**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**

The data set was found on Kaggle, and can be accessed using the following link <https://www.kaggle.com/datasets/umerhaddii/us-collegiate-sports-dataset> .

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.

For the categorical variables, such as city\_txt, state\_cd, sector\_name, we replaced the missed values with the “Unknown” value using the imputation dictionary function.

The Table 1.1 depicts the updated dataset.

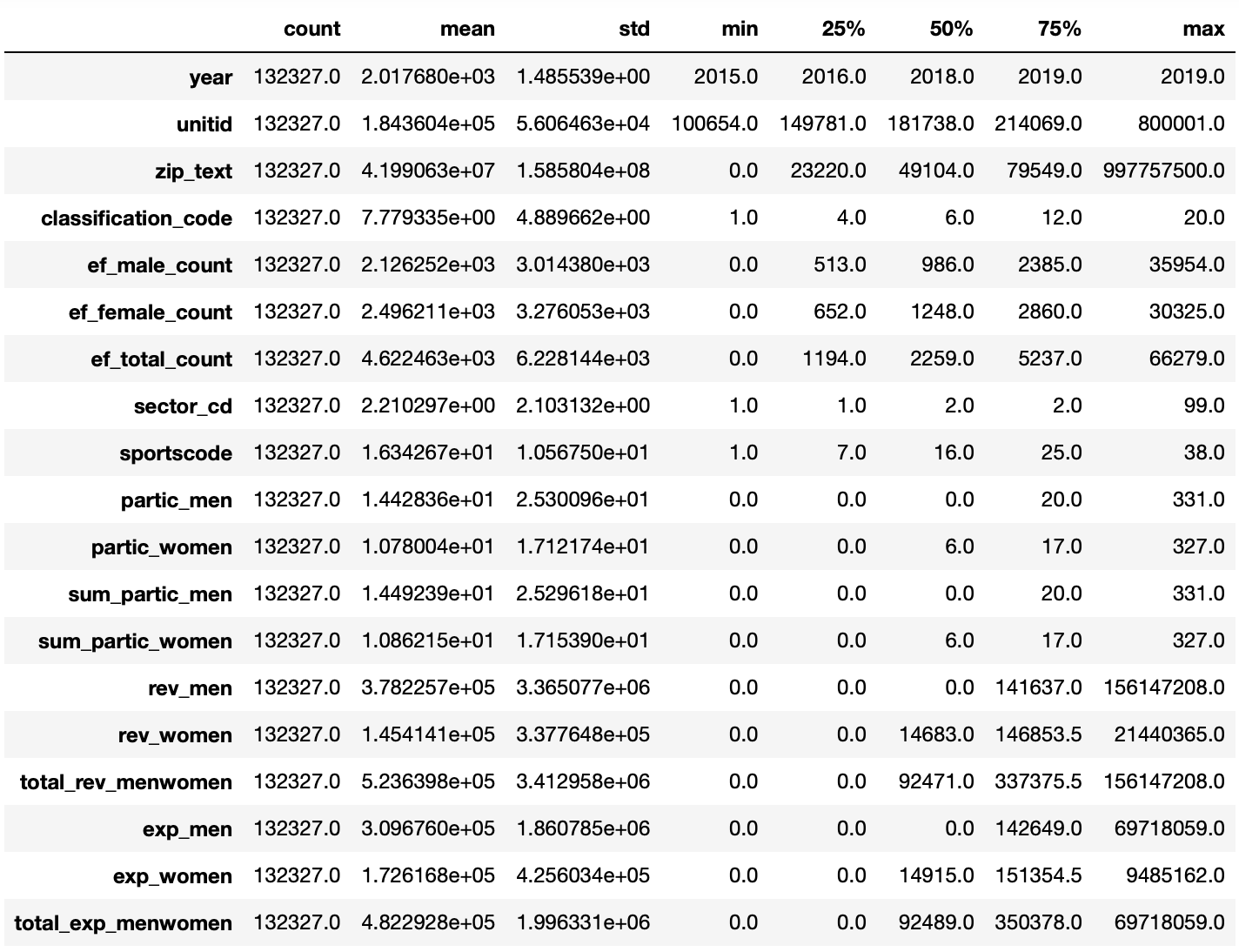
**Table 1.1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *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% |
| 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 | 0% |
| partic\_women | float64/nominal | 0% |
| sum\_partic\_men | int64/nominal | 0% |
| sum\_partic\_women | int64/nominal | 0% |
| rev\_men | float64/ratio | 0% |
| rev\_women | float64/ratio | 0% |
| total\_rev\_menwomen | float64/ratio | 0% |
| exp\_men | float64/ratio | 0% |
| exp\_women | float64/ratio | 0% |
| total\_exp\_menwomen | float64/ratio | 0% |
| sports | object/nominal | 0% |

1. **Data Set Summary Statistics**

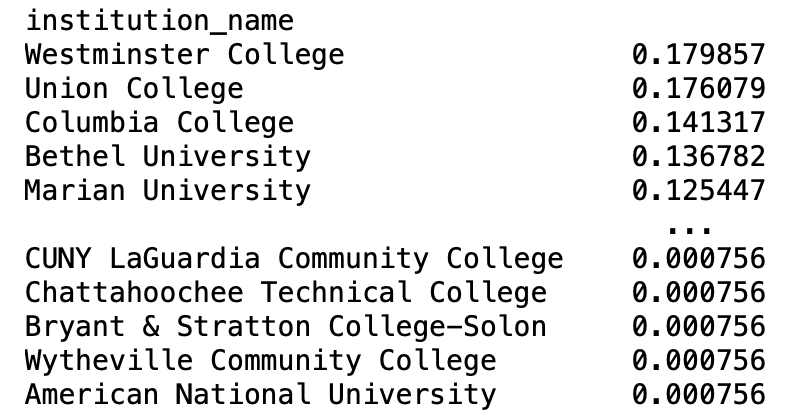
This section will dive deeper into the statistical measurements of the dataset such as count, mean, standard deviation, min, 25th, 50th, 75th, max.

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

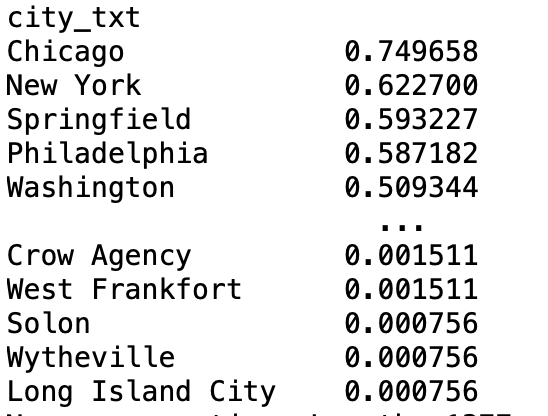
****

Below, is the section that depicts the frequency of each category. The screenshots are the top parts of the tables for each of the categorical variables that show its frequency. The proportion results are shown in the Jupyter Notebook. The remaining parts of some of the tables are also shown in the Jupyter Notebook.

