DATA 606 Data Project Proposal

Nick Oliver

Data Preparation

```
library(dplyr)
library(tidyr)
library(ggplot2)
# load data
donationsRaw <- read.csv("https://media.githubusercontent.com/media/fivethirtyeight/data/master/science-

#exclude parties that are not rep or dem
# select the only 3 columns we need from the original dataset
donations <- donationsRaw %>%
  filter(cand_pty_affiliation == "REP" | cand_pty_affiliation == "DEM") %>%
  select(cand_pty_affiliation, cleanedoccupation, state) %>%
  rename(party=cand_pty_affiliation, occupation=cleanedoccupation)
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Is your occupation in the field of science & mathematics or the state you live in a better predictor of what political party you would donate to?

For example, am I more confident that an individual would donate to a democratic candidate because they lived in California or because they were a software engineer?

Cases

What are the cases, and how many are there?

There are 380387 observations in this dataset

Data collection

Describe the method of data collection.

The Federal Election Commission of the United States receives this data directly from political campaign committees, candidates, political action committees and other political organizations required by federal law to submit such information.

About campaign finance data

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The data was obtained from Fivethirtyeight's GitHub Here but was originally obtained from the United States Federal Election Commission website

The data was collected between the years of 2006 and 2016. The Fivethirtyeight pre-filtered the data to relevant science and mathematics based occupations.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

What was the political party affiliation of the candidate or organization a person donated to

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

What was the person's state of residence and what was the person's occupation (data set is already pre-filtered to the sciences)

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Number of unique occupations & number of unique states

```
length(unique(donations$occupation))
```

```
## [1] 7932
```

```
length(unique(donations$state))
```

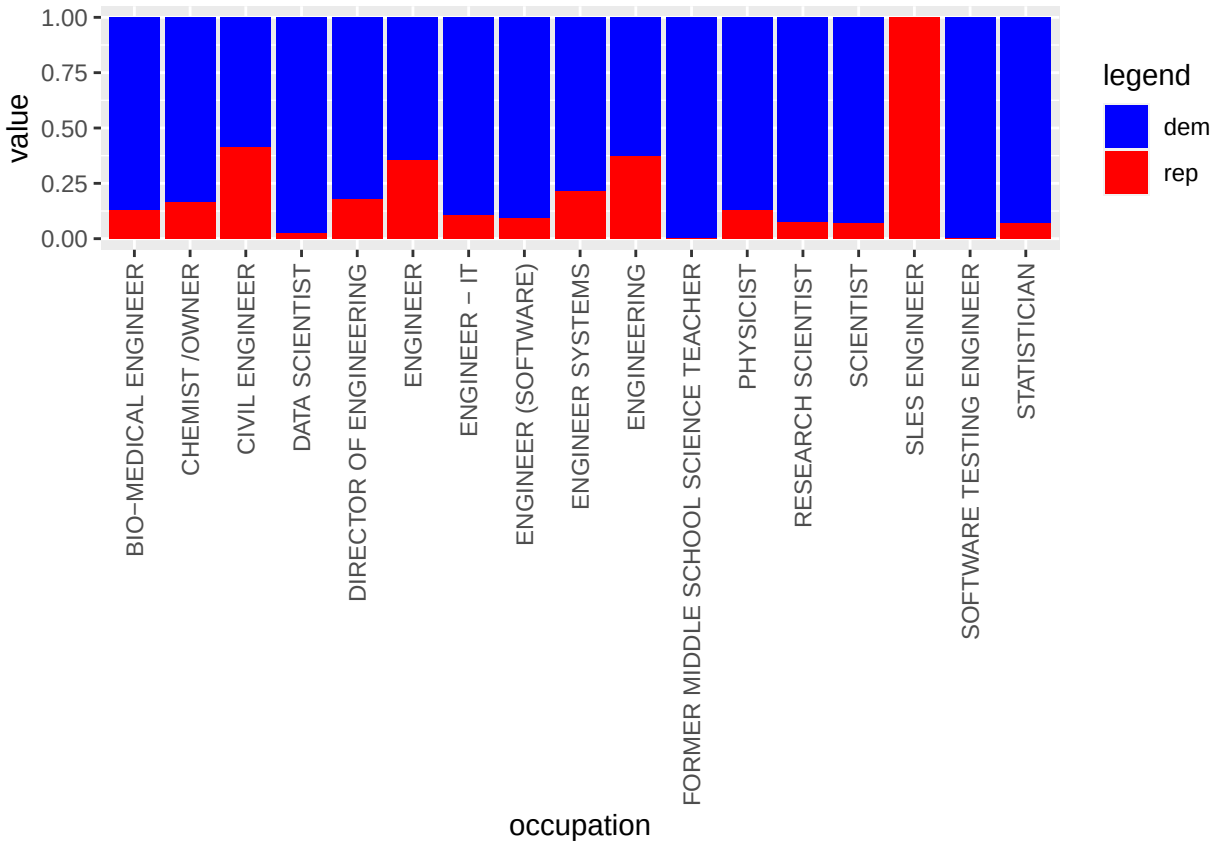
```
## [1] 63
```

There are 7,932 unique occupations in the dataset and 63 states. The fact that there is 63 states means that they likely include different united states territories in the data as well

Lets look at one occupation and see what the breakdown of donations looks like for republicans and democrats

