Jacob Dineen

IST 718 – Advanced Info Analytics

Lab 1

1/24/2018

The following is a high-level recording of our analysis concerning the coaches dataset, which was merged with graduation rates and stadium size, matched to the best of our ability. A foray into Python, this lab tested my ability to learn on the fly, at a rapid pace, and get to know the libraries that are most common to aspiring and professional data scientists. Answers below are derived from summary statistics and predictive techniques noted in my Jupyter workbook.

Okay

o What is the recommended salary for the Syracuse football coach?

We can use a linear equation to derive salary output based on a series of predictor variables. – Because we've noticed that graduation rates don't have statistically significance in our multiple regression, we leave them out to improve our goodness of fit. The below is output from an ordinary least squares model calculated in Python. GSR, FGR and if a school was private were left out of the two models below due to their low p values. Worth mentioning is the lack of collinearity seen amongst our features when constructing a correlation matrix, which is an assumption that linear regression makes- I believe this would classify OLS/MLR as a parametric model because it has a finite set of parameters.

✓

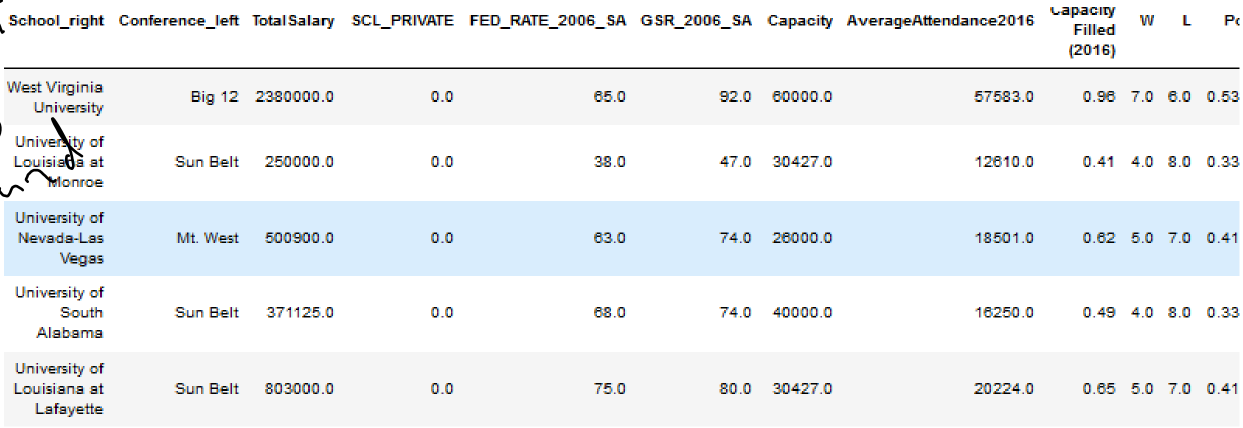
we don want collinearity

Through the extrapolation of data from various sources, we ~~could~~ created a dataframe containing attendance, wins/losses, stadium capacity and all pertinent data pertaining to the original coaches file.

Variables include: School, Conference, Private/Not Private, Fed Rate, GSR, Capacity, Avg Attendance, Wins, Losses and Winning Percentage.

Think about changing column headings

consider giving background



School_right	Conference_left	TotalSalary	SCL_PRIVATE	FED_RATE_2006_SA	GSR_2006_SA	Capacity	AverageAttendance2016	Capacity Filled (2016)	W	L	Pc
West Virginia University	Big 12	2380000.0	0.0	65.0	92.0	60000.0	57583.0	0.96	7.0	6.0	0.53
University of Louisiana at Monroe	Sun Belt	250000.0	0.0	38.0	47.0	30427.0	12610.0	0.41	4.0	8.0	0.33
University of Nevada-Las Vegas	Mt. West	500900.0	0.0	63.0	74.0	26000.0	18501.0	0.62	5.0	7.0	0.41
University of South Alabama	Sun Belt	371125.0	0.0	68.0	74.0	40000.0	16250.0	0.49	4.0	8.0	0.33
University of Louisiana at Lafayette	Sun Belt	803000.0	0.0	75.0	80.0	30427.0	20224.0	0.65	5.0	7.0	0.41

How many obs?

