Tarea12

Yimmy Eman

2022-07-10

```
library(tidyverse)
```

```
head(diamonds)
```

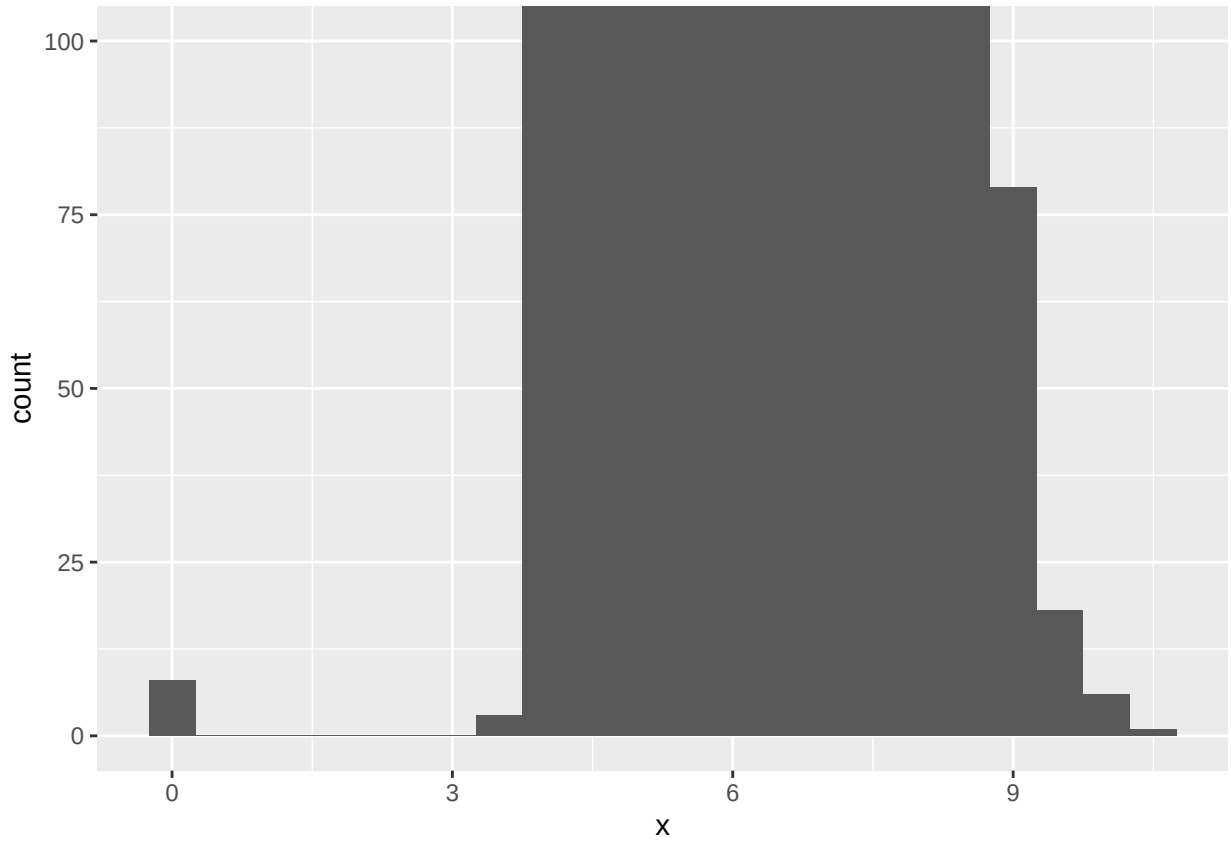
```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

Pregunta 1

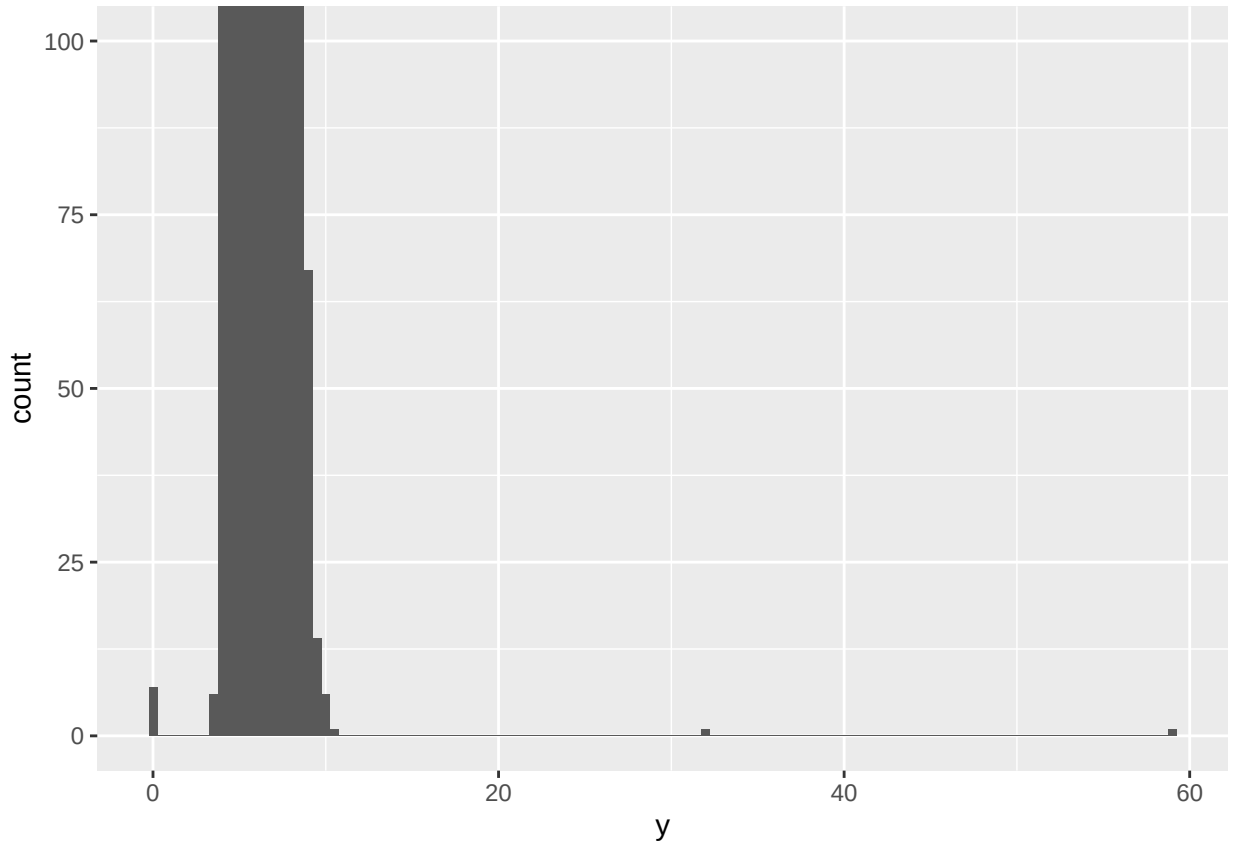
Explora la distribución de las variables x, y, z del dataset de diamonds. ¿Qué podemos inferir? Busca un diamante (por internet por ejemplo) y