Reed McElfresh
IST 718: Big Data Analytics
Lab 3

Note: the code for this lab was written in Google Colab and can be found at:
https://colab.research.google.com/drive/1asKGzs5mYT_kZCY4dDdPsgErFtZBVswt
A GitHub Repository for the Lab can be found at: https://github.com/rmcelfresh/IST718Lab3

Introduction

The NCAA has come under criticism in the last several years due to the hardships faced by student athletes whose performances generate millions of dollars for their colleges while being barred from receiving monetary compensation beyond their scholarships (NCAA Should Pay, 2016). However, while student athletes are fighting for any compensation, the coaches leading their teams are coming under scrutiny due to the amount of money they are receiving in both salary and bonuses (Farmer, A & Pecorino, P., 2010). Based on these two factors, it has been determined that the best way to determine the salary of the Syracuse University Football Coach will be to use a data driven approach by looking at the factors that contribute to salaries of other Division 1 Football coaches in the United States.

Analysis

About the Data

Data was collected from several sources. The core dataset was taken from a GitHub repository created by Jon Fox, a professor at Syracuse University (Fox, GitHub). This dataset included team names, current conference, coach, and six different measurements of pay. For the purposes of the analysis, the following variables were retained: School, Conference, Coach, Total Pay, and Bonus Paid. Additionally, a calculated variable was created, adding together the Total Pay and Bonus Paid to create PayBlusBonus. This data was then cleaned to allow it to be used in later analysis.

Wikipedia was used to collect information on stadium size (List Of NCAA…, 2019). The data was scrapped using Pandas and BeautifulSoup and then cleaned so it could be used in later analysis.

Graduation Success Rate (GSR), is a measure of graduation rate of student athletes. It is a derivation of the Federal Graduation Rate (FGR) that was designed to account of the mobility of college athletes that isn't accounted for by the FGR (Why the GSR…, 2014). Both the GSR and FGR will be used to predict compensation. The data was collected from the NCAA database and copied to Microsoft Excel for pre-processing. The csv file was then uploaded to a public GitHub to be imported into Python (GitHub 1). Additional cleaning was then completed in Python.

Team Records for the 2006 season were collected from Google (NCAA 2006, 2006). Again, the data was copied to Microsoft Excel for pre-processing. The csv file was then uploaded to a

public GitHub to be imported into Python (GitHub 2). Additional cleaning was then completed in Python.

Football Team and School Sports revenue was collected from the U.S. Department of Education's Office of Postsecondary Education's Equity in Athletics Data Analysis Tool. A query of the dataset for 2006 data was exported as a CSV for initial pre-processing in Excel. This was then uploaded to GitHub and imported to Python for additional cleaning (GitHub 3).

All datasets were set to have a common name for the school which was then used to create a dataset for analysis. Four schools were dropped from the dataset because they didn't have a GSR figure:
Charlotte
Georgia State
South Alabama
Texas-San Antonio

Three additional schools were dropped from the analysis due to not having a count of total undergrads or football and team revenue for 2006:
Army
Navy
Old Dominion

Eight schools weren't included in the dataframe of 2006 win/loss records. In order to not delete too many observations, these schools had the average of each win/loss variable input used. The schools were:
Appalachian State
Coastal Carolina
Georgia Southern
Liberty
Massachusetts
Notre Dame
Texas State
Western Kentucky

A boxplot was created to view the distribution of coaches' total pay plus bonus and is show as Figure 1. All Coaches' were in the expected distribution besides one, Alabama's Nick Saban who made $8,807,000 of which $500,000 was a bonus.

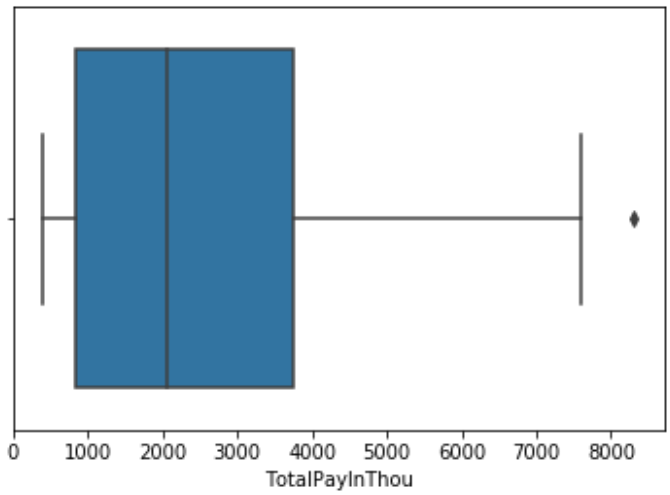

Figure 1: Distribution of All Coaches' Pay in Thousands of Dollars.