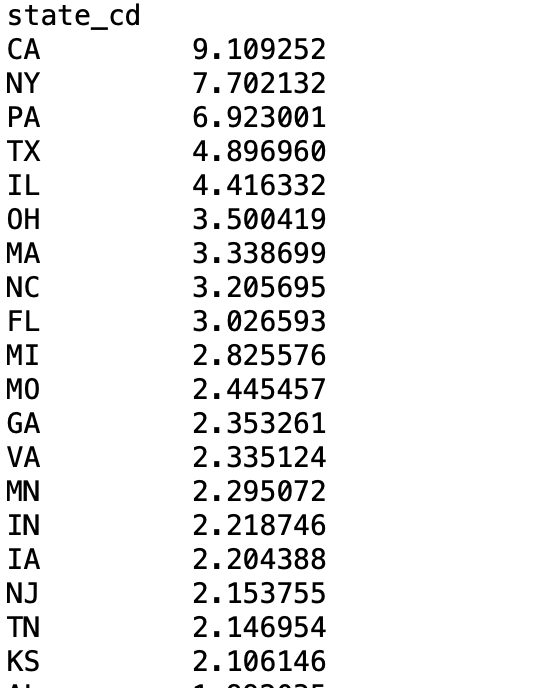
**Table 3: Frequencies within the Institution Name Variable**



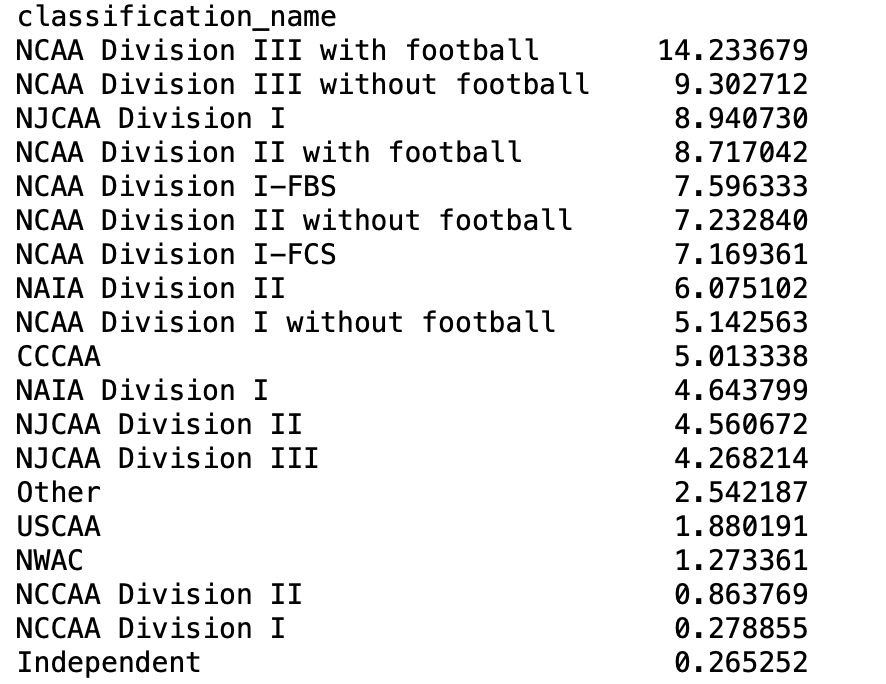
**Table 3.a: Frequencies within the City Name Variable**

**

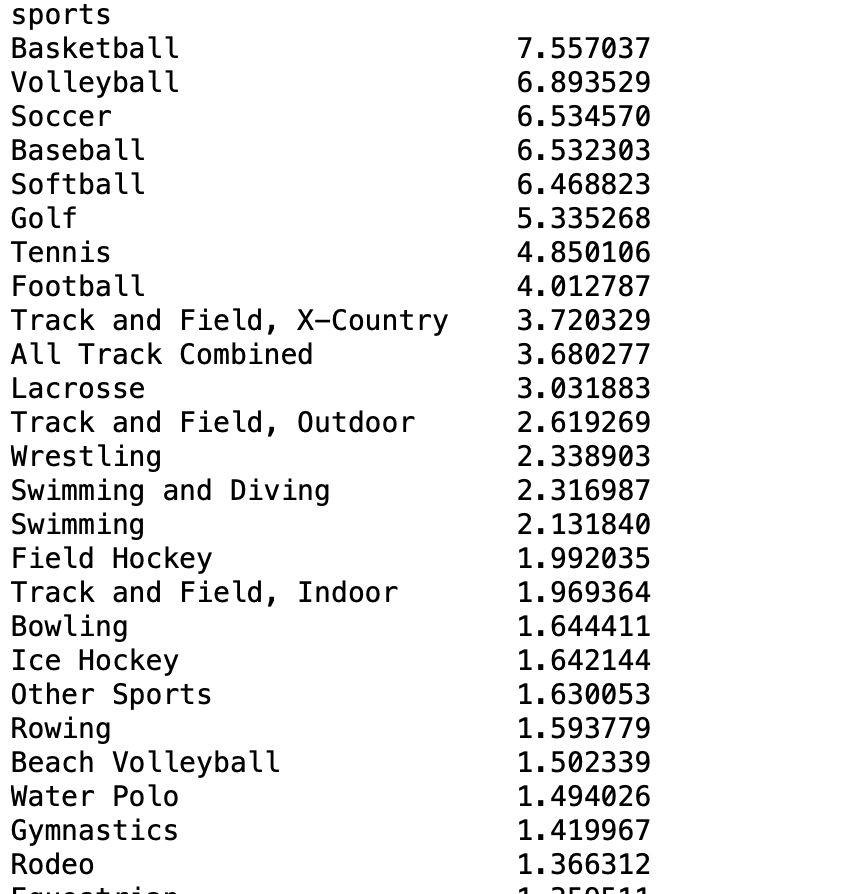
**Table 3.b: Frequencies within the State Abbreviation Variable**



