**Práctica Diamonds**

*Hacer uso del dataset "diamonds" que contendrá el precio (entre otras variables interesantes) de unos 54.000 diamantes.*

Objetivo: *Realizar distintos tipos de análisis estadístico de sus variables para intentar averiguar algún tipo de comportamiento oculto aparentemente en los datos.*

*Para ello se marcan los siguientes pasos: tipos de variables, medidas de posición central, medidas de dispersión, distribución y relación entre ellas, más análisis de regresión.*

*Los diferentes indicadores presentes en el dataset "diamonds" son los siguientes:*

* *Precio en dólares americanos (price)*
* *Peso del diamante (carat)*
* *Calidad del corte (cut: Fair, Good, Very Good, Premium, Ideal)*
* *Color del diamante (color: desde D el mejor hasta J el peor)*
* *Claridad del diamante (clarity: desde el peor I1, SI2, SI1, VS2, VS1, VVS2, VVS1, hasta el mejor IF)*
* *Longitud en mm (x)*
* *Ancho en mm (y)*
* *Profundidad en mm (z)*
* *Porcentaje total de profundidad (depth)*
* *Anchura de la parte superior de diamante con relación al punto más ancho (table)*

*Responde cada bloque cubriendo al menos lo indicado:*

* *Muestra representativa.*
  + *Selecciona una muestra representativa para "cut"*
* *Análisis de las variables*
  + *Análisis descriptivo de las variables: Tipo de variable, distribución y representación*
  + *Detección de casos atípicos y su tratamiento*
* *Inferencia*
  + *Calcula un intervalo de confianza para la media de "carat" y "depth"*
  + *Formula un test de hipótesis*
* *Relaciones entre las variables*
  + *Muestra las relaciones que existen entre variables (dependencia, anova, correlación)*
* *Análisis de regresión*
  + *Formular un modelo de regresión y analiza los resultados*
  + *Muestra los residuos y analiza los resultados*
  + *Aplica una transformación a la regresión y analiza los resultados*
  + *Interpreta los coeficientes estandarizados de la regresión*

A continuación se desarrollan los bloques indicados en el enunciado de la página anterior. Queda adjunto en fichero en R con las implementaciones llevadas a cabo para el análisis del dataset diamonds.

# Muestra representativa

Para tener una muestra representativa de “cut” hay que extraer una porción aleatoria de los datos que no resulte sesgada. Por ello se ha utilizado la función sample(). El dataframe diamonds contiene 53940 filas. Se ha extraído un 25% de datos de “cut”.

> summary(dt$cut)

Fair Good Very Good Premium Ideal

1610 4906 12082 13791 21551

> cutSample <- summary(sample(dt$cut, nrow(dt)\*0.25, rep = F))

> cutSample

Fair Good Very Good Premium Ideal

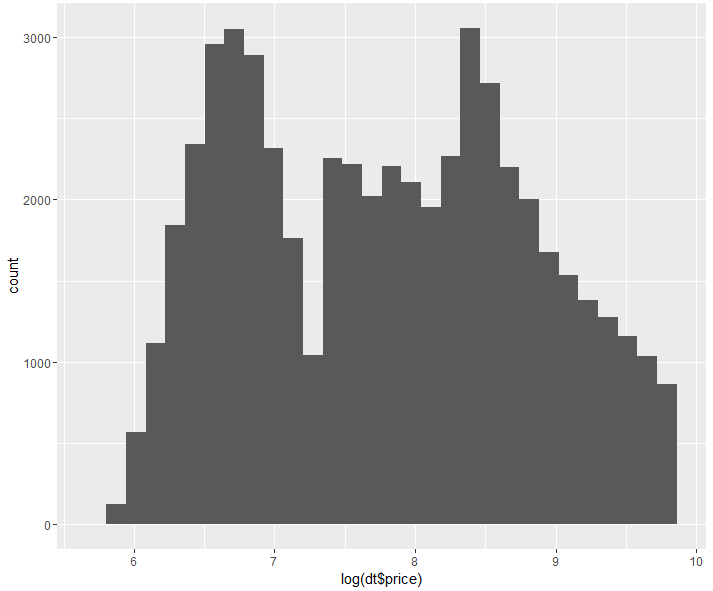
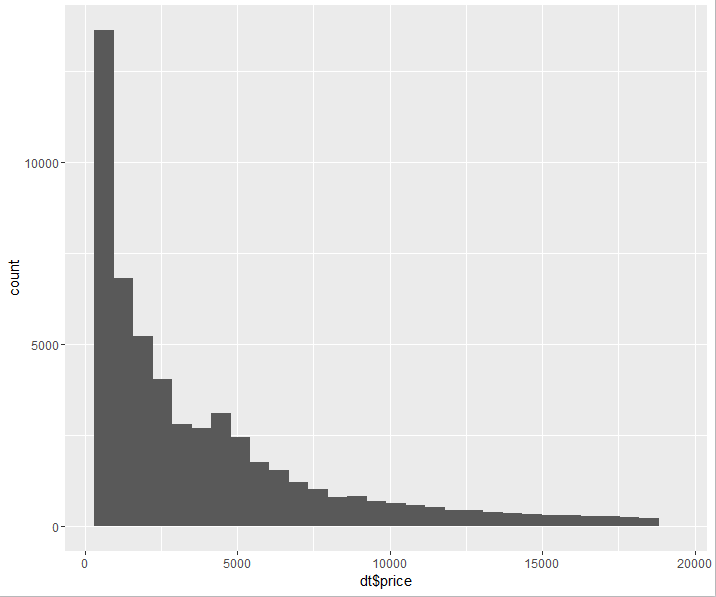
393 1256 3027 3498 5311

