Understanding Broadway Gross: Key Factors and Growth Strategies*

Positive Effects of Pricing, Seats, Performances, and Holidays, with Limited Impact During Tony Award Months

Xuanel Zhou

November 26, 2024

This study examines the key factors influencing Broadway's weekly gross revenue using datasets from Playbill and The Broadway League. The result shows that ticket pricing, seating capacity, performance frequency, and holiday weeks positively impact gross, while Tony Award months show limited effects. Despite steady long-term growth, disruptions such as the 2016 snowstorm and COVID-19 highlight the industry's vulnerability to external shocks. These findings provide actionable strategies, including holiday-focused marketing, and enhanced engagement during award seasons, to sustain growth and optimize Broadway's financial performance.

Table of contents

*Code and data are available at:https://github.com/Isazhou13/Broadway_Gross

1 Introduction

Broadway is a cultural and economic landmark in New York City, contributing significantly to the city's global reputation and economy. For stakeholders, understanding the drivers of Broadway revenue is essential, as it directly informs strategic decisions on pricing, scheduling, and resource allocation to maximize profitability and sustain operations. This study aims to identify and quantify the key factors influencing Broadway's weekly gross revenue, using datasets provided by Playbill and The Broadway League.

The estimand in this paper represents the weekly gross revenue of Broadway shows. Specifically, it examines how variables such as ticket price, theater seating capacity, number of performances, whether it is a holiday week, whether it falls in a Tony Award month, and the year influence weekly gross revenue. While the true estimand is unknown, it is approximated using the available data and model assumptions. Defining the estimand ensures the model and methodology align with the research objective and facilitates addressing potential biases from data limitations or methodological choices.

The analysis, conducted using a multiple linear regression model, reveals that ticket pricing, seating capacity, performance frequency, and holiday weeks positively impact Broadway's weekly gross revenue. Conversely, the limited impact of Tony Award months highlights opportunities for growth. While steady long-term growth is observed, disruptions such as the 2016 snowstorm and COVID-19 emphasize Broadway's vulnerability to external shocks.

These findings emphasize both the opportunities and challenges Broadway faces in maximizing its revenue. Identifying the factors that drive revenue and highlighting periods of underperformance provide a foundation for adapting to evolving market conditions. For instance, additional potential can be realized through strategies such as holiday packages, collaborations with tourism boards, and enhanced on-site experiences. Addressing the low impact of Tony Award months could involve exclusive ticket packages and increased media engagement to leverage the event's visibility. In an era of rapid digitalization and shifting consumer preferences, Broadway must innovate and adopt targeted strategies to sustain growth and maintain its position as a global cultural icon.

The remainder of this paper is structured as follows: Section 2 provides an overview of the dataset and visualizations of the variables. Section 3 details the modeling approach, including predictor selection and the regression model structure. Section 4 presents the primary findings and outcomes of the model. Finally, Section 5 explores the key findings, highlights the main takeaways, addresses the study's weaknesses, and outlines potential next steps.

2 Data

2.1 Overview

This study utilizes R packages (R Core Team 2023) for data cleaning and analysis, incorporating libraries from tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), knitr (Xie 2024), arrow (Richardson et al. 2024), here (Müller 2020) and scales (Wickham, Pedersen, and Seidel 2023). The **tidyverse** package is employed for data manipulation, such as filtering, grouping, and reshaping datasets. **ggplot2** is used for data visualization, creating a range of plots to explore and present trends in the data. **knitr** facilitates the integration of analysis and reporting by generating dynamic reports in R Markdown. **arrow** is used for efficient handling of large datasets, particularly for reading and writing data in Parquet format to optimize storage and processing. The **here** package ensures reproducibility by standardizing file paths in the project directory, making it easier to locate and load data files. Finally, **scales** is employed to enhance the readability of plots by formatting axes, labels, and scales, ensuring clarity in the presentation of numeric and categorical data.

The dataset, sourced from **Playbill**, reflects Broadway weekly grosses as reported by theatres affiliated with The Broadway League. The data cleaning process involved grouping observations and removing missing values. The final dataset consists of 14,519 observations across 10 variables: `week_ending`, `week_number`, `weekly_gross`, `avg_ticket_price`, `seats_in_threatre`, `performances`, `year`, `month`, `holiday_week` and `Tony_Award`.

2.2 Measurement and Considerations

The dataset for this study originates from Playbill, which collects information about Broadway shows through direct reporting and partnerships with industry organizations. A significant portion of the data, particularly box office grosses and performance metrics, is provided by The Broadway League, a national trade association for the Broadway industry.

The transformation of real-world phenomena into the dataset begins with data collection. The Internet Broadway Database (IBDB), managed by The Broadway League, compiles comprehensive records of Broadway productions from sources such as theater programs, which detail cast and crew information. Additional data is supplemented by newspaper and magazine reports, theatrical textbooks, interviews with theater professionals, and The League's archival records. This information is then processed into key metrics, including ticket sales, attendance, revenue, and theater utilization, which are published weekly. Before release, The League's research department verifies the data's accuracy and removes personal information to ensure privacy.

However, converting real-world phenomena into dataset entries inevitably results in the loss of important contextual details. For example, shows offering premium tickets or standing-room sales may disproportionately inflate weekly grosses compared to productions without

these options. While the dataset includes a top ticket price metric, it does not specify the number of tickets sold at this price, leaving gaps in understanding revenue dynamics. Similarly, discounts and promotions, such as rush tickets or group sales, can lower the average ticket price, potentially underestimating the perceived value of tickets sold. Additionally, while the dataset provides weekly metrics, it lacks details about whether performances occurred on weekdays or weekends, further limiting insights into revenue patterns. These limitations underscore the challenges of translating complex industry data into simplified metrics for analysis.

2.3 Outcome variables

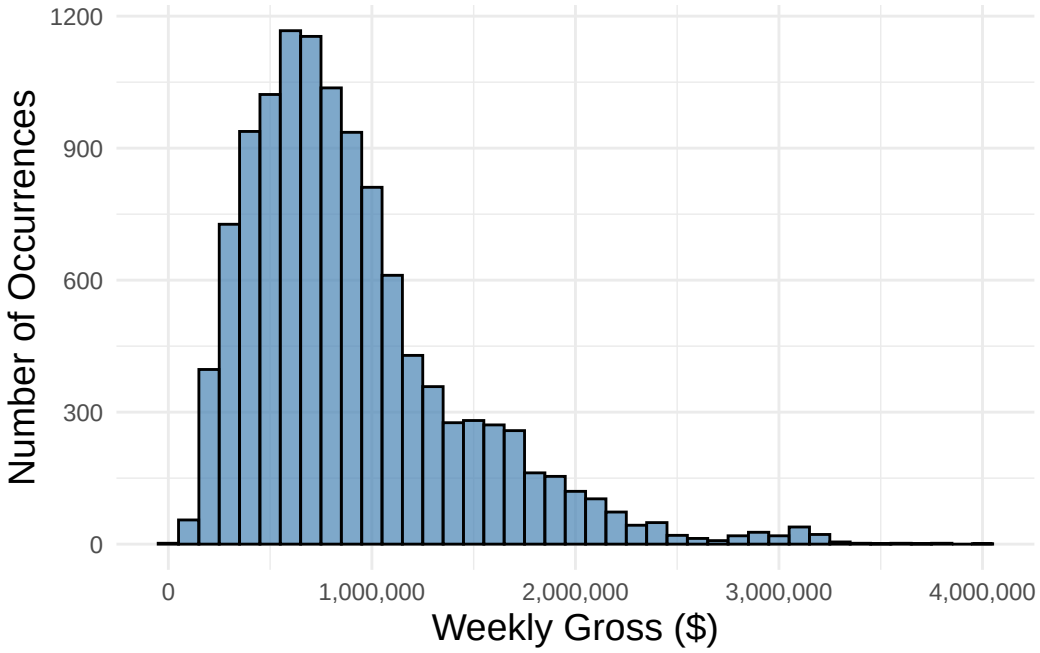

Figure 1: Distribution of Broadway weekly gross.

