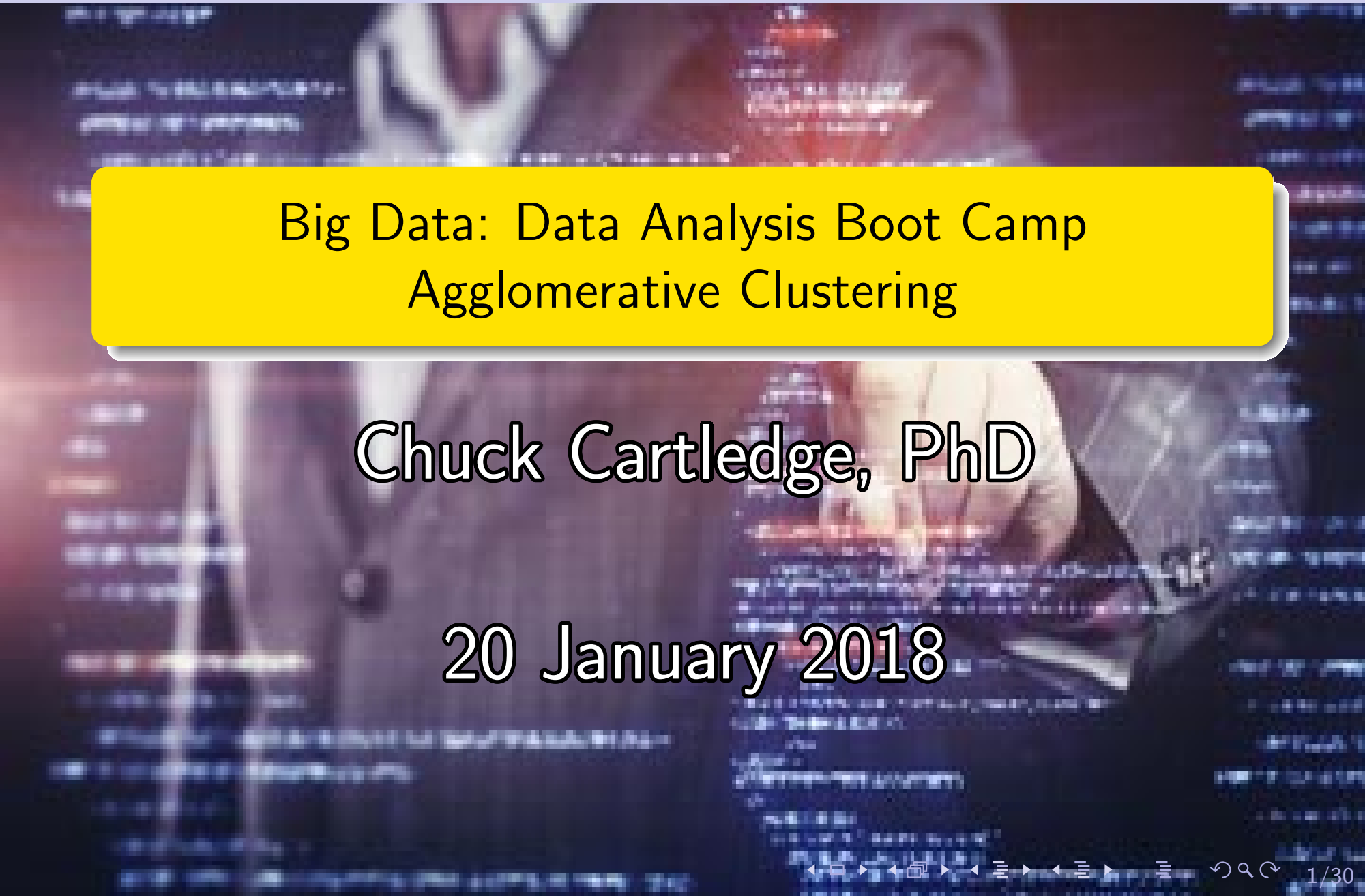


Big Data: Data Analysis Boot Camp
Agglomerative Clustering

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

What are we going to cover?

We're going to talk about:

- Different types of clustering approaches,
- How different measurement techniques affect clustering, and
- How clusters can interact.

Constructive (agglomerative) or deconstructive (divisive) approach?

Basic approaches:[2]

- **Agglomerative** each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy
- **Divisive** all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy

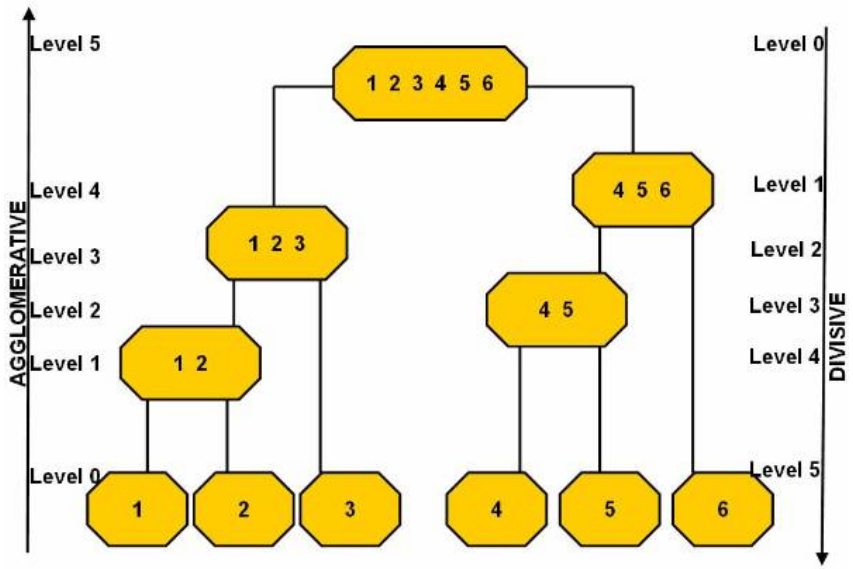


Image from [1].

Same image.

