Assignment 3

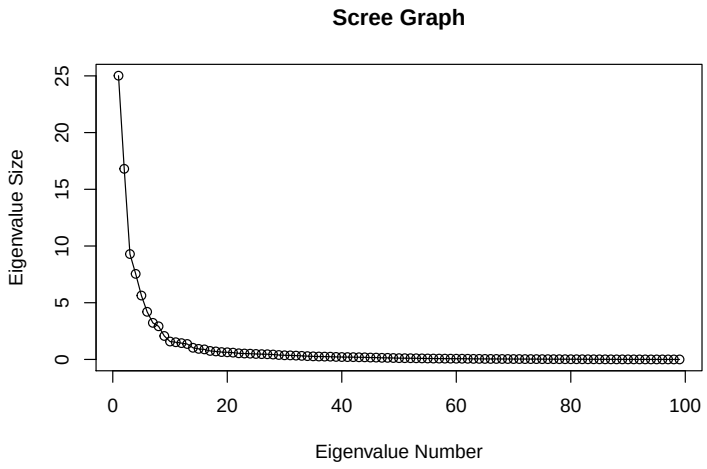
Alejo Pérez Gómez

Assignment 3

1.

First we will scale all the variables and implement the PCA. We will not include the variable State as it is categorical data and we will only apply scaling and PCA on continuous variables.

```
## Load data and convert to numeric
data_com <- sapply(data_com, as.numeric)

#Scaling features excluding state and target
X <- scale(data_com[,c(-1, -101)])
cov_mat <- cov(X)

cov_mat_eigen <- eigen(cov_mat)

#proportion of variance explained for each component
proportion_variation <- cov_mat_eigen$values/sum(cov_mat_eigen$values)

plot(cov_mat_eigen$values, xlab = 'Eigenvalue Number', ylab = 'Eigenvalue Size',
     main = 'Scree Graph')
lines(cov_mat_eigen$values)
```



The percentage of the variance accounted by the first two principal components is 42.2%. We calculated that out of summing the two first terms in our vector of variance explanations, dividing eigen values by total variance.

To achieve an explanation of the variance of 0.955%, 35 PC are needed.

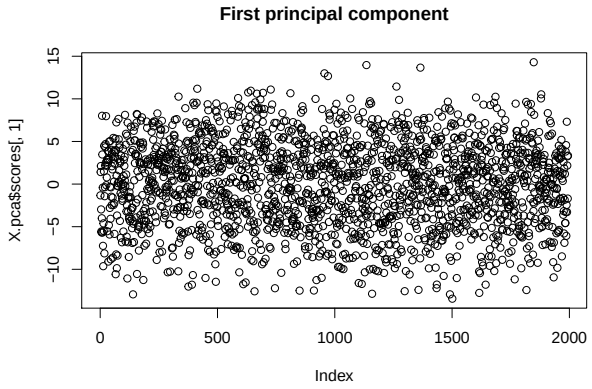
2.

In this section PCA will be used but with function `princomp()`

```
X.pca <- princomp(X)
```

Now the plot for the first principal component will be shown.

```
plot(X.pca$scores[,1], main="First principal component" )
```



