Kay Quiballo | MSDS 411 | 04.10.2023

MSDS 411 Assignment 1: Political Consulting through Primary Component Analysis

Abstract. *An executive summary of the research.*

Unsupervised learning methods such as principal component analysis and exploratory factor analysis were leveraged in the context of political campaigning. Using the 2019 Pew Research dataset on political opinion, we explore the different groupings and clusters that define the public's perspective. Our goal is to extract as much information as possibly from the dataset and make meaningful interpretations relating to political opinion and potential campaign points. These analyses emphasize the following polarizing topics: politicians, economics, religion, and racial discrimination. Depending on how clusters react, a campaign can take advantage of a target audience to bring in more votes.

Introduction. *Why are you conducting this research?*

As a political consultant agency in 2019, we were tasked with assisting politicians with their political campaigns in the upcoming 2020 election. We want to make sure that our stakeholders are leveraging points that they stand for that attract a certain audience or political party. If a political campaign has unfavorable points to a political party, it may drive away votes in an election. As a part of the data team in this political consultant agency, we are tasked with analyzing the political opinion survey conducted by Pew Research. With this dataset, we want to advise our stakeholders to leverage specific topics they support in a campaign in order to gain votes. At the same time, they may need to avoid emphasizing specific topics they support to avoid losing votes from their target demographic.

Literature review. *Who else has conducted research like this?*

In the Encyclopedia of Bioinformatics and Computational Biology, this academic journal references research conducted by Krippendorff that utilizes clustering for political campaigns. Since the 1960's clustering has been used in various research fields, and political science is not the exception. (Elsevier Science 2018). Krippendorff conducted multivariate analyses for human communication research by using clustering. Similarly, in our analysis, we will be using clustering to analyze political opinion.

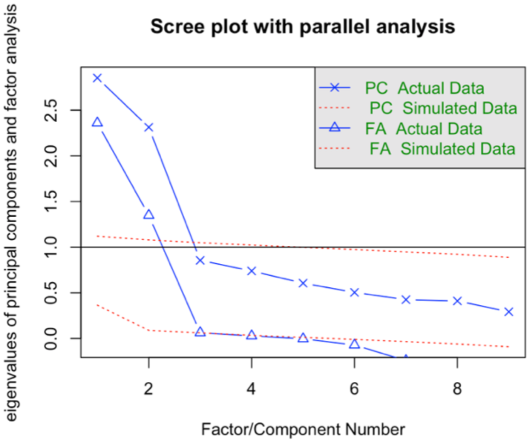
Methods. *How are you conducting the research?*

The Pew Research Dataset consists of 1503 phone interviews that were conducted in the United States in 2019. Phone interviews asked subjects their opinions on certain political topics that included policy, economics, and government. The subset of data used in this paper consists of 30 variables on political opinion that were asked to all participants. From these 30 complete variables, they were divided into binary responses based on the options selected to implement a common scale which resulted in 100 binary variables for the primary component analysis (PCA) and exploratory factor analysis (EFA).

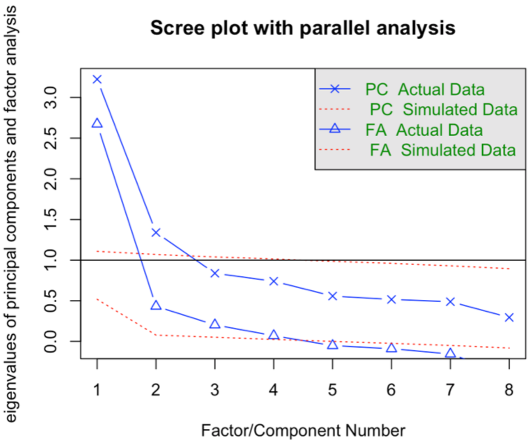
One of the response options for the political opinion variables is “Don’t Know/Refused to Answer” which is coded as a 9. There are benefits to including this data in the analyses like not potentially removing meaningful data. However, some research excludes these types of responses to draw to light insights from responses that gave valid answers. With this in mind, I decided to run the PCA and EFA with the full dataset and separately with a subset of variables that excludes the “9” option to see if any additional insight can be provided.

