Homework 9

Carmen Canedo

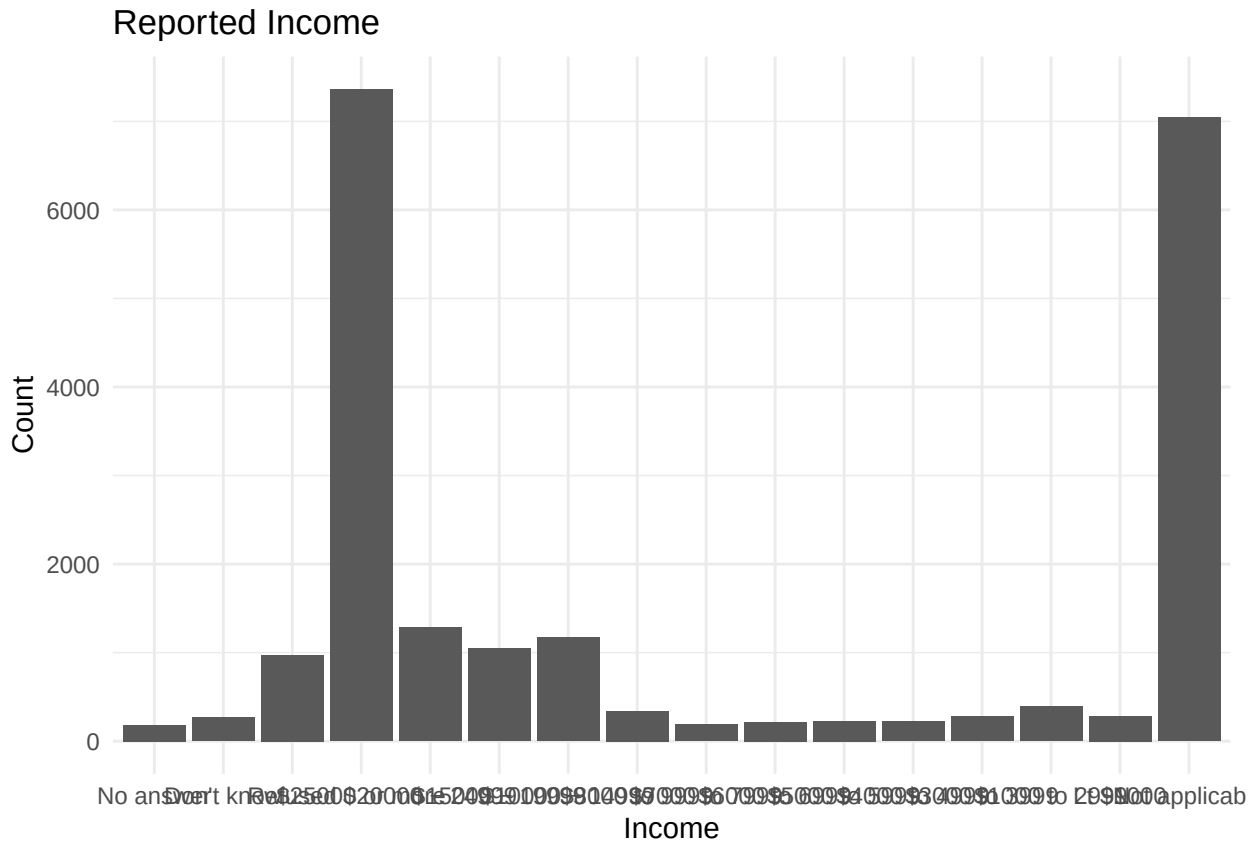
28 November 2020

```
# Loading libraries
source("setup.R")
library(lubridate)
```

Exercise 1

Question 1: Explore the distribution of `rincome` (reported income). What makes the default bar chart hard to understand? How could you improve the plot?

Here is a default bar graph of `rincome`.



This is hard to understand because:

1. The labels on the x-axis are illegible because they are overlapping.
2. Includes "No answer" values

To improve it, I would use `coord_flip()` so that the labels are legible, remove non-response values, and sort count in ascending order.

Question 2: What is the most common religion? Most common `partyid`?

