Parin Patel

Predicting Coaching Staff Salary – Comparing to Graduation Rates.

Introduction:

College football is an important cultural and social aspect of an American college student's life. From campus newsletters to orientation walks, there is always some mention of the school's football team or stadium. In fact, many students use the success of the universities football team as a deciding factor in which college to apply and accept admissions to. Therefore, universities spend heavily on making sure their team are well equipped and talented enough to win championships. To achieve this goal, they need to hire successful coaches, which can often come with a heavy price. And while some coaches make more than college chancellors, the revenue they bring in through tickets and television time is incomparable. This makes coaches a good value to the university they coach at.

However, it's important to consider that the value of a university ideally should be its academic programs and intellectual rigor. Therefore , while a coach may be be considered a "good value" because he brings in money and notoriety to the college; should his salary be contingent on the graduation success rates also? This is an extremely important question to consider, especially since states like California are considering allowing NCAA (college athletes) to make money from endorsements and sponsorships.

The goal of this study is to predict the recommended salary for a Syracuse Football Coach and determine what the single biggest impact on his salary would be. In addition, we wanted to compare his recommended salary to if Syracuse was still in the Big East or Big 10 Conferences. Additionally, we will look at what schools were dropped in our analysis, and why. We will then compare graduation rates to projected salary and determine the accuracy of our model.

Analysis:

Data Pre-Processing: Importing Datasets and Cleaning

In total, the following four datasets were imported throughout our study:

1. Coaches – a file that had data related to Pay for Coaches and their conference, and associated school.
2. Stadium- a file that was scraped from a Wikipedia page (https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums) that lists all the NCAA football stadiums and their capacity.
3. GSR (Graduation Rate Data) – which was downloaded from https://web3.ncaa.org/aprsearch/gsrsearch website and contains Graduation rates for football programs.

4. Coaches Win and Loss Records- This file was created from this website : https://www.sports-reference.com/cfb/coaches/a-index.html . It lists all NCAA football coaches and their Wins and Losses and their last associated schools.

All files were cleaned and then merged to form a master sheet for analysis.

Cleanup for all files was pretty standard, with the Coaches, Stadium, and Coaches Win and Loss Records files requiring the most cleaning. For examples, in these datasets, we removed rows with "—" listed and had cells regarding pay remove the "$" symbol. In addition, we removed any rows that were blank. Finally, any unnecessary columns were removed, as well. Columns had to be renamed, in most datasets, because the original contained spaces between words or symbols. For example "State/Province" became "State". In addition, for the stadium datasets, we had to convert the object columns in the data frame to numeric. In addition, this dataset required more complex symbol removal.

Before we started to merge files, we had to make sure our datasets were in the correct format and check to see if we could merge based on comparable variables. So our first check was to make sure our stadiums dataset could be merged with our coaches dataset along the lines of school and teams. We used lambda and fuzzy string matching to help us achieve this. Specifically, the levenshtein ratio and distance was calculated to find the distance between two string values. If the ratio_calc is true, then the function will compute the levenshtein distance ratio of similarity between two strings. The function would then print the matches of schools where our levenshtein distance is greater than 0.7.

When the datasets were merged we first merged the coaches dataset with the Coaches Win and Loss Records dataset by Coaches. Prior to merging these files with the stadium data., a temporary stadium object was created and then merged with the Coaches+CoachesWin/LossRecords dataset. This temporary copy was created for the process of merging to preserve the original stadium file, in addition to creating a checking to make sure the stadium and coaches file could be merged by Team and School . Finally, the Graduation Success Rate dataset was merged with the previously combined dataset by "School" to result in the final merged product. This final merged dataset was then further cleaned to remove any rows where TotalPay was either 0 or less than 0 and any rows where there was a duplicate school.

I realized at this point that I wanted to have a way to rank the data by coaches win-loss % (Pct). We added this to the datatable, and then created a top 25 column where if the coach and school had a rank of 25 or more, then they were given the value 0. Finally, our final steps for cleaning our merged data included removing all null values, renaming columns, and subsetting the variables so only the following columns were left in the final dataset:

- Rank- Rank of the School/Coach by Pct
- School- NCCAA University or College.
- Conference – Althetic conference college is in.
    - Big12
    - Big10

  - ACC
  - Pac-12
  - SEC
  - Mt.West
  - Ind.
  - AAC
  - C-USA
  - Sun Belt
  - MAC
- Coach – NCAA Football Coach
- TotalPay- Total Compensation excluding Bonus
- Capacity – Stadium seating capacity based on School Stadiums
- GSR- Graduation Success Rate
- Top25- based on Pct.
- Pct – Win/Loss Percentage
- Stadium- Home Team Stadium, based on School.
- Yrs- Years coach was with that school.

Exploratory Analysis:

We started by graphing a Correlation Heat Map to compare the variable (Figure 1 and Table 1 below). We can see that Rank,Pct, and Top25 have the highest negative correlation. This would make sense since Top25 is based off of Rank, which is based off Pct. As rank increases, we would expect the Pct to decrease since the highest rank (#1) would have the highest score. In the future we should make the rank descending with the highest ranking number being #100 or something similar.   We can also see that TotalPay and Rank are pretty strongly correlated, in addition to Rank and Yrs. This latter one is interesting, it shows that as a coach with less experience is less likely to have a higher rank. Some of the strongest positive correlations come between TotalPay and Capacity of the stadium. This supposes that it is possible that a coaches salary may be partly influenced by the size of the stadium. Since larger stadiums have more seats, therefore more tickets to sell.

