PEORIA Project
2020 Election Predictions Model Results

Super Tuesday- Democratic Party

Sanders Leads in Super Tuesday Models By Varying Margins

Innovative Model Incorporates Social Media Variable

of Twitter Mentions to Yield Ranges of Likely Results

MARCH 2020 REPORT FOR SUPER TUESDAY

Our Research Questions

The purpose of this PEORIA research initiative is to construct, test, and adjust a predictive model of election results which includes a social media variable. Our current focus:

1. How best can Twitter mentions, a variable that measures buzz among people paying close attention to a contest, be incorporated into a model?
2. How important are previous election results in affecting buzz and outcomes? We compare predictions from two models: one that includes a “momentum” variable keyed to results and one that does not.
3. How well do our Twitter models fare against traditional models that would use polls as a key predictor instead of Twitter Mentions? Can the Twitter model help to overcome known limitations in polling?

Our Research Questions

Current State of Models

1. How best can Twitter mentions, a variable that measures buzz among people paying close attention to a contest, be incorporated into a model?
   - It improves some fundamental models. See appendix B.
2. How important are previous election results in affecting buzz and outcomes? We compare predictions from two models: one that includes a “momentum” variable keyed to results and one that does not.
   - Currently, the basic model has outperformed the momentum model. See appendix C.
3. How well do our Twitter models fare against traditional models that would use polls as a key predictor instead of Twitter Mentions? Can the Twitter model help to overcome known limitations in polling?
   - Our twitter-based models are not outperforming poll-based models. Using both in a model is problematic as doing so presents a multicollinearity problem. See appendix F.

Our Key Three or Four Variables

Our models predict a candidate's performance based on three or four factors (depending on the model): Twitter mentions, cash on hand, endorsements, and performance in the last nomination contest.

While we are aware that in important ways the Twitter universe does not necessarily reflect the electorate, the quantity of **Twitter Mentions** is a good proxy for the "buzz" a candidate is getting within the wider electorate, and reflects the activity of important opinion leaders.

Cash on Hand reflects the strength of the candidate in the "money primary."

Endorsements indicate each candidate's strength within the party, which speaks to the debate over whether the party decides the outcome of the nomination.

Performance in the Last Nomination Contest is the vote share received in the immediately preceding primary or caucus.

Two Models that Depict the Possibilities after Biden's SC Win

We present two models predicting the outcome of the Super Tuesday Primary. Both models predict the individual vote share of each state weighted by the number of delegates of both states:

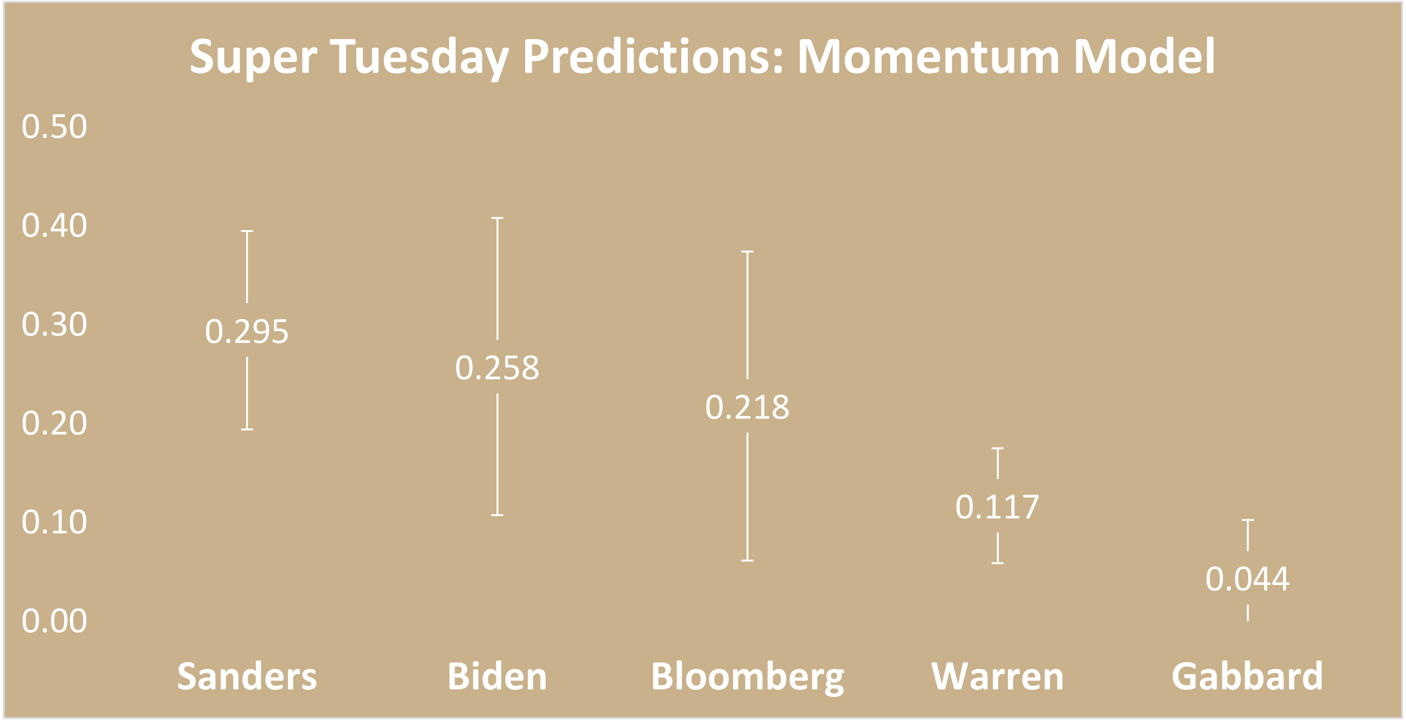
The first model captures the "momentum" of the race, incorporating the results of the most previous primary election. Thus, Sanders and Biden have approximately equal chance of winning, with Biden boosted from SC results.

The second basic model, without accounting for momentum, predicts a Sanders win in Super Tuesday. However, due to January cash on hand, the prediction for Bloomberg is for a strong showing. Moreover, many states have early voting, so late-breaking momentum shifts (such as that for Biden after his performance in South Carolina) will not affect the large number of ballots already cast.

Large error estimates occur for Biden (in the Momentum model) and Bloomberg (in both models) due to recent electoral results and financial reporting (respectively).

It is likely that the actual vote share may fall somewhere in between our models, as early voters may match more closely with the basic model, while later voters will be more affected by momentum.

