A simple R Notebook for exploring the ANES cumulative time series data

This is an R Markdown Notebook.

1. Introduction

The *American National Election Studies* (ANES) are surveys of voters in the U.S. on the national scale. For each presidential election since 1948, ANES collects responses from respondents both before and after the election. The goal of ANES is to understand political behaviors using systematic surveys. ANES's data and results have been routinely used by news outlets, election campaigns and political researchers.

The *Time Series Cumulative Data* of ANES include answers, from respondents from different years, on selected questions that have been asked in three or more ANES' *Time Series* studies. Tremendous amount of efforts have been put into data consolidation as variables are often named differently in different years.

A rule of thumb for analyzing any data set is to understand its study design and data collection process first. You are strongly encouraged to read the *codebooks*.

2. Access to ANES Data

Step 2.1: Register to access ANES dataset.

To access the data, you should register at ANES's website and accept its terms of use, especially committing to "use these datasets solely for research or statistical purposes and not for investigation of specific survey respondents."

Step 2.2: Download the ANES Time Series Cumulative Data

Once you are logged into ANES's website, you should be able to download the data file. You can use ASCII, DTA or SAV. In this notebook, we use the *SAV* format. The downloaded file is a zip file, you should move all unzipped files into the `data` folder of your project 1's local folder.

3. Data processing for this R Notebook.

The following code blocks prepare a processed data set and save it in the `output` folder. The data folder should be only used for storing raw `data`. All processed data should be saved in the `output` folder. The notion here is that one can delete files from the output folder and reproduce them by re-running the codes.

Step 3.1 Checking `R` packages for data processing

From the packages' descriptions:

- `tidyverse` is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures;
- `haven` enables R to read and write various data formats used by other statistical packages. `haven` is part of the `tidyverse`.
- `devtools` provides a collection of package development tools.

- `RColorBrewer` provides ready-to-use color palettes.
- `DT` provides an R interface to the JavaScript library DataTables;
- `ggplot2` a collection of functions for creating graphics, based on The Grammar of Graphics.

Step 3.2 Import raw ANES data

We will be working with the SAV format of the raw ANES data, downloaded from this page, once you are registered *and* logged in. This is a saved data file from SPSS. We will use the `read_sav` function from the `haven` package.

Read more about importing SPSS data into R.

```r
library(haven)
anes_dat <-
    read_sav("../data/anes_timeseries_cdf.sav")
```

Some basic data summaries: there are 59944 respondents and 1029 variables.

```r
barplot(table(anes_dat$VCF0004),
        las=2,
        main="number of respondents over the years")
```

number of respondents over the years
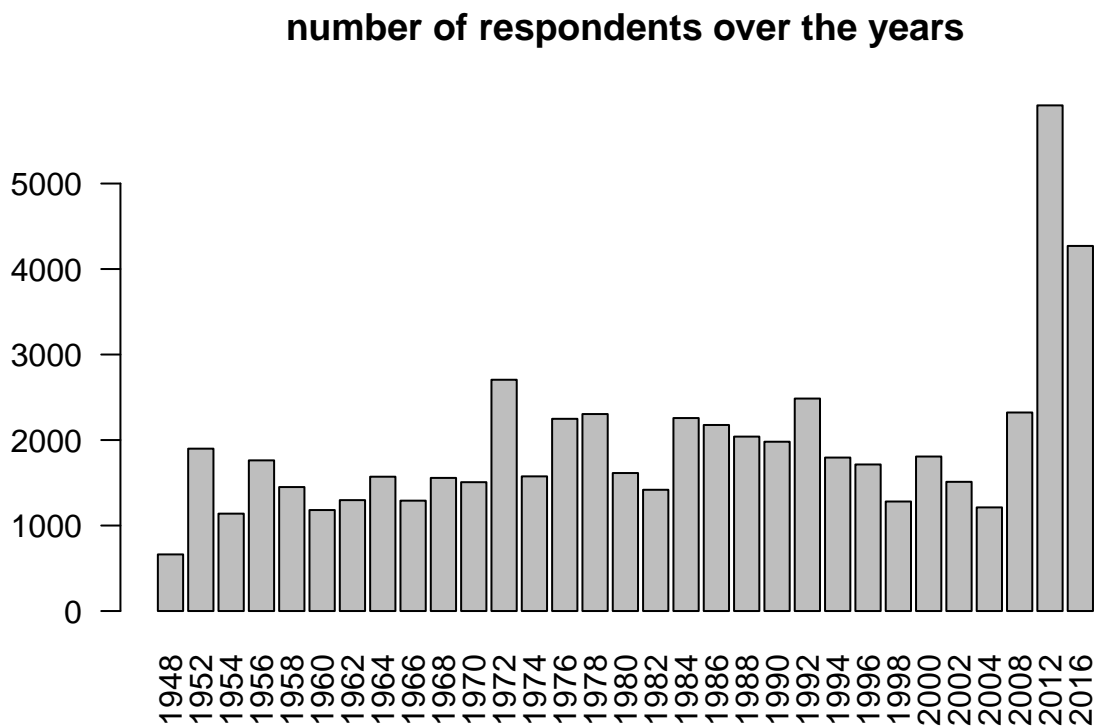
Some variables are asked nearly all the years and some are asked only a few years.

```r
anes_NAs=anes_dat%>%
  summarise_all(list(na.mean=function(x){
                                 mean(is.na(x))}))
anes_NAs=data.frame(nas=unlist(t(anes_NAs)))
```

```
ggplot(anes_NAs, aes(x=nas)) +
  geom_histogram(color="black",
                 fill="white",
                 binwidth=0.05)+
  labs(title="Fractions of missing values")
```


