Lecture Assignment 17

Viraj Vijaywargiya

2022-05-27

```
library(tidyverse)
```

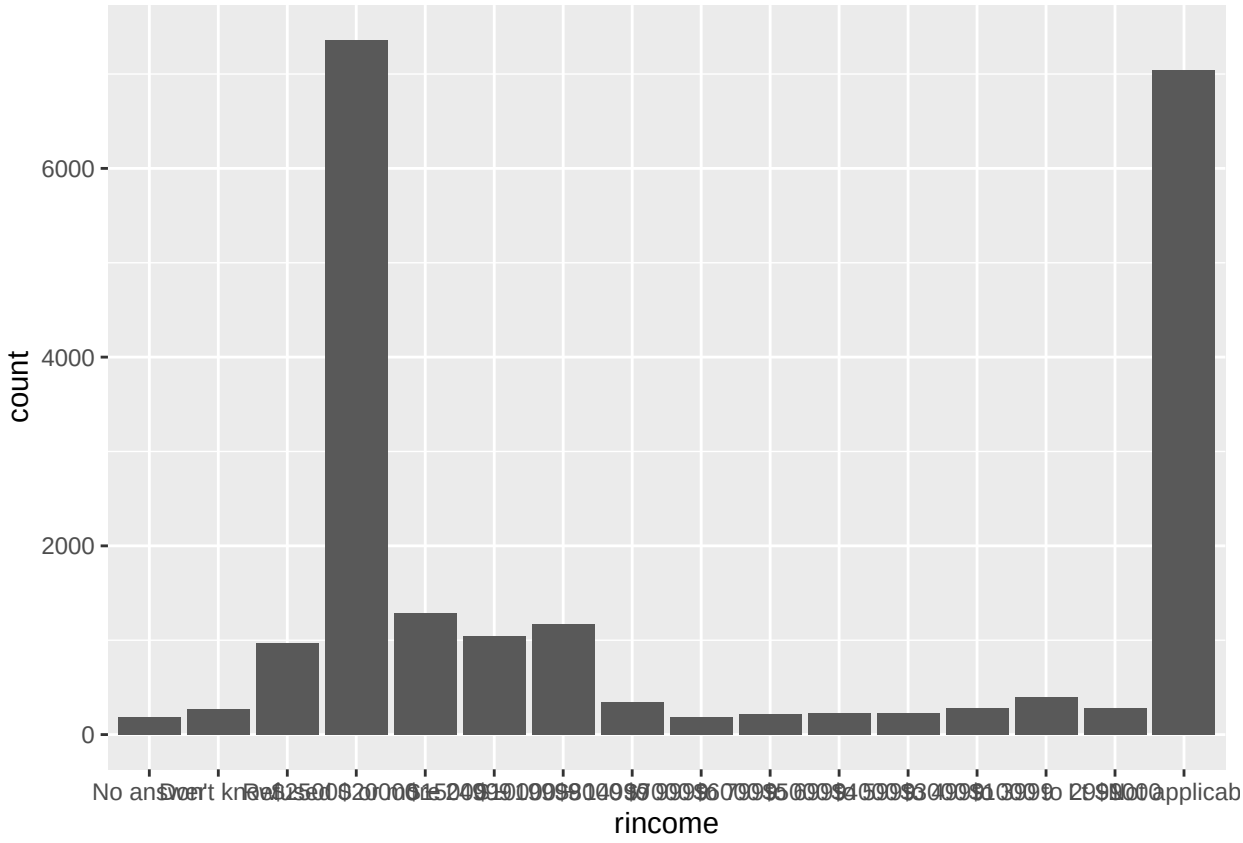
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

