Question 1.

Step 1 --- Read in the Dataset

The first step is to setup the environment and read in the uscrime.txt dataset:

```
# Clear the environment
rm(list = ls())

# Comment in set.seed(33) to repeat results
set.seed(33)

# Load crime data into a data frame
data_df <- read.table("uscrime.txt", header=TRUE)
```

Step 2 --- Perform PCA on the Dataset

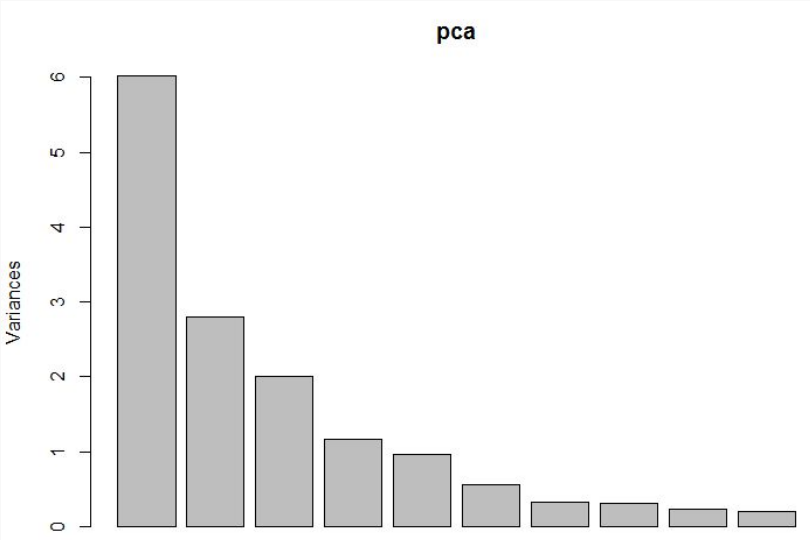
I use prcomp() to perform PCA on the uscrime dataset:

```
# Perform PCA on data_df
pca <- prcomp(data_df[,1:15], scale = TRUE)
summary(pca)

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
Standard deviation     0.55444 0.48493 0.44708 0.41915 0.35804 0.26333
Proportion of Variance 0.02049 0.01568 0.01333 0.01171 0.00855 0.00462
Cumulative Proportion  0.94191 0.95759 0.97091 0.98263 0.99117 0.99579

                         PC14    PC15
Standard deviation     0.2418 0.06793
Proportion of Variance 0.0039 0.00031
Cumulative Proportion  0.9997 1.00000
```

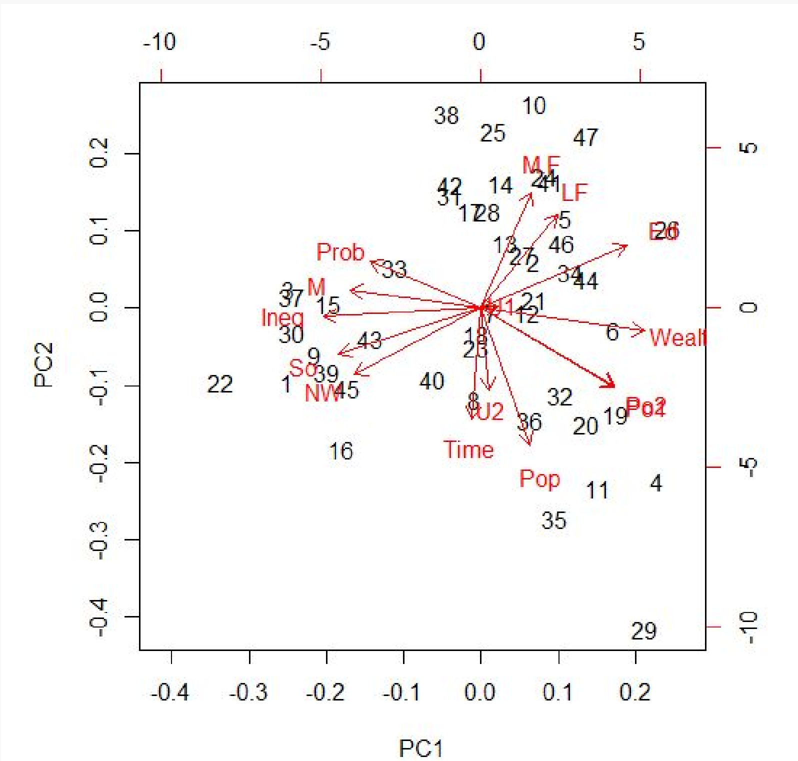
Step 3 --- Visualize PCA Results

I visualized the prcomp() results:

```
# Visualize PCA results
plot(pca)
```




```
biplot(pca)
```



Step 4 --- Extract First Four Principal Components

After creating the PCA, I extract the first four principal components and create a new dataset to use for the linear regression model:

```
# Extract the 1st four components and append Crime
pca_df <- data.frame(cbind(pca$x[,1:4],data_df$Crime))
names(pca_df) <- c('PC1','PC2','PC3','PC4','Crime')
```

