

# Análisis de Datos

## Tema 2 - Análisis Exploratorio de los Datos

### 2.3 Análisis Exploratorio de los Datos

Roi Naveiro

# Análisis Exploratorio de los Datos

- Aprenderemos cómo explorar los datos de forma sistemática...
- ...usando las herramientas aprendidas (visualización y transformación)
- El Análisis Exploratorio de los datos es un ciclo que consiste en
  1. Generar preguntas sobre los datos
  2. Responderlas visualizando, transformando y modelizando
  3. Usar lo aprendido para refinar preguntas o crear nuevas cuestiones

# Análisis Exploratorio de los Datos

- Objetivo: desarrollar entendimiento acerca de los datos
- Útil hacerlo a través de preguntas
  - Permiten centrarnos en una parte de los datos
  - Permiten decidir qué gráficos, modelos y transformaciones utilizar
- Proceso **creativo**

# Análisis Exploratorio de los Datos

- Proceso **creativo**: no hay reglas para determinar qué preguntas son las más útiles.
- Dos tipos de preguntas (casi siempre) interesantes:
  - ¿Qué tipo de variación está presente en mis variables?
  - ¿Qué tipo de co-variación aparece entre mis variables?

# Un poco de jerga

- **Variable:** propiedad que puede ser medida
- **Valor:** estado de la variable cuando se mide. ¡Puede cambiar si se repite la medida!
- **Observación:** conjunto de medidas tomadas en condiciones similares (e.g. sobre un mismo sujeto). Varios valores, cada uno asociado a diferentes variables
- **Datos Tabulares:** conjunto de valores, cada uno asociado a una variable y una observación. Variables en columnas, observaciones en filas.

# EDA: Variación

# Variación

*Variación es la tendencia de los valores de una variable a cambiar de medida a medida*

Esta variación puede estar asociada a distintos fenómenos:

- **Error de medición** al medir una cantidad constante (velocidad de la luz)
- Medición de misma variable en **diferentes sujetos**
- Medición de una misma variable, en el mismo sujeto, en **tiempos diferentes**
- ...

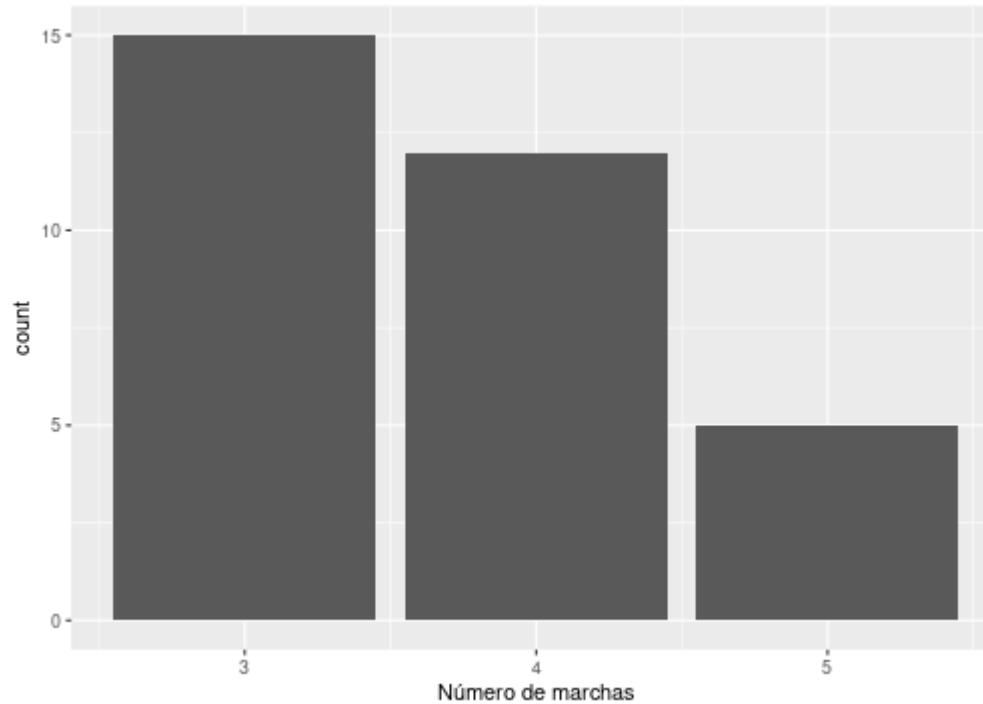
El patrón de variación da información interesante...

... que se destila visualizando la **distribución de valores de la variable**

# Variación: variable categórica

Recordemos cómo visualizar la distribución de una **variable categórica**

```
library(ggplot2)
# Representamos número de marchas en datos mtcars
ggplot(data=mtcars, aes(x=gear)) +
  geom_bar() +
  labs(x = "Número de marchas")
```



# Variación: variable categórica

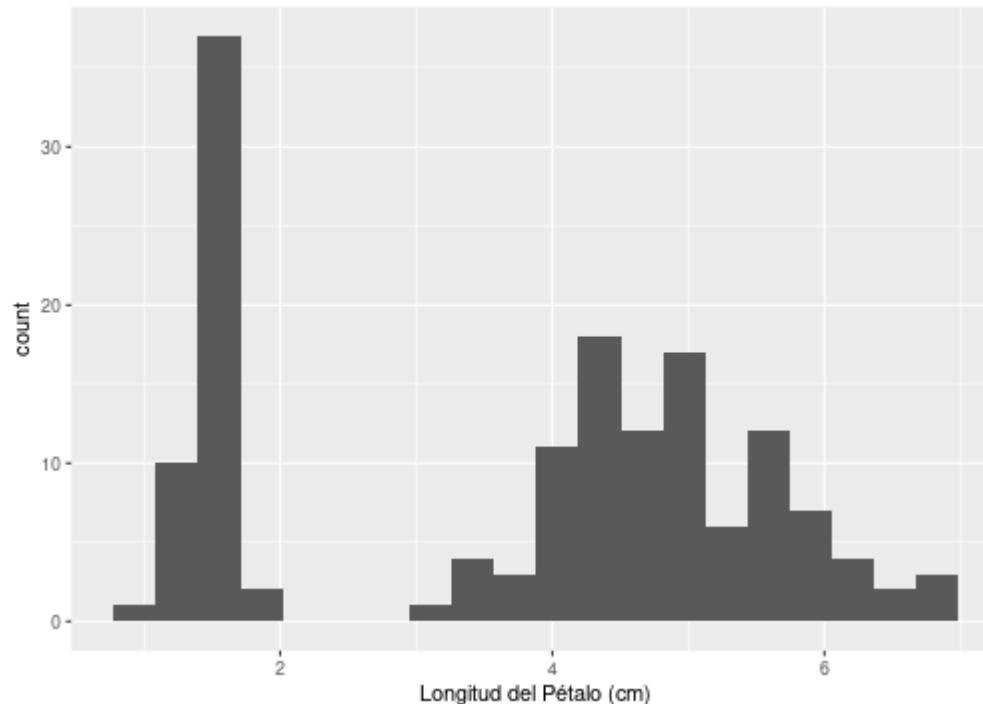
La altura de las barras muestran el número de observaciones. ¿Cómo obtendrías este número?

```
library(tidyverse)
mtcars %>% group_by(gear) %>% summarise(n())  
  
## # A tibble: 3 × 2
##       gear `n()`
##   <dbl> <int>
## 1     3     15
## 2     4     12
## 3     5      5
```

# Variación: variable continua

También habíamos visto cómo visualizar la distribución de variables continuas

```
ggplot(data=iris, aes(x=Petal.Length)) +  
  geom_histogram(bins = 20) +  
  labs(x = "Longitud del Pétalo (cm)")
```



# Variación

¿Qué debemos buscar en gráficos de variación? ¿Qué preguntas hacer?

