L10 Factors

Data Science I (STAT 301-1)

YOUR NAME

Contents

Overview

The goal of this lab is to learn and understand how to deal with factors in R, specifically within the tidyverse (although many of these methods can be applied with base R functions as well). Factors are used to work with categorical variables, or variables that have a fixed and known set of possible values. We'll use the `forcats` package, which provides tools for dealing with categorical variables. This package is not part of the core tidyverse, so you'll need to install it.

For more information on the `forcats` package, see `forcats` tidyverse homepage.

Datasets

We will be using the `gss_cat` dataset that is included in the `forcats` package. To view the documentation for the dataset, use `?gss_cat`.

Exercises

Please complete the following exercises. Be sure that your solutions are clearly indicated and the document is neatly formatted.

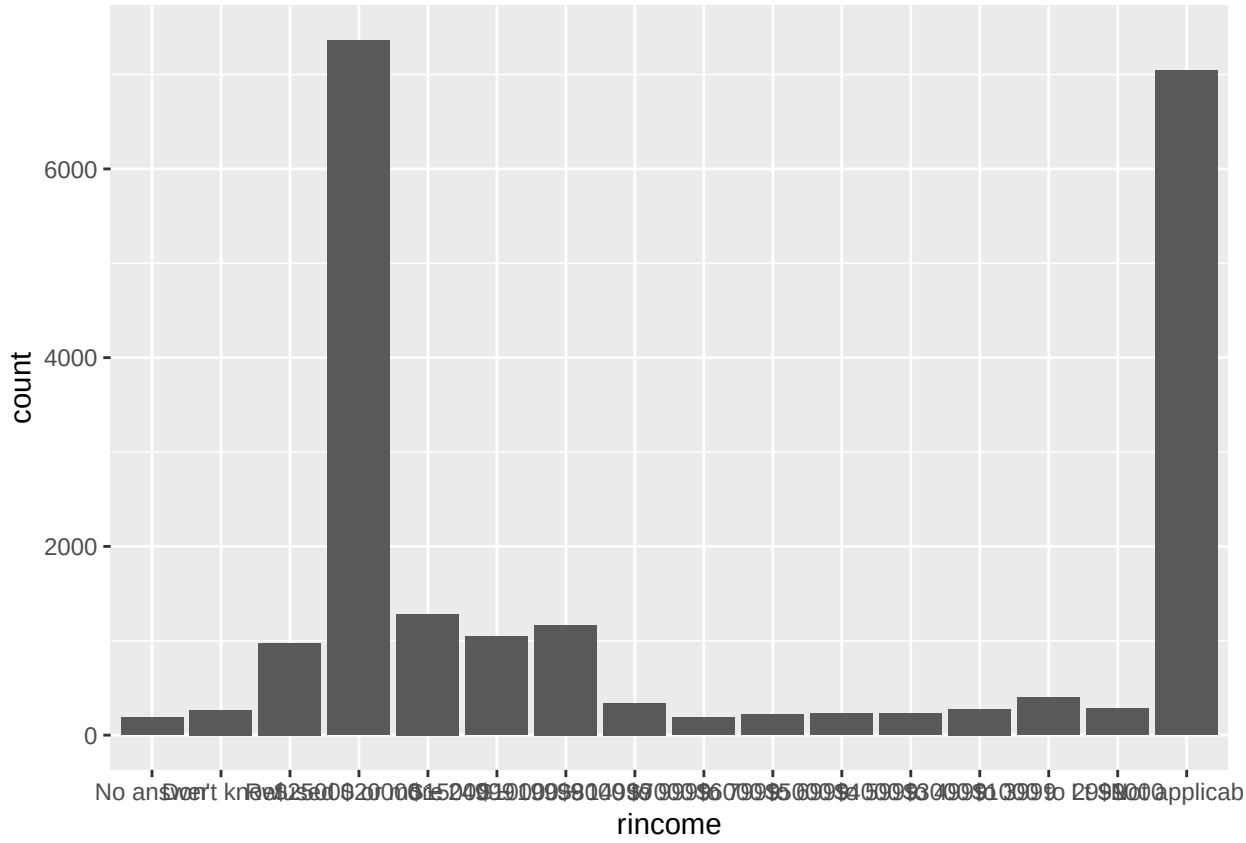
Load Packages You should always begin by loading all necessary packages towards the beginning of your documents. Assume that that all necessary packages have been installed. User should be able to determine if a package needs to be installed either through knowing their R repository or an error message. **Your code should never have install commands.**

