Ocean Data Analysis with R Programming for Early Career Ocean Professionals (ECOPs) (Asia)

Mohamad Lukman Aidid bin Mohd Yusoff

2023-11-05

Assignment. Lesson 3: Clustering

1. Use the kmeans() function to perform k-means clustering on the data. What number of clusters did you choose and why?
2. Use the hclust() function to conduct hierarchical clustering on the data. What do you notice about the way the data points are clustered?
3. Use the cutree() function to extract the cluster assignments for each data point. How well are the clusters distributed?

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(stats)
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.3.2
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.2
```

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

```r
setwd('C:/Users/Administrator/Desktop/R/')
```

Question 1:

```
obis_malaysia <- read.csv("C:/Users/Administrator/Desktop/R/obis_malaysia/Occurrence.csv")
head(obis_malaysia)
```

```
##                                     id                           dataset_id
## 1 b268a4cb-4594-486e-947a-8d4aa1669b0c 2ed3cb38-e033-491d-95a7-5ae37f1f1904
## 2 15bce48e-3176-4025-b7c2-daea5f5624c4 2ed3cb38-e033-491d-95a7-5ae37f1f1904
## 3 dcc5cf4d-5426-47b8-8305-6dc591c5d4ae 36b58f28-0a03-4447-b688-e4eed56afa3d
## 4 9dd63c0c-054a-4e27-9da7-fa70728652b1 36b58f28-0a03-4447-b688-e4eed56afa3d
## 5 55c2a71b-f590-453f-ac41-25429f26cee3 7005b764-2feb-436a-8f17-0b41c6cd8435
## 6 11523180-c456-42b1-8a60-3f3217618f53 7005b764-2feb-436a-8f17-0b41c6cd8435
##   decimallongitude decimallatitude   date_start     date_mid     date_end
## 1         -74.0000         39.0000 1.057622e+12 1.057622e+12 1.057622e+12
## 2         -74.0000         39.0000 1.057622e+12 1.057622e+12 1.057622e+12
## 3         -66.7833         41.4667 8.983872e+11 8.983872e+11 8.983872e+11
## 4         -66.7833         41.4667 8.983872e+11 8.983872e+11 8.983872e+11
## 5          69.6830         22.4000 2.481408e+11 2.481408e+11 2.481408e+11
## 6          73.0000         10.0000 4.844448e+11 4.844448e+11 4.844448e+11
##   date_year        scientificname originalscientificname minimumdepthinmeters
## 1      2003 Merluccius bilinearis  Merluccius bilinearis                   NA
## 2      2003 Merluccius bilinearis  Merluccius bilinearis                   NA
## 3      1998  Calanus finmarchicus   Calanus finmarchicus                    0
## 4      1998  Calanus finmarchicus   Calanus finmarchicus                    0
## 5      1977            Ulva linza     Enteromorpha linza                   NA
## 6      1985            Ulva linza     Enteromorpha linza                   NA
##   maximumdepthinmeters depth coordinateuncertaintyinmeters      flags dropped
## 1                   NA    NA                            NA {NO_DEPTH}       0
## 2                   NA    NA                            NA {NO_DEPTH}       0
## 3                   50    25                            NA         {}       0
## 4                   50    25                            NA         {}       0
## 5                   NA    NA                            NA                  0
## 6                   NA    NA                            NA                  0
##   absence shoredistance bathymetry   sst   sss marine brackish freshwater
## 1       0         53558         37 15.17 32.96      1        0          0
## 2       0         53558         37 15.17 32.96      1        0          0
## 3       0        222884         70 11.12 32.42      1        0          0
## 4       0        222884         70 11.12 32.42      1        0          0
## 5       0           837         -1 26.73 35.05      1       NA         NA
## 6       0         69042       2012 28.89 34.95      1       NA         NA
##   terrestrial taxonrank aphiaid redlist_category superdomain domain  kingdom
## 1           0   Species  158962               NT       Biota     NA Animalia
## 2           0   Species  158962               NT       Biota     NA Animalia
## 3           0   Species  104464                        Biota     NA Animalia
## 4           0   Species  104464                        Biota     NA Animalia
## 5          NA   Species  234474                        Biota     NA  Plantae
## 6          NA   Species  234474                        Biota     NA  Plantae
##      subkingdom infrakingdom     phylum phylum_division subphylum_subdivision
## 1                              Chordata
## 2                              Chordata
## 3                            Arthropoda
## 4                            Arthropoda
## 5 Viridiplantae                             Chlorophyta         Chlorophytina
## 6 Viridiplantae                             Chlorophyta         Chlorophytina
##    subphylum   infraphylum   parvphylum      gigaclass megaclass     superclass
## 1 Vertebrata Gnathostomata Osteichthyes Actinopterygii              Actinopteri
## 2 Vertebrata Gnathostomata Osteichthyes Actinopterygii              Actinopteri
## 3  Crustacea                                                     Multicrustacea
## 4  Crustacea                                                     Multicrustacea
## 5
## 6
##         class  subclass  infraclass subterclass superorder      order suborder
## 1   Teleostei Teleostei                                    Gadiformes
## 2   Teleostei Teleostei                                    Gadiformes
## 3    Copepoda           Neocopepoda              Gymnoplea  Calanoida
## 4    Copepoda           Neocopepoda              Gymnoplea  Calanoida
## 5 Ulvophyceae                                                 Ulvales
## 6 Ulvophyceae                                                 Ulvales
##   infraorder parvorder superfamily       family    subfamily supertribe tribe
## 1                                  Merlucciidae Merlucciinae         NA
## 2                                  Merlucciidae Merlucciinae         NA
## 3                                     Calanidae                      NA
## 4                                     Calanidae                      NA
## 5                                      Ulvaceae                      NA
## 6                                      Ulvaceae                      NA
##   subtribe      genus subgenus section subsection series             species
## 1       NA Merluccius                                 NA Merluccius bilinearis
## 2       NA Merluccius                                 NA Merluccius bilinearis
## 3       NA    Calanus                                 NA  Calanus finmarchicus
## 4       NA    Calanus                                 NA  Calanus finmarchicus
## 5       NA       Ulva                                 NA            Ulva linza
## 6       NA       Ulva                                 NA            Ulva linza
##   subspecies natio variety subvariety forma subforma type modified language
## 1               NA                 NA             NA   NA                NA
## 2               NA                 NA             NA   NA                NA
## 3               NA                 