Technical Report: Final Project DS 5110: Introduction to Data Management and Processing

Team Members: Thomas Simmons, Nicholai Platonoff, Aidan Lee

December 3, 2024

1 Introduction

This report documents our final project for DS 5110: Introduction to Data Management and Processing. The project focuses on designing and implementing a database management system for a movie theater, enabling advanced data analysis and reporting to support decision-making. The motivation for this project stems from the need for movie theaters to utilize data-driven insights to build and enhance the customer experience, improve operational efficiency, and maximize overall revenue.

The primary objectives of this project include:

- Designing a comprehensive relational database to model the operations of a movie theater, including data on movies, showtimes, customers, employees, concessions, tickets, and reviews.
- Generating realistic datasets for analysis, including the application of various statistical distributions to simulate real-world scenarios.
- Writing a variety of SQL queries to extract actionable insights, such as identifying popular movies, tracking revenue, analyzing attendance trends, and summarizing customer demographics.
- Exploring advanced reporting techniques, including Power BI visualizations, to present findings in an accessible and impactful manner.

The scope of the project covers the complete data pipeline, starting from schema design, dataset generation, and query implementation, to producing detailed analytical reports. The database schema supports various aspects of theater operations, such as movie scheduling, ticket sales, concessions management, customer feedback, and employee performance. The inclusion of attributes like promotions, discounts, and demographic data enables richer analysis.

In the following sections, we will expand on the methods, results, and conclusions drawn from the project, demonstrating the potential of structured data management and processing in transforming theater operations into a data-driven enterprise.

2 Literature Review

There is a solid amount of existing research relevant to the design and implementation of database management systems in the context of movie theaters. This review explores methodologies for database schema design, data generation, analytics, and visualization techniques, speaking on findings and identifying gaps.

Relational databases have been extensively studied for their ability to manage structured data efficiently. Edgar F. Codd's relational model (1970) laid the foundation for modern database design, emphasizing normalization to reduce redundancy and ensure data integrity. For movie theaters, prior research demonstrates the use of relational models to represent entities such as movies, customers, tickets, and schedules.

However, gaps remain in adapting relational schema designs for dynamic theater operations. For instance, the integration of real-time data, such as customer feedback or last-minute schedule changes, poses challenges to traditional database models. Additionally, schema designs lack scalability when working on multi-theater operations or complex datasets involving customer demographics, concessions, and reviews.

SQL-based analytics are crucial for extracting insights from structured data. Research emphasizes techniques such as indexing, query optimization, and partitioning to improve performance when dealing with large datasets. Within the context of theaters, analysis often focuses on metrics like attendance trends, revenue generation, and customer segmentation. While methodologies for optimizing SQL queries are well-documented, gaps persist in integrating multiple data sources (attendance, revenue, demographics) for comprehensive insights. Existing studies rarely explore advanced techniques, such as predictive modeling or combining structured data with unstructured inputs such as customer reviews.

Existing research provides valuable methodologies for database design, data simulation, and analysis. However, several gaps hinder their application in the context of movie theaters:

- Domain-Specific Applications: Limited studies address the specific needs of theaters, such as handling dynamic scheduling, customer feedback, and concessions data.
- Integrated Approaches: There is a lack of comprehensive frameworks that

combine database design, simulation, analysis, and visualization to address operational challenges holistically.

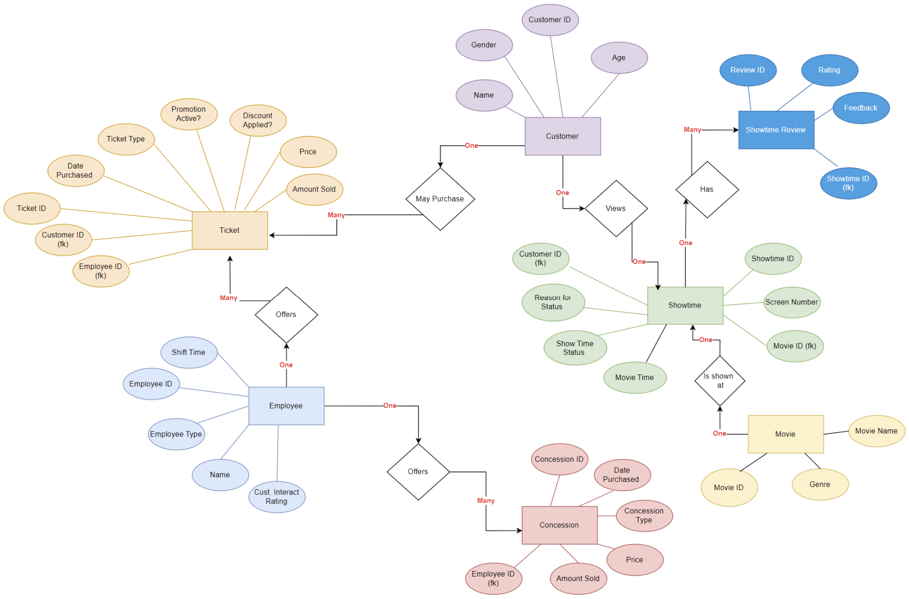
- Scalability and Real-Time Data: Current methodologies often overlook the need for scalability and real-time updates in multi-theater or high-volume environments.

This project aims to address the identified gaps by focusing on domain-specific challenges and incorporating advanced analytics to enhance theater operations. By leveraging tailored datasets, optimized queries, and interactive visualizations, the project contributes to bridging the gap between theoretical research and practical applications in the entertainment industry.

3 Methodology

3.1 ERD Diagram

Before starting the implementation, we created an ERD diagram to outline the design of the database listed below:



Tables: Customer, Employee, Concession, MovieInfo, ShowTime, ShowTime Review, and Ticket.

3.2 Summary Of Tables

- **Customer:** Contains information about each individual customer. Information tracked: Customer ID, Name, Gender, and Age.
- **Employee:** Contains information about all employees which includes: Employee ID, Shift Time, Name, Employee Type, and Customer Interaction Rating.
- **Concession:** Contains information about each transaction purchased at the Concession Stand. Information Tracked: Concession ID, Concession Type, Price, Amount Sold, Date Purchased, and Employee ID
- **MovieInfo:** Contains information about each movie that is shown at the movie theater. Information tracked: Movie ID, Movie Name, and Genre.
- **ShowTime:** Contains information about each possible movie showtime at the theater. Information tracked: Showtime ID, Screen Number, Movie Time, Showtime Status, Reason for Status, Movie ID, and Customer ID.
- **ShowTime Review:** Contains information about customer feedback at the theater. Information tracked: Review ID, Rating, Feedback, and Showtime ID.
- **Ticket:** Contains information about the ticket transaction order of the customer. Information tracked: Ticket ID, Date Purchased, Ticket Type, Price, Amount Sold, Promotion Active, Discount Applied, Employee ID, and Customer ID.

