Laboratory Exercise #1

Introduction

As introduced in the second case study, the purpose of this lab is to evaluate the factors that affect an NCAA Division I Football head coach and predict a salary range for the Syracuse head coach. This evidenced-based approach to setting the coach's salary expectations stems from the idea that athletic departments in large universities have deviated from their mission of graduating students.

The three aspects of a coach I will attempt to cover is: team performance (wins/points), recruitment and academic performance.

Methodology

The original data set for this lab was given in .csv format, it contains data about head coach pay and bonus structure in 2018, downloaded from an USA Today report. This data set was enhanced with additional information that targets the three main aspects of a head coach's job (in my opinion) which are: to increase/maintain attendance, to win games, and to graduate his student-athletes.

The Data

Seven additional data sets were scrapped and joined together. They include:

- AP_Top25: The associated press's top 25 ranked teams for the season, scrapped from ESPN. It contains:
  - Season-end ranks for the top 25 ranks
  - # of votes received
- NCAA Div. 1 Football stats: several statistics related to the football team's performance in the current season (2018) and the previous season (2017). Mainly wins, losses and points scored also scrapped from ESPN.
  - Conf_Points_For
  - Conf_Points_Against
  - Total_Points_For
  - Total_Points_Against
  - Conf_Wins
  - Conf_Losses
  - Home_Wins
  - Home_Losses
  - Away_Wins
  - Away_Losses
  - Wins_2018
  - Losses_2018

- Coach data 2018: this dataset contains statistics about a head coach's career data and when the coach had his first season with the team scrapped from Wikipedia. It attempts to inform the model about the tenure of the coach.
    - First_Season 🡐 First season with the team (year)
    - W 🡐 Current season wins
    - L 🡐 Current season losses
    - W% 🡐 Current winning percentage
    - Career W
    - Career L
    - Career W%
- Stadium size: information about the school's stadium size and when it was opened, scrapped from collegegridirons.com
    - Capacity
    - Opened
- Recruitment class: information about the quality of the recruiting class scraped from 247sports.
  Note: since this data is supposed to relate recruitment data to the performance of team in 2018, recruitment rankings are added for 2019.
    - RecRank_2019
    - 5-stars_2019
    - 4-stars_2019
    - 3-stars_2019
    - Total_commits_2019
    - Avg_RecScore_2019 🡐 247sports proprietary avg. recruitment class score
- Graduation Success Rate: data about graduation rates for Division 1 schools, 2008-2011 cohorts combined.
    - Fed_rate
    - GSR

Data exploration and munging

A combination of fuzzy matching and manual inspection was used to create a 'matching data frame' which contained all the different names from the different data sets indexed to the coaches9 original data set. These were appended together using left joins.

Each variable set of data was evaluated at in terms of its correlation with the dependent variable (Total Pay) in order to select the features that showed some relationship with it.

Wins and Points Scored

Wins and points scored showed interesting relationships between the coaches total Pay and among themselves. There is a bit of a linear relationship between Total Pay and all the variables in this data set. However since all of these show strong relationships with each other, I picked the two that showed the highest R-squared.