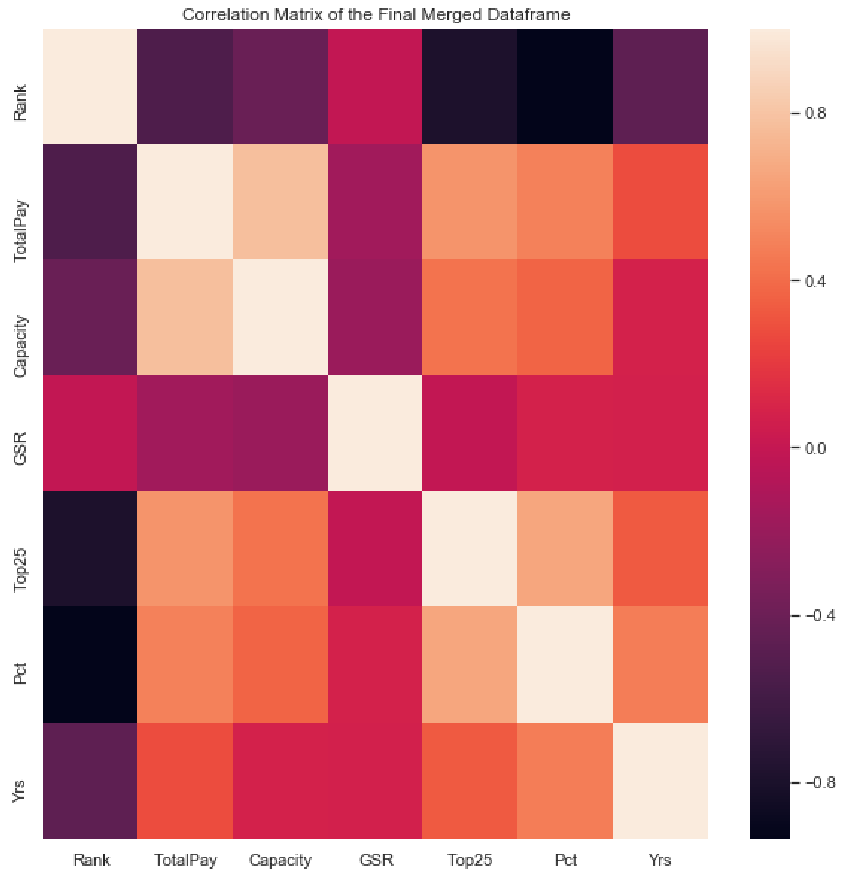
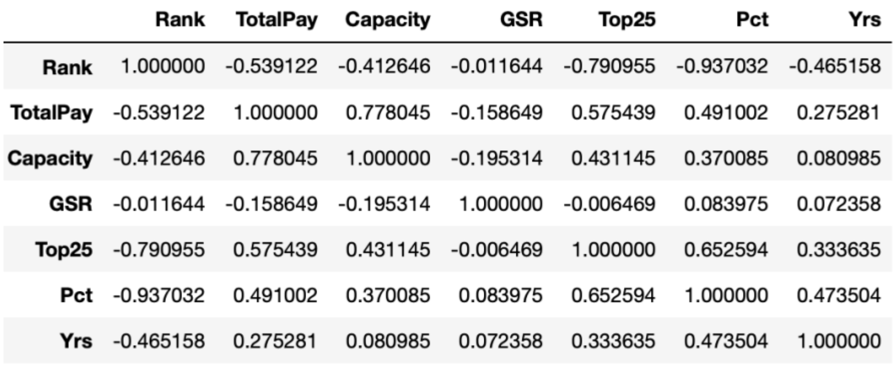


Figure 1: Correlation Matrix of Final Merged Dataframes

Table 1: Correlation ratios of Final Merged Dataframe Variables



	Rank	TotalPay	Capacity	GSR	Top25	Pct	Yrs
Rank	1.000000	-0.539122	-0.412646	-0.011644	-0.790955	-0.937032	-0.465158
TotalPay	-0.539122	1.000000	0.778045	-0.158649	0.575439	0.491002	0.275281
Capacity	-0.412646	0.778045	1.000000	-0.195314	0.431145	0.370085	0.080985
GSR	-0.011644	-0.158649	-0.195314	1.000000	-0.006469	0.083975	0.072358
Top25	-0.790955	0.575439	0.431145	-0.006469	1.000000	0.652594	0.333635
Pct	-0.937032	0.491002	0.370085	0.083975	0.652594	1.000000	0.473504
Yrs	-0.465158	0.275281	0.080985	0.072358	0.333635	0.473504	1.000000

From Figure 2 and 3 below we can see that the total pay of all coaches averages around $1.0 mil to $2.5 mil, with coaches in the Top25 making more. Coaches in the Top25 average around $4.0 mil to $5.0 mil. This is a significant difference more. However, it’s important to note that coaches in the Top25 also have, on average, an overall larger stadium capacity.(See Figure 4).

As mentioned earlier, this means more ticket revenue for the teams. On average, we see a difference in capacity being around 250,000.

