# From Career Nightmares to Cinematic Scares: A Data-Driven Horror Movie Strategy

Angelo Filiol de Raimond

2025-05-25

## Table of contents



This Quarto document supports the academic report *From Career Nightmares to Cinematic Scares*, providing data exploration, model-based insights, and strategic recommendations for horror movie production.

# 1 Data Preparation and Market Overview

## 1.1 Data Import and Cleaning

```
library(tidyverse)
library(tidytuesdayR)
library(stringr)
library(forcats)

tuesdata <- tt_load('2024-10-29')
monster_movies <- tuesdata$monster_movies

horror_data <- monster_movies %>%
  filter(str_detect(genres, "Horror")) %>%
  filter(!is.na(num_votes), !is.na(average_rating), !is.na(runtime_minutes)) %>%
  distinct(primary_title, .keep_all = TRUE)
```

## 1.2 Variable Construction

```
horror_data <- horror_data %>%
  mutate(
    sub_genre = str_remove(genres, "Horror,?\\s*"),
    sub_genre = ifelse(sub_genre == "", "Horror", sub_genre),
    duration_group = case_when(
      runtime_minutes < 60 ~ "<60",
      runtime_minutes < 90 ~ "60-89",
      runtime_minutes < 120 ~ "90-119",
      TRUE ~ "120+"
    ),
    log_votes = log1p(num_votes)
  )
```

## 1.3 Key Measures and Variable Roles

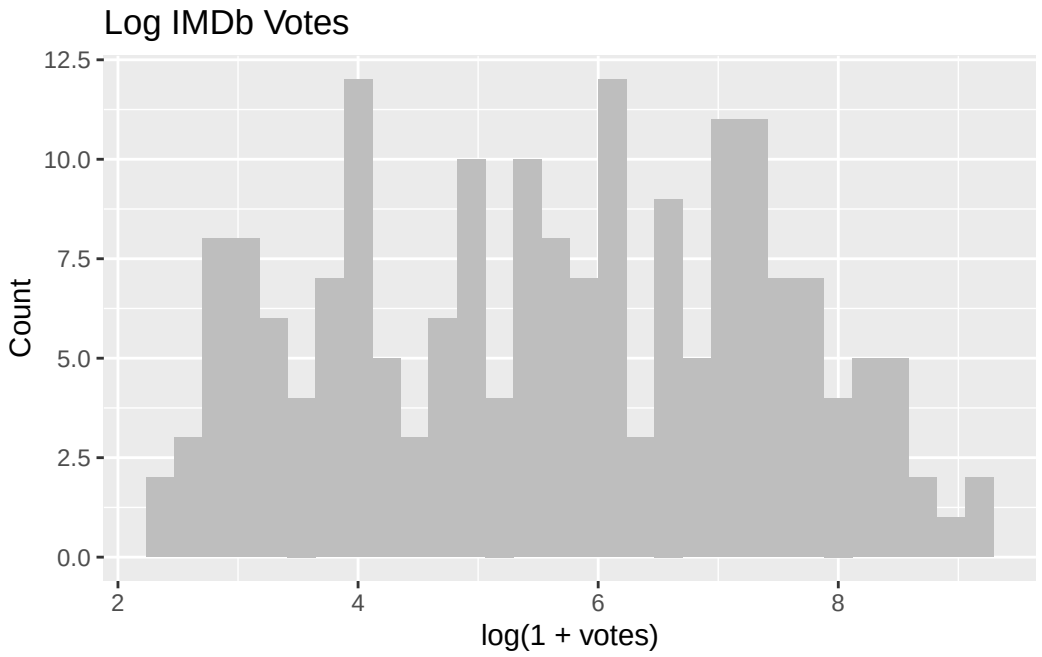
- **log_votes**: proxy for popularity/exposure (performance measure).
- **average_rating**: indicator of perceived quality.
- **Strategy variables**: `sub_genre`, `duration_group`, `runtime_minutes`.
- **Control variable**: `average_rating`.

