Analysis of Disney Movies

Edward Eckle

The Magic of the Box Office: A Decades-Long Analysis of Disney Movies

A Data Analysis - By Edward Eckle

Since the release of * Snow White and the Seven Dwarfs * in 1937, The Walt Disney Studios has transitioned from a risky experiment in feature length animation to a dominant global media powerhouse. Over nearly a century, Disney has defined and redefined family entertainment, navigating through the "Golden Age," the "Renaissance" of the 90s, and the modern era of blockbuster acquisitions.

However, the film industry is not just built on magic, it is built on data. Understanding how different genres, MPAA ratings, and release cycles affect the bottom line is crucial for understanding Disney's longevity.

The goal of this analysis is to explore the commercial trajectory of Disney's film catalog. By utilizing a data set of 579 movies released between 1937 and 2016, this project aims to:

- **Visualize the growth of output:** Track how the volume of Disney releases has shifted across the decades.
- **Identify Revenue Drivers:** Determine which genres (e.g., Musical vs. Adventure) have historically yielded the highest returns.
- **Analyze the "Inflation Factor":** Compare raw total gross against inflation-adjusted figures to see how the "classics" stack up against modern-day blockbusters like the Marvel Cinematic Universe.
- **Examine Demographic Trends:** Analyze how MPAA ratings correlate with box office success.

(Limitations: The data set is made using publicly available information, and not all conclusions may be accurate.)

---

1. Setup and Data Loading, Feel free to skip to section 2

Inclusion of needed libraries

```
library(tinytex)
library(patchwork)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.2     v tibble    3.3.1
## v lubridate 1.9.4     v tidyr     1.3.2
## v purrr     1.2.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(collapse)
```

```
## collapse 2.1.6, see ?‘collapse-package‘ or ?‘collapse-documentation‘
##
## Attaching package: 'collapse'
##
## The following object is masked from 'package:tidyr':
##
##     replace_na
##
## The following object is masked from 'package:stats':
##
##     D
```

```
library(ggthemes)
library(ggeffects)
```

Loading our data set
The data set used in this project contains the following key values:

`move_title` The title of the film
`release_date` The date the film debuted in theaters.
`genre` The thematic category of the movie
`mpaa_rating` The age-appropriateness rating.
`total_gross` Raw domestic box office earnings.
`inflation_adjusted_gross` The earnings adjusted to modern dollar values.

```r
df_raw <- read.csv("./disney_movies.csv")
head(df_raw)
```

```
##                        movie_title release_date     genre mpaa_rating
## 1 Snow White and the Seven Dwarfs   1937-12-21   Musical           G
## 2                       Pinocchio   1940-02-09 Adventure           G
## 3                        Fantasia   1940-11-13   Musical           G
## 4               Song of the South   1946-11-12 Adventure           G
## 5                      Cinderella   1950-02-15     Drama           G
## 6    20,000 Leagues Under the Sea   1954-12-23 Adventure            
##   total_gross inflation_adjusted_gross
## 1   184925485               5228953251
## 2    84300000               2188229052
## 3    83320000               2187090808
## 4    65000000               1078510579
## 5    85000000                920608730
## 6    28200000                528279994
```

Data Preparation & Feature Engineering

To unlock the temporal insights hidden within the data set, our first step is to transform the `release_date` column into a more versatile format. By converting this field into a formal `Date` object, we can derive new features, specifically **Year**, **Month**, and **Decade**.

```r
df_clean <- df_raw |>
  filter(!is.na(release_date)) |>
  mutate(release_date = ymd(release_date)) |>
  mutate(release_year = year(release_date),
         release_month = month(release_date),
         decade = floor(release_year / 10) * 10
  ) |>
  relocate(c(release_year, release_month), .after=release_date) |>
  relocate(decade, .before=release_date)

# Check which were missing. me from the future its none :)
missing_dates <- df_raw |>
  filter(is.na(release_date))

head(missing_dates)
```

```
## [1] movie_title              release_date             genre
## [4] mpaa_rating              total_gross              inflation_adjusted_gross
## <0 rows> (or 0-length row.names)
```