```
set.seed(13371)
occupationDonationSubset <- donations %>%
  group_by(occupation) %>%
  filter(occupation %in% sample(donations$occupation, 25)) %>%
  summarize(
      dem = sum(party == "DEM"),
      rep = sum(party == "REP")
      ) %>%
  pivot_longer(!occupation)
occupationDonationSubset %>%
   ggplot(aes(fill=name, y=value, x=occupation)) +
    geom_bar(position="fill", stat="identity") +
  scale_fill_manual("legend", values = c("dem" = "blue", "rep" = "red")) +

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

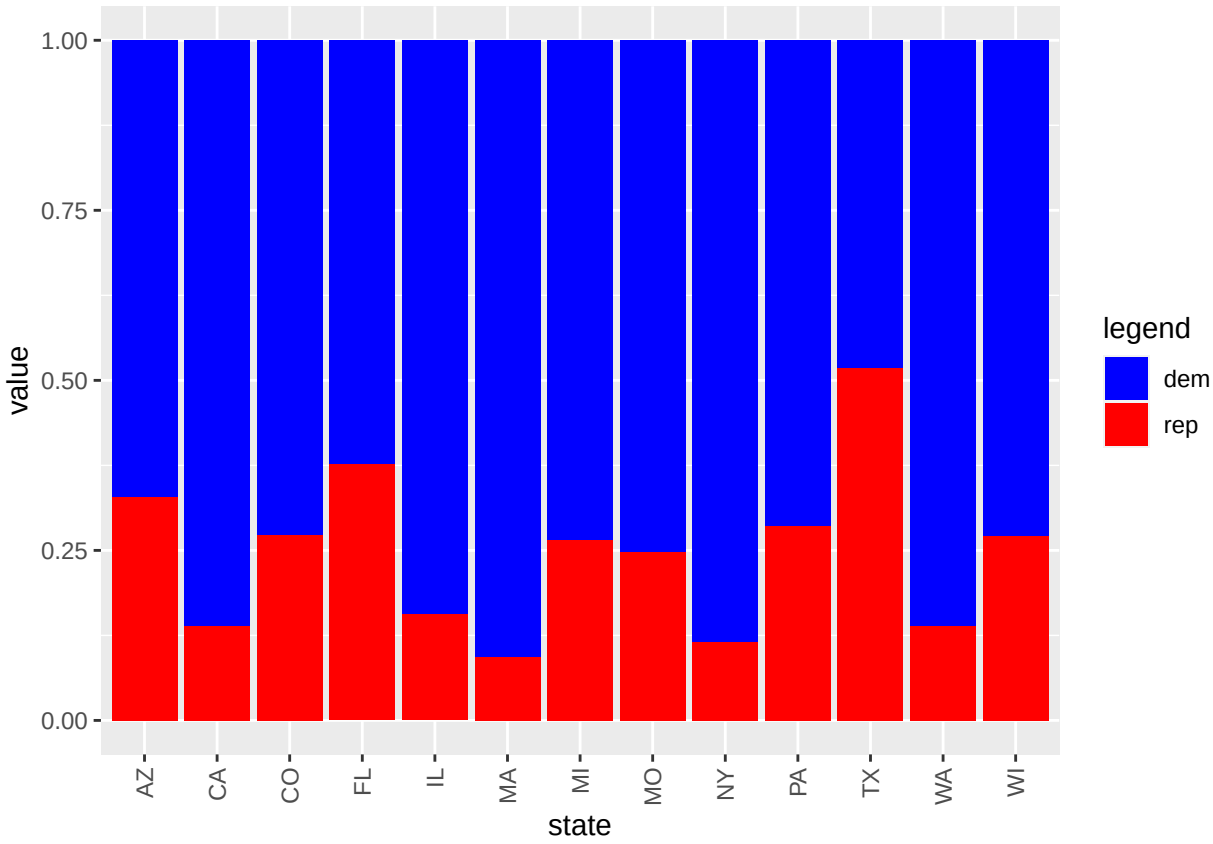


Using the bar chart we can see that there is a decent mix of different preferences per occupation

