PREDICTING PEOPLE'S POLITICAL LEANINGS BASED ON DEMOGRAPHICS

Authors: Monica Swartz, Zoe Meers, Angie Dinh

December 15th, 2016

This project is an extension from the final project for Multiple Regression class. I would like to thank Monica Swartz and Zoe Meers for working with me during that time.

In the class project, we chose a linear regression model based on our intuition. The model violates the Equality of Variance assumption for linear regression and explains very little of the variation in data (low R^2). We concluded that it was because we did not include enough variables, causing omitted variable bias.

In this extension of the class project, I implement a modified version of the backward selection algorithm to select important variables and validate the appropriate features with context reasoning. The new model satisfies the assumptions for the linear regression and explains more of the variability in the data (higher R^2).

For the full code and documentation of this project, please refer to the Appendix section at the end of the report. The .Rmd source code file is also available in the same folder.

Abstract:

This paper examines the political leanings of individuals from a wide variety of countries around the world. Using data gathered from the Comparative Study of Electoral Systems, we examined whether income, religiosity, education, country, location, occupation, marital status or age at the time of the survey could influence how left-wing or right-wing the survey respondent was. We conducted a linear regression model and our findings suggest that religiosity, country, income and education significantly influence one's political ideology. For instance, high levels of education suggest that one is more left-leaning than those who are not educated or have a very low level of education. Our project was limited by the amount of variables we had - it would be impossible to include all of the predictors of what makes someone left-leaning or right-wing. Furthermore, we met all assumptions except for equality of variance not met. The model better

predicted the ideologies of those who we would characterize as being in the center, rather than those who were on the extreme-left or extreme-right.

Introduction

Our world is becoming increasingly politically polarized (Pew Research Center, 2014). With the increase in far-right ideology on the mainstream political stage here in the United States and elsewhere around the world, we thought it would be interesting to see what factors may influence one person's position on the left-right political spectrum. We were interested in what makes a person politically left-wing, centrist or right-wing as this may explain why people are moving further towards the political extremes. This topic is extremely pertinent in the United States following the 2016 election but also has implications for other societies, particularly in European nations such as the UK and France.

We can draw some interesting conclusions from our model about what actually influences political opinions. The role of location, education, and gender, amongst other variables, in influencing people's positions on the left-right scale may not be entirely surprising. The assumption, at least in the United States, is that those on the right have a tendency to be less educated, poorer and to live in more rural areas (Pew Research Center, 2016). We have tested these claims in our model - is someone who is less educated more conservative, for instance? The quantification of normative claims in political science helps prove commonly-held beliefs in the discipline and also provides context for those who are interested in why certain groups of people tend towards one direction in politics and not the other.

Data

Our data came from the Comparative Study of Electoral Systems, a group of election study teams from various countries that survey participants in their country following national elections. They use a common module of questions about elections and politics along with voting, demographic, district and electoral system variables. The sample was collected by national election study bodies in each country. For instance, in the United States the survey was conducted with the ANES (the American National Election Studies). The data was structured so that each respondent was an observation with each variable corresponding to a survey question. There were 28 surveys taken from 2011-2016 in 24 different countries. A few countries, such as Mexico, conducted the

survey multiple times as they had a federal election several times during this period. The data was organized by alphabetical order of countries, starting with Australia and finishing with the United States. In the original data set, there were over 51,663 observations. As this is a common survey sent out to all countries, comparative analysis across nations can easily be achieved. The raw data was available in several formats - we used the csv file.

Model Selection Methodology

Because there are many potentially significant variables in this model, I perform model selection using a combination of the automatic approach - backward elimination - and the intuition check - looking closely at each variables and see if the automatic selection makes sense.

Step 1: Backward Elimination Algorithm

There is are built-in functions in R, call stepwise regressions, that performs automatic variable selection. They combine variables in different ways and choose the best model. However, when dealing with factor variables with many levels, the built-in functions considers each level as a separate variable, thus gives results of "best models" without all the levels in the data (for example, having Master degree is included but having High School Diploma is not included in the model). I do not think that this is the best way for model selection, because we are also interested in the overall relationship of a variable across all levels (for example, the overall relationship between education and political spectrum across all levels of education). Also, after preliminary cleaning of unrelated variables, our data still has 38 variables, and each of which have from six to over twenty levels. Thus treating each level as a separate variable gives a very complicated stepwise regression result.

Therefore, I decided to implement my own version of the back-ward elimination algorithm, using the p-value of overall significance of each variable (including many levels), as opposed to the p-value for individual levels. I also use AIC instead of Cp for information criterion. AIC is the information criterion that balances R^2 and model simplicity, because complicated model have higher R^2 but tend to overfit and perform poorly on new data. I also include calculation of the cross-validation error into the algorithm and use it as one of the information criteria. I want to choose models, high R^2, low AIC, and low cross-validation mean squared error. The detail of the algorithm is shown below:

1. Run the linear regression on all the remaining variables in the dataset

2. Run ANOVA table and choose the variable with the largest overall significance p-value

3. If that largest p-value is greater than 0.1, remove that variable from the dataset. I use 0.1 instead of 0.05 to reduce the risk of automatically rejecting important variables.

4. Repeat until the largest p-value is less than 0.1, which means all variables should be considered to be included. I now want to choose the best model among all these significant models.

5. To do that, continue with the process, but now also print out the variable names, R-squared, and AIC.

6. Stop when the number of variables reach the minimum I want to include

Step 2: Variable Analysis

From backward elimination, I have the list of significant variables and potential models. Before evaluating models, I do a qualitative check to see:

1. If any variables labeled as "significant" should not be included, either due to lack of relevance in interpretation, or high/perfect multicollinearity with other models

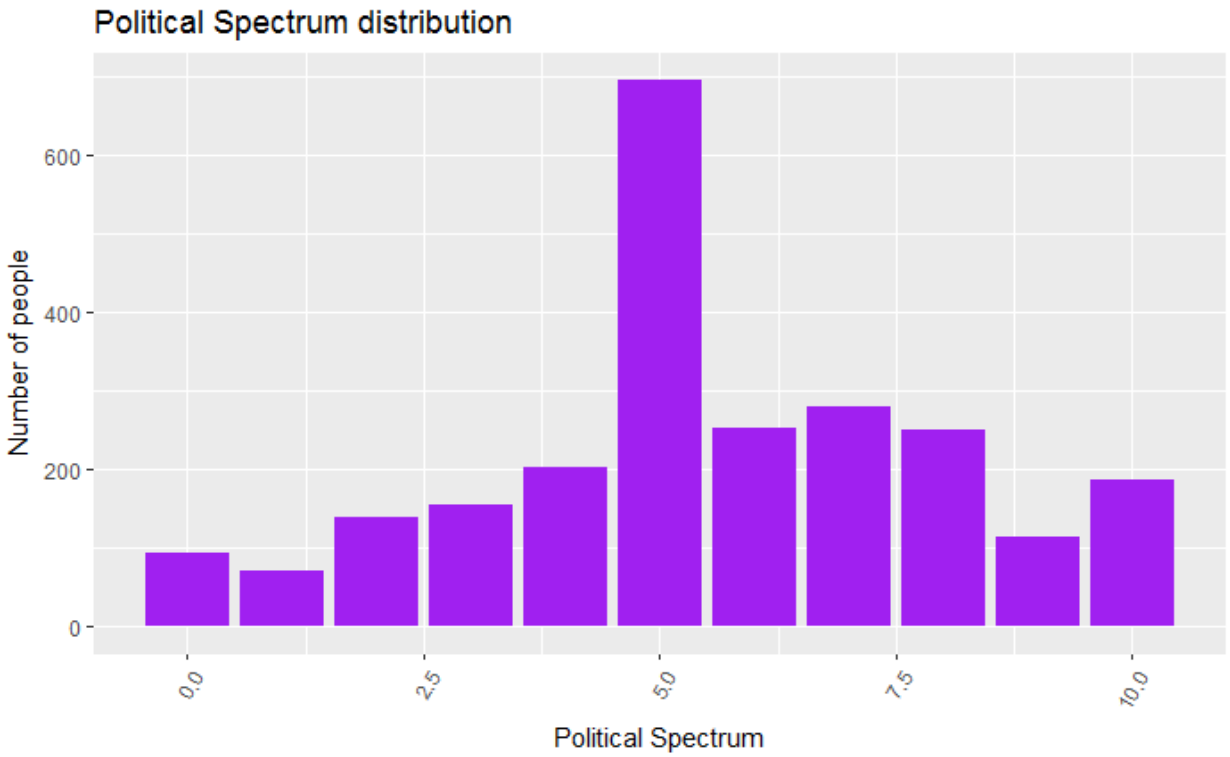
2. If any of the excluded variables can be significant.

After doing all the above steps, I decide on a model which I will elaborate in the following sections.

Descriptive Analysis

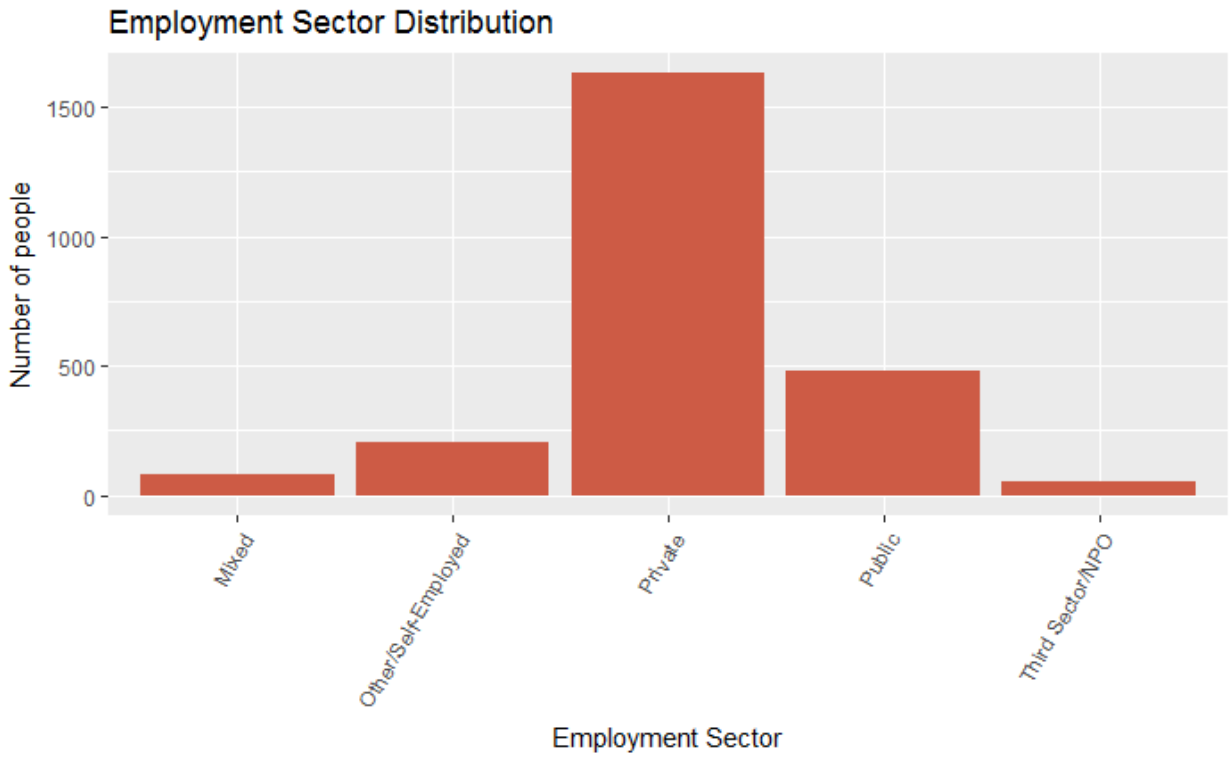
After performing variable selection and data cleaning, the final sample had 2441 observations. I show the distribution of the response variable in the plots below.

Dependent variable (Political spectrum):

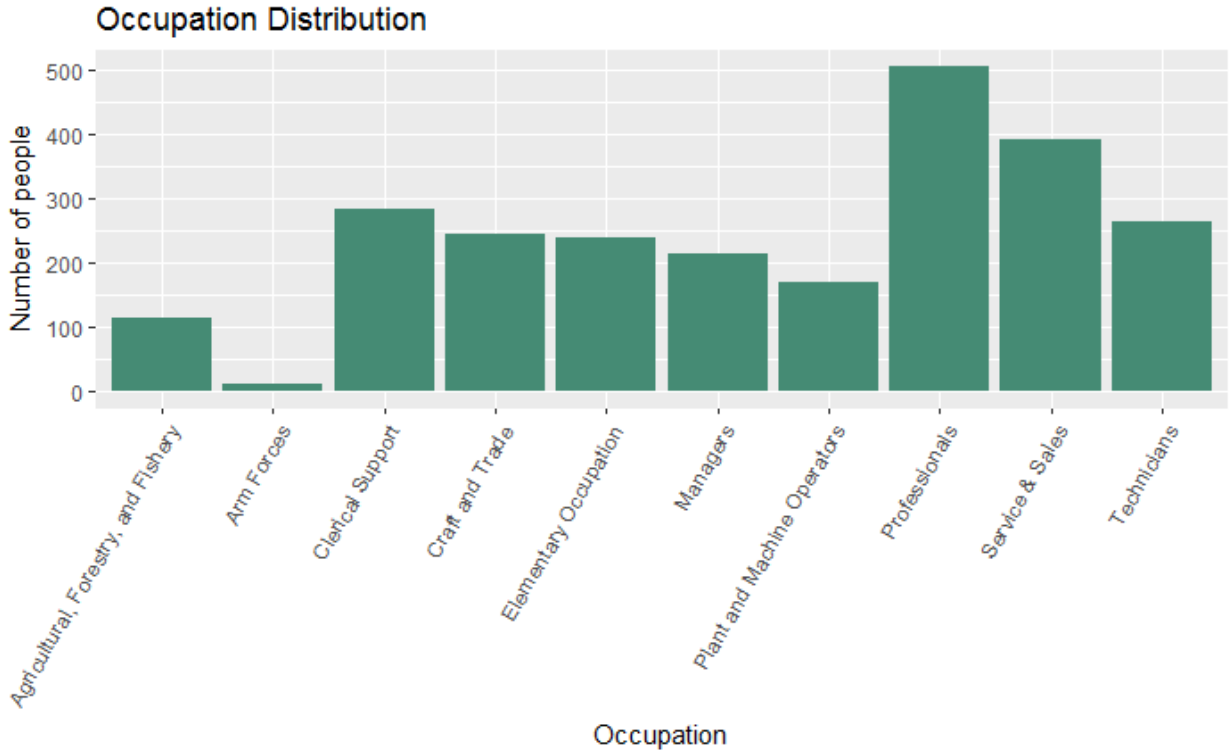


The majority of the observations placed themselves in the center of our spectrum, at 5. I will speak more about the implications of this later in this paper.

