Assignment 4

Sneh Patel (20525801), Parth Shah (20759634)
Gunjan Shah (20761635), Neha Bhatia (20762558)

2nd April 2019

Group Name: Fantastic 4
Dataset: We will be using data from past United States presidential elections collected from The American Presidency Project (https://www.presidency.ucsb.edu/) to run an analysis.

We first read and identify the dataset and look at the variables. The dataset has 13 features. It describes the election data of U.S. It gives the vote share of each party, in each state of that election year, contry state and also, describes the challenger and incumbent party at the time of election.

The United States has a total of 50 states from which data exists. However, only 11 of the states are Confederacy states and 25 of them are Union states. This clearly doesn't add up to 50. These missing states in the dataset are coded as NA for the civil_war variable. This is because these states were not a part of the United States at the time of the civil war.

We averaged the data over the year and states for analysis.

```
## # A tibble: 4 x 8
##   incumbent_party civil_war    mean    sd     n    se se_high se_low
##   <fct>           <fct>       <dbl> <dbl> <int> <dbl>   <dbl>  <dbl>
## 1 democrat        confederacy  45.9  3.43    11 1.03     47.0   44.9
## 2 democrat        union        52.4  4.83    25 0.966    53.3   51.4
## 3 republican      confederacy  58.1  2.24    11 0.676    58.8   57.5
## 4 republican      union        50.0  4.24    25 0.849    50.8   49.1
```

The above mentioned data shows the average vote share over the year of parties in contry states.

Analysis

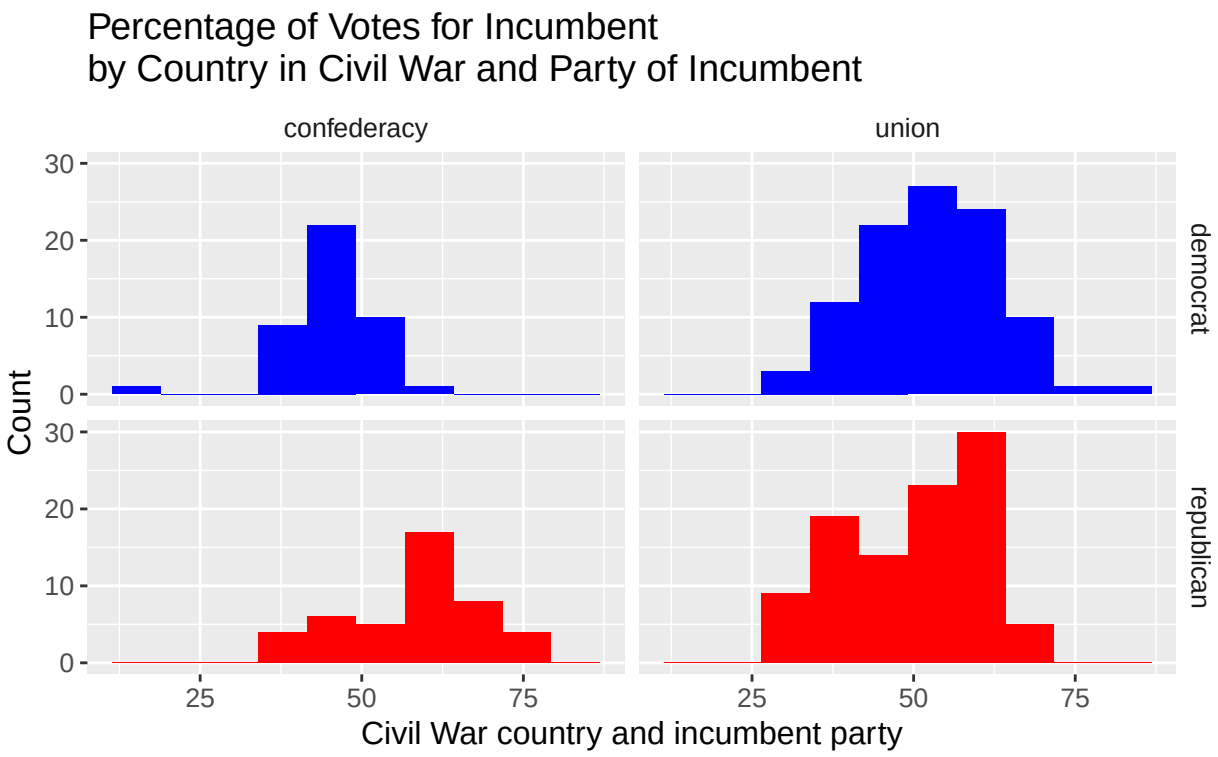
We want to assess the following:
1. Does the incumbent party (either Republican or Democrat) have an effect on the percentage of vote share during elections?
2. How do the civil war (either Confederate or Union) states vote for incumbents?
3. Does there exist some interaction between incumbency and civil war states?

