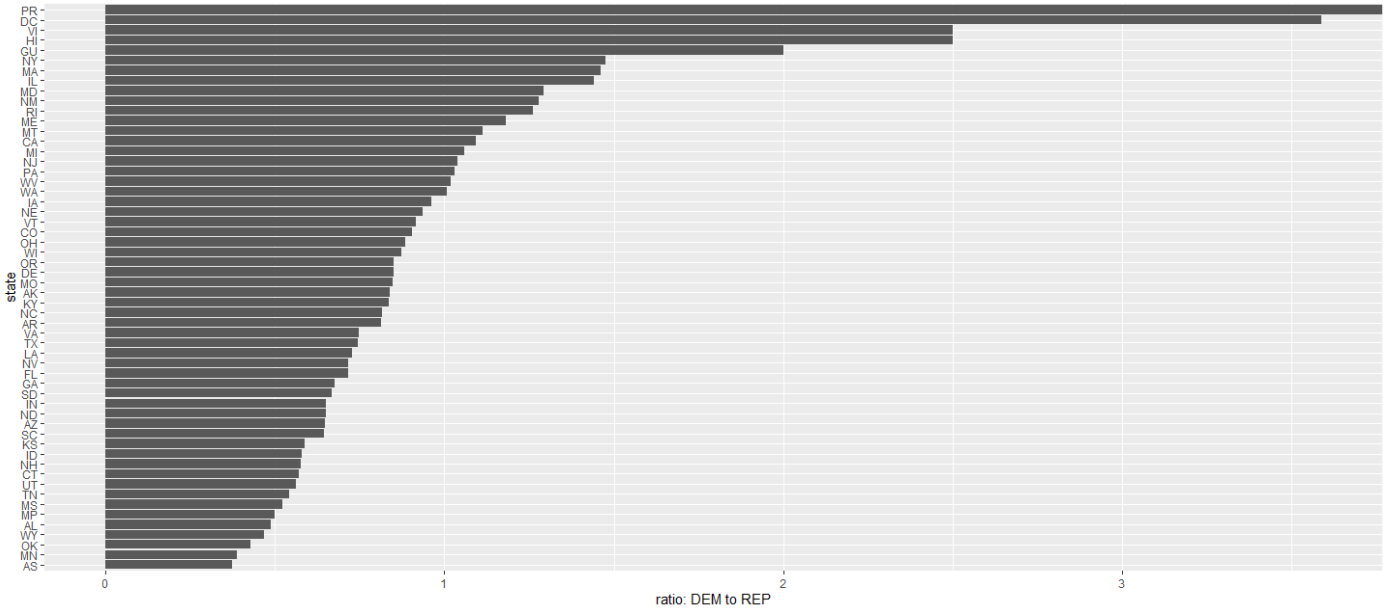
Assignment 2

Megan Rachel Santagata
Nathan Van

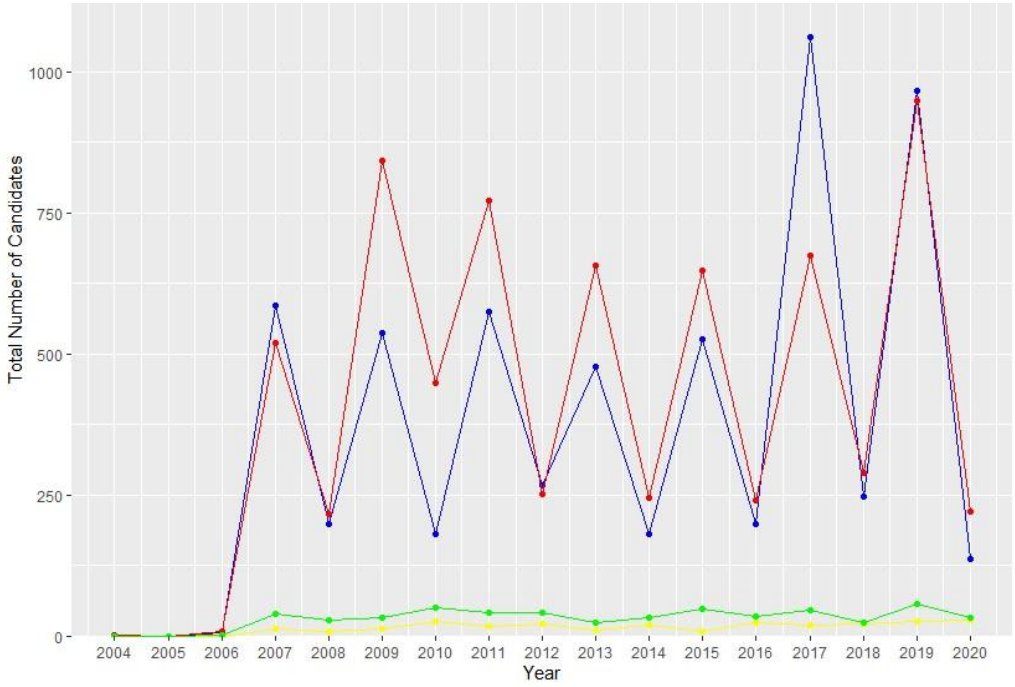
Part 1)

