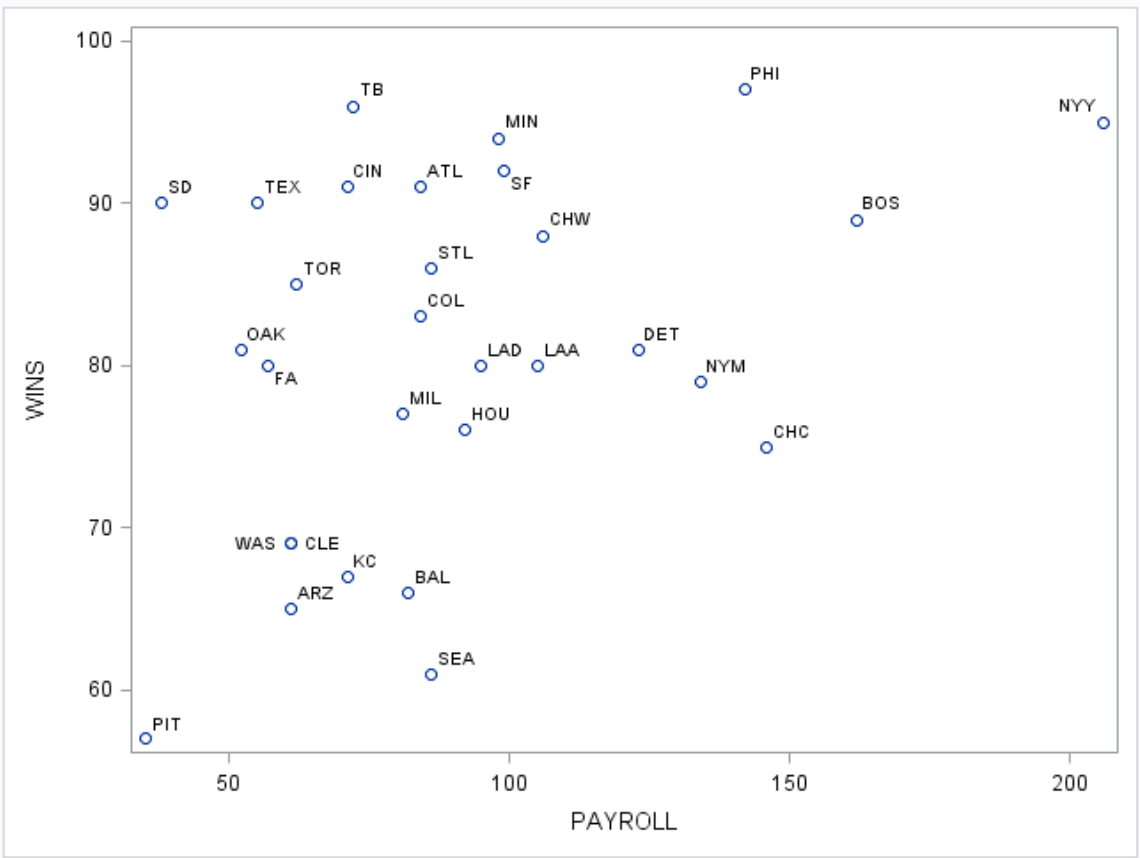
Name: James Tsai
Section: MSDS6371-401
Date: 10/24/15

1. **Draw a scatterplot of the data using software (required!). 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 at the value of the correlation coefficient.**




```
DATA BASEBALL;
   INFILE 'C:\Users\james\Documents\My SAS Files\9.4\Baseball_Data.csv' DLM=','
FIRSTOBS=2;
   INPUT TEAM $ PAYROLL WINS;
RUN;

PROC SGPLOT DATA=BASEBALL;
   SCATTER X=PAYROLL Y=WINS / DATALABEL=TEAM;
RUN;
```

- The correlation looks to be positive as we can draw a positive slope through the data points on the scatterplot.

- The relationship looks weak, as the data points don't really form a cohesive line. Around the 80 and 90 wins, the payroll doesn't seem to have much of an effect as we can almost draw a straight line across.
- The relationship looks slightly linear, but it's very hard to tell visually.
- My guess would be around 0.25.

