Cluster Analysis on Suicides in the US from 1999-2016

Chris Cole

Why do this?

Why do this?

- Which States stand out?

Why do this?

- Which States stand out?
- Why do they stand out?

Why do this?

- Which States stand out?
- Why do they stand out?
- Have suicides gone up since 1999?

Why do this?

- Which States stand out?
- Why do they stand out?
- Have suicides gone up since 1999?
- Will clusters group states together?

Why do this?

- Which States stand out?
- Why do they stand out?
- Have suicides gone up since 1999?
- Will clusters group states together?
- Commonalities between clusters?

Hypothesis?

Hypothesis?

- States with larger populations (or more suicides) will be clustered together

Our data

Our data

- *N* = 51 (Includes the nation’s capital)

Our data

- *N* = 51 (Includes the nation’s capital)
- Number of Deaths for that year

So what are we going to do?

So what are we going to do?

- Cluster Analysis!

So what are we going to do?

- Cluster Analysis!
- Hierarchical

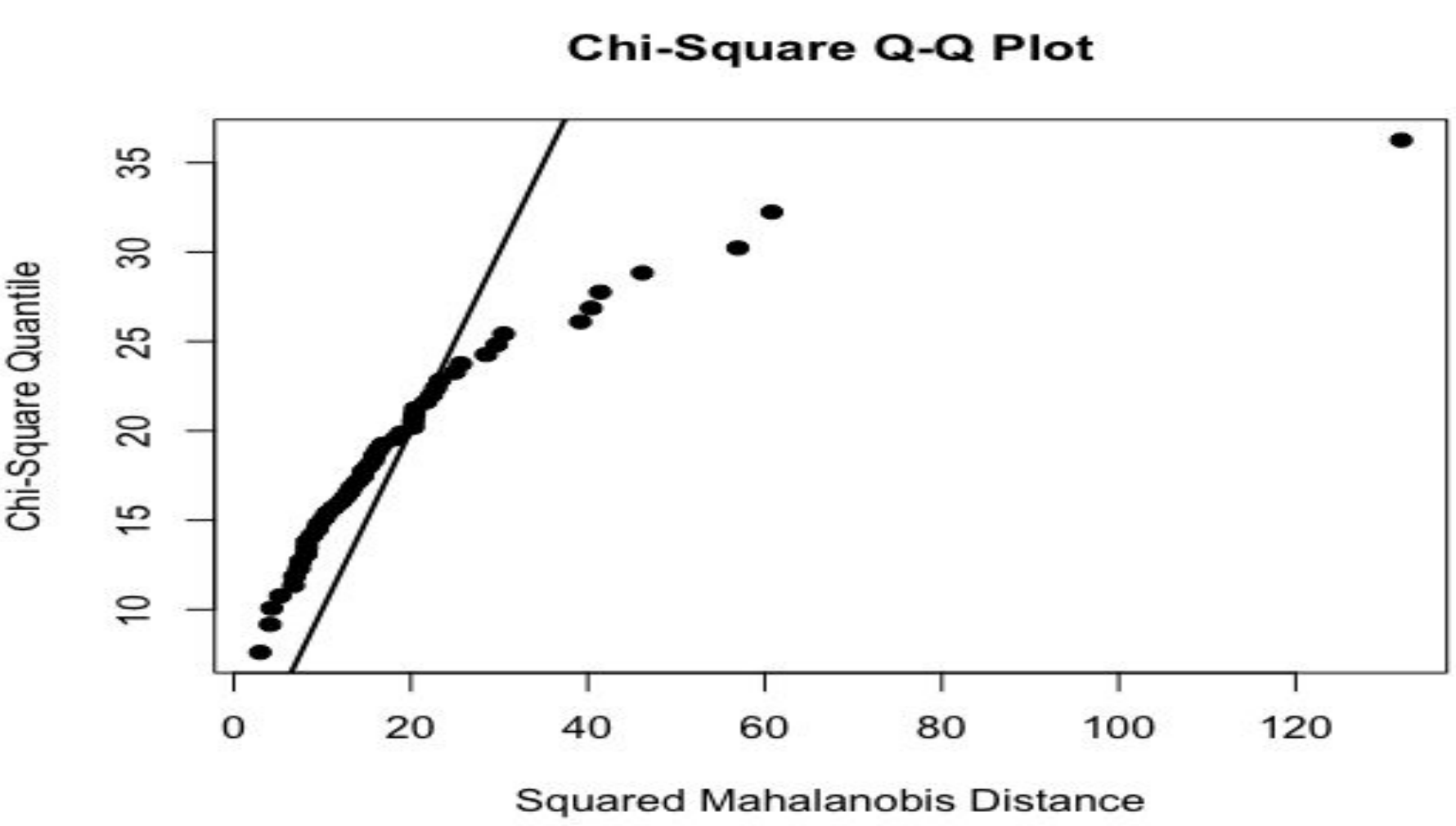
So what are we going to do?

