

July 25, 2021

Quinn E Knudsen

IST 718

Syracuse University

The PRice is Right

Modeling to Predict Syracuse’s Football Coach Salary

**Table of Contents**

**Introduction**……………………………………………………………………………......pg 2

Areas of Focus …………………………………………………………….……......pg 2

**About the Data..**…………………………………………………………………..….….…pg 2

Data Overview ...……………………………………………………………….…...pg 2

**Exploratory Data Analysis & Correlations**……………………………………..….….…pg 3

Salary Distribution..…………………………………………………………….…...pg 3

Correlation Table..…………………………………………………………….….....pg 3

Scatterplots………..…………………………………………………………….…..pg 4

**Modeling**....………………………………………………………………………..………..pg 5

Final Model……….…..…………………………………………………….…...….pg 5

Model Error………..…………………………………………………….…...……...pg 6

**Answering Key Questions**…………………………………………………………….…...pg 6

Recommended Salary.………………………………………………………….…...pg 6

Conference Change..………………………………………………………….….....pg 6

Dropped Schools…..…………………………………………………………….…..pg 6

Graduation Rate Impact……………………………………………………….….....pg 7

Model Evaluation.....………………………………………………………….….....pg 8

Feature Importance..…………………………………………………………….…..pg 8

**Bonus Questions**……………………………………………………………………….…...pg 9

Geographic Visualization……………………………………………………….…...pg 9

Hierarchical Model..……………………………………..…………………….….....pg 11

**Conclusion & Future Directions..**…………………………………………………….…...pg 14

**Introduction**

Welcome to THE U-niversity located in Syracuse New York where it has been asked of the data science team to solve some business questions that will help the athletic director determine the correct pay for their new football coach. The program director has outlined a few key questions to address:

* What is the recommended salary for the Syracuse football coach?
* What would his salary be if we were still in the Big East? What if we went to the Big Ten?
* What schools did we drop from our data and why?
* What effect does graduation rate have on the projected salary?
* How good is our model?
* What is the single biggest impact on salary size?

**About the Data**

An initial dataset was provided containing 129 football programs with variables containing information on the coach name, conference, salary, total pay, bonus, bonus payout, assistant pay and contract buyout. This dataset was scrubbed and prepared for analysis creating an initially smaller data frame containing the outcome variable for this project ‘Total Pay’, conference, coach and school.

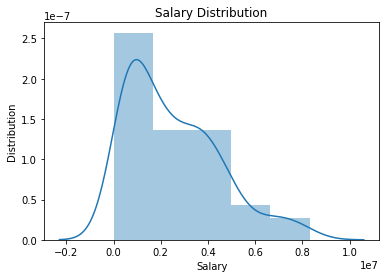
Taking the now narrow dataset, seven additional data sources were joined to bring theoretically relevant variables that may assist in the data modeling and subsequent salary recommendation. A synopsis of this can be found in Table 1 below. Data was joined based upon keys uniquely identifying data relationships. Data prep was needed to assist these merges including some manually found information on the coach’s record, school program revenue and other missing data items that emerged through the preparation process. From the original 129 schools 110 remained. Syracuse was dropped from the training and testing data as a hold out validation school for prediction.