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



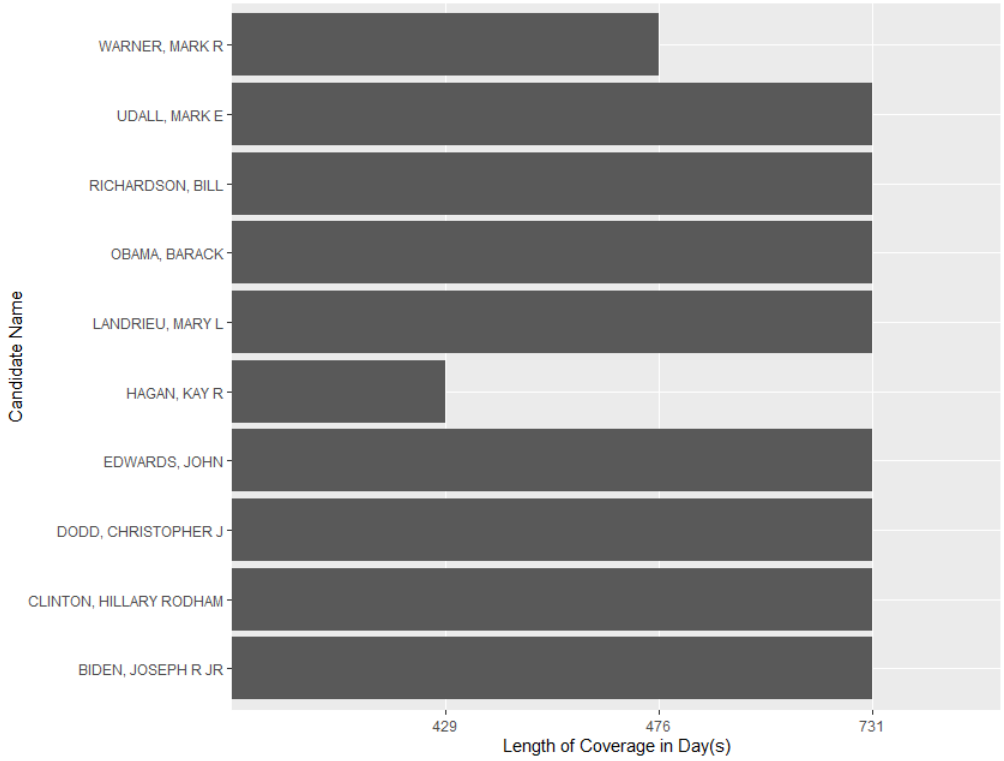
Description: The histogram plot describes the ratio of the number of Democrat candidates to the number of Republican candidates in each state. Less than 1 represents republican majority while greater than 1 represents democrat majority. PA, WA, ME, NJ, MT, AS, RI, CA, HI, IL, DE**,** MA, NY, NM, GU, 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.



