

# College's predicted profit by sport

Standard Deviants - Ava Exelbirt, Yura Heo, Claire Li

2023-11-07

## Introduction and data

In our research project, we aim to investigate and understand the factors influencing the profit 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 determinants of collegiate basketball profit, which can be of significant interest to educational institutions and further local policy makers and sports enthusiasts. Understanding the factors that contribute to profit generation in collegiate basketball sports can inform decision-making, investment strategies in sports, and future planning for universities and colleges involved in sports programs. We chose to do profit as our response variable because expenditures and revenue are closely related, so expenditure would overpower the other predictor variables if we were to predict revenue or expenditure. Therefore, we will create a new variable profit calculated from revenue minus expenditures and predict this.

Our primary research question is as follows: **In the school year 2019 to 2020, how can we predict the profit (revenue- expenditure) in USD of the collegiate sport basketball using participation rate, school sector name, gender ratio, total count of students, percent of expenditures towards women's sports, and school classification name.**

Our hypothesis is as follows: **Participation rate, sector name, and gender ratio will be the most influential predictors for the total profit generated by the collegiate sport basketball in USD.**

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<sup>1</sup>.

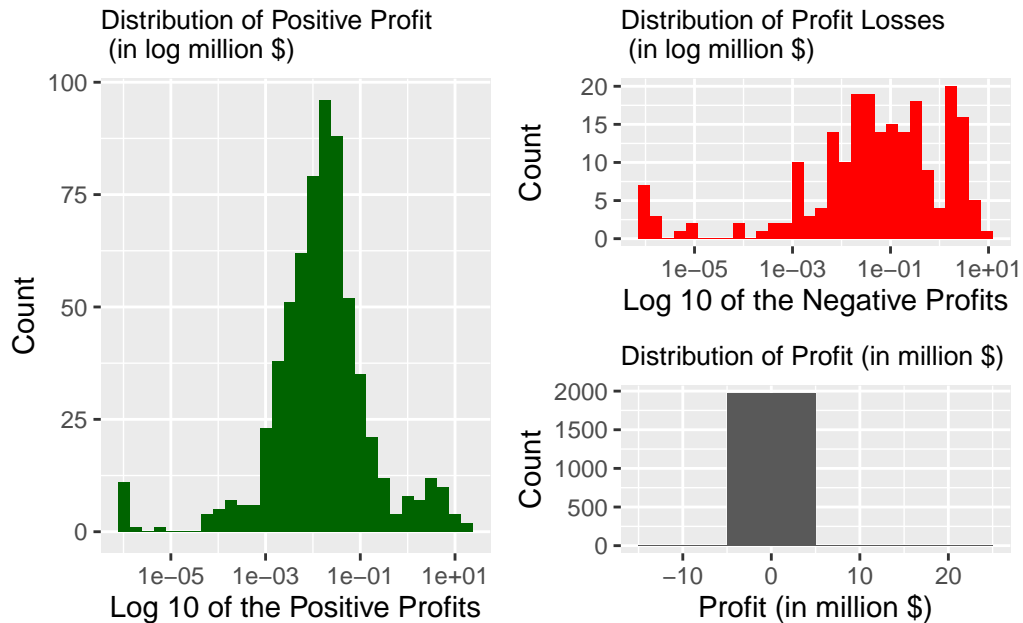
This data is submitted annually from colleges to the EADA. All co-educational postsecondary institutions that receive Title IV funding that have intercollegiate athletics programs are required by the Equity in Athletics Disclosure Act to submit this data. The original data files are also created immediately 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 project. The csv file from Tidy Tuesday contains thousands of observations from years 2015-2019, but we have adjusted the data file to only include the 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, 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.

The key variables we will use are **participation rate** which is the total percentage of men and women students who participate in sports, **sector school name** which is the type of school for example, public, 4-year or above, **gender ratio** which is the male population count divided by female population count, **total count of students** which is the total amount of students enrolled in the college, **women expenditure percent** which is the percentage of expenditures that goes towards woman's sports (which we calculated by taking a school's expenditures of women and dividing it by the school's total expenditures), **school classification name** which is a school's sports classification for example NCAA Division I-FCS. Then, our response variable is **profit** which is calculated by the total revenue of the school for basketball minus the expenditures of the school for basketball.

