Assignment 1: PEW 2017 Survey Data Analysis

Jonah Muniz

Abstract

When analyzing the political beliefs of a population people are primarily categorized as either liberal or conservative. Although this categorization could be accurate, majority of people lie somewhere between the two on their political views. Additional factors could also influence how someone votes during an election. This analysis will analyze survey data and perform variable reduction techniques in order to discover where someone lies on the liberal or conservative spectrum and any additional factors that are important to understand political affinity.

Keywords: Variable Reduction, Multidimensional Scaling, Eigenvalues, Eigenvectors, Principal Components, Scree Plots

Introduction

The conventional approach when categorizing the political environment in the USA, especially during 2019, is that there are two dominating political parties, Republican and Democratic. Although there are two party's majority of people's political beliefs range between the two. Leveraging the PEW 2019 survey dataset, the data will be analyzed to see where an observation lands on the spectrum of liberal to conservative. This will be used to help the democratic party target people who lie in the middle or lean towards liberal to vote democratic in their next election. Using exploratory data analysis and feature engineering through variable reduction techniques overall political sentiment can be analyzed on a spectrum to see how liberal or conversative respondents are.

Literature Review

A similar political analysis was performed analyzing tweets on twitter to classify whether a person is a Democrat or Republican (Chouhbi 2020). Twitter handles following political party leaders' tweets were analyzed and used to train and test a neural net algorithm to classify whether the twitter handle was a Democrat or Republican. The goal of this analysis was to be able to predict whether someone was likely to vote Democrat or Republican by analyzing their tweets.

Methods

To be able to analyze the PEW 2019 survey dataset, the dataset was obtained from pewresearch.com. The dataset from pewresearch.com contains 129 variables and 1503 observations. The survey was conducted over the phone leveraging a standardized script. The dataset is downloaded as a spss file.

To ingest the survey dataset into R the read.spss function from the Foreign package will be used (Team 2020). The dataset will be converted to a data frame during ingestion into R. Analyzing the survey data and leveraging the documentation on how the survey was conducted all survey questions can be found between variables 26-98. A separate dataset was created to capture only the survey questions and responses for all 1503 observations. This new dataset was named pewopinion and contained 73 variables and 1503 observations.

Now that a dataset of only the questions is compiled, each question is analyzed to see if all observations have answered the question. A new dataset is created to include only questions in the survey that all observations have answered. This will allow analysis to be performed as there will be a datapoint for each variables for each observation. This dataset is named pewwork and contains 30 variables and 1503 observations. A matrix was then created to create binary indicator variables for all 30 variables. This converts 30 multi-choice variables to 100 binary indicator

variables. The matrix is then converted to a data frame named pewdf consisting of 100 variables and 1503 observations.

Exploratory data analysis is conducted on pewdf to understand the amount of survey responses for each choice per question and their distribution. All data cleansing activities are performed during this time before leveraging the data frame to conduct feature engineering and variable reduction analysis.

A correlation matrix is created leveraging the cleaned version of pewdf. A parallel analysis is then performed leveraging the fa.parallel and scree functions from the Psych package in R. Eigen values and vectors are then calculated leveraging the eigen function from the Base package. The scree plots and eigen values are then used to determine the number of principal components to keep in the PCA and factor analysis. PCA and factor analysis is performed on the correlation matrix leveraging the principal and fa function from the Psych package. The fa function is used with zero rotation and a varimax rotation.

Each principal component is then analyzed, and each binary indicator is then analyzed. All loadings that are greater than .50 are kept and considered as a descriptor of the principal component. Each principal component is then a new feature of the dataset reducing the number of variables in the dataset while maintaining the necessary information. Observations are analyzed and visualized using multidimensional scaling techniques to visualize observations along the principal components.

Results

Following the methods above, tables for each question in the survey data were created to analyze count of responses per each question. Figure 1 in the appendix shows a sample of tables

analyzed. Q69(VOL) Other/Depends was never selected as a response across all 1503 observations so it was removed from the data set to allow for the proper analysis to be performed. The new pewdf_cleaned dataset now has 99 binary indicator variables and 1503 observations.