In the final dataset, the Conference provided in the Coaches Data from the GitHub repository and the 2006 Conference from the Records data were both included. Boxplots of the Total Pay including Bonuses were created and are shown in Figures 2 and 3. In figure 2, Syracuse is a member of the ACC and in Figure 3 they are a member of the Big East.

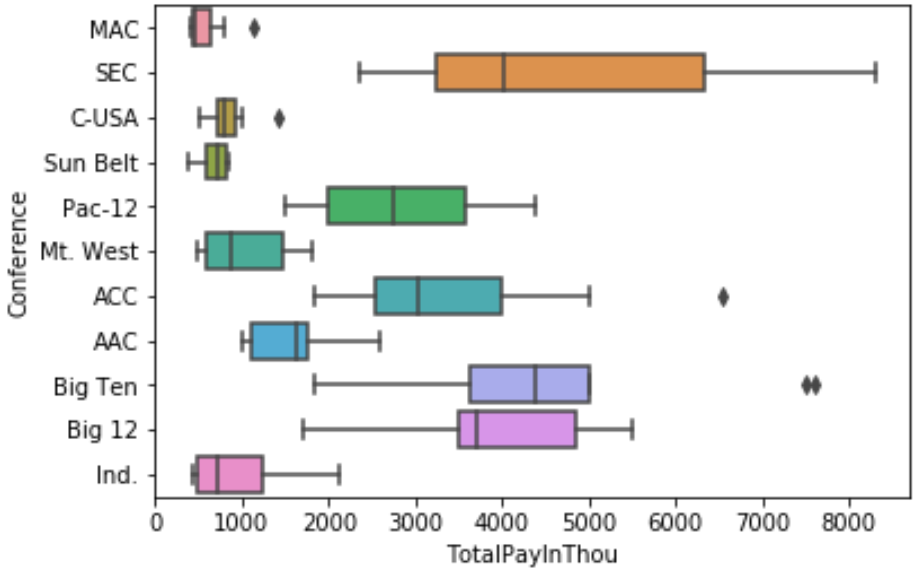

Figure 2: Current Conference Pay Distribution (In Thousand $)

