How Your Health May Be Related to What Your Governor Says

Exploring relationships between regional indicators and social determinants of health and the emotional sentiments of elected leaders

IST 707 - Data Analytics

John Christman
Christina DaSilva
Christopher Hart
Jorge Martinez

THE iSCHOOL
Syracuse University

Introduction

The power of the spoken word is leveraged in nearly every motivational approach toward improving your life and achieving your goals. Speech is equally critical to success in politics for informing and motivating your constituents. Does this common thread result in a linkage between your health and the message of our State Governors?

Modern approaches to health care focus on a holistic view identifying multiple factors that contribute to a person's physical and mental wellness. An individual's physical health impacts their mental well-being and vice versa[1]. External factors, commonly referred to as social determinants, also influence overall health including a basic subsistence income, proper housing, and access to quality health care that supports checkups, preventative care, and when necessary, appropriate medical attention.

At the beginning of each year, elected Governors across the United States deliver their "state of the state" — an opportunity to lay out policy priorities and share sentiments related to the health and welfare of constituents. Of particular interest are the spoken words and whether they reflect, or are contrary to, the state of health and well-being in the region.

There are multiple indicators that can provide insight into the health and well-being of a geographic populace. And researchers across the country spend countless hours tracking trends and attempting to determine localized needs that require attention for adequacy of care, support, and population welfare. These indicators can often be categorized so as to provide a generalized view of any region, at any given time. Commonly, measures of physical and mental health are coupled with select social determinants of health, such as health insurance coverage and employment status, to signal low or high degrees of well-being in a regional setting.

This paper examines the relationship between these two occurrences — a speech and a state of health and well-being. The general inquiry is to consider if the emotions and sentiments selected and expressed by elected leaders may persuade or hinder the general health and well-being of those they were elected to serve.

[1] Retrieved from: https://psycnet.apa.org/record/1994-35904-001

Analysis and Models

About the Data

Data Set #1: The State Of The State Of The States, What America's Governors Are Talking About

The first dataset retrieved for this inquiry originates from the GitHub account of the popular political media and trends website, FiveThirtyEight (June 2019) .[2] The data includes a corpus of text files of all 50 Governors' 2019 state of the state speeches ('statespeechesCorpus') along with an index ('state-speeches-index.csv') that provides a listing of each of the 50 speeches, one for each state, as well as the name and party of the state's governor and a link to an official source for the speech. If an official government source could not be found, FiveThirtyEight provides a URL link to a news media source that has a transcript of the speech. Finally, a reference analysis file ('state-speeches-words.csv') is made available and contains every one-word phrase that is mentioned in at least 10 speeches and every two or three-word phrase that is mentioned in at least five speeches after a list of stop-words is removed. The reference file also contains the results of a chi^2 test that shows the statistical significance of associated p-value of phrases.

To prepare for initial sentiment analysis and create working datasets for learning algorithms to be applied, a Document Term Matrix function is utilized to handle the loading of the corpus of speech text files in order to transform the resulting 'SimpleCorpus' format into a matrix data structure for cleaning and normalization. During this process, two variables are created to remove common English 'stop words' (as defined within an available library in R) as well as the names of states if they appeared in speech text. In addition, the Document Term Matrix function is used to generate counts (as numeric variables) of words in each speech observation that are between 3-10 letters in length ('speechDTM'). The occurrence of words within each speech observation can be applied for sentiment analysis that aims to identify specific emotions and sentiments that were expressed in the Governor speeches.

Preparing the resulting matrix format for selected modeling techniques, inclusive of associated learning algorithms, requires a series of data transformations. First, the speechDTM SimpleCorpus is

[2] The State Of The State Of The States, What America's Governors Are Talking About. https://fivethirtyeight.com/features/what-americas-governors-are-talking-about (June 2019)

transformed into a matrix data structure containing all speech (observations) and feature attributes (‘speechMAT’) in order to handle the numeric units that are generated to indicate the count of word occurrences in each speech observation. With quantitative variables, certain calculations can be highly influenced by applied units of measure. Datasets that include “mixed” units of measurements can return biased or skewed results if not standardized or normalized. Commonly, data is standardized to a calculated means or standard deviation, or is normalized to the same measurement scale, in order to create comparative variables[3].

For the state Governors’ speeches matrix data, the quantitative numeric values reflected the count occurrence of a variable word within an essay .txt file (observation). To normalize this dataset for initial analysis and then eventually for training and testing data to be applied through unsupervised and classification learning methods, the quantitative count attributes are normalized to a value of 0-1. Certain learning algorithms (such as SVM and kNN) are more sensitive to the scale of data than others since the distance between the data points is very important. In order to avoid this problem, the dataset is transformed to a common scale (between 0 and 1) while keeping the distributions of variables the same. This is often referred to as *min-max scaling*—converting each data point to a normalized data point. This is handled with the following code:

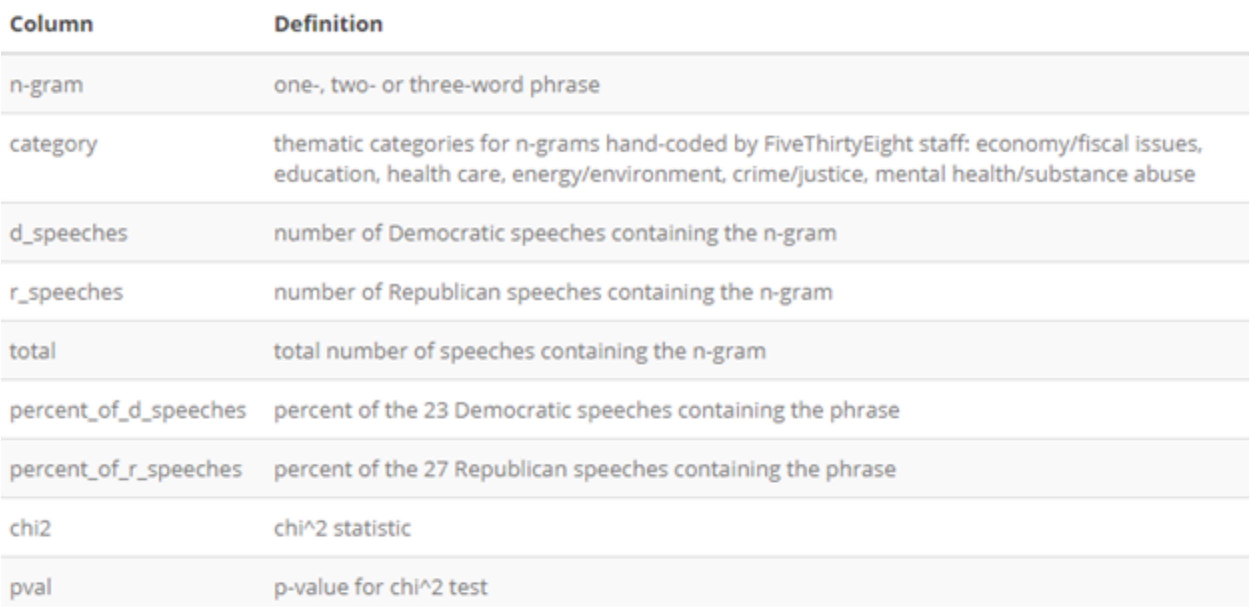
```
speechMAT_Norm <- apply(speechMAT, 1, function(x) round(x/sum(x),3))
```

To assist in the exploration of this data, the additional reference files (‘state_speeches_index.csv’ and ‘state_speeches_words.csv’) are loaded and used to expand the base normalized data frame of speech observations by binding new “class” nominal factor attributes with information specific to state name, name of the Governor, and political party affiliation for each observation (speech) row (‘speechDF_Reporting’). These classes, or labels, are applied when conducting sentiment analysis and when training supervised learning models for predicting a certain class value or outcome.

In addition, the reference files are used to create a data dictionary dataframe, specific to the FiveThirtyEight words file, (Table 1) to be used during analysis activities.

[3] Practical Guide to Clustering Algorithms & Evaluation in R: https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/clustering-algorithms-evaluation-r/tutorial

Table 1: Data Dictionary for state_speeches_words.csv’



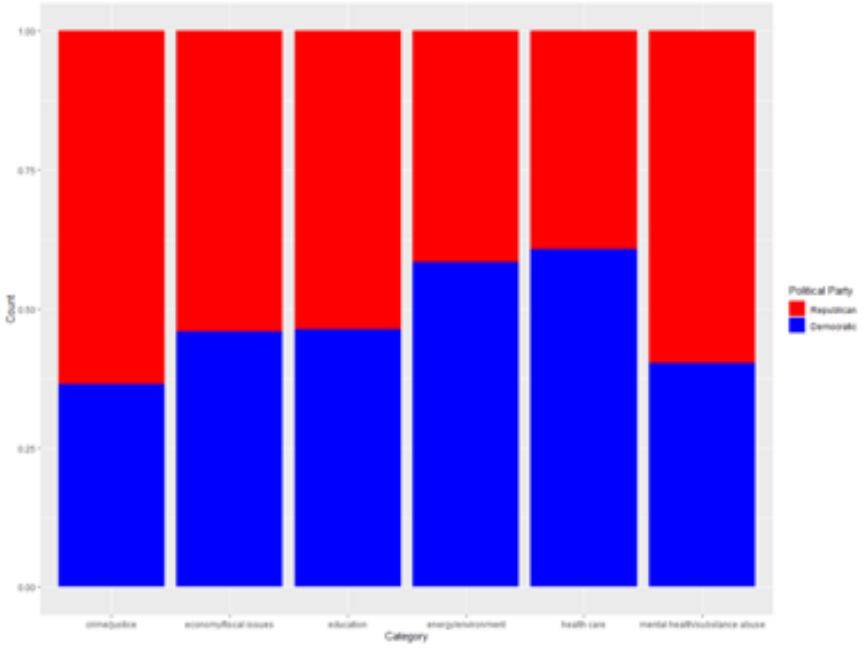
Column	Definition
n-gram	one-, two- or three-word phrase
category	thematic categories for n-grams hand-coded by FiveThirtyEight staff: economy/fiscal issues, education, health care, energy/environment, crime/justice, mental health/substance abuse
d_speeches	number of Democratic speeches containing the n-gram
r_speeches	number of Republican speeches containing the n-gram
total	total number of speeches containing the n-gram
percent_of_d_speeches	percent of the 23 Democratic speeches containing the phrase
percent_of_r_speeches	percent of the 27 Republican speeches containing the phrase
chi2	chi^2 statistic
pval	p-value for chi^2 test

The transformed corpus datasets for Governor speeches are labeled:

- **speechDF_Reporting** - Normalized DATAFRAME with all observations + Nominal factors: state, governor, party, and .txt filename
- **speechMAT_Norm** – Data structure MATRIX with all .txt speeches with normalized % word occurrence freq.
- **speechDF_Norm** - Normalized DATAFRAME with all observations + State label (nominal factor) and Speech txt title (character)
- **indexDF** - Dataframe of all 50 speeches with attributes from Data Dictionary
- **wordsDF** - Dataframe of speech phrases with chi^2 test that shows the statistical significance of and associated p-value of identified phrases (from FiveThirtyFive analysis)

Finally, to assist in providing context for working with the Governor speeches dataset, the thematic categories that were initially identified by the FiveThirtyEight research team are plotted for further reference during project inquiries. The categories, identified and plotted according to political party affiliation, are presented in Figure 1. Of importance is the reminder that, beyond the words, emotions, and sentiments that elected leaders may express at any given time, they belong to specific parties that represent certain ideologies. As illustrated, each party appears to place emphasis on different issues that can directly impact the health and well-being of people living in state communities.

Figure 1: Thematic categories by Governor political party



Data Set #2 State Health Indicators

The second dataset retrieved for this inquiry originates from the County Health Rankings & Roadmaps program, a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute[4]. The program tracks health measures at the county level, with the goal of providing a foundation to improve health within local communities. Measures include metrics regarding health outcomes, health behaviors, clinical care, social and economic environment, as well as physical environment. An excel file with all indicators tracked was downloaded from the County Health Rankings website. Included in the original file for each record were the following attributes (Table 2), which were associated with each of the four health indicators that were chosen for exploration.

[4] Data retrieved from: https://www.countyhealthrankings.org/about-us (June 2019)

Table 2: Glossary for Attributes Applicable to All State Health Indicators

Variable	Data Type	Description
FIPS	Integer	Federal Information Processing Standard county code
state	Factor	Full US state name
county	Factor	Full county name within a US state

The four indicators chosen for analysis include poor physical health days, poor mental health days, uninsured, unemployment.

Poor Physical Health Days, 2019 County Health Rankings

Data Source: Behavioral Risk Factor Surveillance System (2016)

Description of the Data: Average number of physically unhealthy days reported in the past 30 days (age-adjusted).

Measuring health-related quality of life (HRQoL) helps characterize the burden of disabilities and chronic diseases in a population. In addition to measuring how long people live, it is also important to include measures of how healthy people are while alive – and people's reports of days when their physical health was not good are a reliable estimate of their recent health.

Table 3: Glossary for Poor Physical Health Days Attributes (Denoted as x1)

Variable	Data Type	Description
x1Days	Numeric	Physically Unhealthy Days. Average number of physically unhealthy days reported in past 30 days (age-adjusted)
x1CIL	Numeric	95% Confidence Interval - Low
x1CIH	Numeric	95% Confidence Interval - High
x1Z	Numeric	County Z-scoring for standardization

Poor Mental Health Days, 2019 County Health Rankings

Data Source: Behavioral Risk Factor Surveillance System (2016)

Description of the Data: Average number of mentally unhealthy days reported in the past 30 days (age-adjusted).

Self-reported health status is a widely used measure of people’s health-related quality of life. In addition to measuring how long people live, it is important to also include measures that consider how healthy people are while alive. Further, reports of days when mental health was not good is a reliable estimate of recent health.

Table 4: Glossary for Poor Mental Health Days Attributes (Denoted as x2)

Variable	**Data Type**	**Description**
x2Days	Numeric	Mentally Unhealthy Days. Average number of mentally unhealthy days reported in past 30 days (age-adjusted)
x2CIL	Numeric	95% Confidence Interval - Low
x2CIH	Numeric	95% Confidence Interval - High
x2Z	Numeric	County Z-scoring for standardization

Uninsured, 2019 County Health Rankings

Data Source: Small Area Health Insurance Estimates (2016)

Description of the Data: Percentage of population under age 65 without health insurance.

Lack of health insurance coverage is a significant barrier to accessing needed health care and to maintaining financial security.

Table 5: Glossary for Uninsured Attributes (Denoted as x3)

Variable	Data Type	Description
x3Num	Integer	Number uninsured
x3Per	Integer	Percentage of population under age 65 without health insurance
x3CIL	Integer	95% Confidence Interval - Low
x3CIH	Integer	95% Confidence Interval - High
x3Z	Numeric	County Z-scoring for standardization

Unemployment, 2019 County Health Rankings

Data Source: Bureau of Labor Statistics (2017)

Description of the Data: Percentage of population ages 16 and older unemployed but seeking work.

The unemployed population experiences worse health and higher mortality rates than the employed population.

Table 6: Glossary for Unemployment Attributes (Denoted as x4)

Variable	Data Type	Description
x4Num	Integer	Number unemployed
x4Tot	Integer	Total labor force number
x4Per	Numeric	Percentage of population ages 16 and older unemployed but seeking work
x4Z	Numeric	County Z-scoring for standardization

Data Ingestion and Missing Values

Data cleansing and preparation for analysis begins with the original .CSV file including the health indicators identified above, as well as the FIPS identifier, state, and county data, which is loaded into R and stored into a data frame named cleanHealth. The data types of each attribute are reviewed as they were entered utilizing the str() function, and all appear to be assigned appropriately. Missing values are checked for on a global level by calling the is.na() function on the data set, and then summing the results, returning 314 missing values. In order to gain a more granular understanding of where these missing values lie, the following function is called to identify the total number of missing values per column:

```
sapply(cleanHealth, function(x) sum(is.na(x)))
```

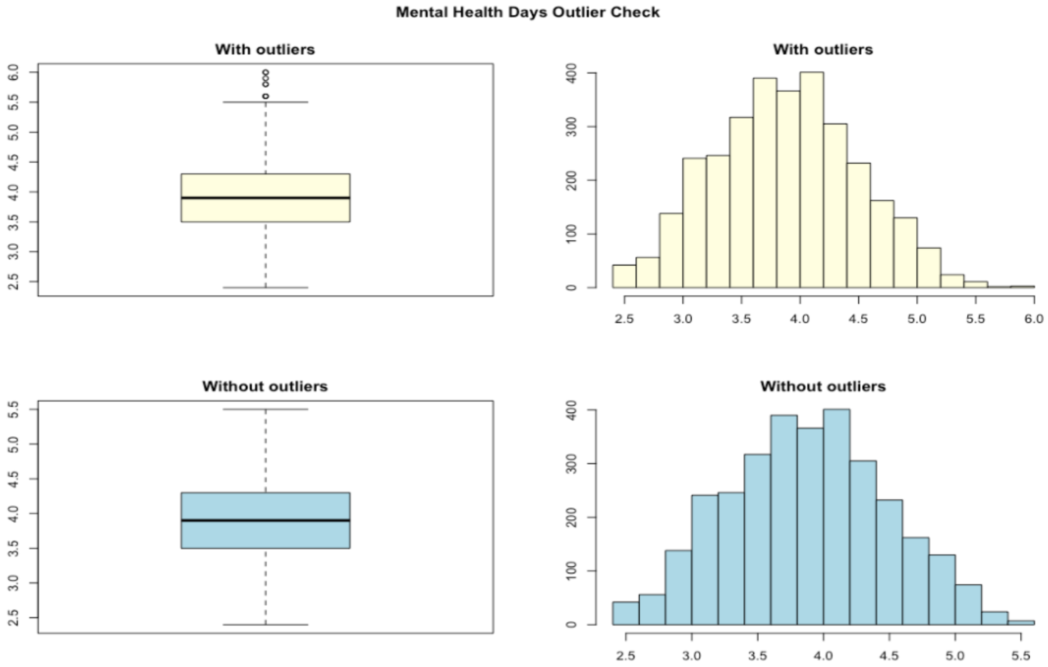
Individual rows in the cleanHealth data frame with missing values are reviewed by calling the rows in the dataframe where is.na() is true for the indicators specified previously.

