Visualization Part I: Homework

FRB/Howard Instructors

Date Assigned: Friday September 1, 2017

Date Due: Thursday September 7, 2017 by 11:59 pm

Introduction

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 questions that ask you for commentary in addition to or in place of a plot please write comments in the code right below your code for the corresponding plot.

For each plot you produce assign the plot object to a variable associated with the corresponding questions. i.e:

```
question1a <- ggplot(...) + ...
```

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 and stock_closings csv files in the ggplot Homework folder. You will have to read it in by yourself.

Additionally: be sure that you `library()` the `ggplot2` and `tidyr` packages for this homework.

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

Question 1a: (5 points)

- Start your answer script by writing a useful comment at the top including your name, the assignment, and the date
- Use the library function to attach the `ggplot2` package.

Question 1b: (5 points)

- Read in our treasuries and stocks data from `treasuries_long.csv` and `stock_data_long.csv`
- Take a look at the classes of the columns in the data, do you need to manipulate any of them from character to numeric etc?

Stock Portfolio Analysis

For this next section we will concentrate on learning about different types of geoms available in ggplot as well as how to deal with overlapping data. For this segment we will need the stock_data_long and treasury_long csv files.

Question 2a (10 points)

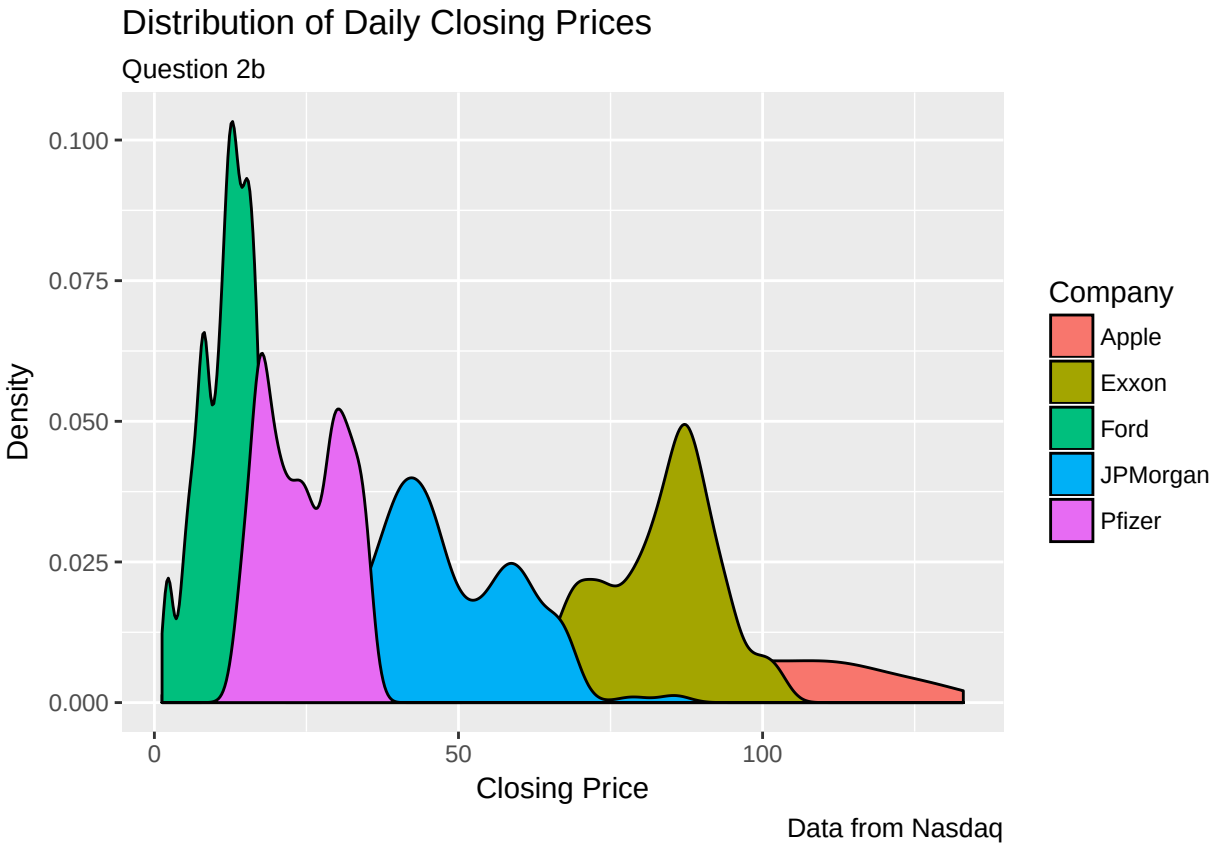
Let's start by just looking at the stock_data_long csv to get a sense of the data.