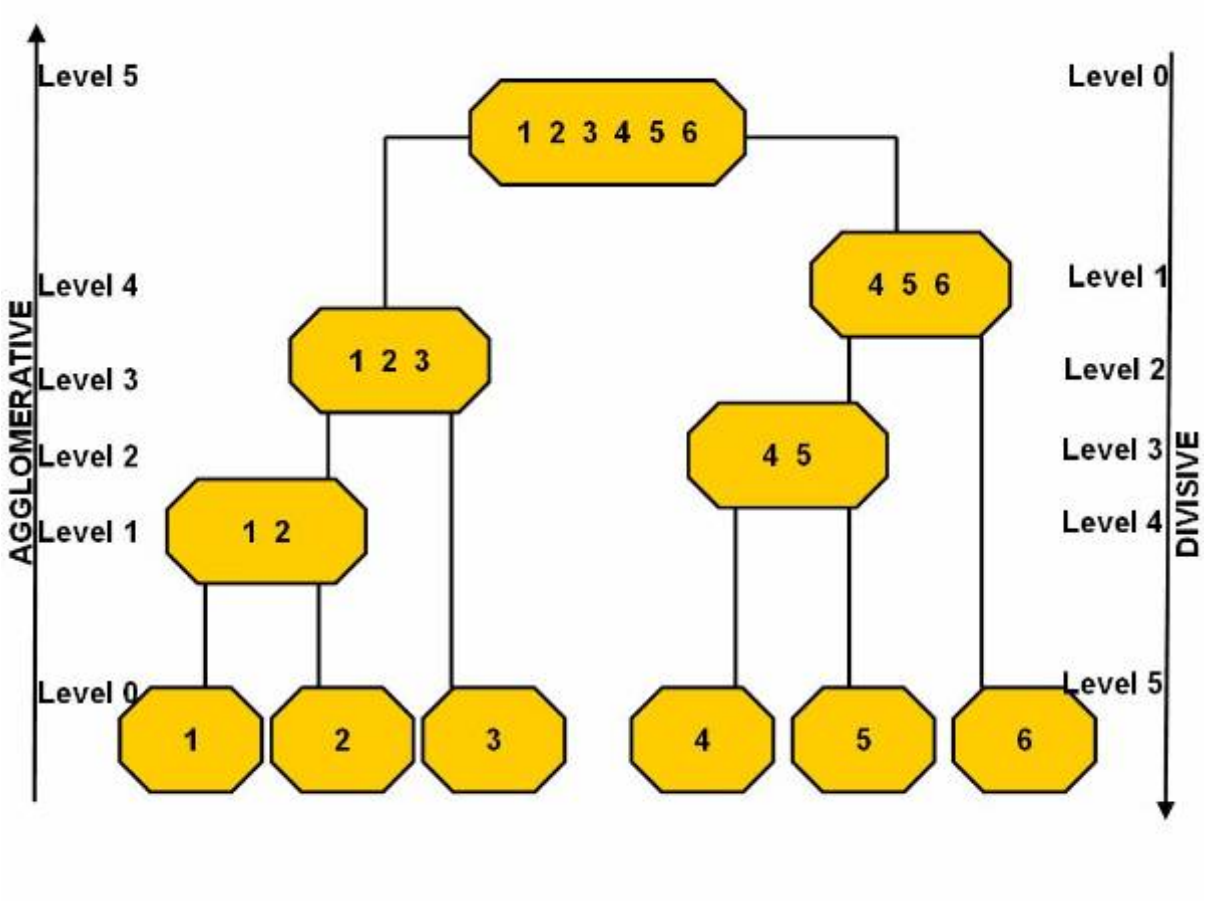


Image from [1].

Agglomerative and divisive are SLOW

Big O complexity of both approaches:[2]

- **Agglomerative** $O(n^2 \log(n))$
- **Divisive** $O(2^n)$

"Blind" clustering is too slow for large datasets. Special cases may have complexity of $O(n^2)$.

Types of clusterings[3]:

- Hierarchical versus partitional: clusters are nested or not
- Exclusive vs. overlapping vs. fuzzy: clusters can overlap or not, or membership is not a binary value
- Complete versus partial: all members are assigned to a cluster or not

Types of clusters[3]:

- Well-separated: each member is closer to the centroid than any other member not in the cluster
- Prototype-based: members are "closer" to the cluster prototype (or medoid) vice any other prototype
- Graph based: members have a "connection" (edge or arc) to other members in the cluster
- Density-based: a cluster is an area of high density surrounded by areas of low density
- Shared-property: members share some common attribute

Data characteristics that will affect clustering[3]:

- High dimensionality: in high dimensioned space, Euclidean distance becomes less useful
- Size: many clustering algorithms do not scale well
- Sparseness: members with many attributes, may have zero values that might be important
- Noise and outliers: may cause members to be incorrectly added to a cluster
- Types of attributes and data set: categorical, quantitative, binary, discrete, continuous
- Scale: units of attribute measures may affect cluster assignment
- Mathematical properties of the data space: ideas of mean, Euclidean distance, density may be applicable to the data space and therefore will affect clustering

Considerations when choosing a clustering algorithm[3]:

- Type of clustering
- Type of cluster
- Characteristics of a cluster
- Characteristics of the dataset and attributes
- Noise and outliers
- Number of data objects
- Number of attributes
- Cluster description
- Algorithmic considerations

Each consideration will affect the outcome.

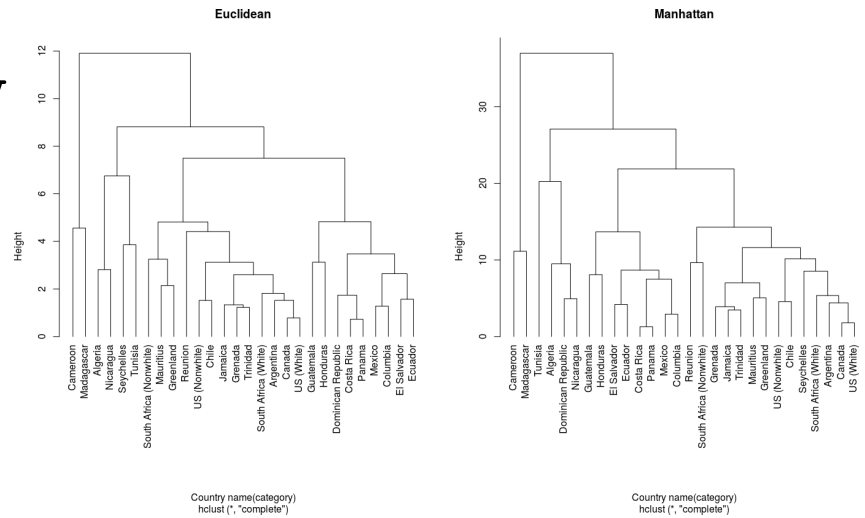
Our clustering explorations

We will be exploring agglomerative hierarchical clusters using `hclust()`. Because:

- Our data support this approach,
- It is easy to understand,
- It is the text.