Part 15.3.1

Question 1

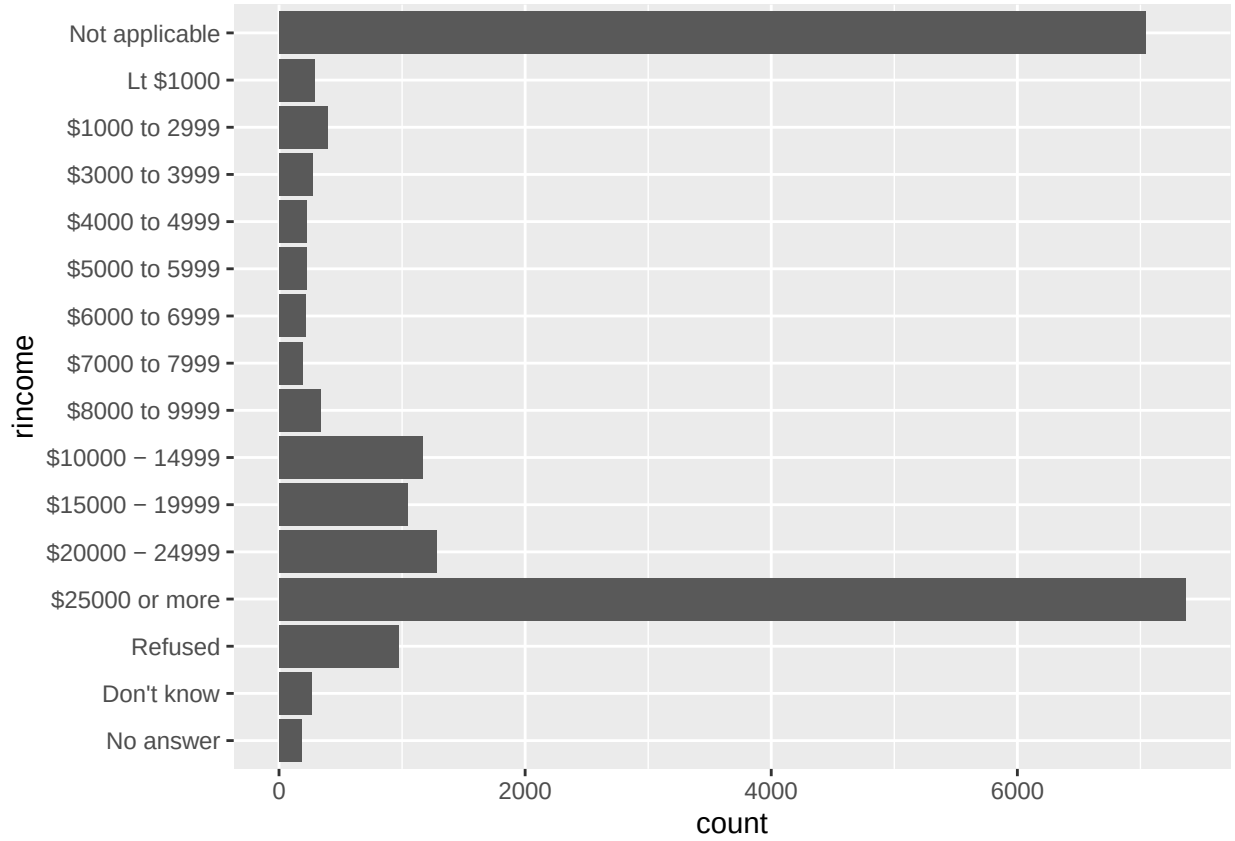
Exploring the distribution of **rincome** with default bar chart,

```
ggplot(gss_cat, aes(rincome)) +
  geom_bar()
```



With the default bar chart, the overlapping of labels on the x axis makes it hard to understand. We can improve the plot by flipping the x and y axis.

```
ggplot(gss_cat, aes(rincome)) +
  geom_bar() +
  coord_flip()
```



Question 2

```
gss_cat %>%
  count(relig) %>%
  arrange(desc(n)) %>%
  head(1)
```

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

"Protestant" is the most common relig.

```
gss_cat %>%
  count(partyid) %>%
  arrange(desc(n)) %>%
  head(1)
```

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

"Independent" is the most common partyid.