NA             NA   NA                NA
## 4               NA                 NA             NA   NA                NA
## 5               NA                 NA             NA   NA                NA
## 6               NA                 NA             NA   NA                NA
##   license rightsholder accessrights bibliographiccitation references
## 1      NA           NA           NA                               NA
## 2      NA           NA           NA                               NA
## 3      NA           NA           NA                               NA
## 4      NA           NA           NA                               NA
## 5      NA           NA           NA                               NA
## 6      NA           NA           NA                               NA
##   institutionid collectionid datasetid institutioncode collectioncode
## 1            NA           NA        NA           RUMFS   OTTE-03-0037
## 2            NA           NA        NA           RUMFS   OTTE-03-0038
## 3            NA           NA        NA         Zoogene
## 4            NA           NA        NA         Zoogene
## 5            NA           NA        NA             NIO            168
## 6            NA           NA        NA             NIO            152
##   datasetname ownerinstitutioncode    basisofrecord informationwithheld
## 1                               NA HumanObservation                  NA
## 2                               NA HumanObservation                  NA
## 3                               NA HumanObservation                  NA
## 4                               NA HumanObservation                  NA
## 5                               NA HumanObservation                  NA
## 6                               NA HumanObservation                  NA
##   datageneralizations          dynamicproperties materialsampleid occurrenceid
## 1                  NA observedindividualcount=1;               NA
## 2                  NA observedindividualcount=1;               NA
## 3                  NA                                          NA
## 4                  NA                                          NA
## 5                  NA                                          NA
## 6                  NA                                          NA
##   catalognumber
## 1
## 2
## 3
## 4
## 5        NIO186
## 6        NIO167
##
## 1 location=OT 2;gear=OTTE;mesh=6;length=4;areasampled=1;heading=1;towdir=;sudo=7.8;bodo=6.78;susal=3
## 2        location=STA15;gear=OTTE;mesh=6;length=4;areasampled=1;heading=3;towdir=UP CRK;sudo=6.54;bo
## 3
## 4
## 5
## 6
##   recordnumber             recordedby recordedbyid individualcount
## 1                                               NA              NA
## 2                                               NA              NA
## 3                                               NA              NA
## 4                                               NA              NA
## 5            6 Dr. Arvind G. Untawale           NA              NA
## 6            0                Unknown           NA              NA
##   organismquantity organismquantitytype  sex lifestage reproductivecondition
## 1               NA                   NA             NA                    NA
## 2               NA                   NA             NA                    NA
## 3               NA                   NA             NA                    NA
## 4               NA                   NA             NA                    NA
## 5               NA                   NA Male        NA                    NA
## 6               NA                   NA Male        NA                    NA
##   behavior establishmentmeans occurrencestatus preparations disposition
## 1       NA                 NA
## 2       NA                 NA
## 3       NA                 NA
## 4       NA                 NA
## 5       NA                 NA
## 6       NA                 NA
##   othercatalognumbers associatedmedia associatedreferences associatedsequences
## 1                  NA              NA                                       NA
## 2                  NA              NA                                       NA
## 3                  NA              NA                                       NA
## 4                  NA              NA                                       NA
## 5                  NA              NA                                       NA
## 6                  NA              NA                                       NA
##   associatedtaxa organismid organismname organismscope associatedoccurrences
## 1             NA         NA           NA            NA                    NA
## 2             NA         NA           NA            NA                    NA
## 3             NA         NA           NA            NA                    NA
## 4             NA         NA           NA            NA                    NA
## 5             NA         NA           NA            NA                    NA
## 6             NA         NA           NA            NA                    NA
##   associatedorganisms previousidentifications organismremarks eventid
## 1                  NA                      NA
## 2                  NA                      NA
## 3                  NA                      NA
## 4                  NA                      NA
## 5                  NA                      NA
## 6                  NA                      NA
##   parenteventid samplingprotocol samplesizevalue samplesizeunit samplingeffort
## 1            NA                               NA             NA             NA
## 2            NA                               NA             NA             NA
## 3            NA                               NA             NA             NA
## 4            NA                               NA             NA             NA
## 5            NA                               NA             NA             NA
## 6            NA                               NA             NA             NA
##              eventdate eventtime startdayofyear enddayofyear year month day
## 1 2003-07-08T12:57:00Z        NA             NA           NA 2003     7   8
## 2 2003-07-08T12:57:00Z        NA             NA           NA 2003     7   8
## 3 1998-06-21T12:00:00Z        NA             NA           NA 1998     6  21
## 4 1998-06-21T12:00:00Z        NA             NA           NA 1998     6  21
## 5 1977-11-12T12:00:00Z        NA             NA           NA 1977    11  12
## 6 1985-05-09T12:00:00Z        NA             NA           NA 1985     5   9
##   verbatimeventdate habitat fieldnumber fieldnotes eventremarks locationid
## 1                                    NA         NA           NA         NA
## 2                                    NA         NA           NA         NA
## 3                                    NA         NA           NA         NA
## 4                                    NA         NA           NA         NA
## 5                                   155         NA           NA         NA
## 6                                   