Politics Are Afoot!

w203: Statistics for Data Science

The Setup

There is *a lot* of money that is spent in politics in Presidential election years. Like, a lot, a lot. Estimates and analysis from the US Federal Election Comission, puts the total amount at about $14,400,000,000 ($14.4 billion USD). For context, Twitter's 2020 annual revenue was about $3,500,000,000 ($3.5 billion USD).

The work

Install the package, `fec16`.

```
## install.packages('fec16')
```

This package is a compendium of spending and results from the 2016 election cycle. In this dataset are 9 different datasets that cover:

- `candidates`: candidate attributes, like their name, a unique id of the candidate, the election year under consideration, the office they're running for, etc.
- `results_house`: race attributes, like the name of the candidates running in the election, a unique id of the candidate, the number of `general_votes` garnered by each candidate, and other information.
- `campaigns`: financial information for each house & senate campaign. This includes a unique candidate id, the total receipts (how much came in the doors), and total disbursements (the total spent by the campaign), the total contributed by party central committees, and other information.

Your task

Your task is to describe the relationship between spending on a candidate's behalf and the votes they receive.

If it is helpful to structure your response, you might want to place yourself into a scenario where you are advising a person or business about whether they should make a political donation. While the benefits that accrue as a result of a successful investment are unclear, you can be quite sure that investing with **no** return (i.e. more spending does not increase the chances of winning) is a bad idea.

Your work

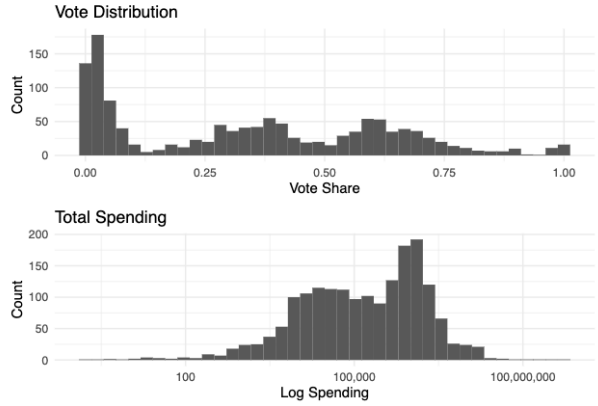
- We want to keep this work *relatively* constrained, which is why we're providing you with data through the `fec16` package. It is possible to gather all the information from current FEC reports, but it would require you to make a series of API calls that would pull us away from the core modeling tasks that we want you to focus on instead.
- Throughout this assignment, limit yourself to functions that are within the `tidyverse` family of packages: `dplyr`, `ggplot`, `patchwork`, and `magrittr` for wrangling and exploration and `base`, `stats`, `sandwich` and `lmtest` for modeling and testing. You do not *have* to use these packages; but try to limit yourself to using only these.
- Our choice to encourage you to use only these packages is to try to cut down on the amount of searching that you do: to help you avoid looking for the *"one package that does the thing I need it to do."* Certainly,

such a package exists, but it will very likely be more productive for you to write things yourself than to try and find it for this homework.

```
candidates    <- fec16::candidates
results_house <- fec16::results_house
campaigns     <- fec16::campaigns
```

1. What does the distribution of votes and of spending look like?

1. (3 points) In separate histograms, show both the distribution of votes (measured in `results_house$general_percent` for now) and spending (measured in `ttl_disb`). Use a log transform if appropriate for each visualization. How would you describe what you see in these two plots?



Answer: The first plot shows the vote share based on percentage. It it not normally distributed with a large spike around 0 (Under achieving candidates) and a heavy tail around 100% for those who dominated the race. The second plot looks at the total spending of campaigns transformed using the log function and it is more normal in shape but still with a heavy tails.

