1992 U.S. Presidential election

Ali Seada and Paul Lovis Maximilian Trüstedt

24 6 2021

Read the data into R environment

```
library(pacman)
p_load(ggplot2,   # reportable graphs
       cowplot,   # arranges ggplot graphs nicely
       stargazer, # nice tables
       glmnet,    # for regularization (lasso, ridge, elastic net)
       pROC)      # ROC AUC
rm(list=ls())
vote<-read.csv("vote92.csv")
str(vote)
```

```
## 'data.frame':    909 obs. of  10 variables:
##  $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ vote       : Factor w/ 3 levels "Bush","Clinton",..: 1 1 2 1 2 2 3 1 1 3 ...
##  $ dem        : int  0 0 1 0 0 1 1 0 0 0 ...
##  $ rep        : int  1 1 0 1 0 0 0 1 1 1 ...
##  $ female     : int  1 1 1 0 1 1 1 0 1 0 ...
##  $ persfinance: int  1 0 0 0 0 -1 1 0 1 0 ...
##  $ natlecon   : int  0 -1 -1 -1 -1 -1 0 0 -1 0 ...
##  $ clintondis : num  4.0804 4.0804 1.0404 0.0004 0.9604 ...
##  $ bushdis    : num  0.102 0.102 1.742 5.382 11.022 ...
##  $ perotdis   : num  0.26 0.26 0.24 2.22 6.2 ...
```

```
summary(vote)
```

```
##        X            vote          dem              rep             female      
##  Min.   :  1   Bush   :310   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:228   Clinton:416   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :455   Perot  :183   Median :0.0000   Median :0.0000   Median :0.0000  
##  Mean   :455                 Mean   :0.4884   Mean   :0.4301   Mean   :0.4752  
##  3rd Qu.:682                 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :909                 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##   persfinance          natlecon         clintondis         bushdis       
##  Min.   :-1.000000   Min.   :-1.0000   Min.   : 0.0004   Min.   : 0.1024  
##  1st Qu.:-1.000000   1st Qu.:-1.0000   1st Qu.: 0.9604   1st Qu.: 0.4624  
##  Median : 0.000000   Median :-1.0000   Median : 1.0404   Median : 1.7424  
##  Mean   :-0.009901   Mean   :-0.6722   Mean   : 3.5062   Mean   : 3.3793  
##  3rd Qu.: 1.000000   3rd Qu.: 0.0000   3rd Qu.: 4.0804   3rd Qu.: 5.3824  
##  Max.   : 1.000000   Max.   : 1.0000   Max.   :16.1600   Max.   :18.6620  
##     perotdis      
##  Min.   : 0.2401  
##  1st Qu.: 0.2401
```

```
##  Median : 2.2201
##  Mean   : 2.1710
##  3rd Qu.: 2.2801
##  Max.   :12.1800
```

Preprocess the data, preparing it for the modeling

```
vote$dem <-as.factor(vote$dem)
vote$rep <-as.factor(vote$rep)
vote$female <-as.factor(vote$female)
vote$persfinance <-as.factor(vote$persfinance)
vote$natlecon <-as.factor(vote$natlecon)
# ??should be numeric by definition in doc LETS SEE
str(vote)
```

```
## 'data.frame':    909 obs. of  10 variables:
##  $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ vote       : Factor w/ 3 levels "Bush","Clinton",..: 1 1 2 1 2 2 3 1 1 3 ...
##  $ dem        : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 2 1 1 1 ...
##  $ rep        : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 1 2 2 2 ...
##  $ female     : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 2 1 2 1 ...
##  $ persfinance: Factor w/ 3 levels "-1","0","1": 3 2 2 2 2 1 3 2 3 2 ...
##  $ natlecon   : Factor w/ 3 levels "-1","0","1": 2 1 1 1 1 1 2 2 1 2 ...
##  $ clintondis : num  4.0804 4.0804 1.0404 0.0004 0.9604 ...
##  $ bushdis    : num  0.102 0.102 1.742 5.382 11.022 ...
##  $ perotdis   : num  0.26 0.26 0.24 2.22 6.2 ...
```

We decided to change some of the numeric variables to factors, because it makes more sense to have them as factors than as numeric variables.

