Project Red Scare

Daniel Truver

5/03/2018

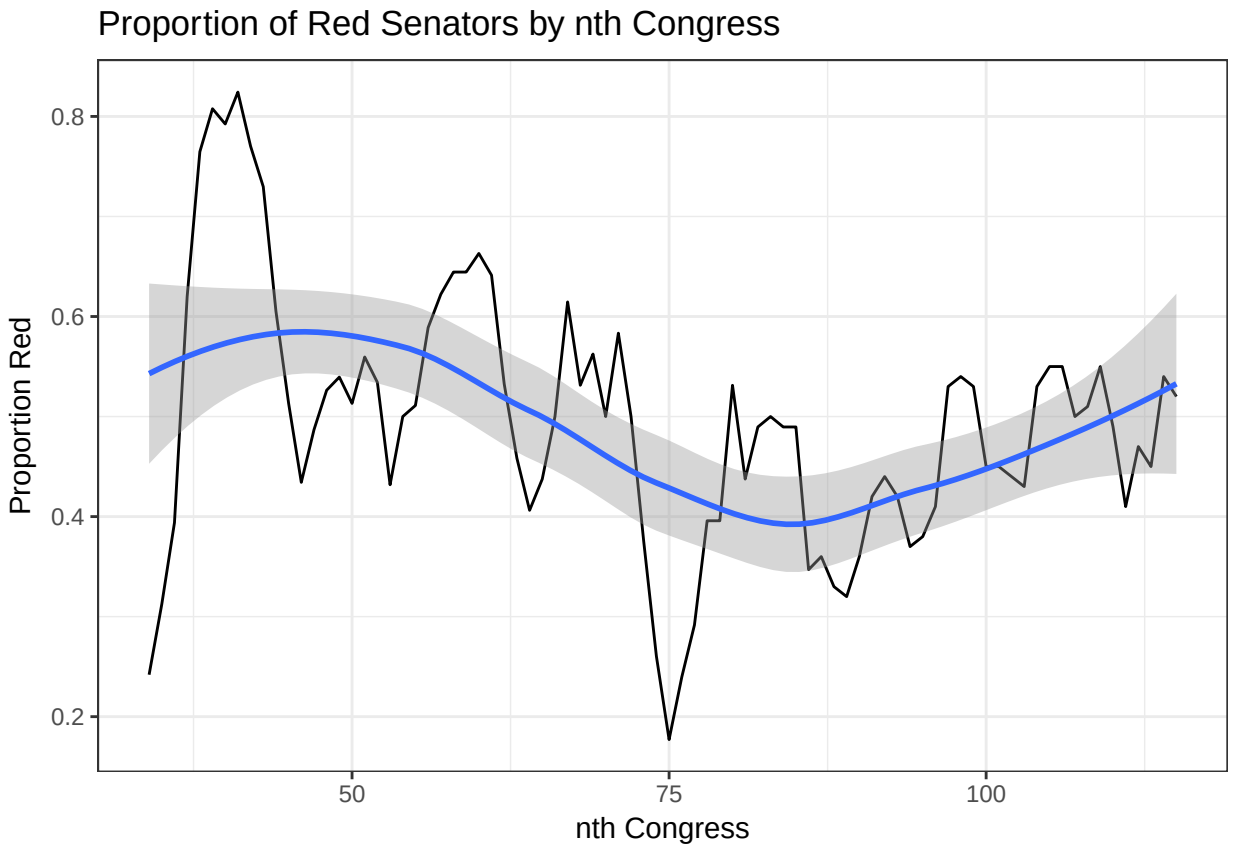
Introduction

After the presidential election of 2016, and once they got through some of the depression, democrats predicted that the 2018 midterm elections would see a blue wave. Of all the times I have heard this claim, I have never seen a statistical model that predicts it. The goal of this project is to construct a time series to model the proportion of seats held by republicans in the House and Senate of the U.S. Congress. We will construct a set of predictors for the mean and acount for additional variation with AR and MA terms. The eventual goal is to predict the proportion of republicans in the 116th Congress.

EDA

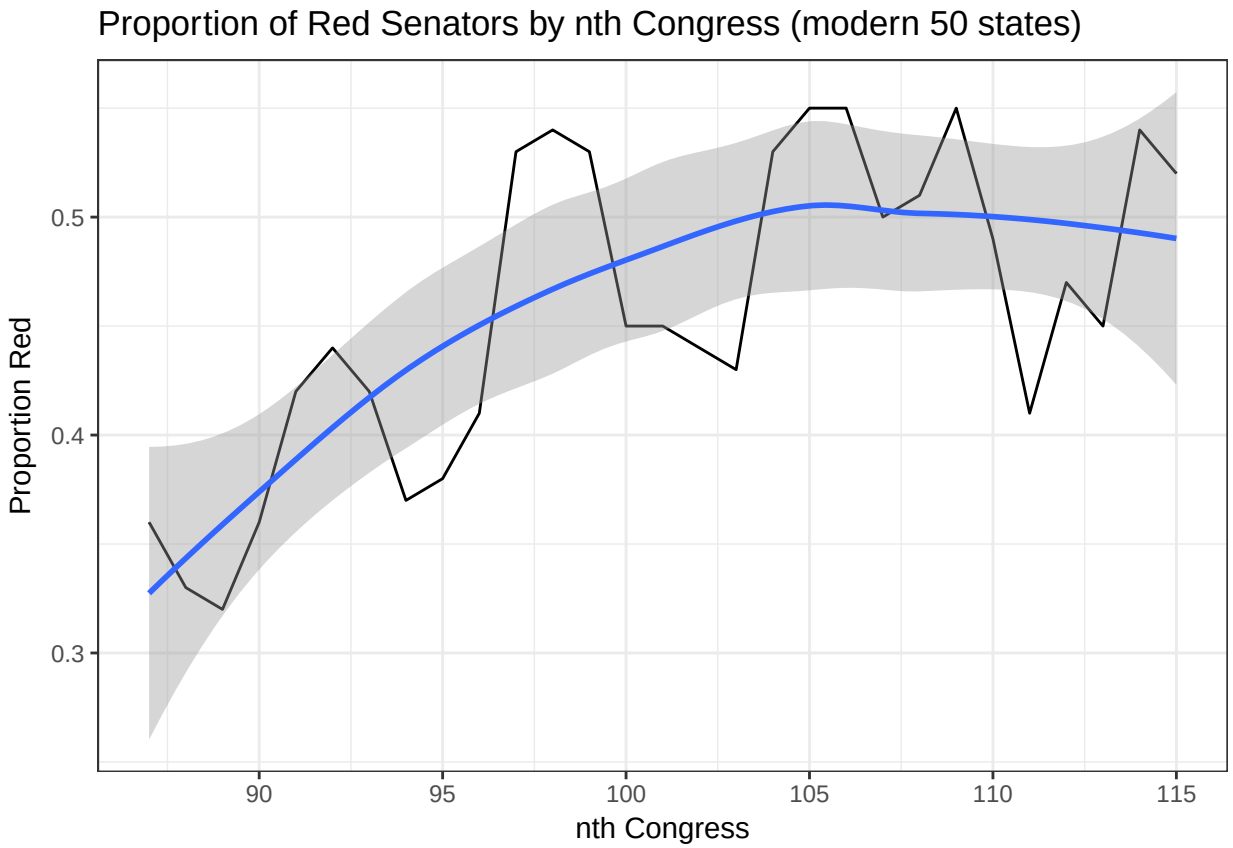
The following data come from the Brooking's institute and were last update in September of 2017. Our first concern should be if a time series model is appropriate. We also want to know if the intended process is stationary and what seasonal trends appear.

```
partisan = read.csv("partisan_congress.csv", stringsAsFactors = FALSE)
partisan$Number.of.representatives[53:54] = c(436,437)
partisan = partisan[1:82,]
for (col in names(partisan) %>% .[!.%in%c("year")] ){
  partisan[,col] = partisan[,col] %>%
    gsub("[^0-9\\.]", "", .) %>%
    as.integer()
}
partisan.full = partisan = partisan %>%
  mutate(red_sen = Republican_sen/Number.of.senators) %>%
  mutate(red_rep = Republican_rep/Number.of.representatives) %>%
  mutate(year_in = as.integer(str_extract(year, "\\d\\d\\d\\d")))
ggplot(data = partisan,
       aes(x = congress, y = red_sen)) +
  geom_line() +
  ggtitle("Proportion of Red Senators by nth Congress") +
  xlab("nth Congress") + ylab("Proportion Red") +
  geom_smooth()
```



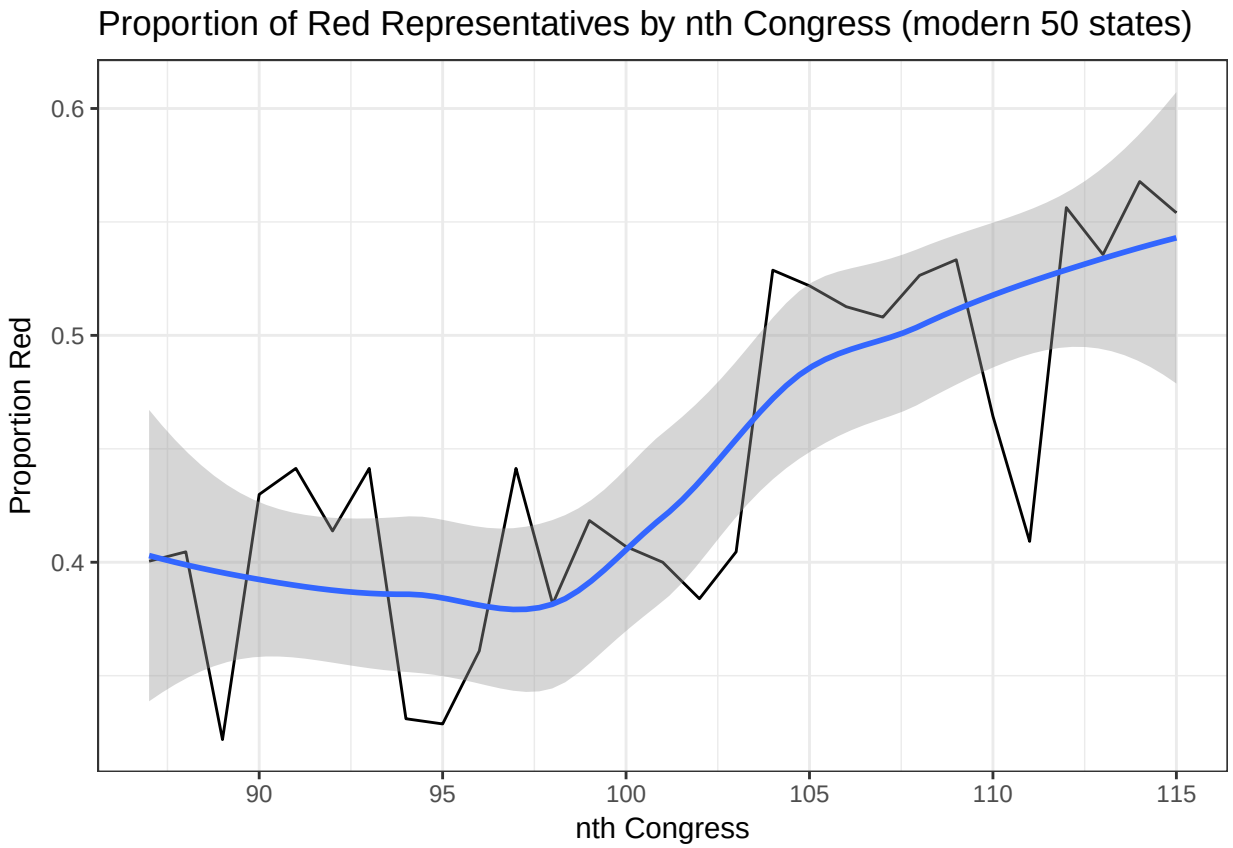
The above figure shows the proportion of Senate seats occupied by republicans. There does seem to be dependence in time. This figure is all years for which we have data, but moving forward, we will only incorporate years for which all 50 modern states have representatives in the Senate. The adjusted figure is below.

```
partisan = partisan %>% filter(Number.of.senators == 100)
ggplot(data = partisan,
       aes(x = congress, y = red_sen)) +
  geom_line() +
  ggtitle("Proportion of Red Senators by nth Congress (modern 50 states)") +
  xlab("nth Congress") + ylab("Proportion Red") +
  geom_smooth()
```



