Problem:

For the data generated in Problem 7.7, perform a principal components analysis on X1,X2,...,X9. Compare the results with what is known about the population.

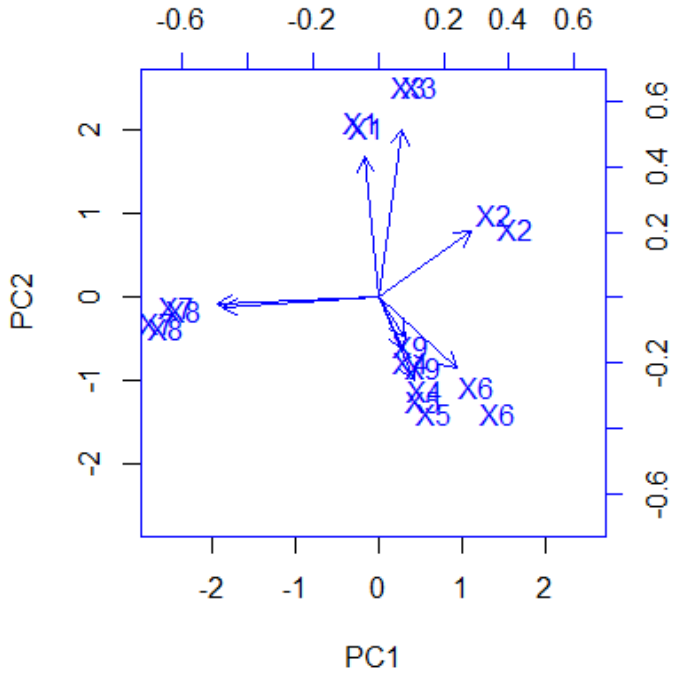
Code and Output:

```
library("xtable")
set.seed(36541)
X1<-rnorm(100, mean = 0, sd = 1)
X1 <- 5*X1
X1
set.seed(43893)
X2<-rnorm(100, mean = 0, sd = 1)
X2 = 3*X2
X2
set.seed(45671)
X3<-rnorm(100, mean = 0, sd = 1)
X3 = X1 + X2 + 4*X3
X3
set.seed(65431)
X4<-rnorm(100, mean = 0, sd = 1)
X4 = X4
X4
set.seed(98753)
X5<-rnorm(100, mean = 0, sd = 1)
X5 = 4*X5
X5
set.seed(78965)
X6<-rnorm(100, mean = 0, sd = 1)
X6 = X5 - X4 + 6*X6
X6
set.seed(67893)
X7<-rnorm(100, mean = 0, sd = 1)
X7 = 2*X7
X7
set.seed(34521)
X8<-rnorm(100, mean = 0, sd = 1)
X8 = X7 + 2*X8
X8
set.seed(98431)
X9<-rnorm(100, mean = 0, sd = 1)
X9 = 4*X9
X9
set.seed(67895)
Y<-rnorm(100, mean = 0, sd = 1)
Y = 5 + X1 + 2*X2 + X3 + 100*Y
```

```
cmat1=cbind(X1,X2,X3,X4,X5,X6,X7,X8,X9)
cmat1 = cor(cmat1)
cmat1_table<- xtable(cmat1)
View(cmat1_table)
```

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1.00000000	-0.06075012	0.67082320	-0.01951675	0.17175762	0.02234367	0.12788004	0.11233388	0.02907846
X2	-0.06075012	1.00000000	0.34686301	-0.06968350	0.01379695	0.08942669	-0.15491798	-0.04708558	0.01415194
X3	0.67082320	0.34686301	1.00000000	-0.08930995	0.12947873	0.02704916	0.09859133	0.04788110	-0.01114966
X4	-0.01951675	-0.06968350	-0.08930995	1.00000000	0.07566002	-0.12423338	-0.09054126	-0.09104812	-0.08759530
X5	0.17175762	0.01379695	0.12947873	0.07566002	1.00000000	0.48986171	0.11558495	0.16185961	0.13488075
X6	0.02234367	0.08942669	0.02704916	-0.12423338	0.48986171	1.00000000	-0.06007157	-0.04877537	-0.02382420
X7	0.12788004	-0.15491798	0.09859133	-0.09054126	0.11558495	-0.06007157	1.00000000	0.70683153	0.02856373
X8	0.11233388	-0.04708558	0.04788110	-0.09104812	0.16185961	-0.04877537	0.70683153	1.00000000	-0.06394655
X9	0.02907846	0.01415194	-0.01114966	-0.08759530	0.13488075	-0.02382420	0.02856373	-0.06394655	1.00000000

