A Statistical Analysis of How Budget Sizes of U.S. Youth Theater Companies Relate to Rating of Production Values

Word Count: 2867

Date: 20/12/24

Introduction

Storytelling is an art as old as time itself. For me, there has always been one specific form of storytelling that stood out: theater. As a kid, getting up on stage fascinated and terrified me. However, at 16, I can say with confidence that performing has become second nature. I've participated in shows with a variety of different theater companies of varying budget sizes, and it always intrigues me: **How does budget size affect the priorities of the leaders' of youth theater companies when it comes to production of plays?**

I will be using statistical modeling and analysis, specifically bar graphs, 1-variable statistics, lines of regression, and Pearson's Correlation Coefficient to answer this question. All of this data was collected through a study I found from ARTSPRAXIS magazine, a peer-reviewed journal that focuses on applied theater in education. The study, conducted by Matt Omasta and Aubrey Felty, asked 121 leaders of youth theater companies to rank their level of agreement with given statements on production values on a scale of 1 (strongly disagree) to 4 (strongly agree). (Omasta). The statements were as follows:

"In general, it is important that the shows my theater presents and/or produces…"

- Demonstrate aesthetic excellence ("*Aesthetic Excellence*")
- Are highly entertaining ("*Highly Entertaining*")
- Align with school curriculum ("*Align with School Curriculum*")
- Address social issues and diversity ("*Addresses Social Issues*")

Analysis of this data will show ties between the production values and a theater company's need to produce money through ticket sales. The knowledge of dependency on ticket sales makes budget size a significant factor when it comes to the inner workings of any given theater company. If a company isn't receiving that revenue, a potential contributing factor to the low budget, they will not devote much time or attention to values such as *Aligning With School Curriculum*, or *Addressing Social Issues.*

Limitations

A couple limitations to the data that I have access to are due to the fact that I didn't gather my own data. This data is a collection of mean rankings, which translates to a loss of intricacy within the data set.

Overall Data and Calculations

The table below shows the data on annual budgets grouped by size, and the mean rankings of the four production values.

		Category A (250,000 <)	Category B (250,000 to $999,999)	Category C (1 million - 2.99 million)	Category D (≥ $3 million)
Aesthetic excellence		3.72	3.81	3.81	3.90
Highly entertaining		3.39	3.46	3.48	3.26
Align School Curricula		2.50	2.77	2.88	2.71
Social issues and Diversity		3.28	3.50	3.57	3.68

Figure 1a. N=121

Below is a visual representation of the above data, and shows how different production values were ranked across different budget sizes. The letters directly correspond to the budget categories.

