Analysis of Ticket Sales for CityVille Arts Theatre

Tyler Oh, Breanna Blackwell

Jul 28, 2023

1. Introduction

The initial purpose of this report was to maximize profits for Cityville Arts Theatre through data analysis using 25 years of Broadway ticket sales data (from 1990 to 2016) provided by the theater. This data included the show name, year of show, total gross revenue, number of performances, and type of show (musical, play, or special). From this information, we realized we lacked critical data like performance costs and ticket prices, leading us to reconsider our purpose. After studying the situation, we decided to retain only the total gross revenue, number of performances, and type of show for analysis. The main question now was: Does the type of show influence the total gross revenue for Broadway shows at Cityville Arts Theatre?

In this scenario, we used the type of show and number of performances as explanatory variables, with total gross revenue as the response variable. In the analysis stage, we applied methods like Transformation, linear regression, ANOVA test, and Tukey test.

In the initial analysis, we found that the total gross revenue and the number of performances did not follow a normal distribution. Therefore, we used transformation to make log(Total.gross) and log(number of performance) conform to normal distribution for smooth testing. To explore the combined impact of the number of performances and type of show on total gross income, we analyzed it using a linear regression model, also hoping to predict future ticket sales. Since the type of show is a categorical variable, we used ANOVA to check if there were significant differences in the average total gross revenue for different types of shows like musicals, plays, and specials. In addition, ANOVA tests can also help analyze variation in data to determine variation between and within different groups. This helped the theater understand if the type of

show was a major factor influencing total gross variation. As a follow-up to the ANOVA test, the Tukey test was used mainly to perform pairwise comparisons and pinpoint specific significant differences in total gross income between types of shows like musicals, plays, and specials. By accurately identifying these differences and controlling for Type I errors, it provided the theater with more targeted information to make wiser decisions and plan.

After a series of data analyses, we have concluded that there are significant differences in the average total gross revenue among different types of shows. Among these three types, musicals make the most money, followed by special shows, and then play shows. Both linear regression analysis and ANOVA test indicate that there is a significant difference in the average total gross revenue for at least one type of shows. Our Tukey test analysis further found that the average total gross revenue from musicals is significantly higher than that from special shows and plays, and the revenue from special shows is also significantly higher than that from plays. Therefore, different types of shows have a significant impact on the total gross revenue.

2. Statistical Methods

The provided dataset contains information on 1458 Broadway shows that played over 25 years, from 1990 to 2016. The dataset consists of variables, including the show's name, the year it played, the show type, total gross revenue (in dollars), and the number of performances.

We clean the data by excluding entries with 0 performances yet showing a total gross income, ensuring the accuracy of average gross income performance. After omitting these entries, our sample consists of 1428 observations.

Our response variable is total gross revenue, which is the money made from sales before deducting expenses. The explanatory variables are the number of performances, and the type of show, categorized as either musical, play or special shows.

We use linear regression to investigate the relationship among total gross income, the number of performances, and show type. We plot residuals to assess normality assumptions and perform data transformations when necessary. Additionally, we conduct an ANOVA test to determine significant differences in the log mean total gross revenue across show types.

3. Results

We start by categorizing the show type explanatory variables into three levels: musical, play, and special. This categorization facilitates an in-depth exploration of how each show type impacts the response variable, total gross revenue.

Coefficients	Estimate	SE	P-value
Intercept	987188	583919	0.0911
Play	-4268544	604999	2.68e-12
Special	-1424242	1351466	0.2921
Number of Performances	88144	2026	< 2e-16

Table 1. Summary of Fitted Linear Model

As demonstrated in Table 1, the coefficient for the play show type is -4268544. It implies an estimated $4,268,544 lower total gross revenue for plays compared to musicals on average. Also, special show total gross revenue is estimated to be $1,424,242 lower than that of musicals

on average. However, since the p-value for the play show type is greater than 0.05, we cannot confidently confirm that this difference is reliable. Overall, our test highlights the significant influence of show type (play or special) on generated gross revenue. In particular, plays and special shows exhibit relatively lower revenue compared to musicals at the theatre.

