THE DATA:

How does a college or university determine how much to pay their football coach? Does the conference decide it? Or is there more that goes into it? Syracuse University is asking those questions and would also like to know if the coach were to move to another conference, such as the Big East of Big Ten, what would the salary be then.

They have an initial dataset named coaches9.csv, which contains the following variables:

	School	Conference	Coach	SchoolPay	TotalPay	Bonus	BonusPaid	AssistantPay	Buyout
0	Air Force	Mt. West	Troy Calhoun	885000	885000	247000	--	$0	--
1	Akron	MAC	Terry Bowden	$411,000	$412,500	$225,000	$50,000	$0	$688,500
2	Alabama	SEC	Nick Saban	$8,307,000	$8,307,000	$1,100,000	$500,000	$0	$33,600,000
3	Alabama at Birmingham	C-USA	Bill Clark	$900,000	$900,000	$950,000	$165,471	$0	$3,847,500
4	Appalachian State	Sun Belt	Scott Satterfield	$712,500	$712,500	$295,000	$145,000	$0	$2,160,417

Figure 1: coaches9.csv

Upon examination of coaches' names and schools, it was determined that this dataset is from either 2018 or 2019. This will be an important factor for when additional data is added to this set. Thus, if more data is to be added to this set, these are the years that should be focused on if possible.

Before adding additional data, this dataset should be cleaned and prepped for merging with other datasets. Upon first examination, it is discovered that there are $ signs and commas that will need to be removed, double dashes that should be changed to zeros, and the datatype will need to be changed to integers for the following variables: SchoolPay, TotalPay, Bonus, BonusPaid, AssistantPay, and Buyout.

```
(129, 9)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129 entries, 0 to 128
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   School        129 non-null    object
 1   Conference    129 non-null    object
 2   Coach         129 non-null    object
 3   SchoolPay     129 non-null    object
 4   TotalPay      129 non-null    object
 5   Bonus         129 non-null    object
 6   BonusPaid     129 non-null    object
 7   AssistantPay  129 non-null    object
 8   Buyout        129 non-null    object
```

```
dtypes: object(9)
memory usage: 9.2+ KB
None
                  School Conference              Coach    SchoolPay  \
0              Air Force   Mt. West       Troy Calhoun       885000
1                  Akron        MAC       Terry Bowden    $411,000
2                Alabama        SEC         Nick Saban  $8,307,000
3  Alabama at Birmingham      C-USA         Bill Clark    $900,000
4      Appalachian State   Sun Belt  Scott Satterfield    $712,500
5                Arizona     Pac-12       Kevin Sumlin  $1,600,000

      TotalPay        Bonus  BonusPaid AssistantPay        Buyout
0       885000       247000         --          $0            --
1    $412,500     $225,000    $50,000          $0      $688,500
2  $8,307,000   $1,100,000   $500,000          $0   $33,600,000
3    $900,000     $950,000   $165,471          $0    $3,847,500
4    $712,500     $295,000   $145,000          $0    $2,160,417
5  $2,000,000   $2,025,000         --          $0   $10,000,000
```

Figure 2: coaches9.svs in-depth examination

After cleaning and prepping the data:

	School	Conference	Coach	SchoolPay	TotalPay	Bonus	BonusPaid	AssistantPay	Buyout
0	Air Force	Mt. West	Troy Calhoun	885000	885000	247000	0	0	0
1	Akron	MAC	Terry Bowden	411000	412500	225000	50000	0	688500
2	Alabama	SEC	Nick Saban	8307000	8307000	1100000	500000	0	33600000
3	Alabama at Birmingham	C-USA	Bill Clark	900000	900000	950000	165471	0	3847500
4	Appalachian State	Sun Belt	Scott Satterfield	712500	712500	295000	145000	0	2160417

