STAT 4400 Take Home Test I Key (75%)

1. You are working as a statistical assistant for the ABC baseball team at Austin of Texas and plan to do analysis on players' performance. You have gathered data on 100 players, measuring 11 variables on each, which represent various characteristics of the players: Salary (in thousands of dollars), batting average, on-base percentage, number of runs, number of hits, number of doubles, number of home runs, number of walks, number of strike-outs, number of errors, and player's name (in quotation marks). The data set is named baseball.csv.

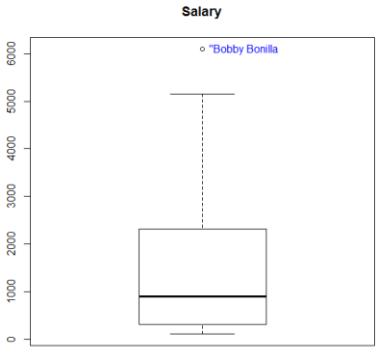
- Players' batting averages are calculated as the ratio of number of hits to the number of hits plus the number of outs.
- On-base percentage is the ratio of number of hits plus the number of walks to the number of hits plus the number of walks plus the number of outs.
- A batting average above 0.300 is very good; OBP above 0.400 is excellent.

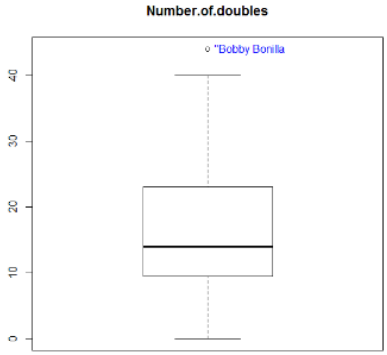
  The questions that the team boss want to hear from you are:

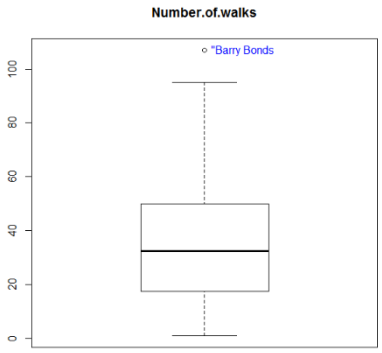
  a. (3 pts) Are there particular players that are highly unusual in terms of the measured characteristics? If so, identify them.

The eleven variables in the data set can be classified into three groups. The players name, salary, and performance.

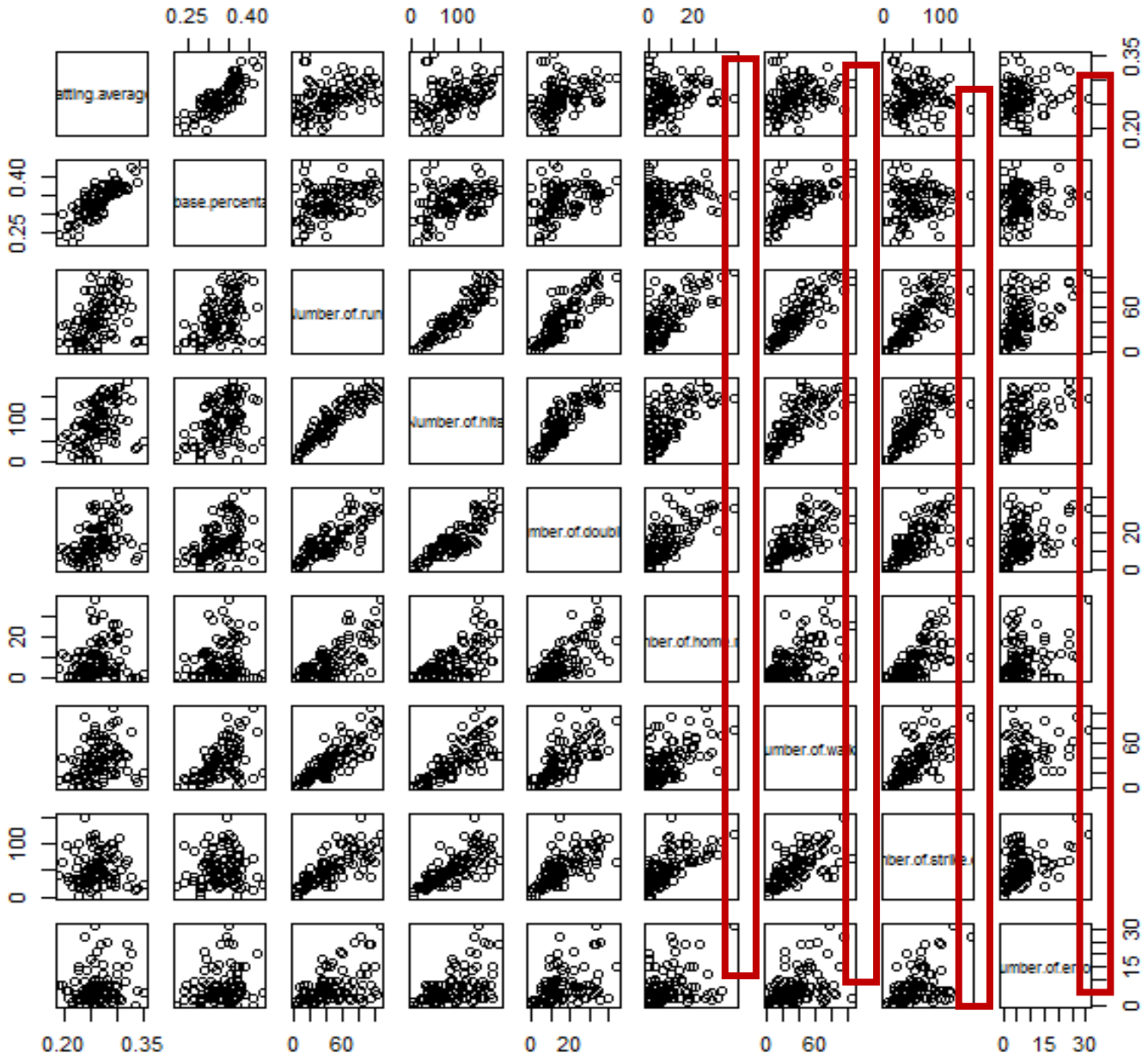
- The boxplot of salary shows that Bobby Bonilla is the highest paid player out of the 100 sampled. When we closely examine the observation of Bobby, he has highest scores for number of doubles, batting average and number of walks. That is, Bobby Bonilla is an unusual observation.