Doing the same for states

```
stateDonationSubset <- donations %>%
  group_by(state) %>%
  filter(state %in% sample(donations$state, 25)) %>%
  summarize(
      dem = sum(party == "DEM"),
      rep = sum(party == "REP")
      ) %>%
  pivot_longer(!state)
stateDonationSubset %>%
   ggplot(aes(fill=name, y=value, x=state)) +
    geom_bar(position="fill", stat="identity") +
  scale_fill_manual("legend", values = c("dem" = "blue", "rep" = "red")) +

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Summarizing how many data points we have per occupation yields interesting results.

There appears to be a very large spread across the data with the mean being just 47.96 but the standard deviation being 1,602. The occupation with the largest amount of data is `ENGINEER` with 125,699 observations in the data set.

This is not surprising as there is such a large amount of occupations in the data set but engineer is likely a generic title than many people fall under.

```
occupationDonationCount <- donations %>%
  group_by(occupation) %>%
  summarize(count = n())

summary(occupationDonationCount$count)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      1.00      1.00      2.00     47.96      5.00 125699.00
```

```
print(paste("Standard Dev: ", sd(occupationDonationCount$count)))
```

```
## [1] "Standard Dev:  1602.45654524643"
```

```
occupationDonationCount %>% slice_max(count)
```

```
## # A tibble: 1 x 2
##   occupation  count
##   <chr>       <int>
## 1 ENGINEER   125699
```

Doing the same for states shows that the similarly large spread with a a much larger standard deviation given the smaller number of states.

This is likely due to the inclusion of some very small by population territories. I used `slice_min` to find that the Northern Mariana Islands only have a single data point. With a population of around 57,000 it is not surprising there is almost no data for this area.

Because of this for my project I will likely be limiting my analysis to a number of larger states which would have more data.

```
stateDonationCount <- donations %>%
  group_by(state) %>%
  summarize(count = n())

summary(stateDonationCount$count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     904    2294    6038    6800   81446
```

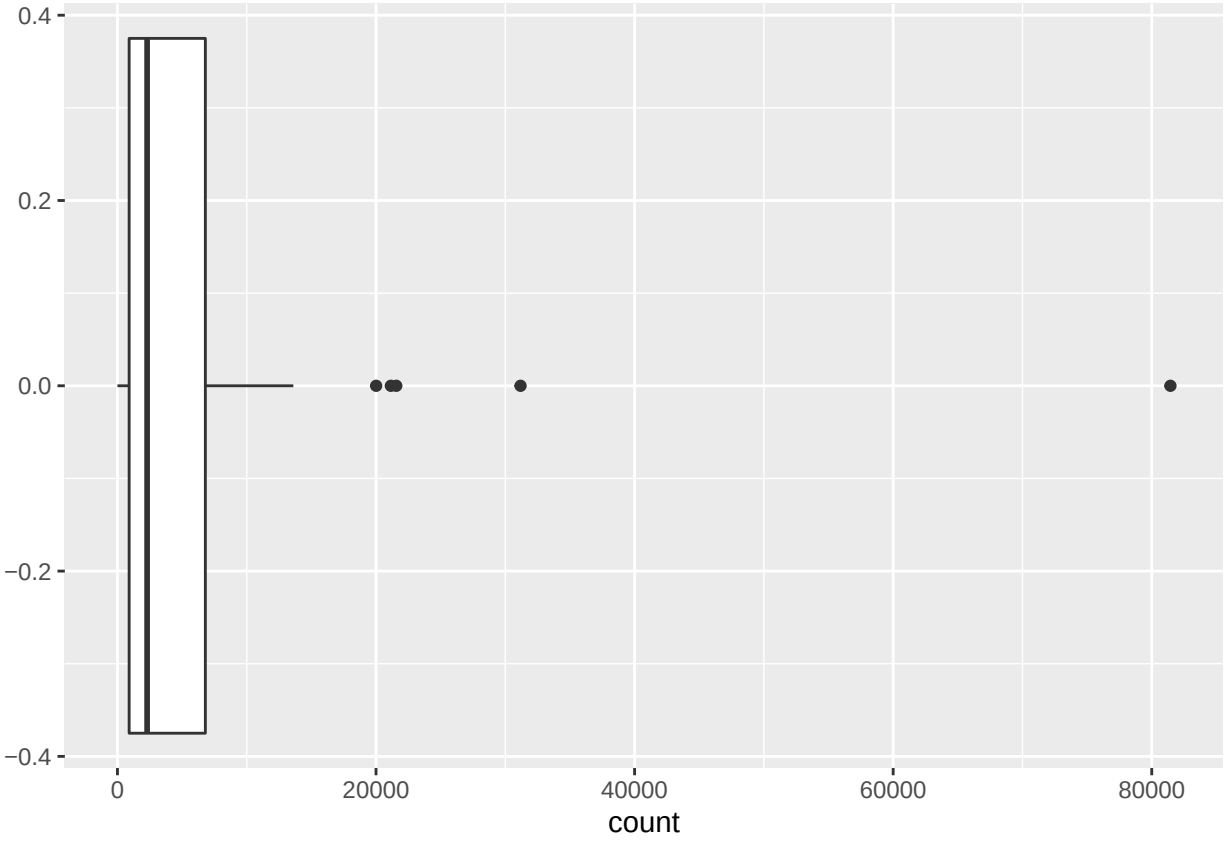
```
print(paste("Standard Dev: ", sd(stateDonationCount$count)))
```