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

Exercise 1 (Website: 15.3.1 Ex. 1)

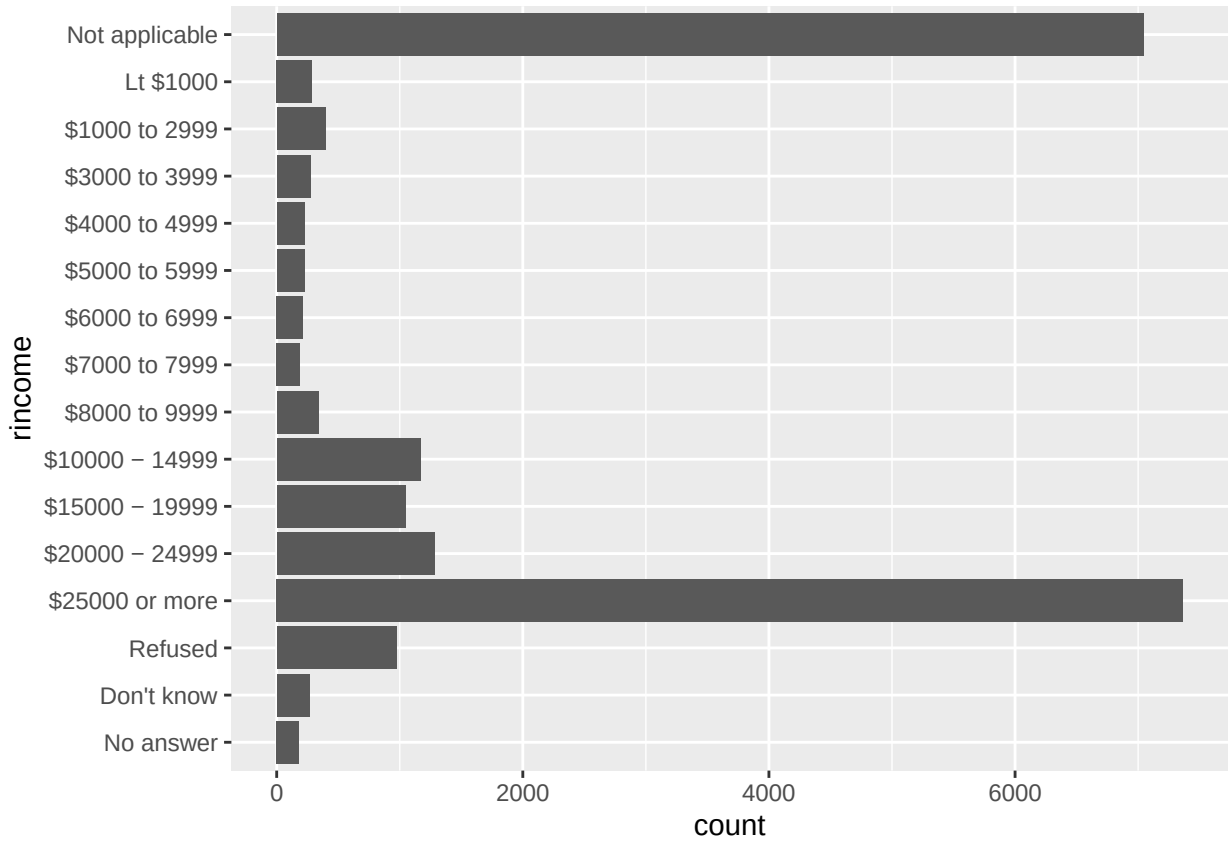
Create a bar chart to explore the distribution of `rincome` (reported income). What makes the default bar chart hard to understand? Improve the bar chart.

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



There are too many factors, and they overlap on the x axis so that you can't read it. To improve it, I'll just flip the axes.

