Visualización con R

Miguel Angel Escalante Serrato

Octubre 13, 2020

Contents

Introducción

La graficación es una manera eficiente de resumir, y mostrar información. Es fundamental entender el contexto para tener más información con respecto a lo que se está graficando. Muchas veces es mejor hacer varias gráficas simples en lugar de una muy compleja.

No hay teoría muy complicada con respecto a las gráficas, de hecho no hay mucha en lo absoluto. Usualmente cuando se revisan las distintas gráficas y formas, usualmente libros y cursos se siguen a otras cosas complejas. (léase cursos de estadística, probabilidad, etc.) Lo que vemos actualmente nos deja mucho que desear de cómo se usan las gráficas y más aún nos revela la necesidad de ahondar más en estos temas y revisarlos más a fondo.

John Turkey, nos dice en cuatro frases el verdadero propósito de la visualización de información:

1. Las gráficas son para análsisis cualitativos o descriptivos y quizá semi cuantitativos, nunca para análisis profundo cuantitativo.

2. Las gráficas son para comparaciones, comparaciones entre grupos, no para acceder a cantidades particulares.
3. Las gráficas son para impactar, de primera instancia, mover percepción, transmitir información, no se debería de pensar mucho para llegar a la conclusión.
4. Las gráficas deberían de reportar anáisis de datos trabajado, fino y cuidadoso. No se debería intentar que la gráfica sea en si el análisis. Las gráficas están para fortalecer el análisis, no para fundamentarlo.

Análisis de Datos con Gráficos.

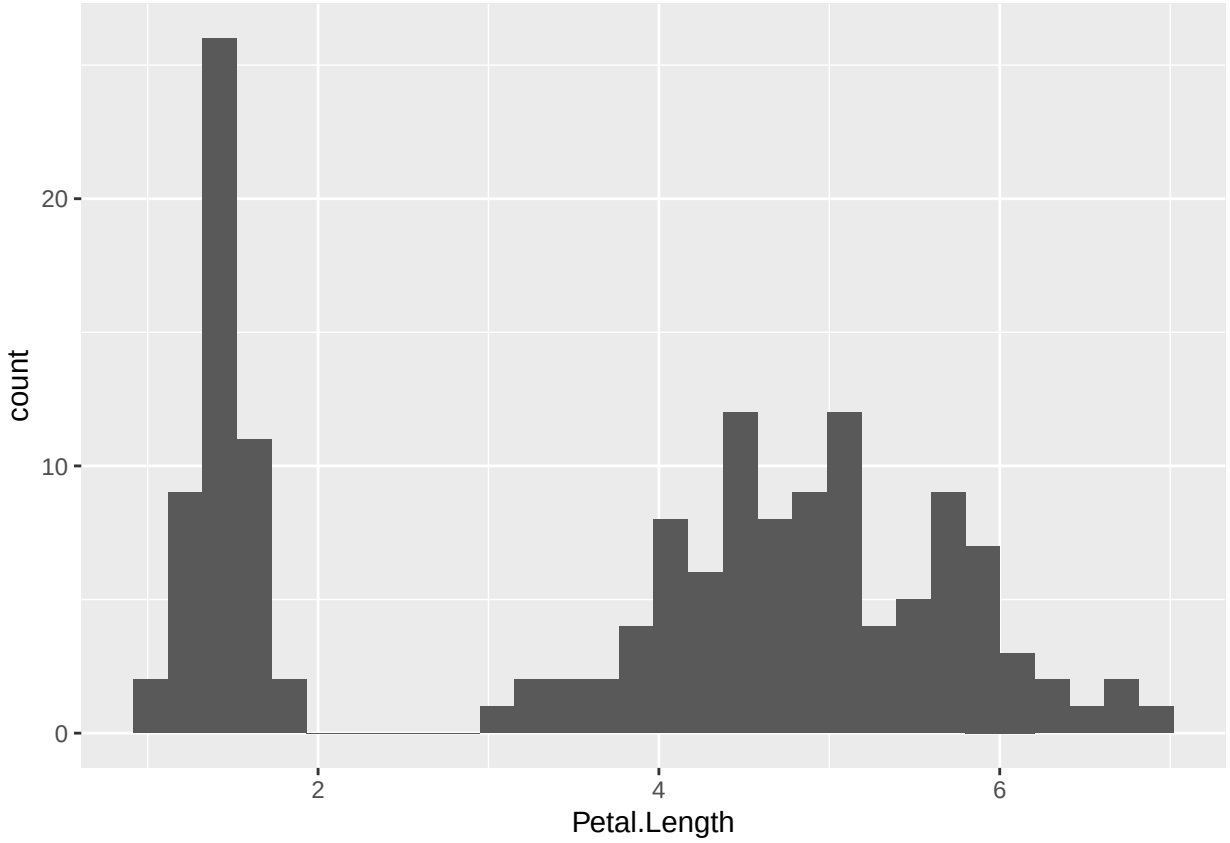
Ahora veremos ejemplos de visualizaciones que se pueden hacer para entender más a profundidad los datos.

Iris

Del conjunto de datos de Fisher, Iris, que contiene información de medidas de plantas usado por Fisher para ilustrar análisis de discriminante linear. Es de los conjuntos de datos más usados para ejemplos.