The outcome variable of interest for this study is the weekly gross revenue for Broadway productions. The distribution, shown in Figure 1, exhibits a right-skewed pattern. Most weeks have revenues ranging between $500,000 and $2,000,000, with a declining frequency as revenue increases. The mode of the distribution appears around $2,000,000. A small number of weeks report revenues exceeding $3,000,000, which may represent potential outliers or rare events that warrant further investigation. These high-revenue weeks could stem from unusual circumstances, such as special promotions, holiday seasons, or significant market shifts.

2.4 Predictor variables

2.4.1 Year

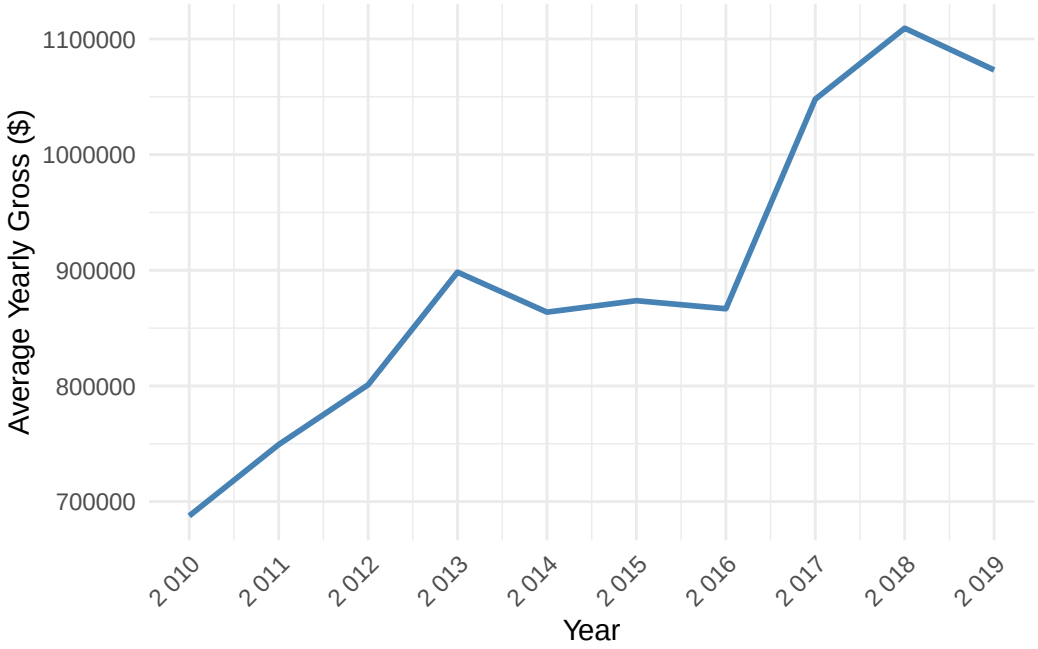

Figure 2: Yearly Gross Revenue Trends for Broadway Productions, spanning from 2010 to 2019.

Figure 2 captures the Broadway industry's growth over the decade. The yearly averages were calculated by averaging the weekly gross revenue across all Broadway shows for each year. The chart shows a steady increase in average yearly gross revenue, reflecting overall industry growth. From 2013 to 2015, revenues remain relatively stable, followed by sharp growth from 2016 onward, peaking in 2019. A slight dip in 2020 likely reflects the impact of the COVID-19 pandemic on Broadway operations.

Figure 3 provides detailed information on weekly gross revenues, showing that most values clusters predominantly in the range of $500,000 to $2,000,000, which aligns with the histogram results shown in Figure 1. However, there is a gradual upward trend in the maximum weekly gross over the years, indicating that the Broadway industry experienced growth during this period.

Earlier years (2010–2012) show a compact distribution of weekly gross. From 2015 onward, the range widens, with both higher peaks and broader variation between the lowest and highest values, reflecting increased performance variability among Broadway shows. Notably, some

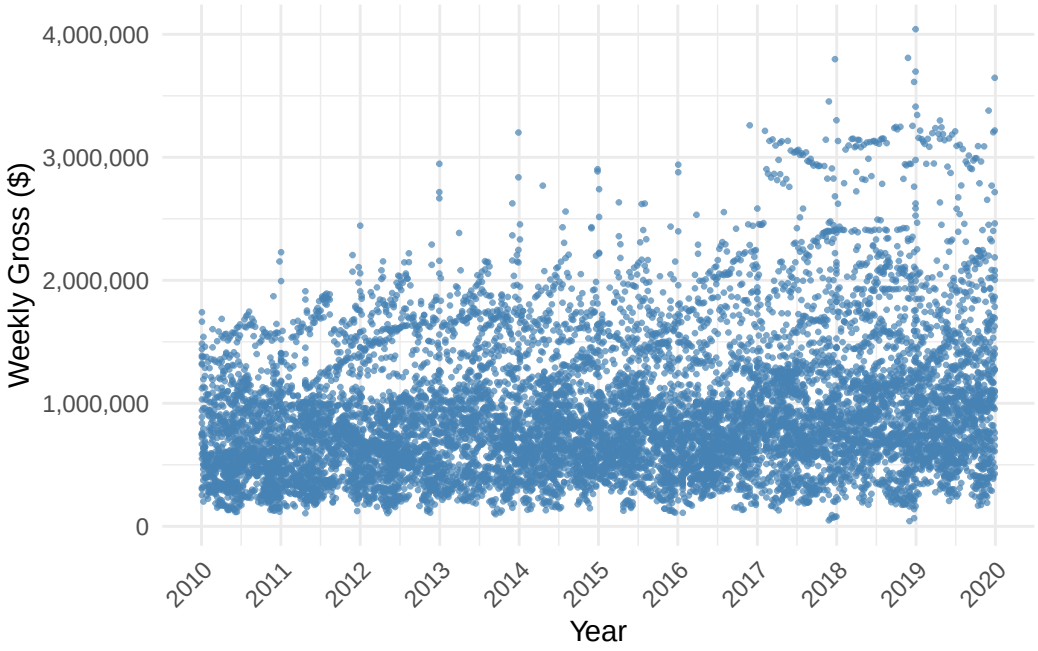

Figure 3: Weekly grossfor each Broadway show over time, spanning from 2010 to 2019. Each point represents the gross for a particular week for a specific show.

shows at some weeks after 2015 exceed $3,000,000, indicating more outliers and growing disparities between top-performing shows and others, likely influenced by audience preferences, pricing, and industry growth. These peaks are more commonly observed at the end of the year, suggesting a seasonal effect, likely tied to holiday demand and special events.

