Lights, Camera, Regression:
Variable Effects on the Gross Weekly Sales for Broadway Shows

J. Burke, E. Courtemanche, M. Johnson
ST 625-1
Spring 2025

Lights, Camera, Regression: Variable Effects on the Gross Weekly Sales for Broadway Shows

J. Burke, E. Courtemanche, M. Johnson

Abstract

This analysis examined which factors influence weekly gross profit for Broadway shows. It focused on the effects of top ticket price, the presences of previews (pre-show performances), and the show itself. It specifically looked at the data for shows *The Lion King*, *The Phantom of the Opera*, and *Wicked*. We applied a multiple linear regression model using publicly available data from Kaggle that was sourced from the company Playbill from 1985 to 2020. Our analysis revealed that higher top ticket prices were associated with greater profits, suggesting a benefit to premium pricing strategies. *The Lion King* contributed more to profit than the other shows, and having previews were found to negatively impact overall sales.

Introduction

Broadway, the iconic professional theatre district in New York City, has put on entertaining performances for audiences for years. From the emotionally provoking plots, to the elaborate costumes and makeup, to captivating song and dance numbers, Broadway has turned performances into an experience that attendees are thrilled to attend. But behind the scenes, Broadway is also a big money-maker for producers and investors alike; last week alone, its theaters made a gross profit of almost $47 million (BroadwayWorld, n.d.). Given the high costs of staging a production and running a physical theater, it is crucial that decision-makers understand what drives ticket sales in order to maximize revenue and maintain investors' interest.

For our analysis, we examine three key factors that may influence weekly gross profit: top ticket price, whether the show included preview performances, and the specific production itself. We focus on three popular Broadway shows—*The Lion King*, *The Phantom of the Opera*, and *Wicked*—using historical data from 1985 to 2020, originally sourced from Playbill. A multiple linear regression model is used to evaluate the impact of these variables on weekly earnings.

The report begins with a description of the dataset and methodology, followed by a discussion of our findings and their implications. We conclude with limitations of the study and recommendations for future research.

Data Characteristics

Our original dataset came from the public platform Kaggle and was sourced from Playbill (Mostipak, 2022). The data includes reports about all Broadway shows ending each week from 1985 to 2020. To clean our dataset, we removed data columns that were not relevant to our research question (i.e., number of seats sold, which theater the performance was being shown in, etc.). We also created a subset of our data to only focus on our shows of interest (*The Lion King*, *The Phantom of the Opera*, and *Wicked*) and converted the numerical variable of show previews into a binary variable of "Yes" or "No" based on whether the number of previews listed was

greater than zero ("Yes") or equal to zero ("No"). This left us with N = 3,698 rows of show data to work with.

Here is a summary of our dataset variables:

- show: The name of the show (*The Lion King*, *The Phantom of the Opera*, or *Wicked*).
- weekly_gross: The weekly box office gross for the individual show.
- top_ticket_price: The highest price of the tickets sold.
- previews: "Yes" if the show had a preview performance, "No" if it did not.

The Lion King had an average weekly gross of $1.44M (SD: $511K), *The Phantom of the Opera* averaged $746K (SD: $206K), and *Wicked* led with $1.60M (SD: $350K). All distributions were roughly symmetric with high-end outliers. On average, *Wicked* earned the most per week, followed by *The Lion King* and then *Phantom* (see Appendix A-B).

Top ticket prices were right-skewed for *The Lion King* (mean: $184, SD: $68) and *Phantom* (mean: $158, SD: $68), but left-skewed for *Wicked* (mean: $270, SD: $37), which had low-end outliers at $100. *Wicked* had the highest top ticket prices on average, then *The Lion King*, then *Phantom* (see Appendix C-D).

Shows with previews had slightly left-skewed weekly grosses (mean: $605K, SD: $205K), while without previews were slightly right-skewed (mean: $1.16M, SD: $525K) with a few upper outliers. On average, shows without previews grossed more weekly (see Appendix E-F).

