Politics Are Afoot!

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.

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).

Our task

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

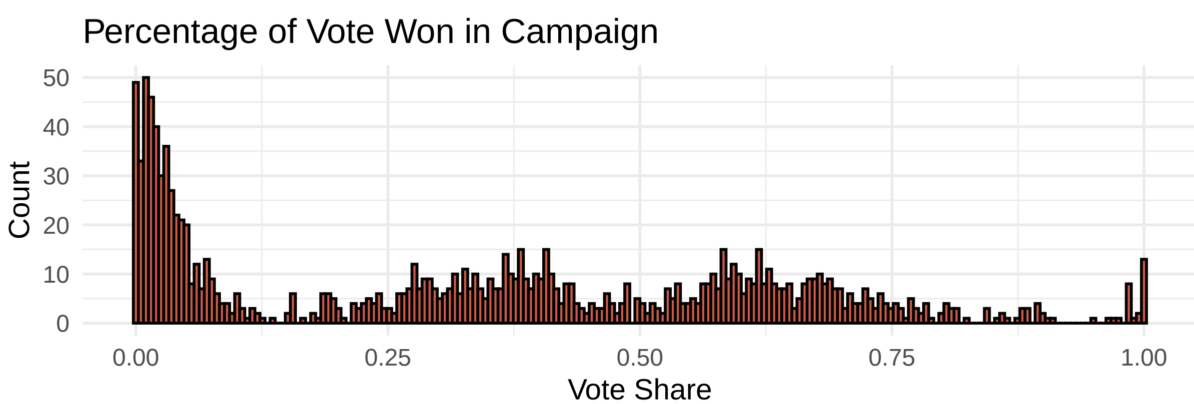
Work

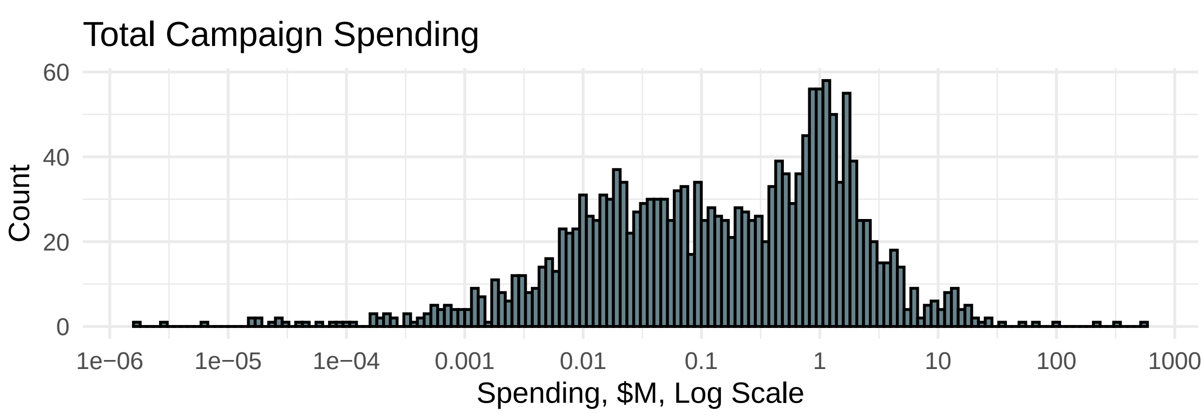
```
library(tidyverse)
library(magrittr)
library(ggplot2)
library(patchwork)
library(sandwich)
library(lmtest)
library(fec16)
theme_set(theme_minimal())
knitr::opts_chunk$set(dpi = 300)
```

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

1. Distribution of votes and of spending:

```
histogram_of_vote <- results_house %>%
  ggplot() +
  aes(x = general_percent) +
  geom_histogram(bins = 200,
    fill="coral3", color = "black") +
  labs(
    title = 'Percentage of Vote Won in Campaign',
    x = 'Vote Share', y = 'Count')
histogram_of_spending <- campaigns %>%
  ggplot() +
  aes(x = log10(ttl_disb)) +
  geom_histogram(bins = 150,
    fill="lightblue4", color = "black") +
  labs(
    title = 'Total Campaign Spending',
    x     = 'Spending, $M, Log Scale',
    y     = 'Count') +
  scale_x_continuous(
    breaks=seq(0, 10, 1),
    labels = 10^(seq(0,10,1)-6))
histogram_of_vote / histogram_of_spending
```



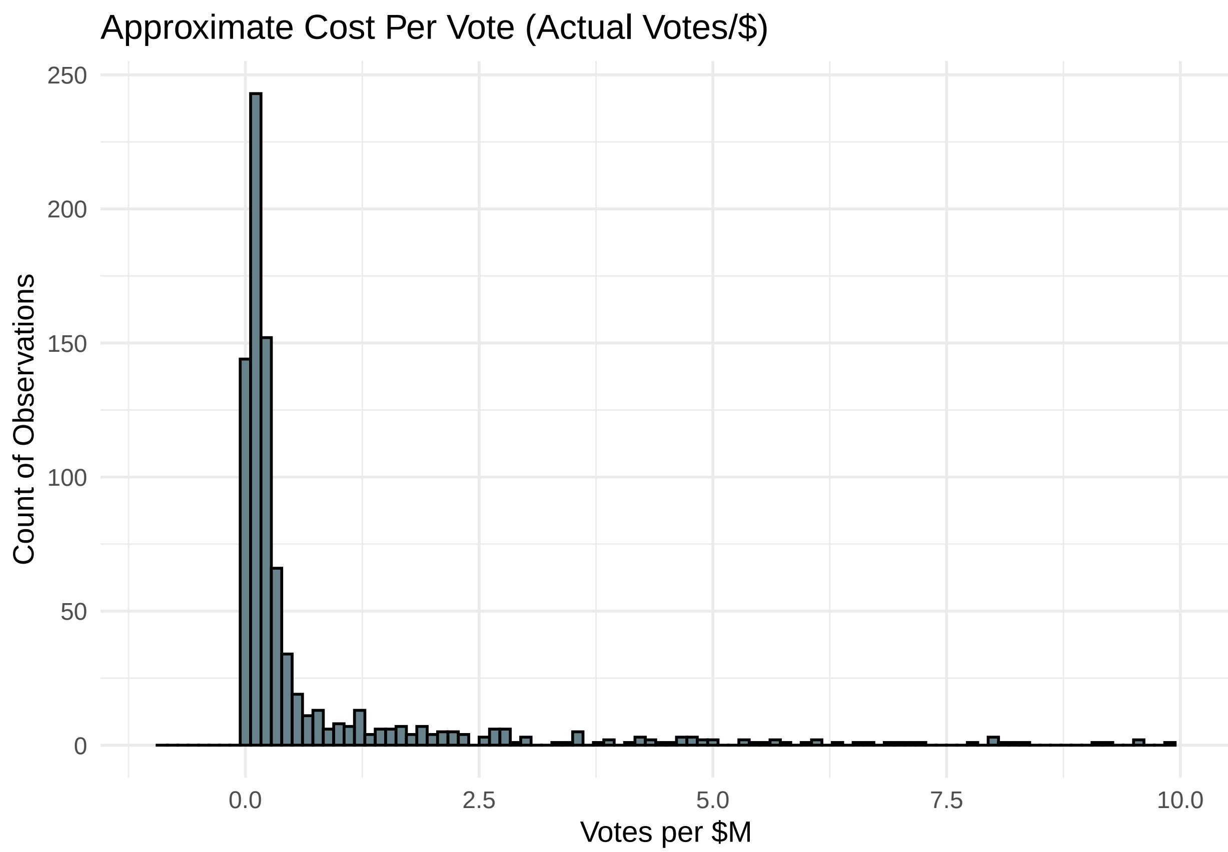


The vote distribution shows four modes, the greatest being candidates who receive a small (<10%) share of the vote. There are a few who receive nearly all the vote, and two groups that receive 5-10% more or less than half the votes.

The spending plot shows spending on campaigns mostly in the $0.1M to $3M range, with a large peak around $1M.

