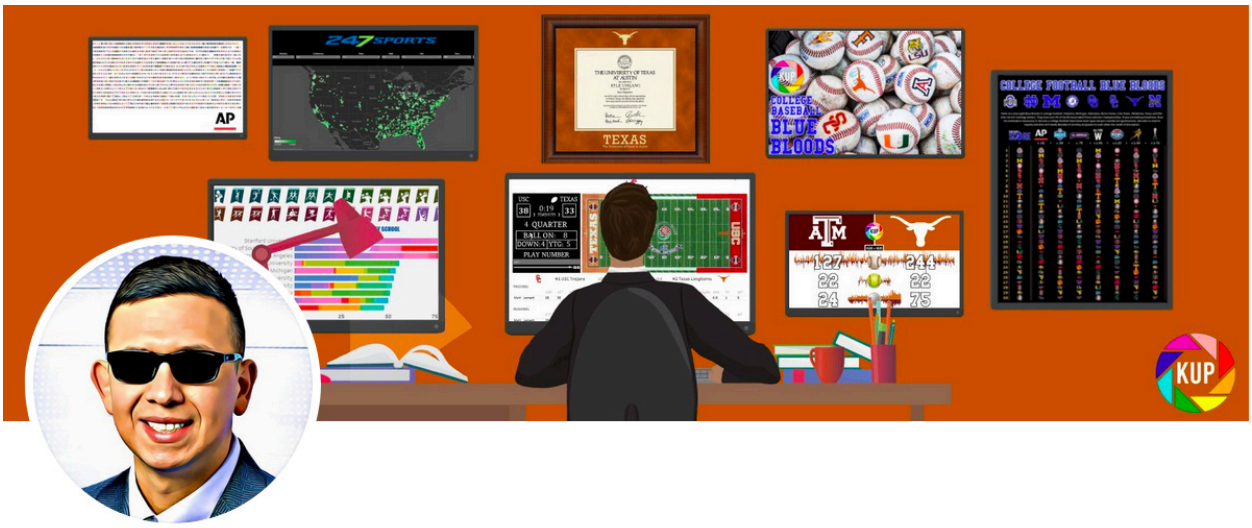
College Football
Blue Bloods

Kyle Umlang
IST 687 | Applied Data Science | 2019-0703

Table of Contents

Introduction

I have an obsession with College Football. It has gone from the simple joy of watching weekly games of my favorite sports team (The Texas Longhorns) to a serious sports analytics hobby on Twitter and hopefully one day, a profession.



Kyle Umlang
@kyleumlang

Data Analyst | Graphic Designer | Dad | College Football Stats Expert | Longhorn | MS Data Science - Syracuse '21 | Mr. Data | MOF

bit.ly/TableauKU Born September 3, 1985 Joined July 2011

210 Following **5,550** Followers

One topic that has always fascinated me is the subject of "Blue Bloods", which is a list of schools that – in different eras – have taken turns dominating the sport of college football. There actually aren't that many that have been consistently great over time, so the list is very short.

Most college football stats enthusiasts have a good sense for which schools belong in that category, but I'm hoping to use R to help confirm which schools should be considered "Blue Bloods".

If you research "College Football Blue Bloods", you will most likely find the same 8 schools in the discussion: Alabama, USC, Texas, Oklahoma, Michigan, Ohio State and Notre Dame.

ESPN published an article in 2016 where it listed their Blue Blood choices:

BLUE BLOODS

T1. Alabama Crimson Tide | 10

The Tide own 11 national titles (including the most recent one). They play in the ever-competitive SEC, where they have won 25 conference titles, 12 more than the next closest team. Alabama also has more bowl wins (35) than any team in the country.

T1. Notre Dame Fighting Irish | 10

It's hard to argue with the highest all-time winning percentage in all of college football, which is the perch on which Notre Dame sits. At 892-313, the Irish have won more than 73 percent of their games. They have also won eight national championships and generated seven Heisman Trophy winners.

T1. Ohio State Buckeyes | 10

Ohio State also has seven Heisman Trophies, including the only two-time winner in running back Archie Griffin (1974, 1975). The Buckeyes have six titles to their names, half of which were won by Woody Hayes, who led the program for 28 seasons. The Buckeyes have won 35 Big Ten titles.

T1. Oklahoma Sooners | 10

The Sooners have won 72 percent of their games since taking the field in 1895, and they've won 75 percent of their conference games. With seven national titles and 14 undefeated seasons, Bob Stoops' program is a lock for blueblood status.

T1. USC Trojans | 10

The Trojans would be part of the seven-Heisman club were it not for Reggie Bush's vacated trophy. They have won seven national titles and boast 33 bowl wins, the No. 2 mark in the country behind Alabama.

6. Michigan Wolverines | 9.92

The Wolverines fall short of their fellow blue bloods in national championships with only two. But while Notre Dame has the best winning percentage, Michigan has won more games overall by a margin of 33. Current coach Jim Harbaugh will try to widen the gap.

7. Texas Longhorns | 9.83

The Longhorns have won four titles and are one of eight teams to have won 70 percent or more of their games all-time. Texas won or shared 25 conference championships during its time in the Southwest Conference but has only won three in the Big 12.

8. Nebraska Cornhuskers | 9.5

Relatively speaking, Nebraska has enjoyed more recent success. The first of the Cornhuskers five national championships came in 1970. They went back to back in 1970-71 and again in 1994-95. They've also produced three Heisman winners including 2001's winner, quarterback Eric Crouch.

The overall objective of this analysis is to determine if these specific 8 schools are consistently in the Top in the following categories: SRS, SOS, AP, Win%, Conf Win% and Point Differential ll-Time. In doing so, I'm hoping to confirm if they deserve the title of "Blue Blood". Using Sports Viz Sunday's monthly sports related dataset about College Football and a combination of factors used in determining past Blue Bloods lists, I will answer the following questions:

Business Questions

My main objective was to determine which of the 8 Blue Bloods did/did not deserve the title based on the following questions:

- Which schools lead in overall wins and overall conference wins?
- Which schools lead in number of AP ranks?
- Which schools lead in overall win percentage and conference win percentage?
- Which schools lead in overall avg point differential?
- Which schools lead in SOS and Avg SRS?
- Which schools lead over all for all categories combined and do they match the Blue Bloods named by ESPN

Data Acquisition, Cleansing, Transformation, Munging

Data Acquisition Process

First off, it was difficult getting my hands on such a great file full of CFB metrics, but luckily for me Sports Viz Sunday recently had a month dedicated to College Football and they had a great dataset to use for free with registration to the Data World website. Their Combined Seasons dataset had everything I needed in it for my analysis. The only column I removed was a column called "Notes", which had notes about that team from a particular season. It served no purpose, so I removed it before reading in the .xls file into R.

