**DSA210 Final Report**

**Project:** *Seasonal Trends in Movie Viewership and Genre Popularity in Turkey*

**Student:** Hilal Öngel

ID: 32425

**Date:** 29.05.2025

**Supervised by**: Selim Balcısoy

**Introduction**

This project investigates the relationship between seasons, the number of moviegoers, and the popularity of movie genres in Turkey. It combines Exploratory Data Analysis (EDA), hypothesis testing (ANOVA), and a supervised machine learning model (Random Forest Regressor) to uncover trends and make predictions. The final model explains over 71% of the variance in weekly moviegoer numbers, demonstrating strong predictive power from seasonal and genre-based variables. To ensure methodological rigor, alternative models such as Linear Regression and Decision Tree Regressor were also implemented. However, Random Forest consistently outperformed them, particularly in terms of R² score, justifying its selection as the final model.

**Motivation**

The motivation for this project stems from the observation that cinema attendance exhibits notable fluctuations throughout the calendar year. Understanding how these patterns correspond to genre popularity and audience volume holds practical relevance for film distributors, marketers, and cultural analysts. On a personal level, the project aligns with the researcher's interests in cinema, data-driven behavioral analysis, and the strategic application of data science in real-world cultural contexts.

**Hypotheses**

* **H1:** There is a statistically significant difference in the average weekly number of moviegoers across seasons.
* **H2:** Specific film genres exhibit higher popularity in particular seasons (e.g. Action and Adventure in Summer; Drama and Thriller in Winter).

**Data Sources and Preparation**

The study utilizes two primary data sources:

* **Box Office Turkey:** Weekly Top 10 box office reports for the years 2022, 2023, and 2024, including audience counts.
* **TMDB (The Movie Database) API:** Metadata including genre classification.

Seasonal classification was operationalized using the following schema based on Turkish seasons:

* Winter: Weeks 1–8 and 49–52
* Spring: Weeks 9–21
* Summer: Weeks 22–34
* Autumn: Weeks 35–48

**Exploratory Data Analysis (EDA)**

This project applies several Exploratory Data Analysis (EDA) techniques to understand the relationship between seasons, movie attendance, and genre popularity.

**1. Grouping and Aggregation**

* Calculated average weekly moviegoer attendance across seasons.
* Summarized total viewership by genre and season.
* Tools: groupby(), mean(), sum(), pivot().

**2. Data Cleaning and Preprocessing**

* Movie title cleaning (removal of release dates, extra spaces).
* Genre matching using TMDB API.
* Exploding multi-genre entries so that each genre appears in a separate row.
* Season categorization: Assigned each week to one of the four seasons (Winter, Spring, **Summer, Autumn).**

**3. Statistics**

* Use of statistical measures like mean, sum, and median.
* Supported the identification of seasonal fluctuations and genre preferences.

**4. Data Visualization**

* Bar Charts: Visualized average weekly attendance by season.
* Heatmaps: Showed total viewership per genre across seasons.
* Stacked Bar Charts: Illustrated genre contribution to seasonal attendance.
* Boxplots: Displayed weekly viewership distribution across seasons.
* Tools: matplotlib, seaborn.

