Politics Are Afoot!

Da Qi Ren

The Setup

There is *a lot* of money that is spent in politics in Presidential election years. So far, estimates have the number at about $11,000,000,000 (11 billion USD). For context, in 2019 Twitter's 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

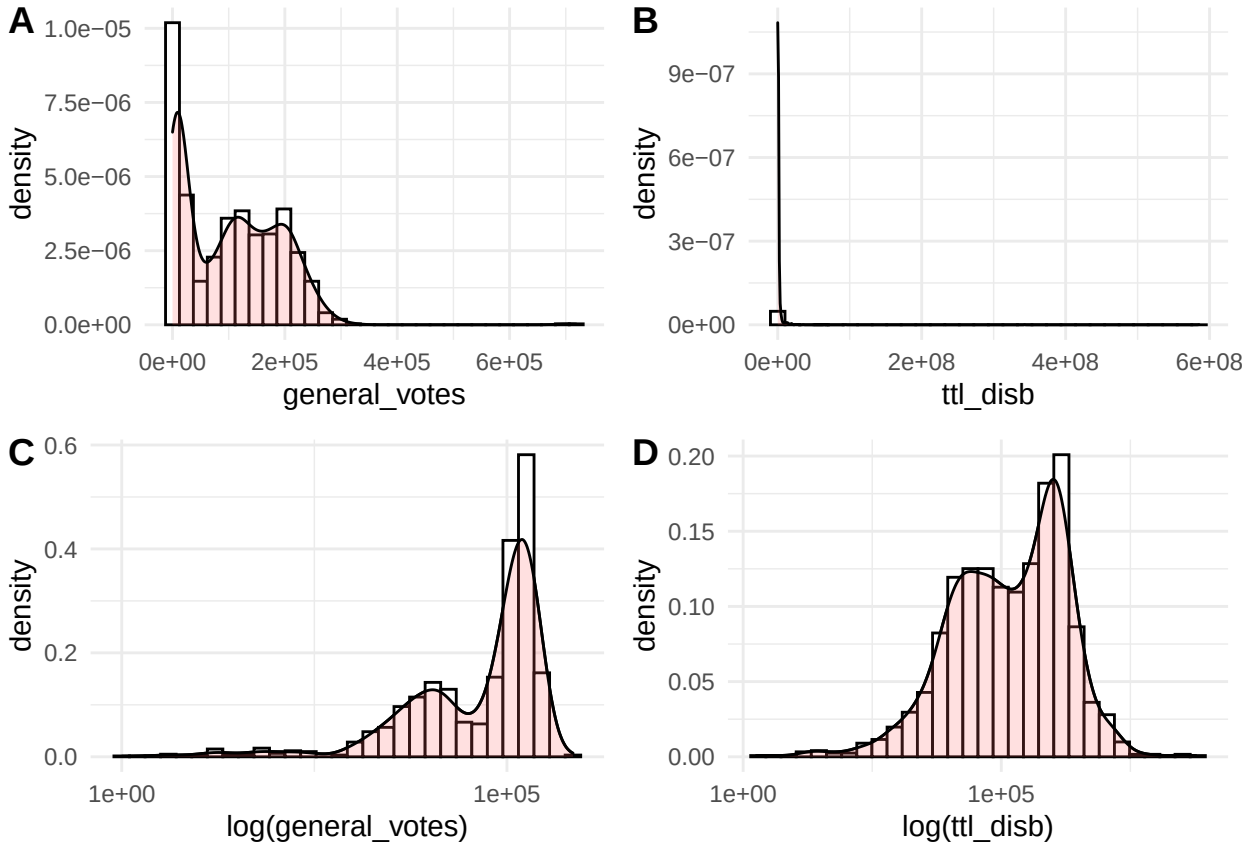
Describe the relationship between spending on a candidate's behalf and the votes they receive.

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.

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?



