Module 8 – Portfolio Project

Christopher Rashidian

Colorado State University Global

MIS 540: Introduction to Business Intelligence

Dr. Kimberly A. Ford, D.M.

November 1, 2020

Module 8 – Portfolio Project

As the United States head into the November 3, 2020 General Election, one cause for much scrutiny across all platforms of media is the approval ratings of President Donald J. Trump. As those from the Right Wing of the political spectrum, including Trump himself, claim that Trump has the highest approval ratings of any sitting President. The data collection and the source of these ratings has the possibility, much like most of his claims, to contradict him by providing empirical evidence against his lay to claim.

Studies have shown that considering a national crisis, the President's response to such a crisis would have their approval ratings carried away well above seventy-five percent such as the case with George W. Bush after the September 11, 2001 terrorist attacks, or plummeting to less than thirty percent as was the case with Jimmy Carter during the 1979 Iranian Hostage crisis. The response of President Trump during the Coronavirus-19 crisis stemming from November 2019 to present. As such, the dataset being utilized for this assignment is the President Approval Polls from FiveThirtyEight (Our Data 2018).

The approval that President Trump claims to have is subjective to his false impressions of his delusions of grandeur, exasperated with violent mood changes, risky behavior, and lack of compassion during the COVID-19 pandemic, as detailed by his niece, Mary L. Trump, Ph.D. in her 2020 biography diving into the psychological motivations to these claims, and as such, the empirical measures of his approval ratings will vary between the news sources and the institutions that conducts these surveys based on their pool of participants (Trump, 2020).

As such, it could be assumed that those who subscribe right leaning news sources such as Fox News, Daily Mail, or World News Daily would reflect almost favorable results of the approval ratings whereas Bloomberg or Associated Press will reflect a neutral or balanced bias in the

population surveyed. As such, the problem is the selection of what survey sources should be utilized when comparing President Trump's approval ratings.

Data Discovery

The data discovery requires the analyst to detect patterns and outliers without creating a predictive or analytic model. This would allow us to understand the relationship of the data and would provide us deeper insights of the data at hand. When initially exploring the data, the column headers are as detailed below:

1. president: Which President is being surveyed on
2. subgroup: The group which was sampled
3. modeldate: Data retrieval date
4. startdate: Date in which the survey was started
5. enddate: Date in which the survey was ended
6. pollster: Surveying agency
7. grade: this is the reliability of the result
8. samplesize: this is the sample of the population surveyed
9. population: classification of registered voters (rv), adults (a), likely voters (lv), and voters (v)
10. weight: this is the estimation of the weight of the survey relative to the entire population
11. influence: the is the influences of the weight of the survey results, relative to the entire population
12. approve: The percentage of those who approve
13. disapprove: The percentage of those whom disapprove

14. adjusted_approve: The percentage of those who approve, adjusted for the weight and influence
15. adjusted_disapprove: The percentage of those whom disapprove, adjusted for the weight and influence
16. multiversions: this is missing many fields
17. tracking: this is missing many fields
18. url: where the survey is housed
19. poll_id: this is the identifying poll id
20. question_id: this is the unique identifier
21. createddate: this is the date in which the data was created
22. Year Created: this is the year in which the data was created
23. timestamp: this is the date in which the data was last updated

Data Preparation

The data preparation phase covers all activities to create the working dataset for further processing and analysis. The analysis of the raw data allows us to identify patterns and establish the relationship between variables. The data reviewed in this study encompasses the data that has been collected from January 23, 2017, three days after President Trump's inauguration, through October 27, 2020, seven days before the November 3, 2020 national general election. Contained in this dataset were classification of the poll, assigning a letter grading system from A, meaning the highest in trust in the survey conducted, to a D, meaning the lowest in trust rating from the conducted survey along with assigning the conductors within these categories, with highest ranking reputable sources such as CBS News, Fox News, and the Washington Post, to the lowest ranking reputable survey conductors such as SurveyMonkey. One important thing to make notation

of is that there are 536 surveys conducted without a letter grade. Table one below and on the following pages detail the summary statistics classified by the letter grade given and the survey conductor.

Table 1.

Summary Statistics of the Approval Ratings Classified by Letter Grade

grade	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Median
A	642	adjusted_approve disapprove	41.8684112 52.2392212	2.8136384 3.0927956	31.7094780 43.0000000	49.3281890 63.0000000	42.0926960 52.0000000
B	8605	adjusted_approve disapprove	41.4178513 53.4337230	2.7584380 3.0055615	32.9479300 35.0000000	52.6749390 63.0000000	41.4316000 53.0000000
C	5523	adjusted_approve disapprove	41.7962345 52.9254210	2.4715057 2.7298579	32.6148690 37.0000000	54.7977090 64.6000000	41.7977090 53.0000000
D	620	adjusted_approve disapprove	41.3002604 53.7035484	2.6125865 2.4694041	32.6603010 44.0000000	48.0667100 62.0000000	41.5568750 54.0000000
N/A	536	adjusted_approve disapprove	41.7363430 55.8690858	3.8883774 5.4691895	25.1115860 43.0000000	51.8456490 75.9000000	42.3383920 56.0000000

When one pollster in the C range was selected at random, Gravis Marketing, a few searches yielded the results of "The Worst Poll in America," who did not hone the right criteria for their polling, and admitted to changing the criteria on the same poll during the course of the survey of voters who planned to participate in the 2014 Midterm Primary for the United States Senator for the State of Kentucky. When identifying the reliability of other surveys conducted by other pollsters, SurveyMonkey allows the same user to take the same survey many times, thus skewing the results.

