Visualization Part II: Homework

FRB/Howard Instructors

Date Assigned: Friday October 13, 2017

Date Due: Thursday October 19, 2017 by 11:59 pm

Introduction - Read the following instructions

For this homework you are going to use ggplot to reproduce the following plots. For each question I have generated what the answer should look like, you have to figure out how to use ggplot to make it!

For this assignment you will need to use Rmarkdown which we discussed last week. You will need to produce a .Rmd file which creates a pdf document containing the code you use for the assignments, the output plots, and answers to your free response questions. You must submit to us the .Rmd file as well as the pdf it generates.

Additionally, be sure to include in the subtitle of the chart the question that the plot is created for. Your charts must all be appropriately titled, (chart titles must be descriptive and accurate for the data displayed), and axes/guides must be labeled appropriately as well.

You must use ggplot for creating your plots. This assignment will be worth 100 points.

Before beginning this assignment I'd recommend reading chapters 3 and 28 of R for Data Science by Hadley Wickham at r4ds.had.co.nz

Data

For this homework you will need to read in the treasuries, stock_closings, national_election_results, and oh_elections_results csv files in the ggplot Homework folder. You will have to read it in by yourself.

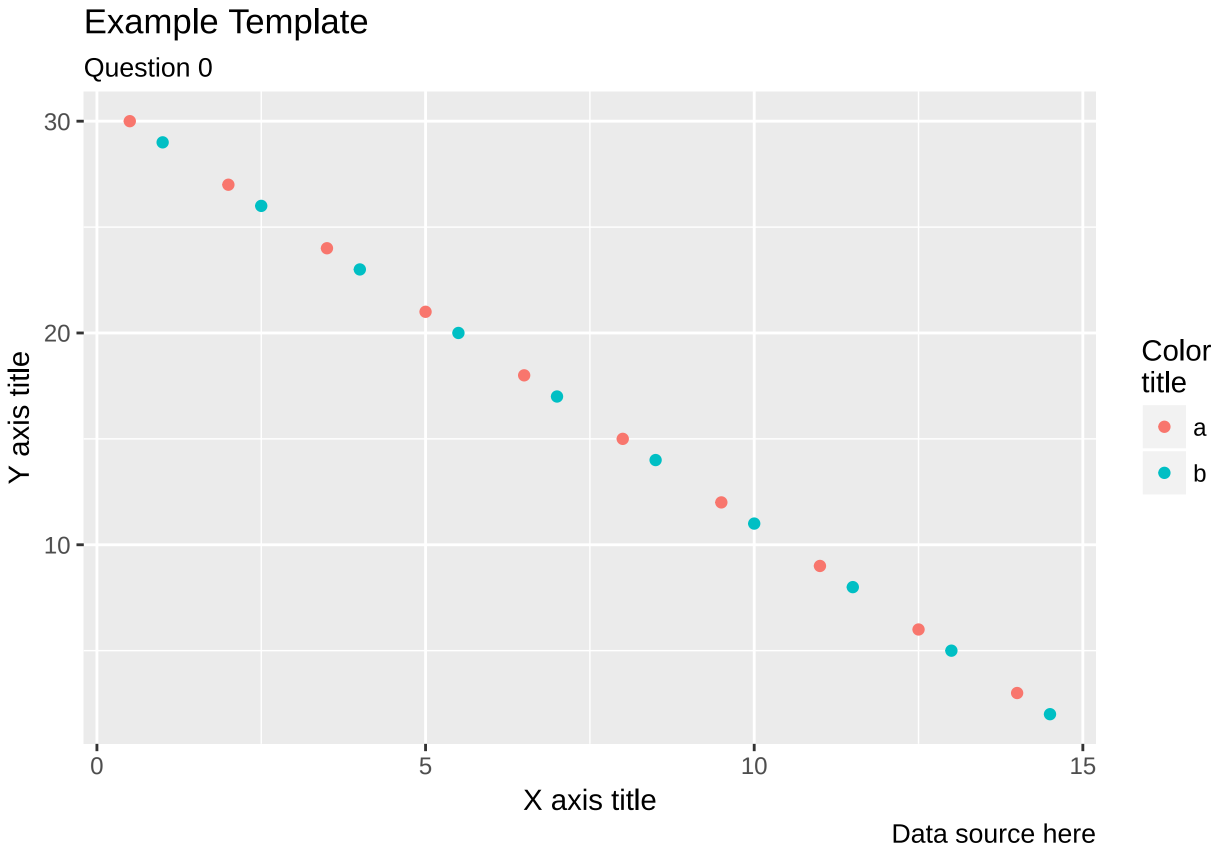
Additionally: be sure that you `library()` the `ggplot2`, `dplyr`, `stringr`, `lubridate`, `maps`, and `tidyr` packages for this homework.

Hint: Don't forget that you can add layers to already created plots!

Example Answer

This page should serve as a template for you regarding how to write your .Rmd file and what your pdf file should look like. The answer for each question should have the following: code used to set up the data/make the plot, the plot itself, the answer to any free response question if applicable. Don't worry if you use more than one page answering a question.

```r
# Read in/set up your data (I will use random data for this)
ex_data <- data.frame(group = rep(c("a", "b", "c"), 10),
                      x = seq(0.5, 15, by = 0.5),
                      y = seq(30, 1, by = -1))
# Use any needed packages to set up your data for plotting
question0_data <- ex_data %>% filter(group != "c")
# now plot your data
question0 <- question0_data %>%
    ggplot(aes(x = x, y = y, color = group)) +
    geom_point() +
    labs(x = "X axis title", y = "Y axis title",
         title = "Example Template", subtitle = "Question 0",
         color = "Color\ntitle", caption = "Data source here")
question0
```



Free response answer: We can see from this chart that observations in groups “a” and “b” have identical trends. This plot is severly lacking as we do not know what the variables shown represent.

