Lab 6

Taylor Blair

Math 141, Week 6

Due: Before the time when your Week 7 lab would meet

Note:

- There are no labs during Week 7.
- You must still turn in this lab at the usual time.
- We **strongly** encourage you to finish this lab before taking the mid-term exam.

Goals of this lab

1. Conduct exploratory data analysis.
2. Build linear regression models.
3. Interpret the estimated coefficients of a linear regression model.
4. Understand how the inclusion of multiple explanatory variables can impact the utility of a particular explanatory variable.
5. Practice manually collecting and entering data for analysis.
6. Learn how to import data into RStudio from a Google Sheet.
7. Think about the role and limitations of statistical forecasting in elections and elsewhere.

Useful Math Notation in RMarkdown Documents

You can add mathematical formula to your knitted document by placing the math between dollar signs: β_o. You can add centered equations with double dollar signs:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

If there is a mathematical symbol you want to include and we haven't shown you an example in lab or in-class, feel free to ask on Slack or the internet. (If asking the internet, you want search for "latex code for insert_your_desired_symbol".)

Problems

Problem 1

In this problem, we will try to answer the seemingly simple question: **How much does a fireplace cost?** To do this, we will use the dataset `SaratogaHouses`, which contains data on homes in Saratoga Springs, New York in 2006.

```
library(mosaicData)
data(SaratogaHouses)
```

```
# Look at the help file to learn more about the variables
?SaratogaHouses
```

a. Load the libraries (packages) you will need to solve this problem.

```
library(tidyverse)
library(dplyr)
```

b. Create a new variable for whether or not a house has a fireplace.

```
SaratogaHouses$has_fire <- SaratogaHouses$fireplaces!=0
```

c. For both categories of the variable you created in b, determine the average price of a house. Using that information, answer our original question: **How much does a fireplace cost?**

```
no_fire = mean(SaratogaHouses$price[SaratogaHouses$has_fire==FALSE])
cat("Mean cost without a fireplace: $", no_fire)
```

```
## Mean cost without a fireplace: $ 174653.4
```

```
yes_fire = mean(SaratogaHouses$price[SaratogaHouses$has_fire==TRUE])
cat("Mean cost with a fireplace: $", yes_fire)
```

```
## Mean cost with a fireplace: $ 239914
```

```
cat("Difference in cost or cost of a fireplace: $", yes_fire - no_fire)
```

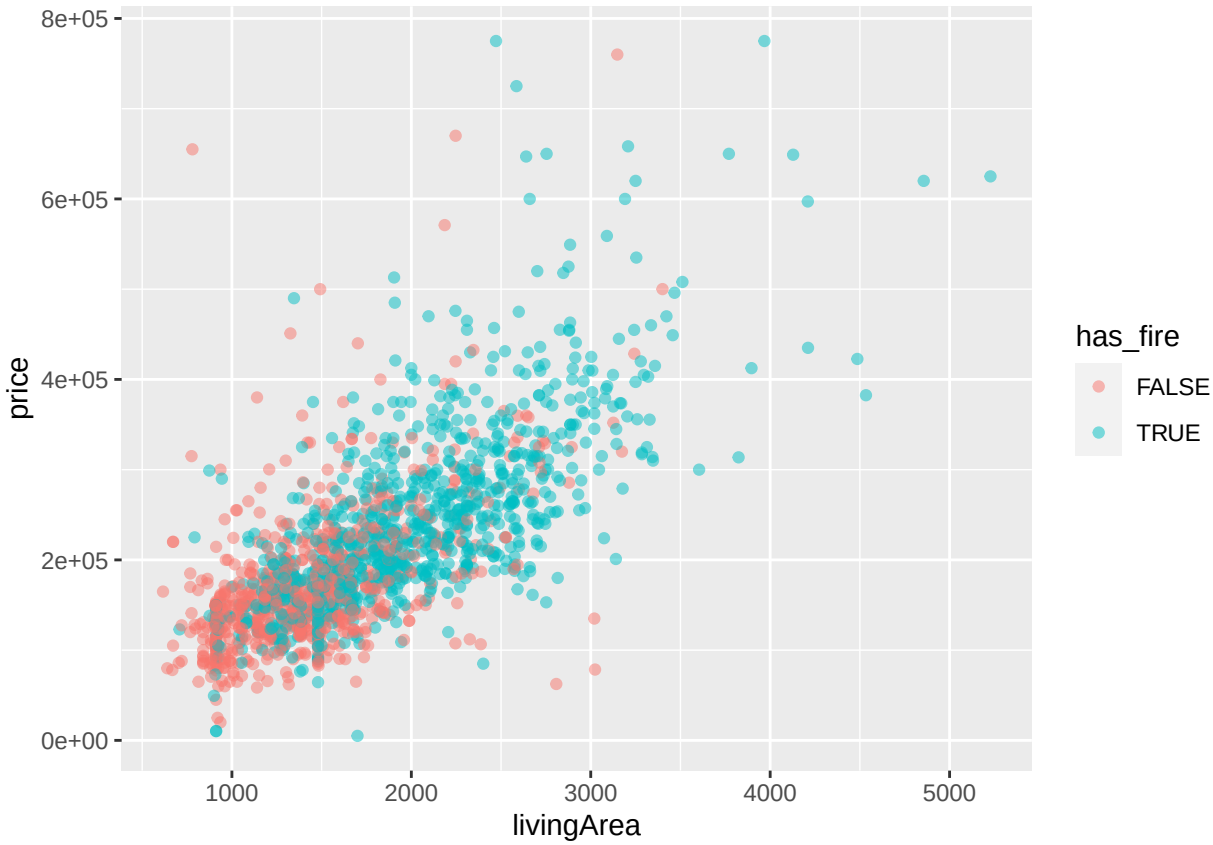
```
## Difference in cost or cost of a fireplace: $ 65260.61
```

