Financial Contributions to Presidential Campaigns by CA in 2016

> **Tip**: One of the requirements of this project is that your code follows good formatting techniques, including limiting your lines to 80 characters or less. If you're using RStudio, go into Preferences > Code > Display to set up a margin line to help you keep track of this guideline!

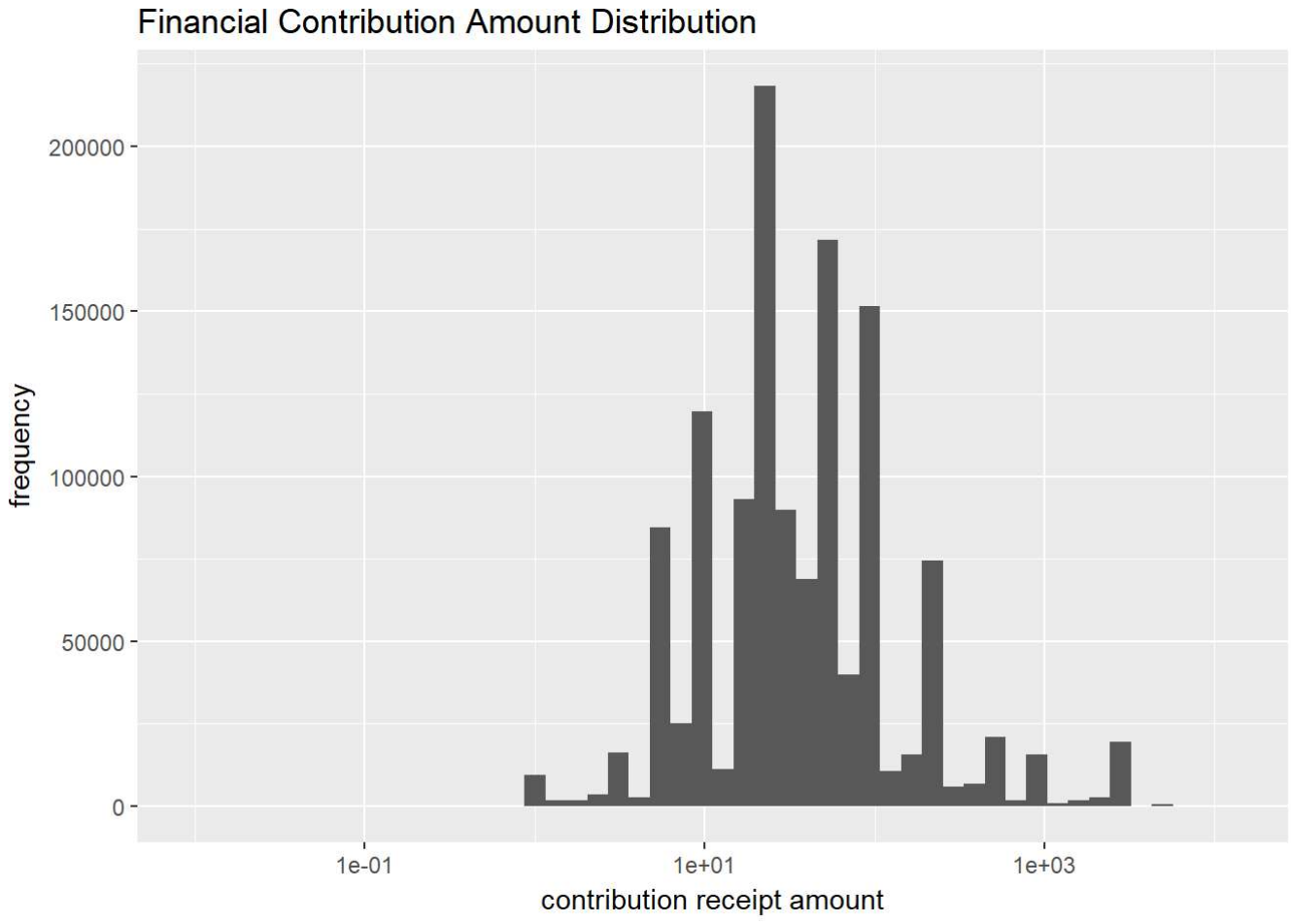
```
## [1] "D:/Udacity-DAND/project 4"
```

The dataset was a collection of financial contributions to presidential campaigns by CA in 2016. It contained 19 variables and 130,4346 observations in this dataset. And most variables are factor variables.

