2019 ANES Analysis: Descriptives with Skimr()

Damon C. Roberts

8/6/2020

Notes:

Description: Getting Descriptive Statistics of the cleaned 2019 ANES Pilot

Files:

Input: 'Data/anes__pilot__2019__dta/anes-2019-cleaned.dta'

Output:

Dependencies:

Input file cleaned using anes-cleaned-*.R files.

```r
here::here()
```

```
## [1] "C:/Users/damon/Dropbox/Viability and Race/code"
```

```r
source('anes-2019-analysis-descriptives.R')
names(dat$age) <- "Age"
names(dat$pid7) <- "PID"
names(dat$gender) <- "Male"
names(dat$educ) <- "Education"
names(dat$race) <- "Race"
names(dat$white) <- "White"
names(dat$black) <- "Black"
names(dat$ideo5) <- "Ideology"
names(dat$fttrump) <- "Feeling Thermometer: Trump"
names(dat$ftobama) <- "Feeling Thermometer: Obama"
names(dat$ftbiden) <- "Feeling Thermometer: Biden"
names(dat$ftwarren) <- "Feeling Thermometer: Warren"
names(dat$ftsanders) <- "Feeling Thermometer: Sanders"
names(dat$ftbuttigieg) <- "Feeling Thermometer: Buttigieg"
names(dat$ftharris) <- "Feeling Thermometer: Harris"
names(dat$ftblack) <- "Feeling Thermometer: Black"
names(dat$ftwhite) <- "Feeling Thermometer: White"
names(dat$vote20dem) <- "Primary Participation"
names(dat$vote20cand) <- "Candidate Support"
names(dat$vote20jb) <- "Vote Biden"
names(dat$vote20ew) <- "Vote Warren"
names(dat$vote20bs) <- "Vote Sanders"
names(dat$lcself) <- "Ideology: Own"
names(dat$lcd) <- "Ideology:Democrat"
names(dat$lcr) <- "Ideology:Republican"
names(dat$pop1) <- "The people have power"
names(dat$pop2) <- "Politicians do what is right"
names(dat$pop3) <- "Trust Politicians"
names(dat$att1) <- "Attend Church"
names(dat$att2) <- "Frequency of church attendance"
names(dat$pew_religimp) <- "Importance of religion"
names(dat$pew_churatd) <- "Frequency of church attendance"
names(dat$sentence) <- "Sentencing"
names(dat$newsint) <- "Interest in political news"
names(dat$electable) <- "Electability"
names(dat$raceid) <- "Racial Identity importance"
names(dat$racework) <- "Linked Fate"
```

Demographics

```r
demographic.subset <- dat %>%
  select(age, pid7, gender, educ, race, white, black, ideo5)

demographic.subset %>%
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	3165
Number of columns	8
Column type frequency:	
numeric	8
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	52.03	17.15	20	37	56	65	94	
pid7	0	1	0.18	2.63	-3	-2	0	2	8	
gender	0	1	0.48	0.50	0	0	0	1	1	
educ	0	1	3.50	1.49	1	2	3	5	6	
race	0	1	1.67	1.27	1	1	1	2	8	
white	0	1	0.70	0.46	0	0	1	1	1	
black	0	1	0.11	0.31	0	0	0	0	1	
ideo5	2	1	3.39	1.44	1	2	3	4	6	

Feeling Thermometers

```
ft.subset <- dat %>%
  select(fttrump, ftobama,ftbiden,ftwarren,ftsanders,ftbuttigieg,ftharris,ftblack,ftwhite)
ft.subset %>%
  skim()
```

Table 3: Data summary

Name	Piped data
Number of rows	3165
Number of columns	9
Column type frequency:	
numeric	9
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
fttrump	34	0.99	43.87	41.43	0	2	38	91	100	
ftobama	20	0.99	53.53	37.54	0	11	59	91	100	
ftbiden	51	0.98	42.15	33.44	0	7	42	70	100	
ftwarren	132	0.96	40.51	34.36	0	4	41	71	100	
ftsanders	45	0.99	42.15	34.88	0	5	42	73	100	
ftbuttigieg	313	0.90	38.66	30.40	0	7	41	61	100	
ftharris	213	0.93	35.30	30.18	0	4	35	58	100	
ftblack	5	1.00	70.86	23.78	0	51	73	91	100	
ftwhite	6	1.00	70.91	22.82	0	51	73	90	100	

Prospective Voting