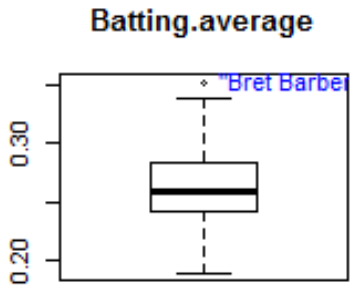


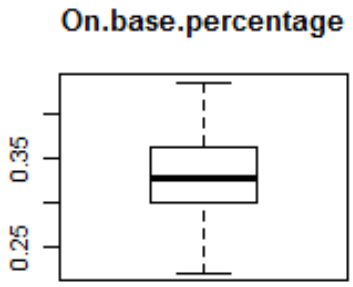
- There are more than three variables related to the performance of the players. The scatter plots matrix shows potential outliers in number of home run, number of errors except the outliers of variables caused by Bobby.

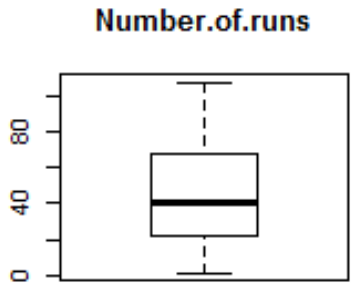


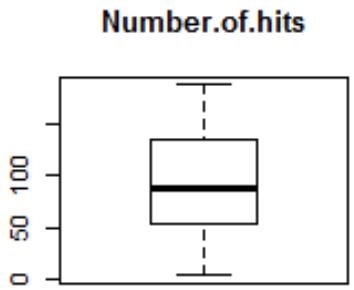
The boxplots in the following show the potential outliers of higher values for each variable related to the performance. Delino Deshields has a higher batting average and number of strike outs, Barry Bounds has a high number of walks, and Bret Barberie has a high batting average. There are a quite a few players with exceptionally high numbers of home runs and errors.

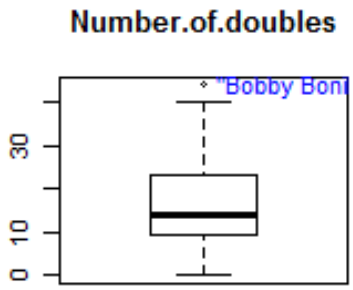
There are no potential outliers for on base percentage, number of runs and number of hits.