- Cluster Analysis!
- Hierarchical
- K-means

Assumptions?

Assumptions?

- Are the data multivariate normal?


Chi-Square Q-Q Plot
Chi-Square Quantile
10
15
20
25
30
35
0
20
40
60
80
100
120
Squared Mahalanobis Distance

Lets get started...

Lets get started...

- Let’s get the means for number of deaths per year

Lets get started...

- Let's get the means for number of deaths per year

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Death Count	572.529	575.490	600.431	620.686	617.333	636.059	639.941	652.941	678.392

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016
Death Count	706.569	723.706	752.235	774.863	796.078	806.843	839.725	866.529	881.667

Lets get started...

- Now let’s get the ranges for number of deaths per year

Lets get started...

- Now let's get the ranges for number of deaths per year

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Range of Deaths	3047	2946	2791	3197	3361	3335	3173	3304	3566

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016
Range of Deaths	3732	3794	3872	3959	3856	3987	4162	4133	4254

Standardize Variables

Standardize Variables

- First we need the number of deaths for each state for each year

Standardize Variables

- First we need the number of deaths for each state for each year
- Then divide that by the range (max-min) of that specific year

Standardize Variables

- First we need the number of deaths for each state for each year
- Then divide that by the range (max-min) of that specific year
- Do this 18 times...

Hierarchical Analysis

Hierarchical Analysis

- Euclidean Distance

Hierarchical Analysis

- Euclidean Distance
- Ward Method

Stopping rules...

Stopping rules...

1. Duda

Stopping rules...

1. Duda

Number of Clusters	Duda_Critical	Duda_Value
2	.719	.143
3	.783	.263
4	.696	.290
5	.732	.443
6	.363	.012
7	.539	.217
8	.617	.434
9	.676	.304
10	.650	.402
11	.617	.346
12	.498	1.246
13	.363	.437
14	.443	.333
15	.228	5.866

Stopping rules...

1. Duda
2. CH-Index

Stopping rules...

1. Duda
2. CH-Index

Number of Clusters	CHindex
2	51.175
3	185.638
4	190.257
5	214.022
6	207.767
7	487.420
8	612.349
9	680.482
10	673.418
11	680.196
12	718.734
13	767.039
14	874.342
15	940.256

Stopping rules...

1. Duda
2. CH-Index
3. C-Index

Stopping rules...

1. Duda
2. CH-Index
3. C-Index

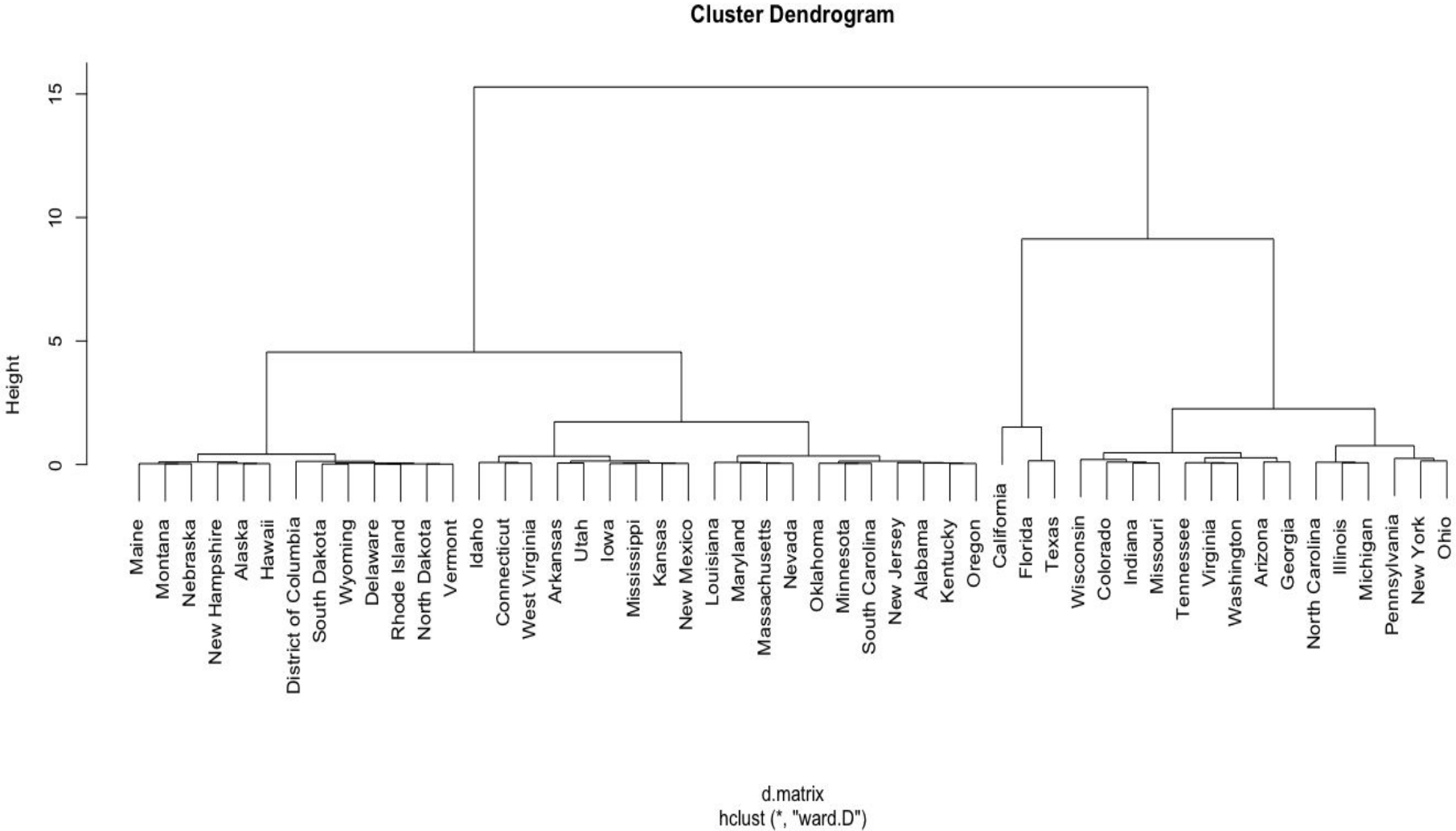
Number of Clusters	Cindex
2	.119
3	.240
4	.170
5	.129
6	.092
7	.098
8	.088
9	.078
10	.075
11	.067
12	.058
13	.053
14	.050
15	.046