Theming (30 points)

In lecture we discussed the theming mechanism of ggplot and I showed you that we can customize themes to make for individualized presentations. Over the course of question 1 you will build your own customized theme and use it to display plots for the rest of the homework. As you build your custom theme be sure that your displays are easily readable and understandable.

For this section we will look at how the stock data trends with the treasury data. You will need to create a dataset that merges the stock data with the treasury data by finding an average monthly closing price for each stock and merging the monthly data with the treasury data. (Hint, you will want to create a month column in the stock data of the format "%Y-%m-01" so that it is easy to merge these dates with the treasuries data).

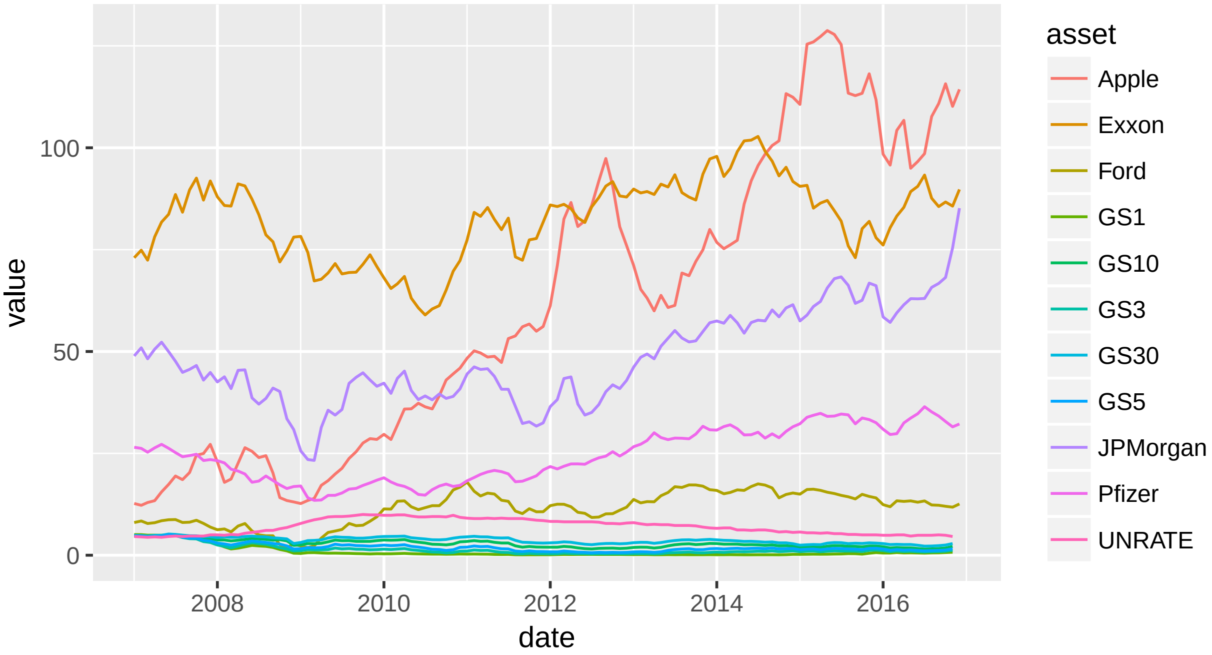
Question 1a. (10 points)

Read in the treasuries.csv and stocks.csv files.

- Find the average monthly closing price for each stock
  - (You will not need the columns whose names start with "r" for now)
- Merge your monthly stock data with your treasuries data
  - Remember back to the lesson when we used the lubridate package to work with aggregating time series data
  - Your dataset should only have data for the months with both stock and bond data
- Save your merged monthly data as `question1a`
- Don't forget to answer the free response question below

Let's start our analysis by looking at a line chart of our securities and stock returns over time.

```
question1a %>%
    gather(key = "asset", value = value, -date) %>%
    ggplot(aes(x = date, y = value, color = asset)) +
    geom_line()
```