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`.)

```
d1 <- dplyr::inner_join(results_house, campaigns, by = NULL)
```

```
## Joining, by = "cand_id"
```

```
d1[d1 == -Inf] <- 0
nrow(d1)
```

```
## [1] 1342
```

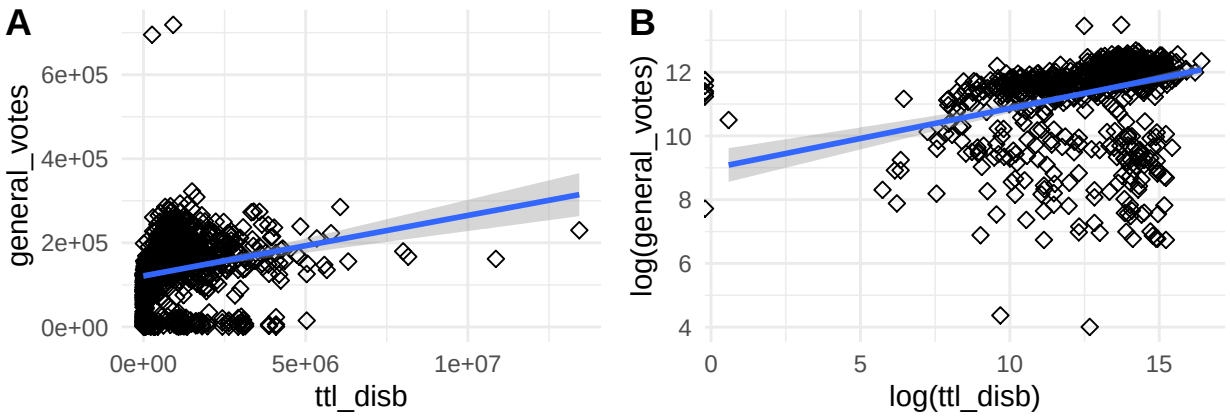
```
summary(d1)
```

```
##     state           district_id          cand_id          incumbent
##  Length:1342        Length:1342        Length:1342        Mode :logical
##  Class :character   Class :character   Class :character   FALSE:895
##  Mode  :character   Mode  :character   Mode  :character   TRUE :447
##
##
##
##
##     party           primary_votes    primary_percent    runoff_votes
##  Length:1342        Min.   :     1   Min.   :0.00015   Min.   : 1096
##  Class :character   1st Qu.:  8650   1st Qu.:0.19158   1st Qu.: 1464
##  Mode  :character   Median : 21299   Median :0.42257   Median : 8206
##                     Mean   : 32227   Mean   :0.48844   Mean   :11274
##                     3rd Qu.: 45638   3rd Qu.:0.78382   3rd Qu.:20082
##                     Max.   :326988   Max.   :1.00000   Max.   :25322
##                     NA's   :291      NA's   :292       NA's   :1330
##  runoff_percent   general_votes    general_percent     won
##  Min.   :0.3427   Min.   :    55   Min.   :0.0000   Mode :logical
##  1st Qu.:0.4624   1st Qu.: 88229   1st Qu.:0.3087   FALSE:850
##  Median :0.5000   Median :142597   Median :0.4773   TRUE :492
##  Mean   :0.5000   Mean   :136932   Mean   :0.4597
##  3rd Qu.:0.5376   3rd Qu.:198290   3rd Qu.:0.6406
##  Max.   :0.6573   Max.   :718591   Max.   :1.0000
##  NA's   :1330     NA's   :462      NA's   :463
##   footnotes          cand_name           cand_ici            pty_cd
##  Length:1342        Length:1342        Length:1342        Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :1.607
##                                                           3rd Qu.:2.000
##                                                           Max.   :3.000
##
##  cand_pty_affiliation  ttl_receipts      trans_from_auth       ttl_disb
##  Length:1342          Min.   :       0   Min.   :       0   Min.   :       0
##  Class :character     1st Qu.:   46612   1st Qu.:       0   1st Qu.:   46147
##  Mode  :character     Median :  398962   Median :       0   Median :  379570
```

```
