Final Project

Daisha Drayton, Paula Del Rio & TaNia Rowland

2025-10-09

Final Project

Use the college data set Download *college* data set from the course to perform clustering analysis for schools located in Georgia using Principle Component Analysis (PCA) with K-means and hierarchical clustering methods.

The data comes from the U.S Department of Education and includes about 1,270 colleges and universities.

Packages use for this project: *tidyverse*, *stats*, *factoextra*, *patchwork*, *ggplot2*, and *ggrepel*

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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 errors
```

```
library(stats)
library(factoextra)
```

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

```
library(patchwork)
library(ggplot2)
library(ggrepel)
```

1. Importing the Data

```
college <- read_csv("college.csv", col_types = "nccfffffnnnnnnnnn")
```

```
head(college)
```

```
## # A tibble: 6 x 17
##       id name    city  state region highest_degree control gender admission_rate
##    <dbl> <chr>   <chr> <fct> <fct>  <fct>          <fct>   <fct>           <dbl>
## 1 102669 Alaska~ Anch~ AK    West   Graduate       Private CoEd            0.421
## 2 101648 Marion~ Mari~ AL    South  Associate      Public  CoEd            0.614
## 3 100830 Auburn~ Mont~ AL    South  Graduate       Public  CoEd            0.802
## 4 101879 Univer~ Flor~ AL    South  Graduate       Public  CoEd            0.679
## 5 100858 Auburn~ Aubu~ AL    South  Graduate       Public  CoEd            0.835
## 6 100663 Univer~ Birm~ AL    South  Graduate       Public  CoEd            0.857
## # i 8 more variables: sat_avg <dbl>, undergrads <dbl>, tuition <dbl>,
## #   faculty_salary_avg <dbl>, loan_default_rate <dbl>, median_debt <dbl>,
## #   lon <dbl>, lat <dbl>
```

```
names(college)
```

```
##  [1] "id"                 "name"               "city"
##  [4] "state"              "region"             "highest_degree"
##  [7] "control"            "gender"             "admission_rate"
## [10] "sat_avg"            "undergrads"         "tuition"
## [13] "faculty_salary_avg" "loan_default_rate"  "median_debt"
## [16] "lon"                "lat"
```

2. Explore and Prepare the Data

As requested by the instructions we will filter only the schools from *Georgia*

```
college_ga <- college %>%
  filter(state == "GA") %>%
  column_to_rownames(var = "name")
```

For the segmentation we will only use numerical data.

```
num_cols <- c("admission_rate","sat_avg","undergrads","tuition",
              "faculty_salary_avg","loan_default_rate","median_debt", "lon", "lat")
college_ga_num <- college_ga[, num_cols]
```

2.1 Checking for missing values

```
colSums(is.na(college_ga_num))
```

```
##     admission_rate            sat_avg         undergrads            tuition
##                  0                  0                  0                  0
## faculty_salary_avg  loan_default_rate        median_debt                lon
##                  0                  1                  0                  0
##                lat
##                  0
```

```r
sum(is.na(college_ga_num))
```

```
## [1] 1
```

There is one "NA" in *loan_default_rate*, we will do a median transformation to take care of the missing value

```r
college_ga_num$loan_default_rate[is.na(college_ga_num$loan_default_rate)] <-
  median(college_ga_num$loan_default_rate, na.rm = TRUE)
```

2.2 Scaling the Data

```r
summary(college_ga_num)
```

```
##  admission_rate      sat_avg       undergrads       tuition
##  Min.   :0.2181   Min.   : 760   Min.   :  295   Min.   : 3394
##  1st Qu.:0.4912   1st Qu.: 930   1st Qu.: 1399   1st Qu.: 5839
##  Median :0.5944   Median : 990   Median : 3244   Median :11394
##  Mean   :0.6099   Mean   :1016   Mean   : 5881   Mean   :15381
##  3rd Qu.:0.7624   3rd Qu.:1081   3rd Qu.: 6715   3rd Qu.:24780
##  Max.   :0.9664   Max.   :1400   Max.   :26738   Max.   :45008
##  faculty_salary_avg loan_default_rate  median_debt         lon
##  Min.   : 3511      Min.   :0.01700   Min.   :12350   Min.   :-85.36
##  1st Qu.: 5423      1st Qu.:0.06000   1st Qu.:21500   1st Qu.:-84.39
##  Median : 6149      Median :0.08450   Median :24250   Median :-84.16
##  Mean   : 6448      Mean   :0.09228   Mean   :24318   Mean   :-83.77
##  3rd Qu.: 6914      3rd Qu.:0.12700   3rd Qu.:27000   3rd Qu.:-83.36
##  Max.   :12358      Max.   :0.20400   Max.   :38185   Max.   :-81.10
##       lat
##  Min.   :30.83
##  1st Qu.:32.46
##  Median :33.58
##  Mean   :33.29
##  3rd Qu.:34.02
##  Max.   :34.98
```

As we can see the variables have very different mean and the media since wider ranges tent to have a disproportionate impact on the calculation we will scale the values prior to building the model.

```r
X <- scale(college_ga_num)
```

3. PCA

```r
pca_fit <- prcomp(X, scale. = FALSE)
summary(pca_fit)
```

```
## Importance of components:
##                            PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7681 1.3325 1.1327 0.9954 0.81060 0.70212 0.58270
## Proportion of Variance 0.3474 0.1973 0.1426 0.1101 0.07301 0.05477 0.03773
## Cumulative Proportion  0.3474 0.5446 0.6872 0.7973 0.87030 0.92507 0.96280
##                            PC8     PC9
## Standard deviation     0.45139 0.36200
## Proportion of Variance 0.02264 0.01456
## Cumulative Proportion  0.98544 1.00000
```

The variance explained by each principal component is obtained by squaring these:

```
pr.var <- pca_fit$sdev^2
pr.var
```

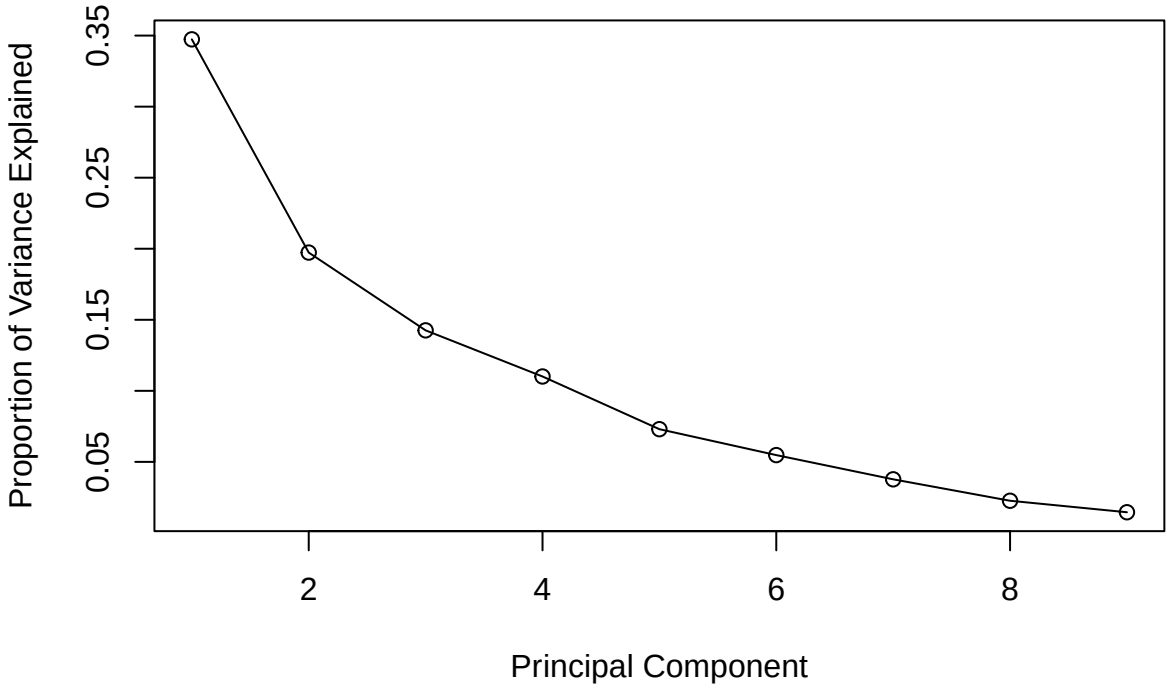
```
## [1] 3.1261968 1.7756489 1.2829960 0.9907812 0.6570681 0.4929742 0.3395421
## [8] 0.2037509 0.1310418
```

To obtain the proportion of the variance explained by each principal component we divide each variance by the total variance:

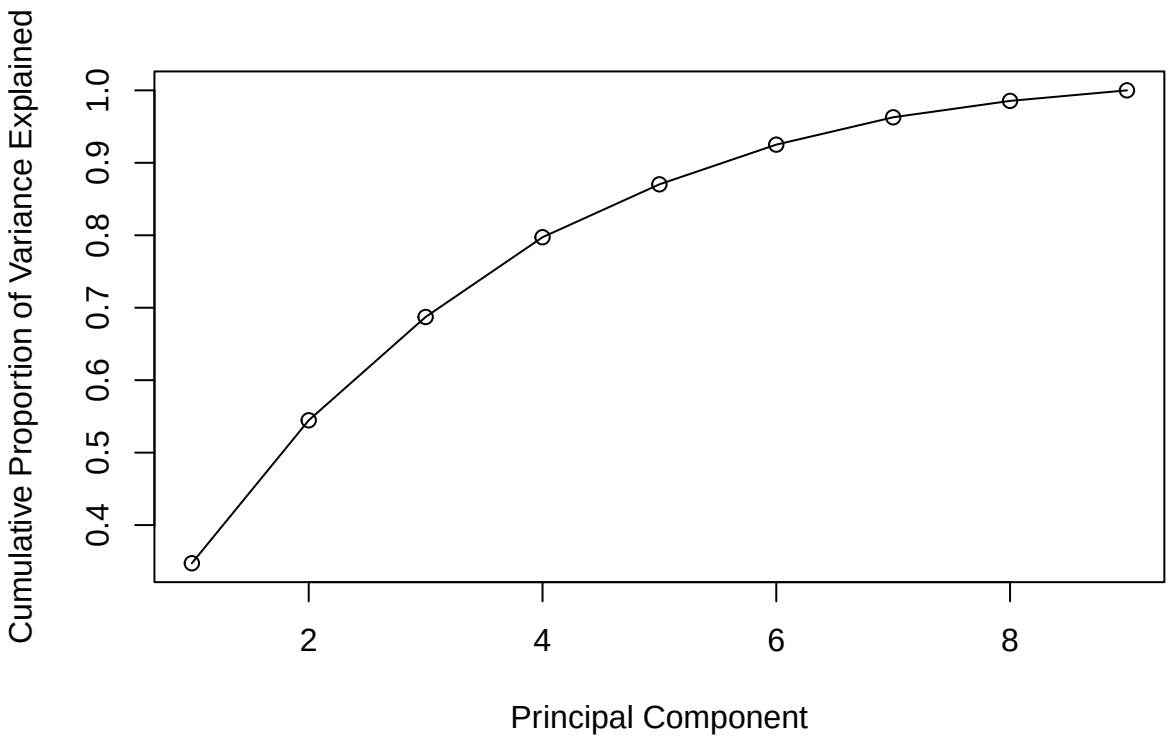
```
pve <- pr.var / sum(pr.var)
pve
```

```
## [1] 0.34735520 0.19729433 0.14255511 0.11008680 0.07300757 0.05477491 0.03772690
## [8] 0.02263899 0.01456020
```

```
plot(pve, type="o", ylab="Proportion of Variance Explained", xlab="Principal Component")
```




```
plot(cumsum(pve), type="o", ylab="Cumulative Proportion of Variance Explained",
     xlab="Principal Component")
```



Loadings:

```
round(pca_fit$rotation[,1:2], 3)
```

```
##                       PC1    PC2
## admission_rate     -0.038 -0.045
## sat_avg             0.520 -0.106
## undergrads          0.264 -0.513
## tuition             0.280  0.489
## faculty_salary_avg  0.462 -0.163
## loan_default_rate  -0.489  0.001
## median_debt        -0.130  0.388
## lon                -0.163 -0.385
## lat                 0.289  0.398
```

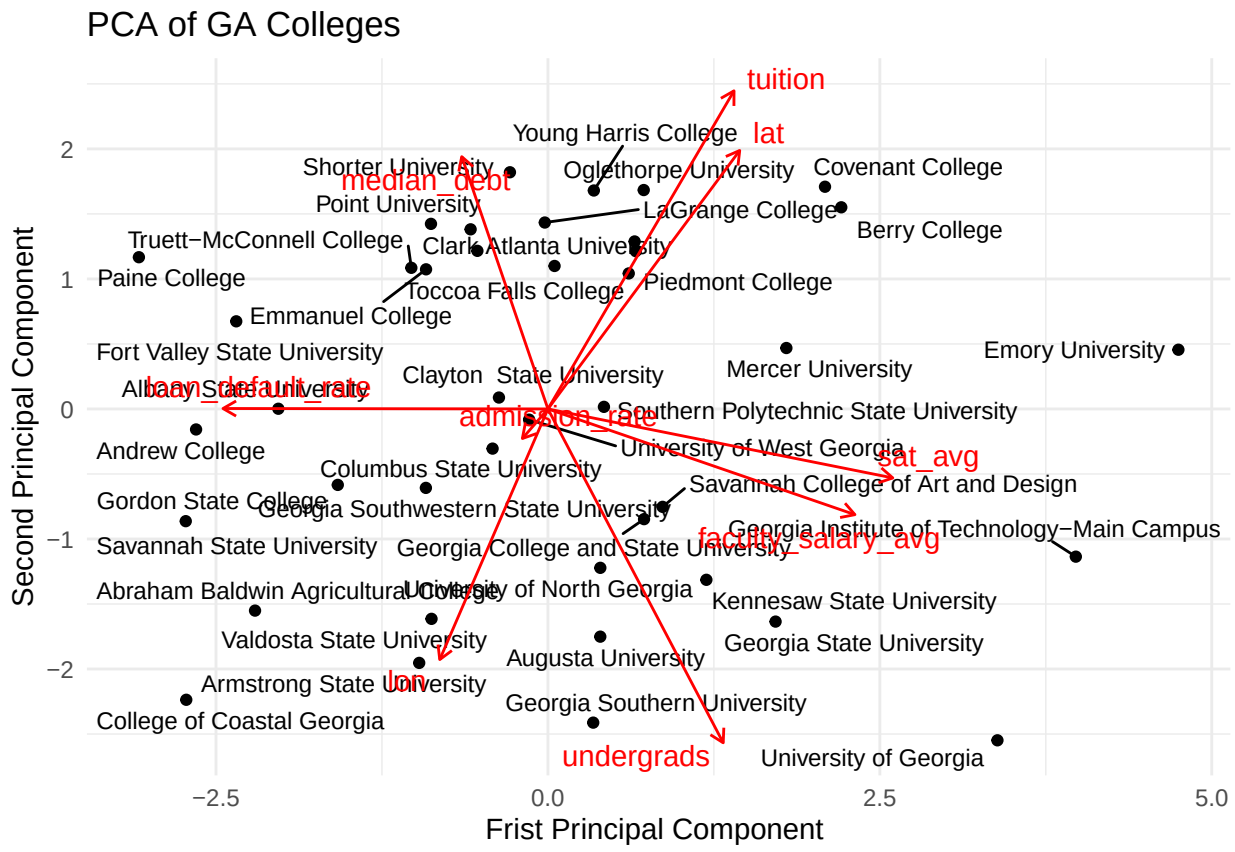
```
pca_df <- data.frame(PC1 = pca_fit$x[, 1],
                     PC2 = pca_fit$x[, 2],
                     id = row.names(X))

loadings <- as.data.frame(pca_fit$rotation)
loadings$Variable <- rownames(loadings)


scores <- as.data.frame(pca_fit$x)
scores$id <- rownames(X)
```

```
loadings_scaled <- loadings
loadings_scaled[, 1:2] <- loadings_scaled[, 1:2] * 5

ggplot(scores, aes(x = PC1, y = PC2)) +
  geom_point() +
  geom_text_repel(aes(label = id), size = 3) +
  geom_segment(data = loadings_scaled,
               aes(x = 0, y = 0, xend = PC1, yend = PC2),
               arrow = arrow(length = unit(0.2, "cm")),
               color = "red") +
  geom_text_repel(data = loadings_scaled,
                  aes(x = PC1, y = PC2, label = Variable),
                  color = "red", size = 4) +
  labs(title = "PCA of GA Colleges",
       x = "Frist Principal Component ",
       y = "Second Principal Component") +
  theme_minimal()
```



Find extremes

```
scores <- as.data.frame(pca_fit$x[,1:2])
colnames(scores) <- c("PC1", "PC2")
scores$name <- college_ga$name
scores[order(scores$PC1)[1:5], c("PC1", "PC2")]
```

```
##                                    PC1        PC2
## Paine College                -3.080609  1.1670289
## Savannah State University    -2.727814 -0.8628158
## College of Coastal Georgia   -2.723972 -2.2365185
## Andrew College               -2.650081 -0.1576927
## Fort Valley State University -2.347487  0.6738999
```

