Introducción a R y RStudio

Francisco Charte

En esta sesión nos ocuparemos de:

- R y RStudio
  - ¿Por qué aprender R?
  - Herramientas de trabajo
  - Ejecución de tareas habituales
- Introducción a R
  - Tipos de datos fundamentales en R
  - Cómo cargar conjuntos de datos
  - Fundamentos de análisis exploratorio de datos

R y RStudio

¿Por qué aprender R?

¿Por qué aprender R?

- Es el lenguaje más usado para análisis de datos

- Conocimiento de R altamente valorado

Fuente KDnuggets

R, 46.9% share (38.5% in 2014)
RapidMiner, 31.5% (44.2% in 2014)
SQL, 30.9% (25.3% in 2014)
Python, 30.3% (19.5% in 2014)
Excel, 22.9% (25.8% in 2014)
KNIME, 20.0% (15.0% in 2014)
Hadoop, 18.4% (12.7% in 2014)
Tableau, 12.4% (9.1% in 2014)
SAS, 11.3 (10.9% in 2014)
Spark, 11.3% (2.6% in 2014)

Fuente Dice Tech Salary Survey

AVERAGE SALARY FOR **High Paying Skills and Experience**

SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%

Además …

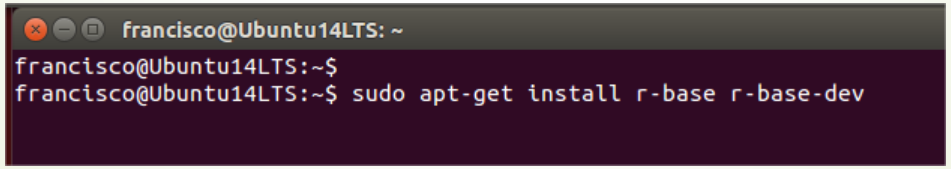
- R es *Open Source* (multiplataforma, libre, abierto)
- Gran número de paquetes disponibles (CRAN/GitHub)
- Extensa comunidad de usuarios
- Ciclo completo de trabajo:
  - Implementación de algoritmos
  - Preparación de datos
  - Análisis de resultados
  - Generación de documentación

R y RStudio

Herramientas de trabajo

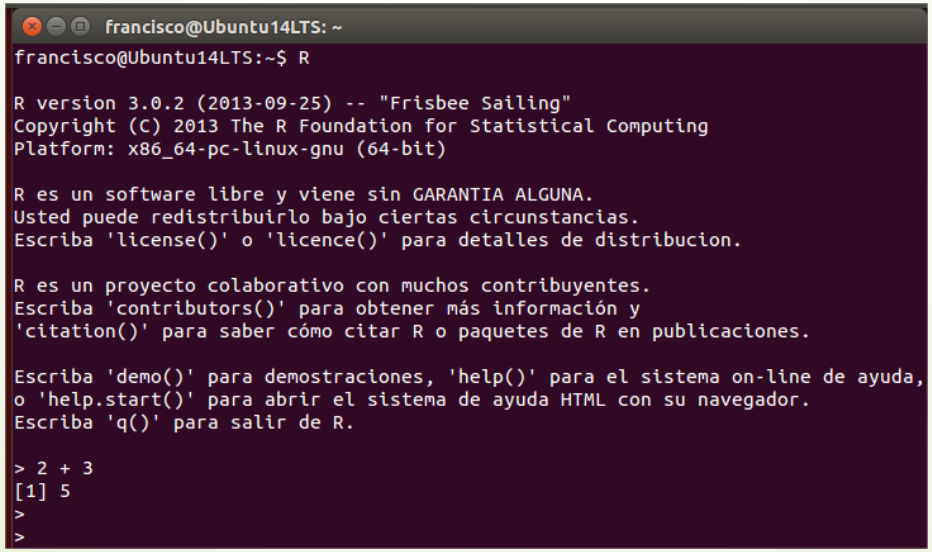
Herramientas de trabajo - R

- Binarios disponibles para Linux, OS X y Windows
- Descarga desde http://www.r-project.org/
- Disponible en repositorios Linux



