

# Exploracion de Datos

*Juan Salamanca*

*30 de marzo de 2017*

```
library(ggplot2)
library(knitr)
```

Nota: Se omiten acentos por compatibilidad con sistema operativo

Cargar librerías

El dataset a utilizar esta en la librería **ggplot2** La descripción general del dataset es:

---

## Prices of 50,000 round cut diamonds

### Format

A data frame with 53940 rows and 10 variables

### Description

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

### Details

price. price in US dollars (\$326–\$18,823) carat. weight of the diamond (0.2–5.01) cut. quality of the cut (Fair, Good, Very Good, Premium, Ideal) colour. diamond colour, from J (worst) to D (best) clarity. a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best)) x. length in mm (0–10.74) y. width in mm (0–58.9) z. depth in mm (0–31.8) depth. total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43–79) table. width of top of diamond relative to widest point (43–95)

```
# descripción general del dataset
str(diamonds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame': 53940 obs. of 10 variables:
## $ carat   : num 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
## $ cut      : Ord.factor w/ 5 levels "Fair" < "Good" < ...: 5 4 2 4 2 3 3 3 1 3 ...
## $ color    : Ord.factor w/ 7 levels "D" < "E" < "F" < "G" < ...: 2 2 2 6 7 7 6 5 2 5 ...
## $ clarity  : Ord.factor w/ 8 levels "I1" < "SI2" < "SI1" < ...: 2 3 5 4 2 6 7 3 4 5 ...
## $ depth    : num 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
## $ table   : num 55 61 65 58 58 57 57 55 61 61 ...
## $ price   : int 326 326 327 334 335 336 336 337 337 338 ...
## $ x       : num 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
## $ y       : num 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
## $ z       : num 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

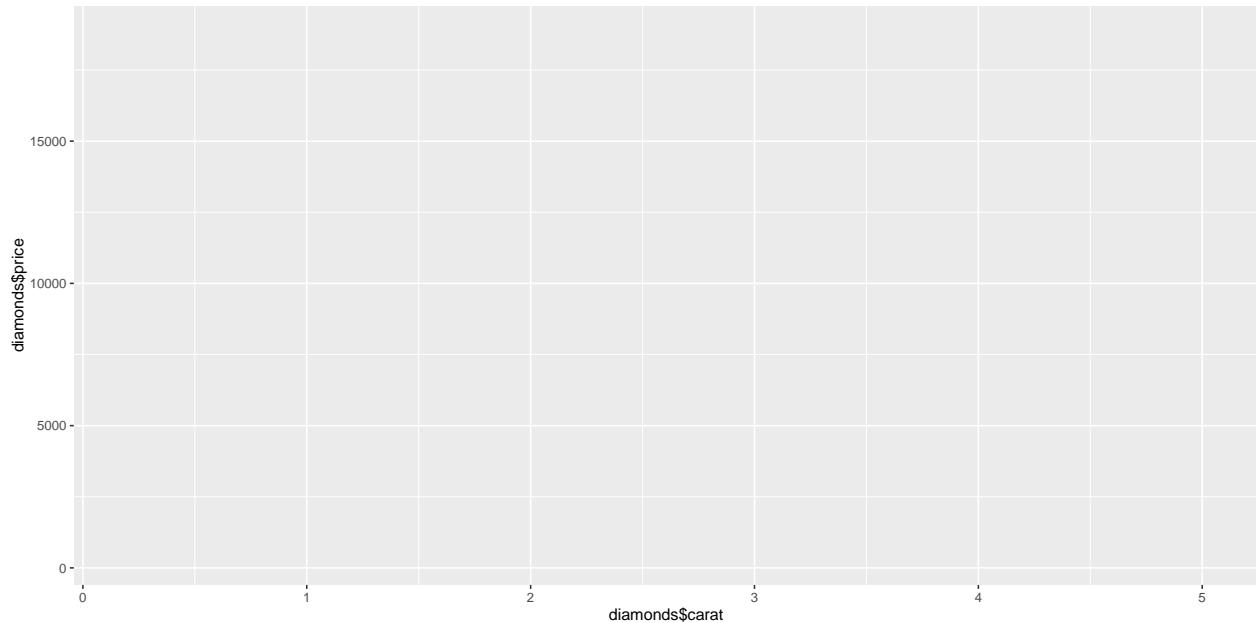
---

Qué tipos de datos encuentra en el dataset?

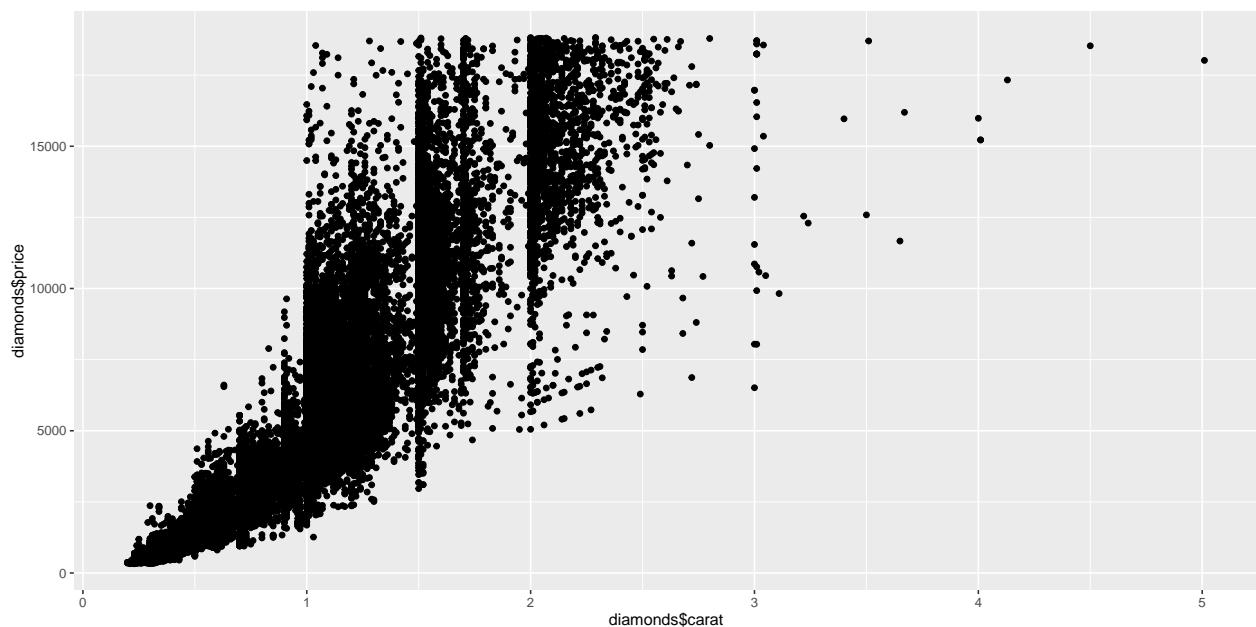
## Diagrama de dispersion (Scatter Plot)

Muestra la relacion entre de dos variables numericas

