# Project 2

Marques Chacon

2/20/2020

## SVD Analysis of Partisanship in U.S. Congress

### Introduction – Loading the SVD Visualizations

As someone who enjoys politics, as well as diving into data analyses, I was curious about the partisanship of representatives and senators in the United States Congress. This would lead me to consider some questions that I sought to answer – How does partisanship compare from the 90th Congress vs. the 116th Congress? How does partisanship vary across states? Do certain states have more "extreme" partisans? Are there any discernible trends among different age groups? In order to resolve these questions, I first performed Singular Value Decomposition (SVD) for the four datasets to extract the data:

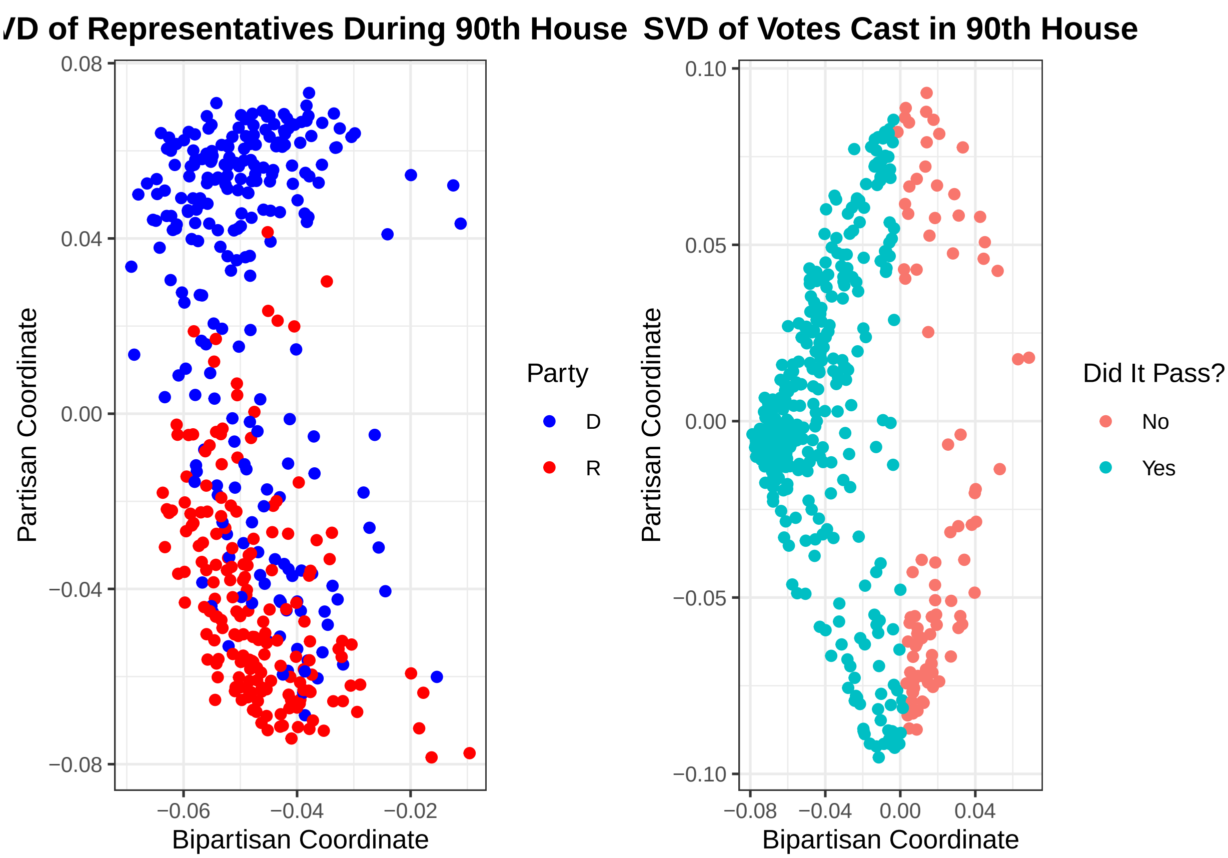



Figure 1: Figure 1: SVD Composition of the 90th House

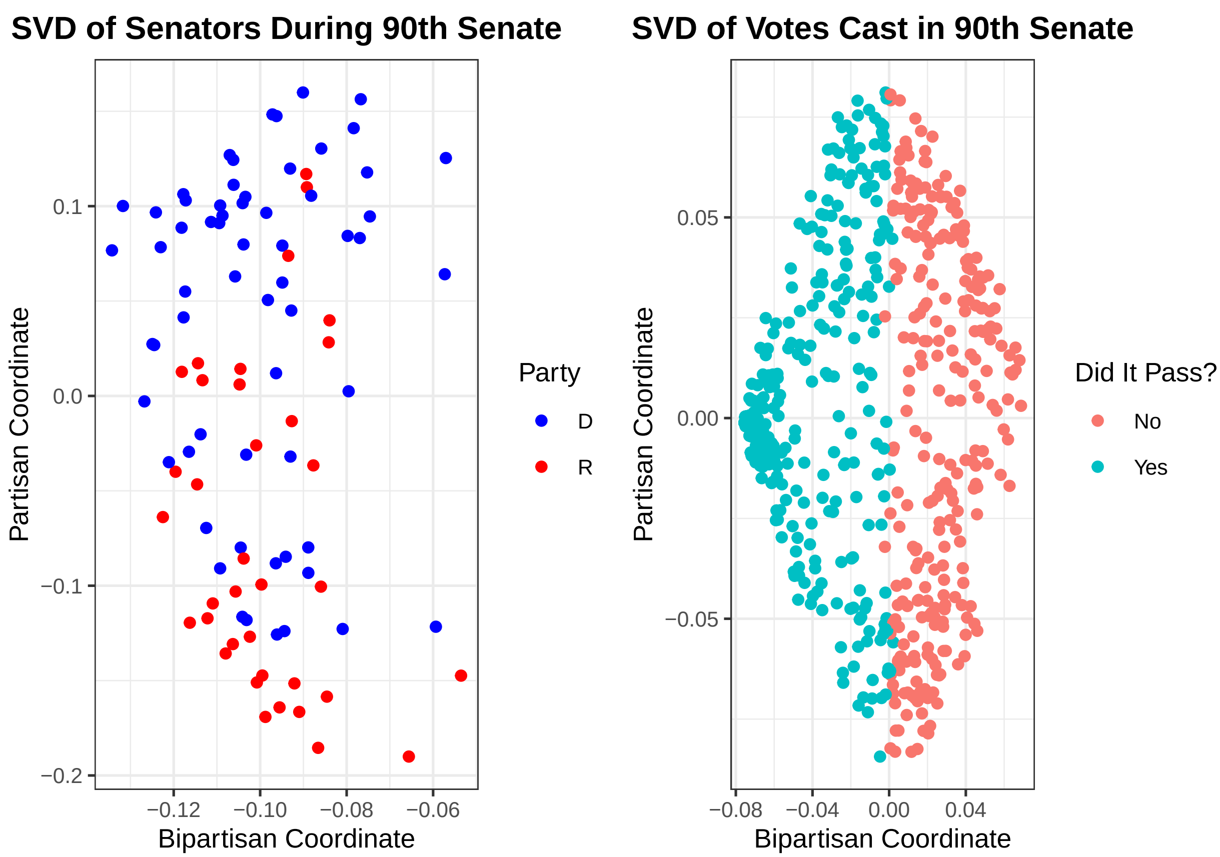



Figure 2: Figure 2: SVD Composition of the 90th Senate

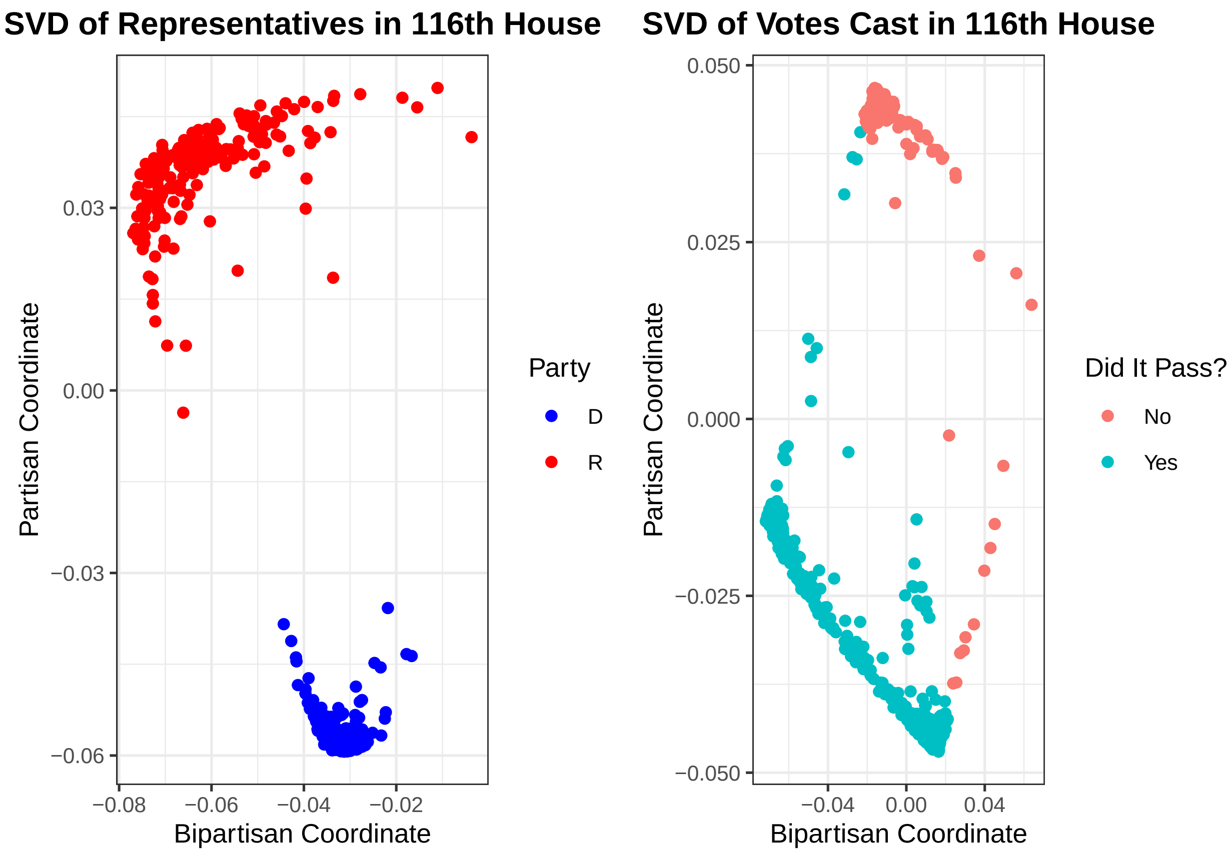



Figure 3: Figure 3: SVD Composition of the 116th House

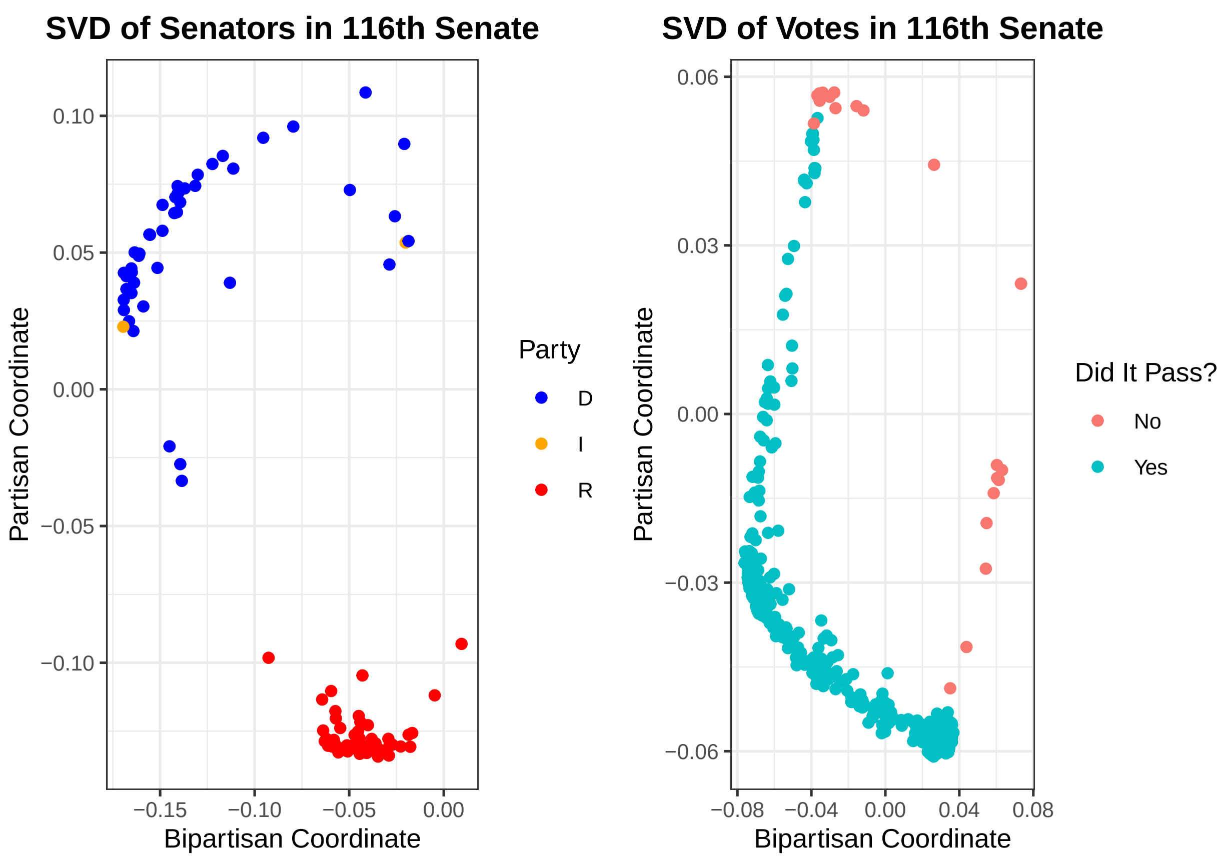



Figure 4: Figure 4: SVD Composition of the 116th Senate

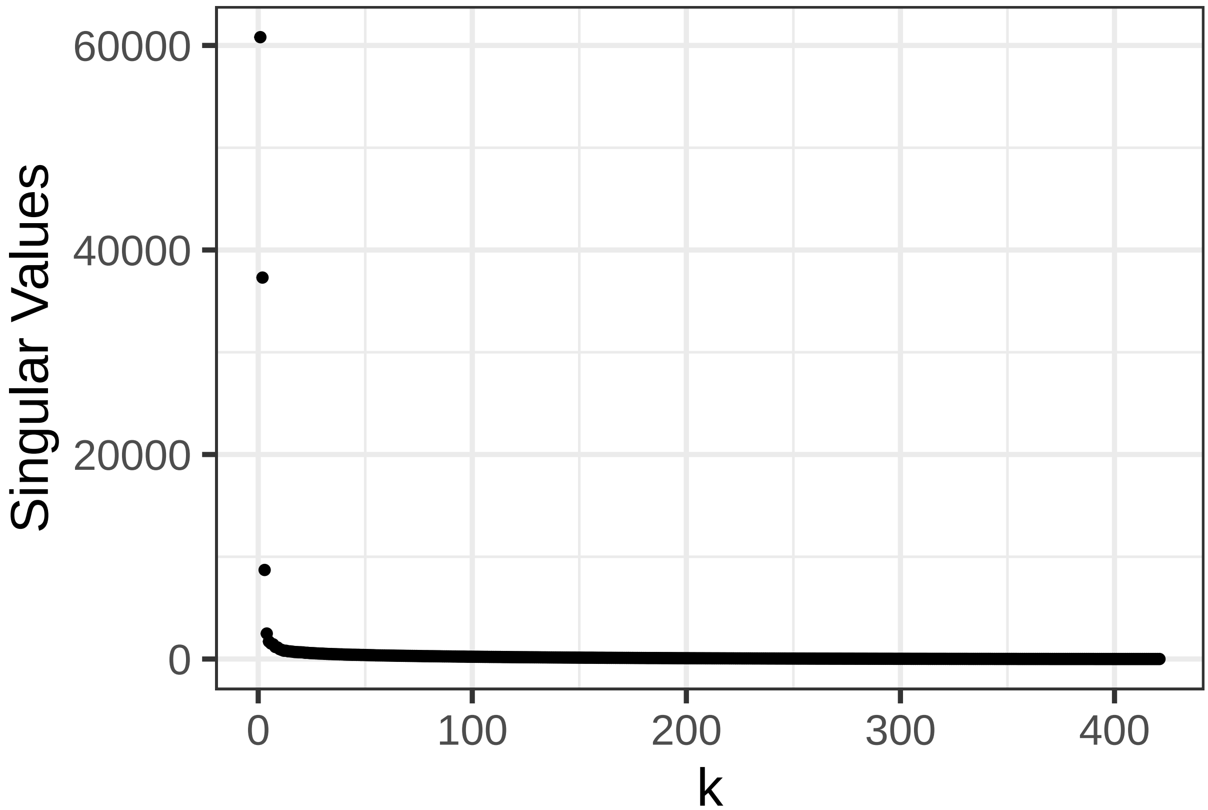
The first thing to note is that when constructing these visualizations, I filtered the datasets differently depending on if I was analyzing the clusters of