Preparing and Exploring our Election Panel Data

Chris Kardish and Marie Agosta

November 8, 2016

To reiterate briefly, we posit that predictors of individual voter turnout that commonly appear in academic literature will have different impacts on Millennial voters than older generations owing to significant differences in attitude, composition, and other areas. Some – such as education and party identification – will have less of an impact as measured in a logit regression, while others – such as gender, race, candidate attachment, employment status – will have greater predictive power. To assess these differences, the same set of variables will be applied to different generational age bands using a logistical regression model.

Our data source is the 2012 Times Series Study from American National Election Studies, a joint Stanford University and University of Michigan collaboration that is widely cited in academic literature. The study surveyed the same respondents both before and after the 2012 U.S. presidential election. There were approximately 6,000 respondents who were contacted via the internet and face to face. The database is comprehensive inscope, covering myriad demographic variables along with attitudes, opinions, and dispositions. While there is a 2013 follow-up with the same panel, the variables are not of interest and merging would require losing more than 3,000 respondents, making the validity of our finndings less robust.

Variables used include age, highest level of education, race, employment status, gender, past voting decisions, strength of party identification, sense of civic duty, and strength of candidate attachment. In the interest of word count and sheer length we will not include descriptive statistics for each variable in this document, but for those without tables, graphs, or other figures, we will provide greater detail about how they are coded and what exactly they measure.

Because ANES requires login information and its ZIP folders contain many variables, it is not possible to access the data using only a URL and tempfile function.

While the data set is quite clean, we begin with some minor things, such as recoding for readability of results and to better align with how we'll use each variable. Also, we can ultimately pare down this enormous data set to only variables of use, as demonstrated below.

```r
#rm(list = ls()) this command clears your environment
#setwd('C:/Users/Chris/Documents/GitHub/pair3')

library(foreign)
anes <- read.dta("anes_timeseries_2012_Stata12.dta")

#table(anes$dem_age_r_x) #age
anes <- anes[!(anes$dem_age_r_x <= -2 | anes$dem_age_r_x >= 88),]
#deletes outliers

#table(anes$dem_edugroup_x) #education
levels(anes$dem_edugroup_x) <- c("Refused", "Don't know", "Data missing",
                                 "Below high school", "High school",
                                 "Some post-high", "Bachelor", "Graduate")
#the above #applies new levels for improved readability later on

#table(anes$dem_raceeth_x) #race
levels(anes$dem_raceeth_x) <- c("Data missing", "White", "Black",
                                "Asian, Hawaiian, or Pacif Islr",
                                "Native American or Alaska Native",
                                "Hispanic", "Other")
```

```
#the above applies new levels for improved readability later on

#table(anes$dem_empstatus_initial) #employment
anes$employed <- anes$dem_emptype_work == "Working now" #creates dummy
#for whether respondent is employed

#table(anes$gender_respondent_x) #gender
anes$female <- anes$gender_respondent_x == "2. Female" #creates dummy gender #variable

#table(anes$interest_voted2008) #voting in past election (2008)
levels(anes$interest_voted2008) <- c("Refused", "Don't know", "Yes", "No")

#table(anes$pid_strong) #strong party ID
levels(anes$pid_strong) <- c("Refused", "Don't know", "Error", "Inapplicable",
                             "Strong", "Not very strong")
anes$pid_dummy <- anes$pid_strong == "Strong" #creates dummy

#table(anes$preswin_dutyst) #voting as civic duty
levels(anes$pid_strong) <- c("Refused", "Don't know", "Inapplicable",
                             "Very strongly", "Moderately strongly",
                             "A little strongly")
anes$duty <- anes$preswin_dutyst == "Very strongly"

#table(anes$postvote_prefprstr) #strong preference for candidate
anes$strong_pref <- anes$postvote_prefprstr == "Strong"

#combining needed variables in smaller data set
anes_small <- cbind(anes$dem_age_r_x, anes$dem_edugroup_x, anes$dem_raceeth_x,
                    anes$employed, anes$female, anes$interest_voted2008,
                    anes$pid_dummy, anes$preswin_dutyst, anes$strong_pref)
```