- A fireplace costs $65,260 on avarage. This does not control for other variables (location and house size), so I am assuming that will come next. I would typically do a K-nearest neighbour analysis disregarding wheter or not a house has a fireplace, as well as cost, then

---

d. Produce a plot of `price` and `livingArea` with different colors based on whether or not the house has a fireplace. What does the plot tell us about houses with fireplaces, in terms of size and price?

```
ggplot(data = SaratogaHouses, mapping = aes(x= livingArea, y = price, color = has_fire) ) +
  geom_point(alpha=0.5)
```



- Larger places have a higher likelihood of having a fireplace than a smaller house.

---

e. Build a linear regression model for `price` using your fireplace variable and `livingArea`. Assume equal slopes in the model.

```
lm(data = SaratogaHouses, price ~ livingArea + has_fire)
```

```
##
## Call:
## lm(formula = price ~ livingArea + has_fire, data = SaratogaHouses)
##
## Coefficients:
##  (Intercept)    livingArea  has_fireTRUE
##      13599.2         111.2        5567.4
```

f. Interpret the coefficients in the above model. Make sure to state whether the intercept makes sense in this context.

- Without fire place
  - $price \approx 111.2(livingarea) + 13599.2$
  - Intercept of: $ 13599.2
- With fire place
  - $price \approx 111.2(livingarea) + 13599.2 + 5567.4$
  - Intercept of: $ 19166.6
  - $5,567.40 more than a house without a fireplace when they have the same fireplace

g. **Now that we are controlling for house size, how does your answer to our original question about the cost of a fireplace change?** Use a plot you already created to help explain why the answer changes when we control for living area.

- We were likely looking at a coorelation, not causation. Looking at the graph above, houses that do not have a fire place have a domain of [600, 3500] compared to [700, 5200]. If we thought of the two as having a completley uniform, linear, and equal distributions than wethe range of the second would be skewed more expensive. When we controlled for area we were no longer comparing two lists as if they had the same range with a different distribution.

---

h. Build a linear regression model for `price` using your fireplace variable and `livingArea`. Assume the slopes in the model to vary.

```
#"No fireplace linear model"
print(lm(data = SaratogaHouses[SaratogaHouses$has_fire==FALSE, ], price~livingArea))
```

```
##
## Call:
## lm(formula = price ~ livingArea, data = SaratogaHouses[SaratogaHouses$has_fire ==
##     FALSE, ])
##
## Coefficients:
## (Intercept)   livingArea
##    40901.29        92.36
```

```
#"With fireplace linear model"
print(lm(data = SaratogaHouses[SaratogaHouses$has_fire==TRUE, ], price~livingArea))
```

```
##
## Call:
## lm(formula = price ~ livingArea, data = SaratogaHouses[SaratogaHouses$has_fire ==
##     TRUE, ])
##
## Coefficients:
## (Intercept)   livingArea
##      3290.9        119.2
```

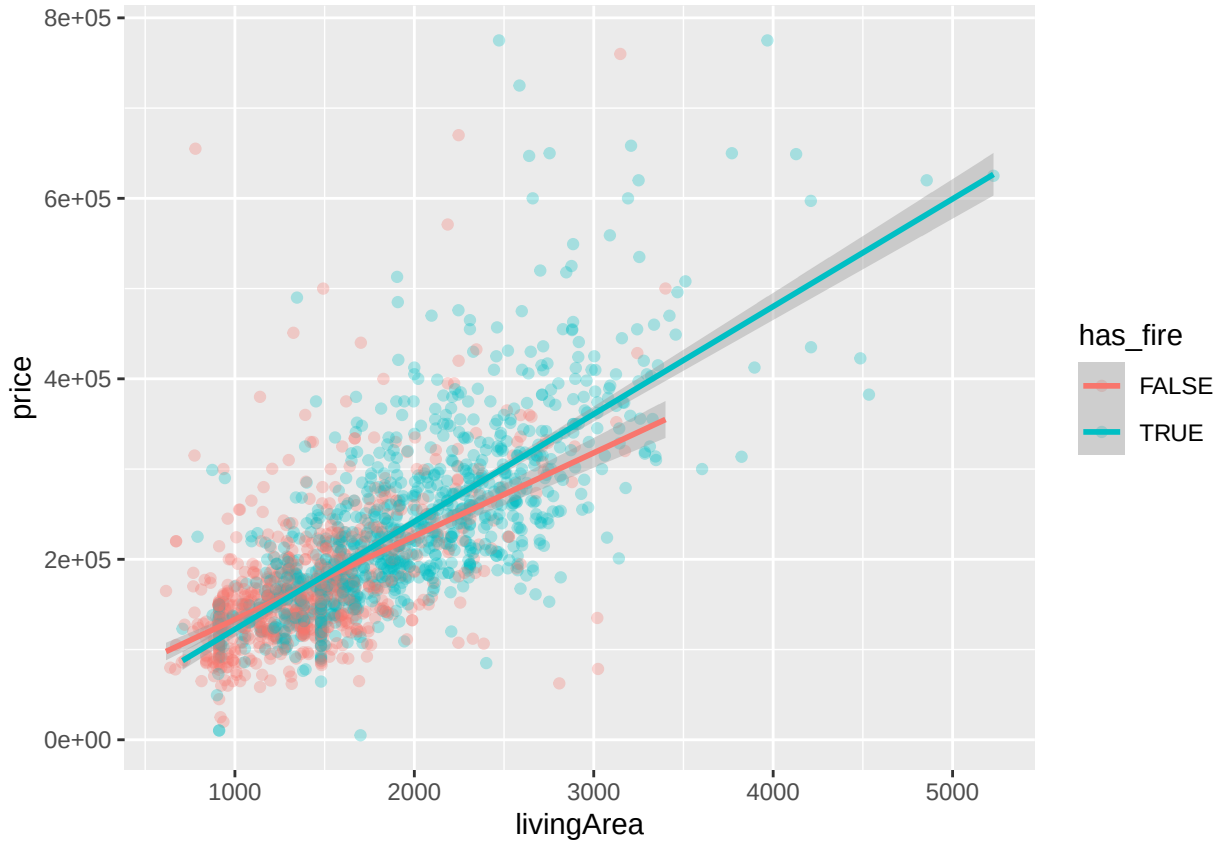
i. Based on your model above, what is the estimated slope for houses with fireplaces? What is the estimated slope for houses without fireplaces?

