Tarea20

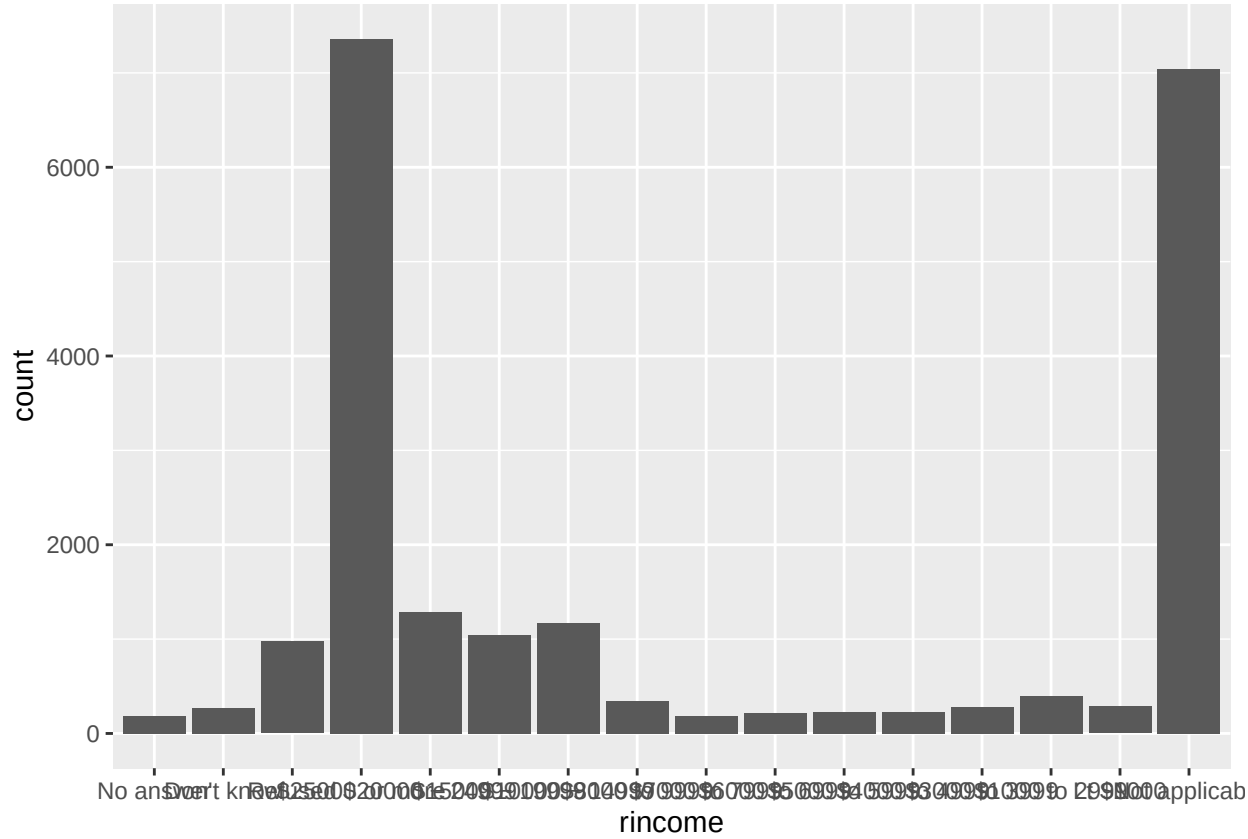
Yimmy Eman

2022-07-16

Pregunta 1

1. Explora la distribución de rincome (reported income). ¿Por qué el barchart es difícil de interpretar y qué podrías hacer para mejorar la visualización del dato?