3.3 Creation of the Dataset

There was not a lot of data cleaning necessary for this project. The only data cleaning that was needed was to remove any duplicated rows created by the Faker package. Although, the Faker package does try its best to not print out duplicates due to the large dataset being created there were some duplicate rows that were produced and had to be removed afterwards.

In order to obtain a dataset that properly fit the project, we would have to generate a synthetic dataset. As mentioned earlier, the tables that were created were the Customer Data, Employee Data, Concession Data, Movie Info, ShowTime Data, ShowTime Review Data, and Ticket. The tables were created based on the ERD Diagram created during the pre-planning process of the project.

For the generation of the dataset, we mainly used Python and the Faker package to create synthetic data for the movie database. Although most of the information like employee names and customer names are generated randomly, there were weights applied on several columns such as Ticket types in the Ticket table and Screen Condition in the ShowTime Review table to make the data more realistic to a real movie theater database.

In this generated dataset, our movie theater contains 5 different movie screening rooms with screening times ranging from 9 am - 12 am showing 30 different types of movies through January 2023 to January 2024. The movies genres in this data set are Action, Animation, Biography, Comedy, Crime, Drama, Sci-Fi, Thriller, and Western. Moreover, there were three different ticket types that customers can buy which are: Standard, 3D, and VIP all ranging from different price points. The movie theater has over 120 employees who have the responsibility of either a Box Office, Concession Stand, or Manager.

The Employee and Movie table have around 30 - 120 row entries while the other five tables have over 10,000 row entries.

Lastly, each of the tables were converted into a panda data frame and exported as a csv file. Once the csv files were created, they were all loaded one by one into the pre created sqlite database called: MovieTheater.db.

3.4 Analytical Techniques

The Analytical Techniques that were used throughout the project was SQL Querying and Visualization Dashboards. More specifically, the Visualization tools used were Microsoft Power Business Intelligence and Tableau. The group spent a majority of the weeks running SQL queries on the MovieTheater Database and exporting the results into the Visualization tools to create graphics that hold statistical importance about the data.

4 Results

Top 10 Most Popular Movies: - Blockbusters dominated attendance due to their critical acclaim and franchise appeal. - Mid-tier movies also performed well, demonstrating opportunities to diversify programming.

Daily Attendance Patterns: - Weekends and holidays showed peak attendance, aligning with expectations. - Certain weekdays had unexpectedly high attendance, warranting further investigation.

Revenue Insights: Total Revenue by Movie: - Revenue disparities highlight the impact of premium ticket pricing (3D, VIP). - Moderate-attendance movies with higher pricing generated significant revenue.

Monthly Revenue: - August had the highest revenue, attributed to summer vacation and favorable movie releases. - Concessions contributed significantly, with popcorn generating the highest sales.

Demographics by Genre: - Younger audiences preferred Action and Sci-Fi; older audiences leaned toward Drama and Biographical films. - Gender preferences varied significantly by genre.

Peak Hours: - The busiest period was from 5:00 PM to 6:00 PM. - Recommendations: Allocate more employees during peak times to enhance service efficiency.

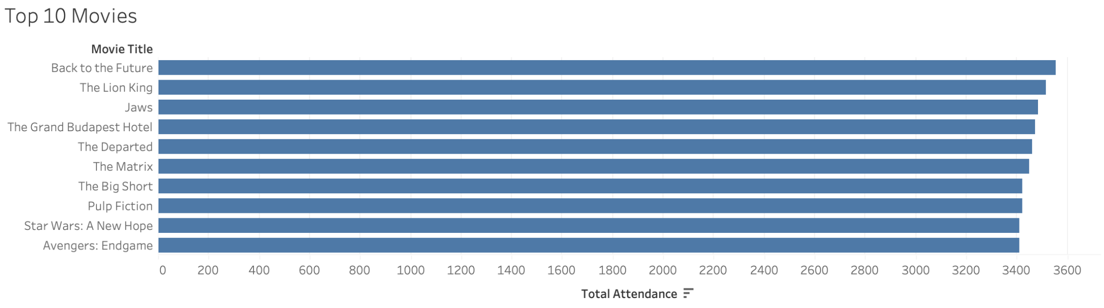
Notable Trends: - Niche genres performed better than anticipated, reflecting local audience preferences. - Unexpected trends, such as high weekday attendance, suggest opportunities for tailored promotions.

5 Discussion

The results of our analysis provided important insights into movie attendance trends, revenue generation, and audience demographics. These findings offer several implications for theater operations, marketing strategies, and customer engagement, as seen below.

5.1 Implications of Key Findings

The Top 10 Most Popular Movies analysis revealed that blockbusters with widespread critical acclaim or franchise appeal dominate attendance figures. This aligns with industry expectations that high-budget, heavily marketed films attract the largest audiences. However, mid-tier movies also featured prominently, suggesting opportunities for theaters to diversify programming by including niche genres with loyal audiences.

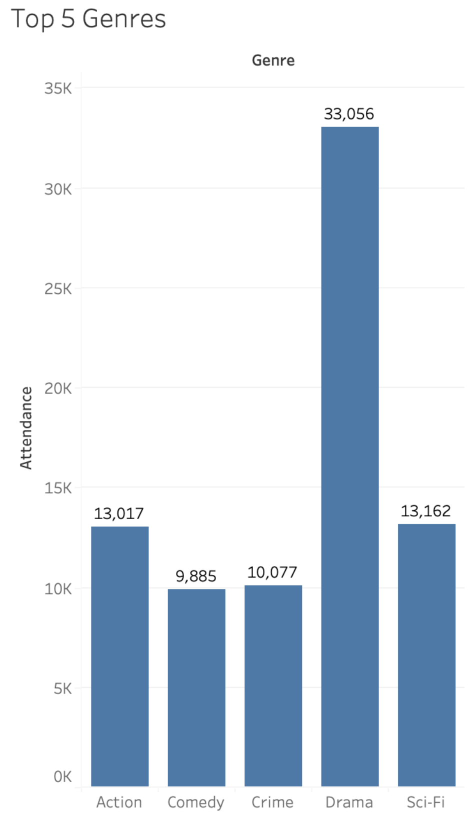


Sum of Total Attendance for each Movie Title.