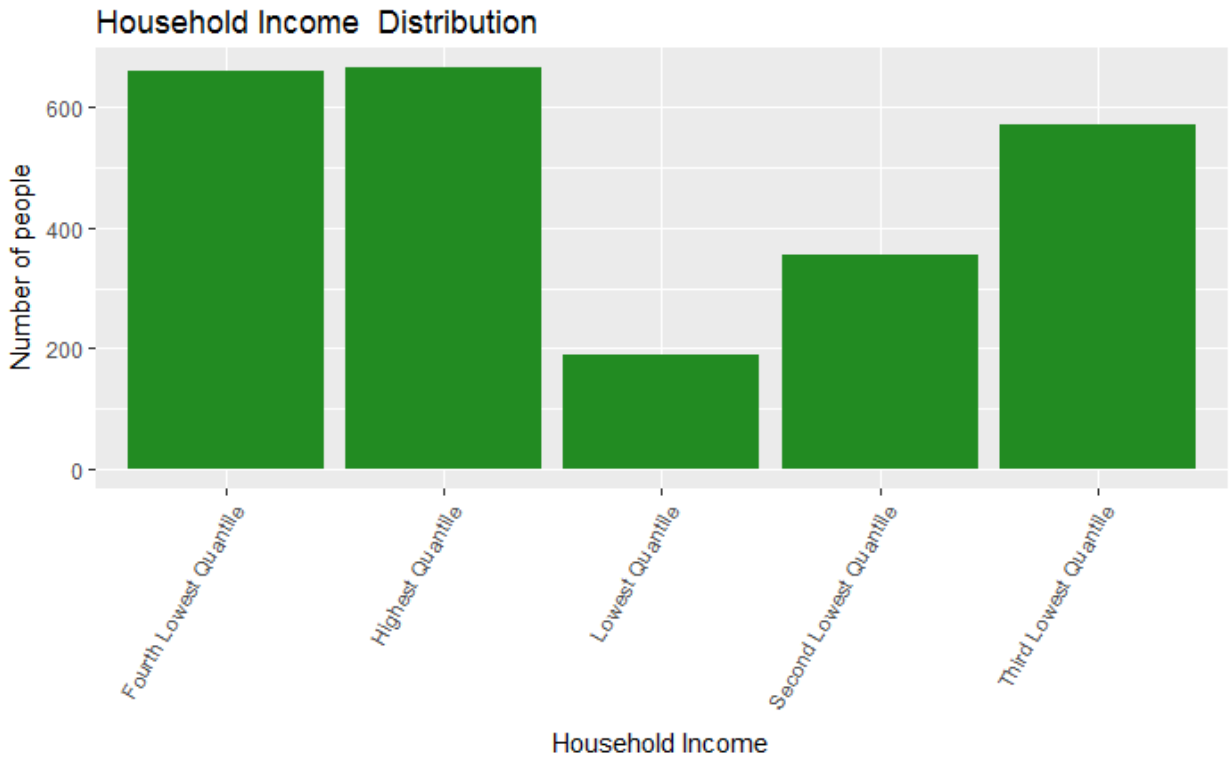
Independent variable: Demographics



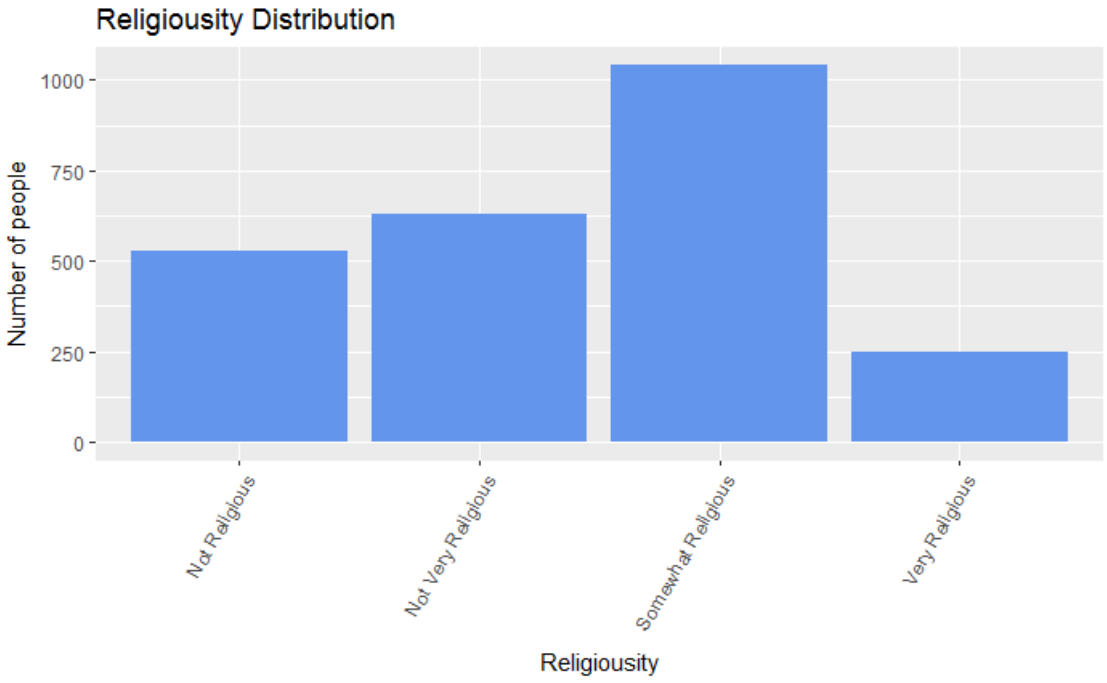
Most people work in the private sector

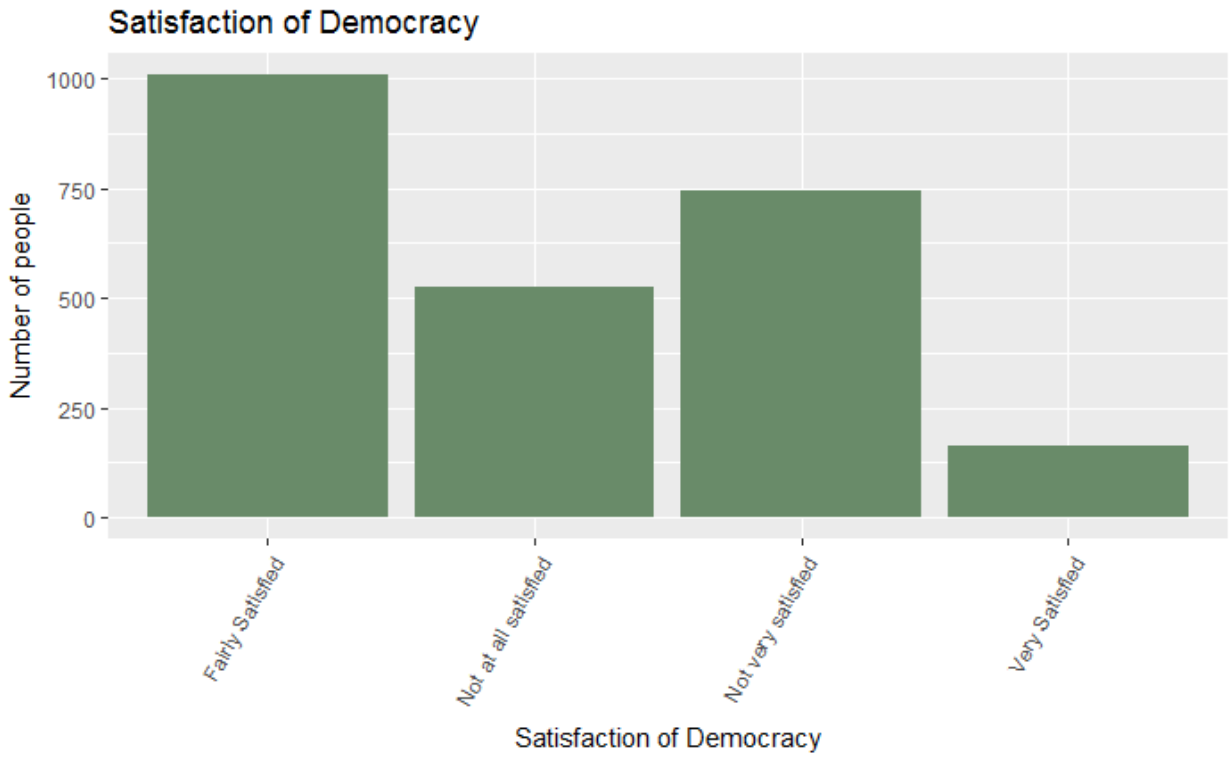


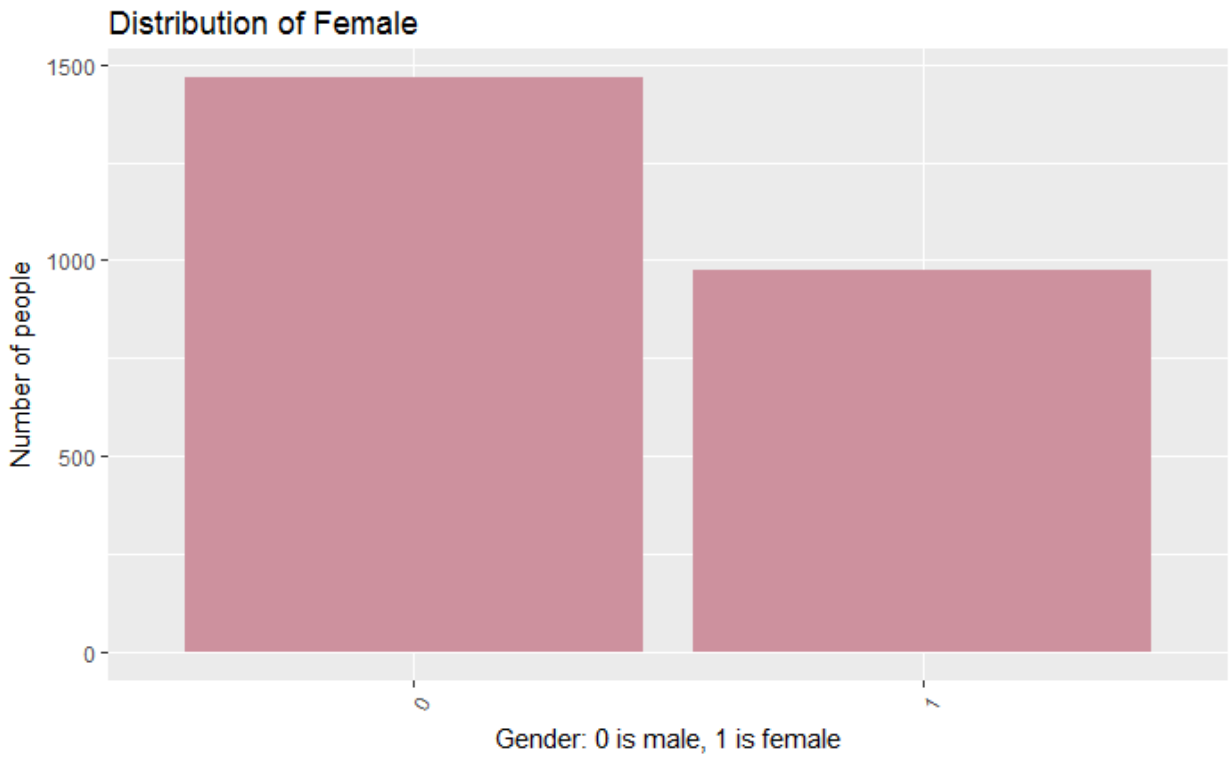
ctor.

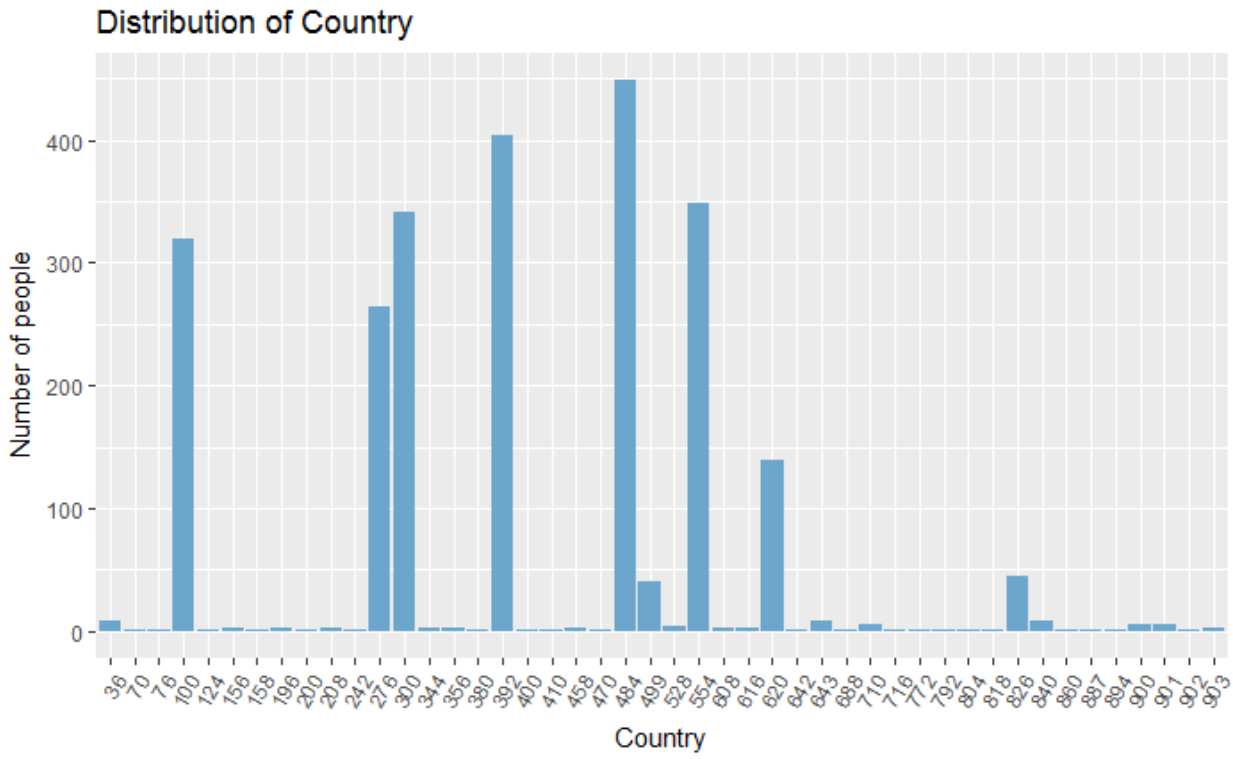


The lowest income group is not well-represented in this sample.

