1992 U.S. Presidential election

Ali Tarek Maher Ibrahim Ali Seada and Paul Lovis Maximilian Trüstedt

24 6 2021

Read the data into R environment

```
## '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 ...
```

```
##        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
```

This data set is about the 1992 USA presidential elections. In the data set we have 909 rows/observations with 10 variables. We only have 1 Factor, vote that has the levels Bush, Clinton and Perot. This determines what candidate the respondent voted for in the 92 election.
There are also 5 integer variables: dem, rep, female, persfinance and natlecon. The first two are 1 if the respondent identified with the democratic (dem) or the republican (rep) party. The variable female is 1 if the voter in the data set was female. While persfinance describes the change in personal wealth of the voter, natlecon is how the voter viewed the change in the national economy. Those variables are -1 if there was a negative change, 0 if there was no change and 1 if the financial or economic change was conceived as positive.

The rest of the variables (which are 3) are numeric: clintondis, bushdis and perotdis. Those determine the squared ideological distance of the respondent from the respective party.

Preprocess the data, preparing it for the modeling

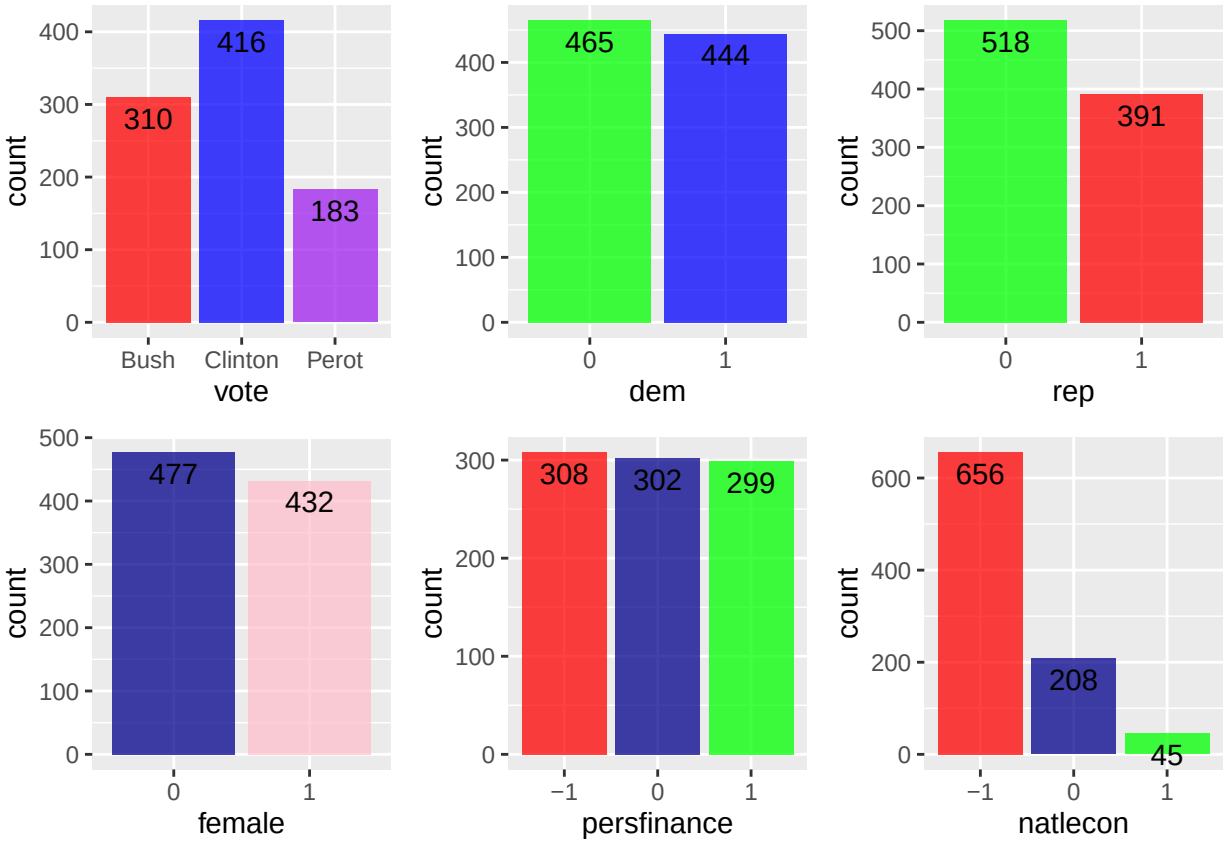
```
## 'data.frame':    909 obs. of  12 variables:
##  $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ vote       : Factor w/ 3 levels "Bush","Clinton",..: 1 1 2 1 2 2 3 1 1 3 ...
##  $ dem        : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 2 1 1 1 ...
##  $ rep        : Factor w/ 2 levels "0","1": 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 ...
##  $ vote_num   : num  1 1 2 1 2 2 3 1 1 3 ...
##  $ polID      : Factor w/ 3 levels "1","2","3": 3 3 2 3 1 2 2 3 3 3 ...
```