A correlation matrix was then created using the pewdf_cleaned dataset determining the correlation between each binary indicator in the dataset. Figure 2 in the appendix is a sample of the correlation matrix of pewdf_cleaned for questions 1, 2, 19, and 20 multiple-category items. If Q69(VOL) Other/Depends was not removed from the dataset the correlation matrix would not be able to be created due to the inability to determine a correlation with 0 or undefined.

A scree plot with parallel analysis was then created calculating eigenvalues of both the principal components and factor analysis. Figure 3 in the appendix shows the Scree plot with parallel analysis. An eigenvalue by index plot was also created, see Figure 4 in the appendix. Reviewing the scree plot and eigenvalue plot, 3 factors were determined to be the ideal amount. Analyzing the 3 eigenvalues you can see how many variables each principal component captures, Figure 5 in the appendix. As can be seen in Figure 5, the first two eigen values capture 5 variables and the third captures about 3.

Now that the number of factors has been determined the 3 factors were calculated utilizing the fa function from the Pysch package. Figure 6 in the appendix shows the 3 factors and a sample of their values for each loading. All loadings with a value greater than 0.50 are determined significant descriptor of the factor. Figure 7 in the appendix shows a table outlining the factor loadings of each factor. The first factor can be considered to capture how liberal or conversative an observation is. Those who responded that they were disappointed with the way things are going in the country and disapprove of the way Donald Trump is handing his job as president

would be considered more liberal than those who responded the opposite for example. Factor 2 captures observation's feeling towards economic and racial matters. This factor captures how indifferent someone is to respond to these types of questions and is a proxy on how they may feel politically. The third factor captures the level of trust an observation has with the government. A multidimensional scaling plot was created to show the potential clusters of responses and their correlation across the three principal components.

Conclusion

Analyzing pewdata survery data can enable an organization like the democratic party to analyze what the key factors are that can describe an individual's political alignment and how to better tailor their message to specific groups. Through the above analysis we have uncovered 3 key factors that describe the pewdata survey and each observations political beliefs.

My recommendation for my client is to target messages of trust to those who fall more liberal in the views, factor 1, and those who have middle to little trust in the government, factor 3. This will hopefully change their attitude towards the government and will persuade them to vote for the democratic politician on the ballot.

Appendix

```
> table(pewwork$q68b)

  Friendly toward religion    Neutral toward religion Unfriendly toward religion   (VOL) Don't know/Refused
                       816                        414                        203                         70
> table(pewwork$q68d)

  Friendly toward religion    Neutral toward religion Unfriendly toward religion   (VOL) Don't know/Refused
                       351                        904                        174                         74
> #Other/Depends has a count of 0 for q69
> table(pewwork$q69)

                                  Should keep out of political matters
                                                                   841
Should express their views on day-to-day social and political questions
                                                                   610
                                                 (VOL) Other/Depends
                                                                     0
                                               (VOL) Don't know/Refused
                                                                    52
> table(pewwork$q70)

                 Approve               Disapprove (VOL) Don't know/Refused
                     588                      703                      212
> table(pewwork$q71)

              Very well            Somewhat well             Not too well          Not at all well
                     470                      542                      257                      169
(VOL) Don't know/Refused
                      65
> table(pewwork$q75)

                   A lot                 A little          Nothing at all (VOL) Don't know/Refused
                    1232                      195                       65                       11
```

