Analysing Hierarchical Model for Course Project Data

The data of the project contains 5 factors:

Sex (2 levels) Race (4 levels) Age (5 levels) Education (3 levels) State (51 levels) The total number of combinations of levels is 6120. Look at the data. Check the variables notation:

```
dataPath <- 'C:/users/jasonlyn/Downloads/Bayesian/Final_Project'
dat<-read.csv(paste(dataPath,"MScA_32014_BayesianMethods_CourseProjectData.csv",sep="/"))
head(dat)
```

```
##      sex    race   age      education state y
## 1 1.Male 1.White 18-24    1.NoCollege    GA 0
## 2 1.Male 1.White 25-34    1.NoCollege    AZ 0
## 3 1.Male 1.White 25-34  2.SomeCollege    SD 0
## 4 1.Male 1.White 18-24 3.CollegeOrMore   SC 0
## 5 1.Male 1.White 18-24 3.CollegeOrMore   SC 0
## 6 1.Male 1.White 18-24 3.CollegeOrMore   SC 0
```

```
unique(dat$sex)
```

```
## [1] 1.Male   2.Female
## Levels: 1.Male 2.Female
```

```
unique(dat$race)
```

```
## [1] 1.White    2.Black    3.Hispanic 4.Other
## Levels: 1.White 2.Black 3.Hispanic 4.Other
```

```
unique(dat$age)
```

```
## [1] 18-24 25-34 35-44 45-54 55+
## Levels: 18-24 25-34 35-44 45-54 55+
```

```
unique(dat$education)
```

```
## [1] 1.NoCollege     2.SomeCollege   3.CollegeOrMore
## Levels: 1.NoCollege 2.SomeCollege 3.CollegeOrMore
```

```
unique(dat$state)
```

```
##  [1] GA AZ SD SC AL VA KS TN IA ME AR WA CT OH PA MA NH MD WI NE MS CA NY
## [24] DE MN MI ND ID HI IN VT FL OK UT NM KY LA WY DC RI IL OR NJ MT MO CO
## [47] NV WV AK NC TX
## 51 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY
```

After running OBAMA with the these data and the model obtain a Markov chain posterior date for 870 parameters including 2-way
Loading [MathJax]/jax/output/HTML-CSS/jax.js

interactions. Each Markov chain of the stan object obama_fit has length 36000.

Explore the fitted model object. It is not necessary to reproduce the results shown below, but if you see significant differences, please, report.

MODEL NOT RERUN (Saved model below):
##

model <- stan_model(file=paste(dataPath,"obama.complete.ext.stan",sep="/"))

obama_fit <- sampling(model, data=list(N = length(dat$y), y = dat$ y, sex = as.integer(dat$sex), NSex = nlevels(dat$ sex), race = as.integer(dat$race), NRace = nlevels(dat$ race), age = as.integer(dat$age), NAge = nlevels(dat$ age), education = as.integer(dat$education), NEducation = nlevels(dat$ education), state = as.integer(dat$state), NState = nlevels(dat$ state)), pars=c('b_0','b_sex', 'b_race', 'b_age', 'b_education', 'b_state', 'b_sex_race', 'b_sex_age', 'b_sex_education', 'b_sex_state', 'b_race_age', 'b_race_education', 'b_race_state', 'b_age_education', 'b_age_state', 'b_education_state', 'var_0', 'var_sex', 'var_race', 'var_age', 'var_education', 'var_state', 'var_sex_race', 'var_sex_age', 'var_sex_education', 'var_sex_state', 'var_race_age', 'var_race_education', 'var_race_state', 'var_age_education', 'var_age_state', 'var_education_state', 'nu', 'sigma'), control=list(adapt_delta=0.99, max_treedepth=12), iter=1000, chains = 4, cores = 4, verbose = F)