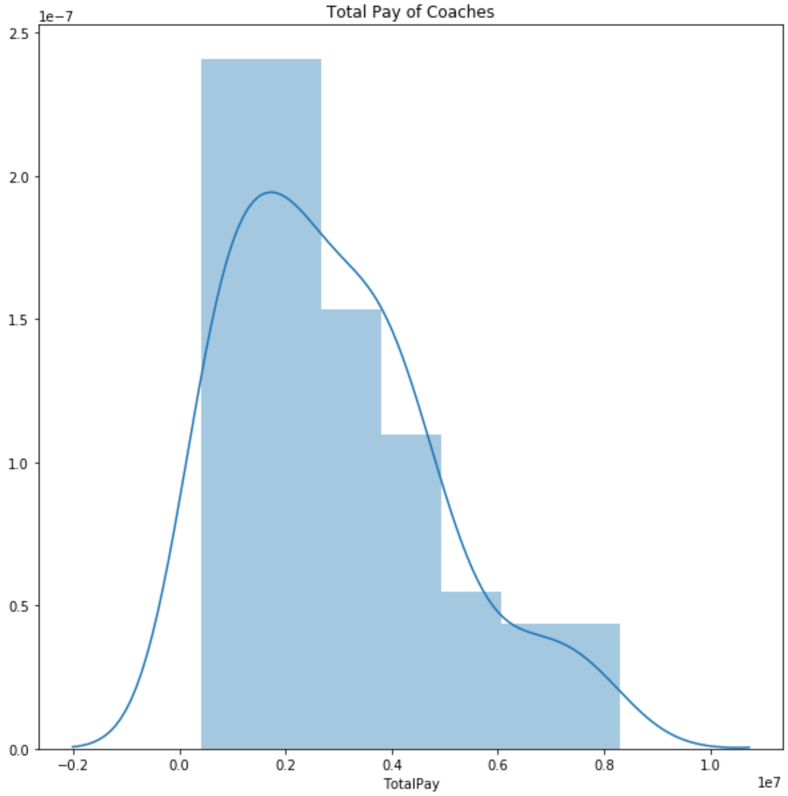


Figure 2:Total Pay of Coaches

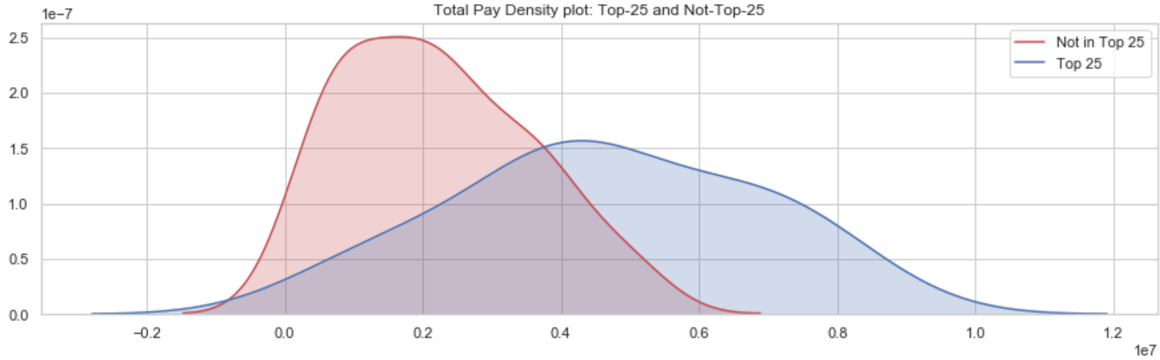


Figure 3:Total Pay Density plot: Top-25 and Not-Top-25

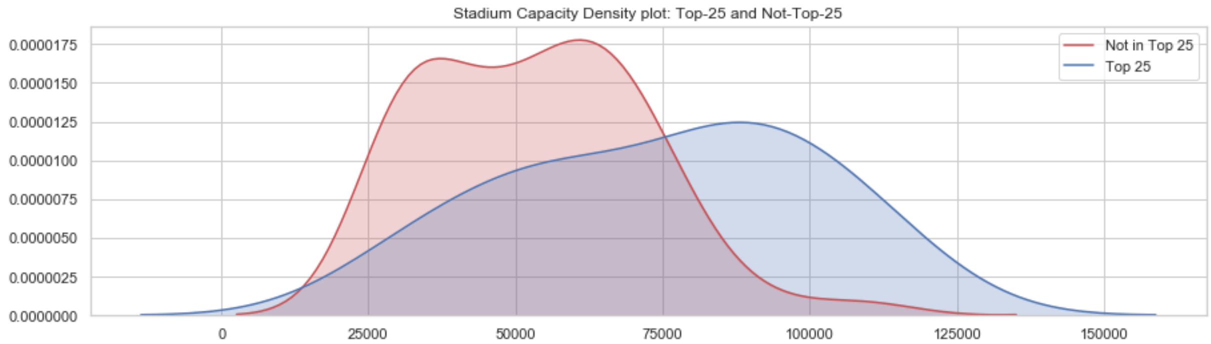


Figure 4:Stadium Capacity Density plot: Top-25 and Not-Top-25

Next, we decided to explore the relationship between the Top25 and their conferences (see figure 5). We can see that the top-25 most frequently occurring conference is the SEC. It is followed by the Big12 and then Big Ten. For Not-Top25, the most frequently occurring conference is the Big Ten. We can see from figure 6 that the TotalPay of the coaches follows the trend, where the SEC, Big12, and Big Ten have the highest overall TotalPay.

