Syracuse University-iSchool

Coaches' Salaries

Predicting Coaches' Salaries of NCAA Football Programs

Alejandro Mora
1-27-2020

Problem

Coaches Salaries can be debated amongst many people within a school's athletic program and even within fans across the world. Some people say coaches are paid too much, too little, or think they themselves can coach better when being paid as much as they are given the opportunity. To many circumstances, you can argue the coach should get paid by how good they perform in a span of seasons or the past season. This is usually how professional athletes are regarded when they bring in better records for teams and want the athletes to stay and play for them. The better the team performs, the more audience that is attracted to them and leads to more merchandise sales, ticket sales, popularity, and more. Based on personal experience going to a division 1 NCAA program university, many people would not want to go to the school's football games as the team had a struggling record for many years. Every year that I attended the university the stadium never met capacity, even at rivalry games; given that the stadium held over 30,000, it was a struggle to sell tickets. This ties back into the head coaches' salary as the question is; How much should they get paid? Many factors can go into the decision and some can be reasons not readily available to the public, things like program budget, revenue, cutbacks, biases, and so forth. But with things available to view, there can be an estimate on a few variables to help in the decision process. Many programs can use methods like linear regression models for predictions.

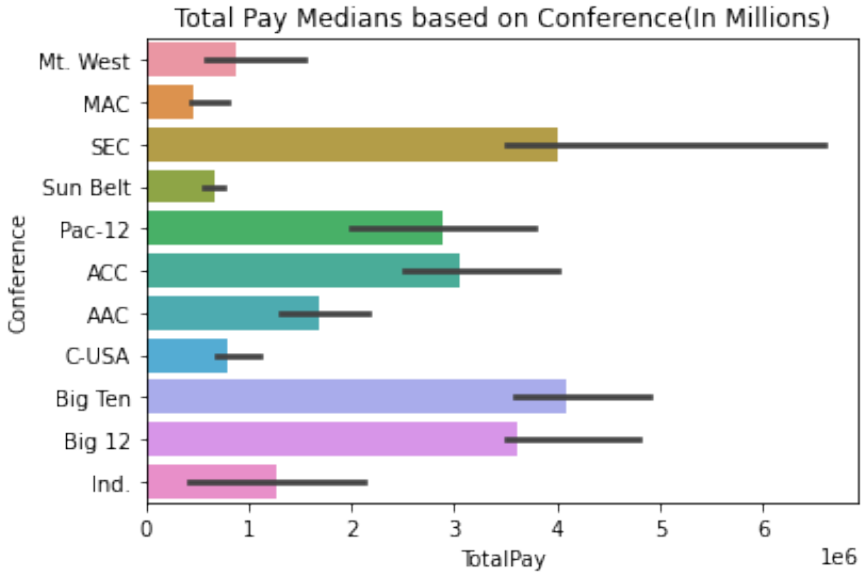
Data

For the big picture, a variety of variables are recommended to get a better estimate of coaches' pay. Multiple sources must be acquired as not one data set contains a different amount of non-similar data. One data set will contain school information and coach salary, another will contain stadium sizes of school's football programs, another will take NCAA Graduation Success Rate data from schools, and the last one will contain rankings and records of the programs themselves(Preferably from the year before since a salary can be pre-determined on what has happened and not what can happen). Some schools are omitted as there is an insignificant amount of data for their contribution to the analysis itself (e.g. not having a TotalPay value available or other variables not being readily available).

Full Variable List:

School	Conference	Coach	TotalPay	Bonus
BonusPaid	Buyout	GSR(Graduation Success Rate)	FGR(Federal Graduation Rate)	Stadium_Capacity
Rank	APRank	Wins	Loses	OSRS(Offensive SRS)
DSRS(Defensive SRS)	SRS(Average Point Differential and Strength of Schedule)			

Some things to consider when looking at a full data set with this many variables is what is important and what cannot be included in the linear regression equation. Conference is a variable worth looking at for analysis when looking at TotalPay:



The number of differences between conferences is quite large while also showing large amounts of standard error within the larger conferences like SEC, Big Ten, Big 12, and ACC. This is a variable showing a potential good variable for analysis later.

