|  |
| --- |
| Simon Fraser University |
| Assignment 1 |
| STAT 445 |
|  |
| **Kun Yang 301178299** |
| **1/17/2014** |

|  |
| --- |
|  |

**STAT 445/645 Assignment #1**

1. You will be provided with a slightly altered version of the file on major league baseball statistics. Convert the file into an appropriate format, and read it into R.

### Read the file on major league baseball statistics into R

**setwd("C:\\Users\\Kun\\Desktop\\homework")**

**getwd()**

**MLB<-read.csv("Major League Baseball Main Stats Altered 2014.csv")**

**MLB**

1. Use R to calculate
   1. the mean vector

**(round(colMeans(MLB[,2:6]),3))**

* 1. the variance-covariance and correlation matrices.

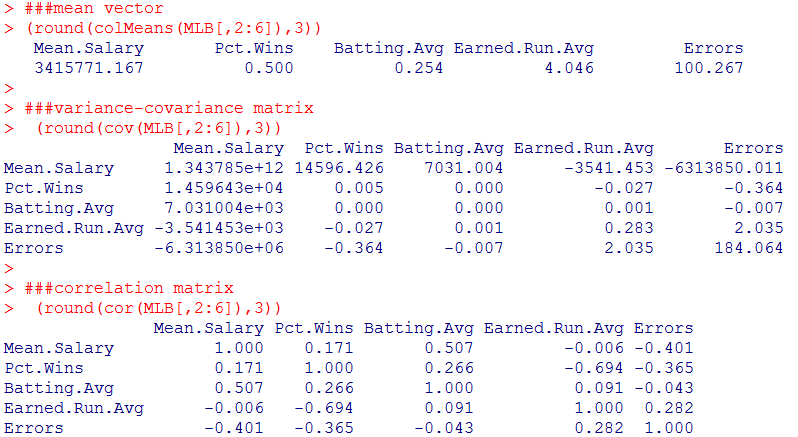
###variance-covariance matrix

**(round(cov(MLB[,2:6]),3))**

###correlation matrix

**(round(cor(MLB[,2:6]),3))**

Print all of the above with appropriate text labels and the summary statistics rounded off to 3 decimal points of accuracy.



1. For the same data, produce all pairs of two-variable scatterplots, but with the entries in the lower-left triangle below the diagonal containing estimates of the correlation coefficients as demonstrated in class.

###all pairs of two-variable scatterplots

**pairs(MLB[2:6])**

###function panel.cor

**panel.cor <- function(x, y, digits=3, prefix="", cex.cor)**

**{**

**usr <- par("usr"); on.exit(par(usr))**

**par(usr = c(0, 1, 0, 1))**

**r <- abs(cor(x, y))**

**txt <- format(c(r, 0.123456789), digits=digits)[1]**

**txt <- paste(prefix, txt, sep="")**

**if(missing(cex.cor)) cex <- 0.8/strwidth(txt)**

###hypothesis testing with null hypothesis: "true correlation is equal to zero"