The greatest distribution of the survey grades falls within the B and C range, which contains reputable sources such as Gallup, the Pew Research Center, and the Public Religion Research Institute. One other observation is that the distribution of the standard deviation tends to be centered close to the mean when reviewing the distributions down the grading scale, however, those surveys ranked with a D have a wide array of standard deviations between the approval and

disapproval. The type of population responding to these survey questions could be indicative of the quality of the total survey results. When reviewing the summary statistics of the type of population that has been surveyed, the surveys have broken down the population of the same size on the following classifiers: (i) registered voters ("rv)", (ii) adults ("a"), (iii) likely voters ("lv"), and voters ("v"). The summary statistics for the type of population within the sample size surveyed is detailed within Table 2.

Table 2.

Summary Statistics of Surveys Based on Voter Type.

population	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Median
a	7900	adjusted_approve	40.7659510	2.7157200	25.1115860	51.8456490	40.9389070
		disapprove	53.4944051	3.3604387	35.0000000	75.9000000	53.0000000
lv	3370	adjusted_approve	42.1785098	2.8109554	32.6148690	54.7977090	41.7977090
		disapprove	52.0859456	2.6748638	41.0000000	62.0000000	52.0000000
rv	4654	adjusted_approve	42.5064247	2.1678856	33.2217510	52.1805830	42.4559710
		disapprove	53.8544177	2.5419821	37.0000000	63.0000000	54.0000000
v	2	adjusted_approve	40.0416980	0.6539903	39.5792570	40.5041390	40.0416980
		disapprove	55.0000000	0	55.0000000	55.0000000	55.0000000

When reviewing the summary statistics on the population of the voter type, it could be concluded that only two surveys were conducted with only voters, in which the standard deviations are not that great, which means that the data is clustered around the mean, whereas when compared to those surveyed of adults, the standard deviation is more spread out meaning that the opinions have differed through the course of the presidency. As such, a review of the survey data is to be examined, separated for each year in which Trump was President. Table three on the following page details the approval rating, broken out by the year the survey was conducted.

Table 3.

Summary Statistics of Surveys for Each Year.

Year Conducted	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Median
2017	2996	adjusted_approve disapprove	40.2486307 54.5836248	3.6203130 4.0940539	31.7094780 35.0000000	54.7977090 68.0000000	39.5749390 55.0000000
2018	3590	adjusted_approve disapprove	41.9899098 53.1056620	2.4614620 2.9595915	25.1115860 43.0000000	51.8456490 75.9000000	41.9934710 53.0000000
2019	4063	adjusted_approve disapprove	42.0595100 53.1758799	1.9245577 2.4448499	35.3770780 43.0000000	48.7977090 65.0000000	41.9879250 53.0000000
2020	5277	adjusted_approve disapprove	41.6678249 52.8043263	2.5479643 2.6740125	34.1030560 42.0000000	51.3676120 65.0000000	41.4559710 53.0000000

The mean of the approval ratings has remained static through the course of the prior four years, however, an anomaly that has been noted would be the standard deviation contained within the year 2019, which is indicative of relatively steady results in the survey aggregation. One additional item to highlight is the amount of surveys have increased steadily, approximately by about 500 surveys, from 2017 through 2019; however, the surveys conducted from January 1, 2020 through October 27, 2020 have increased by nearly double, and is projected to have nearly three times the amount of surveys conducted for 2020. One item to note is that the increase in the conducting of surveys could be due to the 2020 General Election cycle in the United States.

The items which have been reviewed during the data exploration phase has identified three major variables that appear to have a significant impact on the presidential approval ratings: (i) grade, (ii) population, and (iii) year.