```
# Revise el lienzo vacio y defina los datos que quiere asignar a X y Y
grafica <- ggplot(diamonds, aes(x=diamonds$carat, y=diamonds$price))
grafica
```

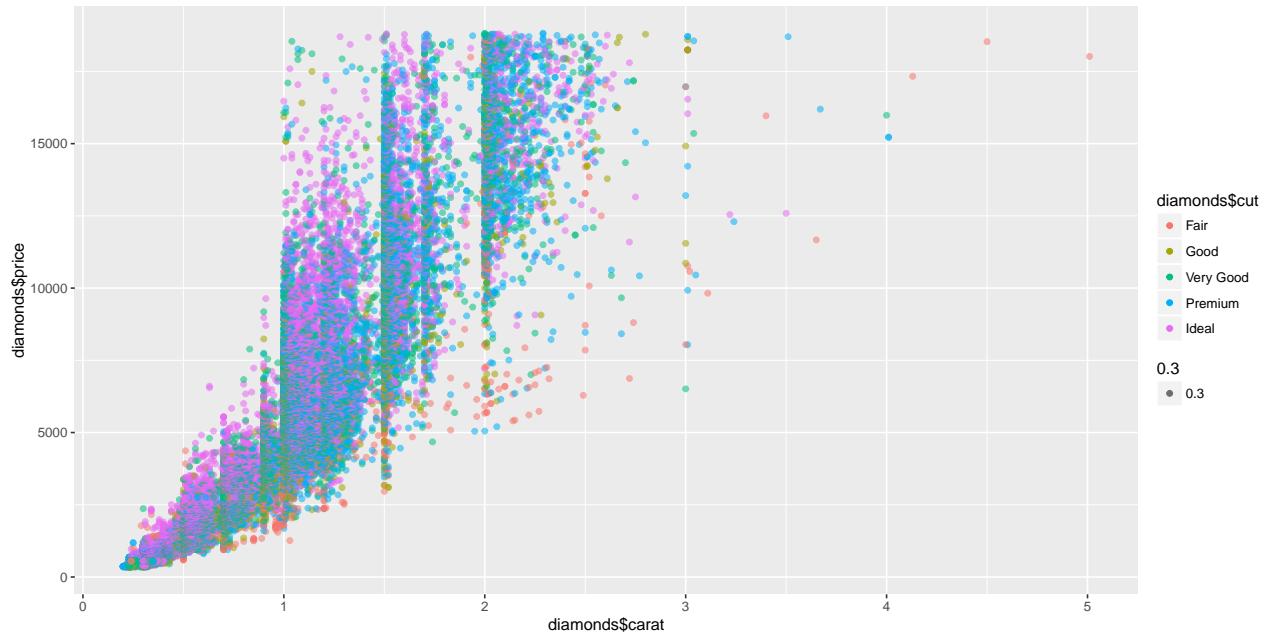


