Matthew Powers, Tara Moore, and Clare McAuliffe

SI 370

2 December 2021

[**Link to Presentation**](https://docs.google.com/presentation/d/1bcx5DcfdNTe28AyQHxu4SjnIbP5aDdNSoyWXMXW3whc/edit?usp=sharing)

**Which Statistics Predict Performance the Most in the Big Ten and the SEC?**

**Background & Purpose for Project:**

Michigan Football has taken college football by storm this season, from starting unranked to now playing for a Big Ten Championship as the second best team in the country (and beating Ohio State along the way). How could this have happened? How did analysts let this great Michigan Football team slip through the cracks when evaluating teams in the preseason? What if there was a way to predict this outcome before the season? Are there certain statistics that will point towards a team’s success or failure? As assumed, a team’s total wins and losses from the previous season are good indicators for how they will perform, however, what if how many times a team rushes the football can help predict their season’s outcome? Or, are there intangible aspects that no one can predict?

We have used [college football data](https://www.kaggle.com/jeffgallini/college-football-team-stats-2019?select=cfb13.csv) from the 2014 - 2020 seasons to compile statistics and rankings from recent seasons in hopes of answering these questions. The data incorporates over 110 Division I FBS football teams with over 140 different statistics. These statistics measure teams’ offensive, defensive, and special teams performance. **Using this college football data, we will investigate whether there are certain statistics that can predict a team’s performance. If so, we will create a machine learning model to predict what the rankings for the 2020 season should have looked like, focusing on Michigan’s conference and division.**

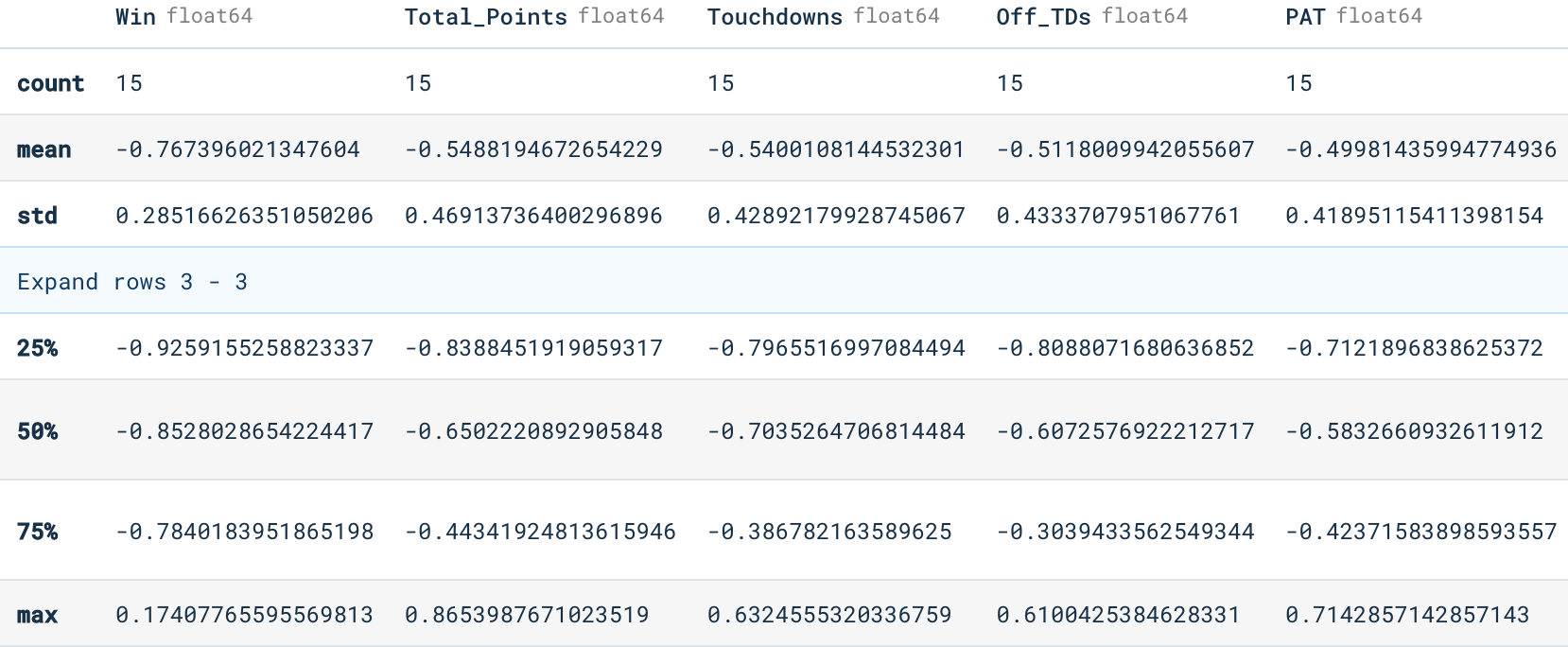
**Focusing in on the Big Ten and SEC Conferences:**

As there are many teams and conferences within college football, it is hard to incorporate all teams and factors into one model; different conferences have different styles of play and teams have changed conferences (ex: Rutgers joined the Big Ten in 2014). With that, our team chose to use the Big Ten and SEC conferences as we felt the two encompass different types of teams and have produced some of the best and worst teams in recent years where it will be helpful to compare and contrast.