Top ticket prices for shows with previews were slightly right-skewed (mean: $86, SD: $13), while those without previews were also right-skewed (mean: $197, SD: $73) with no outliers. Shows without previews had higher top ticket prices on average (see Appendix G-H).

For all shows, weekly gross profit and top ticket price appear to have a moderate, positive linear relationship. It is noteworthy that there is some clustering and vertical variability in this plot around different top ticket prices, which we attributed to Broadway tickets tending to be sold in round values ending in zeroes or fives, as well as other factors affecting gross profit that we did not directly examine (see Appendix I).

Model and Interpretation

The following multiple linear regression model was used to fit the data:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, where

Y = weekly gross in $

X_1 = top ticket price in $

X_2 = 1 if show = *The Phantom of the Opera*, 0 otherwise

X_3 = 1 if show = *Wicked*, 0 otherwise

X_4 = 1 if previews = "Yes", 0 otherwise

ε = error term

The following fitted multiple linear regression model was obtained for the data with top ticket price, show, and previews as the predictors and weekly gross as the response:

$\widehat{Y} = 936604 + 2730.3X_1 - 584498.4X_2 - 77365X_3 - 421859.2X_4$, where

$\widehat{Y}$ = model-predicted weekly gross in $

X_1 = top ticket price in $

X_2 = 1 if show = *The Phantom of the Opera*, 0 otherwise

X_3 = 1 if show = *Wicked*, 0 otherwise

X_4 = 1 if previews = "Yes", 0 otherwise

Obtaining a multiple R-squared value of 0.571, we found that about 57.1% of the variability in weekly gross profit is explained by the top ticket price, the show, if the show had a preview that week, and the model. After adjusting for the number of predictors, about 57.05% of the variability is still explained by these predictors and the model. The model's RMSE is $339,700, meaning that our model's predictions are typically off by about $340K with bigger values and outliers having a bigger impact on this number.

The overall model appears to have significant utility at the 0.05 significance level for predicting weekly gross from top ticket price, show, and previews ($p < 2.2e-16$). In other words, there is convincing evidence that at least one of our predictors (top ticket price, show, and/or previews) is affecting our response (weekly gross).

The model's y-intercept is $936,604.00, so the gross weekly profit we would expect a theater to make independent of any other variables is $937K. However, interpreting the y-intercept would not be as useful, since it is not possible to make positive weekly gross profit if the top ticket price (and therefore all ticket prices) is $0.00. This model estimates that when the top ticket price of a show increases by $1, on average, the weekly gross profit of that show would increase by $2.7K, regardless of which show it is and if it had a preview. These values were both found to be significant ($p < 2e-16$) at the 0.05 significance level.

It was also found that, when top ticket price is held constant and regardless of previews, *The Phantom of the Opera* generates $584K less in weekly gross than if *The Lion King* were showing while *Wicked* generates $77K less than if *The Lion King* were showing. These values are significant at the 0.05 level ($p < 2e-16$ and $p = 1.23e-05$, respectively).

Lastly, it was found to be significant at the 0.05 significance level ($p = 2.29e-04$) that having a preview performance generates $422K less weekly gross profit on average than shows that do not have a preview showing when holding top ticket price and the show itself constant (see Appendix M).

When applying our multiple linear regression model, we investigated the following model assumptions:

1. Linearity: The relationship between top ticket price and weekly gross appears to be linear. The correlation coefficient is 0.579, indicating a moderately strong, positive linear relationship between square footage and price. As previously stated, we did observe some clustering and vertical variability around different top ticket prices, but we attributed this to the even pricing strategy (ending in zeroes or fives) and other factors we did not

directly examine. Because of this, we treated this model assumption as satisfied (see Appendix I-J).

2. Independence, Mean and Constant Variance of Error Term: A residuals vs fitted plot of weekly grosses appears to be mostly centered around zero and shows no discernable pattern. Therefore, these model assumptions are satisfied (see Appendix K).
3. Normality of Error Term: A Q-Q Plot of the residuals appears to follow a relatively linear pattern. It is noteworthy that our tails have a slight skewness, indicating potential (but not overwhelming) violation, so further exploration may be needed in future studies (see Appendix L).

