Student Clustering using DBSCN

Code

Data preprocessing

Hide

```
library(readxl)
library(dplyr)

# stack 3 files into one
df1 <- read_excel('Fall 2021 English 106.xlsx')
df2 <- read_excel('Fall 2021 English 107.xlsx')
df3 <- read_excel('Fall 2021 English 108.xlsx')
df <- rbind(df1, df2, df3)

# only keep the colns we want
df <- df[, c('ID','Acad Level','College','Birth Country Code',
             'TOEFL COMPI','IELTS Overall Band Score',
             'DUOLINGO English','CEPT Total')]

# simplify column name
df <- df %>% rename(Acad = 'Acad Level',
                    BCC = 'Birth Country Code',
                    TOEFL = 'TOEFL COMPI',
                    IELTS = 'IELTS Overall Band Score',
                    DUOLINGO = 'DUOLINGO English',
                    CEPT = 'CEPT Total')

# normalize scores
df$TOEFL <- df$TOEFL / 120
df$IELTS <- df$IELTS / 9
df$DUOLINGO <- df$DUOLINGO / 160
df$CEPT <- df$CEPT / 150

# get average score as a new col
df$score <- rowMeans(df[, c("TOEFL", "IELTS", "DUOLINGO", "CEPT")], na.rm = TRUE)
df <- df[, c('ID','Acad','College','BCC', 'score')]
df <- na.omit(df)
df
```

ID	Acad	College	B…	score
<dbl>	<chr>	<chr>	<chr>	<dbl>
23378554	Freshman	College of Science	CHN	0.6833333
23449622	Junior	College of Soc & Behav Sci	ARE	0.5555556
23507647	Freshman	College of Science	SAU	0.6666667
23562425	Freshman	College of Science	CHN	0.6562500

ID	Acad	College	B...	score
<dbl>	<chr>	<chr>	<chr>	<dbl>
23562571	Freshman	College of Science	CHN	0.5850000
23593854	Freshman	College of Science	MEX	0.6250000
23594310	Freshman	Eller College of Management	IND	0.7750000
23605627	Freshman	College of Science	SAU	0.7222222
23608992	Freshman	Eller College of Management	CHN	0.5833333
23609468	Freshman	College of Science	SAU	0.7222222

1-10 of 364 rows Previous **1** 2 3 4 5 6 ... 37 Next

Hide

```
# convert string to number for computation
df_feature <- df %>%
  mutate(across(c(Acad, College, BCC), as.factor)) %>%
  mutate(across(c(Acad, College, BCC), as.numeric))
df_feature <- df_feature[, c("Acad", "College", "BCC", "score")]
df_feature
```

Acad	College	BCC	score
<dbl>	<dbl>	<dbl>	<dbl>
1	9	8	0.6833333
2	10	2	0.5555556
1	9	42	0.6666667
1	9	8	0.6562500
1	9	8	0.5850000
1	9	28	0.6250000
1	11	21	0.7750000
1	9	42	0.7222222
1	11	8	0.5833333
1	9	42	0.7222222

1-10 of 364 rows Previous **1** 2 3 4 5 6 ... 37 Next

K means ++

Hide

```
set.seed(42) # fix random
kmeans_result <- kmeans(df_feature, centers = 3, nstart = 25, algorithm = "Lloyd", iter.
max = 20)
df$k_cluster <- kmeans_result$cluster
df
```

ID	Acad	College	B...	score	k_cluster
<dbl>	<chr>	<chr>	<chr>	<dbl>	<int>
23378554	Freshman	College of Science	CHN	0.6833333	1
23449622	Junior	College of Soc & Behav Sci	ARE	0.5555556	1
23507647	Freshman	College of Science	SAU	0.6666667	2
23562425	Freshman	College of Science	CHN	0.6562500	1
23562571	Freshman	College of Science	CHN	0.5850000	1
23593854	Freshman	College of Science	MEX	0.6250000	3
23594310	Freshman	Eller College of Management	IND	0.7750000	3
23605627	Freshman	College of Science	SAU	0.7222222	2
23608992	Freshman	Eller College of Management	CHN	0.5833333	1
23609468	Freshman	College of Science	SAU	0.7222222	2

1-10 of 364 rows Previous **1** 2 3 4 5 6 ... 37 Next

DBSCAN

Hide

```
# install.packages("dbscan")
library(dbscan)

result <- dbscan(df_feature, eps = 3, minPts = 8)
df$d_cluster <- result$cluster
df
```

ID	Acad	College	B..	score	k_cluster	d_cluster
<dbl>	<chr>	<chr>	<chr>	<dbl>	<int>	<int>
23378554	Freshman	College of Science	CHN	0.6833333	1	1
23449622	Junior	College of Soc & Behav Sci	ARE	0.5555556	1	0
23507647	Freshman	College of Science	SAU	0.6666667	2	2
23562425	Freshman	College of Science	CHN	0.6562500	1	1
23562571	Freshman	College of Science	CHN	0.5850000	1	1