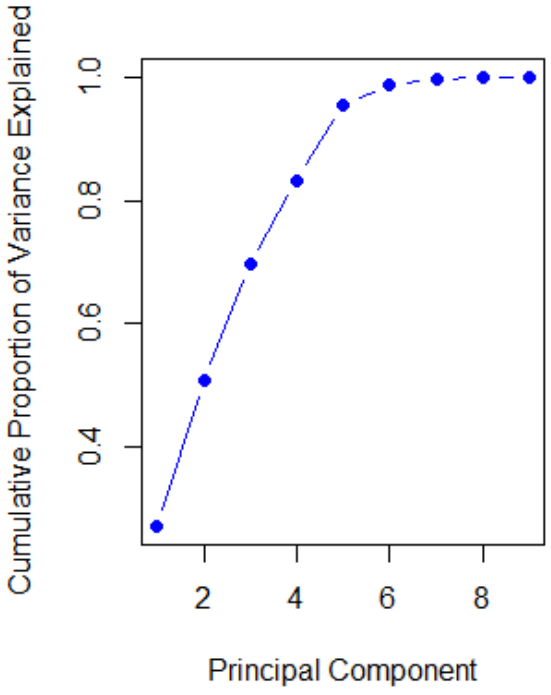
```
fit_model <- prcomp(cmat1_table, scale=T)
summary(fit_model)
names(fit_model)
fit_model$rotation
biplot(fit_model,scale=0,col="blue")
```

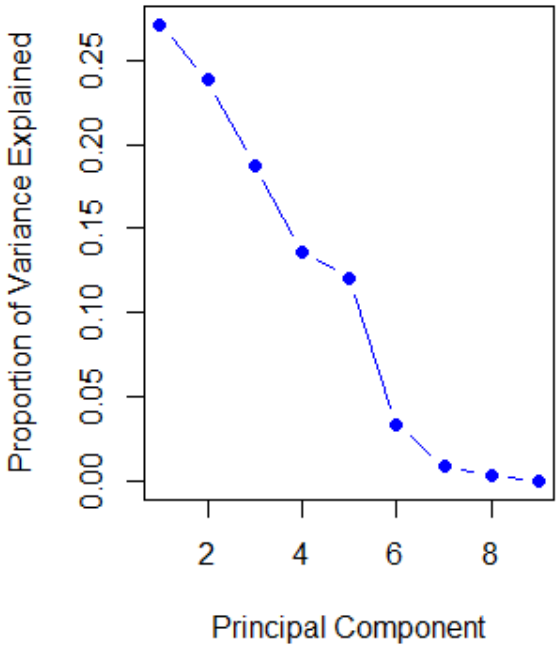


```
> summary(fit_model)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5      PC6      PC7     PC8          PC9
Standard deviation     1.5615 1.4640 1.2997 1.1078 1.0427 0.54847 0.28415 0.1824 1.624e-16
Proportion of Variance 0.2709 0.2382 0.1877 0.1363 0.1208 0.03342 0.00897 0.0037 0.000e+00
Cumulative Proportion  0.2709 0.5091 0.6968 0.8331 0.9539 0.98733 0.99630 1.0000 1.000e+00
```

```
stddev <- fit_model$sdev
var1 <- stddev^2
propvar <- var1/sum(var1)
propvar
plot(propvar, xlab = "Principal Component", ylab = "Proportion of Variance Explained",type = "b",col="blue")
plot(cumsum(propvar), xlab = "Principal Component",ylab = "Cumulative Proportion of Variance Explained",type
     = "b",col="blue")
```

```
> propvar
[1] 2.709086e-01 2.381566e-01 1.877024e-01 1.363469e-01 1.207938e-01 3.342449e-02 8.971284e-03 3.695950e-03
[9] 2.929294e-33
```





From the above plots we can see the factors that contribute towards variability in data. The above plots show that the 6 components contribute towards 97% variance in data.