|  |  |
| --- | --- |
| **Data** | **Variables** |
| [Bowl Wins](https://www.sports-reference.com/cfb/leaders/wins-post-coach-career.html) | **Bowl W** |
| [Coach Wins](https://www.sports-reference.com/cfb/leaders/wins-coach-career.html) | **Age; Career Yrs.; Career W; Career L; Career Win %; School Yrs.; School W; School L; School Win %** |
| [Program Revenue](https://sports.usatoday.com/ncaa/finances) | **Program Revenue** |
| [Stadium Capacity](https://www.collegegridirons.com/comparisons-by-capacity/) | **Capacity** |
| [Team Ratings](https://www.sports-reference.com/cfb/years/2017-team-offense.html) | **AP High Rank, AP Preseason; AP Rank End; Advanced Passing; Advanced Defense; Advanced Special Teams** |
| [Team Advanced Ratings](https://www.sports-reference.com/cfb/years/2018.html) | **Team W; Team L; Win %; Conference W; Conference L; Conference Win %; PPG Off; PPG Def;** [**SRS**](https://www.sports-reference.com/blog/2015/03/srs-calculation-details/) **(Scoring + SOS);** [**SOS**](https://bleacherreport.com/articles/152199-how-do-you-figure-out-strength-of-schedule) **(Strength of Schedule)** |
| [Graduation Ratings](https://web3.ncaa.org/aprsearch/gsrsearch) | **State; GSR (Graduation Success Rate); FGR (Federal Graduation Rate)** [**Details**](https://bleacherreport.com/articles/1731218-college-football-schools-with-the-best-and-worst-graduation-rates) |

*Table 1.* A list of data sources and variables brought into final data frame.

**Exploratory Data Analysis & Correlations**

Beyond descriptive statistics, data visualizations were leveraged to get a high-level sense of the data beginning with the outcome variable Total Pay which displayed a skew towards high earning coaches. This indicated a data transformation might be useful when modeling (see Figure 1 below).



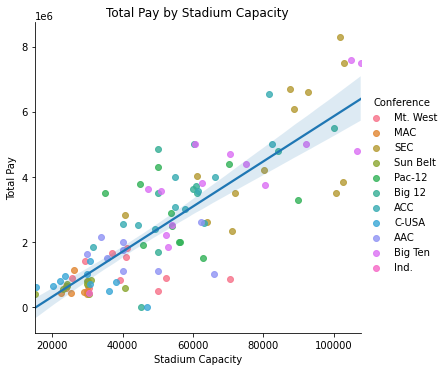
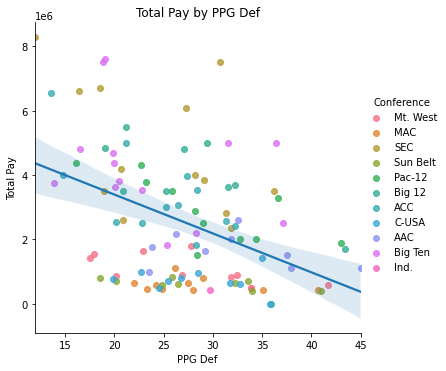
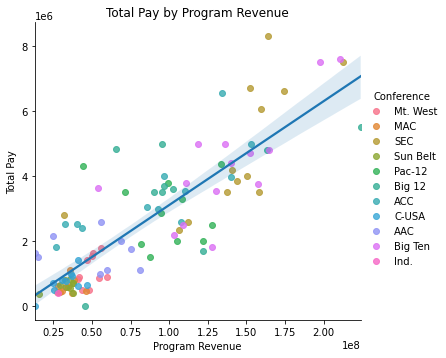
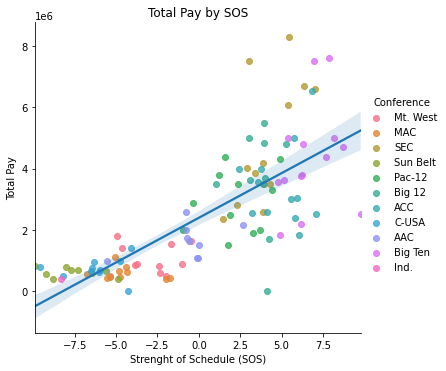
*Figure 1.* A histogram of total pay displays some skewedness towards high pay bins.