2. Exploring the relationship between spending and votes.

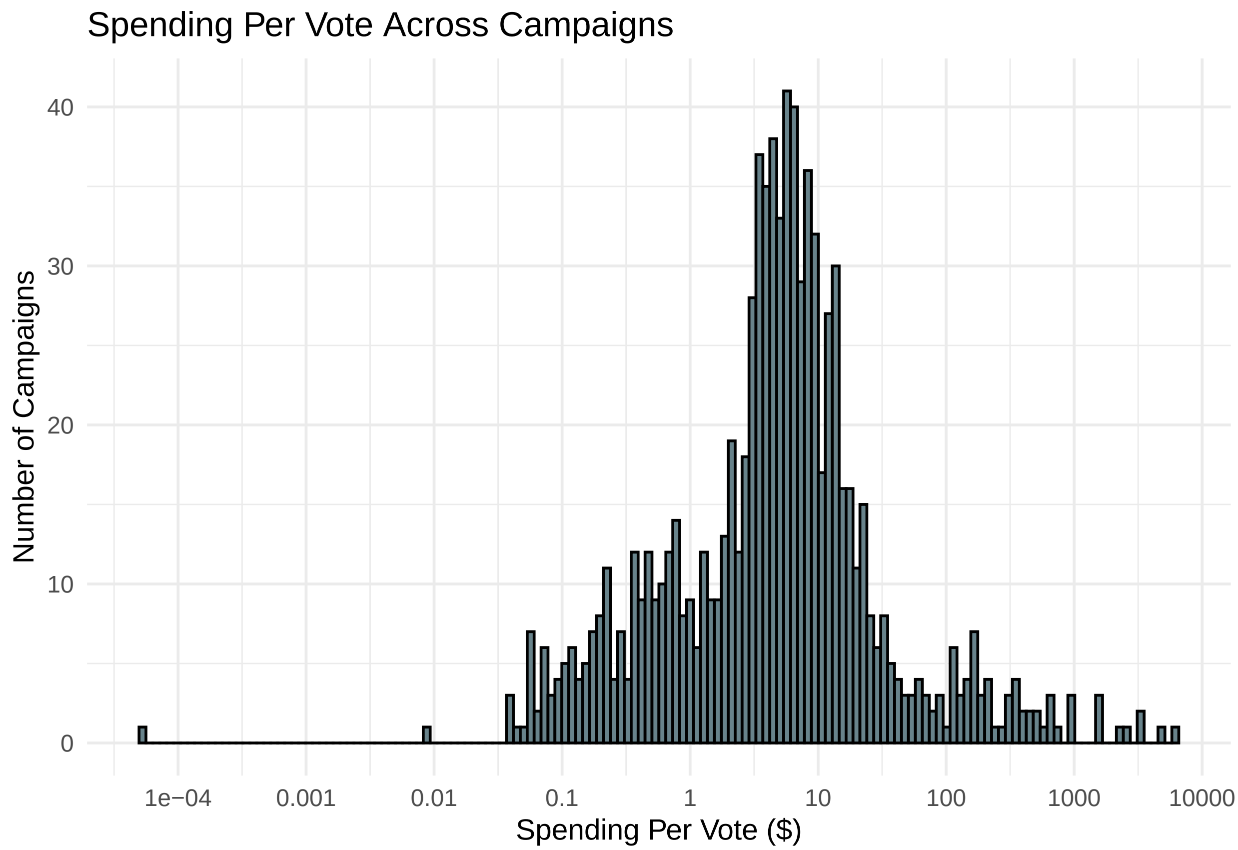
```
d <- inner_join(results_house, campaigns, by = 'cand_id')
# engineer reused features
d <- d %>%
  mutate(
    cost_per_vote = general_votes / ttl_disb,
    party = case_when(
      cand_pty_affiliation == "REP" ~ "Republican",
      cand_pty_affiliation == "DEM" ~ "Democrat",
      TRUE                          ~ "Other")
  )
d %>%
  ggplot(aes(x = cost_per_vote)) +
  geom_histogram(bins=100,
    fill="lightblue4", color = "black") +
  xlim(c(-1, 10)) +
  labs(
    title = 'Approximate Cost Per Vote (Actual Votes/$)',
    x     = 'Votes per $M',
    y     = 'Count of Observations'
  )
```



I have restricted the data to campaigns that got votes in the general election. I have done this by removing the rows for which general votes was marked “NA” in the data. There were data for 1898 campains in total. Of those, there were finance data for 1342 of them, and of those 880 received votes in the general election. The difference in 1342 and 880 were candidate who received votes in the primary, but did not participate in the general election. The total number of campaigns with votes in neither the general election and the primary was 4.

The histogram is above, but I find this to be very misleading. The “Cost per Vote” metric is actually upside-down. Based on its calculation, it should be labelled Votes/$. Here is a better graph for Cost Per Vote (log transformed). Note the peak around $5/vote. Also, “spending” is a better term than cost, as it reflects association better than “cost” which implies causation.

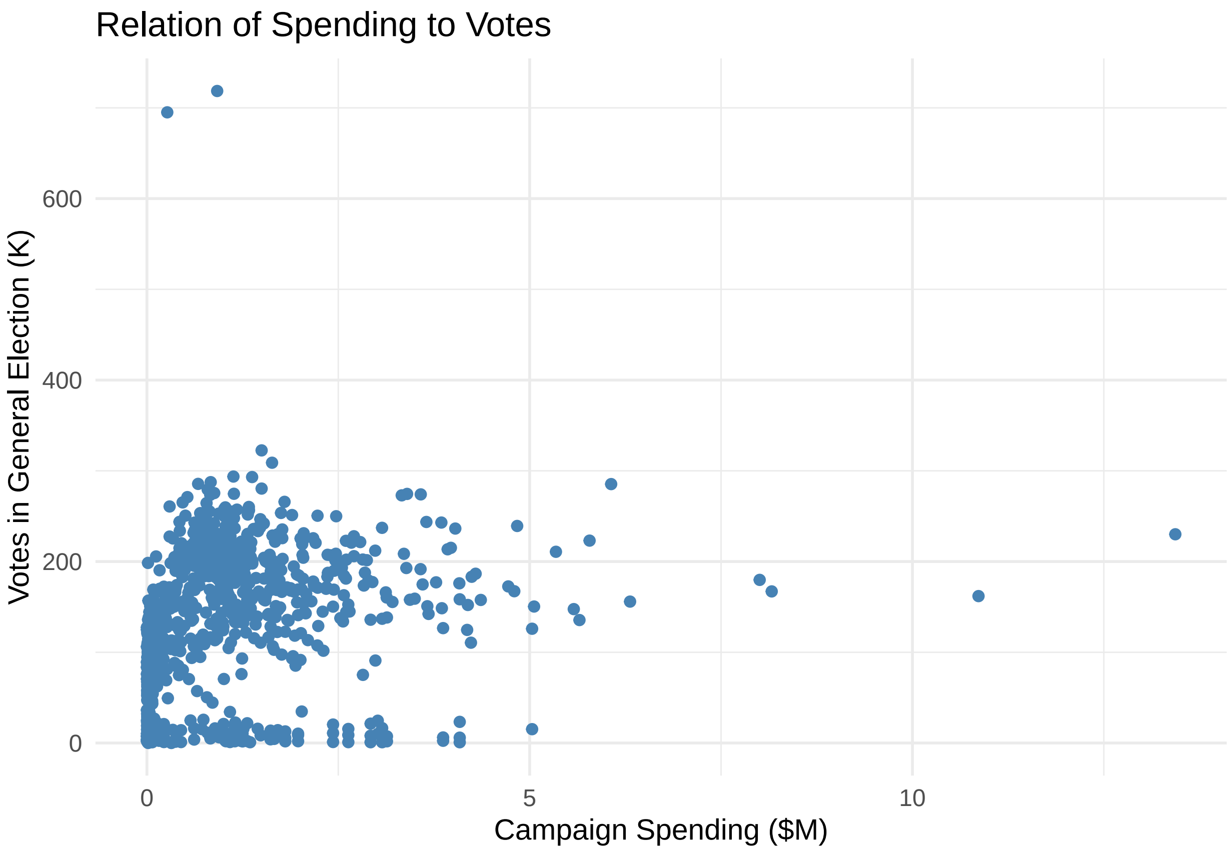
```
d %>%
  mutate(Real_cost_per_vote = ttl_disb/general_votes) %>%
  ggplot(aes(x = log10(Real_cost_per_vote ))) +
  geom_histogram(bins=150,
    fill="lightblue4", color = "black") +
  labs(
    title = 'Spending Per Vote Across Campaigns',
    x     = 'Spending Per Vote ($)',
    y     = 'Number of Campaigns') +
  #  this is how we label the axis meaningfully for log10 variables
  scale_x_continuous(
    breaks=seq(-4, 10, 1),
    labels = 10^seq(-4,10,1))
```



There seem to be a lot of candidates with a zero cost-per-vote. What might this mean?