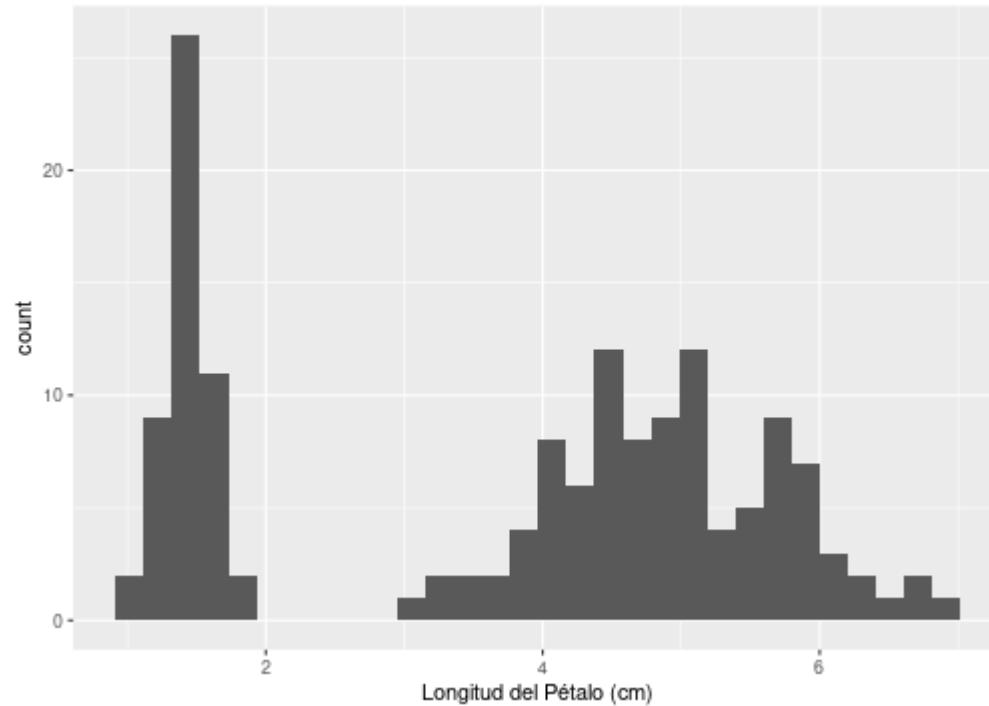
- Valores típicos
- Valores atípicos

# Variación: valores típicos

- ¿Qué valores son los más comunes?
- ¿Cuáles los menos comunes?
- ¿Existe algún patrón inusual?

# Ejemplo: longitud de pétalo

```
ggplot(data=iris, aes(x=Petal.Length)) +  
  geom_histogram() +  
  labs(x = "Longitud del Pétalo (cm)")
```



# Ejemplo: longitud de pétalo

Los *clusters* de valores similares sugieren que existen subgrupos en los datos. Para entenderlos, conviene que preguntarse

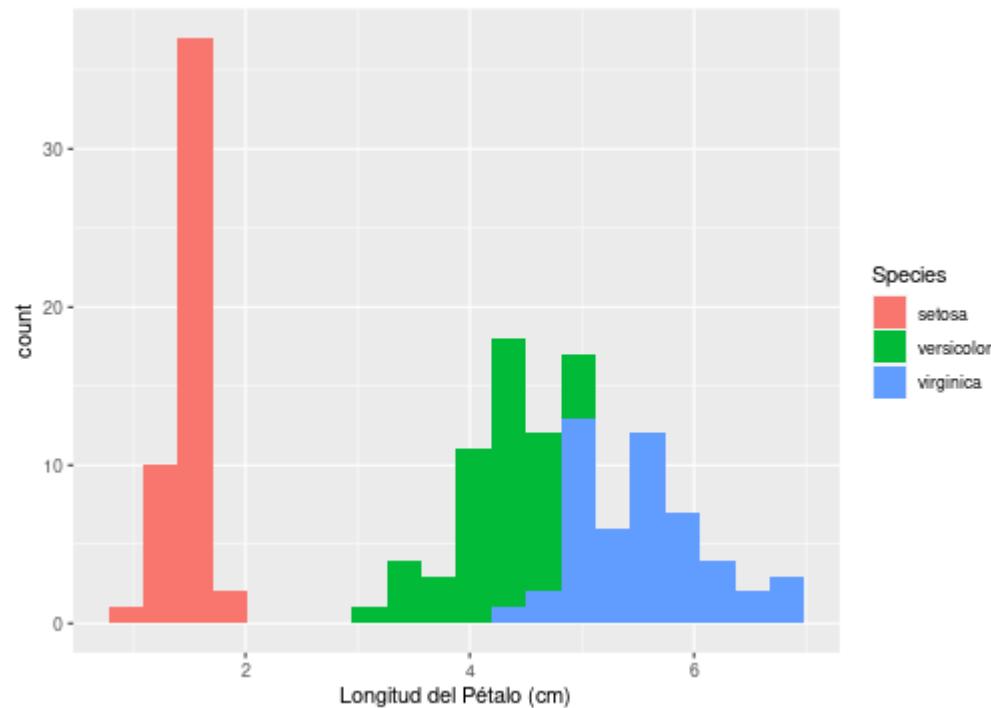
- ¿Qué tienen en común las observaciones de cada subgrupo?
- ¿Qué diferencia a las observaciones de distintos subgrupos?
- ¿Cómo podemos explicar los subgrupos?
- Existe un grupo de flores con pétalos cortos (<2cm) y otro con pétalos largos (>2cm)
- ¿Explica la variable especia esta separación?

# Variación: variable continua

- Si se quieren visualizar varios histogramas en la misma gráfica, se pueden usar colores
- A veces más claro usar **geom\_freqpoly()**
- Igual que histograma, pero pintado con líneas poligonales