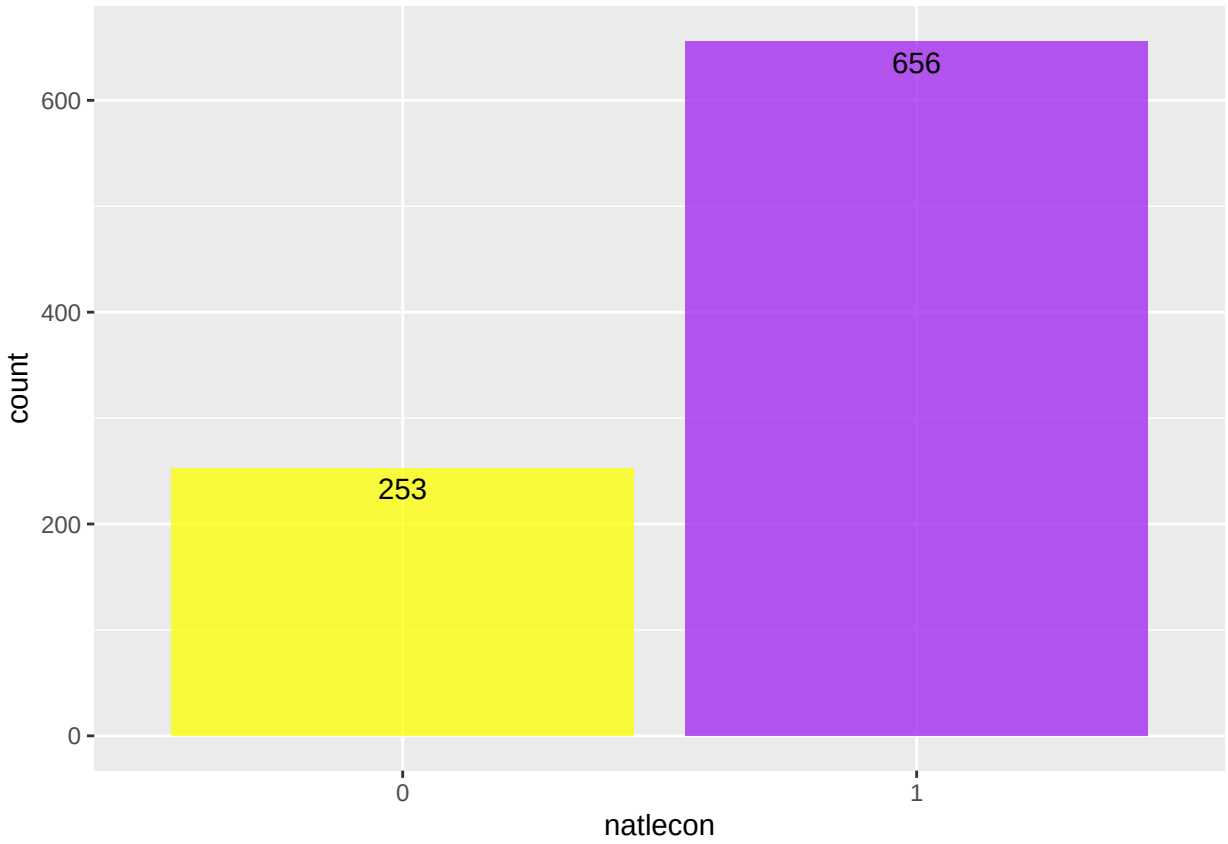
- treat missing values

```
colSums(is.na(vote))
```

