UNIT 8 HW

Americans love their baseball. Even so, there is a concern that teams with more money to spend on players have more success than teams that have less money to spend. The payroll (X) (in millions of dollars) and the number of games won in the season (out of 162) (Y) are provided in the table below for all of the 30 major league teams. The numbers are from the 2010 regular season. We can use these data to illustrate statistical methods for drawing inferences about correlations.

Team	Payroll	Wins	Team	Payroll	Wins	Team	Payroll	Wins
NYY	206	95	LAD	95	80	KC	71	67
BOS	162	89	HOU	92	76	TOR	62	85
CHC	146	75	SEA	86	61	ARZ	61	65
PHI	142	97	STL	86	86	CLE	61	69
NYM	134	79	ATL	84	91	WAS	61	69
DET	123	81	COL	84	83	FA	57	80
CHW	106	88	BAL	82	66	TEX	55	90
LAA	105	80	MIL	81	77	OAK	52	81
SF	99	92	TB	72	96	SD	38	90
MIN	98	94	CIN	71	91	PIT	35	57

1. Provide a scatterplot of the data using both SAS and R. Looking at the scatterplot, do you expect the correlation to be positive, negative, or close to 0? Why? Is the relationship between team payroll and number of wins strong, moderate, or weak? Is the relationship linear? Take a guess of the value of the correlation coefficient.

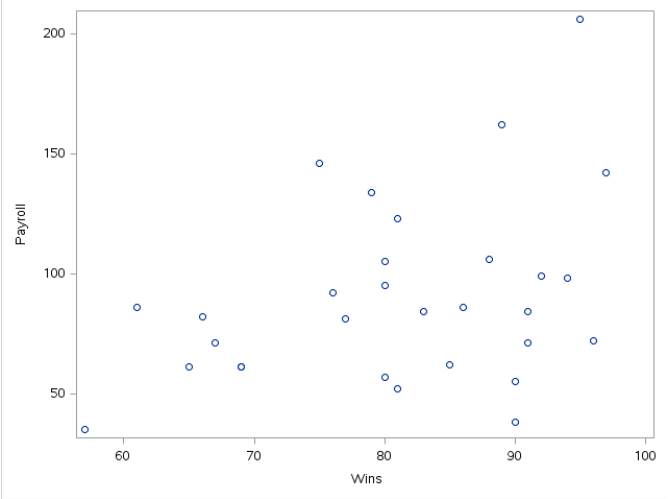
```
proc import datafile = '/home/chec0/New Folder/Baseball_Data.csv'
 out = baseball
 dbms = CSV
 ;

proc print data = baseball;
run;

proc sgscatter data = baseball;
```

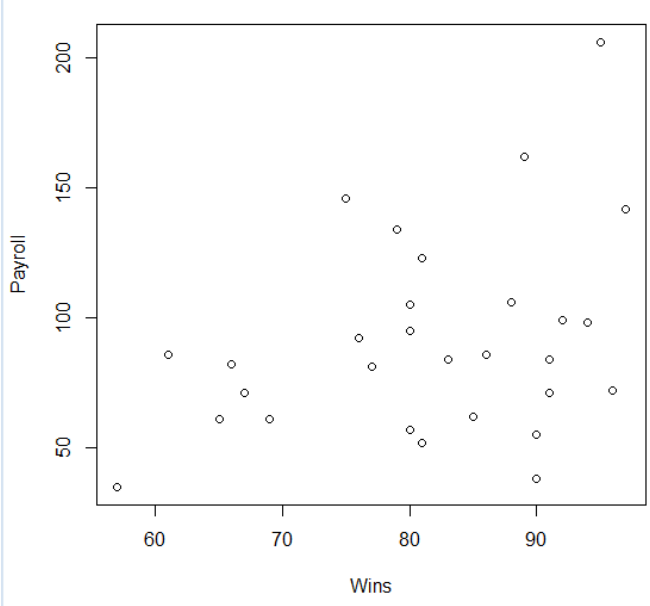
plot Payroll*Wins;

run;



baseball <- read.csv("/Baseball_Data.csv")

plot(baseball$Wins, baseball$Payroll, xlab='Wins', ylab='Payroll')