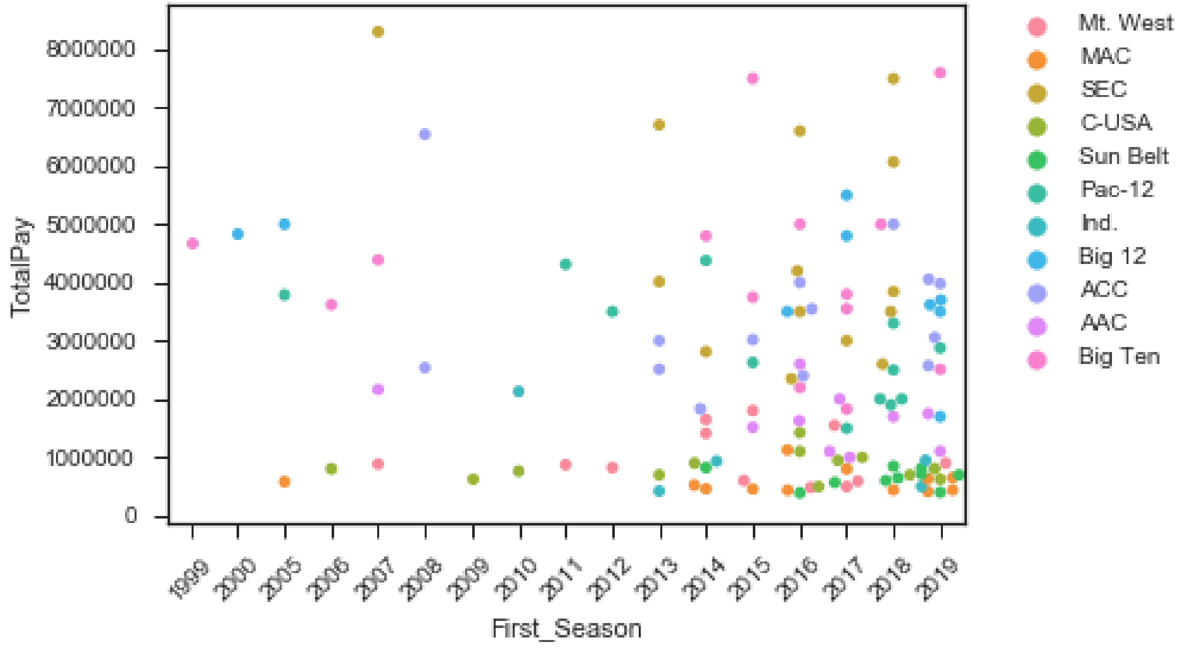
For 2018: Home wins and Away losses show the strongest correlation

1.	Home_Wins_2018	0.433400
2.	Wins_2018	0.382158
3.	Total_Points_For_2018	0.366312
4.	Conf_Points_For_2018	0.324572
5.	Conf_Wins_2018	0.285672
6.	Away_Wins_2018	0.075977
7.	Conf_Points_Against_2018	-0.041440
8.	Conf_Losses_2018	-0.194567
9.	Total_Points_Against_2018	-0.256755
10.	Home_Losses_2018	-0.261566
11.	Losses_2018	-0.356480
12.	Away_Losses_2018	-0.439801

For 2017, the relationship is very similar. Home wins and away losses were selected for the final dataset.

Coach tenure and career stats:

Through exploration of this dataset it was found that the web scrapper for the stats breaks mid-dataset for some reason and half of the values are incorrect. Therefore only the names and start year could be used. However, Total Pay does show a relationship with this one variable:



It seems to capture the idea that a tenured coach should get paid more.

Recruiting Stats

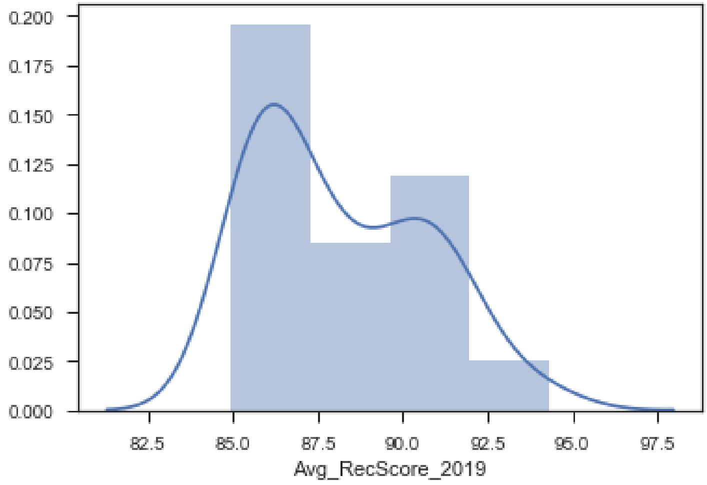
The correlation coefficient (r-squared) for the recruitment with Total Pay is the highest out of the scrapped variables on average. Most show upwards of .6 with the 247 Sports data. Especially the variable that talks about the recruiting class ranking:

1. Avg_RecScore_2019	0.690903
2. 5-stars_2019	0.645297
3. 4-stars_2019	0.629687
4. Total_commits_2019	0.299239
5. 3-stars_2019	-0.564128
6. RecRank_2019	-0.654865

From this data set I selected the Rec_Score, the scores lend themselves to discretization better than the ranks.

This dataset, however, contains a lot of missing values (about 60% of schools in the coaches' dataset are not present here) so the distribution of the variables was studied to understand how to replace said values:

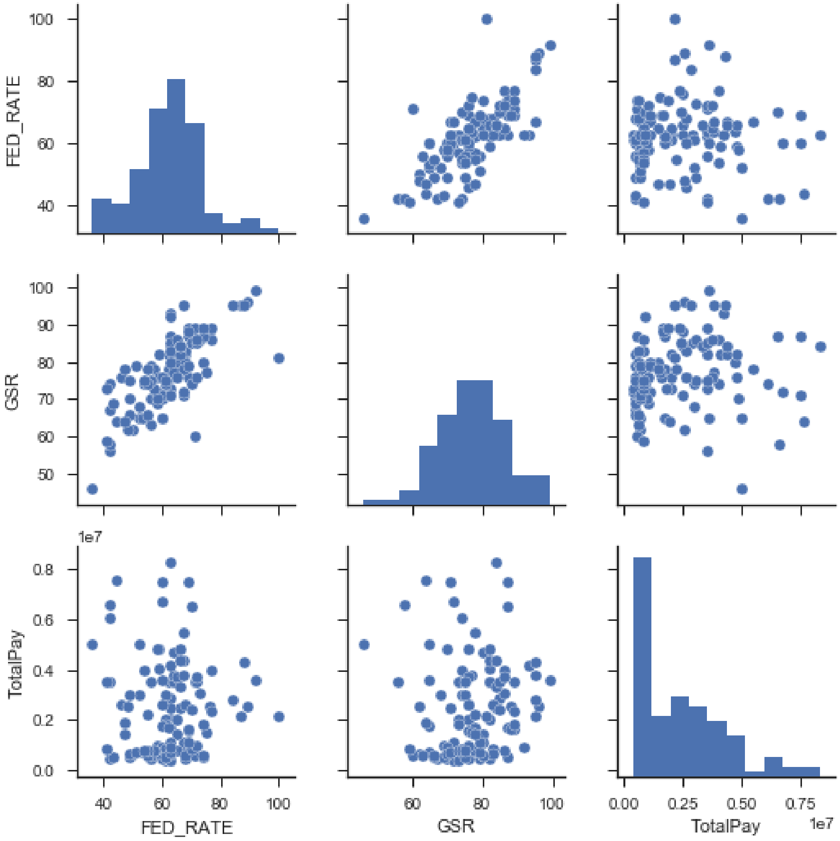
The rec_score would lend itself well to replacement with the median since most of the values fall around there… but it would not capture the idea that all values that are not here are worse.



The null counts for 4 and 5 star recruits will be replaced with 0's basically saying these schools do not have them. It should help the model by showing how special it is to have these type of recruits.

