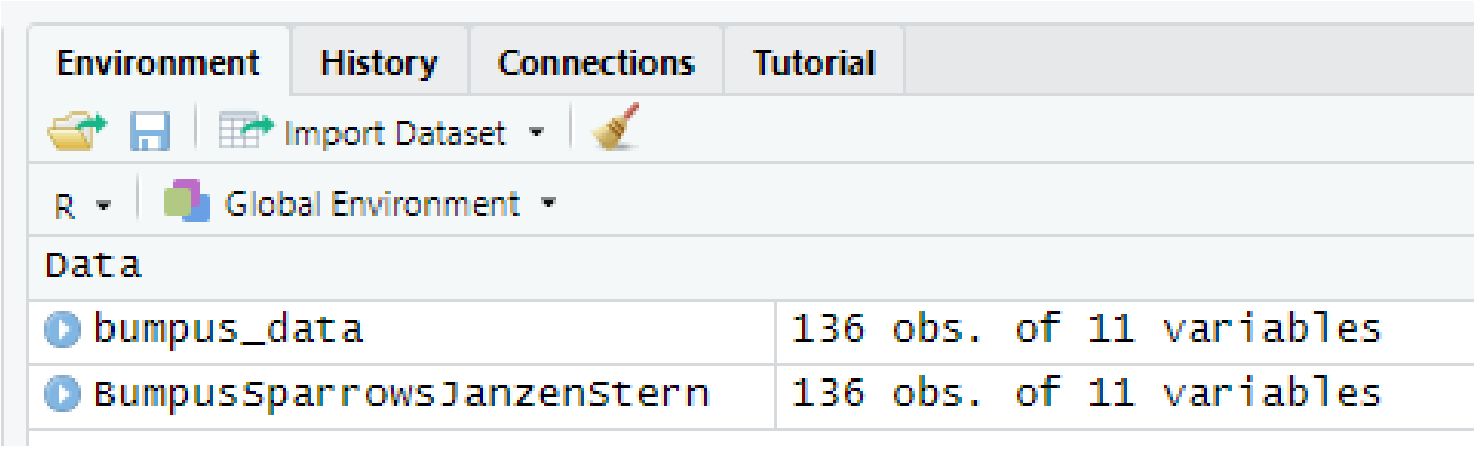
NA: Default

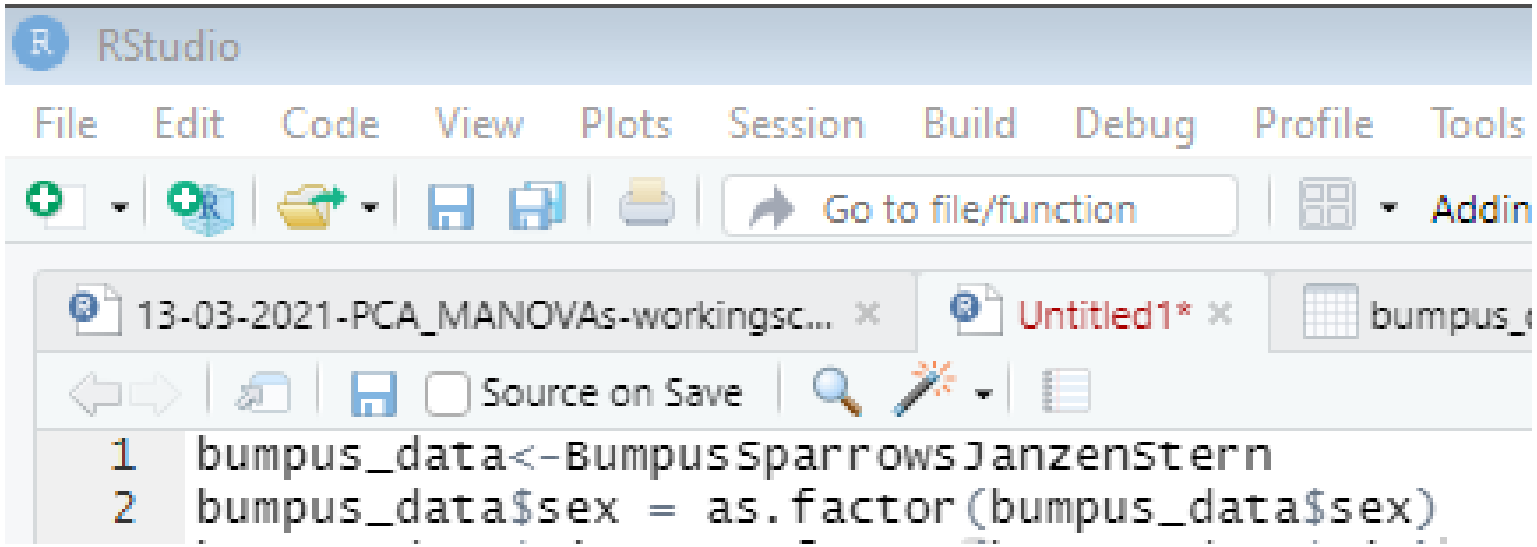
Code Preview:

```
library(readr)
BumpusSparrowsJanzenStern <- read_table2("C:/Users/Victor/Downloads/PCA analysis/BumpusSparrowsJanzenStern.txt")
View(BumpusSparrowsJanzenStern)
```

