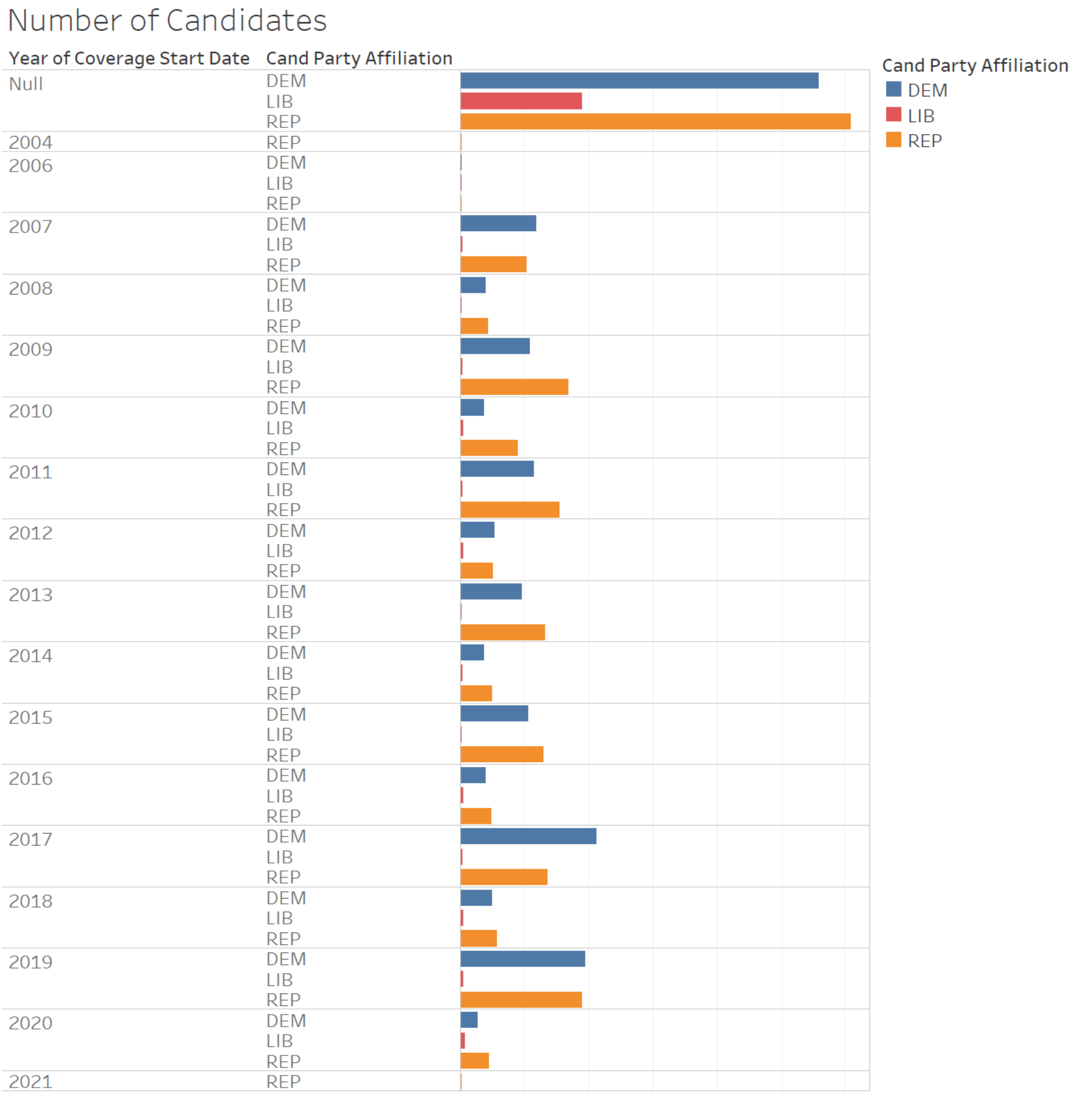
Enze Hu

Part 1: Explore the dataset

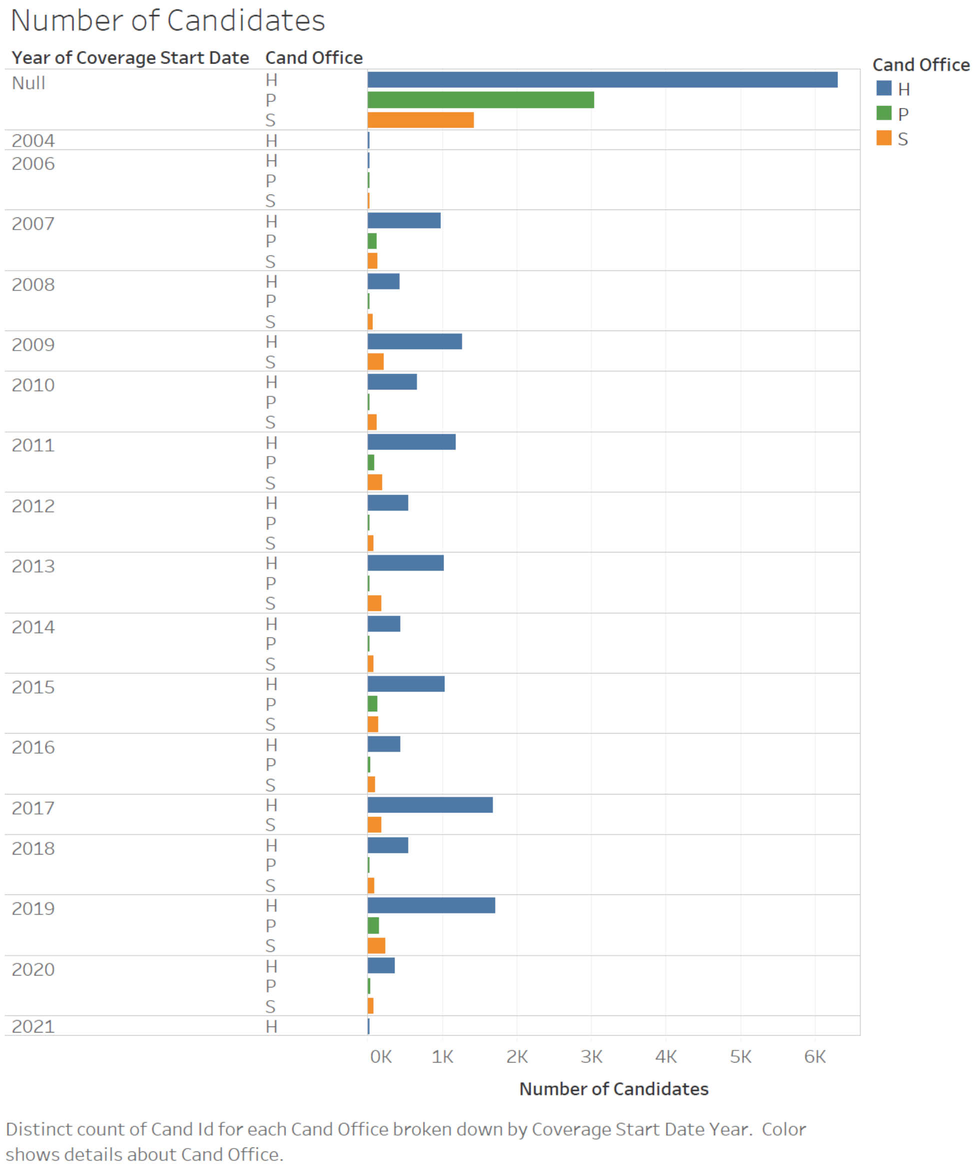
Figure 1: Number of Candidates by Year coverage and Party Affiliation



Distinct count of Cand Id for each Cand Party Affiliation broken down by Coverage Start Date Year. Color shows details about Cand Party Affiliation. The view is filtered on Cand Party Affiliation, which keeps DEM, LIB and REP.

Democracy and republican are the two main party, followed by the liberty party. All other parties are quite small and has relatively low number of candidates, and there are too many of small parties, so we void them here to keep the chart neat. Democracy party has more candidates in 2019, 2017; all the other years Republicans has more candidates, including when the years are null.