ID	Acad	College	B..	score	k_cluster	d_cluster
<dbl>	<chr>	<chr>	<chr>	<dbl>	<int>	<int>
23593854	Freshman	College of Science	MEX	0.6250000	3	3
23594310	Freshman	Eller College of Management	IND	0.7750000	3	3
23605627	Freshman	College of Science	SAU	0.7222222	2	2
23608992	Freshman	Eller College of Management	CHN	0.5833333	1	1
23609468	Freshman	College of Science	SAU	0.7222222	2	2

1-10 of 364 rows Previous **1** 2 3 4 5 6 … 37 Next

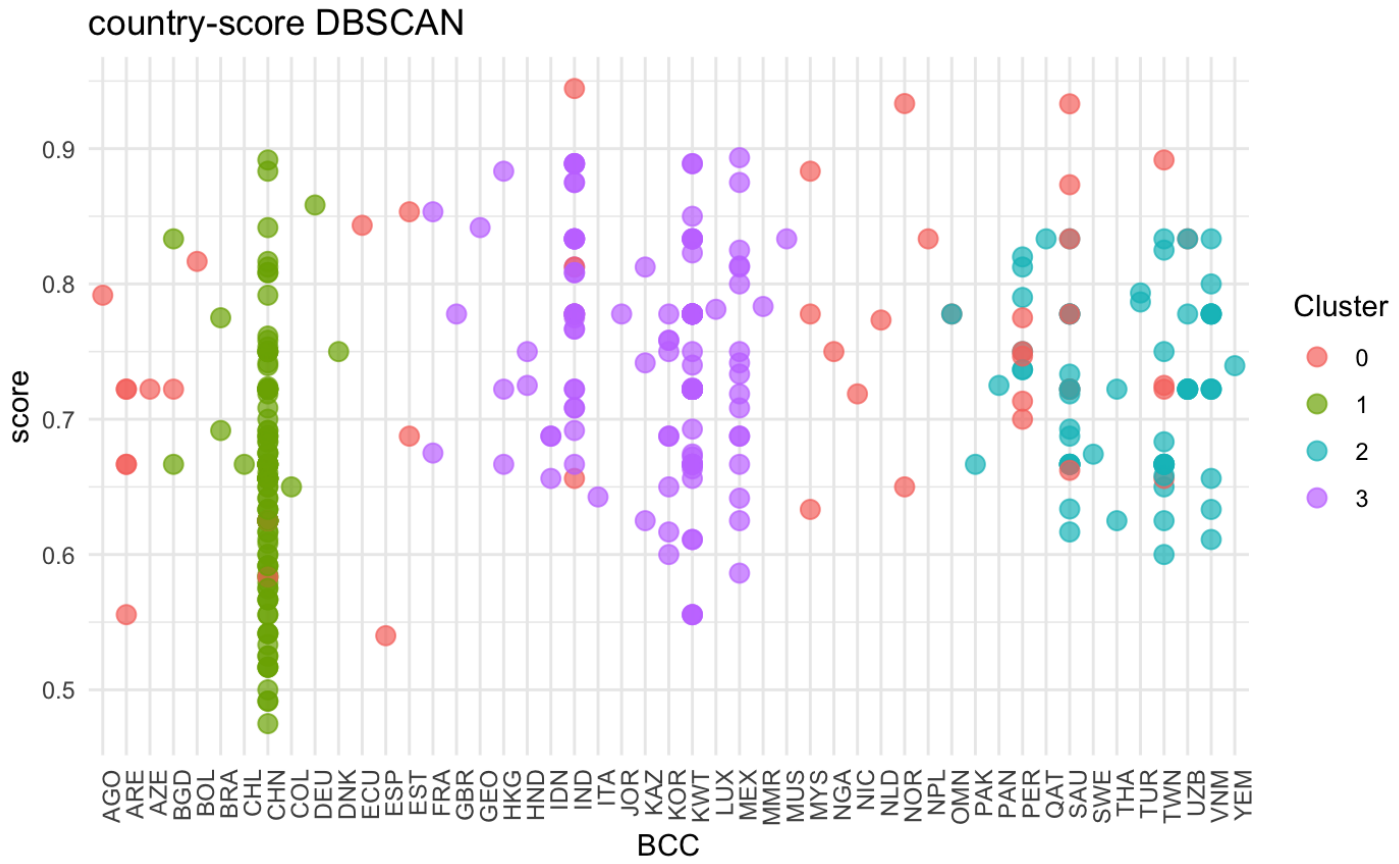
Hide

```
NA
```

view country-score

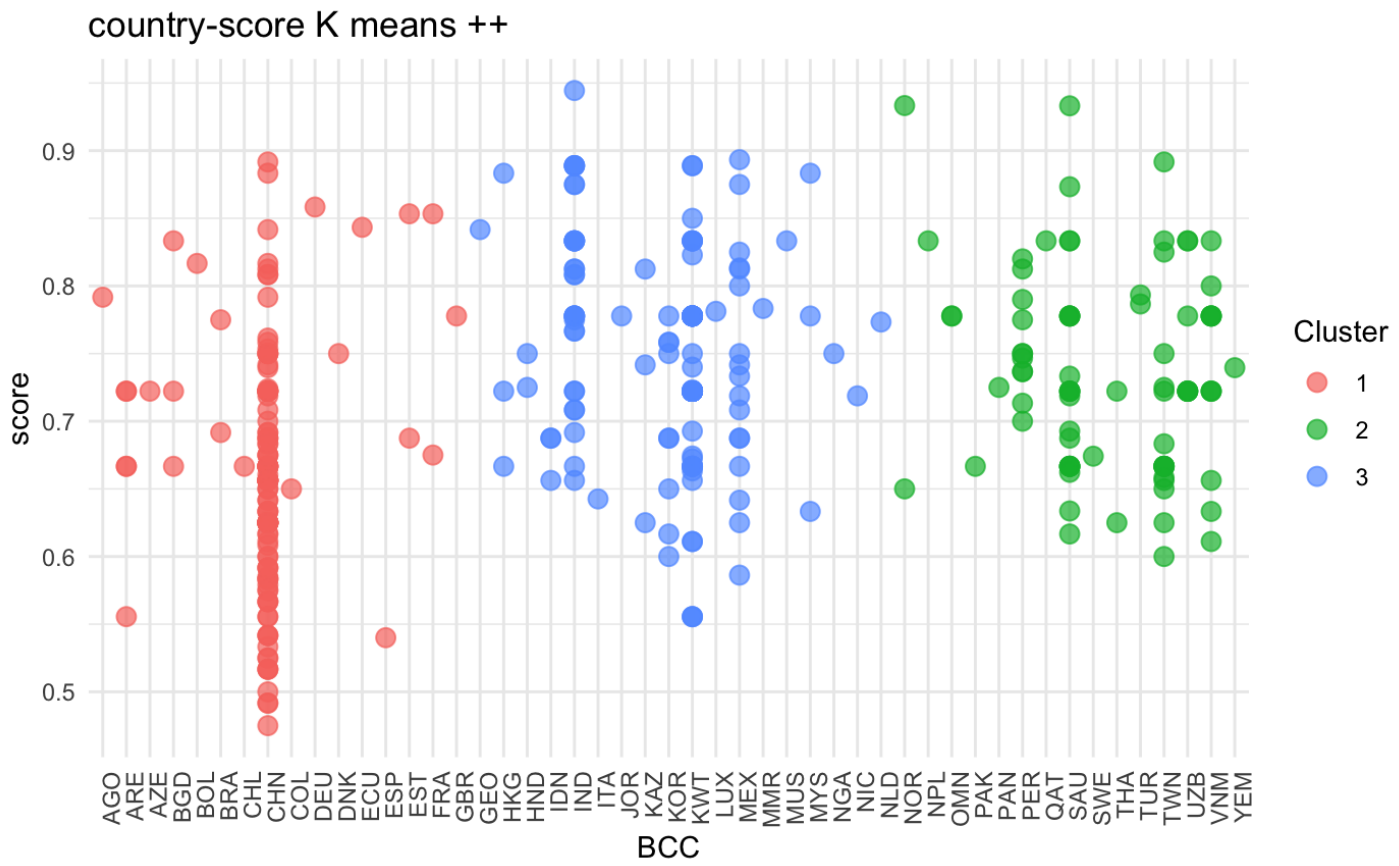
