Visualización con R

Miguel Angel Escalante Serrato

Octubre 15, 2020

Contents

Variables Categóricas

A comparación de las variables continuas, las variables categóricas tienden a estar más limitadas en el rango de gráficas a utilizar, las opciones más utilizadas con las categorías son las gráficas de barras y los pie charts.

A veces se sugiere que se pongan puntos con cierto ruido alrededor de los valores, sin embargo no necesariamente se ayuda a cuando se tienen muchos datos, aunque puede ser que se pueda ver algo. Sin embargo tampoco es tan recomendado.

En las varibles categóricas usualmente no hay tanta información pero es importante verlos para seguir adelante dentro de análisis más complejos, hay veces que información sencilla y concreta de variables categóricas se pierde en gráficos más complejos.

Características

- Patrones inesperados de resultados: Encontrar categorías con más observaciones de las esperadas.

- Distribuciones desequilibradas: Podría ser que haya un sesgo en la toma de las observaciones, sólo un sexo, sólo un grupo de edad. Distintas características que pueden ser dominantes en el estudio.
- Categorías extras: En el estudio se tienen categorías como “M” y “F”, pero dentro de su base observan “male”, “female”, “m”,“f”, “ma”, “fe”, etc.
- Experimentos desbalanceados: Existe la posibilidad que en el experimento no se tenga un área de valores posibles dentro de la muestra. Es importante conocer las características de la información que se tiene para poderla tomar en cuenta dentro del modelado.
- Muchas categorías: Esto es especialmente problemático cuando se tienen preguntas abiertas: ¿Quién es su artista o grupo preferido?
- No sabe, no contestó, errores, vacíos: En alguna encuesta que trabajé se tenían las siguientes características:
  - No contestó
  - No contestó (espontáneo)
  - No quiso contestar.
  - NA

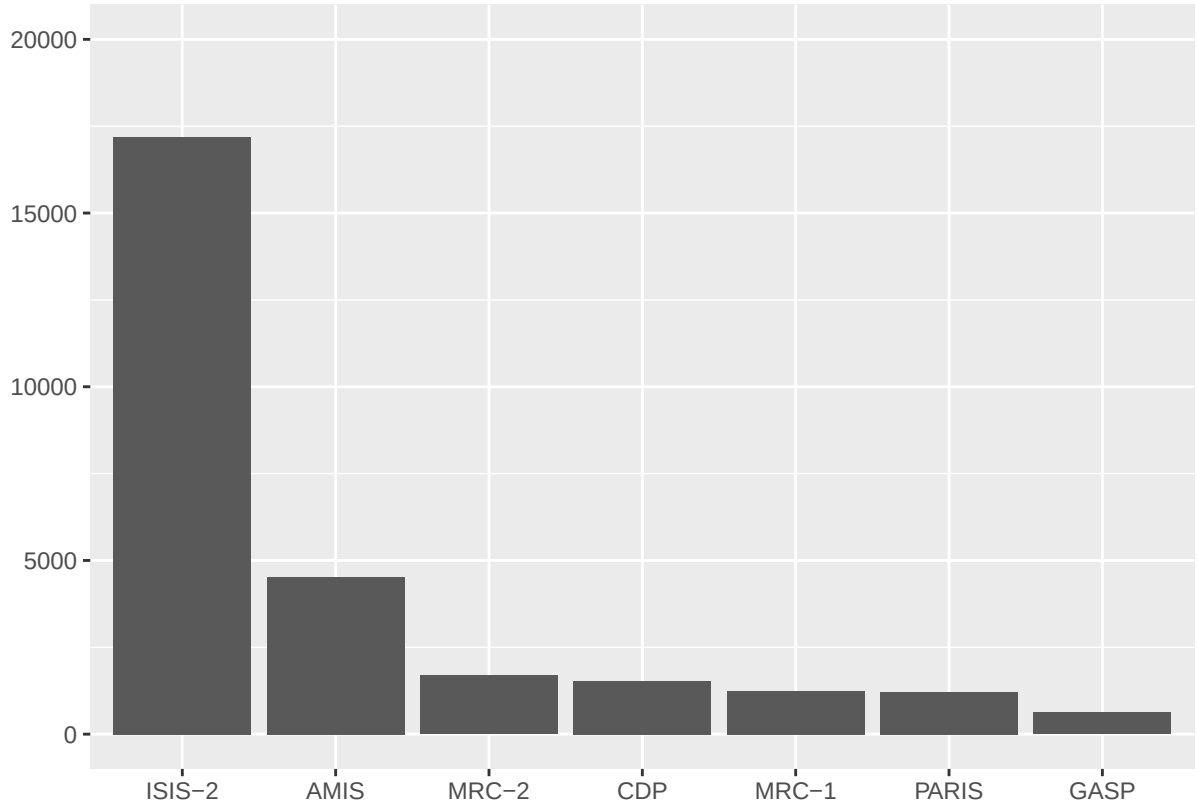
Variables Nominales

Las variables nominales son aquellas que no tienen orden en sus categorías; el color de pintura, marca de un auto, distintos experimentos, etc.