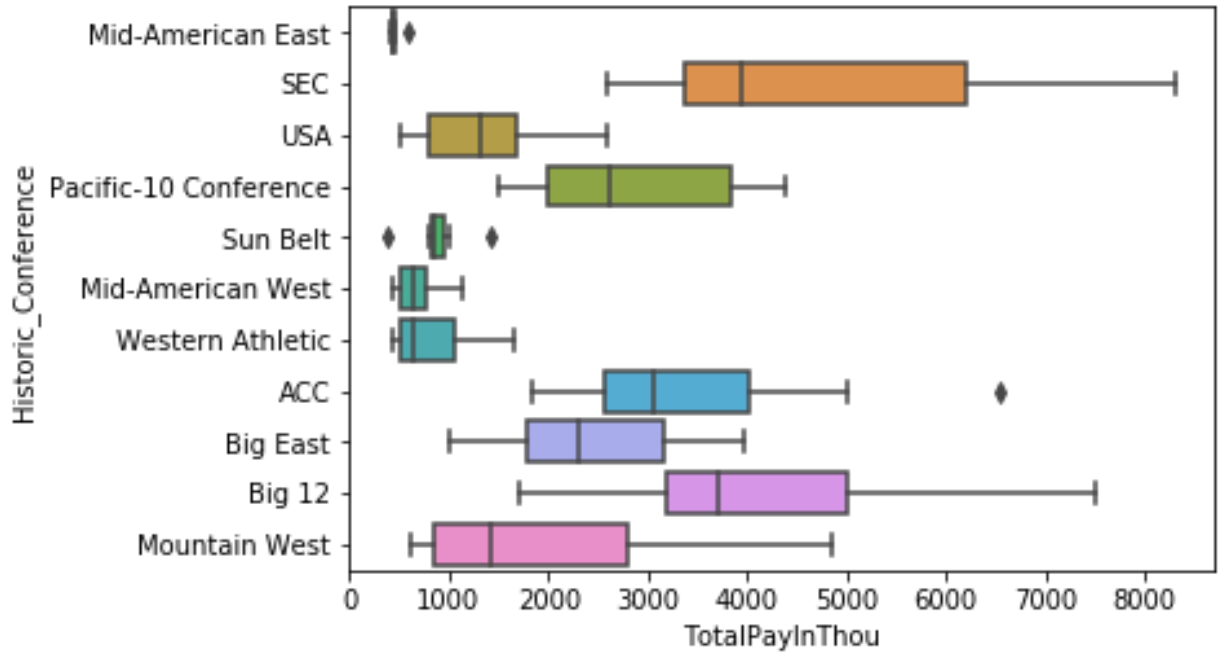

Figure 3: Historic Conference Pay Distribution (In Thousand $)

Data Analysis

The initial analysis was done by creating a heatmap of the correlation of the different variables in the dataframe which is shown in Figure 4. PayPlusBonus was strongly positively correlated to: FootballRevenue, TotalSportsRevenue, and Capacity. There was a slight positive correlation with Total undergrads, OverallWins and HomeWins. There was a slight negative correlation with OverallLoss and AwayLoss. There was virtually no correlation with GSR, FGR, ConferenceWins and AwayWins.

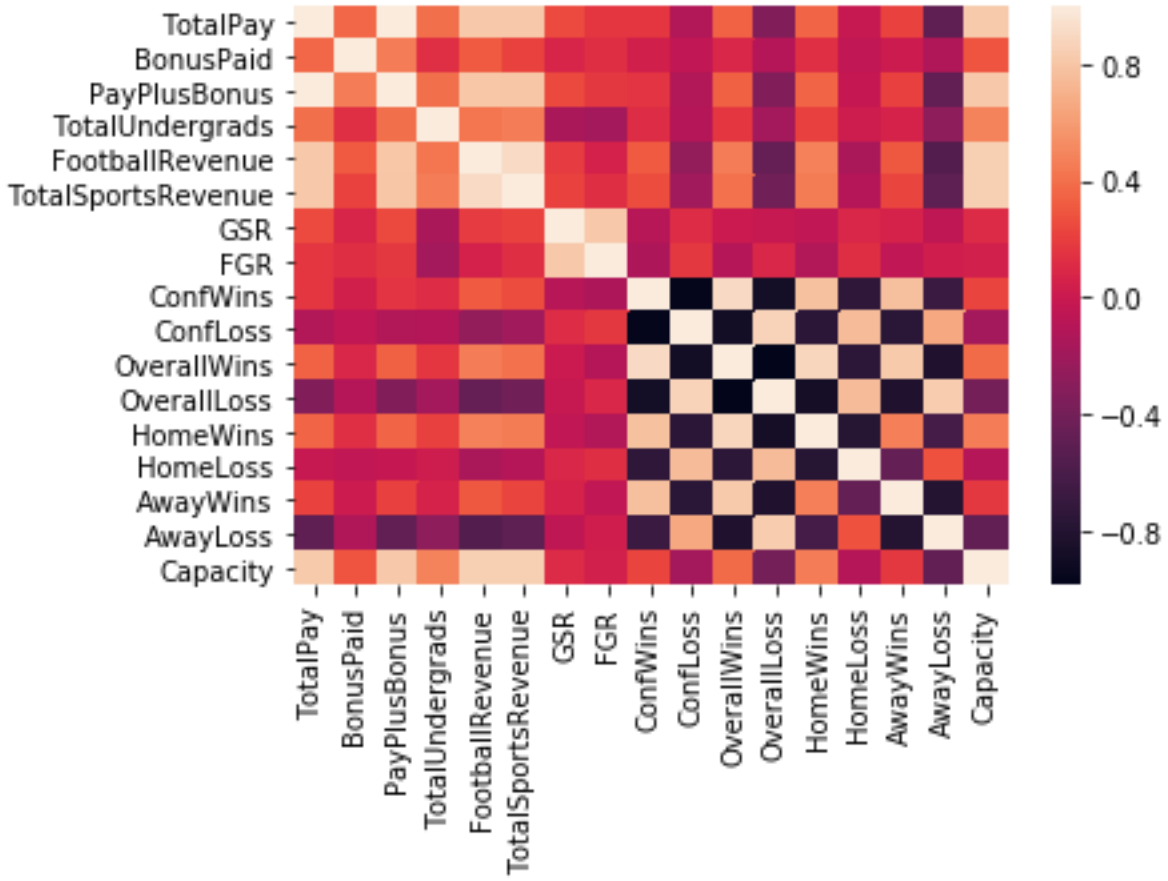

Figure 4: Heat Map of Correlations

A Pair Plot was created which confirmed these findings. It is available in the GitHub Repository that accompanies this paper.