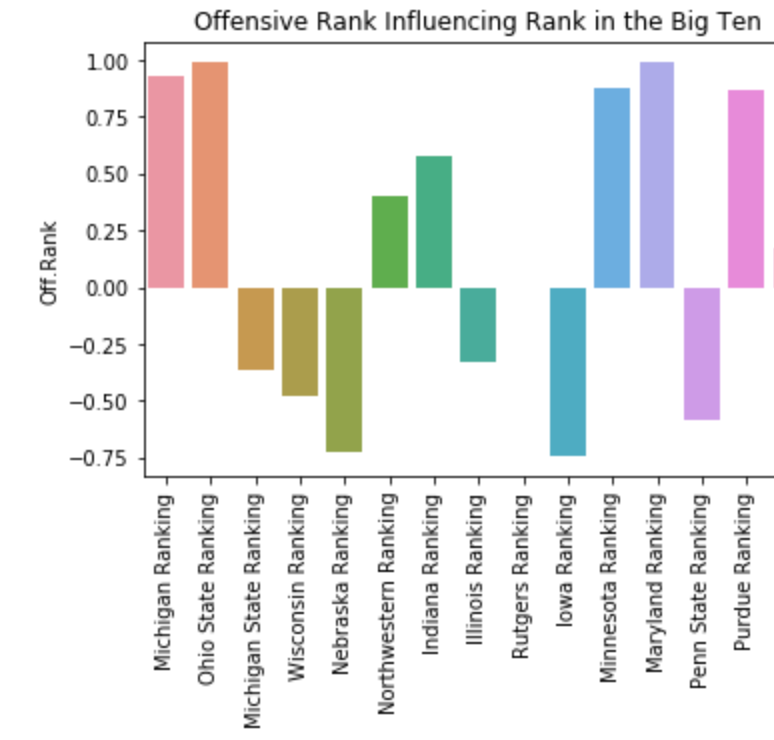
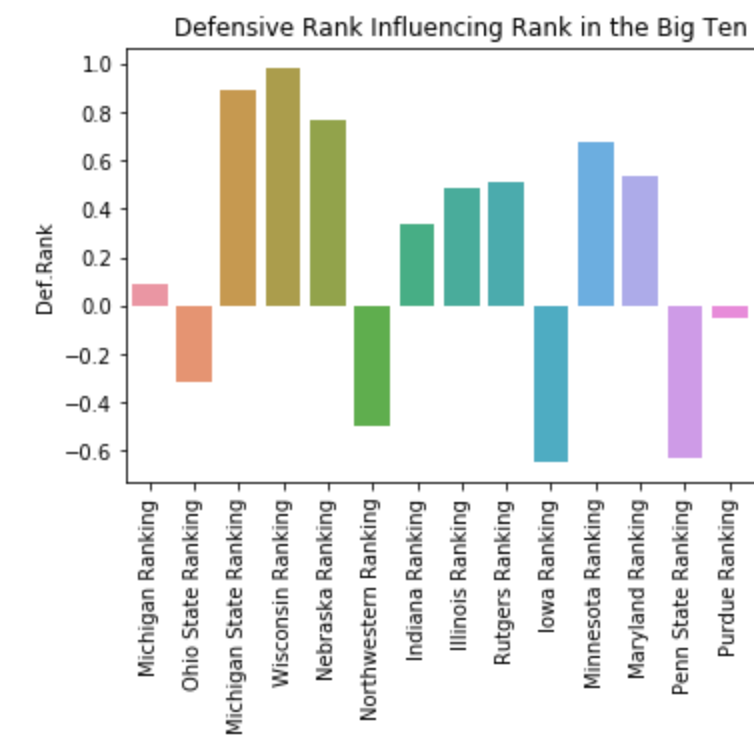
**Pre-Processing:**

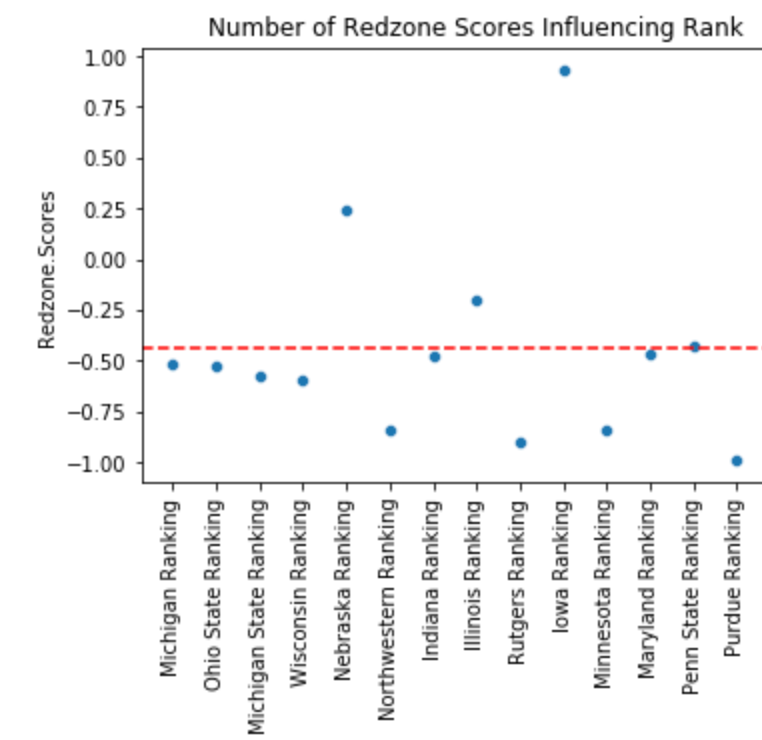
In order to evaluate the teams’ performance, we merged each csv into one dataframe to get all years together. Next, we split the “Team” column into “Team” and “Conference” columns to better categorize teams (ex: using .split(), “Michigan (Big Ten)” became “Michigan” and “Big Ten”). We were able to find the teams within the Big Ten and SEC and their rankings within each conference for each season from 2016 - 2019. In this process we also created a column that broke down the conference into divisions as well, for example, SEC East. We chose to use more recent years (with sufficient data points) to better represent the evolution of the conferences for 2020.

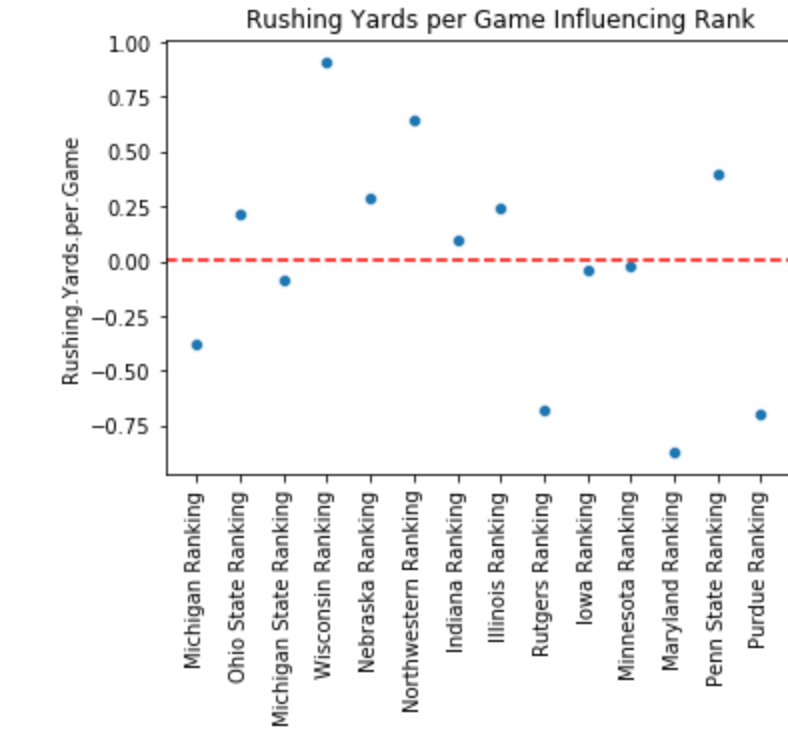
After that, we created a Big Ten dataframe and an SEC dataframe to move forward with our analysis. We found the correlation between a team’s ranking and all of the other variables to understand which statistics were the most influential in a team’s rank. For both conferences, we merged every team’s correlation with their rank and each statistic. In having all teams together, we took the mean correlation for each statistic with rank. This helped us to understand what the average correlation was between rank and statistic to identify the most influential statistic for each conference. In future planning this will help build the machine learning model to show what the 2020 season should have looked like if the season was not affected by COVID.