Como se puede observar, se respeta la proporción de la muestra (cutSample) con respecto a los datos de “cut”.

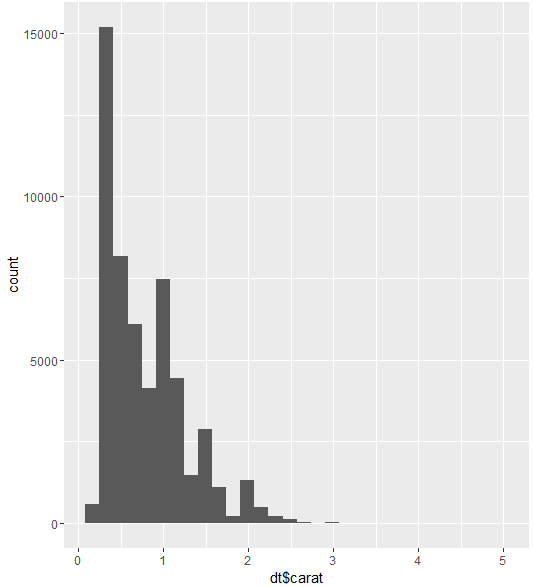
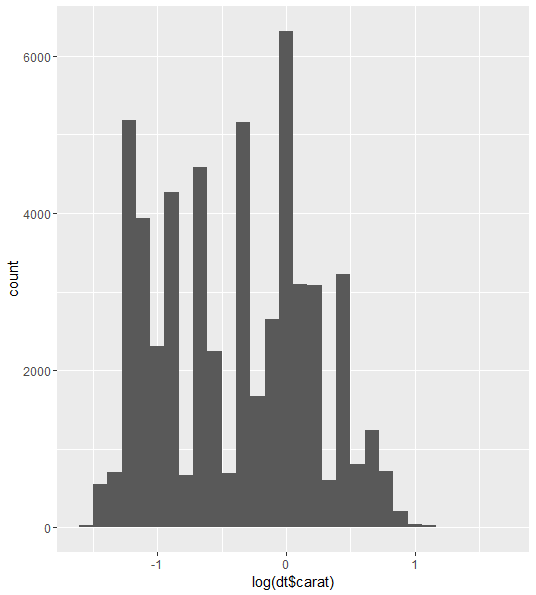
\*\* El parámetro “rep” se deja a False para que no se llegue a dar ninguna repetición en la generación de la muestra \*\*

