North Carolina Voting Data Analysis

Team Multicollinearity

Abhisek Baral, Ravitashaw Bathla, Yiran (Becky) Chen, Nathan Warren, Jiayue (JY) Xu

10/29/2019

Summary

A study of the residents of North Carolina across different counties was performed to analyze the voting turnout rate across different demographics. It was observed that the results were statistically significant in determining the odds of turnout for each person based upon different demographics. People older than 40 years had a significantly higher turnout rate for voting than people below 40 years of age. Across the population, males were less likely to vote than females by 40 percent. However, the difference between voter turnout narrowed between the genders as the age increased until an age of 65 years. Older men (66+) were more likely to vote than older women. Across parties, the increase in odds from males to females are two times more for Democrat compared to Unaffiliated, and three times more compared to Libertarians or Republicans. People with undesignated gender had higher voting turnout than other genders but they represented only 1.3 percent of the total population.

1. Introduction

This study aims to analyze 2016 presidential voter data obtained from the North Carolina State Board of Elections for 20 counties (Appendix 1) within the state across different demographic. In addition, probability of voting across different counties is analyzed. We will further identify how voter turnout rates differ across demographics groups for different party affiliations.

Section 2 describes the data transformation and the exploratory data analysis to understand voter turnout based on individual factors. In Section 3, model building is described, and the final model is presented. The evaluation of the final model and the statistical significance of individual parameters are also discussed in Section 4. Section 5 presents summarized results that were obtained from model selection and conclusions inferred from the relationship. The limitations of the model are described in the final section.

2. Data

Two separate datasets were used - registered voter data (*registered*) and the actual voting data (*voted*) across different demographics of North Carolina state for all the counties. The datasets represent the aggregated actual voter counts and total voter registerations across different demographic groups like county, precinct, age group, gender etc.

Data Preparation and Cleaning

The 'registered' data (*voter_stats_20161108.txt*) has voting information including total number of registered voters for different demographics. In contrast, the 'voted' data (*history_stats_20161108.txt*) has information including total number of voters who actually voted. In order to calculate voter turnout, the two datasets

had to be joined together to calculate the turnout rate of voting across different demographic and geographic sub-groups.

Before joining the two datasets, several steps were performed to remove the missing values from both the datasets. This comprised of approximately 4.1 percent of the data combined. The following columns were removed since they did not play any significant roles in determining turnout rate of voters - election_date, update_date, stats_type, voting_method, and voting_method_desc. Removing these undesired columns resulted in multiple records within the same demographic and geographic sub-groups. Hence, the 'voted' data had to be aggregated, based upon the remaining 8 columns (county_desc, precinct_abbrv, vtd_abbrv, party_cd, race_code, ethnic_code, sex_code, age) common in both the datasets. This aggregation step provided us with voting count for unique demographic and geographic groups across the state.

The resulting voted data from the previous aggregation step was then joined with the registered data resulting in a single dataset- with the total number of registered voters and number of voters who actually voted, separated by different demographic for all the counties.

For the purpose of this study, 20 out of the 100 counties were selected at random for analysis after data cleaning and preparation (ref Appendix 1). The random sample of 20 counties selected which represents 12 percent of the total population of 100 counties of North Carolina.

There were some inconsistencies found in the dataset as there were 363 sub-groups which had more votes than voters registered. It is assumed that all the voters voted within their own registered precincts. Therefore inconsistencies in these sub-groups were removed which consisted of 1.4% of the population data in these 20 counties.

Data Transformation

A new calculated field 'non-voter' was added to the final data set for exploratory data analysis and model building. This field represents the number of people who were registered to vote but did not actually vote in the 2016 elections.

Exploratory Data Analysis

The overall voting turnout for the 20 counties of North Carolina was 67.5 percent. The voter turnout across the 20 counties was almost normally distributed. The median of the voter turnout rate across these 20 counties was 70 percent with a standard deviation of 6 percent. A majority of the counties have a voting turnout rate around the median and there are a few counties that deviate significantly from the median. CHATHAM has a much higher voting turnout rate, close to 80 percent. ROBESON and ONSLOW counties have a low voting turnout rate, around 55 percent. This suggests that county difference is likely to affect voter turnout rate, and certain counties may share similarities in voter turnout rate, which may stem from geographical adjacencies or historical reasons.

Within each county, the spread of the voting turnout across precincts varies. Alamance county has the largest voting turnout rate across different precincts ranging from 58 percent to 87 percent. Graham county had lowest voting turnout rate varying across different precincts from 64 percent to 72 percent. The number of precincts in counties varies from 4 to 40 with a median of 17 precincts. There were around 45 precincts across which the actual number of voters was less than 500 people across different demographics. This implies that there can be similarities among different precincts of the same counties and the relationship of voter turnout within the county can be shared where the population of registered and actual voters is small.

The overall distribution of male and female population was slightly skewed towards female at 54 percent of the total population. However, the voting turnout rate across male and female genders was 66 percent and 68 percent, respectively. 1.3 percent of the population did not disclose their gender. The voting turnout rate for this group of people was relatively high at 72 percent.