- treat missing values

```
vote[vote==-999|vote==""]<-NA
colSums(is.na(vote))
```

```
##           X        vote         dem         rep      female persfinance
##           0           0           0           0           0           0
##    natlecon  clintondis     bushdis    perotdis
##           0           0           0           0
```

There are no missing values in this data set

- handle sparse classes of categorical predictors

```
table(vote$vote)
```

```
##
##    Bush Clinton   Perot
##     310     416     183
```

```
table(vote$dem)
```

```
##
##   0   1
## 465 444
```

```
table(vote$rep)
```

```
##
```

```
##   0   1
## 518 391
```

```
table(vote$female)
```

```
##
##   0   1
## 477 432
```

```
table(vote$persfinance)
```

```
##
##  -1   0   1
## 308 302 299
```

```
table(vote$natlecon) # ?? leave fusion of 0 and 1, technically sparse
```

```
##
##  -1   0   1
## 656 208  45
```

```
vote$natlecon[vote$natlecon==1]<-0
vote$natlecon[vote$natlecon==-1]<-1
vote$natlecon=droplevels(vote$natlecon)
table(vote$natlecon)
```

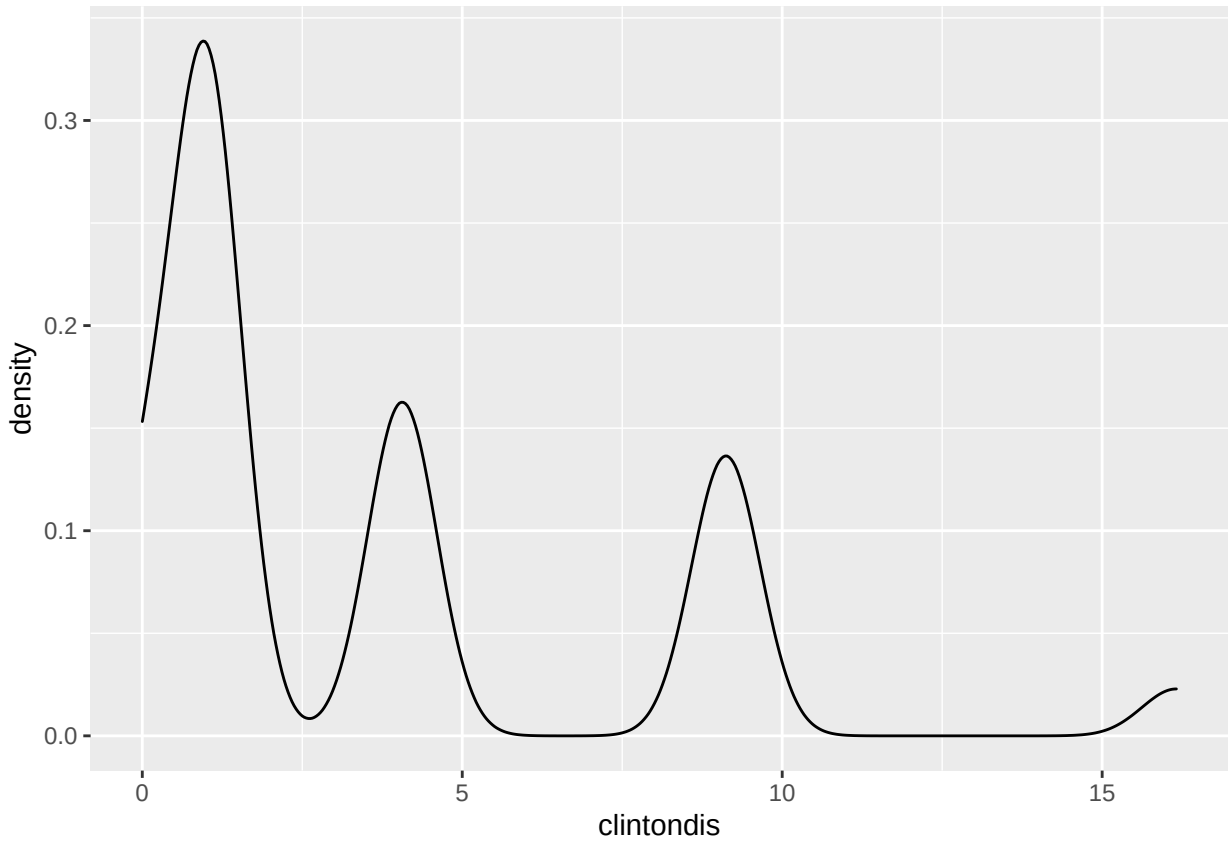
```
##
##   0   1
## 253 656
```

X The categorical variables are simplified enough and don't need anything else to be done to them. OR IF UPDATE LEFT O We leave everything as is except for natlecon which has a sparse class regarding the level 1. As solution we combine 0 and 1 as the level 0, meaning national economic conditions have gotten better or stayed the same over the last 12 months. Level -1 gets changed to 1 as well which now means that conditions have gone better.
The change from -1 to 1 is executed just because it is more common to have levels 0 and 1 instead of 0 and -1.

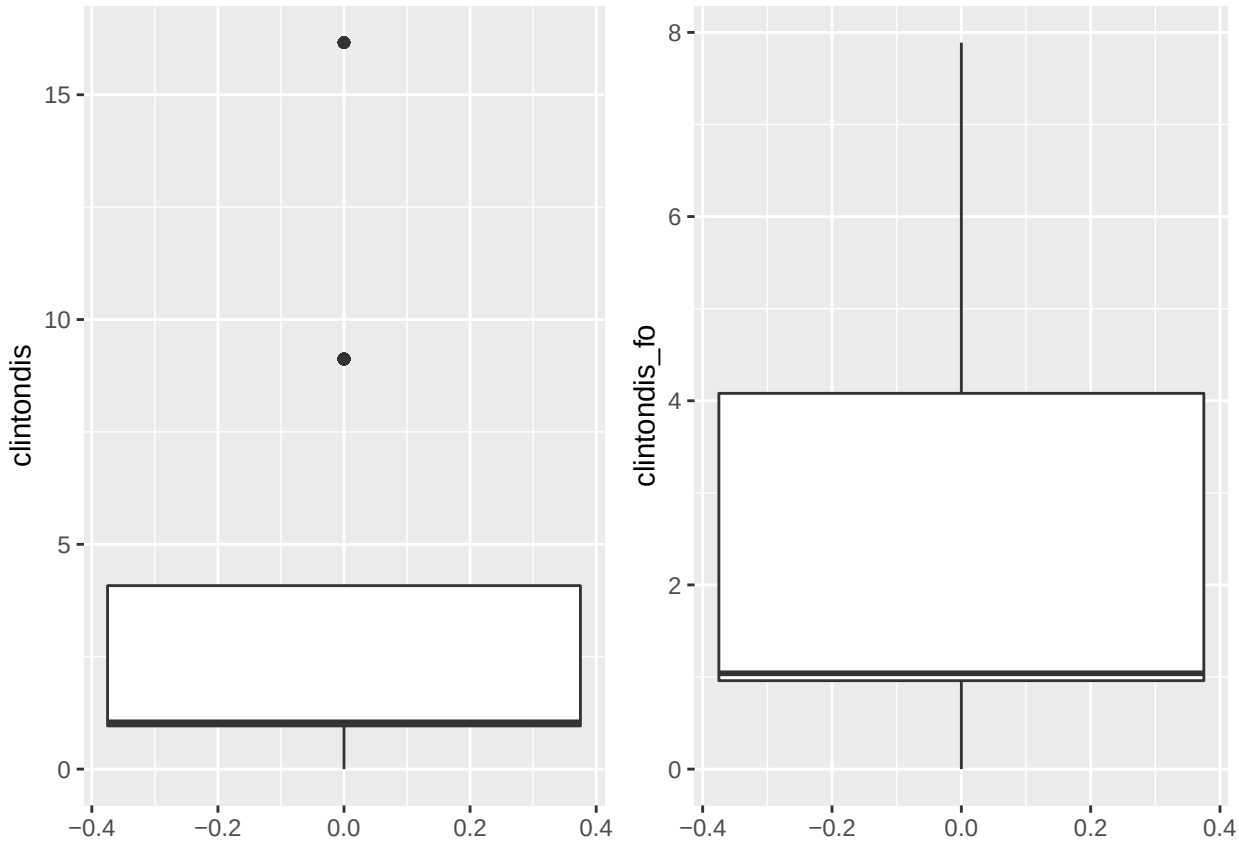
- take care of outliers, treat the skewed distributions and create new features

```
zScores <- function(var){
    mu <- mean(var)
    sd <-  sd(var)
    scores <- (var - mu)/sd
    return(scores)
}

# treating clintondis
ggplot(vote,aes(clintondis))+geom_density() # ??is this considered skewed; ??also very common
```