Herramientas de trabajo - R

- Trabajo interactivo mediante línea de comandos

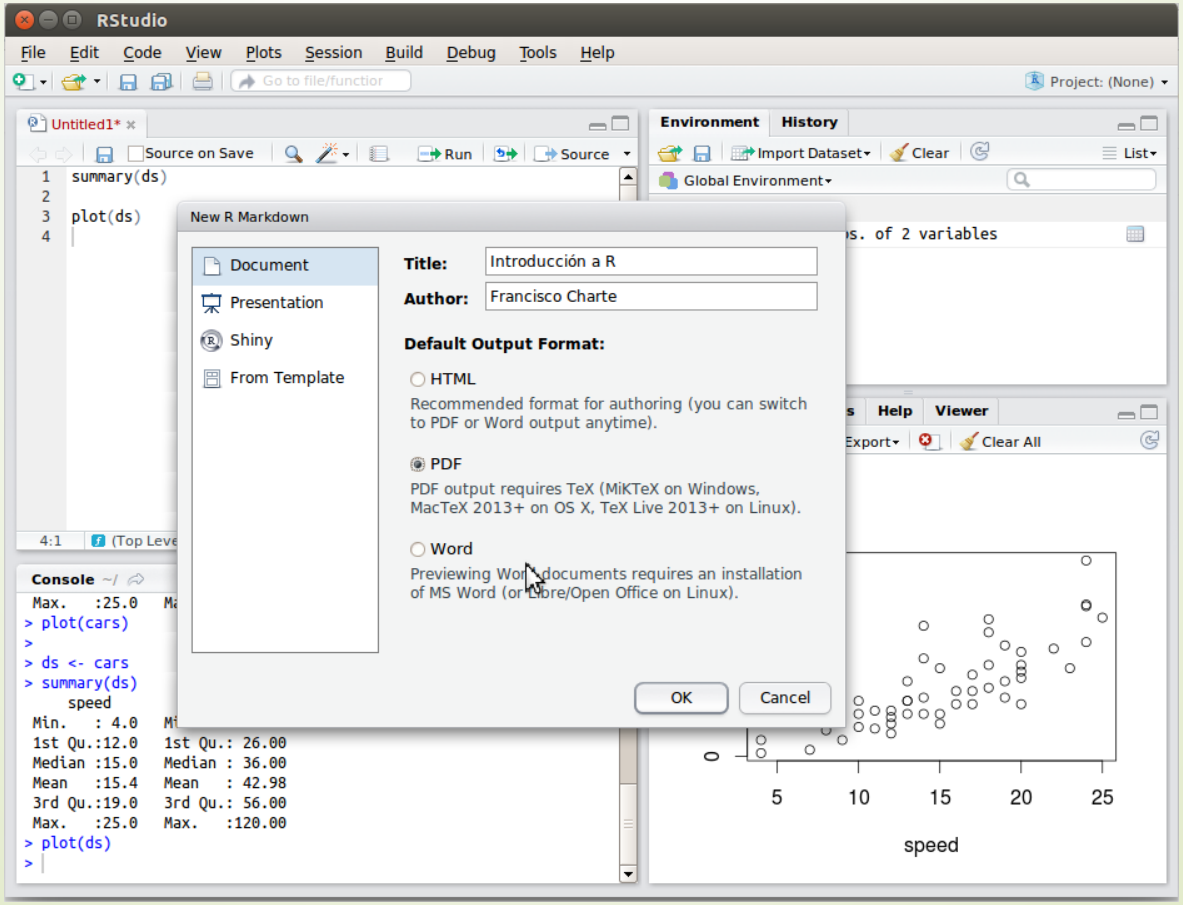


Herramientas de trabajo - RStudio

- Binarios disponibles para Linux, OS X y Windows
- Descarga desde http://www.rstudio.com/
- Licencia Open Source y comercial
- IDE estándar para trabajar con R
- **Será la herramienta que usemos en el curso**