decide qué dimensiones pueden ser aceptables para las medidas de longitud, altura y anchura de un diamante.

Sol: Los diamantes con x=0 , y<2 y y>30, z= 0 y z>30 son los datos atípicos que deben tenerse en cuenta.

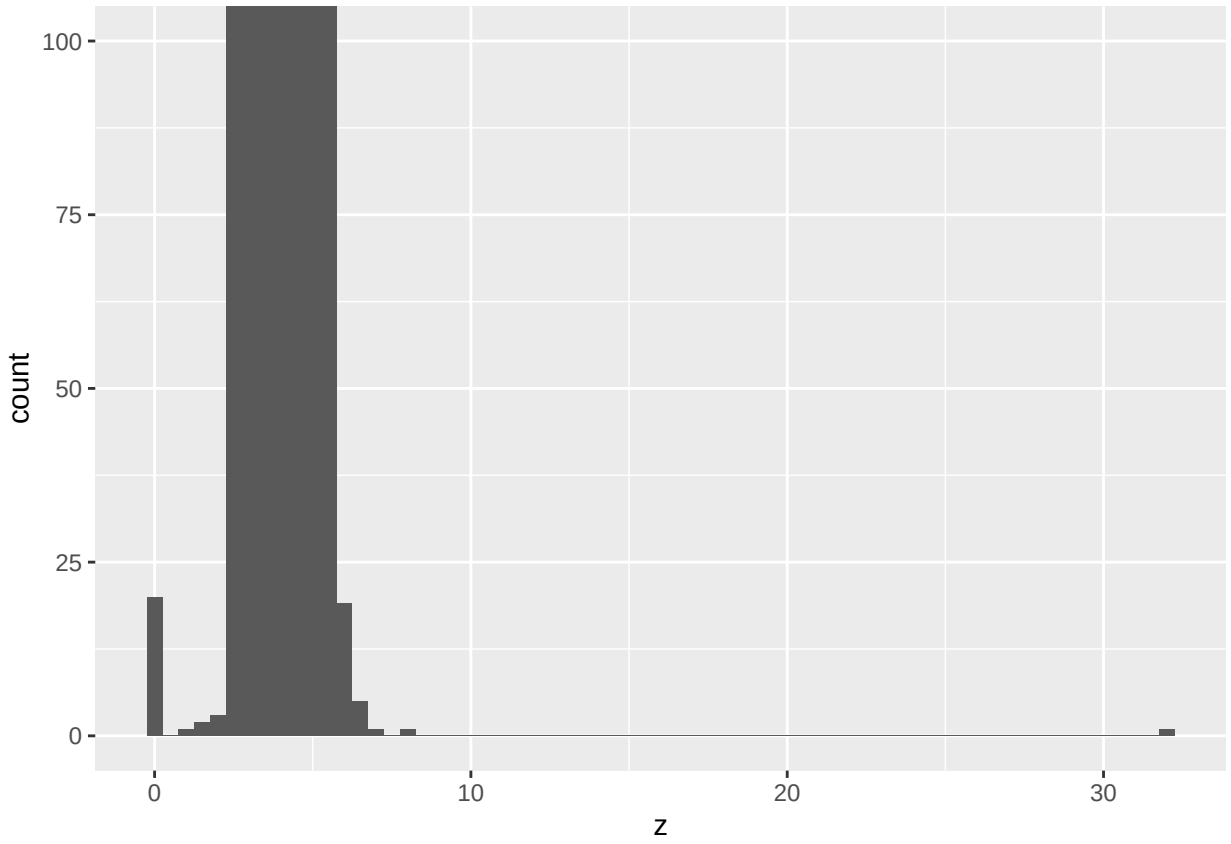
```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = x), binwidth = 0.5) +
  coord_cartesian(ylim = c(0,100))
```