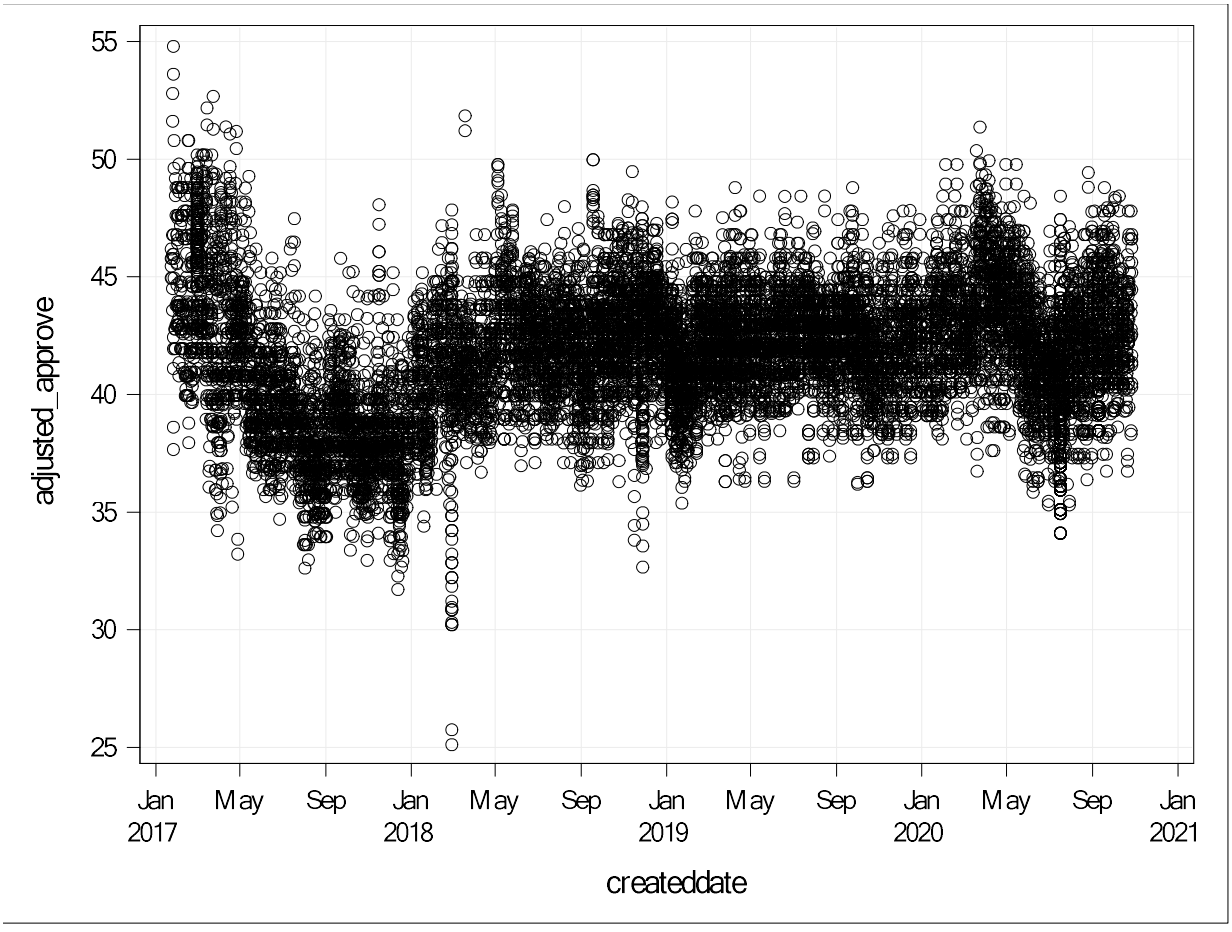
Model Planning

Upon exploration of the data, the model would need to be built around the parameters of the data that has been presented to address the business problems. The first thing to do is to create a scatter plot in order to identify relationships that would establish a correlation between the approval ratings and time based on an array of the different variables explored during the data

exploration phase. The two variables that will be focused in the creation and analysis of the scatterplots are the grading system of each survey conducted, and the type of population in which they have surveyed and its linear relationship from January 23, 2017 through October 27, 2020. Figure one will detail the approval ratings of President Trump in this period. Figures two through seven detail the linear correlation of the disapproval ratings as a whole and via grading tier during this period. Figures eight through twelve detail the linear correlation of the disapproval ratings as a whole and via population that has been sampled during this period the surveys were conducted.

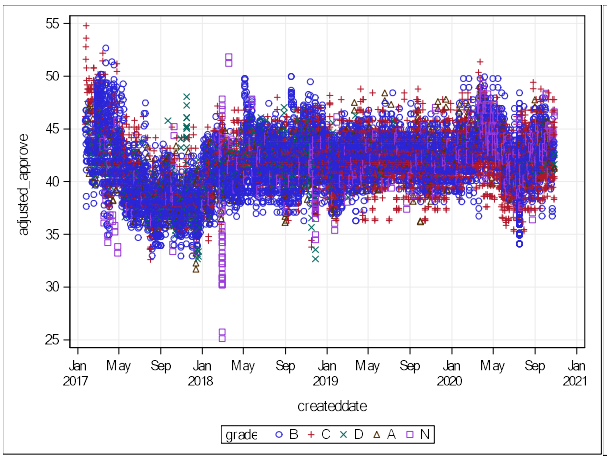
Figure 1.

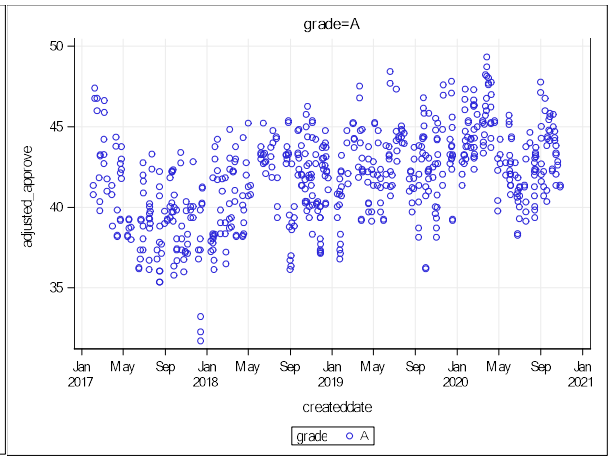
Scatter Plot of the Approval Ratings from January 23, 2017 through October 27, 2020.

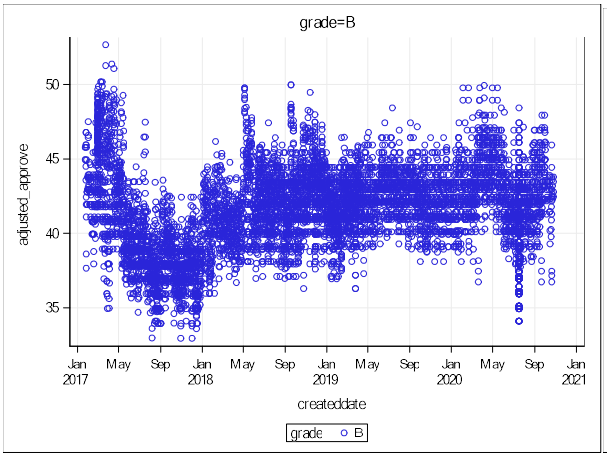


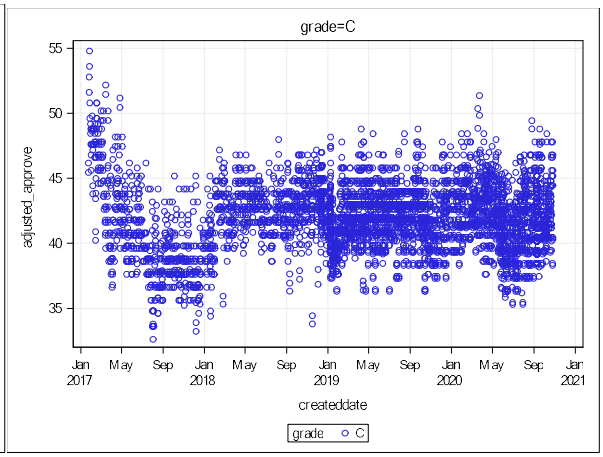
Figures 2 through 7.