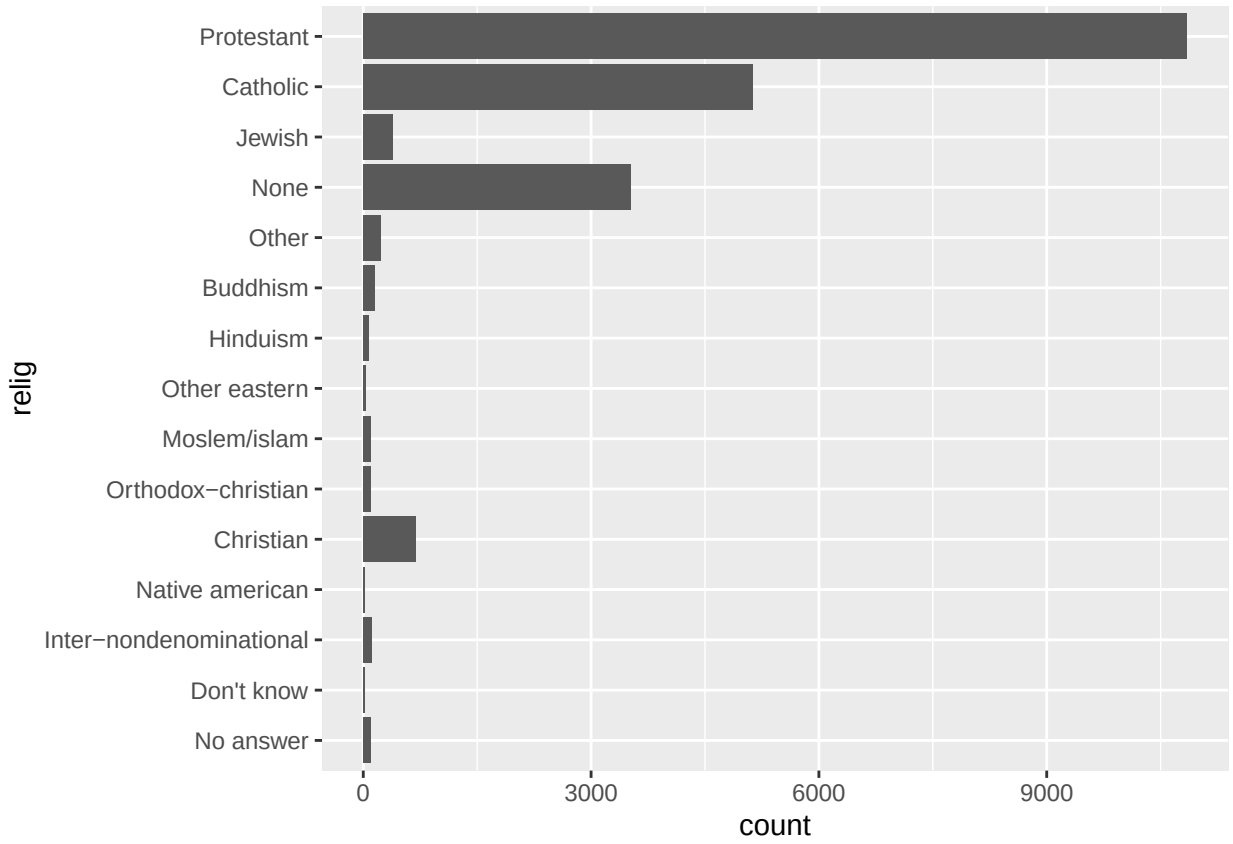
```
gss_cat %>%
  ggplot(mapping = aes(rincome)) +
  geom_bar() +
  coord_flip()
```