Figure 2: Number of Candidates by year, subset by Candidates Office



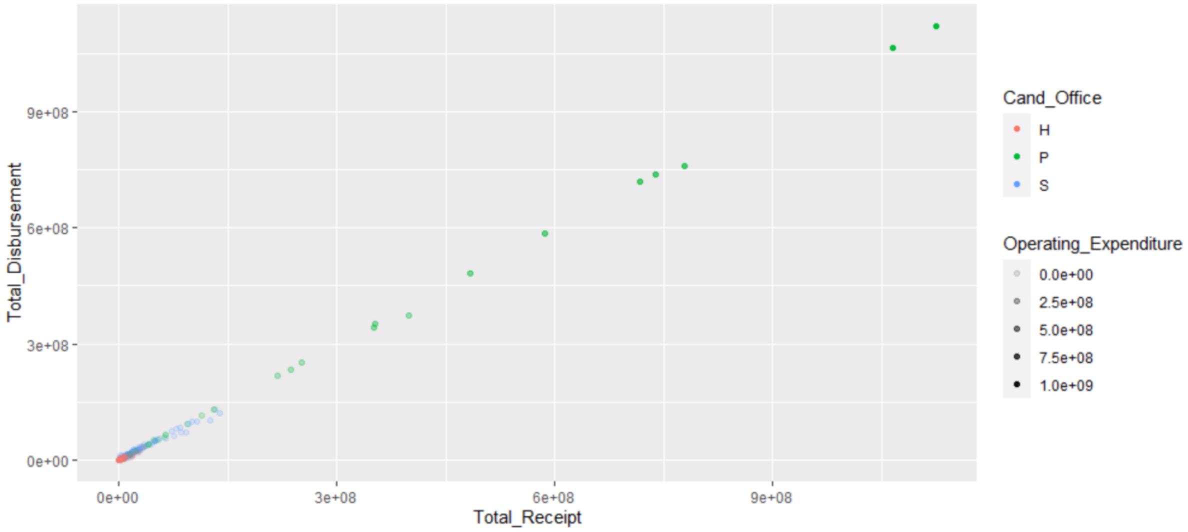
Here we use a similar approach as the previous one, only to explore the candidates office subset. We can see that the House has most candidates, followed by senate, and president office.

We want to know the relative relationship between number of candidates registered for House, President, and Senate. Except the Null data, we can see that the 2019 has the highest amount of registered House candidates, followed by 2017.

2019 also has the highest number of president candidates, followed by 2015.

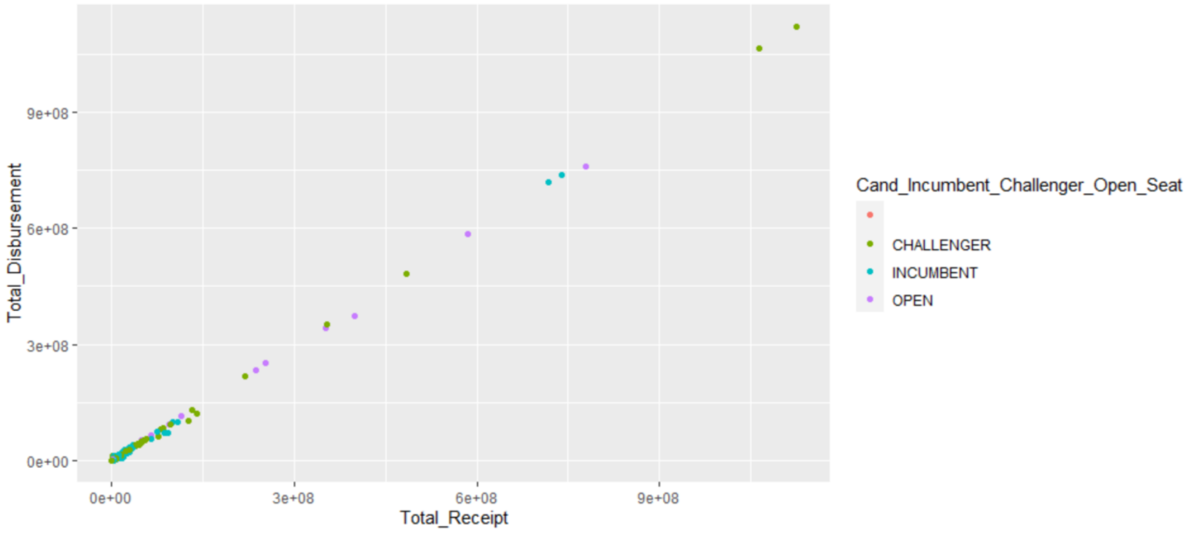
As for senator, 2019 has the highest amount of registered senator, followed by 2009.