The Total Revenue by Movie analysis highlighted significant revenue disparities between high-attendance movies and those with premium ticket pricing. For example, some movies with moderate attendance generated more revenue due to higher ticket prices for special formats (e.g., 3D, VIP). This underscores the importance of pricing strategies and premium offerings in revenue optimization.

Daily attendance patterns from the Daily Attendance Summary visualization showed spikes during weekends and holidays, which align with existing research in the literature review. This supports the notion that theaters should focus promotional efforts during these periods to maximize attendance. However, some weekdays showed unexpectedly high attendance, which may warrant further investigation into specific promotions or events that influenced these trends.

The Top 5 Genres by Attendance visualization confirmed that Action, Drama, and Sci-Fi genres dominate theater attendance. Genres like Animation performed particularly well in family-oriented theaters, suggesting that targeted marketing strategies for different audience segments could enhance profitability.

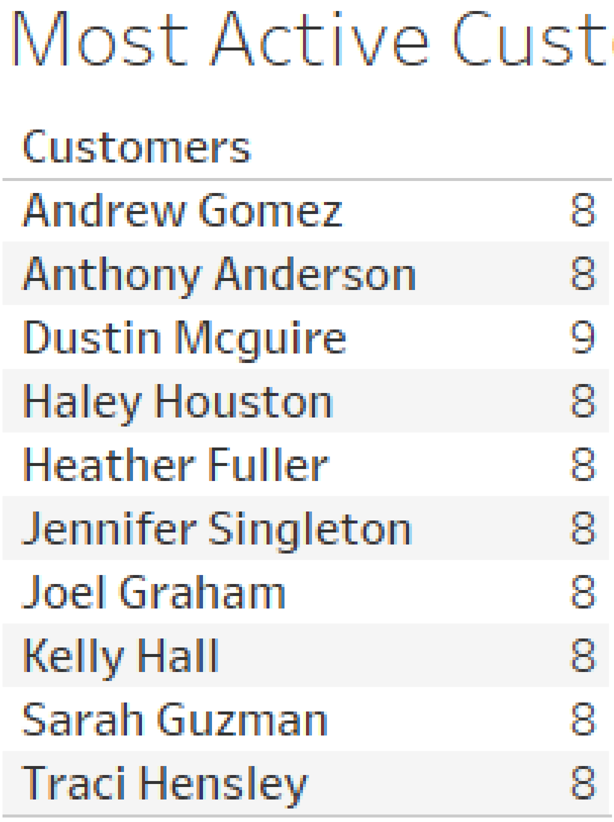


Sum of Attendance for each Genre. The marks are labeled by sum of Attendance.

The Customer Demographics by Movie Genre analysis demonstrated that genres appeal to distinct demographic groups. For instance, younger audiences were more likely to watch Action and Sci-Fi movies, while older audiences preferred Drama and Biographical films. Gender preferences also varied significantly, with certain genres attracting predominantly male or female audiences.

These insights could inform tailored promotional campaigns and programming decisions.

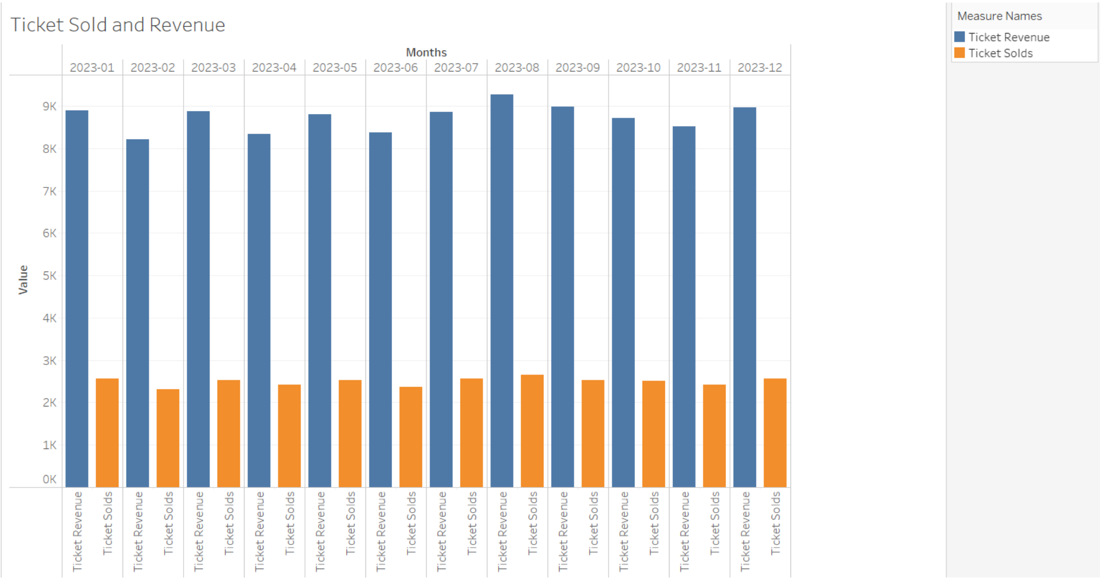
The Most Active Customers analysis highlights the individuals who are the most frequent visitors at the move theater. Although Dustin Mcguire was the most frequent visitor of the 2023-24 year, that statistic can change really quickly based off what movies are being shown at the theater during the time. People will be more influenced to go to the movies more often if the movies being shown match their interests more.