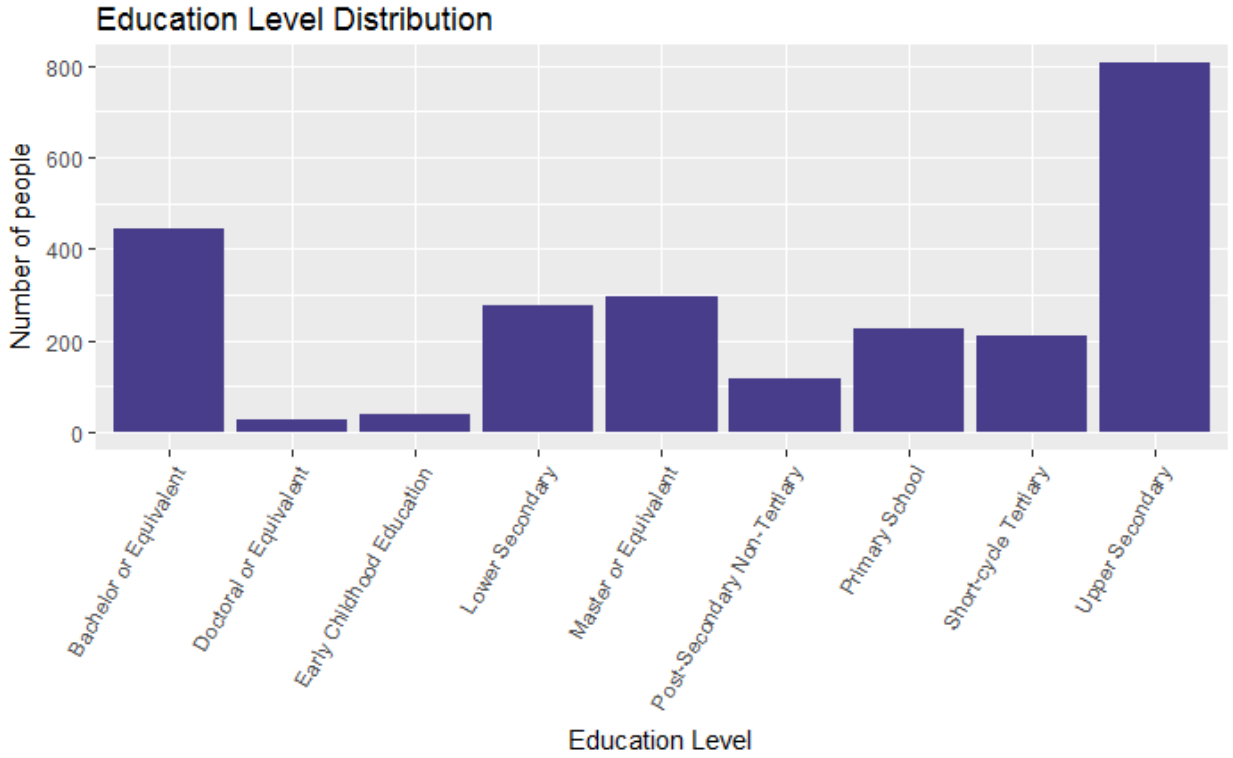



Satisfaction of Democracy
Number of people
1000
750
500
250
0
Fairly Satisfied
Not at all satisfied
Not very satisfied
Very Satisfied
Satisfaction of Democracy


Distribution of Female
Number of people
1500
1000
500
0
0
1
Gender: 0 is male, 1 is female

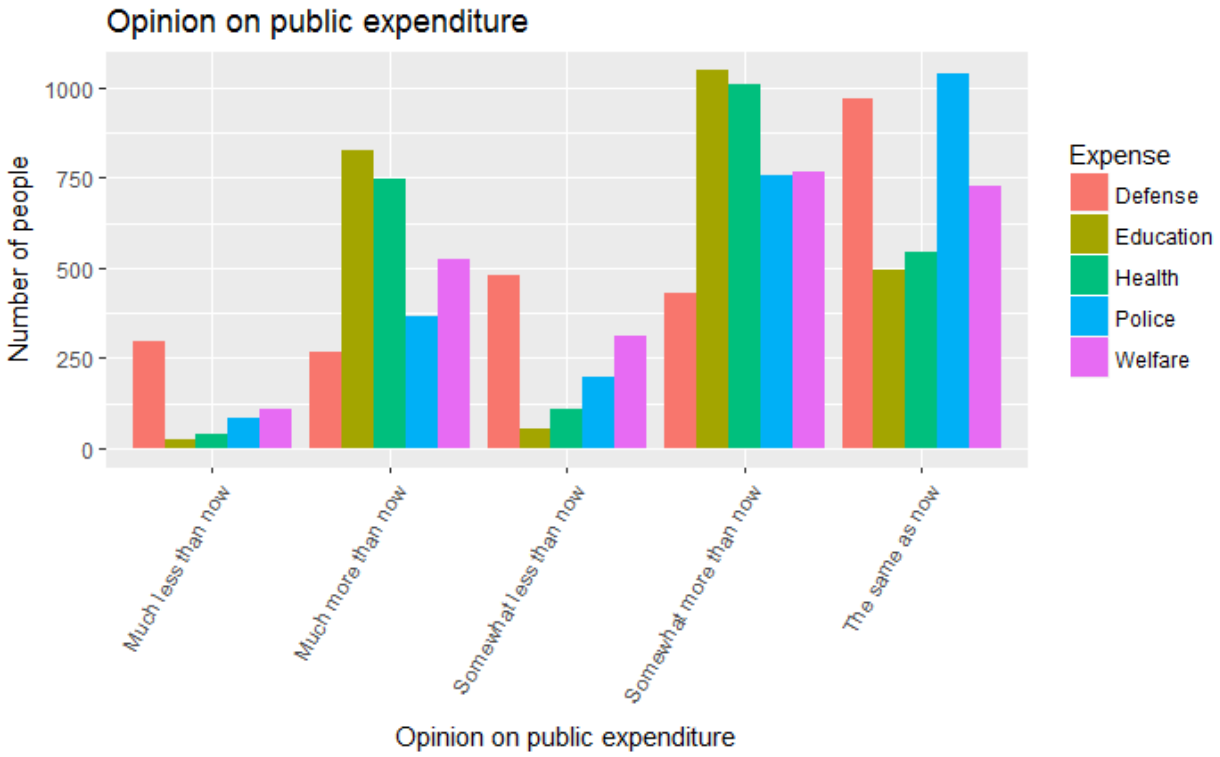


A few countries are well-represented in this sample, others are not.



Independent Variable: Opinions about politics

Beside demographics, the model also includes variables indicating people's political opinions, such as whether the respondent thinks that the government should increase spending on welfare, police, defense, and whether he/she think that the government's job is to reduce inequality. It also includes the respondent's opinion about the economy (better, same, or worse than before) and his/her satisfaction with democracy in his/her country.



Most people prefer an increase in public expenditures on all areas of expense, particularly health.

Model and Model Evaluation:

The model is as follow:

$$\widehat{Left\_Right\_Wing} = \beta 1\ Education + \beta 2\ Union\_Membership + \beta 3\ Union\_Membership\_Family + \beta 4\ Employment\_Status + \beta 5\ Occupation + \beta 6\ Employment\_Sector + \beta 7\ Employment\_Status\_Spouse + \beta 8\ Household\_Income + \beta 9\ Religiosity + \beta 10\ Public\_Expense\_Welfare + \beta 11\ Public\_Expense\_Police + \beta 12\ Public\_Expense\_Health + \beta 13\ Public\_Expense\_Defense + \beta 14\ Public\_Expense\_Education + \beta 15\ Optimism\_Standard\_Living + \beta 16\ Opinion\_Economy + \beta 17\ Opinion\_Gov\_Reduce\_Inequality + \beta 18\ Satisfaction\_Democracy + \epsilon$$

All the independent variables are categorical variables, so this is an ANOVA model and we do not need to test for Linearity. To evaluate this model, we first test the assumptions of Independence, Normality, and Equality of Variance.

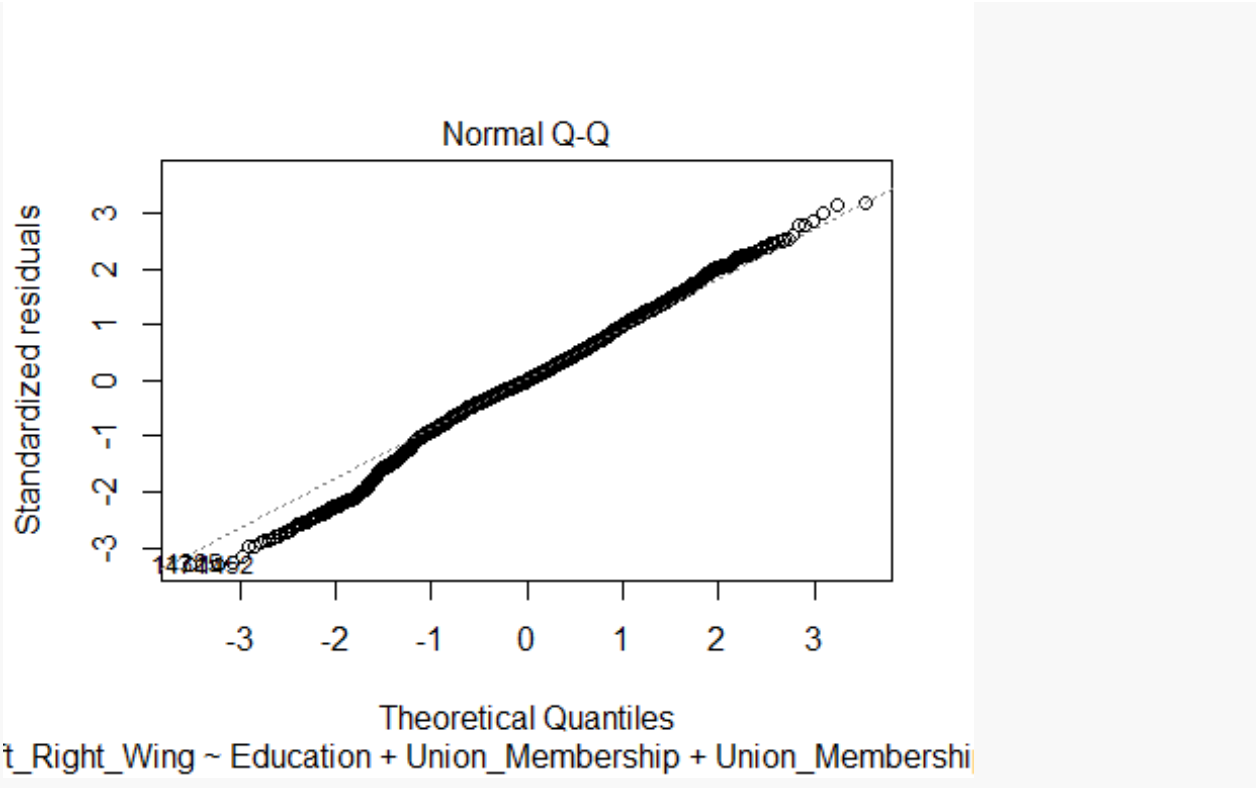
Test for Equality of Variance Assumption

```
ncvTest(reg)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1529629    Df = 1     p = 0.6957198
```

According to this test, we do not find evidence of inequality of variance. The assumption holds.

Test for Normality Assumption

```
plot(reg,which=c(2))
```



This seems to be normal, since the data goes back to follow the normal line at the end of the quantiles.

The Independence assumption cannot be assured because I do not know how the data was collected.

Cross-validation prediction error

The mean squared error of this model, after performing cross-validation, is 5.050172. Therefore, the root mean squared error, which indicates how off the model predicts from the real data, is 2.247. This is a very high error rate; thus the model is not very good for prediction. However, the model still carry value for interpretation and research, as it gives insights about people's political views.