Stopping rules...

1. Duda
2. CH-Index
3. C-Index
4. “Best” partition

Stopping rules...

1. Duda
2. CH-Index
3. C-Index
4. “Best” partition
5. Dendrogram


Cluster Dendrogram
Height
0
5
10
15
Maine
Montana
Nebraska
New Hampshire
Alaska
Hawaii
District of Columbia
South Dakota
Wyoming
Delaware
Rhode Island
North Dakota
Vermont
Idaho
Connecticut
West Virginia
Arkansas
Utah
Iowa
Mississippi
Kansas
New Mexico
Louisiana
Maryland
Massachusetts
Nevada
Oklahoma
Minnesota
South Carolina
New Jersey
Alabama
Kentucky
Oregon
California
Florida
Texas
Wisconsin
Colorado
Indiana
Missouri
Tennessee
Virginia
Washington
Arizona
Georgia
North Carolina
Illinois
Michigan
Pennsylvania
New York
Ohio
d.matrix
hclust (*, "ward.D")

So how many clusters do we choose?

So how many clusters do we choose?

- Let’s stick with 3 clusters (for now)

K-means

K-means

- Euclidean

Stopping rules...

Stopping rules...

1. CH-Index

Stopping rules...

1. CH-Index

Number of Clusters	kCHindex
2	98.971
3	185.638
4	212.794
5	263.110
6	409.560
7	512.131
8	472.612
9	426.247
10	683.995
11	616.904
12	556.666
13	499.669
14	534.792
15	583.466

Stopping rules...

1. CH-Index
2. C-Index

Stopping rules...

1. CH-Index
2. C-Index

Number of Clusters	kCindex
2	.293
3	.240
4	.164
5	.189
6	.130
7	.097
8	.095
9	.095
10	.075
11	.075
12	.077
13	.081
14	.073
15	.064

Stopping rules...