# Análisis de las variables

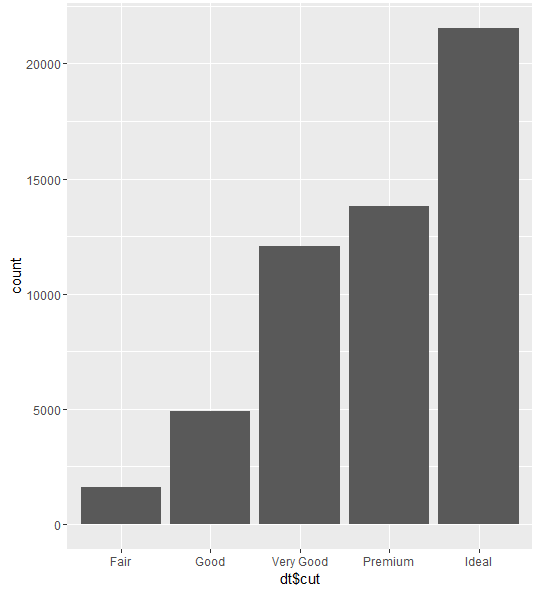
* A continuación se realiza el añálisis de las distribuciones y su representación para cada columna del dataframe “diamonds”.
* Precio en dólares americanos (price): Se trata de una variable cuantitativa continua que con distribución normal. Su representación del recuento del precio de menor a mayor en el eje de abscisas descubre que en condiciones normales va disminuyendo la los precios caros. La misma gráfica logarítmica muestra mejor esa aproximación a la distribución normal.



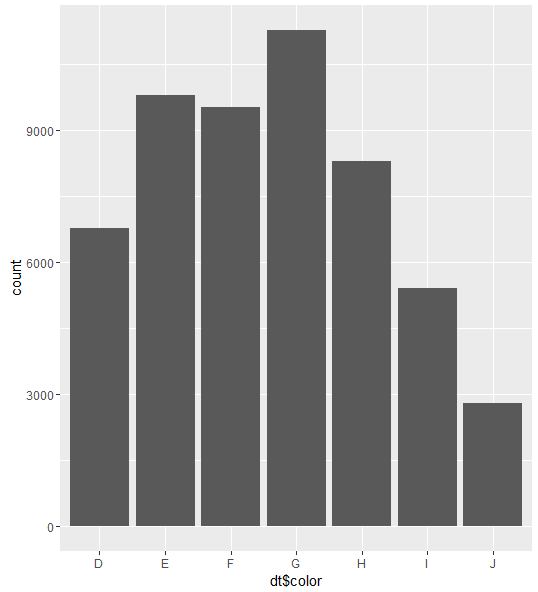
* Peso del diamante (carat): Se trata, al igual que el precio en dólares, de una variable cuantitativa continua que se puede asociar a una distribución normal. Su representación muestra un nivel de variación menos típico que en “price”.

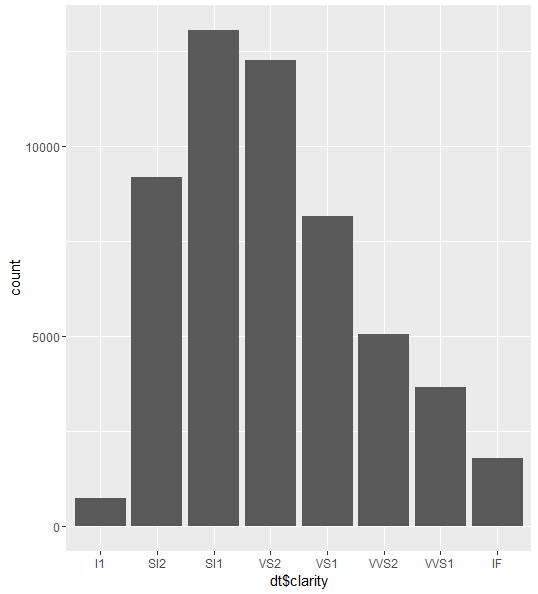
* Calidad del corte (cut: Fair, Good, Very Good, Premium, Ideal): “cut” es un tipo de variable cualitativa (o categórica) nominal. Distribución normal.



* Color del diamante (color: desde D el mejor hasta J el peor). Variable categórica ordinal. Distribución normal.



* Claridad del diamante (clarity: desde el peor I1, SI2, SI1, VS2, VS1, VVS2, VVS1, hasta el mejor IF). Mismas características que el color.



* El resto de variables son medidas / porcentajes: cuantitativas continuas.
  + Longitud en mm (x)
  + Ancho en mm (y)
  + Profundidad en mm (z)
  + Porcentaje total de profundidad (depth)
  + Anchura de la parte superior de diamante con relación al punto más ancho (table)
* Detección de casos atípicos (outliers) y su tratamiento. Se ha implementado una función que extrae los datos atípicos extremos tomando los datos de la columna del dataframe para poder posteriormente eliminarlos si fuese necesario. En el caso siguiente se muestra dicha función y la extracción de ellos a la columna del precio.

