PCA of Students' Dropout and Academic Success Dataset from Kaggle

Jeremy Haakenson

2024-06-10

The data can be found at: https://www.kaggle.com/datasets/mattop/predict-students-dropout-and-academic-success

Load packages.

Read in data.

```
aca1 = read.csv('academic.csv')
```

Check for NAs.

```
sum(is.na(aca1))
```

```
## [1] 0
```

There are no missing values.

Look at data.

```
skim(aca1)
```

Table 1: Data summary

Name	aca1
Number of rows	4424
Number of columns	37
Column type frequency:	
character	1
numeric	36
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Target	0	1	7	8	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Marital.status	0	1	1.18	0.61	1.00	1.00	1.00	1.00	6.00	
Application.mode	0	1	18.67	17.48	1.00	1.00	17.00	39.00	57.00	
Application.order	0	1	1.73	1.31	0.00	1.00	1.00	2.00	9.00	
Course	0	1	8856.64	2063.57	33.00	9085.00	9238.00	9556.00	9991.00	
Daytime.evening.attendance	0	1	0.89	0.31	0.00	1.00	1.00	1.00	1.00	
Previous.qualification	0	1	4.58	10.22	1.00	1.00	1.00	1.00	43.00	
Previous.qualification..grade.	0	1	132.61	13.19	95.00	125.00	133.10	140.00	190.00	
Nacionality	0	1	1.87	6.91	1.00	1.00	1.00	1.00	109.00	
Mother.s.qualification	0	1	19.56	15.60	1.00	2.00	19.00	37.00	44.00	
Father.s.qualification	0	1	22.28	15.34	1.00	3.00	19.00	37.00	44.00	
Mother.s.occupation	0	1	10.96	26.42	0.00	4.00	5.00	9.00	194.00	
Father.s.occupation	0	1	11.03	25.26	0.00	4.00	7.00	9.00	195.00	
Admission.grade	0	1	126.98	14.48	95.00	117.90	126.10	134.80	190.00	
Displaced	0	1	0.55	0.50	0.00	0.00	1.00	1.00	1.00	
Educational.special.needs	0	1	0.01	0.11	0.00	0.00	0.00	0.00	1.00	
Debtor	0	1	0.11	0.32	0.00	0.00	0.00	0.00	1.00	
Tuition.fees.up.to.date	0	1	0.88	0.32	0.00	1.00	1.00	1.00	1.00	
Gender	0	1	0.35	0.48	0.00	0.00	0.00	1.00	1.00	
Scholarship.holder	0	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00	
Age.at.enrollment	0	1	23.27	7.59	17.00	19.00	20.00	25.00	70.00	
International	0	1	0.02	0.16	0.00	0.00	0.00	0.00	1.00	
Curricular.units.1st.sem..credited.	0	1	0.71	2.36	0.00	0.00	0.00	0.00	20.00	
Curricular.units.1st.sem..enrolled.	0	1	6.27	2.48	0.00	5.00	6.00	7.00	26.00	
Curricular.units.1st.sem..evaluations.	0	1	8.30	4.18	0.00	6.00	8.00	10.00	45.00	
Curricular.units.1st.sem..approved.	0	1	4.71	3.09	0.00	3.00	5.00	6.00	26.00	
Curricular.units.1st.sem..grade.	0	1	10.64	4.84	0.00	11.00	12.29	13.40	18.88	
Curricular.units.1st.sem..without.evaluations.	0	1	0.14	0.69	0.00	0.00	0.00	0.00	12.00	
Curricular.units.2nd.sem..credited.	0	1	0.54	1.92	0.00	0.00	0.00	0.00	19.00	
Curricular.units.2nd.sem..enrolled.	0	1	6.23	2.20	0.00	5.00	6.00	7.00	23.00	
Curricular.units.2nd.sem..evaluations.	0	1	8.06	3.95	0.00	6.00	8.00	10.00	33.00	
Curricular.units.2nd.sem..approved.	0	1	4.44	3.01	0.00	2.00	5.00	6.00	20.00	
Curricular.units.2nd.sem..grade.	0	1	10.23	5.21	0.00	10.75	12.20	13.33	18.57	
Curricular.units.2nd.sem..without.evaluations.	0	1	0.15	0.75	0.00	0.00	0.00	0.00	12.00	
Unemployment.rate	0	1	11.57	2.66	7.60	9.40	11.10	13.90	16.20	
Inflation.rate	0	1	1.23	1.38	-0.80	0.30	1.40	2.60	3.70	
GDP	0	1	0.00	2.27	-4.06	-1.70	0.32	1.79	3.51	

All variables except Target are numeric.