representatives or the clusters of votes. For the clusters of partisans, I decided to remove all NA values, which meant having to remove some people from the sample. I chose this instead of setting `NA`'s to `0`, since a `0` may not accurately reflect someone's position on a roll call, and that would compromise the legitimacy of our data. Conversely, when analyzing the votes, I chose to set all `NA`'s to `0`, as it would not affect the status of the vote.

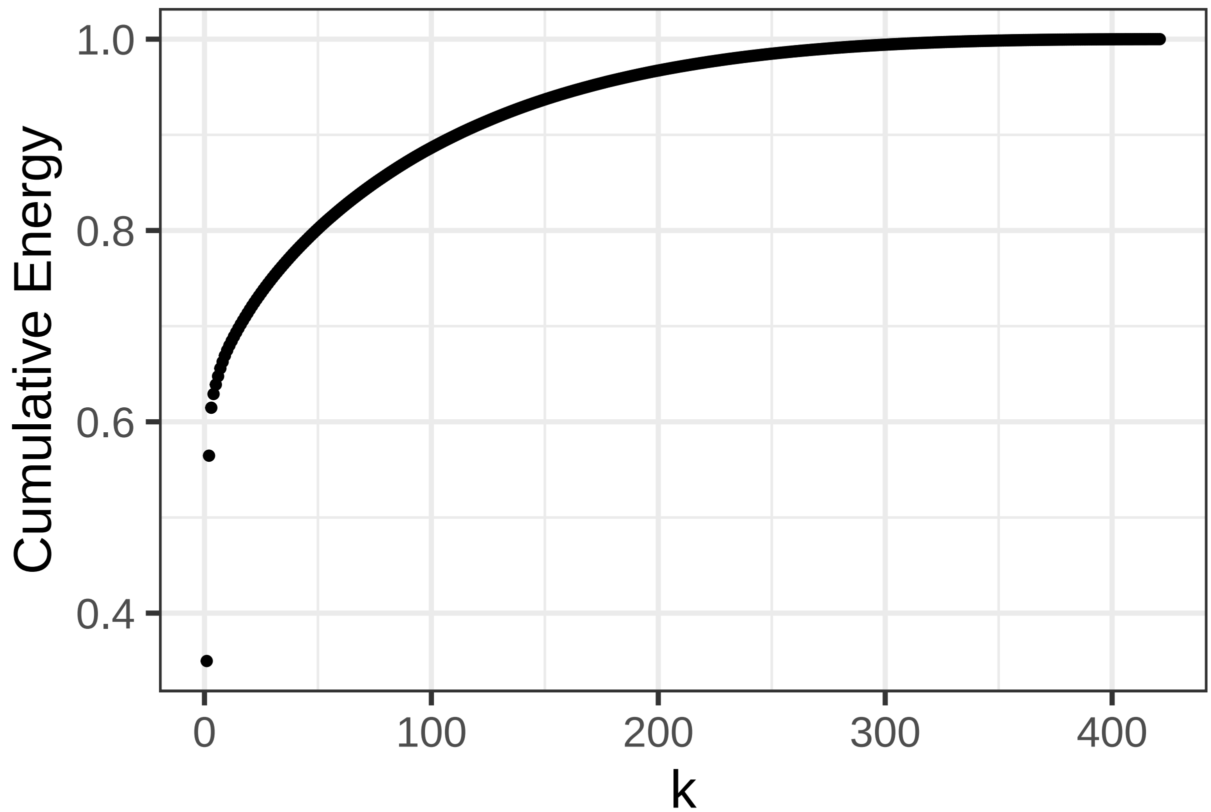
As we can see from the preliminary analyses of these graphs, there are distinct clusters of Democrats vs. Republicans, as well as votes that passed vs. votes that did not.

