Project Proposal

Standard Deviants - Ava Exelbirt, Yura Heo, Claire Li

Introduction

We aim to investigate and understand the factors influencing the revenue generated by collegiate sport, specifically basketball, in the school year 2019 to 2020, We will be utilizing a dataset sourced from Tidy Tuesday, which provides a comprehensive collection of observations related to collegiate sports, including information about the schools, classifications, sport types, and various quantitative variables describing the sport players and the schools' financial investment in sports programs in the dataset called "sports.csv". Our primary motivation for this research is to gain insights into the factors contributing to collegiate basketball revenue, which can be of significant interest to educational institutions and further local policy makers and sports enthusiasts. This can inform decision-making, investment strategies in sports, and future planning for universities and colleges involved in sports programs.

Our primary research question is as follows: **In the school year 2019 to 2020, how can we predict the revenue in USD of the collegiate sports basketball using participation rate, type of school, expenditure, gender, total count of students, and school classification name.**

Our hypothesis is as follows: **Expenditure, type of school, and gender will be the most influential predictors for the total revenue generated by the collegiate sports basketball in USD.**

We anticipate that expenditure will be a key predictor of total revenue of the sports basketball because schools often allocate significant financial resources to their sport programs, scholarships, and marketing. As these investments increase, the expectation is that they will have a direct impact on the overall revenue generated through different resources such as ticket sales. We also expect that the type of school will be a strong predictor of revenue because different types of schools have varying levels of resources, alumni support, and participation rates in sports depending on school size and program. Gender is another predictor to consider because sports could be hugely dependent on the demographics of players.

- Collegiate Sports Dataset: https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-03-29

Data description

The data set was taken from TidyTuesday and was originally scraped from Equity in Athletics Data Analysis (EADA), a sector of the US Department of Education. The data is available on an online database found on the (EADA) website[1].

This data is submitted annually from colleges to the EADA. All co-educational postsecondary institutions that receive Title IV funding that have intercollegiate athletics programs are required by the Equity in

[1] https://ope.ed.gov/athletics/#/datafile/list

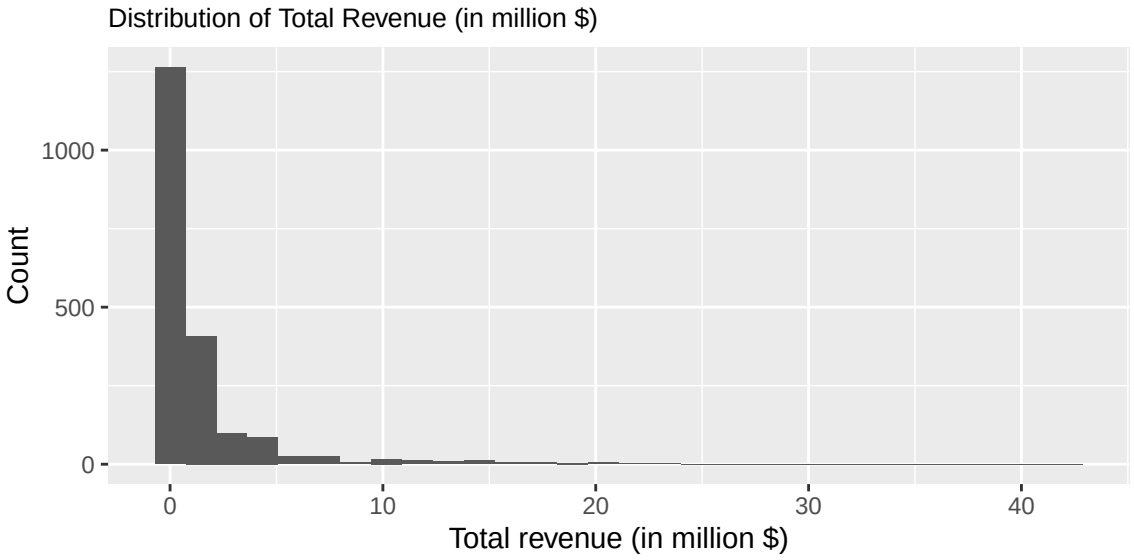
Athletics Disclosure Act to submit this data[2]. The original data files are also created immediately after the data collection for each school. These data are collected annually starting from 2003 to 2022, but for our specific data analysis we will only look at data taken from the school year of 2019 by cleaning the data to a new csv file which we will use for the rest of the projecr. The csv file from Tidy Tuesday contains thousands of observations from years 2015-2019, but we have adjusted the data file to only include year 2019 and will filter the data to only include basketball as a sport, as this is our population we are analyzing.