How many vari?

consider using labels for each figure or table => this allows reference later in the document

Good

But school gives us a weird impact

Linear equation that will be used below:

TotalSalary ~ School + Conference + Capacity + AverageAttendance2016 + W + L + Intercept

Big East

Y^= -152800(1) + -416400(1) + 12.4563(49250) + 8.6718 (32805) + 105900(4) + 88250 (8) - 199200

Y^= $1,259,151.17

Y= $1,259,276.0

Error= $124.83

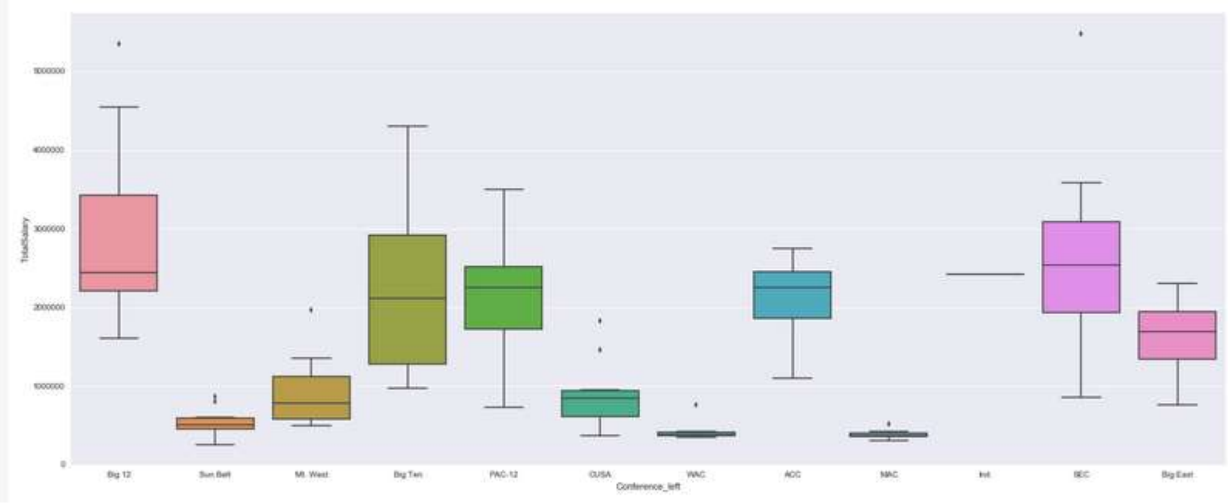
Our prediction, based on the above predictors, is that the recommended salary for the Syracuse football coach is $1,259,151.17. If we look back at the coaching data set, we can see that Doug Marrone, the Syracuse head football coach made $1,259,276.00 in the year the salaries were recorded. Our residual, or error, is $124, or within 1% of the actual value. ✓

Big Ten

Y^= -152800(1) + -199700(1) + 12.4563(49250) + 8.6718 (32805) + 105900(4) + 88250 (8) - 199200

Y^= $1,475,851.174

If we were to place Syracuse into the Big 10, Doug Marrone would see an 17% rise in his salary, based on the variance in this dataset. Looking back at one of graphs generated during our data gathering and summary statistics phase, we can see that coaches in the Big 10 have a higher average salary than coaches in the Big East- Perhaps strength of schedule and nationally broadcasted games comes into play here. ✓



*Reference Notebook for a bar graph with average salaries by Conference. ✓

*At the time the data was collected, Syracuse was noted as being in the Big East.

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

We were forced to drop certain schools that did not have salary information for our coaches. Additionally, it should be noted that we employed the use of fuzzy matching to align graduation rates, along with stadium capacity with our original data set, and that some matching may not be precise, although an eye test confirmed that accuracy was high.

what were the results? # of success

✓ **Schools dropped from data include:** Temple, Miami, Pittsburgh, Tulane, Tulsa, Brigham Young, Stanford and Vanderbilt. These schools were spread out across different conferences and their absence, while noted, shouldn't make or break or ability to predict total salary.

o What effect does graduation rate have on the projected salary?

Graduation Rate data was sourced from the NCAA site provided, and was deemed to have a moderate impact on the model. Both Graduation Success Rate and Federal Graduation Rate, regarding student athletes, were used in some iteration of our OLS models. GSR had a moderate impact on our response, while Fed Rate had a negative impact, although both variables were deemed to be less than statistically significant (low p values), and were left out of our final model to improve our adjusted R2. I think it would be difficult to derive a clear-cut causal relationship between these predictors and the response variable in the real world. It may be idealistic to believe that coaches of major college programs are truly concerned about their students' academic success rate, and even more so that the person paying their salary does.

