Andrew Alford
7/22/2023
IST 718 Big Data Analytics

Lab 1 Report

In this report, I do a deep dive analysis into college football coach salaries within the United States. Throughout the report, I utilize the OSEMIN method in order to follow a logical process of completing the project. The format of the report will likewise follow that of the method mentioned above, in that I will describe in detail the different aspects of the method one by one. I want to give a special thanks to Bill Steel for helping me with compiling the multiple CSV files, Ryan Summers for helping me understand some of the statistical principles behind the modeling, and naturally many online websites such as stackoverflow and geeksforgeeks for helping me troubleshoot coding issues.

Obtain

The first step in the project was to obtain all the necessary data and compile it into one complete data frame, upon which one could become more familiar with it and start doing analysis on it. To start, I downloaded the CSV file "Coaches9.csv" and read it into a jupyter notebook. The completely unscrubbed data frame looked like this upon reading it in:

	School	Conference	Coach	SchoolPay	TotalPay	Bonus	BonusPaid	AssistantPay	Buyout
0	Air Force	Mt. West	Troy Calhoun	885000	885000	247000	--	$0	--
1	Akron	MAC	Terry Bowden	$411,000	$412,500	$225,000	$50,000	$0	$688,500
2	Alabama	SEC	Nick Saban	$8,307,000	$8,307,000	$1,100,000	$500,000	$0	$33,600,000
3	Alabama at Birmingham	C-USA	Bill Clark	$900,000	$900,000	$950,000	$165,471	$0	$3,847,500
4	Appalachian State	Sun Belt	Scott Satterfield	$712,500	$712,500	$295,000	$145,000	$0	$2,160,417
...	...	...	...	...	...	...	...	...	...
124	West Virginia	Big 12	Dana Holgorsen	$3,605,000	$3,617,500	$1,000,000	$90,000	$0	$7,150,000
125	Western Kentucky	C-USA	Mike Sanford Jr.	$800,000	$805,850	$400,000	$0	$0	$1,200,000
126	Western Michigan	MAC	Tim Lester	$800,000	$800,000	$346,500	$39,250	$0	$800,000
127	Wisconsin	Big Ten	Paul Chryst	$3,750,000	$3,750,000	--	$290,000	$0	$6,000,000
128	Wyoming	Mt. West	Craig Bohl	$1,412,000	$1,412,000	$450,000	$236,000	$0	$8,016,667

129 rows × 9 columns

Scrub

After some time considering what aspects of the data frame needed to be clean, I settled on removing the commas and dollar signs to make numeric analysis easier in the future. Additionally, instead of a "--" string representing a null value, I replaced all of them with NaN values. I noticed that there were four schools which did not have any data within them, so I removed those four schools, which were Baylor, Brigham Young, Rice, and Southern Methodist. I then used some quick string formatting to get a feeling for how many missing values were in the frame. I was happy to see that most of the columns did not have any missing values aside from three, Bonus, BonusPaid, and Buyout, that had 17, 36, and 17 missing values respectively. Lastly, for this first data set, I converted all numeric values (all columns except for School and Conference) to floats or integers so that future statistical analysis would be possible.

Next, in accordance with the assignment guidelines, I started combining other data sets to the coaches data frame. In total I added five additional data sets: StadiumSize, GSR and FGR, Revenue, Wins and Losses, and Coordinates. The stadium size added a column that displayed each school's stadium capacity, the GSR and FGR columns looked at graduation rates, revenue looked at each school's total revenue, the wins and losses displayed each school's football team's wins and losses, and lastly the coordinates showed the geographic location of each school.

Finding the correct information was the most difficult part of the obtain section of the project, as only a combination of data scrubbing, excel magic, and manual input was successful in getting these five complete data sets. Luckily, I had a colleague, Bill Steel, who was an immense help in finding and cleaning these additional tables. Adding the data sets within the Jupyter notebook was very easy, all that was needed was to read in the csv, add the required column to the main data frame, and make sure that its values were the proper numeric type. Upon adding all these five data sets, the main data frame now looked like this:

Coach	SchoolPay	TotalPay	Bonus	BonusPaid	AssistantPay	Buyout	StadiumSize	GSR	FGR	Revenue	Wins	Losses	Latitude	Longitude
Terry Bowden	411000	412500	225000.0	50000.0	0	688500.0	30000	74	76.0	12354872.0	528	583	41.072570	-81.508384
Nick Saban	8307000	8307000	1100000.0	500000.0	0	33600000.0	101821	89	66.0	140831439.0	953	335	33.207490	-87.550392
Scott Satterfield	712500	712500	295000.0	145000.0	0	2160417.0	24150	81	57.0	194500000.0	655	349	36.211515	-81.685506
Blake Anderson	825000	825000	185000.0	25000.0	0	300000.0	30964	75	52.0	8593341.0	490	518	35.848990	-90.667695
Gus Malzahn	6700000	6705656	1400000.0	375000.0	0	32143750.0	87451	83	65.0	128960499.0	793	464	32.602362	-85.488911
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Mike Leach	3500000	3500000	725000.0	75000.0	0	4900000.0	35117	86	57.0	51125157.0	567	579	46.731968	-117.160586
Dana Holgorsen	3605000	3617500	1000000.0	90000.0	0	7150000.0	60000	83	53.0	34050353.0	772	522	39.652220	-79.955175
Mike Sanford Jr.	800000	805850	400000.0	0.0	0	1200000.0	22113	75	65.0	13764592.0	607	421	36.984877	-86.459014
Tim Lester	800000	800000	346500.0	39250.0	0	800000.0	30200	74	40.0	12863908.0	593	473	42.285755	-85.601004
Craig Bohl	1412000	1412000	450000.0	236000.0	0	8016667.0	29181	90	53.0	20381998.0	556	595	41.311936	-105.569065

Explore

In order to explore this data, I started with some relatively rudimentary techniques for seeing what I was working with. To start, I did a simple string format check to see how many rows there were in the data frame.

```
There are 125 unique schools in the coaches data set
```

Next, knowing that the conference column would play a key role in future analysis, I wanted to check and see how many unique conferences there were.

```
There are 11 unique conferences in the coaches data set, they are: ['Mt. West', 'MAC', 'SEC', 'C-USA', 'Sun Belt', 'Pac-12', 'I
nd.', 'ACC', 'AAC', 'Big Ten', 'Big 12']
```