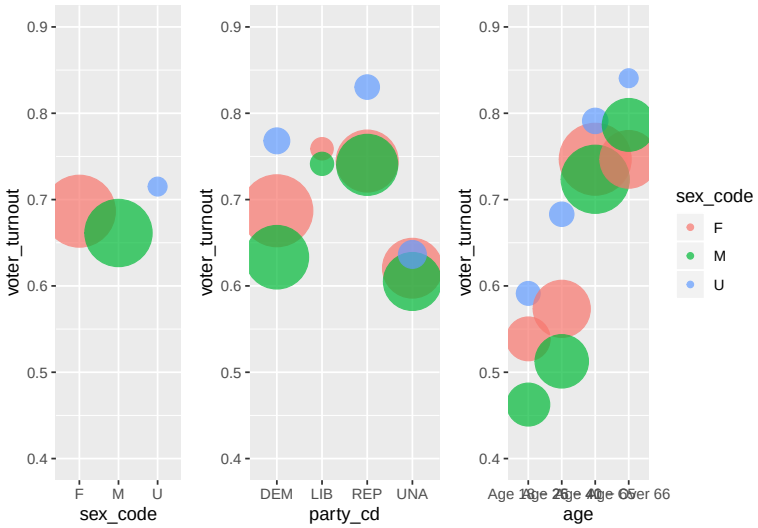
The population of North Carolina has a median age of 38.8 years (reference https://datausa.io/profile/geo/north-carolina). The sample data roughly mimics the population of North Carolina with 66 percent of population above 41 years of age. 23 percent of the people are in age group 26-40, and 10 percent of the total population is in the 18-25 age group. There is a large difference in voter turnout for people above and below 40 years of age. People older than 40 have a turnout rate around 75 percent, and those under 40 years of age have a turnout rate of around 54 percent. This suggests that older people are more likely to vote than younger people.

Across different age groups, the female population has a higher turnout rate than males except for the age group of people over 66 years old. The difference in voter turnout between genders narrowed as age increases until the age of 65. Combining historical contexts, it might be the case that this was the first time in the history of the United States that a woman was running for the presidential elections leading to a higher voting turnout rate for females. On the other hand, older men aging above 66 are more likely to vote than women at the same age group.

The race of the registered voters is predominantly White (71 percent) and Black (20 percent). The average turnout rate across these two races is 69 percent and 64 percent, respectively. The other races have varying turnout rates with the lowest for American Indians at 49 percent and the highest for the Mixed race group at 79 percent. However, the mixed group race is only a very small number of the total population.

The voter population is evenly spread out across different party affiliations, except for people affiliated to the Libertarian party. Voter turnout rate was highest at 74 percent for Republicans, followed by Democrats at 66 percent. Approximately 28 percent of the total population was unaffiliated to any party. The unaffiliated voter turnout rate was low, around 61 percent.

The voting turnout between genders varied across different party affiliations. For Democrats, females have a distinctly higher turnout rate compared to males. For all other party affiliations, including those Unaffiliated, both genders have similar turnout rates.



3. Model

Model Building Process

The methodology adopted in the model building was initialized with the variable selection of all demographic variables using AIC, BIC, and accuracy as the judgement criteria. Based on the EDA, there was a significant difference in turnout for different counties. Therefore, the variable 'county' was included as a classifier to account for variations across counties. As each of the five demographic subgroups (sex, age, race, ethnic and

party) was of key research interest, they were first included as a fixed effect. These variables were retained for the model building as all of them proved to be significant.

The interaction of sex with party and sex with age were subsequently added one by one on top of the base model and were tested for significance via change in deviance tests. When including precinct as a nested level within county, there was convergence problem likely due to a lack of data in some sub-groups, and therefore precinct is not included in the final model.

Model statistics of AIC, BIC, Accuracy, Sensitivity, and Specificity were also used to validate and compare each intermediate model. For each model, the predicted turnout rate for each demographic and geographic sub-groups was computed. The number was then decomposed into individual turnout status based on total registered voter count of each subgroups. Similarly, the observed individual turnout status was also computed for each sub-groups. The confusion matrix was obtained by comparing the predicted and observed for each subgroup. It was shown that by incorporating the two interaction terms, there was noticeable increase in accuracy, but decrease in the AIC and BIC score.

Final Model

All demographic variables (sex, age, race, ethnic, party) and the interactions of (party with sex) and (age with sex) as fixed effects and county as random effect are included in the final model, all of which were consistent with previous EDA findings of each predictor. The final model can predict the true positives and true negatives proportionally with an overall accuracy of 91.9 percent. The optimal sensitivity and specificity suggest that the model is 94.3 percent correct at identifying true positives (i.e. actual_voters), and 87.0 percent correct identifying true negatives (i.e. non-voters).

$$
\begin{gathered}
logit(Pr[vote_i = 1]) = \beta_0 + \gamma_{0m[i]}^{county} + \beta_1 sex_i + \\
\beta_2 age_i + \beta_3 race_i + \beta_4 ethnic_i + \beta_5 party_i + \\
\beta_6 sex_i : party_{cd_i} + \beta_7 sex_i : age_i \\
\gamma_{0m[i]}^{county_{desc}} \sim N(0, \tau^2) \\
\epsilon_{i,j} \overset{iid}{\sim} N(0, \sigma^2)
\end{gathered}
$$

4. Results

Based on the multilevel logistic regression output, a non-Hispanic or non-Latino white female aged 18-25 from Democratic party has 55 percent chance of turning out to vote, which is set as the baseline reference averaged across all different counties. The voting turnout rate differs by counties, with the standard deviation of county as random effects of 0.2355. Among the 20 random sampled counties, CHATHAM has the highest turnout rate of 66 percent compared to baseline, while ONSLOW and CHEROKEE having the lowest voting turnout rates of 42 percent and 43 percent respectively. The rest of the counties can be grouped into two groups where 5 (BURKE, GRAHAM, GATES, SWAIN, ROBESON) out of 20 counties are below the baseline, and 11 (YANCEY, BLADEN, LINCOLN, ALAMANCE, CASWELL, DUPLIN, JONES, ROCKINGHAM, WILSON, CATAWBA, VANCE) out of 20 counties are above the baseline.

Holding all other factors constant, males have generally 40% lower odds of voting than females. Regarding the interaction between sex and party, the odds of voting for female is around 66% more than male for Democratic party, while for other parties including Libertarians, Republicans, and Unaffiliated, the odds of voter turnout for female is around 23%, 26%, and 31% more than males respectively. The increase in odds from males to females are two times more for Democrat compared to Unaffiliated, and three times more compared to Liberaltarians or Republicans. For the interaction between sex and age, the difference in odds of voting between genders decrease as age increases until age of 65, which agrees with previous EDA findings.