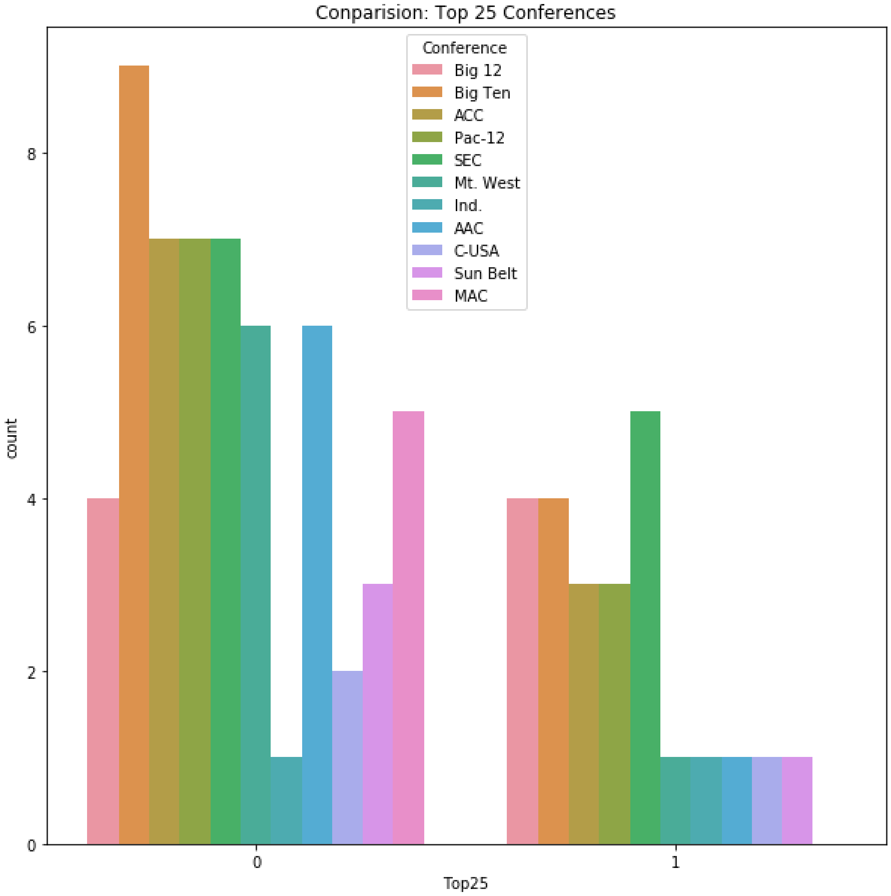


Figure 5:Comparision: Top 25 Conferences

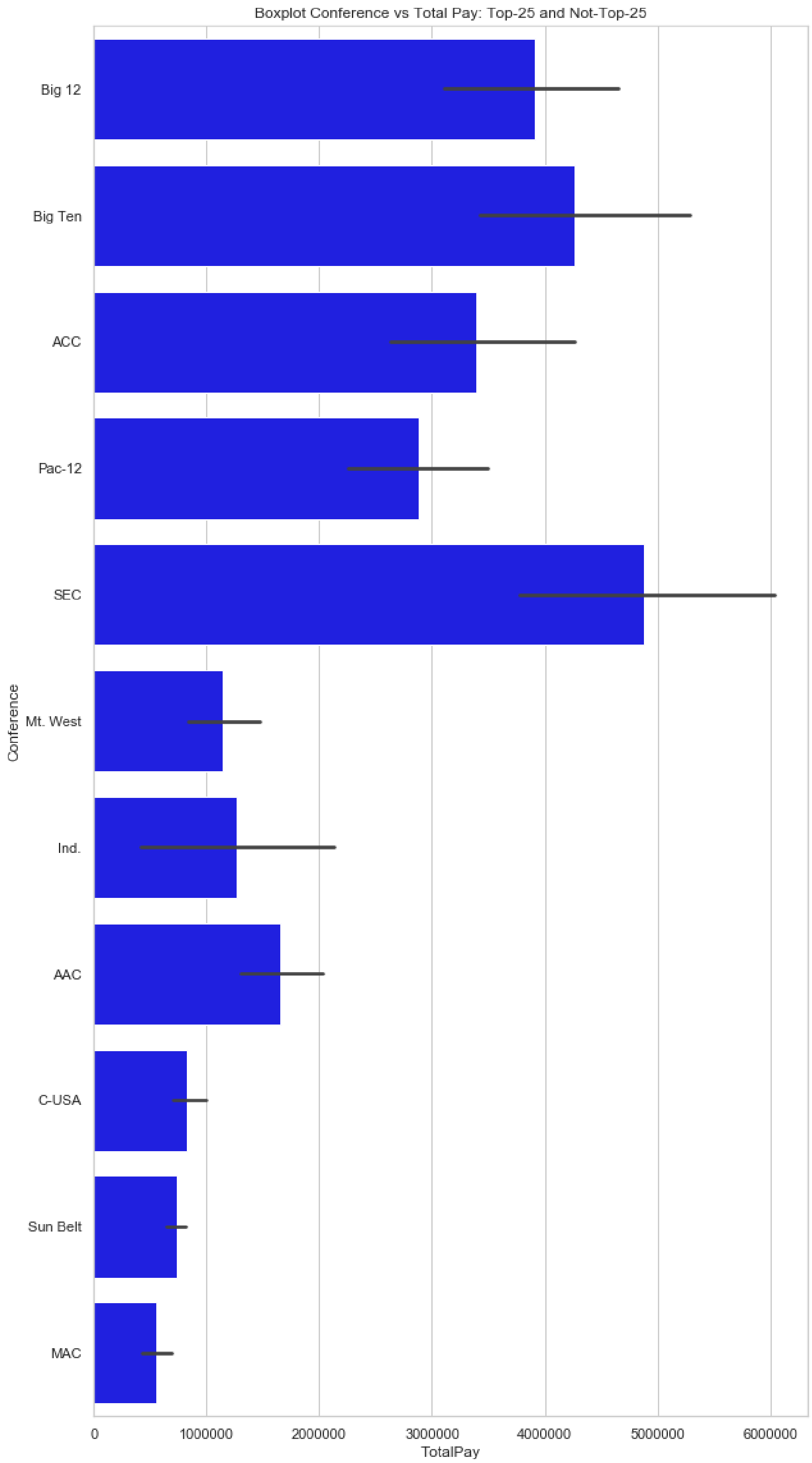


Figure 6: Boxplot Conference vs Total Pay: Top-25 and Not-Top-25

Next, we explored the relationship between graduation success rates and Rank, TotalPay, and conference. Interestingly, right off the bat in Figure 7, we can easily see that graduation rates are only slightly higher for higher ranking schools. From the scatterplot (figure 9), we can see that the most of the values are clustered around the middle-left-hand side of the plot. This makes sense considering that our calculated mean of GSR (Figure 8) is 63.26. The scatterplot shows us a few more trends: First that the highest paid coach has a graduation rate well below the mean. In addition, we see that of the two schools with the two highest graduation, one of them is actually in an independent conference (Notre Dame) and the other one is in the ACC (Wake Forest). Both of these schools' coaches make well below the mean TotalPay amounts, which is about 2.9 million . From Figure 10 we can see a stark difference to the box plot comparing pay to conference (Figure 6). Instead, Figure 10 shows us that the conferences with the highest GSR is not the SEC or Big 12 or Big 10. Instead, the ACC and the Ind. Conferences have the highest GSR, while the SEC the lowest GSR.

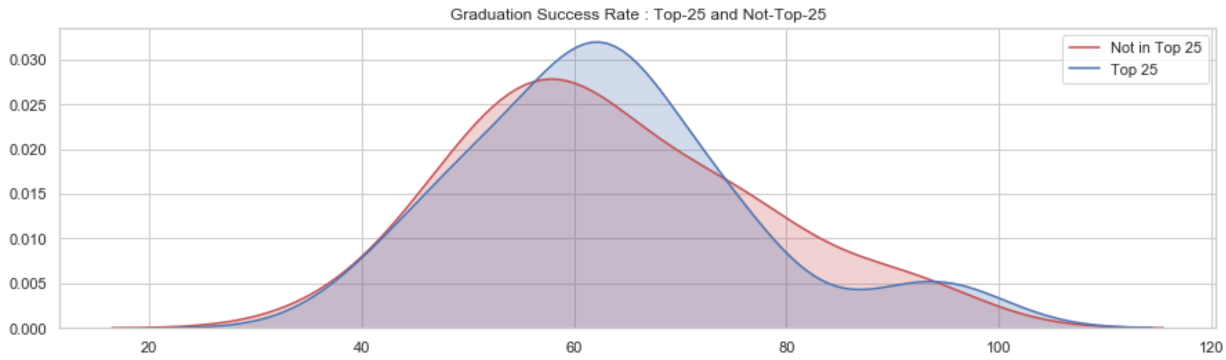
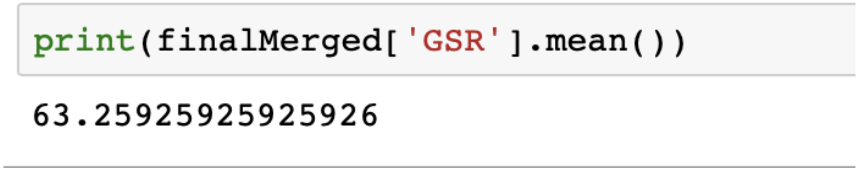


Figure 7: Graduation Success Rate : Top-25 and Not-Top-25'



```
print(finalMerged['GSR'].mean())

63.25925925925926
```

