@ NYC Data Science Academy

(2nd August 2017)

Senti-Meter

by

Shivakumar Ranganathan (*Kumar*)

1 Overview

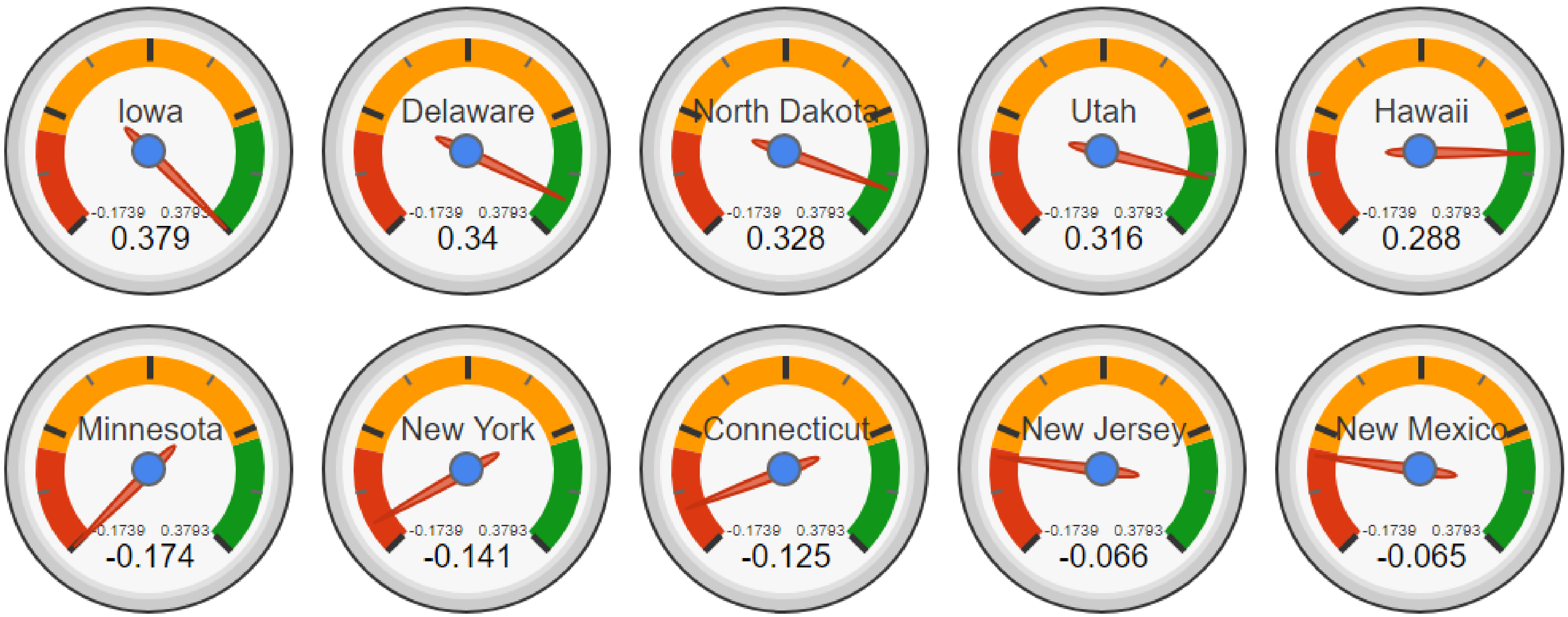
2 Methodology

3 Data Visualization & Analysis

4 Conclusions & Future Work

1 Overview

- Senti-Meter=Sentiments (tweets) + Measurement (use of positive and negative words)



1 Overview

- Objective(s)—

  a) **Scrape** tweets using **Selenium**

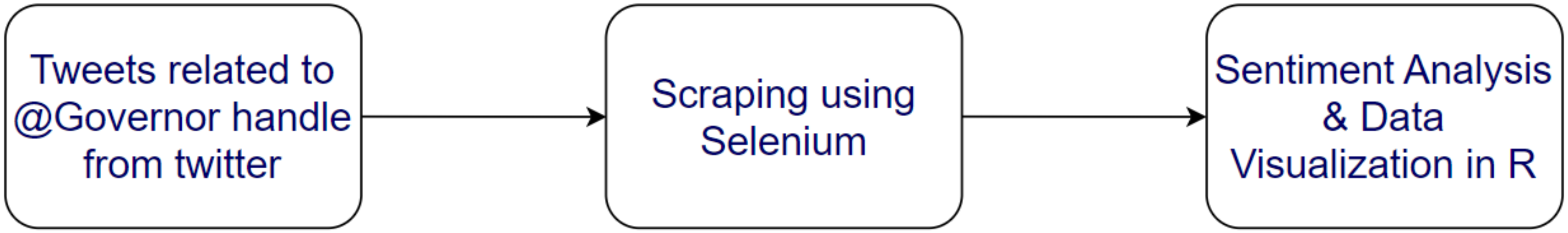
  b) **Quantify** the **sentiment** of people towards **governors** of all 50 States in US.

  c) **Rank** the Governors based on a sentiment analysis of the tweets and **compare** the **rankings** with the recently published Morning Consult Governor Approval rankings[1]

  d) Provide **insight** on **performance** that pertain to **age**, **party affiliation** and **gender** of governors

[1] https://morningconsult.com/governor-approval-ratings-july-2017/

2 Methodology



- Web scraping—

  a) **Selenium** used to scrape 43,737 tweets [target was to extract 1000 tweets for each governor]

  b) All **sampled tweets** had **reference** to @Governor (twitter verified account) handle. This kept out irrelevant tweets and tweets from people with similar names as the Governors

2 Methodology

- Sentiment Analysis & Data Visualization—

    a) **R-Studio** was used for postprocessing the scraped data

    b) Each tweet was cleaned up (using ‘gsub’, ‘tolower’, ‘str_split’) and the individual words were matched with a **word repository** to obtain a sentiment score [+1 if positive, 0 if neutral and -1 if negative]. For example—

```
> score = as.integer(unlist(score.sentiment("Web scraping
is exciting and fun", pos, neg)))

> score
[1] 2

> score = as.integer(unlist(score.sentiment("Web scraping
is okay", pos, neg)))

> score
[1] 0
> score = as.integer(unlist(score.sentiment("Web scraping
can be challenging at times", pos, neg)))

> score
[1] -1
```

2 Methodology

c) The word repository contains a list of 2006 positive words and 4784 negative words[2,3] and can be downloaded from—
https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