Summary—Greater Impacts

It should not be surprising that as the top ticket price increases, so do weekly gross sales. The more expensive a ticket is for a show, the higher the show's sales will be, even if it is filling fewer seats than a show with cheaper tickets. What is worth looking into, however, is *why* ticket prices are so high, as well as attempting to understand current trends in ticket pricing. Ticket prices are high because "Putting on a show is costly... 'You […] can't make it without a director, or a costume team, or a set designer, and so on'" (Pratt, 2017). Therefore, it makes sense for theaters to price their tickets higher in order to break even and return a higher weekly gross. However, it is also important to note that "By pricing tickets so high, thousands of potential buyers have been eliminated from the market" (Pratt, 2017). If ticket prices continue to increase, more and more patrons will be alienated from the theater market. Because of this, once a theater breaks even on a show, it may be beneficial for the theater to lower its ticket prices or begin offering more severely tiered seating prices.

We also analyzed the impact of theater previews on gross weekly ticket sales. A live theater preview is when shows offer tickets at a very discounted price while the show is still in its workshop phase. This is to receive audience feedback and make changes that will be more appealing to regular audiences once the show opens and can be viewed by critics. Unfortunately, it seems that previews are now doing more harm than good, likely due to social media culture. While critics are not allowed to review shows until opening night, there is nothing stopping regular patrons from voicing their own opinions on media platforms. Because most shows are still in production during preview phases, the reviews tend to be negative. These negative reviews can doom a show before it even opens, deterring patrons from attending regular performances with regular ticket prices, ultimately leading to lower weekly gross sales.

Therefore, it is recommended that theaters do not offer previews. If previews are absolutely crucial to improve a show's potential reach, theaters should either vet the patrons attending, have them sign NDAs, or limit attendance to friends and families of the cast and crew.

Lastly, we looked at the importance of picking the right show to put on. We only considered musicals in our analysis, since these tend to be more popular than plays. We picked one modern popular musical (*Wicked*), one classic musical (*The Phantom of the Opera*), and one musical based on pre-existing popular IP (*The Lion King*). What we found was that both *Wicked* and *The Phantom of the Opera* yield lower weekly gross sales than *The Lion King*. This is likely due to *The Lion King* being Disney property, meaning it's well-known by all audiences, and it being a family-friendly show could bring in whole families instead of just individual viewers. Of the remaining two shows, *Wicked* had less distance between its weekly gross sales versus *The Lion King* at -$77K. This makes sense, as *Wicked* is a very well-known show and is incredibly

popular, allowing it to bring in consistently large audiences. *The Phantom of the Opera* comes third, likely due to classics being a bit outdated in this modern era and having been around for long enough that most semi-regular theater goers have already seen them. Taking all of this into account, it is recommended that theaters select shows based on pre-existing popular media. Other popular Disney musical properties include *Frozen, Descendents, Cinderella,* and *Newsies.* The next most lucrative option would be more modern popular musicals such as *Hamilton, Six, Dear Evan Hanson,* or *Hadestown.* The least lucrative option would be to show a classic musical such as *The Sound of Music, Cats, Chicago,* or *Cabaret.*

Summary—Model Limitations

As previously mentioned, we only looked at popular musicals. This means that our model does not consider any plays or non-popular musicals.

Certain measures of utility and assumptions were not as strong as we would have liked from this model. The model's multiple R-squared value is .571. This is a lower value than desired but is still passable considering our individual variables are statistically significant. While all our assumptions passed, the clustered nature of our Top Ticket Price vs. Weekly Gross graph makes it difficult to truly decipher linearity, even though individual regression lines display linearity.

There are many factors that play into weekly gross sales beyond the ones that we explored. These can include variables such as theater location, theater condition, time of year, theater capacity, and more.

The dataset we worked with does not take inflation into account for variables associated with ticket pricing. For example, long-running, classic shows like *The Phantom of the Opera* were also operating during a time when ticket prices were much lower than they are today. This could potentially affect the overall data for these types of shows, bringing down their weekly gross values.

