Factors

Here are some examples from Chapter 15. The examples are related to the General Social Survey from NORC at the Unversity of Chicago.

```r
library(tidyverse)
library(forcats)

gss_cat
```

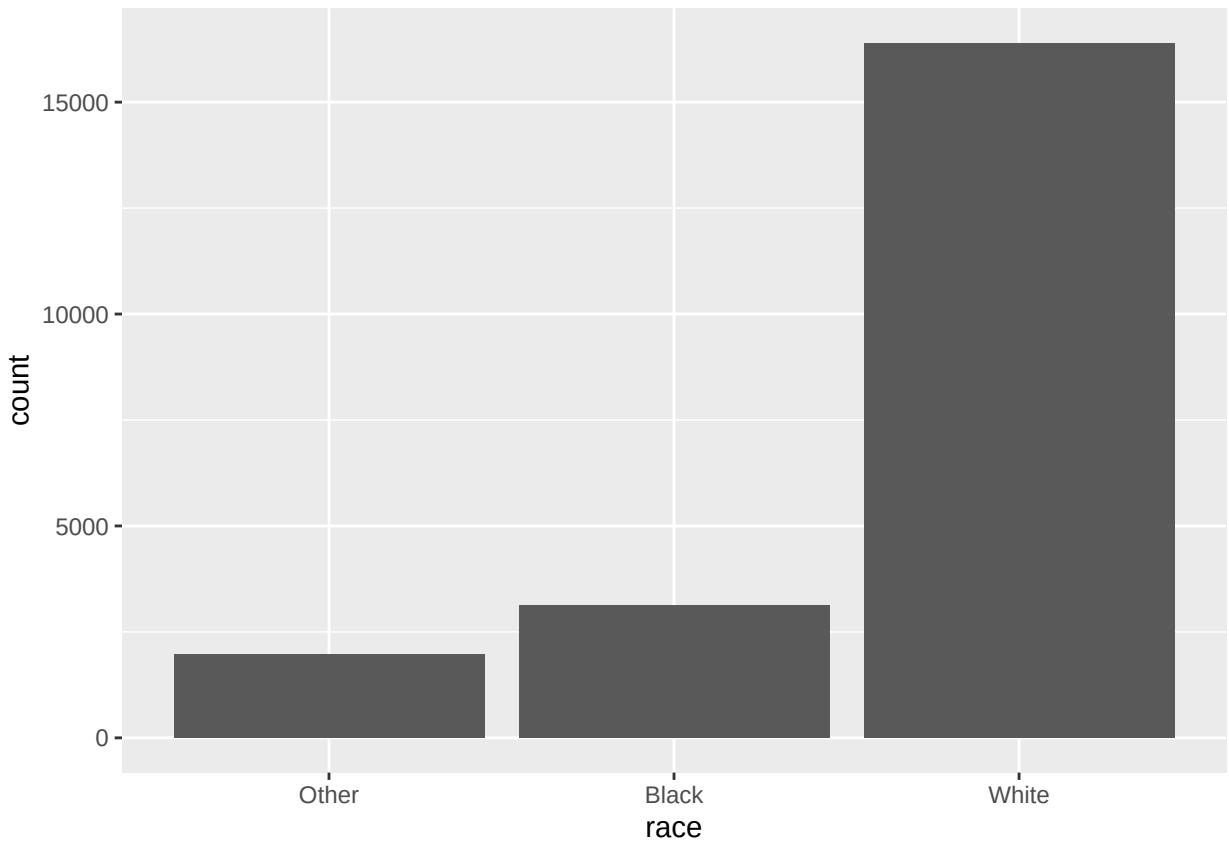
```
## # A tibble: 21,483 x 9
##     year marital     age race  rincome   partyid    relig   denom   tvhours
##    <int> <fct>     <int> <fct> <fct>     <fct>      <fct>   <fct>     <int>
##  1  2000 Never ma~    26 White $8000 to~ Ind,near ~ Protes~ Southe~      12
##  2  2000 Divorced     48 White $8000 to~ Not str r~ Protes~ Baptis~      NA
##  3  2000 Widowed      67 White Not appl~ Independe~ Protes~ No den~       2
##  4  2000 Never ma~    39 White Not appl~ Ind,near ~ Orthod~ Not ap~       4
##  5  2000 Divorced     25 White Not appl~ Not str d~ None    Not ap~       1
##  6  2000 Married      25 White $20000 -~ Strong de~ Protes~ Southe~      NA
##  7  2000 Never ma~    36 White $25000 o~ Not str r~ Christ~ Not ap~       3
##  8  2000 Divorced     44 White $7000 to~ Ind,near ~ Protes~ Luther~      NA
##  9  2000 Married      44 White $25000 o~ Not str d~ Protes~ Other         0
## 10  2000 Married      47 White $25000 o~ Strong re~ Protes~ Southe~       3
## # ... with 21,473 more rows
```

```r
gss_cat %>%
  count(race)
```

```
## # A tibble: 3 x 2
##   race      n
##   <fct> <int>
## 1 Other  1959
## 2 Black  3129
## 3 White 16395
```

