01_exploratorio

Ita Santiago

15/8/2020

Propinas

Los siguientes datos fueron registrados en un restaurante durante cuatro días consecutivos:

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------
--------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------
--------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork)

## Lee los datos
tips <- read_csv("tips.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   total_bill = col_double(),
##   tip = col_double(),
##   sex = col_character(),
##   smoker = col_character(),
##   day = col_character(),
##   time = col_character(),
##   size = col_double()
## )
```

```
glimpse(tips)
```

```
## Rows: 244
## Columns: 8
```

```
## $ X1         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, 1...
## $ total_bill <dbl> 16.99, 10.34, 21.01, 23.68, 24.59, 25.29, 8.77,
26.88, 1...
## $ tip        <dbl> 1.01, 1.66, 3.50, 3.31, 3.61, 4.71, 2.00, 3.12,
1.96, 3....
## $ sex        <chr> "Female", "Male", "Male", "Male", "Female", "Male",
"Mal...
## $ smoker     <chr> "No", "No", "No", "No", "No", "No", "No", "No",
"No", "N...
## $ day        <chr> "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun",
"Sun", ...
## $ time       <chr> "Dinner", "Dinner", "Dinner", "Dinner", "Dinner",
"Dinne...
## $ size       <dbl> 2, 3, 3, 2, 4, 4, 2, 4, 2, 2, 2, 4, 2, 4, 2, 2, 3,
3, 3,...
```

```
## Recodificar nombres y niveles
propinas <- tips %>%
  rename(cuenta_total = total_bill,
         propina = tip, sexo = sex,
         fumador = smoker,
         dia = day, momento = time,
         num_personas = size) %>%
  mutate(sexo = recode(sexo, Female = "Mujer", Male = "Hombre"),
         fumador = recode(fumador, No = "No", Si = "Si"),
         dia = recode(dia, Sun = "Dom", Sat = "Sab", Thur = "Jue", Fri =
"Vie"),
         momento = recode(momento, Dinner = "Cena", Lunch = "Comida"))
%>%
  select(-sexo) %>%
  mutate(dia  = fct_relevel(dia, c("Jue", "Vie", "Sab", "Dom")))
propinas
```

```
## # A tibble: 244 x 7
##       X1 cuenta_total propina fumador dia   momento num_personas
##    <dbl>        <dbl>   <dbl> <chr>   <fct> <chr>          <dbl>
##  1     1        17.0     1.01 No      Dom   Cena               2
##  2     2        10.3     1.66 No      Dom   Cena               3
##  3     3        21.0     3.5  No      Dom   Cena               3
##  4     4        23.7     3.31 No      Dom   Cena               2
##  5     5        24.6     3.61 No      Dom   Cena               4
##  6     6        25.3     4.71 No      Dom   Cena               4
##  7     7         8.77    2    No      Dom   Cena               2
##  8     8        26.9     3.12 No      Dom   Cena               4
##  9     9        15.0     1.96 No      Dom   Cena               2
## 10    10        14.8     3.23 No      Dom   Cena               2
## # ... with 234 more rows
```

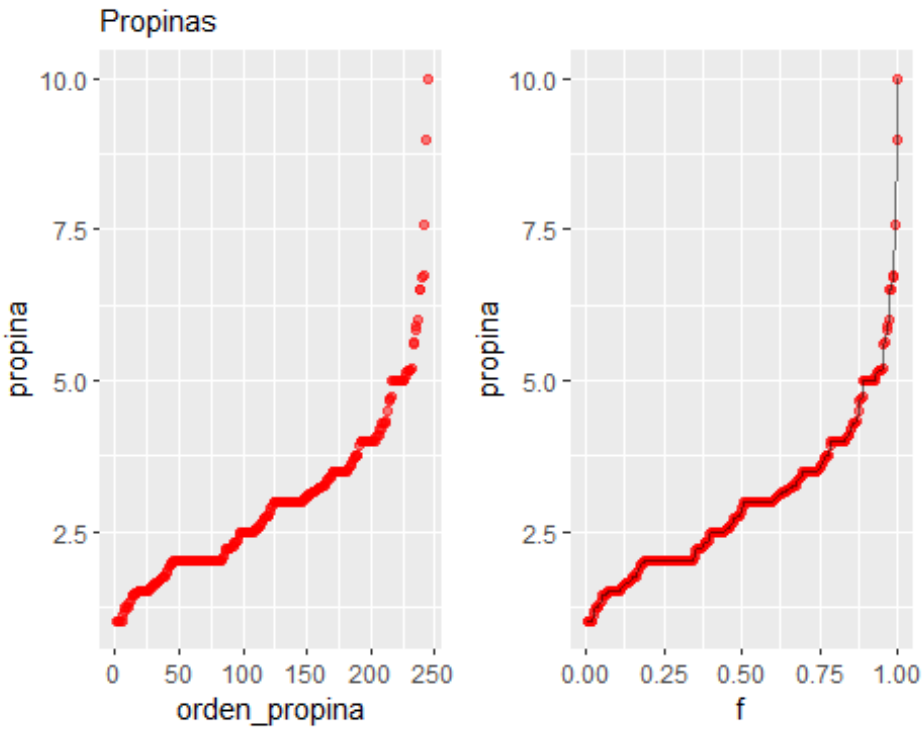
1. Calcula percentiles de la variable propina junto con mínimo y máxio