Question 3

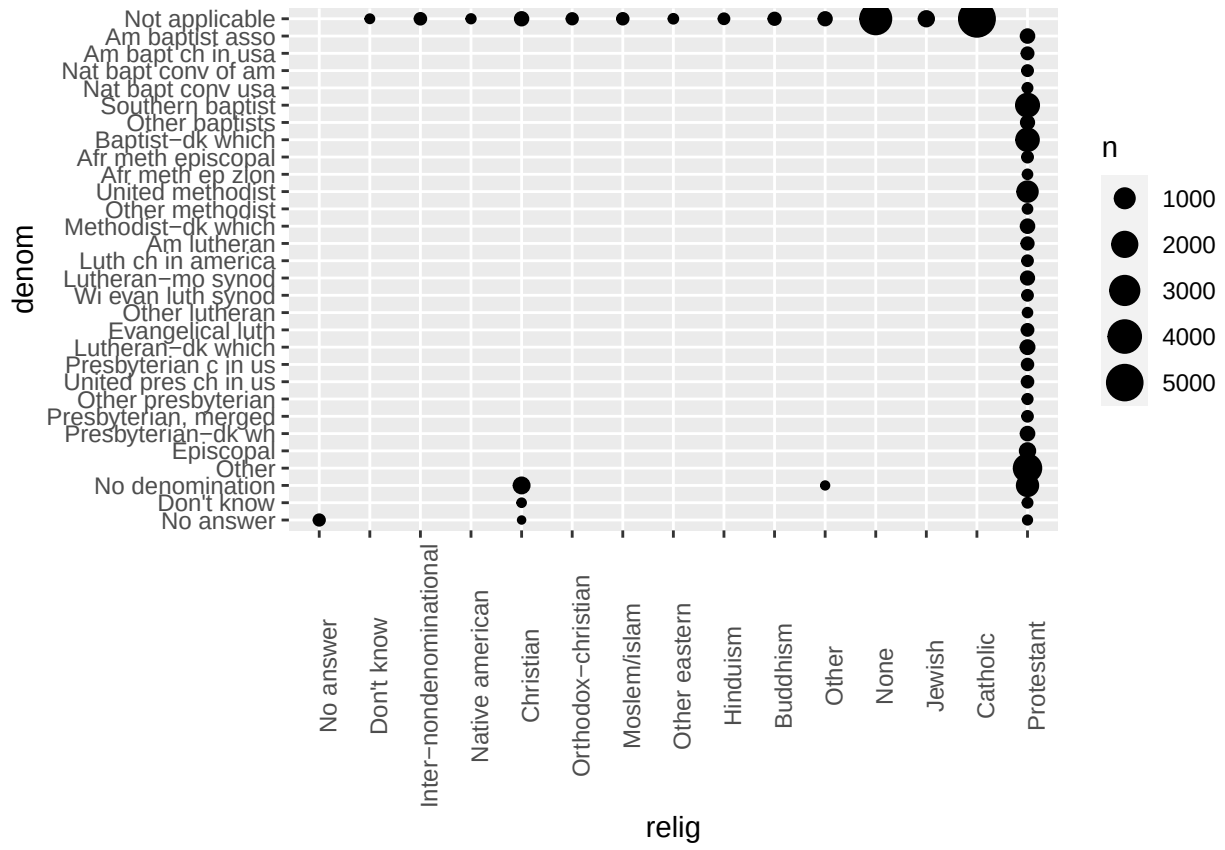
Finding out with a table,

```
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"
```

From this we can see that the relig, denom applies to, is "Protestant".

```
gss_cat %>%
  count(relig, denom) %>%
  ggplot(aes(x = relig, y = denom, size = n)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90))
```



From the scatterplot above, we can also see that the relig, denom applies to, is "Protestant".

