Melissa Mosier

4/30/2023

## Obtain

I knew I would want to create a key for the Schools and Conferences in order to link the files together. I used the Coaches file provided on S2U to match college names using an INDEX(array,MATCH(“\*”&A2&”\*”, array, 0)) code in Excel before uploading. There are other ways to do this, but I wanted to have eyes on the data in Excel before importing to Python where it can be difficult to see the finer details in the data for cleansing. Any major adjustments could be made later in Python, but I wanted to be sure the correct schools were matching up first.

Here are the files I used and their column information:

**File Name: Coaches**

|  |  |  |
| --- | --- | --- |
| Column Name | Explanation | Example |
| School\_Coaches | Categorical. | Houston |
| **SchoolName\_ncaadata** | Categorical. Derived.  School names found in ncaa\_data file using index(match()) in Excel; for join function | University of Houston |
| **SchoolName\_footballDL** | Categorical. Derived.  School names found in footballDL file using index(match()) in Excel; for join function | University of Houston |
| **SchoolName\_stadiums** | Categorical. Derived.  School names found in stadiums file using index(match()) in Excel; for join function | Houston |
| **SchoolName\_WinsLosses** | Categorical. Derived.  School names found in WinsLosses file using index(match()) in Excel; for join function | Houston |
| ConfCode | Categorical. Conference Code, abbreviation of Conference name in the original dataset | AAC |
| Conference | Categorical. Derived.  Matched to full-name of conference, as found in other files like ncaa\_data | American Athletic Conference |
| Coach | Coach name: First Name Last Name | Major Applewhite |
| SchoolPay | Numeric. Text | $1,750,000 |
| TotalPay | Numeric. Text | $1,750,000 |
| Bonus | Numeric. Text | $782,500 |
| BonusPaid | Numeric. Text | $20,000 |
| Buyout | Numeric. Text | $2,112,500 |

Notes:

* Source: Provided on S2U.
* Deleted AssistantPay column because all values were 0.

**File Name: ncaa\_data**

|  |  |  |
| --- | --- | --- |
| Column Name | Explanation | Example |
| SchoolName\_ncaadata | Renamed from: School  School Name | Alabama A&M University |
| Conference | Full name of the conference | Southwestern Athletic Conf. |
| State | State | AL |
| GSR | Numeric. Graduation Success Rate.  Accounts for transfers in/out. | 66 |
| FGR | Numeric. Federal Graduation Rate.  Measures the percentage of students who complete a BA/BS from their initial school (transfer = failure). | 54 |

Notes:

* Source: [GSR FGR Scores by School and Conference - 2021 - 2022 Season](https://web3.ncaa.org/aprsearch/gsrsearch)
  + ncaa\_data: ncaa\_data.xlsx
* [Click here for more information on calculating NCAA graduation rates.](https://ncaaorg.s3.amazonaws.com/research/gradrates/RES_HowGradRateCalculated.pdf)
* Deleted columns for Cohort Year (all entries were 2015) and Sport (all entries were Football).

**File Name: WinsLosses**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Explanation** | **Example** |
| Conference | Conference Code, similar to what was found in coaches file. | C-USA |
| ConfSUB | Additional conference information. | East |
| SchoolName\_WinsLosses | Renamed from: School  School Name | Western Kentucky |
| OverallW | Numeric. Wins in all games | 12 |
| OverallL | Numeric. Losses in all games | 2 |
| OverallPCT | Numeric. Percent Wins / Total  = (OverallW) / (OverallW + OverallL) | 0.857 |
| ConfW | Numeric. Wins in all games within Conference | 8 |
| ConfL | Numeric. Losses in all games within Conference | 0 |
| ConfPCT | Numeric. Percent Wins / Total Conferences  = (ConfW) / (ConfW + ConfL) | 1 |
| PPGOff | Numeric. Offense | 44.3 |
| PPGDef | Numeric. Defense | 25.9 |
| SRS | Numeric. Simple Rating System – Team rating | 9.62 |
| SOS | Numeric. Strength of Schedule – Combined winning percentages of opponents played | -4.02 |

Notes:

* Source: [SRCFB Website](https://www.sports-reference.com/cfb/years/2015.html)
* Deleted column for Year (all entries were 2015).

