Week 10 Exercise: Cluster Analysis

Calum Webb

27/11/2021

In this practical activity we will be using cluster analysis to try and explore whether there are underlying clusters of crime incidence in US states and English Community Safety Partnerships, using data from the CORGIS Dataset Project (https://corgis-edu.github.io/corgis/csv/state_crime/) and the Office for National Statistics (https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/recordedcrimedatabycommunitysafetypartnershiparea). The researchers in this activity are interested in finding out if states/community partnerships simply cluster into 'high' or 'low' crime, or whether there are clusters of specific types of crime (assault, murder, sexual crime, theft, etc.)

First, you will be asked to follow along and interpret the output from some code analysing clusters of crime in US states. Then, you will be asked to use this code as a template to explore clusters of crime in community safety partnerships in England.

Part I: Clusters of Crime Types in US States

Start by loading (or installing and then loading) the relevant libraries used for cluster analysis.

```
# Don't forget to install any packages you don't have installed.
#install.packages("tidyverse")
#install.packages("cluster")
#install.packages("factoextra")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

First, the researchers read in the data and save it in an object called `usa_data`

```
usa_data <- read_csv("state_crime_rates.csv")
```

```
## Rows: 51 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): state
## dbl (6): property_burglary, property_larceny, property_motor, violent_assaul...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
usa_data
```

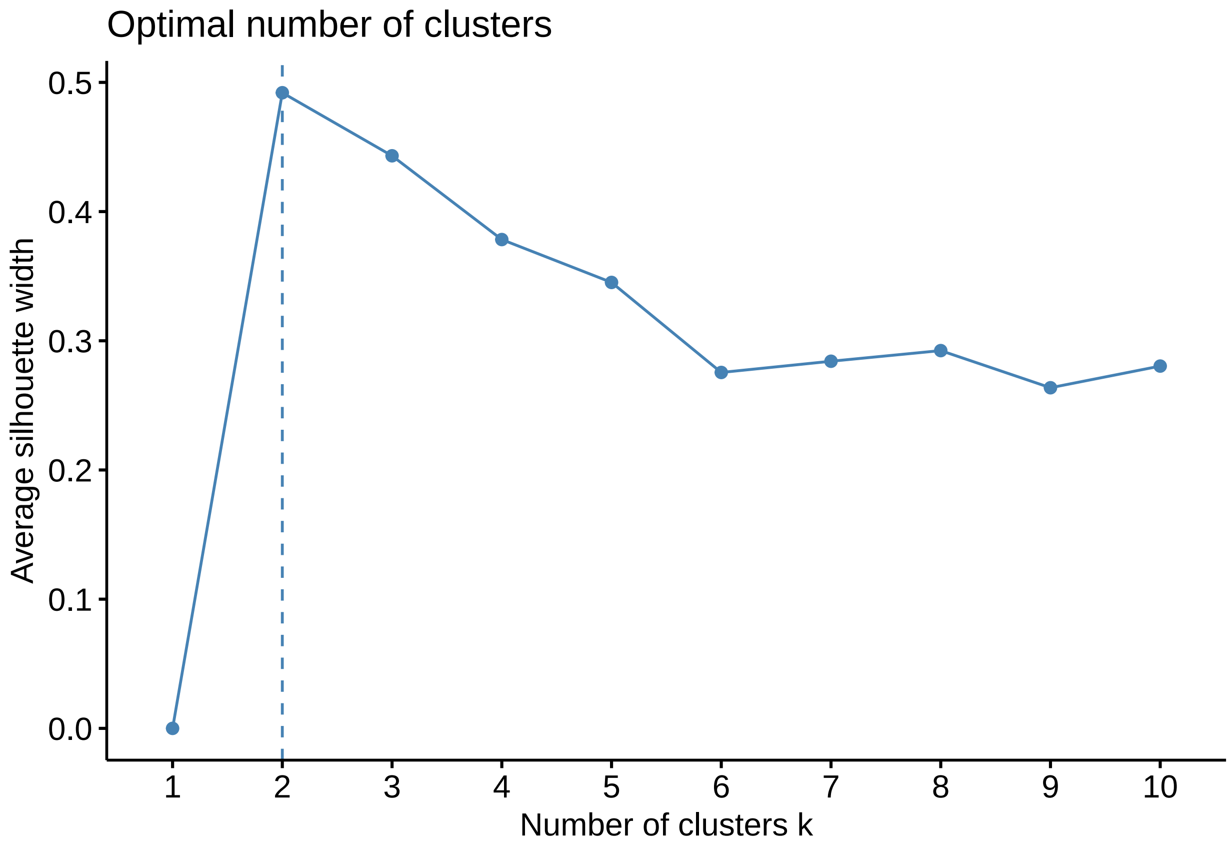
```
## # A tibble: 51 x 7
##    state       property_burglary property_larceny property_motor violent_assault
##    <chr>                   <dbl>            <dbl>          <dbl>           <dbl>
##  1 Alabama                  532.            1886.           256.             381
##  2 Alaska                   487.            2066            358.             596
##  3 Arizona                  394.            1797.           249.             312.
##  4 Arkansas                 600.            2013.           246.             448.
##  5 California               386.            1586.           359.             267.
##  6 Colorado                 348.            1858.           384              246.
##  7 Connecticut              181.            1079.           167.             105
##  8 Delaware                 305.            1783.           165.             305.
##  9 Florida                  295.            1669.           182.             258.
## 10 Georgia                  372.            1780.           224.             232
## # ... with 41 more rows, and 2 more variables: violent_murder <dbl>,
## #   violent_robbery <dbl>
```

Next, the researchers decide that because their variables of interest are all continuous they will start by using k-means to try and identify relevant clusters. They start by removing any non-numeric variables from their dataset, keeping only the numeric ones.

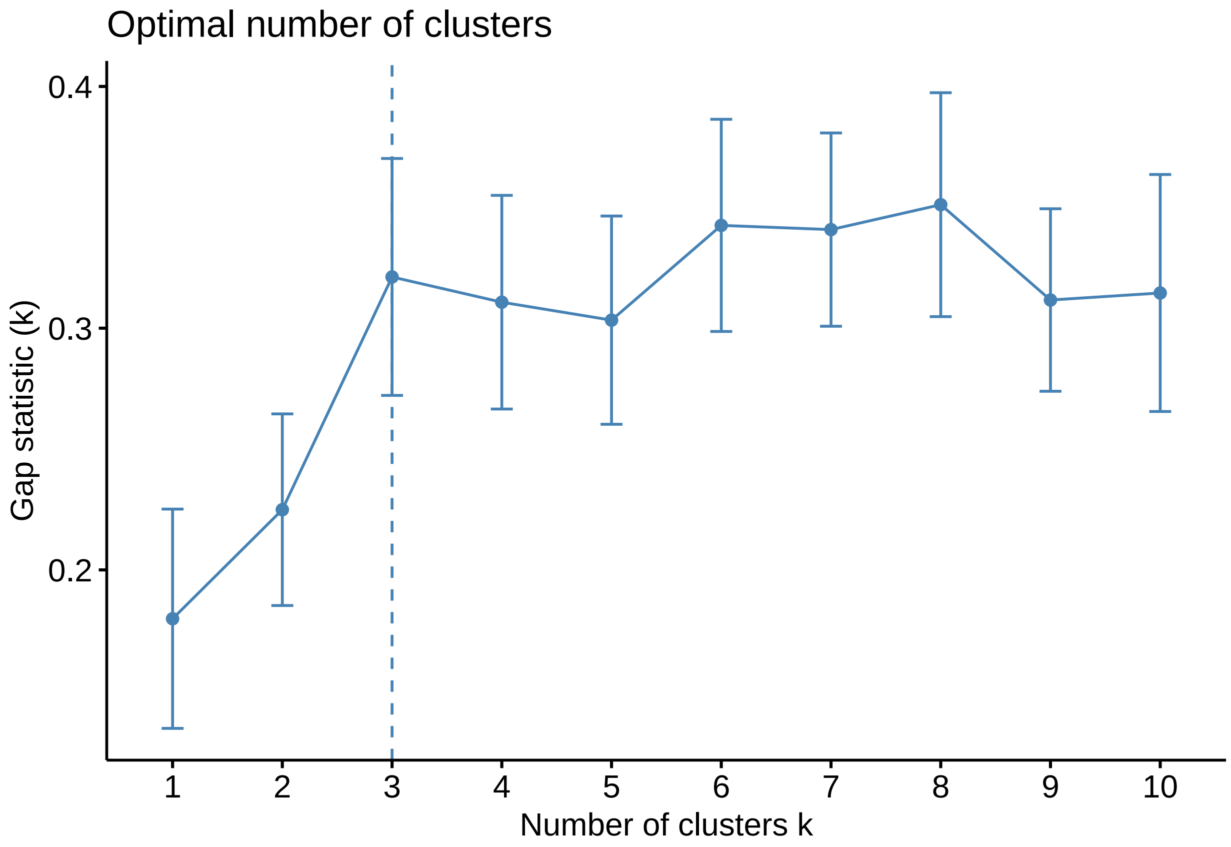
```
usa_data_prepped <- usa_data %>%
  select(-state)
```

They start by using the `factoextra` package, and the `fviz_nbclust` function to try and determine how many clusters they should try to identify.