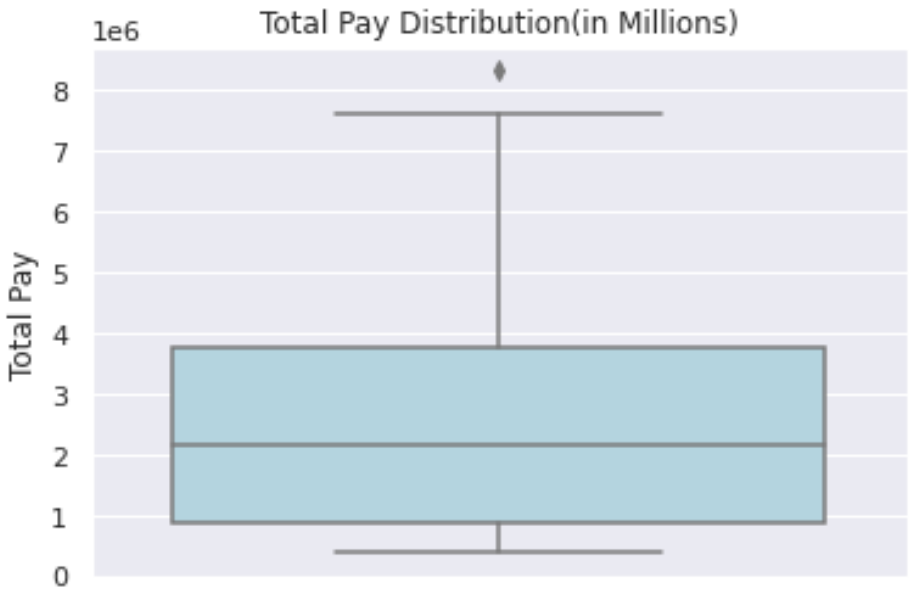
When looking at descriptive statistics within the TotalPay itself amongst all coaches, there are huge gaps between the salaries however the central tendency measures seem to rightly skew the distribution:

Mean Total Pay of Coaches:
2,570,260.33

Median Total Pay of Coaches:
2,129,638.00

Minimum Total Pay of Coaches:
390,000.00

Maximum Total Pay of Coaches:
8,307,000.00

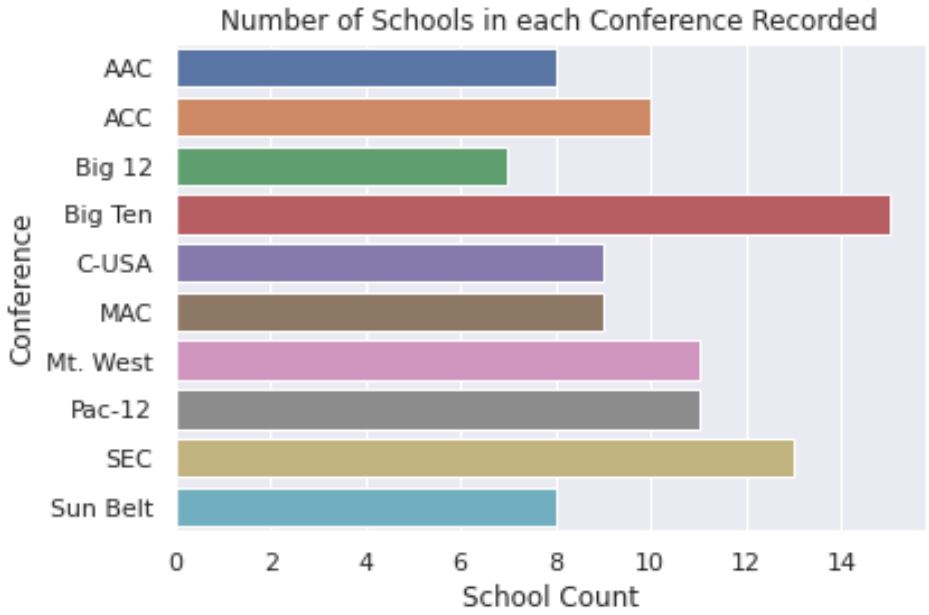


Displayed Vertically (Bottom = Left Hand Side, Top = Right Hand Side

Since there is some evidence of conferences and their variabilities of pay, there must be a check on whether the variable is valid for analysis. There needs to be a count of schools for each conference to make sure the data is about even.

