

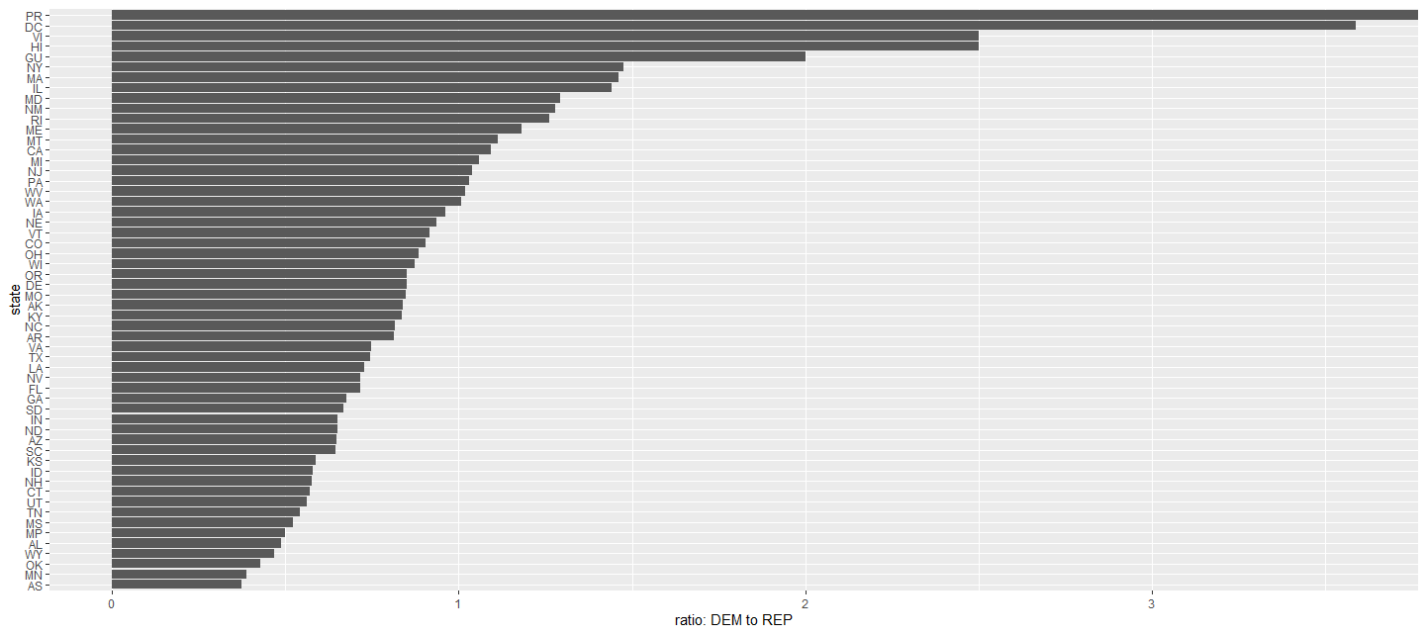
Assignment 2

Megan Rachel Santagata

Nathan Van

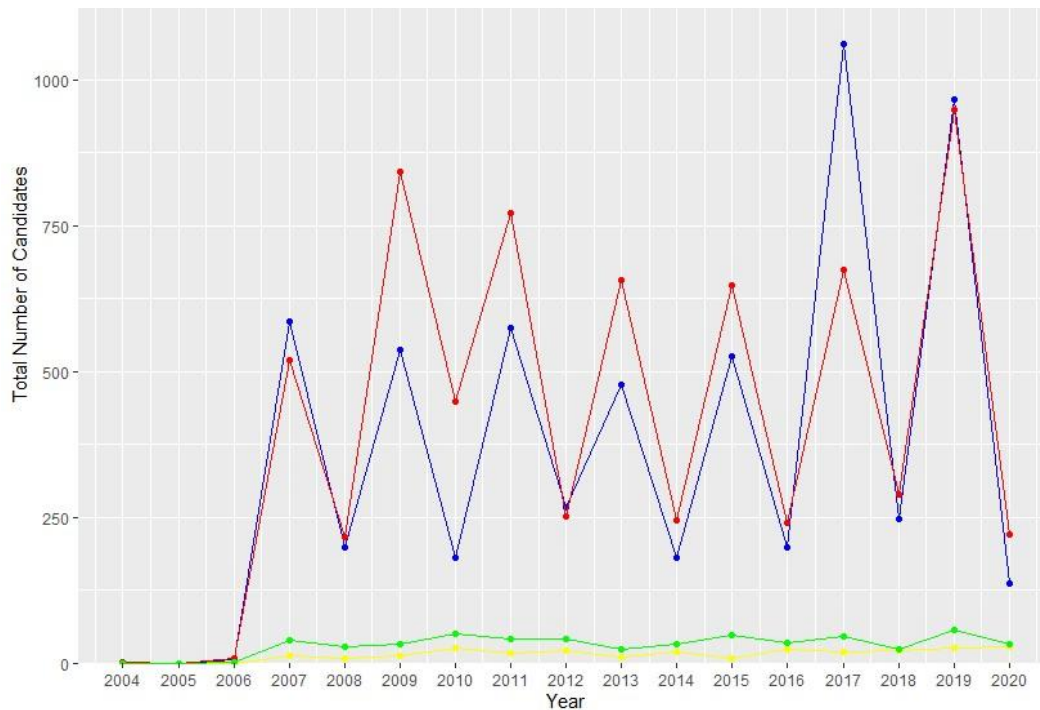
Part 1)

Visualization 1: The Democrat to Republican ratio in each of the 50 U.S states and 14 U.S territories.



