Homework 9

Peyton Hall

04/11/2025

```
library(readxl)
```

```
Car_data <- read_excel("~/Desktop/STAT 301/Week 10/Car data.xlsx")
# Car_data
```

Question 1 Code

```
# a)
# select only numeric variables for PCA
car_numeric <- Car_data[, c("DealerCost", "Price", "Engine", "MPG", "Weight")]
car_scaled <- scale(car_numeric) # standardize the numeric data
pca_result <- prcomp(car_scaled) # perform PCA
eigenvalues <- pca_result$sdev^2 # (squared standard deviations of PCs)
eigenvalues
```

```
## [1] 3.15196410 1.17968208 0.41449846 0.22261167 0.03124368
```

```
# b)
# get proportion of variance explained
variance_explained <- eigenvalues / sum(eigenvalues)
sum(variance_explained[1:2]) # sum of first two components
```

```
## [1] 0.8663292
```

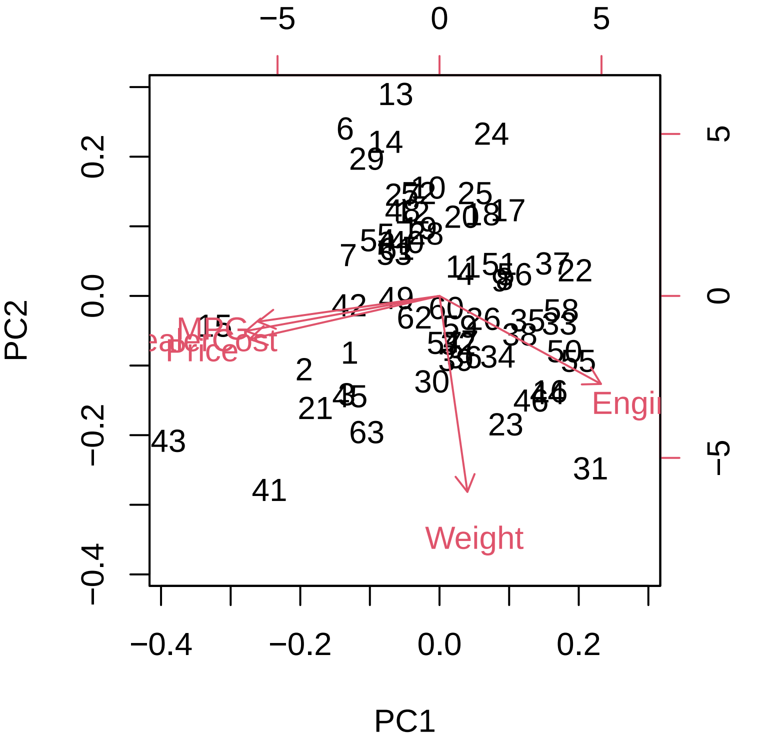
```
pca_result <- prcomp(~DealerCost + Price + Engine + MPG + Weight, data = Car_data, scale = TRUE)
result <- summary(pca_result)
result$importance
```

```
##                             PC1      PC2       PC3       PC4       PC5
## Standard deviation     1.775377 1.086132 0.6438155 0.4718174 0.1767588
## Proportion of Variance 0.630390 0.235940 0.0829000 0.0445200 0.0062500
## Cumulative Proportion  0.630390 0.866330 0.9492300 0.9937500 1.0000000
```

```
# d)
result$rotation
```

```
##                   PC1        PC2        PC3        PC4          PC5
## DealerCost -0.53032576 -0.1568396 -0.3946416 -0.1278723 -0.722539022
## Price      -0.52028236 -0.1937809 -0.4537840 -0.1009969  0.689662933
## Engine      0.44215861 -0.3939680 -0.5429699  0.5934762 -0.047483917
## MPG        -0.49670709 -0.1152040  0.4676786  0.7219739  0.006365367
## Weight      0.07641133 -0.8771360  0.3532684 -0.3162089 -0.002675738
```

```
# f)
biplot(pca_result)
```



```
library(readxl)
personal_test_scores <- read_excel("~/Desktop/STAT 301/Week 10/personal test scores.xlsx")
# personal_test_scores
library(factoextra)
```

```
## Loading required package: ggplot2
```

