Kwasi Mensah

IST 718

Big Data Analytics

Lab Exercise 1

**Objective:**

1) Obtain data and understand data structures and data elements.

2) Scrub data using scripting methods, to include debugging, for data manipulation in R and other tools.

3) Explore data using essential qualitative analysis techniques including descriptive statistics.

4) Model relationships between data using the appropriate analytical methodologies matched to

the information and the needs of clients and users.

5) Interpret the data, model, analysis, and findings. Communicate the results in a meaningful way.

**Introduction**:

This exercise provides an opportunity to demonstrate our ability to combine data sets and produce meaningful analysis. We will use the datasets that were given and those that were manually acquired in order to try to predict the best pay and what may be the possible main predictor or determinant in deciding the amount each NCAA football coach earns in salary. We will perform an analysis to determine a recommendation for Syracuse’s next football head coach.

**Question:**

“How can we recommend the best salary for our next head football coach?” by use of visualizations and models.

**Data Sets Used for Analysis:**

*‘mdf’ Data Frame (123 Rows x 23 Columns)*



**Data Obtained:**

* I first started with the ‘coaches9’ data set, which we obtained from the 2SUBDA repository on GitHub.
  + 129 Rows x 9 Columns & 0 Null values
* I next used the link provided to obtain the stadium csv file from GitHub and decided on the combined dataset named. ‘stadium\_geocode’
  + 253 Rows x 11 Columns & 107 Null values
* I manually created ‘gradrate’ csv file by using the NCAA website and copying and pasting the data into an excel csv file.
  + 257 Rows x 7 Columns & 20 Null values
* I created two data sets manually rather than scraping them online with python for example.
* I decided to look online and search for the win loss records for the NCAA 2022 Football season and copy and pasted the NCAA Division 1 teams record into excel named ‘winrate’.
  + 131 Rows x 5 Columns & 0 Null values
* Once the data sets were collected, I merged all four of the csv files in python by merging them on the like column “Team” which was later renamed to “School” and named this new data frame for analysis “mdf”.

**Data Cleaning (Scrub):**

* Used Excel's filter feature to sort by individual conference.
* Manually edited the GSR dataset Team column to match the formatting of the Team column in the other data frame.
* Compared the conferences on the ‘coaches9’ csv and ‘gradrate’ csv and manually edited the Team column to match the same format in order to join the data.
* Edited the conference format to match that of both the coach data frames and the stadium data frames when analyzing them in the future.
* Verified that the number of schools per conference matched in each csv file.
* Manually changed the values in the team and conference columns in the ‘stadium’ csv file to match the ‘coaches9’ csv file to maintain observations when merging data frames on python.
* Used Excel's find and replace feature to remove unnecessary words and extra spaces in the "Team" column of the csv.
* Changed "Independent" to "Ind." to match all the data frames to be combined.
* Manually removed extra spaces found in cells.
* Some schools were not found in certain datasets, so they were removed or edited accordingly.
  + James Mad, Temple, not in ‘coaches 9’ compared to ‘winrate.
  + Alabama at Birmingham, not in ‘winrate’ compared to ‘coaches9’.
* Manually corrected some school names to match across all data frames.
  + Miami (OH) changed to Miami (Ohio),
  + Middle Tenn to Middle Tennessee
  + Wash State to Washington State
  + UNLV to Nevada-Las Vegas
  + USC to Southern California
  + BYU to Brigham Young
  + Central Fl to central Florida
  + VA Tech to Virginia Tech

I decided to do some of these steps manually to avoid losing observations when merging data frames in python.