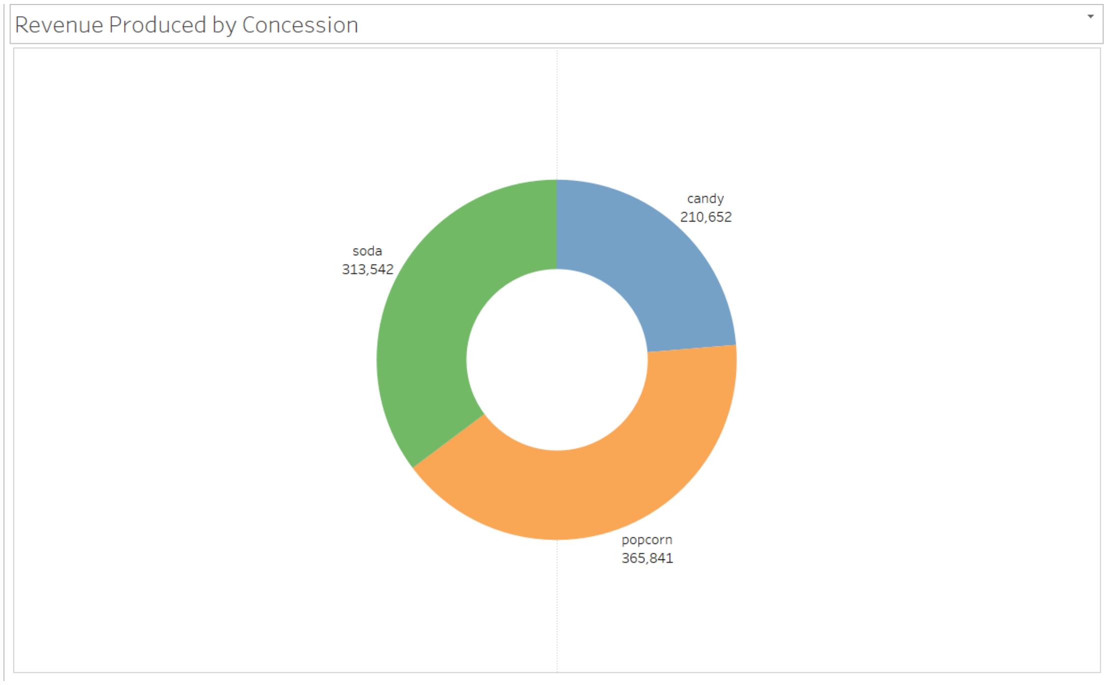
Most Active Customers

Customers	
Andrew Gomez	8
Anthony Anderson	8
Dustin Mcguire	9
Haley Houston	8
Heather Fuller	8
Jennifer Singleton	8
Joel Graham	8
Kelly Hall	8
Sarah Guzman	8
Traci Hensley	8

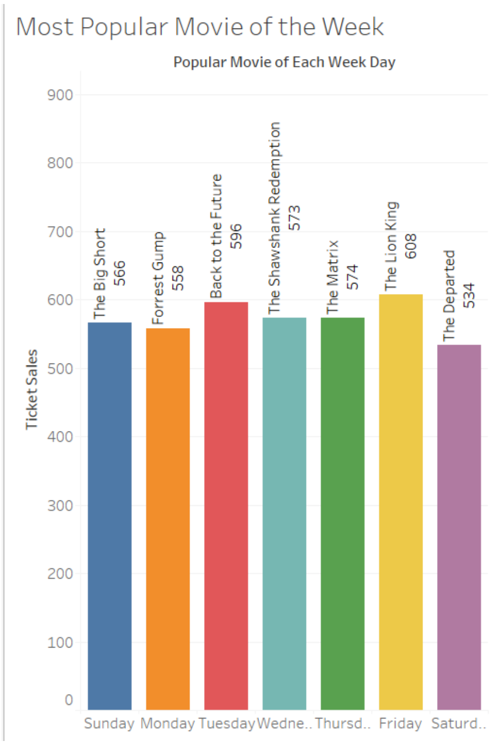
The Monthly Revenue Summary analysis demonstrates that the month with the highest amount of revenue is the month of August. This reason being so could possibly be since the movie genres are more favored to the younger demographic many of their younger audience has more opportunities to head to the movie theaters during this time since summer break is in session. A movie theater environment also gives young adults a great place to hang out with friends at an affordable cost.



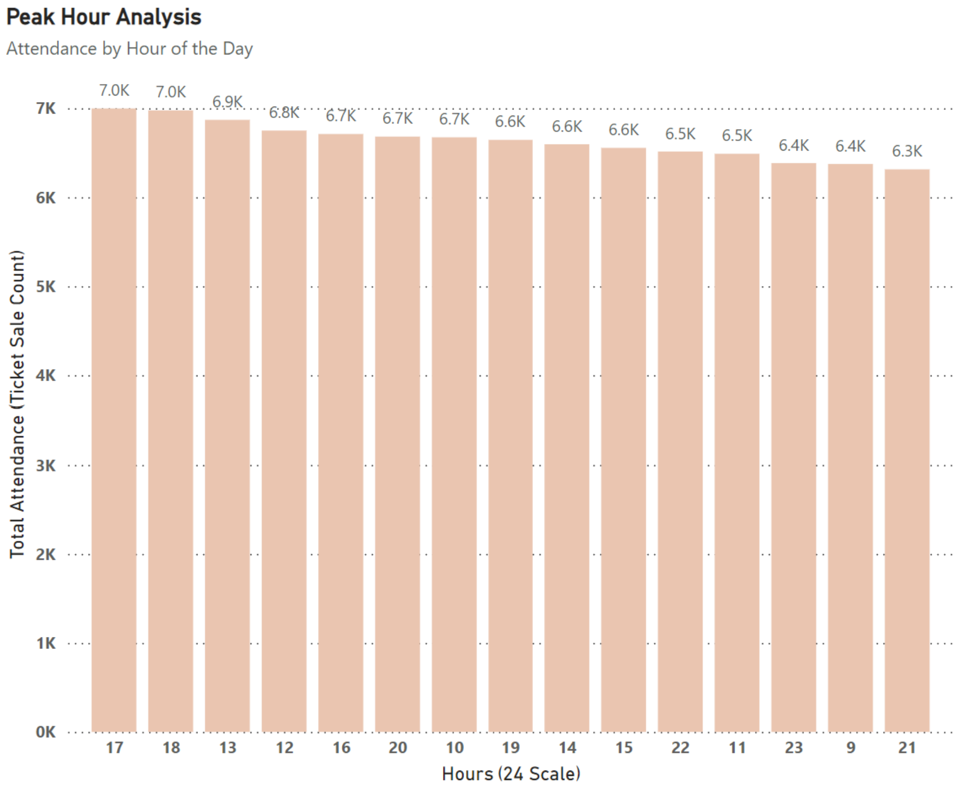
The total concession revenue by item visualizations displays the amount of revenue gained per each concession item at the stand. The items that are served at the concession are: popcorn, candy, and soda. Based off the visualization, we can identify that the popcorn item brings in the most revenue at the concession. It is not really a surprise here because popcorn has been a staple movie snack for a few decades now starting from 1938. Since popcorn is a hot commodity to have when watching a movie, movie theaters tend to up charge the price of the snack at an attempt to gain more money from their customers while still meeting their demands as well.