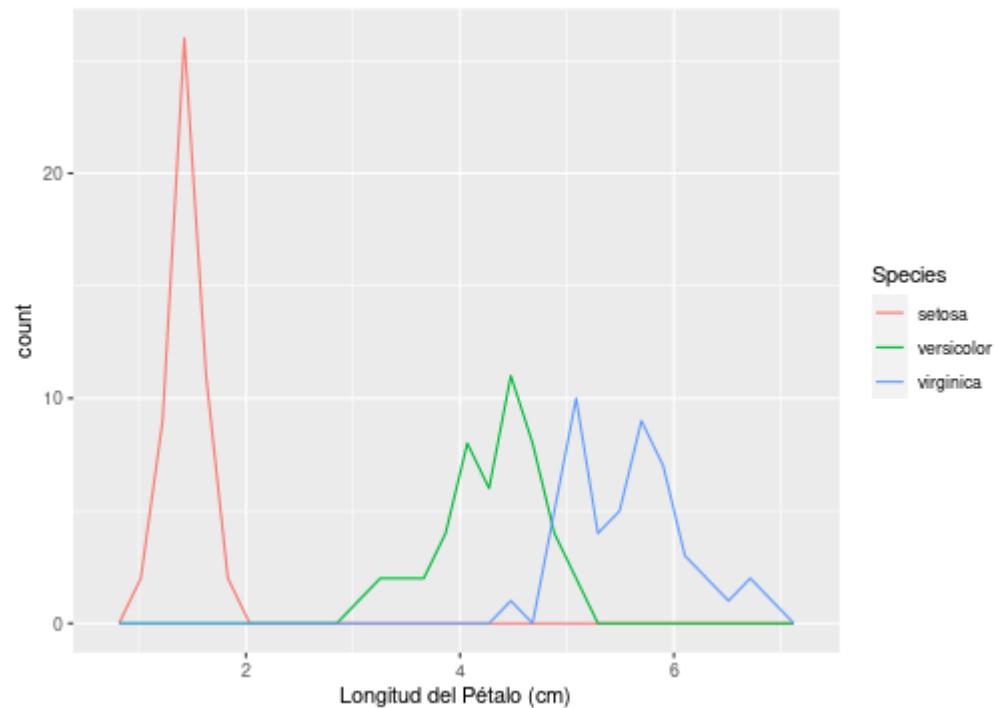
# Variación: variable continua

```
ggplot(data=iris, aes(x=Petal.Length, fill=Species)) +  
  geom_histogram(bins = 20) +  
  labs(x = "Longitud del Pétalo (cm)")
```



# Variación: variable continua

```
ggplot(data=iris, aes(x=Petal.Length, color=Species)) +  
  geom_freqpoly() +  
  labs(x = "Longitud del Pétalo (cm)")
```



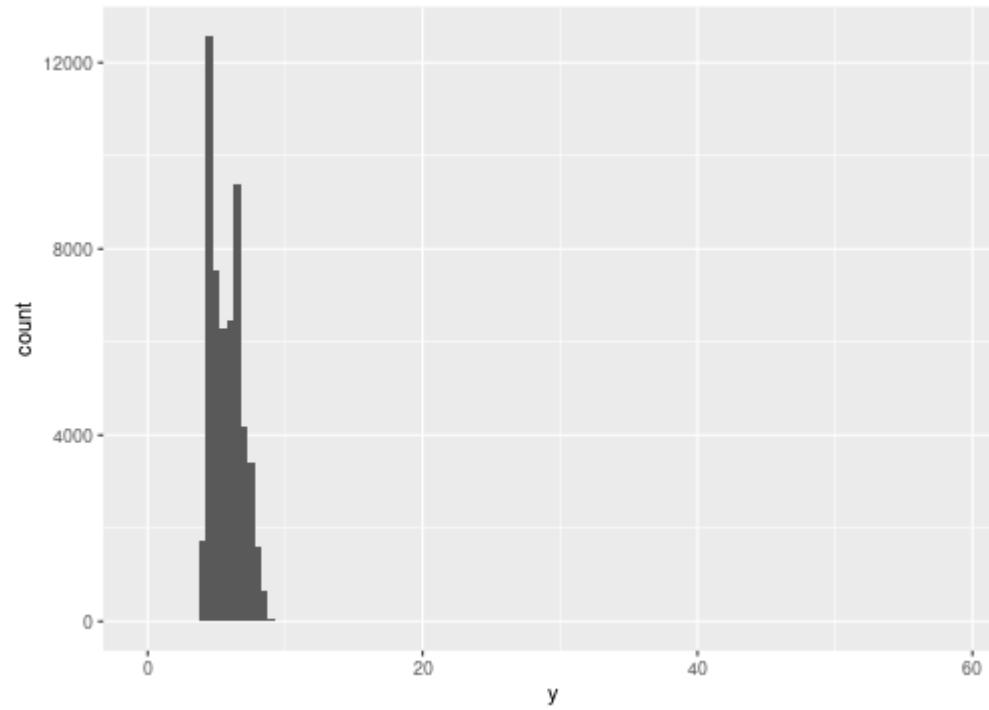
# Variación: valores atípicos

- Encontrar **valores atípicos** (outliers) Tan importante como estudiar los valores típicos
- Valores atípicos:
  - A veces, errores de medida
  - Otras, importantes descubrimientos!

# Ejemplo: anchura de los diamantes

Estudiemos la distribución de la anchura de los diamantes en el dataset **diamonds** (más información en `?diamonds`)

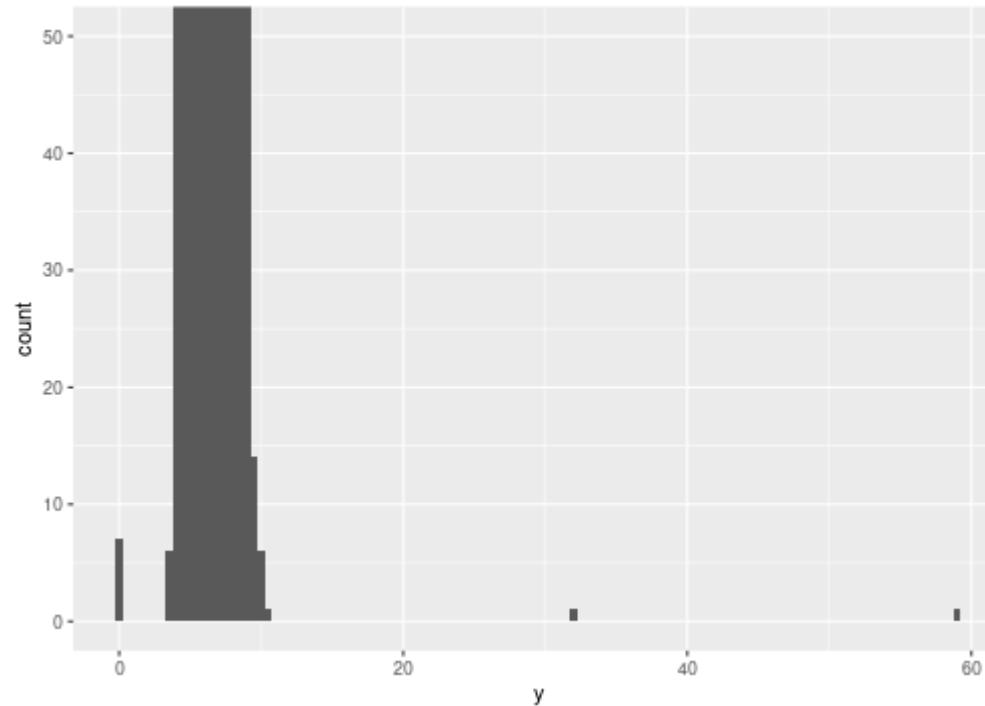
```
ggplot(diamonds, aes(x=y)) +  
  geom_histogram(binwidth = 0.5)
```



# Ejemplo: anchura de los diamantes

Aparentemente nada raro, pero...

```
ggplot(diamonds, aes(x=y)) +  
  geom_histogram(binwidth = 0.5) +  
  coord_cartesian( ylim = c(0,50) )
```



# Ejemplo: anchura de los diamantes

Parece ser que hay valores atípicos en torno al 0, 30 y 60.