Before we analyze this any further, we note that these graphs represent different aspects of our SVD analyses. Remember, SVD is a decomposition of a matrix $\mathbf{X}$ into a unique combination of matrices $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$, such that $\mathbf{X} = \mathbf{UDV}^T$.In this scenario, the coordinate axes of the graphs on the left (highlighting the clusters of representatives) correspond to the left singular vectors $\mathbf{u}$ of the matrix $\mathbf{U}$. The columns of $\mathbf{U}$ in general describe abstract "concepts" which may influence the votes of House and Senate members. For instance, the first two left singular vectors of $\mathbf{U}$ represent the level of partisanship and bipartisanship that is assigned to a representative or senator, based off the SVD computation of the data.

On the other hand, the graphs on the right hand side represent all the votes that were cast for that session. These plots are associated with the $\mathbf{V}$ matrix in our SVD composition. This matrix is responsible for predicting the different roll call votes to each of our defined concepts. In other words, it is assigning a level of bipartisanship and partisanship to each vote. Of course, votes with high bipartisanship are more likely to pass since bipartisanship measures how much of the House or Senate the vote captured.

There is also the third matrix of SVD composition, $\mathbf{D}$, which we have not discussed. This matrix is a diagonal matrix which measures the "strength" of each concept that we discussed earlier. The reason we take the first two left singular vectors, is because they represent the two most impactful concepts in our analysis. In fact, we can plot the strength of each individual concept and find that the two most powerful concepts capture a majority of the variation in our data. The graphs below highlight this for our 90th House data, although they are generalizable for the other datasets too:

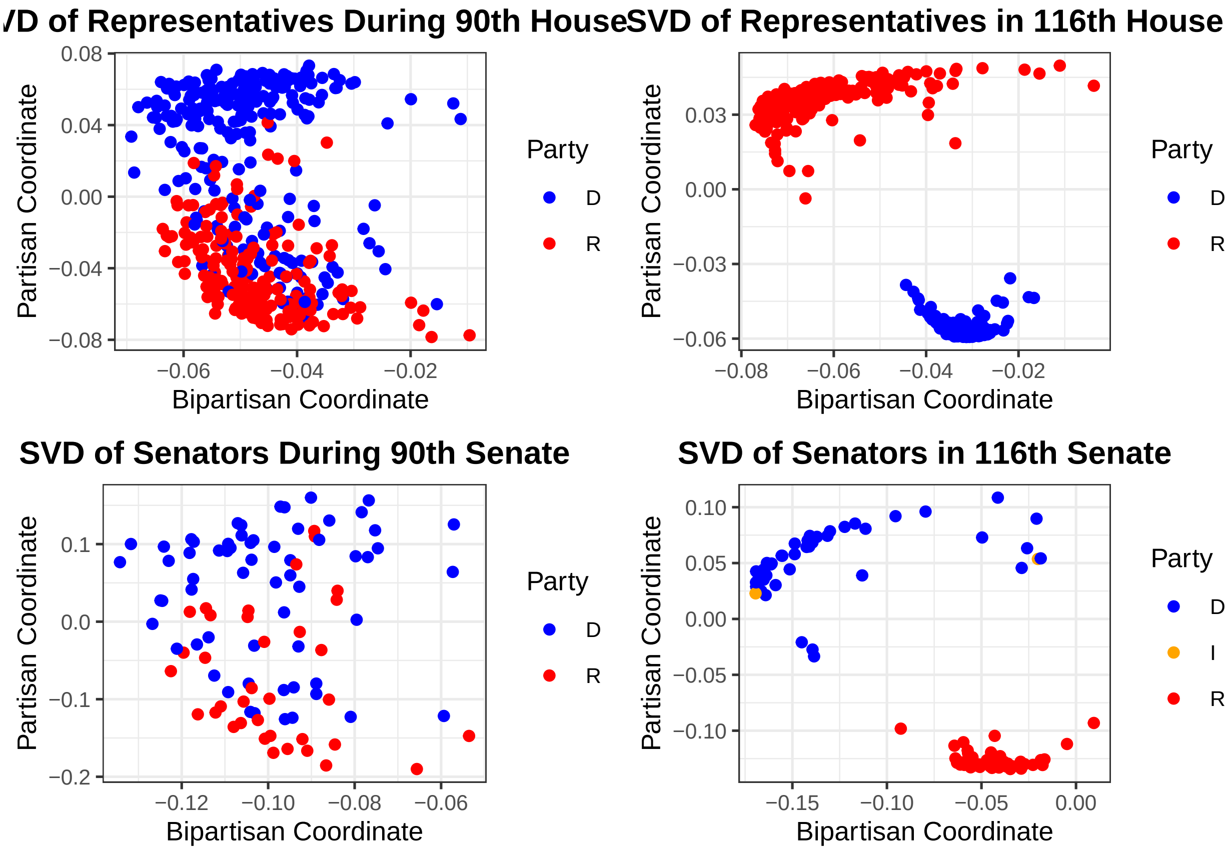

Singular Values
60000
40000
20000
0
0
100
200
300
400
k




As we can see from these graphs, just the first two singular values retain a significant portion (around 60%) of the total energy. This is also why our graphs are fairly accurate – much of the variation is due to the levels of bipartisanship and partisanship a vote has.

### Analysis of the SVD Data

Now that we have loaded the SVD visualizations, we can proceed with our analysis. As was hinted earlier, there are noticeable clusters for each data visualization. In particular, we can see that the level of partisanship has grown over the 50 years between the 90th and 116th Congresses:




This definitely matches the perception of many Americans who have felt that the country has gotten more divided. One other thing to note between these two time periods is that in addition to parties atomizing, the House of Representatives seems to have switched positions between the Democrats and Republicans compared to their positions back in the 90th House. This could be a reflection of party shifts that have occurred within the past 50 years.

We have analyzed partisanship in Congress on a grand scale, but we now focus on partisanship across several parameters. First, we will take a look at how different states differ in their levels of partisanship. For this next analysis, I decided to use the 4 most populous states – California, Texas, New York, Florida – to analyze. This was mostly because they had sufficient representatives in Congress to analyze any particular trend.