The most common religion is Protestantism.

```
## # A tibble: 3 x 2
##   relig          n
##   <fct>      <int>
## 1 Protestant 10846
## 2 Catholic    5124
## 3 None        3523
```

The most common political party is Independent. However, the levels for each party are split, so I wonder if they are combined if it would change which is the most common political party.

```
## # A tibble: 5 x 2
##   partyid               n
##   <fct>             <int>
## 1 Independent         4119
## 2 Not str democrat    3690
## 3 Strong democrat     3490
## 4 Not str republican  3032
## 5 Ind,near dem        2499
```

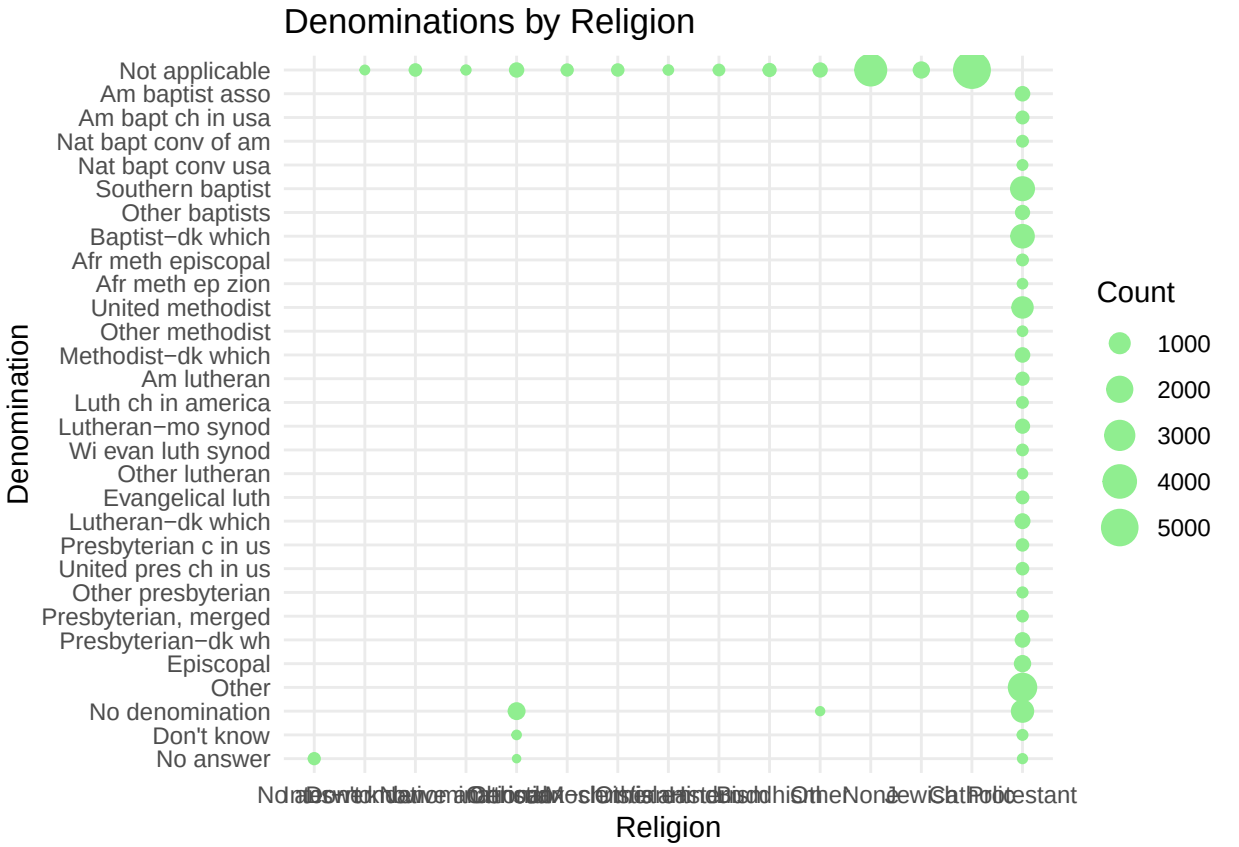
Question 3: Which `relig` does `denom` apply to? How can you find out with a table? How can you find out with a vizualization?

To find out what `relig denom` belongs to, we can check the levels to see if they give us any further details.

```
##  [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"
```

The levels are all Protestant denominations, so we can assume that `denom` belongs to `Protestant`.