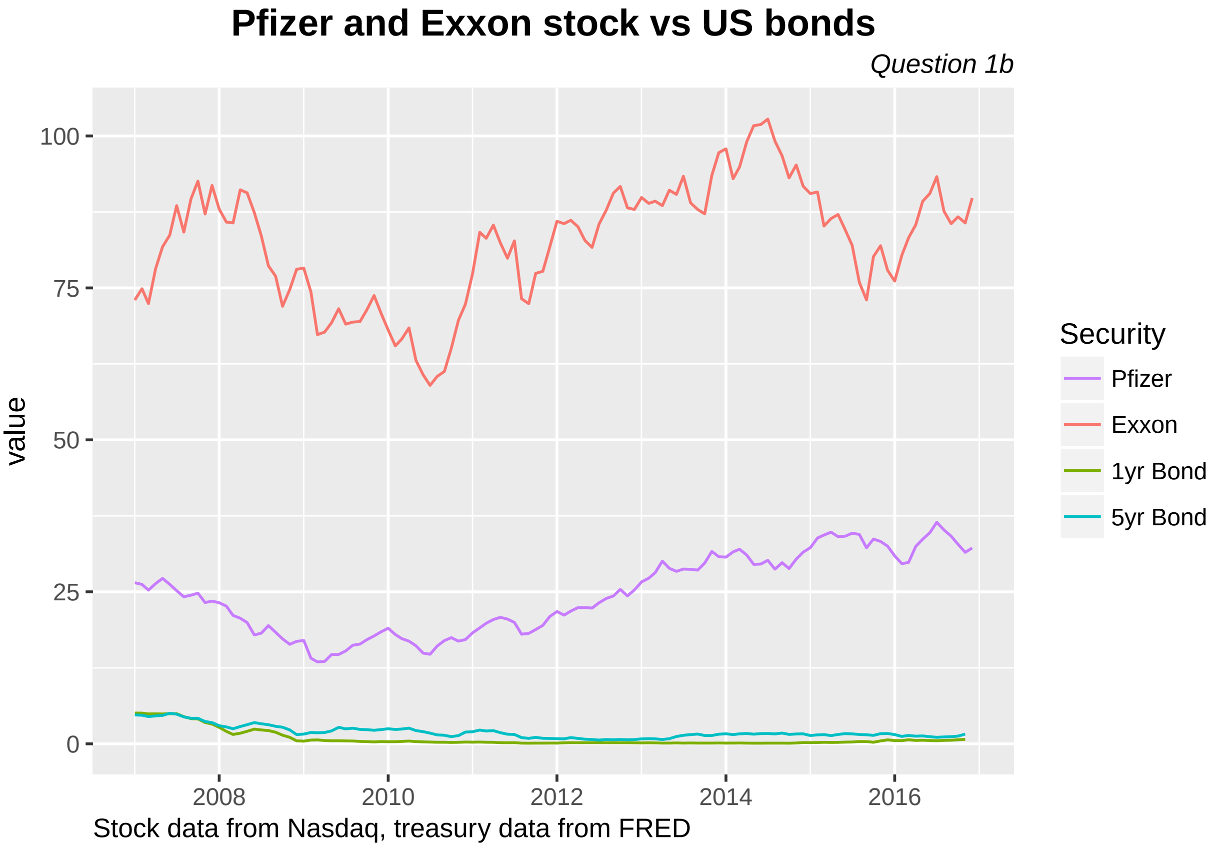


This is a terrible chart! Discuss at least 3 glaring problems with this chart and how you would fix these issues.

Question 1b (10 points)

For question 1b, your task is to recreate the below chart and help to fix some of the issues you discussed above.

- You should only plot data for Pfizer, Exxon, 1 year securities, and 5 year securities (4 lines)
- Be sure to add the correct title, subtitle, and caption
- For your color legend, make sure that you have the correct labels as shown below, and put the legend in the same order as below
- Use what we learned in class to update the following:
  - title - centered, size 14, and bold
  - subtitle - size 10, adjusted to the far right side of the plot
  - caption - adjusted to the far right side of the plot



Question 1c (10 points)

From a formatting perspective, our chart looks much better.

- We still have a fundamental problem: We are trying to compare prices with rates, this does not make sense.

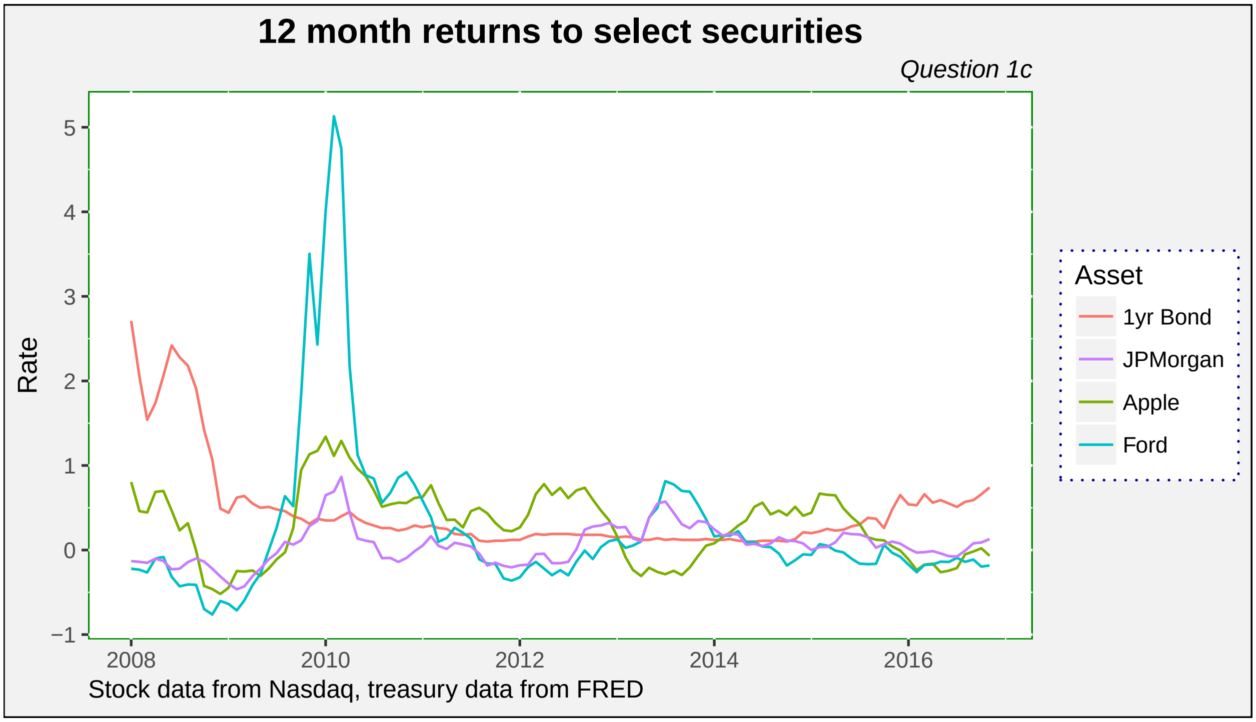
- Luckily we can you use our monthly closing price data to find the 12 month percent change in prices. We can compare this number to our annual interest rate
  - Say we have a closing price for December 2012. To find the 12 month percent change we would do (December 2012/December 2011)-1 which would give us our 12 month return rate
  - Hint: Check out `lag()` and think about how you can use it with `mutate()` to calculate this difference

Your first step for this question will be to calculate this 12 month return rate data for our stock data in question1a. Note that since quesion1a has data only from 2007 onward and we need 12 months of data in order to find the 12 month return rate, your output should only have data for 2008 onward.