**5. Statistical Testing (Supporting EDA)**

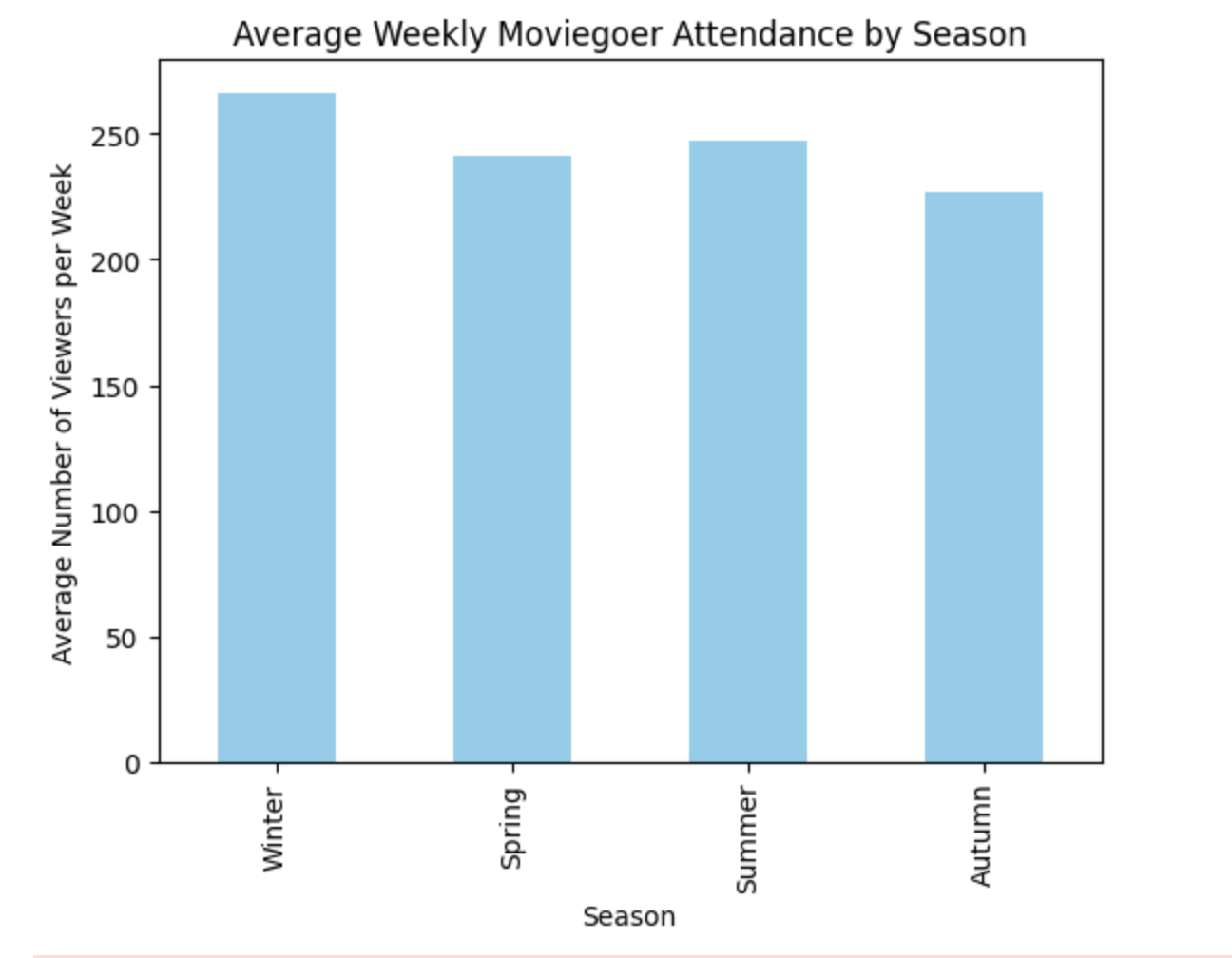
* Applied ANOVA (Analysis of Variance) to statistically test whether the seasonal differences in attendance are significant.
* Interpreted F-statistic and p-value to confirm or reject the null hypothesis.
* This statistical testing strengthened the insights obtained from the visual analysis.

