# ISYE6502-Homework6

```
UScrime_data = read.table(file="~/Desktop/ISYE6501-\ Introduction\ to\ Analytics\ Mod
eling/Fall2020hw6/uscrime.txt", header = TRUE,stringsAsFactors = FALSE)
str(UScrime_data)
```

```
## 'data.frame':    47 obs. of  16 variables:
##  $ M     : num  15.1 14.3 14.2 13.6 14.1 12.1 12.7 13.1 15.7 14 ...
##  $ So    : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed    : num  9.1 11.3 8.9 12.1 12.1 11 11.1 10.9 9 11.8 ...
##  $ Po1   : num  5.8 10.3 4.5 14.9 10.9 11.8 8.2 11.5 6.5 7.1 ...
##  $ Po2   : num  5.6 9.5 4.4 14.1 10.1 11.5 7.9 10.9 6.2 6.8 ...
##  $ LF    : num  0.51 0.583 0.533 0.577 0.591 0.547 0.519 0.542 0.553 0.632 ...
##  $ M.F   : num  95 101.2 96.9 99.4 98.5 ...
##  $ Pop   : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW    : num  30.1 10.2 21.9 8 3 4.4 13.9 17.9 28.6 1.5 ...
##  $ U1    : num  0.108 0.096 0.094 0.102 0.091 0.084 0.097 0.079 0.081 0.1 ...
##  $ U2    : num  4.1 3.6 3.3 3.9 2 2.9 3.8 3.5 2.8 2.4 ...
##  $ Wealth: int  3940 5570 3180 6730 5780 6890 6200 4720 4210 5260 ...
##  $ Ineq  : num  26.1 19.4 25 16.7 17.4 12.6 16.8 20.6 23.9 17.4 ...
##  $ Prob  : num  0.0846 0.0296 0.0834 0.0158 0.0414 ...
##  $ Time  : num  26.2 25.3 24.3 29.9 21.3 ...
##  $ Crime : int  791 1635 578 1969 1234 682 963 1555 856 705 ...
```

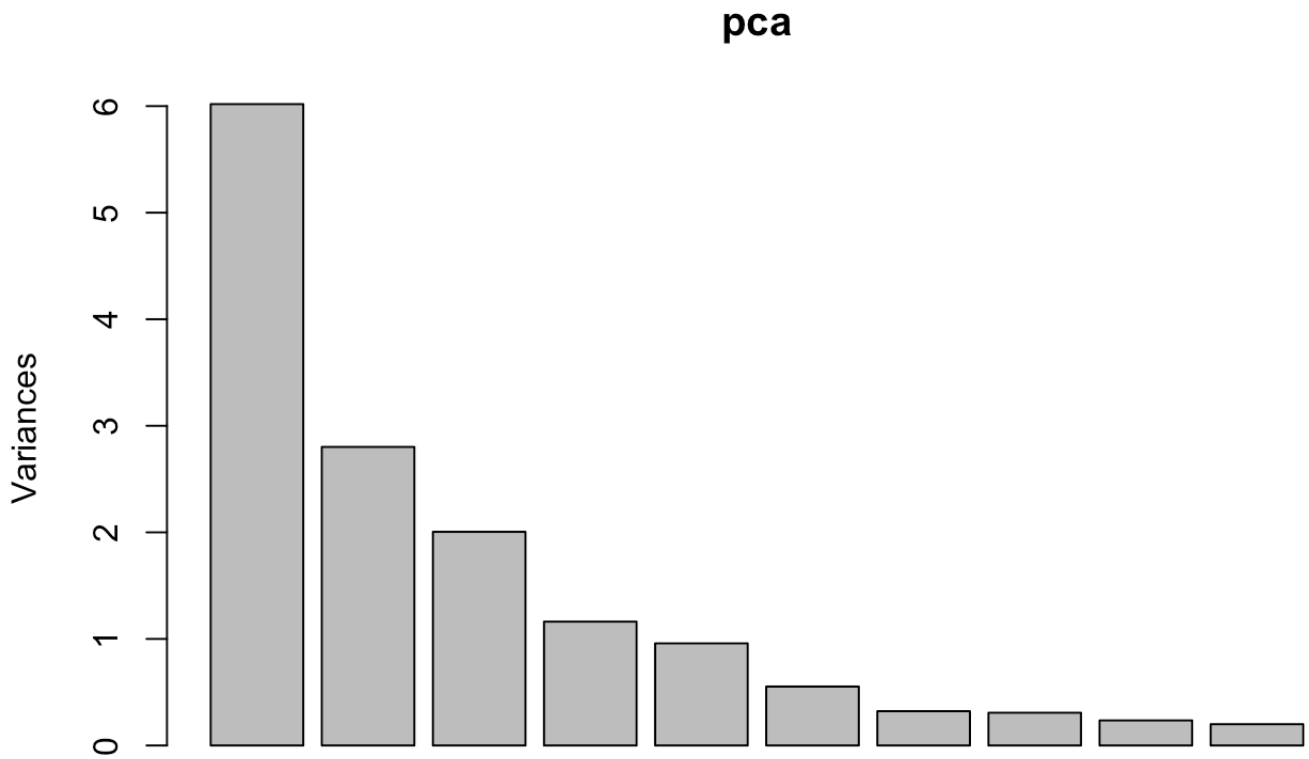
```
pca<-prcomp(UScrime_data[,-16], center = TRUE, scale=TRUE)
summary(pca)
```

```
## 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   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
```