Once you have calculated your 12 month return stock data plot the 12 month return rates for Ford, JPMorgan, Apple, and 1 year bonds over time.

- Be sure to provide appropriate labels and orders in your color guide
- Be sure to provide correct title, subtitle, etc…

- In addition to the text formatting of question1b, we are going to continue customizing the visualization
  - Change the plot background color to be "gray95" and give the background a solid black border
  - Change the panel background color to be "white" and give it a solid border of the color "green4"
  - Change the legend background color to "white", give it a dotted "blue4" border of size 2
- Call your output plot "question1c"



-->

Mapping Part I (40 points)

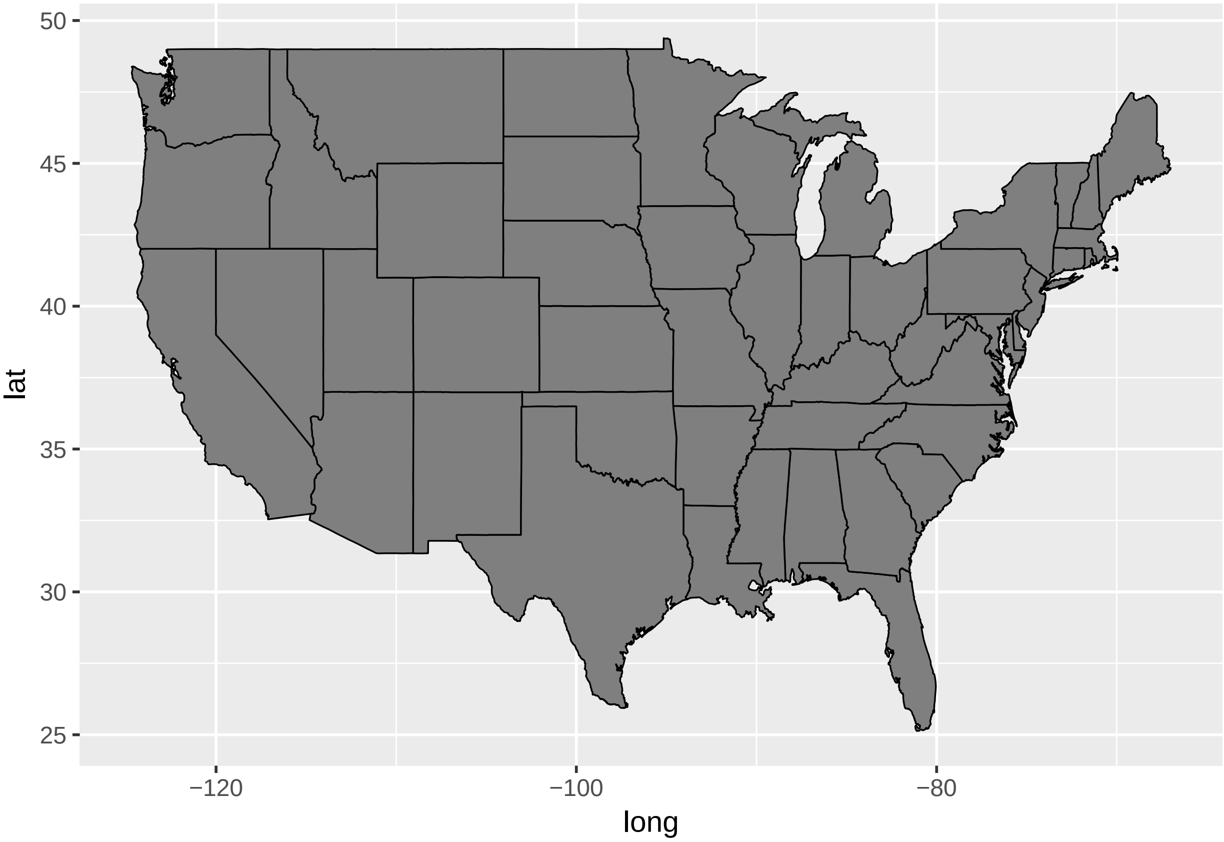
Heatmaps allow us to show location based data and trends. In this homework we are going to analyze the election results of the 2016 election. In this question we will look at the national results before later looking at the results for the state of Ohio.

Disclaimer: We will not be including Alaska and Hawaii on our maps for this exercise.

You will need the national_election_results csv for this exercise.

- We will also need to get the latitude and longitude data.
  - To set up this data, use the map_data function in the ggplot2 package
  - make sure you have the `maps` library loaded
  - for the `map` argument use "state".