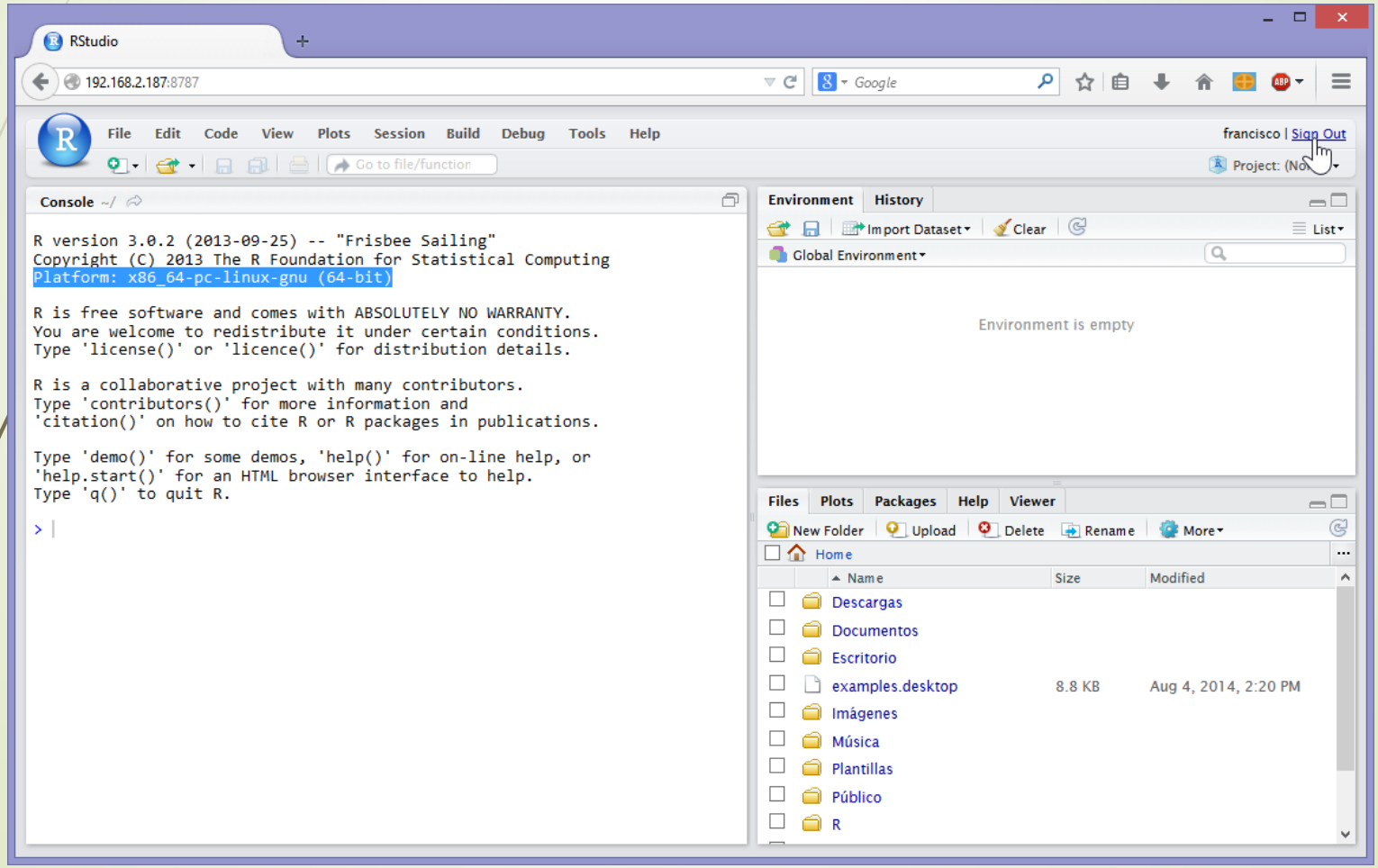
Herramientas de trabajo - RStudio

- IDE completo para trabajar con R
- Gestión de proyectos y paquetes
- Editor de scripts
- Acceso a objetos del entorno
- Visualización de gráficas y ayuda
- Consola de R integrada



