Jacob Arluck

Football and Regression: The Predictors of Coach Pay

Introduction

College football is big business. Though the athletes themselves earn no compensation beyond their scholarships, the best-paid coaches draw base salaries that nearly reach into the eight digits. In most states, the highest paid public employee is the head coach of the leading state university. Nonetheless, not all coaches, even in the top-flight Football Bowl Subdivision of the NCAA, draw such exorbitant salaries, with the lowest paid earning a mere $390,000. This inquiry seeks to clarify what factors best predict the salaries paid to coaches, with a particular eye towards determining how much Syracuse University should pay for a hypothetical new coach.

Locating Data and Preprocessing

Four sources of data were used in preparing this report:

- A data set of coach salaries from 2018, provided with the assignment. 3 schools (out of 130) were missing the response variable (School Pay). Since this was a very small fraction of the total set, I felt comfortable dropping those rows rather than imputing them. Additional preprocessing included standardizing salary columns as floating-point , rather than strings.
- A data set of graduation rates for football student-athletes, from the NCAA's database at https://web3.ncaa.org/aprsearch/gsrsearch . Due to the vagaries of the web interface it was necessary to download the resulting web page and them read it using panda's read_html() function.
- The 2019 win-loss percentage for all teams. (2019 was used rather than 2020 owing to the ongoing nature of the current season as well as the disruptions of the COVID-19 pandemic).This was collated by a user on Kaggle and was downloaded from https://www.kaggle.com/jeffgallini/college-football-team-stats-2019 . The data set included a considerable variety of statistics (passing and running yards, and many more) but only the column for overall record was used. Some parsing was needed to convert records of the form "13-3" into a win-loss percentage.
- A dataset of football stadium sizes and locations, taken from a Github project at https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv . This dataset included all Division I stadia, but was filtered for only Division 1-A (Football Bowl Subdivision) schools. It did not require any other preprocessing.

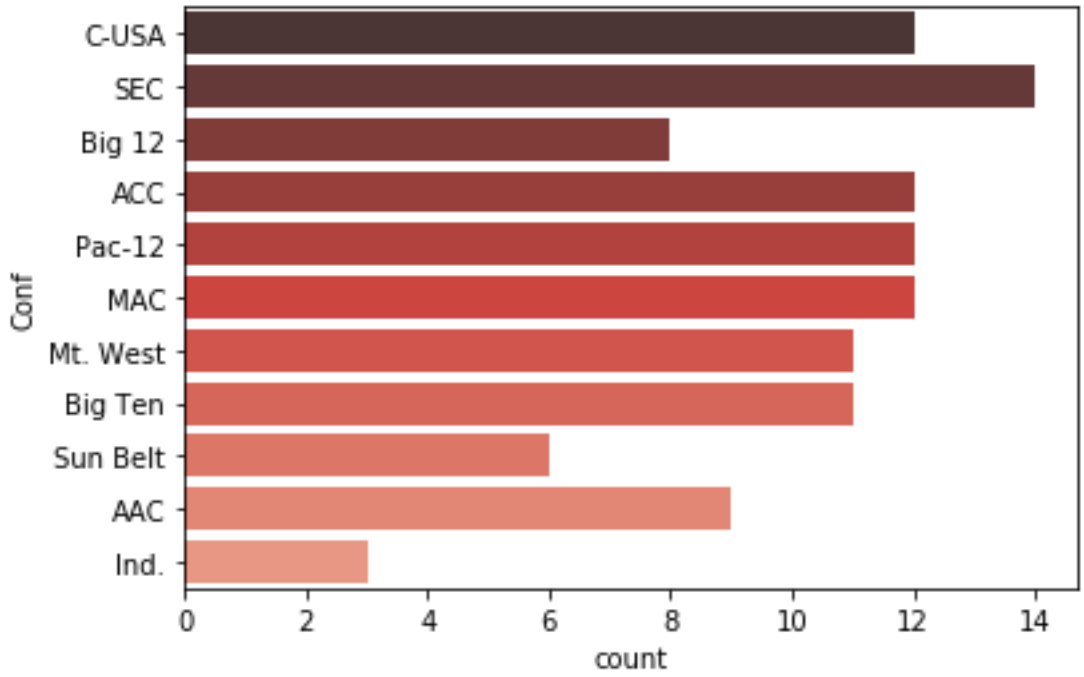
All four data sets were then merged. Since the datasets were not consistent in their identification of schools (with some using the full names and others shorthand), this was a difficult