Hide

```
library(ggplot2)
ggplot(df, aes(x = BCC, y = score, color = as.factor(d_cluster))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "country-score DBSCAN",
       color = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Hide

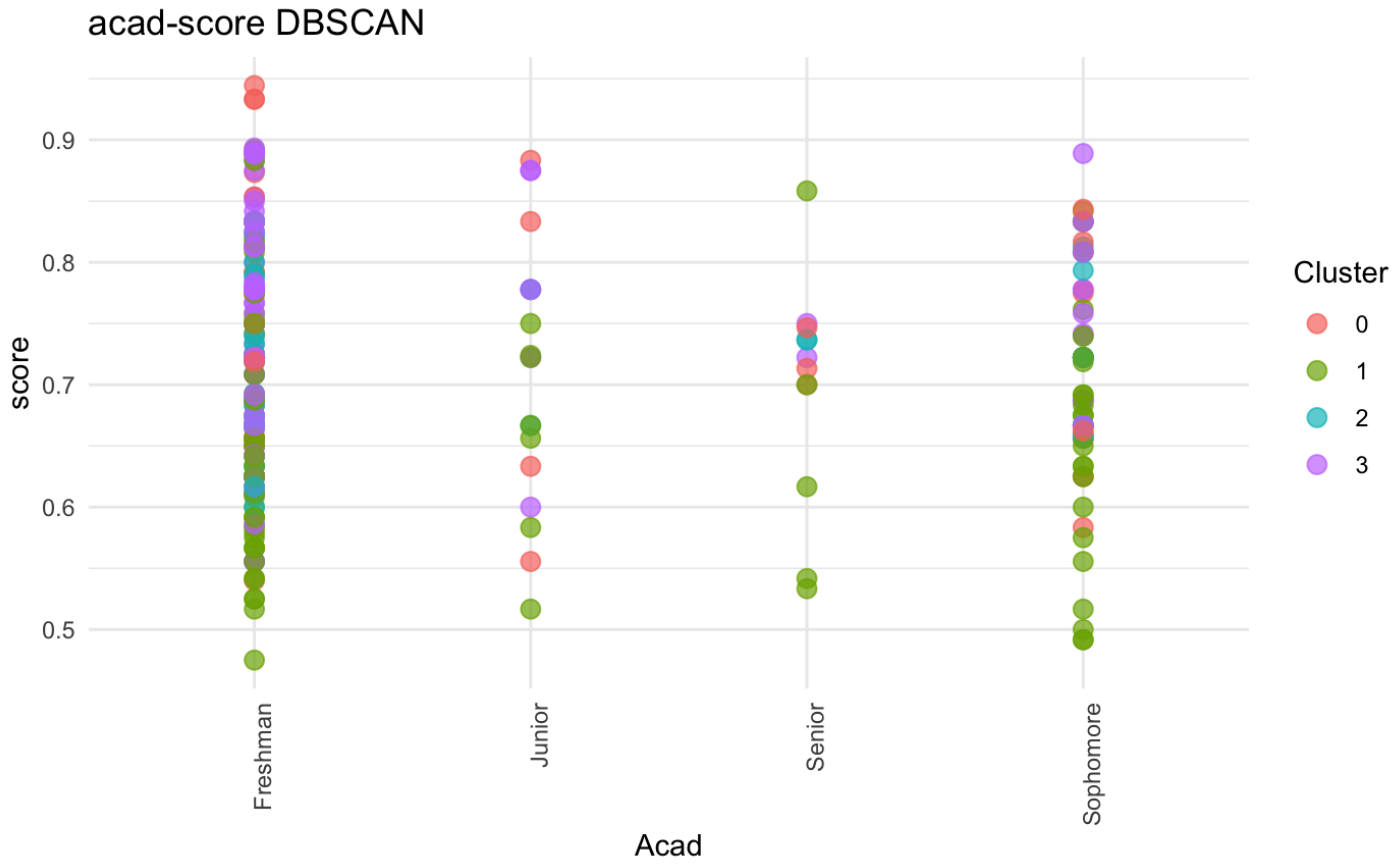
```
library(ggplot2)
ggplot(df, aes(x = BCC, y = score, color = as.factor(k_cluster))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "country-score K means ++",
       color = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



