PCA – US Crimes Data

```
#setwd("C:/Users/beena/Downloads/Analytics Modelling")

uscrime<-read.table("uscrime.txt",header=TRUE)
View(uscrime)
```

Applying PCA to US Crimes dataset

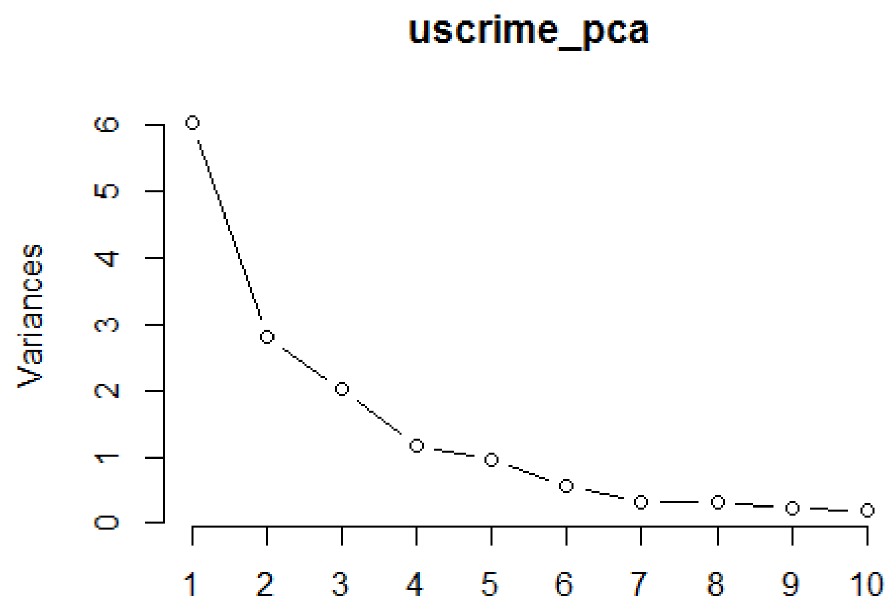
```
uscrime_pca<- prcomp(uscrime[1:15],scale = TRUE)
#### Viewing the First 4 principal components
head(uscrime_pca$x[,1:4])

##             PC1        PC2         PC3         PC4
## [1,] -4.199284 -1.0938312 -1.11907395  0.67178115
## [2,]  1.172663  0.6770136 -0.05244634 -0.08350709
## [3,] -4.173725  0.2767750 -0.37107658  0.37793995
## [4,]  3.834962 -2.5769060  0.22793998  0.38262331
## [5,]  1.839300  1.3309856  1.27882805  0.71814305
## [6,]  2.907234 -0.3305421  0.53288181  1.22140635

summary(uscrime_pca)

## Importance of components%s:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377
## Proportion of Variance 0.4013 0.1868 0.1337 0.07748 0.06389 0.03688
## Cumulative Proportion  0.4013 0.5880 0.7217 0.79920 0.86308 0.89996
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.56729 0.55444 0.48493 0.44708 0.41915 0.35804
## Proportion of Variance 0.02145 0.02049 0.01568 0.01333 0.01171 0.00855
## Cumulative Proportion  0.92142 0.94191 0.95759 0.97091 0.98263 0.99117
##                           PC13   PC14    PC15
## Standard deviation     0.26333 0.2418 0.06793
## Proportion of Variance 0.00462 0.0039 0.00031
## Cumulative Proportion  0.99579 0.9997 1.00000

plot(uscrime_pca,type="line")
```



From the summary and plot we can see that the first 4 principle componensts explain most of the variability of the data, hence we use the first 4 components to create our model

```
### Extracting the first 4 principle components
PCA_uscrimes<-as.data.frame(cbind(uscrime_pca$x[,1],uscrime_pca$x[,2],uscrime_pca$x[,3],uscrime_pca$x[,4],uscrime$Crime))
#### linear regression model with principal components
model_pca<-lm(V5~.,data=PCA_uscrimes )
summary(model_pca)

##
## Call:
## lm(formula = V5 ~ ., data = PCA_uscrimes)
##
## 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 ***
## V1             65.22      20.22   3.225  0.00244 **
## V2            -70.08      29.63  -2.365  0.02273 *
## V3             25.19      35.03   0.719  0.47602
## V4             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
```

Now we calculate the coefficients in terms of the original variables in our model using the eigen vectors from PCA. We iterate across the top 4 PCA coefficients and multiply the beta coefficients obtained from the linear regression model with each of the top 4 PCA values.

```
transformed_coeff<-c()
for(x in 1:(ncol(uscrime)-1)) {
  iter <- 0

 for (i in 1:4) {

    iter <- iter + model_pca$coefficients[i+1]*uscrime_pca$rotation[x,i]

  }

  transformed_coeff <- rbind(transformed_coeff, c(colnames(uscrime)[x],iter))
}
transformed_coeff <- as.data.frame(transformed_coeff)
colnames(transformed_coeff) <- c("Variable", "PCA Coefficient")
print(as.data.frame(transformed_coeff))
```

```
##    Variable   PCA Coefficient
## 1         M -21.2779630823314
## 2        So  10.2230912160043
## 3        Ed  14.3526100868343
## 4       Po1  63.4564258306081
## 5       Po2  64.5579741936575
## 6        LF -14.0053491046701
## 7       M.F -24.4375717582785
## 8       Pop   39.830667209046
## 9        NW  15.4345453322952
## 10       U1 -27.2222812613964
## 11       U2  1.42590219642975
## 12   Wealth  38.6078553183368
## 13     Ineq -27.5363479781423
## 14     Prob  3.29570747307768
## 15     Time -6.61261565979637
```

Explanation:

We get a R^2 value of 30.91% from the multiple regression model using 4 Principle components whereas from the previous question we get R^2 value of 76.59%. We see that the PCA model with 4 components does not explain as much of the variability hence R^2 is low as compared to the Multiple Linear regression model of the previous question. If we use all the Principal Components for our Multiple Linear Regression Model, we get a much higher R^2 value as compared to Linear Regression without using PCA. Basically not all of the variation is explained by just 4 Principal components hence we get a low R^2 value.