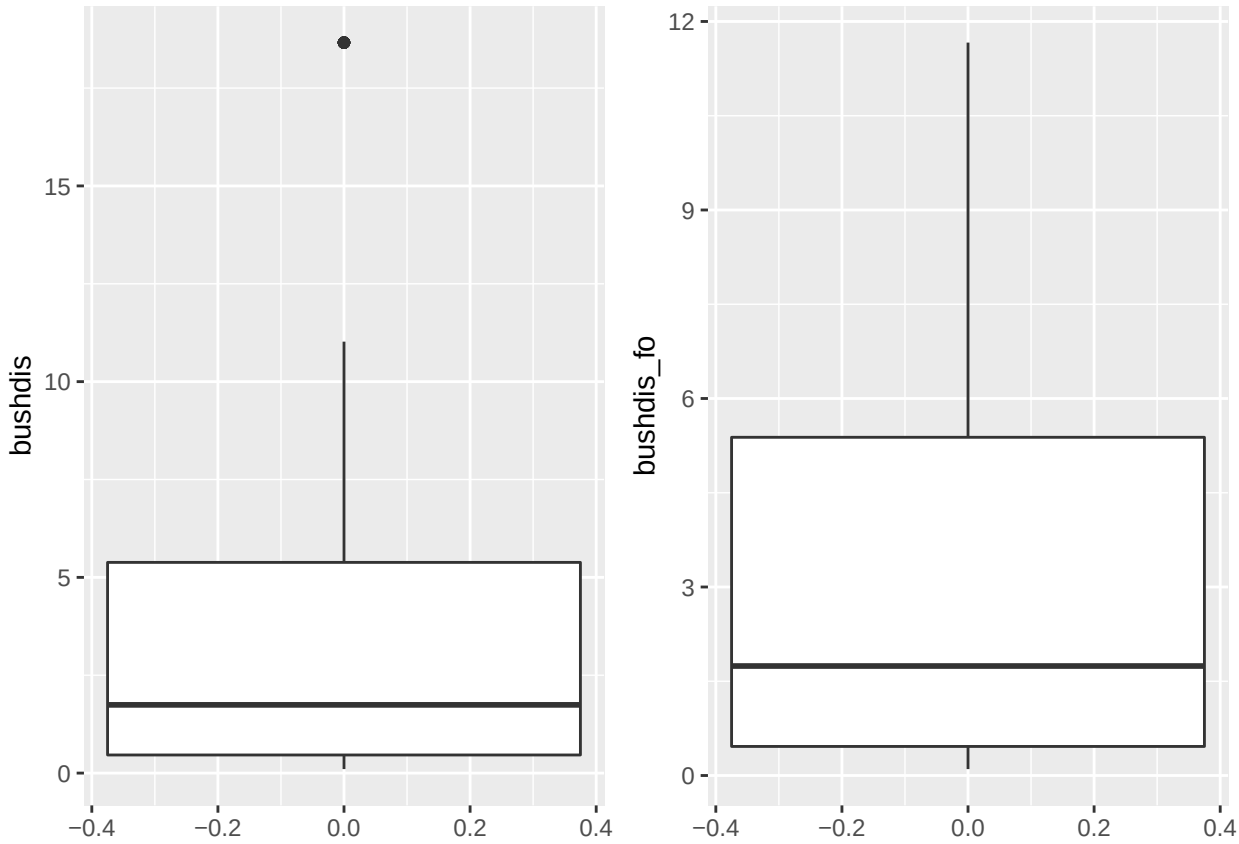

```
tp1<-ggplot(vote,aes(clintondis))+geom_boxplot()+coord_flip()
vote$clintondis_fo<-vote$clintondis
vote$clintondis_fo[zScores(vote$clintondis_fo)>1]<-
    round(mean(vote$clintondis_fo))+sd(vote$clintondis_fo)
tp2<-ggplot(vote,aes(clintondis_fo))+geom_boxplot()+coord_flip()
plot_grid(tp1,tp2,ncol=2)
```




```
# !!make this shit beautiful
# !!reduce the outlier fixes with function
# ??passable results
```

