IST 718
Aneeka Latif
Lab 3

Introduction

Collegiate football is a highly lucrative sport, averaging as much as $2.7 billion dollars in revenue per year (Smith, 2019). Coach's salaries can range from just under a million dollars to tens of millions of dollars. The National Collegiate Athletics Association categorizes schools into conferences and divisions based on location and school sizeespectively. In this dataset we will explore the schools in Division I which tend to include teams that have higher budgets, more exhaustive programs, and better facilities than other divisions. While it may make sense that those involved in sports that generate such high revenue are entitled to high paying salaries, there is a large variance in salaries between coaches within the same division. Coach's salaries are not only a result of their school's revenue, in this study we will investigate the different attributes affecting coach's salaries.

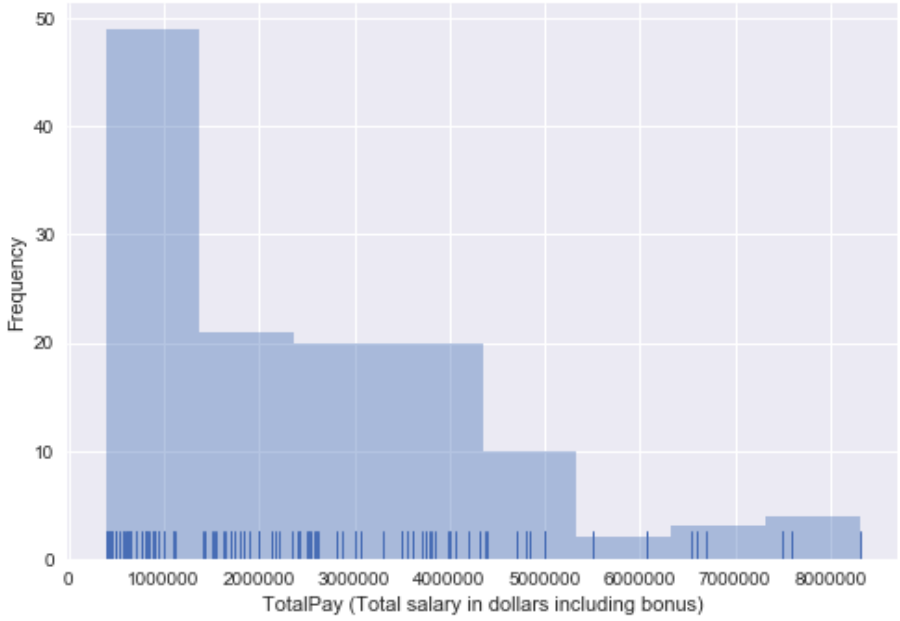
About the data

The base dataset is a set of coaches, schools, conferences, and their pay broken down into base school pay, bonus potential, bonus paid out, and the total pay overall. From this, I referenced the College GridIrons page (*"College Football Stadium Comparisons"*) and added stadium information, as well as the stadium capacity by scraping the text from the webpage. Next, I added win/loss information available on the NCAA page for 2015-2016 year by importing the data from an excel file *("Winning Percentage").* In the same way, I added the graduation rate from the school as well as the federal graduation rate. These metrics differ slightly because the federal graduation rate does not account for transfer students (*"Graduation Rates."*).

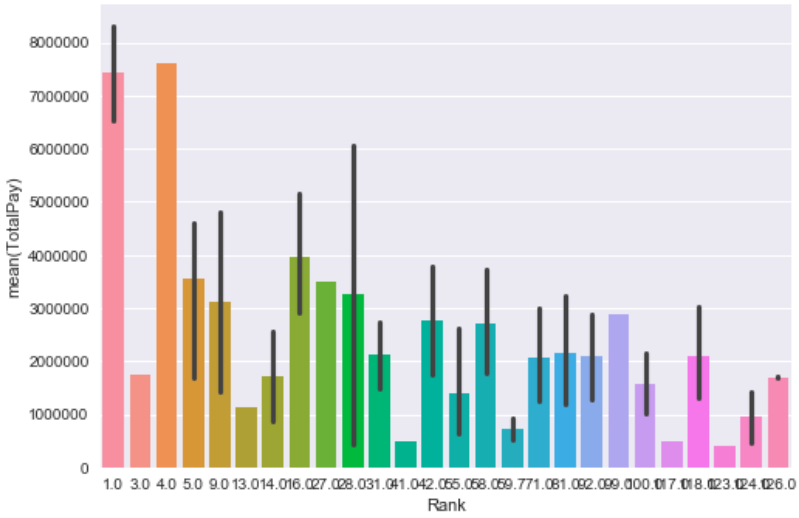
In terms of cleaning the data, the majority of time was spent making sure the school names were standardized and matched in each table I was trying to join – the names became my unique joining identifier/key in every additional sheet of information. Additionally, I needed to explicitly transform numeric values as numeric, and also remove "$" and commas from monetary values.