2. Exploring the relationship between spending and votes.

2. (3 points) Create a new dataframe by joining `results_house` and `campaigns` using the `inner_join` function from `dplyr`. (We use the format `package::function` – so `dplyr::inner_join`.) Does this data frame contain all the data that was present in the two frames that you're joining together, or has some data been dropped? As you're manipulating data, keep a keen eye for what is, and what is not making it through your data → analysis → reporting pipeline.

```
d <- inner_join(results_house, campaigns, by ='cand_id')
d
```

```
## # A tibble: 1,342 x 37
##    state district_id cand_id   incumbent party primary_votes primary_percent
##    <chr> <chr>       <chr>     <lgl>     <chr>         <dbl>           <dbl>
##  1 AL    01          H4AL01123 TRUE      REP           71310          0.601
##  2 AL    01          H6AL01060 FALSE     REP           47319          0.399
##  3 AL    02          H0AL02087 TRUE      REP           78689          0.664
##  4 AL    02          H6AL02142 FALSE     REP           33015          0.278
##  5 AL    02          H6AL02159 FALSE     REP            6856          0.0578
##  6 AL    02          H6AL02167 FALSE     DEM              NA         NA
##  7 AL    03          H2AL03032 TRUE      REP           77432          0.760
##  8 AL    03          H6AL03157 FALSE     REP           24474          0.240
##  9 AL    03          H4AL03061 FALSE     DEM              NA         NA
## 10 AL    04          H6AL04098 TRUE      REP           86660          0.812
```

```
## # ... with 1,332 more rows, and 30 more variables: runoff_votes <dbl>,
## #   runoff_percent <dbl>, general_votes <dbl>, general_percent <dbl>,
## #   won <lgl>, footnotes <chr>, cand_name <chr>, cand_ici <chr>, pty_cd <dbl>,
## #   cand_pty_affiliation <chr>, ttl_receipts <dbl>, trans_from_auth <dbl>,
## #   ttl_disb <dbl>, trans_to_auth <dbl>, coh_bop <dbl>, coh_cop <dbl>,
## #   cand_contrib <dbl>, cand_loans <dbl>, other_loans <dbl>,
## #   cand_loan_repay <dbl>, other_loan_repay <dbl>, debts_owed_by <dbl>, ...
```

```
campaigns
```

```
## # A tibble: 1,898 x 25
##    cand_id   cand_name             cand_ici pty_cd cand_pty_affili~ ttl_receipts
##    <chr>     <chr>                 <chr>     <dbl> <chr>                   <dbl>
##  1 H6AK00045 YOUNG, DONALD E       I             2 REP                   1103562.
##  2 H6AK00235 LINDBECK, STEVE       <NA>          1 DEM                   1102310.
##  3 H4AL01123 BYRNE, BRADLEY ROBER~ I             2 REP                   1367470.
##  4 H6AL01060 YOUNG JR, LARRY DEAN  C             2 REP                    178767.
##  5 H6AL02167 MATHIS, NATHAN        C             1 DEM                     36844
##  6 HOAL02087 ROBY, MARTHA          I             2 REP                   1404260.
##  7 H6AL02142 GERRITSON, REBECCA (~ C             2 REP                    206908.
##  8 H6AL02159 ROGERS, ROBERT L      C             2 REP                     25382
##  9 H4AL03061 SMITH, JESSE TREMAIN  C             1 DEM                      9810
## 10 H2AL03032 ROGERS, MICHAEL DENN~ I             2 REP                   1139022.
## # ... with 1,888 more rows, and 19 more variables: trans_from_auth <dbl>,
## #   ttl_disb <dbl>, trans_to_auth <dbl>, coh_bop <dbl>, coh_cop <dbl>,
## #   cand_contrib <dbl>, cand_loans <dbl>, other_loans <dbl>,
## #   cand_loan_repay <dbl>, other_loan_repay <dbl>, debts_owed_by <dbl>,
## #   ttl_indiv_contrib <dbl>, cand_office_st <chr>, cand_office_district <chr>,
## #   other_pol_cmte_contrib <dbl>, pol_pty_contrib <dbl>, cvg_end_dt <date>,
## #   indiv_refunds <dbl>, cmte_refunds <dbl>
```

```
results_house
```

```
## # A tibble: 2,110 x 13
##    state district_id cand_id   incumbent party primary_votes primary_percent
##    <chr> <chr>       <chr>     <lgl>     <chr>         <dbl>           <dbl>
##  1 AL    01          H4AL01123 TRUE      REP           71310           0.601
##  2 AL    01          H6AL01060 FALSE     REP           47319           0.399
##  3 AL    02          HOAL02087 TRUE      REP           78689           0.664
##  4 AL    02          H6AL02142 FALSE     REP           33015           0.278
##  5 AL    02          H6AL02159 FALSE     REP            6856           0.0578
##  6 AL    02          H6AL02167 FALSE     DEM              NA          NA
##  7 AL    03          H2AL03032 TRUE      REP           77432           0.760
##  8 AL    03          H6AL03157 FALSE     REP           24474           0.240
##  9 AL    03          H4AL03061 FALSE     DEM              NA          NA
## 10 AL    04          H6AL04098 TRUE      REP           86660           0.812
## # ... with 2,100 more rows, and 6 more variables: runoff_votes <dbl>,
## #   runoff_percent <dbl>, general_votes <dbl>, general_percent <dbl>,
## #   won <lgl>, footnotes <chr>
```