**test <- cor.test(x,y)**

**Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,**

**cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),**

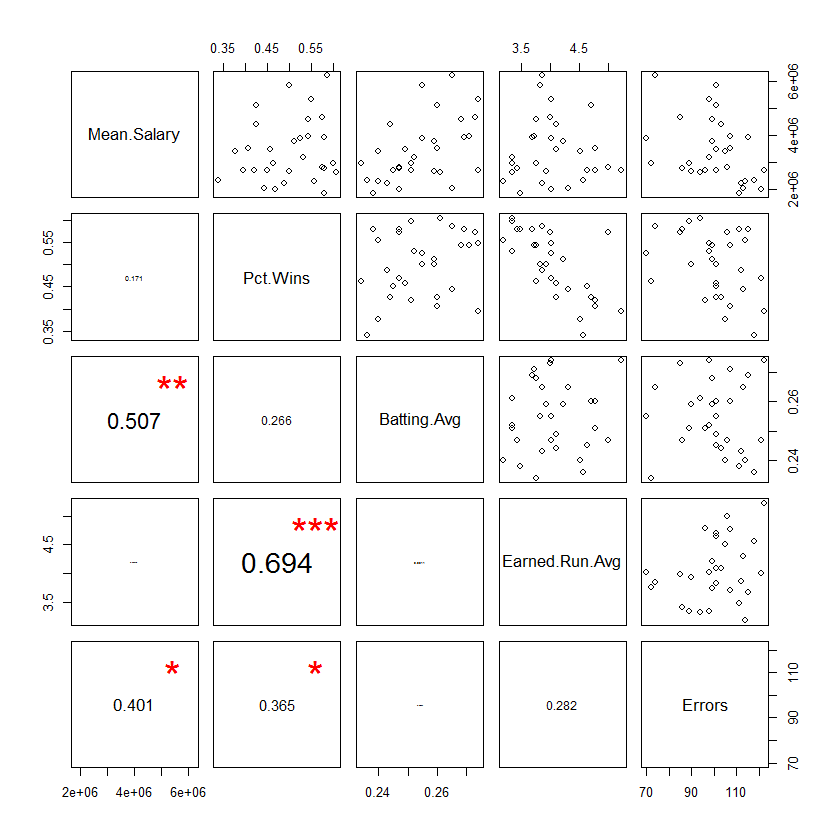
**symbols = c("\*\*\*", "\*\*", "\*", ".", " "))**

**text(0.5, 0.5, txt, cex = cex \* r)**

**text(.8, .8, Signif, cex=cex, col=2)**

**}**

**pairs(MLB[2:6],lower.panel=panel.cor)**



1. Use R to plot boxplots for each of the variables in this dataset. Do these plots highlight any apparent outliers in this dataset?

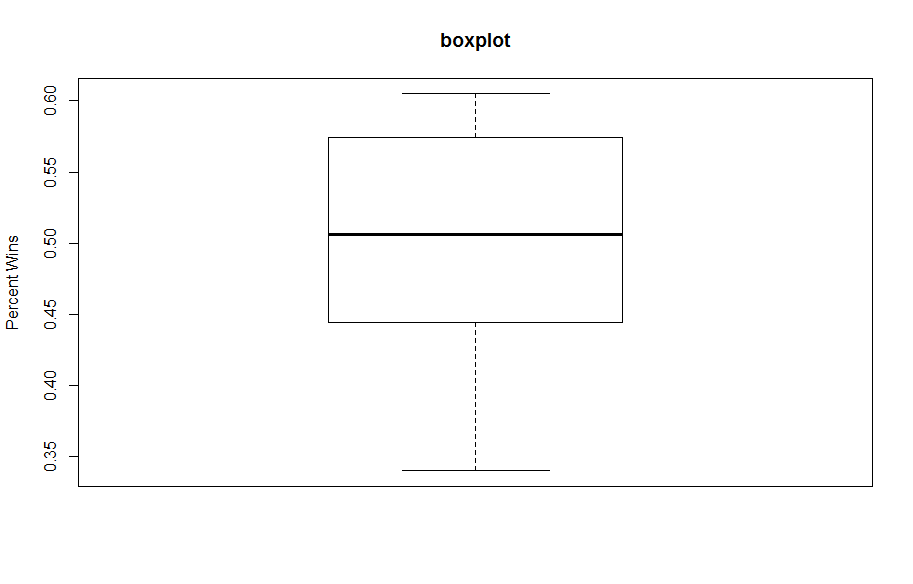
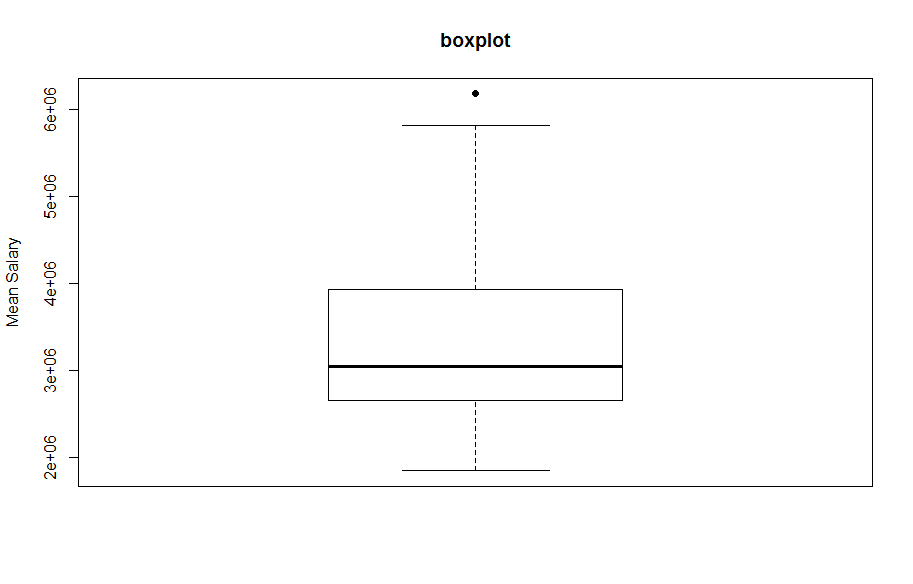
**boxplot(MLB[,2], pch=19, main="boxplot", ylab="Mean Salary")**

