Homework 5

Cory Costello

Due: Nov 1, 2017

Part 1: read in data and count observations and variablles

The first thing we need to do is read in the data. I'll then pipe the data to the `clean_names` function.

```
gss_2014 <- import("GSS2014merged_R6.sav") %>%
  clean_names()
```

Next, we'll get the number of rows and columns; we can use `dplyr::count()` for the number of observations, and `ncol()` from base R to get the number of column. Note that `nrow()` from base R could be used instead of `dplyr::count()`).

```
count(gss_2014)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  3842
```

```
ncol(gss_2014)
```

```
## [1] 970
```

It looks like there are 3842 observations in the data and 970 variables. So this is a pretty big dataset.

Part 2: Subset variables with select

Now we'll go ahead and subset the data so we have just the columns we want to work with. The columns we want are: year, id, age, educ, sex, race, res16, income, partyid, polviews, vote08, vote12, pres08, and pres12.

```
gss_2014_subset <- gss_2014 %>%
  select(id, year, age, educ, sex, race, res16, income, partyid, polviews, vote08, vote12, pres08, pres
```

Now I'm going to turn everything but id, year, age, and education into factors using `rio::factorize`, which turns labelled variables (like the ones in this dataset) into factors. I would have done it in the pipeline above (after select), but that was leading to age and education being missing (it looks like they have labels that only apply to certain values).

I'm going to do this with the base R subseting syntax `[]`. I couldn't figure out an efficient tidyverse solution (that wouldn't require finding all of the labels, and setting them manually).

```
gss_2014_subset[,-(1:4)]<- factorize(gss_2014_subset[,-(1:4)])
```

Part 3: Filter out participants who didn't vote in '08 and/or '12

Now we'll filter for people that voted in 08 and 12. I'll do this by filtering for people that have "Voted" for their value in vote08 and vote12 (note, this will still have some people missing on presidential vote).

```
gss_2014_subset_vote08_12 <- gss_2014_subset %>%
  filter(vote08 == "Voted" & vote12 == "Voted")
```

We started with a sample of $N = 3842$ participants. After eliminating everyone that didn't vote in 2008 and/or 2012 ($N = 2330$), we ended up with our final sample of $N = 1512$.

Part 4: Recode presidental vote variable

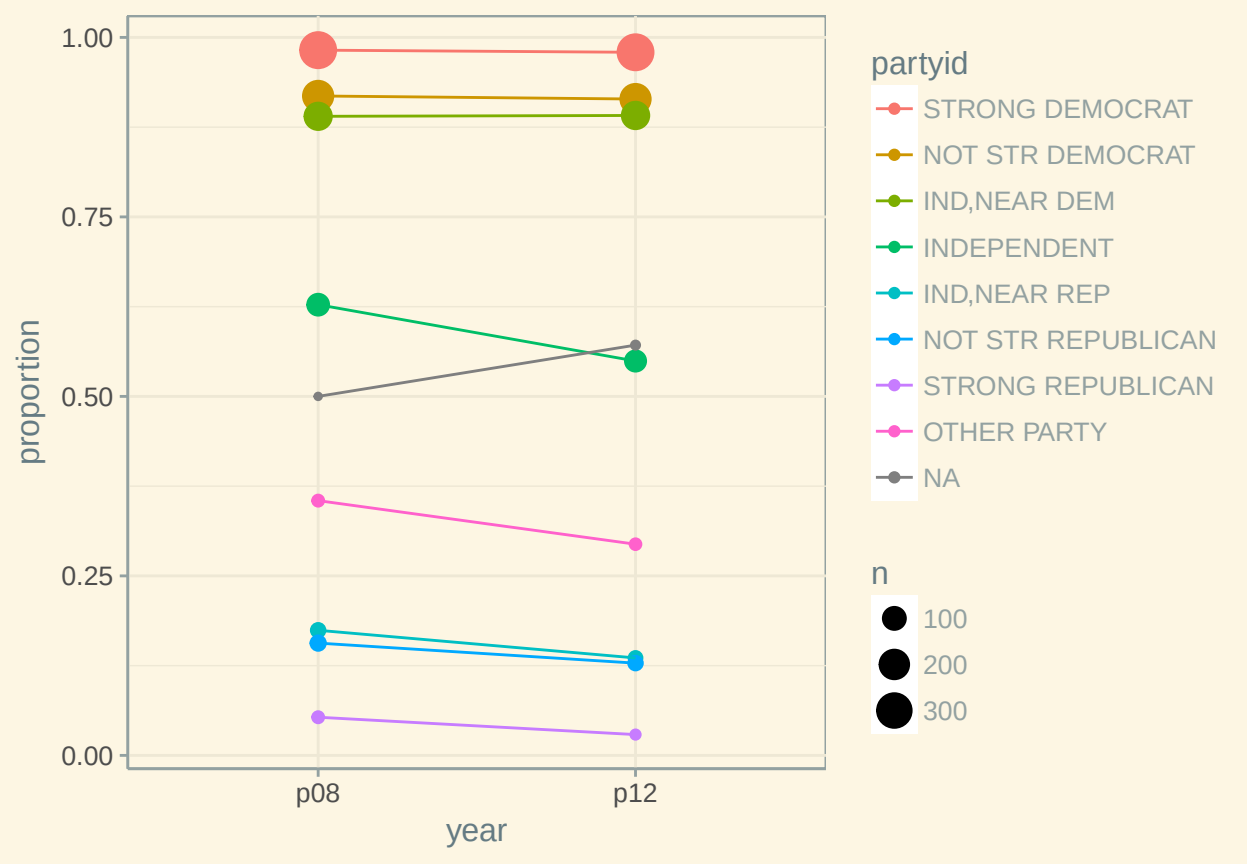
Now we'll recode the pres08 and pres12 variable to be equal to 1 when they voted for Obama, and equal to 0 if they voted for anyone else. This essentially changes from tracking who they voted for to tracking whether or not they voted for Obama (the eventual winner).