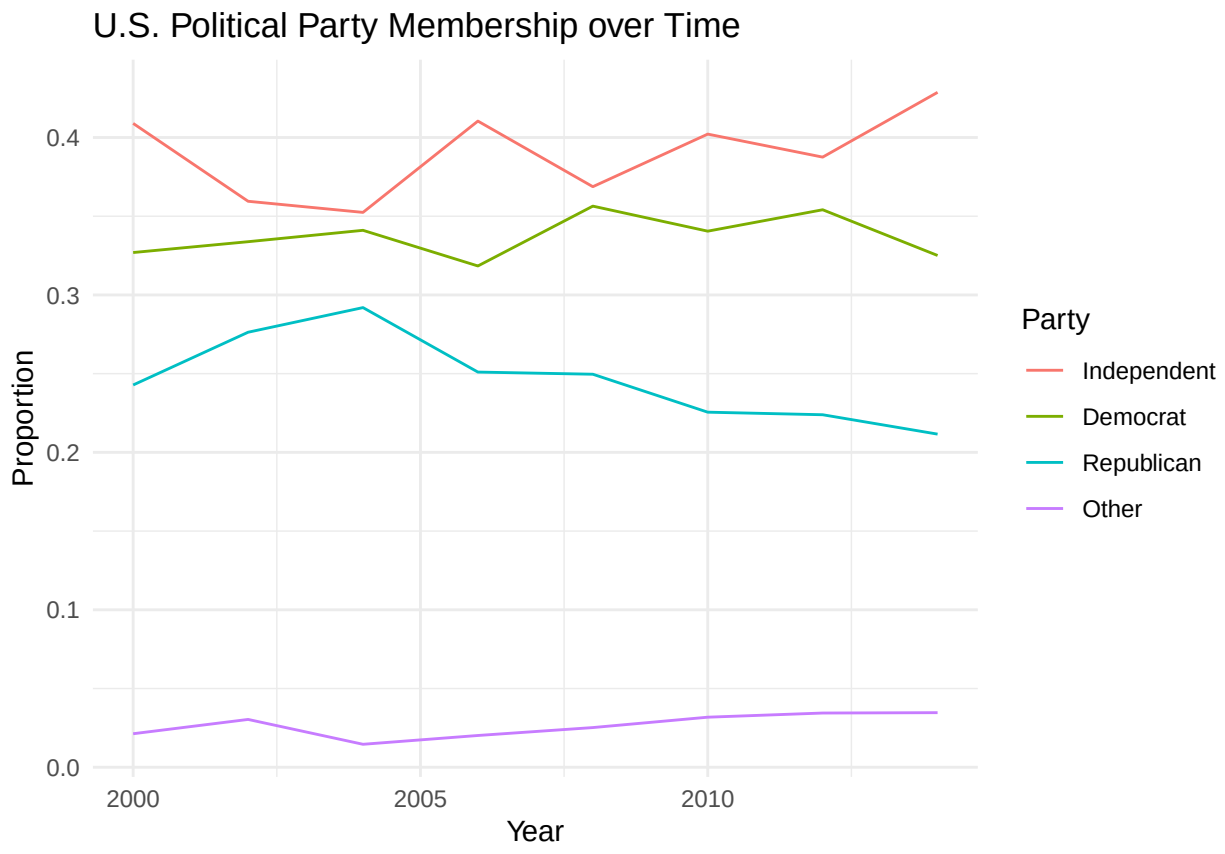
To find out with a visualization, a scatterplot can show the count of each denomination by religion, so we will use that below.



Question 4: How have the proportions of people identifying as Democrat, Republican, and Independent changed over time?

The levels for the parties are split, so if we are going to look at their overall change, we need to combine them using `fct_collapse()`, and then we need to calculate the proportion of them by year.

Then we can create a line graph that plots the proportion of party membership overtime.



It appears that there has been an increase in individuals identifying as Independent, a decrease of people identifying as Republican, and the proportion of Democrats has remained relatively the same.

Question 5: How could you collapse `rincome` into a small set of categories?

The levels for `rincome` are not broken down into even groups, so to decrease the number of categories, I am going to try to aim for each category to increase by about $5000.

The first thing to do would be to collapse the non-responses into a group called `Other`. Then we will need to use `str_c()` from the `stringr` package to dilineate the new levels.