## 1.4 Summary Statistics

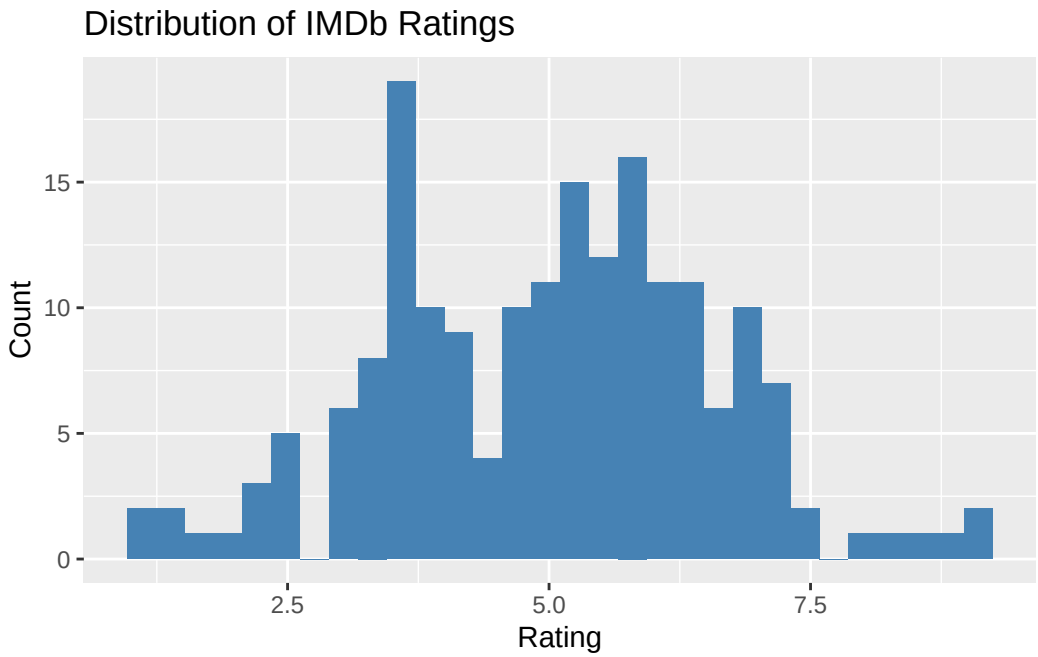
```
 average_rating    num_votes        runtime_minutes
 Min.   :1.200   Min.   :   10.0   Min.   :  3.00
 1st Qu.:3.750   1st Qu.:   55.5   1st Qu.: 73.00
 Median :5.200   Median :  289.0   Median : 83.00
 Mean   :5.006   Mean   :  985.0   Mean   : 82.78
 3rd Qu.:6.150   3rd Qu.: 1168.0   3rd Qu.: 91.50
 Max.   :9.200   Max.   :10106.0   Max.   :295.00
```

## 1.5 Distribution Visuals

```
ggplot(horror_data, aes(x = log_votes)) +
  geom_histogram(bins = 30, fill = "gray") +
  labs(title = "Log IMDb Votes", x = "log(1 + votes)", y = "Count")
```

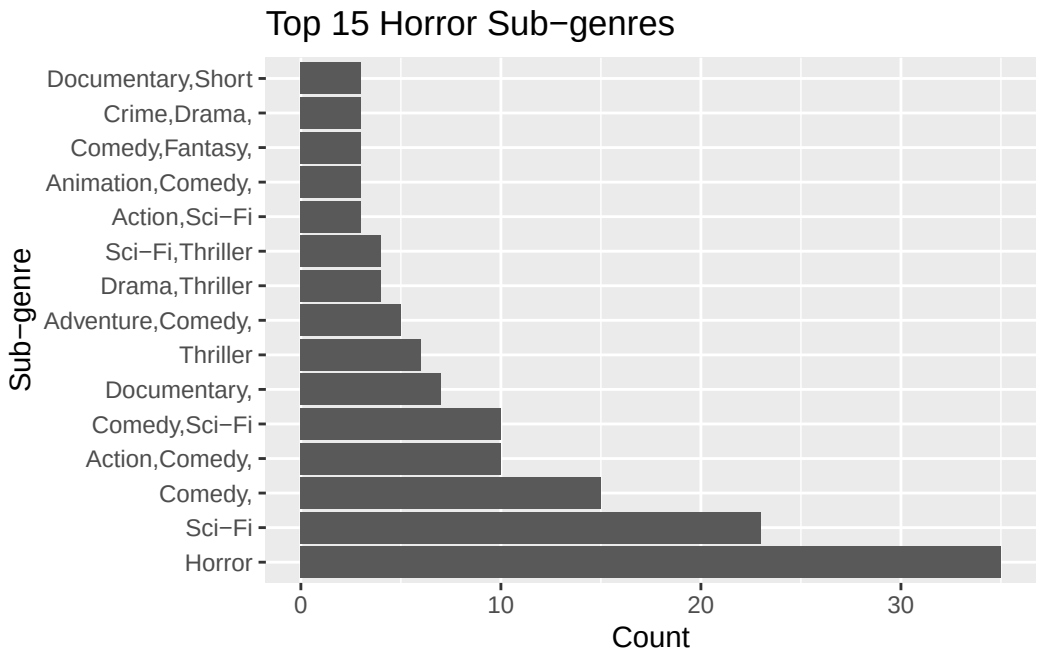



```
ggplot(horror_data, aes(x = average_rating)) +
  geom_histogram(bins = 30, fill = "steelblue") +
  labs(title = "Distribution of IMDb Ratings", x = "Rating", y = "Count")
```




```
top_subs <- horror_data %>%
  count(sub_genre, sort = TRUE) %>%
  slice_head(n = 15) %>%
  pull(sub_genre)

horror_data %>%
  filter(sub_genre %in% top_subs) %>%
  ggplot(aes(x = fct_infreq(sub_genre))) +
  geom_bar() +
  coord_flip() +
  labs(title = "Top 15 Horror Sub-genres", x = "Sub-genre", y = "Count")
```