Un ejemplo de visualización:

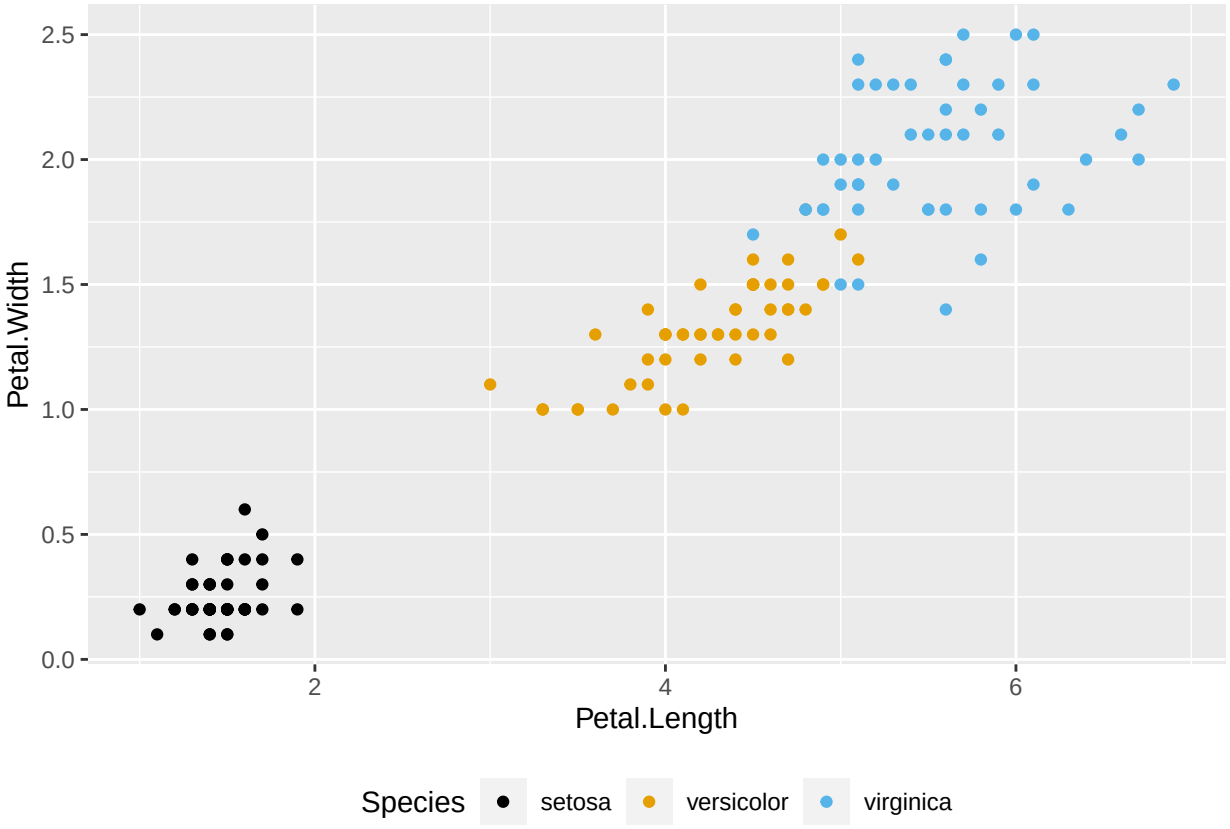
```
ggplot(iris, aes(Petal.Length)) + geom_histogram()
```



Podemos ver que hay al menos dos gupos de flores según su longitud de pétalo.

Sin embargo si ponemos una capa más de complejidad en la visualización podemos observar un poco más acerca de estos.

```
library(ggthemes)
ggplot(iris, aes(Petal.Length, Petal.Width, color=Species)) +
       geom_point() + theme(legend.position="bottom") +
       scale_colour_colorblind()
```



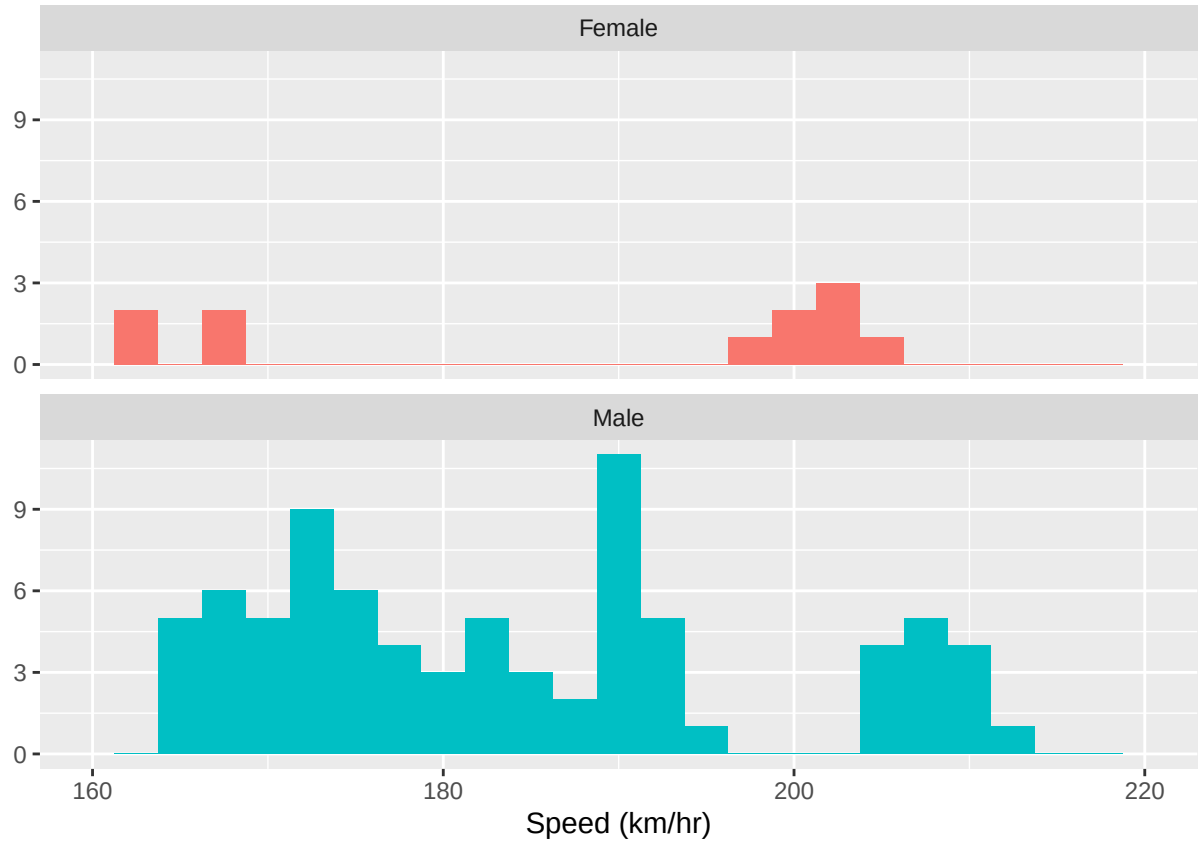
Con esta nueva visualización podemos ver una diferencia entre tres especies de plantas y cómo hay una clara relación entre la longitud y amplitud del pétalo.