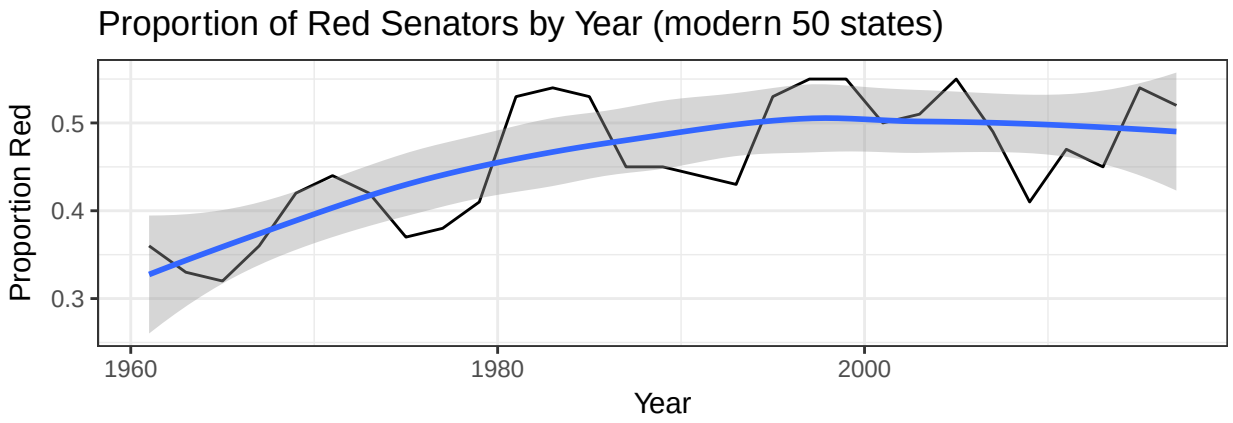
We lose a large fraction of our data by making this change, but we are most interested in modeling the modern trends in Congress anyway, and our target predictions involve all 50 states. We see a different overal trend in these data than we saw when including all previous years. There does still appear to be autocorrelation. Let's take a look at the House.

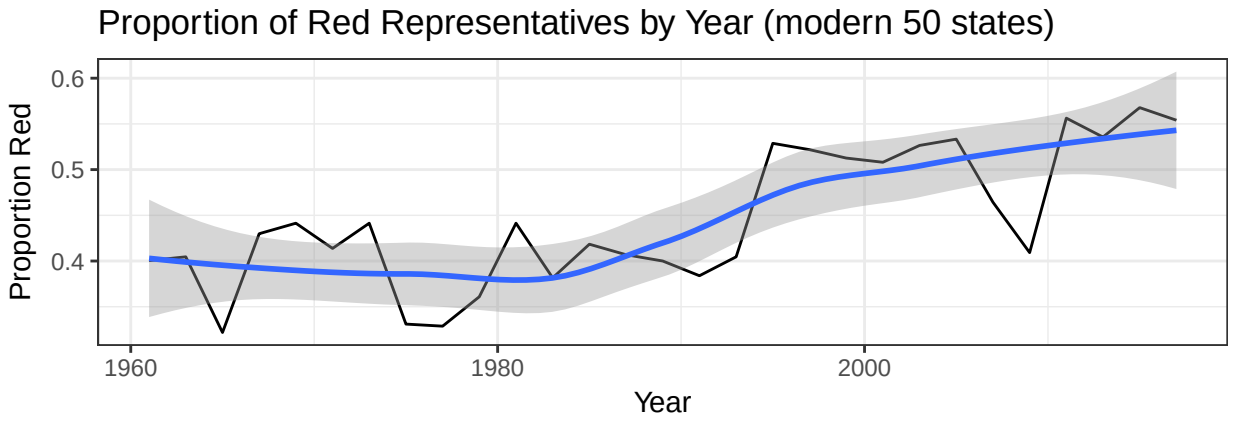
```
ggplot(data = partisan,
       aes(x = congress, y = red_rep)) +
  geom_line() +
  ggtitle("Proportion of Red Representatives by nth Congress (modern 50 states)") +
  xlab("nth Congress") + ylab("Proportion Red") +
  geom_smooth()
```



The trend is incresing as with the senate, but the pattern is not the same. The trend by year for both bodies is below.

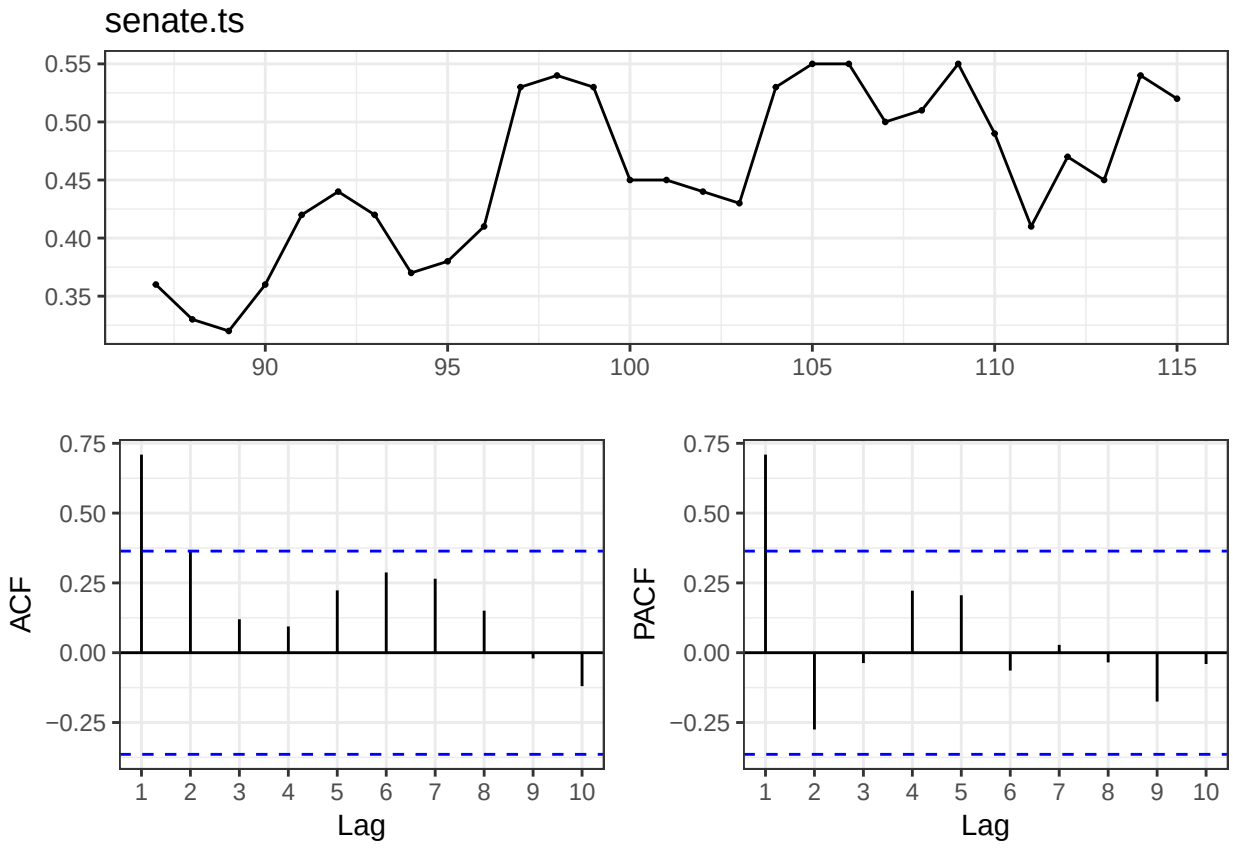
```
library(gridExtra)
g_rep = ggplot(data = partisan,
       aes(x = year_in, y = red_rep)) +
  geom_line() +
  ggtitle("Proportion of Red Representatives by Year (modern 50 states)") +
  xlab("Year") + ylab("Proportion Red") +
  geom_smooth()
g_sen = ggplot(data = partisan,
       aes(x = year_in, y = red_sen)) +
  geom_line() +
  ggtitle("Proportion of Red Senators by Year (modern 50 states)") +
  xlab("Year") + ylab("Proportion Red") +
  geom_smooth()
grid.arrange(g_sen, g_rep, nrow = 2)
```