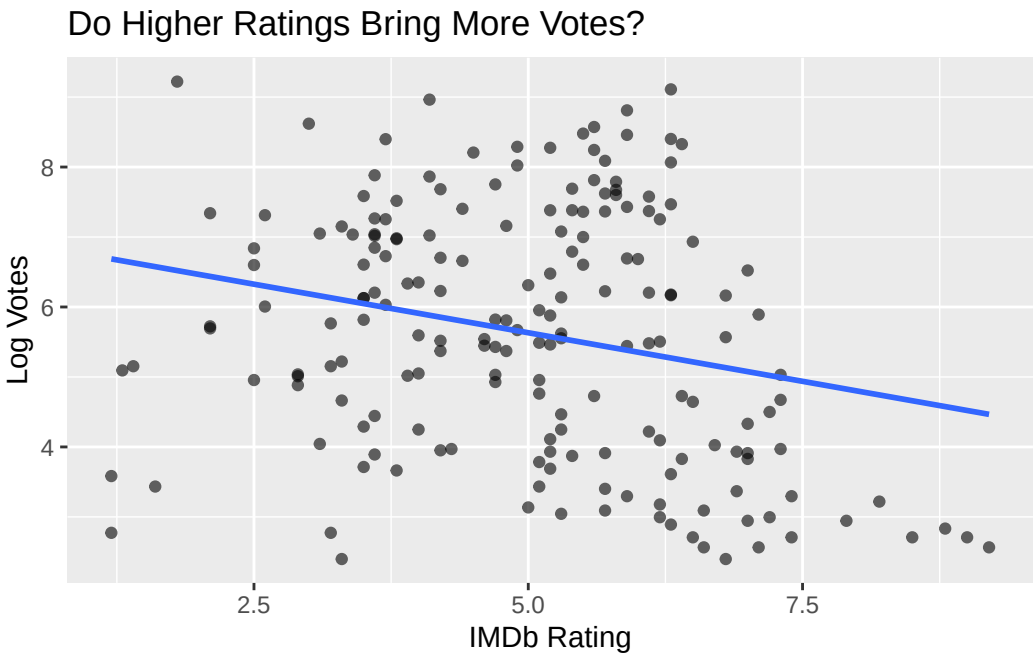

## 2 Analysis of Success Factors

### 2.1 Ratings and Popularity

```
cor(horror_data$average_rating, horror_data$log_votes)
```

```
[1] -0.2472498
```

```
ggplot(horror_data, aes(x = average_rating, y = log_votes)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Do Higher Ratings Bring More Votes?", x = "IMDb Rating", y = "Log Votes")
```




## 2.2 Linear Model

```
model <- lm(log_votes ~ average_rating + runtime_minutes + sub_genre + duration_group, dat

library(broom)
library(knitr)

tidy(model) %>%
  kable(
    digits = 3,
    caption = "Linear Model Coefficients",
    booktabs = TRUE
  )
```