Next, all variables were tested for correlation strength with the Total Pay to determine, considering the absolute value of correlation, the variables that might offer the most indication of Total Pay. Table 2 below shows top correlations with Total Pay.

|  |  |
| --- | --- |
| **Variable** | **Correlation Strength** |
| **Program Revenue** | **0.84** |
| **Stadium Capacity** | **0.84** |
| **SOS** | **0.76** |
| **SRS** | **0.72** |
| **AP High** | **-0.69** |
| **AP Preseason** | **-0.67** |

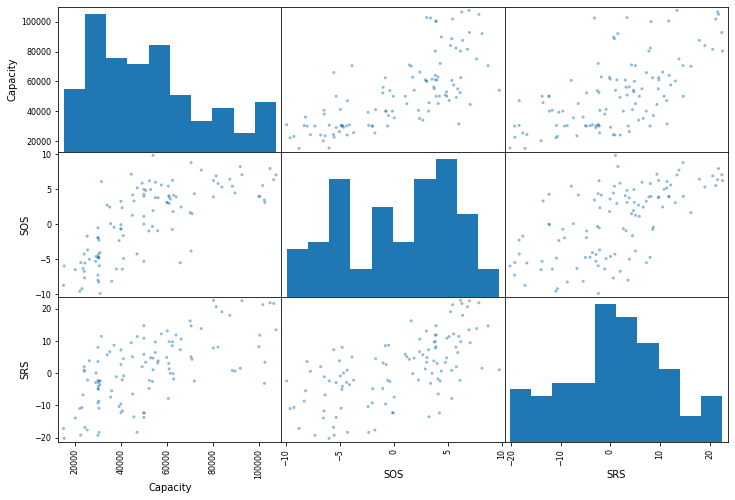
*Table 2.* Variables most correlated with Total Pay in descending order of absolute strength.

Of the top correlations, relationships between Total Pay are displayed in the scatterplots below to visually depict the strength of the linear relationship. Values are colored by conference which can be highly influential on coach pay.



*Figure 2.* Scatterplots demonstrating correlations with Total Pay amongst influential variables.

Intervariable correlations suggest that multicollinearity may be a problem and a PCA may be useful in modeling for future predictions. Below, a scatter matrix displays relationships and distribution of key variables from the correlations.



*Figure 3.* Scatter matrix of top predictors based upon inferences from the correlations.

**Modeling**

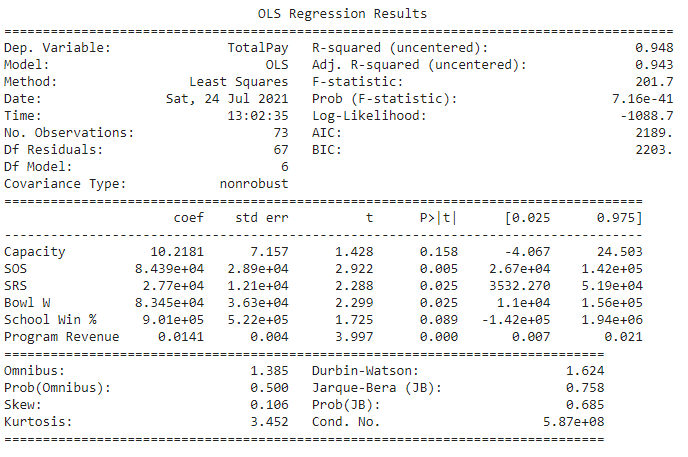
Conference was ‘dummified’ into binary variables results in a final data frame containing 109 rows and 64 columns to be used for modeling. The data, excluding Syracuse, was then split into training and testing sets so that 33% of the data would be used to test.

Several models were tested with including a model that contained all columns, a pseudo stepwise regression model removing noisy predictors and a model containing just the top correlation variables to name a few. Following Occam’s Razor and in line with the actual model performance, the simple model containing just the top correlation variables and those found to provide significant incremental predictive power was chosen. This model was found to be the best generalizable model on the test set.

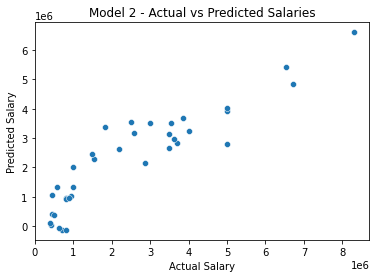
**Chosen model:**

Total Pay ~ Stadium Capacity + SOS + SRS + Bowl W + School Win % + Program Revenue

This model was found to have an adjusted of 0.94 and all independent variables neared a significant threshold.