A new data frame called *healthIndicators_DF* is created, including only columns with statistical data. Then, each row from the data set with a missing value is called for review. One row with missing values for uninsured and unemployed measures is found, as well as sixty-one rows with missing z-scores. A new data frame called 'healthIndicatorsWithStats_DF' is created to include all data by county, yet exclude the row with missing uninsured and unemployment measure data, as well as those with missing z-score data.

Outliers

A function is written to evaluate outliers present in a given variable within a dataset, calling these, as well as a description of the variable name as parameters. The function returns four plots: a boxplot and histogram of the variables including outliers, as well as both plots excluding outliers. The function also returns the number of outliers identified, proportion of values that are outliers, the mean of the outliers, mean of the attribute values if outliers are included, as well as removed. This function is called on each of the health indicator columns to review and identify any outliers for potential removal.

Figure 2: Outlier results for poor mental health days indicator variable



After review, it is decided to keep outliers within the data set as there is not any evidence of error in these values, and their removal would potentially skew the data.

Aggregation

A new dataframe, healthIndicators_byState, is created and populated with the data from the healthIndicators_DF data, aggregated by state. The following statement is used to facilitate aggregation with the ddply() function from the plyr package:

```
ddply(healthIndicators_DF,.(state), numcolwise(sum))
```

At this point, it is decided to exclude health indicator attributes with percentage values (x3Per, x4Per, X5Per), as these cannot be aggregated by state with the information available as described in the standardization explanation. Initially another health indicator was included for analysis (severe housing problems - X5), but ultimately was excluded since population count data is not available for normalizing the data set.

Relative-to-Population Standardization

A final data preparation and transformation step is required for the health indicators data to be applied during modeling. In general, the retrieved data set included health reporting data at the county level for all states. This reporting included average days for indicators X1 and X2, and then total population counts and percentages for indicators X3 and X4. In order to derive state-level reporting values for each attribute and standardize to comparable units for multivariate analysis, an additional data file is sourced from the U.S. Census Bureau for 2019 estimated population numbers in each state. This is then used to create standard numeric indicator measurement units for each attribute that are values "relative-to-population".

Of importance is the recognition that state-level values can not be derived by merely "averaging the averages", which is reported by county for indicators X3 and X4, without knowledge of each county-level population size. Therefore, a sum of reported county-level counts (for example, the total number of uninsured persons from all counties for a state) is calculated for X3 and X4 attributes in each of the 50 state observations and then divided by the reporting state population. This provides a standard indicator value that is proportionally relative to the state population and, thus, is comparative across the data set.

Health indicators X1 and X2 report the average number of poor physical and mental health days in the past 30-days. To standardize these values, the sum of total days reported by counties within each state and then divided by 30.42 which is the average number of days in a month. The standard comparable unit is then described as the number of days in any given month that a person within the state is having a poor physical or mental health experience.

A new data frame is created, 'healthIndicators_byState_Relative', to record the relative-to-population features for each reporting state.

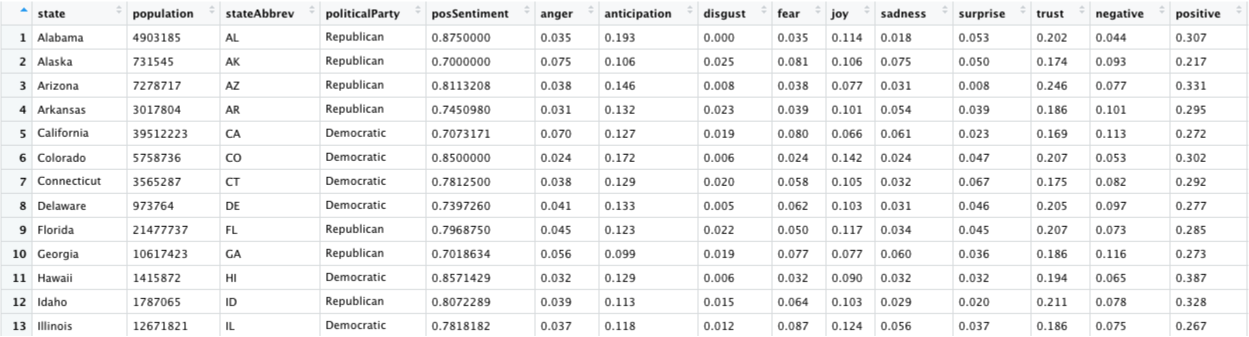
Data Set #3: Resulting Master Data Set with Speech Sentiment Analysis, and Health Indicators by State

Discretization

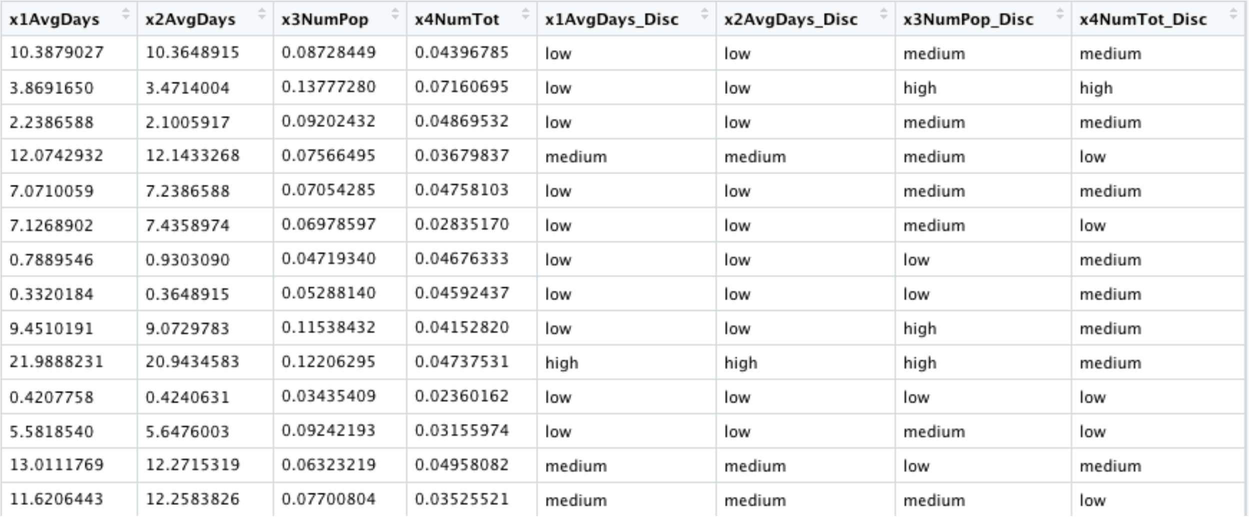
Two functions are written to create discretized versions of the health indicator and sentiment variables, respectively. Both functions take in a parameter for the variable to be discretized. The minimum and maximum variable values are calculated, and then used to conditionally apply a value of low, medium, or high, based on whether the variable's value falls within the bottom third of values (low), middle third of values (medium), or top third of values (high). While this method is applied for discretizing the health indicator variables, the second function applied to sentiment variables takes in an additional parameter to append the variable name to each given result.

Master Data Set

The final, master data set is a data frame named *state_emotions_health*, which provides the basis for all models to be generated. In addition to the relative-to-population standardization previously described for health indicator data by state, the master data set includes normalized data on attributes derived from text (speech) sentiment analysis. This normalized value is described as the percentage of words within a state speech that reflected an emotion or a sentiment attribute (see sentiment analysis in the next section for full description). A sample of this data set can be found below:



	state	population	stateAbbrev	politicalParty	posSentiment	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	Alabama	4903185	AL	Republican	0.8750000	0.035	0.193	0.000	0.035	0.114	0.018	0.053	0.202	0.044	0.307
2	Alaska	731545	AK	Republican	0.7000000	0.075	0.106	0.025	0.081	0.106	0.075	0.050	0.174	0.093	0.217
3	Arizona	7278717	AZ	Republican	0.8113208	0.038	0.146	0.008	0.038	0.077	0.031	0.008	0.246	0.077	0.331
4	Arkansas	3017804	AR	Republican	0.7450980	0.031	0.132	0.023	0.039	0.101	0.054	0.039	0.186	0.101	0.295
5	California	39512223	CA	Democratic	0.7073171	0.070	0.127	0.019	0.080	0.066	0.061	0.023	0.169	0.113	0.272
6	Colorado	5758736	CO	Democratic	0.8500000	0.024	0.172	0.006	0.024	0.142	0.024	0.047	0.207	0.053	0.302
7	Connecticut	3565287	CT	Democratic	0.7812500	0.038	0.129	0.020	0.058	0.105	0.032	0.067	0.175	0.082	0.292
8	Delaware	973764	DE	Democratic	0.7397260	0.041	0.133	0.005	0.062	0.103	0.031	0.046	0.205	0.097	0.277
9	Florida	21477737	FL	Republican	0.7968750	0.045	0.123	0.022	0.050	0.117	0.034	0.045	0.207	0.073	0.285
10	Georgia	10617423	GA	Republican	0.7018634	0.056	0.099	0.019	0.077	0.077	0.060	0.036	0.186	0.116	0.273
11	Hawaii	1415872	HI	Democratic	0.8571429	0.032	0.129	0.006	0.032	0.090	0.032	0.032	0.194	0.065	0.387
12	Idaho	1787065	ID	Republican	0.8072289	0.039	0.113	0.015	0.064	0.103	0.029	0.020	0.211	0.078	0.328
13	Illinois	12671821	IL	Democratic	0.7818182	0.037	0.118	0.012	0.087	0.124	0.056	0.037	0.186	0.075	0.267



x1AvgDays	x2AvgDays	x3NumPop	x4NumTot	x1AvgDays_Disc	x2AvgDays_Disc	x3NumPop_Disc	x4NumTot_Disc
10.3879027	10.3648915	0.08728449	0.04396785	low	low	medium	medium
3.8691650	3.4714004	0.13777280	0.07160695	low	low	high	high
2.2386588	2.1005917	0.09202432	0.04869532	low	low	medium	medium
12.0742932	12.1433268	0.07566495	0.03679837	medium	medium	medium	low
7.0710059	7.2386588	0.07054285	0.04758103	low	low	medium	medium
7.1268902	7.4358974	0.06978597	0.02835170	low	low	medium	low
0.7889546	0.9303090	0.04719340	0.04676333	low	low	low	medium
0.3320184	0.3648915	0.05288140	0.04592437	low	low	low	medium
9.4510191	9.0729783	0.11538432	0.04152820	low	low	high	medium
21.9888231	20.9434583	0.12206295	0.04737531	high	high	high	medium
0.4207758	0.4240631	0.03435409	0.02360162	low	low	low	low
5.5818540	5.6476003	0.09242193	0.03155974	low	low	medium	low
13.0111769	12.2715319	0.06323219	0.04958082	medium	medium	low	medium
11.6206443	12.2583826	0.07700804	0.03525521	medium	medium	medium	low

Transactions Data Set

Additional transaction data sets are generated for association rules mining models. Two transaction data sets are created, using the as() function with parameters of the discretized data frame to be used, as well as type *transaction*:

- **healthInd_party:** includes state, political party, and discretized health indicators
- **healthPartyTransactions:** includes only political party and discretized health indicators

Models for Analysis

Speech text mining for sentiment analysis

Identify sentiment and attitudinal attributes from state Governors

When setting out to examine a potential relationship between linguistic behaviors and individual health and well-being, a common starting point is often sentiment analysis. Broadly speaking, sentiment analysis is the interpretation and classification of emotions (positive, negative, and neutral) within text data using text mining techniques. The intention for this project is to apply such techniques to interpret and better understand the general nature, tone, and sentiment in each state Governor's speech. When people listen to such an address, they use their understanding of the emotional intent of words to infer whether a speech, in whole or in part, is positive or negative, or perhaps characterized by some other more nuanced emotion like surprise or disgust. Text mining techniques allow for this type of analysis on emotional content from text programmatically[5].

A central tool is the selection of a "lexicon" to be applied for the intended nature of the analysis or inquiry. These lexicons contain English words where words are assigned scores for positive/negative sentiment, and emotions like joy, anger, sadness, and so forth. For the Governor speeches, given the nature of the inquiry, the NRC Emotion Lexicon[6] is selected for conducting sentiment analysis. The NRC is a list of words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

[5] Sentiment analysis with tidy data. Retrieved from: https://www.tidytextmining.com/sentiment.html

[6] NRC Word-Emotion Association Lexicon. Retrieved from http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

The applied lexicon identifies words in a dataset in a binary fashion (0 (not associated) or 1 (associated)) and then categorizes those words, in this case in a speech, as being a positive or negative sentiment or an emotion described as anger, anticipation, disgust, fear, joy, sadness, surprise, or trust.

To prepare for applying the NRC lexicon for sentiment analysis, a function is written in R to transform each speech observation into a long format vector of words:

```
stateWords <- function(stateName) {
  statewords <- subset(speechDF_Norm, speechDF_Norm$State == stateName)
  statewords$State <- NULL
  statewords$Speech <- NULL
  statewords[statewords==0] <- NA
  statewords <- statewords[, colSums(is.na(statewords)) != nrow(statewords)]
  indexState=length(statewords)-1
  statewords <- statewords %>% gather(Word, Occurence, 2:indexState)
  statewords <- statewords$Word
  return(statewords)
}
```

When complete, a second function is written to count NRC associations for emotions and sentiments within each observational speech, according to the identified words. (A complete table listing for association counts can be referenced in supplemental R code, *IST 707 Project File_Master Code 20200316.R*)

```
    stateEmotions <- function(STwords) {
      emotions <- get_nrc_sentiment(STwords)
      emo_bar = colSums(emotions)
      emo_sum = data.frame(count=emo_bar, emotion=names(emo_bar))
      emo_sum$emotion = factor(emo_sum$emotion,
levels=emo_sum$emotion[order(emo_sum$count, decreasing = TRUE)])
      return(emo_sum)
    }
```

Finally, the matrix data format capturing NRC associations is converted into a working data frame that includes additional attributes identifying the affiliated political party and state abbreviation for each speech observation. This data processing allows for examination of each reporting emotion and sentiment by state as illustrated in the collection labeled Figure 3.