Females have 66% higher odds of voting than males for people aged from 18 to 25, 56% higher for people aged from 26 to 40, and 34% higher for those aged from 41 to 65. On the other hand, the odds of voting for males overtakes females by 10% for people aged over 66.

Compared with white people, black or African Americans have 3 percent higher odds of voting, Asians 21 percent higher, mixed race 1.8 times higher, American Indians 35 percent higher, and other races 9 percent higher. Regarding ethnics, Hispanic or Latinos have 29 percent higher odds of voting compared with non-Hispanic or non-Latino people. Compared with democratic, given other covariates being constant, Libertarians have 100 percent higher odds of voting, Republicans 35 percent higher, and those unaffiliated 16 percent lower. Regarding age groups, compared with people aged 18 to 25, for people aged between 26 to 40, they have 16 percent higher odds of voting, while people aged from 41 to 65 2.4 times higher, and people aged above 67 2.3 times higher. It can be seen that older generation has double odds of voting compared to the younger generation.

5. Conclusion

It is found that the overall probability or odds of voting differs to a relatively large extent across different counties in 2016. Further, the five main demographic variables (sex, age, race, ethnic, party) and two interaction terms (sex and age, sex and party) are mostly significant. Generally, males have 40 percent lower odds of voting than females. The older generation have higher odds of voting where people above 40 have two times higher odds of voting compared to younger people. People who are mixed races or Asians have higher odds of voting compared to white people, 1.8 times and 21 percent respectively. Hispanic or Latinos have 29 higher odds of voting compared with non-Hispanic or non-Latino people. Compared to democrats, libertarians have higher odds of voting with 100 percent, Republicans 35 percent more while the unaffiliated have lower odds. Across parties, the increase in odds from males to females are two times more for democrat compared to Unaffiliated, and three times more compared to Liberaltarians or Republicans.

6. Limitations

There were several limitations in the study which may have affected the accuracy of identifying the factors which influence the voter turnout rate. Firstly, the age was categorically represented into 4 groups with unknown distribution within each age group, which could lead to biased result of the significance of age on voting turnout rate. Secondly, there were counties which had higher number of actual voters than registered for those counties. This suggests that gerrymandering might happen in all counties causing vote counts not representative of true vote counts of registered voters in that county. There were about 1.3 percent of people who did not disclose their gender and were assumed to be a third-gender category, which may not hold in reality. In addition, there could have been other external factors like advertising budget across different counties for the election or the accessibility of polling stations for people. These factors could also impact the final results if such information was provided.

7. Appendix

Appendix 1

Counties

ALAMANCE, ROBESON, CATAWBA, GATES, LINCOLN, ROCKINGHAM, ONSLOW, VANCE, BLADEN, WILSON, DUPLIN, NORTHAMPTON BURKE, GRAHAM, CHATHAM, CASWELL, YANCEY, JONES, SWAIN, CHEROKEE

Appendix 2

Code

```
# load libraries
library(dplyr)
library(ggplot2)
library(gridExtra) # for grid.arrange(): combine ggplots
library(lme4) # for glmer(): hierarchical logistics regression
library(rstan)
```

```
## Loading required package: StanHeaders
```

```
## rstan (Version 2.19.2, GitRev: 2e1f913d3ca3)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
```

```
##
## Attaching package: 'rstan'
```

```
## The following object is masked from 'package:arm':
##
##     traceplot
```

```
library(brms)
```

```
## Loading required package: Rcpp
```

```
## This version of Shiny is designed to work with 'htmlwidgets' >= 1.5.
##     Please upgrade via install.packages('htmlwidgets').
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Loading 'brms' package (version 2.10.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

##
## Attaching package: 'brms'

## The following object is masked from 'package:rstan':
##
##     loo

## The following object is masked from 'package:e1071':
##
##     rwiener

## The following object is masked from 'package:lme4':
##
##     ngrps

## The following object is masked from 'package:survival':
##
##     kidney
```

```
library(sjPlot) # for tab_model()
library(pROC) # for roc()
library(arm) # for binnedplot()
library(e1071)
library(caret) # for confusionMatrix()
library(rlist) # for list.append()
library(lattice) # for dotplot()
library(kableExtra) # for kable()
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
##### Data cleaning #####

### step1: load data

# set pwd to git directory
getwd()
```

```
## [1] "/Users/N1/team-project-2-estrogen-bioassay-and-voting-in-nc-team-multicollinearity/Reports"
```

```
setwd('../')
getwd()
```

```
## [1] "/Users/N1/team-project-2-estrogen-bioassay-and-voting-in-nc-team-multicollinearity"
```

```
# load actual voters data --> 734126 rows
voted <- read.delim("/Users/N1/team-project-2-estrogen-bioassay-and-voting-in-nc-team-multicollinearity/
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

```
summary(voted)
```

```
##       county_desc     precinct_abbrv     vtd_abbrv
##  WAKE       : 81711   02     :  5480          : 29977
##  MECKLENBURG: 69893   04     :  4848   04     :  6215
##  GUILFORD   : 41792   13     :  4421   02     :  5569
##  FORSYTH    : 31034   14     :  4330   06     :  5349
##  CUMBERLAND : 27589   11     :  4289   05     :  5144
##  DURHAM     : 22439   03     :  4279   03     :  4834
##  (Other)    :459668   (Other):706479   (Other):677038
##                                  age         party_cd     race_code
##  Age < 18 Or Invalid Birth Dates:     1   DEM:281251   A: 30568
##  Age 18 - 25                     :143983   LIB: 15153   B:151494
##  Age 26 - 40                     :197257   REP:183579   I: 14236
##  Age 41 - 65                     :242494   UNA:254143   M: 22940
##  Age Over 66                     :150391                O: 61254
##                                                         U: 86230
##                                                         W:367404
##  ethnic_code sex_code    total_voters        election_date
##  HL: 69748    :     7   Min.   :  1.000   11/08/2016:734126
##  NL:379552   F:359661   1st Qu.:  1.000
##  UN:284826   M:321380   Median :  2.000
##              N:     1   Mean   :  6.495
##              U: 53077   3rd Qu.:  5.000
##                         Max.   :710.000
##
##    stats_type          update_date     voting_method
##  history:734126   12/28/2016:734126   O      :323141
##                                       V      :255293
##                                       M      : 81693
##                                       U      : 32234
##                                       P      : 22031
##                                       C      : 10564
##                                       (Other):  9170
##          voting_method_desc voted_party_cd
##  ABSENTEE ONESTOP :323141      :    47
##  IN-PERSON        :255293   DEM:282682
##  ABSENTEE BY MAIL : 81693   LIB: 15120
##  ABSENTEE CURBSIDE: 32234   REP:183208
##  PROVISIONAL      : 22031   UNA:253069
##  CURBSIDE         : 10564
##  (Other)          :  9170
```