Herramientas de trabajo - RStudio

- Instalable en un servidor web
- Accesible desde el navegador
- Multiusuario
- Idéntica interfaz de RStudio



R y RStudio

Ejecución de tareas habituales

Ejecución de tareas habituales

▸`tareasHabituales.R`

- Acceso a la documentación
  - `help('source'), vignette('grid'), demo('image')`
- Ruta de trabajo
  - `getwd(), setwd()`
- Espacio de trabajo
  - `save(), save.image(), load()`
- Instalación de paquetes
  - `install.packages(), library()`

Introducción a R

Tipos de datos fundamentales en R

Tipos de datos simples

▶`tiposDatos.R`

- numeric
  - Enteros :: `1024, -3`
  - Punto flotante :: `3.1415927`
  - Notación exponencial :: `3.85e6`
  - Otros :: `Inf, NaN`

- integer :: `as.integer(numeric)`
- complex :: `1+2i`
- character :: `'R', "Hola"`
- logical :: `TRUE, FALSE, NA`

Uso de variables

▶tiposDatos.R

- Asignación
  - `a = 1024 | a <- 1024 | 1024 -> a`
- Obtención de clase y tipo
  - `class(a) # numeric | typeof(a) # double`
- Comprobación de tipo
  - `is.numeric(a), is.character(a), is.integer(a), is.infinite(a), is.na(a)`
- Objetos en el espacio de trabajo
  - `ls(), rm(var), str(var), save(var,file = arch), save.image(), load()`

Vectores

▶tiposDatos.R

- Definición
  - `diasMes <- c(31,29,31,30,31,30,31,31,30,31,30,31)`
  - `dias <- c('Lun','Mar','Mié','Jue','Vie','Sáb','Dom')`
  - `quincena <- 16:30`
  - `semanas <- seq(1, 365, 7)`
  - `rep(T, 5)`
- Obtención número de elementos
  - `length(dias)`
- Acceso a elementos
  - `dias[2]` `# Segundo elemento del vector`
  - `dias[-2]` `# Todos los elementos menos el segundo`
  - `días[c(3, 7)]` `# Elementos 3 y 7`

Matrices

▶tiposDatos.R

- Definición
  - `mes <- matrix(1:35, nrow = 5)`
  - `mes <- matrix(1:35, ncol = 7, byrow = T)`
- Obtención número de elementos
  - `length(mes)` | `nrow(mes)` | `ncol(mes)`
- Acceso a elementos
  - `mes[2, ]  # 2ª fila completa`
  - `mes[ ,2]  # 2ª columna completa`
  - `mes[2, 5] # 5ª columna de la segunda fila`
  - `fix(mes)  # Editar elementos en la matriz`

Factors

▶tiposDatos.R

- Definición
  - `herramientas <- factor('Consola','RStudio')`
  - `fdias <- factor(días)`
  - `tam <- ordered(c('Ligero','Medio','Pesado'))`
- Obtención niveles
  - `nlevels(fdias)`
  - `levels(días)`
- Relación de orden (factors ordenados)
  - `tam[2] < tam[1]  # FALSE`

Data Frames

▶tiposDatosII.R

- Definición
  - `df <- data.frame(vect1, …, vectN)`
  - `df <- data.frame(matrix)`
  - `df <- data.frame(col1 = tipo(N), …,`
    `colN = tipo(N)`
- Ejemplo
  - `df <- data.frame(Dia = fdias,`
    `Estimado = rep(c(T, F), 7),`
    `Lectura = rnorm(14, 5))`
- Obtención número de elementos
  - `nrow(mes)`
  - `ncol(mes)`

