Question 9.1

Using the same crime data set `uscrime.txt` as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Question 8.2. You can use the R function `prcomp` for PCA. (**Note** that to first scale the data, you can include `scale. = TRUE` to scale as part of the PCA function. Don't forget that, to make a prediction for the new city, you'll need to unscale the coefficients (i.e., do the scaling calculation in reverse)!)

Last week, I constructed a web app using the R package Shiny that also answer this week's questions. This Shiny webapp presents two main pages: "Regression" and "PCA." In the Regression page, users can select any subset of 15 predictors to model Crime via linear regression. The app displays the model's coefficients, R-squared values, a prediction for a new hypothetical city, and a regression plot. GGplot2 was used to

In the PCA page, users see principal component analysis on the selected numeric predictors. A biplot is displayed, and an RMSE-based reconstruction error table helps determine how many components best represent the original data. This addresses model complexity, showing how dimensionality reduction can highlight key factors and potentially mitigate overfitting. It was built on the R method 'prcomp' and scaling was included. Changing analysis based on principal components is left as an exercise for the reader, but all information is presented such that the user can make an educated decision. This has the benefit of allowing the user to consider external context.

Together, the two pages provide a complete analysis workflow for this US Crime dataset. My code is present in the uploaded file 'regression.R'.

Attached to the bottom of this writeup are screenshots from the webapp's two pages.

As an aside, I found the PCA biplot fascinating. I didn't expect Time to span the y axis so substantially, and the x axis so minimally.

US Crime Data: Regression & PCA

Select which predictors to use in the linear model (check boxes are horizontal). Then see regression outputs, diagnostic plots, a prediction for the new city, and PCA results (including RMSE by number of PCs).

Choose Predictors for Crime Model:

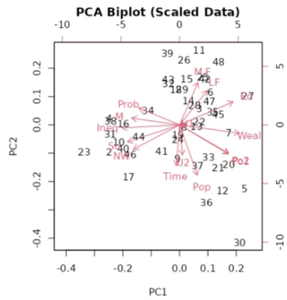
☑ M ☑ So ☑ Ed ☑ Po1 ☑ Po2 ☑ LF ☑ M.F ☑ Pop ☑ NW ☑ U1 ☑ U2 ☑ Wealth ☑ Ineq ☑ Prob ☑ Time

Regression | PCA

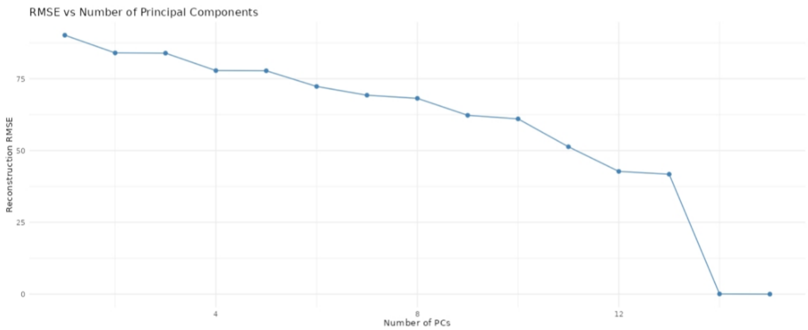
PCA Summary

```
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
```

PCA Biplot



RMSE by Number of Principal Components



Minor Writeup:

Interpretation: As we increase the number of principal components, the RMSE for reconstructing the original data generally decreases. The first few components often capture most variance, so we see a rapid drop in RMSE initially. Additional components refine the fit but may yield diminishing returns.

To choose the 'best' number of components, look for the 'elbow' in the chart or a minimal RMSE that balances simplicity and accuracy. Each principal component corresponds to a linear combination of the original predictors, emphasizing those with the greatest variance.

US Crime Data: Regression & PCA

Select which predictors to use in the linear model (check boxes are horizontal). Then see regression outputs, diagnostic plots, a prediction for the new city, and PCA results (including RMSE by number of PCs).

Choose Predictors for Crime Model:

☑ M ☑ So ☑ Ed ☑ Po1 ☑ Po2 ☑ LF ☑ M.F ☑ Pop ☑ NW ☑ U1 ☑ U2 ☑ Wealth ☑ Ineq ☑ Prob ☑ Time

Regression | PCA

Selected Predictors:

```
[1] "M"      "So"     "Ed"     "Po1"    "Po2"    "LF"     "M.F"    "Pop"
[9] "NW"     "U1"     "U2"     "Wealth" "Ineq"   "Prob"   "Time"
```

Model Coefficients

Term	Coefficient
(Intercept)	-5984.29
M	87.83
So	-3.80
Ed	188.32
Po1	192.80
Po2	-109.42
LF	-663.83
M.F	17.41
Pop	-0.73
NW	4.20
U1	-5827.10
U2	167.80
Wealth	0.10
Ineq	70.67
Prob	-4855.27
Time	-3.48

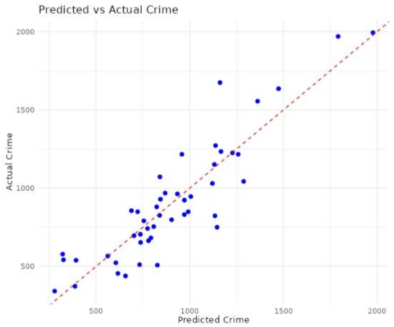
Model Summary

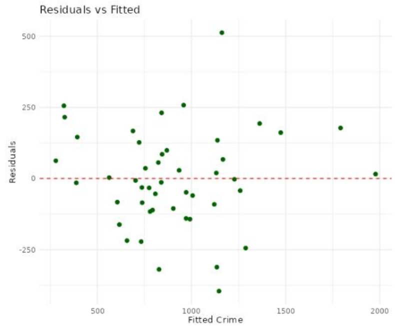
```
Call:
lm(formula = Crime ~ ., data = df_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-395.74  -98.09   -6.69  112.99  512.67

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
M            8.783e+01  4.171e+01   2.106 0.043443 *
So          -3.803e+00  1.488e+02  -0.026 0.979765
Ed           1.883e+02  6.209e+01   3.033 0.004861 **
Po1          1.928e+02  1.061e+02   1.817 0.078892 .
Po2         -1.094e+02  1.175e+02  -0.931 0.358830
LF          -6.638e+02  1.470e+03  -0.452 0.654654
M.F          1.741e+01  2.035e+01   0.855 0.398995
Pop         -7.330e-01  1.290e+00  -0.568 0.573845
NW           4.204e+00  6.481e+00   0.649 0.521279
U1          -5.827e+03  4.210e+03  -1.384 0.176238
U2           1.678e+02  8.234e+01   2.038 0.050161 .
Wealth       9.617e-02  1.037e-01   0.928 0.360754
Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
Time        -3.479e+00  7.165e+00  -0.486 0.630708
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 209.1 on 31 degrees of freedom
Multiple R-squared:  0.8031,    Adjusted R-squared:  0.7078
F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

R-Squared Values

R_Squared	Adj_R_Squared
0.80	0.71

Regression Diagnostic Plots





Prediction for New City

Predicted_Crime
155.43

I.