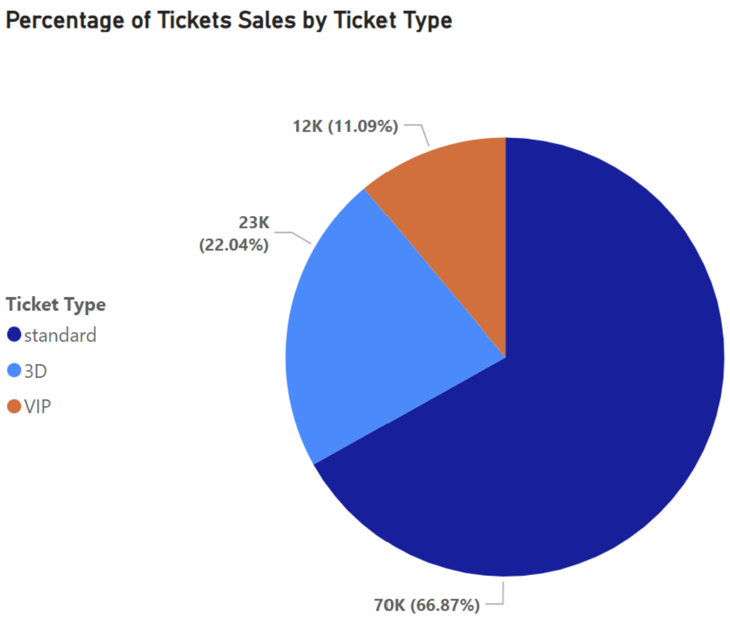
The analysis about which movies have the highest sales on each day of the week confirms the genre claim that was made earlier in the article about Action, Drama, and Sci-Fi genres being dominant at the box office. Since from Monday - Sunday through out the year, the top movie of each day of the week has always been either a Action, Drama, or Sci-Fi movie in the top spot. The strategy team of the movie theater could potentially use this insight and insert more movies of these genres to attract more people to come more often and spend more money since they understand their audience's preference.



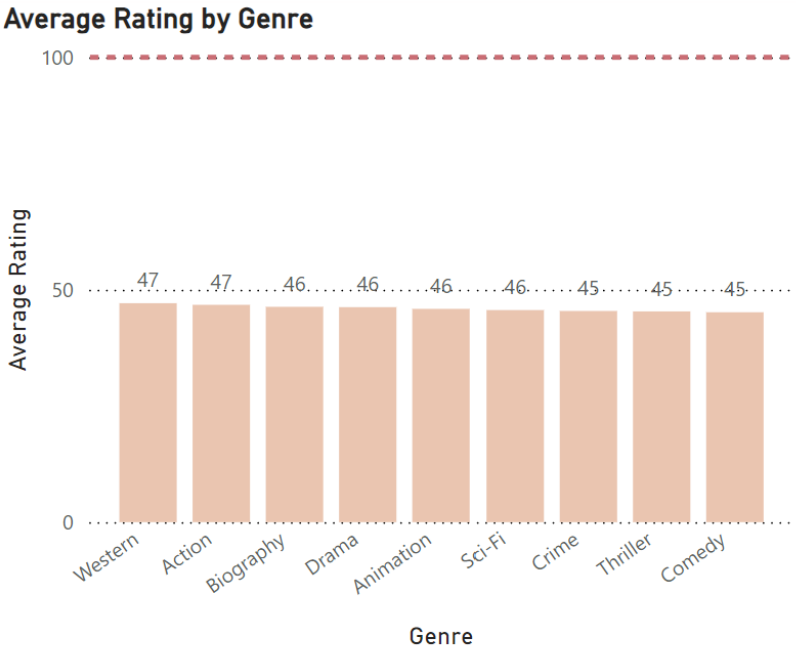
The Peak Hour analysis reveals what time of day is the most busiest for the movie theater. The busiest time at the movie theater is between 5:00 PM - 6:00 PM. The strategy team can use this information to allocate more employees to take more shifts at that time to ensure that the evening rush runs smoothly and minimize the potential loss of customers due to the long lines.



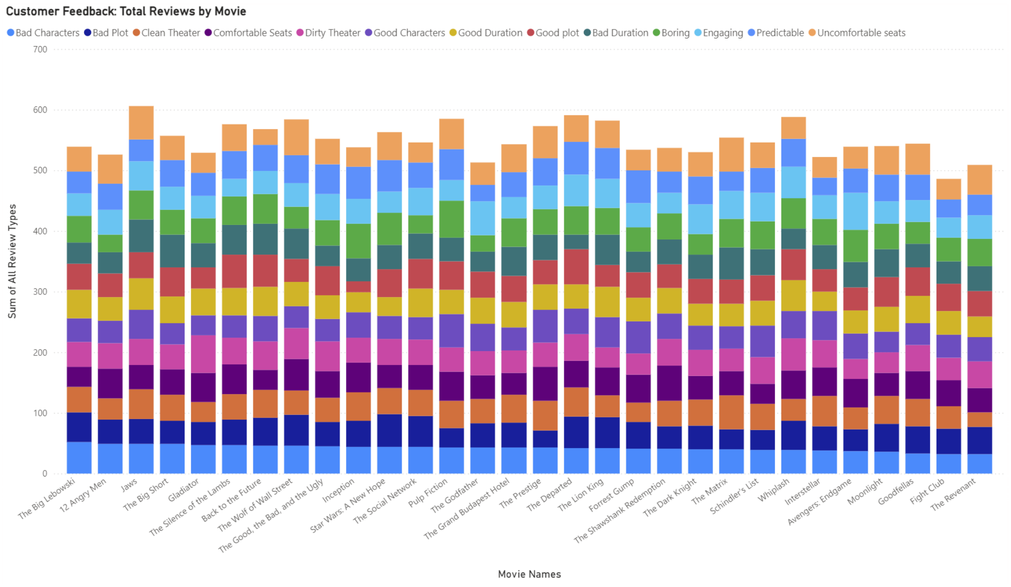
The Ticket Sales by Type analysis demonstrated that the expectation was true: a majority of standard tickets were sold (roughly 70 percent). A sales team could use this basic metric multiple times to track how different promotional strategies and discounts effect these percentages. Below is corresponding pie-chart.



Average Rating by Genre and Average Rating by Movie Analysis: The results for both show averages for all genres and all movies to be around 50 out of 100. This is because our data was generated from a uniform distribution. In the real world, this analysis would show different results. It is expected that some movies and some genres are more popular than others. A movie theater could leverage this information to show popular movies and genres more often than less popular ones. Below is the average rating by genre bar chart. The average rating by movies is included in the "Additional Figures" section.

