My title[*]

My subtitle if needed

First author Another author

1 December 2023

First sentence. Second sentence. Third sentence. Fourth sentence.

```
election_data <- read.csv(here("outputs/data/election_data_clean.csv"))
covid_data <- read.csv(here("outputs/data/covid_data_clean.csv"))
acs_data <- read.csv(here("outputs/data/acs_data_clean.csv"))
covid_election_data <- read.csv(here("outputs/data/covid_election.csv"))
data <- read.csv(here("outputs/data/merged_data.csv"))
republican_data <- data %>%
  filter(party == "Republican")
```

1 Introduction

You can and should cross-reference sections and sub-sections.

The remainder of this paper is structured as follows. Section 2....

2 Data

In this paper, the data used mainly consist the voting patterns for 2020 US Federal Election, COVID infection rate and socio-economic variables in each county of US. The voting data is taken from the MIT Election Data Science Lab which includes the voting for each party in county level at each Federal Election from 2000 to 2020. For the COVID data, it is taken from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The socio-economic data is taken from American Community Survey.

[*] Code and data are available at: LINK.

2.1 MIT Election Data Science Lab

The MIT Election Data Science Lab is a Lab at MIT which collect and analyzing the election data to supports advances in election science. Their collected data is not limited to the presidential election results, but also covers the midterm or US senate results in either state or county level.

The data abstracted from MITEDSL in this paper is the "County Presidential Election Returns 2000-2020." It contains the number of votes for each party and the total votes at each US county from 2000 to 2020 US Federal Election. In this paper, I will only use the data for the 2020 US Election and focus on the counties on the mainland of US. The Table 1 summarizes the important variables.

Table 1: The summary of voting patterns in 2020 US Federal Presidential Election

Party	Candidate	Total Votes	Mean Pct Vote	Median Pct Vote
Democrat	JOSEPH R BIDEN JR	80974742	0.33	0.30
Republican	DONALD J TRUMP	73988825	0.65	0.68
Libertarian	JO JORGENSEN	1798188	0.01	0.01
Other	OTHER	790507	0.01	0.00
Green	OTHER	378867	0.00	0.00

The Table 1 summarizes the voting patterns for each party during the 2020 US Presidential Election. Undoubtedly, the two most popular parties are the Democrat and the Republican. However, it seems that even though Republican has almost double mean percent vote than the Democrat, the total votes for Republican is less than the Democrat. This pattern may indicate that more people living at the high-population states such as California votes for Biden. This finding aligns with the fact that Biden defeated Trump in 2020. I will explain more latter.

2.2 Center for Systems Science and Engineering at JHU

The Center for Systems Science and Engineering at JHU is the center at the Department of Civils and Engineering to collect the local, national and global multidimensional data including medicine, health care, disaster response etc. During the pandemic, they collected the US and global COVID cases and deaths and report them on their GitHub. Their data is summarized by daily reports, ranging from April 12, 2020 to March 9, 2023.

To illustrate the impact of COVID on the 2020 Election to the greatest extend, the COVID data used in this paper will be the daily report on November 3, 2020 which is the Election

Table 2: **?(caption)**

```
# A tibble: 2 x 6
  winning_party   case deaths   inf  mort income
  <chr>          <dbl>  <dbl> <dbl> <dbl>  <dbl>
1 Democrat      10855.  309.  2939.  76.6 83976.
2 Republican     1482.   29.7 2939.  55.7 69969.
```

Day [citation]. This report contains the aggregated cases and deaths for each each country and counties in US.

2.3 American Community Survey (ACS)

The American Community Survey is the survey conducted by the U.S. Census Bureau. The survey contains a variety of socio-economic variables for each county. Although there are different data tables from ACS, I will take 2021 five years estimate of DP02, DP03 and DP05.

These three tables cover the social, economic and demographic characteristics of each county in US. By referring my previous research where I found the variables which have the highest correlations to the COVID mortality rate, I will use the same variables in this paper. That said, I will extract the data regarding the educational attainment from DP02, especially the proportion of people having at least a bachelor degree. In terms of economic perspective, I will take the proportion of people with private health insurance and the mean household income. Lastly, regarding the demographic factors, I will use the total population, the proportions of children and individuals above 85. I will also take the White and Black people percentage as well.