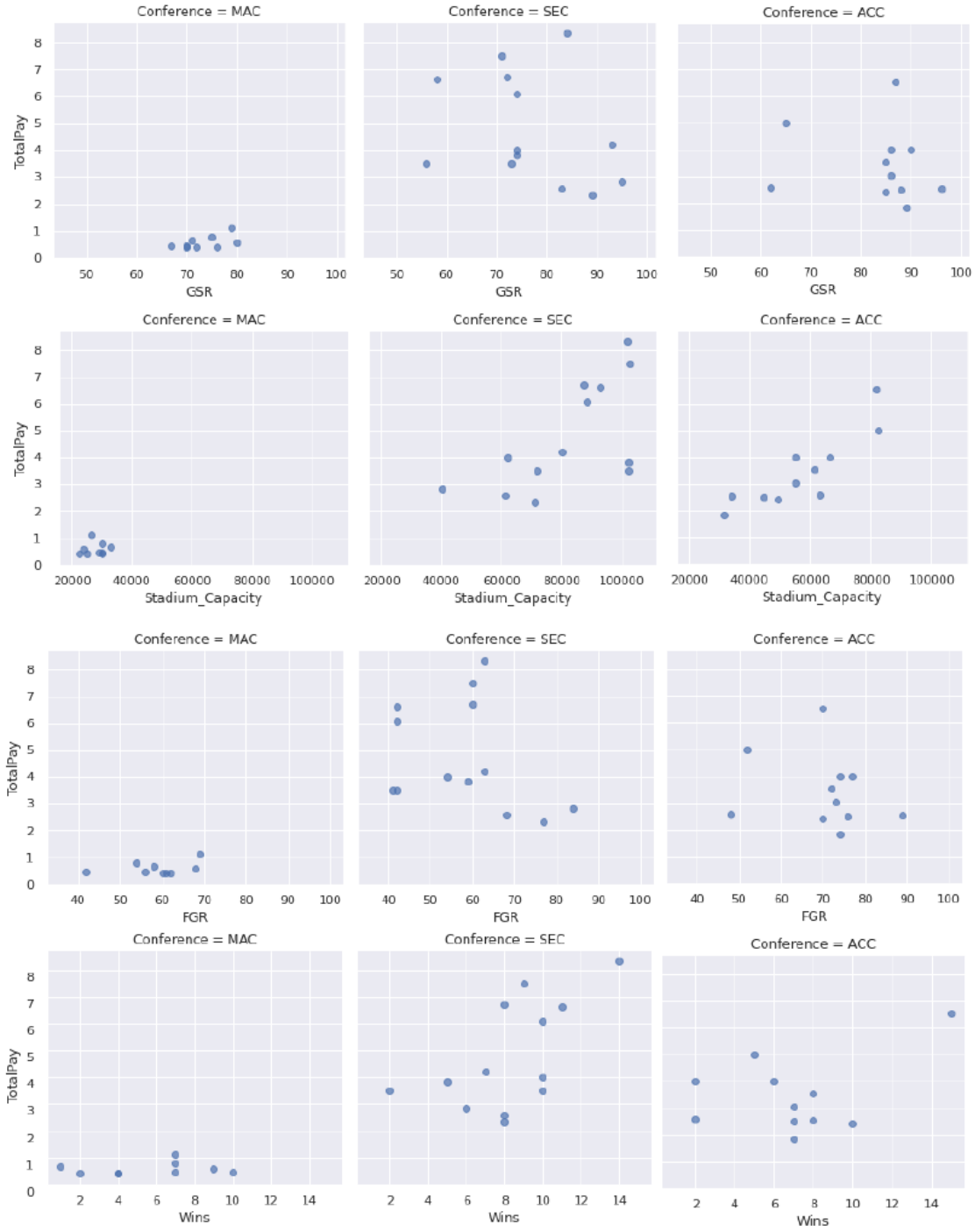
	Conference	School
0	AAC	8
1	ACC	11
2	Big 12	7
3	Big Ten	14
4	C-USA	9
5	Ind.	2
6	MAC	9
7	Mt. West	11
8	Pac-12	11
9	SEC	13
10	Sun Belt	8

The data seems about even but there is the conference “Ind.” That is showing a low number of schools. Many Independent schools were taken out of the data sets as they could not provide a TotalPay value and there are only 2 available to view. To avoid the data being too skewed, they can be removed as there are still 10 conferences that can help in analysis.