We plot the residuals against fitted values (Figure 1) to assess the normality diagnostics. The y-axis of the plot ranges from -2e+07 to 6e+07, indicating higher values within the dataset. The normal Q-Q plot in Figure 2 shows a deviation from the x=y line in the upper right quadrant, suggesting a potential violation of the normality assumption. Consequently, we transform the

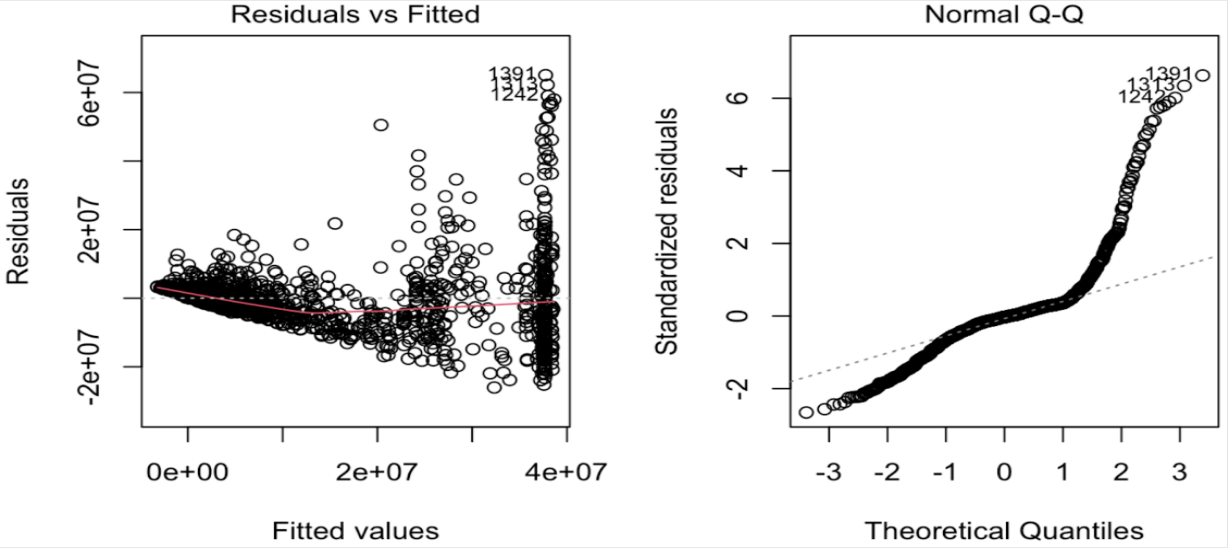

Figure 1. Residuals vs Fitted Values plot

Figure 2. Normal Q-Q plot

data by taking the logarithm of one or both numerical variables. Through experimentation, we determine that the best-fitted model for our analysis is achieved by applying the logarithm to both the number of performances and total gross revenue.

Moreover, both plots show numerous outliers, which indicates potential deviations from the assumptions of the linear regression model. These outliers can impact normality assumptions,

leading to skewed or heavily tailed distributions in the residuals that could potentially affect the model accuracy. However, it is important to note that these outliers, identified through normality diagnostics, represent values of total gross. Hence, we proceed our analysis by keeping these outliers.

Coefficients	Estimate	SE	P-value
Intercept	11.21588	0.07285	< 2e-16
log(Number of Performances)	1.01122	0.01415	< 2e-16
Play	-0.56207	0.03476	< 2e-16
Special	-0.22284	0.08472	0.00862

Table 2: Summary of fitted model after log-transformation

Following that, we fit our initial regression model with a log-log transformation. As shown in Table 2, this transformation results in a significantly improved fit. A relatively higher Multiple R-squared value (83.92%) suggests that the log-log transformation enhances the model's ability to explain variance in the log-transformed total gross revenue. Besides, the coefficients from Table 2 provides enhanced insights into the relationships among the variables. In particular, we can now illustrate that log total gross revenue decreases by approximately 0.56 units when considering the play show type. This conclusion was not achievable since the p-value of the play show type was not above the threshold of 0.05. However, the log-transformation yields more meaningful and statistically significant results that allow us to build prediction models.

