Example Data Report: State-Level Newspaper Coverage of Health Care Reform, August 2009

Jaime E. Settle[*]

August 17, 2014

Abstract

This document is an example data report designed to show you some of the possibilities for the kinds of tables and figures that are useful for describing and exploring a dataset. Ultimately, you want to move toward inference, or assessing the relationship between two or more variables, by asserting and testing clear hypotheses. The strongest research in the social sciences is designed and analyzed in a way to support *causal* inference, or understanding a cause and effect relationship between two or more variables. Often, however, before this inference can be designed or assessed, it is critical to simply "get a handle" on the nuances of the data you will be using. Descriptive statistics and a thorough summary of a dataset are important first steps in any research project.

The dataset used in this report is the Obamacare Newspaper Dataset, generated by SNaPP Lab research assistants Joanna Borman, Will Evans and Gabe Manion during the 2013-2014 school year. This report will summarize and describe patterns in the article-level dataset. Future research will explore this dataset, as well as other datasets that can be derived from it or matched to it, such as state-level or newspaper-level datasets.

[*] jsettle@wm.edu

Description of the Data Generation Process

The data were generated by downloading articles from the Access World News database.[1] Research assistants followed a defined search procedure to return all articles in a state that were topically related to health care published in a state-level newspaper during August 2009. The search parameters were designed to maximize article recall at the expense of precision, aligning with the procedures outlined in Stryker et al. (2006).

First Field
("ACA" OR "affordable care act" OR "patient protection and affordable care act" OR "PPACA" OR "obama care" OR "obamacare" OR "obama-care")
OR **Second Field**
("health care reform" OR "healthcare reform") AND ("nation*")
OR **Third Field**
("obama" OR "barack obama" OR "obama administration" OR "president" OR "president obama" OR "democrat*" OR "clinton" OR "kennedy" OR "pelosi" OR "reid" OR "republican*" OR "mccain" OR "boehner" OR "congress" OR "house" OR "senate" OR "111th" OR "111th Congress") AND ("healthcare reform" OR "health care reform")
OR **Fourth Field**
("abortion" OR"access" OR "benefit*" OR "coverage" OR "quality" OR "wellness" OR "uninsured" OR "underinsured" OR "cost" OR "expensive" OR "lifetime benefit maximum" OR "employer" OR "fee for service" OR "fee-for-service" OR "long term care" OR "long-term care" OR "managed care" OR "pay for performance" OR "pay-for-performance" OR "payment bundl*" OR "premium" OR "public" OR '"public option" OR "single payer" OR "single-payer" OR "universal coverage" OR "townhall" OR "town hall" "socialized medicine" OR "ration*" OR "death panel*" OR "communis*" OR "sociali*") AND ("healthcare reform" OR "health care reform")

Table 1: Terms used to generate newspaper article database.

The exact process for downloading and sending the articles is described elsewhere, but the process resulted in 25,509 articles for the 50 states and Washington, D.C. These articles were then processed using Python and R into a single dataset, where each row is an article and each column contains meta information about the article, such as the headline, date, etc. A separate "document by word" matrix was created where each column in the dataset represents one of the [XXXXX] words that appeared in the entire corpus of the newspaper articles. The *OCnatlnews* data file described in this report has selected key words of interest appended to the meta datafile:

[1] Thanks to Jake Lewitz for his work designing the search term parameters and Chris Coelho for pilot testing the article downloading process.

Variables

Variable Summary

The full codebook of the dataset can be found elsewhere, but an abbreviated version is provided below in Table 2 and Table 7:

Variable Name	Variable Description
articleID	The 4-5 digit alphanumeric identification showing the state and document number within the state data. Example: AK19
doc	The document number within the state data. Example: 19 Note: this is NOT a unique variable
totalwords	The total number of words in the article, including the headline, byline, and other information
Statename	The state in which the newspaper was published.
Headline	The headline of the article
DateLine	The unformatted date of the article
Year	All values for this variable read 2009
Month	All values for this variable read August
Day	The date in August 2009 on which the article was published
NewsService	The name of the newspaper in which the article was published
Byline	The author of the article
Section	The section of the newspaper in which the article appeared
statedoc	A second alphanumeric identification variable including the full state name and document number. Example: Alaska_19
id	The unique identification number for the article in the dataset
article_id	The same unique identification number for the article (used to merge together the meta data and word data
Columns with key words	The number of times a key word appears in the article

Table 2: Variable name and brief descriptions.

Missing Data

Variable Name	Unique Values	Missing Values
articleID	25509	0
doc	2374	0
totalwords	2095	0
Statename	51	0
Headline	19341	1
DateLine	1403	1040
Year	1	0
Month	1	0
Day	32	2
NewsService	2163	1
Byline	7313	10213
Section	3869	3250
statedoc	25509	0
id	25498	12
article_id	25473	37

Table 3: Missingness in key variables in the *OCnatlnews* dataset.

The number of unique values and missingness for a variable is an important first step in understanding your dataset Table 3.

In most datasets, you want to have an *unique identifying variable* for each unit of analysis in the dataset so that you can match in other variables at the same unit of analysis. Assessing the unique values for a variable is a good verification to make sure that every row (every case) in your dataset is unique. In the OCnatlnews data, the *articleID* variable is the only variable for which we have complete data with a completely unique identifier. The other id variables have unique values for all cases that are not missing. The *id* variable is the unique document code that originated in the meta data file. There were 12 articles that appeared in the meta data but did not appear in the document-by-word matrix, shown in Table 4. The 10 articles from Georgia are the result of a processing error. The article from Arizona did not have any text in the body of the article, and the article in Michigan is a processing artifact and is not actually an article.

articleID	doc	Statename	statedoc	id	article_id
AZ106	115	Arizona	Arizona_115		
GA352	368	Georgia	Georgia_368		
GA353	369	Georgia	Georgia_369		
GA354	370	Georgia	Georgia_370		
GA355	371	Georgia	Georgia_371		
GA356	372	Georgia	Georgia_372		
GA357	373	Georgia	Georgia_373		
GA358	374	Georgia	Georgia_374		
GA359	375	Georgia	Georgia_375		
GA360	376	Georgia	Georgia_376		
GA361	377	Georgia	Georgia_377		
MI453	470	Michigan	Michigan_470		

Table 4: Missing data from processing errors.