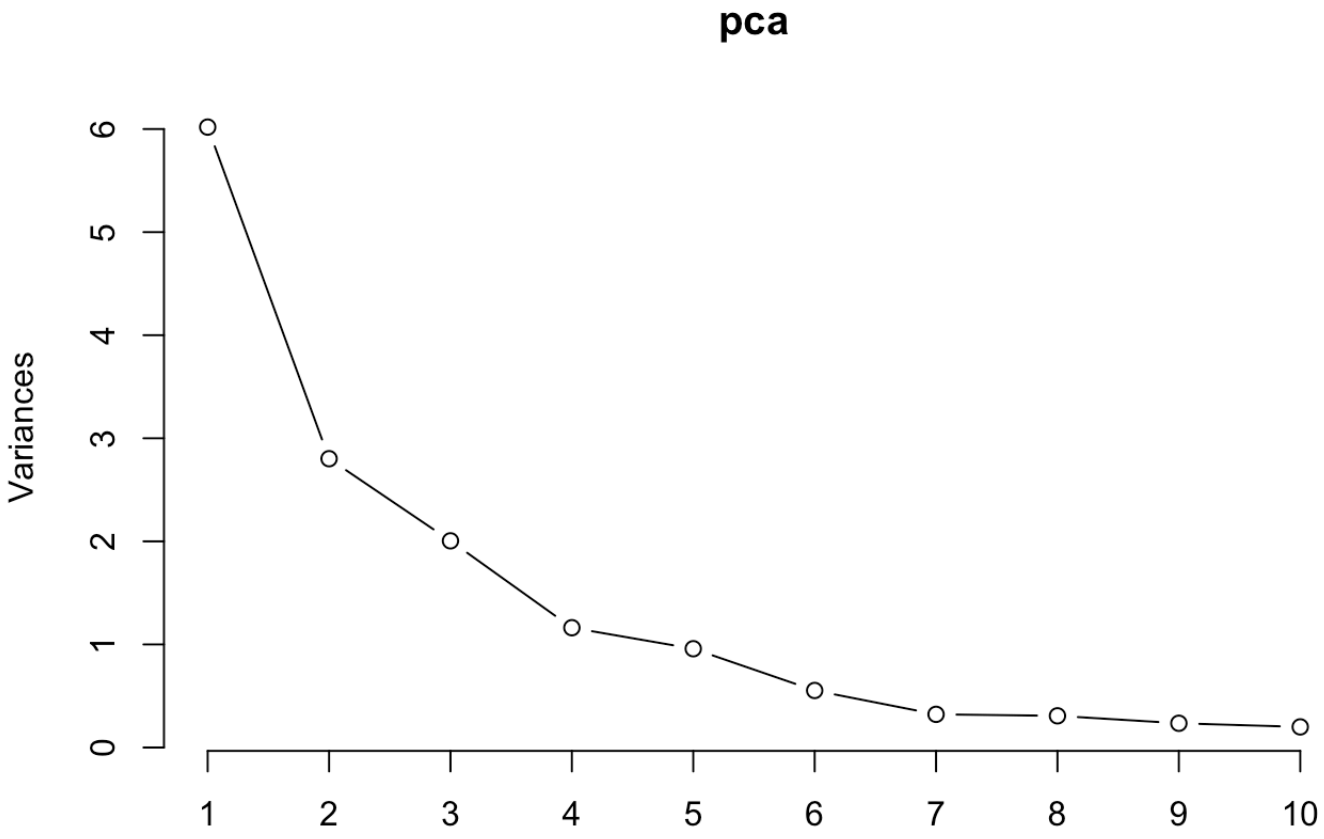
```
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