```
# asigne el tipo de geometria a usar
grafica <- grafica + geom_point()
# muestre el resultado
grafica
```

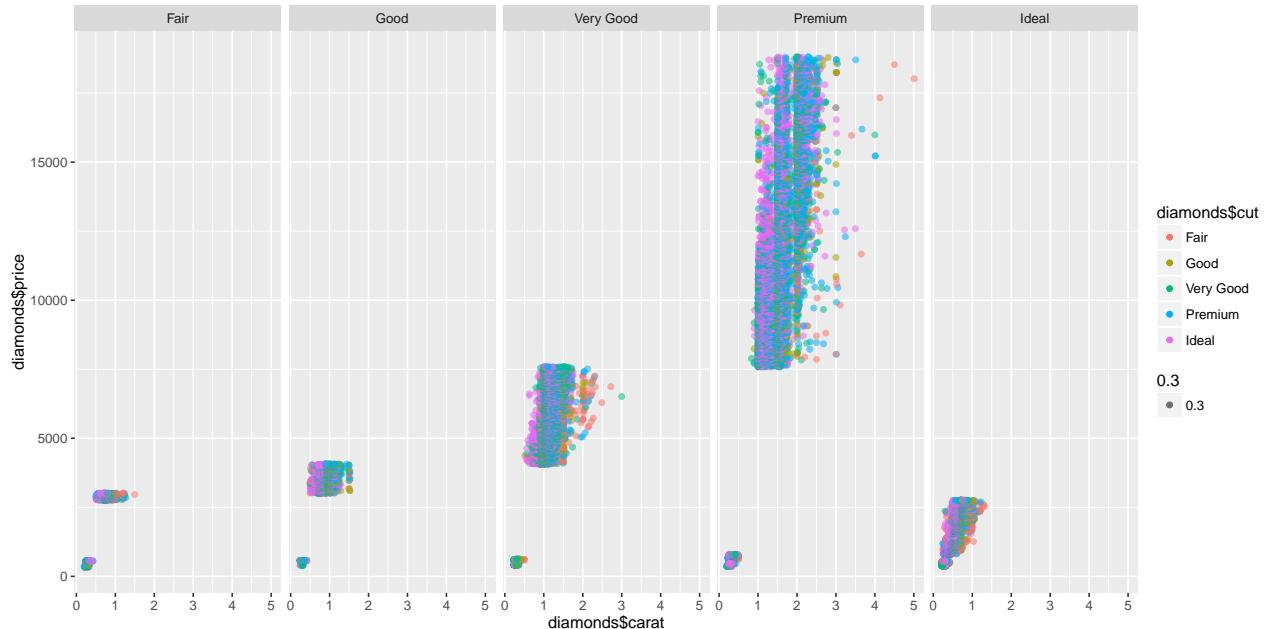


Encuentre patrones con la variables categóricas del dataset

