Narrative

Jasmine Sanders

05/10/2020

Brief substantive background

Grounded in the principles of talent and personal effort, competitive sport has been heralded as one of the few remaining meritocratic societal structures (Loland 2007). Specifically, college athletics has historically been seen as a vehicle for marginalized students to realize their postsecondary dreams. Yet, this perception contradicts the shifting trend in collegiate sports. Recent data from the NCAA (2016) reveals a steep decline in first-generation college athletes across nearly all Division I sports, but particularly in men's basketball, where first-generation college athletes plunged from 28 percent in 2010 to 19 percent in 2015. This final project is a part of a larger mixed-methodology study that aims to identify the mechanisms of inequality contributing to the decline in first-generation college athletes on Division I men's basketball teams, specifically probing the role of the NCAA's Academic Progress Rate (APR) policy.

Drawing on insights from organizational theory (Meyer and Rowan 1977; DiMaggio and Powell 1983), particularly Tilly's (1999) model of durable inequality, which suggests organizations "rarely set out to manufacture inequality" rather unequal access arises as an unintended consequence of their attempt to address other organizational problems, I posit that a byproduct of the NCAA's increasing focus on APR has made coaches more hesitant to take on the risk of academically borderline student athletes. In addition to state-of-the-art training equipment, top-notch coaches, and well-resourced athletic facilities, elite private and preparatory academies are often known for their high academic standards and access to rigorous coursework, making student athletes of these institutions presumably less of an APR threat for Division I teams. Since students from disadvantaged backgrounds are less likely to have access to high-quality educational opportunities, the NCAA's APR policy consequentially impacts students from low socioeconomic backgrounds and/or students who identify as the first in their family to attend college.

There is no publicly available data that identifies NCAA players by high school type, nor have existing literatures empirically addressed this gap. This project utilizes webscraping techniques to create an original dataset that tracks high school and college characteristics of Division I men's basketball players from 2010 to 2020. By analyzing patterns in students' high school type (public vs. private), I can identify NCAA recruitment changes that potentially shrunk opportunities for first-generation college students.

Research Goals

1. High school type trends of NCAA recruits from 2010-2020
2. Top producing high schools of NCAA talent (top 10)
3. Top colleges drawing in NCAA recruits (top 10)

Main Challenges & Solutions

1. Webscraping: To collect student athlete data for NCAA athletes, I scraped data from ESPN's Top 100 Recruiting Database (http://www.espn.com/college-sports/basketball/recruiting/playerrankings) for years 2010-2020. Initially, I planned to scrape the roster pages of a randomly selected group of NCAA Division I basketball teams from 2010-2020. However, it would be an extremely time intensive and laborious process to it would be scrape a different website for each school. The website I selected addresses that issue by having a collection of data for the range of desired years in one place. While the

website addressed one challenge, it presented others. Despite the ESPN data being all in one place in a table format, it was not formatted for extraction in the way we've utilized webscraping techniques in this course. The contents of the table were separated into even and odd rows, and the row information was embedded within individual urls. However, I was able to use BeautifulSoup to parse the html data from the even and odd rows, then combine them into one data frame. From there I was able to loop the code to extract the necessary data from each individual athlete url.

2. Looping year as a string: While I set up the Python code to loop year as a string in the webpage (from 2010-2020), I was unable to successfully scrape multiple pages into a single data frame. I attempted several combinations of code to loop the scraped data, including the sleep() function from the time module and the randint() function from the random module, unfortunately neither worked. For the sake of this project, I re-ran the original Python script simply changing the years to scrape data for each class and write to an individual csv. Then I combined each year's csv to a single document with years from 2010-2020. Moving forward, I will revisit this to ascertain how to revise the Python code to scrape from multiple websites in a single data frame.

Future directions

As aforementioned, this webscraping component is integral to the overall project as it details private vs. public school attendance of NCAA players over a decade. Findings from this project provides key context on the phenomenon of declining Division I first-generation college student athletes. In the future, I will refine webscraping coding to more efficiently capture each class. Additionally, I will revisit kable data tables to address formatting aesthetics.

While my dataset fulfills the research goals of my project, there are limitations that could be addressed in the future. Given that the data only captures the top 100 NCAA recruits from 2010-2020, this dataset is but a small subset of all NCAA Division I players. It also doesn't include high school athletes who weren't recruited and signed. Future research should seek to capture the characteristics of broader populations of NCAA Division I men's basketball players, as well as those who weren't selected. This will provide a more holistic picture of defining characteristics that lead to inequality within the NCAA.