Summary—Further Exploration & Concluding Remarks

The dataset we worked with was very robust with great potential for further exploration. One of the variables we originally wanted to investigate was the week number of the year in which a show ran in order to learn more about how seasonality affects weekly gross. We would've then grouped these weekly numbers together by season (fall, winter, spring, summer) to investigate if there was a significant difference in gross weekly sales according to time of year. There is also potential in further studies to utilize prediction and confidence intervals in R that allow one to isolate and compare the weekly gross of specific inputs. These can also be used to investigate specific questions theater owners may have.

By implementing a well-researched and innovative sales strategy, our goal is not only to boost immediate sales figures but also to establish a sustainable framework for Broadway's continued growth and success.

References

BroadwayWorld. (n.d.). *Broadway grosses*. BroadwayWorld. https://www.broadwayworld.com/grosses.cfm

Mostipak, J. (2022). *Broadway weekly grosses* [Dataset]. Kaggle. https://www.kaggle.com/datasets/jessemostipak/broadway-weekly-grosses

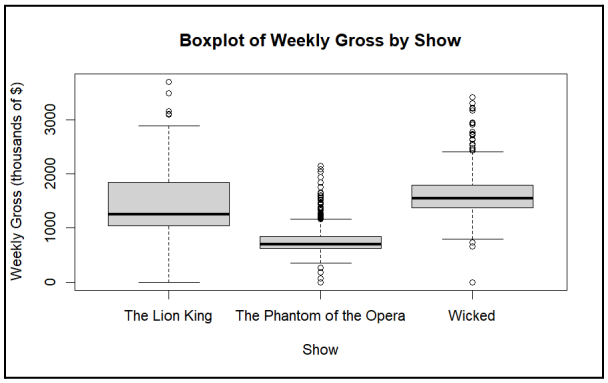
Pratt, B. (2017, December 7). *Unreasonably high ticket prices limit live theater to elite, wealthy viewers*. University Wire. ProQuest. http://ezp.bentley.edu/login?url=https://www.proquest.com/wire-feeds/unreasonably-high-ticket-prices-limit-live/docview/2116315033/se-2

Appendix

A. Data Summary for Weekly Gross by Show

Weekly Gross ($)	Lion King	Phantom of the Opera	Wicked
Mean	1,438,938	746,269	1,595,613
Median	1,254,044	696,080	1,540,868
Standard Deviation	510,630	206,071	350,186
Minimum	0	0	0
Maximum	3,696,974	2,146,126	3,411,819

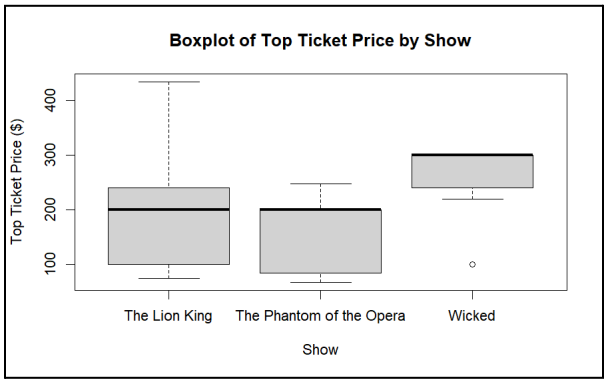
B. Boxplot of Weekly Gross by Show



C. Data Summary for Top Ticket Price by Show

Top Ticket Price ($)	Lion King	Phantom of the Opera	Wicked
Mean	184.37	158.33	270.25
Median	200.00	200.00	300.00
Standard Deviation	68.21	59.39	36.62
Minimum	75.00	67.50	100.00
Maximum	434.00	247.00	300.00