Ordinary Least Squares Regression was used to analyze the effect of these variables on a Coaches' salary plus bonus. The final model included the following variables: Conference, FGR, FootballRevenue and Capacity.

This model was able to predict 77.8% of the total salary plus bonus for the coach with a p-value for the F-Statistic of $3.62e^{-30}$. The model did suggest that there is strong multicollinearity in this model. Removing FootballRevenue and Capacity cleared this warning which indicates that conference is tied to both revenue and capacity. This changed the Adjusted R-Squared to 62.9%. Alternatively, removing Conference and Capacity resulted in an Adjusted R-Squared of 66.9%. However, for further analysis all the variables were kept.

Other versions of the model used OverallWins, OverallLoss, GSR, and TotalUndergrads. Each of these variables were found to have p-values greater than 0.1 and were considered to be non-significant.

Using TotalRevenue as opposed to FootballRevenue was insignificant, changing the Adjusted R-squared to 77.5.

Using GSR as opposed to FGR again was insignificant, changing the Adjusted R-Squared to 77.6.

The final model is as follows:

OLS Regression Results

Dep. Variable:	PayPlusBonus	R-squared:	0.803
Model:	OLS	Adj. R-squared:	0.778
Method:	Least Squares	F-statistic:	32.03
Date:	Sun, 28 Apr 2019	Prob (F-statistic):	3.62e-30
Time:	02:57:46	Log-Likelihood:	-1752.8
No. Observations:	116	AIC:	3534.
Df Residuals:	102	BIC:	3572.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.63E+05	6.85E+05	-0.967	0.336	-2.02E+06	6.97E+05
Conference[T.ACC]	9.82E+05	4.19E+05	2.341	0.021	1.50E+05	1.81E+06
Conference[T.Big 12]	1.34E+06	4.66E+05	2.868	0.005	4.12E+05	2.26E+06
Conference[T.Big Ten]	1.18E+06	4.58E+05	2.579	0.011	2.73E+05	2.09E+06
Conference[T.C-USA]	-4.38E+05	4.36E+05	-1.004	0.318	-1.30E+06	4.27E+05
Conference[T.Ind.]	-9.81E+05	5.90E+05	-1.662	0.1	-2.15E+06	1.89E+05
Conference[T.MAC]	-4.17E+05	4.40E+05	-0.949	0.345	-1.29E+06	4.55E+05
Conference[T.Mt. West]	-3.14E+05	4.26E+05	-0.737	0.463	-1.16E+06	5.31E+05
Conference[T.Pac-12]	3.67E+05	4.32E+05	0.849	0.398	-4.90E+05	1.22E+06
Conference[T.SEC]	1.18E+06	4.63E+05	2.553	0.012	2.64E+05	2.10E+06
Conference[T.Sun Belt]	-3.44E+05	4.70E+05	-0.733	0.465	-1.28E+06	5.87E+05
FGR	1.84E+04	8944.39	2.055	0.042	642.444	3.61E+04
FootballRevenue	0.0438	0.011	4.018	0	0.022	0.065
Capacity	21.3845	8.027	2.664	0.009	5.463	37.306

Omnibus:	6.528	Durbin-Watson:	1.862
Prob(Omnibus):	0.038	Jarque-Bera (JB):	6.009
Skew:	0.507	Prob(JB):	0.0496
Kurtosis:	3.466	Cond. No.	3.08e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.08e+08. This might indicate that there are strong multicollinearity or other numerical problems.