Graduation Rates (GSR)

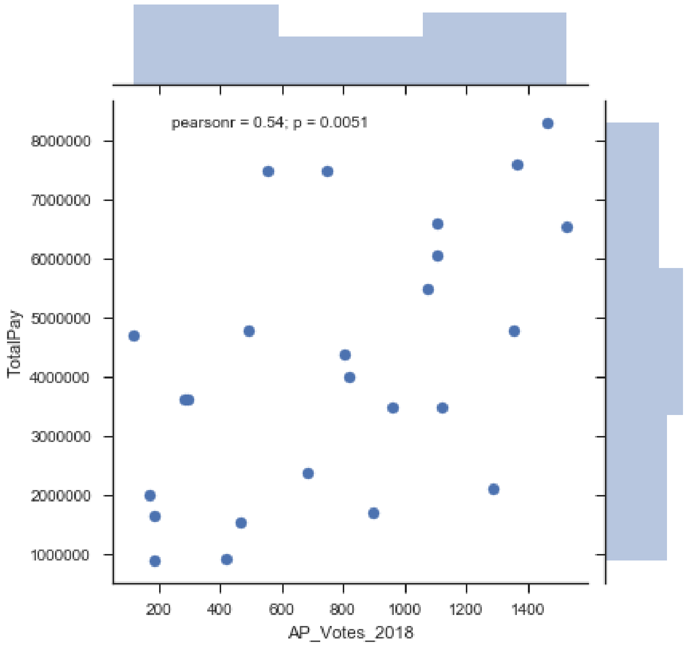
This variable also has decent correlation with total pay, it does seem like it is much different from the federal graduation rate and, due to their mutual linear relationship, only one should be chosen:



The distribution of the GSR variable should allow its null values to be replaced with the mean/median since it looks like a normal distribution.

AP Top25

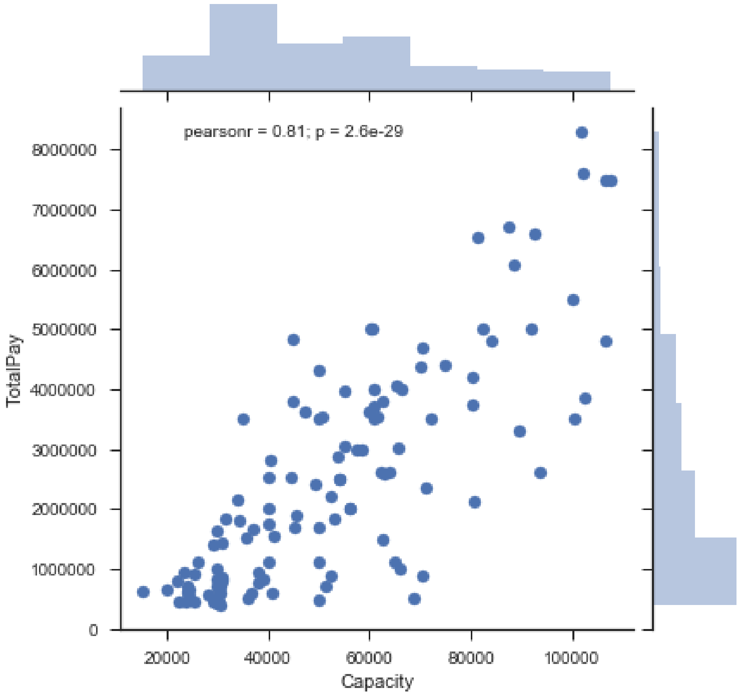
The AP Top 25 data, similar to the recruiting class ranks, shows a loose positive relationship with Total Pay.



But this data will be hard to model through regression due to the very large presence of nulls (only 25 teams). Most likely it would benefit from discretization for algorithmic modeling vs the statistical modeling.

Stadium Size

Stadium size showed very high correlation with Total pay but showed a high chance for error within the regression model.