Vamos a extraerlos con **filter**

```
diamonds %>%
  filter(y < 3 | y > 20)
```

```
## # A tibble: 9 × 10
##   carat cut      color clarity depth table price     x     y     z
##   <dbl> <ord>    <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1 1     Very Good H     VS2     63.3    53  5139    0     0     0
## 2 1.14 Fair       G     VS1     57.5    67  6381    0     0     0
## 3 2     Premium   H     SI2     58.9    57 12210  8.09  58.9  8.06
## 4 1.56 Ideal      G     VS2     62.2    54 12800    0     0     0
## 5 1.2   Premium   D     VVS1    62.1    59 15686    0     0     0
## 6 2.25 Premium   H     SI2     62.8    59 18034    0     0     0
## 7 0.51 Ideal      E     VS1     61.8    55 2075   5.15  31.8  5.12
## 8 0.71 Good       F     SI2     64.1    60 2130    0     0     0
## 9 0.71 Good       F     SI2     64.1    60 2130    0     0     0
```

# Ejemplo: anchura de los diamantes

- Parece que hay diamantes de medida 0. Datos incorrectos.
- Hay diamantes muy grandes, pero no demasiado caros...

# Variación: valores atípicos

- Suele ser útil repetir análisis con y sin outliers, para medir su efecto
- Si el efecto es mínimo y no sabemos su origen, reemplazar con **NA**
- Si no, solo quitarlos de manera **justificada** e informar de ello

# EDA: Valores ausentes

# Valores atípicos

Ante valores atípicos que se quieran eliminar, dos opciones:

- Eliminar fila entera (mucho pérdida de información)

```
diamonds2 <- diamonds %>%
  filter(y>=3 & y<=20)
```

- Reemplazar valores atípicos con **NA**

```
diamonds2 <- diamonds %>%
  mutate(y = ifelse(y<3 | y>20, NA, y))
```

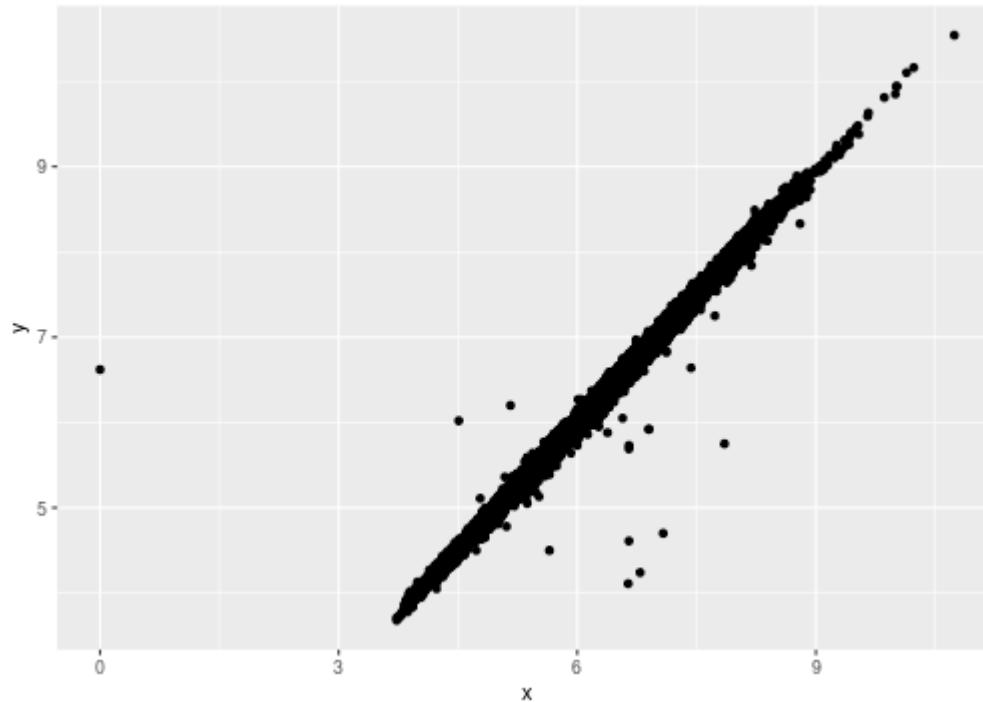
- NOTA: primer argumento de **ifelse()** es lógico, segundo valor si **TRUE** y tercero, valor si **FALSE**

# Valores ausentes

Los valores ausentes se indican con un warning

```
ggplot(data=diamonds2, mapping = aes(x=x, y=y)) +  
  geom_point()
```

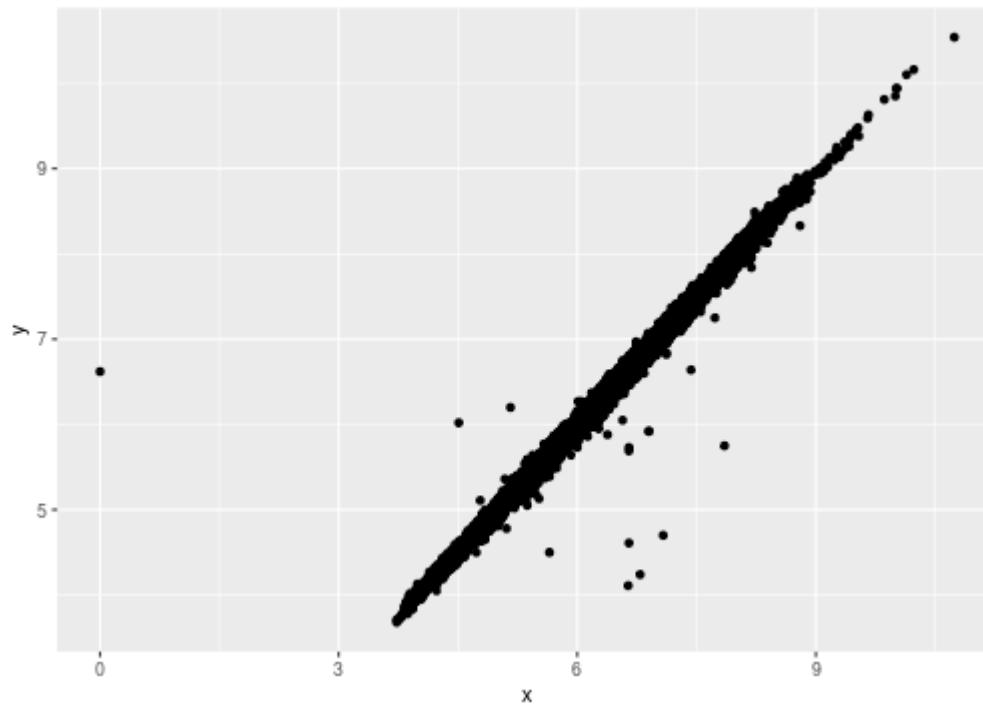
```
## Warning: Removed 9 rows containing missing values (geom_point).
```



# Valores ausentes

Puede eliminarse con **na.rm = TRUE**

```
ggplot(data=diamonds2, mapping = aes(x=x, y=y)) +  
  geom_point(na.rm = T)
```



