R meets psychology

Sesión 1. Introducción al ggplot2

El ejercicio desarrollado en este material se ha preparado haciendo uso de la base de datos “Familia y roles de género 2012”, del IOP, que se encuentra para descargar en el siguiente enlace.

<http://iop-data.pucp.edu.pe/busqueda/encuesta/71>?

El IOP Data tiene varias bases de datos que puedes revisar, descargar, y desarrollar con esta plantilla.

El paquete ggplot2

* Con el R es posible obtener el mismo resultado usando diferentes "caminos"
* El paquete ggplot es uno de los entornos gráficos del R
* Permite elaborar un gráfico a partir de un proceso de acumulación de ***capas*** o ***layers***.
* Tiene un cierto nivel de complejidad pero se obtienen resultados muy profesionales.

Referencias bibliográficas

Textos disponibles en la biblioteca de CCSS de la PUCP para el uso de ggplot2

* Chang, Winston. 2012. *R Graphics Cookbook*. Sebastopol, CA: O’Reilly Media.
* Field, Andy P. 2012. *Discovering statistics using R*. London; Thousand Oaks, Calif: Sage.
* Wickham, Hadley. 2009. *Ggplot2: elegant graphics for data analysis*. New York: Springer.

Capas o layers en ggplot2

Un gráfico en ggplot2 puede tener varias capas, en su conjunto, las capas forman el gráfico al combinar:

* Un data frame y las variables a ser graficadas
* Una o varias capas indicando, entre otros:
  + El tipo de objeto a graficar o "geom" (barra, línea, punto, etc.)
  + Las transformaciones estadísticas a los datos
  + La posición de los objetos en el gráfico
* Una escala para cada variable a ser graficada
* Un sistema de coordenadas
* La especificación de "facetas" del gráfico

Cargamos los datos de trabajo

Descomprimir y grabar el archivo SPSS en el directorio de trabajo de R

# Importar la base de datos del SPSS a un data frame de R  
library(foreign)  
genero <- as.data.frame(read.spss("IOP\_1212\_01\_B.sav"))

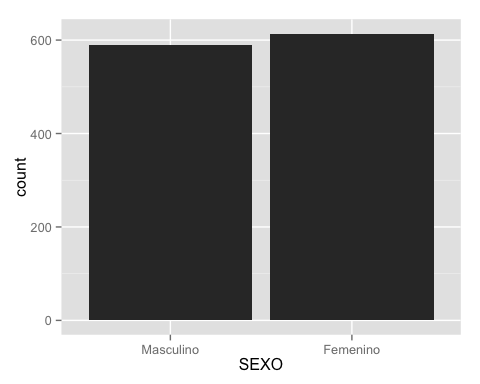
## re-encoding from UTF-8

Un gráfico simple: Gráfico de barras

El esquema básico de la gramática de ggplot es:

ggplot(data.frame, aes(x = variable)) + geom\_*forma*()

library(ggplot2)  
ggplot(genero, aes(x = SEXO)) + geom\_bar()

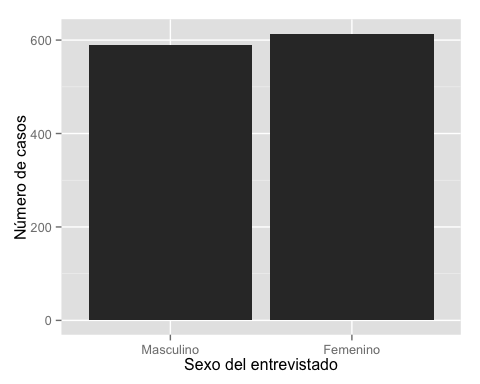


Podemos guardar el gráfico en un objeto y añadir más capas:

gr1 <- ggplot(genero, aes(x = SEXO)) + geom\_bar()  
gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos")