* We next dropped the 'Conference\_x', 'Conference\_y', 'State\_x','Sport', 'AssistantPay', 'Cohort Year' columns in python to reduce noise when creating our models.
* Count any number of n/a in the data sets.
  + Replace the 3 n/a found in the ‘FGR’ column with the mean of the ‘FGR’ column.
  + Drop the ‘Expanded’ column due to it having 22 n/a’s.
* Check the data type of the columns and change any as needed.
  + Win\_Percent, SchoolPay, TotalPay, Bonus, BonusPaid, Buyout to float object.
* Removed any special characters from the columns that were integers.
* Clean and separate the “Win-Loss Record” to make the ‘W’ and ‘L’ columns and drop the “Win-Loss Record” column.
* Drop the 4 schools (Rice, Baylor, Southern Methodist, and Brigham Young) that had the “TotalPay” value missing in that column.
* Create a school id and state id column for correlation visualization.

**Exploration & Interpretation**

Figure 1

Figure 1 is a snapshot of the statical properties of our merged data frame that would be used for our analysis. This is a data set after cleaning was performed. We see the min and max value in each column, as well as the mean, count, and standard deviation for each column. Lastly, we see the 25% (lower quartile), 50%(median), & 75% (upper quartile) percentiles of our columns.

Graphical user interface, application

Description automatically generatedTable

Description automatically generated

Figure 2

A correlation matrix was created for the variables in our data set that makes use of a heat map to show how each variable correlates with one another. We see that in the middle of the heat map that there is a high positive correlation of the variable TotalPay with Capacity, Buyout, BonusPaid, and Bonus.

Chart

Description automatically generated

Figure 3

A Box and whisker plot was created for the Total Pay of the college coaches. We arranged this visualization in descending order. We can see that the top 5 conferences in terms of pay for coaches are the SEC, Big Ten, Big 12, ACC, & the PAC-12. We have noticed some outliers in salaries for each division, and we could possibly do further research with additional data to find out the reasons or conditions that were met to have such high salaries. Lastly, we see that many of the coaches in the NCAA are making over $1,000,000, but five divisions have coaches receiving lower salaries.

Chart, box and whisker chart

Description automatically generated

Figure 4

A horizontal bar chart visualizing a comparison of Syracuse previous coach’s salary if it were a part of the previously disbanded Big East Conference. We see that in this conference in the past Syracuse coaches were in the top 3 in terms of salary earned.

Chart, bar chart

Description automatically generated

Figure 5

Figure 5 is a horizontal bar chart comparing the salaries of coaches in the Big Ten Conference with the addition of Syracuse added to the conference. We see that if in this division Syracuse football coaches would be in the bottom 3 in terms of salary earned when compared with the conference.

Chart, bar chart

Description automatically generated

Figure 6

We see that when compared with the school in Syracuse’s actual conference we see that the coach’s salary is 2nd to last in earn salary when compared to the salaries of the other coaches in the division. This is probably because of the low winning percentage on average the team has had over the years.

Chart, bar chart

Description automatically generated

Figure 7

I created a multiple scatter plot that compares the TotalPay variable with Capacity, GSR, FGR, &Win %. We see that there is a clear positive correlation between the total pay of coaches when looking at the size of the stadium each school has, and the win rate the coach has for the school. We can see there is somewhat of a positive correlation of the Division 1 athletes graduation rates, and a negative correlation when comparing total pay with the federal graduation rate which is understandable being that that rate is a lot lower than the GSR on average.

Chart, scatter chart

Description automatically generated

Figure 8

A box and whisker plot were created to compare the total pay of NCAA football coaches who are in different Tournament divisions. We see that there is a significant difference between the pay of coaches in the FBS (Football Bowl Series) and the FCS (Football Championship Series) divisions. We believe that FBS division coaches make significantly more than FCS coaches, and that FCS coaches’ maximum salaries are barely more than FBS lower quartile salaries. (Lower 25% of FBS coach salaries.)

=Chart, bar chart, box and whisker chart

Description automatically generated

Modeling:

We created a copy of our original “mdf” data frame and named it “tdf” to create models for analysis. Next, we created multiple Regression models to try to understand any relationship between the variables in our dataset between the dependent variables ‘TotalPay’ and various combinations of independent variables. In addition, we will use one of best models to try and predict what the recommended salary should be for the next Syracuse football coach. We added a new column ‘runiform’ that’s creates random observations of 0’s and 1’s. We split our data frame into a training and testing set where our training set has any rows that is 33% or more and for the training where ‘runiform’ is less than 33%. Lastly, we fit and create our Ordinary Least Square Regression model and investigate our results for the 7 initial models created.