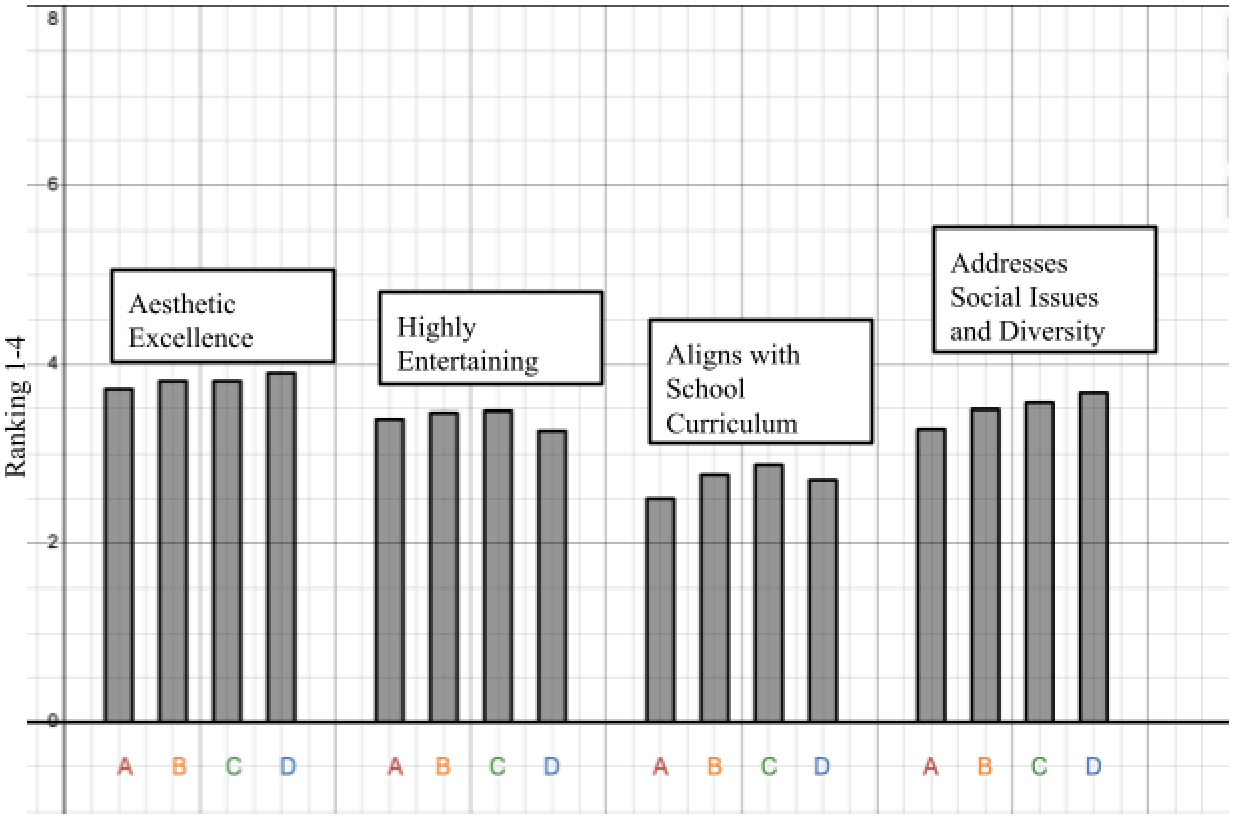

Figure 1b.

The graph above shows how different budget categories (A, B, C, D) rank the production values (*Aesthetic Excellence, Highly Entertaining, Alignment with School Curriculum, and Addressing Social Issues*) In this graph, one can look at how the ranking of the same values differ across budgets. While *Aesthetic Excellence* ranks the highest overall in any budget category, the most interesting observations can be found when looking at the remaining three production values.

Standard Deviation Sample Calculations

Calculating the standard deviation between the different rows will show how far the rankings are spread out from one another across the different production values.

The formula for sample standard deviation is as follows.

$$\sqrt{\frac{\Sigma_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}$$

This formula calculates the sample standard deviation by summing each of the squared differences between a data value and the sample mean should be added up. For the first column, the different rankings for aesthetic excellence, first the sample mean is taken for these averaged rankings of one production value. The formula for finding the mean is as follows:

$$\overline{x} = \frac{\Sigma x_i}{n}$$

Step One maps out the numbers that one substitutes into the above equation. We will take the sum of x_i, which in this context means all of our rankings for the given production value, which then gets divided by n, which is the number of rankings available. Taking the sum is represented by Σ.

$$1.\overline{x} = \frac{(3.72+3.81+3.81+3.90)}{4}$$

$$2.\overline{x} = \frac{(15.24)}{4}$$

Step Two is fairly self-explanatory. 15.24 is the sum of all of the rankings.

$$3.\,\overline{x} = 3.81$$

After calculating the sample mean, shown in Step Three, we can calculate the sample standard deviation. x_i represents each specific ranking, and $\overline{x}$ represents the sample mean. When we calculate standard deviation, we square the difference between the data and the sample mean for each value, and then take their sum (Σ). Then, it's all divided by $n - 1$, which is the number of observations, minus one. Finally, the square root of this value is taken, thus being left with the sample standard deviation.

$$Sx = \sqrt{\frac{\Sigma_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}$$

Step One applies the above formula to our values.

$$1.\sqrt{\frac{(3.72-3.81)^2+(3.81-3.81)^2+(3.81-3.81)^2+(3.90-3.81)^2}{4-1}}$$

In order, to get to Step Two, the sample mean is subtracted from each ranking, as well as subtracting one from the n value, in order to get 3.

$$2.\sqrt{\frac{(-0.09)^2+(0)^2+(0)^2+(0.09)^2}{3}}$$

Step Three shows the result of squaring the values.