Univariate Plots Section

receipt amount distribution

```
## [1] 19
```

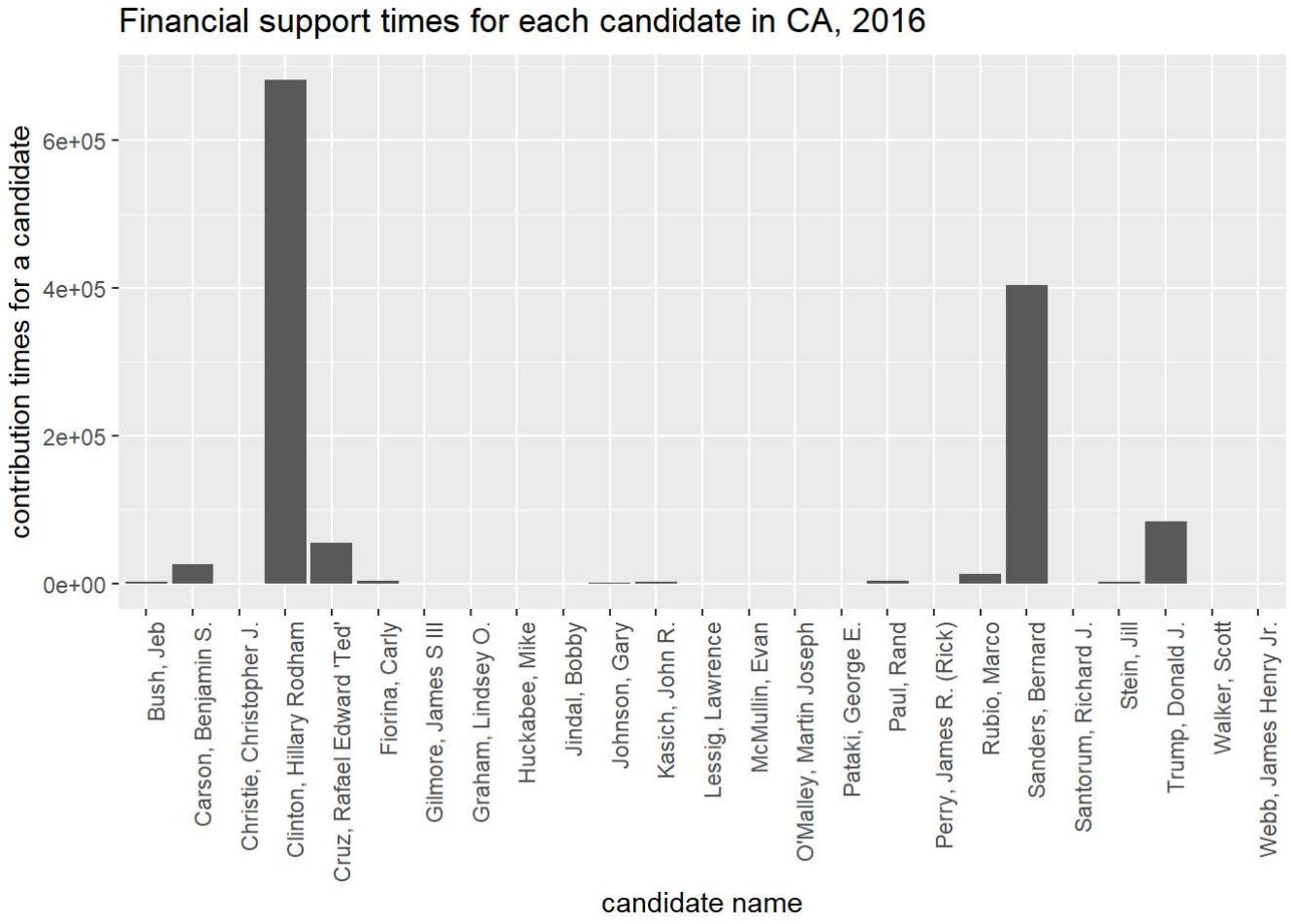


committee & candidate

```
## [1] 25
```

```
## [1] 25
```

```
## [1] 25
```

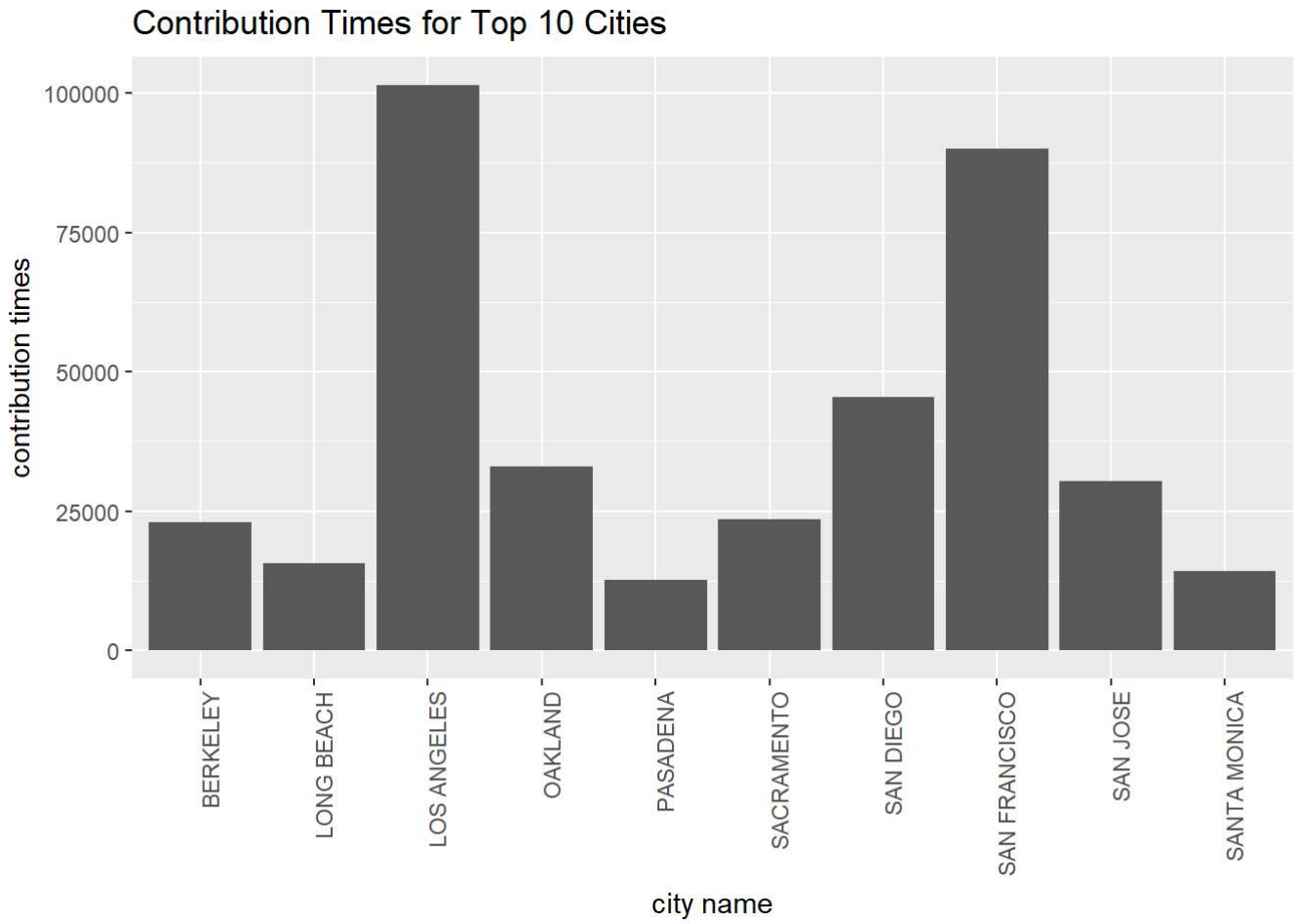



```
## [1] 16
```

contributing cities

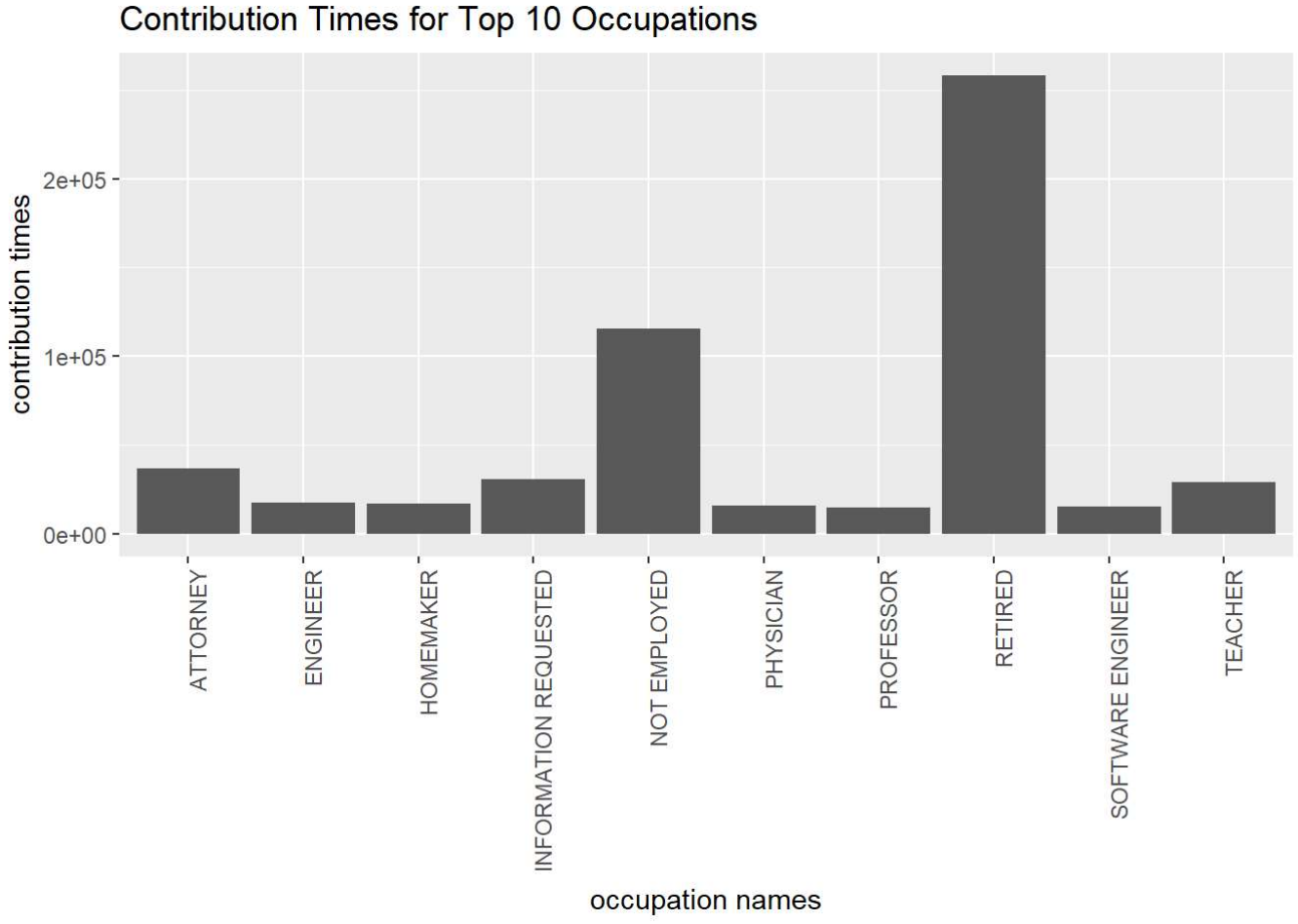
```
## [1] 2517
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

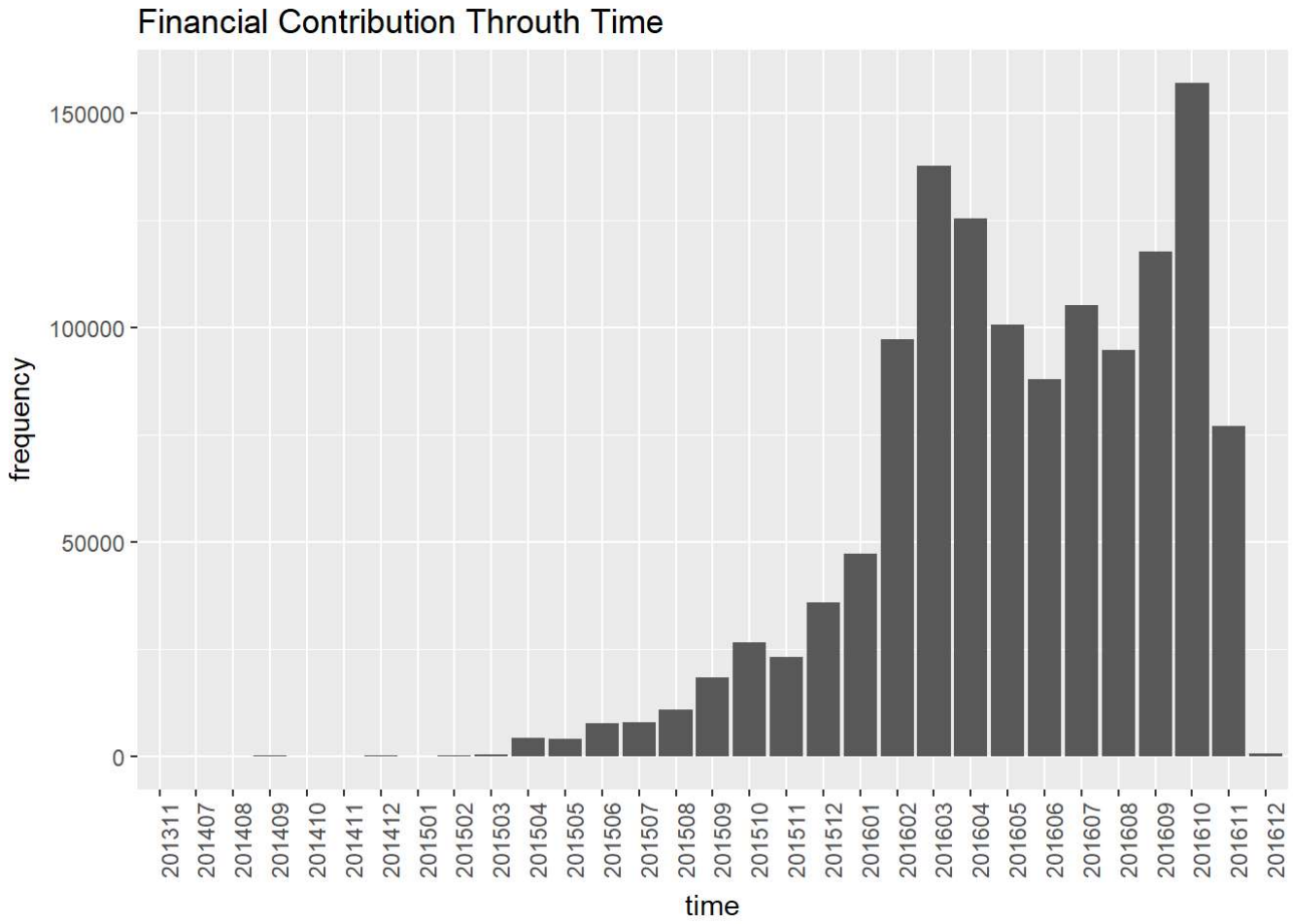


occupation

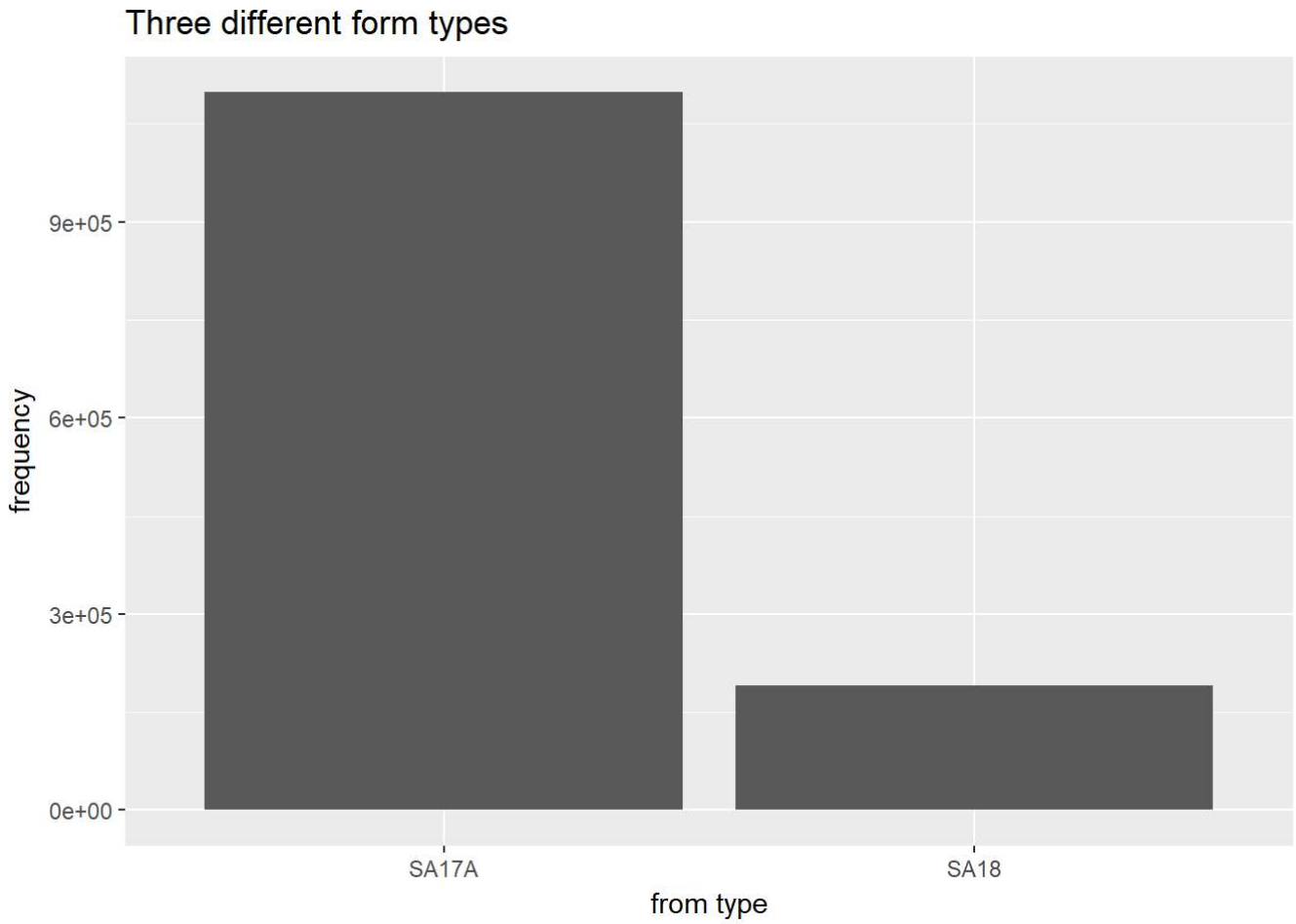
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



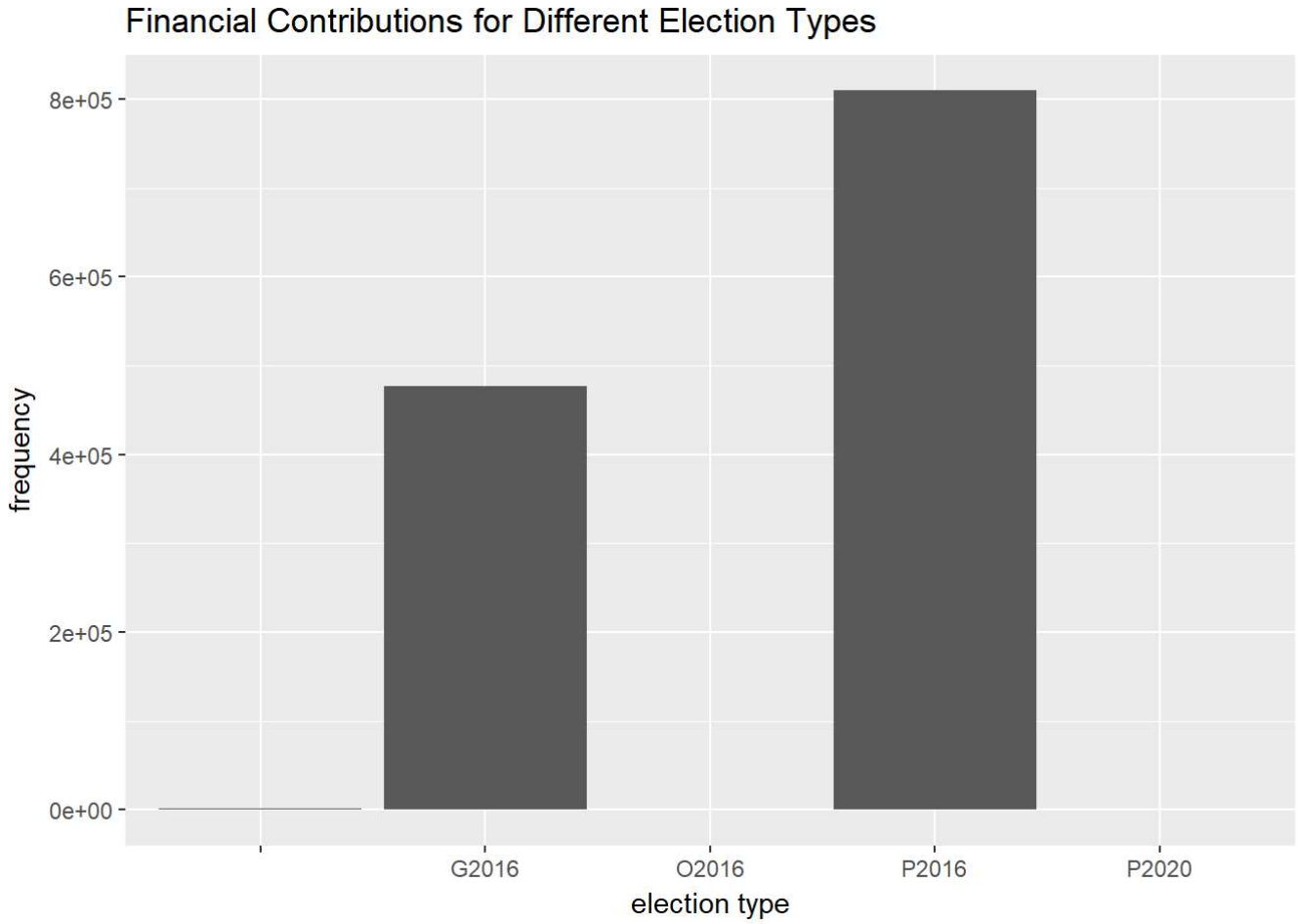
contribution time distribution



form type



election type



Univariate Analysis

Your idea?

1. I'm interested to find how was each candidate being supported financially in presidential campign in CA and want to figure out who was the most popular candidate in CA, 2016.
2. I'm also interested in using financial support data to predict the proportion of votes in general election in CA， 2016.

What is the structure of your dataset?

And most of the variables are factor variables, like contributing city, contributing name and etc. It's not a cross section dataset(snapshot taken at a given time). It recorded every financial contribution during the election process.