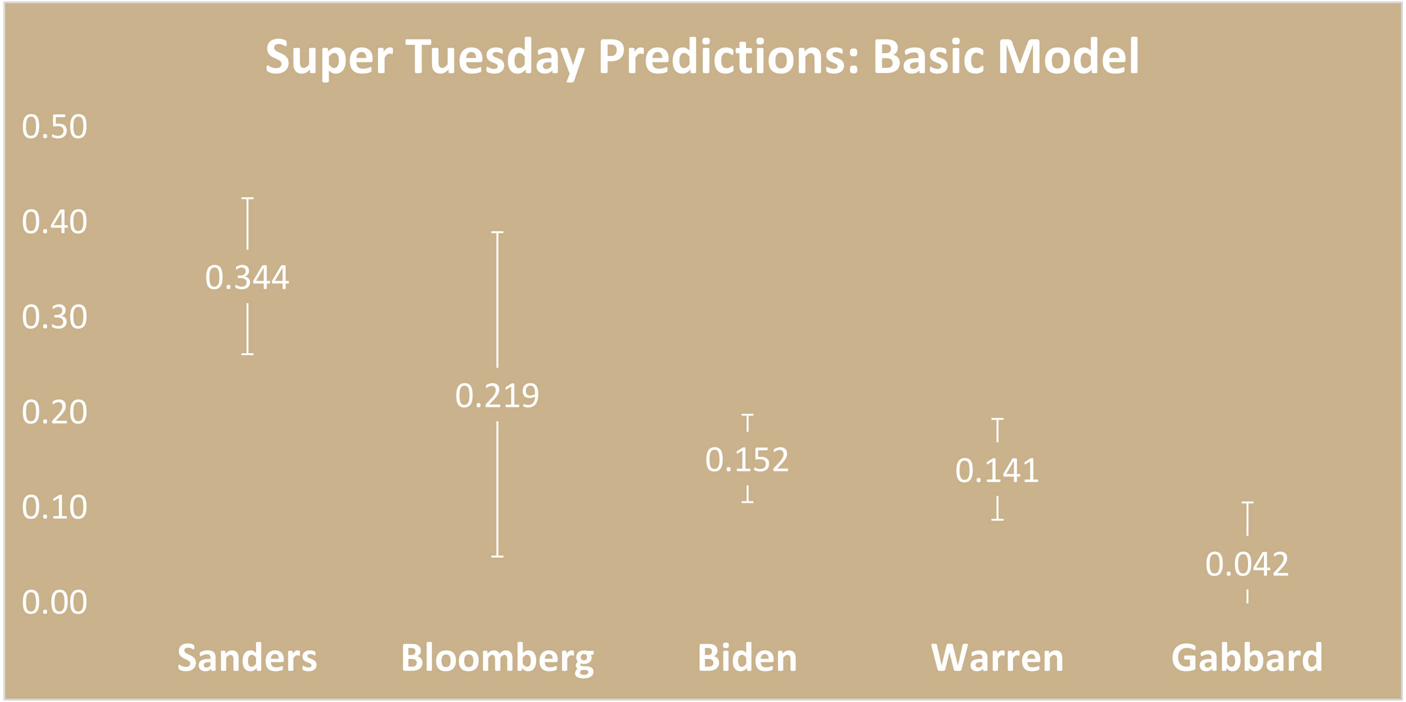
Predicting Super Tuesday Delegate Share: The Momentum Model



Candidate	Average Predicted Vote Share	Lower Bound	Upper Bound
Sanders	**0.295**	0.194	0.330
Biden	**0.258**	0.108	0.408
Bloomberg	**0.218**	0.062	0.374
Warren	**0.117**	0.059	0.175
Gabbard	**0.044**	-0.014	0.103

The chart and table report the predicted Super Tuesday vote share weighted by delegates for each candidate. For example, Bernie Sanders is predicted to receive 29.5% of the vote share. The bars indicate the upper and lower bounds for the prediction (95% confidence interval).

Predicting Super Tuesday Delegate Share: The Basic Model



Candidate	Average Predicted Vote Share	Lower Bound	Upper Bound
Sanders	**0.344**	0.262	0.426
Bloomberg	**0.219**	0.049	0.390
Biden	**0.152**	0.106	0.198
Warren	**0.141**	0.088	0.194
Gabbard	**0.042**	-0.021	0.106

The chart and table report the predicted Super Tuesday vote share weighted by delegates for each candidate. For example, Bernie Sanders is predicted to receive 34.4% of the weighted vote share. The bars indicate the upper and lower bounds for the prediction (95% confidence interval).

Explanation of Models

What Our Models Do

Our models predict the Super Tuesday vote share weighted by the pledged delegates from each state for the Democratic candidates using three or four predictor variables generated by an equation estimated through an Ordinary Least-Squares (OLS) multiple regression. See the following pages for equations.

How We Predict Vote Share

In order to predict each candidate's vote share, we input the latest variable data (see below) into the regression model to generate an estimate as well as an upper- and lower-bound for the predicted performance of each candidate.

Twitter Mentions: Measured as the number of mentions on Twitter for each candidate as a percentage share of the total number of mentions for all candidates within the party. The data for these models were tallied through one month leading up to the week prior to the date of the contest. We note that overall Twitter mentions are not the same as the tone of those mentions – for example, much recent Twitter chatter about Sanders has been negative. However, the total volume of chatter is indicative of his strong position in the field. Moreover, we find measurements of tone on Twitter to be problematic (both those supplied by Crimson Hexagon as well as a classification model). Source for data: Crimson Hexagon.

