Lab Assignment Ten

Noah Gallego

2024-10-11

Lab Assignment 10: PCA Analysis on the MTCars Dataset

Import Libraries

```
library(dplyr)      # For data manipulation
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)    # For plotting
library(ggrepel)    # For Last Plot
library(factoextra) # For Scree Plot
```

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

Exploring the Dataset

```
data("mtcars")
df = mtcars
# Remove Unessecary Columns
df = df %>% select(c("mpg", "disp", "hp", "drat", "wt", "qsec"))

# Display First Few Columns
head(df)
```

```
##                    mpg disp  hp drat    wt  qsec
## Mazda RX4         21.0  160 110 3.90 2.620 16.46
## Mazda RX4 Wag     21.0  160 110 3.90 2.875 17.02
## Datsun 710        22.8  108  93 3.85 2.320 18.61
## Hornet 4 Drive    21.4  258 110 3.08 3.215 19.44
## Hornet Sportabout 18.7  360 175 3.15 3.440 17.02
## Valiant           18.1  225 105 2.76 3.460 20.22
```

```r
# Scale Dataset for PCA
df_scaled = scale(df)
apply(df_scaled, 2, sd)
```

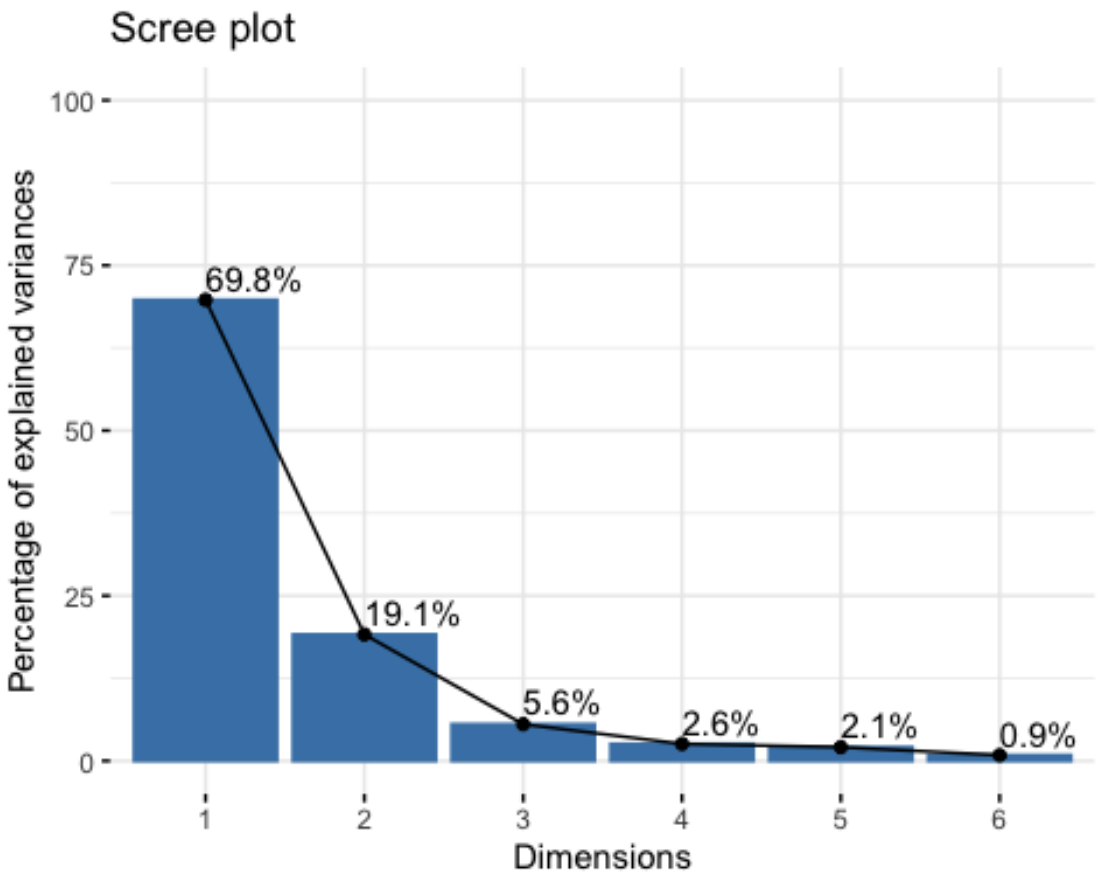
```
##  mpg disp   hp drat   wt qsec
##    1    1    1    1    1    1
```