Handling Missing Categorical Data

To ensure our visualizations are clean and that no data is excluded due to missing labels, we need to address the ***NA*** values in the categorical columns. Specifically, the `genre` and `mpaa_rating` columns contain several missing entries. Rather than dropping these rows and losing the associated financial data, we will encode them to “Unknown” and “Not Rated.”

```r
df_clean <- df_clean |>
  mutate(
    # Catch both NA values and empty strings
    genre = case_when(
      is.na(genre) | genre == "" ~ "Unknown",
      TRUE ~ genre
    ),
    mpaa_rating = case_when(
      is.na(mpaa_rating) | mpaa_rating == "" ~ "Not Rated",
      TRUE ~ mpaa_rating
    ),
    # Convert to factors inside the same mutate call
    genre = as.factor(genre),
    mpaa_rating = as.factor(mpaa_rating)
  )

head(df_clean)
```

```
##                        movie_title decade release_date release_year
## 1 Snow White and the Seven Dwarfs   1930   1937-12-21         1937
## 2                       Pinocchio   1940   1940-02-09         1940
## 3                        Fantasia   1940   1940-11-13         1940
## 4               Song of the South   1940   1946-11-12         1946
## 5                      Cinderella   1950   1950-02-15         1950
## 6    20,000 Leagues Under the Sea   1950   1954-12-23         1954
##   release_month     genre mpaa_rating total_gross inflation_adjusted_gross
## 1            12   Musical           G   184925485               5228953251
## 2             2 Adventure           G    84300000               2188229052
## 3            11   Musical           G    83320000               2187090808
## 4            11 Adventure           G    65000000               1078510579
## 5             2     Drama           G    85000000                920608730
## 6            12 Adventure   Not Rated    28200000                528279994
```

Streamlining Financial Data

To facilitate a smoother analysis, we will reorganize our data set by moving the core financial metrics total_gross and inflation_adjusted_gross to more accessible positions within the data frame. Additionally, we will perform a final quality check to remove any records with missing or zero-value financial data.

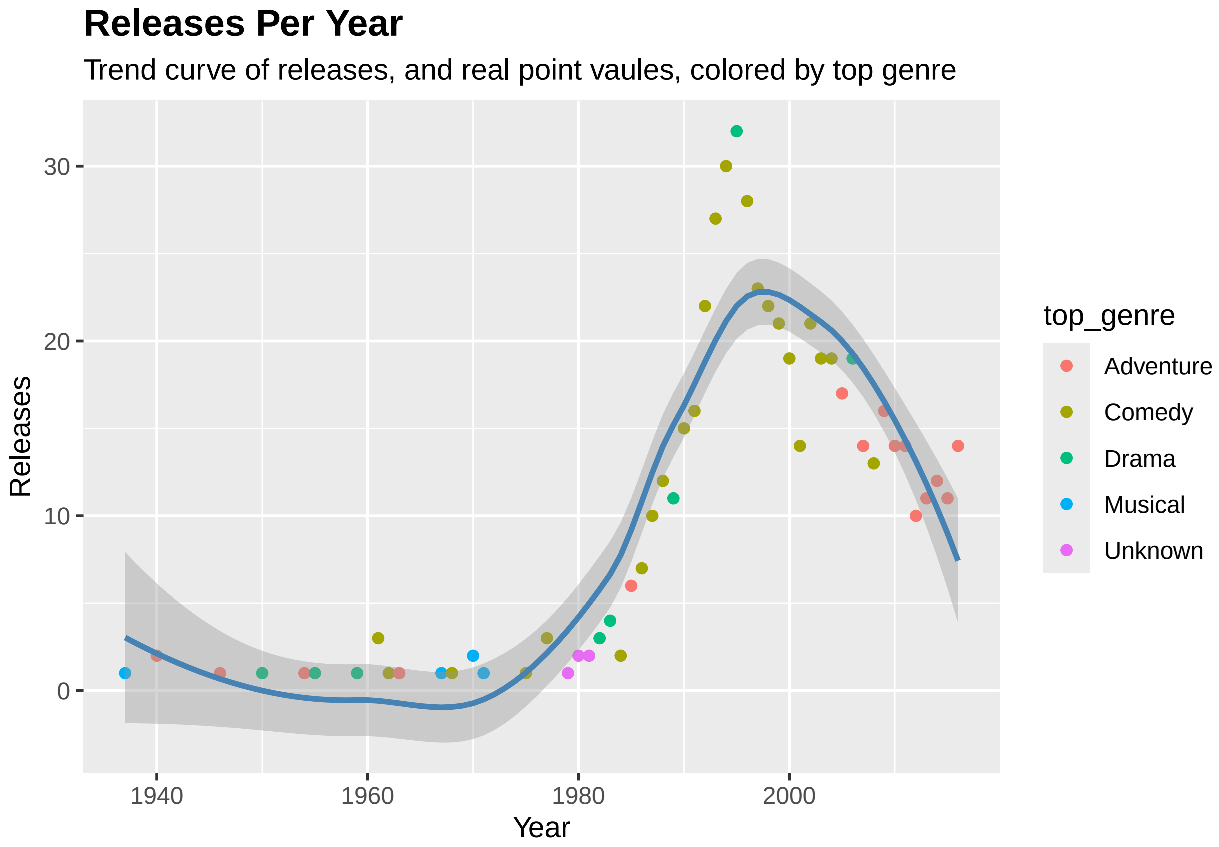
```r
df_clean <- df_clean |>
  relocate(c(total_gross, inflation_adjusted_gross), .before=decade)

df_clean <- df_clean |>
  filter(total_gross > 0, inflation_adjusted_gross > 0)

head(df_clean)
```

```
##                        movie_title total_gross inflation_adjusted_gross decade
## 1 Snow White and the Seven Dwarfs   184925485               5228953251   1930
## 2                       Pinocchio    84300000               2188229052   1940
## 3                        Fantasia    83320000               2187090808   1940
## 4               Song of the South    65000000               1078510579   1940
## 5                      Cinderella    85000000                920608730   1950
## 6    20,000 Leagues Under the Sea    28200000                528279994   1950
##   release_date release_year release_month     genre mpaa_rating
## 1   1937-12-21         1937            12   Musical           G
## 2   1940-02-09         1940             2 Adventure           G
## 3   1940-11-13         1940            11   Musical           G
## 4   1946-11-12         1946            11 Adventure           G
## 5   1950-02-15         1950             2     Drama           G
## 6   1954-12-23         1954            12 Adventure   Not Rated
```