Figure 1: *Exploratory data analysis of counts of question responses*

	q1Dissatisfied	q1(VOL) Don't know/Refused	q2Disapprove	q2(VOL) Don't know/Refused	q19Some of the time	q19Only now and then	q19Hardly at all	q19(VOL) Don't know/Refused	q20Frustrated	q20Angry	q20(VOL) Don't know/Refused
q1Dissatisfied	1.000000000	-0.309964965	0.501939950	-0.074091755811421855551541957	0.0016040353	0.0051205615	-0.077384110	-0.0267863468	0.134939531	0.1539675863	-0.0606373553
q1(VOL) Don't know/Refused	-0.309964965	1.000000000	-0.071144975	0.121738301016601810355943777	0.0256187896	0.0100133503	0.054288735	0.0684401006	-0.005298403	-0.0973871349	0.1163803896
q2Disapprove	0.501939950	-0.071144975	1.000000000	-0.236741405574955543489323 87	0.0655969775	0.0509931026	-0.053740857	-0.0330506047	0.086252155	0.0684606766	-0.0620283291
q2(VOL) Don't know/Refused	-0.074091756	0.121738301	-0.236741406	1.000000000000000000000000000	0.0303420198	0.0895273733	0.162426019	0.0761704328	-0.050003920	-0.0548738858	0.1321120270
q19Some of the time	0.001604035	0.025618790	0.065596977	0.030342019785160330436 20226	1.0000000000	-0.1494843899	-0.120233637	-0.0338034295	0.055227465	-0.0830232936	0.0017401863
q19Only now and then	0.005120561	0.010013350	0.050993103	0.089527373296534346058805 91	-0.1494843899	1.0000000000	-0.073598125	-0.0206919550	-0.055264310	-0.0386919512	-0.0025652071
q19Hardly at all	-0.077384110	0.054288735	-0.053740857	0.162426018701269908595108 40	-0.1202336375	-0.0735981246	1.000000000	-0.0166430021	-0.134499730	0.0091123294	0.1002289837
q19(VOL) Don't know/Refused	-0.026786347	0.068440101	-0.033050605	0.076170432840543519925624 59	-0.0338034295	-0.0206919550	-0.016643002	1.0000000000	-0.041116554	-0.0353647433	0.0962931613
q20Frustrated	0.134939531	-0.005298403	0.086252155	-0.050003919933784890528016 83	0.0552274648	-0.0552643099	-0.134499730	-0.0411165541	1.000000000	-0.6102552035	-0.2189658562
q20Angry	0.153967586	-0.097387135	0.068460677	-0.054873885757721743705594 97	-0.0830232936	-0.0386919512	0.009112329	-0.0353647433	-0.610255204	1.0000000000	-0.0959046096
q20(VOL) Don't know/Refused	-0.060637355	0.116380390	-0.062028329	0.132112026991865333380715 70	0.0017401863	-0.0025652071	0.100228984	0.0962931613	-0.218965856	-0.0959046096	1.0000000000

Figure 2: *Sample correlation matrix of the binary indicator variables for question responses*