```
vote.subset <- dat %>%
  select(vote20dem, vote20cand,vote20jb,vote20ew,vote20bs)
vote.subset %>%
  skim()
```

Table 5: Data summary

Name	Piped data
Number of rows	3165
Number of columns	5
Column type frequency:	
numeric	5
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
vote20dem	2	1.00	-0.07	0.90	-1	-1	0	1	1	
vote20cand	1770	0.44	4.40	2.70	1	1	5	7	9	
vote20jb	0	1.00	0.66	1.65	-1	-1	1	1	4	
vote20ew	0	1.00	0.70	1.65	-1	-1	1	1	4	
vote20bs	0	1.00	0.61	1.63	-1	-1	1	1	4	

Guesses of Ideological Placement

```
ideo_place.subset <- dat %>%
  select(lcself, lcr, lcd)
ideo_place.subset %>%
  skim()
```

Table 7: Data summary

Name	Piped data
Number of rows	3165
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
lcself	4	1	0.12	1.97	-3	-2	0	2	3	
lcr	7	1	1.57	1.50	-3	1	2	3	3	
lcd	3	1	-1.68	1.46	-3	-3	-2	-1	3	

Populism

```
pop.subset <- dat %>%
  select(pop1, pop2, pop3)
pop.subset %>%
  skim()
```

Table 9: Data summary

Name	Piped data
Number of rows	3165
Number of columns	3

Table 9: Data summary

Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pop1	2	1	0.78	1.09	-2	0	1	2	2	
pop2	2	1	0.95	1.01	-2	0	1	2	2	
pop3	0	1	-0.85	1.09	-7	-2	-1	0	2	

Church

```r
church.subset <- dat %>%
  select(att1, att2, pew_religimp, pew_churatd)
church.subset %>%
  skim()
```

Table 11: Data summary

Name	Piped data
Number of rows	3165
Number of columns	4
Column type frequency:	
numeric	4
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
att1	1	1.00	0.40	0.49	0	0	0	1	1	
att2	1887	0.40	2.36	1.28	1	1	2	4	5	
pew_religimp	0	1.00	2.75	1.19	1	2	3	4	4	
pew_churatd	64	0.98	4.20	1.76	1	3	5	6	6	

Political Attitudes (on Sentancing, political interest, traits for electability)

```r
polatt.subset <- dat %>%
  select(sentence, newsint, electable)
```

```
polatt.subset %>%
  skim()
```

Table 13: Data summary

Name	Piped data
Number of rows	3165
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
sentence	1586	0.50	3.32	1.88	1	2	3	4	7	
newsint	134	0.96	3.27	0.96	1	3	4	4	4	
electable	1769	0.44	1.80	0.78	1	1	2	2	4	

Race (race, linked fate, working together)