Train: 89 Rows x 26 Columns

Test: 34 Rows x 26 Columns

Single Regression:

m1 = 'TotalPay ~ Win\_Percent'

Table

Description automatically generated

* The Adjusted R-squared of M1 was .14, or 14% of the variation in 'TotalPay' can be explained by 'Win\_Percent'.
* The F-Statistic, which tells us how significant overall our model has a score of 15.35 with a p-value of .000177 which is less than .05 telling us our model is statistically significant.
* The Coefficient tells that that every single increase in win percentage there is a $37440 increase to Total Pay
* The proportion of variance for this model was 0 which tells us that this model is not a good predictor for new data.

m2= 'TotalPay ~ Capacity'

Table

Description automatically generated

* The Adjusted R-squared of M2 was .674, or 67.4% of the variation in 'TotalPay' can be explained by 'Capacity.
* The F-Statistic, which tells us how significant overall our model has a score of 182.9 with a p-value of 4.19e-23 which is less than .05, telling us our model is statistically significant.
* The Coefficient tells that that every single increase in stadium is a $64.52 increase to Total Pay
* The proportion of variance for this model was .571 or 57.1% which tells us that this model may or may not be a good predictor for new data.

Multiple Regression:

(The variables that had the highest correlation to the Total pay Dependent variable.)

m3 = 'TotalPay ~ Capacity + Win\_Percent + Buyout +Bonus+ BonusPaid '

A black and white document

Description automatically generated with low confidence

* The Adjusted R-squared of M3 was .791, or 79.1% of the variation in 'TotalPay' can be explained by the independent variables chosen.
* The F-Statistic, which tells us how significant overall our model has a score of 67.57 with a p-value of 8.40e-28 which is less than .05 telling us our model is statistically significant.
* The most significant Coefficient was ‘Win\_Percent’ tells that that every single in win percentage there is a $9,383 increase in Total Pay
* The proportion of variance for this model was .638 or 63.8% which tells us that this model may or may not be a good predictor for new data.

m4 = 'TotalPay ~ Win\_Percent + Capacity + GSR+FGR'

Table

Description automatically generated

* The Adjusted R-squared of M4 was .691, or 69.1% of the variation in 'TotalPay' can be explained by the independent variables chosen.
* The F-Statistic, which tells us how significant overall our model has a score of 50.20 with a p-value of 1.64e-21 which is less than .05 telling us our model is statistically significant.
* The most significant Coefficient was ‘GSR’ tells that for every single increase in Graduation Rates there is a $28,674 increase in Total Pay
* The proportion of variance for this model was .548 or 54.8% which tells us that this model may or may not be a good predictor for new data.

m5 = 'TotalPay ~ Win\_Percent + Capacity+ Buyout +GSR'

Table

Description automatically generated

* The Adjusted R-squared of M5 was .802, or 80.2% of the variation in 'TotalPay' can be explained by the independent variable chosen.
* The F-Statistic, which tells us how significant overall our model has a score of 90.01 with a p-value of 1.49e-29 which is less than .05, telling us our model is statistically significant.
* The most significant Coefficient was ‘GSR’ tells for every single increase in Graduation Rates there is a $28,025 increase in Total Pay
* The proportion of variance for this model was .681 or 68.1% which tells us that this model may be a good predictor for new data.

m6 = 'TotalPay ~ Win\_Percent + Capacity+ Buyout+ FGR'

Table

Description automatically generated

* The Adjusted R-squared of M6 was .794, or 79.4% of the variation in 'TotalPay' can be explained by the independent variables chosen.
* The F-Statistic, which tells us how significant overall our model has a score of 85.82 with a p-value of 7.42e-29 which is less than .05 telling us our model is statistically significant.
* The most significant Coefficient was ‘FGR’ tells that for every single increase in Federal Graduation Rates there is a $17,525 increase in Total Pay
* The proportion of variance for this model was .689 or 68.9% which tells us that this model may be a good predictor for new data.