**Big Ten Analysis:**

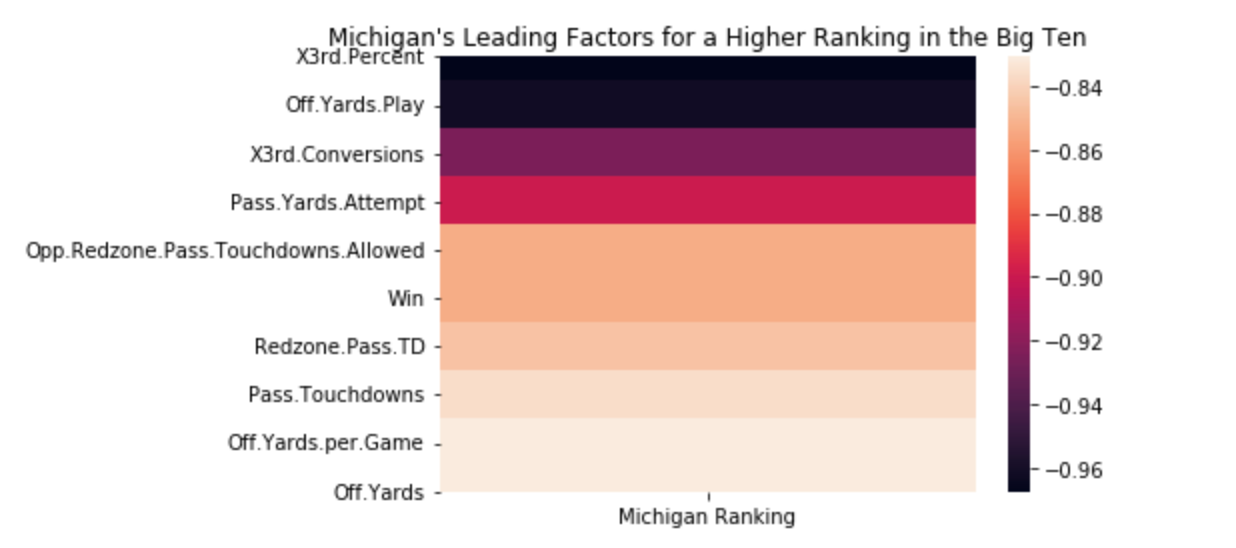
Above, we used .describe() to get a statistical summary of the correlation between ranking and statistics for each team to understand the distribution of the values. Between each statistic, the spread of values vary. For example, the standard deviation for “Total\_Points” is almost double the standard deviation for “Win.” This means that some statistics are more influential throughout the whole conference, while some are just significant for certain teams.



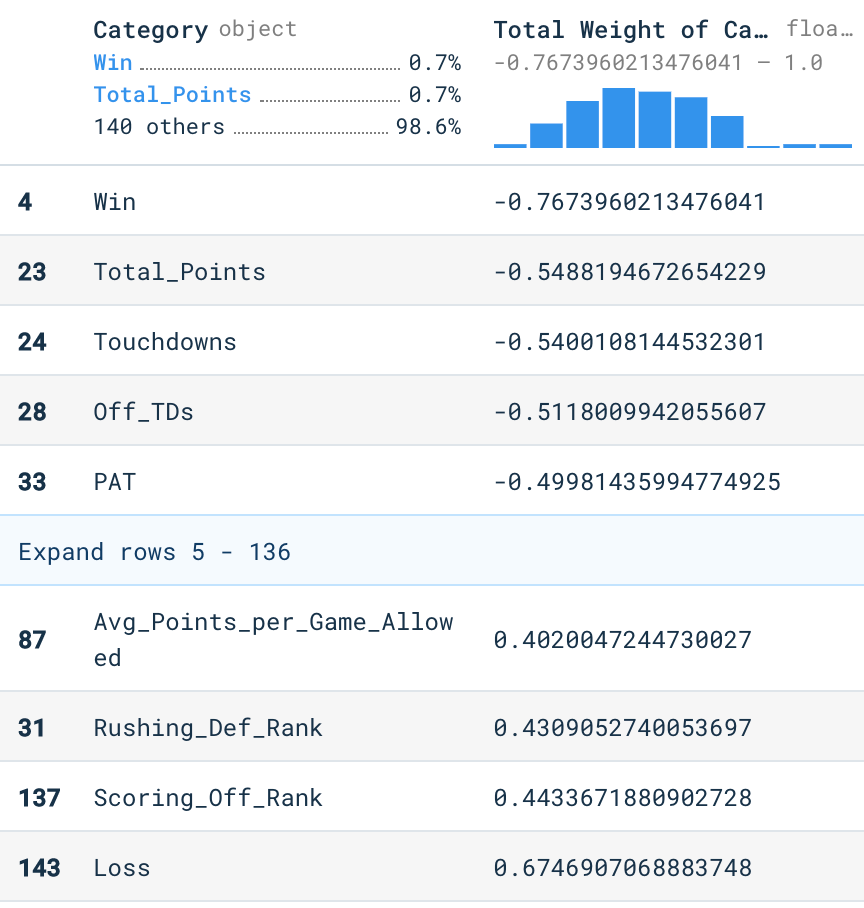
With that, the graphs above show each team in the Big Ten’s “Offensive and Defensive Rank” correlated with their conference ranking. As we saw in the statistical summary, the ranks have a wide spread, meaning some of the teams’ offensive and defensive rank didn’t necessarily show how they were going to be ranked in their conference. This can be explained through an upset, as a team with an overall worse defensive rank beating out the best team in the conference.