- Without fireplaces
  - $price \approx 92.36(livingarea) + 40901.29$
- With fireplaces
  - $price \approx 119.20(livingarea) + 3290.90$

---

j. Produce (or reproduce if you already produced it in an earlier problem) a plot that contains `price`, `livingArea`, and your fireplace variable and the linear regression lines. Based on this plot and your analyses, do fireplaces cost "more" or "less" for larger houses than for smaller houses? Justify your answer.

```
ggplot(data = SaratogaHouses, mapping = aes(x= livingArea, y = price, color = has_fire) ) +
  geom_point(alpha=0.3)+ geom_smooth(method = lm)
```



- Fireplaces do cost more for more expensive houses. Looking at the equations, the one for with a fireplace has a higher intercept and steeper slope.

Problem 2

Note: This problem was provided by the Early Voting Information Center.

As all of you know, there is a US presidential election coming up in a few weeks. You have likely seen the election forecasts put out by websites like FiveThirtyEight or the Economist. In this problem, you will gather data and create your own (much simpler) forecast for the 2020 Election. In the process, you'll think about the ways in which this kind of forecasting is useful, and some of the ways it might not be.

a. First you will need to collect data on the outcomes of recent elections (from 1960 to 2016). A Google Sheet with columns for the year, incumbent president and incumbent party can be found here. Make a copy for yourself by choosing the "Make a Copy" option from the "File" drop-down menu. Once you have copied the sheet, press the green 'share' button in the top right corner and change the link permissions to "Anyone with a link."

To make things simple, we will be attempting to predict/forecast the **percentage of the popular vote** for the candidate from the incumbent party. This data can be found in many places, such as this Wikipedia article. Make sure you record the vote share of the **incumbent party**, not the winning candidate, in each election since 1960. **Note: Don't put spaces in the variables names.**

b. Now you need to collect some explanatory variables. Let's start with the presidential approval rating. For example, for the 2016 election, you will provide President Obama's approval rating. Here are two potential sources: Gallup, Wikipedia.

You can find approval rating listed as an average (over the course of the president's time in office) or as a time series such that you could record the approval of each president during a particular month. Choose one method and add it to your spreadsheet. Explain why you chose it along with the relationship you think approval rating has with incumbent party vote share.

- I am adding the difference in approval rating from entering office and month before election ~~Because I am tired, need to do a hum paper, and finish this panic attack~~ because I believe it .
- January 20th of year after innaguration is enter office
- As close to a month before election

---

c. Next we'll want some kind of economic indicator. Possible options include unemployment rate, GDP growth, wage growth and others. Government sources like the Census Bureau and the Bureau of Labor Statistics will be the best source for this data. Again, describe your choice, why you picked it, and how you expect it to relate to the response variable.

- *I want to add that I was initially going to look at unemployment rate of incumbent office, or change, or what not, but it is 9:38 PM on Wenesday, and I wasn't able to do that yesterday as I was in the middle of a panic attack which ended in me crying on a toilet.* **tl;dr: Had a panic attack wasn't able to look at the variable I had intially planned on finding.**
- **I am looking at life expactancy at birth**, *specifically change in life expactancy since previous election*
- I argue life expectancy **IS** an economic indicator.
- When it is lower it suggests a high rate of child mortality which is a burden on a system and forces families to have more children.
- In addition it often reflects current strifes, lack of access to healt care.
- When it is higher it suggests there are fewer childrens, and longer until retirement.
- I sourced my data from the Worldbank
- *I considered looking at 18 year olds life expectancys, but I could not find the data*

---

d. Finally, pick a categorical variable of your choosing. It can be something you believe to be relevant to the election or something silly that you don't expect to have a relationship to the election. Regardless, you should once again explain your choice and how predictive (or not predictive) you expect it to be.