139         NA           NA         NA
##   highergeographyid highergeography continent waterbody islandgroup island
## 1                NA              NA        NA                    NA     NA
## 2                NA              NA        NA                    NA     NA
## 3                NA              NA        NA                    NA     NA
## 4                NA              NA        NA                    NA     NA
## 5                NA              NA        NA                    NA     NA
## 6                NA              NA        NA                    NA     NA
##   country countrycode stateprovince county municipality    locality
## 1                  NA                                NA
## 2                  NA                                NA
## 3                  NA                                NA
## 4                  NA                                NA
## 5   India          NA       Gujarat                  NA     Pirotan
## 6   India          NA   Lakshadweep                  NA Lakshadweep
##   verbatimlocality verbatimelevation minimumelevationinmeters
## 1               NA                NA                       NA
## 2               NA                NA                       NA
## 3               NA                NA                       NA
## 4               NA                NA                       NA
## 5               NA                NA                       NA
## 6               NA                NA                       NA
##   maximumelevationinmeters verbatimdepth minimumdistanceabovesurfaceinmeters
## 1                       NA            NA                                  NA
## 2                       NA            NA                                  NA
## 3                       NA            NA                                  NA
## 4                       NA            NA                                  NA
## 5                       NA            NA                                  NA
## 6                       NA            NA                                  NA
##   maximumdistanceabovesurfaceinmeters locationaccordingto locationremarks
## 1                                  NA                  NA              NA
## 2                                  NA                  NA              NA
## 3                                  NA                  NA              NA
## 4                                  NA                  NA              NA
## 5                                  NA                  NA              NA
## 6                                  NA                  NA              NA
##   verbatimcoordinates verbatimlatitude verbatimlongitude
## 1                  NA               NA                NA
## 2                  NA               NA                NA
## 3                  NA               NA                NA
## 4                  NA               NA                NA
## 5                  NA               NA                NA
## 6                  NA               NA                NA
##   verbatimcoordinatesystem verbatimsrs geodeticdatum coordinateprecision
## 1                       NA          NA                                NA
## 2                       NA          NA                                NA
## 3                       NA          NA                                NA
## 4                       NA          NA                                NA
## 5                       NA          NA                                NA
## 6                       NA          NA                                NA
##   pointradiusspatialfit footprintwkt footprintsrs footprintspatialfit
## 1                    NA                        NA                  NA
## 2                    NA                        NA                  NA
## 3                    NA                        NA                  NA
## 4                    NA                        NA                  NA
## 5                    NA                        NA                  NA
## 6                    NA                        NA                  NA
##   georeferencedby georeferenceddate georeferenceprotocol georeferencesources
## 1              NA                NA                   NA                  NA
## 2              NA                NA                   NA                  NA
## 3              NA                NA                   NA                  NA
## 4              NA                NA                   NA                  NA
## 5              NA                NA                   NA                  NA
## 6              NA                NA                   NA                  NA
##   georeferenceverificationstatus georeferenceremarks geologicalcontextid
## 1                             NA                  NA                  NA
## 2                             NA                  NA                  NA
## 3                             NA                  NA                  NA
## 4                             NA                  NA                  NA
## 5                             NA                  NA                  NA
## 6                             NA                  NA                  NA
##   earliesteonorlowesteonothem latesteonorhighesteonothem
## 1                          NA                         NA
## 2                          NA                         NA
## 3                          NA                         NA
## 4                          NA                         NA
## 5                          NA                         NA
## 6                          NA                         NA
##   earliesteraorlowesterathem latesteraorhighesterathem
## 1                         NA                        NA
## 2                         NA                        NA
## 3                         NA                        NA
## 4                         NA                        NA
## 5                         NA                        NA
## 6                         NA                        NA
##   earliestperiodorlowestsystem latestperiodorhighestsystem
## 1                           NA                          NA
## 2                           NA                          NA
## 3                           NA                          NA
## 4                           NA                          NA
## 5                           NA                          NA
## 6                           NA                          NA
##   earliestepochorlowestseries latestepochorhighestseries
## 1                          NA                         NA
## 2                          NA                         NA
## 3                          NA                         NA
## 4                          NA                         NA
## 5                          NA                         NA
## 6                          NA                         NA
##   earliestageorloweststage latestageorhigheststage lowestbiostratigraphiczone
## 1                       NA                      NA                         NA
## 2                       NA                      