Cash on Hand: Measured as a percentage share of the total cash on hand for all candidates within the party. The most recent data were for January, 2020. Source for data: FEC.gov

Endorsements: Measured as the total number of endorsements for each candidate by US Senators, members of the US House of Representatives, former Presidents and Vice Presidents, former presidential candidates from the current election cycle who had dropped out of the race, elected statewide officials, state legislative leaders, and mayors of large cities. Source for data: FiveThirtyEight.com

Performance in the Last Nomination Contest : Measured as each candidate's share of the total vote within the party in the immediately preceding caucus or primary.

How We Chose Our Model

To find the best fitting model, we used campaign data from 2012 and 2016 for the predictor variables above with Super Tuesday vote share weighted by the delegates from each state for each year as the dependent variable. Several models and independent variables were tested in a series of trials. Independent variables included those used here as well as many other variables used in previously published studies (see Appendix E). Independent variables that performed consistently well across different models (with large and significant coefficients) were selected for inclusion. Types of models included OLS, beta, longitudinal (using Q1 through Jan cash on hand as well as monthly twitter mentions), lasso, ridge, logistic, partial least squares, and principal component regressions. At this point, the model with the lowest RMSE while maintaining the highest possible R^2 was chosen for this report (in this case, OLS regression with the variables noted).

Descriptive Table of Variables and Regression Model for Super Tuesday: The Momentum Model

Means, Standard Deviations, and Correlations

Variable	*M*	*SD*	*1*	*2*	*3*
1. Twitter Mentions	0.250	0.183			
2. Cash on Hand	0.250	0.233	0.522		
3. Endorsements	265.583	349.788	0.300	0.739	
4. Last Primary Vote Share	0.246	0.206	0.764	0.668	0.707

Summary of Momentum Model of Super Tuesday Predictions

Variable	*Estimate*	*SE B*	*β*
Intercept	0.02965	0.02571	
Twitter Mentions	0.29716	0.16212	0.34944
Cash on Hand	0.24910	0.10901	0.37164
Endorsements	-0.00002	0.00009	-0.04607
Last Primary Vote Share	0.31381	0.18046	0.41509

adj $R^2 = 0.899$, $F(4,7) = 25.54$

$^{*}p < 0.05$. $^{**}p < 0.01$.

- The equation representing the model is:
- Predicted Vote Share = 0.02965 + (0.29716 * Twitter Mentions) + (0.24910 * Cash on Hand) + (-0.00002 * Endorsements) + (0.31381 * Most Previous Results)
- We can interpret the Twitter coefficient as such: As one candidate increases their share of Twitter by 1%, their vote share is predicted to increase by .0029716.

Descriptive Table of Variables and Regression Model for Super Tuesday: The Basic Model

Means, Standard Deviations, and Correlations

Variable	*M*	*SD*	*1*	*2*
1. Twitter Mentions	0.250	0.183		
2. Cash on Hand	0.250	0.233	0.522	
3. Endorsements	265.583	349.788	0.300	0.739

Summary of Basic Model of Super Tuesday Predictions

Variable	*Estimate*	*SE B*	*β*
Intercept	0.02853	0.02877	
Twitter Mentions	0.52371**	0.10801	0.61583
Cash on Hand	0.21990	0.12056	0.32808
Endorsements	0.00009	0.00007	0.19976

adj R^2 *= 0.874, F(3,8) = 26.37*

*p < 0.05. **p < 0 .01.

- The equation representing the model is:
- Predicted Vote Share = 0.02853+ (0.52371 * Twitter Mentions) + (0.21990 * Cash on Hand) + (0.00009 * Endorsements)
- We can interpret the Twitter coefficient as such: As one candidate increases their share of Twitter by 1%, their vote share is predicted to increase by 0.0052371.

Thanks for reading!
Come back each week for new predictions!

