REGULAR ARTICLE

Political Polarization and Covid-19 Presidential Approval Ratings

Samuel P Dummer*

Correspondence:
dummer22s@ncssm.edu
North Carolina School of Science and Mathematics, 1219 Broad St., 27705 Durham, NC
Full list of author information is available at the end of the article
*NCSSM Online Program

Abstract

Political polarization is a problem in the United States, especially in recent years. It causes many issues, one of these being increased disputes between opposing parties, the Republican and Democratic parties. Another main issue with polarization is the biases it causes within the party. These biases were measured using Covid-19 approval ratings of Presidents. We started by creating a few graphs to visualize the data quickly, then ran Bayesian Information Criterion calculations to find variables that had a causal relationship with the approval rating. Then, we ran a k-NN algorithm to test the efficacy of the variables and lastly created a few more graphics based on our findings. In all terms of things, it was found that political parties were the leading cause of variation in the approval ratings, with Democrats having a mean approval rating of 12% under Trump and 86% under Biden and Republicans with an approval rating of 81% under the Trump administration and 24% under Biden. Therefore, it supports the idea that political polarization played a role in the Covid-19 approval rating data.

Keywords: coronavirus; bias; polarization

1 Introduction

Since March of 2020, the *coronavirus* has led to over 777 thousand deaths in the United States. This deadly virus has caused the country to go into quarantine and has placed immense pressure on the President. Any decision he sways the public's opinion heavily. Through polls, the President's approval ratings can be observed, reflecting how his actions have shifted it. The problem with this is that there are political biases affecting the levels of approval rates. This is due to the political polarization caused by the presence of a two-party system within the United States, something the founding fathers had feared ([1]).

1.1 Political Polarization

Political polarization is the idea that over time members of opposite parties idea's drift further and further away from each other, increasing conflict and dislike between the political parties. In a study done by the Pew Research center, they found that since 1994 there has been a major increase of unfavorable views of the other party ([2]). The data suggests that there was about a 20% increase in the amount people in each party that viewed the other one unfavorably. 538 found a similar result in a poll they conducted. They found that on a scale from 0 to 100 (0 being very unfavorable and 100 being favorable) the sentiments towards the other party in the United States decreased by about 20 points since about 1980 ([3]). This gap was much larger than any other country that they had polled. This decrease has

been incredibly drastic since the closest country to the United States was the United Kingdom with a 10 point drop.

A similar result was found in another study except they found that average American saw their party as much better than the opposing party. On a scale from 0-100 they were rating their party about 45 points higher, almost a 20 point increase from 1978 ([4]). Additionally, another study done found that between 1972 and 2004 there hadn't been any major shift in polarization between parties([5]). This may seem contradictory to the earlier statement, but since 2004 Pew Research center has found a significantly larger decrease in people who identify as centrists dropping from 49% to 39% ([2]). Polarization has caused immense bias within political polling. When a voter is presented with a poll of a President of the opposite party, there is a much higher chance they disapprove of their actions. This is a major problem since this skews data and does not represent how well the President is doing in actuality. While doing research, there wasn't a lot of information about how political polarization has directly affected polling, but since disapproval of the opposing party is directly related to political polls there is a lot of overlap.

1.2 Description of Data

In our specific research, we looked at *coronavirus* presidential approval rating for the past few years gathered by FiveThirtyEight. With this data we can see many variables such as the different parties, pollsters, dates, sample sizes, and the question asked. Using this data, we analyzed what caused the most variation in the approval rating. There were a few factors that caused this variation, the larger ones being political party and population polled, but the pollster had some influence in variation, but no causation with approval rating. When looking at how much party caused variation between approval ratings.

1.3 Summary of Results

In Figure 1 one can even observe up to 80% approval ratings for the president of their own party. This is an example of political polarization. The voters in the polls gave the President they voted for a higher rating for the way they handled the *coronavirus* even though the actions taken have change that drastically between Presidents. This shows a bias within politics towards members of their own party. This bias was shown in a journal article written by Delia Baldassarri and Andrew Gelman. They found that due to political polarization individuals are much more likely to assign themselves and label themselves to a party and not differ from it [6]. This labeling causes greater divergence between parties and removes any grey area between them. This means that people will become less and less likely to end up going against the ideas of the figurehead in charge of their party. This helps to explain the immense bias that is present throughout the data within this data set. The Presidents end up having immense support from their party and immense opposition from the other party. This will generally stay true no matter if the decisions the Presidents make stray somewhat from the parties platform.