- Print out the first 3 rows of the dataset and remark on the data.
  - (Frequency, is it all the same format, how long is the series, is it sorted the way we want, what type of work will we need to do on the data in order to use it?).
- We want to make a line plot of the daily closing prices for our stocks, so we want all observations where the value in the company column is one of: (Pfizer, Apple, Ford, JPMorgan, Exxon).
  - Remember how we used the %in% operator in class, you will want to do this to pull out our relevant data
  - Make sure you don't pull out observations where the company name starts with "r"
- Construct an appropriately labelled line plot of the daily closing prices for each company differing color by company.
  - Be sure to include a caption indicating data source (stock data source is Nasdaq).
  - In a comment below the chart discuss your takeaways.

Question 2b (8 points)

So let's look at the same data in another way.

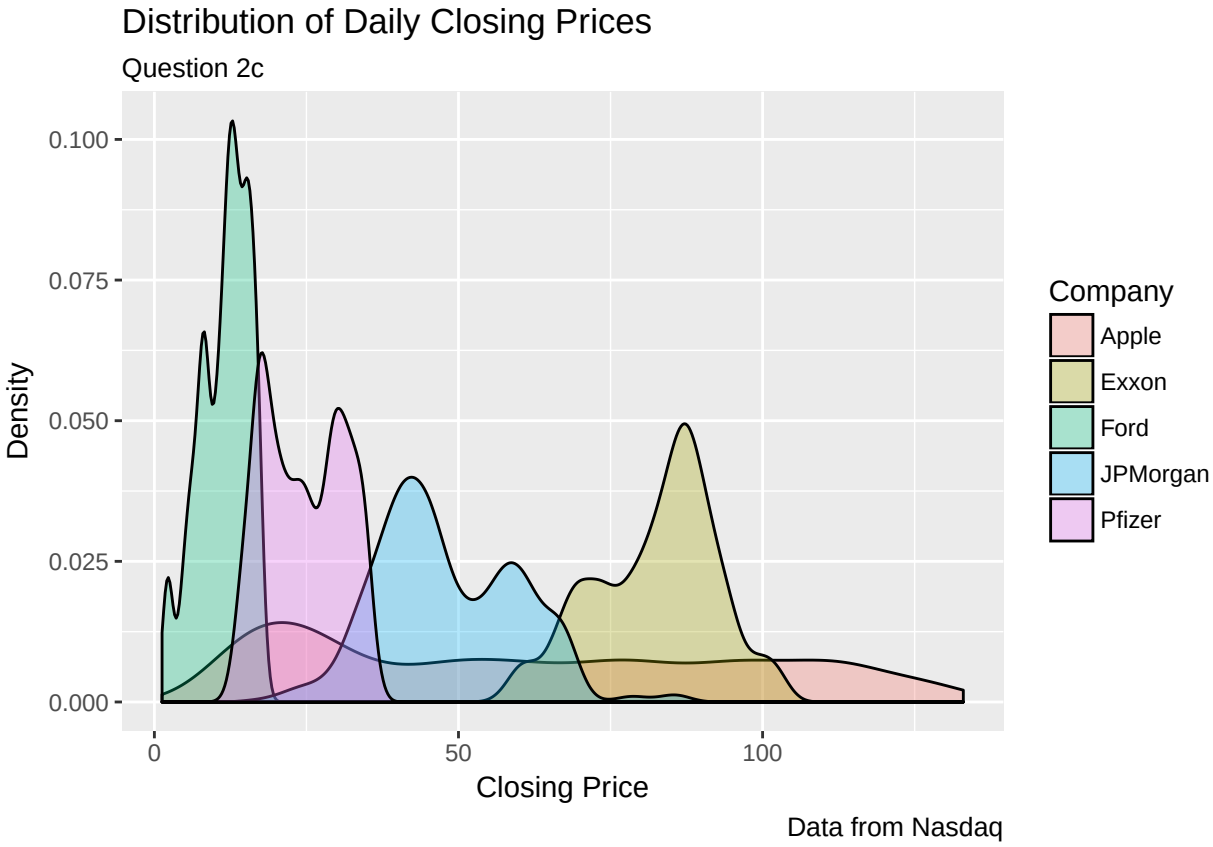
- Now let's look at the distribution of closing prices for each company over the period.
  - Instead of regular line plots, we will make smoothed histograms using geom_density.
    - Set the fill for each density distribution to be the different company.
- Take a look at the chart, what does this chart show? What are some problems with this display?