Figure 3: coaches9.svs cleaned

Now this data set is ready for additional data. The areas of focus for additional data are, graduation rates, win loss records, stadiums, and rankings. The graduation rates were obtained from NCAA and was for the school year 2017-2018. The rankings were also obtained from the NCAA, but it is the 2020 rankings. The win loss records and list of stadiums were both obtained from Wikipedia and they are through 2019. Once theses datasets have also been cleaned and prepped using the same steps as the initial dataset, they can be merged. It should be noted that during the merging of the datasets, the biggest challenge was how the college or university was named in the sets. For example, in the coaches9 data set the University of Texas-El Paso is Texas-El Paso, yet in other datasets it is UTEP or University of Texas at El Paso. As these differences were discovered they where corrected thus enabling a better matching process. The final dataset has 129 observations and 24 variables.

```
(129, 24)
<class 'pandas.core.frame.DataFrame'>
Int64Index: 129 entries, 0 to 90
Data columns (total 24 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   School        129 non-null    object
 1   Conference    129 non-null    object
 2   Coach         129 non-null    object
 3   SchoolPay     129 non-null    int32
 4   TotalPay      129 non-null    int32
 5   Bonus         129 non-null    int32
 6   BonusPaid     129 non-null    int32
 7   AssistantPay  129 non-null    int32
 8   Buyout        129 non-null    int32
 9   Won           129 non-null    int64
 10  Lost          129 non-null    int64
 11  Tied          129 non-null    int64
 12  Pct.          129 non-null    float64
 13  Years         129 non-null    int64
 14  Total Games   129 non-null    int64
 15  Stadium       129 non-null    object
 16  City          129 non-null    object
 17  State         129 non-null    object
 18  Capacity      129 non-null    object
 19  Cohort Year   129 non-null    int64
 20  Year          129 non-null    object
 21  Sport         129 non-null    object
 22  GSR           129 non-null    int64
 23  FGR           126 non-null    float64
dtypes: float64(2), int32(6), int64(7), object(9)
memory usage: 22.2+ KB
None
             School Conference          Coach  SchoolPay  TotalPay   Bonu
s  \
0     San Jose State   Mt. West  Brent Brennan     590424    590424  21000
0
44  Nevada-Las Vegas   Mt. West   Tony Sanchez     600000    600000  26000
0
6     Louisiana Tech      C-USA     Skip Holtz     700000    700000  39500
0
2    San Diego State   Mt. West     Rocky Long     872576    873576  72000
0
25  Georgia Southern   Sun Belt  Chad Lunsford     650000    650000  29500
0
1         New Mexico   Mt. West      Bob Davie     822690    823740  34000
0

    BonusPaid  AssistantPay   Buyout  Won  ...  Total Games  \
0           0             0  1476060  491  ...         1049
```

```
44          0             0   950000  558  ...         1085
6           0             0  2508333  627  ...         1127
2       95000             0   447412  565  ...         1024
25          0             0    83333   36  ...          119
1           0             0  1303294  488  ...         1127

                       Stadium         City  State  Capacity Cohort Year
\
0                CEFCU Stadium     San Jose     CA     30456        2011
44             Mackay Stadium         Reno     NV     26000        2011
6           Joe Aillet Stadium       Ruston     LA     28109        2011
2   Dignity Health Sports Park       Carson     CA     27000        2011
25         Center Parc Stadium      Atlanta     GA     25000        2011
1           Dreamstyle Stadium  Albuquerque     NM     39224        2011

         Year     Sport GSR   FGR
0   2017-2018  Football  80  74.0
44  2017-2018  Football  63  56.0
6   2017-2018  Football  65  53.0
2   2017-2018  Football  74  70.0
25  2017-2018  Football  62  50.0
1   2017-2018  Football  59  41.0
```

Figure 4: coachesjoined in-depth examination

VISUALIZATIONS:

Once the data has been merged, the process for understanding the dataset through visualizations should be done before building the model. The first chart shows a box plot with scatter plot of the points represented on each box plot. From this, it can be seen that the SEC has much higher coaches' salaries. The Big 12, ACC, Big Ten, and Pac-12, seem to have similar salary groupings, yet it should be noted that both the ACC and Big Ten have outliers that are above $6.5 million in salary. This could be due to these fews school having more successful programs than the rest of the conferences, such as Clemson university's football program in the ACC.

