Tarea: Introducción a la Exploración de los Datos

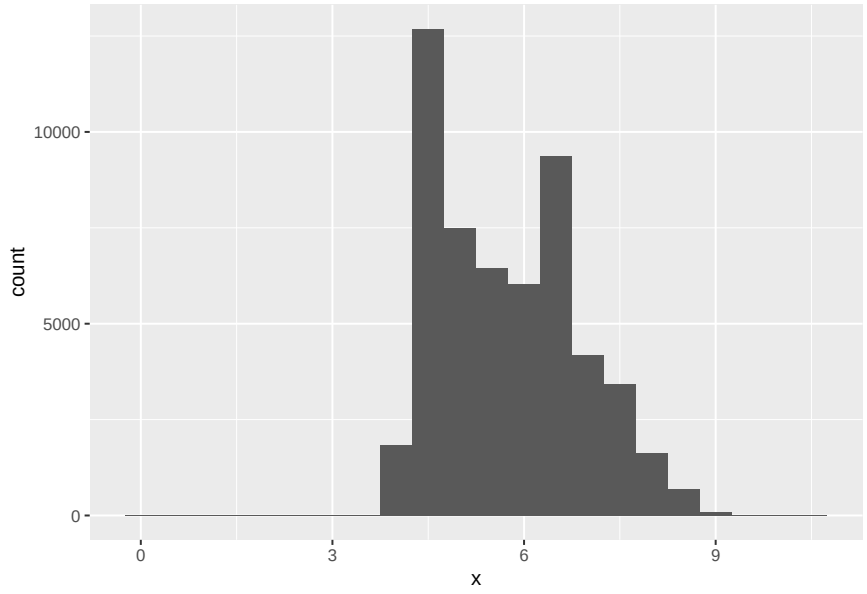
Jesus Mudarra Luján

2023-09-24

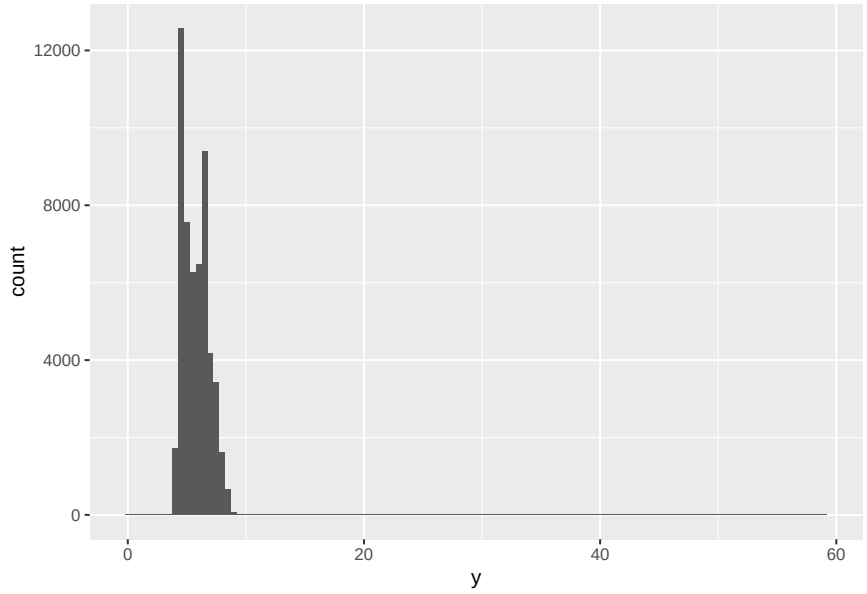
Pregunta 1

Explora la distribución de las variables `x`, `y`, `z` del dataset de `diamonds`. ¿Qué podemos inferir?

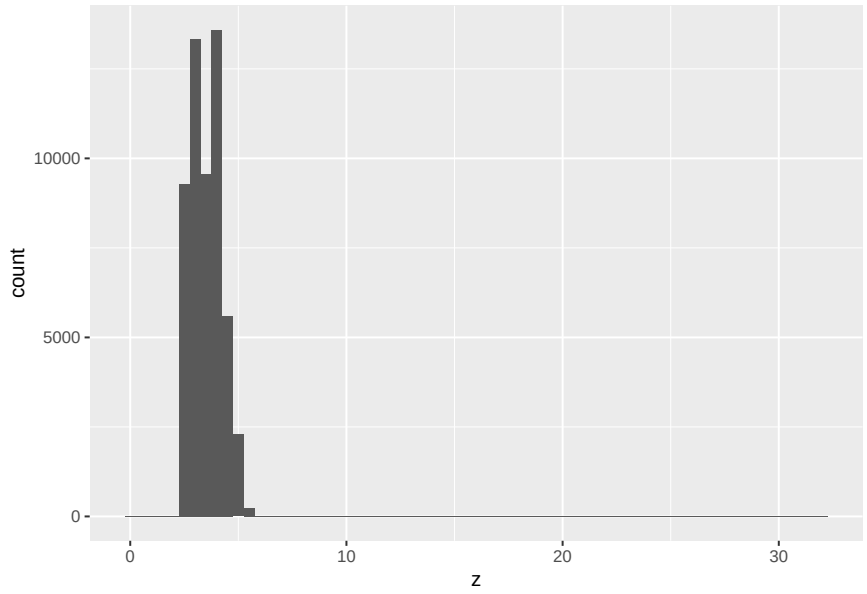
```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = x), binwidth = 0.5)
```