Answer: These two data sets have a different number of rows: campaigns has 1898 and results_house has 2110. The join has resulted in 1342 rows meaning there is some data that has been lost. However there is still a rather large sample size to work with.

3. (3 points) Produce a scatter plot of `general_votes` on the y-axis and `ttl_disb` on the x-axis. What do you observe about the shape of the joint distribution?

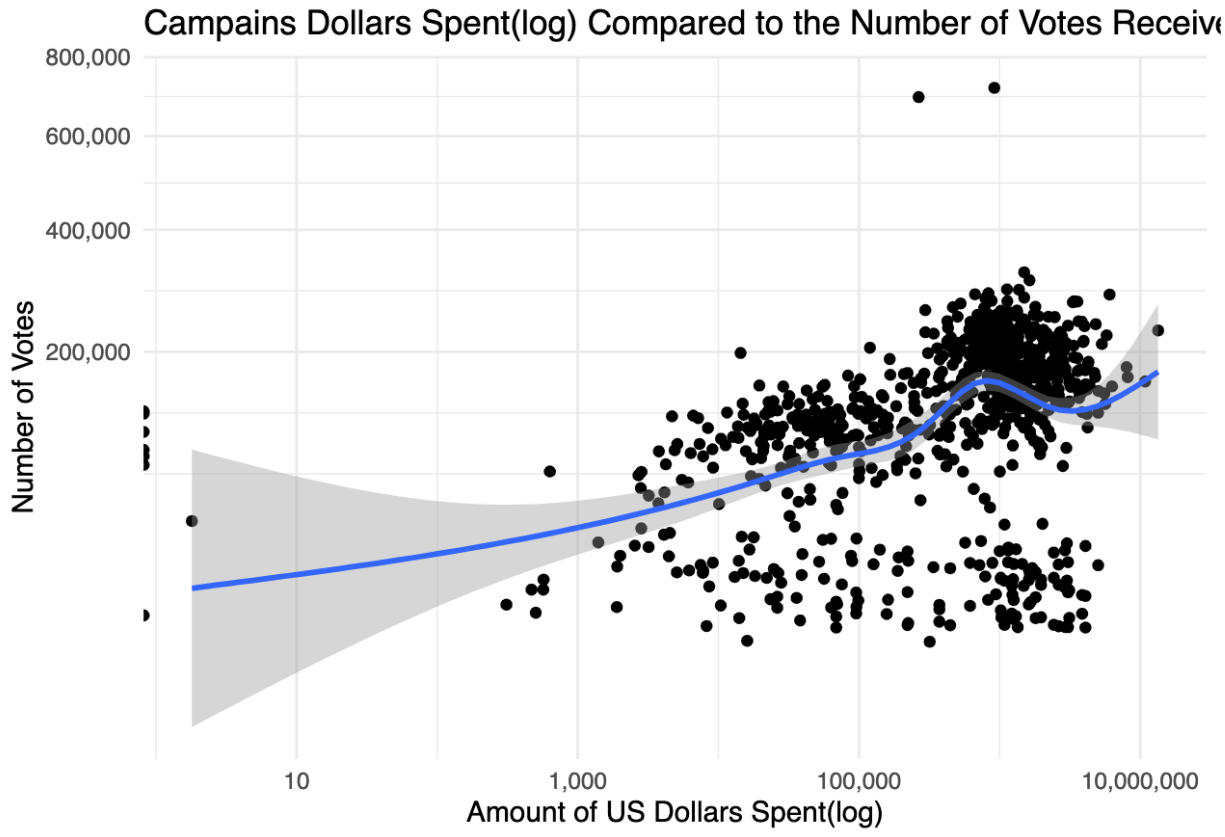
```
p1 <- ggplot(d, aes(x = ttl_disb, y = general_votes)) +
    geom_point() +
    scale_x_log10(labels = comma) +
    scale_y_sqrt(labels = comma) +
    geom_smooth() +
    labs(title = "Campains Dollars Spent(log) Compared to the Number of Votes Received", x = "Amount of
p1
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
## Transformation introduced infinite values in continuous x-axis
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 469 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 462 rows containing missing values (geom_point).
```



Answer: The amount of money spent is relaltively tightly clustered with a few outliers. Some of these points seem to be less than zero this could be do to campaign funds being returned. We can see that there is a point of diminishing returns where more money spent yeilds less votes.