Part 15.4.1

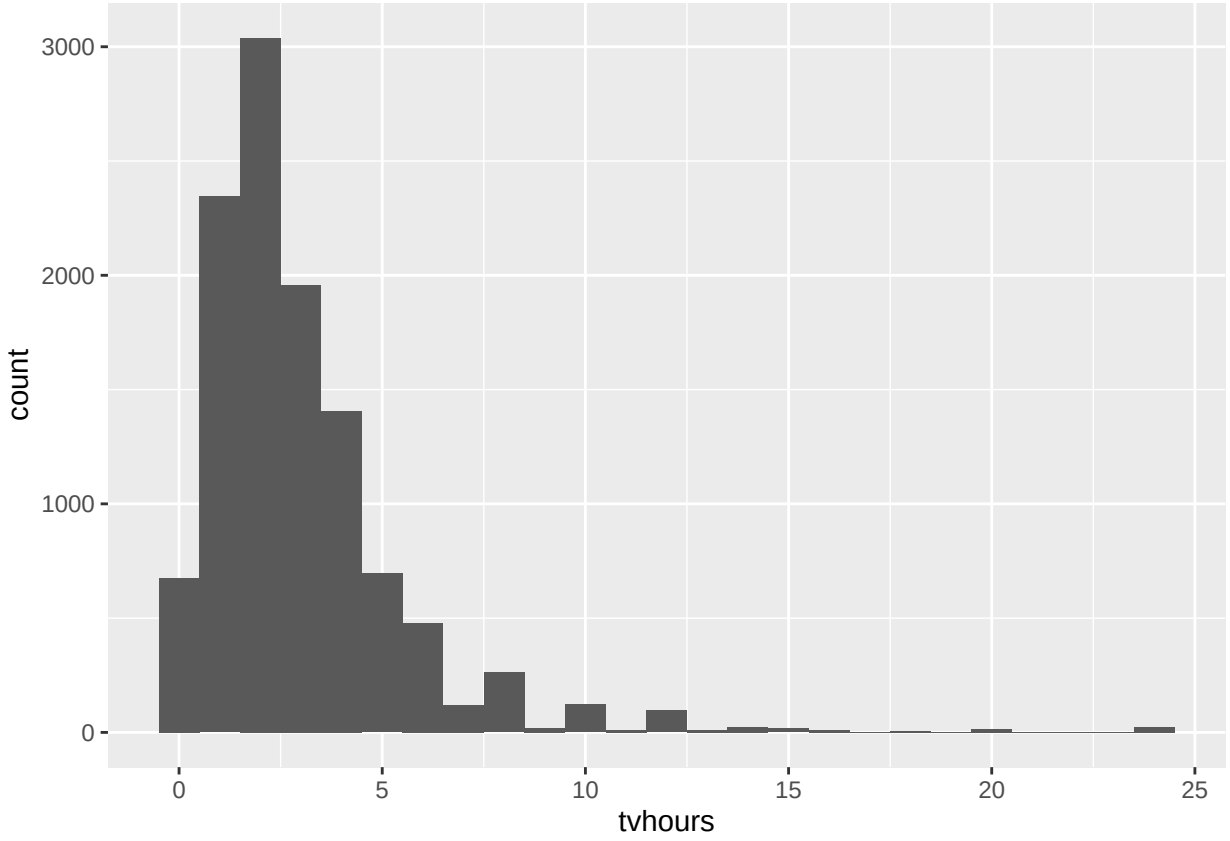
Question 1

Checking the summary and visualizing the data,

```
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
```

```
gss_cat %>%
  filter(!is.na(tvhours)) %>%
  ggplot(aes(x = tvhours)) +
  geom_histogram(binwidth = 1)
```



The tv hours and the mean looks fine to me. However, the potential of mean depends on what it is used for.

Question 2

Factors in gss_cat,

```
keep(gss_cat, is.factor) %>%
  names()
```

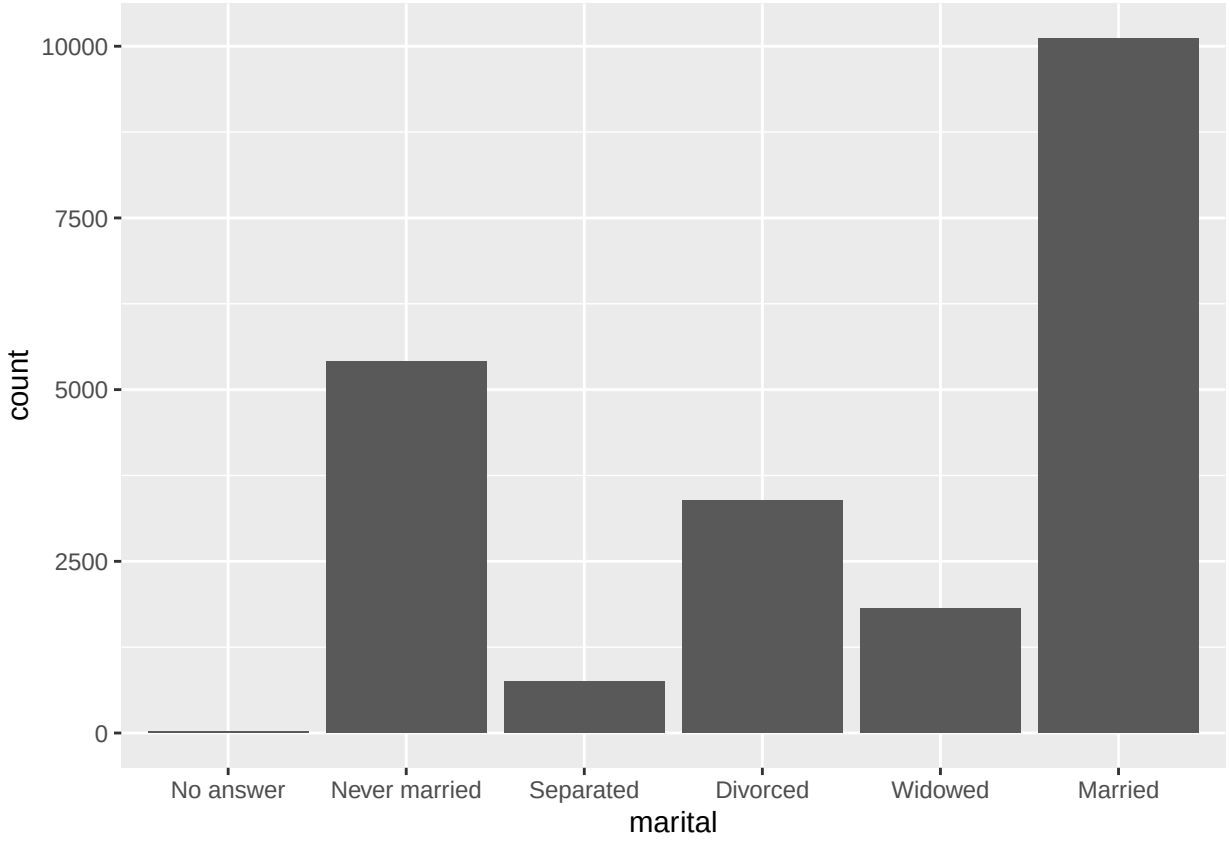
```
## [1] "marital" "race"    "rincome" "partyid" "relig"   "denom"
```