The name is bad. This is votes per dollar… Zero votes per dollar really means that the candidate got very few votes for the money they spent.. That's what the spike on the left side of the first histogram means. This either means that they got few votes, or they spent a great deal on few votes, or both. More on this later.

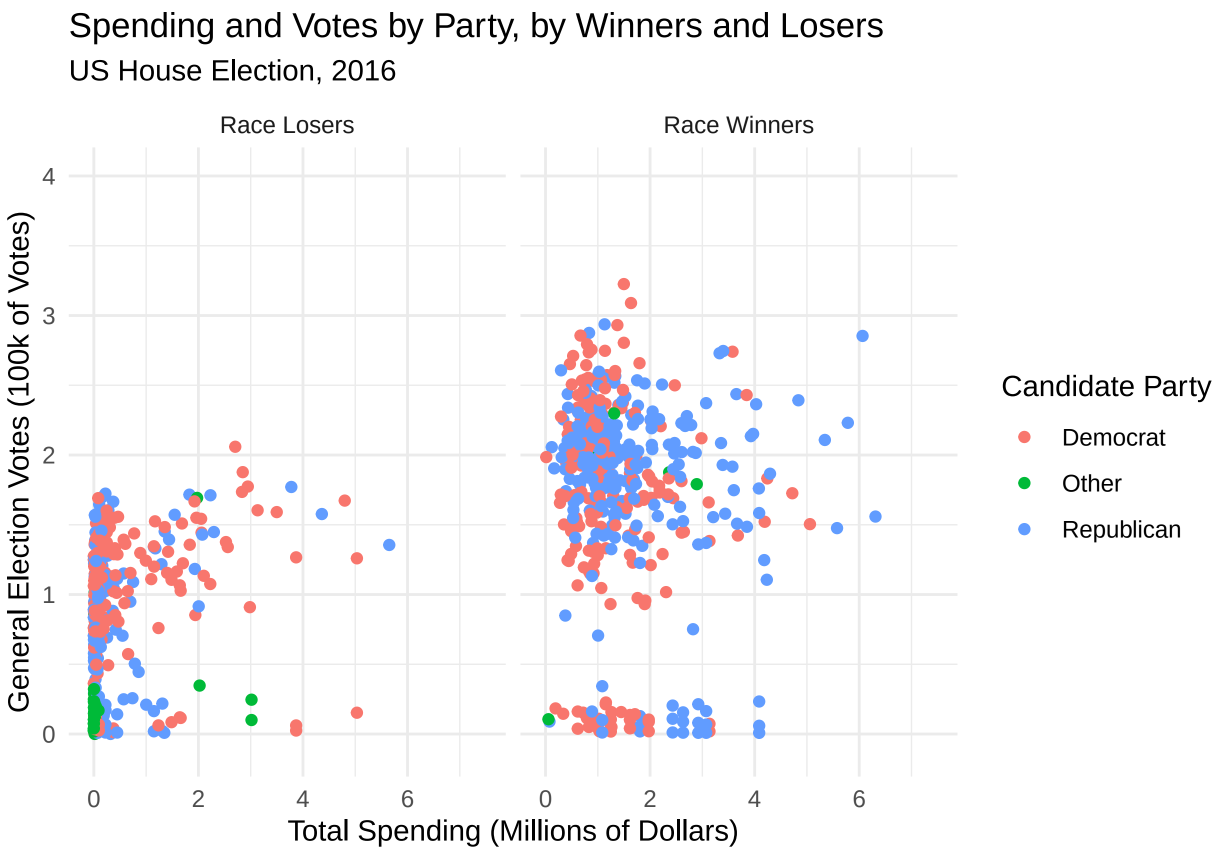
```
d %>%
  ggplot(aes(x = ttl_disb/1000000, y = general_votes/1000)) +
  geom_point(col="steelblue") +
  labs(
    title = 'Relation of Spending to Votes',
    x     = 'Campaign Spending ($M)',
    y     = 'Votes in General Election (K)'
  )
```



There are many campaigns that received very few votes, desite spending millions. These are on the bottom of the graph. There are also many campaigns that received many votes despite very little spending. These are on the left side of the graph. The remaining campaigns show that most campaigns spent around $1M, but a few campaigns spent more than $5M.

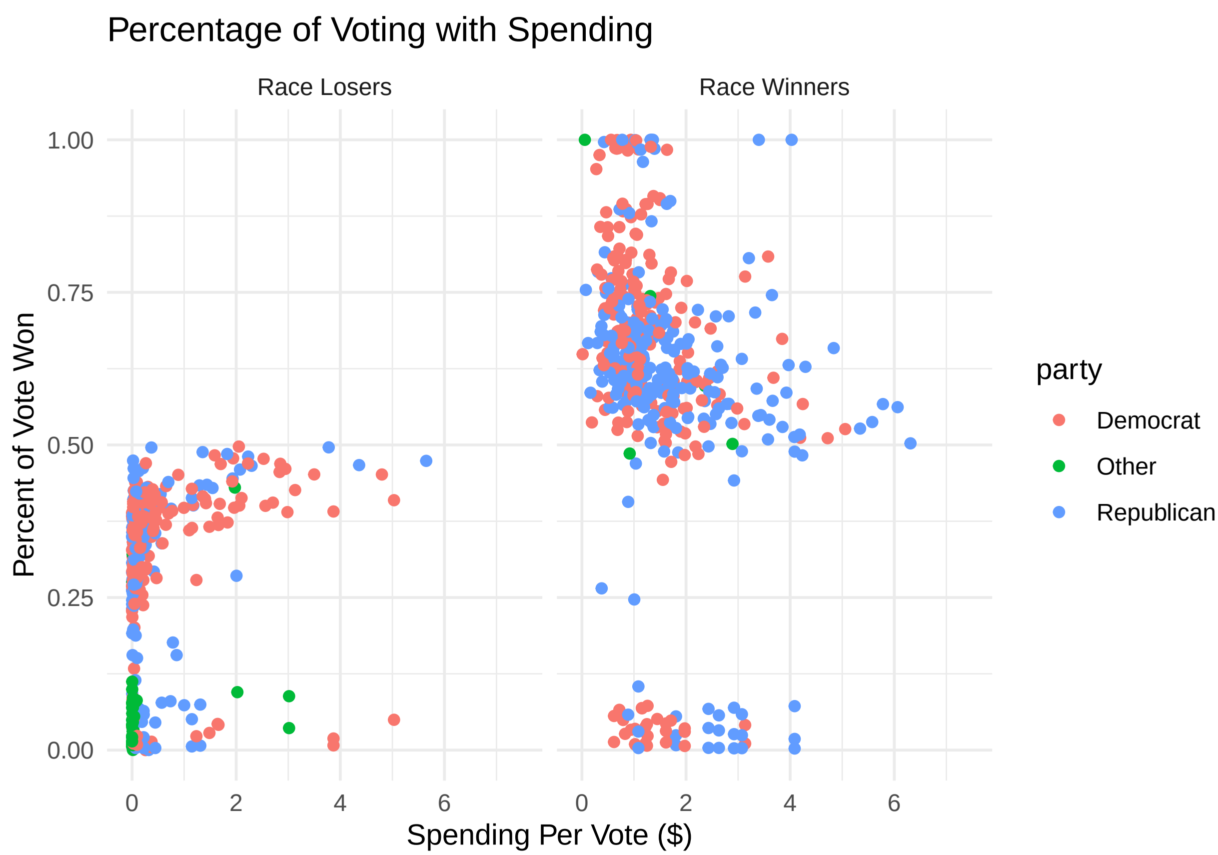
Note: This is a niaive approach to this question, based on accepted the data as they are in the plot. A more accurate analysis is included below.