##                       Mean   :  883177   Mean   :   26408   Mean   :  814754
##                       3rd Qu.: 1290266   3rd Qu.:       0   3rd Qu.: 1154148
##                       Max.   :19852221   Max.   :12374657   Max.   :13433669
##
##  trans_to_auth       coh_bop           coh_cop         cand_contrib
##  Min.   :     0   Min.   : -18681   Min.   : -32074   Min.   :       0
##  1st Qu.:     0   1st Qu.:      0   1st Qu.:      0   1st Qu.:       0
##  Median :     0   Median :      0   Median :   3881   Median :       0
##  Mean   :  7577   Mean   : 150271   Mean   : 218929   Mean   :   21879
##  3rd Qu.:     0   3rd Qu.:  85884   3rd Qu.: 170548   3rd Qu.:    1000
##  Max.   :766500   Max.   :3750024   Max.   :9098873   Max.   :13414225
##
##    cand_loans       other_loans     cand_loan_repay   other_loan_repay
##  Min.   :      0   Min.   :     0   Min.   :      0   Min.   :     0.0
##  1st Qu.:      0   1st Qu.:     0   1st Qu.:      0   1st Qu.:     0.0
##  Median :      0   Median :     0   Median :      0   Median :     0.0
##  Mean   :  56809   Mean   :  1049   Mean   :  12579   Mean   :   638.7
##  3rd Qu.:   9000   3rd Qu.:     0   3rd Qu.:      0   3rd Qu.:     0.0
##  Max.   :8050000   Max.   :350000   Max.   :1655854   Max.   :350000.0
##
##  debts_owed_by     ttl_indiv_contrib cand_office_st     cand_office_district
##  Min.   :  -1786   Min.   :      0   Length:1342        Length:1342
##  1st Qu.:      0   1st Qu.:  21310   Class :character   Class :character
##  Median :      0   Median : 207337   Mode  :character   Mode  :character
##  Mean   :  42528   Mean   : 464597
##  3rd Qu.:  12903   3rd Qu.: 638629
##  Max.   :2795000   Max.   :5975190
##
##  other_pol_cmte_contrib pol_pty_contrib   cvg_end_dt         indiv_refunds
##  Min.   :      0        Min.   :    0   Min.   :2015-08-10   Min.   : -1150
##  1st Qu.:      0        1st Qu.:    0   1st Qu.:2016-12-31   1st Qu.:     0
##  Median :  13700        Median :    0   Median :2016-12-31   Median :   200
##  Mean   : 305670        Mean   : 1230   Mean   :2016-11-30   Mean   :  6617
##  3rd Qu.: 506471        3rd Qu.:  150   3rd Qu.:2016-12-31   3rd Qu.:  5400
##  Max.   :3279747        Max.   :25400   Max.   :2017-01-31   Max.   :227497
##
##   cmte_refunds
##  Min.   :     0
##  1st Qu.:     0
##  Median :     0
##  Mean   :  1093
##  3rd Qu.:   250
##  Max.   :104758
##
```

```
#write.csv(d1, "d1.csv")
```