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 and total student count into the unit of hundreds, so that we can get larger coefficients for better modeling. We create the new response variable of profit and new predictor variable of woman's expenditure percent.

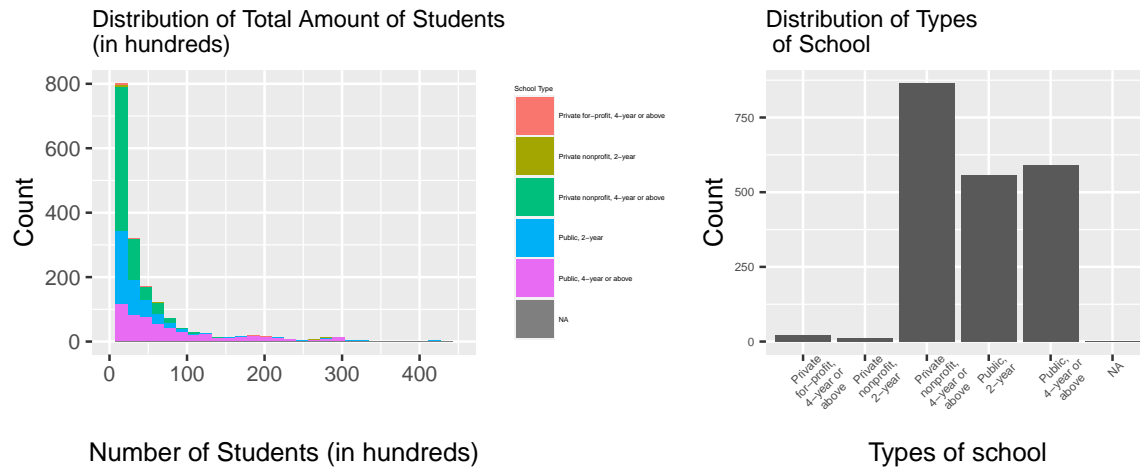
**Distribution of the response variable: total profit generated by college basketball**



```
# 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      0.0274      0.919    -9.67      0         0
  numeric.p75 numeric.p100
    <dbl>      <dbl>
1  0.00417      16.9
```

The distribution of profit is approximately normally distributed with a slight right skew. However, many observations centered around 0\$ in profit. The median of the data is at 0\$, and the mean is about 0.027 million USD. Since the response is approximately normal, we can use the mean as the center of the data. The range is from -9.6738 million USD to 16.8938 million USD. We also split up the distribution into positive and negative profits and took the log of the x axis so better see a more detailed distribution. It seems that positive profits is normally distributed while negative profits are left skewed. There are 57 missing values of profit.