**Table 3.c: Frequencies within the Classification Name Variable**



**Table 3.d: Frequencies within the Sports Variable**



**Table 3.e: Frequencies within the Sector Name Variable**

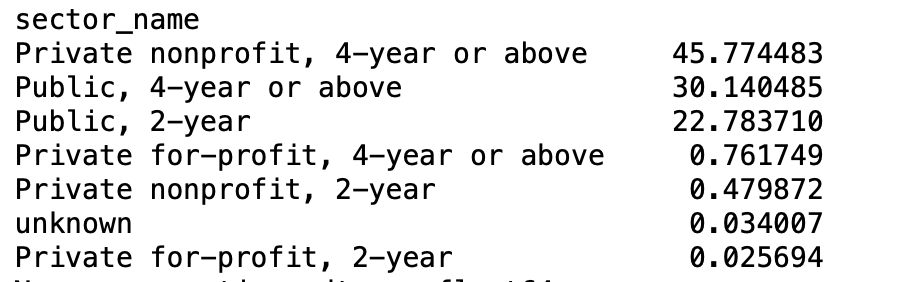
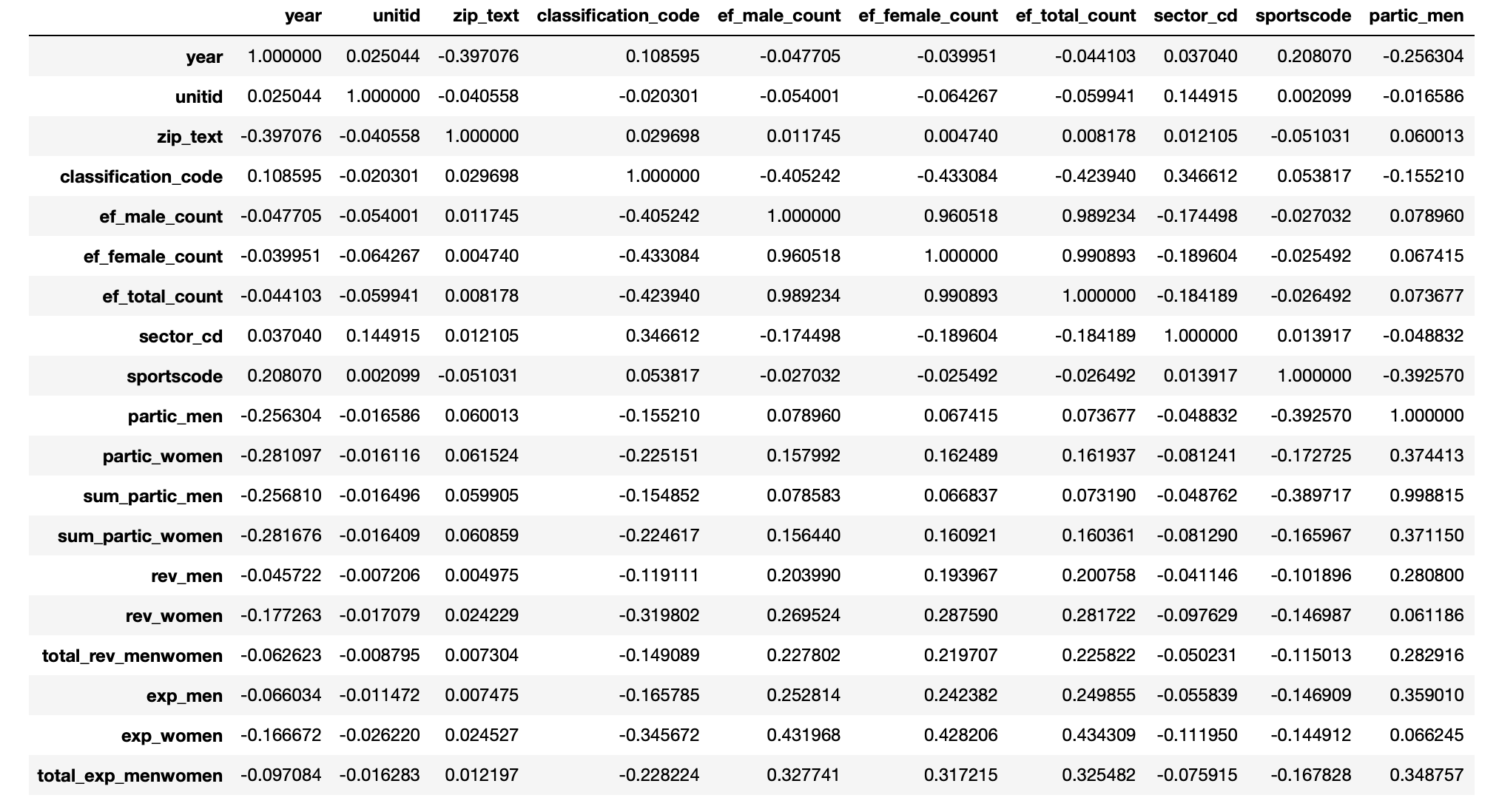


Table 4 shows the correlation of all continuous variables.