$$3.\sqrt{\frac{(0.0081)+(0)\ +(0)\ +(0.0081)}{3}}$$

Step Four shows the sum of the squared values.

$$4.\sqrt{\frac{0.0162}{3}}$$

Step Five shows the result of dividing by $n - 1$

$$5.\sqrt{0.0054}$$

After taking the square root, shown above, we are left with the sample standard deviation, in Step Six.

$$6. Sx = 0.0734846923$$

This value represents the distance of each ranking from the sample mean. If the standard deviation is low, it indicates that the data is clustered around the sample mean value, while data with a high standard deviation is more spread out. In application, these calculations could be used to look at which value would be the most unifying or varied. In the table below (1c) the same process demonstrated above was repeated for all of the production values:

Category	Standard deviation
Aesthetic excellence	0.0735
Highly entertaining	0.0995
Algins	0.1597
Social issues	0.1688

Figure 1c.

The production value with the highest standard deviation in its ranking is *Addresses Social Issues,* and the production value with the lowest deviation in responses is *Aesthetically Excellent.* This calculation allows for a confident assertion that certain values are more varied than others among differing budget sizes.

Pearson's Correlation Coefficient Sample Calculations

However, in order to look at reasons why that might be, it would be important to see if these same values have a correlation in the way that they change with respect to increasing budget sizes. In order to find this, one can use Pearson’s Correlation Coefficient (PCC). The Pearson's Correlation Coefficient measures the strength and direction of a linear relationship between two variables. The formula is as follows.

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}}$$

In the numerator, for each x and y pair, the difference between the sample means is found ($\overline{x}$ or $\overline{y}$, depending) and then multiplied. Then, the sum of all the resulting products is taken. In the denominator, for each x and y pair, again the difference between the sample means and each given x or y value is found. However, then the differences are squared. After taking the square of each, the sum of all the x-value squares is taken. Repeating the same process with the y-values, these summed squares are multiplied together. To begin calculating the PCC, although the sample mean for the rankings of *Aesthetic Excellence* were already calculated, solving for the sample mean for our x-axis values is needed. While the sample mean of the rankings was represented as $\overline{x}$ previously, in this calculation, the sample mean for the rankings will be notated as $\overline{y}$. This is because the ranking values are on the y-axis. In order to find the sample standard deviation for the x-axis, the same process used above is followed.

$$\overline{x} = 2.5$$

As shown below, Step One will simply plug our values into this equation.

$$1.\ r = \frac{\Sigma(1-2.5)(3.72-3.81)+(2-2.5)(3.81-3.81)+(3-2.5)(3.81-3.81)+(4-2.5)(3.90-3.81)}{\sqrt{\Sigma(1-2.5)^2+ (2-2.5)^2+(3-2.5)^2+ (4-2.5)^2\ \cdot\ \Sigma(3.72-3.81)^2+(3.81-3.81)^2+(3.81-3.81)^2+(3.90-3.81)^2}}$$

Step Two subtracts all of the sample means from the data values, otherwise stated as the operations within parentheses.

$$2.\ r = \frac{\Sigma(-1.5)(-0.09)+(-0.5)(0)+(0.5)(0)+(1.5)(0.09)}{\sqrt{\Sigma(-1.5)^2+(-0.5)^2+(0.5)^2+(1.5)^2\cdot\Sigma(-0.09)^2+(0)^2+(0)^2+(0.09)^2}}$$

In Step Three, different operations are carried out on the numerator and the denominator. For the numerator, Step 3 multiplies the values on the top, and on the bottom, each of the differences between sample mean and data value are squared.

$$3.\ r = \frac{\Sigma(0.135)+(0)+(0)+(0.135)}{\sqrt{\Sigma(2.25)+ (0.25)+(0.25)+(2.25)\cdot\Sigma(0.0081)+(0)+(0)+(0.0081)}}$$

Step 4 shows what happens after taking the sum of the multiplied or squared differences.

$$4.\ r = \frac{0.27}{\sqrt{5\cdot 0.0162}}$$

Step 5 multiplies the sums of the x and y squared differences for the denominator.

$$5.\ r = \frac{0.27}{\sqrt{0.081}}$$

Step 6 takes the square root of 0.081

$$6.\quad r = \frac{0.27}{0.2846049894}$$

Step 7 is the final answer, converting the fraction into a decimal.