There is also some missing data originating from problems in the creation of the document-by-word matrix. In addition to the 12 missing articles described above, there are 25 articles in Spanish in the dataset (Table 5). These articles were not processed in the document-by-term matrix, and thus were not assigned the *article_id* variable although they do appear in the meta data and thus were assigned *articleID* and *id* variable values.

The other variables with missing values can be explained by processing artifacts. The *dateline* variable is very messy, but was pre-processed in the meta data so that there is no missing data for the year or month variables which were derived from the dateline. The only two articles with missing *day* values are the single articles in Michigan and Arizona described above. The only article with missing *NewsService* and *Headline* is the Michigan article.

The *Byline* and *Section* variables were reported inconsistently in the database, and thus it was difficult for the automated textual pre-processing to correctly identify them. Any analysis using those variables would need to be cleaned before use.

articleID	doc	Statename	DateLine	statedoc	id	article_id
AR139	147	Arkansas	14 de agosto del 2009	Arkansas_147	782	
AR140	148	Arkansas	14 de agosto del 2009	Arkansas_148	783	
AR58	66	Arkansas	7 de agosto del 2009	Arkansas_66	1133	
AZ106	115	Arizona		Arizona_115		
CA1494	1507	California	18 de agosto del 2009	California_1507	1735	
CA1452	1465	California	8 de agosto del 2009	California_1465	1688	
CA1453	1466	California	27 de agosto del 2009	California_1466	1689	
CA1583	1596	California	29 de agosto del 2009	California_1596	1833	
CA1585	1598	California	18 de agosto del 2009	California_1598	1835	
DC237	237	DC	20 de agosto del 2009	DC_237	5053	
DC247	247	DC	6 de agosto del 2009	DC_247	5064	
DC248	249	DC	13 de agosto del 2009	DC_249	5066	
DC249	250	DC	13 de agosto del 2009	DC_250	5068	
DE05	5	Delaware	14 de agosto del 2009	Delaware_5	5226	
DE49	49	Delaware	28 de agosto del 2009	Delaware_49	5225	
DE50	50	Delaware	28 de agosto del 2009	Delaware_50	5227	
DE51	51	Delaware	28 de agosto del 2009	Delaware_51	5228	
FL442	469	Florida	22 de agosto del 2009	Florida_469	6060	
FL443	470	Florida	22 de agosto del 2009	Florida_470	6062	
GA352	368	Georgia	August 15 2009	Georgia_368		
GA353	369	Georgia	August 15 2009	Georgia_369		
GA354	370	Georgia	August 15 2009	Georgia_370		
GA355	371	Georgia	August 15 2009	Georgia_371		
GA356	372	Georgia	August 15 2009Edition: HOME	Georgia_372		
GA357	373	Georgia	August 15 2009	Georgia_373		
GA358	374	Georgia	August 15 2009	Georgia_374		
GA359	375	Georgia	Georgian (Carrollton GA)	Georgia_375		
GA360	376	Georgia	Herald The (Newnan GA)	Georgia_376		
GA361	377	Georgia	August 15 2009	Georgia_377		
IL903	966	Illinois	30 de agosto del 2009	Illinois_966	9672	
IL216	218	Illinois	23 de agosto del 2009	Illinois_218	8842	
IL1439	1529	Illinois	16 de agosto del 2009	Illinois_1529	8393	
IL904	967	Illinois	30 de agosto del 2009	Illinois_967	9673	
MI453	470	Michigan		Michigan_470		
NV62	66	Nevada	14 de agosto del 2009	Nevada_66	15938	
NV63	67	Nevada	14 de agosto del 2009	Nevada_67	15939	
NV101	106	Nevada	28 de agosto del 2009	Nevada_106	15752	

Table 5: Missing data resulting from Spanish language articles.

Data Exploration

State Information

State	Total Words	Article Count	**State**	Total Words	Article Count
Alaska	53,006	85	**Montana**	131,461	225
Alabama	200,815	322	**North Carolina**	595,494	1,046
Arkansas	307,716	412	**North Dakota**	121,338	187
Arizona	177,375	299	**Nebraska**	182,936	286
California	1,632,022	2,358	**New Hampshire**	138,656	212
Colorado	403,971	666	**New Jersey**	367,317	511
Connecticut	402,258	560	**New Mexico**	143,288	205
DC	211,525	271	**Nevada**	131,808	215
Delaware	63,128	105	**NewYork**	534,533	839
Florida	853,926	1,163	**Ohio**	318,885	494
Georgia	423,488	618	**Oklahoma**	186,428	305
Hawaii	101,734	153	**Oregon**	274,392	403
Iowa	299,112	531	**Pennsylvania**	794,550	1,276
Idaho	178,426	281	**Rhode Island**	108,103	151
Illinois	1,211,524	1,815	**South Carolina**	414,590	611
Indiana	355,393	547	**South Dakota**	61,461	102
Kansas	186,705	299	**Tennessee**	315,628	455
Kentucky	185,814	288	**Texas**	828,643	1186
Louisiana	143,078	237	**Utah**	178,505	315
Massachusetts	958,257	1326	**Virginia**	355,730	502
Maryland	243,663	347	**Vermont**	197,138	265
Maine	52,393	104	**Washington**	378,714	602
Michigan	349,690	588	**Wisconsin**	208,732	292
Minnesota	295,705	407	**West Virginia**	135,566	190
Missouri	380,857	594	**Wyoming**	44,130	69
Mississippi	133,163	189			

Table 6: The volume of newspaper coverage by state.