```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = y), binwidth = 0.5) +
  coord_cartesian(ylim = c(0,100))
```



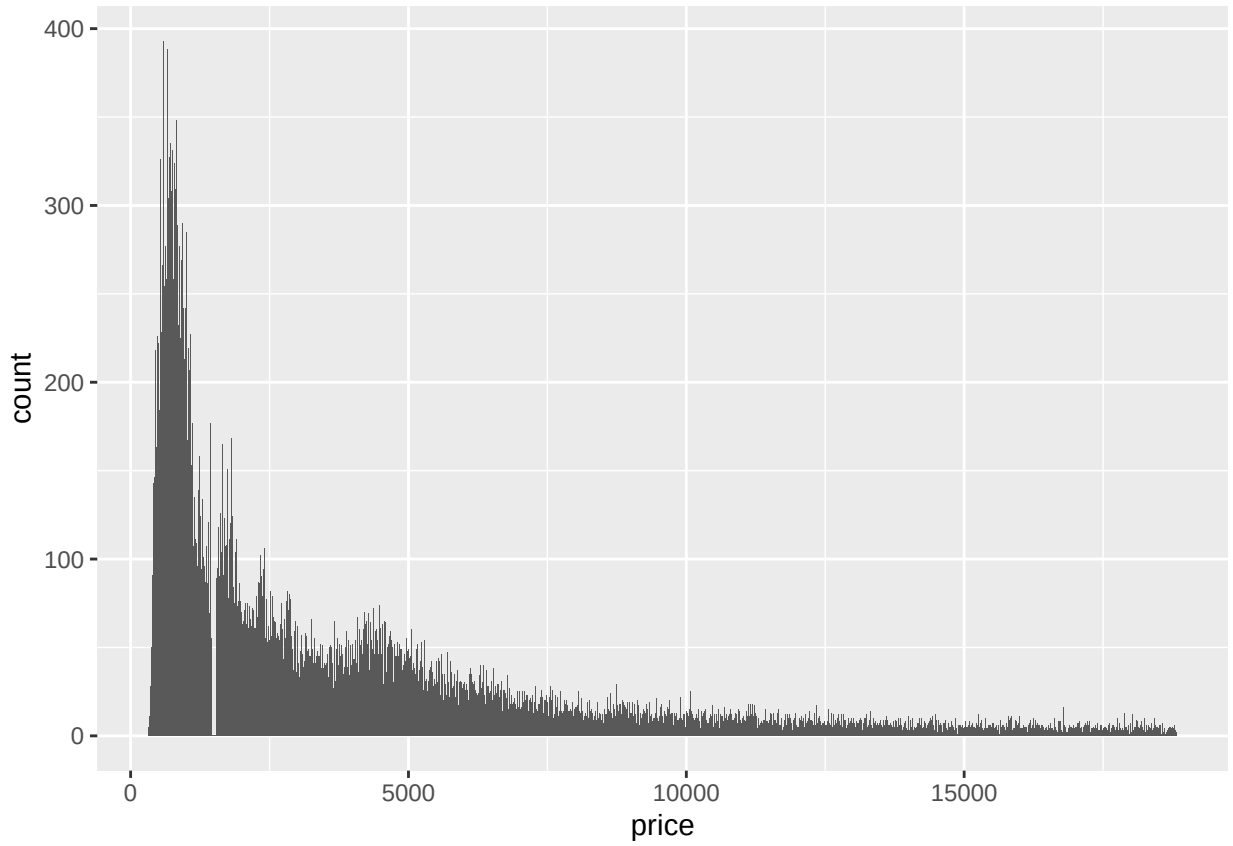