NA                         NA
## 3                       NA                      NA                         NA
## 4                       NA                      NA                         NA
## 5                       NA                      NA                         NA
## 6                       NA                      NA                         NA
##   highestbiostratigraphiczone lithostratigraphicterms group formation member
## 1                          NA                      NA    NA        NA     NA
## 2                          NA                      NA    NA        NA     NA
## 3                          NA                      NA    NA        NA     NA
## 4                          NA                      NA    NA        NA     NA
## 5                          NA                      NA    NA        NA     NA
## 6                          NA                      NA    NA        NA     NA
##   bed identificationid identifiedby identifiedbyid dateidentified
## 1  NA               NA                          NA
## 2  NA               NA                          NA
## 3  NA               NA                          NA
## 4  NA               NA                          NA
## 5  NA               NA      Unknown             NA
## 6  NA               NA      Unknown             NA
##   identificationreferences identificationremarks identificationqualifier
## 1                       NA                    NA                      NA
## 2                       NA                    NA                      NA
## 3                       NA                    NA                      NA
## 4                       NA                    NA                      NA
## 5                       NA                    NA                      NA
## 6                       NA                    NA                      NA
##   identificationverificationstatus typestatus taxonid
## 1                               NA                 NA
## 2                               NA                 NA
## 3                               NA                 NA
## 4                               NA                 NA
## 5                               NA                 NA
## 6                               NA                 NA
##                             scientificnameid acceptednameusageid
## 1 urn:lsid:marinespecies.org:taxname:158962                  NA
## 2 urn:lsid:marinespecies.org:taxname:158962                  NA
## 3 urn:lsid:marinespecies.org:taxname:104464                  NA
## 4 urn:lsid:marinespecies.org:taxname:104464                  NA
## 5 urn:lsid:marinespecies.org:taxname:145967                  NA
## 6 urn:lsid:marinespecies.org:taxname:145967                  NA
##   parentnameusageid originalnameusageid nameaccordingtoid namepublishedinid
## 1                NA                  NA                NA                NA
## 2                NA                  NA                NA                NA
## 3                NA                  NA                NA                NA
## 4                NA                  NA                NA                NA
## 5                NA                  NA                NA                NA
## 6                NA                  NA                NA                NA
##   taxonconceptid acceptednameusage parentnameusage originalnameusage
## 1             NA                NA              NA                NA
## 2             NA                NA              NA                NA
## 3             NA                NA              NA                NA
## 4             NA                NA              NA                NA
## 5             NA                NA              NA                NA
## 6             NA                NA              NA                NA
##   nameaccordingto namepublishedin namepublishedinyear higherclassification
## 1              NA              NA                  NA                   NA
## 2              NA              NA                  NA                   NA
## 3              NA              NA                  NA                   NA
## 4              NA              NA                  NA                   NA
## 5              NA              NA                  NA                   NA
## 6              NA              NA                  NA                   NA
##   specificepithet infraspecificepithet verbatimtaxonrank
## 1                                                     NA
## 2                                                     NA
## 3                                                     NA
## 4                                                     NA
## 5                                                     NA
## 6                                                     NA
##   scientificnameauthorship vernacularname nomenclaturalcode taxonomicstatus
## 1                                                        NA              NA
## 2                                                        NA              NA
## 3                                                        NA              NA
## 4                                                        NA              NA
## 5                J. Agardh                               NA              NA
## 6                J. Agardh                               NA              NA
##   nomenclaturalstatus taxonremarks
## 1                  NA           NA
## 2                  NA           NA
## 3                  NA           NA
## 4                  NA           NA
## 5                  NA           NA
## 6                  NA           NA
```

```
str(obis_malaysia)
```

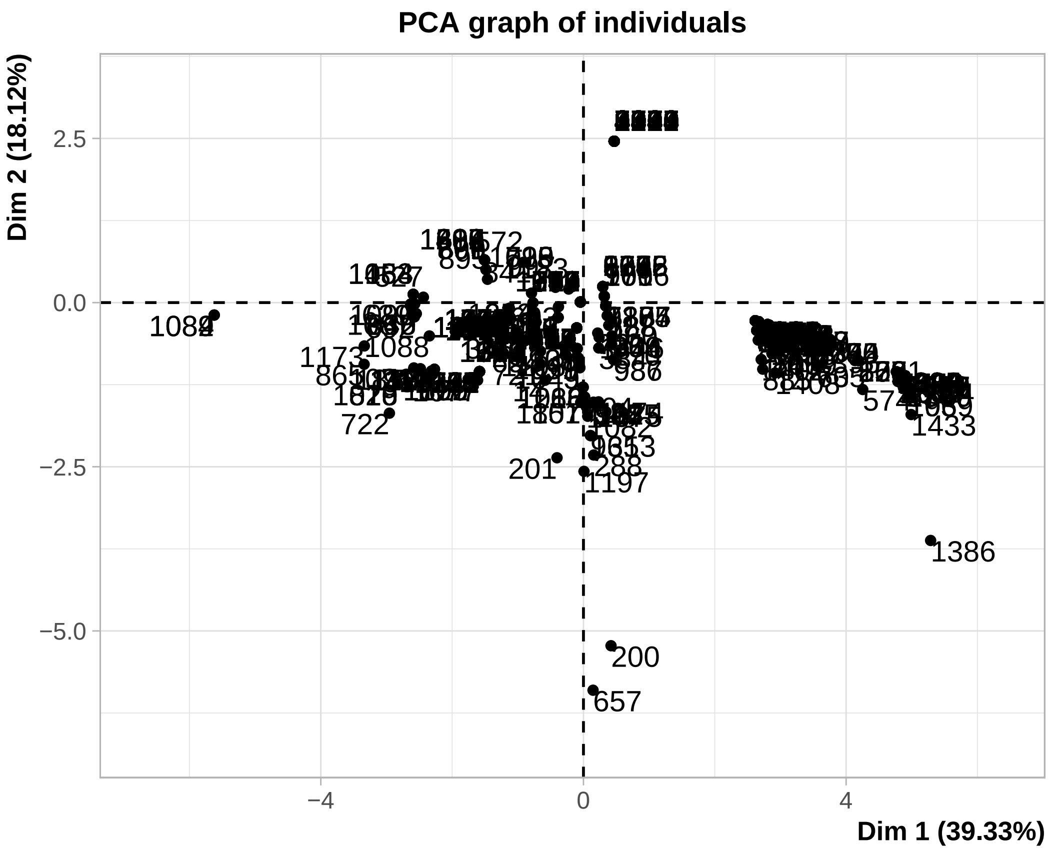
```
## ’data.frame’:    51974 obs. of  226 variables:
##  $ id                                 : chr  "b268a4cb-4594-486e-947a-8d4aa1669b0c" "15bce48e-3176-4
##  $ dataset_id                         : chr  "2ed3cb38-e033-491d-95a7-5ae37f1f1904" "2ed3cb38-e033-4
##  $ decimallongitude                   : num  -74 -74 -66.8 -66.8 69.7 ...
##  $ decimallatitude                    : num  39 39 41.5 41.5 22.4 ...
##  $ date_start                         : num  1.06e+12 1.06e+12 8.98e+11 8.98e+11 2.48e+11 ...
##  $ date_mid                           : num  1.06e+12 1.06e+12 8.98e+11 8.98e+11 2.48e+11 ...
##  $ date_end                           : num  1.06e+12 1.06e+12 8.98e+11 8.98e+11 2.48e+11 ...
##  $ date_year                          : int  2003 2003 1998 1998 1977 1985 1985 2006 NA 2007 ...
##  $ scientificname                     : chr  "Merluccius bilinearis" "Merluccius bilinearis" "Calanus
##  $ originalscientificname             : chr  "Merluccius bilinearis" "Merluccius bilinearis" "Calanus
##  $ minimumdepthinmeters               : num  NA NA 0 0 NA NA NA NA NA NA ...
##  $ maximumdepthinmeters               : num  NA NA 50 50 NA NA NA NA NA NA ...
##  $ depth                              : num  NA NA 25 25 NA NA NA NA NA NA ...
##  $ coordinateuncertaintyinmeters      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ flags                              : chr  "{NO_DEPTH}" "{NO_DEPTH}" "{}" "{}" ...