Ski

En el campeonato de Sk de 2011, se tomaron medidas de velocidad y el headline decía lo siguiente: "El mejor tiempo lo tuvo un hombre y el peor tiempo lo tuvo una mujer."

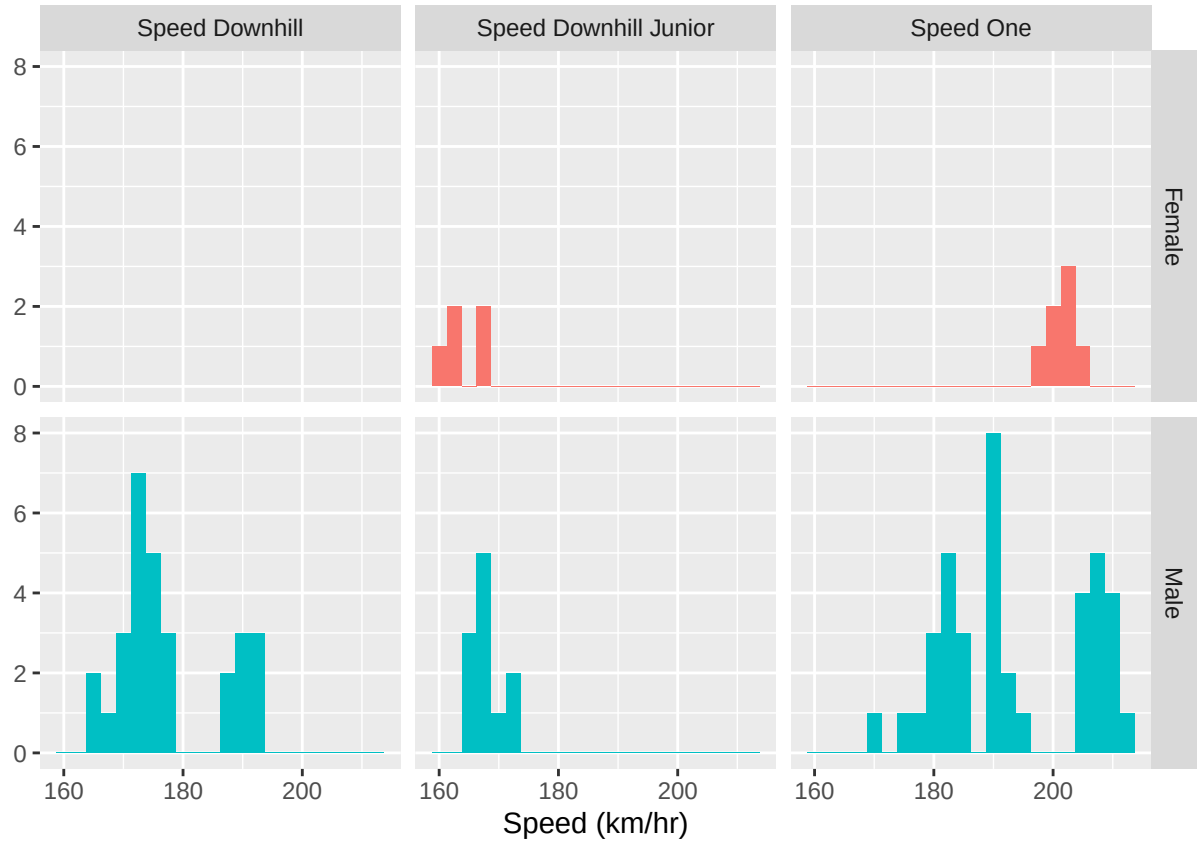
¿Pueden ver qué está mal?

```
data(SpeedSki, package = "GDAdata")
ggplot(SpeedSki, aes(x=Speed, fill=Sex)) + xlim(160, 220) +
       geom_histogram(binwidth=2.5) + xlab("Speed (km/hr)") +
       facet_wrap(~Sex, ncol=1) + ylab("") +
       theme(legend.position="none")
```



Resulta que había varias categrorías, y también resultó que participaron muchas menos mujeres, también está el hecho que en una categoría no participaron mujeres en lo absoluto:

```
ggplot(SpeedSki, aes(Speed, fill=Sex)) +
       geom_histogram(binwidth=2.5) + xlab("Speed (km/hr)") +
       ylab("") + facet_grid(Sex~Event) +
       theme(legend.position="none")
```



Es importante dejar claro lo que tenemos en los datos, siempre se puede mostrar una versión de la realidad sesgada; sin embargo para entender lo que sea que estudiemos hay que ver todas las posibles maneras de mostrarlo.

Contexto

El análisis de gráficas no puede quedar solo, ya lo dije y debe de quedar grabado: Todo resultado obtenido de los gráficos debe de ser verificado con análisis estadístico. Usualmente se usa la graficación para verificar resultados estadísticos, sin embargo también la estadística nos debe de dar luz sobre los resultados que aparecen en las visualizaciones. Ver es creer, pero probar y verificar es convincente.

Las visualizaciones nos ayudan a relvelar estructura más que detalles, para ver diferencias considerables más que para entender las diferencias finas. Si se requieren valores exactos se deben de pensar en tablas. Las tablas con datos y las gráficas no son competidores sino complementarios. Cuando pensamos en reportes impresos, es difícil decidir en qué poner, si tabla o gráficas, sin embargo, con un reporte digital, se puede pensar en incluír las dos opciones.

La importancia de la visualización de datos puede ser subestimada, sin embargo no es nada fácil. Una visualiación bien hecha puede cambiar perfectamente la visión de un fenómeno a estudiar. El problema puede ser entender el problema y buscar una visualización que plasme lo que queremos decir y el entendimiento del fenómeno. No hay una teoría ni un método establecido, más bien guías y ciertas reglas que podemos estudiar.