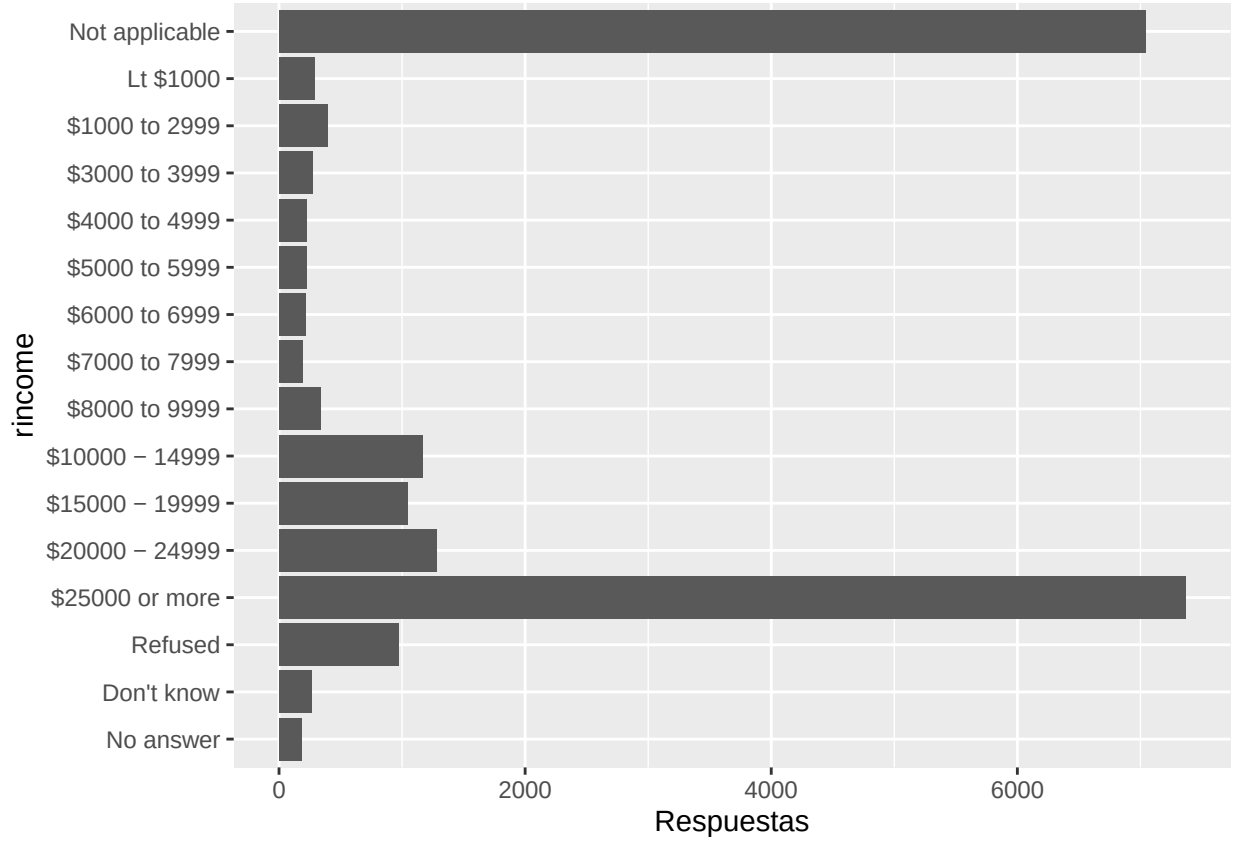
```
gss_cat %>%
  ggplot(aes(rincome)) +
  geom_bar()
```




```
# Imposible visualizar la observación.
```

```
# Aplicando un coord_flip para girar el gráfico mejoraría la lectura

gss_cat %>%
  ggplot(aes(rincome)) +
  geom_bar() +
  coord_flip() +
  labs(y = "Respuestas")
```



2. Identifica la religión relig más común del dataset. Identifica también el partido partyid más común del dataset.

```
(gss_cat %>%
  count(relig) %>%
  arrange(desc(n)))[1,]
```

```
## # A tibble: 1 x 2
##   relig          n
##   <fct>      <int>
## 1 Protestant 10846
```

```
(gss_cat %>%
  count(partyid) %>%
  arrange(desc(n)))[1,]
```

```
## # A tibble: 1 x 2
##   partyid         n
##   <fct>       <int>
## 1 Independent  4119
```