From there, simply to get a feel for the numeric columns, I did a .describe() method for each numeric column in the data frame. The results were as follows:

```
count    1.250000e+02
mean     2.410301e+06
std      1.881377e+06
min      3.900000e+05
25%      8.015040e+05
50%      1.831580e+06
75%      3.605000e+06
max      8.307000e+06
Name: SchoolPay, dtype: float64
```

```
count    1.250000e+02
mean     2.417061e+06
std      1.885752e+06
min      3.900000e+05
25%      8.058500e+05
50%      1.900008e+06
75%      3.617500e+06
max      8.307000e+06
Name: TotalPay, dtype: float64
```

```
count    1.080000e+02
mean     8.690469e+05
std      6.339712e+05
min      5.000000e+04
25%      3.915000e+05
50%      7.700000e+05
75%      1.150000e+06
max      3.100000e+06
Name: Bonus, dtype: float64
```

```
count    8.900000e+01
mean     1.495296e+05
std      2.373974e+05
min      0.000000e+00
25%      2.000000e+04
50%      6.500000e+04
75%      1.800000e+05
max      1.350000e+06
Name: BonusPaid, dtype: float64
```

```
count    125.0
mean       0.0
std        0.0
min        0.0
25%        0.0
50%        0.0
75%        0.0
max        0.0
Name: AssistantPay, dtype: float64
```

```
count    1.080000e+02
mean     8.136523e+06
std      1.041392e+07
min      0.000000e+00
25%      1.200000e+06
50%      4.018758e+06
75%      1.070750e+07
max      6.812500e+07
Name: Buyout, dtype: float64
```

```
count       125.000000
mean      52518.928000
std       22953.985867
min       15000.000000
25%       30964.000000
50%       50000.000000
75%       65500.000000
max      107601.000000
Name: StadiumSize, dtype: float64
```

```
count    125.000000
mean      81.448000
std        8.611186
min       54.000000
25%       75.000000
50%       82.000000
75%       88.000000
max       97.000000
Name: GSR, dtype: float64
```

```
count    122.000000
mean      62.786885
std       10.116862
min       30.000000
25%       57.000000
50%       64.000000
75%       69.000000
max       90.000000
Name: FGR, dtype: float64
```

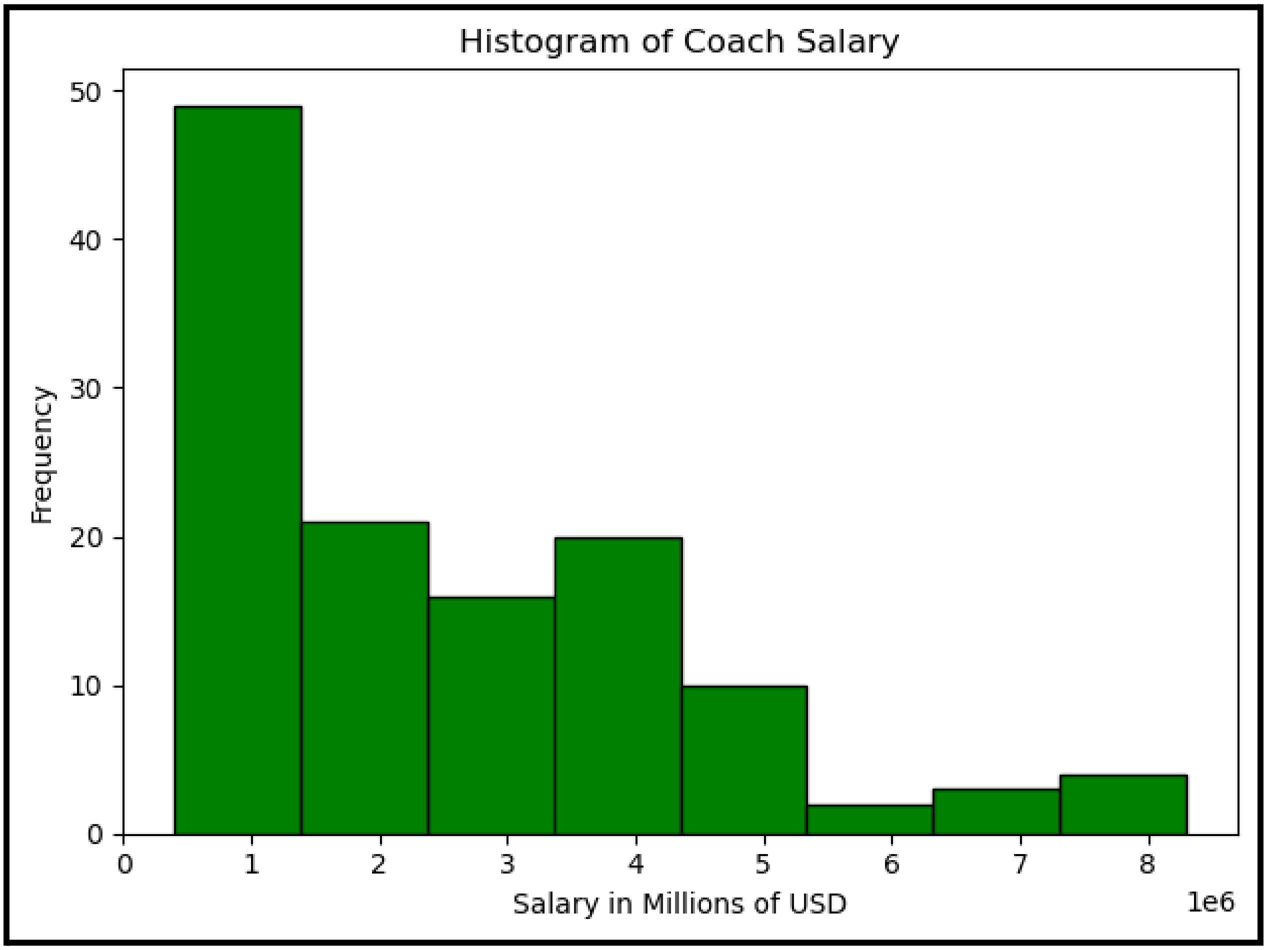
```
count    1.100000e+02
mean     5.188111e+07
std      4.072004e+07
min      6.682465e+06
25%      1.502958e+07
50%      4.777673e+07
75%      7.175769e+07
max      1.945000e+08
Name: Revenue, dtype: float64
```

```
count    125.000000
mean     587.056000
std      201.284236
min       42.000000
25%      523.000000
50%      604.000000
75%      711.000000
max      989.000000
Name: Wins, dtype: float64
```

```
count    125.000000
mean     475.688000
std      150.150485
min       71.000000
25%      420.000000
50%      498.000000
75%      583.000000
max      704.000000
Name: Losses, dtype: float64
```