Questions may be directed to the authors:

Meg O'Neill
megoneill@gwu.edu
Data Scientist for The PEORIA Project,
Graduate School of Political Management,
The George Washington University

Michael Cornfield
corn@gwu.edu
Associate Professor and Director of The PEORIA Project,
Graduate School of Political Management
The George Washington University

Todd Belt
tbelt@gwu.edu
Professor and Director of the
Political Management Master's Program,
Graduate School of Political Management
The George Washington University

Lara Brown
larambrown@gwu.edu
Associate Professor and Director of the
Graduate School of Political Management,
The George Washington University

Table of Contents for Appendices

Appendix A: Twitter Data

2012 Candidate	Twitter Mentions	Share within Party
Mitt Romney	248230	0.320
Rick Santorum	203436	0.262
Ron Paul	196965	0.254
Newt Gingrich	127232	0.164
Fred Karger	259	0.000

2016 Candidate	Twitter Mentions	Share within Party
Donald Trump	1794146	0.481
Bernie Sanders	1264727	0.551
Ted Cruz	1199357	0.322
Hillary Clinton	1030050	0.449
Marco Rubio	426797	0.115
Ben Carson	208389	0.056
John Kasich	97918	0.026

2020 Candidate	Twitter Mentions	Share within Party
Donald Trump	4428180	0.993
Bernie Sanders	3998430	0.513
Elizabeth Warren	1545778	0.198
Joe Biden	1481891	0.190
Michael Bloomberg	639217	0.082
Tulsi Gabbard	126384	0.016
Bill Weld	30400	0.007
Rocky De La Fuente	347	0.000

Data tallied for one month exactly one week prior to the primary, but because Super Tuesday fell on different days each year, the time period did not match for each year.
For 2012, data was tallied from January 28 – February 28.
For 2016, data was tallied from January 23 – February 23.
For 2020, data was tallied from January 25 – February 25.

Appendix B: Can Twitter "Buzz" Be Incorporated in a Prediction Model?

GW | GSPM

- Take Away: Fundamental models are slightly improved by adding Twitter data.
- Our basic model was able to explain on average 80.4% of the variance in 2012 and 2016 vote share (aggregated: IA, NH, NV, SC). Additionally, the average 2020 prediction error (RMSE) was 8.5%.
- Mayer's primary model (albeit not originally intended for individual states), was able to explain 77.1% of the variance in vote share. The average prediction error was 11.6%.
    - The model is improved by the addition of twitter data. Specifically, with twitter it explains 84.4% of the variance in the 2012 and 2016 vote share. The average prediction error with twitter the prediction error decreases to 9.1%.
- Dowdle and Adkin's model (also not intended for individual states) was able to explain 91.7% of the variance in 2012 and 2016 vote share. The 2020 prediction error was 6.4%.
    - The model's 2020 prediction were slightly improved with Twitter with a prediction error of 5.7%. It explained less of variance of the 2012 and 2016 vote share at 91.5%
- But, for a great review on why Twitter data should not be used, see Gayo-Avello's 2012 critique on Twitter data in models which mentions the fact Twitter does not map directly onto voters, the problems with using aggregated numbers from Twitter, and more. Also, for a discussion of how the Twitterverse differs from the electorate, see Belt & Cornfield, 2019.

Appendix C: Does momentum improve models of primaries?

Model	Prediction Error 2020	Variance Explained 2012 & 2016
Basic Model	8.5%	80.4%
Momentum Model	9.0%	77.9%

In our models, the "momentum" model – accounting for the most previous results in the primary schedule, performs worse. It has more prediction error and explains less variance in previous years.

Appendix D: How do the models compare in predicting the primaries?

Model	Iowa	New Hampshire	Nevada	South Carolina	Super Tuesday	Aggregated States
Basic	0.926 0.067	0.882 0.079	0.817 0.060	0.520 0.132	0.874	0.804 0.085
Momentum	N/A – first primary	0.877 0.082	0.806 0.061	0.534 0.127	0.899	0.779 0.090
Mayer's Model **	0.677 0.119	0.845 0.151	0.836 0.115	0.601 0.080	0.897	0.771 0.116
Mayer's Model with Twitter	0.907 0.077	0.923 0.108	0.878 0.082	0.618 0.097	0.895	0.844 0.091
Dowdle and Adkin's Model **	0.983 0.038	0.958 0.058	0.827 0.084	0.891 0.077	0.928	0.917 0.064
Dowdle and Adkin With Twitter	0.986 0.037	0.955 0.058	0.838 0.064	0.881 0.067	0.916	0.915 0.057
fivethirtyeight **	N/A 0.036 (Jan 31)	N/A 0.043 (Feb 6)	N/A 0.040 (Feb 20)	N/A 0.034 (Feb 23)	N/A	0.038