2. Exploratory Data Analysis

A. Changes across historical output.

```
yearly_releases <- df_clean |>
  group_by(release_year) |>
  summarise(releases=n(), top_genre = fmode(genre))

ggplot(yearly_releases, aes(x=release_year, y=releases, color=top_genre)) +
  geom_point() +
  geom_smooth(method="loess", formula=y~x, color="steelblue") +
  labs(
    title="Releases Per Year",
    subtitle="Trend curve of releases, and real point vaules, colored by top genre",
    x="Year",
    y="Releases",
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 14)
  )
```



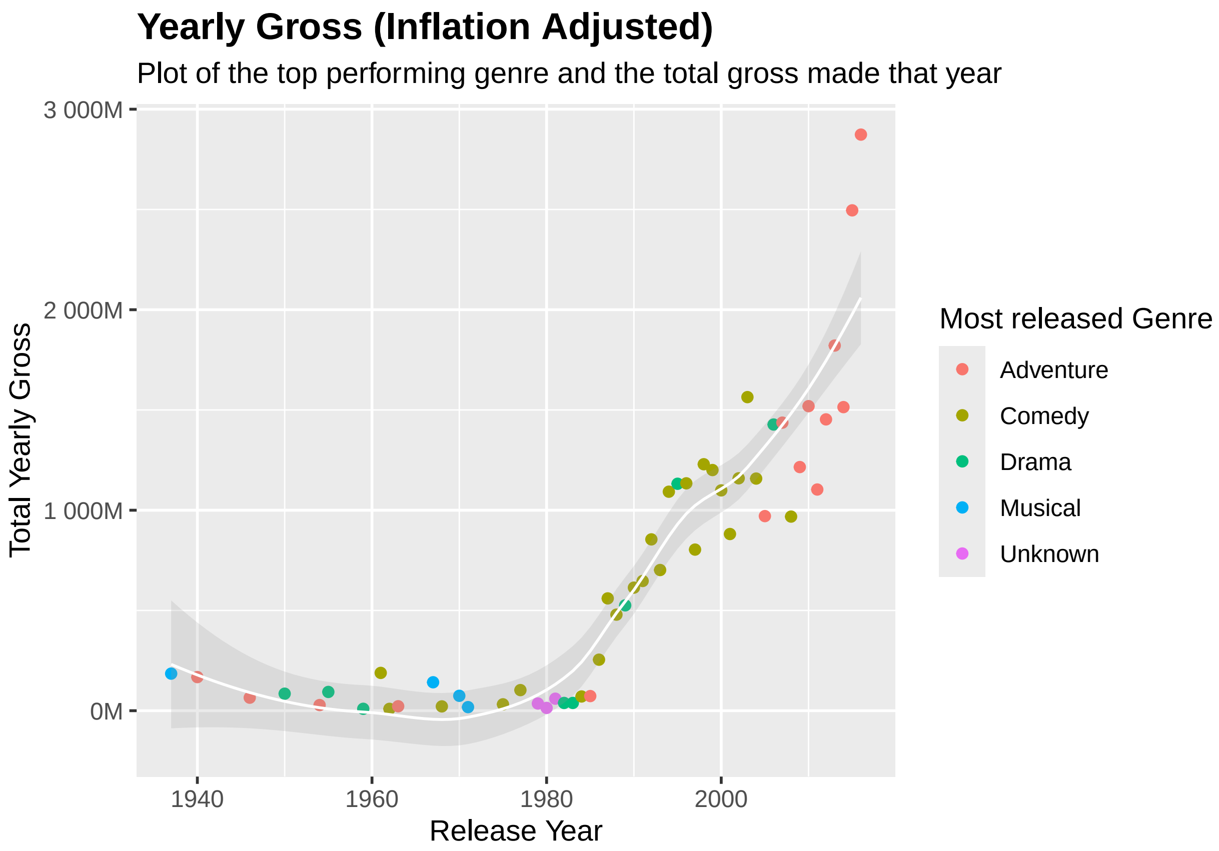
From the late 1930s through the 1960s, releases remain consistently low, generally fluctuating between zero and three per year. This period is characterized by **sparse output** and **minimal genre clustering**, suggesting either limited production overall or incomplete historical coverage. Beginning in the late 1970s, there is a noticeable upward shift. Releases increase steadily through the 1980s, marking the start of a significant growth phase.