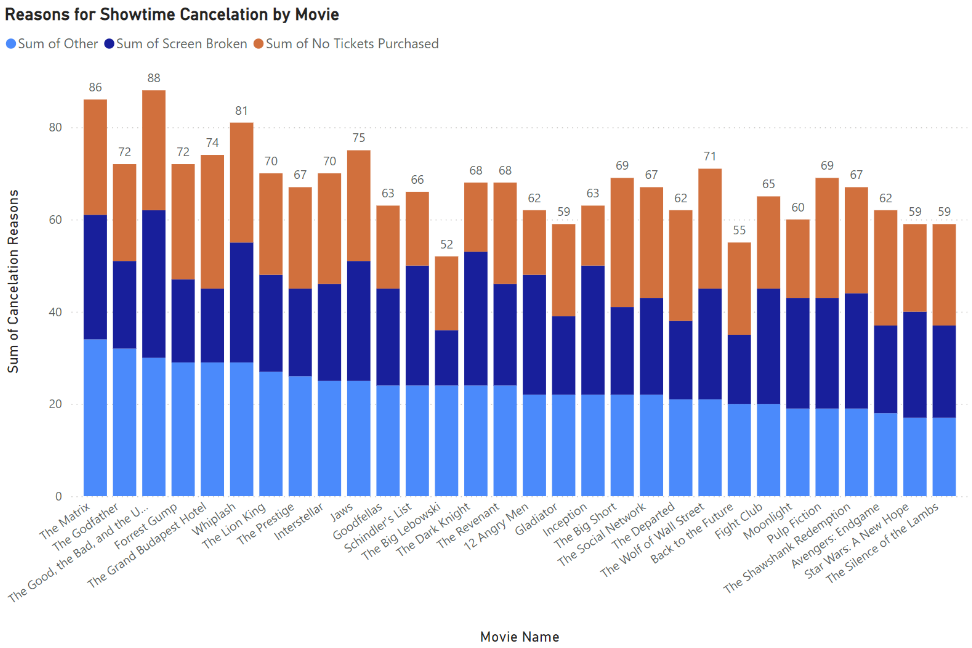


Customer Showtime Feedback Analysis: Customers were asked to fill out a survey after they finished a showing of a particular movie. They would select from a preset options list that reflect quality of movie or quality of theater room. Examples include: "good characters", "bad characters", "good seats", "bad seats", etc. Again, since these reviews were generated from a uniform distribution, all screening reviews are the roughly the same across all movies. In reality, they would be different. Management would be able to identify screening rooms that were often rated dirty or uncomfortable to correct issues. They could also uses these reviews to see which movies are the most popular in order to screen them more. The bar chart is shown below

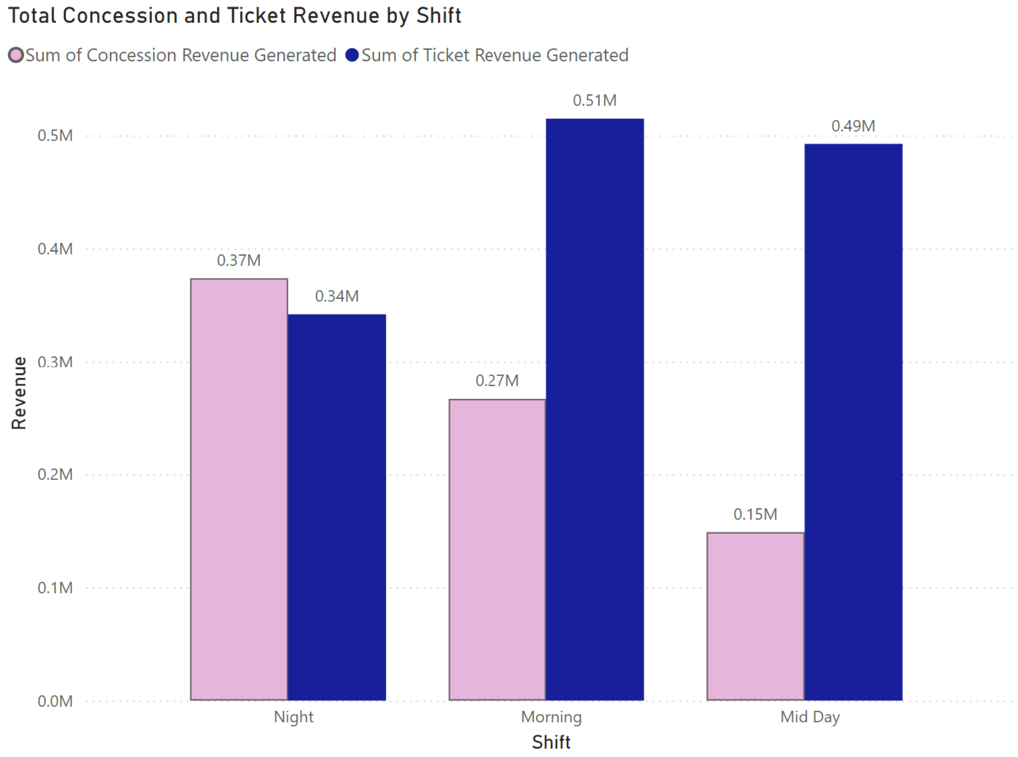


Discount Usage Summary: It was discovered that 29,707 discounts and promotions were applied over the year. Of those discounts and promotions, customers saved $737,140 dollars. A finance team could use these figures to understand how these discount usages effected the theater budget and if the discounts were over or under-used. There is no figure required for this statistic.

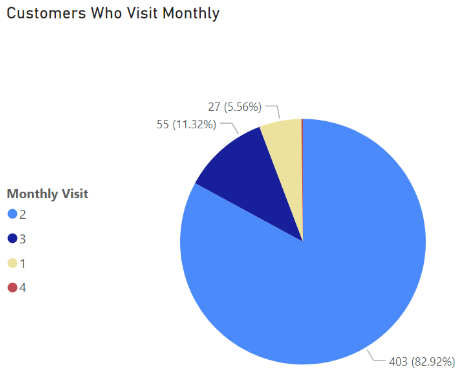
Reasons for Showtime Cancellation: Employees had three reasons to mark why they canceled a movie's showing. 1: There were no tickets purchased, 2: The screen was broken, 3: marked "other". Similarly, We also tracked why showing's were rescheduled. Both of these analysis would help a movie theater company better understand where maintenance failed to upkeep the theater, and where unpopular showings occurred leading to cancellation. The cancellation figure is below, while the very similar reschedule graph is in "additional figure" section.