Figure 3: NRC emotions and sentiments for U.S. Governors' speeches (2019)

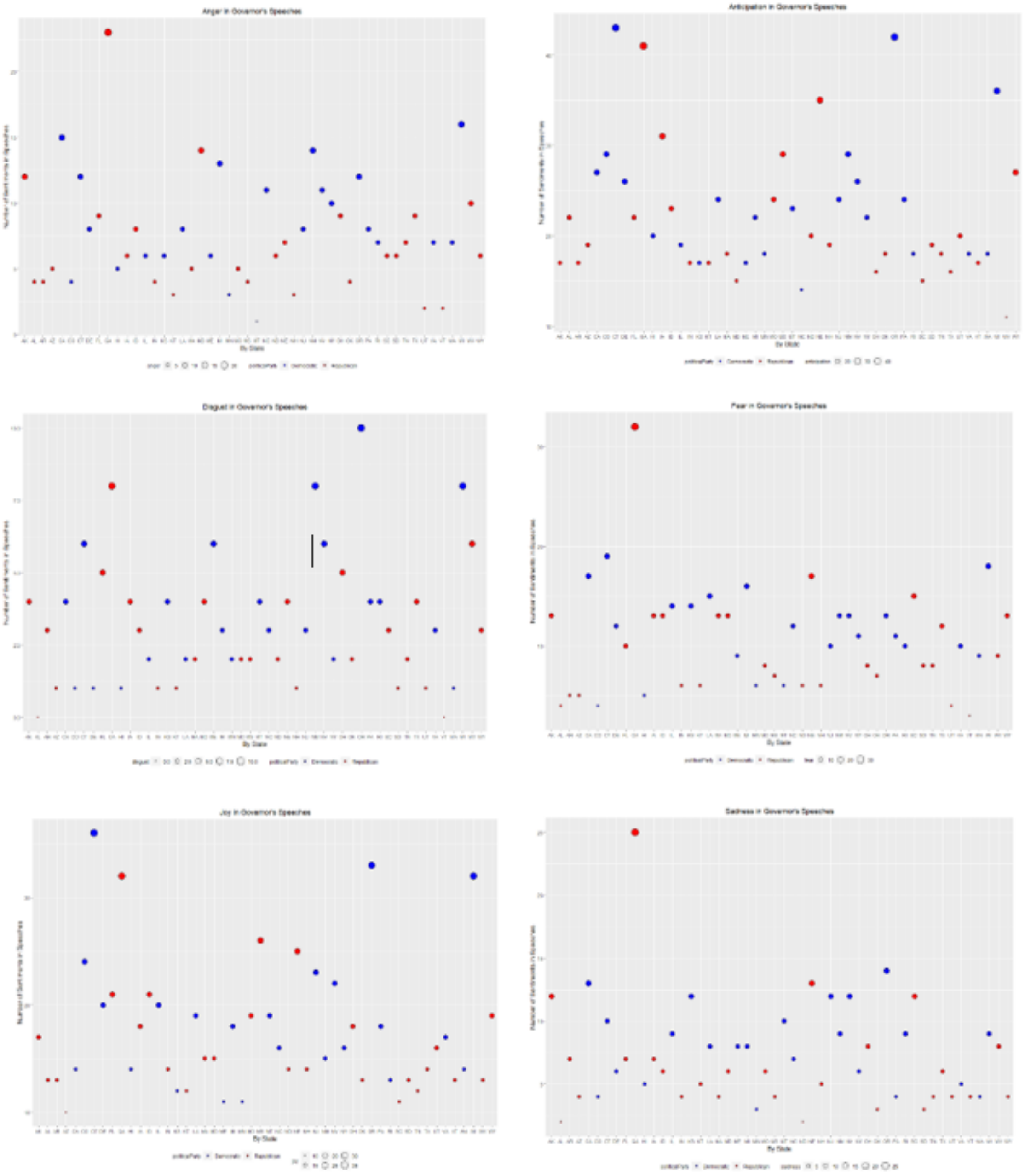
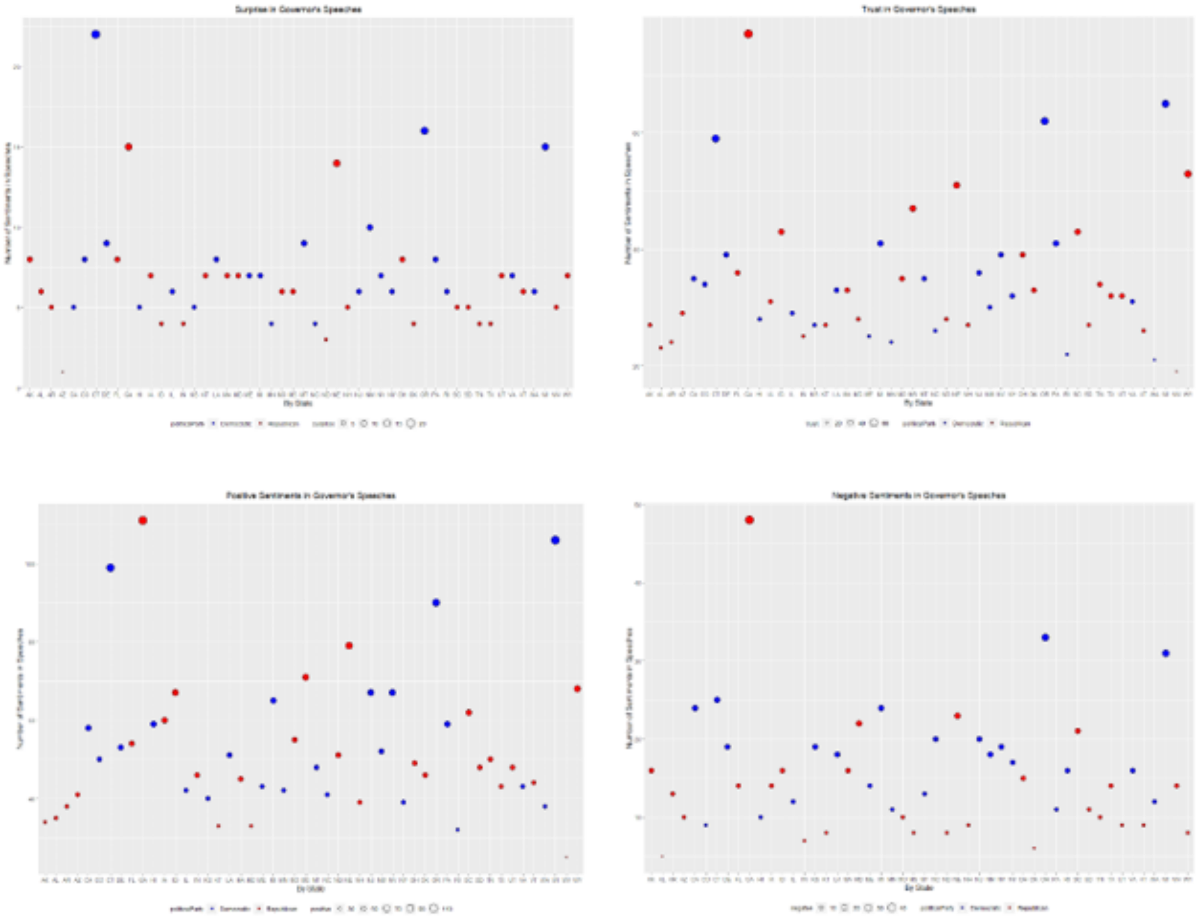


Figure 3 (continued): NRC emotions and sentiments for U.S. Governors' speeches (2019)



Association Rule Mining with the Apriori Algorithm (Unsupervised Learning)

Association Rules (AR) Mining is applied to examine the co-occurrence frequency and probability of certain items appearing in a "transaction". In this case, the co-occurrence or frequency of certain sentiments from a speech with selected indicators of health or political party affiliation. To apply the apriori algorithm in the R package, 'arules', the unsupervised learning model requires discretized transactions data in order to generate a comparative set of rules for transactional examination. In general, the resulting "rules" associate attributes or groups of attributes with a resulting attribute based on preset values for support, confidence and a minimum rule length.

Support describes the number of times the group of attributes, also described as the Left-Hand Side (LHS) and the resulting attribute, called the Right-Hand Side (RHS), appear together divided by the total number of opportunities. The group of attributes can contain one or more elements. The mathematical formula is described below where N is an attribute, M is the resulting attribute and T is the total number of opportunities.

$$\texttt{Support}(\{N_1, N_2..N_x\} \texttt{->} \{M\} = \texttt{Count}(\{N_1, N_2..N_x,M\})/T$$

Confidence describes the number of times the group of attributes and the result appear together divided by the number of times the LHS group of attributes appears without the result. The formula is described as:

$$\texttt{Confidence } (\{N_1, N_2..N_x\} \texttt{->} \{M\} = \texttt{Support count } (\{N_1, N_2..N_x,M\})/ \texttt{ Support count}(\{N_1, N_2..N_x\})$$

The model also produces a third measure to evaluate the rules called Lift. Lift describes the probability of the LHS and the RHS together divided by the probability of the LHS times the probability of the RHS. Mathematically this can be written as:

$$Lift(\{N_1, N_2..N_x\} \texttt{->} \{M\} = Support(\{N_1, N_2..N_x\} \texttt{->} \{M\}/(Support(\{N_1, N_2..N_x\})\ x\ Support\{M\}$$

Co-occurrence frequencies and relationships among sentiments and health indicators

The AR modeling is applied initially to identify rules based on health indicators (physical, mental, insurance coverage and employment), the Governor's political party affiliation, and the overall sentiment of the "state of the state" speech. Several models are run to narrow down the resulting rules. Table 7 shows the modeling parameters as well as the resulting rule count. The model also allows for a specification of minimum rule size. For this inquiry, the minimum rule size remained constant at 2. Evaluation targets for parameters are set at confidence greater than 0.08 (80%), support above 0.15 (15%), and a lift greater than 1.

Table 7: Health, political party, and sentiment Association Rules parameters

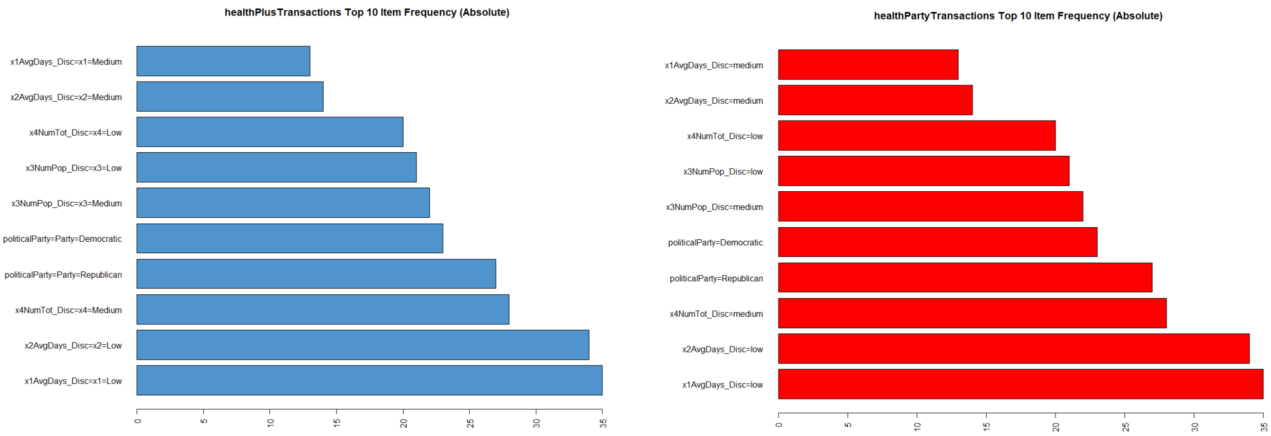
Support	Confidence	Minimum Length	Rule Count
0.2	0.8	3	31

Additional configurations were attempted to specify the left-hand side for the following attributes (the parameters are held consistent with a general rules search):

- `High sentiment`
- `Low sentiment`
- `High physical health`
- `Low physical health`
- `High mental health`
- `Low mental health`
- `High uninsured population proportion`
- `Low uninsured population proportion`
- `High unemployment population proportion`
- `Low unemployment population proportion`

Finally, a more complete view of the frequency of items in the ‘healthPlusTransactions’ and ‘healthPartyTransactions’ data is provided in Figures 4-5. Here, an examination of the top 10 items in the “transactions” show the higher frequency of co-occurrence for poor physical and mental health days (X1 and X2) in association with both moderate unemployment (X4) and political parties.

Figures 4-5: Top 10 item frequency in health-related transactions data



Co-occurrence frequencies and relationships among political parties and health indicators

Associations were also sought among individual political parties and selected health indicators, without regard to Governor speech sentiment. Again, with the apriori algorithm, association rule mining was run with the listed parameters (Tables 8-9). The minimum rule length was decreased to two, after finding few rules when running the algorithm with higher values for this parameter.

Table 8: Political party and health indicator rule parameters: Right-hand side political party = Republican

Support	Confidence	Minimum Length	Rule Count
0.2	0.8	2	7

Table 9: Political party and health indicator rule parameters: Right-hand side political party = Democratic

Support	Confidence	Minimum Length	Rule Count
0.2	0.8	2	35

Classification modeling (Unsupervised and Supervised learning)

Clustering analysis for patterns of similarities and dissimilarities among sentiment, parties, and health indicators

Clustering techniques use distance measures to decide the similarities or dissimilarities among data objects or observations. The central idea is that the closer objects are by distance—in this case, the distance between observations which are defined as observations related to sentiment, political party, and health—the more similar they are and can be grouped or categorized as clusters together. The center of each resulting cluster of observations can also be measured and the distance between different cluster centers (centroids) can further be used to identify similarities and dissimilarities among a selected collection of observations.

Identifying a specific distance measure for clustering analysis largely depends on the type of variables present in the normalized dataset. The distance measures (how far or different two objects are) applied for numeric/nominal (quantitative) variables are different then measures of similarity (how close or similar two objects are) that are used for categorical (qualitative) variables.

The clustering techniques identified for exploring a potential relationship among political parties, speech sentiment, and health attributes include the K-Means (partitioning approach using the 'kmeans' algorithm in R) and HAC (hierarchical approach using the 'hclust' algorithm) modeling functions. As with other common clustering techniques, the default distance measure, given the normalized numeric variables in the dataset, is the Euclidean distance measure. When Euclidean distance is applied, observations with high feature values (e.g. high % occurrence of positive sentiments in a speech) are clustered together in comparison to observations with low feature values.

K-means Clustering Model

K-means modeling is applied to the master dataset with class labels, attribute variables, and discretized health indicator numeric factor classes consisting of 50 observations with 7 variables. The cluster count is set to two to reflect the differentiation of High/Low and Democrat/Republican splits that exist in the groups. K-means chooses a quantity of centroids, in this case two, and iterates across the data calculating the distances from each point, in this case the word percentages. Cluster centers are revised until the centroids do not change. Cluster assignments, individual word clusters and document clusters were generated to visualize the data.

```
model_h <- kmeans(disc2_test, centers =  2)
```

Hierarchical Clustering Model

Hierarchical clustering (HAC) is run on the master dataset with class labels, attribute variables, and discretized health indicator numeric factor classes consisting of 50 observations with 7 variables. The dataset is discretized into min-mid-max levels with distance calculations generated utilizing the default, Euclidean and Manhattan distance calculations. Clustering is run against each of the distance measures using the Ward.D and centroid methods. HAC produces tree structures known as dendrograms with different distance measures and clustering techniques resulting in various dendrograms for result analysis. To assist in the analysis, each dendrogram is cut into two clusters with states identified by color according to their political party affiliation.

```
distdiscHealth_E <- dist(disc2_HAC, method="euclidean")

distdiscHealth_M <- dist(disc2_HAC, method="manhattan")

discHac_E <- hclust(distdiscHealth_E, method = "ward.D")

discHac_E_C <- hclust(distdiscHealth_E, method = "centroid")
```

Decision Trees Classification

Decision trees are supervised classification learning models which requires the data to be split into training and test sets. Critical to the success of a decision tree is a sufficient data sample size. The discretize data set created for the project only has 50 rows of data with resulting training/test set splits further reducing the data size.

A sample of the master data set is created using the sample.int function. In order to obtain reproducible results, a seed value is set, with values described at the end of this section. The parameters are specified with the number of rows of the data set as the n value, the size of the sample was 80% of that n value, and sample values were not replaced. This subset of the master data set (80% of rows) is assigned to the train variable, and the inverse of this data set (the other 20% of rows) is assigned to the test variable. The train:test ratio is then calculated using the length of the sample divided by the number of rows in the initial *state_emotions_health* data frame. Then, the label variable is removed from the testing data set and saved into a new variable, testLabeled. The remaining, unlabeled data is then placed into a 'testUnlabeled' variable.

The training model is then created using the rpart() function to train the decision tree algorithm. The formula parameter is designated as *politicalParty ~.* (class label for political party classification is the output variable, with all other health indicator attributes as the inputs). The data set is the train data set established in the previous paragraph, and the method is of type class. The control parameter includes the rpart.control function with a cp value of 0.03 and minsplit value of two. This is saved in a new variable named *train_tree*. The summary function is then called on the *train_tree* variable.

```
train_tree <- rpart(politicalParty ~ x1AvgDays, x2AvgDays, data=train,
   method="class", control=rpart.control(cp=0.08, minsplit=2))
```

```
summary(train_tree)
```

After training the algorithm, it is then tested using the 'predict()' function in R. Parameters include the train_tree variable as the object, the test data set from the first paragraph of this section as the newdata, and class as the type. This result is saved in a new variable called *predicted*.

```
predicted = predict(train_tree, test, type="class")
```

To visualize the decision tree for analysis, the 'fancyRPlot()' function is called with the *train_tree* variable, and title with the main parameter, to display the training decision tree. A confusion matrix is also created with the predicted and true label. The accuracy (effectiveness of the trained classification model) is calculated from the confusion matrix by adding up the diagonal values in the matrix, dividing them by the total number of values in the matrix, and multiplying by 100.

Finally, 10-fold cross validation is completed on this model by enclosing the code outlined above within a function, with a parameter value from one to ten. The sampling seed value noted previously is set to the parameter value multiplied by ten, which increases with each function call. After the model is run ten times, the overall accuracy of the model is calculated by finding the average of each of the ten accuracies. This is the overall accuracy for the model in terms of classifying political parties by selected health indicators.