2.4 Data Cleaning

Combining the three data sources, **?@tbl-covid** shows the COVID cases and deaths, as well as the infection and mortality rate for the county voting for the Democrat and Republican party, respectively. Surprisingly, the average case and deaths of COVID of the counties voting for the Democrat is significantly higher than than the Republican party. However, the values of infection and mortality rate does not show the same dramatic difference. It also emphasizes that the counties (states) voting for the Democrat are population intensive and the counties having more residence generally have higher mean household income level. We can observe this from the average income where the "Democratic" counties have about $13k mean income higher than the "Republic" ones.

Figure 1 compares the distribution of county income levels for the two parties. In general, counties voting for the Democrat has a higher income levels than the Republican. In addition, we can observe that the counties are intensively concentrated around income levels $50k to

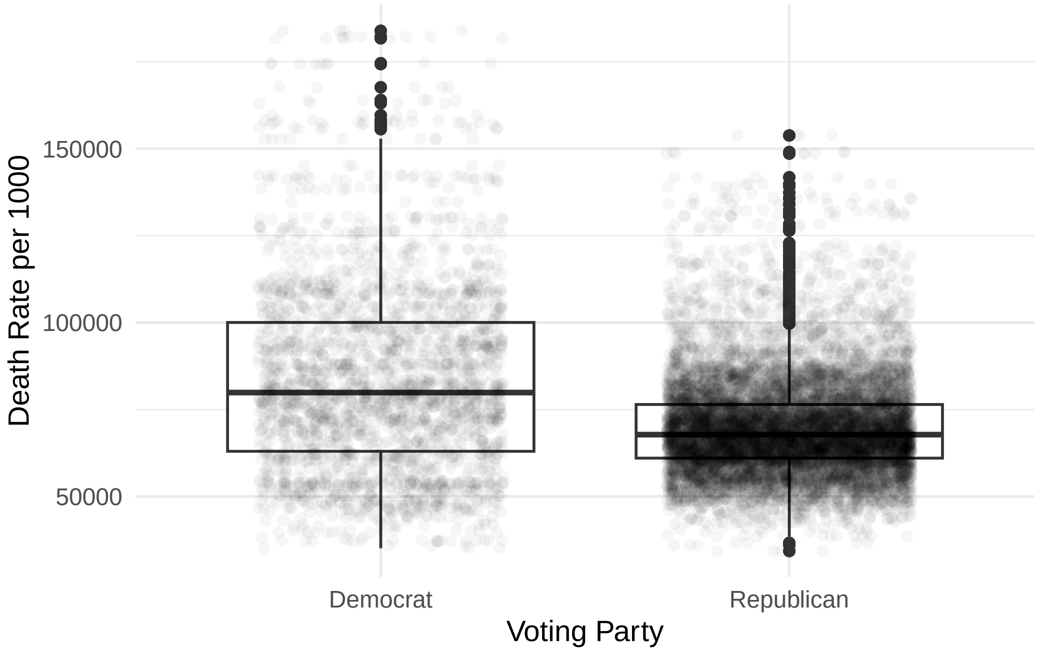

Figure 1: Summary of income levels for counties voting for Democratic and Republican

$75k, compared to the Democrat where the counties are approximately uniform distributed at each income level. Furthermore, there is barely no counties having at least 150k voting for the Republican party. Both **?@tbl-covid** and Figure 1 indicates that the Democrat is in favor of rich counties but poorer counties for Republican. Consequently, the richer counties (states) usually have more electoral votes than the poorer ones, this may provide insighes why Trump lost.

2.5 Data Limitation

Although the merged data covers the voting patterns, socio-economic characteristics and COVID infection rate by county, some of the counties are missing. From the above two maps, the counties on the middle west shows a grey color, indicating that the data for these counties are absent. This absence of data may increase the bias and hence influence the accuracy of the data analyses.

3 Methods

This paper aims to conduct the casual inference between the 2020 voting patterns and COVID, to find whether Trump can contribute his loss to the COVID. To find the casual effect in the