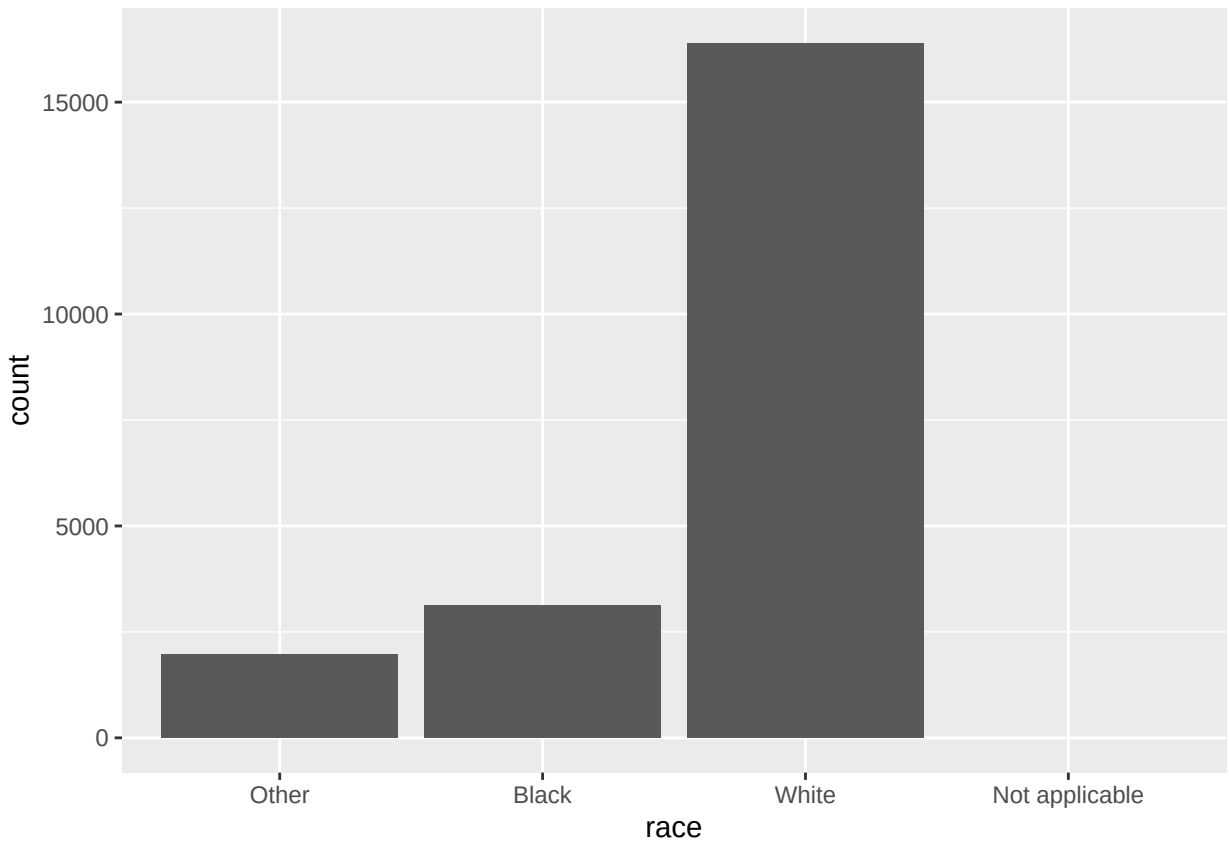
Factor variables are used to make bar charts. The *geom_bar()* counts the observations in each level of the factor.

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



Forcing NAs.

```
ggplot(gss_cat, aes(race)) +
  geom_bar() +
  scale_x_discrete(drop = FALSE)
```