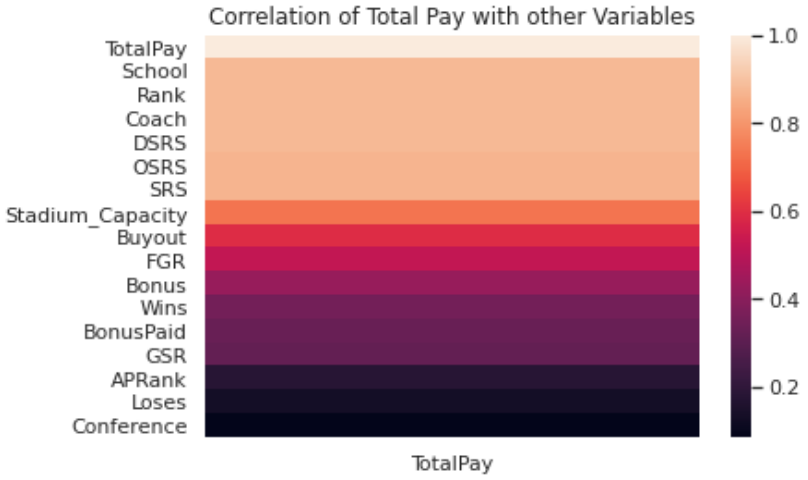


Scatterplots can help show if there are any correlations between the TotalPay and other variables. Separating them by conference can also see trends in the data. The MAC(Mid-American Conference) showed a lower coach salary value than others so we can the see the correlation between the other variables on TotalPay. The MAC schools showed consistency when looking at stadium capacity and GSR while FGR and Wins seemed to have little impact on

Pay. Showing other conferences based on their TotalPay medians, The larger conference (SEC) shows higher pay when wins goes up, stadium capacity, and GSR. The FGR statistic shows a little more randomness than the others as even a low FGR Rate had high paying salaries. The ACC shows higher pay when looking at stadium capacity and GSR but seems more random on wins and FGR.



To see a full correlation comparison between the TotalPay and other variables, a matrix can be made to see what has higher impact on the Pay.



Some variables are to not be considered while having high correlation with the TotalPay itself in a linear equation model. Variables like; School, Coach, Bonus, BonusPaid are all correlated on the fact that they are solely based on the TotalPay but do not affect it. The prediction of coach salary is based on other variables regarding the school paying a coach a certain amount regardless of who it is.

