APPLICATIONS OF DATA ANALYTICS: VISUALIZATION AND CLUSTER ANALYSIS OF GOVERNMENTAL DATA – TWO CASE STUDIES

ESSAY 2

OBJECTIVES

- Since data analytics is one way to explore the data and to help uncover hidden relationships
  - In these case studies we plan to explore the literature for the use of emerging data mining techniques in auditing
    - ✓ In particular, cluster analysis & visualization techniques as supportive tools to gain more insights into data.
- Conduct two case studies:
  1) Rutgers AICPA Data Analytics Research Initiative (RADAR): A Case Study.
     - ✓ Facilitate the integration of different data analytics tools and techniques into the audit process.
  2) Visualization and Clustering Analytics of U.S. states' on budgeting.
     - ✓ Information on U.S. States.

CONTRIBUTION

- We show how visualization and data clustering techniques could be used on governmental data and to help gain more information about financial statements & budgeting.

INTRODUCTION

- **Data mining** is the process of gaining insights and identifying interesting patterns and trends from data stored in large databases in such a way that the insights, patterns, and trends are previously unknown, statistically reliable, and actionable
  - Meaning that some decisions could be taken to exploit the knowledge, Sharma & Panigrahi (2013).
- **Cluster analysis** as a data mining approach can help find similar objects in data.
  - Kaufman & Rousseeuw (2009) have defined cluster analysis as ***"the art of finding groups in data."***

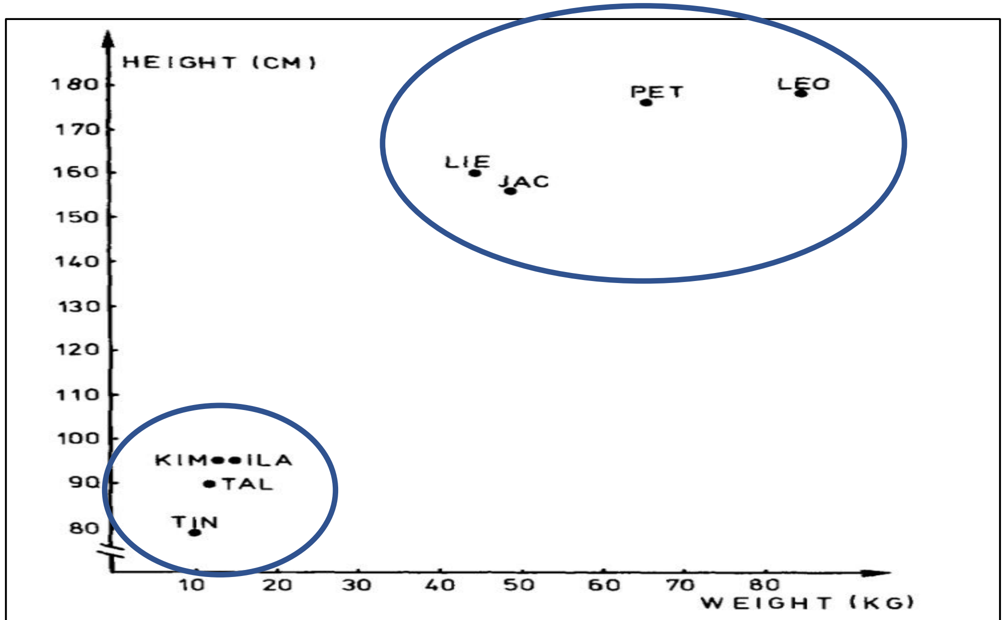


Figure 1 A plot of eight objects.

CLUSTER ANALYSIS OVERVIEW

- ***K*-means** Clustering:
  - *K*-means algorithm (MacQueen, 1967) is one of the most common and efficient data mining methods
    - *k*-means clustering - basically, the concept of "birds of a feather flock together.", McPherson et al. (2001).
  - It uses centroids to form clusters by optimizing the within clusters' squared errors.
  - Groups a dataset into *k* partitions known as clusters:
    - Choose a value for *k*, the total number of clusters to be determined.
    - Choose *k* instances (data points) within the dataset at random. These are the initial clusters' centers.
    - Scan through the list of *m* observations, then assign each observation to its nearest cluster's center.
    - Each cluster's center is then updated to be the average of the new observations assigned.
    - Repeat the previous two-steps iteratively until there are no more reassignments.
- **Hierarchical** Clustering:
  - In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters.

➢ <u>Both *k*-means and hierarchical clustering methods are unsupervised.</u>

HIERARCHICAL CLUSTERING

- Strategies for hierarchical clustering generally fall into two types:
- **Agglomerative (HAC):** This is a "bottoms up" approach based on similarities:
  - Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive (HDC):** This is a "top down" approach:
  - All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

1. RUTGERS AICPA DATA ANALYTICS RESEARCH INITIATIVE (RADAR): A CASE STUDY

- **RADAR Vision:** facilitate the integration of data analytics into audit process, and demonstrate through research how this can lead to advancement in the accounting profession.
- **Data:** RADAR Data.
  - U.S. States Financial Statements.
  - Average of the years were used: (FY 2000 – FY 2016).
  - Per Capita basis.
- **The variables** used in the analysis are as follow:
  1. Total General Fund Revenues.
  2. Excess (Deficiency) of Revenues over Expenditures.
  3. Total Operating Expenses.
  4. Education Expenses.
  5. Net Change in Fund Balance.
  6. General Fund Total Other Financing Sources.
  7. General Fund Transfers to Other Funds.
  8. General Fund Transfers from Other Funds.
  9. Pension Expense.

✓ *Cluster Analysis:*

- *K*-means cluster analysis.
- Hierarchical cluster analysis.

