**US Collegiate Sports Dataset**

**Exploratory Analysis**

Yeva Kramarova, [ykramarova@bellarmine.edu](mailto:ykramarova@bellarmine.edu)

John Little, [jlittle3@bellarmine.edu](mailto:jlittle3@bellarmine.edu)

1. **INTRODUCTION**

This data set contains the comprehensive information on collegiate sports programs across various institutions in the United States from years 2015-2019. It includes data on student enrollment, sports participation, revenue, and expenditures, categorized by gender and sport. The dataset can be used to analyze trends, financial aspects, and gender disparities in collegiate sports. The variables are as follows:

- year: Year.

- unitId: School ID.

- institution\_name: School name.

- city\_txt: City name.

- state\_cd: State abbreviation.

- zip\_text: Zip of school.

- classification\_code: Code for school clasification.

- classification\_name: School classification.

- classification\_other: School classification otehr.

- ef\_male\_count: Total male students.

- ef\_female\_count: Total female student.

- ef\_total\_count: Total students for binary male/female gender.

- sector\_cd: Sector code.

- sector\_name: Sector name.

- sportscode: Sport code.

- partic\_men: Participation men.

- partic\_women: Participation women.

- partic\_coed\_men: Participation for coed men.

- pactic\_coed\_women: Participation for coed women.

- sum\_partic\_men: Sum of participation for men.

- sum\_partic\_women: Sum of participation for women.

- rev\_men: Revenue in USD for men.

- rev\_women: Revenue in USD for women.

- total\_rev\_menwomen: Total revenue for both.

- exp\_men: Expenditures in USD for men.

- exp\_women: Expenditures in USD for women.

- total\_exp\_menwomen: Total expenditure for both.

- sports: Sport name.

1. **DATASET DESCRIPTION**

This dataset includes the total number of male and female students enrolled in each institution, providing insights into the gender distribution of the student body. Participation data is broken down by gender and sport, allowing for analysis of gender representation in different sports. Revenue and expenditures for men's and women's sports are detailed, enabling financial analysis of sports programs. Institutions are classified by type and sector, which helps in comparing different categories of schools (e.g., NCAA Division I, II, III).

The data set contains 132,327 rows and 28 columns with various data types. A complete listing is shown in Table 1.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| year | int64/interval | 0% |
| unitid | int64/nominal | 0% |
| institution\_name | object/nominal | 0% |
| city\_txt | object/nominal | 0.03% |
| state\_cd | object/nominal | 0.03% |
| zip\_text | float64/nominal | 0.03% |
| classification\_code | int64/nominal | 0% |
| classification\_name | object/nominal | 0% |
| classification\_other | object/nominal | 98.73% |
| ef\_male\_count | int64/ratio | 0% |
| ef\_female\_count | int64/ratio | 0% |
| ef\_total\_count | int64/ratio | 0% |
| sector\_cd | int64/nominal | 0% |
| sector\_name | object/nominal | 0.03% |
| sportscode | int64/nominal | 0% |
| partic\_men | float64/nominal | 53.25% |
| partic\_women | float64/nominal | 47.94% |
| partic\_coed\_men | float64/nominal | 99.42% |
| partic\_coed\_women | float64/nominal | 99.42% |
| sum\_partic\_men | int64/nominal | 0% |
| sum\_partic\_women | int64/nominal | 0% |
| rev\_men | float64/ratio | 53.25% |
| rev\_women | float64/ratio | 47.94% |
| total\_rev\_menwomen | float64/ratio | 34.15% |
| exp\_men | float64/ratio | 53.25% |
| exp\_women | float64/ratio | 47.94% |
| total\_exp\_menwomen | float64/ratio | 34.15% |
| sports | object/nominal | 0% |

After analyzing the dataset, we found that several variables, such as classification\_other, partic\_coed\_men, and partic\_coed\_women, had nearly 100% missing values. Based on this observation, we decided to drop these variables (as shown in step:4). Upon further inspection, we noticed that a few remaining variables still had a significant proportion of missing values. However, after closely examining the dataset, we realized that the "NaN" values actually represent zeros rather than missing data.

For example, in the variable partic\_women, which indicates the number of female participants in a given sport, some universities, such as Alabama A&M University, show 99 male participants in football but "NaN" for female participants. This pattern of "NaN" values essentially reflects cases where no female participation occurred. Given this, we concluded that it would be reasonable to replace all "NaN" values with zeros, as the lack of participation is equivalent to a value of zero. This decision is reflected in step: 6.