##  $ dropped                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ absence                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ shoredistance                      : int  53558 53558 222884 222884 837 69042 69042 -1328 83 -555
##  $ bathymetry                         : num  37 37 70 70 -1 ...
##  $ sst                                : num  15.2 15.2 11.1 11.1 26.7 ...
##  $ sss                                : num  33 33 32.4 32.4 35 ...
##  $ marine                             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ brackish                           : int  0 0 0 0 NA NA NA 0 NA NA ...
##  $ freshwater                         : int  0 0 0 0 NA NA NA 0 NA NA ...
##  $ terrestrial                        : int  0 0 0 0 NA NA NA 0 NA NA ...
##  $ taxonrank                          : chr  "Species" "Species" "Species" "Species" ...
##  $ aphiaid                            : int  158962 158962 104464 104464 234474 234474 234474 510565
##  $ redlist_category                   : chr  "NT" "NT" "" "" ...
##  $ superdomain                        : chr  "Biota" "Biota" "Biota" "Biota" ...
##  $ domain                             : logi  NA NA NA NA NA NA ...
##  $ kingdom                            : chr  "Animalia" "Animalia" "Animalia" "Animalia" ...
##  $ subkingdom                         : chr  "" "" "" "" ...
##  $ infrakingdom                       : chr  "" "" "" "" ...
##  $ phylum                             : chr  "Chordata" "Chordata" "Arthropoda" "Arthropoda" ...
##  $ phylum_division                    : chr  "" "" "" "" ...
##  $ subphylum_subdivision              : chr  "" "" "" "" ...
##  $ subphylum                          : chr  "Vertebrata" "Vertebrata" "Crustacea" "Crustacea" ...
##  $ infraphylum                        : chr  "Gnathostomata" "Gnathostomata" "" "" ...
##  $ parvphylum                         : chr  "Osteichthyes" "Osteichthyes" "" "" ...
##  $ gigaclass                          : chr  "Actinopterygii" "Actinopterygii" "" "" ...
##  $ megaclass                          : chr  "" "" "" "" ...
##  $ superclass                         : chr  "Actinopteri" "Actinopteri" "Multicrustacea" "Multicrus
##  $ class                              : chr  "Teleostei" "Teleostei" "Copepoda" "Copepoda" ...
##  $ subclass                           : chr  "Teleostei" "Teleostei" "" "" ...
##  $ infraclass                         : chr  "" "" "Neocopepoda" "Neocopepoda" ...
##  $ subterclass                        : chr  "" "" "" "" ...
##  $ superorder                         : chr  "" "" "Gymnoplea" "Gymnoplea" ...
##  $ order                              : chr  "Gadiformes" "Gadiformes" "Calanoida" "Calanoida" ...
##  $ suborder                           : chr  "" "" "" "" ...
##  $ infraorder                         : chr  "" "" "" "" ...
##  $ parvorder                          : chr  "" "" "" "" ...
##  $ superfamily                        : chr  "" "" "" "" ...
##  $ family                             : chr  "Merlucciidae" "Merlucciidae" "Calanidae" "Calanidae" .
##  $ subfamily                          : chr  "Merlucciinae" "Merlucciinae" "" "" ...
##  $ supertribe                         : logi  NA NA NA NA NA NA ...
##  $ tribe                              : chr  "" "" "" "" ...
##  $ subtribe                           : logi  NA NA NA NA NA NA ...
##  $ genus                              : chr  "Merluccius" "Merluccius" "Calanus" "Calanus" ...
##  $ subgenus                           : chr  "" "" "" "" ...
##  $ section                            : chr  "" "" "" "" ...
##  $ subsection                         : chr  "" "" "" "" ...
##  $ series                             : logi  NA NA NA NA NA NA ...
##  $ species                            : chr  "Merluccius bilinearis" "Merluccius bilinearis" "Calanus
##  $ subspecies                         : chr  "" "" "" "" ...
##  $ natio                              : logi  NA NA NA NA NA NA ...
##  $ variety                            : chr  "" "" "" "" ...
##  $ subvariety                         : logi  NA NA NA NA NA NA ...
##  $ forma                              : chr  "" "" "" "" ...
##  $ subforma                           : logi  NA NA NA NA NA NA ...
##  $ type                               : logi  NA NA NA NA NA NA ...
##  $ modified                           : chr  "" "" "" "" ...
##  $ language                           : logi  NA NA NA NA NA NA ...
##  $ license                            : logi  NA NA NA NA NA NA ...
##  $ rightsholder                       : logi  NA NA NA NA NA NA ...
##  $ accessrights                       : logi  NA NA NA NA NA NA ...
##  $ bibliographiccitation              : chr  "" "" "" "" ...
##  $ references                         : logi  NA NA NA NA NA NA ...
##  $ institutionid                      : logi  NA NA NA NA NA NA ...
##  $ collectionid                       : logi  NA NA NA NA NA NA ...
##  $ datasetid                          : logi  NA NA NA NA NA NA ...
##  $ institutioncode                    : chr  "RUMFS" "RUMFS" "Zoogene" "Zoogene" ...
##  $ collectioncode                     : chr  "OTTE-03-0037" "OTTE-03-0038" "" "" ...
##  $ datasetname                        : chr  "" "" "" "" ...
##  $ ownerinstitutioncode               : logi  NA NA NA NA NA NA ...
##  $ basisofrecord                      : chr  "HumanObservation" "HumanObservation" "HumanObservation
##  $ informationwithheld                : logi  NA NA NA NA NA NA ...
##  $ datageneralizations                : logi  NA NA NA NA NA NA ...
##  $ dynamicproperties                  : chr  "observedindividualcount=1;" "observedindividualcount=1
##  $ materialsampleid                   : logi  NA NA NA NA NA NA ...
##  $ occurrenceid                       : chr  "" "" "" "" ...
##  $ catalognumber                      : chr  "" "" "" "" ...
##  $ occurrenceremarks                  : chr  "location=OT 2;gear=OTTE;mesh=6;length=4;areasampled=1;
##  $ recordnumber                       : chr  "" "" "" "" ...
##  $ recordedby                         : chr  "" "" "" "" ...
##  $ recordedbyid                       : logi  NA NA NA NA NA NA ...
##  $ individualcount                    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ organismquantity                   : logi  NA NA NA NA NA NA ...