For preprocessing we decided to drop any states that have an NA for the civil_war variable.

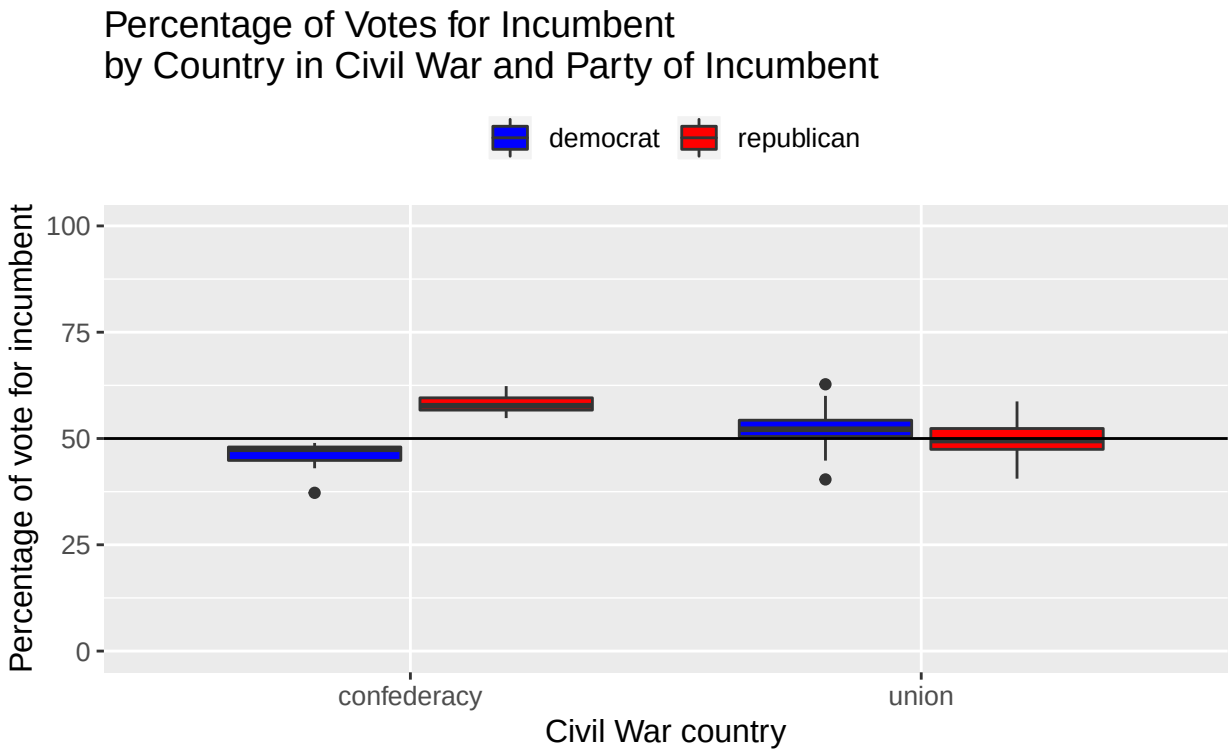
For the analysis, we want to conduct ANOVA between 3 variables:
1. "party" (Democrat or Republican)
2. "civil war country" (Union or Confederate)
3. "incumbency of party" (Incumbent or Non-Incumbent)