What is/are the main feature(s) of interest in your dataset?

- The financial contribution amount distribution is kind of a normal distribution in log scale. And most contributions' amounts are between tens to hundreds.
- There were 25 candidateds being financially supported during the presidential campaign in 2016. The Democratic was well supported compared to the other parties. And Clinton, Hillary Rodham and Sanders, Bernard received top 2 financial contributions ranked by frequency.
- There were more than 2500 cities contributed in this presidential campaign in CA in 2016, of which Los Angeles, San Francisco and San Diego were top 3 contributing cities ranked by contribution times.
- It was really surprised to find that retired men and not employed people contributed for most times among all occupations.
- When looking at the total contribution amount for each month, we could figure out that most contributions were made in 2016. And there were two contribution peaks, one in 201603(during primary election), and the other in 201610(one month before the general election).
- Actullay, most contributions were made for primary election and general election.
- Most contributions came from indivials, and a few were transferring from authorized committees.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

??can't understand

Did you create any new variables from existing variables in the dataset?

There are a few changes I made to the original dataset. (1) I changed the format of the contribution date and also creates another column(called `contb_receipt_dt_yr_mo`) representing the year and month. (2) I combined the committee id and candidate name to from a new column to check whether committees and candidates were correlated one by one.

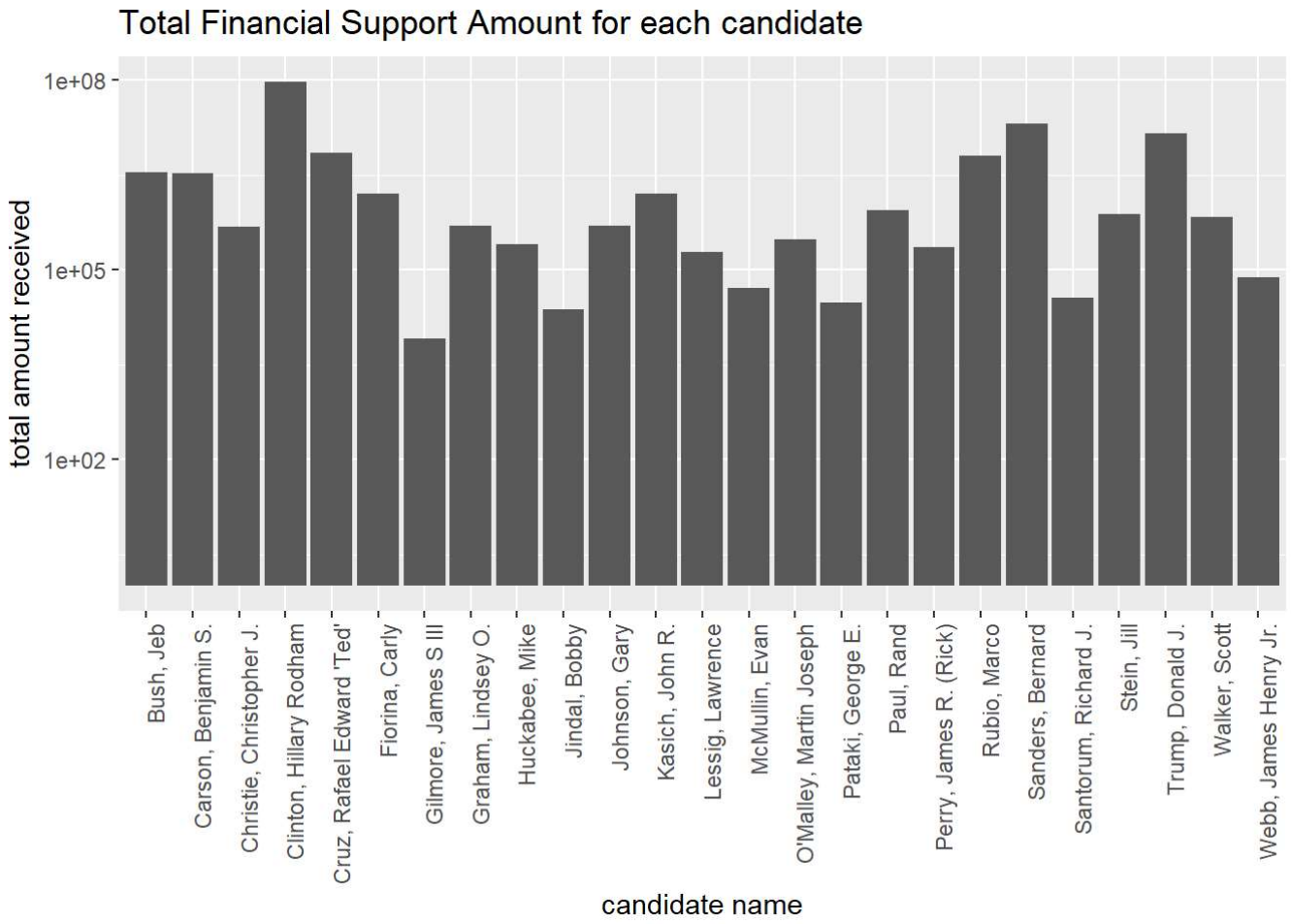
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or

change the form
of the data? If so, why did you do this?

1. There were some records with negative contribution receipt amount in the dataset(original_financial). So I excluded all negative values in mny dataset and I used modified dataset(financial) for my exploration.
2. And retired man contributed most times of all occupations, which really surprised me.

Bivariate Plots Section

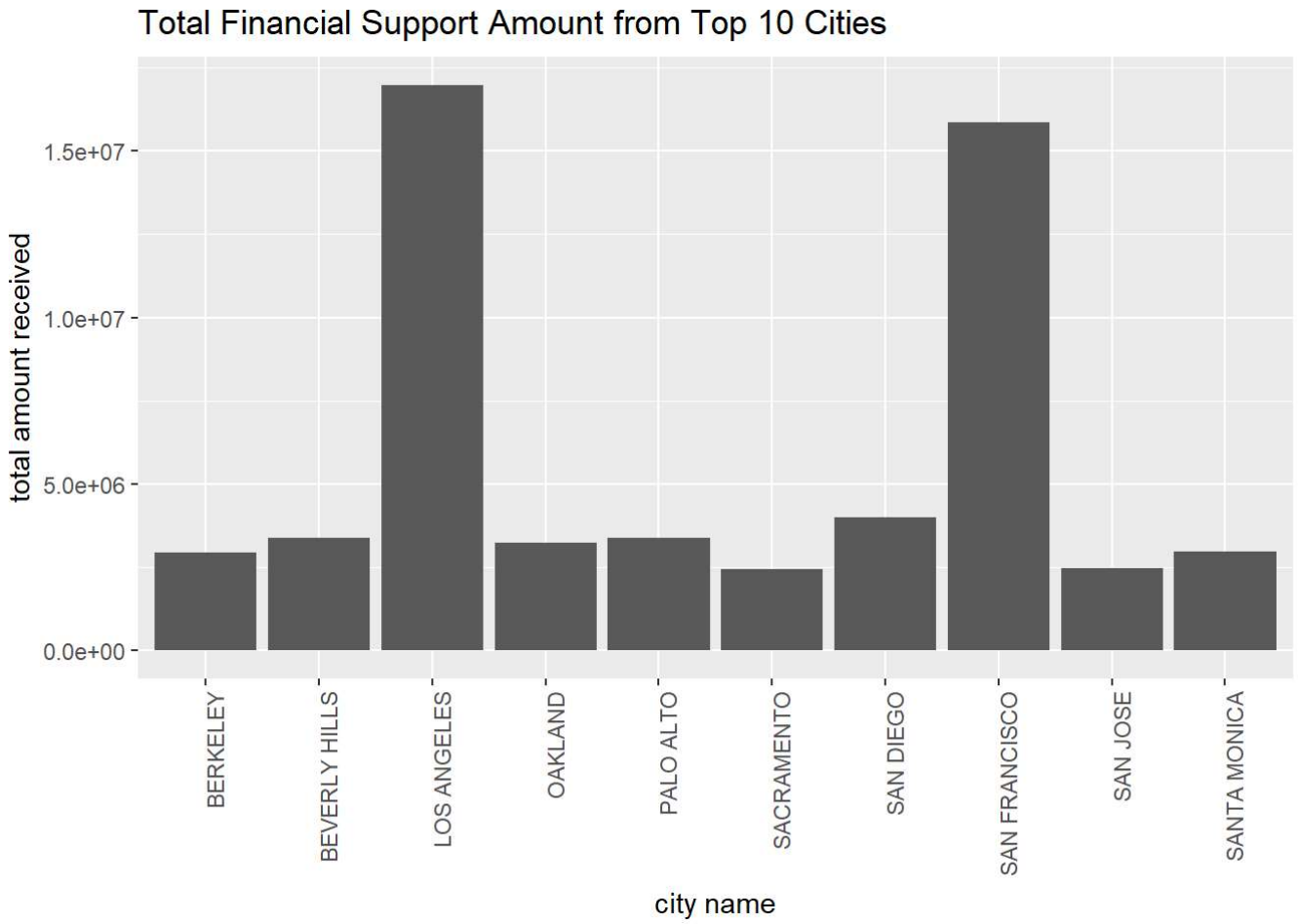
candidate & receipt amount



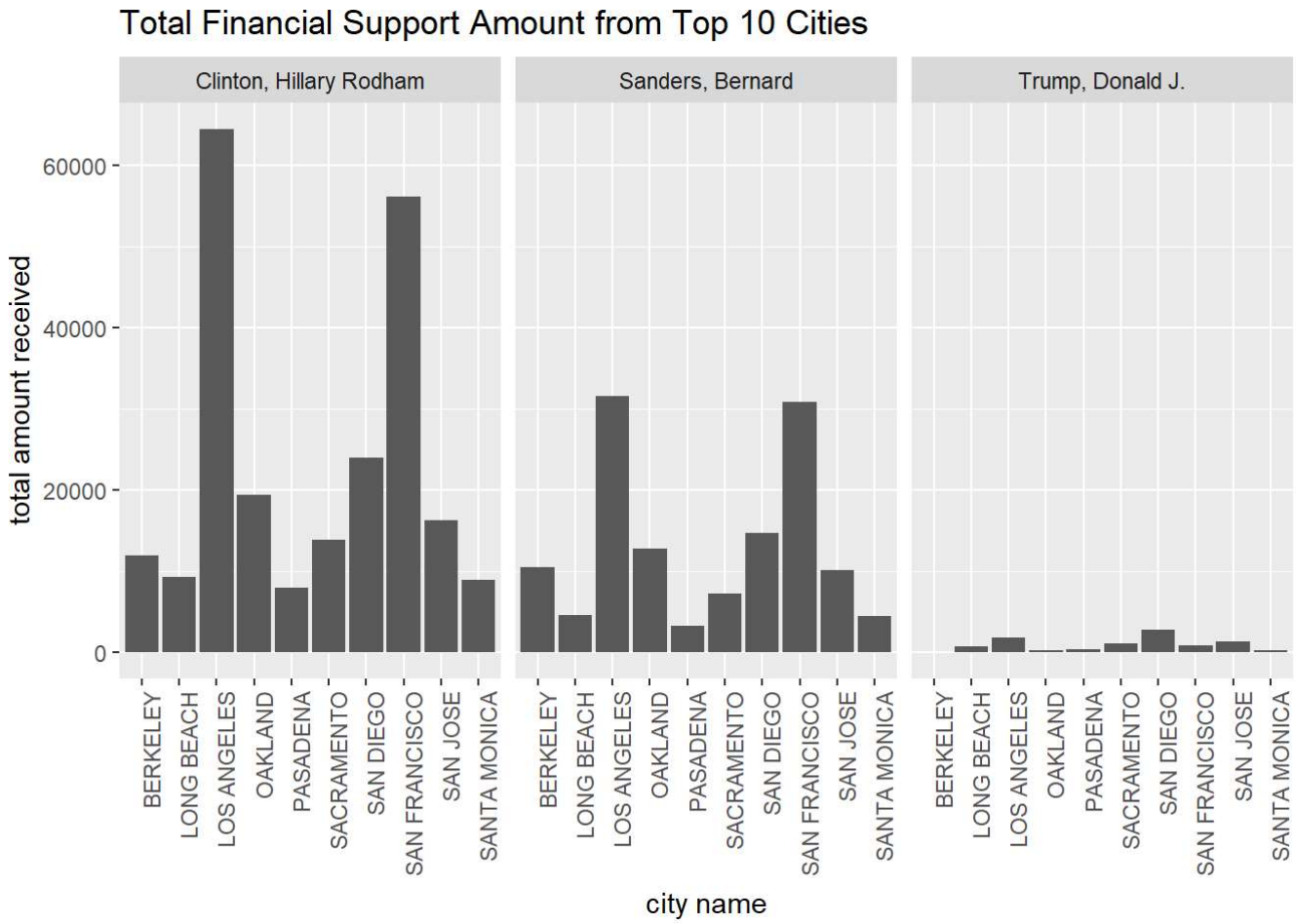