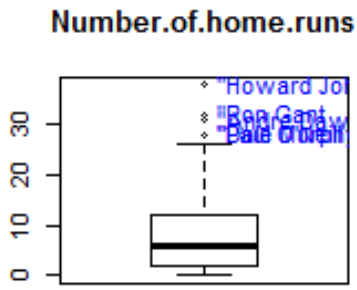


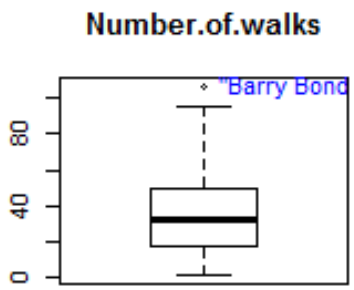


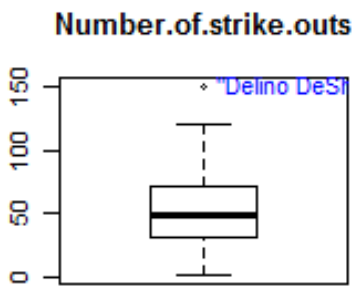


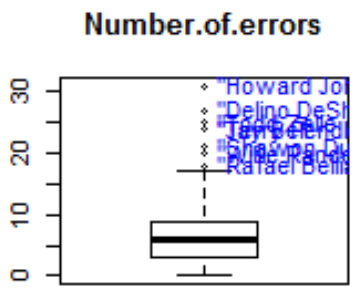












Outliers

Barry Bond for salary, number of walk, number of doubles

Batting.average	Number.of.doubles	Number.of.home.runs	Number.of.walks	Number.of.strike.outs	Number.of.errors
Bret Barberie	Bobby Bonilla	Andre Dawson	Barry Bonds	Delino Deshields	Shawon Dunston
		Howard Johnson			Delino Shields
		Dale Murphy			Howard Johnson
		Ron Gant Willie Randolph			Willie Randolph
		Paul O'Neill Jay Bell			Willie Randolph
					Todd Zeile
					Terry Pendleton

b. (3 pts) Are there notable associations/relationships between some of the variables? If so, describe them.

- The matrix scatter plots shows strong linear associations between Players' batting averages and on base percentage, number of runs and number of hits. The correlation matrix shows us more accurate high associations among the variables of number of runs, number of hits, number of double.
  Cor(batting averages, on base percentage)= 0.81, cor(number of runs, number of hits)=0.926, cor(number of runs, number of double)=0.832, cor(number of hits, number of double)=0.88.

	Salary	Batting average	On base percenta	Number of runs	Number of hits	Number of double	Number of home r	Number of walks	Number of strike
Batting average	0.224								
On base percenta	0.297	0.810							
Number of runs	0.606	0.353	0.431						
Number of hits	0.582	0.422	0.409	0.926					
Number of double	0.494	0.395	0.385	0.832	0.880				
Number of home r	0.574	0.141	0.183	0.707	0.645	0.699			
Number of walks	0.578	0.259	0.551	0.839	0.779	0.710	0.541		
Number of strike	0.343	-0.007	0.082	0.756	0.746	0.682	0.685	0.655	
Number of errors	0.158	0.133	0.146	0.470	0.490	0.390	0.289	0.394	0.452

b. (3 pts) Is there a way to graphically represent the raw data for the 100 players and draw conclusions about the data set from such a graph?
The boxplot used above provides the verbalizations of data restricted in univariates respects. The matrix scatter plots shows the association between two variables only. Start plots might be used to make a graphical representation of each player, but they are often hard to find the trend, compare or interpret when the number of variables are more than five. Overall, graphical displays are not sufficient to draw conclusions about the data set.

c. (5 pts) Baseball provides a rare opportunity to judge the value of an employee. In this case, a player -- by standardized measures of performance. The question is, can we find a less number of items to describe at least 90% of variations of the original data set or the original number of characteristics?
The salary is not nothing to do with performance, so nine variables (removing salary) will be considering in the analysis. Even although there are nine variables used to measure the performance in the data set, we have found that some of the variables, such as number of hits, number of runs, number of double, and two continuous variables are highly related. Principal components analysis is a procedure for identifying a smaller number of uncorrelated variables, from a large set of data and principal components analysis can explain the maximum amount of variance with the fewest number of principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.3047	1.2833	0.8772	0.69512	0.57424	0.46823	0.37993	0.23087	0.20517
Proportion of Variance	0.5902	0.1830	0.0855	0.05369	0.03664	0.02436	0.01604	0.00592	0.00468
Cumulative Proportion	0.5902	0.7732	0.8587	0.91236	0.94900	0.97336	0.98940	0.99532	1.00000

R output shows the first three principal components accounts for 85.87% of variance of data set, and the first four principal components accounts for 91.23% of variation of the data set. This is an acceptably large percentage. That is, two or three principal components can be used for describing the player's performance well.

d. (10 pts) If so, what are those items? Is there convenient interpretation of any of the items?
The large value of number of error