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

Question 2 Code

```
# a)
pca_result2 <- prcomp(~Anxiety + Agoraphobia + Adventure + Sociability, data = personal_test_scores, sca
result2 <- summary(pca_result2)
get_eig(pca_result2)
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 2.97689356       74.4223390                    74.42234
## Dim.2 0.95723625       23.9309061                    98.35325
## Dim.3 0.04422886        1.1057215                    99.45897
## Dim.4 0.02164134        0.5410334                   100.00000
```

```
# b)
eig_values <- get_eig(pca_result2)
eig_values[1:2, "variance.percent"] # variance explained by PC1 and PC2
```

```
## [1] 74.42234 23.93091
```

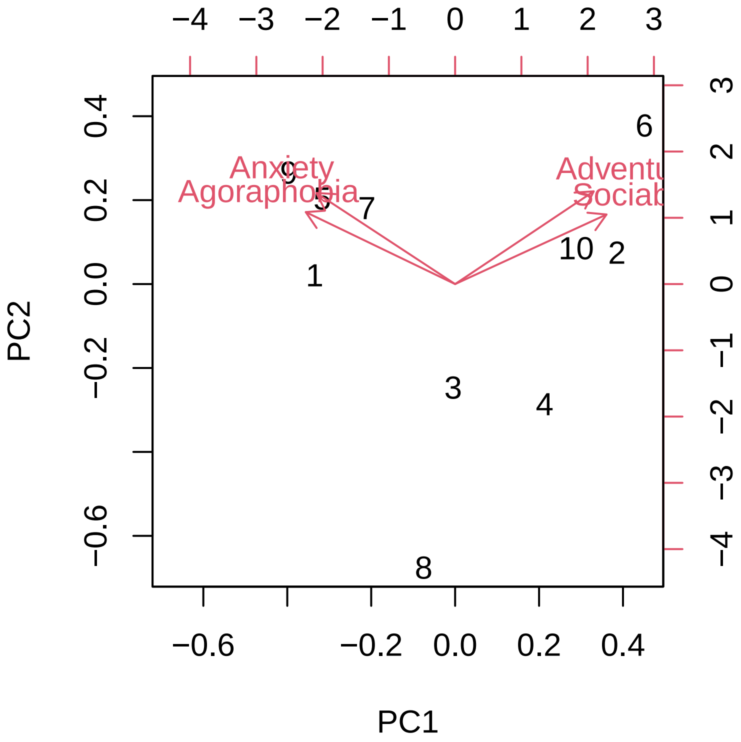
```
# get the total variance explained by the first two components
sum(eig_values[1:2, "variance.percent"])
```

```
## [1] 98.35325
```

```
# d)
pca_result2$rotation
```

```
##                    PC1       PC2        PC3        PC4
## Anxiety     -0.4796322 0.5552569  0.6673019 -0.1278702
## Agoraphobia -0.5163307 0.4385256 -0.6842992  0.2698750
## Adventure    0.4790256 0.5651897 -0.2457594 -0.6250579
## Sociability  0.5233451 0.4242000  0.1613847  0.7211930
```

```
# e)
biplot(pca_result2)
```



```r
library(readxl)
premium_and_discount_bond <- read_excel("~/Desktop/STAT 301/Week 10/premium and discount bond.xlsx")
# premium_and_discount_bond
```

Question 3 Code

```r
# remove the Date column and scale the data
bond_data <- premium_and_discount_bond[, -1] # remove 'Date'
pca_bond <- prcomp(bond_data, scale. = TRUE)

# a)
library(factoextra)
get_eig(pca_bond)
```