#Se utiliza el comando xlab para añadir la etiqueta del eje X

#Se utiliza el comando Ylab para añadir la etiqueta del eje Y



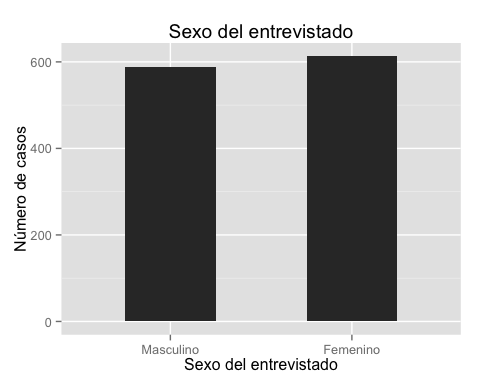
gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos") +   
 ggtitle("Sexo del entrevistado")

#Se utiliza el comando main para añadir un titulo



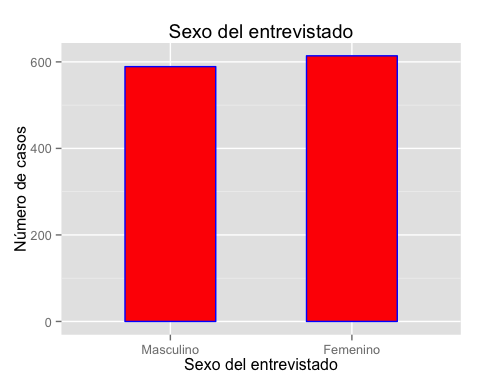
Podemos cambiar cómo se ven algunos elementos, por ejemplo, que las barras sean más pequeñas, en este caso la mitad (0.5) del tamaño por defecto:

gr1 <- ggplot(genero, aes(x = SEXO)) + geom\_bar(width=0.5)  
gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos") +   
 ggtitle("Sexo del entrevistado")



Con la opción "colour" y "fill" en el comando **geom\_bar** podemos cambiar el color del controno de las barras y de su relleno

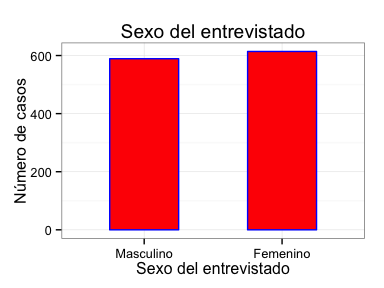
gr1 <- ggplot(genero, aes(x = SEXO)) + geom\_bar(width=0.5, colour="blue", fill="red")  
gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos") +   
 ggtitle("Sexo del entrevistado")



Temas

Los temas (theme) son un conjunto de opciones predefinidas sobre la apariencia de los objetos en ggplot. El tema por defecto del ggplot dibuja el gráfico sobre un fondo gris. Podemos cambiarlo a blanco y negro añadiendo el comando theme\_bw()

gr1 <- ggplot(genero, aes(x = SEXO)) + geom\_bar(width=0.5, colour="blue", fill="red")  
gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos") +   
 ggtitle("Sexo del entrevistado") + theme\_bw()



Complicando un poco más el asunto...

Hacer un gráfico de barras de sexo, pero con porcentajes en el eje vertical. Primero preparar los datos:

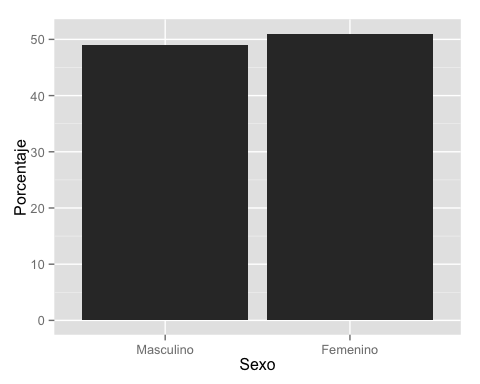
tab.sex1 <- as.data.frame(prop.table(table(genero$SEXO))\*100)  
tab.sex1

## Var1 Freq  
## 1 Masculino 48.96  
## 2 Femenino 51.04

colnames(tab.sex1) <- c("Sexo", "Porcentaje")