- Is the US at war with another country? Has The US just Entered or completed a war?
- Looking at what has happened in the 4 years prior to an election. Only counting wars were blood is shed and the war crosses over two presidential terms
- Using wars listed on Wikipedia
- Represented with
  - `N` for No war
  - `S` for Starting
  - `O` for Ongoing
  - `E` for End

- Wars that are included are:
  - **Vietnam War**
    - **Begins**: Several numbers, but the US provided support in *1961*, so although there is not an official congressional support until the following election, I mark this as the beggining.
    - **Ends**: *1975*
    - Wikipedia: Vietnam War
  - **Afghanistan War**
    - **Begins**: *2001*
    - **Ends**: In spite of the massive mission accomplished sign, this is still ongoing.
    - [Afghanistan War](https://en.wikipedia.org/wiki/War_in_Afghanistan_(2001%E2%80%93present)

---

e. Now you can import your data using the package `gsheet`. Copy and paste the sharing url for your sheet below and run the R chunk. Check your loaded to make sure that everything looks as it should. (Remember to change `eval` to `TRUE`.)

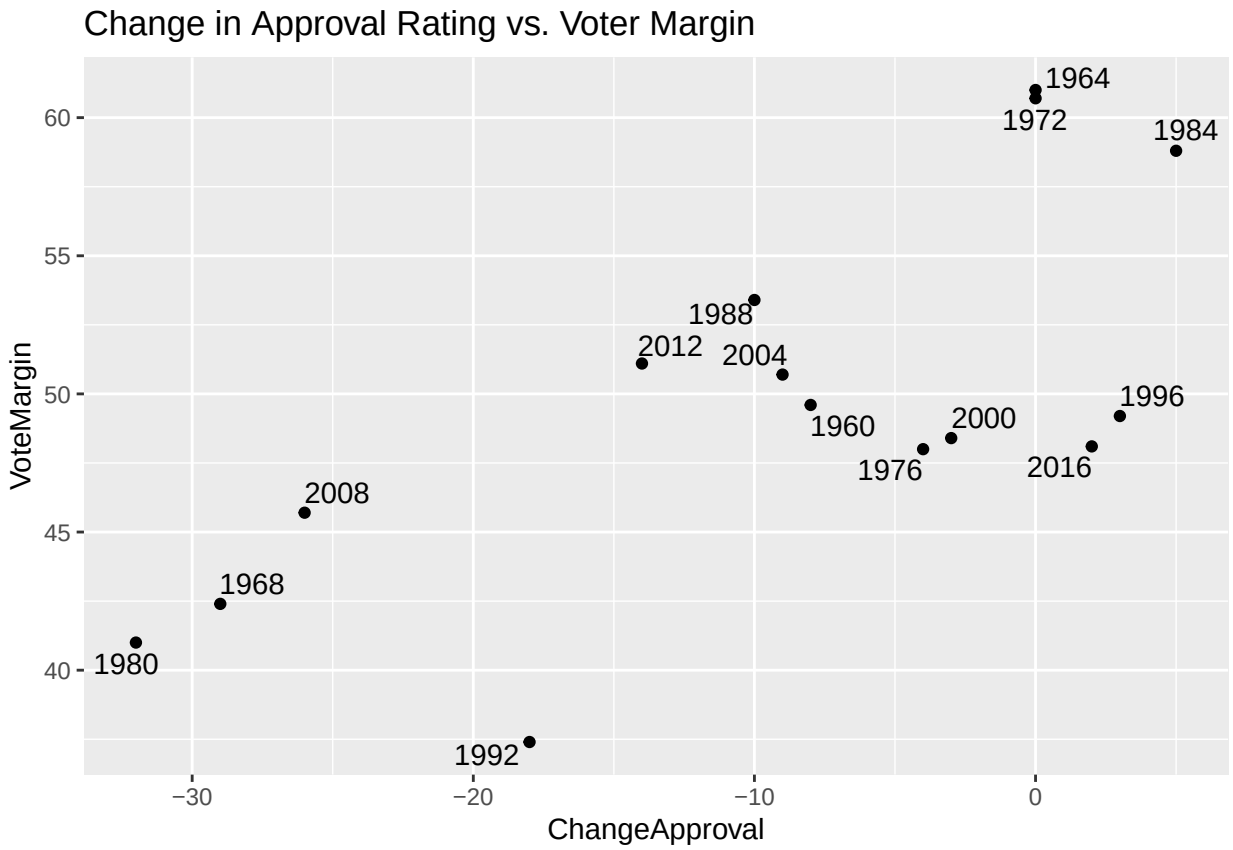
```
library(gsheet)
library(ggrepel)
library(ggcorrplot)

url <- "https://docs.google.com/spreadsheets/d/1TZg32WGmer3iAQ63CrKzKh61c4mpw9K93DTsN651WAA/edit#gid=0"
my_pres_data <- gsheet2tbl(url)
my_pres_data
```