For **marital**,

```
levels(gss_cat$marital)
```

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

```
ggplot(gss_cat, aes(marital)) +
  geom_bar()
```



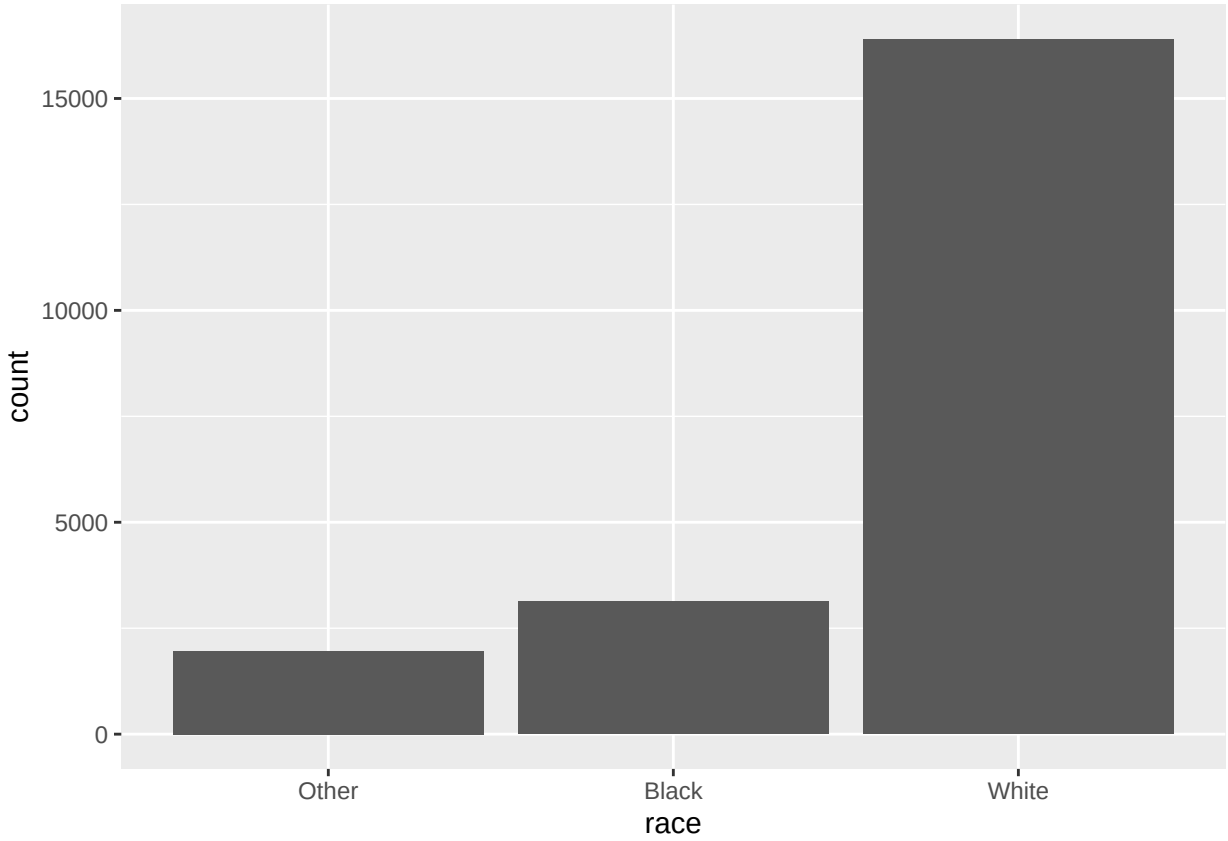
The order of the levels for **marital** can be principled or arbitrary. This is because the way it is ordered does make a little sense but at the same time, it is not very principled.

For **race**,

```
levels(gss_cat$race)
```

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

```
ggplot(gss_cat, aes(race)) +
  geom_bar()
```