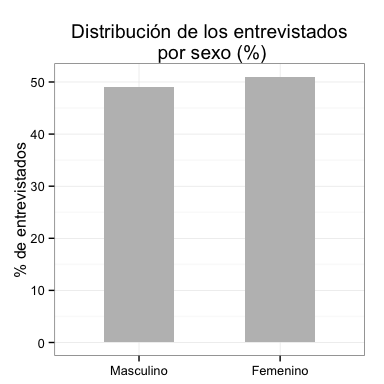
Ahora el gráfico:

ggplot(tab.sex1, aes(x=Sexo, y=Porcentaje)) + geom\_bar(stat="identity")



Afinamos el gráfico:

gr2 <- ggplot(tab.sex1, aes(x=Sexo, y = Porcentaje)) +   
 geom\_bar(stat="identity", width=0.5, fill = "grey")   
  
gr2 + xlab(NULL) + ylab("% de entrevistados") +   
 ggtitle("Distribución de los entrevistados\n por sexo (%)") + theme\_bw()



Otra manera de llegar al mismo resultado

library(scales) # requiere instalar el paquete "scales"  
gr2.a <- ggplot(genero, aes(SEXO)) +   
 geom\_bar(aes(SEXO, (..count..)/sum(..count..)), width=0.5, fill = "grey")   
  
gr2.a + scale\_y\_continuous(labels=percent) + xlab(NULL) +   
 ylab("% de casos") + ggtitle("Distribución de los\n entrevistados por sexo") + theme\_bw()

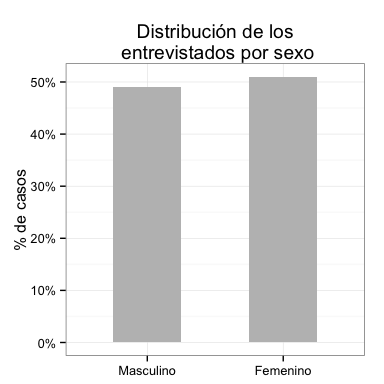
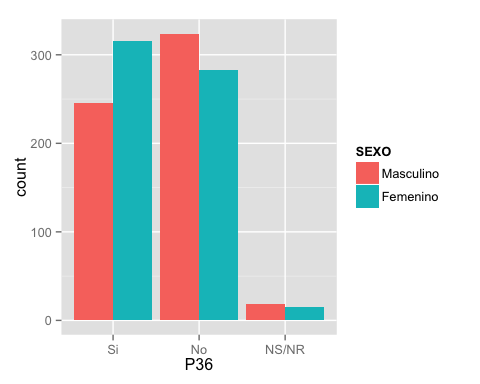


Gráfico de barras múltiples

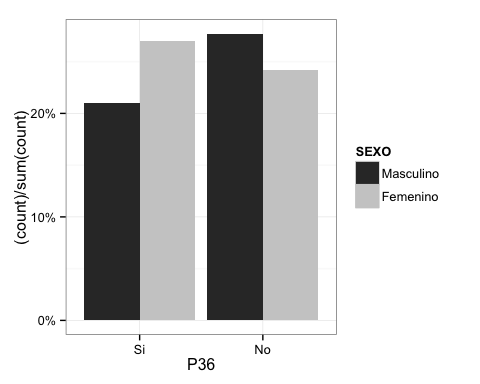
La distribución de frecuencias de la pregunta P36 según sexo

ggplot(genero, aes(P36, fill=SEXO)) + geom\_bar(position="dodge")



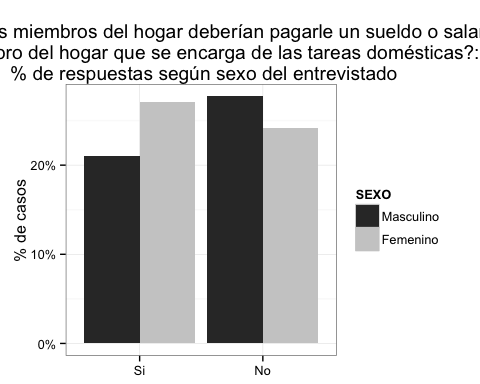
¿Cómo quitamos "NS/NR" y establecemos que la escala del eje vertical sean %?