$$7.\ r = \frac{0.27}{0.2846049894} = 0.9487$$

This process was repeated for the rest of the values in this data table, with both r and R^2 values being included. The r value will show us the strength of the relationship between our values, and what direction they are going.

	r-value	R^2 value
Aesthetically Excellent	0.9487	0.9
Highly Entertaining	-0.4803	0.2307
Aligns with School Curriculum	0.5983	0.3579
Addresses Social Issues and Diversity	0.9713	0.9435

Figure 1d

Line of Regression Sample Calculations

The line of regression will show how much the ranking of production values will change given a one unit change in the budget size. The equation to calculate the line of regression consists of separate equations for slope and intercept. The equation for slope is as follows.

$$b_1 = \frac{(n(\Sigma xy) - (\Sigma x)(\Sigma y))}{(n(\Sigma x^2) - (\Sigma x)^2)}$$

Organizing into a table helps the calculated values. Below are the x and y values we start off with.

x_1	y_1
1	3.72
2	3.81
3	3.81
4	3.90

For the line of regression slope calculations, finding xy, x^2, and y^2 for each row is needed, as well as the sum of all the columns, after they have been calculated.

	x_1	y_1	xy	x^2	y^2
	1	3.72	$1 \cdot 3.72 = 3.72$	$(1)(1) = 1$	$(3.72)(3.72)\ 13.8384$
	2	3.81	$2 \cdot 3.81 = 7.62$	$(2)(2) = 4$	$(3.81)(3.81)\ 14.5161$
	3	3.81	$3 \cdot 3.81 = 11.43$	$(3)(3) = 9$	$(3.81)(3.81)\ 14.5161$
	4	3.90	$4 \cdot 3.90 = 15.6$	$(4)(4) = 16$	$(3.90)(3.90)\ 15.21$
Σ	10	15.24	38.36	30	58.0806

Figure 1e.

$$b_1 = \frac{(n(\Sigma xy) - (\Sigma x)(\Sigma y))}{(n(\Sigma x^2) - (\Sigma x)^2)}$$

In this equation, n represents the number of observations. For these calculations, we only have 4. Since the values needed for this equation have already been calculated in the table above, the next step is substituting them in.

$$b_1 = \frac{(4(38.36)-(10)(15.24))}{(4(30)-(10)^2)}$$

$$b_1 = \frac{1.04}{20}$$

$$b_1 = 0.054$$

In order to determine the y- intercept, to complete the line of regression calculations, the equation:

$$b_0 = \frac{((\Sigma y)(\Sigma x^2)-(\Sigma x)(\Sigma xy))}{(n(\Sigma x^2)-(\Sigma x)^2)}$$

$$b_0 = \frac{((15.24)(30)-(10)(38.36))}{(4(30)-(10)^2)}$$

$$b_0 = \frac{73.6}{20}$$

$$b_0 = 3.68$$

To construct the line of regression for the *Aesthetic Excellence* values, the formula is as follows:

$$y = b_1x + b_0$$

$$y = 0.054x + 3.68$$

Using the same equation, all the lines of regression for our data are organized into the table below.

Production value	**Line of regression**
Aesthetic Excellence	$y = 0.054x + 3.68$
Highly Entertaining	$y =- 0.037x \ + 3.49$
Aligns with School Curriculum	$y = 0.074x + 2.53$
Addresses Social Issues and Diversity	$y = 0.127x + 3.19$

Figure 1f.

To analyze the results of the sample calculations in a way that helps answer the research question, each production value should be evaluated separately. First, the analysis examines *Aesthetically Excellent*, then *Highly Entertaining*, *Aligns with School Curriculum*, and *Addresses Social Issues and Diversity*.

1-Aesthetic Excellence

The underlying survey asked respondents "On a scale of 1-4, how important to you is it that your theater company produces a play that is aesthetically excellent?"