Step 1: Initial Analysis

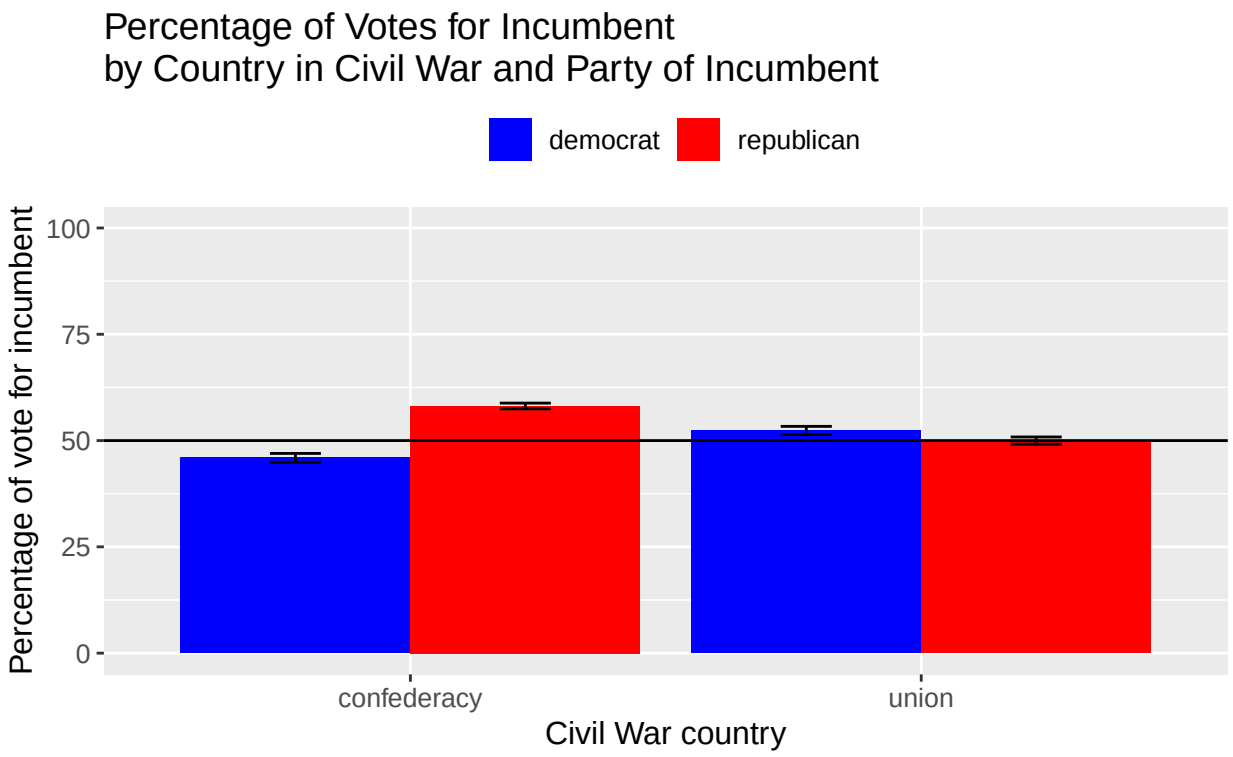
Below is the histogram plot for the vote share for each party in each of the civil war states (Union or Confederate). The x-axis denotes the vote share party in percentage (binned at 10) and the y-axis denotes the frequency of times each vote share percentage occurred from 1964 - 2012 (from the dataset).



The boxplots below show a party's performance (in vote share percentage) when they are incumbent in either states (Union or Confederate).



We can see that Republicans tend to do better than Democrats in Confederate states and Democrats do slightly better than Republicans in Union states. The below barplot shows this more clearly:



Step 2: Assumptions

For ANOVA, the following assumptions have to be met:
1. Homoscedasticity
2. Independent observations
3. Distributions within groups are normally distributed

We will test assumption 1. after building the model. As seen by histogram above, assumption 3. has been met. We confirmed this by generating the Q-Q Plot. For assumption 2. we will proceed by assuming that the data has independent observations.

Step 3: Testing

```
## BUILD MODELS ####
# ANOVA (base R)
incumbent.aov = aov(perc_incumbent_mean ~ incumbent_party * civil_war, data = data_stats)

incumbent.aov_sum = summary(incumbent.aov)
incumbent.aov_sum
```

```
##                           Df Sum Sq Mean Sq F value   Pr(>F)
## incumbent_party            1   77.2    77.2   4.526    0.037 *
## civil_war                  1   11.0    11.0   0.647    0.424
## incumbent_party:civil_war  1  811.5   811.5  47.581 2.17e-09 ***
## Residuals                 68 1159.7    17.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ezANOVA
incumbent.ezanova = ezANOVA(data.frame(data_stats),
                            dv = perc_incumbent_mean,
                            wid = state,
                            within = incumbent_party,
                            between = civil_war,
                            type = "III")
```

```
## Warning: Data is unbalanced (unequal N per group). Make sure you specified
## a well-considered value for the type argument to ezANOVA().
```

```r
incumbent.ezanova
```

```
## $ANOVA
##                       Effect DFn DFd         F            p p<.05
## 2                  civil_war   1  34  3.878659 5.709558e-02
## 3           incumbent_party   1  34 11.756521 1.605528e-03     *
## 4 civil_war:incumbent_party   1  34 25.954476 1.295247e-05     *
##           ges
## 2 0.009422014
## 3 0.240669449
## 4 0.411667383
```