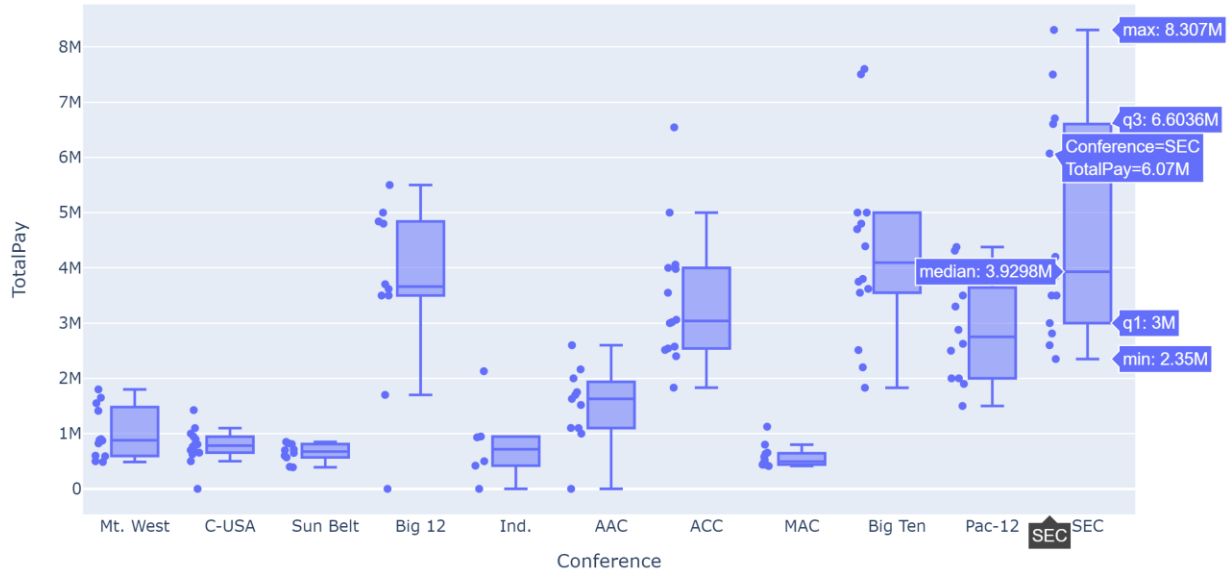

Figure 5: box and scatterplots of sum of TotalPay by Conference

The next visualization was also the box plot with scatter plot, yet for this chart the GSR or Graduation Success Rate was used in in place of Total Pay. This was done to get a better understanding of how the GSR compared across the conferences. This plot shows that the majority of conferences in this dataset had an average GSR is between 70 to 80 percent. The conferences that had an average above 80 percent were the ACC and Big Ten, both of which were in the second group for TotalPay by Conference above. It should be noted that there are many student athletes that do leave before they graduate to pursue a professional career in the NFL, while they are not the majority, they can still affect the outcome of the GSR. It seems that GRS may not be a strong influencer on coaches' TotalPay.