Perform PCA

```r
# Perform PCA
pca = prcomp(df_scaled)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5     PC6
## Standard deviation     2.0463 1.0715 0.57737 0.39289 0.3533 0.22799
## Proportion of Variance 0.6979 0.1913 0.05556 0.02573 0.0208 0.00866
## Cumulative Proportion  0.6979 0.8892 0.94481 0.97054 0.9913 1.00000
```

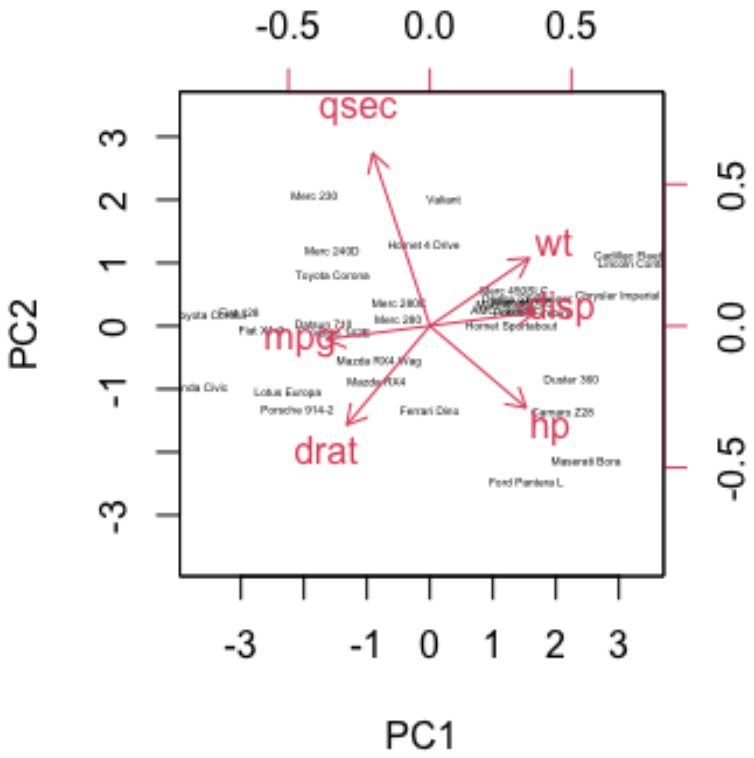
```r
# Display Proportion of Variance
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 100))  # Scree plot showing variance explained by components
```




```r
# Examine Loadings
pca$rotation
```

```
##              PC1         PC2         PC3         PC4         PC5         PC6
## mpg  -0.4586835 -0.05867609  0.19479235 -0.78205878  0.1111533 -0.35249327
## disp  0.4660354  0.06065296 -0.09688406 -0.60001871 -0.2946297  0.56825752
## hp    0.4258534 -0.36147576 -0.14613554 -0.12301873  0.8057408 -0.04771555
## drat -0.3670963 -0.43652537 -0.80049152 -0.02259258 -0.1437714  0.11277675
## wt    0.4386179  0.29953457 -0.41776208 -0.10438337 -0.2301541 -0.69246040
## qsec -0.2528320  0.76284877 -0.34059066 -0.04268124  0.4218755  0.24152663
```

```
# Display Biplot
biplot(pca, scale = 0, cex = c(0.3, 1))
```