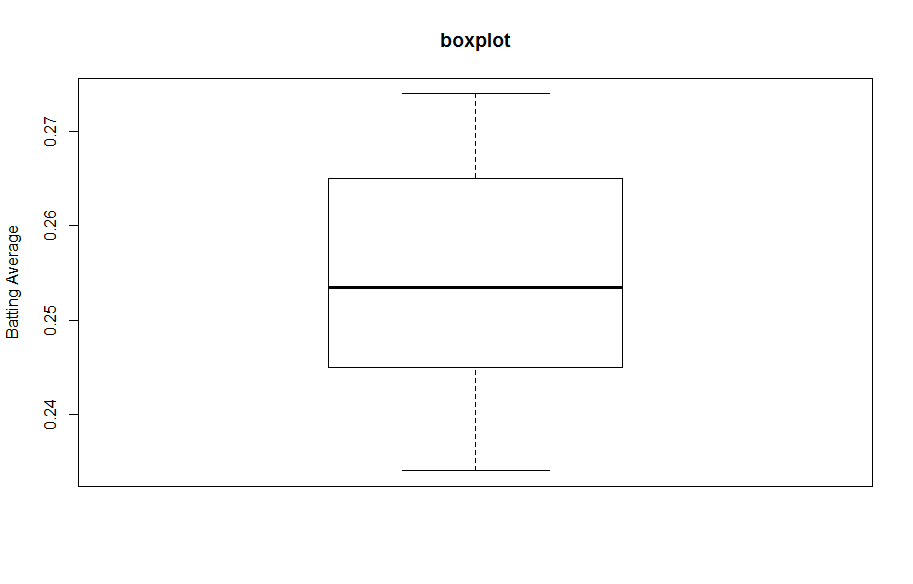
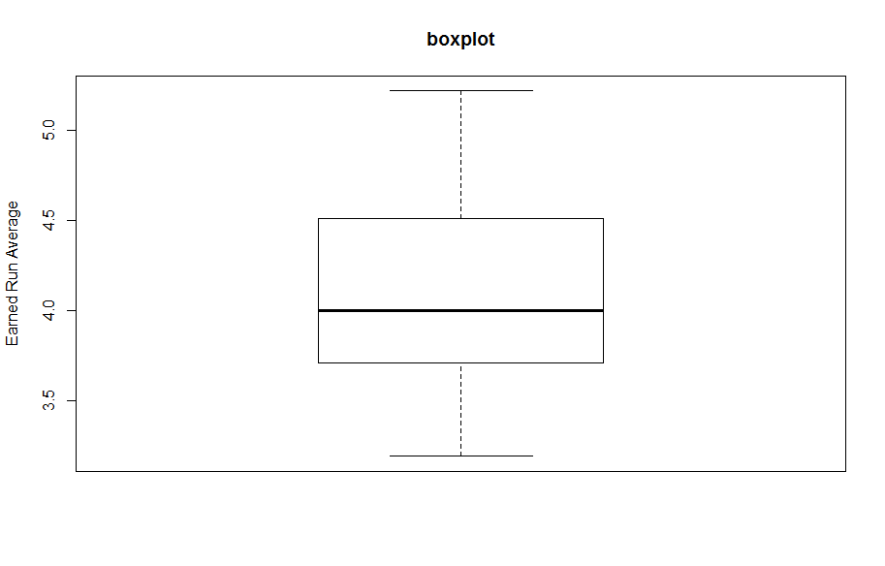
**boxplot(MLB[,3], main="boxplot", ylab="Percent Wins")**

