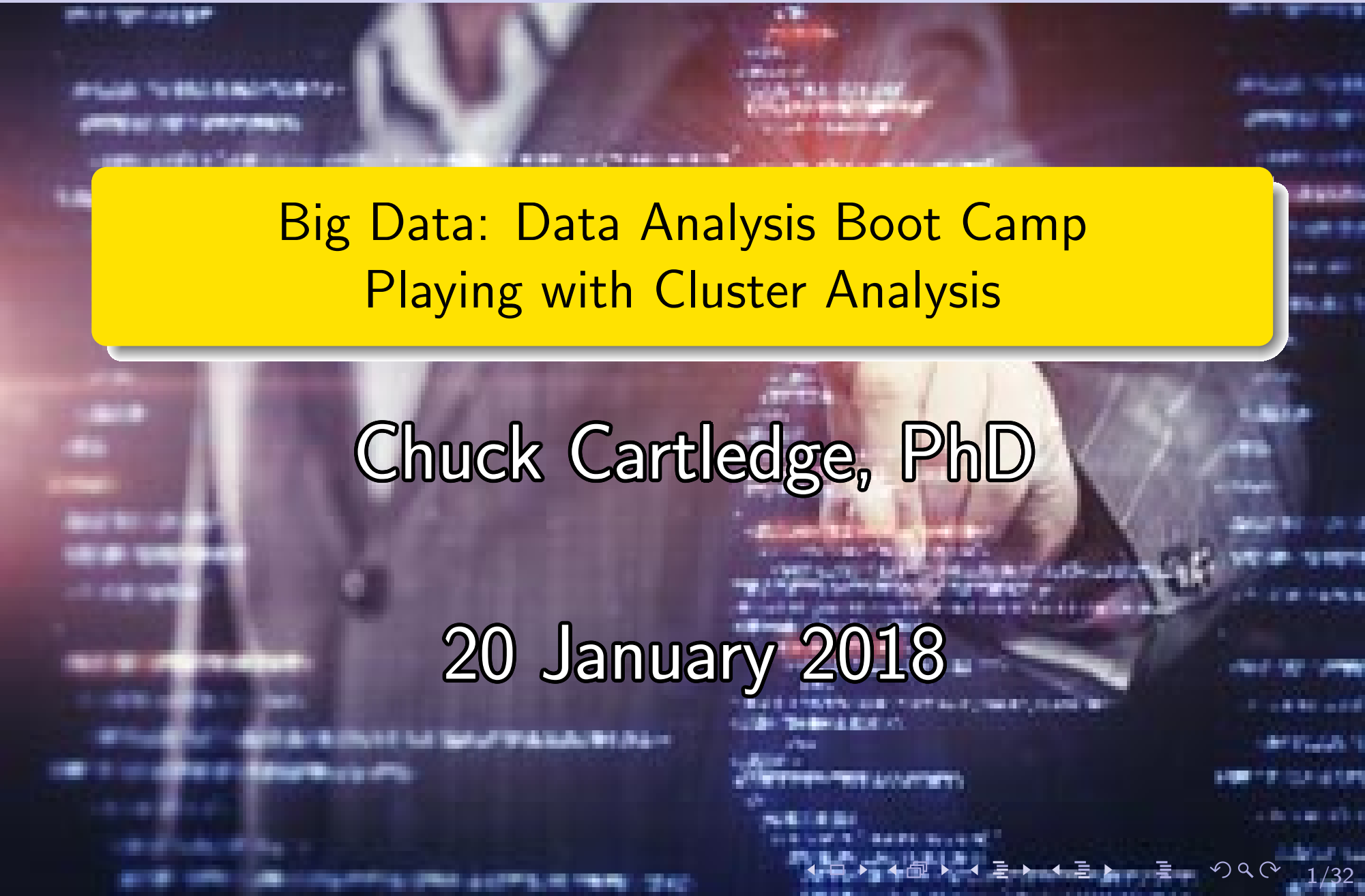


Big Data: Data Analysis Boot Camp
Playing with Cluster Analysis

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

What are we going to cover?