Quality Assessment

After reading in the .xls file to RStudio, I ran the dim() function which gave me the total numbers of rows and columns (13486, 20). I then ran the sum() function to show each column, it's data type and a few other attributes. I can see that for all the numeric attributes, it also displays min, 1st quartile, median, mean, 3rd quartile and max values. This allowed me to ensure that I did not need to clean my data and that the column names were clear and informative.

```
      Year            Rk            School              Conf              Ovr W       
 Min.   :1869   Min.   :  1.00   Length:13486       Length:13486       Min.   : 0.000  
 1st Qu.:1933   1st Qu.: 25.00   Class :character   Class :character   1st Qu.: 3.000  
 Median :1962   Median : 53.00   Mode  :character   Mode  :character   Median : 5.000  
 Mean   :1961   Mean   : 55.43                                        Mean   : 5.393  
 3rd Qu.:1990   3rd Qu.: 83.00                                        3rd Qu.: 7.000  
 Max.   :2018   Max.   :145.00                                        Max.   :16.000  

     Ovr L            Ovr T           Ovr PCT          Conf W          Conf L      
 Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.000  
 1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:0.3640   1st Qu.:0.000   1st Qu.:0.000  
 Median : 4.000   Median :0.0000   Median :0.5450   Median :2.000   Median :2.000  
 Mean   : 4.556   Mean   :0.3329   Mean   :0.5373   Mean   :2.217   Mean   :2.217  
 3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.:0.7080   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :13.000   Max.   :5.0000   Max.   :1.0000   Max.   :9.000   Max.   :9.000  

     Conf T          Conf PCT          Off              Def              SRS         
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.00   Min.   : 0.00   Min.   :-36.080  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:13.40   1st Qu.:10.00   1st Qu.: -5.640  
 Median :0.0000   Median :0.3330   Median :19.20   Median :17.00   Median :  1.820  
 Mean   :0.1254   Mean   :0.3477   Mean   :19.81   Mean   :17.42   Mean   :  1.454  
 3rd Qu.:0.0000   3rd Qu.:0.6250   3rd Qu.:25.50   3rd Qu.:23.80   3rd Qu.:  8.710  
 Max.   :4.0000   Max.   :1.0000   Max.   :68.70   Max.   :60.50   Max.   : 34.770  
                                                                   NA's   :3        
      SOS               AP Pre          AP High          AP Rank        Differential     
 Min.   :-18.41000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :-58.500  
 1st Qu.: -4.21000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: -5.400  
 Median :  0.12000   Median : 0.000   Median : 0.000   Median : 0.000   Median :  2.500  
 Mean   : -0.06405   Mean   : 1.254   Mean   : 2.378   Mean   : 1.473   Mean   :  2.377  
 3rd Qu.:  4.17000   3rd Qu.: 0.000   3rd Qu.: 0.000   3rd Qu.: 0.000   3rd Qu.: 10.000  
 Max.   : 18.73000   Max.   :25.000   Max.   :25.000   Max.   :25.000   Max.   : 68.300  
 NA's   :3
```

Data Cleansing

I knew there was no missing data in our dataset since there was a mean for each numeric column. The only column removed was “Notes” and it was done before reading in the data.

Data Dictionary

Below are the definitions for all fields in the Combined Seasons dataset that I chose to use.

Variables	Data Type	Description
year	year	The year In which a season of football was played
rk	integer	What rank the team in question finished at the end of the season
school	string	Team name
conf	string	Team conference name
ovr_w	integer	Overall wins for a particular season
over_l	integer	Overall losses for a particular season
over_t	integer	Overall ties for a particular season
over_pct	decimal	Win percentage for a particular season
conf_w	integer	Overall conference wins for a particular season
conf_l	integer	Overall conference losses for a particular season
conf_t	integer	Overall conference ties for a particular season
conf_pct	decimal	Conference win percentage for a particular season
off	decimal	Avg points scored by a team in a particular season
def	decimal	Avg points scored against a team in a particular season
srs	decimal	Simple Rating System (a team rating that takes into account average point differential and strength of schedule)
sos	decimal	Strength of Schedule (the difficulty or ease of a team's opponent as compared to other teams)
ap_pre	string	Preseason AP Ranking (the ranking given by the Associated Press at the beginning of the season)
ap_high	integer	Highest AP Ranking achieved in a particular season
ap_rank	integer	Final AP Ranking (the ranking given by the Associated Press at the end of the season)
differential	decimal	Point Differential (avg points scored minus avg points scored against)

About the Data

The str() function gave me a brief overview and displayed the first few elements for each column in our dataset.

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       13486 obs. of  20 variables:
 $ Year         : num  1869 1869 1870 1870 1870 ...
 $ Rk           : num  1 2 1 2 3 2 1 4 3 5 ...
 $ School       : chr  "Princeton" "Rutgers" "Princeton" "Rutgers" ...
 $ Conf         : chr  "Ind" "Ind" "Ind" "Ind" ...
 $ Ovr W        : num  1 1 1 1 0 1 1 1 1 0 ...
 $ Ovr L        : num  1 1 0 1 1 0 0 2 1 1 ...
 $ Ovr T        : num  0 0 0 0 0 0 0 1 1 0 ...
 $ Ovr PCT      : num  0.5 0.5 1 0.5 0 1 1 0.375 0.5 0 ...
 $ Conf W       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Conf L       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Conf T       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Conf PCT     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Off          : num  6 3 6 4 3 3 4 2.8 2.7 0 ...
 $ Def          : num  3 6 2 4.5 6 0 1 2.5 3 6 ...
 $ SRS          : num  0 0 7 0 -7 4.9 8.4 -2.1 1.4 -9.1 ...
 $ SOS          : num  -0.5 0.5 0 0 0 -2.1 1.4 -0.35 1.4 -2.1 ...
 $ AP Pre       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AP High      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AP Rank      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Differential: num  3 -3 4 -0.5 -3 3 3 0.3 -0.3 -6 ...
```

Next, I used the head() function which allowed me to see the first 10 rows of the dataset.

```
# A tibble: 10 x 20
    Year    Rk School Conf  `Ovr W` `Ovr L` `Ovr T` `Ovr PCT` `Conf W`
   <dbl> <dbl> <chr>  <chr>   <dbl>   <dbl>   <dbl>     <dbl>    <dbl>
 1  1869     1 Princ… Ind         1       1       0     0.5          0
 2  1869     2 Rutge… Ind         1       1       0     0.5          0
 3  1870     1 Princ… Ind         1       0       0     1            0
 4  1870     2 Rutge… Ind         1       1       0     0.5          0
 5  1870     3 Colum… Ind         0       1       0     0            0
 6  1872     2 Yale   Ind         1       0       0     1            0
 7  1872     1 Princ… Ind         1       0       0     1            0
 8  1872     4 Colum… Ind         1       2       1     0.375        0
 9  1872     3 Rutge… Ind         1       1       1     0.5          0
10  1872     5 Steve… Ind         0       1       0     0            0
# … with 11 more variables: `Conf L` <dbl>, `Conf T` <dbl>, `Conf
#   PCT` <dbl>, Off <dbl>, Def <dbl>, SRS <dbl>, SOS <dbl>, `AP
#   Pre` <dbl>, `AP High` <dbl>, `AP Rank` <dbl>, Differential <dbl>
```

I used the tail() function next, which displayed the first and last five rows of data which helped verify that there was no need to remove any rows.