2.4.2 Average price of tickets sold

Figure 4 shows the majority of average ticket prices are concentrated around $100, with a peak frequency between $80 and $120. The right skewed distribution, indicating that while most ticket prices are relatively affordable, there are instances of higher ticket prices extending beyond $300, with a few exceeding $500.

2.4.3 Theatre seat capacity

Figure 5 illustrates the distribution of theatre sizes on Broadway, measured by the number of seats. The most common theatre size is around 1,000 seats, with a significantly higher frequency compared to other sizes.

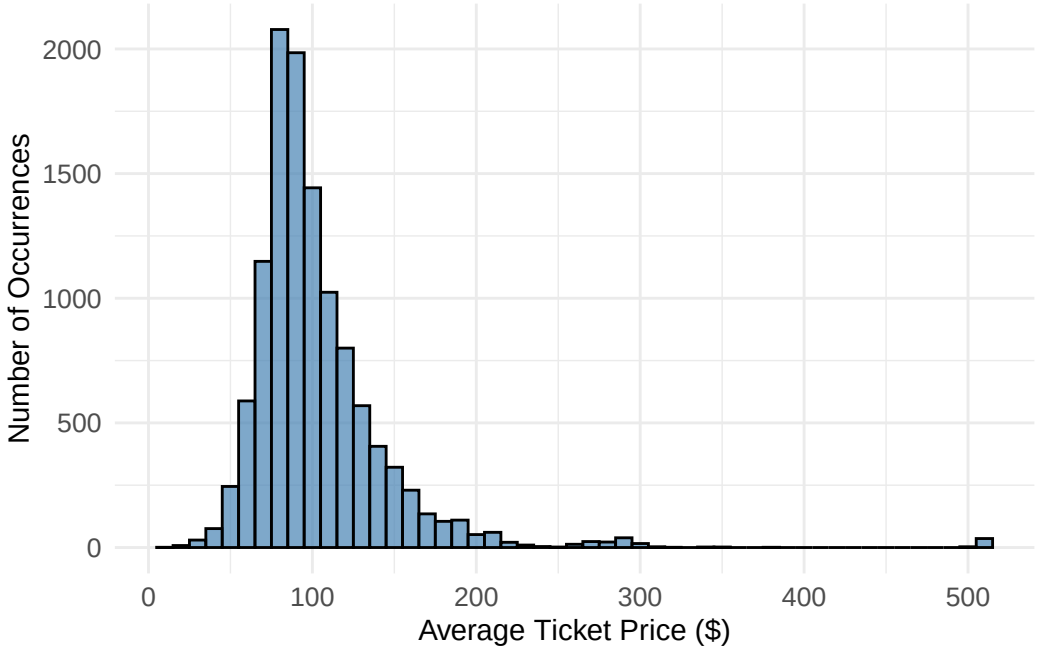

Figure 4: Distribution of Average Broadway Ticket Prices

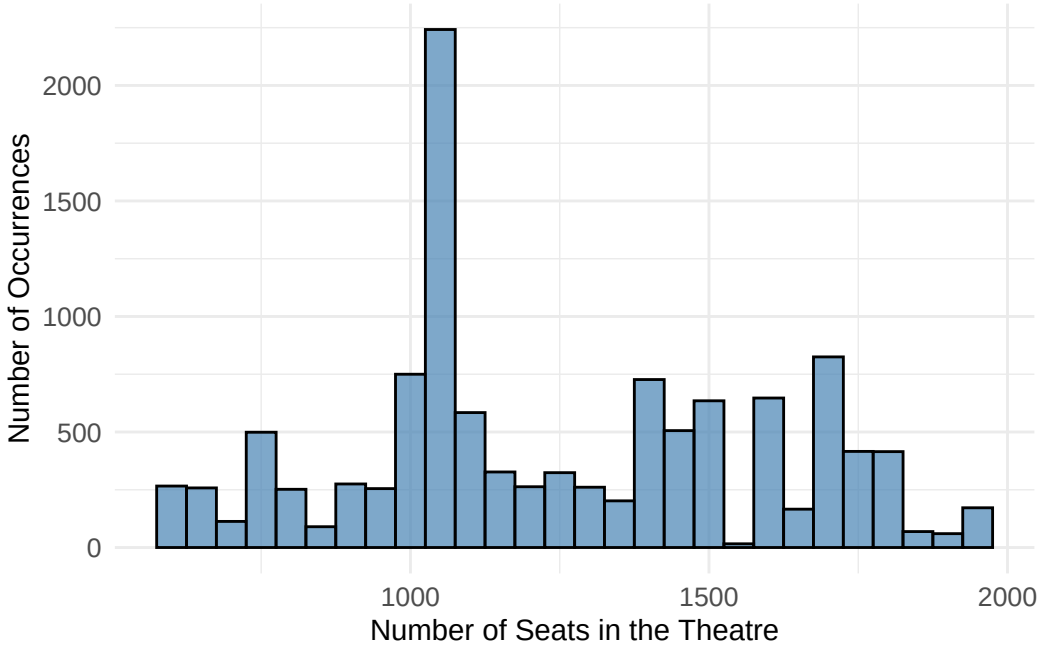

Figure 5: Distribution of Theatre Sizes Based on Number of Seats

2.4.4 Number of performances in the week

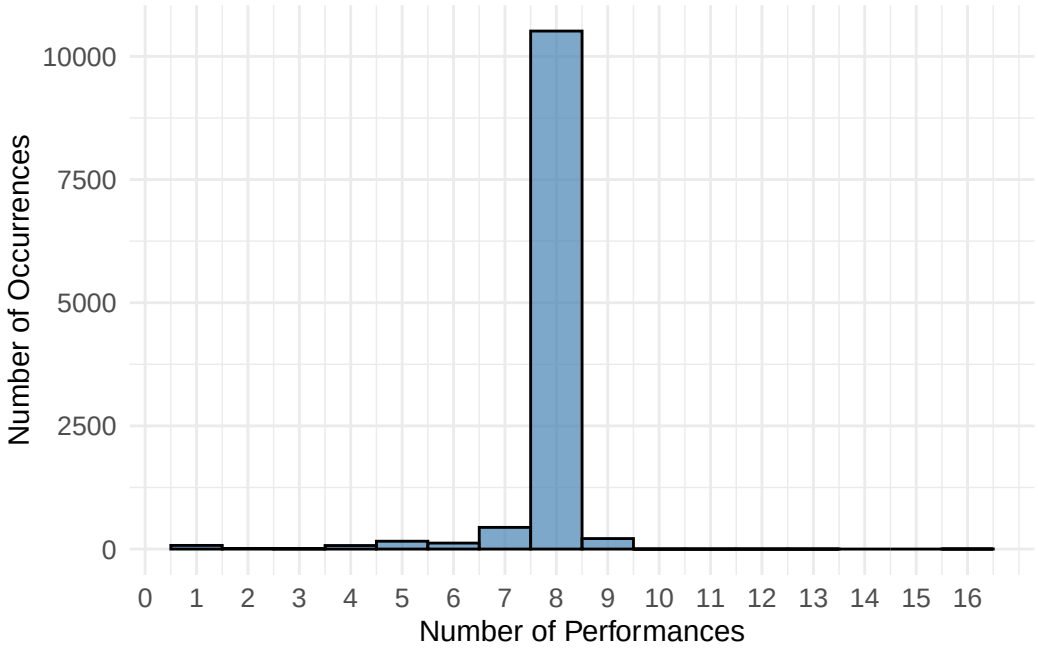

Figure 6: Distribution of Performances

Figure 6 highlights the consistency in Broadway's performance schedules, with the vast majority of shows performing 8 times per week, reaching a frequency of over 10,000 occurrences. A small number of shows have fewer or occasionally more than 8 performances per week, likely due to special events, holidays, or production adjustments.