```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = z), binwidth = 0.5) +
  coord_cartesian(ylim = c(0,100))
```



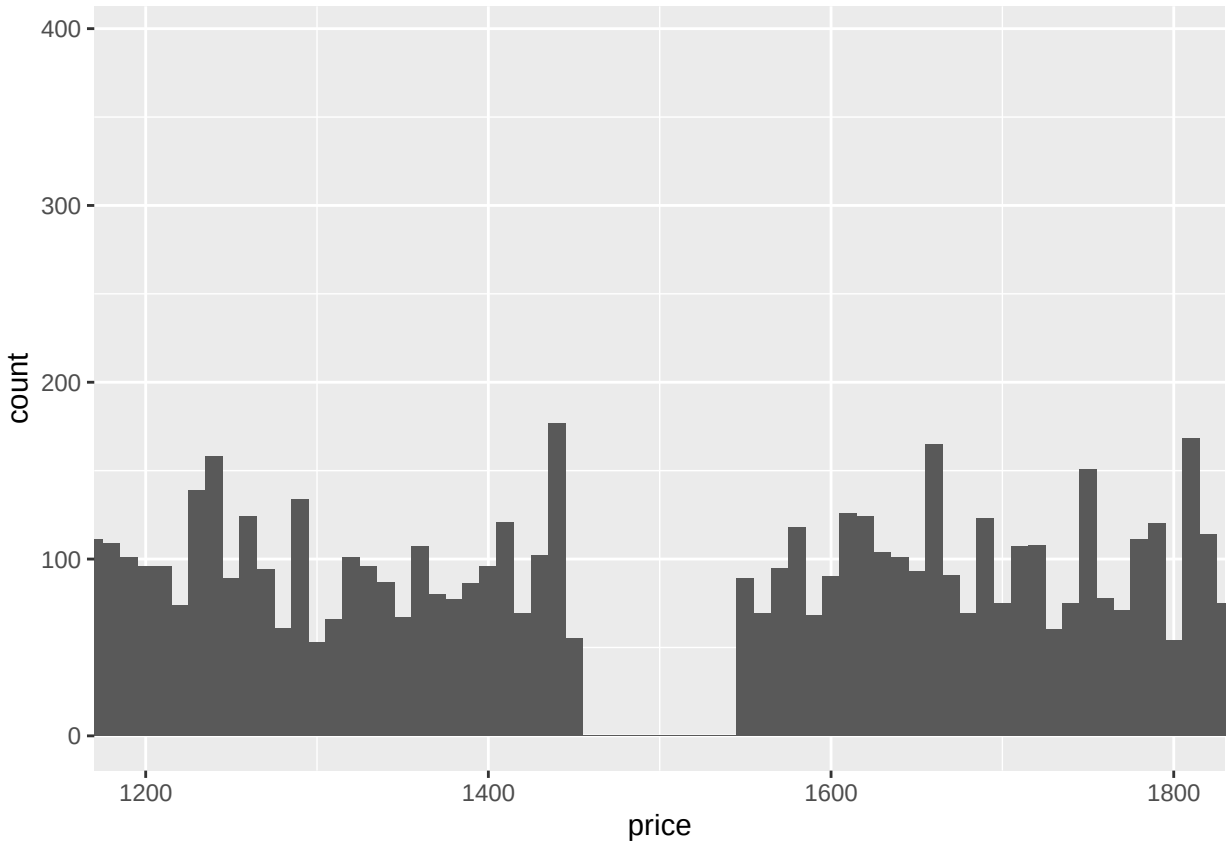
Pregunta 2

Explora la distribución del precio (price) del dataset de diamonds. ¿Hay algo que te llame la atención o resulte un poco extraño? Recuerda hacer uso del parámetro binwidth para probar un rango dispar de valores hasta ver algo que te llame la atención.

