Question 9.1

After applying Principal Component Analysis (PCA) to the uscrime dataset from Question 8.2, there was a significant increase in predicted accuracy for the test data. Figure 1 details the entire code used to predict the PCA values and the final model. In the last model, before PCA, the predicted accuracy from the model generated with all predictors was 155. This value lies far below the lowest value in the given uscrime dataset of 342, and in combination with the notion that the data was overfitted w.r.t the model, it is easily observed that this predicted value was inaccurate. However, after applying PCA to the uscrime dataset it produces a result that correlates with each predictor detailing the proportion to which the variance is distributed. The principal components and subsequently, the proportion of variances, are aligned with the data in a 1-1 fashion such that PC1 = M (the first predictor of the uscrime dataset), PC2 = So (2nd predictor of the uscrime dataset), and so on and so forth. Each proportion of variance is visualized in Figure 2 following the same pattern.

By observing the cumulative variance values generated in Figure 3, we can determine which predictors are the most important to the model and dataset. Extracting the variances that have a smaller effect on the cumulative variance can reduce the dimensionality of the model and thus increase its predictive potential. This notion is witnessed in Figures 4, 5, and 6 where are the final outputs of the model with the principal components deemed most influential in the prior figures. I did incorporate a histogram of the residuals computed in Figure 4 so that the fitting of the model with the data could be visualized (per Using Linear Regression for Predictive Modeling in R). According to the source previously linked, the histogram should have a fairly symmetrical distribution around zero, and in my case this seems to be present. Additionally, they should also possess the quality of being normally distributed around zero, essentially meaning that there should be a bell curve type distribution. Again, the model ran does have this quality as seen in Figure 5. Finally, Figure 6 is the result of the predicted output with the model only taking into account the first 7 predictors (M, So, Ed, Po1, Po2, LF, M.F). This result was determined to be 629, and falls perfectly within the range of the original uscrime dataset of [342, 1993].

