Barbara Schweitzer

Lab 1

To begin the analysis of the Coaches data set, it was first necessary to gather more valuable data. This other data included Graduation Success Rate and the Federal Graduation Rate, the teams win and loss numbers, and the stadium capacity information. To combine the different types of data, it was most effective to load all of the data into Excel and match the newly gathered values with the existing Coaches data set. By matching the school names in the different data sets, a final data set was created with the values for each school present, where available. Based on the final data set, it seemed most valuable to use the Total Pay variable as the target for the model evaluation. It was then necessary to remove four schools, due to their lack of pay information. These schools were Southern Methodist University, Baylor University, Rice University, and Brigham Young University. Finally, to end the data preparation, dummy variables for the conferences were added, to aid in model creation later in the data analysis. For each conference, a variable was created with a 1 if the school was a member of this conference, and a 0 if they were not a member.

Next, the data analysis in Python began with loading the necessary packages and importing the data. Then, it was necessary to change the data type of the Total Pay variable from a Pandas Series to a float data type. This would allow the values to be treated as numbers for the subsequent data analysis. A variable was also created for the Total Pay in thousands, which would allow the values to be visualized in a more understandable way in the graphics.

The first step in the analysis was to look at the summary statistics for the Total Pay variable. The mean was $2,430,077.35 and the median was $1,900,008.00. The minimum pay value was $390,000.00, with the maximum pay being $8,307,000.00. The range of pay was then $7,917,000.00, making the spread of data values very large. To view the distribution of pay more accurately, below is a boxplot of the coachs’ total pay, along with a histogram of the data. There is one clear outlier in the data, which is the coach at the University of Alabama, Jim Harbaugh.

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

For a clearer understanding of the pay distribution between the conferences, below are boxplots of the total pay by each conference, ordered by decreasing median values.

Chart, box and whisker chart

Description automatically generated

Clearly, the Big 10 Conference has the highest median pay value, and the Mid-American Conference has the lowest median pay. The Southeastern Conference, though it does not have the highest median value, the coaches in this group have the largest range of pay, and also the highest wages in the league.

To again visualize the distribution of wages within the leagues, a swarm plot of total pay by conference, again in decreasing median order, is below. The first five conferences, Big 10, SEC, Big 12, ACC, and Pac-12 are part of the Power Five Conference. This conference consists of the “best of the best” college football teams. The coaches, being part of this conference, have clearly benefitted from their coaching skills and the credibility that comes with this conference, as they are some of the highest earners among all 11 conferences.

Chart, scatter chart

Description automatically generated

The next visuals show the relationships between a coach’s total pay and their teams wins, losses, graduation rates, and stadium sizes. The first graphic is a Swarm Plot of Total Pay and the Graduation Success Rate at the specific colleges. There is no clear relationship present between these two variables, which will become more apparent in later analysis.

Chart, scatter chart

Description automatically generated

Similarly for the Federal Graduation Rate, there is no relationship between the variables and the plotted points are randomly scattered throughout the graphic.

Chart, scatter chart

Description automatically generated

With the Stadium Capacity, there is a very clear positive relationship between the variables. As the stadium capacity increases, the total pay of the coach generally increases as well.

Chart, scatter chart

Description automatically generated

Finally, the relationship between the wins and losses of the teams and the coach’s total pay are evaluated. As expected, as the number of wins rises, the pay of the coach increases in most cases. Similarly, as the number of losses increases, the pay of the coach decreases in most cases. This created plots that are nearly mirror images, with slight discrepancies caused by the differing number of games played by each team in one season.

Chart, scatter chart

Description automatically generated

The final graphic is a correlation matrix between the conference, total pay, GSR, FGR, stadium capacity, and win/loss data. Based on this graphic, there is a slight negative correlation between the different conferences. There are also both positive and negative correlations between each of the conferences and the total pay. The largest correlation is present between the win and loss data, which is expected.