Data Frames

▶tiposDatosII.R

- Selección y proyección de datos
  - `df[5, ]` # 5ª fila
  - `df[ ,3]` # 3ª columna
  - `df[c(-3,-6), ]` # Menos 3ª y 6ª fila
  - `df$Lectura` # 3ª columna
  - `df$Lectura[5]` # 5ª fila de 3ª col.
  - `df[, c('Dia','Lectura')]` # Columnas 1 y 3
  - `df[df$Estimado == F, ]` # Filas condición
  - ```
    # Selección de filas y columnas
    df[df$Estimado == F & df$Lectura > 3,
       c('Dia', 'Lectura')]
    ```
    ```

Introducción a R

Cómo cargar conjuntos de datos

Cargar datos CSV

▶cargaDatos.R

```
datosCSV <- read.table(
 file     = "miArchivo.csv",
 header   = T,
 sep      = ",",
 dec      = ".",
 quote    = "\"")
```

```
read.csv("miArchivo.csv")  # sep=",", dec="."
```

```
read.csv2("miArchivo.csv") # sep=";", dec=","
```

```
read.delim("miarchivo.txt") # sep = "\t"
```

Cargar datos Excel

▶`cargaDatos.R`

- Múltiples posibilidades
  - Exportar desde Excel a CSV
  - Copiar datos al portapapeles
  - Leer archivo Excel desde R
- Paquetes R para trabajar con archivos Excel
  - XLConnect
    - `datos <- readWorksheetFromFile('archivo.xls', sheet=n)`
  - xlsx
    - `datos <- read.xlsx('archivo.xlsx',sheetName = n, rango)`
- `vignette(paquete)` # Abrir el manual asociado

Cargar datos ARFF (Weka)

▶`cargaDatos.R`

- Paquete `foreign`
  - Funciones para leer múltiples formatos de archivo
  - `read.arff('dataset.arff')`
- Paquete RWeka
  - Interfaz completa entre R y Weka
    - Leer y escribir archivos ARFF
    - Acceso a algoritmos de clasificación, agrupamiento, etc.
  - `read.arff('dataset.arff')`

Datasets integrados

▶cargaDatos.R

- Lista de datasets
  - data()

```
Data sets in package 'datasets':

AirPassengers                Monthly Airline Passenger Numbers 1949-1960
BJsales                      Sales Data with Leading Indicator
BJsales.lead (BJsales)       Sales Data with Leading Indicator
BOD                          Biochemical Oxygen Demand
CO2                          Carbon Dioxide Uptake in Grass Plants
ChickWeight                  Weight versus age of chicks on different diets
DNase                        Elisa assay of DNase
EuStockMarkets               Daily Closing Prices of Major European Stock Indices,
Formaldehyde                 Determination of Formaldehyde
HairEyeColor                 Hair and Eye Color of Statistics Students
Harman23.cor                 Harman Example 2.3
Harman74.cor                 Harman Example 7.4
Indometh                     Pharmacokinetics of Indomethacin
InsectSprays                 Effectiveness of Insect Sprays
JohnsonJohnson               Quarterly Earnings per Johnson & Johnson Share
LakeHuron                    Level of Lake Huron 1875-1972
LifeCycleSavings             Intercountry Life-Cycle Savings Data
Loblolly                     Growth of Loblolly pine trees
Nile                         Flow of the River Nile
Orange                       Growth of Orange Trees
OrchardSprays                Potency of Orchard Sprays
PlantGrowth                  Results from an Experiment on Plant Growth
Puromycin                    Reaction Velocity of an Enzymatic Reaction
Seatbelts                    Road Casualties in Great Britain 1969-84
Theoph                       Pharmacokinetics of Theophylline
```

Introducción a R

Fundamentos de análisis exploratorio de datos

Contenido del dataset

▶analisisExploratorio.R

- Estructura interna de la variable
  - `str(variable)`
- Resumen del contenido
  - `summary(variable)`
- Exploración del contenido
  - `head(variable)  |  tail(variable)`
  - `variable[filas, columnas]`
  - `variable$columna`
  - `variable$columna[which(condición)]`
    - `iris$Sepal.Length[which(iris$Species == 'versicolor')]`

Estadística descriptiva

▶analisisExploratorio.R

- Funciones básicas (operan sobre vectores)
  - `mean` `# media`
  - `median` `# mediana`
  - `var` `# varianza`
  - `sd` `# desviación estándar`
  - `max` `# máximo valor`
  - `min` `# mínimo valor`
  - `range` `# rango de valores`
  - `quantile` `# cuartiles`

- Para estructuras complejas
  - `lapply(iris[ ,1:4], mean)` `# Aplicar a cada columna`

Agrupamiento de datos

▶analisisExploratorio.R

- Tabla de contingencia con número de combinaciones
  - Longitud de sépalo según especie