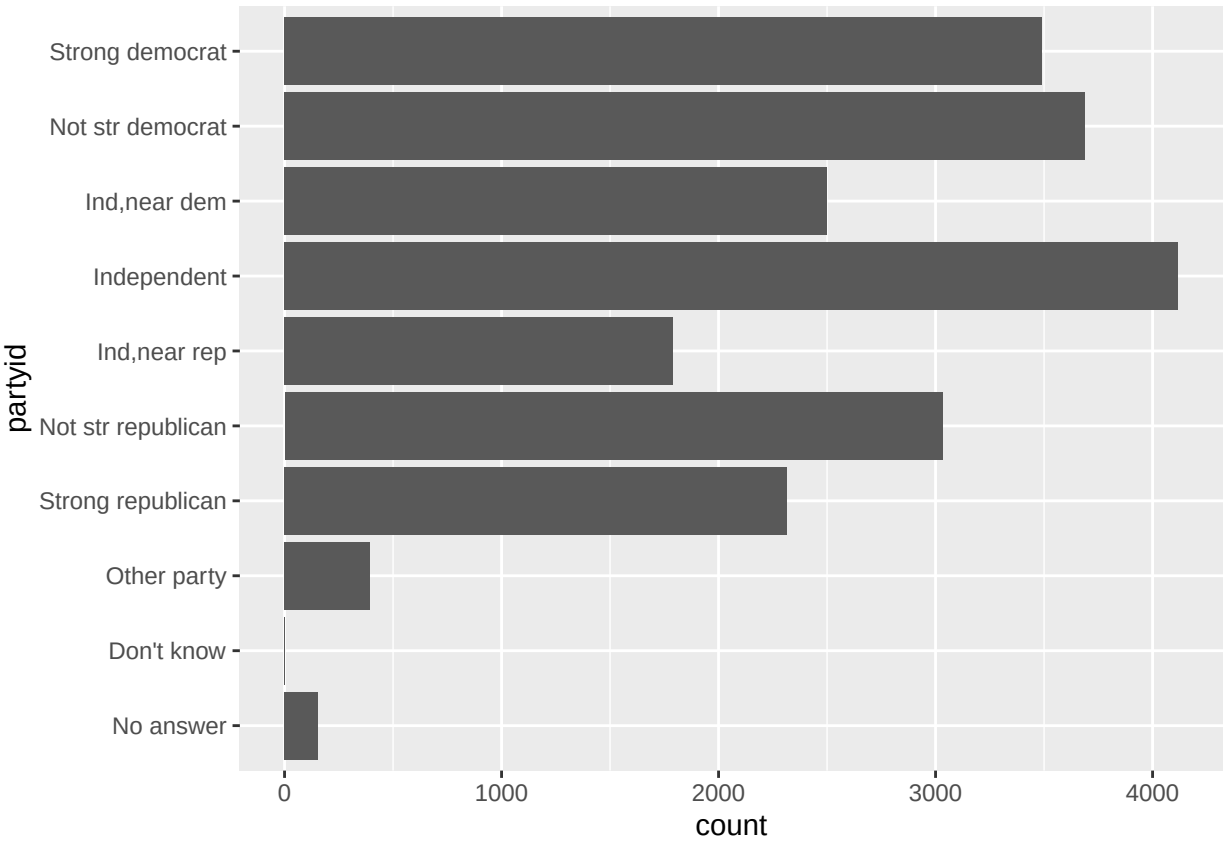
Exercise 2 (Website: 15.3.1 Ex. 2)

What is the most common `relig` in this survey? What's the most common `partyid`?

```
gss_cat %>%
  ggplot(mapping = aes(relig)) +
  geom_bar() +
  coord_flip()
```




```
gss_cat %>%
  ggplot(mapping = aes(partyid)) +
  geom_bar() +
  coord_flip()
```