1. CH-Index
2. C-Index
3. “Best” Partition

Make centroids based on our HCA

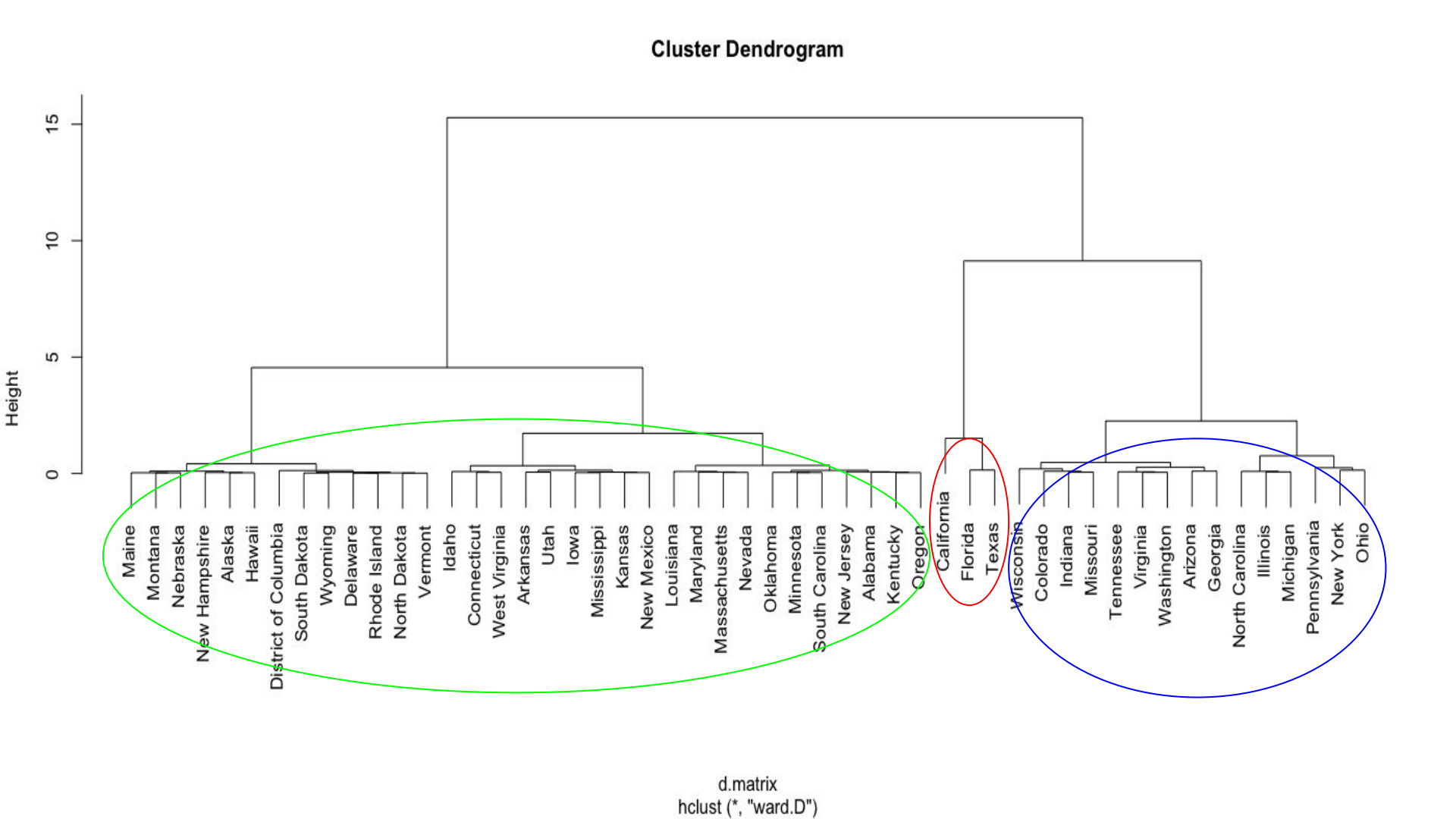
Make centroids based on our HCA

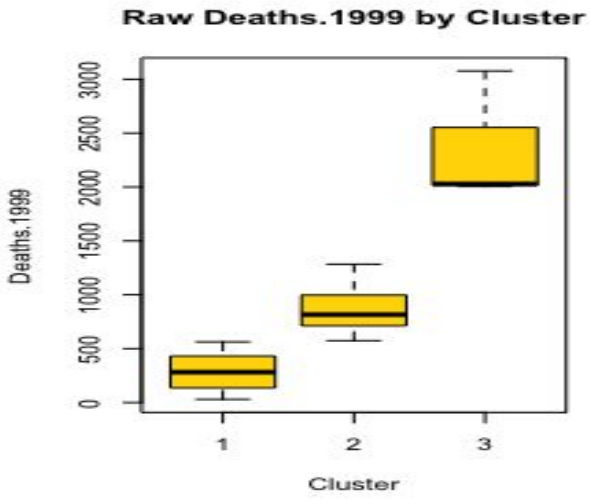
- Use K-means with the center being our three centroids

Clusters	States
Cluster 1	California, Florida, Texas
Cluster 2	Arizona, Colorado, Georgia, Illinois, Indiana, Michigan, Missouri, New York, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, Washington, Wisconsin
Cluster 3	Alabama, Alaska, Arkansas, Connecticut, Delaware, District of Columbia, Hawaii, Idaho, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Minnesota, Mississippi, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, North Dakota, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Utah, Vermont, West Virginia, Wyoming

