Cereal PCA

Taylor Anderson

4/25/2020

```
library(factoextra)

## Loading required package: ggplot2

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

library(tidyverse)

## -- Attaching packages ---------------------------------------------------
------------------------------------------------------ tidyverse 1.3.0 --

## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------------------------
------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(dplyr)
library(ggplot2)

cereals_data <- read_csv("cereal.csv")

## Parsed with column specification:
## cols(
##   name = col_character(),
##   mfr = col_character(),
##   type = col_character(),
##   calories = col_double(),
##   protein = col_double(),
##   fat = col_double(),
##   sodium = col_double(),
##   fiber = col_double(),
##   carbo = col_double(),
##   sugars = col_double(),
##   potass = col_double(),
##   vitamins = col_double(),
##   shelf = col_double(),
##   weight = col_double(),
```

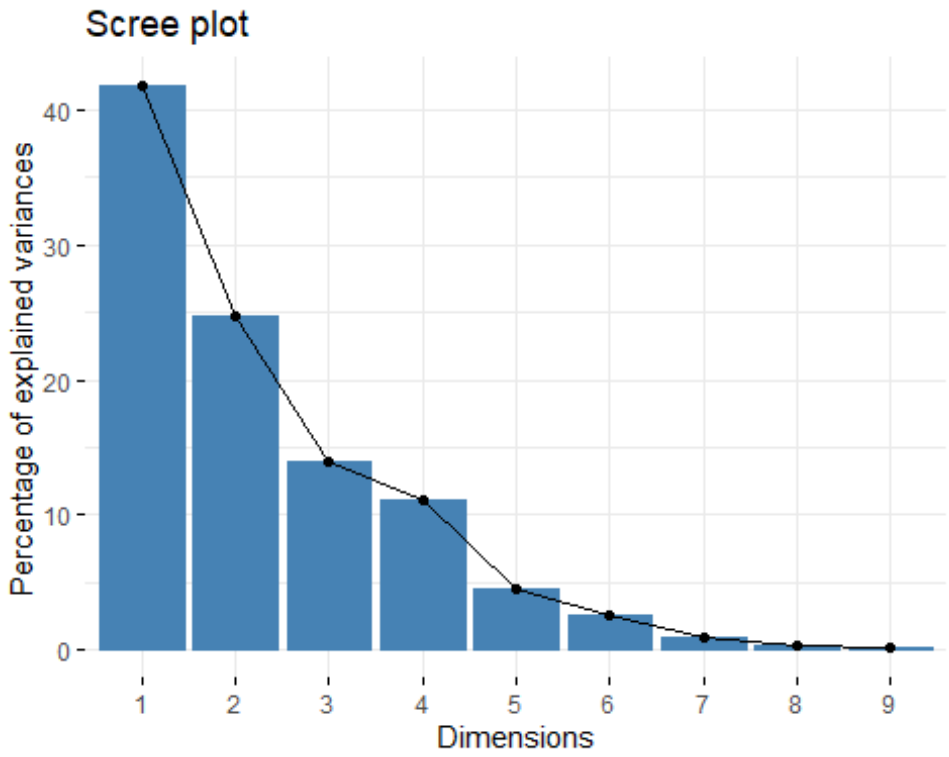
```
##   cups = col_double(),
##   rating = col_double()
## )

colnames(cereals_data) <- c("Name", "Manufacturer", "Type", "Calories",
"Protein", "Fat", "Sodium", "Fiber", "Carbohydrates", "Sugar", "Potassium",
"Vitamins", "Shelf", "Weight", "Cups", "Rating")

# Create feature with full manufacturer name
cereals_data$Manufacturer_Name <- cereals_data$Manufacturer

cereals_data$Manufacturer_Name <- gsub(pattern = "P", replacement = "Post", x
= cereals_data$Manufacturer_Name)
cereals_data$Manufacturer_Name <- gsub(pattern = "A", replacement = "American
Home Food Products", x = cereals_data$Manufacturer_Name)
cereals_data$Manufacturer_Name <- gsub(pattern = "G", replacement = "General
Mills", x = cereals_data$Manufacturer_Name)
cereals_data$Manufacturer_Name <- gsub(pattern = "K", replacement =
"Kellogs", x = cereals_data$Manufacturer_Name)
cereals_data$Manufacturer_Name <- gsub(pattern = "N", replacement =
"Nabisco", x = cereals_data$Manufacturer_Name)
cereals_data$Manufacturer_Name <- gsub(pattern = "Q", replacement = "Quaker
Oats", x = cereals_data$Manufacturer_Name)
cereals_data$Manufacturer_Name <- gsub(pattern = "R", replacement = "Ralston
Purina", x = cereals_data$Manufacturer_Name)
cereals_data$Manufacturer <- factor(cereals_data$Manufacturer)

# Replace negative values with NA
cereals_data$Carbohydrates[cereals_data$Carbohydrates < 0] <- NA
cereals_data$Sugar[cereals_data$Sugar < 0] <- NA
cereals_data$Potassium[cereals_data$Potassium < 0] <- NA

# Add nutritionals per ounce
cereals_data$Calories_oz <- cereals_data$Calories / cereals_data$Weight
cereals_data$Protein_oz <- cereals_data$Protein / cereals_data$Weight
cereals_data$Fat_oz <- cereals_data$Fat / cereals_data$Weight
cereals_data$Sodium_oz <- cereals_data$Sodium / cereals_data$Weight
cereals_data$Fiber_oz <- cereals_data$Fiber / cereals_data$Weight
cereals_data$Carbohydrates_oz <- cereals_data$Carbohydrates /
cereals_data$Weight
cereals_data$Sugar_oz <- cereals_data$Sugar / cereals_data$Weight
cereals_data$Potassium_oz <- cereals_data$Potassium / cereals_data$Weight
cereals_data$Vitamins_oz <- cereals_data$Vitamins / cereals_data$Weight

library(standardize)
# Create subset for PCA
PCA_data <- cereals_data %>%
  select(Name, Manufacturer_Name, Calories = Calories_oz, Protein =
Protein_oz, Fat = Fat_oz, Sodium = Sodium_oz, Fiber = Fiber_oz, Carbohydrates
= Carbohydrates_oz, Sugar = Sugar_oz, Potassium = Potassium_oz, Rating)
```

```r
# Remove observations with NAs
PCA_data <- PCA_data[complete.cases(PCA_data),]

PCA_cereals <- prcomp(PCA_data[, 3:11], scale. = TRUE)

# Obtain Summary of PCA
summary(PCA_cereals)
```