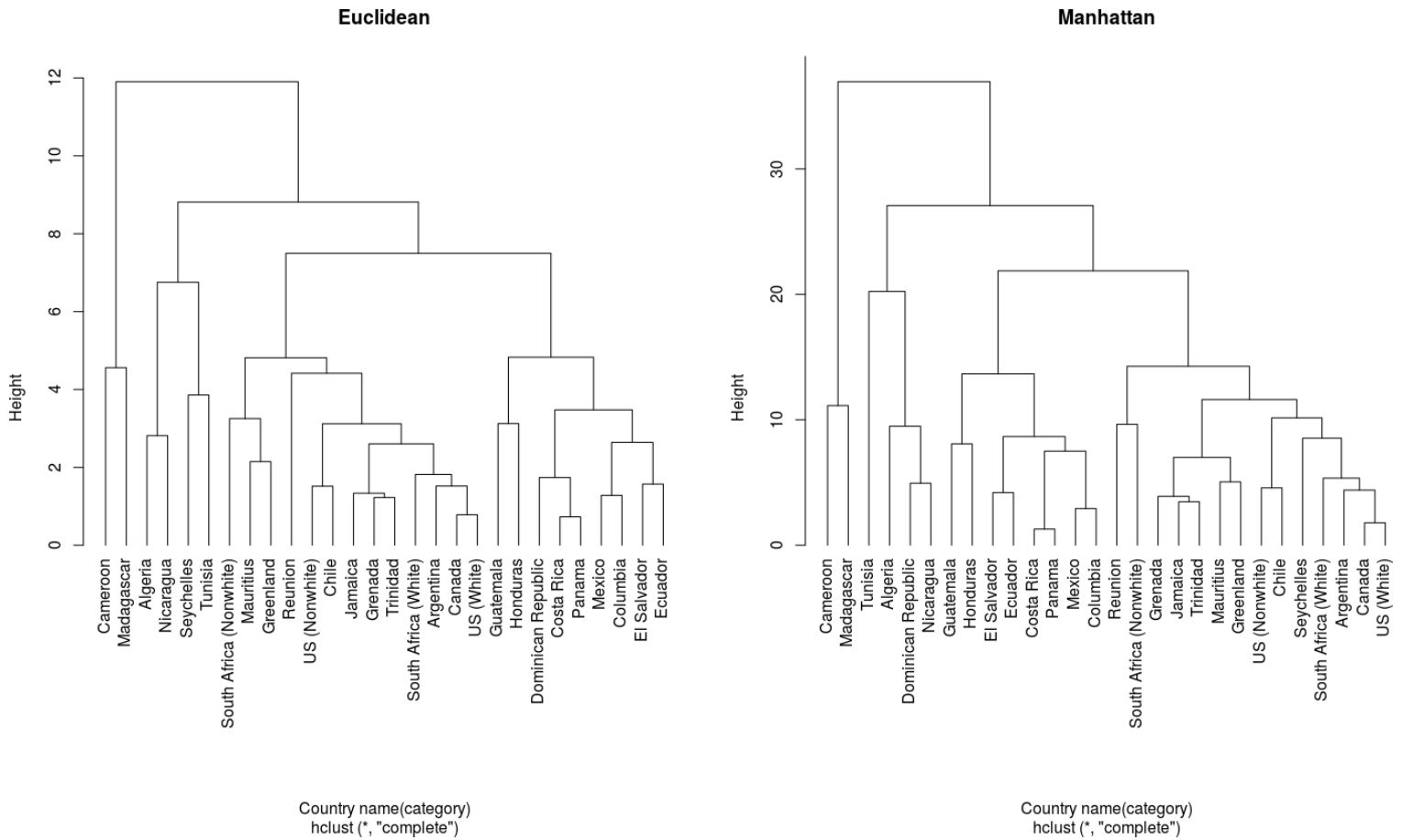
Life expectancy from text

1. Load the attached file `chapter-05-life-expectancy` into the editor
2. Execute the entire file (highlight and press the "run" button)
3. Execute: `main(FALSE)` in the console



Attached file.

Same image.

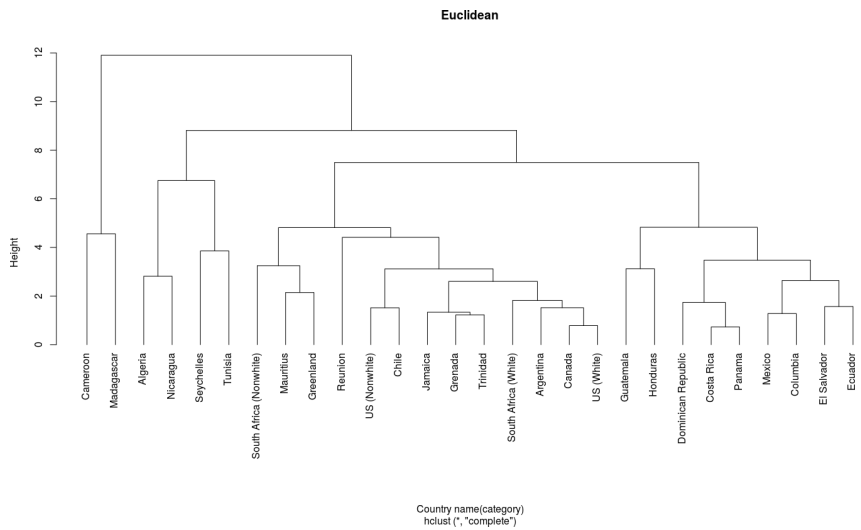


Attached file.