2. VISUALIZATION AND CLUSTERING ANALYTICS OF U.S. STATES: A CASE STUDY

By: Zamil S. Alzamil, Deniz Appelbaum, William Glasgall and Miklos A. Vasarhelyi

- **Data:** Volcker's Survey Results Data (Average Grades, 2015 - 2017).
  - How the U.S. states score on an annual basis on budgeting.
  - "Truth and Integrity in State Budgeting: What is the Reality?.", November 2, 2017.
- **Using five-variables:**
  1. Budget Forecasting.
  2. Budget Maneuvers.
  3. Legacy Costs.
  4. Reserve Funds.
  5. Transparency.
- **Methodology:**
  a. Data Visualization.
  b. Data Analytics: *k*-means & hierarchical cluster analysis.

DATA VISUALIZATION

Variables Correlation Coefficient

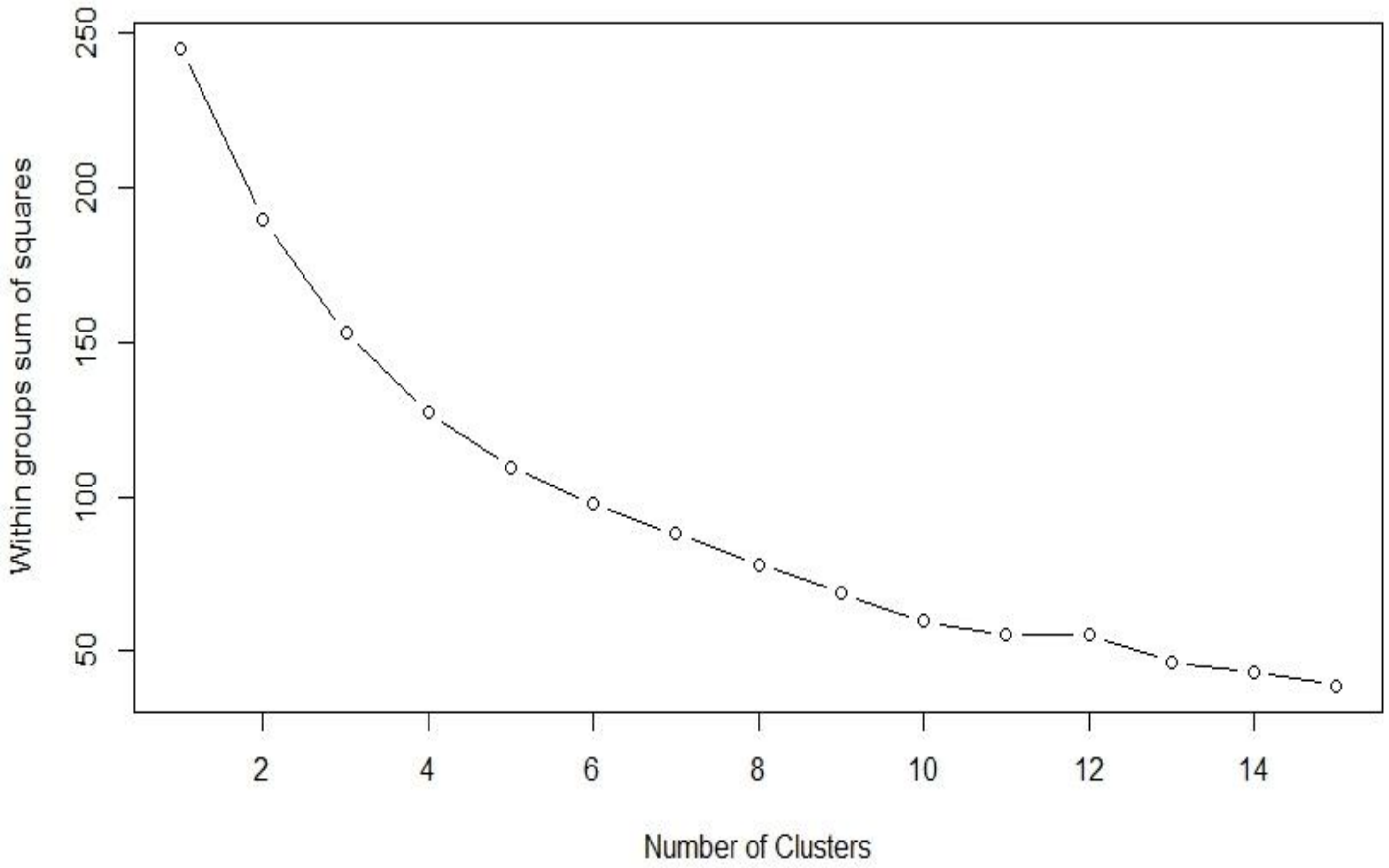
First we establish that there is a moderate correlation (relationship) between the variables of legacy costs and budget maneuvers (~0.512)