3. Party Differenciation



Another way to present this information is to look at the percentage of votes won versus spending per vote.

```
d %>%
  select(
    state, district_id, cand_id, party, incumbent, # place vars
    primary_percent, general_percent, won, general_votes,        # performance vars
    ttl_disb) %>%                                                 # spending vars
  mutate(
    race_winner = case_when(
    won == TRUE ~ 'Race Winners',
    won != TRUE ~ 'Race Losers')) %>%
  ggplot() +
    aes(
      x = ttl_disb/ 1e6 , y = general_percent,
      color = party) +
    geom_point() +
    facet_wrap(facets = vars(race_winner), nrow = 1, ncol = 2, ) +
    scale_x_continuous(limits = c(-0.1, 7.5)) +
  labs(
    title = 'Percentage of Voting with Spending',
    x     = 'Spending Per Vote ($)',
    y     = 'Percent of Vote Won'
  )
```



The data points in the bottom right part of the graph (and the bottom left of the graph above this one) require some investigation. They are quite suspicious. How is it possible to spend $1000 per vote and get such a small percentage of the vote?

```
d %>%
  filter(
    state=="NY",
    district_id=="01",
    cand_id =="H8NY01148") %>%
  select(
    state, district_id, cand_id, cand_name,
    party, general_votes, general_percent, cand_pty_affiliation,
    ttl_disb)
```

```
## # A tibble: 4 x 9
##   state district_id cand_id   cand_name     party  general_votes general_percent
##   <chr> <chr>       <chr>     <chr>         <chr>          <dbl>           <dbl>
## 1 NY    01          H8NY01148 ZELDIN, LEE M Repub~        158409         0.489
## 2 NY    01          H8NY01148 ZELDIN, LEE M Repub~         23327         0.0720
## 3 NY    01          H8NY01148 ZELDIN, LEE M Repub~           843         0.00260
## 4 NY    01          H8NY01148 ZELDIN, LEE M Repub~          5920         0.0183
## # ... with 2 more variables: cand_pty_affiliation <chr>, ttl_disb <dbl>
```

After some investigation, these data are mostly for races in CT and NY. In some states (CT, NY), a single candidate can run under multiple parties, so the candidate can get votes under a number of parties. This is the story behind those data points in the bottom right. To restate the problem, in the data, there are some candidates who have multiple rows of data for the *same campaign* under different parties.

Lee Zeldin, a candidate in the First District of New York, who identifies as a Republican, has votes from four party affiliations. According to the FEC website (https://www.fec.gov/resources/cms-content/documents/partylabels2012.pdf), these are:

- **R** Republican;
- **CRV** Conservative;
- **REF** Reform; and,
- **IDP** Independence.

Note that the votes are by party, but the disbursement value is the same for all rows, so the votes (and percentage votes) need to be summed, but the disbursement is value from a single row.

I group the votes together as a single campaign and use total disbursements to place candidate on these graphs. The suspicious data at the bottom is gone.

```
# in some states, a candidate can represent multiple parties -
# group those lines in a single line for each state/district_id/cand_id
d_single_candidates <- d %>%
  mutate(
    primary_votes   = zoo::na.fill(primary_votes, 0), # make primary votes 0 if NA
    general_votes   = zoo::na.fill(general_votes, 0), # make primary votes 0 if NA
    general_percent = zoo::na.fill(general_percent, 0), # make primary votes 0 if NA
    WL = case_when(
      won == FALSE ~ "Race Losers",
      won == TRUE ~  "Race Winners"),
    INCU = ifelse(incumbent == "TRUE", "Incumbent", "Challenger")) %>%
  group_by(state, district_id, cand_id) %>%
  summarize(
    general_votes   = sum(general_votes),
    general_percent = sum(general_percent),
    runoff_votes    = sum(runoff_votes),
    cand_name       = first(cand_name),
    primary_votes   = sum(primary_votes),
    cand_pty_affiliation = first(cand_pty_affiliation),
    ttl_receipts    = max(ttl_receipts),
    incumbent       = first(incumbent),
    INCU            = first(INCU),
    WL              = first(WL),
    pol_pty_contrib = max(pol_pty_contrib),
    other_pol_cmte_contrib = max(other_pol_cmte_contrib),
    party           = first(party),
    ttl_disb        = max(ttl_disb),
    ttl_indiv_contrib = max(ttl_indiv_contrib)) %>%
  mutate(
    log_dollars_per_vote = log10((ttl_disb+1)/(general_votes+1)),
    valid_general = FALSE)
```

Now, let's take a look at the single candidate data, especially those with very low votes and high expeditures. We need to remove write-in candidates who would have few votes, but whose votes would correspond to higher spending from the primary. In other words, we will remove data for candidates who lost in the primary and whose general election votes do not reflect the relationship we wish to model.

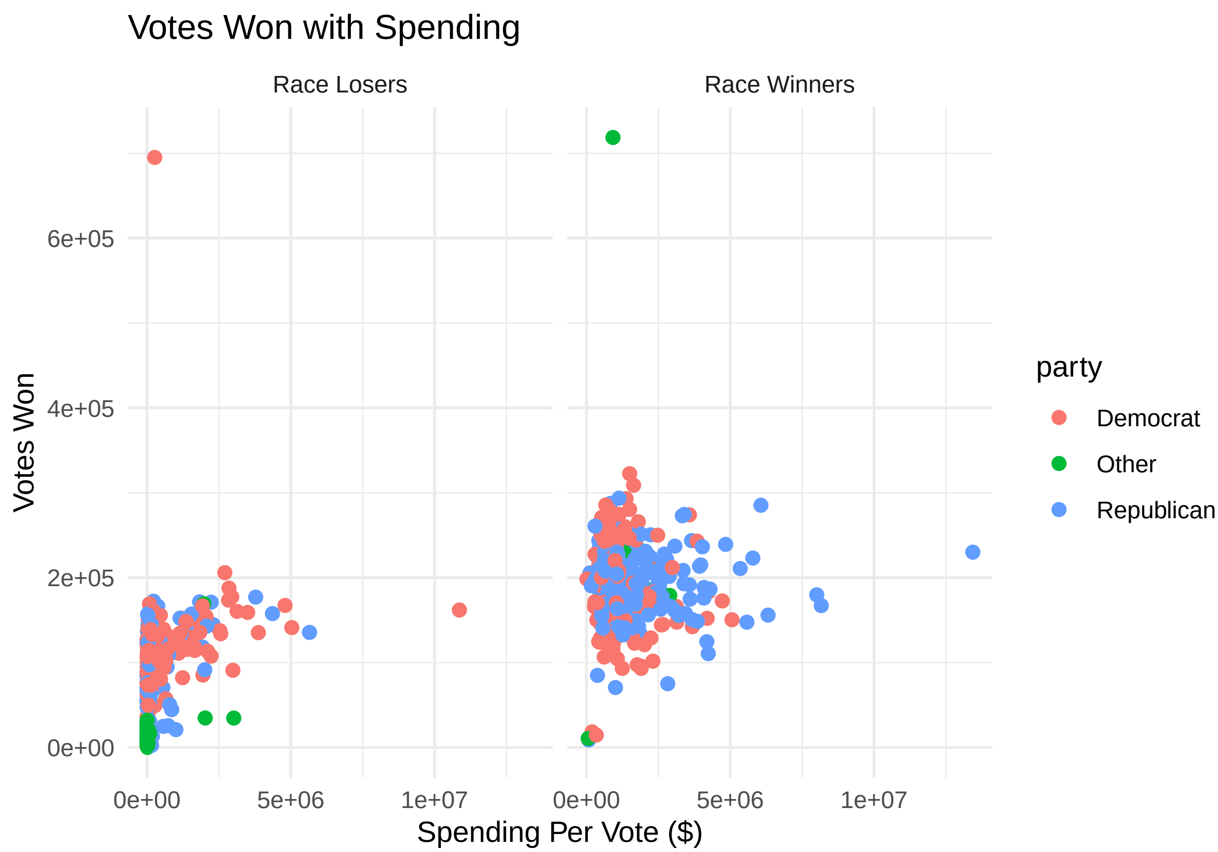
There are six of these.

```
#  Data wrangling - lot's of it...
#
#  for viewing the runoff loser data
d_runoff_losers <- d_single_candidates %>%
```