As we can see from these graphs, the data points are sparse, so it is difficult to be certain about any conclusion that we draw from this. However, these graphs seem to show that Democrats in California are highly partisan, while the same could be said of Republicans in Texas. Florida has wider variation among party colleagues, and New York seems to have the lowest level of partisanship. However, with the lack of available data, we cannot draw any distinct conlusions from this graph.

We can look further into this analysis by looking at the partisanship of the same four states, except with the

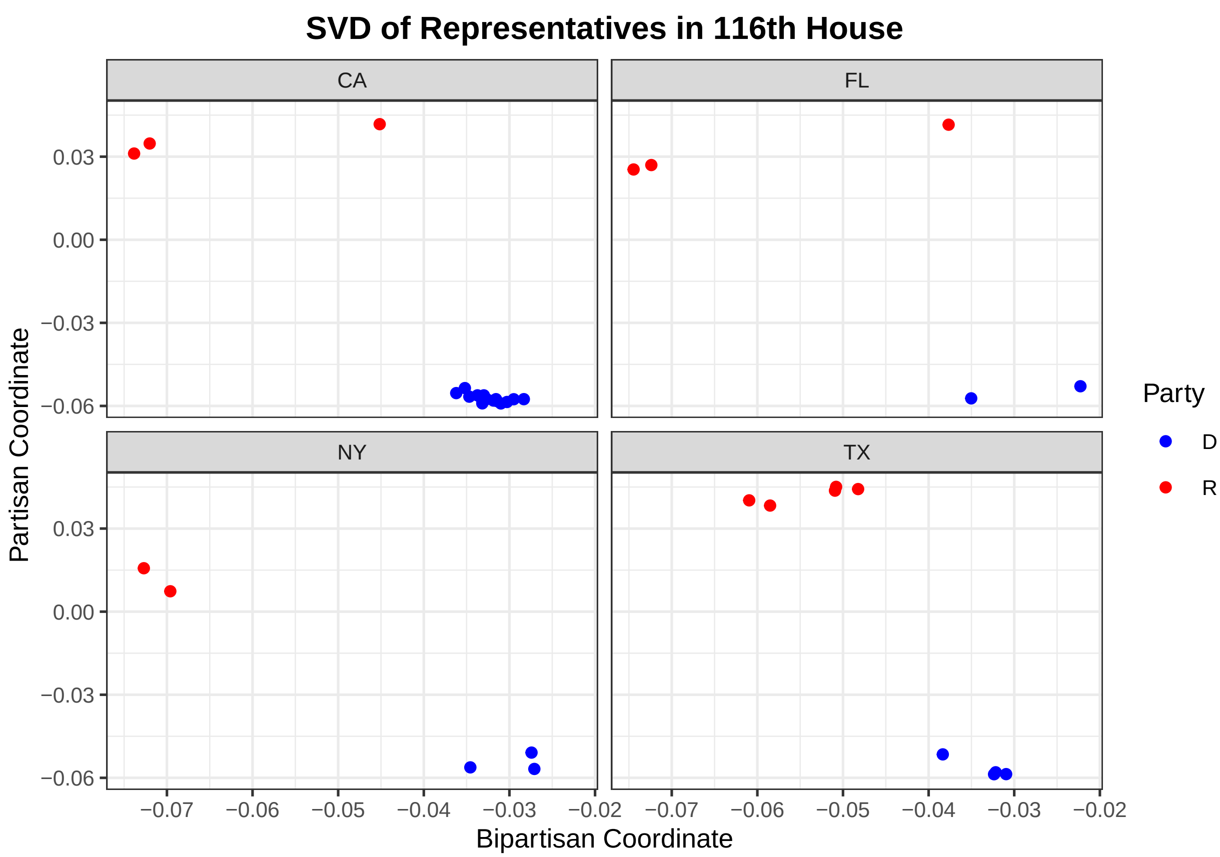
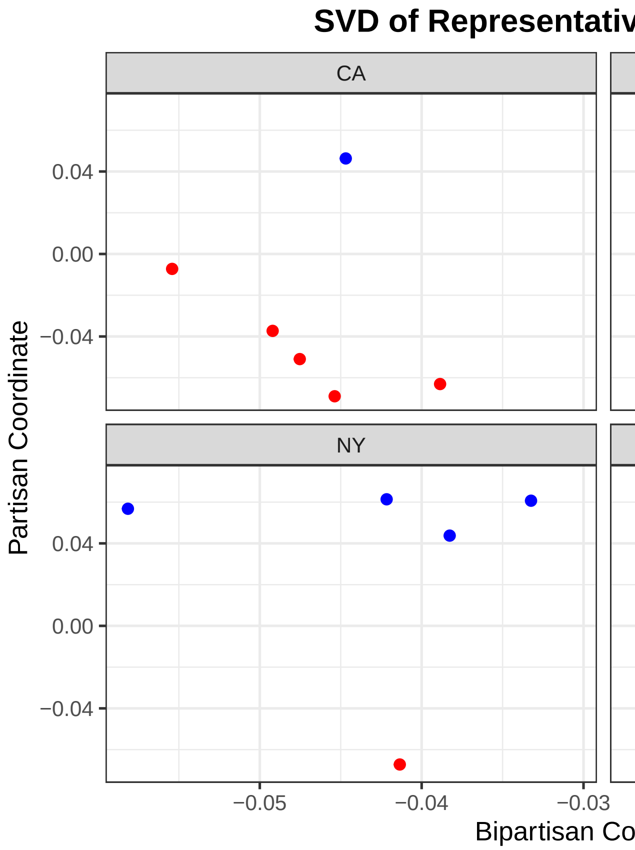



Figure 5: Figure 7: Partisanship of the 116th House for Representatives from CA, TX, NY, FL




90th House instead of the 116th House, to see if there are any changes.

The takeaways from the above graph largely mirror the takeaways from our overall Congressional analysis. There seems to have been an increase in partisanship for each state over the past 50 years, as well as some party re-alignment. In particular, states such as Texas and California used to have a majority of their representatives be Democrat and Republican respectively, while the opposite is the case today.

While we have analyzed the partisanship for these individual states, it is also worth exploring whether these states have any "extreme" partisans (i.e. partisans which deviate from the rest of the party). To analyze this, I took the overall SVD composition of the 116th House and overlayed the data points which correspond to representatives for the given state.

From this data, it doesn't appear that any of the states have extreme partisans. Texas and Florida seem to have Republicans who are more "extreme" compared to the rest of their party, although ths significance of that is lessened by the fact that the Republicans vary more widely across both the partisan and biparisan axes than the Democrats. California and New York have more Democrats than Republicans, as expected.