The model substituting historic conference is as follows:

OLS Regression Results

Dep. Variable:	PayPlusBonus	R-squared:	0.785
Model:	OLS	Adj. R-squared:	0.751
Method:	Least Squares	F-statistic:	23.26
Date:	Sun, 28 Apr 2019	Prob (F-statistic):	2.40e-22
Time:	03:08:51	Log-Likelihood:	-1465.6
No. Observations:	97	AIC:	2959.
Df Residuals:	83	BIC:	2995.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.04E+06	8.47E+05	-1.232	0.222	-2.73E+06	6.41E+05
Historic_Conference[T.Big 12]	4.04E+05	4.28E+05	0.943	0.348	-4.48E+05	1.26E+06
Historic_Conference[T.Big East]	-3.66E+05	4.57E+05	-0.802	0.425	-1.27E+06	5.42E+05
Historic_Conference[T.Mid-American East]	-1.09E+06	5.71E+05	-1.91	0.06	-2.23E+06	4.50E+04
Historic_Conference[T.Mid-American West]	-7.24E+05	5.50E+05	-1.317	0.191	-1.82E+06	3.69E+05
Historic_Conference[T.Mountain West]	-2.31E+05	4.85E+05	-0.477	0.635	-1.20E+06	7.34E+05
Historic_Conference[T.Pacific-10 Conference]	-7.66E+05	4.15E+05	-1.847	0.068	-1.59E+06	5.90E+04
Historic_Conference[T.SEC]	-3.52E+04	4.54E+05	-0.077	0.938	-9.38E+05	8.68E+05
Historic_Conference[T.Sun Belt]	-6.68E+05	5.05E+05	-1.323	0.19	-1.67E+06	3.37E+05
Historic_Conference[T.USA]	-9.01E+05	4.47E+05	-2.013	0.047	-1.79E+06	-1.09E+04
Historic_Conference[T.Western Athletic]	-8.82E+05	4.86E+05	-1.813	0.073	-1.85E+06	8.55E+04
FGR	3.25E+04	1.02E+04	3.194	0.002	1.23E+04	5.28E+04
FootballRevenue	0.0628	0.013	4.772	0	0.037	0.089
Capacity	22.143	9.025	2.454	0.016	4.193	40.093

```
==============================================================================
Omnibus:                        8.424   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.015   Jarque-Bera (JB):                8.152
Skew:                           0.609   Prob(JB):                       0.0170
Kurtosis:                       3.729   Cond. No.                     2.80e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.8e+08. This might indicate that there
are strong multicollinearity or other numerical problems.
```

Results

Based on the results from the final model with the current conferences, it is worthwhile for coaches to work with the school to increase graduation rate, as every 1% increase in FGR results in them receiving approximately $32,500 in additional compensation.

However, this bump is minuscule compared to the effect of different conferences, playing in the Big 12 increases a coach's salary $1.3 Million however playing in an independent conference can decrease it $980,000 or playing in the USA Conference, decreases a coach's salary $438,000.

Based on these two tables, with Syracuse currently a member of the ACC, the coach should be paid: $3.2 Million per the below:

	Coefficient	Multiplier	Resulting Pay
Intercept	-6.63E+05	1	$ (662,900.00)
Conference[T.ACC]	9.82E+05	1	$ 981,900.00
Conference[T.Big 12]	1.34E+06	0	$ -
Conference[T.Big Ten]	1.18E+06	0	$ -
Conference[T.C-USA]	-4.38E+05	0	$ -
Conference[T.Ind.]	-9.81E+05	0	$ -
Conference[T.MAC]	-4.17E+05	0	$ -
Conference[T.Mt. West]	-3.14E+05	0	$ -
Conference[T.Pac-12]	3.67E+05	0	$ -
Conference[T.SEC]	1.18E+06	0	$ -
Conference[T.Sun Belt]	-3.44E+05	0	$ -
FGR	1.84E+04	64	$ 1,176,320.00
FootballRevenue	0.0438	$14,866,061	$ 651,133.47
Capacity	21.3845	49,250	$ 1,053,186.63
		Total Pay:	**$ 3,199,640.10**

This suggests the coach should receive a 33% raise over the salary reported in the original dataset.

Moving to the Big 10 from the ACC would push the recommended salary to $3.55 Million, or a 48% bump over his current salary:

	Coefficient	Multiplier	Resulting Pay
Intercept	-6.63E+05	1	$ (662,900.00)
Conference[T.ACC]	9.82E+05	0	$ -
Conference[T.Big 12]	1.34E+06	1	$ 1,337,000.00
Conference[T.Big Ten]	1.18E+06	0	$ -
Conference[T.C-USA]	-4.38E+05	0	$ -
Conference[T.Ind.]	-9.81E+05	0	$ -
Conference[T.MAC]	-4.17E+05	0	$ -
Conference[T.Mt. West]	-3.14E+05	0	$ -
Conference[T.Pac-12]	3.67E+05	0	$ -
Conference[T.SEC]	1.18E+06	0	$ -
Conference[T.Sun Belt]	-3.44E+05	0	$ -
FGR	1.84E+04	64	$ 1,176,320.00
FootballRevenue	0.0438	$14,866,061	$ 651,133.47
Capacity	21.3845	49,250	$ 1,053,186.63
		Total Pay:	**$ 3,554,740.10**