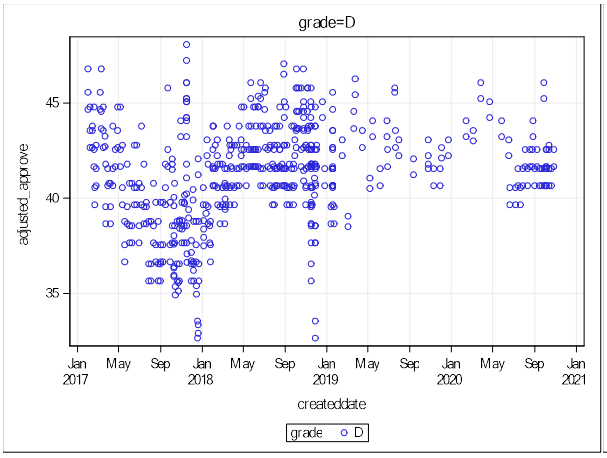
Scatter Plots of the Approval Ratings of President Trump as a Whole and via Grade.

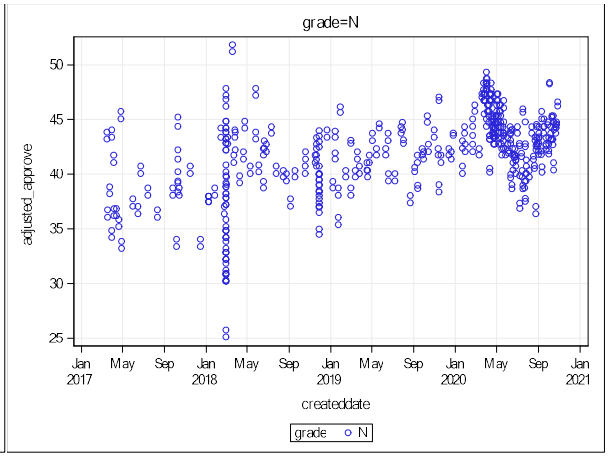






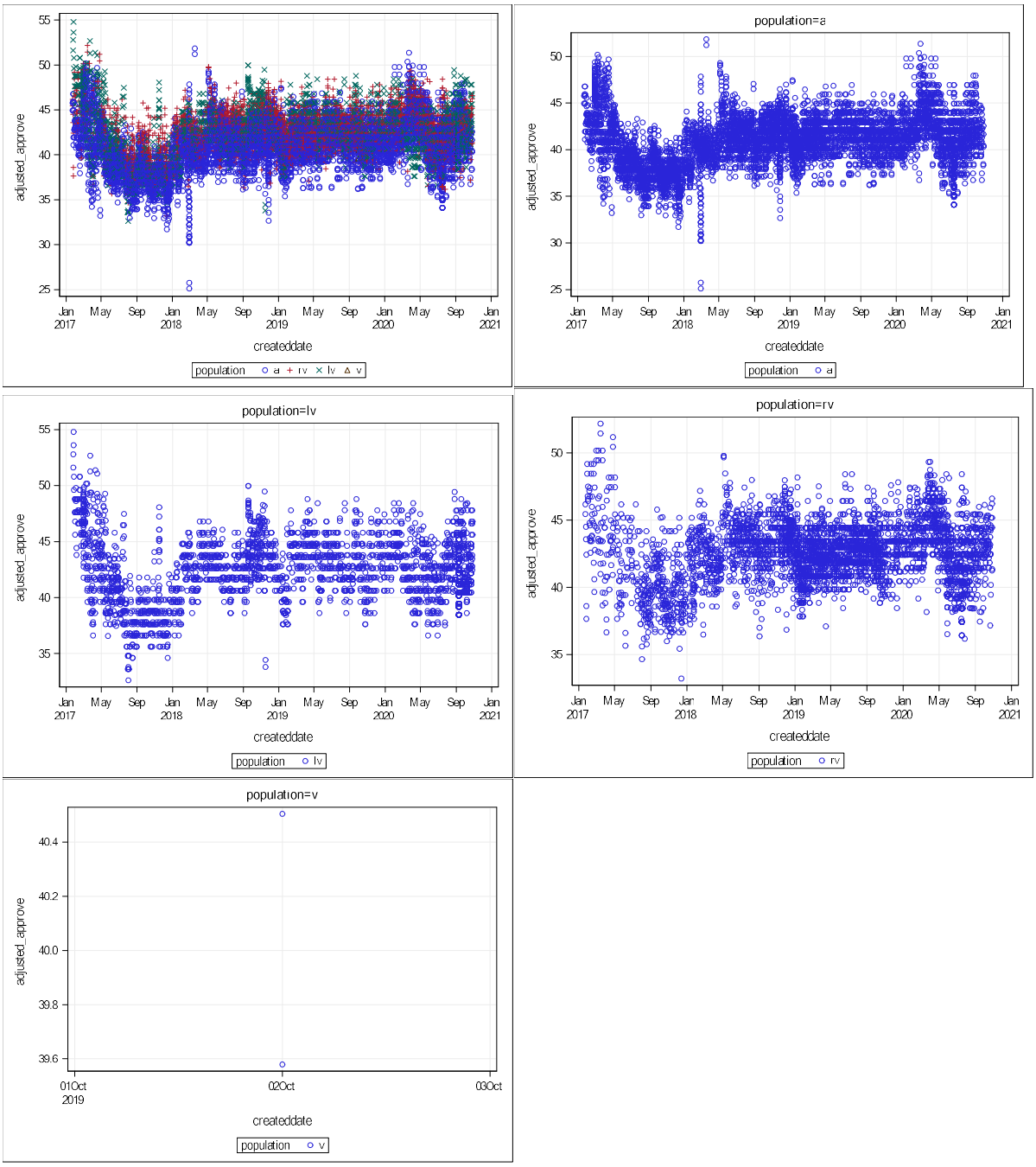






Figures 8 through 12.