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Batting.average	-0.1924945	0.64451367	0.041011485	-0.41268129	0.1433047	-0.25362632	0.12815066	0.24506756	-0.46283578
On.base.percentage	-0.2275589	0.61619377	0.005291105	0.24533007	-0.3759436	-0.12000880	-0.22102290	-0.28498631	0.47170253
Number.of.runs	-0.4138803	-0.05174249	-0.051618412	0.10267399	0.1113135	0.08859030	0.58281425	-0.64545514	-0.19051423
Number.of.hits	-0.4112329	-0.02393977	0.004067615	-0.04561061	0.4312542	0.02802299	0.30010231	0.42815440	0.60669138
Number.of.doubles	-0.3913541	-0.02657516	-0.152324815	-0.22558748	0.3699855	0.38673683	-0.65557864	-0.22129768	-0.08265908
Number.of.home.runs	-0.3258526	-0.22381246	-0.350001669	-0.46818861	-0.6642450	0.14970426	0.09910239	0.16486035	0.03732826
Number.of.walks	-0.3740852	0.01848417	-0.039771824	0.67510806	-0.1497212	0.22602387	-0.04682752	0.42475852	-0.38242658
Number.of.strike.outs	-0.3424792	-0.36141740	-0.043855431	0.07407379	0.0233340	-0.82330684	-0.24469834	-0.04688135	-0.06497276
Number.of.errors	-0.2327740	-0.14104789	0.919999968	-0.15980278	-0.1966919	0.10835298	-0.05759958	-0.00518689	-0.01456670

Let $Y_1, Y_2,\ Y_3$ and Y_4 be the three principal components. For convenient interpretation, the eigenvectors from R output are multiplied by -1. Then we have

$$Y_1 \approx 0.414z_{runs} + 0.411z_{hits} + 0.391z_{doubles} + .326z_{homeruns} + .374z_{walks} + 0.34z_{outs} + linear\ comb\ of\ other\ vars(small)$$

$$Y_2 \approx -.644z_{bavg} - .616z_{obpage} + .36z_{outs} + linear\ comb\ of\ other\ vars(small)$$

$$Y_3 \approx .35z_{runs} - .92z\_errs + + linear\ comb\ of\ other\ vars(small)$$

$$Y_4 \approx 0.41z_{baverge} + 0.46z_{runs} - .675z_{walks} + linear\ comb\ of\ other\ vars(small)$$

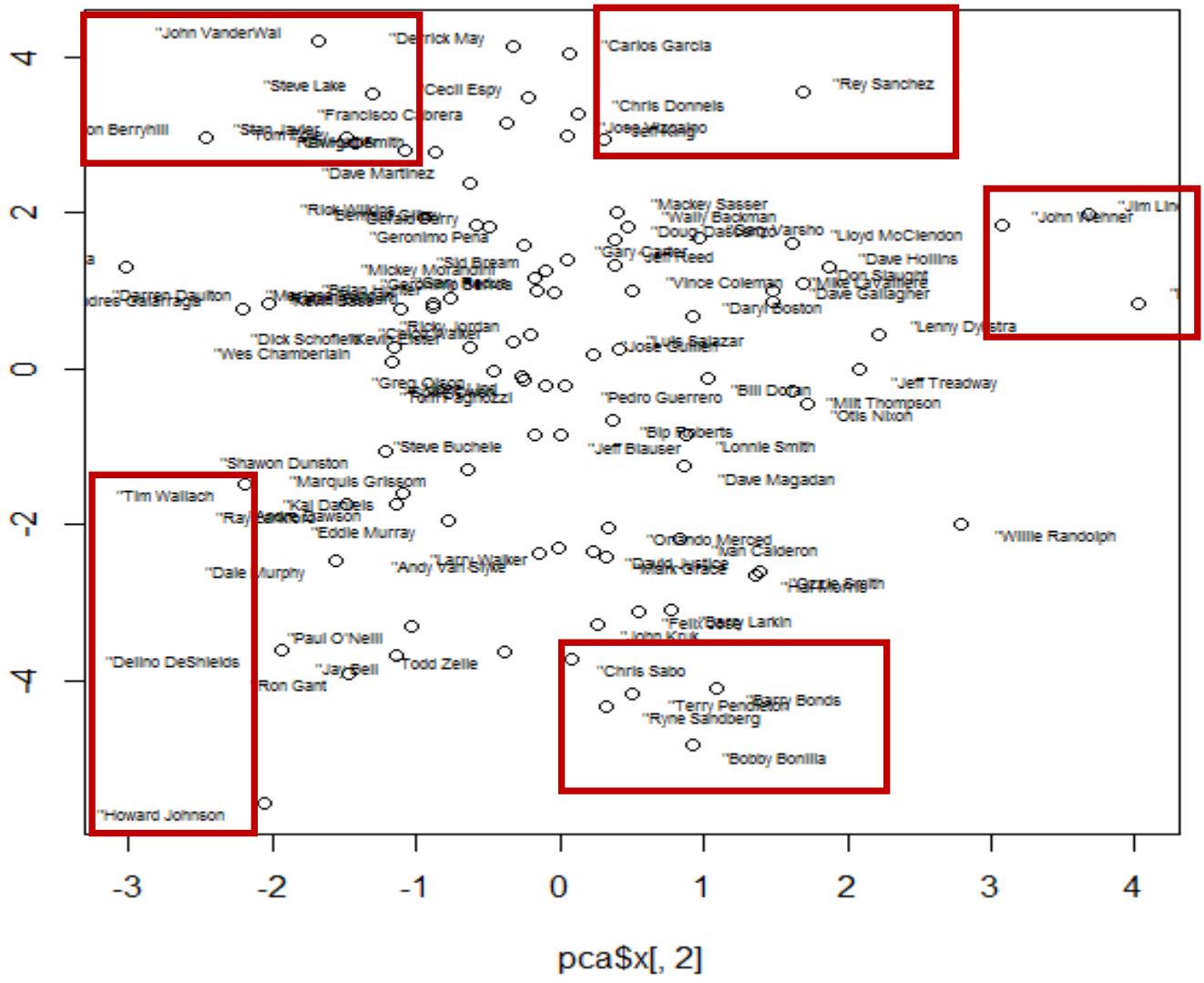
The first principal component is correlated with six of the standardized variables. The first principal component increases with increasing number of runs, number of hits number of doubles, number of double, number of walk and number of strike outs (strange here). This suggests that these six criteria vary together. If one increases, then the remaining ones tend to as well. This component can be viewed as weighted average of various successes.