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?

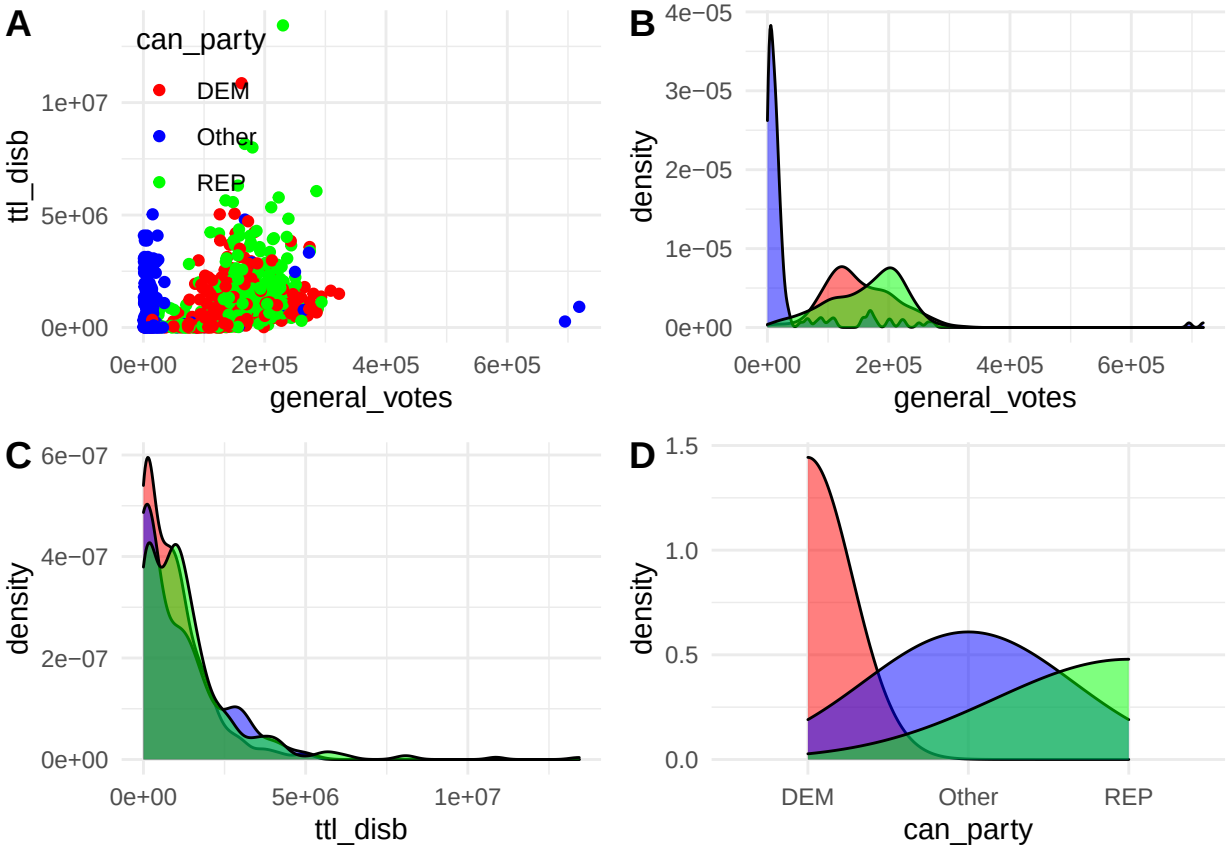


4. (3 points) Create a new variable to indicate whether each individual is a "Democrat", "Republican" or "Other Party".

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

```
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?



Produce a Descriptive Model

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.

```
##
##  studentized Breusch-Pagan test
##
## data:  sdat
## BP = 472.23, df = 57, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = general_votes ~ ttl_disb + can_party)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162812  -50839    -463   37128  645725
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.634e+05  4.080e+03  40.061  < 2e-16 ***
## ttl_disb     1.163e-02  1.864e-03   6.238 6.88e-10 ***
## can_party   -5.062e+04  3.213e+03 -15.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69240 on 877 degrees of freedom
## Multiple R-squared:  0.2602, Adjusted R-squared:  0.2585
## F-statistic: 154.2 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
## [1] 0.01738939
```

```
## [1] 880
```

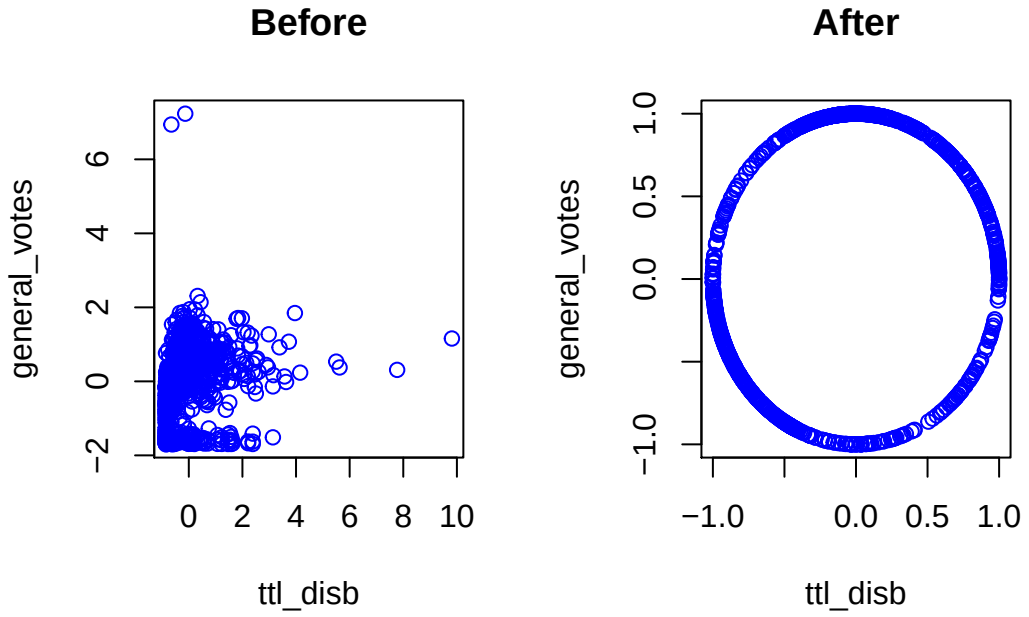
```
## [1] 6
```

```
## [1] 15.30266
```

```
## [1] 0.009144436
```

```
##
## Call:
## lm(formula = general_votes ~ ttl_disb + can_party, weights = 1/abs(e))
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -404.99 -214.86   -1.48  194.43  808.03
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.641e+05  1.112e+03  147.50   <2e-16 ***
## ttl_disb     1.155e-02  7.035e-04   16.41   <2e-16 ***
## can_party   -5.268e+04  1.149e+03  -45.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 226.1 on 877 degrees of freedom
## Multiple R-squared:  0.7956, Adjusted R-squared:  0.7951
## F-statistic:  1707 on 2 and 877 DF,  p-value: < 2.2e-16
```