There was one other question that I was curious about when analyzing this dataset – Does age play a significant role in determining the partisanship of a representative? If so, how? I wanted to explore this question because I was curious if any particular generational divide existed in Congress. To help answer this question, I decided to re-classify the "born" category to more arbitrary generational lines. In particular, I took the median birth year and divided the representatives and senators with half being older and half being younger. My rationale was that if there is a significant divide among older and younger partisans, then it would show up in these graphs:

Looking at these graphs, it doesn't appear that age plays an effect on an individual's partisanship and bipartisanship. Younger politicians are pretty well distributed along with older politicians. There is no significant clustering along age boundaries, so we do not have sufficient evidence to think there is an actual difference.

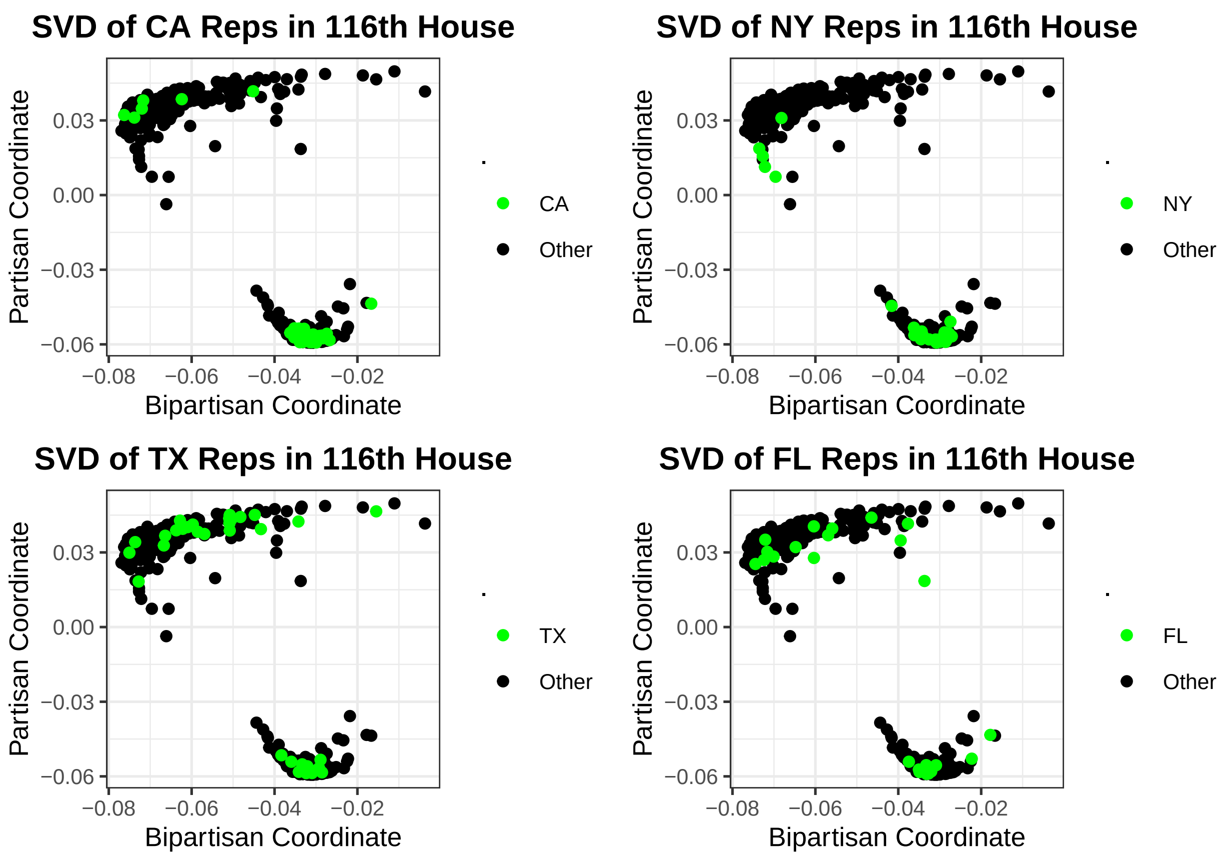



Figure 6: Figure 9: Partisanship of Representatives from CA, TX, NY, FL compared to overall House Partisanship

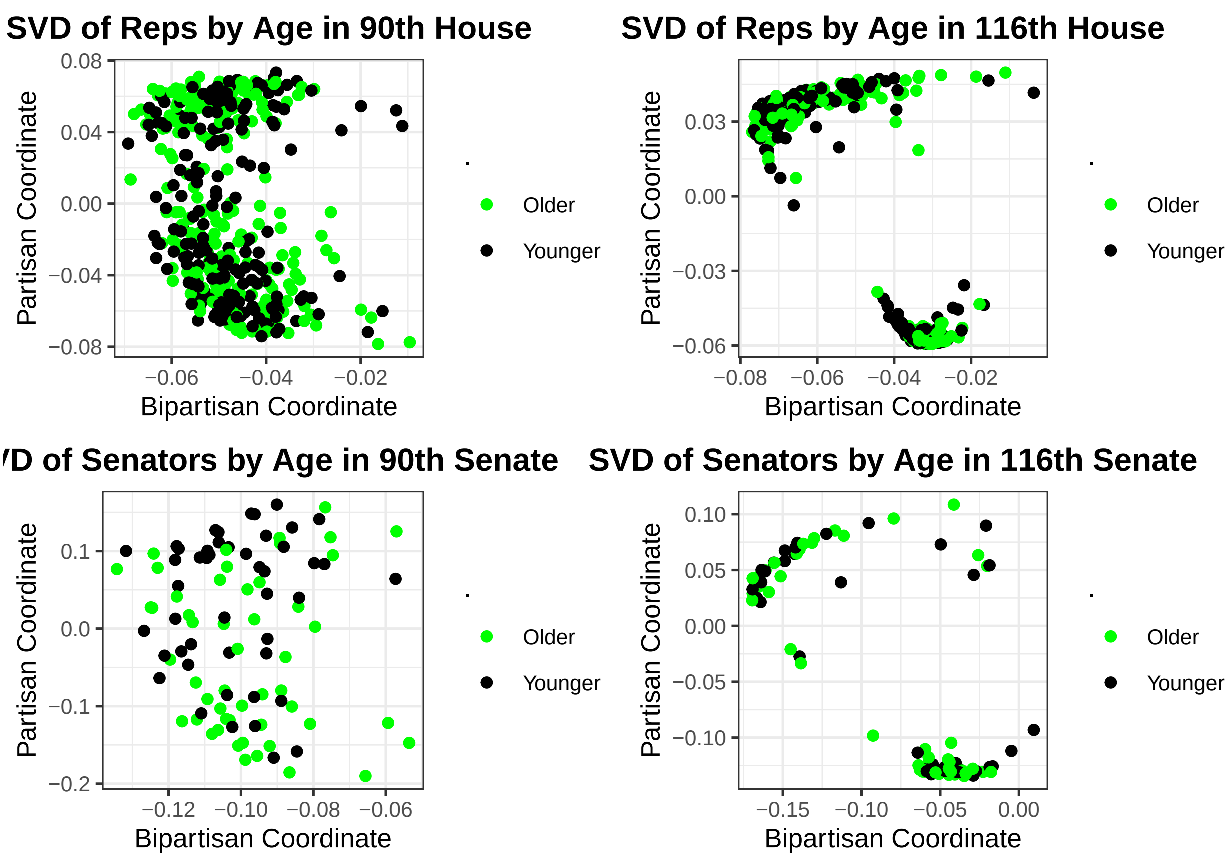



Figure 7: Figure 10: Partisanship of Representatives by Age

## Conclusion

Most of my analyses did not seem to note anything particularly significant. However, despite the lack of conclusive evidence for age affecting partisanship, future analyses can focus on subdividing the partisans along more age groups, as a division based on half the population may still have too much variance. Similarly, we can continue the state-based analysis by analyzing more states, and perhaps in a broader sense, regions. Ultimately, this analysis has made me even more curious to about the data. As someone who likes to identify trends, this project illuminated areas that have piqued my interest further.