**Distributions of total amount of students (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         0         41.7       59.5        0.5       10.6       20.1
  numeric.p75 numeric.p100
    <dbl>      <dbl>
1     46.3     663.
```

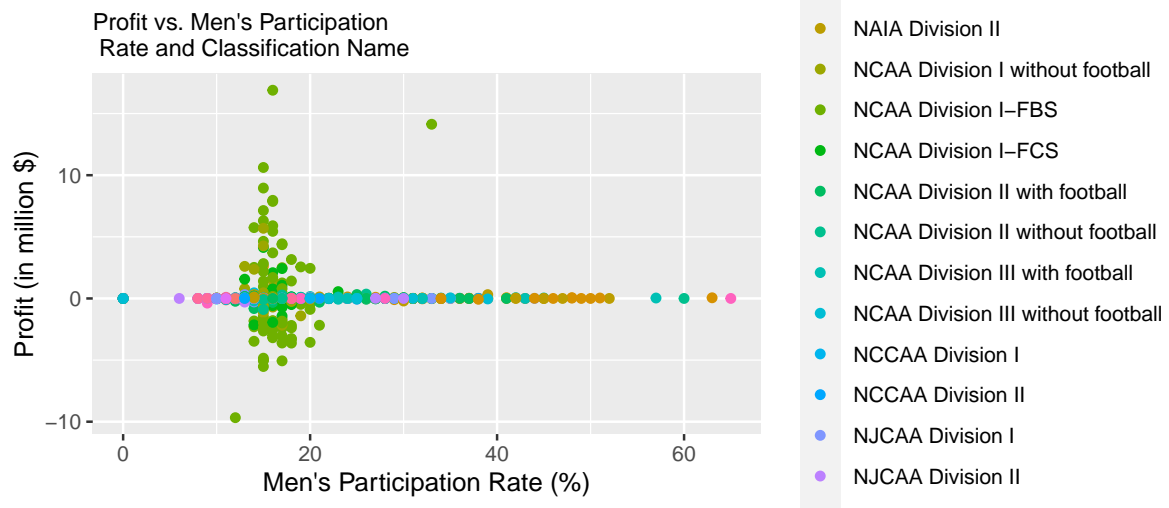
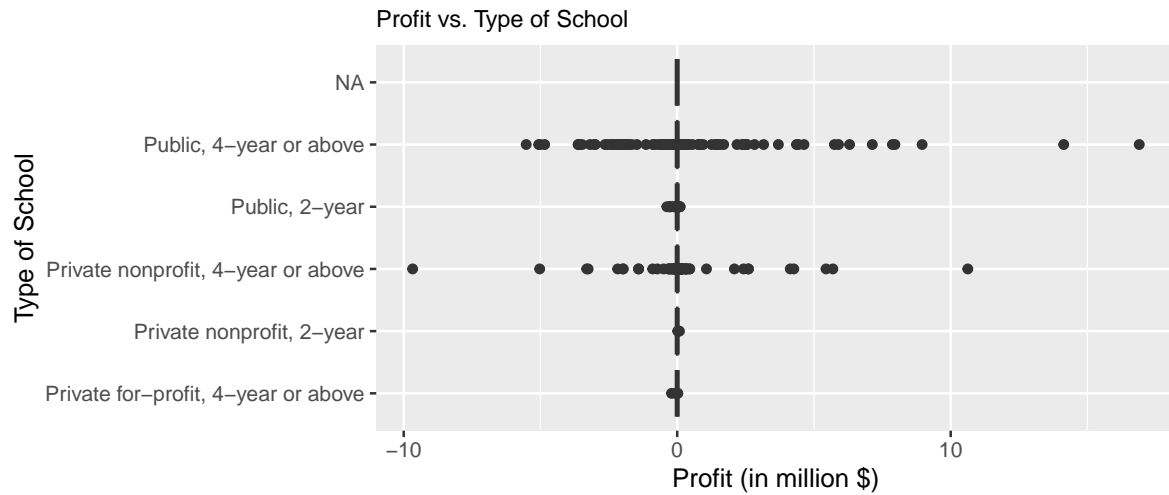
	n	p	nmiss
Private for-profit, 4-year or above	22	0.010757946	1
Private nonprofit, 2-year	11	0.005378973	NA
Private nonprofit, 4-year or above	865	0.422982885	NA
Public, 2-year	556	0.271882641	NA
Public, 4-year or above	591	0.288997555	NA

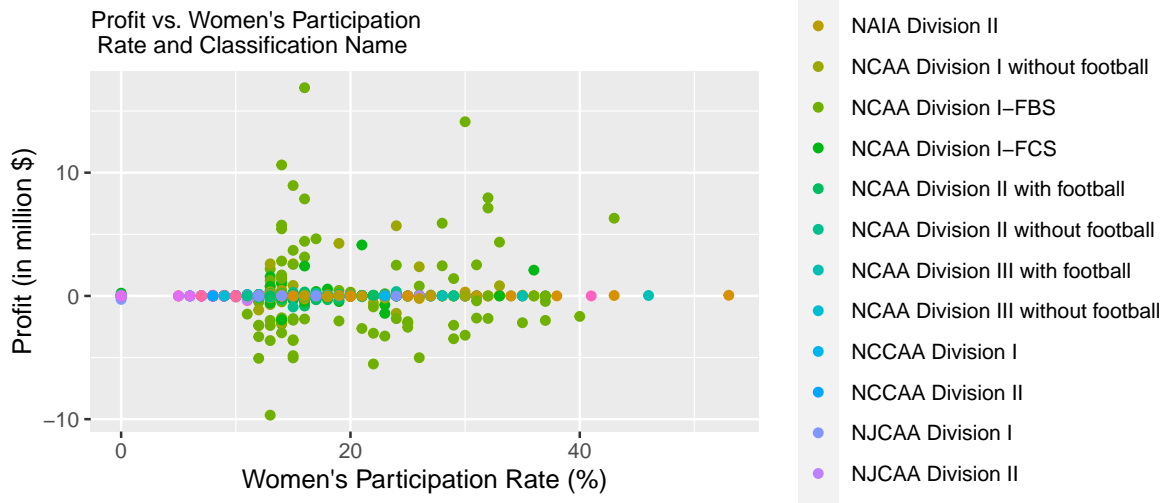
The distribution of total amount of students on college sports is right-skewed with most of the amount of student values in the lower range, while a number of observations have very high values that make them outliers. Given the apparent skewness, the center is the median of 2,005.5 students. Since the distribution is skewed, the IQR is used as a more reliable measure of spread which is  $Q3 - Q1 = 4,633 - 1,060.25 = 3,572.75$  students. You can also see that most of the schools with very high number of students are public, 4 year or above colleges.

There are 5 types of schools. 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 in types

of school.

## Relationship between Profit and a Categorical and Quantitative Predictor

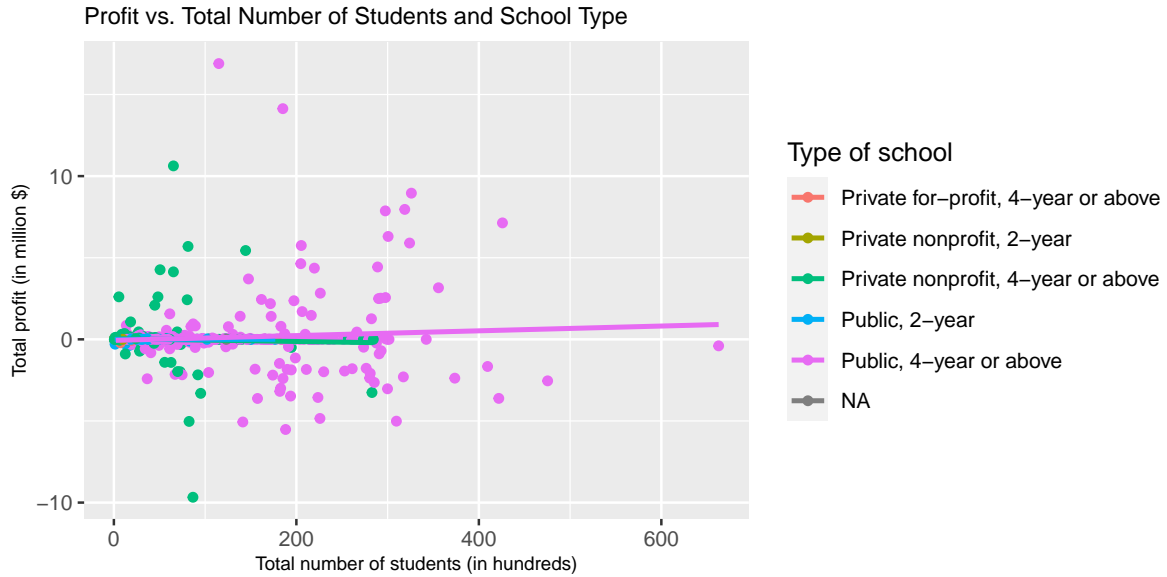