```
scores[order(scores$PC1, decreasing = TRUE)[1:5], c("PC1", "PC2")]
```

```
##                                                PC1        PC2
## Emory University                             4.748790  0.4555156
## Georgia Institute of Technology-Main Campus 3.976255 -1.1358549
## University of Georgia                        3.385661 -2.5481515
## Berry College                                2.209565  1.5508122
## Covenant College                             2.086948  1.7090429
```

3.1 PCA Conclusion

We can see that the first principal component explains 35% of the variance in the data, the next principal component explains 20% of the variance and the third about 14% and the fourth 11%. We will suggest to choose the 4 first components for dimensionality reduction as they explained about 80% which is good balance between simplicity and information retention.

Screen and cumulative proportion plots confirm that the first three PCs capture the major structure, while four PCs give a reasonable low-dimensional representation.

Loadings: PC1 (selectivity/resources axis):

High positive loadings: SAT average (0.52), faculty salary (0.46), tuition (0.28), undergrads (0.26).

Negative loading: loan default rate (–0.49).

Interpretation: separates selective, well-resourced schools (high SATs, tuition, faculty pay) from less selective schools with higher default rates.

PC2 (size/cost axis):

Positive loadings: tuition (0.49), median debt (0.39), latitude (0.40).

Negative loadings: undergrads (–0.51), longitude (–0.39).

Interpretation: distinguishes small/private, higher-tuition schools with greater student debt from larger public universities.

Schools at extremes of PCs: Low PC1 (high default, low SAT/tuition): Paine College, Savannah State, College of Coastal Georgia, Andrew College, Fort Valley State.

High PC1 (elite, high SAT/tuition/faculty pay): Emory, Georgia Tech, University of Georgia, Berry College, Covenant College.

Plots

Proportion of Variance Explained: The y-axis = variance explained by each individual PC.

The x-axis = principal component number.

Shown in plot:

PC1 explains the most (~0.35, or 35%).

PC2 explains the next largest (~0.20, or 20%).

After PC2, the variance explained by each additional PC gets much smaller (PC3 ~15%, PC4 ~11%, etc.).

The "elbow" occurs around PC2. The curve flattens out afterwards. This means most useful information is already captured by PC1 and PC2.

Cumulative PVE Plot: The y-axis = total variance explained when adding up PCs.

The x-axis = number of PCs included.

Shown in plot:

With PC1 alone, about 35% of variance is explained. With PC1 + PC2,we reach ~55% (more than half of the total variance). With PC3, we reach ~70%. By PC4, we are near 80%.

After that, each PC only adds a small extra amount.

Scree plot: Big drop after PC2, later PCs don't add much.

Biplot: PC1 (x-axis) and PC2 (y-axis) are the first two principal components.

They explain the largest chunks of variance in the data (~ 34.7% + 19.7% ~ 54% total).

The plot shows how the schools (black numbers) and the variables (red arrows) relate in this 2-D space.

Red arrows = variables (loadings): Direction: shows which way the variable "pulls" the data. Length: longer arrows = variable contributes more to the PCs. Angle: arrows pointing in similar directions = variables positively correlated. Opposite directions = negative correlation.

Examples from the plot: sat_avg, faculty_salary_avg, and undergrads all point right: they're positively correlated.

loan_default_rate points left: it's negatively correlated with SAT/faculty pay.

tuition and median_debt point upward: they contribute to PC2.

Black numbers = schools (observations): Each number is one Georgia college Their position shows their scores on PC1 and PC2.

A school's location near an arrow means it's strongly influenced by that variable.

Examples from plot: Schools far to the right (West Georgia, Gordon State, University of Georgia): high SAT, faculty salary, large undergrad counts.

Schools far left (Augusta, Clark, and Fort Valley): high loan default rates, lower SAT/faculty salary. Schools higher on PC2 (Armstrong, , GA Tech): higher tuition & student debt, further north (lat). Schools lower on PC2 (Abraham Baldwin, Georgia College): more undergrads, further south.

PC1 (horizontal): "Selectivity/resources" axis. Separates elite, well-funded schools (right side) from less selective schools with higher defaults (left side).

PC2 (vertical): "Size/cost" axis. Separates small, expensive schools with higher debt (top) from larger, cheaper public schools (bottom).

Overall: Red arrows tell you how the original variables map into PC space.

Black numbers show how each school scores on those composite axes.

Where they line up tells you what "drives" each school's position.

4. K-means

We will use the first two components to perform the k-means and the clustering method

```
pc_data <- pca_fit$x[, 1:2] #extracting the first two components
```

4.1 Optimal K

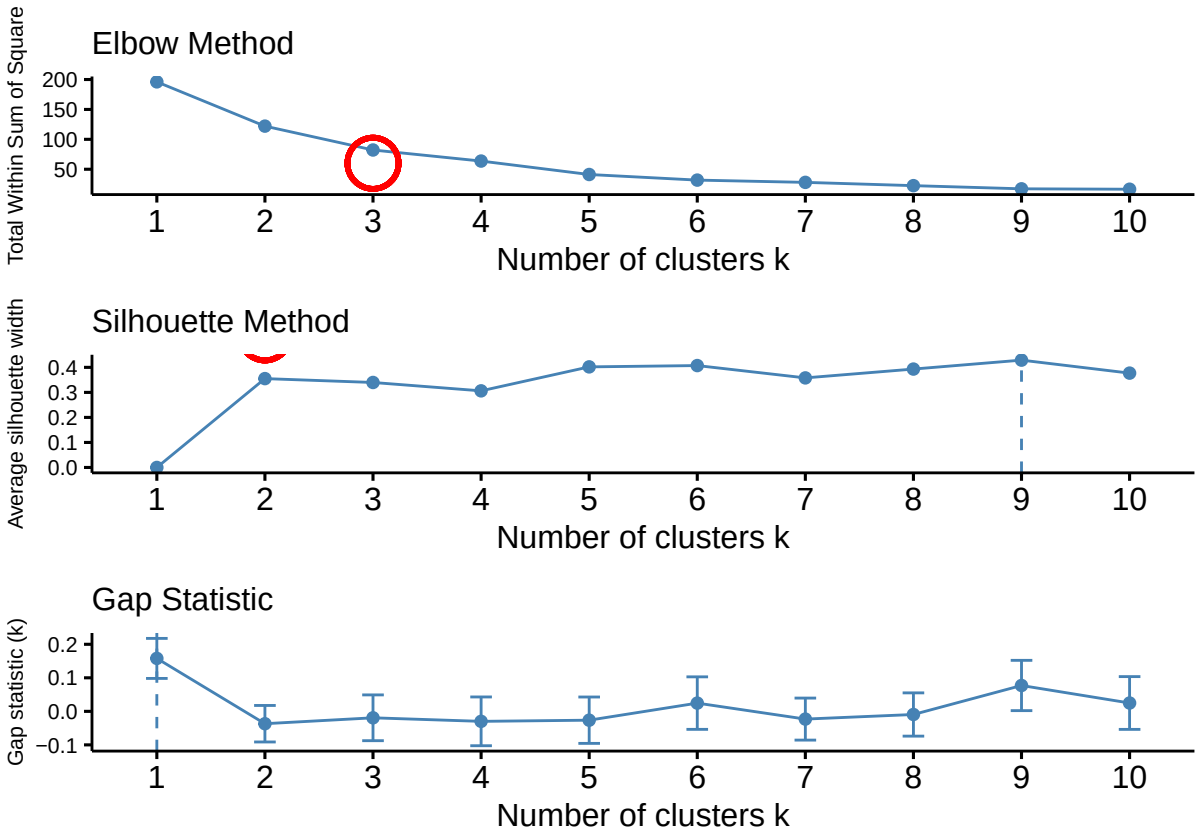
We are now going to try to cluster the data, for this we need to identify the optimal k. We will use the three different approaches learned in this module: Elbow Method, Silhouette Method, and Gap Statistic.