Question 2c. (8 points)

All the density distributions overlap! We need a way to make each curve somewhat transparent so that we can see all the curves beneath it.

- What was our solution to this problem in the lecture?
- Implement that same solution in your code to display a density chart where it is possible to see the underlying distributions.
- Based on this chart which company has the largest variability in terms of closing prices?
- Which company looks like it has the smallest variability?
- The return on a stock is the difference between the price you buy the stock at and the price you sell the stock for, based on this chart can we say anything about the return on the stock for any of these companies?
  - What further data would we need in order to discuss returns?
- The percentage return on a stock is the percentage difference between the price you buy the stock at and the price you sell the stock at, based on this chart, which company would the possibility of the largest percentage return assuming you bought the stock at it's lowest point and sold at the highest?
  - Which company would have the second largest possible percentage return?



Question 2d (12 points)

We can also specify alpha in a scale_fill_manual call in the values argument. Here we will instead call alpha in the scale function and we will also specify the colors we want.

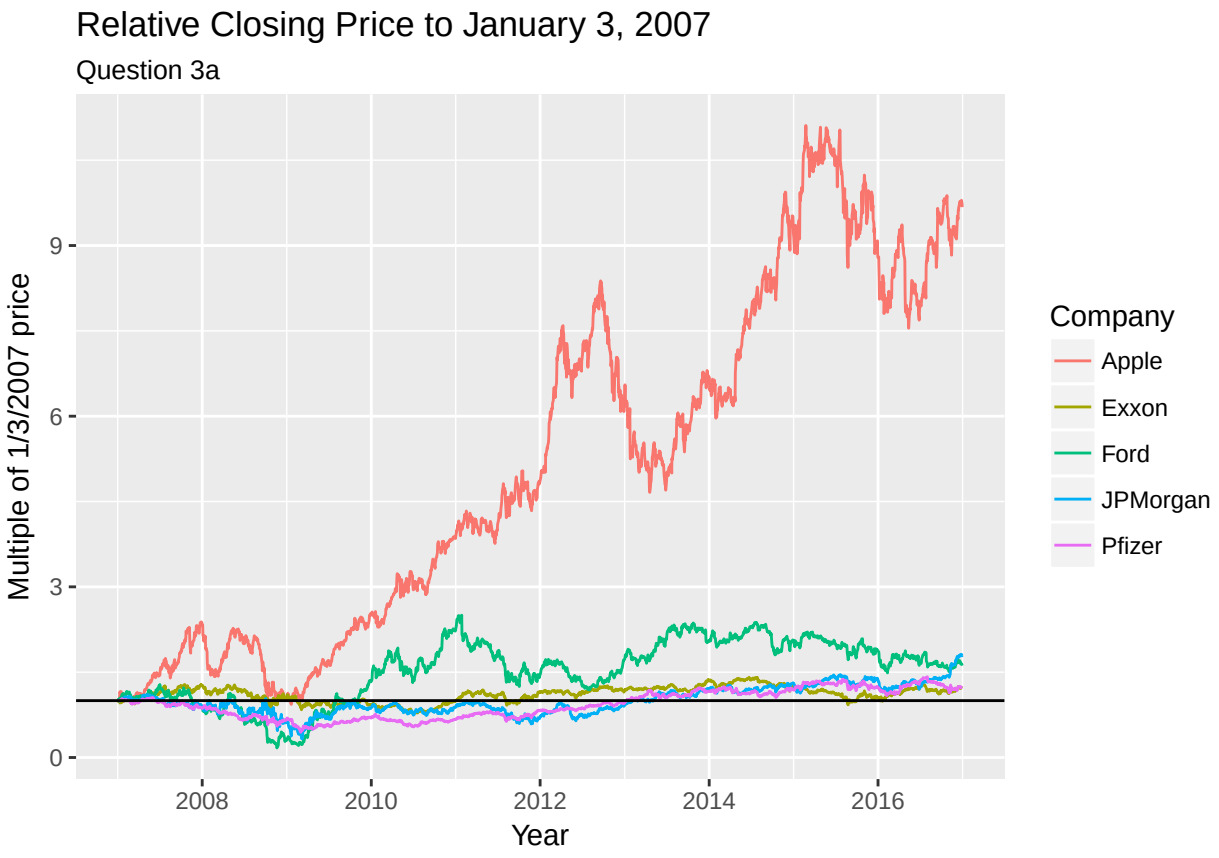
- alpha takes arguments of the form `alpha(colour, value)`.
  - If we do not specify the `colour` argument the defaults are used.
- This time specify the following vector of colors: coral, darkgoldenrod1, cyan, forestgreen, hotpink.
  - Provide an alpha value of 0.45.
- Below write your takeaways from the plot:
  - Does this distribution tell you anything about the quality of investment each stock would be?
  - In evaluating a stock, what do you think an "ideal" distribution would be if one exists?
    - If one does not exist, how do you think that a density chart such as this could help someone in making a decision of which stock to buy