##  $ organismquantitytype               : logi  NA NA NA NA NA NA ...
##  $ sex                                : chr  "" "" "" "" ...
##   [list output truncated]
```

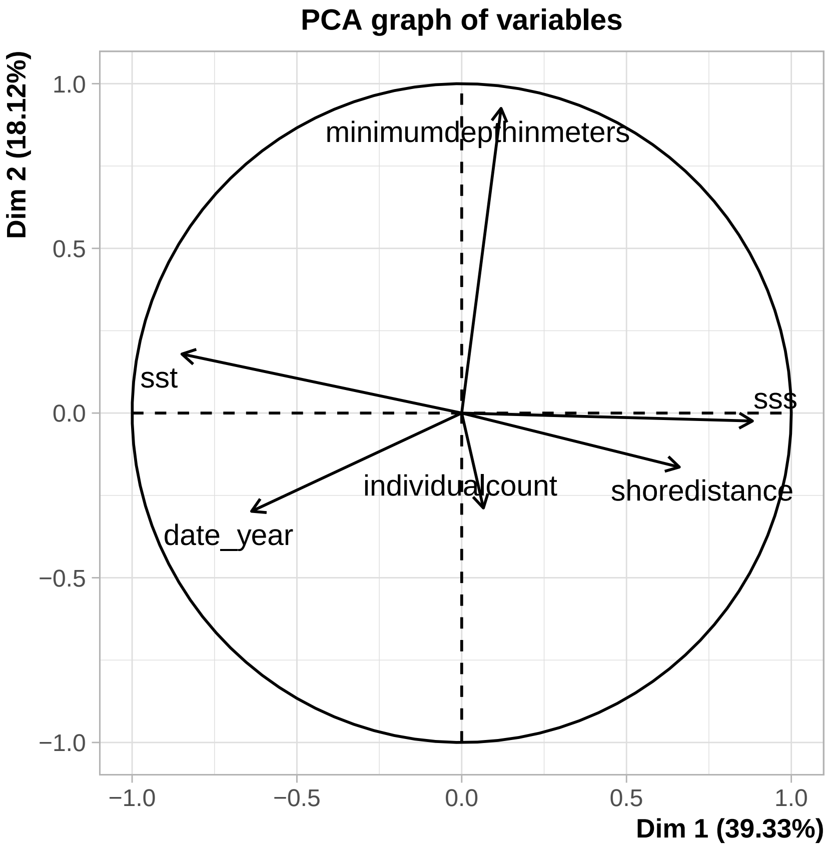
Question 1:

```
data <- obis_malaysia %>% select(date_year, minimumdepthinmeters, shoredistance, sst, sss, individualcou
clusters <- kmeans(data, 3)
print(clusters)
```

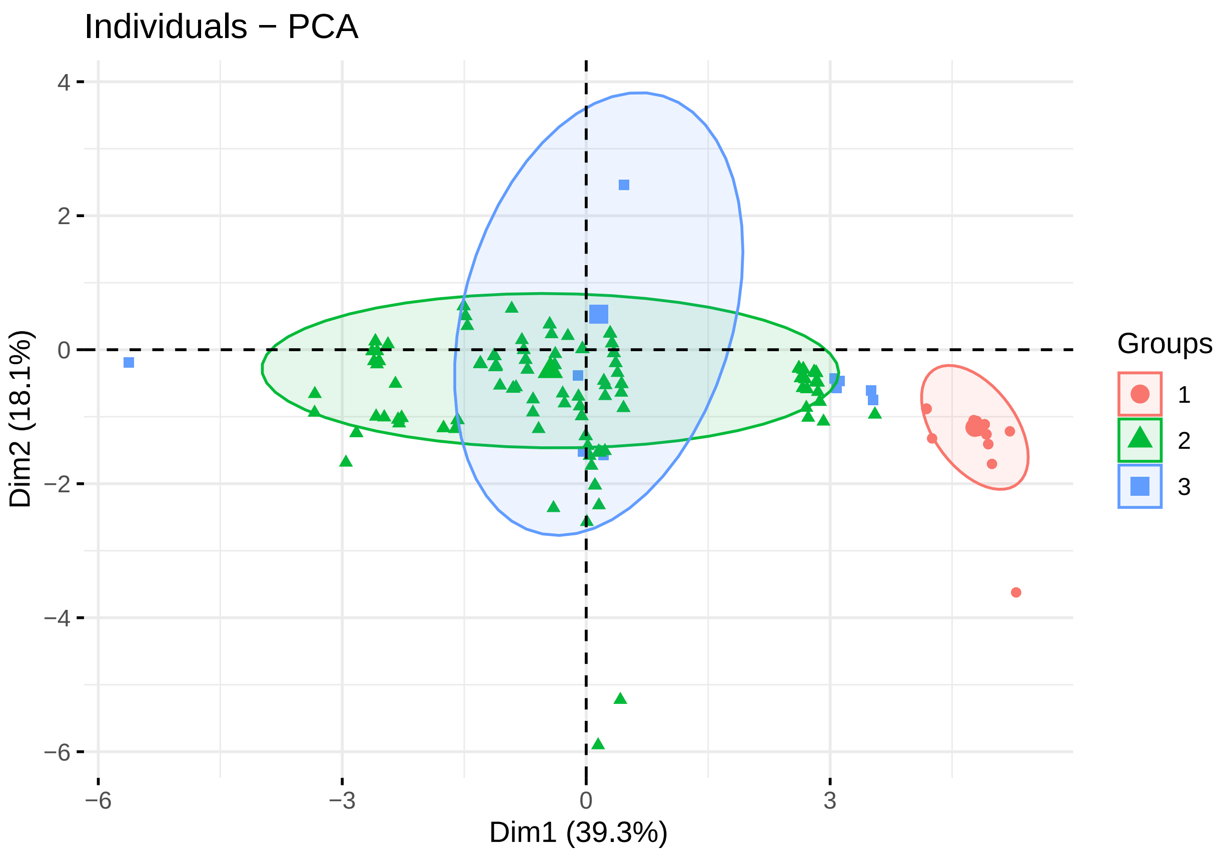
```
## K-means clustering with 3 clusters of sizes 55, 801, 590
##
## Cluster means:
##   date_year minimumdepthinmeters shoredistance      sst      sss
## 1  2004.000              0.00000   172983.8364 28.88800 32.41964
## 2  2007.337              0.00000     -152.1236 29.19596 30.69876
## 3  2007.288              4.90678    11083.0068 29.18827 31.15047
##   individualcount
## 1        1.636364
## 2        1.429463
## 3        1.003390
##
## Clustering vector:
##    [1] 2 3 2 2 2 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 3 3 3 3 3 2 2 2 2 2 2
##   [38] 2 2 2 2 3 2 3 3 3 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 3 3 3 3 3 2 2 2 2 3 3 3 2
##   [75] 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [112] 2 2 2 2 3 3 3 3 3 3 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 2 2 2
##  [149] 2 2 2 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 1 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [186] 3 3 3 3 3 3 2 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [223] 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 2 2 2 2 3 1 1 1 3
##  [260] 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 2 2 3 2 3 3 3 3 3 3
##  [297] 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 2 2 2 3 3 3 3 3 3 3 3 2 2 2 2 2
##  [334] 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 3 3 3 3 3 3
##  [371] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
##  [408] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3
##  [445] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [482] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [519] 2 2 2 2 2 2 3 2 2 2 3 3 3 3 3 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [556] 3 3 3 3 2 2 2 2 2 2 2 3 3 3 3 3 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [593] 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
##  [630] 3 