```
race.subset <- dat %>%
  select(race, raceid, racework)
race.subset %>%
  skim()
```

Table 15: Data summary

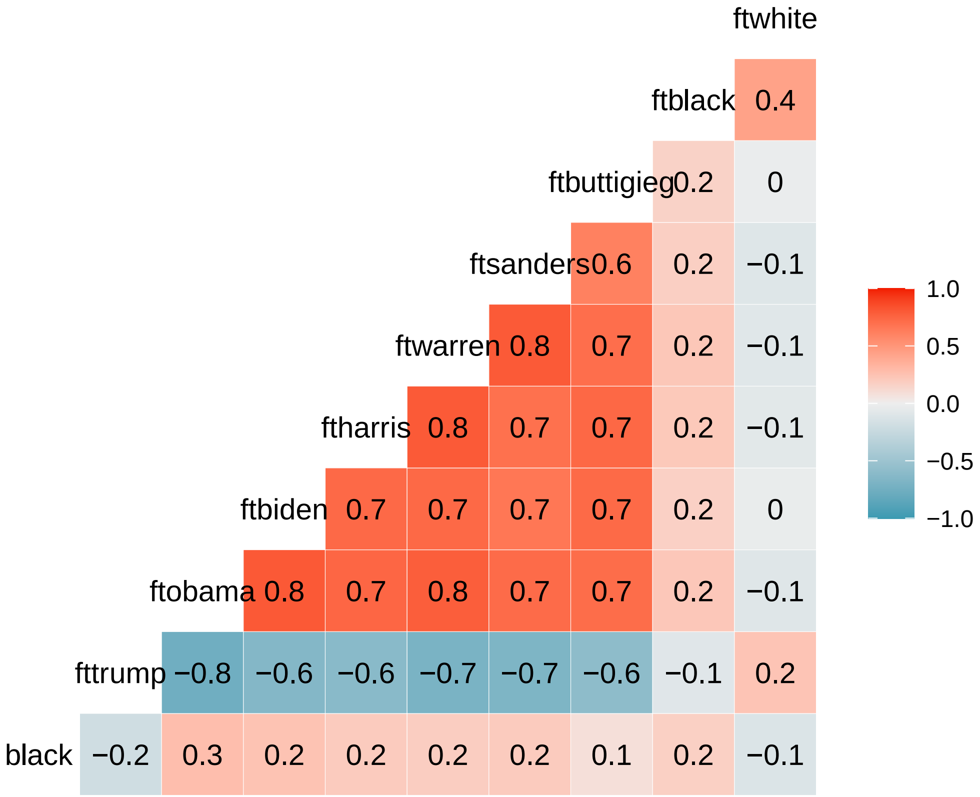
Name	Piped data
Number of rows	3165
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

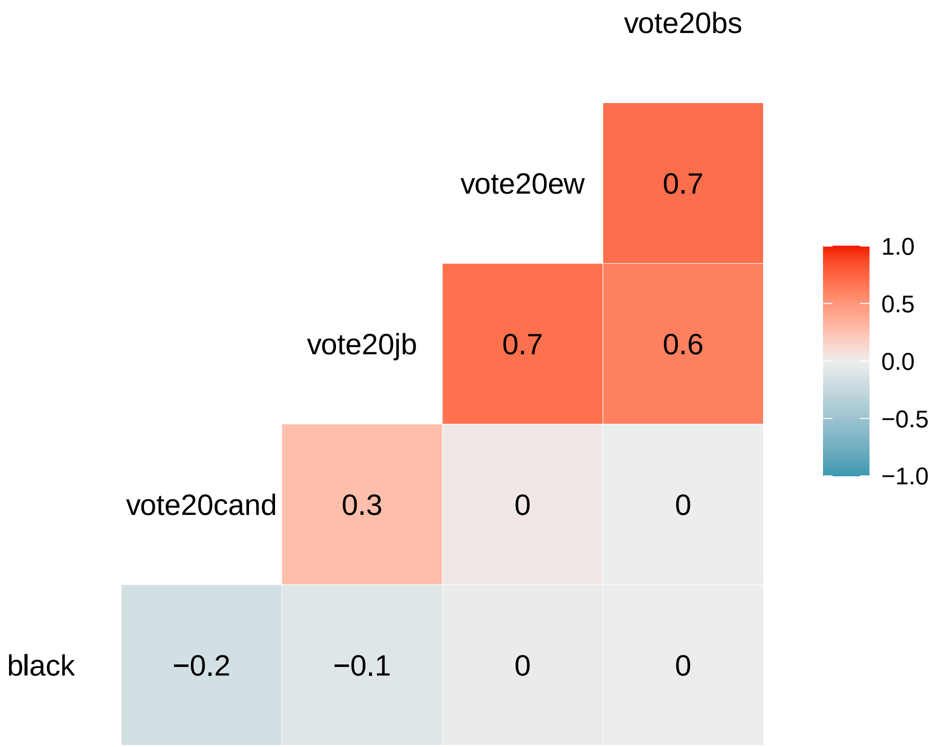
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
race	0	1.00	1.67	1.27	1	1	1	2	8	
raceid	114	0.96	2.70	1.45	1	1	3	4	5	
racework	120	0.96	2.94	1.49	1	1	3	4	5	

Exploring Some Relationships