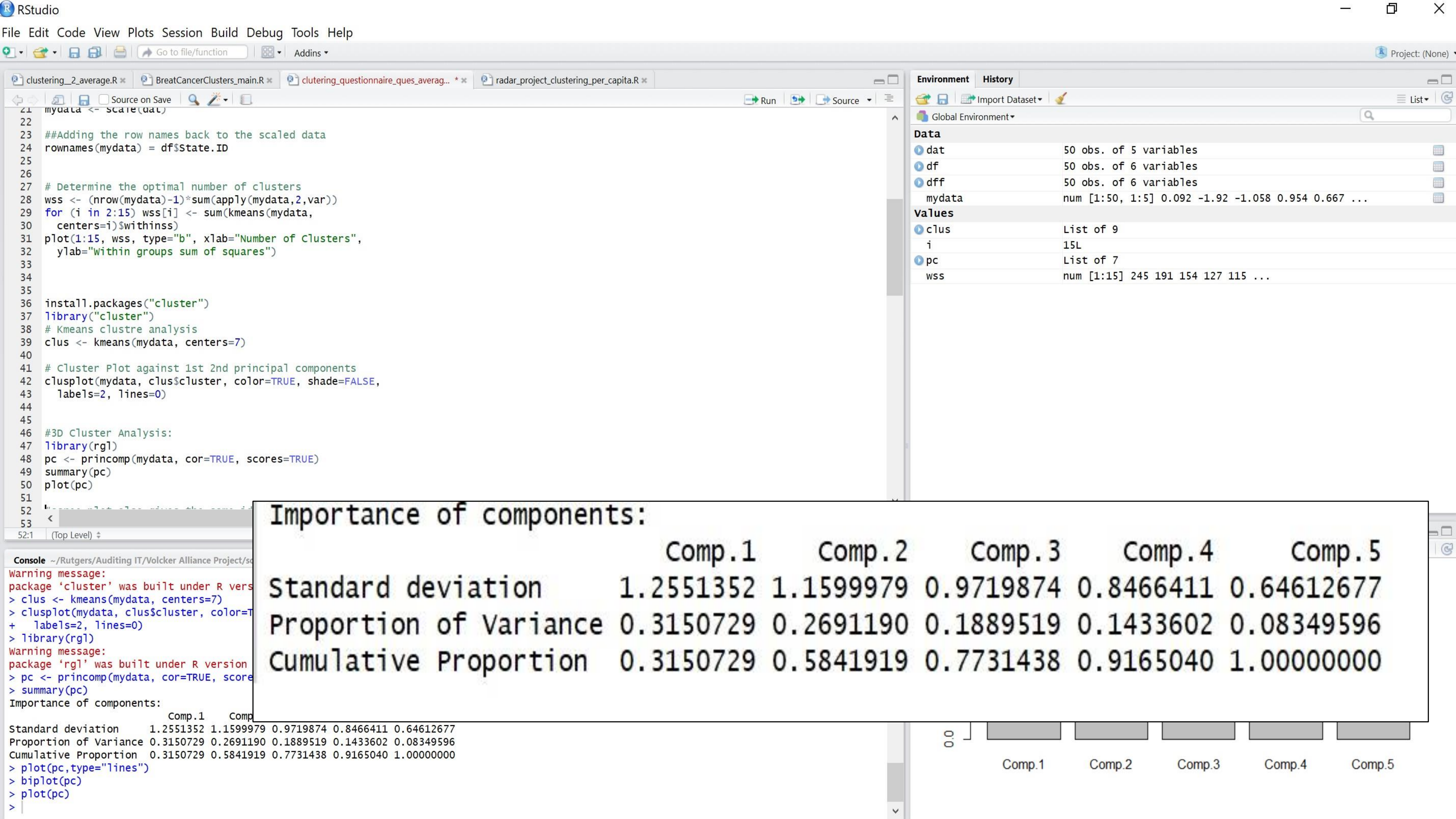
	Budget.Forecasting	Budget.Maneuvers	Legacy.Costs	Reserve.Funds	Transparency
Budget.Forecasting	1.000000000	-0.007919089	-0.03613848	0.25110021	0.18377649
Budget.Maneuvers	-0.007919089	1.000000000	0.51272449	0.22466741	-0.11578494
Legacy.Costs	-0.036138475	0.512724489	1.00000000	0.02784838	0.04485754
Reserve.Funds	0.251100213	0.224667415	0.02784838	1.00000000	0.09371242
Transparency	0.183776490	-0.115784941	0.04485754	0.09371242	1.00000000

- This analysis could assist in:
  - More insights into the survey results data.
  - Assist in selecting appropriate variables to build models.