Results. *What did you learn from the research?*

Scree plot from *unfiltered* full dataset



Scree plot from *subsetted* dataset



From the Scree plots above, the full dataset for both the PCA and EFA should have 2 factors as indicated by the eigenvalues > 1. Similarly, the subset for the PCA should have 2 factors, but the EFA should only have 1 factor based on the eigenvalues.

Our goal in the PCA/EFA is to extract as much information as possible from the Pew Research dataset based on the linear relationships between political opinion variables. The more components, or groups of variables that we analyze, we gain marginally less explainability of variance in the dataset.

Using all variables from either the full or subsetted dataset, I first ran the PCA/EFA and looked at the loadings and eliminated variables with loadings less than 0.5. This is to identify variables that contribute more to explanation of variance in each component. The process was iterated until no more variables had loadings less than 0.5. Some key statistics that I pulled from each analysis includes cumulative variance (>0.90 is best for variance explanation in the dataset) and root mean square residual (<0.10 is best indication a good fit). I analyzed various numbers of principal components (2, 3 and 6) and principal axes (1, 2, and 3) with both the full and subsetted dataset. Although objectively there is no "best" option, each analysis points towards similar results.

Primary Component Analysis

Analysis	Dataset	Variables and Loadings	# Principal Components	Results

PCA	Includes "Don't Know/Refused"
	First 7 eigenvalues:
	3.22, 1.34, 0.84, 0.74, 0.56, 0.52, 0.49
		PC1 PC2
		q01_2 0.66 0.16
		q02_2 0.85 0.18
		q50d_2 -0.62 -0.26
		q50d_9 -0.16 0.56
		q61a_9 -0.27 0.80
		q61b_9 -0.27 0.76
		q61c_9 -0.24 0.77
		q68a_2 -0.65 -0.21
		q70_2 0.81 0.16
			2 principal components (**scree + eig suggested*)	Cum. Var = 0.57
				RMSR = 0.09
				Fit = 0.92

		PC1 PC2 PC3
		q01_2 0.66 0.16 0.26
		q02_2 0.85 0.18 0.02
		q50d_2 -0.62 -0.26 0.30
		q50d_9 -0.16 0.56 -0.74
		q61a_9 -0.27 0.80 0.03
		q61b_9 -0.27 0.76 0.26
		q61c_9 -0.24 0.77 0.28
		q68a_2 -0.65 -0.21 0.04
		q70_2 0.81 0.16 0.08
			3 principal components	Cum. Var = 0.67
				RMSR = 0.08
				Fit = 0.93

		PC1 PC2 PC3 PC4 PC5 PC6 PC7
		q01_2 0.66 0.16 0.26 0.52 0.28 -0.32 -0.12
		q02_2 0.85 0.18 0.02 -0.06 0.12 0.18 0.00
		q50d_2 -0.62 -0.26 0.30 -0.27 0.62 0.02 -0.02
		q50d_9 -0.16 0.56 -0.74 0.08 0.31 -0.03 -0.05
		q61a_9 -0.27 0.80 0.03 0.01 0.04 0.08 0.09
		q61b_9 -0.27 0.76 0.26 -0.01 -0.06 -0.18 0.40
		q61c_9 -0.24 0.77 0.28 -0.06 -0.11 0.17 -0.45
		q68a_2 -0.65 -0.21 0.04 0.61 0.01 0.37 0.09 q70_2
		0.81 0.16 0.08 -0.07 0.15 0.40 0.17
			6 principal components
			(*min PCA for 90% cum. var*)
				Cum. Var = 0.92
				RMSR = 0.04
				Fit = 0.98

	Omitted "Don't Know/Refused"
	First 6 eigenvalues:
	3.22, 1.34, 0.84, 0.74, 0.56, 0.52
		PC1 PC2
		q01_2 -0.65 0.10
		q02_2 -0.83 0.15
		q50a_2 0.59 -0.10
		q50d_2 0.68 -0.13
		q61a_2 0.41 0.75
		q61b_2 0.25 0.83
		q68a_2 0.66 -0.09
		q70_2 -0.79 0.14
			2 principal components (**scree + eig suggested*)	Cum. Var = 0.57
				RMSR = 0.10
				Fit = 0.92