One of the extensions of this project is to explore the relationship between state-level variables and newspaper coverage. The Obamacare research assistants have identified a number of state-level indicators they would like to evaluate for possible inference in explaining the variation between the states in the rate and type of coverage of health care reform. These indicators can largely be broken up into three categories:

- **Policy indicators:** These variables will factor in how states responded to the ACA, as well as measures of their pre-ACA health care system. Examples of these variables include: a. Whether states signed onto lawsuits against the ACA; b. State decisions to expand Medicaid, and to what income level; c. State insurance exchanges, including data on sign-ups after websites went live; d. Medicare and Medicaid spending (total and per-capita) before and after ACA

- **Political indicators:** These variables will assess the ideology of the elected officials and the public in the states, as well as aggregate measures of public opinion in each state during the summer and fall of 2009.

- **Health/Demographic variables:** These measures focus on various characteristics of the population of states to see if these relate to public opinion on the ACA, including: a. Mortality and infant-mortality rates; b. State-specific life expectancies; c. State health spending as a percent of GDP, including public and private; d. Rates of uninsured in each state, pre and post-ACA.

These are just a few examples of potential variables the team may investigate at the state level. Based on a thorough literature review of the predictors of media coverage and the consequences of media coverage on public opinion, the team will develop and test hypotheses to explain the causes and effects of state-level media coverage in this time period.

Temporal Exploration

Another way to explore the data is to look at patterns over time. While there is little analytic traction to be gained from this kind of data visualization, it is a good way to check for patterns you expect in the data. For example, looking at Figure 1, it appears that there is a difference in the number of articles that are posted during the week versus the number that are posted during the weekend. We can visualize that difference in Figure 2. This is a good "face validity" check, in the sense that there are good reasons to think that newspapers cover more policy news during the week than they do on the weekend

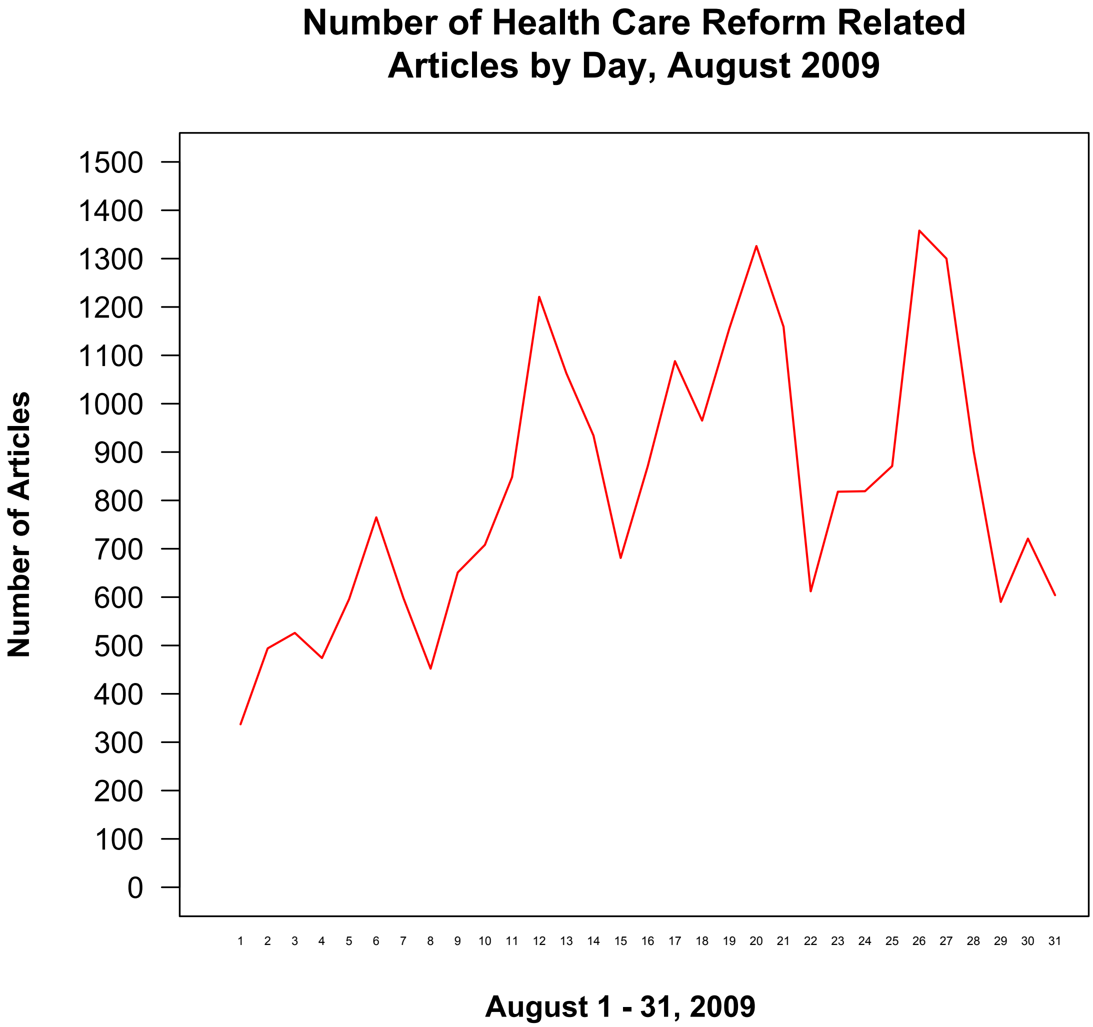

Figure 1: The number of articles published about health care reform each day in August 2009