```
# cree nuevamente el objeto ggplot con el dataset a mostrar. Si no lo crea nuevamente va a adicionar una grafica <- ggplot(diamonds, aes(x=diamonds$carat, y=diamonds$price))
# Asigne el campo del dataset que quiere usar para asignar color
grafica <- grafica + geom_point(aes(color = diamonds$cut, alpha = 0.3))
# muestre el resultado
grafica
```



Que atributo categorico tiene mayor relacion con el precio? Si hay dos parecidos que tipo de grafica y con que campos en cada ejecree que podria realizar para responder esta pregunta con mas precision? Por ejemplo algo asi:

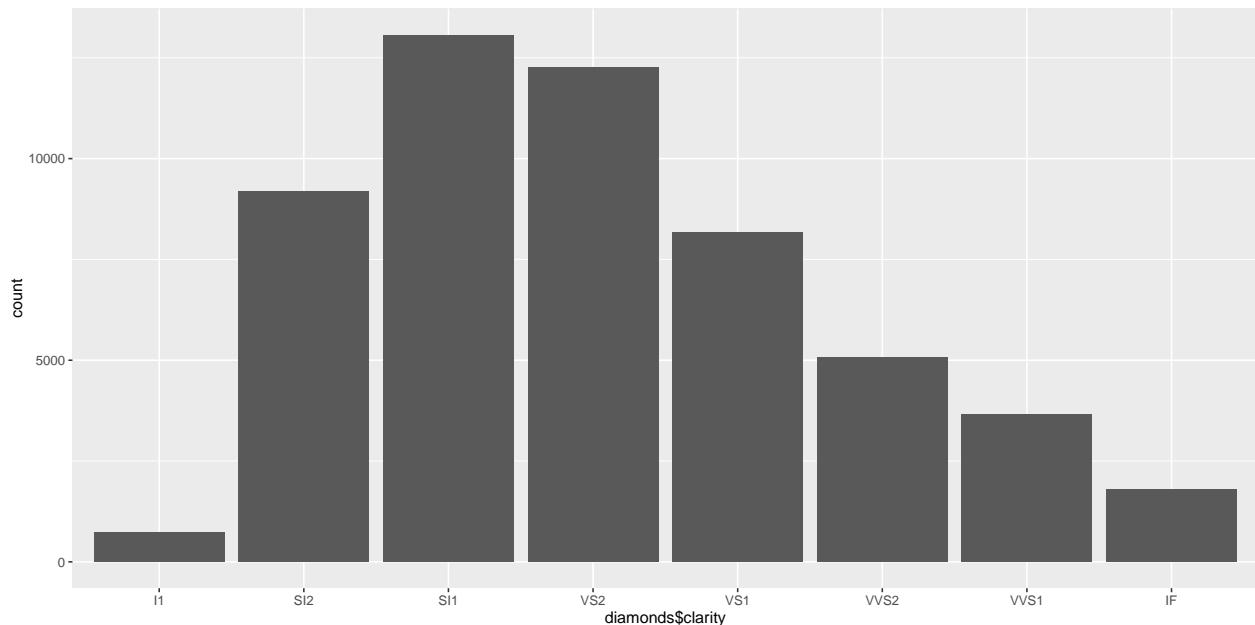


## Diagrama de barras (Bar)

Muestra la cuenta de casos por categorías. Similar a un histograma pero con valores categóricos (discretos) en el eje X. Note que no debe asignar valores en Y porque se calculan automáticamente