```
# A tibble: 6 x 20
   Year    Rk School Conf  `Ovr W` `Ovr L` `Ovr T` `Ovr PCT` `Conf W` `Conf L` `Conf T` `Conf PCT`
  <dbl> <dbl> <chr>  <chr>   <dbl>   <dbl>   <dbl>     <dbl>    <dbl>    <dbl>    <dbl>      <dbl>
1  2018   125 Georg… Sun …       2      10       0     0.167        1        7        0      0.125
2  2018   126 Arkan… Sun …       8       5       0     0.615        5        3        0      0.625
3  2018   127 Louis… Sun …       7       7       0     0.5          5        3        0      0.625
4  2018   128 Louis… Sun …       6       6       0     0.5          4        4        0      0.5
5  2018   129 South… Sun …       3       9       0     0.25         2        6        0      0.25
6  2018   130 Texas… Sun …       3       9       0     0.25         1        7        0      0.125
# … with 8 more variables: Off <dbl>, Def <dbl>, SRS <dbl>, SOS <dbl>, `AP Pre` <dbl>, `AP
#   High` <dbl>, `AP Rank` <dbl>, Differential <dbl>
```

Demographic Statistics

The dataset consists of 298 teams. 168 of those teams no longer play or are not in the Division I FBS, so I will not use them in my analysis. Out of the remaining 130 FBS teams, 7/8 Blue Bloods play in conferences that belong to the Power 5 and 1/8 are Independent, which means they are not in any conference. These 70 of teams (5 Conferences + Independent teams) are the best teams in the country. I will only be calling on data from teams inside of this 70. The dataset covers almost every season played for every team going back as far as 1869.

Data Analysis and Modeling Functions Used

sqldf function

All the rows of data have made it very difficult to read, so I had to subset the various totals and averages by school and then by conferences using the sqldf() function as well as one big data frame with all teams and conferences. Now I will be able to plot and analyze the 70 teams easier.

Rbind function

I'm ready now to do a deep dive into my college football data by creating my own Conference data frames containing the schools summarized info within them by using the rbind() function. By combining all the school vectors into their own conference data frames, I can now perform further analysis using various modeling and data analysis functions.

lm function

Wins have always been the main criteria used in determining Blue Blood status, so it will be my dependent variable. After creating my conference data frames, I used the lm() function to make linear regression models to determine if a team's Overall Wins are statistically significant in relation to Conf Wins, SOS, SRS, WinPCT, ConfPCT, APRanks and Diff independent variables.

summary function

Using the summary() function, I was able to determine that out of all the independent variables, SRS had the highest R-squared value at .7257 while Conference Win Pct had the lowest at .1782.

Using all the independent variables resulted in an adjusted R-squared value of .9724. In addition, all variables appear to be significant based on their extremely low p values.

Wins vs ConfWins

```
Call:
lm(formula = Wins ~ ConfWins, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-383.91  -71.30  -10.99   53.96  449.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 389.1194    38.0908  10.216 2.27e-15 ***
ConfWins      0.7938     0.1252   6.341 2.13e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.9 on 68 degrees of freedom
Multiple R-squared:  0.3716,    Adjusted R-squared:  0.3624
F-statistic: 40.21 on 1 and 68 DF,  p-value: 2.134e-08
```

Wins vs SOS

```
Call:
lm(formula = Wins ~ SOS, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-326.63  -95.45   -1.85   86.94  263.66

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  525.172     20.603  25.491  < 2e-16 ***
SOS           34.665      5.526   6.274 2.81e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.4 on 68 degrees of freedom
Multiple R-squared:  0.3666,    Adjusted R-squared:  0.3573
F-statistic: 39.36 on 1 and 68 DF,  p-value: 2.811e-08
```

Wins vs SRS

```
Call:
lm(formula = Wins ~ SRS, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-256.60  -48.93   12.47   59.50  171.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  482.144     13.987   34.47   <2e-16 ***
SRS           25.921      1.932   13.41   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.82 on 68 degrees of freedom
Multiple R-squared:  0.7257,    Adjusted R-squared:  0.7217
F-statistic: 179.9 on 1 and 68 DF,  p-value: < 2.2e-16
```

Wins vs WinPct

```
Call:
lm(formula = Wins ~ WinPct, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-510.71  -19.60   36.11   54.76   96.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -236.53      74.65  -3.168   0.0023 **
WinPct       1506.48     131.33  11.471   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.65 on 68 degrees of freedom
Multiple R-squared:  0.6593,    Adjusted R-squared:  0.6543
F-statistic: 131.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

Wins vs ConfPct

```
Call:
lm(formula = Wins ~ ConfPct, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-537.44  -77.06   -4.00   75.97  394.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   438.52      47.98   9.141 1.88e-13 ***
ConfPct       419.65     109.27   3.840 0.000272 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148.6 on 68 degrees of freedom
Multiple R-squared:  0.1782,    Adjusted R-squared:  0.1661
F-statistic: 14.75 on 1 and 68 DF,  p-value: 0.0002723
```

Wins vs APRanks

```
Call:
lm(formula = Wins ~ APRanks, data = FBS)

Residuals:
     Min       1Q   Median       3Q      Max
-230.538  -74.698   -8.227   55.074  256.889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2880    67.0899   0.466    0.642
APRanks       5.4854     0.6235   8.798 7.81e-13 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112.1 on 68 degrees of freedom
Multiple R-squared:  0.5324,    Adjusted R-squared:  0.5255
F-statistic: 77.41 on 1 and 68 DF,  p-value: 7.811e-13
```

Wins vs Diff

```
Call:
lm(formula = Wins ~ Diff, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-422.76  -23.27   30.66   52.23  144.35

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  500.918     14.226   35.21   <2e-16 ***
Diff          28.863      2.385   12.10   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.27 on 68 degrees of freedom
Multiple R-squared:  0.683,     Adjusted R-squared:  0.6783
F-statistic: 146.5 on 1 and 68 DF,  p-value: < 2.2e-16
```

Wins vs All Variables

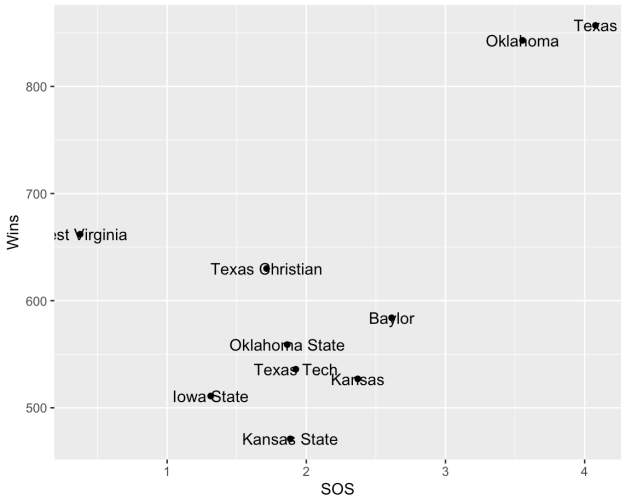
```
Call:
lm(formula = Wins ~ ConfWins + SOS + SRS + WinPct + ConfPct +
    APRanks + Diff, data = FBS)

Residuals:
    Min      1Q  Median      3Q     Max
-59.396 -11.612   5.363  18.025  53.739

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -202.1176   176.6152  -1.144   0.2569
ConfWins       0.3125     0.1334   2.342   0.0224 *
SOS          -21.6259    16.2508  -1.331   0.1881
SRS           19.3017    16.1813   1.193   0.2375
WinPct       639.3111   338.3174   1.890   0.0635 .
ConfPct     -171.6512    90.8342  -1.890   0.0635 .
APRanks        3.8250     0.2636  14.510   <2e-16 ***
Diff          -2.5740     8.0988  -0.318   0.7517
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.04 on 62 degrees of freedom
Multiple R-squared:  0.9752,    Adjusted R-squared:  0.9724
F-statistic: 347.8 on 7 and 62 DF,  p-value: < 2.2e-16
```

ggplot function

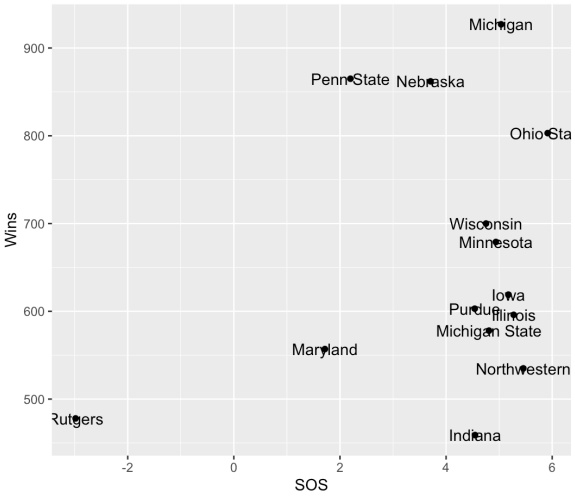
I wanted to plot some scatterplots to see if any of the 8 Blue Bloods stood out as outliers or shared similarities using ggplot().

