**IST 718**

**LAB 1 ASSIGNMENT**

Matthew L. Pergolski

Professor Jillian Lando

**Overview**

In the IST 718 Lab 1 assignment, the O-S-E-M-IN method will be conducted to recommend a specific salary of Syracuse’s next Head Football Coach. In order to provide the analysis, a brief explanation of the data science method will be required.

* O
  + Obtain: In the obtaining section, Data Acquisition will be discussed and referenced.
* S
  + Scrub: In the scrubbing section, Data Cleaning will be discussed and referenced.
* E
  + Explore: In the exploring section, Data Exploration will be discussed and referenced.
* M
  + Model: In the modeling section, Data Modeling techniques will be discussed – the workings of our linear model will be introduced and referenced.
* IN
  + Interpret: In the interpreting section, we will summarize the results and provide the overall recommendation to the stakeholder.

To achieve the recommendation, a modeling technique of ‘Ordinary Least Squares’ (OLS) will be conducted. All in all, OLS is an optimization strategy that aids in finding a straight line as close as possible to data points in a linear regression model. The main predictor that will be used as the dependent variable in the analysis will be ‘TotalPay.’ The independent variables will be discussed and introduced in the following sections.

**Data Acquisition**

Multiple data sets were considered in the analysis. They are the following:

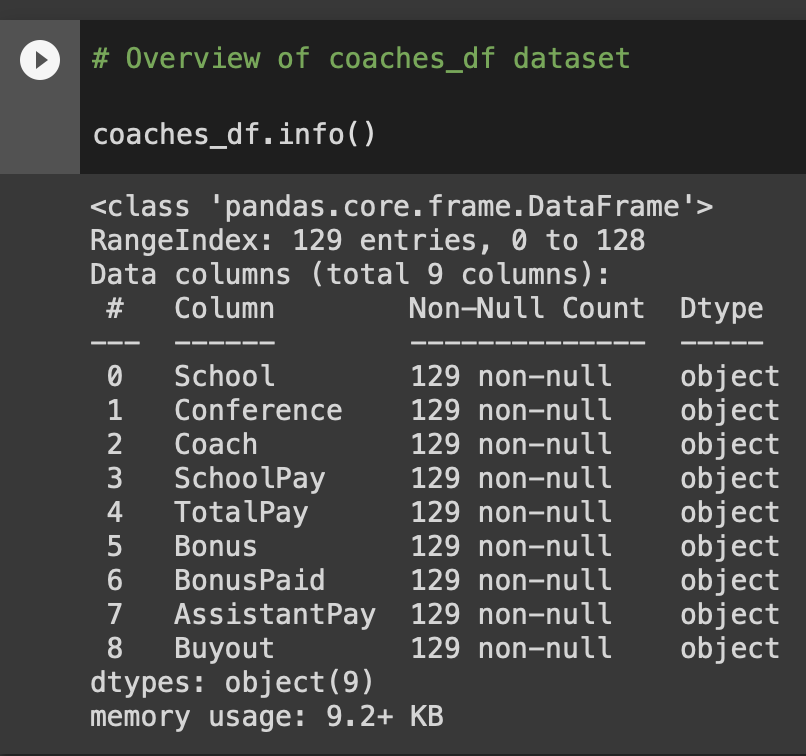
* Coaches
  + https://raw.githubusercontent.com/2SUBDA/IST\_718/68273222b88e35e1d7ef4f2c874e667ee64dca60/Coaches9.csv
* Stadiums
  + https://www.collegegridirons.com/comparisons-by-capacity/https://www.teamrankings.com/ncf/trends/win\_trends/
* Teamrecord
  + https://www.teamrankings.com/ncf/trends/win\_trends/
* Grad
  + /content/sample\_data/gradyear.xlsx
  + https://www.ncaa.org/sports/2016/12/14/shared-ncaa-research-data.aspx

Most datasets were able to be loaded with the pandas library directly from the web. For the ‘Grad’ datasets, this was downloaded directly from the ‘Shared NCAA Research Data’ site in excel format.