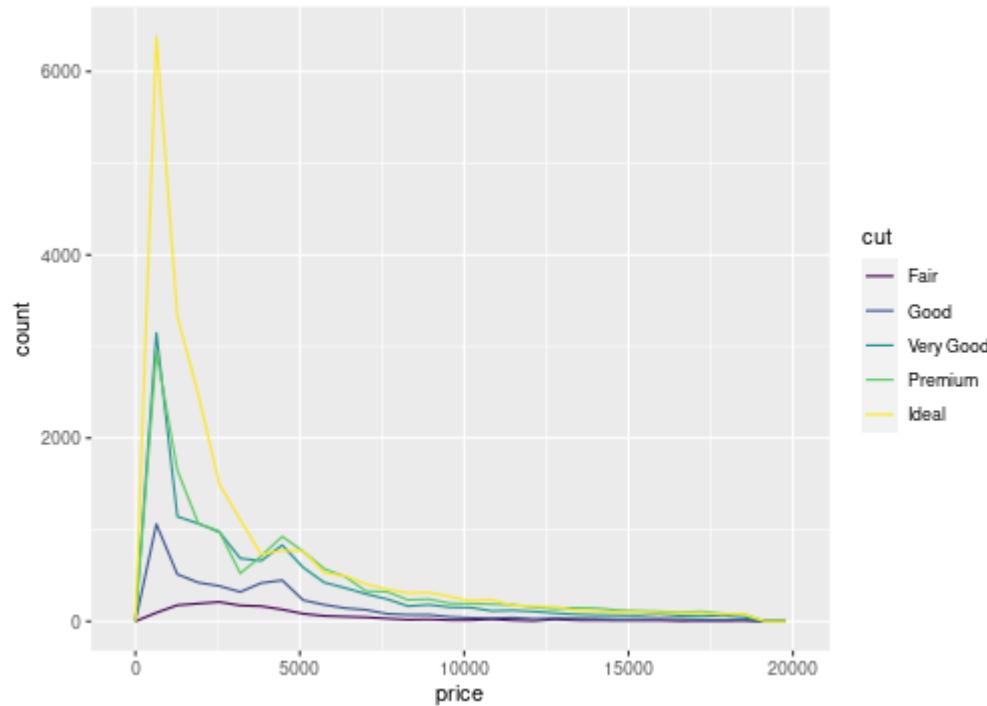
# EDA: Covariación

# Covariación

- Tendencia de los valores de dos o más variables a cambiar conjuntamente
- La manera de visualizar covariación, depende del tipo de variables (como vimos)
  - Categórica - Contínua
  - Categórica - Categórica
  - Contínua - Contínua

# Covariación: Categórica - Contínua

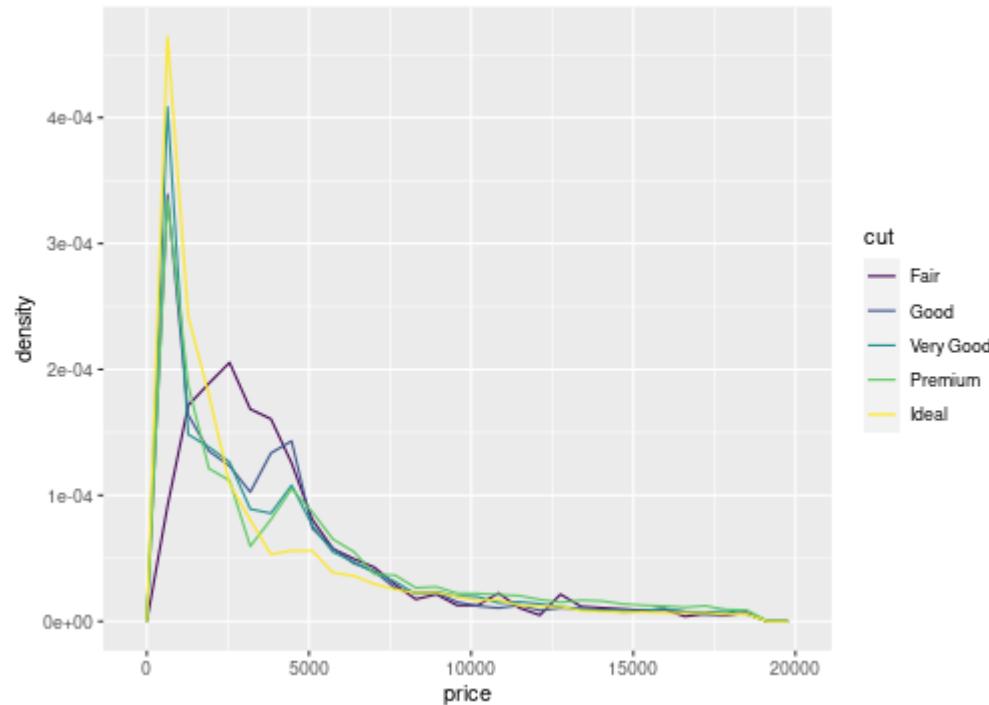
```
ggplot(data=diamonds, aes(x=price, color=cut)) +  
  geom_freqpoly()
```



# Covariación: Categórica - Contínua

..density.. pinta densidad en eje y, que es la cuenta estandarizada para que el área de cada polígono sea 1

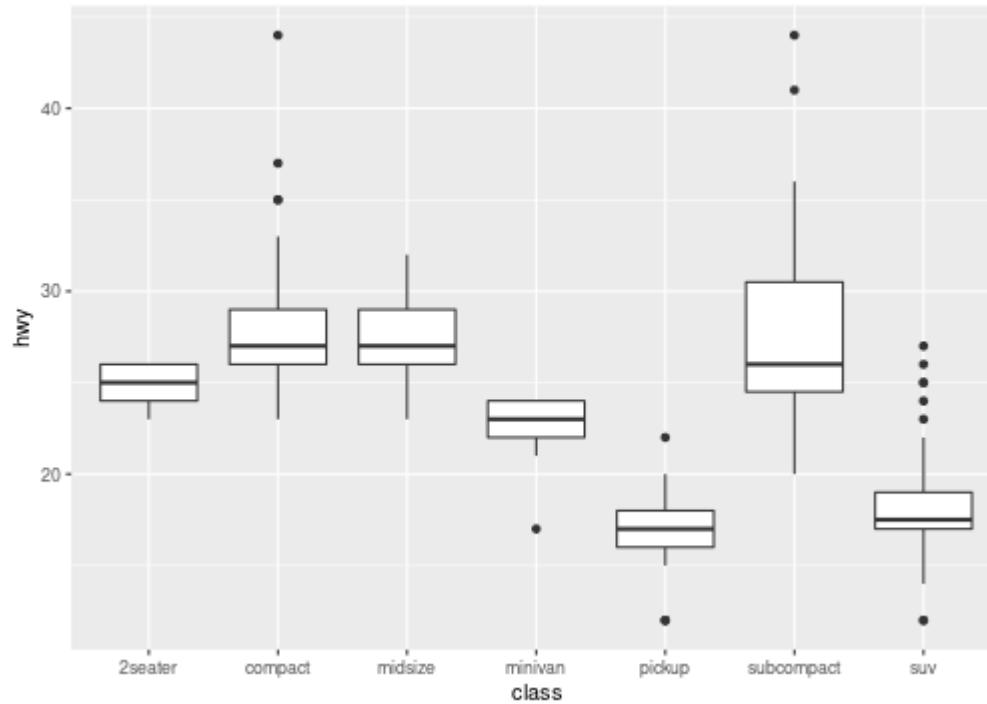
```
ggplot(data=diamonds, aes(x=price, y=..density.., color=cut)) +  
  geom_freqpoly()
```



# Covariación: Categórica - Contínua

¿Qué otra forma hay de visualizar covariación entre variables categóricas y continuas?

