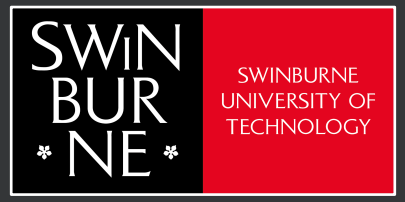
STA30005:

MULTIVARIATE ANALYSIS

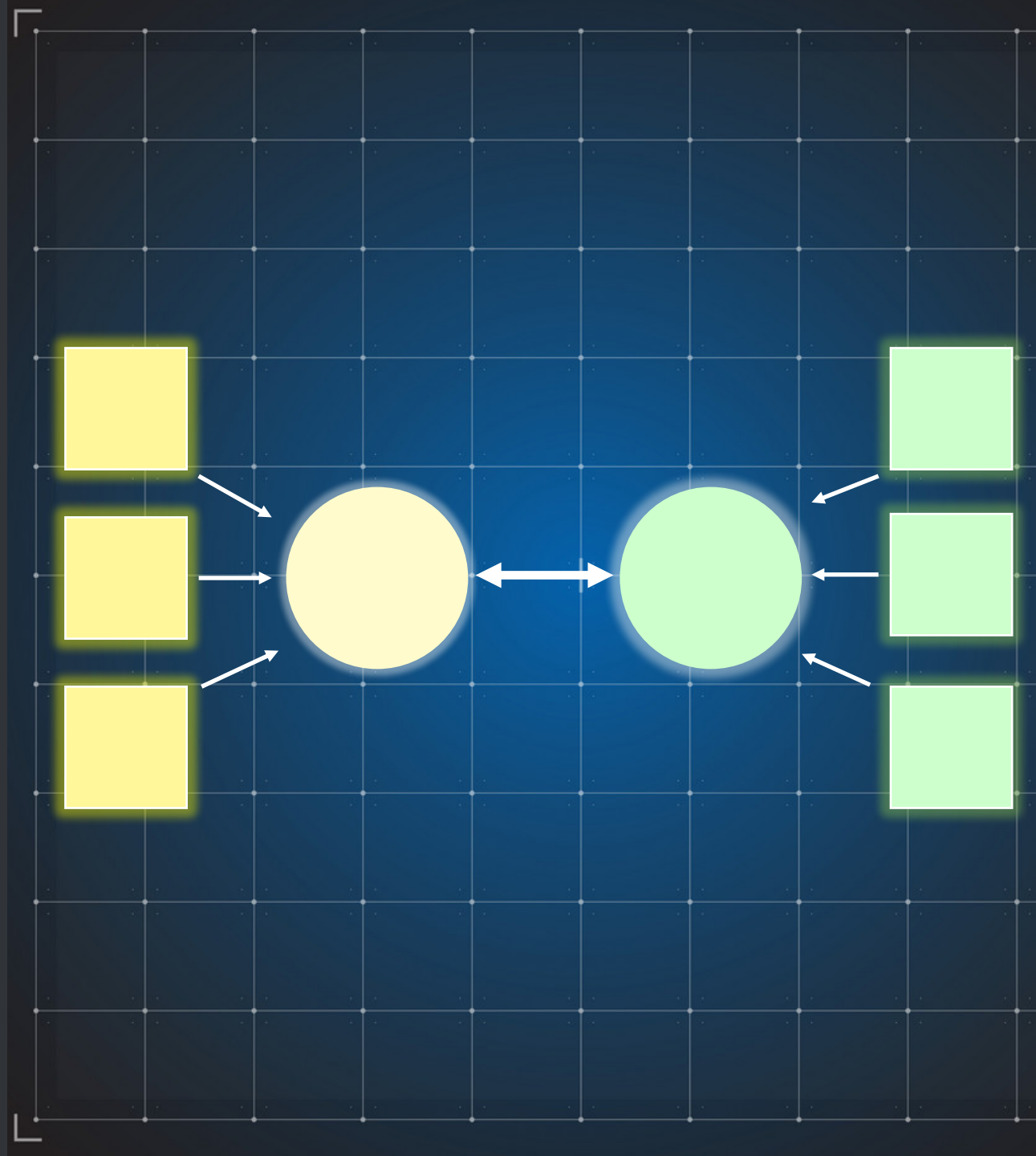


WEEK 10:

CANONICAL CORRELATION

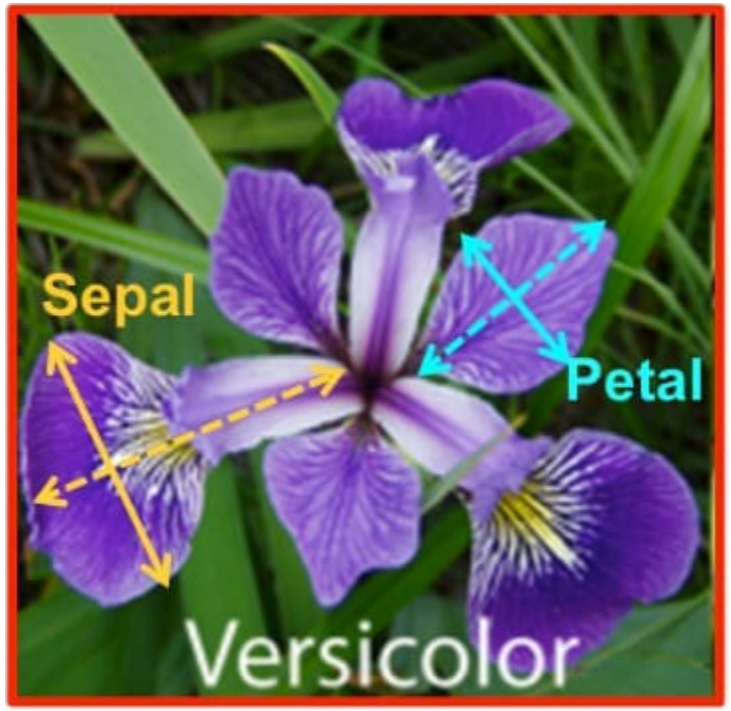
Tutorial Slides

STA30005: Multivariate Analysis



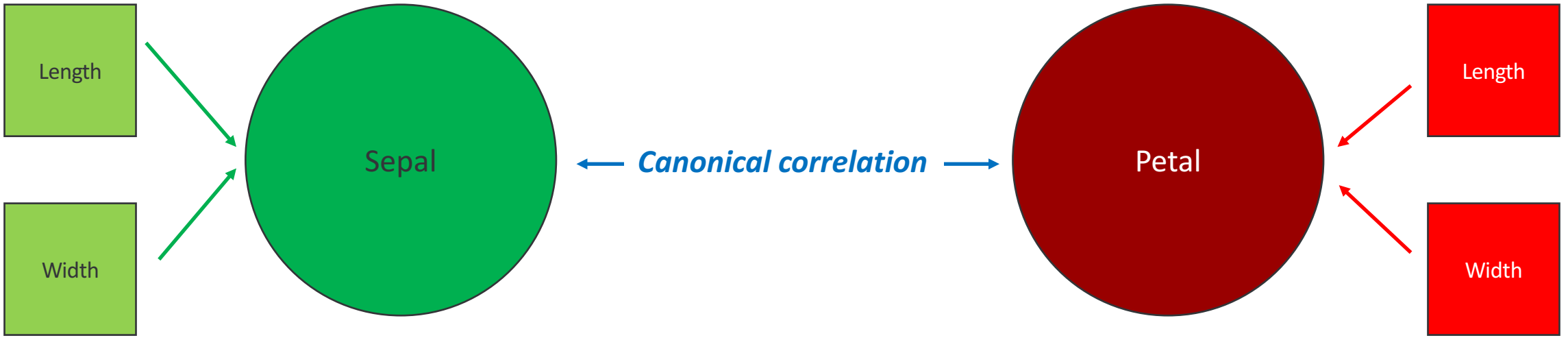
THE DATA

The data file iris contains the measurements for 150 iris flowers regarding the length and width of their sepals and petals. Naturally, the sepal length and width should be related (as well as the petal length and width). In other words, there is collinearity among the variables. This makes multiple regression inappropriate and other methods (i.e. canonical correlations) would be required.



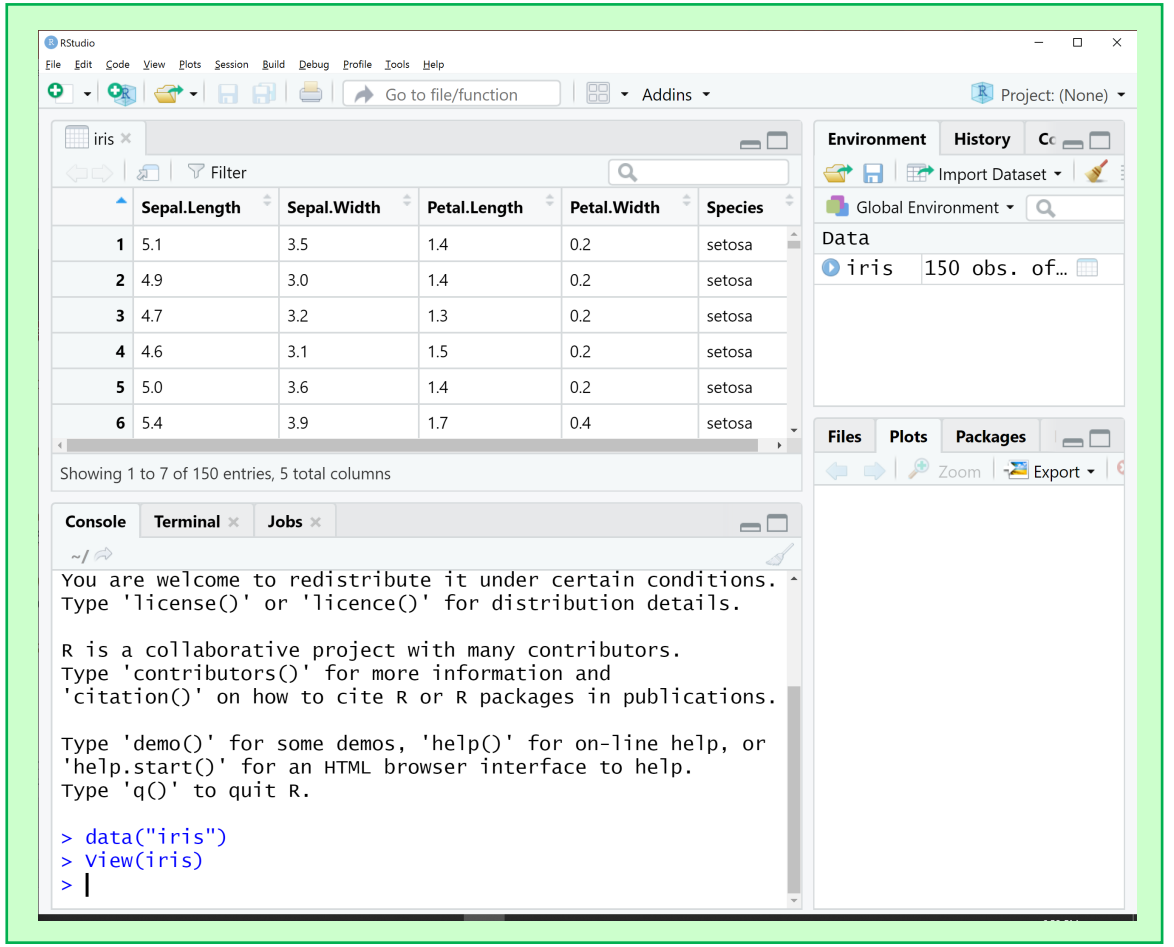
THE MODEL

In this model we believe there are two variates (Sepal and Petal), both of which have two variables: length and width. Conduct a canonical correlation on the iris data and determine the main elements for this model



