Q4 Federal Elections Exercise

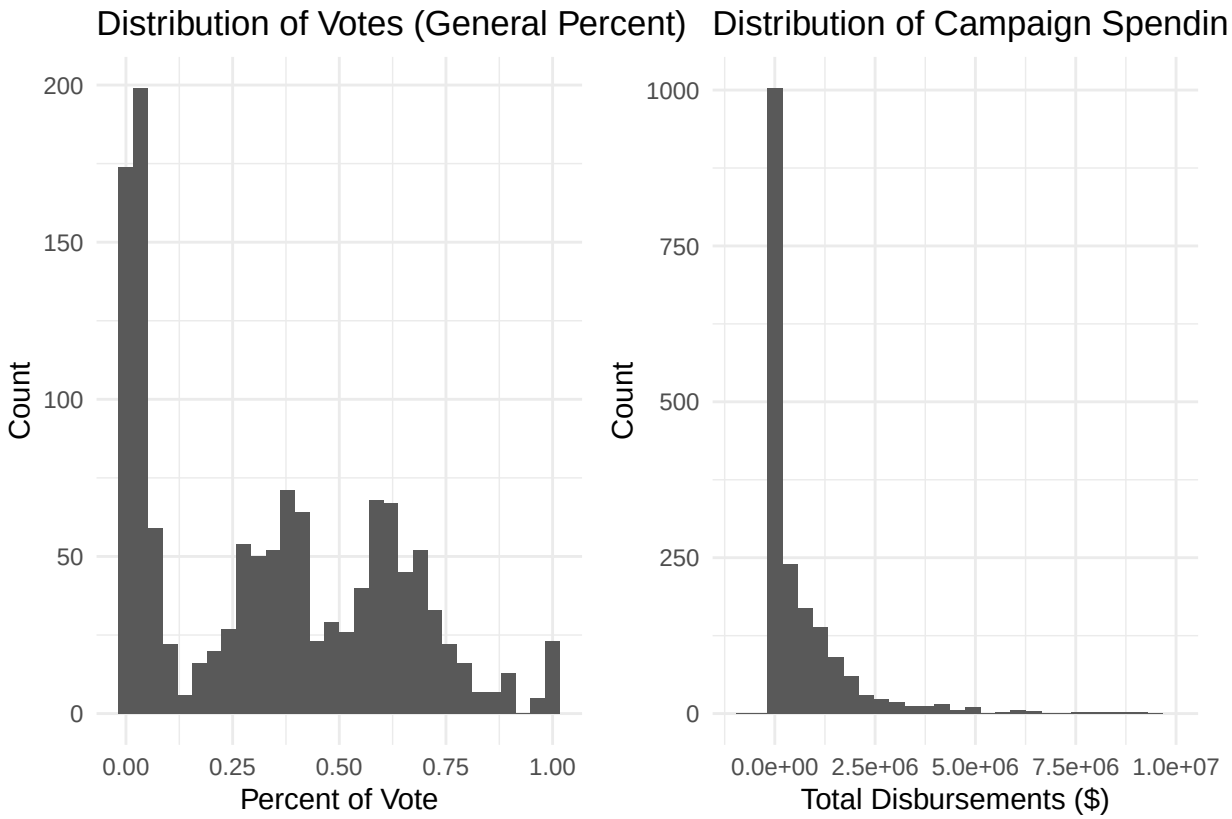
Q4.1 Exploring - Make Histograms

```
general_percent_chart <- results_house %>%
  ggplot(aes(x = general_percent)) +
  geom_histogram() +
  labs(title = "Distribution of Votes (General Percent)",
       x = "Percent of Vote",
       y = "Count") +
  theme_minimal()

ttl_disb_chart <- campaigns %>%
  ggplot(aes(x = ttl_disb)) +
  geom_histogram() +
  xlim(-1000000, 10000000) +
  labs(title = "Distribution of Campaign Spending (Total Disbursements)",
       x = "Total Disbursements ($)",
       y = "Count") +
  theme_minimal()

general_percent_chart | ttl_disb_chart
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Multi-modal distribution, so there's likely data that's smushed-together on one graph
```

Q4.2 Exploring - Build a Data Frame pt 1

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

This comes with 1342 rows of data.