Modeling

If applicable, an ideal model would include every and any variable to be included in the equation to "get the best results". In this case including all variables would gather the best fit to the predicted salaries, however there is a large chance for your hypothesis being wrong as all variable P-values would be extremely high.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.894
Model:                           OLS   Adj. R-squared:                  0.437
Method:                Least Squares   F-statistic:                     1.955
Date:               Wed, 27 Jan 2021   Prob (F-statistic):              0.319
Time:                       22:58:36   Log-Likelihood:                -252.90
No. Observations:                  17   AIC:                             533.8
Df Residuals:                       3   BIC:                             545.5
Df Model:                          13
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------------------------
Intercept                -3.961e+05   1.53e+07     -0.026      0.981    -4.9e+07
4.82e+07
Conference[T.ACC]         -6.04e+05   2.71e+06     -0.223      0.838   -9.24e+06
8.03e+06
```

Conference[T.Big 12]	9.682e+04	1.98e+06	0.049	0.964	-6.19e+06 6.39e+06
Conference[T.Big Ten]	8.224e+04	3.71e+06	0.022	0.984	-1.17e+07 1.19e+07
Conference[T.C-USA]	-2.686e-10	2.81e-08	-0.010	0.993	-8.96e-08 8.91e-08
Conference[T.MAC]	2.923e-09	6.5e-08	0.045	0.967	-2.04e-07 2.1e-07
Conference[T.Mt. West]	-3.356e+06	4.87e+06	-0.690	0.540	-1.88e+07 1.21e+07
Conference[T.Pac-12]	1.683e+06	3.57e+06	0.471	0.670	-9.68e+06 1.3e+07
Conference[T.SEC]	1.702e+06	3.05e+06	0.558	0.616	-8e+06 1.14e+07
Conference[T.Sun Belt]	4.368e-10	7.21e-09	0.061	0.955	-2.25e-08 2.34e-08
Stadium_Capacity	60.3910	69.632	0.867	0.450	-161.210 281.992
Wins	2.837e+05	1.52e+06	0.187	0.864	-4.55e+06 5.12e+06
GSR	-4.375e+04	8.85e+04	-0.494	0.655	-3.25e+05 2.38e+05
FGR	4.972e+04	1.29e+05	0.385	0.726	-3.61e+05 4.6e+05
Loses	-1.252e+06	1.78e+06	-0.703	0.533	-6.92e+06 4.42e+06
APRank	2.201e+05	3.67e+05	0.599	0.591	-9.49e+05 1.39e+06
Rank	-5.723e+04	2.4e+05	-0.238	0.827	-8.22e+05 7.08e+05
SRS	-5.437e+04	4.16e+05	-0.131	0.904	-1.38e+06 1.27e+06

Some variables also can skew the equation and have no real significance. Some variables include APRank, Losses OSRS, DSRS and FGR. The APRank only has observations for 25 teams which means that ¾ of the data set has no value and cannot be estimated. OSRS and DSRS values are split up from the actual SRS value so we can try to leave them out to eliminate repetition and FGR, as seen from the scatterplots, can be random for some conferences so it can be a potential variable to disregard.

OLS Regression Results

Dep. Variable:	TotalPay	R-squared:	0.866
Model:	OLS	Adj. R-squared:	0.830
Method:	Least Squares	F-statistic:	23.62
Date:	Wed, 27 Jan 2021	Prob (F-statistic):	1.84e-17
Time:	22:58:44	Log-Likelihood:	-988.77
No. Observations:	66	AIC:	2008.
Df Residuals:	51	BIC:	2040.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025 0.975]
Intercept	8.978e+05	2.1e+06	0.427	0.671	-3.32e+06 5.12e+06

Conference[T.ACC]	4.952e+05	6.52e+05	0.760	0.451	-8.14e+05 1.8e+06
Conference[T.Big 12]	3.012e+05	8.01e+05	0.376	0.708	-1.31e+06 1.91e+06
Conference[T.Big Ten]	1.1e+06	9.08e+05	1.212	0.231	-7.22e+05 2.92e+06
Conference[T.C-USA]	-7.197e+05	6.06e+05	-1.188	0.240	-1.94e+06 4.97e+05
Conference[T.MAC]	-8.071e+05	5.85e+05	-1.379	0.174	-1.98e+06 3.68e+05
Conference[T.Mt. West]	-1.04e+06	5.68e+05	-1.830	0.073	-2.18e+06 1.01e+05
Conference[T.Pac-12]	2.768e+05	6.3e+05	0.439	0.662	-9.89e+05 1.54e+06
Conference[T.SEC]	5.615e+05	8.66e+05	0.648	0.520	-1.18e+06 2.3e+06
Conference[T.Sun Belt]	-1.038e+06	6.15e+05	-1.688	0.097	-2.27e+06 1.96e+05
Stadium_Capacity	35.8961	8.136	4.412	0.000	19.563 52.229
Wins	5069.4959	1.32e+05	0.038	0.969	-2.59e+05 2.69e+05
GSR	-1.542e+04	1.37e+04	-1.126	0.265	-4.29e+04 1.21e+04
SRS	1.071e+05	6.56e+04	1.631	0.109	-2.47e+04 2.39e+05
Rank	1.377e+04	1.49e+04	0.922	0.361	-1.62e+04 4.37e+04

There was no major R-squared value change, and the likelihood of variables went up. The AIC value did significantly go up however we can modify another model to see what we can compare to. Our Wins variable still seems to have a high P-Value showing not a lot of influence on our equation.

OLS Regression Results

Dep. Variable:	TotalPay	R-squared:	0.863
Model:	OLS	Adj. R-squared:	0.832
Method:	Least Squares	F-statistic:	27.81
Date:	Wed, 27 Jan 2021	Prob (F-statistic):	1.22e-18
Time:	22:58:52	Log-Likelihood:	-989.60
No. Observations:	66	AIC:	2005.
Df Residuals:	53	BIC:	2034.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025 0.975]
Intercept	-4.95e+05	1.11e+06	-0.446	0.657	-2.72e+06 1.73e+06
Conference[T.ACC]	4.103e+05	6.12e+05	0.670	0.506	-8.18e+05 1.64e+06
Conference[T.Big 12]	4.305e+05	6.6e+05	0.652	0.517	-8.94e+05 1.75e+06
Conference[T.Big Ten]	1.051e+06	7.88e+05	1.333	0.188	-5.3e+05 2.63e+06
Conference[T.C-USA]	-6.198e+05	5.68e+05	-1.091	0.280	-1.76e+06 5.19e+05