```
quantile(propinas$propina)
```

```
##      0%     25%     50%     75%    100% 
##  1.0000  2.0000  2.9000  3.5625 10.0000
```

2. Haz una gráfica de cuantiles de la variable propina

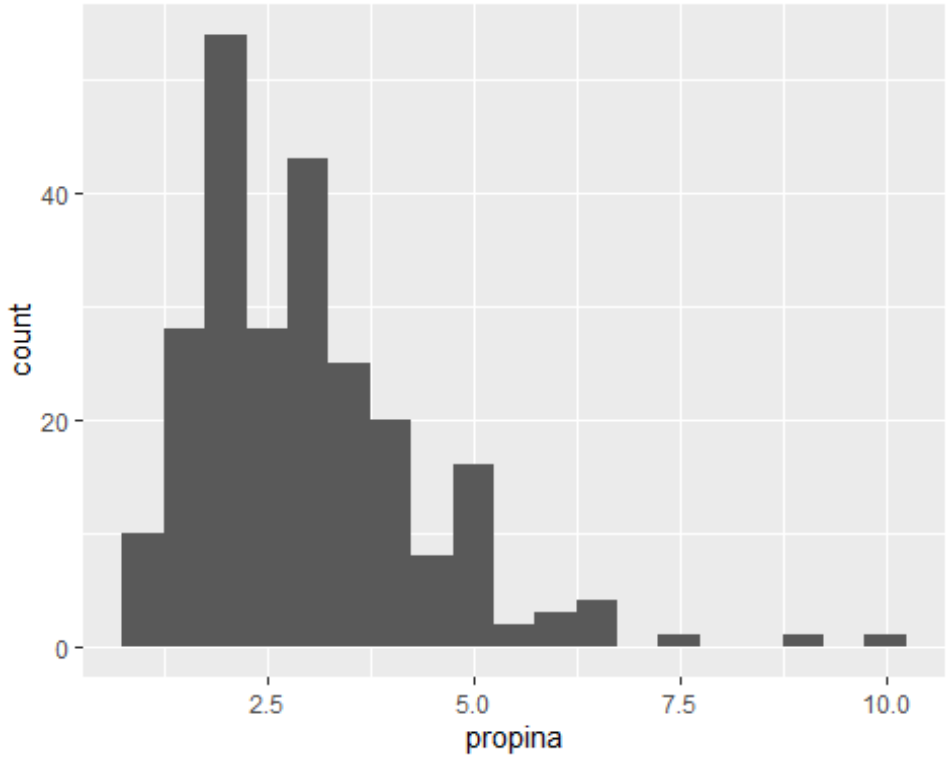
```
propinas <- propinas %>%
  mutate(orden_propina = rank(propina, ties.method = "first"),
         f = orden_propina / n())