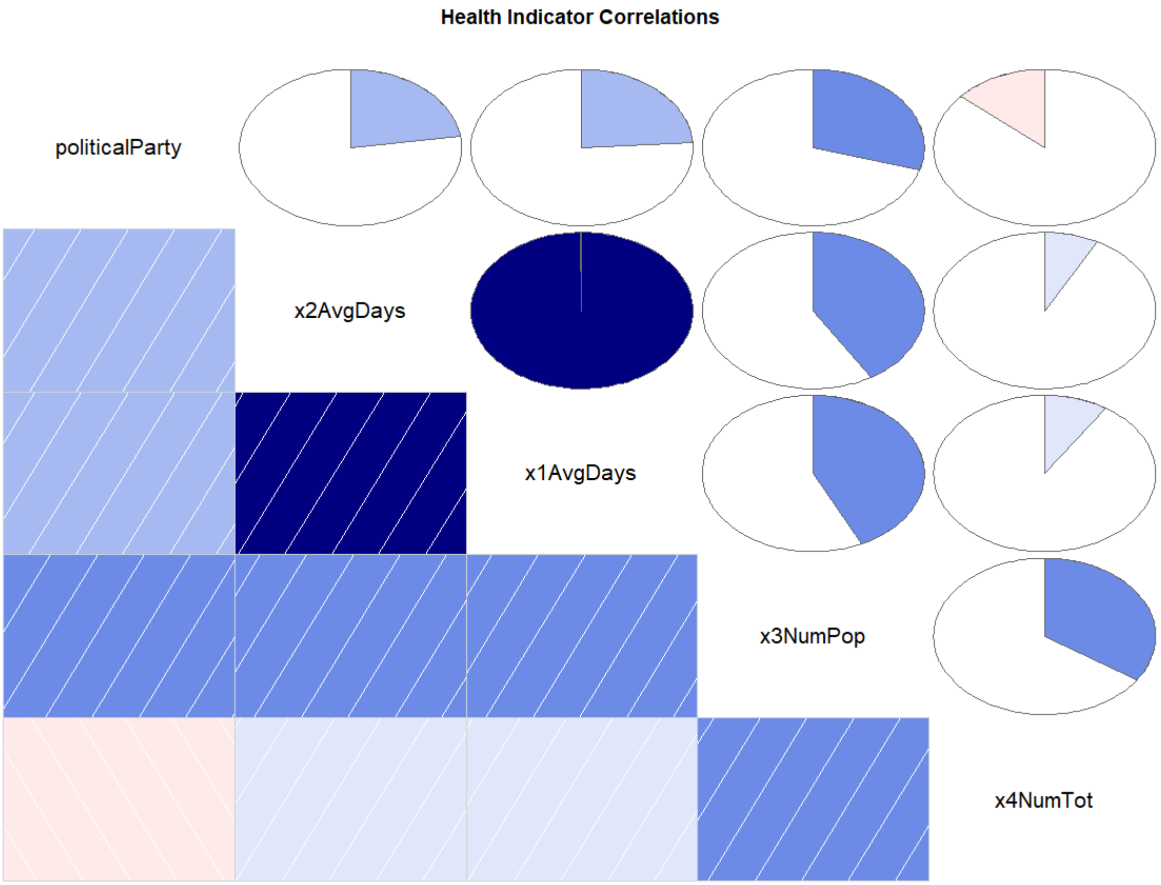
Classification prediction modeling (Supervised Learning)

Classification algorithms (naïve Bayes, SVM) for building and testing applied predictive learning models

Naive Bayes

In general, the principle behind naïve Bayes is the Bayes theorem which is used to create "classifiers" based on conditional probability - the probability of an event occurring based on information about the event in the past. For modeling, there is an assumption that all input variables are independent of one another. For this reason, correlation is evaluated among the health indicator variables as illustrated in Figure 6. The 'corrgram' function R can be interpreted as the darker the blue color shading, the closer the correlation is to 1 (the darker red is closer to -1). The circles similarly show a fuller circle when a correlation is closer to one, and an emptier circle is closer to 0. Based on this analysis, both x1AvgDays and x2AvgDays could not both be used for naive Bayes modeling because they are not independent attributes in the dataset.

Figure 6: Correlations among health indicators in the dataset.



Training and testing data sets required for the naive Bayes modeling, and the sampling method, is the same as described previously for the Decision Tree model. The training model is then created using the naiveBayes() function. The parameters include the formula with political party as the output variable, and the four health indicators as the input variables, the data from the training data set created previously, and laplace smoothing set to 1 to account for zero values. (The laplace smoother adds a small number to each of the counts in the frequencies for each attribute, which ensures that each feature has a nonzero probability of occurring for each class.)

```
NB_Training_Model <- naiveBayes(politicalParty ~., data=train, laplace=1)
```

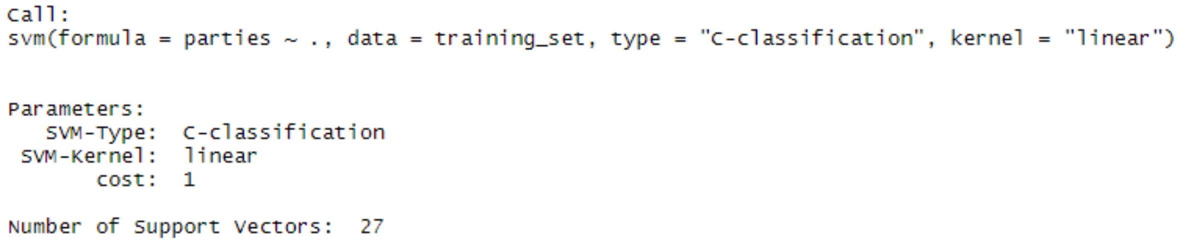
Support Vector Machines (SVM) Model

Support Vector Machines are linear classifiers. Each row of data is identified as a vector and the algorithm iterates through the data to identify the necessary vectors to define a separation of data. The lower the number of vectors needed to classify the results, the stronger the model. SVMs optimize the calculations by ignoring non-support vectors when performing the prediction decisions. SVMs have the ability to map data into a higher dimensional space allowing data that is inseparable in two dimensions to become separable. Additionally, the model used is a soft-margin SVM that allows for data within the margins in order to avoid over-fitting.

SVMs can only differentiate between 2 groups. In order to address multi-class problems, n(n-1)/2 SVMs need to be created. The support vector machine model supports pre-defined kernels to transform the data. Additionally a cost value is defined to reduce the impact of data within the defined margin. SVMs have a high tolerance to noise, are scalable and allow probabilistic predictions. SVMs have similar requirements for data size as decision trees classification models.

Using different parameters allows for exploration of the best fitting SVM classifier to the training dataset. A good starting point for SVMs is to use a “linear” or “radial” kernel (mathematical function) and manually adjust the cost in powers by 10 to find the best value as evaluated on a testing dataset. Overall, the goal is to identify modeling with the highest accuracy of classification prediction that requires the lowest number of “support vectors”. This represents the best fitting model.

The first SVM model is trained from the master data set for political party classification (class = ‘parties ~.’) and applies kernel = “linear” with cost = 1 (default) to generate a classifier with 27 support vectors:



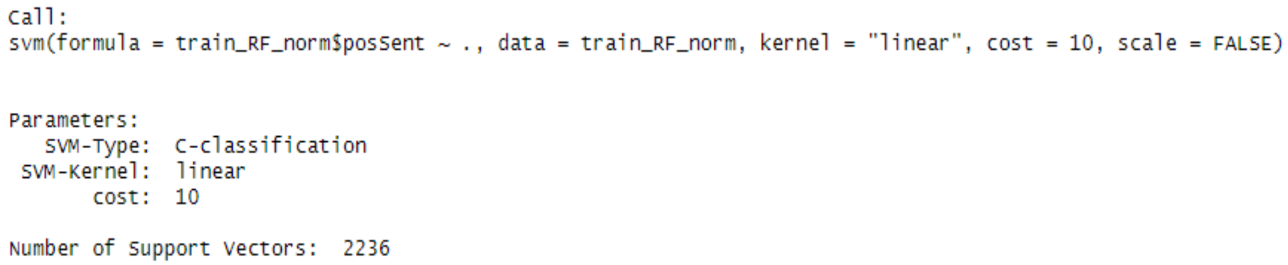
```
Call:
svm(formula = parties ~ ., data = training_set, type = "C-classification", kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  27
```

A second, more expansion, modeling is undertaken by returning to the datasource and creating a revised county-level data set (3,140 observations of health indicator attributes by state) that includes a discretized class label indicating the level of positive sentiment (low, medium, high) in the state

Governor's speech. Similar to the first model parameters, the second SVM model is trained for positive sentiment classification, based on county-level health indicator measures, and applies kernel = "linear". With cost indicated at 10, the model generates a classifier with 2,239 support vectors.



```
Call:
svm(formula = train_RF_norm$posSent ~ ., data = train_RF_norm, kernel = "linear", cost = 10, scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  10

Number of Support Vectors:  2236
```

K Nearest Neighbor (KNN) Model

K nearest neighbor models identify and categorize data by calculating distance between data points. Like SVMs, it is a linear classifier. KNN is classified as a lazy learner and works with no class labels on both the training and test data. This is one of the benefits of the model in that no previous assumptions are made. This works well when the decision functions are very complicated.

KNN is sensitive to training data with a high degree of noise and it utilizes all attributes as part of the classification. This results in a high computational cost and nearly all computation takes place in the prediction step. Prior to running the model, a new data set was generated by copying the training data set without labels. The labels were saved to determine model accuracy. The KNN algorithm is tuned by varying the number of labels that are considered to build the classes. This variable is noted as k.

The initial value of k was chosen as the square root of the number of rows (1400) rounded to the ones place. Additional models are tested with two times k and k divided by 2.

```
    k <- round(sqrt(nrow(discTrainN)))

    dblk <- k*2

    halfk <- k/2

    kNNdiscTrain <- class::knn(train=discTrainNoLabelN, test=discTestNoLabelsN,
cl=discTrainLabelsN ,k = k, prob=TRUE)
```

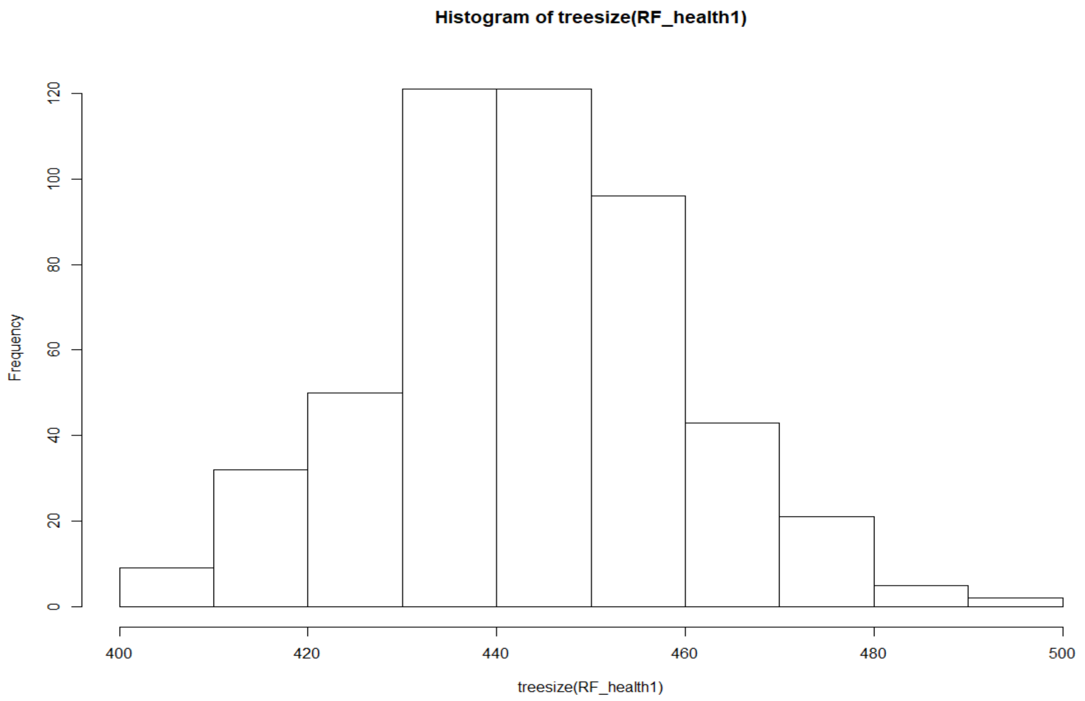
Random Forest Model

Random Forest is the final supervised learning model applied for examining project data questions. This particular model generates multiple models and then averages the individual results to produce a final, stronger model. Each of the individual sub-models is a decision tree. The approach reduces variance and avoids overfitting of the model. Random Forests work well with large data sets and high dimensional problems. However, it has the potential to over-fit noisy data sets and the model has fewer controls than other models. In general, Random Forest calls for the generation of multiple modeling activities that then combine (average) output rules for classification of data according to labeled factors of inquiry interest.

Given the characteristics required for Random Forest training, the expanded data set assembled from county-level data for secondary SVM applications is utilized to consider if select normalized health indicator measures can be used as prediction attributes for political party affiliation. Figure 6 shows the distribution of "terminal node" counts in the resulting model as an indication of the optimization that may require attention to avoid overfitting the classification model.

```
RF_health1 <- randomForest(party ~., data=train_DF_norm)
```

Figure 7: Correlations among health indicators in the dataset.



Results

The models identified for analysis were selected to assist in addressing core data questions aimed at examining the prospective relationship between the words expressed by elected leaders, their political party affiliation, and the general health and well-being of those they were elected to serve. Each line of inquiry is presented along with corresponding results derived from applied data mining techniques and classification algorithms leveraged to for the project objectives.

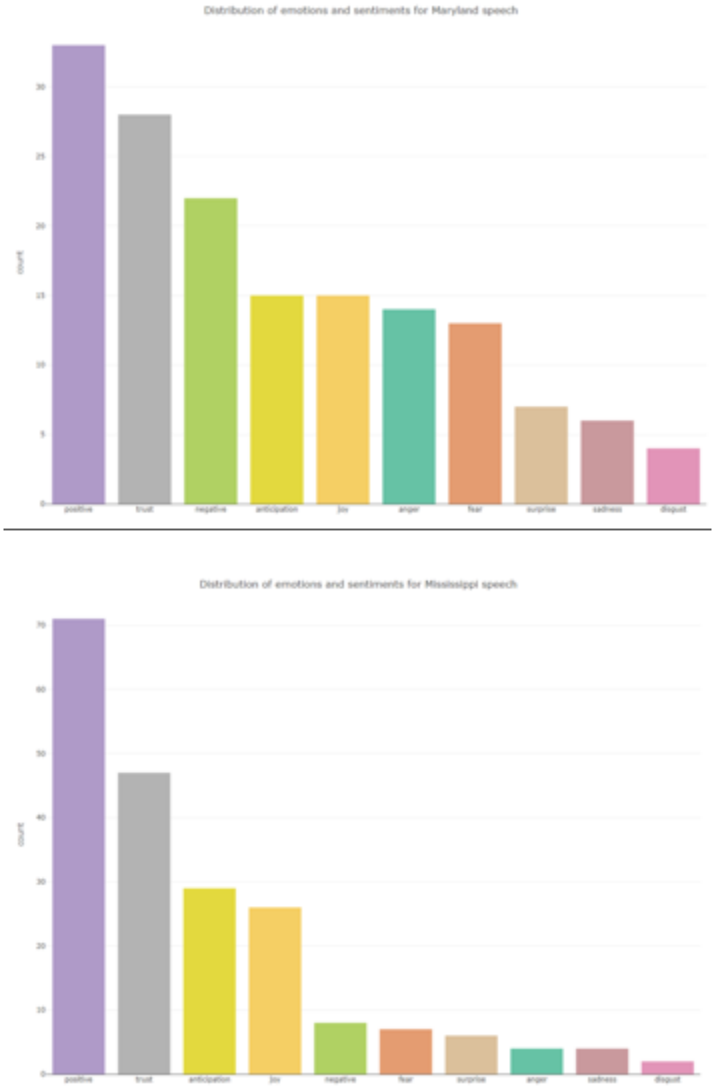
What our Governors are saying

What emotions and sentiments are conveyed by our elected leaders when addressing those living in their state?

Model 1: Speech text mining for sentiment analysis

An area of interest for the project was the utilization of sentiment analysis to define the specific emotions and sentiments that were conveyed by elected leaders when addressing their constituents in 2019. As illustrated in the earlier modeling of sentiment frequency, Governor speeches were largely positive in nature which can be attributed to their political role. Further analysis, however, was applied to recognize the emotions that characterized a diversity within "positive" speeches. In this case, the lowest positive sentiment occurrence in a speech was from the State of Maryland (60.0%) and the highest occurrence was from the State of Mississippi (89.9%). Figure 8 shows that the differences were driven by variances in two key emotions – anger and fear. So, while overall speech sentiment may have been positive in nature, there were underlying emotions that may indirectly influence the mental health and well-being of those living within the states.