Given that our model will ultimately isolate the effcts of our chosen turnout predictors on various age groups, it is useful to break our giant panel data set into smaller data frames by age bands. This will entail having one age band exclusively for Millennials, who, at the time of the 2012 election, were 18-32 years of age, based on the defnition used by the Pew Research Center and other organizations. The next age band will be Generation X, people born between 1965-1979, which places them between the ages of 33 and 47. Next comes the Baby Boomers, who were born between 1946 and 1964 and ran from ages 48 to 66 at the time of the 2012 election.The Silent Generation, born between 1925 and 1945, accounts for the age band of 67 to 87. But first, let's see what the age variable actually looks like.

Table 1: Frequency of Age

Statistic	N	Mean	St. Dev.	Min	Max
.	5,828	49.26	16.65	17	87

Then we take a look at the frequency of observations by age.

We have all observations accounted for, with a mean age of respondent of 48, a fairly sizable standard deviation. We did have some outliers, one of which makes zero sense and has to be because of reporting error. 60 observations are coded with an age of -2. But those were dropped during the cleaning performed earlier. Additionally, because there are so few observations in the data set who are older than 87, the cuttoff age for the Silent Generation, we dropped those rows as well, because obtaining real results for this age group with so few observations in a logit model would be impossible. We do this by creating four different age bands

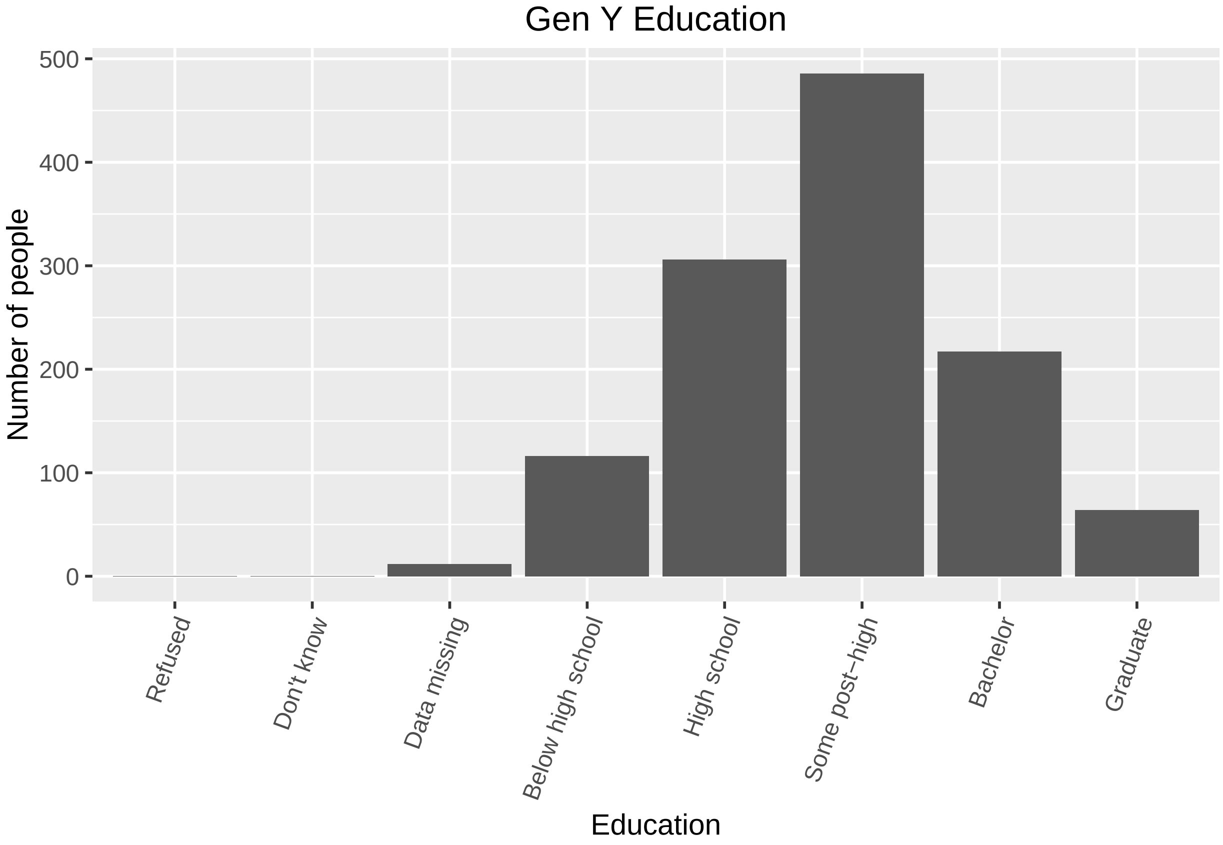
Table 2:

Age	Frequency
17	2
18	43
19	61
20	77
21	59
22	95
23	81
24	93
25	94
26	72
27	78
28	85
29	96
30	82
31	95
32	90
33	107
34	82
35	85
36	72
37	85
38	84
39	79
40	95
41	106
42	101
43	78
44	102
45	67
46	76
47	115
48	100
49	108
50	118
51	117
52	146
53	131
54	129
55	127
56	155
57	114
58	159
59	116
60	120
61	131
62	114
63	123
64	97
65	130
66	98
67	120
68	90
69	82
70	90
71	75
72	57
73	69

representing the four different generations above, which will exclude values younger than 18 and older than 87, when we eventually run logit models.

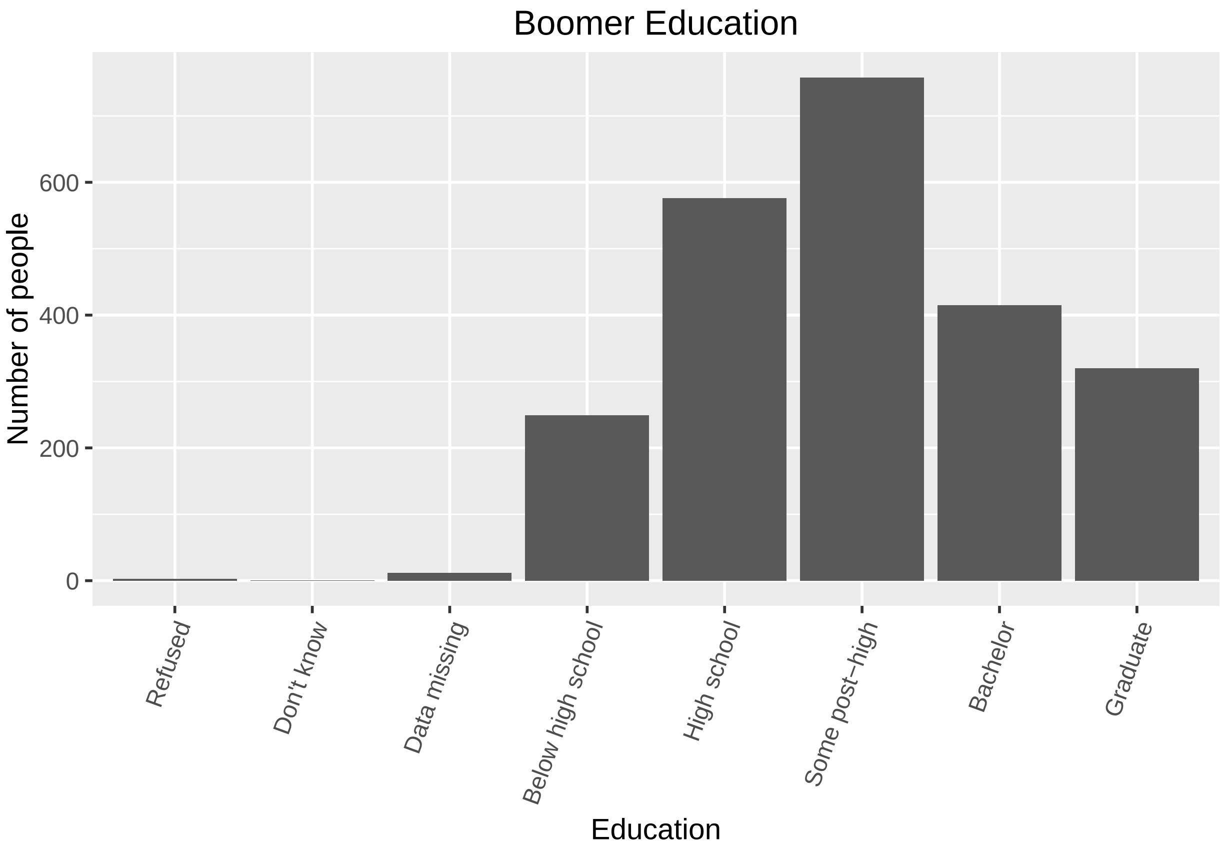
```
anes_genY <- subset(anes, anes$dem_age_r_x > 17 & anes$dem_age_r_x < 33)
#creates Millennial subset
anes_genX <- subset(anes, anes$dem_age_r_x > 32 & anes$dem_age_r_x < 48)
#creates Generation X subset
anes_boomer <- subset(anes, anes$dem_age_r_x > 47 & anes$dem_age_r_x < 67)
#creates Baby Boomer subset
anes_silent <- subset(anes, anes$dem_age_r_x > 66 & anes$dem_age_r_x < 88)
#creates Silent Generation subset
```