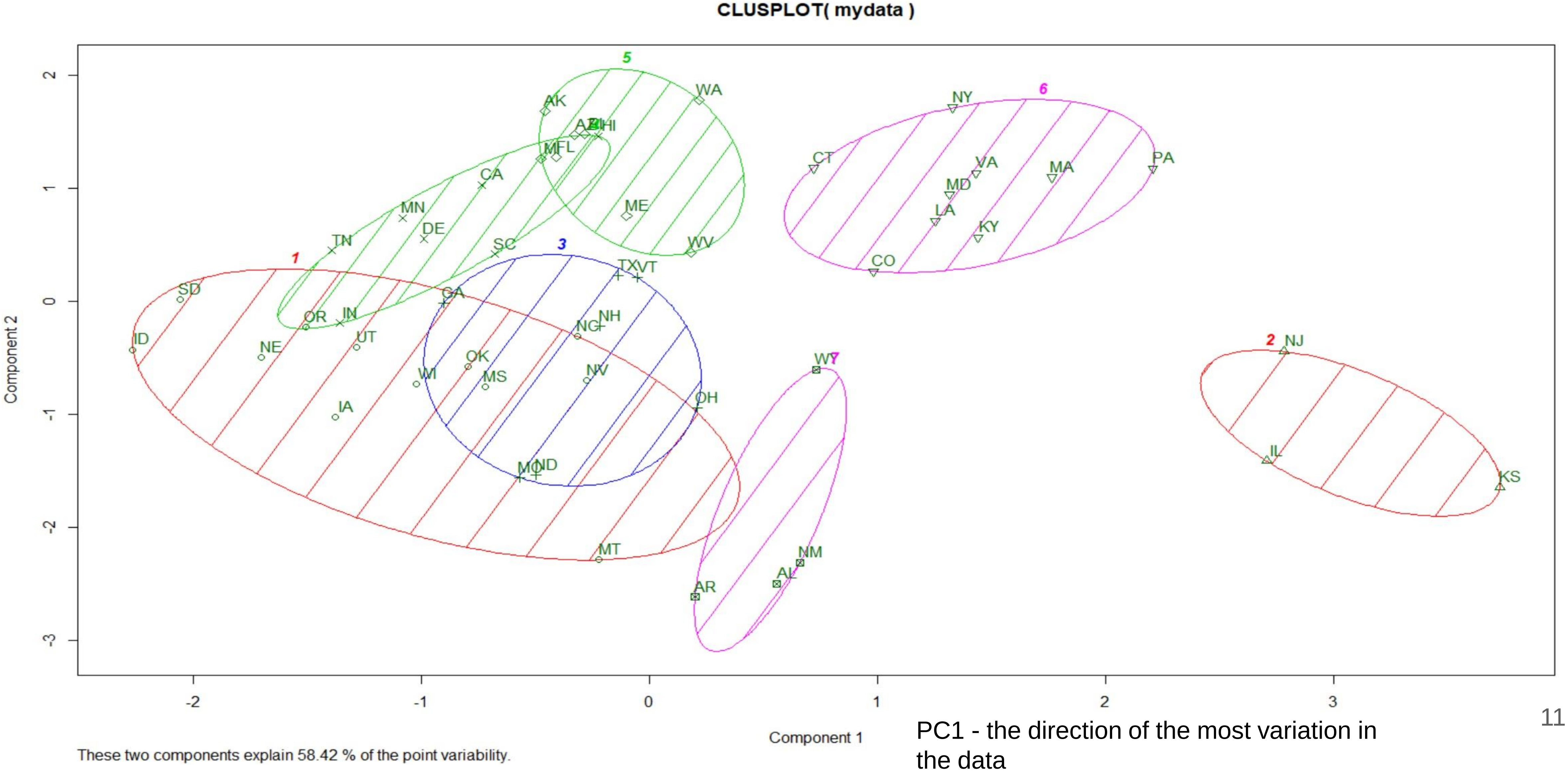
DATA ANALYTICS

- We explore the data by means of clustering:
  - how are the states similar with one another regarding their budgetary practices?
  - May we find previously unknown relationships and patterns with cluster analysis.
- The figure on the right side shows that 7 clusters would be a good fit.
- This method is called “the within clusters sum of squares” or the Elbow method which is a method of interpretation and validation of consistency of points within each cluster. It is performed by computing the within clusters sum of squares designed to help determine the optimal number of clusters.




RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Project: (None)
clustering_2_average.R
BreatCancerClusters_main.R
clutering_questionnaire_ques_averag...
radar_project_clustering_per_capita.R
Source on Save
Run
Source
mydata <- scale(dat)
##Adding the row names back to the scaled data
rownames(mydata) = df$State.ID
# Determine the optimal number of clusters
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
ylab="Within groups sum of squares")
install.packages("cluster")
library("cluster")
# Kmeans clustre analysis
clus <- kmeans(mydata, centers=7)
# Cluster Plot against 1st 2nd principal components
clusplot(mydata, clus$cluster, color=TRUE, shade=FALSE,
labels=2, lines=0)
#3D Cluster Analysis:
library(rgl)
pc <- princomp(mydata, cor=TRUE, scores=TRUE)
summary(pc)
plot(pc)
52:1
(Top Level)
Environment
History
Import Dataset
List
Global Environment
Data
dat 50 obs. of 5 variables
df 50 obs. of 6 variables
dff 50 obs. of 6 variables
mydata num [1:50, 1:5] 0.092 -1.92 -1.058 0.954 0.667 ...
Values
clus List of 9
i 15L
pc List of 7
wss num [1:15] 245 191 154 127 115 ...
Console ~/Rutgers/Auditing IT/Volcker Alliance Project/s
Warning message:
package ‘cluster’ was built under R vers
> clus <- kmeans(mydata, centers=7)
> clusplot(mydata, clus$cluster, color=T
+ labels=2, lines=0)
> library(rgl)
Warning message:
package ‘rgl’ was built under R version
> pc <- princomp(mydata, cor=TRUE, score
> summary(pc)
Importance of components:
Comp.1 Comp
Standard deviation 1.2551352 1.1599979 0.9719874 0.8466411 0.64612677
Proportion of Variance 0.3150729 0.2691190 0.1889519 0.1433602 0.08349596
Cumulative Proportion 0.3150729 0.5841919 0.7731438 0.9165040 1.00000000
> plot(pc,type="lines")
> biplot(pc)
> plot(pc)
0.0
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Importance of components:
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Standard deviation 1.2551352 1.1599979 0.9719874 0.8466411 0.64612677
Proportion of Variance 0.3150729 0.2691190 0.1889519 0.1433602 0.08349596
Cumulative Proportion 0.3150729 0.5841919 0.7731438 0.9165040 1.00000000

K-MEANS CLUSTERING: Representation of Clusters Solution



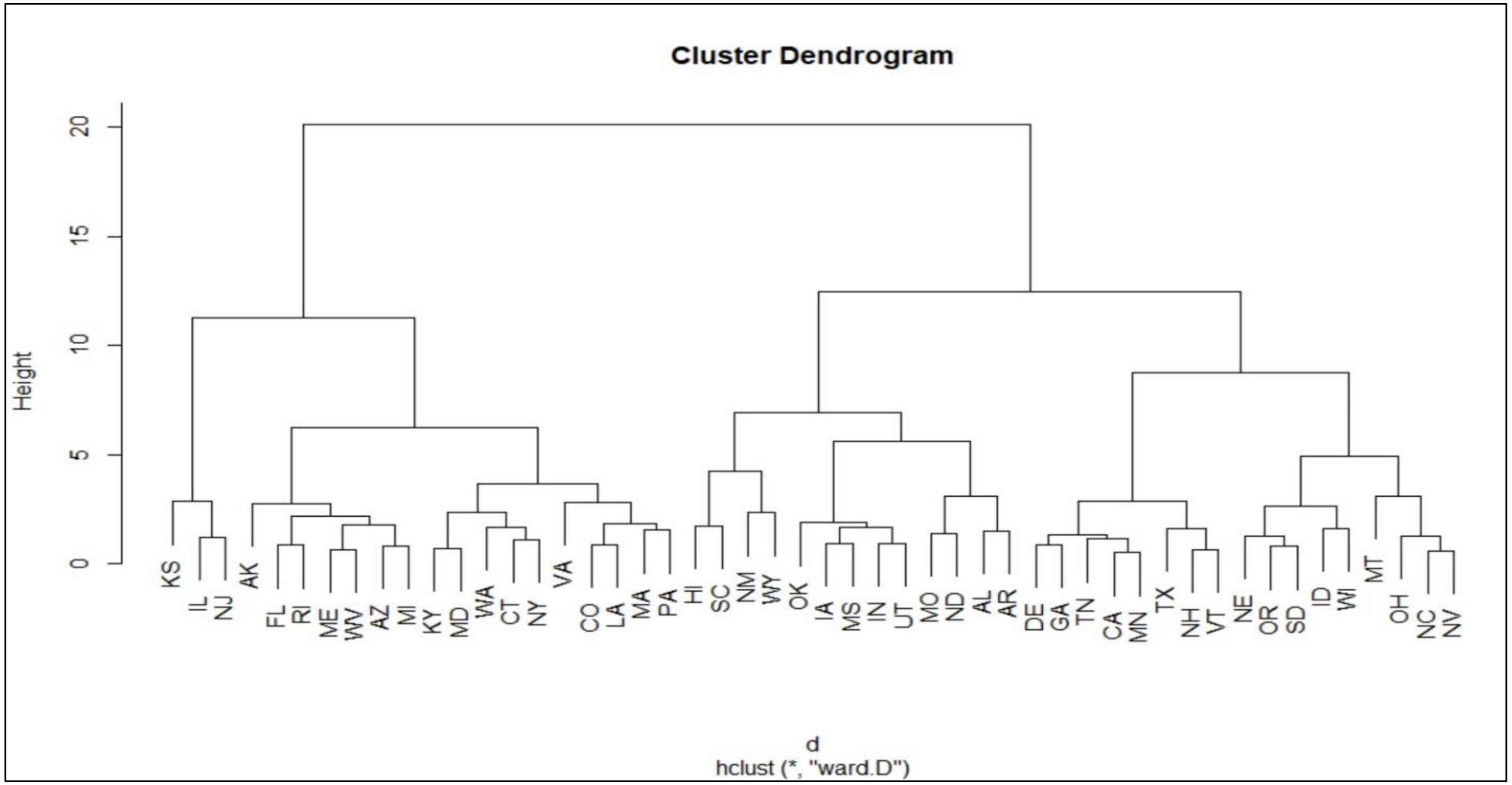
PC1 - the direction of the most variation in the data