Aspirinas

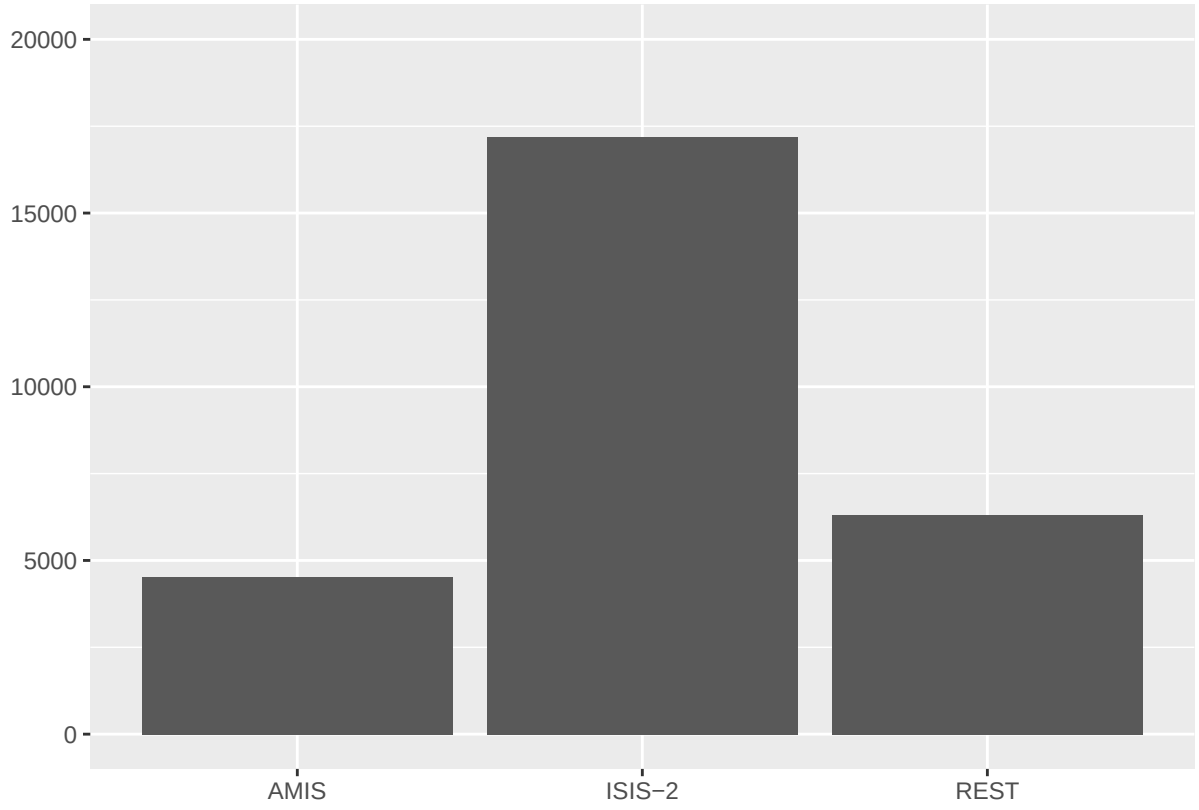
Experimento de uso de aspirinas después de un infarto a miocardio para detener la muerte de los pacientes.

```
data(Fleiss93, package="meta")
Fleiss93 <- within(Fleiss93, {
                   total <- n.e + n.c
                   st <- reorder(study, -(total)) })
ggplot(Fleiss93, aes(st, total)) + geom_bar(stat="identity") +
       xlab("") + ylab("") + ylim(0,20000)
```



Un experimento contiene una concentración muy importante de los experimentos por lo que si agrupamos todos los que contengan menos de 2000 individuos, tenemos el siguiente gráfico:

```
Fleiss93 <- within(Fleiss93, st1 <- as.character(study) )
Fleiss93 %<>% mutate(st1 = ifelse(total>2000,st1, "REST"))
ggplot(Fleiss93, aes(st1, total)) + geom_bar(stat="identity") +
       xlab("") + ylab("") + ylim(0,20000)
```