```
# load registered voters data --> 514844 rows
registered <- read.delim("/Users/N1/team-project-2-estrogen-bioassay-and-voting-in-nc-team-multicollinea
summary(registered)
```

```
##       county_desc        election_date    stats_type     precinct_abbrv
##  WAKE       : 54343   11/08/2016:514848   voter:514848   02     :  3946
##  MECKLENBURG: 50284                                      04     :  3344
##  GUILFORD   : 31643                                      13     :  3183
##  CUMBERLAND : 21425                                      03     :  3076
##  FORSYTH    : 21062                                      15     :  3070
##  DURHAM     : 15491                                      05     :  3063
##  (Other)    :320600                                      (Other):495166
##    vtd_abbrv      party_cd       race_code      ethnic_code sex_code
##  04     :  4238   DEM:187511   W      :180532   HL: 81358   F:235968
##  02     :  3938   LIB: 20055   B      : 98031   NL:221671   M:226908
##  06     :  3859   REP:122440   U      : 81335   UN:211819   U: 51972
##  05     :  3735   UNA:184842   O      : 70521
##  03     :  3489                A      : 33891
##  13     :  3329                M      : 31981
##  (Other):492260                (Other): 18557
##           age          total_voters
##  Age 18 - 25:122352   Min.   :   1.00
##  Age 26 - 40:150020   1st Qu.:   1.00
##  Age 41 - 65:156036   Median :   2.00
##  Age Over 66: 86440   Mean   :  13.45
##                       3rd Qu.:   8.00
##                       Max.   :1598.00
##
```

```r
### step 2: remove unnecessary rows

# remove duplicate rows(if any), and save into new df
voted_cleaned <- unique(voted)
registered_cleaned <- unique(registered)
# remove na (if any)
voted_cleaned <- na.omit(voted_cleaned)
registered_cleaned <- na.omit(registered_cleaned)
# remove rows with empty string "" and " "
voted_cleaned <- voted_cleaned[!apply(voted_cleaned, 1, function(x) any(x=="")),] # 704102 rows (1140 r
voted_cleaned <- voted_cleaned[!apply(voted_cleaned, 1, function(x) any(x==" ")),] # 704095 rows (7 rows
registered_cleaned <- registered_cleaned[!apply(registered_cleaned, 1, function(x) any(x=="")),] # 5137
registered_cleaned <- registered_cleaned[!apply(registered_cleaned, 1, function(x) any(x==" ")),] # 513
# drop unused levels
voted_cleaned <- droplevels(voted_cleaned)
registered_cleaned <- droplevels(registered_cleaned)

### step 3: rename and remove unnecessary columns

# drop voted party_cd column (duplicate with voted_party_cd), election_date, update_date,
# stats_type, voting_method, voting_method_desc column
voted_cleaned <- subset(voted_cleaned, select = -c(party_cd, election_date, update_date,
                                                   stats_type, voting_method, voting_method_desc))
# drop registered election_date, stats_type column
registered_cleaned <- subset(registered_cleaned, select = -c(election_date, stats_type))

# rename voted voted_party_cd to party_cd to align naming with registered
colnames(voted_cleaned)[colnames(voted_cleaned)=="voted_party_cd"] <- "party_cd"
# rename voted total_voters to actual_voters to avoid same name with registered
```

```r
colnames(voted_cleaned)[colnames(voted_cleaned)=="total_voters"] <- "actual_voters"

### step 4: check every common (joint) variable if levels match (8 variables in total)

# for loop for every variable
to_remove_variable_list = c("total_voters")
variable_list = colnames(registered_cleaned)
variable_list <- variable_list[-match(to_remove_variable_list, variable_list)]
check_level <- data.frame(variable_list, check_level = rep(FALSE, length(variable_list)))

for (variable in variable_list){
  index = match(c(variable), variable_list)
  check_level[index,2] <- all(levels(voted_cleaned[,variable]) == levels(registered_cleaned[,variable]))
}
```

```
## Warning in levels(voted_cleaned[, variable]) ==
## levels(registered_cleaned[, : longer object length is not a multiple of
## shorter object length
```

```
## Warning in levels(voted_cleaned[, variable]) ==
## levels(registered_cleaned[, : longer object length is not a multiple of
## shorter object length
```

```r
check_level
```

```
##    variable_list check_level
## 1    county_desc        TRUE
## 2 precinct_abbrv        TRUE
## 3      vtd_abbrv        TRUE
## 4       party_cd        TRUE
## 5      race_code        TRUE
## 6    ethnic_code        TRUE
## 7       sex_code       FALSE
## 8            age       FALSE
```

```r
# age and sex_code levels do not match
# removed extra levels in voted
voted_cleaned <- voted_cleaned[!(voted_cleaned$age == "Age < 18 Or Invalid Birth Dates"), ] # 704094 ro
voted_cleaned <- voted_cleaned[!(voted_cleaned$sex_code == "N"), ] # 704093 rows (1 rows removed)
# drop unused levels
voted_cleaned <- droplevels(voted_cleaned)
registered_cleaned <- droplevels(registered_cleaned)
# re-run for loop to check if levels match