CONT'D

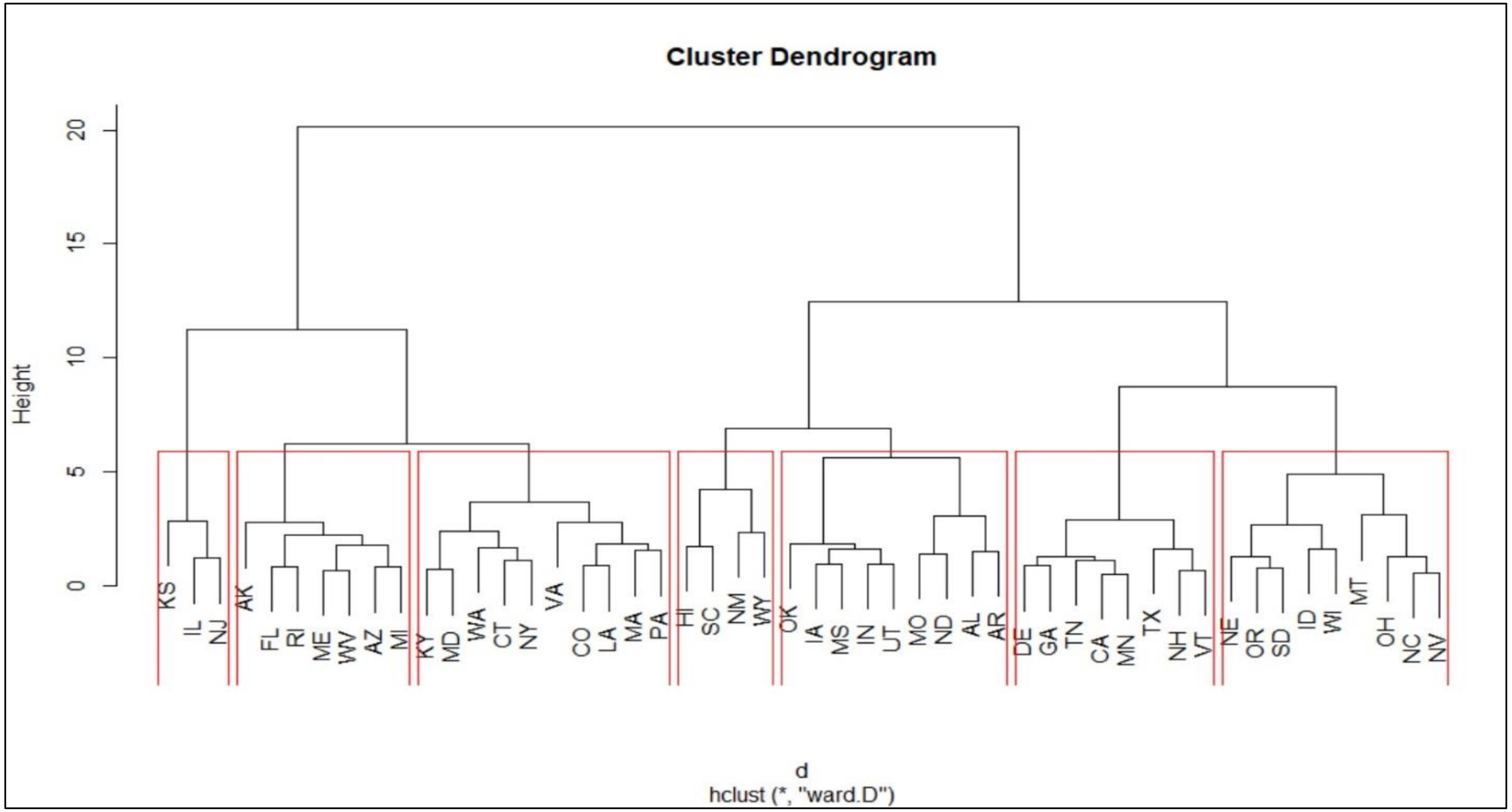
- As shown from the previous figure, the states are clustered as follow (based on their scores of these five variables):
  1. Budget Forecasting.
  2. Budget Maneuvers.
  3. Legacy Costs.
  4. Reserve Funds.
  5. Transparency.