```
# Datos

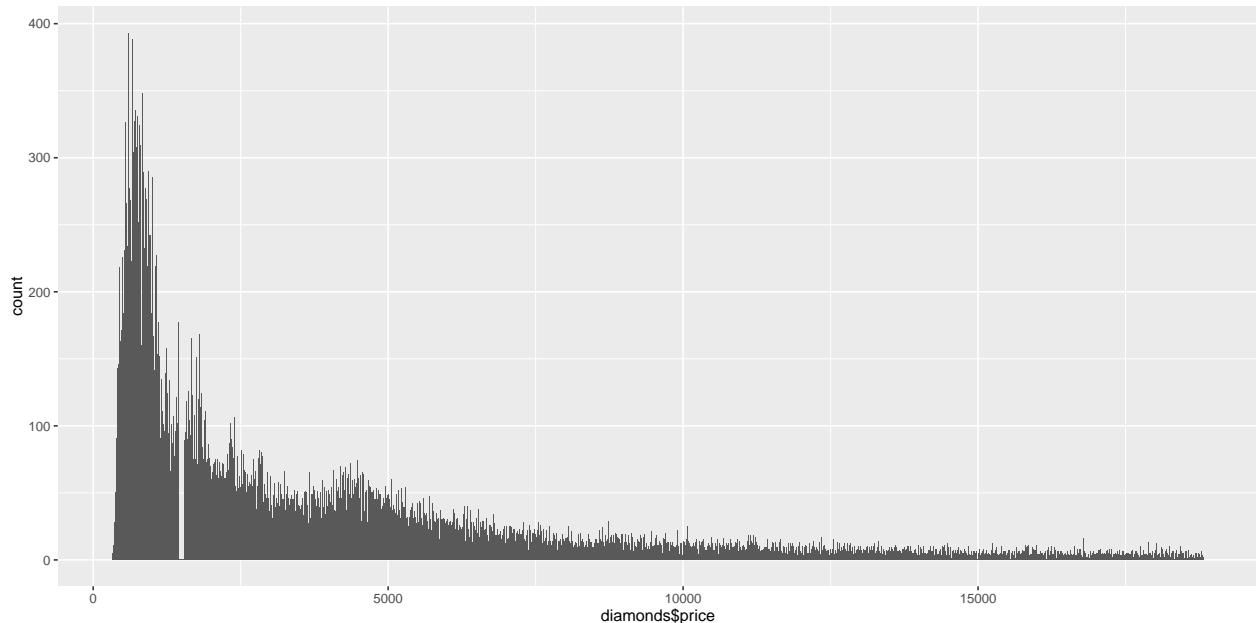
```



## Histograma (Histogram)

Muestra la distribución por contenedores calculados a partir de una variable continua. Note que no debe asignar valores en Y porque se calculan automáticamente