g_orden <- ggplot(propinas, aes(y = propina, x = orden_propina)) +
  geom_point(colour = "red", alpha = 0.5) +
  labs(subtitle = "Propinas")
g_cuantiles <- ggplot(propinas, aes(y = propina, x = f)) +
  geom_point(colour = "red", alpha = 0.5) +
  geom_line(alpha = 0.5) +
  labs(subtitle = "")
g_orden + g_cuantiles
```



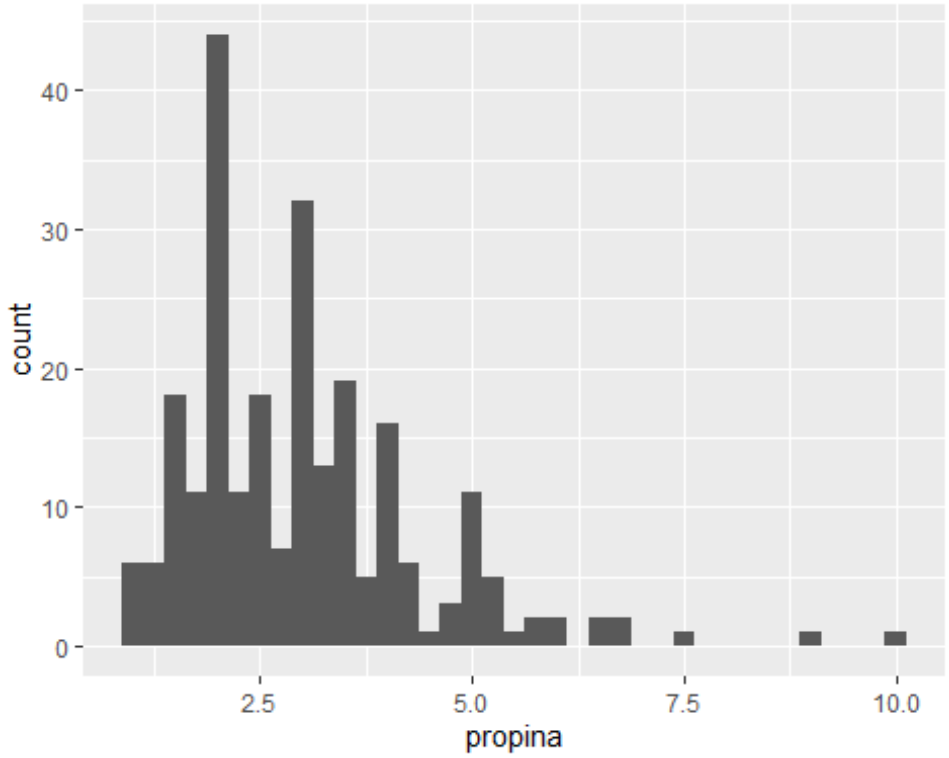
3. Haz un histograma de la variable propinas

Ajusta distintos anchos de banda

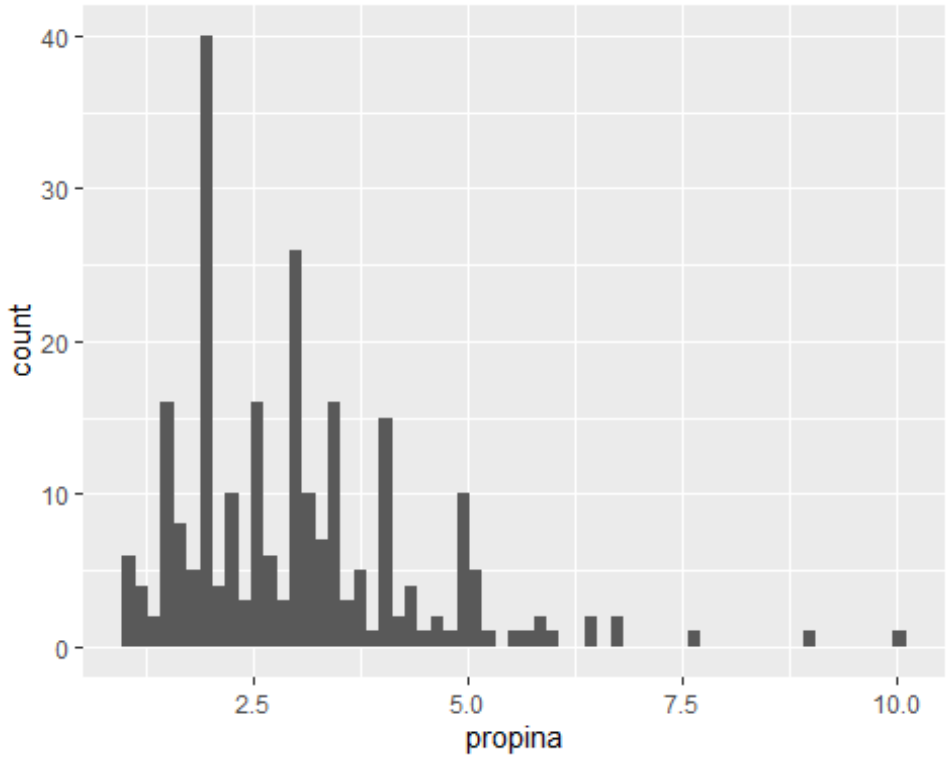
```
ggplot(propinas, aes(x = propina)) +
  geom_histogram(binwidth = 0.5)
```