**Data Cleaning**

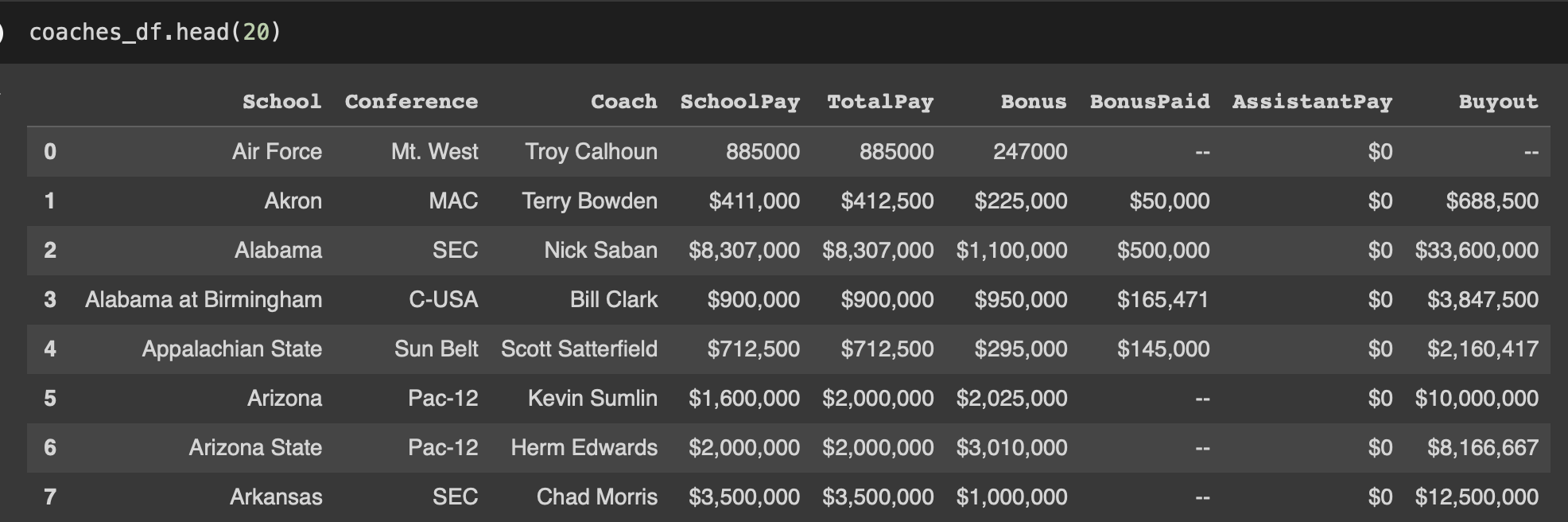
In the ‘Data Cleaning’ phase, each dataset was inspected and transformed as needed.

Coaches

In looking at the ‘Coaches’ dataset, the .info() method was invoked and provided the following results:  
  


This method allowed me to inspect datatypes as well as any null-values; it also provides an overview of the shape of the dataframe – calling out the number of rows and columns, respectively.

Although no nulls were technically found, the .head() method gave insight into a null-like value: the “--“.



As a result, the “--“ values were converted to null values; the null values were then replaced by zeros, and zeros were replaced with the mean of the respective numerical column.

The datatypes were reset for all numerical columns, converting from ‘object’ to ‘float.’

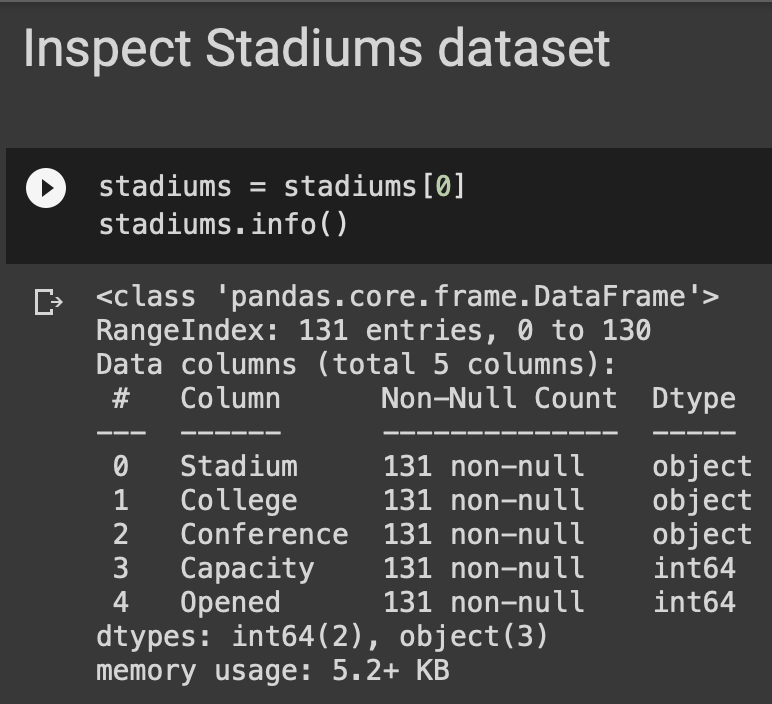
The ending result looked as follows:

Text

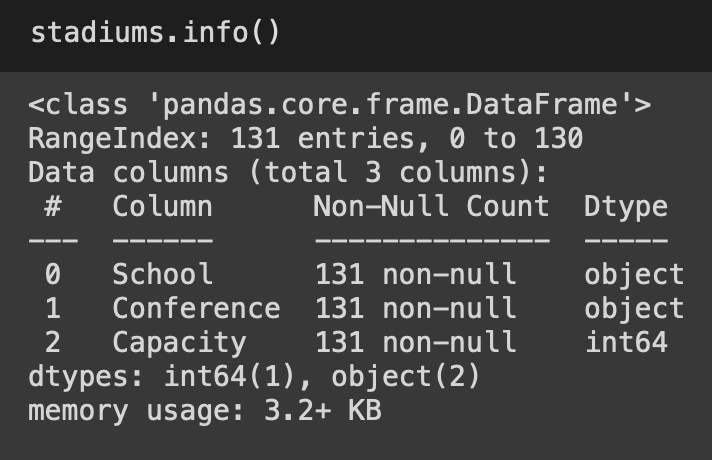
Description automatically generated

Stadiums

The Stadiums dataset was then inspected. Similarly, the .info() method was invoked to see an overview:



No null or null-like values were observed within this dataset. The datatypes also appeared to be satisfactory. In hopes to eventually merge all datasets together, the ‘College’ column was renamed to ‘School’ to conform to the ‘Coaches’ dataset. The number of columns were then reduced, as the actual Stadium names (as well as year Opened) were not needed for the analysis. The final dataset appeared as this:

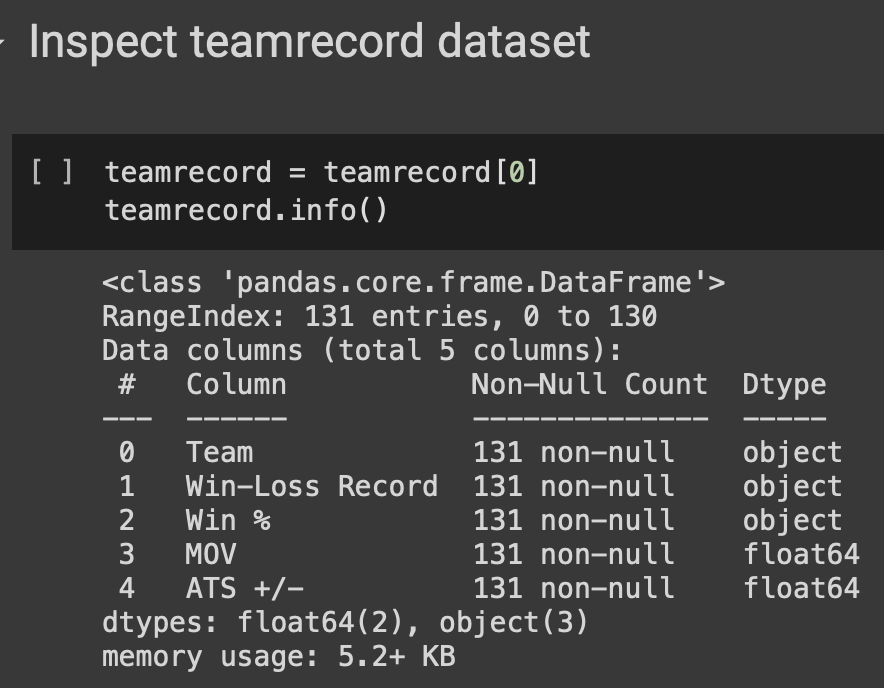


\*\*An import sidenote to introduce: School names appeared to have variations in spelling or abbreviation. This was not immediately transformed manually. Instead, a ‘fuzzy match’ that attempts to merge school names based on similarity score was used later in the analysis.\*\*