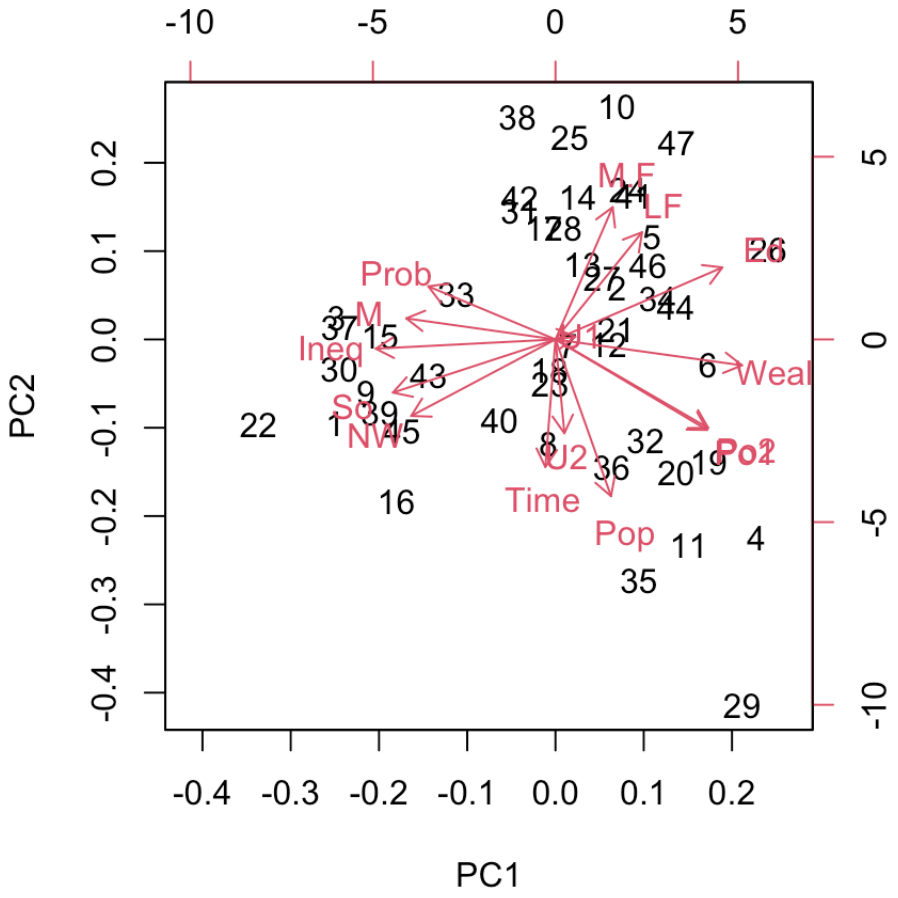
```
plot(pca)
```