**File Name: stadiums**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Explanation** | **Example** |
| Stadium |  | Michigan Stadium |
| City |  | Ann Arbor |
| State |  | MI |
| SchoolName\_stadium | Renamed from: Team | Michigan |
| Conference |  | Big Ten |
| Capacity | Numeric. | 107601 |
| Built | Year. | 1927 |
| Expanded | Year. | 2015 |
| Div | Delete me. | fbs |
| Latitude | Float. | 42.26587 |
| Longitude | Float. | -83.7487 |

Notes:

* Source: [gboeing Github page](https://github.com/gboeing/data-visualization/blob/main/ncaa-football-stadiums/data/stadiums-geocoded.csv?fbclid=IwAR3GI7EfGMABzBMDOEATqbvYnNzG7P0Ghi8INIpKrc_kicnvbSqOWh63crI)

**File Name: FootballDL**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Explanation** | **Example** |
| SCL\_UNITID | UNITID assigned to post-secondary institution | 100654 |
| SchoolName\_footballDL | Renamed from: SCL\_NAME  Official name of institution | Alabama A&M University |
| SCL\_DIV\_14 |  | 1 |
| SCL\_SUB\_14 |  | 2 |
| CONFNAME\_14 |  | Southwestern Athletic Conf. |
| D1\_FB\_CONF\_14 |  | Southwestern Athletic Conf. |
| SCL\_HBCU | Numeric. Text | 1 |
| SCL\_PRIVATE | Numeric. Text | 0 |
| MULTIYR\_APR\_RATE\_1000\_RAW | Numeric. Four year Academic Progress Rate on 1000 point scale. Raw value. | 932 |
| MULTIYR\_APR\_RATE\_1000\_CI | Numeric. Four year Academic Progress Rate adjusted based on upper confidence interval boundary. | 945 |
| MULTIYR\_APR\_RATE\_1000\_OFFICIAL | Numeric. Four year Academic Progress Rate officially used for penalty. | 932 |
| MULTIYR\_ELIG\_RATE | Numeric. Four year Eligibility Rate on 1000 point scale. | 0.886097 |
| MULTIYR\_RET\_RATE | Numeric. Four year Retention Rate on 1000 point scale. | 0.961279 |
| MULTIYR\_SQUAD\_SIZE | Numeric. Four year Squad Size. | 321 |
| 2014\_APR\_RATE | Renamed from: APR\_RATE\_2014\_1000  Numeric. Academic Progress Rate | 905 |
| 2014\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2014  Numeric. Eligibility Rate | 0.810811 |
| 2014\_RET\_RATE | Renamed from: RET\_RATE\_2014  Numeric. Retention Rate | 0.97973 |
| 2014\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2014  Numeric. Squad Size *\*From this point on, all APR, ELIG, RET, and NUM\_OF\_ATHLETES are referring to these same statistics.* | 78 |
| 2013\_APR\_RATE | Renamed from: APR\_RATE\_2013\_1000 | 933 |
| 2013\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2013 | 0.904459 |
| 2013\_RET\_RATE | Renamed from: RET\_RATE\_2013 | 0.954839 |
| 2013\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2013 | 86 |
| 2012\_APR\_RATE | Renamed from: APR\_RATE\_2012\_1000 | 940 |
| 2012\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2012 | 0.893333 |
| 2012\_RET\_RATE | Renamed from: RET\_RATE\_2012 | 0.959732 |
| 2012\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2012 | 78 |
| 2011\_APR\_RATE | Renamed from: APR\_RATE\_2011\_1000 | 951 |
| 2011\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2011 | 0.93662 |
| 2011\_RET\_RATE | Renamed from: RET\_RATE\_2011 | 0.950704 |
| 2011\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2011 | 79 |
| 2010\_APR\_RATE | Renamed from: APR\_RATE\_2010\_1000 | 931 |
| 2010\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2010 | 0.916031 |
| 2010\_RET\_RATE | Renamed from: RET\_RATE\_2010 | 0.9375 |
| 2010\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2010 | 69 |
| 2009\_APR\_RATE | Renamed from: APR\_RATE\_2009\_1000 | 935 |
| 2009\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2009 | 0.909091 |
| 2009\_RET\_RATE | Renamed from: RET\_RATE\_2009 | 0.946565 |
| 2009\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2009 | 74 |
| 2008\_APR\_RATE | Renamed from: APR\_RATE\_2008\_1000 | 930 |
| 2008\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2008 | 0.901639 |
| 2008\_RET\_RATE | Renamed from: RET\_RATE\_2008 | 0.95 |
| 2008\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2008 | 64 |
| 2007\_APR\_RATE | Renamed from: APR\_RATE\_2007\_1000 | 919 |
| 2007\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2007 | 0.85906 |
| 2007\_RET\_RATE | Renamed from: RET\_RATE\_2007 | 0.979866 |
| 2007\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2007 | 75 |
| 2006\_APR\_RATE | Renamed from: APR\_RATE\_2006\_1000 | 941 |
| 2006\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2006 | 0.911111 |
| 2006\_RET\_RATE | Renamed from: RET\_RATE\_2006 | 0.962963 |
| 2006\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2006 | 70 |
| 2005\_APR\_RATE | Renamed from: APR\_RATE\_2005\_1000 | 921 |
| 2005\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2005 | 0.875862 |
| 2005\_RET\_RATE | Renamed from: RET\_RATE\_2005 | 0.965517 |
| 2005\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2005 | 74 |
| 2004\_APR\_RATE | Renamed from: APR\_RATE\_2004\_1000 | 871 |
| 2004\_ELIG\_RATE | Renamed from: ELIG\_RATE\_2004 | 0.86014 |
| 2004\_RET\_RATE | Renamed from: RET\_RATE\_2004 | 0.881119 |
| 2004\_NUM\_OF\_ATHLETES | Renamed from: NUM\_OF\_ATHLETES\_2004 | 74 |
| 2015\_PUB\_AWARD | Numeric. Renamed from: PUB\_AWARD\_15  Public recognition award. 1 = received; 0 did not. *\*From this point on, PUB\_AWARDs are referring to these same statistics.* | 0 |
| 2014\_PUB\_AWARD | Renamed from: PUB\_AWARD\_14 | 0 |
| 2013\_PUB\_AWARD | Renamed from: PUB\_AWARD\_13 | 0 |
| 2012\_PUB\_AWARD | Renamed from: PUB\_AWARD\_12 | 0 |
| 2011\_PUB\_AWARD | Renamed from: PUB\_AWARD\_11 | 0 |
| 2010\_PUB\_AWARD | Renamed from: PUB\_AWARD\_10 | 0 |
| 2009\_PUB\_AWARD | Renamed from: PUB\_AWARD\_09 | 0 |
| 2008\_PUB\_AWARD | Renamed from: PUB\_AWARD\_08 | 0 |
| 2007\_PUB\_AWARD | Renamed from: PUB\_AWARD\_07 | 0 |
| 2006\_PUB\_AWARD | Renamed from: PUB\_AWARD\_06 | 0 |