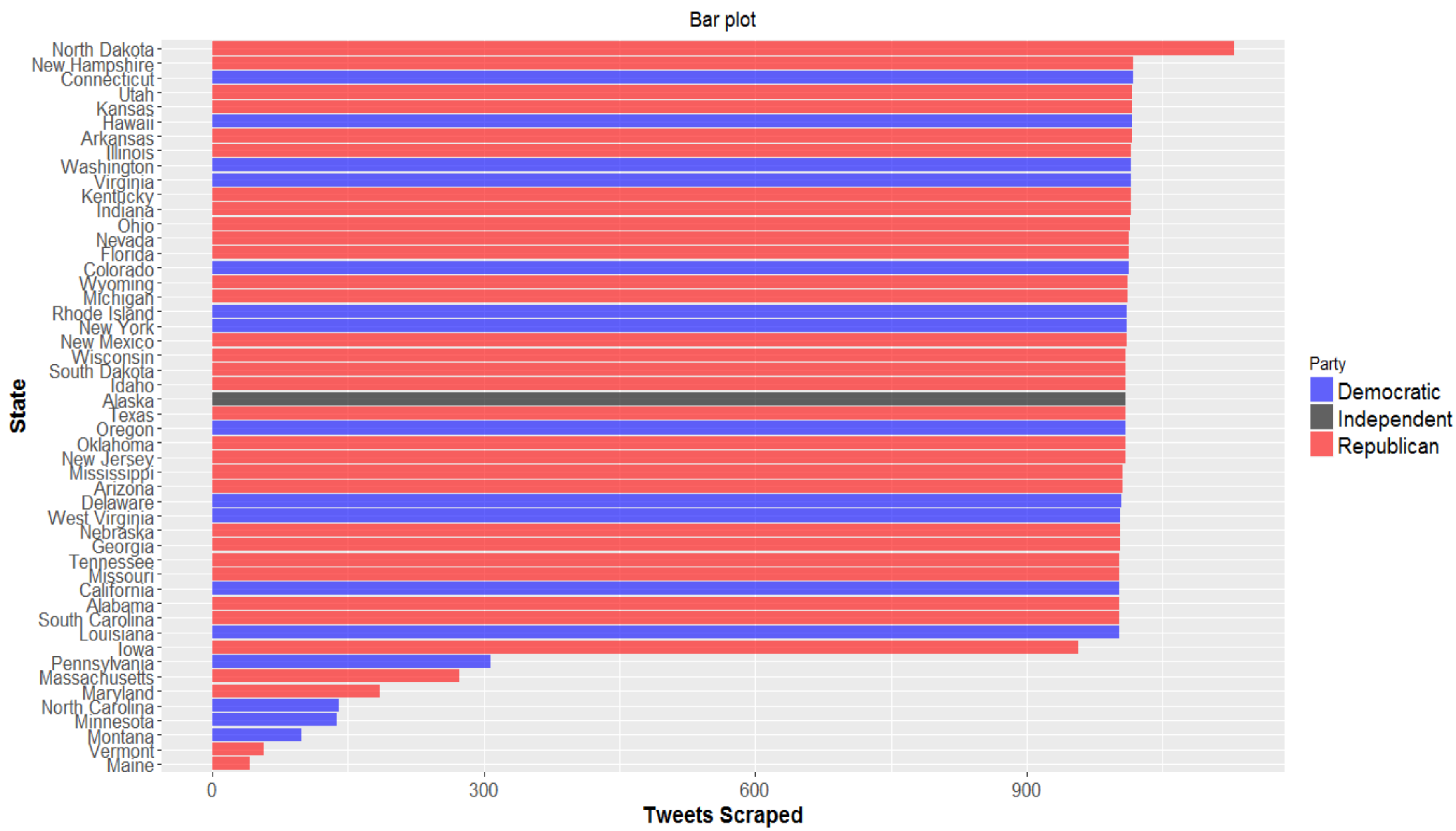
---

[2] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[3] Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

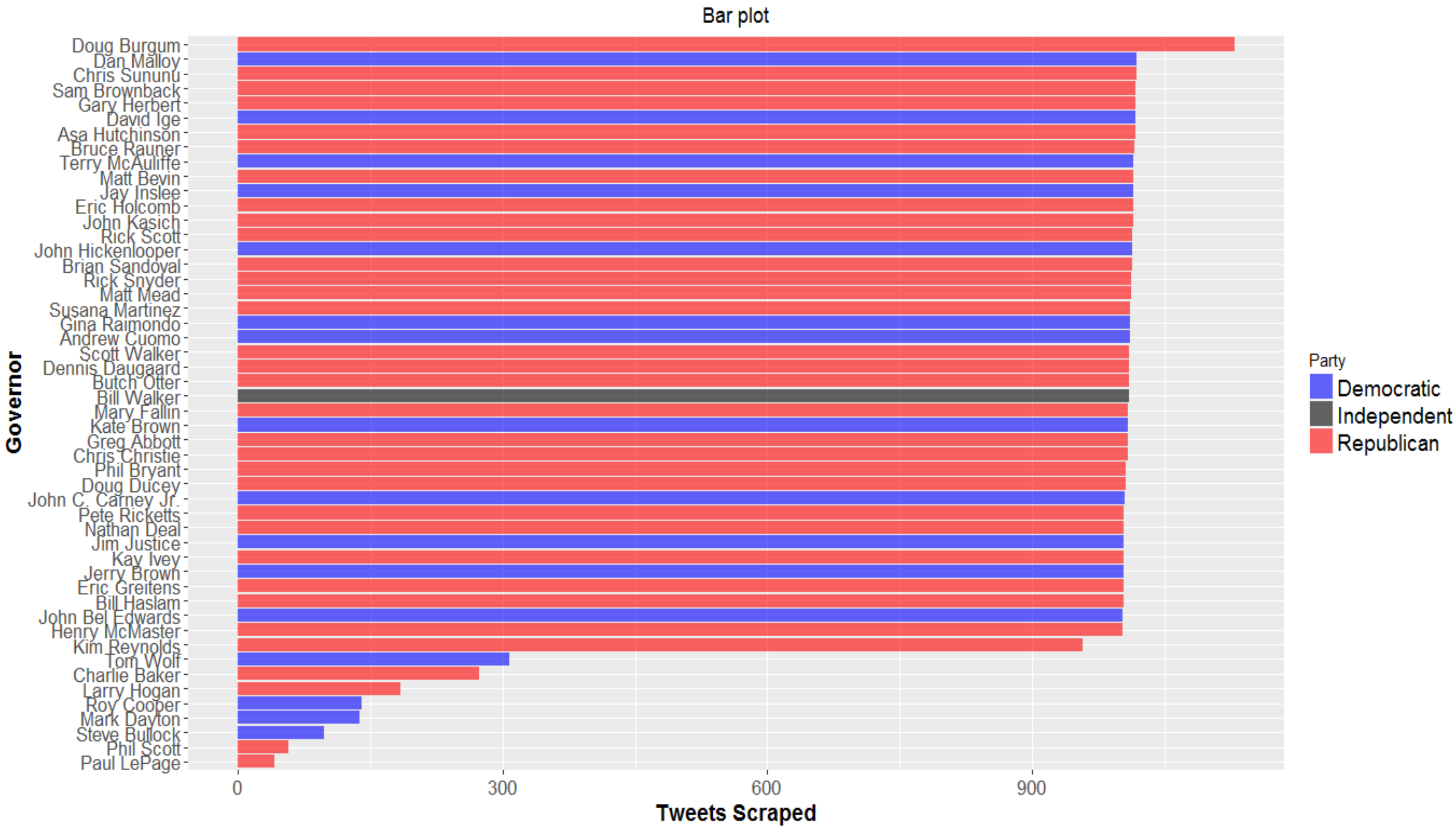
3 Data Visualization & Analysis

- Tweets scraped listed by **state** (43,737 tweets in total)