2.4.5 Holiday Week

Holiday weeks were identified based on the major holidays of the year and include the following: New Year's Week: The week containing January 1st.

- Independence Day Week: The week of July 4th.
- Labor Day Week: The week containing Labor Day, which falls on the first Monday of September.
- Thanksgiving Week: The week containing the fourth Thursday of November.
- Christmas Week: The week containing December 25th.

Non-holiday weeks are defined as all other weeks that do not overlap with these predefined holiday periods.

Figure 7: Comparison of Average Weekly Gross Between Holiday and Non-Holiday Weeks

The weekly gross revenue was aggregated by grouping the data into holiday and non-holiday weeks. The average weekly gross revenue for each category was then calculated by summing all weekly gross revenues within the category and dividing by the total number of weeks in that category.

As shown in Figure 7, the average weekly gross revenue during holiday weeks is slightly higher than during non-holiday weeks, though the difference is relatively small. This suggests that holiday weeks may attract slightly larger audiences or higher ticket sales, due to increased leisure time and tourism. However, non-holiday weeks still maintain a comparable level of average weekly gross, demonstrating consistent revenue throughout the year.

2.4.6 Tony Award Period

The Tony Awards, formally the Antoinette Perry Award for Excellence in Broadway Theatre, are the highest honors in Broadway, typically held in June to mark the end of the Broadway season. Due to COVID-19, the 2020 awards were postponed to September.

The variable Tony Award Month refers to the month of the awards, which often sees heightened media attention and anticipation. Off-Award Months include all other months, serving as a baseline for comparison.

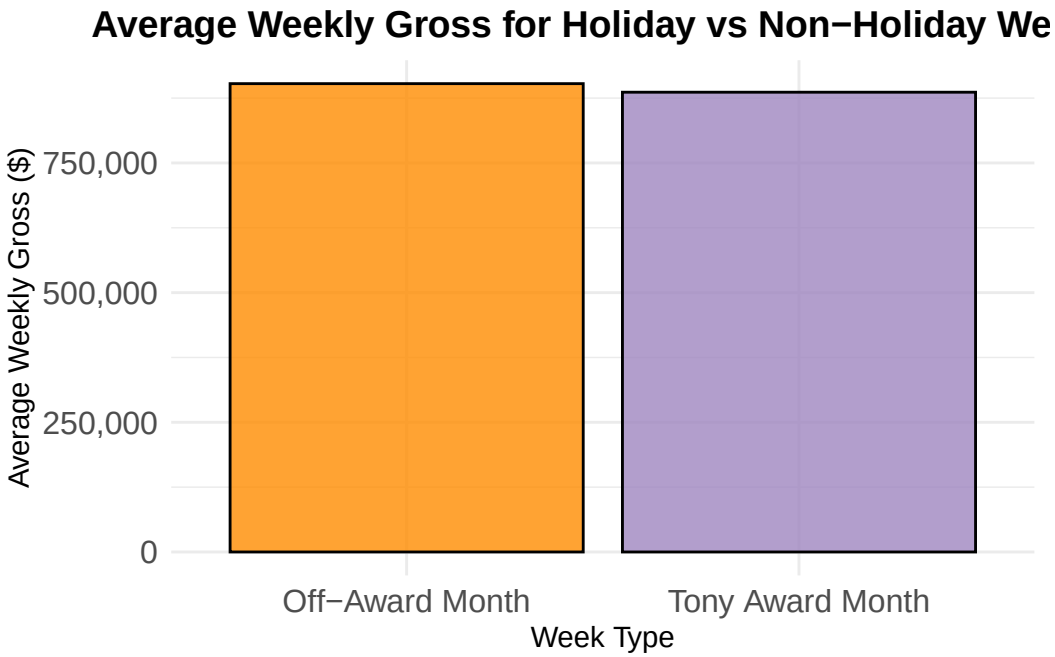

Figure 8: Comparison of Weekly Gross Revenue During Tony Award and Non-Award Months

Interestingly, Figure 8 shows that the average weekly gross revenue during off-award months is slightly higher than during the Tony Award month. This unexpected outcome may stem from factors like limited new production launches or audience focus shifting toward the awards themselves rather than attending shows.

3 Model

3.1 Model Selection

Since the objective of this study is to help Broadway strategically manage and make decisions to increase gross revenue, it is crucial to select a model that effectively explains how various factors influence revenue. To achieve this goal, several models were evaluated before selecting the final approach, including simple linear regression, decision trees, and multiple linear regression, each with its own strengths and weaknesses.

Simple Linear Regression is straightforward to implement, computationally efficient, and easy to interpret. It provides a clear understanding of the linear relationship between one independent variable and the dependent variable. However, the model is limited to analyzing the

effect of a single predictor at a time and cannot account for the influence of multiple interacting variables. This oversimplification can lead to omitted variable bias in the context of a multivariable dataset like this one.

Despite Decision Trees excel at capturing non-linear relationships and interactions between variables, they are prone to overfitting, especially in moderately sized datasets, leading to poor generalization on unseen data.

Based on these considerations, **Multiple Linear Regression** was chosen. While it has limitations, such as reliance on key assumptions and sensitivity to outliers, MLR strikes a balance between simplicity, interpretability, and the ability to analyze the combined effects of multiple predictors. Its coefficients provide a clear quantification of the impact of each independent variable on weekly gross revenue, making it well-suited for understanding the underlying relationships in the data.

To ensure robust evaluation and avoid overfitting, the cleaned dataset was divided into training (70%) and testing (30%) subsets. The training set was used to fit the model, while the testing set was reserved for evaluating the model's performance on unseen data.

3.2 Model Set-up

Key Assumptions:

- Linearity: The relationship between the dependent variable and each predictor is linear.
- Independence: Observations are independent of one another.
- Homoscedasticity: Variance of residuals is constant across all levels of the independent variables.
- Normality of Residuals: The residuals are normally distributed.
- No Multicollinearity: Independent variables are not highly correlated with each other.

Background details and diagnostics are included in **?@sec-model-details**.

The model in this study is designed to predict weekly gross revenue using the following predictors:

- Average Ticket Price (`avg_ticket_price`): The average price of tickets sold during the week.
- Seats in the Theater (`seats_sold`): The seating capacity of the theater.
- Number of Performances (`performances`): The total number of performances held during the week.
- Holiday Week Indicator (`holiday_week`): Whether the show was performed during a holiday week (1 = holiday week, 0 = otherwise).
- Tony Award Month Indicator (`Tony_Award`) : Whether the show was performed during the Tony Award month (1 = Tony Award month, 0 = otherwise).
- Year (`year`): The year in which the show was performed, ranging from 2010 to 2019.