```
corr1 <- dat %>%
  select(black, fttrump, ftobama, ftbiden, ftharris, ftwarren,ftsanders,ftbuttigieg,ftblack,ftwhite)

ggcorr(corr1, label = TRUE)
```

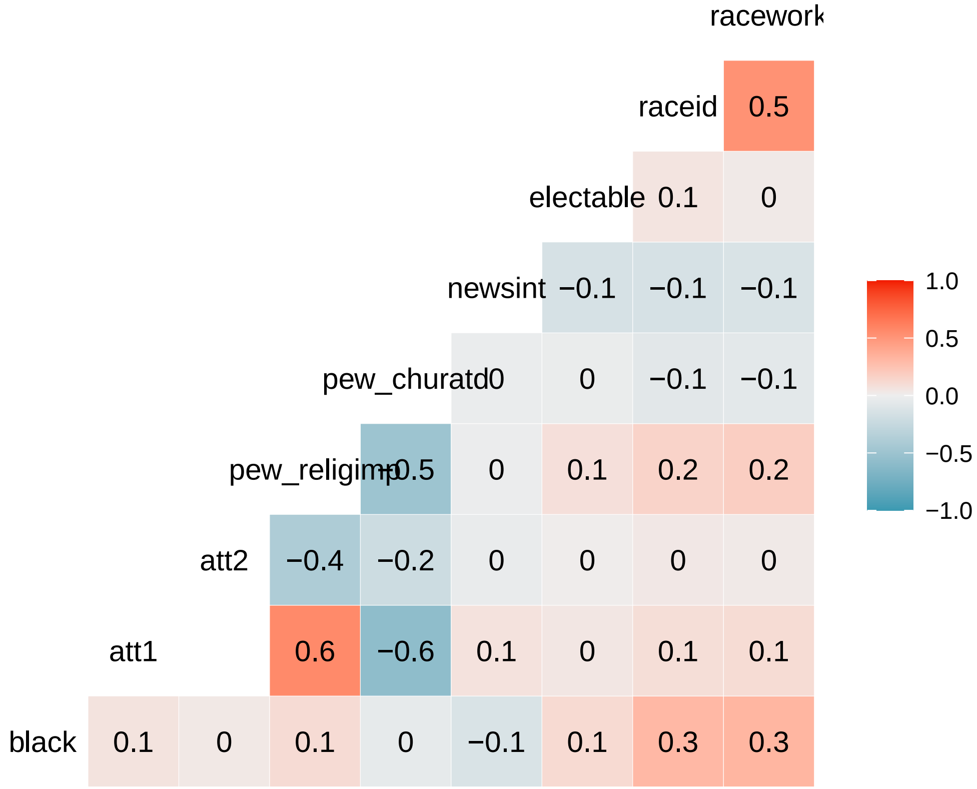


```
corr2 <- dat %>%
  select(black, vote20cand, vote20jb, vote20ew,vote20bs)

ggcorr(corr2, label = TRUE)
```



```
corr3 <- dat %>%
  select(black, att1, att2, pew_religimp, pew_churatd, newsint, electable, raceid, racework)