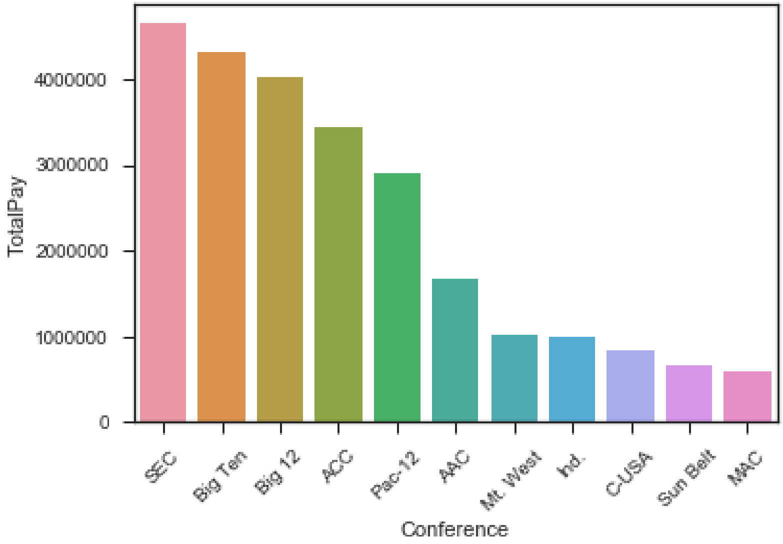


Modeling

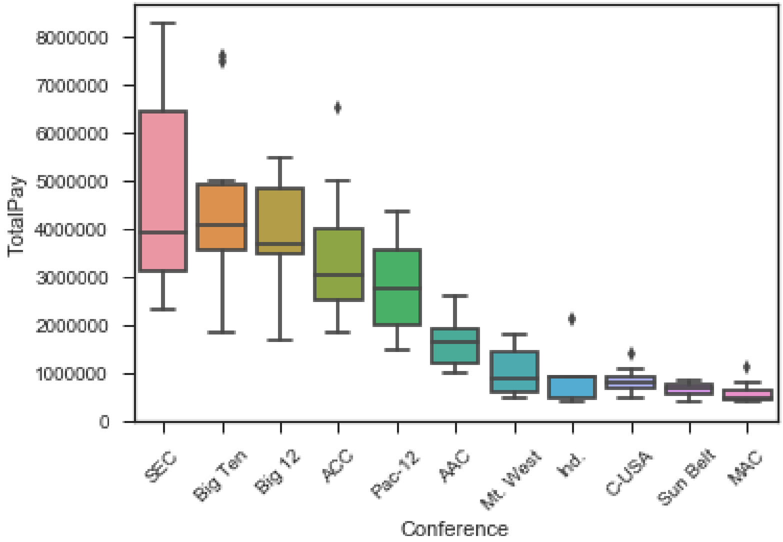
OLS regression models were built using the statsmodel package. These models suffered from severe overfitting and did not perform well on training data. Relatively high (upwards of .6) adjusted r-square values on the full data set drop under .2 when the data gets split for training/testing.

Results and Findings

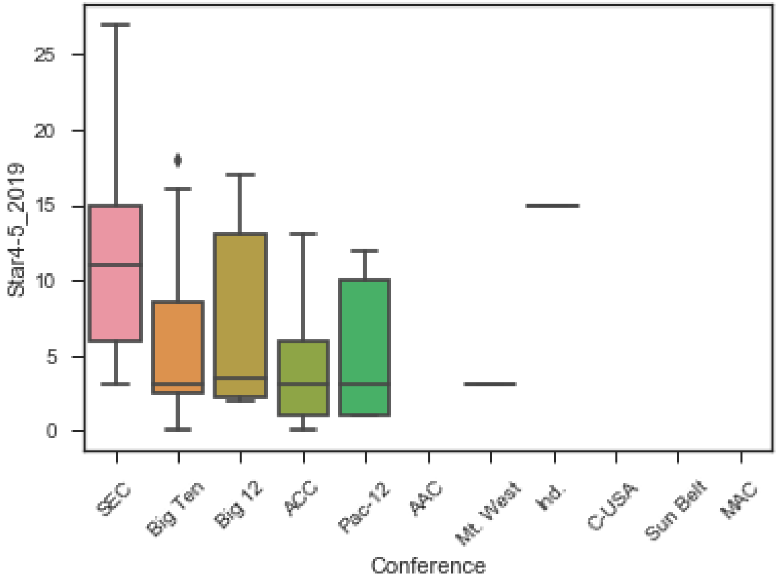
In terms of salary, the power-5 conferences have a clear lead on the rest of the conferences with, as expected, the SEC at the top:



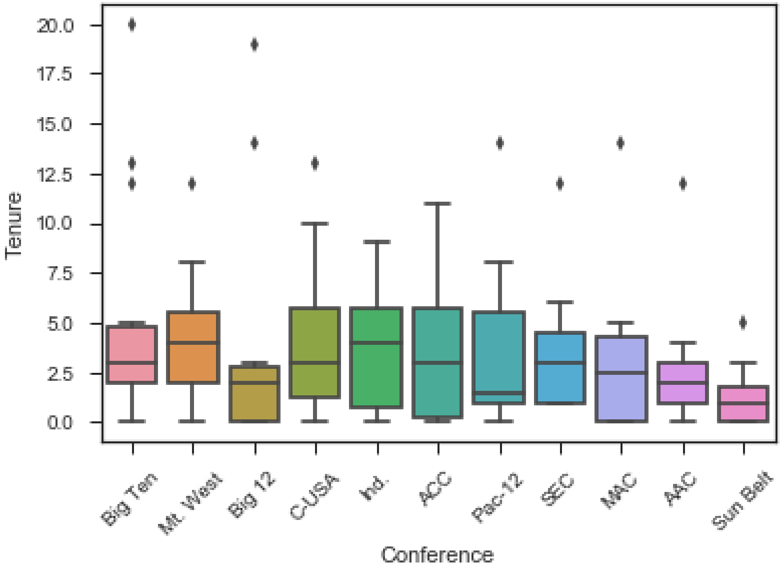
There are outliers in some of the conferences, especially the Big Ten (OSU and Michigan) and the ACC (Clemson), where some of the coaches make significantly more than their peers:



Unsurprisingly, the SEC also commands the highest number of 4 and 5 star recruits, also notice how outside of the power 5 there are almost no starred recruits:



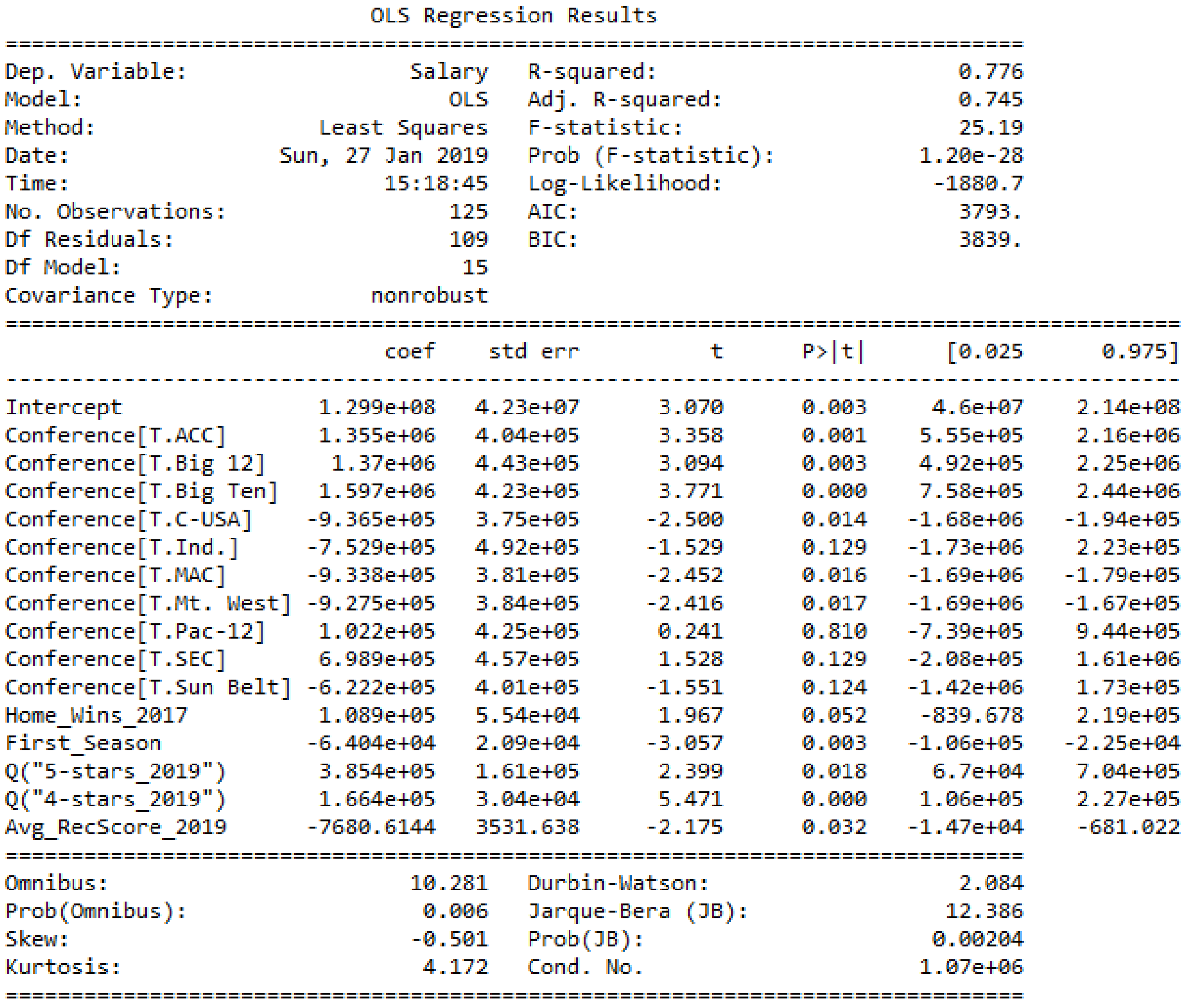
One surprising stat when looking at the data was that these conferences was how the highest paid conferences don't necessarily have the most tenured coaches:



The OLS model

Although not fit for prediction, mostly due to high outliers and missing data, the model explains a lot about which variables are important to the variable “Salary” (TotalPay – Bonus). This model did not respond well to discretization so all the variables were passed in as continuous except for conference.

The best model contained mostly aspects of recruiting with some for tenure and performance:
Salary ~ Conference + Home_Wins_2017 + First_Season + Q("5-stars_2019") + Q("4-stars_2019") + Avg_RecScore_2019'