```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = price), binwidth = 10)
```



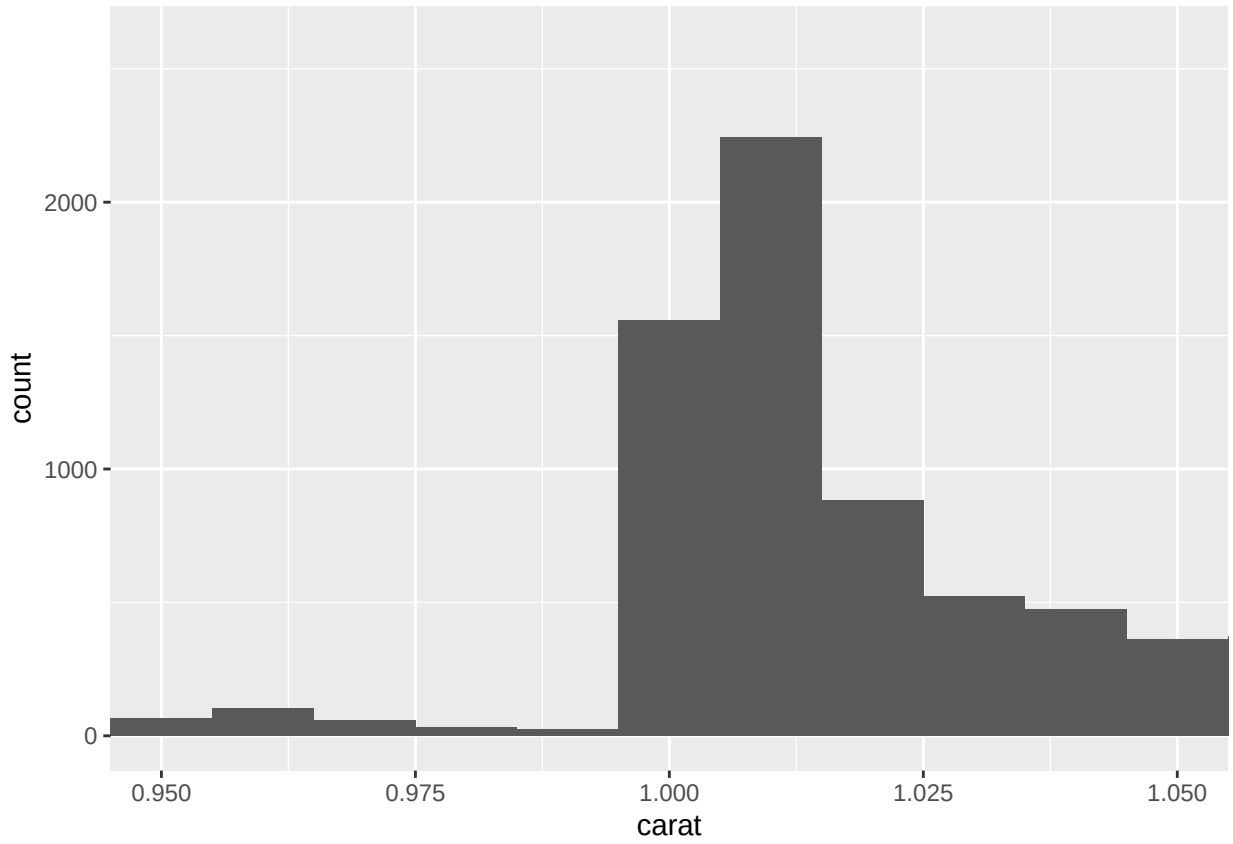

```
#Haciendo Zoom
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = price), binwidth = 10) +
  coord_cartesian(xlim = c(1200,1800))
```



Pregunta 3

¿Cuántos diamantes hay de 0.99 kilates? ¿Y de exactamente 1 kilate? ¿A qué puede ser debida esta diferencia?