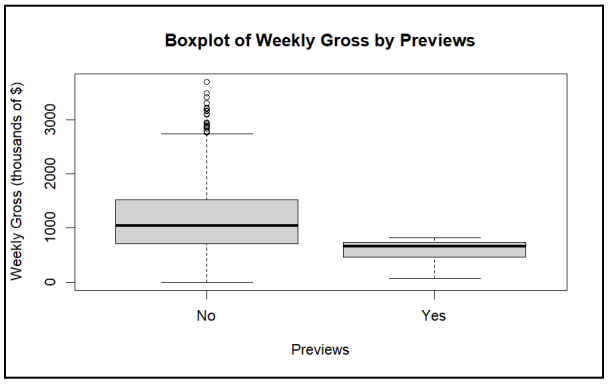
D. Boxplot of Top Ticket Price by Show



E. Data Summary for Weekly Gross by Preview

Weekly Gross ($)	Yes	No
Mean	604,779	1,163,007
Median	662,666	1,042,563
Standard Deviation	204,917	524,933
Minimum	71,291	0
Maximum	821,248	3,696,974

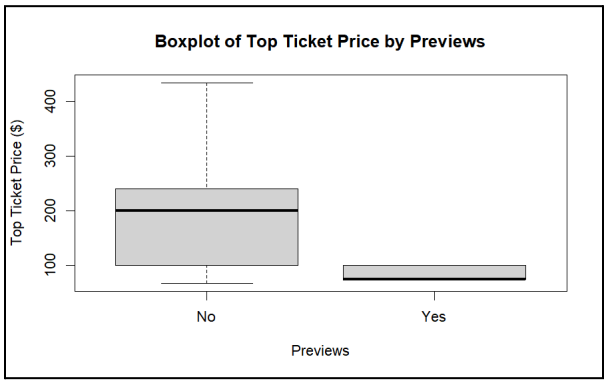
F. Boxplot of Weekly Gross by Previews



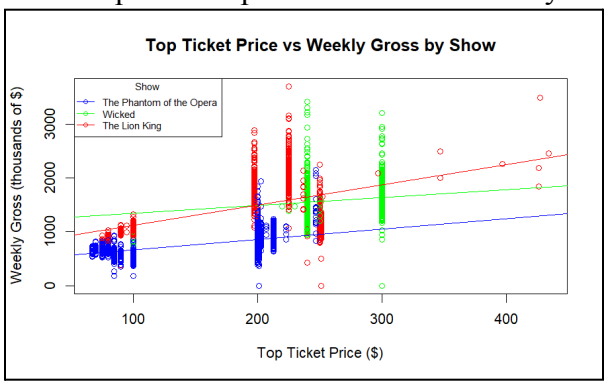
G. Data Summary for Top Ticket Price by Show

Top Ticket Price ($)	Yes	No
Mean	86.11	196.68
Median	75.00	200.00
Standard Deviation	13.18	73.23
Minimum	75.00	67.50
Maximum	100.00	434.00

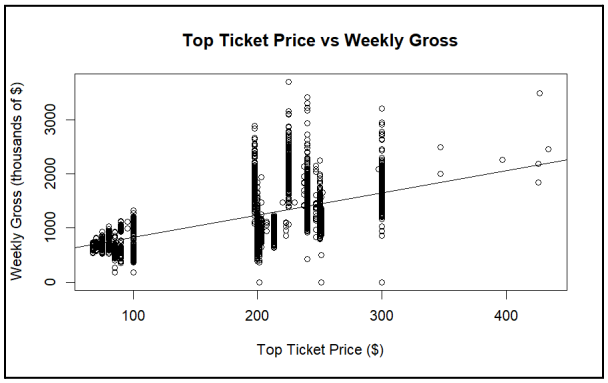
H. Boxplot of Top Ticket Price by Previews



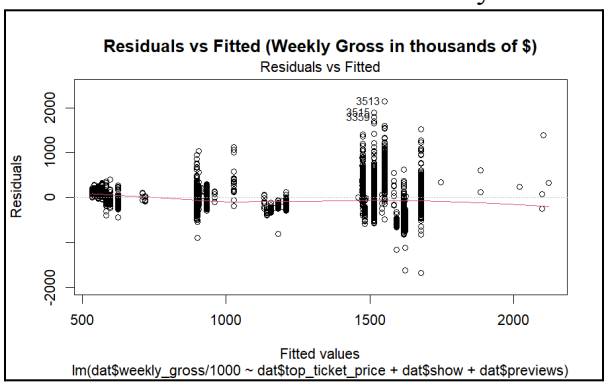
I. Scatterplot of Top Ticket Price vs Weekly Gross by Show



