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)!)

ANSWER:
I first ran a PCA analysis on the data using prcomp

```
myPCA <- prcomp(crime[,1:15], scale = TRUE)
myPCA
summary(myPCA)
```

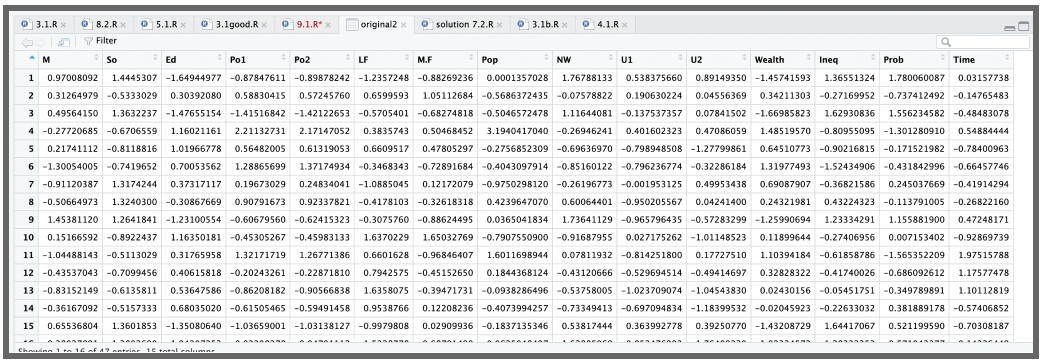
I used the summary of the PCA analysis to determine which principal components were most important in explaining the data. I determined that PC1 - PC9 were useful since it captured 95% of the variance

```
> summary(myPCA)
Importance of components:
                          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
```

I then used the rotation and x of the PCA model to restate the PCA in terms of the original variables.

```
pcaroation2 <- myPCA$rotation[,1:9]
standardized2 <- myPCA$x[,1:9]
original2 <- t(pcaroation2 %*% t(standardized2))
```

The output of original2 looks like this:



	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time
1	0.97008092	1.4445307	-1.64944977	-0.87847611	-0.89878242	-1.2357248	-0.88269236	0.0001357028	1.76788133	0.538375660	0.89149350	-1.45741593	1.36551324	1.780060087	0.03157738
2	0.31264979	-0.5333029	0.30392080	0.58830415	0.57245760	0.6599593	1.05112684	-0.5686372435	-0.07578822	0.190630224	0.04556369	0.34211303	-0.27169952	-0.737412492	-0.14765483
3	0.49564150	1.3632237	-1.47655154	-1.41516842	-1.42122653	-0.5705401	-0.68274818	-0.5046572478	1.11644081	-0.137537357	0.07841502	-1.66985823	1.62930836	1.556234582	-0.48483078
4	-0.27720685	-0.6706559	1.16021161	2.21132731	2.17147052	0.3835743	0.50468452	3.1940417040	-0.26946241	0.401602323	0.47086059	1.48519570	-0.80955095	-1.301280910	0.54884444
5	0.21741112	-0.8118816	1.01966778	0.56482005	0.61319053	0.6609517	0.47805297	-0.2756852309	-0.69636970	-0.798948508	-1.27799861	0.64510773	-0.90216815	-0.171521982	-0.78400963
6	-1.30054005	-0.7419652	0.70053562	1.28865699	1.37174934	-0.3468343	-0.72891684	-0.4043097914	-0.85160122	-0.796236774	-0.32286184	1.31977493	-1.52434906	-0.431842996	-0.66457746
7	-0.91120387	1.3174244	0.37317117	0.19673029	0.24834041	-1.0885045	0.12172079	-0.9750298120	-0.26196773	-0.001953125	0.49953438	0.69087907	-0.36821586	0.245037669	-0.41914294
8	-0.50664973	1.3240300	-0.30867669	0.90791673	0.92337821	-0.4178103	-0.32618318	0.4239647070	0.60064401	-0.950205567	0.04241400	0.24321981	0.43224323	-0.113791005	-0.26822160
9	1.45381120	1.2641841	-1.23100554	-0.60679560	-0.62415323	-0.3075760	-0.88624495	0.0365041834	1.73641129	-0.965796435	-0.57283299	-1.25990694	1.23334291	1.155881900	0.47248171
10	0.15166592	-0.8922437	1.16350181	-0.45305267	-0.45983133	1.6370229	1.65032769	-0.7907550900	-0.91687955	0.027175262	-1.01148523	0.11899644	-0.27406956	0.007153402	-0.92869739
11	-1.04488143	-0.5113029	0.31765958	1.32171719	1.26771386	0.6601628	-0.96846407	1.6011698944	0.07811932	-0.814251800	0.17727510	1.10394184	-0.61858786	-1.565352209	1.97515788
12	-0.43537043	-0.7099456	0.40615818	-0.20243261	-0.22871810	0.7942575	-0.45152650	0.1844368124	-0.43120666	-0.529694514	-0.49414697	0.32828322	-0.41740026	-0.686092612	1.17577478
13	-0.83152149	-0.6135811	0.53647586	-0.86208182	-0.90566838	1.6358075	-0.39471731	-0.0938286496	-0.53758005	-1.023709074	-1.04543830	0.02430156	-0.05451751	-0.349789891	1.10112819
14	-0.36167092	-0.5157333	0.68035020	-0.61505465	-0.59491458	0.9538766	0.12208236	-0.4073994257	-0.73349413	-0.697094834	-1.18399532	-0.02045923	-0.22633032	0.381889178	-0.57406852
15	0.65536804	1.3601853	-1.35080640	-1.03659001	-1.03138127	-0.9979808	0.02909936	-0.1837135346	0.53817444	0.363992778	0.39250770	-1.43208729	1.64417067	0.521199590	-0.70308187