Figure 8: Distribution of emotions for low (MD) and high (MS) positive sentiments



To further examine the expressed words and emotions of elected leaders, and the potential relationship to the general health and well-being of state residents, the distribution of emotions presented in Figure 9 reflect Governor speeches from states that have a high prevalence of both Poor Physical Health Days (30-day average number) and Poor Mental Health Days (30-day average number)—the states of Oklahoma and West Virginia. As illustrated, the varied distribution and skew toward positivity and trust would signal that expressed emotions may not be an influencing factor on health. However, a look at the actual words that were used in the speeches (Figure 10) of states with poor health indicators (Oklahoma and West Virginia), in comparison to a state with good health indication, shows that Governor speeches from states with poorer health were focused on largely administrative or uninspiring themes while healthier states (demonstrated by South Dakota) appear to offer more inspiring and growth-oriented sentiments.

Figure 9: Distribution of emotions for states with poor health indicators (Oklahoma and West Virginia)

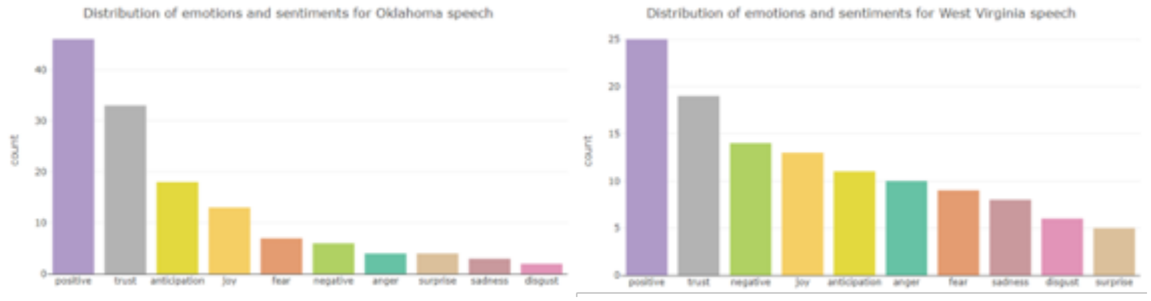
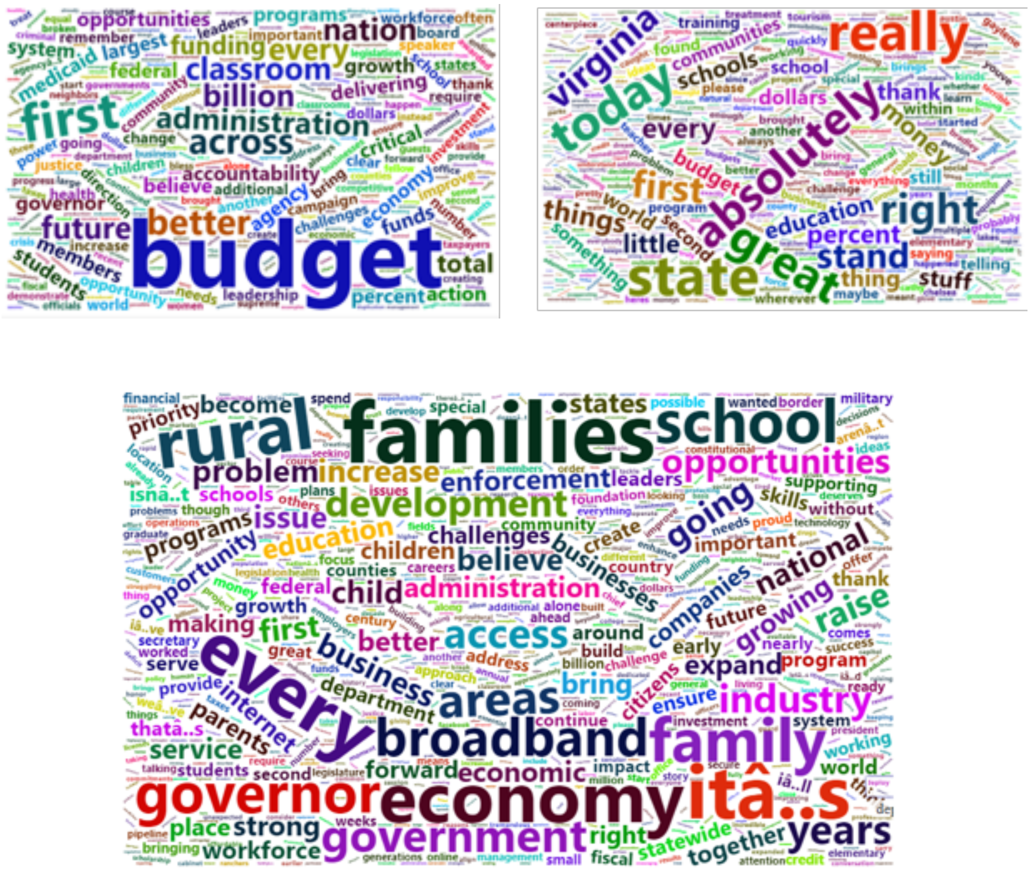


Figure 10: Word Clouds for states with poor health indicators (top: Oklahoma and West Virginia) in comparison to a healthier reporting state (bottom: South Dakota)

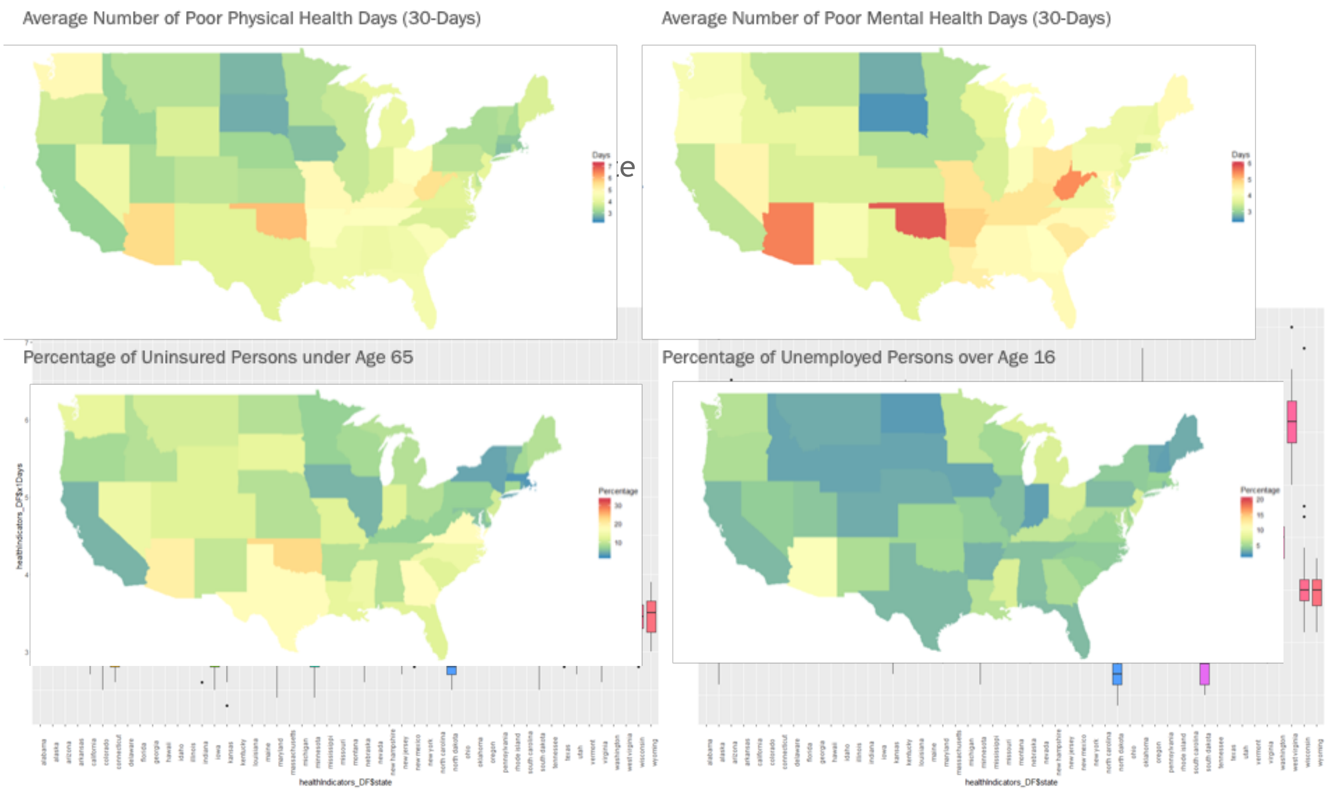


How we are feeling

Do political party positions, as expressed through their words, elicit certain emotions that may have an impact on our health and well-being?

The health indicator values chosen for evaluation were explored by state to identify whether health outcomes vary by location. As seen in Figures 11-14, states with darker blue shading have better health outcomes, where states shaded with more red hues have poorer outcomes. Oklahoma has the highest average number of poor physical and mental health days on average, as well as the highest percentage of unemployed. West Virginia is not far behind with the second highest average number of poor physical and mental health days. The distribution of these health outcomes is further explored in the boxplots in Figure 15. Again, West Virginia and Oklahoma show the highest rates of poor physical and mental health days. In contrast, South Dakota shows the lowest values for these indicators. This supports the findings in the previous section, where major themes in Governors' speeches are reflective of the health of their state, as seen in Figure 10.

Figures 11-14: US Maps with Health Indicator Values by State


Average Number of Poor Physical Health Days (30-Days)
Average Number of Poor Mental Health Days (30-Days)
Days
Days
Percentage of Uninsured Persons under Age 65
Percentage of Unemployed Persons over Age 16
Percentage
Percentage
healthIndicators_DF$x1Days
healthIndicators_DF$state
healthIndicators_DF$state

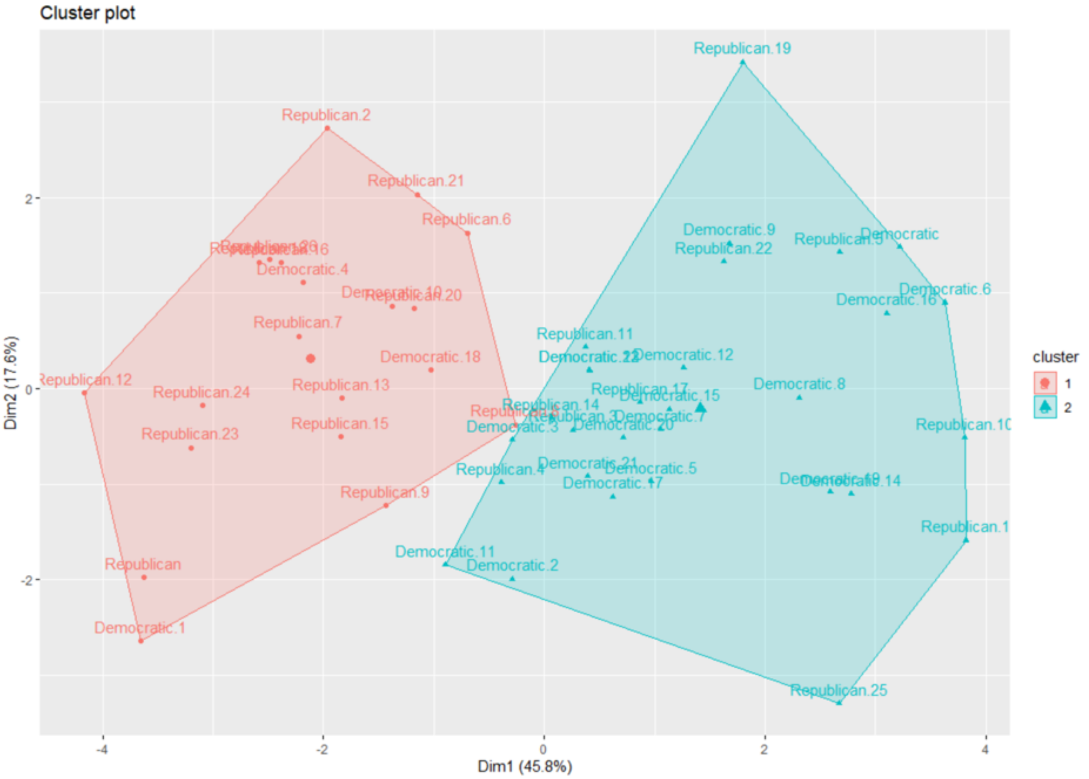
The sentiments of political parties

Do certain emotions and sentiments, as shared by our leaders, persuade or constrain our general health and well-being? Can certain expressed emotions predict our general health and well-being status?

Model 2a: K-means Clustering Model

As illustrated in Figure 16, political affiliation across all 50 states is clustered into two groups: a red cluster and a blue cluster representing the republican and democratic parties, respectively. Using the K-means algorithm, the optimal number of clusters with the greatest drop in Total Within Sum of Squares is two, thus used for the number of centroids within the algorithm, and conveniently the same number of parties in this analysis. Several states have been classified as republican within the democratic (blue) cluster, which brings about the accuracy of the model or an interesting insight on similar emotions conveyed across these states.

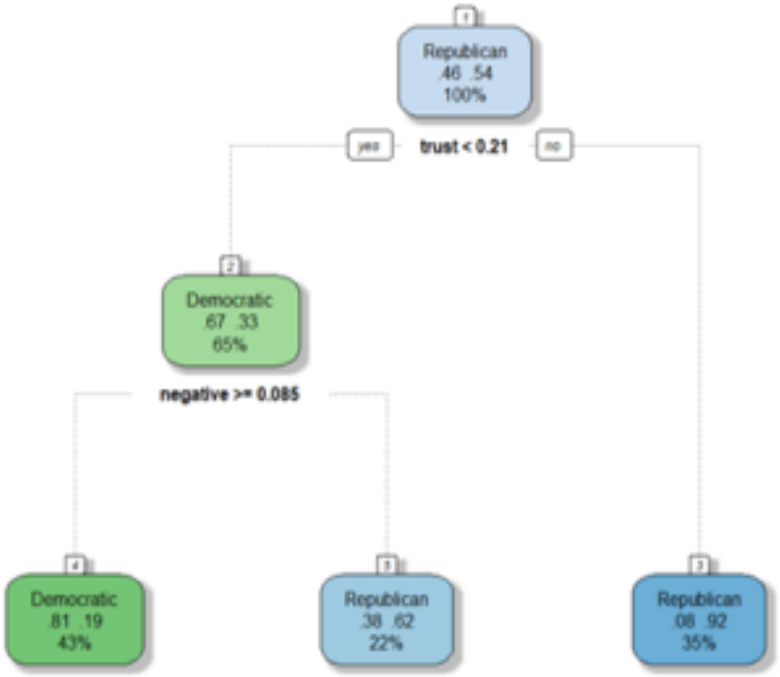
Figure 16: Political Affiliation Classified by Emotion



Model 2b: Decision Trees

Using rpart’s decision tree, in targeting the key emotions and sentiments in what would identify a state’s political affiliation, trust and negativity were found to be the key emotion and sentiment, respectively. If a state’s governor expressed trust with a frequency rate of less than or equal to 21%, or greater than 21% and expressed negativity with a frequency rate of less than 8.5%, their respective state would be considered republican; otherwise democratic. 57% of the states were classified as republican whereas the other 43% were classified as democratic, as seen in Figure 17.