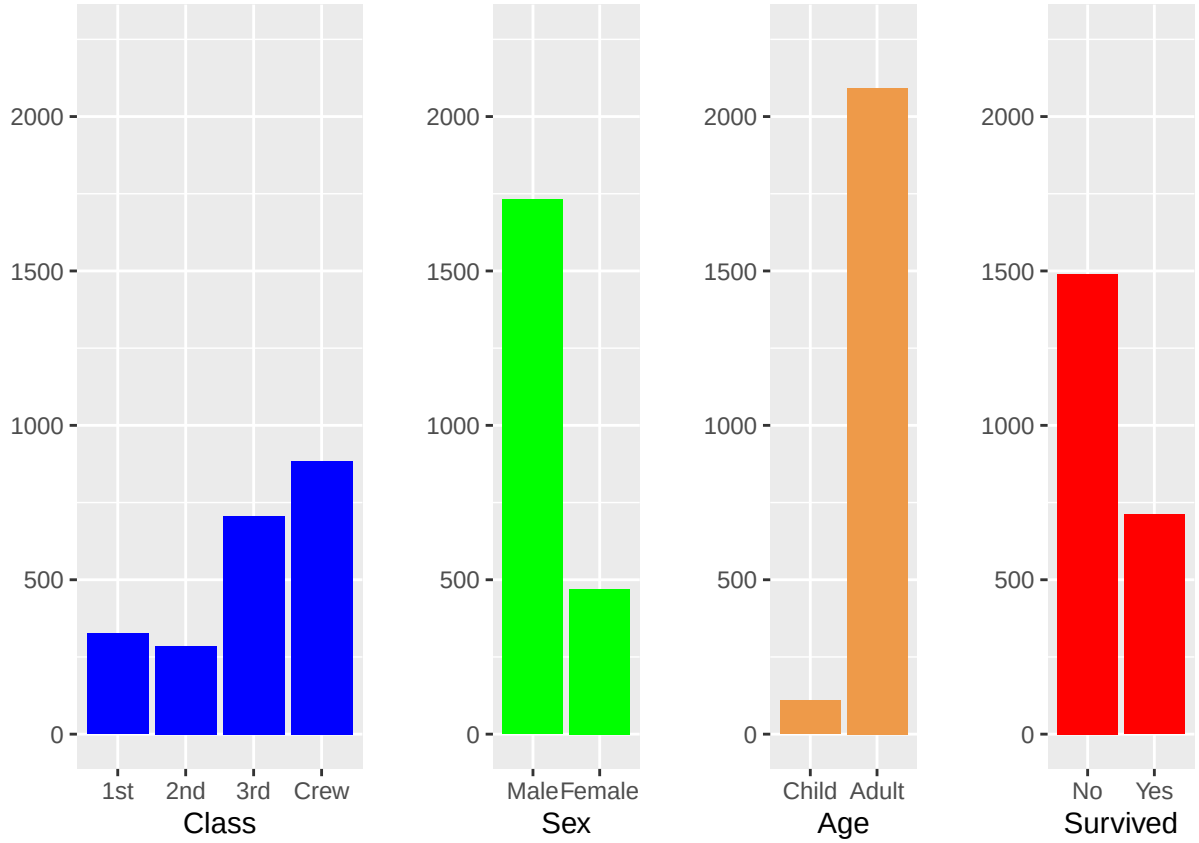


Si agrupamos todos los estudios menores a 2000 individuos, se tiene otro tipo de visualización mucho más clara y también se entiende bastante claramente el hecho que el experimento ISIS-2 contiene mucha más poblacioón que todos los demás.

Titanic

Dentro de la base de datos que ya vimos en la clase pasada, veremos que si hacemos los distintos gráficos a misma escala con distintas vriables tenemos una historia más interesante que sólo el conteo de muertes:

```
library(gridExtra)
Titanic1 <- data.frame(Titanic)
p <- ggplot(Titanic1, aes(weight=Freq)) +
            ylab("") + ylim(0,2250)
cs <- p + aes(Class) + geom_bar(fill="blue")
sx <- p + aes(Sex) + geom_bar(fill="green")
ag <- p + aes(Age) + geom_bar(fill="tan2")
su <- p + aes(Survived) + geom_bar(fill="red")
grid.arrange(cs, sx, ag, su, nrow=1, widths=c(3, 2, 2, 2))
```



Deberías de tener una idea en la cabeza de lo que quieres mostrar antes de graficar cualquier cosa, de esta forma es más fácil que te sorprendas de las cosas que puedas encontrar.

Discutan el expectativas vs gráfico.

Variables Ordinales

Muchas veces cuando se levanta una encuesta, se tiene en cuenta un rango de valores posibles para que el encuestado "califique" la variable que medimos, "En la escala del 1 al 5 ....".

Datos discretos, conteos y enteros:

Para estas visualizaciones podemos tomar en cuenta la ocurrencia de distintas variables.

TL;DR

- Los gráficos de barras son una manera muy sencilla de graficar y mostrar información, sin embargo aunque haya simpleza en la manera de presentarlos pueden mostrar cosas que no se esperaban y nos pueden dar un *insight* insteresante.
- Los gráficos de barras se pueden usar para graficar variables nominales, ordinales y discretas.
- El orden de las barras afeta la manera de ver los resultados.

Ejercicios:

Máquinas tragamonedas

En el paquete DAAG, está el dataset vlt, muestren si los símbolos que aparecen tienen la misma frecuencia, o no. Ejes comparables.