By doing this, we ensure the dataset accurately represents the data without misinterpreting "NaN" values as missing information.

The Table 1.1 depicts the updated dataset.

**Table 1.1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| year | int64/interval | 0% |
| unitid | int64/nominal | 0% |
| institution\_name | object/nominal | 0% |
| city\_txt | object/nominal | 0.03% |
| state\_cd | object/nominal | 0.03% |
| zip\_text | float64/nominal | 0.03% |
| classification\_code | int64/nominal | 0% |
| classification\_name | object/nominal | 0% |
| ef\_male\_count | int64/ratio | 0% |
| ef\_female\_count | int64/ratio | 0% |
| ef\_total\_count | int64/ratio | 0% |
| sector\_cd | int64/nominal | 0% |
| sector\_name | object/nominal | 0.03% |
| sportscode | int64/nominal | 0% |
| partic\_men | float64/nominal | 53.25% |
| partic\_women | float64/nominal | 47.94% |
| sum\_partic\_men | int64/nominal | 0% |
| sum\_partic\_women | int64/nominal | 0% |
| rev\_men | float64/ratio | 53.25% |
| rev\_women | float64/ratio | 47.94% |
| total\_rev\_menwomen | float64/ratio | 34.15% |
| exp\_men | float64/ratio | 53.25% |
| exp\_women | float64/ratio | 47.94% |
| total\_exp\_menwomen | float64/ratio | 34.15% |
| sports | object/nominal | 0% |

**Table 2: Summary Statistics for US Collegiate Sports Dataset**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| year | 132327.0 | 2017.68 | 1.49 | 2015.0 | 2016.0 | 2018.0 | 2019.0 | 2019.0 |
| unitid | 132327.0 | 184360.4 | 56064.63 | 100654.0 | 149781.0 | 181738.0 | 214069.0 | 800001.0 |
| zip\_text | 132327.0 | 41990630.0 | 158580400.0 | 0.0 | 23220.0 | 49104.0 | 79549.0 | 997757500.0 |
| classification\_code | 132327.0 | 7.78 | 4.89 | 1.0 | 4.0 | 6.0 | 12.0 | 20.0 |
| ef\_male\_count | 132327.0 | 2126.25 | 3014.38 | 0.0 | 513.0 | 986.0 | 2385.0 | 35954.0 |
| ef\_female\_count | 132327.0 | 2496.21 | 3276.05 | 0.0 | 652.0 | 1248.0 | 2860.0 | 30325.0 |
| ef\_total\_count | 132327.0 | 4622.46 | 6228.14 | 0.0 | 1194.0 | 2259.0 | 5237.0 | 66279.0 |
| sector\_cd | 132327.0 | 2.21 | 2.10 | 1.0 | 1.0 | 2.0 | 2.0 | 99.0 |
| sportscode | 132327.0 | 16.34 | 10.57 | 1.0 | 7.0 | 16.0 | 25.0 | 38.0 |
| partic\_men | 132327.0 | 14.43 | 25.30 | 0.0 | 0.0 | 0.0 | 20.0 | 331.0 |
| partic\_women | 132327.0 | 10.78 | 17.12 | 0.0 | 0.0 | 6.0 | 17.0 | 327.0 |
| sum\_partic\_men | 132327.0 | 14.49 | 25.30 | 0.0 | 0.0 | 0.0 | 20.0 | 331.0 |
| sum\_partic\_women | 132327.0 | 10.86 | 17.15 | 0.0 | 0.0 | 6.0 | 17.0 | 327.0 |
| rev\_men | 132327.0 | 378225.7 | 3365077.0 | 0.0 | 0.0 | 0.0 | 141637.0 | 156147208.0 |
| rev\_women | 132327.0 | 145414.1 | 337764.8 | 0.0 | 0.0 | 14683.0 | 146853.5 | 21440365.0 |
| total\_rev\_menwomen | 132327.0 | 523639.8 | 3412958.0 | 0.0 | 0.0 | 92471.0 | 337375.5 | 156147208.0 |
| exp\_men | 132327.0 | 309676.0 | 1860785.0 | 0.0 | 0.0 | 0.0 | 142649.0 | 69718059.0 |
| exp\_women | 132327.0 | 172616.8 | 425603.4 | 0.0 | 0.0 | 14915.0 | 151354.5 | 9485162.0 |
| total\_exp\_menwomen | 132327.0 | 482292.8 | 1996331.0 | 0.0 | 0.0 | 92489.0 | 350378.0 | 69718059.0 |