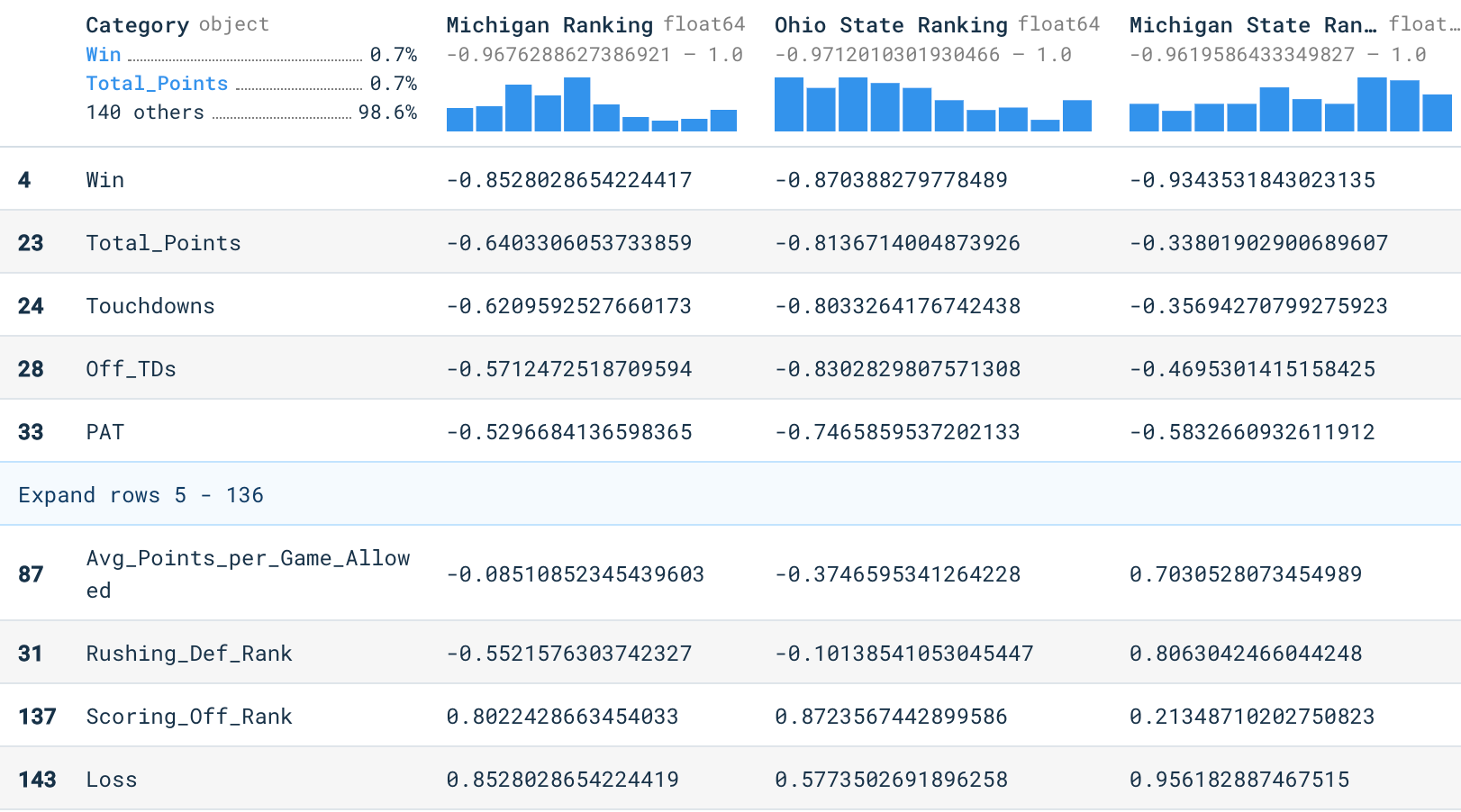


The redline shows the average correlation between each of the two statistics (Number of Rushing Yards per Game and Redzone Touchdowns) and each team’s ranking above. While each graph has its variability, we see that there is a major difference in average correlation; Rushing Yards per Game has a 0.00 correlation, which shows that overall there is no relationship while Redzone Touchdowns has a correlation of about -.500 showing the indirect relationship. This shows that there are better statistics to use to more accurately predict rankings in the Big Ten.



The heat map above shows the top ten statistics that have influenced a better ranking for Michigan in the Big Ten. The biggest factor is the team’s ability to convert on 3rd Down (“X3rd.Percent”), which makes sense as converting helps keep offensive drives alive. It is interesting to note that the majority of the top ten statistics that contribute to a better ranking for Michigan are related to offensive performance.

In finding the correlation between each variable and rank, the relationship is either positive or negative. A negative correlation is associated with better rankings as it shows that a higher statistic connects to a lower ranking in the conference. A positive correlation is associated with a not as good ranking as it shows a higher rank statistic (ex: Points Allowed Per Game) connects to a higher ranking in the conference.



Above to the left (in “Total Weight of Category”) shows the top five statistics for a better Big Ten ranking and the top five statistics for a worse Big Ten ranking. The top five statistics shown above for a better Big Ten ranking are expected, as they are associated with scoring more points (“Win” - “PAT”). However, as we moved into the top 6 - 10 variables, we found that variables such as “Passing Yards Attempted”, “Rushing Attempts”, and “First Down Passes” were very influential in the Big Ten ranking as well. On the flipside, the top five statistics shown above for a worse Big Ten ranking are expected as they are associated with allowing more points. For example, a high Rushing Defense ranking means that a team is allowing a lot of rushing yards and touchdowns; this means they are more likely losing games (leading to a worse ranking). The right visual above shows some specific team correlations that were used in calculating the average to the left. As seen, some teams had stronger correlations than others for certain statistics. It was important to identify this because finding the average most influential statistics would lead to more accuracy than basing it off of 1 or 2 teams.