Fractions of missing values
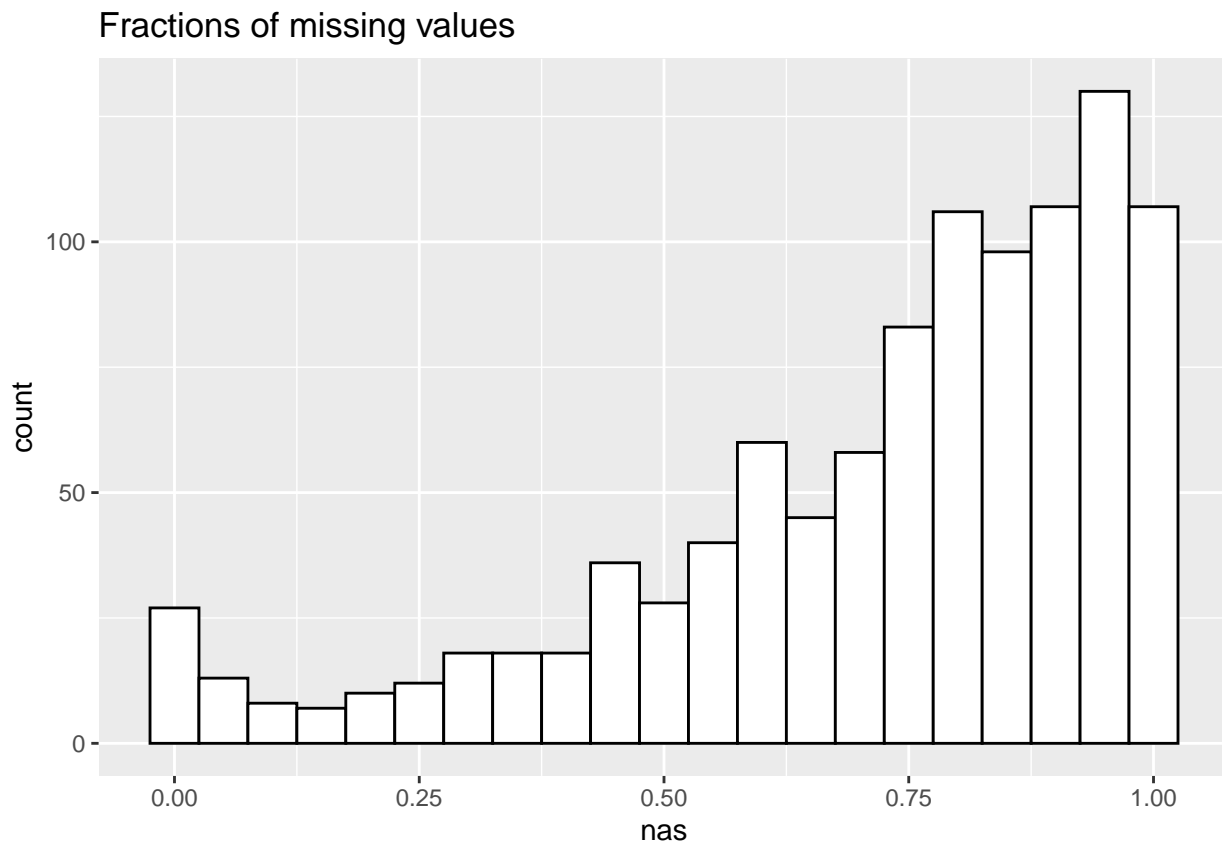
Step 3.3 Process variables for analysis