✓ ✓ ?

o How good is our model?

With our final model built, we see an R2 of 69.1%, and an adjusted R2 of 64.1%. We could achieve a slightly higher R2 with more variables, but our adjusted R2 would subsequently fall. It would make sense that we don't have close to enough data to eliminate unknown causes of variance in our dataset. It would have been useful to have data on each coaches' current record as coach of their existing team, as an aggregate, and perhaps bowl games attended and won, seeing as how the league seems to heavily reward performance on the football field. Length of tenure also comes to mind, as the general trend of years coached in relation to salary appears to be linear. Another interesting piece of data that would have been nice, and that I will potentially circle back to once my Python skills are up to par, would be layering in schedule ranking per college. I would venture to assume that schools with more difficult schedules, or schools playing teams that are nationally ranked and generate more revenue on broadcast would have much more disposable income to distribute to coaches.

Great

o What is the single biggest impact on salary size?

The biggest impact on salary size, with the predictor variables that were utilized in this lab, was Conference, but stadium size was the variable with the lowest p value (highest significance to

Do conference deals for $, matter?

✓

our model), and had the potential to be the one of the biggest driver of salary in an upward direction, along with average attendance and Wins (meaning stadium size increased at an exponential rate, and the value for conference was simply its resulting coefficient).

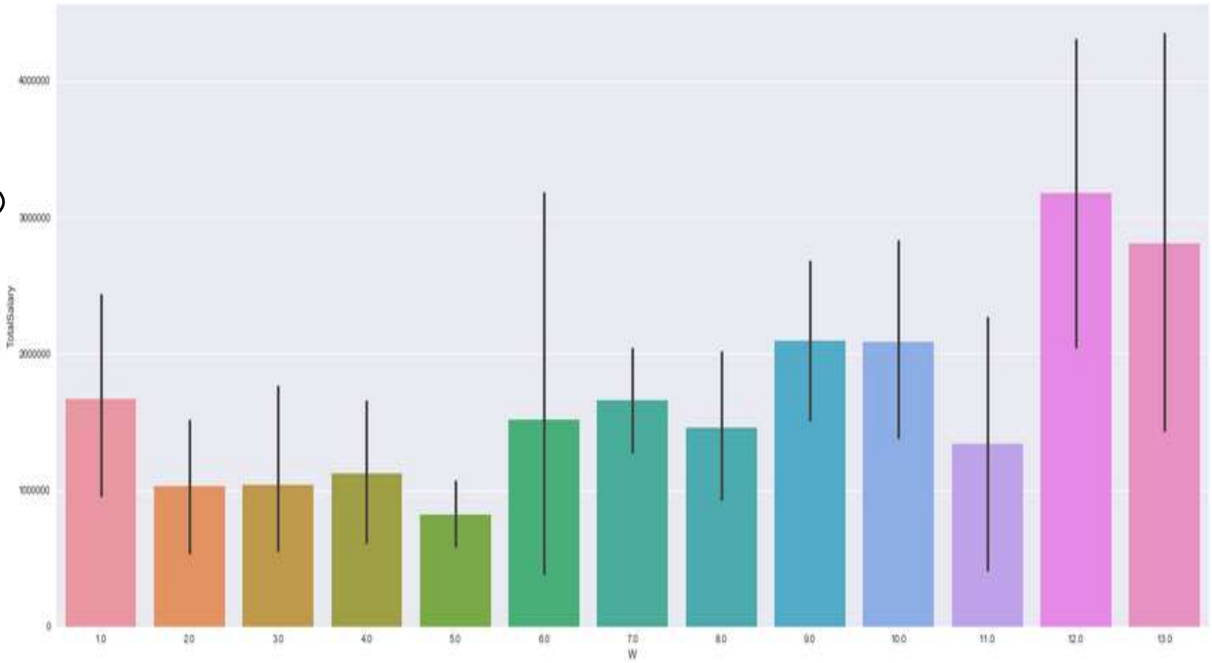
Consider Starting here

Using summary statistics, boxplots, lattice plots and bar plots, we can shape some initial views on the distribution of our data, and relationships between variables. A block of code in my Jupyter script contains many different visualizations relating to our final merged dataset. For instance, a direct linear relationship can be seen between wins (2017) and salary. We could also see that coaches in the SEC were the highest paid of any conference, on average.