Looking at the relationships above, it seems that the medians of each type of school's profit is around \$0. We can also see that Public, 4 year or above colleges have the highest recorded profit in USD, and that Private nonprofit, 4 year or above colleges have the lowest profit in USD.

Looking at the participation rate graphs, there seems to be slightly more participation of women than of men, as many of the data points are more spread out to the right of the graph for women while there is a conglomerate of data points centered around 20% for men's participation rate. There seems to be a very weak almost negative linear relationship between participation rate versus profit, as most of the data points are centered around \$0 profit regardless of participation rate. However, looking at the mens participation rate specifically, it seems schools classified with sports as NCAA Division I-FBS have the highest profits and highest range of profits. Schools classified as NAIA Division II and NCAA Division II without football seem to have high mens participation rates. For womens participation rate, it shows the same thing as mens, however there tends to be greater participation rate of women than men in NCAA Division I-FBS schools.

**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.

## Methodology

We plan to use a multiple linear regression to predict the profit (revenue - expenditure) in USD of the collegiate sport basketball using participation rate, school sector name, gender ratio, percentage of women expenditure, total count of students, and school classification name. For the predictor gender ratio, we will mutate the data and divide the total male student count by total female student count. For the predictor percentage of women expenditure, we will mutate the data and divide a school's woman's expenditure by total expenditure.

For the response variable, we decided to predict the profit, subtracting the expenditure variable from the revenue variable, instead of predicting the revenue variable and using the expenditure variable as a predictor because we anticipate that the revenue variable and the expenditure variable will have strong correlations. Expenditure will be strongly correlated with the revenue of the sport basketball because schools often allocate significant financial resources to their sports 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.

For this reason, we decided that using expenditure to predict the revenue will not produce a meaningful result in choosing the best predictors as expenditure may already significantly affect the revenue, muting the effects of all the other predictors.

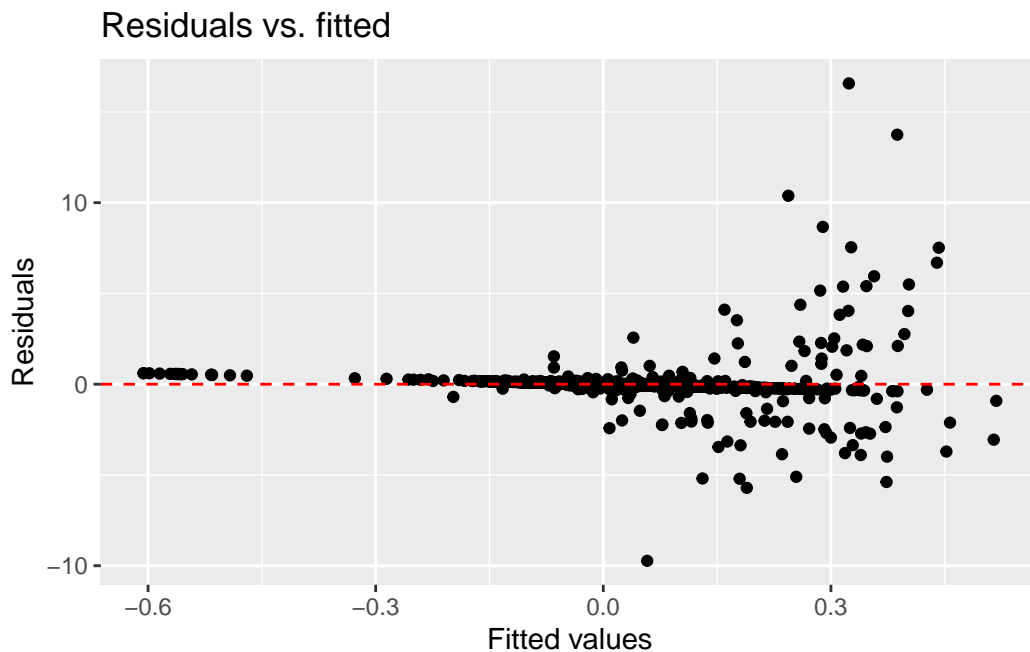
For the predictors, we anticipate that **participation rate** will be a key predictor of the total profit of the sport basketball because schools may allocate more money to this sport if there

is more participation from the students. We also expect that the type of school (**sector name**) will be a strong predictor of profit because different types of schools have varying levels of resources, alumni support, and participation