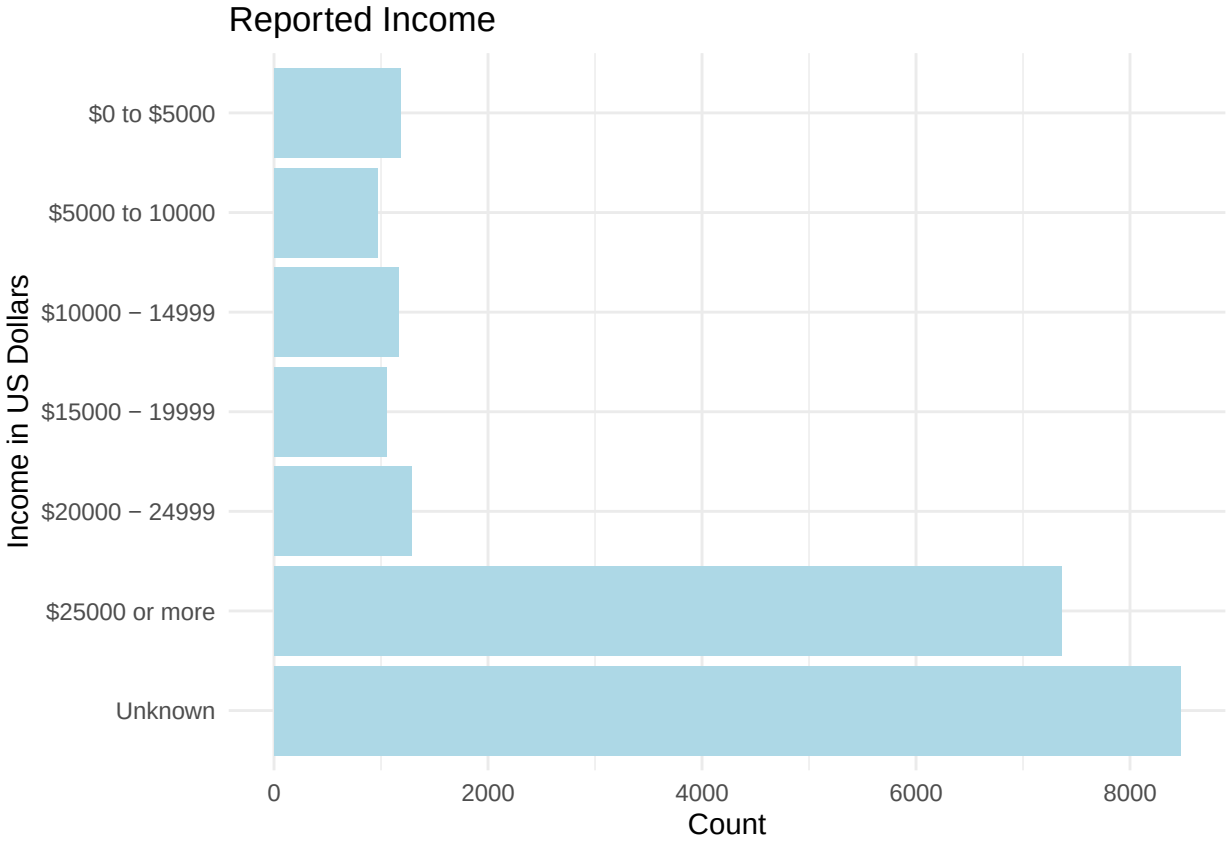
I will be using the following levels:

- $0 to $5000
- $5000 to $10000

```
## # A tibble: 10 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 Unknown    Independent Protesta~ No deno~       2
##  4  2000 Never ma~    39 White Unknown    Ind,near r~ Orthodox~ Not app~       4
##  5  2000 Divorced     25 White Unknown    Not str de~ None      Not app~       1
##  6  2000 Married      25 White $20000 - ~ Strong dem~ Protesta~ Souther~      NA
```

```
##  7  2000 Never ma~    36 White $25000 or~ 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 $25000 or~ Not str de~ Protesta~ Other          0
## 10  2000 Married      47 White $25000 or~ Strong rep~ Protesta~ Souther~       3
```

Using the recommendations for what I would change in Question 1, this is what our graph looks like now.



