Homework 22

Ziyao Yang

4/21/23

In this problem we illustrate how PCA can help improve linear regression through dimension reduction. Here we use data from https://howlongtobeat.com/, a website that tracks how long it takes to complete different story-based video games. Please use the dataset video_game_lengths.csv.(on Canvas) In this data set, we have different video games with their review score (from 0 to 100) and how long it takes to complete the game. We consider games with two play styles ("main story" versus "complete") and three different measures ("leisure", "median", and "rushed") for players who complete the game slowly, normally, or quickly. This gives us 6 different measures of how long it takes to complete the video game.

```
library(dplyr)
```

```
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
```

```
df <- read.csv("video_game_lengths.csv")
```

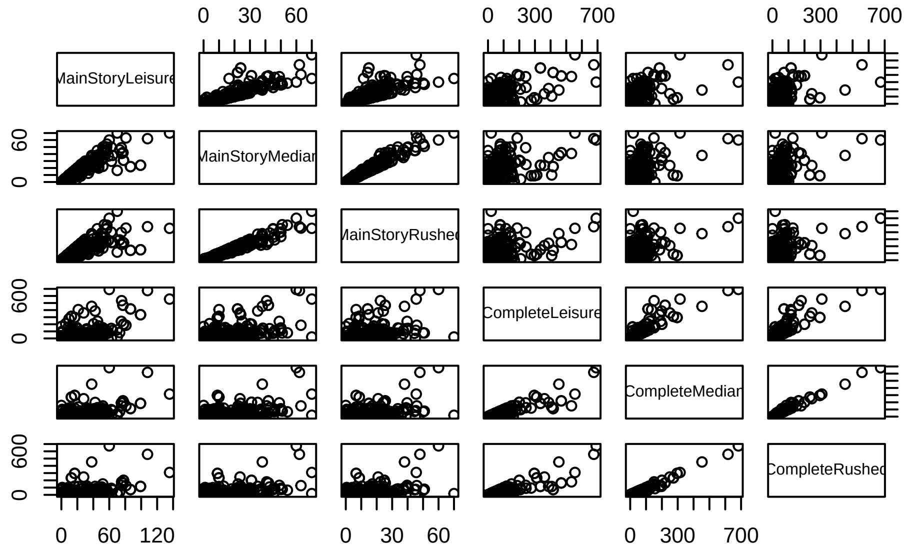
[C] Compute the pairwise correlations among the following 6 play length variables from the video_game_lengths.csv dataset. Read the documentation of pairs function and then use this function to create a matrix of pairwise scatter plots for the following 6 variables. Describe the patterns you observed in the pairwise correlations and scatter plots.

"MainStoryLeisure" "MainStoryMedian" "MainStoryRushed" "CompleteLeisure" "CompleteMedian" "CompleteRushed"

```r
# Compute pairwise correlations
correlations <- cor(df[,4:9])
print(correlations)
```

```
                 MainStoryLeisure MainStoryMedian MainStoryRushed
MainStoryLeisure        1.0000000       0.9128193       0.8525426
MainStoryMedian         0.9128193       1.0000000       0.9792983
MainStoryRushed         0.8525426       0.9792983       1.0000000
CompleteLeisure         0.6932078       0.5760248       0.5050873
CompleteMedian          0.6285759       0.5762545       0.5365953
CompleteRushed          0.5884774       0.5543003       0.5248587
                 CompleteLeisure CompleteMedian CompleteRushed
MainStoryLeisure       0.6932078      0.6285759      0.5884774
MainStoryMedian        0.5760248      0.5762545      0.5543003
MainStoryRushed        0.5050873      0.5365953      0.5248587
CompleteLeisure        1.0000000      0.9138510      0.8805134
CompleteMedian         0.9138510      1.0000000      0.9884127
CompleteRushed         0.8805134      0.9884127      1.0000000
```

```r
# Create a matrix of pairwise scatter plots
pairs(df[,4:9])
```