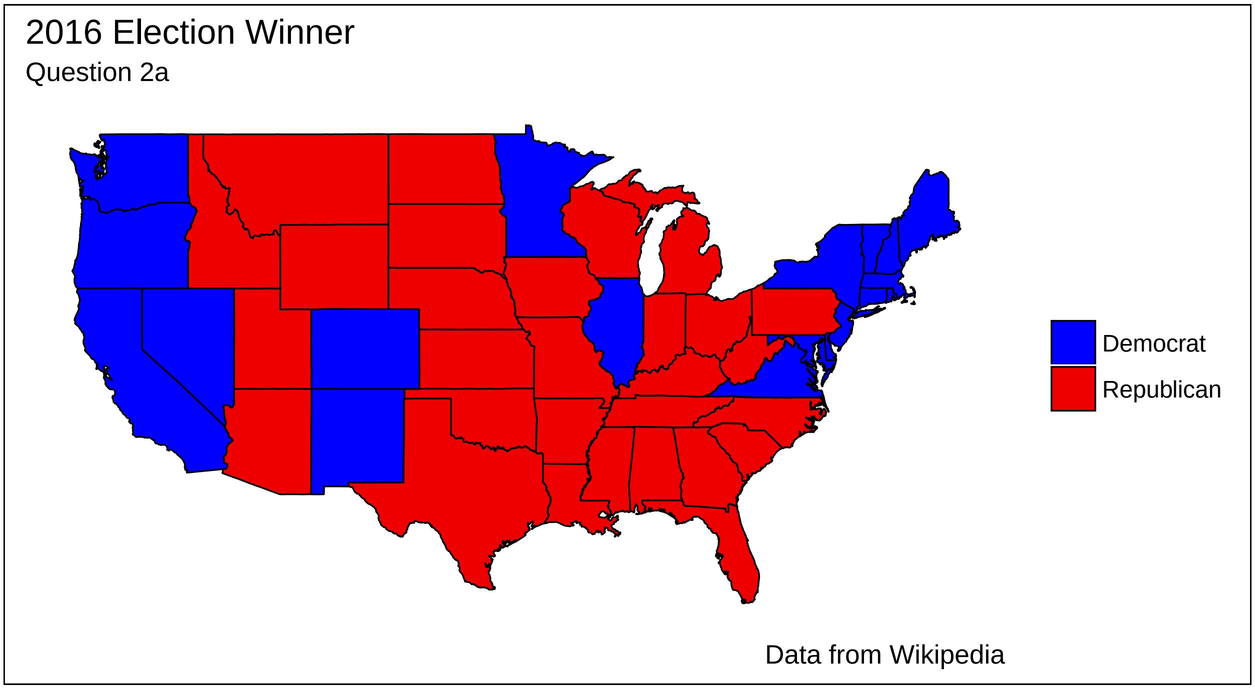
```
states <- map_data("state")

states %>%
    ggplot(aes(x = long, y = lat, group = group)) +
    geom_polygon(fill = "gray50", color = "black", size = 0.3)
```



Question 2a (15 points)

Read in the national election results data if you have not done so already. Using the columns showing the vote totals for each party, create a column which shows the party which won the state.

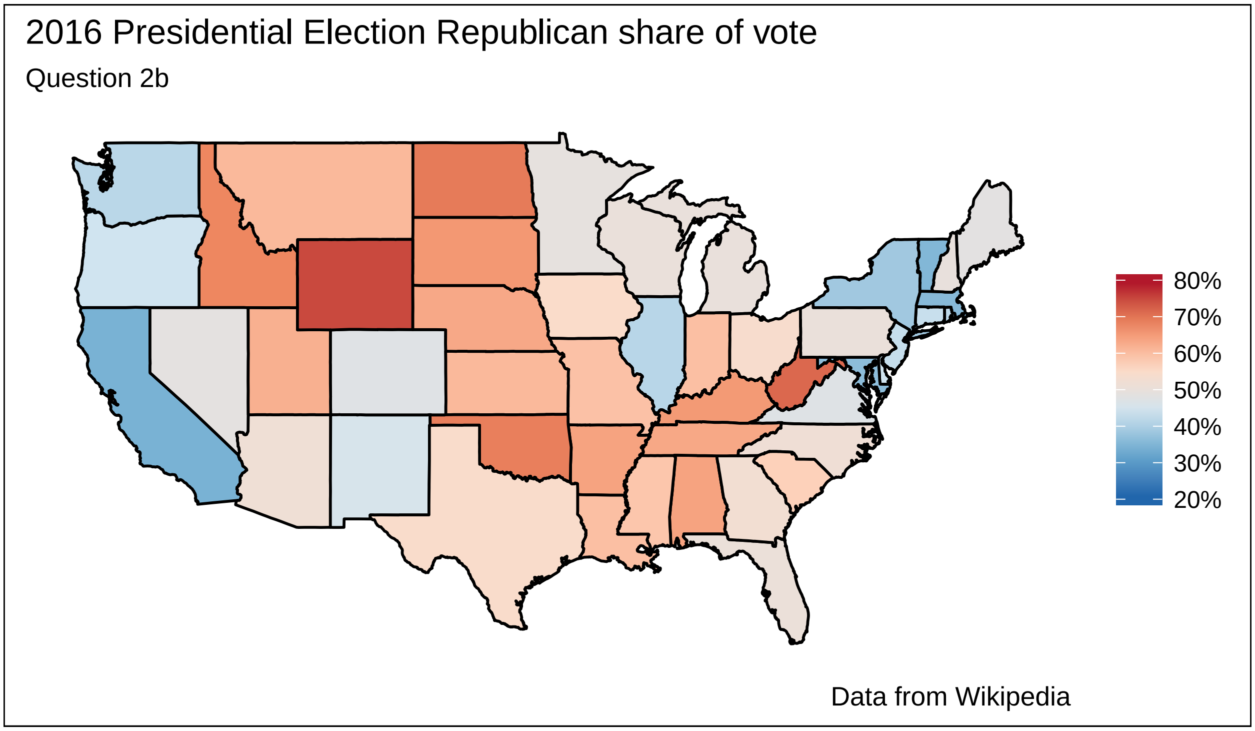
- Once you calculate the party to win the state, merge your election results data with the states data so we can make a heatmap showing the winner of each state.
- For this plot you do not need to have x and y labels
- Since we are showing a map, be sure to use the function discussed in class that sets our coordinate axes appropriately for displaying map data
- Additionally, you will want to turn off the panel background and all tick marks and axis text
  - use `theme()` and `element_blank()` to do so
- Make sure to you have appropriate title, subtitle, and caption
- For your fill guide make sure that states in which the Democratic candidate won are blue while states in which the Republican candidate won are colored "red2"



Question 2b (10 points)

Instead of just looking at the outright winner, let's see how close the results were. For question 2b take a look at the percentage of the total vote in each state won by the Republican candidate.