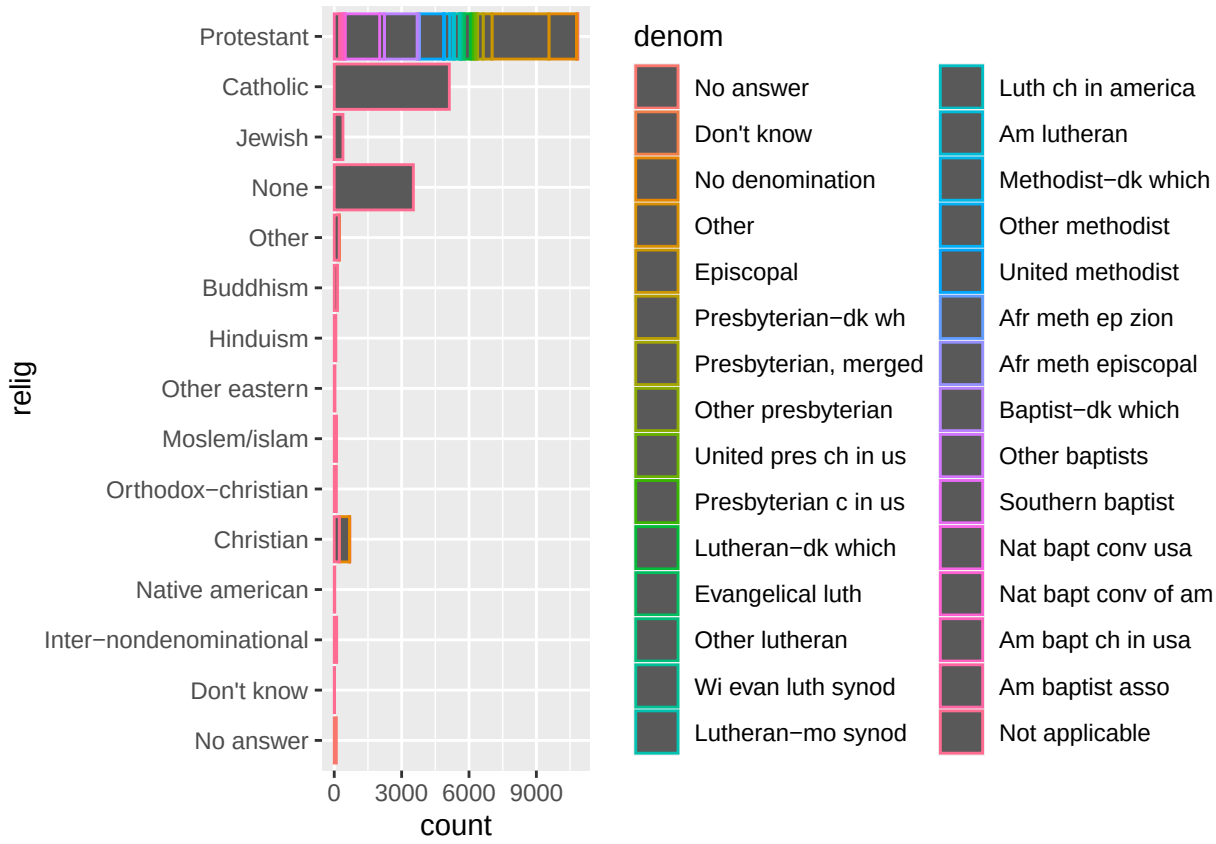


Protestant is the most common religion in the dataset. Independent is the most common party in the data set.

Exercise 3 (Website: 15.3.1 Ex. 3)

Which `relig` does `denom` (denomination) apply to? Create a visualization to find out.

```
gss_cat %>%
  ggplot(mapping = aes(relig)) +
  geom_bar(mapping = aes(color = denom)) +
  coord_flip()
```



I could definitely improve this graphic, but it clearly shows that Protestant is the religion for which `denom` is associated.

Exercise 4 (Website: 15.4.1 Ex. 4)

There are some suspiciously high numbers in `tvhours`. Since the mean is not robust to outliers, it is not a good summary of this variable. Create a graphic similar to the one below, but use a more appropriate summary of `tvhours`.