- Many variables has very strong correlations with other variables

[C] Use the lm function to build a linear regression model that can predict ReviewScore from the 6 play length variables listed above. Show the regression output table.

```
lm <- lm(ReviewScore ~ . -Title - Year, df)
summary(lm)
```

```
Call:
lm(formula = ReviewScore ~ . - Title - Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-46.896  -7.896   1.991   8.909  24.221

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      65.89570    0.46727 141.023  < 2e-16 ***
MainStoryLeisure  0.17705    0.07926   2.234  0.02567 *
MainStoryMedian   0.38566    0.27163   1.420  0.15593
MainStoryRushed  -0.38335    0.26245  -1.461  0.14437
CompleteLeisure   0.04719    0.01725   2.736  0.00631 **
CompleteMedian    0.05136    0.06675   0.769  0.44179
CompleteRushed   -0.10523    0.06374  -1.651  0.09904 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.19 on 1205 degrees of freedom
Multiple R-squared:  0.1199,    Adjusted R-squared:  0.1155
F-statistic: 27.36 on 6 and 1205 DF,  p-value: < 2.2e-16
```

[C] Use the svd function to perform PCA on the data matrix formed by the 6 play length variables listed above. Plot the proportion of total variation explained by PCs as a function of the number of PCs. Compute the proportion of total variation explained by the first two PCs.

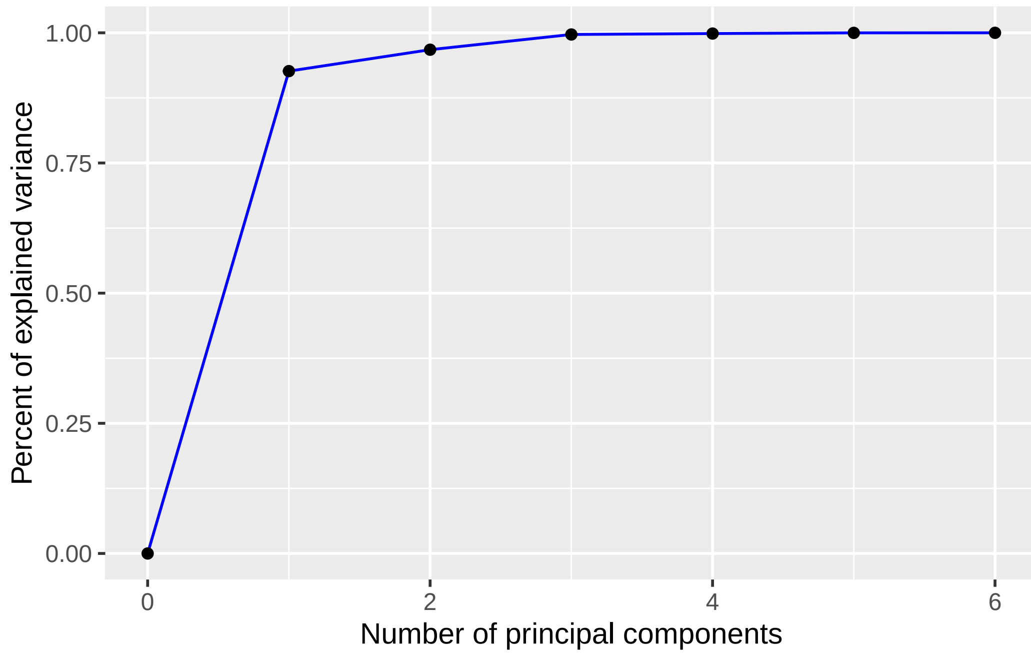
```
# MainStoryLeisure + MainStoryMedian + MainStoryRushed + CompleteLeisure + CompleteMedian
x_svd <- svd(df[,4:9])
x_svd$d
```

```
[1] 3217.81683  677.93421  571.44143  141.54207  122.05267   32.66728
```

```
sum(x_svd$d)
```

```
[1] 4763.454
```

```
library(ggplot2)
# plot reduction in variance curve
ggplot(
data=data.frame(
n_pc=0:6,
pcvar=c(0,
cumsum(x_svd$d^2) /
sum(x_svd$d^2))
),
mapping=aes(n_pc, pcvar)
) +
geom_line(color="blue", group=1) +
geom_point() +
xlab("Number of principal components") +
ylab("Percent of explained variance")
```




```
(x_svd$d[1] + x_svd$d[2])/sum(x_svd$d)
```

```
[1] 0.8178416
```