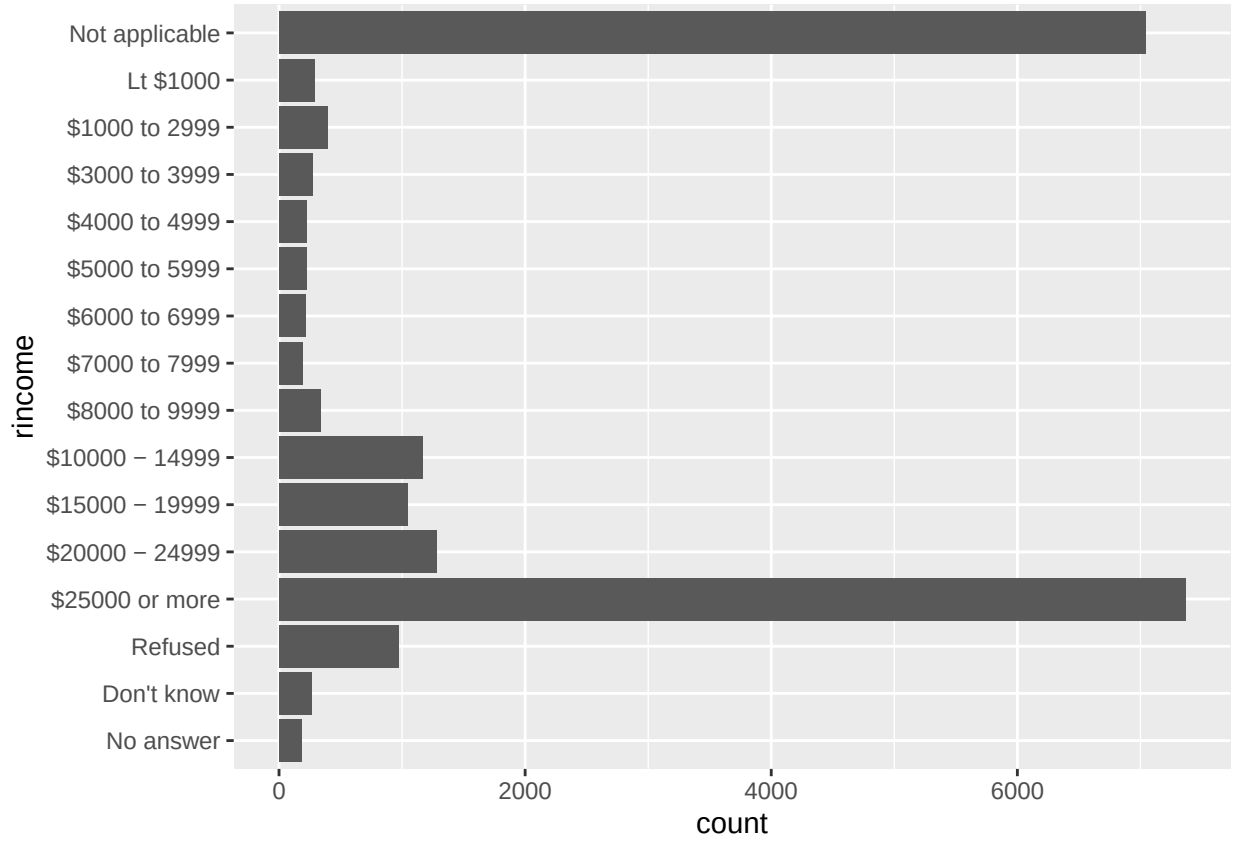
The order of the levels for **race** is principled as it is ordered by counts.

For **rincome**,

```
levels(gss_cat$rincome)
```

```
##  [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"
```

```
ggplot(gss_cat, aes(rincome)) +
  geom_bar() +
  coord_flip()
```



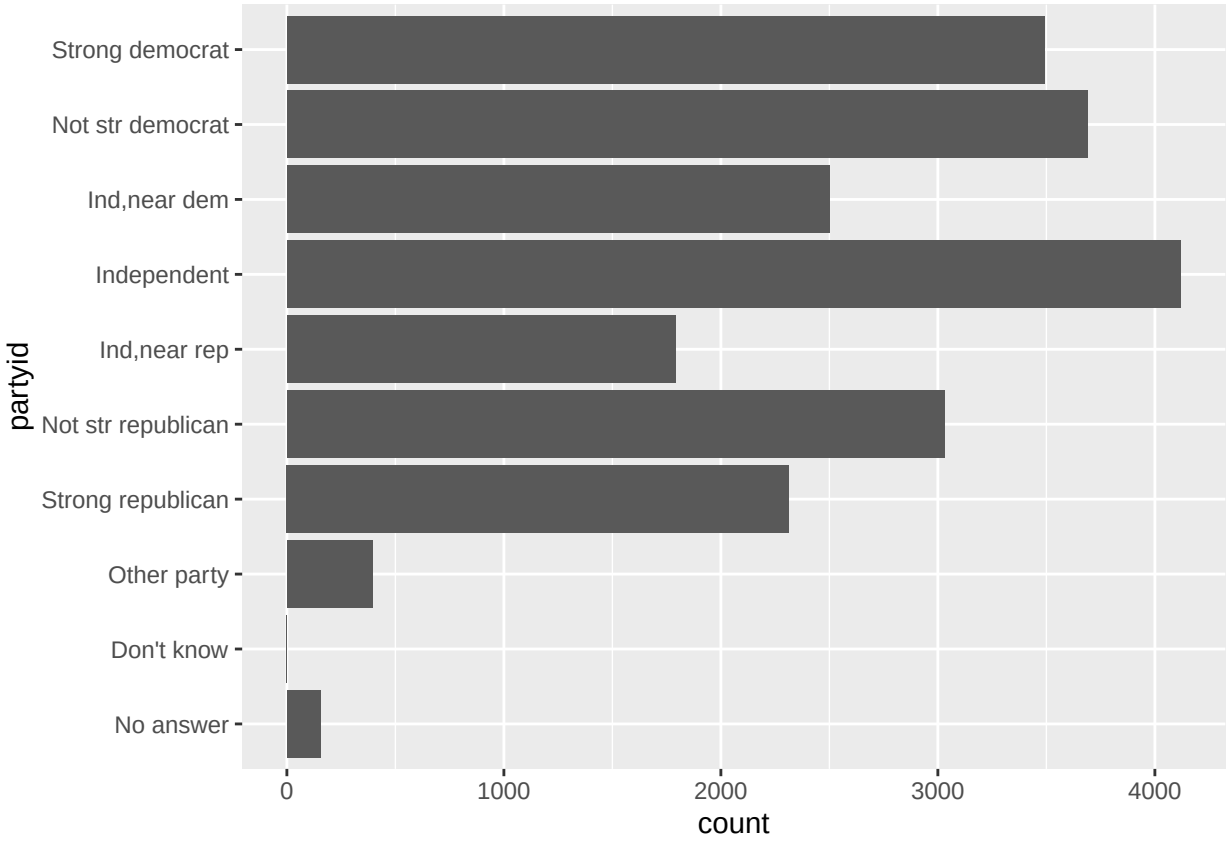
The order of the levels for **rincome** is principled if we exclude "Not applicable", "Refused", "Don't know", and "No answer". The **rincome** is ordered in decreasing order of income.