Using the historic conferences, the coach should make $2.7 Million:

	Coefficient	Multiplier	Resulting Pay
Intercept	$ (1,043,000.00)	1	$ (1,043,000.00)
Historic_Conference [T.Big 12]	$403,900.00	0	$ -
Historic_Conference [T.Big East]	$ (366,000.00)	1	$(366,000.00)
Historic_Conference [T.Mid-American East]	$(1,091,000.00)	0	$ -
Historic_Conference [T.Mid-American West]	$(724,200.00)	0	$ -
Historic_Conference [T.Mountain West]	$(231,400.00)	0	$ -
Historic_Conference [T.Pacific-10 Conference]	$ (765,700.00)	0	$ -
Historic_Conference [T.SEC]	$ (35,180.00)	0	$ -
Historic_Conference [T.Sun Belt]	$ (668,000.00)	0	$ -
Historic_Conference [T.USA]	$ (900,600.00)	0	$ -
Historic_Conference [T.Western Athletic]	$ (881,900.00)	0	$ -
FGR	$ 32,530.00	64	$2,081,920.00
FootballRevenue	$0.06	$ 14,866,061	$ 933,588.63
Capacity	$22.14	49250	$ 1,090,542.75
		Total Pay:	**$2,697,051.38**

This suggests the coach should receive a 12% raise over the salary reported in the original dataset.

Conclusion

Overall, this model is adequate, it explains a over 75% of the variability in college football coaches' salaries. As described above, there is multicollinearity associated with this output due to the conference system grouping schools with similar profiles together. Removing variables to

remove the artificial inflation of the Adjusted R-Squared resulted in a model that was able to predict 66.9% of the variability in coaches' pay.

Additional variables that could increase our predictive power include a five year look back at team performance as one year's worth of records likely doesn't tell us how the school is viewing the team's value to their overall goals. Additionally, looking at the coach's record may be more valuable than the teams, as the average tenure of a college football coach is only 3.8 years (Gaines & Nudelman, 2017).

Based on our model, the current Syracuse football coach is underpaid by 33% and should be making $3,199,640.10. However, before giving Dino Babers this raise, Syracuse may want to consider using some of this money to lobby the NCAA to create a stipend for the college athletes whose hard work generated the $43.7 Million in Total Sports Revenue the school received in 2006.

Works Cited

Farmer, A., & Pecorino, P. (2010). Is the Coach Paid too Much?: Coaching Salaries and the NCAA Cartel. Journal of Economics & Management Strategy, 19(3), 841–862. https://doi-org.libezproxy2.syr.edu/10.1111/j.1530-9134.2010.00271.x

Fox, Jon (2019). GitHub: https://github.com/2SUBDA/IST_718/Coaches9.csv

Gaines, C., & Nudelman, M. (2017, Dec 6). Most college football players will be forced to change head coaches at least once in their career. *Business Insider* Retrieved from https://www.businessinsider.com/college-football-players-coaches-recruiting-2017-12

GitHub 1: https://github.com/rmcelfresh/IST718Lab3/GSR.csv

GitHub 2: https://github.com/rmcelfresh/IST718Lab3/Record.csv

GitHub 3: https://github.com/rmcelfresh/IST718Lab3/FootballRevenue.csv

List of NCAA Division I FBS football stadiums. (2019). *Wikipedia* Retried From https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums

NCAA 2006 Football Season (2006). *Google* Retrieved from https://www.google.com/search?q=NCAA+2006+Football+Season&oq=NCAA+2006+Football+Season&aqs=chrome..69i57j69i60j0l4.8698j1j4&sourceid=chrome&ie=UTF-8#sie=lg;/m/0bwyv6;6;/m/012hfxch;st;fp;1;;

NCAA should pay student athletes. (2016, Apr 27). *University Wire* Retrieved from https://search-proquest-com.libezproxy2.syr.edu/docview/1784655849?accountid=14214

Why the GSR is a Better Methodology. (2014, Oct 28). *NCAA* Retrieved from http://www.ncaa.org/about/resources/research/why-gsr-better-methodology