m7 = 'TotalPay ~ State\_ID + Capacity + Win\_Percent + MOV+ GSR+ Buyout+ BonusPaid '

A picture containing text, receipt, screenshot

Description automatically generated

* The Adjusted R-squared of M6 was .798, or 79.8% of the variation in 'TotalPay' can be explained by the independent variables chosen.
* The F-Statistic, which tells us how significant overall our model has a score of 50.69 with a p-value of 5.33e-27 which is less than .05, telling us our model is statistically significant.
* The most significant Coefficient was ‘GSR’ tells that for every single increase in Graduation Rates there is a $29,488 increase in Total Pay
* The proportion of variance for this model was .693 or 69.3% which tells us that this model may be a good predictor for new data.

Bonus Figures

Bonus 1

The same visualization that was created in Figure 3, but instead of being in descending order by salary they are in ascending order by the median salary per conference. When organized like this we can see two groups of three conferences that have close to similar median salaries for their football coaches.

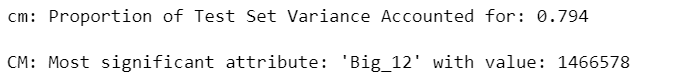
Chart, box and whisker chart

Description automatically generated

Bonus 2

A close-up of a document

Description automatically generated with medium confidence



* The Adjusted R-squared of CM was .881, or 88.1% of the variation in 'TotalPay' can be explained by the independent variables chosen.
* The F-Statistic, which tells us how significant overall our model has a score of 39.27 with a p-value of 2.60e-29 which is less than .05 telling us our model is statistically significant.
* The most significant Coefficient was ‘Big\_12’ tells that being a part of this conference there may be an associated $1,467,000 increase in Total Pay
* The proportion of variance for this model was .794or 79.4% which tells us that this model may or may not be a good predictor for new data.

Conclusion:

In conclusion when using our models to predict the recommended salary for the next Syracuse head football coach, we recommend a salary around $2,705,965.01 for the ACC conference. If Syracuse were still apart of the disbanded Big East conference the recommended estimate salary would be $2,077,458.39. Lastly, if we recommend a salary for Syracuse’s head coach if they were apart of the Big Ten would be $2,959,692.18. We see from our different models that used the Graduation Rate of Division 1 Athletes as an independent variable often time was the most significant. The Graduation Rate on average for every single increment causes around a $28,0000 increase in coaches’ salaries. We do see that the capacity of a Stadium is a big factor that plays into the coaches’ salaries. In addition, we saw that the win rate of the coaches, the tournament league, and the graduation rate of their athletes are the main factors that go into determining a coach’s salary in the NCAA. We feel that with more additional datasets with assistant coaches pay, the usual conference leaders, the average income in the surround areas for each school, etc. we can do more research in determining an effective recommendation for NCAA football coaches’’ salaries.

Appendix:

[Shared NCAA Research Data - NCAA.org](https://www.ncaa.org/sports/2016/12/14/shared-ncaa-research-data.aspx)

[Graduation Success Rate (ncaa.org)](https://web3.ncaa.org/aprsearch/gsrsearch)

[2020RES\_APR2019Codebook.pdf](http://s3.amazonaws.com/ncaa.org/documents/2020/5/19/2020RES_APR2019Codebook.pdf)

[data-visualization/ncaa-football-stadiums/data at main · gboeing/data-visualization · GitHub](https://github.com/gboeing/data-visualization/tree/main/ncaa-football-stadiums/data)

[GitHub - 2SUBDA/IST\_718: Basic Repository for 2SU Version of IST 718](https://github.com/2SUBDA/IST_718)

<https://seaborn.pydata.org/>

<https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/>

<https://www.teamrankings.com/ncf/trends/win_trends/>