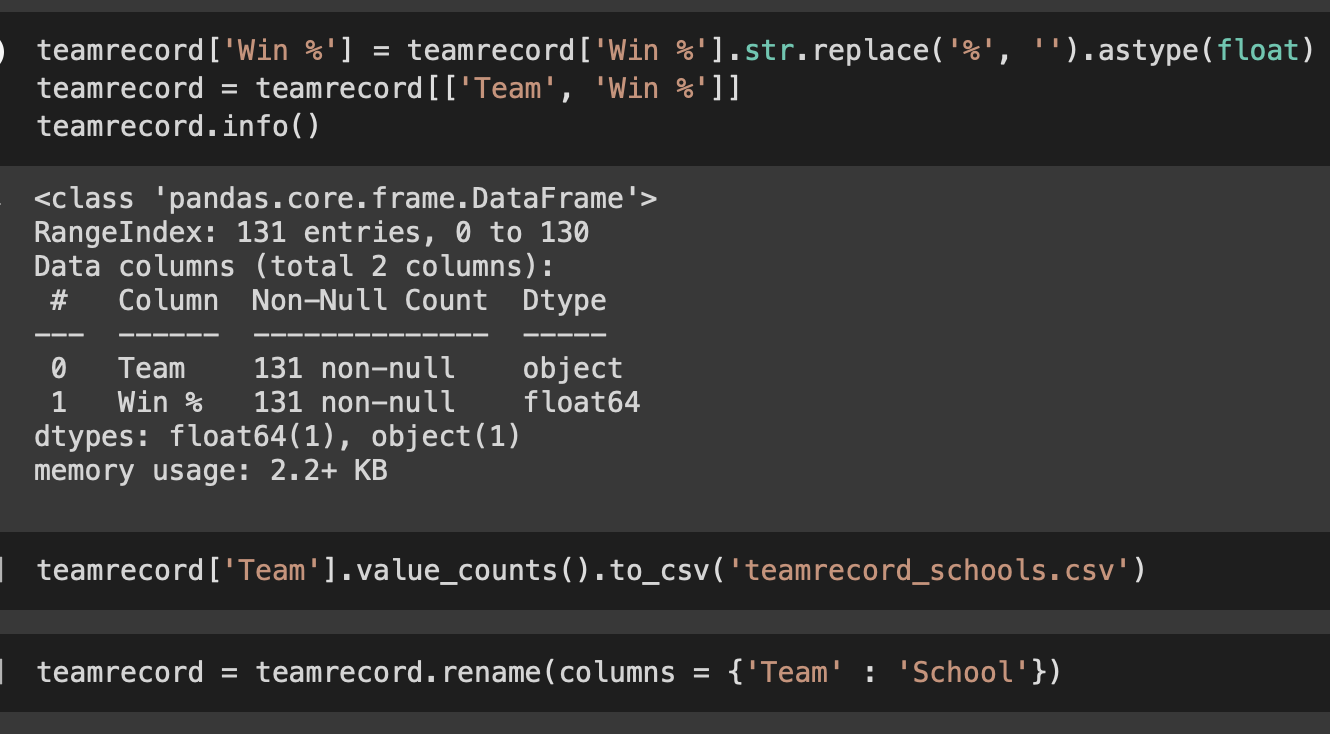
Team Record

As for the team record dataset, no observed null or null-like values were detected; however, the datatypes for winning percentage appeared to be incorrect. In addition to this, more columns were determined to not be required for the analysis, and were thus dropped. A before and after view can be seen in the figures below. In brief, the ATS, MOV, and Win-Loss Record were dropped, as ‘Win %’ was chosen as the main statistic for potential use as an independent variable in the model.

Before:



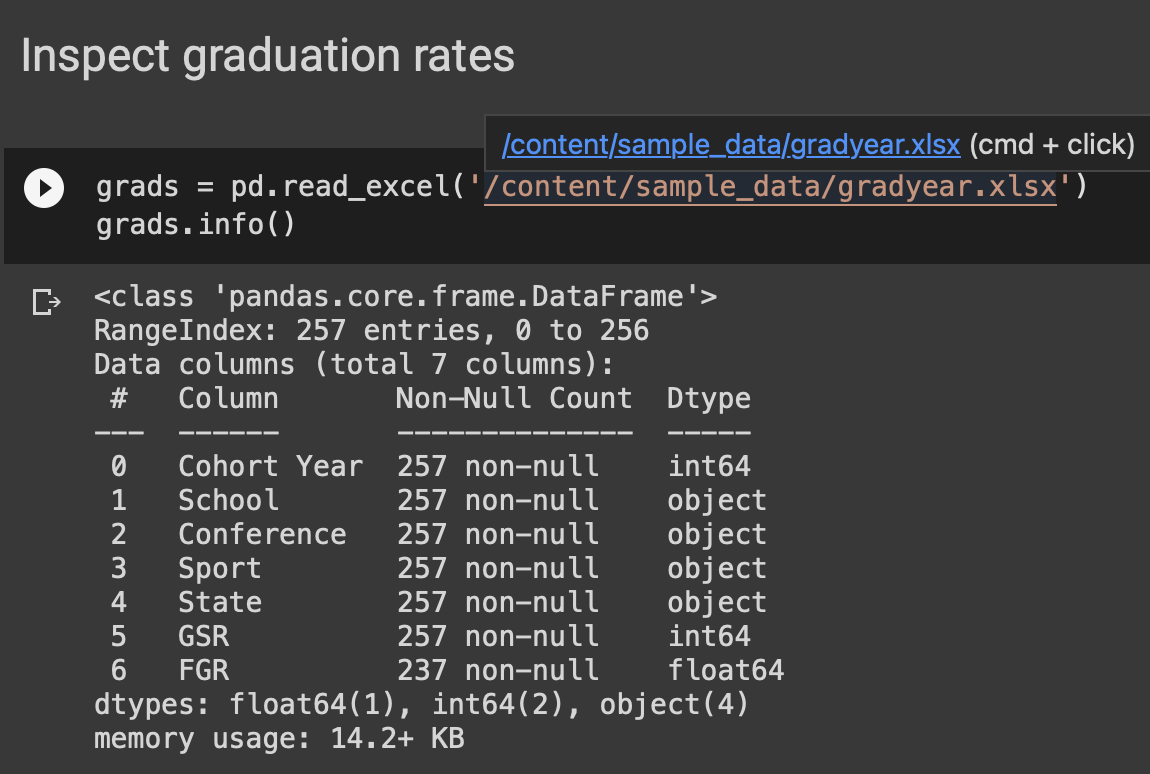
After:

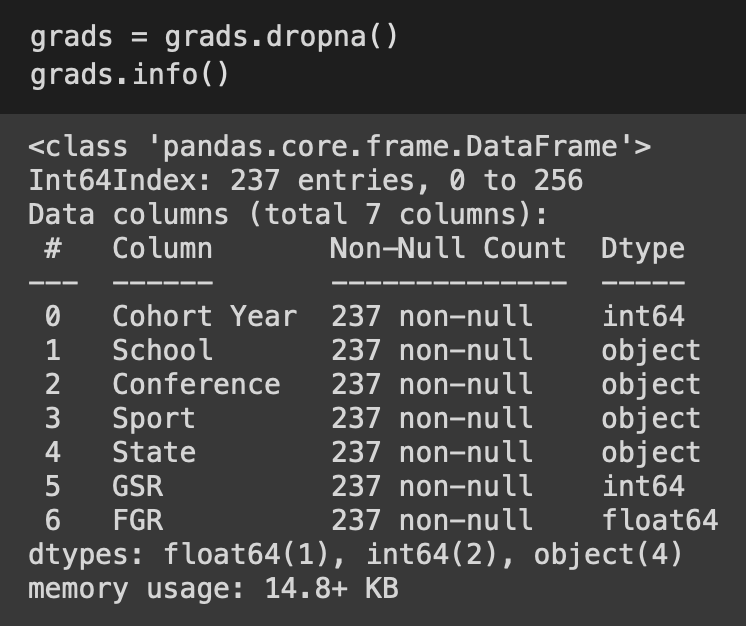


\*\*An import sidenote to introduce: School names appeared to have variations in spelling or abbreviation. This was not immediately transformed manually. Instead, a ‘fuzzy match’ that attempts to merge school names based on similarity score was used later in the analysis.\*\*

Grades

The final dataset used in the analysis included data on graduation rates. Similarly to the previous datasets, the .info(), .head(), and other methods were used to gather an overview and preview the data.



This time, null-values were detected in the FGR column name. These values were dropped and the datatypes were not modified. The ending results looks as follows:  
  