Description: The histogram plot describes the ratio of the number of Democrat candidates to the number of Republican candidates in each state. Less than 1 represents republican majority while greater than 1 represents democrat majority. PA, WA, ME, NJ, MT, AS, RI, CA, HI, IL, DE, MA, NY, NM, GU, VI, DC, PR have Democrat majority while the rest have Republican majority. Megan mainly worked on this graph

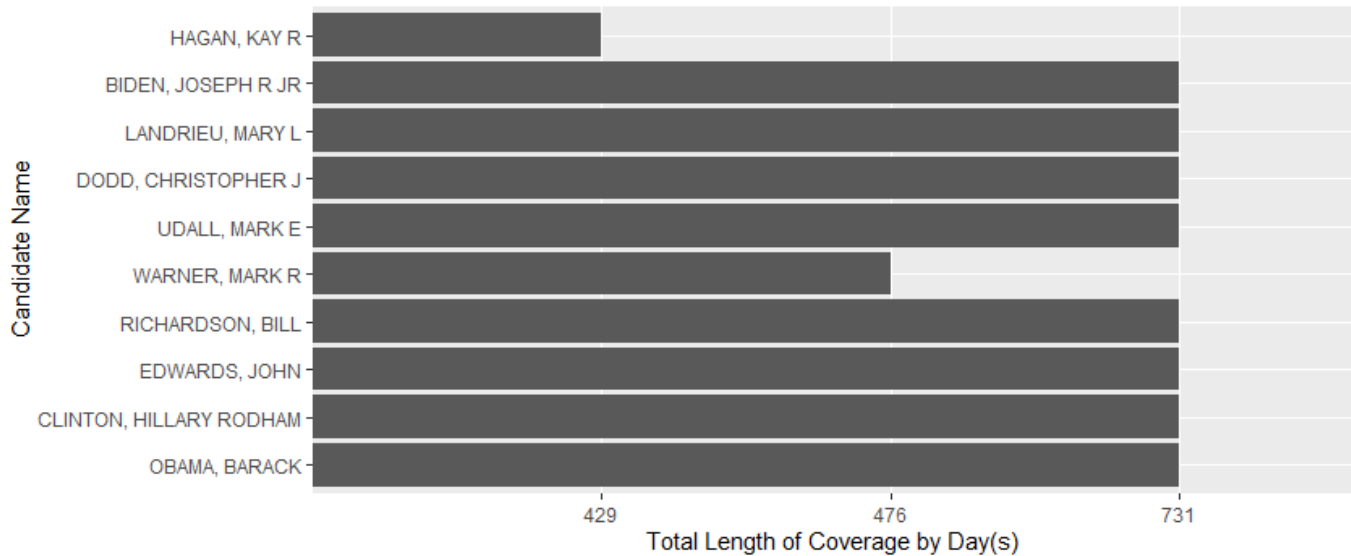
Visualization 2: The total number of candidates in each political party per year.



Description: From 2006 to 2008 (during Bush's presidency) there were more Democrats (**blue**) than Republicans (**red**), from 2008 to 2016 (during Obama's presidency) there were more Republicans (**red**) than Democrats (**blue**), from 2016 to 2018 there were more Republicans (**red**) than Democrats (**blue**), and finally, from 2018 to 2020 the number of Democrat (**blue**) candidates and Republican (**red**) candidates were almost the same. Both Libertarian (**yellow**) and Independent (**green**) numbers had always been lower than Democrats (**blue**) and Republicans (**red**). Meanwhile, between the two smallest parties, the total number of Independent

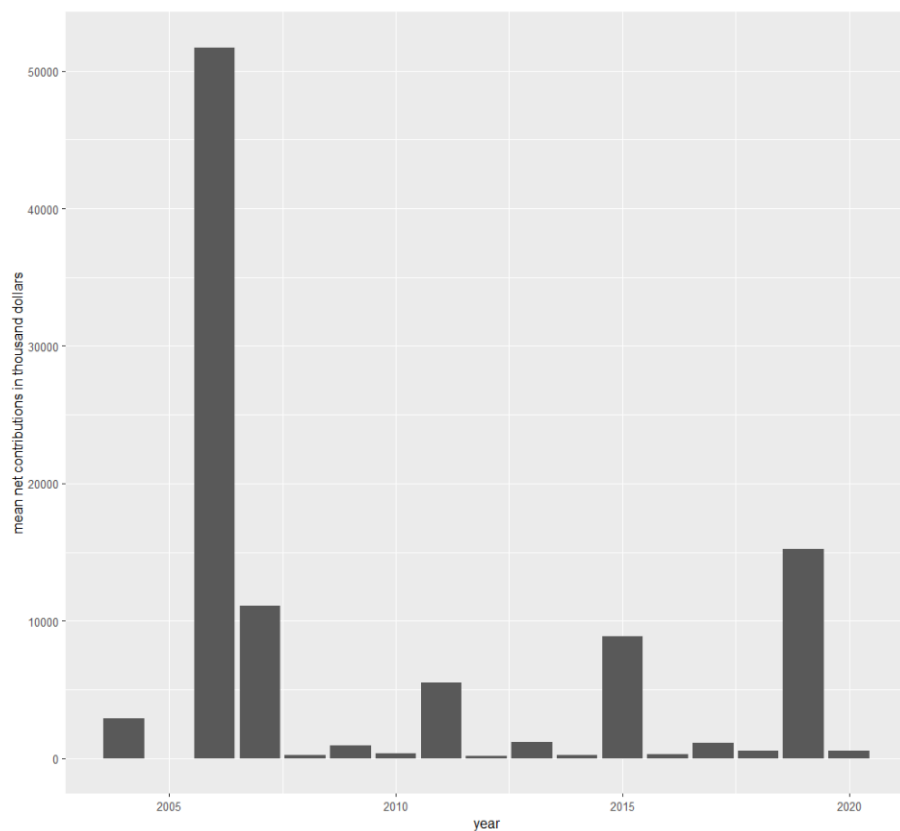
(green) candidates has been consistently higher than the total number of Libertarian (yellow) since 2006. (Nathan)

Visualization 3: The top 10 democratic candidates with the highest total contribution ranked by their total contributions and their length of coverage in day(s).