The second principal component increases with decreasing batting average and on base percentage. This component can be viewed as another measure of unsuccessful at bat: how lower values of batting average and on base percentage. (Cody label it as the batting ability, Brandon labels it as batting performance). Since the sign is arbitrary for solving the eigenvectors, we might ignore the sign if the interpretation is more convinence.

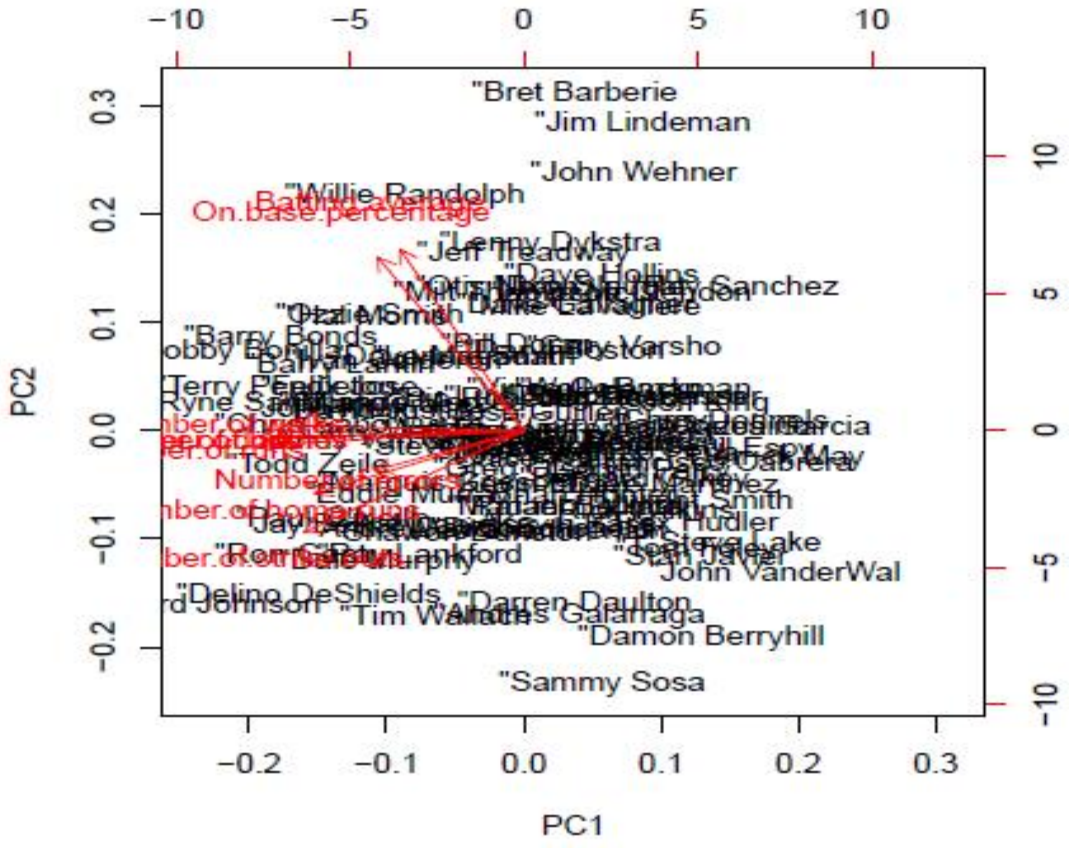
The third principal component increases with increasing decreasing number of errors.