The model takes the form:

$$\begin{aligned}
\text{weekly_gross}_i &= \beta_0 + \beta_1 \cdot \text{avg_ticket_price}_i + \beta_2 \cdot \text{seats_in_theatre}_i && (1)\\
&\quad + \beta_3 \cdot \text{performances}_i + \beta_4 \cdot \text{holiday_week}_i && (2)\\
&\quad + \beta_5 \cdot \text{Tony_Award}_i + \beta_6 \cdot \text{year}_i \epsilon_i && (3)\\
\epsilon_i &\sim \text{Normal}(0, \sigma^2) && (4)
\end{aligned}$$

Where:

$$\begin{aligned}
\beta_0 &\text{ is the intercept term} && (5)\\
\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 &\text{ are the coefficients for each predictor} && (6)\\
\sigma^2 &\text{ is the variance of the error term} && (7)
\end{aligned}$$

We run the model in R (R Core Team 2023) using the `stats` package from base R. The lm() function from this package was used to fit the multiple linear regression model.

3.3 Interpretation of Coefficients

- Intercept (β_0): Represents the predicted value of the weekly gross revenue when all predictors are 0. Since it is unrealistic for all predictors the intercept serves as a theoretical starting point for the model.
- Average Ticket Price (β_1): Measures the change in weekly gross revenue for each one-unit increase in average ticket price, assuming other predictors remain constant.
- Seats in the Theater (β_2):Represents the additional revenue associated with a one-unit increase in the seating capacity of the theater, holding other factors constant.
- Number of Performances (β_3): Indicates the increase in weekly gross revenue for each additional performance in the week, controlling for other variables.
- Holiday Week (β_4): Quantifies the difference in weekly gross revenue between holiday weeks and non-holiday weeks, all else being equal.
- Tony Award Month (β_5): Captures the difference in weekly gross revenue between weeks during the Tony Award month and other months, assuming other predictors remain unchanged.
- Year (β_6): Reflects the annual trend, representing the average change in weekly gross revenue for each one-year increase, controlling for other variables.

3.4 Model justification

We expect a negative relationship between average ticket price and weekly gross revenue, as Broadway tickets are a form of entertainment and not a necessity. As ticket prices increase, demand is likely to decrease significantly because consumers may prioritize essential expenses over discretionary spending, especially for shows that are not well-known or during off-peak periods.

The seating capacity of the theater and the number of performances are also expected to have a positive relationship with weekly gross revenue, as larger venues or more performances each week can accommodate more audience members, resulting in higher total revenue.

For holiday weeks, we predict a positive effect on weekly gross revenue, as holidays often bring increased audience demand due to leisure time and tourism. Conversely, the effect of Tony Award months is less certain, as the difference in average weekly gross revenue during Tony Award months compared to other months is minimal.

Finally, the year variable is included to account for long-term trends, with an expected positive effect reflecting the steady growth of Broadway's gross revenue over the years.

4 Results

The results of the multiple linear regression model is summarized in the Table 1. The result matches with some of our expectations while deviating in other areas.

As anticipated, average ticket price, seating in yheater, and number of performances exhibit significant positive relationships with weekly gross revenue. Specifically, increases in ticket prices and theater size lead to higher gross revenue, consistent with the notion that larger venues and premium ticketing contribute to revenue growth. Similarly, the significant positive impact of the number of performances aligns with our expectation that additional shows increase overall revenue by accommodating more audiences.

The findings for holiday weeks also meet initial expectations, showing a positive relationship with weekly gross revenue. This result suggests that increased demand during holiday periods effectively boosts ticket sales, contributing to higher overall revenue.

However, the results for Tony Award months diverge from our initial expectations. Contrary to the assumption that Tony Award months would boost revenue due to increased demand or publicity, the coefficients for these variables are negative, indicating a slight decrease in weekly gross revenue during these periods.

Lastly, the year variable exhibits a small but positive effect, consistent with the expectation of a gradual upward trend in Broadway revenue over the years. This reflects the broader growth of the industry, potentially due to factors such as inflation. Overall, the model confirms some of

our hypotheses while highlighting areas where the data does not fully align with expectations, prompting further discussion and exploration in the next section.

Table 1: Regression Results for Weekly Gross. This table displays the coefficients, standard errors, and t-statistics for the multiple linear regression model used to analyze the factors influencing weekly gross for Broadway.

	Coefficient	Standard Error	t-Statistic
(Intercept)	-15932708.16	1451587.35	-10.98
avg_ticket_price	9168.19	48.14	190.45
seats_in_theatre	699.82	5.87	119.23
performances	70688.48	2457.02	28.77
holiday_week	16683.82	5863.93	2.85
Tony_Award	-7140.66	6937.47	-1.03
year	7178.16	721.16	9.95

5 Discussion

5.1 Key Findings

This paper presents a multiple linear regression analysis to examine the factors influencing weekly gross on Broadway. Key predictors, including average ticket price, seating capacity, and number of performances, were found to significantly impact revenue, while holiday weeks and Tony Award months showed unexpected or non-significant effects. These findings highlight the complexity of audience behavior and the importance of operational decisions in optimizing revenue.

5.2 Balancing Pricing, Seating Capacity, and Performance Frequency

Operational strategies such as ticket pricing, seating capacity, and performance frequency emerged as key drivers of weekly gross revenue, offering actionable opportunities for Broadway to increase earnings. The strong impact of average ticket price underscores the importance of implementing demand-based pricing strategies, particularly for popular or limited-run productions, where higher ticket prices can significantly boost revenue with minimal impact on attendance. For less popular shows, optimizing affordability could help maintain audience retention while still contributing to overall revenue.

Similarly, maximizing theater utilization through strategic scheduling and venue allocation is crucial for increasing revenue. Larger venues and additional performances can expand audience capacity, but Broadway managers must evaluate whether demand can sufficiently meet

the increased supply. Additionally, managers must balance these strategies with logistical considerations such as performer availability, production costs, and preserving the quality of the audience experience. By carefully aligning pricing strategies with capacity optimization, Broadway can enhance revenue while ensuring sustainable long-term growth.

5.3 Opportunities in Holiday Weeks and Tony Award Months

The model results indicate that while holiday weeks align with the expectation of increased revenue due to higher audience availability, the effects of Tony Award months challenge the assumption that they inherently boost revenue through increased media attention.

During holiday weeks, families often travel or engage in special outings. New York City, home to Broadway, becomes a global attraction with iconic events like the Rockefeller Center Christmas Tree Lighting, holiday markets, and the New Year's Eve Ball Drop in Times Square. This festive atmosphere draws millions of tourists, creating a significant opportunity for Broadway to boost gross revenue and position itself as a must-visit holiday experience.

To maximize this potential and leverage its prime location, Broadway could adopt targeted strategies to make attending shows more convenient and appealing for holiday visitors. For example, collaborating with hotels, travel agencies, and tourism boards to create holiday packages that include show tickets could firmly establish Broadway as an essential part of the holiday itinerary. Additionally, enhancing the on-site experience with seasonal concessions and festive elements could further help Broadway attract more visitors during this lucrative season.

