Rohini Shrivastava

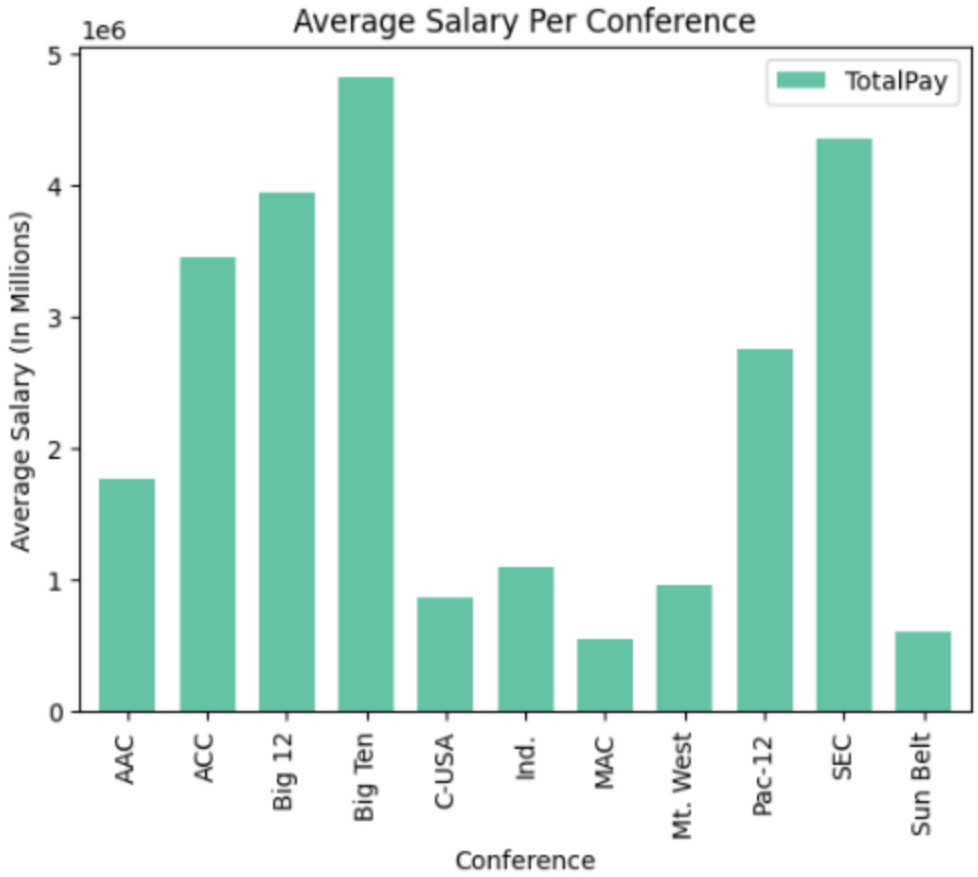
Week 3 Lab

To begin this lab, I took various data from multiple csv files and merged them together to create our dataframe. We used Coaches.csv that was provided to us, along with a list of Staiums exported from Wikipedia, percentage of alumni who donate taken from a github repo, GSR/FED33 graduation data taken from NCAA, and a win/loss ratio for 2019 taken from a Sports Reference website. We merged the data using the 'College' column. Then we converted our numerical objects into the int64 data type so we can analyze it.