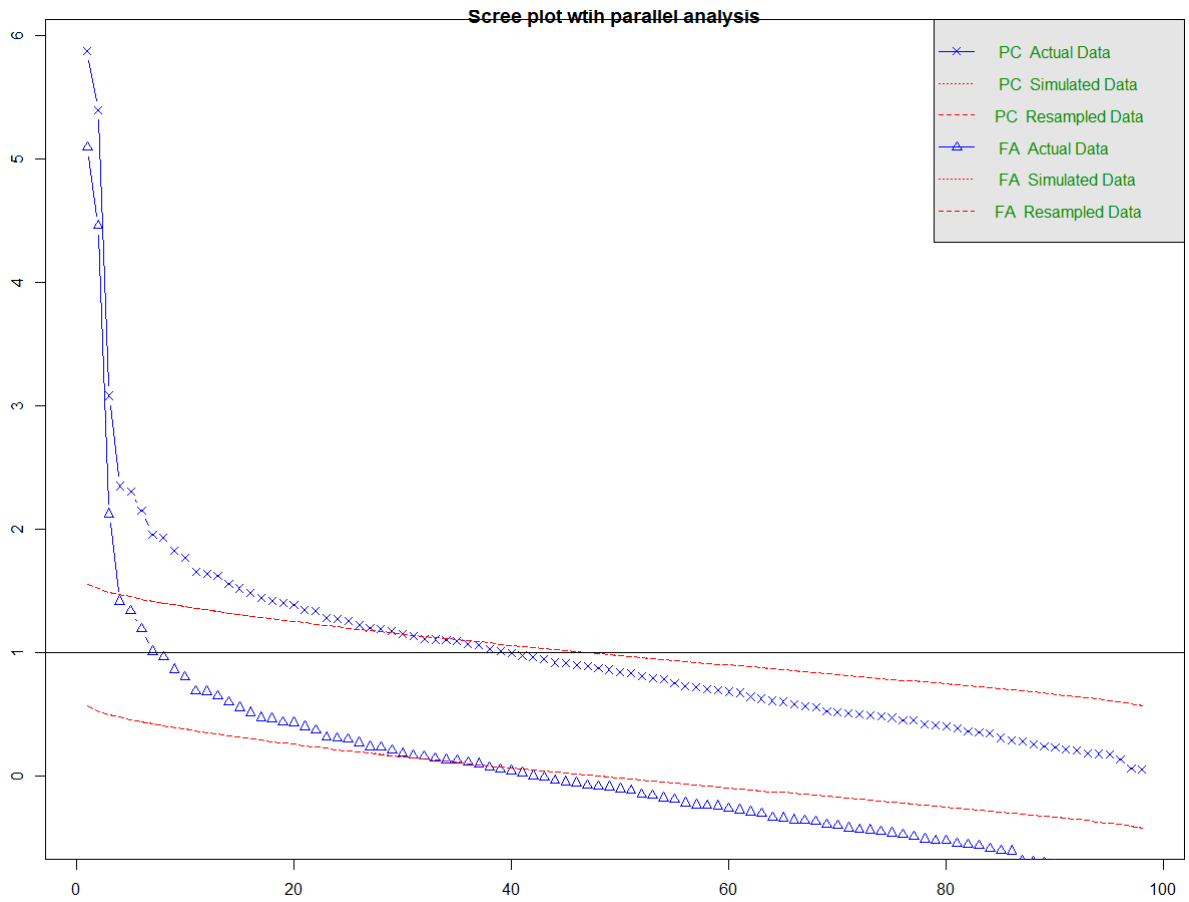

Figure 3: *Scree plot with parallel analysis performed for both principal component and factor analysis*