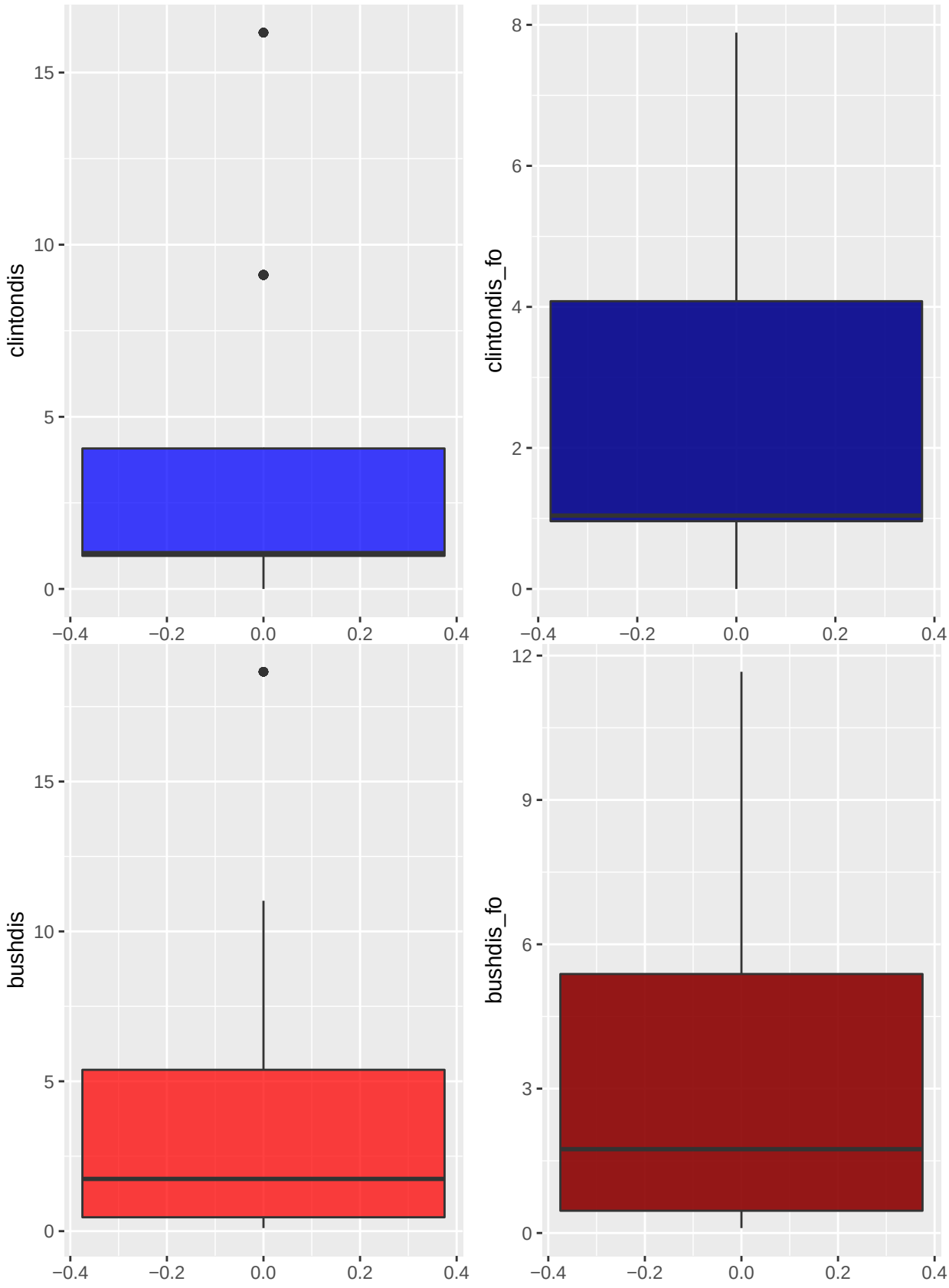
```
##           X        vote         dem         rep      female persfinance
##           0           0           0           0           0           0
##    natlecon  clintondis     bushdis    perotdis    vote_num       polID
##           0           0           0           0           0         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

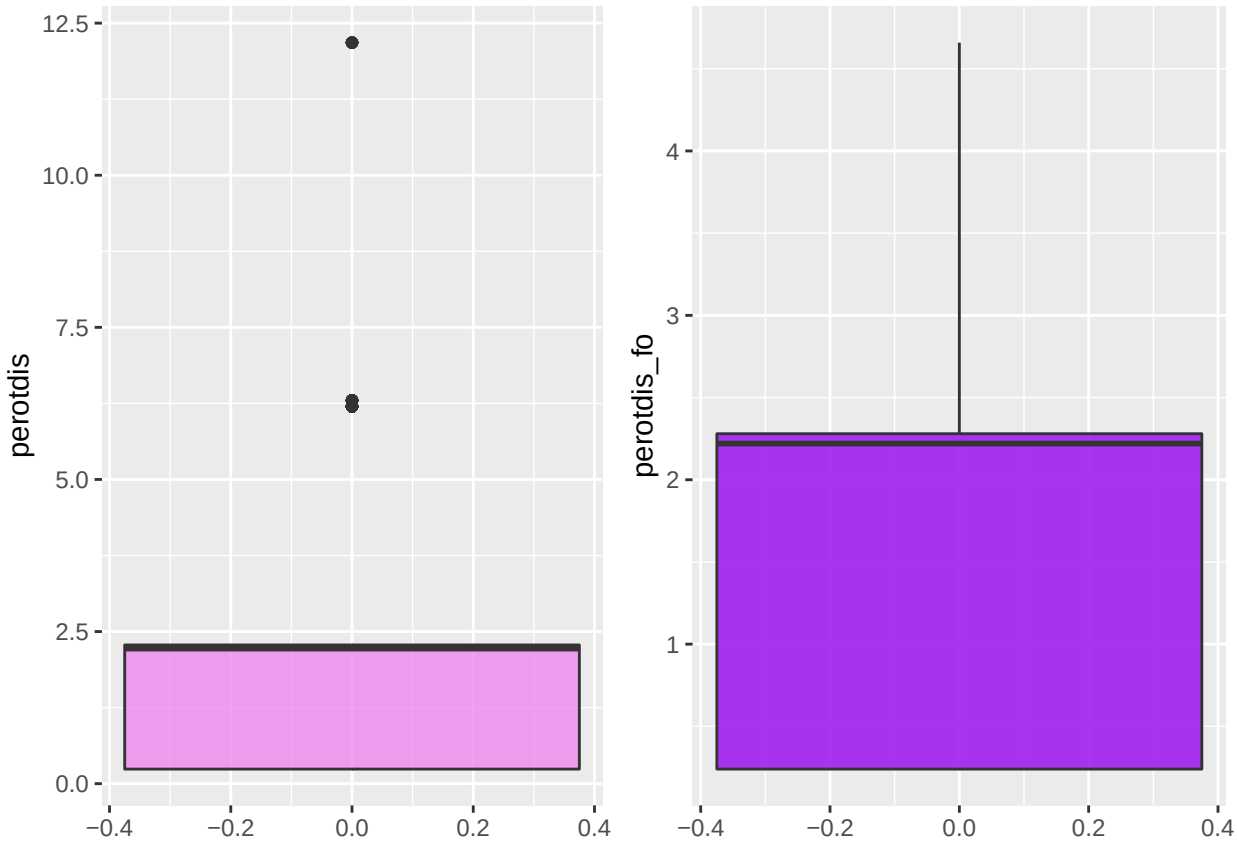
- handle sparse classes of categorical predictors





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

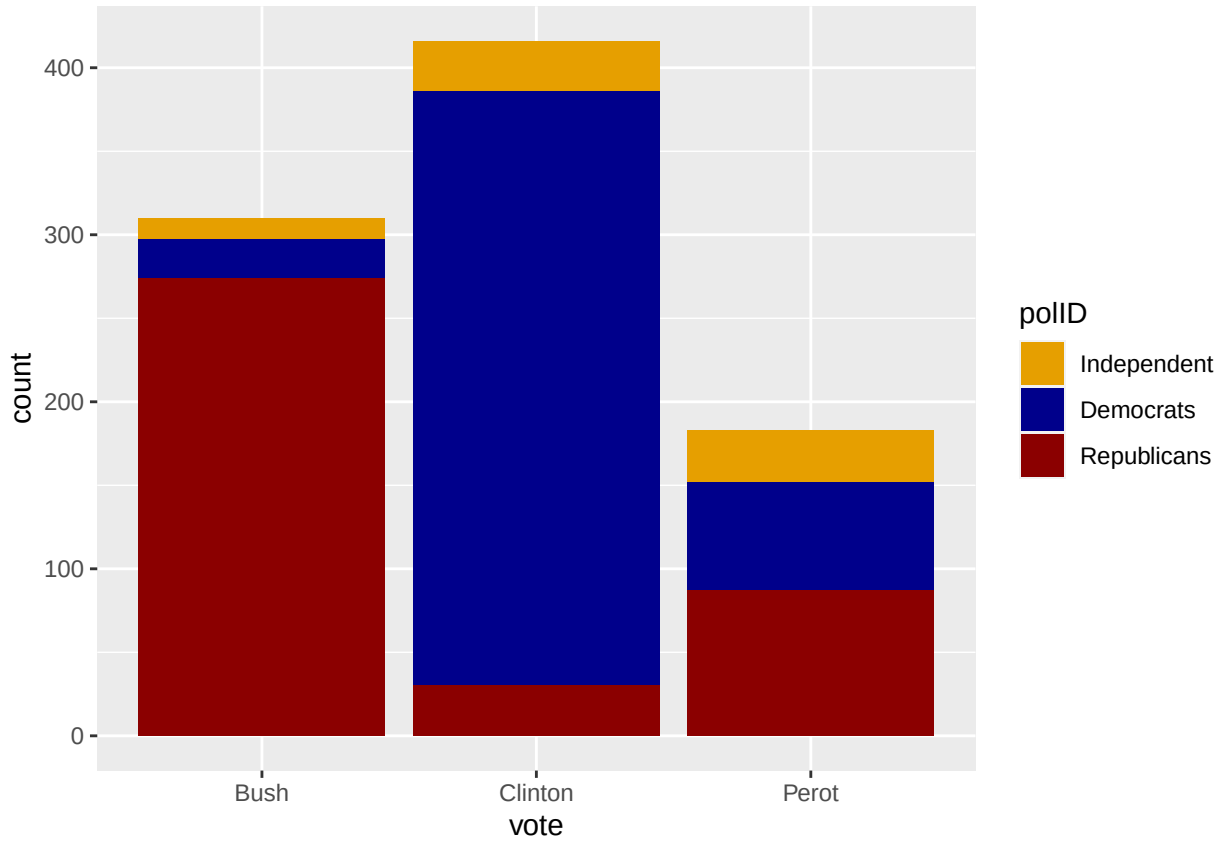

clintondis
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4
clintondis_fo
8
6
4
2
0
bushdis
bushdis_fo
12
9
3

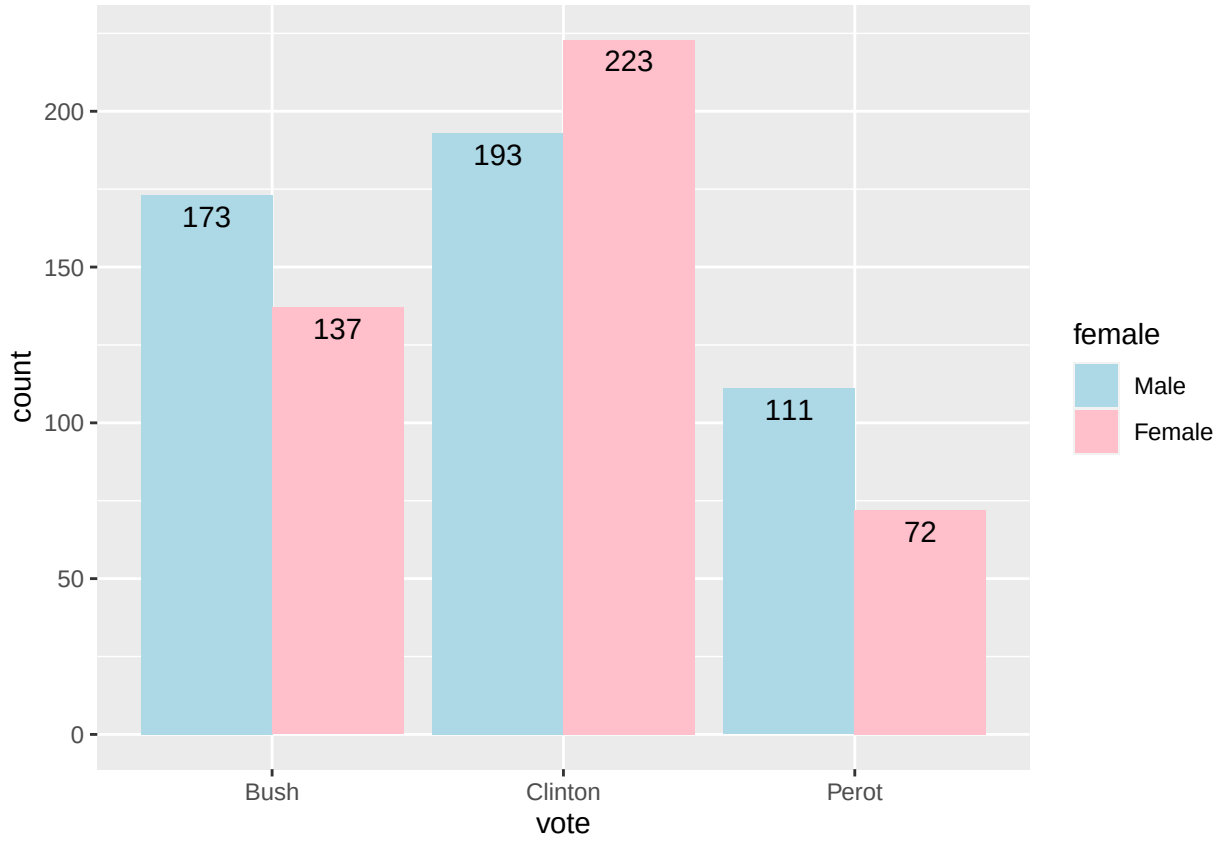


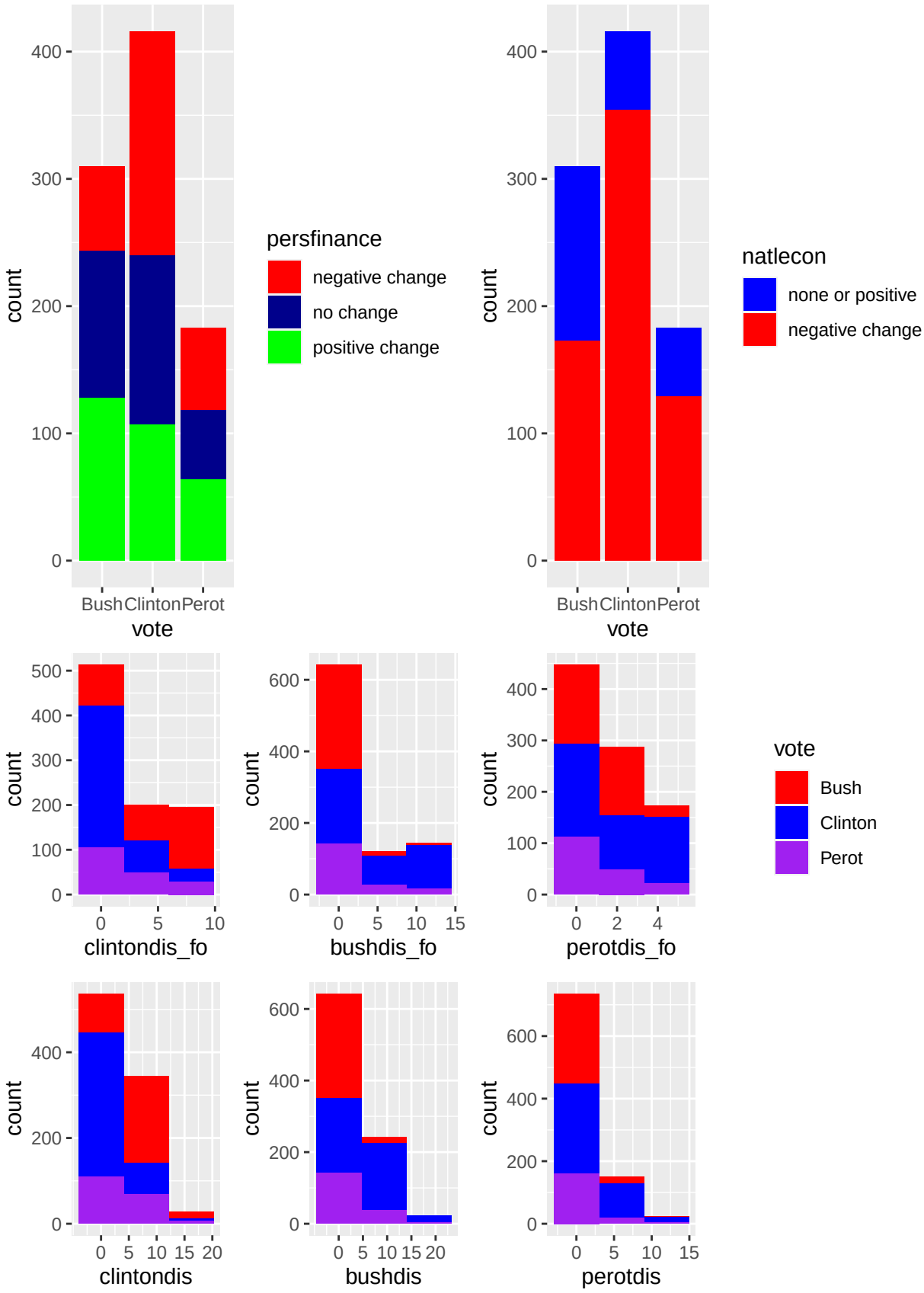
In the beginning we decided to change some of the numeric variables to factors, because it makes more sense to have them as categorical than as numeric variables. Also this way, we can see, that there are no problems with the categorical variables regarding wrong values, because all provided levels are described by the given data set definition. Additionally we created a categorical variable called polID to summarize which political party the respondent is identifying himself with. Afterwards we decided to start looking at the missing values and treating them, but we couldn't find any explicit or implicit missing values. Looking at the categorical values we only decided to change natlecon since 1 was a sparse class. We combined 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 was changed to 1 as well which now means that conditions have gotten 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.