When used to predict values in the test set, it was found that this model overfit on the training data, but was still about to achieve an adjusted R-squared of 0.81 with somewhat generous RMSE: $881,097 and MAE: $710,496. This indicates a fairly large confidence interval for which the predicted salary will be based off of. When negotiating with the chosen coach, it would be advisable to consider a range rather than a specific number.



*Figure 4.* Scatterplot of predicted model results versus the actual shows a fairly linear pattern.

**Answering the Key Questions**

* What is the recommended salary for the Syracuse football coach?

The aforementioned model that displayed both generalizability and simplicity was chosen to predict Syracuse’s new coach salary. This model estimates that the Syracuse coach salary should be $2,190,066. Using the validation data, the actual coach salary is $2,401,206 indicating a short offer of about $200K. When negotiating with this coach, it would be expected that the salary should fall within the range of $1,758,548 to $2,621,584 using the model confidence interval estimations.

|  |  |  |
| --- | --- | --- |
| **Salary Min** | **Salary Target** | **Salary Max** |
| $1,758,548 | $2,190,066 | $2,621,584 |

* What would his salary be if we were still in the Big East? What if we went to the Big Ten?

If Syracuse were in the Big 10 instead of the ACC, the salary would be: $2,167,572 or about $22,494 more. The was determined by applying the hierarchical model (including the wide dummy conference variables) and changing the conference code for Syracuse from ACC to Big Ten. The model was not trained on Big East data, so a different prediction was not made.

* What schools did we drop from our data and why?

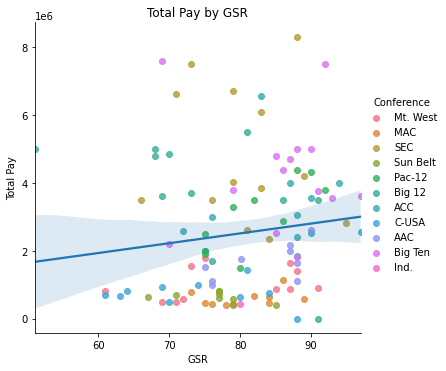
A total of 19 schools were dropped from the analysis do to sparsity of the variables of interest for these schools. Missing data restricted their inclusion and while more data was imputed than desired, it still left a sizeable sample of data to train and test on. There are no known reasons why the imputed data may differ from the overall population of college football programs.

|  |
| --- |
| **School** |
| Alabama at Birmingham |
| Army |
| Bowling Green |
| Brigham Young |
| Central Florida |
| Liberty |
| Louisiana-Lafayette |
| Massachusetts |
| University of Miami (Florida) |
| Miami (Ohio) |
| Middle Tennessee |
| Mississippi |
| Nevada-Las Vegas |
| Notre Dame |
| Pittsburgh |
| Southern California |
| Southern Methodist |
| Texas-El Paso |
| University of Texas at San Antonio |

*Table 3.* Schools removed from analysis due to missing data.

* What effect does graduation rate have on the projected salary?

Graduation rate, although relevant for the overall program success as judged by the NCAA (in theory) and the university had little to do with total coach pay. GSR and FGR had weak and unmeaningful relationships and offer little predictive power. Figure 5 below demonstrates the weak relationship between GSR and total pay.



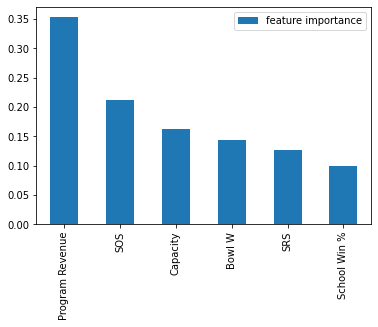
*Figure 5*. A weak relationship between GSR and pay demonstrates insignificance for this factor.