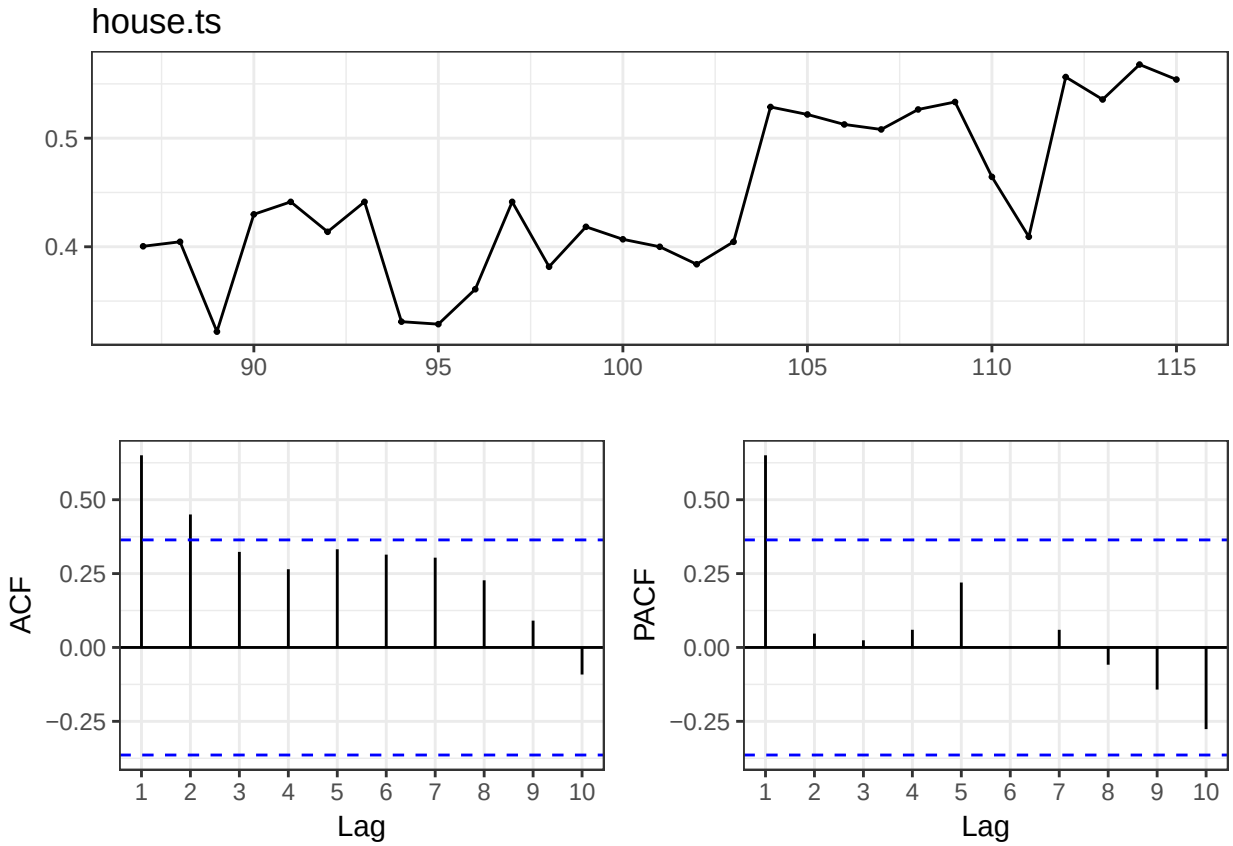


Technical EDA

```
library(forecast)
senate.ts = ts(data = partisan %>% select(red_sen), start = 87)
ggtsdisplay(senate.ts)
```




```
house.ts = ts(data = partisan %>% select(red_rep), start = 87)
ggtsdisplay(house.ts)
```

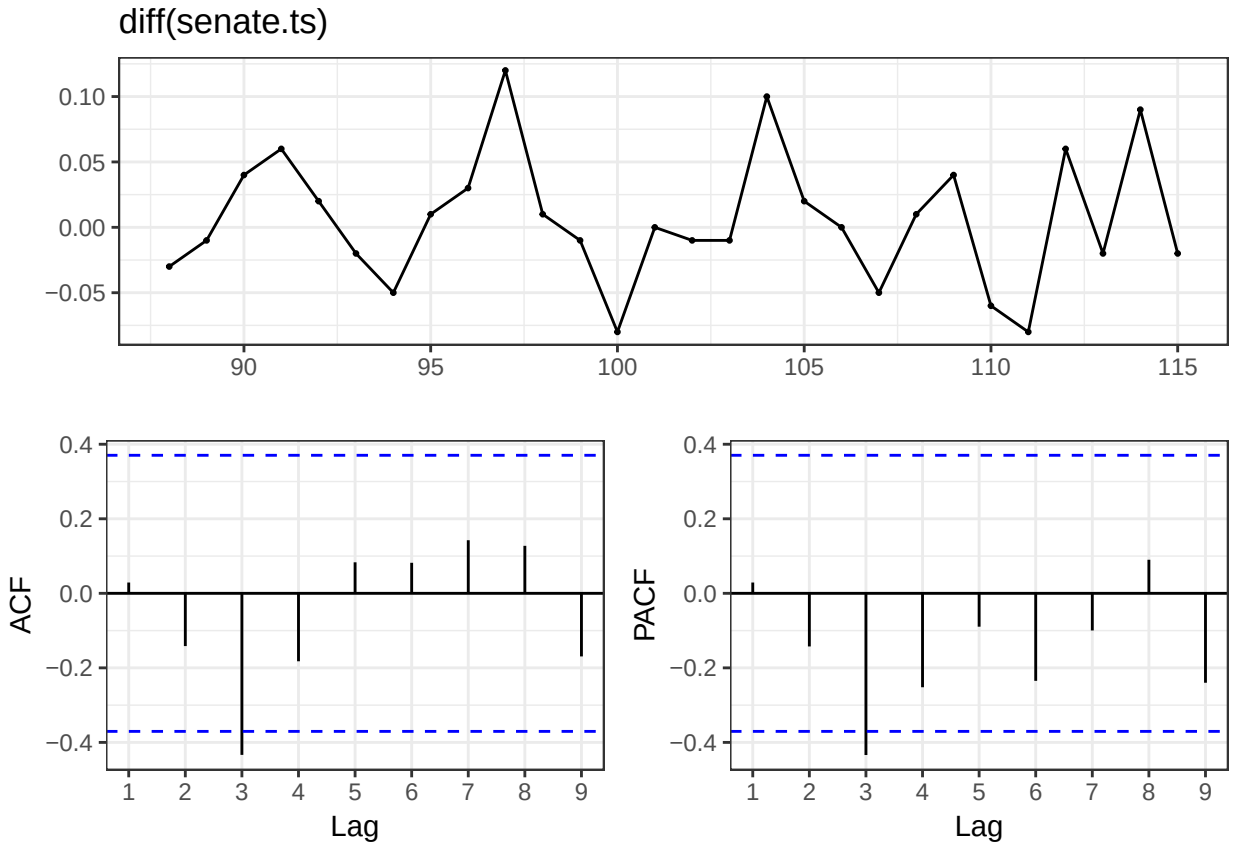


There is autocorrelation in both of these series, but it does not look the same. There could also be some stationarity issues. We're going to ignore R's suggested significance levels and just look at the ACF and PACF spikes; we have to remember that we do not have an abundance of data. The Senate's ACF has a strange wave pattern to it, but it seems to die out at lag 2. Similary, the Senate's PACF spikes are lag 1 and lag 2.

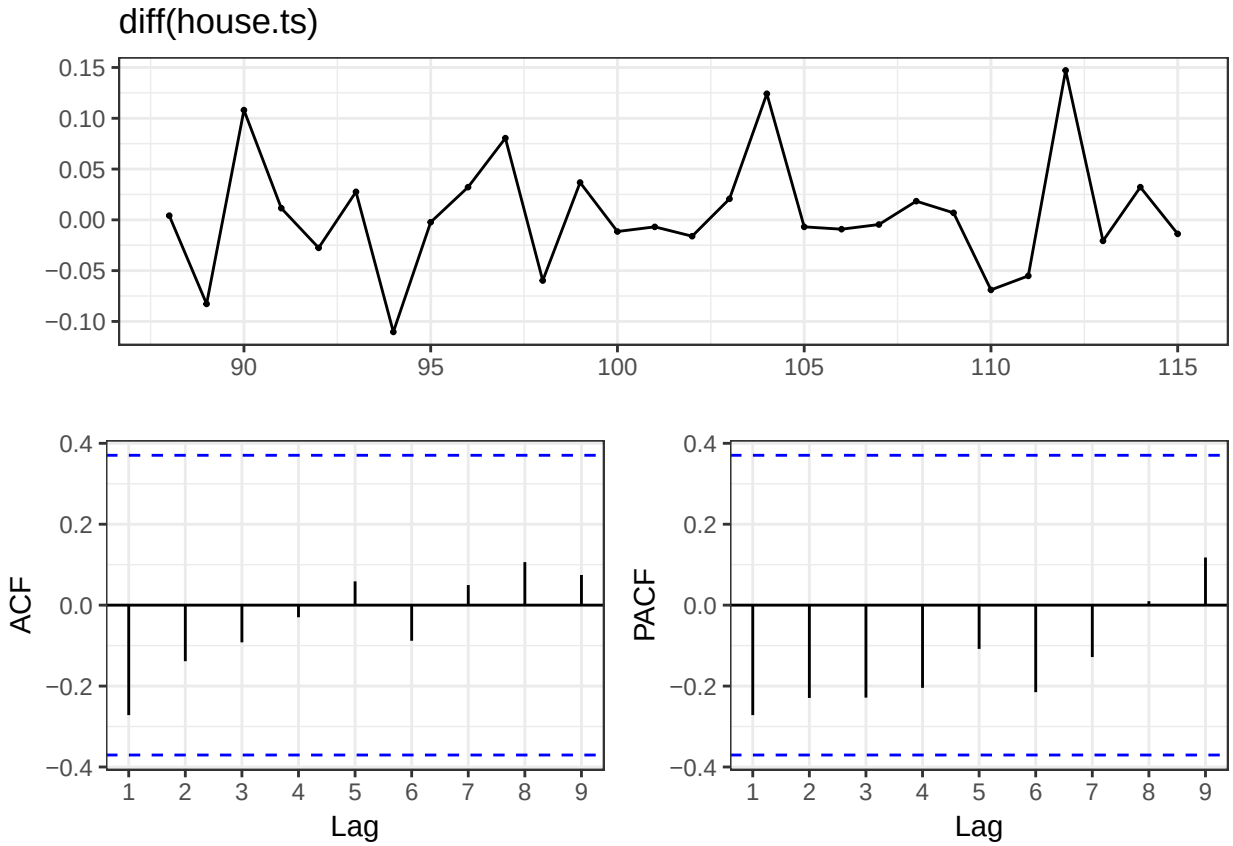
The House's ACF takes a lot longer to die off to the same level as the Senate's, and it still has the wave pattern. The House also shows PACF spike at 5 and 10, possibly a seasonal component?