Pregunta 2

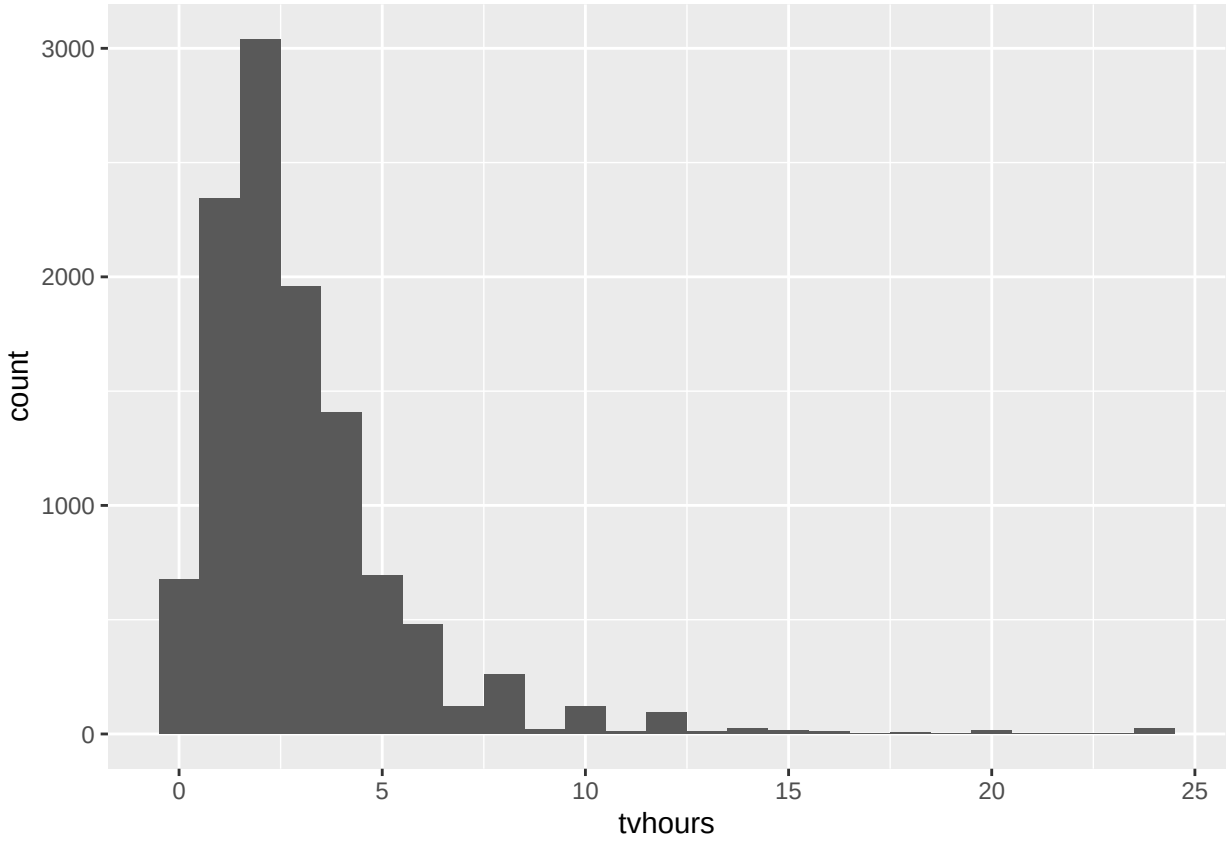
1. Investiga algunos números raros del campo tvhours. Deduce si la media es el mejor estadístico para resumir información de dicho campo.

```
summary(gss_cat$tvhours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   2.000   2.981   4.000  24.000   10146
```

```
# Algunos número son bastantes altos, es raro ver una persona 24 horas seguida viendo tv

gss_cat %>%
  filter(!is.na(tvhours)) %>% # Quitemos los valores NA
  ggplot(aes(tvhours))+
  geom_histogram(binwidth = 1)
```