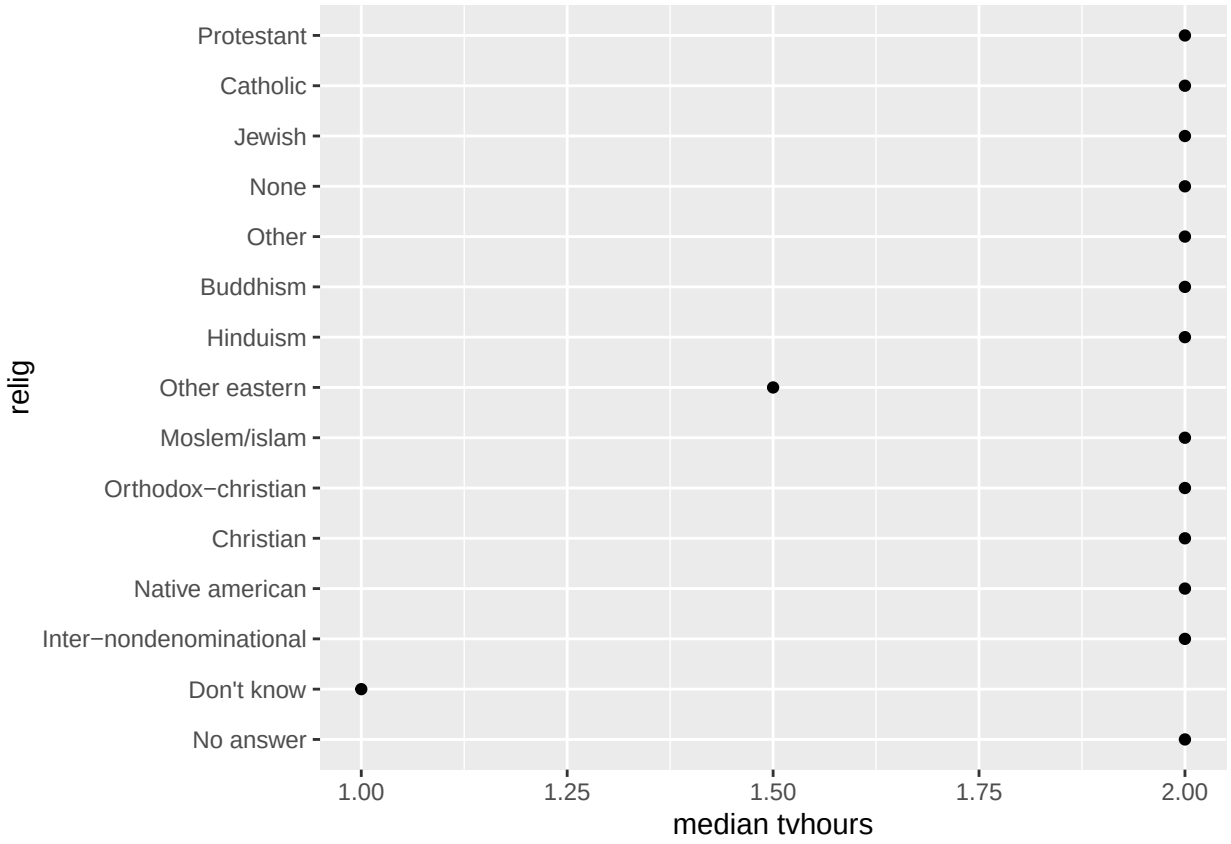
```
relig_summary <- gss_cat %>%
  group_by(relig) %>%
  summarise(
    age = mean(age, na.rm = TRUE),
    tvhours = median(tvhours,na.rm = TRUE),
    n = n()
  )

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

```
gss_cat %>%
  group_by(relig) %>%
  summarize(med = median(tvhours, na.rm = TRUE)) %>%
  ungroup() %>%
  ggplot() +
```

```
  geom_point(mapping = aes(x = relig, y = med)) +
  coord_flip() +
  ylab('median tvhours')
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



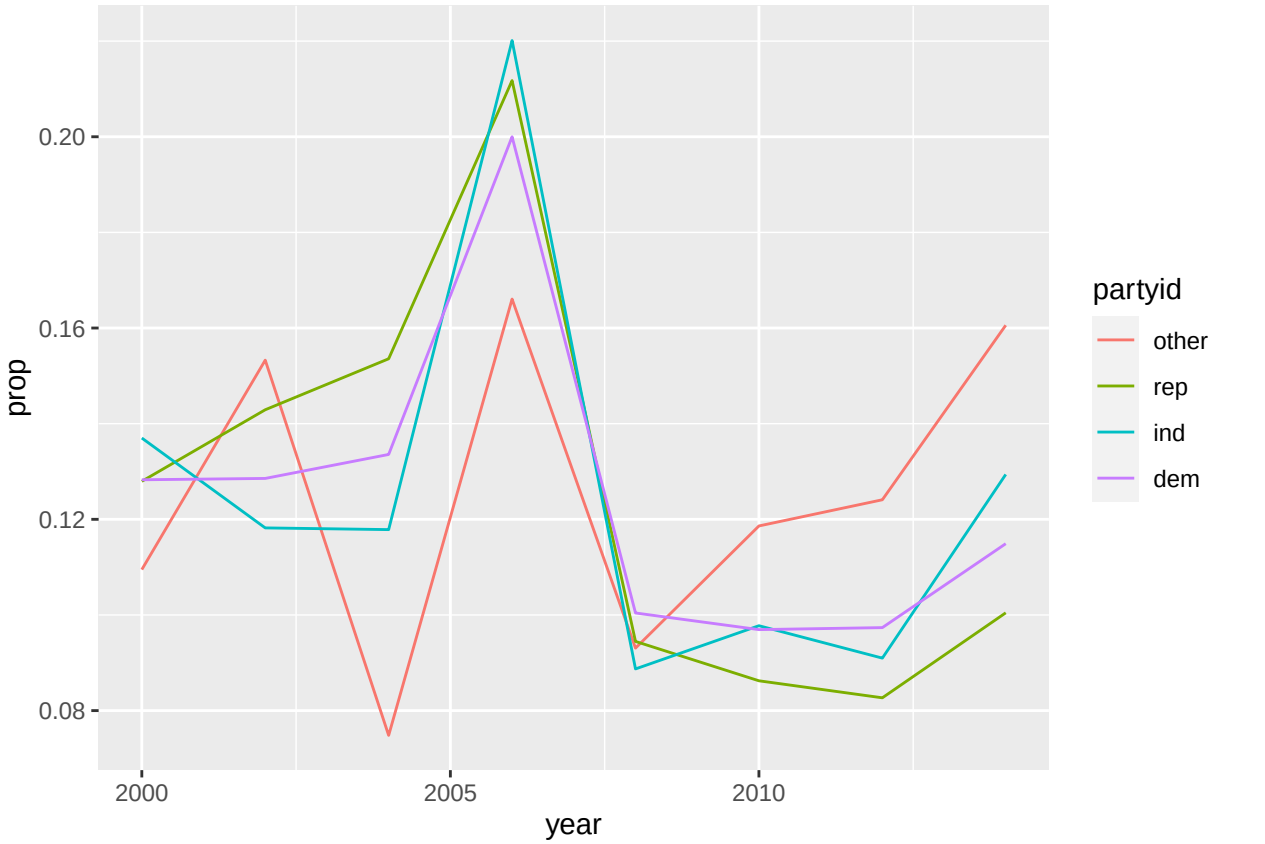
By taking the median of tvhours instead of the mean, we see that every religion except for 'Other eastern' and 'Dont know' has a median `tvhours` of 2.