```
# Graficamente
grafica <- ggplot(diamonds, aes(x=diamonds$price))
# Cambie el valor del ancho de la barra.
grafica <- grafica + geom_histogram(binwidth = 10)
grafica
```

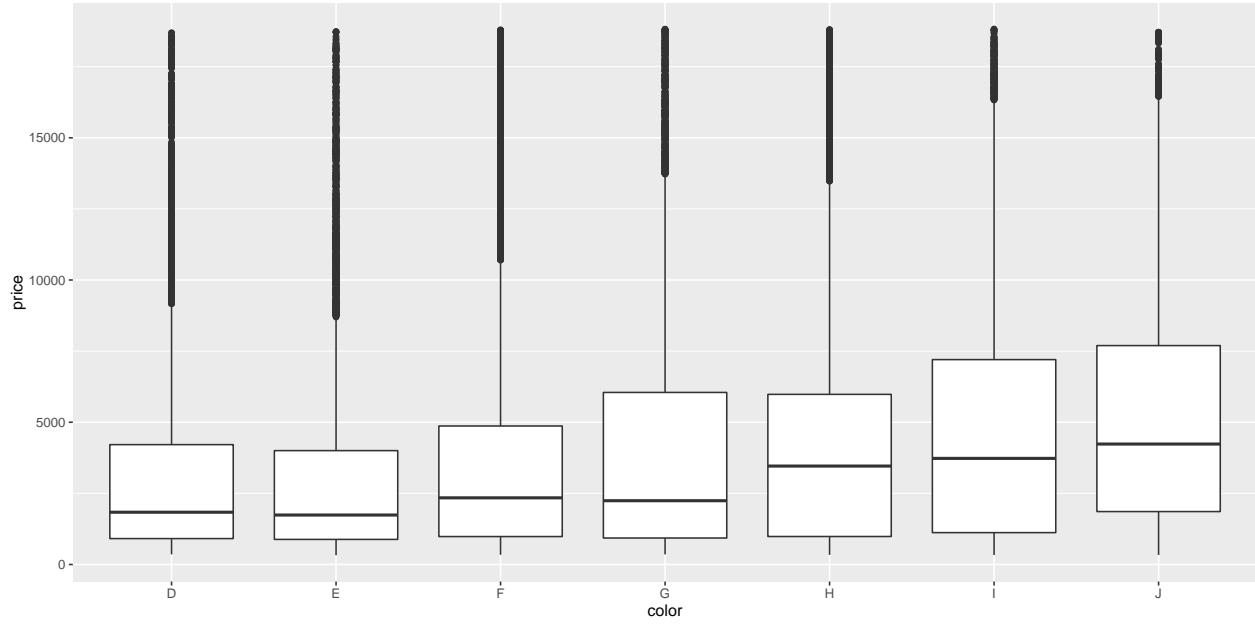


## Diagrama de Caja y Bigote (Box Plot)

Muestra una distribución de la siguiente forma: [https://en.wikipedia.org/wiki/Box\\_plot#/media/File:Boxplot\\_vs\\_PDF.svg](https://en.wikipedia.org/wiki/Box_plot#/media/File:Boxplot_vs_PDF.svg)

La mediana en el medio, el valor mínimo de Q1 y el máximo de Q3 definen el rango intercuartil. Los bigotes corresponden a +/- 1.5 veces el rango intercuartil. Por fuera de estos límites están los valores atípicos.

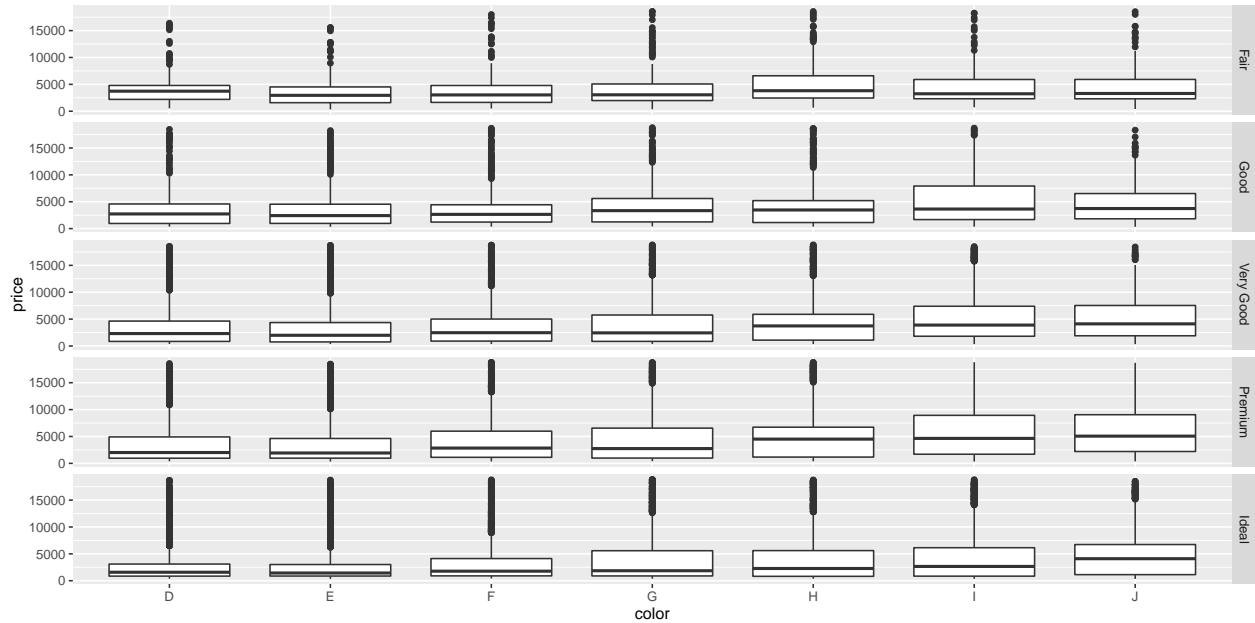
```
# Cree un lienzo con doc coordenadas, en el eje X debe ser catgorica
grafica <- ggplot(diamonds, aes(x=color, y=price))
# Haga el boxplot
grafica <- grafica + geom_boxplot()
grafica
```



Esta gráfica muestra muchos *outliers* o valores atípicos. Que interpreta de esto?

Ahora veamos boxplots relacionando dos variables categóricas.