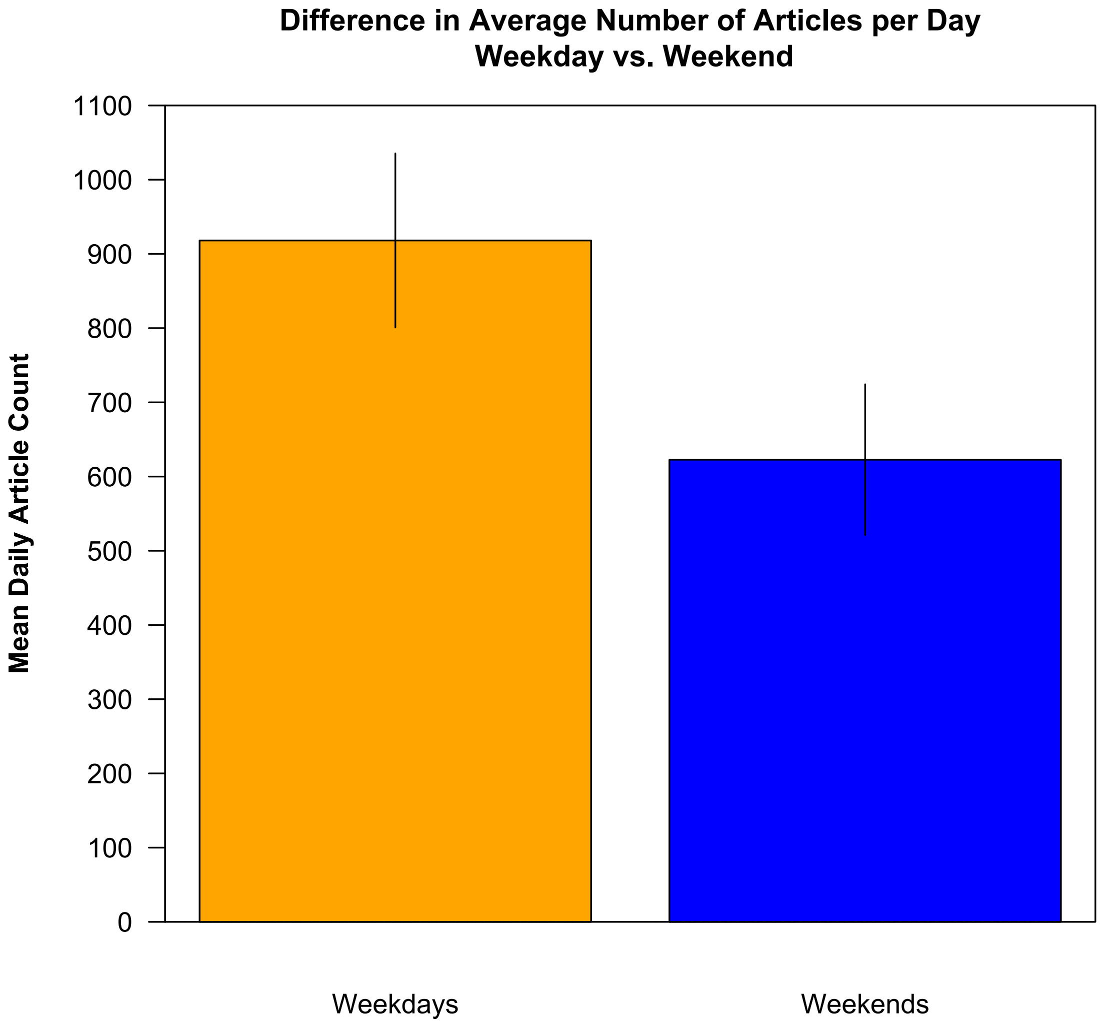

Figure 2: Newspapers published more articles about health care reform on the weekdays compared to weekend days.

Framing

The core interest in the project is to identifying differences in the way that the debate over health care reform was portrayed. For this initial exploration in the data, we take a very simple and straightforward approach: identifying key words that are conceptually related to different frames already discussed in the literature on Obamacare, depicted in Table 7.

Frame	Key Words
Neutral	"health" , "healthcare" , "healthcare_reform" , "healthcare_reforms" "healthcarereform" , "healthcares" , "healthinsurance" , "healthreform" "reform" , "obama" , "obamas"
Nazi	"nazi" , "nazis" , "nazism" , "hitler" , "swastika"
Sociali*	"socialism" , "socialist" , "socialistic" , "socialists" , "socialization" , "socialize" "socialized" , "socialized_medicine" , "socializes" , "socializing"
Ration*	"ration" , "rationed" , "rationed_care" , "rationing" , "rations"
Death Panel	"death" AND ("panel" OR "panels"), "death_panels"
Universal	"universal" ,"public_option" , "single_payer"
Constitutional	"liberty" , "constitution"
Fiscal	"bankrupt"
Town Hall	"town" AND ("hall" OR "halls") , "town_hall" , "town_halls"

Table 7: Frames and key words in OCnatlnews dataset.

We visualize this data in Figure 3. Again, we cannot make any analytic inference from this data. However, it appears that coverage about town halls dominated the news most days during the month. One further way to evaluate this would be to annotate the plot showing the timing the key events, such as the date that Sarah Palin first used "death panel" on her Facebook newsfeed.