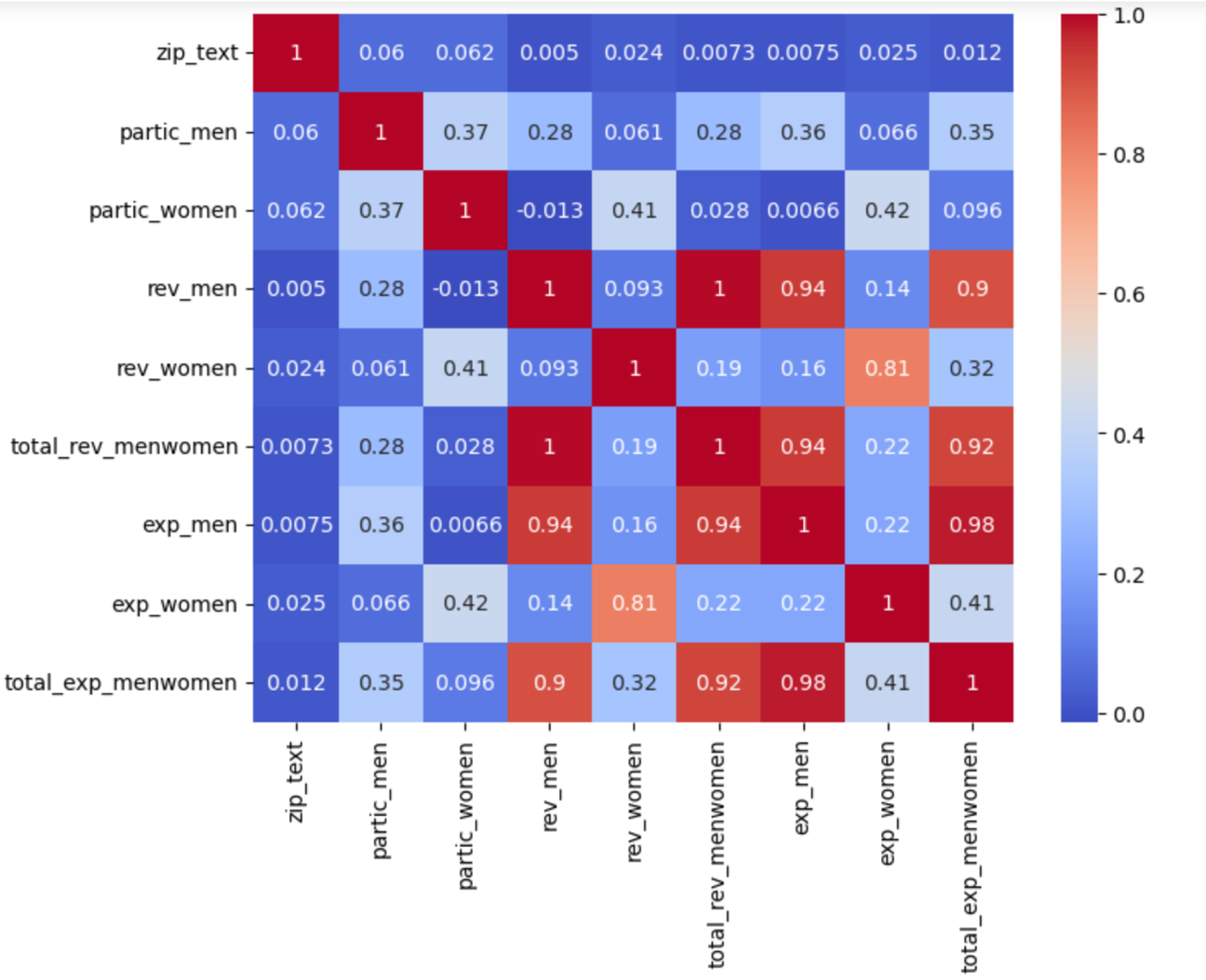
**Table 4: Correlation Table**

****

****

A **correlation heatmap** visually displays the **strength and direction** of relationships between multiple variables in a dataset. Each cell in the heatmap represents the **correlation coefficient** between two variables, which ranges from **-1** to **+1. The closer the number to +1, the stronger the relationship between the variables. The closer the number to -1, the stronger the negative relationship between the variables.**