Figure 3: Total_Disbursement vs Total_receipt, By Cand_Office & Operating_expenditure



In this figure we want to find that if candidates' office have an impact on their total_receipt, total_disbursement, and their operating_expenditure. It's easy to tell that the green dots, which represents president candidates, have the highest total_receipt and disbursement, as well as highest operating_expenditure. The red dots (which represent house candidates has the lowest total receipt, total disbursement and operating expenditure. Senators are in the middle.

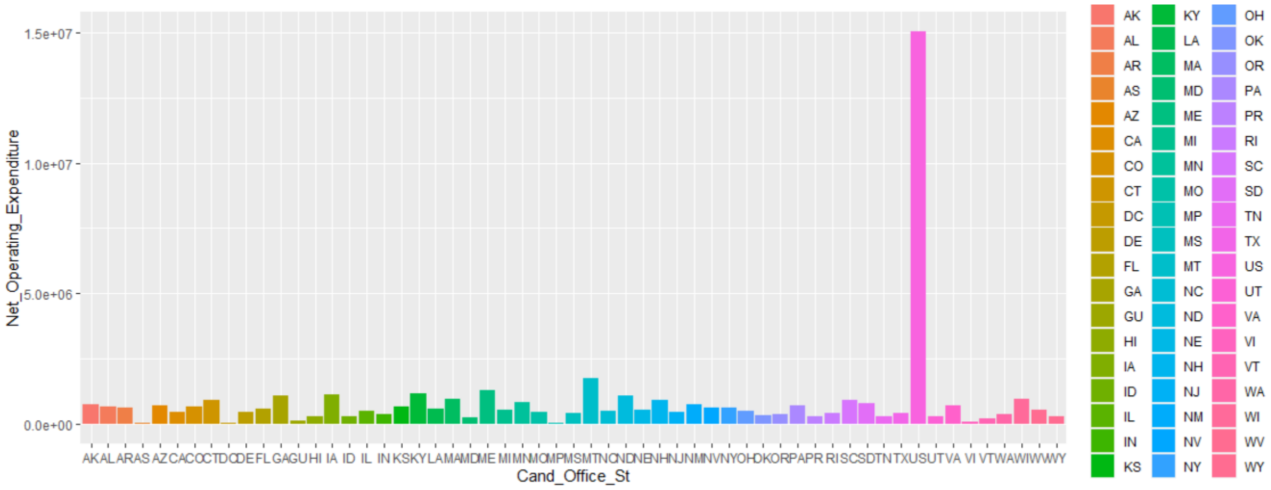
Figure 4: Total_receipt vs total_disbursement grouped by Cand_incumbent_challenger open seat



In this chart we are looking at how challenger, incumbent, and open positon differ from each other.

Here we can found that the challenger have the highest total receipt and disbursement, which are president candidates. Open and incumbent are more mixture together, hard to differentiate.

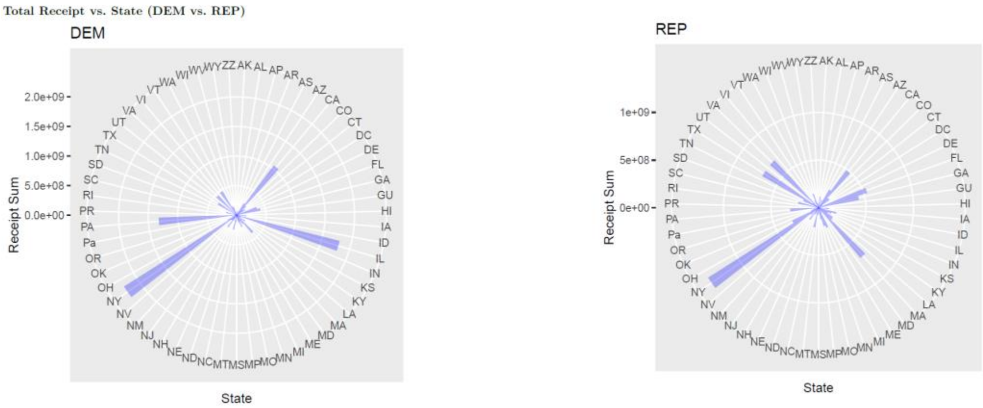
Figure 5:



There is no doubt that the U.S. in total has the highest net operating expenditure.