Next step was taking care of the outliers, treating skewed distributions and creating new features. There are a few outliers in the variables clintondis, bushdis and perotdis. We fixed those outliers and saved the fixed data within variables called [original_var_name]_fo. The ending "fo" is derived from "fixed outliers". Moving forward we didn't find any other problems regarding the data set so we decided to move on.

- explore the relationships between predictors and the target


400
300
200
100
0
count
Bush
Clinton
Perot
vote
pollD
Independent
Democrats
Republicans


200
150
100
50
0
count
173
137
193
223
111
72
Bush
Clinton
Perot
vote
female
Male
Female


count
vote
Bush
Clinton
Perot
persfinance
negative change
no change
positive change
natlecon
none or positive
negative change
clintondis_fo
bushdis_fo
perotdis_fo
clintondis
bushdis
perotdis

Next, we decided to look at the relationships between the predictors and the target, here our target variable being vote. Looking at the votes for Perot, most prominently visible is, that half of his votes came from republican voters, which Bush lost. Also Bush got the least votes from voters who didn't clearly align with either the democratic or the republican party. Those undecided voters, that didn't vote for Bush about equally voted for Clinton and Perot. Not only did more of them vote for Clinton instead of Bush, but also more democrats voted for Clinton than republicans for Bush. To top it off, even more republicans voted for the democratic as for the republican party, though it must be noted that more respondents reported aligning with the democratic party in the first place.