What is happening?

1. Rows (countries) that are "close" to each other are clustered (agglomerated)
2. How far the centroid is from each row is the "height" value (height can be thought of as "distance")
3. Clusters that are "close" to each other are clustered (hence hierarchy)

At any time, a horizontal line will give clusters and membership.

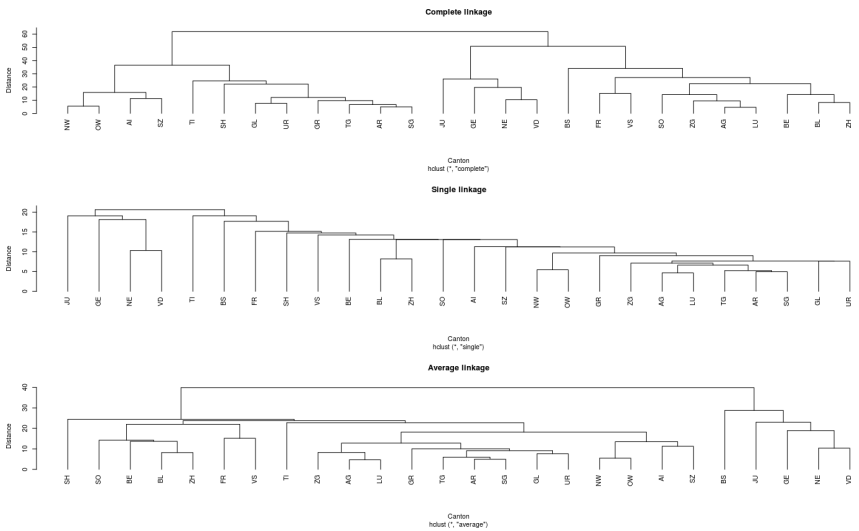


Exploring Swiss voting results

Analyzing a common datafile with different clustering approaches.

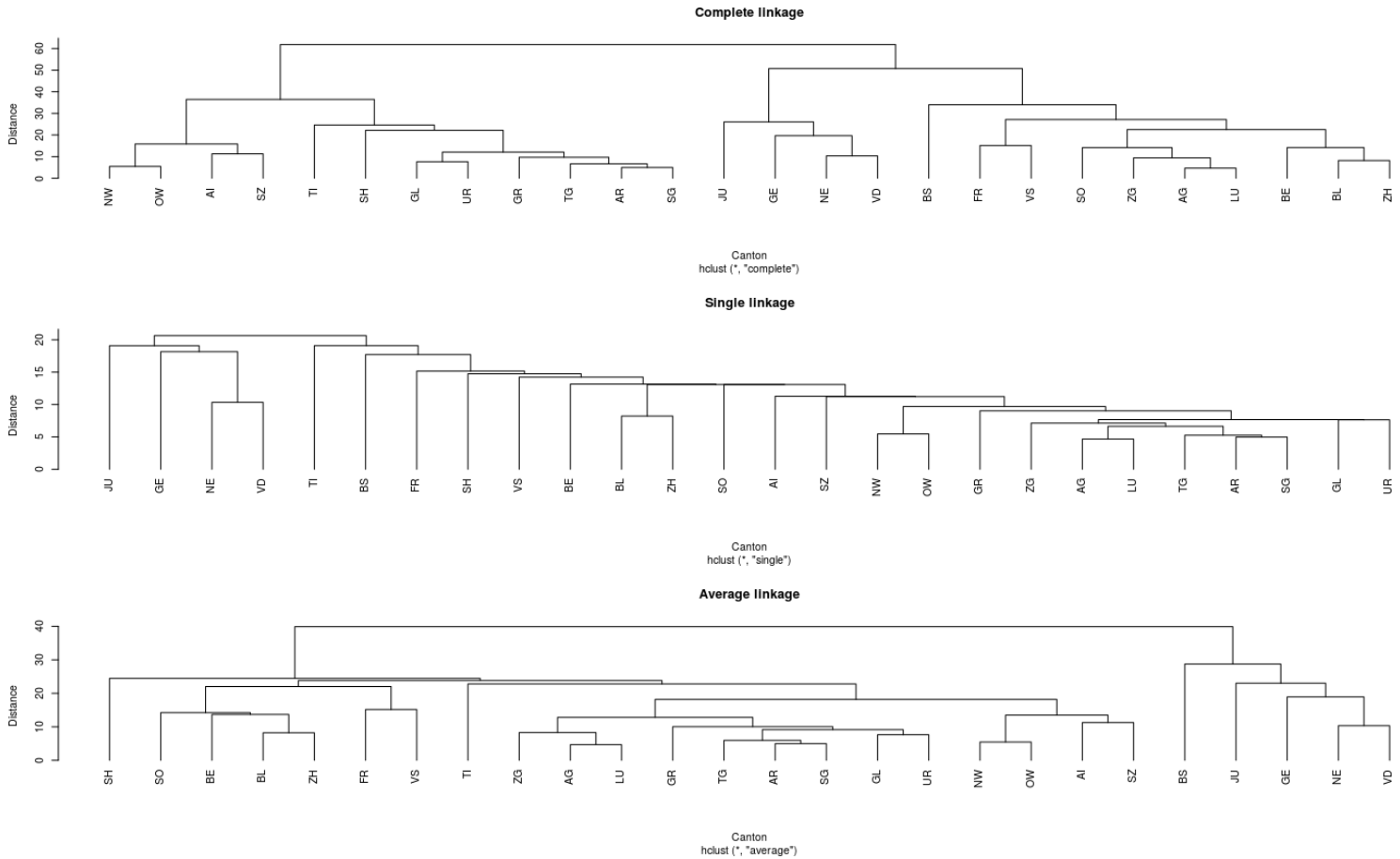
- **complete**: produces compact clusters
- **single**: more inclusive clusters
- **average**: clusters between compact and inclusive

Program in attached file (chapter-05-swiss-voting.R).



