Categorial Variables and Factors Assignment

Tyler Kephart

2024-08-18

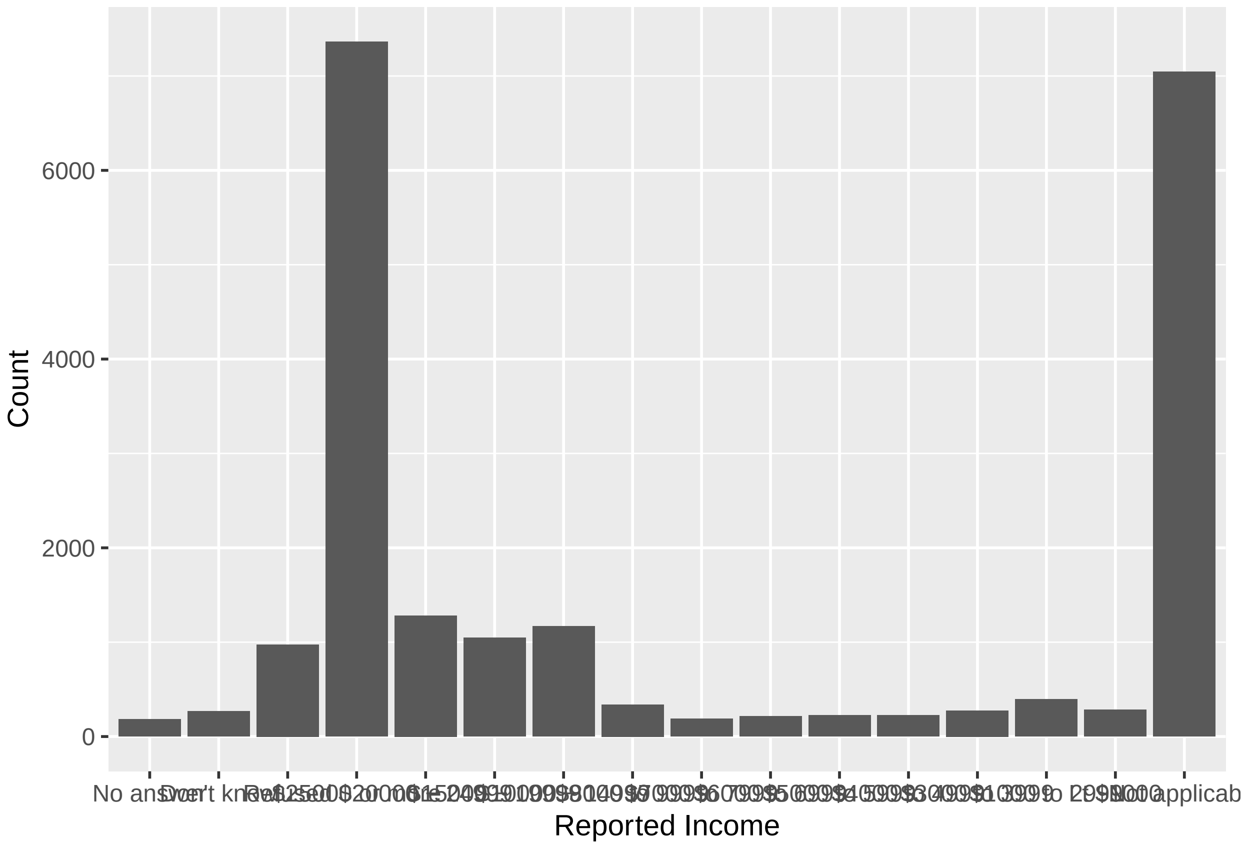
Loaded libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