```
# utilice facetas para ver todas las opciones . Debe usar variables categoricas.
grafica <- grafica + facet_grid(cut~.)
grafica
```



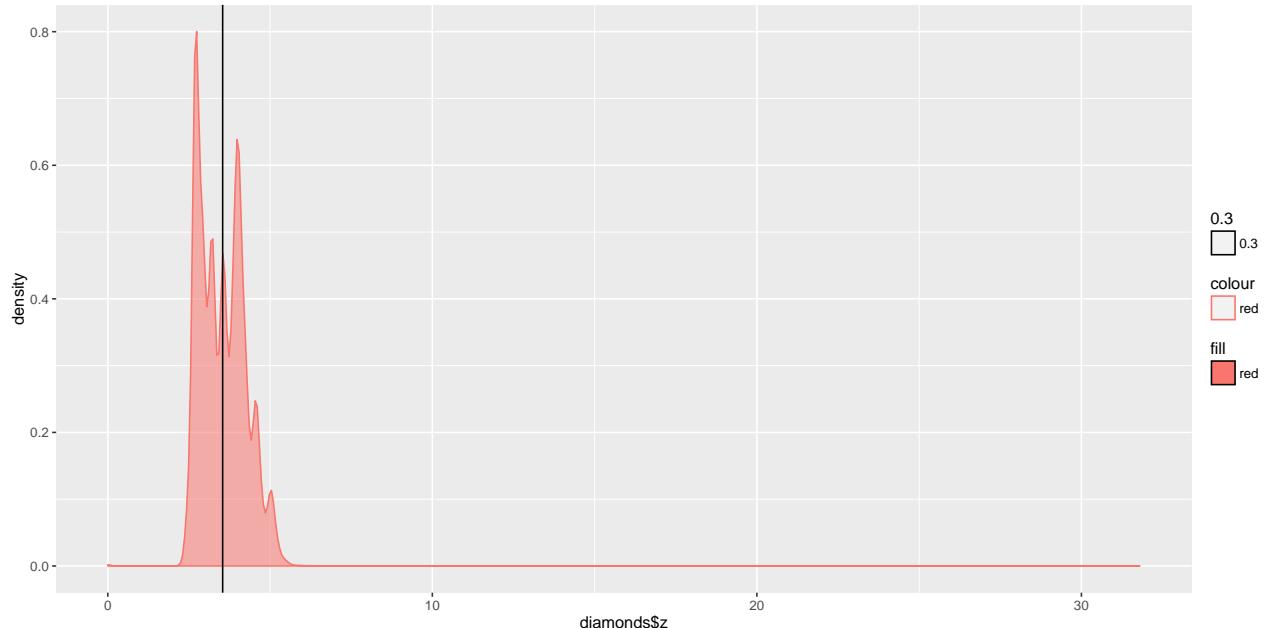
## Curva de densidades (density)

Muestra una distribución similar al histograma pero en forma de curva continua