For **partyid**,

```
levels(gss_cat$partyid)
```

```
##  [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"
```

```
ggplot(gss_cat, aes(partyid)) +
  geom_bar() +
  coord_flip()
```



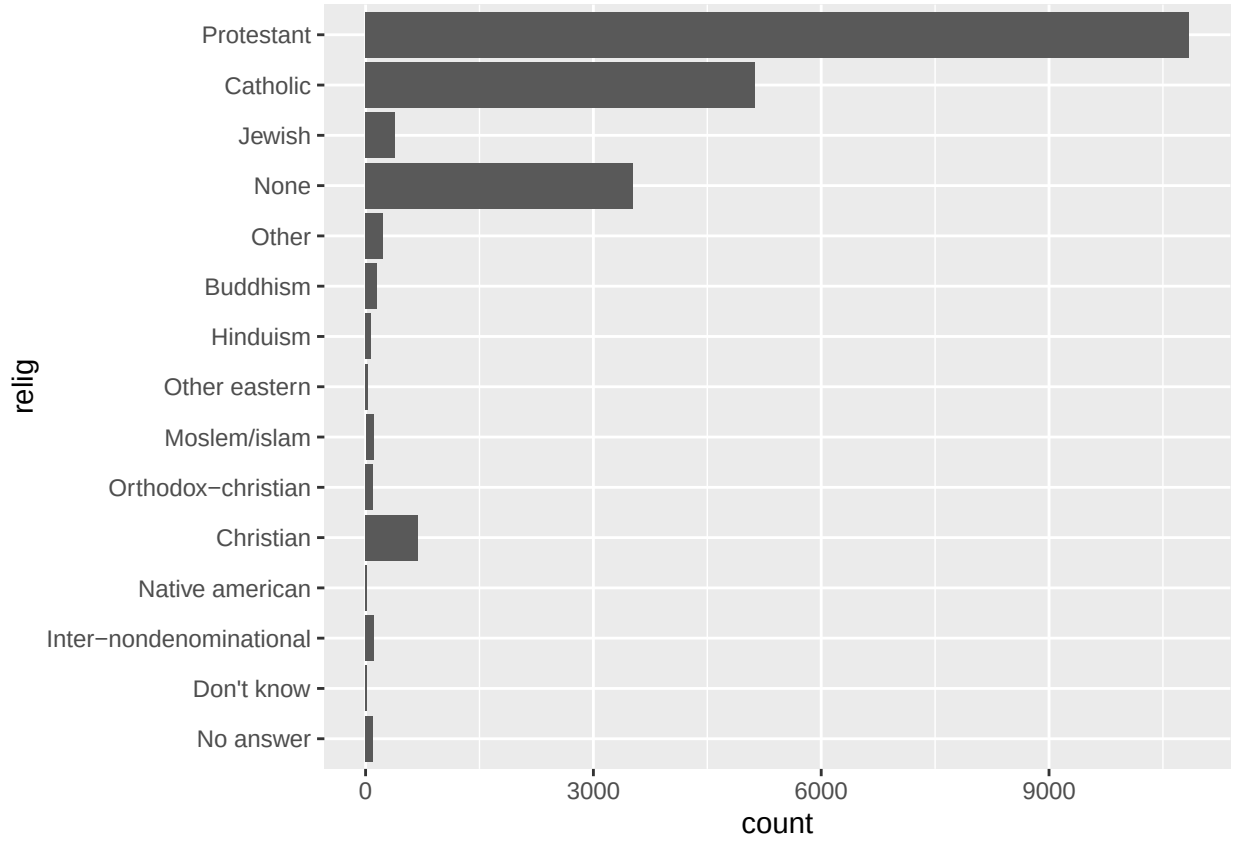
The order of the levels for **partyid** is principled as it is ordered from "Strong republican" to "Strong democrat".