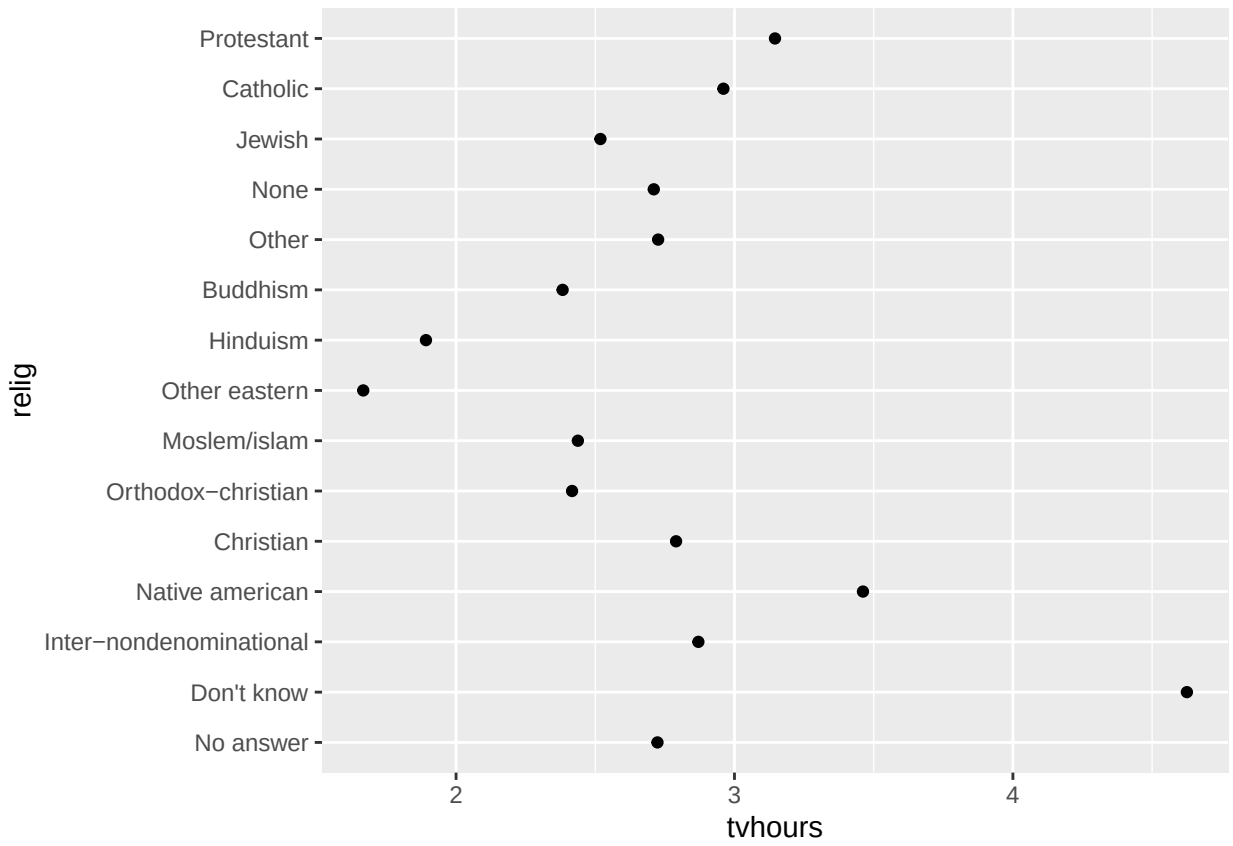


Modifying the order of a factor.

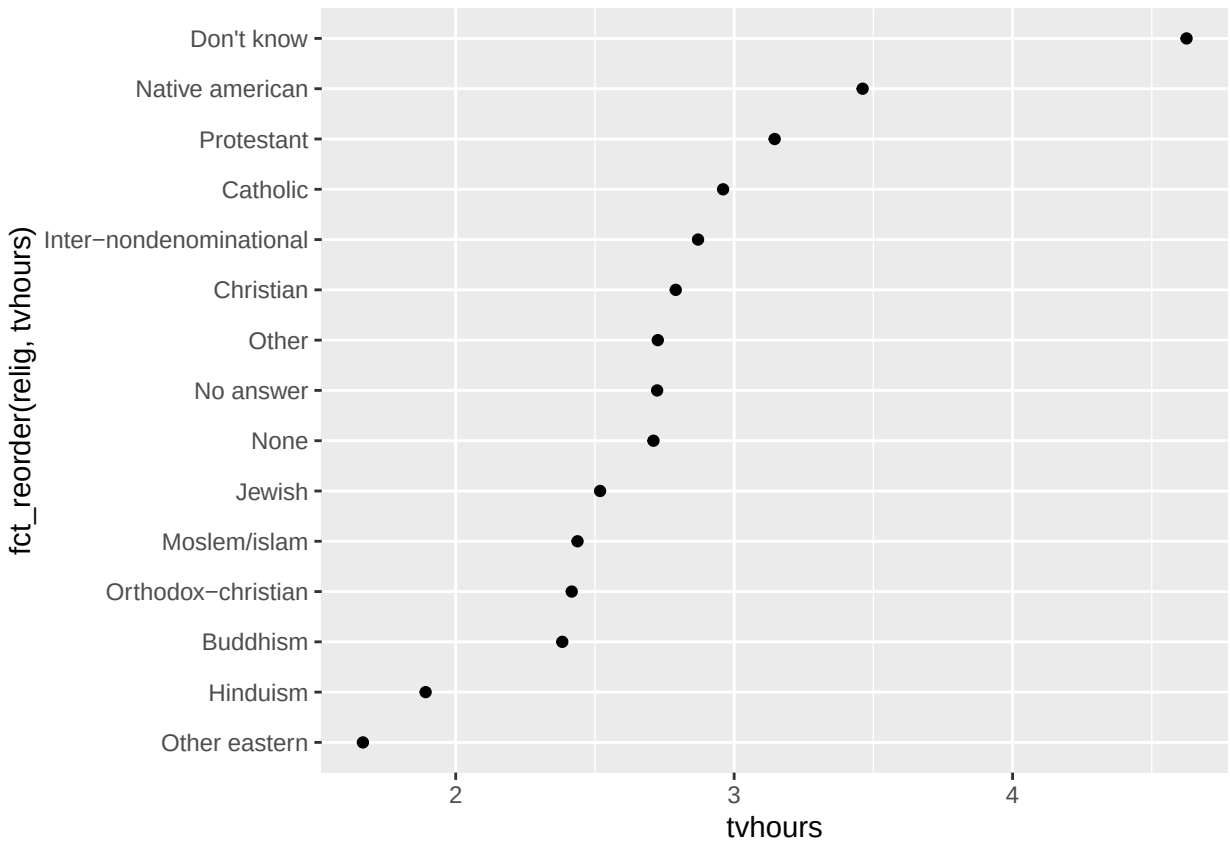
Examine tv watch time by religion.

```
relig_summary <- gss_cat %>%
  group_by(relig) %>%
  summarise(
    age = mean(age, na.rm = TRUE),
    tvhours = mean(tvhours, na.rm = TRUE),
    n = n()
  )

relig_summary %>% ggplot(aes(tvhours, relig)) + geom_point()
```