Results and Interpretation

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: Left_Right_Wing
##                            Df  Sum Sq Mean Sq F value    Pr(>F)
## Education                   8   285.1  35.640  7.3503 1.072e-09 ***
## Union_Membership            1    40.5  40.497  8.3521 0.0038877 **
## Union_Membership_Family     1    45.2  45.190  9.3199 0.0022922 **
## Employ_Status               9   118.9  13.216  2.7256 0.0036577 **
## Occupation                  9   164.6  18.284  3.7708 9.899e-05 ***
## Public_Private_Employ       4   267.7  66.935 13.8046 3.934e-11 ***
## Employ_Status_Spouse        9   101.5  11.273  2.3250 0.0132717 *
## Household_Income            4   212.1  53.023 10.9353 8.720e-09 ***
## Religiousity                3   372.1 124.048 25.5836 2.773e-16 ***
## Birth_COuntry               7   755.5 107.922 22.2577 < 2.2e-16 ***
## Public_Expense_Welfare      4   169.8  42.457  8.7563 5.150e-07 ***
## Public_Expense_Police       4   102.5  25.628  5.2856 0.0003086 ***
## Public_Expense_Health       4   114.9  28.718  5.9228 9.682e-05 ***
## Public_Expense_Defense      4   128.1  32.032  6.6063 2.767e-05 ***
## Public_Expense_Edu          4    57.0  14.246  2.9382 0.0194894 *
## Improv_Standard             3   139.8  46.616  9.6141 2.623e-06 ***
## Economy                     2   150.0  75.005 15.4690 2.117e-07 ***
## Gov_Action_Dif_Income       4   137.8  34.449  7.1047 1.104e-05 ***
## Satisfaction_Democrat       3   245.3  81.751 16.8603 7.775e-11 ***
## Residuals                2353 11409.1   4.849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First of all, our ANOVA test shows that all of our variables are statistically significant at the overall level. That is, without taking into account specific levels of education (high school or college), income (highest or second highest), and other variables, the model says that there is a significant overall relationship between people's political leanings and their income, religiosity, education, country where they are from, occupation, gender, and marital status. However, we are more interested in the significance and magnitude of the relationship at different levels of income and education, as well as different occupations, countries, and other factors. Thus a summary table

showing the detailed coefficients at every level is needed, which is shown below. Only the significant variables are included.

Independent Variable	**Coefficients**	**p-value**	
(Intercept)	6.32	0.00	***
Education: Bachelor or Equivalent	-1.00	0.01	*
Education: Doctoral or Equivalent	-1.16	0.05	.
Education: Lower Secondary	-0.73	0.06	.
Education: Master or Equivalent	-0.83	0.05	.
Education: Primary School	-0.92	0.02	*
Education: Upper Secondary	-0.83	0.03	*
Family Member in Union	-0.37	0.02	*
Employment Status: Unemployed	-0.94	0.05	.
Employment Sector: Mixed	0.87	0.00	**
Employment Sector: Private	0.51	0.00	***
Employment Sector: Third Sector/NPO	1.04	0.00	**
Spouse's employment status: Disabled	-1.59	0.01	**
Household_Income: Second Lowest Quantile	-0.40	0.06	.
Religiosity: Not Very Religious	0.47	0.00	***
Religiosity: Somewhat Religious	0.61	0.00	***
Religiosity: Very Religious	1.22	0.00	***
Country of birth: Bulgaria	0.90	0.00	***
Country of birth: Germany	-1.00	0.00	***
Country of birth: Mexico	1.58	0.00	***
	0.75	0.00	**

Public_Expense_Welfare: Much less than now			
Public_Expense_Welfare: Somewhat less than now	0.70	0.00	***
Public_Expense_Police: Much more than now	0.52	0.00	**
Public_Expense_Health: Much more than now	-0.43	0.01	**
Public_Expense_Health: Somewhat less than now	-0.65	0.01	**
Public_Expense_Health: Somewhat more than now	-0.33	0.01	*
Public_Expense_Defense: Much less than now	-0.47	0.00	**
Think it is likely that their standard of living will improve	0.33	0.03	*
Think that economy is better	0.30	0.03	*
Think that economy is worse	-0.30	0.01	**
Government's job is to improve equality: Somewhat Disagree	0.50	0.01	**
Government's job is to improve equality: Strongly Agree	-0.35	0.02	*
Satisfaction of Democracy: Fairly Satisfied	-0.44	0.02	*
Satisfaction of Democracy: Not at all satisfied	-1.38	0.00	***
Satisfaction of Democracy: Not very satisfied	-0.74	0.00	***

We can interpret from this model that even though there is a relationship between income and political views, as well as occupation and political views, no particular level of income or occupation has a significant correlation.

The relationship between religiosity and political conservativeness is strong. We would expect that more religious people are more right-wing, holding everything else constant. Specifically, very religious people are expected be 1.22 points more right-wing than non-religious people, holding everything else constant. That number is 0.61 for “somewhat religious” people,” and 0.47 for not very religious people. Interestingly, even “not very religious” people are predicted to be more right-wing than those who are completely non-religious.

This model tells us that in general, highly educated people are expected to be more left-wing than people with early childhood education, holding everything else constant. For instance, we would expect people with a bachelor's degree to be 1 point more "left wing" than people with early childhood education, holding everything else constant.

Perhaps the strongest predictor in our model is the country where people are from. Looking at other coefficients, we can see that holding everything else constant, people from stereotypically "conservative" countries such as Bulgaria and Mexico are expected to be more "right-wing": 0.90 points more right-wing for the case of Bulgaria and 1.58 points more for the case of Israel. Germans are expected to be 1 point more liberal than people from other countries.

This model also tells us that people who works in the private sector, mixed sector, or third sector are usually more right-wing than people who works in the public sector. It also shows that people who have family members in unions or disabled spouse are also expected to be more liberal. The same case is also expected of unemployed people.

As for the opinion about government's role and decisions, there is a strong correlation between people's political spectrum and how they view government policy, in particular public expenditures. People who want public welfare to decrease tend to be more right-wing, and people wishing a reduction in expense in defense and police tend to be more left-wing. Those who believe that it is the government's job to reduce inequality are also expected to be more liberal. This agrees very well with liberal and conservative ideologies.

There is also a strong relationship between people's political spectrum and their optimism in life. More right-wing people tend to be more satisfied with the current democracy. They are also expected to be more optimistic about the future standard of living and that the economy is getting better. Is that because of the situations in the sampled countries does not satisfy the liberals, or is that because left-wing people are more involved in activism and more demanding of change?

To sum up, the results from our model confirm prior political research and common consensus that liberalism is widespread among people that are highly educated, live in progressive countries, and work in the public sector, while conservatism is highly correlated with religiosity.

Limitations

This model is an improvement from the prior class project as it now meets the assumption for the linear regression and has a higher R^2 – an improvement from 11% to 21%. However, it still carries very low predictive power, with a high error rate of 2.247. Firstly, most people put themselves into the middle in category 5. We believe that this is a result of people wanting to choose the neutral, least-extreme option. We believe that we cannot account for all factors that may influence one's position on the political spectrum, which is a very hard problem to model mathematically. Going forward with this topic, we need to collect more data on different aspects that can influence people's political viewpoints, or perhaps try a different predictive model.

Conclusion

To sum up, our model confirms a lot of prior research and stereotypes about the types of people that are liberal and conservative. However, it also shows that many things come in to play in influencing people's political viewpoints that cannot be limited to demographics. It is the beliefs about the best government policies that makes one liberal or conservative, and such beliefs may or may not be influenced by one's demographic background.

References

Comparative Study of Electoral Systems (www.cses.org). CSES MODULE 4 THIRD ADVANCE RELEASE [dataset]. June 22, 2016 version.
doi:10.7804/cses.module4.2016-06-22

Pew Research Center, April 2016, "A Wider Ideological Gap Between More and Less Educated Adults"

Pew Research Center, June 2014, "Political Polarization in the American Public"

Code Appendix: Multiple Regression Final Project extension

Ngoc Dinh, Zoe Meers, Monica Swartz

November 20, 2016

1. Data cleaning

```
cses4 <- read_csv("C:/Users/stuadmin/Desktop/Projects/cses4_csv/cses4.csv")
```

This data has 448 variables and 51663 observations.

```
dim(cses4)
```

```
## [1] 51663   448
```

Many variables are completely irrelevant to this research topic, which are removed from the dataset.

```
getUnusedVars <- function(col_name){
  variable_number = as.numeric(str_sub(col_name,2,5))
  return((variable_number<2001 & variable_number!=1004 & variable_number!=1006) |
      variable_number > 3017 |
      variable_number %in% c(3005,3006,3007,3008,3011,3012,3013,3015,3016,3021,3025))
}

remove=lapply(names(cses4),getUnusedVars)

political_data=cses4[-which(remove==TRUE)]
```

I also choose to include survey year (to calculate the age at the time of the survey) and country:

```
political_data= political_data %>%
  mutate(Survey_Year=as.integer(str_sub(D1004,-4,-1)),
         Country=D1006_NAM,
         D1006_NAM=NULL,
         D1004=NULL,
         D1006_UN=NULL,
         D1006=NULL)
```

Structure and name

```
dim(political_data)
```

```
## [1] 51663    54
```

After removing irrelevant variable sections, there are now 54 variables that we will use to select the best model.

```
names(political_data)
```

```
##  [1] "D2001_M"     "D2001_Y"     "D2002"       "D2003"       "D2004"
##  [6] "D2005"       "D2006"       "D2007"       "D2008"       "D2009"
## [11] "D2010"       "D2011"       "D2012"       "D2013"       "D2014"
## [16] "D2015"       "D2016"       "D2017"       "D2018"       "D2019"
## [21] "D2020"       "D2021"       "D2022"       "D2023"       "D2024"
## [26] "D2025"       "D2026"       "D2027"       "D2028"       "D2029"
## [31] "D2030"       "D2031"       "D2032"       "D2033"       "D2034"
## [36] "D3001_1"     "D3001_2"     "D3001_3"     "D3001_4"     "D3001_5"
## [41] "D3001_6"     "D3001_7"     "D3001_8"     "D3002"       "D3003_1"
## [46] "D3003_2"     "D3003_3"     "D3004"       "D3009"       "D3010"
## [51] "D3014"       "D3017"       "Survey_Year" "Country"
```

The variable names are not clear and need to be renamed to something more descriptive.

From the codebook, I know that missing/privacy restricted values in this data are coded as 7-9, 95-99, 995-999, 9995-9999, 99995-99999 depending on the value range of the variable. These numbers are outside the range of the variable values; therefore, leaving these values unchanged can create high leverage points. I create a function to set coded NA values back to NA.

```
setNA <- function(column){
  # Set coded NA values back to NAs
  #   Args: the dataset
  #   Returns: the new dataset, with coded NA values back as NAs.
    if (max(column)==9){
      column[which(column>6)] <- NA
    } else if (max(column)==99){
      column[which(column>94)] <- NA
    } else if (max(column)==999){
      column[which(column>994)] <- NA
    } else if (max(column)==9999){
      column[which(column>9994)] <- NA
    } else if (max(column)==99999){
      column[which(column>99994)] <- NA
    }
  return(column)
}

political_data=lapply(political_data, setNA)

str(political_data)
```

```
## List of 54
##  $ D2001_M    : int [1:51663] NA NA NA NA NA NA NA NA NA NA ...
##  $ D2001_Y    : int [1:51663] 1953 1948 1953 1994 1962 1984 1967 1983 1943
 1937 ...
##  $ D2002      : int [1:51663] 1 1 2 2 2 1 2 1 1 2 ...
##  $ D2003      : int [1:51663] 3 7 7 6 7 7 8 7 3 9 ...
##  $ D2004      : int [1:51663] 1 4 1 4 1 1 1 1 3 1 ...
##  $ D2005      : int [1:51663] 1 1 2 2 2 2 1 2 NA NA ...
```

```
…
```

All variables are currently in numeric types, but most of them are actually categorical (factor). I create a function to change categorical variables to factor type and leave numeric variables unchanged. Of the remaining variables, only birth year and political spectrum (left/right wing) are numeric

```
convertToFactor <- function(mydata){
  # Convert categorical variables to factor, leaving quantitative variables u
nchanged
  # Args: the dataset
  # Returns: the modified data
  data_with_factor <- mydata %>%
  mutate_each(funs(as.factor)) %>%
  mutate(D2001_Y=as.numeric(as.character(D2001_Y)),
         D3014=as.numeric(as.character(D3014)),
         D2021=as.numeric(as.character(D2021)))
  return(data_with_factor)
}
political_data=convertToFactor(as.data.frame(political_data))
str(political_data)
```

For convenience while doing the backward elimination algorithm, I move the dependent variable(Left/Right Wing, or D3014 in the codebook) to the first column and rename it to Left_Right_Wing

```
data_x=political_data%>%
  mutate(D3014=NULL)
D3014=political_data$D3014
political_data=cbind(D3014, data_x)
political_data = political_data%>%
  rename("Left_Right_Wing"=D3014)
```

My preliminary variable check show that there are a lot of missing values in the data. If I remove all observations with at least one variable missing, there will be no data left to use. Hence I need to remove variables that have too many missing values, since values with are usually not useful. The challenge here is that the more variable I include, the smaller the sample size gets, but if I include few variables, I risk removing useful variables. After some trial and error, I choose to remove variables that have more than 18000 missing values, which leaves me with 38 variables and 1308 observations

```
checkNumNALimit <- function(column, NA_limit){
  # Remove variables that have the number of missing values above a certain l
imit
  # Args:
  #   mydata: the dataset
  #   NA_limit: the maximum number of NA values allowed in a variable
  # Returns: the new dataset: missing values are removed and variables with t
```

```
he number of missing values
  # above the limit are removed
    return (sum(is.na(column))>NA_limit)
}

col_to_remove=lapply(political_data, function(x) checkNumNALimit(x,25000))
political_data=political_data[-which(col_to_remove==TRUE)]
str(political_data)
```