```
#Elbow Method
p1 <- fviz_nbclust(pc_data, kmeans, method = "wss") + geom_point(
  shape = 1,
  x = 3,
  y = 60,
  colour = "red",
  size = 8,
  stroke = 1.5
) + ggtitle("Elbow Method")
```

```
#Silhouette Method

p2 <- fviz_nbclust(pc_data, kmeans, method = "silhouette") + geom_point(
  shape = 1,
  x = 2,
  y = 0.53,
  colour = "red",
  size = 8,
  stroke = 1.5
) + ggtitle("Silhouette Method")
```

```
#Gap Statistic
p3 <- fviz_nbclust(pc_data, kmeans, method = "gap_stat") + geom_point(
  shape = 1,
  x = 1,
  y = 0.57,
  colour = "red",
  size = 8,
  stroke = 1.5
) + ggtitle("Gap Statistic")
```



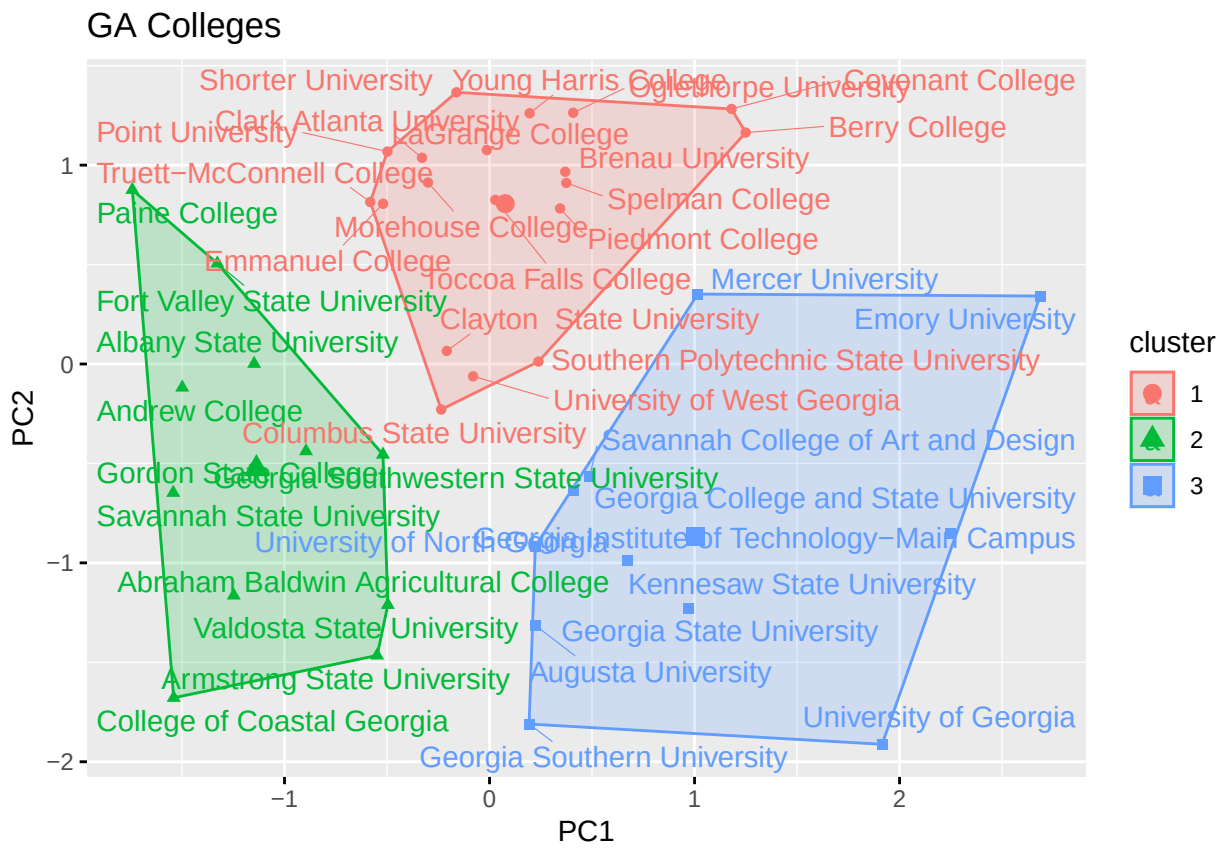
- The Elbow Method is at k = 3, This suggests 3 clusters as a good balance between variance explained and model simplicity.
- Silhouette Method The average silhouette width is highest at k = 2, and slightly lower but stable for k = 3–5. This indicates 2 clusters may be optimal, but 3 could still be reasonable if you prioritize interpretability or if other methods support it.
- Gap Statistic Here, the largest gap is at k = 1, but since clustering with only one group is meaningless, and other values are similar afterward, this method is inconclusive or weakly supports k = 2–3.

We will go with $k = 3$.

4.2 Cluster Visualization

Let's now visualize the results to see which cluster each of our mall customer belong to.