* How good is our model?

Overall, this model accounted for a significant amount of variance in the coach pay but had room for error as indicated by the RMSE: $881,097 and MAE: $710,496. The model was fairly close in estimating the actual Syracuse salary, but should not be used a standalone determinant at this point without improvements to narrow the distance from the model line of best fit to the residuals. On the surface, the model performed well in explaining the data, but in practical application the error makes this hard to utilize without other domain considerations. This model should be improved upon and used as a data point in decision-making rather than a silver bullet solution.

* What is the single biggest impact on salary size?

The existing data was then normalized to compare the relative impact of each variable on total pay. This was done to create equivalent units of measurement. Figure 6 shows that program revenue is the most important variable in the model which makes sense theoretically and empirically due to the high correlations found during the data exploration. It appears that schools that have the money and resources to pay top tier coaches also spend the most money in a ‘pay to play’ scenario. This creates a dilemma of causality, however, because empirically this study assumes that revenue predicts salary when in fact the relationship is likely much more complex than that.

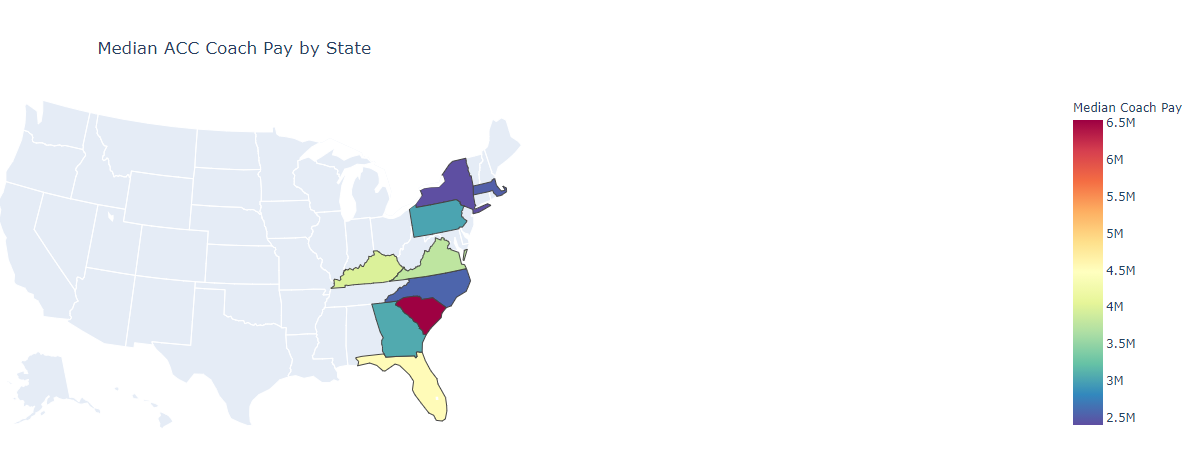


*Figure 6*. Normalized data frame helps demonstrate relative feature importance.