I rename the variables into more descriptive names

```
political_data = political_data %>%
  rename("Birth_Year"=D2001_Y,
         "Gender"=D2002,
         "Education"=D2003,
         "Marital_Status"=D2004,
         "Union_Membership"=D2005,
         "Union_Membership_Family"=D2006,
         "Farmer_Assosication"=D2008,
         "Employ_Status"=D2010,
         "Occupation"=D2011,
         "Public_Private_Employ"=D2013,
         "Employ_Status_Spouse"=D2015,
         "Household_Income"=D2020,
         "Number_Household"=D2021,
         "Num_Young_Child"=D2022,
         "Religious_Attendance"=D2024,
         "Religiousity"=D2025,
         "Religious_Denomination"=D2026,
         "Region"=D2028,
         "Rural_Urban"=D2031,
         "District"=D2032,
         "Birth_COuntry"=D2033,
         "Public_Expense_Health"=D3001_1,
         "Public_Expense_Edu"=D3001_2,
         "Public_Expense_UnEmpl"=D3001_3,
         "Public_Expense_Defense"=D3001_4,
         "Public_Expense_Pension"=D3001_5,
         "Public_Expense_Business"=D3001_6,
         "Public_Expense_Police"=D3001_7,
         "Public_Expense_Welfare"=D3001_8,
         "Improv_Standard"=D3002,
         "Economy"=D3003_1,
         "Gov_Action_Dif_Income"=D3004,
         "Power_Make_Dif"=D3009,
         "Vote_Make_Dif"=D3010,
         "Satisfaction_Democrat"=D3017) %>%
           mutate(Age=as.numeric(as.character(Survey_Year)) - as.numeric(as.c
haracter(Birth_Year)), Birth_Year=NULL, Survey_Year=NULL)
```

```
final_political_data=na.omit(political_data)
dim(final_political_data)

## [1] 1412   37
```

Model Selection

1. Automatic Variable Selection:

Step 1: Backward Elimination Algorithm

I run the algorithm the first time to remove completely irrelevant variables. I set the p-value cut-off to be 0.1, to eliminate the risk of removing relevant variables

```
getIndexFromName <- function(mydata,name){
  # Return the index of a given variable in the dataset
  # Args:
  #   mydata: the data set
  #   name: the name of the variable
  # Returns:
  #   The index of that variable
  result=which(names(mydata)==name)
  return (result)
}

backwardElim <- function(mydata, min_var, kfold){
  # Perform backward elimination to choose the best model
  # Args:
  #   mydata: the dataset
  #   min_var: the minimum number of variables we want to include
  # Returns: the data with the minimum number of variables
  # Prints out the variables, R-squared, and AIC for each model.

  min_var_reached=FALSE
  data_model_selection=mydata
#while the number of variables are more than the minimum number of variables
  while(!min_var_reached){

    #perform regression on all the variables
    regression=lm(Left_Right_Wing ~., data=data_model_selection)

    #get the table of overall significance (of each factor variables across a
ll levels), variable name, and p-value of overall significance
    anova_model=anova(regression)
    var_name=row.names(anova_model)[-length(row.names(anova_model))]
    p_value=anova_model[1:length(anova_model[,1])-1,5]

    #create a table of variable name and their corresponding overall signific
ance p-value
```

```
    p_value_table=data.frame(var_name, p_value)

    #find the least significant variable: its index, name, and p-value
    least_significant=which.max(p_value_table$p_value)
    least_significant_name=p_value_table$var_name[least_significant]
    max_p_value=p_value_table$p_value[least_significant]

    #if the number of variables is equal to the minimum, end loop
    if (length(p_value)==min_var){
      min_var_reached=TRUE
    }

    #otherwise, set the p-value of the least significant variable to 0 in the
 p-value table
    p_value_table$p_value[least_significant]=0

    #Find the least significant variable in the data table by name, then remo
ve it
    index_removed=getIndexFromName(data_model_selection, least_significant_na
me)
    data_model_selection=data_model_selection[-index_removed]

    #If the least significant p-value is less than 0.05, meaning that all var
iables are signficant,
    #print out variable names, R-squared, and AIC for model selection
    if (max_p_value < 0.1){
      print("Variables:")
      print(var_name)
      print(paste("Adj R-squared:",summary(regression)$adj.r.squared,sep="
"))
      print(paste("AIC:",AIC(regression),sep=" "))
      if (kfold){
        set.seed(20)
        cv.error.10=rep(0 ,10)
        for (i in 1:10){
          glm.fit=glm(Left_Right_Wing ~. ,data=data_model_selection)
          cv.error.10[i]=cv.glm(data_model_selection, glm.fit ,K=10) $delta
[1]
        }
        print("MSE cross-validation")
        print(mean(cv.error.10))
      }
    }
  }
  return (data_model_selection)
}
new_data=backwardElim(final_political_data,25,kfold=FALSE)

## [1] "Variables:"
##  [1] "Gender"                    "Education"
```

```
##  [3] "Union_Membership"         "Union_Membership_Family"
##  [5] "Farmer_Assosication"      "Employ_Status"
##  [7] "Occupation"               "Public_Private_Employ"
##  [9] "Employ_Status_Spouse"     "Household_Income"
## [11] "Religious_Attendance"     "Religiousity"
## [13] "Religious_Denomination"   "District"
## [15] "Birth_COuntry"            "Public_Expense_Health"
## [17] "Public_Expense_Edu"       "Public_Expense_Defense"
## [19] "Public_Expense_Pension"   "Public_Expense_Police"
## [21] "Public_Expense_Welfare"   "Improv_Standard"
## [23] "Economy"                  "Gov_Action_Dif_Income"
## [25] "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.283231219227458"
## [1] "AIC: 6123.91777888863"
```

Step 2: Variable Analysis

The list of significant variables listed. Variables that can be included are modified into more descriptive level names.

```
political_data=subset(political_data, select=c(Left_Right_Wing, Gender, Educa
tion, Union_Membership, Union_Membership_Family, Employ_Status, Occupation, P
ublic_Private_Employ, Employ_Status_Spouse, Household_Income, Number_Househol
d, Religiousity, Birth_COuntry, Public_Expense_Health, Public_Expense_Edu, Pu
blic_Expense_Defense, Public_Expense_Pension, Public_Expense_Police, Public_E
xpense_Welfare, Improv_Standard, Economy, Gov_Action_Dif_Income, Satisfaction
_Democrat))

final_political_data=na.omit(political_data)

dim(final_political_data)

## [1] 2441   23

any(is.na(final_political_data))

## [1] FALSE

final_political_data$Female=ifelse(final_political_data$Gender==1,0,1)
final_political_data$Gender=NULL

any(is.na(final_political_data))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Education=case_when(
                                 Education==1 ~ "Early Childhood Education",
                                 Education==2 ~ "Primary School",
                                 Education==3 ~ "Lower Secondary",
                                 Education==4 ~ "Upper Secondary",
                                 Education==5 ~ "Post-Secondary Non-Tertiary",
```

```
                                Education==6 ~ "Short-cycle Tertiary",
                                Education==7 ~ "Bachelor or Equivalent",
                                Education==8 ~ "Master or Equivalent",
                                Education==9 ~ "Doctoral or Equivalent",
                                Education==96 ~ "No Education",
                                is.na(Education) ~ "Missing")))
any(is.na(final_political_data))


final_political_data <- with(final_political_data,
  mutate(final_political_data,
         Union_Membership=case_when(
           Union_Membership==1 ~ 1,
           Union_Membership==2 ~ 0)))

any(is.na(final_political_data))


final_political_data <- with(final_political_data,
  mutate(final_political_data,
         Union_Membership_Family=case_when(
           Union_Membership_Family==1 ~ "1",
           Union_Membership_Family==2 ~ "0"
           ,
           is.na(Union_Membership_Family) ~ "Missing")))

any(is.na(final_political_data))


final_political_data <- with(final_political_data,
  mutate(final_political_data,
         Employ_Status=case_when(
           Employ_Status==1 ~ "Full-time",
           Employ_Status==2 ~ "Part-time",
           Employ_Status==3 ~ "Less than 15 hours",
           Employ_Status==4 ~ "Help Family",
           Employ_Status==5 ~ "Unemployed",
           Employ_Status==6 ~ "Student",
           Employ_Status==7 ~ "Retired",
           Employ_Status==8 ~ "Housewife",
           Employ_Status==9 ~ "Disabled",
           Employ_Status==10 | Employ_Status ==11 | Employ_Status==12 ~ "Othe
r",
           is.na(Employ_Status) ~"Missing")))

any(is.na(final_political_data))


final_political_data <- with(final_political_data,
```

```
  mutate(final_political_data,
         Gov_Action_Dif_Income=case_when(
          Gov_Action_Dif_Income==1 ~ "Strongly Agree",
          Gov_Action_Dif_Income==2 ~ "Somewhat Agree",
          Gov_Action_Dif_Income==3 ~ "Neither",
          Gov_Action_Dif_Income==4 ~ "Somewhat Disagree",
          Gov_Action_Dif_Income==5 ~ "Strongly Disagree",
           is.na( Gov_Action_Dif_Income) ~"Missing")))
any(is.na(final_political_data))


final_political_data <- with(final_political_data,
  mutate(final_political_data,
         Employ_Status_Spouse=case_when(
           Employ_Status_Spouse==1 ~ "Full-time",
           Employ_Status_Spouse==2 ~ "Part-time",
           Employ_Status_Spouse==3 ~ "Less than 15 hours",
           Employ_Status_Spouse==4 ~ "Help Family",
           Employ_Status_Spouse==5 ~ "Unemployed",
           Employ_Status_Spouse==6 ~ "Student",
           Employ_Status_Spouse==7 ~ "Retired",
           Employ_Status_Spouse==8 ~ "Housewife",
           Employ_Status_Spouse==9 ~ "Disabled",
           Employ_Status_Spouse==10 | Employ_Status_Spouse==11 | Employ_Statu
s_Spouse==12~ "Other",
           is.na(Employ_Status_Spouse) ~"Missing")))
any(is.na(final_political_data))


final_political_data = final_political_data %>%
  mutate(Occupation=as.numeric(as.character(Occupation)))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Occupation=case_when(
                                    Occupation <100 ~ "Arm Forces",
                                    Occupation >=100 & Occupation < 200 ~ "Mana
gers",
                                    Occupation >=200 & Occupation < 300 ~ "Prof
essionals",
                                    Occupation >=300 & Occupation < 400 ~ "Tech
nicians",
                                    Occupation >=400 & Occupation < 500 ~ "Cler
ical Support",
                                    Occupation >=500 & Occupation < 600 ~ "Serv
ice & Sales",
                                    Occupation >=600 & Occupation < 700 ~ "Agri
cultural, Forestry, and Fishery",
                                    Occupation >=700 & Occupation < 800 ~ "Craf
t and Trade",
```

```
                                  Occupation >=800 & Occupation < 900 ~ "Plan
t and Machine Operators",
                                  Occupation >=900 & Occupation <= 962 ~ "Ele
mentary Occupation",
                                  is.na(Occupation) ~ "Wrong Value")))

any(is.na(final_political_data))


final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Private_Employ=case_when(
                               Public_Private_Employ==1 ~ "Public",
                               Public_Private_Employ==2 ~ "Private",
                               Public_Private_Employ==3 ~ "Mixed",
                               Public_Private_Employ==4 ~ "Third Sector/NPO",
                               Public_Private_Employ==5 | Public_Private_Employ
 == 6 ~ "Other/Self-Employed",
                               is.na(Public_Private_Employ) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Household_Income=case_when(
                               Household_Income==1 ~ "Lowest Quantile",
                               Household_Income==2 ~ "Second Lowest Quantile",
                               Household_Income==3 ~ "Third Lowest Quantile",
                               Household_Income==4 ~ "Fourth Lowest Quantile",
                               Household_Income==5 ~ "Highest Quantile",
                               is.na(Household_Income) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Religiousity=case_when(
                               Religiousity==1 ~ "Not Religious",
                               Religiousity==2 ~ "Not Very Religious",
                               Religiousity==3 ~ "Somewhat Religious",
                               Religiousity==4 ~ "Very Religious",
                               is.na(Religiousity) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Expense_Health=case_when(
                               Public_Expense_Health==1 ~ "Much more than now",
                               Public_Expense_Health==2 ~ "Somewhat more than n
ow",
                               Public_Expense_Health==3 ~ "The same as now",
                               Public_Expense_Health==4 ~ "Somewhat less than n
ow",
                               Public_Expense_Health==5 ~ "Much less than now",
                               is.na(Public_Expense_Health) ~ "Missing")))
```

```
final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Expense_Edu=case_when(
                               Public_Expense_Edu==1 ~ "Much more than now",
                               Public_Expense_Edu==2 ~ "Somewhat more than now
",
                               Public_Expense_Edu==3 ~ "The same as now",
                               Public_Expense_Edu==4 ~ "Somewhat less than now
",
                               Public_Expense_Edu==5 ~ "Much less than now",
                               is.na(Public_Expense_Edu) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Expense_Defense=case_when(
                               Public_Expense_Defense==1 ~ "Much more than now
",
                               Public_Expense_Defense==2 ~ "Somewhat more than
now",
                               Public_Expense_Defense==3 ~ "The same as now",
                               Public_Expense_Defense==4 ~ "Somewhat less than
now",
                               Public_Expense_Defense==5 ~ "Much less than now
",
                               is.na(Public_Expense_Defense) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Expense_Pension=case_when(
                               Public_Expense_Pension==1 ~ "Much more than now
",
                               Public_Expense_Pension==2 ~ "Somewhat more than
now",
                               Public_Expense_Pension==3 ~ "The same as now",
                               Public_Expense_Pension==4 ~ "Somewhat less than
now",
                               Public_Expense_Pension==5 ~ "Much less than now
",
                               is.na(Public_Expense_Pension) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Expense_Police=case_when(
                               Public_Expense_Police==1 ~ "Much more than now",
                               Public_Expense_Police==2 ~ "Somewhat more than n
ow",
                               Public_Expense_Police==3 ~ "The same as now",
                               Public_Expense_Police==4 ~ "Somewhat less than n
ow",
```