# Esta función devuelve los outliers (valores atípicos extremos)

# Si se quisieran los atípicos leves, habría que pasar type # con valor 1.5

atipicos <- function(data, type=3) {

# Q1 es el del 25%

q1 = quantile(data,0.25)

# Q1 es el del 75%

q3 = quantile(data, 0.75)

iqr = q3 - q1

# Identificar outliers

extremoMax = (iqr \* type) + q3

extremoMin = q1 - (iqr \* type)

result <- which(data > extremoMax | data < extremoMin)

}

temp <- atipicos(dt$price)

price <- dt$price

price <- price[-temp]

# Muestra los datos sin outliers

boxplot(price)

\*\* Definimos los valores atípicos (extremos) como < Q1 – 3\* IQR ó valor > Q3 + 3 \* IQR, siendo IQR = Q3 -Q­1 (rango intercuartílico) \*\*

# Inferencia

* Calcula un intervalo de confianza para la media de "carat" y "depth"

Planteemos un intervalo de confianza para carat al 95% tomando una muestra de 36800 diamantes.

> muestra1 <- sample(dt$carat, 36800, rep=F)

> mean(muestra1)

[1] 0.7972478

> sd(muestra1)

[1] 0.4732049

En el caso de esta muestra, el resultado tiene como media  = 0.7972478 y como desviación típica σ = 0.4732049. Ahora hay que buscar el intervalo para p=0.95.

Por tanto, Z será 1.96

> mean(muestra1) - (1.96 \* sd(muestra1) / sqrt(length(muestra1)))

[1] 0.792413

> mean(muestra1) - (1.96 \* sd(muestra1) / sqrt(length(muestra1)))

[1] 0.8020827

|  |
| --- |
| El intervalo de confianza de una muestra representativa de 36800 diamantes para carat con p = 0.95 está en (0.792413, 0.8020827). |

Por otra parte, con depth, se calculará un intervalo de confianza al 90% tomando una muestra de 27000 diamantes.

> muestra2 <- sample(dt$depth, 27000, rep=F)

> mean(muestra2)

[1] 61.73619

> sd(muestra2)

[1] 1.416325

En el caso de esta muestra, el resultado tiene como media  = 61.73619 y como desviación típica σ= 1.416325. Ahora hay que buscar el intervalo para p=0.90.

Por tanto, Z será 1.645

> mean(muestra2) - (1.645 \* sd(muestra2) / sqrt(length(muestra2)))

[1] 61.72201

> mean(muestra2) + (1.645 \* sd(muestra2) / sqrt(length(muestra2)))

[1] 61.75036

|  |
| --- |
| Por tanto, el intervalo de confianza de una muestra representativa de 36800 diamantes para carat con p = 0.95 está en (61.72201, 61.75036). |

* Formula un test de hipótesis

Determinar si se puede aceptar que los diamantes de la muestra tienen una media de 0.8 quilates (carat).

> t.test(muestra1,mu = 0.8,conf.level = 0.95)

One Sample t-test

data: muestra1

t = -1.1157, df = 36799, p-value =

0.2646

alternative hypothesis: true mean is not equal to 0.8

95 percent confidence interval:

0.7924129 0.8020827

sample estimates:

mean of x

0.7972478

Determinar si los diamantes tienen una media del 62% de profundidad (depth).

> t.test(muestra2,mu = 62,conf.level = 0.9)

One Sample t-test

data: muestra2

t = -30.607, df = 26999, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 62

90 percent confidence interval:

61.72201 61.75036

sample estimates:

mean of x

61.73619

\*\* t.test es una función que… \*\*

# Relaciones entre las variables

* + *Muestra las relaciones que existen entre variables (dependencia, anova, correlación)*
* *Dependencia: demostrarla con diagramas de dispersión la correlación*

*http://www.ditutor.com/estadistica\_2/correlacion\_estadistica.html*

ANOVA: ¿¿

<https://www.youtube.com/watch?v=-yQb_ZJnFXw>

# Análisis de regression

* + *Formular un modelo de regresión y analiza los resultados*
  + *Muestra los residuos y analiza los resultados*
  + *Aplica una transformación a la regresión y analiza los resultados*
  + *Interpreta los coeficientes estandarizados de la regresión*

<http://www.listendata.com/2015/09/linear-regression-with-r.html?m=1>