The dataset we will use from the data scraped on Tidy Tuesday include many observations regarding collegiate sports. These observations include variables such as the name of the city which a school is in, the state, the school name, the classification of the school (like whether it is NCAA Division I, II, or III), the type/sector name of the school (like 4 year accredited university), sport, and many quantitative variables regarding characteristics about the specific school and their collegiate spending. Most of the quantitative variables are split between men and women, having two different observations for the same variable. For example, there is total male population, total female population, participation rate of women, participation rate of men, participation rate for coed sports for men, participation rate for coed sports for women, revenue for men, revenue for women, expenditures for men, and expenditures for women. There are also observations for the total amount for each of these above variables which includes both men and women; for example, total expenditures for both men and women together. The observations are therefore both quantitative and categorical and measure characteristics of different school's spending, revenue, populations, locations, and sports.

Initial exploratory data analysis

For data cleaning, we need to filter for observations that have `Basketball` as the sports and filter out missing values. we also need to filter out observations that have missing values of the predictor variables we need in the model. We need to create new variables by mutation to turn variables involving revenue and expenditure into the unit of millions, so that we can get larger coefficients for better modeling.

Distribution of the response variable: total revenue generated by college sports

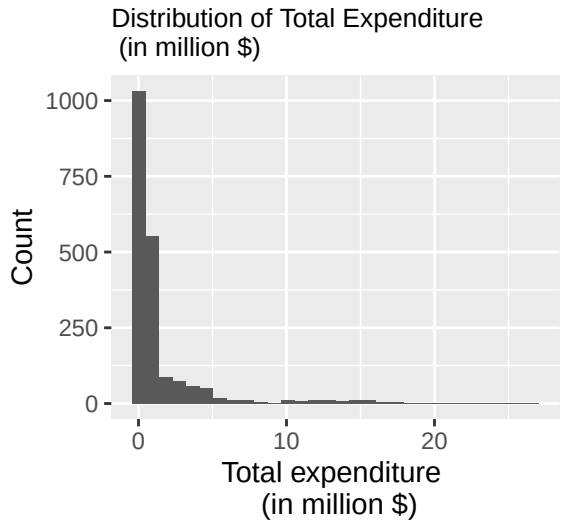


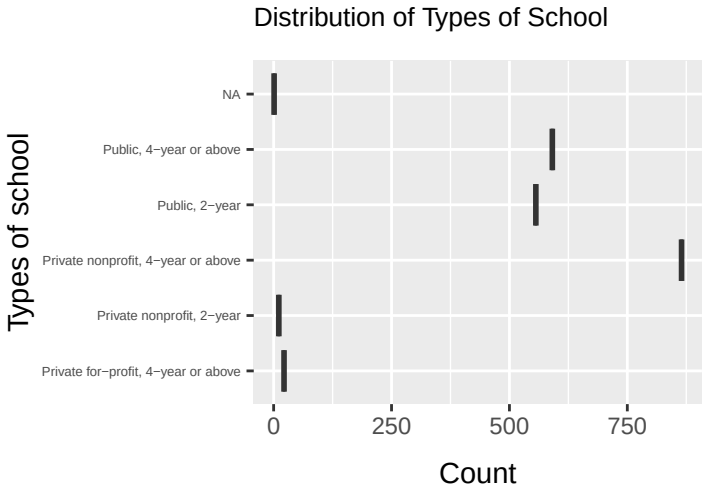
[2]https://knightnewhousedata.org/about-the-data#:~:text=The%20data%20is%20available%20via,student%20aid%20programs)

```
# A tibble: 1 x 8
  n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
      <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
1        57         1.54       3.39    0.00226       0.193       0.431
  numeric.p75 numeric.p100
        <dbl>        <dbl>
1        1.10         42.2
```

The distribution is right-skewed, as most observations are at the lower end of the x axis with a number of outliers having very high values. The center is the median of 0.431111 million dollars due to the apparent skewness. Since the distribution is skewed, the IQR is used as a more reliable measure of spread which is Q3 - Q1 = 1.098081 - 0.19334 = 0.904741 million dollars. There are 57 missing values of total revenue.