rates in sports depending on school size and program. **Gender ratio** is another predictor to consider because sports could be hugely dependent on the demographics of players. Lastly, we included woman expenditure percentage as a predictor because we were curious to see if the proportion of spending on woman's sports would influence the sports profit of a school.

Additionally, we anticipate interaction effects which we will do further analysis in our results. Specifically, we believe the participation rate and the school sector type will have a correlation. This is because we expect the participation rate to be very dependent on the levels of investment, competitive levels, institutional culture, and student demographics which vary based on the type of institution(sectorname).

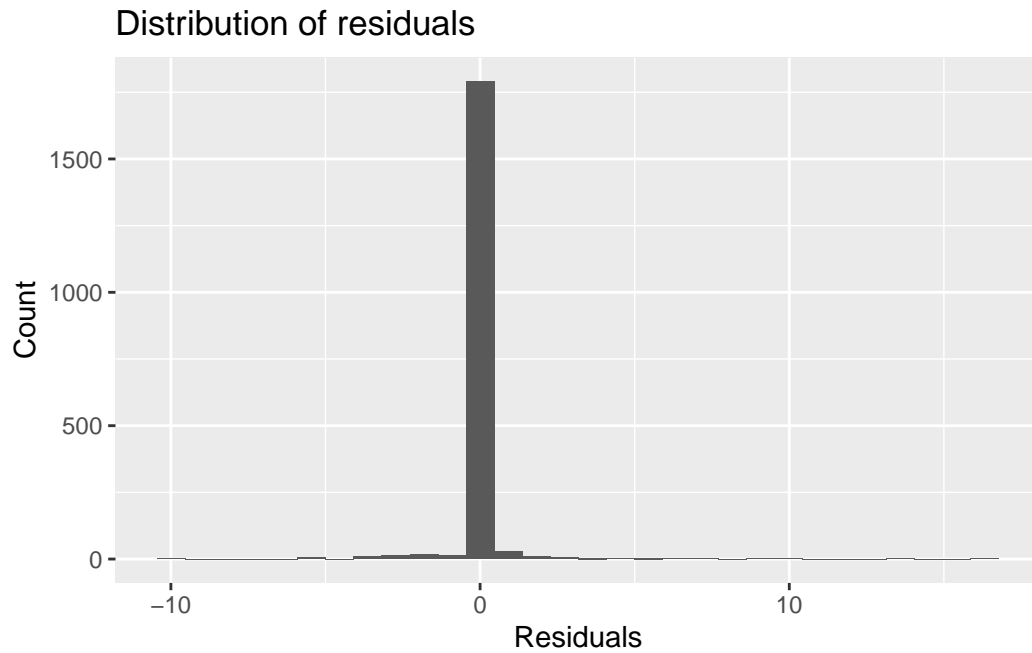
We first check the model's conditions.

*Checking Model Conditions:*



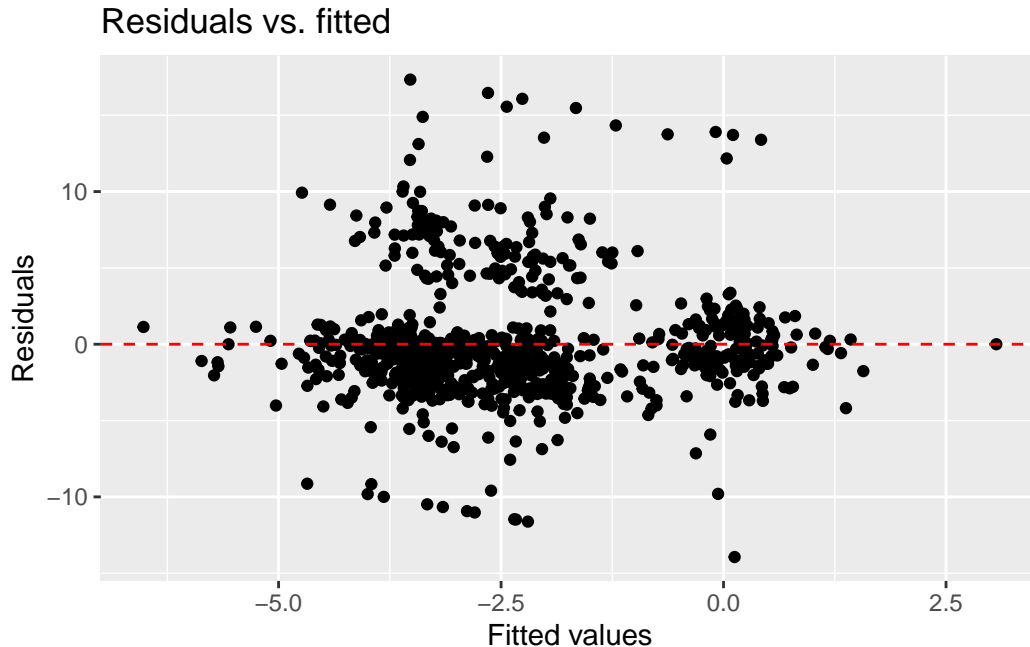
Constant variance is met because there is an even amount of residuals above and below the horizontal line. Linearity may not be met. We can see that there is a conglomerate of points that shape a almost negative line around the y intercept. However, there is a random scatter of points around this. This should be kept in mind and show that a linear model may not be the best regression model to use.





We can see the normality condition is satisfied as the residuals are normally distributed and the sample has more than 30 observations. Independence is also satisfied because of how the data was taken. Collegiate information about one school is independent of another school.

Since the linearity is not satisfied, we log transformed the response variable `profit`.



To avoid overfitting the data, we then split the dataset into the training and the test data, and we used cross validation using 20 folds. Then we used our recipe to fit our models. Using the models, we calculated AIC and BIC to evaluate each model fit.

To prepare the variables for our analysis, we plan to use a recipe. First, we plan to drop all the NA values using `step_naomit()`. We decided to drop NA values to ensure that the values we do use are accurate. For example, the data set has observations with NA values in women's participation rate. However, when the sum participation rate for women is calculated, the data set represents this sum with 0. This 0 can skew the regression model when the true participation rate of women may not be 0. Since we are using total participation rate in our model, it is best to drop NA values. Also, considering that the data set has 775 observations after all NA's are dropped, we still have a lot of reliable data to use for modeling. Then, we plan to use `step_zv()` to remove all predictors that contain only a single value. We plan to create dummy variables using `step_dummy()` for all nominal predictors which are Classification name and Sector name. Lastly, we will be using `step_center()` to mean center our quantitative predictors.