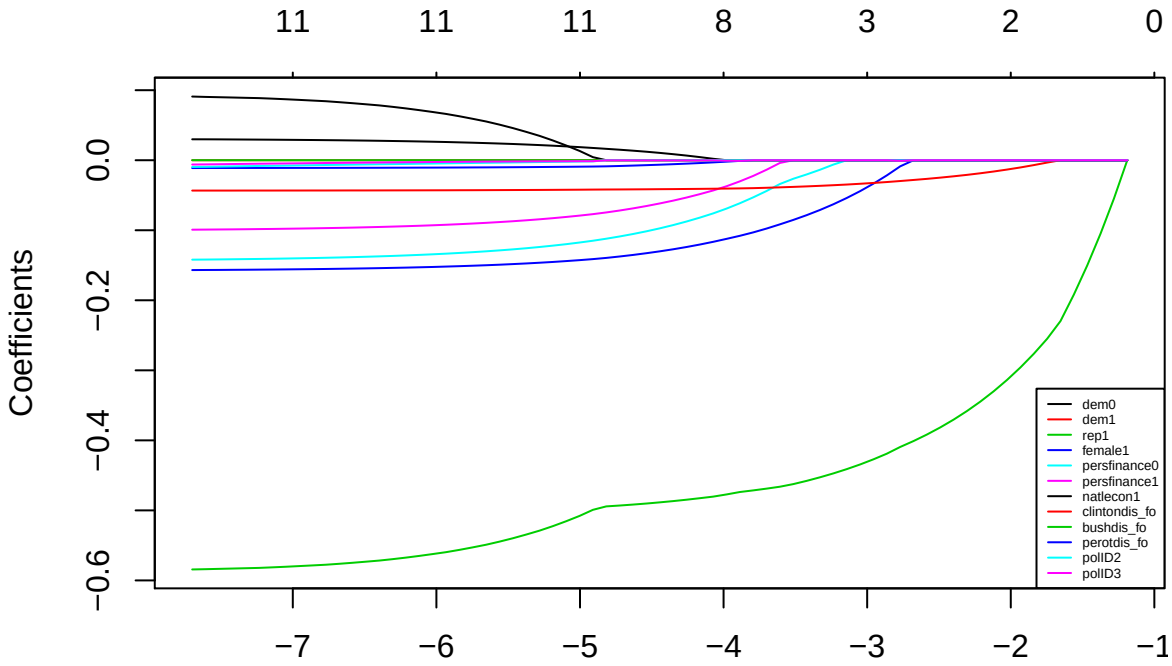
Both Bush and Perot had significantly more male than female voters, though that could also be, because more respondents in this data set are male than female. Although on the one hand Bush and Perot had more male than female voters, Clinton on the other hand received more female than male voters regardless of the gender bias of the data set.

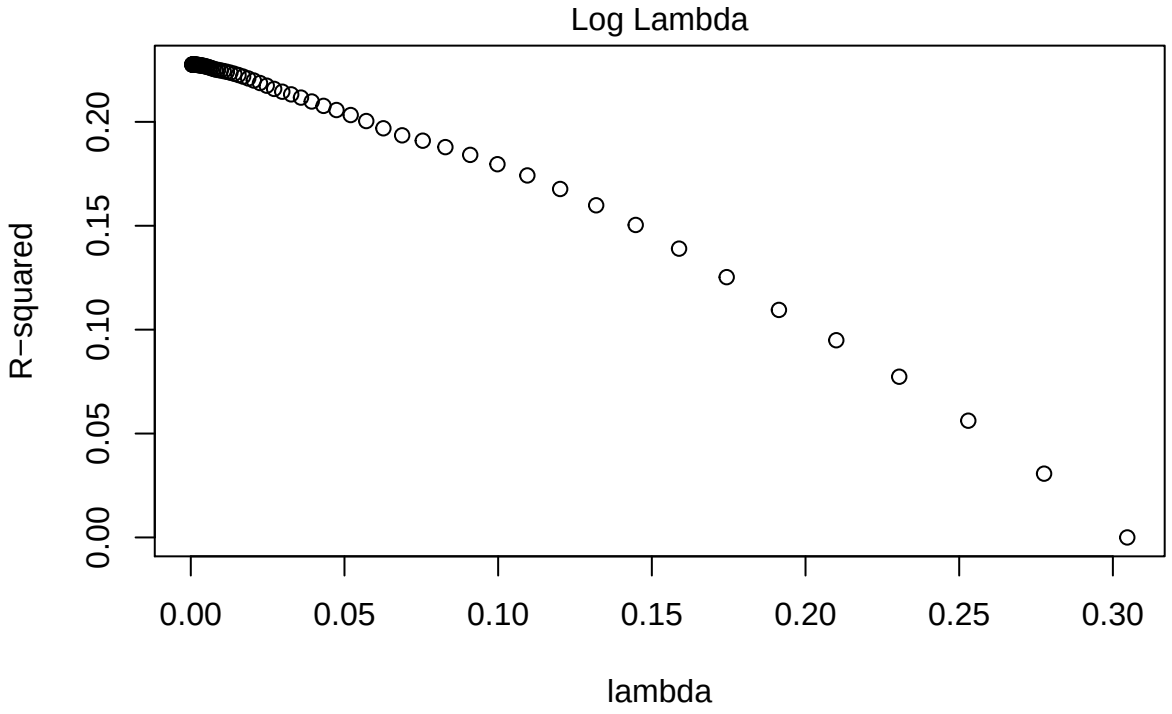
Most of the people, who felt that their own financial situation had worsened, voted for Clinton, while people that had a positive change regarding their personal finances voted for Bush. While even most of those, who felt no personal financial change voted for Clinton, more who had a change for the worst voted for Clinton than for Bush and Perot combined. Perots source of votes are mostly balanced regarding the respondents personal financial situation, but those who felt like the national economy had gotten worse, were much more likely to vote for Perot, than those who felt no change or an increase in national wealth. That being said, those who felt the national economy was getting worse, seem to overwhelmingly have voted for Clinton. Even despite the mentioned bias regarding Perot, Clinton received about as much votes from those who felt a negative change in the national economy as Perot and Bush together, but more people, that were comfortable with the change of the national economy voted for Bush, than for Clinton and Perot combined.

Those who ideologically identified themselves most with the candidate of the democratic party unsurprisingly mostly voted for Clinton. Moving away from the optimal ideological alignment of the respondent with Clinton, more and more voters mostly chose Bush over Clinton, as well as Perot.