Convert Target from character to categorical.

```
aca2 = aca1 %>%
  mutate(Target = as.factor(Target))
```

Although all features are encoded as numeric, some of them represent categorical data. I will drop these features for PCA.

```
aca3 = aca2[c(7, 13, 20, 22:37)]
```

Scale the data.

```
aca.scale = scale(aca3[1:18])
```

Perform PCA.

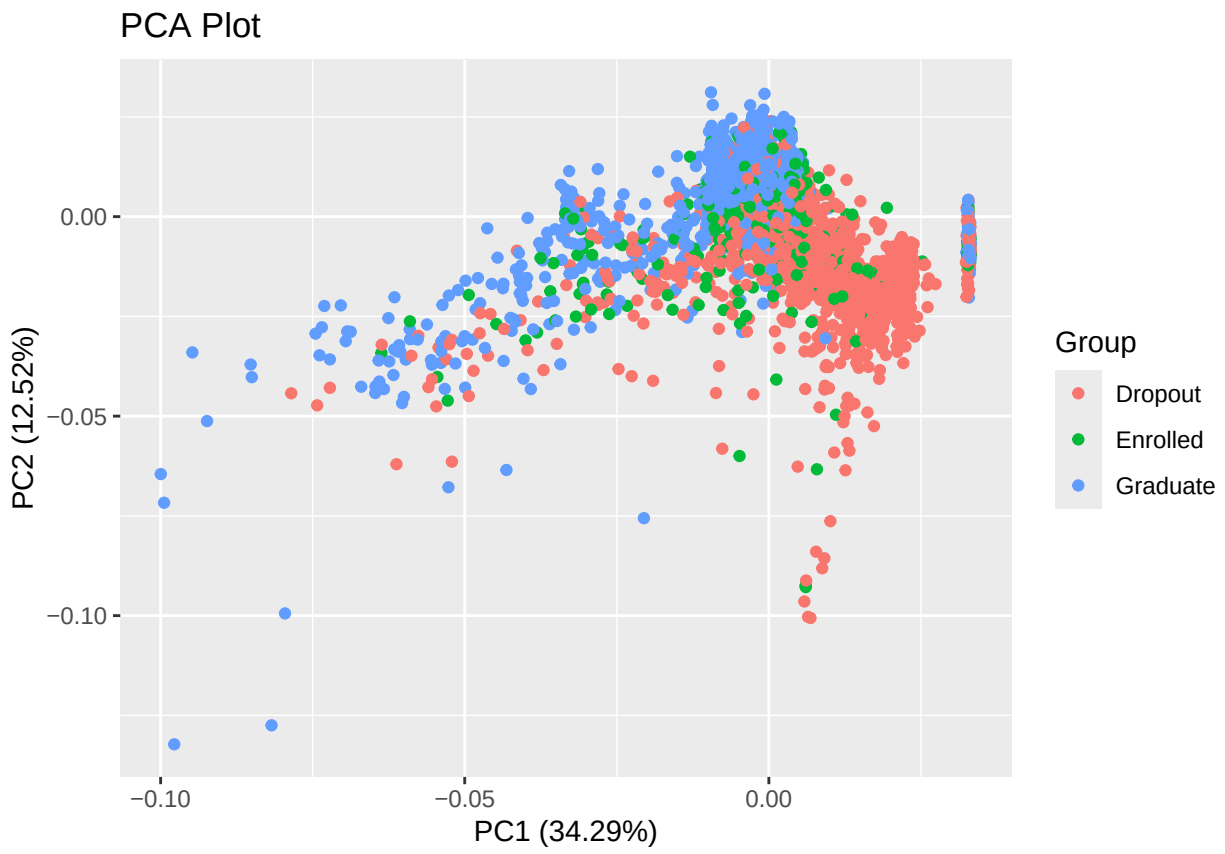
```
aca.pca = prcomp(aca.scale)
summary(aca.pca)
```

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4844 1.5011 1.27133 1.23484 1.14573 1.00481 0.93463
## Proportion of Variance 0.3429 0.1252 0.08979 0.08471 0.07293 0.05609 0.04853
## Cumulative Proportion  0.3429 0.4681 0.55788 0.64260 0.71553 0.77162 0.82015
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.84767 0.79391 0.66307 0.64104 0.60613 0.46180 0.41598
## Proportion of Variance 0.03992 0.03502 0.02443 0.02283 0.02041 0.01185 0.00961
## Cumulative Proportion  0.86006 0.89508 0.91951 0.94234 0.96275 0.97459 0.98421
##                           PC15    PC16    PC17    PC18
## Standard deviation     0.36779 0.30361 0.18908 0.14515
## Proportion of Variance 0.00751 0.00512 0.00199 0.00117
## Cumulative Proportion  0.99172 0.99684 0.99883 1.00000
```

The first 12 principal components explain over 96% of the variance.

Plot PCA.