process. I used the fuzzymatcher library to match rows with sufficiently similar names and conferences, but this matching was not perfect. The most expeditious way to correct the data set was to export the combined dataframe as a csv and review it in Excel. This analysis revealed about 15 bad joins; such rows were deleted, leaving a respectable 111 rows out of the original 130. At this stage duplicate columns were also deleted. (Another approach would have been to do all preprocessing and joining in Excel, but that's hardly suitable for an advanced analytics course). The amended data set was then read back into a pandas dataframe.

A notable limitation of this data set is the inconsistency of year. The salary data is from 2018, the win-loss record from 2019, the graduation data is based on the cohort that matriculated in 2006, and the stadium data is undated (although stadia presumably vary less from year to year). This means that we should be wary about inferring causality, particularly for the relationship between salary and win-loss record. Do better coaches get rewarded with more pay, or is it merely that higher-ranking programs feel obliged to pay for more expensive coaches? With this data set, it is difficult to say.

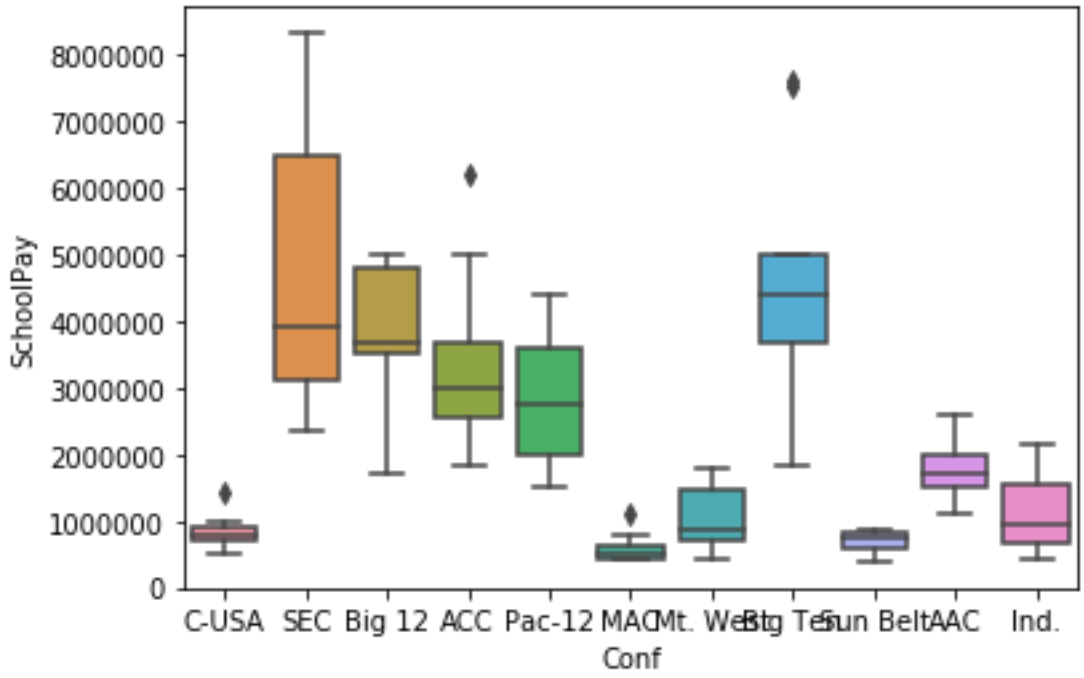
EDA and Graphical Interpretation

Since some rows were dropped during preprocessing, it is important to make sure that no one conference was unduly affected:

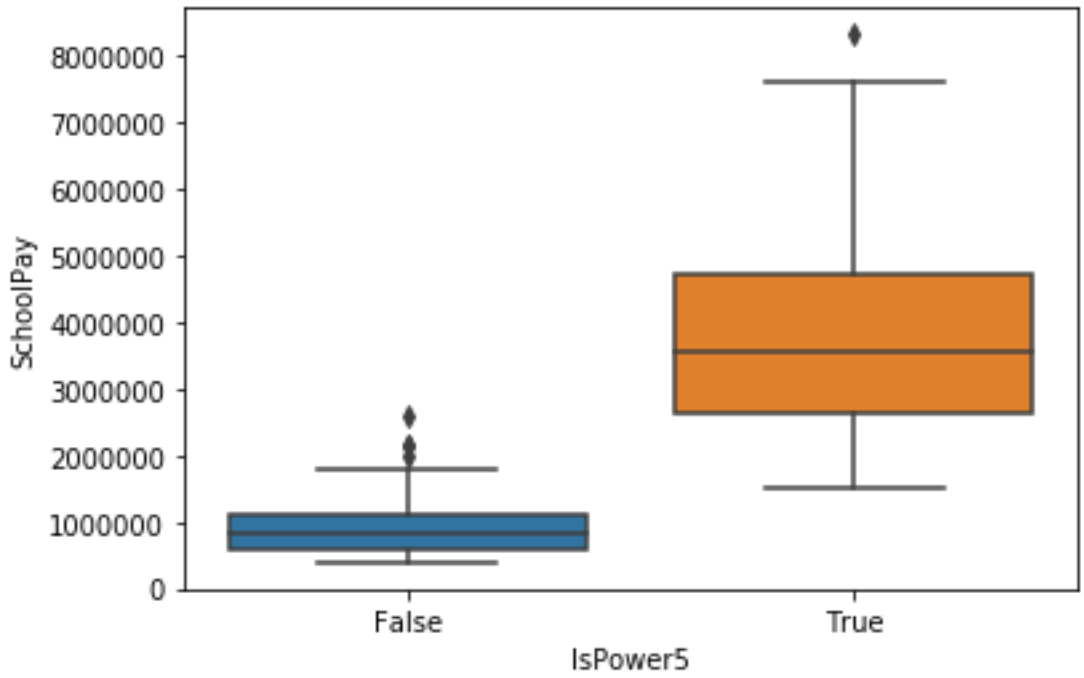


Some conferences were more affected than others, but no conference is so reduced as to make inferences about it impossible. (There shortest bar is for the 3 independent schools, all of which were included in the data).

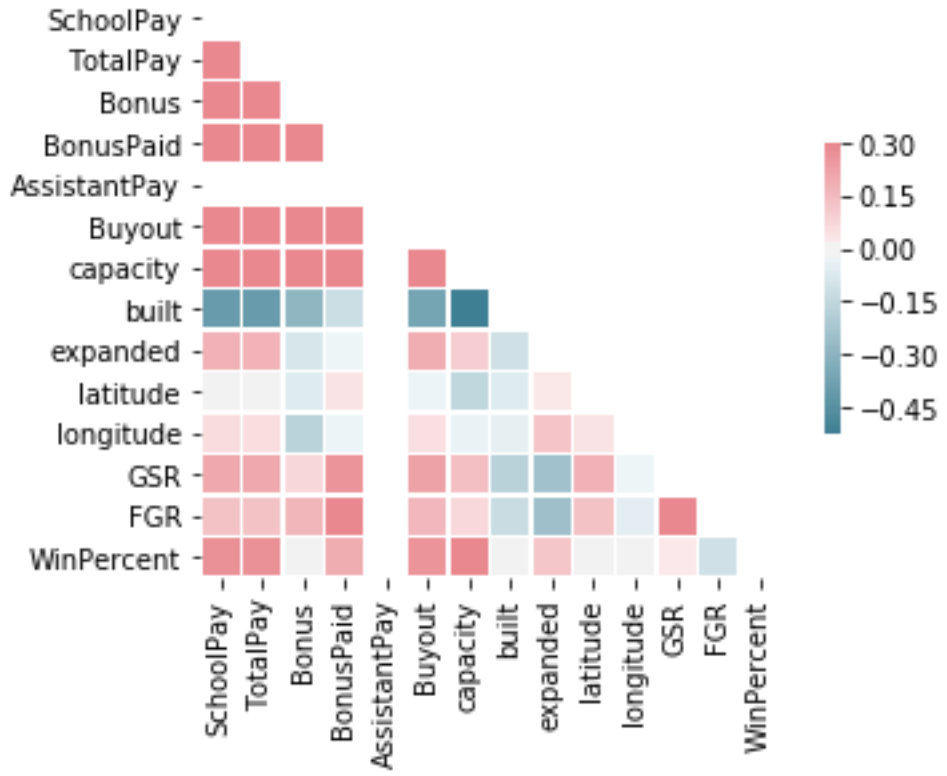
To what extent do conferences vary in the salaries paid to coaches? The differences are quite stark:



The most lucrative conferences completely dominate the smaller ones, with no overlap in salary. Considering the fundamental distinction between the “Power 5” conferences and the others, the distinction is even starker:

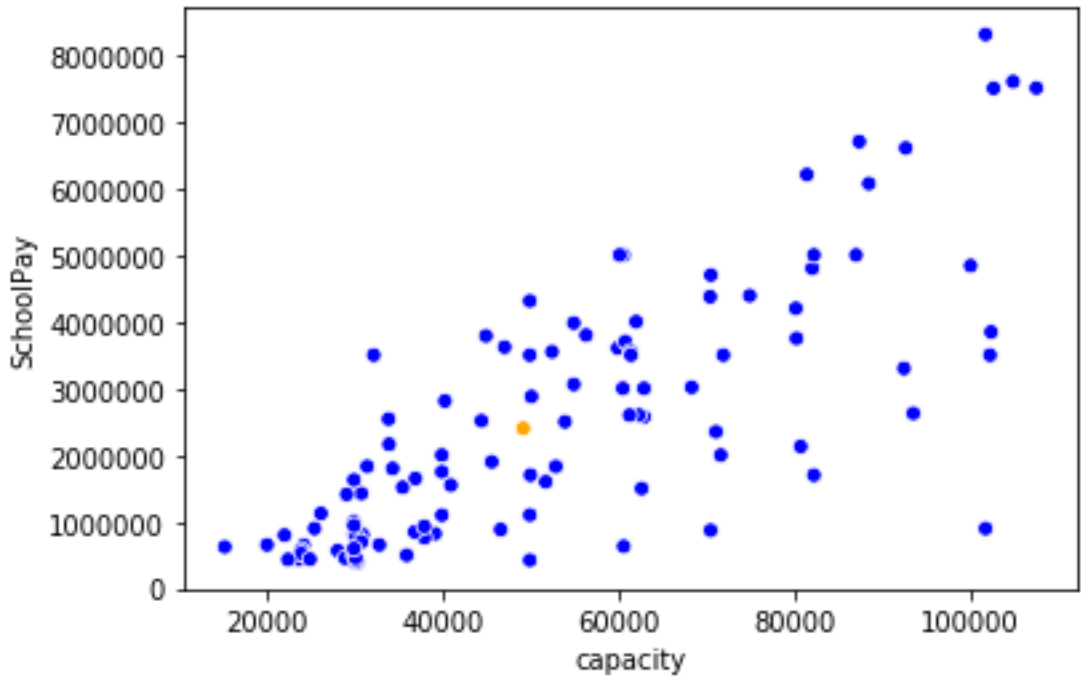


But there are many variables beyond conference that may affect coach pay. First, I prepared a heat map of correlations between all quantitative data:



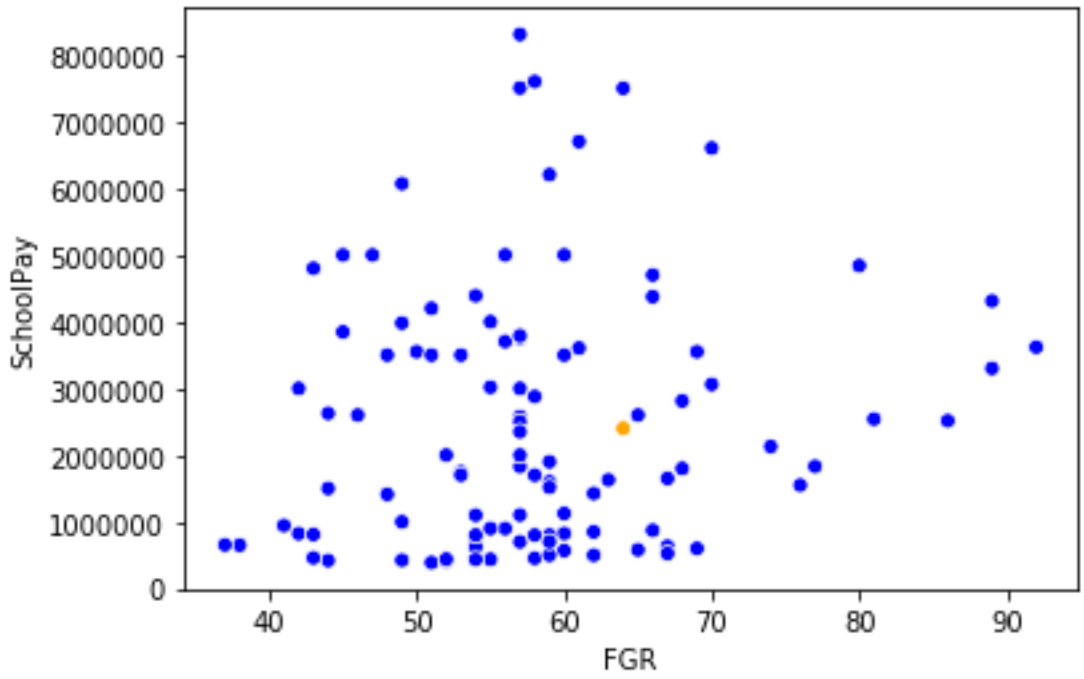
These correlations are not enormous – positive correlations top out at 0.30; negative at -0.45 – but still striking. It is unsurprising that the various components of pay in addition to base salary are all positively correlated with it. The strongest negative correlation is that between the year a stadium was built and its capacity. This reflects the stadium criteria of a different era, when sheer seating capacity was more important than providing features like luxury boxes.

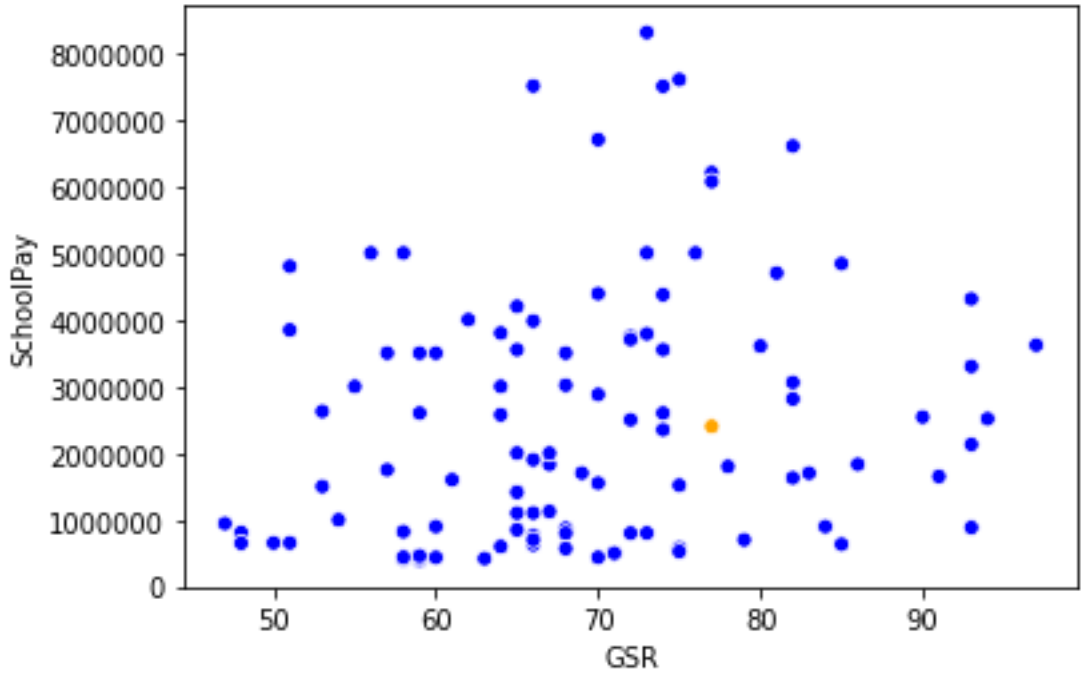
Next, I considered the relationship between coach salary and stadium size. Presumably, teams with bigger stadia have larger fan bases and hence more money to spend on luring top talent.