Relative Stock Closings

Question 3a (14 points)

Let's now turn to looking at ways of comparing growth among the companies instead of looking at raw closing price values.

- One way to do this is to find the relative closing price of each company relative to itself at the start of the period.
  - I.E. the closing price on day one would be 1.0, `stocks$Apple[1]/stocks$Apple[1]` and each other day would be a percentage of the first day, `stocks$Apple[n]/stocks$Apple[1]`.
  - The observations in the stock_data_long.csv file where the value in the company column begin with the lowercase letter r show the results of these computations.
- Make a line-graph of these relative closing prices similar to question 2a.
  - Use `geom_hline` with a yintercept of 1 to add a solid black line at a y value of 1
- If you could go back to January 2007 with $100, which stock of these should you buy if you wanted to sell for the most profit?
- Which stock showed the second highest relative closing prices on average over the time period shown, does this surprise you?
- What does a y value greater than 1 signify, less than 1?



Question 3b (5 points)

Now with our relative price data we can look at our density plots again but this time use the relative data.

- Show the new graph of the density distributions for the relative closing prices.
- How does this chart add to our current understanding of the relative returns on our chosen stocks?
  - If it does not add anything, why not?

Question 3c (9 points)

Now make the barcharts for the relative price data.

- Again do not show the fill scale. Make sure that there is a mark at the value of 1.0 on the y-axis.
  - Is this a valuable chart, what does this chart show us that the barplots of the raw closing price do not?
  - What do you take away from the data displayed in this chart, what investment decisions, if any do you think this chart points to?

Question 3d (12 points)

What if there was way we could combine a boxplot with a density distribution. Great news! There is, it's called a violin plot.

- Using the relative closing price data from the boxplots above, make a violin plot of the data.
- Below discuss what a violin plot is able to show that a traditional boxplot does not, do you prefer this plot over the boxplot? Why or why not?
  - Does this plot lead you to draw differnt conclusions than the boxplot?

Challenge Question

We will start by reviewing how to add chart elements to a graphic by using ggplots layering grammar. Let's take a look at the unemployment data in the treasuries.csv file.

Question 4a. (0 points)

Using the UNRATE data in the treasuries file create an appropriately labeled graph of the unemployment rate over time from February 1977 onward. What is the value of this chart? Does the chart point to anything about the cyclicality of employment? What further questions do you have upon viewing this chart? Answer these questions as comments below your code for the plot.