CANONICAL CORRELATION

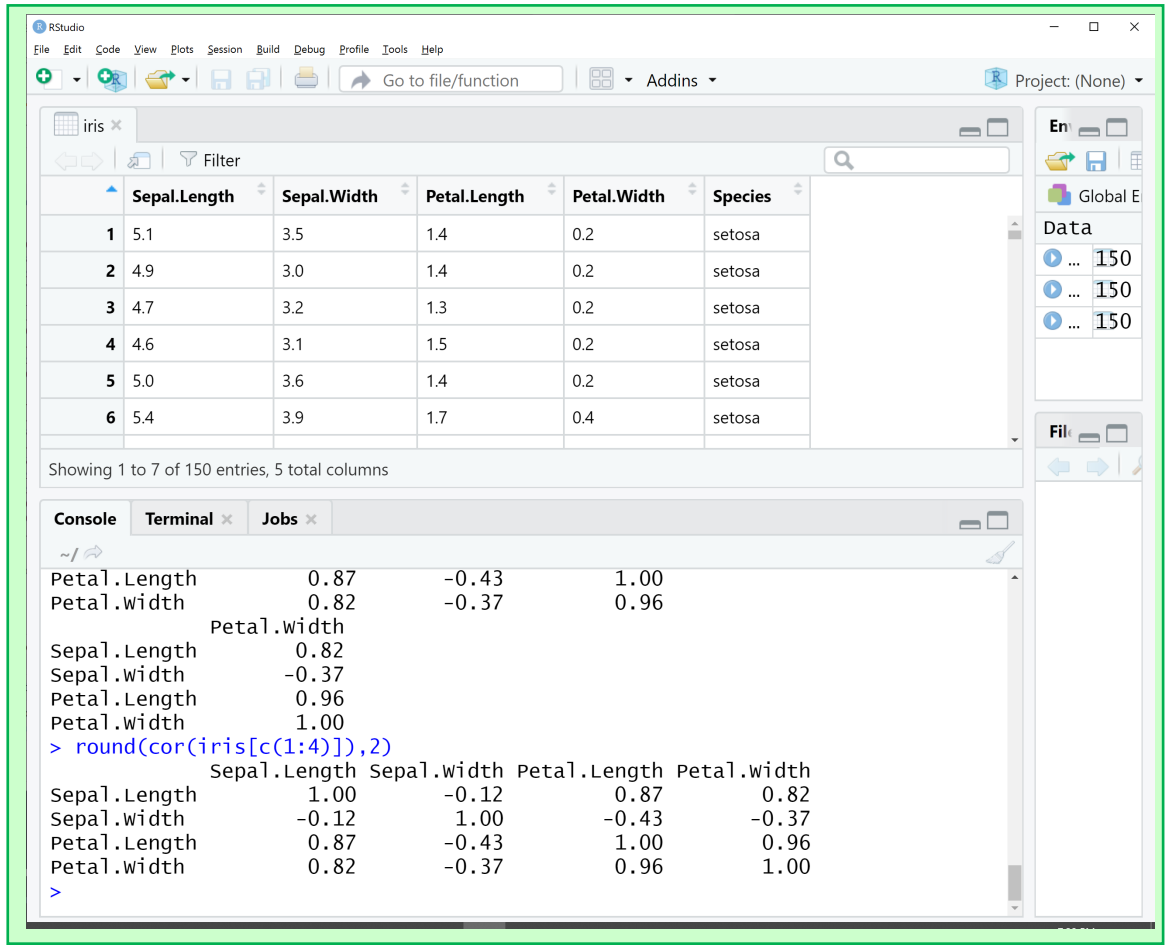

```
# Load the data (this is a base data set for R)
data("iris")

# Show the data
View(iris)

# Define the first variate (sepal) based on columns 1 and 2
sepal<-iris[,1:2]

# Define the second variate (petal) based on columns 3 and 4
petal<-iris[,1:2]

# Install / Load the required packages
Install.packages("yacca")
library(yacca)
Install.packages("CCA")
library(CCA)
```



CANONICAL CORRELATION


```
# Check the bivariate correlations
round(cor(iris[c(1:4)]),2)
```

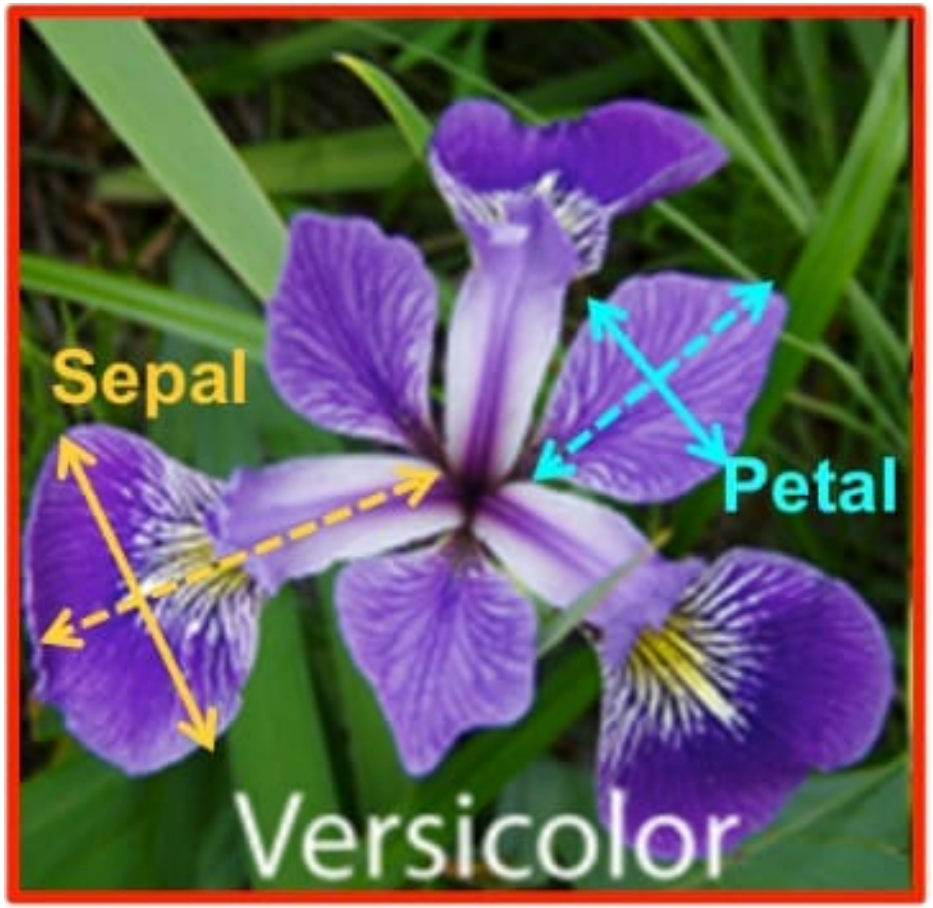
This gives us an indication of the correlations within the variates:



CANONICAL CORRELATION

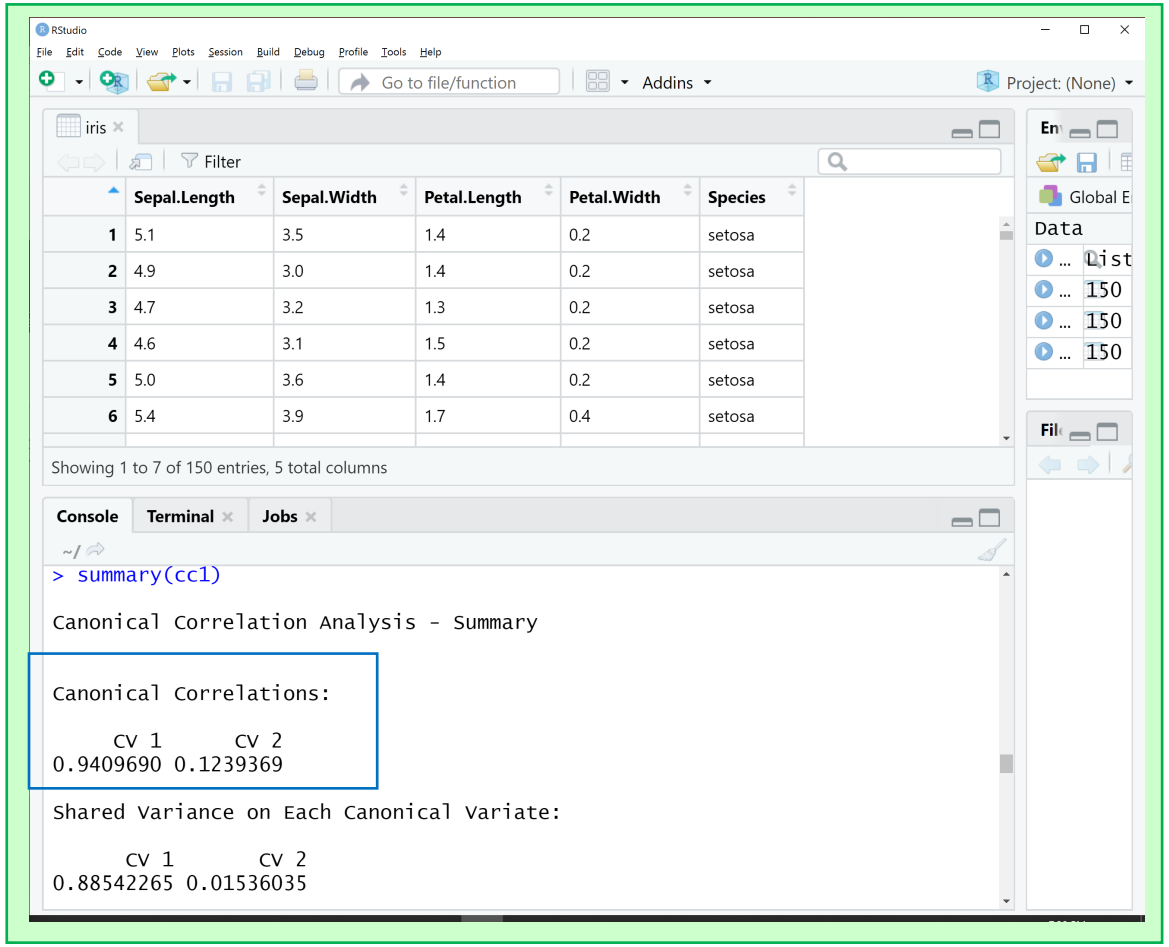

```
# Check the bivariate correlations
round(cor(iris[c(1:4)]),2)
```

This gives us an indication of the correlations within the variates:



CANONICAL CORRELATION

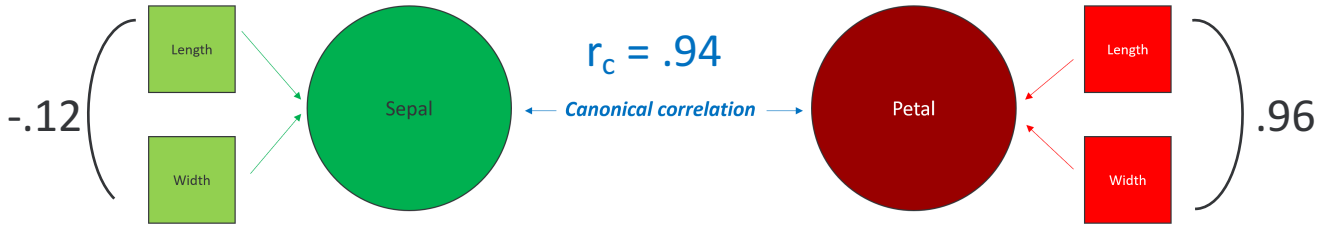

```
# Define and create your canonical correlation model
cc1<-cca(sepal,petal)

# Show the canonical correlation values for your model
summary(cc1)
```