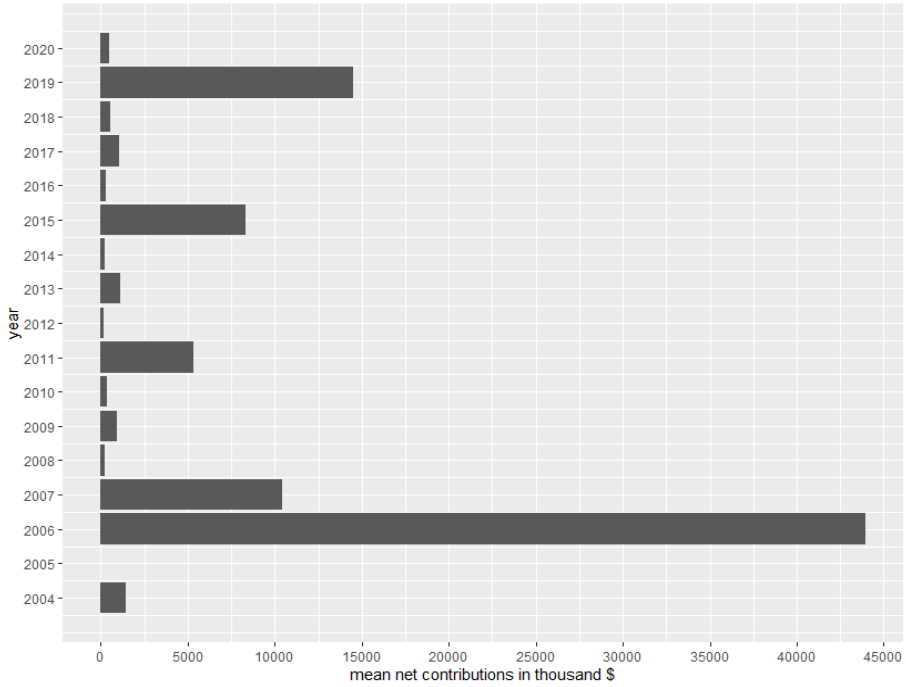
Description: From 2006 to 2008 there were more Democrats (**blue**) than Republicans (**red**), from 2008 to 2016 there were more Republicans (**red**) than Democrats (**blue**), from 2016 to 2018 there were more Republicans (**red**) than Democrats (**blue**), and finally, from 2018 to 2020 the number 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 mainly worked on this graph.

Visualization 3: The top 10 candidates with the highest total contribution ordered by their length of coverage.



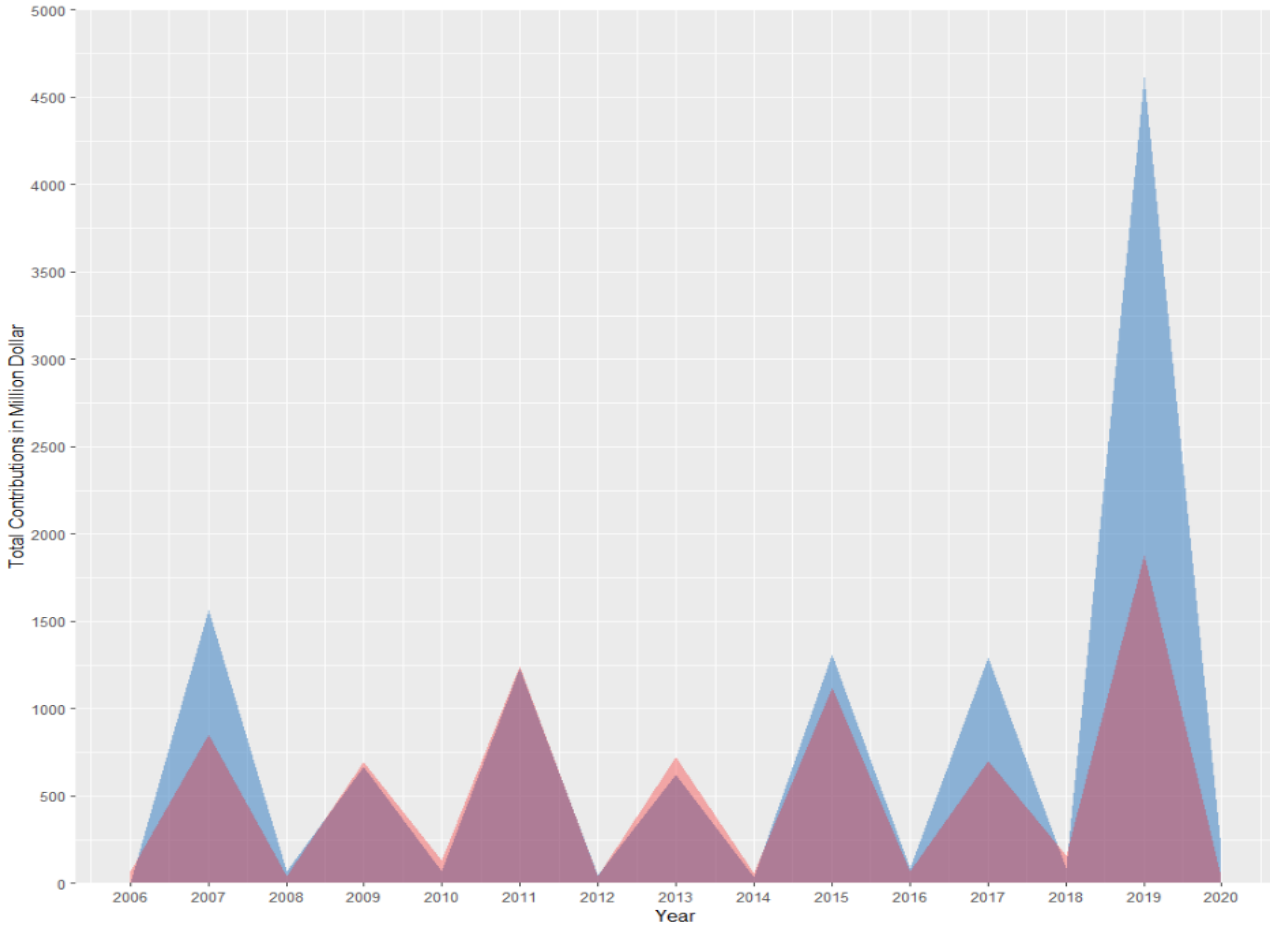
Description: This bar graph shows the top 10 Democrat candidates with the highest total contributions. We compute the coverage length for each of these candidates by subtracting the Coverage_End_Date by the Coverage_Start_Date using the seq() function to get the number of days each candidate gets for coverage. Nathan mainly worked on this graph.