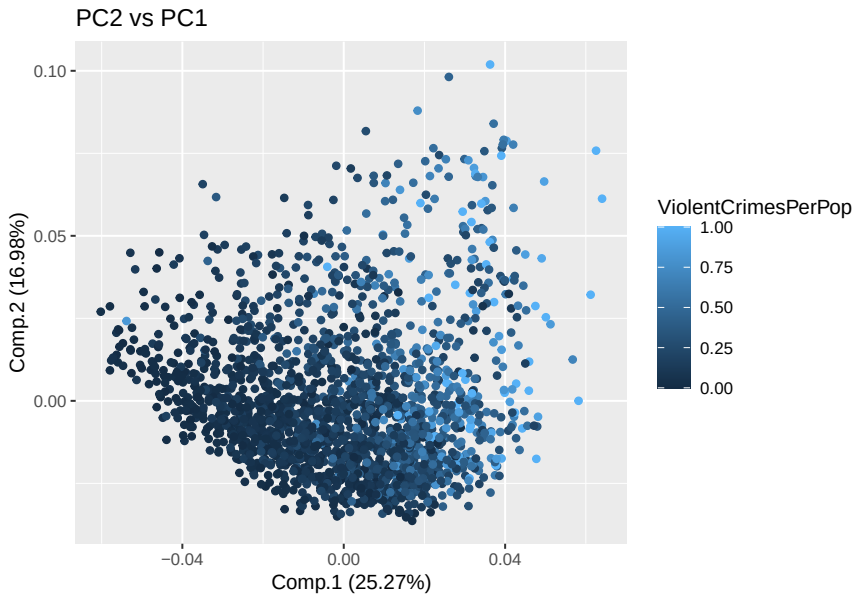
There are a total of 13 features that contribute notably to the first PC (using a threshold of 0.15) are the following (in absolute value).

	Value of contribution
medFamInc	0.1832453
medIncome	0.1819115
PctKids2Par	0.1755956
pctWInvInc	0.1749060
PctPopUnderPov	0.1738039
PctFam2Par	0.1727358
PctYoungKids2Par	0.1716485
perCapInc	0.1694056
pctWPubAsst	0.1647161
PctHousNoPhone	0.1640624
PctNotHSGrad	0.1619471
PctUnemployed	0.1587331
PctTeen2Par	0.1515750

The 5 most contributing features in regards to the first PC are **median family income, median household income, percentage of kids in family housing with two parents, percentage of households with investment and percentage of people under the poverty level**. These features can be arguably related to the crime level in communities. We could assume a correlation between these variables and criminality. It can be likely then, that an individual that happen to commit a crime registers poor values in terms of

some of the aforementioned features disregarding a cause-effect approach. In addition, that could lead us to think that those variables can connected when relating them to criminality.

Here is presented a plot of 2 fist PCA colored according the violent crimes per population. As can be seen, we could separate observations in clusters if we establish a threshold for the criminality rate.

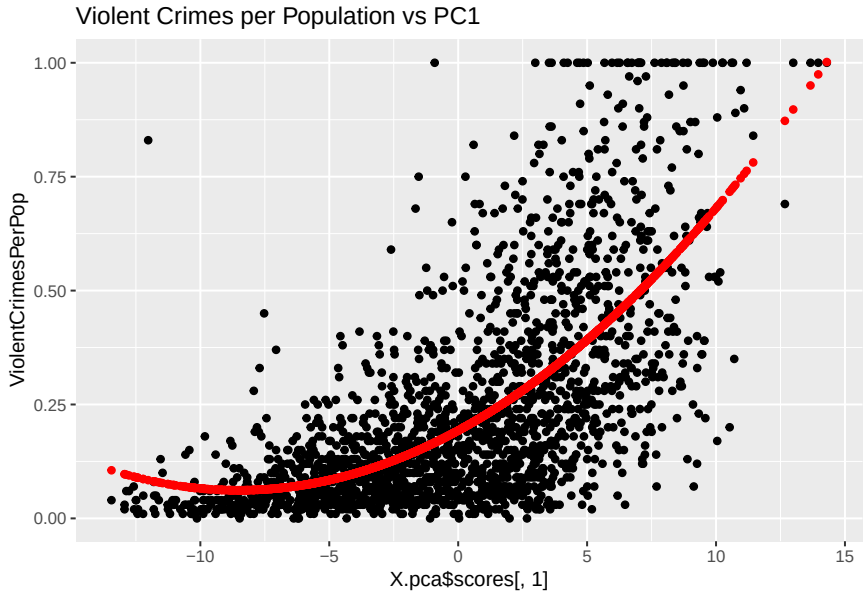


3.

In this section, a quadratic linear model is to be applied using the violent crime rate as a target and PC1 as feature.