Load the fitted model object.

```
library(rstan)
```

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

```
## Loading required package: ggplot2
```

```
## 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)
```

```
## For improved execution time, we recommend calling
## Sys.setenv(LOCAL_CPPFLAGS = '-march=native')
## although this causes Stan to throw an error on a few processors.
```

```
library(HDInterval)
load(paste(dataPath,"fit_Obama.Rdata",sep="/"))
```

Launch shinystan() to explore the chains for convergence.

```
#library(shinystan)
#launch_shinystan(obama_fit)
```

Extract chains for further analysis.

```
OBAMA <-rstan::extract(obama_fit)
names(OBAMA)
```

```
##  [1] "b_0"                 "b_sex"               "b_race"
##  [4] "b_age"               "b_education"         "b_state"
##  [7] "b_sex_race"          "b_sex_age"           "b_sex_education"
## [10] "b_sex_state"         "b_race_age"          "b_race_education"
## [13] "b_race_state"        "b_age_education"     "b_age_state"
## [16] "b_education_state"   "var_0"               "var_sex"
## [19] "var_race"            "var_age"             "var_education"
## [22] "var_state"           "var_sex_race"        "var_sex_age"
## [25] "var_sex_education"   "var_sex_state"       "var_race_age"
## [28] "var_race_education"  "var_race_state"      "var_age_education"
## [31] "var_age_state"       "var_education_state" "nu"
## [34] "sigma"               "lp__"
```

Find parameters which are significantly different from zero: zero does not belong to 95% HDI. Show selected parameters.

```
sum.obama_fit<-rstan::summary(obama_fit)[[1]]
dim(sum.obama_fit)
```

```
## [1] 870  10
```

```
selection<-apply(sum.obama_fit[,c(4,8)],1,function(z) findInterval(0,z)!=1)
head(sum.obama_fit[selection,c(4,8)])
```

```
##                   2.5%       97.5%
## b_0        0.41169031  0.58327189
## b_sex[1]  -0.19051337 -0.05765717
## b_sex[2]   0.05765717  0.19051337
## b_race[1] -1.35940713 -1.17897800
## b_race[2]  2.11186703  2.50663040
## b_race[3] -0.64314864 -0.36492290
```

The model parameters are re-centered to satisfy additional constraint $\sum_j \beta_j = 0$, so, for example slopes of gender predictor are:

```
sum.obama_fit[2:3,1]
```

```
##   b_sex[1]   b_sex[2]
## -0.1214598  0.1214598
```

```
sum(sum.obama_fit[2:3,1])
```