Among all the states, MT(MONTANA) has the highest net operating expenditure, followed by ME (MAINE).

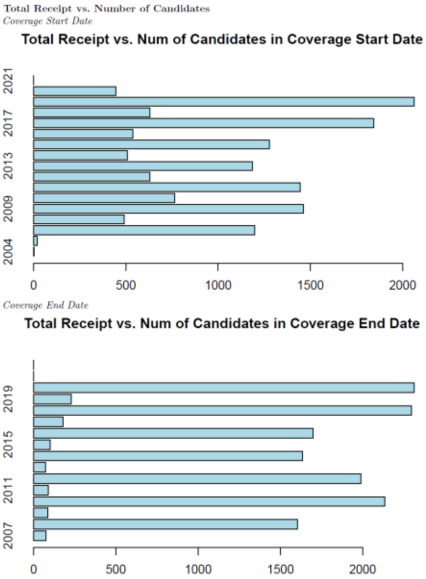
Figure 6:Total receipt Vs States grouped by parties (DEM VS REP)



We can see that for the democracy party, the New York state has the highest total receipt, followed by IL, PA and CA.

For the republican party, the highest total receipt is still the New York state, followed by TX and VA, then FL and CA.

Figure 7: total receipt vs No. of Candidates in coverage start date



Enze Hu

These 2 charts show the total receipt vs number of Candidates in different coverage periods. 2019 has the highest amount of total receipt.

Part 2: hypothesis:

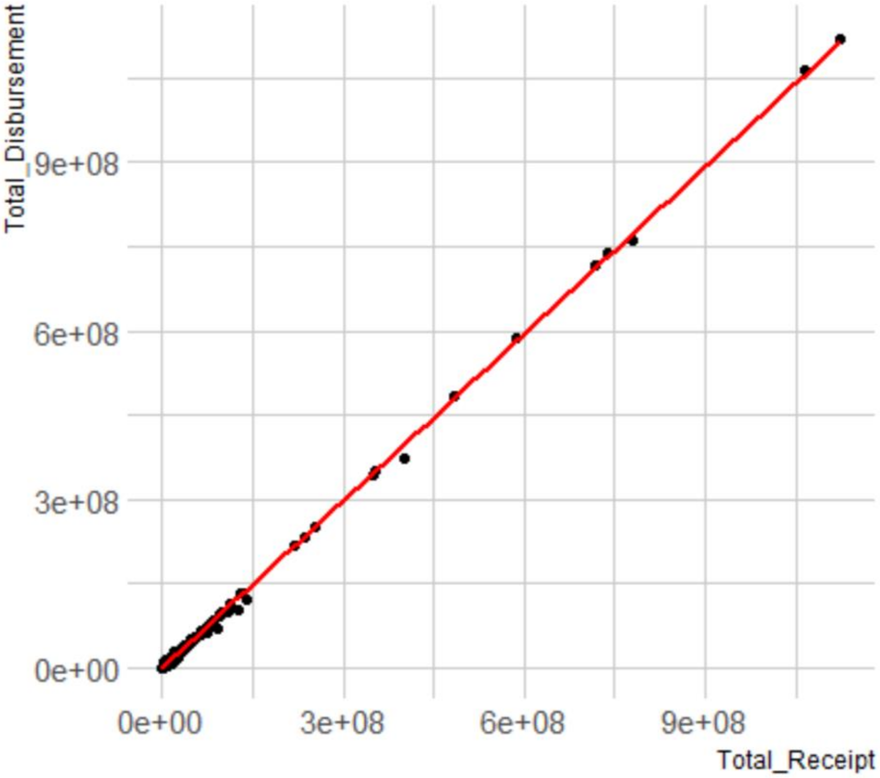


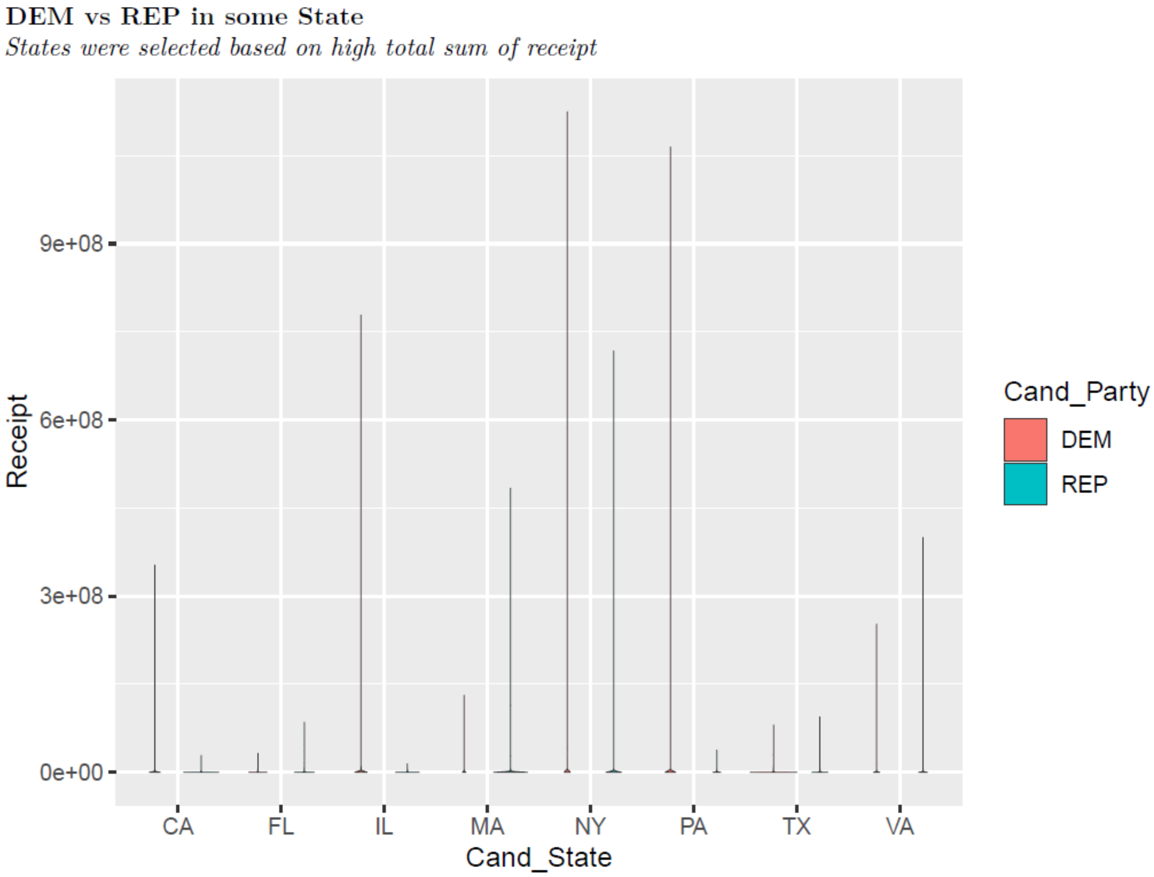
Figure 8: Correlation check for Total_receipt and Total_Disbursement

Regarding to the Total_Receipt, minimum is 0, median is 2.28e+03, mean is 8.89e+05, max is 1.125e+9.

For the total disbursement, min is -1458, which we suppose is a typo here, median is 2150, mean is 865930, max is 1121170037.

Enze Hu

Our hypothesis here is the total disbursement has linear correlation with total receipt. First, we plot the basic scatter plot, then we added linear trend. As we can see here, most of the candidates has a total receipt under $3e+08, with only several candidates with higher amount of receipt money. Almost all the candidates will spend all of their receipt money, with only a few exceptions.



At last, we can check the DEM vs Rep in different Cand_state. NY has the highest number of DEM receipt, and the Rep receipt is also pretty high. PA state has the largest difference between DEM and REP, and we can inter that DEM will always win PA state.