```r
ggplot(propinas, aes(x = propina)) +
  geom_histogram(binwidth = 0.25)
```

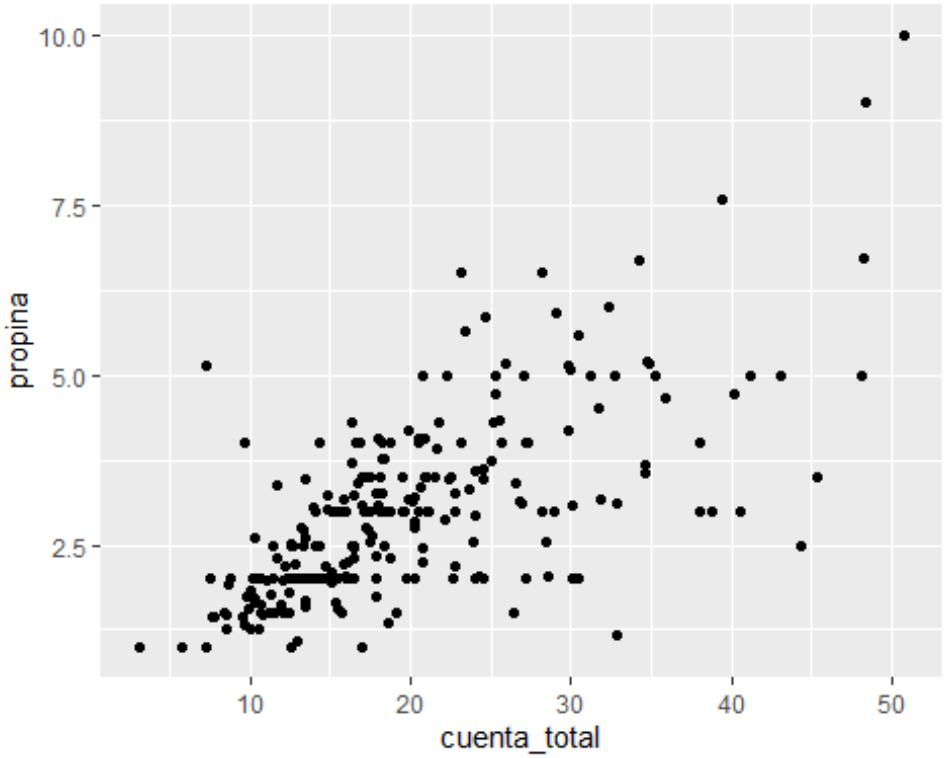


```
ggplot(propinas, aes(x = propina)) +
  geom_histogram(binwidth = 0.15)
```



4. Haz una gráfica de cuenta total contra propina