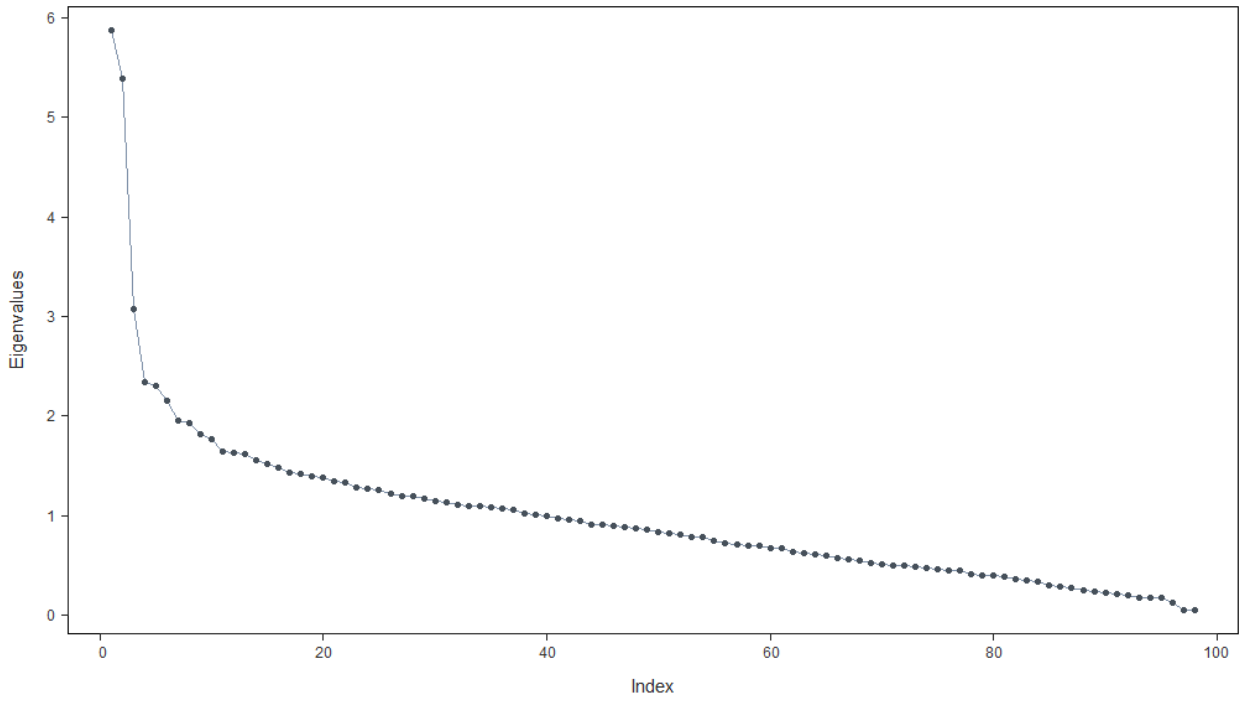

Figure 4: *Eigenvalues plot*

```
 [1] 5.87543273 5.38994879 3.07887621 2.34743780 2.30049641 2.14866101 1.95541774 1.92896597 1.82491265 1.76324009 1.64930470
[12] 1.63435408 1.61741876 1.55664829 1.51816416 1.47888769 1.43839865 1.41494354 1.40132432 1.38214501 1.34348362 1.33279939
[23] 1.27842178 1.27204844 1.25310100 1.22103157 1.19924964 1.19134878 1.17155936 1.14946468 1.13357855 1.10953712 1.10251563
[34] 1.09686219 1.08882411 1.06670978 1.05660815 1.02627474 1.01138292 0.99607332 0.97318043 0.95991191 0.94516824 0.91609764
[45] 0.91467562 0.89790576 0.88759816 0.87275851 0.85847007 0.84112840 0.82921278 0.80941973 0.78981236 0.78051649 0.75019707
[56] 0.72613403 0.71663823 0.70325610 0.69538372 0.68011671 0.67302007 0.64002675 0.62370170 0.60703885 0.59969086 0.57643328
[67] 0.56297763 0.55366677 0.52362173 0.51579368 0.50410601 0.49704299 0.48615643 0.48050514 0.46842663 0.45162143 0.44591994
[78] 0.41753731 0.40610450 0.39731571 0.38596867 0.36201297 0.34909986 0.34156822 0.30698979 0.28578881 0.27711704 0.25351442
[89] 0.23893336 0.23148822 0.21266394 0.20623010 0.17793044 0.17567217 0.17327989 0.13143770 0.05803074 0.04813094
```