Attached file.

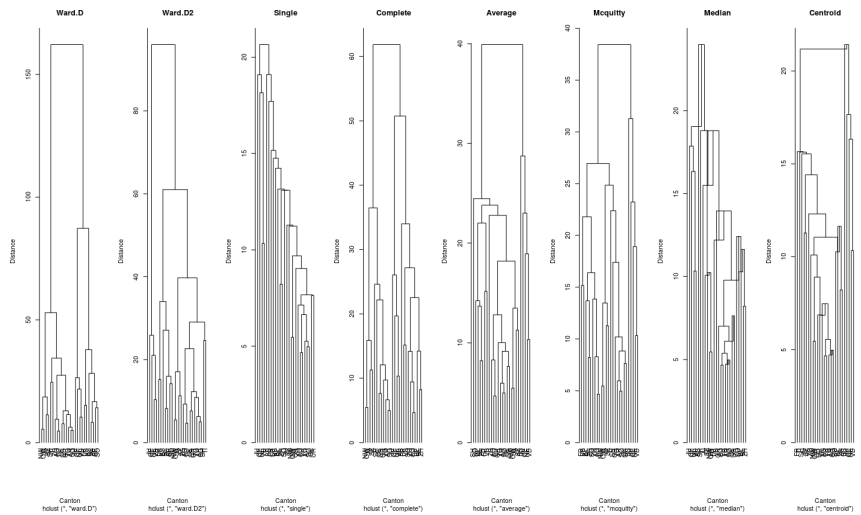
Same image.



Attached file.

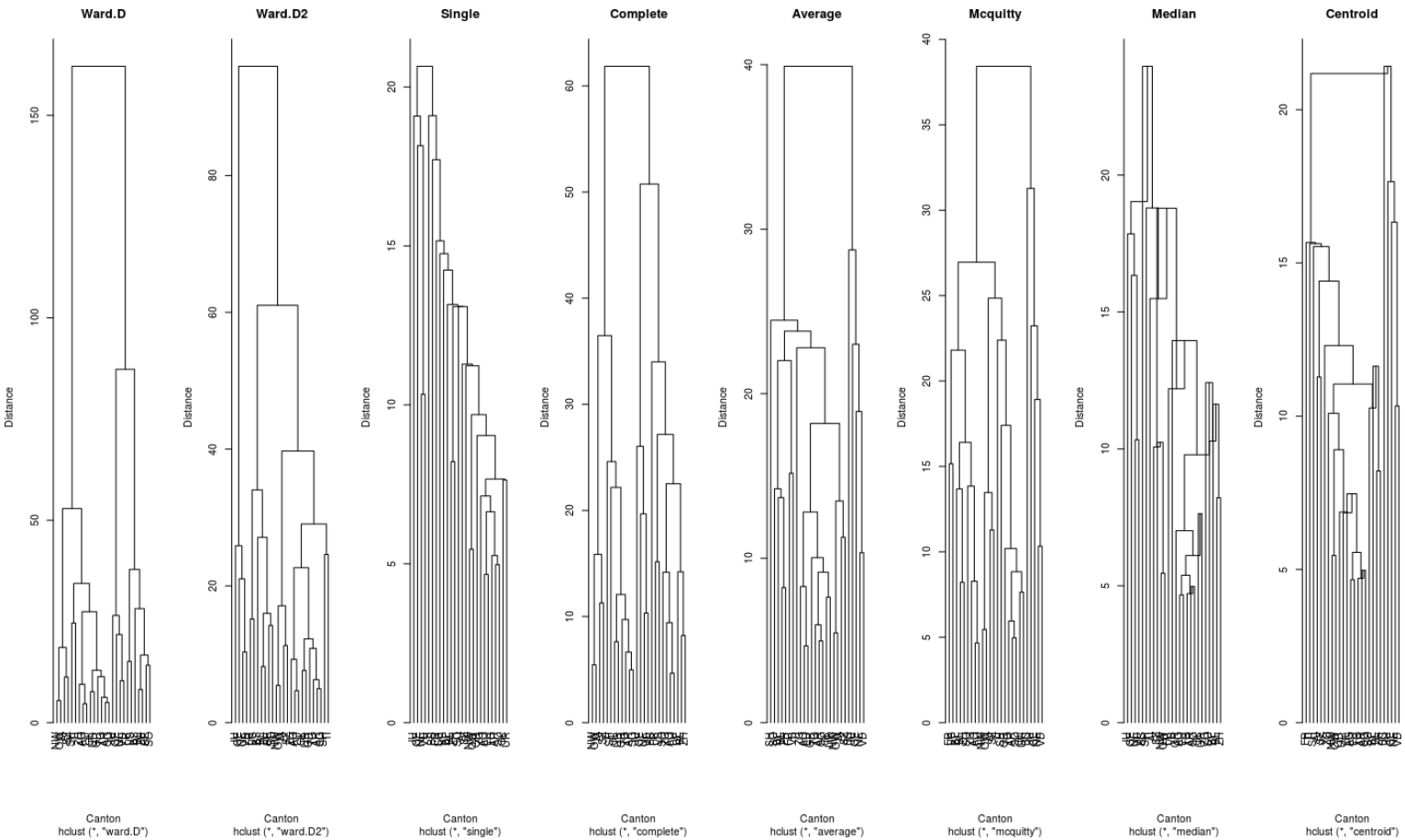
Swiss voting data in different forms

1. Load the attached file `chapter-05-swiss-voting.R` into the editor
2. Execute the entire file (highlight and press the "run" button)
3. Execute: `main(FALSE)` in the console



Attached file.

Same image.