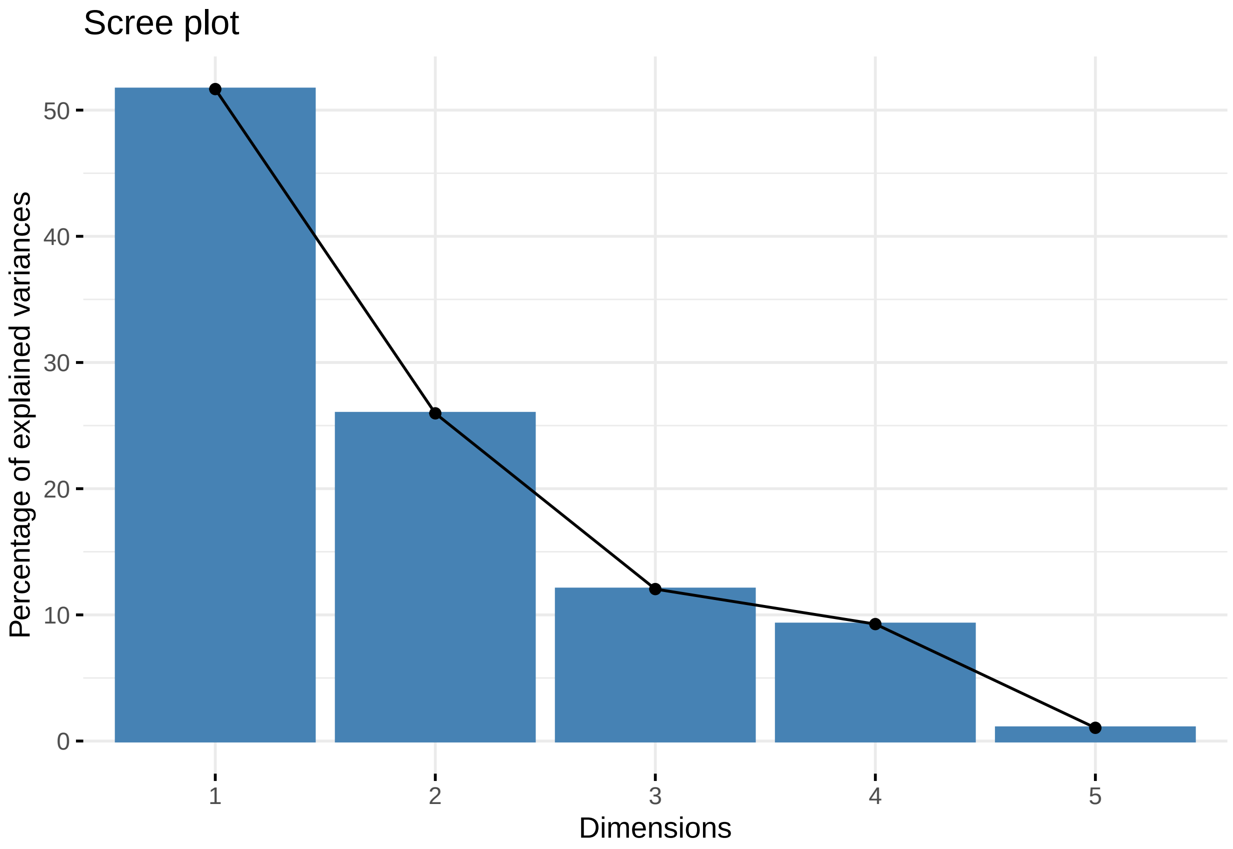
```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 2.58344602        51.668920                    51.66892
## Dim.2 1.29851146        25.970229                    77.63915
## Dim.3 0.60215645        12.043129                    89.68228
## Dim.4 0.46348464         9.269693                    98.95197
## Dim.5 0.05240143         1.048029                   100.00000
```

```r
# b)
summary(pca_bond)
```

```
## Importance of components:
```

```
##                           PC1    PC2    PC3    PC4     PC5
## Standard deviation     1.6073 1.1395 0.7760 0.6808 0.22891
## Proportion of Variance 0.5167 0.2597 0.1204 0.0927 0.01048
## Cumulative Proportion  0.5167 0.7764 0.8968 0.9895 1.00000
```