### step 5: aggregate voted after removing voting_method

# check for duplicate rows
any(duplicated(registered_cleaned))
```

```
## [1] FALSE
```

```
any(duplicated(voted_cleaned))
```

```
## [1] TRUE
```

```
# aggregate actual_voters based on 8 demo variables in voted, resave into new df --> 403343 rows
aggregate_list = list()
for (variable in variable_list){
  aggregate_list <- list.append(aggregate_list, voted_cleaned[,variable])
}
voted_cleaned_agg <- aggregate(voted_cleaned$actual_voters, by=aggregate_list, FUN=sum)
colnames(voted_cleaned_agg) <- c(variable_list, "actual_voters")

### step 6: merging and sample

# merge voted_cleaned_agg with registered_cleaned --> 397687 rows with 10 variables
voting2016 <- inner_join(voted_cleaned_agg, registered_cleaned)
```

```
## Joining, by = c("county_desc", "precinct_abbrv", "vtd_abbrv", "party_cd", "race_code", "ethnic_code"
```

```
# select 20 random counties --> 48720 rows
set.seed(1000)
sample_county <- sample(unique(voting2016$county_desc), 20, replace=F)
sample_voting2016 <- voting2016[is.element(voting2016$county_desc, sample_county),]
# remove unused levels
sample_voting2016 <- droplevels(sample_voting2016)
# data summary
summary(sample_voting2016)
```

```
##    county_desc    precinct_abbrv    vtd_abbrv     party_cd    race_code
##  ALAMANCE: 6064   30     :  477   11     :  558   DEM:18481   A: 1943
##  CATAWBA : 5663   13     :  472   NE22   :  490   LIB: 1300   B: 9997
##  ONSLOW  : 4888   11     :  435   30     :  477   REP:12029   I: 2214
##  ROBESON : 4531   07     :  420   13     :  472   UNA:16910   M: 1852
##  LINCOLN : 3342   20     :  385   07     :  420               O: 5072
##  BURKE   : 3173   EN03   :  372   20     :  385               U: 7082
##  (Other) :21059   (Other):46159   (Other):45918               W:20560
##  ethnic_code sex_code           age        actual_voters
##  HL: 5991    F:23162   Age 18 - 25:10331   Min.   :  1.00
##  NL:22976    M:21361   Age 26 - 40:13298   1st Qu.:  1.00
##  UN:19753    U: 4197   Age 41 - 65:15871   Median :  2.00
##                        Age Over 66: 9220   Mean   : 11.61
##                                            3rd Qu.:  8.00
##                                            Max.   :571.00
##
##   total_voters
##  Min.   :  1.00
##  1st Qu.:  1.00
##  Median :  3.00
##  Mean   : 17.17
##  3rd Qu.: 13.00
##  Max.   :676.00
##
```

```r
str(sample_voting2016)
```

```
## 'data.frame':    48720 obs. of  10 variables:
##  $ county_desc   : Factor w/ 20 levels "ALAMANCE","BLADEN",..: 1 1 1 1 15 1 1 15 5 9 ...
##  $ precinct_abbrv: Factor w/ 330 levels "0001","0003",..: 40 56 57 68 75 81 87 101 104 127 ...
##  $ vtd_abbrv     : Factor w/ 334 levels "0001","0003",..: 38 52 53 64 71 76 82 96 99 120 ...
##  $ party_cd      : Factor w/ 4 levels "DEM","LIB","REP",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ race_code     : Factor w/ 7 levels "A","B","I","M",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ ethnic_code   : Factor w/ 3 levels "HL","NL","UN": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex_code      : Factor w/ 3 levels "F","M","U": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age           : Factor w/ 4 levels "Age 18 - 25",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ actual_voters : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ total_voters  : int  2 1 1 1 1 1 1 1 1 1 ...
```

```r
### step 7: remove rows with actual_voters > total_voters

# new column for number of non voters
sample_voting2016$non_voters <- sample_voting2016$total_voters - sample_voting2016$actual_voters
# 363 rows with -ve non voter counts: actual voters more than total
nrow(sample_voting2016[sample_voting2016$non_voters<0,])
```

```
## [1] 363
```

```r
# remove and save --> 48357 rows
sample_voting2016 <- sample_voting2016[!(sample_voting2016$non_voters<0), ]

### step 8: export clean datasets to csv

# write.csv(sample_voting2016, file = "sample_voting2016.csv")
# write.csv(voting2016, file = "voting2016.csv")
# write.csv(voted_cleaned_agg, file = "voted_cleaned_agg.csv")
# write.csv(voted_cleaned, file = "voted_cleaned.csv")
# write.csv(registered_cleaned, file = "registered_cleaned.csv")