```
table(iris$Sepal.Length, iris$Species)
```

  - Valoración de vendedores según moneda

```
table(ebay$sellerRating, ebay$currency)
```

- Agrupamiento y selección
  - Separar los casos por especie de flor

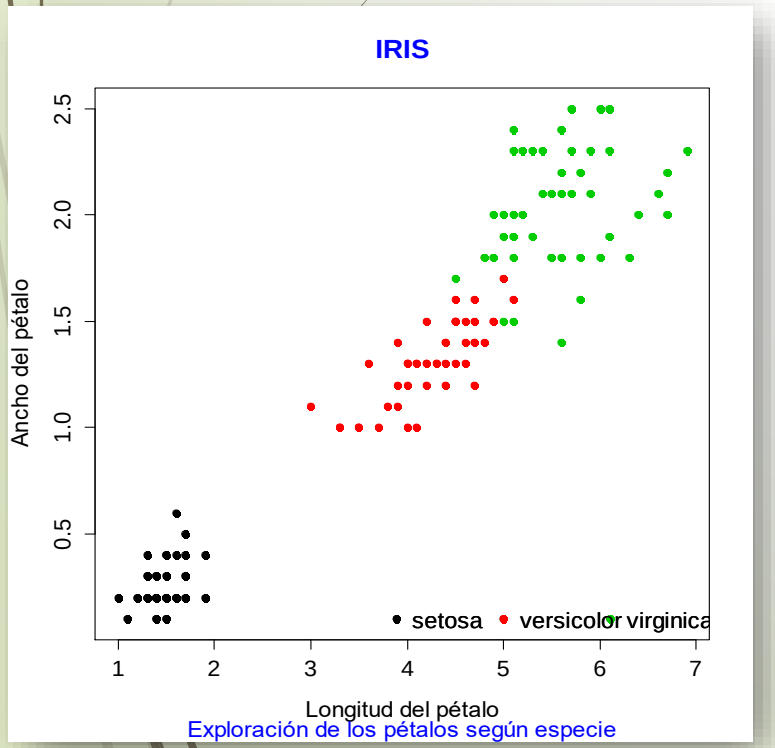
```
split(iris, iris$Species)
```

  - Obtener elevación, pendiente y clase de filas que cumplan condición

```
subset(covertype, slope > 45 & soil_type == '1',
       select = c(elevation, slope, class))
```

Exploración visual - scatterplot

- analisisExploratorio.R



Exploración de los pétalos según especie

```
plot(iris$Petal.Length,
     iris$Petal.Width,
     col = iris$Species,  pch = 19,
     xlab = 'Longitud del pétalo',
     ylab = 'Ancho del pétalo')

title(main = 'IRIS',
      sub = 'Pétalos según specie',
      col.main = 'blue',
      col.sub = 'blue')