- For this map you will want to use the `scale_fill_distiller()` function and the "RdBu" palette
- What does this map say about the distribution of political beliefs across the country? Do you see any regional groupings that all lean the same way or does the distribution of voting seem random?



2016 Presidential Election Republican share of vote

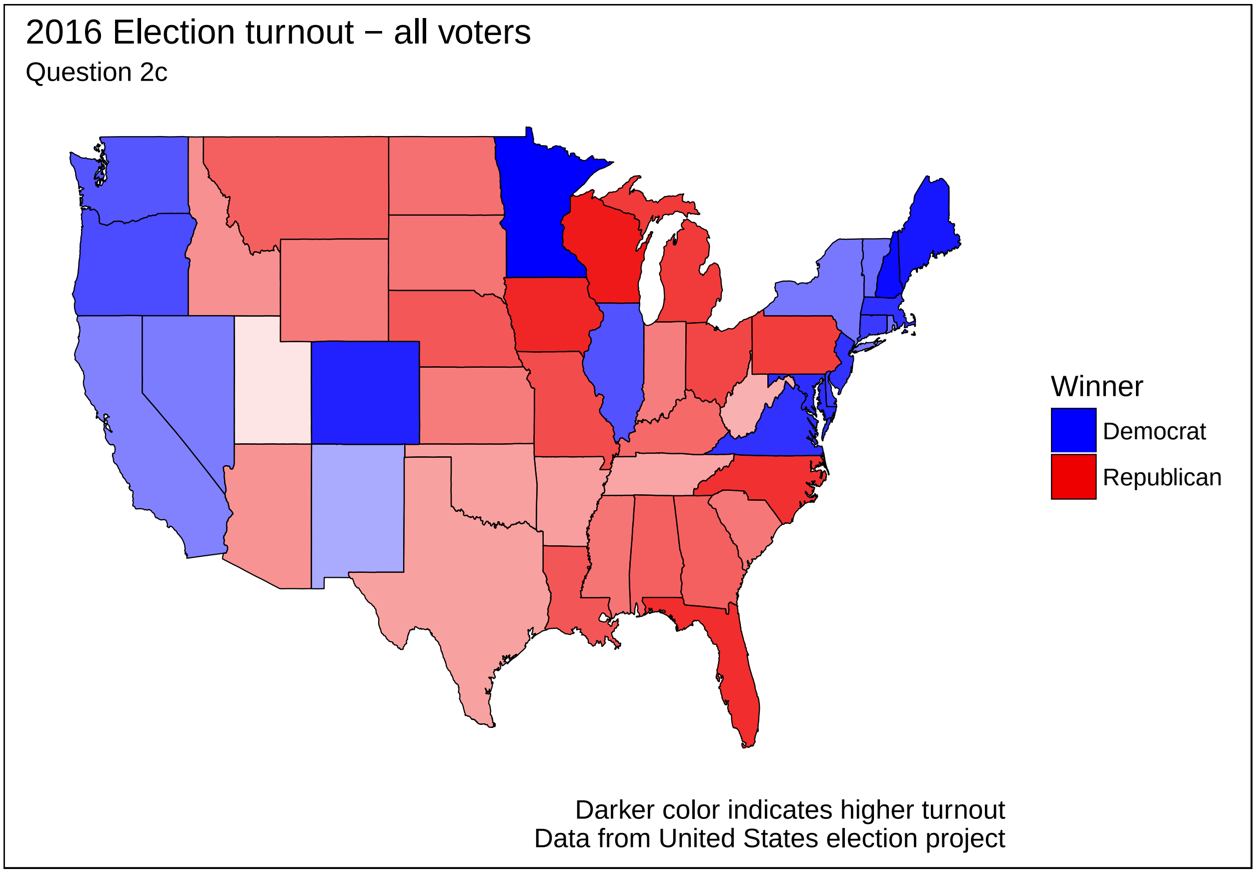
Question 2b

Data from Wikipedia

Question 2c (15 points)

Finally, let's take a look at the percentage of the voter eligible population to vote in each state. Use the VEP column in the national election results data as the total number of eligible voters in each state.

- You will have to calculate the total percentage of eligible voters to vote yourself using the vote count data
- For this exercise we are going to vary two aspects of the map
  - The fill color will correspond to the party to win the state
  - The saturation (alpha) of the fill will correspond to the percentage of the state's eligible population to actually vote
    - (Note that we turned off the guide for alpha)
- What takeaways does this map provide in terms of turnout in US elections?
- Do you think this is a clear visualization? What other ways of showing the share of eligible voters data could we have used to communicate this same idea?



