HW 9

Runze Wang

You can run the following code to prepare the analysis.

```
library(r02pro)     #INSTALL IF NECESSARY
library(tidyverse)  #INSTALL IF NECESSARY
library(MASS)
my_ahp <- ahp %>% dplyr::select(gar_car, liv_area, lot_area, bsmt_area, gar_area, oa_qual, sale_price, 
  na.omit()
my_ahp_x <- my_ahp %>% dplyr::select(-sale_price)
my_ahp_y <- my_ahp %>% dplyr::select(sale_price)
```

```
summary(cars)
```

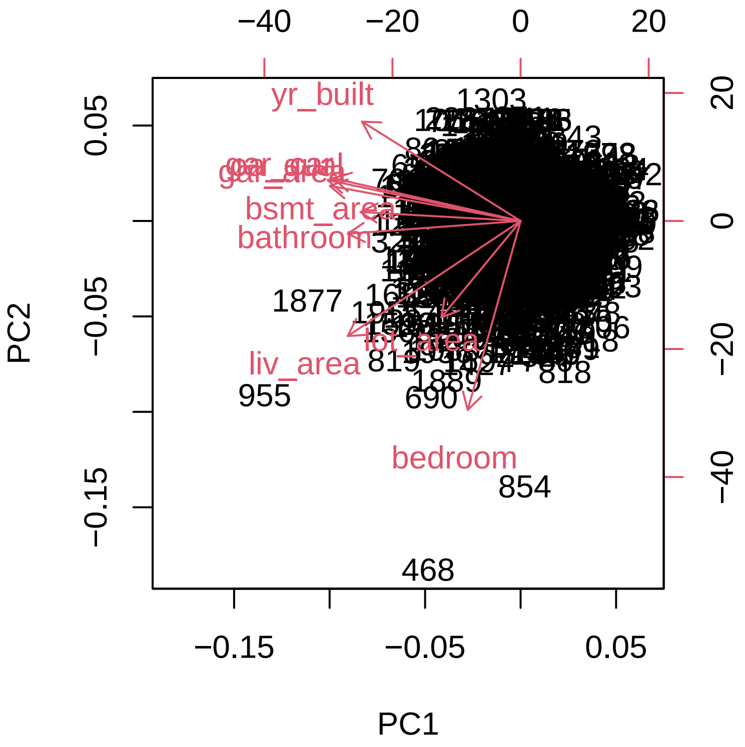
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

0.1 1. Conduct PCA on `my_ahp_x` with `scale = TRUE`.

```
pca_result <- prcomp(my_ahp_x, scale = TRUE)
```

0.2 a. Create a biplot.

```
biplot(pca_result)
```



0.3 b. Plot the Proportion of Variance Explained and the Cumulative Proportion of Variance Explained.