For **relig**,

```
levels(gss_cat$relig)
```

```
##  [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"
```

```
ggplot(gss_cat, aes(relig)) +
  geom_bar() +
  coord_flip()
```



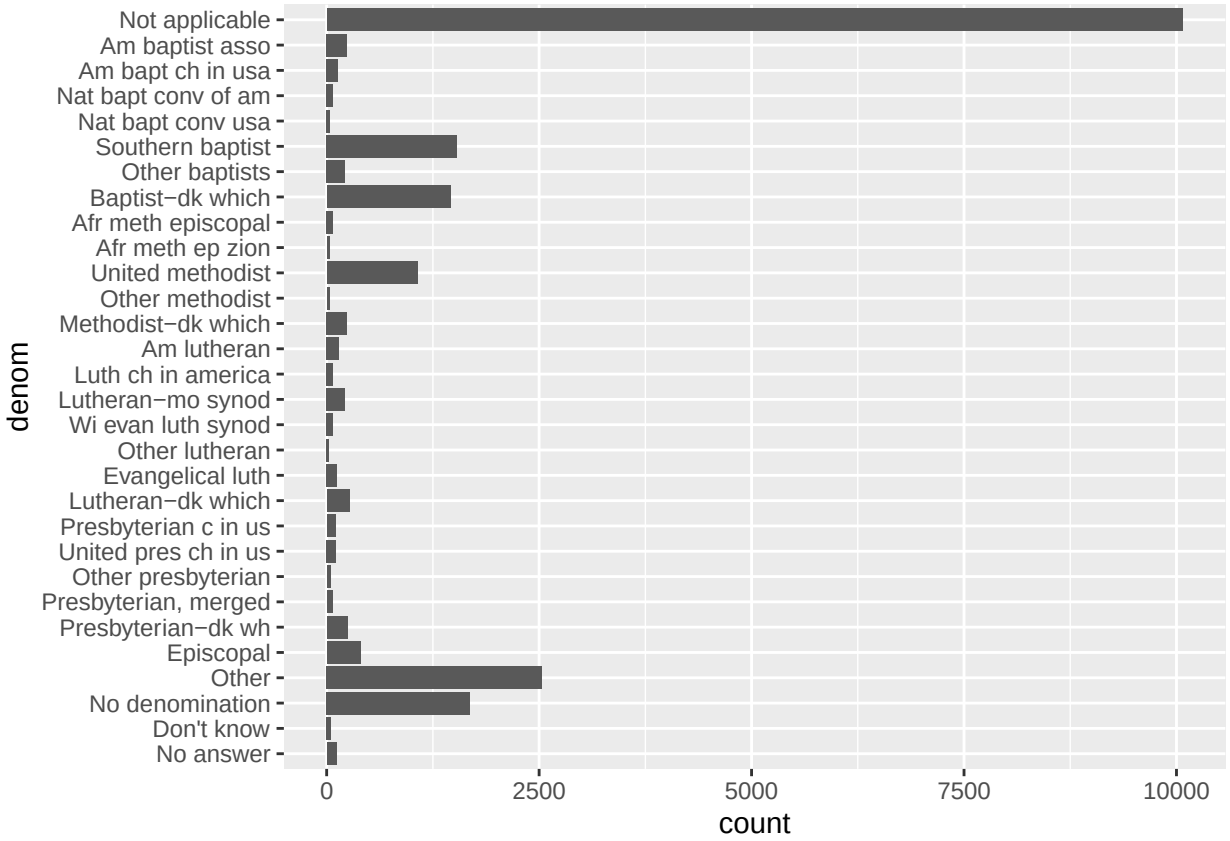
The order of the levels for **relig** is arbitrary as there is no natural ordering.

For **denom**,

```
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"
```

```
ggplot(gss_cat, aes(denom)) +
  geom_bar() +
  coord_flip()
```



The order of the levels for **denom** is also arbitrary as there is no natural ordering.