Description: This bar graph shows the top 10 Democrat candidates with the highest total contribution ranked by their total contributions. We display their length of coverage in days, we computed this number by subtracting the Coverage_Start_Date by the Coverage_End_Date. Obama being the Democrat candidate with the most total contribution and Hagan being the Democrat candidate with the least total contribution among the top 10 Democrat candidates with the highest total contribution. The plot shows no patterns correlations with being the candidate with the most total contribution and length of coverage. (Nathan)

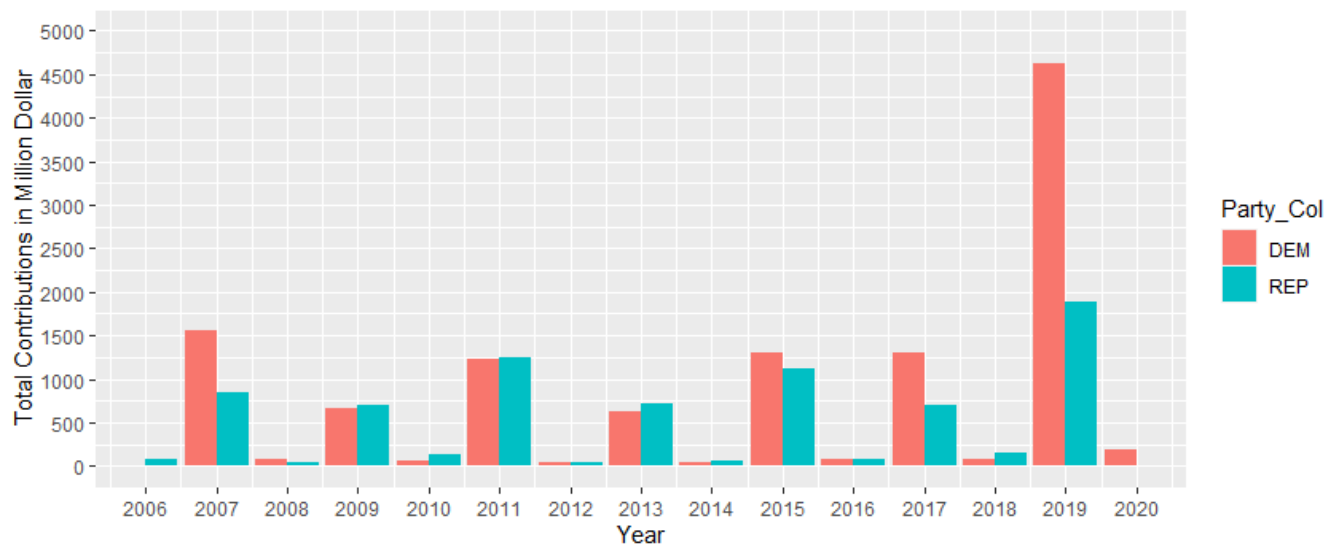
Visualization 4: Mean contribution for each year over time (2004 to 2020)



Description: This graph shows the average (mean) net contributions for all political candidates for each year. This includes the political candidates who received zero contributions. Contributions particularly spiked in 2006 and 2019, with the former being having the highest mean net contributions of any year.

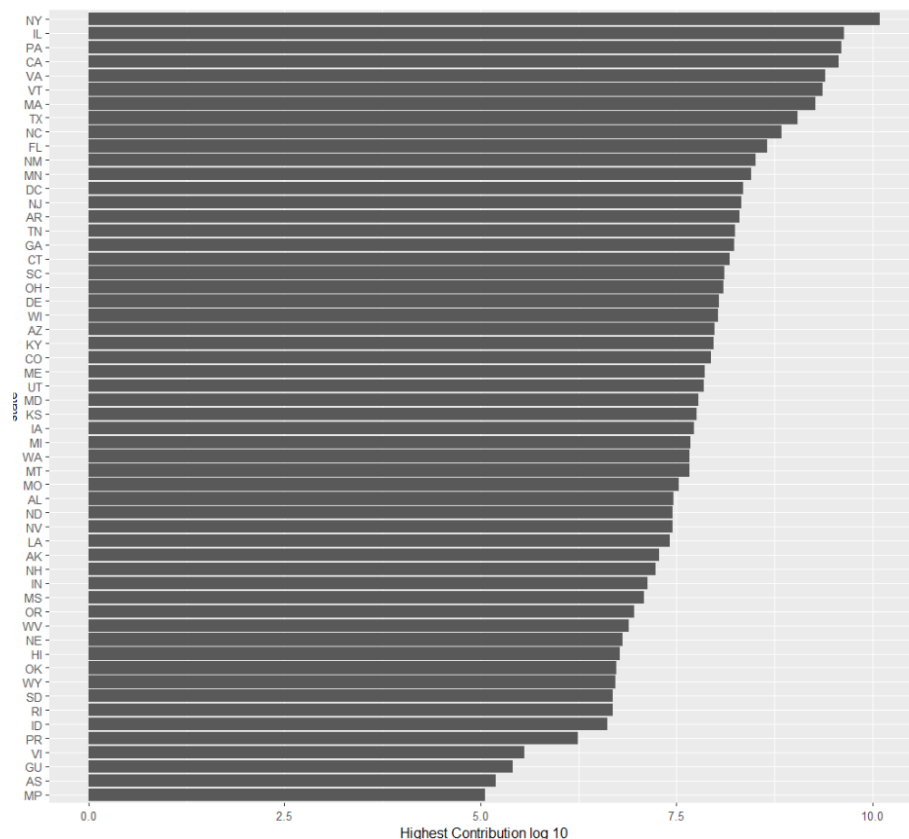
Most of the spikes come in the years before election years. 2006 is a notable outlier, with the highest mean net contributions, but not coming before an election year. Megan mainly worked on this graph.