For Tony Award months, the data indicates that Broadway gross remains unaffected, challenging the belief that the awards inherently drive revenue through increased media attention. While the Tony Awards generate significant publicity, the focus tends to be on a limited number of nominated or winning shows, leaving many other shows relatively unnoticed. This concentrated attention may not translate into broader attendance increases across the industry. To address this, Broadway could collaborate with media outlets to feature live broadcasts, behind-the-scenes content, and interviews with nominees to create excitement around both the awards and the shows themselves. Offering exclusive ticket packages for Tony-nominated shows during the awards season could further incentivize attendance and increase revenue.

5.4 Sustaining Broadway's Growth in a Rapidly Changing World

The significant positive relationship between year and weekly gross revenue reflects the steady growth of Broadway over the past decade, driven by factors such as rising ticket prices, increased tourism, and the industry's ability to attract global audiences. However, this growth trajectory also have some exceptional cases, and not valid for all time period. For instance, the sharp decline in 2016 weekly gross during winter was influenced by a severe snowstorm that forced theaters to close, disrupting public transportation and preventing both audiences and

performers from attending. Similarly, the slight downturn in 2019 coincided with the onset of the COVID-19 pandemic, which had a profound impact on Broadway operations, highlighting the industry's vulnerability to external shocks.

Even with a generally positive and sustainable long-term growth trend, Broadway must adapt and innovate to keep pace with the rapid changes in today's society. The rise of streaming platforms and the growing popularity of short-form video content present new challenges to live theater, demanding fresh approaches to attract and engage modern audiences. By embracing digital innovations while emphasizing the unique, immersive experience of live performances, Broadway can maintain its growth and continue thriving as a global cultural hub.

5.5 Weaknesses

While the paper provides valuable insights into revenue drivers, it primarily focuses on direct factors such as ticket price, seating capacity, and performance frequency. The dataset lacks critical information such as marketing efforts, show popularity, repeat attendance rates, or audience demographics. As a result, the analysis misses deeper insights into other potential variables that could offer a more comprehensive understanding of revenue dynamics.

The model assumes linear relationships between predictors and weekly gross revenue. However, factors such as ticket price or performance frequency may demonstrate diminishing returns or threshold effects that are not reflected in the analysis. Furthermore, complex interactions between variables are not explored. For instance, changes in ticket price could influence consumers' willingness to buy, potentially altering demand and revenue in ways a linear model cannot fully capture.

Additionally, the model assumes uniform effects of predictors across all Broadway shows, which may oversimplify the analysis. Different types of productions (e.g., musicals, plays, blockbusters, or niche shows) likely have distinct revenue dynamics that are not adequately addressed by a single model.

Moreover, the model does not account for external factors such as economic conditions, competitor pricing, or shifts in tourism trends, all of which could significantly influence Broadway revenue. These omissions limit the model's ability to fully explain variations in weekly gross revenue and highlight areas for future improvement.

Another weakness of this study is its exclusive focus on analyzing gross revenue without considering the associated costs of producing and running Broadway shows. While understanding revenue drivers is crucial, gross revenue alone does not provide a complete picture of financial performance. Factors such as production expenses, marketing costs, theater rental fees, and operational overhead significantly impact profitability but are not accounted for in this analysis. Ignoring costs limits the ability to assess the net financial health of Broadway productions and could lead to incomplete or misleading conclusions about the success of certain strategies or shows.

5.6 Next Steps

Incorporating additional variables, such as marketing expenditures, customer demographics, and customer experiences, could provide a deeper understanding of the factors influencing revenue. These information could be gathered in the future through detailed surveys aimed at collecting data on audience preferences, spending habits, and satisfaction levels. With this information, Broadway could develop more targeted marketing strategies tailored to different audience groups, allowing for customized promotions, pricing models, and programming that better address the unique needs of each segment, ultimately maximizing engagement and revenue.

Future analyses could also employ more advanced models to address the limitations of linear assumptions. For example, exploring non-linear relationships and interaction effects between variables, such as the interplay between ticket prices and theater size or the varying impact of holiday weeks on revenue, could reveal important dynamics.

Additionally, conducting separate analyses for different types of productions, such as traditional shows versus new or experimental productions, would help identify the unique revenue drivers for each category. This segmentation would enable more precise strategy recommendations tailored to the specific needs and performance characteristics of different types of Broadway shows. Also, combining these advanced analyses with cost considerations would offer a holistic approach to optimizing Broadway's economic sustainability.

A Appendix: Idealized Methodology and Survey

A.1 Objective and Overview

The purpose of this appendix is to outline an idealized methodology and survey design to supplement the current analysis of Broadway's weekly gross revenue. By incorporating additional survey data and addressing gaps in the existing dataset, this methodology aims to enhance the understanding of revenue drivers, audience behaviors, and market dynamics. The survey is designed to capture demographic, behavioral, and experiential data that complement the operational metrics already analyzed in the study.

A.2 Core Objectives

- To gather audience demographics (e.g., age, income, location) to understand their influence on ticket purchases and attendance patterns. This information can be integrated into the model to enhance its predictive accuracy.
- To identify purchasing behaviors, such as repeat attendance, pricing sensitivity, and group sales, providing a foundation for refining marketing and pricing strategies.
- To assess audience motivations and satisfaction, focusing on factors driving show selection and overall theater experience, offering guidance on how to improve Broadway's organizational and operational strategies.

A.3 Sampling Strategy

Our sampling strategy utilizes **random sampling** to ensure unbiased representation of Broadway audiences. This approach randomly selects participants from the overall population of Broadway ticket buyers and attendees, minimizing potential biases and ensuring the sample reflects the diversity of the audience.

Random sampling provides a robust foundation for generalizable analysis, capturing a wide range of perspectives and behaviors without targeting specific subgroups. By employing this method, the study aims to produce reliable and comprehensive insights into the factors influencing Broadway's attendance and revenue dynamics.

A minimum sample size of 1,000 responses per quarter is proposed to ensure statistical robustness across strata and account for seasonality.

A.4 Recruitment Strategy

Participants will be recruited through a combination of:

- **In-Theater Surveys**: Surveys distributed at participating Broadway theaters, targeting audiences immediately after the performance for real-time feedback.
- **Online Platforms**: Email invitations to ticket buyers, links embedded in ticket purchase confirmations, and social media outreach via official Broadway channels.
- **Tourism Partnerships**: Collaborations with NYC tourism boards, hotels, and travel agencies to engage out-of-town visitors.
- **Incentives**: Offering discounts on future ticket purchases or exclusive merchandise as incentives for survey participation.

A.5 Data Validation and Quality Assurance

- **Data Triangulation**: Responses will be cross-referenced with transactional data from ticket sales to validate self-reported behaviors.
- **Pretesting**: Pilot surveys will be conducted to test question clarity, response rates, and overall survey design effectiveness.
- **Anonymity and Privacy**: Personal data will be anonymized to encourage honest responses and comply with privacy regulations.

A.6 Survey Design Considerations

The survey is designed to be comprehensive but concise, ensuring high response rates while capturing actionable data. Key design considerations include:

- **Question Types**: A mix of multiple-choice, Likert scale, and open-ended questions to balance structured data collection with qualitative insights.
- **Clarity**: Simple, jargon-free language to ensure accessibility for diverse audiences.