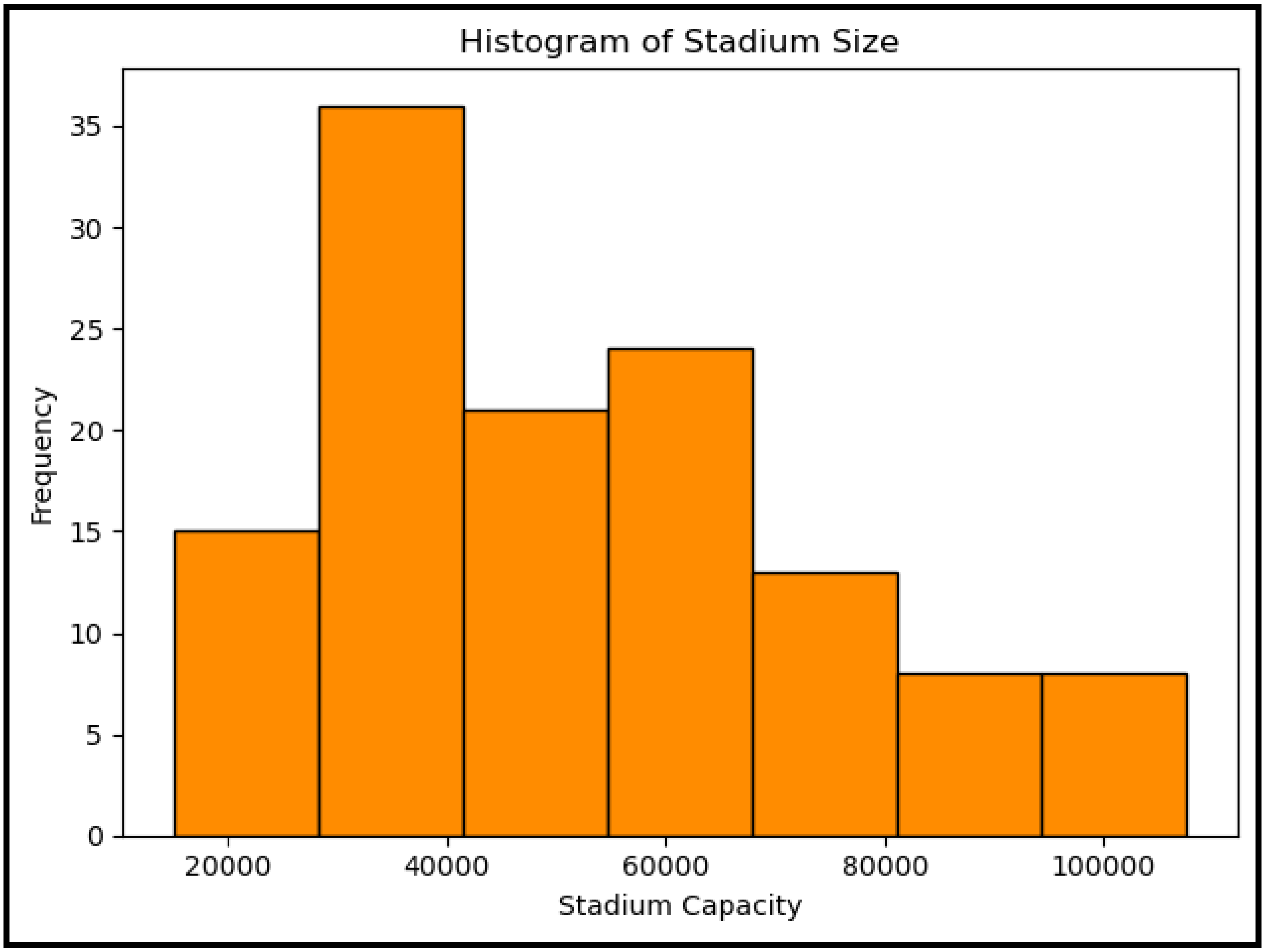
Of course, none of this is actual analysis on the data itself, but this aspect of the project rather serves as a point of reference that I could come back to in future. As it turned out, this ended up being invaluable largely because this information was needed many multiple times during the model and interpretation portions of the project.

At this point, I started to create some visualizations that would start to tell a story about the shape of the data and some insights that could come from it. The first step towards this was to create some basic histograms on coach salary, stadium size, and revenue.