view acad-score

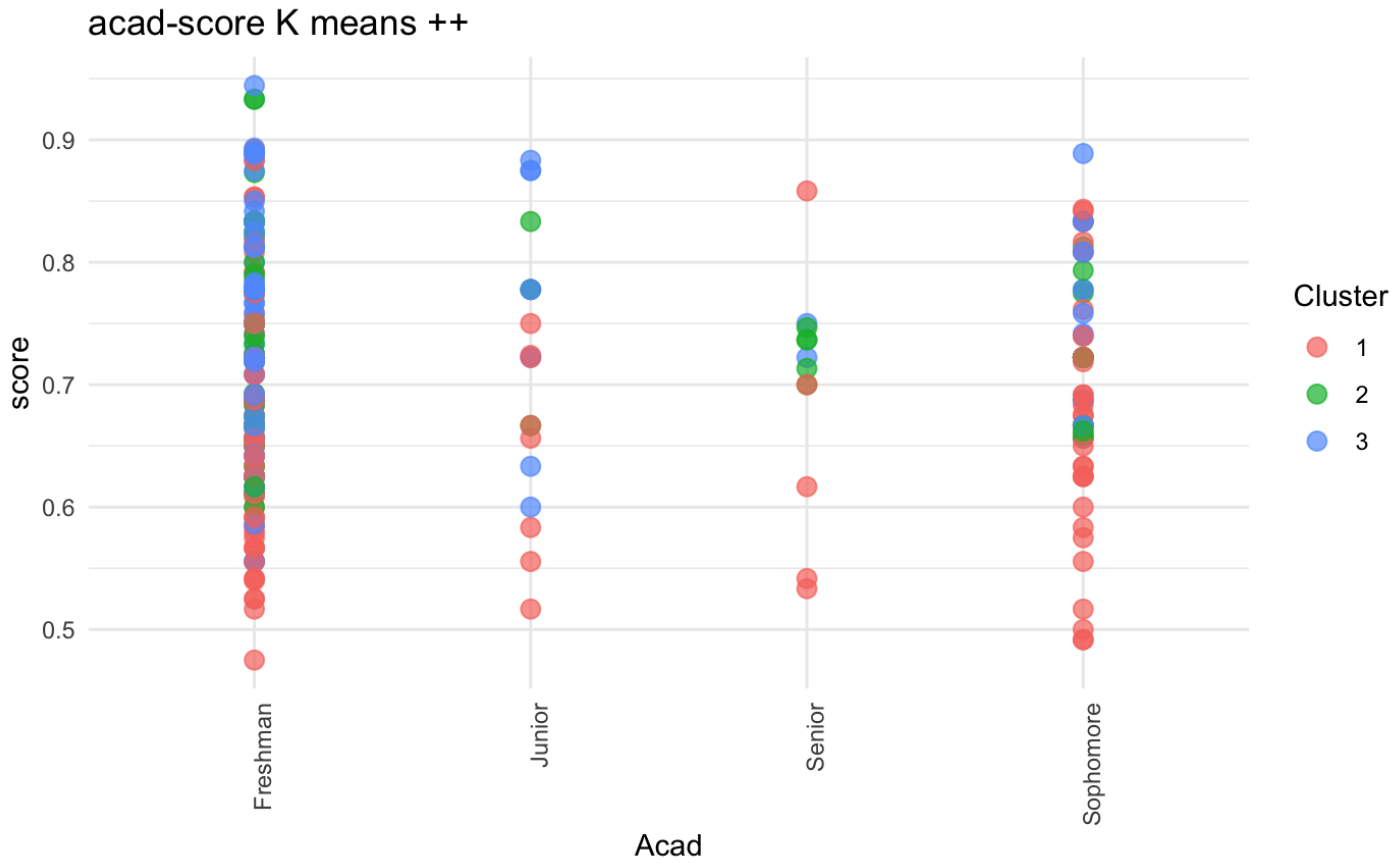
Hide

```
ggplot(df, aes(x = Acad, y = score, color = as.factor(d_cluster))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "acad-score DBSCAN",
       color = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Hide