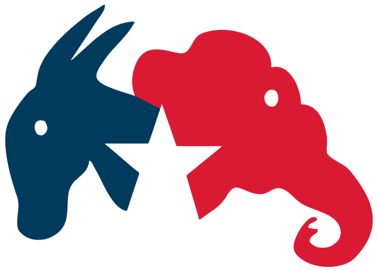
black = variance explained in 2012 and 2016, higher numbers correspond to better explanation of past performance

red = prediction error, lower numbers correspond to better predictions

** citation on next page

Although the basic model preforms well, Dowdle and Adkin's model outperforms both models with Twitter information.

However, Nate Silver's Fivethirtyeight team's prediction model outperforms all others and, unlike Mayer and Dowdle and Adkin's models was designed to predict individual state outcomes.



Appendix E: Citations

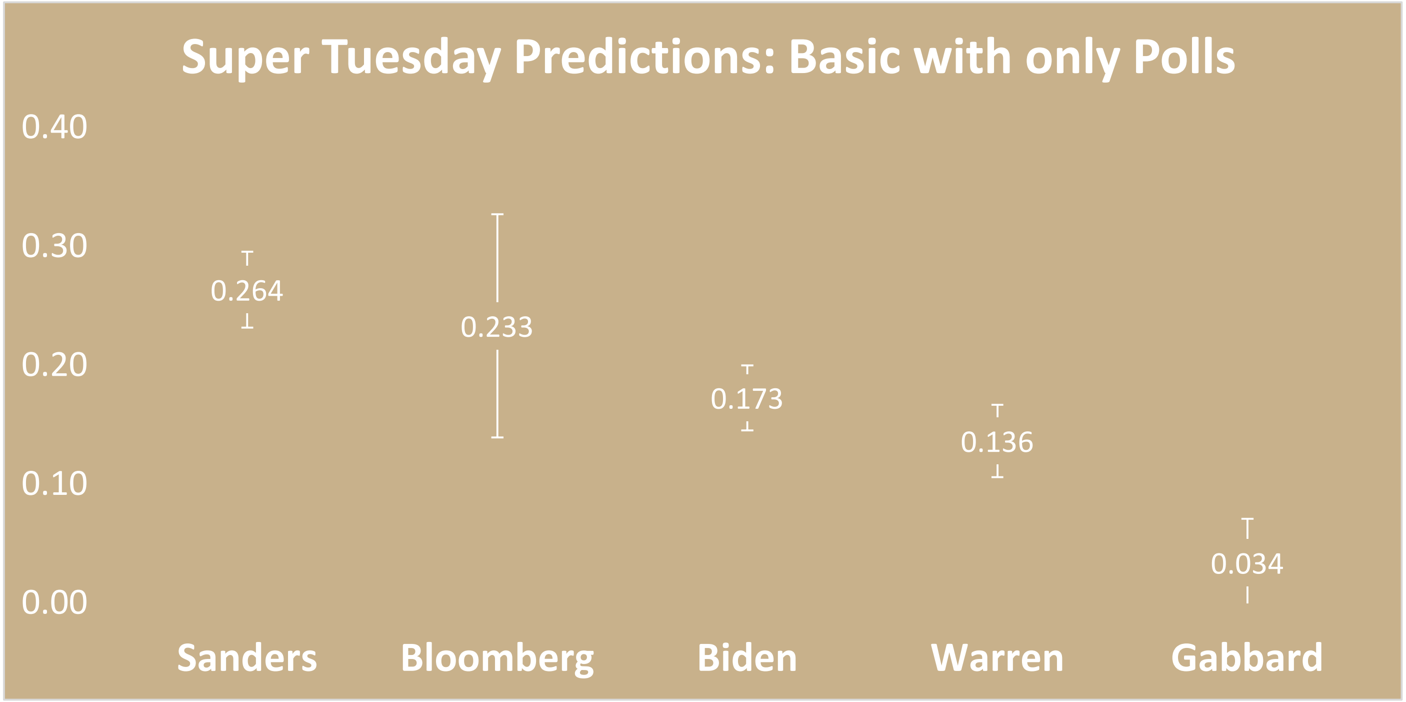
- Mayer model, see: https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/forecasting-presidential-nominations-or-my-model-worked-just-fine-thank-you/38859B253BF81BDFE81189E481AC0C42