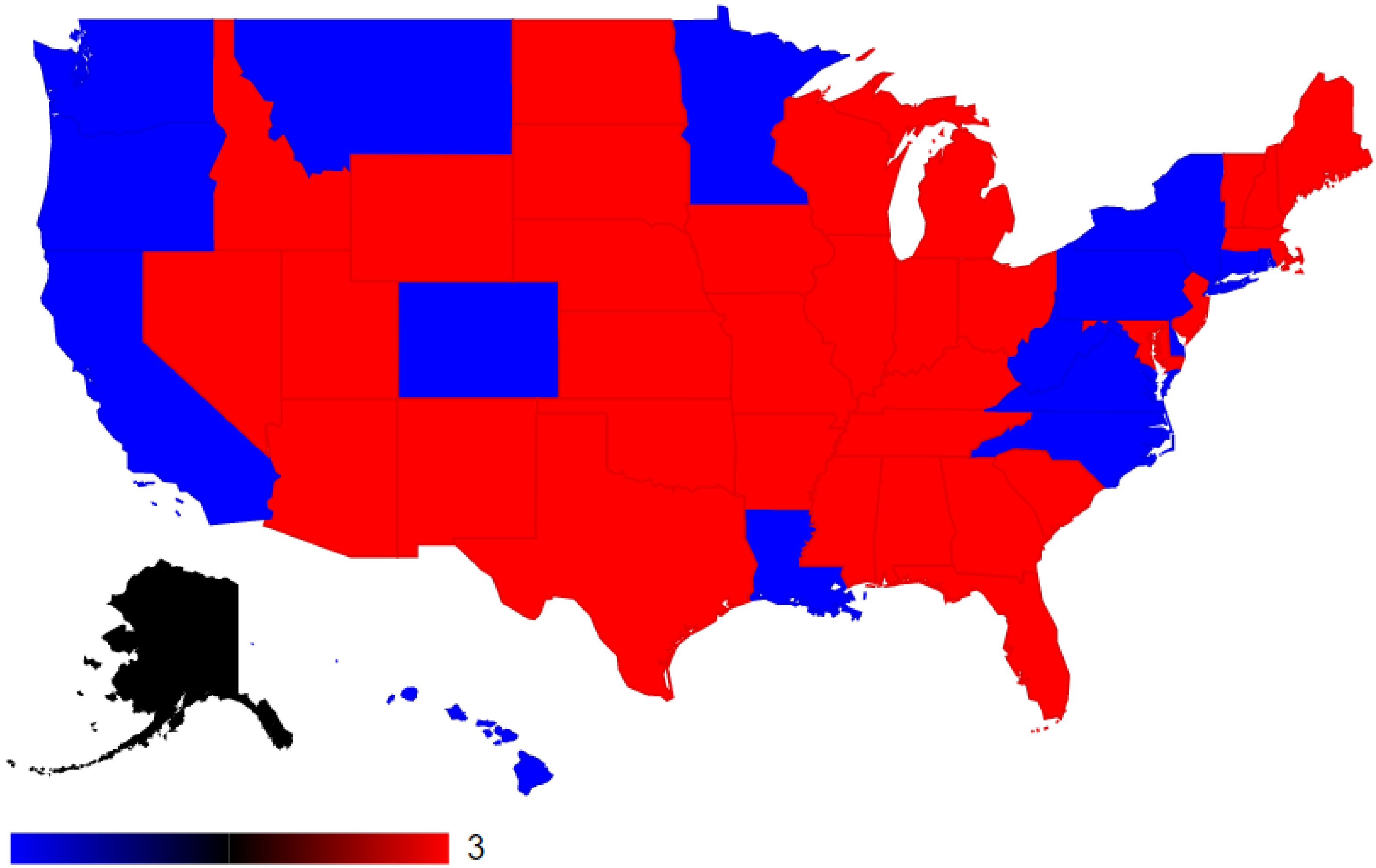
3 Data Visualization & Analysis

- Tweets scraped by name of **governor** (43,737 tweets in total)



3 Data Visualization & Analysis

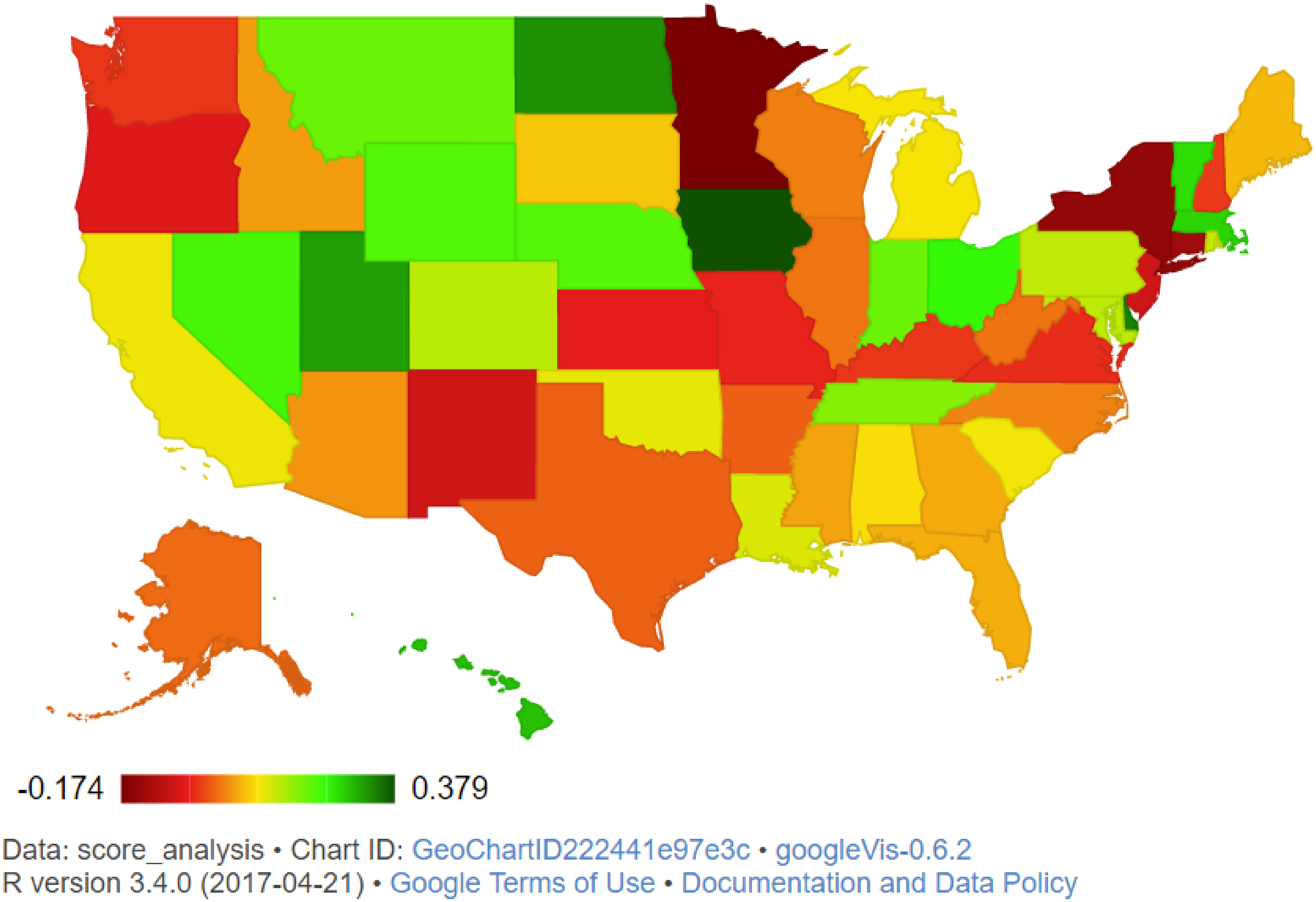
- **Party** affiliation of current **governors** [Republican, Democrat, Independent]