```
set.seed(1234)
k_clust <- kmeans(pc_data, centers = 3, nstart = 25)
fviz_cluster(
  k_clust,
  data = pc_data,
  main = "GA Colleges",
  repel = TRUE
)
```



4.3 K-means Conclusion:

The colleges are grouped into six distinct clusters based on different characteristics: *SAT average*, *number of undergrad students*, *annual tuition*, *average monthly faculty salary*, *percentage of students who default their loan*, *median amount of debt for graduating student*

Cluster 1 (Red): Smaller, private liberal arts colleges and HBCU likely similar in size and mission.

Cluster 2 (Green): Regional public colleges and smaller state universities—more spread out, but still distinct.

Cluster 3 (Blue): Large research universities (R1, R2) and tech-focused institutions—clearly differentiated by scale and academic profile.

Even though your earlier diagnostics gave mixed signals (Silhouette and Gap Statistic), the PCA plot shows clear spatial separation among three groups. That's a strong visual confirmation that k = 3 is meaningful and interpretable.

5. Hierarchical Clustering

The *hclust()* function is use for hierarchical clustering. For the following example we will continue to use the data created. Let's plot the hierarchical clustering dendrogram using complete, single, and average linkage clustering

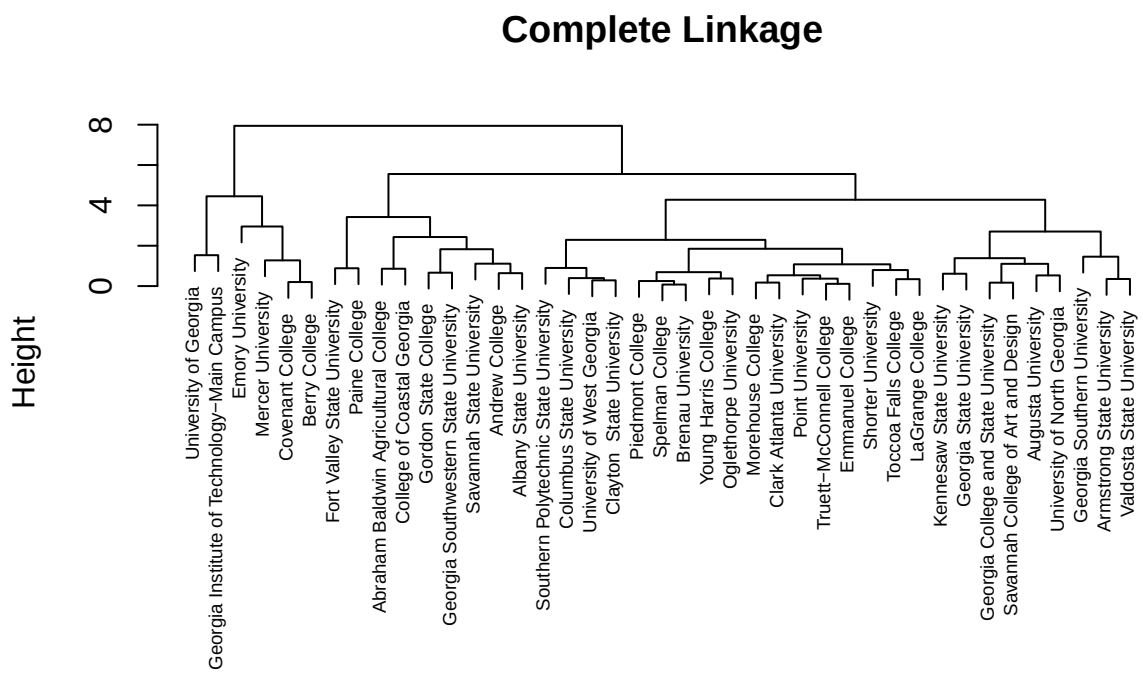
```
hc.complete <- hclust(dist(pc_data), method = "complete")
```

```
hc.average <- hclust(dist(pc_data), method = "average")
```

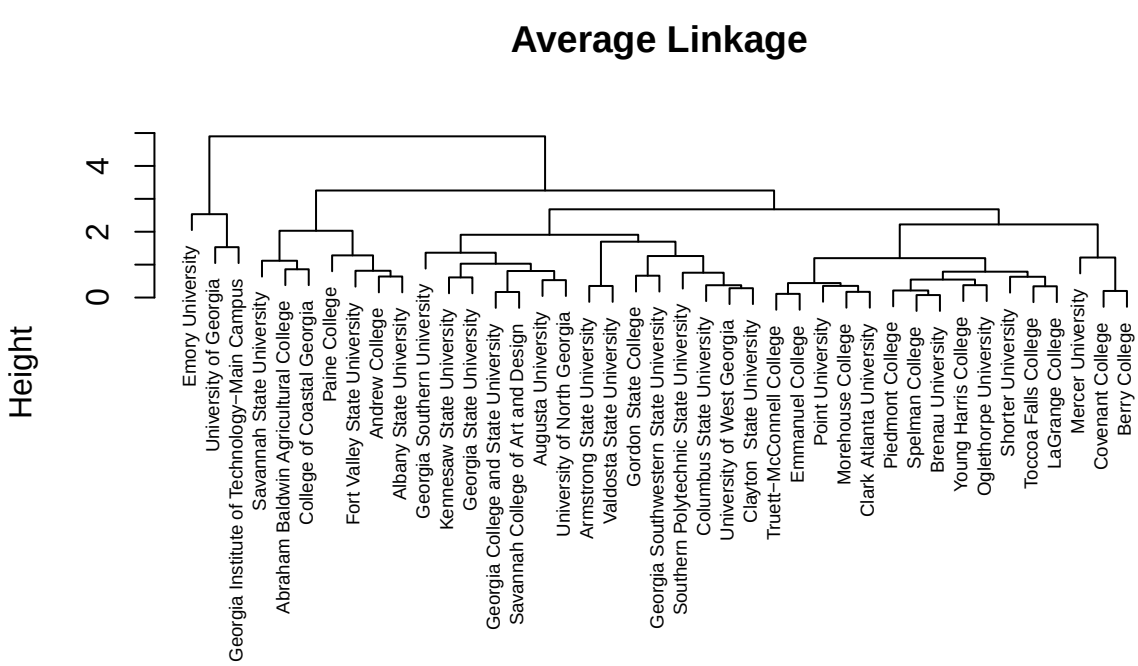
```
hc.single <- hclust(dist(pc_data), method = "single")
```

We can now plot the dendrograms. The numbers at the bottom of the plot identify each observation.

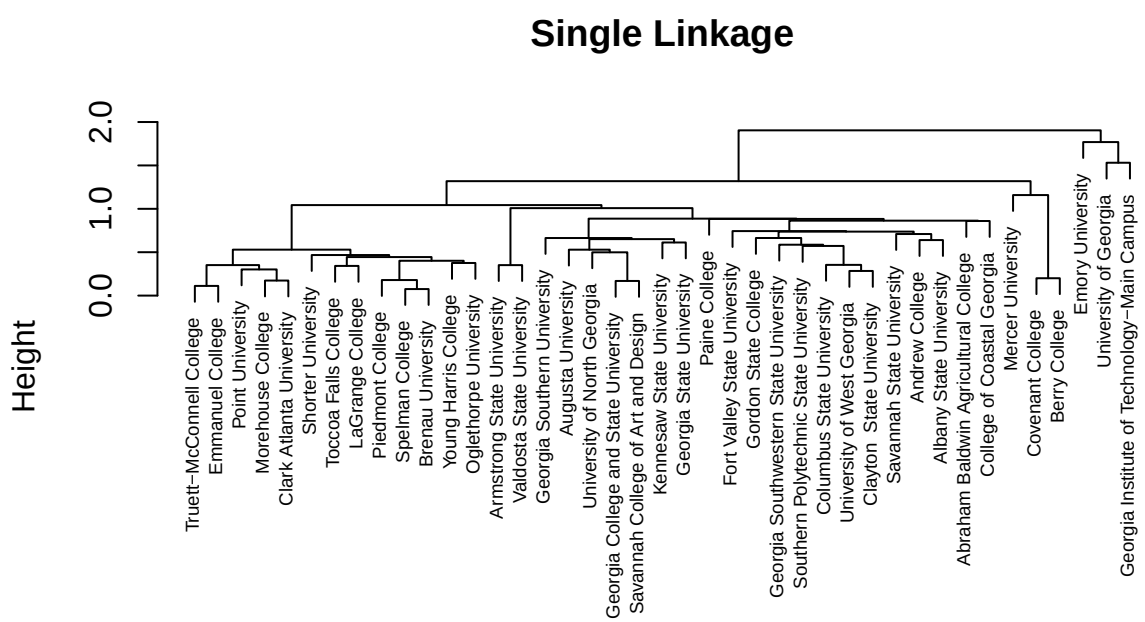
```
plot(hc.complete, main = "Complete Linkage",
     xlab ="", sub ="", cex = .6)
```