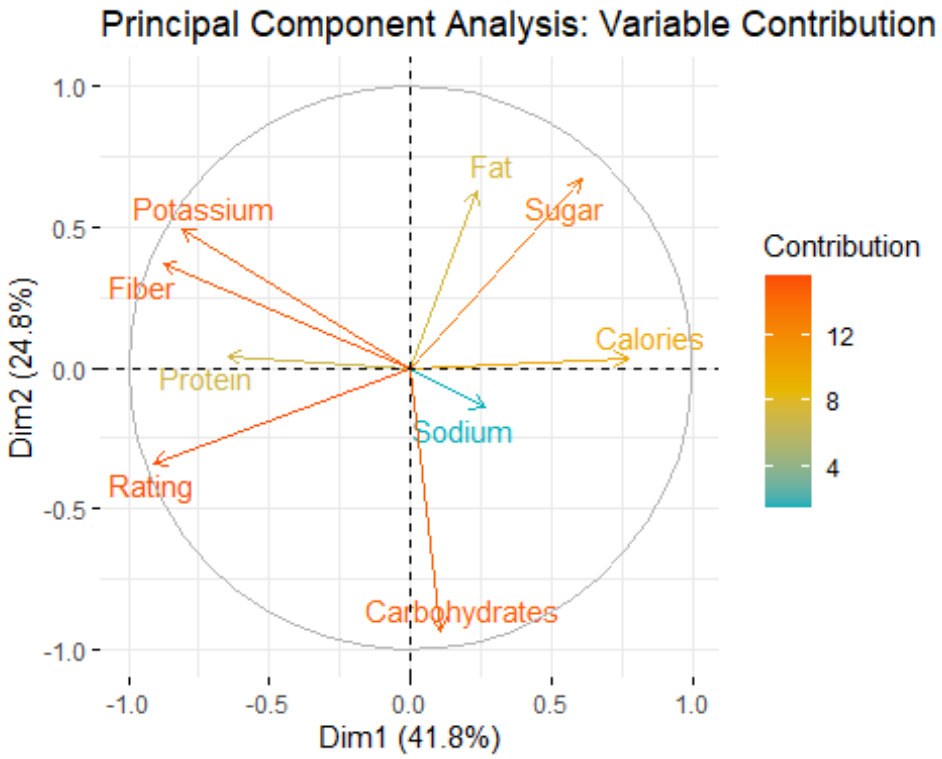
```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.940 1.4927 1.1203 0.9999 0.63540 0.47572 0.29284
## Proportion of Variance 0.418 0.2476 0.1395 0.1111 0.04486 0.02515 0.00953
## Cumulative Proportion  0.418 0.6655 0.8050 0.9161 0.96094 0.98608 0.99561
##                           PC8     PC9
## Standard deviation     0.17150 0.10049
## Proportion of Variance 0.00327 0.00112
## Cumulative Proportion  0.99888 1.00000
```

```r
PCA_stand <- scale(PCA_data[, 3:11])
PCA_cereals_stand <- prcomp(PCA_stand, scale. = TRUE)

# Obtain Scree Plot
fviz_eig(PCA_cereals_stand)
```