Data: score_analysis • Chart ID: GeoChartID22243a82102 • googleVis-0.6.2
R version 3.4.0 (2017-04-21) • Google Terms of Use • Documentation and Data Policy

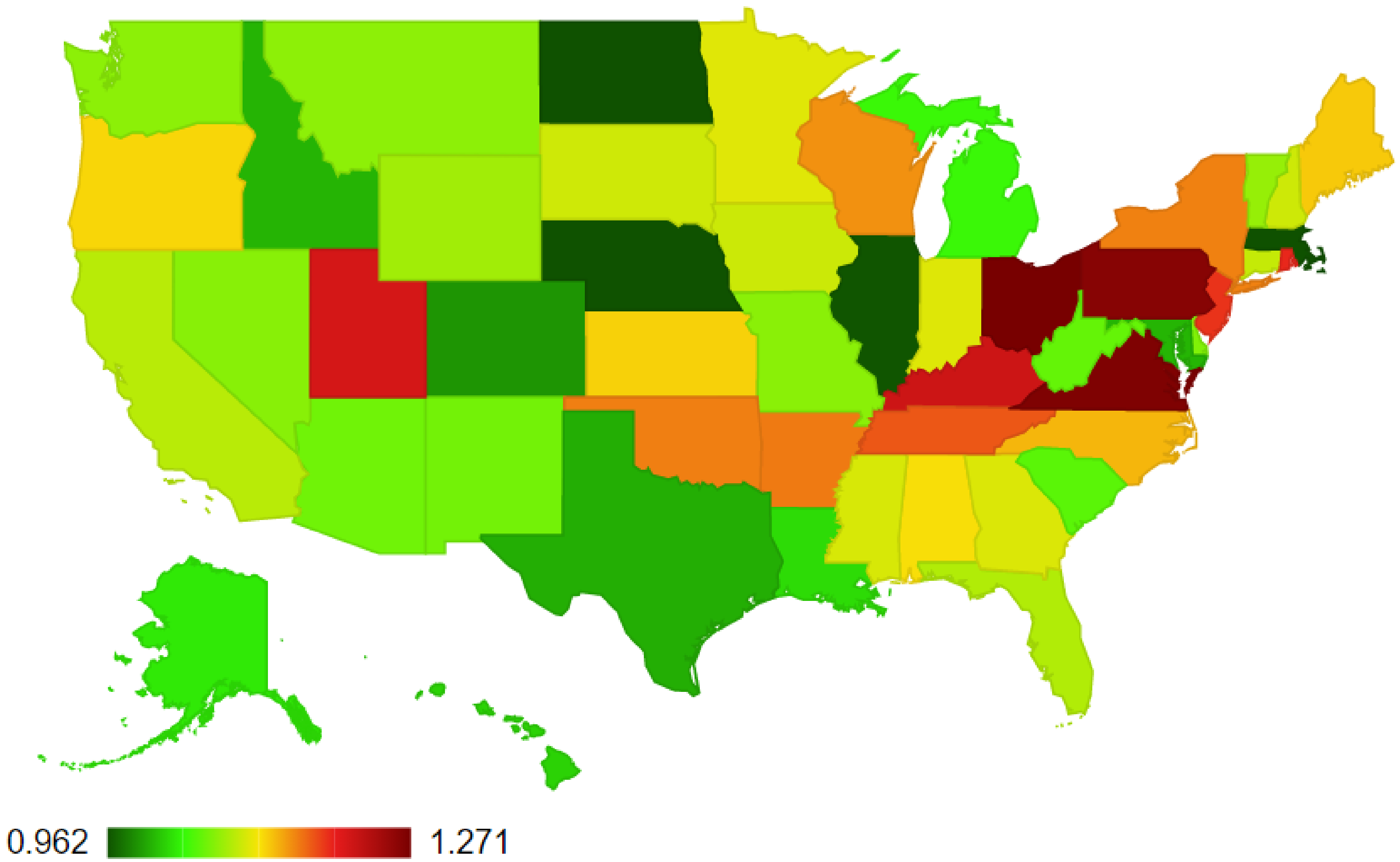
3 Data Visualization & Analysis

- **Mean** sentiment score of **governors**



3 Data Visualization & Analysis

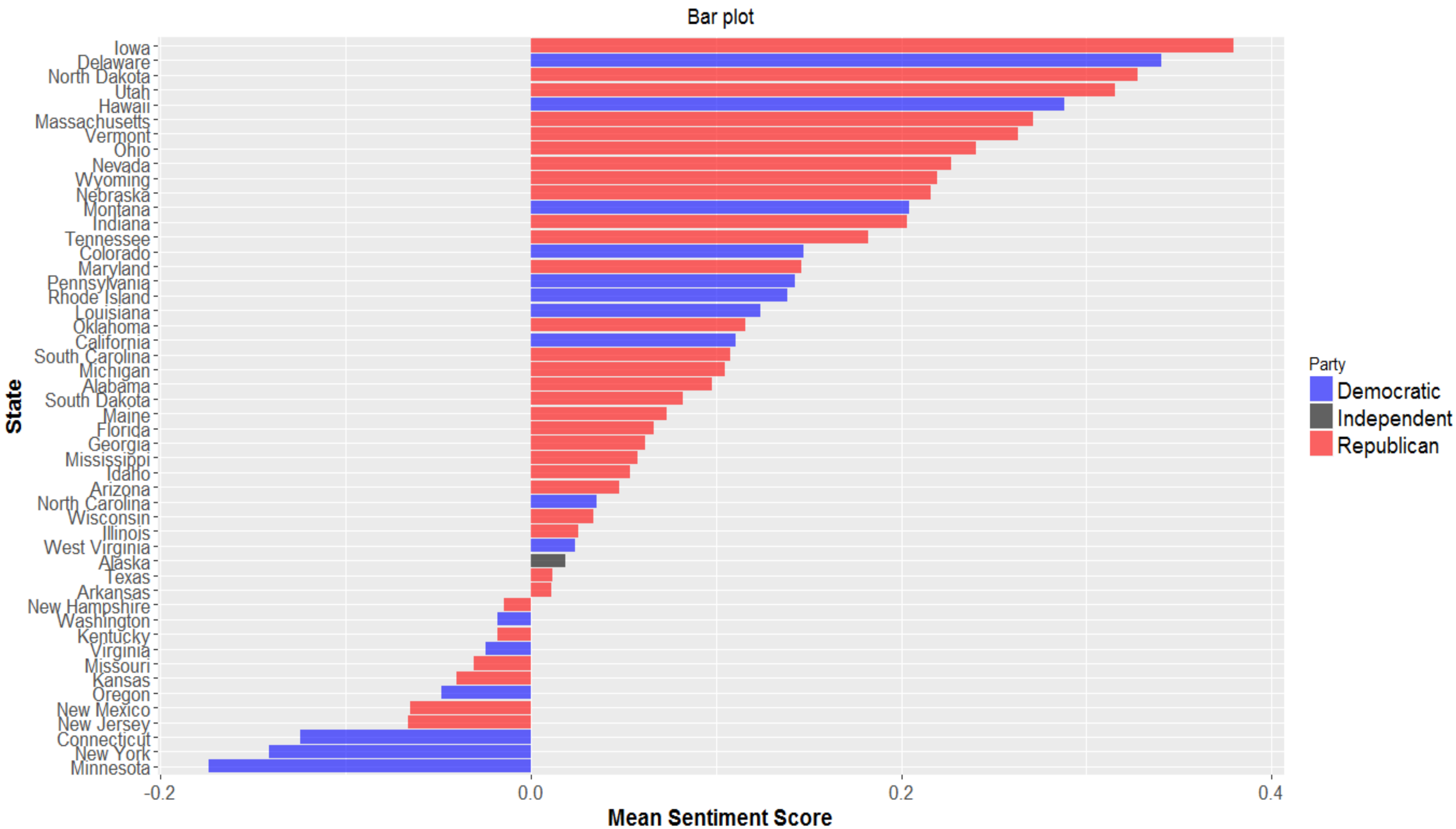
- **Standard deviation** on sentiment score of **governors**