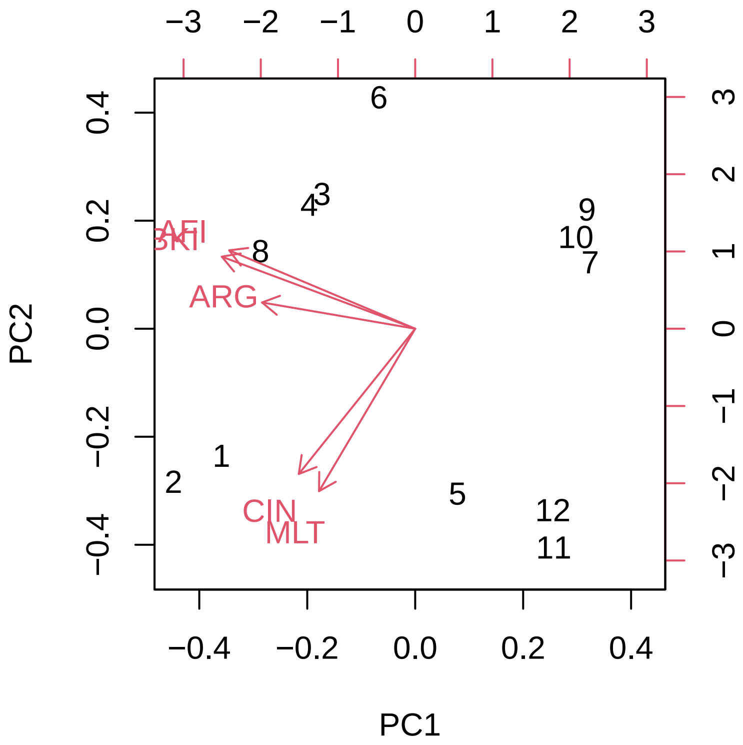
```
# c)
fviz_eig(pca_bond)
```



```
# d)
pca_bond$rotation
```

```
##            PC1        PC2         PC3         PC4         PC5
## CIN -0.3385763 -0.5954612 -0.03884496  0.72283513 -0.08241626
## BKI -0.5623764  0.2953799 -0.26089382  0.04858906  0.72529787
## ARG -0.4454192  0.1076249  0.88272650 -0.08009771 -0.06630893
## AFI -0.5408593  0.3212190 -0.37079173 -0.08593139 -0.67780443
## MLT -0.2795921 -0.6658918 -0.11718743 -0.67922649  0.05774810
```

```
# e)
biplot(pca_bond)
```



```r
library(readxl)
Heptathlon <- read_excel("~/Desktop/STAT 301/Week 10/Heptathlon.xlsx")
Heptathlon
```

```
## # A tibble: 25 x 6
##    Name           Hurdles Highjump  Shot Longjump Run800
##    <chr>            <dbl>    <dbl> <dbl>    <dbl>  <dbl>
##  1 Joyner-Kersee     12.7     1.86  15.8     7.27   129.
##  2 John              12.8     1.8   16.2     6.71   126.
##  3 Behmer            13.2     1.83  14.2     6.68   124.
##  4 Sablovskaite      13.6     1.8   15.2     6.25   132.
##  5 Choubenkova       13.5     1.74  14.8     6.32   128.
##  6 Schulz            13.8     1.83  13.5     6.33   126.
##  7 Fleming           13.4     1.8   12.9     6.37   133.
##  8 Greiner           13.6     1.8   14.1     6.47   134.
##  9 Lajbnerova        13.6     1.83  14.3     6.11   136.
## 10 Bouraga           13.2     1.77  12.6     6.28   135.
## # i 15 more rows
```

Question 4 Code

```r
library(factoextra)

# remove the Name column and scale the numeric data
hep_data <- Heptathlon[, -1]
```