```
plot(pca,type ="l")
```




```
biplot(pca)
```




```
#Because each eigenvalue is roughly the importance of its corresponding eigenvector, the proportion of variance explained is the sum of the eigenvalues of the features yo u kept divided by the sum of the eigenvalues of all features.
range(UScrime_data$Crime)
```

```
## [1]  342 1993
```

```
pca$rotation
```

```
##                PC1         PC2           PC3         PC4         PC5
## M      -0.30371194  0.06280357  0.1724199946 -0.02035537 -0.35832737
## So     -0.33088129 -0.15837219  0.0155433104  0.29247181 -0.12061130
## Ed      0.33962148  0.21461152  0.0677396249  0.07974375 -0.02442839
## Po1     0.30863412 -0.26981761  0.0506458161  0.33325059 -0.23527680
## Po2     0.31099285 -0.26396300  0.0530651173  0.35192809 -0.20473383
## LF      0.17617757  0.31943042  0.2715301768 -0.14326529 -0.39407588
## M.F     0.11638221  0.39434428 -0.2031621598  0.01048029 -0.57877443
```

```
## Pop     0.11307836 -0.46723456  0.0770210971 -0.03210513 -0.08317034
## NW     -0.29358647 -0.22801119  0.0788156621  0.23925971 -0.36079387
## U1      0.04050137  0.00807439 -0.6590290980 -0.18279096 -0.13136873
## U2      0.01812228 -0.27971336 -0.5785006293 -0.06889312 -0.13499487
## Wealth  0.37970331 -0.07718862  0.0100647664  0.11781752  0.01167683
## Ineq   -0.36579778 -0.02752240 -0.0002944563 -0.08066612 -0.21672823
## Prob   -0.25888661  0.15831708 -0.1176726436  0.49303389  0.16562829
## Time   -0.02062867 -0.38014836  0.2235664632 -0.54059002 -0.14764767
##                PC6         PC7         PC8         PC9        PC10        PC11
## M      -0.449132706 -0.15707378 -0.55367691  0.15474793 -0.01443093  0.39446657
## So     -0.100500743  0.19649727  0.22734157 -0.65599872  0.06141452  0.23397868
## Ed     -0.008571367 -0.23943629 -0.14644678 -0.44326978  0.51887452 -0.11821954
## Po1    -0.095776709  0.08011735  0.04613156  0.19425472 -0.14320978 -0.13042001
## Po2    -0.119524780  0.09518288  0.03168720  0.19512072 -0.05929780 -0.13885912
## LF      0.504234275 -0.15931612  0.25513777  0.14393498  0.03077073  0.38532827
## M.F    -0.074501901  0.15548197 -0.05507254 -0.24378252 -0.35323357 -0.28029732
## Pop     0.547098563  0.09046187 -0.59078221 -0.20244830 -0.03970718  0.05849643
## NW      0.051219538 -0.31154195  0.20432828  0.18984178  0.49201966 -0.20695666
## U1      0.017385981 -0.17354115 -0.20206312  0.02069349  0.22765278 -0.17857891
## U2      0.048155286 -0.07526787  0.24369650  0.05576010 -0.04750100  0.47021842
## Wealth -0.154683104 -0.14859424  0.08630649 -0.23196695 -0.11219383  0.31955631
## Ineq    0.272027031  0.37483032  0.07184018 -0.02494384 -0.01390576 -0.18278697
## Prob    0.283535996 -0.56159383 -0.08598908 -0.05306898 -0.42530006 -0.08978385
## Time   -0.148203050 -0.44199877  0.19507812 -0.23551363 -0.29264326 -0.26363121
##               PC12        PC13        PC14          PC15
## M       0.16580189 -0.05142365  0.04901705  0.0051398012
## So     -0.05753357 -0.29368483 -0.29364512  0.0084369230
## Ed      0.47786536  0.19441949  0.03964277 -0.0280052040
## Po1     0.22611207 -0.18592255 -0.09490151 -0.6894155129
## Po2     0.19088461 -0.13454940 -0.08259642  0.7200270100
## LF      0.02705134 -0.27742957 -0.15385625  0.0336823193
## M.F    -0.23925913  0.31624667 -0.04125321  0.0097922075
## Pop    -0.18350385  0.12651689 -0.05326383  0.0001496323
## NW     -0.36671707  0.22901695  0.13227774 -0.0370783671
## U1     -0.09314897 -0.59039450 -0.02335942  0.0111359325
## U2      0.28440496  0.43292853 -0.03985736  0.0073618948
## Wealth -0.32172821 -0.14077972  0.70031840 -0.0025685109
## Ineq    0.43762828 -0.12181090  0.59279037  0.0177570357
## Prob    0.15567100 -0.03547596  0.04761011  0.0293376260
## Time    0.13536989 -0.05738113 -0.04488401  0.0376754405
```