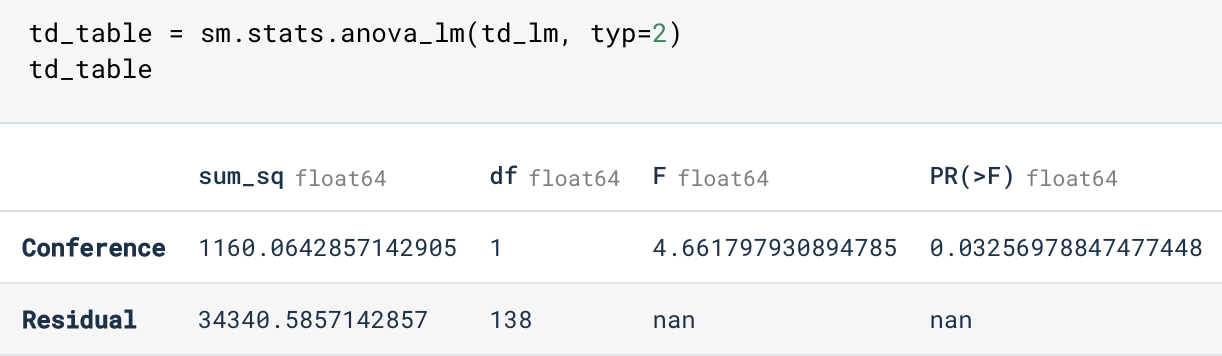
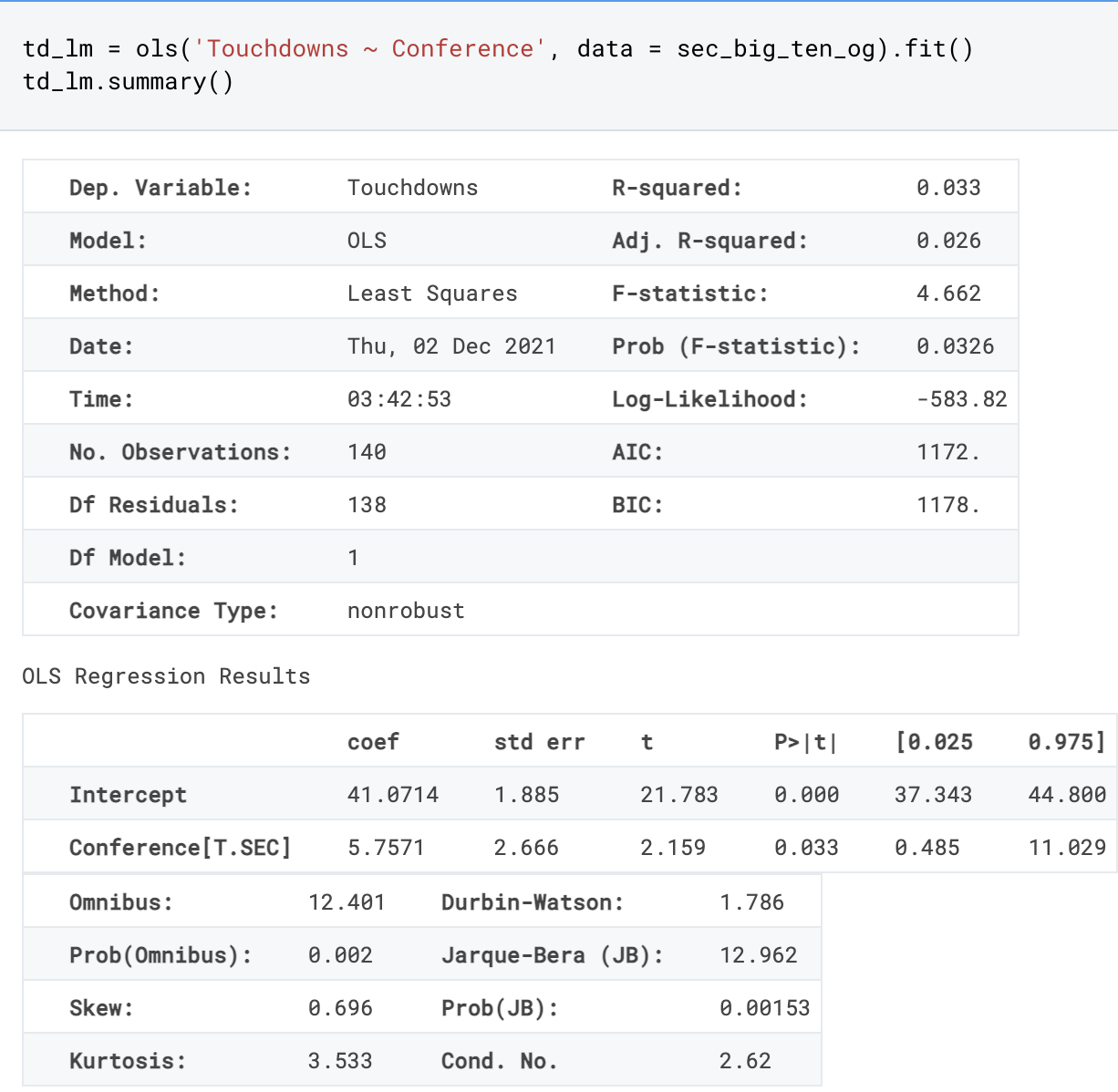
From finding the statistics that are linked to the strongest correlation with the teams’ ranking, we can use these variables to predict what the 2020 football season should’ve looked like. It is important to use these variables, and exclude others because it will help us create a more accurate model.