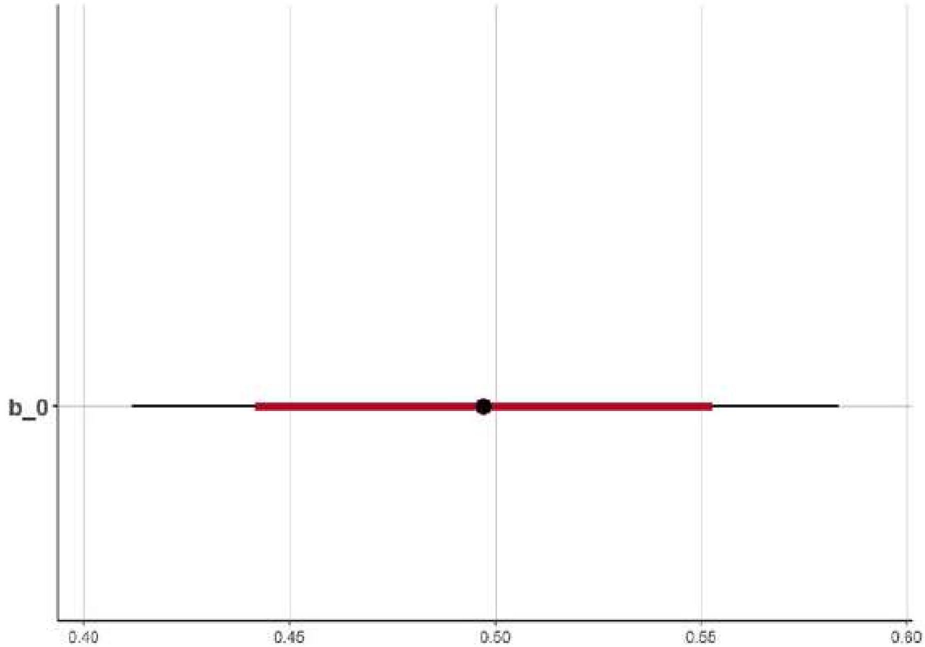
```
## [1] 1.387779e-17
```

As a result estimated intercept is close to 0.5:

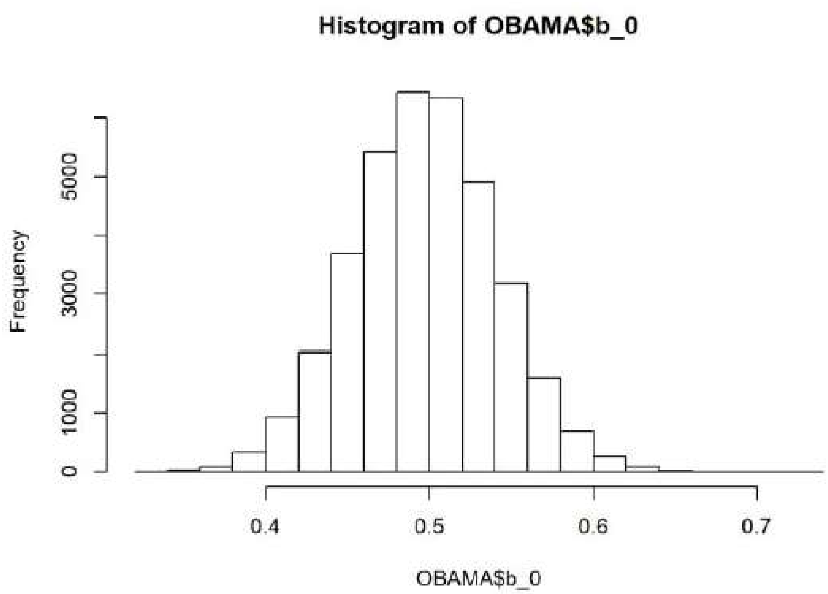
```
plot(obama_fit,pars=c("b_0"))
```

```
## ci_level: 0.8 (80% intervals)
```

```
## outer_level: 0.95 (95% intervals)
```




```
hist(OBAMA$b_0)
```




```
mean(OBAMA$b_0)
```

