Gráficos en R

Javier Adanaqué

March 24, 2017

Resumen

Se muestran los gráficos aprendidos en clase usando data salarial en un contexto académico. Se agregan algunas conclusiones básicas basadas en los gráficos con el único objetivo de darle sentido a los gráficos, sin intentar ser exhaustivos al respecto.

Introducción

El objetivo del presente artículo es mostrar los gráficos aprendidos en clase. Para ello se está usando la data de salarios otorgada en clase junto a otros datasets.

Exploraremos los datos gráficamente, mostrando el código usado para llegar a ellos.

Cabe recalcar que de ninguna manera se pretende ser exhaustivo con el análisis o conclusiones, sólo se busca replicar los gráficos y aprender de ellos.

Data

Para trabajar los gráficos, se está usando la data `SALARIOS.txt`, ubicada entre los datasets distribuidos para la clase.

```
salarios <- read.table("data/SALARIOS.txt", header = T)
head(salarios)
```

rank	discipline	yrs.since.phd	yrs.service	sex	salary
Prof	B	19	18	Male	139750
Prof	B	20	16	Male	173200
AsstProf	B	4	3	Male	79750
Prof	B	45	39	Male	115000
Prof	B	40	41	Male	141500
AssocProf	B	6	6	Male	97000

Como se puede observar en las primeras seis observaciones, nos muestra información salarial en un ámbito académico.

Descripción de variables:

- `rank`: Variable categórica con niveles: AssocProf (Associate Professors), AsstProf (Assistant Professors) and Prof (Professors).
- `discipline`: Variable categórica con niveles: A ("theoretical" departments) o B ("applied" departments).
- `yrs.since.phd` años desde que obtuvo el PhD.
- `yrs.service` años de servicio.
- `sex`: Variable categórica con niveles: Female o Male.
- `salary` salario de nueve meses, en dólares.

Distribución de Datos Univariados

Para entender nuestra data, uno de los primeros pasos suele ser explorar cada una de nuestras variables, de manera univariada. Para ello tenemos varias opciones; sin embargo, acá nos concentraremos en las siguientes:

- Histogramas
- Boxplots
- Beanplot
- Vioplot

Histogramas

Los histogramas agrupan una variable numérica en intervalos y nos devuelve la cantidad de observaciones en cada intervalo, representando cada cantidad en la altura de las barras:

```
# R Nativo
hist(salarios$salary,
     main = "Salarios de Profesores Universitarios en USA",
     xlab = "Salario en USD", ylab = "Frecuencia")
```

```
# Lattice
library(lattice)
histogram(~salary, data = salarios, type = "count",
          main = "Salarios de Profesores Universitarios en USA",
          xlab = "Salarios en USD", ylab = "Frecuencia")
```

