Explore IMDb Movies

Team: Nuggests-Forever

2024-03-03

Outline

- Introduction
- Data description
- Results from EDA and modeling
- Interactive Visualization
- Q&A

Introduction

- Our goal is to uncover insights into the evolving landscape of film and television ratings over time
- Specifically, we are interested in high-rated directors' professions and genres preferences
- To predict ratings based on given features
  - XG boost
- To visualize trends change interactively
  - Observable notebook

Description Of The Data

- IMDb is an online dataset updated daily, containing information scraped since the beginning of this quarter.
- Link: https://developer.imdb.com/non-commercial-datasets/
- The dataset consists of seven different compressed packaged:
  - title.akas.tsv.gz
  - title.basics.tsv.gz
  - title.crew.tsv.gz
  - title.episode.tsv.gz
  - title.principals.tsv.gz
  - title.ratings.tsv.gz
  - name.basics.tsv.gz

Results

Top 10 directors by avg_rating and num_movies

```
# A tibble: 10 x 4
   directors primaryName          num_movies avg_rating
   <chr>     <chr>                     <int>      <dbl>
 1 nm1337210 Chuck O'Neil               2510       6.59
 2 nm3766090 Doug Walker                1199       6.92
 3 nm0123273 James Burrows               957       7.46
 4 nm3005544 James Rolfe                 796       7.09
 5 nm0669120 Michael Pearlman            750       7.92
 6 nm9072464 Luke Lerdwichagul           741       7.12
 7 nm1206693 Dave Diomedi                721       6.74
 8 nm0635620 Russell Norman              693       6.94
 9 nm1347153 Tyler Perry                 689       6.94
10 nm1121649 Brad Jones                  681       7.89
```

- Average rating for movies is 6.8981331.

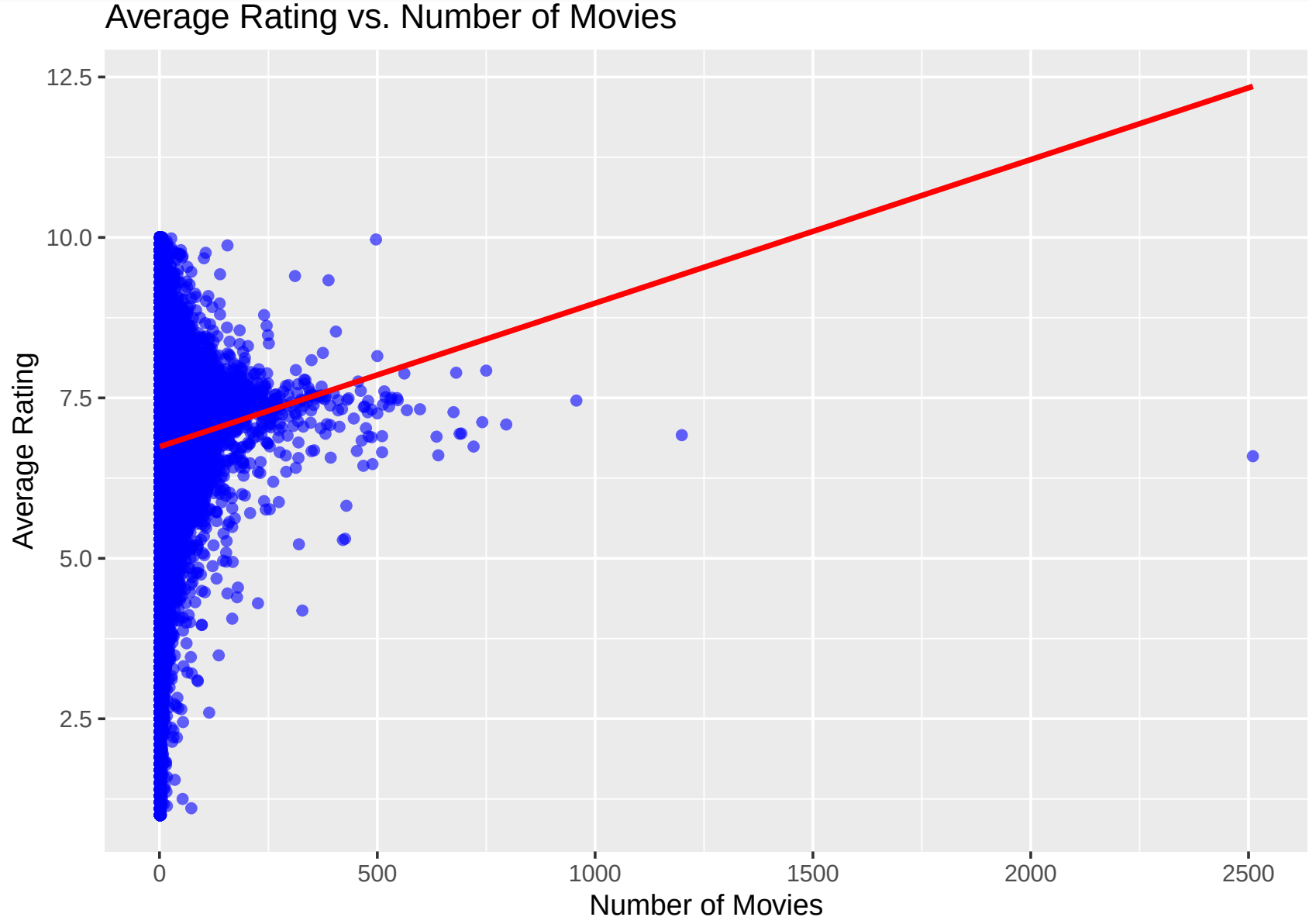
Correlation between avg_rating and num_movies:

- Strong but Limited

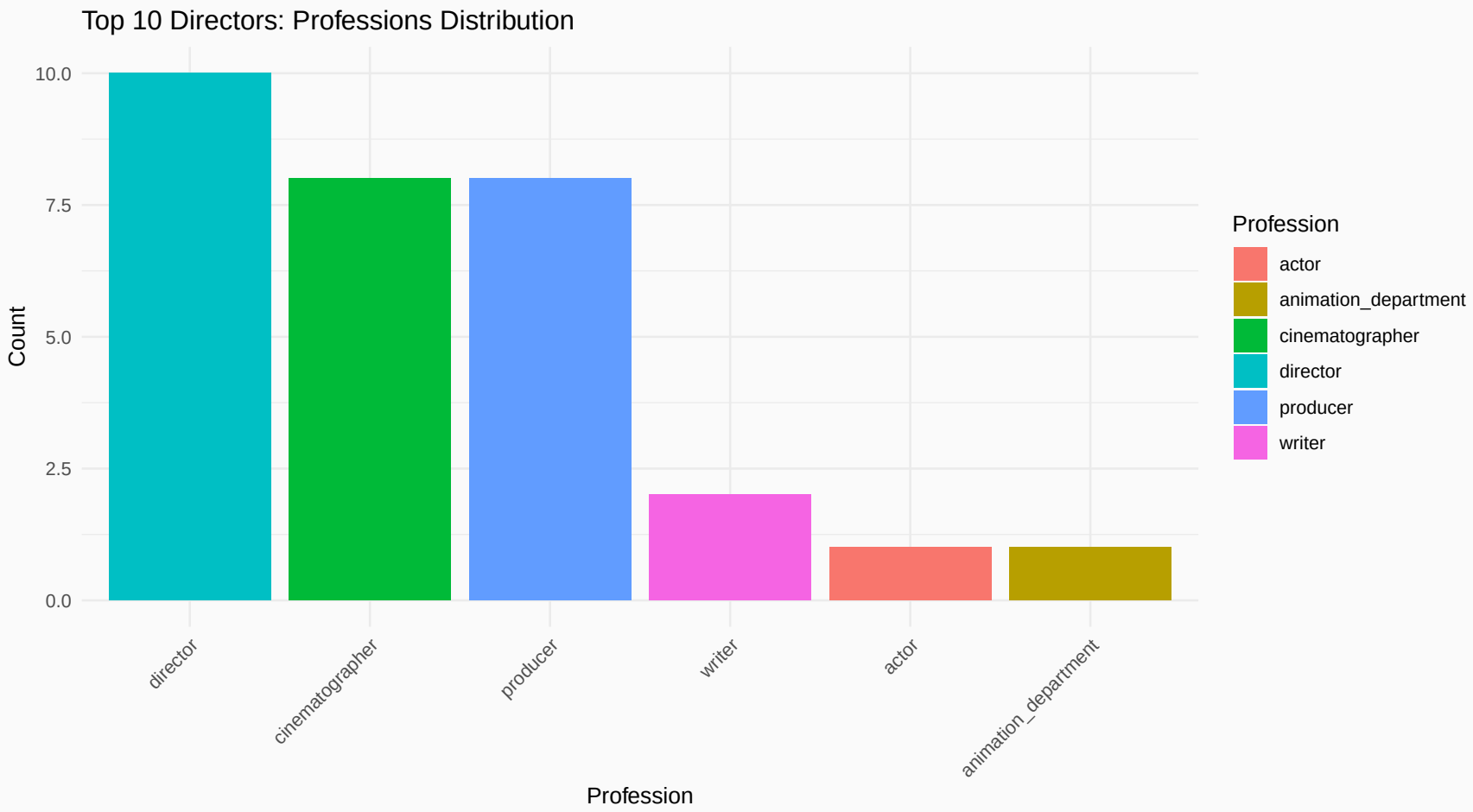
	Coefficient	P_Value	R_Squared
(Intercept)	6.738804250	0.000000e+00	0.0008297688
num_movies	0.002237122	3.869807e-38	0.0008297688

- Coefficient: 0.0022
- Both coefficients highly statistically significant
- Limited explanatory power: R-squared $= 0.0008$