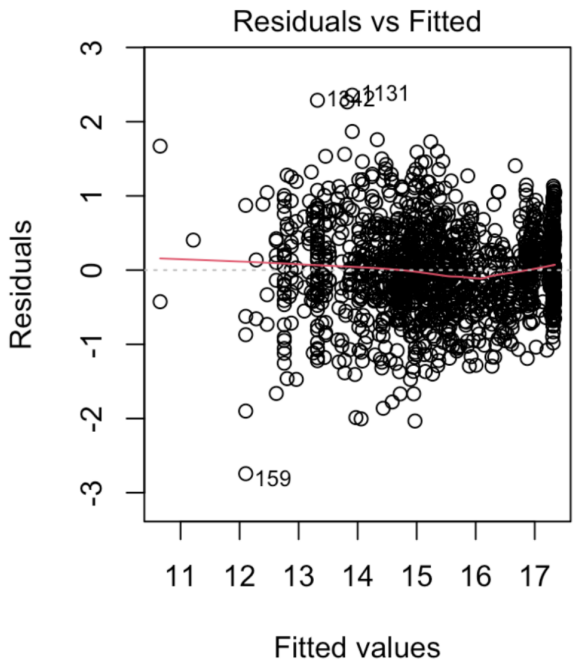

Figure 3: Residuals vs Fitted Plot After Transformation

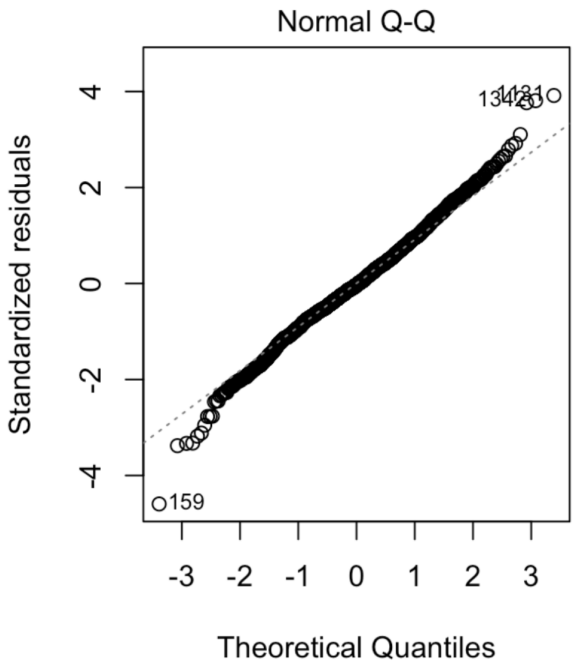

Figure 4: Q-Q Plot After Transformation

Figure 3 displays an even distribution of residuals around the mean line. It indicates that our model's assumptions of normality are well met. In Figure 4, the log-log transformation significantly improves the distribution of standardized residuals, resulting in a more normal pattern despite a few outliers. The near 45-degree alignment in the normal Q-Q plot also indicates a strong approximation of normality, confirming the effectiveness of the transformation in capturing the data's underlying distribution.

Model	RSS	F-Statistics	P-value
Reduced (Without Show Type)	613.40	132.82	< 2e-16
Full (With Show Type)	516.96		

Table 3: ANOVA Testing comparing Reduced Model and Full Model

Using transformation, we perform an ANOVA analysis to determine if there is a statistical difference in mean log total gross income across show types, while accounting for the log number of performances. As previously explained, since the dataset represents total gross income rather than total profit, we make an adjustment for the log number of performances.

We start by fitting a model with both the log number of performances and show type as explanatory variables, and then fitting a reduced model containing only the log number of performances as the explanatory variable.

Upon generating an ANOVA table that compares the fit of the full and reduced models, we observe an F-statistic of 132.82. Although F-statistics of 25.492 from ANOVA testing before transformation was significant, there is a noticeable enhancement in the fit of the full model compared to the reduced model. Furthermore, the p-value being less than the significance level of 0.05 provides strong evidence that the improvement is statistically significant.

Based on the low p-value and substantial F-statistics, we can confidently conclude that the full model, which includes the additional explanatory variable show type, significantly enhances the model's ability to explain the variation in log total gross revenue compared to reduced model.

The resulting p-value of 2.2e-16 provides evidence, at the 5% significance level, of a significant association between the show type and log total gross revenue. Furthermore, the exceptionally small p-value indicates that there are noticeable differences in the mean log total gross revenue across at least one of the show types.

Contrast	Estimate	SE	Lower Confidence Interval Limit	Upper Confidence Interval Limit	p-value
Musical - Play	0.562	0.0348	0.4805	0.644	<.0001
Musical - Special	0.223	0.0847	0.0241	0.422	0.0234
Play - Special	-0.339	0.0819	-0.5314	-0.147	0.0001

Table 4. Contrast Table

To determine which show type has a difference mean log total gross revenue, we analyze the differences between each pair of show types with a post-hoc Tukey test for a family-wise confidence level of 95% for a family of three estimates. Table 4 shows the p-values associated with the differences in means for each pair with all of the corresponding p-values being less than 0.05.