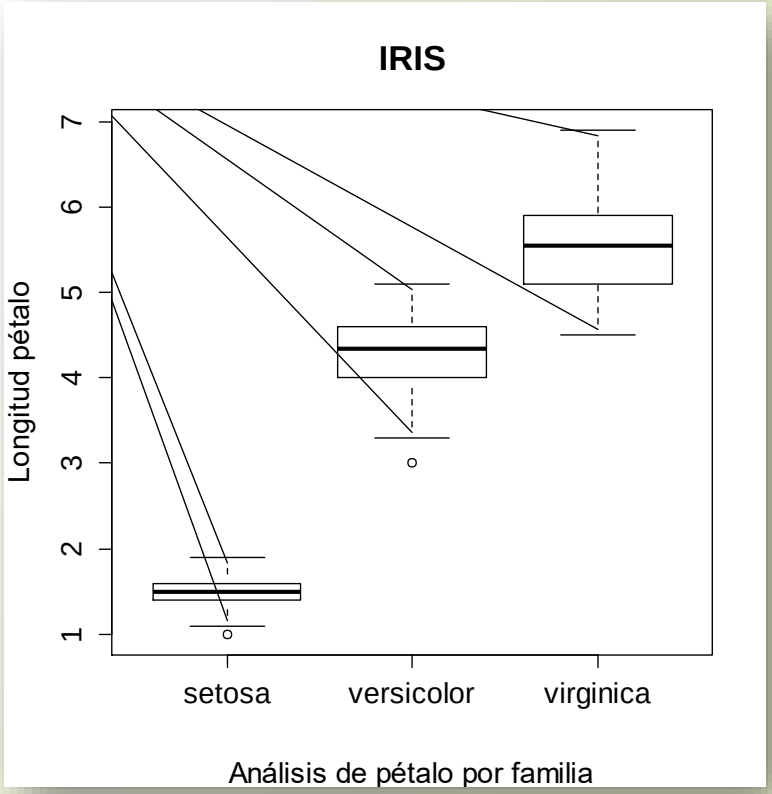
legend("bottomright",
  legend = levels(iris$Species),
  col =  unique(iris$Species),
  ncol = 3, pch = 19, bty = "n")
```

Exploración visual - cajas y bigotes

- analisisExploratorio.R

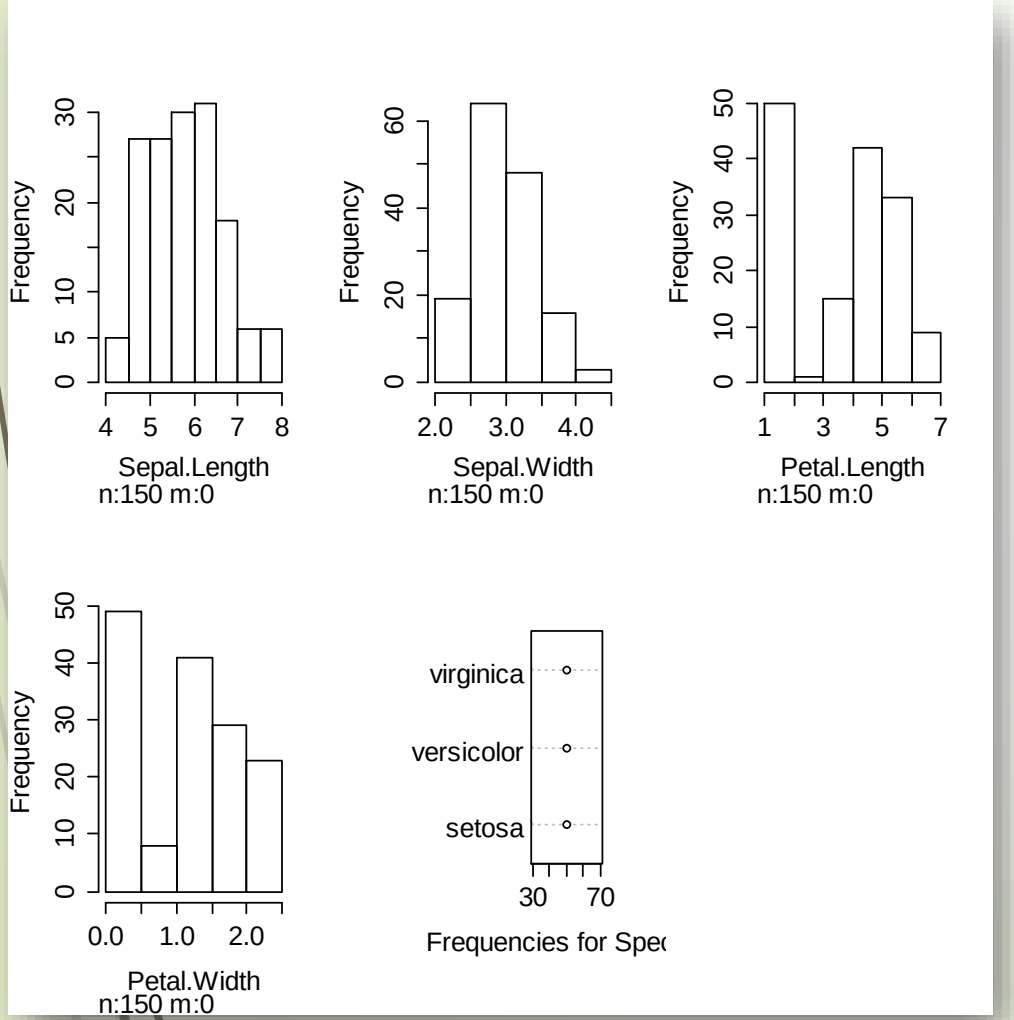
```
boxplot(iris$Petal.Length ~ iris$Species)