```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.01)+
  coord_cartesian(xlim = c(0.95,1.05))
```



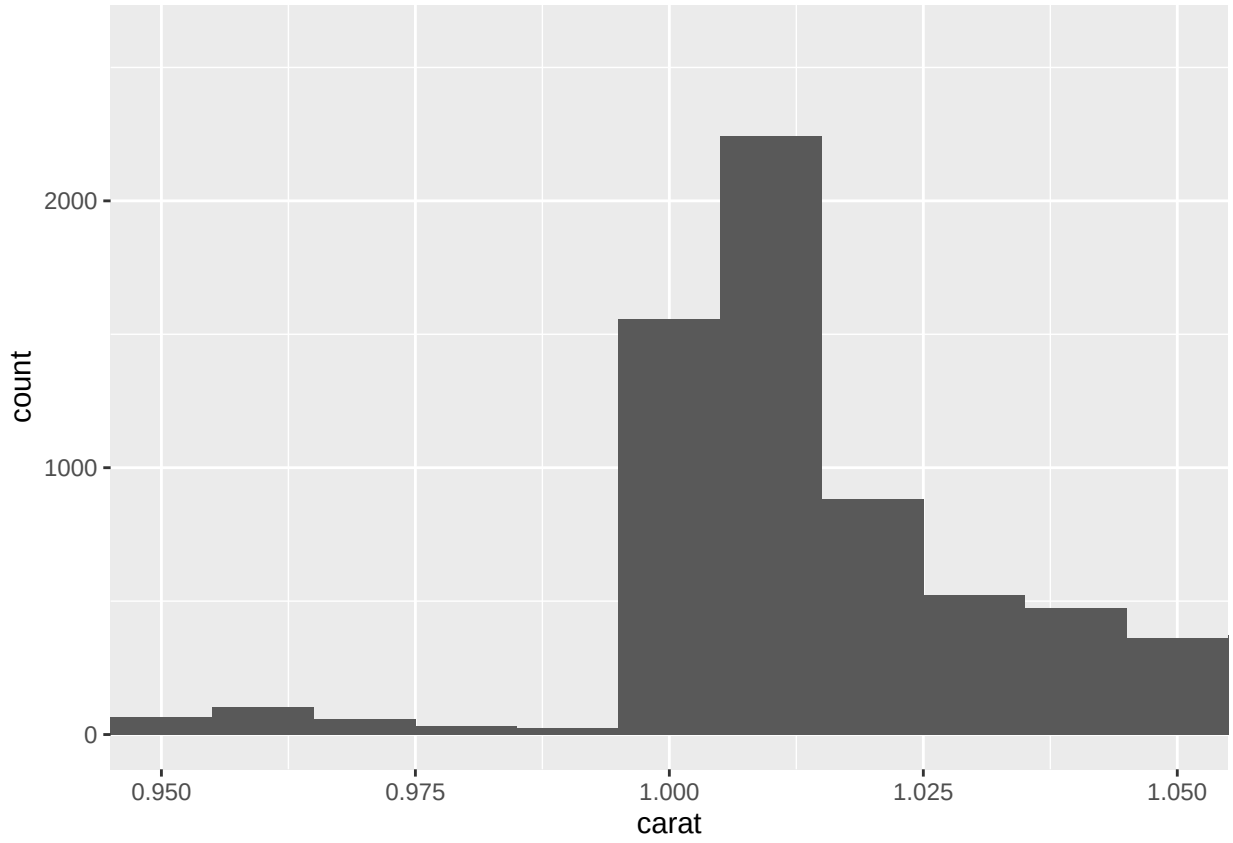

```
diamonds %>%
  filter(between(carat, 0.95, 1.05)) %>%
  count(cut_width(carat, 0.01))
```

```
## # A tibble: 11 x 2
##    `cut_width(carat, 0.01)`     n
##    <fct>                    <int>
##  1 [0.945,0.955]               65
##  2 (0.955,0.965]              103
##  3 (0.965,0.975]               59
##  4 (0.975,0.985]               31
##  5 (0.985,0.995]               23
##  6 (0.995,1.01]              1558
##  7 (1.01,1.02]               2242
##  8 (1.02,1.03]                883
##  9 (1.03,1.04]                523
## 10 (1.04,1.05]                475
## 11 (1.05,1.06]                361
```

Pregunta 4

Compara y contrasta el uso de las funciones coord_cartesian() frente xlim() y ylim() para hacer zoom en un histograma. ¿Qué ocurre si dejamos el parámetro binwidth sin configurar? ¿Qué ocurre si hacemos zoom y solamente se ve media barra?