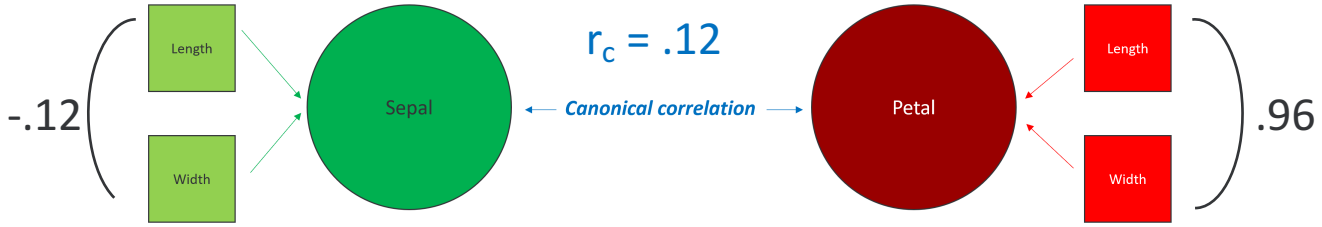


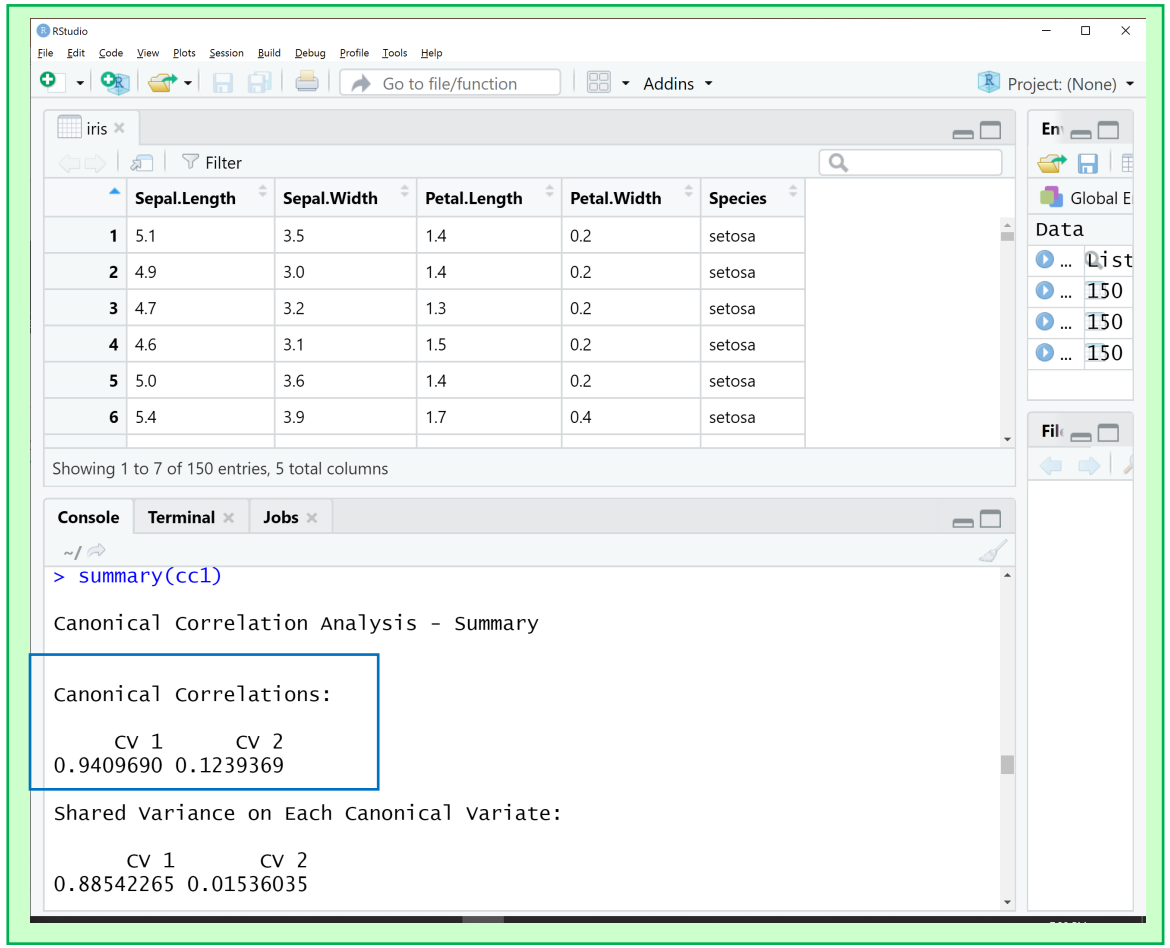
CANONICAL CORRELATION


```
# Define and create your canonical correlation model
cc1<-cca(sepal,petal)

# Show the canonical correlation values for your model
summary(cc1)
```

* Note: scroll up (as this provides a lot of output)