Scatterplot for Correlation



Top 10 directors' professions distribution



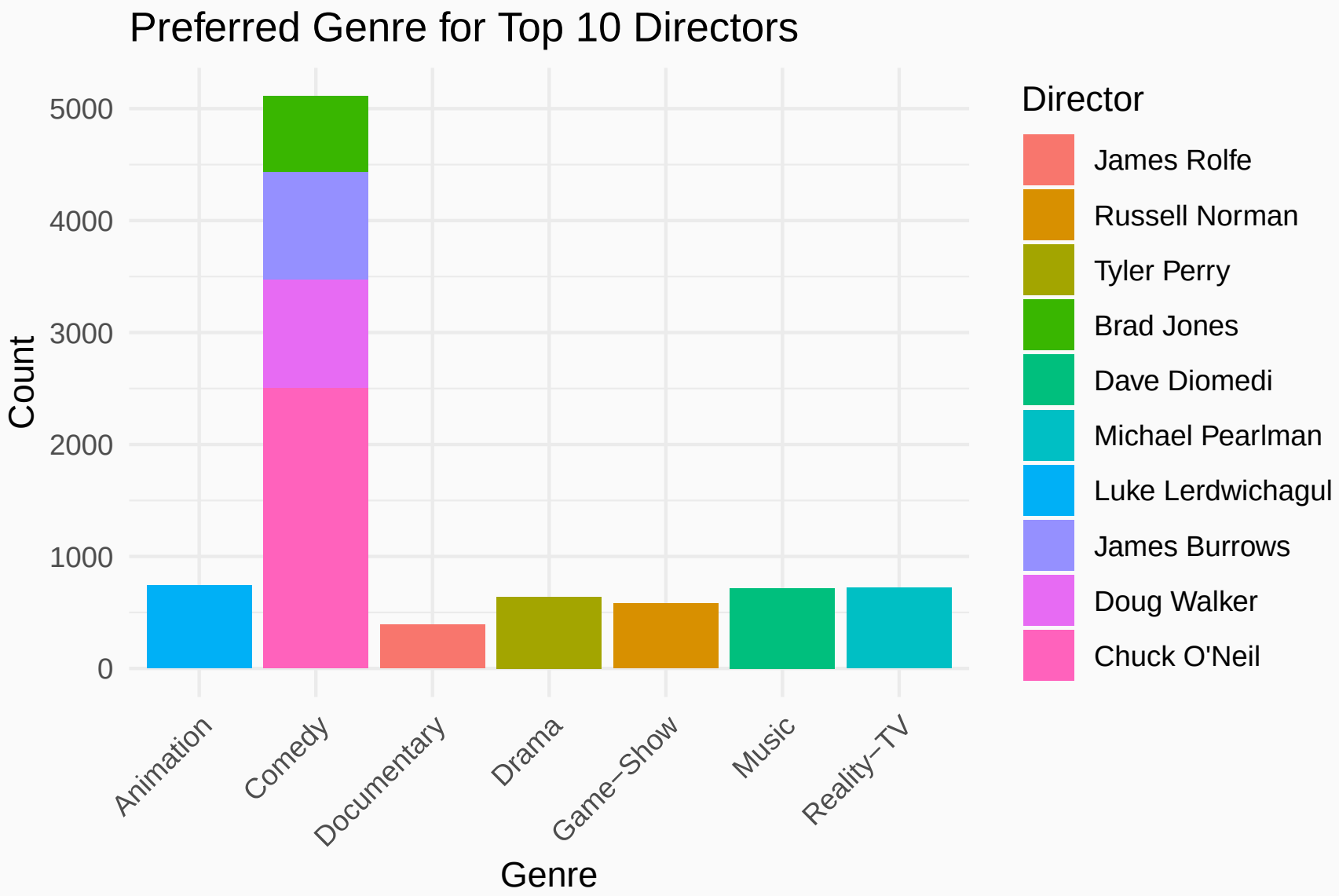
Percentage of directors among the top 10 directors is approximately 33.3%.