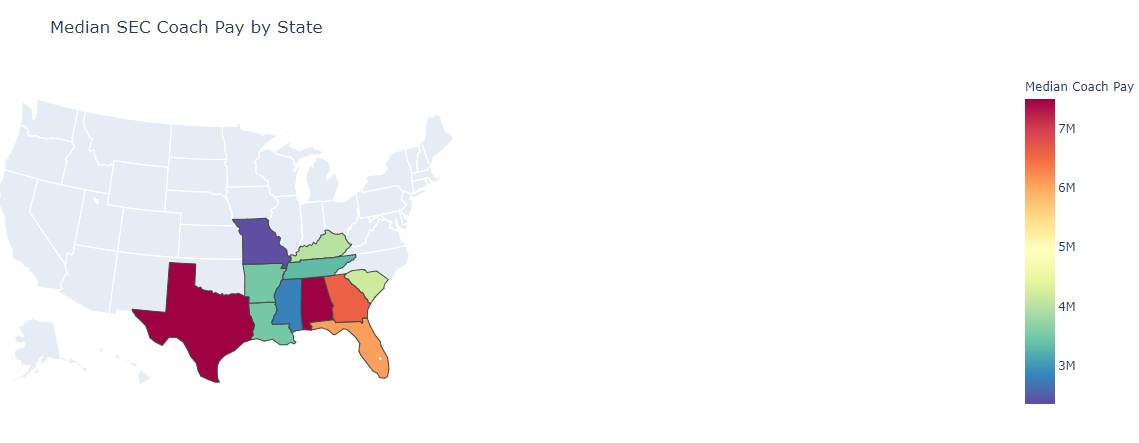
**Bonus Questions**

* Develop a geographic visualization that in your view best depicts the conferences ‘median salary

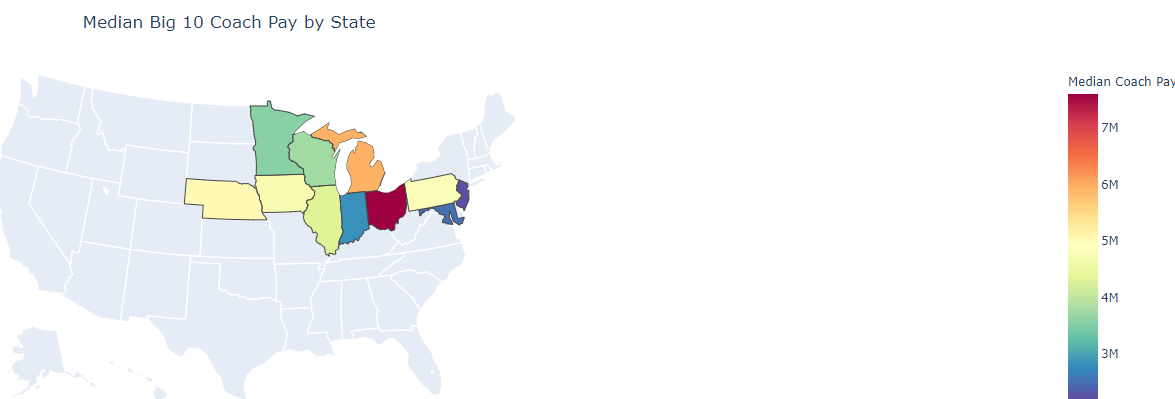
Graphic visuals were created to demonstrate the median pay disparity by conference using the ACC, SEC and Big Ten as examples in Figures 7-10.



*Figure 7*. A geographical representation of the ACC and median coach pay.

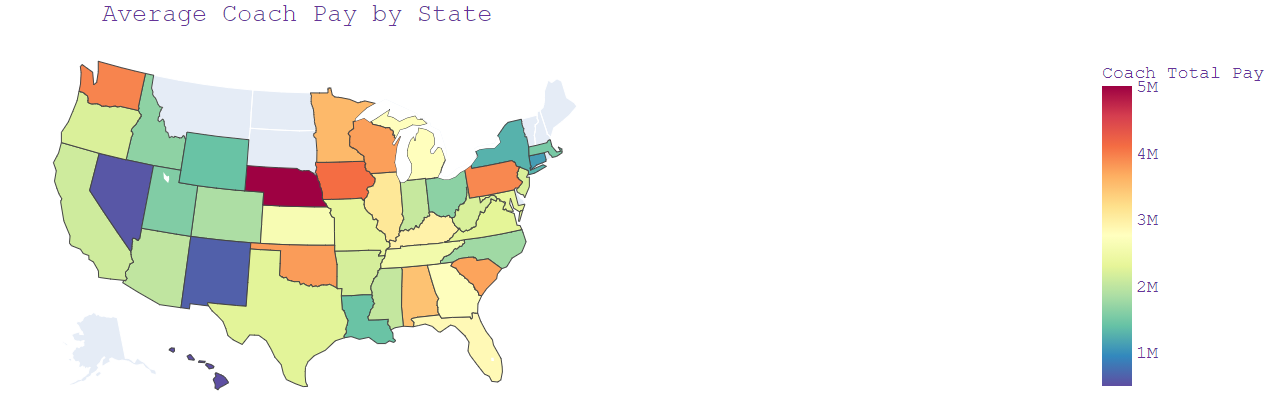


*Figure 8*. A geographical representation of the SEC and median coach pay.