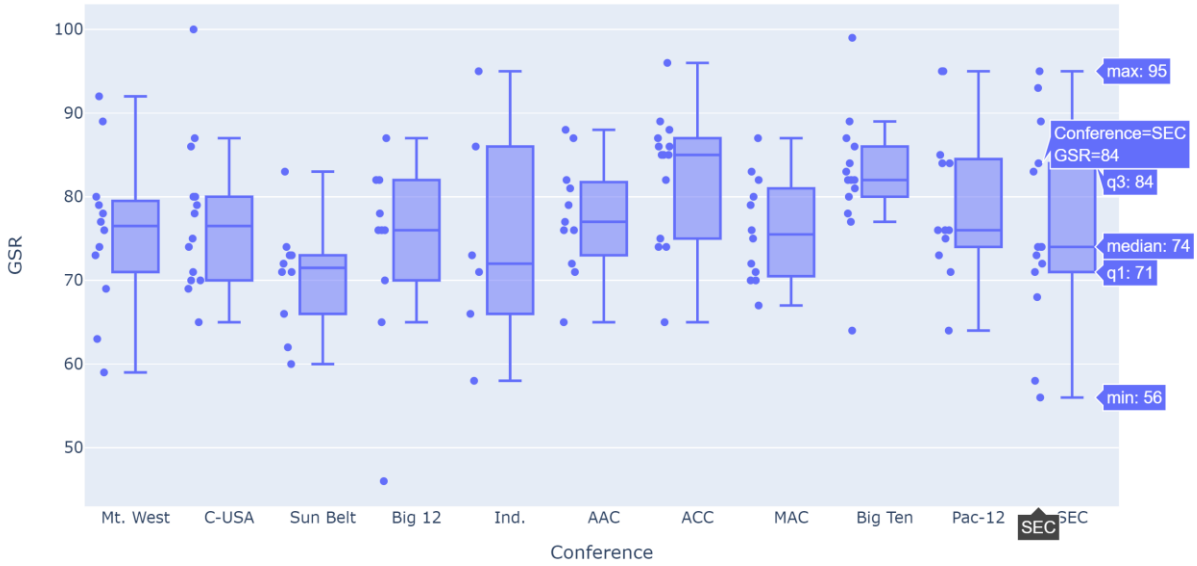

Figure 6: box and scatterplots of GSR by Conference

Another area that was examined, was the total amount of wins for each conference. To accomplish this, a bar chart was created showing the amount of wins by the conference. While the SEC and Big Ten tied for first with around 10,000 wins, they were followed by the ACC with around 9,000 and the Pac-12 with just under 8,000. These conference were also some of the top conferences for TotalPay for their coaches. While wins could be a strong influencer on coaches' salaries, it could also be a strong influencer on the universities games being broadcasted thus brining in extra income into the university. While not in this dataset, that extra income could also determine coaches' TotalPay especially with colleges and universities having to take possible budgets cuts into consideration due to COVID-19 creating shortages in state and federal budgets.