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2
##  [667] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 2 2 2 3 3 2 2 2 2 2 2 2
##  [704] 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 3
##  [741] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3
##  [778] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 3 2 2 2 2 2 2 2
##  [815] 2 2 2 2 2 2 3 3 3 3 3 3 3 1 1 1 1 1 1 1 3 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3
##  [852] 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 1 1 1 1 1 1 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [889] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [926] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3
##  [963] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 2 2 2
## [1000] 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 1 2 3 3 2 2
## [1037] 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 2 2 3 3 3 3 3 3 2 2 2 1 3 2 2 2
## [1074] 2 3 2 2 2 1 2 2 2 2 3 2 2 2 2 3 2 2 3 3 3 3 2 3 3 3 3 3 2 2 3 3 2 2 3 3 3
## [1111] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2
## [1148] 2 2 2 2 2 2 2 2 2 3 1 2 2 2 1 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2
## [1185] 2 3 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2
## [1222] 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3
## [1259] 3 3 3 3 3 3 3 3 3 2 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [1296] 3 2 3 2 2 2 2 2 2 3 3 3 2 2 1 2 2 2 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2
## [1333] 2 2 3 3 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2
## [1370] 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 2 2 2 2 2 2 2 2 2 2
## [1407] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 2 1 1 2 2 2 2 2 2 2 2 2
## [1444] 2 2 2
##
## Within cluster sum of squares by cluster:
## [1]  8658862204   126783314 29992679975
##  (between_SS / total_SS =  97.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
pca_new <- PCA(data)
```





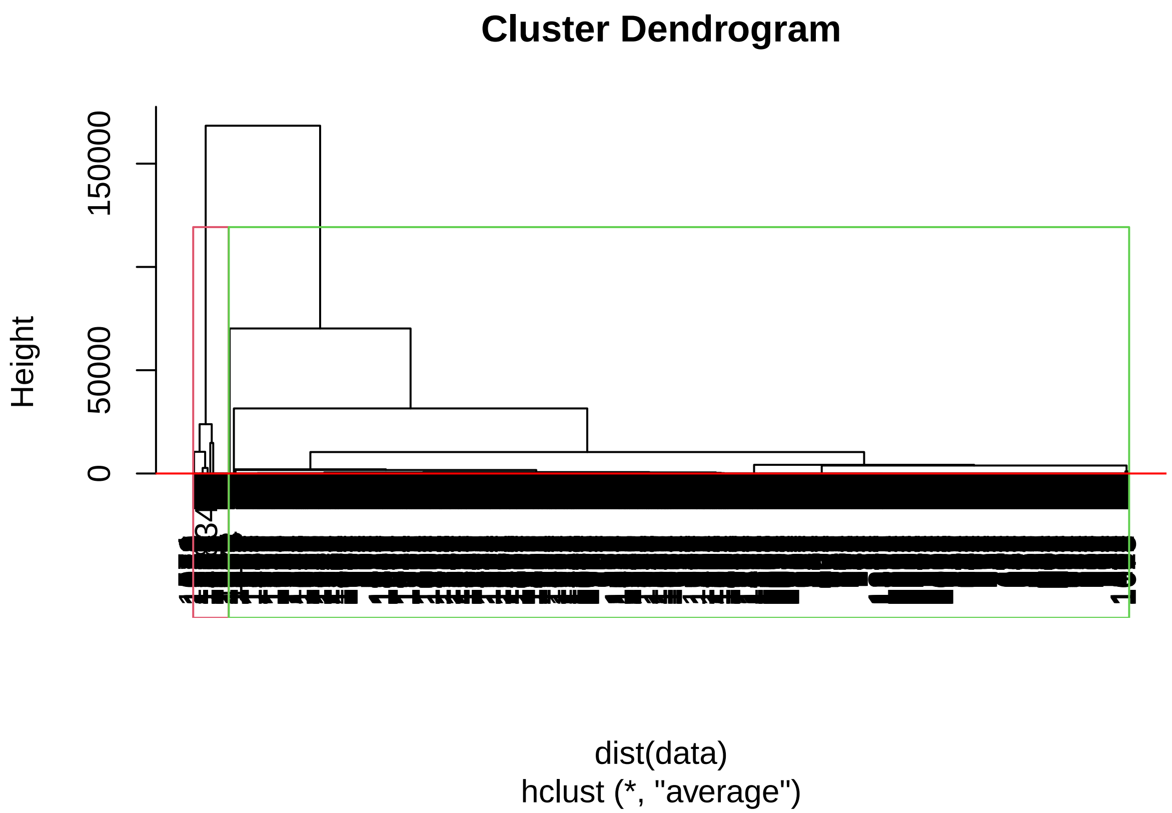
```
fviz_pca_ind(pca_new, label="none", habillage = as.factor(clusters$cluster), addEllipses = T)
```



I choose 3 as the number of clusters for this k-mean clustering because it seems to provide the most parsimonious way of clustering the data and that it doesn't have much overlap between the clusters.