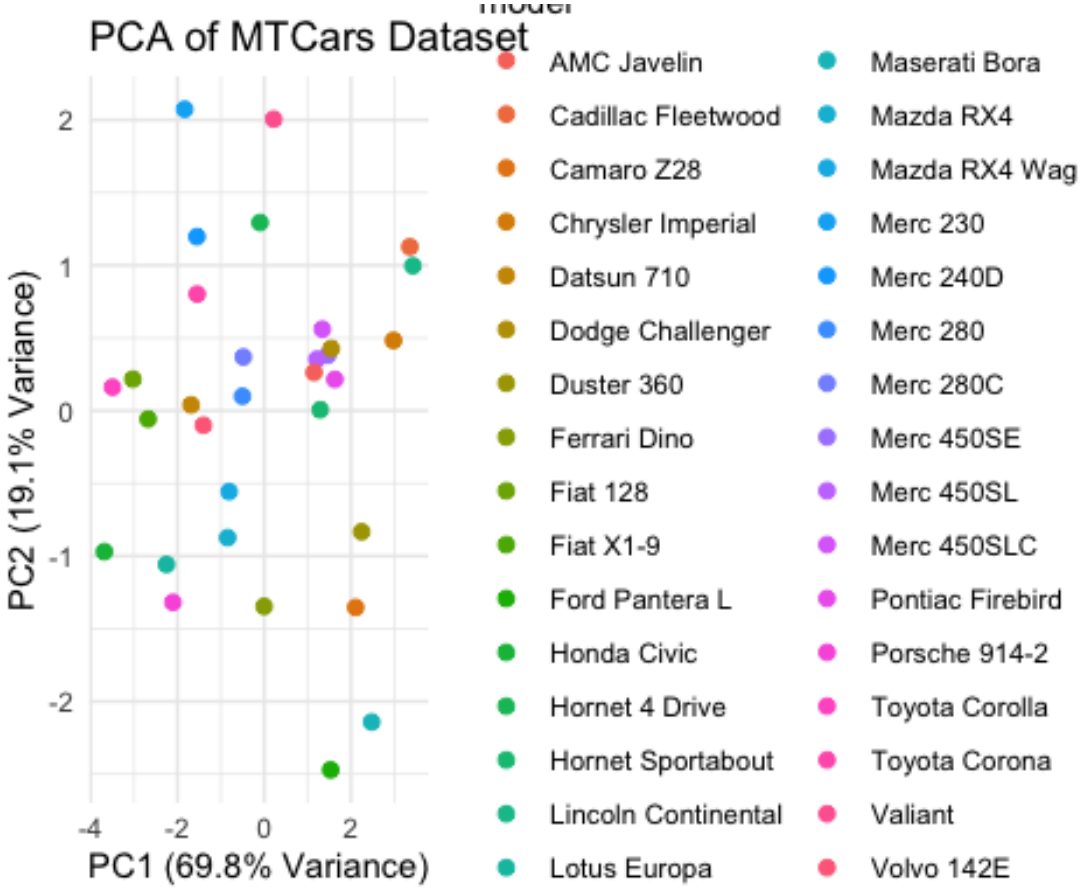


Diagnostic Plots & Visualizing Principal Components

```
# Create a data frame with the principal components and model
pca_df = data.frame(
  pca$x, model = rownames(df)
)

variance_explained = pca$sdev^2 / sum(pca$sdev^2)
# Examine PCA Components on a 2D Plane
ggplot(pca_df, aes(x = PC1, y = PC2, color = model)) +
  geom_point(size = 2) +
  ggtitle("PCA of MTCars Dataset") +
  xlab(paste0("PC1 (", round(variance_explained[1] * 100, 1), "% Variance)"))
```

```
+
  ylab(paste0("PC2 (", round(variance_explained[2] * 100, 1), "% Variance)"))
+
  theme_minimal()
```