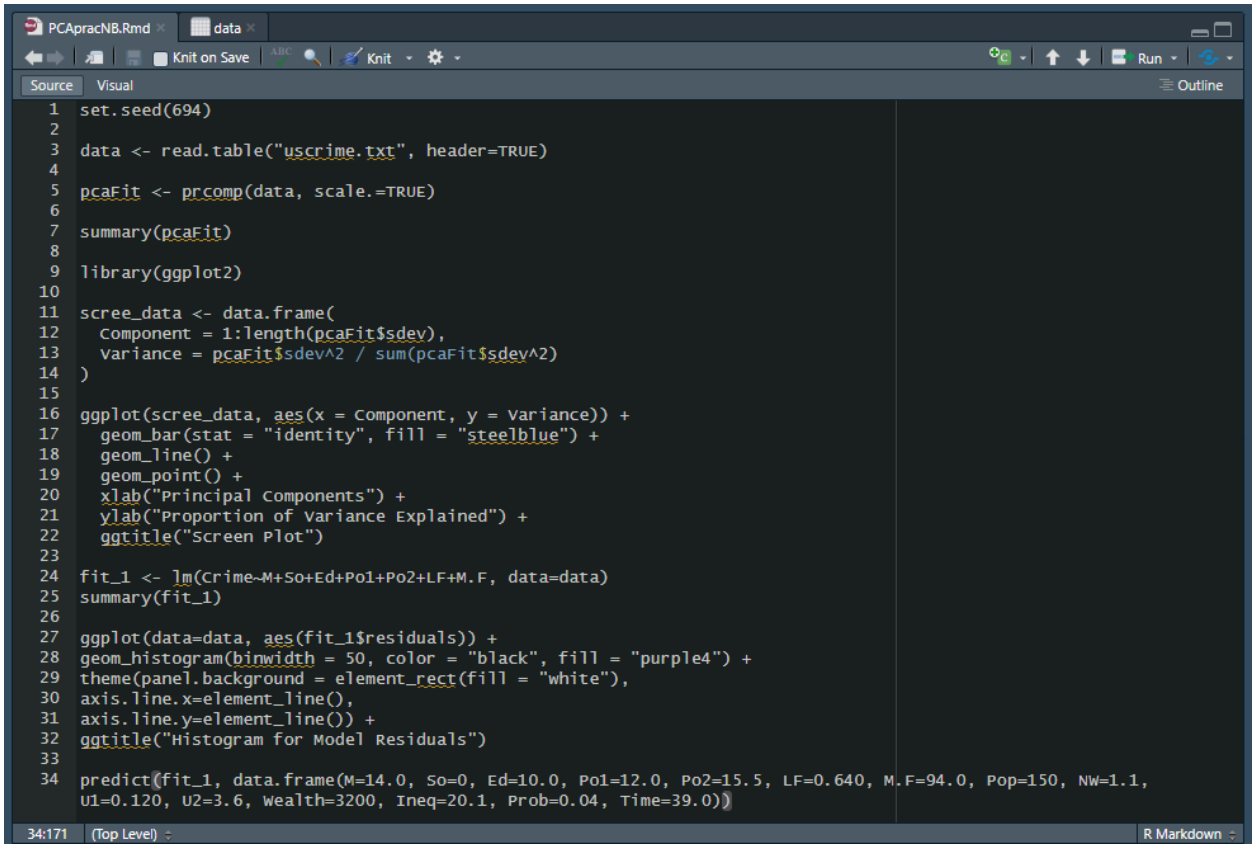

Figure 1. This is a screenshot of the final code used to run the homework assignment. Below are more detailed figures of what each snippet produces.