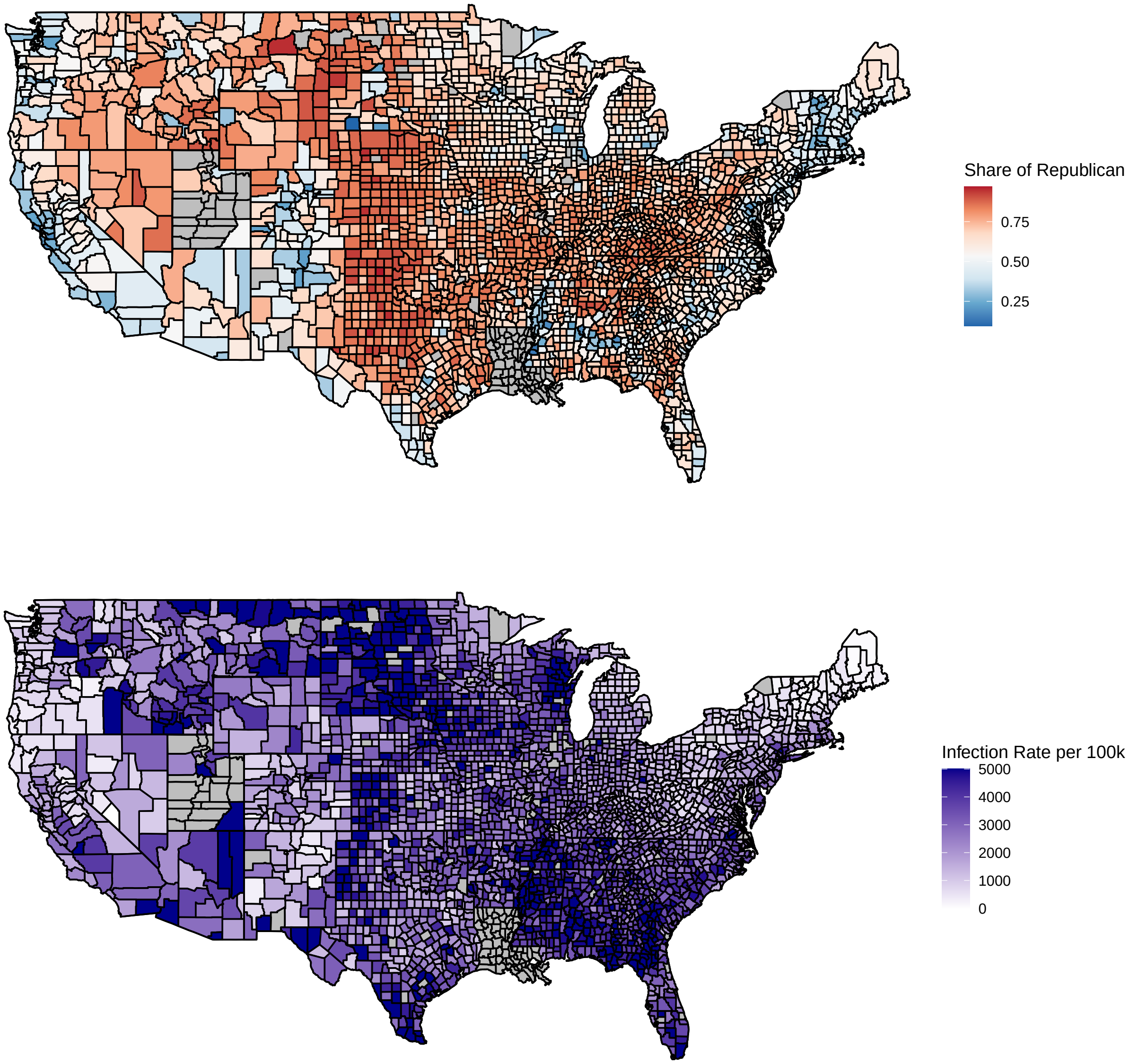

Figure 2: The ratio of votes for the Republican and the infection rate per 100k in each county

observational study, I will implement the methods of propensity scores to find the change of votes for Trump due to the pandemic. Based on the results, I will also conduct a counterfactual analysis to see what would be happened if there was no COVID,

First and foremost, we need to define the treatment. The treatment in an experiment is something that will be artificially imposed to a certain groups of people to see the effect of interested variable with and without this treatment. Therefore the people with the treatment is called treatment group and the rest are called the control group. The settings of treatment and control group is to test whether the imposed factor really have impacts on the interested variable. However, since this is an observational study, it is unrealistic to impose a factor manually. In order to find the the casual effect of COVID on Trump's voting, I will choose the treatment to be whether a county has a high infection rate.

3.1 Propensity Score

Propensity score is a technique that can help us to decide which county would be in the treatment group. That said, the propensity score in this study is the probability of a county having a high infection rate given a set of socioeconomic variables. By using the propensity score, it can ensure that the two counties with similar propensity score having the same distribution of the observed socio-economic variables such as income level even though they are in the different group. This is fantastic as we can have may pairs from treatment of control group which have similar covariates by fixing propensity scores. The key drawback that we can not estimate the casual effect from observational studies is addressed.