2. Para cada factor de gss_cat identifica si el orden de los niveles es arbitrario o está bien definido.

```r
# Primero vemos los que son factores
str(gss_cat)
```

```
## tibble [21,483 x 9] (S3: tbl_df/tbl/data.frame)
##  $ year   : int [1:21483] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ marital: Factor w/ 6 levels "No answer","Never married",..: 2 4 5 2 4 6 2 4 6 6 ...
##  $ age    : int [1:21483] 26 48 67 39 25 25 36 44 44 47 ...
##  $ race   : Factor w/ 4 levels "Other","Black",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ rincome: Factor w/ 16 levels "No answer","Don't know",..: 8 8 16 16 16 5 4 9 4 4 ...
##  $ partyid: Factor w/ 10 levels "No answer","Don't know",..: 6 5 7 6 9 10 5 8 9 4 ...
##  $ relig  : Factor w/ 16 levels "No answer","Don't know",..: 15 15 15 6 12 15 5 15 15 15 ...
##  $ denom  : Factor w/ 30 levels "No answer","Don't know",..: 25 23 3 30 30 25 30 15 4 25 ...
##  $ tvhours: int [1:21483] 12 NA 2 4 1 NA 3 NA 0 3 ...
```

```r
# Entre lo que son factores tenemos:
levels(gss_cat$marital)
```

```
## [1] "No answer"     "Never married" "Separated"     "Divorced"
## [5] "Widowed"       "Married"
```

```r
levels(gss_cat$race) # No hay orden
```

```
## [1] "Other"          "Black"          "White"          "Not applicable"
```

```r
levels(gss_cat$rincome) # De mayor a menor
```

```
##  [1] "No answer"      "Don't know"     "Refused"        "$25000 or more"
##  [5] "$20000 - 24999" "$15000 - 19999" "$10000 - 14999" "$8000 to 9999"
##  [9] "$7000 to 7999"  "$6000 to 6999"  "$5000 to 5999"  "$4000 to 4999"
## [13] "$3000 to 3999"  "$1000 to 2999"  "Lt $1000"       "Not applicable"
```

```r
levels(gss_cat$partyid) # están ordenados a partir de más republicano a más demócrata.
```

```
##  [1] "No answer"          "Don't know"         "Other party"
##  [4] "Strong republican"  "Not str republican" "Ind,near rep"
##  [7] "Independent"        "Ind,near dem"       "Not str democrat"
## [10] "Strong democrat"
```

```r
levels(gss_cat$relig) # No tiene orden
```

```
##  [1] "No answer"               "Don't know"
##  [3] "Inter-nondenominational" "Native american"
##  [5] "Christian"               "Orthodox-christian"
##  [7] "Moslem/islam"            "Other eastern"
##  [9] "Hinduism"                "Buddhism"
## [11] "Other"                   "None"
## [13] "Jewish"                  "Catholic"
## [15] "Protestant"              "Not applicable"
```