2 Data Preparation and Modeling

The data for this work was obtained from `https://github.com/fivethirtyeight/covid-19-polls/archive/master.zip`.

2.1 Data Cleaning

To start off the data preparation, we read in the data and had to clean it up some. This was because there were many missing values that we did not want. To get rid of these values, we identified the column that had the most missing values and deemed it as unimportant. Next, we removed variables that didn't have any impact on our research such as the URL of the poll.

2.2 Exploratory Data Analysis

Following this, we created many different filtered sets of the original data set to help easily create a few graphics of the *coronavirus* approval ratings of each president. This helped us to easily separate the variables within the graphics to help make them clearer. Once the newly created data sets were ready we create some graphics (see Figure 1). These graphics helped to show many trends which helped identify the most influential variables on the data. To help continue the Exploratory Data Analysis(EDA) we performed Bayesian Information Criterion (BIC) analyses. These analyses are helpful in finding any causal relationship. To find this there is a specific equation that is used:

$$BIC = k * \ln(n) - 2\ln(L) \tag{1}$$

This equation helps us find the variables that have a causal relationship with approval ratings. This means that changes in the variable cause there to be specific changes in approval rating. In this data set specifically we used it on all the variables we believed had some sort of relationship.

2.3 Analysis of Data

The next step we took was to see how well the variables worked in predicting the subject(a.k.a. the President being polled). To do so we used the k-NN algorithm also known as the k-nearest neighbors algorithm. This algorithm pairs new values based on the neighbors. It takes the average or in some cases most common values from the number of neighbors. This number of neighbors is specified through the variable k. There are many ways to calculate or find the predicted value. For continuous values, Euclidean distance is used and the mean is found, but for discrete values, the overlap method is generally used. This method find the k number of closest variables and then assigns the mode of the k number of variables to the value being predicted. There are some ways to select a good value of k, but we decided to just try again and again to find a value that seemed to work best for the set.

To test the algorithm the data set was split into two sets: a training set and a testing set. Generally, the sets are split by 60-70% of the data for the training set and 30-40% of the data for the testing set. The training set is used to train the algorithm and then the algorithm checks itself with the testing set. To check itself, it will take the values from the testing set and use the algorithm to predict what values in a variable will be. In our specific data set, we will be predicting the type of glass. We chose to split the data set up by 70% and 30%.

Additionally, this is a seed-based algorithm so we must set a seed to be able to run the algorithm. We create the training set and testing set then run the algorithm. Once our results were obtained, we create a cross table to all our values we obtained versus the actual values (see Figure 2).

2.4 Final Graphics

Lastly we created a few more models based on the new information obtained. We created a density plot of the approval ratings based on party and pollster. The reason we chose pollster was because we knew there wasn't a causal relationship, but it was somewhat close so we wondered whether or not there was any correlation at all. Lastly we created a set of tables of the mean approval rating based on population and party since population was another value that seemed to be very close.

3 Results

3.1 Preliminary Analysis

Right off the bat, we theorized that within the results of the *coronavirus* polls, there were some bias within the questions since not all the pollsters asked the same question. So we graphed the values based on these results, but found nothing of significance. There was still some sort of trend going on within the data so we decided to explore further. If the bias in the data wasn't caused by the question what could it be caused by? After some testing, we found that the trend we observed was a trend between the political parties. This specific trend is visible in Figure 1. During the Trump part of the administration during the *coronavirus* quarantine many of the approval rating results from the Republicans were coming through as 70% to 80% approval while the approval rating from the Democrats seems to lean towards the 10% to 20%. On to of this, the third party that was polled, the Independent party, had results that were usually in the middle, so around 40% to 50%. These results were basically completely swapped when Biden came into office. The Democratic approval rating immediately jumped to 80% and the Republican approval rating dropped to 10%. The only results that stayed relatively the same were those of the Independent party. There was another interesting trend. The shape of the trend over time of the approval rating of the opposite party of the President and the Independent party were usually more similar than that of the same party as the President. This trend is interesting but there doesn't seem to be an explanation. It was very surprising that the questions asked by the pollsters had no correlation or trend with approval rating even though the question is so important within polling. Lastly, there does seem to be a trend over time which does help to show how much the population as a who is reacting to it, but there still may be problems when it comes to the percentages given. Based on these results. the only real possible use of the polls would be to see how each party seems him over time rather than at a certain point. This is because the trend is observable and not as biased by the political polarization. This means that the overall approval rating doesn't actually show what the population as a whole thinks of him since the trend are all separate and additionally since all the parties are being taken accounted for all of the skewed results effect the results of everyone combined.