**Findings:**

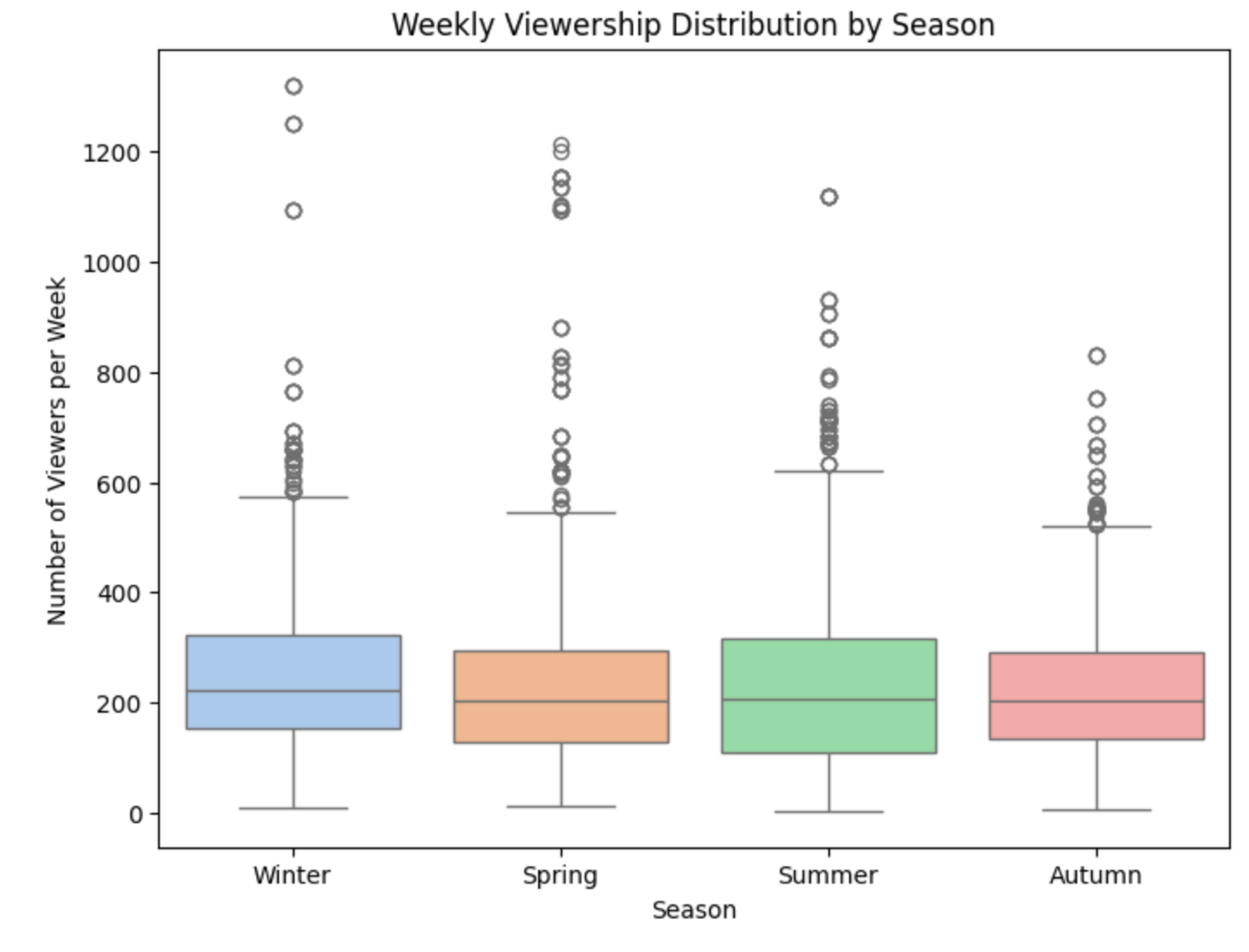
* Winter months showed elevated median attendance, with lower variance
* Summer months favored Action and Adventure films
* Comedy maintained a relatively stable distribution year-round, with slight peaks in Spring and Summer

**Hypothesis Testing**

* One-way ANOVA (Analysis of Variance) was applied to test whether the average weekly movie attendance differs across seasons.
* F-statistic: 8.08
* p-value: 0.0000
* Since the p-value is below 0.05, the null hypothesis was rejected, it is confirmed that seasonal differences in average weekly movie attendance are statistically significant, which supports Hypothesis 1.