We attempt differencing.

```
ggtsdisplay(diff(senate.ts))
```

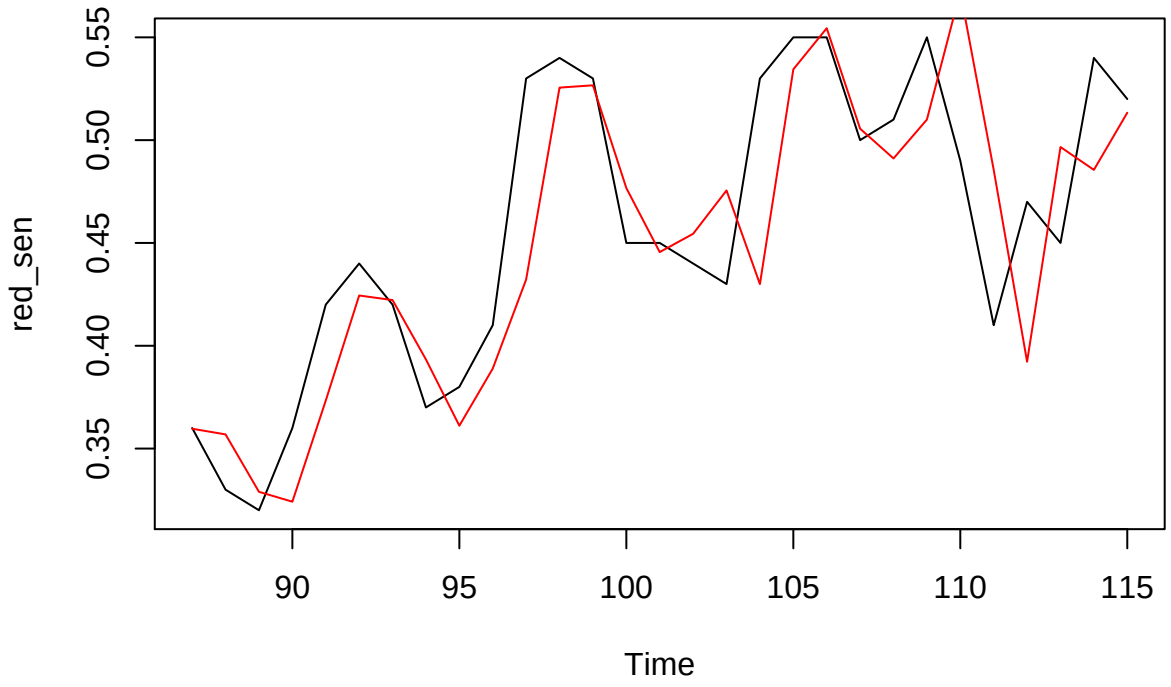



```
ggtsdisplay(diff(house.ts))
```

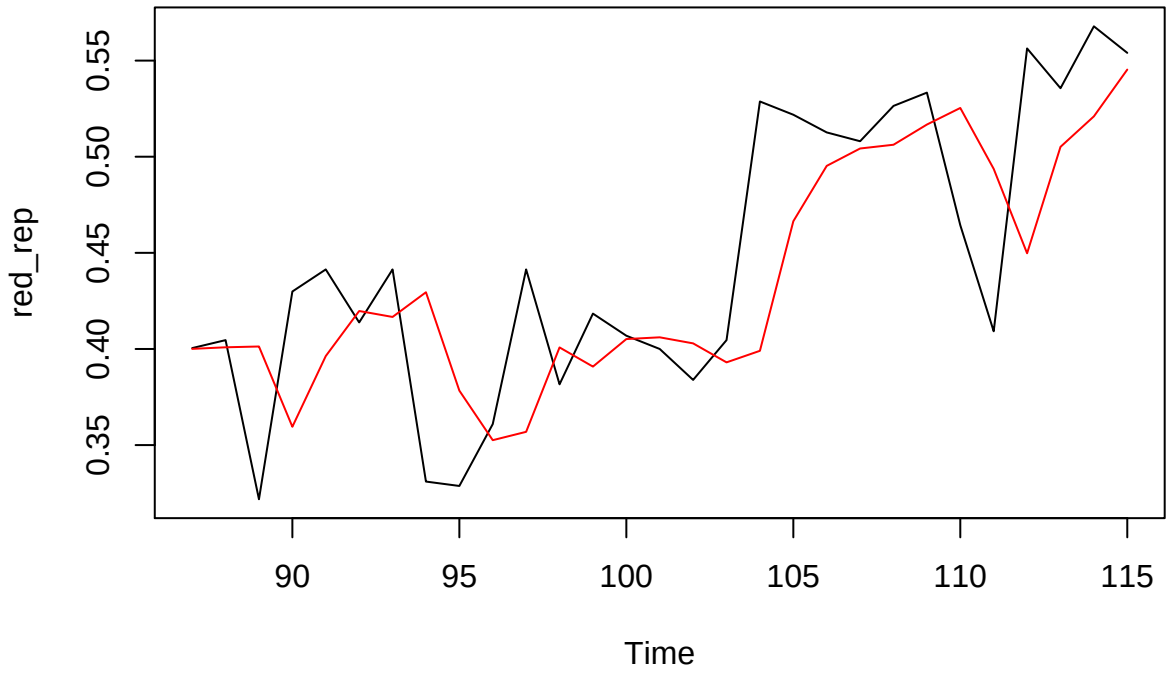


Differencing the Senate seems to reveal as seasonal component every 3 congresses. If we recall our high school government classes, this makes sense; Senate seats only come up for re-election every six years (3 session of Congress).

```
Senate.arima = Arima(senate.ts, order = c(0,1,0), seasonal = list(order = c(1,0,0), period = 3))
{
  plot(senate.ts)
  # points(time(fitted(basicSenate))-deltat(basicSenate), basicSenate$fitted,
  #        col = "red", type = "l")
  points(Senate.arima$fitted, col = "red", type = "l")
}
```




```
basicHouse = Arima(house.ts, order = c(0,1,1))
{
  plot(house.ts)
  points(basicHouse$fitted, col = "red", type = "l")
}
```



From these initial, basic models, we see much better results for the Senate than we do for the house. The fit is also not as tight as we would like on either. Now begins the hard part, finding other predictors.

Other Predictors

What influences the election of Congress? We begin with the ones we hear about consistently: what is the party affiliation of the president, and is it a midterm election?

```
partisan$midterm = c(rep(c(0,1), 14), 0)
partisan$red_pres = c(0,0,0,0,1,1,1,1,0,0,1,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1)
partisan$red_pres_in_election = c(1,partisan$red_pres[-29])
```

The columns created above indicate whether voters chose the congress in a midterm election, whether the president who served at the same time as the Congress was a republican, and whether the president at the time of the Congress's election was republican. We will fit a linear model to see what happens.

```
senateByPres = lm(red_sen ~ red_pres + red_pres_in_election + midterm + red_pres_in_election:midterm,
                  data = partisan)
houseByPres = lm(red_rep ~ red_pres + red_pres_in_election + midterm + red_pres_in_election:midterm,
                 data = partisan)
summary(senateByPres)
```

```
##
## Call:
## lm(formula = red_sen ~ red_pres + red_pres_in_election + midterm +
##     red_pres_in_election:midterm, data = partisan)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.125714 -0.044231  0.005769  0.047143  0.122308
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.42769    0.03363  12.717 3.73e-12 ***
## red_pres                       0.07404    0.03633   2.038   0.0527 .
## red_pres_in_election          -0.02346    0.03633  -0.646   0.5245
## midterm                        0.02802    0.04279   0.655   0.5188
## red_pres_in_election:midterm -0.04343    0.06503  -0.668   0.5105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07001 on 24 degrees of freedom
## Multiple R-squared:  0.1681, Adjusted R-squared:  0.0295
## F-statistic: 1.213 on 4 and 24 DF,  p-value: 0.3313
```

```
summary(houseByPres)
```

```
##
## Call:
## lm(formula = red_rep ~ red_pres + red_pres_in_election + midterm +
##     red_pres_in_election:midterm, data = partisan)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.126826 -0.050246  0.001432  0.048933  0.111002
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.44866    0.03357  13.364  1.3e-12 ***
## red_pres                       0.04587    0.03626   1.265    0.218
## red_pres_in_election          -0.05459    0.03626  -1.505    0.145
```

```
## midterm                        0.03147    0.04272   0.737    0.468
## red_pres_in_election:midterm -0.05598    0.06491  -0.862    0.397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06989 on 24 degrees of freedom
## Multiple R-squared:  0.2303, Adjusted R-squared:  0.102
## F-statistic: 1.795 on 4 and 24 DF,  p-value: 0.1628
```

To my surprise, the only covariate with a reasonable confidence interval is `red_pres` in the senate; the indication is that the Senate tends to lean republican when the president is a republican. There was not a significant relationship between the red proportion of the Congress and the affiliation of the president in office at the time of Congressional elections. Midterm elections also did not have a significant relationship to the composition of either branch. We need to expand our collection of predictors.