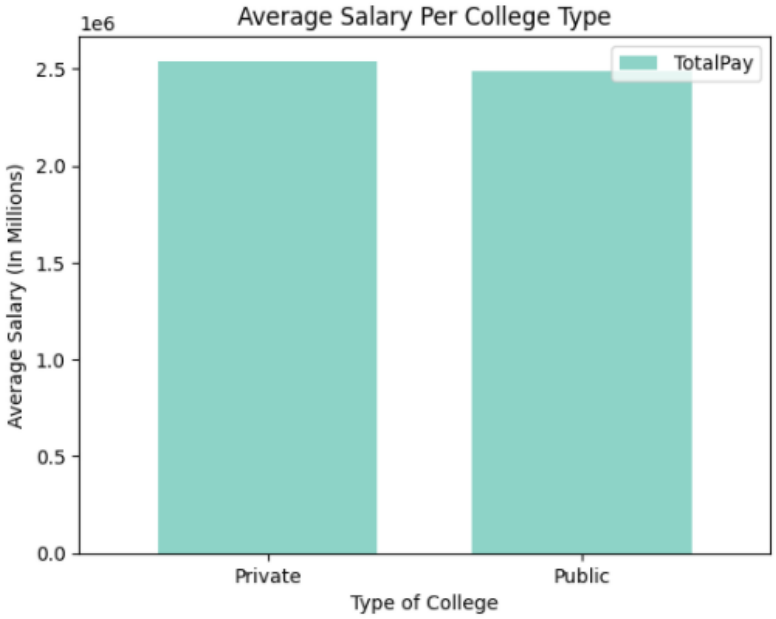
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 78 entries, 0 to 102
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   College         78 non-null     object
 1   Conference      78 non-null     int64
 2   Coach           78 non-null     object
 3   SchoolPay       78 non-null     int64
 4   TotalPay        78 non-null     int64
 5   Bonus           78 non-null     int64
 6   BonusPaid       78 non-null     int64
 7   AssistantPay    78 non-null     int64
 8   Buyout          78 non-null     int64
 9   Stadium         78 non-null     object
 10  Capacity        78 non-null     int64
 11  W               78 non-null     int64
 12  L               78 non-null     int64
 13  SCL_DIVISION    78 non-null     int64
 14  SCL_CONFERENCE  78 non-null     object
 15  SCL_PRIVATE     78 non-null     int64
 16  GSR_2006_SA     78 non-null     int64
 17  FED_N_2006_SA   78 non-null     int64
 18  perc.alumni     78 non-null     int64
dtypes: int64(15), object(4)
memory usage: 12.2+ KB
```

We did some basic analysis on the data set we had created. The average pay for coaches overall was $2,402,868, with the maximum salary being 7,600,000 and minimum being 400,000. For all the schools the average stadium size was 50,643. Average GSR was 81.19 and FED was 69.44. On average 19% of alumni donated back to the school.

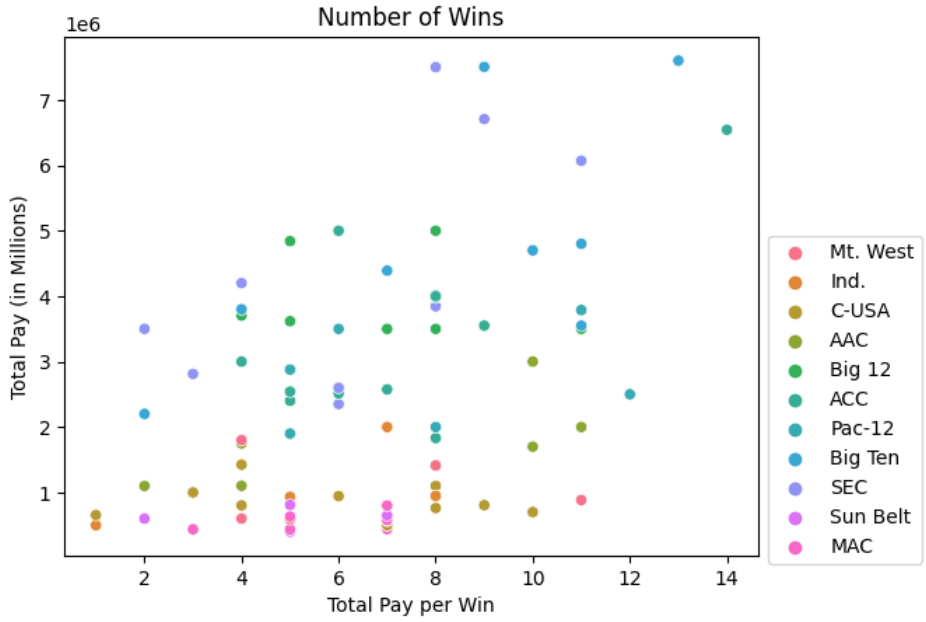
Now this does not tell us much, so I went further and divided the averages into categories. I looked at average Salary per Conference and found that the Big Ten had the highest average salary,