- We now see the contents of the file name
- The file name may be too long; not ideal for coding
- Let's rename:

bumpus_data<-BumpusSparrowsJanzenStern





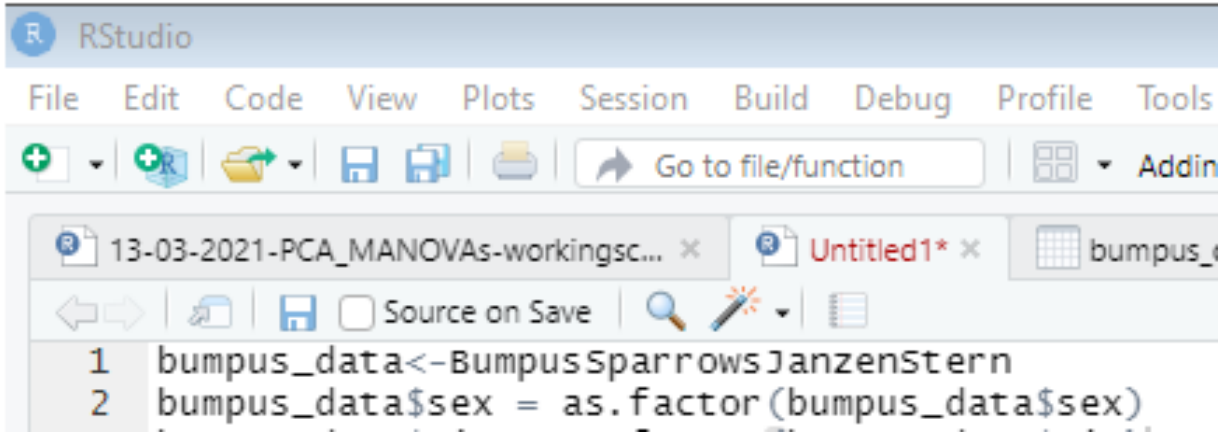
- We can now use the autofill feature by using **$ before the dataset**

  e.g., data$....

Run Descriptive Statistics Function

Change sex and fit to factorial:

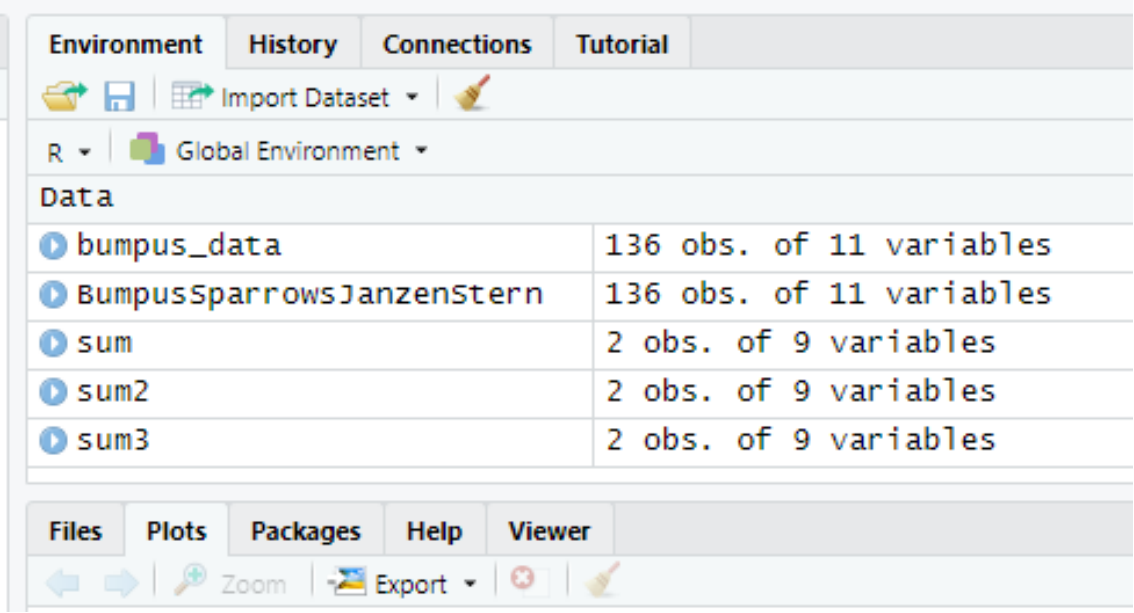
```
bumpus_data$sex = as.factor(bumpus_data$sex)
bumpus_data$Fit = as.factor(bumpus_data$Fit)
```