Since all independent variables of my data set were deemed significant, I decided to start with the most important outside of Wins, which is Strength of Schedule. Using scatterplots, I plotted SOS and Wins for every conference and then the entire FBS. As you can see, all 8 Blue Bloods are in the upper right hand corners of their conferences. Even in the FBS scatter plot, all 8 Blue Bloods appear in the top right corner.

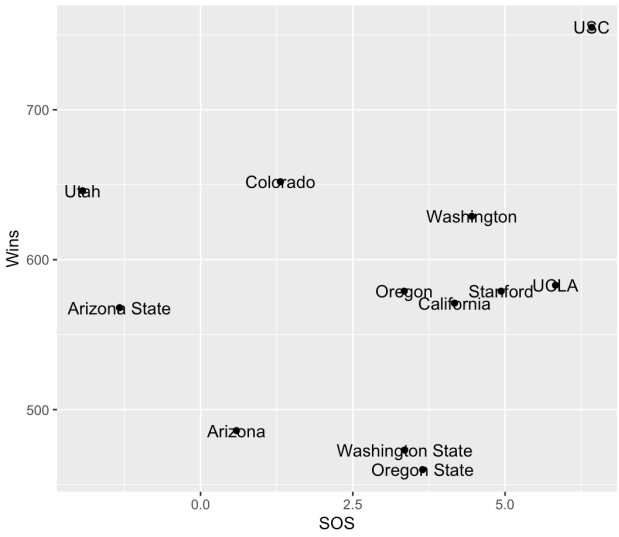
Big12



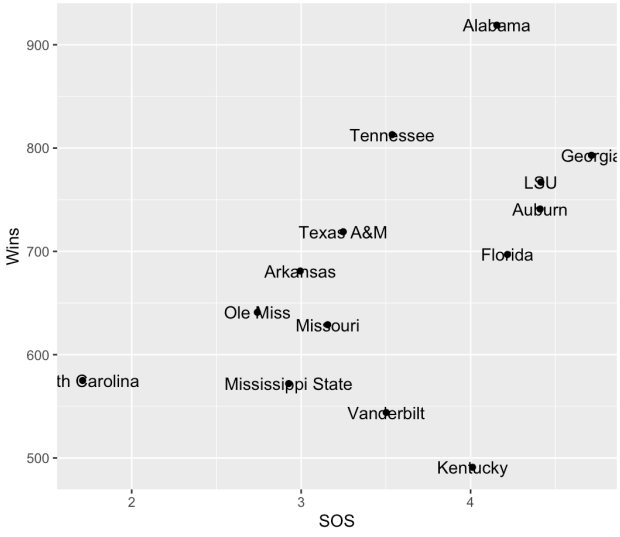
BigTen



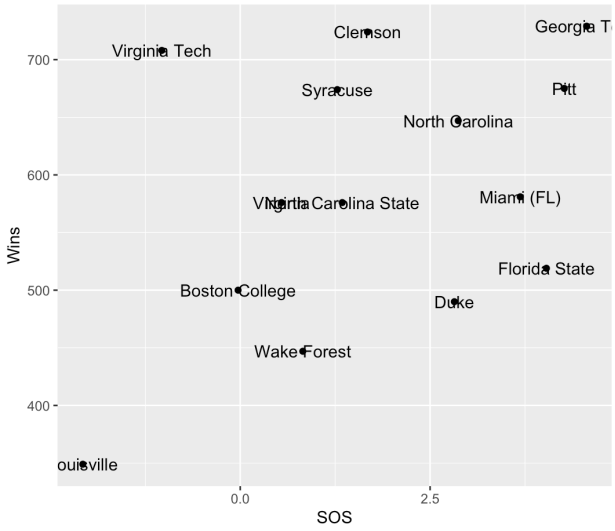
Pac12



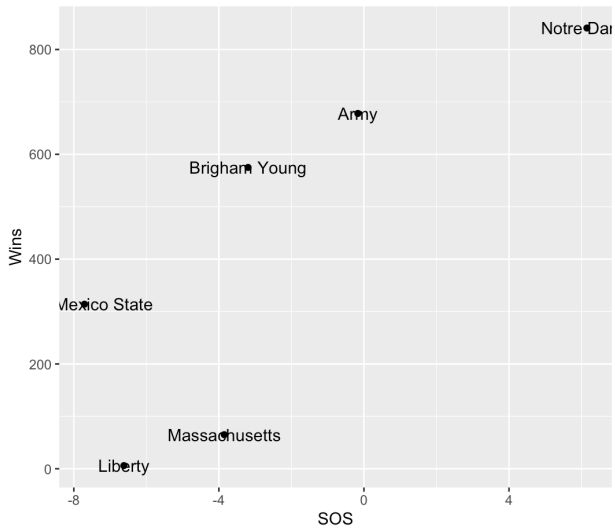
SEC



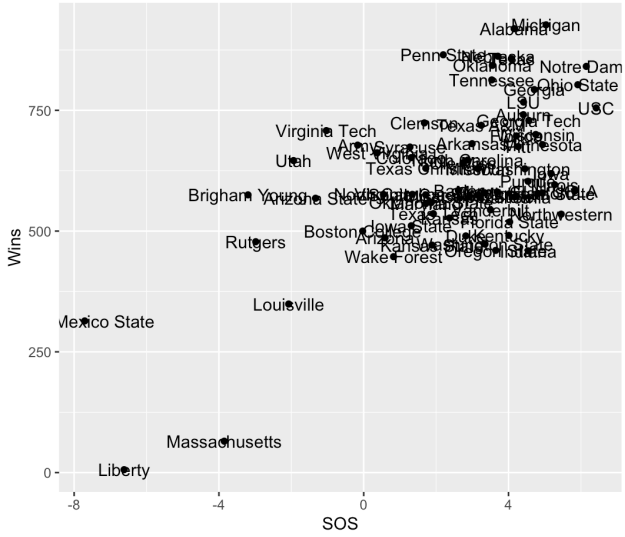
ACC



IND

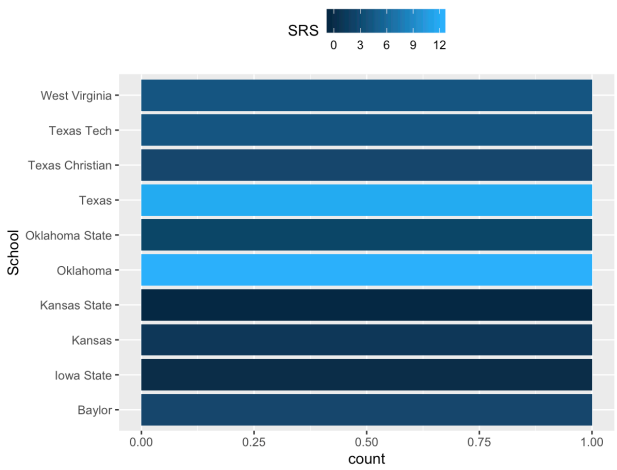


FBS

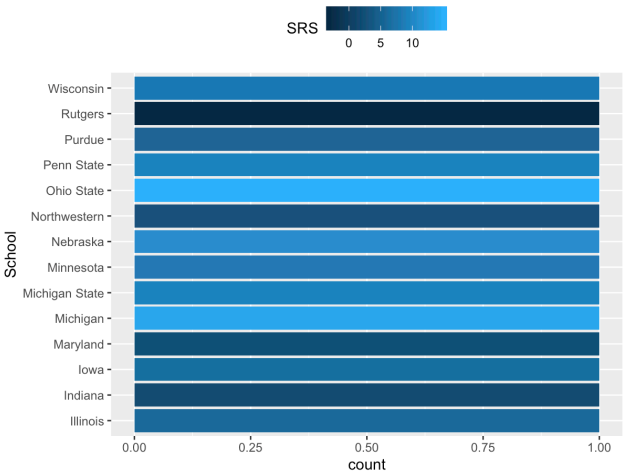


Since SRS had the highest R-Squared out of all the independent variables, I decided to see if I could tell anything about the 8 Blue Bloods regarding where they rank in SRS in their respected divisions next.