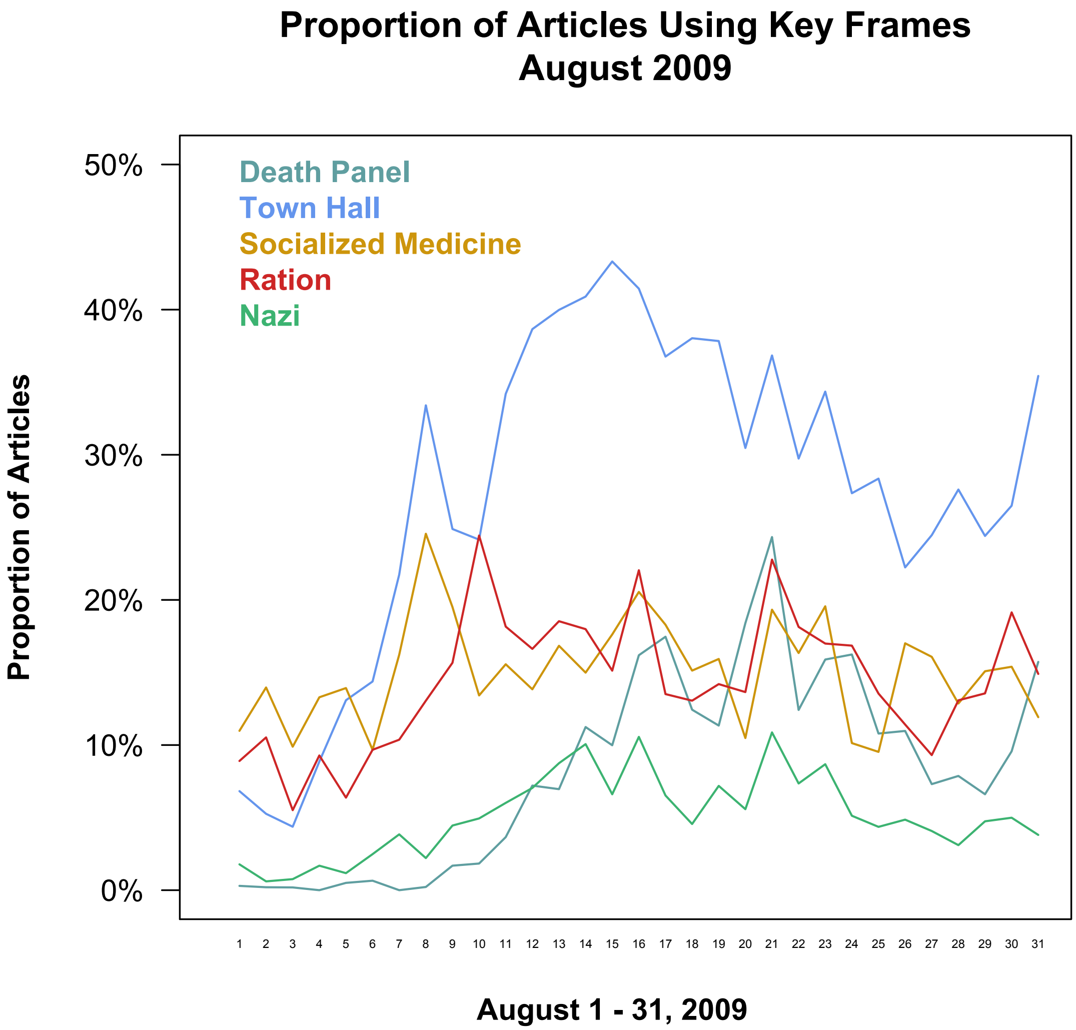

Figure 3: The number of articles each day that used one of the key words associated with a frame of interest

Variation in the Use of Frames?

Once we have identified the frames that were used in the debate, we hope to explore why some frames were selected more than others. Moving in this direction will require aggregating the data in some way, either at the level of the newspaper or the level of the state. Doing so will require certain coding decisions, and it is important to have a grasp on the distribution of the number of articles at each value of the variable to which you are aggregating.

For example, Figure 4 shows that most newspapers published very few articles during the month, but a handful of newspapers published over 100 articles. Any decision about aggregating the newspaper will require some baseline threshold for the number of articles that a newspaper must have published in order to meaningfully evaluate its use of frames. For the data in Figure 4, we examine all newspapers that published at least five articles during the month. The histograms show the variation in the proportion of articles published at the newspaper-level (Figure 5) or state-level (Figure 6) that employ the keywords for each frame.

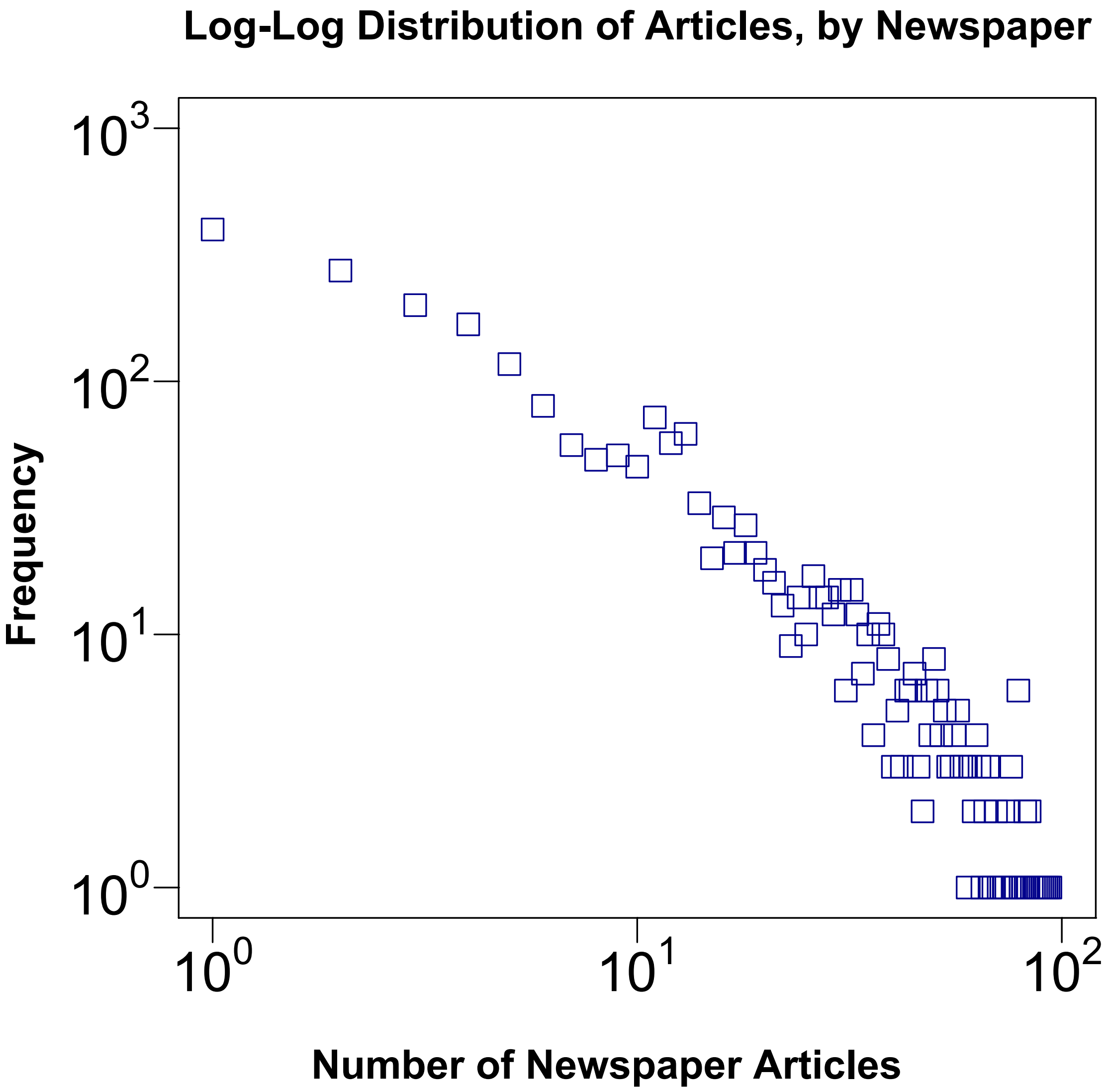

Figure 4: The log distribution of the number of articles published in each newspaper