```
# Silhouette Plot
fviz_nbclust(x = usa_data_prepped,
             FUNcluster = kmeans,
             method = "silhouette")
```



```
# Gap statistic plot
fviz_nbclust(x = usa_data_prepped,
             FUNcluster = kmeans,
             method = "gap")
```



- Interpret the above plots: what is the optimal number of clusters according to the silhouette statistic and what is the optimal number of clusters according to the gap statistic?

The optimal number of clusters according to the silhouette statistic is 2, whereas the optimal number of clusters according to the gap statistic is 3.

The researchers decide that they will create a 2-cluster solution (as suggested by the silhouette plot), as well as a 3-cluster solution suggested by the gap statistic. They use the `kmeans` to first estimate the clusters, they then visualise the clusters using the `fviz_cluster` function.

```
# run kmeans analysis
set.seed(2021)
usa_k2 <- kmeans(usa_data_prepped, centers = 2)
usa_k2
```

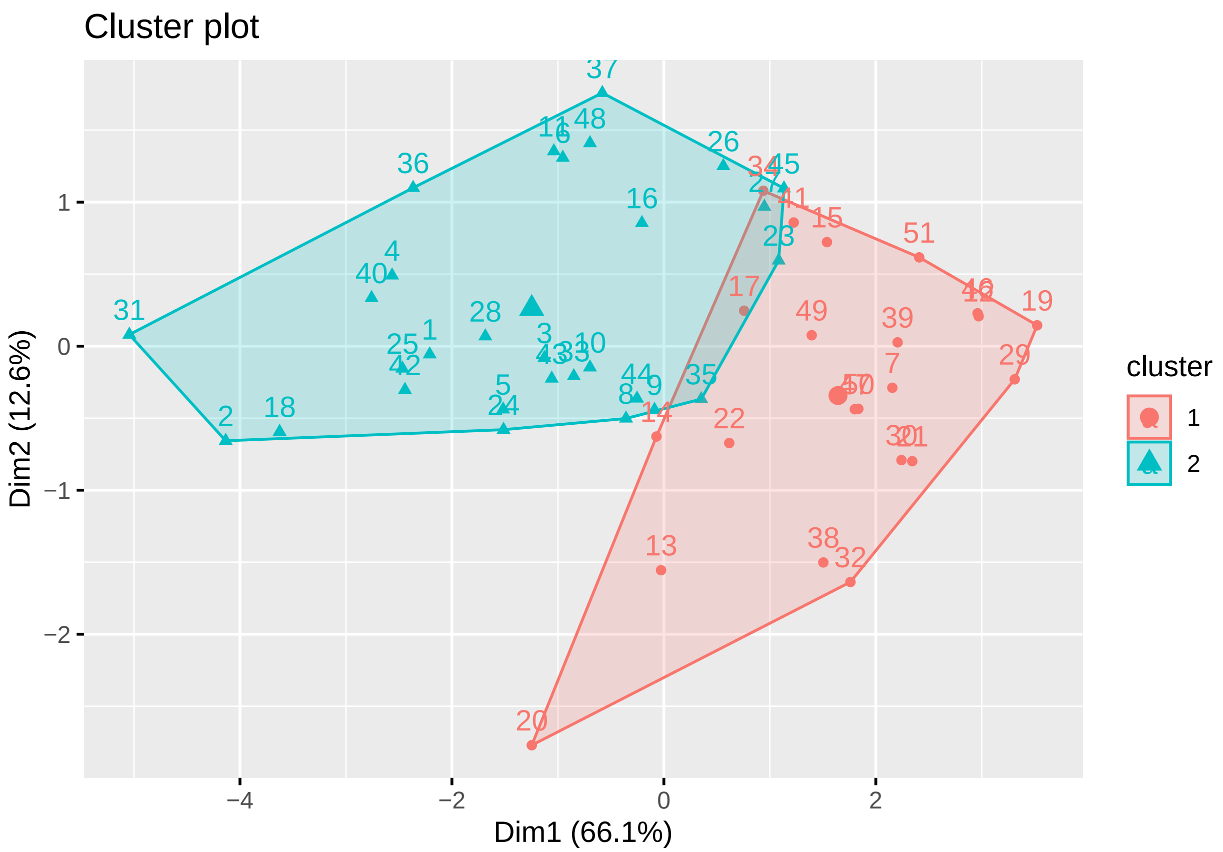
```
## K-means clustering with 2 clusters of sizes 22, 29
##
## Cluster means:
##   property_burglary property_larceny property_motor violent_assault
## 1          240.0727         1186.655       134.7500        186.6682
## 2          428.4034         1802.272       272.4793        298.9207
##   violent_murder violent_robbery
## 1       3.609091        48.48636
## 2       5.672414        73.93448
##
## Clustering vector:
```

```
##  [1] 2 2 2 2 2 2 1 2 2 2 2 1 1 1 1 2 1 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 2 1 2 2 2 1
## [39] 1 2 1 2 2 2 2 1 1 2 1 1 1
##
## Within cluster sum of squares by cluster:
## [1]  877435.5 2349097.2
##  (between_SS / total_SS =  63.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