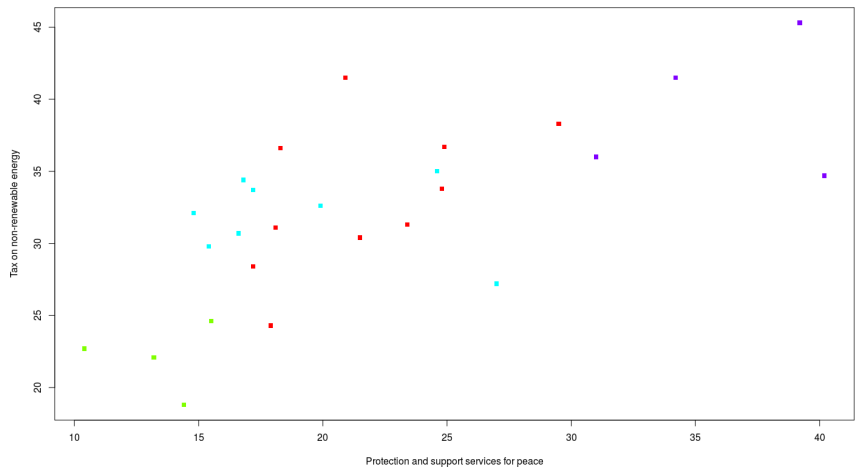
Attached file.

Interesting things about clusters

Some interesting things:

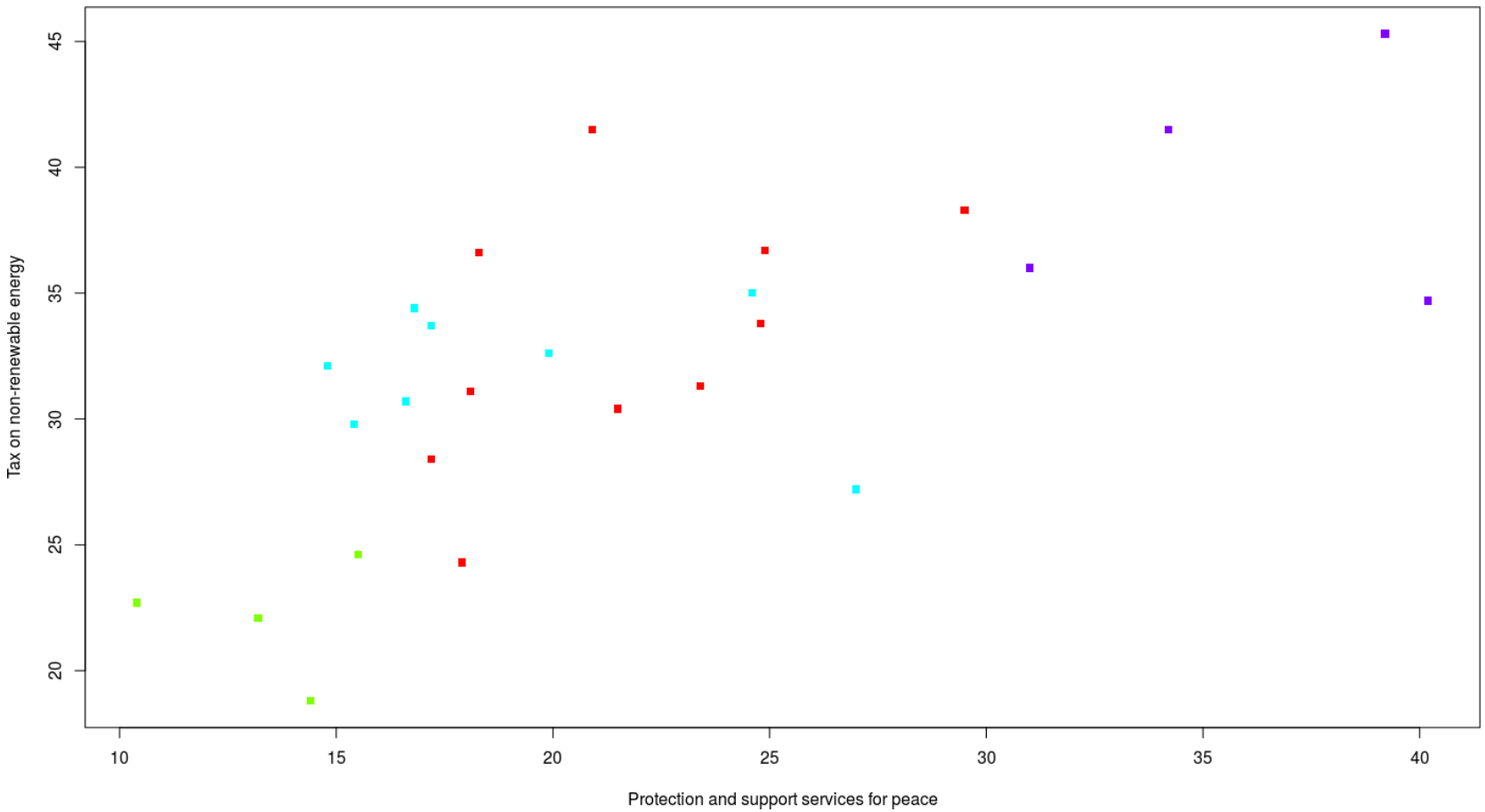
- Data was "cut" based on "complete" clustering and $k = 4$
- Colors match the cluster each cannon is assigned to
- There are two well separated clusters
- There are two over-lapping clusters

Based on the clusterings, more investigation into why there are similarities and differences is warranted.



Attached file.

Same image.

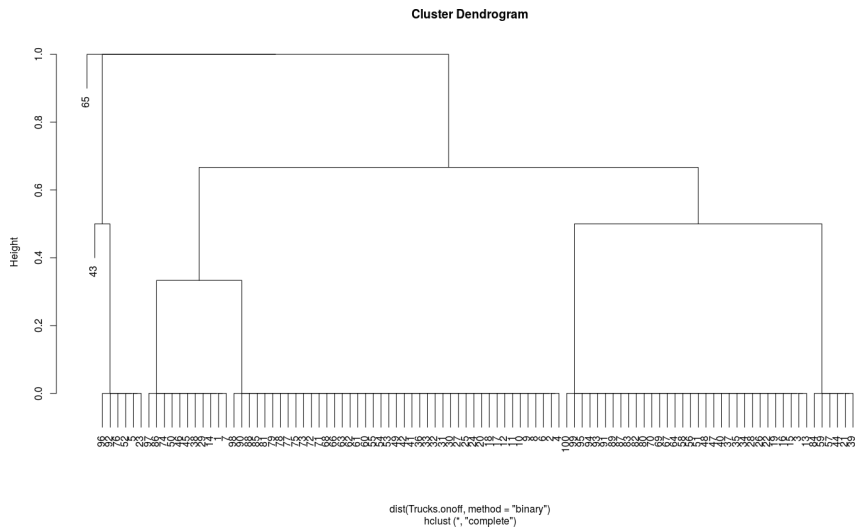


Attached file.