Distributions of total expenditure (potential quantitative predictor variable) and type of school (potential categorical predictor variable)






```
# A tibble: 1 x 8
  n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
      <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
1        57         1.52       3.09    0.00386       0.188       0.434
  numeric.p75 numeric.p100
        <dbl>        <dbl>
1        1.10         26.6
```

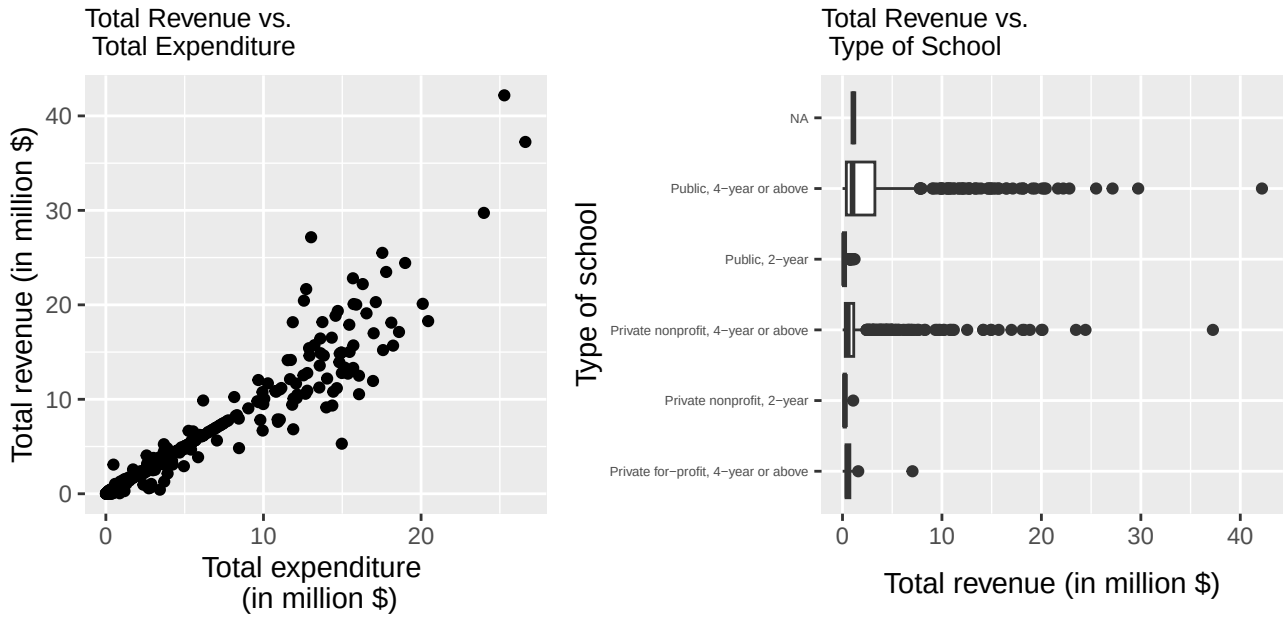
```
                                    n           p nmiss
Private for-profit, 4-year or above  22 0.010757946     1
Private nonprofit, 2-year            11 0.005378973    NA
Private nonprofit, 4-year or above  865 0.422982885    NA
Public, 2-year                      556 0.271882641    NA
Public, 4-year or above             591 0.288997555    NA
```

The distribution of total expenditure on college sports is right-skewed with most total expenditure values in the lower range, while a number of observations have very high values that make them outliers. Given

the apparent skewness, the center is the median of 0.433728 million dollars. Since the distribution is skewed, the IQR is used as a more reliable measure of spread which is Q3 - Q1 = 1.100745 - 0.188125 = 0.91262 million dollars. There are 57 missing values of total revenue.

here are 5 types of school. Private for-profit, 2-year and private nonprofit, 2-year have low numbers of observations. Private nonprofit, 4-year or above, public, 2 year, and public, 4-year or above have comparably higher number of observations than the other two, with private nonprofit, 2-year having the highest number of observations. There is 1 missing value of type of school.

