Choate Data First Look

Theodore Dounias

5/4/2018

Merge

I have already completed this step, and the files are consolidated in a full document. For the sake of replicability, I might need to repeat this exercise.

I will also create a second dataset here that is significantly smaller, as for the diagnostic operations I do not need all the data.

```
diagnostic_vrf <- CO_2017_VRF_full %>%
  select(1:40)

rm(CO_2017_VRF_full)
```

Diagnostics

The total amount of records

```
length(unique(diagnostic_vrf$VOTER_ID))
```

```
## [1] 3734303
```

Note that every voter ID listed in this record is unique; no-one has their ID registered twice. There appear to be 43.387 files missing from the Colorado reported numbers. This is unlikelly to be due to an entry error, since each individual file is between 120k and 140k observations.

There are also 463,902 less unique voter ID's present in the Voting History file than there are in the voter information file.

The total amount of records in specific counties

Divided into counties, the counts for each county are:

```
summary(as.factor(diagnostic_vrf$COUNTY))
```

```
##       Adams     Alamosa    Arapahoe   Archuleta        Baca        Bent
##      270303        9849      410546       10506        2767        2983
##     Boulder  Broomfield     Chaffee    Cheyenne Clear Creek     Conejos
##      237091       49163       14562        1388        7916        5402
##    Costilla     Crowley      Custer       Delta      Denver     Dolores
##        2708        2082        3922       21434      450616        1669
##     Douglas       Eagle      Elbert     El Paso     Fremont    Garfield
##      237659       34703       19707      445708       30231       35632
##      Gilpin       Grand    Gunnison    Hinsdale    Huerfano     Jackson
##        4927       11522       13081         748        5081        1207
##   Jefferson       Kiowa  Kit Carson        Lake    La Plata     Larimer
##      422362        1029        4884        4916       43261      250626
##  Las Animas     Lincoln       Logan        Mesa     Mineral      Moffat
```

```
##       10302        3133       12755      115109         808        9716
##   Montezuma    Montrose      Morgan       Otero       Ouray        Park
##       19323       27210       16012       12020        4438       14091
##    Phillips      Pitkin     Prowers      Pueblo  Rio Blanco  Rio Grande
##        3182       14959        7213      109434        4413        7863
##       Routt    Saguache    San Juan  San Miguel    Sedgwick      Summit
##       19662        4280         719        6281        1787       26344
##      Teller  Washington        Weld        Yuma
##       19552        3335      182156        6015
```

Boulder

```
boulder_vrf_diag <- diagnostic_vrf %>%
  filter(COUNTY == "Boulder")

nrow(boulder_vrf_diag)
```

```
## [1] 237091
```

El Paso

```
elpaso_vrf_diag <- diagnostic_vrf %>%
  filter(COUNTY == "El Paso")

nrow(elpaso_vrf_diag)
```

```
## [1] 445708
```

Denver

```
denver_vrf_diag <- diagnostic_vrf %>%
  filter(COUNTY == "Denver")

nrow(denver_vrf_diag)
```

```
## [1] 450616
```

Total Active/Inactive

```
summary(as.factor(diagnostic_vrf$VOTER_STATUS))
```

```
##   Active Inactive     NA's
##  3167730   566572        1
```

Totals for Categorical Vars

Merge

Since this is an initial analysis, I will conduct a merge exclusively for the latest presidential election, and make calculations from there.

```
CO_2017_VHist_full$ELECTION_DATE <- mdy(CO_2017_VHist_full$ELECTION_DATE)

pres_votes <- CO_2017_VHist_full %>%
  filter(year(ELECTION_DATE) == 2016) %>%
  filter(ELECTION_TYPE == "General")


pres_votes <- pres_votes %>%
  filter(month(ELECTION_DATE) == 11) %>%
  select(2, 3, 8)
```

```
election_data_16 <- left_join(diagnostic_vrf, pres_votes, by = "VOTER_ID")
```

Interestingly enough, there is an amount of voters in the 2016 election that are not registered, since the left join returns more observations than were initially present in the diagnostic VRF dataset. However, the number of unique voter IDs is still the same:

```
length(unique(election_data_16$VOTER_ID)) - length(unique(diagnostic_vrf$VOTER_ID))
```

```
## [1] 0
```

Diagnostics on Election Year 2016

a) Turnout

The total amount of votes cast in the 2016 presidential election in Colorado are 2809019 according to my data, which is less than the reported 2,855,960.

For turnout, the voter history file should count the total number of individuals who voted in the 2016 presidential general. Therefore turnout should be that number over the total amount of registrants, or the total voting age population.

Something interesting that happens is that, if I use the effective date of registration to filter out all those effectivelly registered after the 2016 presidential election, the number I am left with is larger than the total votes.

```
#Turnout on effective date
diagnostic_vrf$EFFECTIVE_DATE <- mdy(diagnostic_vrf$EFFECTIVE_DATE)

turnout_calc_eff_date <- diagnostic_vrf %>%
  select(2, 9, 34) %>%
  filter(EFFECTIVE_DATE < as.Date("2016-11-08"))

nrow(turnout_calc_eff_date)
```

```
## [1] 2666824
```

```
nrow(pres_votes)
```

```
## [1] 2809019
```

If, instead, I use date of registration, I get the following results:

```
#Turnout by registration
diagnostic_vrf$REGISTRATION_DATE <- mdy(diagnostic_vrf$REGISTRATION_DATE)

turnout_calc_reg_date <- diagnostic_vrf %>%
  select(2, 9, 33) %>%
```

```
  filter(EFFECTIVE_DATE < as.Date("2016-11-08"))

nrow(turnout_calc_reg_date)
```

```
## [1] 2666824
```

Note that this, again, is a smaller number than the total votes cast. This for me implies that there have been individuals purged off of the voter rolls, who voted in 2016 but are not registared anymore.

This also means that I cannot accuratelly calculate turnout over gross total of registrants on election day. I can approximate this with the amount of registered active voters in 2017, when the snapshot of the VRF data was taken:

```
#Very approximate turnout over active voters
nrow(pres_votes)/sum(diagnostic_vrf$VOTER_STATUS == "Active", na.rm = TRUE)
```

```
## [1] 0.8867609
```

```
#Approximate Turnout over registered voters
nrow(pres_votes)/length(diagnostic_vrf$VOTER_STATUS)
```

```
## [1] 0.7522204
```

The reported values are 85% and 74.9% respectivelly.

I take the figure for voting age population from the Census Bureau, in order to calculate one last turnout statistic:

```
#Turnout over voting age population

vpop <- 3896986

nrow(pres_votes)/vpop
```

```
## [1] 0.7208183
```

This number is higher than the reported turnout calculated in this way, which is 71.3%.

b) Turnout in Counties

Since I have already seen some discrepancy between the data and the official reports, calculating the turnout here to see if it aligns is probably an unnecessary exercise–it obviously will not. However, if I wished to do so, the steps would be the following:

```
el_paso_pres_votes <- pres_votes %>%
  filter(COUNTY_NAME == "El Paso")
```

And then I would use the number of entries here, and a similar filtering on the VRF to obtain a turnout statistic.