ggcorr(corr3, label = TRUE)
```

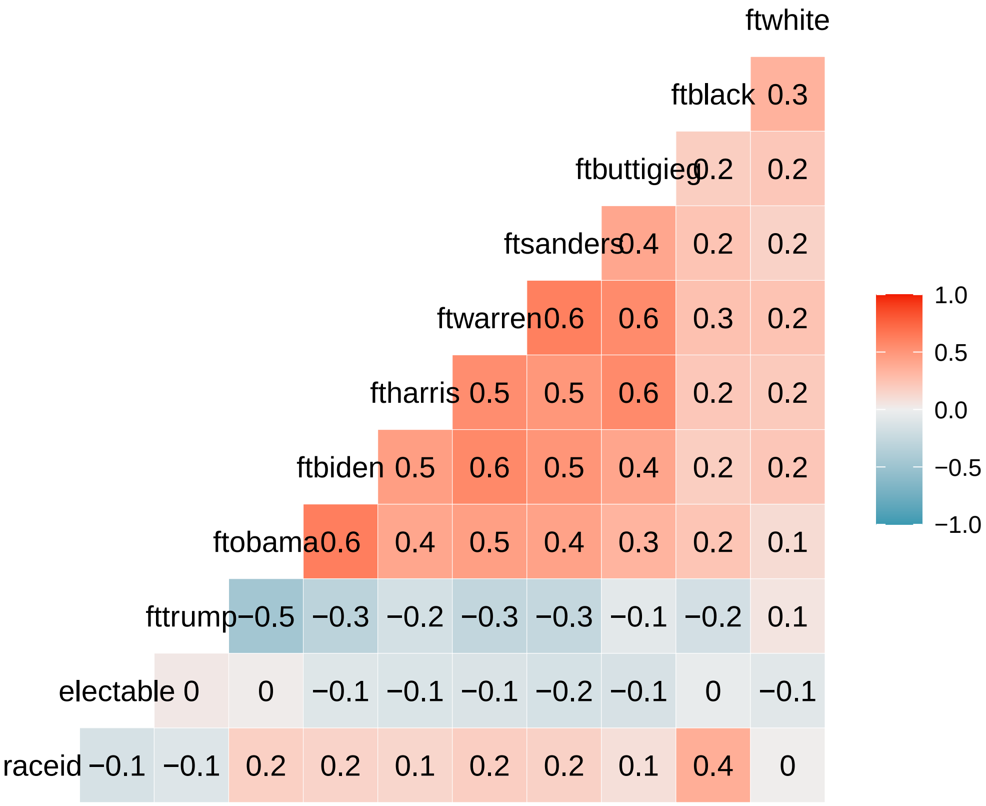
```
## Warning in cor(data, use = method[1], method = method[2]): the standard
## deviation is zero
```



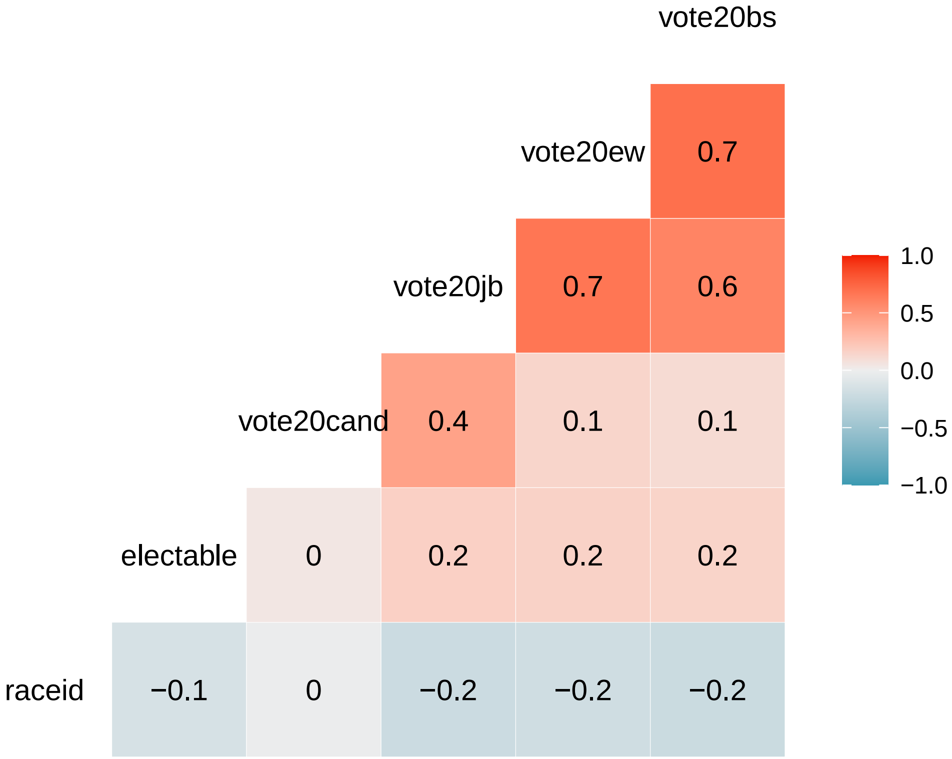
Additional Skimming (Black Respondents Only)