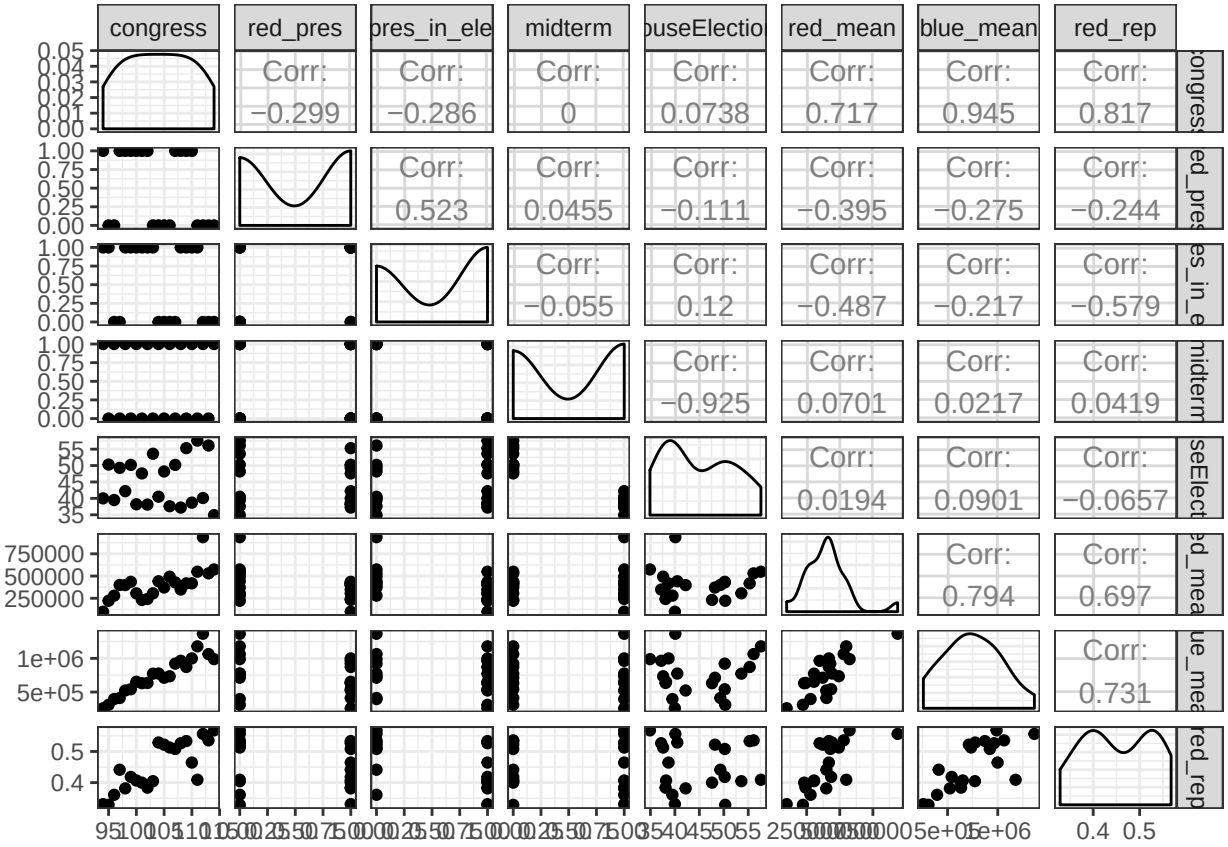
```
# source: the Brookings institute
# 115th results from the Office of the Clerk of the House of Representatives
house_by_region = read_csv("vitalstats_ch1_tbl3.csv") %>%
  filter(Congress >= 87) %>%
  filter(Party == "R") %>%
  filter(Region != "Total Seats") %>%
  .[order(.$Congress),]
house_by_region$Percent = as.numeric(house_by_region$Percent)
house_election_turnout = read_csv("vitalstats_ch2_tbl1.csv") %>%
  select(Year, HouseElections) %>%
  filter(Year >= 1960) %>%
  mutate(congress = 86+1:nrow(.))
house_election_turnout[29,] = c(2016,54.7,115)
partisan = inner_join(partisan, house_election_turnout, by = "congress")
pres_party_mid_loss = read_csv("vitalstats_ch2_tbl4.csv") %>%
  filter(Year >= 1960)
spending_house = read_csv("vitalstats_ch3_tbl2.csv") %>%
  filter(CandidateType == "All", ExpenditureType %in% c("MeanDem", "MeanRep"))
spending_house_red = spending_house %>%
  filter(ExpenditureType == "MeanRep") %>%
  .[order(.$Year),]
spending_house_blue = spending_house %>%
  filter(ExpenditureType == "MeanDem") %>%
  .[order(.$Year),]
spending_house = data_frame(congress = 93+1:nrow(spending_house_red),
                            Year = spending_house_red$Year, red_mean = spending_house_red$Dollars,
                            blue_mean = spending_house_blue$Dollars)
spending_senate = read_csv("vitalstats_ch3_tbl5.csv") %>%
  filter(IncumbStatus == "All", ExpenseType %in% c("MeanDem", "MeanRep"))
spending_senate_red = spending_senate %>%
  filter(ExpenseType == "MeanRep") %>%
  .[order(.$Year),]
spending_senate_blue = spending_senate %>%
  filter(ExpenseType == "MeanDem") %>%
  .[order(.$Year),]
spending_senate = data_frame(congress = 93+1:nrow(spending_senate_red),
                            Year = spending_senate_red$Year, red_mean = spending_senate_red$Dollars,
                            blue_mean = spending_senate_blue$Dollars)
```

This is a selection from the hundred or so datasets available from the Brookings Institue vital statistics on

Congress. These particular chapters were the most relevant to elections, which are the focus of this project. Much of the data is not available for all years of interest, so we will need multiple models to come together in an ensemble at the end.

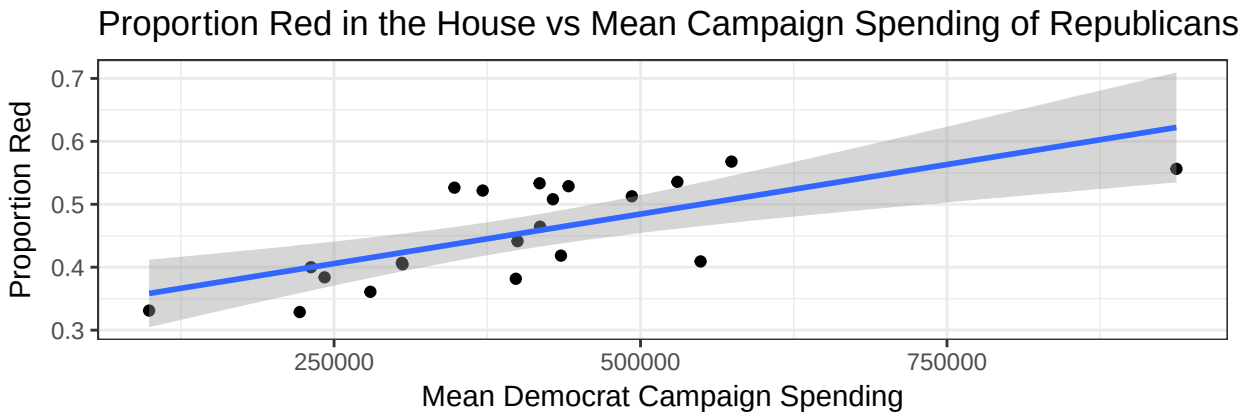
The House is up first since it has more pathologies to explain than the Senate. We begin by looking at voter turnout and candidate funding. The data we have for these variables starts at 1974. We must be careful to not let p approach too close to n given the limited amount of data.

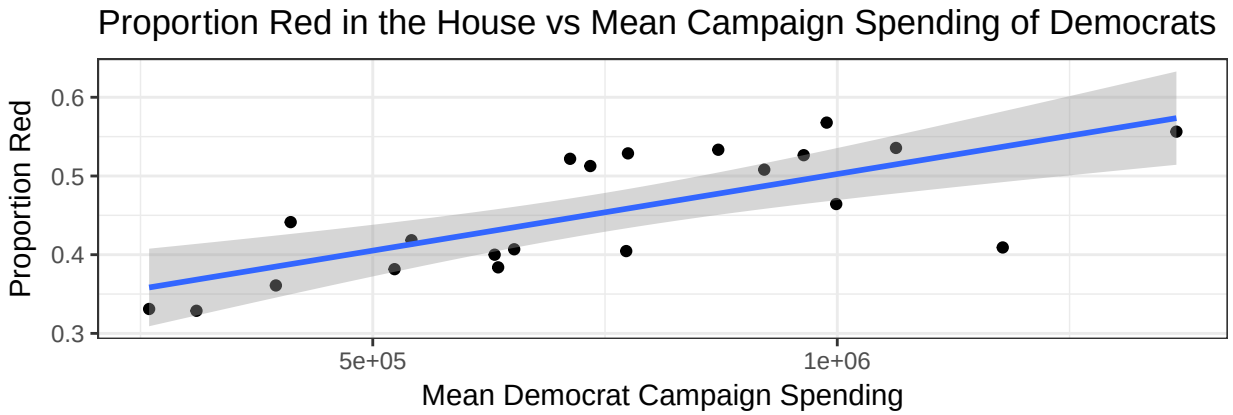
```
from1974_df_house = partisan %>%
  filter(94 <= congress & congress <= 114) %>%
  cbind(.,red_mean = spending_house$red_mean, blue_mean = spending_house$blue_mean) %>%
  select(congress,red_pres,red_pres_in_election,midterm,HouseElections,red_mean,blue_mean,red_rep)
ggpairs(data = from1974_df_house)
```



We should probably highlight these two relations.

```
g_red = ggplot(data = from1974_df_house, aes(x = red_mean, y = red_rep)) +
  geom_point() + geom_smooth(method = "lm") +
  ggtitle("Proportion Red in the House vs Mean Campaign Spending of Republicans") +
  xlab("Mean Democrat Campaign Spending") + ylab("Proportion Red")
g_blue = ggplot(data = from1974_df_house, aes(x = blue_mean, y = red_rep)) +
  geom_point() + geom_smooth(method = "lm") +
  ggtitle("Proportion Red in the House vs Mean Campaign Spending of Democrats") +
  xlab("Mean Democrat Campaign Spending") + ylab("Proportion Red")
grid.arrange(g_red, g_blue, nrow = 2)
```

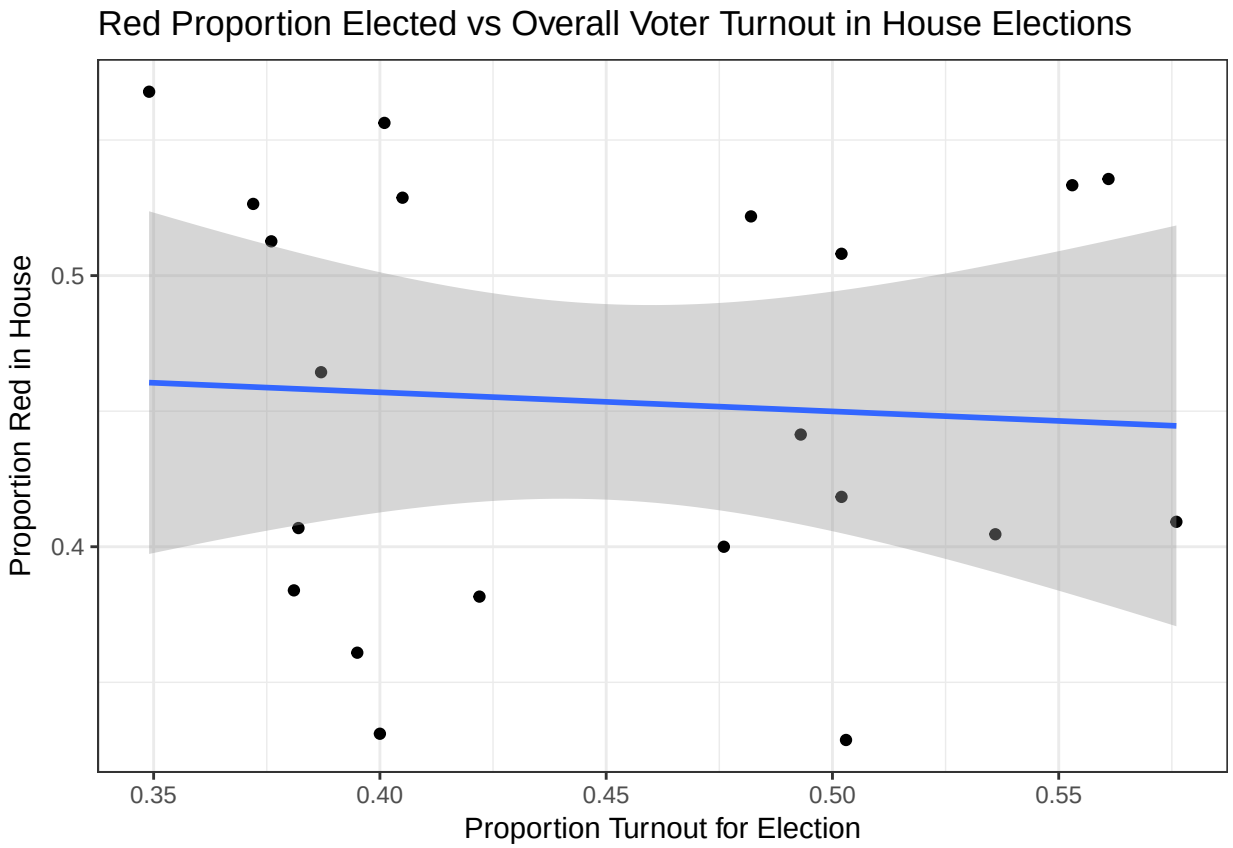




The first graph may seem to indicate that Republican represenation increases with an increase in campaign spending, but we must keep in mind that this is a correlative relationship as Republican representation also increases as Democrtic spending increases. Democrats also tend to spend more money than republicans for House elections on average. A point of high leverage for the republicans occurs in the 2010 midterm elections.