Looking at British truck accident data

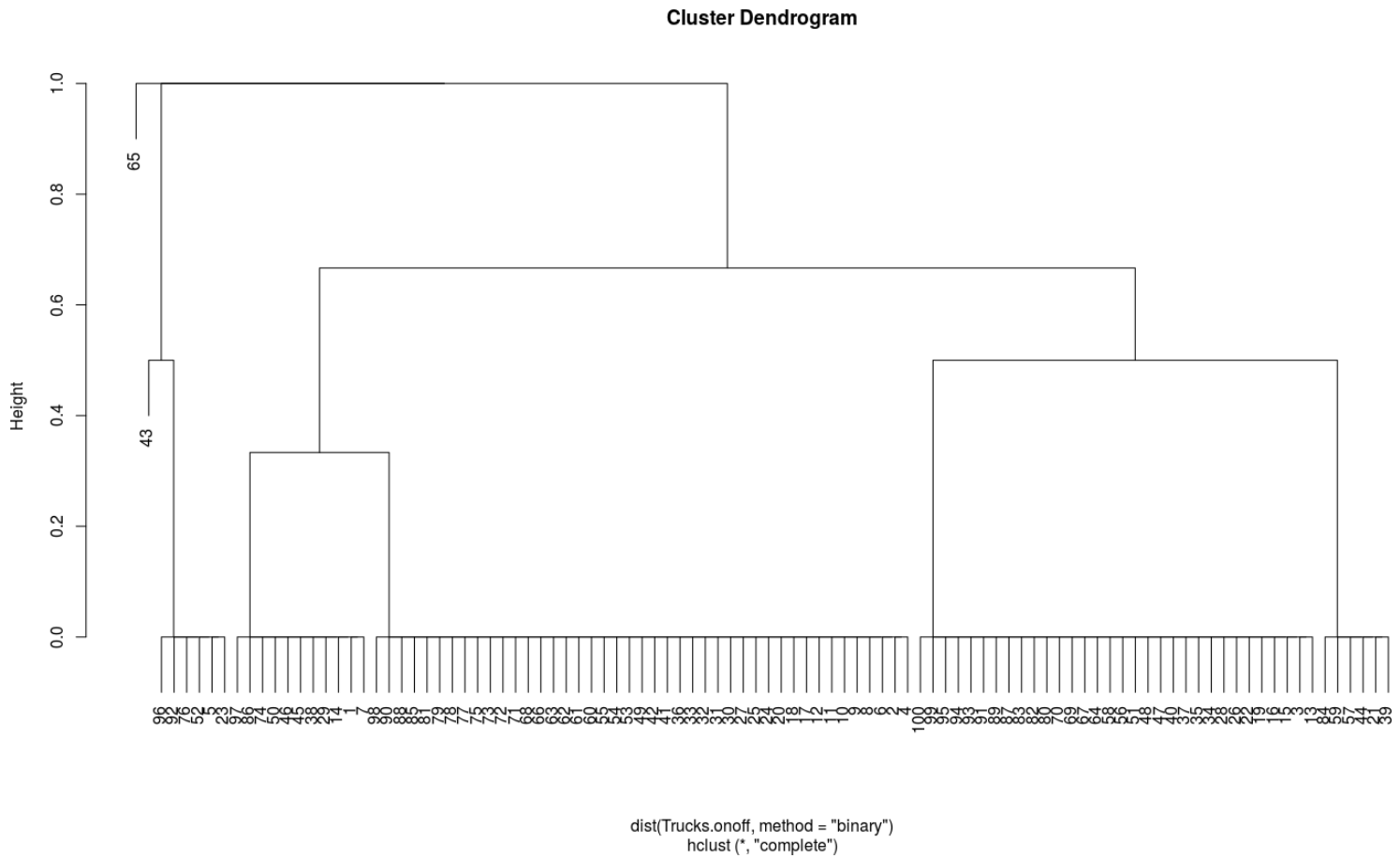
Some notes:

- Data was collected in Britain before and after new safety regulations
- Raw data is mostly categorical (two values)
- Categorical data converted to binary
- Distance is still valid measurement



Attached file.

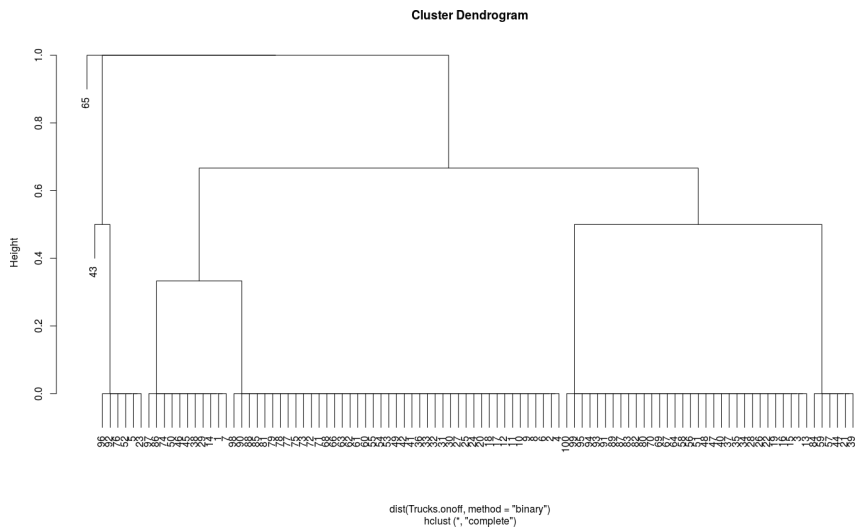
Same image.



Attached file.