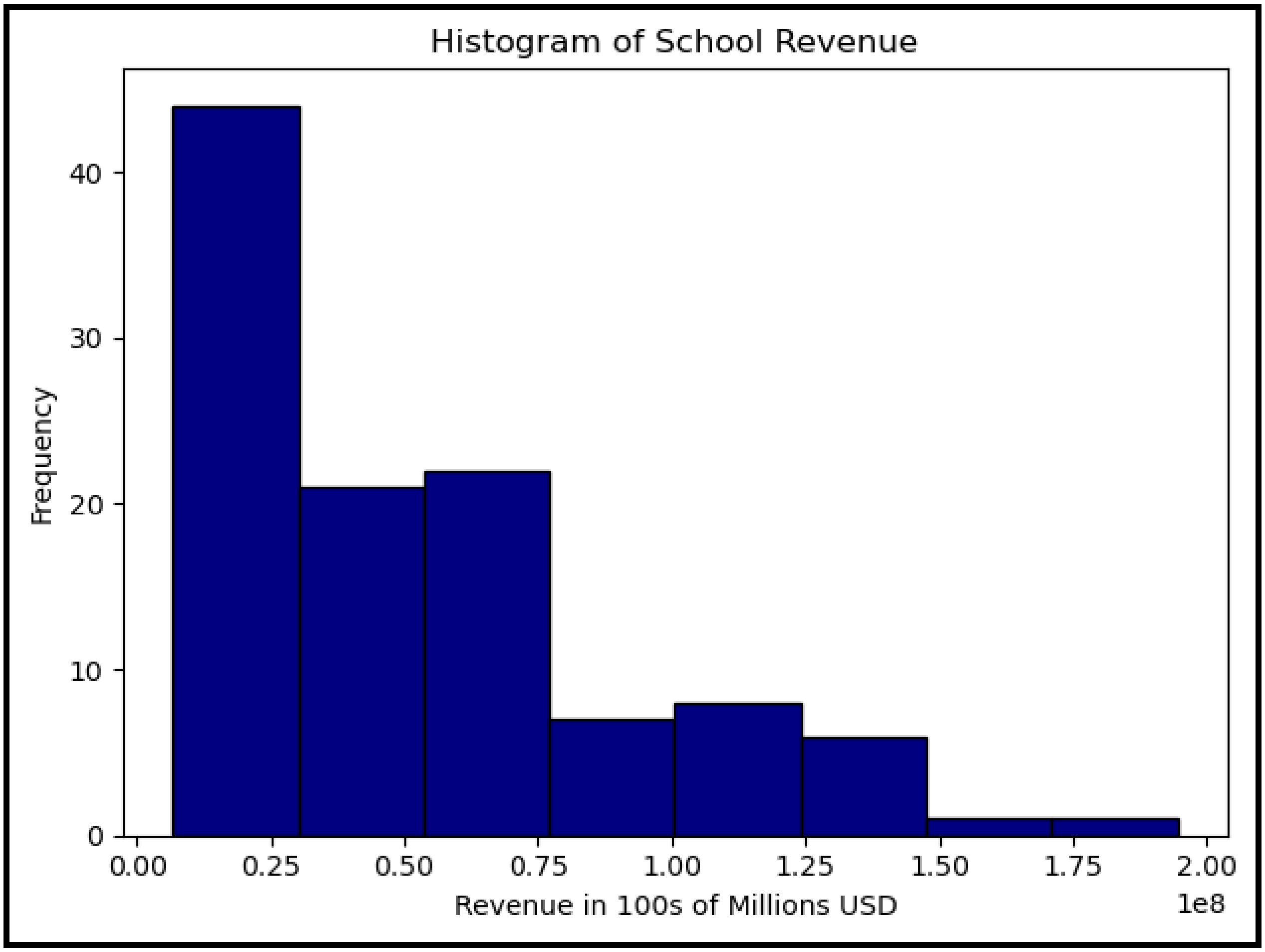
This plot demonstrates that there can be quite a good bit of variation in coach salary, with a large percentage of coach salaries hovering above and below one million. Clearly, however, the vast majority of coach salaries rest somewhere between $500,000 and $5,250,000, with only a minority of salaries going above that range.

The next histogram looks at the distribution of stadium sizes.



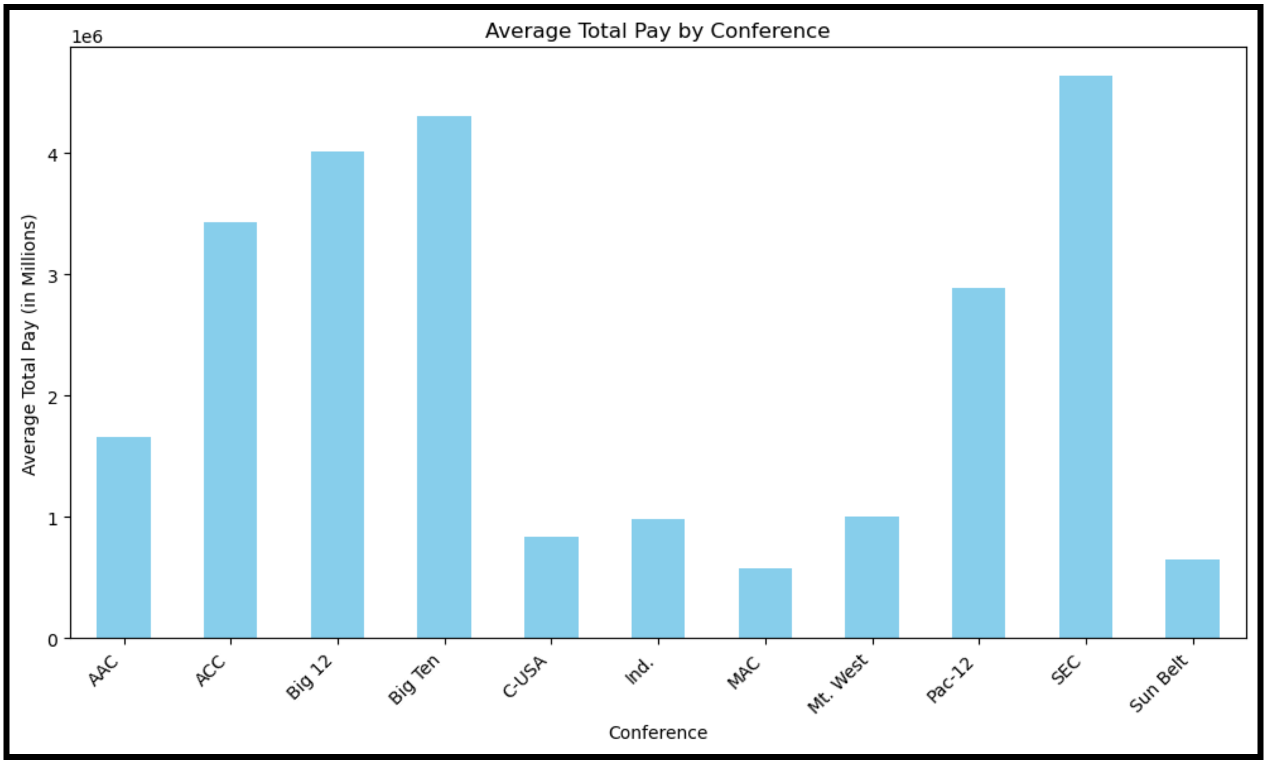
The plot for stadium sizes generally fits a normal distribution, starting at around the 20,000 mark and tailing out around above the 100,000 mark. With the average stadium size standing at 52,518, stadiums skew more towards the smaller side of the range rather than the larger side.

Looking at revenue, it paints a slightly different story. Each school's revenue, which is measured in the hundreds of millions of USD, starts off high in the left side of the plot and slowly tapers off as it moves to the right.