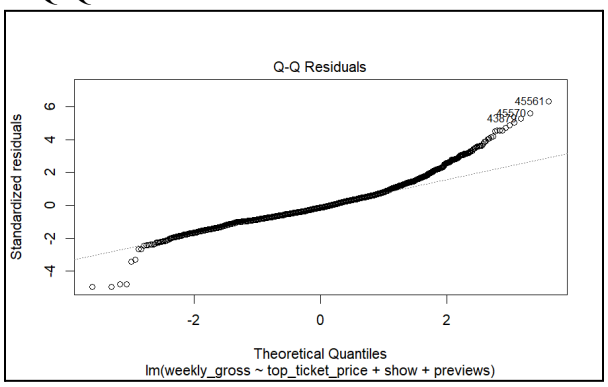
J. Scatterplot of Top Ticket Price vs Weekly Gross



K. Residuals vs Fitted Plot for Weekly Gross



L. Q-Q Plot of Residuals



M. R Output for MLR

```
Call:
lm(formula = weekly_gross ~ top_ticket_price + show + previews,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1678339  -215336   -44897   163115  2146045

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 936604.0    21455.6  43.653  < 2e-16 ***
top_ticket_price              2730.3      102.7  26.578  < 2e-16 ***
showThe Phantom of the Opera -584498.4    14009.5 -41.722  < 2e-16 ***
showWicked                    -77365.0    17667.5  -4.379 1.23e-05 ***
previewsYes                  -421859.2   114367.2  -3.689 0.000229 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 339700 on 3304 degrees of freedom
  (389 observations deleted due to missingness)
Multiple R-squared:  0.571,     Adjusted R-squared:  0.5705
F-statistic:  1099 on 4 and 3304 DF,  p-value: < 2.2e-16
```

Test 1: H_0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ and H_A: $at\ least\ one\ of\ the\ \beta_i's \neq 0$, where

β_0 = the coefficient variable for weekly gross

β_1 = the coefficient variable for top ticket price

β_2 = the difference in average weekly gross between *The Phantom of the Opera* and *The Lion King* when top ticket price and previews remain the same

β_3 = the difference in average weekly gross between *Wicked* and *The Lion King* when top ticket price and previews remain the same

β_4 = the difference in average weekly gross between having a preview and not having a preview when top ticket price and show remain the same