Focus Areas:

- Demographics: Age, income, location, gender, and education level.
- Attendance Behavior: Frequency of attendance, motivations, and purchasing habits.
- Experience Metrics: Satisfaction levels, likelihood of future attendance, and feedback on show quality.
- Pricing Sensitivity: Perceptions of ticket affordability and willingness to pay for premium experiences.

A.6.1 Broadway Audience Survey Form

Introduction:

Thank you for taking the time to participate in this survey. Your feedback will help us understand Broadway audiences better and improve the overall experience.

Please Note:

- All responses will be kept strictly confidential.
- Your participation is entirely voluntary.
- We kindly request that you answer all questions honestly and to the best of your knowledge.
- The survey is estimated to take approximately 10 minutes to complete.
- If you have any inquiries or concerns regarding this survey, please don't hesitate to contact the research team at **isabella.zhou@mail.utoronto.ca**.

Your contribution to this study is greatly appreciated! Each participant will be entered into a lottery with a chance to win a **$100 gift card redeemable for any Broadway show of your choice**.

Section 1: About You

What is your age ?

- Under 18
- 18–24
- 25–34
- 35–49
- 50–64
- 65 or older

What sex were you assigned at birth, on your original birth certificate?

- Female
- Male

How do you currently describe yourself (mark all that apply)?

- Female
- Male
- Transgender
- I use a different term [free-text]

Where do you currently live?

- Manhattan

- Brooklyn
- Queens
- Bronx
- Staten Island
- Other New York City Borough: _____
- Outside NYC (please specify): _____

What is your annual household income (before taxes)?

- Less than $25,000
- $25,000–$49,999
- $50,000–$74,999
- $75,000–$99,999
- $100,000–$149,999
- $150,000–$249,999
- $250,000 or more

How far in advance do you typically purchase Broadway tickets?

- Less than a week
- 1–2 weeks
- 3–4 weeks
- 1–2 months
- 3 months or more

Section 2: Broadway Attendance and Ticket Purchases

How many Broadway shows have you attended in the past 12 months?

- None
- 1–2
- 3–5
- 6–10
- 11 or more

Have you seen this show before?

- Yes
- No

What motivates you to attend Broadway shows? (Check all that apply)

- Specific performers
- Storyline or theme of the show
- Award recognition (e.g., Tony Awards)
- Recommendations from friends or family

- Discounts or promotions
- Other (please specify): _____

How much did you pay for your most recent Broadway ticket (including fees)?

- Less than $99
- $100–$149
- $150–$199
- $200–$249
- $250–$299
- $300–$349
- $350 or more

How do you usually purchase tickets?

- Online via website or app
- At the theater box office
- Through group sales or subscriptions
- Other (please specify): _____

Section 3: Experience and Preferences

How satisfied were you with the last Broadway show you attended?

- Very satisfied
- Satisfied
- Neutral
- Dissatisfied
- Very dissatisfied

What factors most influenced your satisfaction? (Check all that apply)

- Quality of performance
- Cast or performers
- Theater amenities (e.g., seating, concessions)
- Ticket price relative to experience
- Other (please specify): _____

What would improve your Broadway experience? (Check all that apply)

- Lower ticket prices
- Easier ticket access
- More diverse productions
- Enhanced theater amenities
- Other (please specify): _____

When do you typically attend Broadway shows? (Check all that apply)

- Weekdays
- Weekends
- Holiday weeks (e.g., Thanksgiving, Christmas, New Year's Week, (please specify): _____
- Tony Award months (typically June)

Where do you typically hear about Broadway shows? (Check all that apply)

- Social Media (e.g., Instagram, Facebook, Twitter)
- Online ads or newsletters
- Word of mouth
- Print media (e.g., newspapers, magazines)
- Other (please specify): _____

Section 4: Future Engagement

How likely are you to attend another Broadway show in the next 6 months?

- Very likely
- Somewhat likely
- Neutral
- Somewhat unlikely
- Very unlikely

What factors might prevent you from attending more Broadway shows?

- High ticket prices
- Lack of time
- Lack of interest in available productions
- Other (please specify): _____

What additional suggestions do you have to improve the Broadway experience?

End Message:

Thank you for participating! Your feedback will play a vital role in enhancing the Broadway experience for future audiences. If you have any questions about this survey or would like to be informed about the results, please contact **isabella.zhou@mail.utoronto.ca**

B Appendix: Additional Broadway Surveys and Data

Besides the dataset used for this study, Broadway also conducts several key surveys to gather comprehensive data on various aspects of Broadway theatre. However, the information and datasets collected through these surveys are not publicly accessible for analysis. If this data were made available, merging it with the current dataset could provide a more robust model for gross analysis. These surveys include **The Demographics of the Broadway Audience**, which is designed to provide insights into the composition and behavior of Broadway theatregoers.

- **Sampling Method**: The survey is typically conducted annually, with performances selected quarterly to ensure a representative sample of Broadway's diverse offerings. This approach captures seasonal variations in the audience and provides a balanced dataset
- **Recruitment Method**:
  - In-Person Distribution: Questionnaires are handed out to audience members during selected performances.
  - Online Access: Audiences can complete the survey online via a QR code provided at performances or through a link sent post-event.
  - Wi-Fi Login Prompt: Patrons connecting to a theatre's Wi-Fi are also invited to participate in the survey. These varied recruitment strategies aim to ensure a high response rate and broad representation of attendees.

Broadway also records and analyzes additional information, such as Broadway's Economic Contribution to New York City, the Audience for Touring Broadway, and the Economic Impact of Touring Broadway. These surveys and studies help capture broader insights into Broadway's influence and reach, contributing to a more comprehensive understanding of its role in the cultural and economic landscape.