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



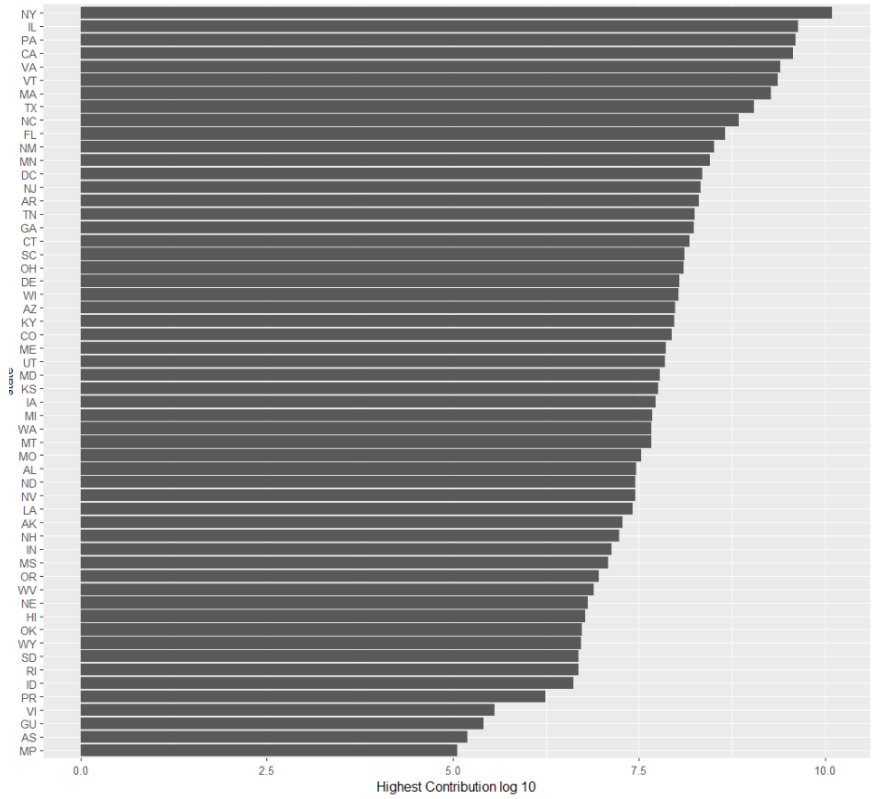
Description: This graph shows the average (mean) net contributions for all political candidates for each year. This includes the political candidates who received zero contributions. Contributions particularly spiked in 2006 and 2019, with the former being having the highest mean net contributions of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