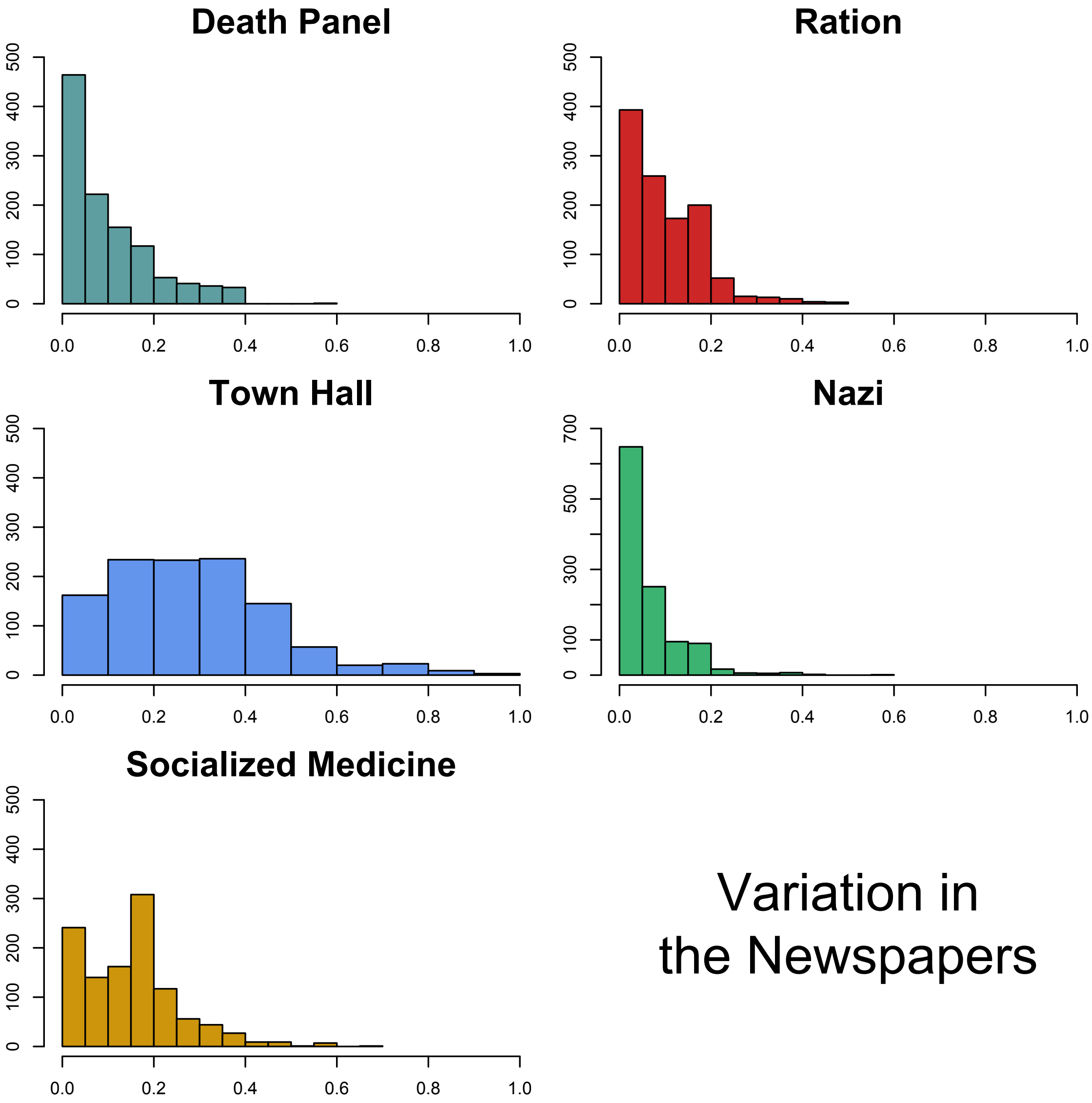

Figure 5: Variation in the use of frames by newspapers. The x-axis shows the proportion of all articles published by the newspapers that used one or more key words for each frame

The important takeaway point from these figures is that there is sufficient variation at both the state and newspaper levels to merit studying the use of frames as independent or dependent variables. A very cursory example of that analysis is shown here. Table 8 shows the correlations between three key measures of ideology in the states and the proportion of the articles published in the state that used a particular frame. None of the correlations are statistically significant from 0, however the correlation between the "Town Hall" frame and the ideology of the citizens in the state is close. That pattern is shown in Figure 7, and visualized in a different type of plot in **??**.

	Citizen Ideology		ADA		DW Nominate	
	Pearson's r	p-value	Pearson's r	p-value	Pearson's r	p-value
Death Panel	0.15	0.28	0.12	0.40	0.09	0.55
Town Hall	-0.22	0.12	-0.06	0.68	-0.09	0.54
Socialized Medicine	-0.02	0.91	-0.12	0.42	-0.12	0.41
Ration	0.07	0.63	0.06	0.69	0.05	0.75
Nazi	-0.07	0.62	-0.01	0.92	0.00	0.99

Table 8: Correlations between measures of state ideology and use of key frames.

While the lack of relationship may seem discouraging, it shouldn't be. There are many explanations for the lack of a direct correlation. For example, future research could explore whether conservative states were more likely to report on the town halls, controlling for the number of town hall meetings that occurred in the state.