One thing we should also note is this relationship between voter turnout and red represenation in the house.

```
ggplot(data = from1974_df_house, aes(x = HouseElections/100, y = red_rep)) +
  geom_point() + geom_smooth(method = "lm") +
  ggtitle("Red Proportion Elected vs Overall Voter Turnout in House Elections") +
  xlab("Proportion Turnout for Election") + ylab("Proportion Red in House")
```



What we should take away from this graph is that voter turnout does not display any relation to the composition of the House elected by said voters. Whether this is an indictment of the "silent majority" theory is left to the reader to interpret. We should also note that the large gap of no points between 0.425 and 0.475 in the figure above is the split in turnout between midterm and presidential election years.

Now let's hit this monster with an linear model and see what happens.

```
from1974_lm_house = lm(red_rep ~ red_pres+red_pres_in_election+red_mean+
                         midterm + HouseElections,
                       data = from1974_df_house)
summary(from1974_lm_house)
```

```
##
## Call:
## lm(formula = red_rep ~ red_pres + red_pres_in_election + red_mean +
##     midterm + HouseElections, data = from1974_df_house)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.070820 -0.041587 -0.008972  0.037818  0.095253
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.321e-01  2.610e-01   2.039  0.05947 .
## red_pres                 2.404e-02  3.090e-02   0.778  0.44868
## red_pres_in_election -5.148e-02  3.459e-02  -1.488  0.15741
## red_mean                 2.815e-07  9.123e-08   3.086  0.00753 **
## midterm                 -4.754e-02  7.221e-02  -0.658  0.52028
```

```
## HouseElections       -3.331e-03  5.247e-03  -0.635  0.53512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05642 on 15 degrees of freedom
## Multiple R-squared:  0.5984, Adjusted R-squared:  0.4645
## F-statistic:  4.47 on 5 and 15 DF,  p-value: 0.01079
```

With this set of predictors, our most significant correlation to `red_rep` is `red_mean`, the variable indicating the mean spending of republican candidates in the election. Note that the value may appear small, but it is in dollars. Unfortunately, if we throw `blue_mean`, the democratic spending, into the model, we get:

```
from1974_lm_house = lm(red_rep ~ red_pres+red_pres_in_election+blue_mean+
                         midterm + HouseElections,
                       data = from1974_df_house)
summary(from1974_lm_house)
```

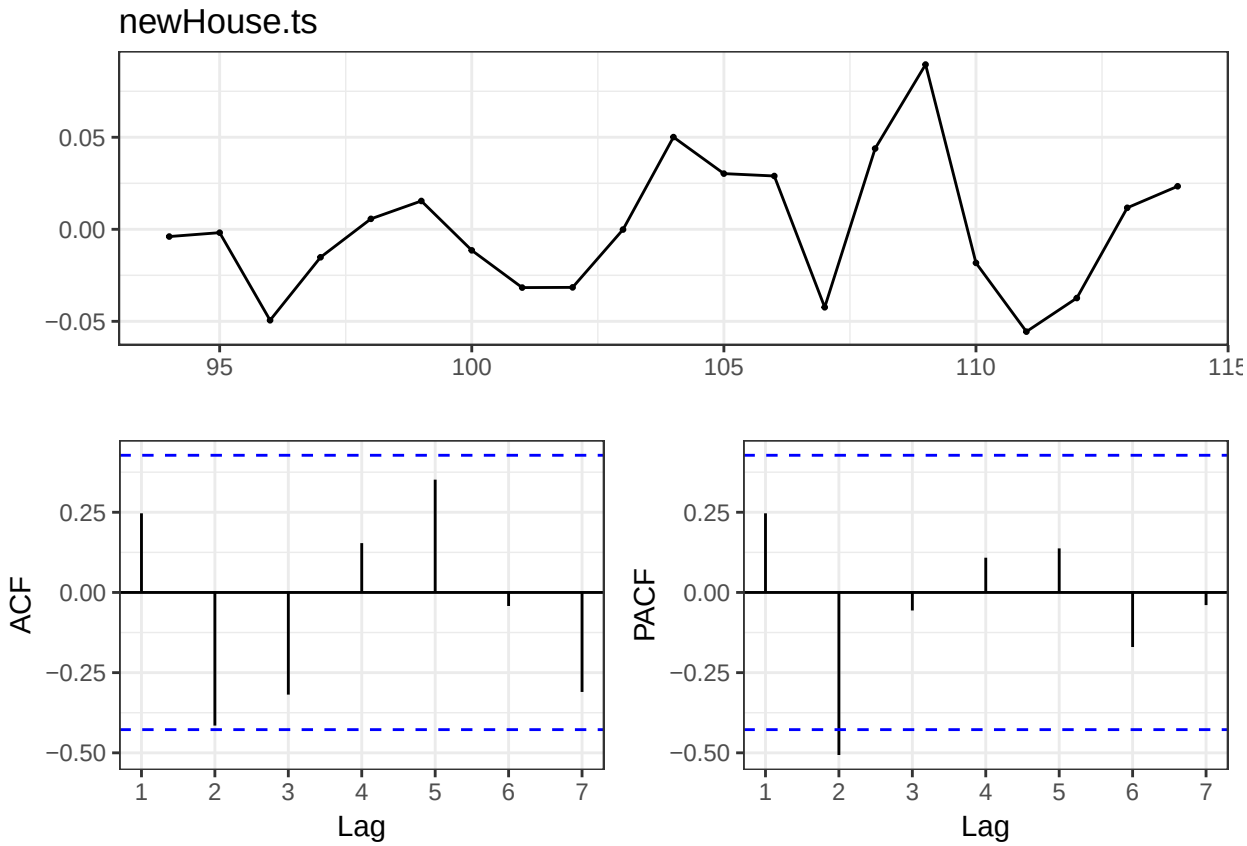
```
##
## Call:
## lm(formula = red_rep ~ red_pres + red_pres_in_election + blue_mean +
##     midterm + HouseElections, data = from1974_df_house)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055614 -0.031556 -0.001834  0.023375  0.089519
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.680e-01  1.947e-01   2.918  0.01060 *
## red_pres               2.758e-02  2.308e-02   1.195  0.25060
## red_pres_in_election -7.456e-02  2.362e-02  -3.157  0.00651 **
## blue_mean              1.919e-07  3.561e-08   5.388 7.53e-05 ***
## midterm               -5.895e-02  5.330e-02  -1.106  0.28615
## HouseElections        -4.422e-03  3.888e-03  -1.137  0.27321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04211 on 15 degrees of freedom
## Multiple R-squared:  0.7764, Adjusted R-squared:  0.7018
## F-statistic: 10.41 on 5 and 15 DF,  p-value: 0.0001852
```

Democratic spending is also positively correlated with more republicans in the house. This model, however, also shows a significant negative correlation between an incumbant red president at the time of the election and proportion of red represenation in the House.

The interpretation for the above models, with respect to money, seems to be that when there is more money involved, republicans are more likely to gain seats in the House. Again, this is not a causal model. Democrats should not look at this and think, "well, if we spend less money, the republicans will lose ground."

Let's take a look at the time series for this subset of the data for which we have the additional predictors.

```
house_structure = from1974_lm_house$fitted.values
newHouse.ts = ts(data = from1974_df_house$red_rep - house_structure,
                 start = min(from1974_df_house$congress))
ggtsdisplay(newHouse.ts)
```



Would you look at that, it's not entirely awful like it was before. Let's see what R thinks of it.

```
newHouse.arima = auto.arima(newHouse.ts)
newHouse.arima
```

```
## Series: newHouse.ts
## ARIMA(0,0,0) with zero mean
##
## sigma^2 estimated as 0.001266:  log likelihood=40.25
## AIC=-78.51   AICc=-78.3   BIC=-77.46
```

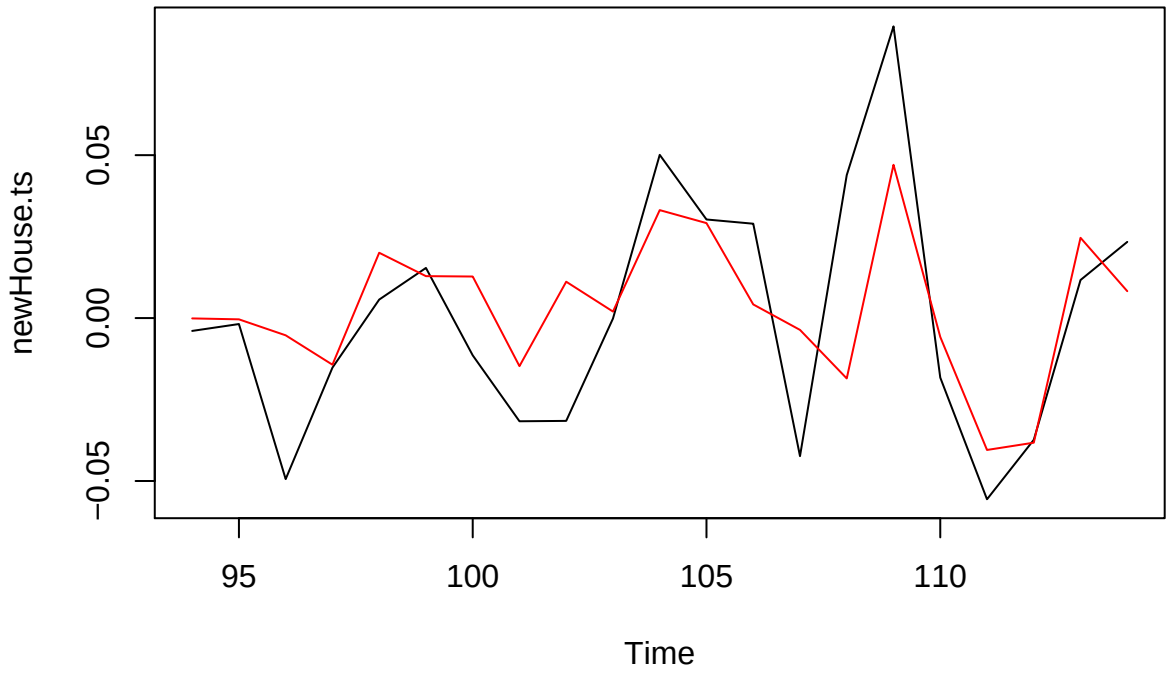
Apparently nothing. Here are some configurations I tried.

```
newHouse.arima = Arima(newHouse.ts,order = c(0,0,2),
                       seasonal = list(order = c(1,0,1), period = 3 ))
newHouse.arima
```

```
## Series: newHouse.ts
## ARIMA(0,0,2)(1,0,1)[3] with non-zero mean
##
## Coefficients:
##          ma1      ma2    sar1     sma1    mean
##       0.3950  -0.4944  0.4662  -0.9999  0.0014
## s.e.  0.2707   0.2713  0.2660   0.5065  0.0033
##
## sigma^2 estimated as 0.0008602:  log likelihood=44.69
## AIC=-77.38   AICc=-71.38   BIC=-71.11
```