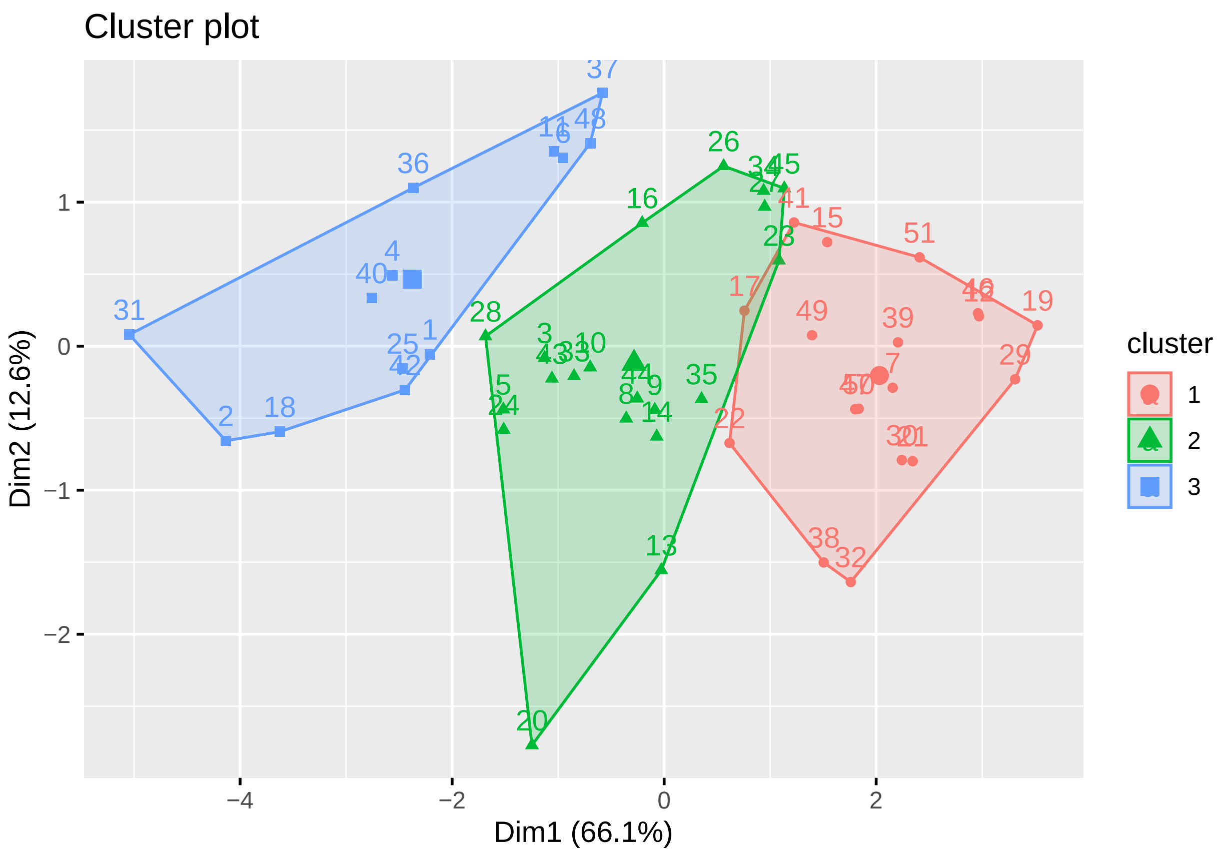
```
# run kmeans analysis
set.seed(2021)
usa_k3 <- kmeans(usa_data_prepped, centers = 3)
usa_k3
```

```
## K-means clustering with 3 clusters of sizes 18, 20, 13
##
## Cluster means:
##   property_burglary property_larceny property_motor violent_assault
## 1          225.8389         1130.656       121.7278        174.5278
## 2          357.2300         1604.145       223.4400        244.2050
## 3          499.6615         1995.200       323.5769        365.3692
##   violent_murder violent_robbery
## 1       3.061111        39.63333
## 2       4.945000        74.72000
## 3       6.915385        77.15385
##
## Clustering vector:
##  [1] 3 3 2 3 2 3 1 2 2 2 3 1 2 2 1 2 1 3 1 2 1 1 2 2 3 2 2 2 1 1 3 1 2 2 2 3 3 1
## [39] 1 3 1 3 2 2 2 1 1 3 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 485677.8 634619.6 730596.7
##  (between_SS / total_SS =  79.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# visualise clusters - 2 cluster
fviz_cluster(usa_k2,
             data = usa_data_prepped)
```



```
# visualise clusters - 3 cluster
fviz_cluster(usa_k3,
             data = usa_data_prepped)
```



- Approximately what proportion of variance could be explained by the cluster membership solution chosen in the 2-cluster and 3-cluster solutions?

The proportion of variance explained in the 2 cluster solution was around 63.4%, whereas the proportion explained in the 3 cluster solution was 79%.

- Do the clusters look well defined? Are there any states that may have been misclassified by the algorithm?

Both 2-clusters and 3 cluster solutions appear to have some overlap. For example, in the two cluster solution the states on row 17, 23, 27, 41, 45, and 34 are all clustered together on the border of the two clusters and could therefore be valid for either.

The 3-cluster solution seems to have slightly more unique clusters, although at the same point (around where cluster 41 is) there appears to be some overlap.

The researchers then decide to add the cluster membership to the original data and then explore how the clusters differ in mean values of each crime rate.

```
# Add cluster results to data and save in `usa_data_results`
usa_data_results <- usa_data %>%
  mutate(
    cluster_k2 = usa_k2$cluster,
    cluster_k3 = usa_k3$cluster
  )

# Summarise all numeric variables with their mean
# Tip: Uncomment the %>% view() section of the code to view all output
```

```
usa_data_results %>%
  group_by(cluster_k2) %>%
  summarise_all(~mean(., na.rm = TRUE)) # %>% view()
```

```
## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## # A tibble: 2 x 9
##   cluster_k2 state property_burglary property_larceny property_motor
##        <int> <dbl>             <dbl>            <dbl>          <dbl>
## 1          1    NA              240.            1187.           135.
## 2          2    NA              428.            1802.           272.
## # ... with 4 more variables: violent_assault <dbl>, violent_murder <dbl>,
## #   violent_robbery <dbl>, cluster_k3 <dbl>
```

```
usa_data_results %>%
  group_by(cluster_k3) %>%
  summarise_all(~mean(., na.rm = TRUE)) # %>% view()
```

```
## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## # A tibble: 3 x 9
##   cluster_k3 state property_burglary property_larceny property_motor
##        <int> <dbl>             <dbl>            <dbl>          <dbl>
## 1          1    NA              226.            1131.           122.
## 2          2    NA              357.            1604.           223.
## 3          3    NA              500.            1995.           324.
## # ... with 4 more variables: violent_assault <dbl>, violent_murder <dbl>,
## #   violent_robbery <dbl>, cluster_k2 <dbl>
```

- Describe and label the two kinds of clusters found in the 2-cluster solution.

The two cluster solution appears to be broken into high-crime and low-crime clusters. Cluster 1 has higher average incidence of all forms of recorded crime whereas cluster 2 has lower average incidence of all forms of crime.

- Describe and label the three kinds of clusters found in the 3-cluster solution.

The three cluster solution appears to have clustered the states into low, middling, and high incidence of crime, with cluster 1 being the lowest and cluster 3 being the highest.

The researchers then decide to check whether they find similar results when using hierarchical cluster analysis.

They decide that since all of their data is continuous, they will create a dissimilarity matrix based on Euclidean distance.

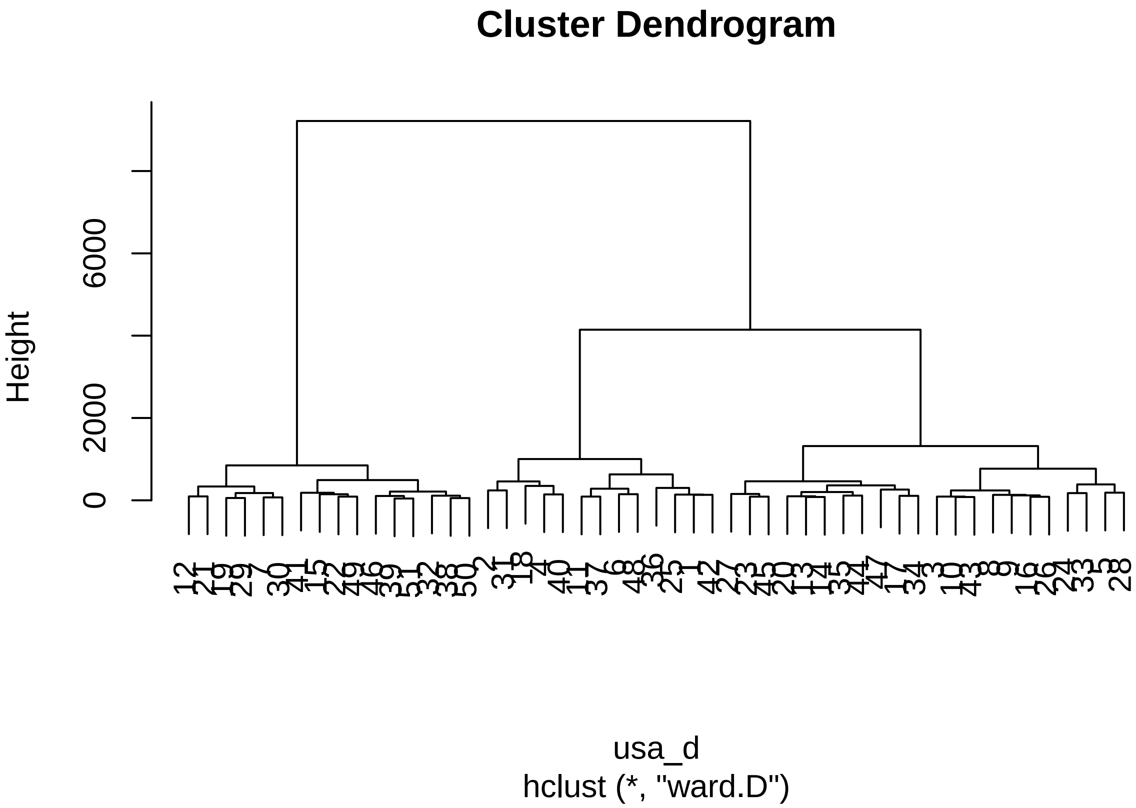
```r
usa_d <- daisy(usa_data_prepped, metric = "euclidean")
```

They decide to use two different methods for hierarchical cluster analysis: the Ward's linkage method and complete linkage method.