Loading the truck accident software

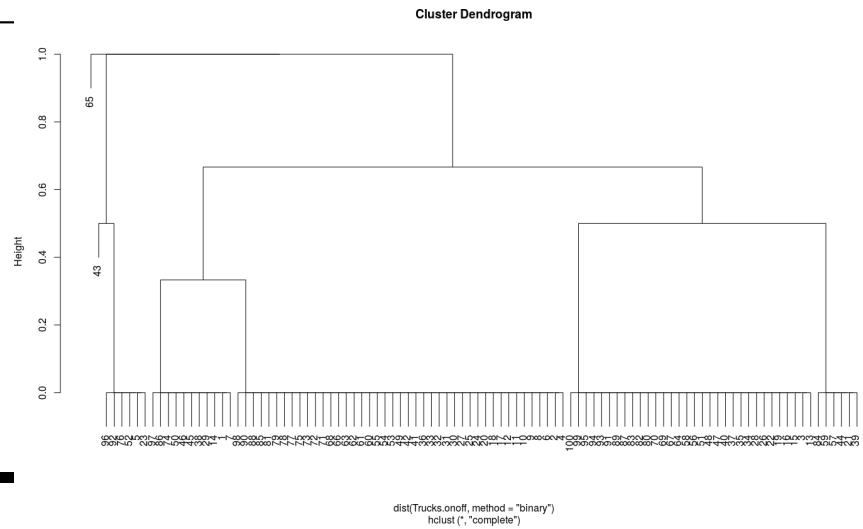
1. Load the attached file `chapter-05-trucks.R` into the editor
2. Execute the entire file (highlight and press the "run" button)
3. Execute: `main(FALSE)` in the console



More detailed investigation

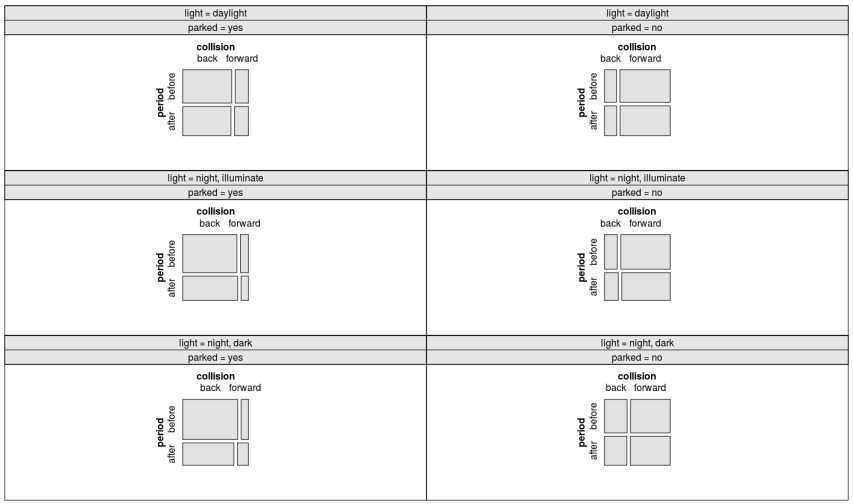
Index	period	collision	parked
65	0	1	0
43	1	0	0
96	1	1	0
97	1	1	1
100	0	0	1
84	0	1	1

Only 6 of possible 8 combinations.



A final look

1. Load the attached file `chapter-05-trucks-help.R` into the editor
2. Execute the entire file



From `*help[R](Trucks)*`

Some simple exercises to get familiar with data clustering

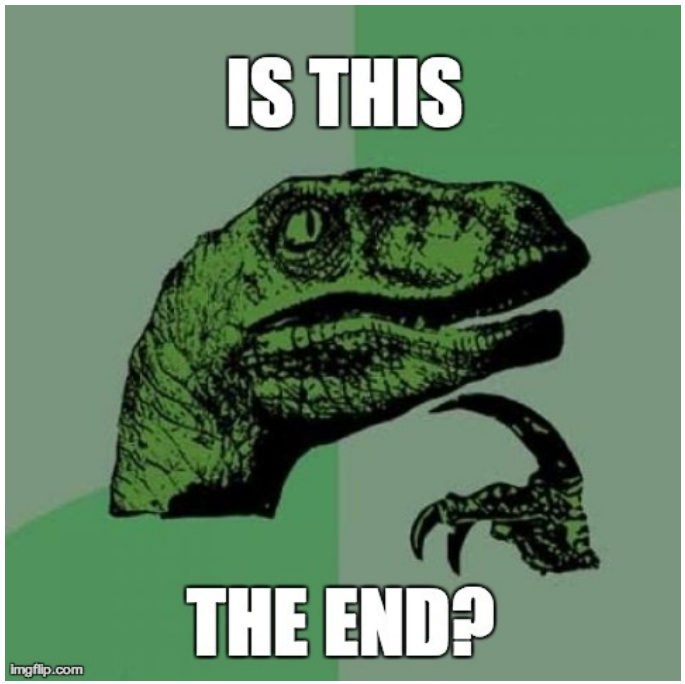
1. What are the "height" values for the life expectancy plots (requires looking at code)?
2. Change the life expectancy plots to display raw values.
3. Change the Swiss voting example to use Manhattan distance. Which distance measurement is better/best? Why?

Q & A time.

Q: How did you get into artificial intelligence?
A: Seemed logical – I didn't have any real intelligence.

What have we covered?

- Looked at different types of clustering approaches
- Looked at different types of clusters and data characteristics
- Looked at clustering considerations
- Looked at how distance computations can affect clustering
- Looked at how clusters can interrelate
- looked at continuous and binary data clustering



Next: LPAR Chapter 9, linear regression

References (1 of 1)

[1] Erin Shellman, BI Tech CP303 - Data Mining, http://erinshellman.github.io/data-mining-starter-kit/, 2017.

[2] Wikipedia Staff, Hierarchical clustering, https://en.wikipedia.org/wiki/Hierarchical_clustering, 2017.

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

Files of interest

1. Revised life-expectancy clustering (chapter-05-life-expectancy.R)
2. Revised Swiss voting clustering
3. Revised truck accidents
4. Truck accidents (a final look)
5. R library script file