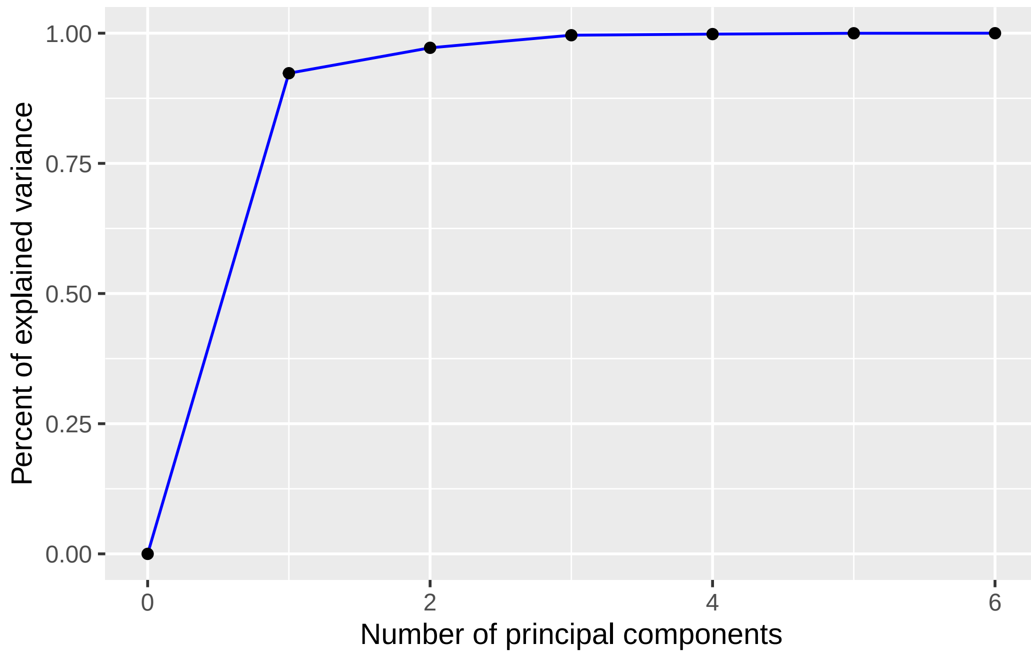
[C] Repeat the previous part by using the prcomp function. Compare the results with those from the previous part.

```
x_prcomp <-  prcomp(df[,4:9])
summary(x_prcomp)
```

```
Importance of components:
                           PC1      PC2      PC3     PC4     PC5     PC6
Standard deviation     84.3200 19.38517 13.65539 4.05940 3.49000 0.93832
Proportion of Variance  0.9232  0.04879  0.02421 0.00214 0.00158 0.00011
Cumulative Proportion   0.9232  0.97195  0.99616 0.99830 0.99989 1.00000
```

```
ggplot(
data=data.frame(
n_pc=0:6,
```

```r
pcvar=c(0,
cumsum(x_prcomp$sdev**2) /
sum(x_prcomp$sdev**2))
),
mapping=aes(n_pc, pcvar)
) +
geom_line(color="blue", group=1) +
geom_point() +
xlab("Number of principal components") +
ylab("Percent of explained variance")
```




```r
(x_prcomp$sdev[1] + x_prcomp$sdev[2]) / sum(x_prcomp$sdev)
```

```
[1] 0.8240492
```

[C] Use the lm function to build a linear regression model that can predict ReviewScore from the first two PCs. Show the regression output table.

```r
# add new columns to data.frame with principal component projections
df = cbind(
df,
as.matrix(df[, 4:9]) %*% as.matrix(x_prcomp$rotation)
)
# fit model using PC projections as predictors
summary(
lm(ReviewScore ~ PC1 + PC2,
data=df)
)
```

```
Call:
lm(formula = ReviewScore ~ PC1 + PC2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.196  -8.138   1.804   9.148  26.048 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 67.196298   0.389648 172.454  < 2e-16 ***
PC1         -0.038330   0.004208  -9.108  < 2e-16 ***
PC2          0.117524   0.018305   6.420 1.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.35 on 1209 degrees of freedom
Multiple R-squared:  0.09315,	Adjusted R-squared:  0.09165 
F-statistic: 62.09 on 2 and 1209 DF,  p-value: < 2.2e-16
```