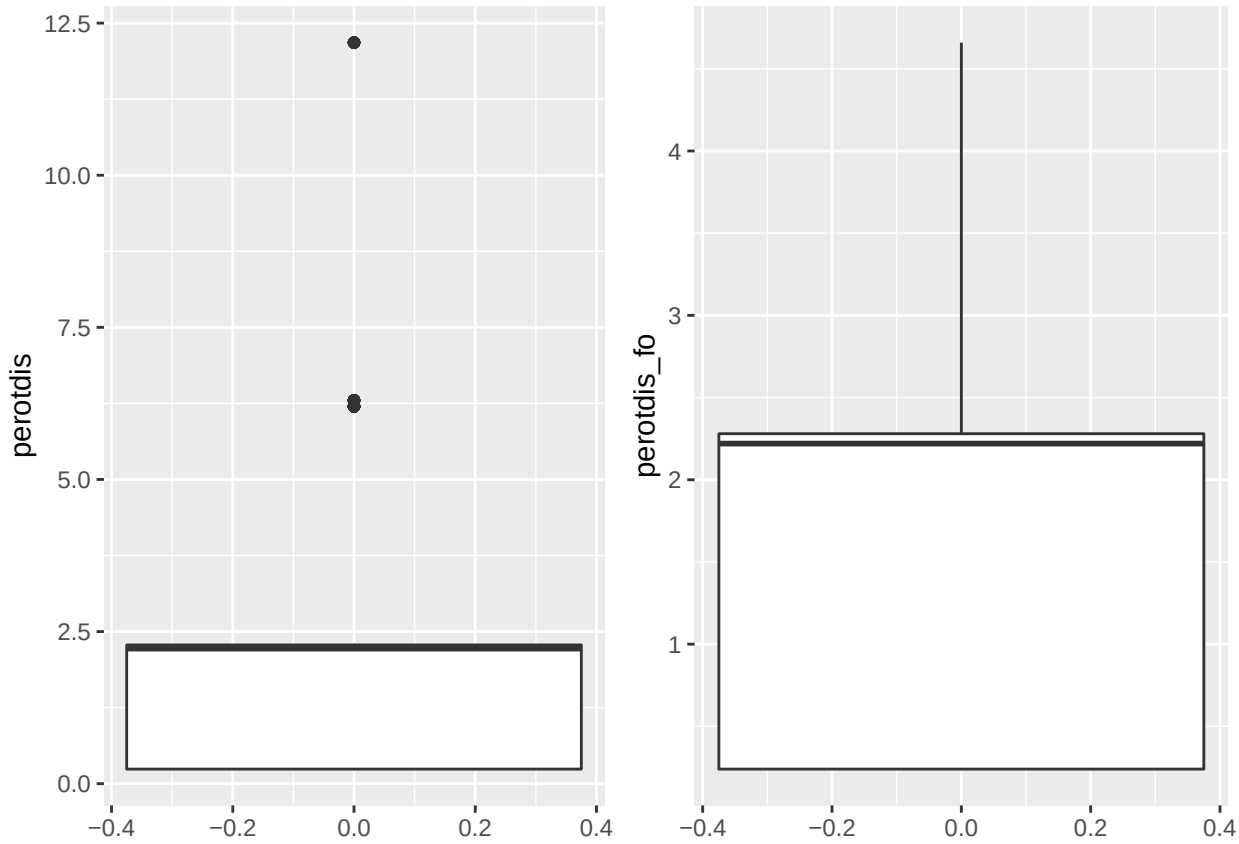
INSERT

```
# treating bushdis
tp1<-ggplot(vote,aes(bushdis))+geom_boxplot()+coord_flip()
vote$bushdis_fo<-vote$bushdis
vote$bushdis_fo[zScores(vote$bushdis_fo)>2]<-
    round(mean(vote$bushdis_fo))+2*sd(vote$bushdis_fo)
tp2<-ggplot(vote,aes(bushdis_fo))+geom_boxplot()+coord_flip()
plot_grid(tp1,tp2,ncol=2)
```