Visualization 5: The total contribution amount for Democrat vs. Republican from 2006 to 2020.



Description: This graph displays the total contributions for democratic and republican candidates each year. We can see that during election campaign periods, there were spikes in the total contributions across both Democrat (**blue**) and Republican (**red**) parties. (Nathan)

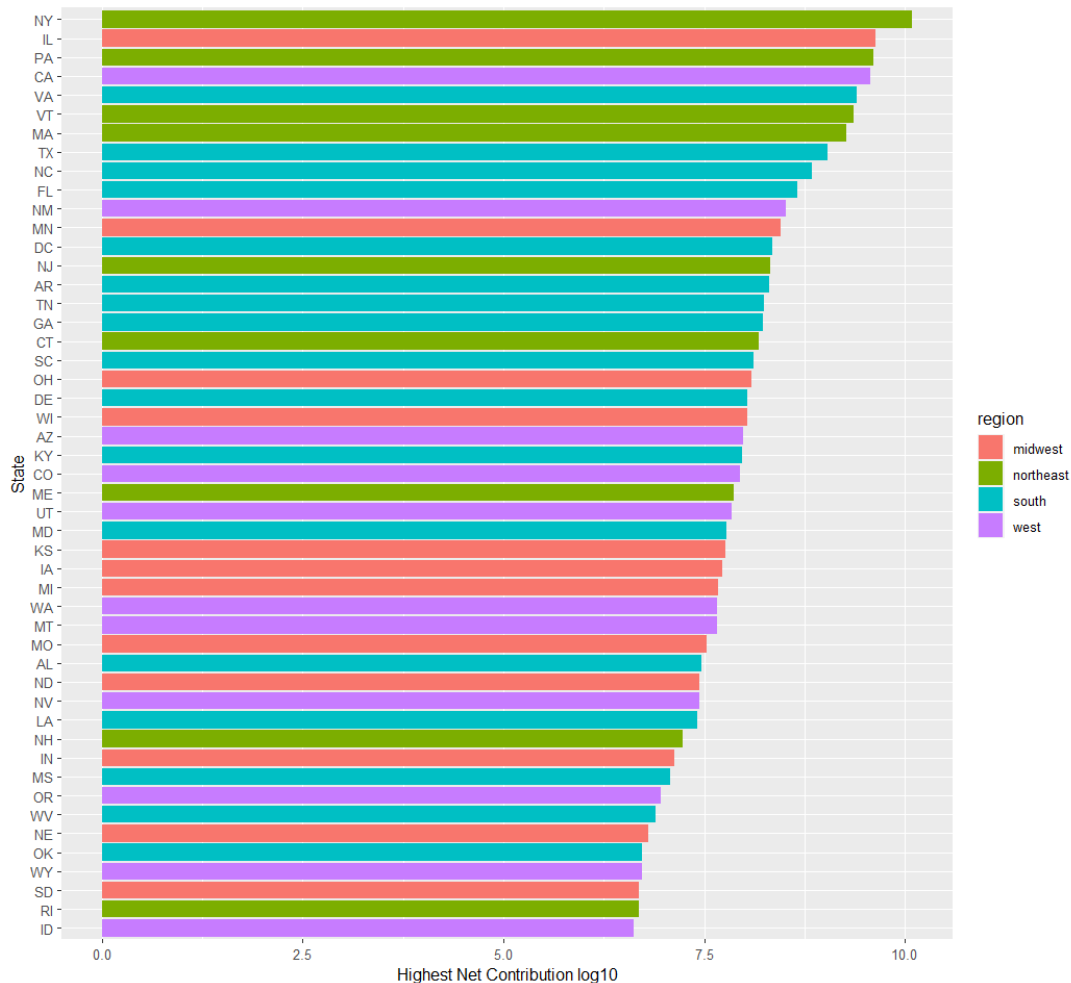
Visualization 6: Highest Net Contribution by State



Description: This graph tracks the highest contributions (max) for candidates from each state. This graph uses data from 2004 to 2020. Contributions are measures in log 10 to make sure that they were able to all be viewed on the same graph. New York, Illinois, Pennsylvania, California, and Virginia are the top five highest ranking states. I thought it was surprising that New York out performed other states by such a large margin, but I remembered that Donald Trump, the most recent president, is from New York which may explain the large lead the state has.