2. **Find the correlation between team payroll and the number of wins – no fair going back and changing your answer to the previous question! (SAS Note: you can get correlations from SAS using PROC CORR; VAR X1 X2 X3;, where X1, X2, and X3 are the variables for which you want correlations. This does all combinations of pairwise correlations.)**

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

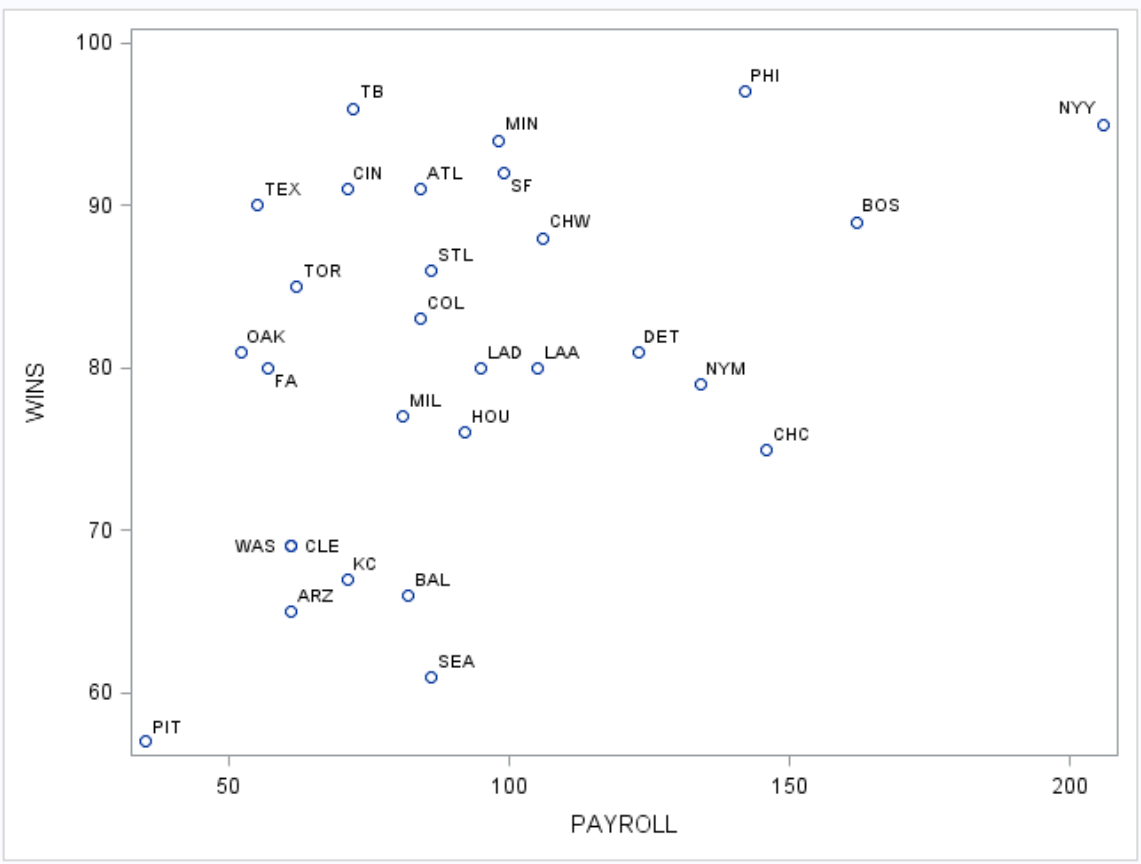
Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	PAYROLL	WINS
PAYROLL	1.00000	0.36623 0.0465
WINS	0.36623 0.0465	1.00000

```
PROC CORR DATA=BASEBALL;
   VAR PAYROLL WINS;
RUN;
```

- The correlation coefficient calculated by SAS is: 0.36623.

3. **San Diego (SD) has a payroll of 38 million, yet has 90 wins – more than Boston does. Delete SD from the data and rerun the analysis. You can use SAS. How does the correlation change?**



The CORR Procedure

2 Variables:	PAYROLL WINS

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PAYROLL	29	92.03448	37.52379	2669	35.00000	206.00000
WINS	29	80.68966	11.06508	2340	57.00000	97.00000

Pearson Correlation Coefficients, N = 29
Prob > |r| under H0: Rho=0

	PAYROLL	WINS
PAYROLL	1.00000	0.42555 0.0214
WINS	0.42555 0.0214	1.00000

```
PROC SGPLOT DATA=BASEBALL;
   SCATTER X=PAYROLL Y=WINS / DATALABEL=TEAM;
   WHERE (TEAM ^= "SD");
RUN;

PROC CORR DATA=BASEBALL;
   VAR PAYROLL WINS;
   WHERE (TEAM ^= "SD");
RUN;
```

- The correlation coefficient calculated by SAS is: 0.42555. The correlation has increased substantially after removing “SD” from the data set.

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

   - His argument is weak because he is trying to infer causality from only two data points. He is purposely picking data points to try to prove his point. We can argue for example, if we chose another pair of teams, ATL vs. CLE, the higher payroll does result in more wins. However it would not be a statistically sound statement. In conclusion, while it is most likely there are a lot more factors than just payroll that affects performance, he does not prove his case conclusively.

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

   - This would be the whole population for the entire payroll in the major league baseball.
   - This data cannot be considered as random sample as it is only from the 2010 regular season and can be considered one slice of the data. Since randomization is not involved, we cannot infer causality and can only apply the conclusion and scope to the 2010 season based on a 5% p-value.