Daño gastrointestinal

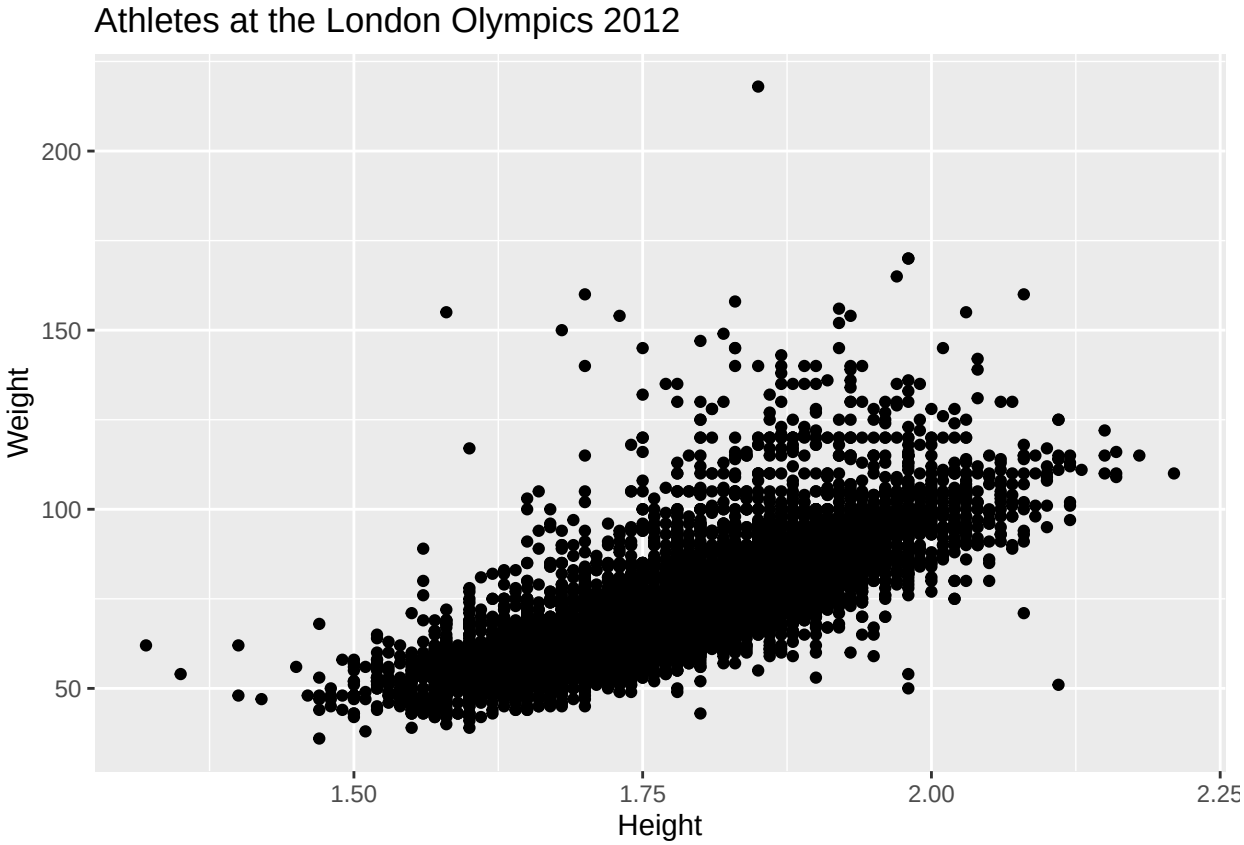
Del conjunto de datos `Lanza` del paquete `HSAUR2`, hay cuatr estudios, dibujen un gráfico para comparar los tamaños de los estudios, ¿son iguales? El resultado se mide en la variable classificación, ¿qué opinan?

Estructura, dependencias, relaciones y asociaciones.

Buscaremos explicar pares de variables y sus relaciones, para esto un scatterplot es la manera más sencilla y común de hacerlo. El scatterplot nos ayuda a buscar infomración que no necesariamente es obvia de alguna tabla o de los resúmenes de la información. El verdadero valor de estos gráficos reside en ver las interacciones entre dos variables, también nos ayudan a buscar valres extremos. A pesar de la información que nos pueda arrojar un scatterplot se recompenda que se sigan graficando las distribuciones univariadas para entender el proceso.

Para las olimpiadas de 2012 se muestra el scatterplot entre altura y peso:

```
data(oly12, package="VGAMdata")
ggplot(oly12, aes(Height, Weight)) + geom_point() +
       ggtitle("Athletes at the London Olympics 2012")
```



¿Qué buscamos en los Scatterplots?