Data: score_analysis • Chart ID: GeoChartID222438cf7667 • googleVis-0.6.2
R version 3.4.0 (2017-04-21) • Google Terms of Use • Documentation and Data Policy

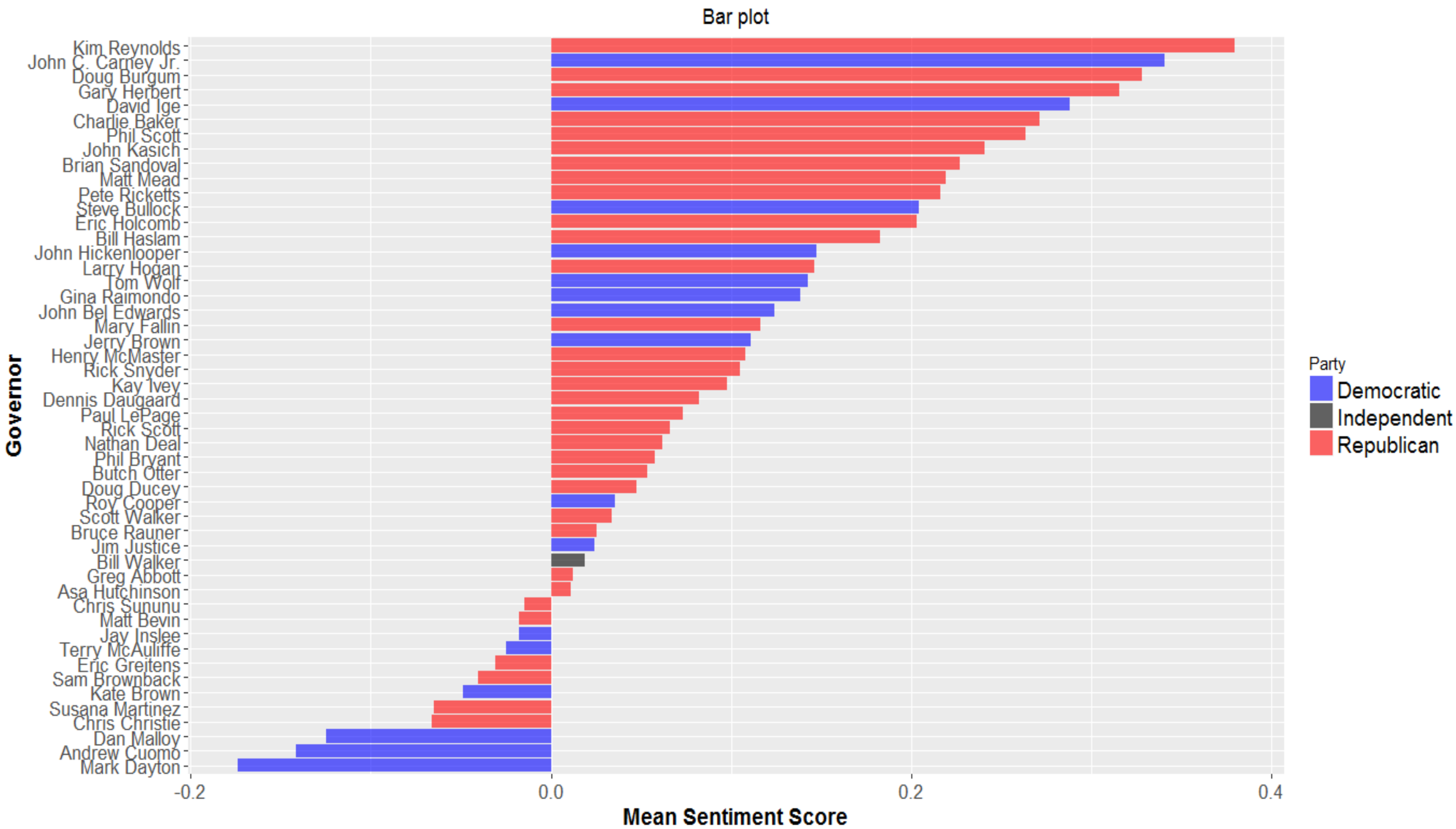
3 Data Visualization & Analysis

- **Bar plot** with mean sentiment score for each **state**



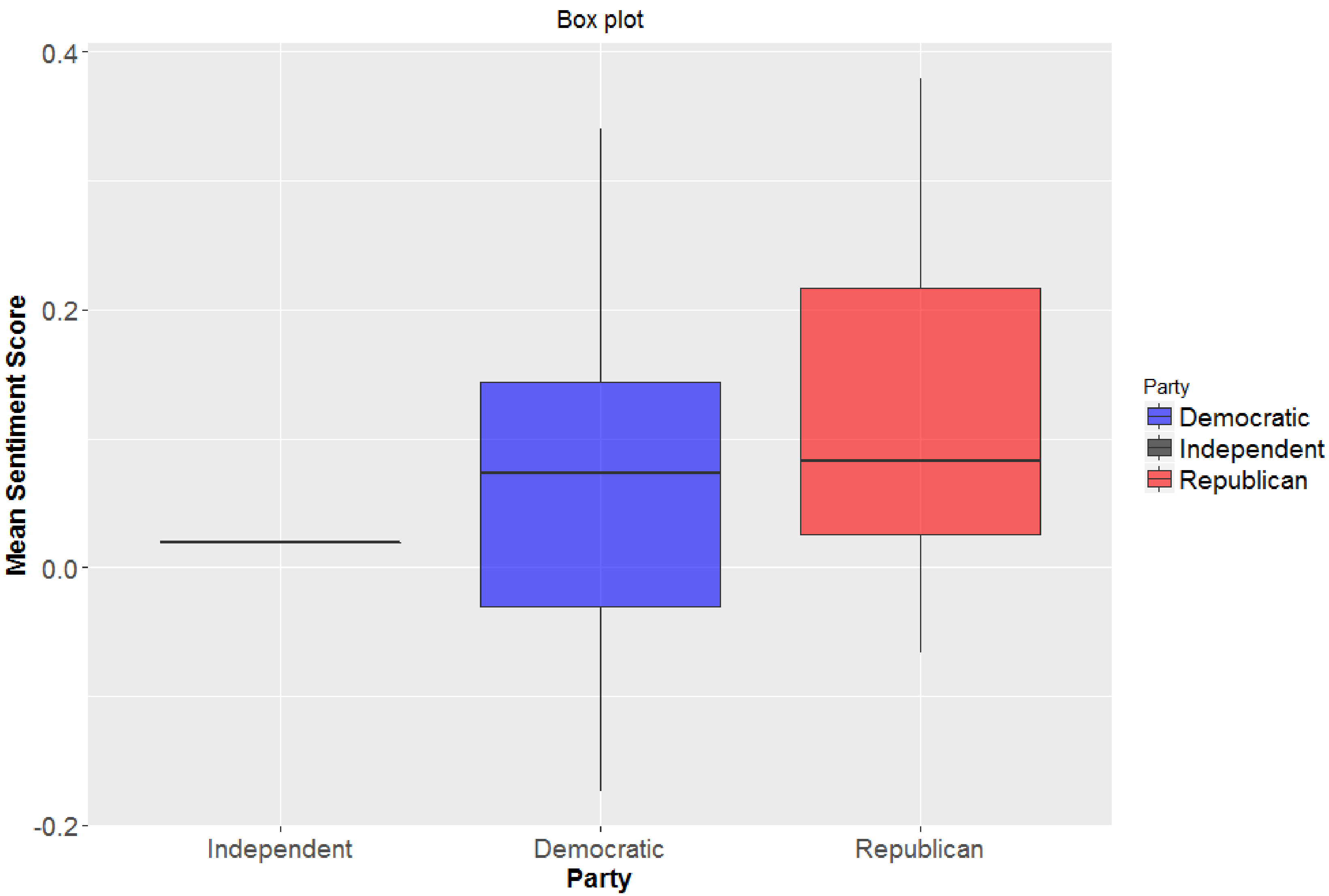
3 Data Visualization & Analysis

- **Bar plot** with mean sentiment score for each **governor**



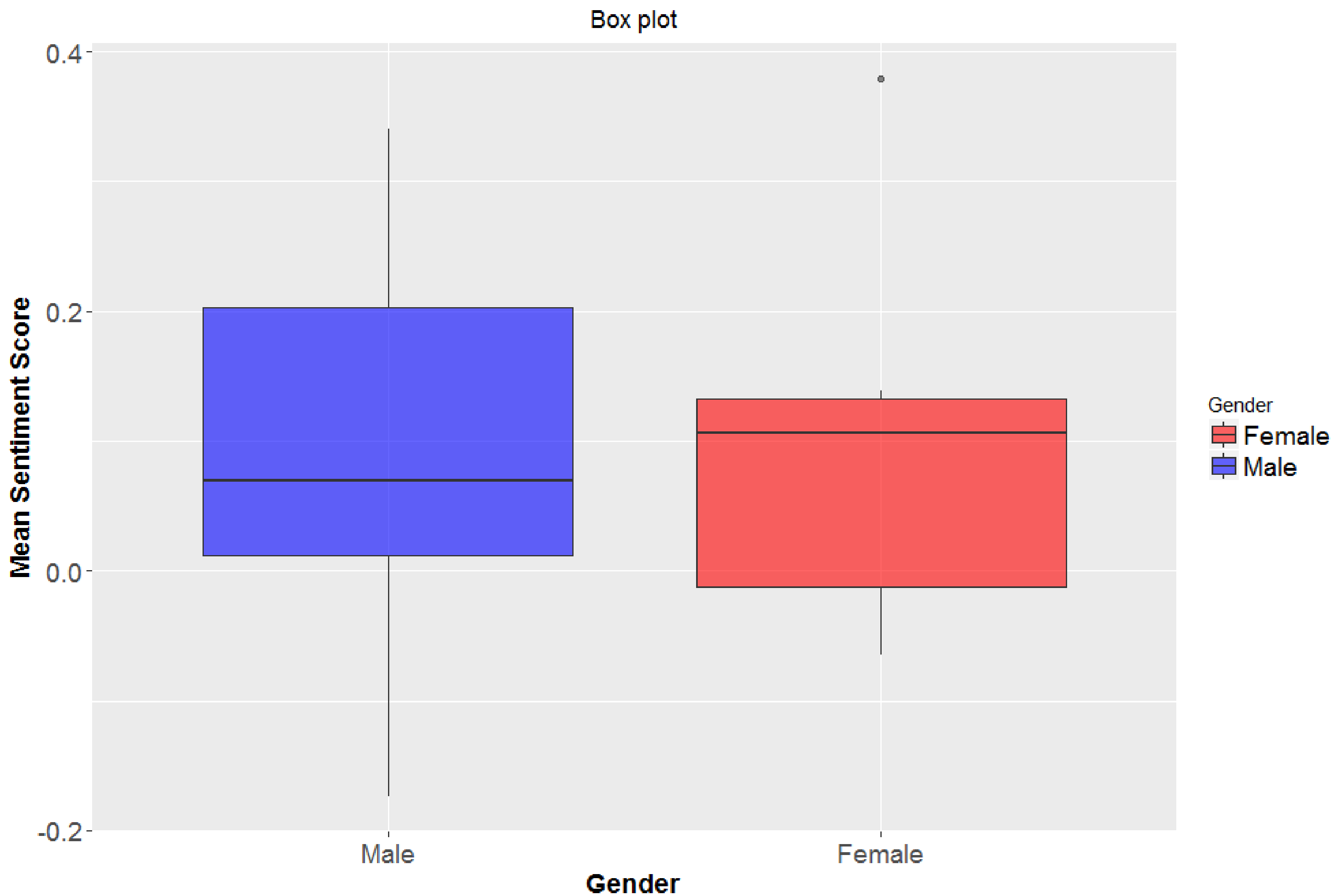