## Appendix

```r
# Loads the necessary libraries to perform the data visualizations and manipulations
library(tidyverse)
library(ggrepel)
library(gridExtra)

# Reads the .csv files containing the raw congressional data and initializes dataframes
house_90 <- read_csv("https://raw.githubusercontent.com/bryandmartin/STAT302/master/docs/Projects/proje
senate_90 <- read_csv("https://raw.githubusercontent.com/bryandmartin/STAT302/master/docs/Projects/proj
house_116 <- read_csv("https://raw.githubusercontent.com/bryandmartin/STAT302/master/docs/Projects/proj
senate_116 <- read_csv("https://raw.githubusercontent.com/bryandmartin/STAT302/master/docs/Projects/pro

# 90th House SVD Analysis ------------------------------------------

# Drops the NA values from the dataframe to avoid biasing our data points
house_90_no_missing <- drop_na(house_90)

# Creates an SVD matrix from the data
house_90_svd <- svd(house_90_no_missing[, -1:-4])

# Creates a dataframe to be used in our plot
house_90_rep <- data.frame("x" = house_90_svd$u[, 1],
                           "y" = house_90_svd$u[, 2],
                           "Party" = house_90_no_missing$party_code,
                           "state" = house_90_no_missing$state_abbrev)

# creates a dotplot showing the clustering from the SVD analysis on the Representatives in the 90th Hou
ggplot(house_90_rep, aes(x = x, y = y, color = Party)) +
  geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  labs(title = "SVD of Representatives During 90th House",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Replaces all NA's to 0s in the dataframe so that we can perform an accurate SVD analysis on votes
house_90_zeros <- house_90 %>%
  replace(., is.na(.), 0)

# Create an SVD Matrix
house_90_zero_svd <- svd(house_90_zeros[, -1:-4])
```

```r
# Determines whether or not a vote was passed by checking if the sum over the column is positive or not
pass_fail <- colSums(house_90_zeros[, -1:-4]) > 0

# Creates a dataframe to be used for plotting the SVD of votes
house_90_votes <- data.frame("x" = house_90_zero_svd$v[, 1],
                             "y" = house_90_zero_svd$v[, 2],
                             "success" = pass_fail)

# Creates a dot plot showing the SVD of votes cast in the 90th House
ggplot(house_90_votes,
       aes(x = x, y = y, color = success)) +
  geom_point() +
  scale_color_discrete(name = "Did It Pass?",
                       labels = c("No", "Yes")) +
  labs(title = "SVD of Votes Cast in 90th House",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# A Vector of the squared singular values
sing_vals2_house_90 <- house_90_svd$d^2

# A dataframe containing the singular values and cumulative energy vectors
energy_house_90_df <- data.frame("sing_vals" = sing_vals2_house_90,
                                 "energy" = cumsum(sing_vals2_house_90/sum(sing_vals2_house_90)))

# Plots the strength of each singular value
ggplot(energy_house_90_df, aes(x = 1:nrow(energy_house_90_df), y = sing_vals)) +
  geom_point() +
  theme_bw(base_size = 20) +
  labs(y = "Singular Values", x = "k")

# Plots the cumulative energy retained with each new singular value
ggplot(energy_house_90_df, aes(x = 1:nrow(energy_house_90_df), y = energy)) +
  geom_point() +
  theme_bw(base_size = 20) +
  labs(y = "Cumulative Energy", x = "k")

# 90th Senate SVD Analysis ------------------------------------

# Drop NA's to analyze left singular vectors
senate_90_no_missing <- drop_na(senate_90)

# Create SVD Matrix
senate_90_svd <- svd(senate_90_no_missing[, -1:-4])

# Create dataframe for plotting left singular vectors
senate_90_ppl <- data.frame("x" = senate_90_svd$u[, 1],
                            "y" = senate_90_svd$u[, 2],
                            "Party" = senate_90_no_missing$party_code)

# dotplot depicting SVD analysis of left singular vectors (bipartisanship and partisanship of senators)
```

```r
ggplot(senate_90_ppl, aes(x = x, y = y, color = Party)) +
  geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  labs(title = "SVD of Senators During 90th Senate",
       x = "Bipartisan Coordinate",
       y ="Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Converts NA's to 0's
senate_90_zeros <- senate_90 %>%
  replace(., is.na(.), 0)
# Generates SVD matrices
senate_90_zero_svd <- svd(senate_90_zeros[, -1:-4])
# Check if a vote passes
sen_90_pass_fail <- colSums(senate_90_zeros[, -1:-4]) > 0
# create dataframe
senate_90_votes <- data.frame("x" = senate_90_zero_svd$v[, 1],
                              "y" = senate_90_zero_svd$v[, 2],
                              "success" = sen_90_pass_fail)

# dotplot showing SVD analysis of votes
ggplot(senate_90_votes,
       aes(x = x, y = y, color = success)) +
  geom_point() +
  scale_color_discrete(name = "Did It Pass?",
                       labels = c("No", "Yes")) +
  labs(title = "SVD of Votes Cast in 90th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# 116th House SVD Analysis ---------------------------------------

# Drop NA's
house_116_no_missing <- drop_na(house_116)
# Create SVD matrices
house_116_svd <- svd(house_116_no_missing[, -1:-4])
# Create dataframe for SVD analysis on left singular vectors
house_116_ppl <- data.frame("x" = house_116_svd$u[, 2],
                            "y" = house_116_svd$u[, 1],
                            "Party" = house_116_no_missing$party_code,
                            "state" = house_116_no_missing$state_abbrev)

# Dotplot showing SVD analysis of Representatives
ggplot(house_116_ppl, aes(x = x, y = y, color = Party)) +
  geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  labs(title = "SVD of Representatives in 116th House",
       x = "Bipartisan Coordinate",
       y ="Partisan Coordinate") +
  theme_bw(base_size = 10) +
```

```r
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Convert NA's to zeros
house_116_zeros <- house_116 %>%
  replace(., is.na(.), 0)
# Create SVD matrices
house_116_zero_svd <- svd(house_116_zeros[, -1:-4])
# Check if a vote passes
house_116_pass_fail <- colSums(house_116_zeros[, -1:-4]) > 0
# Create dataframe analyzing right singular vectors (votes)
house_116_votes <- data.frame("x" = house_116_zero_svd$v[, 2],
                              "y" = house_116_zero_svd$v[, 1],
                              "success" = house_116_pass_fail)

# Dotplot showing SVD analysis of votes
ggplot(house_116_votes,
       aes(x = x, y = y, color = success)) +
  geom_point() +
  scale_color_discrete(name = "Did It Pass?",