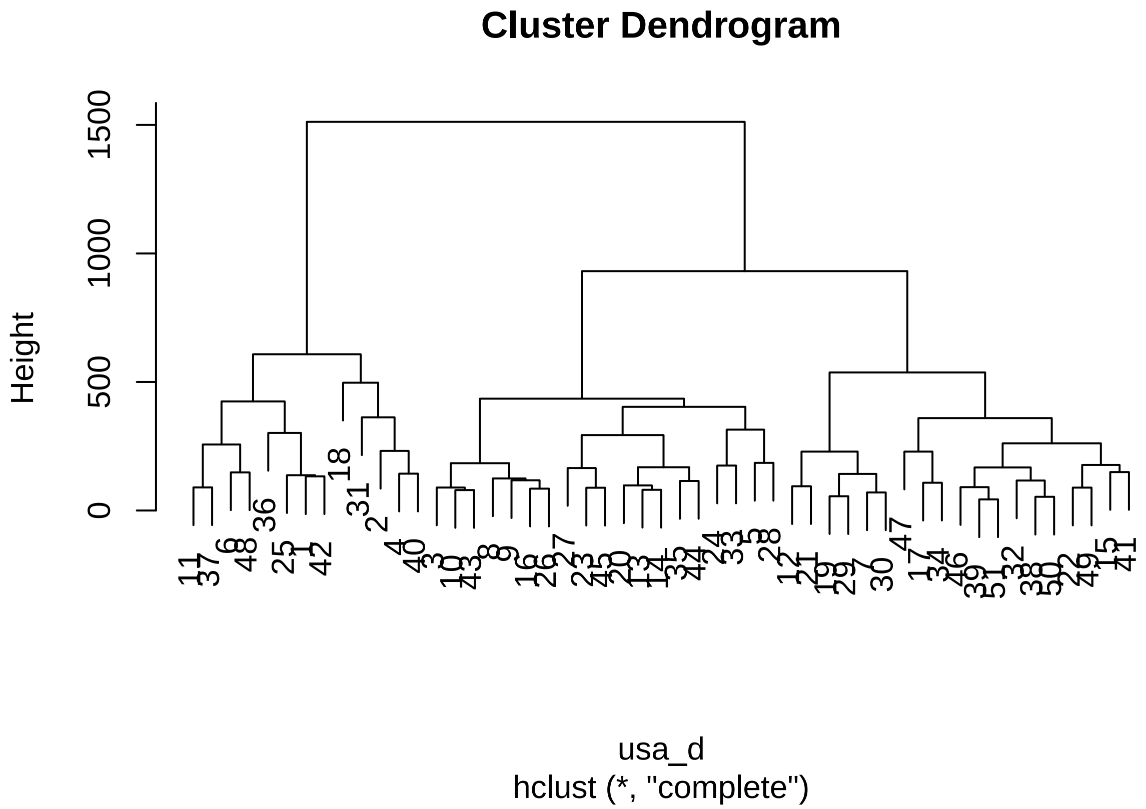
```r
set.seed(2021)
usa_ward     <- hclust(d = usa_d, method = "ward.D")
set.seed(2021)
usa_complete <- hclust(d = usa_d, method = "complete")
```

Then then visualise a dendrogram of their results.

```r
plot(usa_ward)
```



```r
plot(usa_complete)
```



- Based on the two dendrograms, how many different clusters might be reasonable to extract from the data and why?

HCA using Ward linkage seems to show three well-defined clusters, as does HCA using complete linkage.

The researchers decide to test a 3-cluster (Ward and complete) solution to their data clustering. They use the `cutree` function to achieve this.

```
usa_ward_k3 <- cutree(usa_ward, k = 3)
usa_comp_k3 <- cutree(usa_complete, k = 3)
```

They add the cluster results to their data and generate some descriptive statistics for each solution.

```
usa_data_hca_results <- usa_data %>%
  mutate(
    ward_k3 = usa_ward_k3,
    comp_k3 = usa_comp_k3
  )

# Results for 3 cluster ward
usa_data_hca_results %>%
  group_by(ward_k3) %>%
  summarise_all(~mean(., na.rm = TRUE)) # %>% view()
```

```
## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA
## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## # A tibble: 3 x 9
##   ward_k3 state property_burgla~ property_larceny property_motor violent_assault
##     <int> <dbl>            <dbl>            <dbl>          <dbl>           <dbl>
## 1       1    NA             500.            1995.           324.            365.
## 2       2    NA             348.            1580.           219.            234.
## 3       3    NA             222.            1104.           115.            180.
## # ... with 3 more variables: violent_murder <dbl>, violent_robbery <dbl>,
## #   comp_k3 <dbl>
```

```
# Results for 3 cluster complete
usa_data_hca_results %>%
  group_by(comp_k3) %>%
  summarise_all(~mean(., na.rm = TRUE)) # %>% view()
```

```
## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## Warning in mean.default(., na.rm = TRUE): argument is not numeric or logical:
## returning NA

## # A tibble: 3 x 9
##   comp_k3 state property_burgla~ property_larceny property_motor violent_assault
##     <int> <dbl>            <dbl>            <dbl>          <dbl>           <dbl>
## 1       1    NA             500.            1995.           324.            365.
## 2       2    NA             358.            1615.           223.            246.
## 3       3    NA             232.            1145.           128.            176.
## # ... with 3 more variables: violent_murder <dbl>, violent_robbery <dbl>,
## #   ward_k3 <dbl>
```

- Write a description and labels for the 3-cluster solution found through Ward's linkage

The Ward linkage HCA appears to show three clusters of states that could be labelled high crime (1), medium crime (2), and low crime (3).

- Write a description and labels for the 3-cluster solution found through complete linkage

Complete data linkage appears to have found clusters very similar to Ward linkage and k-means. With the exception of violent robbery, cluster 1 represents the highest rates of crime on average, cluster 2 represents the middle rates of crime, and cluster 3 represents low rates of crime.

- How would you summarise the research? Did the researchers find evidence that state-level crime fell into distinct categories of crimes committed, or did clusters largely reflect rates of all crimes?

Cluster analysis did not seem to suggest that, at the state level, communities face very different clusters characterised by types of crime. Rather, the clustering of crime was associated with the prevalence of all forms of crime rather than specific forms.

Part II: Clusters of Crime Types in English Community & Safety Partnerships

Now we'll explore whether we find similar or different results for crime rates in English Community and Safety Partnerships.

- Start by loading the "england_crime_rates.csv" data into R using the read_csv function. Save the result to an object called english_crime.

```
english_crime <- read_csv("england_crime_rates.csv")
```

```
## Rows: 300 Columns: 10

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): community_safety_partnership_code, community_safety_partnership_name
## dbl (8): violence_against_the_person, sexual_offences, robbery, theft_offenc...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
english_crime
```

```
## # A tibble: 300 x 10
##    community_safety_~ community_safety~ violence_agains~ sexual_offences robbery
##    <chr>              <chr>                        <dbl>           <dbl>   <dbl>
##  1 E22000001          Bath and North E~             18.3            1.89    0.54
##  2 E22000002          Bristol, City of              34.5            3.13    1.91
##  3 E22000003          North Somerset                25.6            2.01    0.49
##  4 E22000369          Somerset                      25.1            2.27    0.37
##  5 E22000006          South Gloucester~             19.5            1.9     0.44
##  6 E22000009          Bedford                       29.1            2.46    0.86
##  7 E22000353          Central Bedfords~             17.6            1.36    0.51
##  8 E22000010          Luton                         32.8            2.27    1.18
##  9 E22000013          Cambridge                     30.4            2.83    1.08
## 10 E22000014          East Cambridgesh~             19.0            1.76    0.24
## # ... with 290 more rows, and 5 more variables: theft_offences <dbl>,
## #   vehicle_offences <dbl>, criminal_damage_and_arson <dbl>,
## #   drug_offences <dbl>, public_order_offences <dbl>
```

- Check the kinds of variables in the data and decide whether you could use k-means, Hierarchical Cluster Analysis, or both methods for exploring clusters of crime.

Both k-means and Hierarchical Cluster Analysis could be used for analysing this data as all variables of interest are continuous.