title(main = 'IRIS',
  ylab = 'Longitud pétalo',
  sub = 'Análisis de pétalo por familia')
```



Exploración visual - histograma

- `analisisExploratorio.R`

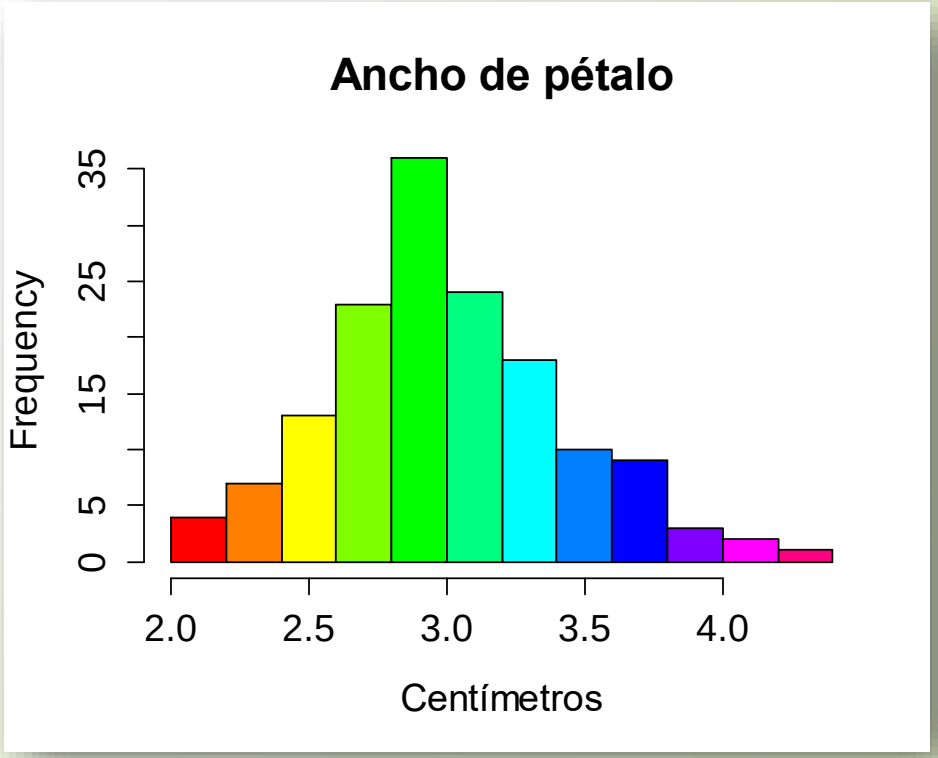



```
hist(iris)
```

Exploración visual - histograma

- analisisExploratorio.R

```
hist(iris$Sepal.Width, breaks = 12,
     col = rainbow(12),
     main = 'Ancho de sépalo',
     xlab = 'Centímetros')
```



Introducción a R y RStudio

Francisco Charte