Scatter Plots of the Approval Ratings of President Trump as a Whole and via Population.



While reviewing the data that has been compiled, the overall approval rating of President Trump appears to be falling between forty and fifty percent from all survey grades and voter types. When reviewing the survey clusters falling within the grade type, the majority of the data with the grade of A has a positive correlation, however, when reviewing the survey grades of B and C, it appears to be static between 40 to 45 percent approval. When the grading of the data decreases, the polarity of the survey approval ratings identifies many of the outliers and giving an appearance of a positive rating outlook.

When reviewing the data from the different types of population, the population of adults sampled spear to be much closer together, thus providing more reliability in the survey that was conducted. When reviewing the polls of likely voters and registered voters, the tend to fall in line with the rest of the data set; however, the population sample of voters is extremely polarizing with two samples conducted.

In order to determine the validity of the surveys conducted, there is ultimately a biased opinion based on the questions, the sample size, the selection of the participants comprising of the sample size. In order to root these potential items out, items that come into question regarding the validity of the surveys and the true approval ratings are:

1. Would the population type that has been sampled would have any type of bias with regards to the approval ratings and the reli? Based upon this business question, the null hypothesis would be:

   H_0: The sampling of different types of population would have an effect of the approval ratings,

   whereas the alternative hypothesis would be:

   H_A: The population type does not have any impact on the approval ratings.

2. How would the grading of the survey data affect reliability of the presidential approval data? Based upon this business question, the null hypothesis would be:

   H_0: The different grade types of the survey has an impact on the reliability of the survey collection of the presidential approval surveys,

   whereas the alternative hypothesis would be:

   H_A: The types of population sampled do not have any impact on the approval ratings.

3. How does the timing of conducting the survey affect reliability of the presidential approval data? Based upon this business question, the null hypothesis would be:

   H_0: The time in which the survey was conducted has an impact on the reliability of the survey collection of the presidential approval surveys,

   whereas the alternative hypothesis would be:

   H_A: The timing of the survey does not have any impact on the approval ratings.

The type of test that will be conducted to test the validity of the hypothesis in relation to the business questions would be the Analysis of Variance. This test will assess means of a continuous variable in two or more independent comparison groups. The technique to test for a difference in more than two independent means is an extension of the two independent samples procedure discussed previously which applies when there are exactly two independent comparison groups. The ANOVA technique applies when there are two or more than two independent groups (Sullivan). The table details the results of the ANOVA test for both the grade of the survey

Table 4.

Results of the ANOVA examination for the Grade, Population, and Year Conducted.

Source	DF	Type I SS	Mean Square	F Value	ρr > F	Accept or Reject H_0
grade	4	598.8077047	149.7019262	20.44	<.0001	Accept
population	3	10440.70855	3480.23618	519.11	<.0001	Reject
Year Conducted	3	6887.964445	2295.988148	331.44	<.0001	Reject

The table above details the results of the linear regression model, including the F Value, or > F, and whether the Null Hypothesis should be accepted or rejected. When reviewing the distribution of the Grade and the approval rating in the boxplot in Figure Thirteen, those with the letter grade of A had the majority of the outliers way below the Standard Deviation, while the majority of the outliers indicating the polarization of differing approval ratings were concentrated in the B and C range. When reviewing the Least Squares Means in Figure Fourteen, the data appears to be widely distributed in the graph that was produced, however, when reviewing the graph in the perspective of a scale from one to one hundred, it appears to be static. Figure Fifteen details the significance level for those surveys conducted, with the heavy reliability focused on those approval ratings between 41.5 percent and the grade levels of B and D. Overall, the Null Hypothesis would be accepted in this case with caution as the F Value is the lowest value when compared to variables.

When reviewing the distribution of the Population and the approval rating in the boxplot in Figure Sixteen, many of the outliers were distributed above and below the Standard Deviation for all population types. When reviewing the Least Squares Means in Figure Seventeen, the data appears to be widely distributed in the graph that was produced, however, when reviewing the graph in the perspective of a scale from one to one hundred, it appears to be static. Figure Eighteen details the significance level for those surveys conducted, with the heavy reliability focused on

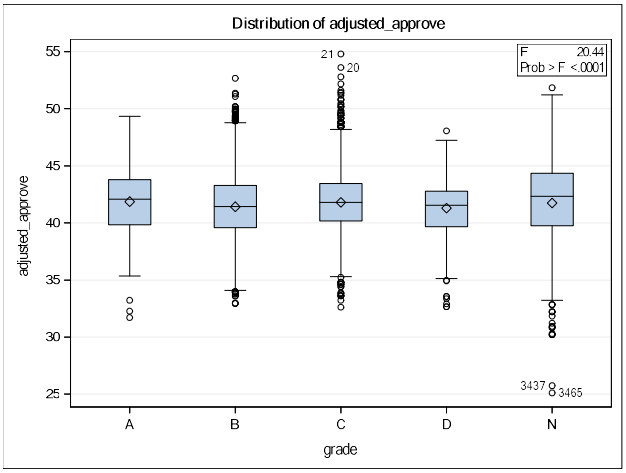
those approval ratings between 42.5 percent and the population of adults and likely voters. The Null Hypothesis would be rejected in this as the F Value is the greatest, and the Alternative Hypothesis of the population type does not have any effect of the approval ratings of President Trump.