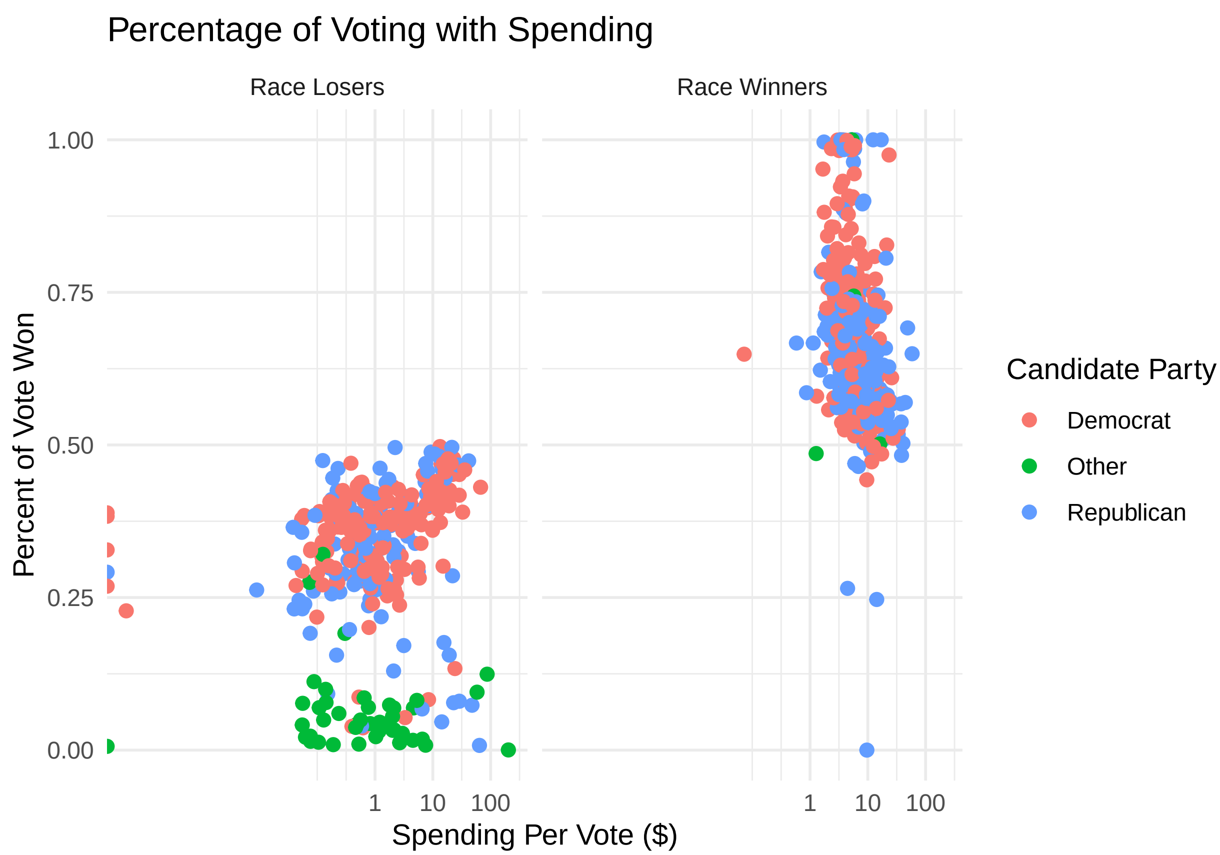
```r
  filter(!is.na(runoff_votes)) %>%
  group_by(state, district_id, cand_pty_affiliation)  %>%
  slice_min(order_by = runoff_votes, n = 1) %>%
  ungroup()
#for use in filtering out these candidates
runoff_losers <- d_runoff_losers %>%
  select(cand_id) %>%
  unlist
#  six candidates lost in the runoff but got general votes - remove those
d_single_candidates_2 <- d_single_candidates %>%
  filter(!cand_id %in% runoff_losers)
# now remove candidates that lost multi-party or single-party primaries
# then remove the rest of the candidates that got no general election votes
mp_states <- c("CA","WA")
# mark candidates that lost the primary - in some states this is top in party
#   697 of these
won_primary <- d_single_candidates %>%
  filter(! state %in% mp_states) %>%
  group_by(state, district_id, cand_pty_affiliation) %>%
  slice_max(order_by = primary_votes, n = 1) %>%
  ungroup() %>%
  select(cand_id) %>%
  unlist()
#  mutate(won_party_primary = TRUE, valid_general) %>%
#  select(state, district_id,cand_id, won_party_primary, valid_general)

# in california, etc , top two candidates from any party go to general
# 122 of these
d_won_MP_primary <- d_single_candidates %>%
  filter(state %in% mp_states) %>%
  group_by(state, district_id) %>%
  slice_max(order_by = primary_votes, n = 2) %>%
  ungroup()
# Postma is missing from the DB, so the WA10 race should
#  not include the 2nd DEM as an MP winner.  Remove her.
won_MP_primary <- d_won_MP_primary %>%
  select(cand_id) %>%
  filter (cand_id != "H2WA10048") %>%
  unlist
# should be 818 of these - check
 d_single_candidates_general <- d_single_candidates %>%
   filter(
     ((!(state %in% mp_states)) & (cand_id %in% won_primary)) |
       ((state %in% mp_states) & (cand_id %in% won_MP_primary))) %>%
   # finally, remove the rows associated with no votes in the general election
   #  I have found several of these to be the result of missing candidates in the
   #   campaign db that caused candidates who were not in the general election
   #  to show up there.
   filter(general_votes > 0)
```

Finally, remove the rows associated with no votes in the general election I have found several of these to be the result of missing candidates in the campaign db that caused candidates who were not in the general election to show up there.

```
d_single_candidates_general %>%
  ggplot() +
  aes(x = ttl_disb,y = general_votes, color = party) +
  geom_point(size=2) +
  facet_wrap(vars(WL)) +
  labs(
    title = 'Votes Won with Spending',
    x = 'Spending Per Vote ($)',
    y = 'Votes Won'
  )
```



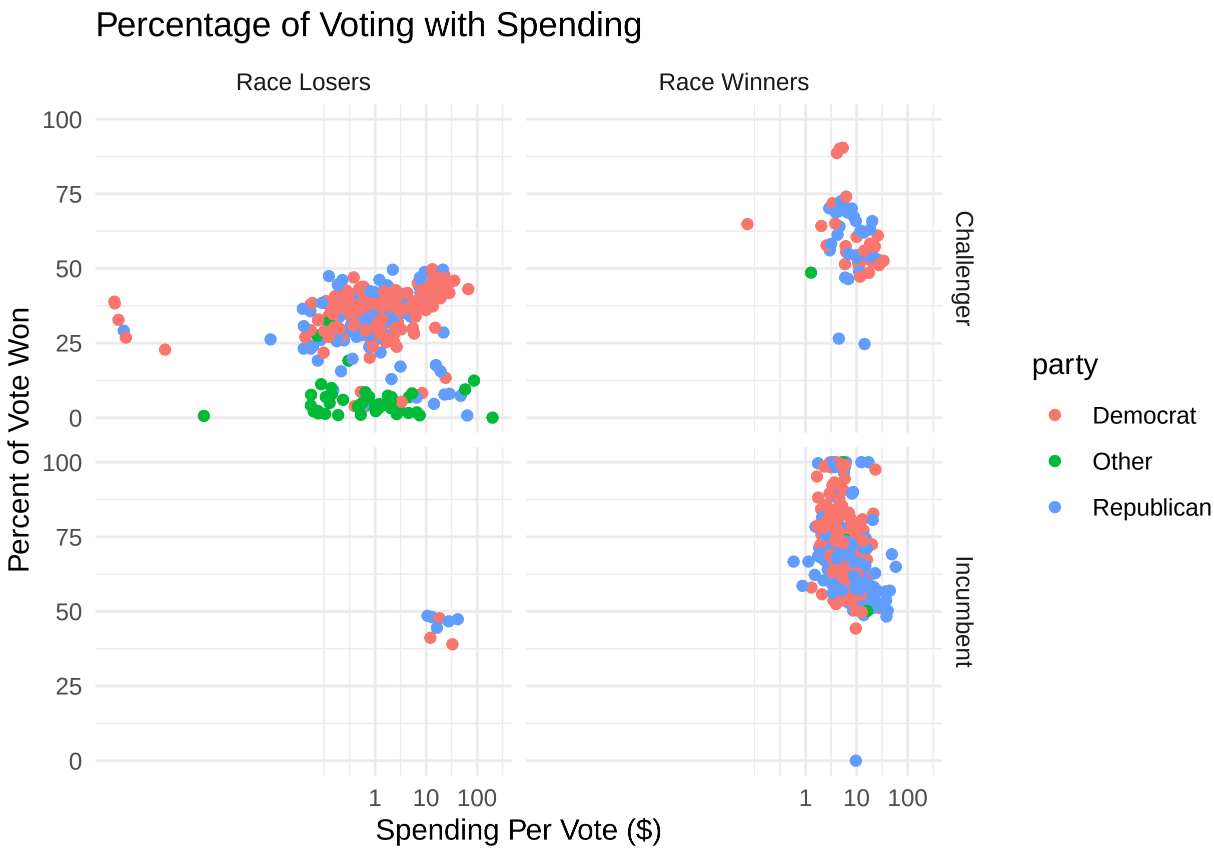

```
d_single_candidates_general %>%
  filter(!is.na(general_votes)) %>%
  ggplot() +
    aes(
      x = log10( ttl_disb / general_votes ),
      y = general_percent,
      color = party) +
    geom_point(size=2) +
    facet_wrap(vars(WL)) +
    xlim(-1,6) +
    scale_x_continuous(breaks=seq(0, 6, 1), labels = 10^seq(0,6,1)) +
    labs(
      title = 'Percentage of Voting with Spending' ,
      x = "Spending Per Vote ($)" ,
      y = "Percent of Vote Won",
      color = 'Candidate Party')
```



This makes a great deal more sense.