Cluster	Members
#1	ID, SD, NE, IA, UT, OR, WI, OK, MS, NV, NC, MT
#2	NJ, IL, KS
#3	TX, VT, GA, MO, ND, OH, NH
#4	TN, MN, DE, CA, HI, SC, IN
#5	AK, WA, AZ, FL, ME, WV, MI, RI
#6	CT, NY, PA, MA, VA, MD, LA, KY, CO
#7	NM, AL, AR, WY

Hierarchical Clustering: A dendrogram Representation of Clusters Solution



CONT'D


Cluster Dendrogram
Height
0
5
10
15
20
KS IL NJ AK FL RI ME WV AZ MI KY MD WA CT NY VA CO LA MA PA HI SC NM WY OK IA MS IN UT MO ND AL AR DE GA TN CA MN TX NH VT NE OR SD ID WI MT OH NC NV
d
hclust (*, "ward.D")

CONT'D

- As shown from the previous figure, the states are clustered as follow:

Cluster	Members
#1	KS, IL, NJ
#2	AK, FL, RI, ME, WV, AZ, MI
#3	KY, MD, WA, CT, NY, VA, CO, LA, MA, PA
#4	HI, SC, NM, WY
#5	OK, IA, MS, IN, UT, MO, ND, AL, AR
#6	DE, GA, TN, CA, MN, TX, NH, VT
#7	NE, OR, SD, ID, WI, MT, OH, NC, NV

COMPARING CLUSTERING RESULTS

Cluster	K-means	Hierarchical
#1	ID, SD, NE, IA, UT, OR, WI, OK, MS, NV, NC, MT	**KS, IL, NJ**
#2	**NJ, IL, KS**	AK, FL, RI, ME, WV, AZ, MI
#3	TX, VT, GA, MO, ND, OH, NH	KY, MD, WA, CT, NY, VA, CO, LA, MA, PA
#4	TN, MN, DE, CA, HI, SC, IN	HI, SC, NM, WY
#5	AK, WA, AZ, FL, ME, WV, MI, RI	OK, IA, MS, IN, UT, MO, ND, AL, AR
#6	CT, NY, PA, MA, VA, MD, LA, KY, CO	DE, GA, TN, CA, MN, TX, NH, VT
#7	NM, AL, AR, WY	NE, OR, SD, ID, WI, MT, OH, NC, NV

DISCUSSION

- The states that populate each cluster of the hierarchical method are moderately different from *k*-means clusters
  - Except: KS, Ill, NJ
- Their similarities affirm that the clus … are well distributed.
- Man… ised to … with NJ and Ill.
  - Th… budgeta… little publicity about KS.

BUDGET FORECASTING

STATE	GRADE
Alabama	D-
Illinois	D-
Kansas	D-
North Dakota	D-

BUDGET MANEUVERS

STATE	GRADE
Illinois	D
Kansas	D
New Jersey	D
New York	D
Pennsylvania	D
Virginia	D

LEGACY COSTS

STATE	GRADE
Hawaii	D-
Illinois	D-
Kansas	D-
Massachusetts	D-
New Jersey	D-
Pennsylvania	D-
Texas	D-
Virginia	D-
Wyoming	D-