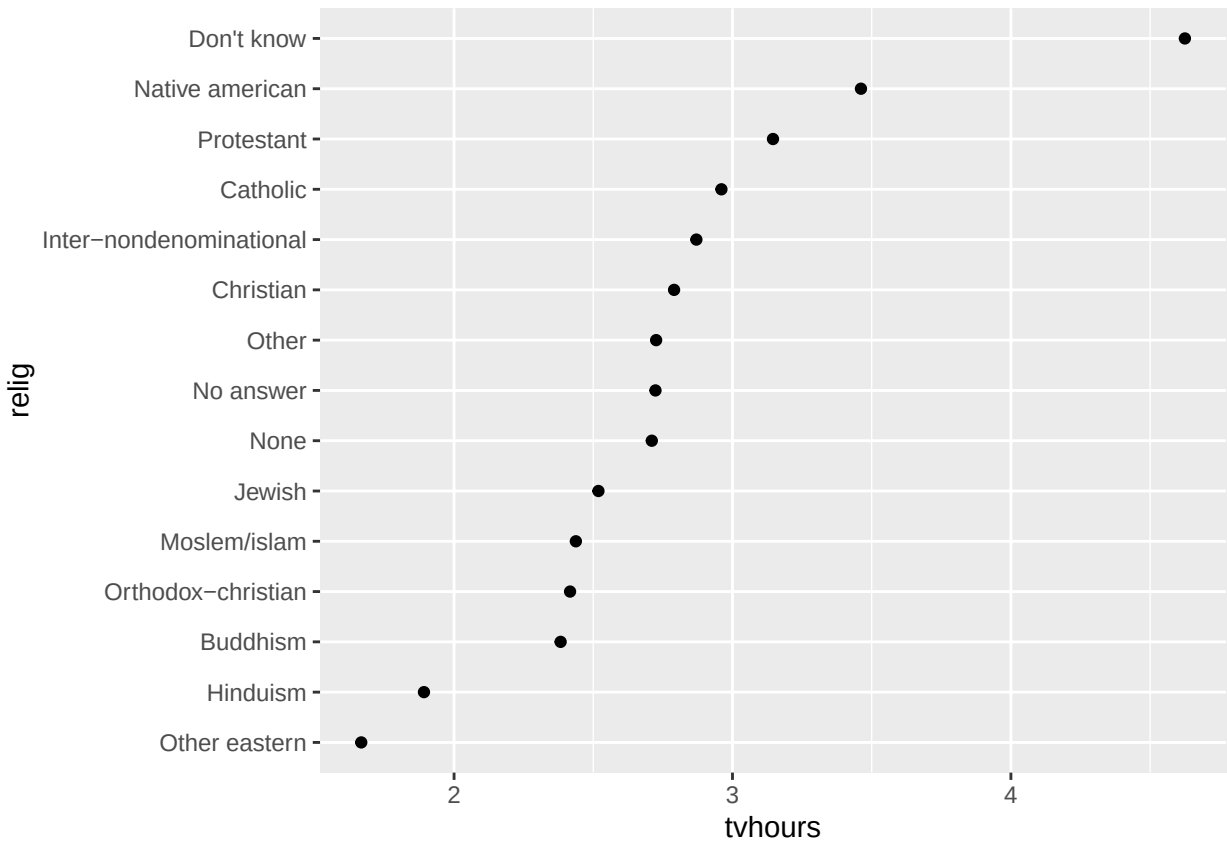

```
relig_summary %>% ggplot(aes(tvhours, fct_reorder(relig, tvhours))) +
  geom_point()
```



The *fct_reorder()* functon should be used in a mutate statement.

Same as the last code.

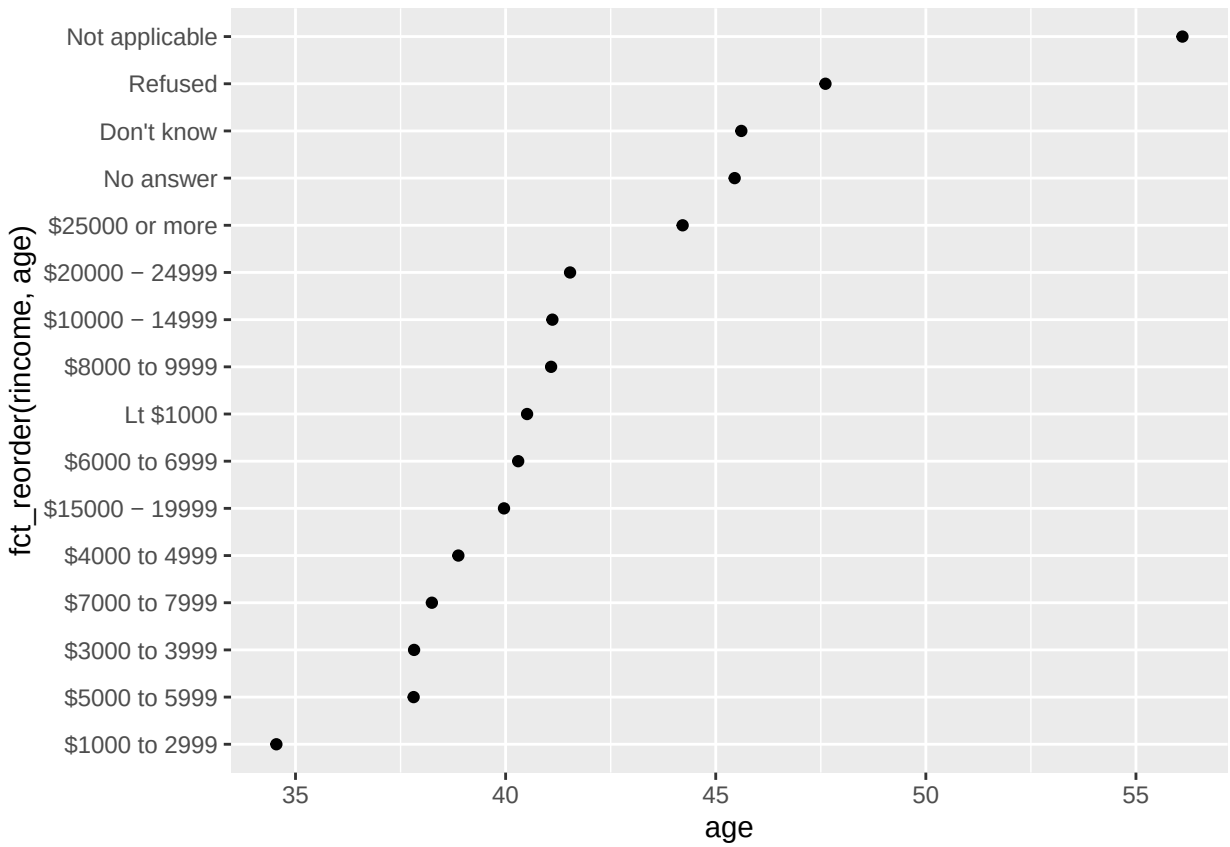
```
relig_summary %>%
  mutate(relig = fct_reorder(relig, tvhours)) %>%
  ggplot(aes(tvhours, relig)) +
    geom_point()
```