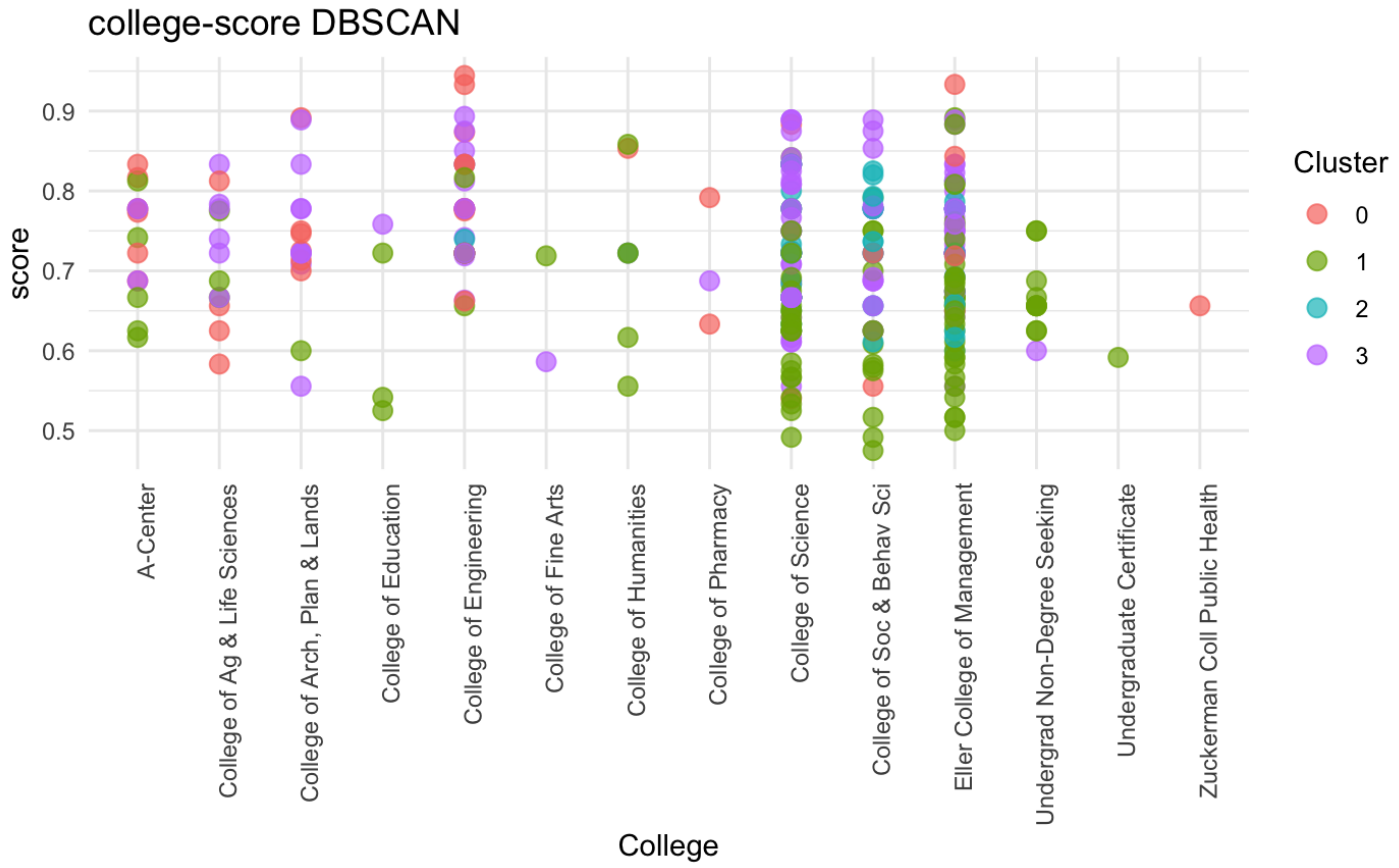
```
library(ggplot2)
ggplot(df, aes(x = Acad, y = score, color = as.factor(k_cluster))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "acad-score K means ++",
       color = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