The relationships between total revenue and each of the predictors

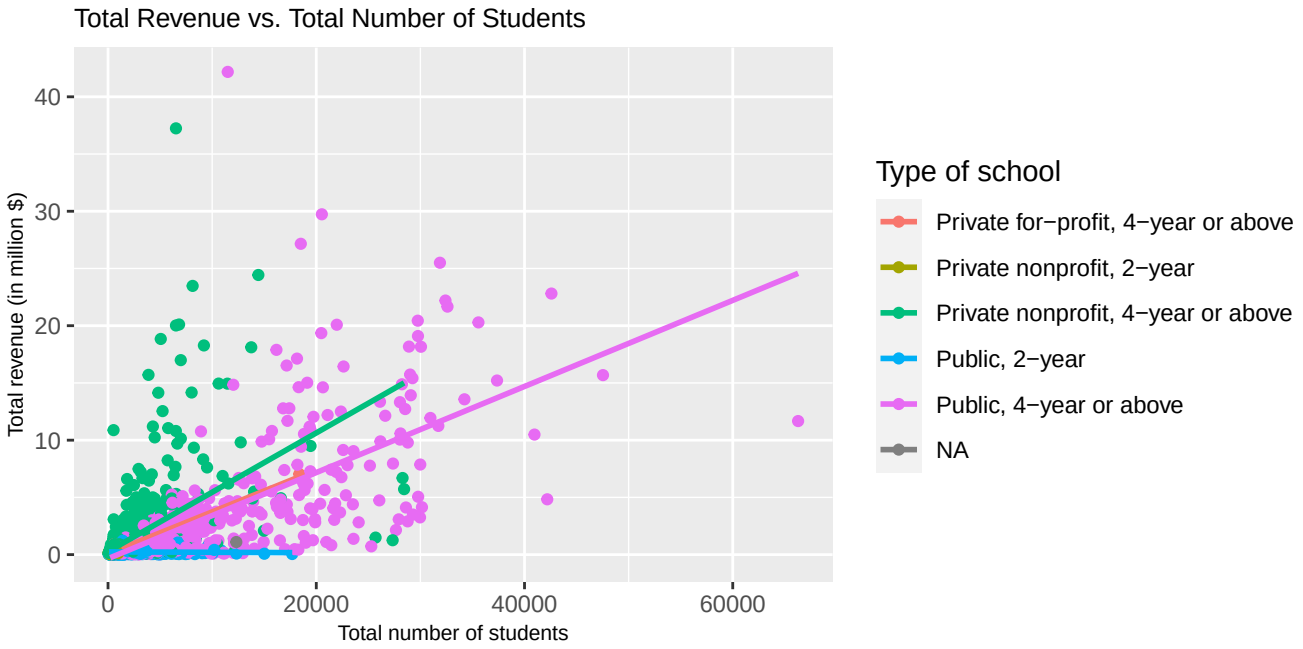



```
[1] 0.9641812
```

There is a positive, linear relationship between total expenditure and total revenue. The correlation between these two variables is 0.9641812, indicating a strong strength. There are more observations with total expenditure in the lower range, mostly concentrating between 0 and 20 million dollars. There are potential outliers where total expenditure is above 20 million dollars.

Public, 4 year or above and private nonprofit, 4-year or above have larger spreads and higher median of total revenue compared to the other types of school. The two types of school also have higher numbers of high values in total revenue, which are outliers exceeding the spread of most observations compared to the rest. Public, 4 year or above has the observation with the highest value of total revenue.

A potential interaction effect between total number of student and type of school



The lines are not parallel indicating there is an interaction effect. The slope of total number of student differs based on the type of school.

Analysis approach

Response variable: total_rev_menwomen, which is the total revenue in USD for men and women collegiate sports.

Potential predictors: participation rate (sum_partic_men and sum_partic_women), type of school (sector_name), sport division (classification_name), expenditures (total_exp_menwomen), gender, and total count of students (ef_total_count).

We plan to use multiple linear regression as our regression model technique to analyze the relationship between the multiple potential predictors and the response variable.

Data dictionary

The data dictionary can be found here