Voters, who ideologically identified more with the candidate of the republican party did not vote as decisively for Bush as the democrats for Clinton. Not only did more voters, aligning with Bush, vote for Perot but also much more for Clinton than the voters, aligning with Clinton, for Bush. Again, moving away from the ideological alignment of the voters with Bush, this trend is much more prominently visible, voters absolute decisively voting Clinton instead of anyone else. But recognizing those majorities, we have to keep in mind that most of the respondents as a whole did vote Clinton, so majorities in favour of Clinton are to be expected.
The same trend is visible in the histogram regarding ideological alignment with Perot. While those, who aligned more with Perots ideology tended to vote Bush a little more often than Clinton, moving away from the ideological alignment we again see a strong tendency towards Clinton. The small trend towards Bush by voters aligning with Perot is not surprising, because Perot later was considered a republican, but in 92 disagreed with Bush on some things, most prominently regarding war among other topics. But even those who align with Perot the most, did not end up voting for Perot. That is because a vote for Perot would not determine presidency, because America has a two-party voting system in which only two presidential candidates can be voted for and Perot in 92 was a third party candidate, that couldn't be elected president.

Building the models

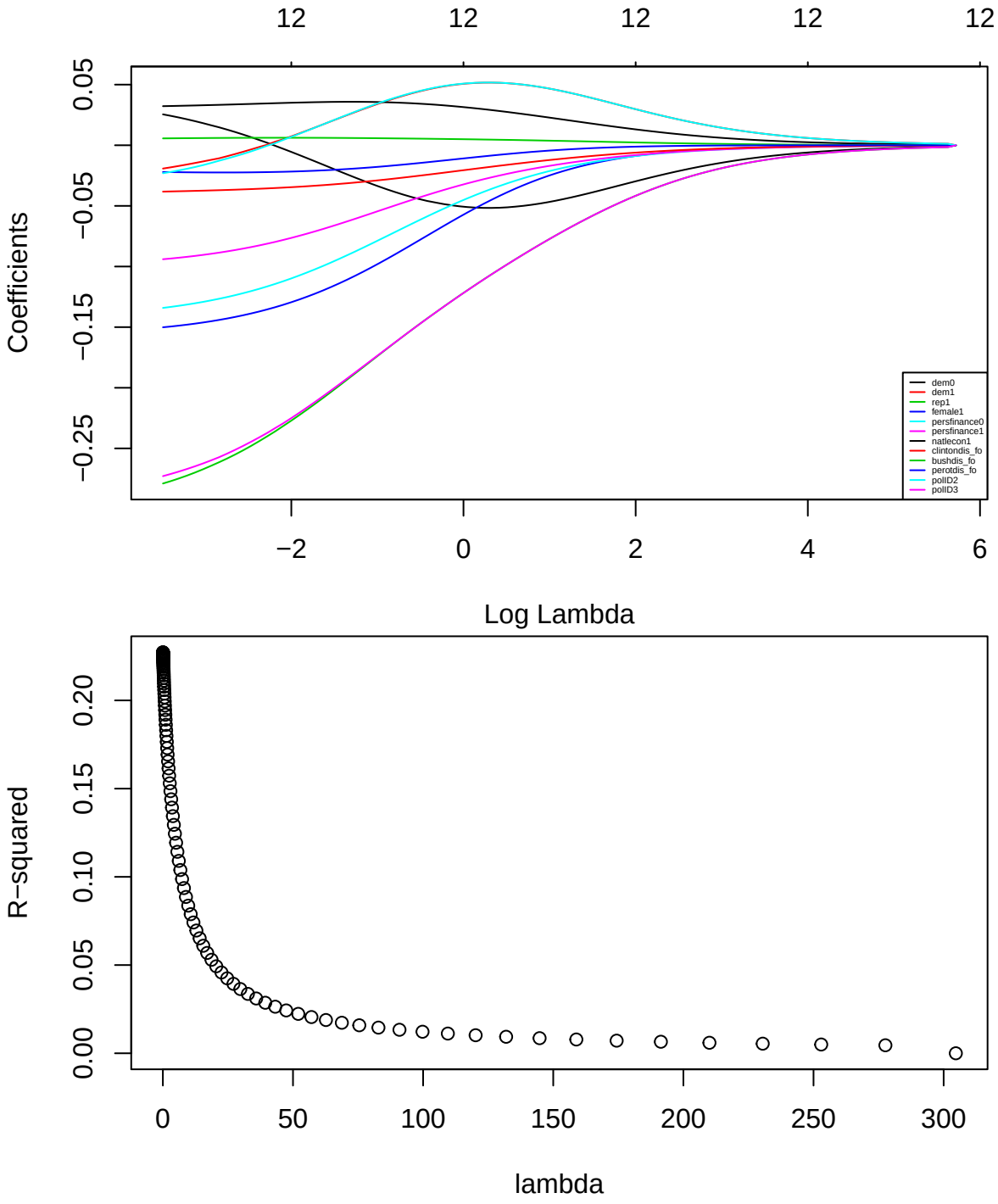
1. Lasso