```
##    can_party      general_votes       ttl_disb          state
##  Min.   :0.0000   Min.   :    55   Min.   :       0   Length:880
##  1st Qu.:0.0000   1st Qu.: 88229   1st Qu.:  102276   Class :character
##  Median :1.0000   Median :142597   Median :  830659   Mode  :character
##  Mean   :0.7727   Mean   :136932   Mean   : 1084565
##  3rd Qu.:1.0000   3rd Qu.:198290   3rd Qu.: 1527533
##  Max.   :2.0000   Max.   :718591   Max.   :13433669
##     novotes             nodisb
##  Min.   :-1.00000   Min.   :-1.0000
##  1st Qu.:-0.65905   1st Qu.:-0.7263
##  Median : 0.07400   Median :-0.2163
##  Mean   : 0.07698   Mean   :-0.1272
##  3rd Qu.: 0.90077   3rd Qu.: 0.4287
##  Max.   : 1.00000   Max.   : 1.0000
```

```
##    can_party      general_votes       ttl_disb          state
##  Min.   :0.0000   Min.   :    55   Min.   :       0   Length:880
##  1st Qu.:0.0000   1st Qu.: 88229   1st Qu.:  102276   Class :character
##  Median :1.0000   Median :142597   Median :  830659   Mode  :character
##  Mean   :0.7727   Mean   :136932   Mean   : 1084565
##  3rd Qu.:1.0000   3rd Qu.:198290   3rd Qu.: 1527533
##  Max.   :2.0000   Max.   :718591   Max.   :13433669
```