We're going to talk about:

- Cluster analysis, and what its good for.
- How there is more than one way to measure distance.
- If you have a lot of data, what is the correct number of clusters?

Lots of words.

"Cluster analysis is a statistical technique used to identify how various units – like people, groups, or societies – can be grouped together because of characteristics they have in common. Also known as clustering, it is an exploratory data analysis tool that aims to sort different objects into groups in such a way that when they belong to the same group they have a maximal degree of association and when they do not belong to the same group their degree of association is minimal.

Unlike some other statistical techniques, the structures that are uncovered through cluster analysis need no explanation or interpretation – it discovers structure in the data without explaining why they exist."

A. Crossman [1]

Picking it apart.

- "...can be grouped together..." – implies a way to compare different pieces of data
- "...sort different objects into groups..." – decide group membership
- "...the structures that are uncovered..." – the cluster analysis has no preconceived ideas
- "...discovers structure in the data without explaining why they exist." – clusters may in fact not exist

A little deeper.

- Need to define characteristics necessary to define group membership
- Need to define order that items are considered for group membership
- Need to define how many groups/clusters there are
- Recognize that group membership may not have meaning

Down the "rabbit hole"

- What are the important items to use to define group membership (size, weight, time, location, textual content, . . .)
- Adding a new member changes the "characteristics" of the group, so adding new members in different order may result in different groups
- How to choose number of groups? Easy cases are: 1 (all members belongs to a single group), and n (each member belongs to its own group). Selecting the number of groups between 2 and $n - 1$ is hard.

One of these is easy, the others are hard.

As Alice falls further . . .

- How to determine when to add a new member?
  1. One at a time (selected in order, or random, and if random then how to ensure results are repeatable)
  2. All at once by keeping two copies of the current group membership
- How to choose number of groups?
  1. Have subject matter expert (SME) provide guidance
  2. Brute force from 1 to n
  3. How to know the "right" number of groups

Slippery slopes

- How to define the center of a cluster?
  1. Median or mean of all members (may not match a group member)
  2. Select group member nearest median or mean
- How to measure distance from candidate member to cluster?
  1. Over 1,000 different ways to measure distance[2]
  2. Measure distance to:
     1. Cluster center
     2. Closest cluster member
     3. All cluster members

What to do about units of measurement?

If interested in clustering people based on their height, weight, and waist, then

- Can't directly compare weight and others attributes (different units)
- Can't directly compare height and waist (different ranges of values)
- How to make a cluster out of things that have more than 2 attributes?

Simple approaches to handling numerical data

- Convert all attribute data to the same units (feet to inches, pounds to ounces, etc.)
- Normalize the data between 0 and 1:

$$x_{normalized} = \frac{x_{raw} - min(x_{all})}{max(x_{all}) - min(x_{all})}$$

- Compute the $z - score$ for a data point. A $z - score$ is the number of standard deviations from the mean a data point is.

$$z - score = \frac{x_{raw} - \mu(x_{all})}{\sigma(x_{all})}$$

Simple numerical distances

Based on the L_p notation.
If we have two vectors $x = (x_1, x_2, x_3, ..., x_n)$ and $y = (y_1, y_2, y_3, ..., y_n)$, then the distance between the two is:

$$d(x, y) = \|x - y\| = \left(\sum_1^n |x_i - y_i|^r\right)^{\frac{1}{r}}$$

Where r is chosen by the user.

- $r = 1$ the Manhattan distance (the number of city blocks you have to walk to get from one place to another), sometimes also known as the Hamming distance[3]
- $r = 2$ standard Euclidean distance
- $r = \infty$ Supermum, the maximum difference between any attribute of the vectors

Simple approaches to handing textual data

We are interested in a few things:

- How often does the term t appear in a document d
- How many documents N in the corpus D

Combining those ideas towards finding "important" terms in the corpus.
A little math:

$$
\begin{aligned}
f(t,d) &= \textbf{frequency of term } t \textbf{ in document } d \\
tf(t,d) &= 1 + \log(f(t,d))(\textbf{log term frequency}) \\
idf(t,D) &= \log(\frac{N}{d \epsilon D : t \epsilon d})(\textbf{inverse document frequency}) \\
tfidf(t,d,D) &= tf(t,d) * idf(t,D)
\end{aligned}
$$

Problems (opportunities) unique to numerical data.