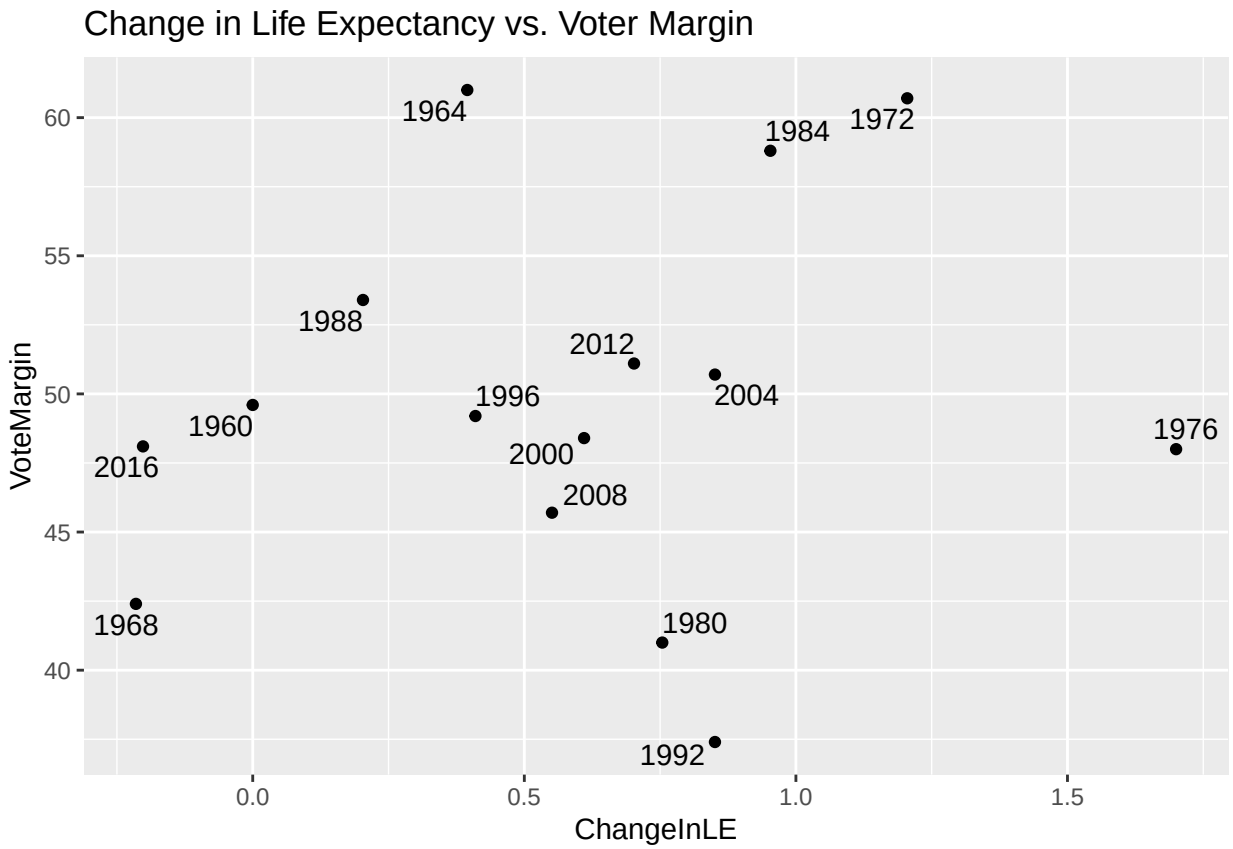
```
## # A tibble: 15 x 10
##     Year IncumbentParty IncumbentPresid~ VoteMargin EnterOffice BeforeElection
##    <dbl> <chr>          <chr>                <dbl>       <dbl>           <dbl>
##  1  2016 D              Barack Obama          48.1          50              52
##  2  2012 D              Barack Obama          51.1          66              52
##  3  2008 R              George W. Bush        45.7          51              25
##  4  2004 R              George W. Bush        50.7          57              48
##  5  2000 D              Bill Clinton          48.4          61              58
##  6  1996 D              Bill Clinton          49.2          54              57
##  7  1992 R              George H. W. Bu~      37.4          51              33
##  8  1988 R              Ronald Reagan         53.4          63              53
##  9  1984 R              Ronald Reagan         58.8          51              56
## 10  1980 D              Jimmy Carter          41            66              34
## 11  1976 R              Gerald Ford           48            51              47
## 12  1972 R              Richard Nixon         60.7          59              59
## 13  1968 D              Lyndon B. Johns~      42.4          71              42
## 14  1964 D              Lyndon B. Johns~      61            72              72
## 15  1960 R              Dwight D. Eisen~      49.6          73              65
## # ... with 4 more variables: ChangeApproval <dbl>, LifeExpectancy <dbl>,
## #   ChangeInLE <dbl>, ForeignConflict <chr>
```

f. Create graphs comparing the incumbent vote share to your explanatory variables. Comment on any relationships displayed in the graphs.