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




```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = z), binwidth = 0.5)
```



En las variables `y` y `z` hay presencia de outliers muy pronunciados. También, se observa que la mayoría de valores en la coordenada `x` se acumulan en 6mm, en la coordenada `y` en 5mm, y en la coordenada `z` en 3mm.

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.

Ejemplo con el zafiro ovalado:

- Longitud (x): Un zafiro ovalado típico podría tener una longitud de alrededor de 6-7 mm.
- Anchura (y): La anchura de zafiro ovalado típico suele ser de aproximadamente 4-5 mm.

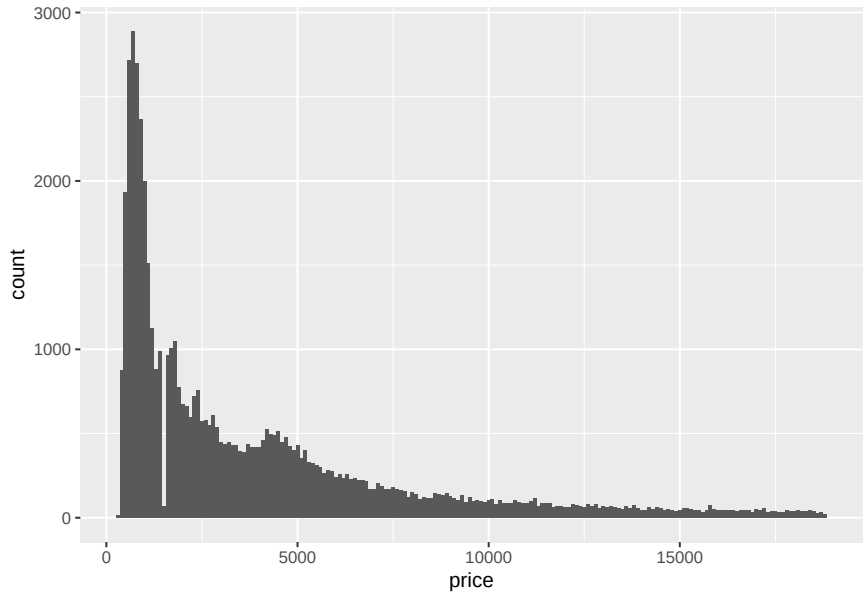
- Profundo (z): En general, los zafiros suelen tener una profundidad que varía entre el 40% y el 80% de su longitud, dependiendo de su forma y estilo de corte, aproximadamente 2.5-5 mm.

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(data = diamonds) +
  geom_histogram(mapping = aes(x = price), binwidth = 100)
```



Se observa una caída exponencial en el precio del diamante. Se concluye que los diamantes cuanto más caros menos unidades hay. Llama la atención la ausencia de valores en aproximadamente 1500$.

Pregunta 3

¿Cuántos diamantes hay de 0.99 kilates? ¿Y de exactamente 1 kilate?

