Question 9.1

Using the same crime data set `uscrime.txt` as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Question 8.2. You can use the R function `prcomp` for PCA. (**Note** that to first scale the data, you can include `scale. = TRUE` to scale as part of the PCA function. Don't forget that, to make a prediction for the new city, you'll need to unscale the coefficients (i.e., do the scaling calculation in reverse)!)

Answer 9.1

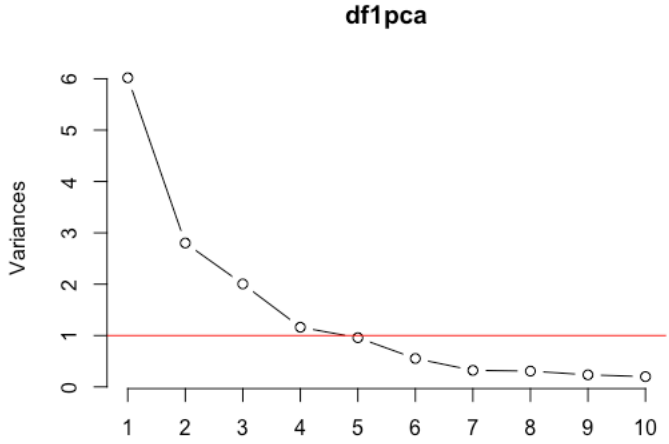
```
# Clear global environment, load and preview dataset
> rm(list=ls())
> df1 <- read.delim("/Users/…/uscrime.txt", header = T, stringsAsFactors = F)
> head(df1)
    M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
    Time Crime
1 26.2011   791
2 25.2999  1635
3 24.3006   578
4 29.9012  1969
5 21.2998  1234
6 20.9995   682

# Perform PCA on the entire dataset with scaled data. Preview the summary of components.
> df1pca <- prcomp(df1[,1:15], center = T, scale. = T)
> summary(df1pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377 0.56729
Proportion of Variance 0.4013 0.1868 0.1337 0.07748 0.06389 0.03688 0.02145
Cumulative Proportion  0.4013 0.5880 0.7217 0.79920 0.86308 0.89996 0.92142
                           PC8     PC9    PC10    PC11    PC12    PC13   PC14
Standard deviation     0.55444 0.48493 0.44708 0.41915 0.35804 0.26333 0.2418
Proportion of Variance 0.02049 0.01568 0.01333 0.01171 0.00855 0.00462 0.0039
Cumulative Proportion  0.94191 0.95759 0.97091 0.98263 0.99117 0.99579 0.9997
                          PC15
Standard deviation     0.06793
```

Proportion of Variance 0.00031
Cumulative Proportion 1.00000

Scree plot shows the variance explained by each principle components. Kaiser rule recommended picking PCs that can explain at least 80% of the variance. Therefore I'd pick the first five principle components as suggested in the following plot.

```
> screeplot(df1pca, type = "line")
> abline(h=1, col="red")
```



Append the first five principle components to the original dataset to perform the following regression model.

```
> df2 <- cbind(df1pca$x[,1:5], df1[,16])
> head(df2)
        PC1       PC2       PC3       PC4       PC5
[1,] -4.199284 -1.0938312 -1.11907395  0.67178115  0.05528338  791
[2,]  1.172663  0.6770136 -0.05244634 -0.08350709 -1.17319982 1635
[3,] -4.173725  0.2767750 -0.37107658  0.37793995  0.54134525  578
[4,]  3.834962 -2.5769060  0.22793998  0.38262331 -1.64474650 1969
[5,]  1.839300  1.3309856  1.27882805  0.71814305  0.04159032 1234
[6,]  2.907234 -0.3305421  0.53288181  1.22140635  1.37436096  682
```

Preform regression modeling using the first five principle components. The R^2 value with 0.6452 is higher than the result from the last assignment which is 6-factor cross validated model with R^2 value of 0.6031.

```
> model1 <- lm(df2[,6]~., data = as.data.frame(df2[,1:5]))
> summary(model1)

Call:
lm(formula = df2[, 6] ~ ., data = as.data.frame(df2[, 1:5]))
```

```
Residuals:
   Min     1Q  Median     3Q    Max
-420.79 -185.01   12.21  146.24  447.86

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   905.09      35.59  25.428  < 2e-16 ***
PC1            65.22      14.67   4.447 6.51e-05 ***
PC2           -70.08      21.49  -3.261  0.00224 **
PC3            25.19      25.41   0.992  0.32725
PC4            69.45      33.37   2.081  0.04374 *
PC5          -229.04      36.75  -6.232 2.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 244 on 41 degrees of freedom
Multiple R-squared:  0.6452,  Adjusted R-squared:  0.6019
F-statistic: 14.91 on 5 and 41 DF,  p-value: 2.446e-08
```

6-factor cross validated model from the last assignment has a predicted value of 1304. The new prediction model for the criteria given in the last assignment has a new predicted value of 1389.

```
> pred_df <- data.frame(M = 14.0,So = 0,Ed = 10.0,Po1 = 12.0,Po2 = 15.5,LF = 0.640,M.F = 94.0,Pop = 150,NW = 1.1,U1 = 0.120,U2 = 3.6,Wealth = 3200,Ineq = 20.1,Prob = 0.04,Time = 39.0)
> df3 <- data.frame(predict(df1pca, pred_df))
> result <- predict(model1, df3)
> result
       1
1388.926
```