An unsupervised machine learning analysis for votes percentages in USA elections.
Big Data Analysis for Economics and Finance 1st intermediate project

Big Data Analysis Report

Clustering

Author: Stefano Blando

FACOLTÀ DI ECONOMIA

Cluster analysis

- *votes.repub*

Analist:

BLANDO Stefano | 0334525

Supervisor:

Professor Alessio Farcomeni

Doctor Federico Roscioli

2023/2024

Contents

INTRODUCTION 3

DATA PREPROCESSING 4

Data exploration and cleaning 4

Distance Metric Calculation 4

CLUSTER ANALYSIS 4

Elbow method 4

Principal Component Analysis 4

VISUALIZATION 5

Pairs 5

Heatmap 5

K-MEANS 5

Validation 5

TRIMMING 5

CTL curves 5

Trimming 5

CONCLUSION 6

INTRODUCTION

This report presents a comprehensive analysis of the voting patterns for the Republican party in the United States presidential elections from 1856 to 1976. The data, sourced from the votes.repub dataset, encompasses voting records across all 50 states.

The primary objective of this analysis is to identify and interpret distinct clusters within the data, which may reveal underlying patterns or trends in Republican voting behaviour over the selected period. To achieve this, we employ the k-means clustering algorithm, a method for partitioning a dataset into k distinct, non-overlapping subgroups.

Before conducting the cluster analysis, we first preprocess the data by selecting records from 1900 onwards and removing any states with missing values. This ensures the integrity and reliability of our subsequent analysis.

We then determine the optimal number of clusters and the appropriate trimming level using ctlcurves. This step is crucial in enhancing the robustness of our k-means solution against potential outliers.

The final part of our analysis involves reporting and interpreting the “optimal” k-means solution. This includes detailing the centroids (i.e., the center points of each cluster), identifying any trimmed units, and describing the sizes and compositions of the clusters.

DATA PREPROCESSING

Data exploration and cleaning

The initial exploration of the dataset involved a meticulous examination of its structure using str.

Further insights were gained by inspecting the first and last rows with head and tail. To streamline subsequent analyses, relevant columns were extracted for focused investigation.

An essential part of data preprocessing, the count of missing values per row was calculated using rowSums(is.na()). Subsequently, the na.omit function was employed to clean the dataset, ensuring that rows with missing values were removed. This step is critical for robust analysis, as it enhances the dataset's quality.

To gain a statistical understanding of the dataset, the standard deviation and mean for each year were computed using apply. This provided valuable insights into the variability and central tendency of the data over time.

Distance Metric Calculation

A sophisticated approach to understanding relationships between observations involved the calculation of correlation-based distance using the Pearson method (get_dist). The resulting distance matrix was then visualized through an interactive dendrogram using fviz_dist. This visualization offers an intuitive representation of dissimilarity within the dataset.

CLUSTER ANALYSIS

Elbow method

The Elbow Method is a crucial step in determining the optimal number of clusters. A loop was implemented to apply the K-means algorithm for different numbers of clusters.

The resulting plot of wcss against the number of clusters aided in identifying the elbow point, providing insights into the optimal cluster number.

Principal Component Analysis

PCA, a dimensionality reduction technique, was applied to the cleaned dataset using prcomp. This transformation allowed for the identification of key features and patterns in the data. Visualization of PCA results through various fviz_pca functions provided a comprehensive understanding of variable contributions and relationships in reduced dimensions.

VISUALIZATION

Pairs

Pairs plots were generated both before and after PCA using the pairs function. These plots provided a visual representation of relationships between variables, aiding in the identification of potential clusters or patterns within the data.

Heatmap

An interactive heatmap created with heatmaply offered a dynamic visualization of the cleaned dataset. This visualization technique allowed for the exploration of intricate relationships and patterns that might not be immediately evident through traditional statistical summaries.

K-MEANS

The K-means clustering algorithm was applied to the cleaned dataset using the kmeans function. The resulting clusters were visualized through fviz_cluster, providing immediate insights into the grouping of observations based on similarity.

Validation

Silhouette analysis, residuals analysis, and Dunn index computation were employed to rigorously validate the quality of the clustering results. This multi-faceted approach ensured a thorough evaluation of the chosen number of clusters and the reliability of the clustering outcome.

TRIMMING

CTL curves

Classification Trimmed Likelihood (CTL) curves were generated to visualize the impact of trimming on the dataset. This visual representation was crucial for understanding the trade-off between retaining data points and achieving robust clustering results. CTL curves provide insights into the optimal level of trimming for a balanced analysis.

Trimming

Trimming was applied using the tkmeans function, with a specified alpha value derived from the observation of the CTL curves. Visualization of the trimmed clusters through fviz_cluster offered a clear understanding of how the trimming operation influenced the distribution of data points. This step provided insights into the robustness of the clustering outcome after the application of trimming.

CONCLUSION

In summary, the analysis of the votes.repub dataset encompassed meticulous data preprocessing, robust cluster identification using elbow method, insightful Principal Component Analysis (PCA), and effective visualization techniques.

The application of K-Means clustering, coupled with thorough validation, ensured the reliability of identified clusters.