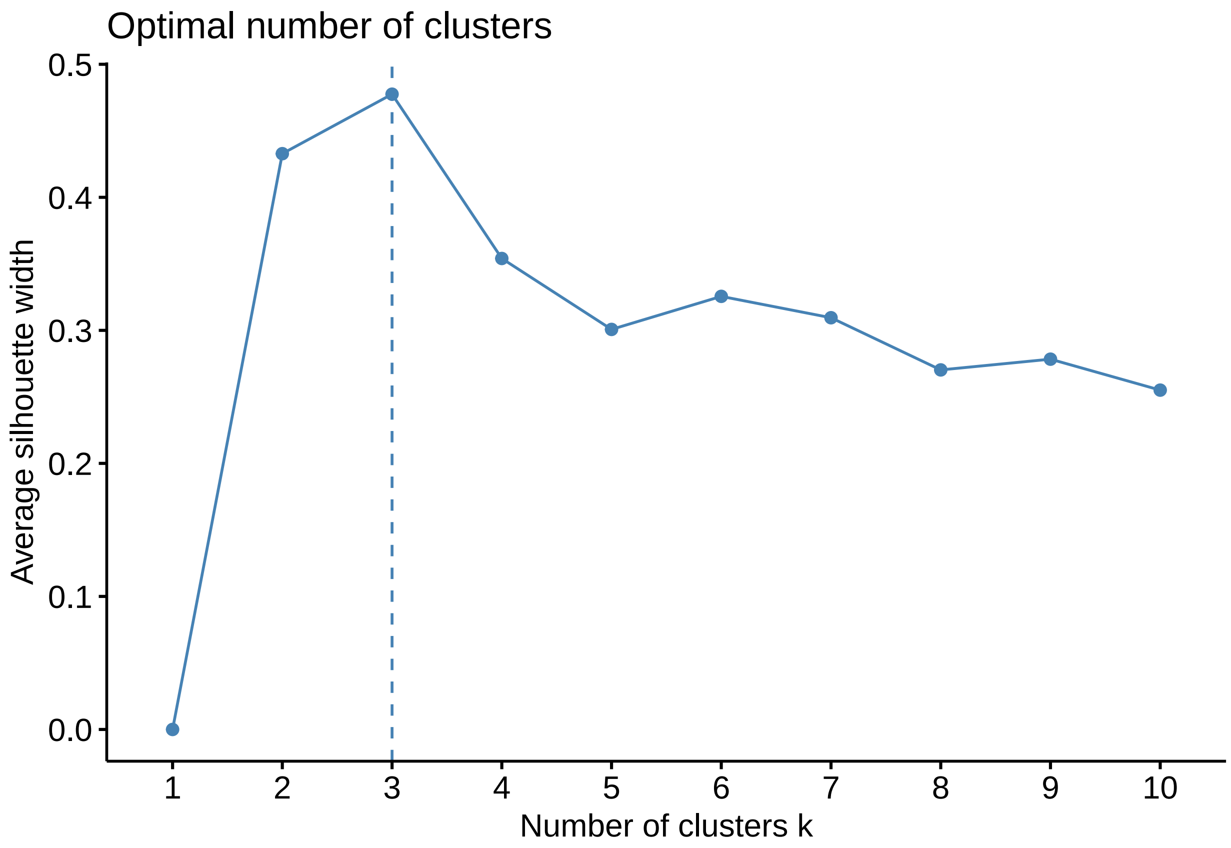
- Create a version of the data that contains only the numeric type variables and store it in an object called english_crime_prepped so that it can be used for k-means and HCA.

```
english_crime_prepped <- english_crime %>%
  select_if(is.numeric)
```

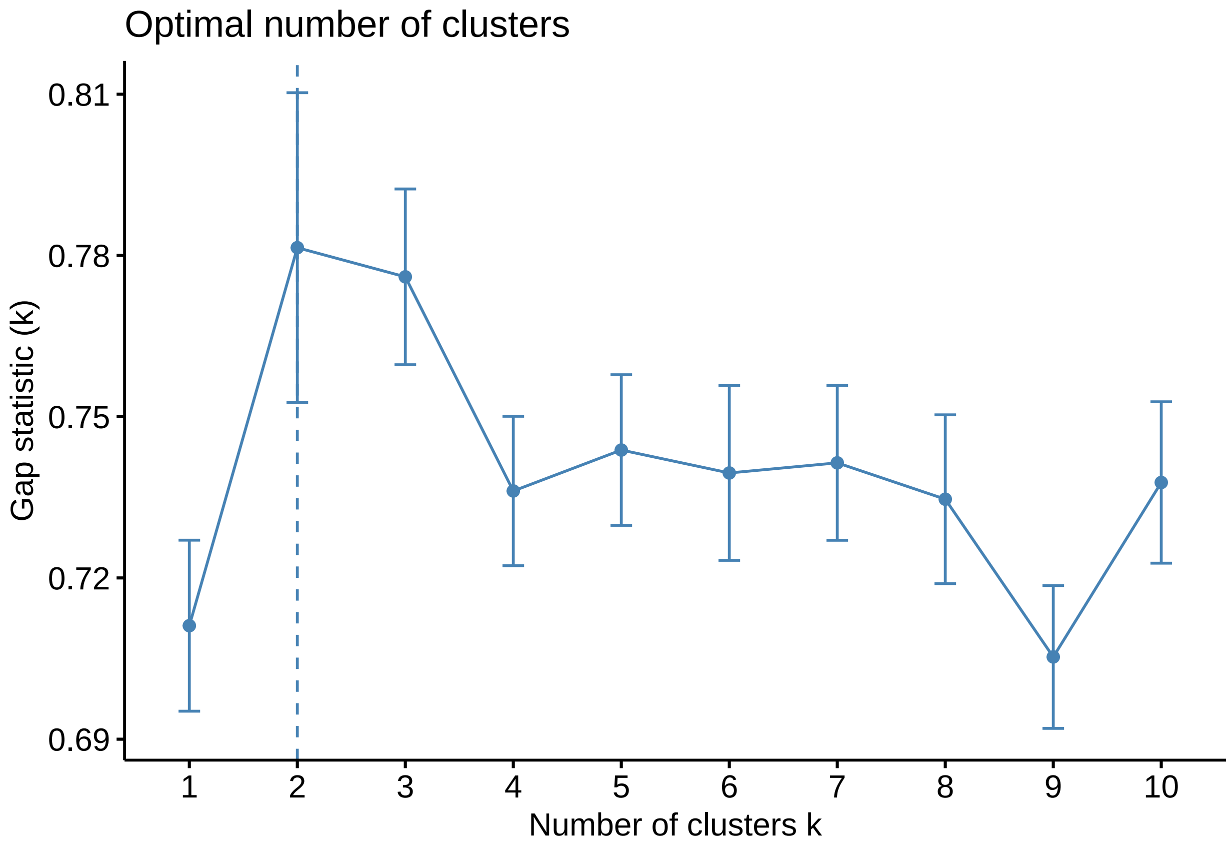
Let's start with k-means analysis.

- Use the fviz_nbclust function from the factoextra package to identify the optimal cluster solution under both the silhouette method and the gap statistic method.

```
fviz_nbclust(english_crime_prepped, FUNcluster = kmeans, method = "silhouette")
```



```
fviz_nbclust(english_crime_prepped, FUNcluster = kmeans, method = "gap")
```



- How many clusters do the silhouette and gap statistic methods recommend respectively?