To do this, I'm going to use mutate, reuse the variable labels (so that the old ones are overwritten), and then use an ifelse statement that assigns a value of 1 to cases where their value for presidental vote is Obama (e.g., when pres08 is equal to Obamma; same for pres12) and assigns a 0 to all other cases. Side note, I **love** the `ifelse()` function and am happy to find out it can be used in mutate! (ps. I needed a place to use bold, which is why this is here).

```
gss_2014_subset_vote08_12 <- gss_2014_subset_vote08_12 %>%
  mutate(pres08 = ifelse(pres08 == "Obama", 1, 0),
         pres12 = ifelse(pres12 == "Obama", 1, 0))
```

Part 5: Proportion of votes for Obama per party

```
gss_2014_subset_vote08_12 %>%
  select(id, partyid, pres08, pres12) %>%
  gather(year, vote, -id, -partyid) %>%
  mutate(year = recode(year, "pres08" = "p08"),
         year = recode(year, "pres12" = "p12")) %>%
  group_by(partyid, year) %>%
  summarize(n = sum(vote, na.rm = TRUE),
            proportion = mean(vote, na.rm = TRUE)) %>%
  ggplot(aes(x = year, y = proportion, color = partyid, group = partyid)) +
  geom_point(aes(size = n)) +
  geom_line()
```



It looks like the proportion of people who identify as democrats (to any degree) voting for Barack Obama didn't change from 2008 to 2012; all three groups were very close to the ceiling (nearly 100% of those people voted for Obama in '08 and '12). It looks like the proportion of people who identify as Republican (to any degree) who voted for Obama consistently declined, as did the proportion of Independents voting for Obama. It looks like the people who's party identification was entered as na were the only group that had a greater proportion of Obama votes in '12 than '08.

As an unrelated note, using the size of the points is a really cool idea; made me think of a meta-analysis by (Briley and Tucker-Drob 2014), which has 3 really beautiful graphs at figure 3. I'll end with a quote from that paper (Briley and Tucker-Drob 2014, 1327):

> Individual differences in patterns of thoughts, feelings, and behavior tend to stabilize over development. Along with increases in phenotypic stability, genetic and environmental influences both increase in stability with age. Near age 30, genetic stability ap- proaches unity, and true environmental stability slowly increases across the majority of the life span to reach similar levels of stability in old age. The

References

Briley, D.a., and Elliot M Tucker-Drob. 2014. "Genetic and environmental continuity in personality development: A meta-analysis." *Psychological Bulletin* 140 (5): 1303–31. doi:10.1037/a0037091.