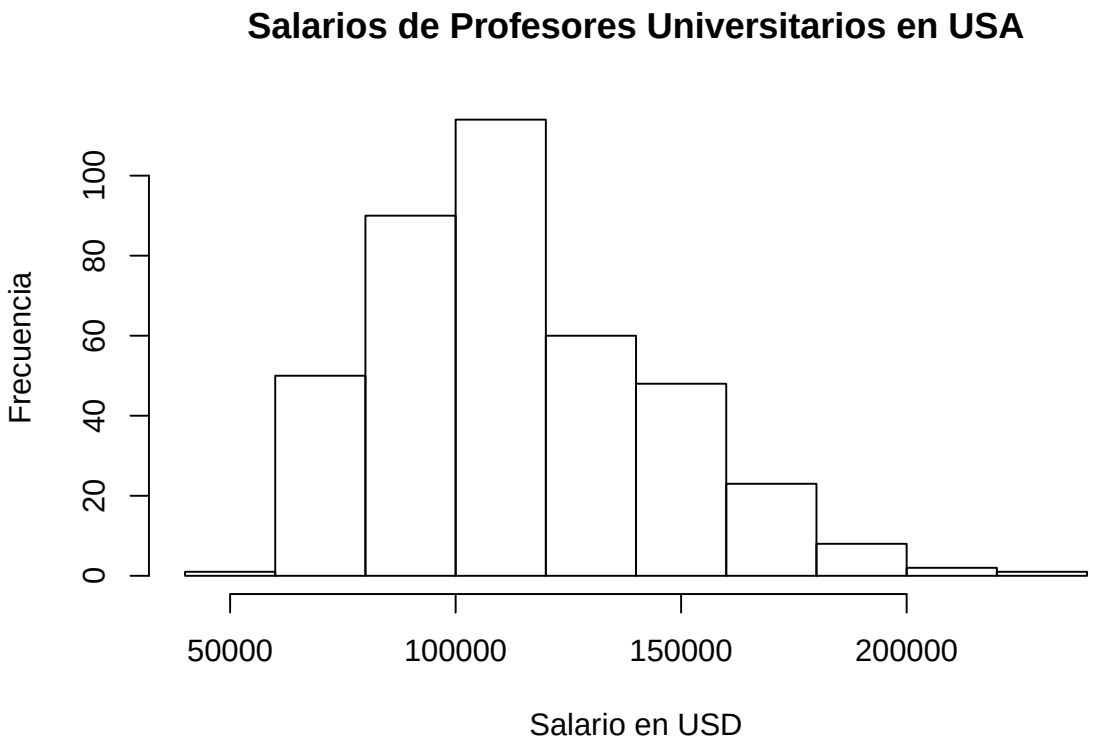


Figura 1: Histograma de salarios usando R nativo

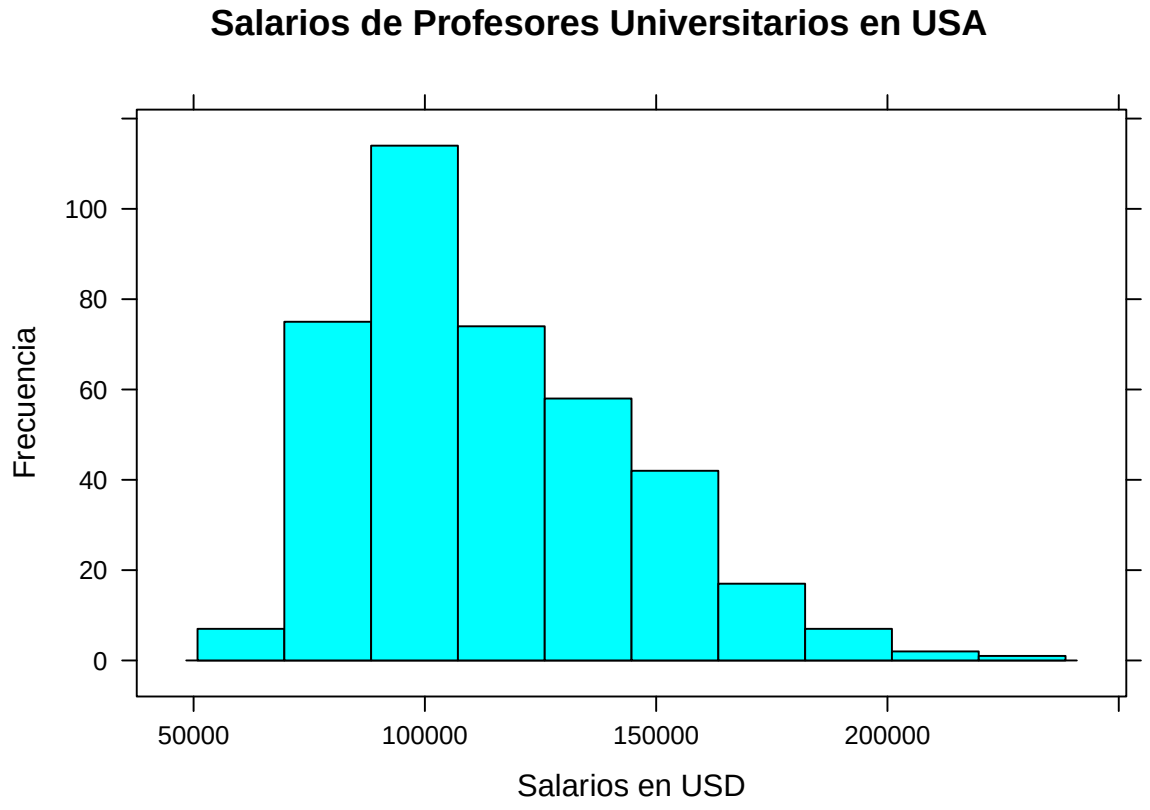


Figura 2: Histograma de salarios usando el paquete Lattice

Se puede observar que la distribución es positivamente asimétrica, resultado común casi siempre que trabajamos con data salarial. Esto se debe a que siempre existen unos pocos que ganan mucho más que la mayoría. En este caso en particular, la mayoría se encuentra alrededor de 107300 (la mediana).

A veces es útil observar la distribución condicionada a alguna otra variable categórica. Observemos cómo se distribuyen los salarios condicionados al rango del profesor/investigador:

```
histogram(~salary | rank, data = salarios, type = "count",
          main = "Salarios de Profesores Universitarios en USA",
          xlab = "Salarios en USD", ylab = "Frecuencia")
```