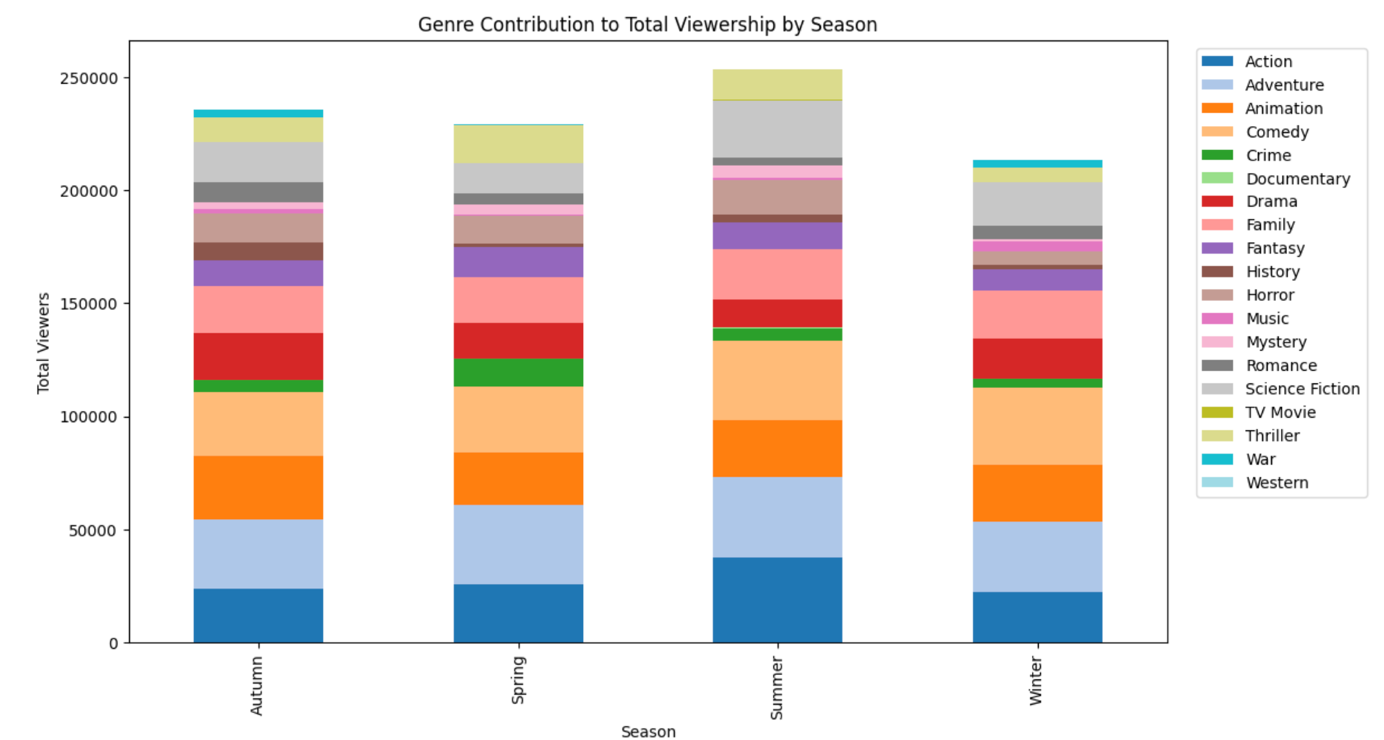
* The **F-statistic** reflects the ratio of between-group variance (seasonal differences) to within-group variance (weekly fluctuations).
* The result indicates that at least two seasons significantly differ in their average audience numbers.
* This conclusion was visually supported by the **boxplot**, showing **Winter** as the season with the highest median weekly attendance and relatively low variability.