Now use the summary function in FSA package:

```
sum=Summarize(len ~ sex, data = bumpus_data)
sum2=Summarize(wt ~ sex, data = bumpus_data)
sum3=Summarize(stern ~ sex, data = bumpus_data)
```

Let's look at the tables generated

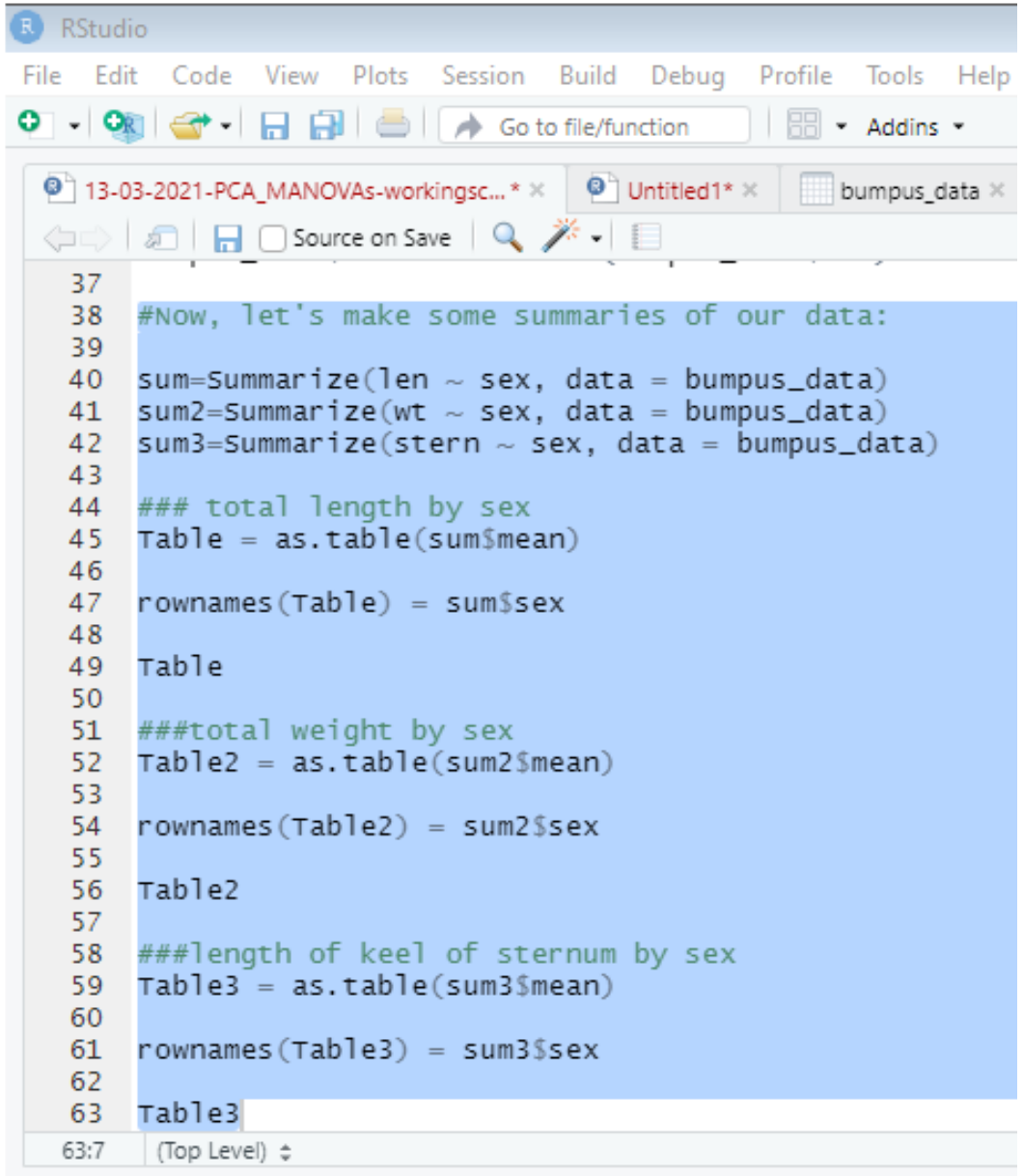


Plotting data: extract data

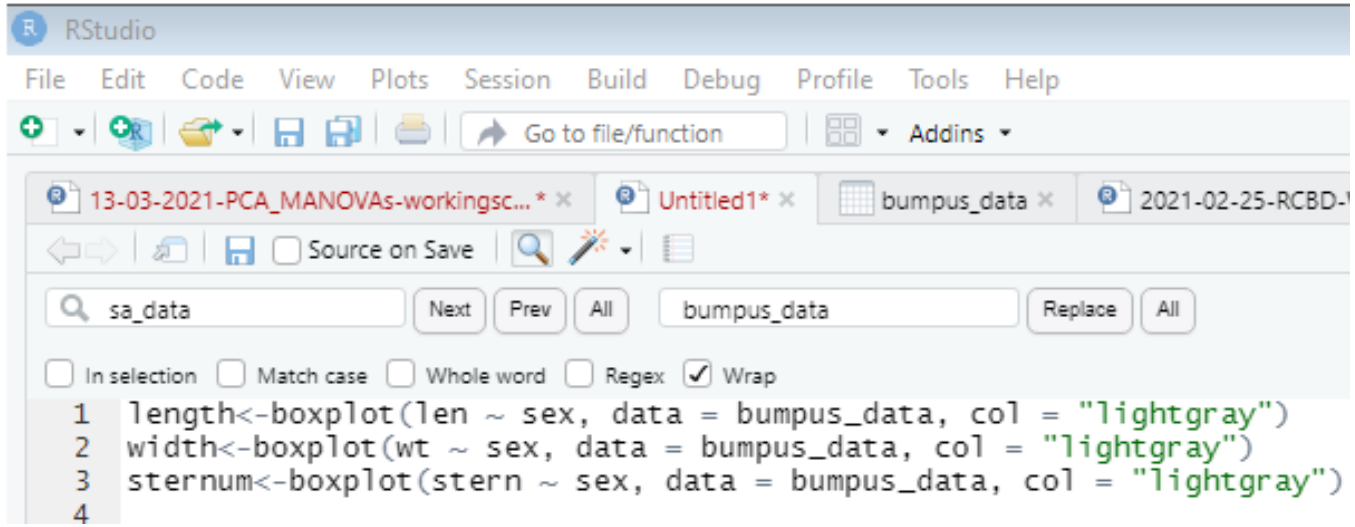
- Data is best plotted, let's make some boxplots

- First extract the mean from the table:

```
Table = as.table(sum$mean)

rownames(Table) = sum$sex

Table

Table2 = as.table(sum2$mean)

rownames(Table2) = sum2$sex

Table2

Table3 = as.table(sum3$mean)

rownames(Table3) = sum3$sex

Table3
```



Plotting data



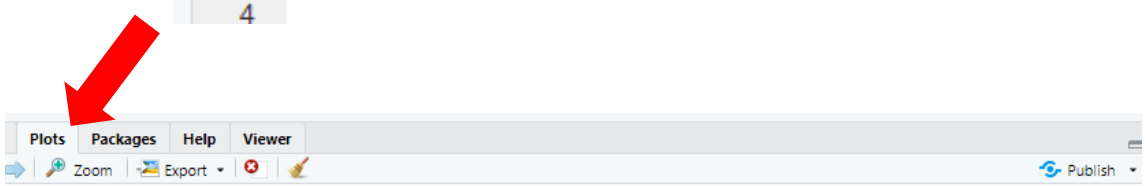
- Plot the boxplot :