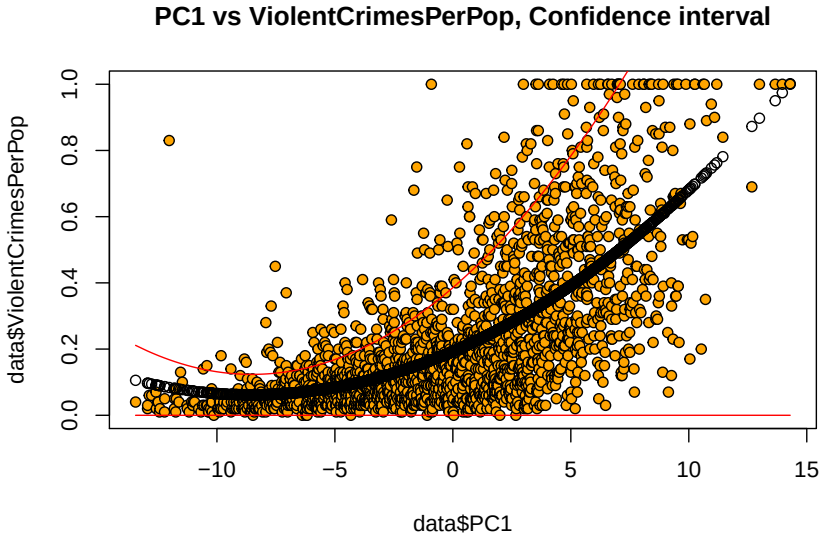
```
X_data <- as.data.frame(cbind(X, data_com[,101]))
colnames(X_data)[100] <- "ViolentCrimesPerPop"

model = lm(ViolentCrimesPerPop ~ poly(X.pca$scores[,1], degree = 2), data= X_data)
```

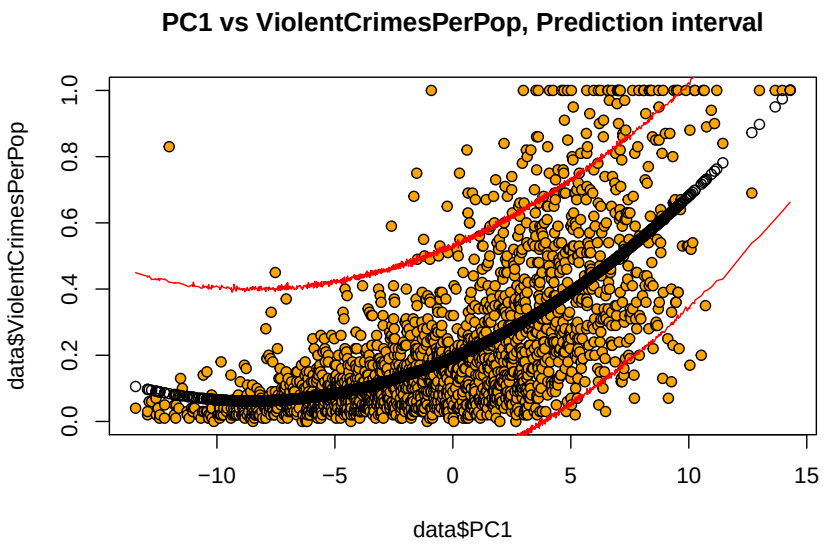
A plot is shown down below presenting the observations scattered in black (PC1 vs Violent Crimes per Population) and the predictions in red. It is visible how the model has been able to capture the underlying relationship between two variables emulating the curve inside the cloud of points.



4.

Parametric Bootstrap will be used in this point to estimate the confidence and prediction bands from the model in section 3. In case of the prediction interval, the plotted lines envelop closely the fitted predictions and fall inside the cloud of point observations, bordering the edges though. The plot of confidence interval shows lines that imitate the curve the fitted predictions as well. The lines are closer to the predicted values at the beginning, whereas as ViolentCrimesPerPop increases they move away. The bottom line is straight, yet to be fixed.




PC1 vs ViolentCrimesPerPop, Prediction interval
data$ViolentCrimesPerPop
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
15
data$PC1