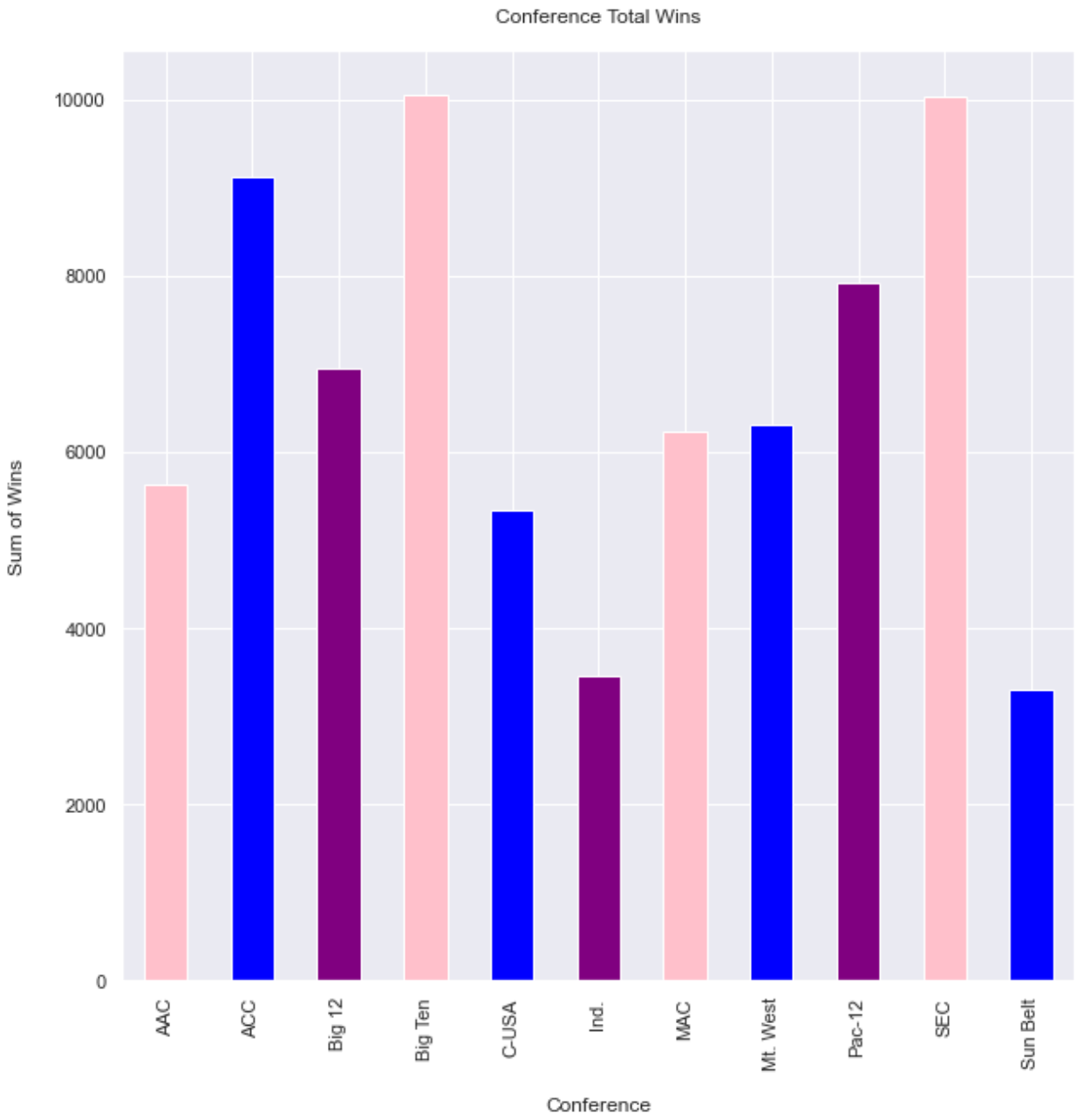

Figure 7: bar chart of Wins by Conference

The final plot is a scatter plot showing the wins and total games of each university and conference. From this scatter plot, it may be observed that again the tops conferences are the ACC, SEC, Big Ten, and Big 12. These are also the top conferences when it comes conferences when it comes to coaches' salaries, thus concluding that wins may be an influencer on how much a coach is indeed paid.

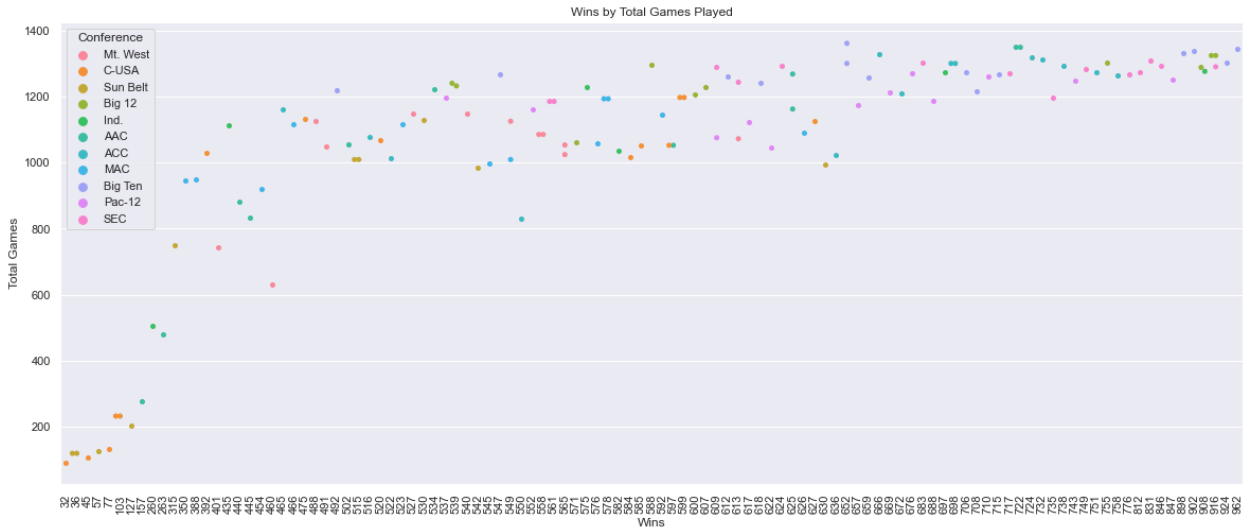

Figure 7: scatterplots of Wins by Total Games and Conference

MODEL:

The final step is building the model. There were quite a few challenges in building this model which could be due to the data not formatted right, multicollinearity within a variable, such as the conference variable, or another numerical issue. The following model was the best that obtained with this dataset. The R-squared value is 0.701 with the Prob (F-statistic) is under 0.05. It should be noted that under the conference, some of the conference have a p-value of less than 0.05, while some are above.