We can also split the data by Winner/Loser and Challenger/Incumbent to look for patterns.

```
d_single_candidates_general %>%
#  filter(won=="TRUE") %>%
#  arrange(desc(total_percent))
  ggplot(aes(x=log_dollars_per_vote, y=general_percent*100)) +
  geom_point(aes(col=party), size=1.5) +
  scale_x_continuous(breaks=seq(0, 3, 1),labels = 10^seq(0,3,1)) +
    labs(
    title = 'Percentage of Voting with Spending',
    x      = 'Spending Per Vote ($)',
    y      = 'Percent of Vote Won') +
  facet_grid(INCU ~WL)
```



It is also informative to look at the winners and the source of their campaign contributions. If we focus in on Winners for whom individual contributions were low, we can see that their parties contributed funds to give them a fundraising advantage.

```
just_winners <- d_single_candidates_deltas %>%
    filter(WL=="Race Winners")
just_winners_needing_help <- just_winners %>%
  filter(per_vote_indiv_contrib_advantage < 0)
d_x <- 0.16
```

```
#
# plt0 <- just_winners %>%
#   filter(not(delta_votes == 0)) %>%
#   ggplot(aes(x=per_vote_receipts_advantage, y=delta_pct, col=INCU)) +
# #  ggplot(aes(x=delta_disb, y=delta_votes, col=INCU)) +
#   geom_point() +
#   xlim(-5,25)+
#   xlab("Winner's Receipts Advantage ($/vote)")+
#   ylab("Wining Margin") +
#   theme(axis.title = element_blank())
#
plt1 <- just_winners %>%
  filter(not(delta_votes == 0)) %>%
  ggplot(aes(x=per_vote_spending_advantage, y=delta_pct, col=INCU)) +
  geom_point() +
  xlim(-5,25)+
  xlab("Winner's Spending Advantage ($/vote)")+
  ylab("Wining Margin")
```

```
#  theme(axis.title.x = element_blank())+
#  theme(legend.position="none")
#
#
# plt2 <- just_winners %>%
#   filter(not(delta_votes == 0)) %>%
#   ggplot(aes(x=per_vote_indiv_contrib_advantage, y=delta_pct, col=INCU, alpha=I(need_help))) +
# #  ggplot(aes(x=delta_disb, y=delta_votes, col=INCU)) +
#   geom_point() +
#   xlim(-5,25)+
# #  xlab("Winner's Individidual Contribution Advantage ($/vote)")+
#   ylab("Wining Margin")+
#   theme(axis.title.x = element_blank(),
#        legend.position = c(1, 1),
#        legend.justification = c(1, 1))
#
#
# plt2b <- just_winners %>%
#   filter(not(delta_votes == 0)) %>%
#   filter(per_vote_indiv_contrib_advantage <0) %>%
#   ggplot(aes( y=delta_pct, col=INCU)) +
# #  ggplot(aes(x=delta_disb, y=delta_votes, col=INCU)) +
#   geom_segment(aes(x=per_vote_indiv_contrib_advantage,
#                   xend=(delta_pol_pty_contrib + delta_other_pol_cmte_contrib)/general_votes-d_x,
#                   yend=delta_pct),
#               size=0.3,col="grey50",
#                   arrow = arrow(length = unit(0.2, "npc"),type="closed")) +
#     geom_point(aes(x=per_vote_indiv_contrib_advantage)) +
#   geom_point(aes(x=(delta_pol_pty_contrib + delta_other_pol_cmte_contrib)/general_votes)) +
#   xlim(-5,25)+
#   ylim(0,.22)+
#   scale_y_continuous(breaks=seq(0, .2, .2))+
#   ylab("Wining Margin")+
#   theme(axis.title.x = element_blank())+
#   theme(legend.position="none")
```

```
plt2c <-
  ggplot() +
  geom_point(
    data = just_winners,
    mapping = aes(
      x = per_vote_indiv_contrib_advantage,
      y = delta_pct,
      color = INCU,
      alpha = I(need_help))) +
  geom_point(
    data = just_winners,
    mapping = aes(
      x = (delta_pol_pty_contrib + delta_other_pol_cmte_contrib)/general_votes,
      y = delta_pct,
      color = INCU,
      alpha = I(need_help))) +
  geom_segment(
    data = just_winners_needing_help,
```