**SEC Analysis:**

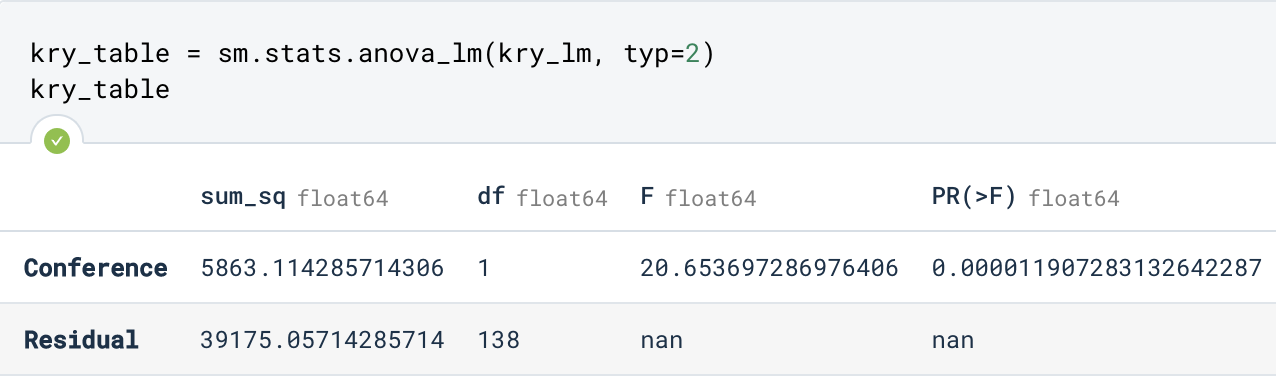
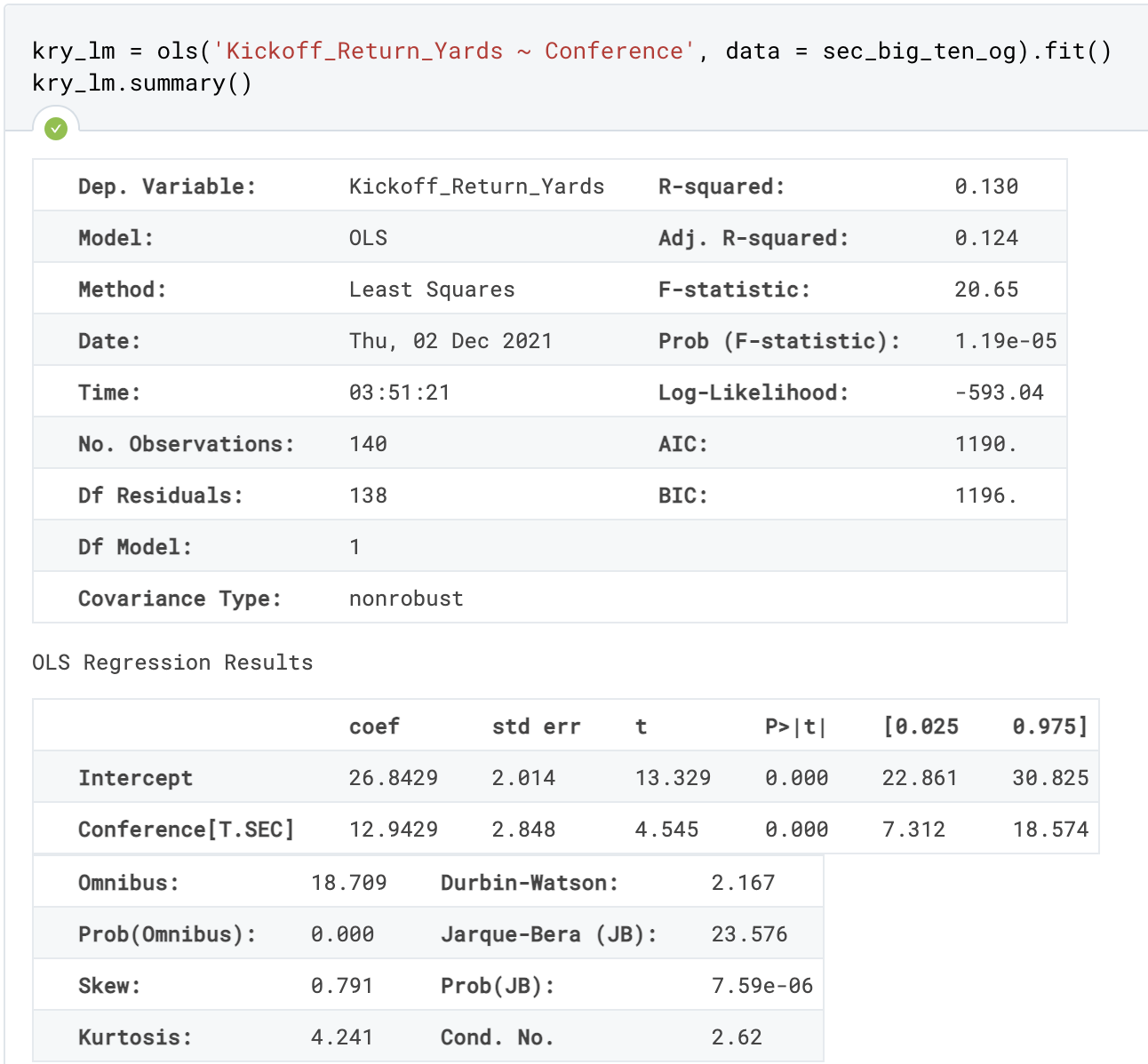
Like the Big Ten, we used .describe() to statistically summarize the correlation within the SEC between each team’s rank and statistics. The average correlation with “Win” is almost identical to the Big Ten’s, which makes sense; there is also variability between the statistic’s influence as well.

**Difference in Statistics Influence in SEC vs. Big Ten:**

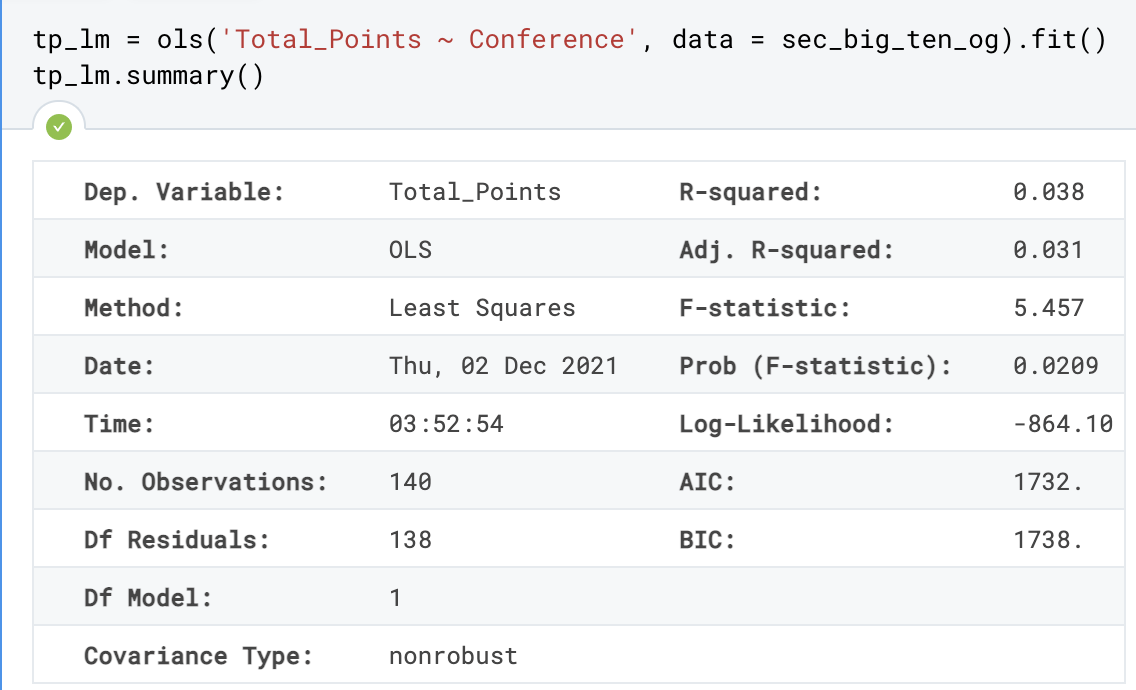
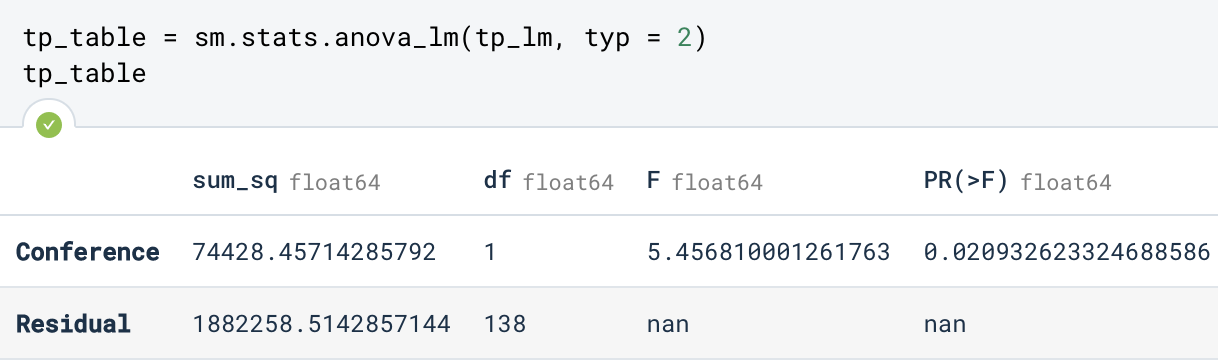
Knowing that the SEC and B1G conferences have different rules, style of plays, stadiums, weather, environments, etc., we hypothesized that there would be differences in their correlations, and which factors would have the biggest influence on their rankings. In order to test this hypothesis, we ran OLS Regressions and ANOVA tests for a few of the factors.