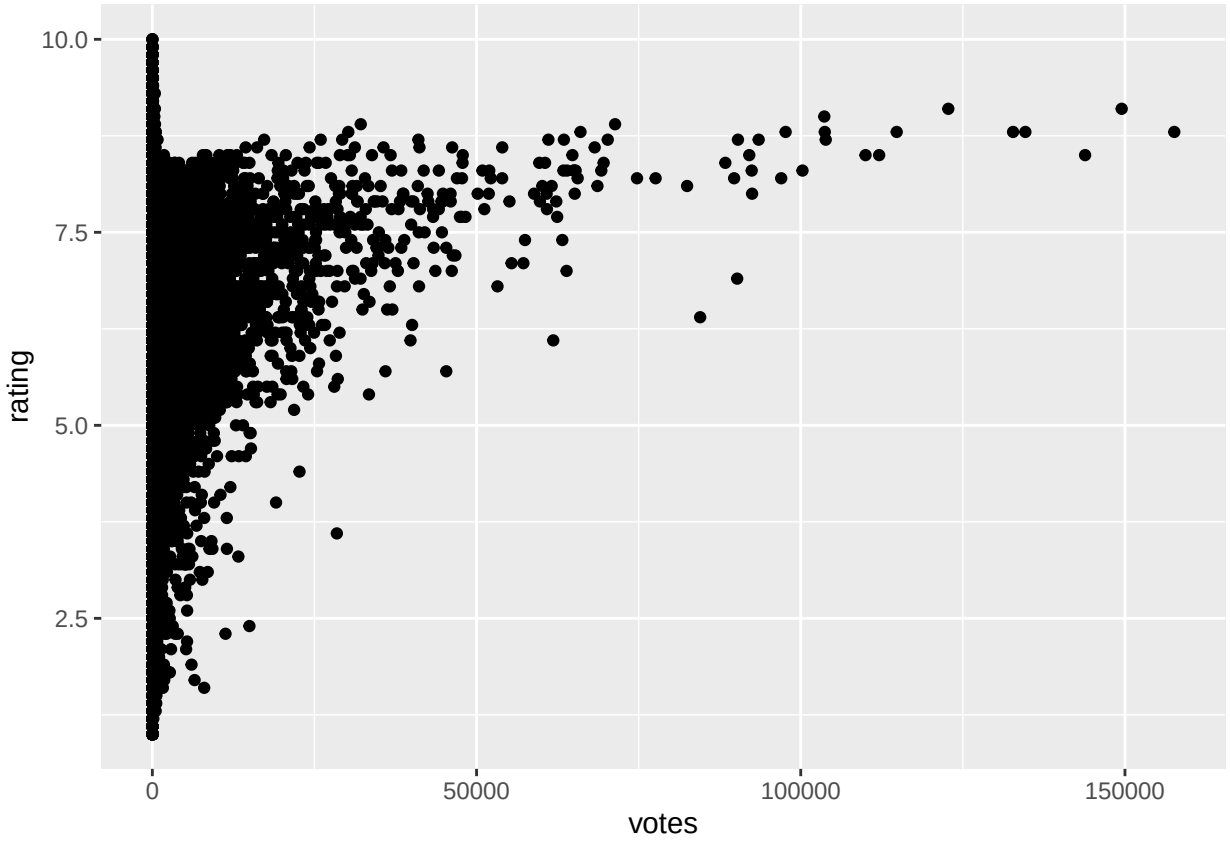
Acerca de correlación

- Relaciones entre variables, puede que haya una fuerte relación entre las variables, buscamos causalidad aunque debe quedar claro que es muy complicado establecer esta relación en general.
- Asociaciones, grupos de variables que puedan estar relacionados entre si, sin ser afectados por la misma causa directa.

- Valores extremos: Puede haber casos de variables extremas en la relación de las variables sin que haya algún extremo en ninguna de las variables individuales.
- Clusters: Ayuda buscar grupos de individuos, como el que vimos de las flores, la clase pasada.
- Gaps: Ocasionalmente algunas parejas de vlores son complicadas que aparezcan juntas,
- Barreras: barreras naturales dentro de la variable, por ejemplo: no se puede tener más edad laboral que la edad actual del individuo.
- Condicionales: hay veces que es más faćil entender al variable con subgrupos de otra variable, ingreso vs jubilación.

Movie ratings

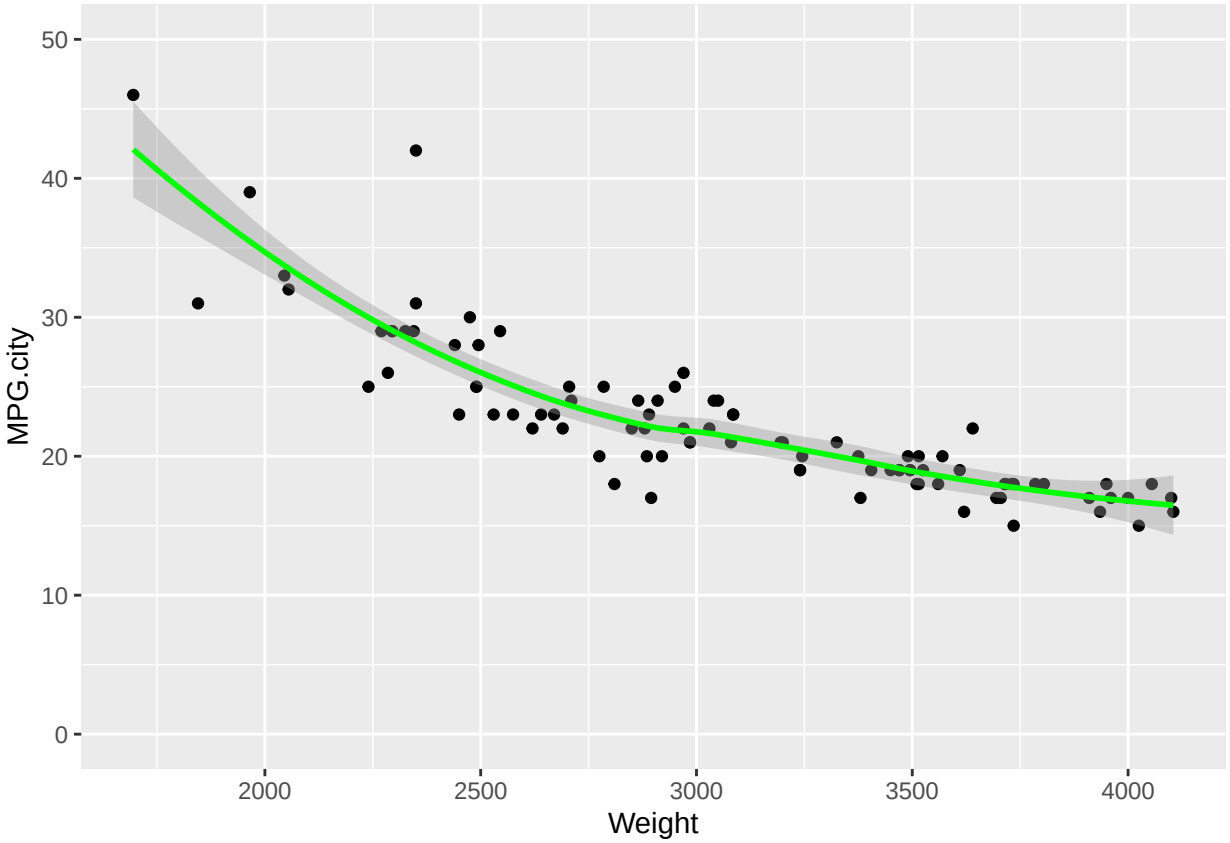
```
library(ggplot2movies)
ggplot(movies, aes(votes, rating)) + geom_point() + ylim(1,10)
```



Discutan:

Líneas y suavizado.