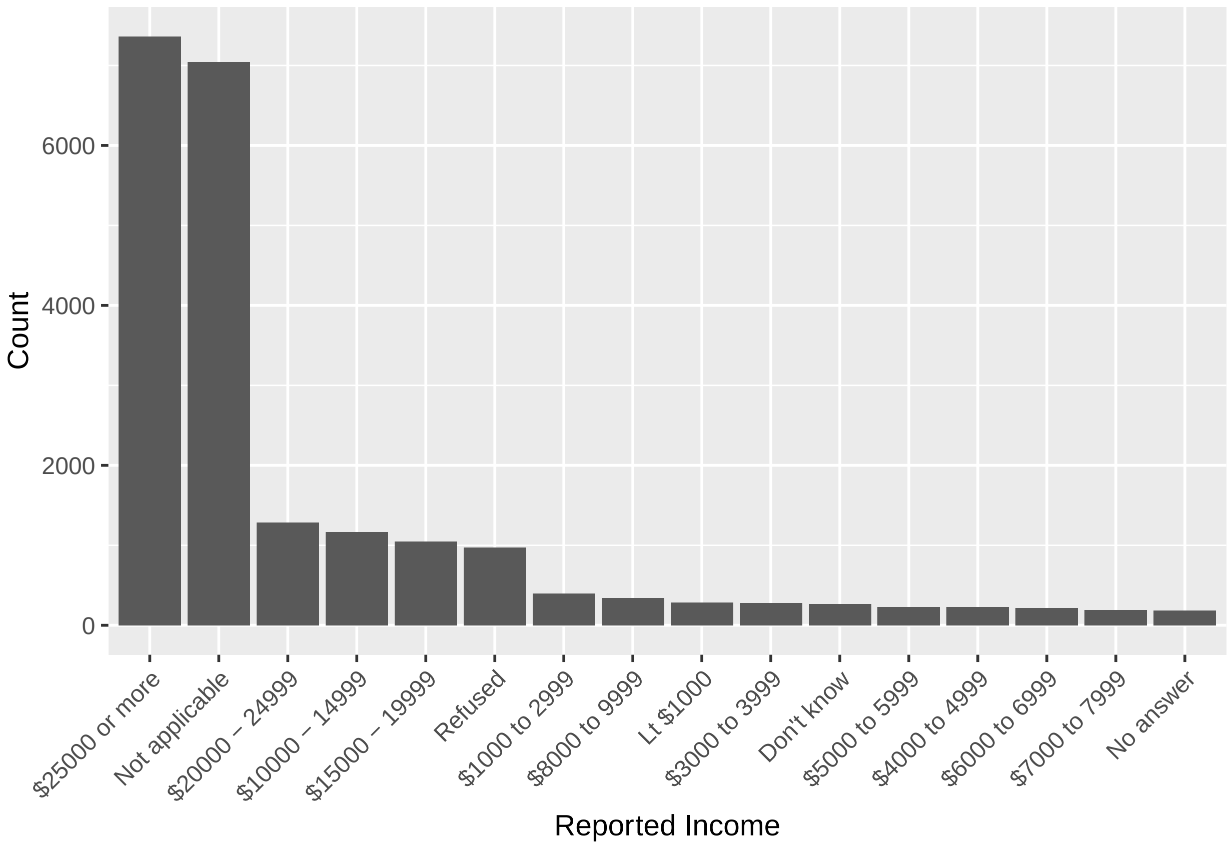
1. From the "forcats" library load gss_cat data. Explore the distribution of rincome (reported income). What makes the default bar chart hard to understand? How could you improve the plot?

```
# default bar chart of rincome
gss_cat %>%
    ggplot(aes(x = rincome)) +
    geom_bar() +
    labs(x = "Reported Income",
         y = "Count")
```



```
# The default bar chart can be hard to understand due to:
# - overlapping labels
# - order of categories

# improved plot
gss_cat %>%
    # reorder categories by frequency
    ggplot(aes(x = fct_infreq(rincome))) +
    geom_bar() +
    labs(x = "Reported Income",
         y = "Count") +
    # angle x variable names for readability
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



2. What is the most common religion? What is the most common partyid?