C Appendix: Model details

C.1 Diagnostics

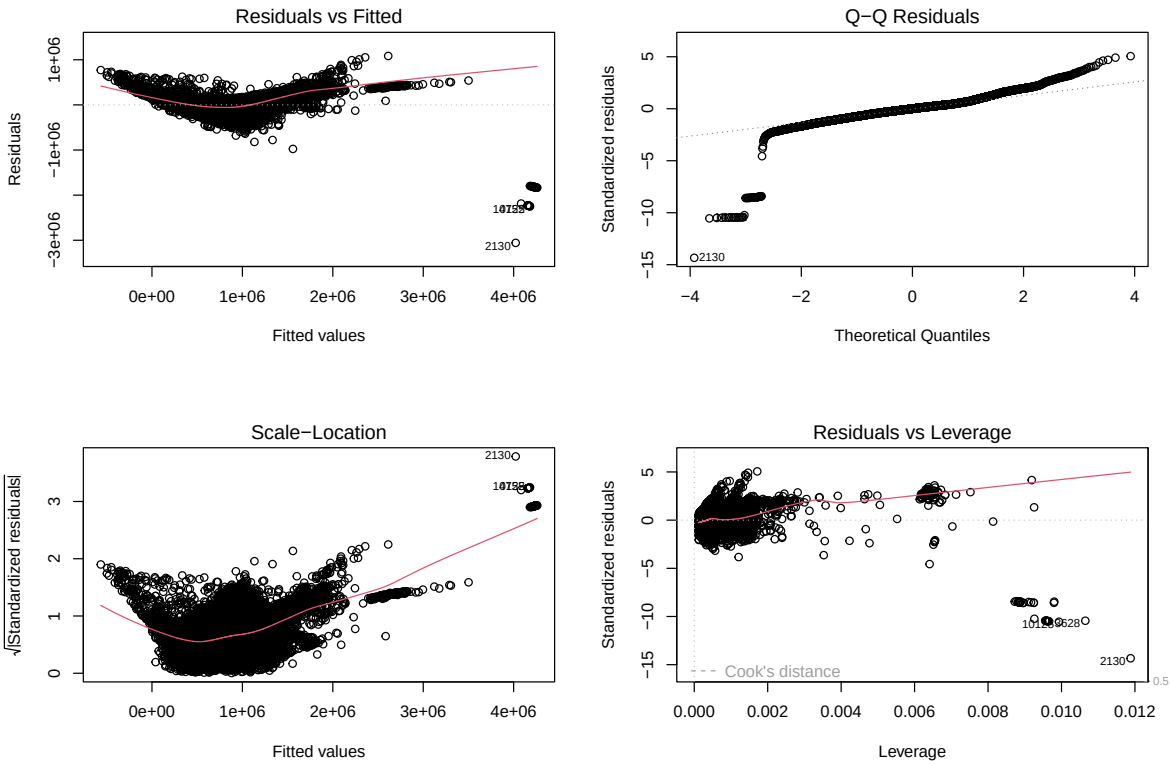

Figure 9: Diagnostic Plots for Regression Model Residuals. Residuals vs Fitted values, Q-Q Plot of residuals, Scale-Location plot, and Residuals vs Leverage, used to assess the assumptions of the multiple linear regression model.

As shown in Figure 9, these diagnostic plots indicate potential violations of the linear regression assumptions, suggesting the need for model refinements or transformations.

The curved pattern in the Residuals vs Fitted plot points to possible non-linearity, while the tails in the Q-Q plot highlight deviations from normality. Additionally, the trend observed in the Scale-Location plot suggests heteroscedasticity. Furthermore, the Residuals vs Leverage plot reveals points with high leverage and large residuals, indicating potential outliers that may disproportionately influence the model.

C.2 Mean Squared Error and Mean Absolute Error on Test Data

Table 2: Calculate Mean Squared Error (MSE) on Test Data

Table 2: Model Performance Metrics

Metric	Value
Mean Squared Error (MSE)	50,906,242,649
Mean Absolute Error (MAE)	142,488
R-squared	0.83

From Table 2, MSE measures the average squared prediction error. MAE represents the average prediction error magnitude. R-squared indicates that the model explains 83% of the variance in weekly gross revenue.

C.3 Multicollinearity Check on the Training Data

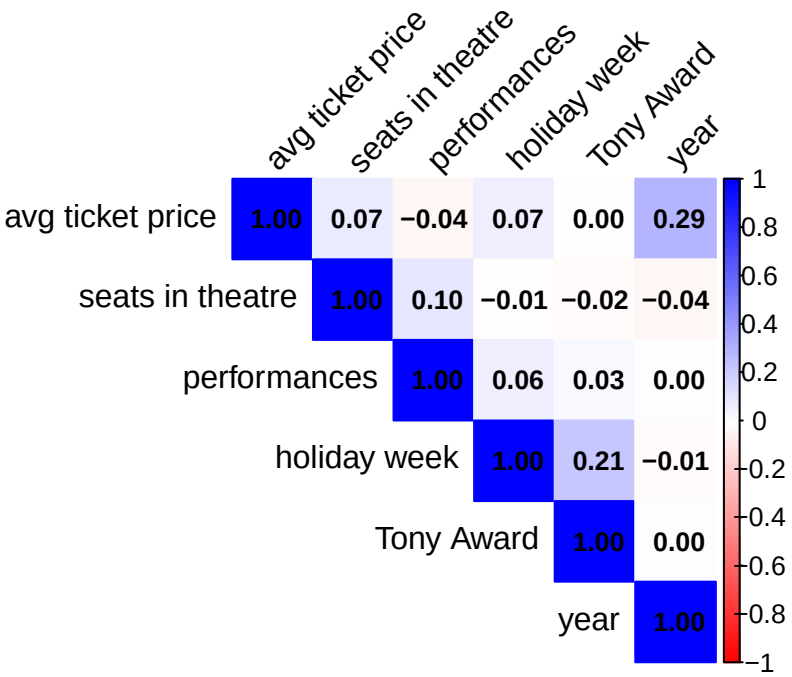

Figure 10: Correlation Matrix between variables

In Figure 10, a correlation check is performed. The matrix confirms the assumption of low multicollinearity among predictors, supporting the reliability of coefficient estimates in the regression model.

D Additional Visulization

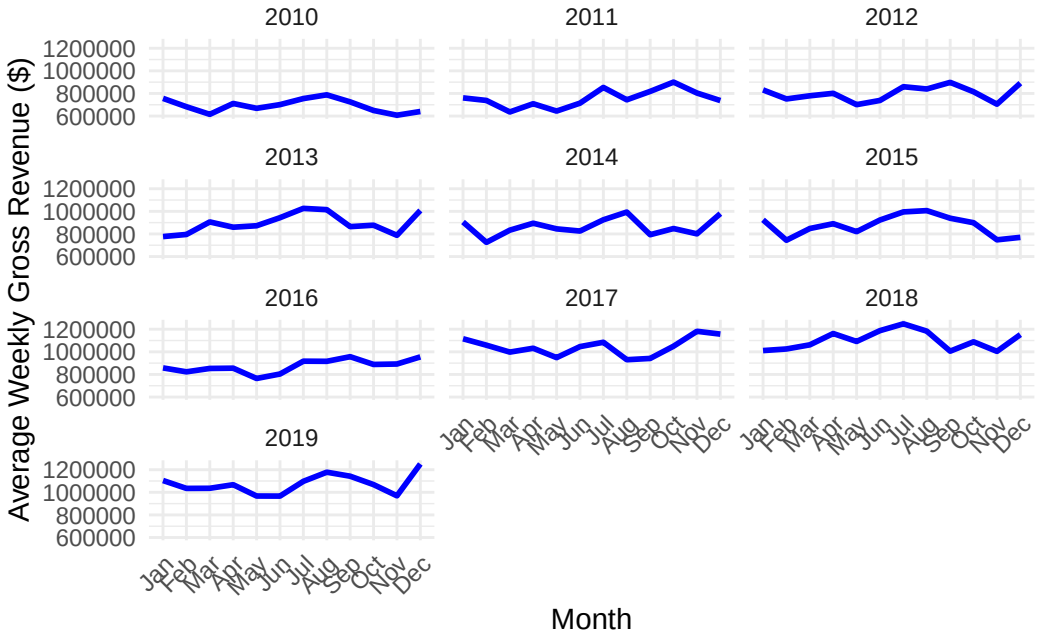

Figure 11: Average Weekly Gross Revenue Trends Across Months, spanning from 2010 to 2019.

E Acknowledgements

Thanks to Open AI and ChatGPT 4o is used to write the paper.

This study utilizes R packages (R Core Team 2023) for data cleaning and analysis, incorporating libraries from tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), knitr (Xie 2024), arrow (Richardson et al. 2024), here (Müller 2020) and scales (Wickham, Pedersen, and Seidel 2023).

References

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.