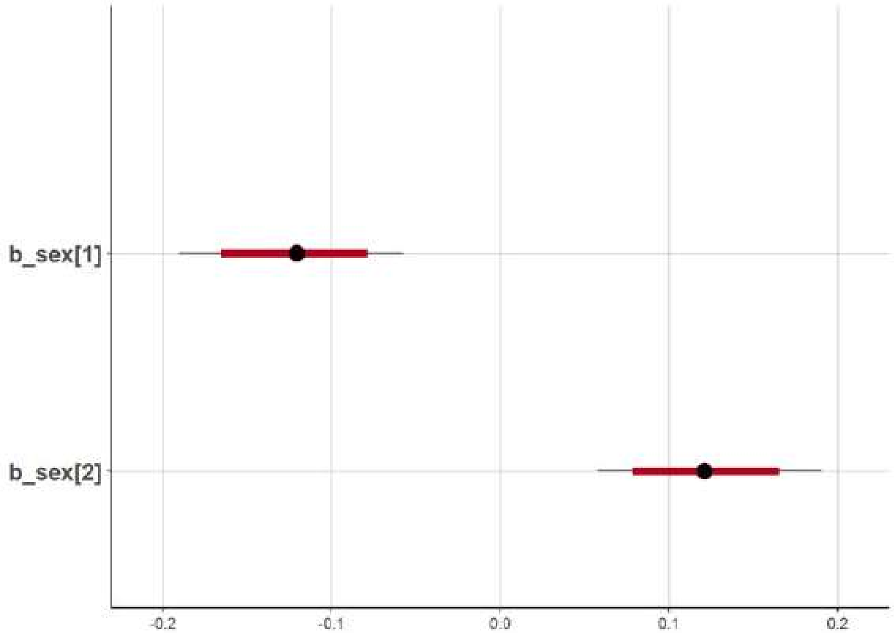
```
## [1] 0.4971177
```

Plot HDIs for the 5 main parameters. For example:

```
plot(obama_fit,pars=c("b_sex"))
```

```
## ci_level: 0.8 (80% intervals)
```

```
## outer_level: 0.95 (95% intervals)
```




```
sum.obama_fit[67:74,c(1,4,8)]
```

```
sum.obama_fit[67:74,c(1,4,8)]
```

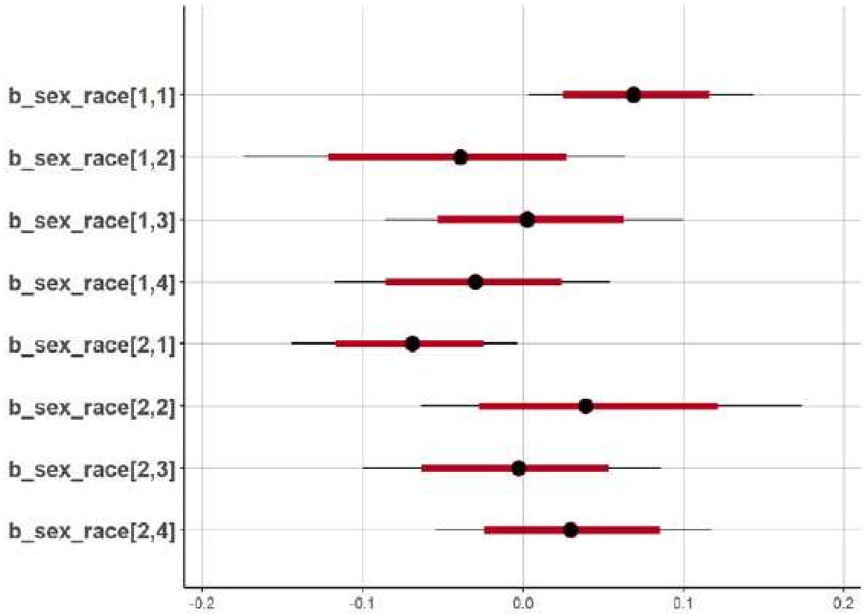
```
##                        mean         2.5%        97.5%
## b_sex_race[1,1]  0.07000647  0.003575282  0.143861731
## b_sex_race[1,2] -0.04379414 -0.174292010  0.063163381
## b_sex_race[1,3]  0.00392710 -0.085982318  0.099682650
## b_sex_race[1,4] -0.03013943 -0.117475813  0.054531941
## b_sex_race[2,1] -0.07000647 -0.143861731 -0.003575282
## b_sex_race[2,2]  0.04379414 -0.063163381  0.174292010
## b_sex_race[2,3] -0.00392710 -0.099682650  0.085982318
## b_sex_race[2,4]  0.03013943 -0.054531941  0.117475813
```

Interpretation of interactions. Look, for example, at interaction between sex and race.

```
plot(obama_fit,pars=c("b_sex_race"))
```

```
## ci_level: 0.8 (80% intervals)
```

```
## outer_level: 0.95 (95% intervals)
```



Two of the coefficients are significantly different from zero:

```
sum.obama_fit[c(67,70),c(1,4,8)]
```

```
##                         mean        2.5%      97.5%
## b_sex_race[1,1]  0.07000647  0.003575282 0.14386173
## b_sex_race[1,4] -0.03013943 -0.117475813 0.05453194
```