```
ggplot(propinas, aes(x= cuenta_total, y = propina)) +
  geom_point()
```



5. Calcula propina en porcentaje de la cuenta total

calcula algunos cuantiles de propina en porcentaje

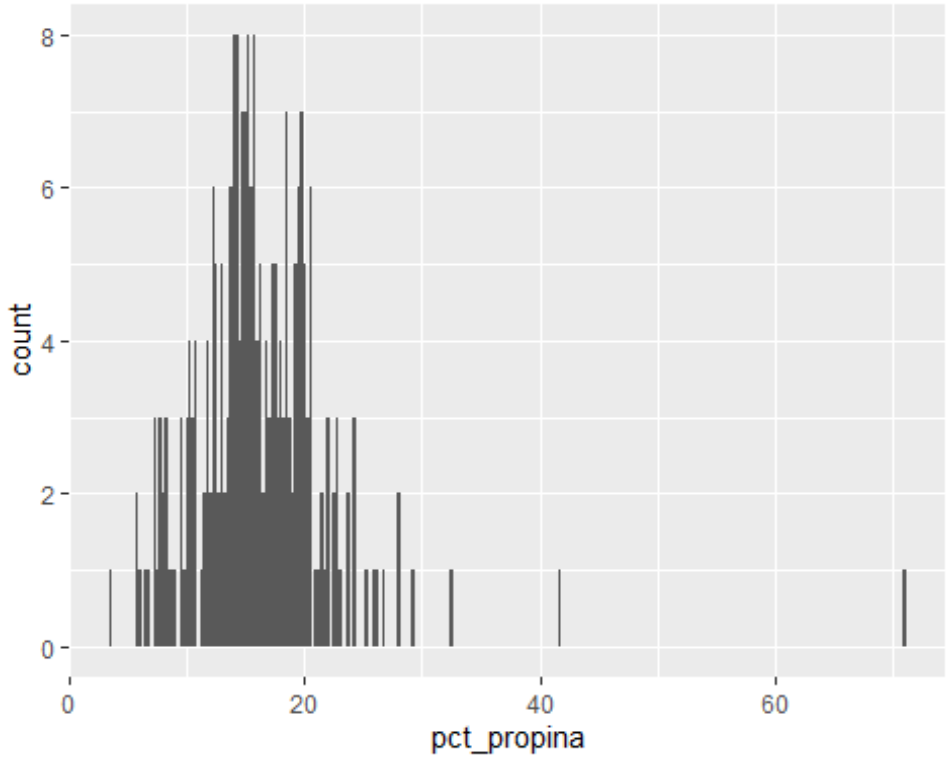
```
propinas <- propinas %>%
  mutate(pct_propina = 100 * propina / cuenta_total)

quantile(propinas$pct_propina)
```

```
##        0%       25%       50%       75%      100%
##  3.563814 12.912736 15.476977 19.147549 71.034483
```

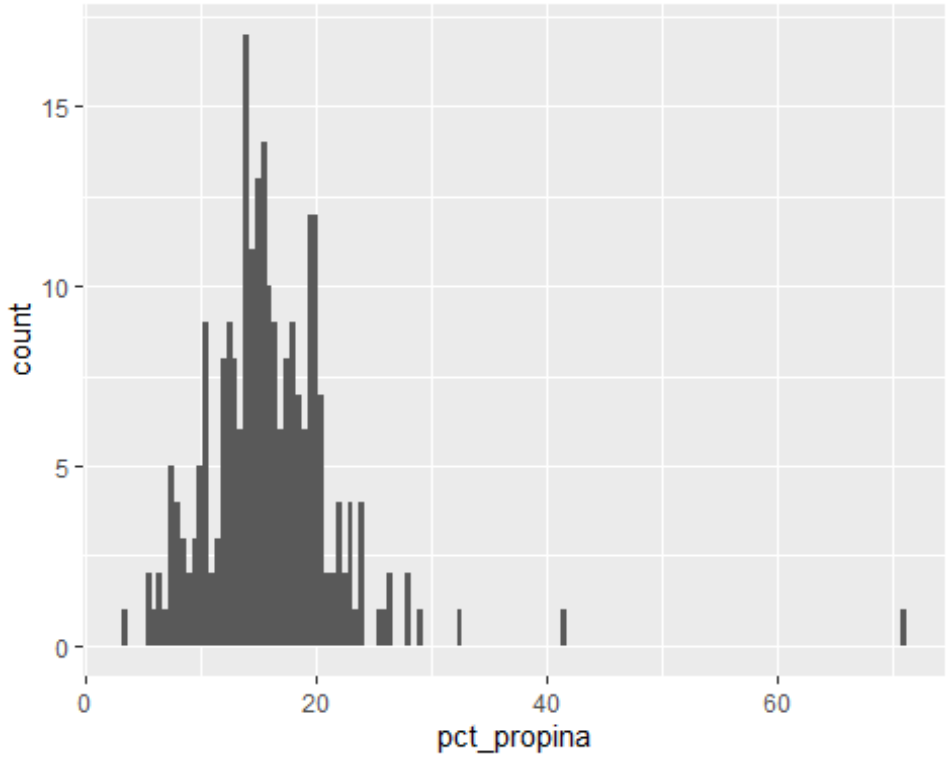
6. Haz un histograma de la propina en porcentaje.

Prueba con distintos anchos de banda.