```
## Warning: Only mse, deviance, mae available as type.measure for Gaussian models;
## mse used instead
```

2. L1-nrom




```
## Warning: Only mse, deviance, mae available as type.measure for Gaussian models;
## mse used instead
```

3. Logistic regressions

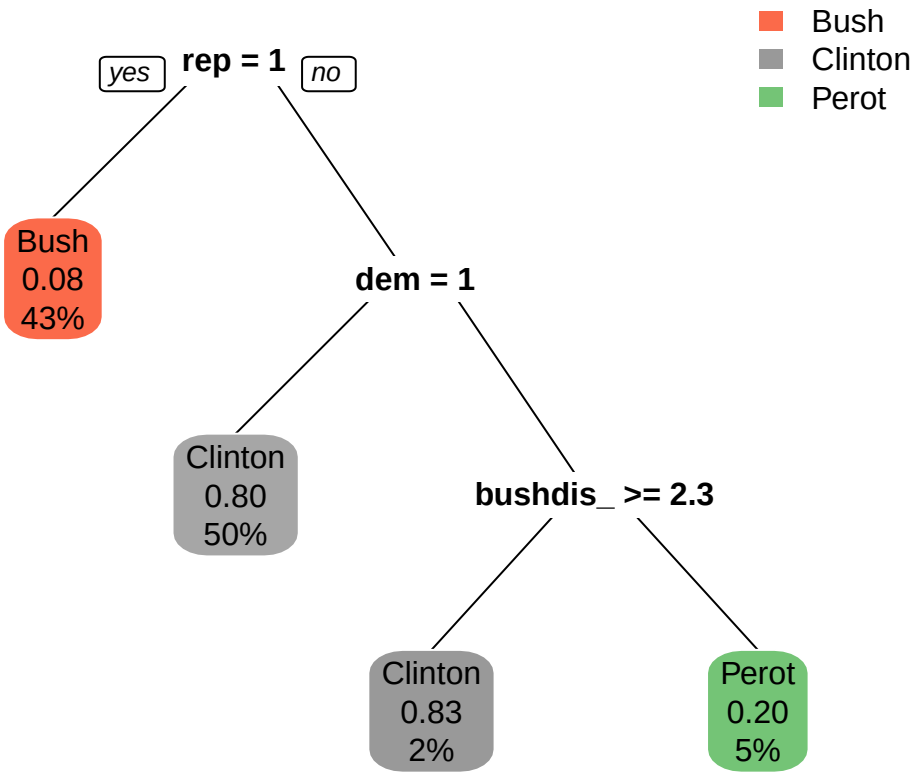
```
# fixed variables
log1 <- glm(vote_num ~ dem + rep + female + persfinance +
              natlecon + clintondis_fo + perotdis_fo + bushdis_fo,
            data = vote)
```