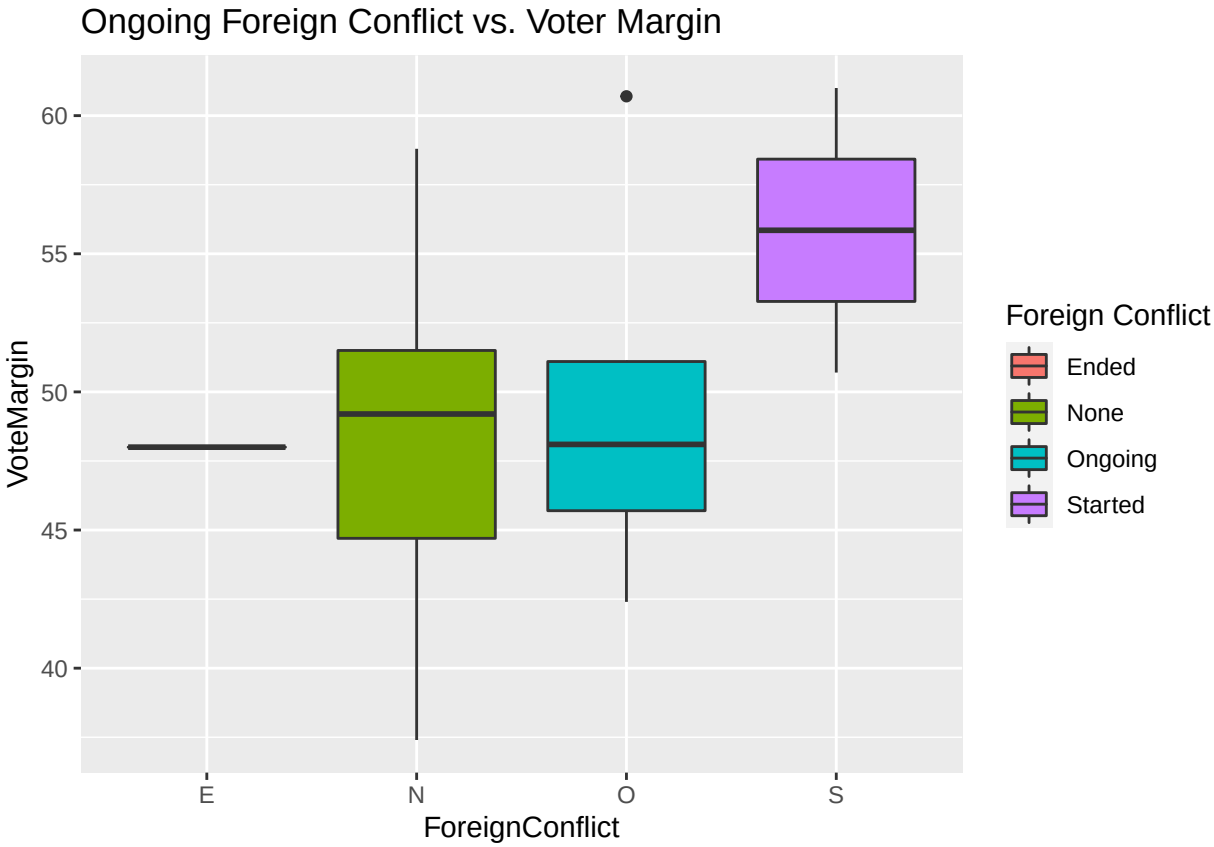
```
ggplot(my_pres_data, mapping =  aes(ChangeApproval, VoteMargin, label=Year) ) +
  geom_point() + geom_text_repel() +
  labs(title = "Change in Approval Rating vs. Voter Margin" )
```




```
ggplot(my_pres_data, aes(ChangeInLE, VoteMargin, label=Year) ) +
  geom_point() + geom_text_repel() +
  labs(title = "Change in Life Expectancy vs. Voter Margin", xlab=("Change in Life Expectancy") )
```




```
ggplot(my_pres_data, aes(ForeignConflict, VoteMargin, fill = ForeignConflict) ) +
  geom_boxplot() +
  labs(title = "Ongoing Foreign Conflict vs. Voter Margin", xlab=("Ongoing Foreign Conflict") ) +
  scale_fill_discrete(name = "Foreign Conflict", labels = c("Ended", "None", "Ongoing", "Started"))
```



- **Change in Approval Rating**
  - This is a particularly strong correlation, it makes sense that a variable the compares presidential approval rating after being elected and before an election would strongly correlate to the incumbent party voter margin.

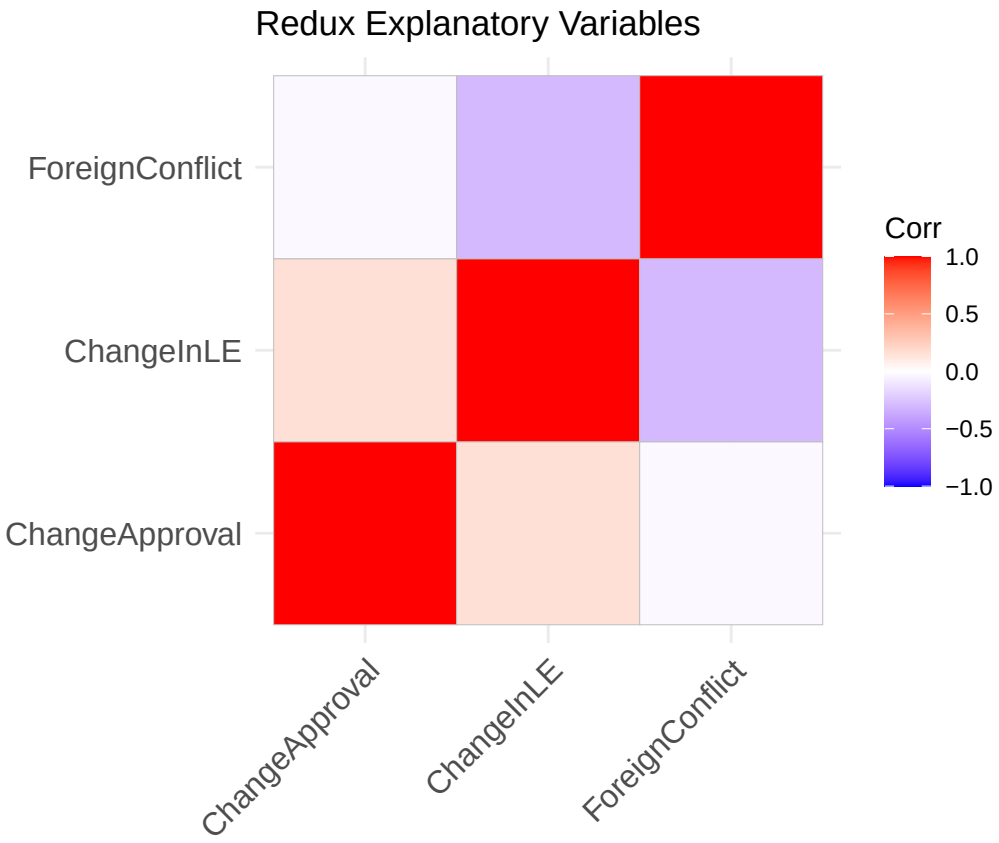
- **Change in Life Expectancy**
  - This graph does not have a particularly strong coorelation, however. When the life expectancy of newborns drops below 0, the vote margin does not exceed 50%
  - *This was going to be used in a descion tree, espcially if I could have found other age groups*
- **Multi term Foreign Conflicts**
  - If there is one lesson, if you want a strong voter margin, start a war. This does only contain two points, so it is not particularly indicative of a pattern.
  - Ending wars is unpopular
  - Not being at war is on avarage better than being at war in terms of reelection, but it has a larger spread (and therefore more risk)

---

g. Also create some graphs of your explanatory variables against each other. Comment on any relationships displayed in the graphs.