RESERVE FUNDS

STATE	GRADE
Kansas	D
Montana	D
New Mexico	D

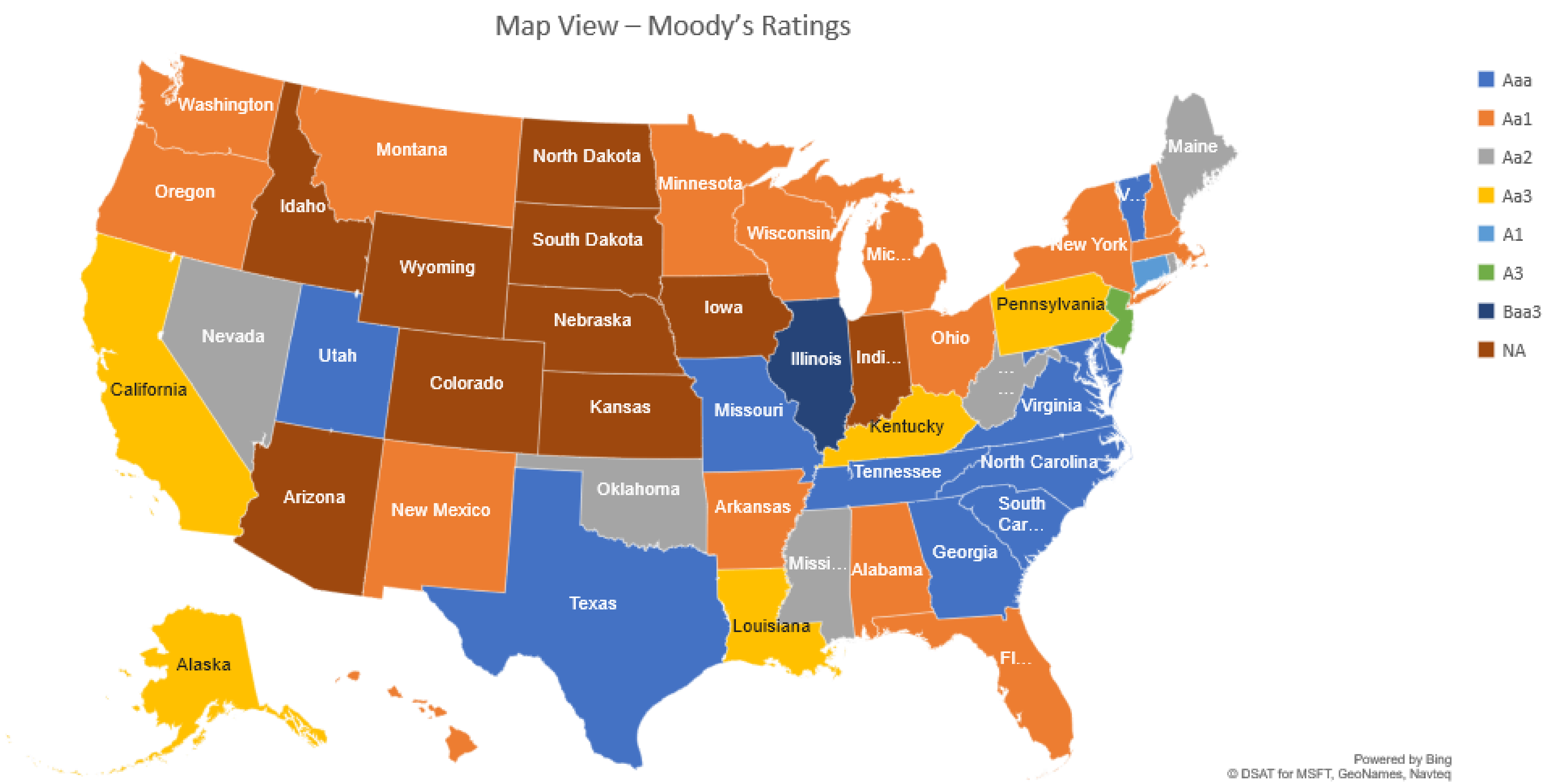
COMPARISONS WITH MOODY'S RATINGS

GRAB

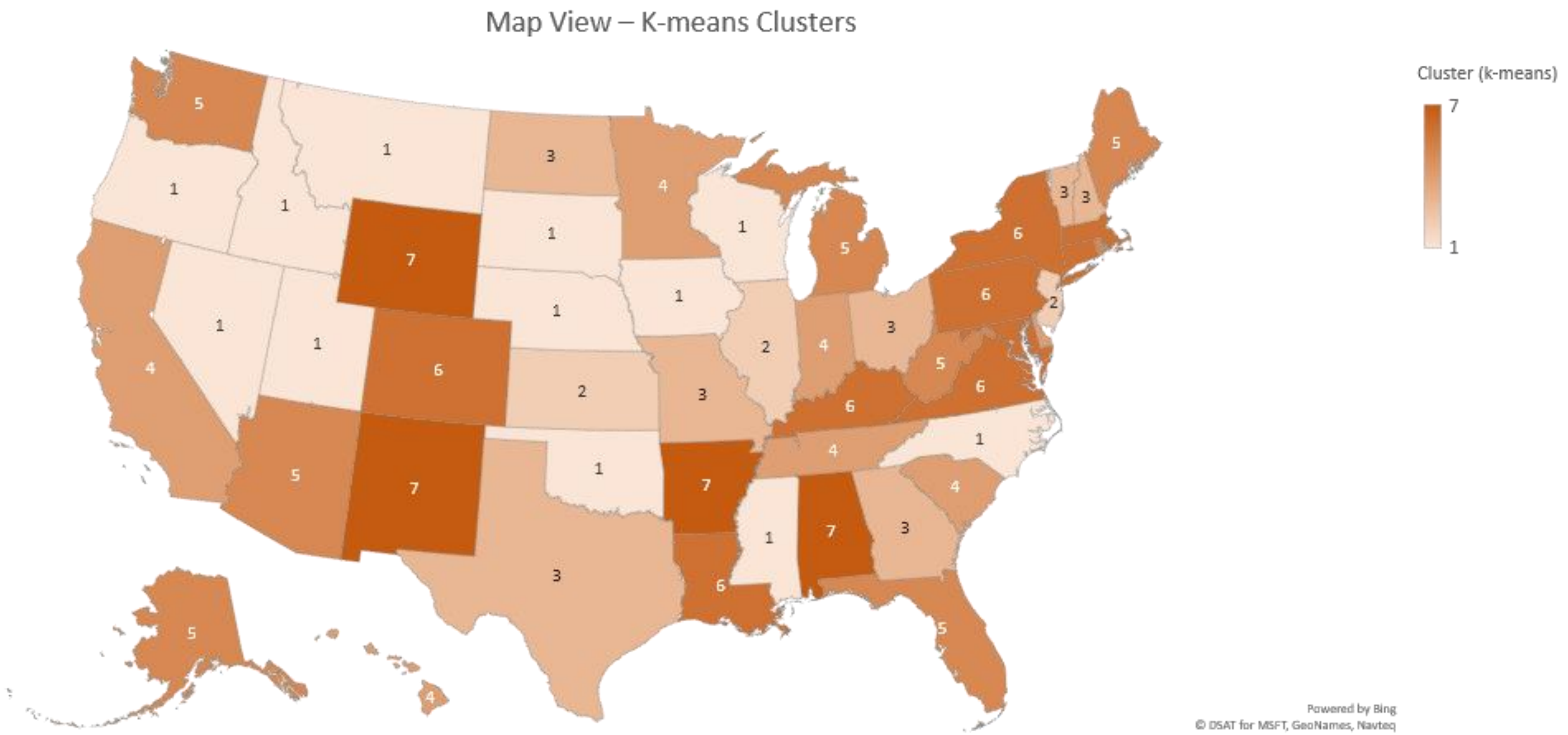
State General Obligation (G.O.) Bond Ratings
U.S. MUNICIPAL SECURITIES
(See MTAX for Individual State Income Tax Rates)