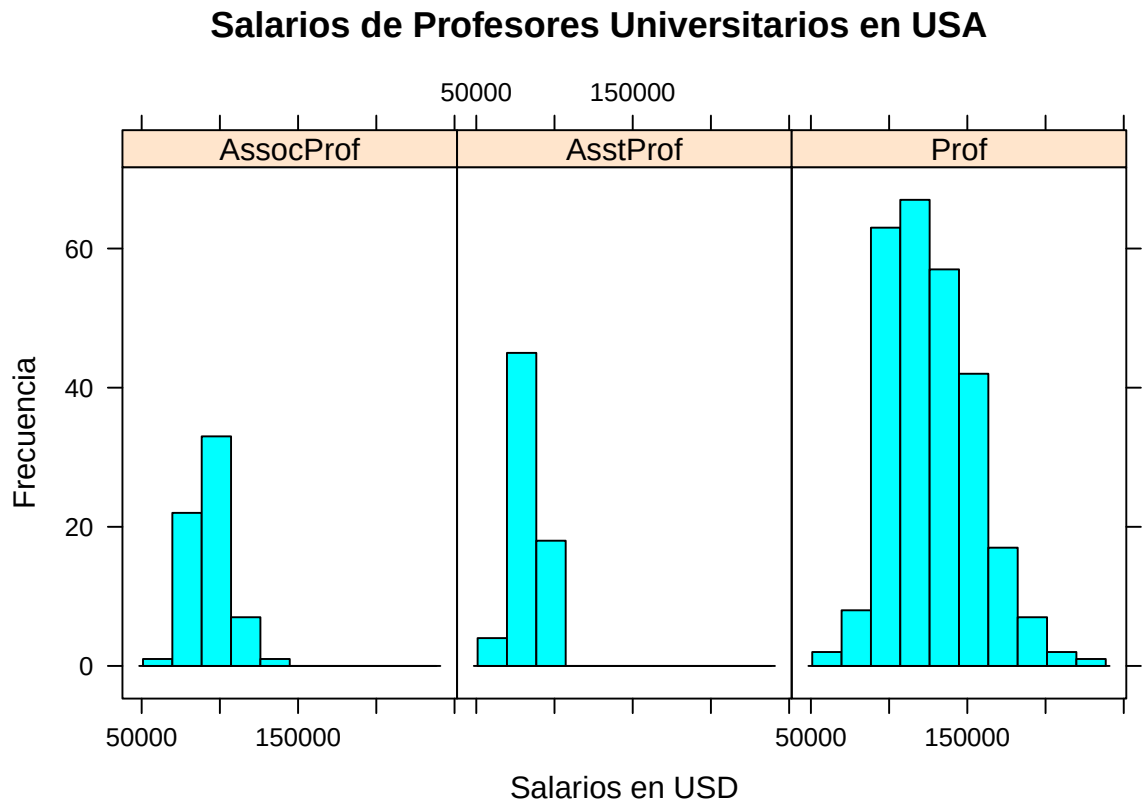


Figura 3: Histograma de salarios según rango

Podemos observar claramente 2 cosas:

- La mayoría de la data se encuentra bajo el rango "Prof" (Professor).

rank	Frecuencia
AssocProf	64
AsstProf	67
Prof	266

- Aquellos que se encuentra bajo el rango "Prof" pueden llegar a ganar mucho más que aquellos que se encuentran en otro rango.

Podríamos seguir analizando más esta distribución, condicionándola a otras variables; sin

embargo, lo dejaremos ahí debido a que ya quedó clara la utilidad de hacer este tipo de gráficos condicionados.

Boxplots

Los Boxplots o, en español, Gráficos de Cajas son gráficos muy útiles que nos muestran la distribución de los datos, pero sin agruparla. Nos brinda, de manera concisa, 5 características de la distribución de interés:

- 1er. cuartil
- Mediana
- 3er. cuartil
- Rango intercuartil (IQR)
- Outliers

```
par(mar = c(0.5, 4, 4, 1))
boxplot(salarios$salary,
        main = "Salarios de Profesores Universitarios en USA",
        xlab = NULL, ylab = "Frecuencia")
```