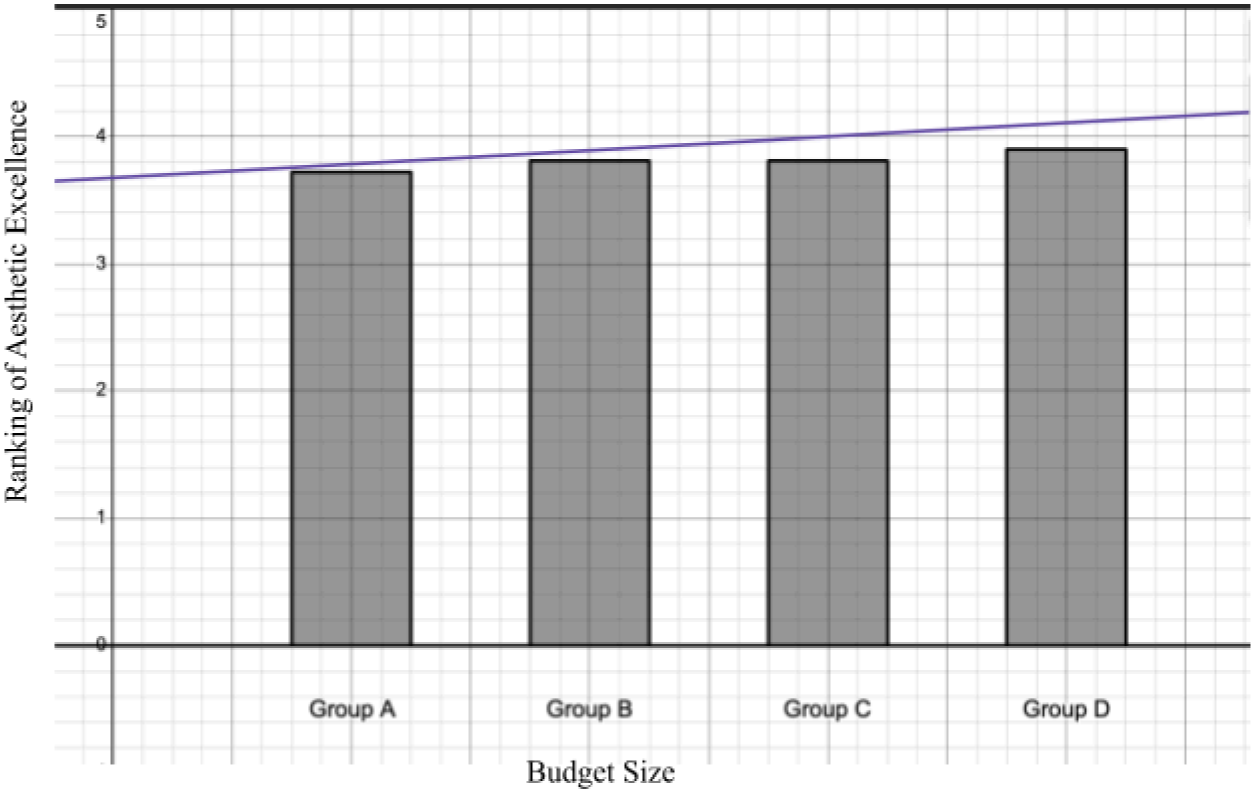

Figure 2a.

x_1	y_1
1	3.72
2	3.81
3	3.81
4	3.90

The graph above is a visual model of the rankings of *Aesthetic Excellence* across budget size, with the corresponding table. From the sample calculations above (Tables 1d, 1f), the r value for this data is 0.9487, and that the R^2 value is 0.9, with the line of regression being

$y = 0.054x + 3.675$

Using the PCC, this data demonstrates a strong positive linear relationship, and the R^2 , with a value of 0.9, shows that the regression line fits the data well. Thus, this strong positive linear relationship represents an accurate way to model the relationship of the rankings of this value.

Deploying the line of regression, if the original collectors of this data were to theoretically poll theaters with budget sizes that are larger than Group D ($3 million), and construct a "Group E", using the line of regression, an estimate of what their ranking of this production value could be anticipated by substitution of the values into the formula. In this equation, the hypothetical bar graph for "Group E" would have an x coordinate of 5, which is the value used in the equation below.

$$y = 0.054x + 3.675$$

$$y = 0.054(5) + 3.675$$

$$y = 3.94$$

However, this value is an extrapolation, and there would be other factors within the theater community, such as values and beliefs that would make this number somewhat difficult to obtain as an average ranking.

2-Highly Entertaining

The underlying survey asked respondents "On a scale from 1-4, how important is it to you that your theater company produces plays that are highly entertaining?"