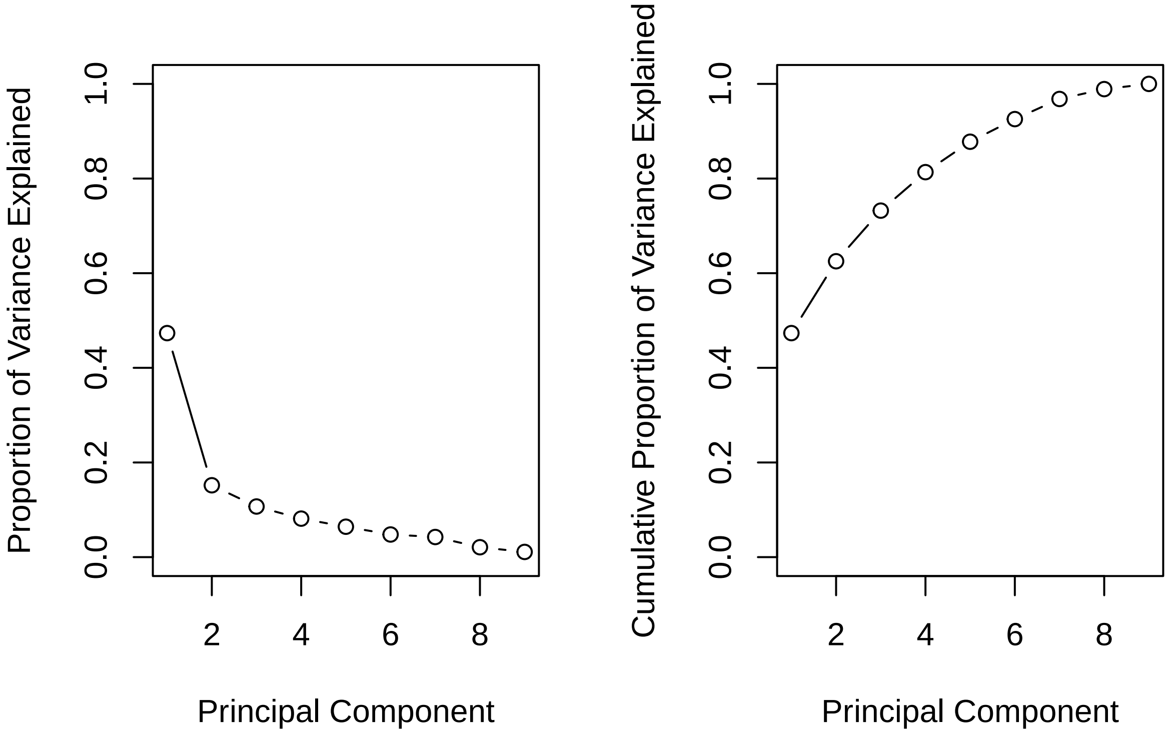
```r
pr.var <- pca_result$sdev^2
pve <- pr.var / sum(pr.var)
pve
```

```
## [1] 0.47340876 0.15182550 0.10697572 0.08133667 0.06426047 0.04781993 0.04252555
## [8] 0.02089640 0.01095101
```

```r
par(mfrow = c(1, 2))

plot(pve, xlab = "Principal Component",
ylab = "Proportion of Variance Explained", ylim = c(0, 1),
type = "b")

plot(cumsum(pve), xlab = "Principal Component",
ylab = "Cumulative Proportion of Variance Explained", ylim = c(0, 1), type = "b")
```



0.4 c. Fit a linear regression of `sale_price` on the first two principle components. What's the R^2?

```
pc_scores <- pca_result$x[, 1:2]
model_pca <- lm(sale_price ~ .,  data = cbind(my_ahp_y, pc_scores))
summary(model_pca)$r.squared
```

```
## [1] 0.7451746
```

0.5 d. Fit a linear regression of `sale_price` on `gar_car` and `liv_area`. What's the R^2?

```
model2 <- lm(sale_price ~ gar_car + liv_area, data = my_ahp)

summary(model2)$r.squared
```

```
## [1] 0.6124356
```

0.6 2. Conduct PCA on `my_ahp_x` with `scale = FALSE` and compare the results of a-c with those of Q1.

```
pca_result2 <- prcomp(my_ahp_x, scale = FALSE)
```

0.7 a. Create a biplot.