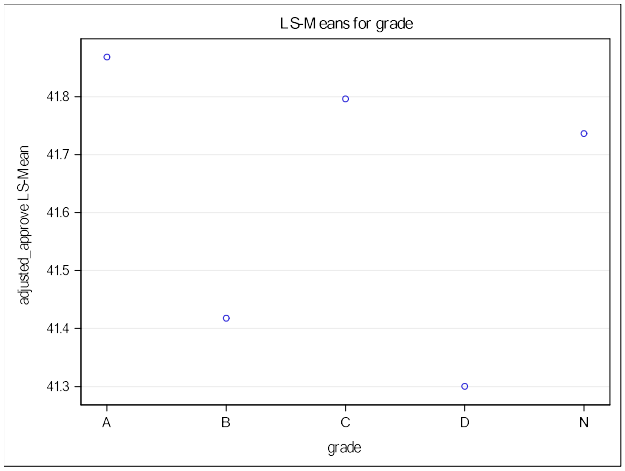
When reviewing the distribution of the Year and the approval rating in the boxplot in Figure Nineteen, the approval ratings appear to be very polarizing based on the occurrences of the outliers outside the Boxplots. When reviewing the Least Squares Means in Figure Twenty, the data appears to be widely distributed in the graph that was produced, however, when reviewing the graph in the perspective of a scale from one to one hundred, it appears to be static. Figure Twenty-one details the significance level for those surveys conducted, with the heavy reliability focused on those approval ratings between 42.0 percent and the year 2017 and 2020, which appears to be the inauguration of the President and the current election cycle. The Null Hypothesis would be rejected in this as the F Value is the greatest, and the Alternative Hypothesis of the year in which the survey was conducted would not have any effect of the approval ratings of President Trump.

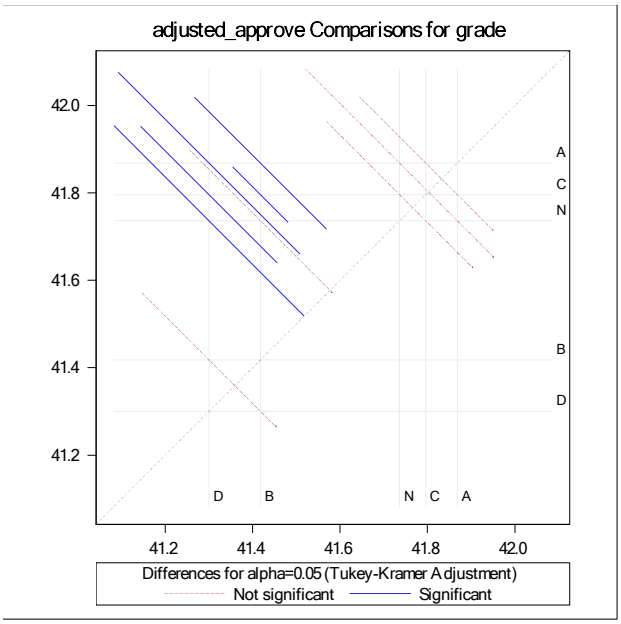
When reviewing the forecasted model detailed in Figure 22, the overall approval ratings for President Trump is predicted to be within the 45 percent range, while the overall predicted ratings within the 95 percent confidence level is projected to be within 39.4 percent to 50.5 percent. This could be indicative based on the response to a variety of current events, including the potential racial unrest, COVID-19 response, approval of the recent Supreme Court Justice Amy Coney Barret, and the ongoing Russian and Iranian interference causing lack of confidence in the voting process.

Figures 13 through 15

Boxplot Detailing the Distribution of the Approval Rating, the Least Square Means, and the Adjustment for Multiple Comparisons Based on the Data Grade.