Now tv watch time by average age.

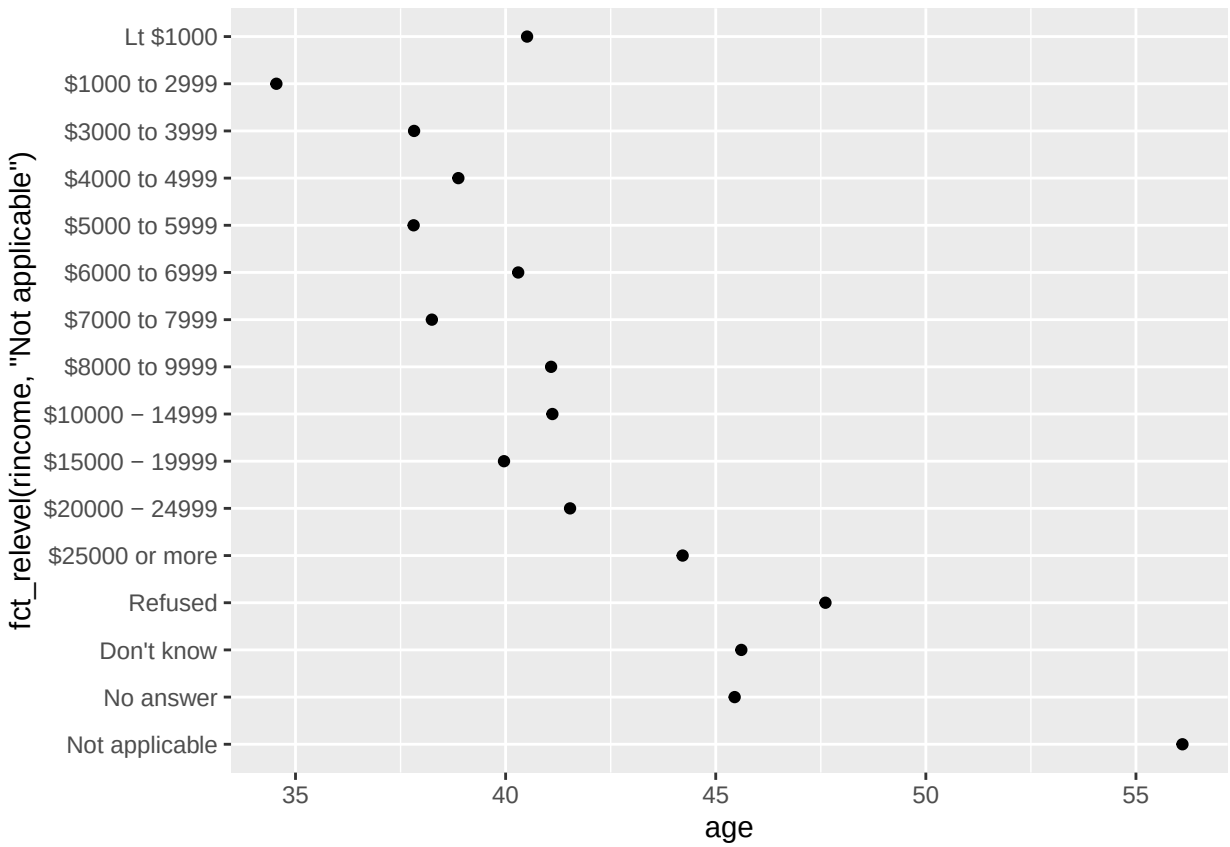
```
rincome_summary <- gss_cat %>%
  group_by(rincome) %>%
  summarise(
    age = mean(age, na.rm = TRUE),
    tvhours = mean(tvhours, na.rm = TRUE),
    n = n()
  )

rincome_summary %>% ggplot(aes(age, fct_reorder(rincome, age))) +
  geom_point()
```