There are a limited number of things you can do with missing numerical data.

- You can ignore the entire "data record"
- You can insert "false" data, either:
  - Compute the average of all "good" data
  - Insert the mode of all "good" data
  - Compute a random number based on the statistics of the "good" data

Options for dealing with "bad"/missing data are limited.

Problems (opportunities) unique to textual data.

There are some interesting problems when dealing with text.

- Capitalizations
- Prefixes and suffixes
- Words that are too common (articles, conjunctives, etc.)
- How words are used (chapter headings, footers, titles, captions, etc.)

Textual analysis presents all sorts of opportunities.

Overview of processing

1. Randomly assign each data member to a cluster
2. Until exit condition met
   1. Compute the distance from each member to all cluster centers
   2. Assign each member to new cluster
   3. Compute new cluster center

Exit conditions can include:

- No members move between clusters
- A maximum number of loops are executed
- Movements are too small to affect overall solution

Source code from the text

Need to do a couple of things:

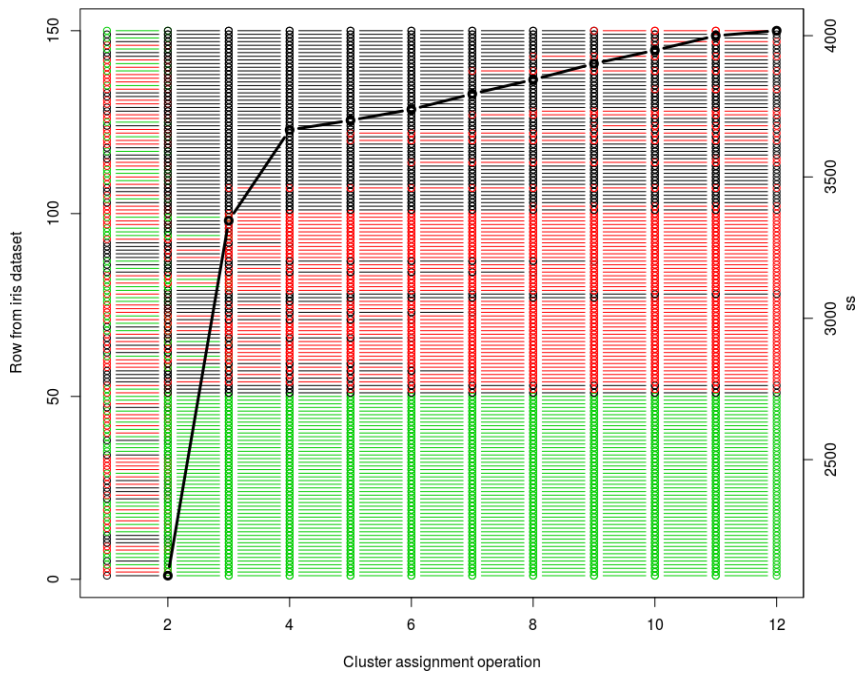
1. Load "chapter-04.R" into the editor
2. Highlight and execute the entire file
3. Understand the output

	1	**2**	**3**
setosa	0	0	50
versicolor	3	47	0
virginica	36	14	0

Clustering was able to correctly label 89% ($0.89 = \frac{50+47+36}{50+47+36+3+14}$) of the flowers.