Now that we have the ability to separate our data by age, let's look at some key descriptive differences between the age groups, starting with education for Millennials.



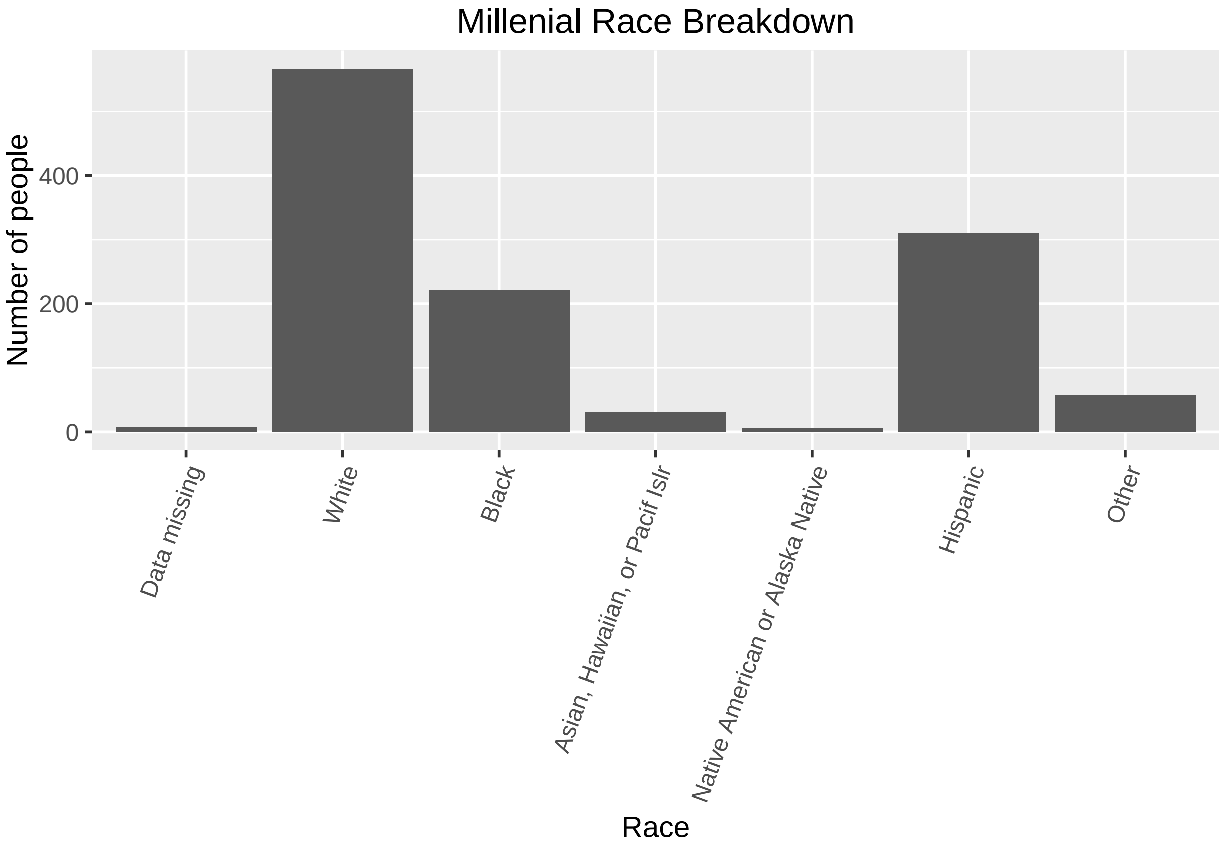
The largest share of Millennial respondents, nearly 500, list some post-high-school education.The secondhighest category is high school education alone.