Comparsion with all directors

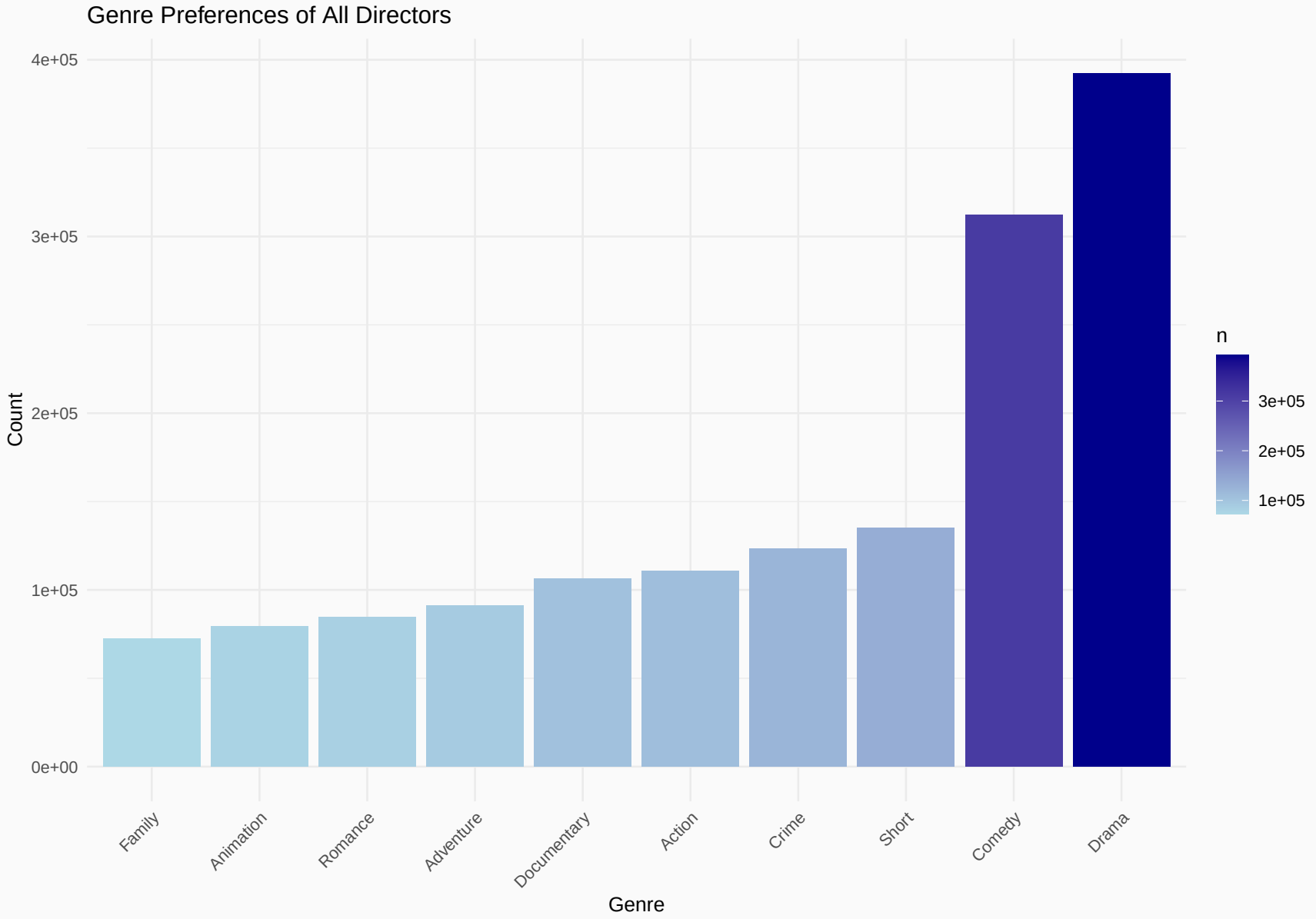
- 5000 randomly draw sample from all directors

primaryProfession	n	ratio_top_10	ratio_sampled
actor	421	0.1600152	0.2012
animation_department	14	0.0053212	0.0414
cinematographer	442	0.1679970	0.0802
director	877	0.3333333	0.9442
producer	853	0.3242113	0.4932
writer	24	0.0091220	0.5356