```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.01)+
  coord_cartesian(xlim = c(0.95,1.05))
```



Pregunta 5

- ¿Qué ocurre cuando hay NAs en un histograma?
- ¿Qué ocurre cuando hay NAs en un diagrama de barras?
- ¿Qué diferencias observas?

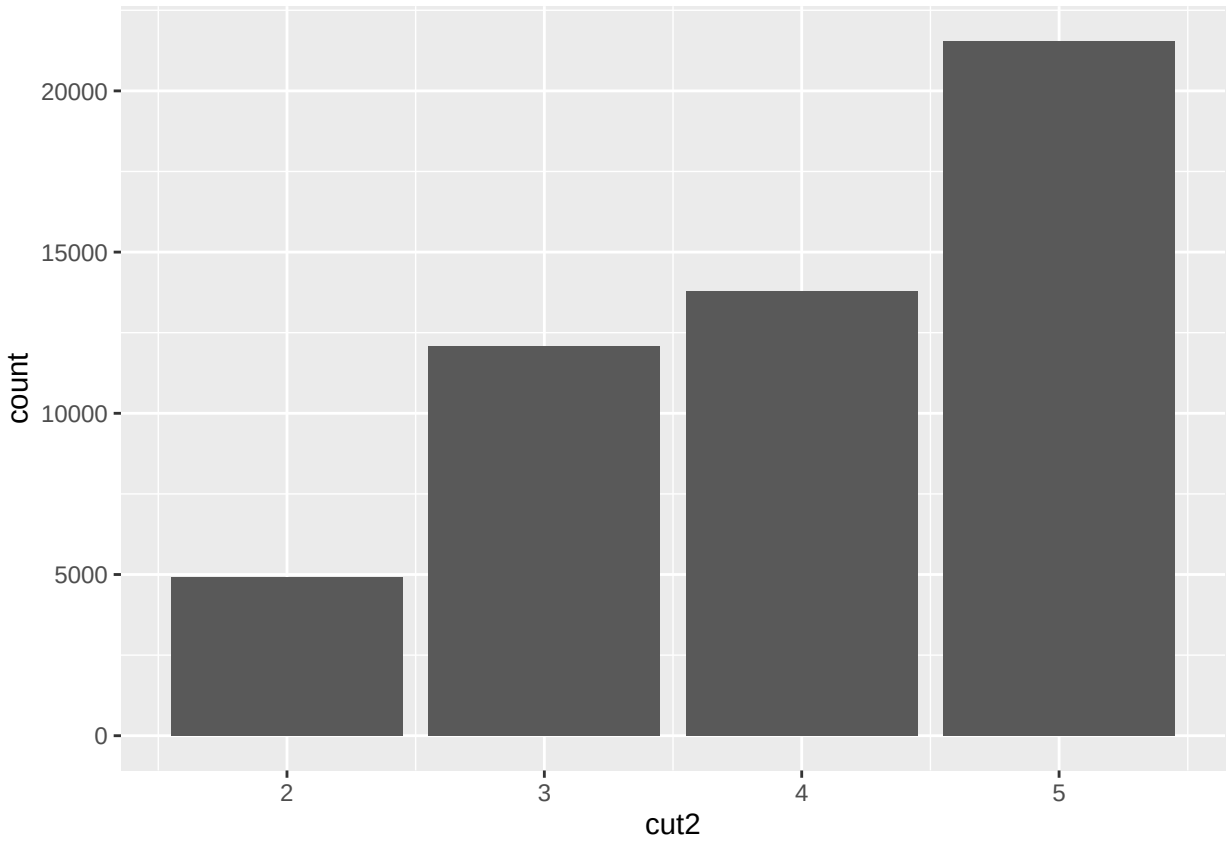
```
good_diamonds <- diamonds %>%
  mutate(y = ifelse(y<2 | y>30, NA, y))

na_diamonds <- good_diamonds %>%
  mutate(cut2 = ifelse(cut == "Fair", NA, cut))

na_diamonds
```

```
## # A tibble: 53,940 x 11
##    carat cut       color clarity depth table price     x     y     z  cut2
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl> <int>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43     5
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31     4
##  3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31     2
```