Appendix: R Code

I appendix all the code I run to generate figure at here.

```
install.packages("dplyr")

install.packages('hrbrthemes')

library(dplyr)

library(ggplot2)

library(hrbrthemes)

election<-read.csv("data1.csv", header = TRUE))

#explore the dataset

head(election)

summary(election)

#plot our dataset

plot(election)

#Error in plot.new() : figure margins too large

# first sight into the dataset, explore the total receipt and total disbursement

plot(election$Total_Receipt, election$Total_Disbursement)

#seems that here is a linear trend

#ggplot election

ggplot(election, aes(x=Total_Receipt, y=Total_Disbursement, col = Cand_Office))+
  geom_point()
ggplot(election, aes(x=Total_Receipt, y=Total_Disbursement, col =
Cand_Incumbent_Challenger_Open_Seat))+
  geom_point()
#We plotted 2 to explore the receipt and disbursement for Cand_office and Challenger/open seat
ggplot(election, aes(x=Total_Receipt, y=Total_Disbursement, col = Cand_Office, alpha =
Operating_Expenditure
))+
  geom_point()
#Bars and box plot
ggplot(election,aes(Cand_Office_St, Net_Operating_Expenditure,fill = Cand_Office_St ))+
  geom_bar(stat = "summary", fun.y = "mean")
ggplot(election,aes(Cand_Office, Net_Operating_Expenditure,fill = Cand_Office))+
  geom_boxplot(fill = "blue", col = "red", notch = TRUE)
election$Cand_Office
ggplot(election, aes(x=Total_Receipt, y=Total_Disbursement)) +
  geom_point()
# Use tableau to generate the subgroup by parties and office since it’s more convenience.
#Basic scatter plot.
p1<- ggplot(election, aes(x=Total_Receipt, y=Total_Disbursement)) +
  geom_point( color="#69b3a2") +
  theme_ipsum()

p1

# with linear trend

p2 <- ggplot(election, aes(x=Total_Receipt, y=Total_Disbursement)) +

  geom_point() +

  geom_smooth(method=lm , color="red", se=FALSE) +

  theme_ipsum()

p2
```

DEM vs REP in some State

\newline

States were selected based on high total sum of receipt

```
library(ggplot2)

library(dplyr)

library(forcats)

library(viridis)

library(hrbrthemes)



c <- c("NY", "IL", "CA", "PA", "MA", "VA", "TX", "FL", "GA")
```

Enze Hu

```
viodata_imp_state <- viodata[viodata$Cand_State %in% c("NY", "IL", "CA", "PA", "MA", "VA", "TX", "FL"), ]


viodata_imp_state %>%
  ggplot(aes(fill = Cand_Party, y = Receipt, x = Cand_State)) +
  geom_violin(position = "dodge", alpha = 5, size = 0.2)
```

Total Disbursement vs. Total Receipt

```
data <- read.csv("data1.csv", header = TRUE)


Start_date <- substring(data$Coverage_Start_Date, 7,10)
End_date <- substring(data$Coverage_End_Date, 7,10)
```

~~~~
Cand_Name <- data$Cand_Name

Cand_Party <- data$Cand_Party_Affiliation

Cand_State <- data$Cand_State

Receipt <- data$Total_Receipt

Disbursement <- data$Total_Disbursement



plotdata <- data.frame(Start_date, End_date, Cand_Name, Cand_Party, Cand_State, Receipt, Disbursement)



plot(plotdata$Receipt, plotdata$Disbursement, xlab = "Total Receipt", ylab = "Total Disbursement", main = "Disbursement vs. Receipt")

abline(lm(plotdata$Disbursement ~ plotdata$Receipt), col = "blue")

```

\newpage

**Total Receipt vs. Year**

\newline

*Total Receipt vs Coverage Start Date*



```{r, echo = FALSE}

library(gtable)

unique_start_date <- as.numeric(unique(plotdata$Start_date))
~~~~

```
unique_start_date <- sort(unique_start_date[!is.na(unique_start_date)])
n_start_date <- length(unique_start_date)
sum_receipt <- vector(mode = "numeric", length = n_start_date)

unique_end_date <- as.numeric(unique(plotdata$End_date))
unique_end_date <- sort(unique_end_date[!is.na(unique_end_date)])
n_end_date <- length(unique_end_date)
sum_receipt_end <- vector(mode = "numeric", length = n_end_date)

for (i in 1:n_start_date) {
  sum_receipt[i] <- sum(plotdata[plotdata$Start_date == unique_start_date[i],]$Receipt, na.rm =
TRUE)
}

for (i in 1:n_end_date) {
  sum_receipt_end[i] <- sum(plotdata[plotdata$End_date == unique_end_date[i],]$Receipt, na.rm =
TRUE)
}


barplot(sum_receipt, main = "Total Receipt vs. Coverage Start Date",
```

~~~~
           xlab = "Year",

       names.arg = unique_start_date,

       col = "lightblue")