p-value < 2.2e-16

Test 2: H_0: $\beta_0 = 0$ and H_A: $\beta_0 \neq 0$

p-value < 2e-16

Test 3: H_0: $\beta_1 = 0$ and H_A: $\beta_1 \neq 0$

p-value < 2e-16

Test 4: H_0: $\beta_2 = 0$ and H_A: $\beta_2 \neq 0$

p-value < 2e-16

Test 5: H_0: $\beta_3 = 0$ and H_A: $\beta_3 \neq 0$

p-value = 1.23e-05

Test 6: H_0: $\beta_4 = 0$ and H_A: $\beta_4 \neq 0$

p-value = 2.29e-04

Code

```
# Data Setup
dat <- read.csv("grosses.csv")
dat <- subset(dat, show %in%
                    c("Wicked", "The Phantom of the Opera", "The Lion King"))
dat$previews <- ifelse(dat$previews > 0, "Yes", "No")
dat$theatre <- NULL
dat$seats_in_theatre <- NULL
dat$performances <- NULL
dat$weekly_gross_overall <- NULL
dat$potential_gross <- NULL
dat$pct_capacity <- NULL
dat$week_number <- NULL
dat$avg_ticket_price <- NULL
dat$seats_sold <- NULL

# Exploratory Analysis
tapply(dat$weekly_gross, dat$show, mean)
tapply(dat$weekly_gross, dat$show, sd)
tapply(dat$weekly_gross, dat$show, fivenum)
boxplot(dat$weekly_gross/1000 ~ dat$show, xlab = "Show",
        ylab = "Weekly Gross (thousands of $)",
        main = "Boxplot of Weekly Gross by Show")

tapply(dat$top_ticket_price, dat$show, mean, na.rm = TRUE)
tapply(dat$top_ticket_price, dat$show, sd, na.rm = TRUE)
tapply(dat$top_ticket_price, dat$show, fivenum)
boxplot(dat$top_ticket_price ~ dat$show, xlab = "Show",
        ylab = "Top Ticket Price ($)",
        main = "Boxplot of Top Ticket Price by Show")

plot(dat$top_ticket_price, dat$weekly_gross/1000, xlab = "Top Ticket Price ($)",
     ylab = "Weekly Gross (thousands of $)",
     main = "Top Ticket Price vs Weekly Gross")
abline(lm(dat$weekly_gross/1000 ~ dat$top_ticket_price))

plot(dat$top_ticket_price, dat$weekly_gross/1000,
     col = ifelse(dat$show == "The Phantom of the Opera", "blue",
     ifelse(dat$show == "Wicked", "green", "red")),
     xlab = "Top Ticket Price ($)",
     ylab = "Weekly Gross (thousands of $)",
     main = "Top Ticket Price vs Weekly Gross by Show")
abline(lm(dat$weekly_gross/1000 ~ dat$top_ticket_price,
          subset = dat$show == "The Phantom of the Opera"), col = "blue")
abline(lm(dat$weekly_gross/1000 ~ dat$top_ticket_price,
```

```
        subset = dat$show == "Wicked"), col = "green")
abline(lm(dat$weekly_gross/1000 ~ dat$top_ticket_price,
        subset = dat$show == "The Lion King"), col = "red")
legend("topleft",
     legend = c("The Phantom of the Opera", "Wicked", "The Lion King"),
      col = c("blue", "green", "red"),
      lty = 1,
      pch = 1,
      title = "Show", cex = 0.7)

tapply(dat$weekly_gross, dat$previews, mean)
tapply(dat$weekly_gross, dat$previews, sd)
tapply(dat$weekly_gross, dat$previews, fivenum)
boxplot(dat$weekly_gross/1000 ~ dat$previews, xlab = "Previews",
       ylab = "Weekly Gross (thousands of $)",
       main = "Boxplot of Weekly Gross by Previews")

tapply(dat$top_ticket_price, dat$previews, mean, na.rm = TRUE)
tapply(dat$top_ticket_price, dat$previews, sd, na.rm = TRUE)
tapply(dat$top_ticket_price, dat$previews, fivenum)
boxplot(dat$top_ticket_price ~ dat$previews, xlab = "Previews",
       ylab = "Top Ticket Price ($)",
       main = "Boxplot of Top Ticket Price by Previews")

# Analysis and Assumptions
mod <- lm(weekly_gross ~ top_ticket_price + show + previews, data=dat)
summary(mod)

cor(dat$weekly_gross, dat$top_ticket_price, use = "complete.obs")

mod2 <- lm(dat$weekly_gross/1000 ~
          dat$top_ticket_price + dat$show + dat$previews)
plot(mod2, 1, main = "Residuals vs Fitted (Weekly Gross in thousands of $)")
plot(mod, 2)
```

Statement of Group Contribution

Group Member	Project Contribution
Jessi Burke	33% — Responsible for background research, identifying project goals, and defining variables.
Erin Courtemanche	33% — Responsible for model interpretation, model limitations, and recommendations for future studies.
Megan Johnson	33% — Responsible for coding, exploratory analysis, and assessing model utility and model assumptions.
All	Contributed to creating and editing the report and slides, ensuring quality and completeness, and actively participated in group meetings by sharing ideas.

Electronic Signatures

Jessica Burke
Erin Courtemanche
Megan Johnson