```
## # A tibble: 1 x 2
##                   cand_nm total_amt
##                    <fctr>     <dbl>
## 1 Clinton, Hillary Rodham  95187058
```

city & receipt amount

```
## # A tibble: 1 x 2
##   contbr_city total_amt
##        <fctr>     <dbl>
## 1 LOS ANGELES  16986730
```



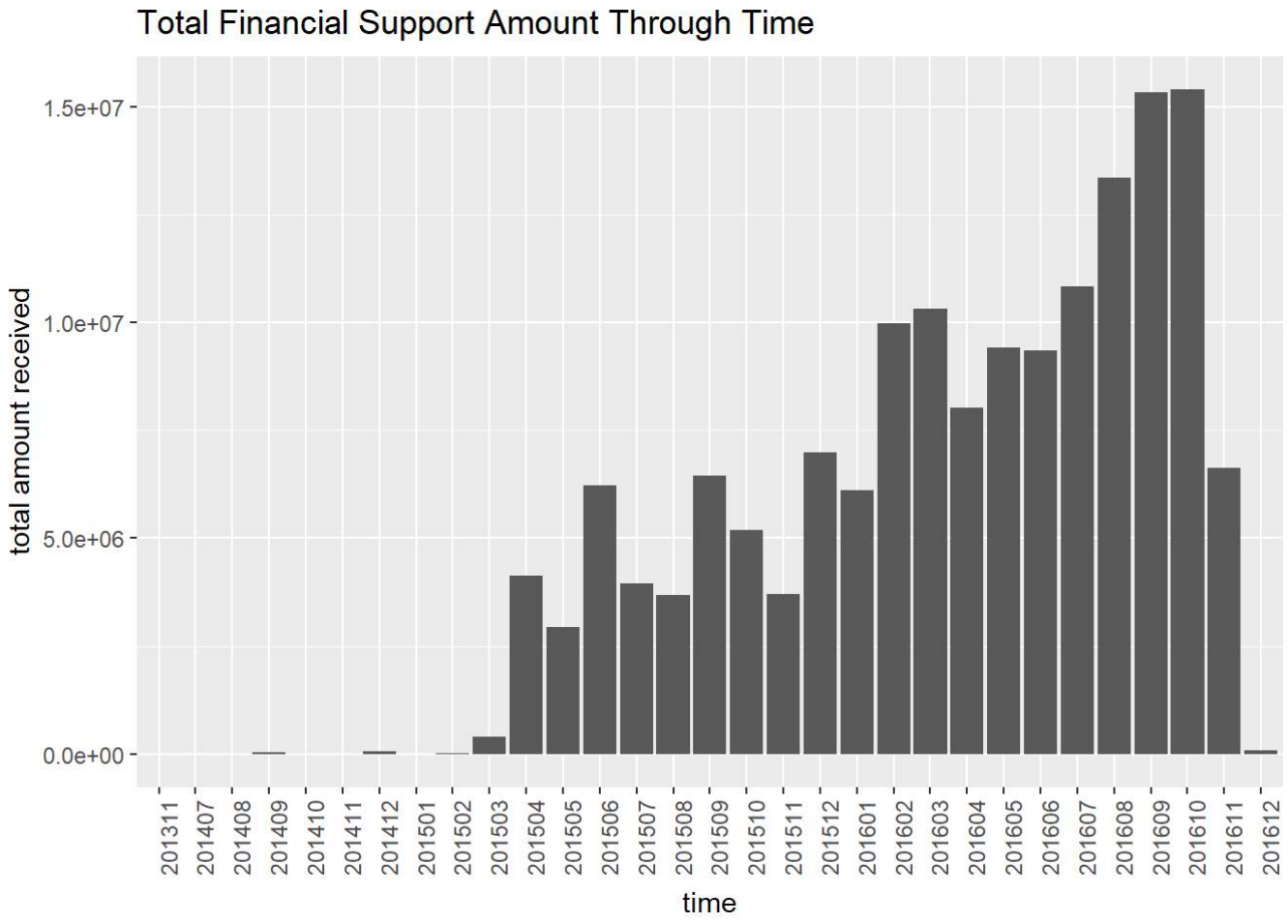