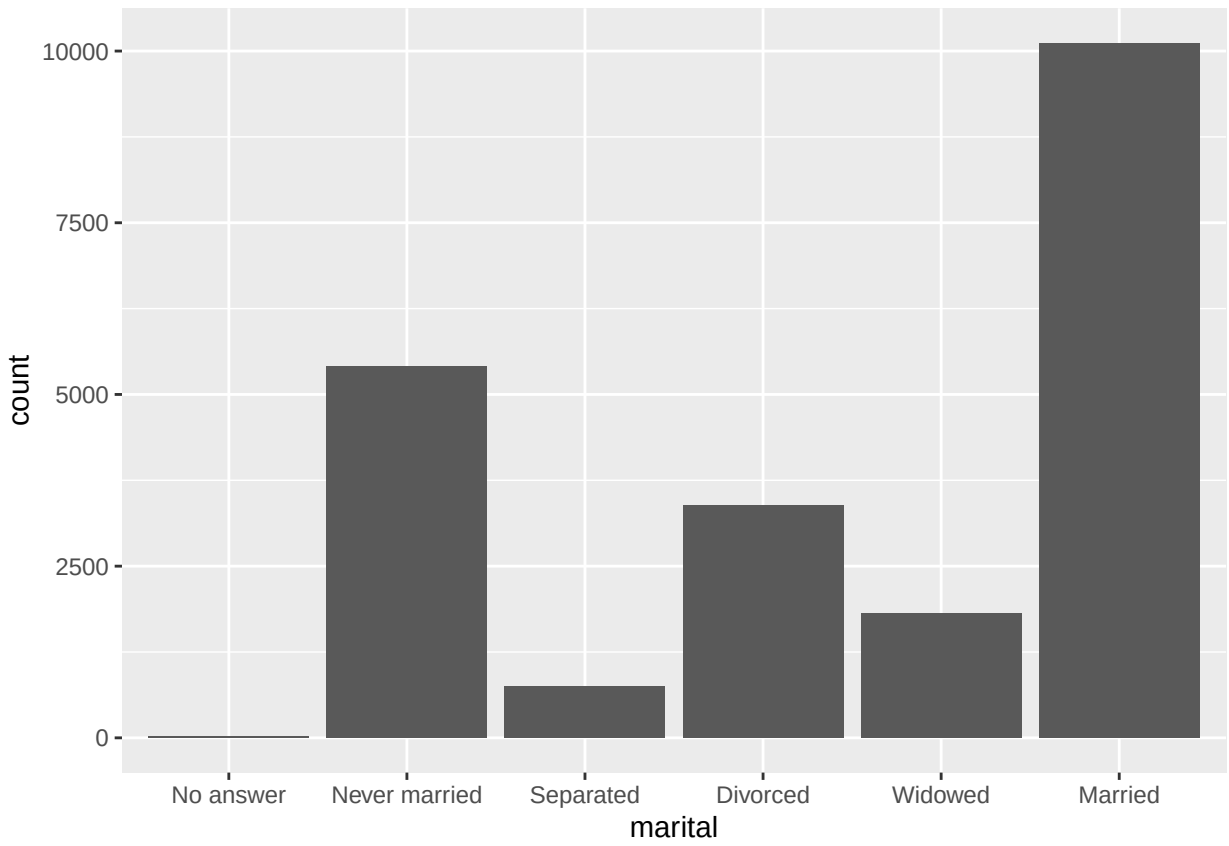
Does this make sense? What is wrong with this plot?

```
rincome_summary %>%ggplot(aes(age, fct_relevel(rincome, "Not applicable"))) +
  geom_point()
```