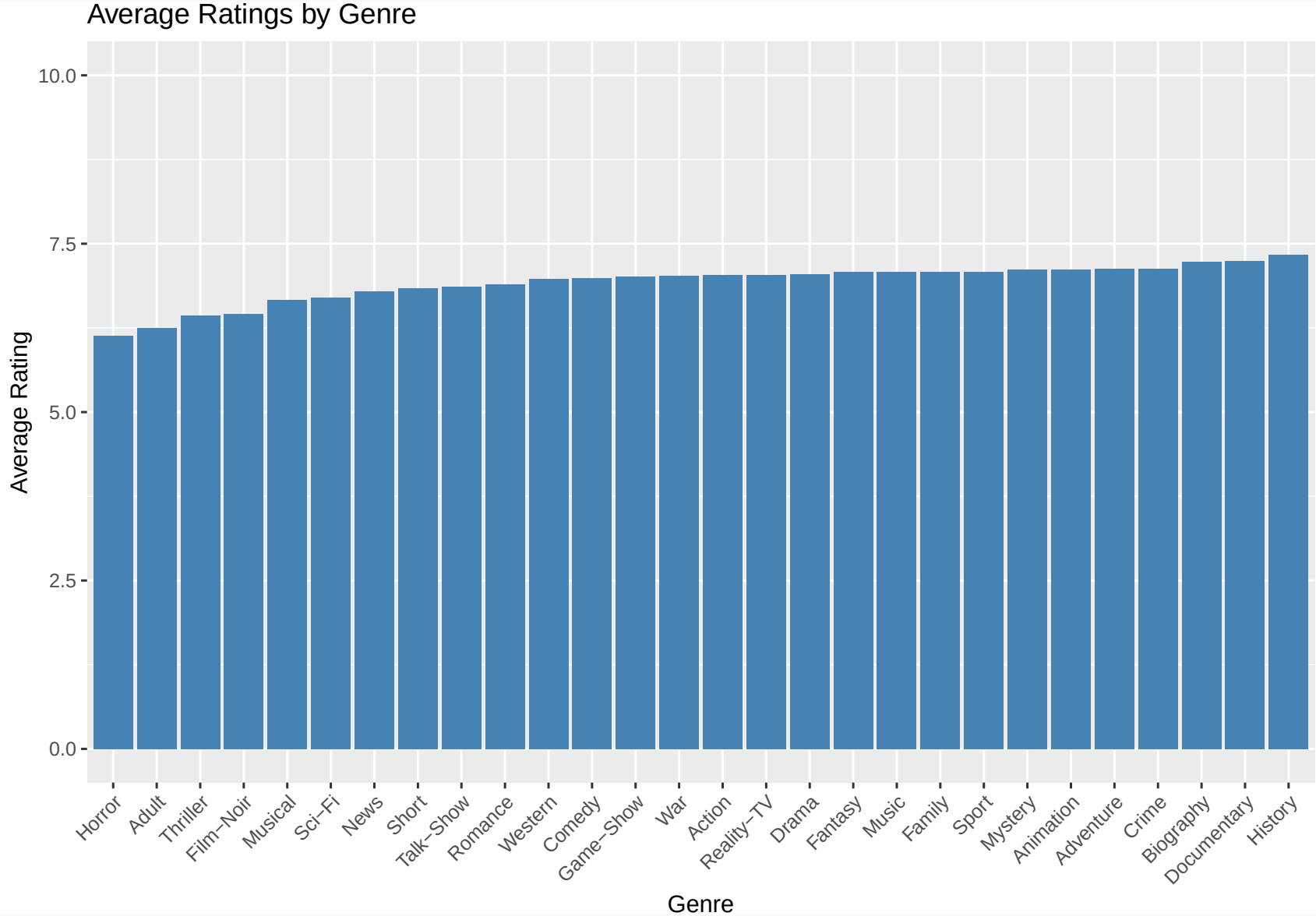
What are the favored genres among top 10 directors?