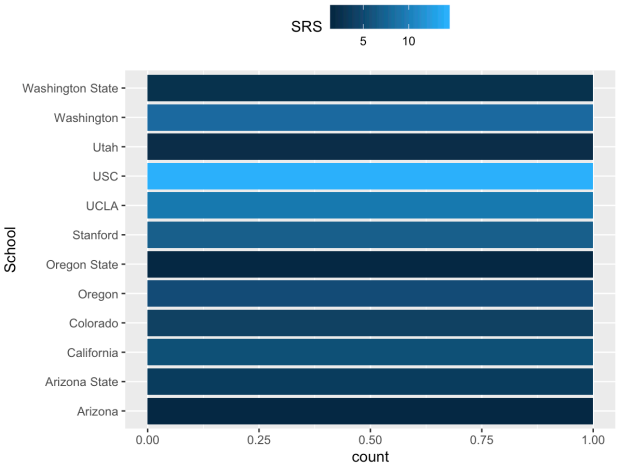
Big12



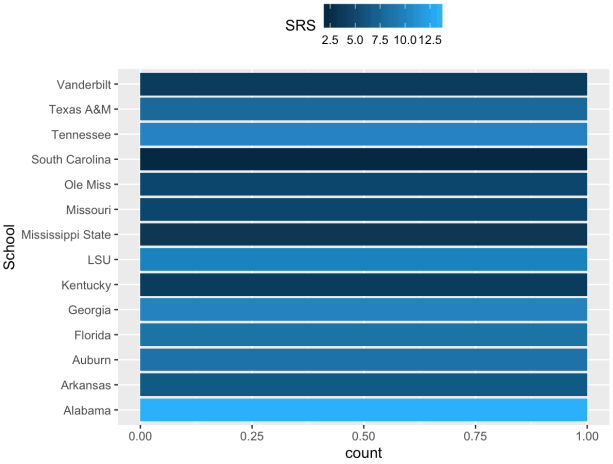
BigTen



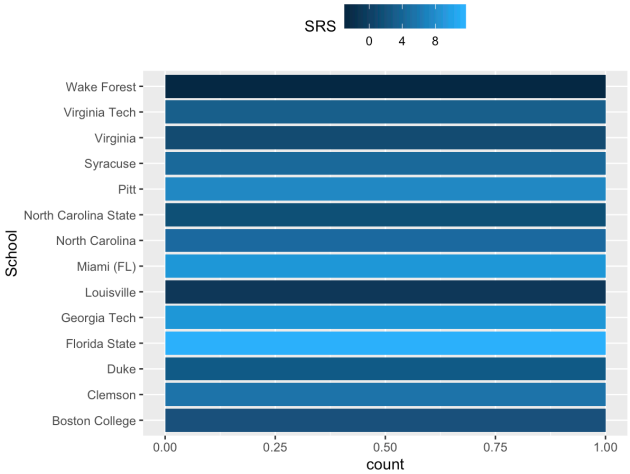
Pac12



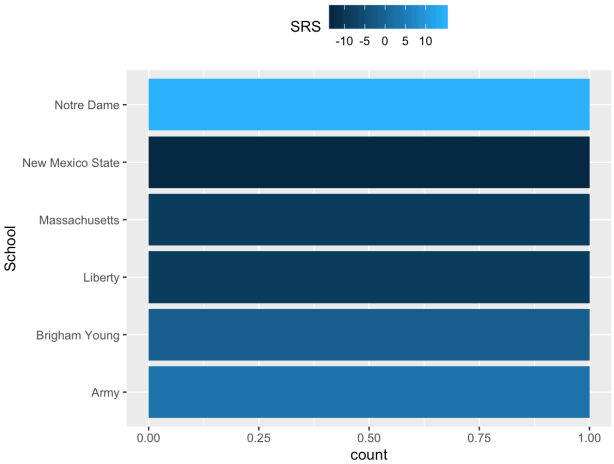
SEC



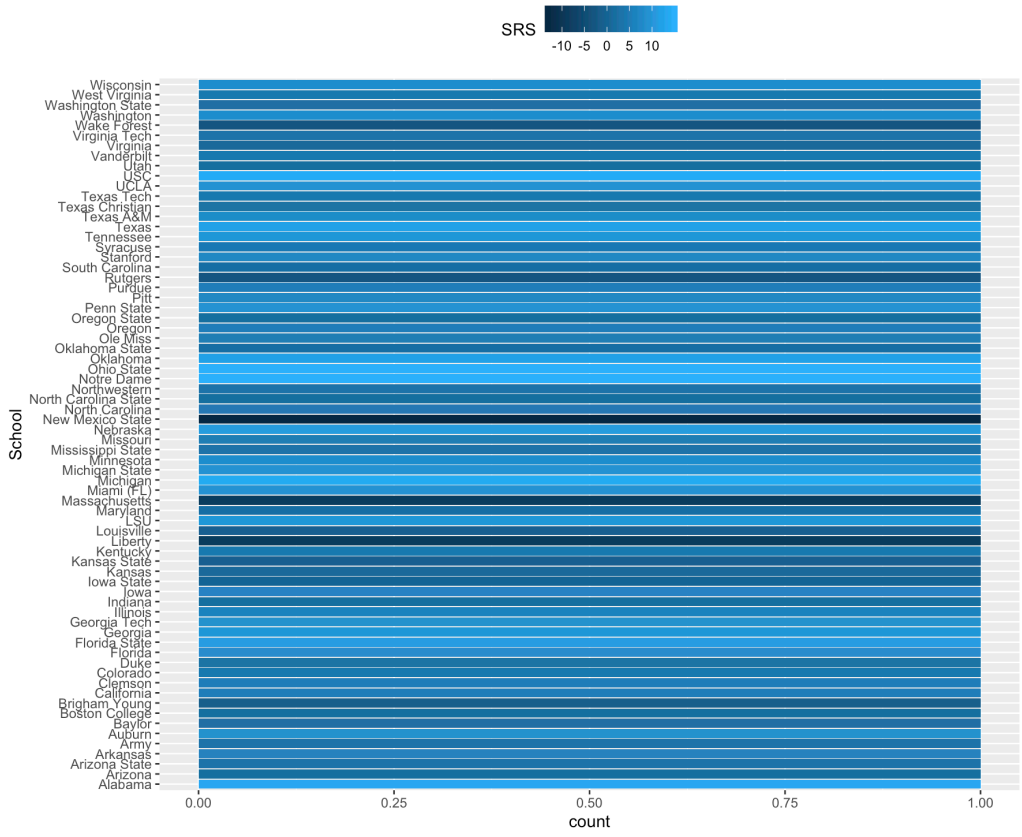
ACC



IND

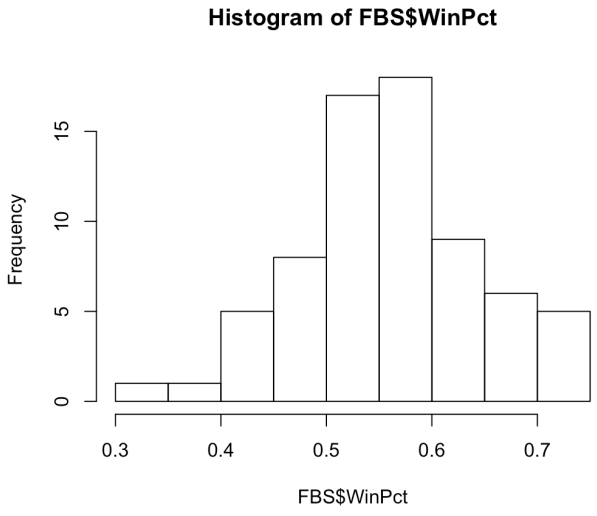


All 8 Blue Bloods dominate and stand out in when comparing SRS among conferences and when compared to the entire FBS as a whole.



hist and order function

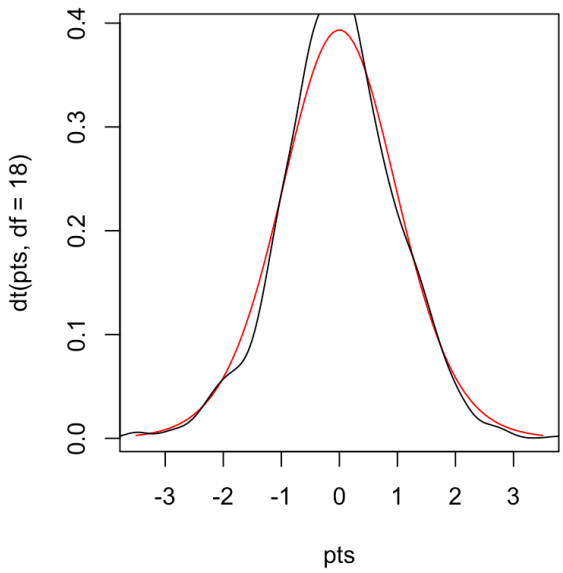
I next used the hist() function to show how win % is distributed among the 70 FBS teams. Using the order() function, I sorted the teams in the FBS by WinPct to see where the 8 Blue Bloods fell. All 8 Blue Bloods fill up the Top 8 spots. I'm beginning to realize why everyone ranks them the way they do. Since ND technically isn't in a conference (IND schools aren't a part of conferences), I decided not to take conf wins or conference win % rankings into consideration.



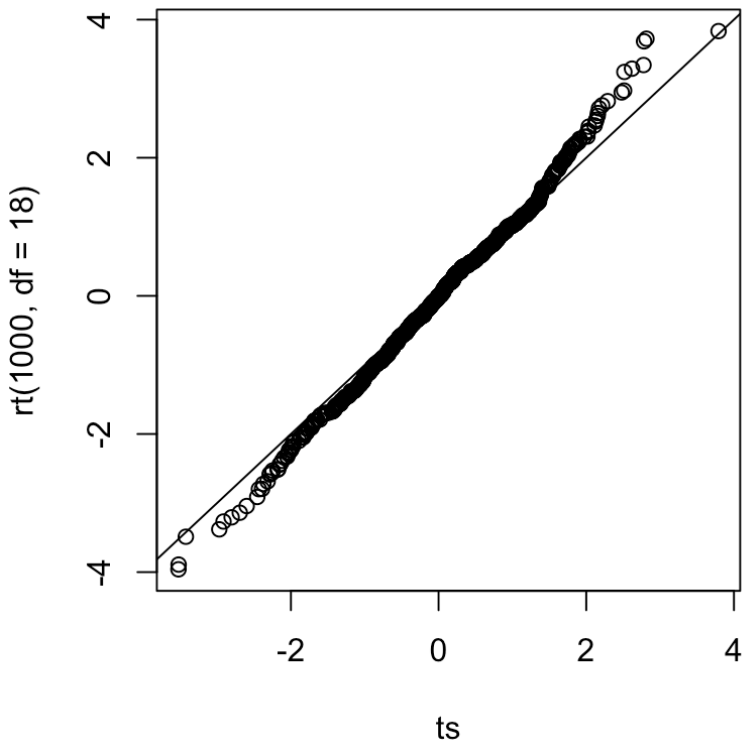
	School	Wins	ConfWins	SOS	SRS	WinPct	ConfPct	APRanks	Diff
25	Ohio State	803	520	5.91392523	14.9603738	0.7378224	0.71578505	107	12.5564486
39	Alabama	919	489	4.15478261	13.4418261	0.7366870	0.61078261	115	13.1179130
66	Notre Dame	841	3	6.14767857	14.7358929	0.7283304	0.01785714	112	11.7824107
27	Michigan	927	512	5.03697674	13.7378295	0.7216279	0.61574419	129	12.7984496
21	Oklahoma	843	502	3.55354545	12.3219091	0.7151273	0.71767273	110	13.7583636
61	USC	755	460	6.42175258	14.1636082	0.6999691	0.71556701	97	10.5157732
15	Texas	857	480	4.07811966	11.8292308	0.6968803	0.61616239	117	10.5431624
32	Nebraska	862	471	3.70831933	10.8342017	0.6885462	0.66474790	119	10.8803361
26	Penn State	865	133	2.19712000	9.1200800	0.6769760	0.13067200	125	9.6900800
47	Tennessee	813	388	3.53701754	10.1019298	0.6745439	0.53785088	114	9.3819298