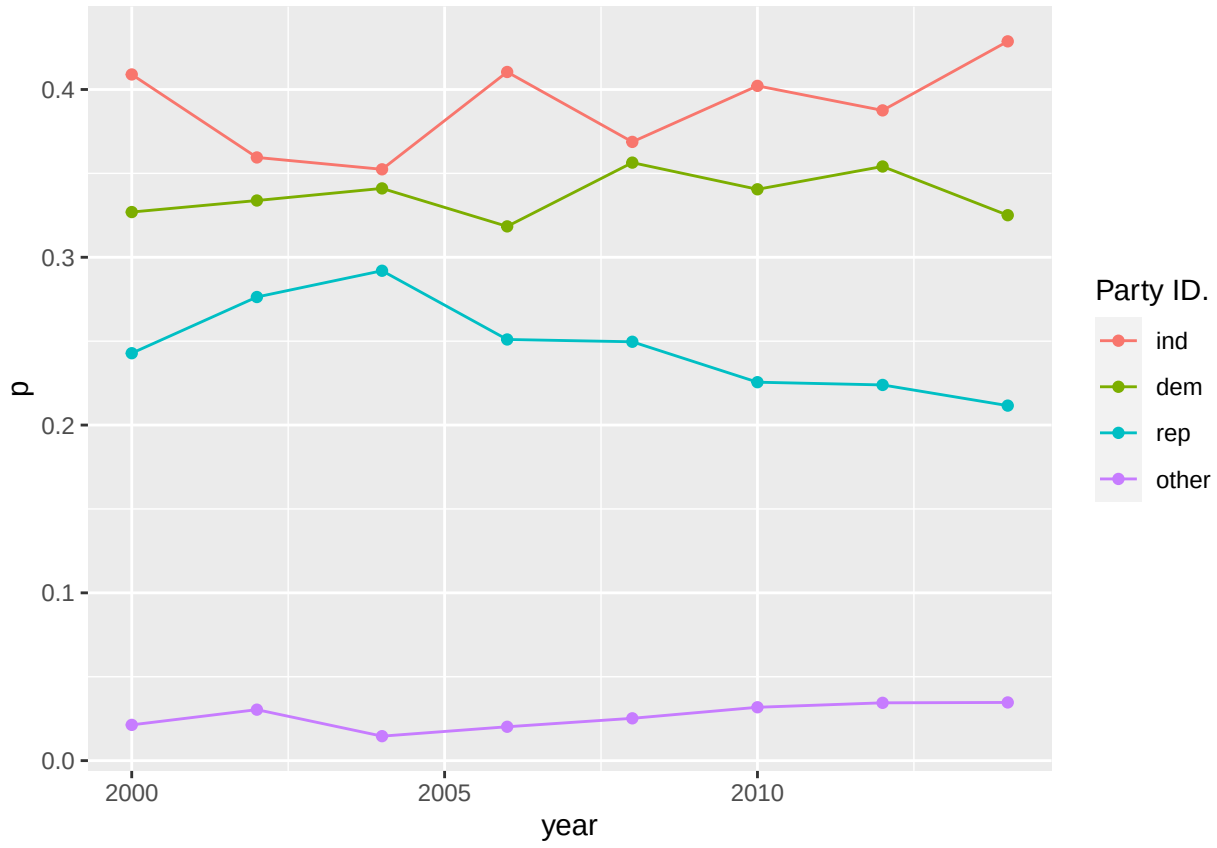
```
levels(gss_cat$denom)
```

```
##  [1] "No answer"            "Don't know"           "No denomination"
##  [4] "Other"                "Episcopal"            "Presbyterian-dk wh"
##  [7] "Presbyterian, merged" "Other presbyterian"   "United pres ch in us"
## [10] "Presbyterian c in us" "Lutheran-dk which"    "Evangelical luth"
## [13] "Other lutheran"       "Wi evan luth synod"   "Lutheran-mo synod"
## [16] "Luth ch in america"   "Am lutheran"          "Methodist-dk which"
## [19] "Other methodist"      "United methodist"     "Afr meth ep zion"
## [22] "Afr meth episcopal"   "Baptist-dk which"     "Other baptists"
## [25] "Southern baptist"     "Nat bapt conv usa"    "Nat bapt conv of am"
## [28] "Am bapt ch in usa"    "Am baptist asso"      "Not applicable"
```