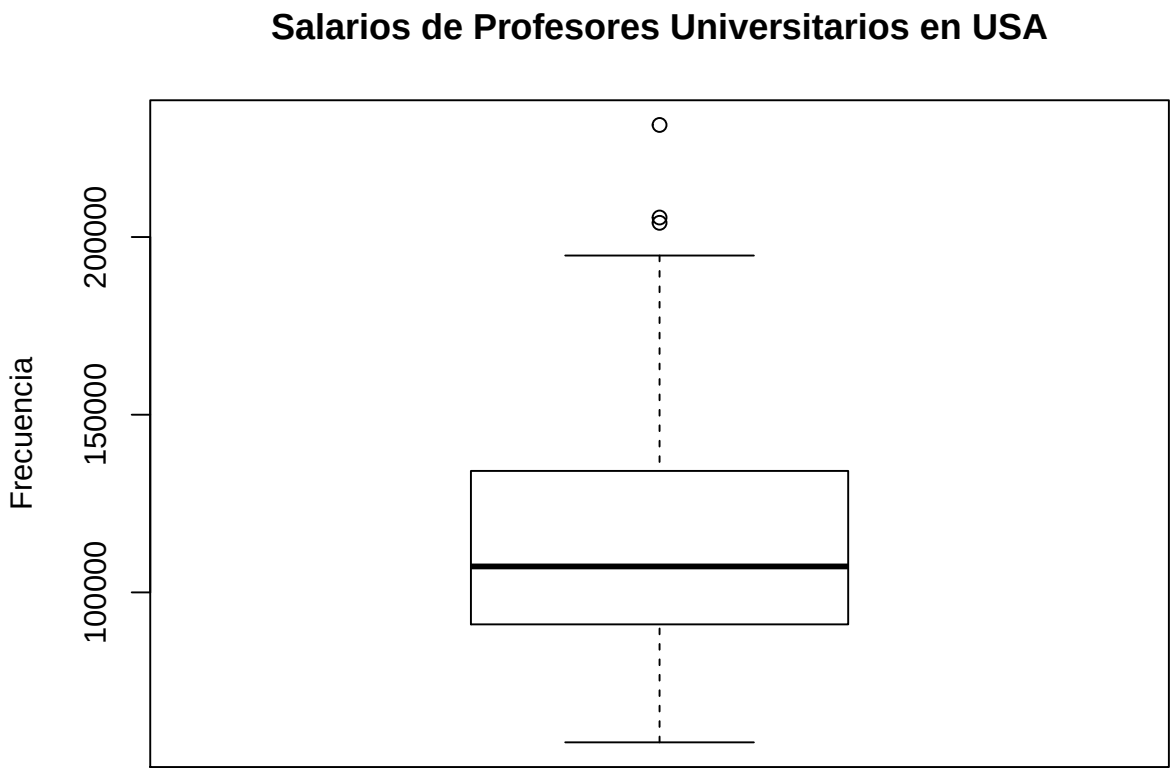


Figura 4: Boxplot con R nativo

```
par(mar = c(5.1, 4.1, 4.1, 2.1))
```

Al igual que con los histogramas, condicionemos nuestros boxplot a una variable categórica. En este caso usaremos una variable distinta, sexo.

```
bwplot(salary~sex, data = salarios,
       main = "Salarios de Profesores Universitarios en USA",
       ylab = "Salarios en USD")
```

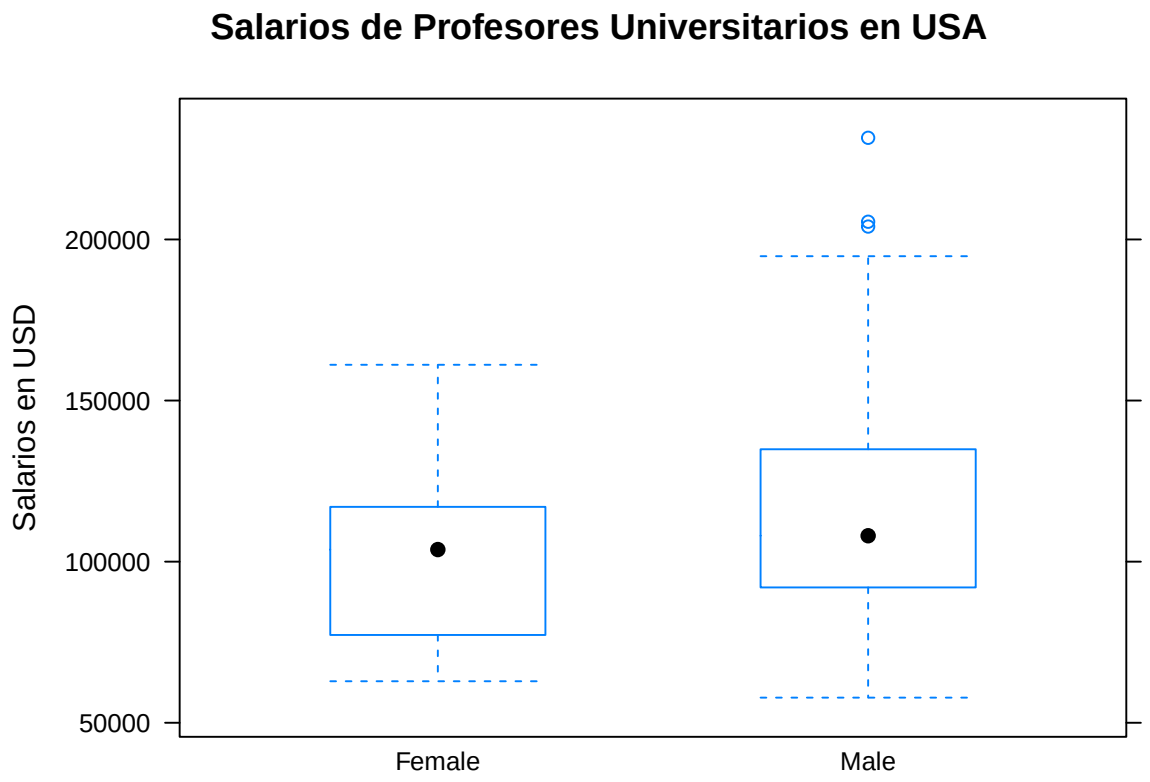


Figura 5: Boxplot de salarios según el sexo del profesor, usando Lattice

Ahora consideremos una 3ra. variable, `discipline`, cuyos niveles son `A` (departamentos "teóricos") y `B` (departamentos "aplicados").

```
bwplot(salary~sex | discipline, data = salarios,
       main = "Salarios de Profesores Universitarios en USA",
       ylab = "Salarios en USD")
```

Claramente, las profesoras (sexo femenino) suelen ganar menos que los profesores (masculino) y las distribuciones son similares entre los departamentos "teóricos" y "aplicados", aunque los "aplicados" (`B`) parecen ganar más.