In the following, we will create a few variables for this tutorial. These variables were selected based on their description in the ANES codebook. You are encouraged to look them up and read about how they were prepared.

First let's look at our data. One advantage of using the SPSS SAV data is that the values are *labelled*. By converting labelled data to factor, you can easily reveal the responses encoded by numbers. We selected four variables for subsequent analysis and save the filtered data sets to the `output` folder.

```
Election_years=as.character(seq(1952, 2016, 4))

anes_use=anes_dat%>%
  mutate(
    year=as_factor(VCF0004),
    turnout=as_factor(VCF0703),
    vote=as_factor(VCF0706),
    race=as_factor(VCF0105a),
    gender=as_factor(VCF0104)
  )%>%
  filter(year %in% Election_years)
```

```r
library(data.table)

data.table(anes_use%>%
           select(year, turnout, vote, race, gender)%>%
           filter(!is.na(turnout))%>%
           sample_n(30))
```

```
##     year                          turnout
## 1: 1996 1. Not registered, and did not vote
## 2: 2016                3. Voted (registered)
## 3: 1976                3. Voted (registered)
## 4: 1960                3. Voted (registered)
## 5: 1952                3. Voted (registered)
## 6: 1984                3. Voted (registered)
## 7: 1952                3. Voted (registered)
## 8: 1972                3. Voted (registered)
## 9: 2012 1. Not registered, and did not vote
## 10: 1952                3. Voted (registered)
## 11: 2000                3. Voted (registered)
## 12: 1984                3. Voted (registered)
## 13: 1988                3. Voted (registered)
## 14: 2016                3. Voted (registered)
## 15: 1992                3. Voted (registered)
## 16: 1972     2. Registered, but did not vote
## 17: 2016                3. Voted (registered)
## 18: 2008                3. Voted (registered)
## 19: 2012                3. Voted (registered)
## 20: 1984                3. Voted (registered)
## 21: 2012                3. Voted (registered)
## 22: 1980 1. Not registered, and did not vote
## 23: 2016                3. Voted (registered)
## 24: 1980                3. Voted (registered)
## 25: 1996     2. Registered, but did not vote
## 26: 2016                3. Voted (registered)
## 27: 1976 1. Not registered, and did not vote
## 28: 1996                3. Voted (registered)
## 29: 2008                3. Voted (registered)
## 30: 2012                3. Voted (registered)
##     year                          turnout
##                                                             vote
## 1: 7. Did not vote or voted but not for president (exc.1972)
## 2:                                              2. Republican
## 3:                                              1. Democrat
## 4:                                              1. Democrat
## 5:                                              1. Democrat
## 6:                                              1. Democrat
## 7:                                              2. Republican
## 8:                                              2. Republican
## 9: 7. Did not vote or voted but not for president (exc.1972)
## 10:                                             1. Democrat
## 11:                                                     <NA>
## 12: 7. Did not vote or voted but not for president (exc.1972)
## 13:                                             2. Republican
```