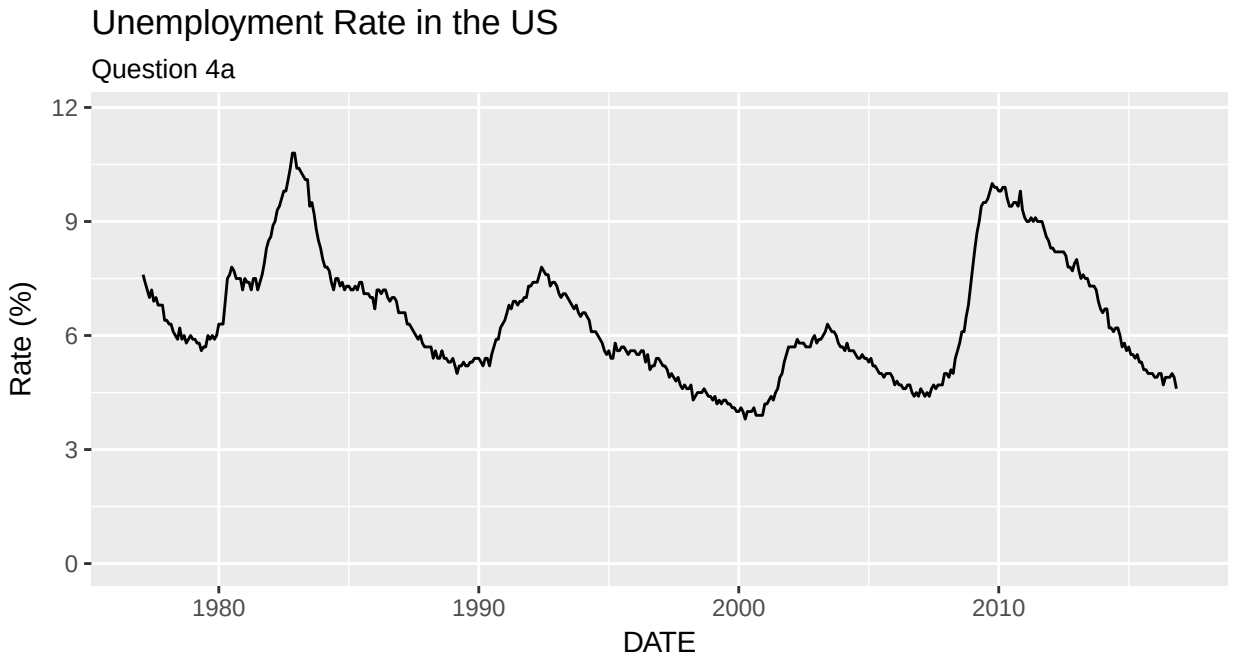
I am including the chart/code below as a template for what your answers should look like throughout this assignment.

```
## basic lineplot of unemployment rate data from the treasuries file

plot.data <- treasuries[treasuries$DATE >= as.Date("1977-02-01") &
                            treasuries$measurement == "UNRATE", ]

question4a <- ggplot(plot.data, aes(x = DATE, y = value)) +
    geom_line() +
    scale_y_continuous(name = "Rate (%)",
                       limits = c(0, max(plot.data$value) + 1)) +
    ggtitle("Unemployment Rate in the US",
            subtitle = "Question 4a")


# The chart seems to hint at underlying cyclicality in employment
# cycles but without context the value of the chart is weak.
# For example, how does this US employment chart compare with other
# similar countries over time.
question4a
```



Question 4b. (4 points)

Let's see how the full range of unemployment data looks by presidential cycle.

- Using the "presidential" dataset included with ggplot2, graph vertical lines for the start of each president's time in office.
  - This dataset is automatically loaded when you ran `library(ggplot2)`, if you run `head(presidential)` you can see the first few rows of the data.frame.

- You will want to use the geom_vline function with the xintercept argument to add the vertical lines.
  - You will need to use the as.numeric function around the start dates to graph them without throwing an error.