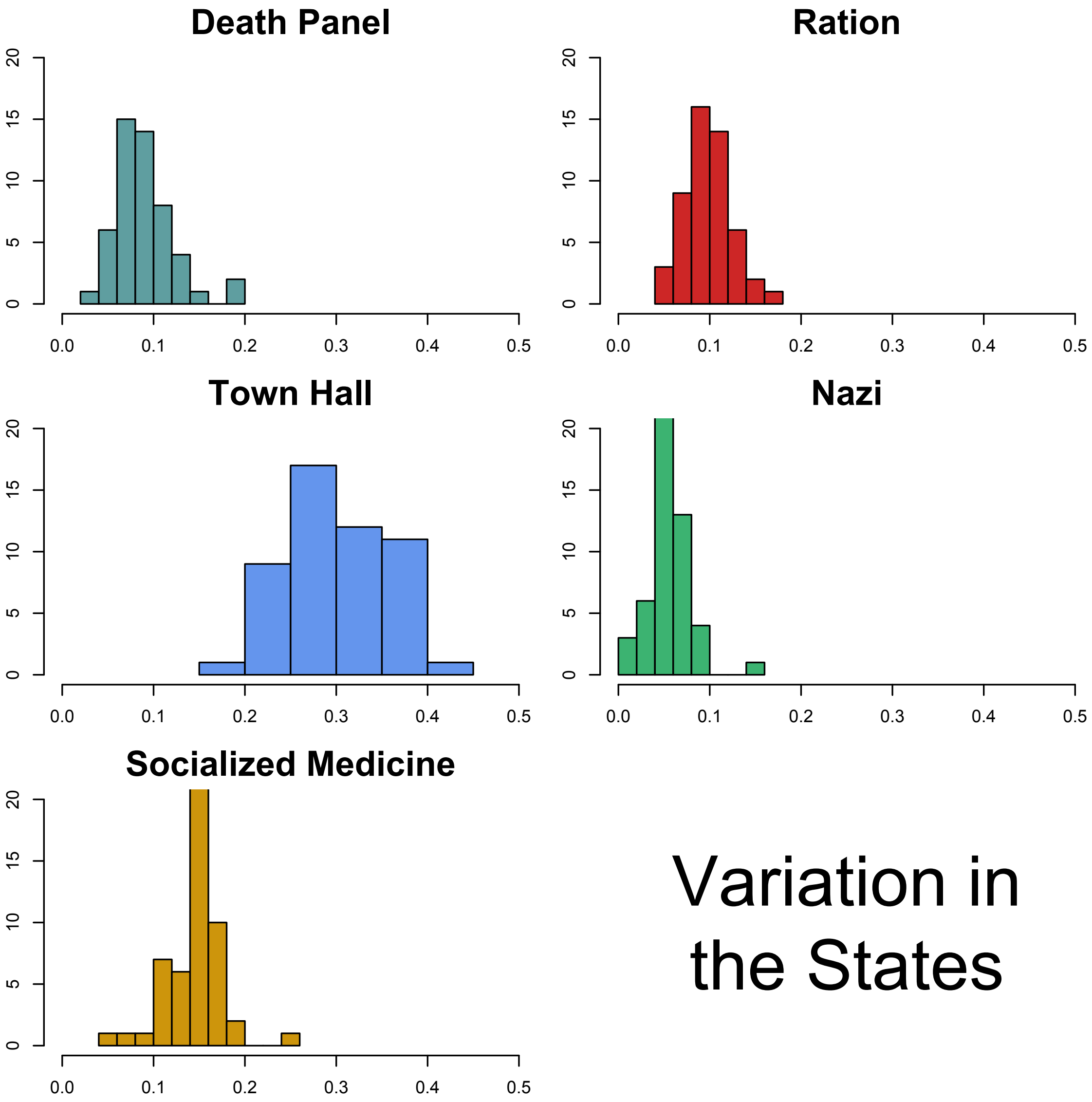

Figure 6: Variation in the use of frames by states. The x-axis shows the proportion of all articles published in each state that used one or more key words for each frame