In addition to that, by finding the propensity scores, we can also reconstruct our assignment mechanism to a strongly ignorable one. An assignment mechanism is "strongly ignorable" if and only if, given the observed covariates, the potential outcomes are independent of the treatment assignment. This indicates that the change of voting for Trump is independent to our assignment that a county has high or low infection rate.

As the propensity score represents the probability of a county having high infection rate, therefore we need to find a logistic regression to predict this probability. On my previous paper, I already find the best model predicting the mortality rate of COVID in each US county. Even though that one was to predict the mortality rate, however, the mortality is correlated with infection rate. Therefore I will take the dependent variables in that model to be the covariates in predicting the propensity score for each county, which is:

$$
\begin{aligned}
\frac{PS}{1-PS} = \beta_0 + & \\
& + \beta_1 \times \text{prop\_higher\_education} \\
& + \beta_2 \times \text{IncomePctile} \\
& + \beta_3 \times \text{no\_insurance} \\
& + \beta_4 \times \text{private\_insurance} \\
& + \beta_5 \times \text{males} \\
& + \beta_6 \times \text{old\_85} \\
& + \beta_7 \times \text{white\_pct} \\
& + \beta_8 \times \text{black\_pct}
\end{aligned}
$$

Using the above model, we can therefore find the PS for each county. However, there are three different ways to use the propensity scores to find the treatment effect.

3.1.1 Propensity Score Matching

The package needed to implement the matching is the `matching`. Using this package, we can match the observations in treatment and control group having the same or similar propensity scores. Then take only take the matched pairs to be the new data set and find the find the treatment effect between them. In addition, we can also check the balance between the treatment and control groups.

However, the most significant drawback is that the new data set can only contain the matched pairs. We may lose a huge amount of observations during this process. Our predictions may also not be accurate.

3.1.2 Propensity Score Stratification

As introduced above that using PSM may cause the waste of data and results in bias. However, propensity score stratification will use the entire data set to avoid the waste. The way to do this is to split the data into different quantiles and fit a model for each subclass with the only dependent variable is the treatment variable. Then calculate the average casual effect for the subclasses.

Even though the PSS can use the entire data set, however, the choice of quantiles will be closely related to the predictions. If we have a large sample and want to split the data into many subclasses, then this method may be computational slow.

3.1.3 Propensity Score Regression Adjustment

The last method is the most straightforward, meaning that we just fit a regression model with treatment variable and propensity score be the two dependent variables.

The above three methods have their unique advantages but the results may vary between different data sets. I will perform both the three methods and to compare which one would be the best.

3.2 Counterfactural Analysis

By using the methodology of propensity score, we are able to access the casual effect of high infection rate on the votings for Trump. However, it can not tell us the election results and we can not make conclusions about the COVID and election. Therefore, I will conduct a counterfactural analysis, meaning that I will re-calculate the votes for Trump and Biden, to see if the final election results would be changed.

The way I will use to re-calculate the votes for the two candidates is to follow the official rule of election. By finding the difference of voting percentage, I will re-allocate the votings for the two parties and calculate the electoral votes in each state using the rule of winner-takes-all.