Visualization 6:



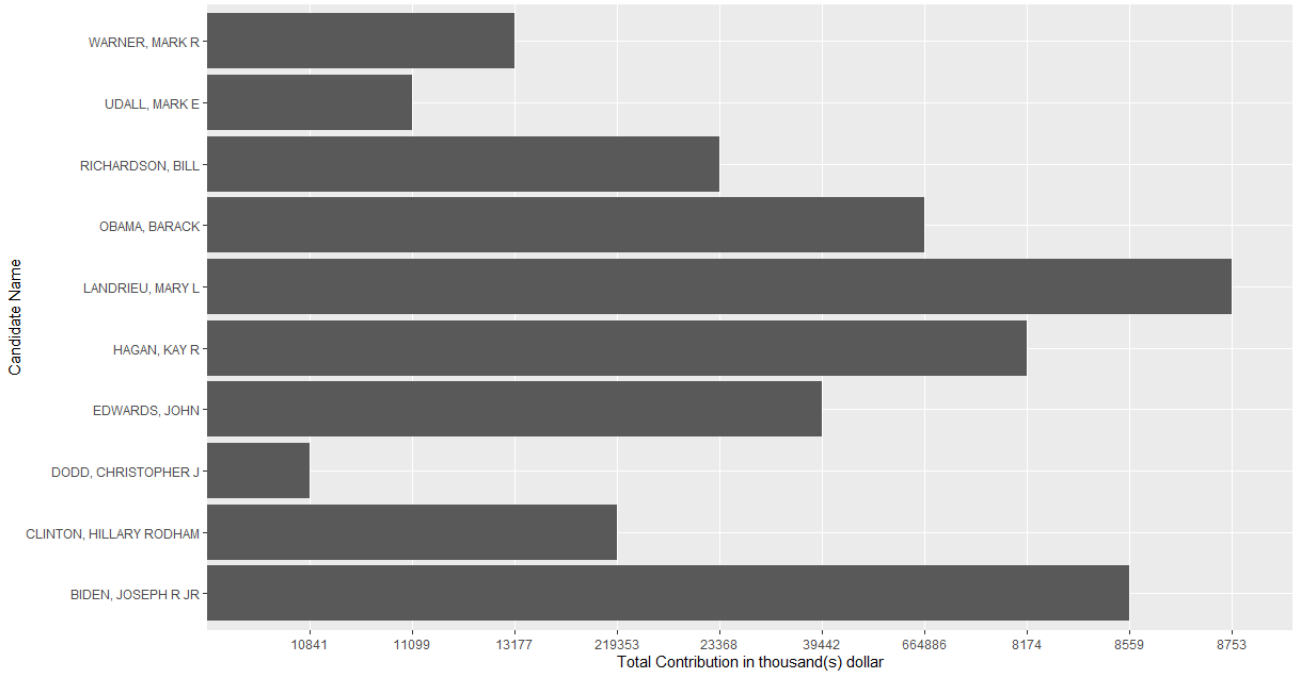
Description: This graph tracks the highest contributions (max) for candidates from each state. This graph uses data from 2004 to 2020. Contributions are measures in log 10 to make sure that they were able to all

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 2) There will be more political candidates of the opposing party than the party of the president currently in office. So, when the president is a Republican, there will be more Democrat candidates, and when the president is a Democrat, there will be more Republican candidates. In visualization 2, we somewhat play out. We see that during 2004 to 2006 when Bush (Republican) was president, there were more Democratic than Republican candidates. Then from 2008 to 2016, during Obama's (Democrat) presidency, there were more republican than Democrat candidates. Since 2016, when Trump (Republican) was elected, there have been more Democratic than Republican candidates. Our hypothesis proved correct, only for odd numbered years, when the number of candidates in both parties peaked. The difference in number of candidates between the two parties was most noticeable in the peak years. There were some years that did not line up with our hypothesis such as 2008, when Bush was still president, there were more republicans, 2012 when Obama was still president and there are more democrats, and 2018 and 2020 when Trump was still president and there were more democrats.