I did not drop any schools from the dataset, rather, I resolved significant missing identifying values such as name or stadium information by researching and changing adding the values as this was a very small dataset. This was only needed for two rows of the data. In the case of missing numeric values, I replaced the missing values with the average value for that attribute. Because Totalpay, SchoolPay, and Bonus/BonusPaid are all highly correlated, moving forward, I only chose to work with TotalPay as it is a cumulation of all variables.

IST 718
Aneeka Latif
Lab 3



Looking at the distribution of TotalPay, it looks like most of our coaches are paid between 1-2 million dollars. The next major group is between 2 million and 4 million; a small number of coaches are paid significantly higher than that.

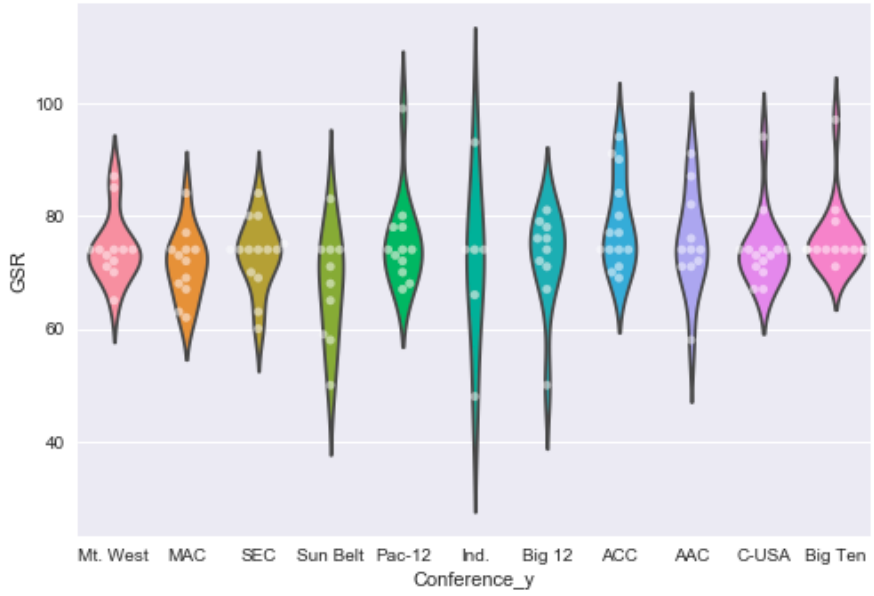


One would think that schools with the highest ranks would have coaches with the highest salaries. Generally, that looks to be the case, the lower ranked schools on the right end side have a lower average mean. The highest ranked schools on the left have higher average mean. However dispersed inbetween the two extremes, both highly ranked schools and poorly ranked schools have coaches with significantly lower than average TotalPay. Rank must not contribute significantly to coach's salary.

IST 718
Aneeka Latif
Lab 3

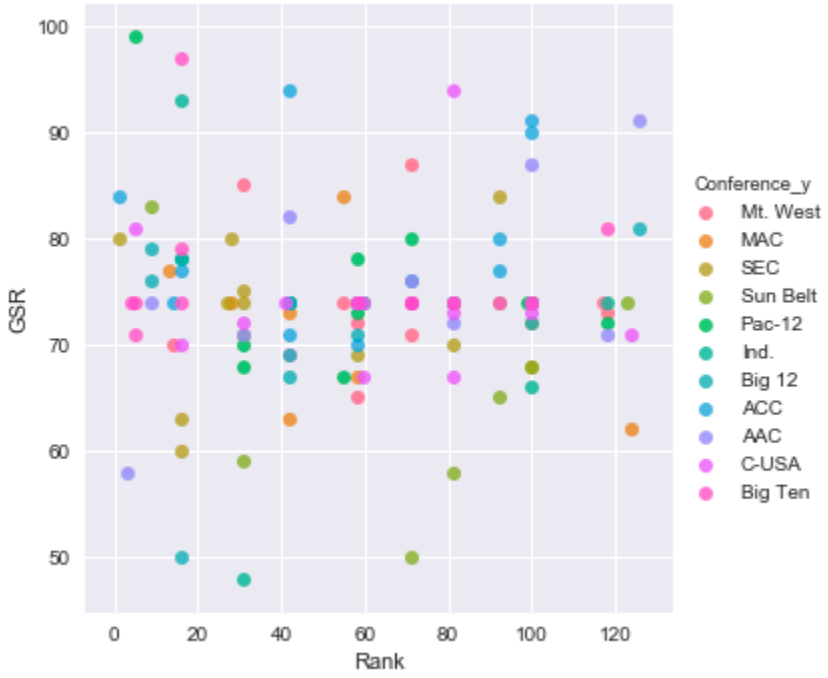
	Stadium	College	Open_date	Capacity	Conference_y	Coach	SchoolPay	TotalPay	Bonus	BonusPaid	AssistantPay	Buyout	Rank	W	L
2	Bryant Denny Stadium	Alabama	1929	101821	SEC	Nick Saban	8307000	8307000.0	1100000	500000	0	33600000	1.0	14.00	1.00
76	Ohio Stadium	Ohio State	1922	104944	Big Ten	Urban Meyer	7600000	7600000.0	775000	350000	0	38058402	4.0	12.00	1.00
57	Michigan Stadium	Michigan	1927	107601	Big Ten	Jim Harbaugh	7504000	7504000.0	1325000	150000	0	17111110	16.0	10.00	3.00
99	Kyle Field	Texas A&M	1904	102733	SEC	Jimbo Fisher	7500000	7500000.0	1350000	--	0	68125000	42.0	8.00	5.00
9	Jordan Hare Stadium	Auburn	1939	87451	SEC	Gus Malzahn	6700000	6705656.0	1400000	375000	0	32143750	58.0	7.00	6.00
33	Sanford Stadium	Georgia	1929	92746	SEC	Kirby Smart	6603600	6603600.0	1150000	1350000	0	27917500	16.0	10.00	3.00