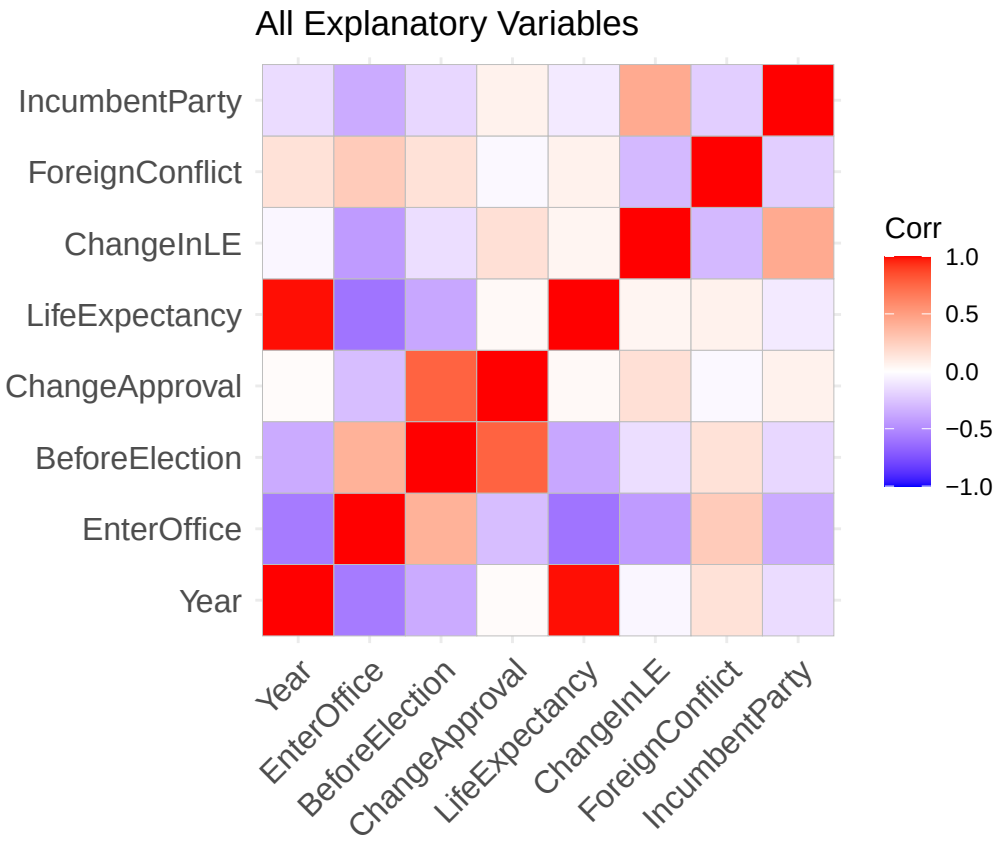
```
explantory_vars <- my_pres_data %>%
  select(ChangeApproval, ChangeInLE)

corr_explan_vars <- explantory_vars
corr_explan_vars$ForeignConflict <- as.numeric(factor(my_pres_data$ForeignConflict))
ggcorrplot(cor(corr_explan_vars)) + labs(title = "Redux Explanatory Variables")
```




```
explantory_vars <- my_pres_data %>%
  select(Year, EnterOffice, BeforeElection, ChangeApproval,
         LifeExpectancy, ChangeInLE)

corr_explan_vars <- explantory_vars
corr_explan_vars$ForeignConflict <- as.numeric(factor(my_pres_data$ForeignConflict))
corr_explan_vars$IncumbentParty <- as.numeric(factor(my_pres_data$IncumbentParty))
ggcorrplot(cor(corr_explan_vars)) + labs(title = "All Explanatory Variables")
```



- **Reduced Explanatory Variables Correltion Graph**
  - There is not a particularly strong correlation with any of the variables, aside from a slight corelation between life expectancy and change in approval.
  - There is a correlation between life expectancy and foreign conflict. *Your odds of dying younger increase when we have been fighting the same war for twenty years.*
- **Full Explanatory Variables Correlation Graph**
  - *Several of these variables are based on other variables in the data, I am skipping over those*
  - Change in life expectancy is linked to incumbent party!?!!?!!?!
  - Entering office approval rating is more correlated to foreign conflicts than either the change in aproval or the end office. Likely because most approval ratings are higher when beggining in office, and fall down and the majority of points are no conflict or ongoing.