- Dowdle and Adkins model, see: https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/forecasting-presidential-nominations-in-2016-wepredictedclintonandtrump/5BAB9897BFFB4D4A9D5B06F9633DC1F9

- Fivethirtyeight, see: https://projects.fivethirtyeight.com/2020-primary-forecast/

- Gayo-Avello's critique, see: https://arxiv.org/abs/1204.6441

Appendix F: Predicting Super Tuesday Delegate Share: The Basic Model with Polls instead of Twitter



Candidate	Average Predicted Vote Share	Lower Bound	Upper Bound
Sanders	**0.264**	0.232	0.296
Bloomberg	**0.233**	0.139	0.327
Biden	**0.173**	0.145	0.200
Warren	**0.136**	0.106	0.167
Gabbard	**0.034**	-0.003	0.071

The chart and table report the predicted vote share weighted by delegates on Super Tuesday for each candidate. For example, Bernie Sanders is predicted to receive 26.4% of the weighted vote share. The bars indicate the upper and lower bounds for the prediction (95% confidence interval).

Appendix G: Descriptive Table of Variables and Regression Model for Super Tuesday: The Basic Model with Polls instead of Twitter

Means, Standard Deviations, and Correlations

Variable	*M*	*SD*	*1*	*2*
1. Polls	0.250	0.173		
2. Cash on Hand	0.250	0.233	0.737	
3. Endorsements	265.583	349.788	0.643	0.739

Summary of Regression of the SC Primary Vote Share Prediction

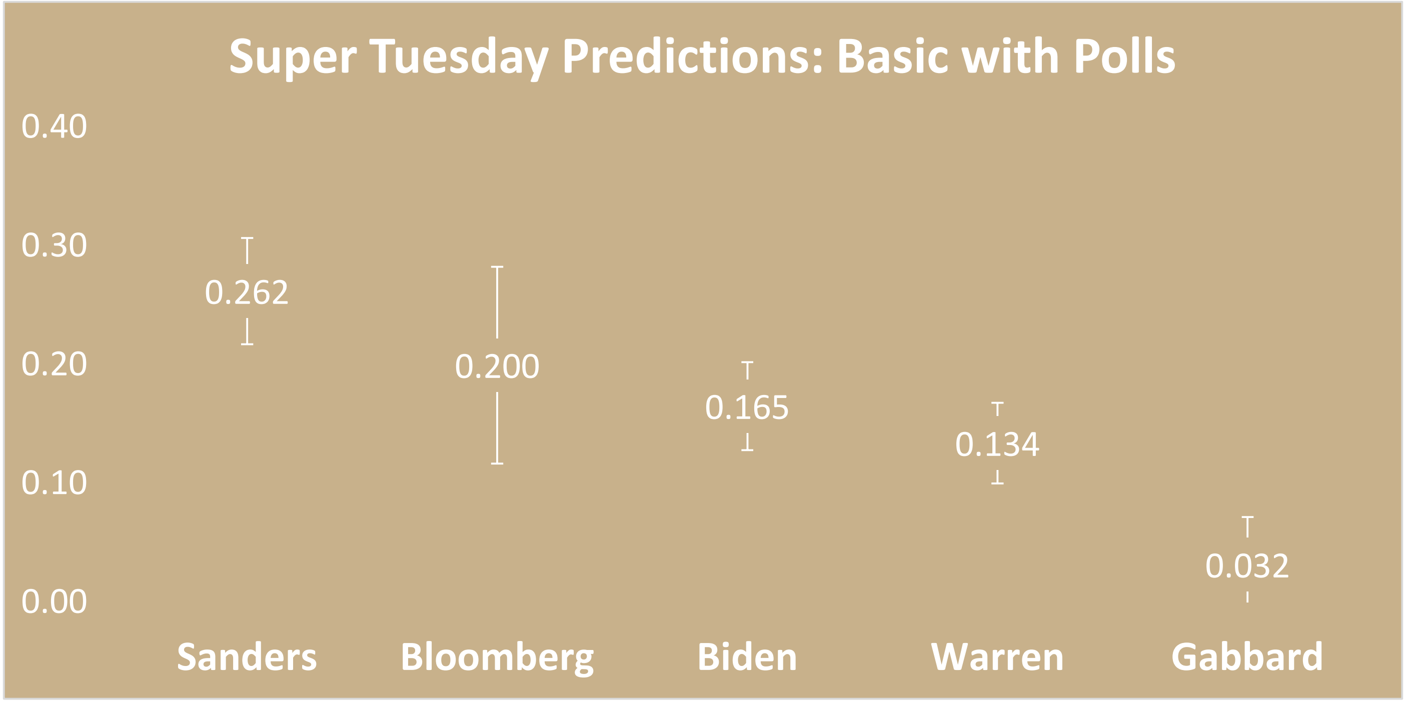
Variable	*Estimate*	*SE B*	*β*
Intercept	0.01526	0.01699	
Polls	0.78581**	0.08488	0.87176
Cash on Hand	0.15564	0.07176	0.23220
Endorsements	-0.00005	0.00004	-0.10540