3.2 Bayesian Information Criterion

We may have found some trends, but we have yet to find out whether they are correlated with each other or there is causation between the variables. To help find the distinction between these we can use the Bayesian Information Criterion analysis which helps us calculate whether there is a possibility or not of causation.

We start off by finding the Bayesian Information Criterion between the approval rating and 1. Doing so give us a value of 27886.63. This will be our base value that we compare the rest to. If the BIC values that we calculate whether or not there is a causation by subtracting the new BIC value for the base and if the number is 10 or greater than there is likely causation. We start off by calculating the BIC of the approval ratings and the party of the people being polled. We got back 27126.8. The difference between this and the base is much greater that 10 meaning that there is very highly likely causation between the two variables. Next, we run many different values such as text, pollster, subject, sample size, and population. After running the BIC calculation we get the values of 29072.66, 28311.59, 27757.09, 27893.34, and 27897.83. These results mean that the only other value that has any sort of causation with the approval ratings is the subject. This makes sense since the subject is the other variable that has any correlation with the political party. This means that depending on your party as President, it will sway the way citizens see you simply by what party you are a part of. The fact that the only variable that led to the causation of different approval rating being party dependent further helps to strengthen the claim that political party and political polarization is what is helping to cause the bias in the polling. Something unexpected that was found while looking at these results was how close sample size and population were to being considered likely causal. This was surprising since we had no real focus on these two variable at all in our research and they ended up being more important than once thought. On top of this, the one variable that we thought would be the biggest cause of change in approval rating, the text/question asked, was actually the least causal of them all. Even the pollster had more impact that the question. That means in a general scheme of things, the pollsters did a very good job in developing questions that would not cause any bias.

3.3 k-Nearest Number Algorithm

Following this, we wanted to test how well these variables were going to perform in a k-NN prediction algorithm. As explained earlier a k-NN algorithm also know as a k-nearest number algorithm is an algorithm that pairs new values based on the neighbors. This means that it takes the average from the number of neighbors. Once this has happened, we create the training and test set. Once this is done we run the algorithm. After running it we can compare the test set with the predicted values. The values do seem to be fairly similar, but it is hard to tell how many were guessed wrong.

This is where we use a cross table. This helps to visualize the results obtained. In our circumstance we are given a 4 by 4 cross table, but the table can be larger based on the range of the variable being predicted. Since there are only two Presidents that have been in charge during the span of the *coronavirus* quarantine, Trump and Biden, there are only two different values that the algorithm tests for. The cross table (see Figure 2) of this specific simulation gave us a success rate of 93.9%. There were 585 out of 623 predictions that were correct. In general, this is very high since in a country without biases caused by political polarization, it should be very hard to predict the President in office at the time based solely on the variables given. Overall, the results found by running this algorithm further give evidence towards there being biases in the polling.

3.4 Final Figures

As a final part of the results we graphed a few density plots to help visualize the distribution of the votes between parties (see Figure 1). In this figure, we can see that it is distributed very similarly to how we described it earlier. The party that is currently the same as the President is that which has the higher approval. Interestingly, the Independent party seems to have a consistently lower approval rating. It is higher than the opposing party, but it is still lower than 50%. This may be because they have no attachment or opposition to the President The key difference we can see in this graph is that the distribution is much more visible. During the Biden presidency there seem to be a larger distribution among the Republican and Independent Voters. This means that they was more variance in their voting and that there were a higher amount of those voters that were giving Biden a slightly higher rating. This may also mean that the approval of Biden's handling of *coronavirus* over his presidency may have shifted very much.

The next variable that I wanted to take a look at was pollster since there was a chance that there could have been some bias within the pollsters themselves and there was a somewhat low BIC meaning that there is an incredibly low chance of causality, but it is still closer than the text. When looking at the graphs, there does not seem to be anything out of he ordinary on it. There are some pollsters that seem to be shifted in some way, but overall it seems as if the distribution is fairly random. Withing each pollster it is also evident that each party is equally represented since there are small peaks around the plots that were grouped by party.