# ********************************************************************************

##### Exploratory data analysis #####

# response variable: percent distribution of voters vs nonvoters: 68:32
sum(sample_voting2016$actual_voters)/sum(sample_voting2016$total_voters)
```

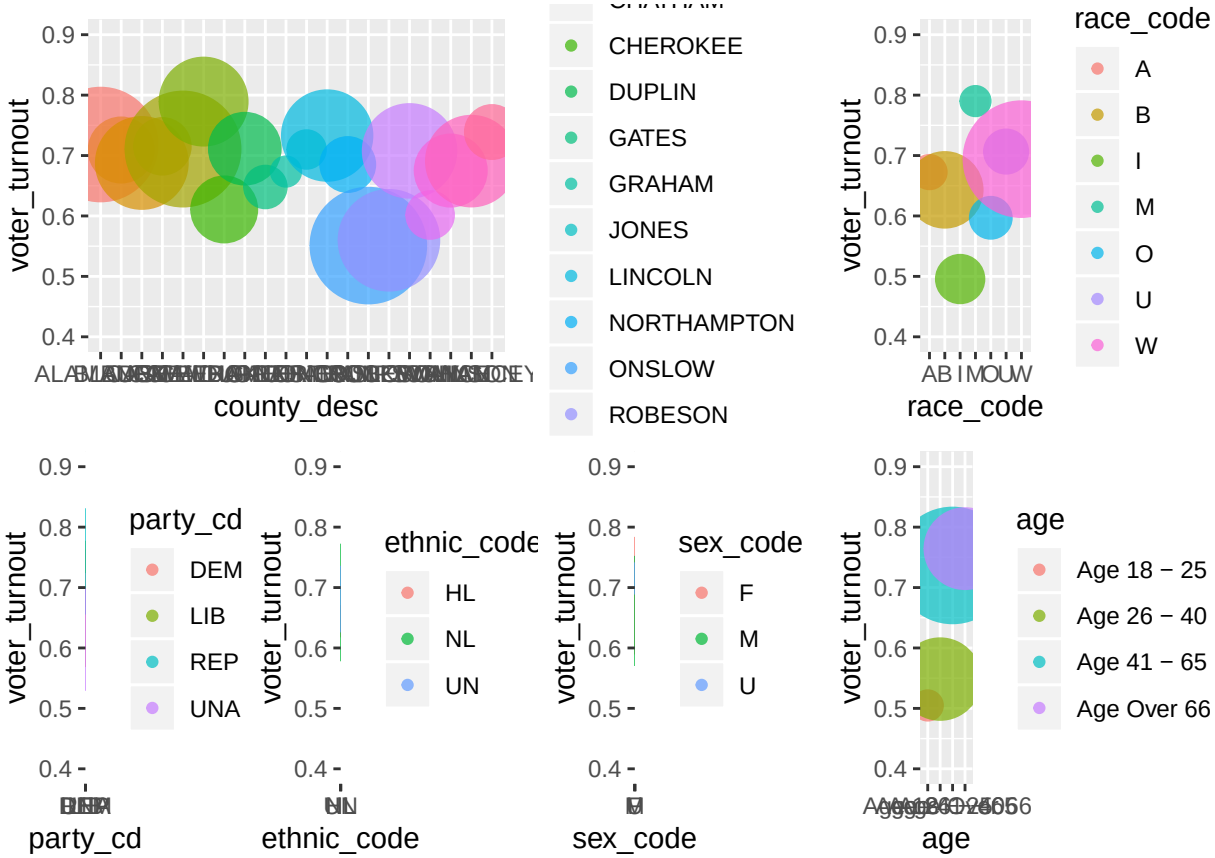
```
## [1] 0.6755987
```

```r
sum(sample_voting2016$non_voters)/sum(sample_voting2016$total_voters)
```

```
## [1] 0.3244013
```

```r
### eda for each of county & 5 demo variables: sex_code, age, ethnic_code, race_code, party_cd

to_remove_predictor_list = c("actual_voters", "total_voters", "non_voters")
```

```
predictor_list <- colnames(sample_voting2016)
predictor_list <- predictor_list[-match(to_remove_predictor_list, predictor_list)]

for (predictor in predictor_list) {

  eda_i <- aggregate(sample_voting2016$actual_voters, by=list(sample_voting2016[,predictor]), FUN=sum)
  colnames(eda_i) <- c(predictor, "actual_voters")
  eda_i$non_voters <- aggregate(sample_voting2016$non_voters, by=list(sample_voting2016[,predictor]), FU
  eda_i$total_voters <- aggregate(sample_voting2016$total_voters, by=list(sample_voting2016[,predictor]]
  eda_i$voter_turnout <- eda_i$actual_voters/eda_i$total_voters
  eda_i$per_total_voters <- eda_i$total_voters/sum(eda_i$total_voters)
  eda_df.name = paste("eda", toString(predictor), sep = "_")
  assign(eda_df.name, eda_i)

  gg_i <- ggplot(eda_i,
         aes(y=voter_turnout,x=.data[[predictor]],color=.data[[predictor]]))+
    geom_point(aes(size=total_voters), alpha=0.7)+
    scale_size_continuous(range = c(5, 20))+
    ylim(0.4, 0.9)+
    xlab(predictor)+
    labs(color = predictor)+
    guides(size=FALSE)
  gg.name = paste("gg", toString(predictor), sep = "_")
  assign(gg.name, gg_i)
}