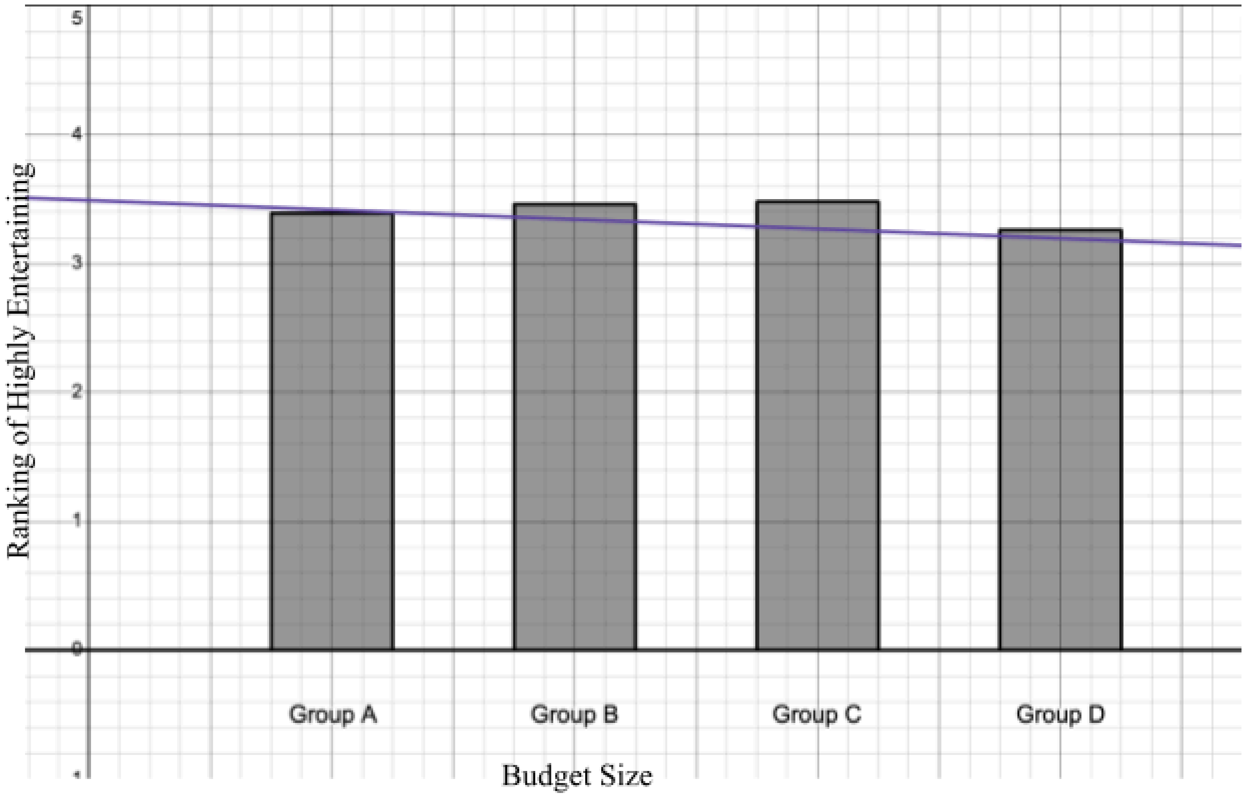

Figure 3a.

x_1	y_1
1	3.39
2	3.46
3	3.48
4	3.26

Figure 3b.

The line of regression for this data, as shown in the Table 1f, above, is $y =- 0.037x + 3.49$, with a r value equaling -0.4803, and an R^2 value of 0.2307. This data set has a weak negative linear relationship, as well as an R^2 value that suggests that this line of regression is not the most accurate way to analyze the relationships of this data set. Therefore, using this line of regression to estimate other theoretical data values would not yield trustworthy results.

3-Aligns with School Curriculum

The underlying survey asked respondents “On a scale of 1-4, how important is it to you that your theater company produces plays that align with school curriculum?”