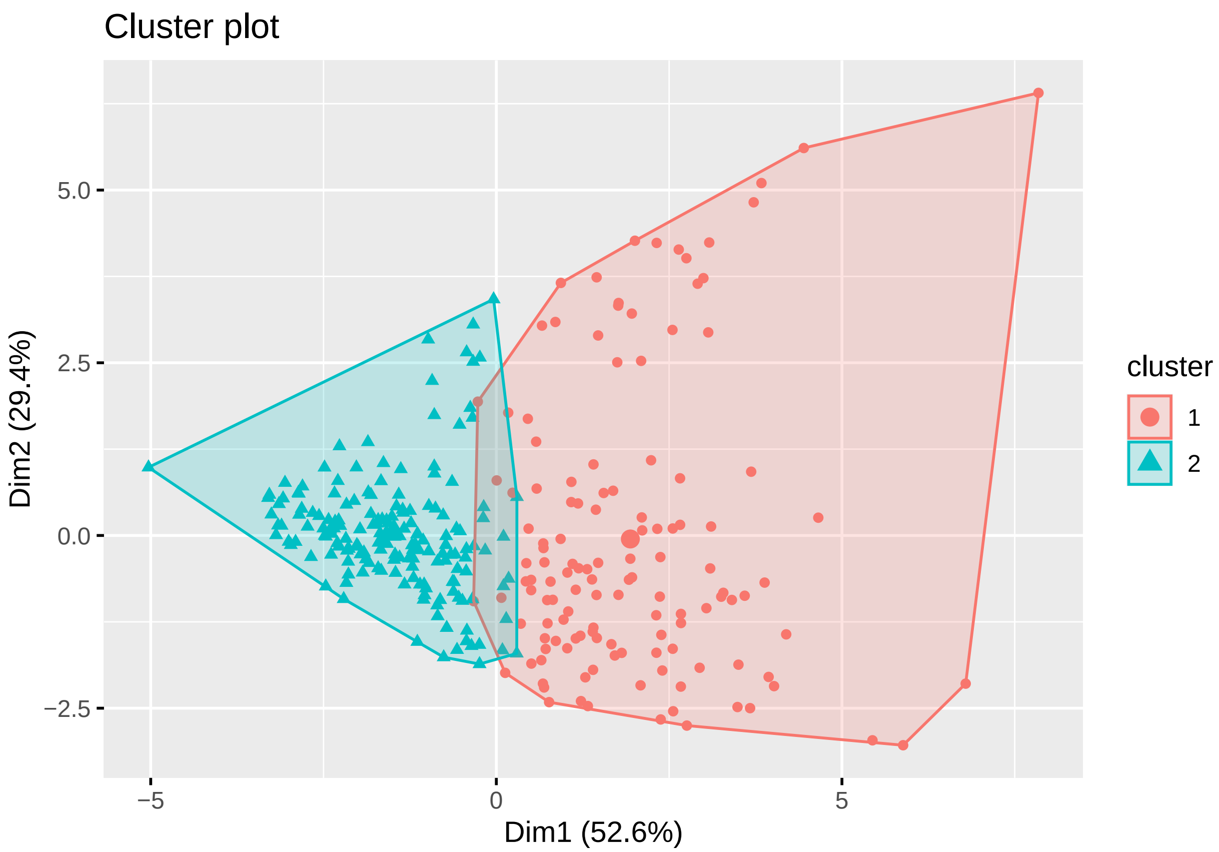
The silhouette method recommends a three cluster solution and the gap statistic recommends a two cluster solution.

- Create a 3-cluster and a 2-cluster solution for the English crime data using the `kmeans` function. Remember to save the results to an object for later use.

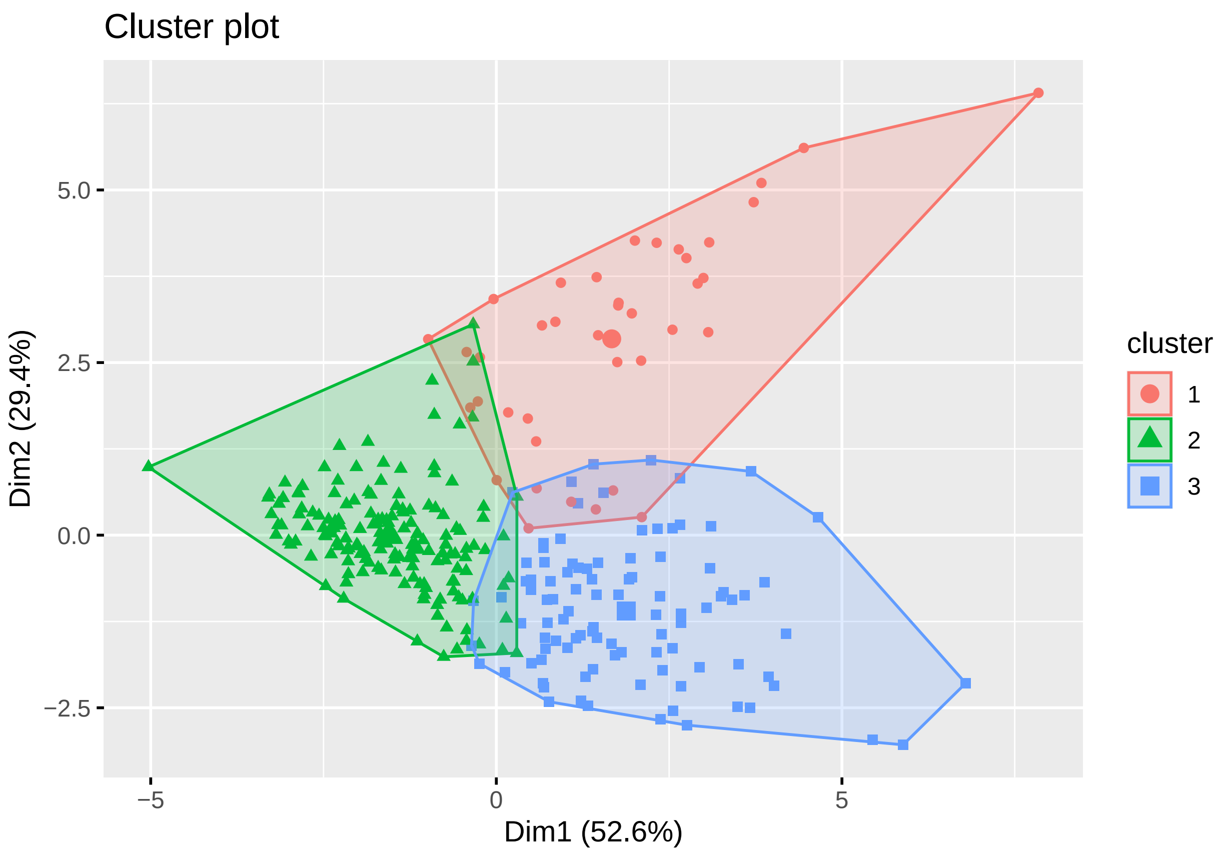
```
set.seed(2021)
english_km2 <- kmeans(english_crime_prepped, centers = 2)
set.seed(2021)
english_km3 <- kmeans(english_crime_prepped, centers = 3)
```

- Visualise the 2 cluster and 3 cluster solutions using the `fviz_clusters` function.

```
fviz_cluster(english_km2, data = english_crime_prepped, geom = "point")
```



```
fviz_cluster(english_km3, data = english_crime_prepped, geom = "point")
```