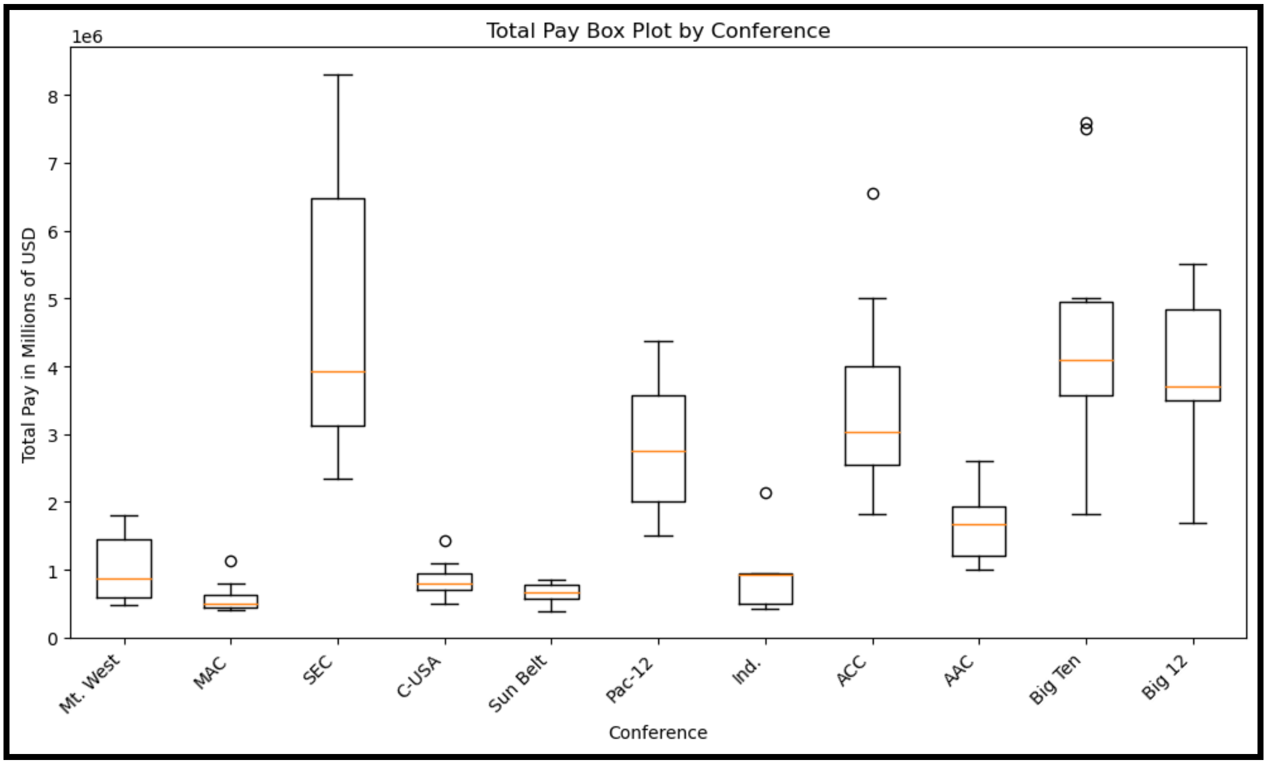
Clearly, most of the universities shown here have football team revenue sizes between $100,000 and $750,000, with only a handful going above $1,500,000.

After doing this bit of exploring with histograms, I decided that a different approach would be useful when looking at more categorical data. Knowing that the football conference played a big part in coach salary, I decided to make a bar chart plotting average coach salary by conference.



In this chart, there are five conferences that have demonstrably higher salaries compared to the others. SEC, Big Ten, Big 12, ACC, and Pac-12 have significantly higher salaries than Mt. West, Ind., C-USA, Sun Belt, or MAC. This information becomes very valuable in the future when analyzing coach salary.

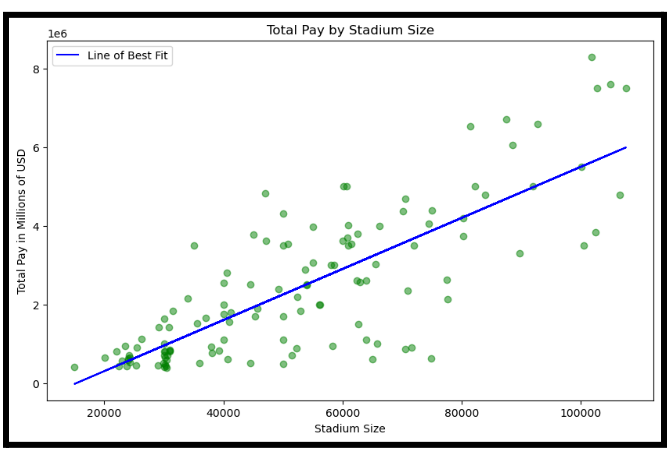
However, a simple bar chart is not the best visualization for coach salary as there exists significant variation within each conference, as well as serious outliers.



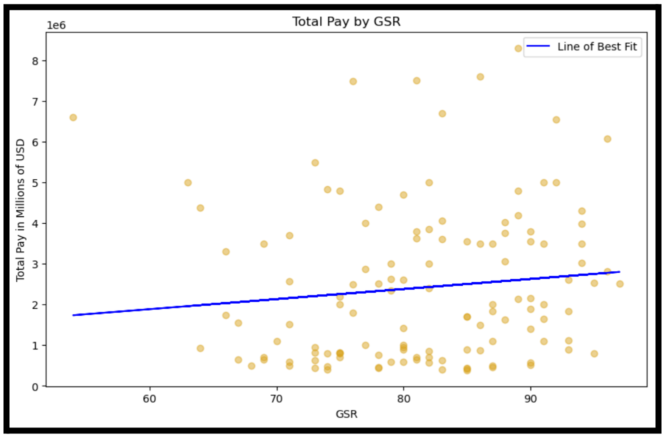
This box and whiskers plot has some very interesting insights. Namely that the conferences with the higher salaries also have greater variation in their salaries, as well as some massive outliers within the ACC and Big Ten. Inversely, the less expensive conferences have much less variation and are, in fact, quite rigid in their salary structure.