3 Data Visualization & Analysis

- **Box plot** with **mean** sentiment score and **party** affiliation



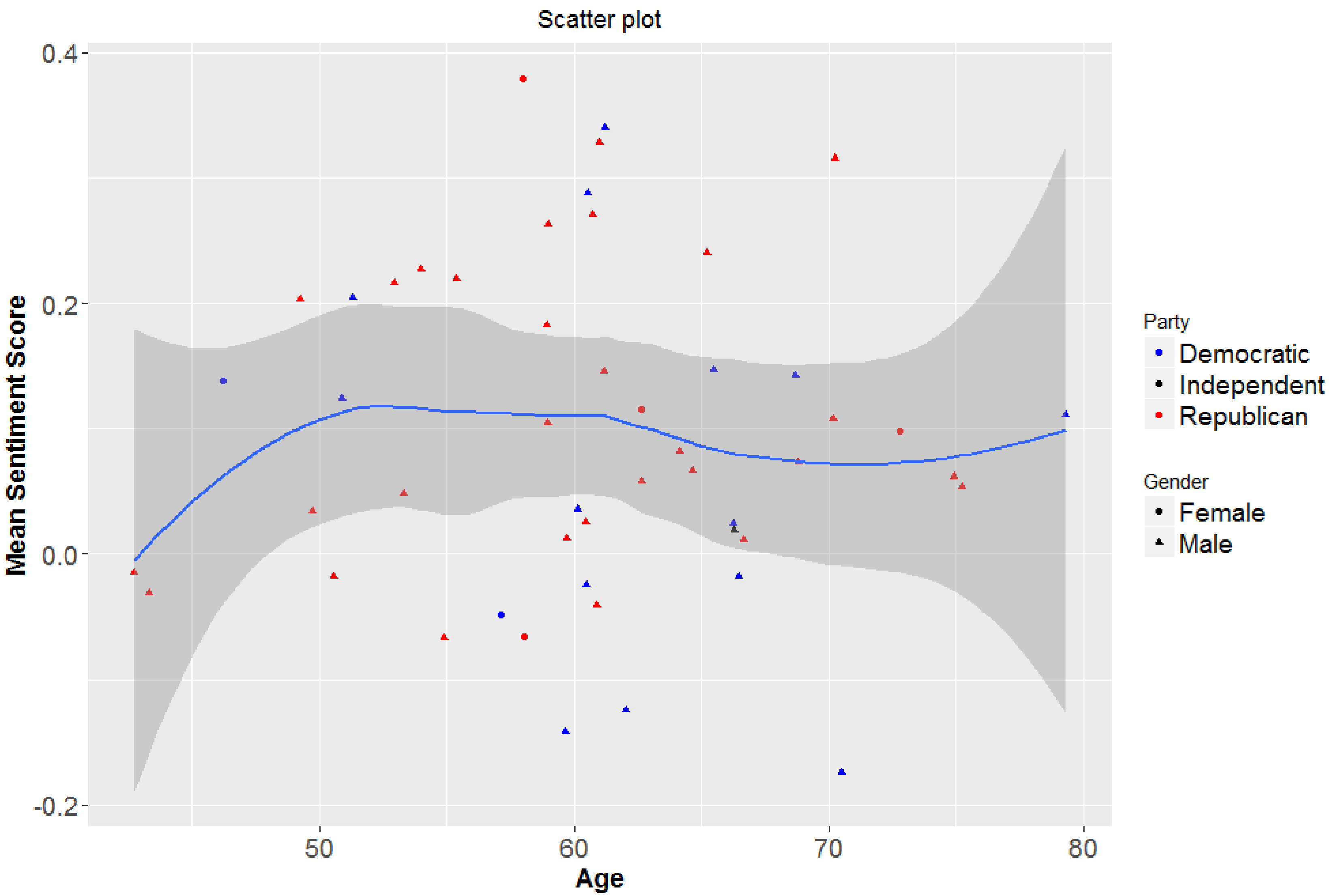
3 Data Visualization & Analysis

- **Box plot** with **mean** sentiment score and **gender** of governor



3 Data Visualization & Analysis

- **Scatter plot** with **mean** sentiment score and **age** of governor



- **ANOVA Test**

$H_0: \mu_1 = \mu_2$

Analysis of Variance Table