```
Conference[T.MAC]        -6.7e+05   5.59e+05     -1.199      0.236   -1.79e+06
4.51e+05
Conference[T.Mt. West]  -9.86e+05   5.62e+05     -1.754      0.085   -2.11e+06
1.42e+05
Conference[T.Pac-12]    2.766e+05   5.87e+05      0.471      0.639   -9.01e+05
1.45e+06
Conference[T.SEC]       5.343e+05   7.27e+05      0.735      0.465   -9.23e+05
1.99e+06
Conference[T.Sun Belt] -8.357e+05    5.8e+05     -1.442      0.155      -2e+06
3.27e+05
Stadium_Capacity          37.4921      7.922      4.733      0.000      21.603
53.381
SRS                     1.136e+05   5.21e+04      2.181      0.034    9150.361
2.18e+05
Rank                    1.545e+04   1.44e+04      1.070      0.289   -1.35e+04
4.44e+04
==============================================================================
Omnibus:                        0.634   Durbin-Watson:                   2.199
Prob(Omnibus):                  0.728   Jarque-Bera (JB):                0.168
Skew:                           0.010   Prob(JB):                        0.919
Kurtosis:                       3.247   Cond. No.                     9.32e+05
==============================================================================
```

Removing Wins and GSR to see how the affect of GSR has on the equation, almost all values around R-Squared and AIC remained consistent. There are lower P-Values all around making the model more likely our hypothesis being true. After training and testing the models with subset data, the model most likely to be used will be the third model containing input variables; Conference, Stadium_Capacity, SRS, and Rank. Using fitted values within the data sets and implementing the models into the entire data set, a salary can be predicted for any school and coach within the set.

Recommendations and Outcomes

Using the model created a recommendation of salary for the head football coach of Syracuse can be as follows:

```
Predicted Pay for Syracuse Coach: $3,450,656
```

Considering this is higher compared to what is actually paid currently, this may be since there are no bonus payments issued to the coach as of now and the model is predicting based on what the other coaches are paid as a total including their bonus payments. If Syracuse were to transfer conferences into the Big Ten, the Syracuse School will have to be inputted into the Big Ten conference in the dataset and use the same model as used before.

```
Predicted Pay for Syracuse Coach if in Big 10: $3,618,445
```

The Pay is slightly larger taking into consideration of the Big Ten conference having a higher median salary amongst coaches compared to the ACC, which Syracuse is currently in.

Considering some schools had to be removed from the original dataset, it can let us see how adding additional information can change our models and make them more accurate. Schools that had to be removed were the independent schools (New Mexico State University,

Norte Dame, Liberty, Massachusetts, Brigham Young, and Army) as they were the only independent schools with TotalPay information and/or no TotalPay information within the conference. They had the potential to skew the salary data with little information. Some schools like Baylor, Rice, and Southern Methodist had to be removed as there were also no TotalPay values and taking medians of salaries within conferences was not a good choice as every school is different with pay.

Some variables that seem to have little impact on Coaches' pay would be things like Graduation Success Rates. As seen with the scatterplots, SEC coaches were highly paid with high GSR and even low GSR. Even in model evaluation, the third model had no GSR variable, yet the R-squared values and AIC had no major change. On the other hand, Stadium_Capacity, had the lowest P-Value within any model, showed pretty good correlation with TotalPay and can be seen increasing the TotalPay amount the higher the capacity went up within most schools. The model used for salary prediction was highly accurate but can be improved. There are multiple areas of affect that can be implemented into the equation to improve accuracy and decrease randomness. Things that can be added to improve model could be, school size, tuition, population near the school's stadium, and school merchandise sales.

Data References:

Coaches Data set and Salaries:

https://github.com/2SUBDA/IST_718/blob/master/Coaches9.csv

Graduation Rates:

https://www.ncaa.org/about/resources/research/shared-ncaa-research-data

Stadium Size:

https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv

Records:

https://www.sports-reference.com/cfb/years/2018-ratings.html