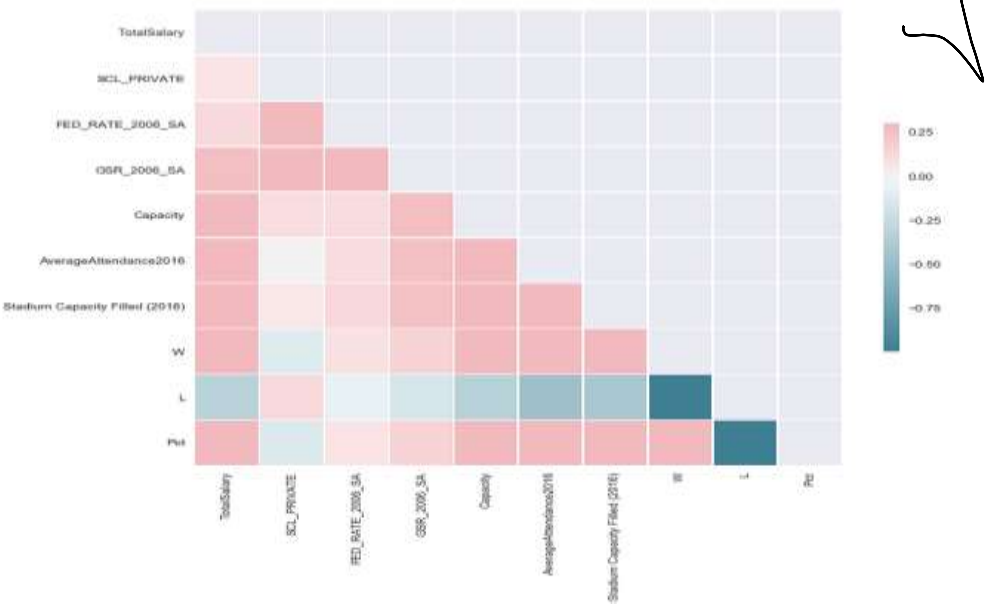
Wins in relation to total salary:

The below shows a linear relationship between wins in relation to total salary accrued by a coach during the specified timeframe.

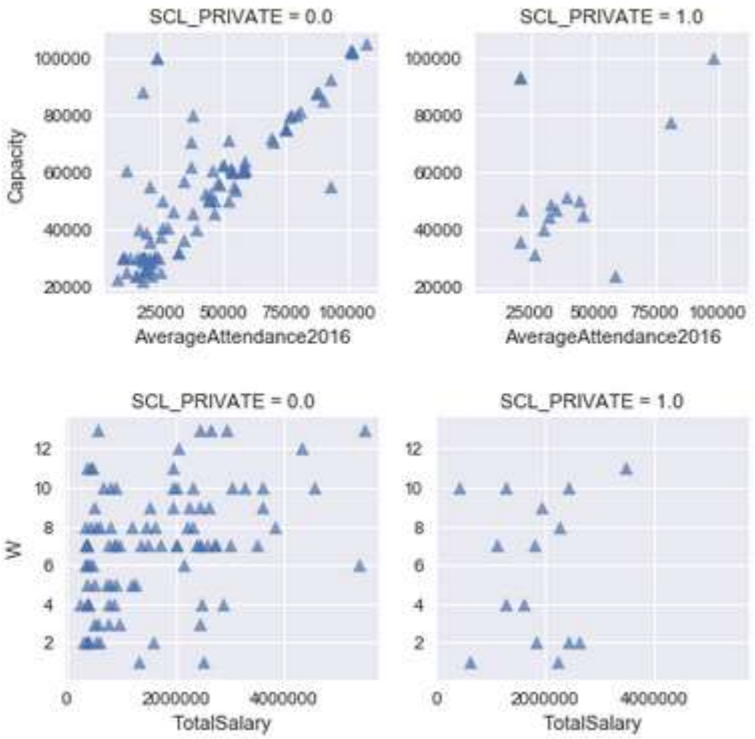
Consider formatting of axis



Correlation plot of relevant numerical variables:

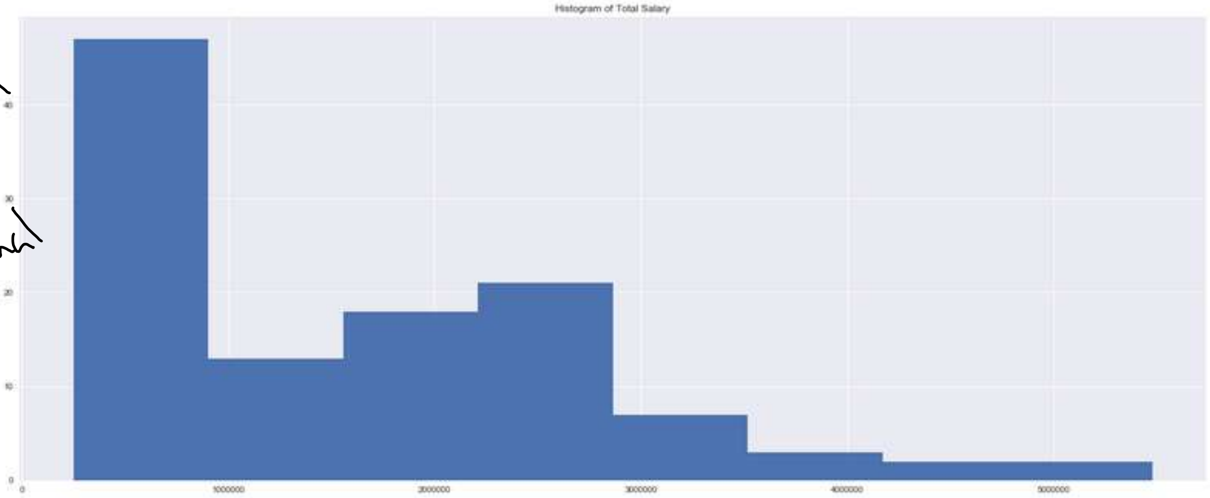


Sample lattice plots used to derive relevant relationships and data structure/distribution:



Histogram of Total Salary w/ auto bins:

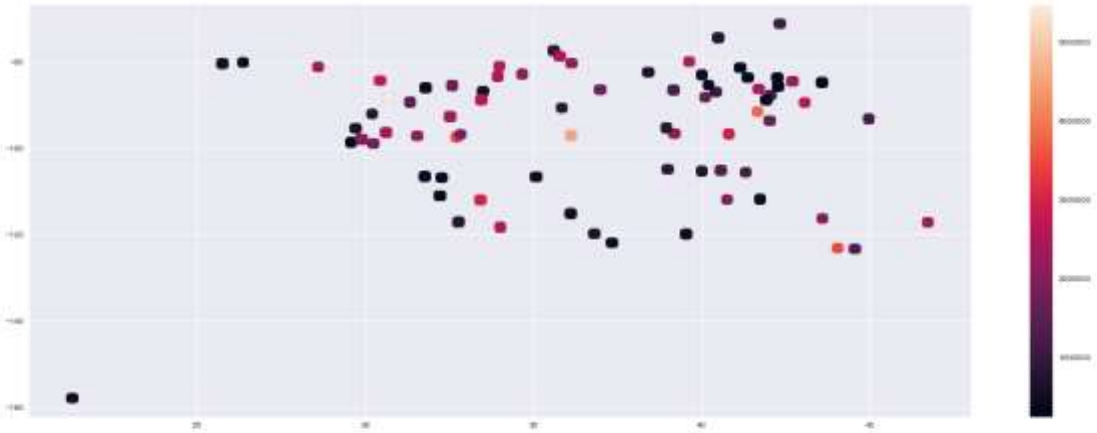
How could we deal with non normal data?



*This was generated post analysis, but helps to show the non normal distribution of the data. Appears to show a long tail towards the right hand side, and heavy kurtosis (peak).

Bonus:

✓ I was able to create a *working* geographical representation of salary in relation to latitude and longitude (output in notebook), but my Python skills are not currently advanced enough to embed them within a border of the United States (I have done something similar in R, but could not find documentation that was comprehendible for a beginner).



All in all, this lab helped me to learn and understand Python syntax and logic, as well as giving me some insight into the Pandas/Numpy/Matplot/Scikit distributions.

Sources:

Matching: https://github.com/RobinL/fuzzymatcher/blob/master/examples.ipynb

Levenshtein: https://www.lfd.uci.edu/~gohlke/pythonlibs/#python-levenshtein

Geocoding tool online: https://www.doogal.co.uk/BatchGeocoding.php

Academic Data: https://www.icpsr.umich.edu/icpsrweb/NCAA/studies/30022#datasetsSection

Stadium Capacity Data:
https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums

Attendance(2016) data: https://www.cbssports.com/college-football/news/college-football-attendance-in-2016-crowds-decline-for-sixth-straight-year/

2017 standings: https://www.sports-reference.com/cfb/years/2017-standings.html