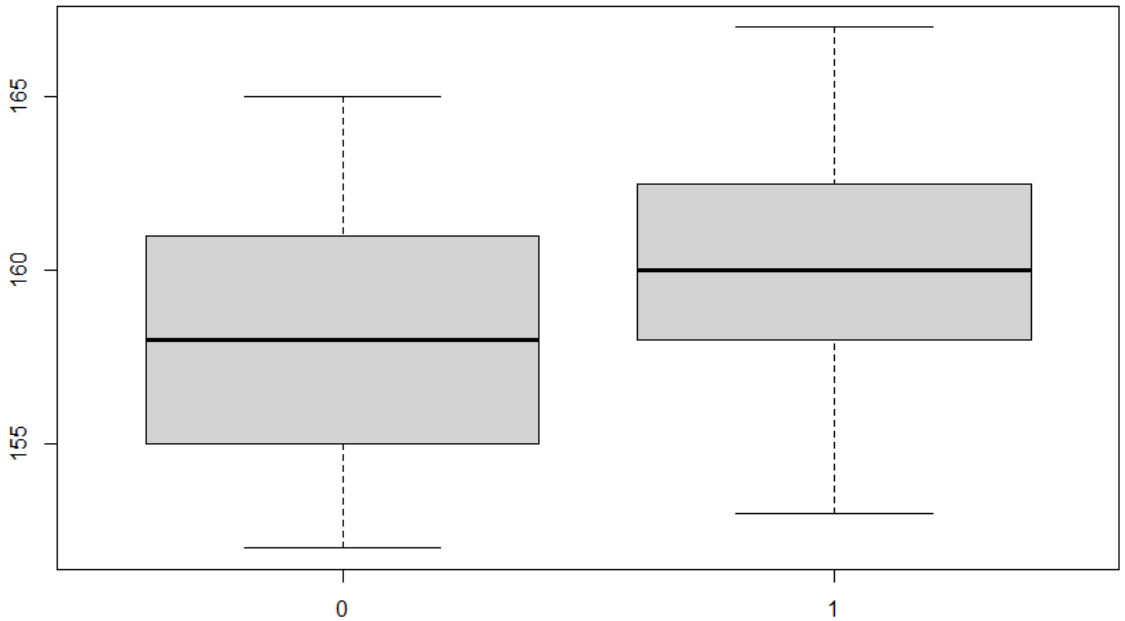
length<-boxplot(len ~ sex, data = bumpus_data, col = "lightgray")

width<-boxplot(wt ~ sex, data = bumpus_data, col = "lightgray")

sternum<-boxplot(stern ~ sex, data = bumpus_data, col = "lightgray")

- Let’s view it in the plot output:





Plotting data

- Barplots:

barplot(Table, ylab="Mean length", xlab="Sex")

barplot(Table2, ylab="Mean width", xlab="Sex")

barplot(Table3, ylab="Mean sternum", xlab="Sex")c

```
barplot(Table, ylab="Mean length", xlab="Sex")
barplot(Table2, ylab="Mean width", xlab="Sex")
barplot(Table3, ylab="Mean sternum", xlab="Sex")
```

Assumptions of normality ????

- A principal components analysis is a multivariate technique that can be used to explore the relationships between variables within a matrix dataset

- The assumptions of normally distributed errors apply to tests, ***but they do not apply*** if you are doing an exploratory analysis for hypotheses formation provided that the variables are not strongly skewed to the right

- The exploration analysis is accomplished with the dimensionality-reduction method
  - Reduce the number of variables in the dataset while explaining all or most of the variance in it

- We begin by normalising/standardising the variables to have a standard deviation of one using parameter scale = T(rue)

- In other words, normalising/standardising allows the range of continuous variables to contribute equally to the variance found in the dataset!

Executing PCA

- prin_comp does not recognize categorical variables

- Recall the original data:

```
bumpus_data<-BumpusSparrowsJanzenStern

#then

prin_comp <- prcomp(~len+ext+wt+head+humer+femur+tibio+skull+stern, scale = T, data=bumpus_data)

## Let's see what prin_comp comes with

names(prin_comp)
```

bumpus_data<-BumpusSparrowsJanzenStern

- Execute the model by calling only the variables you want to see:

prin_comp <- prcomp(~len+ext+wt+head+humer+femur+tibio+skull+stern, scale = T, data=bumpus_data)

- Let’s view prin_comp output:

names(prin_comp)

```
> names(prin_comp)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

Executing PCA

- The mean and standard deviation of the variables used for normalisation prior to implementing PCA are referred to as centre and scale, respectively.

- Output the mean of variables:

  prin_comp$center

```
# Output the mean of variables
prin_comp$center

#outputs the standard deviation of variables
prin_comp$scale
```

```
> prin_comp$center
        len         ext          wt        head       humer       femur       tibio       skull       stern
159.5441176 245.2573529  25.5250000  31.5727941   0.7321618   0.7129338   1.1335662   0.6024118   0.8399338
```

Executing PCA

- Output the standard deviation of variables

prin_comp$scale

```
# Output the mean of variables
prin_comp$center

#outputs the standard deviation of variables
prin_comp$scale
```

```
> prin_comp$scale
       len        ext         wt       head      humer      femur      tibio      skull      stern
3.56083146 5.51227092 1.47521499 0.70547771 0.02322171 0.02415358 0.04074450 0.01498665 0.03964924
```

Executing PCA

- Output the PCA loadings:

prin_comp$rotation