4 Results

Our results are summarized in **?@tbl-modelresults**.

```
Call:
glm(formula = high_infrate ~ prop_higher_education + pctile +
    no_insurance + private_insurance + males + age_85 + white_pct +
    black_pct, data = republican_data)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.0801275  0.2154169   0.372 0.709945
prop_higher_education -0.0066181  0.0012949  -5.111 3.40e-07 ***
pctile                -0.0018417  0.0004839  -3.806 0.000144 ***
no_insurance           0.0193575  0.0021510   9.000  < 2e-16 ***
private_insurance      0.0107016  0.0015025   7.122 1.31e-12 ***
males                  0.0036183  0.0035165   1.029 0.303585
age_85                 0.0078303  0.0078874   0.993 0.320908
white_pct             -0.0062757  0.0010244  -6.126 1.02e-09 ***
```

```
black_pct              0.0046024  0.0010745   4.283 1.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2111345)

    Null deviance: 764.42  on 3088  degrees of freedom
Residual deviance: 650.29  on 3080  degrees of freedom
AIC: 3973

Number of Fisher Scoring iterations: 2
```

```
library(Matching)
```

```
Loading required package: MASS


Attaching package: 'MASS'


The following object is masked from 'package:dplyr':

    select
```

```
##
##  Matching (Version 4.10-14, Build Date: 2023-09-13)
##  See https://www.jsekhon.com for additional documentation.
##  Please cite software as:
##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
##   Software with Automated Balance Optimization: The Matching package for R.''
##   Journal of Statistical Software, 42(7): 1-52.
##
```

```
# Propensity score matching
# mb <- MatchBalance(high_infrate ~ prop_higher_education + pctile +
#                              no_insurance +private_insurance + males + age_85 +
#                              white_pct + black_pct, data = data)
rr <- Match(Y = republican_data$pct_vote, Tr = republican_data$high_infrate,
            X = republican_data$prop_score, M = 1)
# mb <- MatchBalance(high_infrate ~ prop_higher_education + pctile +
```

```r
#                               no_insurance +private_insurance + males + age_85 +
#                               white_pct + black_pct, match.out = rr,
#                    data = republican_data, nboots = 10)
summary(rr)
```

```
Estimate...  -0.011174
AI SE......  0.0079011
T-stat.....  -1.4143
p.val......  0.15728

Original number of observations..............  3089
Original number of treated obs...............  1389
Matched number of observations...............  1389
Matched number of observations  (unweighted).  5522
```

```r
# Propensity score regression adjustment
model_adj <- glm(pct_vote ~ high_infrate + prop_score, data = republican_data)
summary(model_adj)
```

```
Call:
glm(formula = pct_vote ~ high_infrate + prop_score, data = republican_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.702141   0.007277  96.486  < 2e-16 ***
high_infrate  0.014911   0.006234   2.392   0.0168 *
prop_score   -0.132681   0.016134  -8.224 2.88e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.02527324)

    Null deviance: 79.721  on 3088  degrees of freedom
Residual deviance: 77.993  on 3086  degrees of freedom
AIC: -2590.2

Number of Fisher Scoring iterations: 2
```

```r
# Counterfactual Analysis
adjusted_data <- data %>%
  dplyr::select(state, county, fips, party, votes, total_votes, pct_vote, high_infrate)

adjust_function <- function(data){
  adjust_votes <- c()
  for (i in 1:nrow(data)){
    if (data[i,"high_infrate"] == 1){
      if (data[i,"party"] == "Republican"){
        adjust_votes[i] <- round((data[i,"pct_vote"] + 0.036928 ) * data[i,"total_votes"],
      }
      else if (data[i,"party"] == "Democrat"){
        adjust_votes[i] <- round((data[i,"pct_vote"] - 0.036928 ) * data[i,"total_votes"],
      }
      else {adjust_votes[i] <- data[i,"votes"]}
    }
    else {
      adjust_votes[i] <- data[i,"votes"]
    }
  }
  return(adjust_votes)
}



adjusted_data %>%
  group_by(state, party) %>%
  summarise(votes = sum(votes), .groups = "drop")
```

```
# A tibble: 220 x 3
   state party         votes
   <chr> <chr>         <int>
 1 AK    Democrat      11821
 2 AK    Green           157
 3 AK    Libertarian     671
 4 AK    Other           128
 5 AK    Republican    10551
 6 AL    Democrat     849624
 7 AL    Other         32488
 8 AL    Republican  1441170
 9 AR    Democrat     423932
10 AR    Green          2980
```

```
# i 210 more rows
```

```
adjust_function2 <- function(data) {
  adjust_votes <- numeric(nrow(data))  # Initialize the vector with the correct length
  extra_vote_percentage <- 0.036928

  for (i in 1:nrow(data)) {
    if (data[i, "high_infrate"] == 1) {
      total_votes <- data[i, "total_votes"]
      extra_votes <- round(extra_vote_percentage * total_votes, 0)

      if (data[i, "party"] == "Republican") {
        adjust_votes[i] <- data[i, "votes"] + extra_votes
      } else {
        # Calculate total votes for non-Republican parties
        total_non_rep_votes <- sum(data[data$party != "Republican" & data$fips == data[i,

        # Distribute the vote loss proportionally
        if (total_non_rep_votes > 0) {
          party_vote_share <- data[i, "votes"] / total_non_rep_votes
          votes_lost <- round(extra_votes * party_vote_share, 0)
          adjust_votes[i] <- max(0, data[i, "votes"] - votes_lost)
        } else {
          adjust_votes[i] <- data[i, "votes"]
        }
      }
    } else {
      adjust_votes[i] <- data[i, "votes"]
    }
  }

  return(adjust_votes)
}

adjust_vote1 <- adjust_function(adjusted_data)
adjust_vote2 <- adjust_function2(adjusted_data)
adjusted_data$adjust_vote1 <- adjust_vote1
adjusted_data$adjust_vote2 <- adjust_vote2
```

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

6 References