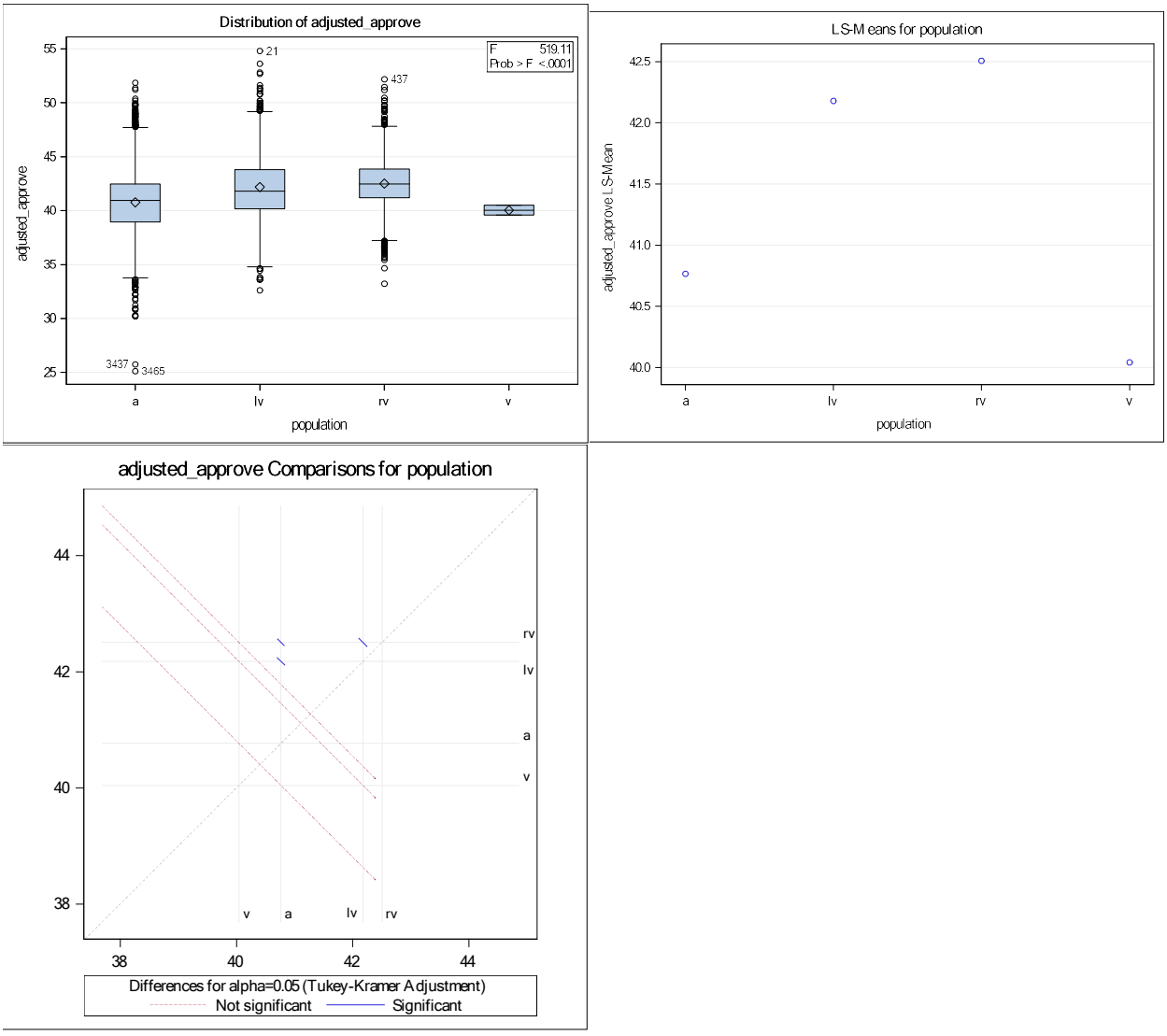






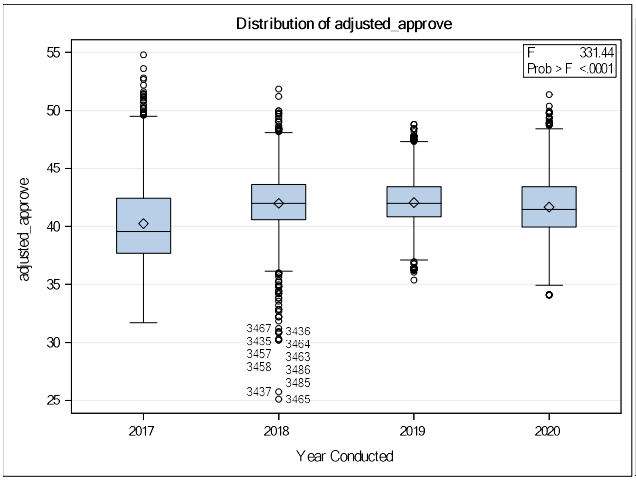
Figures 16 through 18

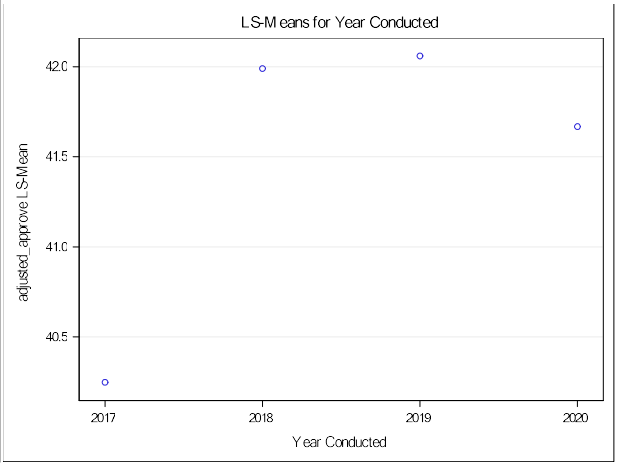
Boxplot Detailing the Distribution of the Approval Rating, the Least Square Means, and the Adjustment for Multiple Comparisons Based on the Population Surveyed.



Figures 19 through 21

Boxplot Detailing the Distribution of the Approval Rating, the Least Square Means, and the Adjustment for Multiple Comparisons Based on the Time Period the Survey Was Conducted.