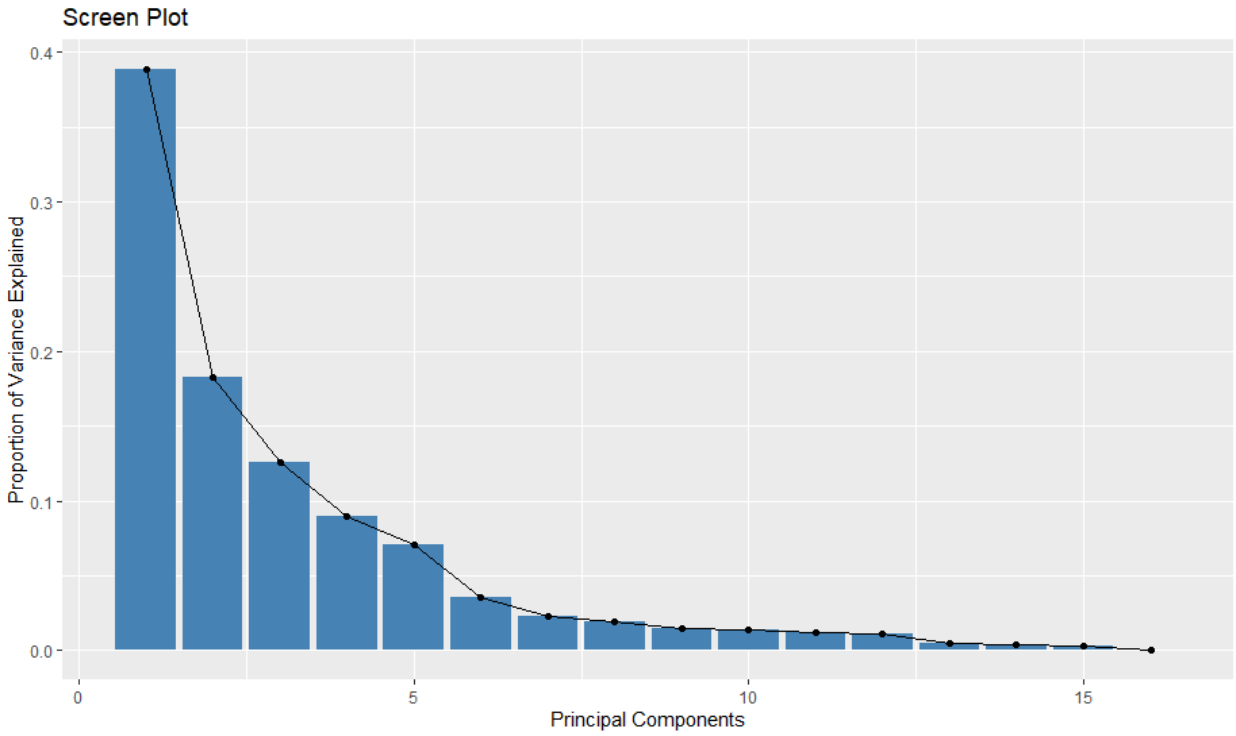

Figure 2. This plot shows the proportion of variance explained by each principal component. It helps in determining how many components should be considered for further analysis.

```
> summary(pcaFit)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.4944 1.7111 1.4208 1.19585 1.06341 0.75087 0.60237 0.55503 0.49244 0.47036 0.43856
Proportion of Variance 0.3889 0.1830 0.1262 0.08938 0.07068 0.03524 0.02268 0.01925 0.01516 0.01383 0.01202
Cumulative Proportion  0.3889 0.5719 0.6981 0.78744 0.85812 0.89336 0.91603 0.93529 0.95044 0.96427 0.97629
                         PC12    PC13    PC14    PC15    PC16
Standard deviation     0.41777 0.29147 0.26063 0.21813 0.06584
Proportion of Variance 0.01091 0.00531 0.00425 0.00297 0.00027
Cumulative Proportion  0.98720 0.99251 0.99676 0.99973 1.00000
>
```

Figure 3. This figure displays the that in practical terms, the first couple principal components (PC1 through PC7) are sufficient to explain most of the variability in the dataset (91.60%), which means we could reduce the dimensionality of the data to these few components with minimal loss of information. This dimensionality reduction can simplify analysis and visualization without significantly compromising the integrity of the data.