Figure 8: Calculated mean of GSR rates

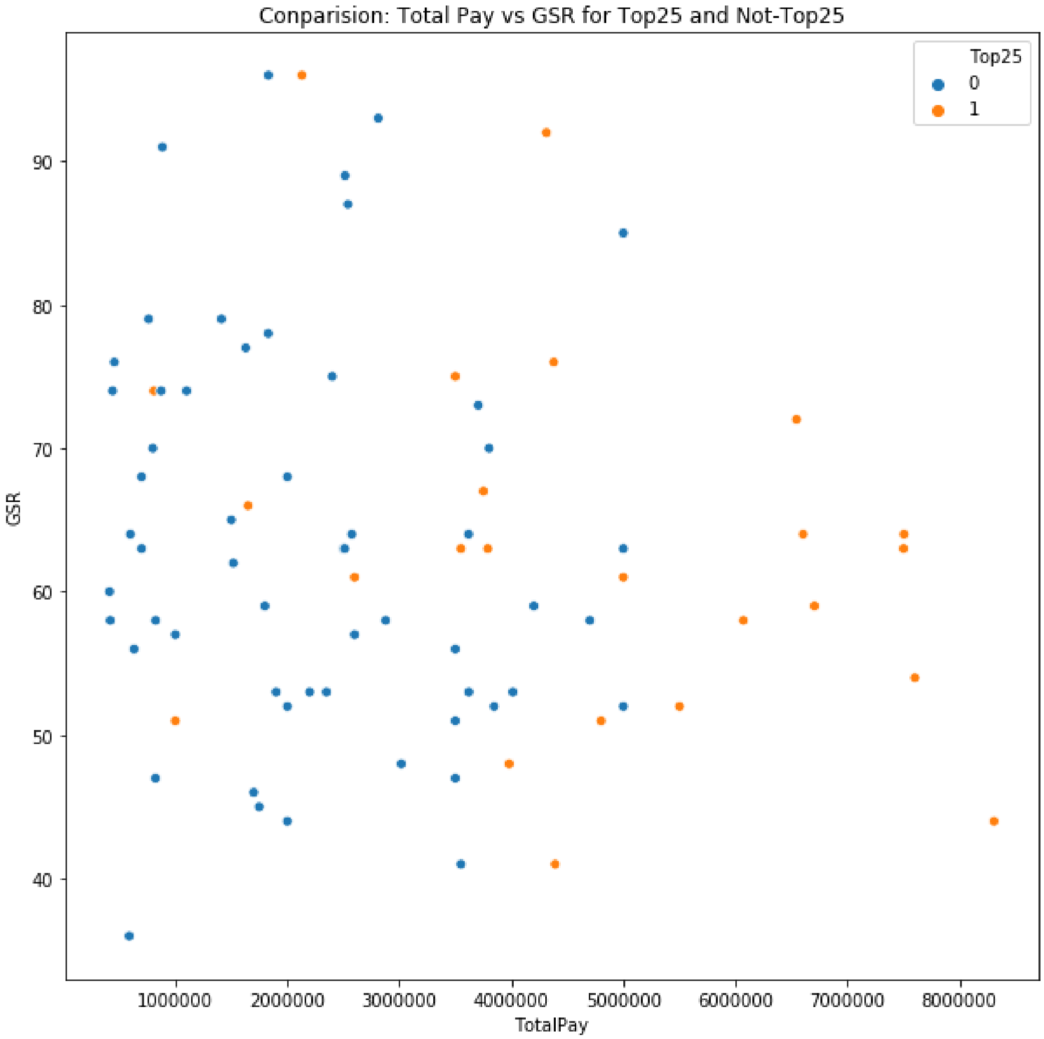


Figure 9: Comparision: Total Pay vs GSR for Top25 and Not-Top25

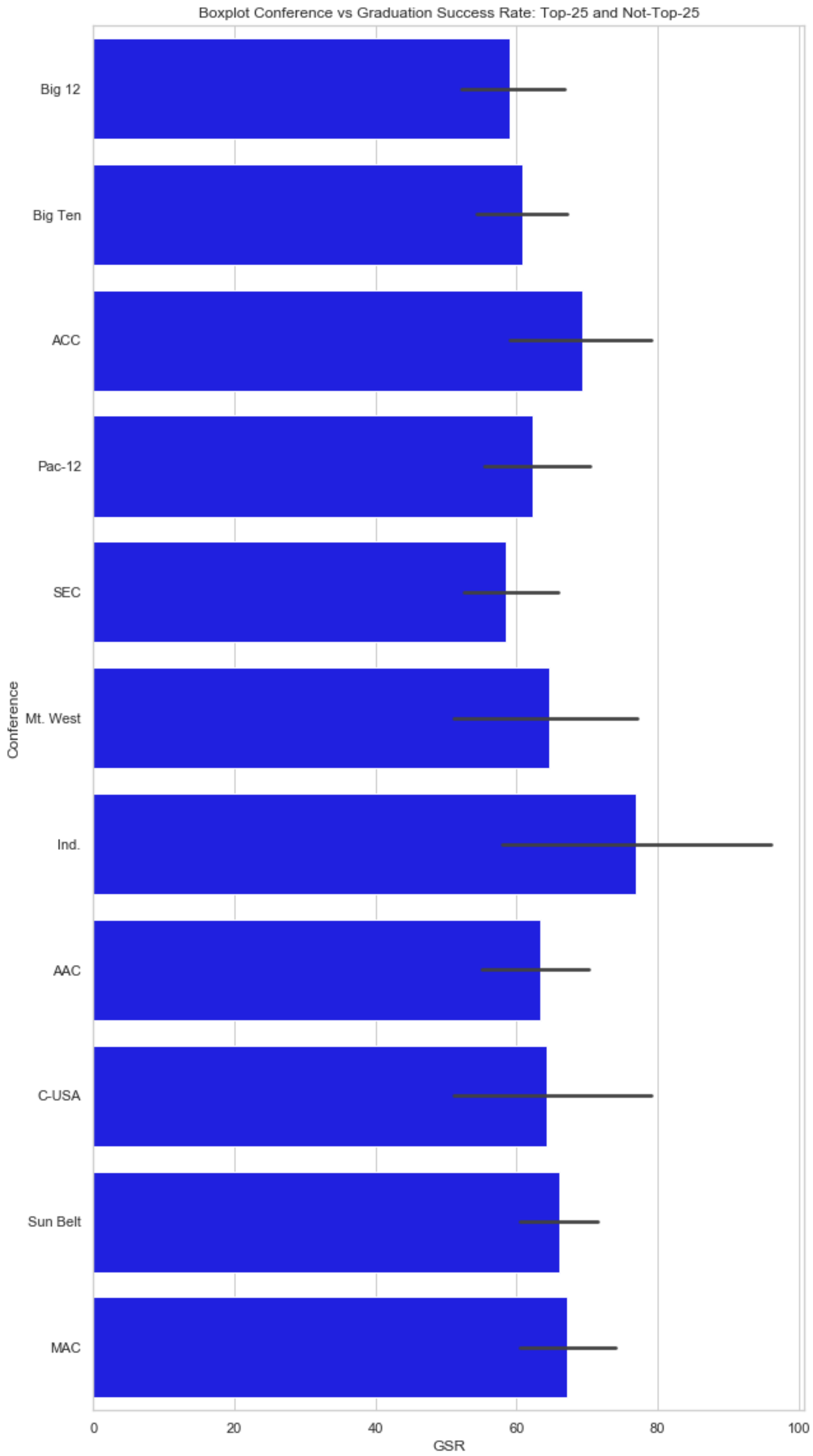


Figure 10: Boxplot Conference vs Graduation Success Rate

Regression Modeling:

We created a training dataset with 2/3 of the original dataset, while the remaining 1/3 was placed in a testing dataset. We used sklearn to condut our Linear Regression Model and used scipy for our ols model (ordinary least squares).

Our first ols model started with several independent variables (Rank + Conference + Top25 + GSR + Capacity + Pct + Yrs + Capacity) while our response variable was TotalPay. In the end, we had narrowed our independent variables to Capacity, GSR. TotalPay, and Top25. TotalPay was still our response variable. For our Linear Regression model, we used Capacity, GSR, and Pct as our Independent variables.

Results:

From our r-squared from our Linear Regression model was 0.46. This shows that our linear regression model is not very good for accounting for variance of the coaches salary with the selected variables.