Figure 5: *Eigenvalues for the correlation matrix depicting the number of variables captured in each principal component*

```
Factor Analysis using method =  ml
Call: fa(r = pewdf_clean[, -1], nfactors = 3, rotate = "none",
    fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
                                                    ML2   ML3   ML1
q1Dissatisfied                                    -0.55  0.05 -0.16
q1(VOL) Don't know/Refused                         0.13  0.13  0.04
q2Disapprove                                      -0.85  0.05 -0.10
q2(VOL) Don't know/Refused                         0.13  0.18  0.08
q19Some of the time                               -0.05  0.01  0.01
q19Only now and then                              -0.03  0.04 -0.04
q19Hardly at all                                   0.05  0.13  0.05
q19(VOL) Don't know/Refused                        0.04  0.30  0.04
q20Frustrated                                     -0.07 -0.10 -0.18
q20Angry                                          -0.10  0.05  0.00
q20(VOL) Don't know/Refused                        0.08  0.16  0.07
q25Most of the time                                0.01 -0.12  0.64
q25Only some of the time                           0.00  0.00 -1.00
q25(VOL) Never                                    -0.06  0.02  0.53
q25(VOL) Don't know/Refused                        0.03  0.26  0.19
q47No, not keeping promises                        0.43 -0.14 -0.02
q47(VOL) Don't know/Refused                        0.01  0.29  0.04
q50aStatement #2                                   0.45 -0.16  0.08
q50a(VOL) Neither/Both equally                    -0.02  0.07  0.02
q50a(VOL) Don't know/Refused                       0.09  0.35  0.05
q50bStatement #2                                   0.42 -0.08  0.04
q50b(VOL) Neither/Both equally                     0.01  0.11  0.00
q50b(VOL) Don't know/Refused                       0.06  0.30  0.03
q50cStatement #2                                   0.20 -0.05 -0.01
q50c(VOL) Neither/Both equally                     0.04  0.17  0.00
q50c(VOL) Don't know/Refused                       0.03  0.22  0.01
q50dStatement #2                                   0.56 -0.16  0.07
q50d(VOL) Neither/Both equally                     0.00  0.05 -0.01
q50d(VOL) Don't know/Refused                       0.09  0.50  0.09
q50eStatement #2                                   0.25 -0.02  0.05
q50e(VOL) Neither/Both equally                     0.04  0.10  0.01
q50e(VOL) Don't know/Refused                       0.06  0.29  0.03
q58Favor                                          -0.03 -0.04  0.03
q58Oppose                                          0.23 -0.06 -0.04
q58Strongly oppose                                 0.22 -0.04  0.03
q58(VOL) Don't know/Refused                        0.15  0.16  0.03
q60Never fair game                                -0.23 -0.08 -0.04
q60(VOL) Other/Depends                             0.04  0.10  0.05
q60(VOL) Don't know/Refused                        0.13  0.36  0.01
q61aSome                                           0.25 -0.13  0.00
q61aOnly a little                                  0.30 -0.07 -0.01
q61aNone at all                                    0.19  0.00  0.06
q61a(VOL) Don't know/Refused                       0.13  0.61  0.05
q61bSome                                           0.10 -0.09 -0.02
q61bOnly a little                                  0.33 -0.11  0.01
q61bNone at all                                    0.21  0.00  0.07
q61b(VOL) Don't know/Refused                       0.13  0.57  0.08
q61cSome                                           0.23 -0.09 -0.01
q61cOnly a little                                 -0.13 -0.08 -0.07
q61cNone at all                                   -0.29  0.01  0.04
q61c(VOL) Don't know/Refused                       0.09  0.56  0.13
q64Moderately fair                                 0.37 -0.15  0.03
q64Not too fair                                   -0.23 -0.03 -0.06
q64Not fair at all                                -0.30  0.08 -0.01
q64(VOL) Don't know/Refused                        0.07  0.38  0.05
q65aSome                                           0.03 -0.08 -0.05
```

Figure 6: *Sample loading values for each variable highlighting the loadings with significance (|loading| > 0.50)*

FA1: Liberal Conservative Spectrum			
Question Number	Question	Response	FA Value
Q1	All in all, are you satisfied or dissatisfied with the way things are going in this country today	Dissatisfied	-0.55
Q2	Do you approve or disapprove of the way Donald Trump is handling his job as President?	Dissapprove	-0.85
Q50	The economic system in this country unfairly favors powerful interests [OR] The economic system in this country is generally fair to most Americans	Statement #2	0.56
Q68	As I name some groups, please tell me whether you feel each one is generally FRIENDLY toward religion, NEUTRAL toward religion, or UNFRIENDLY toward religion. First, do you feel that the Democratic Party is/are generally friendly toward religion, neutral toward religion, or unfriendly toward religion?	Unfriendly toward religio	0.61
Q70	Do you approve or disapprove of the tax law passed by Donald Trump and Congress in 2017?	Dissapprove	-0.77