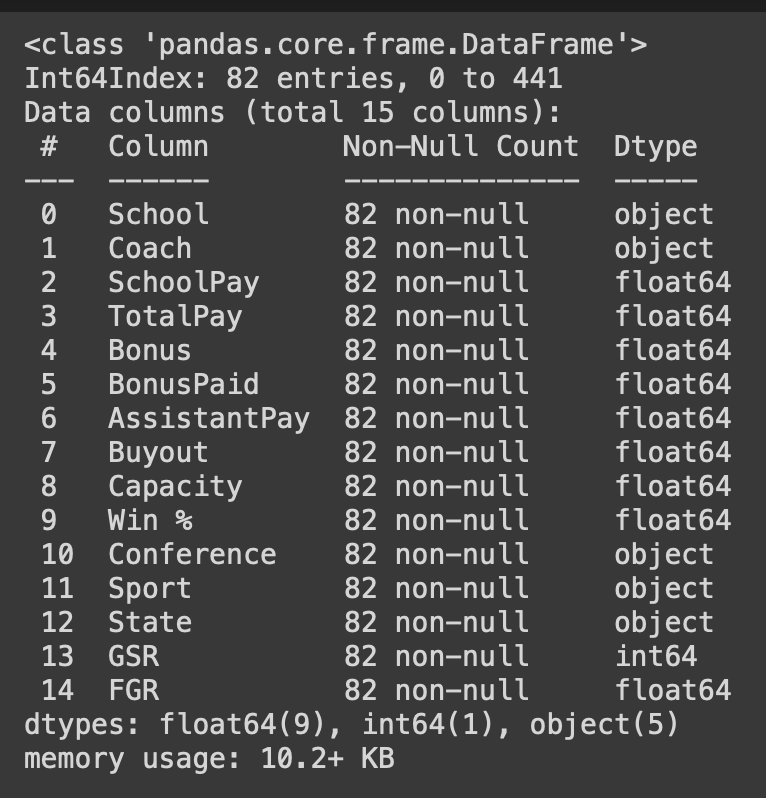
Merging into One, Overall Dataset

Since the goal of the datasets is to use everything in one analysis, a merging of the various data frames will be required for the next phase in the analysis. Rather than manually transforming each school abbreviation through code, a python library called ‘fuzzywuzzy’ was used to use a ‘partial match’ based on similarity. This approach is not always 100% accurate, but was chosen as an exploratory step in this analysis lab assignment.

The process for the partial matching is as follows:

* Derive a 'best match' determined by what’s known as the Levenshtein Distance, which is a measure of the difference between two strings
  + Each string would be the ‘School’ attribute instance in each data frame
* If a match is found, the school name found in the ‘Coaches’ dataset is replaced by the school names found in the Stadiums, Team Record, and Grads datasets

The end results is the following:



The partial matching via the ‘fuzzywuzzy’ python library was not completed with high data retention, as the amount of observations was reduced from ~130 to just ~80. Although this is not an ideal occurrence, it is believed that a reliable model can still be derived based on the final results of this data frame.

All data types in the final data frame are either object, float, or integer.

**Data Exploration**

Now that a final data frame was derived based on the various data sources, exploration can begin.

The exploration is started by checking the distribution of the dependent variable, ‘TotalPay’:

Chart, histogram

Description automatically generated

This looks to be positively skewed, since many of the observations are closer to zero, but a proverbial ‘tail’ is viewed trailing towards larger values. Some would argue taking the logarithm of this variable to convert the distribution type to ‘normal’; however, for the purposes of this exploratory lab, we will see if this negatively affects the results of our model.

Next, before checking the distributions of independent variables, we will gather correlation insights by deriving a matrix along with a heatmap (provided via the seaborn python library).

A picture containing text

Description automatically generated

Although this is helpful, a visualization would aid in a quicker analysis:

Chart, bar chart

Description automatically generated

Based on the correlation matrix visualization, the following independent variables will be trialed for the modeling portion:

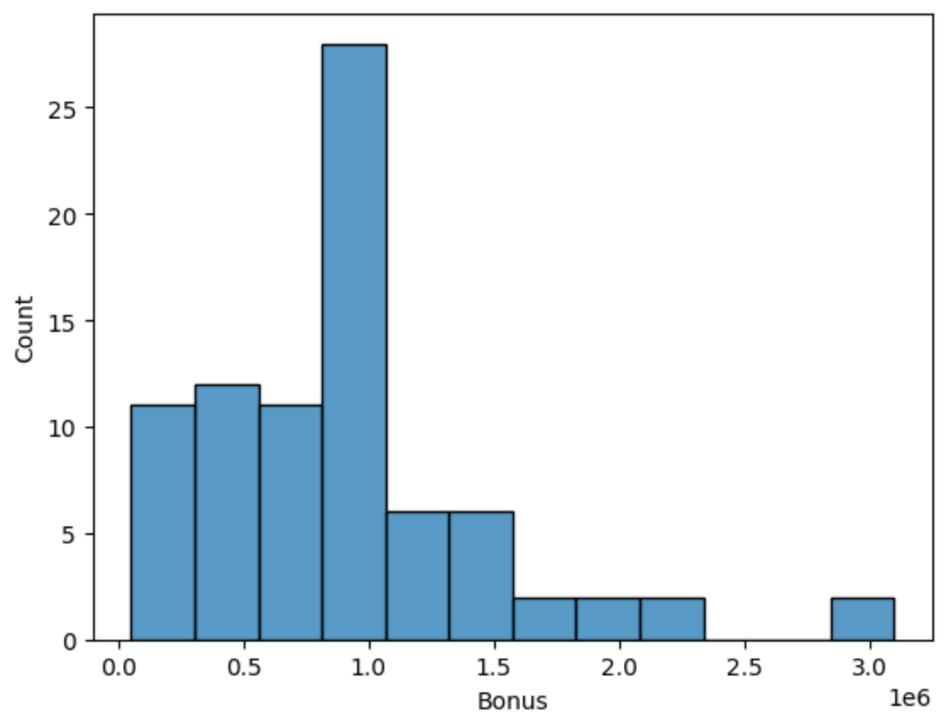
* Buyout
* Capacity
* Bonus
* BonusPaid
* Win %
* GSR
  + Chosen based on lab question set
* FGR
  + Chosen based on lab question set