Looking at a snippet of rows of the top 6 highest paid coaches – the school rankings are 1, 4, 42, 58, and 16 – confirming the idea of rank not being a significant factor of salary.

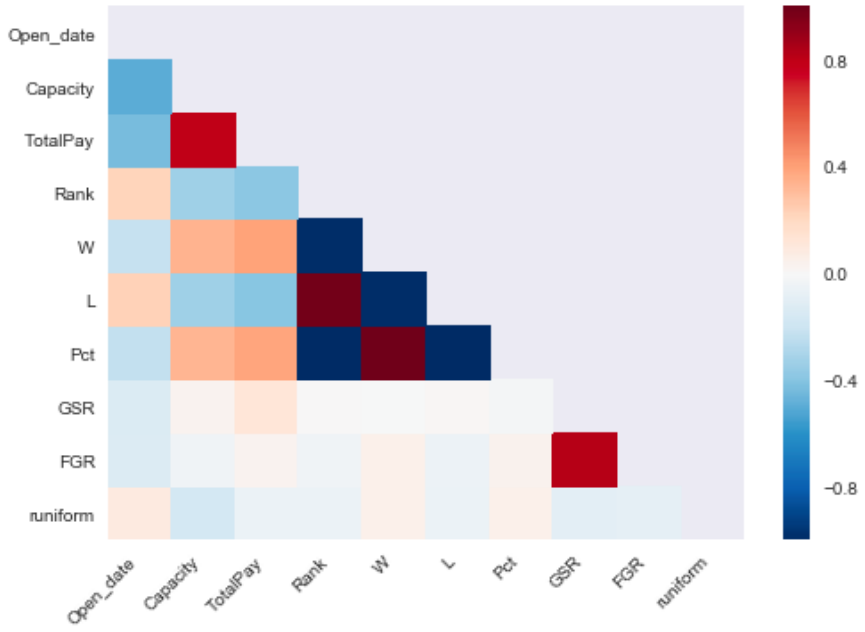


By drawing a violin plot as well as a swarm plot of graduation rates per conference, we can see that Mount West and the Big Ten conferences seem to have the best average graduation rates. The Independent party has the highest variance of graduation rates. Knowing that conferences tend to be based on geographical region, I wondered if we would see a correlation between rank and graduation rates by conference.

IST 718
Aneeka Latif
Lab 3



Plotting the data, it appears that graduation rate does not have any correlation with rank. It looks like the schools in the Big Ten all have the same average graduation rate despite having varied ranks. Schools in the ACC appear to both have generally higher graduation rates and higher ranks. The independent conference has both very high and low graduation rates despite having the same rank.



IST 718
Aneeka Latif
Lab 3

A heatmap showing the correlation of quantitative variables depicts the only strong positive correlation between TotalPay and any variables is with Capacity. The only other moderately positive correlations are with W (number of games won), GSR (graduation rate), and Pct which is the ratio of wins(W) to losses(L) and is strongly correlated with both.

Methods and Models

I used a linear regression model using ordinary least squared for my model. I split the data into a test and training set using the attributes Rank (school rank), GSR (graduation rate), Conference, Pct (Wins to losses/total games), and the stadium capacity to predict total pay.