```
# Create Black respondent only dataset
bdat <- dat[ which(dat$black == 1), ]
```

```
corr1.2 <- bdat %>%
  select(raceid, electable, fttrump, ftobama, ftbiden, ftharris, ftwarren, ftsanders, ftbuttigieg, ftbla
ggcorr(corr1.2, label = TRUE)
```

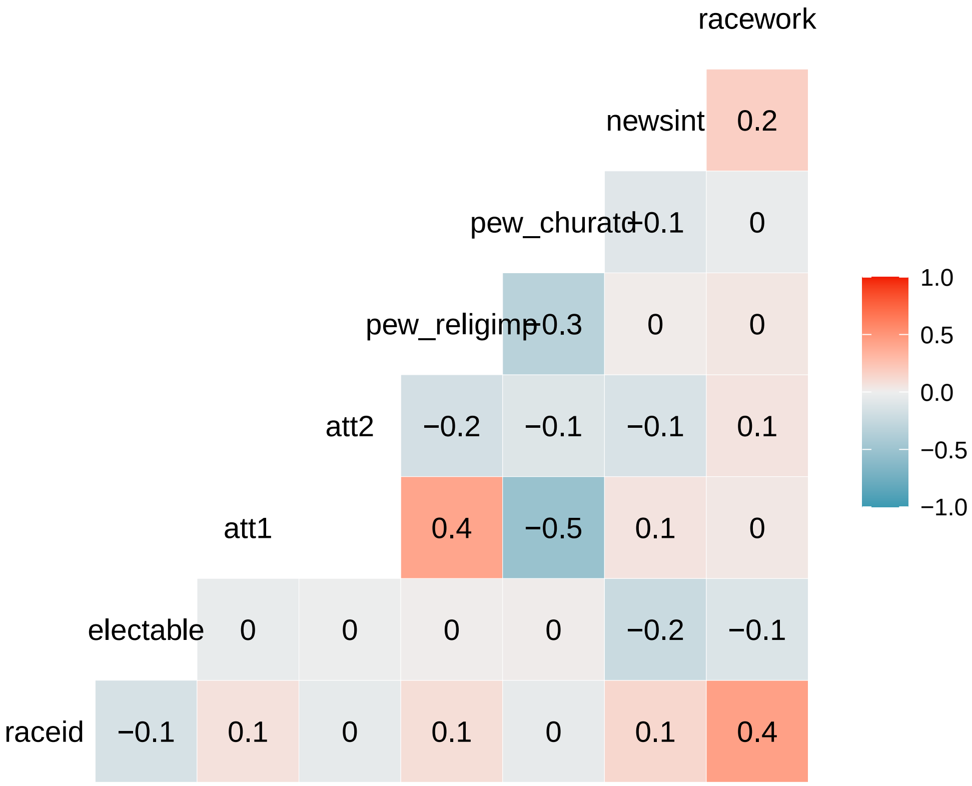


```
corr2.2 <- bdat %>%
  select(raceid,electable, vote20cand, vote20jb, vote20ew, vote20bs)
ggcorr(corr2.2, label = TRUE)
```



```r
corr3.2 <- bdat %>%
  select(raceid, electable, att1, att2, pew_religimp, pew_churatd, newsint, racework)
ggcorr(corr3.2, label = TRUE)
```

```
## Warning in cor(data, use = method[1], method = method[2]): the standard
## deviation is zero
```



Things to consider based on this

- The distributions of the variables are all over the place. We should check scatter plots for heteroskedasticity.
- There seems to be strong r correlations between the 2020 Democratic Presidential Candidates. That isn't really surprising but depending on how much of that is captured in other covariates it may make it really hard to find significant effects - especially with a somewhat small sample of Black Americans.
- Maybe I should play around with an SVM, BART, or maybe a Tree Ensemble after creating a dummies for the different candidates from the vote20cand variable to try to predict voter support - BART may be better since it uses prior distributions in it's optimization algorithm.
- For more traditional regression approaches. I might want to run difference of means tests, then run a parameterization transformation function to dummy it all out and then run an ANOVA.