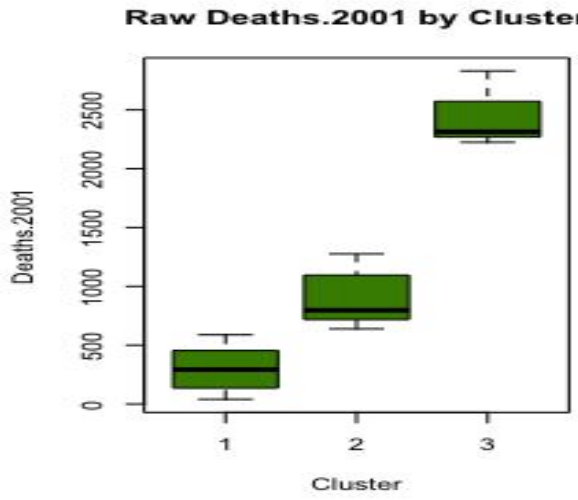

Cluster Dendrogram
Height
0
5
10
15
Maine
Montana
Nebraska
New Hampshire
Alaska
Hawaii
District of Columbia
South Dakota
Wyoming
Delaware
Rhode Island
North Dakota
Vermont
Idaho
Connecticut
West Virginia
Arkansas
Utah
Iowa
Mississippi
Kansas
New Mexico
Louisiana
Maryland
Massachusetts
Nevada
Oklahoma
Minnesota
South Carolina
New Jersey
Alabama
Kentucky
Oregon
California
Florida
Texas
Wisconsin
Colorado
Indiana
Missouri
Tennessee
Virginia
Washington
Arizona
Georgia
North Carolina
Illinois
Michigan
Pennsylvania
New York
Ohio
d.matrix
hclust (*, "ward.D")

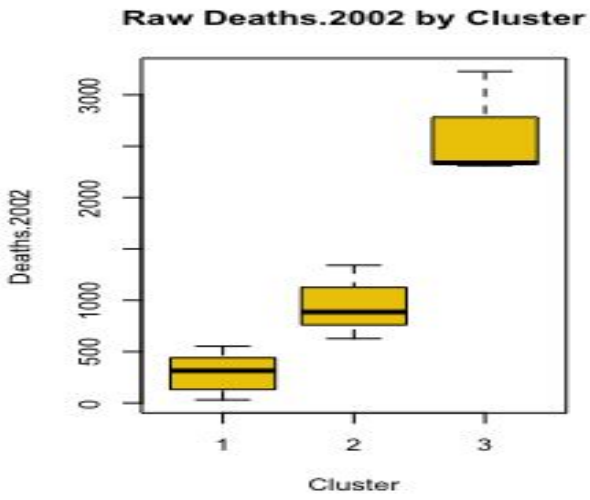

Cluster Dendrogram
Height
0
5
10
15
Maine
Montana
Nebraska
New Hampshire
Alaska
Hawaii
District of Columbia
South Dakota
Wyoming
Delaware
Rhode Island
North Dakota
Vermont
Idaho
Connecticut
West Virginia
Arkansas
Utah
Iowa
Mississippi
Kansas
New Mexico
Louisiana
Maryland
Massachusetts
Nevada
Oklahoma
Minnesota
South Carolina
New Jersey
Alabama
Kentucky
Oregon
California
Florida
Texas
Wisconsin
Colorado
Indiana
Missouri
Tennessee
Virginia
Washington
Arizona
Georgia
North Carolina
Illinois
Michigan
Pennsylvania
New York
Ohio
d.matrix
hclust (*, "ward.D")

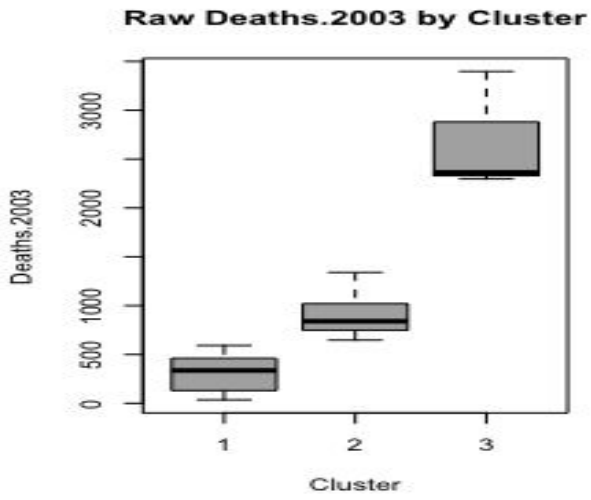

Cluster Dendrogram
Height
0
5
10
15
Maine
Montana
Nebraska
New Hampshire
Alaska
Hawaii
District of Columbia
South Dakota
Wyoming
Delaware
Rhode Island
North Dakota
Vermont
Idaho
Connecticut
West Virginia
Arkansas
Utah
Iowa
Mississippi
Kansas
New Mexico
Louisiana
Maryland
Massachusetts
Nevada
Oklahoma
Minnesota
South Carolina
New Jersey
Alabama
Kentucky
Oregon
California
Florida
Texas
Wisconsin
Colorado
Indiana
Missouri
Tennessee
Virginia
Washington
Arizona
Georgia
North Carolina
Illinois
Michigan
Pennsylvania
New York
Ohio
d.matrix
hclust (*, "ward.D")