```r
# most common religion
gss_cat %>%
    count(relig) %>%
    arrange(desc(n)) %>%
    slice(1)
```

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

```r
# most common partyid
gss_cat %>%
    count(partyid) %>%
    arrange(desc(n)) %>%
    slice(1)
```

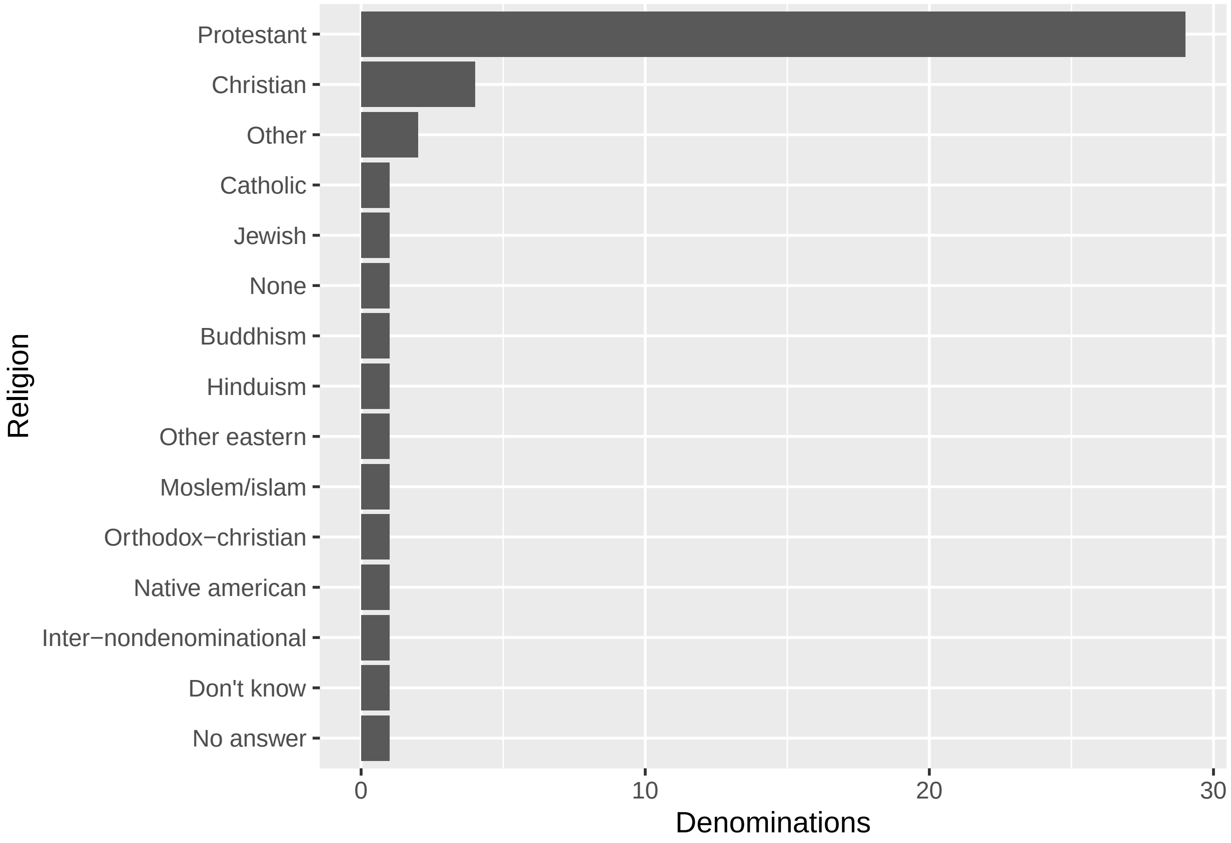
```
## # A tibble: 1 x 2
##   partyid         n
##   <fct>       <int>
## 1 Independent  4119
```

3. Which relig does denom (denomination) apply to? How can you find out with a table? How can you find out with a visualisation?