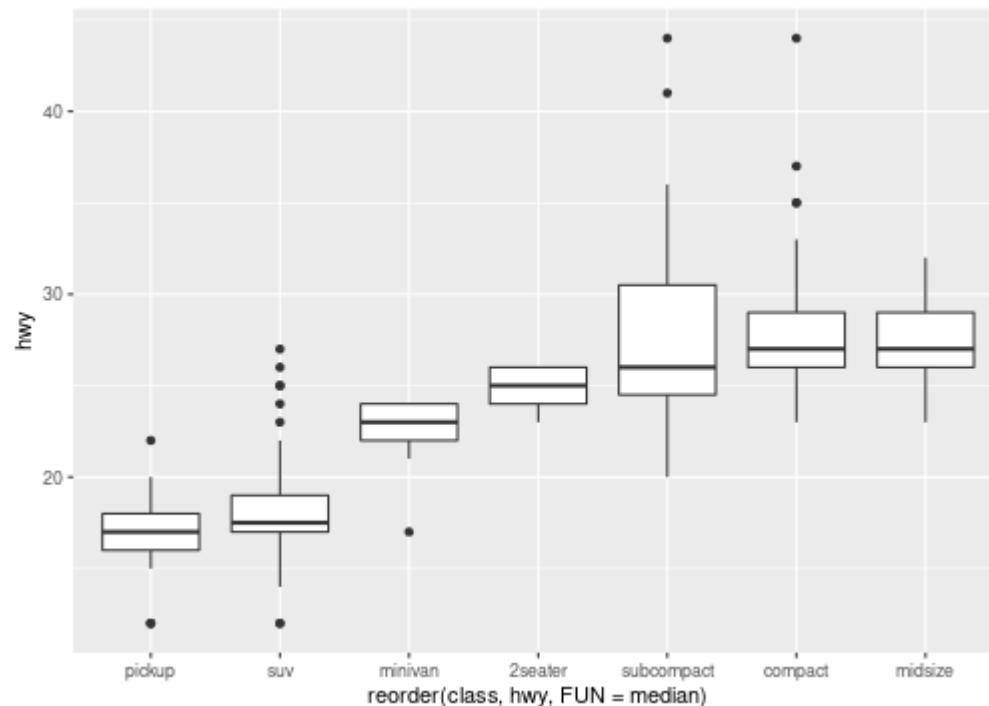
```
ggplot(data = mpg, aes(x = class, y = hwy)) +  
  geom_boxplot()
```



# Covariación: Categórica - Contínua

A veces, conviene reordenar...

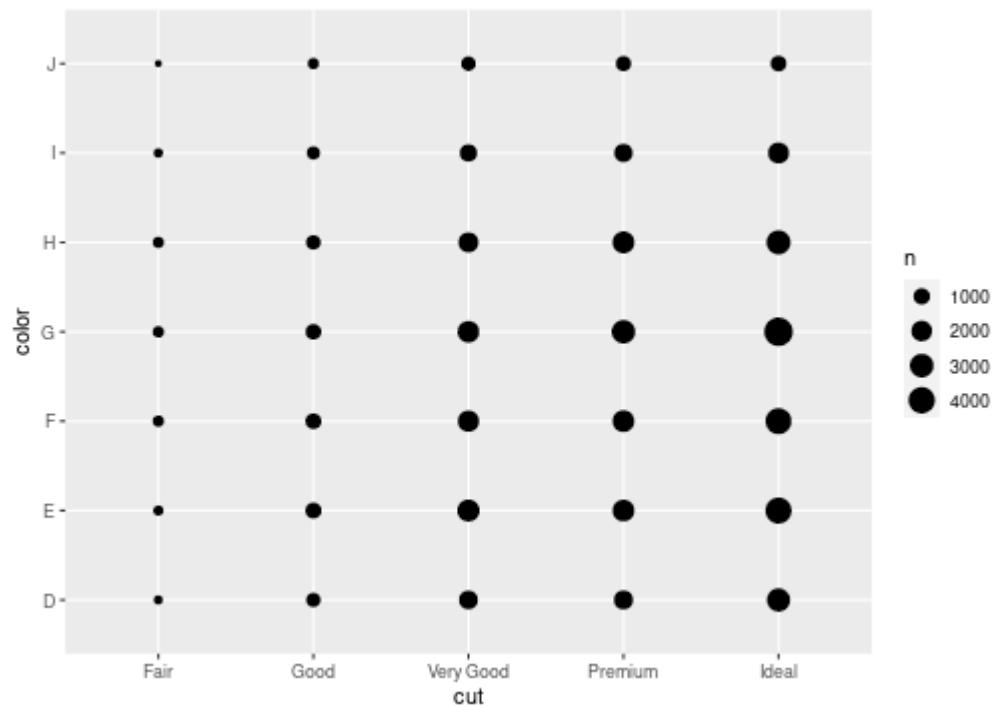
```
ggplot(data = mpg, aes(x = reorder(class, hwy, FUN = median), y = hwy)) +  
  geom_boxplot()
```



# Covariación: Categórica - Categórica

Habíamos visto diagramas de barras. Otra forma:

```
ggplot(data = diamonds, aes(x = cut, y = color)) +  
  geom_count()
```



# Covariación: Categórica - Categórica

¿Cómo harías esto con **dplyr**?

```
diamonds %>%
  group_by(color, cut) %>%
  summarise(n = n())
```

```
## # A tibble: 35 × 3
## # Groups:   color [7]
##       color   cut     n
##       <ord> <ord> <int>
## 1 D      Fair    163
## 2 D      Good    662
## 3 D      Very Good 1513
## 4 D      Premium 1603
## 5 D      Ideal   2834
## 6 E      Fair    224
## 7 E      Good    933
## 8 E      Very Good 2400
## 9 E      Premium 2337
## 10 E     Ideal   3903
## # ... with 25 more rows
```

# Covariación: Categórica - Categórica

También sirve

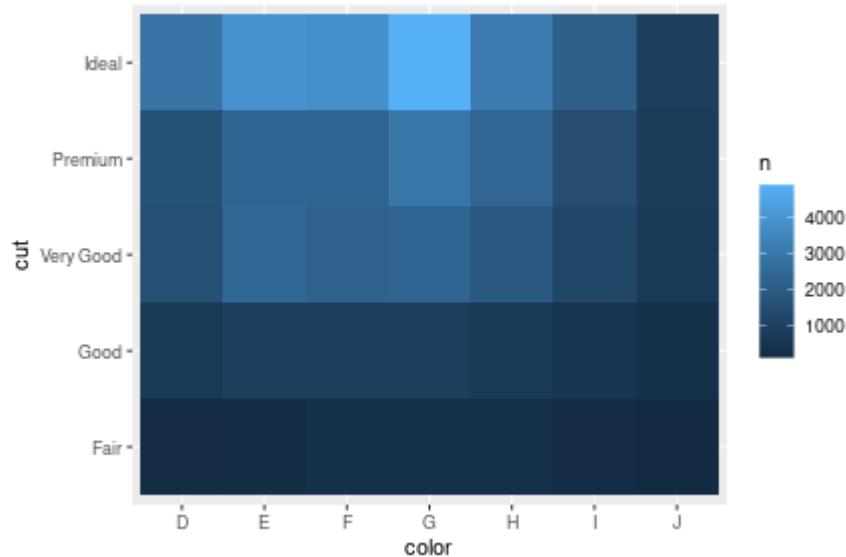
```
library(dplyr)
diamonds %>%
  count(color, cut)

## # A tibble: 35 × 3
##   color     cut      n
##   <ord> <ord>    <int>
## 1 D       Fair     163
## 2 D       Good     662
## 3 D       Very Good 1513
## 4 D       Premium  1603
## 5 D       Ideal    2834
## 6 E       Fair     224
## 7 E       Good     933
## 8 E       Very Good 2400
## 9 E       Premium  2337
## 10 E      Ideal    3903
## # ... with 25 more rows
```