Hypothesis 2: Below is the plot for Visualization 3. We see that the total contribution does not have any correlation with the total length coverage of a candidate. More contribution does not mean more coverage and we can conclude that election coverages were fair and not impacted by financial means.



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

```
#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 != "")
year_data <- subset(myData, Coverage_Start_DateYear != "")
year_data


#graph 1
#how many democrats and republicans in each state?

#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()

dataDEM <- state_data[state_data$Cand_Party_Affiliation=="DEM",]
dataREP <- state_data[state_data$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
# (note: we are counting multiple times if a candidate runs multiple times)
RDdata <- data.frame(state=uniqueStates, demCandNum=demCountVector,
repCandNum=repCountVector)
RDdata

RDdata <- transform(RDdata, newRatios = demCandNum/repCandNum)
RDdata
#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, 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), ]

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()
```

#graph 4

```
#have contributions gotten larger or smaller over time? scatterplots
#plot with a line. use the median or the mode (most probable) it would be a continuous mode. there are
mode estimators: mode est. package in R.
#calculate median.
#Visualization 3
library(tidyverse)

year_data<- read.csv("C:/Users/Megan/OneDrive/school/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
year_data$Coverage_Start_Date <- as.Date(year_data$Coverage_Start_Date, format="%m/%d/%Y")
year_data$Coverage_End_Date <- as.Date(year_data$Coverage_End_Date, format="%m/%d/%Y")

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,]

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

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

ggplot(year_mean, aes(x=year, y=my_mean/1000)) + geom_bar(stat = "identity") +
  labs(x="year", y="mean net contributions in thousand $") +
  scale_x_continuous(breaks = 1*(2004:2020)) +
  scale_y_continuous(breaks = 5000*(0:50000)) +
  coord_flip()
```

#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)))
}

ggplot(totalContributionDem, aes(x=X1, y=X2/1000000)) +
                geom_area(color='transparent', fill='blue') +
                scale_x_continuous(name = 'Year',
                                limits = c(2006, 2020),
                                breaks = 1*(2005:2020)) +
                scale_y_continuous(expand = c(0,0),
                                limits = c(0, 5000),
                                breaks = 500*(0:5000),
                                name = 'Total Contributions in Million Dollar')