Recall that sex[1] is "Male" and sex[2] is "Female", and race[1] means "White" and race[4] means "Other". Main parameters and their corresponding interactions are then:

```
sum.obama_fit[c(1:4,7,67,70),c(1,4,8)]
```

```
##                         mean        2.5%       97.5%
## b_0              0.49711769  0.411690314  0.58327189
## b_sex[1]        -0.12145982 -0.190513365 -0.05765717
## b_sex[2]         0.12145982  0.057657169  0.19051337
## b_race[1]       -1.26909677 -1.359407127 -1.17897800
## b_race[4]       -0.53312966 -0.657803084 -0.40896376
## b_sex_race[1,1]  0.07000647  0.003575282  0.14386173
## b_sex_race[1,4] -0.03013943 -0.117475813  0.05453194
```

Interpret these parameters as influence over odds ratio.

```
(odds<-exp(sum.obama_fit[c(1:4,7,67,70),c(1,4,8)][,1]))
```

```
##             b_0        b_sex[1]        b_sex[2]       b_race[1]
##       1.6439760       0.8856266       1.1291440       0.2810854
##       b_race[4] b_sex_race[1,1] b_sex_race[1,4]
##       0.5867657       1.0725151       0.9703102
```

The baseline odds ratio is 1.643976, meaning that the ratio of people approving Obama as candidate to disapproving him is 1.643976. Among males the odds ratio is lower:

```
prod(odds[1:2])
```

```
## [1] 1.455949
```

And among females it is higher:

```
prod(odds[c(1,3)])
```

```
## [1] 1.856286
```

Approval odds for whites and others are:

```
prod(odds[c(1,4)])
```

```
## [1] 0.4620976
```

```
prod(odds[c(1,5)])
```

```
## [1] 0.9646288
```

Approval odds among white males (sex=1,race=1) is:

```
prod(odds[c(1,2,4,6)])
```

```
## [1] 0.4389225
```

Which is better than in the case of no interaction:

```
prod(odds[c(1,2,4)])
```

```
## [1] 0.409246
```

Note that since interaction b_sex_race[2,1] is not significantly different from zero, odds of approval among white women most likely is just

```
prod(odds[c(1,3,4)])
```

```
## [1] 0.5217748
```

Questions of the project:

Create Groups

```
names(OBAMA)
```

```
##  [1] "b_0"                 "b_sex"               "b_race"
##  [4] "b_age"               "b_education"         "b_state"
##  [7] "b_sex_race"          "b_sex_age"           "b_sex_education"
## [10] "b_sex_state"         "b_race_age"          "b_race_education"
## [13] "b_race_state"        "b_age_education"     "b_age_state"
## [16] "b_education_state"   "var_0"               "var_sex"
## [19] "var_race"            "var_age"             "var_education"
## [22] "var_state"           "var_sex_race"        "var_sex_age"
## [25] "var_sex_education"   "var_sex_state"       "var_race_age"
## [28] "var_race_education"  "var_race_state"      "var_age_education"
## [31] "var_age_state"       "var_education_state" "nu"
## [34] "sigma"               "lp__"
```

```
b0<-OBAMA$b_0
sex<-OBAMA$b_sex
colnames(sex)<-levels(dat$sex)

race<-OBAMA$b_race
colnames(race)<-levels(dat$race)

age<-OBAMA$b_age
colnames(age)<-levels(dat$age)

education<-OBAMA$b_education
colnames(education)<-levels(dat$education)

state<-OBAMA$b_state
colnames(state)<-levels(dat$state)
```

Gender Odds:

```
(Gender<-rbind(mean=exp(apply(sex,2,function(z) mean(z+b0))),exp(apply(sex,2,function(z) hdi(z+b0)))))
```

```
##         1.Male 2.Female
## mean  1.455949 1.856286
## lower 1.294588 1.679686
## upper 1.630791 2.050351
```