```
## [1] "Standard Dev:  11523.5151406649"
```

```
stateDonationCount %>% slice_min(count)
```

```
## # A tibble: 1 x 2
##   state count
##   <chr> <int>
## 1 MP        1
```

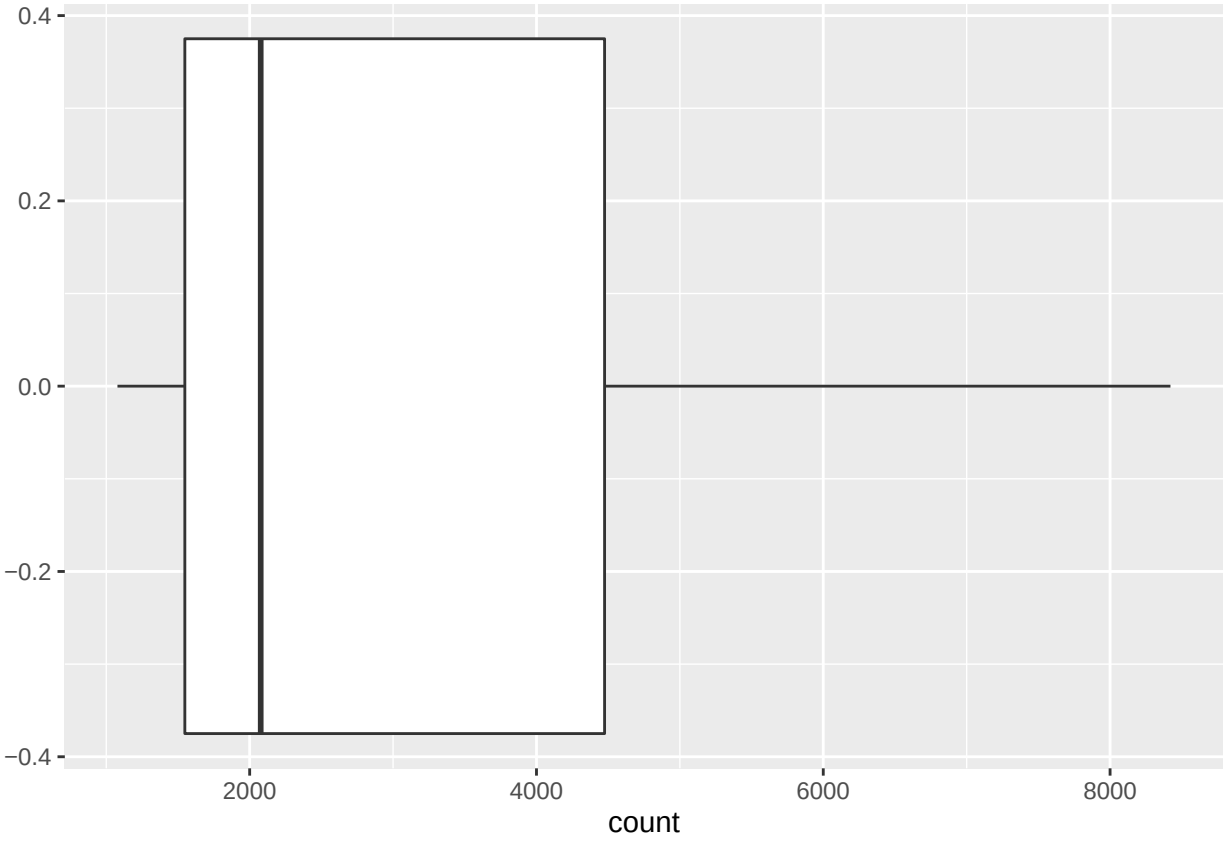
```
stateDonationCount %>%
  ggplot(aes(y = count)) +
  geom_boxplot() + coord_flip()
```



Narrowing the data down to just occupations and states that have between 1,000 and 10,000 observations I get much more reasonable data sets to analyze against.

With 31 of the 7,932 occupations falling in this range and 25 of the 63 states falling in this range.