2016 Election turnout − all voters

Question 2c

Darker color indicates higher turnout
Data from United States election project

Mapping Part II (30 points)

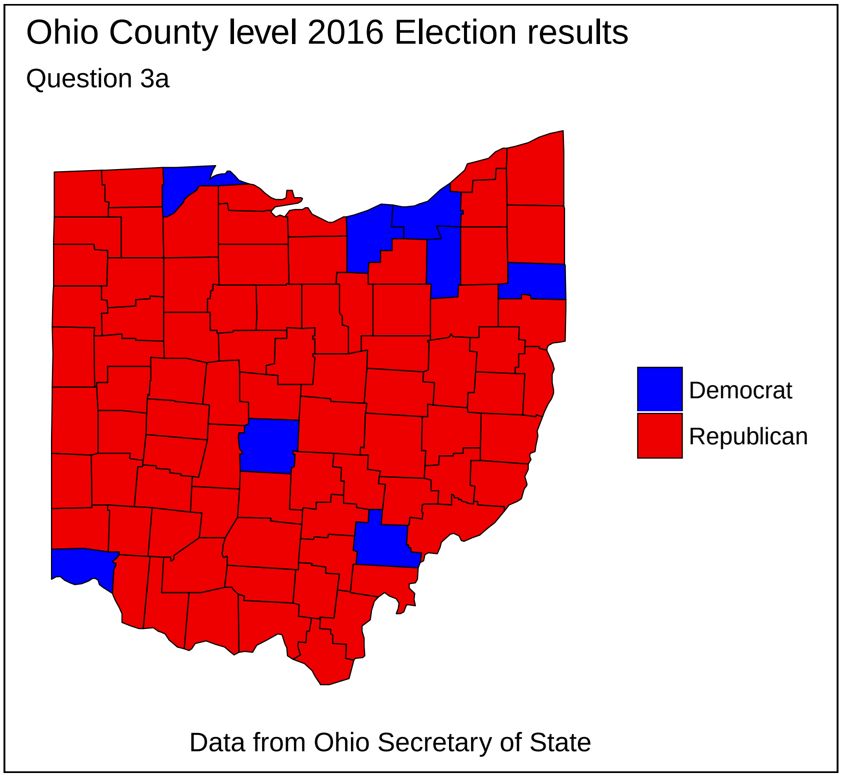
Historically, many states in America vote for the same party in every presidential election. As a result, elections come down to who can win a few battleground states which sometimes vote Democrat and sometimes vote Republican. These "swing" states are the most important for a campaign and tend to be where nominees spend much of their time. One of the most important swing states in the USA is Ohio, a state which, interestingly enough, has voted for the eventual winner of the presidential election in every election since 1964. (John F. Kennedy was the last president to lose Ohio).

The oh_election_results.csv file has some historical election results at the county level.

Question 3a (15 points)

To get the latitude and longitude data needed for the map we will again use the map_data function, but this time use the "county" map data and specify the region as "ohio" to get a data frame with the outline for each Ohio County

- Using the data in the Ohio election results file calculate the winning party in each county and make a map showing the county electoral outcomes.
- Make sure to use appropriate titles and labels, and use the same theme elements as used in the above question showing national election results
- After you create your map calculated the margin by which Donald Trump won the state
    - What percentage of the total Ohio vote did he get?
- Knowing the fraction of total votes in the state won by Donald Trump and looking at the relative number of blue to red counties, what conclusion can you make about counties that voted for Hillary Clinton?


Ohio County level 2016 Election results
Question 3a
🟦 Democrat
🟥 Republican
Data from Ohio Secretary of State

Question 3b (15 points)

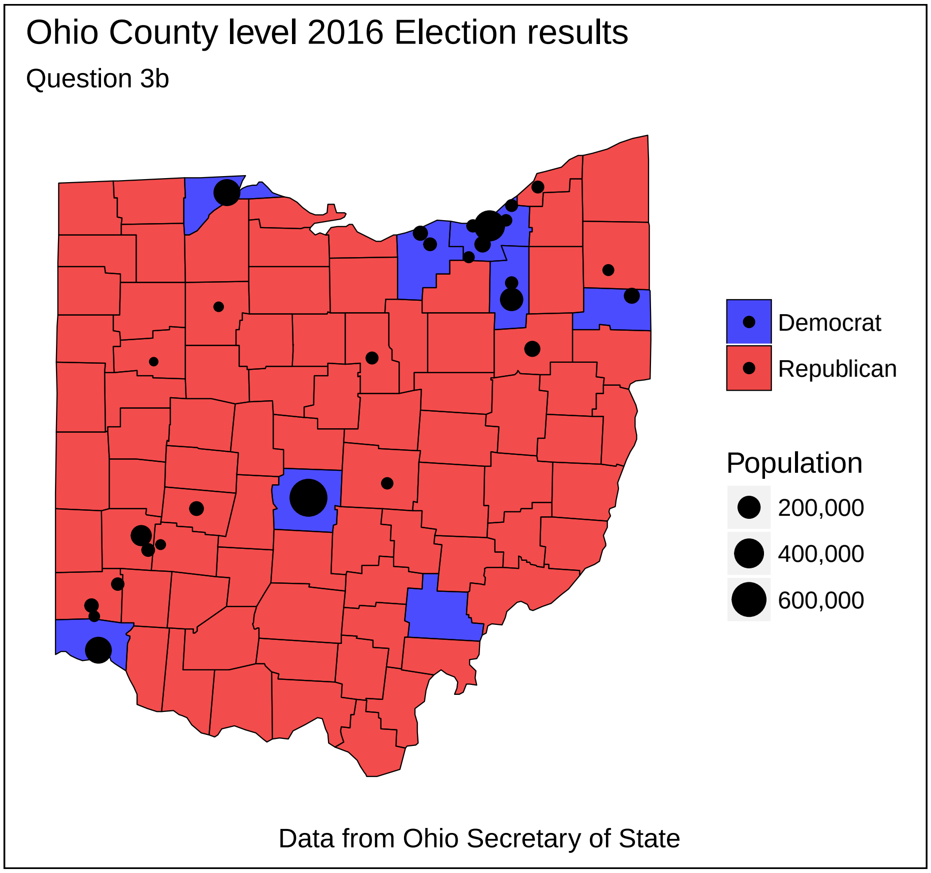
Despite the fact that Donald Trump won the vast majority of counties in the state, his margin of victory in terms of overall votes was not particularly large. We can add on data about city size to our map to get a sense of where and how large the population centers of the state are.