```
ggplot(propinas, aes(x = pct_propina)) +
  geom_histogram(binwidth = 0.25)
```



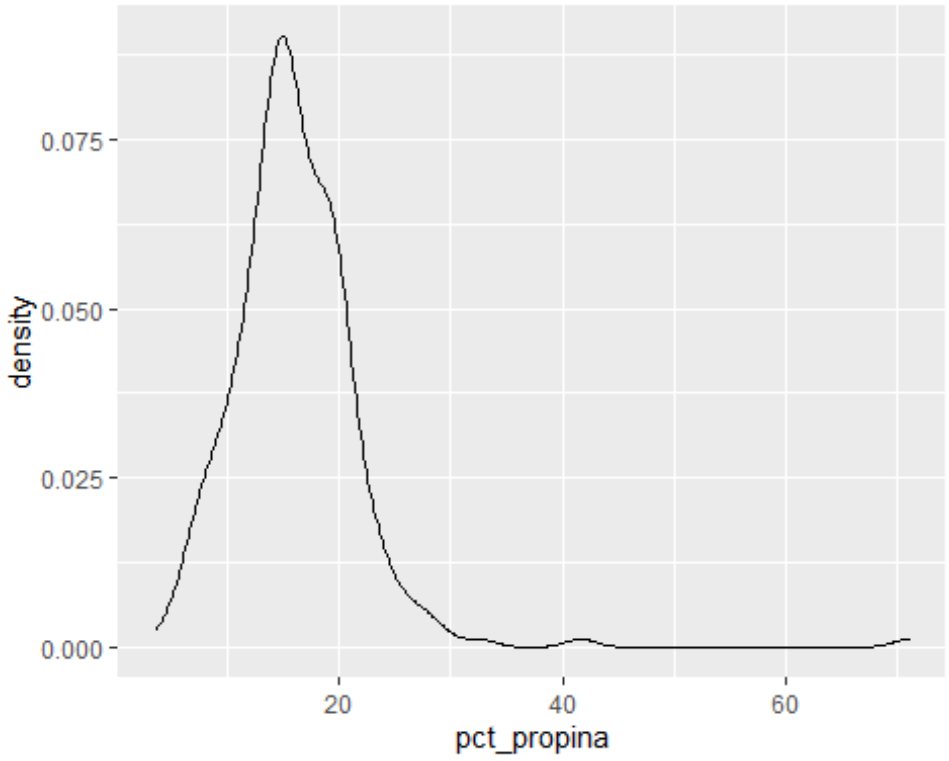

```
ggplot(propinas, aes(x = pct_propina)) +
  geom_histogram(binwidth = 0.5)
```



7. Describe la distribución de propina en pct.

¿Hay datos atípicos?

```
ggplot(propinas, aes(x = pct_propina)) +
  geom_density()
```



Podemos observar que la curva está sesgada hacia la izquierda, después del 25%, podemos inferir que estos serían los porcentajes de propina altos.

8. Filtra los casos con porcentaje de propina muy altos.

¿Qué tipos de cuentas son? ¿Son cuentas grandes o chicas?

```
head(arrange(propinas, desc(pct_propina)))
```

```
## # A tibble: 6 x 10
##      X1 cuenta_total propina fumador dia   momento num_personas
orden_propina
##   <dbl>        <dbl>   <dbl> <chr>   <fct> <chr>          <dbl>
<int>
## 1   173         7.25    5.15 Yes     Dom   Cena               2
229
## 2   179         9.6     4    Yes     Dom   Cena               2
199
## 3    68         3.07    1    Yes     Sab   Cena               1
1
## 4   233        11.6     3.39 No      Sab   Cena               2
```

```
167
## 5   184        23.2     6.5  Yes     Dom   Cena               4
238
## 6   110        14.3     4    Yes     Sab   Cena               2
195
## # ... with 2 more variables: f <dbl>, pct_propina <dbl>
```