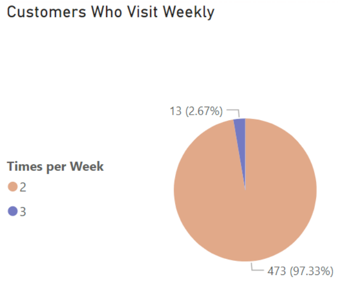


The revenue was analyzed by shift as well. Each shift (morning, midday, night) had their total concession and ticket sale revenues totaled. Management could use these figures to try and help drive sales in the other shifts. Ultimately, morning shift had the highest ticket revenue generation, while night shift had the highest concession revenue. Below is the figure.



Analysis was also done on the number of repeat monthly and weekly customers. Four hundred customers were found to visit the theater the same month at least once. Another four hundred were found to have visited the theater in the same week at least once. These insights could be used to see if the theater is meeting targets of repeat business. Low figure of repeat business could indicate poor customer satisfaction. The figure is below.





5.2 Comparison with Literature Review

The findings corroborate the established literature on audience behavior and revenue generation, particularly in highlighting the importance of blockbuster movies and premium offerings. However, discrepancies were observed in the performance of niche genres, which performed better than expected in our dataset. This could reflect local audience preferences or the inclusion of promotions that influenced attendance.

While the literature suggested a significant impact of pricing on attendance, our results showed that attendance figures were more influenced by movie popularity than ticket pricing. This finding may suggest that theaters have some flexibility in pricing without significantly affecting attendance, though further research is needed to confirm this hypothesis.

The analysis relied on synthetic datasets, which, while realistic, may not capture all real-world complexities. Additionally, some unexpected trends, such as high weekday attendance for specific dates, highlight the need for deeper exploration of contextual factors such as local events, promotions, or weather conditions. Future research could integrate external datasets to provide a more comprehensive analysis.

Advanced techniques such as predictive modeling or sentiment analysis of customer reviews could extend the current findings. These methods could identify potential factors influencing attendance and revenue more effectively, providing actionable insights for how the theaters are managed.

6 Conclusion

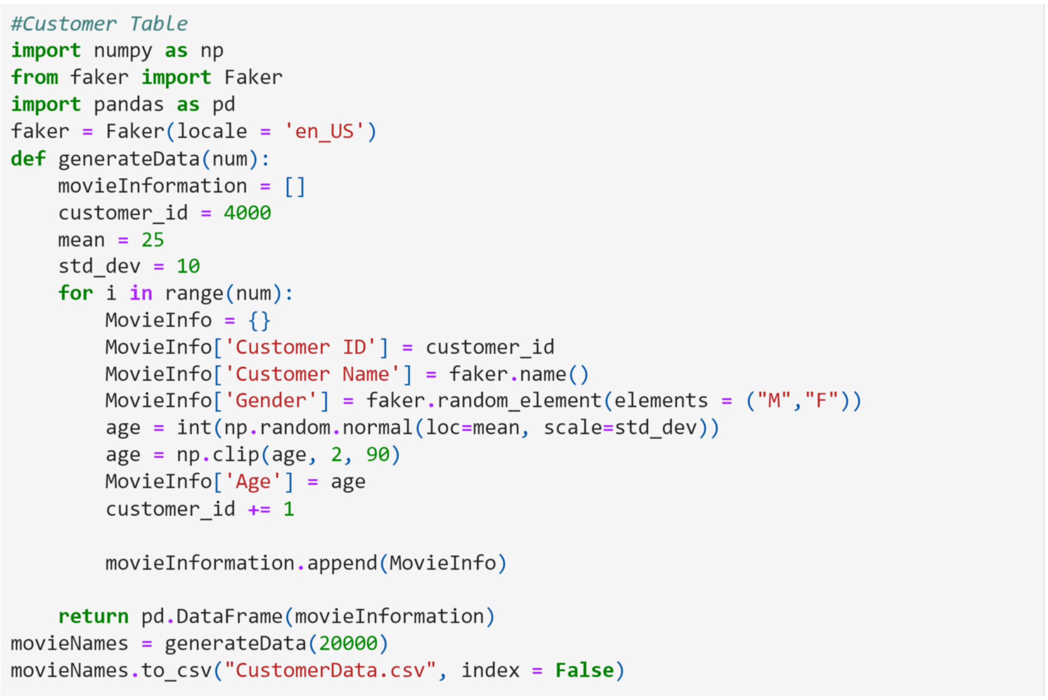
In summary, this project demonstrated the potential of leveraging structured data management and advanced analytics to enhance movie theater operations. By designing a robust relational database schema, generating realistic synthetic datasets, and applying a range of SQL-based analytical techniques, the project successfully provided actionable insights into audience behaviors, revenue optimization, and operational efficiencies. The findings underscore the importance of blockbuster movies, targeted marketing strategies, and premium offerings in driving revenue and customer engagement.

While the results aligned with many established patterns from the literature, they also revealed unique trends, such as the surprising performance of niche genres and unexpected weekday attendance peaks. These insights highlight the value of combining traditional data analysis with contextual exploration to adapt strategies to specific market conditions.

Despite the successes, the limitations in the synthetic dataset and scope of analysis leave room for improvement. Future work could incorporate real-world datasets, integrate external factors such as weather or local events, and employ advanced techniques like predictive modeling or sentiment analysis to provide even more comprehensive insights.

This project illustrates the power of our data-driven decision-making in the entertainment industry and sets a foundation for further potential exploration into scalable, real-time data solutions for multi-theater operations.

7 Appendix A: Code



```
#Customer Table
import numpy as np
from faker import Faker
import pandas as pd
faker = Faker(locale = 'en_US')
def generateData(num):
    movieInformation = []
    customer_id = 4000
    mean = 25
    std_dev = 10
    for i in range(num):
        MovieInfo = {}
        MovieInfo['Customer ID'] = customer_id
        MovieInfo['Customer Name'] = faker.name()
        MovieInfo['Gender'] = faker.random_element(elements = ("M","F"))
        age = int(np.random.normal(loc=mean, scale=std_dev))
        age = np.clip(age, 2, 90)
        MovieInfo['Age'] = age
        customer_id += 1

        movieInformation.append(MovieInfo)

    return pd.DataFrame(movieInformation)
movieNames = generateData(20000)
movieNames.to_csv("CustomerData.csv", index = False)
```

Above is the Faker code that generates the Customer data table entries. Our other data sets are generated with similar code. Some important notes:
The unique customer ID begins at 4000 and is auto incremented.
We are able to generate age data that matches real world expectations by providing a mean and standard deviation for Faker to follow.
The number of data entries to randomly produce is provided as the ”num” variable.