```
## The principal component loading is determined by the rotation measure
## The principal component loading vector is found in each column of the rotation matrix
## This is the most important metric to pay attention to

prin_comp$rotation
```

```
> prin_comp$rotation
            PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8          PC9
len   0.3114224 -0.48427067 -0.05281482  0.43023525 -0.10961330  0.39114345 -0.51423034 -0.14818691 -0.170896937
ext   0.3521077 -0.28200902 -0.33432126  0.23533693 -0.23736889  0.04004504  0.58079761  0.36909115  0.312786128
wt    0.3164368 -0.33334464  0.17386215  0.08082929  0.71636233 -0.46786927  0.03641344 -0.08605999  0.105729679
head  0.3320588  0.12845733  0.33755306 -0.30446584  0.31777742  0.70453061  0.26160802  0.01691163 -0.022880931
humer 0.3831926  0.21481310 -0.24657286  0.03215779 -0.09110199 -0.17331327  0.30835563 -0.40139080 -0.671305903
femur 0.3616831  0.41970039 -0.16305757  0.04230734 -0.07978064 -0.01410798 -0.18475201 -0.48588169  0.622800880
tibio 0.3392715  0.46763813 -0.18573614  0.08570522  0.17226011 -0.07726138 -0.37726707  0.65099252 -0.150338233
skull 0.2928700  0.07093272  0.77380963  0.13806889 -0.46009334 -0.26387380  0.01817180  0.09822807 -0.007558083
stern 0.3001862 -0.33514616 -0.14354290 -0.79438587 -0.24791027 -0.15589398 -0.23790636  0.05902907  0.008094029
```

Executing PCA

- Output the portion of variance explained by the principal components:

summary(prin_comp)

```
##Output the portion of variance explained by the principal components

summary(prin_comp)
```

```
> summary(prin_comp)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.3103 1.0008 0.81567 0.73270 0.67989 0.64192 0.51325 0.44305 0.35295
Proportion of Variance 0.5931 0.1113 0.07392 0.05965 0.05136 0.04579 0.02927 0.02181 0.01384
Cumulative Proportion  0.5931 0.7044 0.77828 0.83793 0.88929 0.93508 0.96435 0.98616 1.00000
```

Executing PCA

- Better to visualize as a screeplot

- Extract the variance attributed to the PCs

```
std_dev <- prin_comp$sdev
```

- Compute variance squared

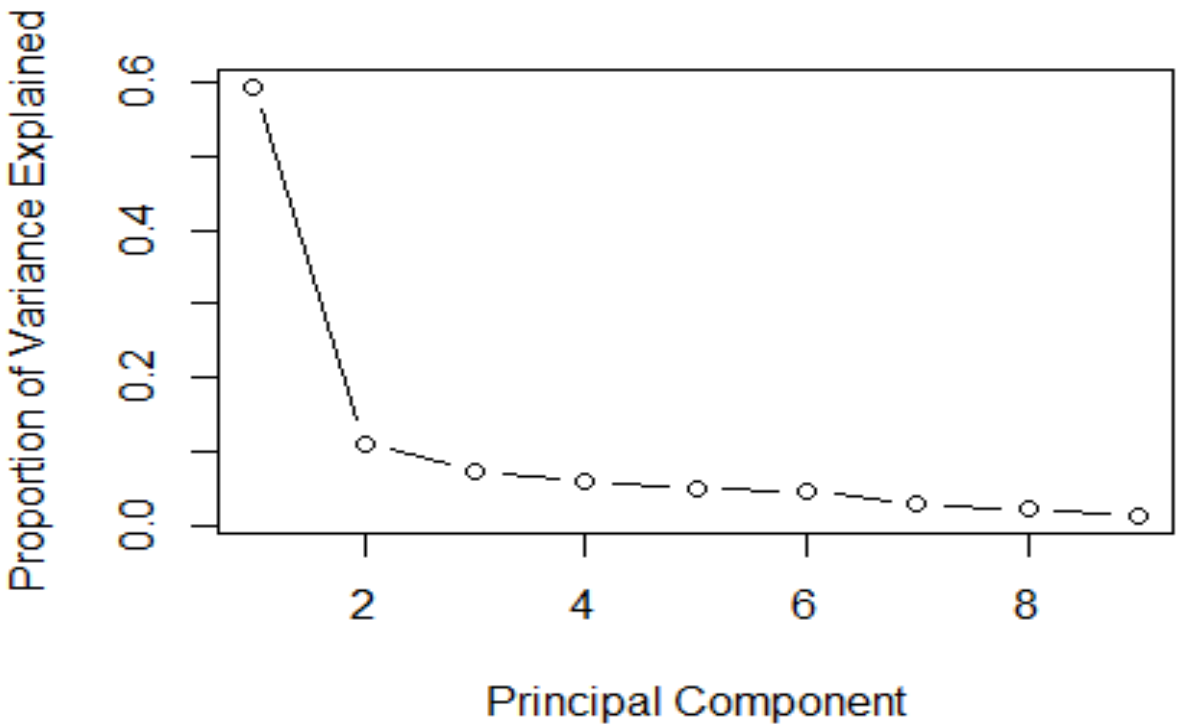
```
pr_var <- std_dev^2
```

- Compute the proportion of variance explained