```
# names(financial)
# summary(financial$cand_nm)
top_3_candidates <- c("Clinton, Hillary Rodham", "Sanders, Bernard", "Trump, Donald J.")

# names(top_10_contributing_cities_by_times)
ggplot(aes(x = contbr_city), data = subset(financial,
      contbr_city %in% top_10_contributing_cities_by_times$contbr_city & cand_nm %in% top_3_candidates
 )) + geom_bar() +
  facet_wrap(~cand_nm) +
  labs(x = "city name", y = "total amount received",
      title = "Total Financial Support Amount from Top 10 Cities") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

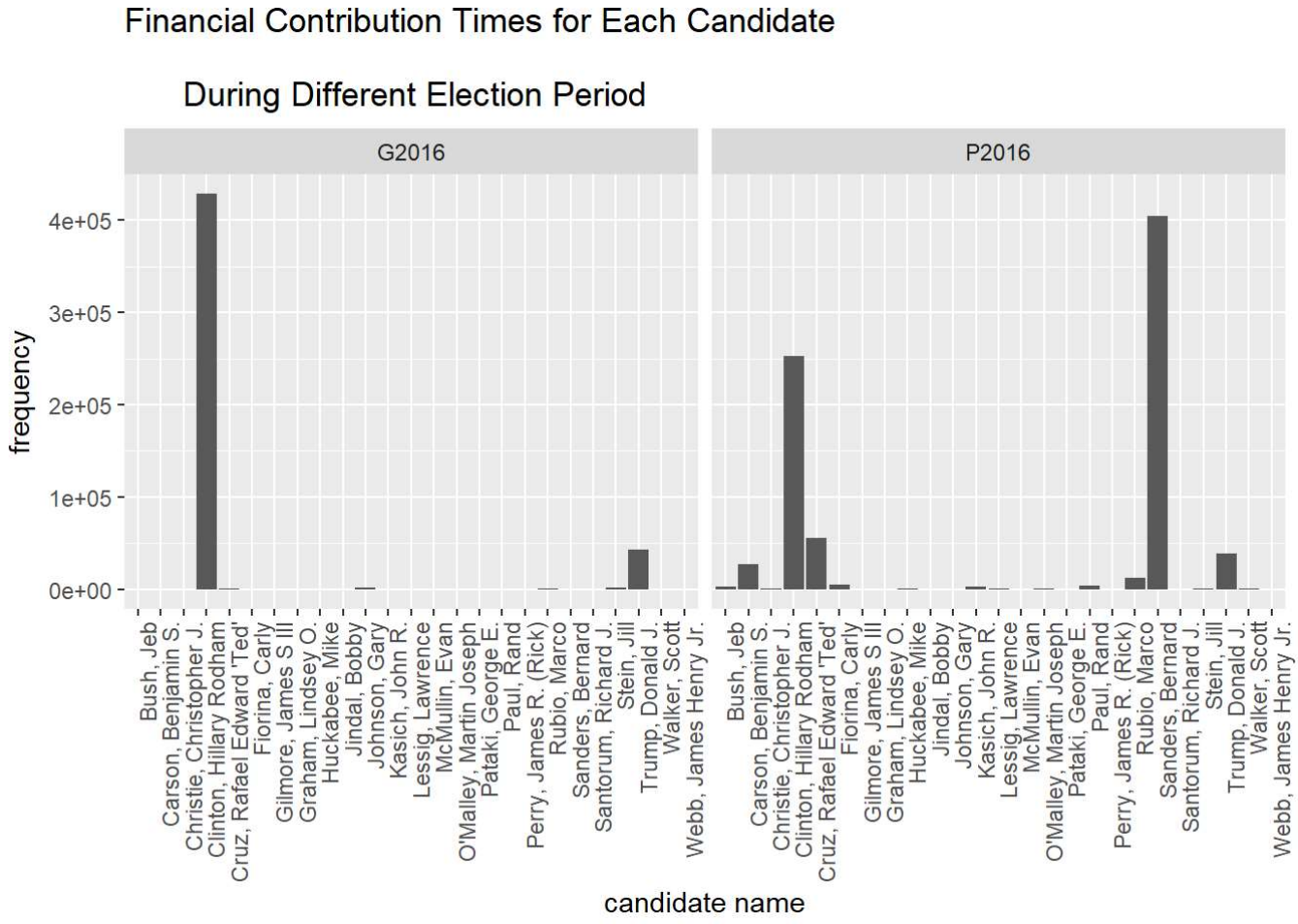



```
# While the big cities are still the big cities
```

contribution time & receipt amount

```
## # A tibble: 1 x 2
##   contb_receipt_dt_yr_mo total_amt
##                    <chr>     <dbl>
## 1                 201610  15395999
```





Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

When doing univariate analysis, I explored the frequency(contribution times) of some variables, like contribution city, candidate name. And when doing bivariate analysis, I primarily changed the frequency(contribution times) to contribution receipt amount. And the most conclusions in univariate analysis didn't vary much, but the distribution changed a little. Like

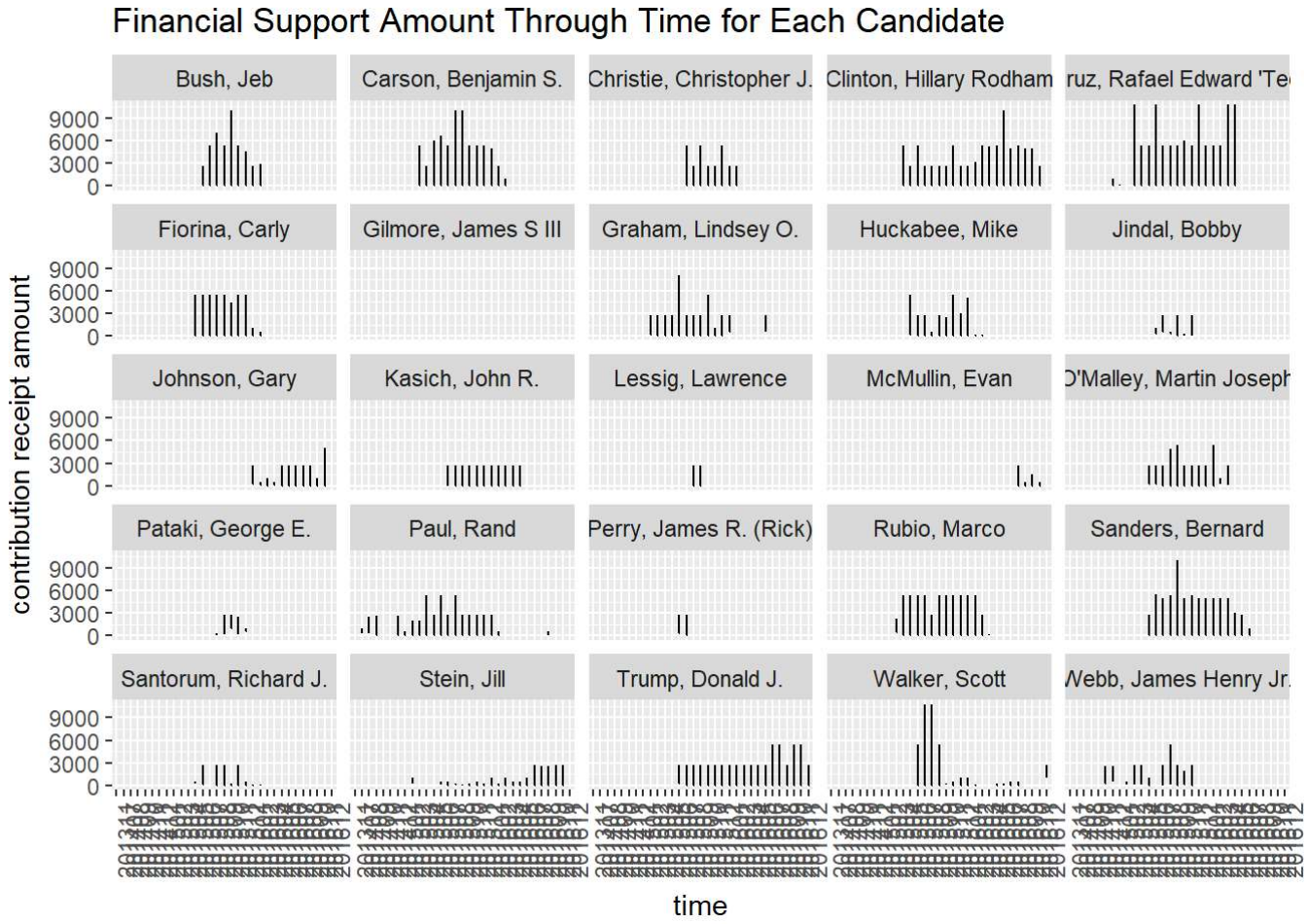
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