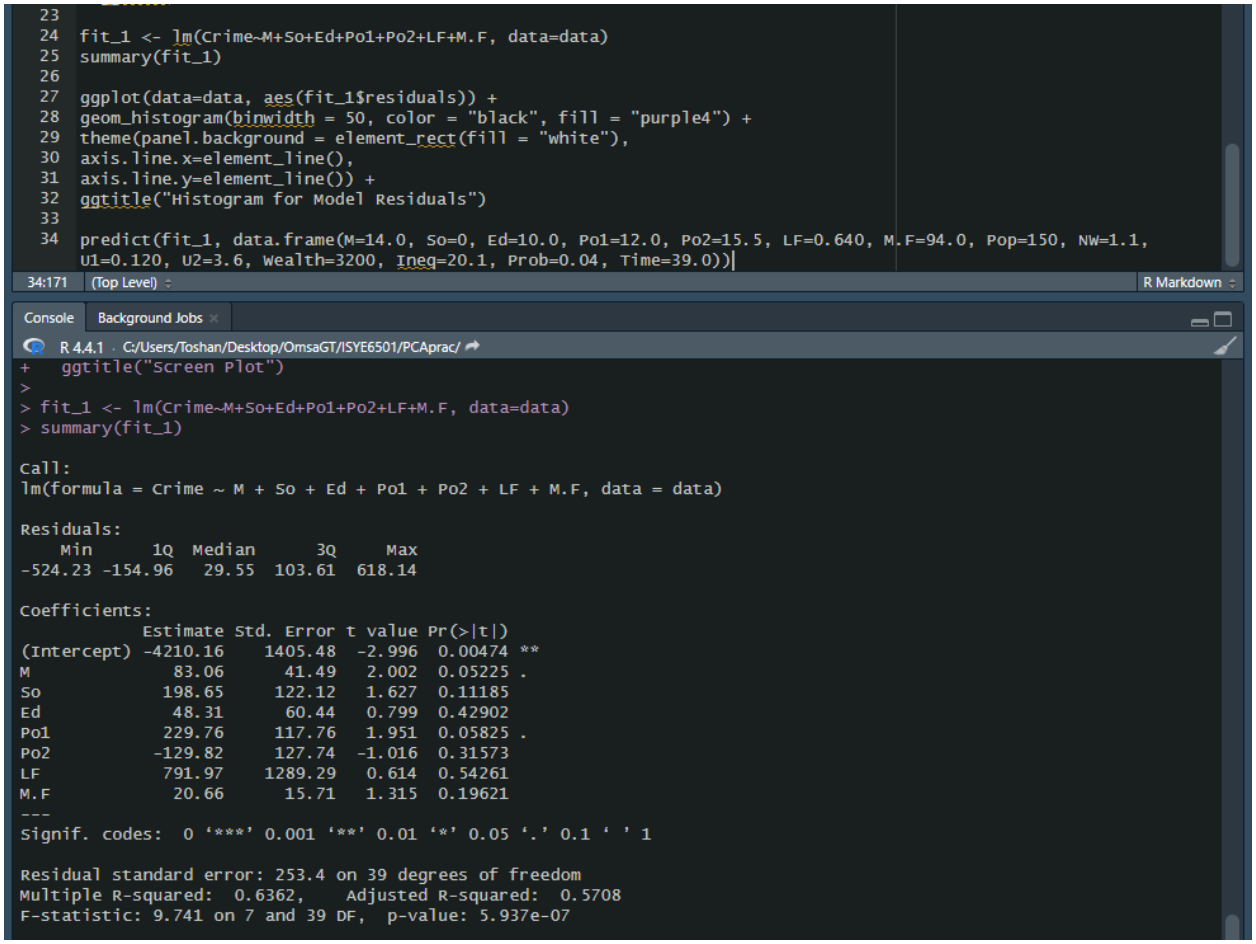

Figure 4. This figure denotes the new model determined by the PCA results (i.e the coefficients and residuals of the model). The sections outside of the coefficients detail how well the model fits the data, and the coefficients dictate how the hypothesis (relatedness of predictors to Crime) is supported.

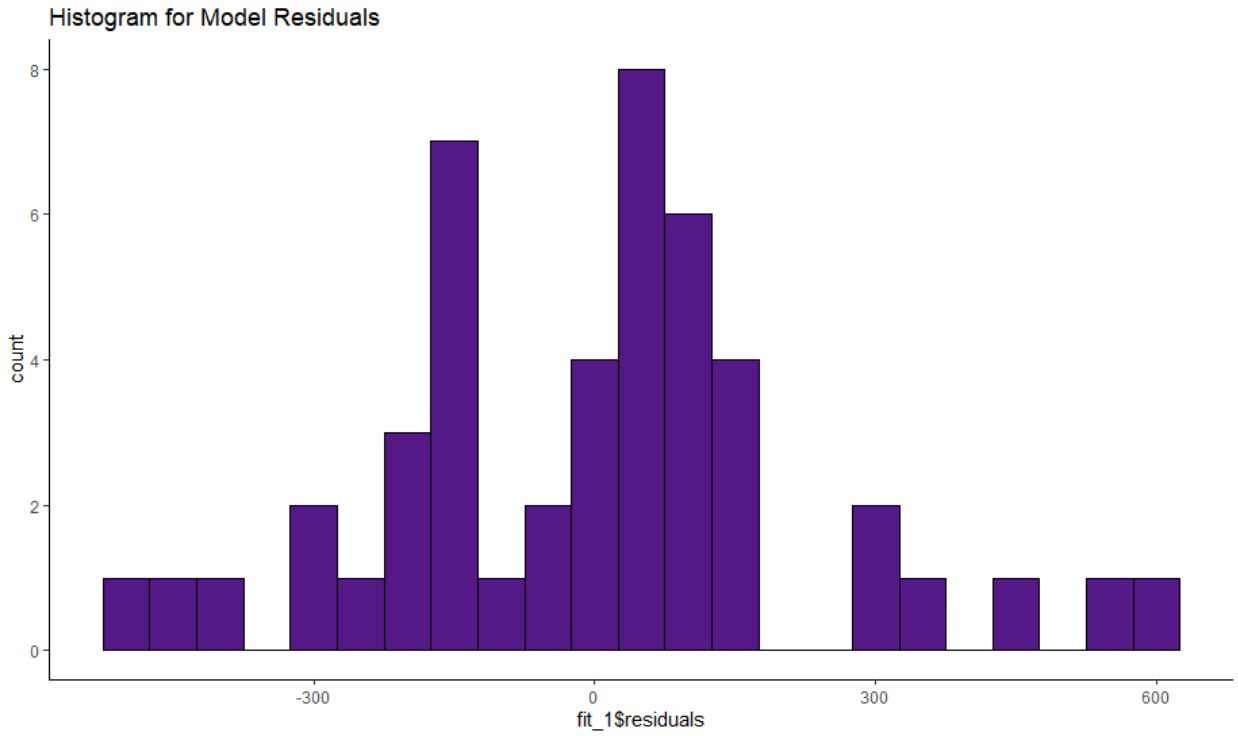

Figure 5. This is a histogram of the residuals calculated in Figure 4. The visualization of these values shows more easily how well the model fits the data.

```
>
> predict(fit_1, 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))
       1
629.1321
>
```

Figure 6. This figure is the final value associated with the new PCA model. This puts the value within the range of the crime data for the uscrime dataset.