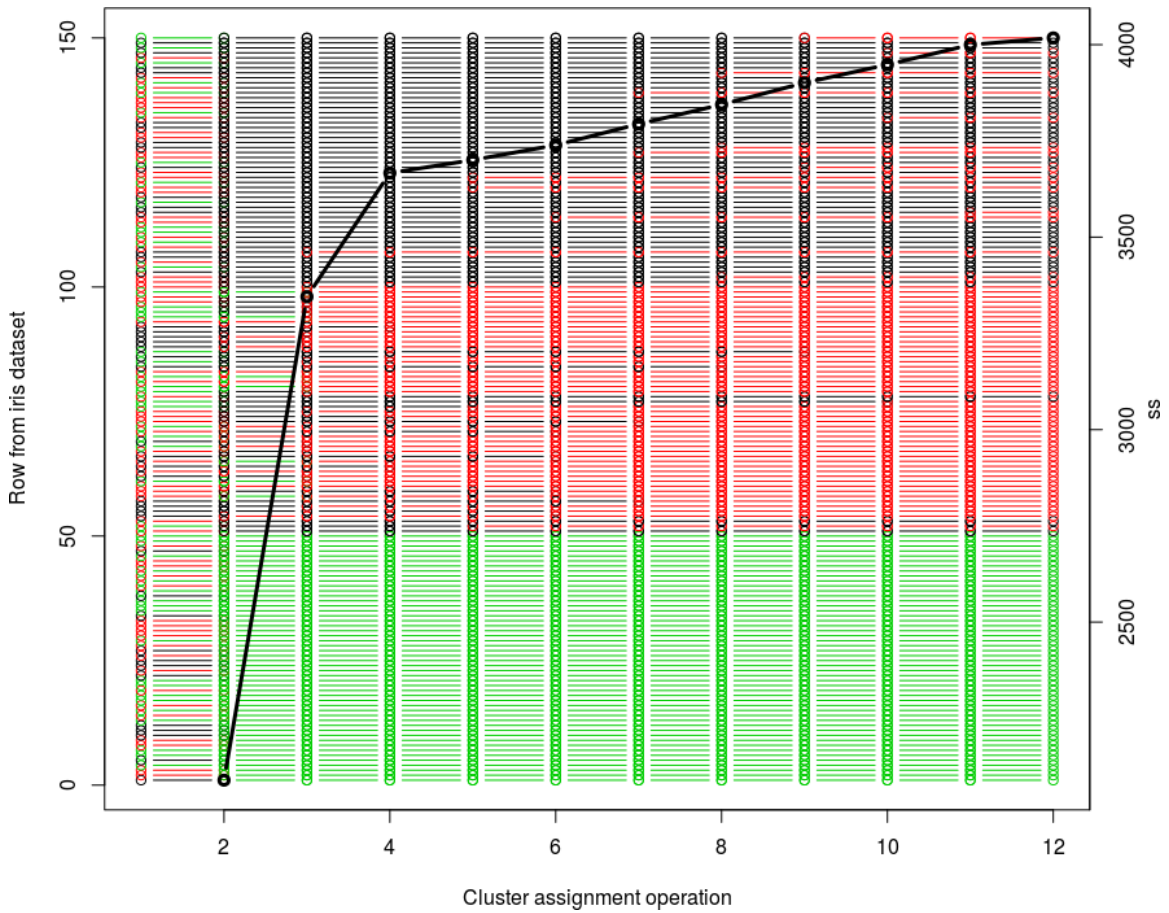
A picture is worth . . .

The clustering algorithm has lots of moving parts, and it would be useful to see them in action.
Load "chapter-04-iris-cluster.R" into the editor and run.
Changes in color mean that an iris data row changed cluster assignment.



Attached file.

Same image.