```
# summary table of denominations per religion
gss_cat %>%
    group_by(relig) %>%
    # counts distinct denominations for each religion
    summarise(count_denom = n_distinct(denom)) %>%
    arrange(desc(count_denom))
```

```
## # A tibble: 15 x 2
##    relig                   count_denom
##    <fct>                         <int>
##  1 Protestant                       29
##  2 Christian                         4
##  3 Other                             2
##  4 No answer                         1
##  5 Don’t know                        1
##  6 Inter-nondenominational           1
##  7 Native american                   1
##  8 Orthodox-christian                1
##  9 Moslem/islam                      1
## 10 Other eastern                     1
## 11 Hinduism                          1
## 12 Buddhism                          1
## 13 None                              1
## 14 Jewish                            1
## 15 Catholic                          1
```

```
# plot of denominations per religion
gss_cat %>%
    group_by(relig) %>%
    # counts distinct denominations for each religion
    summarise(count_denom = n_distinct(denom)) %>%
    ggplot(aes(
            # orders religions by number of denominations
            x = reorder(relig, count_denom),
            y = count_denom)) +
        # bar chart with height of bars being the number of denominations
        geom_col() +
        # flip the axes to make the plot easier to read
        coord_flip() +
        labs(x = "Religion",
            y = "Denominations")
```

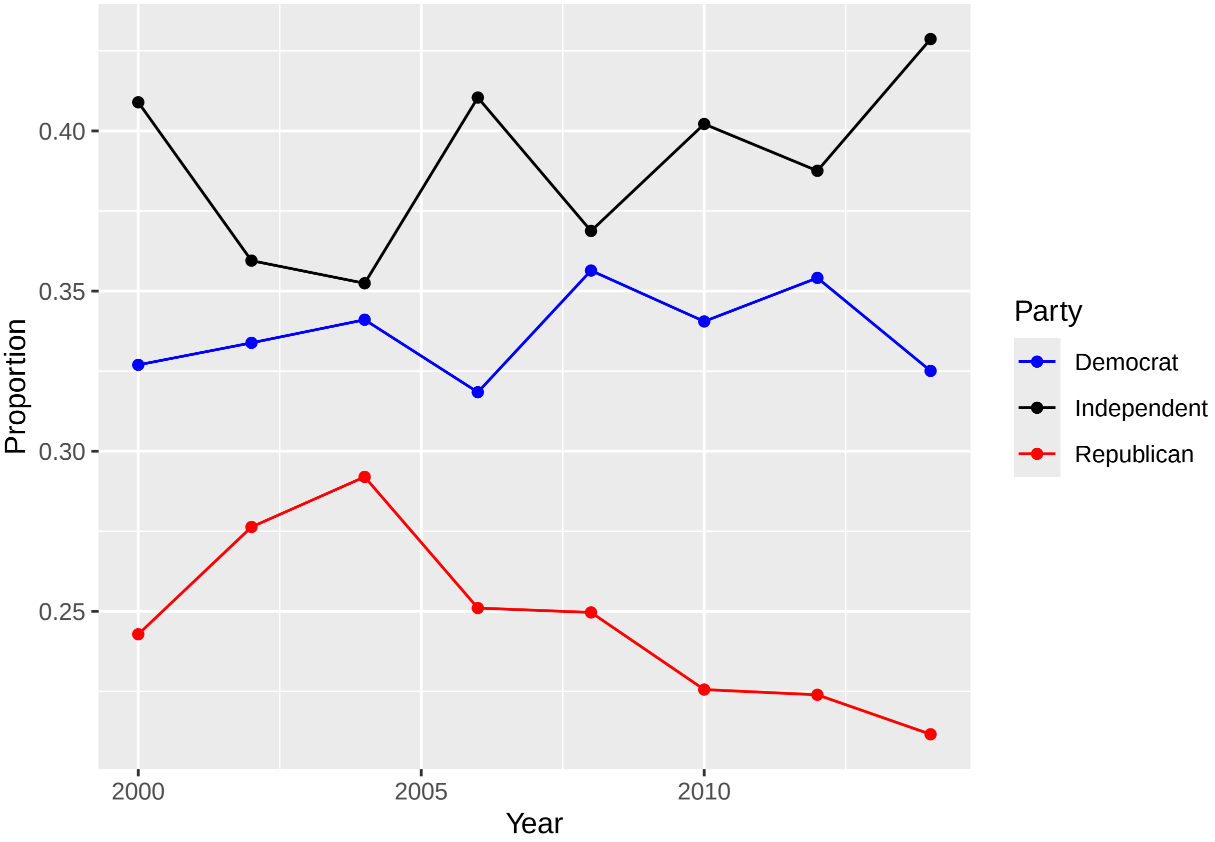


4. How have the proportions of people identifying as Democrat, Republican, and Independent changed over time? Plot a suitable chart.