Let's see how the data look for Baby Boomers.

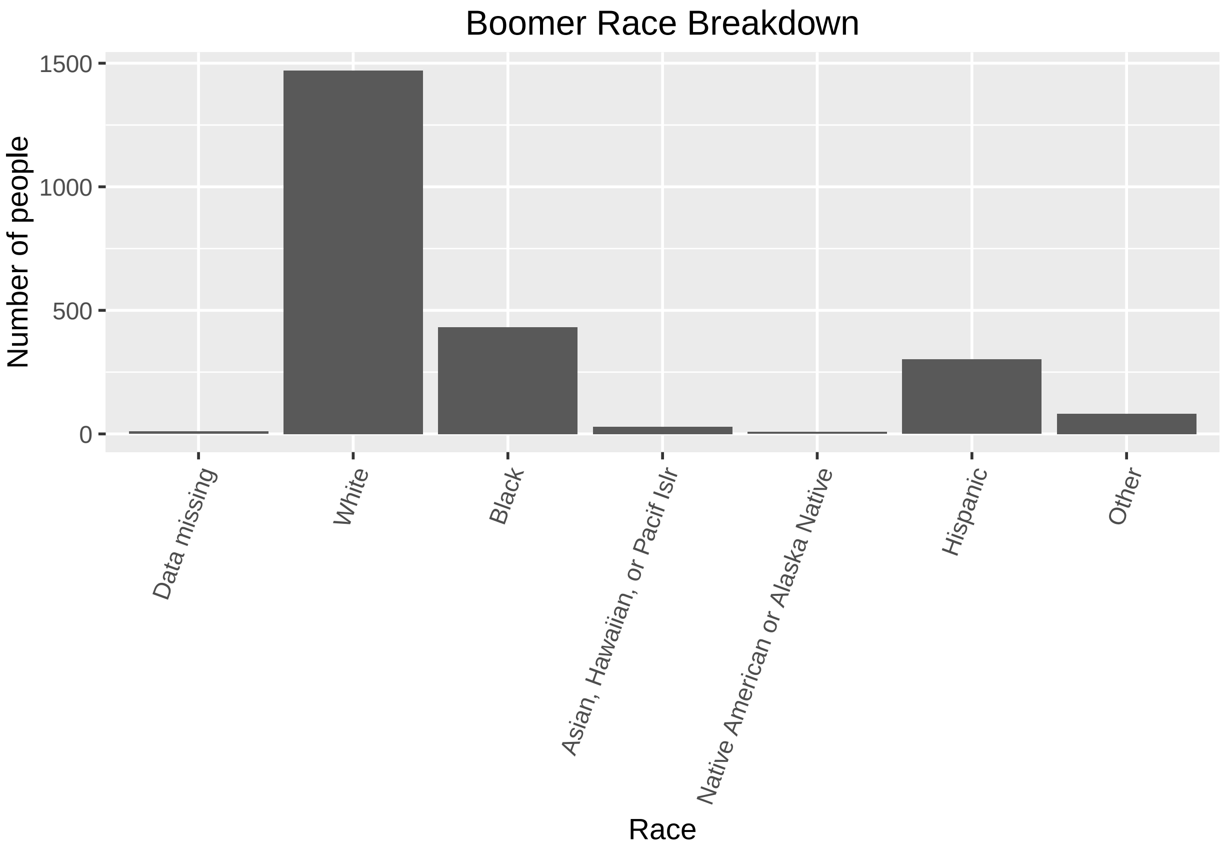


The data actually have a somewhat similar shape and distribution, just slightly different proportions. For one, a lot more Baby Boomers in this sample have graduate-level educations, which may reflect the relative youth of the Millennials in our sample. Now we'llll take a quick look at race by generation, and again we'll begin with Millennials.

Race by Generation



Now, we'll take a look at the Baby Boomers by race.



The difference visually is stark. The relatively lower proportions of black and hispanic people among Boomers is especially striking. Comparing the two groups with exact percentages, we have something like this:

Millennial Race Breakdown

Race Number of people Percentage

Data missing 8 1
White 567 47
Black 221 18
Asian, Hawaiian, or Pacif Islr 31 3
Native American or Alaska Native 6 0
Hispanic 311 26
Other 57 5

Boomer Race Breakdown

Race Number of People Percentage

Data missing 10 0
White 1,471 63
Black 432 19
Asian, Hawaiian, or Pacif Islr 29 1

Native American or Alaska Native 8 0
Hispanic 301 13
Other 82 4

These tables make the racial differences even clearer, showing that 47% of Millennials are non-white in this sample, compared with 33% of Baby Boomers, which generally refects the differences within the true population.

We'll show descriptive statistics for one more demographic variable before moving on to some other variables of interest in the sake of space and word count.

From academic literature we believe employment status will have an especially profound impact on Millennial voters by contributing to a lack of sense of political efficacy. We continue with differences specifically between Millennials and Baby Boomers, showing the following tables.