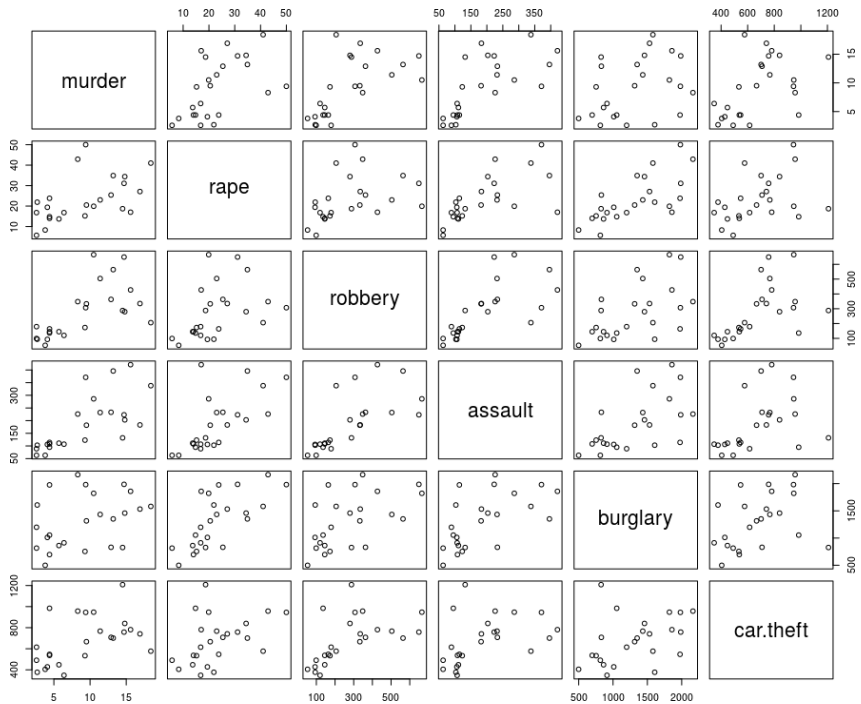
Attached file.

Some basics

There is an R library with some crime statistics from 1970.

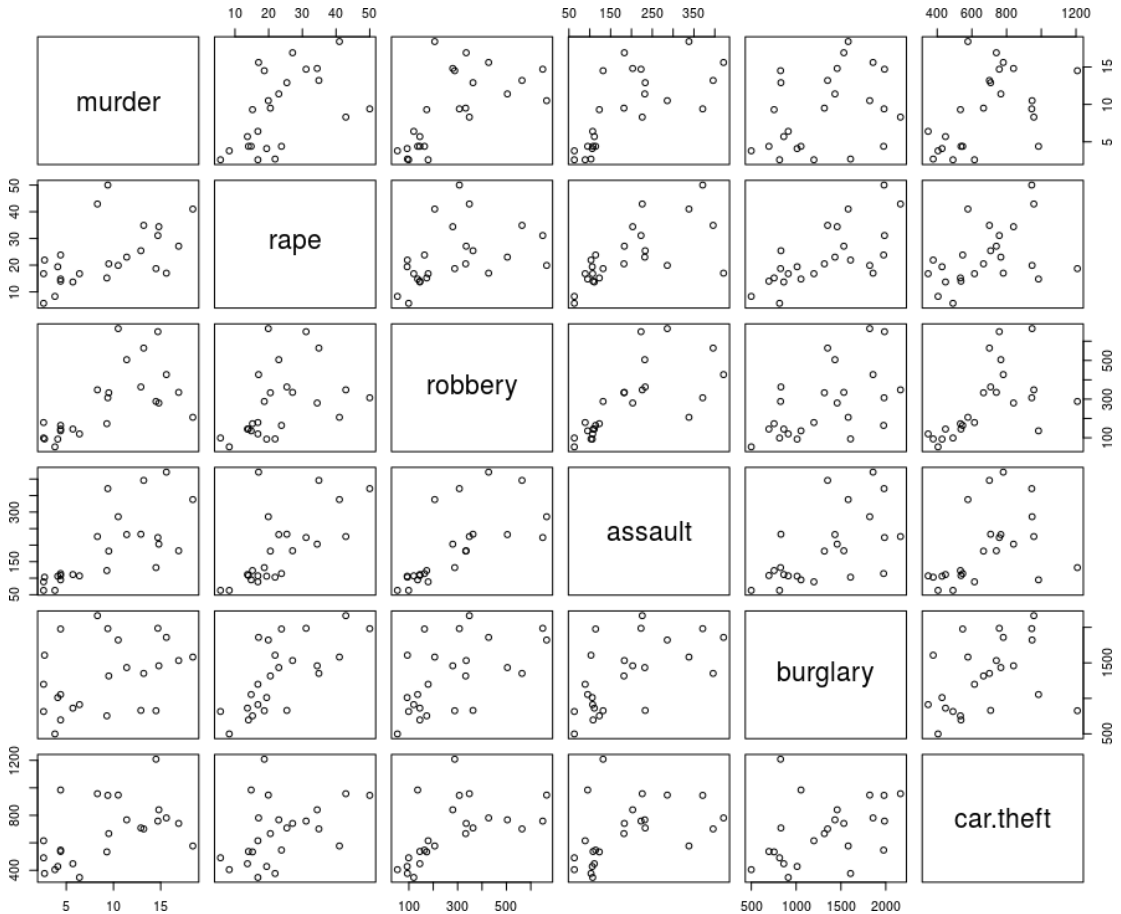
```
library(cluster.datasets)
data(all.us.city.crime.1970)
crime =
all.us.city.crime.1970
plot(crime[5:10])
```