The fourth principal component increase with decreasing the number of walks and increasing number of runs and batting average.

e. (5 pts) Are there any players that are similar or different from each other in any aspects?



We can see players that are similar and different according to PC1 and PC2. The players that are close to each other are similar regrading to the first two PCs. For instance, the players in the bottom box have higher scores in Y_1 and lower scores for Y_2. They are very success at bat. The players in the left top box have lower in Y_1 and higher scores for Y_2. They are not very success at bat.
We might draw biplot to find the answers for the question. Unfortunately, the biplot is not very clear to identify the players.



f. (6 pts)What are those characteristics that is worth to be consider the best predictors of the players' salary of the ABC baseball team?

According to the linear regression output from R. The PC1, 3, 6, 7 and 8, which are orthogonal to each other, are the best predictors of the players' salary. This prediction only explains about 54% of variation in salary. There are another 46% of variation in salary can be explained by unmeasured lurking variables.

```
Call:
lm(formula = Salary ~ pca$x, data = baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-1720.33  -456.32   -58.07   417.23  3117.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1334.050     85.287  15.642  < 2e-16 ***
pca$xPC1    -319.253     37.192  -8.584 2.52e-13 ***
pca$xPC2       3.526     66.793   0.053 0.958012
pca$xPC3    -304.358     97.718  -3.115 0.002470 **
pca$xPC4      67.208    123.313   0.545 0.587090
pca$xPC5    -262.697    149.270  -1.760 0.081825 .
pca$xPC6     686.507    183.067   3.750 0.000312 ***
pca$xPC7     901.106    225.611   3.994 0.000132 ***
pca$xPC8     982.877    371.281   2.647 0.009580 **
pca$xPC9     249.943    417.779   0.598 0.551164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 852.9 on 90 degrees of freedom
Multiple R-squared:  0.5798,    Adjusted R-squared:  0.5377
F-statistic:  13.8 on 9 and 90 DF,  p-value: 1.115e-13
```

If we remove Bobby from the data set, the regression prediction has been slightly improved based on the following output from R. The PC1, 3, 5, 6, 7 and 8 are significant for the prediction.

```
Call:
lm(formula = Salary ~ pca$x, data = baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-1512.61  -398.87   -76.32   375.91  2110.39

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1285.91      77.99  16.488  < 2e-16 ***
pca$xPC1     -284.32      34.17  -8.321 9.51e-13 ***
pca$xPC2      -12.72      60.76  -0.209 0.834701
pca$xPC3     -307.42      88.62  -3.469 0.000808 ***
pca$xPC4       65.66     110.76   0.593 0.554787
pca$xPC5     -310.44     134.87  -2.302 0.023686 *
pca$xPC6      547.74     172.98   3.166 0.002114 **
pca$xPC7      961.23     202.75   4.741 8.03e-06 ***
pca$xPC8     1055.22     333.64   3.163 0.002138 **
pca$xPC9      557.39     381.32   1.462 0.147331
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 776 on 89 degrees of freedom
Multiple R-squared:  0.5966,    Adjusted R-squared:  0.5558
F-statistic: 14.62 on 9 and 89 DF,  p-value: 2.97e-14
```

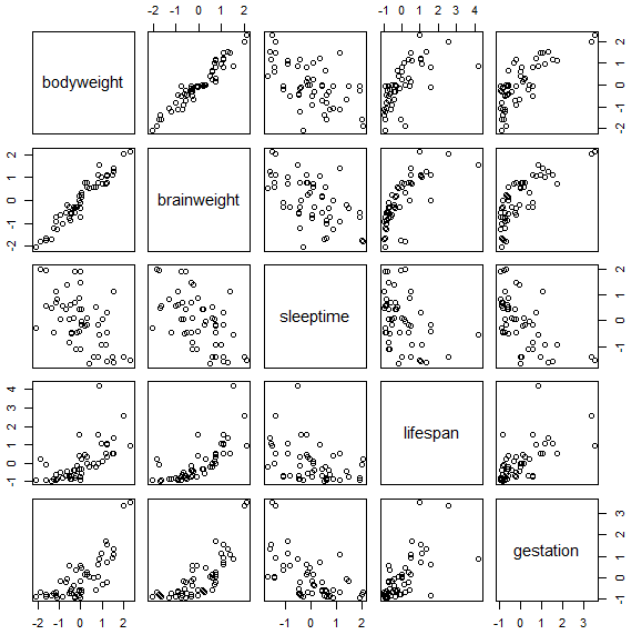
2. The researchers who study of mammals collected a data set on 52 mammals. The variables measured include weight measurements of mammals (body weight and brain weight) and characteristics measurements of mammal (total daily sleep, maximum lifespan, and gestation time). The data file contains the observed values of the 6 variables (plus a labeling column with the names of the species) for 52 mammals. The data set is named mammals.csv. The units of measurements are given below

- species of animal, body weight in kg, brain weight in g, total sleep (hrs/day), maximum life span (years)
- gestation time (days)

NOTE: It is recommended by the ecologists that you use a natural log transformation of the body weight and brain weight variables before doing the analysis.