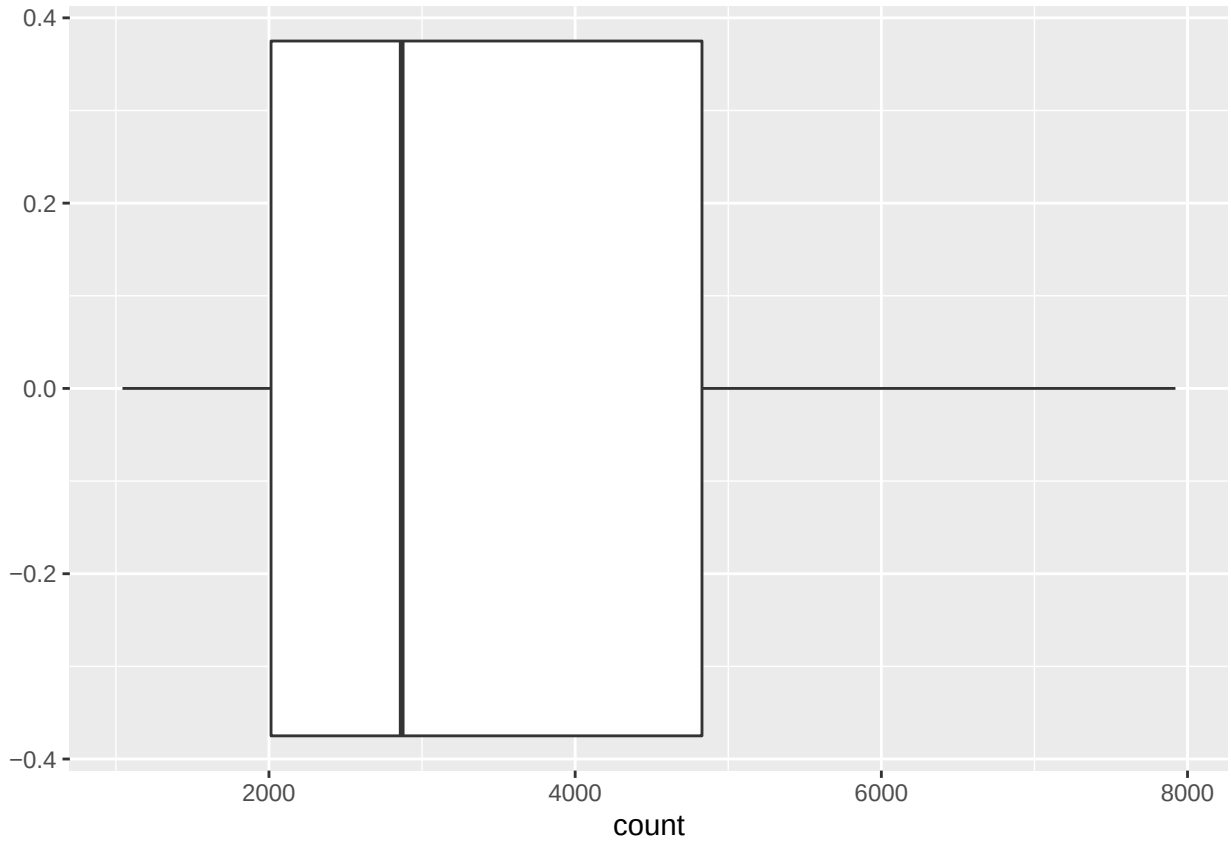
```
occupationDonationCount %>%
  filter(count > 1000 & count < 10000) %>%
    ggplot(aes(y = count)) +
  geom_boxplot() + coord_flip()
```




```
nrow(occupationDonationCount %>%
  filter(count > 1000 & count < 10000))
```

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

```
stateDonationCount %>%
  filter(count > 1000 & count < 10000) %>%
    ggplot(aes(y = count)) +
  geom_boxplot() + coord_flip()
```




```
nrow(stateDonationCount %>%
  filter(count > 1000 & count < 10000))
```

```
## [1] 31
```