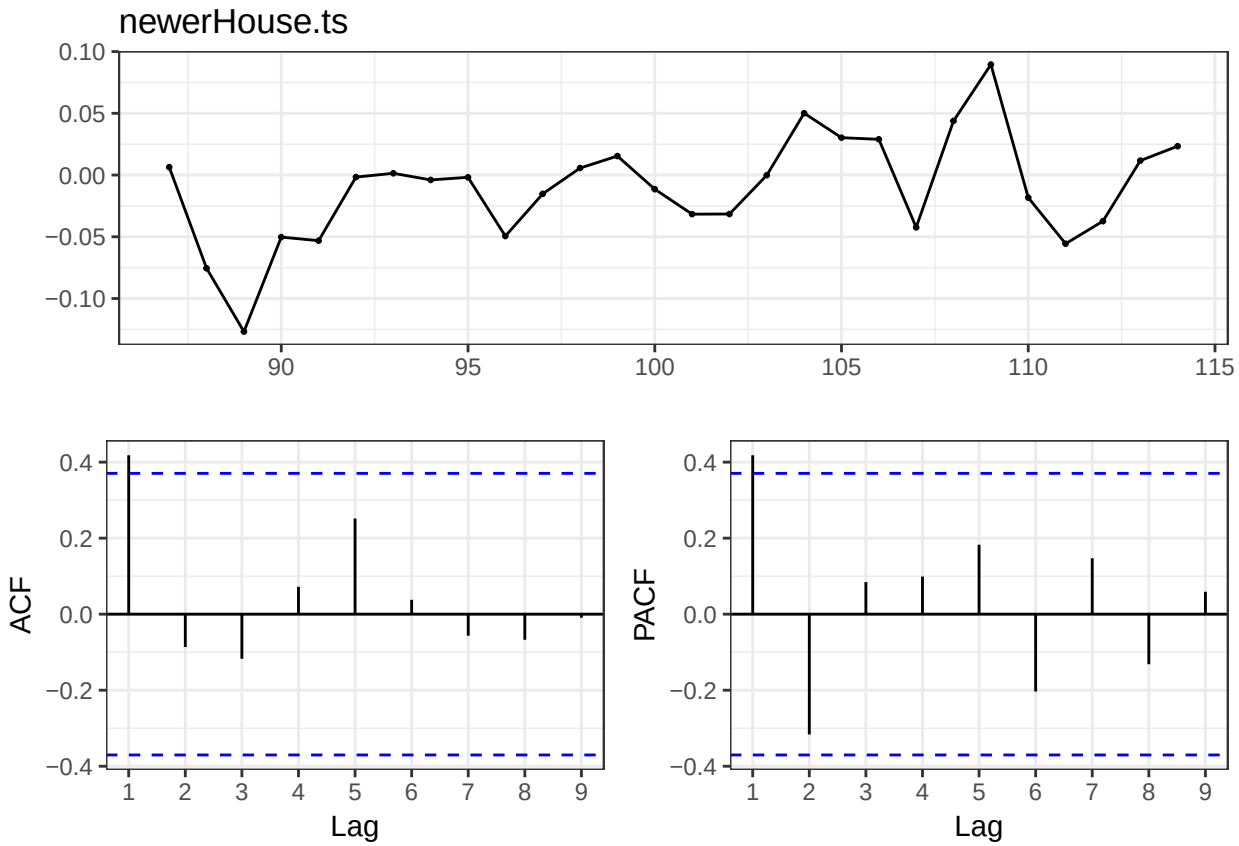
```
{
  plot(newHouse.ts)
```

```
  points(newHouse.arima$fitted, col = "red", type = "l")
}
```



Unfortunately, these predictors only reach from the 94th to the 114th Congress. We will have to fall back on our more basic model for the remaining observations. Congresses 87 to 93 we fit with out previous model based on the afiliation of the president and the timing of the election. Congresses 94 to 114 we fit with the 1974 model and its added predictors.

```
firstModelObs = c(which(partisan$congress < 94))
firstModelStructure = houseByPres$fitted.values[firstModelObs]
house_structure_full = c(firstModelStructure, house_structure)
newerHouse.ts = ts(data = partisan$red_rep[-29] - house_structure_full,
                   start = min(partisan$congress))
ggtsdisplay(newerHouse.ts)
```

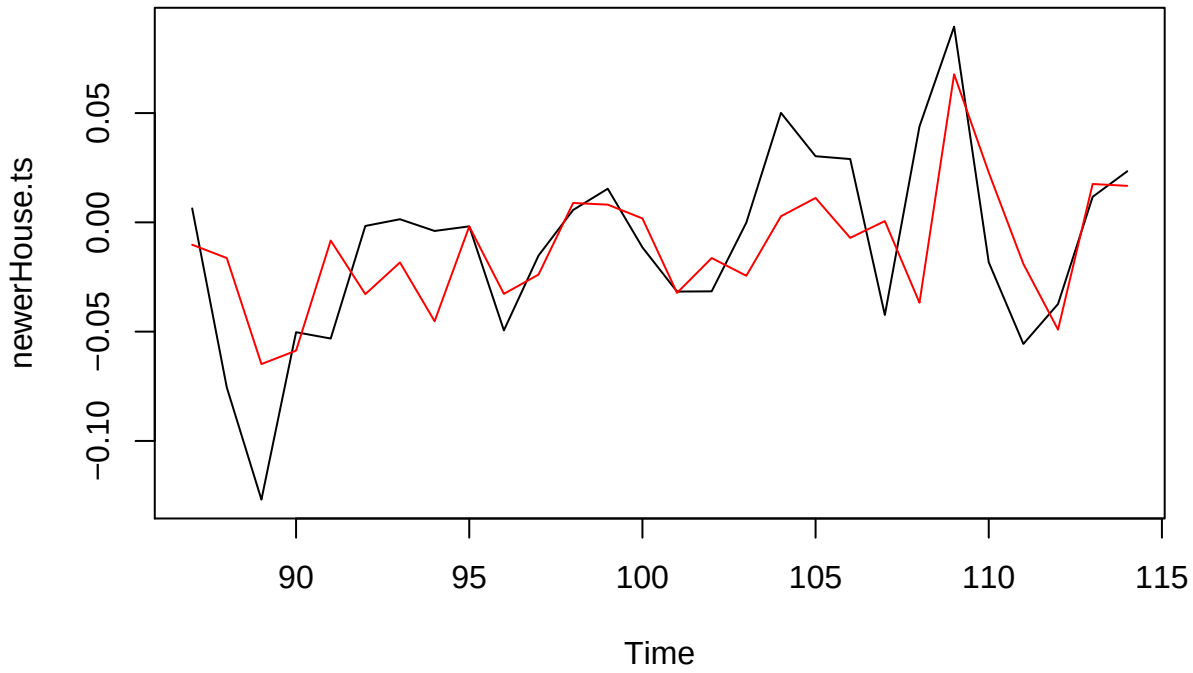


Now, some more ARIMAs fit to this data.

```
newerHouse.arima = Arima(newerHouse.ts, order = c(0,0,1),
                         seasonal = list(order = c(0,0,1), period = 5))
newerHouse.arima
```

```
## Series: newerHouse.ts
## ARIMA(0,0,1)(0,0,1)[5] with non-zero mean
##
## Coefficients:
##          ma1    sma1     mean
##       0.6536  0.4868  -0.0157
## s.e.  0.1623  0.2557   0.0147
##
## sigma^2 estimated as 0.001207:  log likelihood=54.92
## AIC=-101.84   AICc=-100.11   BIC=-96.52
```

```
{
  plot(newerHouse.ts)
  points(newerHouse.arima$fitted, col = "red", type = "l")
}
```

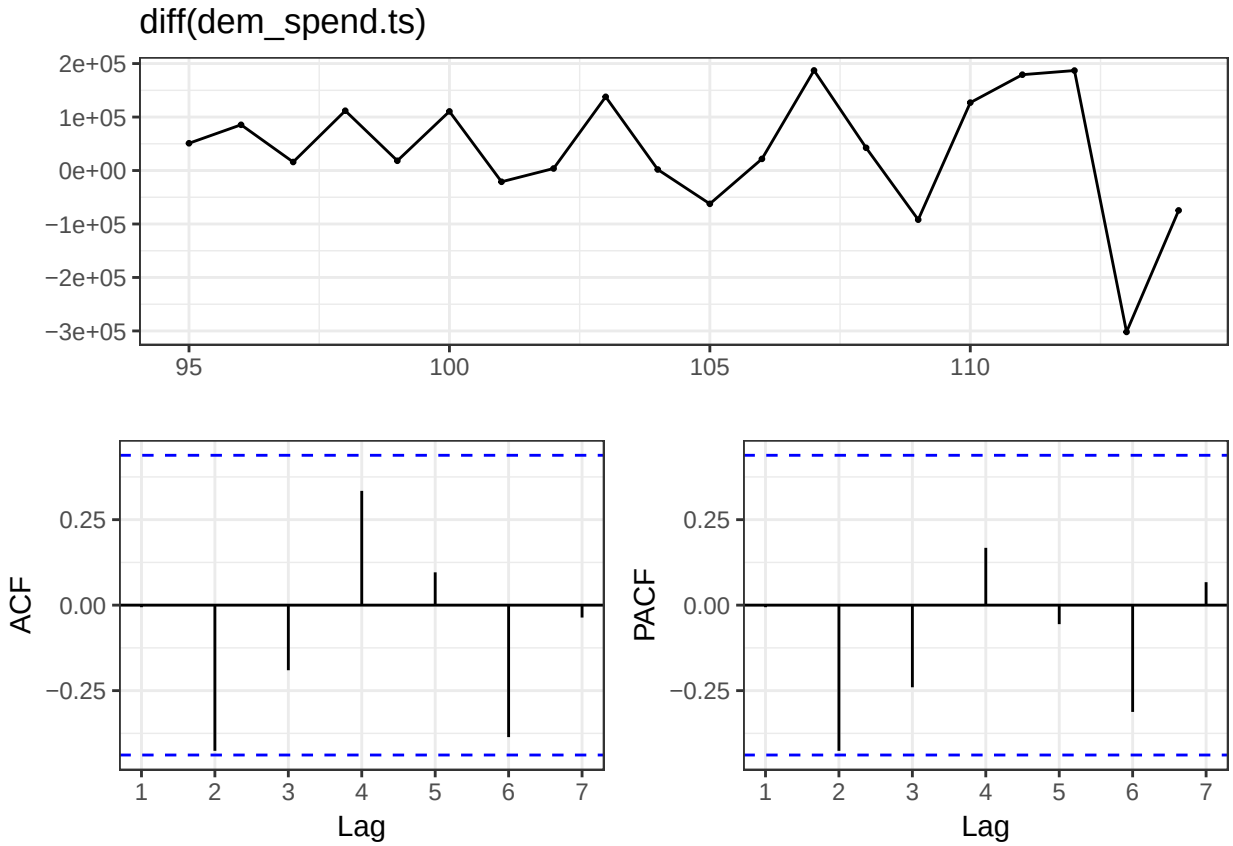


Of all attempts, the model above had the best AIC value. That said, it still has some pretty drunk behavior. At this point however, we have decided to make some predictions.