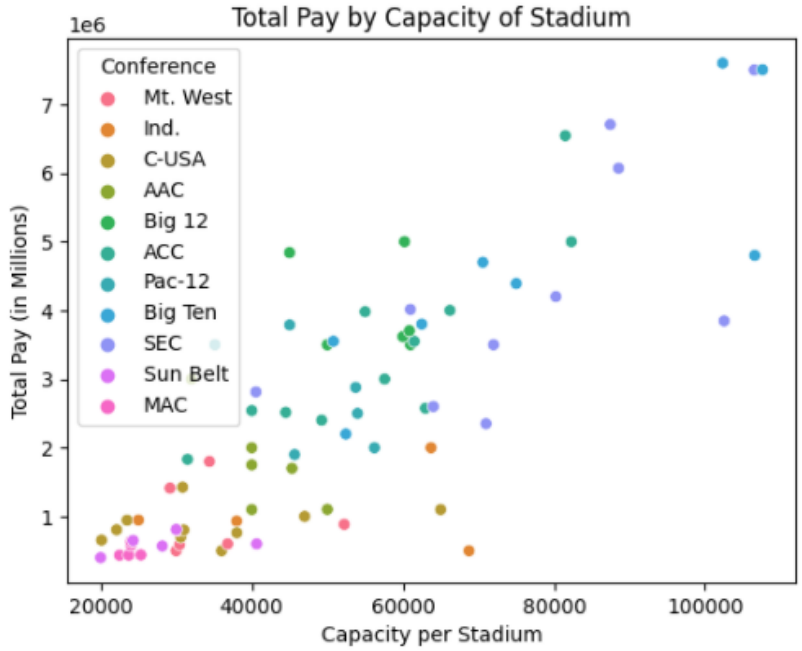
I was also curious to see whether the classification of the school made a difference (i.e. Public vs Private). I created another bar graph for this. We can see that the average salary between the two is about the same.



I then looked at the average salary for a total number of wins. There does not seem to be much of a correlation. We ran a correlation model and found the relationship to be 0.43, indicating one does exist but it is not that closely related.



Lastly a model was created to see total pay vs the stadium capacity. We can see that there seems to be a correlation between the two. After running a correlation model we found the relationship to be 0.83, which gives an indication the two are closely related. Looking at the graph we can see the highest paid and the highest capacity are found for the Big 10 and SEC.



I ran a linear regression model to calculate the salary for a coach based on the entire dataset. We kept School Salary as our dependent variable, with Conference, Number of Wins, Capacity of the Stadium, whether it is a Private/Public school, GSR, FED, and % Donations from alumni as our independent variable.

OLS Regression Results

Dep. Variable:	SchoolPay	**R-squared (uncentered):**	0.866
Model:	OLS	**Adj. R-squared (uncentered):**	0.853
Method:	Least Squares	**F-statistic:**	65.43
Date:	Sat, 17 Oct 2020	**Prob (F-statistic):**	1.96e-28
Time:	03:11:42	**Log-Likelihood:**	-1198.1
No. Observations:	78	**AIC:**	2410.