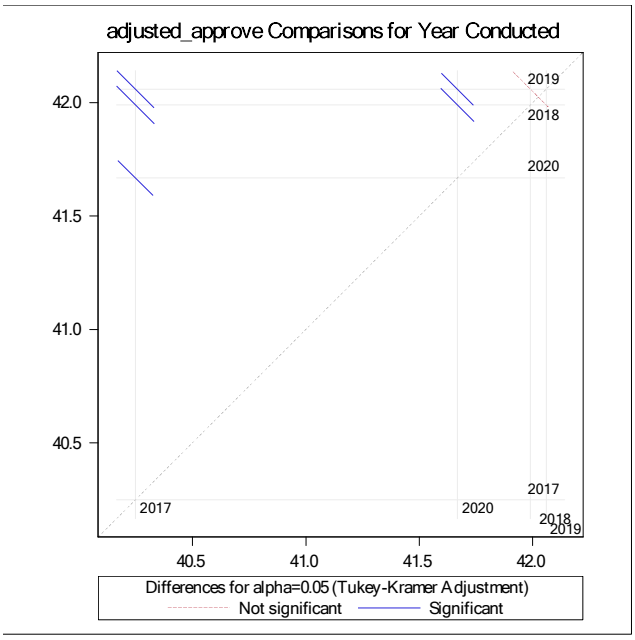
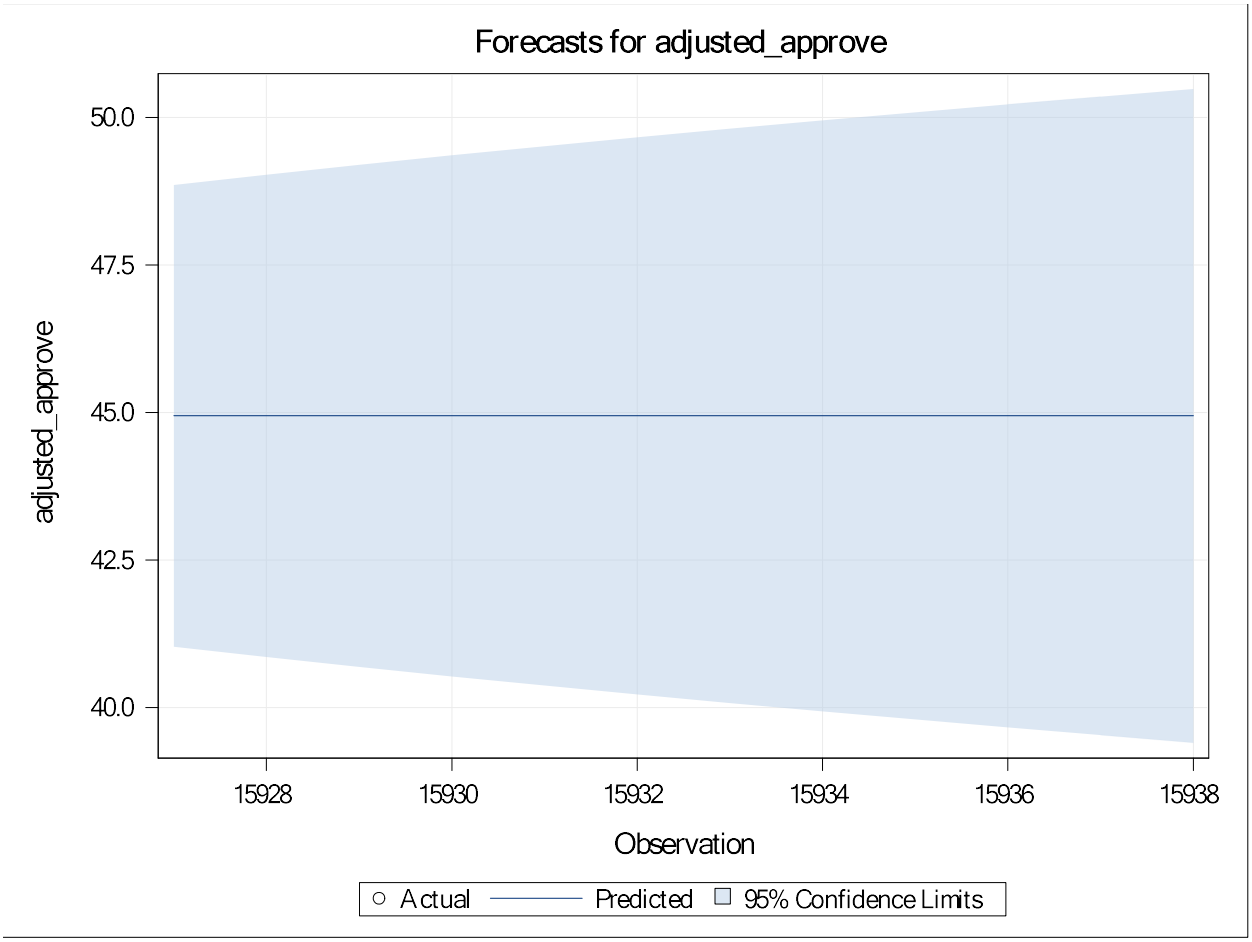


Figure 22

Forecasted Model of President Trump's Approval Ratings Through the End of Year 2020.



Communicate Results

Even through the ANOVA model supports the Null Hypothesis of the survey grade and how it effects the approval ratings, the Exponential Smoothing Model an overall grim rating of President Trump when evaluating the overall ratings, regardless of the population and the Survey Grade. As a company, bracing for the worst should be arranged in such a way that there would still be profits during this pandemic.

When evaluating additional requirements to determine the validity of the survey results, the survey ratings with a Grade of A and B appear to be the most reliable in yielding the most

accurate response. When reviewing those whom conducted the Survey within those grading levels, CBS News, ABC News, Fox News, NBC News, and CNN, most likely reputable sources of news literacy, regardless of political leaning, have the highest grades. Those with the lower rankings do not yield the highest confidence, such as SurveyMonkey, the Kaiser Family Foundation, and Cards Against Humanity, may offer skewed data sources, a particular leaning of political views when selecting the survey population, or the wording of the questions will direct them to answer the question in a predicted manner (Home 2020).

Operationalize

Additional details would need to be examined when evaluating the presidential approval ratings. Obviously, the political party affiliation and political leaning should not be included as the ideology would offer a predicted answer; however, items such as detailing the median age, income level, years of education, years of employment, percent of each gender sample, and other demographic information would provide key insights on to the approval ratings of the current president. Further exploration of the data with these factors incorporated within the data set. This type of demographic information would be relevant to identify the sampling of the population when compared to the average demographic indicators of the United States as a whole.

References

2020 Election Poll. Survey. https://www.surveymonkey.com/r/R8Z2PZ2.

Carroll, J. (2005). President approval vs. favorability ratings. The Gallup Poll Tuesday Briefing, 3.

Cohen, J. E. (2019). Polls and Elections: Presidential Referendum Effects in the 2018 Midterm Election: An Initial Analysis. Presidential Studies Quarterly, 49(3), 669–683. https://doi.org/10.1111/psq.12579

Home. Ad Fontes Media. (2020, September 1). https://www.adfontesmedia.com/.

Our Data. FiveThirtyEight. (2018, February 9). https://data.fivethirtyeight.com/.

Sullivan, L. Introduction. Hypothesis Testing - Analysis of Variance (ANOVA). https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html.

Trump, M. L. (2020). Too Much and Never Enough: How My Family Created the World's Most Dangerous Man. SIMON SCHUSTER.

Weigel, D. (2014, May 21). The Worst Poll in America. Slate Magazine. http://www.slate.com/blogs/weigel/2014/05/21/the_worst_poll_in_america.html.