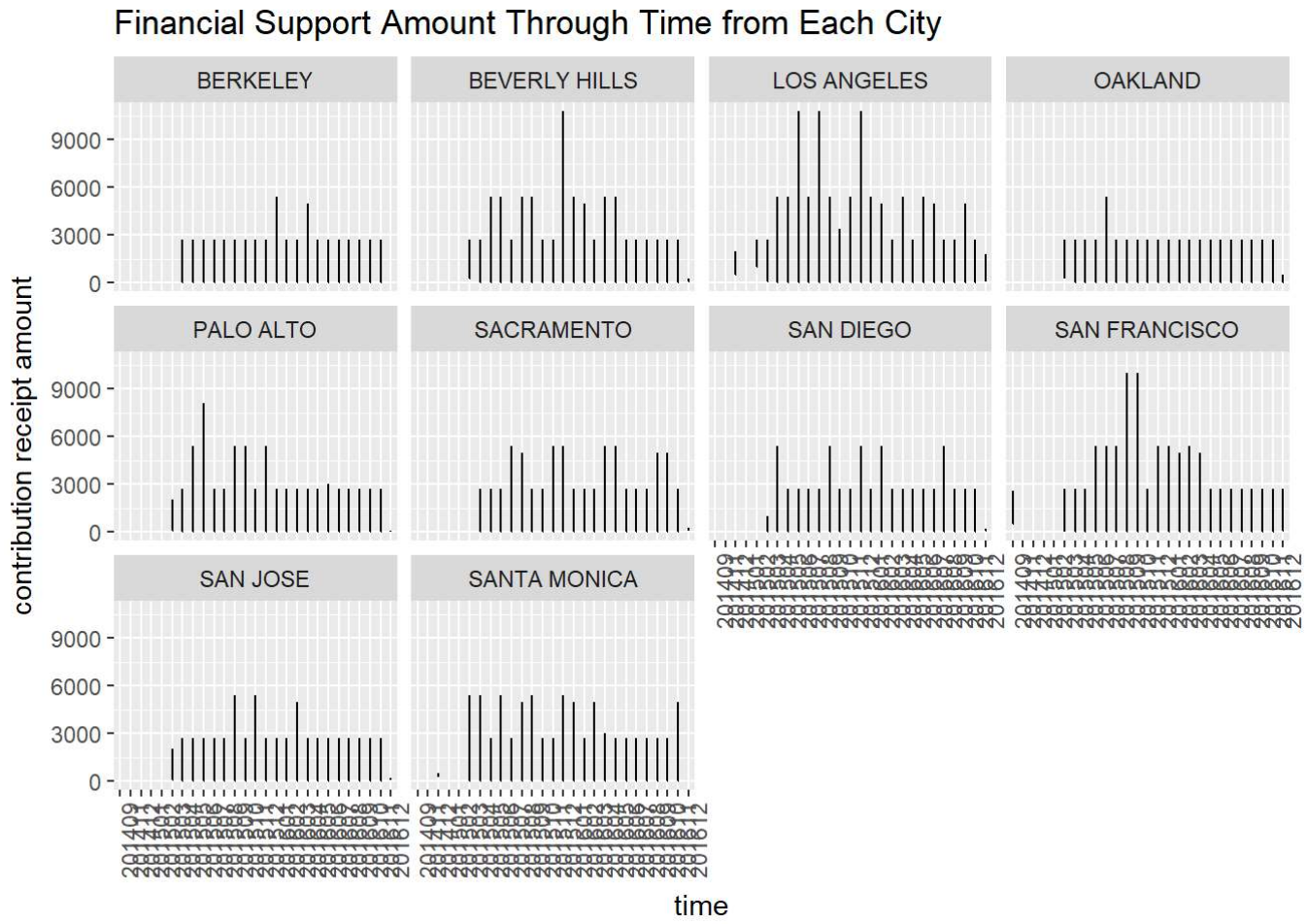
What was the strongest relationship you found?

Since there were only one numerical variable(i.e., contribution receipt amount), so scatterplot was not suitable here. Thus no linear or non-linear relationship were found in this dataset. But We could still see that the contribution receipt amount varied as the candidate or contributing city changed.

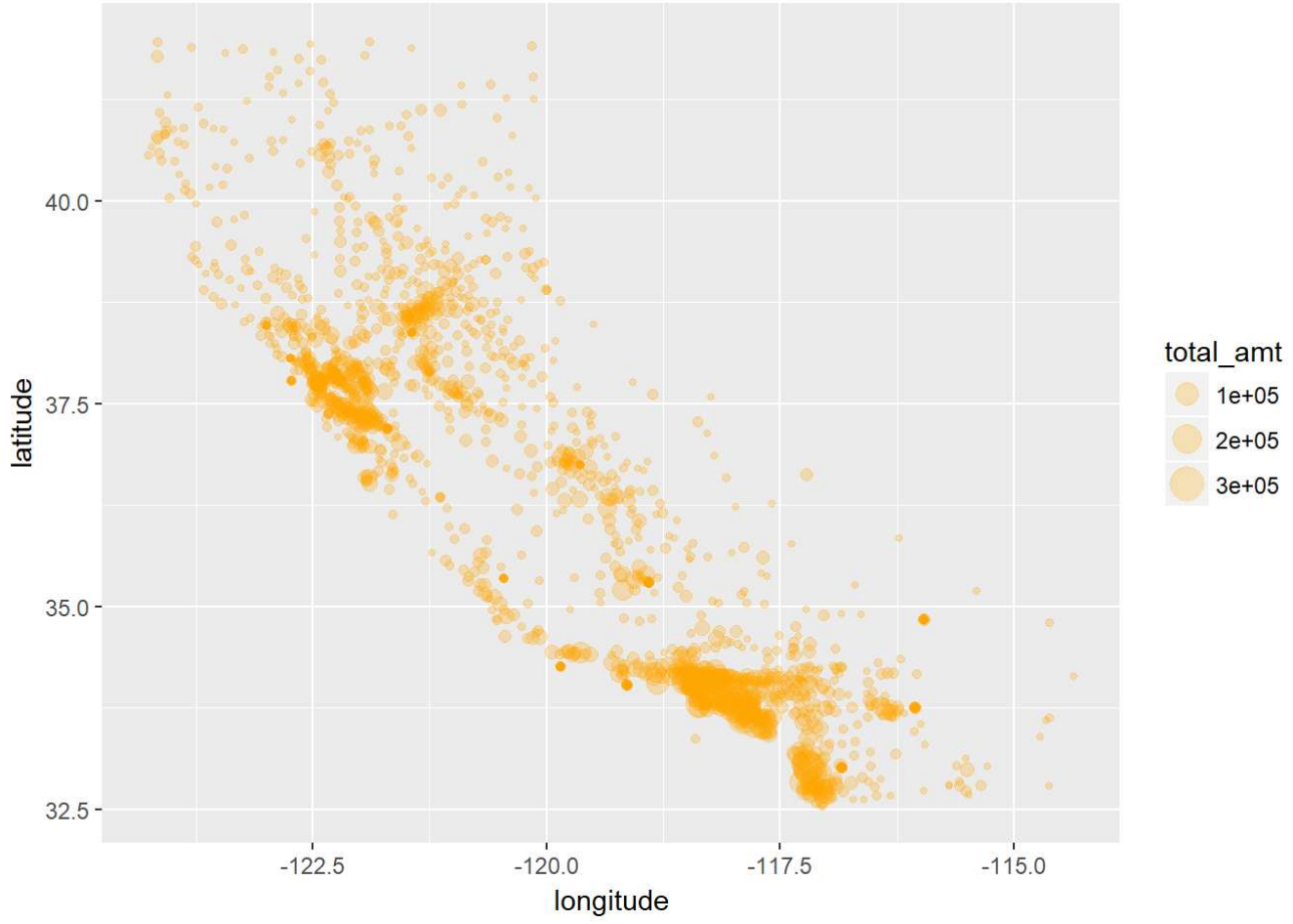
Multivariate Plots Section

> **Tip**: Now it's time to put everything together. Based on what you found in the bivariate plots section, create a few multivariate plots to investigate more complex interactions between variables. Make sure that the plots that you create here are justified by the plots you explored in the previous section. If you plan on creating any mathematical models, this is the section where you will do that.