INSERT

```
# treating perotdis
tp1<-ggplot(vote,aes(perotdis))+geom_boxplot()+coord_flip()
vote$perotdis_fo<-vote$perotdis
vote$perotdis_fo[zScores(vote$perotdis_fo)>1]<-
    round(mean(vote$perotdis_fo))+sd(vote$perotdis_fo)
tp2<-ggplot(vote,aes(perotdis_fo))+geom_boxplot()+coord_flip()
plot_grid(tp1,tp2,ncol=2)
```



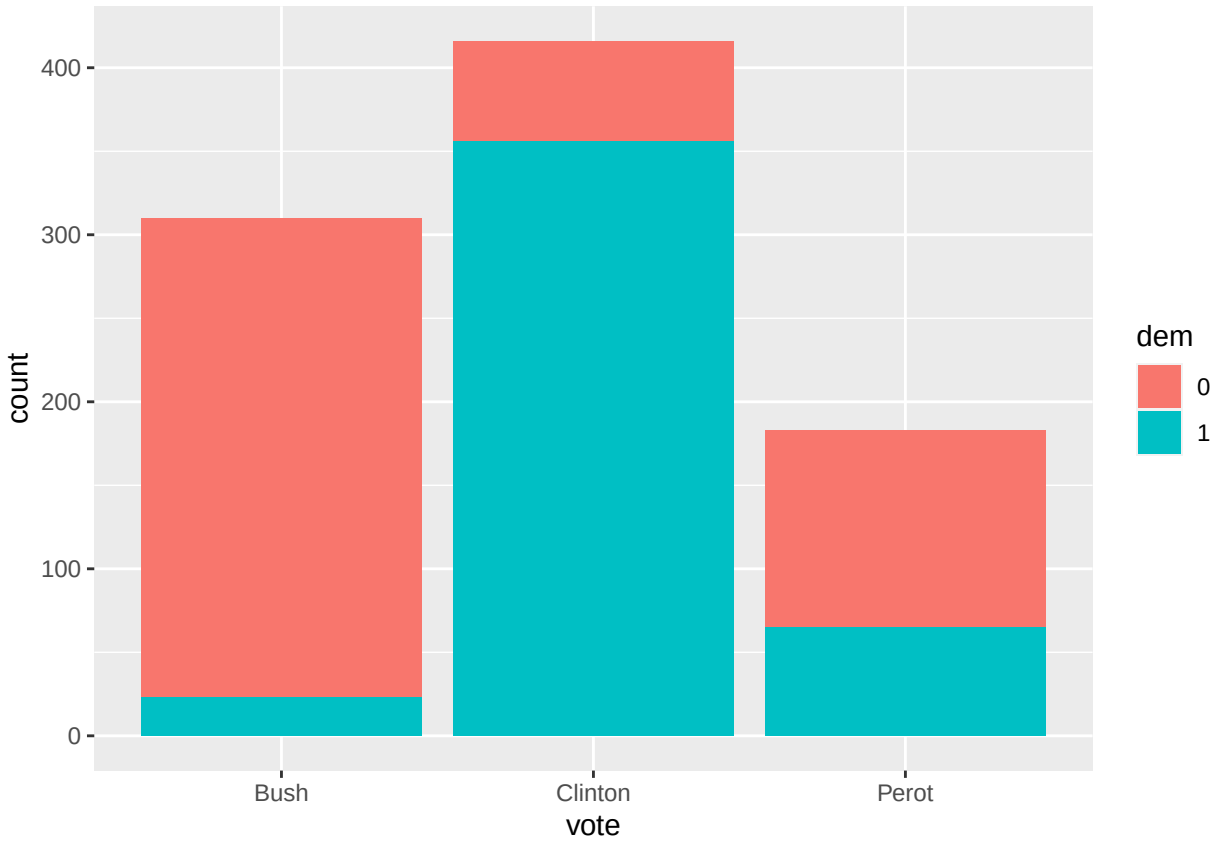
INSERT

- explore the relationships between predictors and the target

```
tcs<-vote$dem
tcs[tcs==1]<-vote$rep[vote$rep==1] # $$tiny prototype of future way of presentation
```

```
## Warning in x[...] <- m: Anzahl der zu ersetzenden Elemente ist kein Vielfaches
## der Ersetzungslänge
```

```
ggplot(vote,aes(vote,fill=dem))+geom_bar()
```



INSERT