> Yet if it does not seem a moment's thought, Our stiching and unstiching has been naught.

Es una analogía interesante a lo que decimos.

Ejercicios.

Velocidad de la luz

Tenemos 5 experimentos con 20 corridas de mediciones de la velocidad de la luz.

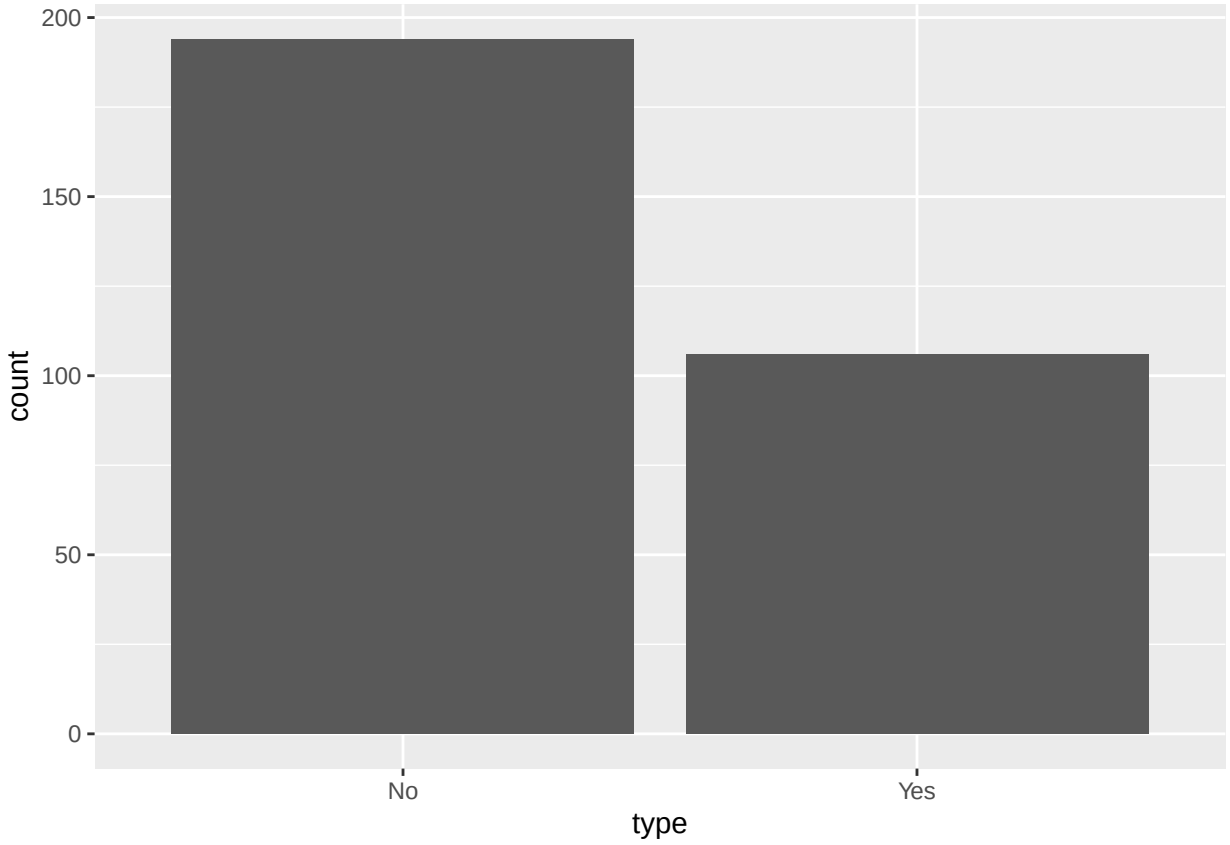
- ¿Cómo exponen los resultados?
- ¿Los resultados se ven equivalentes entre los experimentos?

```
library(MASS)
vl <- data.frame(michelson)
```

Pima Indians

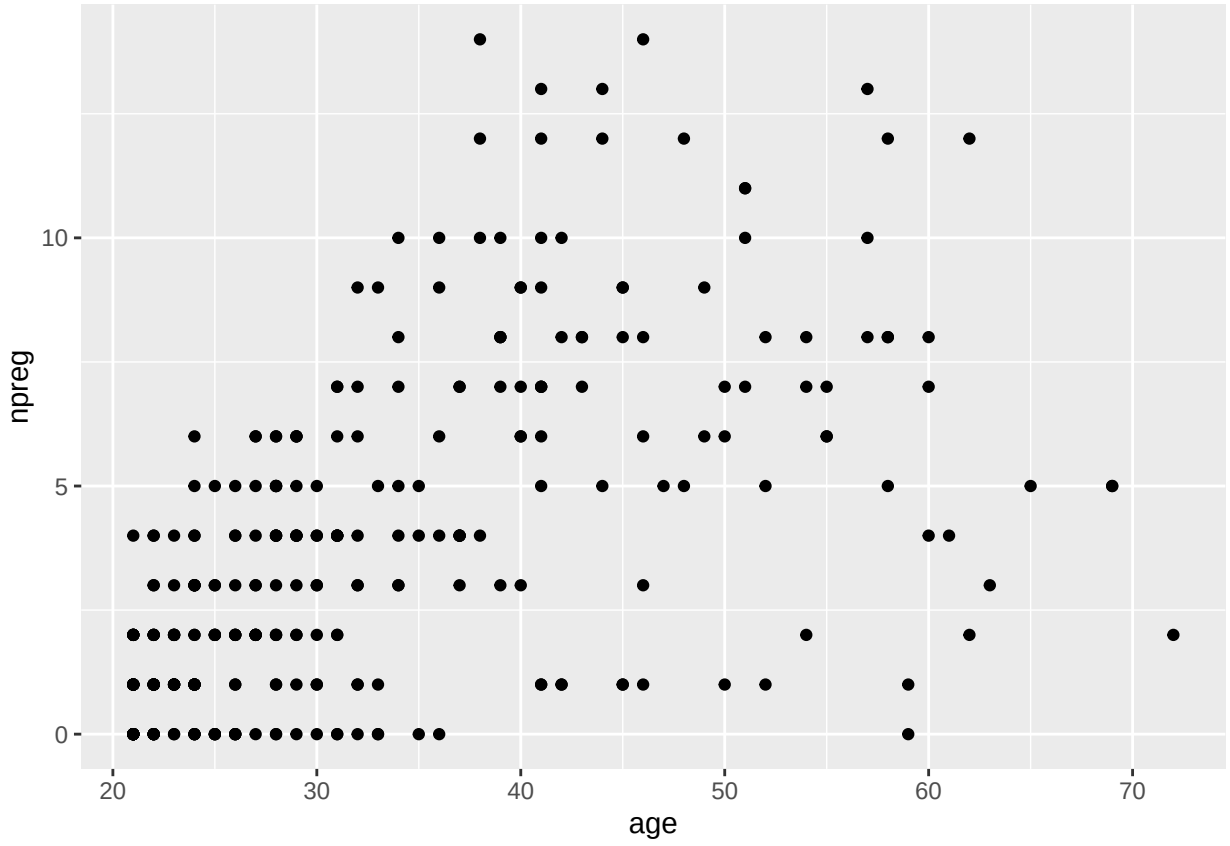
- ¿Qué dice este gráfico de barras?

```
ggplot(Pima.tr2, aes(type))+geom_bar()
```



