ANLY_699_Assignment6

Code

Subhash Pemmaraju

July 19, 2020

Principal Component Analysis

Hide

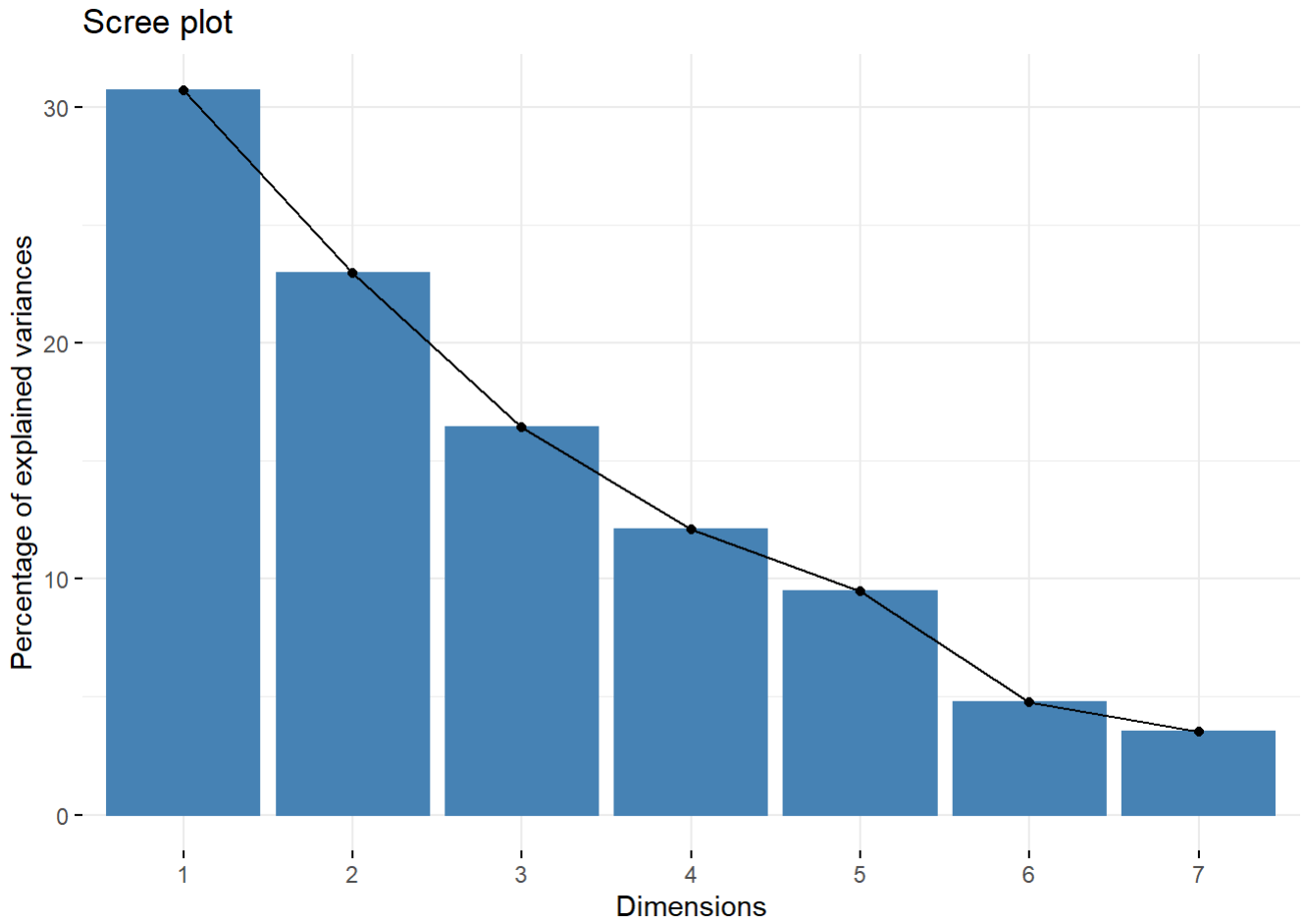
```
clust_data <- merged_data[, c(6,9:19)]
clust_data1 <- clust_data[complete.cases(clust_data),]
#dim(fa_data1)


x<-prcomp(clust_data1[,c(6:12)], retx=TRUE, center=TRUE, scale=TRUE)
summary(x)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4666 1.2683 1.0724 0.9209 0.81477 0.57706 0.49541
## Proportion of Variance 0.3073 0.2298 0.1643 0.1211 0.09484 0.04757 0.03506
## Cumulative Proportion  0.3073 0.5371 0.7014 0.8225 0.91737 0.96494 1.00000
```