ggplot(totalContributionDem, aes(x=X1, y=X2/1000000)) +
  geom_area(color='transparent', fill= alpha("#2C77BF", .5)) +
  scale_x_continuous(name = 'Year',
                limits = c(2006, 2020),
```

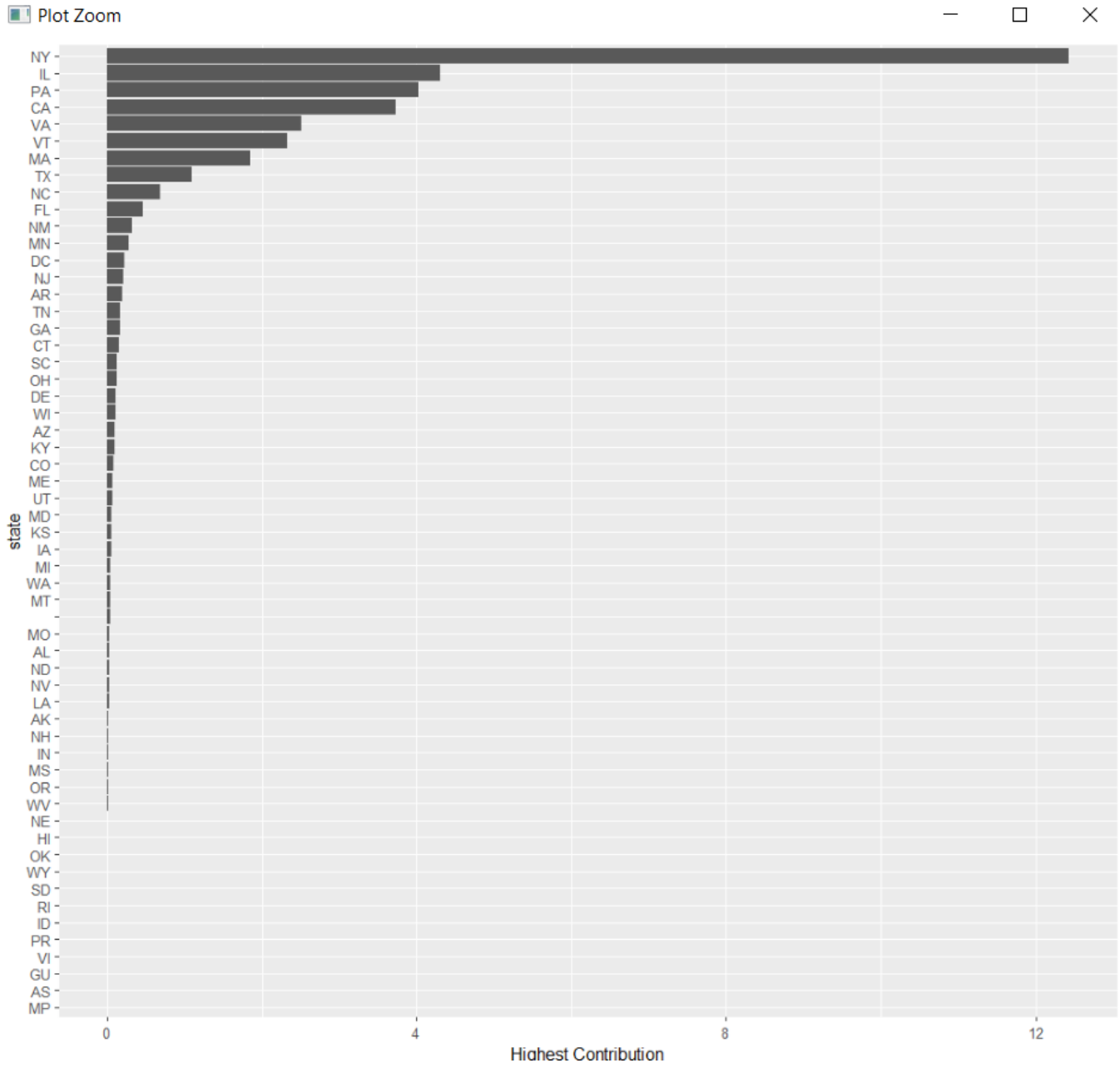
```
                    breaks = 1*(2005:2020)) +
  scale_y_continuous(expand = c(0,0),
                    limits = c(0, 5000),
                    breaks = 500*(0:5000),
                    name = 'Total Contributions in Million Dollar')