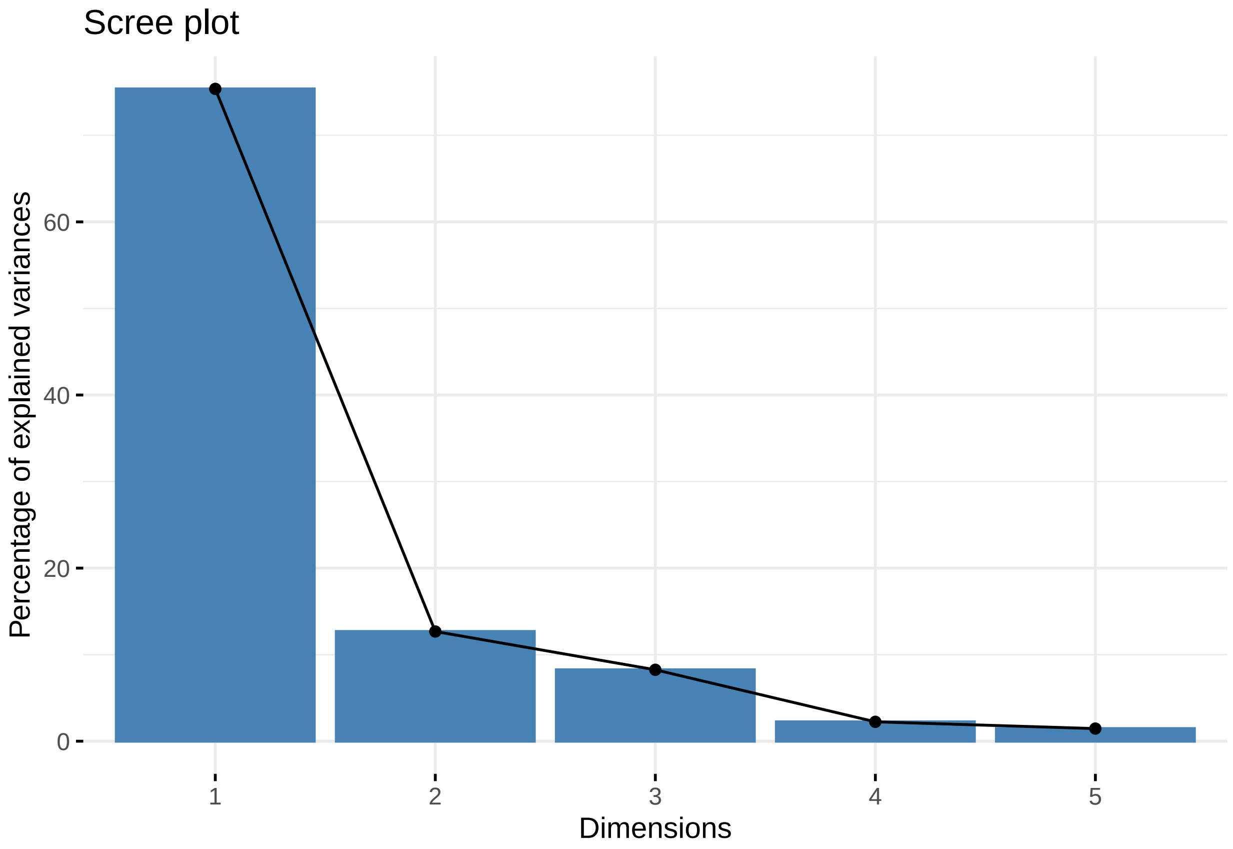
```r
pca_hep <- prcomp(hep_data, scale. = TRUE) # run PCA

# a)
get_eig(pca_hep)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 3.76839300        75.367860                    75.36786
## Dim.2 0.63386703        12.677341                    88.04520
## Dim.3 0.41266100         8.253220                    96.29842
## Dim.4 0.11206776         2.241355                    98.53978
## Dim.5 0.07301122         1.460224                   100.00000
```

```r
# b)
summary(pca_hep)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5
## Standard deviation     1.9412 0.7962 0.64239 0.33477 0.2702
## Proportion of Variance 0.7537 0.1268 0.08253 0.02241 0.0146
## Cumulative Proportion  0.7537 0.8804 0.96298 0.98540 1.0000
```

```r
# c)
fviz_eig(pca_hep)
```