PC2: Undecided/Uninterested in Economic or Racial issues			
Question Number	Question	Response	FA Value
Q61a	Please tell me how much discrimination there is against each of these groups in our society today. Blacks	Don't Know/Refused	0.61
Q50d	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	0.5
Q61b	Please tell me how much discrimination there is against each of these groups in our society today. Hispanics	Don't Know/Refused	0.57
Q61c	Please tell me how much discrimination there is against each of these groups in our society today. Whites	Don't Know/Refused	0.56

PC3: Politcal Trust			
Question Number	Question	Response	FA Value
Q25	How much of the time do you think you can trust the government in Washington to do what is right? Just about always, most of the time, or only some of the time?	Only some of the time	-1
Q25	How much of the time do you think you can trust the government in Washington to do what is right? Just about always, most of the time, or only some of the time?	Never	0.53
Q25	How much of the time do you think you can trust the government in Washington to do what is right? Just about always, most of the time, or only some of the time?	Most of the time	0.64

Figure 7: *Significant factor loadings for the 3 factors*

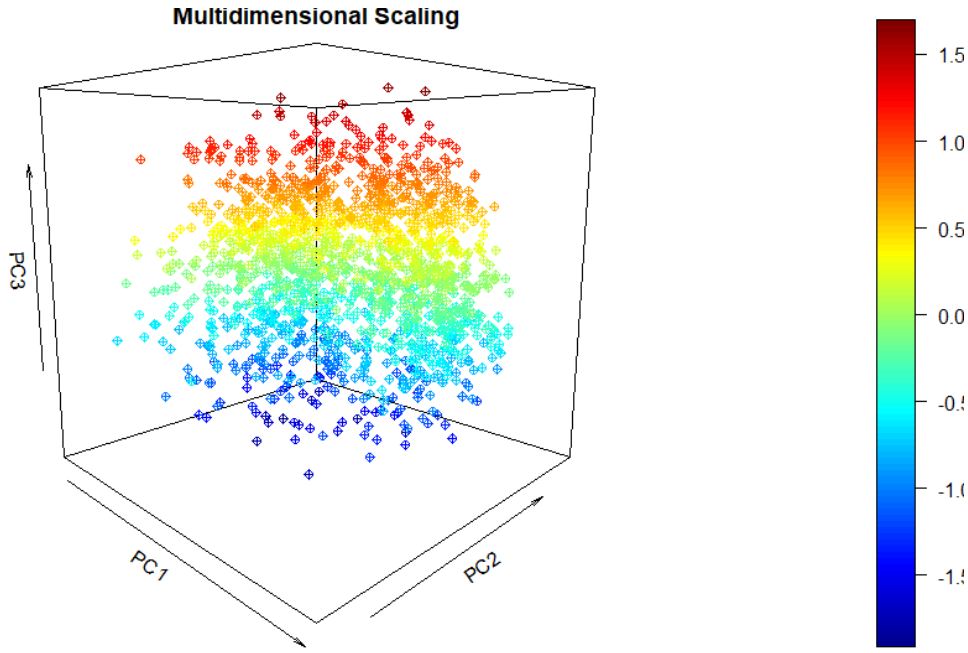

Figure 8: *Multidimensional scaling of the observation's responses across 3 principal components*

References

1. Chouhbi, Kamal. "Are You Democrat or Republican? Let Your Tweets Define You ..." Medium. Towards Data Science, February 5, 2020. https://towardsdatascience.com/are-you-democrat-or-republican-let-your-tweets-define-you-4aa4cadf4bea.
2. Team, R Core. "Read Data Stored by 'Minitab', 's', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'DBase', ... [R Package Foreign Version 0.8-81]." The Comprehensive R Archive Network. Comprehensive R Archive Network (CRAN), December 22, 2020. https://cran.r-project.org/web/packages/foreign/index.html.
3. "Fa.parallel: Scree Plots of Data or Correlation Matrix Compared to Random ``Parallel' Matrices." RDocumentation. Accessed October 5, 2021. https://www.rdocumentation.org/packages/psych/versions/2.1.9/topics/fa.parallel.
4. "Eigen: Spectral Decomposition of a Matrix." RDocumentation. Accessed October 5, 2021. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/eigen.