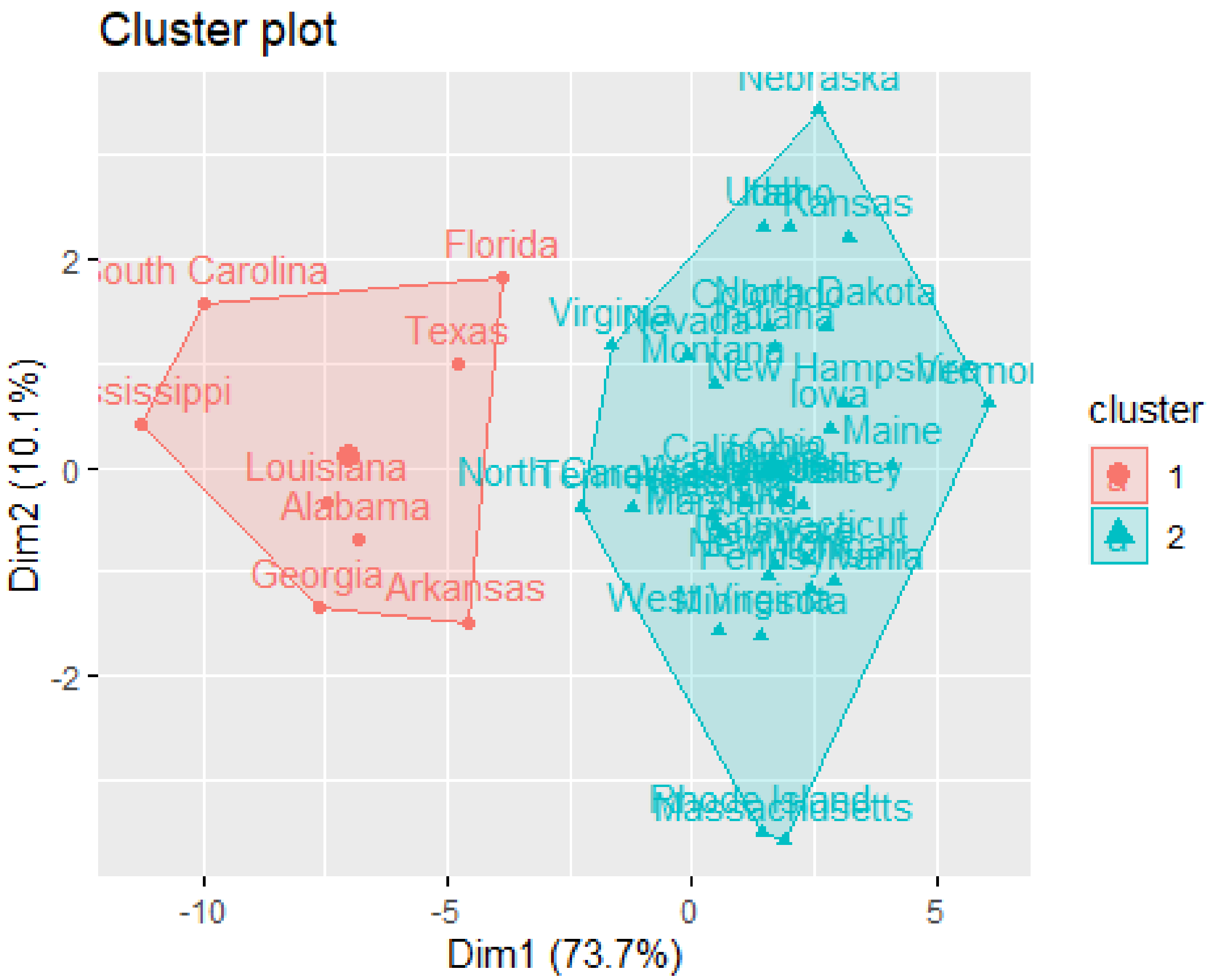
The exploration of Classification Trimmed Likelihood (CTL) curves and trimming added depth to the understanding of cluster robustness: it has been shown that trimming does not provide an additional benefit to the definition of clusters for the data frame, and therefore it has been not applied to the analysis.

The final outcome is a total of 2 distinct clusters, with the following statistics:

- cluster size: [1] 35 8;
- diameter [1] 113.2451 100.6438;
- average distance: [1] 41.36975 62.49704;
- separation: [1] 39.00622 39.00622
- Clustering vector:

Alabama	**Arkansas**	**California**	**Colorado**	**Connecticut**	**Delaware**
2	2	1	1	1	1
Florida	**Georgia**	**Idaho**	**Illinois**	**Indiana**	**Iowa**
2	2	1	1	1	1
Kansas	**Kentucky**	**Louisiana**	**Maine**	**Maryland**	**Massachusetts**
1	1	2	1		
Michigan	**Minnesota**	**Mississippi**	**Missouri**	**Montana**	**Nebraska**
1	1	2	1	1	1
Nevada	**New Hampshire**	**New Jersey**	**New York**	**North Carolina**	**North Dakota**
1	1	1	1	1	1
Ohio	**Oregon**	**Pennsylvania**	**Rhode Island**	**South Carolina**	**Tennessee**
1	1	1	1	2	1
Texas	**Utah**	**Vermont**	**Virginia**	**Washington**	**West Virginia**
2	1	1	1	1	1
Wisconsin					
1					

- Within cluster sum of squares by cluster: [1] 32963.75 15007.72


Cluster plot
Nebraska
Utah
Idaho
Kansas
Florida
South Carolina
Texas
Virginia
Nevada
Colorado
North Dakota
Indiana
Montana
New Hampshire
Vermont
Iowa
ssissippi
Maine
California
Ohio
Louisiana
North Carolina
Tennessee
New Jersey
Maryland
Alabama
Connecticut
Delaware
New York
Michigan
Pennsylvania
Georgia
Arkansas
West Virginia
Minnesota
Rhode Island
Massachusetts
2
0
-2
-10
-5
0
5
Dim1 (73.7%)
Dim2 (10.1%)
cluster
1
2