Table 3: Millennial Employment Breakdown

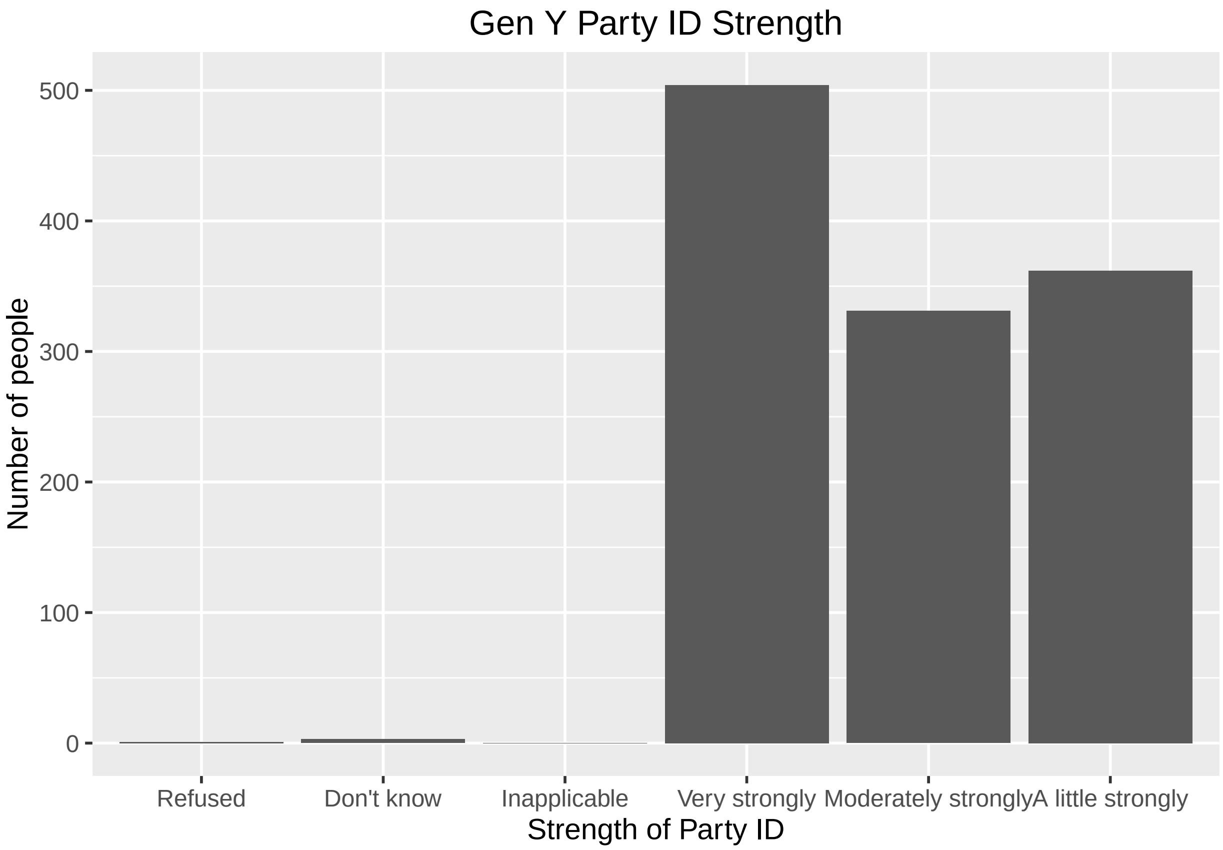
Employment Status	Number of People	Percentage
Refused	1	0
Don't know	1	0
Inapplicable	634	53
Working now	383	32
Temporarily laid off	19	2
Unemployed	49	4
Retired	1	0
Permanently disabled	10	1
Homemaker	34	3
Student	69	6