Positive going with weak relationship between payroll and wins. Not linear and value coefficient around .3

2. Find the correlation between team payroll and the number of wins. (No fair going back and changing your answer to the previous question!) You should do this in both R and SAS.

cor.test(baseball$Payroll, baseball$Wins)

```
        Pearson's product-moment correlation

data:  baseball$Payroll and baseball$Wins
t = 2.0826, df = 28, p-value = 0.04654
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.00686799 0.64181770
sample estimates:
      cor
0.366231
```

proc corr data= baseball;

run;

The CORR Procedure

2 Variables:	Payroll Wins

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Payroll	30	90.23333	38.16812	2707	35.00000	206.00000
Wins	30	81.00000	11.00470	2430	57.00000	97.00000

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

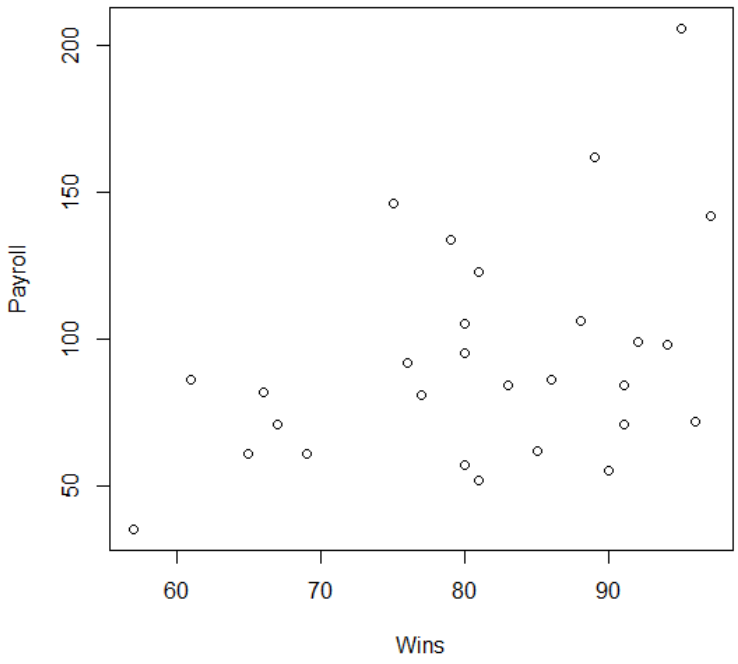
	Payroll	Wins
Payroll	1.00000	0.36623 0.0465
Wins	0.36623 0.0465	1.00000

R correlation is .36623

3. San Diego (SD) has a payroll of $38 million, yet SD has 90 wins – more than Boston does (with a payroll of $162 million). Delete SD from the data and rerun the analysis (scatter plot and correlation value). How does the correlation change? You may use your preference here, R or SAS.

```
        Pearson's product-moment correlation

data:  baseball$Payroll and baseball$Wins
t = 2.4435, df = 27, p-value = 0.02136
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06995422 0.68518874
sample estimates:
      cor
0.4255494
```



Correlation coefficient changed to .42555 so the relationship got stronger without San Diego data.

4. The league commissioner notes that the Texas Rangers (TEX), with one of the lowest payrolls, won 90 games (and were the American League Champions) and the Chicago Cubs (CHC), with the third highest payroll, won only 75 games. He argues that this proves that there is no advantage to teams with a higher payroll. Comment on his argument.

   There is a lot of other variables that can go into wins like injuries for instance, having the highest payroll just shows that you can afford to pay the better players more. Younger players coming into the league could be great players but aren't paid like it yet.

5. What is the population for these data? Can these data be considered a random sample from that population?

   The population is the major league baseball players for the respective teams. It is not a random sample since it is dealing with all the players for each team