```
propinas %>%
  filter(pct_propina > 25)
```

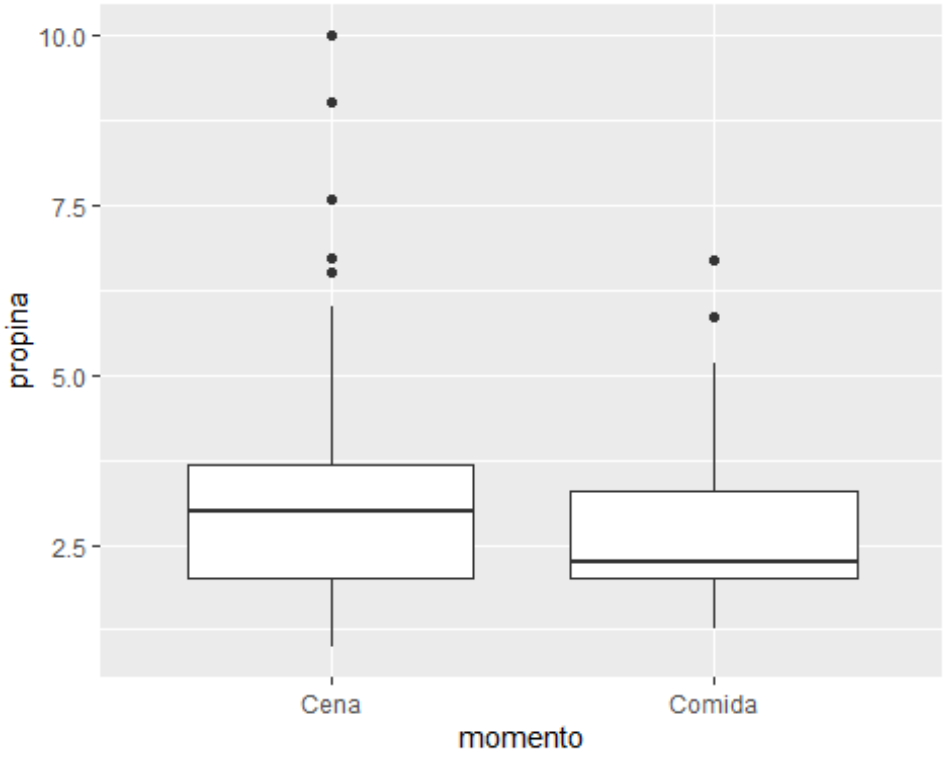
```
## # A tibble: 10 x 10
##       X1 cuenta_total propina fumador dia   momento num_personas
orden_propina
##    <dbl>        <dbl>   <dbl> <chr>   <fct> <chr>          <dbl>
<int>
##  1    52        10.3     2.6  No      Dom   Cena               2
113
##  2    68         3.07    1    Yes     Sab   Cena               1
1
##  3    94        16.3     4.3  Yes     Vie   Cena               2
211
##  4   110        14.3     4    Yes     Sab   Cena               2
195
##  5   150         7.51    2    No      Jue   Comida             2
59
##  6   173         7.25    5.15 Yes     Dom   Cena               2
229
##  7   179         9.6     4    Yes     Dom   Cena               2
199
##  8   184        23.2     6.5  Yes     Dom   Cena               4
238
##  9   222        13.4     3.48 Yes     Vie   Comida             2
172
## 10   233        11.6     3.39 No      Sab   Cena               2
167
## # ... with 2 more variables: f <dbl>, pct_propina <dbl>
```

Se puede observar que en realidad no fueron cuentas grandes y la mayoría son cuentas en la "cena" y de los fumadores

9. Haz una diagrama de caja y brazos para propina en dolares dependiendo del momento (comida o cena)

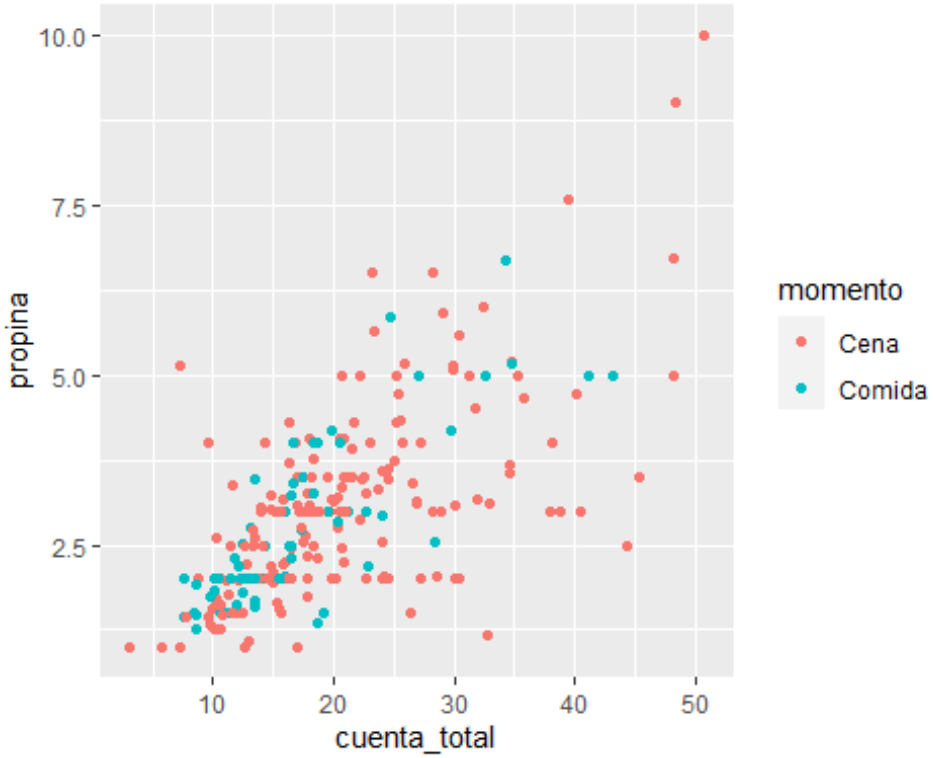
¿Cuál parece más grande? ¿Por qué? Haz otras gráficas si es necesario.