After looking at some of the interesting insights from the histograms and conference breakdowns, I did a number of scatter plots demonstrating TotalPay vs different factors.

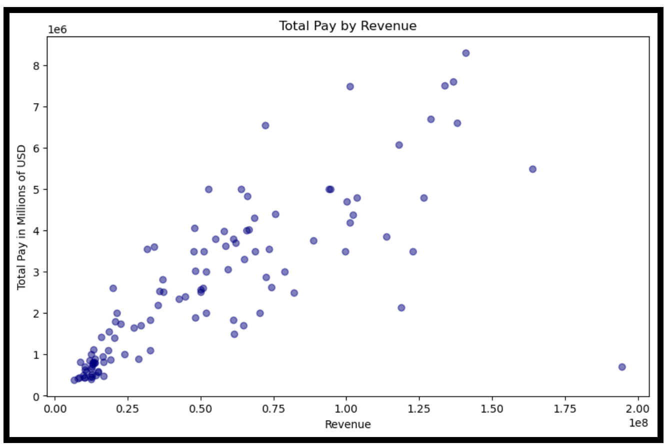
Total Pay by Stadium Size



Total Pay by GSR

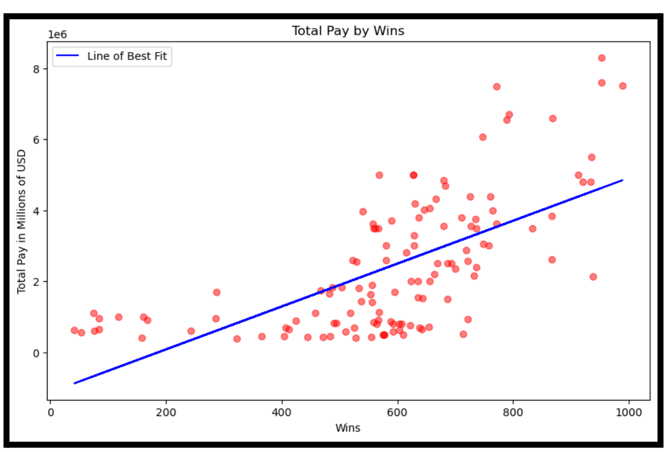


Total Pay by Revenue

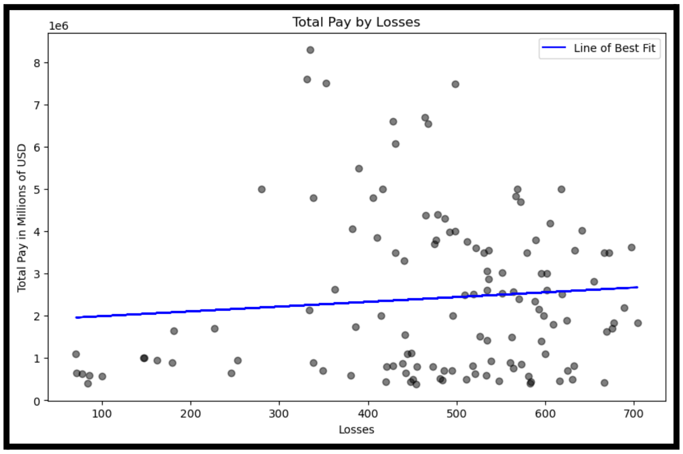


*Unfortunately, I could not find a mathematical way to compute a line of best fit here.

Total Pay by Wins



Total Pay by Losses



Looking at all these factors, we can have a rough outline of which factors have a stronger linear correlation to coach salary, and which ones don't. As was demonstrated in the week 2 case study, stadium size plays a significant role in the coach's salary, with a strong linear relationship. What I found from analysis in my own data set is that revenue and wins also have a strong linear correlation, whereas GSR and Losses do not have a significant relationship to coach salary.

Model

For this project, I used a simple linear regression to predict Syracuse University's football coach salary. In truth, this is the first model that I've ever created, so it took a lot of trial and error to actually get something that worked, and even more trial and error to get something that approached what could be considered a good model. In the end, the highest adjusted R-squared value I was able to achieve was 0.80, which is certainly not bad, but is nonetheless a far shot from the gold standard of 0.95.

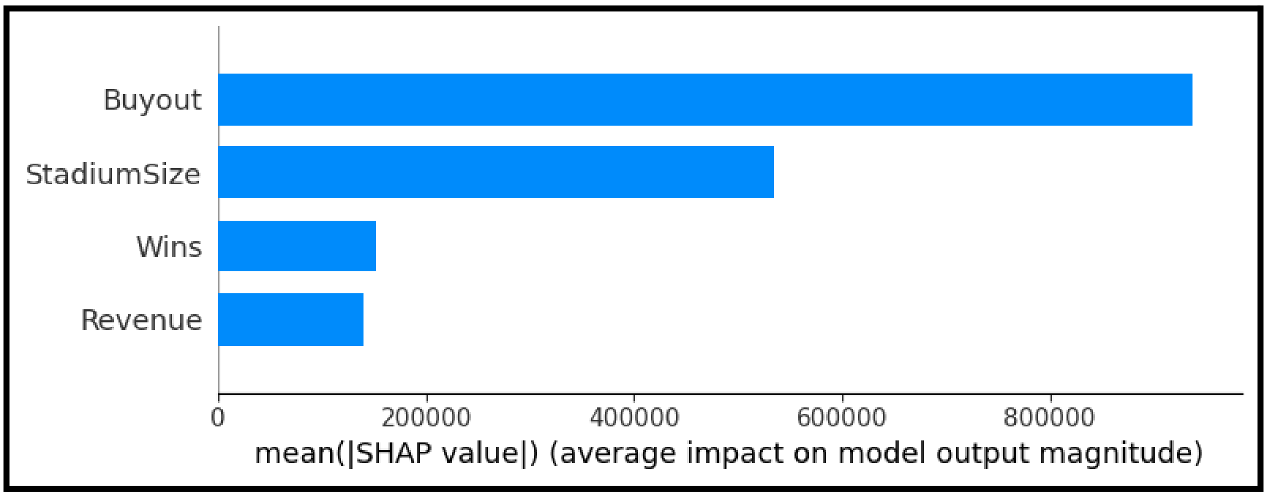
Methodology

Unlike the strategy that some of my fellow students took, I decided to use a popular package called scikit-learn to create my linear model. The first step was to drop all rows that contained any NaN values because it causes an error when doing predictive analysis. In the end, dropping all rows that contained NaN values left me with 80 rows. After this, I split my data set up into predictor and response variables. To start, I put all numeric columns in the X (predictor) variable, and TotalPay in the y (response) variable. I then separated them into test and training sets in accordance with a standard linear regression procedure, with 20% going into the test set and 80% going into the train set. I called my linear regression function and fit the training model. When my results weren't satisfactory, I used the resource that we discussed in class called SHAP to visualize which predictor variables had a better value compared to their peers.