view college-score

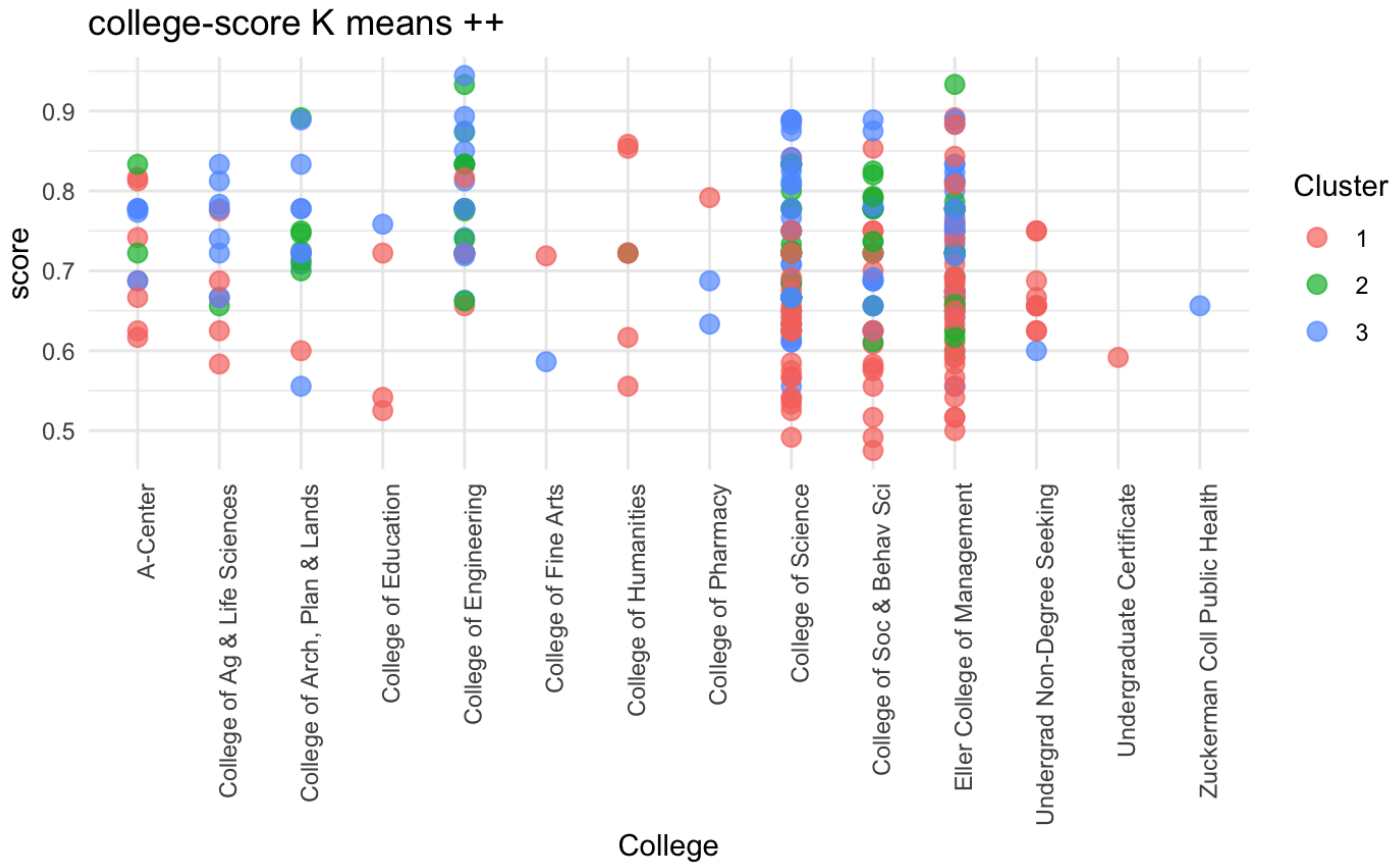
Hide

```
ggplot(df, aes(x = College, y = score, color = as.factor(d_cluster))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "college-score DBSCAN",
       color = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Hide