grid.arrange(gg_county_desc, gg_party_cd, gg_race_code, gg_ethnic_code, gg_sex_code, gg_age,
               widths = c(1, 1, 1, 1.5),
               layout_matrix = rbind(c(1, 1, 1, 3),
                                     c(2, 4, 5, 6)))
```

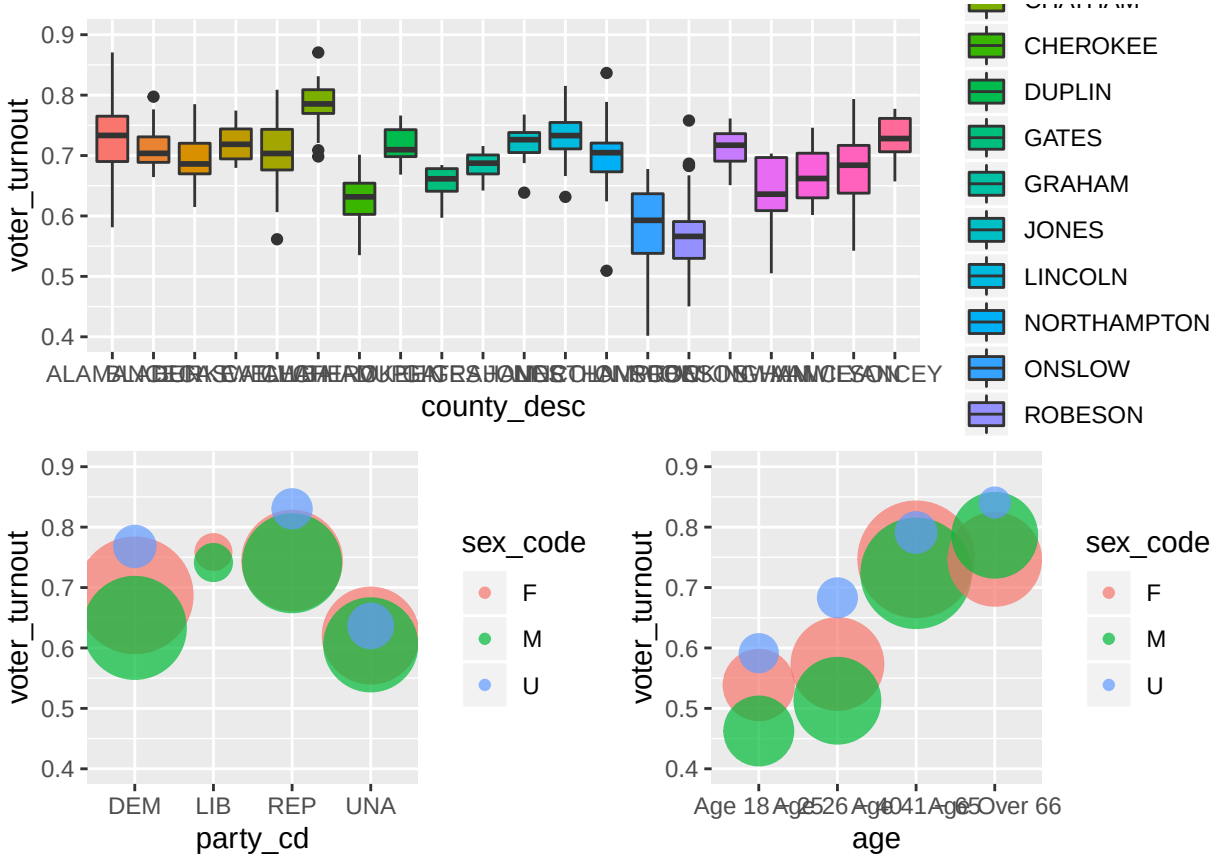



```
### eda for nested location variables: county_desc:precinct_abbrv

nested_list = data.frame(c("county_desc", "precinct_abbrv"))

for (nested in nested_list) {

  var1 = toString(nested[1])
  var2 = toString(nested[2])
  eda_i <- aggregate(sample_voting2016$actual_voters, by=list(sample_voting2016[,var1], sample_voting20
  colnames(eda_i) <- c(var1, var2, "actual_voters")
  eda_i$non_voters <- aggregate(sample_voting2016$non_voters, by=list(sample_voting2016[,var1], sample_
  eda_i$total_voters <- aggregate(sample_voting2016$total_voters, by=list(sample_voting2016[,var1], samp
  eda_i$voter_turnout <- eda_i$actual_voters/eda_i$total_voters
  eda_i$per_total_voters <- eda_i$total_voters/sum(eda_i$total_voters)
  eda_df.name = paste("eda", var1, var2, sep = "_")
  assign(eda_df.name, eda_i)

  eda_count_i <- eda_i %>%
    group_by(.data[[var1]]) %>%
    summarise(no_rows = length(.data[[var1]]))
  eda_count_i <- data.frame(eda_count_i)
  eda_count_df.name = paste("eda_count", var1, var2, sep = "_")
  assign(eda_count_df.name, eda_count_i)

  gg_i <- ggplot(eda_i,
                 aes(y=voter_turnout,x=.data[[var1]],fill=.data[[var1]]))+
```

```
    geom_boxplot()+
    ylim(0.4, 0.9)+
    xlab(var1)+
    labs(fill = var1)+
    guides(size=FALSE)
  gg.name = paste("gg", var1, var2, sep = "_")
  assign(gg.name, gg_i)
}

### eda for interaction term: party_cd:sex_code, age:sex_code

interaction_list = data.frame(c("party_cd", "sex_code"),
                              c("age", "sex_code"))

for (interaction in interaction_list) {

  var1 = toString(interaction[1])
  var2 = toString(interaction[2])
  eda_i <- aggregate(sample_voting2016$actual_voters, by=list(sample_voting2016[,var1], sample_voting20
  colnames(eda_i) <- c(var1, var2, "actual_voters")
  eda_i$non_voters <- aggregate(sample_voting2016$non_voters, by=list(sample_voting2016[,var1], sample_
  eda_i$total_voters <- aggregate(sample_voting2016$total_voters, by=list(sample_voting2016[,var1], samp
  eda_i$voter_turnout <- eda_i$actual_voters/eda_i$total_voters
  eda_i <- eda_i %>%
    group_by(.data[[var1]]) %>%
    mutate(per_total_voters = (total_voters/sum(total_voters)))
  eda_i <- data.frame(eda_i)
  eda_i <- eda_i[order(eda_i[,var1]), ]
  eda_df.name = paste("eda", var1, var2, sep = "_")
  assign(eda_df.name, eda_i)

  gg_i <- ggplot(eda_i,
                 aes(y=voter_turnout,x=.data[[var1]],color=.data[[var2]]))+
    geom_point(aes(size=total_voters), alpha=0.7)+
    scale_size_continuous(range = c(5, 20))+
    ylim(0.4, 0.9)+
    xlab(var1)+
    labs(color = var2)+
    guides(size=FALSE)
  gg.name = paste("gg", var1, var2, sep = "_")
  assign(gg.name, gg_i)
}

grid.arrange(gg_county_desc_precinct_abbrv, gg_party_cd_sex_code, gg_age_sex_code,
             widths = c(1, 1),
             layout_matrix = rbind(c(1, 1),
                                   c(2, 3)))
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
# ********************************************************************************

##### Model Building #####

# glm_table <-data.frame(c("Model 1", "Model 2", "Model 3", "Model 4"),
#                        rep(0,4), rep(0,4),rep(0,4),rep(0,4),rep(0,4),rep(0,4),rep(0,4))
# colnames(glm_table) <- c("Model","AIC","BIC","AUC","Threshold","Accuracy","Sensitivity",
#                          "Specificity")

glm_table <-data.frame(c("Model 1", "Model 2", "Model 3", "Model 4"),
                       rep(0,4), rep(0,4),rep(0,4),rep(0,4),rep(0,4))
colnames(glm_table) <- c("Model","AIC","BIC","Accuracy","Sensitivity",
                         "Specificity")