```
prop_varex <- pr_var/sum(pr_var)
```

- Produce the scree plot

```
plot(prop_varex, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained",
    type = "b")
```



Executing PCA

Let's visualize the PC1 and PC2 in a biplot:

```
biplot(prin_comp, scale = T)
abline(v=0)
abline(h=0)
```

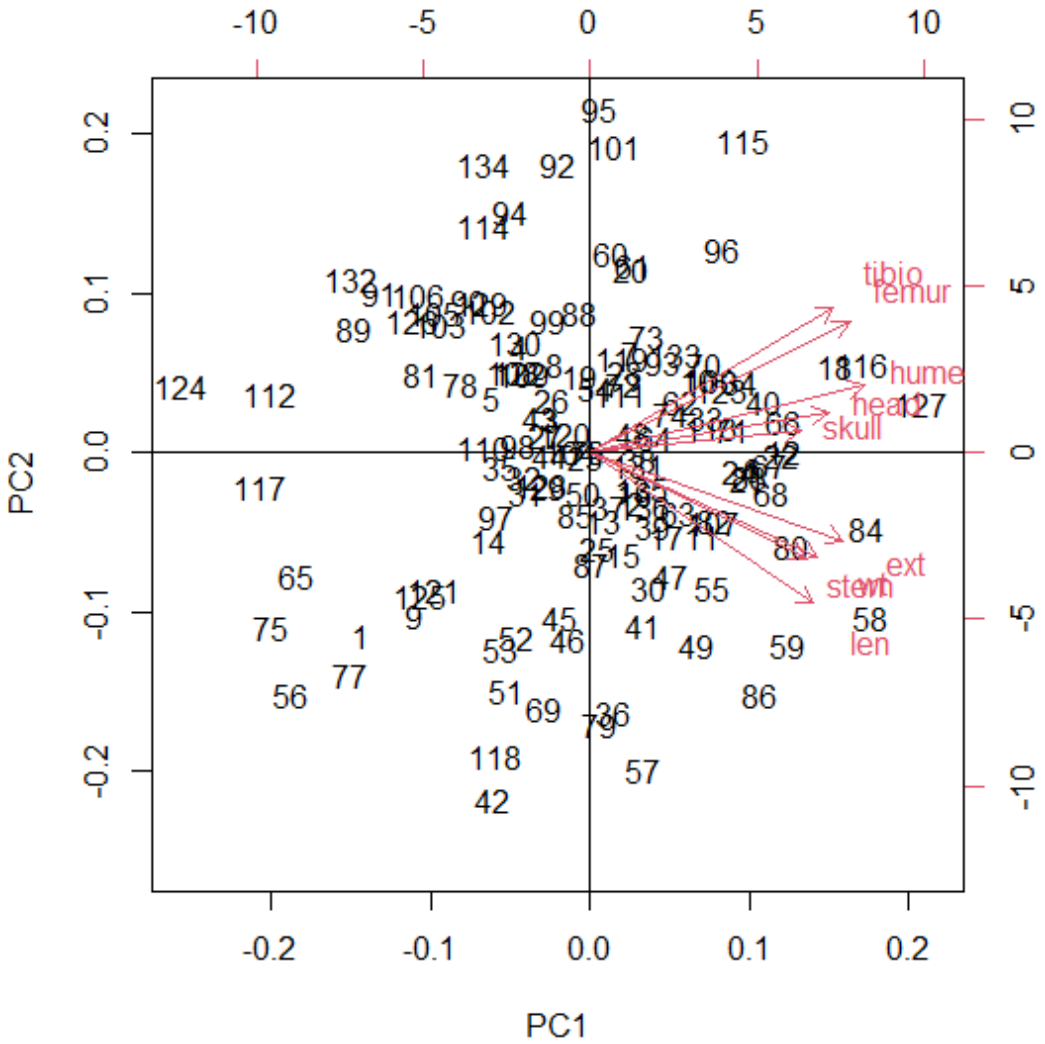
Most of our variance (70.44%) is explained by PC1 (size, 59%) and PC2 (shape, 11%)!

Size and Shape are likely driven by the sex

The east to west direction of the arrows are mostly east, which suggests that most of the size difference is due to sex.

The northward and southward direction of the arrows suggest there are shape differences b/w males and females