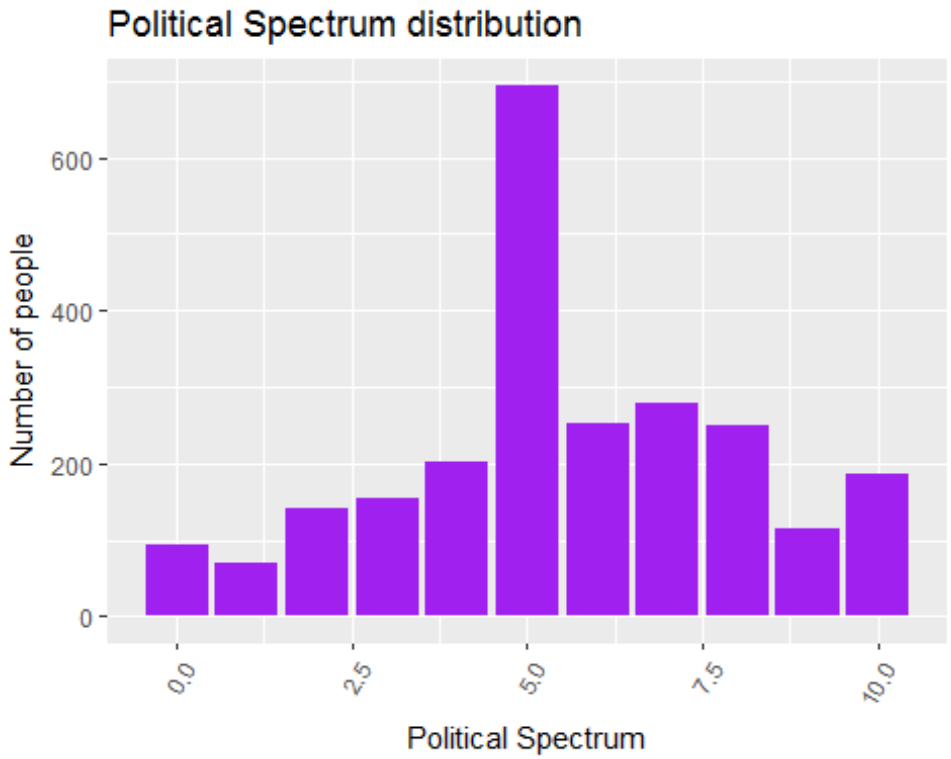
```
                              Public_Expense_Police==5 ~ "Much less than now",
                              is.na(Public_Expense_Police) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Public_Expense_Welfare=case_when(
                              Public_Expense_Welfare==1 ~ "Much more than now
",
                              Public_Expense_Welfare==2 ~ "Somewhat more than
now",
                              Public_Expense_Welfare==3 ~ "The same as now",
                              Public_Expense_Welfare==4 ~ "Somewhat less than
now",
                              Public_Expense_Welfare==5 ~ "Much less than now
",
                              is.na(Public_Expense_Welfare) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Improv_Standard=case_when(
                              Improv_Standard==1 ~ "Very likely",
                              Improv_Standard==2 ~ "Somewhat likely",
                              Improv_Standard==4 ~ "Somewhat unlikely",
                              Improv_Standard==5 ~ "Very unlikely",
                              is.na(Improv_Standard) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                                    Economy=case_when(
                              Economy==1 ~ "Better",
                              Economy==3 ~ "Same",
                              Economy==5 ~ "Worse",
                              is.na(Economy) ~ "Missing")))

final_political_data <- with(final_political_data,
                             mutate(final_political_data,
                              Satisfaction_Democrat=case_when(
                              Satisfaction_Democrat==1 ~ "Very Satisfied",
                              Satisfaction_Democrat==2 ~ "Fairly Satisfied",
                              Satisfaction_Democrat==4 ~ "Not very satisfied",
                              Satisfaction_Democrat==5 ~ "Not at all satisfied
",

                              is.na(Satisfaction_Democrat) ~ "Missing")))
```

Convert character variables to factor

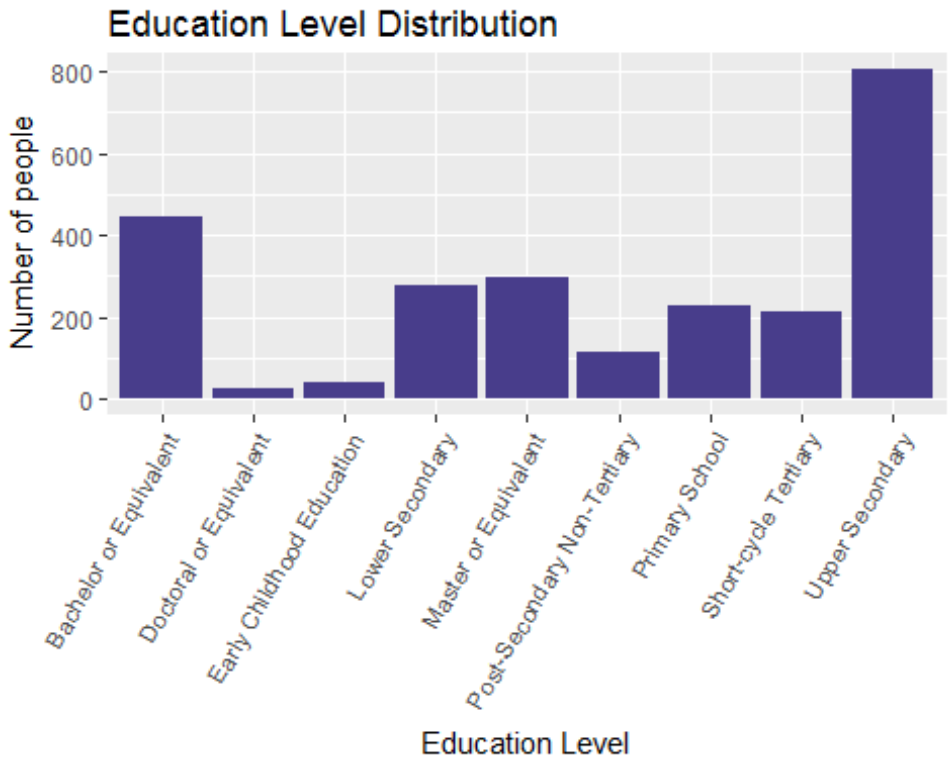
```
final_political_data=final_political_data %>%
  mutate_if(is.character,as.factor)
```

Make the bar charts:

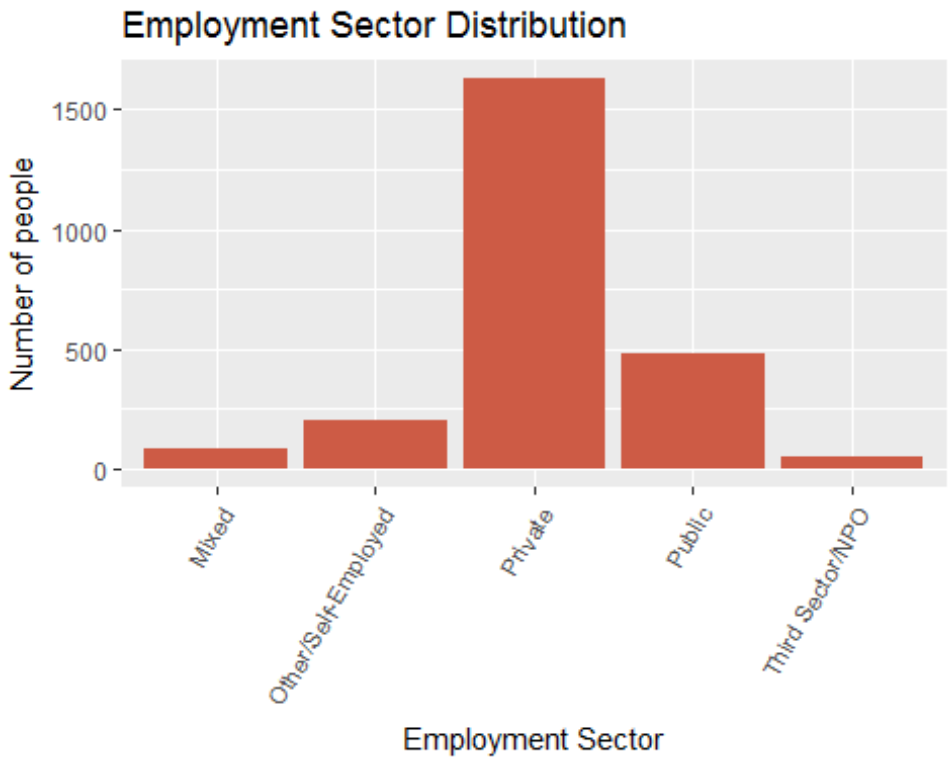
```
barChart <- function(data, variable, x_label, title, color){
  myplot <- ggplot(data, aes(x = variable),
                         y = count)
  myplot + geom_bar(aes(), fill=color) +
  labs(y="Number of people", x=x_label)+
    ggtitle(title) +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))

}

barChart(final_political_data, final_political_data$Left_Right_Wing, "Politic
al Spectrum", "Political Spectrum distribution", "purple")
```



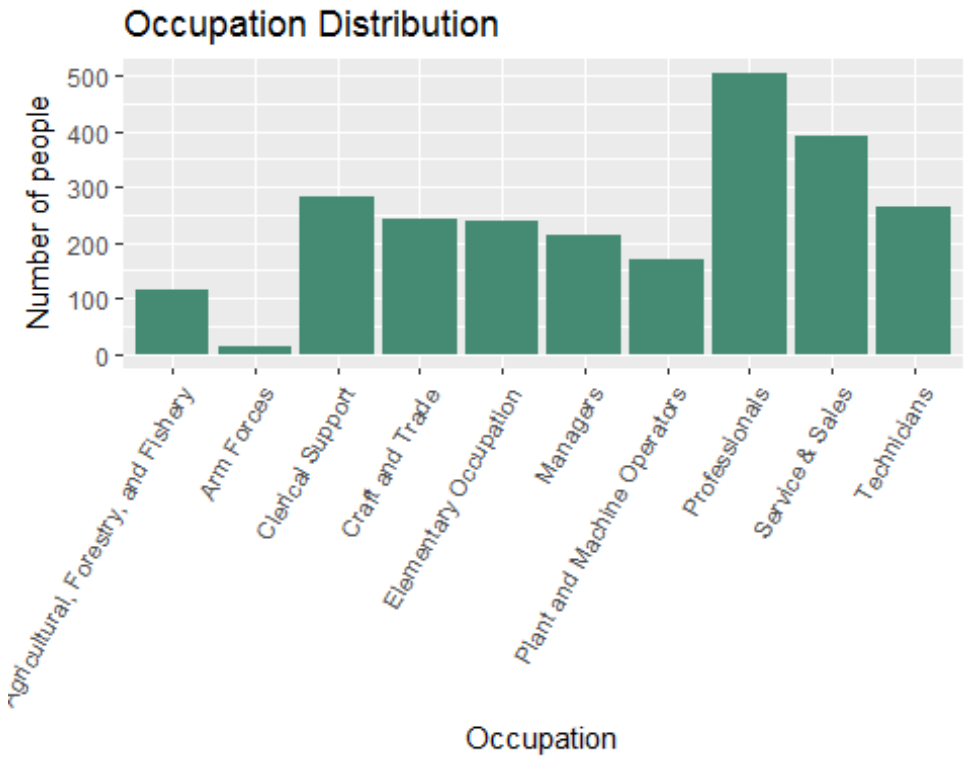

```
barChart(final_political_data, final_political_data$Education, "Education Lev
el", "Education Level Distribution", "darkslateblue")
```



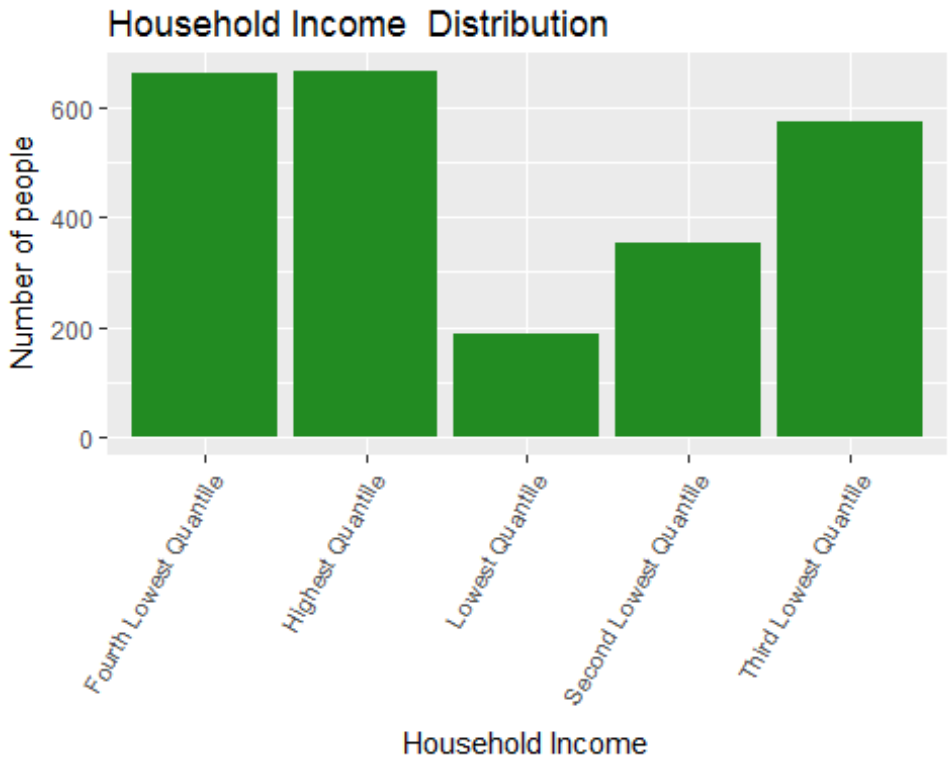

```
barChart(final_political_data, final_political_data$Public_Private_Employ, "E
mployment Sector", "Employment Sector Distribution", "coral3")
```