Pretty basic stuff.



Can we find clusters in the data?

Same image.



Can we find clusters in the data?

Copy and paste into editor:

```
crime.scale = data.frame(scale(crime[5:10]))
set.seed(234)
TwoClusters = kmeans(crime.scale, 2, nstart = 25)
plot(crime[5:10],col=as.factor(TwoClusters\$cluster), main = "2-cluster solution")
ThreeClusters = kmeans(crime.scale, 3, nstart = 25)
plot(crime[5:10],col=as.factor(ThreeClusters\$cluster), main = "3-cluster solution")
FourClusters= kmeans(crime.scale, 4, nstart = 25)
plot(crime[5:10],col=as.factor(FourClusters\$cluster), main = "4-cluster solution")
FiveClusters = kmeans(crime.scale, 5, nstart = 25)
plot(crime[5:10],col=as.factor(FiveClusters\$cluster), main = "5-cluster solution")
```

Peeking into the kmeans object

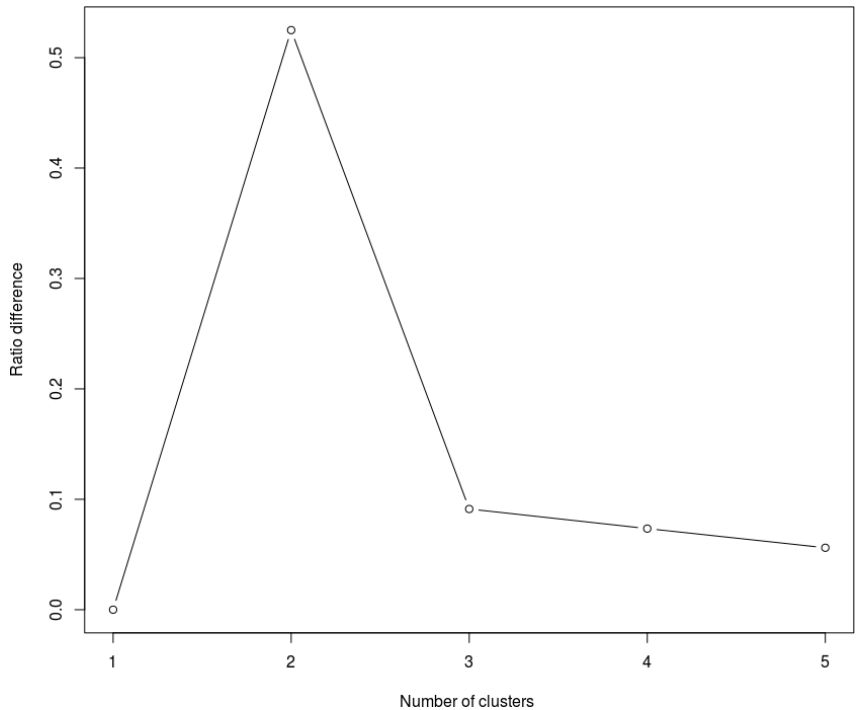
Typing the name of the kmeans object prints out its values:

- **K-means** number of clusters and their size
- **Cluster means** the centroid coordinates when kmeans() ended
- **Clustering vector** the cluster to which each element belongs
- **Within cluster sum of squares by cluster** sum of the squares of the distance of each member from the centroid
- **Available components** components that can be accessed by name or [[]] notation

In the editor, type: `TwoClusters`

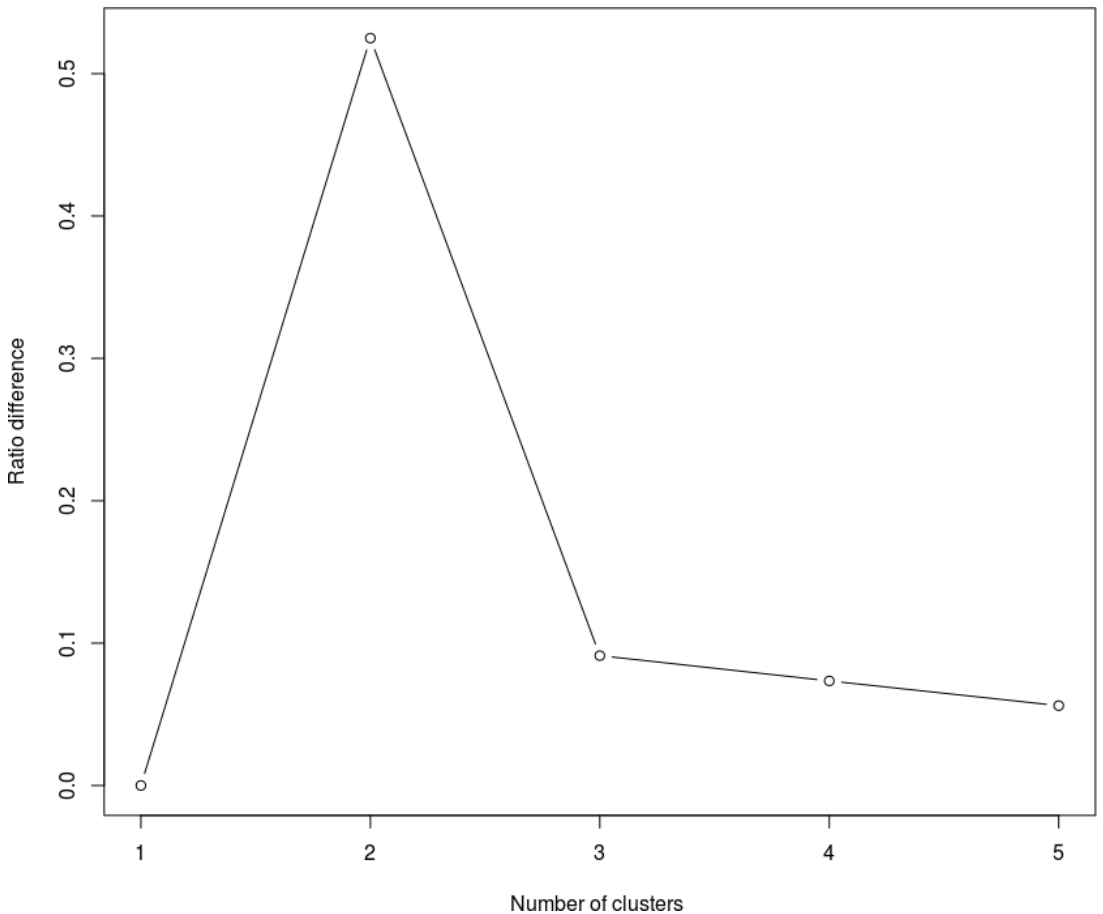
How many clusters are needed?