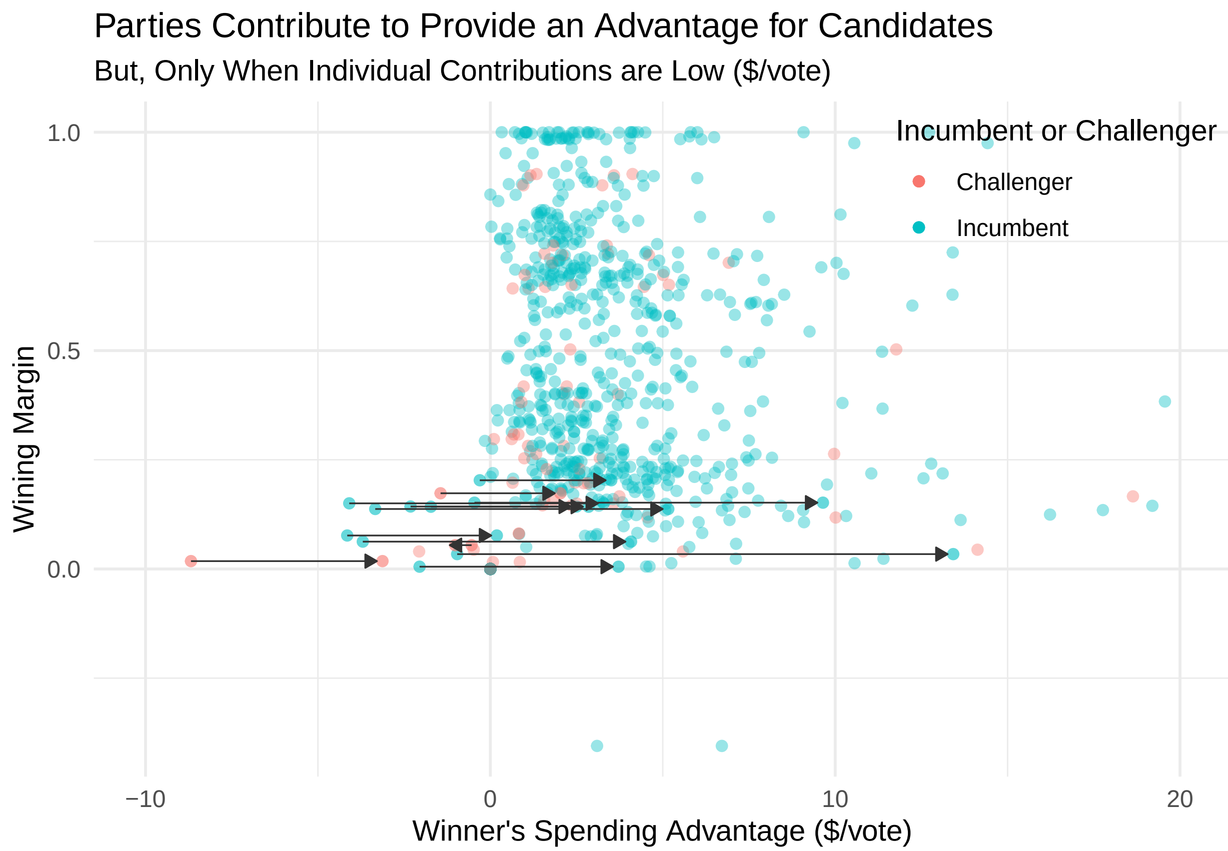
```
    mapping = aes(
      x = per_vote_indiv_contrib_advantage,
      xend=(delta_pol_pty_contrib + delta_other_pol_cmte_contrib)/general_votes - d_x,
      y = delta_pct,
      yend=delta_pct),
    size=0.3,col="grey20",
    arrow = arrow(length = unit(0.02, "npc"),type="closed")) +
  xlim(-10,20) +
  labs(
    title = "Parties Contribute to Provide an Advantage for Candidates",
    subtitle = "But, Only When Individual Contributions are Low ($/vote)",
    x = "Winner's Spending Advantage ($/vote)",
    y = "Wining Margin",
    color = "Incumbent or Challenger")+
  theme(legend.position = c(1, 1),
      legend.justification = c(1, 1))
```

```
# plt3 <- just_winners %>%
#     filter(not(delta_votes == 0)) %>%
# #  filter(not(delta_pol_pty_contrib == 0)) %>%
#   ggplot(aes(y=delta_pct)) +
#   geom_point(aes(x=per_vote_spending_advantage -  per_vote_party_advantage , col=INCU)) +
#   #geom_point(aes(x=log10(pol_pty_contrib)), col="grey20") +
#   #geom_point(aes(x=log10(delta_pol_pty_contrib)), col="grey70") +
#   #geom_point(aes(x=log10(delta_disb) , col=INCU))+
#   xlim(-5,25)+
#   xlab("Winner's Non-Party Spending ($/vote)")+
#   ylab("Wining Margin")
#
#
# plt4 <- just_winners %>%
#   filter(not(delta_votes == 0)) %>%
#   #filter(not(delta_pol_pty_contrib == 0)) %>%
#   ggplot(aes(y=delta_pct)) +
#   geom_point(aes(x=(delta_pol_pty_contrib + delta_other_pol_cmte_contrib)/general_votes, col=INCU)) +
#   #geom_point(aes(x=log10(pol_pty_contrib)), col="grey20") +
#   #geom_point(aes(x=log10(delta_pol_pty_contrib)), col="grey70") +
#   #geom_point(aes(x=log10(delta_disb) , col=INCU))+
#   xlim(-5,25)+
#   xlab("Winner's Party's Contribution Advantage ($/vote)")+
#   ylab("Wining Margin")
#
#
# plt5 <- just_winners %>%
#   filter(not(delta_votes == 0)) %>%
#   #filter(not(delta_pol_pty_contrib == 0)) %>%
#   ggplot(aes(y=delta_pct)) +
#   geom_point(aes(x=(per_vote_spending_advantage - delta_pol_pty_contrib + delta_other_pol_cmte_contri
#   #geom_point(aes(x=log10(pol_pty_contrib)), col="grey20") +
#   #geom_point(aes(x=log10(delta_pol_pty_contrib)), col="grey70") +
#   #geom_point(aes(x=log10(delta_disb) , col=INCU))+
#   xlim(-5,25)+
#   xlab("Winner's Non-Party Advantage ($/vote)")+
#   ylab("Wining Margin")
```

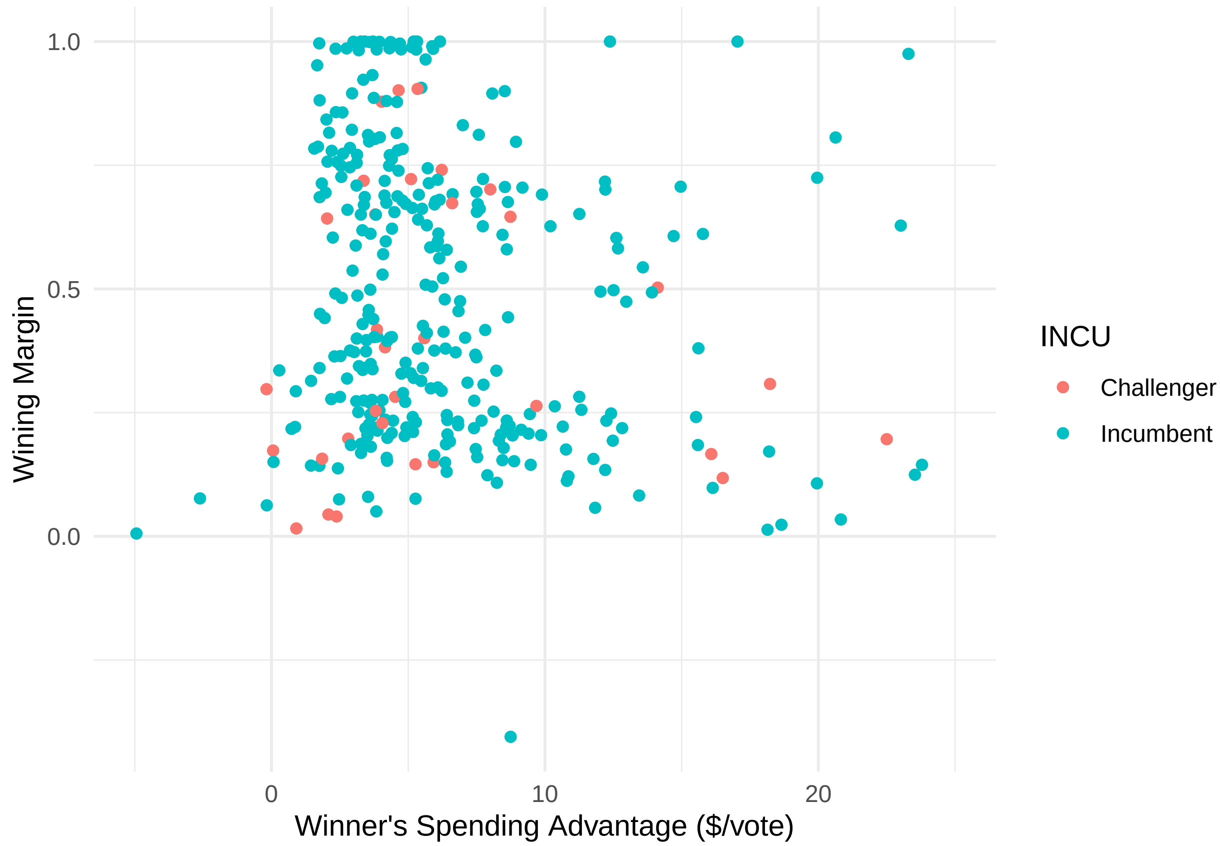
```r
#library(gridExtra)
#grid.arrange(plt2, plt2b, plt1, ncol=1)
#plot_grid(plt2, plt2b, plt1, align = "v", nrow = 3, rel_heights = c(0.41, 0.18, 0.41))
```

```r
plt2c
```

```
## Warning: Removed 4 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_segment).
```



We can also look at the spending advantage of all winning candidates.




```
paste(100 - round(sum(just_winners$per_vote_spending_advantage<0) /
                    length(just_winners$per_vote_spending_advantage)*100,2),
      "percent of winners outspend their rivals. ")
```

```
## [1] "98.4 percent of winners outspend their rivals. "
```

```
t.test(just_winners$per_vote_spending_advantage[just_winners$INCU=="Incumbent"])
```

```
##
##  One Sample t-test
##
## data:  just_winners$per_vote_spending_advantage[just_winners$INCU == "Incumbent"]
## t = 17.37, df = 378, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.026041 6.309186
## sample estimates:
## mean of x
##  5.667613
```

```
t.test(just_winners$per_vote_spending_advantage[just_winners$INCU=="Challenger"])
```

```
##
##  One Sample t-test
##
## data:  just_winners$per_vote_spending_advantage[just_winners$INCU == "Challenger"]
## t = 3.2602, df = 57, p-value = 0.001881
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  1.076342 4.503636
## sample estimates:
## mean of x
##  2.789989
```

From the t-test, we can see that winning Incumbents outspend their opponents by $5.64 per vote, 95% CI [$5.01 - $6.27]. Similarly, winning challengers outspend their opponents but by a lower margin: $2.79 per vote, 95% CI [$1.07 - $4.50].

```
pctt <- 100 - round(sum(just_winners$INCU=="Challenger")/length(just_winners$INCU) *100,2)
paste(pctt, "percent of Incumbents win. ")
```

```
## [1] "86.73 percent of Incumbents win. "
```

Here is an example from NY:

#3 Produce a Descriptive Model

Given what I have observed:

1. 98.4% of winners outspend their closet rival.
2. 86.7% of incumbents win.
3. Political parties of winners make up for a disadvantage in individual constributions with party contributions.
4. Winning incumbents outspend their opponents by $5.64 per vote, 95% CI [$5.01 - $6.27].
5. Winning challengers outspend their opponents but by a lower margin: $2.79 per vote, 95% CI [$1.07 - $4.50].