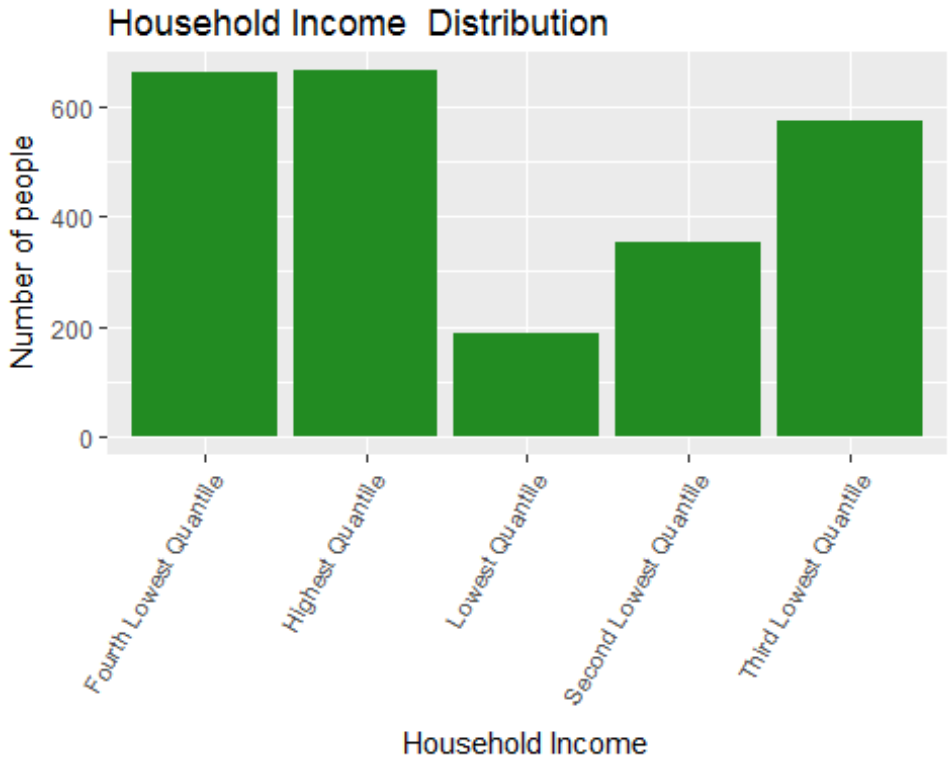
```
barChart(final_political_data, final_political_data$Occupation, "Occupation",
 "Occupation Distribution", "aquamarine4")
```



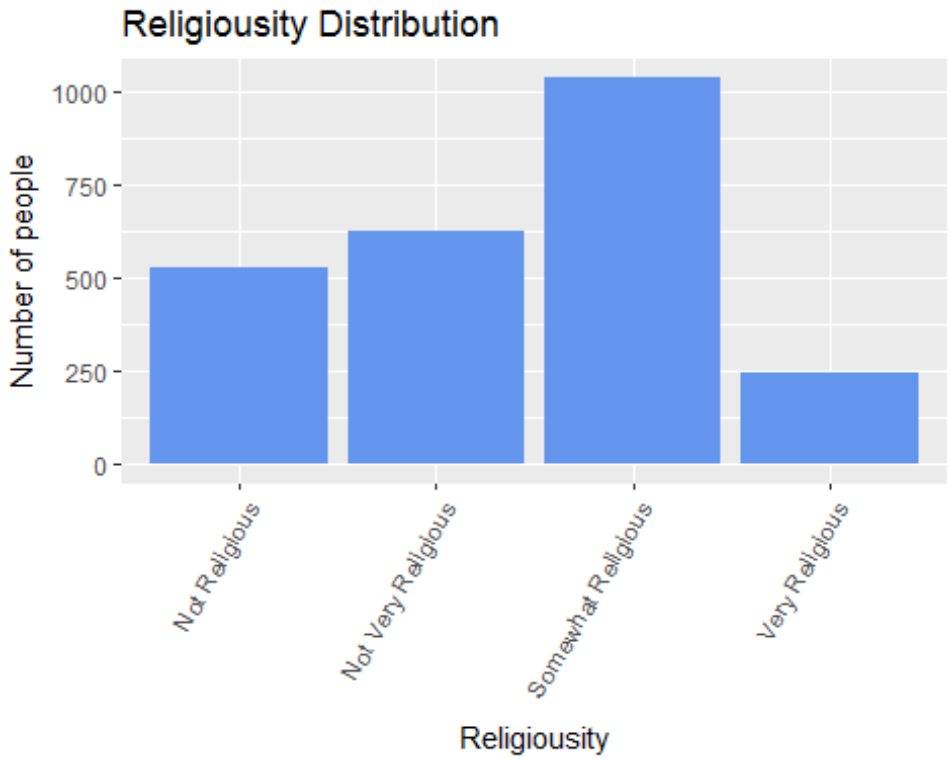

```
barChart(final_political_data, final_political_data$Household_Income, "Househ
old Income", "Household Income  Distribution", "forestgreen")
```



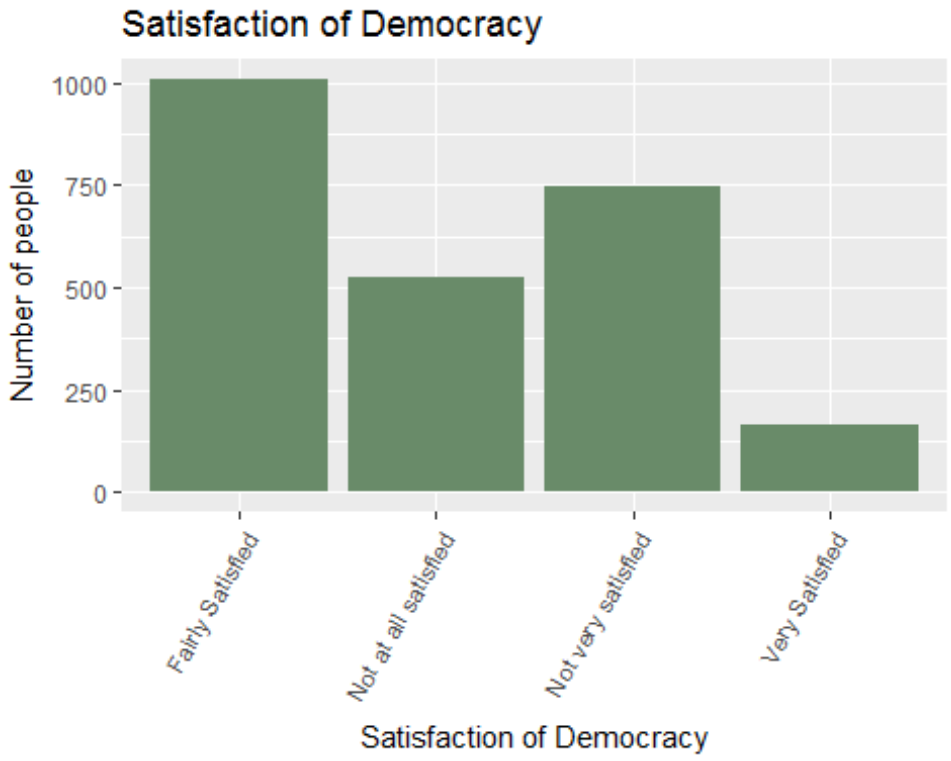

```
barChart(final_political_data, final_political_data$Household_Income, "Household Income", "Household Income  Distribution", "forestgreen")
```



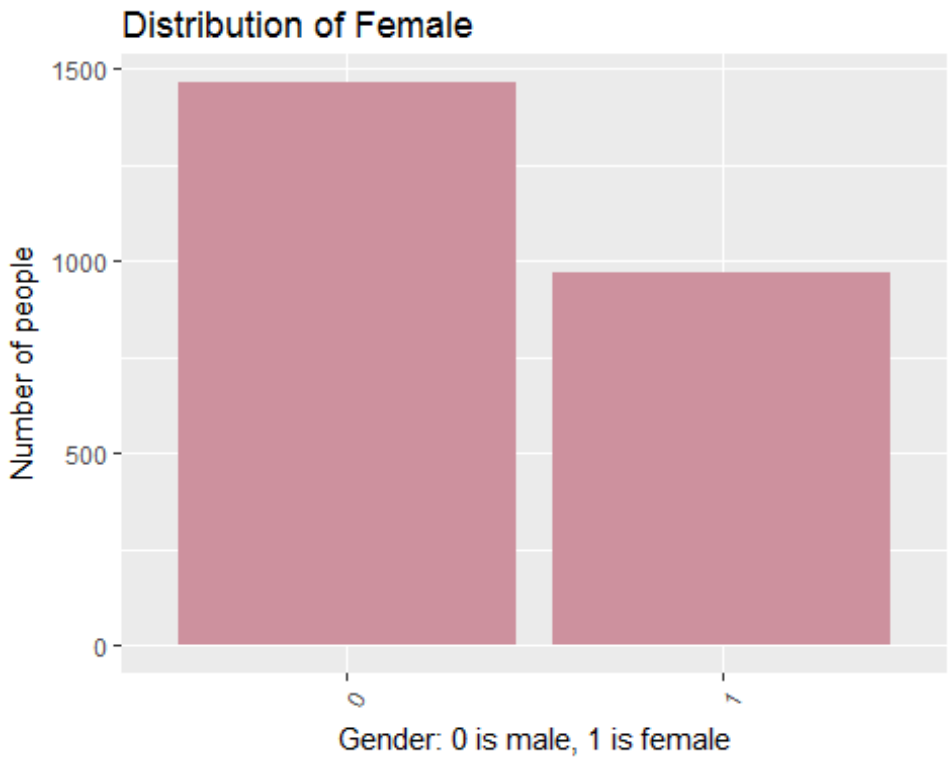
```
barChart(final_political_data, final_political_data$Religiousity, "Religiousity", "Religiousity Distribution", "cornflowerblue")
```



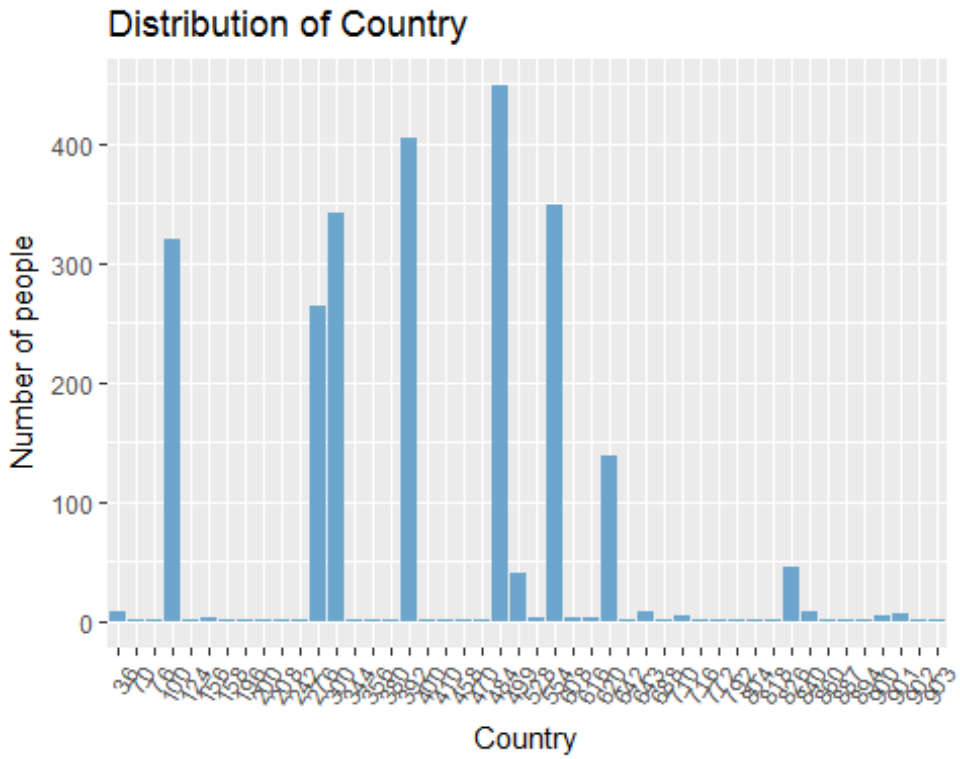

```
barChart(final_political_data, final_political_data$Satisfaction_Democrat, "Satisfaction of Democracy", "Satisfaction of Democracy", "darkseagreen4")
```




```
barChart(final_political_data, as.factor(final_political_data$Female), "Gende
r: 0 is male, 1 is female", "Distribution of Female", "pink3")
```

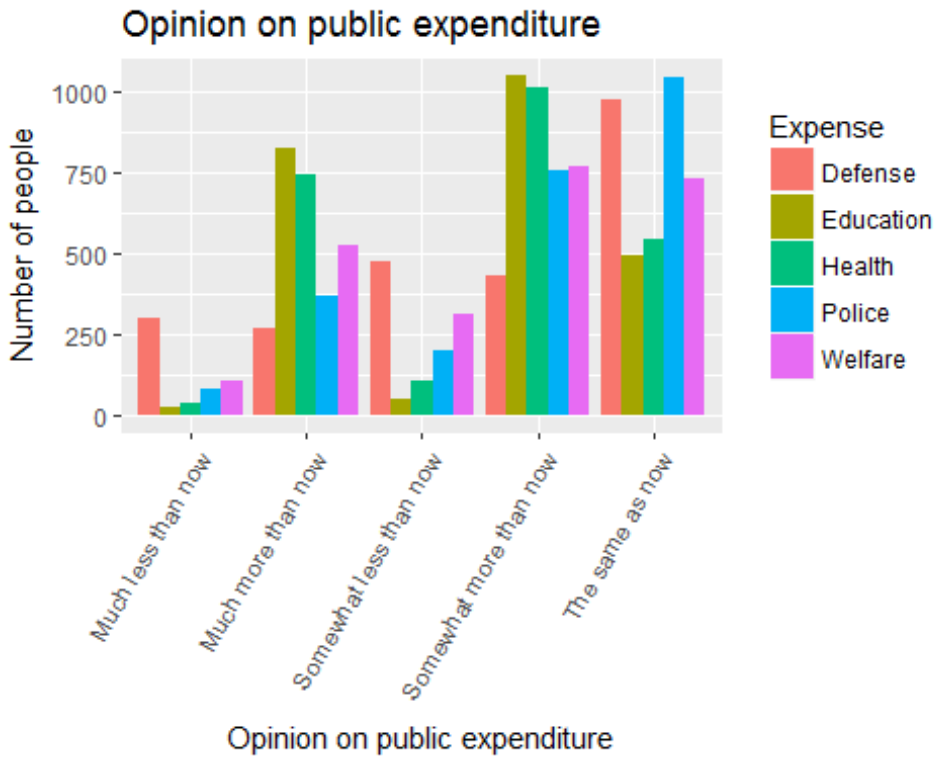


```
barChart(final_political_data, final_political_data$Birth_COuntry, "Country",
 "Distribution of Country", "skyblue3")
```




```
#Create a new table of public expenditure opinion
expenditure_opinion=data.frame(cbind(as.character(final_political_data$Public
_Expense_Health),
                                     as.character(final_political_data$Public
_Expense_Edu),
                                     as.character(final_political_data$Public
_Expense_Defense),
                                     as.character(final_political_data$Public
_Expense_Police),
                                     as.character(final_political_data$Public
_Expense_Welfare)))

names(expenditure_opinion)=c("Health","Education","Defense", "Police", "Welfa
re")

#Reshaping the table
expenditure_opinion <- expenditure_opinion %>%
  gather(key="Expense", value="Opinion")

#Plot it
myplot <- ggplot(expenditure_opinion, aes(fill=Expense, x = Opinion),
                         y = count)
  myplot + geom_bar(aes(), position="dodge")+
  labs(y="Number of people", x="Opinion on public expenditure")+
```

```
    ggtitle("Opinion on public expenditure") +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Set the countries with few observations as "Other":