```r
# original variables
log2 <- glm(vote_num ~ dem + rep + female + persfinance +
              natlecon + clintondis + perotdis + bushdis,
            data = vote)

# both fixed and original variables
log3 <- glm(vote_num ~ dem + rep + female + persfinance +
              natlecon + clintondis + perotdis + bushdis + clintondis_fo +
              perotdis_fo + bushdis_fo,
            data = vote)

pred.log1 <- predict(log1, vote, type = "response")
pred.log2 <- predict(log2, vote, type = "response")
pred.log3 <- predict(log3, vote, type = "response")
```

```
## Warning: extra=106 but the response has 3 levels (only the 2nd level is
## displayed)
```



Naturally after the pre-processing we needed to build some models. So we decided to build 3 types of models Lasso and ridge models, logistic regression models, and decision trees. For our lasso and ridge models we needed to spilt the data set into train (70%) and test (30%) data and we used the features we pre-processed in the previous task (dem , rep , famle, persfinance, natlecon, clintondis_fo, bushdis_fo, petrodis_fo). We then use those train and test sets to train and test our lasso and ridge models. Afterwards we created a couple of logistic regression models. 2 of those models either had categorical variables only or continuous ones only. 1 had both the continuous and the categorical variables that have been processed. The finale 1 had the continuous variables that have not been processed (those without "fo" at the end) and the categorical values. All of our logistic regression models had the same target value which was vote_num. Vote_num is the variable vote that is a numeric instead of a factor. For building a decision tree we also needed to split the data into train and test sets. We then defined some future for the tree to be trained with (vote, dem, rep, female, persfinance, natlecon, clintondis_fo, bushdis_fo, petrodis_fo).

Predictions

```
Accuracy<-function(pred,real,threshold=.5) {
  predClass<-ifelse(pred>threshold,1,0)
  return(sum(predClass==real)/length(real))
}

# Accuracy
(acc1 <- Accuracy(pred = pred.log1, real = vote$vote_num))
```

```
## [1] 0.3410341
```

```
(acc2 <- Accuracy(pred = pred.log2, real = vote$vote_num))
```

```
## [1] 0.3410341
```

```
(acc3 <- Accuracy(pred = pred.log3, real = vote$vote_num))
```

```
## [1] 0.3410341
```

```
# Brier Score
(BS.log1 <- sqrt(mean((vote$vote_num - pred.log1)^2)))
```

```
## [1] 0.6459124
```

```
(BS.log2 <- sqrt(mean((vote$vote_num- pred.log2)^2)))
```

```
## [1] 0.6479702
```

```
(BS.log3 <- sqrt(mean((vote$vote_num - pred.log3)^2)))
```

```
## [1] 0.6436077
```

```
(accLasso <- Accuracy(pred = y.predlog_l1, real = y.test))
```

```
## [1] 0.3516484
```

```
(accLRidge <- Accuracy(pred = y.predlog_r1, real = y.test))
```

```
## [1] 0.3516484
```

```
(BS.logL1 <- sqrt(mean((y.test - y.predlog_l1)^2)))
```

```
## [1] 0.6852876
```

```
(BS.logL2 <- sqrt(mean((y.test - y.predlog_r1)^2)))
```

```
## [1] 0.6852333
```

```
# ----- Evaluating Prediction Quality -----
# Calculate performance with AUC and RMSE
auc(vote.test$vote_num, pred.dt)
```

```
## Warning in roc.default(response, predictor, auc = TRUE, ...): 'response'
## has more than two levels. Consider setting 'levels' explicitly or using
## 'multiclass.roc' instead
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.9299
```

```
( rmse <- sqrt(mean((vote.test$vote_num - pred.dt)^2)) )
```

```
## [1] 1.565776
```

```
Accuracy(pred=pred.dt, real=vote.test$vote_num)
```

```
## [1] 0.01845018
```

Since we created the models we might as well Evaluate them. We will start by looking at our logistic regression models all of them have the same accuracy! So we need to look at the Brier score. There we can see that log3 which is the logistic model with both the processed and original variables is the best one out of all of them. While comparing the lasso and ridge model we hit the same road block. The accuracy score for both of them is the same. Which means we need to look at the brier score, which would mean that the ridge model is the slightly better model. Now when it comes to comparing both the ridge and the log3 model we can say that the best model out of both of them is the log3 model since the brier score of the log3 is smaller than the score of the ridge one. Meaning that the log3 model is more accurate than the ridge model.

The only challenge that we really faced was making the decision tree. For some reason the code we had written didn't seem to work for us at first. But after countless attempts we were able to solve it. So the both of us believe that we did quite well.