MODEL 1

$$Votes = \beta_0 + \beta_1 \cdot Incumbent + \beta_2 \cdot \log(\text{Total disbursements}) + \beta_3 \cdot party$$

β_0 will be the percentage of votes that a non incumbent with no spending advantage gets. β_1 will be the incremental votes that an incumbent gets. β_2 will be the incremental votes resulting increasing disbursements by 1% β_3 will be the incremental votes resulting from being associated with a party.

```
d_single_candidates$INCU <- as.factor(d_single_candidates$INCU)
model1 <- lm(general_votes ~ INCU + log(ttl_disb+1) + party, data=d_single_candidates_general)
summary(model1)
```

```
##
## Call:
## lm(formula = general_votes ~ INCU + log(ttl_disb + 1) + party,
##     data = d_single_candidates_general)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170473  -26695    -285   23828  633686
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6426      12598   0.510    0.610
## INCUIncumbent        54807       4514  12.142  < 2e-16 ***
## log(ttl_disb + 1)     9721       1039   9.354  < 2e-16 ***
## partyOther          -54989       8398  -6.548 1.06e-10 ***
## partyRepublican      -3092       3927  -0.787    0.431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 52770 on 782 degrees of freedom
## Multiple R-squared:  0.4641, Adjusted R-squared:  0.4613
## F-statistic: 169.3 on 4 and 782 DF,  p-value: < 2.2e-16
```

Model 1 does a pretty good job of explaining the varianes in General Election votes by campaign. Incumbency is worth an additional 54,850 votes, while being in Indepent costs a candidate about 55,000 votes as compared with a Democrat. Increasing spending by 1% provides an additional 9712 votes on average. The difference in Party benefit between Republicans and Democrats is not statistically significant.

MODEL 2

$$Votes = \beta_0 + \beta_1 \cdot Incumbent + \beta_2 \cdot \log(\text{Total disbursements}) + \beta_3 \cdot party + \beta_4 \cdot State$$

β_0 will be the percentage of votes that a non incumbent with no spending advantage gets. β_1 will be the incremental votes that an incumbent gets. β_2 will be the incremental votes resulting increasing disbursements by 1% β_3 will be the incremental votes resulting from being associated with a party. β_4 will be the incremental votes resulting from the election being in a specific State.

```
model2 <- lm(general_votes ~ INCU + log(ttl_disb+1) + party + state, data=d_single_candidates_general)
summary(model2)
```

```
##
## Call:
## lm(formula = general_votes ~ INCU + log(ttl_disb + 1) + party +
##     state, data = d_single_candidates_general)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -400259  -21244   -2087   21095  246654
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -15152.4    29958.6  -0.506  0.61316
## INCUIncumbent       57639.0     3585.3  16.076  < 2e-16 ***
## log(ttl_disb + 1)    8592.8      830.3  10.349  < 2e-16 ***
## partyOther         -80179.1     6771.6 -11.840  < 2e-16 ***
## partyRepublican     -1859.1     3010.8  -0.617  0.53711
## stateAL             39146.7    30433.8   1.286  0.19875
## stateAR             34064.3    33139.9   1.028  0.30434
## stateAS            -87832.6    36206.8  -2.426  0.01551 *
## stateAZ             23773.6    29781.7   0.798  0.42498
## stateCA              9341.2    28256.5   0.331  0.74105
## stateCO             57708.9    29901.5   1.930  0.05400 .
## stateCT             37006.7    30933.3   1.196  0.23195
## stateDC            110526.9    48510.6   2.278  0.02299 *
## stateDE            105167.8    39571.4   2.658  0.00804 **
## stateFL             48774.7    28463.3   1.714  0.08703 .
## stateGA             49553.2    29641.8   1.672  0.09501 .
## stateGU            -98506.5    39577.7  -2.489  0.01304 *
## stateHI             29020.4    36128.1   0.803  0.42208
## stateIA             54796.8    31260.7   1.753  0.08004 .
## stateID             51983.2    34306.0   1.515  0.13014
## stateIL             44391.6    28869.3   1.538  0.12456
## stateIN             35130.5    29590.0   1.187  0.23552
## stateKS             14798.5    31271.1   0.473  0.63619
```

```
## stateKY             60440.1    30419.5   1.987  0.04731 *
## stateLA            -27348.9    29422.7  -0.930  0.35293
## stateMA             70993.4    29707.4   2.390  0.01711 *
## stateMD             52892.7    29896.7   1.769  0.07728 .
## stateME             56768.2    34245.7   1.658  0.09781 .
## stateMI             48449.3    29018.2   1.670  0.09543 .
## stateMN             67745.7    29814.0   2.272  0.02336 *
## stateMO             60824.8    30068.1   2.023  0.04345 *
## stateMP            -45721.6    48925.7  -0.935  0.35035
## stateMS             42019.6    32418.0   1.296  0.19532
## stateMT            102167.5    39554.1   2.583  0.00999 **
## stateNC             61759.2    29095.8   2.123  0.03412 *
## stateND             38633.4    36175.3   1.068  0.28590
## stateNE             28077.7    34264.1   0.819  0.41280
## stateNH             15176.4    34294.6   0.443  0.65824
## stateNJ             25666.6    29235.8   0.878  0.38028
## stateNM              2479.5    33092.1   0.075  0.94029
## stateNV              6277.2    30921.0   0.203  0.83919
## stateNY             20233.1    28567.8   0.708  0.47902
## stateOH             47933.6    28915.9   1.658  0.09781 .
## stateOK             32006.1    33121.8   0.966  0.33421
## stateOR             75094.0    31723.2   2.367  0.01819 *
## statePA             63264.2    28945.9   2.186  0.02916 *
## statePR            449290.8    36341.5  12.363  < 2e-16 ***
## stateRI            -12461.9    34294.7  -0.363  0.71643
## stateSC             25494.1    30062.7   0.848  0.39670
## stateSD             52623.5    39541.3   1.331  0.18366
## stateTN             30686.9    30082.2   1.020  0.30802
## stateTX             10093.8    28521.8   0.354  0.72352
## stateUT              4935.3    31263.3   0.158  0.87461
## stateVA             58342.6    29268.6   1.993  0.04660 *
## stateVI           -137362.2    48520.6  -2.831  0.00477 **
## stateVT            105358.6    48497.6   2.172  0.03014 *
## stateWA             38725.9    29673.2   1.305  0.19228
## stateWI             60355.5    29593.6   2.039  0.04176 *
## stateWV             -5518.3    31729.2  -0.174  0.86198
## stateWY             28464.0    36179.8   0.787  0.43169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39540 on 727 degrees of freedom
## Multiple R-squared:  0.7203, Adjusted R-squared:  0.6976
## F-statistic: 31.73 on 59 and 727 DF,  p-value: < 2.2e-16
```

Model 2 does an even better job of explaining the varianes in General Election votes by campaign. The coefficients on Incumbency, and Disbursements are similar to Model 1 (Party=I is -80K). The coefficients on States are mostly not statistically significant, but jointly, they improve adjusted R-square to nearly 70%!

I interpret this as is illustrated in the graphs above. What is important is not just to spend alot, but to outspend your opponent. By doing per-state analysis, we are essentially comparing spending among candidates within each state, making their spending relative to each other more important. It can also reflect the percentage voter turnout in the state, and thus the total number of votes. Further each state has its own rules for primaries and runnoffs, so these differences may be captured in the State coefficients.

I chose not to model spending relative to the opponent, due to causality concerns. That is, outspending requires knowledge of the opponent's spending. It is possible to know this until late in the campaign based on FEC reporting requirements, but there is likely large amounts of spending very late in the campaign that is not based on knowing the final spending by the opponent, as it would not be possible to know this.