```
# categorize partyid into Independent, Democrat, Republican, and Other
gss_cat <- gss_cat %>%
    mutate(parties = case_when(
        partyid %in% c(
            "Ind,near rep","Independent","Ind,near dem") ~ "Independent",
        partyid %in% c(
            "Not str democrat","Strong democrat") ~ "Democrat",
        partyid %in% c(
            "Not str republican","Strong republican") ~ "Republican",
        # catch-all for any unspecified categories
        TRUE ~ "Other"),
    .after = partyid)
# summarize the proportions of parties by year
party_proportions <- gss_cat %>%
    group_by(year, parties) %>%
    summarise(count = n()) %>%
    group_by(year) %>%
    mutate(total = sum(count),
            proportion = count / total) %>%
    ungroup()
```

```
## ‘summarise()‘ has grouped output by ’year’. You can override using the
## ‘.groups‘ argument.
```

```r
# plot the proportions of parties over time
party_proportions %>%
    # only looking at democrat, republican, and independent, filter out others
    filter(parties != "Other") %>%
    ggplot(aes(x = year, y = proportion, color = parties)) +
        geom_line() +
        geom_point() +
        labs(x = "Year", y = "Proportion", color = "Party") +
        # make colors to match party colors
        scale_color_manual(values = c(
                            "Democrat" = "blue",
                            "Republican" = "red",
                            "Independent" = "black"))
```



5. Collapse "rincome" into smaller set of categories?

```r
# define income ranges and collapse rincome into broader categories
gss_cat %>%
  mutate(income_cat = case_when(
    rincome %in% c("Lt $1000") ~ "Low Income",
    rincome %in% c("$1000 to 2999") ~ "Lower Middle Income",
    rincome %in% c("$3000 to 3999", "$4000 to 4999") ~ "Middle Income",
    rincome %in% c("$5000 to 5999", "$6000 to 6999") ~ "Upper Middle Income",
    rincome %in% c("$7000 to 7999", "$8000 to 9999") ~ "High Income",
    rincome %in% c("$10000 - 14999", "$15000 - 19999") ~ "Very High Income",
    rincome %in% c("$20000 - 24999", "$25000 or more") ~ "Top Income",
    TRUE ~ "Unknown"),
    .after = rincome)
```

```
## # A tibble: 21,483 x 11
##     year marital        age race  rincome income_cat partyid parties relig denom
##    <int> <fct>        <int> <fct> <fct>   <chr>      <fct>   <chr>   <fct> <fct>
##  1  2000 Never marri~    26 White $8000 ~ High Inco~ Ind,ne~ Indepe~ Prot~ Sout~
##  2  2000 Divorced        48 White $8000 ~ High Inco~ Not st~ Republ~ Prot~ Bapt~
##  3  2000 Widowed         67 White Not ap~ Unknown    Indepe~ Indepe~ Prot~ No d~
##  4  2000 Never marri~    39 White Not ap~ Unknown    Ind,ne~ Indepe~ Orth~ Not ~
##  5  2000 Divorced        25 White Not ap~ Unknown    Not st~ Democr~ None  Not ~
##  6  2000 Married         25 White $20000~ Top Income Strong~ Democr~ Prot~ Sout~
##  7  2000 Never marri~    36 White $25000~ Top Income Not st~ Republ~ Chri~ Not ~
##  8  2000 Divorced        44 White $7000 ~ High Inco~ Ind,ne~ Indepe~ Prot~ Luth~
##  9  2000 Married         44 White $25000~ Top Income Not st~ Democr~ Prot~ Other
## 10  2000 Married         47 White $25000~ Top Income Strong~ Republ~ Prot~ Sout~
## # i 21,473 more rows
## # i 1 more variable: tvhours <int>
```