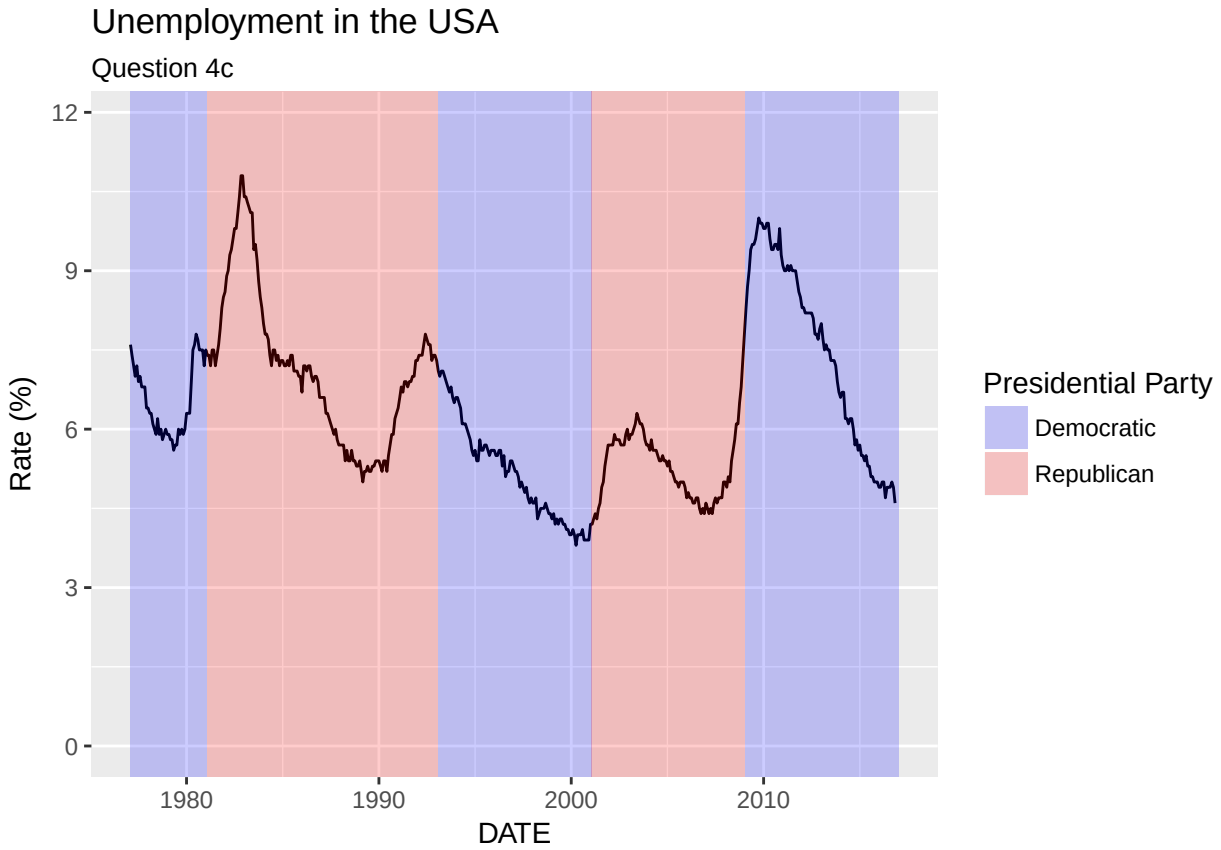
- Be sure to add a caption explaining what the lines are and add text below the chart discussing the value of the chart as well as further questions.

Question 4c (4 points)

Now let's instead take a look at the unemployment rate when different parties controlled the white house instead of different individuals again using the presidential dataset.

- Since our treasury dataset only goes back to 1977, you will want to filter out the rows in the presidential dataset with a "start" value less than January 1, 1977.
  - To use an additional dataset from the main one when making a plot we need to specify our new dataset in the `data =` argument of our `geom_` function.
- You will want to use the geom_rect function in ggplot to create our background.
  - For the geom_rect `aes()` call, your xmin values should be "start" values and your xmax values should be "end" values.
  - Your y-values should be from the bottom of the plot to the top of the plot, you can use the values of `-Inf` and `Inf`.
  - You will likely need to specify an x and y value of `NULL` within the `aes()` call
  - Your fill values should correspond to the presidential "party."
    - Add the following scale_fill_manual call to your plot: `scale_fill_manual( name = "Presidental Party", values = alpha(c("blue", "red"), 0.2))`

Be sure to answer what you think of this plot, do you see any evidence for employment corresponding with the party in power? What further information would you like?



Question 4d (4 points)

We just filled in the background using geom_rect().

- Now use geom_text to fill in the name of the president at the bottom of the chart at the appropriate date.
  - Take a look at `?geom_text` to see the help page for the function.
  - Make sure to angle your text so that each name can be clearly seen.
  - Additionally, add a label in the top right of your chart which says "Unemployment rates vary during presidential cycles."
  - Your label should take up two lines.
    - Use the \n symbol in your text string to insert a line break.