# Covariación: Categórica - Categórica

Con esto

```
iamonds %>%
  count(color, cut) %>%
  ggplot(aes(x = color, y = cut, fill = n))+
  geom_tile()
```

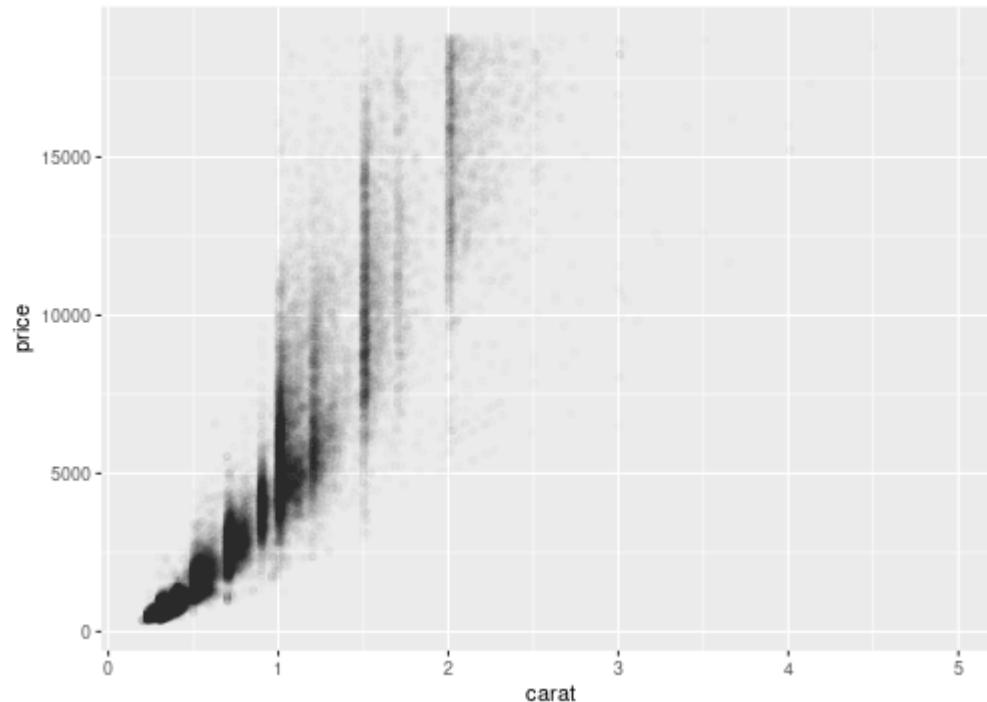


# Covariación: Contínua - Contínua

- Vimos cómo hacerlo con **geom\_point()**
- En datasets muy grandes, puntos solapan
- ¿Idea? Usar **alpha**

# Covariación: Contínua - Contínua

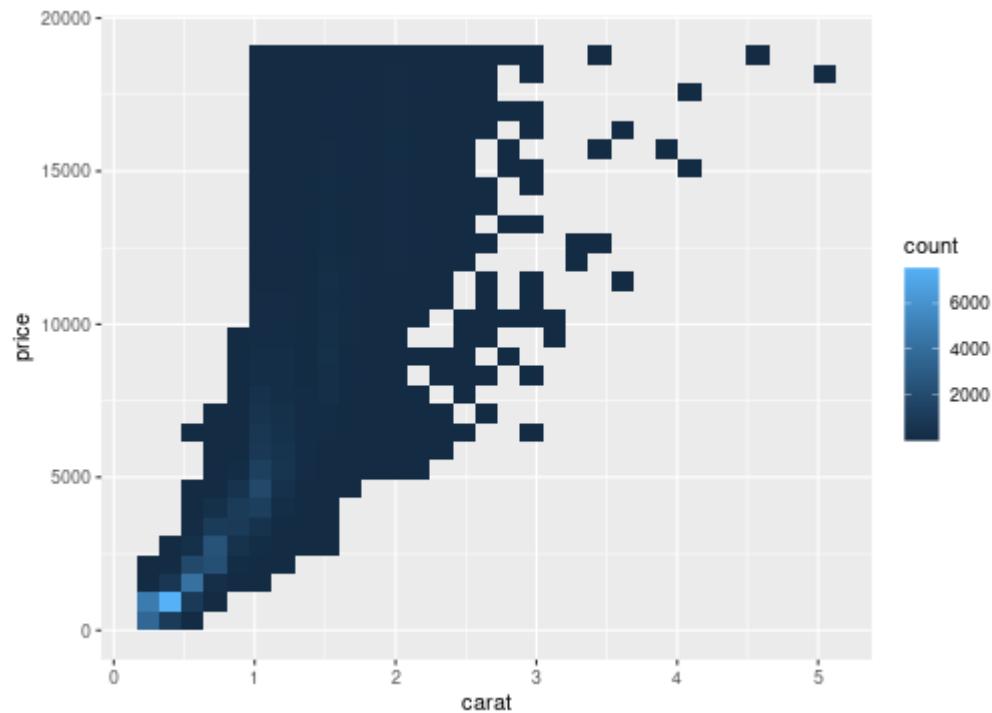
```
ggplot(data=diamonds, aes(x = carat, y = price)) +  
  geom_point(alpha = 0.01)
```