**Correlation Heatmap**

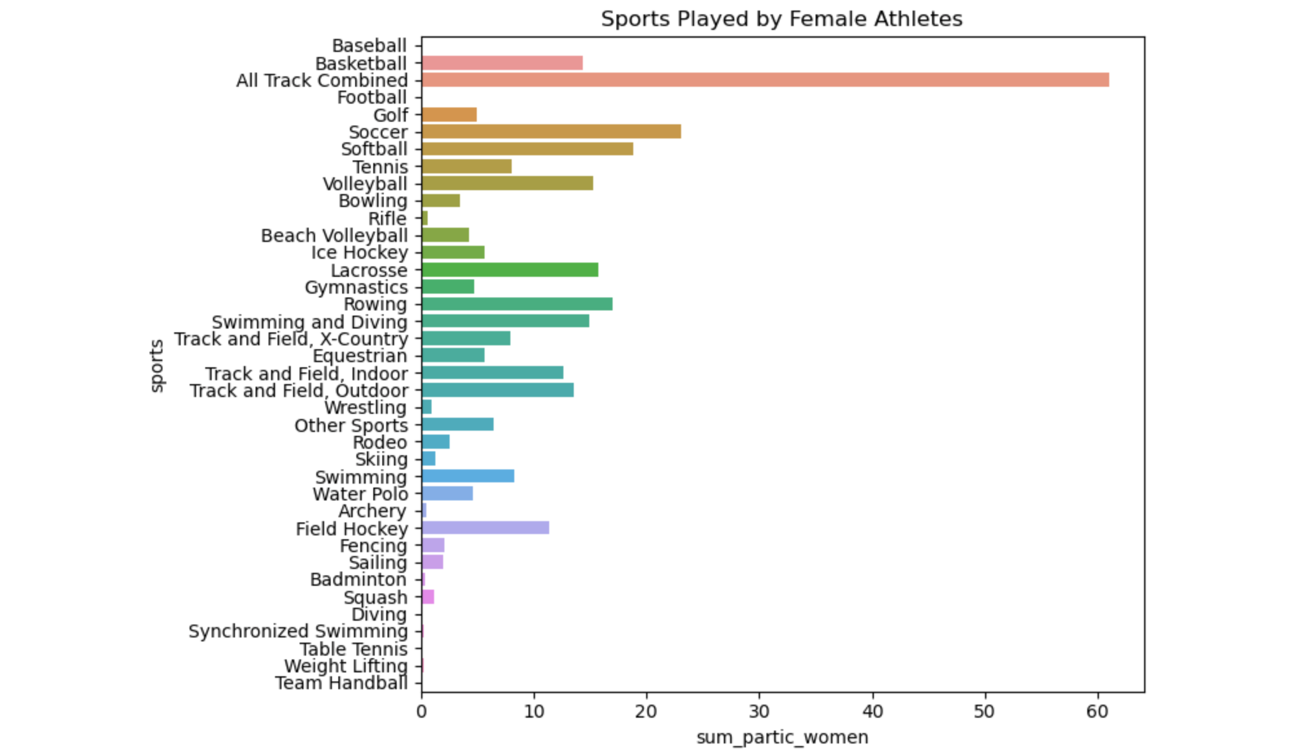


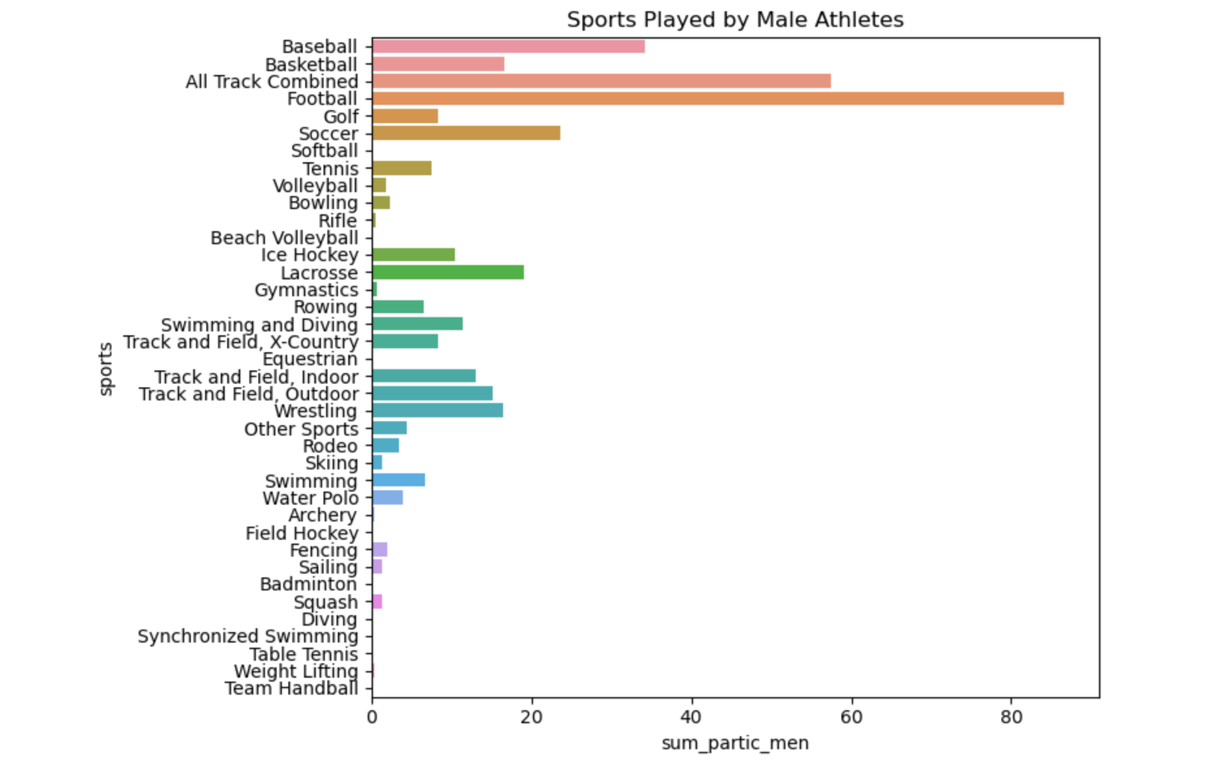
1. **DATA SET GRAPHICAL EXPLORATION**

In this section, different kinds of graphs will help us to become more familiar with the dataset and to discuss interesting distributions, anomalies, or imbalances of the dataset.

The following two charts show the distribution of the sports played by first the female athletes, Figure 1. The distribution of the sports played by the male athletes is shown in Figure 2.

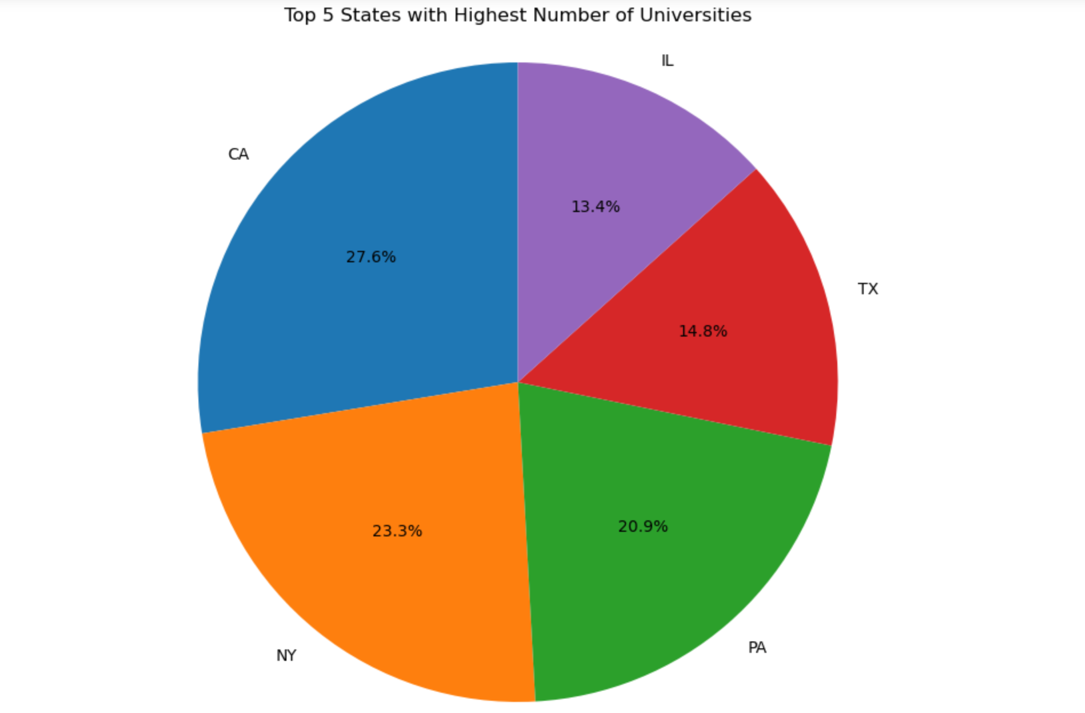
**Figure 1: Sports Distribution in Female Athletes**



**Figure 2: Sports Distribution in Male Athletes** 

The pie chart in the Figure 3 shows which states have the highest number of universities, you can group the data by the state\_cd column, which represents the state abbreviations, and then count the number of institutions per state.

**Figure 3: Top 5 States by University Count**



The Figure 4 is a bar plot showing the total revenue generated by men and women, and it formats the Y-axis to display the values in US dollars with a dollar sign and commas (e.g., $1,000,000) for easier readability.

**Figure 4: Total Revenue by Gender**

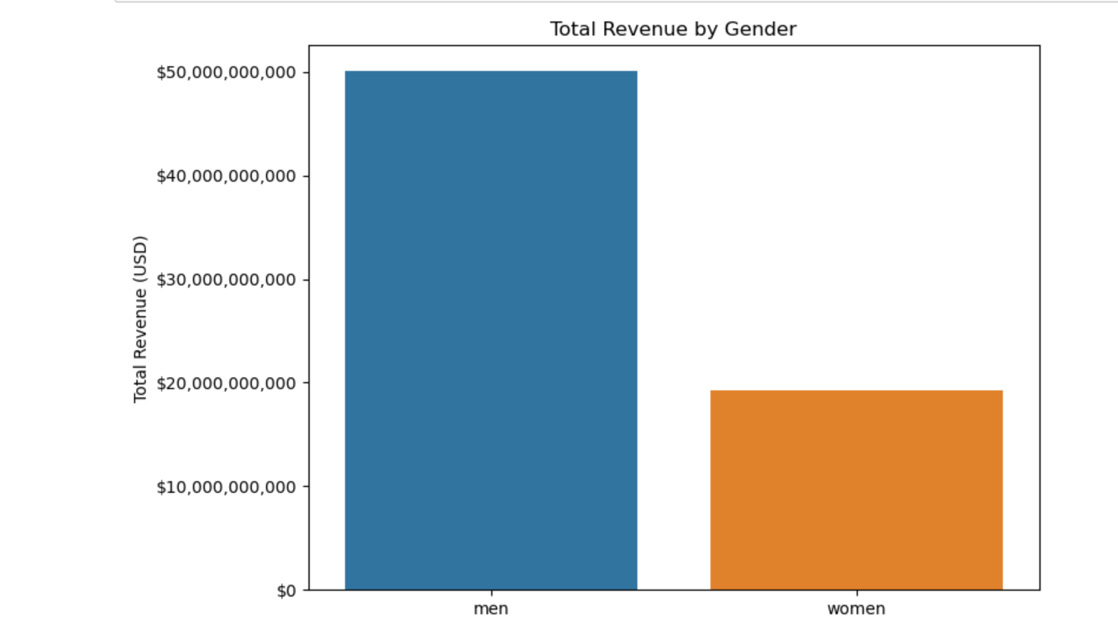


Figure 5 generates a grouped bar chart comparing revenue and expenditure for men and women. It displays two bars for each gender: one for total revenue and one for total expenditure, with the X-axis showing "men" and "women," and the Y-axis representing the amounts in USD.