8 Reference

Codd, E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM, 13(6), 377–387. https://doi.org/10.1145/362384.362685

9 Appendix B: Additional Figures

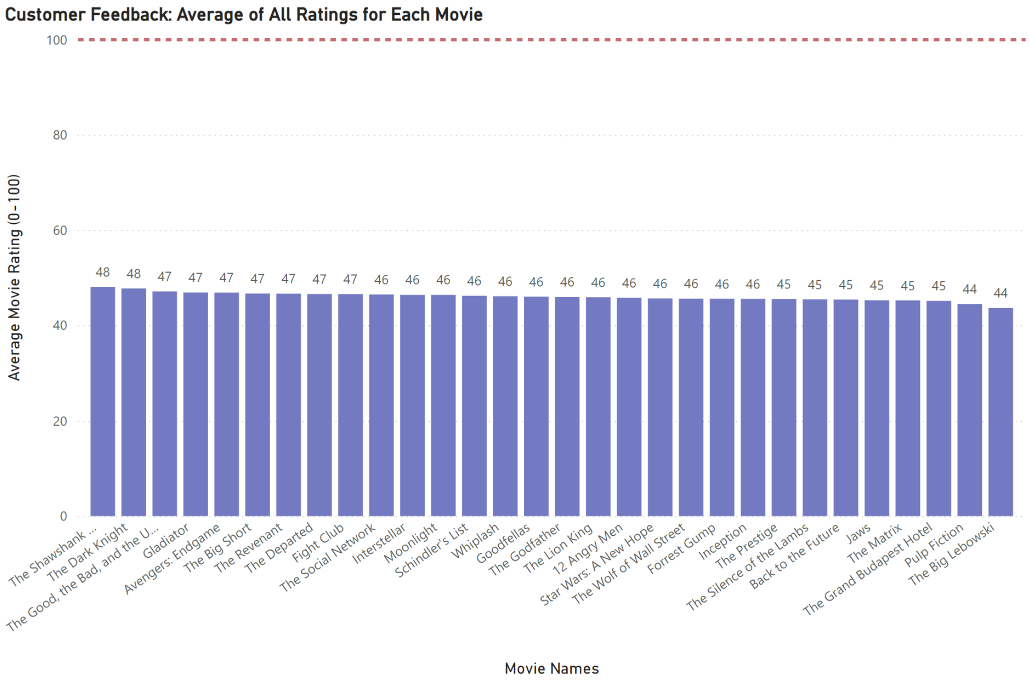


Figure 1: Average Ratings

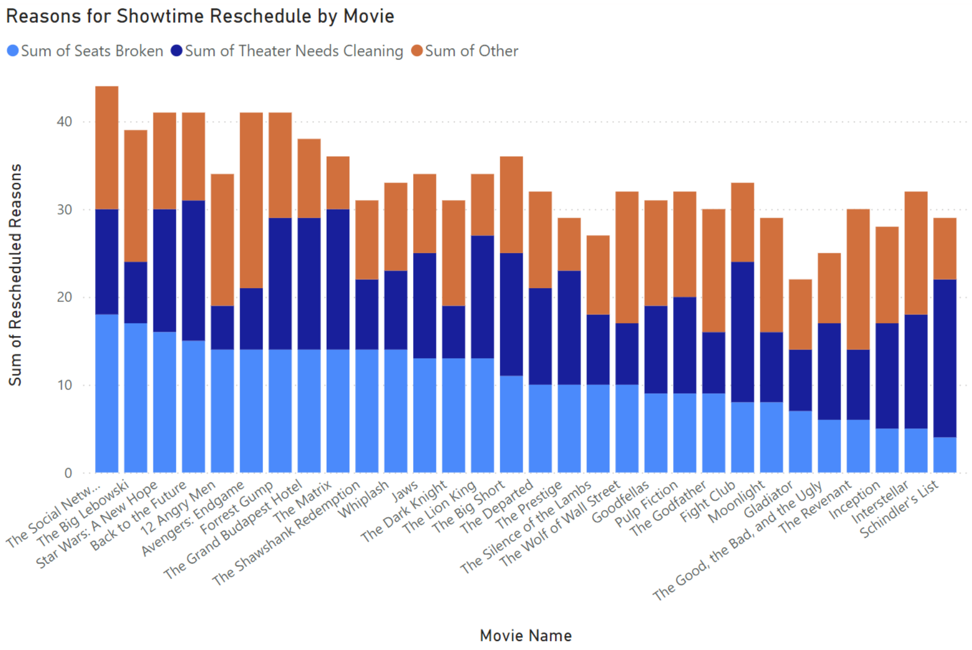


Figure 2: Reasons for Reschedule

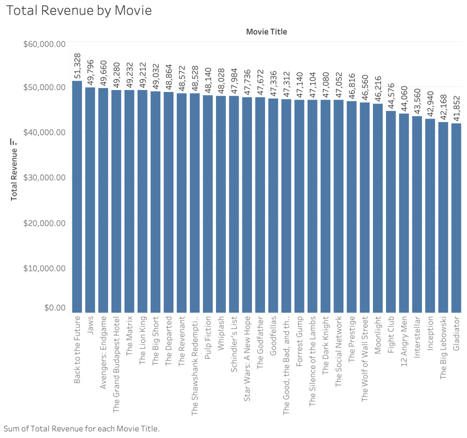
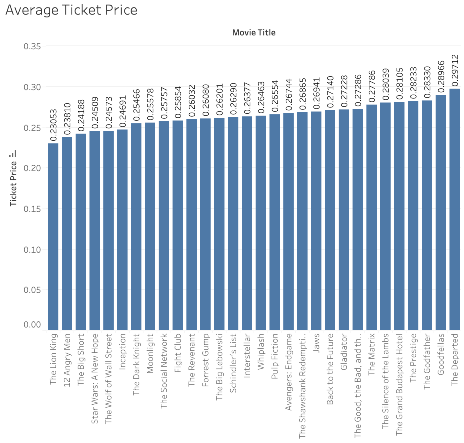


Figure 3: Total Revenue by Movie



Sum of Ticket Price for each Movie Title.

Figure 4: Average Ticket Price