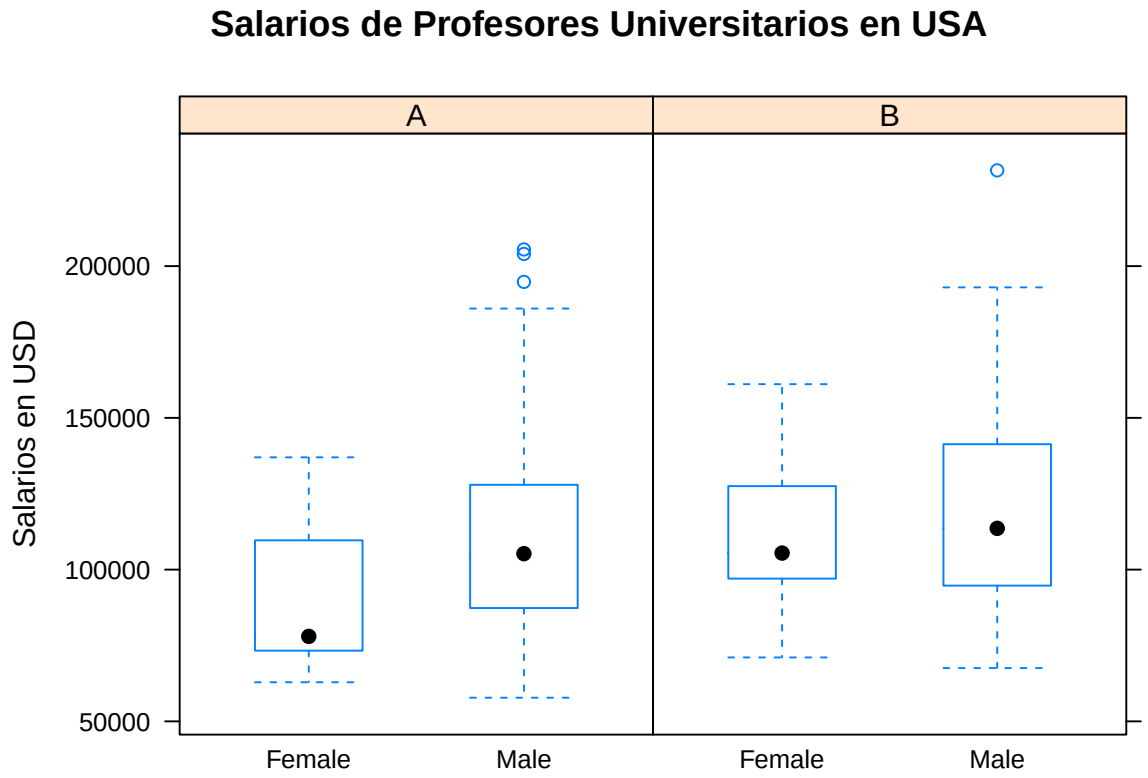


Figura 6: Boxplot de salarios según el sexo y disciplina del profesor, usando Lattice

Beanplot

Los Beanplots son muy similares a los Boxplots, con la excepción de que en vez de graficarte una caja al centro (IQR), te grafican la densidad a lo largo de la distribución. Veamos un ejemplo:

```
par(mfrow = c(1, 2))
library(beanplot)
boxplot(salary~sex, data = salarios, ylim = c(50000, 250000))
beanplot(salary~sex, data = salarios, ylim = c(50000, 250000),
         beanlines = "median", overallline = "median", log = "",
         col = c("black", "white", "black", "red"))
```

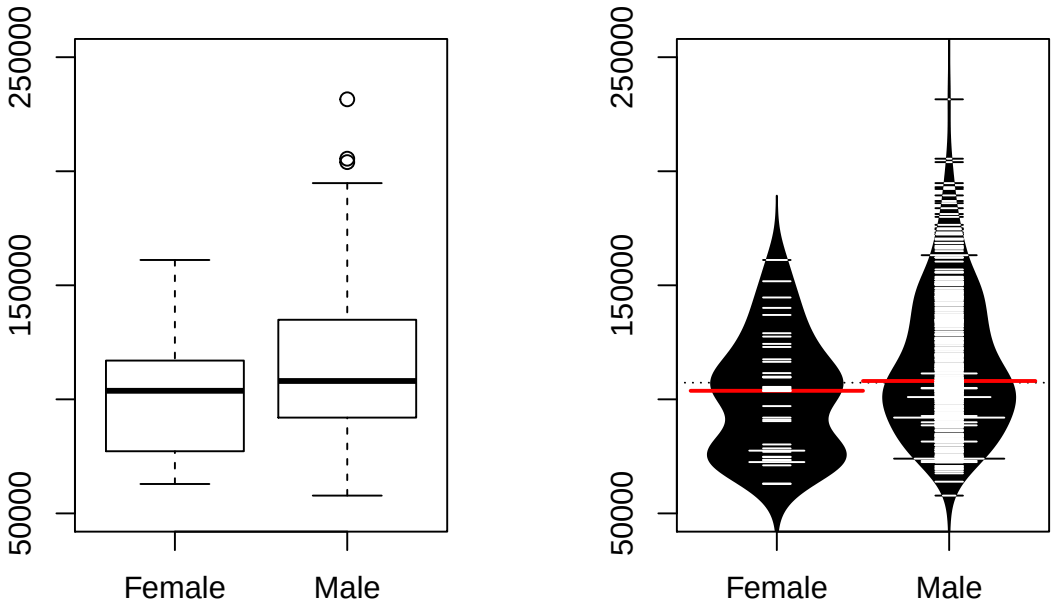


Figura 7: Boxplot (izquierda) y Beanplot (derecha) de salarios según el sexo del profesor.