Race Odds:

```
(Race<-rbind(mean=exp(apply(race,2,function(z) mean(z+b0))),exp(apply(race,2,function(z) hdi(z+b0)))))
```

```
##         1.White  2.Black 3.Hispanic   4.Other
## mean  0.4620976 16.50647  0.9927362 0.9646288
## lower 0.4430622 12.84272  0.8443836 0.8432129
## upper 0.4824002 21.37102  1.1649948 1.1014265
```

Age Odds:

```
(Age<-rbind(mean=exp(apply(age,2,function(z) mean(z+b0))),exp(apply(age,2,function(z) hdi(z+b0)))))
```

```
##          18-24    25-34    35-44    45-54      55+
## mean  2.103702 1.866141 1.605320 1.523512 1.250665
## lower 1.786283 1.616481 1.366515 1.308444 1.081427
## upper 2.474256 2.164906 1.871676 1.775774 1.437299
```

Education Odds:

```
(Education<-rbind(mean=exp(apply(education,2,function(z) mean(z+b0))),exp(apply(education,2,function(z) hdi(z
+b0)))))
```

```
##       1.NoCollege 2.SomeCollege 3.CollegeOrMore
## mean     1.515840      1.673248        1.751752
## lower    1.333525      1.504955        1.546969
## upper    1.732132      1.871543        1.995538
```

State Odds:

```
(State<-rbind(mean=exp(apply(state,2,function(z) mean(z+b0))),exp(apply(state,2,function(z) hdi(z+b0)))))
```

```
##               AK        AL        AR       AZ       CA       CO       CT
## mean  1.3476856 0.9696427 1.1393769 1.690446 1.985228 1.874901 2.077820
## lower 0.8827718 0.6552689 0.7871296 1.194617 1.484320 1.325678 1.459179
## upper 2.0098714 1.4158702 1.6215941 2.352537 2.638478 2.652787 3.003641
##              DC       DE       FL        GA       HI       IA        ID
## mean  3.604544 2.216578 1.725519 1.2345273 3.129431 2.238279 1.3083928
## lower 2.248595 1.473026 1.224880 0.9215636 2.197365 1.516694 0.8760142
## upper 5.771946 3.329536 2.365356 1.6550056 4.438253 3.290229 1.9734203
##              IL        IN        KS       KY        LA       MA       MD
## mean  2.628237 1.0519490 1.0881749 1.619371 0.9720502 2.474350 1.884324
## lower 1.900043 0.7188661 0.7387039 1.095936 0.6975811 1.729469 1.377507
## upper 3.733680 1.5067773 1.5541253 2.358757 1.3587647 3.572645 2.570738
##              ME       MI       MN        MO        MS        MT        NC
## mean  1.889644 1.978596 2.643456 1.4260513 0.9628397 1.1186730 1.3487614
## lower 1.238782 1.358984 1.773180 0.9808641 0.6519408 0.7306527 0.9542182
## upper 2.893011 2.850413 3.910435 2.0746917 1.4208505 1.6911993 1.8963021
##              ND       NE        NH       NJ       NM        NV       NY
## mean  1.567775 1.751783 1.4694228 2.291579 1.604123 1.3754591 2.165378
## lower 1.011873 1.178856 0.9331125 1.669673 1.144317 0.9754387 1.548402
## upper 2.428760 2.603631 2.2804704 3.239776 2.254516 1.9150176 3.005885
##              OH       OK       OR       PA       RI       SC       SD
## mean  2.181995 1.136575 1.554238 2.120849 2.194671 1.233403 1.660243
## lower 1.472890 0.805980 1.080299 1.446736 1.432160 0.974229 1.065935
## upper 3.180823 1.603605 2.262743 3.112789 3.378892 1.571722 2.547570
##               TN        TX        UT       VA       VT       WA       WI
## mean  1.0479435 1.1781476 1.2767905 1.743564 3.071641 2.487922 1.809932
## lower 0.7280519 0.8737884 0.8734067 1.284176 1.821589 1.767373 1.216294
## upper 1.4820015 1.5814508 1.8921464 2.392000 5.118829 3.582052 2.645143
##               WV        WY
## mean  1.0838929 1.1162077
## lower 0.7094964 0.7155686
## upper 1.6773978 1.6997658
```