A way to determine the optimal number of clusters is to vary k and evaluate the output.
The text uses incremental improvement in the ratio between each $k's$ betweeness sum of squares, and the total sum of squares.
The data plotted in the text is inaccurate.



Attached file.
2 clusters has the best ratio.

Same image.



Attached file.

2 clusters has the best ratio.

More explorations into how many clusters are needed.

- The text continues the exploration of how many clusters are needed by looking at life expectancy data from the `cluster.datasets` library
- The `NbClust` function from the `NbClust` library uses a collection of techniques to arrive at a number of clusters
- `NbClust()` then recommends the number of clusters that gets the most votes

Does `NbClust` make a difference?

Yes.
Using the data life expectancy data from the text, and different arguments to the `NbClust` function, different numbers of clusters are determined.

Method	**Distance measurement**			
	euclidean	maximum	manhattan	minkowski
kmeans	3	3	3	3
average	2	2	15	2

How distance is measured, and how membership is decided makes a difference.

Some simple exercises to get familiar with data clustering

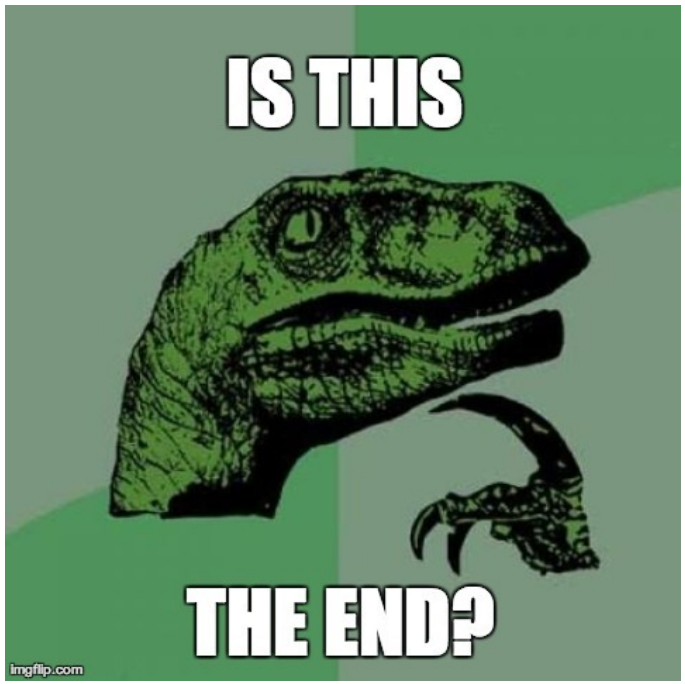
1. What are the clusters in the crime data population and murder rate?
2. Does a scatterplot of population and burglary rate show anything?
3. What kind of correlation exists between white population change and crime rate?

Q & A time.

Q: What was the greatest achievement in taxidermy?
A: The Royal Canadian Mounted Police.

What have we covered?

- Wrote simplistic clustering functions
- Glossed over the idea of distances and how they can be computed and measured
- Explored `kmeans()` as a way to cluster data
- Explored how to find the "correct" number of clusters
- Explored how distance measurements and clustering techniques can and will affect the number of clusters



Next: LPAR Chapter 5, agglomerative clustering

References (1 of 1)

[1] Ashley Crossman,
What Cluster Analysis Is and How You Can Use It in Research,
https://www.thoughtco.com/cluster-analysis-3026694, 2017.

[2] Michel Marie Deza and Elena Deza, Encyclopedia of Distances, Encyclopedia of Distances, Springer, 2009, pp. 1–583.

[3] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Pearson Education India, 2006.

Files of interest

1. Annotated iris histogram
2. Calculus derivation for "normal" distribution
3. YouTube video deriving the "normal" distribution: `https://www.youtube.com/watch?v=ebewBjZmZTw`
4. Cluster source code from Chapter 4
5. Revised cluster code
6. Revised crime cluster code
7. Life expectancy clusters
8. R library script file