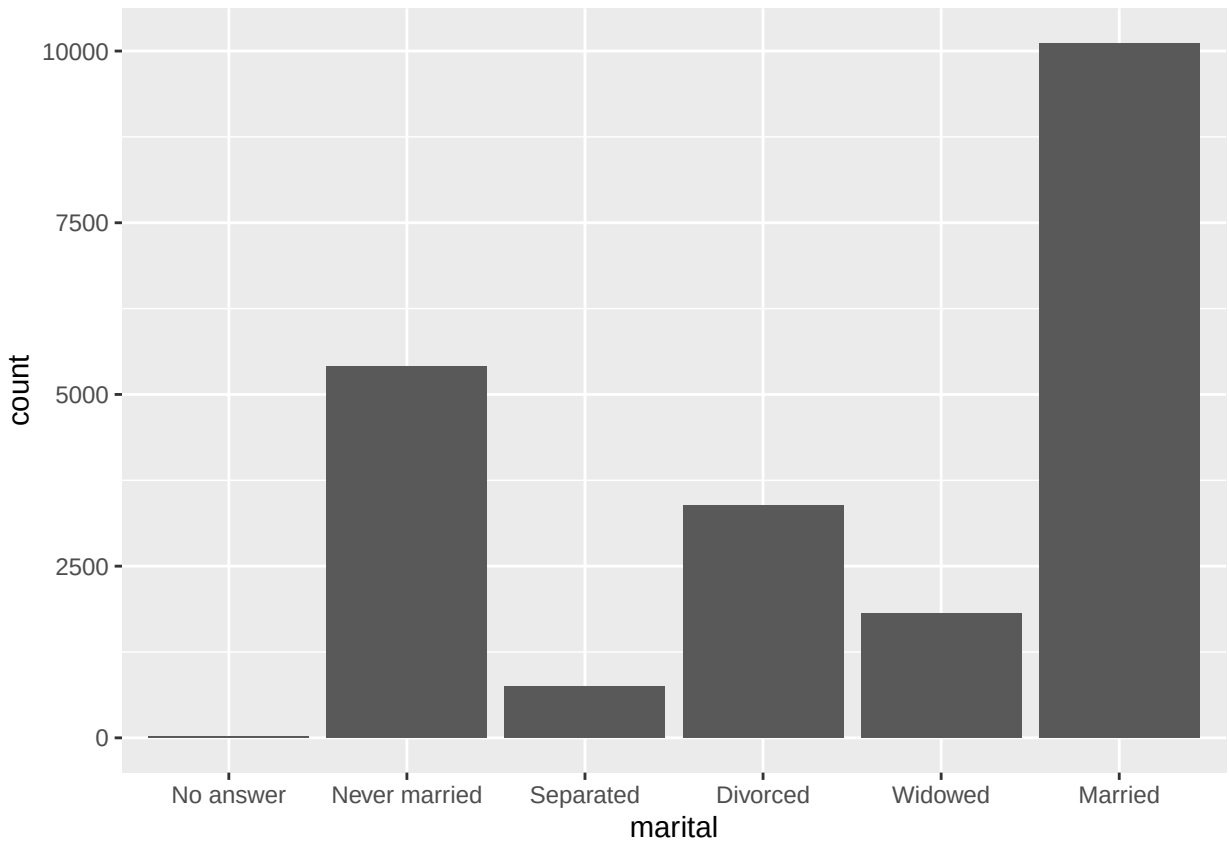


Using *mutate()*

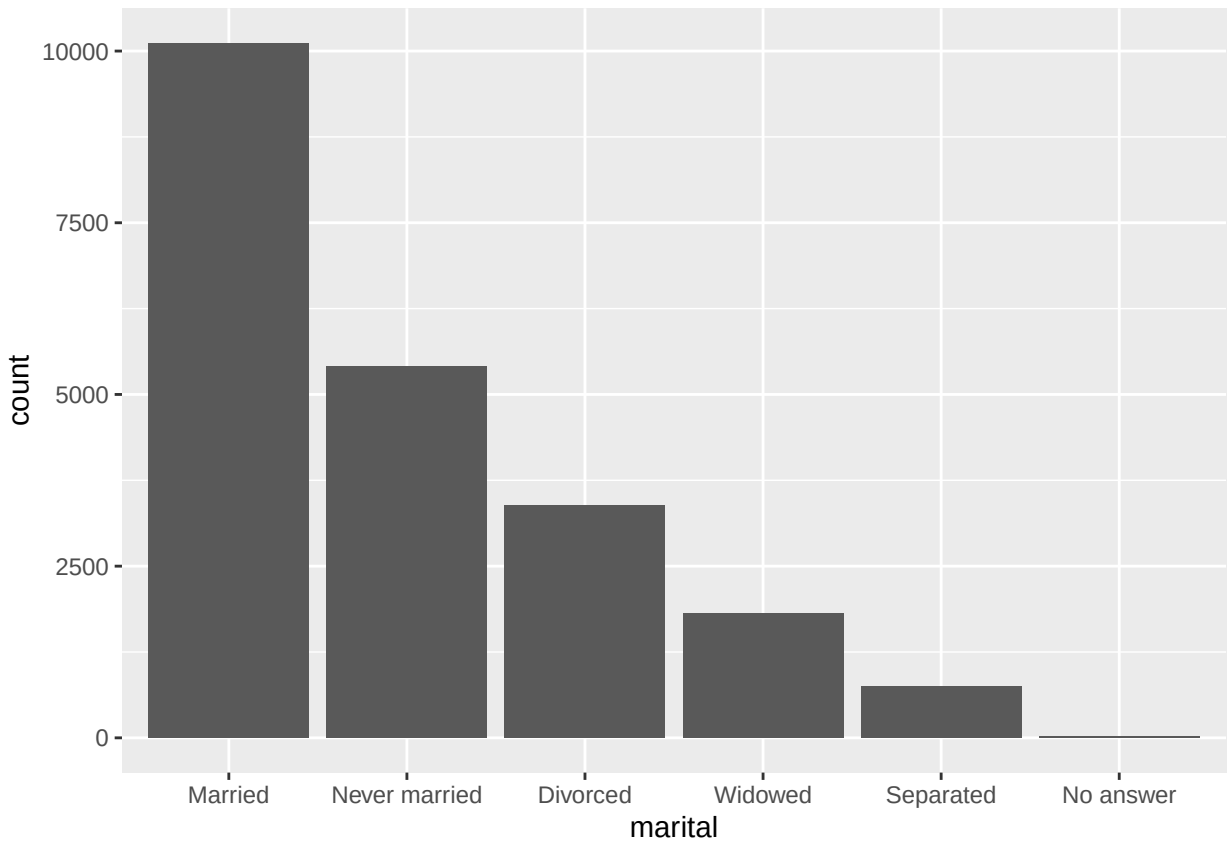
```
gss_cat %>% ggplot(aes(marital)) +
    geom_bar()
```