Histograms were then derived for each independent variable chosen for the modeling of this data:

Buyout:  
Chart, histogram

Description automatically generated

BonusPaid



Capacity

Chart, histogram

Description automatically generated

Win %

Chart, histogram

Description automatically generated

GSR

Chart, histogram

Description automatically generated

FGR

Chart, histogram

Description automatically generated

To summarize, we generally see positively skewed or normal distributions for the independent variables chosen – this could affect the quality of our results in the modeling phase. Modifying some of these distributions with the logarithmic equivalent could improve accuracy.

Last but not least, a pairplot was conducted, also with the seaborn package, to see the relationship between the Y and X variables:  
  
Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Visually speaking, the GSR and FGR data columns from the ‘Grads’ dataset do not appear as though they have a strong relationship with the Y variable; however, since a question related to these data points is specifically in the lab 1 questions, they will be kept in the modeling analysis.

**Data Modeling**

As mentioned in the Overview section of this document, the Ordinary Least Squares (OLS) is a method for estimating unknown parameters in a linear regression model (i.e., prediction data). This modeling technique will be used to ultimately determine the salary Syracuse should offer their next football coach.

Text

Description automatically generated

Conclusions and insights from this modeling technique will be discussed in the next section.

**Conclusions**

Insights from this model can be derived by reviewing the following:

First, the probability of the F-statistic should be reviewed. Since this value is well below the threshold of 0.05, the model can be interpreted and is statistically significant.

Next, the R-squared value can be reviewed. The value of 0.942 indicates that the ~94% of the change in Y, our TotalPay variable, can be explained by the changes in our X variables – Buyout, Capacity, Bonus, BonusPaid, Win%, FGR, and GSR. This is generally a good sign and shows that we may be able to make a ‘good’ prediction with this model.

Following this, checking the validity of the variable coefficients is important. From our model output, it can be viewed that many of the p-value coefficients are above the 0.05 threshold, meaning that they are not statistically significant to the model. More specifically, the Bonus, Win %, GSR, and FGR all have p-values above this threshold. We cannot conclude that these specific variables have any significance to our model. Buyout, Capacity, and BonusPaid appear to be statistically significant, with p-values below the .05 threshold.