---

h. Compute correlation coefficients for all quantitative variables. Discuss how the correlation coefficients do or do not support your conclusions in f and g.

```
explantory_vars <- my_pres_data %>%
  select(Year, EnterOffice, BeforeElection, ChangeApproval,
         LifeExpectancy, ChangeInLE)

corr_explan_vars <- explantory_vars
corr_explan_vars$ForeignConflict <- as.numeric(factor(my_pres_data$ForeignConflict))
corr_explan_vars$IncumbentParty <- as.numeric(factor(my_pres_data$IncumbentParty))
cor(corr_explan_vars)
```

```
##                       Year EnterOffice BeforeElection ChangeApproval
## Year            1.00000000  -0.5693624     -0.3616523     0.02001592
```

```
## EnterOffice     -0.56936236   1.0000000      0.4011335    -0.28041623
## BeforeElection  -0.36165229   0.4011335      1.0000000     0.76678315
## ChangeApproval   0.02001592  -0.2804162      0.7667832     1.00000000
## LifeExpectancy   0.98737936  -0.6000910     -0.3754202     0.02712204
## ChangeInLE      -0.03752415  -0.4274917     -0.1376359     0.15534149
## ForeignConflict  0.15324283   0.2669074      0.1461621    -0.03387659
## IncumbentParty  -0.15464739  -0.3606042     -0.1720395     0.07241885
##                 LifeExpectancy  ChangeInLE ForeignConflict IncumbentParty
## Year                 0.98737936 -0.03752415      0.15324283    -0.15464739
## EnterOffice         -0.60009096 -0.42749171      0.26690739    -0.36060420
## BeforeElection      -0.37542023 -0.13763589      0.14616213    -0.17203948
## ChangeApproval       0.02712204  0.15534149     -0.03387659     0.07241885
## LifeExpectancy       1.00000000  0.05086520      0.07167269    -0.08609833
## ChangeInLE           0.05086520  1.00000000     -0.30457901     0.43695570
## ForeignConflict      0.07167269 -0.30457901      1.00000000    -0.21012763
## IncumbentParty      -0.08609833  0.43695570     -0.21012763     1.00000000
```

- *See pretty graphs and explination above*

---

i. Based on your explanatory analysis, determine an appropriate form for your forecast model. Make sure to consider interaction terms or polynomial terms.

Write it in the following way:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \epsilon$$

where you define what y and each of the x's represents.

- $y \approx \beta_o + \beta_1 x_1 + \beta_2 x_2$
- $y =$ `VoteMargin`
- $\beta_o =$ `ForeignConflict`
- $\beta_1 x_1 =$ `ChangeApproval`
- $\beta_2 x_2 =$ `ChangeInLE`

---

j. Build your model and print out the regression table.

```
BOTUS <- lm(VoteMargin~ ForeignConflict + ChangeApproval+ ChangeInLE, my_pres_data)
```

k. Let's look at the predicted values and the residuals of your model. We can add those two columns to the dataset with the following code (by inserting your model name). Looking at the residuals, how well would you say your model predicts incumbent vote share? Does it correctly predict which party got the majority in the last few elections? Are there elections where the model was particularly wrong?

```
library(modelr)
# Add residuals and predictions
my_pres_data  <- my_pres_data %>%
  add_predictions(BOTUS) %>%
  add_residuals(BOTUS)

ggplot(my_pres_data, aes(pred, VoteMargin, label=Year)) +
  geom_point() + labs(title = "Predicted vs Actual Voter Margin") +
```

```
  geom_text_repel()

plot(my_pres_data$Year, (my_pres_data$pred-my_pres_data$VoteMargin),
     ylab = "Difference Predicted and Actual Margin", xlab= "Year")
```

- Did poorly on the nineties and the 2016 election
- *I am writting this at 1:10 AM on Thursday morning, my brain is doing its best*

---

l. Find current values for your three explanatory variables. Using your model, forecast the the Republican vote share in the 2020 election. Is your result believable? Why/why not?

```
# Still in Afghan war, value is 3 for foreign conflict

# Entering office: 45%
# Closest to 1 month prior election (september): 46%
# Difference is = 1

# Closest to current year in predicted life expect is 2018: 78.7
# 4 yers before 2018 is 2014: 78.841
# Diff life expirience = -0.141

inputs.curr <- data.frame(
  features= c(3, 1, -0.141))

predict(BOTUS, predicted = inputs.curr, interval = "confidence")
```

```
##         fit      lwr      upr
## 1  53.92311 45.80574 62.04048
## 2  50.25480 44.68374 55.82586
## 3  44.97618 38.73459 51.21776
## 4  54.73826 46.37526 63.10126
## 5  50.87925 46.25207 55.50644
## 6  52.67864 47.18382 58.17346
## 7  45.59116 40.07283 51.10949
## 8  46.82776 41.82523 51.83030
## 9  55.13714 48.74775 61.52654
## 10 39.67412 31.93594 47.41230
## 11 48.00000 36.56605 59.43395
## 12 57.40962 48.92472 65.89453
## 13 41.43629 33.50175 49.37082
## 14 56.96174 48.59874 65.32474
## 15 47.01192 41.08102 52.94282
```

- **No**
- Seems really off
- This is a unique election, one that is primarily by mail, and with a president that is extremely unpopular, elected by a minority
- This was going to be a desicion tree, but it is a bit too late for that

---

m. What are some reasons we might want to be cautious with this kind of prediction?

- No election is the same
- **OVERFITTING**!!!!

---

n. What are the circumstances under which models like this one (potentially much more complicated) can be trusted?

- Not enough variables
- We are in the middle of a pandemic
- Voter suppression
- Potentially hostile transition of power
- There is not enough xanax in the sewer runoffs for me to continue listing these. Take some Mefloquine, watch some horror movies for 20 hours, then get back to me on what you dream up and add it onto the list.