		PC1 PC2 PC3
		q01_2 -0.65 0.10 0.15
		q02_2 -0.83 0.15 0.25
		q50a_2 0.59 -0.10 0.67
		q50d_2 0.68 -0.13 0.42
		q61a_2 0.41 0.75 -0.05
		q61b_2 0.25 0.83 0.06
		q68a_2 0.66 -0.09 -0.20
		q70_2 -0.79 0.14 0.29
			3 principal components	Cum. Var = 0.68
				RMSR = 0.09
				Fit = 0.92

		PC1 PC2 PC3 PC4 PC5 PC6
		q01_2 -0.65 0.10 0.15 0.62 0.23 -0.06
		q02_2 -0.83 0.15 0.25 -0.01 -0.05 0.06
		q50a_2 0.59 -0.10 0.67 0.08 -0.38 0.13
		q50d_2 0.68 -0.13 0.42 -0.06 0.54 -0.14
		q61a_2 0.41 0.75 -0.05 0.04 0.16 0.50
		q61b_2 0.25 0.83 0.06 -0.05 -0.13 -0.47
		q68a_2 0.66 -0.09 -0.20 0.58 -0.15 -0.04
		q70_2 -0.79 0.14 0.29 0.00 -0.04 0.04
			6 principal components
			(*min PCA for 90% cum. var*)
				Cum. Var = 0.90
				RMSR = 0.06
				Fit = 0.97

In the PCA, even though the Scree plot suggested only adapting 2 principal components, the cumulative variance shows that 2 components explains ~57% of the variability in the dataset, and 3 components explains ~67-68% of the variability in the dataset (for both the full and subsetted datasets). It isn't until 6 principal components are introduced that the cumulative variance is ideally above 90%. However, the eigenvalues suggest that more than 2 principal components are not needed. The RMSR (<0.10 ideal) and Fit (>0.90 ideal) of each PCA perform adequately.

Exploratory Factor Analysis

Analysis	**Dataset**	**Variables and Loadings**	**# Principal Axes**	**Results**

EFA	Includes "Don't Know/Refused"
	First 3 eigenvalues: 2.85, 2.31, 0.85
		PA1 PA2
		q01_2 0.55 0.14
		q02_2 0.86 0.20
		q50d_2 -0.51 -0.21
		q50d_9 -0.13 0.40
		q61a_9 -0.26 0.75
		q61b_9 -0.26 0.68
		q61c_9 -0.23 0.68
		q68a_2 -0.54 -0.18
		q70_2 0.77 0.17
			2 principal axes (**scree + eig suggested*)	Cum. Var = 0.47
				RMSR = 0.03
				Fit = 0.99

		PA1 PA2 PA3
		q01_2 0.56 -0.01 -0.08
		q02_2 0.88 -0.06 -0.01
		q50d_2 -0.56 0.00 -0.13
		q50d_9 0.00 0.29 0.70
		q61a_9 -0.03 0.73 0.26
		q61b_9 -0.04 0.75 0.09
		q61c_9 -0.01 0.75 0.07
		q68a_2 -0.57 -0.01 -0.02
		q70_2 0.78 -0.05 -0.06
			3 principal axes	Cum. Var = 0.52
				RMSR = 0.02
				Fit = 1.00

	Omitted "Don't Know/Refused"
	First 4 eigenvalues: 2.22, 1.67, 1.14, 0.58
		PA1
		q01_2 0.57
		q02_2 0.87
		q70_2 0.79
			1 principal axis (**scree + eig suggested*)	Cum. Var = 0.57
				RMSR = 0.00
				Fit = 1.00

		ML2 ML1 (varimax, ML)
		q01_2 0.58 -0.10
		q02_2 0.87 -0.01
		q25_2 -0.10 0.63
		q25_3 0.10 -0.99
		q25_4 0.03 0.54
		q70_2 0.79 0.01
			2 principal axes	Cum. Var = 0.57
				RMSR = 0.13
				Fit = 0.86