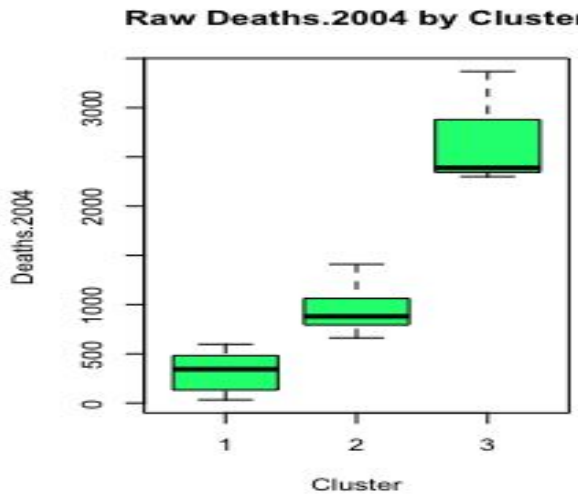

Raw Deaths.1999 by Cluster
Deaths.1999
0
500
1000
1500
2000
2500
3000
1
2
3
Cluster

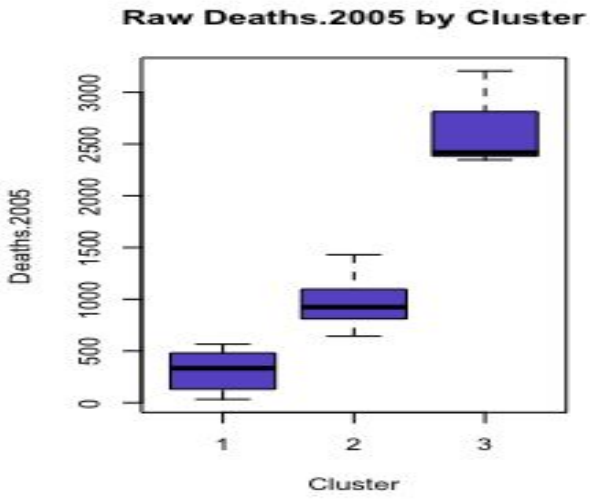

Raw Deaths.2000 by Cluster
Deaths.2000
0
500
1000
1500
2000
2500
3000
1
2
3
Cluster