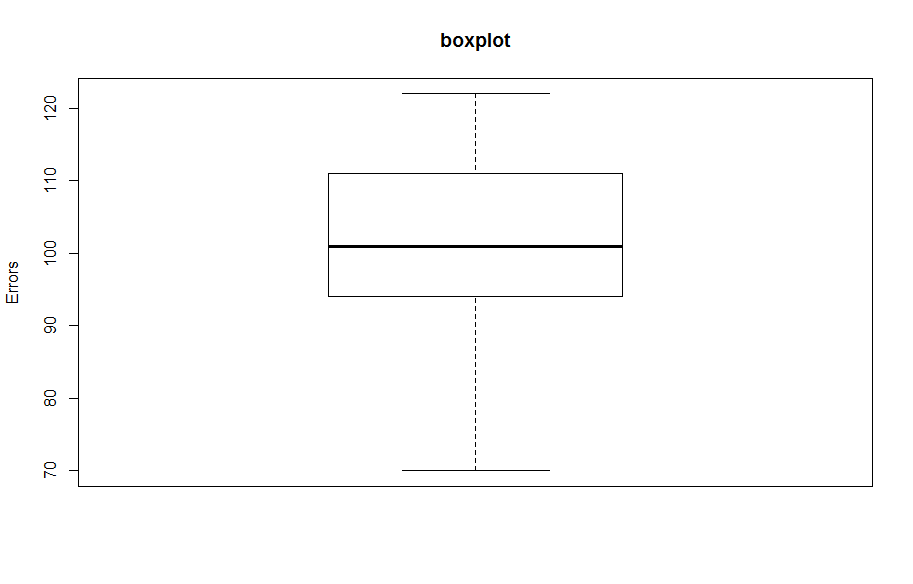
**boxplot(MLB[,4], main="boxplot", ylab="Batting Average")**

**boxplot(MLB[,5], main="boxplot", ylab="Earned Run Average")**

**boxplot(MLB[,6], main="boxplot", ylab="Errors")**







1. Using the function, “bvbox” in the R package, MVA, construct a two-dimensional boxplot for the earned run average and percent wins. Does this plot highlight any potential outliers? If so, for which team(s)?