Table 4: Boomer Employment Breakdown

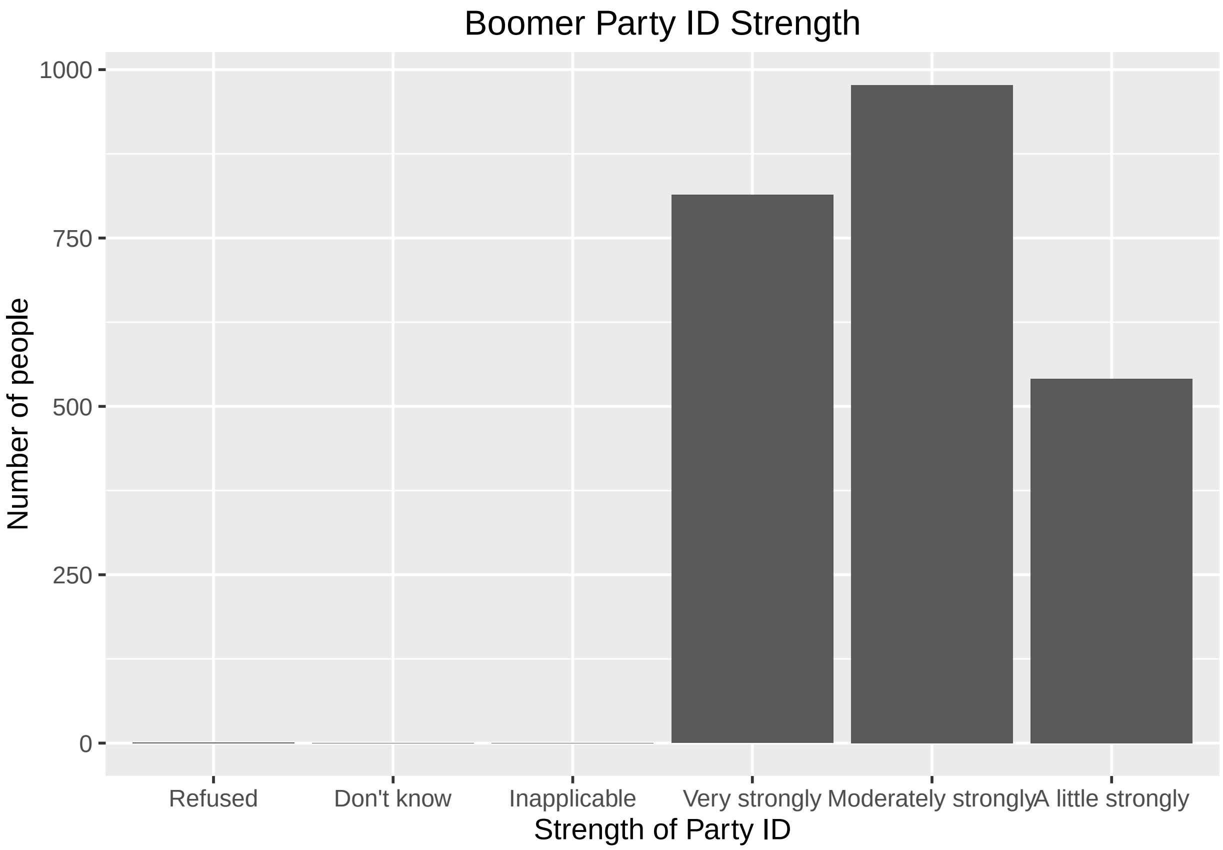
Employment Status	Number of People	Percentage
Refused	2	0
Don't know	2	0
Inapplicable	1,690	72
Working now	340	15
Temporarily laid off	14	1
Unemployed	26	1
Retired	114	5
Permanently disabled	97	4
Homemaker	43	2
Student	5	0

This one is quite surprising, but mostly because so many Boomers are coded as "inapplicable," which gives a false impression of unemployment. According to ANES, observations are coded "inapplicable" when "a question was not asked due to branching or skip patterns that made it inapplicable." Because there are other employment variables, we will substitute one for the final analysis, because the results of predictive statistics would ultimately give false impressions.

The last variable for which we'll provide descriptive statistics is the strenth of party identification, starting with Millennials. The variable does not specify party, but rather tells how much a respondent identifies with his or her party.



For Baby Boomers, the data look like this:



Surprisingly, considering literature routinely touts the unaffiliated status of Millennials, it is the Baby Boomers in this sample who identify less strongly with party in the extreme end of the spectrum.

The variables that will also be in the logistic regression model but haven't been discussed in greater detail yet include: gender, whether a respondent voted in the 2008 election, how strongly a respondent believes voting is a civic duty, and a respondent's strength of preference for his or her chosen candidate.

Gender is coded either 1 for male or 2 for female, but, as was shown earlier, we will be using it as a dummy variable in the model, with female coded as 1.

Whether a respondent voted in the 2008 presidential election is coded either as "refused?," "don't know," "yes", or "no," with the vast majority of respondents saying they did indeed vote.

The strength of the belief that voting is a civic duty ranges from "a little strongly" to "very strongly," with a sizable majority of respondents answering "very strongly," but in actuality most respondents are coded inapplicable. Similarly, most respondents are coded inapplicable for their strength of preference toward their chosen candidate, but more than twice as many people answered as strong than not strong.