```
##     novotes             nodisb           csvotes             csdisb      
##  Min.   :-1.00000   Min.   :-1.0000   Min.   :-1.70236   Min.   :-0.8619  
##  1st Qu.:-0.65905   1st Qu.:-0.7263   1st Qu.:-0.60573   1st Qu.:-0.7806  
##  Median : 0.07400   Median :-0.2163   Median : 0.07045   Median :-0.2018  
##  Mean   : 0.07698   Mean   :-0.1272   Mean   : 0.00000   Mean   : 0.0000  
##  3rd Qu.: 0.90077   3rd Qu.: 0.4287   3rd Qu.: 0.76311   3rd Qu.: 0.3520  
##  Max.   : 1.00000   Max.   : 1.0000   Max.   : 7.23415   Max.   : 9.8139
```

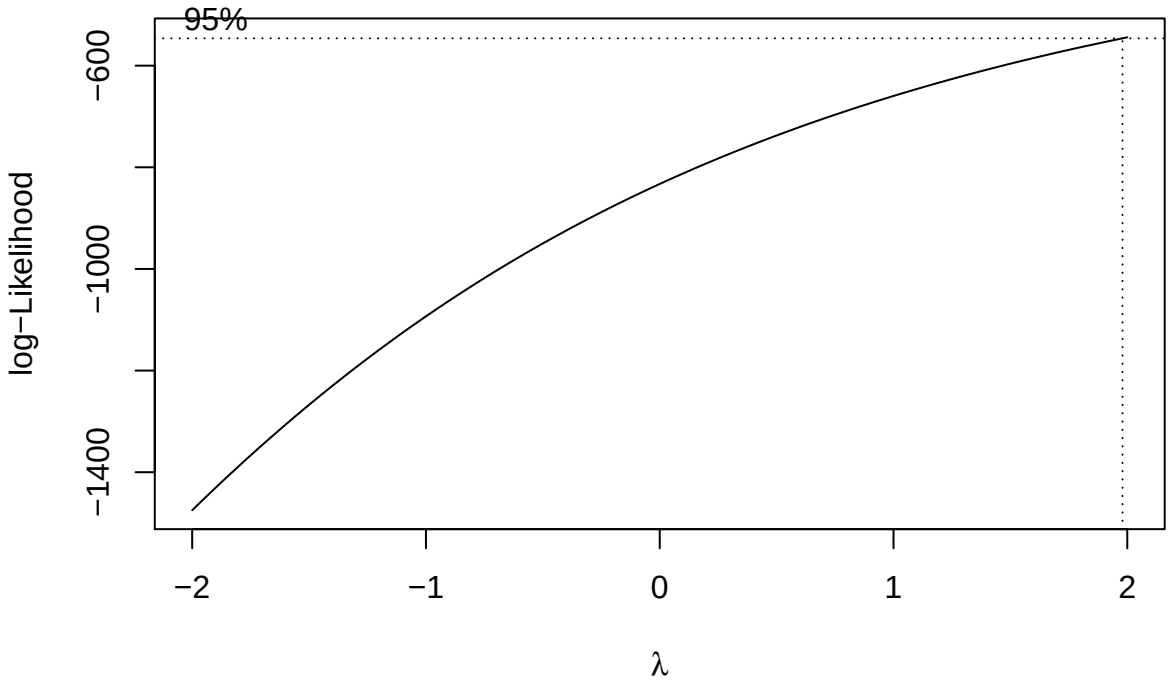
```
##
## Call:
## lm(formula = d2$general_votes ~ d2$ttl_disb + d2$can_party)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -162812  -50839    -463   37128  645725 
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1.634e+05  4.080e+03  40.061  < 2e-16 ***
## d2$ttl_disb   1.163e-02  1.864e-03   6.238 6.88e-10 ***
## d2$can_party -5.062e+04  3.213e+03 -15.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69240 on 877 degrees of freedom
## Multiple R-squared:  0.2602, Adjusted R-squared:  0.2585 
## F-statistic: 154.2 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = d2$novotes ~ d2$nodisb + d2$can_party)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.44064 -0.49643 -0.07907  0.54617  1.46145 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.45438    0.03129  14.521   <2e-16 ***
## d2$nodisb     0.27807    0.03266   8.515   <2e-16 ***
## d2$can_party -0.44263    0.02939 -15.062   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6349 on 877 degrees of freedom
## Multiple R-squared:  0.2653, Adjusted R-squared:  0.2636 
## F-statistic: 158.3 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = d2$logvotes ~ d2$logdisb + d2$logparty)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.3555 -0.1927  0.0741  0.2940  4.1067
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.60108    0.16135  65.704  < 2e-16 ***
## d2$logdisb   0.09767    0.01226   7.968 5.01e-15 ***
## d2$logparty -3.57205    0.10352 -34.507  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.809 on 877 degrees of freedom
## Multiple R-squared:  0.6044, Adjusted R-squared:  0.6035
## F-statistic: 669.9 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = d2$general_votes ~ d2$logdisb + d2$can_party)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -164084  -42521    2037   33117  627966
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20183.4    13565.1   1.488    0.137
## d2$logdisb    11935.7      999.7  11.939   <2e-16 ***
## d2$can_party -46716.3     3070.3 -15.215   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65620 on 877 degrees of freedom
## Multiple R-squared:  0.3354, Adjusted R-squared:  0.3339
## F-statistic: 221.3 on 2 and 877 DF,  p-value: < 2.2e-16
```




```
##         lambda       lik
## [1,] -2.000000 -1474.865
## [2,] -1.959596 -1456.833
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.727   -6.848    1.412    8.771  119.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.737      4.067  29.192   <2e-16 ***
## logdisb        3.029      0.309   9.802   <2e-16 ***
## logparty     -91.757      2.610 -35.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.39 on 877 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:   0.62
## F-statistic: 718.1 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.727   -6.848    1.412    8.771  119.849
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.737      4.067  29.192   <2e-16 ***
## logdisb        3.029      0.309   9.802   <2e-16 ***
## logparty     -91.757      2.610 -35.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.39 on 877 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:   0.62
## F-statistic: 718.1 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.727   -6.848    1.412    8.771  119.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.737      4.067  29.192   <2e-16 ***
## logdisb        3.029      0.309   9.802   <2e-16 ***
## logparty     -91.757      2.610 -35.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.39 on 877 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:   0.62
## F-statistic: 718.1 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
## [1] 0.1645108
```

```
## [1] 880
```

```
## [1] 6
```

```
## [1] 144.7695
```

```
## [1] 0
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty, weights = 1/abs(e))
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -10.184  -2.701   1.057   2.873  10.958
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 118.10513    0.82784  142.67   <2e-16 ***
## logdisb       3.09811    0.06612   46.85   <2e-16 ***
## logparty    -92.24965    0.34882 -264.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.571 on 877 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.9881
## F-statistic: 3.635e+04 on 2 and 877 DF,  p-value: < 2.2e-16
```

6. (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.