```
# A tibble: 2 x 6
  .metric .estimator  mean     n std_err .config
  <chr>   <chr>      <dbl> <int>  <dbl> <chr>
1 rmse    standard    4.14     20  0.180 Preprocessor1_Model1
2 rsq     standard    0.0897    20  0.0189 Preprocessor1_Model1
```

```
# A tibble: 1 x 3
```

```

      mean_adj_rsq mean_aic mean_bic
      <dbl>      <dbl>    <dbl>
1      0.0773     3444.    3545.

# A tibble: 2 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 rmse    standard     4.14    20  0.173 Preprocessor1_Model1
2 rsq     standard     0.0890   20  0.0196 Preprocessor1_Model1

# A tibble: 1 x 3
  mean_adj_rsq mean_aic mean_bic
  <dbl>      <dbl>    <dbl>
1      0.0720     3444.    3527.

```

Our final model uses the predictor of total participation rate, total count of students, gender ratio, and percentage of women expenditure.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.619	0.181	-14.498	0.000
total_partc	0.065	0.032	2.071	0.039
ef_total_count	0.006	0.003	2.318	0.021
gender_r	-0.363	0.394	-0.921	0.358
perc_wom_exp	-7.178	2.013	-3.566	0.000

- Model fit statistics

```

# A tibble: 2 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 rmse    standard     4.16    20  0.155 Preprocessor1_Model1
2 rsq     standard     0.0600   20  0.00880 Preprocessor1_Model1

# A tibble: 1 x 3
  mean_adj_rsq mean_aic mean_bic
  <dbl>      <dbl>    <dbl>
1      0.0467     3447.    3473.

```

The final model has a AIC of 3446.948 and a BIC of 3473.352.

## Results

Our initial model including total participation rate, school sector name, gender ratio, total student count, classification name, and percentage of women expenditure has a AIC of 3443.95 and a BIC of 3545.167. When we reduce the sector name variable, the AIC decreases to 3443.531 and the BIC decreases to 3527.144. We choose our final model that further reduce the classification name variable since it has the lowest AIC and BIC. We decide not to further reduce the terms in the model to have sufficient predictor variables for the response variable.

- Model diagnostics

total_partc	ef_total_count	gender_r	perc_wom_exp
1.028065	1.148082	1.019390	1.174146

There is no variable with  $VIF > 10$  that indicates concerning multicollinearity, so no apparent issue with multicollinearity is found.

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>         <dbl>
1 rmse    standard         5.11
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>         <dbl>
1 rmse    standard         5.06
```

The RMSE for the training data which is 5.111739 while the RMSE for testing data is 5.056632. Since we have a reasonably small difference in the RMSE, there is no apparent sign of model overfit. The model can generalize to new data.

- Interaction Terms

To see if there are interaction terms, we will be using `step_interact()`. Because we ended up reducing the model taking out the sector name variable, we decided to check if there were any interaction effects between gender ratio and percentage of women expenditure with these variables having a connection with gender.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.623	0.181	-14.487	0.000
total_partc	0.064	0.032	2.040	0.042
ef_total_count	0.006	0.003	2.233	0.026
gender_r	-0.350	0.396	-0.883	0.378
perc_wom_exp	-7.676	2.399	-3.199	0.001
gender_r_x_perc_wom_exp	-1.493	3.908	-0.382	0.703

Because the p value for `gender_r:perc_wom_exp` is about 0.703, we can conclude that there are no significant interaction effects between the two variables.

## Discussion + Conclusion

Looking at each predictor's p values, the three significant predictors are total participation rate, total count of students, and percentage of women participation, each with a p-value smaller than 0.05 where we can reject the null hypothesis that the predictor has no significant effect to log-transformed profit.

In terms of the coefficients, we can give the following interpretations: For each additional point in the total participation rate, the school's basketball profit is expected to multiply by a factor of 1.067, holding all else constant; For each additional hundred in the total count of students, the school's basketball profit is expected to multiply by a factor of 1.006, holding all else constant; For each additional point in the percentage of women expenditure, the school's basketball profit is expected to multiply by a factor of 0.001, holding all else constant.

Furthermore, when comparing the AIC and BIC of multiple models to predict profit, we concluded that the best model to fit the data was the model that includes total participation rate, total student count, gender ratio, and percentage of women expenditure. Checking to see if there were any interaction terms between `gender_r` and `perc_wom_exp` by using `step_recipe()`, we concluded that there are no potential interaction effects.

While our study provides a comprehensive analysis of the factors influencing profit in collegiate basketball, there is room for improvement. One of the limitations of our research was that there may be more predictors outside the scope of our dataset that influences the profit of collegiate basketball. Also, the school year 2019-2020 was mainly when COVID 19 affected many schools and especially sports activities. We anticipate that the data collected during this timeline may not be applicable for other years without COVID.

In conclusion, we anticipate that the implications of our research predicting profit generation in collegiate basketball holds profound meaning in multiple domains such as in academia, sports management, and educational governance to foster an environment conducive to the holistic development of student-athletes. As a future direction, we would like to not just look at trends from the school year 2019-2020, but delve into longitudinal trends within collegiate basketball profitability by analyzing data spanning multiple seasons. We would also like to further our study by incorporating external factors that may influence profitability such as global events and changes in sports culture to get a nuanced perspective of our predictive model.