Designing a data pipeline to analyze gender equity in college sports

PPOL 5206: Massive Data Fundamentals

Bridgette Sullivan, Jasmine Jia, Katharyn Loweth, & Maria Bartlett

May 9, 2025

I. Introduction

Gender inequality is a complex issue within U.S. educational institutions. Although in recent years overall more women are graduating with bachelor's degrees than men, significant gender disparities still exist within some academic disciplines like engineering and in activities like sports programs (Hurst, 2024; Fry et al., 2021; NWLC, 2022a). This report demonstrates how applying big data techniques on the Equity in Athletics Disclosure Act (EADA) datasets can help the federal government and universities better monitor the gender divide within collegiate athletics.

II. Policy context

In the United States, gender disparities between men and women are a pervasive and complex issue. Protecting individuals' rights to opportunity is a key aspect of America's national identity. During the civil rights movement of the 1960s and 1970s, Congress enacted legislation that made sex-based discrimination illegal in certain environments. Specifically in 1972, Congress passed Title IX, which prohibits sex-based discrimination in educational institutions that receive federal funding (NWLC, 2022a). This not only includes any admission or classroom activities but also mandates that schools allocate athletic participation opportunities in a nondiscriminatory way (NWLC, 2022b).

Over the last fifty years, Congress and the Department of Education have continued to update the guidance and protocols for Title IX to better monitor its compliance. In 1979, the Department of Education established that institutions' intercollegiate athletics program must meet one of three criteria to be considered in compliance:

1) The number of male and female athletes is substantially proportionate to their respective enrollments
2) The institution has a history and continuing practice of expanding participation opportunities responsive to the developing interests and abilities of the underrepresented sex
3) The institution is fully and effectively accommodating the interests and abilities of the underrepresented sex. (Office of Civil Rights, n.d.)

Several years later, Congress enacted the Equity in Athletics Disclosure Act (EADA) in 1994 to make it easier to track Title IX compliance in higher education institutions. The EADA requires colleges and universities that (1) receive federal financial assistance (Title IV Federal Student Aid) and that (2) sponsor intercollegiate athletics to report data annually on male and female athletics participation, staffing issues for the male and female athletic teams, and the revenues and expenses for the male and female athletic teams to the Department of Education (NCAA, n.d.). The data is then made publicly available through the Department of Education's website (NCAA, n.d.).

III. Problem definition

For our final project, we chose to explore the nuances of gender gap in intercollegiate athletics using the EADA datasets. Specifically, our research is guided by the following questions:

1) Have the disparities between men's and women's sports increased, decreased, or remained stable over the 21-year period? Do these disparities vary by sport?
2) What distinct clusters of institutions emerge considering EADA variables for a single academic year?

IV. Data source

We downloaded the EADA dataset directly from the Department of Education website. Our EADA dataset consists of csv files, one for each academic year from years 2002-2023. Each dataset includes data on coach salaries, roster information, and recruitment expenses for all applicable male and female college athletic programs. There are approximately 4,000 variables captured within a single academic year. Each observation in the dataset represents a unique higher education institution that met the two criteria previously mentioned. While this exact number varies year-to-year, there are approximately 1,800 colleges/universities reflected in a single dataset.

V. Methdology & technical rationale

a. Cloud set-up

Because of the large file size for each dataset, we applied big data techniques to examine and manipulate the EADA data. The 4 steps that we took to prep the data for analysis are as follows:

1. Raw data upload to S3 We begin by storing our raw EADA CSV files in Amazon S3. These datasets contain critical gender equity indicators including team roster sizes, coaching staff compensation, recruitment expenses, and scholarship allocations. Our Schools.csv file contains sport-specific data across institutions, including participation numbers and expenses separated by gender. The instLevel.csv file provides institution-wide metrics such as total athletics revenue and gender proportionality measurements. Both files contain sensitive financial information that benefits from S3's security features, while remaining accessible for our gender equity analysis.

2. Cluster setup with EMR Analyzing gender disparities across hundreds of institutions and dozens of sports requires significant computing power, especially for our unsupervised learning approach. We launched an Amazon EMR cluster to handle the dimensionality reduction and clustering analyses needed to identify patterns in athletic department spending and participation across genders. Our EMR environment included Spark for distributed processing of the equity metrics and Jupyter for interactive analysis. While a single-node cluster proved sufficient for our initial investigation into gender gaps in basketball and track programs, EMR's scalability would accommodate future analyses of all NCAA sports across multiple divisions and years—expanding our understanding of gender equity trends.

3. Launch Jupyter on EMR using SSH Once the EMR cluster is active, we SSH into the EC2 Master node and manually start a Jupyter Notebook server on it. This gives us a familiar, browser-based Python environment where we can fetch data from S3, run our analysis, and visualize the output. This step requires setting up SSH tunneling so that we can access the notebook from our local browser securely.

4. Load & Analyze Data In the notebook, we use boto3 to connect to S3 and load our datasets into Pandas for analysis. We then run exploratory data analysis (EDA), correlation checks, visualizations, and prepare our results for export. Findings of the analysis will be discussed in the next section.

After completing our analytical transformations, the processed gender equity datasets are stored back in our S3 bucket in a structured format. By returning these processed datasets to S3, we maintain a complete data lineage from raw EADA submissions to final analytical outputs. This approach ensures reproducibility and allows team members to access the processed data through various tools, including direct connections from R for the visualization work presented in our exploratory and unsupervised learning sections. The S3 storage also facilitates version control as we refine our equity analyses over time, enabling us to track changes in gender disparities across multiple academic years of athletic department operations.

b. Analytical techniques

1. Exploratory analyses

2. Unsupervised analyses

VI. Findings

VII. Prototype: Institutions-specific reports

VIII. Application & significance

References

Fry, R., Kennedy, B., & Funk, C. (2021, April 1). *STEM Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity.* Pew Research Center. https://www.pewresearch.org/social-trends/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/

Hurst, K. (2024, November 18). *U.S. women are outpacing men in college completion, including in every major racial and ethnic group.* Pew Research Center. https://www.pewresearch.org/short-reads/2024/11/18/us-women-are-outpacing-men-in-college-completion-including-in-every-major-racial-and-ethnic-group/#:~:text=Today%2C%2047%25%20of%20U.S.%20women,from%2025%25%20to%2037%25).

National Collegiate Athletic Association (NCAA). (n.d.) *Gender Equity / Title IX Important Facts.* https://www.ncaa.org/sports/2013/11/21/gender-equity-title-ix-important-facts.aspx

National Women's Law Center (NWLC). (2022a, June 21). *The Battle for Gender Equity in Athletics in Colleges and Universities.* https://nwlc.org/resource/the-battle-for-gender-equity-in-athletics-in-colleges-and-universities/

National Women's Law Center (NWLC). (2022b, June 21). *Quick Facts About Title IX and Athletics.* https://nwlc.org/resource/quick-facts-about-title-ix-and-athletics

Office of Civil Rights. (n.d.). *Intercollegiate Athletics Policy: Three-Part Test – Part Three Q's & A's.* U.S. Department of Education. https://www.ed.gov/laws-and-policy/civil-rights-laws/sex-discrimination/intercollegiate-athletics-policy-three-part-test--part-three-qs-as

Office of Postsecondary Education. (n.d.). *Equity in Athletics Data Analysis.* U.S. Department of Education. https://ope.ed.gov/athletics/#/datafile/list

GitHub repository

The project GitHub repository is available at https://github.com/mariabartlett/massive-data-spring-2025-final-project.