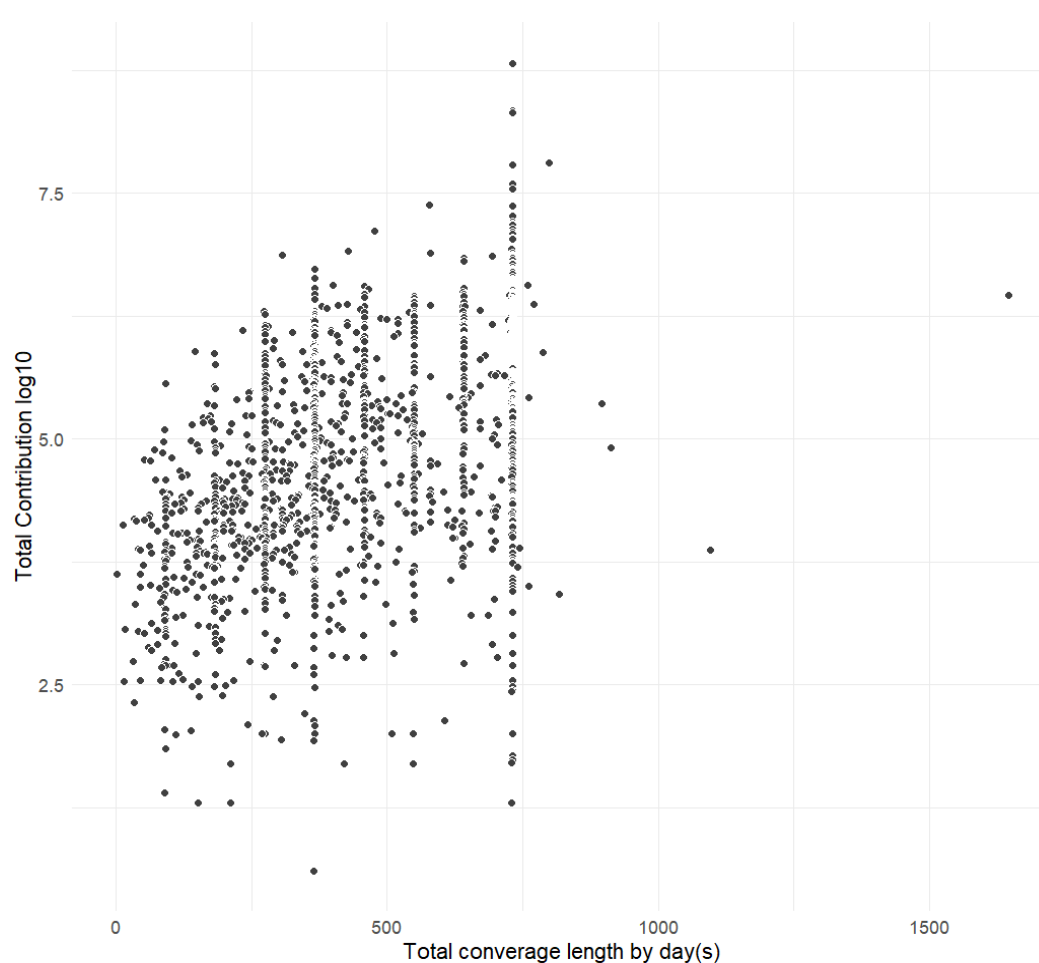
Part 2)

Hypothesis 1: (Visualization 6). The northeast region states candidates receive the highest net contributions. When looking at the uncolored chart, a couple of northeast region states are near the top, with New York coming in first with the highest net contributions by a significant margin. We wanted to see if any trends between different regions could be seen. We colored the graph to reflect which regions the states were in to better visualize the trends.



The graph shows some trends. Both the northeast and south are heavily represented near the top of the graph with the only notable exceptions from the midwest and west being Illinois and California respectively. The midwest states seem to be clumped towards the lower half of the graph. The west is well distributed.

Hypothesis 2: After plotting total contribution and coverage time with Democrats in visualization 3, we wanted to know if there would be a correlation of length of coverage and total contribution over all the candidates of both parties. We predicted that people with less contribution would have more coverage length because it might take longer to file paperwork. Our hypothesis was proven incorrect. We see that the total contribution has a small positive correlation with the total length coverage of a candidate. More contribution does mean slightly more coverage. However, the spread gets bigger as coverage increases. We can see that, since the scatterplot does not form a linear line, the candidates with the least total contributions on this plot also get equal or more coverage than those with the most total contributions. Therefore, we conclude that there's no correlations between the total contributions a candidate gets versus the length of coverage that candidate receives. And, compared to Visualization 3, the top 10 Democrat candidates with the highest total contributions did not get more coverage than other candidates.



R Code and Comments:

```
#import csv file
myData <- read.csv("C:/Users/Megan/OneDrive/school/fec_2008-2022.csv")
#clean data
myData <- myData[myData$Total_Contribution > 0, ] # Quick cleaning to avoid data issues
parties <- unique(myData$Cand_Party_Affiliation) # Create a set with the names of the political parties
myData$Coverage_Start_Date <- as.Date(myData$Coverage_Start_Date, format="%m/%d/%Y") #
Organize the date
format
myData$Coverage_Start_DateYear <- format(myData$Coverage_Start_Date, format="%Y") # Create a
new column with the year
#How to replace a variable in a column, cleaning misspellings
myData$Cand_State[myData$Cand_State == "Pa"] <- "PA"
myData$Cand_State[myData$Cand_State == "AP"] <- "AR"
myData$Cand_State[myData$Cand_State == "ZZ"] <- "MN"
state_data <- subset(myData, Cand_State != "")
state_data
```