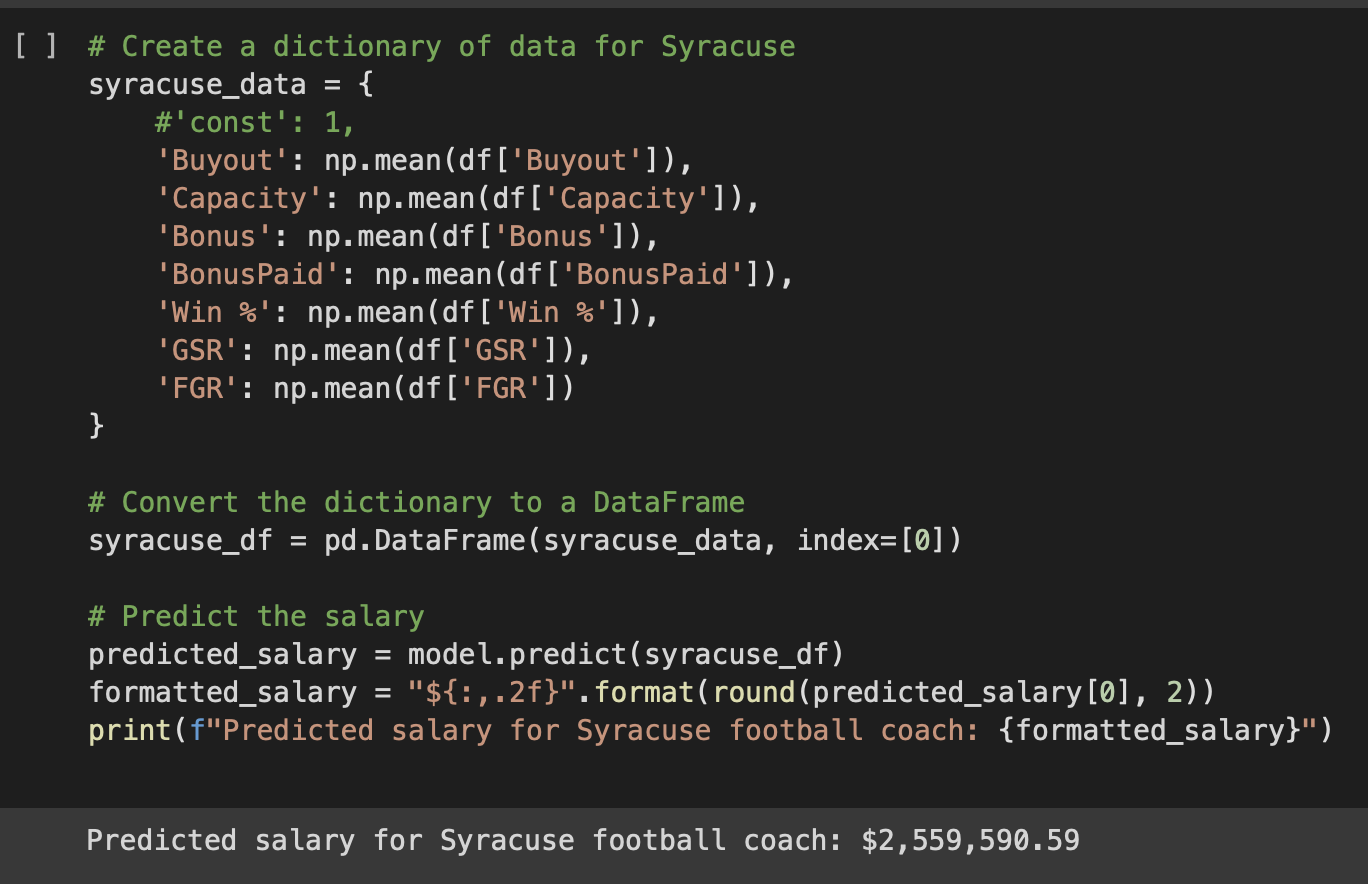
The largest statistically significant coefficient (and coefficient in general) is the capacity variable, which indicates a coaches pay increases ~$31 for each unit increase in a stadium’s capacity.

**Lab Questions**

Lab Question 1

What is the recommended salary for the Syracuse football coach?

To answer this question, a prediction will need to be made with the model developed in the ‘Data Modeling’ section.



Using the model, the predicted salary for the new Syracuse football coach is ~$2.5 million.

Lab Question 2

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

A prediction for each conference in the dataset was made into a table:  
  
Text

Description automatically generated

Although the Big East does not seem to be included in the dataset, the Big 10 predicted salary appears to be ~$3.6 million, which is significantly higher.

Lab Question 3

What schools did we drop from our data and why?

As mentioned earlier in the report, some school names were dropped due to the ‘fuzzywuzzy’ python library partial match function. Any school that was not able to be detected by the similarity partial match functionality of this package was dropped.

All in all, 82 schools/observations were kept in the dataset. The schools that were dropped were the following:



More details can be found in the supporting .ipynb (Jupyter Notebook) file on the 2SU submission page.

Lab Question 4

What effect does graduation rate have on the projected salary?

Based on the previously mentioned p-values for the graduation rate data (GSR, FGR), we cannot conclude these variables are statistically significant, since they are well above the 0.05 threshold – coming in at ~0.8 and ~0.6, respectively. We cannot confirm whether this data influences projected salary.

Lab Question 5

How good is our model?

As previously mentioned, the R-squared value for this model is ~0.94. This means that the ~94% of the change in Y, our TotalPay variable, can be explained by the changes in our X variables – Buyout, Capacity, Bonus, BonusPaid, Win%, FGR, and GSR.

It is a relatively high score, and indicates the variation in the data is explained (mostly) by these variables.

Lab Question 6

What is the single biggest impact on salary size?

The largest impact on salary size is the Capacity variable. As mentioned in the conclusion section, the variable’s p-value is statistically significant and the coefficient value is ~31, meaning a coaches pay increases ~$31 for each unit increase in a stadium’s capacity.