The results above show that the incumbent party variable as well as, the interaction between incumbent party and civil war states has a significant effect on the vote share percentage. Whereas, it shows that civil war has no effect since $p > 0.05$. We should test for homoscedasticity on this model to fulfil assumption 1. from above:

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.3789 0.2567
##       68
```

We can see that the model passes this assumption as we get a p-value that is not significant. This means that we cannot reject the null hypothesis that the population variances are equal (homoscedasticity).

We've seen that there is a significant interaction between incumbent party and civil war states variables. Hence, it only makes sense to follow-up on what this interaction really means. We will proceed to conduct a series of t-tests (four in total) between the following:
1. Union states, Democrat vs. Republican (this will be dependent t-test)
2. Confederate states, Democrat vs. Republican (this will be dependent t-test)
3. Democrat incumbents, Union vs. Confederacy (this will be independent t-test)
4. Republican incumbents, Union vs. Confederacy (this will be independent t-test)

We checked all assumptions on the variables and they meet all assumptions of independent and dependent t-tests.

```
##
##  Paired t-test
##
## data:  data_union_stats$democrat and data_union_stats$republican
## t = 1.3616, df = 24, p-value = 0.186
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.229182  5.995182
## sample estimates:
## mean of the differences
##                   2.383
```

```
##
##  Paired t-test
##
## data:  data_confederacy_stats$democrat and data_confederacy_stats$republican
## t = -7.5325, df = 10, p-value = 1.987e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.799975  -8.586389
## sample estimates:
## mean of the differences
##               -12.19318
```

```
##
##  Welch Two Sample t-test
##
## data:  perc_incumbent_mean by civil_war
## t = -4.5485, df = 26.598, p-value = 0.0001056
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.344743 -3.531985
## sample estimates:
## mean in group confederacy       mean in group union
##                  45.93864                  52.37700
```

```
##
##  Welch Two Sample t-test
##
## data:  perc_incumbent_mean by civil_war
## t = 7.501, df = 32.61, p-value = 1.363e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.92956 10.34608
## sample estimates:
## mean in group confederacy       mean in group union
##                  58.13182                  49.99400
```

As we've done four tests we are going to use Bonferroni correction and divide our original p-value for significance (0.05) by four, giving us a new p-value of 0.0125.

The results show:

1. Union states, Democrat vs. Republican
   The t-test shows a non-significant p-value hence, the means are not significantly different from each other (this can be confirmed by by the boxplots above) between the two parties for the Union states. The effect size is fairly small (r=0.268).

2. Confederate states, Democrat vs. Republican
   The t-test shows a significant p-value hence, the means are significantly different from each other (this can be confirmed by by the boxplots above) between the two parties for the Confederate states. The effect size is very strong (r=0.922).

3. Democrat incumbents, Union vs. Confederacy
   The t-test shows a significant p-value hence, the means are significantly different from each other (this can be confirmed by by the boxplots above) between the two types of civil war states for the Democrats. The effect size is strong(r=0.661).

4. Republican incumbents, Union vs. Confederacy
   The t-test shows a significant p-value hence, the means are significantly different from each other (this

can be confirmed by by the boxplots above) between the two types of civil war states for the Republicans The effect size is strong (r=0.796).

Step 4: Conclusion

There was a significant interaction effect between the incumbent party and contry states, on the percentage of vote. ($F(1, 34) = 25.954476$, $p<.05$).

The Republicans tend to perform better in Confederacy states compared to Democrates ($t = -7.5325$, $df = 10$, $p<0.0125$) and compared to their performace in Union States($t = 7.501$, $df = 32.61$, $p<0.0125$). The vote share for Democrates are higher in Union states than in confederacy ($t = -4.5485$, $df = 26.598$, $p<0.0125$). However, in the Union state, both parties vote share are significantly differentv ($t=1.3616$, $df = 24$, $p>0.0125$).

Contribution of members:

- Sneh Patel (20525801) & Parth Shah (20759634):

1. Analysed, cleaned and merged the dataset
2. Created the hypothesis questions
3. Ran t-tests and checked its assumptions.
4. Wrote the report
5. Understood and checked Gunjan & Neha's work on project

- Gunjan Shah (20761635) & Neha Bhatia (20762558)

1. Filtered dataset to plot graphs and run analysis.
2. Applied the anova.
3. Checked the assumptions of it.
4. Formatted the report.
5. Understood and checked Sneh & Parth's work on project