adj R^2 = 0.958, *F*(3,8) = 83.7

**p* < 0.05. ***p* < 0 .01.

- The equation representing the model is:
- Predicted Vote Share = 0.01526+ (0.78581 * Polls)+ (0.15564 * Cash on Hand) + (-0.00005 * Endorsements)
- We can interpret the Polls coefficient as such: As one candidate increases their share of Polls by 1%, their vote share is predicted to increase by 0.0078581.
- VIF for each of the variables are below 5 and Durbin Watson test is not significant.
- At this point, our use of Twitter instead of polls does not do a better job of predicting outcomes, but we continue to analyze and refine our model as results come in throughout the primary season.

Appendix H: Predicting Super Tuesday Delegate Share: The Basic Model with Polls



Candidate	Average Predicted Vote Share	Lower Bound	Upper Bound
Sanders	**0.262**	0.217	0.307
Bloomberg	**0.200**	0.117	0.282
Biden	**0.165**	0.128	0.202
Warren	**0.134**	0.100	0.168
Gabbard	**0.032**	-0.009	0.072

The chart and table report the predicted vote share weighted by delegates on Super Tuesday for each candidate. For example, Bernie Sanders is predicted to receive 26.2% of the weighted vote share. The bars indicate the upper and lower bounds for the prediction (95% confidence interval).

Appendix I: Descriptive Table of Variables and Regression Model for Super Tuesday: The Basic Model with Polls

GW | GSPM

Means, Standard Deviations, and Correlations

Variable	*M*	*SD*	*1*	*2*	*3*
1. Twitter Mentions	0.250	0.183			
2. Cash on Hand	0.250	0.233	0.522		
3. Endorsements	265.583	349.788	0.300	0.739	
4. Polls	0.250	0.173	0.853	0.737	0.643

Summary of Basic Model of Super Tuesday Predictions Adding Polls

Variable	*Estimate*	*SE B*	β
Intercept	0.01452	0.01788	
Twitter Mentions	0.06988	0.13572	0.08217
Cash on Hand	0.15477	0.07532	0.23090
Endorsements	-0.00003	0.00005	-0.06996
Polls	0.70296**	0.18393	0.77984

adj R^2 *= 0.953, F(4,7) = 57.07*

**p* < 0.05. ***p* < 0 .01.

- The equation representing the model is:
- Predicted Vote Share = 0.01452+ (0.06988 * Twitter Mentions) + (0.15477 * Cash on Hand) + (-0.00003 * Endorsements) + (0.70296 * Polls)
- We can interpret the Twitter coefficient as such: As one candidate increases their share of Twitter by 1%, their vote share is predicted to increase by 0.0006988.
- Multicollinearity is a problem in the model with VIF of Polls = 9.796 and VIF of Tweets = 5.993
- Durbin Watson tests conducted in R are not significant.