- ¿Por qué la parte superior izquierda está vacía?

```
ggplot(Pima.tr2, aes(age,npreg))+geom_point()
```



Titanic

```
head(data.frame(Titanic))
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

- ¿Qué se les ocurre para graficar todo?
- ¿Cuántos pasajeros en total viajaban por clase?
- Exploren una por una las otras variables categóricas del *dataset*

Swiss

Tenemos datos de fertilidad y otras variables socioeconómicas de 47 provincias franco-parlantes en Swiza en 1888.

```
head(data.frame(swiss))
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
```

```
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##               Infant.Mortality
## Courtelary                22.2
## Delemont                  22.2
## Franches-Mnt              20.2
## Moutier                   20.3
## Neuveville                20.6
## Porrentruy                26.6
```

- ¿Cómo conjuntarían todas las variables?
- Hagan gráficas para cada variable. ¿Se ve algo especial o raro?
- Hagan un scatterplot de Fertilidad vs el porcentaje de católicos. ¿Se ve algo interesante?
- ¿Hay relación entre agricultura y Educación?

TL;DR

1. El análisis gráfico se usa para revelar información del dataset, es una herramienta exploratoria.
2. Visualizaciones sencillas pueden dar bastante infomaciónn acerca del fenómeno
3. Formato y escalado son importantes
4. Distintos gráficos nos pueden dar más información acerca de lo obsrvado

Variables Continuas

Una variable continua en principio puede tomar cualquier valor dentro del rango dado $(0, 1)$, por ejemplo. Las variables continuas en la práctica usualmente están redondeadas a algún nivel de precisión. Hay muchas maneras de describir variables continuas, dependiendo del énfasis que se le quiera dar es la elección. Dada la gran cantidad de características que se puedan ocurrir caracterizar, hay muchas maneras de visualizar variables continuas.

Nos enfocaremos en dos:

- Boxplots (Diagrama de caja y brazo)
- Histogramas

Características a buscar

- **Asimetría** Distribuciones simétricas o sesgadas hacia algún lado (distribuciones de ingreso).