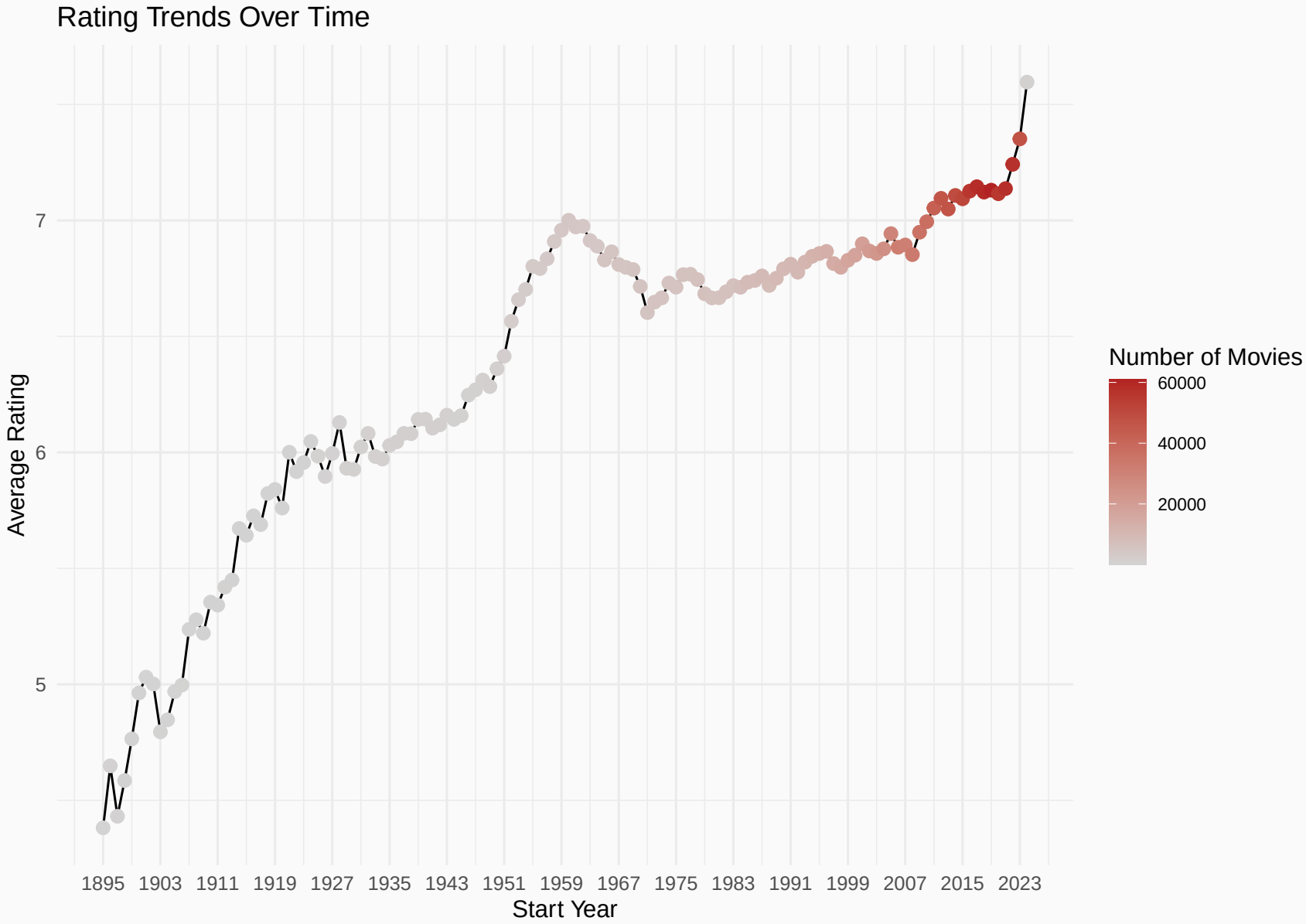
Genre Perference among all directors



IMDB Average Ratings By Genre



IMDB Average Ratings Trends Over Time



IMDB Ratings Prediction Model(Machine Learning)

Xgboost Model introduction:

- Library XGBoost:
  - XGBoost Overview:
    - Stands for eXtreme Gradient Boosting.
    - Widely used and acclaimed for its efficiency and accuracy.
  - Gradient Boosting Explained:
    - A technique that builds models as an ensemble of weak predictors.
    - Typically employs decision trees as the base learners.
  - gbtree Booster:
    - Utilizes tree-based models for prediction and learning.
    - Ideal for capturing complex non-linear relationships in data.

Library data.table

- data.table
  - Handles large datasets more efficiently than base R's data.frame.
  - Provides fast reading and writing of data, significantly reducing data processing time.

Train the model

- Convert character columns to numeric codes
- Randomly split data into training and testing sets
- Set xgboost parameters(eg. eta, max_depth)
- Train the model and make predictions

Calculate & Print Model Performance

```r
# Calculate performance metrics
MAE <- mean(abs(predictions-test_data$averageRating))
RMSE <- sqrt(mean((predictions-test_data$averageRating)^2))
# Print performance metrics
# MAE (Mean Absolute Error)
# RMSE (Root Mean Square Error)
print(paste("MAE:", MAE))
```

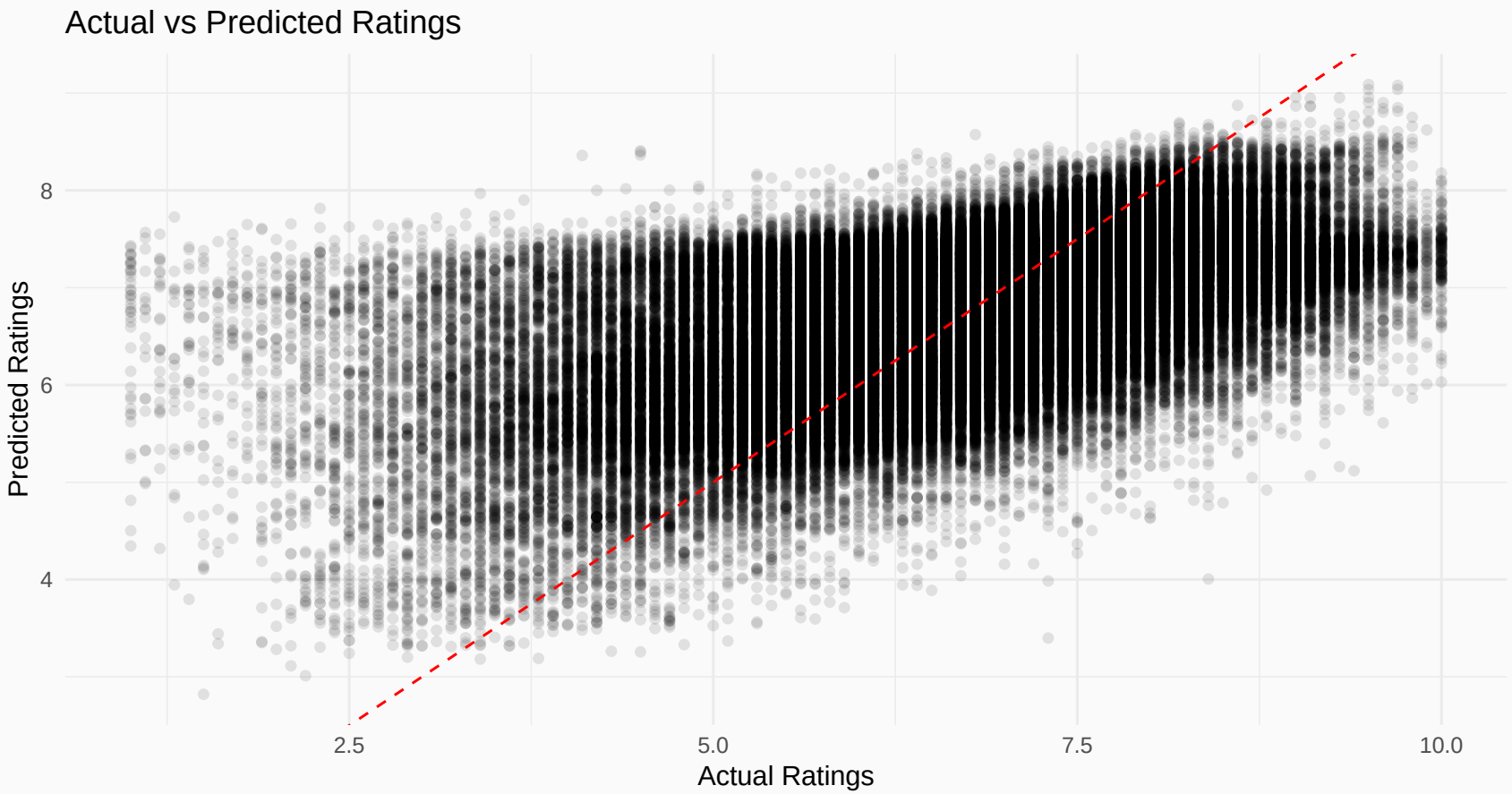
```
[1] "MAE: 0.783337644926456"
```

```r
print(paste("RMSE:", RMSE))
```

```
[1] "RMSE: 1.05568471874053"
```