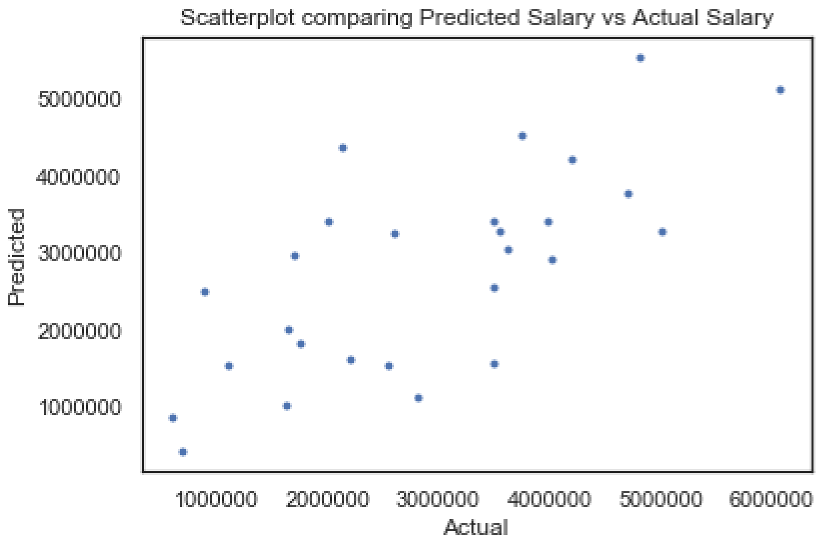
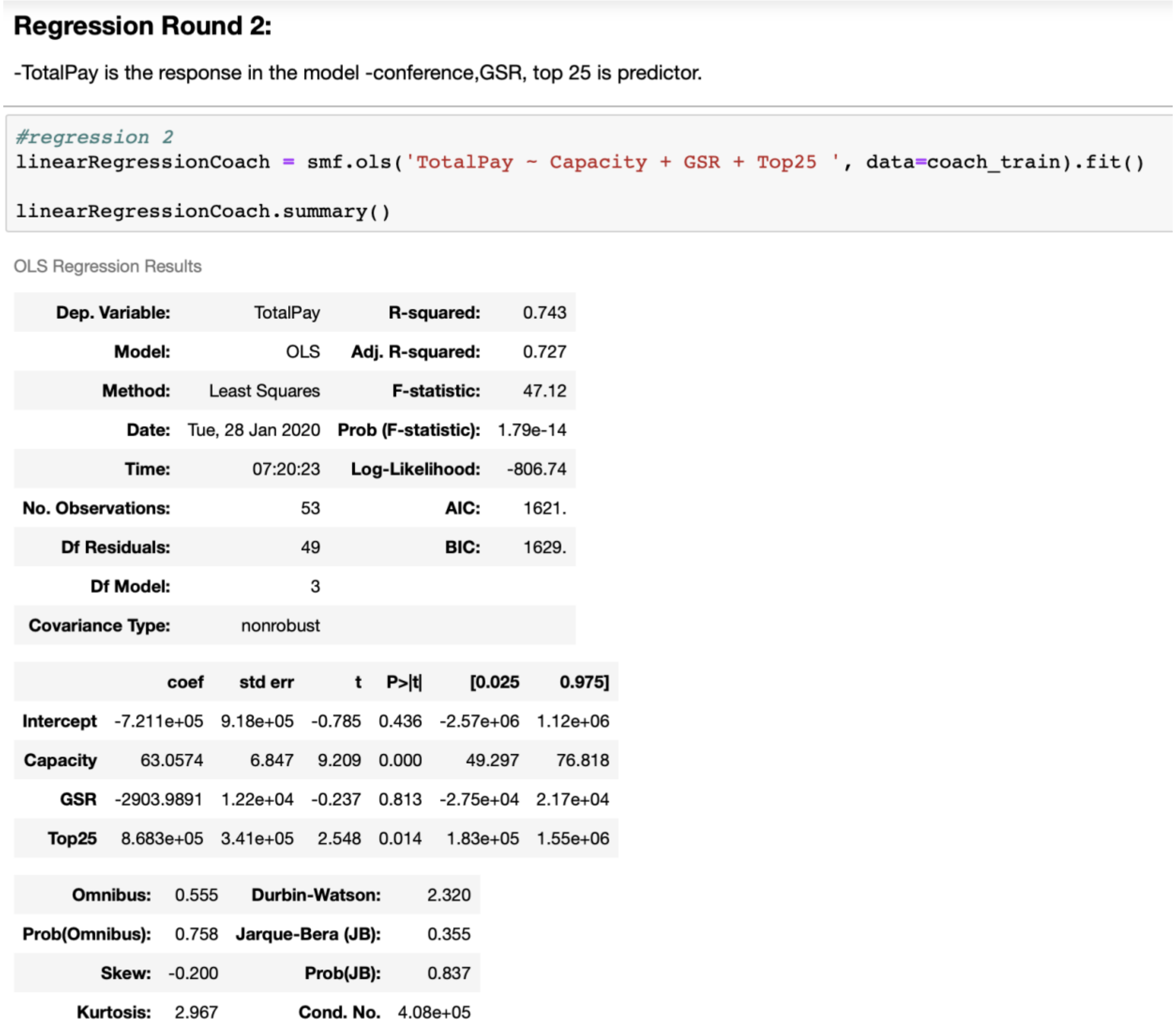


Figure 11: Linear Regression Results Scatterplot (Predicted vs Actual Salary)

From our OLS regression model we were able to generate better insights. We see that our our P-value is1.79e-14. This is much smaller than our alpha of 0.01. Therefore, there is much less than 1% chance that the F-statistic of 47.12 could have occurred by chance under the assumption of a valid Null hypothesis. Therefore, we will reject the Null hypothesis and accept the alternate that our model is significant. And despite a lower than expected R-squared of 0.74, we can somewhat explain the variance in the dependent variable TotalPay better than the Linear Regression model.



Regression Round 2:

-TotalPay is the response in the model -conference,GSR, top 25 is predictor.

```
#regression 2
linearRegressionCoach = smf.ols('TotalPay ~ Capacity + GSR + Top25 ', data=coach_train).fit()

linearRegressionCoach.summary()
```