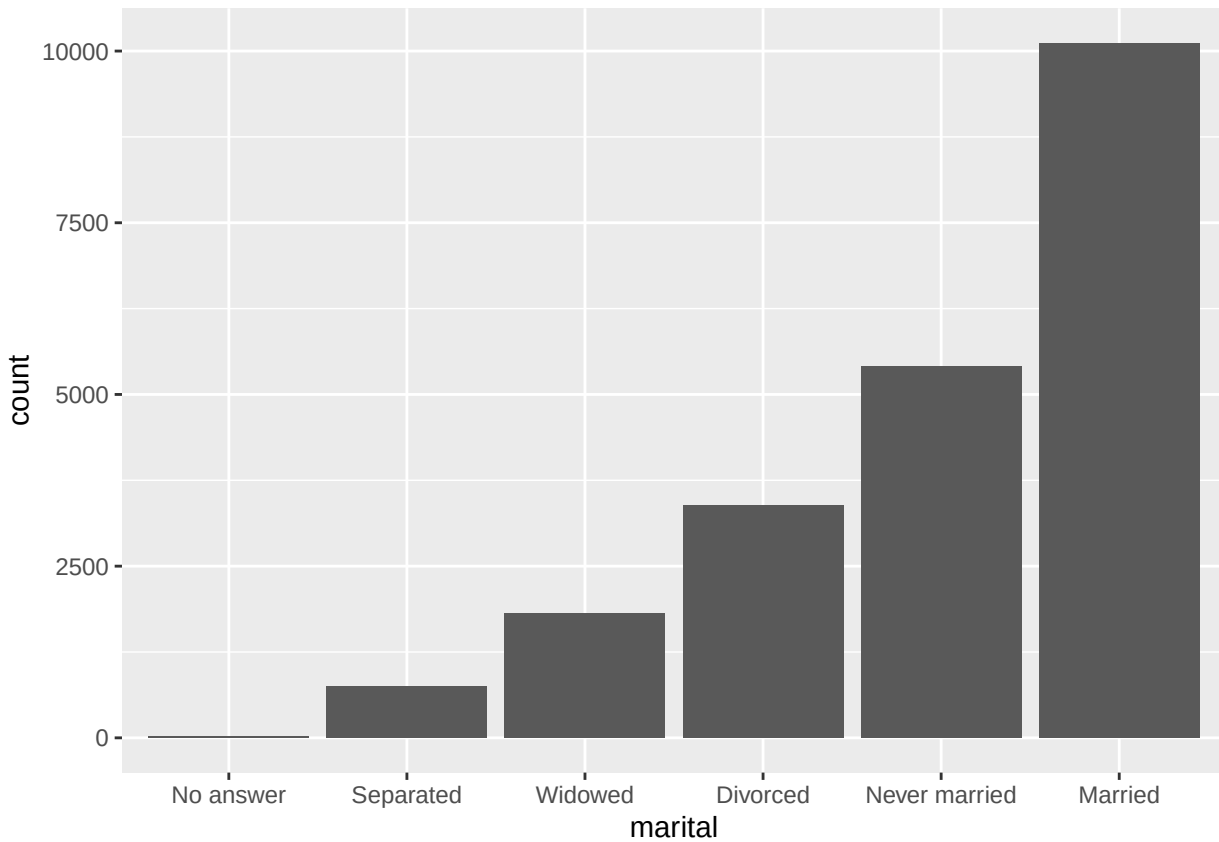

```
gss_cat %>% mutate(marital = marital) %>%
  ggplot(aes(marital)) +
    geom_bar()
```




```
gss_cat %>% mutate(marital = marital %>% fct_infreq()) %>%
  ggplot(aes(marital)) +
    geom_bar()
```




```
gss_cat %>% mutate(marital = marital %>% fct_infreq() %>% fct_rev()) %>%
  ggplot(aes(marital)) +
    geom_bar()
```



Modifying factor levels.

```
gss_cat %>% count(partyid)
```

```
## # A tibble: 10 x 2
##    partyid                n
##    <fct>              <int>
##  1 No answer            154
##  2 Don't know             1
##  3 Other party          393
##  4 Strong republican   2314
##  5 Not str republican  3032
##  6 Ind,near rep        1791
##  7 Independent         4119
##  8 Ind,near dem        2499
##  9 Not str democrat    3690
## 10 Strong democrat     3490
```

Re-coding

```
gss_cat %>%
  mutate(partyid = fct_recode(partyid,
    "Republican, strong"    = "Strong republican",
    "Republican, weak"      = "Not str republican",
    "Independent, near rep" = "Ind,near rep",
    "Independent, near dem" = "Ind,near dem",
    "Democrat, weak"        = "Not str democrat",
    "Democrat, strong"      = "Strong democrat"
  )) %>%
```

```r
  count(partyid)
```

```
## # A tibble: 10 x 2
##    partyid                   n
##    <fct>                 <int>
##  1 No answer               154
##  2 Don't know                1
##  3 Other party             393
##  4 Republican, strong     2314
##  5 Republican, weak       3032
##  6 Independent, near rep  1791
##  7 Independent            4119
##  8 Independent, near dem  2499
##  9 Democrat, weak         3690
## 10 Democrat, strong       3490
```

Other category

```r
gss_cat %>%
  mutate(partyid = fct_recode(partyid,
    "Republican, strong"    = "Strong republican",
    "Republican, weak"      = "Not str republican",
    "Independent, near rep" = "Ind,near rep",
    "Independent, near dem" = "Ind,near dem",
    "Democrat, weak"        = "Not str democrat",
    "Democrat, strong"      = "Strong democrat",
    "Other"                 = "No answer",
    "Other"                 = "Don't know",
    "Other"                 = "Other party"
  )) %>%
  count(partyid)
```

```
## # A tibble: 8 x 2
##   partyid                   n
##   <fct>                 <int>
## 1 Other                   548
## 2 Republican, strong     2314
## 3 Republican, weak       3032
## 4 Independent, near rep  1791
## 5 Independent            4119
## 6 Independent, near dem  2499
## 7 Democrat, weak         3690
## 8 Democrat, strong       3490
```

Collapse a factor

```r
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(partyid)
```

```
## # A tibble: 4 x 2
```

```
##   partyid     n
##   <fct>   <int>
## 1 other     548
## 2 rep      5346
## 3 ind      8409
## 4 dem      7180
```