**Figure 5: Revenue and Expenditure by Gender Comparison**

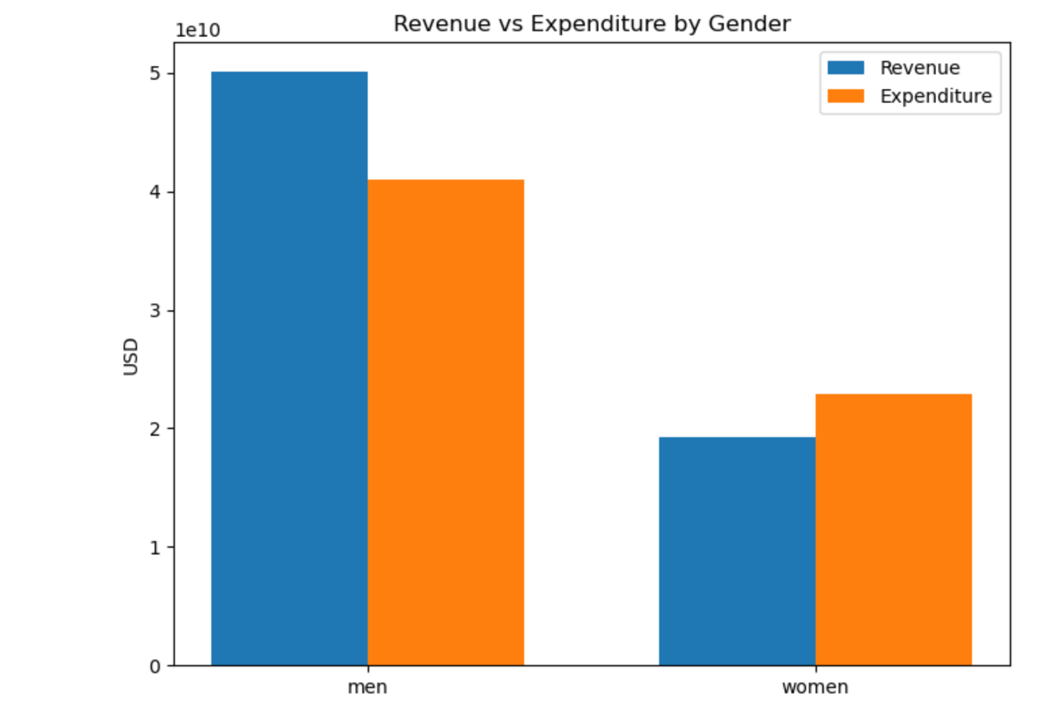
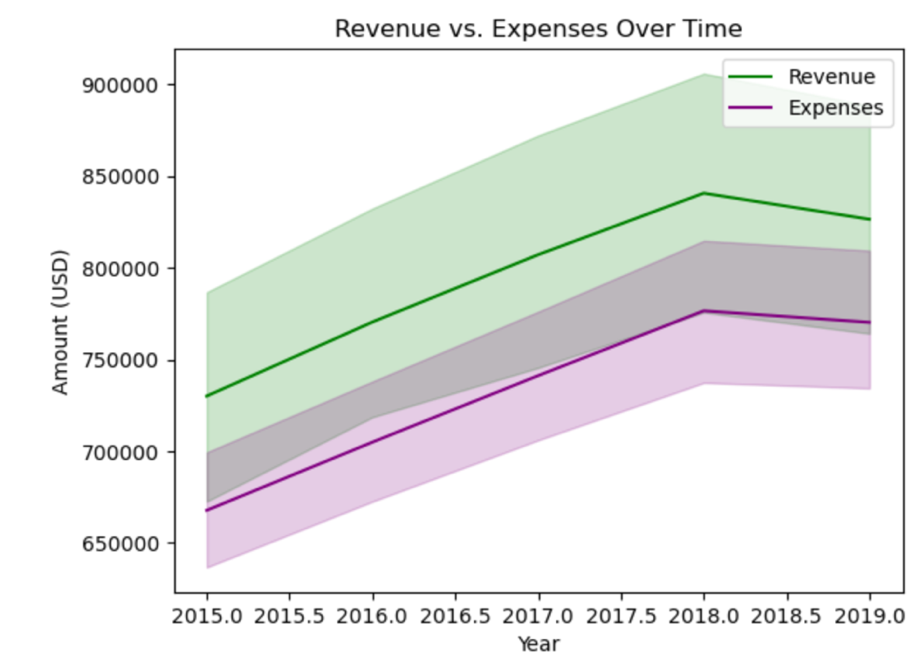


Figure 6 is a two-line plot that compare total revenue and total expenses over time.

**Figure 6: Revenue vs Expenses Over Time in all the US**

****

The Figure 7 is a stacked bar plot showing the total participation in sports by both men and women, where men's participation is represented in blue and women's participation is represented in red.