The following formula should be used for calculating coaches' salaries: y = 247700 + 1584000x + 1664000x + 2293000x – 404400x – 841800x – 942300x – 534800x + 1019000x + 263500x + 2457.0509

```
Proportion of Test Set Variance Accounted for:  0.719
                            OLS Regression Results
==============================================================================
Dep. Variable:                TotalPay   R-squared:                       0.701
Model:                            OLS   Adj. R-squared:                  0.673
Method:                 Least Squares   F-statistic:                     24.96
Date:                Sat, 17 Oct 2020   Prob (F-statistic):           8.96e-26
Time:                        18:59:00   Log-Likelihood:                -1969.8
No. Observations:                 129   AIC:                             3964.
Df Residuals:                     117   BIC:                             3998.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
====
                               coef    std err          t      P>|t|      [0.025      0.
975]
------------------------------------------------------------------------------
----
```

```
Intercept               2.477e+05   4.63e+05      0.535      0.594    -6.7e+05    1.17
e+06
Conference[T.ACC]       1.584e+06   4.47e+05      3.541      0.001    6.98e+05    2.47
e+06
Conference[T.Big 12]    1.664e+06   4.89e+05      3.399      0.001    6.94e+05    2.63
e+06
Conference[T.Big Ten]   2.293e+06   4.58e+05      5.010      0.000    1.39e+06     3.2
e+06
Conference[T.C-USA]    -4.044e+05   4.46e+05     -0.906      0.367   -1.29e+06    4.79
e+05
Conference[T.Ind.]     -8.418e+05   5.54e+05     -1.520      0.131   -1.94e+06    2.55
e+05
Conference[T.MAC]      -9.423e+05   4.54e+05     -2.075      0.040   -1.84e+06   -4.28
e+04
Conference[T.Mt. West] -5.348e+05   4.54e+05     -1.177      0.242   -1.43e+06    3.65
e+05
Conference[T.Pac-12]    1.019e+06   4.64e+05      2.197      0.030       1e+05    1.94
e+06
Conference[T.SEC]       2.635e+06   4.57e+05      5.759      0.000    1.73e+06    3.54
e+06
Conference[T.Sun Belt] -4.086e+05   4.89e+05     -0.835      0.405   -1.38e+06    5.61
e+05
Won                     2457.0509    639.196      3.844      0.000    1191.156    3722
.946
==============================================================================
Omnibus:                        8.693   Durbin-Watson:                   2.103
Prob(Omnibus):                  0.013   Jarque-Bera (JB):               12.142
Skew:                           0.348   Prob(JB):                      0.00231
Kurtosis:                       4.332   Cond. No.                     7.20e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly speci
fied.
[2] The condition number is large, 7.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 8: Final Model

CONCLUSION:

Using the formula above, it is recommended the Syracuse salary be set at $4,350,357.05 which does fall around the third quartile for the ACC which is the division they are in. If they were still in the Big East this salary would have been much lower as that conference was a failing conference which football was pulled from in 2013. If the coach would move to the Big Ten, they would probably stay the same but would have an increased room for growth as they are in the third quartile now, but that same amount is average in the Bog Ten.

DATA SOURCES:

Rankings:

https://www.ncaa.com/rankings/football/fbs/associated-press?utm_campaign=inline-article

Stadiums:

https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums

Win loss records

https://en.wikipedia.org/wiki/NCAA_Division_I_FBS_football_win-loss_records

Graduation rates:

http://www.ncaa.org/about/what-we-do/academics