Problem:

(Continuation of Problem 14.3.) Perform the regression of Y on the principal components. Compare the results with the multiple regression of Y on X1 to X9.

Code and Output:

```
data1<- cbind(X1,X2,X3,X4,X5,X6,X7,X8,X9,Y)
colnames(data1)<- c('X1','X2','X3','X4','X5','X6','X7','X8','X9','Y')
data1<- as.data.frame(data1)
pca_model <- princomp(~X1+X2+X3+X4+X5+X6+X7+X8+X9, cor=TRUE)
pca_model
summary(pca_model)
data2<-cbind(Y,pca_model$scores)
data2<-data.frame(data2)
reg_model<-lm(Y~ .,data=data2)
reg_model
summary(reg_model)
reg_model1<-lm(Y~ .,data=data1)
reg_model1
summary(reg_model1)
```

```
> pca_model
Call:
princomp(formula = ~X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
    cor = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9
1.4170336 1.2851546 1.1831701 1.0547829 1.0173047 0.9501120 0.6354844 0.5461691 0.4337879

 9  variables and  100 observations.
> summary(pca_model)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     1.4170336 1.2851546 1.1831701 1.0547829 1.0173047 0.9501120 0.63548445 0.54616911 0.43378786
Proportion of Variance 0.2231094 0.1835136 0.1555435 0.1236185 0.1149899 0.1003014 0.04487116 0.03314452 0.02090799
Cumulative Proportion  0.2231094 0.4066230 0.5621665 0.6857850 0.8007749 0.9010763 0.94594749 0.97909201 1.00000000
```

```
> reg_model

Call:
lm(formula = Y ~ ., data = data2)

Coefficients:
(Intercept)       Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6       Comp.7       Comp.8       Comp.9
     6.1171       7.4686      -8.2959      11.2394       0.2771      11.8794     -27.9682       6.1593      -0.7068      13.5739

> summary(reg_model)

Call:
lm(formula = Y ~ ., data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-163.001  -48.237    2.004   52.356  149.068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.1171     7.6199   0.803 0.424216
Comp.1        7.4686     5.3773   1.389 0.168288
Comp.2       -8.2959     5.9291  -1.399 0.165196
Comp.3       11.2394     6.4402   1.745 0.084365 .
Comp.4        0.2771     7.2241   0.038 0.969485
Comp.5       11.8794     7.4902   1.586 0.116250
Comp.6      -27.9682     8.0200  -3.487 0.000757 ***
Comp.7        6.1593    11.9906   0.514 0.608738
Comp.8       -0.7068    13.9515  -0.051 0.959707
Comp.9       13.5739    17.5659   0.773 0.441697
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.2 on 90 degrees of freedom
Multiple R-squared:  0.1998,    Adjusted R-squared:  0.1198
F-statistic: 2.497 on 9 and 90 DF,  p-value: 0.01347

> reg_model1

Call:
lm(formula = Y ~ ., data = data1)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5           X6           X7           X8           X9
    7.69015      2.51818     -6.55104      2.80932    -22.01566     -0.72739     -0.34317     -5.15657     -1.24393     -0.05784

> summary(reg_model1)

Call:
lm(formula = Y ~ ., data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-163.001  -48.237    2.004   52.356  149.068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.69015    7.87334   0.977   0.3313
X1            2.51818    2.46577   1.021   0.3099
X2           -6.55104    2.93303  -2.234   0.0280 *
X3            2.80932    1.89273   1.484   0.1412
X4          -22.01566    9.38712  -2.345   0.0212 *
X5           -0.72739    2.41813  -0.301   0.7643
X6           -0.34317    1.30469  -0.263   0.7931
X7           -5.15657    5.49052  -0.939   0.3502
X8           -1.24393    3.76325  -0.331   0.7418
X9           -0.05784    1.91521  -0.030   0.9760
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.2 on 90 degrees of freedom
Multiple R-squared:  0.1998,    Adjusted R-squared:  0.1198
F-statistic: 2.497 on 9 and 90 DF,  p-value: 0.01347
```

From the above summaries we can see that there is no difference on both outputs(residual standard error, multiple R-sqaured, Adjusted R-sqaured, F-statistic, p-value)