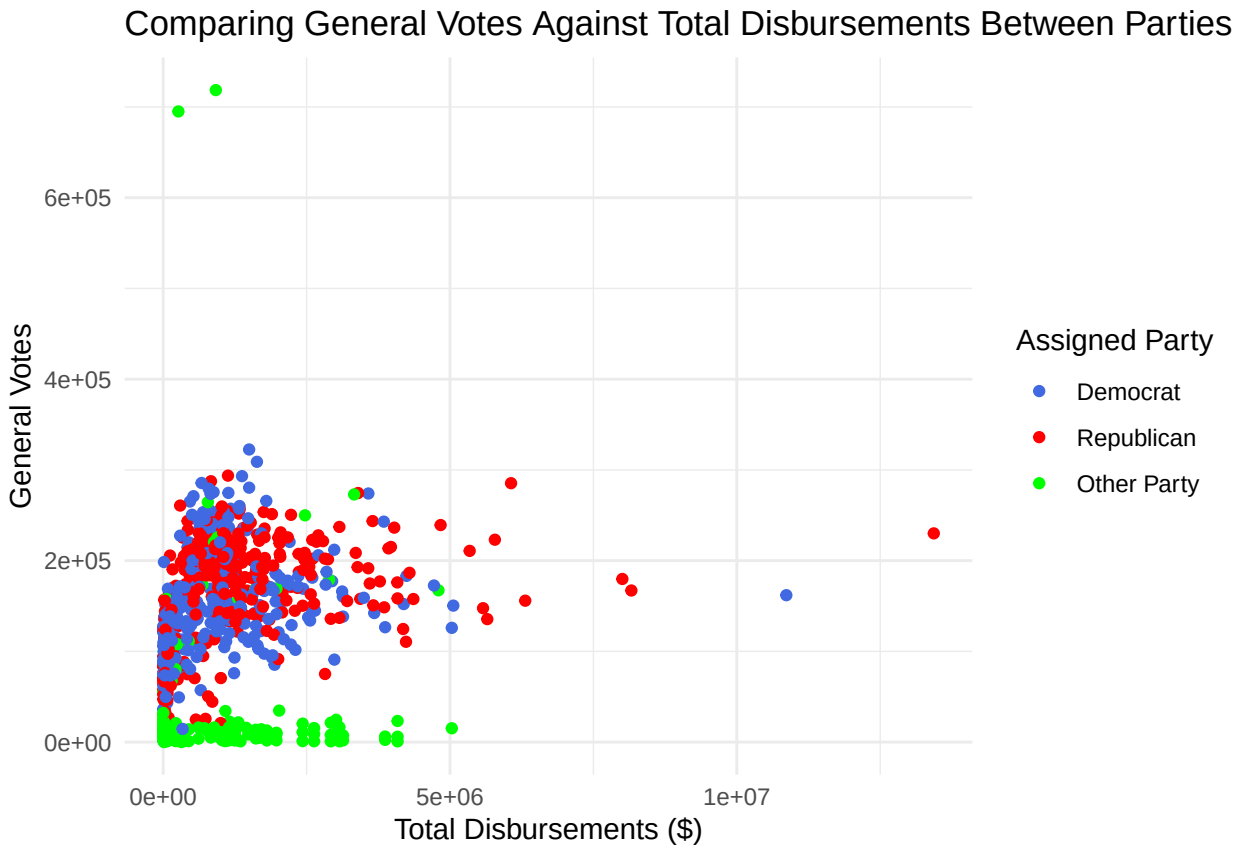
Q4.3 Exploring - Build a Data Frame pt 2

There are 37 columns of data.

Q4.4 Exploring - Indicator Variables and Scatter Plot

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

results_campaigns %>%
  mutate(candidate_party = factor(candidate_party,
                                  levels = c("Democrat",
                                             "Republican",
                                             "Other Party"))) %>%
  ggplot(aes(x = ttl_disb, y = general_votes, color = candidate_party)) +
  geom_point() +
  scale_color_manual(values = c("Democrat" = "royalblue",
                                "Republican" = "red",
                                "Other Party" = "green")) +
  labs(title = "Comparing General Votes Against Total Disbursements Between Parties",
       x = "Total Disbursements ($)",
       y = "General Votes",
       color = "Assigned Party") +
  theme_minimal()
```



Q4.5 Regression - Evaluate large sample assumptions

Regression:

```
ls_model <- lm(general_votes ~ ttl_disb + candidate_party, data = results_campaigns)

# Formula for kurtosis (measure of "tailedness")
ls_model_res <- ls_model$residuals
ls_model_kurtosis <- (sum((ls_model_res - mean (ls_model_res)) ^ 4) /
                        length(ls_model_res)) / (var(ls_model_res) ^ 2) - 3