gr3 <- ggplot(genero[genero$P36!="NS/NR", ], (aes(P36, fill=SEXO))) +   
geom\_bar(aes(P36, (..count..)/sum(..count..)), position="dodge") +   
scale\_y\_continuous(labels=percent) + scale\_fill\_grey() + theme\_bw()  
  
gr3



Completamos el gráfico

gr3 + xlab(NULL) + ylab("% de casos") + ggtitle("¿Los demás miembros del hogar deberían pagarle un sueldo o salario\n al miembro del hogar que se encarga de las tareas domésticas?:\n % de respuestas según sexo del entrevistado")



Usando facets

Las facetas o "facets" en ggplot me permiten reproducir el mismo gráfico en diferentes niveles de un factor. Hagamos un gráfico de la distribución en % de la pregunta P51D, para los diferentes dominios geográficos

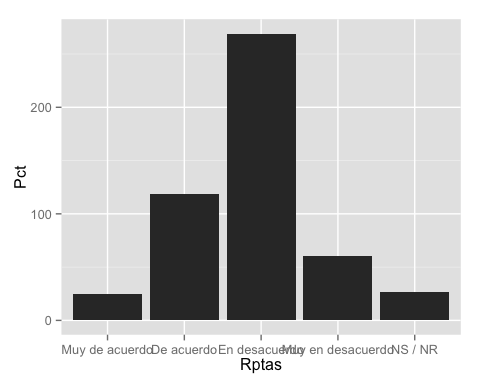
# Primero preparamos los datos en una tabla que convertimos en data frame  
tab.5 <- as.data.frame(prop.table(table(genero$P51D, genero$DOMINIO), 2)\*100)  
tab.5

## Var1 Var2 Freq  
## 1 Muy de acuerdo Lima-Callao 0.8929  
## 2 De acuerdo Lima-Callao 18.7500  
## 3 En desacuerdo Lima-Callao 62.0536  
## 4 Muy en desacuerdo Lima-Callao 13.1696  
## 5 NS / NR Lima-Callao 5.1339  
## 6 Muy de acuerdo Norte 3.4375  
## 7 De acuerdo Norte 25.9375  
## 8 En desacuerdo Norte 55.3125  
## 9 Muy en desacuerdo Norte 10.0000  
## 10 NS / NR Norte 5.3125  
## 11 Muy de acuerdo Sur 7.3469  
## 12 De acuerdo Sur 31.0204  
## 13 En desacuerdo Sur 43.6735  
## 14 Muy en desacuerdo Sur 13.0612  
## 15 NS / NR Sur 4.8980  
## 16 Muy de acuerdo Centro 13.3333  
## 17 De acuerdo Centro 22.8571  
## 18 En desacuerdo Centro 35.2381  
## 19 Muy en desacuerdo Centro 20.9524  
## 20 NS / NR Centro 7.6190  
## 21 Muy de acuerdo Oriente 0.0000  
## 22 De acuerdo Oriente 20.0000  
## 23 En desacuerdo Oriente 72.9412  
## 24 Muy en desacuerdo Oriente 3.5294  
## 25 NS / NR Oriente 3.5294

colnames(tab.5) <- c("Rptas", "Dominio", "Pct")

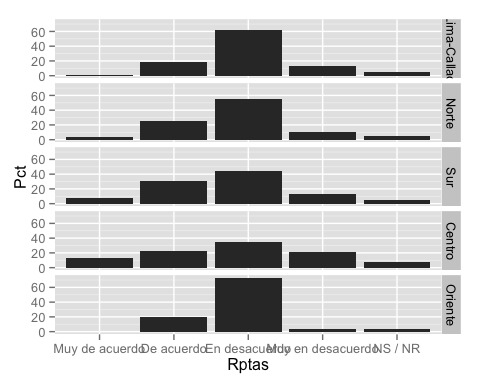
Este sería en gráfico de base:

gr4 <- ggplot(tab.5, aes(x=Rptas, y=Pct)) + geom\_bar(stat="identity")   
gr4



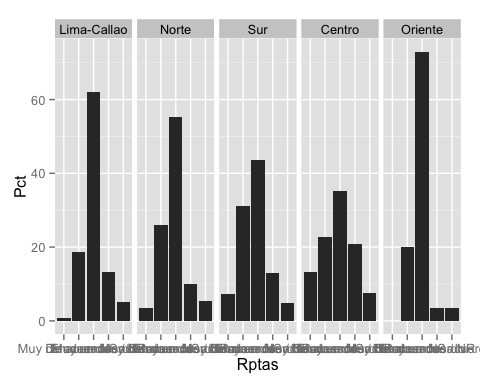
Hacemos las facetas verticales

gr4 + facet\_grid(Dominio ~.)



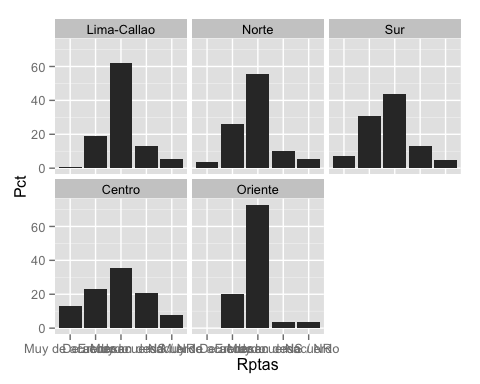
Si las queremos horizontales:

gr4 + facet\_grid(.~ Dominio)



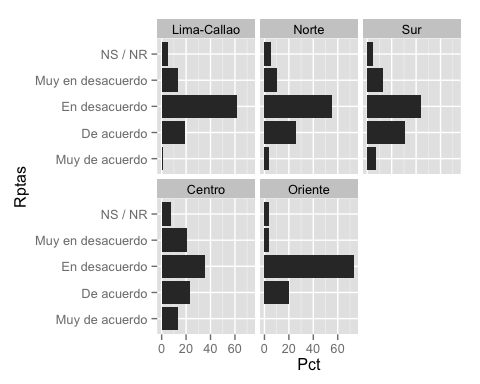
Si queremos que roten a lo largo de columnas y filas:

gr4 + facet\_wrap(~ Dominio)



Podemos rotar el gráfico para que se aprecien mejor las etiquetas de respuestas

gr4 + coord\_flip() + facet\_wrap(~ Dominio)

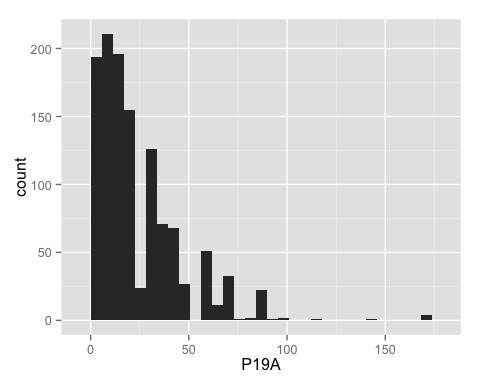


Histogramas

Histograma básico de la pregunta P19A

ggplot(genero, aes(P19A)) + geom\_histogram()

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



Amplitud de los intervalos del histograma

range(genero$P19A, na.rm=TRUE)