		ML3 ML1 ML2
		q01_2 0.55 -0.07 0.22
		q02_2 0.87 -0.06 0.09
		q25_2 0.03 0.28 -0.81
		q25_3 0.00 -0.88 0.48
		q25_4 0.00 0.87 0.49
		q70_2 0.78 -0.03 0.11
			3 principal axes	Cum. Var = 0.75
				RMSR = 0.00
				Fit = 1.00

In the EFA, even though the Scree plot suggested only adapting 1-2 principal axes, the cumulative variance shows that 2 axes explain ~47-57% of the variability in the dataset, and 3 axes explains ~52-75% of the variability in the dataset (for both the full and subsetted datasets). No amount of principal axes will yield a cumulative variance ideally above 90%. However, the eigenvalues suggest that more than 2 principal axes are not needed in the full dataset, and more than 1 principal axis is not needed for the subsetted dataset. The RMSR (<0.10 ideal) and Fit (>0.90 ideal) of each analysis perform adequately.

From both the PCA and EFA, the groupings of variables are quite similar no matter what dataset is used or what number of components are selected. An in-depth analysis of the variable groupings (as well as variable definitions) is covered in the conclusions section.

Conclusions. *So, what does it all mean?*

The following table defines the variables which were coded as 1 if selecting the corresponding option and 0 if otherwise. An F indicates that the variable had a significant loading (>0.50) for a principal component/axis for the full dataset, and S indicates likewise for the subsetted dataset. A negative sign in front of a letter indicates a negative loading. An individual who respond yes to the positive variables and otherwise to the negative variables behave most similarly according to each respective principal component/axis. Those variables, components, and combinations are what we draw insight from for the political consultant agency.

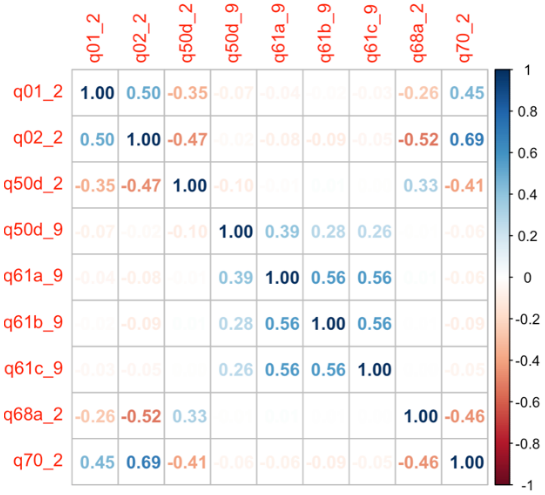
Variables for PCA/EFA

var	summary	description	PCA1	PCA2	EFA1	EFA2
q01_2	country satisfaction	All in all, are you satisfied or dissatisfied with the way things are going in this country today? (Dissatisfied)	-F, S		F, S	
q02_2	trump approval	Do you approve or disapprove of the way Donald Trump is handling his job as President? (Disapprove)	-F, S		F, S	
q25_2	trust government	How much of the time do you think you can trust the government in Washington to do what is right? (most of the time)				S
q25_3	trust government	How much of the time do you think you can trust the government in Washington to do what is right? (only some of the time)				-S
q25_4	trust government	How much of the time do you think you can trust the government in Washington to do what is right? (never)				S
q50a_2	business profits	Business corporations make too much profit [OR] Most corporations make a fair and reasonable amount of profit (Option #2)	S			
q50d_2	economy favors	The economic system in this country unfairly favors powerful interests [OR] The economic system in this country is generally fair to most Americans (Option #2)	F, -S		-F	
q50d_9	economy favors	The economic system in this country unfairly favors powerful interests [OR] The economic system in this country is generally fair to most Americans (Don't know/Refused)		F		
q61a_2	black discrimination	Please tell me how much discrimination there is against Blacks? Would you say there is a lot of discrimination, some, only a little, or none at all? (Some)		S		
q61a_9	black discrimination	Please tell me how much discrimination there is against Blacks? Would you say there is a lot of discrimination, some, only a little, or none at all? (Don't know/Refused)		F		F
q61b_2	hispanic discrimination	Please tell me how much discrimination there is against Hispanics? Would you say there is a lot of discrimination, some, only a little, or none at all? (Some)		S		
q61b_9	hispanic discrimination	Please tell me how much discrimination there is against Hispanics? Would you say there is a lot of discrimination, some, only a little, or none at all? (Don't know/Refused)		F		F