Df Residuals:	71	**BIC:**	2427.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Conference	-8728.8587	4.49e+04	-0.194	0.847	-9.83e+04	8.09e+04
W	5.529e+04	5e+04	1.106	0.273	-4.44e+04	1.55e+05
Capacity	62.3566	6.708	9.296	0.000	48.981	75.732
SCL_PRIVATE	-2.902e+05	4.1e+05	-0.708	0.481	-1.11e+06	5.27e+05
GSR_2006_SA	-1.734e+04	6642.265	-2.610	0.011	-3.06e+04	-4094.248
perc.alumni	-2057.9413	1.39e+04	-0.148	0.883	-2.98e+04	2.56e+04
FED_N_2006_SA	6771.4229	4178.798	1.620	0.110	-1560.864	1.51e+04

Omnibus:	3.320	**Durbin-Watson:**	1.781
Prob(Omnibus):	0.190	**Jarque-Bera (JB):**	2.990
Skew:	-0.194	**Prob(JB):**	0.224
Kurtosis:	3.877	**Cond. No.**	1.69e+05

We had an R-Squared value of 0.853 which seems high. However, we can see that Conference, Wins, Private, Alumni, and FED are not statistically significant. I kept removing the unsignificant variables until I created a model with all variables that were under P = 0.10. Which left us with three variables, Capacity, GSR, and FED.

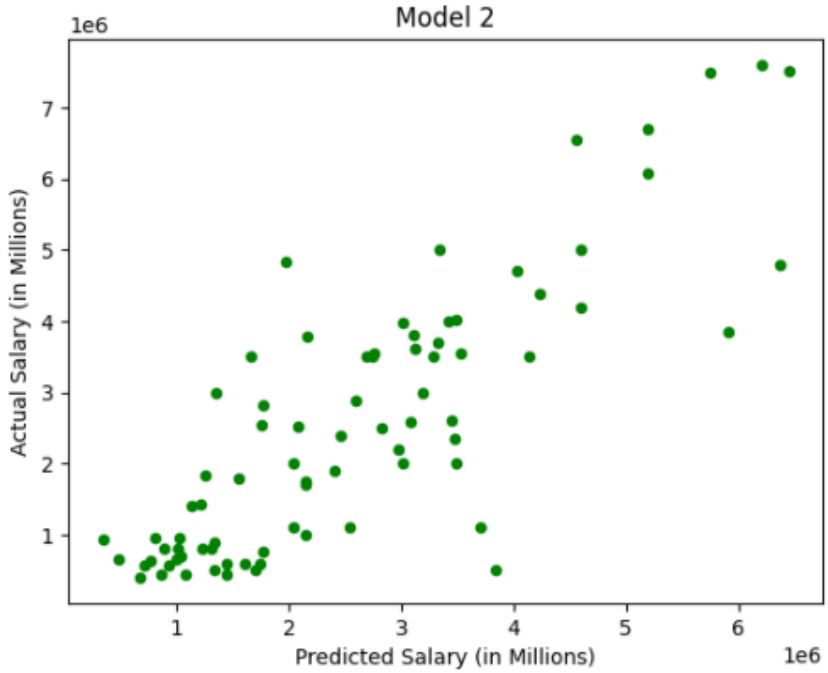
OLS Regression Results

Dep. Variable:	SchoolPay	**R-squared (uncentered):**	0.862
Model:	OLS	**Adj. R-squared (uncentered):**	0.857
Method:	Least Squares	**F-statistic:**	156.6
Date:	Sat, 17 Oct 2020	**Prob (F-statistic):**	3.31e-32
Time:	06:17:40	**Log-Likelihood:**	-1199.1
No. Observations:	78	**AIC:**	2404.
Df Residuals:	75	**BIC:**	2411.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Capacity	64.9551	6.233	10.421	0.000	52.538	77.372
GSR_2006_SA	-1.622e+04	4761.124	-3.406	0.001	-2.57e+04	-6733.163
FED_N_2006_SA	6841.9674	3908.987	1.750	0.084	-945.134	1.46e+04

Omnibus:	5.387	**Durbin-Watson:**	1.771
Prob(Omnibus):	0.068	**Jarque-Bera (JB):**	5.011
Skew:	-0.415	**Prob(JB):**	0.0816
Kurtosis:	3.924	**Cond. No.**	2.25e+03

I decided to run both models and plot the predicted values against the actual values. Model 1 had all the variables, including the ones that were not statistically significant. This gave an R-Squared value of 0.72. The second model only used the significant variables. The R-squared value for the second model was 0.69, showing that even though the model had the significant variables, it was not as accurate as the first one created.