4. (3 points) Create a new variable to indicate whether each individual is a "Democrat", "Republican" or "Other Party".

- Here's an example of how you might use `mutate` and `case_when` together to create a variable.

```
## MIDS 203 students: you can remove this demo code when you knit.
starwars %>%
  select(name:mass, gender, species) %>%
```

```
  mutate(
  type = case_when(
    height > 200 | mass > 200 ~ "large",
    species == "Droid"        ~ "robot",
    TRUE                      ~ "other"
    )
  )
```

Once you've produced the new variable, plot your scatter plot again, but this time adding an argument into the `aes()` function that colors the points by party membership. What do you observe about the distribution of all three variables?

```
party <- d %>%
  mutate(
  party_membership = case_when(
    cand_pty_affiliation == 'REP' ~ 'Republican',
    cand_pty_affiliation == 'DEM' ~ 'Democrat',
    TRUE ~ 'Other Party'
  ))

p2 <- ggplot(party, aes(x = ttl_disb, y = general_votes))+
    geom_point(aes(color=party_membership))+
    scale_x_log10(labels = comma) +
    scale_y_sqrt(labels = comma) +
    geom_smooth()
    labs(title = "Campains Dollars Spent Compared to the # of Votes Received", x = "Amount of US Dollar
```

```
## $x
## [1] "Amount of US Dollars Spent"
##
## $y
## [1] "# of Votes"
##
## $title
## [1] "Campains Dollars Spent Compared to the # of Votes Received"
##
## attr(,"class")
## [1] "labels"
```

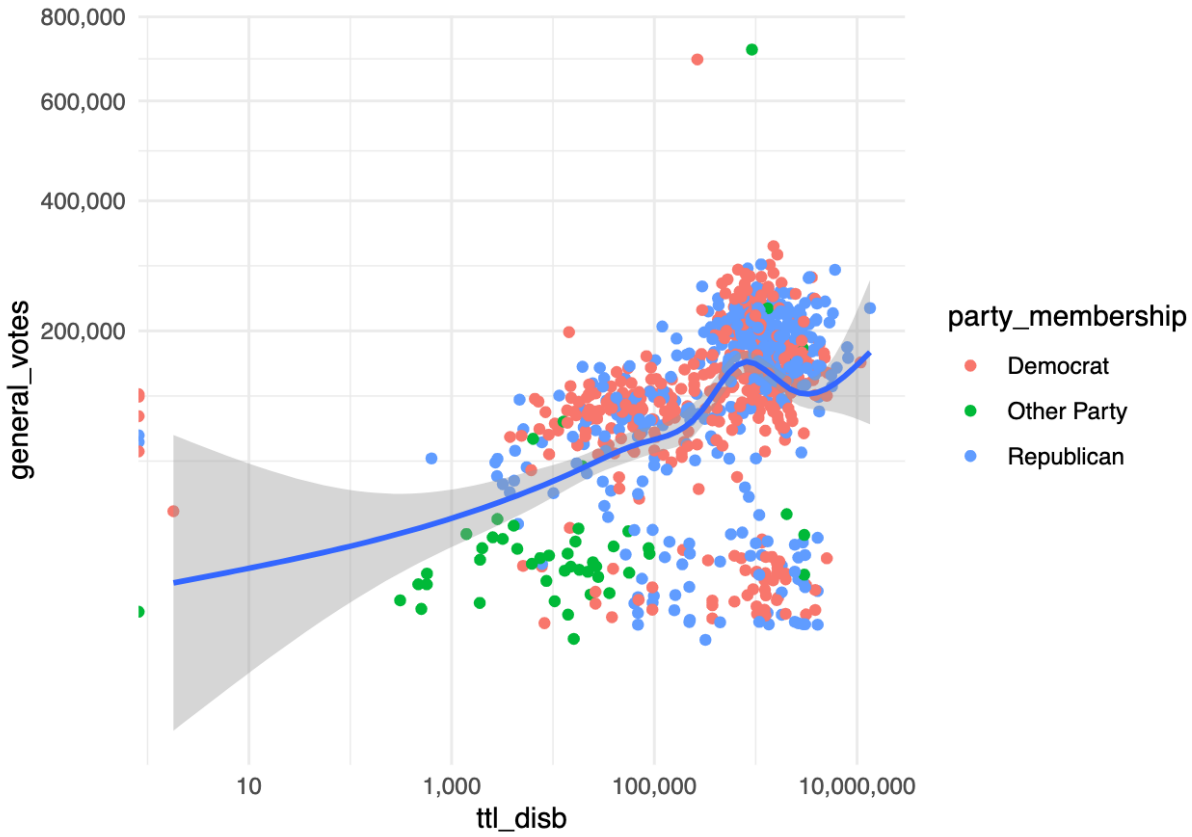
```
p2
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
## Transformation introduced infinite values in continuous x-axis
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 469 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 462 rows containing missing values (geom_point).
```