Question 2:

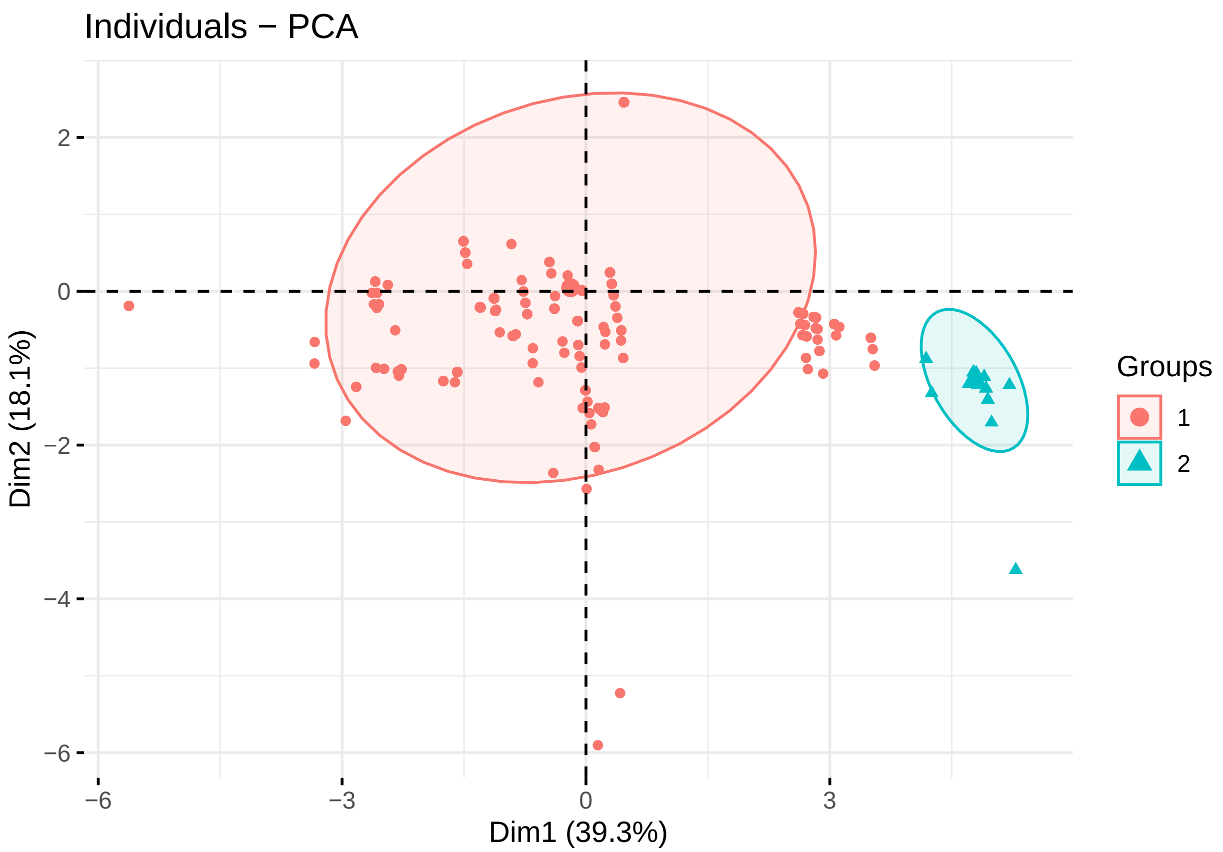
```r
hc <- hclust(dist(data), "ave")
plot(hc)
rect.hclust(hc , k = 2, border = 2:6)
abline(h = 2, col = 'red')
```



```r
clusterCut <- cutree(hc, 2)
print(clusterCut)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [186] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 2 2 1
## [260] 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
## [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [408] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [445] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [482] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [519] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [556] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [593] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [630] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [667] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [704] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [741] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [778] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [815] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [852] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [889] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [926] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [963] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1000] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1
## [1037] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
## [1074] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1111] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1148] 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1185] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1222] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1259] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1296] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1333] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1370] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1407] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1
## [1444] 1 1 1
```

```
fviz_pca_ind(pca_new, label="none", habillage = as.factor(clusterCut), addEllipses = TRUE)
```


Individuals − PCA
2
0
−2
−4
−6
−6
−3
0
3
Dim1 (39.3%)
Dim2 (18.1%)
Groups
● 1
▲ 2