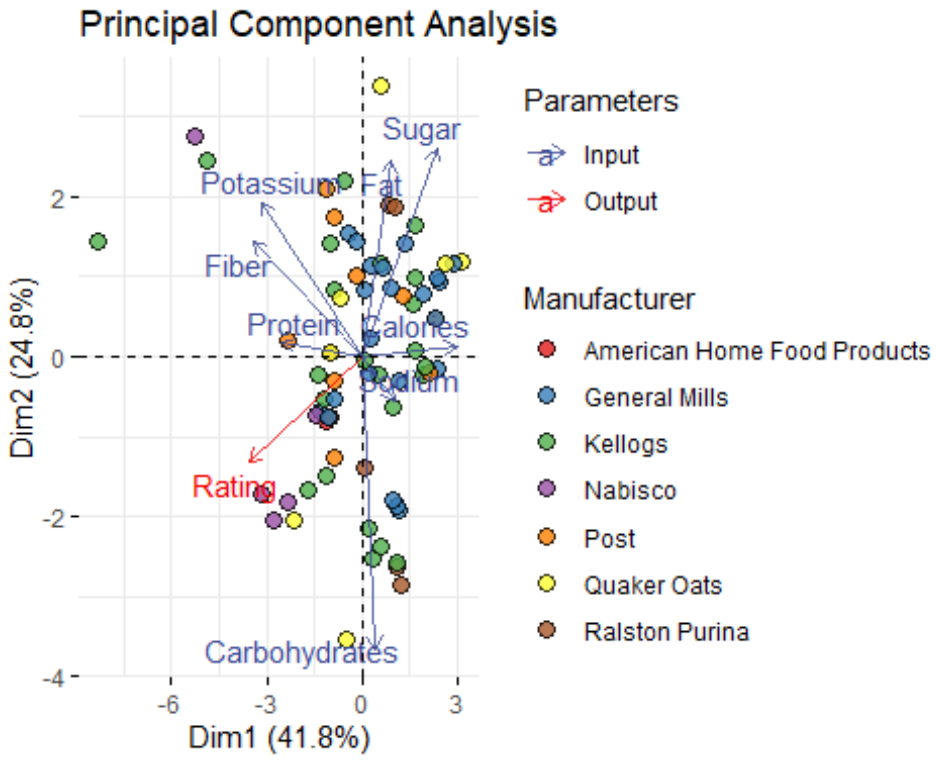
```
## Registered S3 methods overwritten by 'car':
##   method                           from
##   influence.merMod                 lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod          lme4
##   dfbetas.influence.merMod         lme4
```




```r
library(ggpubr)

# PCA Variables
fviz_pca_var(PCA_cereals_stand,
             col.var = "contrib",
             repel = TRUE,
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             title = "Principal Component Analysis: Variable Contribution",
             legend.title = "Contribution"
             )
```




```r
# PCA Biplot: Variables and Individuals
fviz_pca_biplot(PCA_cereals_stand,
                geom.ind = "point",
                pointshape = 21,
                pointsize = 3,
                fill.ind = PCA_data$Manufacturer_Name,
                # col.ind = "Black",
                alpha = 0.8,
                mean.point = FALSE,
                col.var = factor(c("Input", "Input", "Input", "Input",
"Input", "Input", "Input", "Input", "Output")), # Colour inputs and outputs
differently
                repel = TRUE,
                legend.title = list(fill = "Manufacturer", color =
"Parameters"),
                title = "Principal Component Analysis") +
  fill_palette("Set1") + # Palette for individuals
  color_palette(palette = "aaas") # Palette for variables
```


Principal Component Analysis
Sugar
Potassium
Fat
Fiber
Protein
Calories
Sodium
Rating
Carbohydrates
Dim2 (24.8%)
Dim1 (41.8%)
2
0
-2
-4
-6
-3
0
3
Parameters
Input
Output
Manufacturer
American Home Food Products
General Mills
Kellogs
Nabisco
Post
Quaker Oats
Ralston Purina