Answer: This plot provides more context to the relationship of amount of spend as it relates to number of votes by party affiliation. The two major parties (Democrats & Republicans) make up the majority of spend and receive a higher return of votes. However those labeled as "Other Party" can still spend a large amount but get very little number of votes in return with the exceptional outlier such as a Bernie Sanders type of politician.

Produce a Descriptive Model

For this section, rather than us providing you with `'fill in: '` prompts, you can write in whatever way is most effective for you. Please, limit this section to no more than three printed pages. (Your client – aka the TAs – have a finite attention span!)

5. (5 Points) Given your observations, produce a linear model that you think does a good job at describing the relationship between candidate spending and votes they receive. You should decide what transformation to apply to spending (if any), what transformation to apply to votes (if any) and also how to include the party affiliation.

```
model1 <- lm(general_votes ~ ttl_disb + I(party_membership), data=party)
summary(model1)
```

```
##
## Call:
## lm(formula = general_votes ~ ttl_disb + I(party_membership),
##     data = party)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -179027  -37565    -442   51565  662207
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.269e+05  4.222e+03  30.051  < 2e-16 ***
## ttl_disb                        1.207e-02  2.072e-03   5.825 8.04e-09 ***
## I(party_membership)Other Party -8.157e+04  1.134e+04  -7.192 1.37e-12 ***
## I(party_membership)Republican   3.679e+03  5.318e+03   0.692    0.489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76040 on 876 degrees of freedom
##   (462 observations deleted due to missingness)
## Multiple R-squared:  0.1087, Adjusted R-squared:  0.1056
## F-statistic:  35.6 on 3 and 876 DF,  p-value: < 2.2e-16
```

6. (3 points) Evaluate the Large-Sample Linear Model Assumptions

- A large data sample to work with 1,300+ rows
- The sample is IID: We assume IID based on the information provided in the documentation of how the data was collected.
- A BLP exists if the covariance between the features and between each feature and the output variable are finite or they do not have fat tails. BLP is considered unique provided that none of the features can be written as a linear combination of the other features. In this case BLP may not exist because of the skewness/heavytails of our features and output.

7. (3 points) Interpret the model coefficients you estimate.

- Tasks to keep in mind as you're writing about your model:
  - At the time that you're writing and interpreting your regression coefficients you'll be *deep* in the analysis. Nobody will know more about the data than you do, at that point. *So, although it will feel tedious, be descriptive and thorough in describing your observations.*
  - It can be hard to strike the balance between: on the one hand, writing enough of the technical underpinnings to know that your model meets the assumptions that it must; and, on the other hand, writing little enough about the model assumptions that the implications of the model can still be clear. We're starting this practice now, so that by the end of Lab 2 you will have had several chances to strike this balance.

Answer: What can be gleaned from the this model: The omitted category is democrats who didn't spend any money; meaning that democrats who spent less than a dollar ($0.01) received 126,872 votes. Of those party members who identified as "Other Party" could expect to receive 81,566 less votes or a total of 45,306 votes for the same amount of money spent (approx $0.01). However, those candidates who are republican saw an increase of 3,679 more votes than democrats for the same spend.