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.776
Model:                            OLS   Adj. R-squared:                  0.745
Method:                 Least Squares   F-statistic:                     25.19
Date:                Sun, 27 Jan 2019   Prob (F-statistic):           1.20e-28
Time:                        15:18:45   Log-Likelihood:                -1880.7
No. Observations:                 125   AIC:                             3793.
Df Residuals:                     109   BIC:                             3839.
Df Model:                          15                                         
Covariance Type:            nonrobust                                         
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept               1.299e+08   4.23e+07      3.070      0.003     4.6e+07    2.14e+08
Conference[T.ACC]       1.355e+06   4.04e+05      3.358      0.001    5.55e+05    2.16e+06
Conference[T.Big 12]     1.37e+06   4.43e+05      3.094      0.003    4.92e+05    2.25e+06
Conference[T.Big Ten]   1.597e+06   4.23e+05      3.771      0.000    7.58e+05    2.44e+06
Conference[T.C-USA]    -9.365e+05   3.75e+05     -2.500      0.014   -1.68e+06   -1.94e+05
Conference[T.Ind.]     -7.529e+05   4.92e+05     -1.529      0.129   -1.73e+06    2.23e+05
Conference[T.MAC]      -9.338e+05   3.81e+05     -2.452      0.016   -1.69e+06   -1.79e+05
Conference[T.Mt. West] -9.275e+05   3.84e+05     -2.416      0.017   -1.69e+06   -1.67e+05
Conference[T.Pac-12]    1.022e+05   4.25e+05      0.241      0.810   -7.39e+05    9.44e+05
Conference[T.SEC]       6.989e+05   4.57e+05      1.528      0.129   -2.08e+05    1.61e+06
Conference[T.Sun Belt] -6.222e+05   4.01e+05     -1.551      0.124   -1.42e+06    1.73e+05
Home_Wins_2017          1.089e+05   5.54e+04      1.967      0.052    -839.678    2.19e+05
First_Season           -6.404e+04   2.09e+04     -3.057      0.003   -1.06e+05   -2.25e+04
Q("5-stars_2019")       3.854e+05   1.61e+05      2.399      0.018     6.7e+04    7.04e+05
Q("4-stars_2019")       1.664e+05   3.04e+04      5.471      0.000    1.06e+05    2.27e+05
Avg_RecScore_2019      -7680.6144   3531.638     -2.175      0.032   -1.47e+04    -681.022
==============================================================================
Omnibus:                       10.281   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               12.386
Skew:                          -0.501   Prob(JB):                      0.00204
Kurtosis:                       4.172   Cond. No.                     1.07e+06
==============================================================================
```

Through this model we can see how coaches in the big conferences have already a large lead on their peers and how wining home games and recruiting good athletes helps their case.

Discussion and Assignment Questions

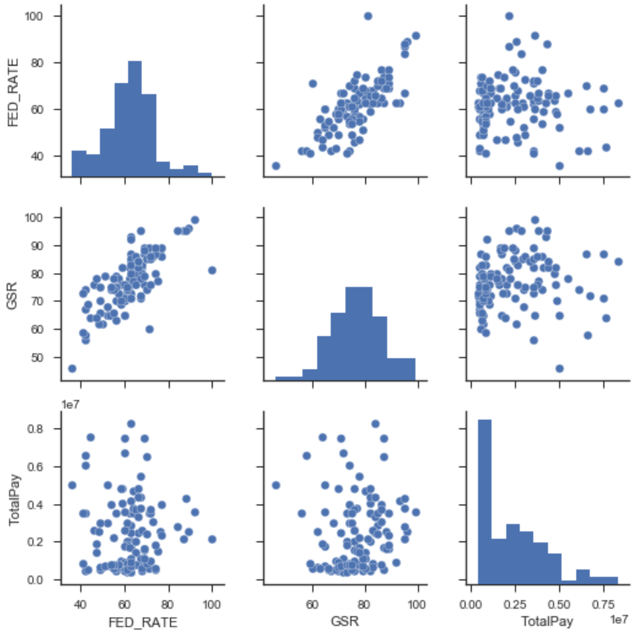
- The recommended salary for the Syracuse coach would be $2,585,960.00. About a 5% raise from the current salary, I think this could be interpreted as the coach deserving a raise because of the recent winning season the team put forth.

- If he went to the Big Ten, he should get a large bump of 18% to $ $2,827,960.00.

- However, if he was still in the Big East (now part of the ACC), he would get significantly less than he's making in the ACC. His salary should be $1,230,960.00.

- Baylor, BYU, Rice and SMU were dropped from the dataset because they did not include salary information. The model cannot learn anything if the category doesn't contain the dependent variable.

- According to the current model, graduation does not significantly affect salary. The two variables do not show any strong, noticeable relationship:



- The model is really good for interpretation (high r-square, p-values ok at the 95% confidence level, the confidence intervals stay on one side of 0) but it does not have good predictive power since it only shows these positive statistics when it gets the entire dataset.

- The biggest impact is the conference. Working in the right conference can net an increase or decrease of almost a million dollars. The number of home wins and the number of 5 star recruits also has a big impact according to my model.