5	Florida State	519	168	4.03200000	11.3209231	0.6732154	0.32307692	65	10.4386154
49	Georgia	793	382	4.71356522	10.0774783	0.6441391	0.51748696	115	7.3726957
43	LSU	767	345	4.41482456	9.8192105	0.6401754	0.48122807	114	8.0932456
14	Miami (FL)	581	139	3.68719512	8.9825610	0.6315488	0.24150000	82	7.0518293
41	Auburn	741	338	4.41043103	8.7835345	0.6202328	0.43816379	116	6.3022414
58	Arizona State	568	319	-1.33129412	3.5464706	0.6127765	0.59617647	85	7.0282353
48	Florida	697	354	4.21738318	8.5590654	0.6078411	0.50928972	107	6.3543925
28	Michigan State	578	282	4.80882979	9.2619149	0.6042447	0.39991489	94	6.5364894
40	Texas A&M	719	353	3.24844828	7.8480172	0.6026724	0.48215517	116	7.0893103
10	Virginia Tech	708	251	-1.03347826	3.2339130	0.6016696	0.36188696	115	5.8806957
13	Georgia Tech	729	309	4.56170940	8.7217094	0.5991966	0.40382051	117	6.7068376
1	Clemson	724	368	1.67581197	5.3715385	0.5969145	0.51476923	117	5.2030769
62	UCLA	583	358	5.82747253	9.3017582	0.5965055	0.57436264	91	4.8657143
59	Utah	646	410	-1.93848214	2.0786607	0.5953036	0.59592857	112	5.8941964
24	West Virginia	662	183	0.37485714	4.0310476	0.5941792	0.29829245	106	5.4533019
53	Washington	629	396	4.45274510	8.1449020	0.5895098	0.56208824	102	5.6070588
29	Wisconsin	700	377	4.75259843	8.1737008	0.5841417	0.48734646	127	5.0809449

t.test, replicate and range function

I used a t-test to determine that the means of Wins and APRanks are equal to each other. After extracting the t-statistic from the output of the t.test() function, I replicated the t-statistic 1000 times in order to plot the theoretical density of it. To get an idea of what range of x values I needed to use for the theoretical density, I viewed the range of my simulated data with the range() function and plotted equally spaced x-values ranging -3.5 to 3.5 along with a line showing the density for my simulated data.