```
autoplot(aca.pca, data = aca3,
         colour = 'Target',
         label = F) +
  ggtitle('PCA Plot') +
  labs(colour = 'Group')
```



Calculate total variance explained by each principal component.

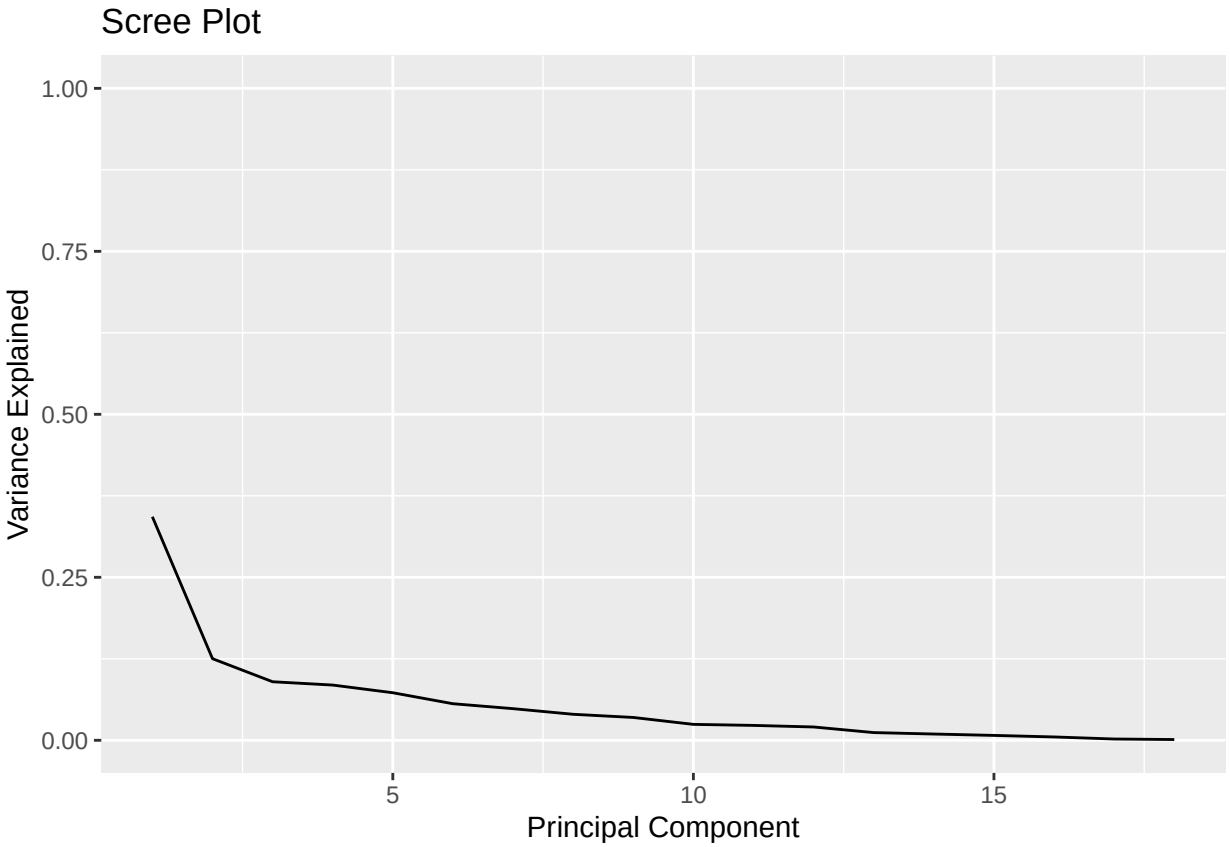
```
var_explained = aca.pca$sdev^2 / sum(aca.pca$sdev^2)
```

Make a dataframe for a scree plot.

```
scree.df = cbind.data.frame(1:length(colnames(aca.pca$x)), var_explained)
colnames(scree.df) = c('PC', 'Var')
```

Make a scree plot.

```
ggplot(scree.df, aes(x = PC, y = Var)) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1)
```


Scree Plot
1.00
0.75
0.50
0.25
0.00
Variance Explained
5
10
15
Principal Component