```
biplot(prin_comp, scale = T)
abline(v=0)
abline(h=0)
```



MANOVA as confirmatory analysis

- A PCA can also be used with a confirmatory analysis, e.g., MANOVA
- If we each response variable separately we leave out the correlation information. If they are correlated, then we can analyze all the response variables in question simultaneously!
- Analyzing the correlations allows us to include the known relationships!
- This is the basis of using a confirmatory approach on multivariate data and for hypothesis formation

Executing MANOVA

- Execute the MANOVA model as follows:

  res.man <- manova(cbind(len,ext,wt,head,humer,femur,tibio,skull,stern) ~ sex, data=bumpus_data)

```
res.man <- manova(cbind(len,ext,wt,head,humer,femur,tibio,skull,stern) ~ sex, data=bumpus_data)
summary(res.man)
```

```
> res.man <- manova(cbind(len,ext,wt,head,humer,femur,tibio,skull,stern) ~ sex, data=bumpus_data)
> summary(res.man)
          Df  Pillai approx F num Df den Df    Pr(>F)
sex        1 0.47549   12.692      9    126 3.357e-14 ***
Residuals 134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Executing MANOVA

- We can call which ones are different using the anova summary function

summary(res.man)

```
> summary.aov(res.man)
 Response len :
            Df  Sum Sq Mean Sq F value    Pr(>F)
sex          1  187.49 187.491  16.483 8.304e-05 ***
Residuals  134 1524.24  11.375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response ext :
            Df Sum Sq Mean Sq F value    Pr(>F)
sex          1 1183.5 1183.54  54.342 1.565e-11 ***
Residuals  134 2918.4   21.78
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Executing MANOVA

- We can call which ones are different using the anova summary function

summary(res.man)

```
             Df Sum Sq Mean Sq F value    Pr(>F)
sex           1 1183.5 1183.54  54.342 1.565e-11 ***
Residuals   134 2918.4   21.78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response wt :
             Df  Sum Sq Mean Sq F value   Pr(>F)
sex           1  18.877 18.8768  9.2009 0.002906 **
Residuals   134 274.918  2.0516
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response head :
             Df Sum Sq Mean Sq F value Pr(>F)
sex           1  0.989 0.98867  2.0012 0.1595
Residuals   134 66.201 0.49403

 Response humer :
             Df   Sum Sq    Mean Sq F value  Pr(>F)
sex           1 0.002536 0.00253568  4.8359 0.02959 *
Residuals   134 0.070263 0.00052435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response femur :
             Df   Sum Sq    Mean Sq F value Pr(>F)
sex           1 0.000079 0.00007898  0.1345 0.7144
Residuals   134 0.078679 0.00058716
```

Executing MANOVA

- We can call which ones are different using the anova summary function

summary(res.man)

```
sex           1 0.002536 0.00253568  4.8359 0.02959 *
Residuals   134 0.070263 0.00052435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response femur :
             Df   Sum Sq    Mean Sq F value Pr(>F)
sex           1 0.000079 0.00007898  0.1345 0.7144
Residuals   134 0.078679 0.00058716

 Response tibio :
             Df   Sum Sq    Mean Sq F value Pr(>F)
sex           1 0.000628 0.00062759  0.3763 0.5406
Residuals   134 0.223488 0.00166782

 Response skull :
             Df    Sum Sq    Mean Sq F value Pr(>F)
sex           1 0.0001662 0.00016619  0.7385 0.3917
Residuals   134 0.0301547 0.00022504

 Response stern :
             Df   Sum Sq   Mean Sq F value    Pr(>F)
sex           1 0.030188 0.0301879  22.221 6.012e-06 ***
Residuals   134 0.182041 0.0013585
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```

Publication quality biplot

```
library(ggplot2)
library(ggfortify)

bp<-autoplot(prin_comp, data=bumpus_data, colour = 'sex',
        loadings = TRUE, loadings.colour = 'blue',
        loadings.label = TRUE, loadings.label.size = 3)
```