With an R-Squared of 0.033, we know that over 3.3% of the total variance can be explained from this data. While this wasn’t incredibly significant, we felt it showed a decent amount of variation. We also have a low p-value of 0.0326, which means that it is statistically significant, under 0.05, and the Touchdowns vary by Conference.

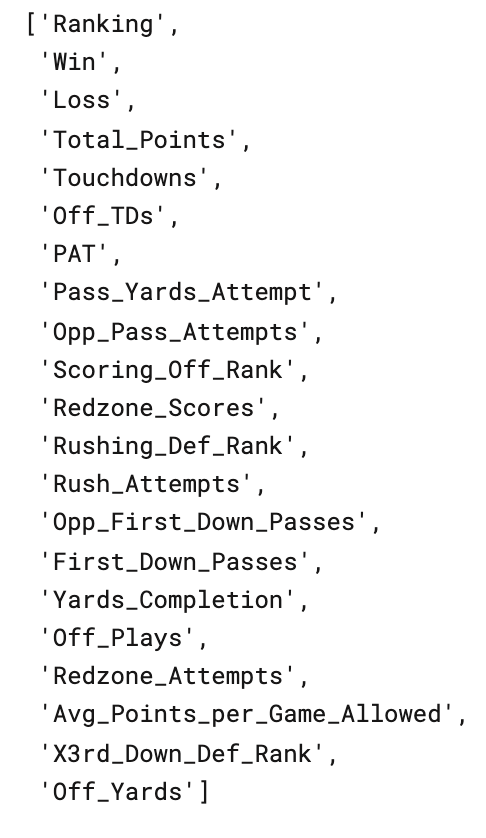
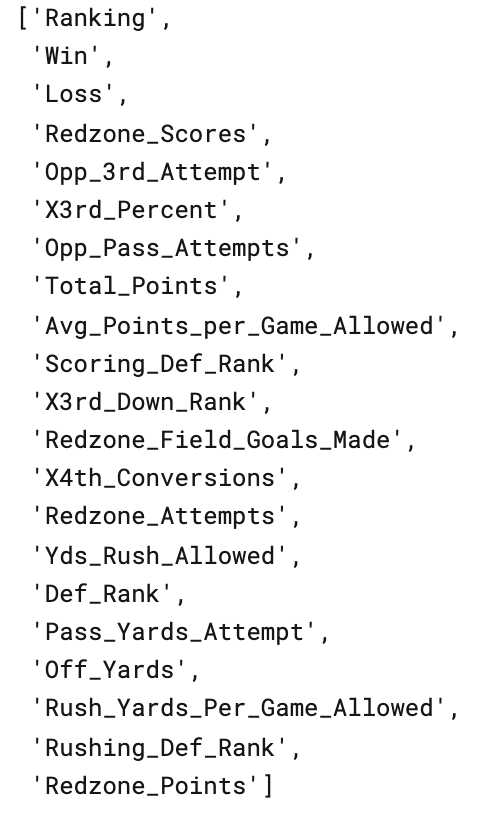


With an R-Squared of 0.130, we know that over 13% of the total variance can be explained from this data which is a significant amount. We also have quite a low p-value of 1.19e-05, which means that it is statistically significant and the Kickoff Return Yards vary by Conference.

****

With an R-Squared of 0.038, we know that over 3.8% of the total variance can be explained from this data which, similar to touchdowns, was not as significant as it could be, but showed variation. We also have quite a p-value of 0.0209, which means that it is statistically significant and the Total Points vary by Conference.

From doing the OLS Regressions and ANOVA tests, we could conclude that there were variations between the two conferences and that there were different variables that held stronger weights between the conferences, and likely the team’s rankings. We wanted to take it a step further and see if the top variables from each conference were the same. To do this, we created absolute values from the correlations and then took the highest ones. After this, we created lists of both to take a closer look and make it more digestible to understand what variables were different against conference.