```
ggplot(propinas, aes(x = propina, y = momento)) +
  geom_boxplot() +
  coord_flip()
```



El diagrama de caja de la cena es más grande, esto se debe a que las propinas de la cena están más dispersos a la mediana de las propinas de la cena. Incluso se puede ver que tiene más datos atípicos.

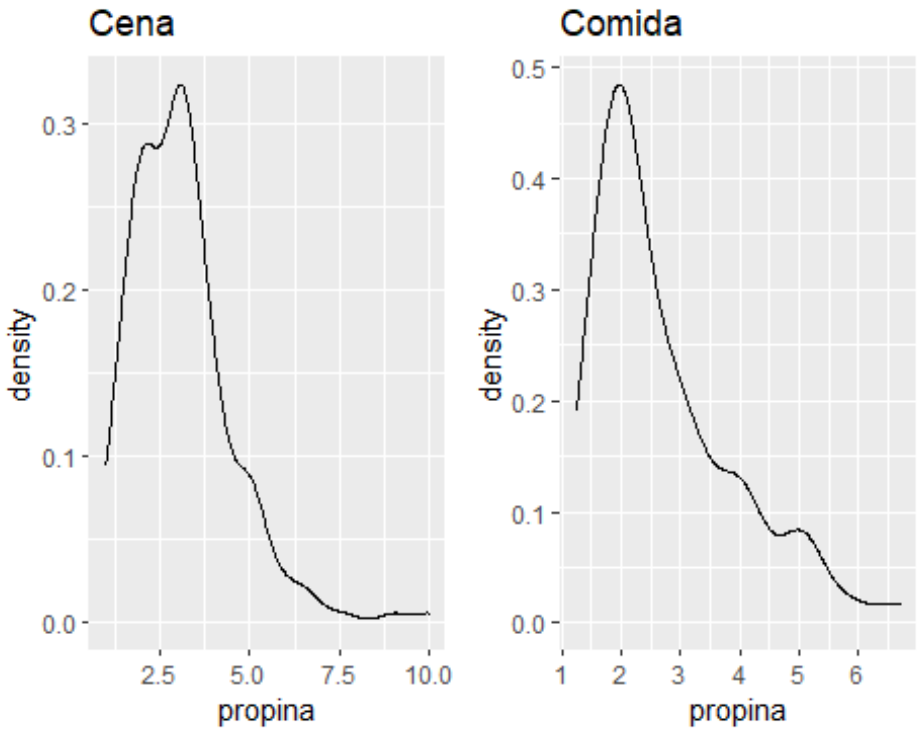
```
ggplot(propinas, aes(x = cuenta_total, y = propina, group = momento)) +
  geom_point(aes(color = momento), size = 1.5)
```



En la gráfica anterior se puede observar que hay más propinas asignadas a las cuentas totales, aunque también podemos ver que son datos más dispersos.

A continuación se graficarán las densidades de las agrupaciones de “Cena” y “Comida”.

```
cena <- propinas %>%
  filter(momento == 'Cena')

comida <- propinas %>%
  filter(momento == 'Comida')

g_cena <- ggplot(cena, aes(x = propina)) +
  geom_density() +
  ggtitle("Cena")

g_comida <- ggplot(comida, aes(x = propina)) +
  geom_density() +
  ggtitle("Comida")

g_cena + g_comida
```



Se puede observar que la curva que describe la distribución de “cena” es más ancha que la de comida, lo cual nos confirma que los datos están más “dispersos”