The questions that the ecologists would like answered include:

(1) (5 pts) Are there notable associations/relationships between some of the variables? (if so, describe them)




```
> pairs(mammals)
> c<-cor(mammals)
> c
             bodyweight brainweight  sleeptime   lifespan  gestation
bodyweight    1.0000000   0.9603793 -0.5512021  0.6476452  0.7711456
brainweight   0.9603793   1.0000000 -0.5644951  0.7247805  0.7791989
sleeptime    -0.5512021  -0.5644951  1.0000000 -0.3784267 -0.5895245
lifespan      0.6476452   0.7247805 -0.3784267  1.0000000  0.6394415
gestation     0.7711456   0.7791989 -0.5895245  0.6394415  1.0000000
```

Both scatter plot matrix and correlation matrix show that

- There is a very strong positive relationship between body weight and brain weight, as might be expected.
- There is also a moderately strong association between bodyweight, lifespan, brain weight, and gestation.
- There is a moderately negative association between sleeptime to all other variables, which may more interesting in the study.

(2) Considering the data set as two groups of data sets: Weight measurement (log body weight and log brain weight) and characteristics measurement (total sleep, life span, and gestation time).

a. (5 pts) Test for independence between two sets of variables.

Even although the correlation matrix and scatter plot matrix show the correlation among some variables, the very first thing for canonical analysis is to determine if there is any relationship between the two sets of variables at all.

- $H0: \rho_1^* = \rho_2^* = 0$, H_a : at least one of them is not zero.

$\emptyset_0^2 = 65.7, df = 6$, which is chi-square distributed with 8 degrees of freedom. We reject null hypothesis if

$$\emptyset_0^2 > \chi_8(0.05)$$

$p - value$ =3.08984-12<0.05 Reject null hypothesis.

$$H_0: \rho_1^* \neq 0, \rho_2^* = 0, \qquad v.s. \qquad H_a: \rho_2^* \neq 0$$

$\emptyset_0^2 = 4.6588, df = 2$, which is chi-square distributed with 2 degrees of freedom. We fail to reject null hypothesis if

$$\emptyset_0^2 > \chi_3(0.05)$$

$p - value$ =0.097>0.05 do not reject null hypothesis.

b. (5 pts) Determine the number of significant canonical variate pairs, if null hypothesis is rejected in part a.

There is only one pair canonical variate which is statistically significant.

c. (5 pts) Compute the correlation between the canonical variate pairs, if null hypothesis is rejected in part a.

Because the eigenvalues of E1 are 0.71175714 and 0.09249827

$$Cov(u_1, v_1) = \sqrt{0.7117} = 0.843, \qquad Cov(u_2, v_2) = 0.304$$

d. (5 pts) Compute the canonical variates from the data.

Method 1: use the method from textbook, that is, the eigenvalues based on $R_{11}^{-1}\ R_{12}\ R_{22}^{-1}\ R_{21}$

$\widehat{U_1} = 0.127z_1 - 0.9918z_2$

```
> (e1 <- eigen(E1))
$values
[1] 0.71175714 0.09249827

$vectors
           [,1]       [,2]
[1,]  0.1275204 -0.7253665
[2,] -0.9918359  0.6883629
```

$\widehat{V_1} = 0.266\, z'_1 - .658z'_2 - 0.703z'_3$

```
> (e2 <- eigen(E2))
$values
[1] 7.117571e-01 9.249827e-02 2.848741e-17

$vectors
           [,1]       [,2]       [,3]
[1,]  0.2664204  0.03251798 -0.84458354
[2,] -0.6584091 -0.71514463 -0.09045225
[3,] -0.7039301  0.69821969 -0.52772818
```

Method 2: use the method from text by Richard A. Johnson, that is the eigenvectors based on $\sum_{11}^{-\frac{1}{2}}\sum_{11}\sum_{22}^{-\frac{1}{2}}\sum_{21}\sum_{11}^{-\frac{1}{2}}$ with respecting the original variables Xs, not the standardized variables Zs.