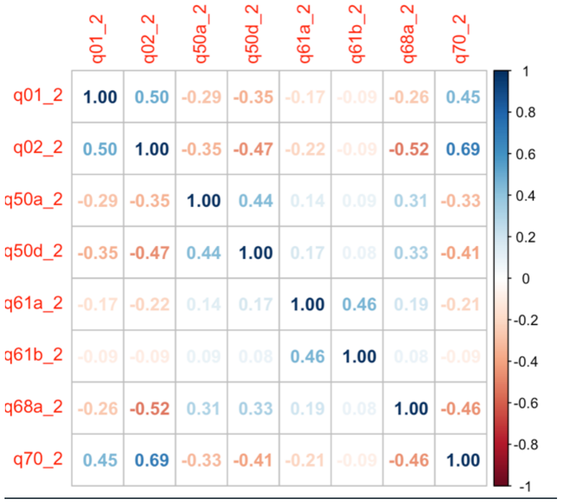
q61c_9	white discrimination	Please tell me how much discrimination there is against Whites? Would you say there is a lot of discrimination, some, only a little, or none at all? (Don't know/Refused)		F		F
q68a_2	democrats and religion	Do you feel The Democratic Party is generally FRIENDLY toward religion, NEUTRAL toward religion, or UNFRIENDLY toward religion. (Neutral toward religion)	F, -S		-F	
q70_2	trump tax law	Do you approve or disapprove of the tax law passed by Donald Trump and Congress in 2017? (Disapprove)	-F, S		F, S	

Note that the table only investigates the 2 most prominent principal components/axes for each PCA/EFA due to the Scree plots' suggestions.

Correlation Plot – Full Dataset



Correlation Plot – Subset Dataset



From the Correlation Plots above, we see that the full dataset has 2 prominent clusters, whereas the subsetted dataset has 2 clusters that are not as prominent.

Variable Breakdown of each PCA/EFA

PCA	PCA	EFA	EFA
Full dataset	Subsetted dataset	Full dataset	Subsetted dataset
PCA1 (-) country satisfaction - dissatisfied (-) trump approval - disapprove (+) economy favors – country is fair (+) democrats and religion - neutral (-) trump tax law - disapprove	PCA1 (+) country satisfaction - dissatisfied (+) trump approval - disapprove (+) business profits – corporations fair (-) economy favors – country is fair (-) democrats and religion - neutral (+) trump tax law - disapprove	EFA1 (+) country satisfaction - dissatisfied (+) trump approval - disapprove (-) economy favors – country is fair (-) democrats and religion - neutral (+) trump tax law – disapprove	EFA1 (+) country satisfaction - dissatisfied (+) trump approval - disapprove (+) trump tax law - disapprove
PCA2 (+) economy favors – don't know (*EFA3 says this is separate) (+) black discrimination – don't know (+) hispanic discrimination - don't know (+) white discrimination – don't know	PCA2 (+) black discrimination – some (+) hispanic discrimination - some	EFA2 (+) black discrimination – don't know (+) hispanic discrimination - don't know (+) white discrimination – don't know	EFA2 (+) trust government – most of the time (-) trust government – sometimes (*EFA3 says this is separate) (+) trust government – never *note: EFA 2 is not suggested from eigenvalues*

Kay Quiballo | MSDS 411 | 04.10.2023

Breakdown Analysis

Component/Axis 1: Trump, Economics, and Religion

The first and main component/axis was aligned across both PCA and EFA as well as across the full and subsetted datasets. This group of variables had the heaviest loading and accounted for the most variance in the datasets.