- By calling the k-means objects created earlier, report the proportion of variance that can be explained by the 2-cluster solution and the 3-cluster solution.

```
english_km2
```

```
## K-means clustering with 2 clusters of sizes 130, 170
##
## Cluster means:
##   violence_against_the_person sexual_offences  robbery theft_offences
## 1                    38.14092        3.176692 1.302154       26.95700
## 2                    22.82553        2.097706 0.359000       14.90271
##   vehicle_offences criminal_damage_and_arson drug_offences
## 1         6.600692                  9.933000      3.911769
## 2         3.623059                  6.525353      2.147353
##   public_order_offences
## 1             10.142538
## 2              5.965235
##
## Clustering vector:
##   [1] 2 1 2 2 2 1 2 1 1 2 2 2 1 2 2 1 1 1 1 1 1 1 2 1 1 2 2 2 2 2 1 1 2 2 2 2 2
##  [38] 2 2 2 2 2 1 2 1 1 2 1 1 2 2 2 2 1 2 1 2 1 1 1 1 2 2 1 1 1 2 2 2 2 1 2 2 1
##  [75] 2 2 1 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1
## [112] 2 1 1 2 2 1 1 1 2 2 1 2 1 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 1 1
## [149] 1 1 2 1 2 2 1 2 1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 1 1 2 1 1 2 1 2 1 1 1 1 2 2
## [186] 1 2 2 1 2 1 1 2 2 2 1 2 2 1 2 2 1 1 1 1 1 1 2 1 1 2 1 1 2 1 2 2 1 2 2 2 2
## [223] 1 1 1 1 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 2
## [260] 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1
## [297] 1 1 2 2
##
## Within cluster sum of squares by cluster:
## [1] 26518.56 10418.02
##  (between_SS / total_SS =  45.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
english_km3
```

```
## K-means clustering with 3 clusters of sizes 39, 163, 98
##
## Cluster means:
##   violence_against_the_person sexual_offences   robbery theft_offences
## 1                    26.44974        2.414359 2.2474359       35.94872
## 2                    22.75982        2.100675 0.3358282       14.51110
## 3                    41.80888        3.398061 0.8971429       23.16898
##   vehicle_offences criminal_damage_and_arson drug_offences
## 1        10.721026                  6.487949      4.844615
## 2         3.436503                  6.542270      2.116196
## 3         5.058571                 11.032449      3.466327
##   public_order_offences
## 1              7.197692
## 2              5.956319
## 3             11.030918
##
## Clustering vector:
##   [1] 2 3 2 2 2 1 2 3 1 2 2 2 3 2 2 3 3 3 3 3 3 3 2 3 3 2 2 2 2 2 3 3 2 2 2 2 2
##  [38] 2 2 2 2 2 3 2 3 1 2 3 3 2 2 3 2 3 2 1 2 3 3 1 3 2 2 3 3 3 2 2 2 2 3 2 2 3
##  [75] 2 2 3 2 2 2 2 2 2 2 2 2 3 3 2 2 2 2 2 1 2 2 3 2 1 2 2 3 3 3 3 3 3 3 3 3 3
## [112] 2 3 3 2 2 3 3 3 2 2 3 2 3 2 2 2 3 2 2 2 2 2 2 3 2 2 2 2 2 2 3 2 2 2 2 3 3
## [149] 3 3 2 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1 1 2 2
## [186] 3 2 2 3 2 3 3 3 2 2 3 2 2 3 2 2 3 3 3 3 3 3 2 3 3 2 1 3 2 3 2 2 1 2 2 2 2
## [223] 3 3 3 1 2 2 2 2 2 2 2 3 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 2
## [260] 2 2 2 2 2 2 2 2 3 1 3 3 2 2 2 2 2 2 2 3 2 2 2 2 2 2 3 3 3 3 3 1 3 3 3 3 3
## [297] 3 3 2 2
##
## Within cluster sum of squares by cluster:
## [1]  5490.214  8558.319 10519.318
##  (between_SS / total_SS =  63.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

- Do you find any particular solution preferable?

The three-cluster solution feels more intuitive as the inclusion of the points at the top of the graph within cluster 1 looks less approporite than an additional cluster. In addition, the variance explained in the three cluster solution was 63.9%, compared to 45.8%, which seems to be an appropriate trade off for the benefit of an additional cluster.

Now, we need to describe the clusters.

- By either calling the clusters, or adding them to the original data and then calculating summary statistic by group, get summary statistics for all of the clusters from the 2-cluster and 3- cluster solution.

```
english_crime_km_results <- english_crime %>%
  mutate(
    cluster_km2 = english_km2$cluster,
    cluster_km3 = english_km3$cluster,
  )

english_crime_km_results %>%
  group_by(cluster_km2) %>%
  summarise_if(is.numeric, mean) # %>% view()
```

```
## # A tibble: 2 x 10
##   cluster_km2 violence_against_the_person sexual_offences robbery theft_offences
##         <int>                       <dbl>           <dbl>   <dbl>          <dbl>
## 1           1                        38.1            3.18   1.30            27.0
## 2           2                        22.8            2.10   0.359           14.9
## # ... with 5 more variables: vehicle_offences <dbl>,
## #   criminal_damage_and_arson <dbl>, drug_offences <dbl>,
## #   public_order_offences <dbl>, cluster_km3 <dbl>
```

```
english_crime_km_results %>%
  group_by(cluster_km3) %>%
  summarise_if(is.numeric, mean) # %>% view()
```

```
## # A tibble: 3 x 10
##   cluster_km3 violence_against_the_person sexual_offences robbery theft_offences
##         <int>                       <dbl>           <dbl>   <dbl>          <dbl>
## 1           1                        26.4            2.41   2.25            35.9
## 2           2                        22.8            2.10   0.336           14.5
## 3           3                        41.8            3.40   0.897           23.2
## # ... with 5 more variables: vehicle_offences <dbl>,
## #   criminal_damage_and_arson <dbl>, drug_offences <dbl>,
## #   public_order_offences <dbl>, cluster_km2 <dbl>
```

- Describe the clusters found in the 2-cluster and 3-cluster solution.

2 Cluster Solution

The two cluster solution appears to show a straightforward split between low crime (cluster 2) and high crime (cluster 1).

3 Cluster Solution

The three cluster solution shows some potentially more interesting results. Cluster 1 appears to be characterised by relatively high property, vehicle, and drug crime (high robbery, high theft, high vehicle offences), but with relatively normal levels of other forms of crime. In contrast, cluster 3 appears to be characterised by high levels of violent crime, sexual offences, public order offences, and criminal damage and arson. Cluster 2 seems to represent relatively low crime. The three clusters could be labelled as follows:

Cluster 1 — High Incidence Property Crime Communities Cluster 2 — Low Overall Crime Communities Cluster 3 — High Incidence Interpersonal Crime Communities

- Is there evidence in either of these cluster solutions of areas being clustered into different types of criminal offences, or do clusters only reflect low or high crime as in the United States?

Yes, as opposed to the US data the three cluster solution appears to show the English community safety partnerships can be classified by prevalence of certain types of crime as well as overall crime.

Let's also see if we get similar results from a hierarchical cluster analysis. Before we can do that, we need to pick an appropriate dissimilarity/distance measure and linkage method. Your answers may start to differ from mine here — that's totally fine and to me expected! A lot of cluster analysis is subjective.

- What might be an appropriate distance measure for this data and why? (Hint: see slide 55)

Manhattan distance may be an appropriate measure of dissimilarity due to the large number of variables included in the data (8).

- What might be an appropriate linkage method for this data and why? (Hint: see slide 58)

Ward, Centroid, and Median linkage should be avoided because the distance matrix will be based on Manhattan and now Euclidean distance. Further, single linkage may not be appropriate as we do not assume a 'chain of command' type hierarchy in the data. As such, complete or average linkage may be most appropriate.

- Create a distance matrix for the English data using the dissimilarity measure of your choice.