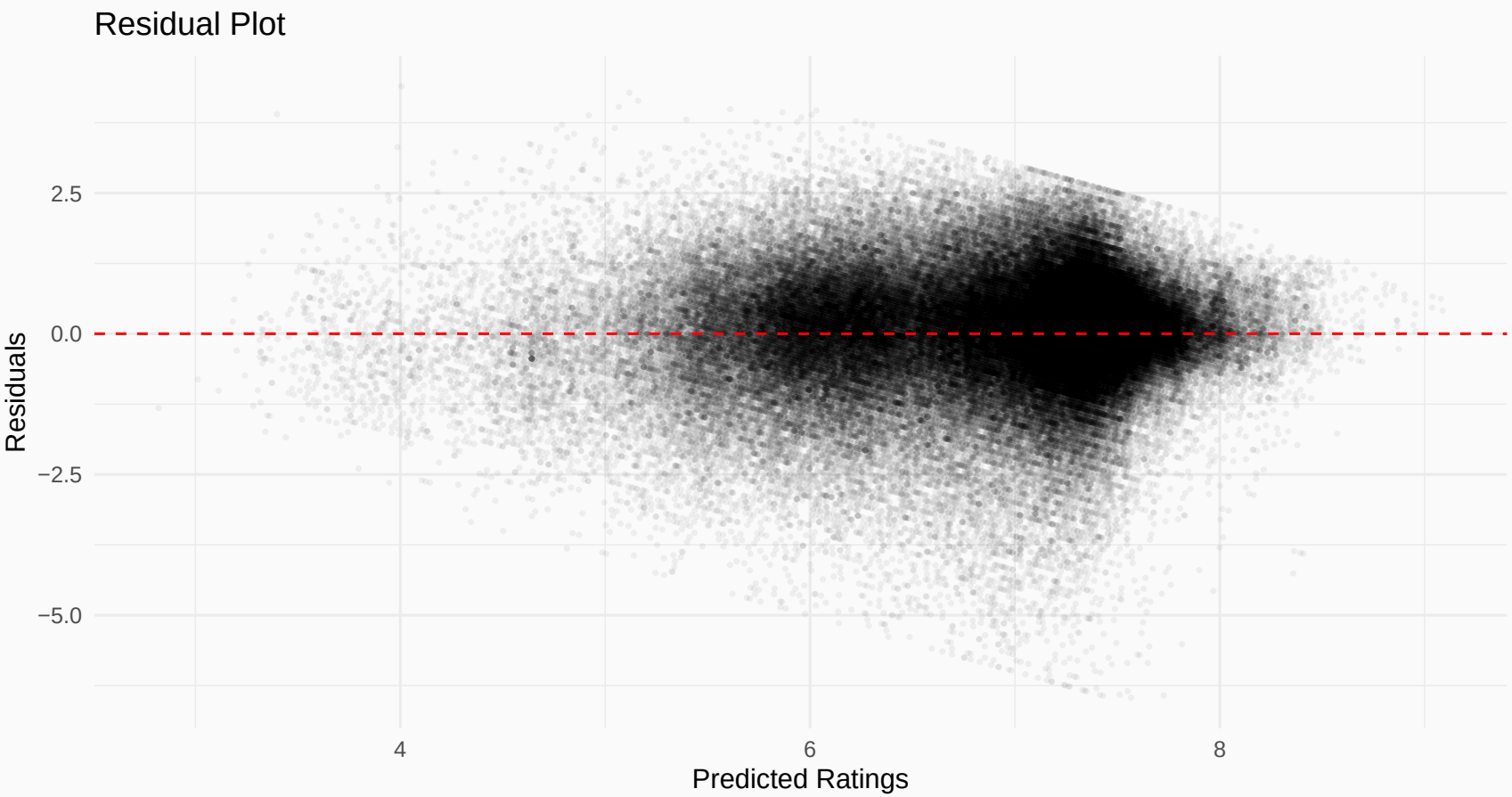
Scatter Plot of Actual vs. Predicted Ratings

- Give us a visual sense of how predicted values compare to the actual values.