```
plot(hc.average, main = "Average Linkage",
     xlab ="", sub ="", cex = .6)
```




```
plot(hc.single, main = "Single Linkage",
     xlab ="", sub ="", cex = .6)
```



To determine the cluster labels for each observation associated with a given cut of the dendrogram, we use *cutree()*. We use 3 as the argument for the function to be able to compare with the K-means using 3 clusters.

```
cutree(hc.complete, 3)
```

```
##                          Augusta University
##                                           1
##                             Andrew College
##                                           2
##                  University of North Georgia
##                                           1
##         Abraham Baldwin Agricultural College
##                                           2
##                   College of Coastal Georgia
##                                           2
##                         Gordon State College
##                                           2
##                        University of Georgia
##                                           3
##                           Emory University
##                                           3
##                   Armstrong State University
##                                           1
##                           Covenant College
##                                           3
##       Georgia Southwestern State University
```

```
##                                         2
##        Georgia College and State University
##                                         1
##                 Georgia Southern University
##                                         1
##                   Kennesaw State University
##                                         1
##                    Georgia State University
##                                         1
##                    Truett-McConnell College
##                                         1
##                        Toccoa Falls College
##                                         1
##                           Piedmont College
##                                         1
##                              Berry College
##                                         3
##                  University of West Georgia
##                                         1
## Georgia Institute of Technology-Main Campus
##                                         3
##                            Mercer University
##                                         3
##                   Columbus State University
##                                         1
##       Southern Polytechnic State University
##                                         1
##                   Valdosta State University
##                                         1
##                        Young Harris College
##                                         1
##                   Clayton  State University
##                                         1
##                             Spelman College
##                                         1
##                         Oglethorpe University
##                                         1
##                            Emmanuel College
##                                         1
##                            LaGrange College
##                                         1
##                           Morehouse College
##                                         1
##          Savannah College of Art and Design
##                                         1
##                            Brenau University
##                                         1
##                           Shorter University
##                                         1
##                             Point University
##                                         1
##                    Clark Atlanta University
##                                         1
##                     Albany State University
```

```
##                                       2
##                     Savannah State University
##                                       2
##                  Fort Valley State University
##                                       2
##                                Paine College
##                                       2
```

```
cutree(hc.average, 3)
```

```
##                            Augusta University
##                                             1
##                               Andrew College
##                                             2
##                   University of North Georgia
##                                             1
##          Abraham Baldwin Agricultural College
##                                             2
##                    College of Coastal Georgia
##                                             2
##                          Gordon State College
##                                             1
##                         University of Georgia
##                                             3
##                              Emory University
##                                             3
##                    Armstrong State University
##                                             1
##                              Covenant College
##                                             1
##         Georgia Southwestern State University
##                                             1
##          Georgia College and State University
##                                             1
##                   Georgia Southern University
##                                             1
##                     Kennesaw State University
##                                             1
##                      Georgia State University
##                                             1
##                      Truett-McConnell College
##                                             1
##                          Toccoa Falls College
##                                             1
##                              Piedmont College
##                                             1
##                                Berry College
##                                             1
##                    University of West Georgia
##                                             1
## Georgia Institute of Technology-Main Campus
##                                             3
##                              Mercer University
##                                             1
```

```
##                   Columbus State University
##                                           1
##       Southern Polytechnic State University
##                                           1
##                   Valdosta State University
##                                           1
##                        Young Harris College
##                                           1
##                   Clayton  State University
##                                           1
##                             Spelman College
##                                           1
##                         Oglethorpe University
##                                           1
##                              Emmanuel College
##                                           1
##                              LaGrange College
##                                           1
##                            Morehouse College
##                                           1
##          Savannah College of Art and Design
##                                           1
##                             Brenau University
##                                           1
##                            Shorter University
##                                           1
##                              Point University
##                                           1
##                     Clark Atlanta University
##                                           1
##                      Albany State University
##                                           2
##                   Savannah State University
##                                           2
##                Fort Valley State University
##                                           2
##                                Paine College
##                                           2
```

```
cutree(hc.single, 3)
```

```
##                             Augusta University
##                                              1
##                                Andrew College
##                                              1
##                  University of North Georgia
##                                              1
##         Abraham Baldwin Agricultural College
##                                              1
##                    College of Coastal Georgia
##                                              1
##                          Gordon State College
##                                              1
##                         University of Georgia
```

```
##                                       2
##                            Emory University
##                                       3
##                   Armstrong State University
##                                       1
##                              Covenant College
##                                       1
##       Georgia Southwestern State University
##                                       1
##        Georgia College and State University
##                                       1
##                 Georgia Southern University
##                                       1
##                    Kennesaw State University
##                                       1
##                     Georgia State University
##                                       1
##                     Truett-McConnell College
##                                       1
##                         Toccoa Falls College
##                                       1
##                             Piedmont College
##                                       1
##                               Berry College
##                                       1
##                   University of West Georgia
##                                       1
## Georgia Institute of Technology-Main Campus
##                                       2
##                             Mercer University
##                                       1
##                    Columbus State University
##                                       1
##       Southern Polytechnic State University
##                                       1
##                    Valdosta State University
##                                       1
##                          Young Harris College
##                                       1
##                   Clayton  State University
##                                       1
##                               Spelman College
##                                       1
##                          Oglethorpe University
##                                       1
##                             Emmanuel College
##                                       1
##                             LaGrange College
##                                       1
##                            Morehouse College
##                                       1
##          Savannah College of Art and Design
##                                       1
##                             Brenau University
```