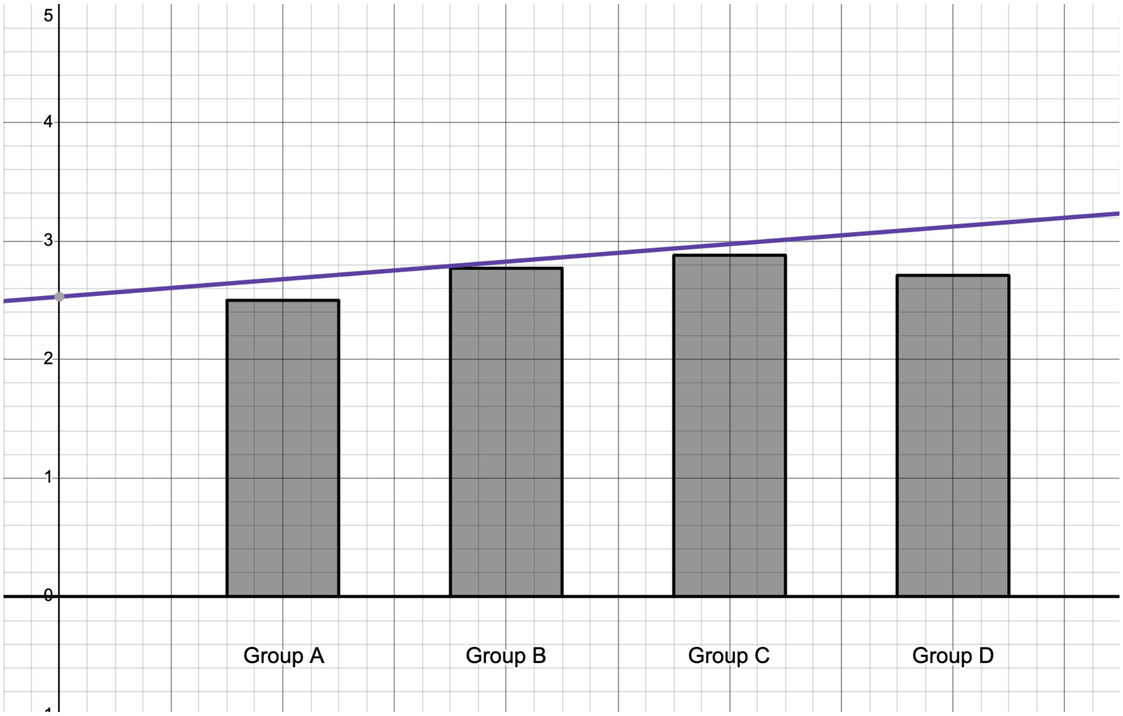

Figure 4a.

x_1	y_1
1	2.50
2	2.77
3	2.88
4	2.71

Figure 4b.

First, this data reflects comparatively low rankings. Also, while the rankings seem to increase along with budget size, at the point of the largest budget size, the value drops back down again.

The line of regression for this data set is $y = 0.074x + 2.53$, and it has a r value of 0.5983, as well as an R^2 value that equals 0.3579. This correlation coefficient indicates that this data set has a moderate positive linear relationship, which shows that the two variables examined (budget size and ranking of *Aligns with School Curriculum)* do not have a strong relationship. With an R^2 value below 0.5, which is typically recommended, this model could not be used to predict other ranking values.

4-Addresses Social Issues and Diversity

The underlying survey asked respondents "On a scale from 1-4, how important is it that your theater company produces shows that address social issues and diversity?