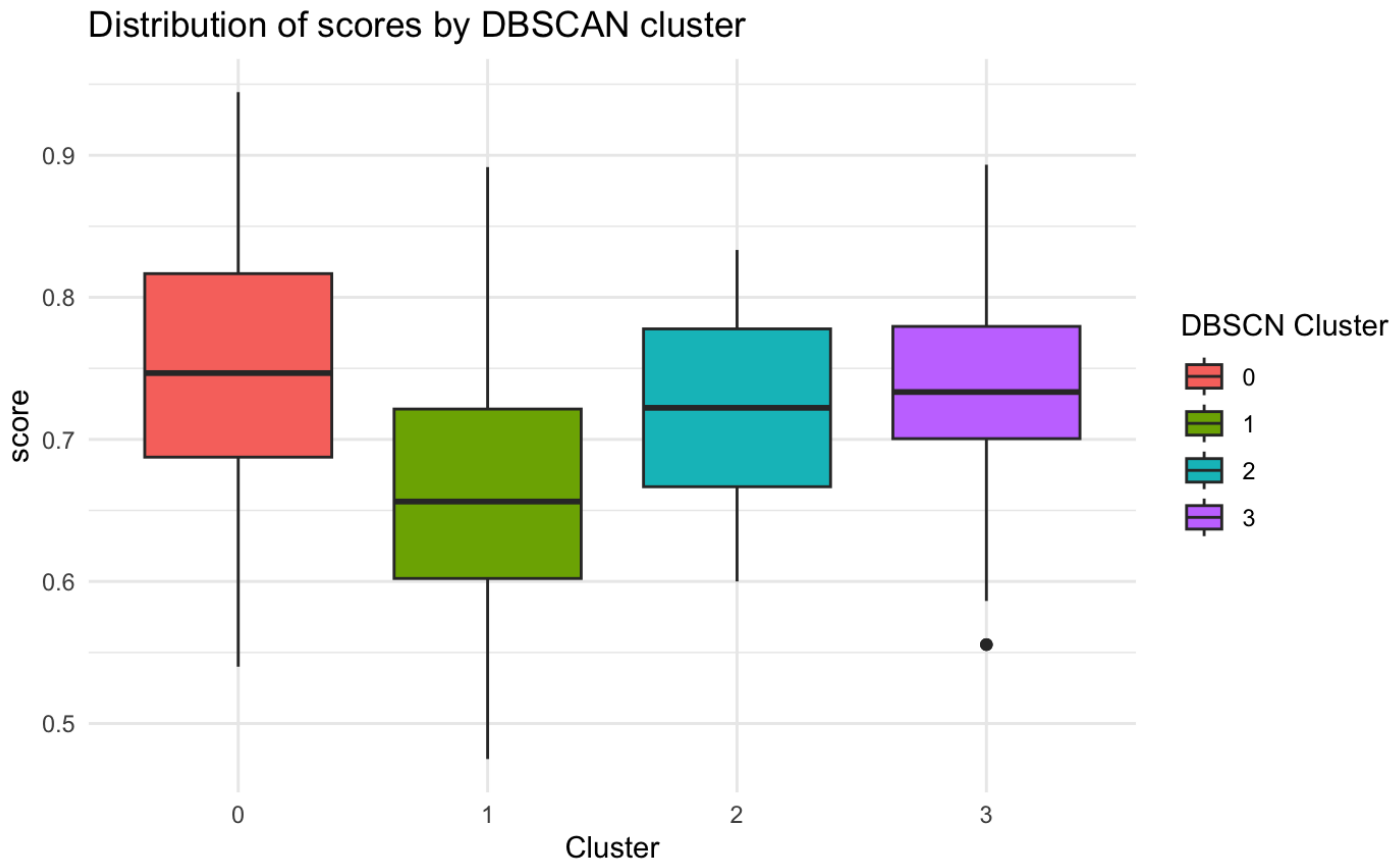
```
library(ggplot2)
ggplot(df, aes(x = College, y = score, color = as.factor(k_cluster))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "college-score K means ++",
       color = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