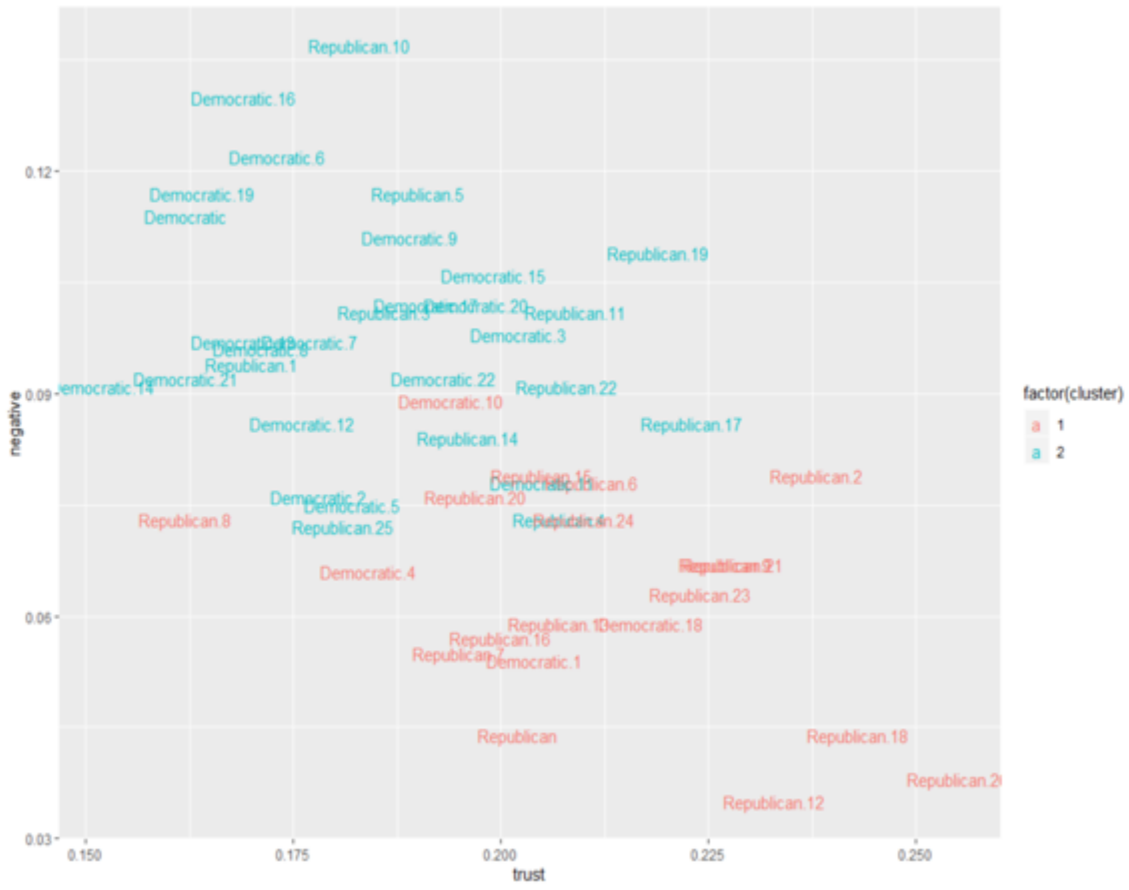
Figure 17: Emotion and Sentiment that Signal State Political Affiliation



Model 2c: K-means Clustering Model

Returning to the k-means algorithm without dimension reduction and only applying the key emotion and sentiment in determining political affiliation, trust and negativity, the cluster plot below, in Figure 18, shows a clear visual on the distinction of these two parties. Again, it is shown that the blue cluster mostly contains democratic parties as the red cluster represents the republican parties. This further validates the two plots above, Figure 16 and 17, in two matters: 1) emotions/sentiments from the governors' health speeches can assist in determining political affiliation of each state; 2) trust and negativity have the greatest weights in this classification.

Figure 18: Clustering Political Party Emotions/Sentiments Trust and Negativity



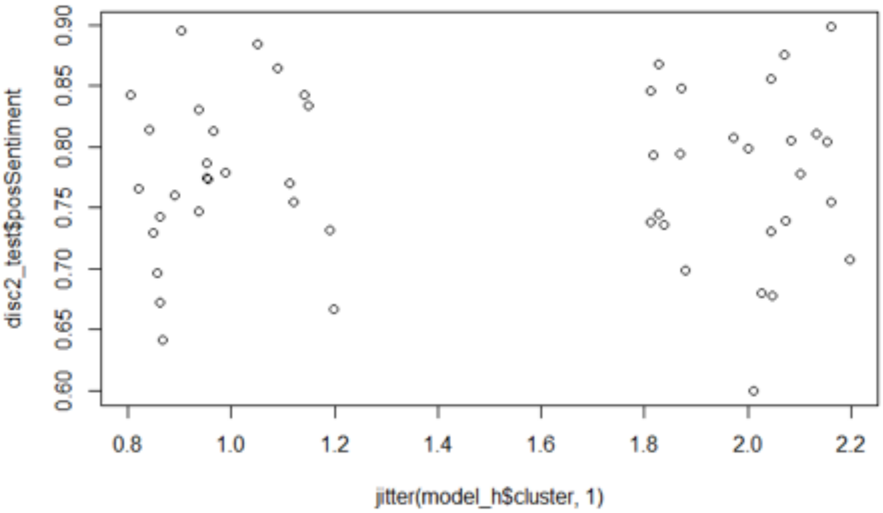
Our health and leader emotions

Do certain emotions and sentiments, as shared by our leaders, persuade or constrain our general health and well-being? Can certain expressed emotions predict our general health and well-being status?

Model 3a: K-means Model for Health, Party and Sentiment

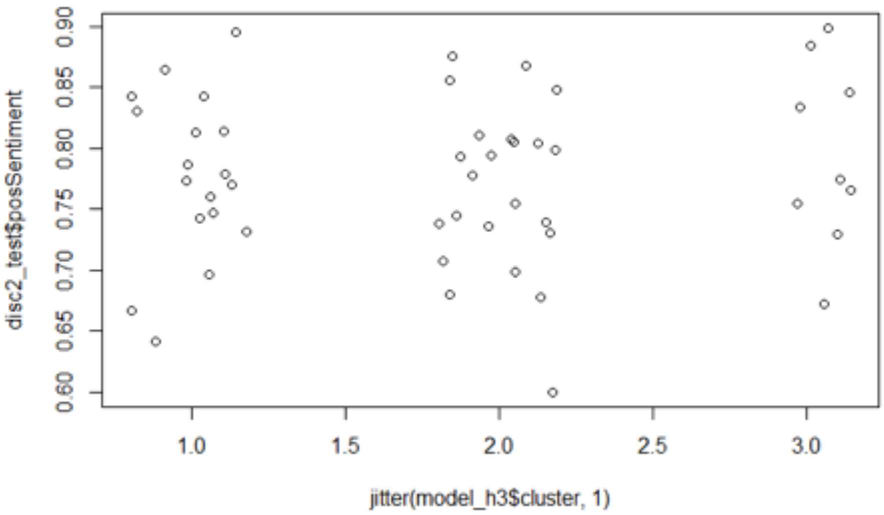
K-means is able to differentiate across all categories. It clusters equally successfully searching for two clusters or three. Given the split between the two parties and the differentiated health and sentiment data (high, medium and low), this result was unexpected. Figure 19 graphically depicts the two clusters based on the sentiment.

Figure 19: K-Means Two Sentiment Clusters



By comparison, Figure 20 depicts the same attribute with three clusters. In this case, there are fewer instances of speeches in the higher sentiment cluster.

Figure 20: K-Means Three Sentiment Clusters



Clustering by the other health attributes produce similar results with fewer vectors in the high cluster as shown in Figures 21 and 22.

Figure 21: K-Means Two Mental Health Days Clusters

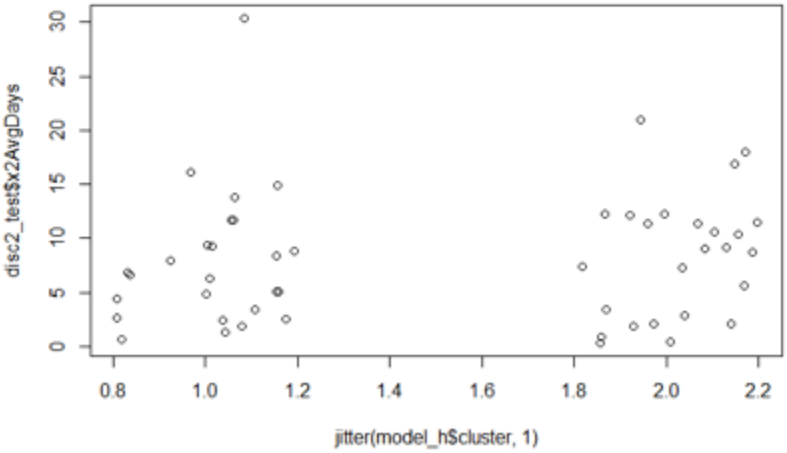
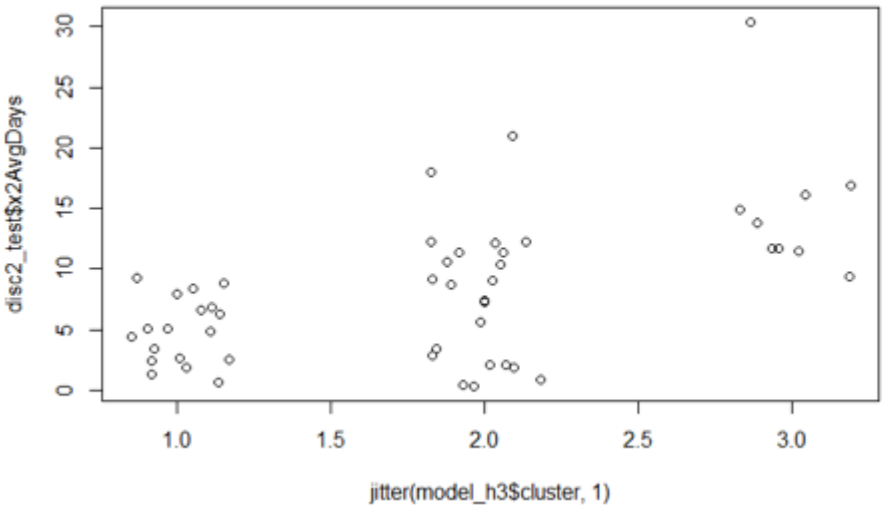


Figure 22: K-Means Three Mental Health Days Clusters



Model 3b: Hierarchical Clustering

The HAC model validates the results of the k-means model. The first analysis is performed using the Euclidean distance (L2) and a Ward.D algorithm producing the dendrogram in Figure 23. Figure 24 displays the results of the Manhattan distance (L1) with the same Ward.D algorithm. The States are not differentiated by either model . This result shows there is no clear differentiation when party, speech sentiment, and health factors are considered together.

Figure 23: HAC using Euclidean Distance and Ward.D

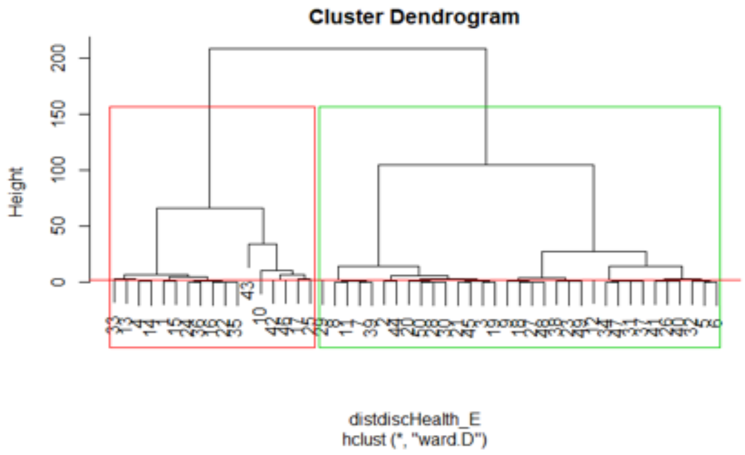
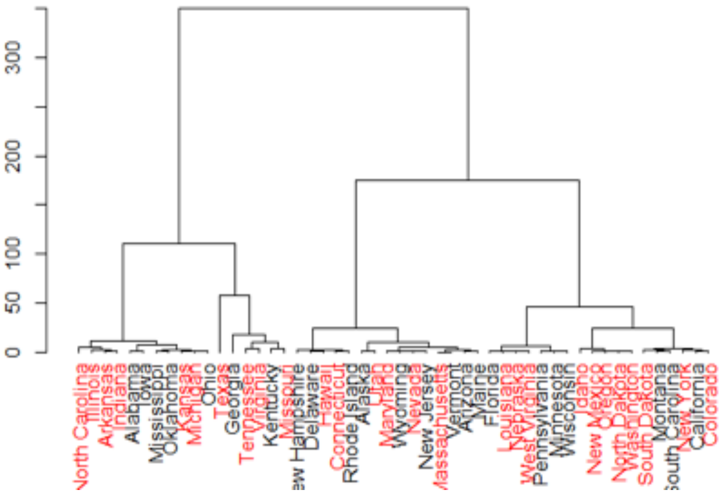


Figure 24: HAC using Manhattan Distance and Ward.D



Model 3c: Decision Tree and Support Vector Machines (SVM) Model

In total there are 3,142 counties or equivalently designated regions in the United States. However, the number of counties varies from state to state. In order to normalize the data, the data is rolled up to the state level. This reduces the total row count to fifty. Splitting the data into training and test reduces each data set to thirty-seven and thirteen, respectively. This quantity was insufficient to produce results.

Model 3d: K Nearest Neighbor (KNN) Model

The KNN model is executed against the non-discretized data set. Separate runs are made to train against the political party, as well as the sentiment. Since the political party is a label, it is converted to a numeric value for the sentiment training. Additionally, three values of K are chosen for this test: K as the square root of the row numbers, two times K and one half K. The accuracy is low for all versions of the model, confirming previous results that prediction of party or speech sentiment based on health values is not effective. Tables 10 and 11, below, identify the accuracy for each value of K when predicting the political party and the sentiment.

Table 10: KNN Accuracy Predicting of sentiment

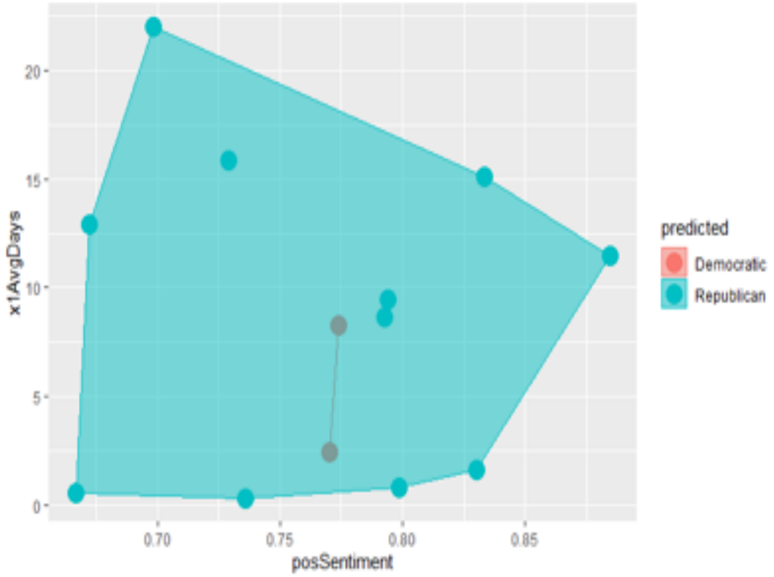
K Value	Accuracy	Kappa	Upper Accuracy	Lower Accuracy	P Value
K/2	30.77%	-0.0086	61.42%	9.10%	0.80
K*2	23.08%	-0.0833	53.81%	5.04%	0.93
SqRT K	38.46%	0.1186	68.42%	11.86%	0.60

Table 11: KNN Accuracy Predicting of Political Party

K Value	Accuracy	Kappa	Upper Accuracy	Lower Accuracy	P Value
K/2	53.85%	0.2041	80.78%	25.13%	0.81
K*2	38.46%	-0.1304	68.42%	13.86%	0.98
SqRT K	53.84%	0.2041	80.78%	25.13%	0.81

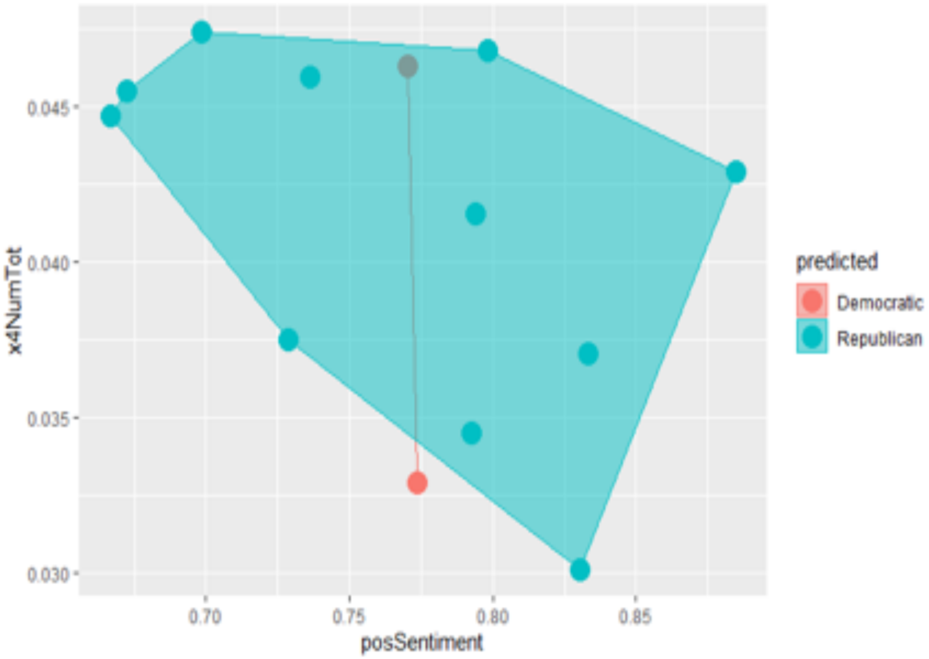
Examining the boundary diagrams further reinforces the overlapping prediction data. Figure 25 plots sentiment against physical health. The boundaries are entirely overlapped.