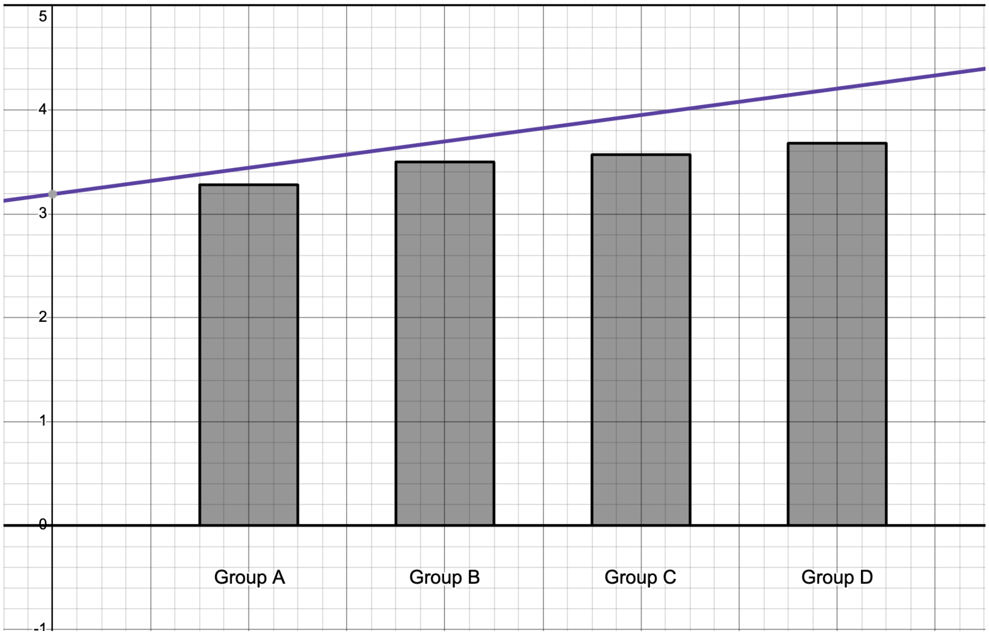

Figure 5a.

x_1	y_1
1	3.28
2	3.50
3	3.57
4	3.68

Figure 5b.

This data has a regression line with an equation of $y = 0.127x + 3.19$. With a r value of 0.9713, this data has a strong positive linear correlation. Furthermore, this strong correlation shows the model fits our data quite well, as the R^2 value equals 0.9435. Given these values, ranking for the aforementioned theoretical "Category E" could be predicted. Again, it will be assigned a x-value of 5.

$$y = 0.127x + 3.19$$

$$y = 0.127\,(5) + 3.19$$

$$y = 3.825$$

Similar to the "Category E" calculator for the production value of *Aesthetic Excellence*, this value is an extrapolation, because 5 is not a value included in our data set. However, one could use it as a basis for an estimate.

Conclusion

The data set used in this paper presents many opportunities for analysis. From the mathematical analysis conducted and the answers reached, I can make connections to my own personal knowledge of youth theater companies.

First, the strong linear correlation found in Section 4: *Addresses Social Issues and Diversity.* The strong positive linear correlation of 0.94 between budget size and ranking of values shows that as the budget size increases, the ranking of that specific production value increases to match it. A potential reason behind the increasing nature of this production value's rankings could simply be lack of ticket sales. For many theater companies, ticket sales are their primary source of revenue. For example, if a lower budget theater company chose to put on a play that features minority voices, or highlights controversial social issues, due to the highly exclusionary canon of theater works, it is likely to be a less well-known play. Theater companies need consumer-driven revenue, and this lesser-known play could drive potential consumers away. The risk for lower budget theater companies with this potential loss of revenue from ticket sales would be a higher risk than for a theater company with a higher budget.

Second, the pattern observed for production values 2 (*Highly Entertaining*) and 3 (*Aligns With School Curriculum*), an increase in rating as the budget sizes go up, until the highest budget size. In fact, if a PCC was calculated for the first three values, on both the 2nd and 3rd data sets, excluding the largest budget, it would show how much more correlated these values would be. For Value 2: *Highly Entertaining*, the new r value would be 0.952. Along with this, for Section 3: *Aligns with School Curriculum*, the new r value would equal 0.9717. By comparison to the original values, they show a dramatic increase in correlation coefficients. The r values for both values are now strong positive linear correlations. Such an analysis could show that higher budget theaters don't need their audiences to find everything they perform to be highly entertaining. Or, for the other set of values, that they don't have to prioritize aligning with the curriculum taught in schools. Furthermore, many low to moderate budget size theater companies could be school based, which may require them to consider the values around school curriculum more than their higher budget counterparts.

Finally, the analysis of this paper points to several meaningful additional research questions. Further exploration could include collecting data on how the different budgets of theater companies are spent. For example, a larger budget theater company could simply mean that they have more employees, and have to pay them more. Along with that, it would be interesting to look at demographics of the lower budget theater companies and the underlying inequities.

Works Cited

Omasta, M., & Felty, A. (2021). Perspectives of theatre for young audiences companies' leaders by budget, region, and longevity:A survey. ArtsPraxis, 8 (2), Appendix A, 1-26.