# fixed effects on all 5 demo variables + random effects on county
formula1 = cbind(actual_voters,non_voters) ~ sex_code + age + race_code + ethnic_code + party_cd + (1|c
# fixed effects on all 5 demo variables & interaction of sex:party + random effects on county
formula2 = update(formula1, ~ . + sex_code:party_cd)
# fixed effects on all 5 demo variables & interaction of sex:party and sex:age
# + random effects on county
formula3 = update(formula2, ~ . + sex_code:age)
# fixed effects on all 5 demo variables & interaction of sex:party and sex:age
# + random effects on county/precinct (nested)
formula4 = update(formula3, ~ . - (1|county_desc) + (1|county_desc/precinct_abbrv))

formula_list = c(formula1, formula2, formula3, formula4)
```

```
model = list()
summary = list()
fixef = list()
ranef = list()
tab_model = list()
newdata = list()
roc = list()
conf_mat = list()

for (formula in formula_list[1]) {
  index = match(c(formula), formula_list)
  # Regression Results
  model_i <- glmer(formula, family=binomial(link="logit"),data=sample_voting2016)
  model <- list.append(model, model_i)
  summary <- list.append(summary, summary(model_i))
  fixef <- list.append(fixef, fixef(model_i))
  ranef <- list.append(ranef, ranef(model_i))
  tab_model <- list.append(tab_model, tab_model(model_i))

  # AIC, BIC
  glm_table$AIC[index] <- AIC(model_i)
  glm_table$BIC[index] <- BIC(model_i)

  # ROC, AUC
  # pred_i <- predict(model_i, type="response")
  # pred_i <- data.frame(cbind(sample_voting2016$actual_voters, sample_voting2016$non_voters, pred=pred_
  #
  # vote_yes <- pred_i[rep(row.names(pred_i), pred_i$actual_voters),]
  # vote_yes$obs <- 1
  # vote_no <- pred_i[rep(row.names(pred_i), pred_i$non_voters),]
  # vote_no$obs <- 0
  # newdata_i <- rbind(vote_yes, vote_no)[, c("pred", "obs")]
  # newdata <- list.append(newdata, newdata_i)
  # roc_i <- roc(newdata_i$obs,newdata_i$pred,plot=T,print.thres="best",legacy.axes=T,
  #            print.auc =T,col="red3")
  # roc.name = paste("roc", toString(index), sep = "")
  # assign(roc.name, roc_i)
  # glm_table$AUC[index] <- auc(roc_i)
  pred_i <- predict(model_i, type="response")
  pred_i <- data.frame(cbind(obs_actual=sample_voting2016$actual_voters,
                             obs_non=sample_voting2016$non_voters,
                             total_voters=sample_voting2016$total_voters, predp=pred_i))
  pred_i$pred_actual <- round(pred_i$total_voters*pred_i$predp)
  pred_i$pred_non <- pred_i$total_voters - pred_i$pred_actual
  pred_i$diff <- pred_i$obs_actual - pred_i$pred_actual

  vote_yes <- pred_i[rep(row.names(pred_i), pmin(pred_i$obs_actual,pred_i$pred_actual)),]
  vote_yes[,c("obs", "pred")] <- c(1,1)
  vote_no <- pred_i[rep(row.names(pred_i), pmin(pred_i$obs_non,pred_i$pred_non)),]
  vote_no[,c("obs", "pred")] <- c(0,0)
  vote_diff <- pred_i[rep(row.names(pred_i), abs(pred_i$diff)),]
  vote_diff[,c("obs", "pred")] <- c(as.numeric(vote_diff$diff>0),1-as.numeric(vote_diff$diff>0))
  newdata_i <- rbind(vote_yes, vote_no, vote_diff)[, c("obs", "pred")]
```

```
  newdata <- list.append(newdata, newdata_i)

  # Threshold, Accuracy, Sensitivity, Specificity
  # glm_table$Threshold[index] <- coords(roc_i, "best", ret = "threshold")
  # glm_table$Accuracy[index] <- coords(roc_i, "best", ret = "accuracy")
  # glm_table$Sensitivity[index] <- coords(roc_i, "best", ret = "sensitivity")
  # glm_table$Specificity[index] <- coords(roc_i, "best", ret = "specificity")

  # Confusion Matrix
  # conf_mat_i <- confusionMatrix(as.factor(ifelse(newdata_i$pred >= as.numeric(glm_table$Threshold[ind
  #                               as.factor(newdata_i$obs),positive = "1")
  conf_mat_i <- confusionMatrix(as.factor(newdata_i$pred),as.factor(newdata_i$obs),positive = "1")
  conf_mat <- list.append(conf_mat, conf_mat_i$table)
  glm_table$Accuracy[index] <- conf_mat_i$overall["Accuracy"]
  glm_table$Sensitivity[index] <- conf_mat_i$byClass["Sensitivity"]
  glm_table$Specificity[index] <- conf_mat_i$byClass["Specificity"]
}

glm_table
```

```
##     Model      AIC      BIC  Accuracy Sensitivity Specificity
## 1 Model 1 149953.6 150111.8 0.9170617    0.941255   0.8666768
## 2 Model 2      0.0      0.0 0.0000000    0.000000   0.0000000
## 3 Model 3      0.0      0.0 0.0000000    0.000000   0.0000000
## 4 Model 4      0.0      0.0 0.0000000    0.000000   0.0000000
```

```
# Binned Residual Plots
# par(mfrow=c(1,1))
# binnedplot(data.frame(newdata[3])$pred, data.frame(newdata[3])$pred-data.frame(newdata[3])$obs, col.p
# test <- dotplot(ranef(model_i, condVar=TRUE))
# class(test)
```