# Variance-covariance matrix to make sure covariances between Xi's are finite
vcov(ls_model)
```

```
##                               (Intercept)      ttl_disb
## (Intercept)                  1.399882e+07 -2.828523e+00
## ttl_disb                    -2.828523e+00  3.034353e-06
## candidate_partyOther Party -1.111328e+07 -2.669924e-01
## candidate_partyRepublican  -1.040906e+07 -1.022467e+00
##                            candidate_partyOther Party candidate_partyRepublican
## (Intercept)                             -1.111328e+07             -1.040906e+07
## ttl_disb                                -2.669924e-01             -1.022467e+00
## candidate_partyOther Party               3.787325e+07              1.145213e+07
## candidate_partyRepublican                1.145213e+07              2.335531e+07
```

```
# Covariance of general_votes and ttl_disb to test X and Y cov
cov(results_campaigns$general_votes, results_campaigns$ttl_disb, use = "complete.obs")
```

```
## [1] 22790357478
```

Assumptions:

1. IID data

- IID data likely does not exist because a relationship exists between those fighting for the same votes in a district. For example, if the total votes for a party is 1 million, and the Democrat has 700,000 of them, the Republicans and individuals from Other Parties can get at most 300,000 votes combined. There's likely also geographic clustering effects, or strategic interactions among those of the same party, further violating IID.

2. Unique BLP exists

- A BLP exists when $\text{cov}[X_i, X_j]$ and $\text{cov}[X_i, Y]$ are finite (no heavy tails): this condition is questionable, as calculating kurtosis (measure of "tailedness") through the formula written above gives a value of 25.96, indicating extremely heavy tails. This might indicate that a BLP may not exist; however, looking at the data itself shows that these heavy tails may not be infinite. Calculating $\text{cov}[X_i, X_j]$ for all the X's, we see in the vcov matrix that the values do appear to finite - though large - and calculating $\text{cov}[X_i, Y]$ for just ttl_disb (since candidate_party is categorical), we see that it's also quite large (2.3e10) - though technically still finite. This leaves some room for debate on whether the BLP does exist, and if we wanted to ensure that it does (for future tests), we may try using a log transform to remove these heavy tails.
- A BLP is unique when there is no perfect collinearity ($E[X^T X]$ is invertible): this condition is likely fulfilled, as any X_i cannot be written as a linear combination of the other X's (and therefore the X's have unique variation).

Q4.5 (Additional code upload and regression summary)

```r
# Code used to create candidate_party from earlier:

# results_campaigns <- results_campaigns %>%
#   mutate(candidate_party = case_when(
#     party == 'REP' ~ 'Republican',
#     party == 'DEM' ~ 'Democrat',
#     TRUE ~ 'Other Party'
#   ))
```

```r
# Regression results:
summary(ls_model)
```

```
##
## Call:
## lm(formula = general_votes ~ ttl_disb + candidate_party, data = results_campaigns)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146242  -38135  -11551   37488  679443
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.401e+05  3.741e+03  37.438  < 2e-16 ***
## ttl_disb                     1.326e-02  1.742e-03   7.614  6.9e-14 ***
## candidate_partyOther Party -1.131e+05  6.154e+03 -18.378  < 2e-16 ***
## candidate_partyRepublican   6.534e+03  4.833e+03   1.352    0.177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64490 on 876 degrees of freedom
##   (462 observations deleted due to missingness)
## Multiple R-squared:  0.3589, Adjusted R-squared:  0.3567
## F-statistic: 163.5 on 3 and 876 DF,  p-value: < 2.2e-16
```

Q4.6 Regression - Build a stargazer table

This first table assumes homoskedasticity (constant variance of errors), using vcov for the standard errors:

```
se <- sqrt(diag(vcov(ls_model)))

stargazer(ls_model,
          type = "latex",
          title = "Linear Regression of General Votes on Campaign Spending and Candidate Party",
          covariate.labels = c("Total Disbursements",
                               "Other Party",
                               "Republican"),
          dep.var.labels = "General Votes",
          se = list(se),
          digits = 2,
          header = FALSE)
```

Table 1: Linear Regression of General Votes on Campaign Spending and Candidate Party

	Dependent variable:
	General Votes
Total Disbursements	0.01^{***}
	(0.002)
Other Party	$-113{,}100.60^{***}$
	(6,154.12)
Republican	6,534.04
	(4,832.73)
Constant	$140{,}075.90^{***}$
	(3,741.50)
Observations	880
R^2	0.36
Adjusted R^2	0.36
Residual Std. Error	64,486.84 (df = 876)
F Statistic	163.50^{***} (df = 3; 876)
Note:	*p<0.1; **p<0.05; ***p<0.01

With an R^2 of 0.36, which means 64% of our variance is unexplained.

A second table with robust standard errors for the large sample assumption is shown below:

This second table assumes heteroskedasticity (non-constant variance of errors) for the large sample assumptions, using vcovHC for robust standard errors:

```
robust_se <- sqrt(diag(vcovHC(ls_model, type = "HC0")))

stargazer(ls_model,
          type = "latex",
          title = "Linear Regression of General Votes on Campaign Spending and Candidate Party",
          covariate.labels = c("Total Disbursements",
                               "Other Party",
                               "Republican"),
          dep.var.labels = "General Votes",
          se = list(robust_se),
          digits = 2,
          header = FALSE)
```

Table 2: Linear Regression of General Votes on Campaign Spending and Candidate Party

	Dependent variable:
	General Votes
Total Disbursements	0.01***
	(0.002)
Other Party	−113,100.60***
	(8,154.31)
Republican	6,534.04
	(4,210.89)
Constant	140,075.90***
	(3,305.61)
Observations	880
R^2	0.36
Adjusted R^2	0.36
Residual Std. Error	64,486.84 (df = 876)
F Statistic	163.50*** (df = 3; 876)
Note:	*p<0.1; **p<0.05; ***p<0.01

Q4.7 Regression - Money's Relationship with Votes

This is done with robust standard errors from the large sample assumptions, though the results are similar if assuming homoskedasticity.

```
coeftest(ls_model, vcov = vcovHC(ls_model, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                               Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)                 1.4008e+05  3.3056e+03  42.3751 < 2.2e-16 ***
## ttl_disb                    1.3262e-02  1.9352e-03   6.8531 1.362e-11 ***
## candidate_partyOther Party -1.1310e+05  8.1543e+03 -13.8700 < 2.2e-16 ***
## candidate_partyRepublican   6.5340e+03  4.2109e+03   1.5517    0.1211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ttl_disb does seem to have a relationship with general votes; specifically, the coefficient of ttl_disb describes how much money you will spend for each additional vote (0.013 votes for every dollar spent). The p-value for the ttl_disb coefficient is 1.362e-11, which is far below the usual significance level of 0.05, indicating strong statistical significance; however, while the coefficient of ttl_disb is statistically significant, its practical significance is small, as an increase of $1 in disbursements leads to the aforementioned 0.013 additional votes.

```
# Using vcov shows similar results in terms of significance
coeftest(ls_model, vcov = vcov(ls_model))
```

```
##
## t test of coefficients:
##
##                               Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)                 1.4008e+05  3.7415e+03  37.4384 < 2.2e-16 ***
## ttl_disb                    1.3262e-02  1.7419e-03   7.6135 6.899e-14 ***
## candidate_partyOther Party -1.1310e+05  6.1541e+03 -18.3780 < 2.2e-16 ***
## candidate_partyRepublican   6.5340e+03  4.8327e+03   1.3520    0.1767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q4.8 Regression - Party's Relationship with Votes

This is done with robust standard errors from the large sample assumptions, though the results are similar if assuming homoskedasticity.

```
short_model <- lm(general_votes ~ ttl_disb, data = results_campaigns)
waldtest(ls_model, short_model, vcov = vcovHC(ls_model, type = "HC0"))
```

```
## Wald test
##
## Model 1: general_votes ~ ttl_disb + candidate_party
## Model 2: general_votes ~ ttl_disb
##   Res.Df Df      F    Pr(>F)
## 1    876
## 2    878 -2 109.08 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the Wald test for the addition of candidate_party is $< 2.2e\text{-}16$, indicating strong statistical significance, which means the addition of candidate_party provides significant additional information about general_votes. Practically, with all other things equal and using the values from Q4.6, being Democrat (intercept) would net roughly 140,000 additional votes, being from the "Other Party" would net roughly 27,000 (140,000 - 113,000) additional votes, and being Republican would net roughly 146,500 (140,000 + 6,500) additional votes, though the result for Republicans is not significant (p-value above 0.05); therefore, while candidate_party does provide additional information, the t-test coefficients are only significant for Democrats and "Other Party" members.

```
# Using vcov shows similar results in terms of significance
waldtest(ls_model, short_model, vcov = vcov(ls_model))
```

```
## Wald test
##
## Model 1: general_votes ~ ttl_disb + candidate_party
## Model 2: general_votes ~ ttl_disb
##   Res.Df Df      F    Pr(>F)
## 1    876
## 2    878 -2 210.58 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```