```




*Total Receipt vs. Coverage End Date*


```{r, echo = FALSE}

barplot(sum_receipt_end, main = "Total Receipt vs. Coverage End Date", xlab = "Year",

       names.arg = unique_end_date,

       col = "lightblue")

```


\newpage

**Total Receipt vs. Number of Candidates**

\newline

*Coverage Start Date*
~~~~

~~~
```{r, echo = FALSE}
num_cand <- vector(mode = "numeric", length = n_start_date)
num_cand_end <- vector(mode = "numeric", length = n_end_date)
for (i in 1:n_start_date) {
  u <- unique(plotdata[plotdata$Start_date == unique_start_date[i],]$Cand_Name)
  num_cand[i] <- length(u)
}

for (i in 1:n_end_date) {
  u <- unique(plotdata[plotdata$End_date == unique_end_date[i],]$Cand_Name)
  num_cand_end[i] <- length(u)
}


barplot(num_cand, main = "Total Receipt vs. Num of Candidates in Coverage Start Date",
        names.arg = unique_start_date,
        col = "lightblue",
        horiz = TRUE)
~~~

~~~
Enze Hu

```

*Coverage End Date*

```{r, echo = FALSE}

barplot(num_cand_end, main = "Total Receipt vs. Num of Candidates in Coverage End Date",

        names.arg = unique_end_date,

        col = "lightblue",

        horiz = TRUE)

```

\newpage

**Total Receipt vs. State (DEM vs. REP)**

```{r, echo = FALSE}

library(ggplot2)

unique_state <- unique(plotdata$Cand_State)
~~~

```
unique_state <- unique_state[unique_state!=""]
state_len <- length(unique_state)
sum_receipt_dem <- vector(mode = "numeric", length = state_len)
sum_receipt_rep <- vector(mode = "numeric", length = state_len)


viodata <- plotdata[(plotdata$Cand_Party == "DEM" | plotdata$Cand_Party == "REP"),]
viodata <- viodata[!is.na(viodata$Cand_Party),]
viodata_dem <- viodata[viodata$Cand_Party == "DEM",]
viodata_rep <- viodata[viodata$Cand_Party == "REP",]
for (i in 1:state_len) {
  sum_receipt_dem[i] <- sum(viodata_dem[viodata_dem$Cand_State == unique_state[i],]$Receipt)
  sum_receipt_rep[i] <- sum(viodata_rep[viodata_rep$Cand_State == unique_state[i],]$Receipt)
}



dem <- data.frame(unique_state, sum_receipt_dem)
rep <- data.frame(unique_state, sum_receipt_rep)


ggplot(dem, aes(x = as.factor(unique_state), y = sum_receipt_dem)) +
  geom_bar(stat = "identity", fill = alpha("blue", 0.3)) +
```

~~~
  coord_polar(start = 0) +
  labs(x = "State",
       y = "Receipt Sum",
       title = "DEM")
```
~~~

~~~
```{r, echo = FALSE}
ggplot(rep, aes(x = as.factor(unique_state), y = sum_receipt_rep)) +
  geom_bar(stat = "identity", fill = alpha("blue", 0.3)) +
  coord_polar(start = 0) +
  labs(x = "State",
       y = "Receipt Sum",
       title = "REP")
```
~~~

\newpage

Enze Hu

DEM vs REP in some State

\newline

States were selected based on high total sum of receipt

~~~
```{r, echo = FALSE, warning=FALSE, message=FALSE}

library(ggplot2)

library(dplyr)

library(forcats)

library(viridis)

library(hrbrthemes)


c <- c("NY", "IL", "CA", "PA", "MA", "VA", "TX", "FL", "GA")

viodata_imp_state <- viodata[viodata$Cand_State %in% c("NY", "IL", "CA", "PA", "MA", "VA", "TX", "FL"), ]


viodata_imp_state %>%

  ggplot(aes(fill = Cand_Party, y = Receipt, x = Cand_State)) +

  geom_violin(position = "dodge", alpha = 5, size = 0.2)
~~~