#graph 1

```
library(ggplot2)
#import csv file
myData <- read.csv("C:/Users/Megan/OneDrive/school/fec_2008-2022.csv")
#clean data
myData <- myData[myData$Total_Contribution > 0, ] # Quick cleaning to avoid data issues
```

```

parties <- unique(myData$Cand_Party_Affiliation) # Create a set with the names of the political parties
myData$Coverage_Start_Date <- as.Date(myData$Coverage_Start_Date, format="%m/%d/%Y") #
Organize the date
format
myData$Coverage_Start_DateYear <- format(myData$Coverage_Start_Date, format="%Y") # Create a
new column with the year
#How to replace a variable in a column, cleaning misspellings
myData$Cand_State[myData$Cand_State == "Pa"] <- "PA"
myData$Cand_State[myData$Cand_State == "AP"] <- "AR"
myData$Cand_State[myData$Cand_State == "ZZ"] <- "MN"
#Taking out data that has no state value
state_data <- subset(myData, Cand_State != "")
#taking out data that has no year value
year_data <- subset(myData, Coverage_Start_DateYear != "")

#simple selection
uniqueStates <- unique(state_data$Cand_State)
uniqueParty <- unique(state_data$Cand_Party_Affiliation)
uniqueCandName <- unique(state_data$Cand_Name)
#creates an empty list
demCountVector <- c()
repCountVector <- c()

#create data frames with only data that has democrat party affiliation
dataDEM <- state_data[state_data$Cand_Party_Affiliation=="DEM",]
#create data frames with only data that has republican party affiliation
dataREP <- state_data[state_data$Cand_Party_Affiliation=="REP",]

#for loop to count the number of dems and reps in each state
for(i in 1:length(uniqueStates)) {
  # note that there will be double counting of candidates if the candidates appear twice
  demCountVector <- c(demCountVector, nrow(dataDEM[dataDEM$Cand_State==uniqueStates[i],]) )
  repCountVector <- c(repCountVector, nrow(dataREP[dataREP$Cand_State==uniqueStates[i],]) )
}
# this new data frame contains the number of DEM and REP candidates for each state
RDdata <- data.frame(state=uniqueStates, demCandNum=demCountVector,
repCandNum=repCountVector)

#This data frame replaces the number of democrats and republicans with a ratio
RDdata <- transform(RDdata, newRatios = demCandNum/repCandNum)
#making a ratio graph
#QUESTION: how to make the bars more spread out
ggplot(RDdata, aes(x=reorder(state, newRatios), y=newRatios)) + geom_bar(stat = "identity") +
  labs(x="state", y="number of candidates") +
  coord_flip()

```

#graph 2

```
#Visualization 2
```

```
library(tidyverse)
```

```
myData <- read.csv('C:/Users/Thanh Van/Documents/R/R/assignment 2/fec_2008-2022.csv')
```

```
#This code formats the date into MM/DD/YYYY format and creates a new column named  
Coverage_Start_DateYear to store the Year
```

```
myData$Coverage_Start_Date <- as.Date(myData$Coverage_Start_Date, format="%m/%d/%Y")  
myData$Coverage_Start_DateYear <- format(myData$Coverage_Start_Date, format="%Y")
```

```
#This code creates a table for Total Democrat Candidates per Year (2004-2020)  
allDem <- data.frame(matrix(ncol=2, nrow=0))
```

```
#This code creates a table for Total Republican Candidates per Year (2004-2020)  
allRep <- data.frame(matrix(ncol=2, nrow=0))
```

```
#This code creates a table for Total Liberal Candidates per Year (2004-2020)  
allLib <- data.frame(matrix(ncol=2, nrow=0))
```

```
#This code creates a table for Total Independent Candidates per Year (2004-2020)  
allInd <- data.frame(matrix(ncol=2, nrow=0))
```

```
#This code appends data into the four tables above for visualization later
```

```
for (x in 2004:2020) {  
  allDem[nrow(allDem) + 1 , ] <- rbind(c(x, nrow(subset(myData, Cand_Party_Affiliation=='DEM'  
    & Coverage_Start_DateYear == x))))  
  
  allRep[nrow(allRep) + 1 , ] <- rbind(c(x, nrow(subset(myData, Cand_Party_Affiliation=='REP'  
    & Coverage_Start_DateYear == x))))  
  
  allLib[nrow(allLib) + 1 , ] <- rbind(c(x, nrow(subset(myData, Cand_Party_Affiliation=='LIB'  
    & Coverage_Start_DateYear == x))))  
  
  allInd[nrow(allInd) + 1 , ] <- rbind(c(x, nrow(subset(myData, Cand_Party_Affiliation=='IND'  
    & Coverage_Start_DateYear == x))))  
}
```

```
ggplot(allDem, aes(x=X1, y=X2)) + geom_line(color='blue') +  
  geom_point(color='blue') +  
  geom_line(data=allRep, color='red') +  
  geom_point(data=allRep, color='red') +  
  geom_line(data=allLib, color='yellow') +  
  geom_point(data=allLib, color='yellow') +  
  geom_line(data=allInd, color='green') +  
  geom_point(data=allInd, color='green') +  
  scale_x_continuous(name = 'Year',  
    limits = c(2004, 2020),  
    breaks = 1*(2004:2020)) +  
  scale_y_continuous(expand = c(0,0),
```

```

name = 'Total Number of Candidates',
limits = c(0, 1100),
breaks = 250*(0:1100))