Notes:

* Source: [NCAA Division I Academic Progress Rate](https://www.icpsr.umich.edu/web/ICPSR/studies/26801/summary?fbclid=IwAR2VQY6AMDY03Ykk0f5j49pI-3Oez7evDR-UmXCMOIWirzi7073gY4_qY_E)
* [Click here for more information on what the columns mean.](https://ncaaorg.s3.amazonaws.com/research/academics/2020RES_APR2019Codebook.pdf)
* As you can see from the pdf attached, I could have gotten a more recent dataset but considering the data provided was from 2014-2015, I didn’t see a reason to have more data than I could use. Granted, this may affect the accuracy of my recommendations, but I’m just going to continue as if we’re in SY2015, when the assignment was likely created.
  + I did delete the following columns from the Excel file before continuing:
    - SPORT\_CODE. It was the Sport identification code; All entries were 4. (4 = Football)
    - SPORT\_NAME. All entries were Football.
    - ACADEMIC\_YEAR. Academic year in which latest data was recorded. All entries were 2014.

## Exploratory Analysis

First, I joined all the numeric information into a dataframe and started exploring it a bit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Histogram: | Plotting a histogram of the TotalPay, I was surprised to find so many coaches making less than $ 1M.  Highest paid coaches:   |  |  |  |  | | --- | --- | --- | --- | | Alabama | Nick Saban | SEC | 8307000 | | Ohio State | Urban Meyer | Big Ten | 7600000 | | Michigan | Jim Harbaugh | Big Ten | 7504000 | | Texas A&M | Jimbo Fisher | SEC | 7500000 | |
| I pulled the top and lowest paid coaches. It looks like coaches in SEC overall get paid more, while MAC get paid less. | Lowest paid coaches:   |  |  |  |  | | --- | --- | --- | --- | | Louisiana-Monroe | Matt Viator | Sun Belt | 390000 | | Coastal Carolina | Joe Moglia | Sun Belt | 400000 | | Akron | Terry Bowden | MAC | 412500 | | New Mexico State | Doug Martin | Ind. | 419640 | |