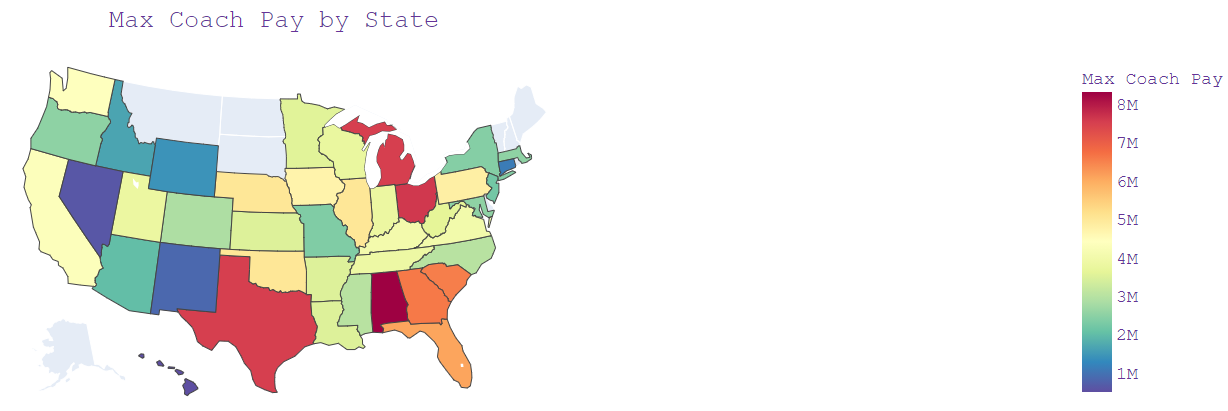


*Figure 9*. A geographical representation of the Big 10 and median coach pay.

In figures 10 and 11, conference was not used to subset into specific geographic areas. Instead, the entirety of the US was compared first by the average pay for each coach by state and then to max pay to see just how large Nick Saban lives relative to his peers.



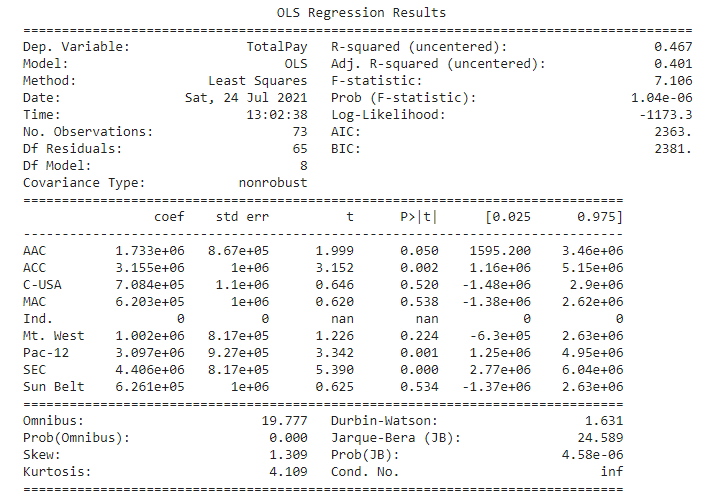
*Figure 10.* A geographical representation of the average coach pay by state.