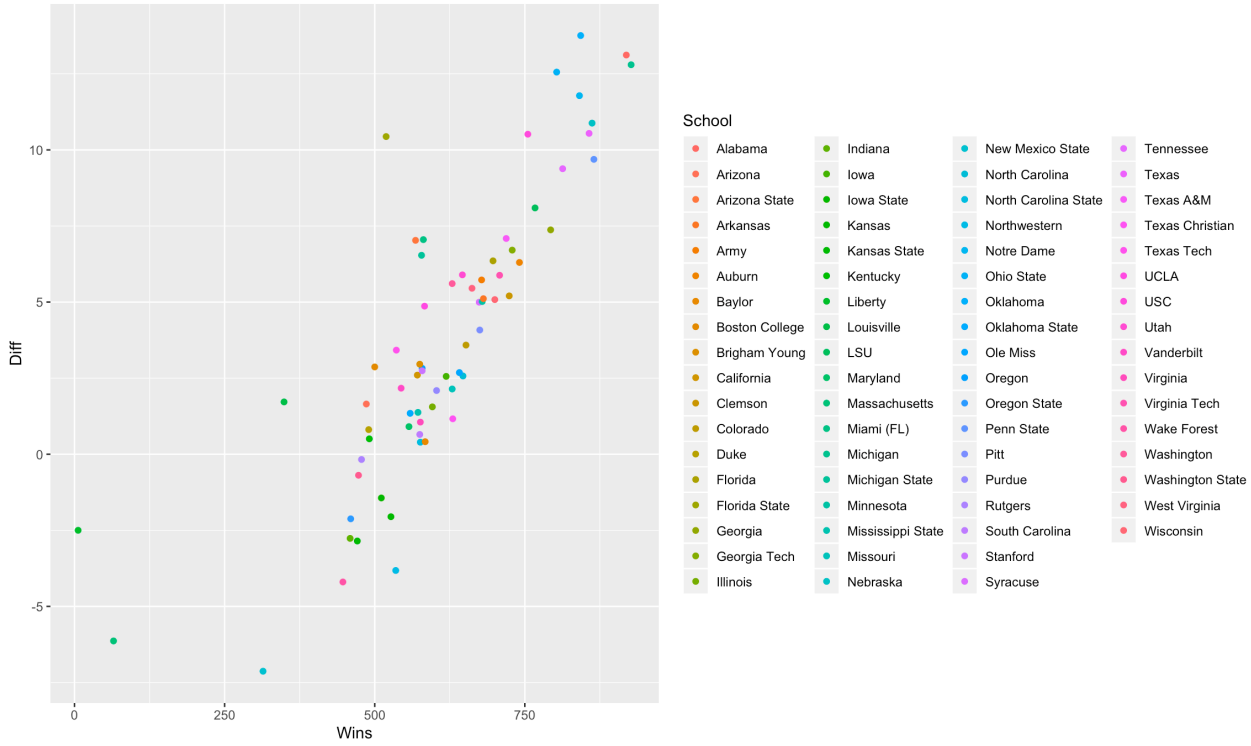


Another way I compared two densities is with a quantile-quantile plot. In this type of plot, the quantiles of two samples are calculated at a variety of points in the range of 0 to 1, and then are plotted against each other. If the two samples came from the same distribution with the same parameters, the result would be a straight line through the origin with a slope of 1 (the various quantiles of the data are identical in the two samples). If the two samples came from similar distributions, but their parameters were different, there would be a straight line, but not through the origin. I produced a quantile-quantile plot using the qqplot() function.



The central points of the graph seem to agree fairly well, but there are some discrepancies near the tails (the extreme values on either end of the distribution). Because the tails of a distribution are so important, another way to test to see if a distribution of a sample follows some hypothesized distribution is to calculate the quantiles of some tail probabilities (using the quantile function) and compare them to the theoretical probabilities from the distribution. The quantiles agree fairly well, especially at the .95 and .99 quantiles. Performing more simulations or using a large sample size for the two groups would probably result in values even closer.

The last variable to dig deeper on is the teams' point differentials. Using ggplot again, I showed the correlation between overall wins and point differential. All the Blue Bloods are grouped together in the top right section again.



Overall Interpretations and Conclusions

Summary of Findings

As a result of the analysis, the following results were discovered.

- The 8 Blue Blood schools are consistently in the top of every variable
- All independent variables have a significant correlation with overall wins based on their extremely low p values.
- SRS is the best individual driver of Wins
- Based on this correlation of Wins and SOS, Penn State and Georgia Tech are making a case for being included according to the scatter plots.
- Penn State, Tennessee and Florida State are making a case for being a Blue Blood based on overall win percentage and point differential
- Tennessee and Florida State are making a case for being a Blue Blood based on SRS
- I did not go into conference wins and win percentage due to Notre Dame not being in a conference, but the other 7 were at the top of the list in those categories as well

Conclusions

The main question we sought to answer for the College Football community was:

1. Do the 8 “Blue Bloods” have an actual claim to their title?

The answer is undoubtedly “Yes”, but I also learned the following from my analysis:

1. Alabama, USC, Texas, Oklahoma, Michigan, Ohio State and Notre Dame rank at the top of every statistical category in the data set
2. Outsiders looking in the “Blue Blood” category are Penn State, Florida State and Tennessee and they deserve their own title, like “Light Blue Bloods”. They are right behind the Top 8 in most every category in the data set

References

Link to the dataset:
Sports Viz Sunday
https://query.data.world/s/zohcgl3nmrgzuronv6qwf7cgyowzyl

ESPN (2016, August 25) Which schools should be considered college football royalty? Retrieved from https://www.espn.com/college-football/story/_/id/17336754/alabama-crimson-tide-notre-dame-fighting-irish-ohio-state-buckeyes-oklahoma-sooners-usc-trojans-lead-list-college-football-blue-bloods

Code Appendix

Importing Data

```
> CFB <- read_excel("CombinedSeasons.xls")
```

Quality Assessment

```
> dim(CFB)
```

```
> summary(CFB)
```

About the Data

```
> str(CFB)
```

```
> head(CFB, 10)
```

```
> tail(CFB)
```

Data Analysis and Modeling Functions Used

rbind function

Big12

```
> Texas<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Texas'")
> Oklahoma<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Oklahoma'")
> Baylor<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Baylor'")
> IowaState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Iowa State'")
> Kansas<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Kansas'")
> KansasState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Kansas State'")
> OklahomaState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS),
avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB
WHERE School = 'Oklahoma State'")
> TCU<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Texas Christian'")
> TexasTech<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Texas Tech'")
> WestVirginia<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'West Virginia'")

> Big12<-rbind(Texas, TexasTech, TCU, IowaState, Kansas, KansasState, Oklahoma,
OklahomaState, Baylor, WestVirginia)

> names(Big12) <- c("School", "Wins", "ConfWins", "SOS", "SRS", "WinPct","ConfPct",
"APRanks", "Diff")
```