The only other variable that seemed to have substantial effect on the data is the population, the type of people being polled (a.k.a. Likely Voter, Adult, Registered Voter). To help model this, we grouped the data by party and population and formed a new data set. This allowed us to create a grouped summary of the approval and disapproval rating for each President (see Table 1 and 2). Again, it is visible in this graph that the parties are much more likely to approve of a President that is the same party as them and disapprove of a President that is a member of the opposite party. The Independent party does sway a little more in Biden's direction, but is overall more consistent that the Democratic and Republican parties. In the switch from Trump to Biden's administration the approval of Democrats grew about 75% and the Republican approval rating fell by about 56%, much lower than the Democrats, but still a significant change.

The next trend that can be observed in these tables is that Likely Voters have, in a general sense, higher disapproval and lower approval rates. The disapproval in generally 4% higher while the approval rating is 5% lower. There is no explanation for this and since it is consistent between the two presidents, than there is no proof of bias within the population.

4 Discussion

4.1 Preliminary Graphs

Throughout all the results, the majority support the claim that political polarization has caused bias in polling. To start off, the scatter plots made (see Figure 1) help to provide us with the separation between the parties approval ratings. In these plots it was found that in the switch from the Trump to the Biden administration,

the Democratic approval rating immediately jumped up to around 80% and the Republican approval rating dropped to about 10%. It would be understandable if there were some outliers with very high approval ratings, but in each graph it is clearly visible that the trend is consistent. This indicates that this is no random occurrence. The data supports the idea that over time the support for ones own party has increase drastically since 1978 ([4]). On top of this, the overall shape of the trend is very much similar between parties, this means that the ideas the citizens had on the President did change similarly overtime. The separation between parties simply means that the basis of someone's views on the President is based on the view of the whole party which continues to dislike the other party more and more over time.

4.2 BIC

Proceeding the preliminary graphs came the Bayesian Information Criterion where we were able to calculate whether or not there was a causal relationship between many of the variables and the approval ratings. In our results, we found that since the BIC of the variable party was 27126.8 and the difference between this and the base is much greater that 10 meaning, there is very highly likely causation between the two variables. We continued to run many different values such as text, pollster, subject, sample size, and population to find whether or not they also had a causal relationship. After running the BIC calculation we get the values of 29072.66, 28311.59, 27757.09, 27893.34, and 27897.83 meaning that the only other variable which had a highly likely chance of being a causal relationship was the subject.

Since the subject (the President that was in office at the time of the poll) also has a relation to party, it means that the only real causal factor causing changes in the approval rating in this data set was political parties, further showing that political polarization is the biggest and only factor affecting the approval ratings. The other variables have very close scores but nowhere as low as party and subject. These variables, such as population, might be closer because there are certain members of the population that will vote much differently than others. There is no major explanation for the other since variables such as sample size and pollster have no relation to political parties. As the parties within the United States continue to diverge as predicted by a study done by FiveThirtyEight the BIC for both party and subject will most likely continue to grow stronger and stronger ([3]).

4.3 k-NN Algorithm and Figures

k-Nearest Neighbor Algorithms were used in this study to test the efficiency of the variables to be able to predict the President in office at the time of the poll. To do so, we separated the data set into 70% of the data and 30% of the data and then used these as training and testing sets to run the algorithm. Once we ran it, it was fairly obvious that the results were fairly accurate, but to help further visualize it, we created a cross table. In this cross table we found that the algorithm had predicted it correct 93.9% of the time.

This is an incredibly high success rate for this data. This high success rate suggests that the variables are indeed very influential in predicting each other. With this information, it would probably by possible to predict the approval ratings of future

polls. The overall approval rating would be hard to calculate, but the approval rating for each party is possible. For the Democrat's, they will most likely have an approval rating in the low 80% and the Republican's approval rating will be around the low 10%.

In continuation to this, the tables made later also provide very similar results. It was found in the table that the average of about and 83% approval rating for Biden from the Democrats and an 11% approval rating for Biden from the Republicans. This is very similar to the projected averages. These tables also show that there is some trend within the population the voters were from, but the trend had no correlation between Presidents.

Lastly, the density graphs created are a great model to show how extreme and distributed the data is. In this case, the data is incredibly centralized based on party concluding that there was not much variation within the party and therefore an incredible amount of bias. The graphs had very high peaks where each party was grouped which means that the variation within the party is incredibly low. These results are similar to those found by researchers, Levi Boxell, Matthew Gentzkow, and Jesse Shapiro. They found that there has been an increase in the rating of their own party versus the other showing more of an affinity to your own party.