Table 1: Linear Model Coefficients

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 4.429 | 1.423 | 3.114 | 0.002 |
| average_rating | -0.024 | 0.089 | -0.265 | 0.791 |
| runtime_minutes | 0.017 | 0.010 | 1.634 | 0.105 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| sub_genreAction,Adventure, | 0.666 | 1.689 | 0.394 | 0.694 |
| sub_genreAction,Animation, | -0.622 | 1.904 | -0.326 | 0.745 |
| sub_genreAction,Comedy, | 0.828 | 1.195 | 0.693 | 0.490 |
| sub_genreAction,Mystery | 0.823 | 1.875 | 0.439 | 0.661 |
| sub_genreAction,Sci-Fi | 2.177 | 1.393 | 1.562 | 0.121 |
| sub_genreAction,Thriller | -1.351 | 1.888 | -0.716 | 0.475 |
| sub_genreAdventure, | 0.542 | 1.879 | 0.288 | 0.774 |
| sub_genreAdventure,Comedy, | 1.746 | 1.308 | 1.335 | 0.184 |
| sub_genreAdventure,Drama, | 0.236 | 1.864 | 0.126 | 0.900 |
| sub_genreAdventure,Mystery | 1.368 | 1.533 | 0.892 | 0.374 |
| sub_genreAdventure,Sci-Fi | 1.792 | 1.520 | 1.179 | 0.241 |
| sub_genreAnimation, | -1.091 | 1.984 | -0.550 | 0.583 |
| sub_genreAnimation,Comedy, | -1.423 | 1.433 | -0.993 | 0.323 |
| sub_genreBiography, | 1.900 | 1.864 | 1.020 | 0.310 |
| sub_genreComedy, | -0.840 | 1.160 | -0.724 | 0.470 |
| sub_genreComedy,Crime, | -2.696 | 1.908 | -1.413 | 0.160 |
| sub_genreComedy,Drama, | 0.289 | 1.526 | 0.189 | 0.850 |
| sub_genreComedy,Family, | 2.082 | 1.536 | 1.356 | 0.178 |
| sub_genreComedy,Fantasy, | 0.165 | 1.398 | 0.118 | 0.906 |
| sub_genreComedy,Musical | 1.838 | 1.528 | 1.203 | 0.231 |
| sub_genreComedy,Mystery | 1.762 | 1.872 | 0.941 | 0.349 |
| sub_genreComedy,Romance | 1.748 | 1.901 | 0.919 | 0.360 |
| sub_genreComedy,Sci-Fi | 1.130 | 1.192 | 0.948 | 0.345 |
| sub_genreCrime,Documentary, | 1.855 | 1.865 | 0.995 | 0.322 |
| sub_genreCrime,Drama, | 1.484 | 1.401 | 1.059 | 0.291 |
| sub_genreCrime,Mystery | 0.070 | 1.863 | 0.038 | 0.970 |
| sub_genreCrime,Sci-Fi | 1.867 | 1.872 | 0.997 | 0.321 |
| sub_genreDocumentary, | -1.183 | 1.268 | -0.933 | 0.353 |
| sub_genreDocumentary,Drama, | -0.427 | 1.859 | -0.230 | 0.819 |
| sub_genreDocumentary,Short | 0.447 | 1.560 | 0.286 | 0.775 |
| sub_genreDrama, | 2.076 | 1.895 | 1.095 | 0.275 |
| sub_genreDrama,Fantasy, | 1.650 | 1.529 | 1.079 | 0.283 |
| sub_genreDrama,Mystery | 1.834 | 1.537 | 1.193 | 0.235 |
| sub_genreDrama,Romance | 1.683 | 1.407 | 1.197 | 0.234 |
| sub_genreDrama,Sci-Fi | 0.382 | 1.535 | 0.249 | 0.804 |
| sub_genreDrama,Thriller | 1.020 | 1.325 | 0.770 | 0.443 |
| sub_genreDrama,Western | 1.243 | 1.859 | 0.669 | 0.505 |
| sub_genreFamily, | -2.310 | 1.887 | -1.225 | 0.223 |
| sub_genreFantasy, | 2.320 | 1.891 | 1.227 | 0.222 |
| sub_genreFantasy,Sci-Fi | 1.624 | 1.859 | 0.874 | 0.384 |
| sub_genreHorror | -0.437 | 1.113 | -0.393 | 0.695 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| sub_genreMusic | 1.902 | 1.860 | 1.023 | 0.308 |
| sub_genreMusic,Short | -1.747 | 2.045 | -0.854 | 0.394 |
| sub_genreMystery | 0.895 | 1.555 | 0.575 | 0.566 |
| sub_genreMystery,Sci-Fi | 1.157 | 1.519 | 0.762 | 0.448 |
| sub_genreMystery,Thriller | 0.815 | 1.427 | 0.571 | 0.569 |
| sub_genreSci-Fi | 1.726 | 1.133 | 1.524 | 0.130 |
| sub_genreSci-Fi,Thriller | 1.785 | 1.318 | 1.355 | 0.178 |
| sub_genreSci-Fi,Western | 0.833 | 1.869 | 0.445 | 0.657 |
| sub_genreShort | -2.022 | 2.028 | -0.997 | 0.321 |
| sub_genreThriller | -0.029 | 1.245 | -0.023 | 0.981 |
| duration_group120+ | -4.104 | 1.608 | -2.553 | 0.012 |
| duration_group60–89 | -0.235 | 0.756 | -0.312 | 0.756 |
| duration_group90–119 | -1.136 | 0.915 | -1.241 | 0.217 |