We now have our original variables and not the principal components. However, this data is still scaled, so I had to use the center output to unscale the data.

```
unscaled_data <- t(t(original2)+ myPCA$center)
```

Now i had my PCA 1-9 unscaled and expressed in the original variables. I was ready to run the linear regression on this data.

```
uscrimePC2 <- cbind(unscaled_data, crime[,16])
modelPCA2 <- lm(V16~., data = as.data.frame(uscrimePC2))
summary(modelPCA2)

test_point <- data.frame(M=14.0, So =0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = .640,
                         M.F = 94.0, Pop = 150, NW = 1.1, U1 = .120, U2 = 3.6,
                         Wealth = 3200, Ineq = 20.1, Prob = 0.040, Time = 39.0)

pred_model <- predict(modelPCA2, test_point)
pred_model
```

There were a few issues with this output, mainly that the summary showed that many of the variables could not be defined because of singularities. I researched this online and learned that it means that the predictor variables are highly correlated (which also seemed to be stated in the office hours).

```
Coefficients: (6 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5185.56    8634.35  -0.601   0.5518
M             156.17      67.67   2.308   0.0267 *
So            231.32      92.98   2.488   0.0175 *
Ed           -116.09     106.83  -1.087   0.2842
Po1          3563.54    1788.64   1.992   0.0538 .
Po2         -3114.45    1751.55  -1.778   0.0836 .
LF             75.51      69.68   1.084   0.2855
M.F            39.65      76.14   0.521   0.6056
Pop           -35.34      70.98  -0.498   0.6215
NW           -284.19     137.50  -2.067   0.0458 *
U1                NA         NA      NA       NA
U2                NA         NA      NA       NA
Wealth            NA         NA      NA       NA
Ineq              NA         NA      NA       NA
Prob              NA         NA      NA       NA
Time              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 239.3 on 37 degrees of freedom
Multiple R-squared:  0.692,     Adjusted R-squared:  0.6171
F-statistic: 9.239 on 9 and 37 DF,  p-value: 3.588e-07
```

Ultimately, the output of the model for the test data point made no sense as it was a negative number (-11,000).