MEASURING ACADEMIC PERFORMANCE ACROSS NCAA DIVISION I ATHLETICS

Sam Beyer
DATA 360
Dr. Stonedahl
April 29th, 2019
Individual Visualization Project

I. Dataset

   a. The dataset was obtained from Kaggle (https://www.kaggle.com/ncaa/academic-scores). It is titled *Academic Scores for NCAA Athletic Programs* and contains every NCAA Division 1 athletic teams' academic scores from 2004 to 2014.
   b. The dataset is 6511 rows by 57 columns. Each row contains the data for a single athletic team at each academic institution. The data in each row includes the athletic teams' APR score, eligibility percentage and retention rate from 2004 to 2014 as well as their 4-year APR score, 4-year eligibility percentage and 4-year retention rate. Each row also contains the teams' athletic conference.
   c. Each individual student athlete on the team calculates a teams' APR score. A student athlete earns 1 point for staying in school or graduating and 1 point for remaining eligible (http://www.ncaa.org/about/what-apr).
   d. The teams' eligibility score is calculated by the percentage of student athletes on the team that are eligible to play in every game in their sports' season. A player can become ineligible due to poor grades and/or failing classes. Thus, a higher eligibility percentage indicates a higher amount of student athletes on the team passing their classes.
   e. A teams' retention rate is calculated by whether or not each student athlete on the team graduates or stays in school. Thus, a higher retention rate indicates a higher percentage of student athletes staying in school or graduating on each team.

II. Questions

   a. On average, how have APR scores changed from the first year of data collection (2004) to the most recent data collection (2014).
   b. On average, do the top academic performing sports teams consistently have high APR scores? On the flip side, do low academic performing sports teams consistently have low APR scores?
   c. How do the athletic conferences compare when considering APR score? Which conferences perform the best over time and which ones are most consistent?

III. Coding and Cleaning the data

   a. School ID, School Type and Sport Code had no impact on my analysis
   b. Removed all sports that were either division 2 or division 3. There was not enough data on D2 or D3, which is why I decided to remove them to make it a division 1 study.
   c. There were several cases in which data was not present. In these cases, the data was listed as "-99". Because I wanted to possibly do time series data, any missing data points would skew the results. Thus, I converted the -99 values to NaN and then removed them from the data frame.
   d. From here, I grouped the data in three different ways. In each way, I took the average of all the values based on how I grouped them.

  i. By Conference: I grouped the values by athletic conference. This gave me an idea of which athletic conferences, as a whole, were academically better than other conferences.

  ii. By School: I grouped the values by academic institution. This gave me an idea of which schools' athletes performed the best academically across all sports.

  iii. By Sport: I grouped the values by sport. This gave me an idea of which sports, in general, performed the best academically. This meant, for example, that every school's football team's academic performances were measured against another sport – for example women's soccer.

e. I decided the most interesting data fell within the by school grouping. I wanted to see which schools, on average, had the best APR scores. I also wanted to see which schools had the worst APR scores. So, I took the top 10 APR scoring institutions and the bottom 10 APR institutions and created their own data frames.

f. I wanted to see if the top 10 APR schools progressed a lot from 2004 to 2014, and also if the bottom 10 APR schools had either progressed or digressed from 2004 to 2014.

g. I merged the bottom 10 and top 10 APR schools to create a data frame of the top and bottom 10 ARP schools.

h. From here, I created a scatter plot comparing each of the institutions 2004 APR score versus their 2014 APR scores.

  i. As you can see in the chart, the high APR schools are clustered in the corner with not much of a difference between 2004 and 2014.

  ii. Meanwhile, the low APR scores are located in the bottom left corner (as expected) but very scattered, meaning some institutions changed for the better and still were in the bottom 10, while some institutions APR score got worse.

i. From here, I decided to add another element to the graph. The size of the dot on the graph also indicated the eligibility percentage of a certain academic institution. Thus, a correlation is shown between APR scores and eligibility percentage, because all of the top 10 institutions have a larger dot than the bottom 10 institutions.

j. I used the seaborn package to graph the data.

k. Next, I wanted to look more into the by conference data. I decided that there are too many conferences, 33 to be exact, to group by conferences. So, I grouped them by the "Power 5 Conferences" which includes the ACC, the Pac-12, the Big-12, the Big Ten and the SEC. These conferences are the most prominent in athletic terms so I wanted to see how they compare to each other in academic ratings. Also, I included the Ivy League in this grouping considering the Ivy League is known for its' academic success.

l. Next, I melted both of my datasets in order to view the APR scores over time. This was very useful considering I was unable to plot the data as a time series prior to melting the data.

m. Finally, I cleaned up all my datasets by renaming the column names so that they were easily readable when they were put into the graphs.