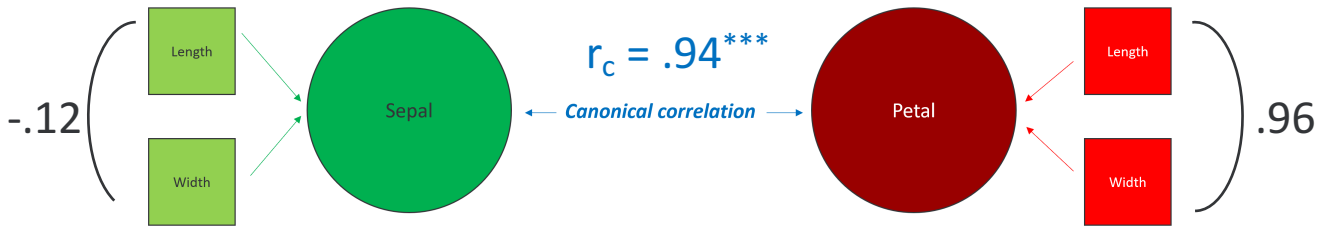


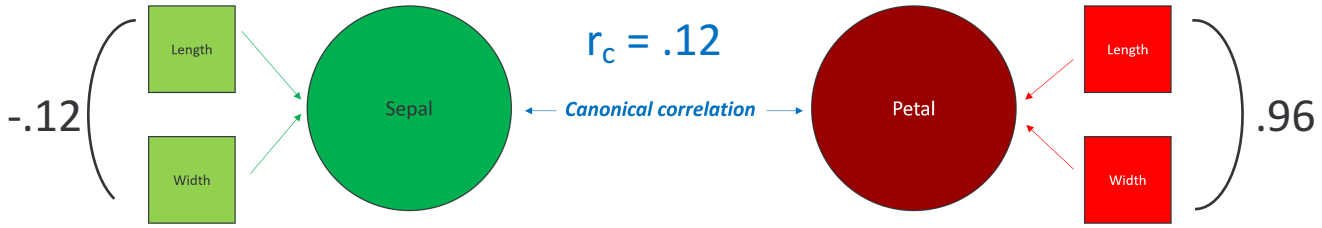


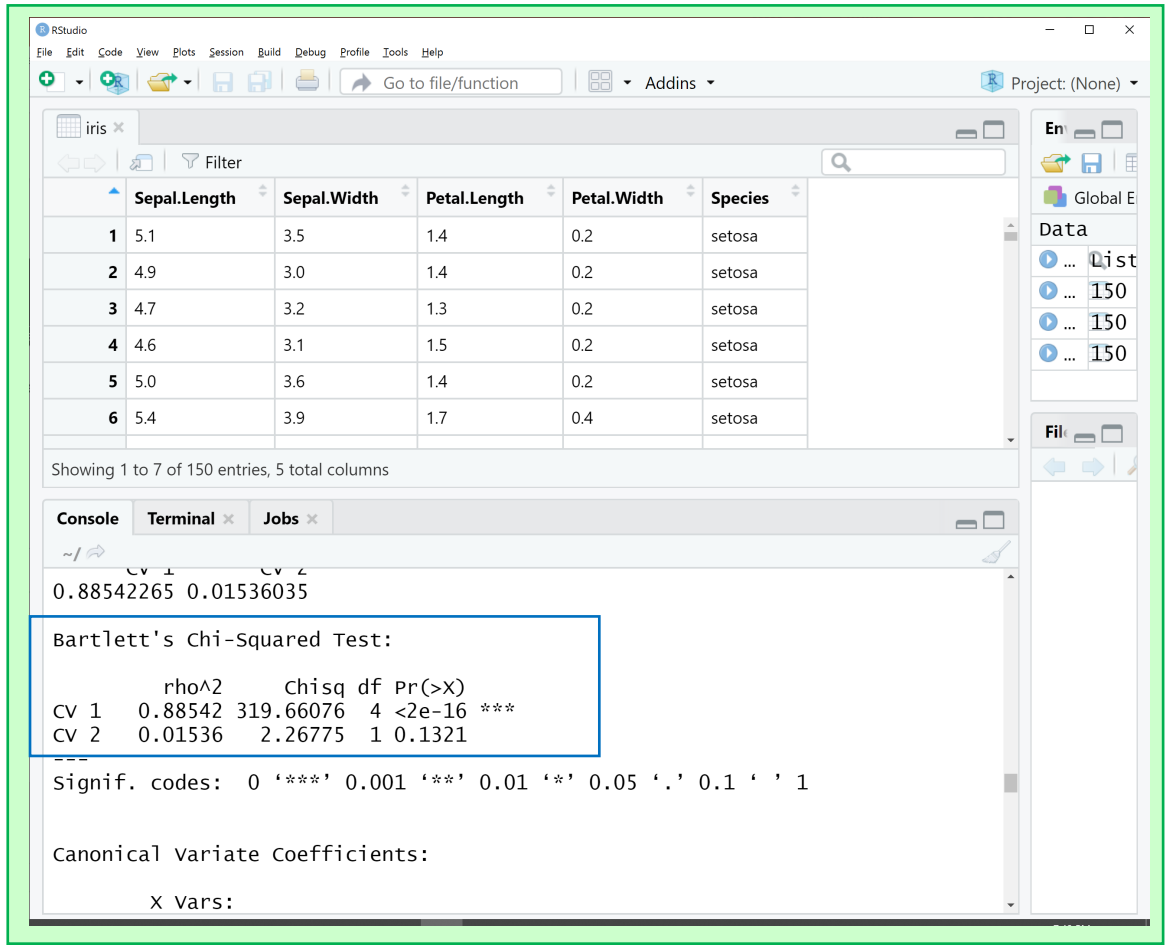


CANONICAL CORRELATION

- Part of the output from the previous command gives us the chi-square tests
- Here only the first canonical correlation is significant, $\chi^2(4) = 319.66$, $p < .001$