```

#graph 3

```

#Visualization 3
library(tidyverse)

myData <- read.csv('C:/Users/Thanh Van/Documents/R/R/assignment 2/fec_2008-2022.csv')

#This code formats the date into MM/DD/YYYY format and creates a new column named
Coverage_Start_DateYear to store the Year
myData$Coverage_Start_Date <- as.Date(myData$Coverage_Start_Date, format="%m/%d/%Y")
myData$Coverage_End_Date <- as.Date(myData$Coverage_End_Date, format="%m/%d/%Y")

#This code creates a table for all democrat names and their corresponding coverage length in days
demCoverageLength <- data.frame(matrix(ncol=3, nrow=0))

#Create a table of all Democrat candidates with a Coverage_Start_Date not na
allDem <- subset(myData, Cand_Party_Affiliation=='DEM' & !is.na(Coverage_Start_Date) &
!is.na(Coverage_End_Date))
#number of rows in the table above to run a for-loop over it
numDem <- nrow(allDem)

#create a table
for (x in 1:numDem){
  total <- length(seq(allDem$Coverage_Start_Date[x], allDem$Coverage_End_Date[x], 'days'))
  name <- allDem$Cand_Name[x]
  contribution <- allDem$Total_Contribution[x]
  demCoverageLength[nrow(demCoverageLength) + 1, ] <- rbind(c(name, total, floor(contribution)))
}

#this code will sort and print the top 10 candidates with the longest (most days) coverage length
demCoverageLength <- demCoverageLength[order(round(as.numeric(demCoverageLength$X3)),
decreasing = TRUE, na.last = TRUE), ]

topTen <- demCoverageLength[1:10, ]
topTen$X3 <- as.numeric(as.character(topTen$X3))
topTen[order(topTen$X3), ]

ggplot(topTen, aes(x=reorder(X1, -X3), y = X2)) +
  geom_bar(stat='identity') +
  labs(x='Candidate Name', y= 'Total Length of Coverage by Day(s) ') +
  coord_flip()

```

#graph 4

```

#graph 4

```



```
#separate the data by each year
```

```
data2004 <- year_data[year_data$Coverage_Start_DateYear==2004,]  
data2005 <- year_data[year_data$Coverage_Start_DateYear==2005,]  
data2006 <- year_data[year_data$Coverage_Start_DateYear==2006,]  
data2007 <- year_data[year_data$Coverage_Start_DateYear==2007,]  
data2008 <- year_data[year_data$Coverage_Start_DateYear==2008,]  
data2009 <- year_data[year_data$Coverage_Start_DateYear==2009,]  
data2010 <- year_data[year_data$Coverage_Start_DateYear==2010,]  
data2011 <- year_data[year_data$Coverage_Start_DateYear==2011,]  
data2012 <- year_data[year_data$Coverage_Start_DateYear==2012,]  
data2013 <- year_data[year_data$Coverage_Start_DateYear==2013,]  
data2014 <- year_data[year_data$Coverage_Start_DateYear==2014,]  
data2015 <- year_data[year_data$Coverage_Start_DateYear==2015,]  
data2016 <- year_data[year_data$Coverage_Start_DateYear==2016,]  
data2017 <- year_data[year_data$Coverage_Start_DateYear==2017,]  
data2018 <- year_data[year_data$Coverage_Start_DateYear==2018,]  
data2019 <- year_data[year_data$Coverage_Start_DateYear==2019,]  
data2020 <- year_data[year_data$Coverage_Start_DateYear==2020,]
```

```
#get the mean data of each year
```

```
mean2004 <-mean(data2004$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2005 <-mean(data2005$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2006 <-mean(data2006$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2007 <-mean(data2007$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2008 <-mean(data2008$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2009 <-mean(data2009$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2010 <-mean(data2010$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2011 <-mean(data2011$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2012 <-mean(data2012$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2013 <-mean(data2013$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2014 <-mean(data2014$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2015 <-mean(data2015$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2016 <-mean(data2016$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2017 <-mean(data2017$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2018 <-mean(data2018$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2019 <-mean(data2019$Net_Contribution, trim = 0, na.rm = FALSE)  
mean2020 <-mean(data2020$Net_Contribution, trim = 0, na.rm = FALSE)
```

```
#QUESTION: why NaN, are there 0 contributions in 2005?
```

```
#create a data frame that lists the years and the means
```

```
year_mean <- data.frame(year=c(2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014,  
2015, 2016, 2017, 2018, 2019, 2020),  
                        my_mean=c(mean2004, mean2005, mean2006, mean2007, mean2008, mean2009,  
mean2010, mean2011, mean2012, mean2013,  
mean2014, mean2015, mean2016, mean2017, mean2018, mean2019, mean2020))
```

```
#create a graph
ggplot(year_mean, aes(x=year, y=my_mean/1000)) + geom_bar(stat = "identity") +
  labs(x="year", y="mean net contributions in thousand dollars")
```