```
##  4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63     4
##  5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75     2
##  6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48     3
##  7  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47     3
##  8  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53     3
##  9  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49    NA
## 10  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39     3
## # ... with 53,930 more rows
```

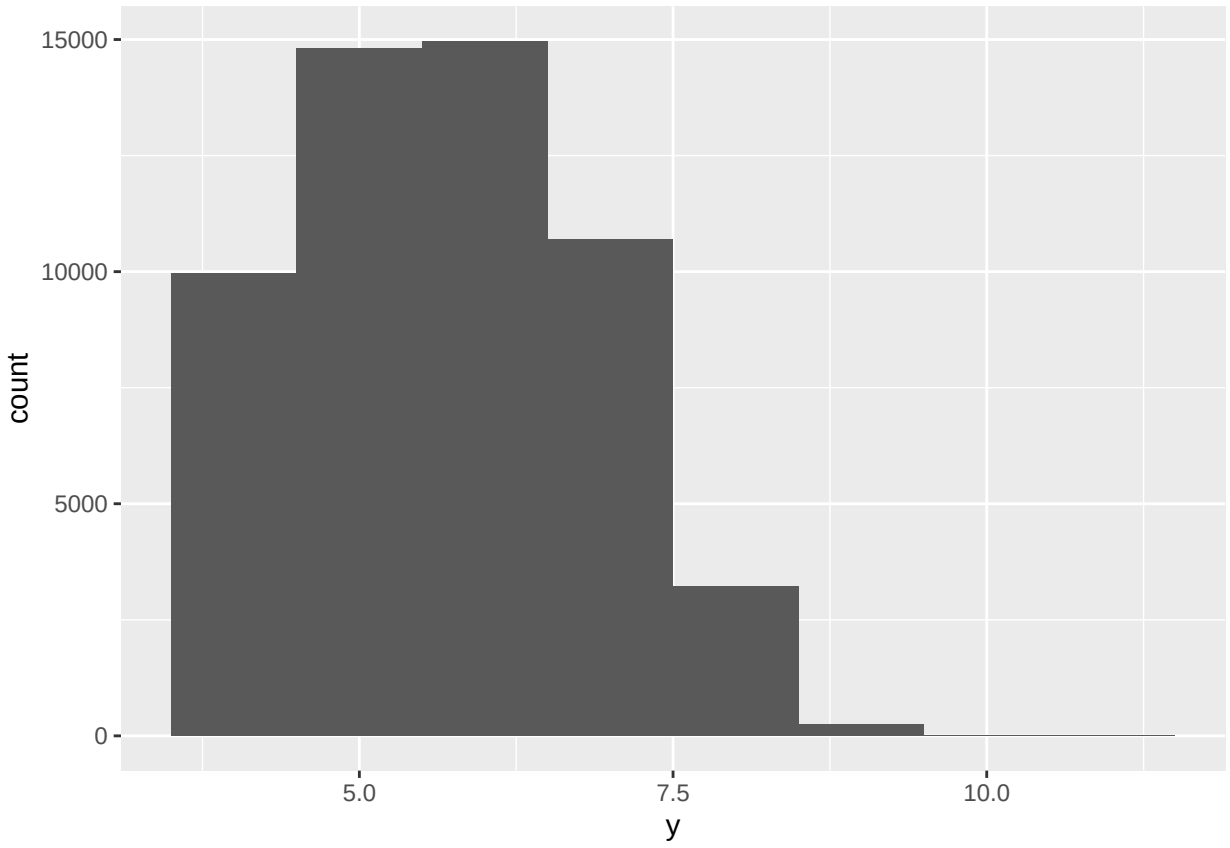
```
ggplot(na_diamonds) +
  geom_bar(mapping = aes(x = cut2))
```

```
## Warning: Removed 1610 rows containing non-finite values (stat_count).
```




```
ggplot(good_diamonds) +
  geom_histogram(mapping = aes(x = y), binwidth = 1)
```

```
## Warning: Removed 9 rows containing non-finite values (stat_bin).
```



Pregunta 6

¿Qué hace la opciónna.rm = TRUEen las funcionesmean()ysum()?

Solución:

Elimina losNAde la suma y del promedio, pero no del dataset sobre el cual se opera.