- ***Outliers*** Valores que están muy lejanos al resto de las observaciones, ojo estos pueden contener información valiosa.
- **Multimodalidad** En caso que haya más de una "joroba" en la distribución de los valores.
- **Gaps** Puede ser natural que no haya un área del rango de las variables (calificaciones de exámenes).
- **Amontonamiento (heaping)** Hay veces que por facilidad los valores se acumulan en un mismo valor particular, o especial. (Hora denuncias)
- **Redondeo** Registro de valores redondeados (edad)
- **Imposibles** Registro de valores que no son posibles (Edades negativas)
- **Errores** Registro incorrecto de valores. (edades de 99 anyone ¿?)

Las visualizaciones como los histogramas y los boxplots nos ayudan a entender estas características, sin embargo reitero que se requiere también un análisis estadístico de los valores.

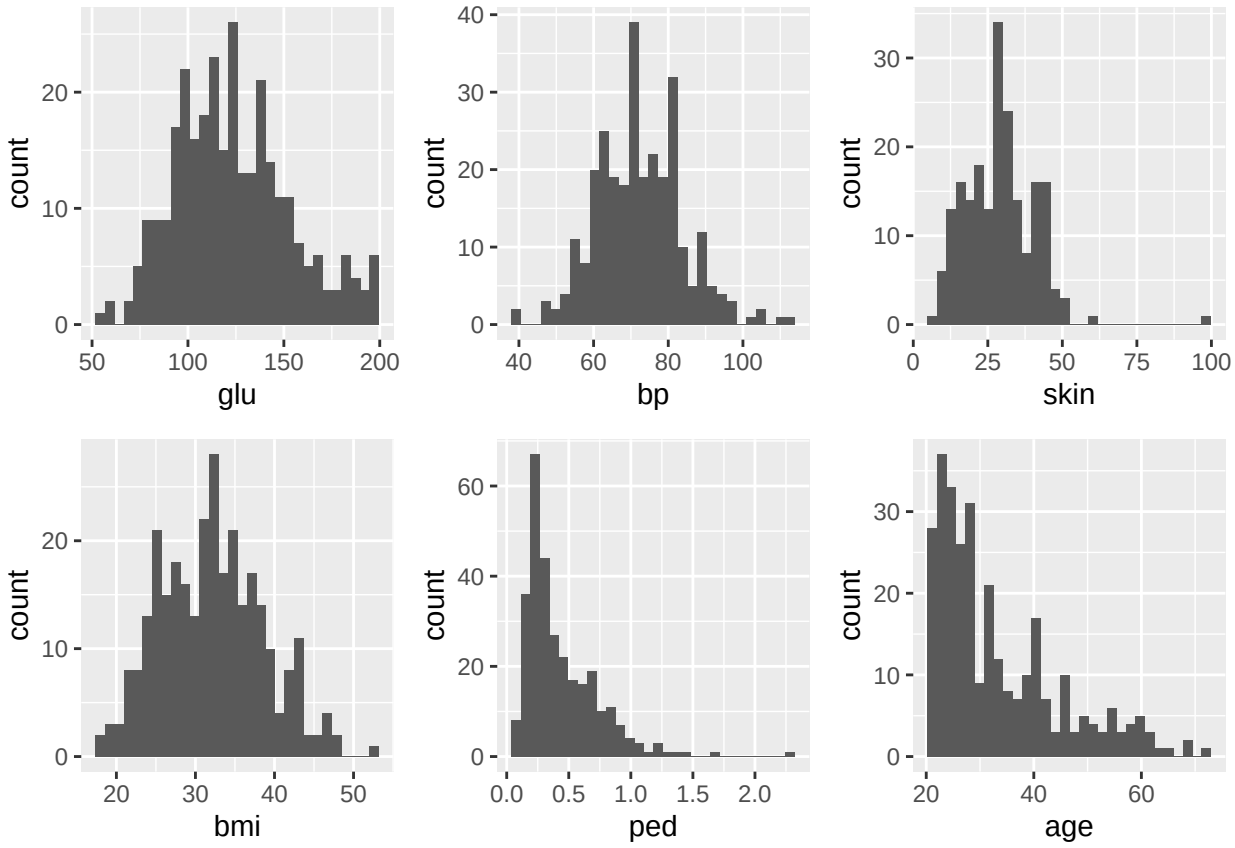
De las estadísticas que describen a una vriable probablemente la media es la medida más común y usada para describir los valores, y usualmente para hacer comparaciones de medias se usa bastante la prueba t. La prueba t depende del supuesto que la variable es normal, aunque este supuesto no es tan riguroso y hay

maneras de datle la vuelta; esto es para explicar el hecho que también hay que hacer pruebas de normalidad (léase qq-plots, por ejemplo).

Ejemplos

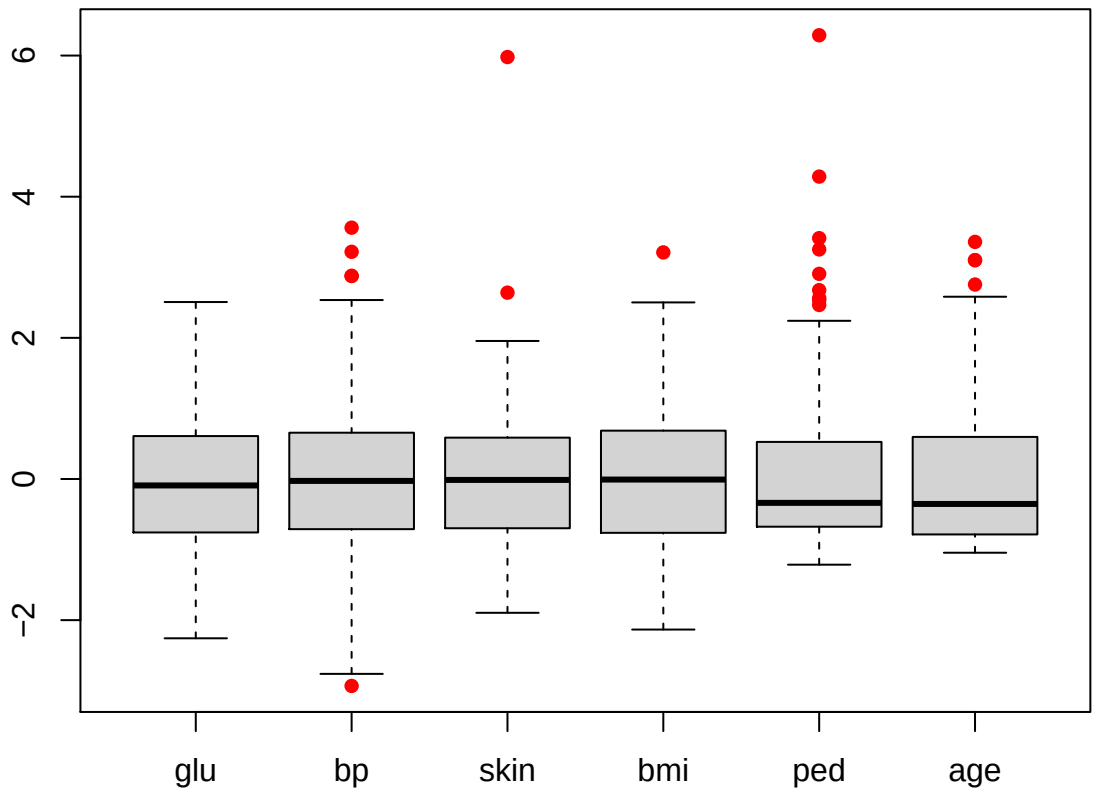
Pima

```
library(gridExtra)
data(Pima.tr2, package="MASS")
h1 <- ggplot(Pima.tr2, aes(glu)) + geom_histogram()
h2 <- ggplot(Pima.tr2, aes(bp)) + geom_histogram()
h3 <- ggplot(Pima.tr2, aes(skin)) + geom_histogram()
h4 <- ggplot(Pima.tr2, aes(bmi)) + geom_histogram()
h5 <- ggplot(Pima.tr2, aes(ped)) + geom_histogram()
h6 <- ggplot(Pima.tr2, aes(age)) + geom_histogram()
grid.arrange(h1, h2, h3, h4, h5, h6, nrow=2)
```



Las distribuciones de variables plasma, presión arterial y bmi parecen ser simétricas, *skin* tiene algunos outliers, la distribucío de *ped* se ve sesgada y con algunos outliers. Podemos ver también que los grupos de edad son jóvenes.