```
## 14:                                                2. Republican
## 15:      3. Major third party candidate (Wallace 1968/Anderson
## 16: 7. Did not vote or voted but not for president (exc.1972)
## 17:                                                  1. Democrat
## 18:                                                2. Republican
## 19:                                                2. Republican
## 20:                                                  1. Democrat
## 21:                                                2. Republican
## 22: 7. Did not vote or voted but not for president (exc.1972)
## 23:                                                2. Republican
## 24:                                                2. Republican
## 25: 7. Did not vote or voted but not for president (exc.1972)
## 26:                                                2. Republican
## 27: 7. Did not vote or voted but not for president (exc.1972)
## 28:                                                2. Republican
## 29:                                                2. Republican
## 30:                                                  1. Democrat
##                                                            vote
##                                       race   gender
##  1: 1. White non-Hispanic (1948-2012) 2. Female
##  2: 1. White non-Hispanic (1948-2012) 2. Female
##  3: 1. White non-Hispanic (1948-2012) 2. Female
##  4: 1. White non-Hispanic (1948-2012)   1. Male
##  5: 1. White non-Hispanic (1948-2012) 2. Female
##  6: 1. White non-Hispanic (1948-2012) 2. Female
##  7: 1. White non-Hispanic (1948-2012)   1. Male
##  8: 1. White non-Hispanic (1948-2012) 2. Female
##  9: 1. White non-Hispanic (1948-2012) 2. Female
## 10: 2. Black non-Hispanic (1948-2012)   1. Male
## 11: 1. White non-Hispanic (1948-2012)   1. Male
## 12: 1. White non-Hispanic (1948-2012)   1. Male
## 13: 1. White non-Hispanic (1948-2012)   1. Male
## 14: 1. White non-Hispanic (1948-2012) 2. Female
## 15: 1. White non-Hispanic (1948-2012)   1. Male
## 16: 1. White non-Hispanic (1948-2012) 2. Female
## 17: 2. Black non-Hispanic (1948-2012)   1. Male
## 18: 1. White non-Hispanic (1948-2012) 2. Female
## 19: 1. White non-Hispanic (1948-2012) 2. Female
## 20: 1. White non-Hispanic (1948-2012)   1. Male
## 21: 1. White non-Hispanic (1948-2012)   1. Male
## 22: 1. White non-Hispanic (1948-2012)   1. Male
## 23: 1. White non-Hispanic (1948-2012) 2. Female
## 24: 1. White non-Hispanic (1948-2012)   1. Male
## 25: 1. White non-Hispanic (1948-2012) 2. Female
## 26: 1. White non-Hispanic (1948-2012)   1. Male
## 27:          5. Hispanic (1966-2012) 2. Female
## 28: 1. White non-Hispanic (1948-2012)   1. Male
## 29: 1. White non-Hispanic (1948-2012) 2. Female
## 30: 2. Black non-Hispanic (1948-2012) 2. Female
##                                       race   gender
```

```r
anes_use = anes_use%>%select(year, turnout, vote, race, gender)

save(anes_use, file="../output/data_use.RData")
```