# Covariación: Contínua - Contínua

Mejor

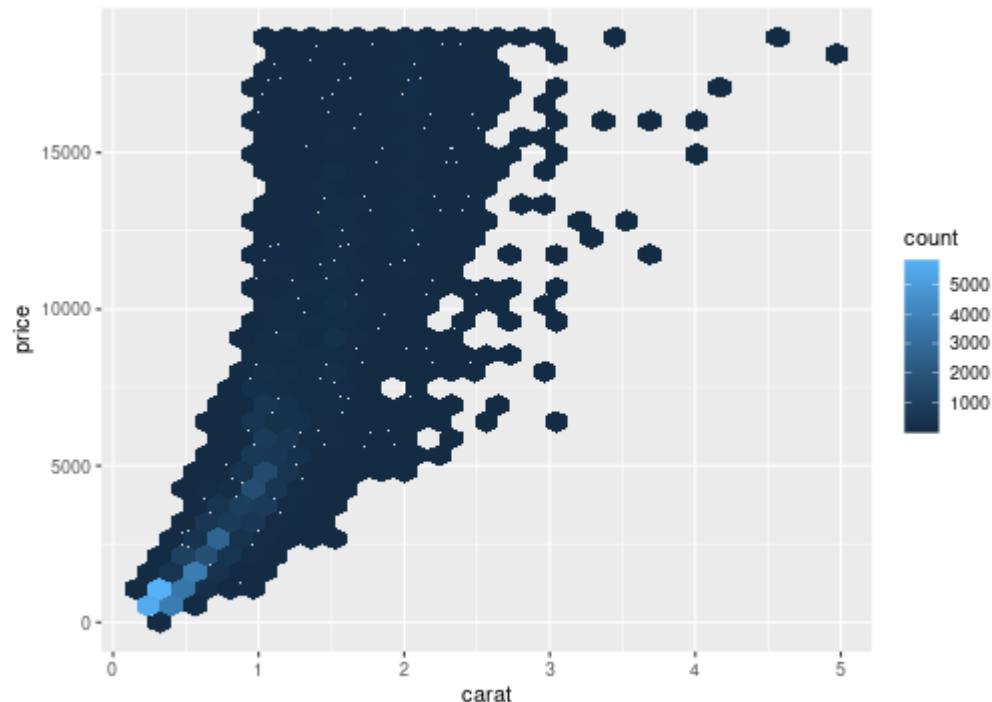
```
ggplot(data=diamonds, aes(x = carat, y = price)) +  
  geom_bin2d()
```



# Covariación: Contínua - Contínua

Mejor

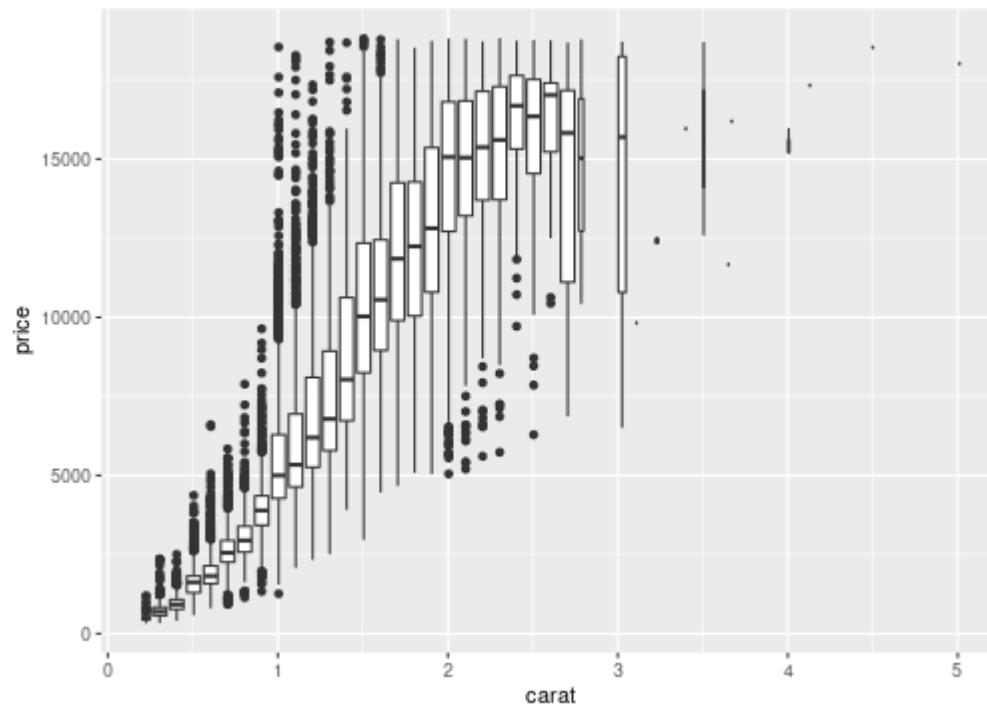
```
# install.packages("hexbin")
ggplot(data=diamonds, aes(x = carat, y = price)) +
  geom_hex()
```



# Covariación: Contínua - Contínua

Otra opción

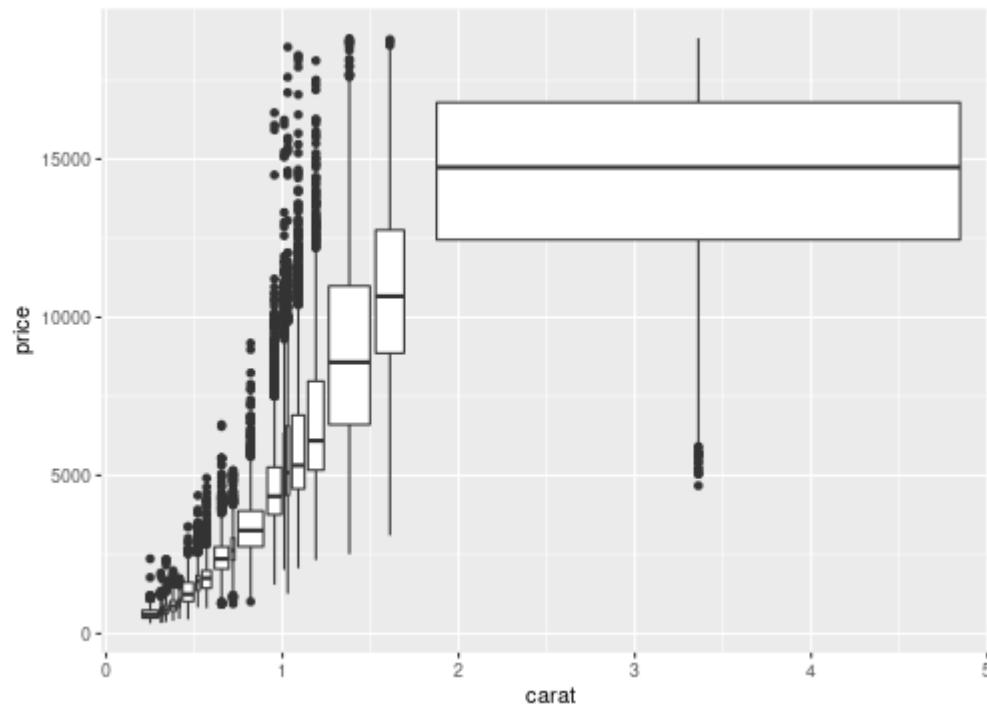
```
ggplot(data=diamonds, aes(x = carat, y = price, )) +  
  geom_boxplot(aes(group = cut_width(carat, 0.1)) )
```



# Covariación: Contínua - Contínua

Otra opción (mismo número de puntos por cada bin)

```
ggplot(data=diamonds, aes(x = carat, y = price, )) +  
  geom_boxplot(aes(group = cut_number(carat, 20)) )
```



# EDA: Patrones y modelos

# Patrones y modelos

Los patrones revelan pistas acerca de relaciones entre variables. Ante uno, cabe preguntarse

- ¿Puede ser debido al azar?
- ¿Cómo se describe la relación que implica?
- ¿Cómo es de fuerte?
- ¿Qué otras variables afectan esta relación?
- ¿Cambian la relación si observamos subgrupos de los datos?

Los **modelos** sirven para extraer patrones de los datos...

... e.g. pueden usarse para eliminar el efecto de una variable y estudiar el efecto restante.

# Bibliografía

- R for Data Science, Wickham and Grolemund (2016)