**Hypothesis 2 Evaluation: Genre Popularity Patterns**

In order to assess **Hypothesis 2**, the distribution of movie genres across seasons was analyzed after exploding multi-genre entries. The results showed clear seasonal genre preferences:

* **Action** and **Adventure** genres were dominated in **Summer**, aligning with blockbuster release strategies.
* **Drama** and **Thriller** genres were more popular in **Winter**, matching seasonal audience preferences for serious and emotional narratives.
* **Comedy** maintained stable popularity across the year but slightly increased in **Spring** and **Summer**.



metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Although no formal statistical test (e.g., chi-square) was applied to genre-season relationships in this project, the descriptive analysis strongly supports Hypothesis 2.  
Future work may apply statistical tests to validate these patterns formally.

**Notes**

This analysis focuses on the Top 10 weekly movies.  
Although this reflects the dominant commercial trends, niche genres or independent films outside of the Top 10 may not be fully represented.  
The methodology can be extended to larger datasets (Top 20, Top 50) or applied to streaming data for future research.

**Machine Learning**

To move beyond descriptive statistics, this project incorporates a supervised learning model to **predict weekly moviegoer numbers**.  
By combining categorical and temporal features with engineered variables, the model learns patterns that influence audience attendance.

**Model Comparison Insights**

To support model selection with empirical evidence, three different models were evaluated:

| **Model** | **MAE** | **RMSE** | **R²** |
| --- | --- | --- | --- |
| Random Forest | 60.29 | 89.09 | 0.7107 |
| Decision Tree | 65.47 | 99.88 | 0.6254 |
| Linear Regression | 71.83 | 105.12 | 0.5962 |

* **MAE (Mean Absolute Error)** indicates how much, on average, the model's predictions deviate from actual values.
* **RMSE (Root Mean Squared Error)** penalizes larger errors more severely and is thus more sensitive to outliers.
* **R² (Coefficient of Determination)** measures how well the model explains the variance in the target variable.

Random Forest achieved the **lowest MAE and RMSE**, as well as the **highest R² score**, indicating not only the most accurate average predictions but also the best ability to capture general patterns in the data. This makes it the most suitable model for the final predictive task.

This project includes a **regression model** that predicts the number of weekly moviegoers using the following features:

* **Season**
* **Genres**
* **Week**, **Year**
* Engineered Features:
  + Movie\_Rank: film’s position within Top 10
  + IsHolidayWeek: whether the week includes a holiday
  + GenreCount: number of genres (1 for exploded data)

**Movie\_Rank**

* **What it is:** Position of a movie in the weekly Top 10 list (1 = highest, 10 = lowest)
* **Why it matters:** Higher-ranked movies typically receive more attention.

df['Movie\_Rank'] = df.groupby(['Year', 'Week']).cumcount() + 1

**IsHolidayWeek**

* **What it is:** 1 if the release falls on a holiday-heavy week (weeks 1–2, 26–31, 52), else 0
* **Why it matters:** Holidays tend to attract more viewers.

holiday\_weeks = list(range(1, 3)) + list(range(26, 32)) + [52]

df['IsHolidayWeek'] = df['Week'].apply(lambda x: 1 if x in holiday\_weeks else 0)

**GenreCount**

* **What it is:** Count of genres assigned (always 1 in this exploded dataset)
* **Why it matters:** Placeholder for future multi-genre improvements

df['GenreCount'] = 1

**Log\_Viewers (Transformed Target)**

* **What it is:** Log-transformed version of Viewers using np.log1p
* **Why it matters:** Reduces skew and improves learning

df['Log\_Viewers'] = np.log1p(df['Viewers'])

**One-Hot Encoded Categorical Features**

* Season → Winter, Spring, Summer, Autumn
* Genres\_List → Action, Drama, Comedy, etc.  
  Used to capture non-numeric relationships.

from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder(sparse=False, handle\_unknown='ignore')

encoded = encoder.fit\_transform(df[['Season', 'Genres\_List']])