Pregunta 3

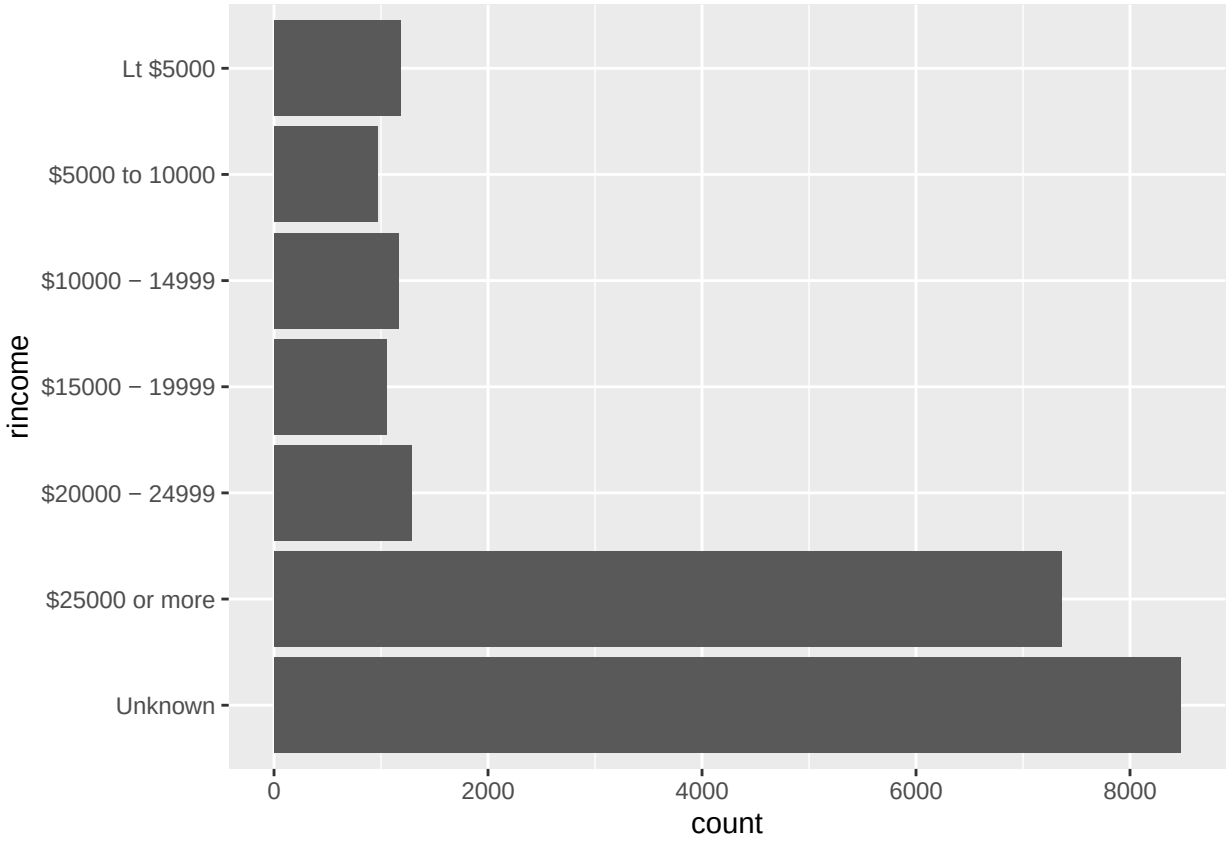
1. Identifica cómo han cambiado las proporciones de los tres partidos políticos Democrat, Republican, y Independent a lo largo del tiempo.

```
gss_cat %>%
  mutate(
    partyid =
      fct_collapse(partyid,
        other = c("No answer", "Don't know", "Other party"),
        rep = c("Strong republican", "Not str republican"),
        ind = c("Ind,near rep", "Independent", "Ind,near dem"),
        dem = c("Not str democrat", "Strong democrat")
      )
  ) %>%
  count(year, partyid) %>%
  group_by(year) %>%
  mutate(p = n / sum(n)) %>%
  ggplot(aes(
    x = year, y = p,
    colour = fct_reorder2(partyid, year, p)
  )) +
  geom_point() +
  geom_line() +
  labs(colour = "Party ID.")
```



2. ¿Cómo podríamos reducir el factor rincome en un conjunto menor de categorías?

```
library("stringr")
gss_cat %>%
  mutate(
    rincome =
      fct_collapse(
        rincome,
        `Unknown` = c("No answer", "Don't know", "Refused", "Not applicable"),
        `Lt $5000` = c("Lt $1000", str_c(
          "$", c("1000", "3000", "4000"),
          " to ", c("2999", "3999", "4999")
        )),
        `$5000 to 10000` = str_c(
          "$", c("5000", "6000", "7000", "8000"),
          " to ", c("5999", "6999", "7999", "9999")
        )
      )
  ) %>%
  ggplot(aes(x = rincome)) +
  geom_bar() +
  coord_flip()
```


rincome
Lt $5000
$5000 to 10000
$10000 − 14999
$15000 − 19999
$20000 − 24999
$25000 or more
Unknown
0
2000
4000
6000
8000
count