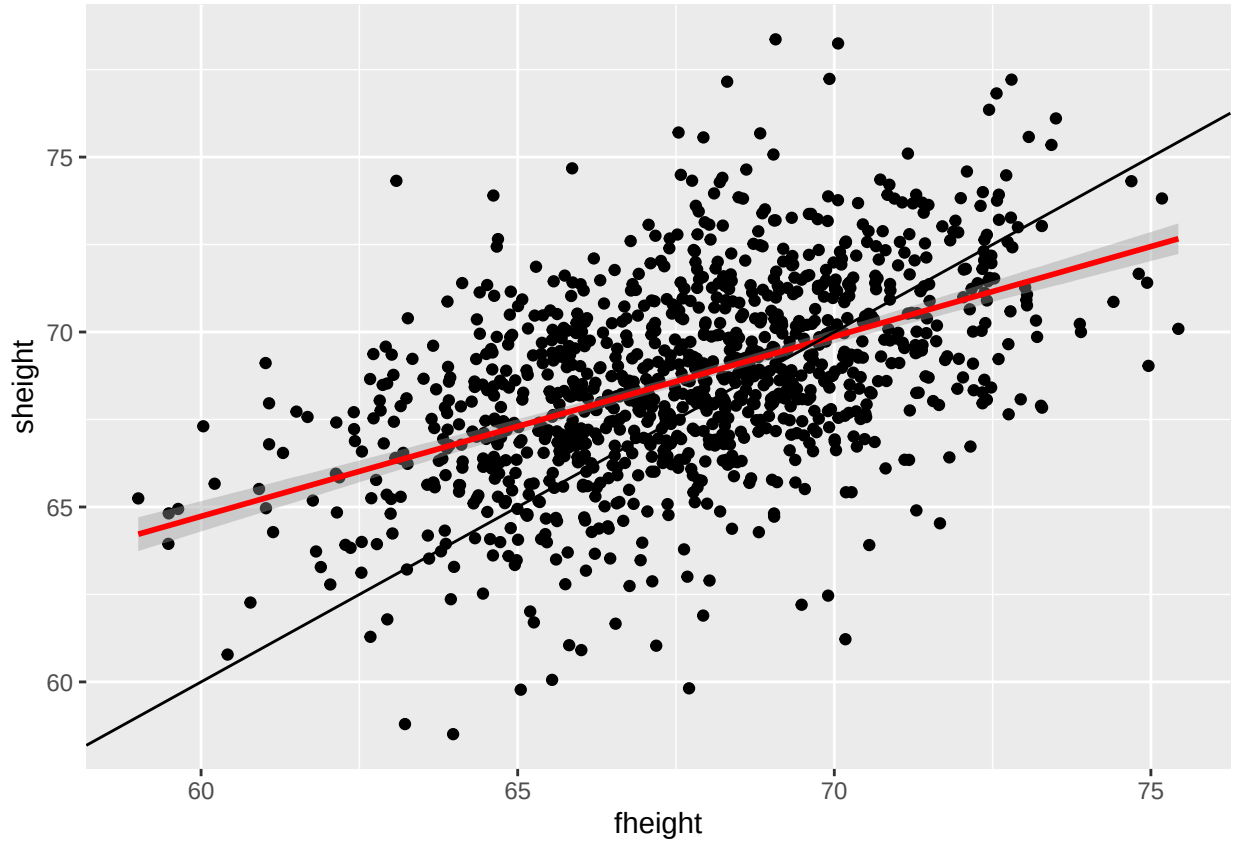
```
data(Cars93, package="MASS")
ggplot(Cars93, aes(Weight, MPG.city)) + geom_point() +
       geom_smooth(colour="green") + ylim(0,50)
```



Vemos cómo conforme aumenta el peso las millas por galón tieneden a disminuír, (autos más grandes usan más combustible)