```
rareLevelAsOther <- function (variable, var_name, mydata, limit){
  freq_table=as.data.frame(tally(~variable, data=mydata))
  variable=as.factor(ifelse(freq_table[match(variable, freq_table[,1]),2]>lim
it, as.character(variable), "Other"))
  index=getIndexFromName(mydata,var_name)
  mydata[,index]=variable
  return(mydata)
}

final_political_data=rareLevelAsOther(final_political_data$Birth_COuntry, "Bi
rth_COuntry", final_political_data, 50)

tally(~final_political_data$Birth_COuntry)
```

Check for multicollinearity:

```
reg_all=lm(Left_Right_Wing~.,data=final_political_data)
vif(reg_all)

##                              GVIF Df GVIF^(1/(2*Df))
## Education                5.726085  8        1.115234
## Union_Membership         1.279795  1        1.131280
## Union_Membership_Family  1.230175  1        1.109132
```

```
## Employ_Status             3.701955  9        1.075423
## Occupation                5.099459  9        1.094730
## Public_Private_Employ     2.857943  4        1.140267
## Employ_Status_Spouse      6.464939  9        1.109255
## Household_Income          2.136925  4        1.099572
## Number_Household          1.194231  1        1.092809
## Religiousity              1.426078  3        1.060939
## Birth_COuntry            44.547211  7        1.311514
## Public_Expense_Health     2.896875  4        1.142198
## Public_Expense_Edu        2.564938  4        1.124954
## Public_Expense_Defense    2.253176  4        1.106877
## Public_Expense_Pension    2.321701  4        1.111030
## Public_Expense_Police     2.143742  4        1.100010
## Public_Expense_Welfare    2.678735  4        1.131075
## Improv_Standard           1.479967  3        1.067518
## Economy                   1.621435  2        1.128431
## Gov_Action_Dif_Income     1.652545  4        1.064803
## Satisfaction_Democrat     1.923781  3        1.115217
## Female                    1.653400  1        1.285846
```

Now I run the backward selection algorithm again and pick the model with high R-squared and low AIC

```
new_data=backwardElim(final_political_data,15,kfold=TRUE)
```

```
## [1] "Variables:"
##  [1] "Education"               "Union_Membership"
##  [3] "Union_Membership_Family" "Employ_Status"
##  [5] "Occupation"              "Public_Private_Employ"
##  [7] "Employ_Status_Spouse"    "Household_Income"
##  [9] "Number_Household"        "Religiousity"
## [11] "Birth_COuntry"           "Public_Expense_Health"
## [13] "Public_Expense_Edu"      "Public_Expense_Defense"
## [15] "Public_Expense_Police"   "Public_Expense_Welfare"
## [17] "Improv_Standard"         "Economy"
## [19] "Gov_Action_Dif_Income"   "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.211934839971681"
## [1] "AIC: 10871.0420516024"
## [1] "MSE cross-validation"
## [1] 5.055799
## [1] "Variables:"
##  [1] "Education"               "Union_Membership"
##  [3] "Union_Membership_Family" "Employ_Status"
##  [5] "Occupation"              "Public_Private_Employ"
##  [7] "Employ_Status_Spouse"    "Household_Income"
##  [9] "Religiousity"            "Birth_COuntry"
## [11] "Public_Expense_Health"   "Public_Expense_Edu"
## [13] "Public_Expense_Defense"  "Public_Expense_Police"
## [15] "Public_Expense_Welfare"  "Improv_Standard"
## [17] "Economy"                 "Gov_Action_Dif_Income"
```

```
## [19] "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.212191262228437"
## [1] "AIC: 10869.2852843427"
## [1] "MSE cross-validation"
## [1] 5.050172
## [1] "Variables:"
##  [1] "Education"               "Union_Membership"
##  [3] "Union_Membership_Family" "Employ_Status"
##  [5] "Occupation"              "Public_Private_Employ"
##  [7] "Employ_Status_Spouse"    "Household_Income"
##  [9] "Religiousity"            "Birth_COuntry"
## [11] "Public_Expense_Health"   "Public_Expense_Defense"
## [13] "Public_Expense_Police"   "Public_Expense_Welfare"
## [15] "Improv_Standard"         "Economy"
## [17] "Gov_Action_Dif_Income"   "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.210963061305591"
## [1] "AIC: 10869.2339351706"
## [1] "MSE cross-validation"
## [1] 5.055429
## [1] "Variables:"
##  [1] "Education"               "Union_Membership"
##  [3] "Union_Membership_Family" "Employ_Status"
##  [5] "Occupation"              "Public_Private_Employ"
##  [7] "Employ_Status_Spouse"    "Household_Income"
##  [9] "Religiousity"            "Birth_COuntry"
## [11] "Public_Expense_Health"   "Public_Expense_Defense"
## [13] "Public_Expense_Welfare"  "Improv_Standard"
## [15] "Economy"                 "Gov_Action_Dif_Income"
## [17] "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.208595770738724"
## [1] "AIC: 10872.6855699707"
## [1] "MSE cross-validation"
## [1] 5.075373
## [1] "Variables:"
##  [1] "Education"               "Union_Membership"
##  [3] "Union_Membership_Family" "Employ_Status"
##  [5] "Occupation"              "Public_Private_Employ"
##  [7] "Household_Income"        "Religiousity"
##  [9] "Birth_COuntry"           "Public_Expense_Health"
## [11] "Public_Expense_Defense"  "Public_Expense_Welfare"
## [13] "Improv_Standard"         "Economy"
## [15] "Gov_Action_Dif_Income"   "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.202360707927753"
## [1] "AIC: 10883.1288449988"
## [1] "MSE cross-validation"
## [1] 5.076092
## [1] "Variables:"
##  [1] "Education"               "Union_Membership_Family"
##  [3] "Employ_Status"           "Occupation"
##  [5] "Public_Private_Employ"   "Household_Income"
```

```
##  [7] "Religiousity"            "Birth_COuntry"
##  [9] "Public_Expense_Health"   "Public_Expense_Defense"
## [11] "Public_Expense_Welfare"  "Improv_Standard"
## [13] "Economy"                 "Gov_Action_Dif_Income"
## [15] "Satisfaction_Democrat"
## [1] "Adj R-squared: 0.201900984741271"
## [1] "AIC: 10883.5650622187"
## [1] "MSE cross-validation"
## [1] 5.05232
```

The second model is the best one. I choose the second model.

Reset the levels of factor variables

```
final_political_data$Household_Income <- relevel(final_political_data$Household_Income, ref="Lowest Quantile")
final_political_data$Religiousity <- relevel(final_political_data$Religiousity, ref="Not Religious")
final_political_data$Education <- relevel(final_political_data$Education, ref="Early Childhood Education")
final_political_data$Occupation <- relevel(final_political_data$Occupation, ref="Elementary Occupation")
final_political_data$Employ_Status <- relevel(final_political_data$Employ_Status, ref="Full-time")
final_political_data$Public_Private_Employ <- relevel(final_political_data$Public_Private_Employ, ref="Public")
final_political_data$Employ_Status_Spouse <- relevel(final_political_data$Employ_Status_Spouse, ref="Unemployed")
final_political_data$Public_Expense_Health <- relevel(final_political_data$Public_Expense_Health, ref="The same as now")
final_political_data$Birth_COuntry <- relevel(final_political_data$Birth_COuntry, ref="Other")

final_political_data$Public_Expense_Defense <- relevel(final_political_data$Public_Expense_Defense, ref="The same as now")
final_political_data$Public_Expense_Welfare <- relevel(final_political_data$Public_Expense_Welfare, ref="The same as now")
final_political_data$Public_Expense_Police <- relevel(final_political_data$Public_Expense_Police, ref="The same as now")
final_political_data$Public_Expense_Edu <- relevel(final_political_data$Public_Expense_Edu, ref="The same as now")
final_political_data$Improv_Standard <- relevel(final_political_data$Improv_Standard, ref="Very unlikely")
final_political_data$Economy <- relevel(final_political_data$Economy, ref="Same")
final_political_data$Satisfaction_Democrat <- relevel(final_political_data$Satisfaction_Democrat, ref="Very Satisfied")
```

Model and Result

Regression Model:

```
reg=lm(Left_Right_Wing ~ Education + Union_Membership + Union_Membership_Fami
ly + Employ_Status + Occupation + Public_Private_Employ+ Employ_Status_Spouse
 + Household_Income + Religiousity + Birth_COuntry + Public_Expense_Welfare+
Public_Expense_Police + Public_Expense_Health + Public_Expense_Defense + Publ
ic_Expense_Edu + Improv_Standard + Economy + Gov_Action_Dif_Income  + Satisfa
ction_Democrat, data=final_political_data)
```

Result

Summary table is already included in the main report.

```
sum_reg=summary(reg)
sum_reg$coefficients
#write.csv(sum_reg$coefficients, "results.csv")

anova(reg)
```

```
## Analysis of Variance Table
##
## Response: Left_Right_Wing
##                            Df  Sum Sq Mean Sq F value    Pr(>F)
## Education                   8   285.1  35.640  7.3503 1.072e-09 ***
## Union_Membership            1    40.5  40.497  8.3521 0.0038877 **
## Union_Membership_Family     1    45.2  45.190  9.3199 0.0022922 **
## Employ_Status               9   118.9  13.216  2.7256 0.0036577 **
## Occupation                  9   164.6  18.284  3.7708 9.899e-05 ***
## Public_Private_Employ       4   267.7  66.935 13.8046 3.934e-11 ***
## Employ_Status_Spouse        9   101.5  11.273  2.3250 0.0132717 *
## Household_Income            4   212.1  53.023 10.9353 8.720e-09 ***
## Religiousity                3   372.1 124.048 25.5836 2.773e-16 ***
## Birth_COuntry               7   755.5 107.922 22.2577 < 2.2e-16 ***
## Public_Expense_Welfare      4   169.8  42.457  8.7563 5.150e-07 ***
## Public_Expense_Police       4   102.5  25.628  5.2856 0.0003086 ***
## Public_Expense_Health       4   114.9  28.718  5.9228 9.682e-05 ***
## Public_Expense_Defense      4   128.1  32.032  6.6063 2.767e-05 ***
## Public_Expense_Edu          4    57.0  14.246  2.9382 0.0194894 *
## Improv_Standard             3   139.8  46.616  9.6141 2.623e-06 ***
## Economy                     2   150.0  75.005 15.4690 2.117e-07 ***
## Gov_Action_Dif_Income       4   137.8  34.449  7.1047 1.104e-05 ***
## Satisfaction_Democrat       3   245.3  81.751 16.8603 7.775e-11 ***
## Residuals                2353 11409.1   4.849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

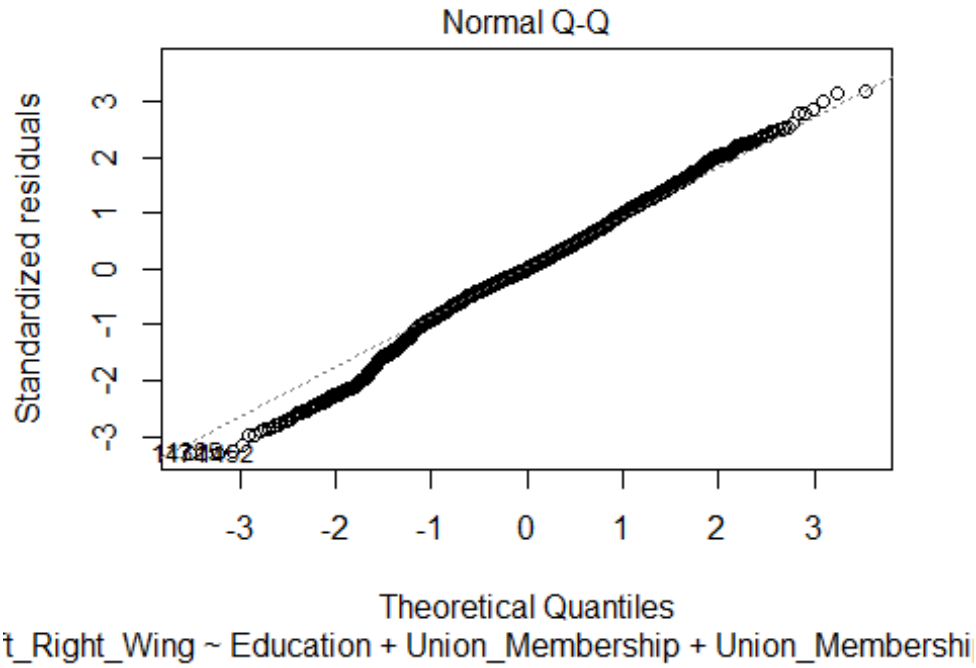
Test for Equality of Variance Assumption

```
ncvTest(reg)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1529629    Df = 1     p = 0.6957198
```

Test for Normality Assumption

```
plot(reg,which=c(2))
```



t_Right_Wing ~ Education + Union_Membership + Union_Membershi