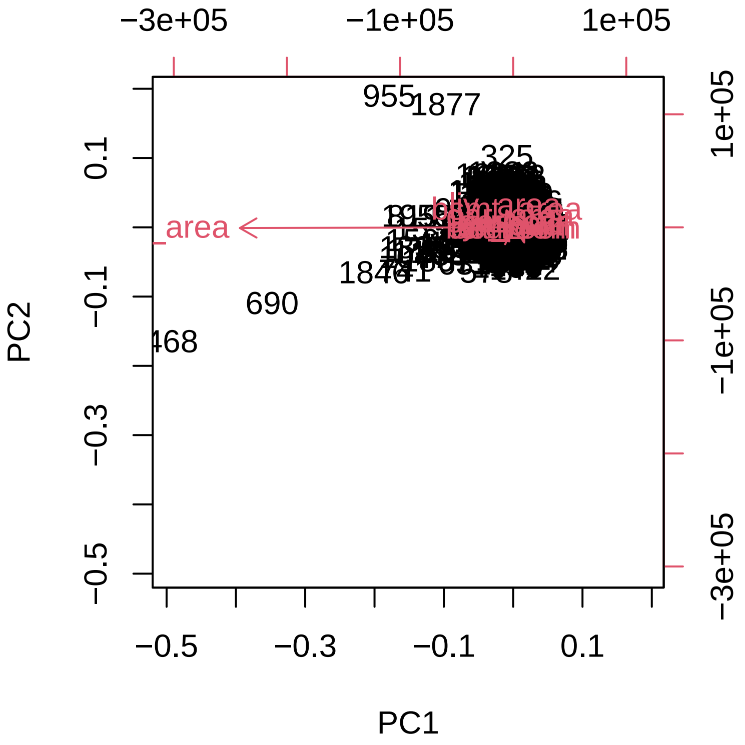
```
biplot(pca_result2)
```

```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```



0.8 b. Plot the Proportion of Variance Explained and the Cumulative Proportion of Variance Explained.

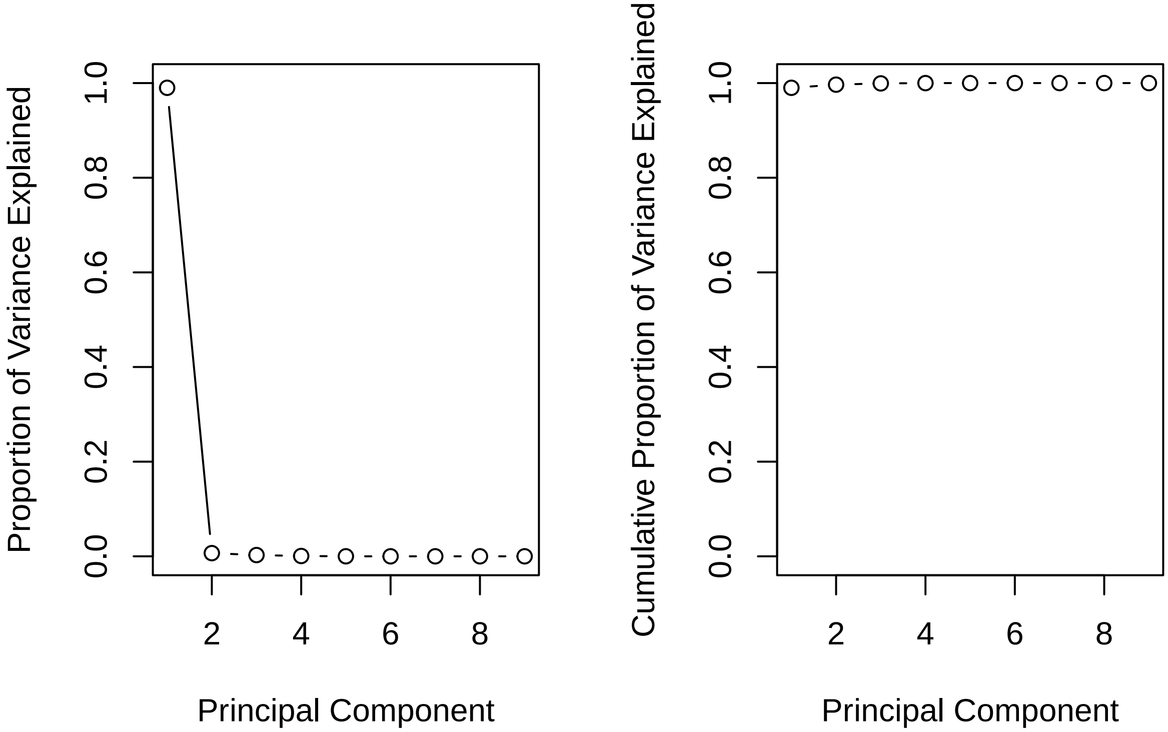
```
pr.var2 <- pca_result2$sdev^2
pve2 <- pr.var2 / sum(pr.var2)
pve2
```

```
## [1] 9.900244e-01 6.670033e-03 2.634425e-03 6.569126e-04 1.417451e-05
## [6] 1.898412e-08 9.247808e-09 7.035001e-09 2.292847e-09
```

```r
par(mfrow = c(1, 2))

plot(pve2, xlab = "Principal Component",
ylab = "Proportion of Variance Explained", ylim = c(0, 1),
type = "b")

plot(cumsum(pve2), xlab = "Principal Component",
ylab = "Cumulative Proportion of Variance Explained", ylim = c(0, 1), type = "b")
```



c. Fit a linear regression of `sale_price` on the first two principle components. What's the R^2?

```r
pc_scores2 <- pca_result2$x[, 1:2]
model_pca2 <- lm(sale_price ~ ., data = cbind(my_ahp_y, pc_scores2))
summary(model_pca2)$r.squared
```

```
## [1] 0.6323622
```

The R square for data, not standardization is less than the data is standardization. For data with standardization, the first component only captures 50% of the feature, and for without standardization is captures almost 90% of the feature.