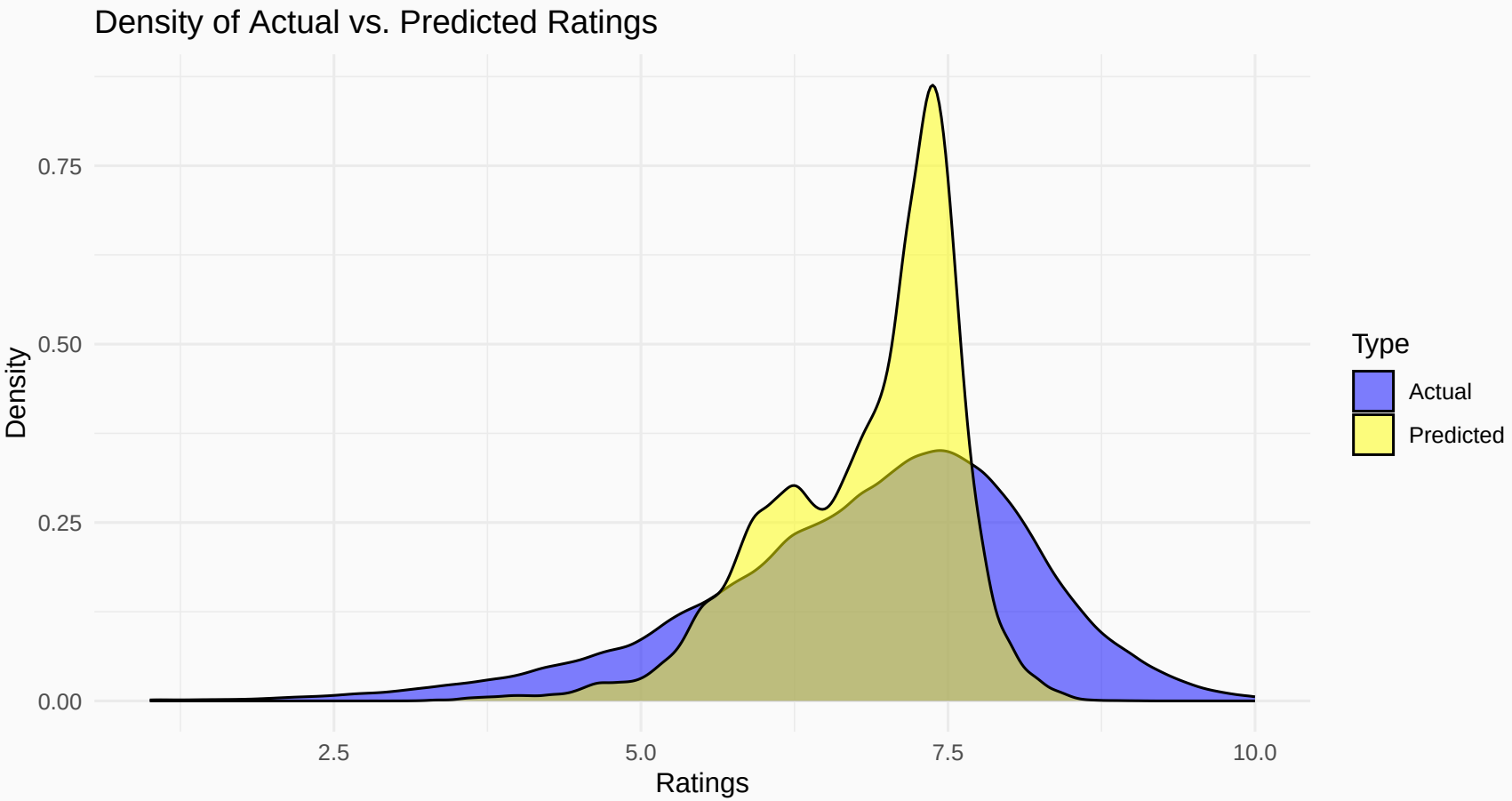
Residual Plot:

- Identify patterns in the errors made by the model.



Density Plot for Actual vs. Predicted Ratings

- Distribution of predicted ratings compares to the actual ratings:



Interactive Visualizations

Click here to visit Example Website

Possible Next Steps

- If We Had More Time:
  - Be more cautious and consistent dealing with data cleaning, reasons for dropping NA.
  - Explore more applicable models for deeper insights and accurate predictions.

Any Question?

Thanks for listening!