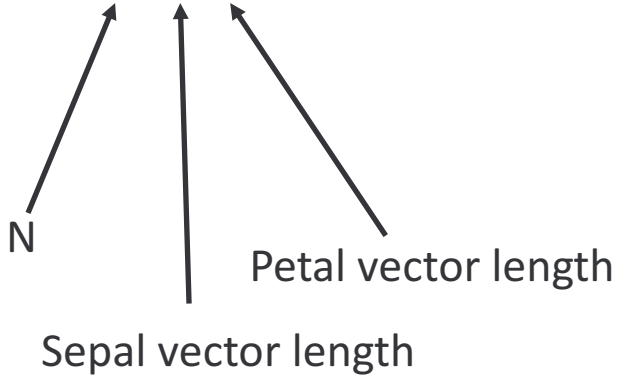


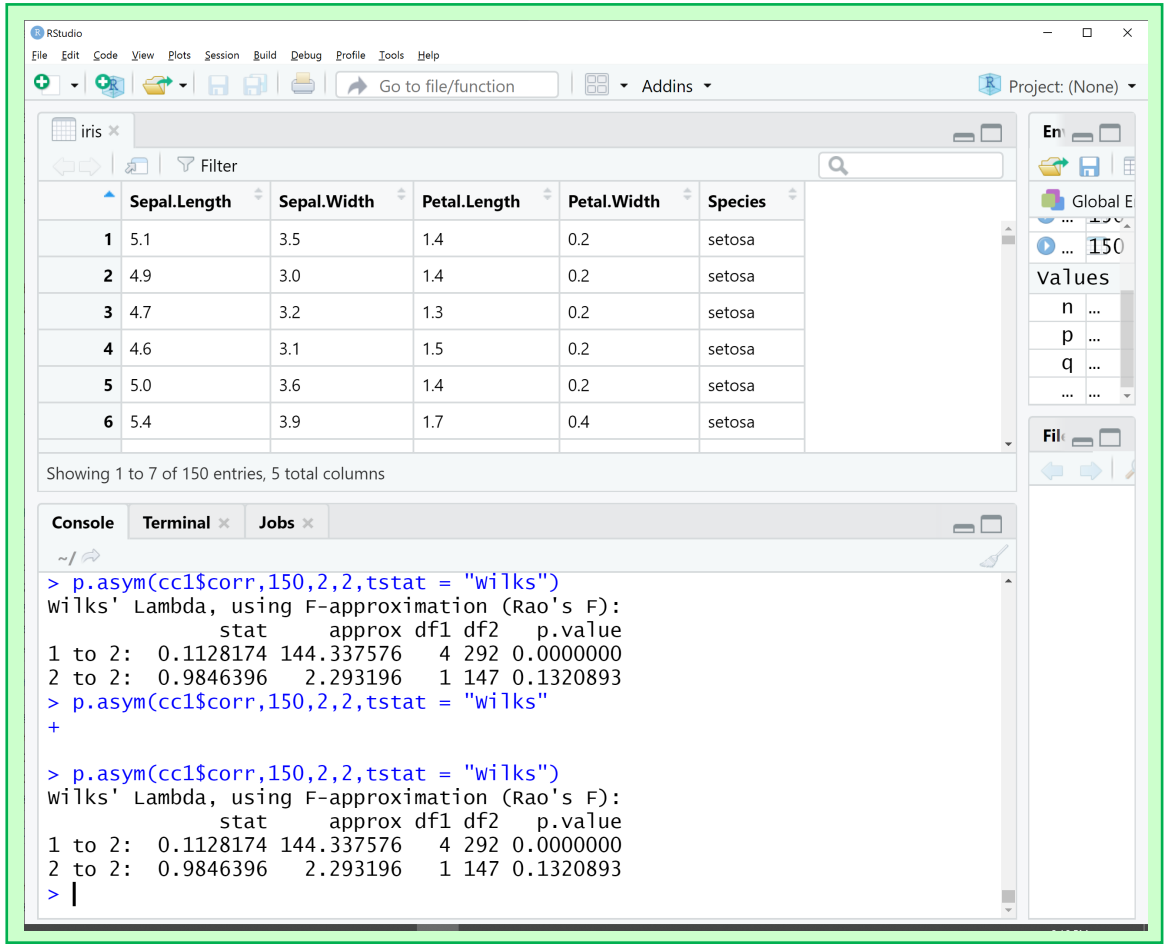


CANONICAL CORRELATION

We can also use other multivariate methods to assess for significance as well

```
# Install / Load the required packages
Install.packages("CCP")
library(CCP)

# Check Wilk's Lambda (or any other method)
p.asym(cc1$corr,150,2,2,tstat = "Wilks")
```