Raw Deaths.2001 by Cluster
Deaths.2001
0
500
1000
1500
2000
2500
1
2
3
Cluster

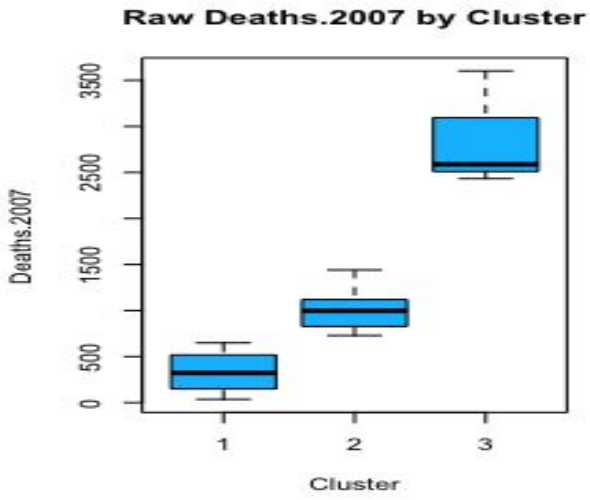

Raw Deaths.2002 by Cluster
Deaths.2002
0
500
1000
2000
3000
1
2
3
Cluster

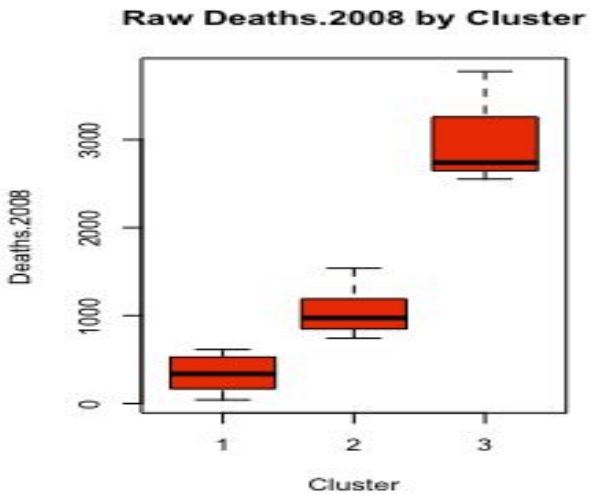

Raw Deaths.2003 by Cluster
Deaths.2003
0
500
1000
2000
3000
1
2
3
Cluster

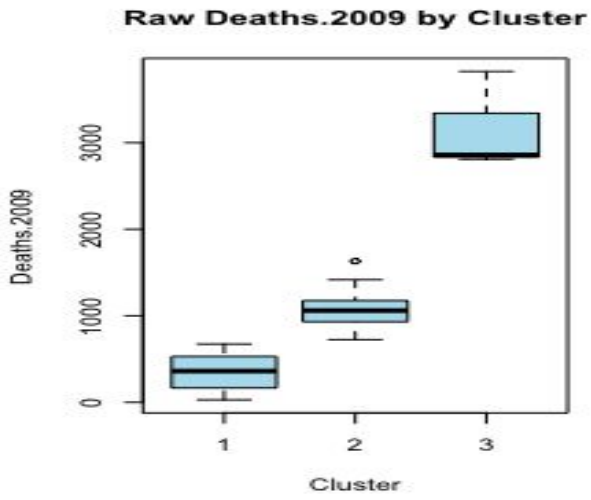

Raw Deaths.2004 by Cluster
Deaths.2004
0
500
1000
2000
3000
1
2
3
Cluster

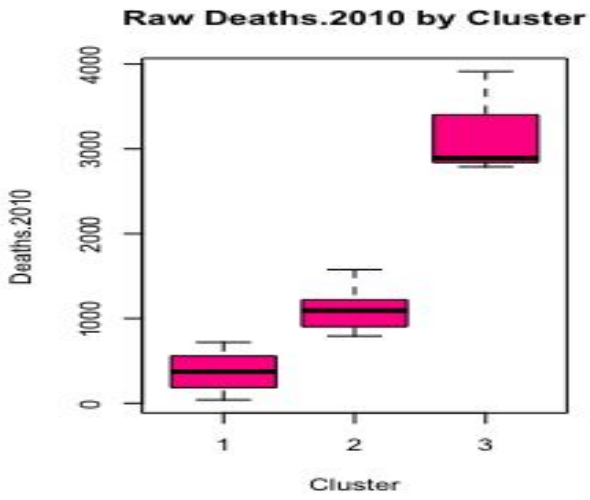

Raw Deaths.2005 by Cluster
Deaths.2005
0
500
1000
1500
2000
2500
3000
1
2
3
Cluster