There is a possibility that the gap between approval ratings continue to rise as time goes on. This is because the longer our country exists as it is the more political polarization will occur as shown in the study done by the Pew Research Center ([2]). Without any control over this matter, it may lead to problems in the future of the country.

5 Conclusion

Political polarization has caused a great influx of bias in the opinions of voters. In polling, the results become biased because those being polled are more likely to be greatly in favor of their party and great against the opposing party. As the negative views of the party rise the lower a chance there is of working together within the government.

The stronger the polarization becomes, the larger the rift between the parties becomes and the more violence and disputes. There are many problems that may follow this such as Congress being at a standstill since no one can compromise. This has already been happening and has prevented large amounts of legislation from being able to pass. To help prevent any of these problems from exponentiating there would have to be reforms to the voting process, specifically reform of the first-past-the-post(FPTP) voting system which is the main of a country having a two-party system. Some voting systems that may help increase the number of parties in elections could be a Rank-Choice system. This allows the voter to list the representative they want in order of favorite to least favorite. This eliminates the idea of "voting against the opposite party." This is because if there is no outright majority, the party with the least votes will be elimintated and the votes will go to the second choice. There are many other ways that the voting system can be reformed, but without any change to the system political polarization will continue to grow.

Availability of data and materials

The data for this work was obtained from
https://github.com/fivethirtyeight/covid-19-polls/archive/master.zip.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

This paper is solely the work of the author. All references are included in the bibliography and are cited appropriately.

Funding

Funding for this program is provided by the North Carolina School of Science and Mathematics, the University of North Carolina General Administration, and the General Assembly for the State of North Carolina.

Acknowledgements

I would like to thank Mr. Gotwals for helping me through the semester and teaching me all about Data Science. He taught me everything I know. This semester with him has been lovely and has helped me discover a greater passion for data science. I would also like to thank my dog for keeping me company for the long hours that I wrote this article. Lastly, I would also like to thank my parents for supporting me and making me want to always do the best I can at any given moment in life.

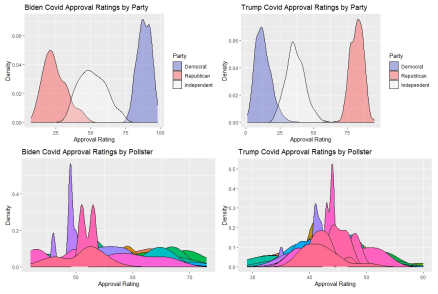
References

1. Drutman, L.: America Is Now the Divided Republic the Framers Feared. https://www.theatlantic.com/ideas/archive/2020/01/two-party-system-broke-constitution/604213/
2. Political Polarization in the American Public. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/
3. Leedrutman: Why The Two-Party System Is Effing Up U.S. Democracy. https://fivethirtyeight.com/features/why-the-two-party-system-is-wrecking-american-democracy/
4. Boxell, L., Gentzkow, M., Shapiro, J.M.: Cross-Country Trends in Affective Polarization (2020). https://www.nber.org/papers/w26669
5. Fiorina, M.P., Abrams, S.J.: Political polarization in the american public. Annu. Rev. Polit. Sci. **11**, 563–588 (2008)
6. Baldassarri, D., Gelman, A.: Partisans without constraint: Political polarization and trends in american public opinion. American Journal of Sociology **114**(2), 408–446 (2008)

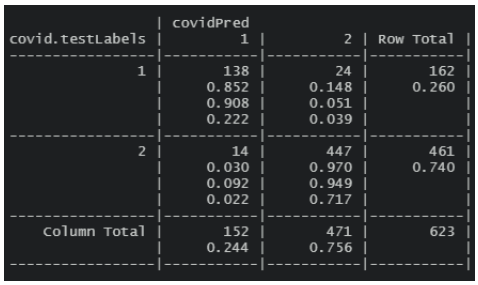
Figures

Figure 1 These are two figures that help to show the distribution of the polling data. Each help to visualize different ideas

Covid-19 Approval Rating Scatter Plots



Covid-19 Approval Rating Density Plots



	covidPred		
covid.testLabels	1	2	Row Total
1	138	24	162
	0.852	0.148	0.260
	0.908	0.051	
	0.222	0.039	
2	14	447	461
	0.030	0.970	0.740
	0.092	0.949	
	0.022	0.717	
Column Total	152	471	623
	0.244	0.756	

Figure 2 Cross table of Prediction Analysis

Tables

Table 1 Summary of Data Grouped by Population and Party While Trump was in Office. This table shows the mean approval rating and disapproval rating for each party and population while Trump was President.