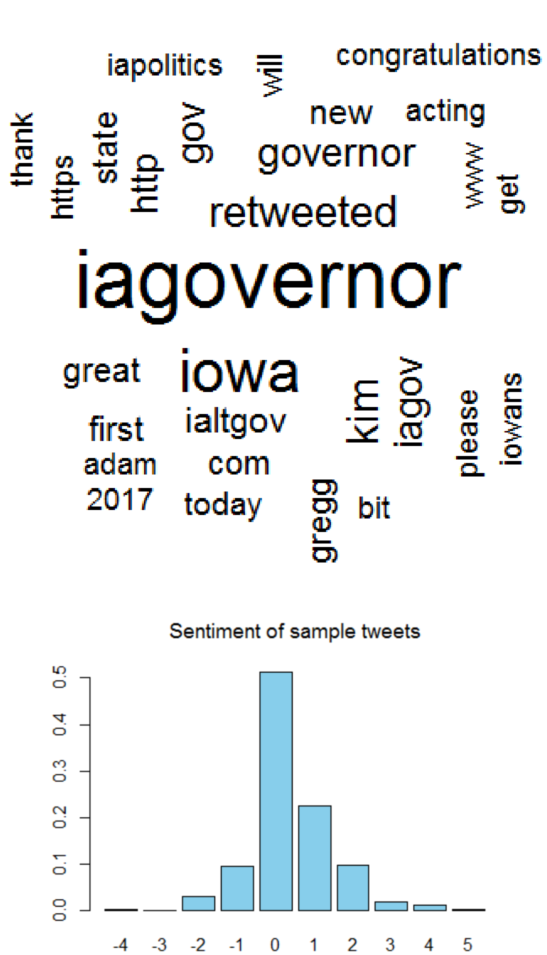
Response: Mean_Sentiment_Score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	0.00043	0.0004314	0.0243	0.8769
Party	2	0.03095	0.0154772	0.8708	0.4255
Age	1	0.00017	0.0001658	0.0093	0.9235
Residuals	45	0.79979	0.0177731		

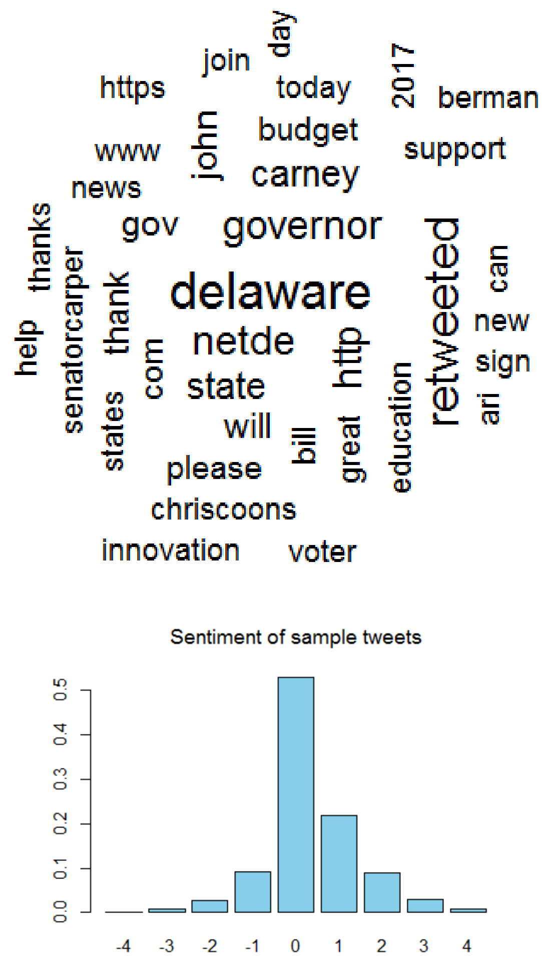
Fail to reject H_0

3 Data Visualization & Analysis

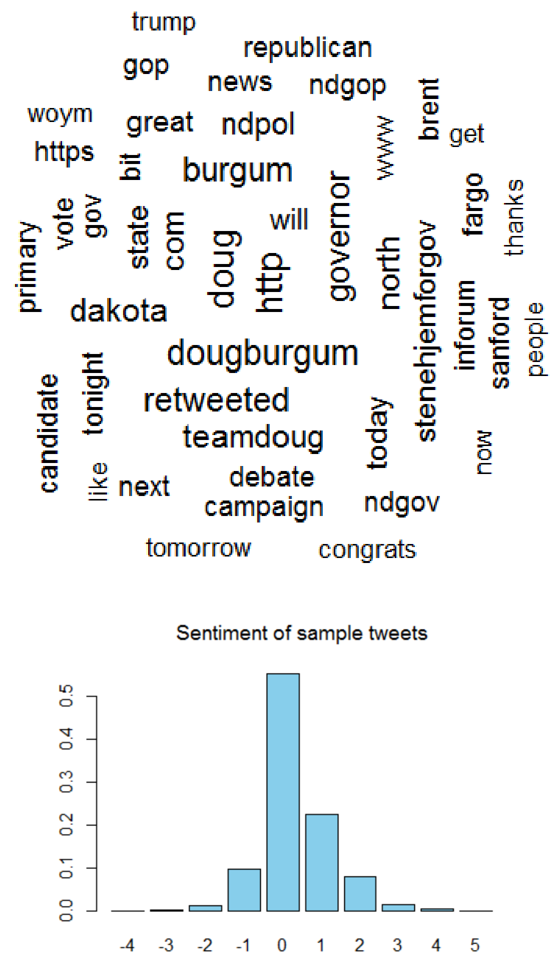
- **Word cloud** of **top** three governors (min. frequency= 25)