Exercise 5 (Website: 15.5.1 Ex. 1)

How have the proportions of people who identify as Democrat, Republican, and Independent changed over time (from 2000 to 2014)? Build a visualization to display this information. You should also write a few sentences to summarize the trends within the graphic.

```
gss_cat %>%
  mutate(partyid = fct_collapse(partyid,
                   dem = c("Not str democrat", "Strong democrat"),
                   rep = c("Strong republican", "Not str republican"),
                   ind = c("Ind,near rep", "Independent", "Ind,near dem"),
                   other = c("No answer", "Don't know", "Other party")
      )) %>%
  count(partyid, year) %>%
  group_by(partyid) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup() %>%
  ggplot(mapping = aes(x = year, y = prop)) +
```

```
  geom_line(aes(color = partyid))
```



From this graphic, we see that all party id's shoot up around 2007, presumably for the 2008 presidential election. After 2008 they all drop significantly, with 'other' growing the fastest post 2010 and the others showing modest growth in proportion.

Exercise 6 (Website: 15.5.1 Ex. 2)

Demonstrate how to collapse `rincome` into a smaller set of categories.

```
gss_cat %>%
  mutate(rincome = fct_collapse(rincome,
                                `<$5000` = c('Lt $1000', '$1000 to 2999',
                                             '$3000 to 3999', '$4000 to 4999'),
                                `$5000 to $9999` = c('$5000 to 5999',
                                                     '$6000 to 6999',
                                                     '$7000 to 7999',
                                                     '$8000 to 9999'),
                                `$10000 to $19999` = c('$10000 - 14999',
                                                       '$15000 - 19999'),
                                `>$20000` = c('$20000 - 24999',
                                              '$25000 or more'),
                                other = c('No answer', "Don't know", 'Refused',
                                          'Not applicable')))
```

```
## # 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 $5000 to ~ Ind,near r~ Protesta~ Souther~      12
##  2  2000 Divorced     48 White $5000 to ~ Not str re~ Protesta~ Baptist~      NA
##  3  2000 Widowed      67 White other      Independent Protesta~ No deno~       2
##  4  2000 Never ma~    39 White other      Ind,near r~ Orthodox~ Not app~       4
##  5  2000 Divorced     25 White other      Not str de~ None      Not app~       1
##  6  2000 Married      25 White >$20000    Strong dem~ Protesta~ Souther~      NA
##  7  2000 Never ma~    36 White >$20000    Not str re~ Christian Not app~       3
##  8  2000 Divorced     44 White $5000 to ~ Ind,near d~ Protesta~ Luthera~      NA
##  9  2000 Married      44 White >$20000    Not str de~ Protesta~ Other          0
## 10  2000 Married      47 White >$20000    Strong rep~ Protesta~ Souther~       3
## # ... with 21,473 more rows
```