**Modeling Pipeline**

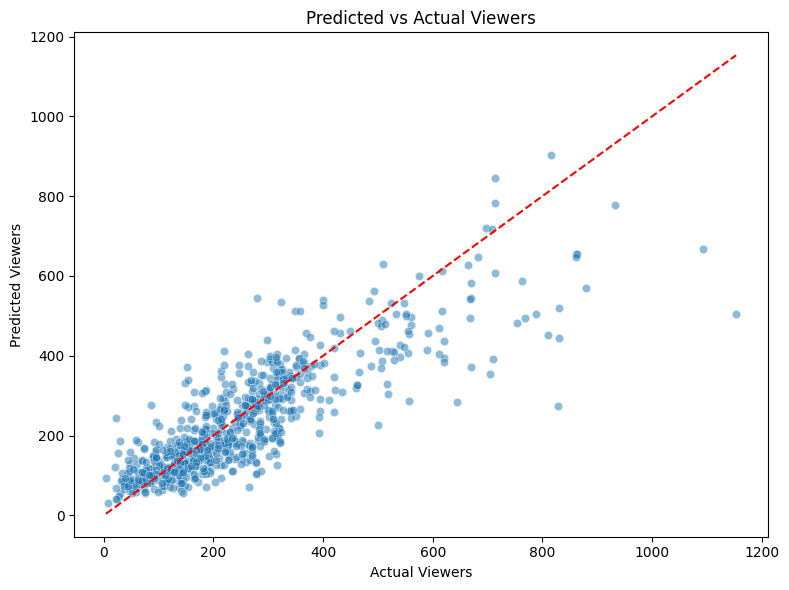
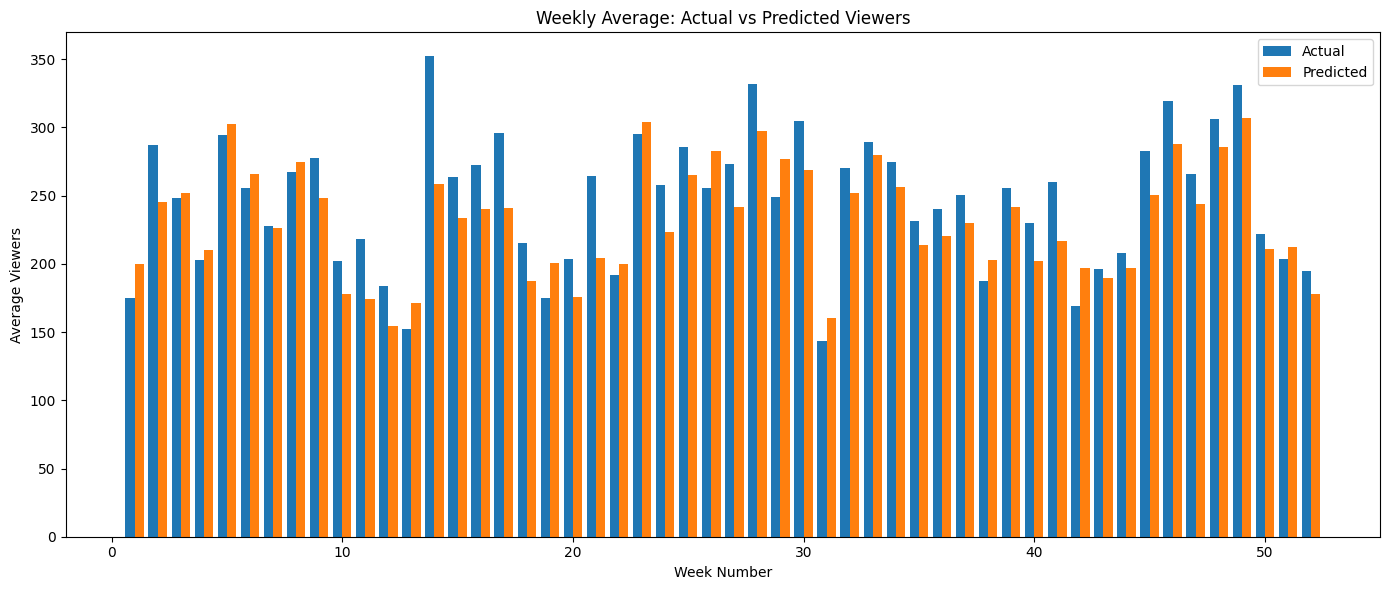
1. Features were engineered and encoded.
2. The target Viewers was log-transformed to improve learning.
3. A **Random Forest Regressor** was trained and evaluated.
4. Predictions were inverse-transformed using np.expm1().
5. Model was evaluated using MAE, RMSE, and R² metrics.