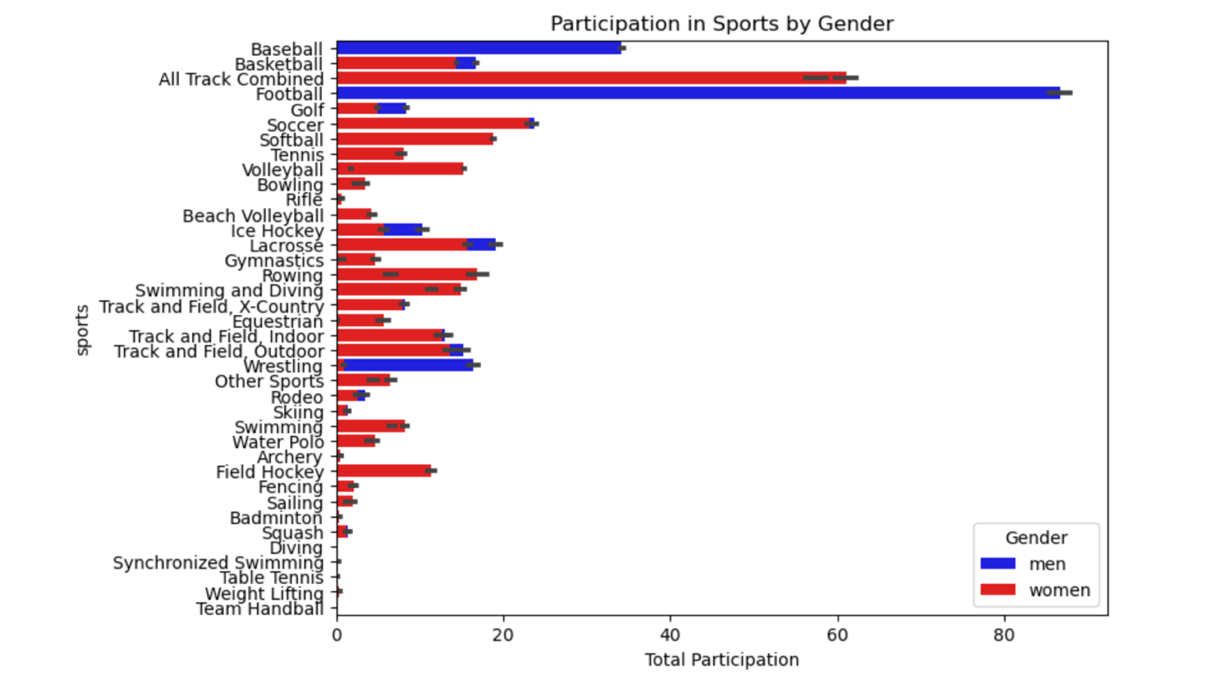
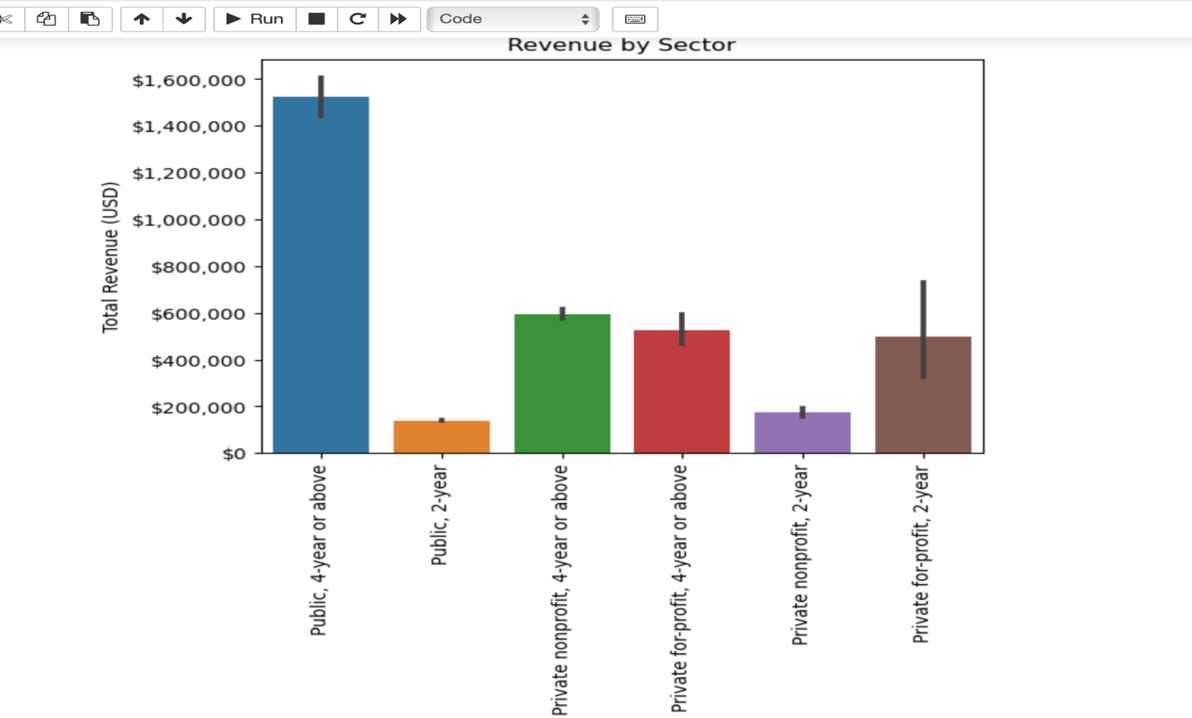
**Figure 7: Gender Sport Participation**

Figure 8 creates a **bar plot** that shows the **total revenue for men and women** by different **sectors** (e.g., public, private, etc.).