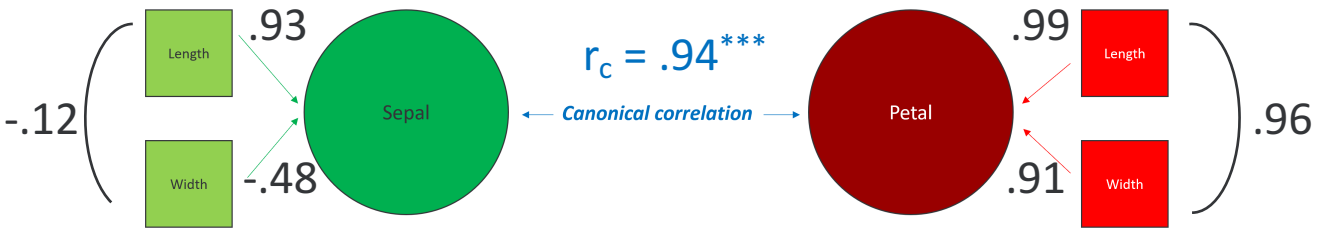


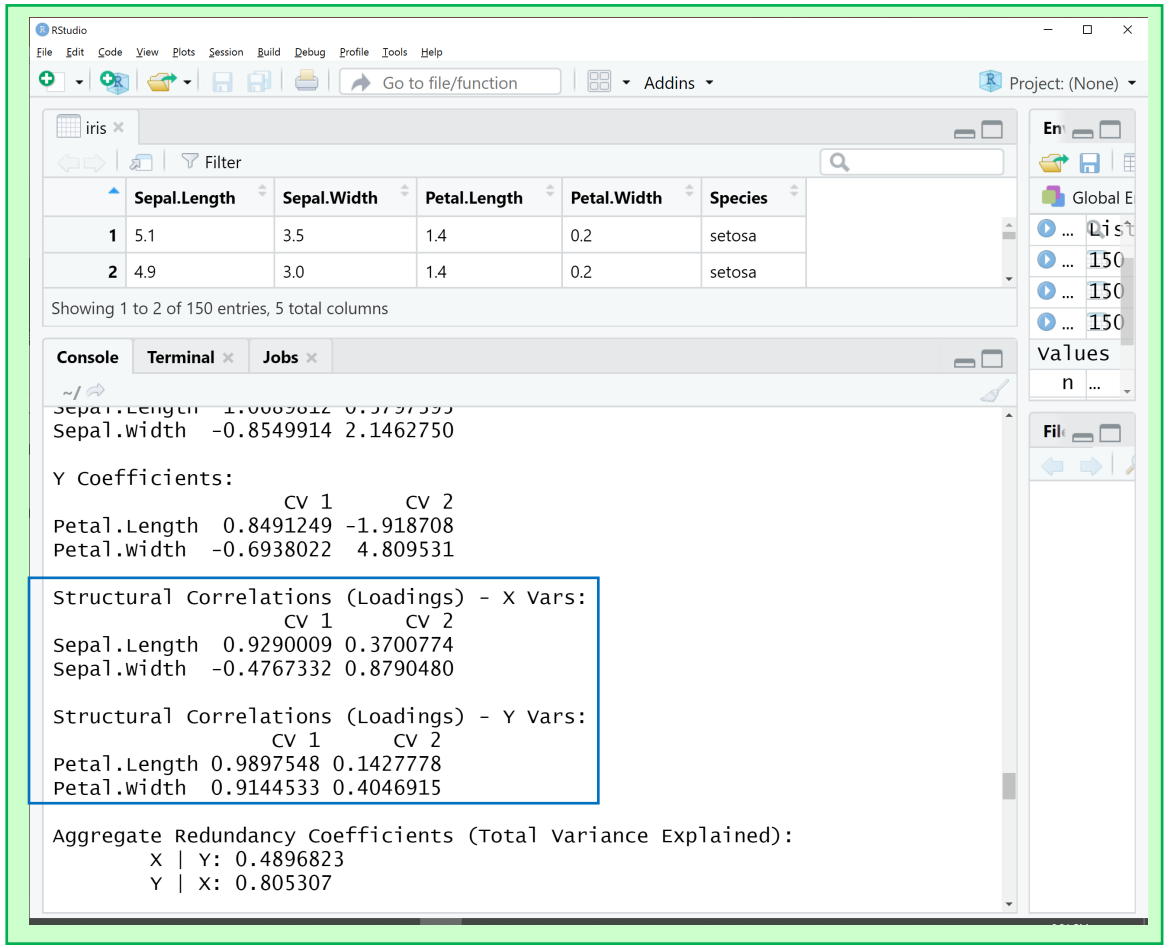


CANONICAL CORRELATION

The yacca package and the cca function we ran earlier provides us with the structural loadings (note: we'll only comment on the first one as the second was not significant)

```
# You can re-run if you lost the previous output
summary(cc1)
```



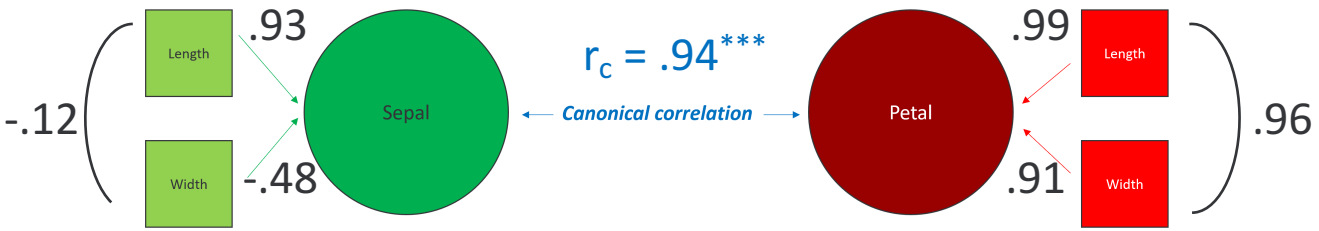


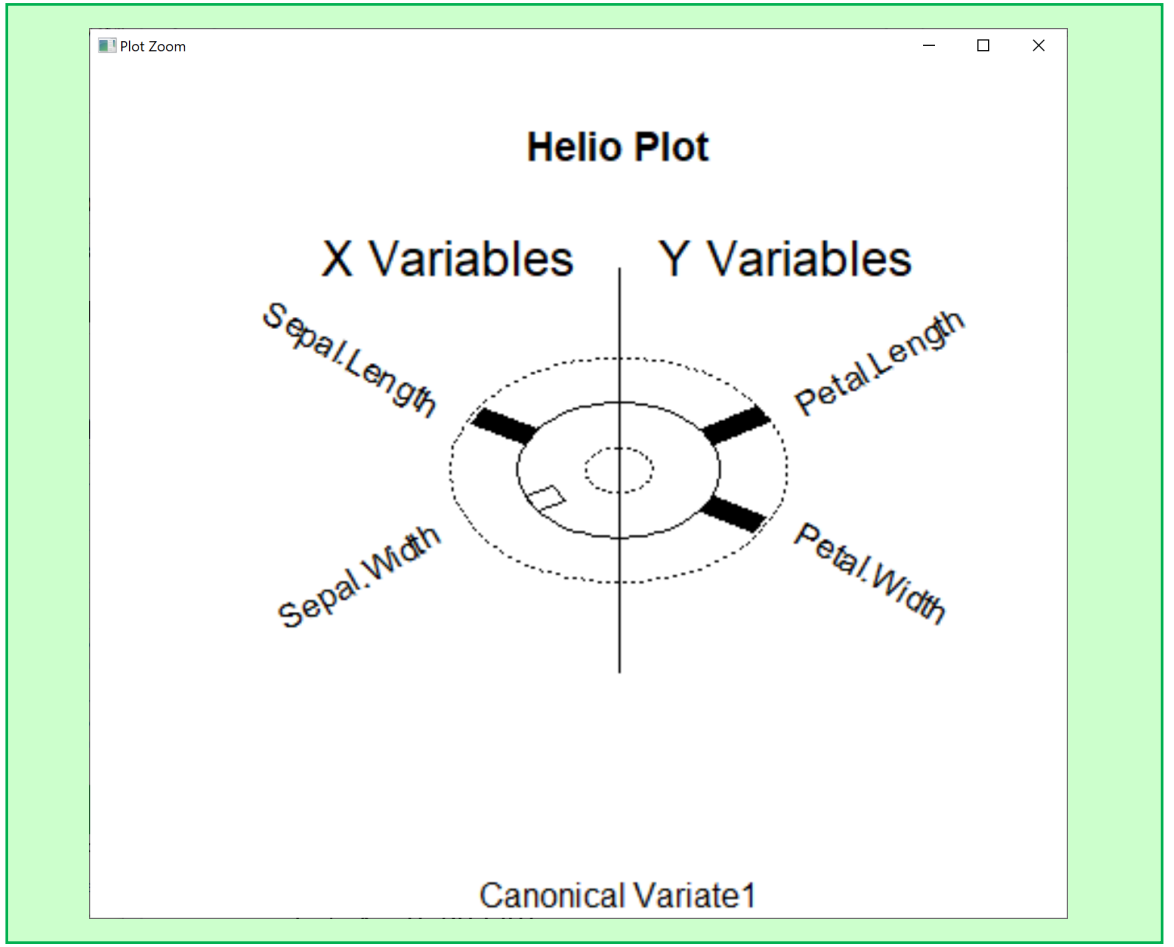
CANONICAL CORRELATION

The yacca package also allows us to create helio plots to visualize our data

```
# Create helio plot
Helio.plot(cc1)
```

Comparing this plot to our model below, what do you think the black bars represent? What about the white?





CANONICAL CORRELATION

Now load the LifeCycleSavings data file (this is a base R file) and repeat what we just did
(Hint: inspect the data frame first, and think about which variables will go to which sets)

About the data:

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.