This turned out to be invaluable because I simply removed the predictor variables that did not matter and ran the model again, which in turn produced a much better result.

Results

Initially, once I used the model.predict() method, I got a pretty rough 0.62 adjusted R-squared value. As mentioned above, I used the SHAP package to get a visual idea of which ones weren't adding to the value of the model. After looking at the results, I decided to only keep Buyout, StadiumSize, Wins, and Revenue, which were clearly the ones with the strongest linear relationship. After keeping only these values, I landed on a satisfactory 0.80 adjusted R-squared value. Here is a breakdown of the value assigned to the predictor variables using the SHAP package:



Report Questions

What is the recommended salary for the Syracuse football coach?

To find the recommended salary for the Syracuse football coach, I first made a dictionary with all the data available for the instance at hand. I converted the dictionary into a data frame and made a prediction on it using the trained model created above. Using some quick string formatting, I made a print statement showing the recommended salary as $2,689,767.20.

```python
#Finding recommended Syracuse coach salary

syracuse_data = {
    'Buyout': [10000000],
    'StadiumSize': [49250],
    'Revenue': [44613716],
    'Wins': [737],
}

# Converting dictionary into data frame
syracuse_df = pd.DataFrame(syracuse_data)

# Making recommendation
predicted_salary_syracuse = model.predict(syracuse_df)

print(f"Recommended Salary for Syracuse: ${predicted_salary_syracuse[0]:,.2f}")
```

```
Recommended Salary for Syracuse: $2,689,767.20
```

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

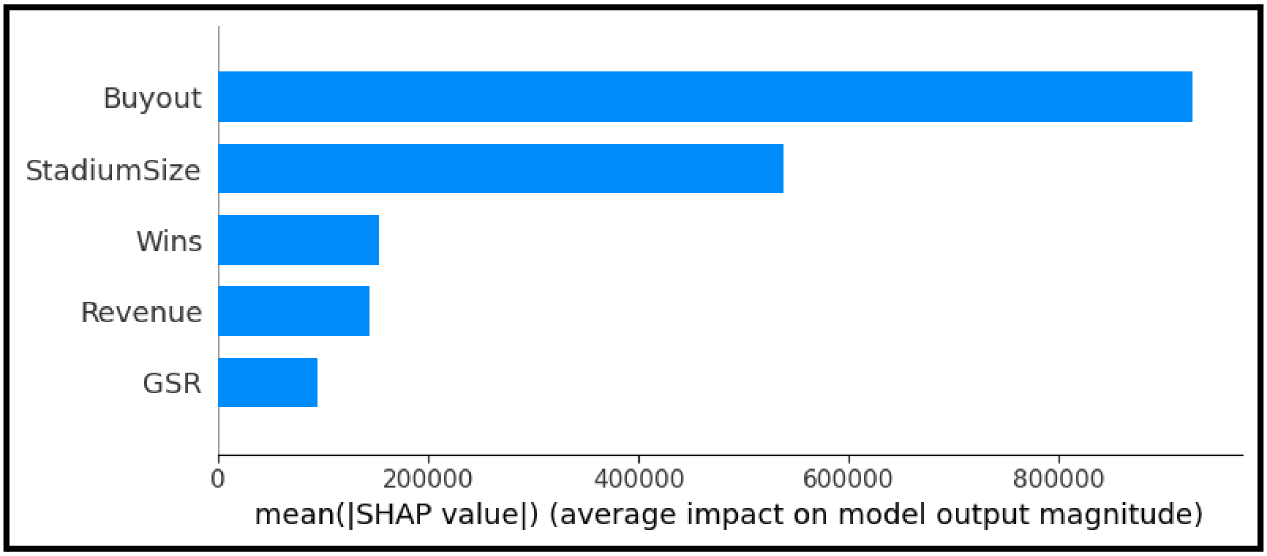
In order to find out what his salary would be if we were in a different conference, I used quite a simple technique to adjust the salary accordingly. In all honesty, I'm not entirely sure if this was the correct method that we were expected to use, but it's the one that I came up with and I believe put us in the correct ball park at the very least. To calculate this, I simply took the average coach salary for the Big East and the Big Ten and converted them to percentages higher or lower compared to the current salary of the Syracuse university coach. The average salary for the Big East was 56.25% smaller than Syracuse's current salary, making the salary of the Syracuse University's coach if we were still in the Big East $1,350,678.38. Applying the same methodology for the Big Ten, which was 19.66% larger, Syracuse University's coach salary if we were in the Big Ten would be $2,873,283.10.