Figure 25: KNN physical Health vs Sentiment by Party



As seen in Figure 26, only when the supporting health attributes of the percentage uninsured and unemployed are examined, does the data partially separate. This is similar to the association rules results where only these attributes were only applied to the Democratic party.

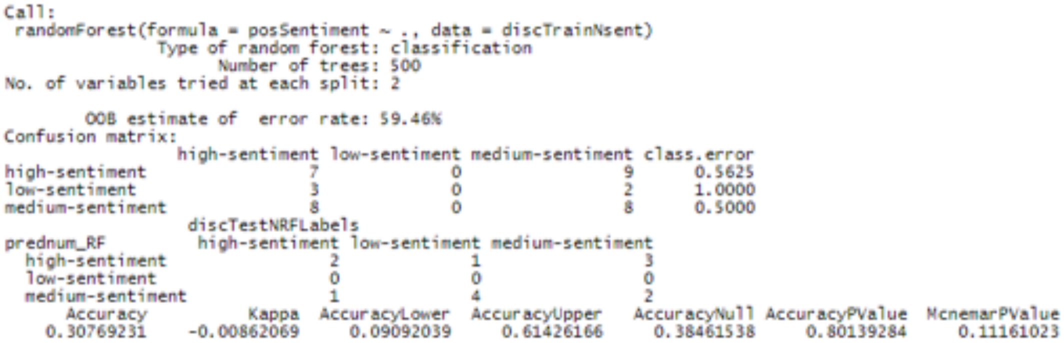
Figure 26: KNN Unemployment vs Sentiment by Party



Model 3e: Random Forest Model Results

The random forest model offers the least amount of tuning options. Similar to the KNN model, random forest processes the un-discretized data set. The same prediction targets of political party and sentiment that were examined in the KNN model are used in the random forest model. The random forest model performs similarly to the KNN models. The random forest models produce a training error estimate and corresponding matrix. Figures 27-30 display these results for the sentiment and political party predictions. The variable importance plots for each prediction model are also displayed. The state ranked highest in each case. When predicting sentiment, the political party ranked the lowest, indicating that health factors have a larger impact when determining the speech sentiment. However, when predicting political party, the sentiment was second only to the state, suggesting that this attribute has the greatest impact on predicting political party.

Figure 27: Random Forest Model detail predicting Sentiment



```
Call:
 randomForest(formula = posSentiment ~ ., data = discTrainNsent)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 59.46%
Confusion matrix:
                 high-sentiment low-sentiment medium-sentiment class.error
high-sentiment                7             0                9      0.5625
low-sentiment                 3             0                2      1.0000
medium-sentiment              8             0                8      0.5000
                  discTestNRFLabels
prednum_RF           high-sentiment low-sentiment medium-sentiment
  high-sentiment                  2             1                3
  low-sentiment                   0             0                0
  medium-sentiment                1             4                2
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull AccuracyPValue  McnemarPValue 
    0.30769231    -0.00862069     0.09092039     0.61426166     0.38461538     0.80139284     0.11161023 
```

Figure 28: Random Forest Variable Importance predicting Sentiment

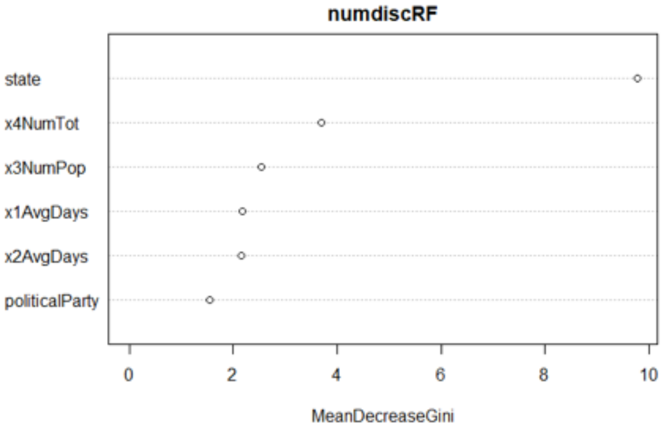
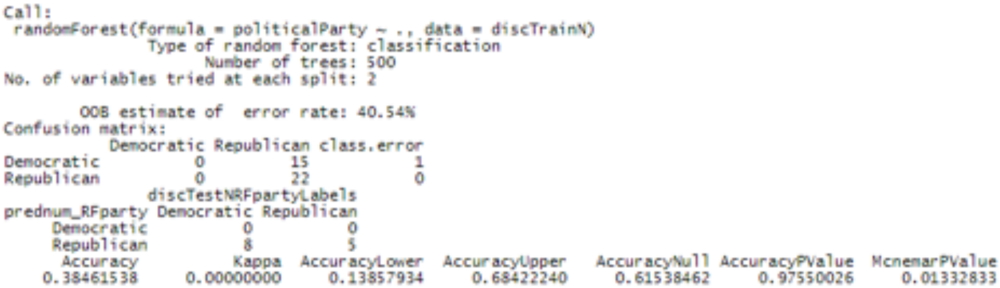


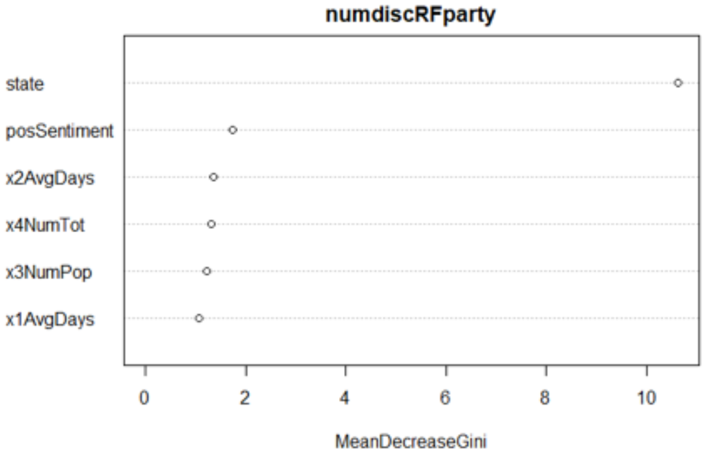
Figure 29: Random Forest Model detail predicting Political party



```
Call:
 randomForest(formula = politicalParty ~ ., data = discTrainN)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 40.54%
Confusion matrix:
           Democratic Republican class.error
Democratic          0         15           1
Republican          0         22           0
               discTestNRFpartyLabels
prednum_RFparty Democratic Republican
     Democratic          0          0
     Republican          8          5
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull AccuracyPValue  McnemarPValue 
    0.38461538     0.00000000     0.13857934     0.68422240     0.61538462     0.97550026     0.01332833 
```

Figure 30: Random Forest Variable Importance predicting Political Party



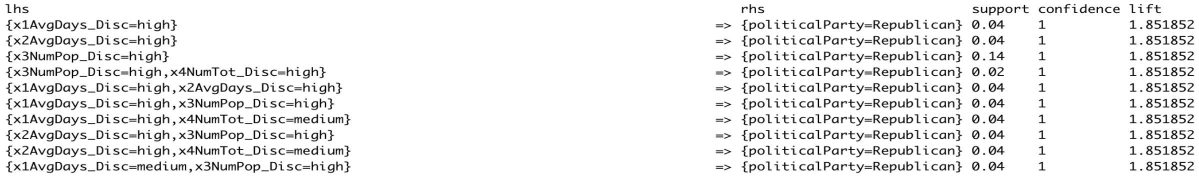
Political parties and our health

Can certain health indicators actually be attributed to a specific political party? Do political parties recognize that they are associated with specific health and well-being conditions in states?

Model 4a: Association Rules Mining - Apriori Algorithm

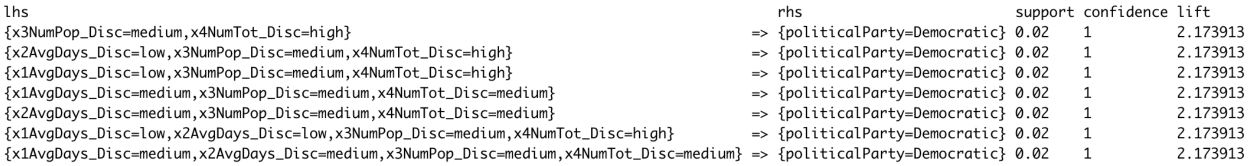
When running the apriori algorithm for association rules mining, to determine whether any association exists among political parties and specific health indicators, rules are generated with strong support, confidence, and lift values. There are apparent associations between political parties and health indicators involving the number of poor physical health days and mental health days per month. The Republican party is associated with high values for these indicators, while the democratic party is associated with low values for these indicators. A sample of the rules generated for each political party can be found in the Figures 31 and 32, as well as network diagrams exhibiting the relationships among the health indicators described and political party, in Figures 33 and 34.

Figure 31: Rules generated with association with the Republican party



```
lhs                                                    rhs                          support confidence lift
{x1AvgDays_Disc=high}                               => {politicalParty=Republican} 0.04    1          1.851852
{x2AvgDays_Disc=high}                               => {politicalParty=Republican} 0.04    1          1.851852
{x3NumPop_Disc=high}                                => {politicalParty=Republican} 0.14    1          1.851852
{x3NumPop_Disc=high,x4NumTot_Disc=high}             => {politicalParty=Republican} 0.02    1          1.851852
{x1AvgDays_Disc=high,x2AvgDays_Disc=high}           => {politicalParty=Republican} 0.04    1          1.851852
{x1AvgDays_Disc=high,x3NumPop_Disc=high}            => {politicalParty=Republican} 0.04    1          1.851852
{x1AvgDays_Disc=high,x4NumTot_Disc=medium}          => {politicalParty=Republican} 0.04    1          1.851852
{x2AvgDays_Disc=high,x3NumPop_Disc=high}            => {politicalParty=Republican} 0.04    1          1.851852
{x2AvgDays_Disc=high,x4NumTot_Disc=medium}          => {politicalParty=Republican} 0.04    1          1.851852
{x1AvgDays_Disc=medium,x3NumPop_Disc=high}          => {politicalParty=Republican} 0.04    1          1.851852
```

Figure 32: Rules generated with association with the Democratic party



```
lhs                                                                               rhs                          support confidence lift
{x3NumPop_Disc=medium,x4NumTot_Disc=high}                                      => {politicalParty=Democratic} 0.02    1          2.173913
{x2AvgDays_Disc=low,x3NumPop_Disc=medium,x4NumTot_Disc=high}                   => {politicalParty=Democratic} 0.02    1          2.173913
{x1AvgDays_Disc=low,x3NumPop_Disc=medium,x4NumTot_Disc=high}                   => {politicalParty=Democratic} 0.02    1          2.173913
{x1AvgDays_Disc=medium,x3NumPop_Disc=medium,x4NumTot_Disc=medium}              => {politicalParty=Democratic} 0.02    1          2.173913
{x2AvgDays_Disc=medium,x3NumPop_Disc=medium,x4NumTot_Disc=medium}              => {politicalParty=Democratic} 0.02    1          2.173913
{x1AvgDays_Disc=low,x2AvgDays_Disc=low,x3NumPop_Disc=medium,x4NumTot_Disc=high}          => {politicalParty=Democratic} 0.02    1          2.173913
{x1AvgDays_Disc=medium,x2AvgDays_Disc=medium,x3NumPop_Disc=medium,x4NumTot_Disc=medium} => {politicalParty=Democratic} 0.02    1          2.173913
```

Figure 33: Network Diagram Visualizing Associations Among the Republican Party and Health Indicators

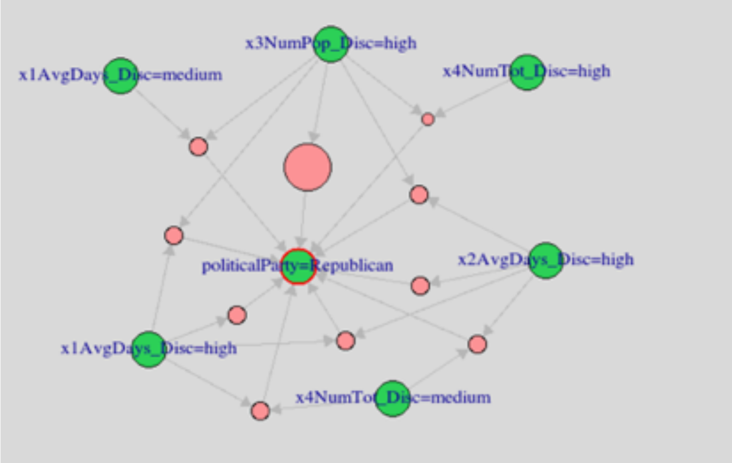
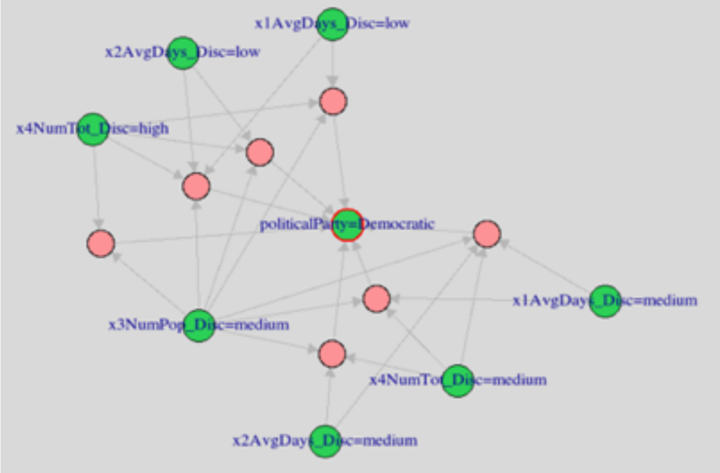


Figure 34: Network Diagram Visualizing Associations Among the Democratic Party and Health Indicators



Model 4b: Decision Trees

Decision trees and confusion matrices are created ten times each with different training and testing configurations, as previously described. A visual of the decision tree with the highest accuracy is seen in Figure 35. Each decision tree indicates the likelihood a political party is classified. The criteria listed at each decision point indicates the frequency of the political party indicated. The percentage in the bottom nodes of each decision tree represents the percentage of records labeled with the political party by following the preceding nodes. All ten confusion matrices can be found in Figure 36, also indicating the percent accuracy of each model. Averaged over the results of the ten models, there is a 58% accuracy rate of prediction.