Predictions for the 116th Congress

When we last left our heroic model, we had not captured the structure of the 115th Congress; the data for mean democratic spending is missing. If we get that we will have a prediction for the mean of the 115th Congress. We will use that to fit an ARIMA to Congresses 87 to 115, and then predict the 116th. To get the mean democratic spending for the Congress 115, we will fit another ARIMA to the data we have on democratic spending. We will predict 2018 midterm election turnout by the mean of midterm election turnout from 1962 to 2014.

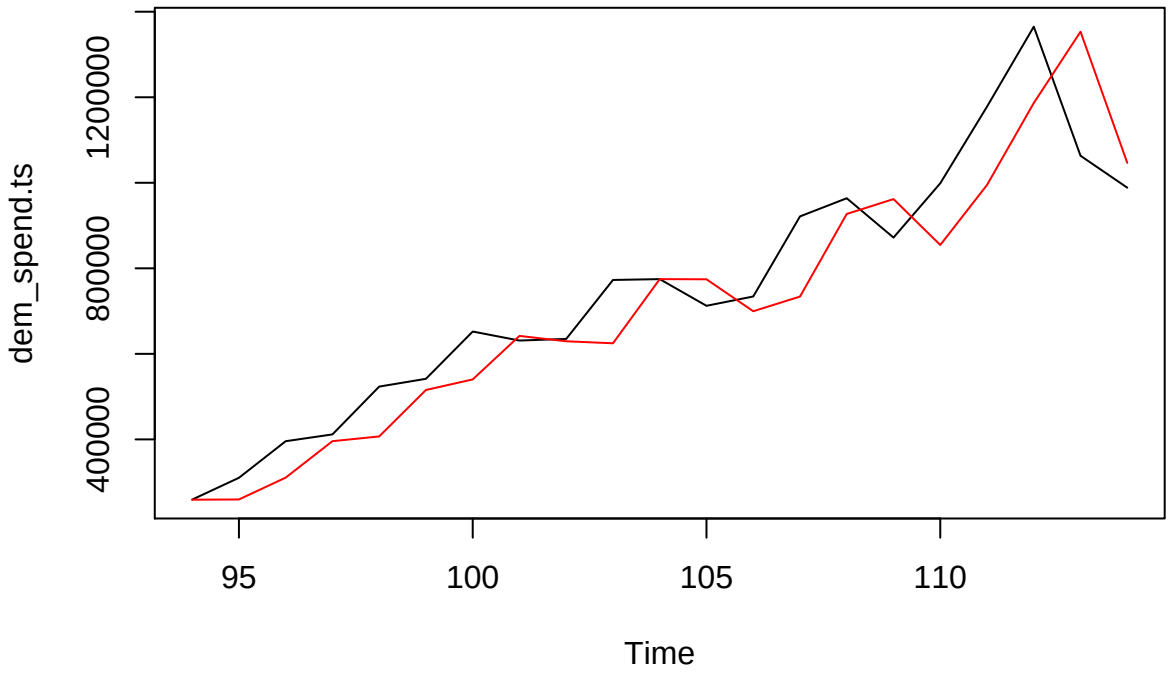
```
dem_spend.ts = ts(data = spending_house$blue_mean, start = 94)
ggtsdisplay(diff(dem_spend.ts))
```



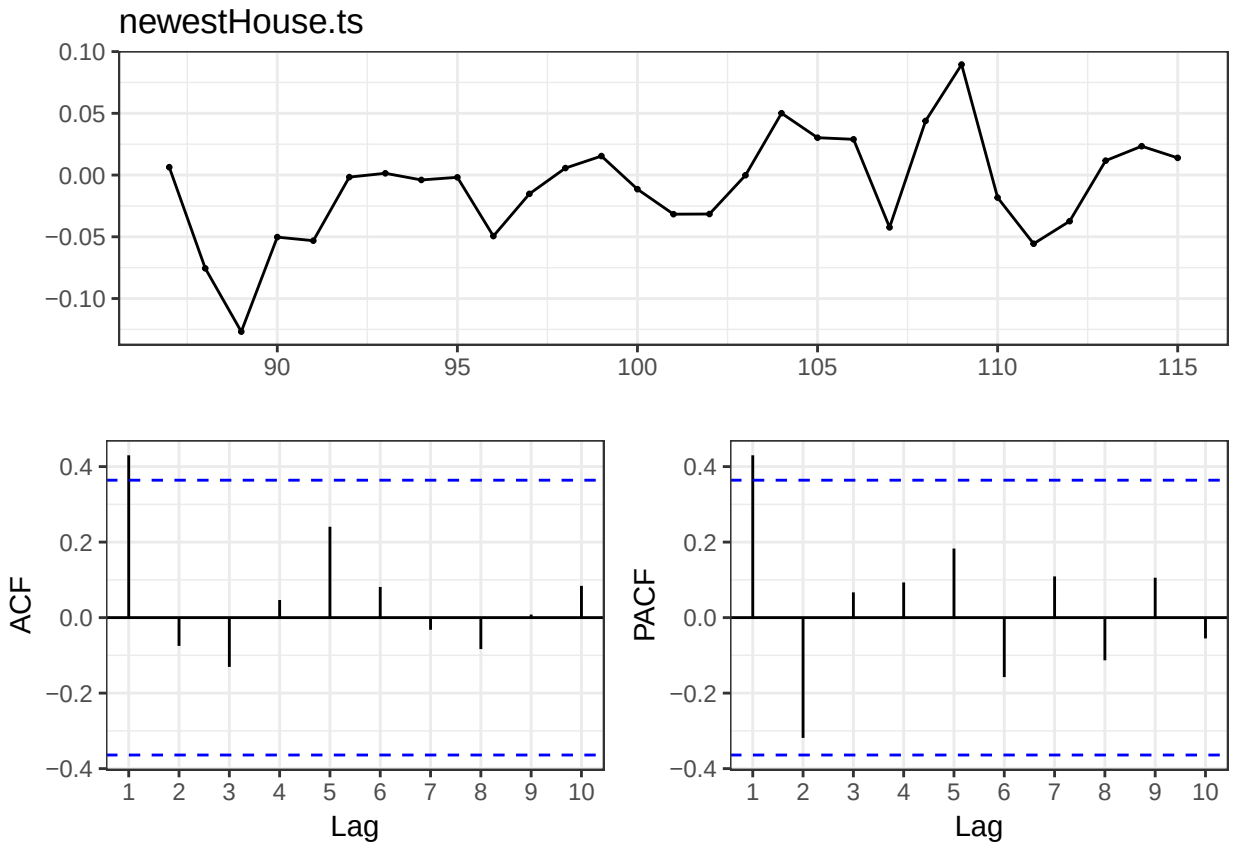

```
dem_spend.arima = Arima(dem_spend.ts, order = c(0,1,0),
                        seasonal = list(order = c(1,0,0), period = 3))
dem_spend.arima
```

```
## Series: dem_spend.ts
## ARIMA(0,1,0)(1,0,0)[3]
##
## Coefficients:
##          sar1
##       -0.0924
## s.e.   0.2757
##
## sigma^2 estimated as 1.486e+10:  log likelihood=-262.1
## AIC=528.19   AICc=528.9   BIC=530.18
```

```
{
  plot(dem_spend.ts)
  points(dem_spend.arima$fitted, col = "red", type = "l")
}
```




```
pred_dem_spend_115 = forecast(dem_spend.arima)$mean
newHouse.df = rbind(from1974_df_house, c(115,1,0,0,54.7,NA,pred_dem_spend_115[1],0.5540230))
structure_115 = predict(from1974_lm_house, newdata = newHouse.df%>%filter(congress==115))
house_structure_full = c(firstModelStructure, house_structure, structure_115)
newestHouse.ts = ts(data = partisan$red_rep - house_structure_full,
                    start = min(partisan$congress))
ggtsdisplay(newestHouse.ts)
```




```
newestHouse.arima = Arima(newestHouse.ts, order = c(0,0,1),
                          seasonal = list(order = c(0,0,1), period = 5))
time_red_rep_116 = forecast(newestHouse.arima)$mean[1]
pred_turnout = partisan %>% filter(midterm == 1) %>% .$HouseElections %>% mean()
newHouse.df = rbind(newHouse.df, c(116,1,1,1,pred_turnout,NA,pred_dem_spend_115[2],NA))
mean_red_rep_116 = predict(from1974_lm_house, newdata = newHouse.df%>%filter(congress==116),
                           interval = "prediction")
full_red_rep_116 = mean_red_rep_116 + time_red_rep_116
```

Table 1: Predicted Red Members of the House in 116th Congress

Lower	Fit	Upper
155.8	199.05	242.3

The Senate predictions will come directly from our ARIMA since it looked decent. Attempting to predict for the 33 seats coming up in 2018 with only 29 Congresses worth of data... very iffy.

```
Senate_red_pred = forecast(Senate.arima)
lwr_sen = round(Senate_red_pred$lower[1,2]*100, 2)
upr_sen = round(Senate_red_pred$upper[1,2]*100, 2)
fit_sen = round(Senate_red_pred$mean[1]*100, 2)
```

	Lower	Fit	Upper

Table 2: Predicted Red Members of the Senate in 116th Congress

	Lower	Fit	Upper
95%	44.07	52.89	61.7