4. A simple analysis

4.1 Who went to vote in the election?

First, we would like to see whether different racial groups have different turnout rates on the election day.

```r
load(file="../output/data_use.RData")
anes_to_race_year = anes_use %>%
  filter(!is.na(race) & !is.na(turnout))%>%
  group_by(year, race)%>%
  count(turnout)%>%
  group_by(year, race)%>%
  mutate(
    prop=n/sum(n)
  )

ggplot(anes_to_race_year,
       aes(x=year, y=prop, fill=turnout)) +
  geom_bar(stat="identity", colour="black") + facet_wrap(~race, ncol=1) + theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Blues")+
  labs(title="How did different racial groups participated in the election \n over the years?")
```

How did different racial groups participated in the election over the years?
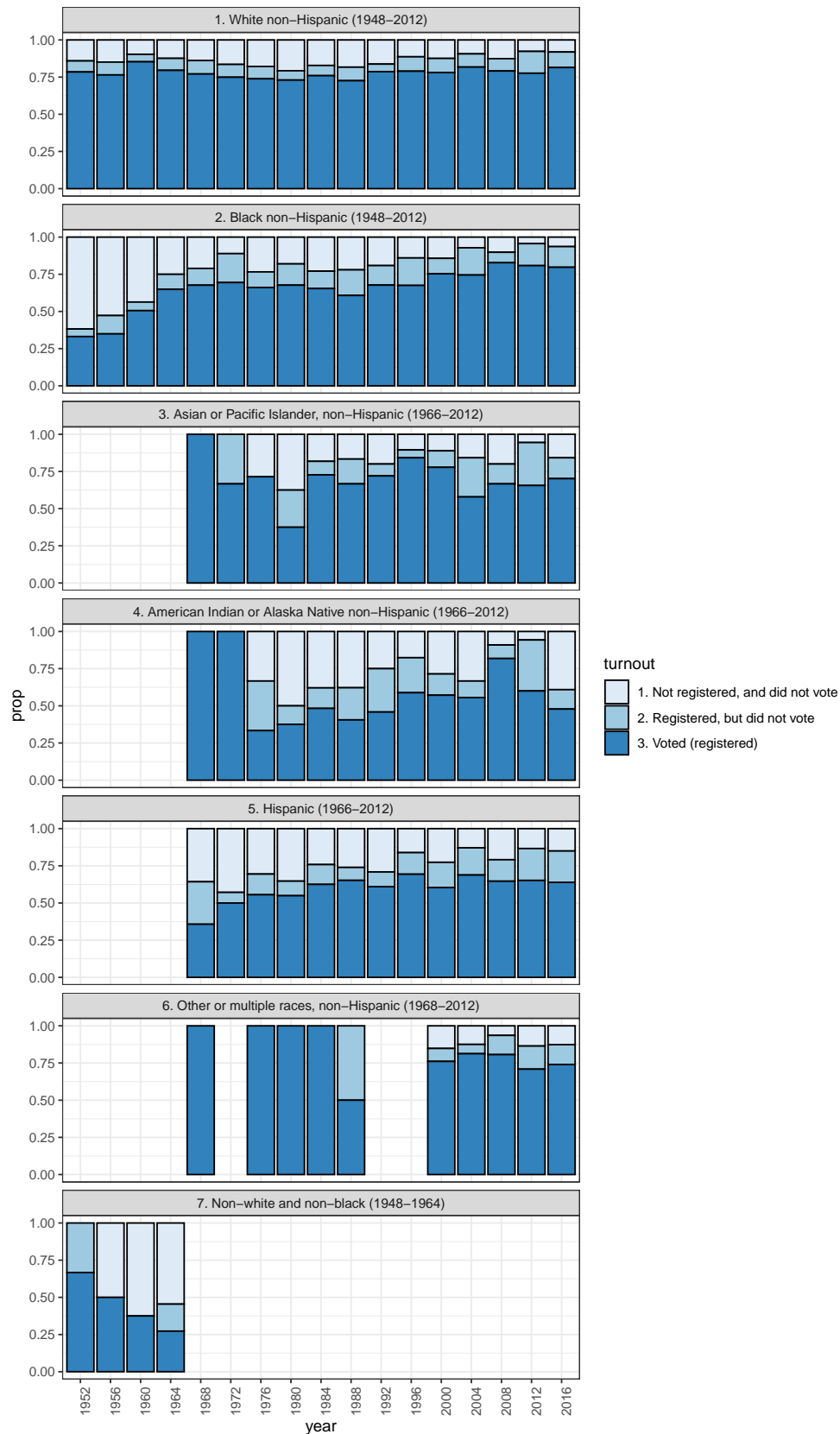
Wait a minute, this is not what we saw in the news!

Looks like other people have noticed it too and wrote a whole paper about this.

4.2 Who did they vote for in the election?

```r
anes_vote_race_year = anes_use %>%
  filter(!is.na(race) & !is.na(vote))%>%
  filter(vote!="7. Did not vote or voted but not for president (exc.1972)")%>%
  group_by(year, race)%>%
  count(vote)%>%
  group_by(year, race)%>%
  mutate(
    prop=n/sum(n)
  )
#%>%
#  filter(vote == "1. Democrat" | vote == "2. Republican")

ggplot(anes_vote_race_year,
       aes(x=year, y=prop, fill=vote)) +
  geom_bar(stat="identity", colour="black")+
  scale_fill_manual(values=c("dodgerblue3", "firebrick2", "gold1", "lightgoldenrod1"))+
  facet_wrap(~race, ncol=1) +
  theme(axis.text.x = element_text(angle = 90))+
  labs(title="Who did racial groups vote for in the election \n over the years?")
```

Who did racial groups vote for in the election over the years?