Figure 35: Decision Tree Model #5

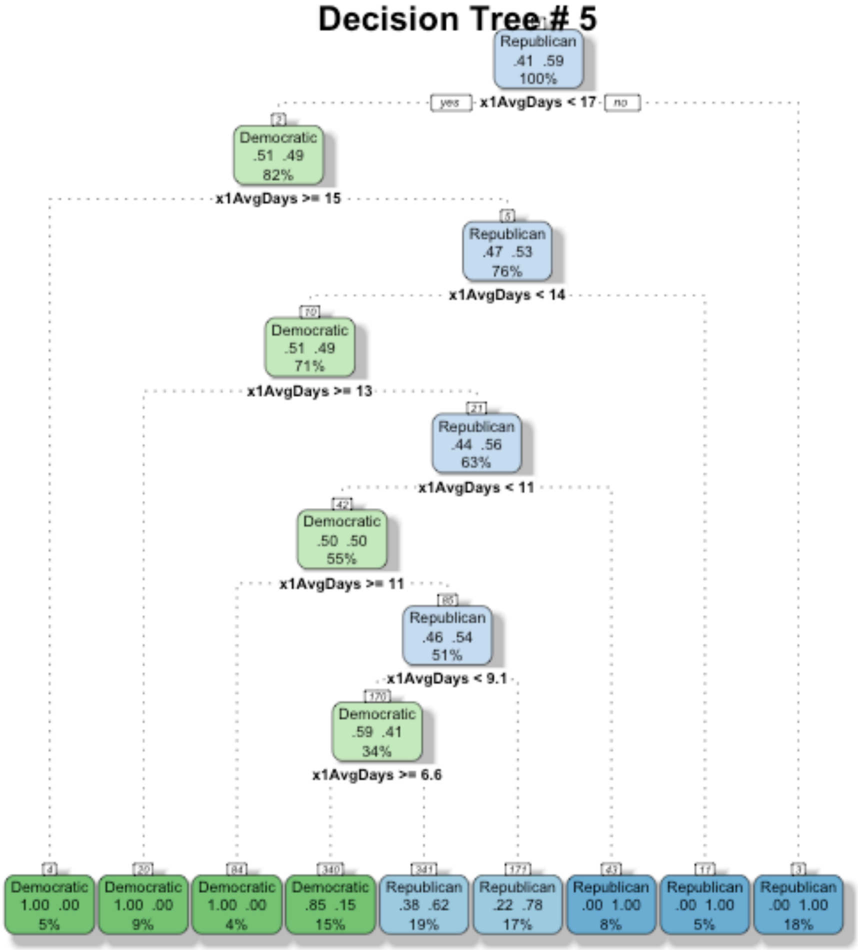
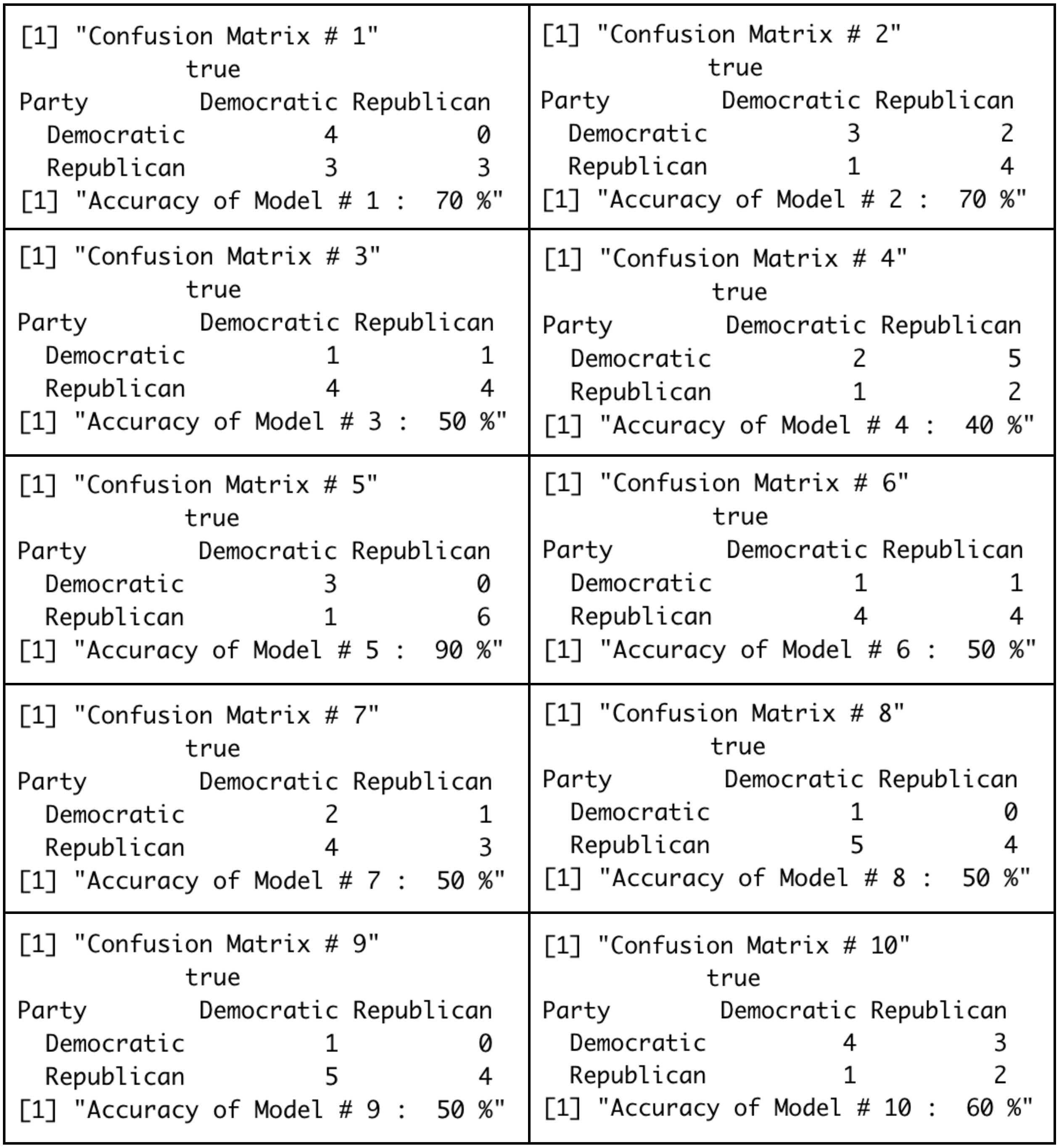


Figure 36: Confusion Matrices for Decision Trees for Political Party Classification



```
[1] "Confusion Matrix # 1"
            true
Party         Democratic Republican
  Democratic           4          0
  Republican           3          3
[1] "Accuracy of Model # 1 :  70 %"
```

```
[1] "Confusion Matrix # 2"
            true
Party         Democratic Republican
  Democratic           3          2
  Republican           1          4
[1] "Accuracy of Model # 2 :  70 %"
```

```
[1] "Confusion Matrix # 3"
            true
Party         Democratic Republican
  Democratic           1          1
  Republican           4          4
[1] "Accuracy of Model # 3 :  50 %"
```

```
[1] "Confusion Matrix # 4"
            true
Party         Democratic Republican
  Democratic           2          5
  Republican           1          2
[1] "Accuracy of Model # 4 :  40 %"
```

```
[1] "Confusion Matrix # 5"
            true
Party         Democratic Republican
  Democratic           3          0
  Republican           1          6
[1] "Accuracy of Model # 5 :  90 %"
```

```
[1] "Confusion Matrix # 6"
            true
Party         Democratic Republican
  Democratic           1          1
  Republican           4          4
[1] "Accuracy of Model # 6 :  50 %"
```

```
[1] "Confusion Matrix # 7"
            true
Party         Democratic Republican
  Democratic           2          1
  Republican           4          3
[1] "Accuracy of Model # 7 :  50 %"
```

```
[1] "Confusion Matrix # 8"
            true
Party         Democratic Republican
  Democratic           1          0
  Republican           5          4
[1] "Accuracy of Model # 8 :  50 %"
```

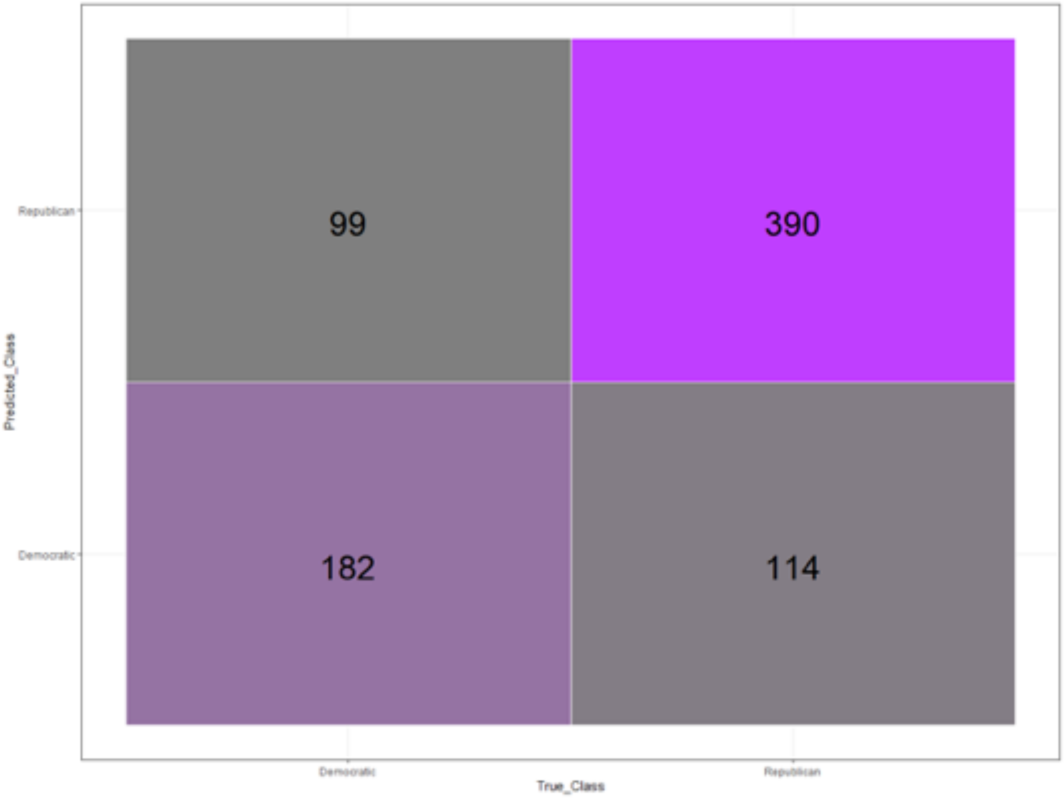
```
[1] "Confusion Matrix # 9"
            true
Party         Democratic Republican
  Democratic           1          0
  Republican           5          4
[1] "Accuracy of Model # 9 :  50 %"
```

```
[1] "Confusion Matrix # 10"
            true
Party         Democratic Republican
  Democratic           4          3
  Republican           1          2
[1] "Accuracy of Model # 10 :  60 %"
```

Model 4c: Random Forest

Using the random forest model, political party affiliation is predicted from health measures of 3,140 counties across the United States, as seen in Figure 37. The selected health indicators predicted political classification with 72.86% accuracy.

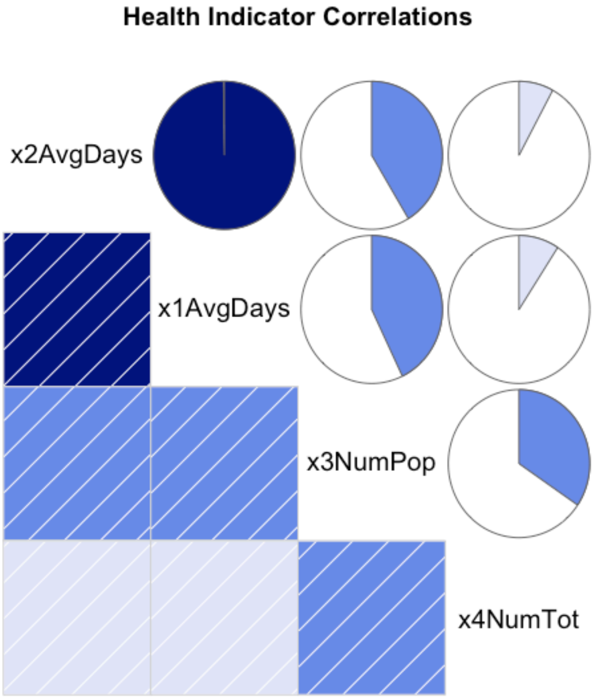
Figure 37: Confusion Matrix for Random Forest Model, Predicting Political Party



Model 4d: Naive Bayes

Correlation is found between the poor physical health days and poor mental health days indicator variables, as seen in Figure 38. For this reason, the naive bayes model is attempted twice. First with the poor mental health days, unemployed, and uninsured variables, and then a second time with poor physical health days, unemployed, and uninsured variables. Each attempt yields zero predictions. This is most likely due to the fact that this data is aggregated at the state level, and there is not enough data to properly train the model.

Figure 38: Correlations Among Health Indicator Variables



Conclusions

A broad exploration on the relationship between regional indicators and social determinants of health and the emotional sentiments expressed by elected leaders offers a unique insight into the intersection of politics and personal well-being. Through a diverse set of analytical questions and inquiries, there is a clear indication that political party affiliations, as represented by the emotions, words, and sentiments of state Governors, can be recognized by certain measures of population health and well-being. There is, however, no substantiating evidence that specific emotions and sentiments when expressed by elected leaders can accurately predict our general health and welfare. Of note is the fact that political sentiments, which are traditionally positive in nature, are an assembly of selected words and themes that can be motivating or demotivating to state constituents. In this manner, what our leaders say may not be predictive of our health but certainly influence our present and future physical and mental well-being.

While emotions and sentiments expressed by political leaders are poor predictors of population health, select indicators of health can be used to classify political parties. Unsurprisingly, correlation is found between the average number of poor mental health days and average number of poor physical health days per month. Association is found between each of these health indicators and political parties. As seen in the maps in Figures 39 and 40, citizens in republican states exhibit poorer physical and mental health days per month on average. In contrast, citizens in democratic states overall have lower numbers of poor mental and physical health days.

Figure 39: US Map showing Poor Physical Health Days and Political Party by State

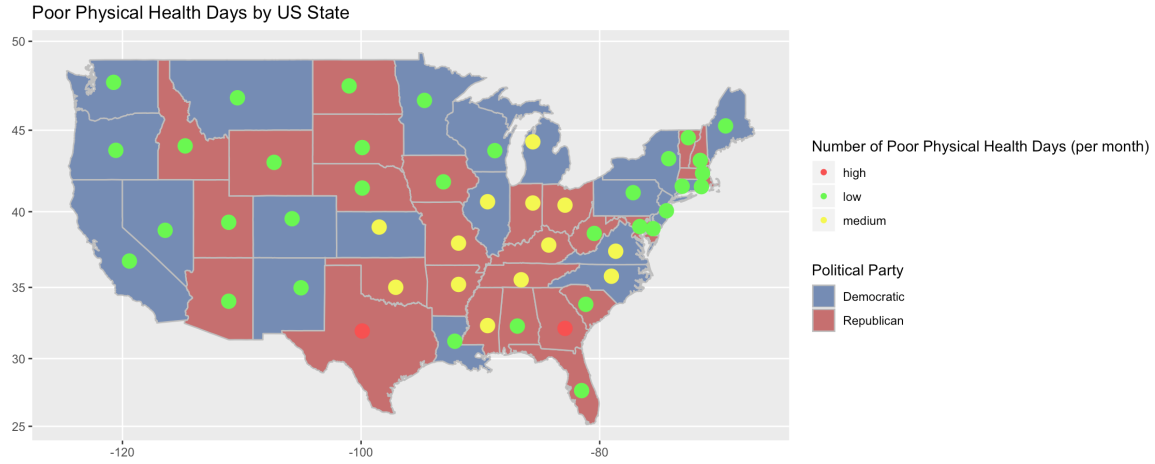
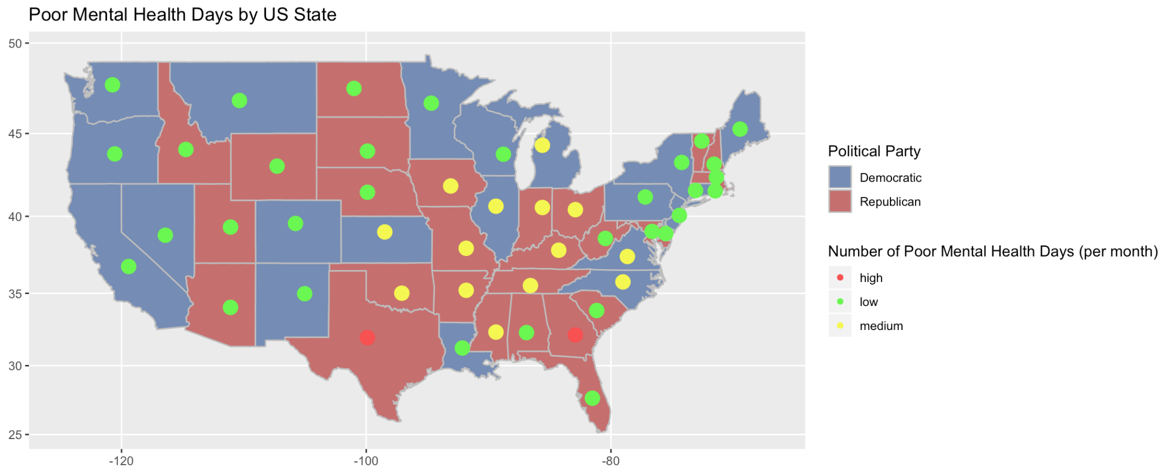

Poor Physical Health Days by US State
50
45
40
35
30
25
-120
-100
-80
Number of Poor Physical Health Days (per month)
● high
● low
● medium
Political Party
🟦 Democratic
🟥 Republican

Figure 40: US Map showing Poor Mental Health Days and Political Party by State


Poor Mental Health Days by US State
50
45
40
35
30
25
-120
-100
-80
Political Party
Democratic
Republican
Number of Poor Mental Health Days (per month)
high
low
medium

It is argued that despite the seemingly constant discussion of health in politics, there is actually an absence of mainstream discussion around the impact of politics on people's health.[7] Rudolf Virchow, who authored many public health publications in the mid 1800s said, "Medicine is a social science, and politics nothing but medicine at a larger scale".[8] While the findings of this analysis did not yield entirely anticipated results, it does add to the greater examination of politics and its influence on public health.

[7] C Bambra et al: *Towards a politics of health* https://doi.org/10.1093/heapro/dah608

[8] J P Mackenbach: *Politics is nothing but medicine at a larger scale: reflections on public health's biggest idea* https://jech.bmj.com/content/63/3/181?ijkey=9f8cc167dc7a154692df57257bcba5e0b6a81dca&keytype2=tf_ipsecsha#ref-9