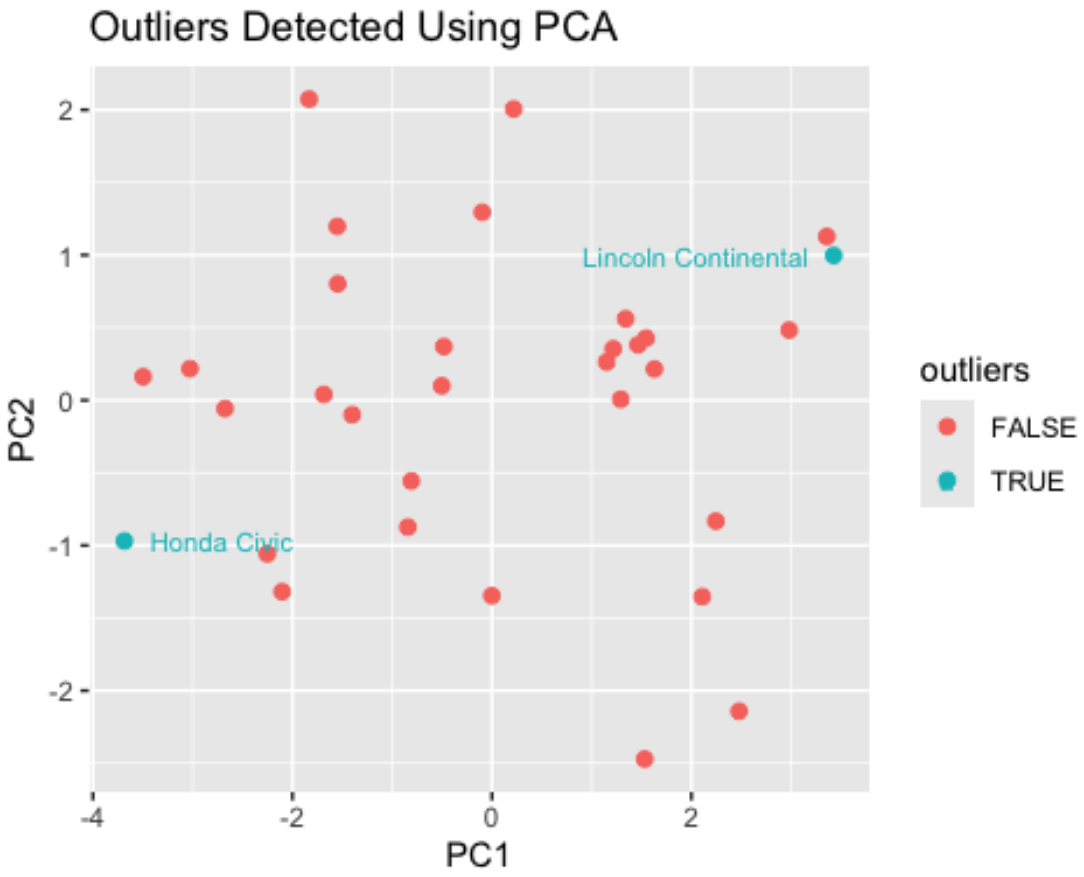


Outlier Detection

```
pca_distances = sqrt(rowSums(pca$x[, 1:2]^2))
outlier_threshold = quantile(pca_distances, 0.95)
outliers = pca_distances > outlier_threshold

pca_df$outliers = outliers

# Plot the PCA with outliers labeled
ggplot(pca_df, aes(x = PC1, y = PC2, color = outliers)) +
  geom_point(size = 2) +
  geom_text_repel(data = subset(pca_df, outliers), aes(label = model),
                  size = 3, box.padding = 0.5, max.overlaps = Inf) +
  labs(title = "Outliers Detected Using PCA")
```


Outliers Detected Using PCA
Lincoln Continental
Honda Civic
PC1
PC2
outliers
FALSE
TRUE