```

# Calcule el promedio
promZ <- mean(diamonds$z)
# Cree el lienzo
grafica <- ggplot(diamonds, aes(x=diamonds$z, fill="red", color = "red", alpha = 0.3))
# agrege la curva de desnsidad con color
grafica <- grafica + geom_density()
# Linea vertical del precio
grafica <- grafica + geom_vline(xintercept=promZ)
# Muestre el resultado
grafica

```



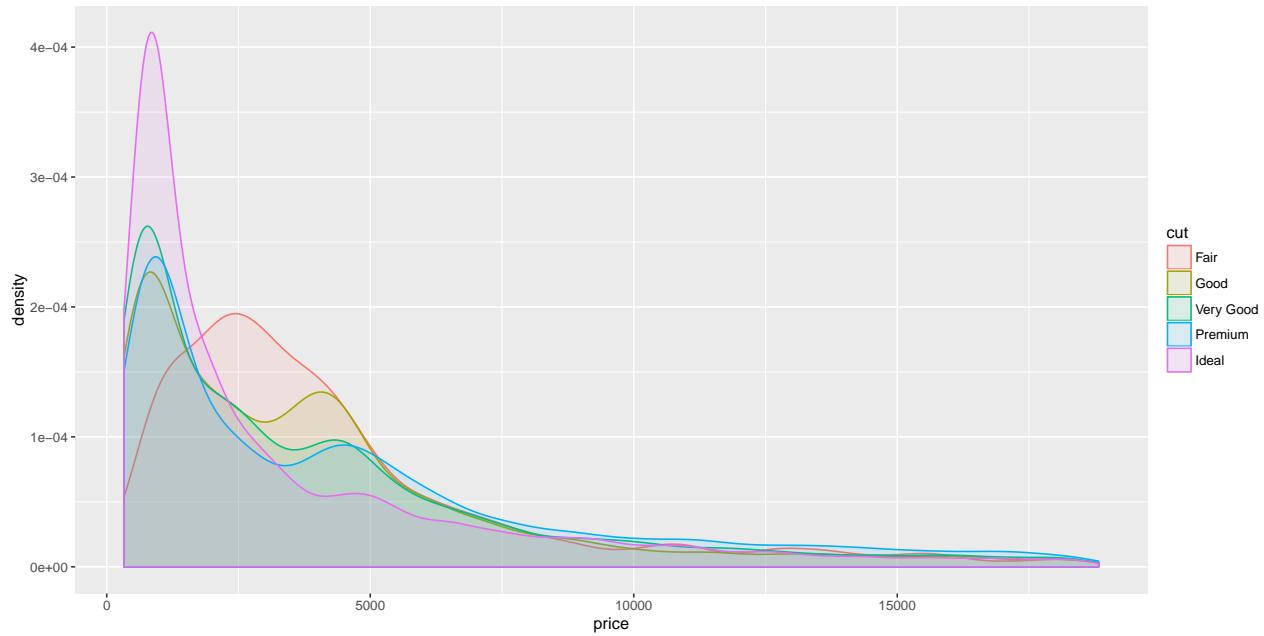
Que encuentra de particular en esta distribución? Que gráfico utilizaría para saber mas acerca de la distribución?

Ahora usemos colores para revisar campos categoricos en funcion de uno numérico

```

grafica <- ggplot(diamonds, aes(price, fill = cut, color = cut))
grafica <- grafica + geom_density(alpha = 0.1)
grafica

```



## Tarea

Haga un diagrama de dispersion en el que describa los valores X,Y y Z del dataset buscando responder: que relacion hay entre el tamaño del diamante y su precio?