view score in cluster

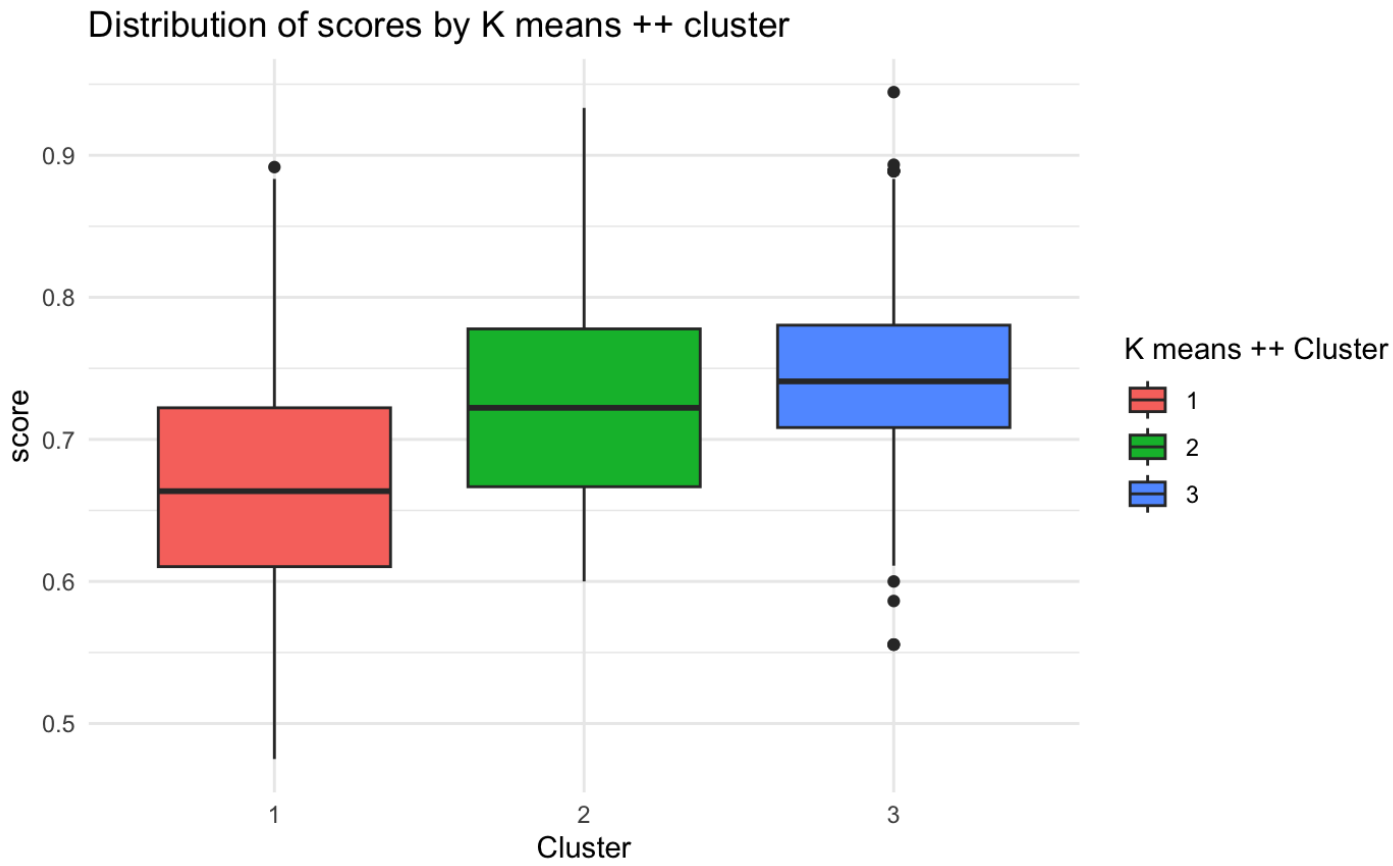
Hide

```
ggplot(df, aes(x = as.factor(d_cluster), y = score, fill = as.factor(d_cluster))) +
  geom_boxplot() +
  labs(title = "Distribution of scores by DBSCAN cluster", x = "Cluster") +
  theme_minimal() +
  scale_fill_discrete(name = "DBSCN Cluster")
```



Hide

```
ggplot(df, aes(x = as.factor(k_cluster), y = score, fill = as.factor(k_cluster))) +
  geom_boxplot() +
  labs(title = "Distribution of scores by K means ++ cluster", x = "Cluster") +
  theme_minimal() +
  scale_fill_discrete(name = "K means ++ Cluster")
```


Distribution of scores by K means ++ cluster
score
0.9
0.8
0.7
0.6
0.5
1
2
3
Cluster
K means ++ Cluster
1
2
3