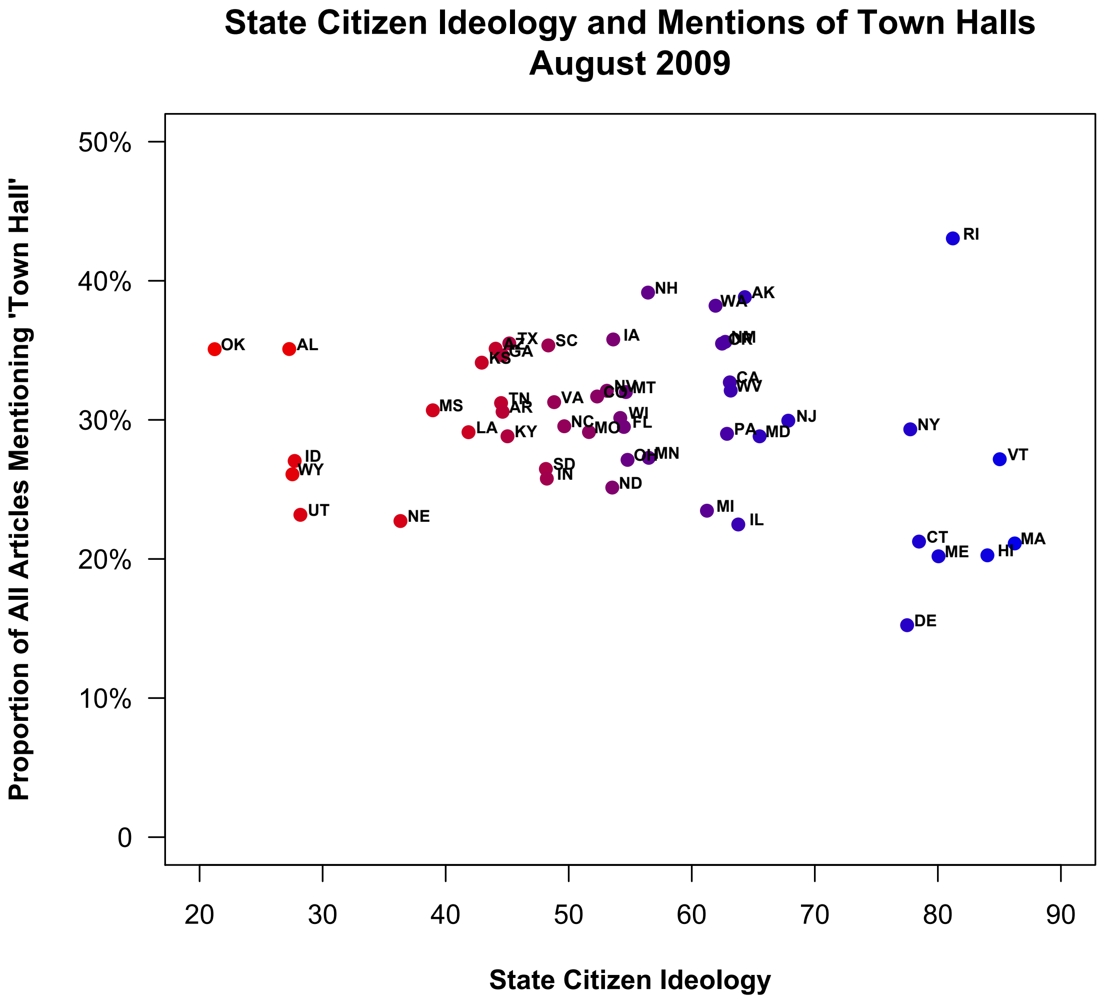

Figure 7: Correlation between state-level ideology and coverage of the town halls.

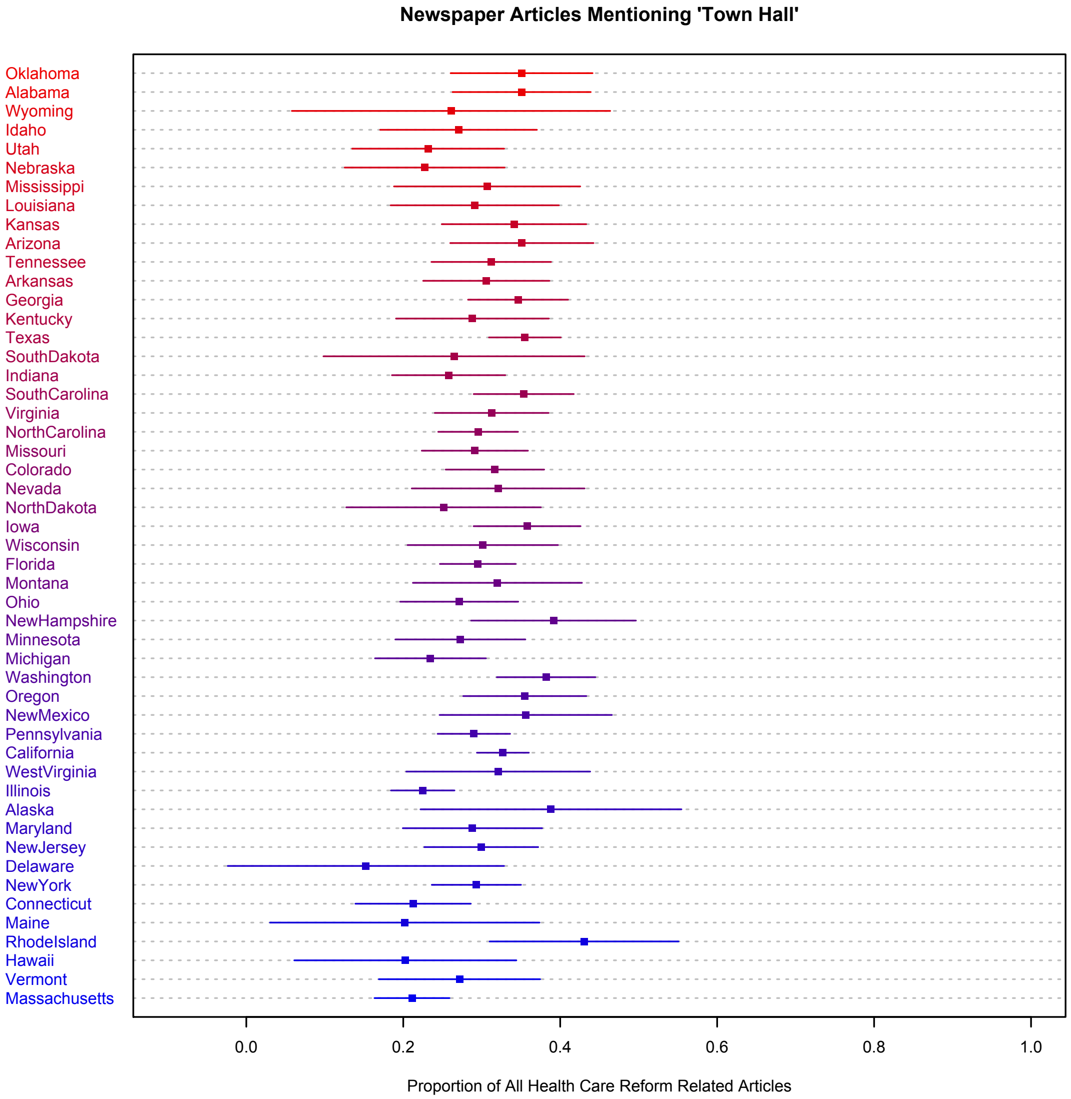

Figure 8: Correlation between state-level ideology and coverage of the town halls.

Conclusions

This example report should give you an idea of the scope and content of an example data report. The exact analysis you include in your report will depend on the nature of your data and the current stage of the project. However, hopefully this has inspired you to learn R and "get your hands dirty" in the data. Happy exploration!

Works Cited

Stryker, Jo Ellen, Ricardo Wray, Robert Hornik and Itzik Yanovitzky. 2006. "Validation of Database Search Terms for Content Analysis: The Case of Cancer News Coverage." *Journalism & Mass Communication Quarterly* 83(2): 413-430.