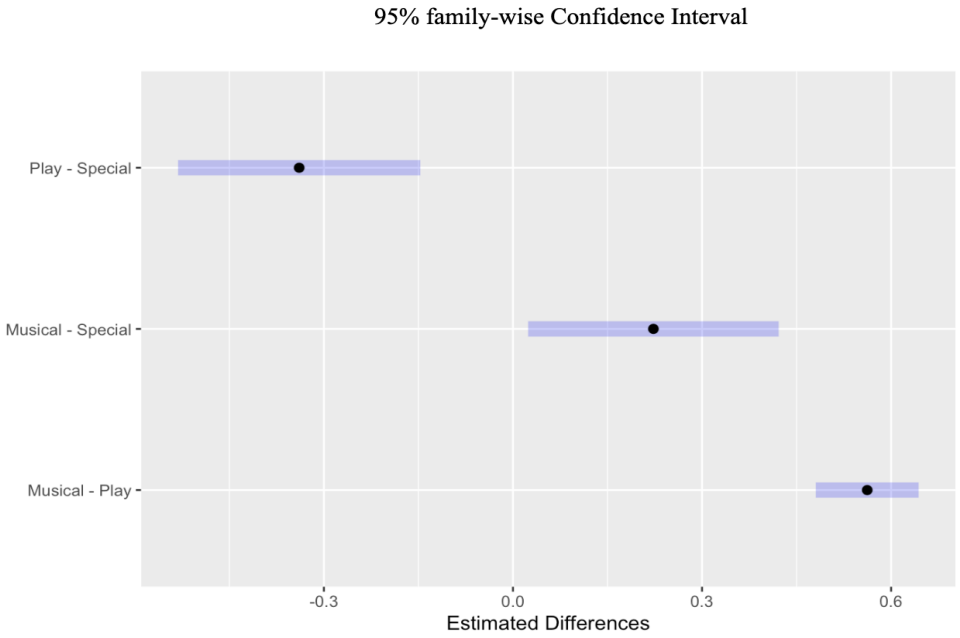

Figure 5. Family-Wise Confidence Intervals for Estimated Differences Between Show Types

The contrast table in Table 4 and the plot in Figure 5 show that none of the confidence intervals contains 0. Therefore, we conclude that 0 is not a reasonable value for any of the differences in means. Specifically, there is a significant difference in the log total gross income among all three show types. Musicals have the highest average total gross income, followed by special shows and then plays.

4. Conclusion

Our analysis aims to understand the factors contributing to the gross total revenue of Broadway performances for Cityville Arts Theatre. Our findings show that the total gross revenue is associated with the type of show. Musicals have the highest total gross revenue on average, followed by special shows and then plays.

City Arts Theatre initially came to us to determine which factors predicted ticket sales to maximize profit. However, there is no available data on the cost of production, net profit for shows, or ticket sales. Hence, we can only draw inferences on the total gross revenue. Without the cost of production, the gross total revenue will continuously increase with the number of performances. Additionally, Cityville Arts Theatre has a say over the type of show but not the specific performance. Therefore, for this analysis to apply to Cityville, we aim to determine if the average total gross income differs across Broadway show types.

This analysis has four main limitations. Firstly, out of the 1428 observations in our sample, 750 were for musicals, 615 were for plays, and only 63 were for special shows. The special show category may be underrepresented in the population, which could affect the accuracy of our estimates and conclusions. Secondly, our target population, Cityville Arts Theatre customers, and our study population are different. Our findings are based on a dataset

that contains statistics on the total gross revenue for Broadway shows, which may not apply to smaller theatres such as Cityville Arts. Thirdly, our dataset is not current; It spans 25 years, from 1990 to 2016. More recent data may result in different conclusions. The fourth limitation is that the dataset lacks information on the cost of production. So while our findings show that musicals generate the highest total gross revenue, if they are the most expensive to produce, they may yield a lower profit than other show types.

For future research, obtaining more recent data, data for smaller theatres, or a wider range of locations may be valuable and produce more accurate and comparable results. As we previously mentioned, given Cityville Arts Theatre's objective of drawing inferences on profit from ticket sales, collecting data on these variables would be valuable in achieving this goal. Additionally, future researchers might consider shortening the time span. While our findings suggest that musicals have the highest total gross revenue based over 25 years, analyzing a subset of the data, for example, from 2006 to 2016, may produce different results. Furthermore, as we note in our method section, we omit the observations that contained 0 for the number of performances but had positive values for total gross income. These values suggest that shows are making money without any performances. This data prompts an interesting question: Do shows make money through other means, such as sponsorship or promotion? Future analyses could include such factors to better understand the total profit.