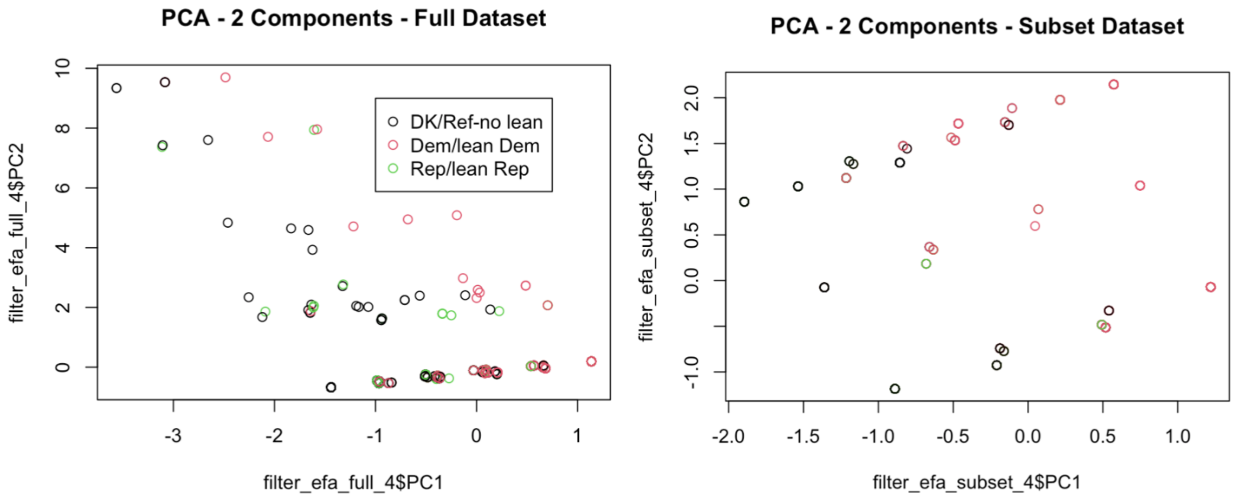
The variables included responding dissatisfied with the country, dissatisfied with Trump, and disapproved of Trump's tax law. Some of the PCA/EFA also had variables about how business corporations are fair, the economy is favors the rich, business corporations are fair, and Democrats are not neutral to religion. The opposite is true where those who did not respond as dissatisfied with the country, Trump, and Trump's tax law also believed that the economy is fair and democrats are neutral towards religion.

Component/Axis 2: Racial Discrimination

The second component/axis helps with explaining variance but not as much as the first. In accordance with the Scree plot and eigenvalues, this second component/axis is included in both PCA (the full and subsetted dataset), and the full EFA.

The variables included responding "Don't Know" to the presence of racial discrimination to white/black/Hispanic individuals. The opposite is true where those who responded "Some" to the presence of discrimination to black/Hispanic individuals were clustered together.

Principal Components by Party Leaning



After graphing the datasets using their primary components, we can see a weak pattern the red points (which are democratic leaning responses) are more towards the right of each graph (higher PCA1 scores) whereas green points (republican leaning) and black points (don't know) are typically more towards the left of the graph (lower PCA1 scores). As a consultant, we would advise our stakeholders to target the clusters of individuals for the party they are running for by relying on appealing to their collective political opinions.

Final Recommendation

Here is an example of implementing the findings from this PCA/EFA to one of our stakeholders. Consider John Johnson who is running for a political position under the Democratic party and wants recommendations for his campaign. His campaign team sent us 3 major points that he stands for and wants to know recommendations from the data team at our political consultant agency.

The 3 points that the campaign sends over are the following:

1) Johnson is a Christian, and hopes to use his background to appeal to other Christians.
2) Johnson does not agree with Trump's tax law and hopes to make changes to tax policy.
3) Johnson wants to improve the economy by focusing on providing new jobs.

From these 3 points we would recommend the following based on the PCA:

1) Johnson should not use his Christian background as a main talking point. Although he wants to win over more votes from the Christian demographic, PCA1 indicates that the target demographic in the Democratic party are not neutral to religion, and would result in the loss of votes from the main demographic he is trying to win votes from.
2) Johnson should vocalize his disapproval of Trump's tax law as this aligns well with PCA1 and his target demographic. He should also emphasize other points he disagrees with Trump on.
3) Johnson may speak on improving the economy but should not make this his main talking point. PCA1 indicates that the target demographic thinks that the economy is generally fair, and emphasizing this in a campaign won't be as effective as the 2nd recommendation.

Resources

Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. (2018). Netherlands: Elsevier Science.