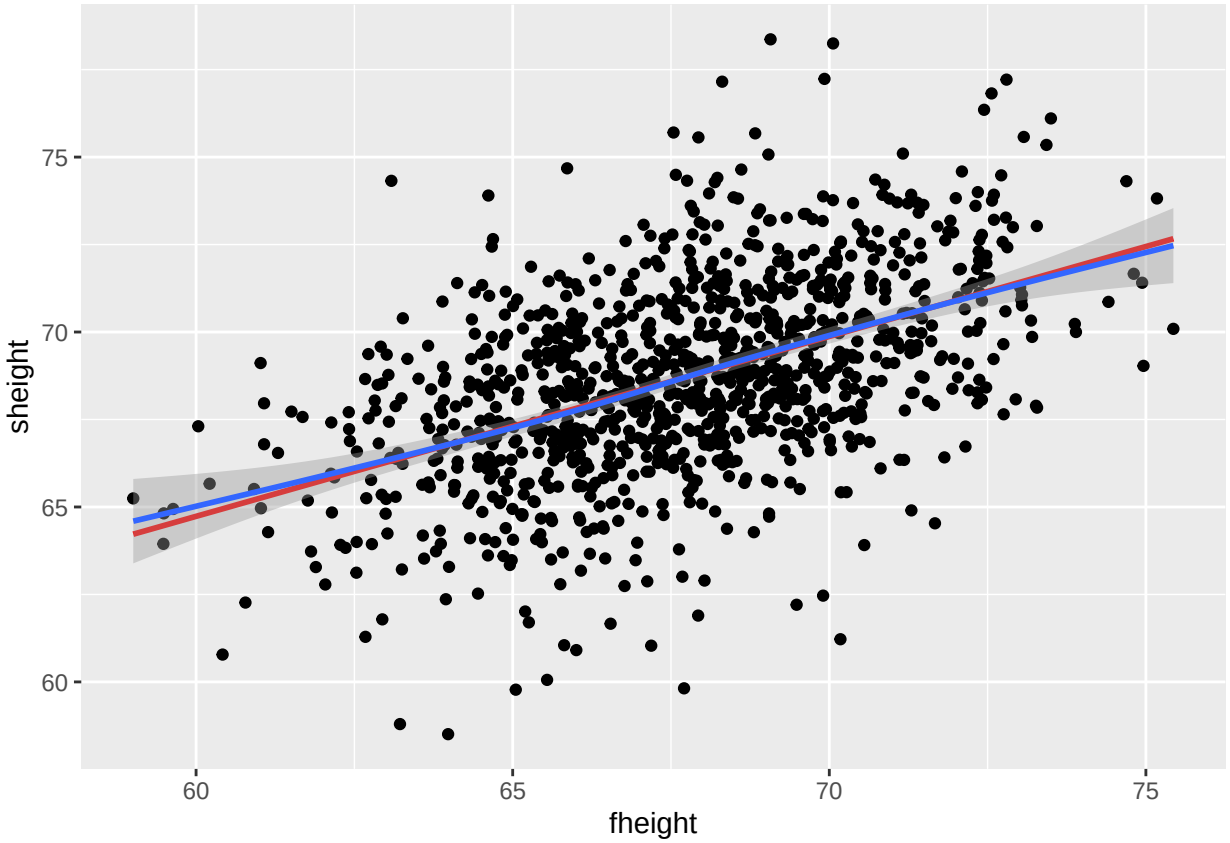
Alturas de las personas

```
data(father.son, package="UsingR")
ggplot(father.son, aes(fheight, sheight)) + geom_point() +
       geom_smooth(method="lm", colour="red") +
       geom_abline(slope=1, intercept=0)
```



Esto nos ayuda a ver que puede que haya una buena relacioń entre las variables, por lo que pintamos la línea $y = x$, para comparar. Ahora, veamos

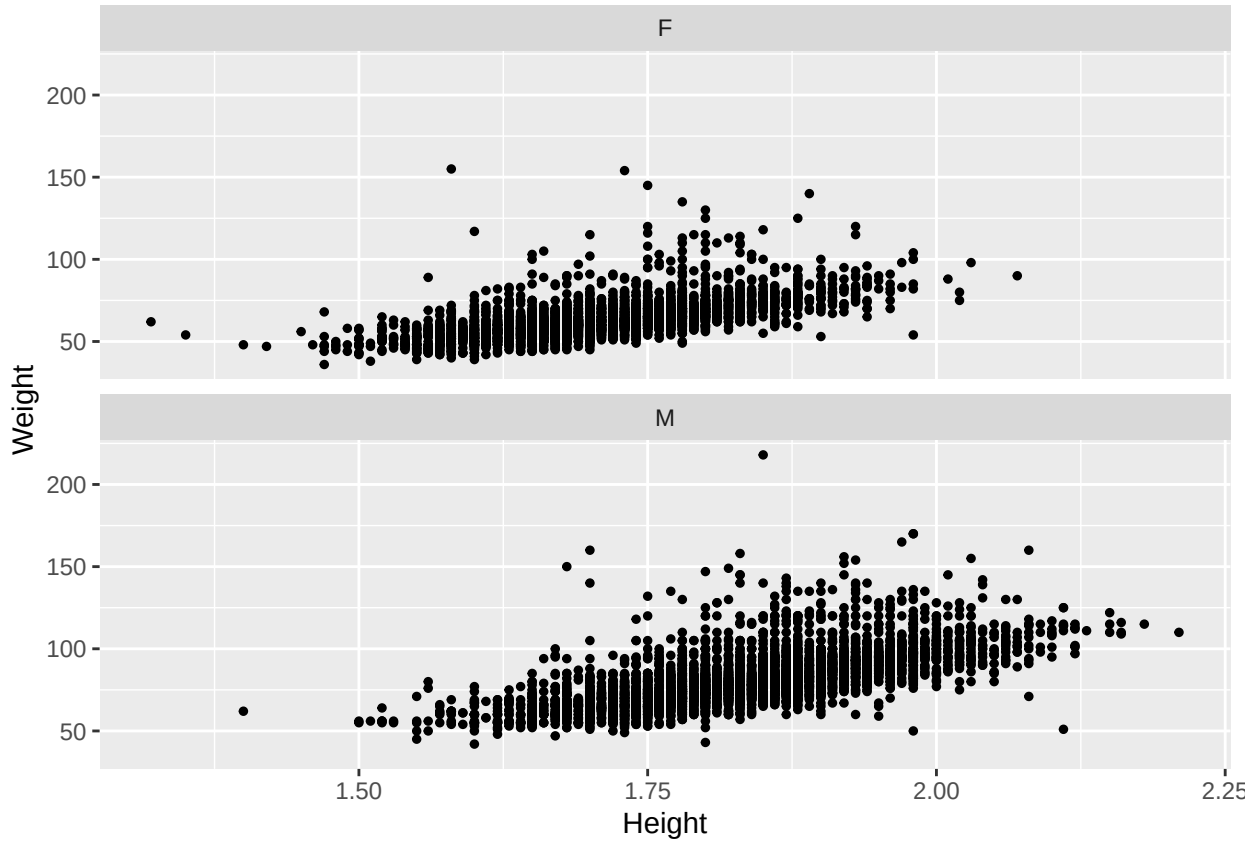
```
data(father.son, package="UsingR")
ggplot(father.son, aes(fheight, sheight)) + geom_point() +
       geom_smooth(method="lm", colour="red", se=FALSE) +
       stat_smooth()
```