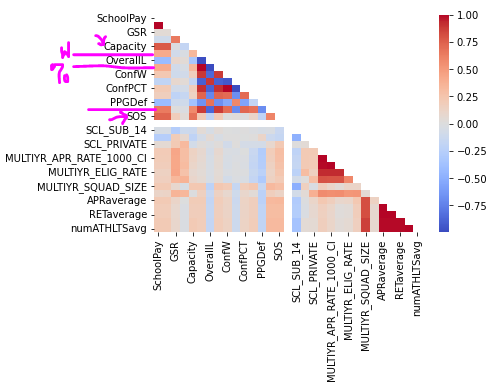
So one variable I may want to include in analysis is **Conference**. Let’s check a few others before building a model.

If most teams are earning 0 awards, maybe the few who earn awards get paid more? Check Public Awards.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Histogram: | |  |  |  |  | | --- | --- | --- | --- | | Duke | David Cutcliffe | ACC | 2540928 | | Northwestern | Pat Fitzgerald | Big Ten | 3619775 | | Rutgers | Chris Ash | Big Ten | 2200000 | | Air Force | Troy Calhoun | Mt. West | 885000 | | Stanford | David Shaw | Pac-12 | 4311543 | |

No noticeable overlap between teams. Maybe not.

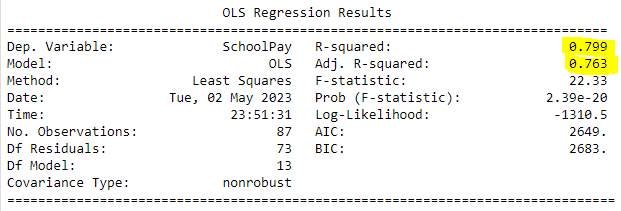
Let’s check all potential relationships via correlation:



Disregard relationship to Total Pay. Strongest correlations to School Pay seem to be Capacity, SRS and SOS. SRS is Ranking – we may disregard that because it may seem obvious that a higher ranked team would have a higher paid coach. SOS is interesting, though: Strength of Schedule refers to the combined winning percentages of opponents played. So playing higher ranked / tougher teams could affect coach pay. Could also look into Overall W as well. OverallPCT can be discounted because it is essentially the same as wins.

So, we’ll focus on Conference, Capacity, SOS, and Wins.

## Regression Model (salary is response var, relevant predictors)



I do want to include a quick note about the scatterplots included:

|  |  |
| --- | --- |
|  | There is a clear relationship between Capacity and Pay. |
|  | And a clear relationship between Capacity and SOS.  This is where I’m concerned there may be collinearity. I thought it was SOS and Wins, but upon removing Wins in the regression analysis, I didn’t see a difference – so there may be a relationship between Capacity and SOS. |
|  | Interesting. This is a weak relationship. |
|  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | ConfCode | confNum | | AAC | 1 | | ACC | 2 | | Big 12 | 3 | | Big Ten | 4 | | C-USA | 5 | | Ind. | 6 | | MAC | 7 | | Mt. West | 8 | | Pac-12 | 9 | | SEC | 10 | | Sun Belt | 11 | | I found confNum to be confounding in my linear model. However, it is clear that teams in the SEC and Big Ten are paid more than other conferences, especially  C-USA, MAC, Mt. West, and Sun Belt.  Changing to SEC or Big Ten conferences could increase coach salary. | |

## Recommendations

Our model appears to be good, explaining over 75% of our data. However….

Multicollinearity is a concern. I tried to avoid this by not including both SRS and SOS, and Wins and PCT. I thought SOS was highly correlated with Wins so, in my code, I tried removing Wins. I got a similar R-squared statistic and the warning that there might still be multicollinearity. It also meant I could not trust the p-values I got.

Based on my current model though, Capacity, SOS, and Conference affect SchoolPay for football coaches. The bigger the stadium, and the more difficult their opponents, the more a football coach will make. **If the Syracuse coach wants to increase their salary, they need to challenge – and win against – tougher opponents. Get better. Get good. As previously stated, playing in the SEC or Big Ten conferences could increase coach salary as well.**