(The orange dot is, of course, Syracuse). The hypothesis was broadly borne out, with an obvious positive association between capacity and pay, although with significant variation. As discussed above, the largest stadiums were mostly built in the 1920s and 1930s (prior to the emergence of the NFL as the premiere attraction for football fans), and schools that were major then may no longer be as dominant, which may explain the dots in the lower right quadrant.

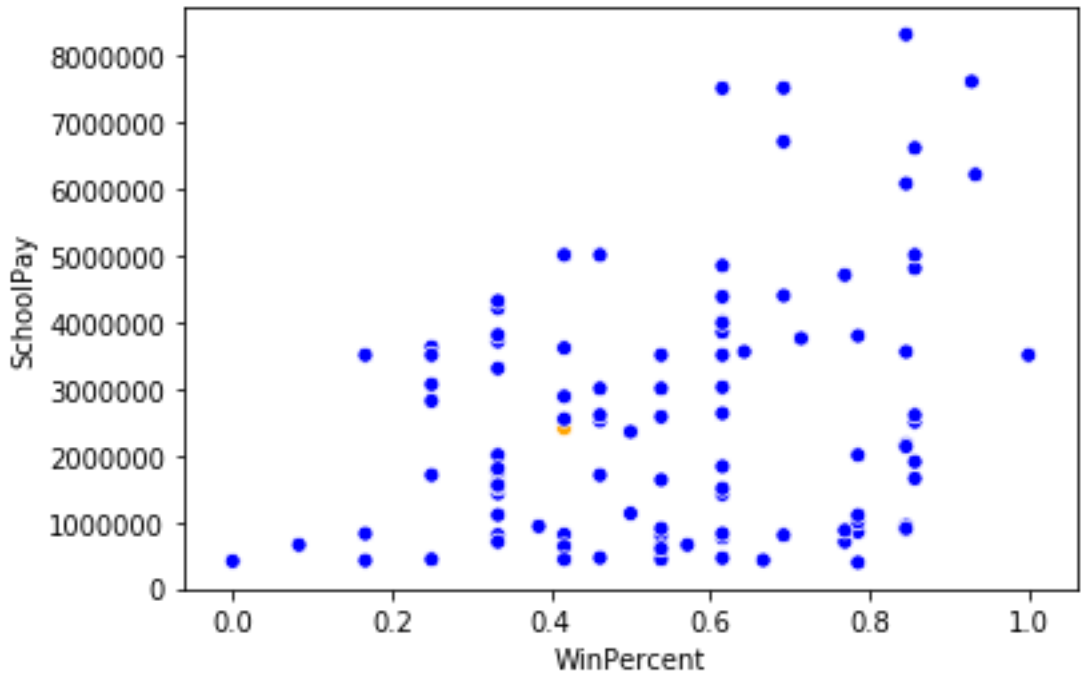
For graduation rate, the link is less obvious:





The graduation rate of student athletes has fairly little relation to coach pay.

Finally, there is a moderate relationship between win rate and pay:



The likely reason that this correlation is only moderate is the intersection with conference – a weak school in a strong conference might be expected to have a more expensive coach than a strong school in a weak conference, but the former school would of course have a lower win rate.