OLS Regression Results

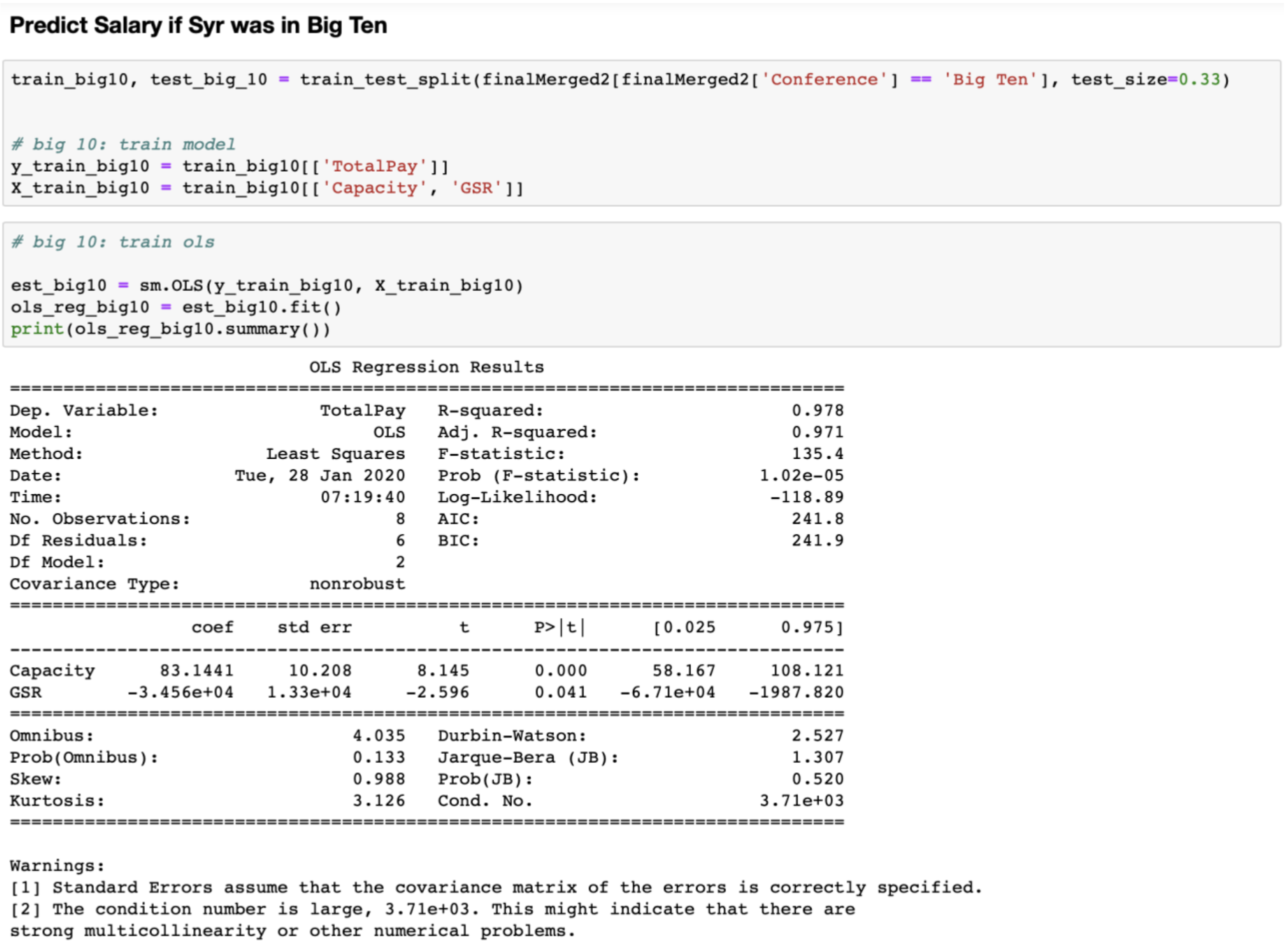
Dep. Variable:	TotalPay	**R-squared:**	0.743
Model:	OLS	**Adj. R-squared:**	0.727
Method:	Least Squares	**F-statistic:**	47.12
Date:	Tue, 28 Jan 2020	**Prob (F-statistic):**	1.79e-14
Time:	07:20:23	**Log-Likelihood:**	-806.74
No. Observations:	53	**AIC:**	1621.
Df Residuals:	49	**BIC:**	1629.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-7.211e+05	9.18e+05	-0.785	0.436	-2.57e+06	1.12e+06
Capacity	63.0574	6.847	9.209	0.000	49.297	76.818
GSR	-2903.9891	1.22e+04	-0.237	0.813	-2.75e+04	2.17e+04
Top25	8.683e+05	3.41e+05	2.548	0.014	1.83e+05	1.55e+06

Omnibus:	0.555	**Durbin-Watson:**	2.320
Prob(Omnibus):	0.758	**Jarque-Bera (JB):**	0.355
Skew:	-0.200	**Prob(JB):**	0.837
Kurtosis:	2.967	**Cond. No.**	4.08e+05

Figure 12: OLS Regression Results

Finally ,we used our ols and linear regression methods to answer the question below about what we would expect a Syracuse coach to make and how the predicted salary would differ if the coach was in the Big Eat or Big 10? Since there were no BigEast conferences in the dataset, we based the prediction off of only the Big10. See Figure 13, below for the summary. Overall, the results yielded a signifigant probability F-Statistic that the overall salary would be much higher (about 3.71 million Total Pay) if Syracuse was in the Big10. In addition to having a p-value much lower than the alpha, the R-squared was found to 0.98.



Predict Salary if Syr was in Big Ten

```
train_big10, test_big_10 = train_test_split(finalMerged2[finalMerged2['Conference'] == 'Big Ten'], test_size=0.33)


# big 10: train model
y_train_big10 = train_big10[['TotalPay']]
X_train_big10 = train_big10[['Capacity', 'GSR']]
```

```
# big 10: train ols

est_big10 = sm.OLS(y_train_big10, X_train_big10)
ols_reg_big10 = est_big10.fit()
print(ols_reg_big10.summary())
```

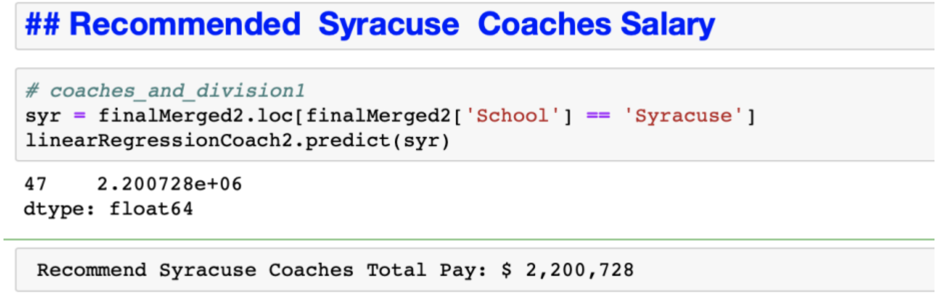
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.971
Method:                 Least Squares   F-statistic:                     135.4
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           1.02e-05
Time:                        07:19:40   Log-Likelihood:                -118.89
No. Observations:                   8   AIC:                             241.8
Df Residuals:                       6   BIC:                             241.9
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Capacity      83.1441     10.208      8.145      0.000      58.167     108.121
GSR        -3.456e+04   1.33e+04     -2.596      0.041   -6.71e+04   -1987.820
==============================================================================
Omnibus:                        4.035   Durbin-Watson:                   2.527
Prob(Omnibus):                  0.133   Jarque-Bera (JB):                1.307
Skew:                           0.988   Prob(JB):                        0.520
Kurtosis:                       3.126   Cond. No.                     3.71e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 13: OLS model predict if SYR was in Big10

Conclusion: Our Analysis Above Answered The Following Questions

1. **What is the recommended salary for a Syracuse football coach?**
   - Recommend Syracuse Coaches Total Pay: $ 2,200,728



## Recommended Syracuse Coaches Salary

```
# coaches_and_division1
syr = finalMerged2.loc[finalMerged2['School'] == 'Syracuse']
linearRegressionCoach2.predict(syr)
```