OLS Regression Results

Dep. Variable:	TotalPay	R-squared:	0.814
Model:	OLS	Adj. R-squared:	0.778
Method:	Least Squares	F-statistic:	22.81
Date:	Tue, 21 Jan 2020	Prob (F-statistic):	3.78e-21
Time:	16:59:57	Log-Likelihood:	-1319.2
No. Observations:	88	AIC:	2668.
Df Residuals:	73	BIC:	2706.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.651e+06	3.87e+06	-1.202	0.233	-1.24e+07	3.06e+06
Conference_y[T.ACC]	7.278e+05	4.52e+05	1.612	0.111	-1.72e+05	1.63e+06
Conference_y[T.Big 12]	1.198e+06	5e+05	2.398	0.019	2.02e+05	2.19e+06
Conference_y[T.Big Ten]	1.049e+06	4.92e+05	2.130	0.037	6.76e+04	2.03e+06
Conference_y[T.C-USA]	-7.384e+05	4.55e+05	-1.624	0.109	-1.64e+06	1.68e+05
Conference_y[T.Ind.]	-2.735e+05	5.57e+05	-0.491	0.625	-1.38e+06	8.36e+05
Conference_y[T.MAC]	-9.391e+05	4.6e+05	-2.042	0.045	-1.86e+06	-2.27e+04
Conference_y[T.Mt. West]	-1.046e+06	4.67e+05	-2.243	0.028	-1.98e+06	-1.17e+05
Conference_y[T.Pac-12]	1.154e+05	4.5e+05	0.256	0.798	-7.82e+05	1.01e+06
Conference_y[T.SEC]	7.89e+05	5.17e+05	1.527	0.131	-2.41e+05	1.82e+06
Conference_y[T.Sun Belt]	-9.124e+05	4.7e+05	-1.941	0.056	-1.85e+06	2.46e+04
Rank	2.337e+04	2.57e+04	0.909	0.366	-2.79e+04	7.46e+04
GSR	1.31e+04	1.28e+04	1.025	0.309	-1.24e+04	3.86e+04
Pct	5.178e+06	4.1e+06	1.263	0.211	-2.99e+06	1.33e+07
Capacity	40.1805	7.617	5.275	0.000	25.000	55.361

Omnibus:	2.021	Durbin-Watson:	2.180
Prob(Omnibus):	0.364	Jarque-Bera (JB):	1.476
Skew:	0.297	Prob(JB):	0.478
Kurtosis:	3.220	Cond. No.	3.23e+06

Looking at the adjusted r squared value, this model fits 77.8% of the data; the accuracy of the model is 77.8%. The root mean squared error is 874934.63; this is the square root of the average error between predicted TotalPay and actual TotalPay. The RMSE is 36.19% of the average TotalPay, so the model isn't optimal by any means, but is satisfactory.

From the model, looking at the coefficients, we can see that the most significant feature, by an overwhelming magnitude is percent games won (Pct) with a coefficient of 5178000. Other important features would be the Conference, graduation rate, and stadium capacity.

Using the model to predict the salary of Syracuse's coach is $2,977,502 which is very close to the actual total pay noted as $2,401,206. If Syracuse was part of the Big Ten conference, the predicted TotalPay would be higher at $3,298,668. On the other hand, if Syracuse had remained in the Big East conference, the predicted total pay would be $2,977,502; this is expected to match the original prediction because the Big East has been succeeded by / is now known as the "American Athletic Conference" or ACC.

IST 718
Aneeka Latif
Lab 3

Graduation rate has a positive correlation with our predicted salaries; higher graduation ranks predict higher salaries. For Syracuse, if the graduation rate were lowered to 63, the predicted TotalPay would be $2,754,747 – lower than the original predicted pay. If the graduation rate was raised to 92, the predicted TotalPay increases to $3,134,742.

Conclusion

The factors that affect a coach's salary are the conference, percentage of games won, graduation rate, school rank, and stadium capacity. The most prominent attributes were the percentage of games won, followed by the conference. While the different conferences are generally organized by region – there does not seem to be any correlation between conference and rank or conference and graduation rates, excluding the Big Ten which all fall average for graduation rates. If a coach were to want to optimize their salary, the model suggests achieving a high percentage of wins, and picking a school that falls under the SEC conference

References

"College Football Stadium Comparisons." *College Gridirons*, www.collegegridirons.com/comparisons-by-capacity/.

"Graduation Rates." *NCAA.org - The Official Site of the NCAA*, 18 Oct. 2019, www.ncaa.org/about/resources/research/graduation-rates.

"Winning Percentage." *NCAA Statistics*, stats.ncaa.org/rankings/change_sport_year_div.

Smith, Chris. "College Football's Most Valuable Teams: Reigning Champion Clemson Tigers Claw Into Top 25." *Forbes*, Forbes Magazine, 28 Dec. 2019, www.forbes.com/sites/chrissmith/2019/09/12/college-football-most-valuable-clemson-texas-am/.

Wolohan, John. "Blog: Protecting Your College Program With a Buyout Clause". Jan 2011. *Athletic Business,* www.athleticbusiness.com/editor-blogs/blog-protecting-your-college-program-with-a-buyout-clause.html.