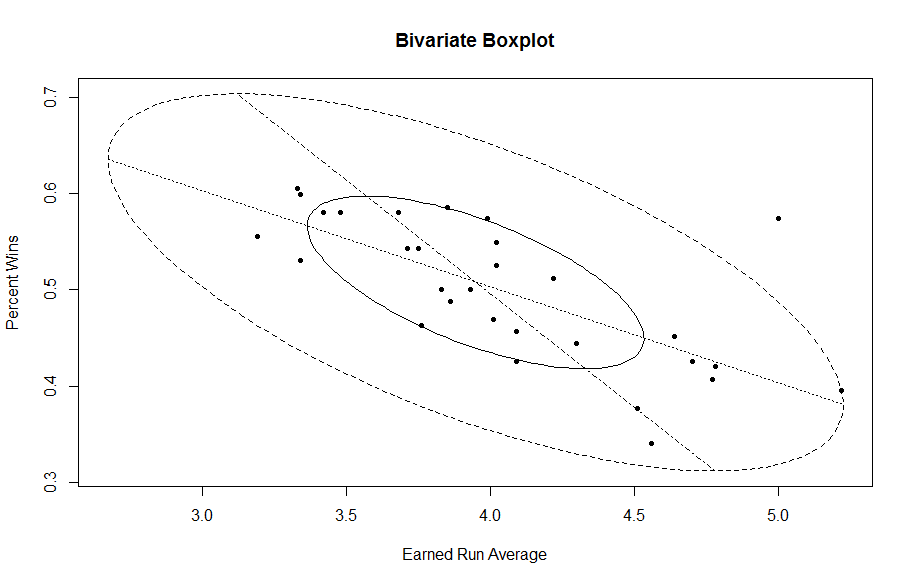
### bivariate boxplot.

**library("MVA")**

**X<-MLB[5]**

**Y<-MLB[3]**

**bvbox(cbind(X, Y),7, pch=20,main="Bivariate Boxplot", xlab="Earned Run Average",ylab="Percent Wins")**



1. Are there any differences in the potential outliers identified in the univariate vs. the bivariate boxplots? If so, explain why you think this happened.

We find a potential outlier from the univariate boxplot of mean salary. Checking the data, we get that “New York Yankees” is the outlier. However, from the bivariate boxplots, we notice that “Baltimore Orioles” is the outlier.

This happened because univariate outliers are extreme values based on single variable, but bivariate outliers are extreme values based on the combination of two variables. Let’s go back into this case. NY Yankees has an extreme value based on single variable--the mean salary. But if we consider two variables, earned run average & percent wins, we find that NY Yankees has a low earned run average and high percent wins. So Yankees are not outliner in bivariate boxplots. But Baltimore Orioles has extreme value this time, which means it has a high earned run average but the team’s percent wins are NOT low. So, Baltimore Orioles are outlier in bivariate boxplots even though it does not have an extreme value in mean salary (or other single variable).

1. The univariate boxplot style
   1. For a normal distribution with mean 2 and standard deviation 1, find the values of the median and the two quartiles.

> ### the median

**> qnorm(0.5, mean=2, sd=1)**

**[1] 2**

> ### the two quartiles

**> qnorm(0.25, mean=2, sd=1)**

**[1] 1.32551**

**> qnorm(0.75, mean=2, sd=1)**

**[1] 2.67449**

* 1. Again for the above normal distribution, find the numerical values of the end-points of the two whiskers.

**> IQR<-qnorm(0.75, mean=2, sd=1)-qnorm(0.25, mean=2, sd=1)**

**> qnorm(0.75, mean=2, sd=1)+1.5\*IQR**

**[1] 4.697959**

**> qnorm(0.25, mean=2, sd=1)-1.5\*IQR**

**[1] -0.697959**

* 1. What is the probability that single value drawn from the above normal distribution will fall outside the range of both the central box and both of the whiskers as calculated above from the actual population?

**> lowerw<-qnorm(0.25, mean=2, sd=1)-1.5\*IQR**

**> pnorm(lowerw, mean=2, sd=1)**

**[1] 0.003488302**

**> upperw<-qnorm(0.75, mean=2, sd=1)+1.5\*IQR**

**> 1-pnorm(upperw, mean=2, sd=1)**

**[1] 0.003488302**

**> pnorm(lowerw, mean=2, sd=1)+(1-pnorm(upperw, mean=2, sd=1))**

**[1] 0.006976603**

* 1. What is the probability that you will obtain at least one observation out of 20 beyond the end of either whisker?

**> p<-pnorm(lowerw, mean=2, sd=1)+(1-pnorm(upperw, mean=2, sd=1))**

**> dbinom(1, 20, p)**

**[1] 0.122153**