```
cor(pca$x) #are all orthogonal to each other,
```

```
##                PC1           PC2           PC3           PC4           PC5
## PC1   1.000000e+00 -1.273307e-16 -1.825724e-16  2.298165e-16 -3.391074e-16
## PC2  -1.273307e-16  1.000000e+00 -5.694249e-16  3.269637e-16 -8.335299e-16
```

```
## PC3  -1.825724e-16 -5.694249e-16  1.000000e+00  1.177395e-16 -1.906912e-16
## PC4   2.298165e-16  3.269637e-16  1.177395e-16  1.000000e+00 -9.226708e-17
## PC5  -3.391074e-16 -8.335299e-16 -1.906912e-16 -9.226708e-17  1.000000e+00
## PC6  -1.459722e-16  4.219478e-16 -7.520921e-16  1.547542e-16  6.022076e-17
## PC7   3.976873e-16  1.540007e-16  2.035710e-16 -5.123996e-16  1.854410e-16
## PC8   6.388541e-16 -6.173812e-17 -7.165046e-17 -1.070185e-15 -6.433073e-16
## PC9   2.470077e-16 -3.807073e-16 -5.893441e-17  4.569382e-16  5.766527e-16
## PC10 -8.449048e-17 -4.552839e-16 -1.456269e-16  3.273781e-16  1.209745e-16
## PC11  1.213205e-16  2.045710e-17  8.169971e-18 -8.690871e-17  1.034889e-15
## PC12  1.662919e-16  1.097279e-16 -5.546615e-16 -5.863430e-16  1.159214e-15
## PC13  1.070330e-16 -8.302804e-16  9.079977e-16  4.193459e-16  2.700256e-16
## PC14  9.443813e-16 -6.262505e-16 -5.086062e-16  1.699532e-16 -1.210316e-17
## PC15  3.677245e-15  3.390845e-15 -3.874069e-15  2.292428e-15  3.579062e-17
##                PC6           PC7           PC8           PC9           PC10
## PC1  -1.459722e-16  3.976873e-16  6.388541e-16  2.470077e-16 -8.449048e-17
## PC2   4.219478e-16  1.540007e-16 -6.173812e-17 -3.807073e-16 -4.552839e-16
## PC3  -7.520921e-16  2.035710e-16 -7.165046e-17 -5.893441e-17 -1.456269e-16
## PC4   1.547542e-16 -5.123996e-16 -1.070185e-15  4.569382e-16  3.273781e-16
## PC5   6.022076e-17  1.854410e-16 -6.433073e-16  5.766527e-16  1.209745e-16
## PC6   1.000000e+00 -2.663864e-16 -1.213255e-16  6.943245e-16  2.552376e-16
## PC7  -2.663864e-16  1.000000e+00  1.364129e-15 -6.240791e-16 -5.487255e-16
## PC8  -1.213255e-16  1.364129e-15  1.000000e+00  3.245495e-16 -1.844524e-16
## PC9   6.943245e-16 -6.240791e-16  3.245495e-16  1.000000e+00 -1.337589e-15
## PC10  2.552376e-16 -5.487255e-16 -1.844524e-16 -1.337589e-15  1.000000e+00
## PC11 -1.090780e-16  1.020271e-16 -7.028380e-16  4.432449e-16  2.883589e-16
## PC12 -2.098893e-17  3.723676e-16  4.344960e-17 -2.141621e-16 -4.547243e-16
## PC13 -8.364523e-16 -2.010077e-16 -2.310523e-16 -2.007507e-16  4.375205e-16
## PC14  3.280329e-16 -1.651300e-16 -1.885520e-16  8.629211e-16  1.261895e-16
## PC15 -2.651521e-15 -5.196469e-16 -1.627361e-16  3.828687e-15  1.382630e-16
##                PC11          PC12          PC13          PC14          PC15
## PC1   1.213205e-16  1.662919e-16  1.070330e-16  9.443813e-16  3.677245e-15
## PC2   2.045710e-17  1.097279e-16 -8.302804e-16 -6.262505e-16  3.390845e-15
## PC3   8.169971e-18 -5.546615e-16  9.079977e-16 -5.086062e-16 -3.874069e-15
## PC4  -8.690871e-17 -5.863430e-16  4.193459e-16  1.699532e-16  2.292428e-15
## PC5   1.034889e-15  1.159214e-15  2.700256e-16 -1.210316e-17  3.579062e-17
## PC6  -1.090780e-16 -2.098893e-17 -8.364523e-16  3.280329e-16 -2.651521e-15
## PC7   1.020271e-16  3.723676e-16 -2.010077e-16 -1.651300e-16 -5.196469e-16
## PC8  -7.028380e-16  4.344960e-17 -2.310523e-16 -1.885520e-16 -1.627361e-16
## PC9   4.432449e-16 -2.141621e-16 -2.007507e-16  8.629211e-16  3.828687e-15
## PC10  2.883589e-16 -4.547243e-16  4.375205e-16  1.261895e-16  1.382630e-16
## PC11  1.000000e+00  1.555555e-16  3.969289e-16 -6.800922e-16  3.893464e-16
## PC12  1.555555e-16  1.000000e+00  1.184215e-16 -1.287411e-16 -3.548408e-16
## PC13  3.969289e-16  1.184215e-16  1.000000e+00  4.443130e-16 -2.885221e-15
## PC14 -6.800922e-16 -1.287411e-16  4.443130e-16  1.000000e+00 -3.562487e-16
## PC15  3.893464e-16 -3.548408e-16 -2.885221e-15 -3.562487e-16  1.000000e+00
```