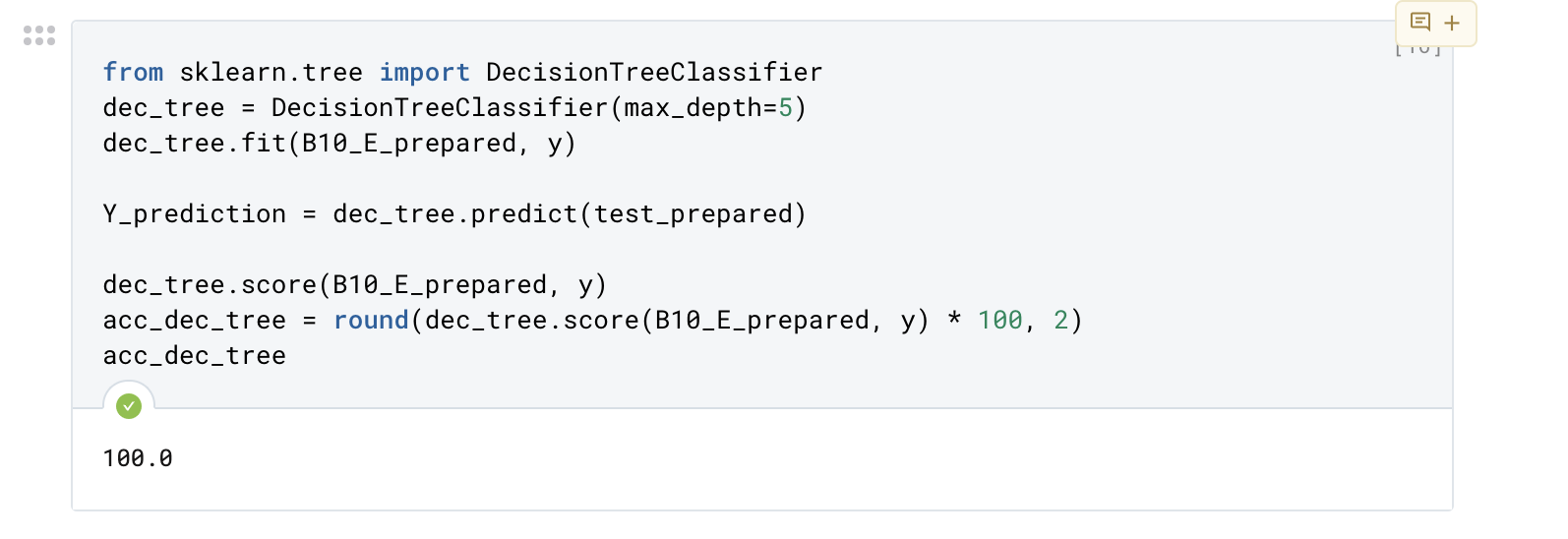
**B10 SEC**

This ultimately led to our decision to not use a uniform model or set of variables as there is a significant enough difference between the two that would lead to inaccurate predictions. From here, we began our machine learning process.

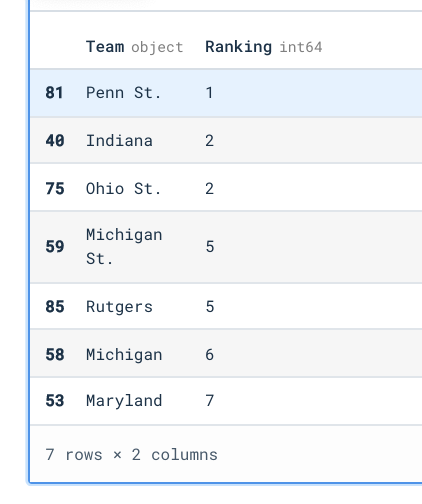
**Machine Learning Set Up to Predict the 2020 Season:**

After we were able to clean the data and do some exploratory analysis, we then created a machine learning model to predict the 2020 rankings. Because each division had their own rankings, we decided to do the model built of the Big Ten East bracket (Obviously Go Blue!). We saw in our previous analysis that different variables held a stronger weight on the correlation with each team's rankings, and gathered the top 20 variables for each conference.In order to create the machine learning model we initially started by exploring our B10 and SEC dataframes from the years 2016-2019. We then condensed this down to just teams in the B10 eastern division because of analysis prior. Once completed, we took our dataset of 28 rows and used a stratified sample to split the data into training and test. This was done in order to make sure there was the same distribution of teams in each set. We resulted in a training set of 21 teams and a test of 7. Following this, we then excluded columns that did not have a high correlation to rankings. We used the list above (excluding ranking) to create an X variable, while our Y\_train variable was y = pd.DataFrame(strat\_train\_set.Ranking), given we want to predict the rankings.

Because these columns differ tremendously in range we wanted to make sure each column had the same weight as others. In order to do this, we used a pipeline that implemented StandardScaler() on all the numerical values we planned on using. This had to be done on our training and test dataset, in addition to the 2020 file we will predict once the model is finalized.

To start we did a linear regression to see the variance we could expect in our predictions. This resulted in our mean square error being .21. Afterwards, we created a logistic regression and decision tree model to see which one would be more accurate in our predictions. Both of these resulted in an accuracy of 100.0. 

Finally, after creating the model we proceeded to predict the 2020 rankings for the B10 East. We first read in the file created earlier that had the statistics cleaned and included the conference they were in. It was important that we pulled a copy of the data prior to adding the rankings since this was what we wanted to predict. To ensure the 2020 data was in the same formatting of the training and testing set, we used the pipeline and reduced the columns on this dataframe as well. After this, we were able to do a .predict() on the data using our logistic regression model and generated rankings. We finalized our conclusions by creating a dataframe of the teams’ names and their rankings, and then exported to a csv.



**Our Model Prediction: Actual Ranking:**

Shown to the left, our model predicted what the 2020 season should look like based on the statistics of the team in the Big Ten East. While there are some differences, our model was able to predict Indiana, Rutgers, and Michigan correctly. For future analysis we could create a model that only predicts unique rankings, however this shows that past statistics make teams compatible. Conferences are split up and want to provide competitiveness for fans, the idea that teams had similar stats in the past affects our models and displays this.

**Challenges Along the Way:**

* We had to replace the “.” with “\_” in title name → “Total.Points” since the variable was being read as a dataframe and when we were doing ols it was referring to a column value instead of the variable.
* Team names had spacing before and after, therefore we had to strip them.
* Team names for the raw data were different then the names listed in the rankings found online. For this we had to cross check and change names to match, and ensure that the length of our merged data frame was the same length as our list of rankings.
* Some teams change conferences so it is hard to get an accurate representation of a larger set of past seasons, for example, Rutgers in 2014. Because of this we had to limit our machine learning model to 2016-2019.
* Too many outside factors and considerations to do all of the team → too many teams and conferences.
* The machine learning model predicts duplicate rankings. Efforts were made to try and change this, however it displays the reason some teams are close given they have performed similarly in the past.

**Considerations for Future Analyses:**

1. **Strength of Schedule**

With using this analyses and model for future work, a big consideration is to tie in confounding variables which could help make the prediction more accurate. For example, some teams in the Big Ten may play higher rank teams meaning their strength of schedule is higher. Creating a strength of schedule multiplier could help increase (or decrease) a team’s rank to show a more accurate depiction. This is important because a team that plays a lot of worse teams may have higher statistics but may not be the better team. In addition, between seasons coaching staffs may change which could also influence how a team performs over the years.

As this analysis is explored deeper, it may be shown that there are other statistics that are more influential and can predict rankings better.

1. **ML Model Predicting Winner of Specific Game**

Another model that would be interesting to create with similar data is one that predicts the winner of a single game. We could use the key variables that correlate to rankings found in our analysis. After this, the model could be trained on games played during that current season. This model would take in two teams, for example Michigan and Ohio St., stats for the current season and then predict a winner in terms of Yes or No.

1. **Incorporating More Variables/Statistics Outside of On-Field Performance**

With this current model, we are incorporating statistics on how the teams perform on the field. However, what if there were more ways to evaluate how a team is performing? For example a future model could consider the ranking of the head coach or assistant coaches. It also could consider how many players left for the NFL Draft and starters who left. Originally we thought the count of graduating seniors would help, however walk-ons and seniors who do not play that much will not affect the overall team's performance. Another statistic that would be helpful is knowing how many starters are returning. Lastly, we could include the number of preseason injuries; a team could lose their star wide receiver to an ACL tear in the preseason, so while their statistics could point towards a fantastic season, realistically there's a higher chance of not reaching that threshold of success.