**Model Performance**

| **Metric** | **Value** |
| --- | --- |
| **MAE** | ~60.29 |
| **RMSE** | ~89.09 |
| **R²** | **0.7107** |

This indicates that the model explains over **71% of the variance** in weekly moviegoer numbers.

**Visual Results**

* **Scatter Plot** of Actual vs Predicted Viewers:  
  [](https://private-user-images.githubusercontent.com/202635447/446548125-ce227513-af6d-4229-87e6-a109a307a37b.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3NDg1MzExMDcsIm5iZiI6MTc0ODUzMDgwNywicGF0aCI6Ii8yMDI2MzU0NDcvNDQ2NTQ4MTI1LWNlMjI3NTEzLWFmNmQtNDIyOS04N2U2LWExMDlhMzA3YTM3Yi5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjUwNTI5JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI1MDUyOVQxNTAwMDdaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT1lY2EwNTVkODk4ODVlNWUzNTdlYjE1ZjhiNDFhMzhmMGRhOGZlNzUyMjczOWEwZGY2YzQyOTcyNDYyNzgzYThiJlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCJ9._5NtSrxzmwg52W4vmlbbOlkkM__TnbvKcJgn08lk8gQ)
* **Bar Chart** of Weekly Average (Actual vs Predicted):  
  [](https://private-user-images.githubusercontent.com/202635447/446548200-870aa296-2b64-460f-a2c8-9e201a9e0c8d.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3NDg1MzExMDcsIm5iZiI6MTc0ODUzMDgwNywicGF0aCI6Ii8yMDI2MzU0NDcvNDQ2NTQ4MjAwLTg3MGFhMjk2LTJiNjQtNDYwZi1hMmM4LTllMjAxYTllMGM4ZC5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjUwNTI5JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI1MDUyOVQxNTAwMDdaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT04Y2QwZWVjODRkYmM1ZjlhMzY1YzZlNjI2YzE4Zjg0NjU0Y2YxNGRlNmNiODc2NDlkOTAwMWYwZTBmOWI4NzExJlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCJ9.cl0jZNGLhUvMiihOJQJckif3HmF20_RAvnJEcLz80ls)

These visuals show that the model performs consistently across weeks and genres.

**CONCLUSION**

**Insights Gained from ML**

* Which features are most predictive? (Movie\_Rank, Genre, Season)
* How close are the model’s predictions to actual viewership?
* Can we use this for real-world scheduling or marketing?

The model adds a **predictive and applied layer** to the previously descriptive analysis.

**Future Work**

To improve and expand this project, the following steps could be studied:

* **Expand Dataset Scope**  
  Instead of Top 10 weekly movies, consider Top 20 or Top 50 to include more diversity and genre variability.
* **Apply Statistical Tests on Genre-Season Relationships**  
  While Hypothesis 2 was supported visually, a chi-square test or regression interaction term could strengthen this finding.
* **Integrate Streaming Data**  
  Analyze how seasonal patterns differ between cinema and digital platforms like Netflix or BluTV.
* **Use Official Holiday Calendars**  
  Improve the IsHolidayWeek feature by incorporating actual national holiday dates rather than manual week tagging.
* **Experiment with Other ML Models**  
  Try advanced ML models for comparison and possibly higher predictive performance.
* **Deploy as Interactive Tool**  
  Build a dashboard where users can input a movie’s genre and release week to see expected attendance.