R output

```
$xcoef
              [,1]       [,2]
bodyweight  -0.04558196  1.115028
brainweight  0.44639041 -1.332317

$ycoef
             [,1]          [,2]
sleeptime -0.040675033  0.011670760
lifespan   0.024972702 -0.063764449
gestation  0.003536027  0.008245048
```

$\widehat{U_1} = -0.046x_1 + 0.446x_2$

$\widehat{V_1} = -0.041y1 + 0.025y2 + 0.004y3$

SAS output

Raw Canonical Coefficients for the Weight Variables

	weight1	weight2
bodyweight	-0.045581964	-1.115028032
brainweight	0.4463904083	1.3323165265

Raw Canonical Coefficients for the Living Scores

	living1	living2
sleeptime	-0.040675033	-0.01167076
lifespan	0.0249727021	0.0637644494
gestation	0.0035360271	-0.008245048

Or

$\widehat{U_1} = 0.147z_1 - 1.1399z_2$

$\widehat{V_1} = 0.1910\, z'_1 - 0.4719z'_2 - 0.5045z'_3$

Standardized Canonical Coefficients for the Weight Variables

	weight1	weight2
bodyweight	-0.1466	-3.5851
brainweight	1.1399	3.4022

Standardized Canonical Coefficients for the Living Scores

	living1	living2
sleeptime	-0.1910	-0.0548
lifespan	0.4719	1.2050
gestation	0.5045	-1.1764

Compare with the method 1 from textbook, there are slight differences in the coefficients.

$\widehat{U_1} = 0.127z_1 - 0.9918z_2$

$\widehat{V_1} = 0.266\, z'_1 - .658z'_2 - 0.703z'_3$

e. (5 pts) Interpret each member of a canonical variate pair using its correlations with the member variables.

The correlations between the weight variables and the canonical variables for living variable are found at the top of the fourth page of the SAS output in the following table:

R output

```
$scores$corr.X.xscores
           [,1]       [,2]
[1,] 0.9481943 0.31769091
[2,] 0.9991655 0.04084554
```

SAS output

Correlations Between the Weight Variables and Their Canonical Variables

	weight1	weight2
bodyweight	0.9482	-0.3177
brainweight	0.9992	-0.0408

Looking at the first canonical variable for weight, we see that two correlations are uniformly large. Therefore, you can think of this canonical variate as an overall measure of living characteristics. For the second canonical variable for weight, none of the correlations is particularly large, and so, this canonical variable yields little information about the data. Again, we had decided earlier not to look at the second canonical variate pairs.

A similar interpretation can take place with the living characteristics scores.

R output

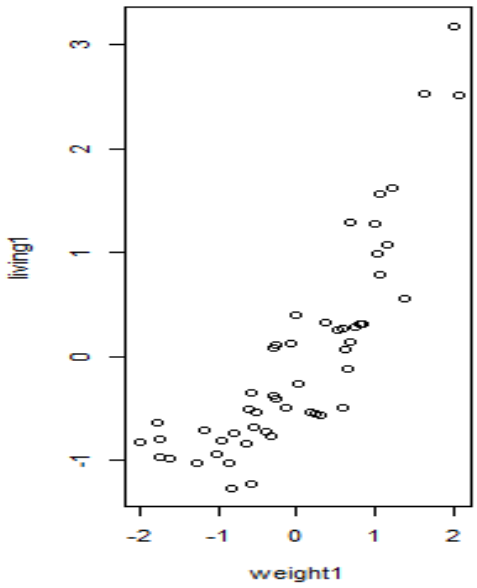
```
$scores$corr.Y.yscores
           [,1]       [,2]
[1,] -0.6669704 -0.1827614
[2,]  0.8667882 -0.4734240
[3,]  0.9188619  0.3736384
```

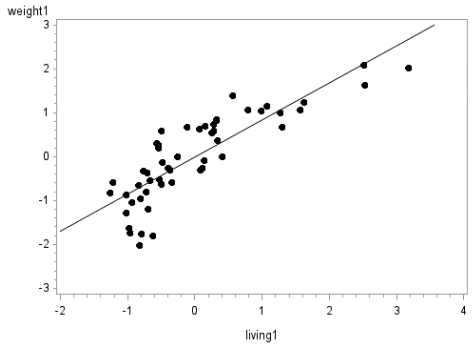
SAS output

Correlations Between the Living Scores and Their Canonical Variables

	living1	living2
sleeptime	-0.6670	0.1828
lifespan	0.8668	0.4734
gestation	0.9189	-0.3736

All correlations are large for the first canonical variable of living characteristic, and this can be thought of as an overall measure of characteristics as well. However, it is negative correlated with sleeptime, which is consistent with scatter plot matrix and correlation matrix in part a.






```
x <- cbind(mammals$bodyweight,mammals$brainweight)

y <- cbind(mammals$sleeptime,mammals$lifespan,mammals$gestation)

cca<-CCA::cc(x,y)

weight1<-cca$scores$xscores[,1]

character1<-cca$scores$yscores[,1]

plot(weight1,character1)
```