Population	Party	Mean Approve	Mean Disapprove
Adult	Democrat	12.8816119402985	83.614447761194
Adult	Republican	82.054	15.0557910447761
Adult	Independent	36.4240384615385	53.6192307692308
Registered Voter	Democrat	12.5357142857143	84.7708333333333
Registered Voter	Republican	81.9952380952381	15.672619047619
Registered Voter	Independent	36.8259259259259	55.9154320987654
Likely Voter	Democrat	11.5408333333333	87.3758333333333
Likely Voter	Republican	85.08	13.7533333333333
Likely Voter	Independent	39.3166666666667	58.0166666666667

Table 2 Summary of Data Grouped by Population and Party While Biden is/was in Office. This table shows the mean approval rating and disapproval rating for each party and population while Biden is/was President.

Population	Party	Mean Approve	Mean Disapprove
Adult	Democrat	87.8869444444444	8.00685185185185
Adult	Republican	22.2073148148148	70.7534259259259
Adult	Independent	49.504854368932	39.9320388349515
Registered Voter	Democrat	89.3076923076923	7.8974358974359
Registered Voter	Republican	27.1538461538462	68.0512820512821
Registered Voter	Independent	53.8076923076923	37.1923076923077
Likely Voter	Democrat	83	14.8
Likely Voter	Republican	26.4	69.2
Likely Voter	Independent	51.4	43.6