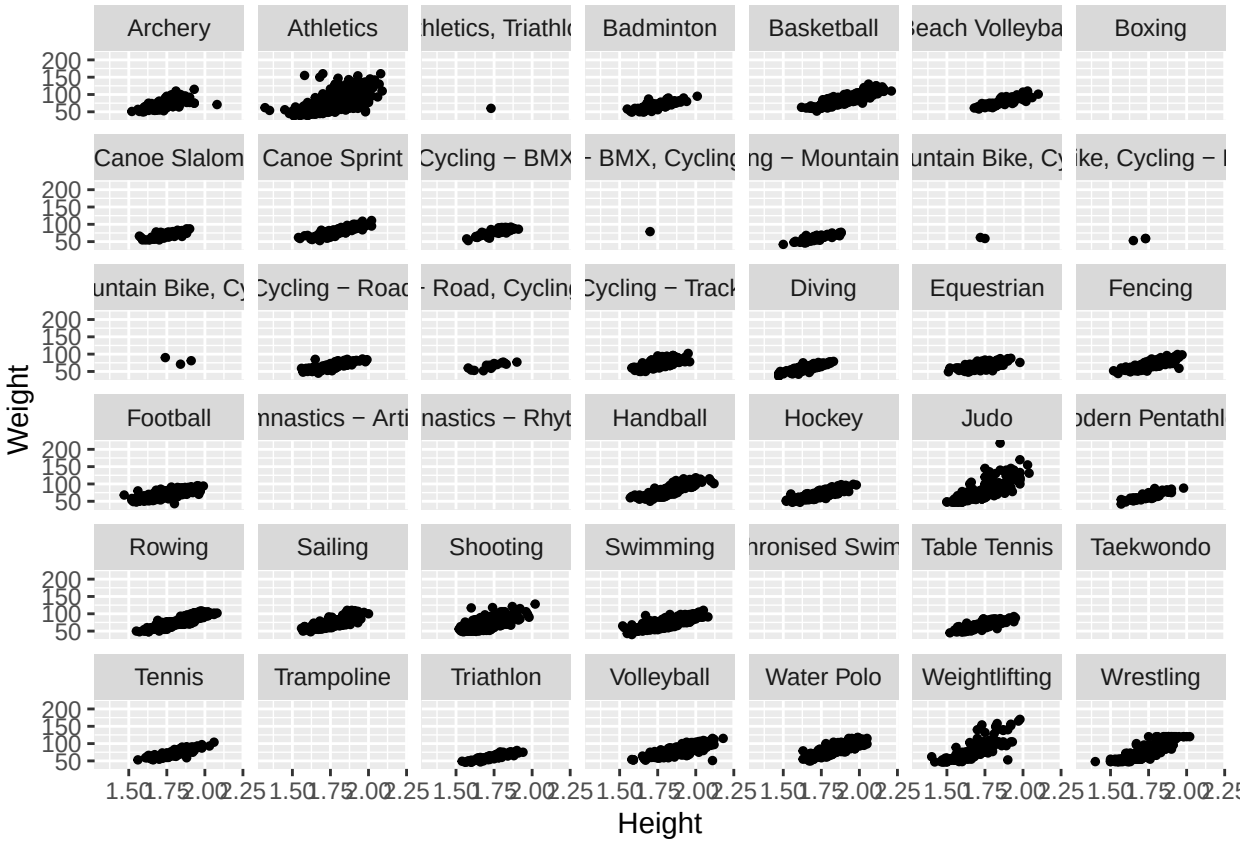
Grupos de categorías

Filtremos por grupos de atletas, hombres y mujeres.

```
ggplot(oly12, aes(Height,Weight))+geom_point(size=1)+facet_wrap(~Sex,ncol=1)
```




```
oly125 <- mutate(oly12,Sport=abbreviate(Sport,12))
ggplot(oly12, aes(Height,Weight))+geom_point(size=1)+facet_wrap(~Sport)
```



TL;DR

- Scatterplots toman diferentes formas y regresan mucha informaciń acerca de la relación de dos variables.
- Agregar los suavizados es trivial y ayuda con el modelado de nuesrta información.

Ejercicios:

Películas:

¿Cómo se ve el scatterplpot si quitamos los que tienen más de 1000 votos? Ahora filtremos los mayores a 9. ¿Pasa algo?

Autos:

Grafiquen 1/MPG.City, vs horsepowerhay ¿una relación lineal? ¿Cuáles son los outliers?

Bancos

Dentro del paquete `Sleuth2`, los datos: `case1202`, tiene tres variables medidas en meses, de: Edad, Seniority, Experience.

¿Qué hay en la matriz de scatterplots? ¿Por?