BigTen

```
> OhioState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Ohio State'")
```

```
> Michigan<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Michigan'")
> PennState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Penn State'")
> MichiganState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS),
avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB
WHERE School = 'Michigan State'")
> Maryland<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Maryland'")
> Indiana<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Indiana'")
> Rutgers<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Rutgers'")
> Northwestern<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Northwestern'")
> Iowa<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Iowa'")
> Wisconsin<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Wisconsin'")
> Purdue<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Purdue'")
> Minnesota<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Minnesota'")
> Nebraska<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Nebraska'")
> Illinois<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Illinois'")

> BigTen<-rbind(OhioState, PennState, Michigan, MichiganState, Wisconsin, Indiana, Illinois,
Nebraska, Iowa, Purdue , Northwestern, Maryland, Rutgers, Minnesota)

> names(BigTen) <- c("School", "Wins", "ConfWins", "SOS", "SRS", "WinPct","ConfPct",
"APRanks", "Diff")
```

Pac12

```
> WashingtonState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Washington State'")
> Washington<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Washington'")
> Stanford<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Stanford'")
> OregonState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Oregon State'")
> Oregon<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Oregon'")
> California<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'California'")
> Utah<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Utah'")
> ArizonaState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Arizona State'")
> Arizona<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School =  'Arizona'")
> USC<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School =  'USC'")
> UCLA<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'UCLA'")
> Colorado<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Colorado'")

> Pac12<-rbind(Washington, WashingtonState, Oregon, OregonState, Arizona, ArizonaState, Utah, Colorado, USC, UCLA,California,Stanford)

> names(Pac12) <- c("School", "Wins", "ConfWins", "SOS", "SRS", "WinPct","ConfPct", "APRanks", "Diff")
```

SEC

```
> Georgia<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Georgia'")
> Florida<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Florida'")
> Kentucky<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Kentucky'")
> Missouri<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Missouri'")
> SouthCarolina<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'South Carolina'")
> Vanderbilt<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Vanderbilt'")
> Tennessee<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Tennessee'")
> Alabama<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Alabama'")
> LSU<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr
PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School =
'LSU'")
> TAMU<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Texas A&M'")
> MississippiState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS),
avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB
WHERE School = 'Mississippi State'")
> Auburn<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Auburn'")
> OleMiss<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Ole Miss'")
> Arkansas<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Arkansas'")

> SEC<-rbind(Alabama, TAMU, Auburn, Arkansas, LSU, OleMiss, MississippiState,
Vanderbilt, Tennessee, Florida, Georgia, SouthCarolina, Missouri, Kentucky)
```

```
> names(SEC) <- c("School", "Wins", "ConfWins", "SOS", "SRS", "WinPct","ConfPct",
"APRanks", "Diff")
```

ACC

```
> Clemson<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Clemson'")
> FloridaState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Florida State'")
> Syracuse<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Syracuse'")
> NorthCarolinaState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS),
avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB
WHERE School = 'North Carolina State'")
> BostonCollege<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS),
avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB
WHERE School = 'Boston College'")
> WakeForest<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Wake Forest'")
> Louisville<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Louisville'")
> Pittsburgh<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Pitt'")
> GeorgiaTech<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Georgia Tech'")
> Virginia<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School =  'Virginia'")
> Miami<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Miami (FL)'")
> VirginiaTech<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Virginia Tech'")
> Duke<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS),
avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE
School = 'Duke'")
```

```
> NorthCarolina<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'North Carolina'")

> ACC<-rbind(Clemson, NorthCarolina, NorthCarolinaState, WakeForest, FloridaState, Syracuse, Louisville, Duke, Virginia, VirginiaTech, BostonCollege, Pittsburgh, GeorgiaTech, Miami)

> names(ACC) <- c("School", "Wins", "ConfWins", "SOS", "SRS", "WinPct","ConfPct", "APRanks", "Diff")
```

IND

```
> Massachusetts<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School =  'Massachusetts'")
> NewMexicoState<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School =  'New Mexico State'")
> Liberty<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Liberty'")
> Army<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Army'")
> BYU<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Brigham Young'")
> NotreDame<-sqldf("SELECT School, sum(`Ovr W`), sum(`Conf W`), avg(SOS), avg(SRS), avg(`Ovr PCT`),avg(`Conf PCT`), count(`AP Rank`), avg(Differential) FROM CFB WHERE School = 'Notre Dame'")

> IND<-rbind(BYU, NotreDame, Massachusetts, Army, NewMexicoState, Liberty)
> names(IND) <- c("School", "Wins", "ConfWins", "SOS", "SRS", "WinPct","ConfPct", "APRanks", "Diff")
```

FBS

```
> FBS<-rbind(ACC, Big12, BigTen, SEC, Pac12,IND)
> ggplot(FBS, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))
```

lm function

```
> model1<-lm(formula=Wins ~ ConfWins, data=FBS)
> summary(model1)
```

```
> model2<-lm(formula=Wins ~ SOS, data=FBS)
> summary(model2)

> model3<-lm(formula=Wins ~ SRS, data=FBS)
> summary(model3)

> model4<-lm(formula=Wins ~ WinPct, data=FBS)
> summary(model4)

> model5<-lm(formula=Wins ~ ConfPct, data=FBS)
> summary(model5)

> model6<-lm(formula=Wins ~ APRanks, data=FBS)
> summary(model6)

> model7<-lm(formula=Wins ~ Diff, data=FBS)
> summary(model7)

> model8<-lm(formula=Wins ~ ConfWins+SOS+SRS+WinPct+ConfPct+APRanks+Diff,
data=FBS)
> summary(model8)
```

Scatter plot ggplot

```
> ggplot(Big12, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))
> ggplot(BigTen, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))

> ggplot(Pac12, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))

> ggplot(SEC, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))

> ggplot(ACC, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))

> ggplot(IND, aes(x = SOS, y = Wins)) + geom_point() + geom_text(aes(label=School))
```

Bar chart ggplot

```
> g1 <- ggplot(ACC, aes(School))
> g1 + geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")

> g2 <- ggplot(Big12, aes(School))
> g2 + geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")

> g3<- ggplot(BigTen, aes(School))
```

```
> g3+ geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")

> g4<- ggplot(Pac12, aes(School))
> g4+geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")

> g5<- ggplot(SEC, aes(School))
> g5+geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")

> g6<- ggplot(IND, aes(School))
> g6+geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")

> g7<- ggplot(FBS, aes(School))
> g7+geom_bar(aes(fill = SRS), position = position_stack(reverse = TRUE)) + coord_flip() +
theme(legend.position = "top")
```

Hist function

```
> hist(FBS$WinPct)
> TopWinPct<- FBS[order(-FBS$WinPct),]
```

T test

```
> ttest<-t.test(FBS$APRanks, FBS$Wins)
> ts = replicate(1000,t.test(rnorm(10),rnorm(10))$statistic)
> range(ts)
> pts = seq(-3.5,3.5,length=100)
> plot(pts,dt(pts,df=18),col='red',type='l')
> lines(density(ts))

> qqplot(ts,rt(1000,df=18))
> abline(0,1)

> quantile(ts,probs)
    90%      95%      99%
1.273961 1.602993 2.173103

> qt(probs,df=18)
[1] 1.330391 1.734064 2.552380
```

gpglot

```
> ggplot(FBS, aes(Wins, Diff, colour = School)) + geom_point()
```