Vioplot

Los Violin Plots son muy similares a los Beanplots, pero no nos muestran las valores observados de manera individual.

```
library(vioplot)
par(mfrow = c(1, 2))
beanplot(salary~sex, data = salarios, ylim = c(50000, 250000),
         beanlines = "median", overallline = "median", log = "",
         col = c("black", "white", "black", "red"))
vioplot(salarios$salary[salarios$sex == "Female"],
        salarios$salary[salarios$sex == "Male"],
        col = "lightgray",
        ylim = c(50000, 250000),
        names = c("Female", "Male"))
```

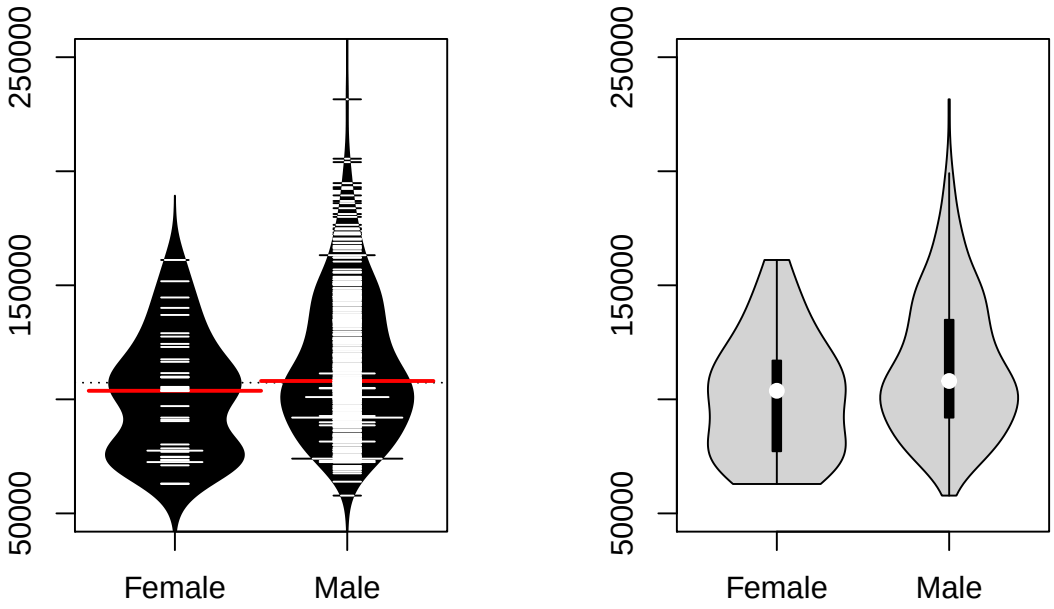


Figura 8: Beanplot (izquierda) y Violin Plot (derecha) de salarios según el sexo del profesor.

Análisis de Datos Multivariantes

Diagramas de Dispersión

Los Diagramas de Dispersión nos permiten observar la relación entre dos variables cuantitativas, graficando cada variable en un eje diferente de nuestro gráfico.

```
plot(salarios$yrs.since.phd, salarios$salary, pch=19,
     xlab = "Años desde el PhD.",
     ylab = "Salario")
```

También podemos hacer el mismo análisis, condicionando la relación a diferentes variables categóricas, e.g., `sexo`:

```
xyplot(salary ~ yrs.since.phd | sex, data = salarios)
```

Observamos claramente, de nuevo, que existen pocas observaciones de mujeres. Sin embargo, en ambos casos observamos una relación positiva entre el salario y los años desde que se obtuvo el PhD., aunque débil debido a la gran dispersión en las observaciones.