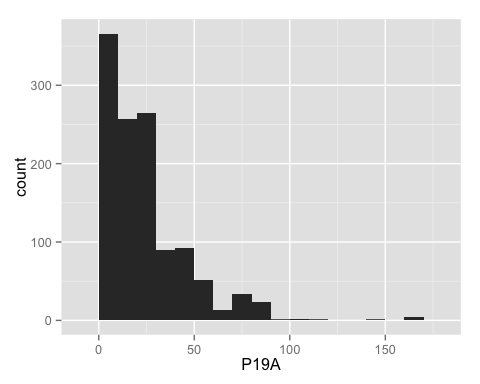
## [1] 0 168

168/30

## [1] 5.6

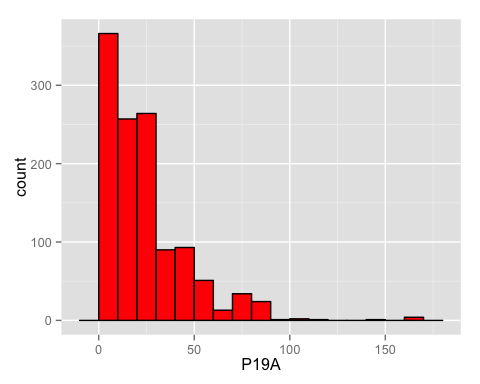
Cambiamos la amplitud del intervalo

ggplot(genero, aes(P19A)) + geom\_histogram(binwidth = 10)



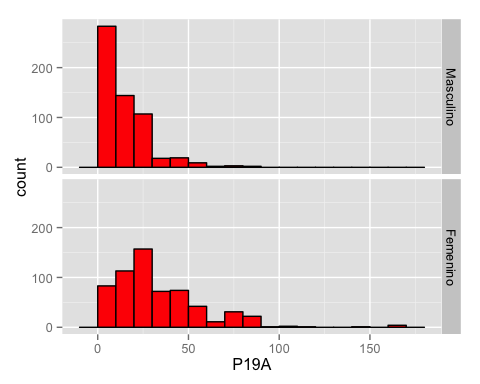
Algunas mejoras:

hist1 <- ggplot(genero, aes(P19A)) + geom\_histogram(binwidth = 10,   
 fill="red", colour="black")  
hist1



Podemos comparar el histograma de los hombres y de las mujeres

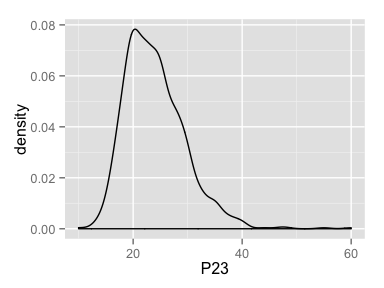
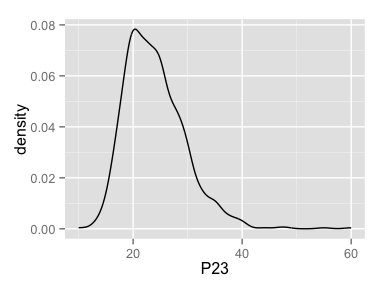
hist1 + facet\_grid(SEXO ~.)



Curvas de Densidad de Kernell

Otra forma de ver la distribución de una variable cuantitativa, sobre la base del cálculo de densidades a partir del histograma:

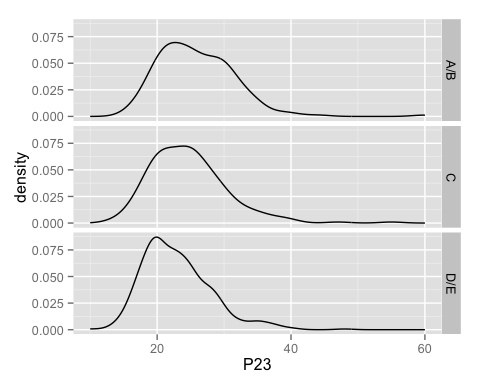
## Las sintaxis nos dan el mismo resultado. La segunda evita dibujar la línea de abajo  
ggplot(genero, aes(P23)) + geom\_density()  
ggplot(genero, aes(P23)) + geom\_line(stat="density")

Curvas de densidad para grupos diferentes

Distribución de la edad en la que se casó según NSE

ggplot(genero, aes(P23)) + geom\_line(stat="density") + facet\_grid(NSEGrup ~.)



Distribución de la edad en la que se casó según NSE y Sexo

ggplot(genero, aes(P23)) + geom\_line(stat="density") + facet\_grid(NSEGrup ~ SEXO)