Step 5 --- Build Model using Principal Components

Having created a new pca dataset, I create a model using lm():

```
# Create lm model using pca dataset
model_pca <- lm(Crime ~., pca_df)

# Display summary
summary(model_pca)

Call:
lm(formula = Crime ~ ., data = pca_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-557.76 -210.91  -29.08  197.26  810.35 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   905.09      49.07  18.443  < 2e-16 ***
PC1            65.22      20.22   3.225  0.00244 ** 
PC2           -70.08      29.63  -2.365  0.02273 *  
PC3            25.19      35.03   0.719  0.47602    
PC4            69.45      46.01   1.509  0.13872    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 336.4 on 42 degrees of freedom
Multiple R-squared:  0.3091,	Adjusted R-squared:  0.2433 
F-statistic: 4.698 on 4 and 42 DF,  p-value: 0.003178
```

<u>Step 6 --- Transform PC Coefficients Back to Original Factors:</u>

I converted the model_pca coefficients back to the original factors:

```
# Convert model_pca coefficients to original factors
coefficients_converted <- (pca$rotation[,1:4] %*%
model_pca$coefficients[2:5])/pca$scale

# Adjust intercept based on pca$center
intercept <- model_pca$coefficients[1] - sum(coefficients_converted *
pca$center)
```

<u>Step 7 --- Predict Crime for Data Point</u>

With the model back into the original factors, I predict Crime for the data point:

```
# New data point that we'll predict Crime for
new_dp <- 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)

# Manually calculate Crime for new_dp using coefficients_coverted and
intercept
### Would like to refactor into more elegant method
Crime <- sum(
  coefficients_converted[1,1] %*% new_dp$M,
  coefficients_converted[2,1] %*% new_dp$So,
  coefficients_converted[3,1] %*% new_dp$Ed,
  coefficients_converted[4,1] %*% new_dp$Po1,
  coefficients_converted[5,1] %*% new_dp$Po2,
  coefficients_converted[6,1] %*% new_dp$LF,
  coefficients_converted[7,1] %*% new_dp$M.F,
  coefficients_converted[8,1] %*% new_dp$Pop,
  coefficients_converted[9,1] %*% new_dp$NW,
  coefficients_converted[10,1] %*% new_dp$U1,
  coefficients_converted[11,1] %*% new_dp$U2,
  coefficients_converted[12,1] %*% new_dp$Wealth,
  coefficients_converted[13,1] %*% new_dp$Ineq,
  coefficients_converted[14,1] %*% new_dp$Prob,
  coefficients_converted[15,1] %*% new_dp$Time,
  intercept
  )

Crime # 1112.678
```

Step 8 -- Compare Results to HW5 Q2:

My model from HW5 Q2 was:

```
Crime ~ -5040.50 + 105.02*M + 196.47*Ed + 115.02*Po1 + 89.37*U2 +
67.65*Ineq - 3801.84*Prob with an Adj. R² of 0.7307
```

The PCA based model is:

```
Crime ~ 1666 - 16.93*M + 21.34*So + 12.83*Ed + 21.35*Po1 + 23.09*Po2 -
346.57*LF - 8.29*M.F + 1.046*Pop + 1.5*NW - 1509.93*U1 + 1.69*U2 +
0.04*Wealth - 6.90*Ineq + 144.95*Prob - 0.93*Time with an Adj. R² of 0.2433
```

HW5-Q2 model predicted Crime of 1,304 for the new_dp; whereas, the PCA model predicts Crime of ~1,113. According to Adj. R^2, the new pca-based model performed much worse than the model specified in HW5 Q2.

APPENDIX --- Full R Script

```
# Clear the environment
rm(list = ls())

# Comment in set.seed(33) to repeat results
set.seed(33)

# Load crime data into a data frame
data_df <- read.table("uscrime.txt", header=TRUE)

# Perform PCA on data_df
pca <- prcomp(data_df[,1:15], scale = TRUE)
summary(pca)

# Visualize PCA results
plot(pca)
biplot(pca)

# Extract the 1st four components and append Crime
pca_df <- data.frame(cbind(pca$x[,1:4],data_df$Crime))
names(pca_df) <- c('PC1','PC2','PC3','PC4','Crime')

# Create lm model using pca dataset
model_pca <- lm(Crime ~., pca_df)

# Display summary
summary(model_pca)
```

```r
# Convert model_pca coefficients to original factors
coefficients_converted <- (pca$rotation[,1:4] %*%
model_pca$coefficients[2:5])/pca$scale

# Adjust intercept based on pca$center
intercept <- model_pca$coefficients[1] - sum(coefficients_converted *
pca$center)

# New data point that we'll predict Crime for
new_dp <- 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)

# Manually calculate Crime for new_dp using coefficients_coverted and
intercept
### Would like to refactor into more elegant method
Crime <- sum(
  coefficients_converted[1,1] %*% new_dp$M,
  coefficients_converted[2,1] %*% new_dp$So,
  coefficients_converted[3,1] %*% new_dp$Ed,
  coefficients_converted[4,1] %*% new_dp$Po1,
  coefficients_converted[5,1] %*% new_dp$Po2,
  coefficients_converted[6,1] %*% new_dp$LF,
  coefficients_converted[7,1] %*% new_dp$M.F,
  coefficients_converted[8,1] %*% new_dp$Pop,
  coefficients_converted[9,1] %*% new_dp$NW,
  coefficients_converted[10,1] %*% new_dp$U1,
  coefficients_converted[11,1] %*% new_dp$U2,
  coefficients_converted[12,1] %*% new_dp$Wealth,
  coefficients_converted[13,1] %*% new_dp$Ineq,
  coefficients_converted[14,1] %*% new_dp$Prob,
  coefficients_converted[15,1] %*% new_dp$Time,
  intercept
  )

Crime # 1112.678
```