Financial Support Amount Through Time for Each Candidate
Bush, Jeb
Carson, Benjamin S.
Christie, Christopher J.
Clinton, Hillary Rodham
ruz, Rafael Edward 'Te
Fiorina, Carly
Gilmore, James S III
Graham, Lindsey O.
Huckabee, Mike
Jindal, Bobby
Johnson, Gary
Kasich, John R.
Lessig, Lawrence
McMullin, Evan
O'Malley, Martin Joseph
Pataki, George E.
Paul, Rand
Perry, James R. (Rick)
Rubio, Marco
Sanders, Bernard
Santorum, Richard J.
Stein, Jill
Trump, Donald J.
Walker, Scott
Webb, James Henry Jr.
9000
6000
3000
0
contribution receipt amount
time



Map distribution Maybe I can group it by candidate, so




```
## Using zoom = 7...
```

Map from URL : http://tile.stamen.com/toner-lite/7/19/47.png

Map from URL : http://tile.stamen.com/toner-lite/7/20/47.png

Map from URL : http://tile.stamen.com/toner-lite/7/21/47.png

Map from URL : http://tile.stamen.com/toner-lite/7/22/47.png

Map from URL : http://tile.stamen.com/toner-lite/7/23/47.png

Map from URL : http://tile.stamen.com/toner-lite/7/19/48.png

Map from URL : http://tile.stamen.com/toner-lite/7/20/48.png

Map from URL : http://tile.stamen.com/toner-lite/7/21/48.png

Map from URL : http://tile.stamen.com/toner-lite/7/22/48.png

Map from URL : http://tile.stamen.com/toner-lite/7/23/48.png

Map from URL : http://tile.stamen.com/toner-lite/7/19/49.png

Map from URL : http://tile.stamen.com/toner-lite/7/20/49.png

Map from URL : http://tile.stamen.com/toner-lite/7/21/49.png

Map from URL : http://tile.stamen.com/toner-lite/7/22/49.png

Map from URL : http://tile.stamen.com/toner-lite/7/23/49.png

Map from URL : http://tile.stamen.com/toner-lite/7/19/50.png

Map from URL : http://tile.stamen.com/toner-lite/7/20/50.png

Map from URL : http://tile.stamen.com/toner-lite/7/21/50.png

Map from URL : http://tile.stamen.com/toner-lite/7/22/50.png

Map from URL : http://tile.stamen.com/toner-lite/7/23/50.png

Map from URL : http://tile.stamen.com/toner-lite/7/19/51.png

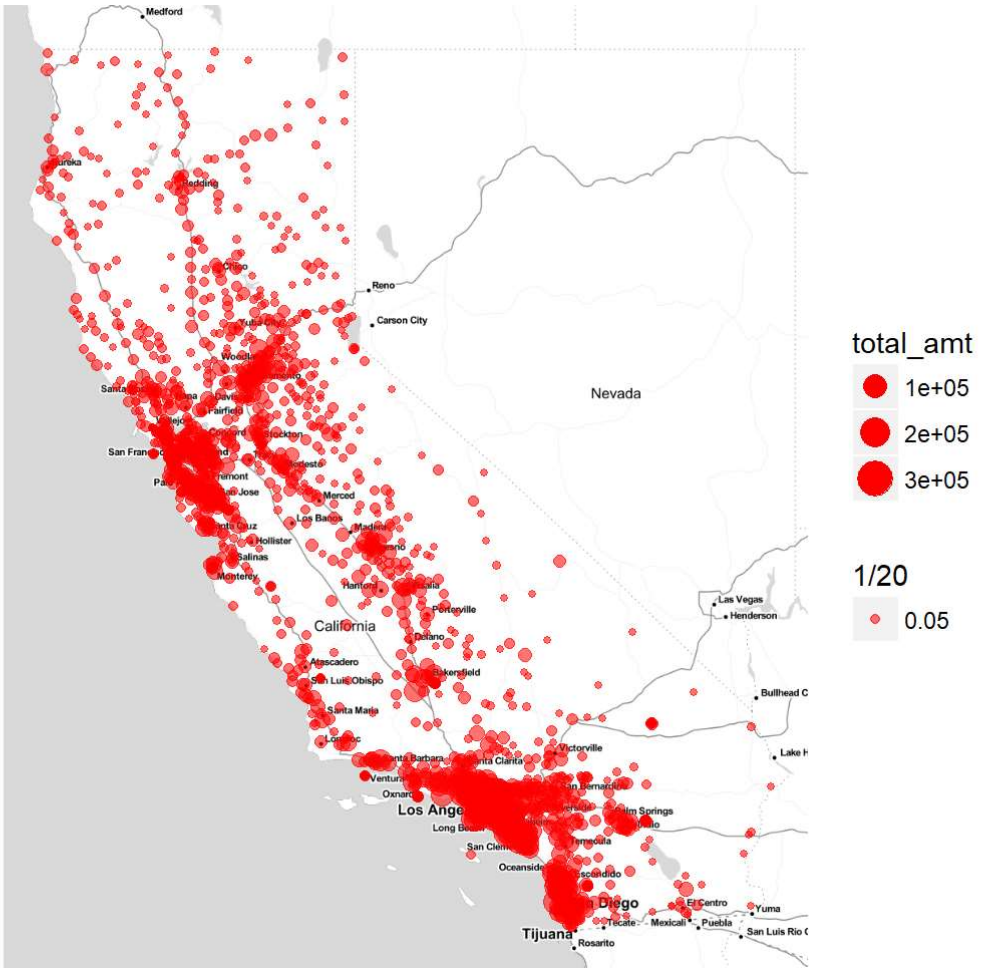
```
## Map from URL : http://tile.stamen.com/toner-lite/7/20/51.png
```

```
## Map from URL : http://tile.stamen.com/toner-lite/7/21/51.png
```

```
## Map from URL : http://tile.stamen.com/toner-lite/7/22/51.png
```

```
## Map from URL : http://tile.stamen.com/toner-lite/7/23/51.png
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```



general election prediction

```
#names(financial)
financial_for_general_election <- subset(financial,
          election_tp == 'G2016' & contb_receipt_amt > 0)
financial_for_general_election_sum <- financial_for_general_election %>%
  group_by(cand_nm) %>%
  summarise(total_amt = sum(contb_receipt_amt),
            n = n())
#head(financial_for_general_election_sum)
#financial_for_general_election_sum

financial_for_general_election_sum$amt_ratio <- financial_for_general_election_sum$total_amt / sum(fin
ancial_for_general_election_sum$total_amt)

financial_for_general_election_sum$times_ratio <- financial_for_general_election_sum$n / sum(financial
_for_general_election_sum$n)
# why above didn't caculate right??
428362 / sum(financial_for_general_election_sum$n)
```

```
## [1] 0.8990815
```

```
# financial_for_general_election_sum
# If we use financial support data for simple predicting, we may get the result that Hillary Clinton w
ould win CA, but the actual ratio(61.73%) was much lower than predicted ones. It may be due to bias(be
cause only rich people would contribute to the candidates and normal people were not considred in this
 prediction, non-respondents bias)
```

Multivariate Analysis

Talk about some of the relationships you observed in this part of the
investigation. Were there features that strengthened each other in terms of
looking at your feature(s) of interest?

Were there any interesting or surprising interactions between features?

OPTIONAL: Did you create any models with your dataset? Discuss the
strengths and limitations of your model.

Final Plots and Summary

Tip: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

Plot One

Description One

Plot Two

Description Two

Plot Three

Description Three

Reflection

Tip: Here's the final step! Reflect on the exploration you performed and the insights you found. What were some of the struggles that you went through? What went well? What was surprising? Make sure you include an insight into future work that could be done with the dataset.

Tip: Don't forget to remove this, and the other **Tip** sections before saving your final work and knitting the final report!