```
##                                       1
##                          Shorter University
##                                       1
##                            Point University
##                                       1
##                     Clark Atlanta University
##                                       1
##                      Albany State University
##                                       1
##                    Savannah State University
##                                       1
##                 Fort Valley State University
##                                       1
##                               Paine College
##                                       1
```

5.1 Hierarchical Clustering Conclusion

Euclidean distance with complete linkage was chosen because it provides the most distinct and balanced separation of colleges into 3 groups. Single linkage over combined the colleges into one large cluster (chaining effect), while average linkage produced less stable boundaries. Complete linkage strikes a balance by forming compact, well-separated clusters.

The complete linkage dendrogram reveals several clear patterns in how the colleges group together. Private universities such as Mercer, Emory, Covenant, and Berry cluster tightly, reflecting strong similarities among liberal arts and private schools. Regional public universities like Georgia Southern, University of North Georgia, Armstrong State, and Valdosta State form their own cohesive group, highlighting comparable college profiles. Large schools like University of Georgia and Georgia Tech, stand apart at higher heights, emphasizing their distinctiveness in size and research focus. You can also see HBCUs such as Spelman, Morehouse, and Clark Atlanta clustering with other smaller liberal arts colleges, highlighting shared characteristics. Overall, the dendrogram illustrates a clear separation between major research universities, regional public institutions, and private or specialized colleges.

```r
hc.out <- hclust(dist(pc_data))
hc.clusters <- cutree(hc.out, 3)
table(hc.clusters)
```

```
## hc.clusters
##  1  2  3
## 26  9  6
```

The cluster distribution, shown by table(hc.clusters), indicates that Cluster 1 contains 26 institutions, Cluster 2 contains 9 institutions, and Cluster 3 contains 6 institutions.

```r
hc.out
```

```
##
## Call:
## hclust(d = dist(pc_data))
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 41
```

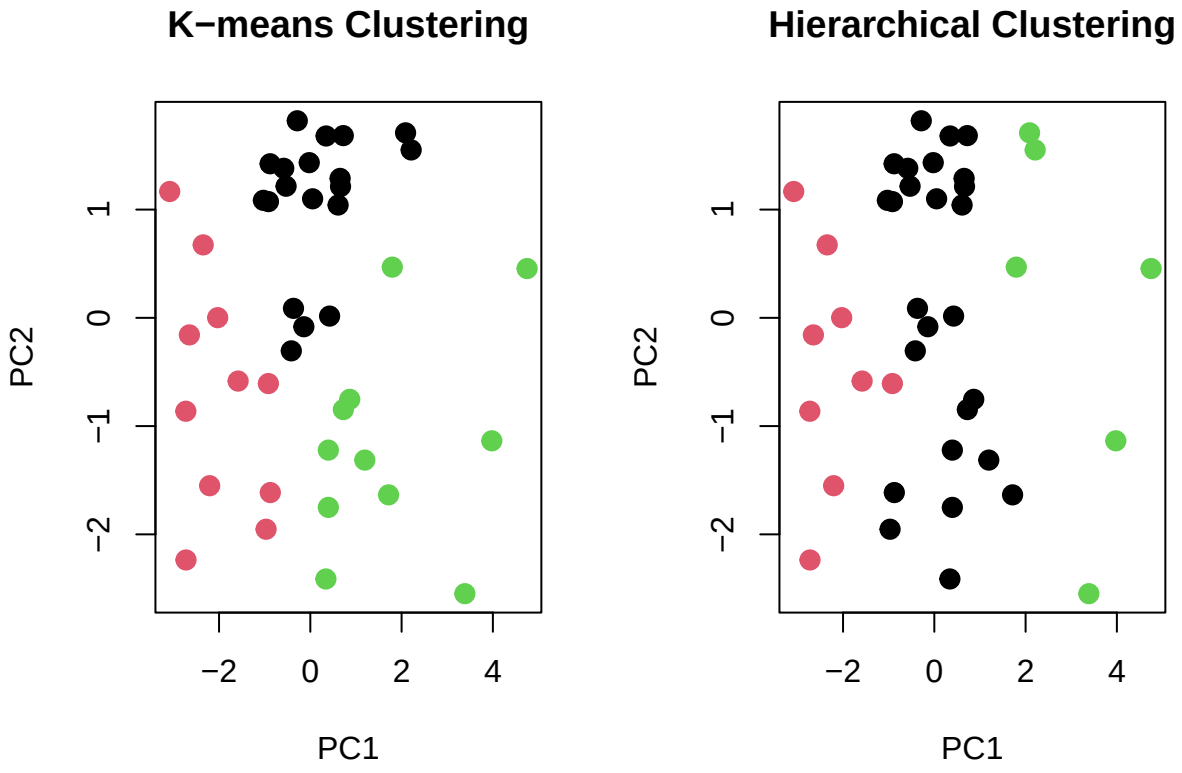
The hc.out results show that hierarchical clustering of all 41 institutions was done using Euclidean distance and complete linkage, forming the basis for the dendrogram and cluster assignments.

6. K-means and Hierarchical Clustering Comparison

```
set.seed(2)
km.out <- k_clust
km.clusters <- km.out$cluster
table(km.clusters, hc.clusters)
```

```
##            hc.clusters
## km.clusters  1  2  3
##           1 17  0  2
##           2  2  9  0
##           3  7  0  4
```

```
par(mfrow = c(1, 2))
plot(pc_data, col = km.clusters, main = "K-means Clustering", pch = 20, cex = 2)
plot(pc_data, col = hc.clusters, main = "Hierarchical Clustering", pch = 20, cex = 2)
```



The comparison between K-means clustering and hierarchical clustering (with dendrogram cut for 3 clusters) reveals an overall agreement rate of 81% (30 out of 37 observations). While both methods identify similar groupings, notable differences exist in cluster assignments. Hierarchical cluster 2 shows near-perfect correspondence with K-means cluster 2 (9/9 observations), indicating both methods identified a highly distinct,

well-separated group. Hierarchical cluster 1 demonstrates moderate agreement, with 17 of 26 observations assigned to K-means cluster 1, though 7 observations were classified into K-means cluster 3. Hierarchical cluster 3 shows the weakest correspondence, with observations split between K-means clusters 1 and 3.

These differences arise from fundamental algorithmic distinctions. K-means uses centroid-based assignment with Euclidean distance, creating spherical clusters, while hierarchical clustering builds a tree structure that can capture more complex cluster shapes depending on the linkage method. The dendrogram cut creates clusters based on hierarchical similarity rather than centroid proximity, leading to different boundary decisions for borderline observations. The results suggest that while both methods capture the overall structure of the data, they disagree on approximately 19% of observations, particularly those near cluster boundaries. This highlights the importance of using multiple clustering approaches to validate results and understand data structure.