```
diamonds %>%
  filter(carat==0.99) %>%
  summarise(count = n())
```

```
# A tibble: 1 x 1
  count
  <int>
1    23
```

```r
diamonds %>%
  filter(carat==1) %>%
  summarise(count = n())
```

```
# A tibble: 1 x 1
  count
  <int>
1  1558
```

hay 23 diamantes de 0.99 kilates y 1558 diamantes de 1 kilate.

¿A qué puede ser debida esta diferencia?

Por la diferencia que existe se entiende que los diamantes cercanos a un kilate se redondea a la unidad. En este caso, hay 23 diamantes que no han sido redondeados.

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?

Dejando el `binwidth` se intenta calcular automáticamente un valor apropiado para los bins en función de los datos proporcionados.

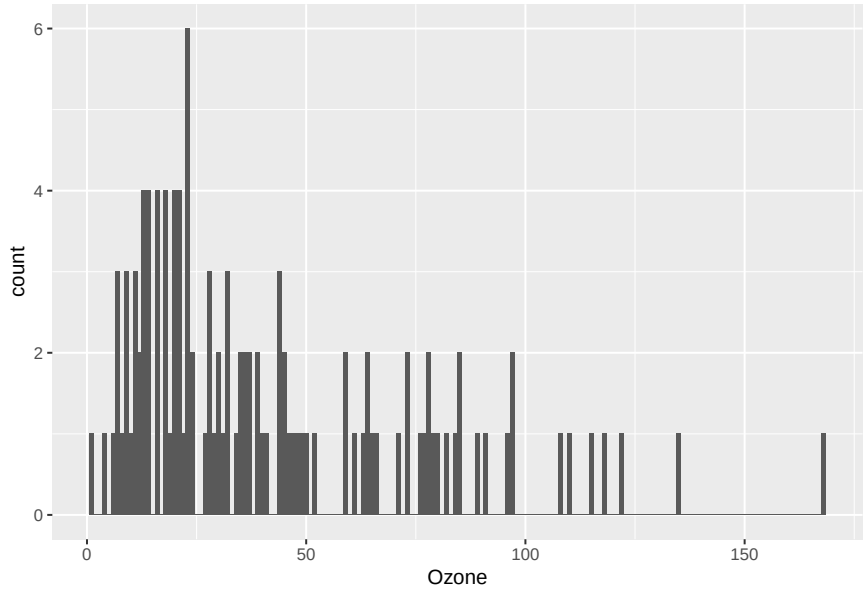
¿Qué ocurre si hacemos zoom y solamente se ve media barra?

Ver solamente media barra indica que los datos dentro de ese rango están muy concentrados o que el ancho utilizado es demasiado grande para mostrar los detalles.

Pregunta 5

- ¿Qué ocurre cuando hay `NAs` en un histograma?

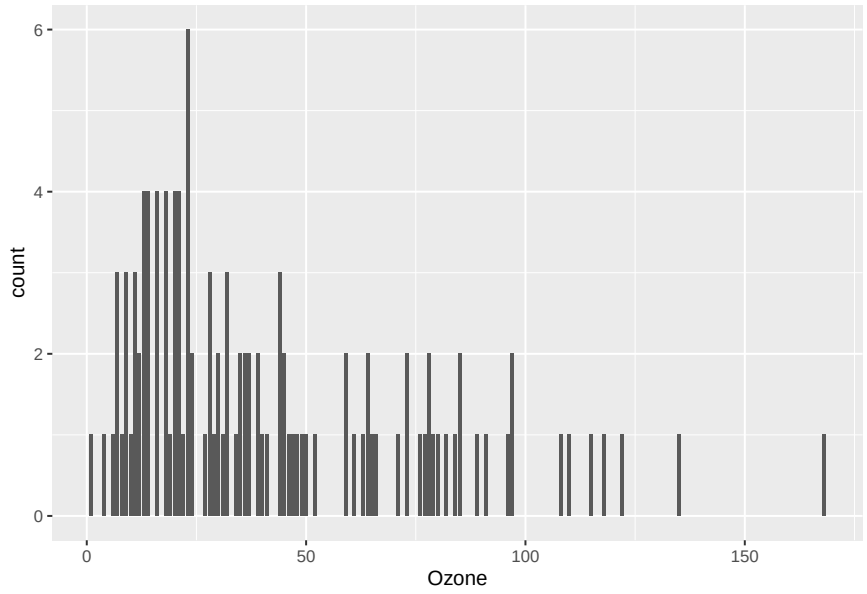
```r
ggplot(data = airquality) +
  geom_histogram(mapping = aes(x = Ozone), binwidth = 1)
```



Utilizando la función `geom_histogram()` de `ggplot` se excluye automáticamente los valores `NAs`.

- ¿Qué ocurre cuando hay `NAs` en un diagrama de barras?

```
ggplot(data = airquality) +
  geom_bar(mapping = aes(x = Ozone))
```



Utilizando la función `geom_bar()` de `ggplot` se excluye automáticamente los valores `NAs`.

- ¿Qué diferencias observas?

Visualmente parece que se no se han realizado los mismos intervalos, da la sensación que en el diagrama de barras hay más presencia de valores faltantes.

Pregunta 6

¿Qué hace la opción `na.rm = TRUE` en las funciones `mean()` y `sum()`?

Indica si los valores `NA` deben eliminarse antes de calcular las funciones media y suma.