The code below pulls data for the cities in Ohio with at least 40,000 people.

```
ohio_cities <- us.cities %>%
    filter(country.etc == "OH") %>%
    select(pop, lat, long)
```

- Your goal is to use the above data on cities in the state to create the map below.
  - Each point represents a city in the state and the size of the point is proportional to the population of the city
  - You will want to accomplish this by using a geom_point layer
    - Note that you may run into errors about certain unspecified aesthetics, remember how we previously told ggplot that no value exists (for example an axis title).
  - Check the `scales` package to find a function that will nicely format our labels for population similar to what we did in class for percentages
- What do you notice about the groupings of cities, especially larger ones, in the state and the election results for those areas?
  - What do you think this says about the urban/rural divide in Ohio?

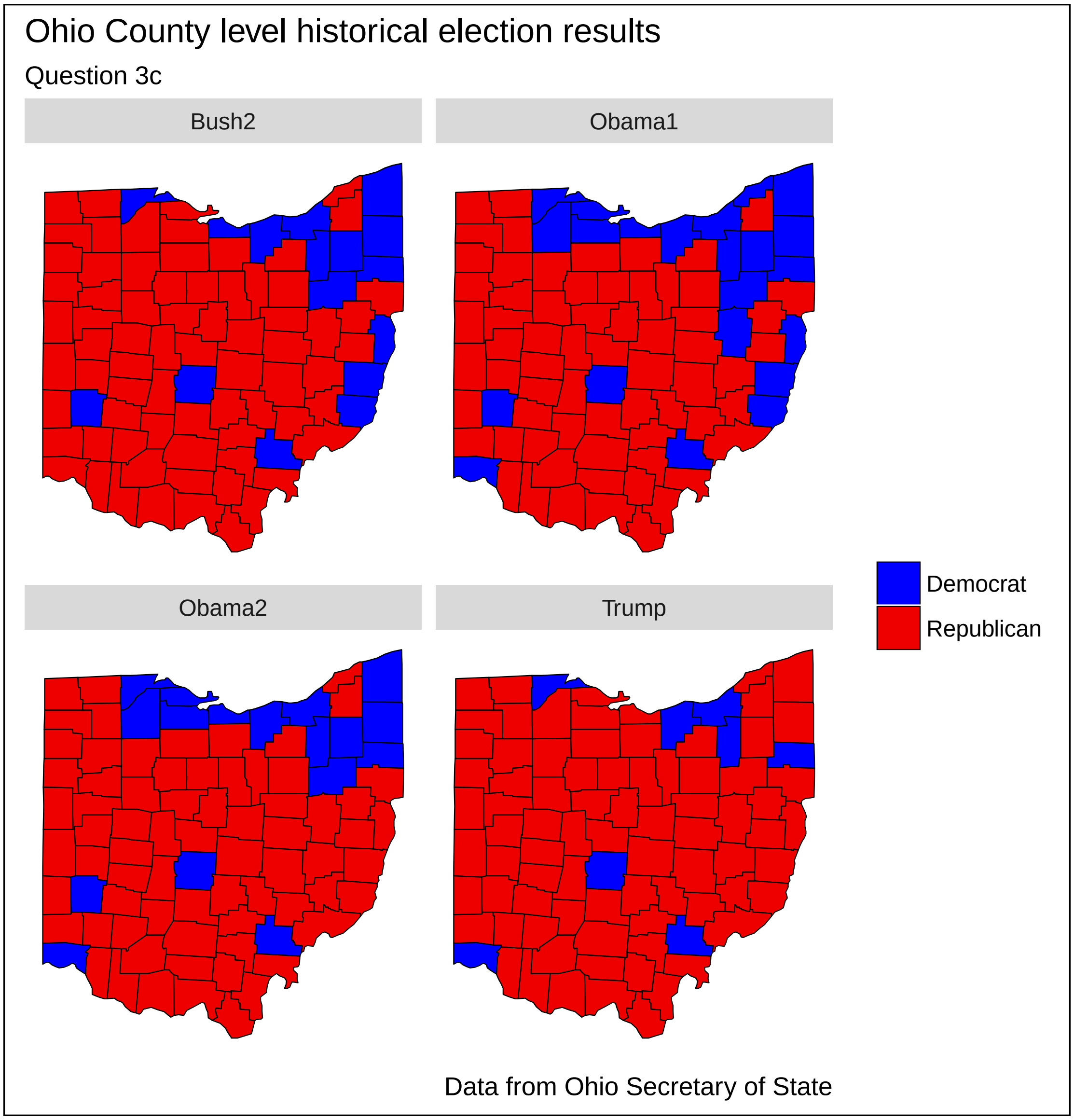
In the version of the map below I lowered the alpha value for the map fill color to increase the contrast of the city dots. You do not need to do this.


Ohio County level 2016 Election results
Question 3b
Democrat
Republican
Population
200,000
400,000
600,000
Data from Ohio Secretary of State

Extra Credit (10 points)

Now that we have seen the result for this past election, let's look at how this election's results compared with previous elections. Our Ohio election results data has multiple columns which include data from past elections. Let's compare this election with the three previous: the first and second elections won by Barack Obama and the second election won by George W. Bush. Here are the two candidates for each election (winner listed first):

- Donald Trump vs. Hillary Clinton (2016)
- Barack Obama vs. Mitt Romney (2012)
- Barack Obama vs. John McCain (2008)
- George W. Bush vs. John Kerry (2004)
- Note that in our dataset, a president running for a second time has a 2 in the column name for data related to that election
  - You will want to compare obama2_count with romney_count


Ohio County level historical election results

Question 3c

Bush2

Obama1

Obama2

Trump

Data from Ohio Secretary of State