State	Moody's	S&P	State	Moody's	S&P	State	Moody's	S&P
ALABAMA	Aa1	AA	KENTUCKY	Aa3	A+	OHIO	Aa1	AA+
ALASKA	Aa3	AA	LOUISIANA	Aa3	AA-	OKLAHOMA	Aa2	AA
ARIZONA			MAINE	Aa2	AA	OREGON	Aa1	AA+
ARKANSAS	Aa1	AA	MARYLAND	Aaa	AAA	PENNSYLVANIA	Aa3	A+
CALIFORNIA	Aa3	AA-	MASSACHUSETTS	Aa1	AA	PUERTO RICO	Ca	D
COLORADO			MICHIGAN	Aa1	AA-	RHODE ISLAND	Aa2	AA
CONNECTICUT	A1	A+	MINNESOTA	Aa1	AA+	SOUTH CAROLINA	Aaa	AA+
D OF COLUMBIA	Aa1	AA	MISSISSIPPI	Aa2	AA	SOUTH DAKOTA		
DELAWARE	Aaa	AAA	MISSOURI	Aaa	AAA	TENNESSEE	Aaa	AAA
FLORIDA	Aa1	AAA	MONTANA	Aa1	AA	TEXAS	Aaa	AAA
GEORGIA	Aaa	AAA	NEBRASKA			UTAH	Aaa	AAA
GUAM		BB-	NEVADA	Aa2	AA	VERMONT	Aaa	AA+
HAWAII	Aa1	AA+	NEW HAMPSHIRE	Aa1	AA	VIRGIN ISLANDS		
IDAHO			NEW JERSEY	A3	A-	VIRGINIA	Aaa	AAA
ILLINOIS	Baa3	BBB-	NEW MEXICO	Aa1	AA	WASHINGTON	Aa1	AA+
INDIANA			NEW YORK	Aa1	AA+	WEST VIRGINIA	Aa2	AA-
IOWA			NORTH CAROLINA	Aaa	AAA	WISCONSIN	Aa1	AA
KANSAS			NORTH DAKOTA			WYOMING		

CONT'D: Moody's Ratings

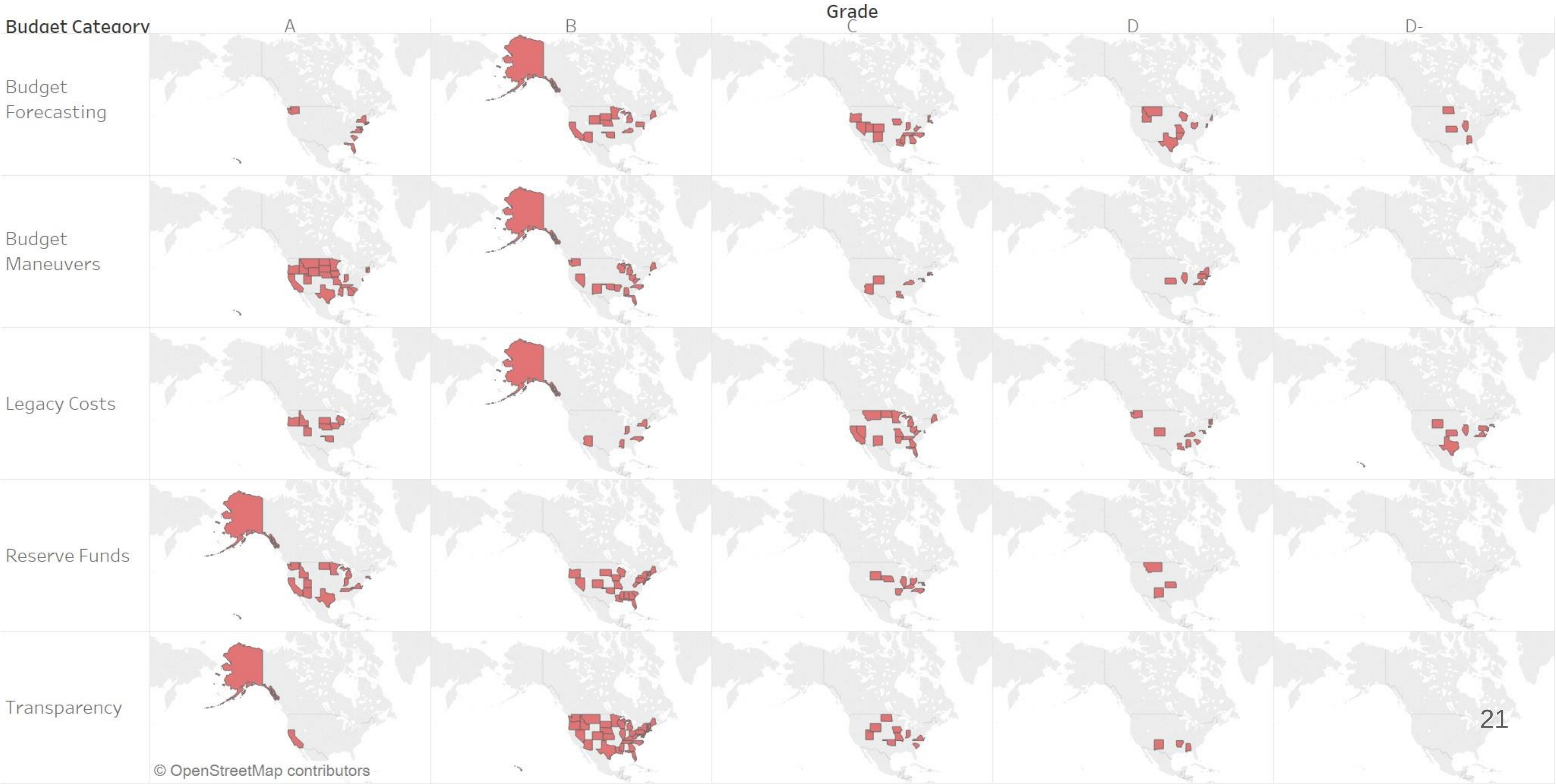


CONT'D: Clustering Results



CONT'D: Volcker's Scores

[states_categories_tablau.twb]



CONCLUSION AND FUTURE WORK

- Cluster analysis is used for grouping and ranking the states.
- Visualization and cluster analysis used in these case studies to get more insight into government data regarding U.S. States financial statements and budgeting.
- The cluster results show that there are some similarities between the two methods, *k*-means and hierarchical, and this could give us an idea about our data quality.
- In addition, we have now clear and unusual patterns and relationships to explore in greater depth.
- Compare the clusters results using external variable, e.g., GDP growth, net population change, public health.
- We plan to explore the literature more on data visualization.