Chart

Description automatically generated

To begin the analysis of the data, training and testing data sets were first created using a 2/3:1/3 split of the data. This resulted in 90 data entries for the training data set and 35 data entries for the testing data set. Next, simple linear regression models were created, using the total pay variable and one of the possible testing variables, including GSR, FGR, wins, stadium capacity, and conference. The models created with GSR and FGR were the least helpful, as they resulted in adjusted R^2 values of 0.032 and -0.012, meaning the GSR and FGR explained 3.2% and -1.2% of the variability in the total pay for coaches. The win variable was slightly more successful, with an adjusted R^2 value of 0.123. The last two models, with stadium capacity and conference were much more helpful. These resulted in adjusted R^2 values of 0.707 and 0.691, respectively. Thus, the stadium capacity and conference had the greatest effect on the total pay for a coach, more specifically the stadium size.

Next, five different models were run using different combinations of these variables to find the best model for the data. First, a model with all of the variables, conference, GSR, FGR, stadium capacity, and wins was run. This model resulted in an adjusted R^2 value of 0.785, thus the variables explained 78.5% of the variability in the coach’s pay. The F-statistic was also significant for this model. The next model run was one with the wins and stadium capacity. This resulted in an adjusted R^2 value of 0.704, which was lower than the last model. Next, a model using the conferences and the wins was created. This model had an adjusted R^2 value of 0.751, which was still lower than the first model. Then, a model with the stadium capacity and the conferences, the most influential variables, was created. This model had an adjusted R^2 value of 0.776, which was still slightly lower than the first model. The final model was created using the stadium capacity, conference, and win variables. This model had an adjusted R^2 of 0.79. Thus, these three variables explained 79% of the variability in the total pay, and it also had a significant F-statistic. This model was the best of the previous models, and thus will be utilized as the model to predict the Syracuse Coach’s pay.

To predict the salary for the Syracuse coach, the final model was used. Taking the coefficients from the model and subbing in the appropriate values, the resulting equation was

68,600 + 30.0464\*(49,250) + (-585,100)\*(0) + 1,342,000\*(0) +

1,310,000\*(0) + (-982,000)\*(0) + (-741,400)\*(0) + (714,300)\*(1) + (-970,200)\*(0) + (-969,200)\*(0) +827,100\*(0) + 1,417,000\*(0) + (-1,094,000)\*(0) + 85,960\*(5).

The 0 values present in the equation are from the dummy variables created for the conferences, thus the equation has 0’s for the conferences that Syracuse is not part of, and a 1 for the ACC. The 49,250 value is the stadium capacity of the Carrier Dome, and the 5 is the number of wins by Syracuse in 2019. Thus, from this equation, the Syracuse coach should have a salary of $2,892,485.20.

The Big East conference disbanded in 2013, and most of its members joined the ACC or AAC. Thus, to project the salary of the Syracuse coach if they had remained in the Big East, the same model will be used, however the equation will have a 1 for both the AAC and ACC coefficients. The resulting equation is

268,600 + 30.0464\*(49,250) + (-585,100)\*(1) + 1,342,000\*(0) +

1,310,000\*(0) + (-982,000)\*(0) + (-741,400)\*(0) + (714,300)\*(1) + (-970,200)\*(0) + (-969,200)\*(0) +827,100\*(0) + 1,417,000\*(0) + (-1,094,000)\*(0) + 85,960\*(5).

The resulting salary projection from this model is $2,307,385.20, which is $585,100 less than if the coach were in their current conference.

Lastly, if the coach were to have joined the Big 10 Conference, the projected salary, using the same equation, would be $3,520,185.20, which is the largest salary projection thus far. The results of the three predictions are shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| School | Current Salary | Predicted in ACC | Predicted in Big East | Predicted in Big 10 |
| Syracuse | $2,401,206.00 | $2,892,485.20 | $2,307,385.20 | $3,520,185.20 |

**Data Obtained From:**

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

<https://www.teamrankings.com/ncf/trends/win_trends/?range=yearly_2019>

<https://web3.ncaa.org/aprsearch/gsrsearch>

**COLAB Link:**

<https://colab.research.google.com/drive/1uk7HIcJxiYh61EKahVI6JTJcCCmndDXj?authuser=1#scrollTo=q_9dgGFiIx0M>