<u>What schools did we drop from our data and why?</u>

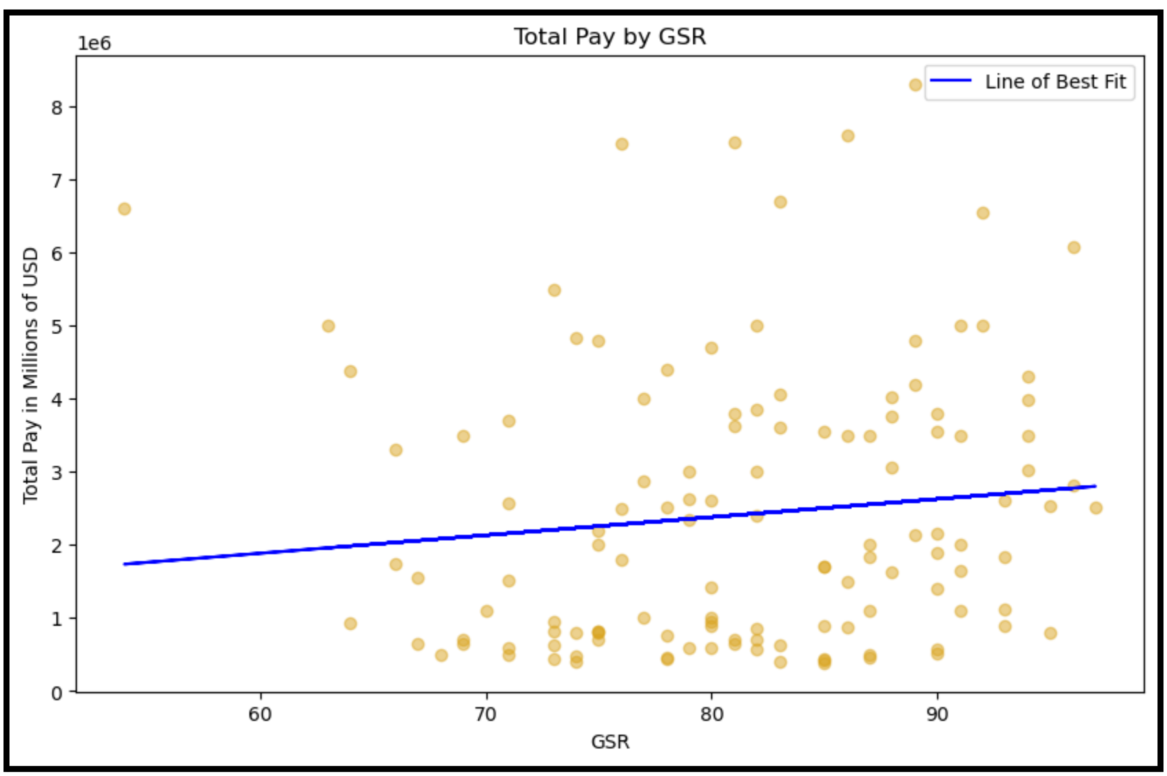
In the end, I dropped four schools from our data set largely because they did not contain any data and were therefore not adding anything to the usefulness of the data. After doing some research with my colleague Bill Steel, we found out that the data was unavailable because they were private universities. However, the data was still available if one went digging deep enough. I decided to not pursue this route partially because I wanted some experience with working with a data set where I had to account for taking some rows out. Judging from the fact that I only took out four rows, I do not believe that it had a significant impact on the final result of my project.

<u>What effect does graduation rate have on the project salary?</u>

When I was building my model, I input GSR as one of the initial factors that the model considered. However, after running all the predictor variables through the SHAP visualization tool, it showed that GSR did not have a significant rate on coach salary.



Additionally, when I did a scatter plot of coach salary by GSR, the line of best fit was almost flat, demonstrating that it did not have a strong effect on the coach salary.



How good is our model?

Overall, for my first model, I would say that it is generally a pretty good model. At its best, the model accounts for 80% of the variance within the coaches salary. Knowing that the gold standard within the industry is 95%, it's not the best that it could possibly be. Nonetheless, I'm satisfied that I was able to improve the model using trial and error as well as start to formulate an understanding of what goes on behind the scenes to make models.

```
#Finding the MSE and R-Squared values

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse:.2f}")
print(f"R-squared: {r2:.2f}")
```

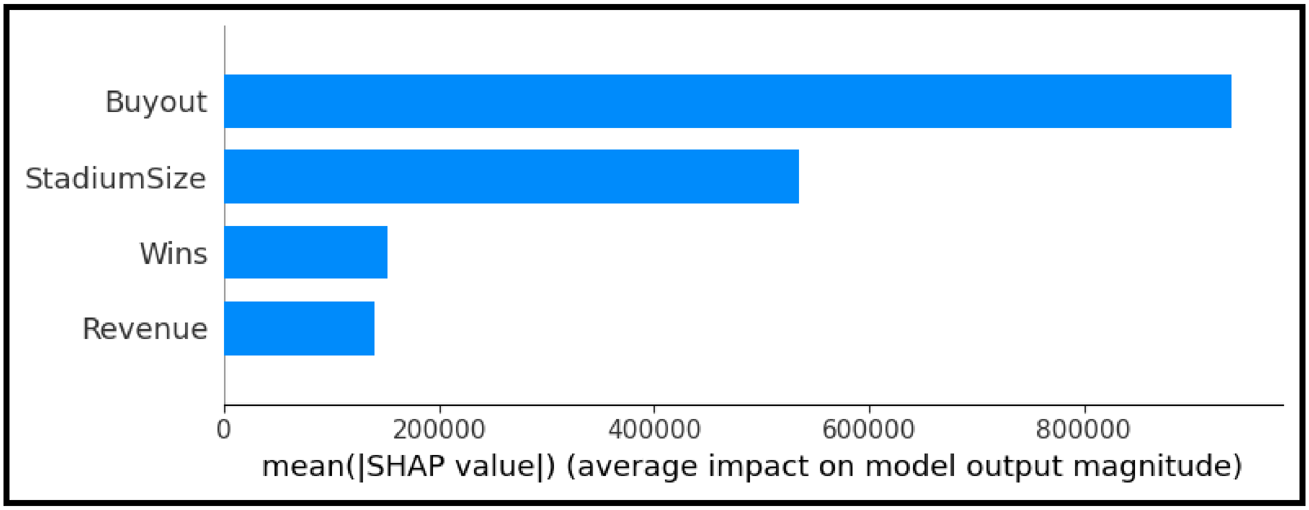
```
Mean Squared Error: 561394329365.90
R-squared: 0.85
```

```
#Adjusted R-squared value

n = len(y_test)
k = X_test.shape[1]

# Calculate the regular R-squared
r2 = r2_score(y_test, y_pred)

# Calculate the adjusted R-squared
adjusted_r2 = 1 - (1 - r2) * (n - 1) / (n - k - 1)

print(f"Adjusted R-squared: {adjusted_r2:.2f}")
```

```
Adjusted R-squared: 0.80
```

What is the single biggest impact on salary size?

Without a doubt, Buyout played far and away the most significant impact on salary size. As shown on the SHAP visualization, Buyout had almost twice the mean average impact on model output magnitude compared to the second highest value of StadiumSize. This was followed by a distant third and fourth of Wins and Revenue respectively. This plot shown below really highlights the significance Buyout, and StadiumSize as well, has on model performance.


Buyout
StadiumSize
Wins
Revenue
0
200000
400000
600000
800000
mean(|SHAP value|) (average impact on model output magnitude)