1. Find groups from which the main support for Obama came in 2012

```
c(max(Gender[1,]),max(Race[1,]),max(Age[1,]),max(Education[1,]),max(State[1,]))
```

```
## [1]  1.856286 16.506465  2.103702  1.751752  3.604544
```

```
data.frame(Category= c('Sex', 'Race', 'Age', 'Education', 'State'), Highest= c('Female', 'Black', '25-34', 'C
olleageOrMore', 'DC'))
```

```
##    Category       Highest
## 1       Sex        Female
## 2      Race         Black
## 3       Age         25-34
## 4 Education ColleageOrMore
## 5     State            DC
```

2. Find groups of the lowest odds of approval

```
c(min(Gender[1,]),min(Race[1,]),min(Age[1,]),min(Education[1,]),min(State[1,]))
```

```
## [1] 1.4559489 0.4620976 1.2506649 1.5158399 0.9628397
```

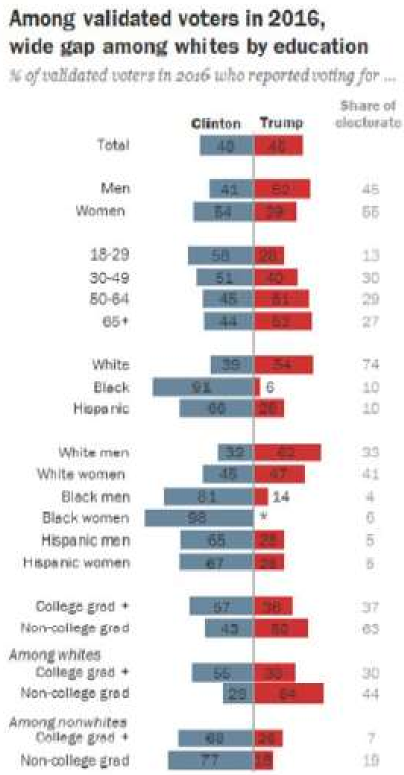
```
data.frame(Category= c('Sex', 'Race', 'Age', 'Education', 'State'), Lowest= c('Male', 'White', '55+', 'NoColl
ege', 'MS'))
```

Loading [MathJax]/jax/output/HTML-CSS/jax.js

```
##    Category    Lowest
## 1       Sex      Male
## 2      Race     White
## 3       Age       55+
## 4 Education NoCollege
## 5     State        MS
```

3. Search for information on main support and no support for Hillary Clinton in 2016 and try to identify the dynamics between 2012 and 2016

Note, raw data is not included in the project. The chart below is from PEW Research Center.



For 2016 Hillary Clinton election:

Main Support: (1) Female (2) Age Group 18-29 (3) Black (4) Collage Graduate (5) DC

No Support : (1) Male (2) Age Group 65+ (3) White (4) Non-College Graduate (5) West Virgina Note: The state results are from the actual election results in 2016.

The overall dynamic is almost identical to the 2012 results. It seems like voters who were strongly likely to vote for Obama would vote for Hillary as well.

4. What else you find interesting in the results?

Based on the 2012 and 2016 results, it seems like the voters are likely to support their party's nominee. Democrats chose a Black Male and White Female, respectively. The same voters who supported the Democratic candidate in 2012 supported the the Democratic candidate in 2016. There is no surprise there. Since the US Presidential Election is determined by "swing" states (and swing districts), we may want to focus the results on swing states (Colorado, Florida, Iowa, Michigan, Minnesota, Ohio, Nevada, New Hampshire, North Carolina, Pennsylvania, Virginia, and Wisconsin). If you examine the 4 states where Obama won and Hillary lost, the dynamics/margin are very close among voters.