```
library(dplyr)
PimaV <- select(Pima.tr2, glu:age)
par(mar=c(3.1, 4.1, 1.1, 2.1))
boxplot(scale(PimaV), pch=16, outcol="red")
```

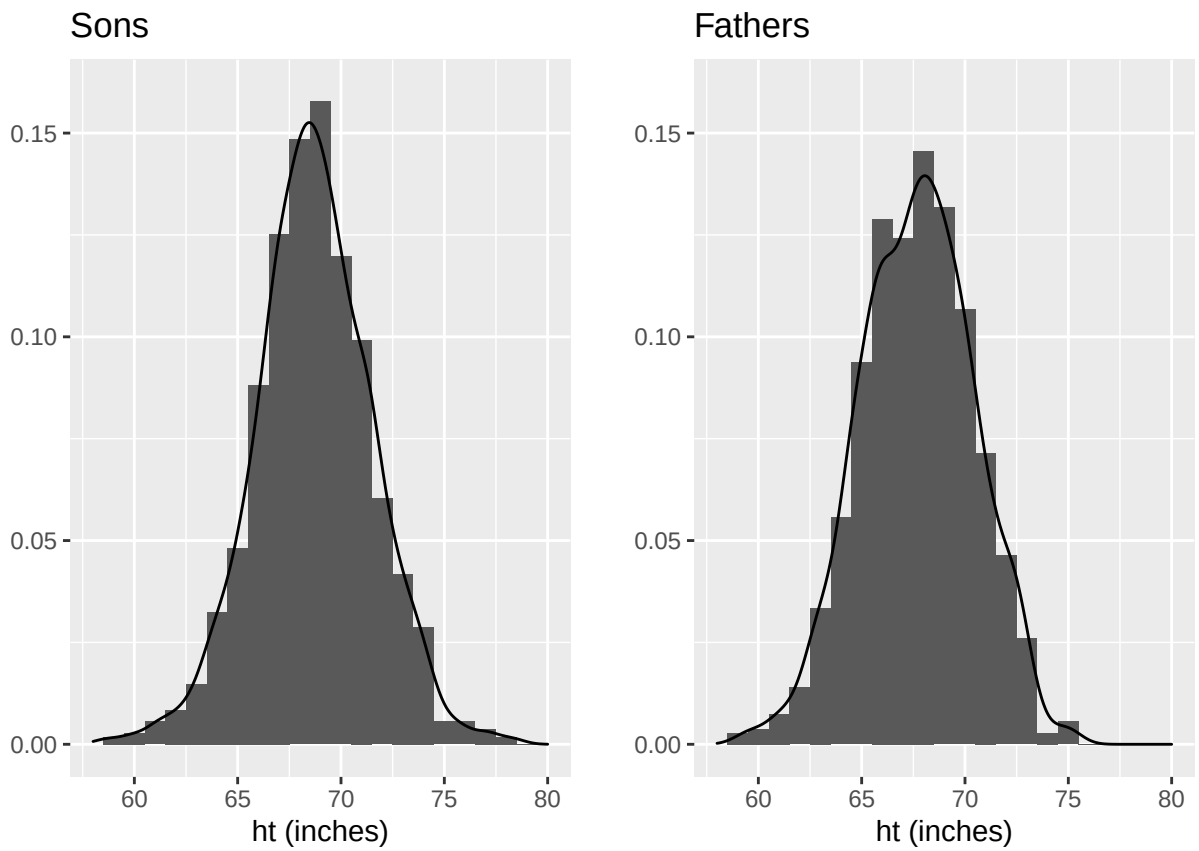


Todo esto se resume muy bien en una sola gráfica de boxplots, que nos ayuda a entender lo que está pasando con una sola vista más compacta.

Pearson

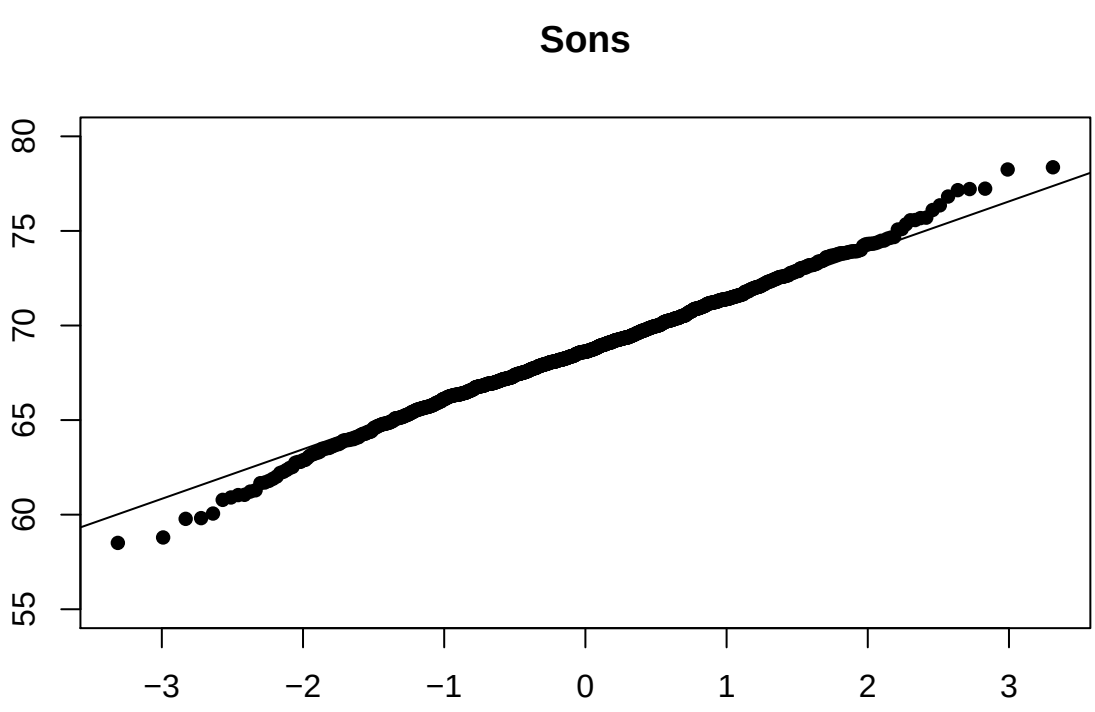
Tomaremos los datos de Karl Pearson, que contiene 1078 alturas apareadas de padres e hijos. Grafiquémoslos en histogramas para entender cómo se ven y en este caso nos interesará la normalidad de las variables:

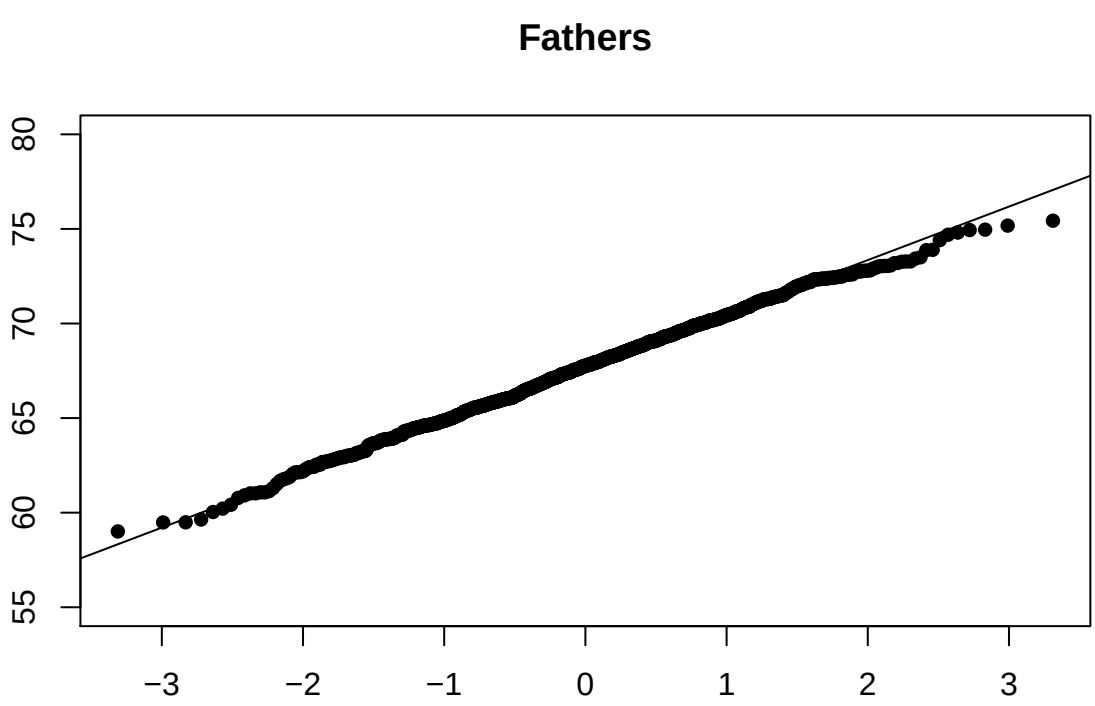
```
data(father.son, package="UsingR")
c2 <- ggplot(father.son, aes(sheight)) +
              geom_histogram(aes(y = ..density..), binwidth=1) +
              geom_density() + xlim(58, 80) + ylim(0, 0.16) +
              xlab("ht (inches)") + ylab("") + ggtitle("Sons")
p2 <- ggplot(father.son, aes(fheight)) +
              geom_histogram(aes(y = ..density..), binwidth=1) +
              geom_density() + xlim(58, 80) + ylim(0, 0.16) +
              xlab("ht (inches)") + ylab("") +
              ggtitle("Fathers")
grid.arrange(c2, p2, nrow = 1)
```



Por lo que se ve de las gráficas y comparando con la imagen que se tiene de normalidad parece que la distribución de las alturas de los hijos es más normal que la de los padres, ahora probaremos esta hipótesis con otra prueba gráfica: qq-plot:

```
with(father.son, {
     qqnorm(sheight, main="Sons", xlab="",
            ylab="", pch=16, ylim=c(55,80))
     qqline(sheight)
     qqnorm(fheight, main="Fathers", xlab="",
            ylab="", pch=16, ylim=c(55,80))
     qqline(fheight)})
```


Sons
80
75
70
65
60
55
−3
−2
−1
0
1
2
3



¿Qué gráficas se pueden ocupara para variables continuas?

1. Histogramas
2. Boxplots
3. dotplot
4. rugplot
5. Aproximación a la densidad
6. Q-Q plot

TL;DR

- Hay muchas características que pueden salir de la frecuencia en las distribuciones.
- No hay una versión óptima por tipo de gráfica, ayuda ver distintas versiones hasta encontrar la más clara.
- Los histogramas ayudan a entender y enfatizar las características de los datos, mientras que las estimaciones de densidad nos ayudan a ver un modelo subyasciente (aunque no siempre)
- Los Boxplots nos ayudan a identificar outliers y comparar distribuciones entre subgrupos de los datos.

Ejercicios variables continuas

Galaxias

Usando el conjunto *galaxies* de MASS, que contiene velocidades para 82 planetas.

1. Hagan histogramas, boxplots y la aproximación de la densidad.
2. Cambien los anchos de las bandas y expliquen cuál es el mejor?
3. Elijan el número de gráficas que se requieren para este experimento

Estudiantes.

El conjunto *survey* contiene información de estudiantes tomando su primer curso de estadística:

1. Hacer el histograma y poner encima la estimación de la densidad, ¿hay bimodalidad?
2. Jueguen con los anchos para tener mejores estimaciones de la densidad, ¿cuál es mejor?
3. Comparen las distribuciones de hombres y mujeres, que compartan la escala, con distintas estimaciones de densidad.

Presupuesto

El conjunto *zuni* del paquete *lawstat*, contiene 3 variabbles, distrito, ingreso por estudiante en dólares y el número de estudiantes.

1. ¿Considerarías el 5% más bajo outliers o extremos?
2. Quitando el 5% inferior hagan el gráfico de estimación de densidad ¿Es simétrico?
3. Hagan un Q-Q plot y comenten si es normal o no.