ggplot(totalContributionDem, aes(x=X1, y=X2/1000000)) +
  geom_area(color='transparent', fill= alpha("#2C77BF", .5)) +
  scale_x_continuous(name = 'Year',
                    limits = c(2006, 2020),
                    breaks = 1*(2005:2020)) +
  scale_y_continuous(expand = c(0,0),
                    limits = c(0, 5000),
                    breaks = 500*(0:5000),
                    name = 'Total Contributions in Million Dollar') +
  geom_area(aes(x=totalContributionRep$X1, y=totalContributionRep$X2/1000000), color = 'transparent',
fill= alpha("red", .3))
```

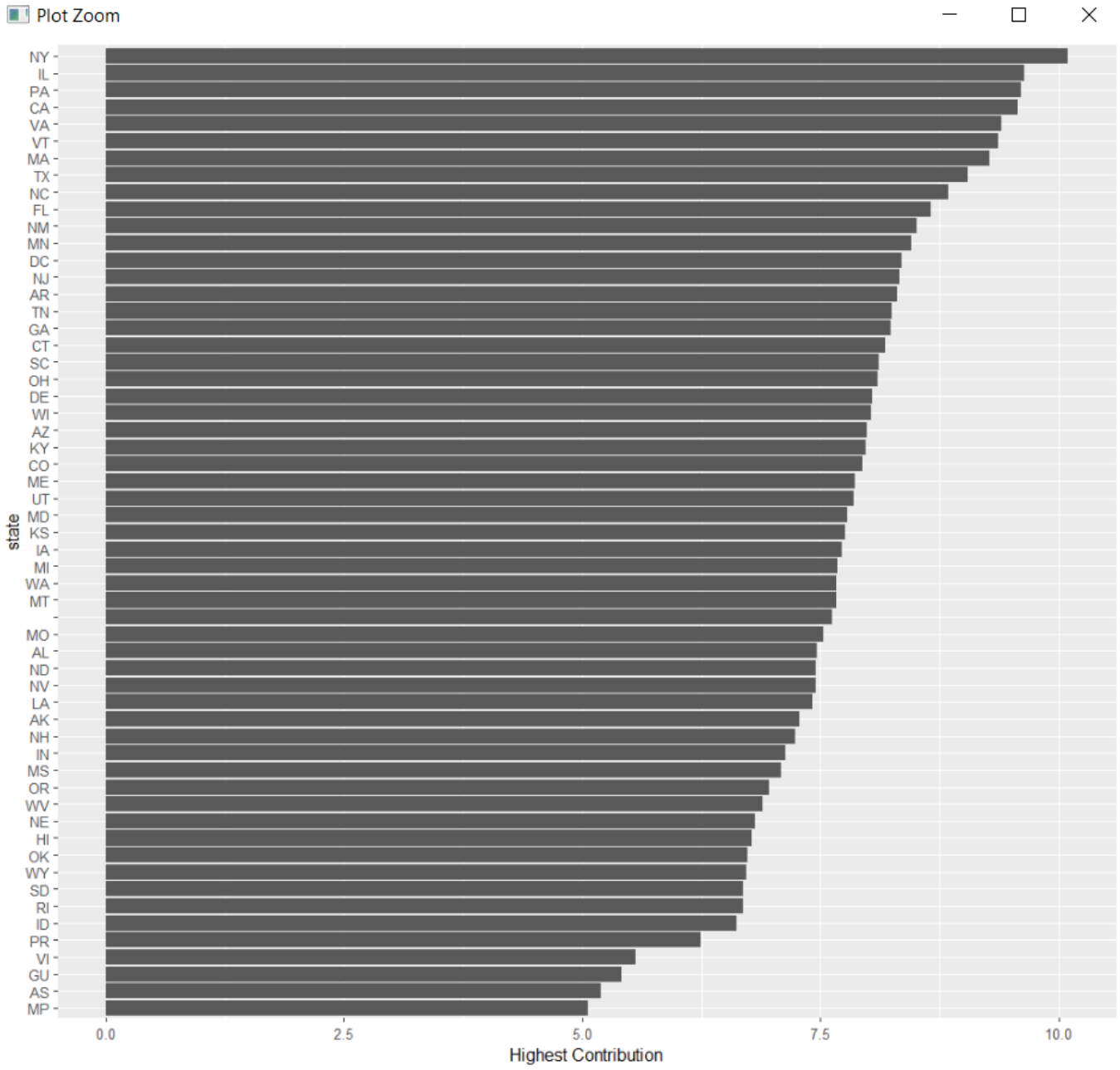
#graph 6

#OR In which state did candidates receive the most net contributions? (average the net contributions and candidates per state)

#run the maximum, write a small loop that runs over each state and find the maximum contribution. What was the maximum contribution in each state?

#what would be the total contribution in each state?


Plot Zoom
NY
IL
PA
CA
VA
VT
MA
TX
NC
FL
NM
MN
DC
NJ
AR
TN
GA
CT
SC
OH
DE
WI
AZ
KY
CO
ME
UT
MD
KS
IA
MI
WA
MT
MO
AL
ND
NV
LA
AK
NH
IN
MS
OR
WV
NE
HI
OK
WY
SD
RI
ID
PR
VI
GU
AS
MP
state
0
4
8
12
Highest Contribution


Plot Zoom
NY
IL
PA
CA
VA
VT
MA
TX
NC
FL
NM
MN
DC
NJ
AR
TN
GA
CT
SC
OH
DE
WI
AZ
KY
CO
ME
UT
MD
KS
IA
MI
WA
MT
MO
AL
ND
NV
LA
AK
NH
IN
MS
OR
WV
NE
HI
OK
WY
SD
RI
ID
PR
VI
GU
AS
MP
state
0.0
2.5
5.0
7.5
10.0
Highest Contribution