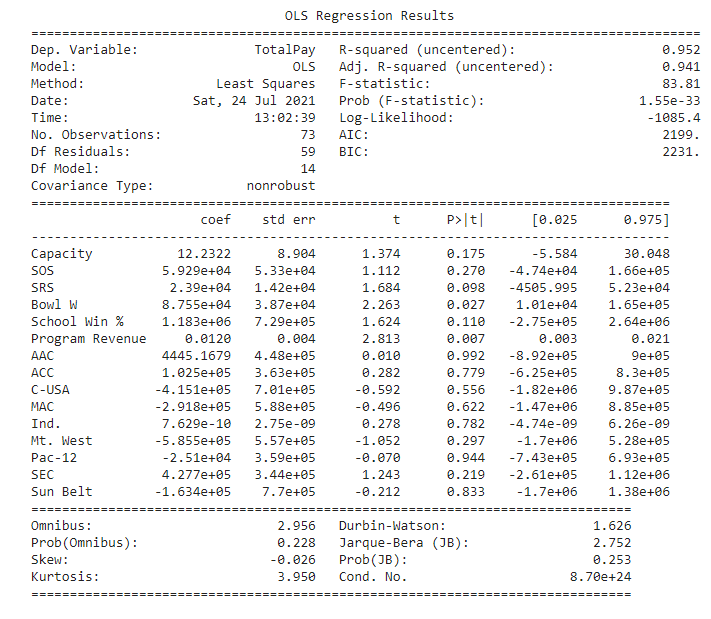


*Figure 11.* A geographical representation of the max coach pay by state.

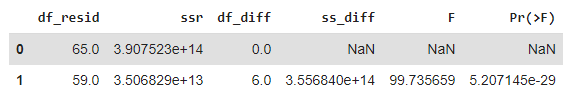
* Fit a hierarchical model based on conference

A hierarchical model was tested using conference wins and the best model after training and testing the data. Below, the conference only model achieved an R-squared value of 0.40.

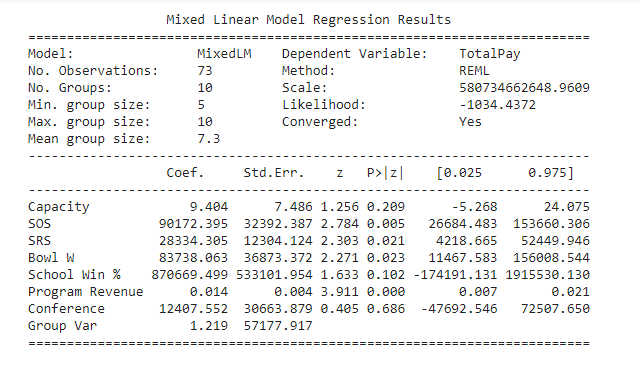


However, after adding in the ‘best model variables, the r-squared was raised to 0.94.

An analysis of variance calculation was run to determine whether or not the gain in variance explained was significant. As displayed below, the non-conference model variables added significant value to the model.



Finally, a mixed linear model saw no improvements over the best model mentioned before.



**Conclusion & Future Directions**

Several models were developed to predict NCAA coach’s salary using eight different sources of information. A lean model was ultimately used with a train R-squared of 0.94 and a test of 0.81. In predicting Syracuse’s coach salary, the model was conversative about $200K and should be used as a factor in negotiating and not a perspective determinant. I would be best to use the model range considering the average error of $700K when sourcing for the role meaning the expected range for the next coach will be within the range of $1.75M-$2.62M.

Many enhancements could be made including resolving some imputation issues that occurred and providing the full rank of AP listings to give greater meaning to that variable. Greater care should be given to exploring the PCA model and more algorithms should be tested including neural networks, boosted and ensemble models. It was suggested that donations might be a strong variable to include and it would be advisable to test that hypothesis as well. This data also may have contained outliers and was not tested to see if Nick Saban and other coaches were outliers as well as testing the assumption that a linear model should be used to model on this data. The np.log-transformation and normalizing the data set did not seem to add value, but further transformation methodologies should be tested to see if gains can be made. This could also lead to a more specified training set that would have less overall generalizability but would like experience performance gains and reduced residual error.