Exercise 2

Question 1: Parse the following date-times

"05/26/2004 UTC 11:11:11.444"
"26 2004 05 UTC 11/11/11.444"

```
## [1] "2004-05-26 11:11:11 UTC"
```

```
## [1] "2004-05-26 11:11:11 UTC"
```

Question 2: Use the appropriate `lubridate` function to parse each of the following dates.

```
## [1] "2010-01-01"
```

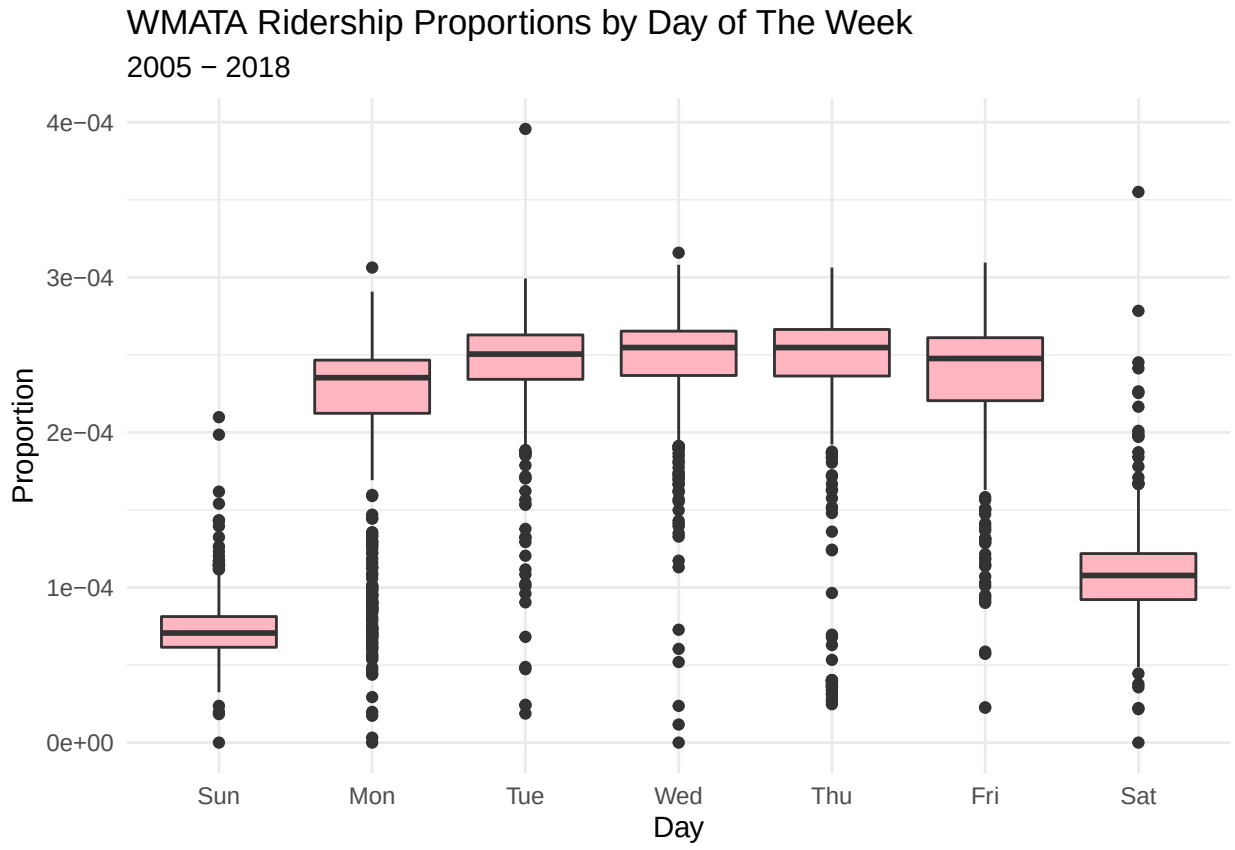
```
## [1] "2015-03-07"
```

```
## [1] "2017-06-06"
```

```
## [1] "2015-08-19" "2015-07-01"
```

```
## [1] "2014-12-30"
```

Question 3: For each month in `wmata_ridership`, calculate the proportion of rides made on a given day of the month. Then make box plots of the proportions of ridership vs. day of the week, *excluding any days from 2004*.



Question 4: How long of a time-span is covered in the WMATA ridership dataset?

```
## [1] "472435200s (~14.97 years)"
```