What is the recommended salary for the Syracuse football coach?

According to our two models, the predicted salary and actual salary were very similar. Our actual salary is $2,401,206. In model 1 our predicted salary is $2,604,285 and in model 2 it is $2,463,928. Anywhere between the original salary and model 1's salary would be recommended for the coach.

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

After looking up what the Big East is, it seems the Football part of the conference split up and because the ACC. Syracuse is currently a part of the ACC, so the coach's salary should be the same as what it is currently ($2,401,206).

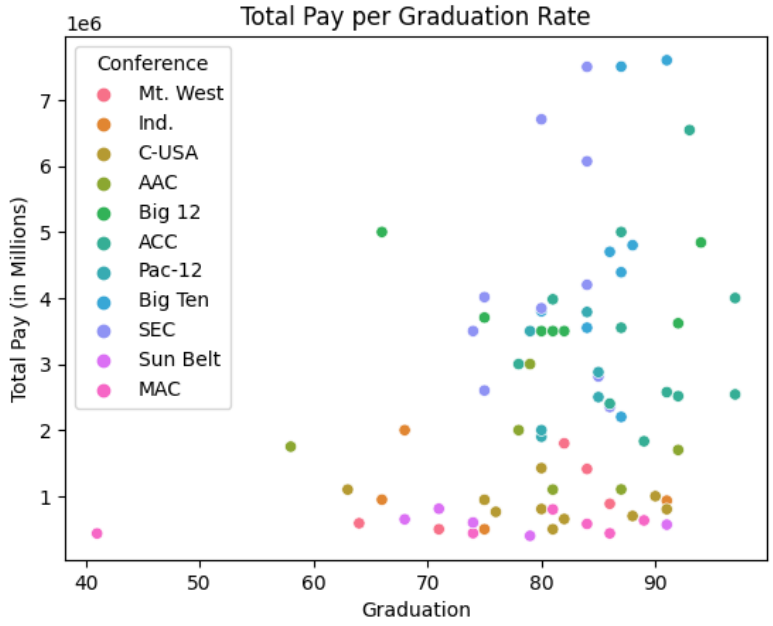
Since we are factoring the conference into our model, we will only be using model 1 to predict his salary. After updating the dataset with Syracuse being in the Big Ten we predicted his salary to be $2,832,029. Earlier we had created a bar graph looking at Average Salary of Coaches vs. Conference the team played in and saw that the Big Ten had the highest average salary. So in this model, it makes sense his salary would be higher than what it currently is.

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

We dropped schools that were not present in all five data sets. This way we do not have partial information for schools, but instead the full scope to accurately determine our model.

What effect does graduation rate have on the projected salary?

By creating a graph showing Total Pay and Graduation we do not see much of a pattern, other than learning its skewed to the right. However, after running the regression model we see that the GSR has a p value of 0.011 (or according to model 2, 0.001). This is lower than our p value of 0.10 so it suggests that the variable is significant to determine the Total Pay for a coach. The coefficient is negative so it is showing that a higher GSR would have a lower salary. That does not make much sense as most people would think there is a positive correlation between the two, but we do see the std for error is very high.



How good is our model?

Both models are pretty accurate. The R-Square for model 1 and 2 were 0.853 and 0.857, respectively. When we predicted the salaries using the models, we were close overall, with the new R-Square values for the Actual Pay and Predicted pay being around 0.72, which is still showing a strong relationship. Our models predict the salary to be a little higher than what it actually is, but looking overall there's not too much of a difference.

What is the single biggest impact on salary size?

Looking back to our regression models, capacity has the smallest P value. There is a very strong positive correlation between the capacity of the stadiums and the salary for the coaches. While there are other variables that have a low P-value, they all have higher standard errors. The error for Capacity is the smallest out of all the variables (significant and not significant).