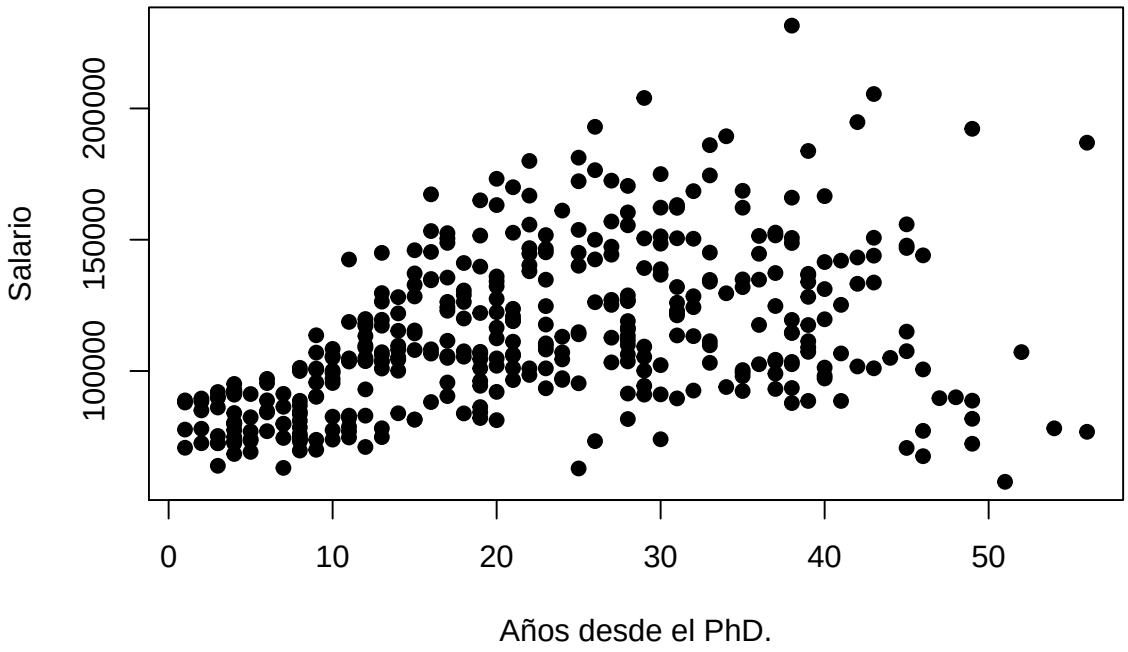


Figura 9: Diagrama de Dispersión simple usando R Nativo

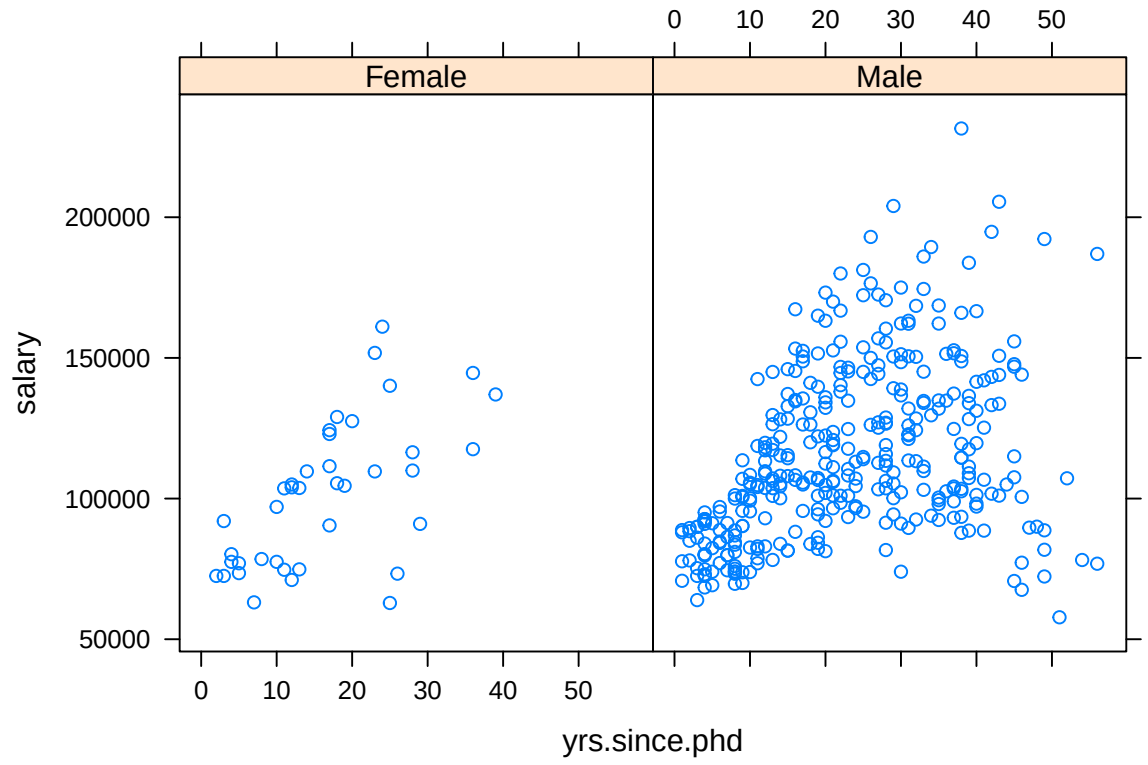


Figura 10: Diagrama de Dispersión según sexo usando Lattice

Gráficos de Correlación.