Analysis

Initially, I tried to fit a linear model using all data in the set, except for the other forms of pay (such as bonus pay), and ID-type variables like the name of the school and the coach. This produced the following regression:

OLS Regression Results

Dep. Variable:	SchoolPay	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	10.53
Date:	Sat, 17 Oct 2020	Prob (F-statistic):	1.36e-10
Time:	18:53:08	Log-Likelihood:	-946.76
No. Observations:	63	AIC:	1932.
Df Residuals:	44	BIC:	1972.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.576e+07	3.42e+07	-1.340	0.187	-1.15e+08	2.31e+07
Conf[T.ACC]	1.475e+06	6.11e+05	2.416	0.020	2.45e+05	2.71e+06
Conf[T.Big 12]	1.897e+06	6.43e+05	2.952	0.005	6.02e+05	3.19e+06
Conf[T.Big Ten]	2.108e+06	6.7e+05	3.147	0.003	7.58e+05	3.46e+06
Conf[T.C-USA]	3.843e+05	1.1e+06	0.348	0.730	-1.84e+06	2.61e+06
Conf[T.Ind.]	6.602e+05	1.22e+06	0.540	0.592	-1.8e+06	3.12e+06

Conf[T.MAC]	-3.42e+05	6.36e+05	-0.538	0.593	-1.62e+06	9.4e+05
Conf[T.Mt. West]	3.956e+05	8.51e+05	0.465	0.644	-1.32e+06	2.11e+06
Conf[T.Pac-12]	1.418e+06	8.94e+05	1.587	0.120	-3.83e+05	3.22e+06
Conf[T.SEC]	1.906e+06	6.1e+05	3.122	0.003	6.76e+05	3.14e+06
Conf[T.Sun Belt]	1.569e+04	7.54e+05	0.021	0.983	-1.5e+06	1.54e+06
capacity	34.7998	9.544	3.646	0.001	15.565	54.035
built	4412.5058	7130.587	0.619	0.539	-9958.248	1.88e+04
expanded	1.786e+04	1.49e+04	1.199	0.237	-1.22e+04	4.79e+04
latitude	1.114e+04	4.31e+04	0.258	0.797	-7.58e+04	9.81e+04
longitude	1.955e+04	1.98e+04	0.986	0.330	-2.04e+04	5.95e+04
GSR	1.585e+04	2.33e+04	0.679	0.501	-3.12e+04	6.29e+04
FGR	8357.0615	2.32e+04	0.360	0.721	-3.84e+04	5.52e+04
WinPercent	1.67e+06	7.98e+05	2.093	0.042	6.2e+04	3.28e+06

Omnibus:	0.589	Durbin-Watson:	1.920
Prob(Omnibus):	0.745	Jarque-Bera (JB):	0.360
Skew:	-0.185	Prob(JB):	0.835
Kurtosis:	3.009	Cond. No.	1.70e+07

Notably, only some of the conferences were individually significant. The Power 5 conferences were all individually significant at the $p < .05$ level (relative to the base case, which was the non-power-5 AAC), but the other non-power conferences were not. Capacity and win percentage were also highly significant. No other predictors, including either metric of graduation rate, was significant. Overall, the model accounted for a quite substantial 81.2% of variation in salary.

I then re-ran the analysis, looking only at those variables found to be significant:

OLS Regression Results

Dep. Variable:	SchoolPay	R-squared:	0.770
Model:	OLS	Adj. R-squared:	0.729
Method:	Least Squares	F-statistic:	18.66
Date:	Sat, 17 Oct 2020	Prob (F-statistic):	6.54e-17
Time:	19:01:53	Log-Likelihood:	-1207.0
No. Observations:	80	AIC:	2440.
Df Residuals:	67	BIC:	2471.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.311e+05	5.03e+05	-0.460	0.647	-1.23e+06	7.72e+05
Conf[T.ACC]	1.259e+06	4.95e+05	2.546	0.013	2.72e+05	2.25e+06
Conf[T.Big 12]	1.553e+06	5.02e+05	3.097	0.003	5.52e+05	2.55e+06
Conf[T.Big Ten]	2.048e+06	4.78e+05	4.284	0.000	1.09e+06	3e+06
Conf[T.C-USA]	-1.729e+04	5.55e+05	-0.031	0.975	-1.12e+06	1.09e+06
Conf[T.Ind.]	-3.965e+05	7.57e+05	-0.524	0.602	-1.91e+06	1.12e+06
Conf[T.MAC]	-5.428e+05	4.75e+05	-1.143	0.257	-1.49e+06	4.05e+05
Conf[T.Mt. West]	-4.562e+05	4.35e+05	-1.048	0.299	-1.33e+06	4.13e+05
Conf[T.Pac-12]	7.84e+05	4.68e+05	1.677	0.098	-1.49e+05	1.72e+06
Conf[T.SEC]	1.446e+06	5.06e+05	2.858	0.006	4.36e+05	2.46e+06
Conf[T.Sun Belt]	-5.918e+05	5.4e+05	-1.095	0.277	-1.67e+06	4.87e+05
capacity	30.4779	7.206	4.229	0.000	16.094	44.862
WinPercent	9.513e+05	5.69e+05	1.672	0.099	-1.85e+05	2.09e+06

Using this model, overall R^2 was lower, but adjusted R^2 – a measure of the degree to which adding new information improves a model more than would be expected due to chance – was nearly identical.

Finally, we consolidated the conference variable into a single dichotomous variable indicating whether the school was in the Power 5 or not:

OLS Regression Results

Dep. Variable:	SchoolPay	R-squared:	0.733
Model:	OLS	Adj. R-squared:	0.722
Method:	Least Squares	F-statistic:	69.44
Date:	Sat, 17 Oct 2020	Prob (F-statistic):	1.02e-21
Time:	19:07:15	Log-Likelihood:	-1213.0
No. Observations:	80	AIC:	2434.
Df Residuals:	76	BIC:	2444.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.862e+05	3.26e+05	-2.107	0.038	-1.33e+06	-3.76e+04
IsPower5[T.True]	1.65e+06	2.99e+05	5.510	0.000	1.05e+06	2.25e+06
capacity	33.3990	6.803	4.909	0.000	19.849	46.949
WinPercent	9.961e+05	5.35e+05	1.862	0.066	-6.93e+04	2.06e+06

Omnibus:	0.842	Durbin-Watson:	1.673
Prob(Omnibus):	0.656	Jarque-Bera (JB):	0.424
Skew:	0.152	Prob(JB):	0.809
Kurtosis:	3.185	Cond. No.	3.09e+05

This performed nearly as well as the previous regression, indicating that the principal difference between conferences is whether that conference is in the Power 5, rather than sub-variation within Power 5 conferences.

The predictors with the greatest impact on salary size were Power-5 status and stadium capacity. The former had the greater impact except when extreme differences in capacity – between 15,000 and 100,000 capacity, say – were considered.

Using the second model, we may finally predict Syracuse's coach's salary. The model predicts a salary of **$2,925,504**. But this is not quite the same as saying that the university should definitely offer this salary to a new coach. It might be wise to try to negotiate as we would for an ordinary job – perhaps we should make an initial offer of a mere $2.2 million, but then be willing to move upwards. Conversely, if there is a causal link between salary and team performance (a question that this analysis cannot resolve, unfortunately), we might want to pay more in the hope of attracting a particularly talented coach.

We may also predict the salary if Syracuse were to change conferences. If Syracuse joined the Big Ten, where salaries average nearly $800,000 more than in the ACC, we would want to pay about $3.7 million. What if the Big East still existed as a football conference, and Syracuse was still a member. This is a harder question to answer, as we have no group of Big East teams to run a comparison to. A rough estimate might involve placing the Big East midway between a Power Five conference and a non-Power Five conference. Taking the midpoint of the Power Five regression coefficient and the non-Power-Five conefficent, we can estimate that a "half Power 5" school would have an average salary about $800,000 lower, implying that Syracuse's coach would only earn about $2.1 million.