```
crime.pca<-cbind(UScrime_data[,16],data.frame(pca$x[,1:5]))
colnames(crime.pca)[1] <- "CrimePCA"
cor(crime.pca)[,1]
```

```
##    CrimePCA         PC1         PC2         PC3         PC4         PC5
##  1.00000000  0.41368481 -0.30331302  0.09223697  0.19357298 -0.57972578
```

```
#regression model on first 5 PCs
model.reg <-lm(CrimePCA~., data = crime.pca)
summary(model.reg)
```

```
##
## Call:
## lm(formula = CrimePCA ~ ., data = crime.pca)
##
## 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
```

```
#rotation converts from  PC original value to, Unscaling
mu <- sapply(UScrime_data[,1:15],mean)
nComp = 15
Xhat = pca$x[,1:nComp] %*% t(pca$rotation[,1:nComp]) #Matrix multiplication
Xhat = scale(Xhat, center = -mu, scale = FALSE)
new.city <- 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 = 0.040,Time = 39.0)
pred.data <- data.frame(predict(pca, new.city))
pred <- predict(model.reg, pred.data)
print(pred)
```

```
##        1
## 1388.926
```

# Analysis

We perform PCA on US crime data with scaling the variables to have standard deviation. The center and scale components correspond to the mean and std dev of the variables. the rotation matrix provide the principal component loadings. We see that there is 15 distinct principal components. the first principal component explains 40% of the variance in the data, the next explains 18% of the variance, and so forth. the plot explained by each components and the variance. However looking at he scree plot, we see that the first five principal components where there is an elbow. This helps and suggests that there may be little benefit to examine through these 5 Principal components; which we used them to model linear regression which R-squared = 0.6452 and after unscaling the data, the predicted value is 1388.9. Comparing to last week HW, R-squared is 0.671 and prediction is 1304. In conclusion, we can say that PCA helped the deliver approximate accuracy with less number of predictors