Additional Files

R Code for this work

```
# Sam Dummer
# November 16, 2021
# covidPresidentDummer.R
#
# In this R Script, we take a look at the approval ratings of the different presidents response to covid since the beginning
#
# We start by cleaning the environment, setting the directory and loading in the libraries
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
library(tidyverse)
library(cluster)
library(factoextra)
library(class)
library(ggvis)
library(gmodels)
library(gridExtra)
library(stringr)
#
```

```
# reading in the data
covid <- read.csv("covid_approval_polls.csv", header=T, na.strings=c(""))
#
# checking to see if there are any na values
clean <- ifelse(complete.cases(covid)==TRUE,1,0)
table(clean)
paste("There are ",dim(covid)[1]-sum(clean), " rows with missing data.")
#
# check head, tail, summary, and structure of data set
head(covid)
tail(covid)
summary(covid)
str(covid)
glimpse(covid)
#
# Remove the columns with the url, sponsor(since there are so many NA values), and the question asked
covid <- covid[c(1:3, 5:8, 10:12)]
#
# now we clean up the population and party column and make it easier to understand
covid$population <- factor(covid$population, levels = c("a", "rv", "lv"), labels = c("Adult", "Registered Voter", "Likely Voter"))
covid$party <- factor(covid$party, levels = c("all", "D", "R", "I"), labels = c("All", "Democrat", "Republican", "Independent"))
#
# we fix the format of the dates
covid$start_date <- as.Date(covid$start_date , format = "%m/%d/%Y")
covid$end_date <- as.Date(covid$end_date , format = "%m/%d/%Y")
#
# lastly we omit the final na values
covid <- na.omit(covid)
summary(covid)
#
# here we create different data sets that can be graphed to see specific trends
covidBiden <- filter(covid, subject == "Biden")
covidTrump <- filter(covid, subject == "Trump")
covidBidenAll <- filter(covid, subject == "Biden", party == "All")
covidAll <- filter(covid, party == "All")
covidTrumpAll <- filter(covid, subject == "Trump", party == "All")
covidNoAll <- filter(covid, party == "Democrat" | party == "Republican" | party == "Independent")
covidBNoAll <- filter(covid, (party == "Democrat" | party == "Republican" | party == "Independent") & subject == "Biden")
covidTNoAll <- filter(covid, (party == "Democrat" | party == "Republican" | party == "Independent") & subject == "Trump")
#
# here are four plots of Trump and Bidens approval rating colored by party
biden_1 <- ggplot(covidBNoAll, aes(start_date, approve, color = party)) + geom_point() + geom_smooth(method = "loess", se = F) +
labs(title = "Biden Covid Approval Ratings By Party", y = "Approval Rating", x = "Date") +
scale_color_manual(values=c("#00BFC4", "#F8766D", "#C77CFF"), name = "Party")
trump_1 <- ggplot(covidTNoAll, aes(start_date, approve, color = party)) + geom_point() + geom_smooth(method = "loess", se = F) +
labs(title = "Trump Covid Approval Ratings By Party", y = "Approval Rating", x = "Date") +
scale_color_manual(values=c("#00BFC4", "#F8766D", "#C77CFF"), name = "Party")
biden_2 <- ggplot(covidBidenAll, aes(start_date, approve)) + geom_point(color = "#7CAE00") +
geom_smooth(method = "loess", se = F, color = "#FF0000") +
labs(title = "Biden Covid Approval Ratings", y = "Approval Rating", x = "Date")
trump_2 <- ggplot(covidTrumpAll, aes(start_date, approve)) + geom_point(color = "#7CAE00") +
geom_smooth(method = "loess", se = F, color = "#FF0000") +
labs(title = "Trump Covid Approval Ratings", y = "Approval Rating", x = "Date")
grid.arrange(biden_1, biden_2, trump_1, trump_2, ncol=2, nrow=2)
#
# here we see if there is any causality between the approval rating and different variables
BIC(lm(covid$approve~1))
BIC(lm(covid$approve~covid$party)) # Yes
BIC(lm(covid$approve~covid$text)) # No
BIC(lm(covid$approve~covid$pollster)) # No
BIC(lm(covid$approve~covid$subject)) # Yes
BIC(lm(covid$approve~covid$sample_size)) # No
BIC(lm(covid$approve~covid$population)) # No
#
# we create a new data set of only numeric values to perform prediction analysis
covidNumeric <- covidNoAll[c(3, 5:9)] %>% mutate(pollster = as.numeric(factor(pollster)),
population = as.numeric(factor(population)), party = as.numeric(factor(party)),
subject = as.numeric(factor(subject)), text = as.numeric(factor(text)))
#
# creating training and testing sets
set.seed(5816497)
split <- sample(2, nrow(covidNumeric), replace=TRUE, prob=c(0.7, 0.3))
```

```
split
covid.training <- covidNumeric[split==1, 1:6]
head(covid.training)
covid.test <- covidNumeric[split==2, 1:6]
head(covid.test)
covid.trainingLabels <- covidNumeric[split==1, 4]
covid.testLabels <- covidNumeric[split==2, 4]
#
# using prediction analysis
covidPred <- knn(train = covid.training, test = covid.test, cl = covid.trainingLabels, k = 10)
covidPred
covid.testLabels
tab <- CrossTable(x = covid.testLabels, y = covidPred, prop.chisq = F)
#
# creating density plots based on party and pollster
BidenDens <- ggplot(covidBNoAll, aes(approve, fill = party)) + geom_density(alpha = .3) +
labs(title = "Biden Covid Approval Ratings", x = "Approval Rating", y = "Density") +
scale_fill_manual(values=c("#0015BC", "#FF0000", "#FFFFFF"), name = "Party")
TrumpDens <- ggplot(covidTNoAll, aes(approve, fill = party)) + geom_density(alpha = .3) +
labs(title = "Trump Covid Approval Ratings", x = "Approval Rating", y = "Density") +
scale_fill_manual(values=c("#0015BC", "#FF0000", "#FFFFFF"), name = "Party")
BidenDensPoll <- ggplot(covidBidenAll, aes(approve, fill = pollster)) + geom_density() +
labs(title = "Biden Covid Approval Ratings", x = "Approval Rating", y = "Density") + theme(legend.position = "none")
TrumpDensPoll <- ggplot(covidTrumpAll, aes(approve, fill = pollster)) + geom_density() +
labs(title = "Biden Covid Approval Ratings", x = "Approval Rating", y = "Density") + theme(legend.position = "none")
grid.arrange(BidenDens, TrumpDens, BidenDensPoll, TrumpDensPoll, ncol=2, nrow=2)
#
#
partyGroupTrumpPop <- covidTNoAll %>% group_by(population, party) %>% summarize(meanapp = mean(approve), meandis = mean(disapprove))
view(partyGroupTrumpPop)
partyGroupBidenPop <- covidBNoAll %>% group_by(population, party) %>% summarize(meanapp = mean(approve), meandis = mean(disapprove))
view(partyGroupBidenPop)
```