#graph 5

```
#Total contribution by each party over time (2004 to 2020)??
#Visualization 5
library(tidyverse)

myData <- read.csv('C:/Users/Thanh Van/Documents/R/R/assignment 2/fec_2008-2022.csv')

#This code formats the date into MM/DD/YYYY format and creates a new column named
Coverage_Start_DateYear to store the Year
myData$Coverage_Start_Date <- as.Date(myData$Coverage_Start_Date, format="%m/%d/%Y")
myData$Coverage_Start_DateYear <- format(myData$Coverage_Start_Date, format="%Y")

#This code creates a table for Total Democrat Candidates per Year (2004-2020)
totalContributionDem<- data.frame(matrix(ncol=2, nrow=0))

#This code creates a table for Total Republican Candidates per Year (2004-2020)
totalContributionRep <- data.frame(matrix(ncol=2, nrow=0))

#This code creates a table for Total Liberal Candidates per Year (2004-2020)
totalContributionLib <- data.frame(matrix(ncol=2, nrow=0))

#This code creates a table for Total Independent Candidates per Year (2004-2020)
totalContributionInd <- data.frame(matrix(ncol=2, nrow=0))

#This code appends data into the four tables above for visualization later
for (x in 2004:2020) {
  totalContributionDem[nrow(totalContributionDem) + 1 , ] <- rbind(c(x, sum(subset(myData,
    Total_Contribution > 0 & Cand_Party_Affiliation == 'DEM' & Coverage_Start_DateYear ==
    x)$Total_Contribution)))

  totalContributionRep [nrow(totalContributionRep ) + 1 , ] <- rbind(c(x, sum(subset(myData,
    Total_Contribution > 0 & Cand_Party_Affiliation == 'REP' & Coverage_Start_DateYear ==
    x)$Total_Contribution)))

  totalContributionLib [nrow(totalContributionLib ) + 1 , ] <- rbind(c(x, sum(subset(myData,
    Total_Contribution > 0 & Cand_Party_Affiliation == 'LIB' & Coverage_Start_DateYear ==
    x)$Total_Contribution)))

  totalContributionInd [nrow(totalContributionInd ) + 1 , ] <- rbind(c(x, sum(subset(myData,
    Total_Contribution > 0 & Cand_Party_Affiliation == 'IND' & Coverage_Start_DateYear ==
    x)$Total_Contribution)))
}
```

```
#how to add columns
```

```
dataset$myNewColumn <- rep(Na, times = 10)
```

```
#how to add two data sets together rbind means row bind (what you want), cbind means column bind
```

```
dataset_new <- rbind(dataset, dataset2)
```

```
totalContributionDem$Party_Col <- rep("DEM", times = nrow(totalContributionDem))
```

```
totalContributionRep$Party_Col <- rep("REP", times = nrow(totalContributionRep))
```

```
totalContributionBoth <- rbind(totalContributionRep, totalContributionDem)
```

```
#replace NA by 1
```

```
#totalContributionBoth$X2[is.na(totalContributionBoth$X2)] <- 1
```

```
totalContributionBoth$X2[totalContributionBoth$X2 == 0] <- 1
```

```
#Make side by side bar graphs.
```

```
#Make side by side bar graphs.
```

```
ggplot(totalContributionBoth, aes(x=X1, y=X2/1000000, fill=Party_Col)) +
```

```
  geom_bar(position="dodge", stat = "identity") +
```

```
  labs(x = "Year", y = "Total Contributions in Million Dollar") +
```

```
  scale_x_continuous(limits = c(2006, 2020),
```

```
                    breaks = 1*(2006:2020)) +
```

```
  scale_y_continuous(limits = c(0, 5000),
```

```
                    breaks = 500*(0:5000))
```

#graph 6

```
#graph 6
```

```
#simple selection
```

```
uniqueStates <- unique(state_data$Cand_State)
```

```
uniqueCandName <- unique(state_data$Cand_Name)
```

```
#get the max of net contribution of each state
```

```
maxContData <- aggregate(Net_Contribution~Cand_State, state_data, max)
```

```
#create a bar graph
```

```
ggplot(maxContData, aes(x=reorder(Cand_State, Net_Contribution), y=log10(Net_Contribution))) +
```

```
  geom_bar(stat = "identity") +
```

```
  labs(x="state", y="Highest Contribution log 10") +
```

```
  coord_flip()
```

Hypothesis 1

```
#adding colors to the bars
```

```
ggplot(maxContData, aes(x=reorder(Cand_State, Net_Contribution), y=log10(Net_Contribution),  
fill=region)) +
```

```
  geom_bar(position="dodge", stat = "identity") +
```

```
  labs(x = "State", y = "Highest Net Contribution log10") +
```

```
  coord_flip()
```

Hypothesis 2

#Visualization 3

library(tidyverse)

#This code formats the date into MM/DD/YYYY format and creates a new column named Coverage_Start_DateYear to store the Year

```
myData$Coverage_Start_Date <- as.Date(myData$Coverage_Start_Date, format="%m/%d/%Y")
```

```
myData$Coverage_End_Date <- as.Date(myData$Coverage_End_Date, format="%m/%d/%Y")
```

#This code creates a table for all democrat names and their corresponding coverage length in days

```
CoverageLength <- data.frame(matrix(ncol=3, nrow=0))
```

#Create a table of all Democrat candidates with a Coverage_Start_Date not na

```
all <- subset(myData, !is.na(Coverage_Start_Date) & !is.na(Coverage_End_Date))
```

#number of rows in the table above to run a for-loop over it

```
my_num <- nrow(all)
```

#create a table

```
for (x in 1:my_num){  
  total <- length(seq(all$Coverage_Start_Date[x], all$Coverage_End_Date[x], 'days'))  
  name <- all$Cand_Name[x]  
  contribution <- all$Total_Contribution[x]  
  CoverageLength[nrow(CoverageLength) + 1, ] <- rbind(c(name, total, contribution))  
}
```

#this code will sort and print the top 10 candidates with the longest (most days) coverage length

```
CoverageLength <- CoverageLength[order(round(as.numeric(CoverageLength$X3)), decreasing =  
TRUE, na.last = TRUE), ]
```

CoverageLength

```
ggplot(demCoverageLength[1:10, ], aes(x=X1, y = X2)) +  
  geom_bar(stat='identity') +  
  labs(x='Candidate Name', y= 'Total converage length by day(s)') +  
  coord_flip()
```

```
ggplot(CoverageLength, aes(x = as.numeric(X2), y = log10(as.numeric(X3)))) +  
  geom_point(pch = 21, fill = "gray25", color = "white", size = 2.5) +  
  scale_x_continuous(name = "Total converage length by day(s)") +  
  scale_y_continuous(name = "Total Contribution log10") +  
  theme_minimal() +  
  theme(text = element_text(size=15))
```