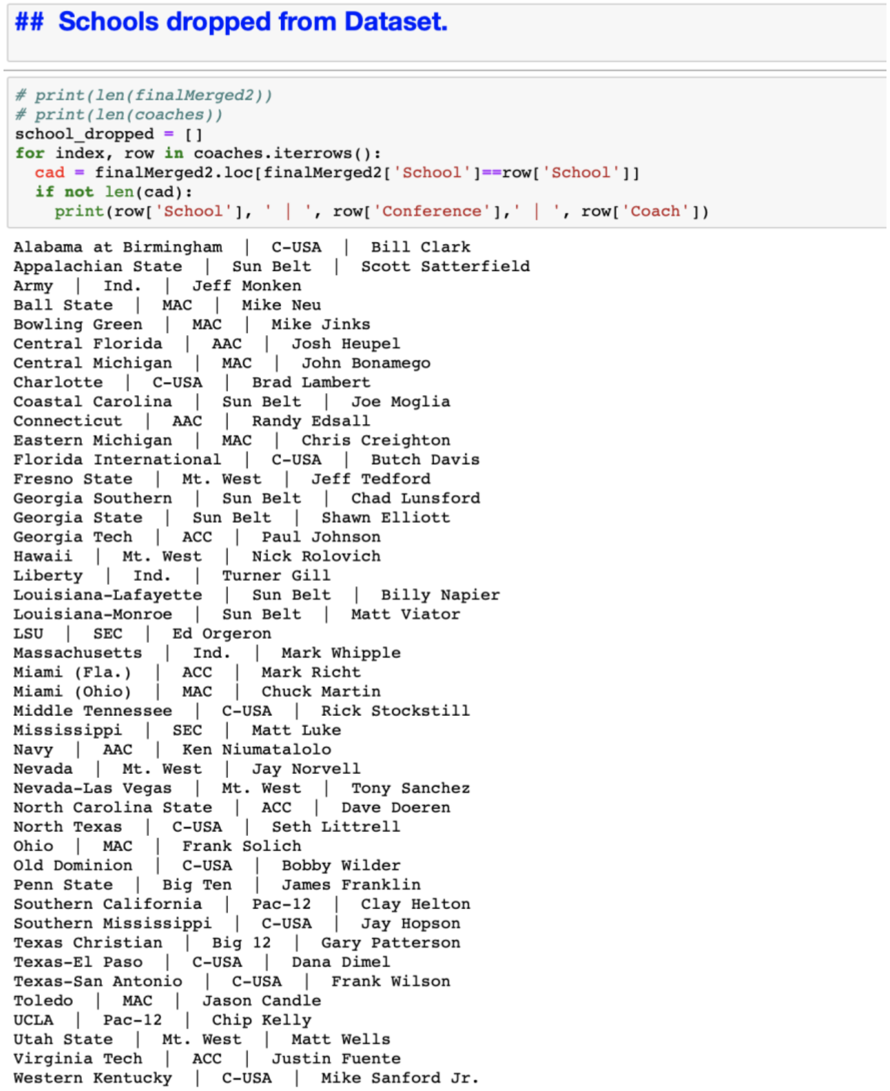
```
47    2.200728e+06
dtype: float64
```

```
Recommend Syracuse Coaches Total Pay: $ 2,200,728
```

Figure 14: Recommended Syr Coach Salary Linear Regression Predict

2. **What would his salary be if Syracuse was still in the Big East? What about Big10?**
    - His salary would be much higher, about $3.71 million (see Figure 13).

3. **What schools were dropped from the data. And why?**
    - The following schools in Figure 15 below were dropped from the dataset for various reasons. When we merged the coaches and stadium datasets by School and HomeTeam, we ran into some naming convention issues. In some cases, the name of the school in one dataset did not match the name of the school in the other, for example one dataset had “University of Alabama at Birmingham” while the other had “University of Alabama Birmingham”. In the future a different method of fuzzy matching should be utilized, or better data quality checks should be in place to make sure the names match. Other schools were dropped due to missing GSR information or Coach information.



Schools dropped from Dataset.

```
# print(len(finalMerged2))
# print(len(coaches))
school_dropped = []
for index, row in coaches.iterrows():
  cad = finalMerged2.loc[finalMerged2['School']==row['School']]
  if not len(cad):
    print(row['School'], ' | ', row['Conference'],' | ', row['Coach'])
```

```
Alabama at Birmingham  |  C-USA  |  Bill Clark
Appalachian State  |  Sun Belt  |  Scott Satterfield
Army  |  Ind.  |  Jeff Monken
Ball State  |  MAC  |  Mike Neu
Bowling Green  |  MAC  |  Mike Jinks
Central Florida  |  AAC  |  Josh Heupel
Central Michigan  |  MAC  |  John Bonamego
Charlotte  |  C-USA  |  Brad Lambert
Coastal Carolina  |  Sun Belt  |  Joe Moglia
Connecticut  |  AAC  |  Randy Edsall
Eastern Michigan  |  MAC  |  Chris Creighton
Florida International  |  C-USA  |  Butch Davis
Fresno State  |  Mt. West  |  Jeff Tedford
Georgia Southern  |  Sun Belt  |  Chad Lunsford
Georgia State  |  Sun Belt  |  Shawn Elliott
Georgia Tech  |  ACC  |  Paul Johnson
Hawaii  |  Mt. West  |  Nick Rolovich
Liberty  |  Ind.  |  Turner Gill
Louisiana-Lafayette  |  Sun Belt  |  Billy Napier
Louisiana-Monroe  |  Sun Belt  |  Matt Viator
LSU  |  SEC  |  Ed Orgeron
Massachusetts  |  Ind.  |  Mark Whipple
Miami (Fla.)  |  ACC  |  Mark Richt
Miami (Ohio)  |  MAC  |  Chuck Martin
Middle Tennessee  |  C-USA  |  Rick Stockstill
Mississippi  |  SEC  |  Matt Luke
Navy  |  AAC  |  Ken Niumatalolo
Nevada  |  Mt. West  |  Jay Norvell
Nevada-Las Vegas  |  Mt. West  |  Tony Sanchez
North Carolina State  |  ACC  |  Dave Doeren
North Texas  |  C-USA  |  Seth Littrell
Ohio  |  MAC  |  Frank Solich
Old Dominion  |  C-USA  |  Bobby Wilder
Penn State  |  Big Ten  |  James Franklin
Southern California  |  Pac-12  |  Clay Helton
Southern Mississippi  |  C-USA  |  Jay Hopson
Texas Christian  |  Big 12  |  Gary Patterson
Texas-El Paso  |  C-USA  |  Dana Dimel
Texas-San Antonio  |  C-USA  |  Frank Wilson
Toledo  |  MAC  |  Jason Candle
UCLA  |  Pac-12  |  Chip Kelly
Utah State  |  Mt. West  |  Matt Wells
Virginia Tech  |  ACC  |  Justin Fuente
Western Kentucky  |  C-USA  |  Mike Sanford Jr.
```

Figure 15: Schools dropped from dataset

4. **What effect does graduation rates have on projected salary?**
    - Interestingly, from Figure 7 to 9 above, we saw that graduation rates actually had very little to do with projected salary. In addition, when we ran our ols model Round 1, we saw that GSR had little significance to TotalPay.

5. **What is the single biggest impact on salary size?**
    - Interestingly, our results showed us that the single biggest impact on salary size was stadium capacity followed by conference. This is likely correlated to our statement earlier that coaches that play in larger stadiums are likely to have programs that generate higher ticket sales, more alumni, higher television viewership, and therefore, more sponsorship and donations. This could mean that their program has a higher salary cap to spend on coaching staff.

Overall, our study showed that the OLS model yielded better results .However, both techniques yielded about the same prediction of what the Syracuse coaches should make as a salary. In addition to if the team moved to the Big10. In the future, the model can be bettered by including more coaching pay data, especially for teams in the Big East and Indep. I would say this analysis is a good start, however, we also need better GSR data. The dataset found on the NCAA portal only provided it for some colleges every year. It would be better to have updated GSR information for every NCAA college with a foodball program.