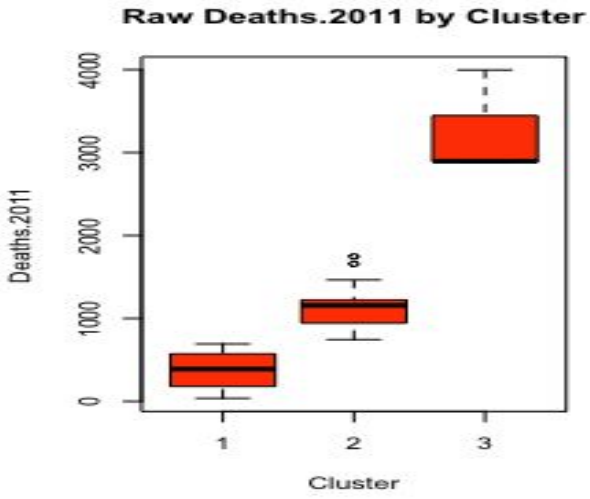

Raw Deaths.2006 by Cluster
Deaths.2006
0
500
1000
2000
3000
1
2
3
Cluster

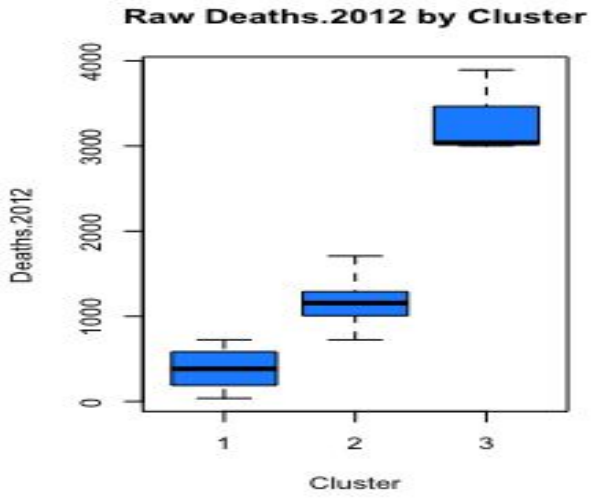

Raw Deaths.2007 by Cluster
Deaths.2007
0
500
1500
2500
3500
1
2
3
Cluster

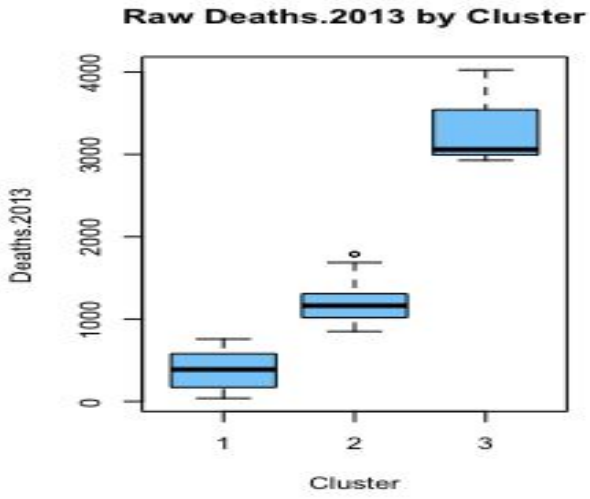

Raw Deaths.2008 by Cluster
Deaths.2008
0
1000
2000
3000
1
2
3
Cluster

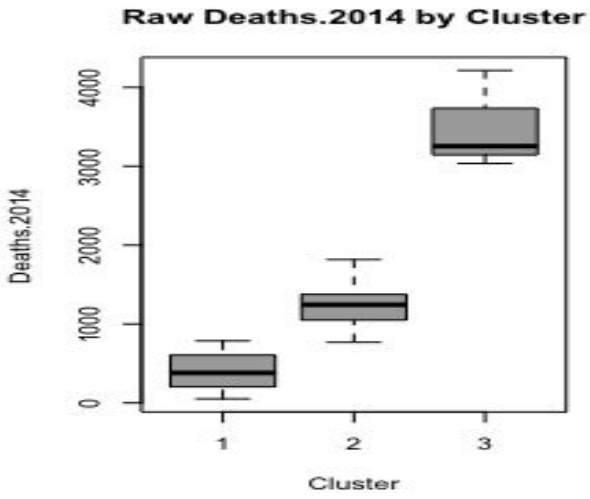

Raw Deaths.2009 by Cluster
Deaths.2009
0
1000
2000
3000
1
2
3
Cluster

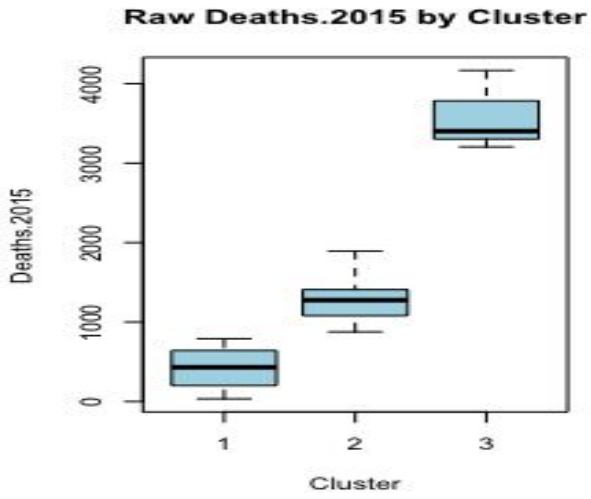

Raw Deaths.2010 by Cluster
Deaths.2010
0
1000
2000
3000
4000
1
2
3
Cluster

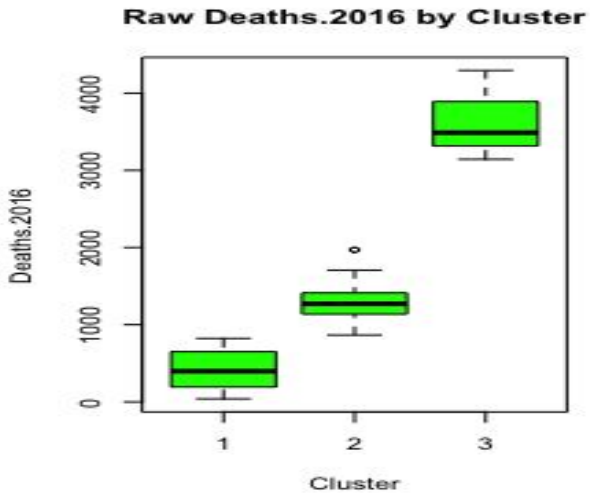

Raw Deaths.2011 by Cluster
Deaths.2011
0
1000
2000
3000
4000
1
2
3
Cluster


Raw Deaths.2012 by Cluster
Deaths.2012
0
1000
2000
3000
4000
1
2
3
Cluster


Raw Deaths.2013 by Cluster
Deaths.2013
0
1000
2000
3000
4000
1
2
3
Cluster


Raw Deaths.2014 by Cluster
Deaths.2014
0
1000
2000
3000
4000
1
2
3
Cluster


Raw Deaths.2015 by Cluster
Deaths.2015
0
1000
2000
3000
4000
1
2
3
Cluster


Raw Deaths.2016 by Cluster
Deaths.2016
0
1000
2000
3000
4000
1
2
3
Cluster

Why do this?

- Which States stand out?
- Why do they stand out?
- Have suicides gone up since 1999?
- Will clusters group states together?

Hypothesis?

- States with larger populations (or more suicides) will be clustered together

Why does this matter?

Questions?