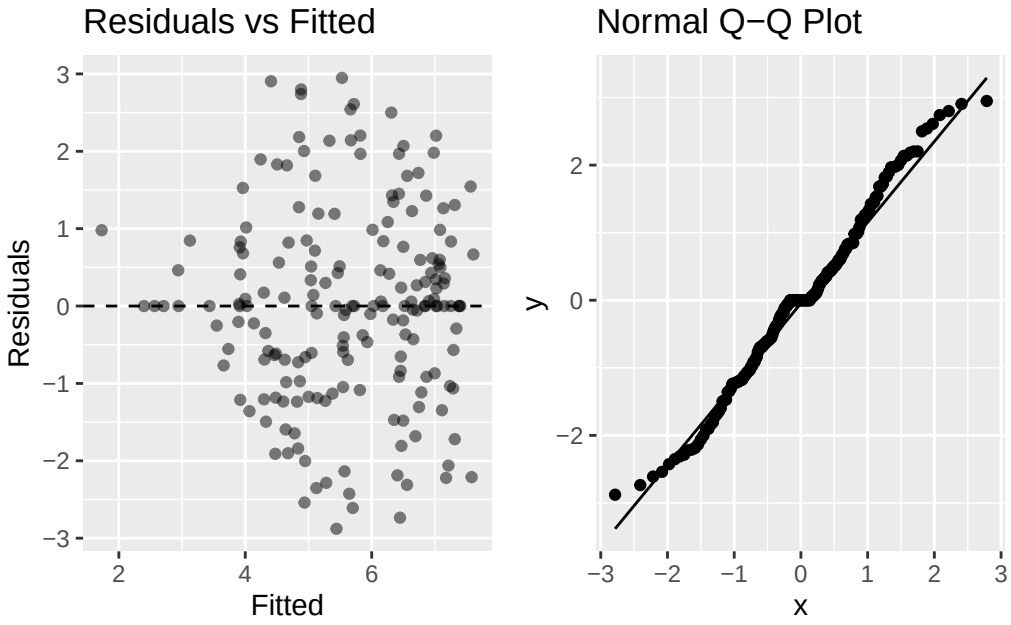
## 2.3 Residual Diagnostics

```
library(patchwork)

p1 <- ggplot(data.frame(fitted = model$fitted.values, resid = model$residuals),
             aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residuals")

p2 <- ggplot(data.frame(resid = model$residuals), aes(sample = resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal Q-Q Plot")

p1 + p2
```




### 2.4 Discussion

- Model explains part of variance, but residuals show slight skewness.
- IMDb rating is not a strong predictor of exposure.
- Sub-genres and duration windows carry most signal.
- Cannot claim causality due to unmeasured confounders and observational design.

## 3 Strategic Recommendation

### 3.1 Recap of Key Findings

| Factor | Takeaway |
|---|---|
| Sub-genre | Comedy/Sci-Fi hybrid horror = more exposure |
| Duration group | 60–89 minutes outperform others |
| Ratings | Do not drive vote count |
| Runtime | Mid-length films preferred |

### 3.2 Recommended Strategy

| Dimension | Recommendation |
|---|---|
| Sub-genre | Horror-Comedy or Horror-Sci-Fi |
| Runtime | 75–85 minutes |
| Format | Designed for streaming (Netflix, etc.) |
| Budgeting | Micro-budget (<500k) + viral marketing |
| Story tone | Workplace horror / dark satire |

### 3.3 Limitations

- IMDb votes = exposure proxy, not financial success.
- No data on distribution, marketing, or cast.
- Confounding unobserved variables may affect model.
- No clear time trend or platform-specific insights.

### 3.4 Future Data and Analysis

To improve predictions and deepen insight, future models should include:

- Budget and revenue breakdowns (to assess ROI)
- Textual sentiment analysis on reviews
- Platform and release date info
- Studio, cast, and marketing spend
- Experimental or longitudinal designs

### 3.5 Final Thoughts

This analysis offers a data-driven entry point into horror film strategy. While not predictive of box office alone, it highlights patterns of visibility and audience attention worth leveraging for a breakout indie success.

---