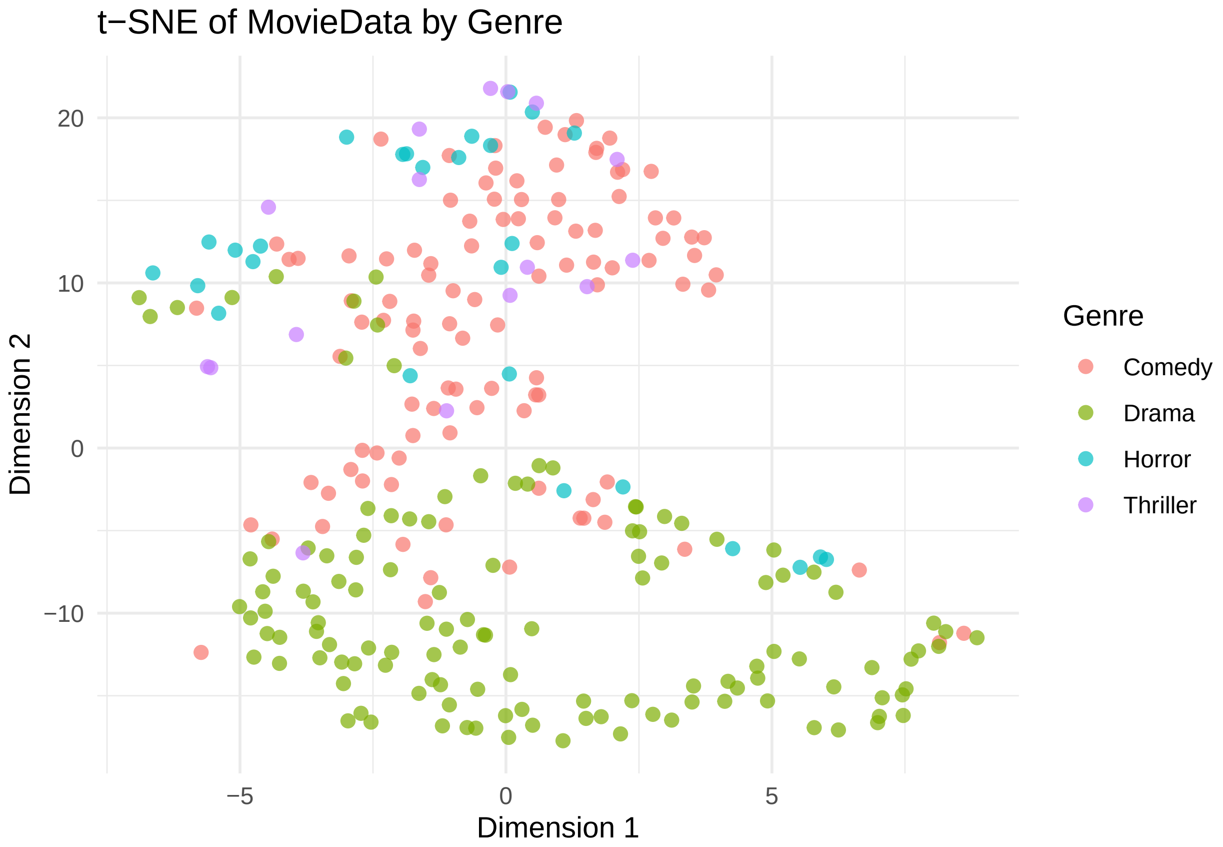
```r
# d)
pca_hep$rotation
```

```
##                  PC1         PC2         PC3        PC4         PC5
## Hurdles   0.4973614 -0.09440773 -0.01137021  0.4476206  0.73703821
## Highjump -0.4351329  0.33166540  0.67840137  0.4877423  0.05036392
## Shot     -0.3820355 -0.81890267 -0.12497869  0.3993114 -0.09153131
## Longjump -0.4944516 -0.12523778  0.09187874 -0.5512653  0.65383329
## Run800    0.4157613 -0.44136058  0.71803964 -0.3136881 -0.13550717
```

```r
library(readxl)
MovieData <- read_excel("~/Desktop/STAT 301/Week 10/MovieData.xlsx")
# MovieData
```

Question 5 Code

```r
# install.packages("Rtsne")
library(Rtsne)
library(ggplot2)

# prepare the numeric data for t-SNE (remove genres column)
tsne_input <- scale(MovieData[, c("rating", "popularity", "vote_average")])

set.seed(123) # reproducibility
tsne_result <- Rtsne(tsne_input, dims = 2, perplexity = 30)

# combine results with genres for plotting
tsne_df <- data.frame(
  X = tsne_result$Y[,1],
  Y = tsne_result$Y[,2],
  Genre = MovieData$genres
)

# t-SNE result
ggplot(tsne_df, aes(x = X, y = Y, color = Genre)) +
  geom_point(size = 2, alpha = 0.7) +
  theme_minimal() +
  labs(title = "t-SNE of MovieData by Genre",
       x = "Dimension 1", y = "Dimension 2")
```


t−SNE of MovieData by Genre
Dimension 2
20
10
0
−10
Dimension 1
−5
0
5
Genre
Comedy
Drama
Horror
Thriller