IOWA



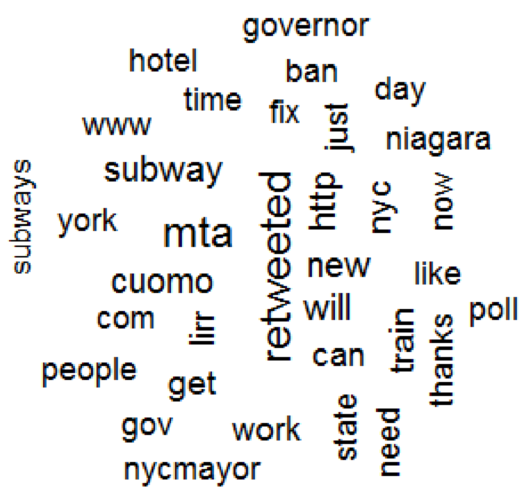
DELAWARE

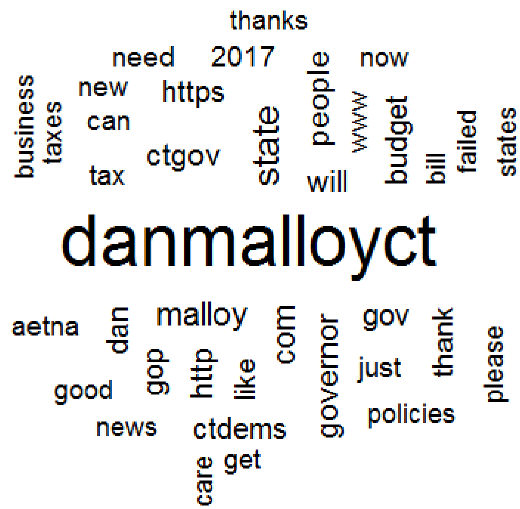


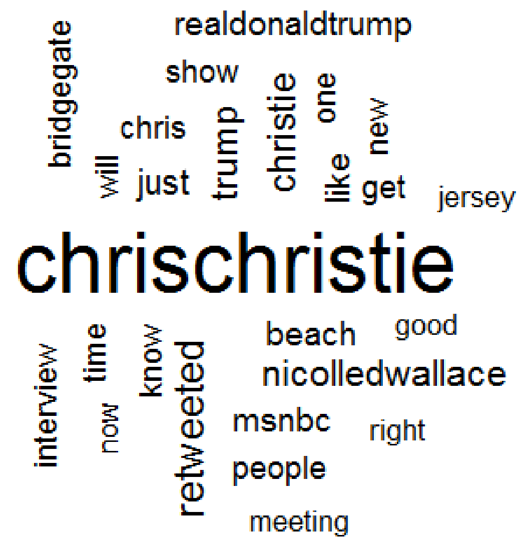
NORTH DAKOTA

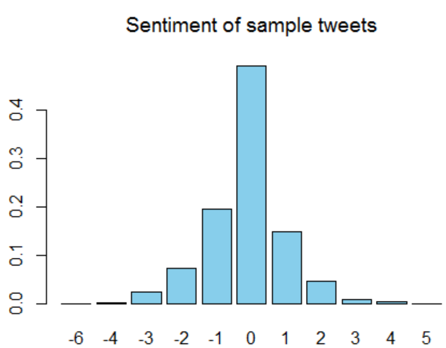
3 Data Visualization & Analysis

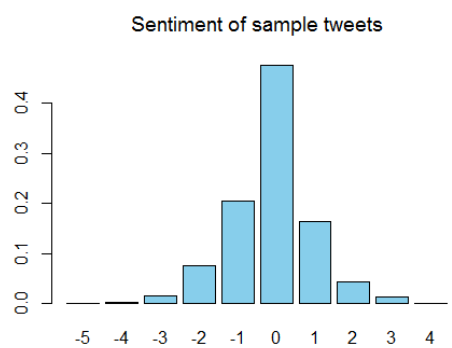
- **Word cloud** of **bottom** three governors (min. frequency= 25)

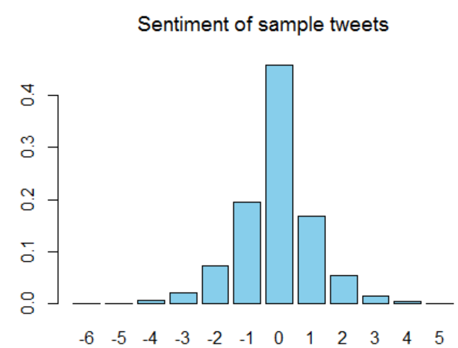












NEW YORK **CONNECTICUT** **NEW JERSEY**

3 Data Visualization & Analysis

- **Key word search—**

Keyword	Count	State
health	40	OH
	38	KY
	32	RI
tax	98	KS
	68	ID
	38	AK
trump	82	NJ
	53	OH
	40	SC
education	43	OK
	31	NM
	31	DE
jobs	37	AL
	37	MI
	36	WV

3 Data Visualization & Analysis

- **Rankings** based on **mean sentiment score** and **comparison** with **Morning Consult**

State	Governor	Party	Tweets Scraped	Twitter Rank (from top)	Morning Consult (from top)
Iowa	Kim Reynolds	Republican	957	1	-
Delaware	John Carney Jr.	Democratic	1005	2	28
North Dakota	Doug Burgum	Republican	1129	3	4
Utah	Gary Herbert	Republican	1017	4	9
Hawaii	David Ige	Democratic	1017	5	36
Massachusetts	Charlie Baker	Republican	273	6	1
Vermont	Phil Scott	Republican	57	7	8
Ohio	John Kasich	Republican	1014	8	20
Nevada	Brian Sandoval	Republican	1013	9	7
Wyoming	Matt Mead	Republican	1012	10	3

State	Governor	Party	Tweets scraped	Twitter Rank (from bottom)	Morning Consult (from bottom)
Minnesota	Mark Dayton	Democratic	138	1	17
New York	Andrew Cuomo	Democratic	1011	2	23
Connecticut	Dan Malloy	Democratic	1018	3	3
New Jersey	Chris Christie	Republican	1009	4	1
New Mexico	Susana Martinez	Republican	1011	5	11
Oregon	Kate Brown	Democratic	1009	6	21
Kansas	Sam Brownback	Republican	1017	7	2
Missouri	Eric Greitens	Republican	1003	8	29
Virginia	Terry McAuliffe	Democratic	1015	9	23
Kentucky	Matt Bevin	Republican	1015	10	13

5 Conclusions and Future Outlook

- **Tweets** can be used as a **quick way to gauge** the approval of governors. It only takes few hours of coding and data processing when compared to conventional polling techniques that require significant $$ and resources.

- The **analysis** can be **improved** by taking **inputs** from **multiple social media** platforms (facebook ,…).

- Sentiment analysis can be **improved** by constantly **updating the list** of positive and negative words as the society evolves with time.

- There is always a **segment of population** who **do not use social media** platforms to express emotions. This will have an impact on the data analysis.

- It would be interesting to study how predictions change as a function of the **sample size** of tweets.

- This study can be extended to gauge the sentiment towards **President and Presidential Candidates** as well.

- With the **right imagination**, sky is the limit!

THANK YOU