Hide

```
fviz_screeplot(x)
```

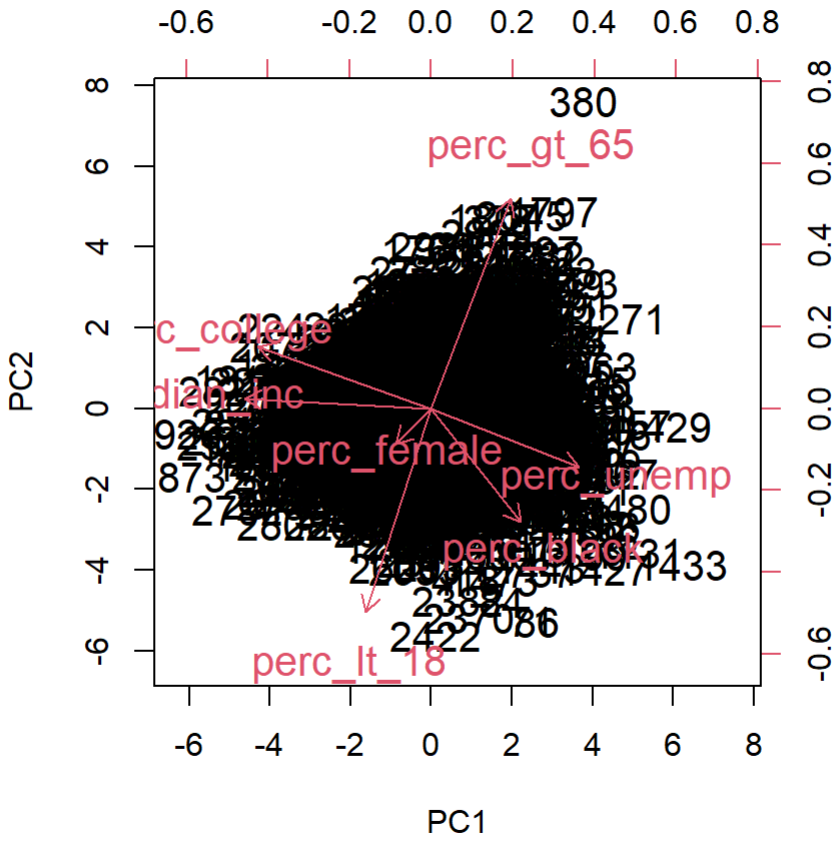


From the proportion of variance, we can see that we need a total of 4 components for over 80% of the variation to be explained. That can also be observed from the screeplot shown above.

Visualization of PCA components

Hide

```
biplot(x,scale=0, cex=1.3)
```



From the biplot, we can see that the variables median_inc, perc_college, perc_unemp contribute the most to PC1. perc_gt_65 and perc_lt_18 contribute the most to PC2. perc_black contribute to both components

Importance of PCA in final project

Using K-means clustering directly on the initial set of demographic variables may not be efficient. PCA can be used to reduce the dimensionality of a bunch of demographic variables. The reduced dimensional principal components can then be used to perform K-means clustering and create multiple clusters that can then be used to perform subsequent analysis.