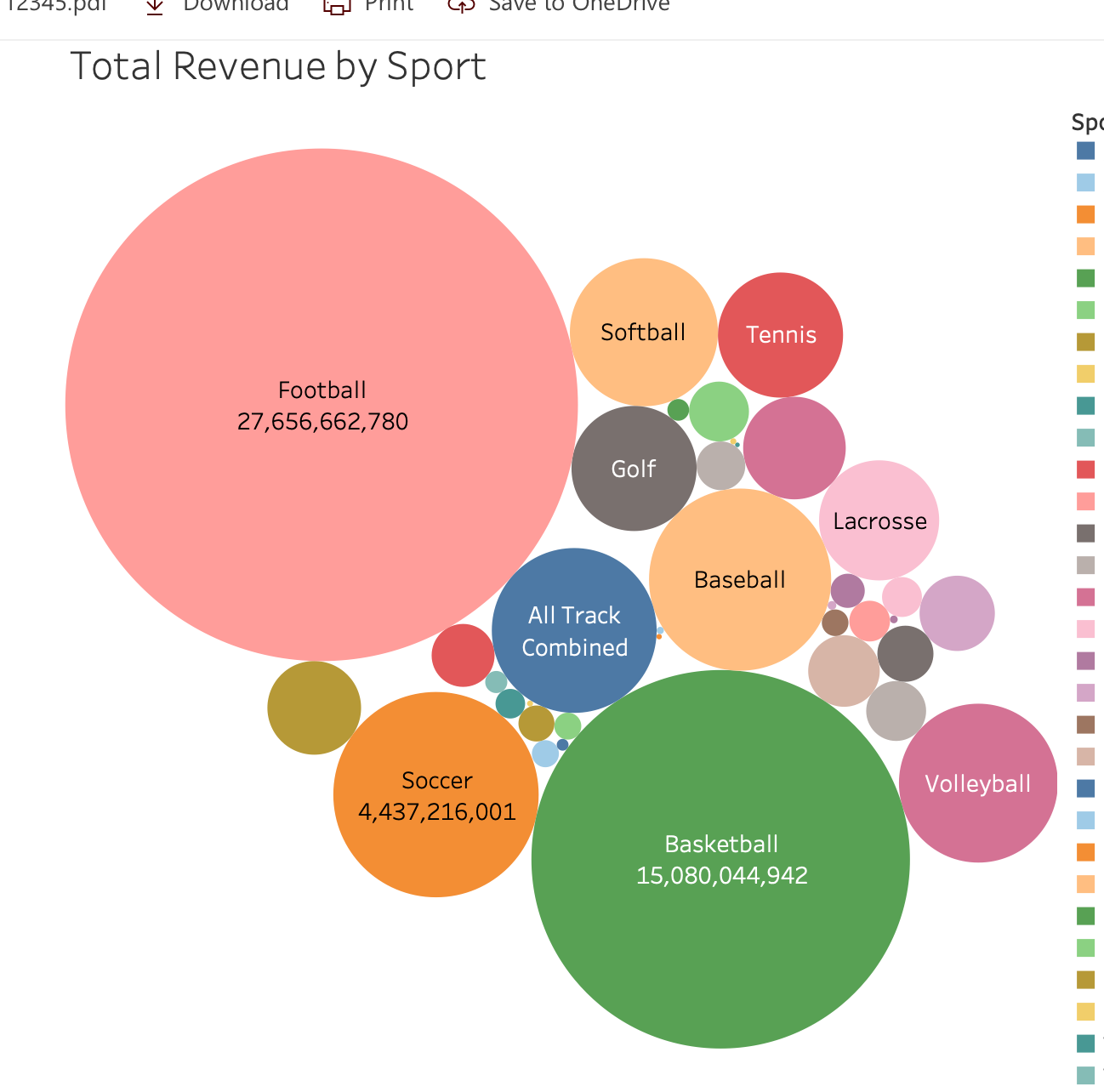
**Figure 8:**

****

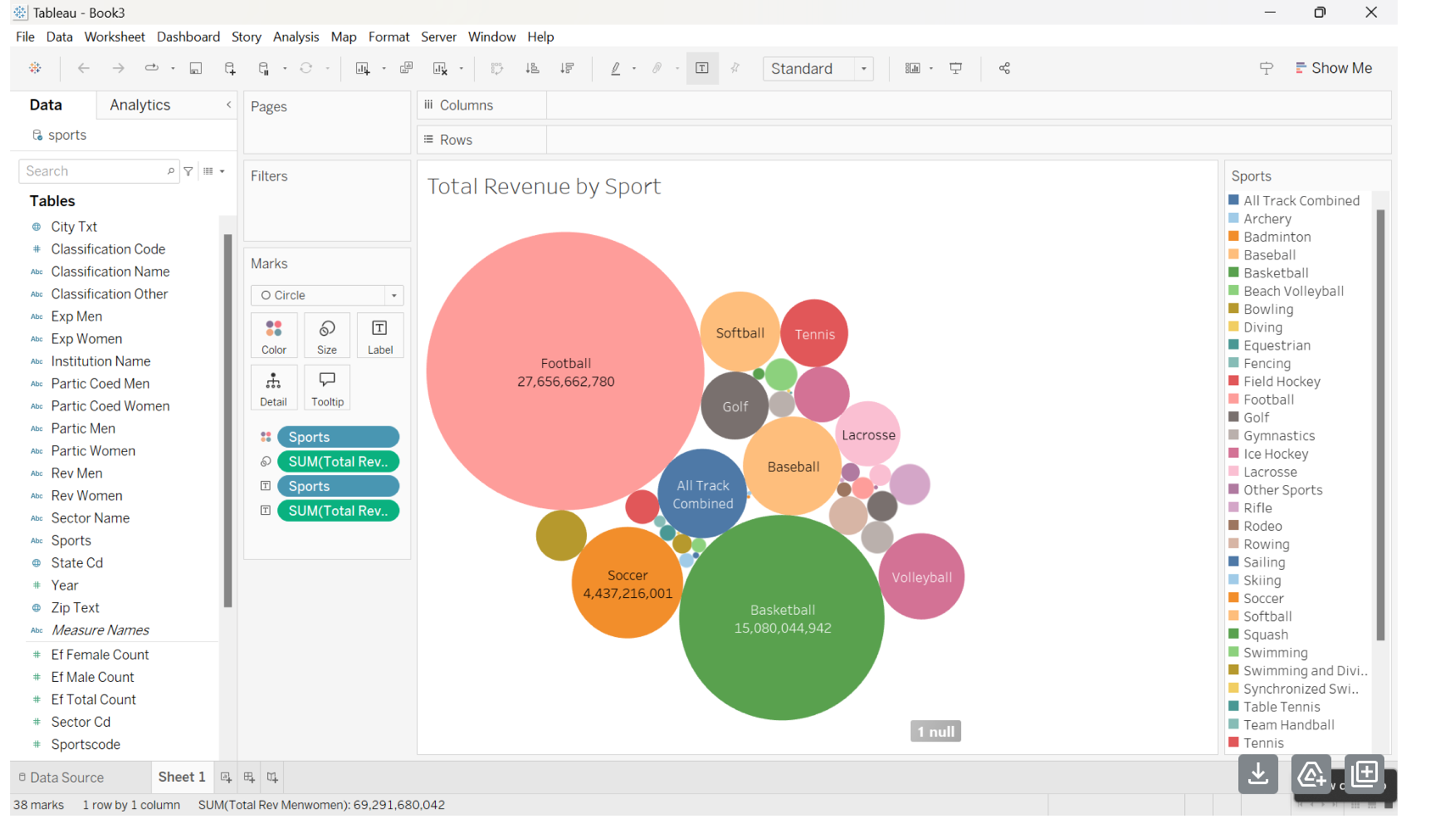
Looking at **Figure 8**and**Table 3.e,** an interesting anomaly stands out: while there are more **private 4-year schools,** it is the **public 4-year schools**that generate the highest revenue.

Figure 9 is packed bubbles that compares the total revenues by sport using the Tableau Software that uses visual analytics to describe data.

**Figure 9: Total Revenue Comparison by Sport**



This is the Tableau work behind the packed bubbles graph.



1. **Summary of Findings**

In conclusion, the US Collegiate Sports Dataset is extensive and encompasses a wide range of categories. From analyzing the graphs, several key insights emerge. For example**, female athletes** are predominantly represented in **track and field events**, while **male athletes** are most commonly involved in **football**. It's also noteworthy that, according to **Table 3.a**, the city of **Chicago** has the highest frequency, even though the state it belongs to is not among the top three most frequently mentioned states in the dataset. Moreover, it is clear that **male sports** generate more revenue, whereas **female sports** tend to have higher expenditures. Another notable finding is that, despite there being more **private 4-year schools**, it is the **public 4-year schools** that bring in the most revenue. And not surprisingly, football brings the most revenue.