                       labels = c("No", "Yes")) +
  labs(title = "SVD of Votes Cast in 116th House",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# 116th Senate SVD Analysis ------------------------------------------

# Drop NA's from dataframe
senate_116_no_missing <- drop_na(senate_116)
# Create SVD matrices
senate_116_svd <- svd(senate_116_no_missing[, -1:-4])
# Create dataframe to analyze left singular vectors (senators)
senate_116_ppl <- data.frame("x" = senate_116_svd$u[, 2],
                             "y" = senate_116_svd$u[, 1],
                             "Party" = senate_116_no_missing$party_code)

# dotplot showing SVD analysis of senators
ggplot(senate_116_ppl, aes(x = x, y = y, color = Party)) +
  geom_point() +
  scale_color_manual(values = c("blue", "orange", "red")) +
  labs(title = "SVD of Senators in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Converts NA's to zeros
senate_116_zeros <- senate_116 %>%
  replace(., is.na(.), 0)
# Creates SVD matrices
senate_116_zero_svd <- svd(senate_116_zeros[, -1:-4])
# Checks if vote passes
```

```r
senate_116_pass_fail <- colSums(senate_116_zeros[, -1:-4]) > 0
# Creates dataframe for analysis on right singular vectors (votes)
senate_116_votes <- data.frame("x" = senate_116_zero_svd$v[, 2],
                               "y" = senate_116_zero_svd$v[, 1],
                               "success" = senate_116_pass_fail)

# dotplot showing SVD analysis of votes
ggplot(senate_116_votes,
       aes(x = x, y = y, color = success)) +
  geom_point() +
  scale_color_discrete(name = "Did It Pass?",
                       labels = c("No", "Yes")) +
  labs(title = "SVD of Votes in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Partisanship of 4 States ------------------------------

# Filter the dataframes for only reps from CA, FL, TX, and NY
ca_ny_tx_fl_116 <- house_116_ppl %>%
  filter(state == c("CA", "FL", "TX", "NY"))
ca_ny_tx_fl_90 <- house_90_rep %>%
  filter(state == c("CA", "FL", "TX", "NY"))

# Dotplots showing partisanship on multiple axes with a plot for each state
ggplot(ca_ny_tx_fl_116,
       aes(x = x, y = y, color = Party)) +
  geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  facet_wrap(~ state) +
  labs(title = "SVD of Representatives in 116th House",
       x = "Bipartisan Coordinate",
       y ="Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
ggplot(ca_ny_tx_fl_90,
       aes(x = x, y = y, color = Party)) +
  geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  facet_wrap(~ state) +
  labs(title = "SVD of Representatives in 90th House",
       x = "Bipartisan Coordinate",
       y ="Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Extreme Partisans for the 4 States ---------------------------------

# Filter and sort the dataframes to only include the selected state and "Other"
ca_filter <- house_116_ppl$state %>%
  fct_other(keep = "CA") %>%
```

```r
  cbind(house_116_ppl[, -4])
ca_sort <- ca_filter[order(ca_filter$., decreasing = TRUE),]
ny_filter <- house_116_ppl$state %>%
  fct_other(keep = "NY") %>%
  cbind(house_116_ppl[, -4])
ny_sort <- ny_filter[order(ny_filter$., decreasing = TRUE),]
tx_filter <- house_116_ppl$state %>%
  fct_other(keep = "TX") %>%
  cbind(house_116_ppl[, -4])
tx_sort <- tx_filter[order(tx_filter$., decreasing = TRUE),]
fl_filter <- house_116_ppl$state %>%
  fct_other(keep = "FL") %>%
  cbind(house_116_ppl[, -4])
fl_sort <- fl_filter[order(fl_filter$., decreasing = TRUE),]

# Generate 4 different dotplots showing partisans in relation to the rest of Congress (one for each sta
ca_reps <- ggplot(ca_sort,
                  aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of CA Reps in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
ny_reps <- ggplot(ny_sort,
                  aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of NY Reps in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
tx_reps <- ggplot(tx_sort,
                  aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of TX Reps in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
fl_reps <- ggplot(fl_sort,
                  aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of FL Reps in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```r
# Plot all 4 graphs in a 2x2 fashion
grid.arrange(ca_reps, ny_reps, tx_reps, fl_reps, nrow = 2, ncol = 2)

# Age Analysis ----------------------------------

# Takes in one of the congressional dataframes (the ones that have no missing data) and calculates the m
sort_by_age_group <- function(congress_data) {
  middle_age <- median(congress_data$born)
  year <- congress_data$born
  year[year <= middle_age] <- "Older"
  age_group <- year %>%
    fct_other(keep = "Older") %>%
    fct_recode(Younger = "Other")
  return(age_group)
}

# Initialize a transformed dataframe with an age_group column that has the values "Younger" or "Older"
house_116_age_grouped <- house_116_no_missing %>%
  sort_by_age_group() %>%
  cbind(house_116_ppl)
senate_116_age_grouped <- senate_116_no_missing %>%
  sort_by_age_group() %>%
  cbind(senate_116_ppl)
house_90_age_grouped <- house_90_no_missing %>%
  sort_by_age_group() %>%
  cbind(house_90_rep)
senate_90_age_grouped <- senate_90_no_missing %>%
  sort_by_age_group() %>%
  cbind(senate_90_ppl)

# Stores dotplots highlighting the age groups of representatives
house_116_age <- ggplot(house_116_age_grouped,
                        aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of Reps by Age in 116th House",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
senate_116_age <- ggplot(senate_116_age_grouped,
                        aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of Senators by Age in 116th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
house_90_age <- ggplot(house_90_age_grouped,
                        aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
```

```r
  labs(title = "SVD of Reps by Age in 90th House",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
senate_90_age <- ggplot(senate_90_age_grouped,
                        aes(x = x, y = y, color = .)) +
  geom_point() +
  scale_color_manual(values = c("green", "black")) +
  labs(title = "SVD of Senators by Age in 90th Senate",
       x = "Bipartisan Coordinate",
       y = "Partisan Coordinate") +
  theme_bw(base_size = 10) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# Plots all four graphs in a 2x2 fashion
grid.arrange(house_90_age, house_116_age, senate_90_age,
             senate_116_age, nrow = 2, ncol = 2)
```