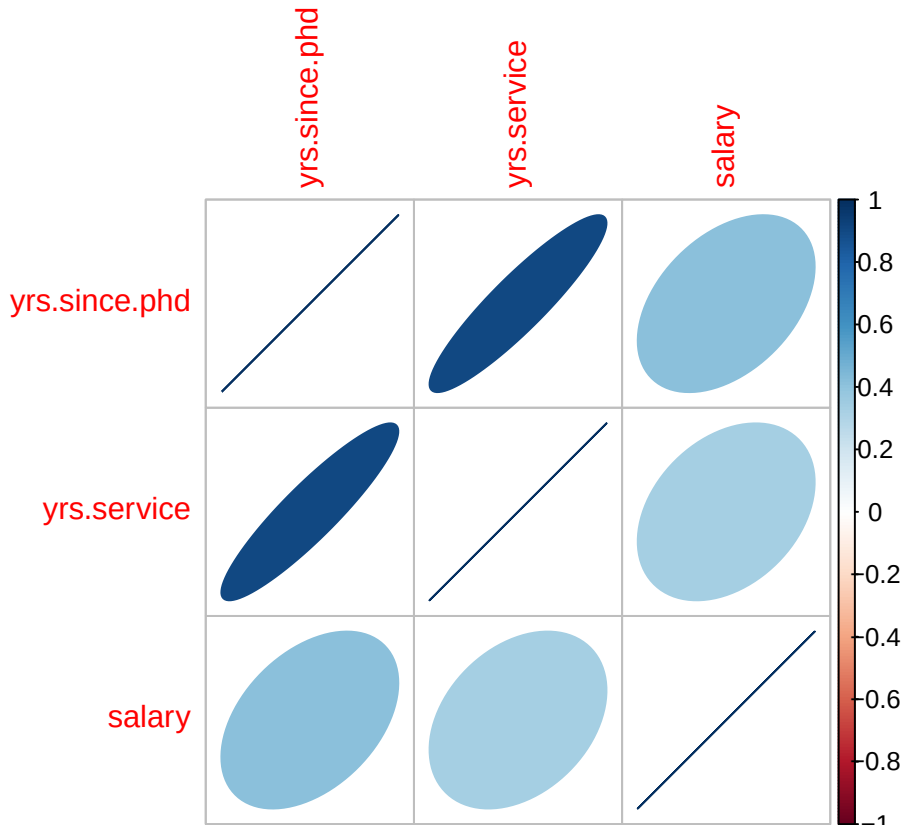
Las matrices de correlación son útiles, pero cuando tenemos muchas variables a veces es útil visualizarlas, y para ello tenemos los gráficos de correlación (`corrplot`).

Lamentablemente, nuestra data de salarios sólo tiene 3 variables cuantitativas, así que usaremos esas para nuestro gráfico de correlación:

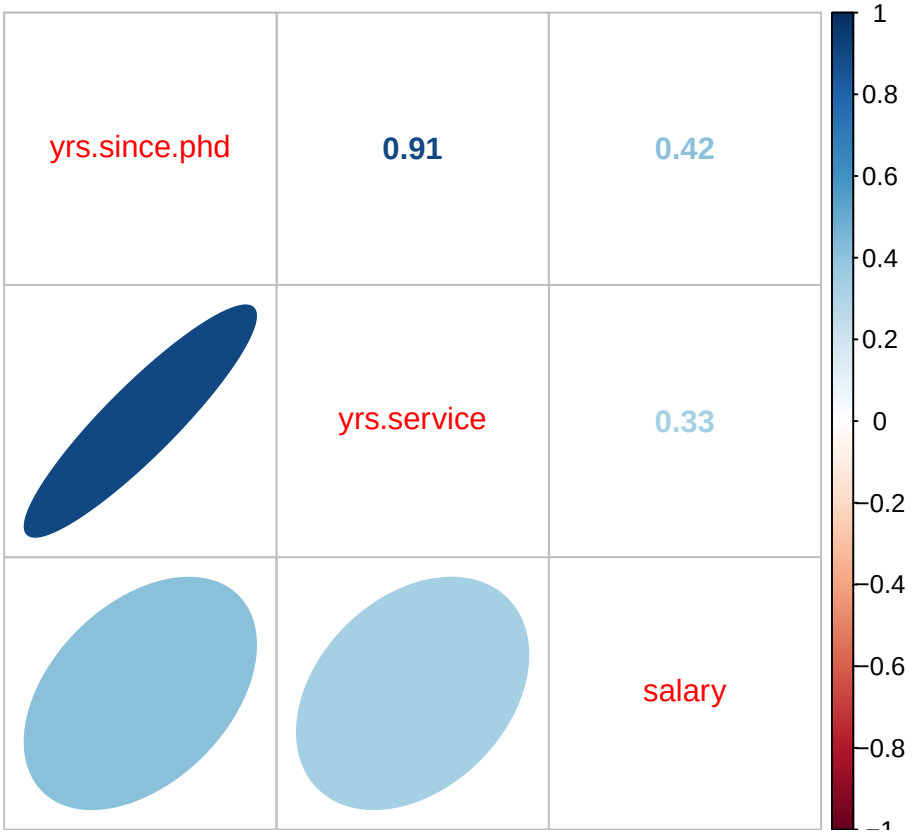
```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.3.3
```

```
cor_salarios <- cor(salarios[, c("yrs.since.phd", "yrs.service", "salary")])
corrplot(cor_salarios, method = "ellipse")
```




```
corrplot.mixed(cor_salarios, lower = "ellipse", upper = "number")
```



Conclusión

El uso de gráficos es muy útil para análisis exploratorio. A nosotros nos sirvió bastante para sacar algunas conclusiones preliminares que luego podríamos analizar más a fondo usando diferentes herramientas/técnicas estadísticas.