```
english_crime_d <- daisy(english_crime_prepped, metric = "manhattan")
```

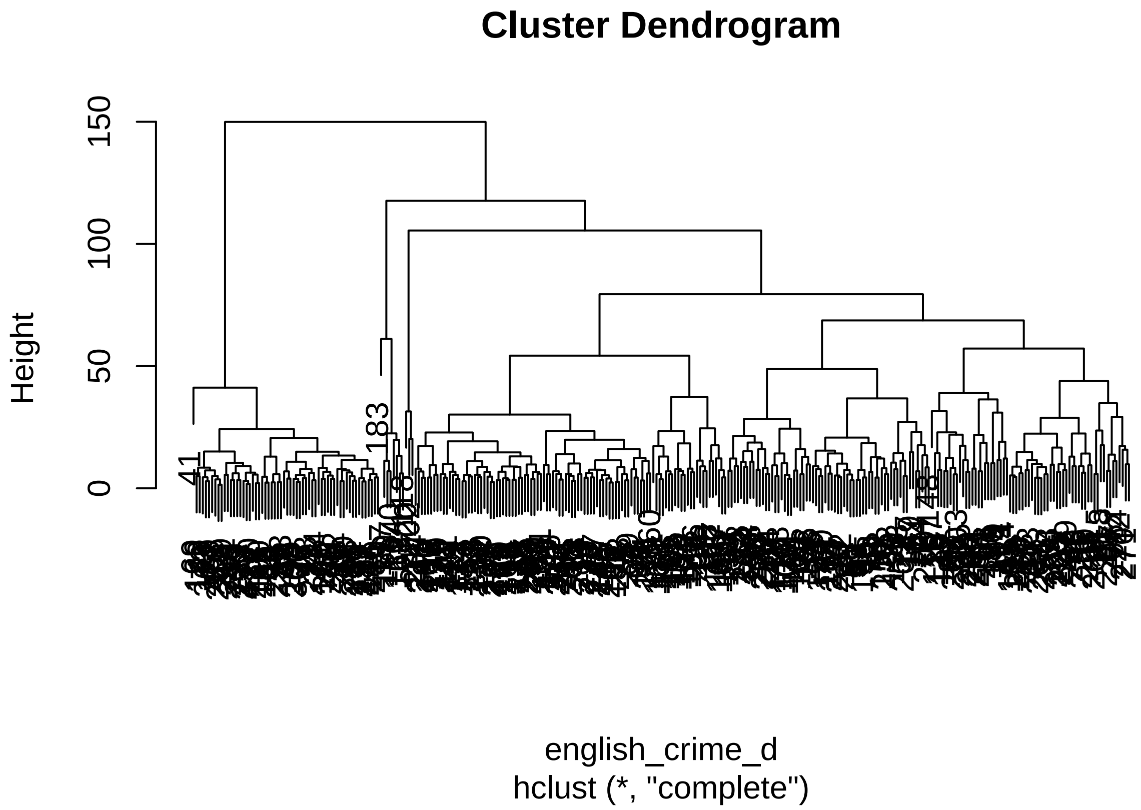
Now we can run the Hierarchical Cluster Analysis algorithm (or algorithms, if trying multiple) that we decided on about.

- Use the `hclust` function to cluster the data according to the linkage method that you chose. Don't forget to save the result to a new object.

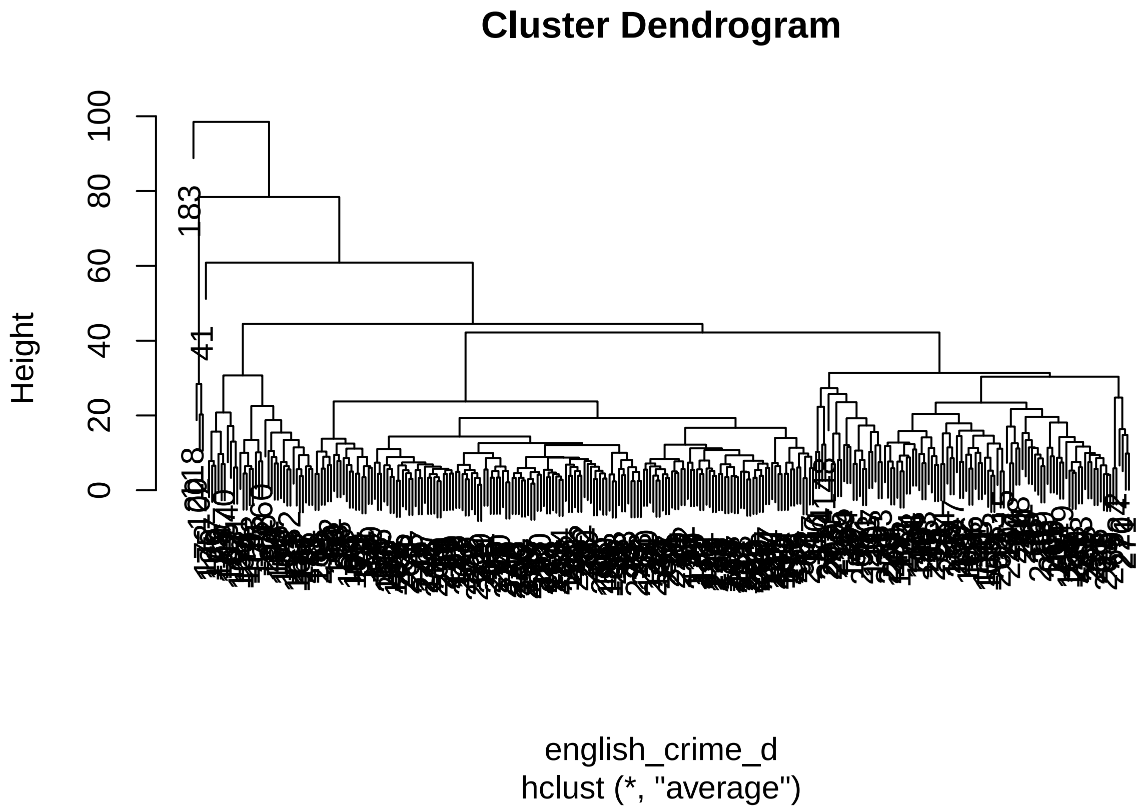
```
set.seed(2021)
english_crime_comp <- hclust(english_crime_d, method = "complete")

set.seed(2021)
english_crime_avg <- hclust(english_crime_d, method = "average")
```

- Plot the results of your HCA with a dendrogram using the `plot` function

```
plot(english_crime_comp) # Maybe 4
```



```r
plot(english_crime_avg) # Maybe 4
```



- Come up with a sensible number of clusters from the above plots that you think the data could be clustered into. Write how many clusters you think there may be in the data based on each dendrogram (if more than one).

Four seems to be an appropriate number of clusters for both dendrograms, though there could be some equally valid larger numbers of clusters. The average linkage dendrogram appears to have two singleton clades (Community 183 & 41), which may indicate outliers or potentially bad fitting clusters. Four also appears to be a defensible number of clusters for the average linkage dendrogram.

Because of the singleton clades in the average linkage, it will not be explored further. The complete linkage will be explored further.

Now we can cut our dendrogram into the number of clusters we believe we identified to explore how we might describe them.

- Use the `cutree` function to cut your dendrogram(s) into the chosen number of clusters.

```
english_hca_comp_results <- cutree(english_crime_comp, k = 4)
```

- Add your cluster solution(s) to the original data using the mutate function and the stored results above.

```
english_crime_hca_results <- english_crime %>%
  mutate(
    hca_comp = english_hca_comp_results,
  )
```

- Now produce some bivariate statistics showing how the crime rates differ by cluster.

```
english_crime_hca_results %>%
  select(-1, -2) %>%
  group_by(hca_comp) %>%
  summarise_all(~mean(., na.rm = TRUE))# %>% view()
```

```
## # A tibble: 4 x 9
##   hca_comp violence_against_the_person sexual_offences robbery theft_offences
##      <int>                       <dbl>           <dbl>   <dbl>          <dbl>
## 1        1                        31.9            2.72   0.781           21.1
## 2        2                        18.8            1.85   0.21            11.2
## 3        3                        61.7            4.84   1.47            35.4
## 4        4                        27.0            2.69   4.30            52.5
## # ... with 4 more variables: vehicle_offences <dbl>,
## #   criminal_damage_and_arson <dbl>, drug_offences <dbl>,
## #   public_order_offences <dbl>
```

- Describe the clusters found above (including from multiple linkage methods, if relevant). If possible, label the clusters found.

Complete Linkage Cluster (4)

Cluster 1: Average violence against person, average sexual offences, relatively low robbery, average theft offences, average vehicle offences, average criminal damage and arson offences, average drug offences, average public order offences. This cluster appears to represent "typical" communities with average levels of crime.

Cluster 2: Low violence against the person, low sexual offences, low robbert, low theft, low vehicle offences, low criminal damage and arson, low drug offences, low public order offences. This cluster appears to represent low-crime communities.

Cluster 3: Very high violence against the person, high sexual offences, average robbery, fairly high theft, average vehicle offences, very high criminal damage and arson, fairly high drug offences, very high public order offences. This cluster appears to represent high-crime communities characterised by interpersonal violent crime and public order violations.

Cluster 4: Average violence against the person, average sexual offences, very high robbert, very high theft, very high vehicle offences, average criminal damage and arson, high drug offences, average public order offences. This cluster appears to represent high-crime communities characterised by material or property-related crime (e.g. theft, robbery, and vehicle crime).

- Do these clusters differ from the clusters found using k-means? Which do you prefer as a typology of crime in English community and safety partnerships and why?

The HCA analysis led to one additional cluster being identified, which split 'low crime' between average and low crime. This may be useful in practice (e.g. in comparing low-crime areas to high crime areas), but equally the distinction is not overly interesting so a simple 3 cluster solution may be appropriate.

Conceptually, I prefer the four cluster solution found by HCA because it does also appear to have made the differences between the property-dominated high-crime areas and the interpersonal violence and disorder-dominated high-crime areas more distinct, with a better reference category in the low-crime cluster for further analysis.

Week 10 Challenge

- Practice using some of the skills we learned in Week 3 (bivariate data visualisation and statistics) to further illustrate the differences and similarities between your favoured cluster analysis of the English crime data. This might make it easier to see the characteristics of clusters than using the means of all variables; it might also show you some interesting differences in terms of variation.