By the early to mid-1990s, the number of releases rises sharply, culminating in a clear peak in the late 1990s to early 2000s, where annual releases reach their highest levels (above 20 per year in several cases). Following this peak, the trend curve shows a gradual decline through the 2000s and into the 2010s. While releases remain higher than pre-1980 levels, the downward trajectory suggests market saturation, changing production dynamics, or shifts in data inclusion over time. Genre-wise, **Comedy** and **Drama** dominate during the peak growth years, contributing heavily to the surge in releases during the 1990s. **Adventure** titles appear more frequently in the later years, while **Musical** and **Unknown** genres remain relatively rare across the entire timeline.

Overall, the graph illustrates a long period of low activity, a rapid expansion beginning around 1980, a pronounced peak around the turn of the millennium, and a subsequent decline, indicating a classic rise-and-fall production cycle rather than steady linear growth.

```r
yearly_releases <- df_clean |>
  group_by(release_year) |>
  summarize(
    total_yearly_gross = sum(total_gross),
    inflation_yearly_adjusted_gross = sum(inflation_adjusted_gross),
    top_genre = fmode(genre),
    n=n(),
  )
```

```r
ggplot(yearly_releases, aes(x=release_year, y=total_yearly_gross, color=top_genre)) +
  geom_point() +
  geom_smooth(linewidth = 0.5, color="white", alpha=0.2, method="loess", formula = y~x) +
  scale_y_continuous(labels = label_number(scale = 1e-6, suffix = "M")) +
  labs(
    title="Yearly Gross (Inflation Adjusted)",
    subtitle = "Plot of the top performing genre and the total gross made that year",
    x = "Release Year",
    y = "Total Yearly Gross",
    color = "Most released Genre",
  ) +
  theme(
    legend.position = "right",
    plot.title = element_text(face = "bold", size = 14)
  )
```



This plot shows total yearly gross revenue (inflation adjusted) over time, with each point colored by the most released (top performing) genre in that year. From the late 1930s through the 1970s, total yearly gross remains relatively low and volatile, rarely exceeding a few hundred million. No single genre consistently dominates during this period, reflecting both lower overall box office scale and a fragmented genre landscape.

A clear structural change occurs beginning in the mid-to-late 1980s. Total yearly gross increases sharply and continues rising into the 1990s, indicating rapid expansion of the film market in real (inflation adjusted) terms. During this growth phase, Comedy frequently appears as the top performing genre, suggesting it played a central role in driving box office gains during this era. Entering the 2000s and especially the 2010s, total yearly gross reaches its highest levels, surpassing one billion and eventually approaching three billion in some years. In this period, **Adventure** becomes the dominant genre, coinciding with the rise of **franchise-driven**, effects-heavy blockbusters that generate substantial box office revenue.

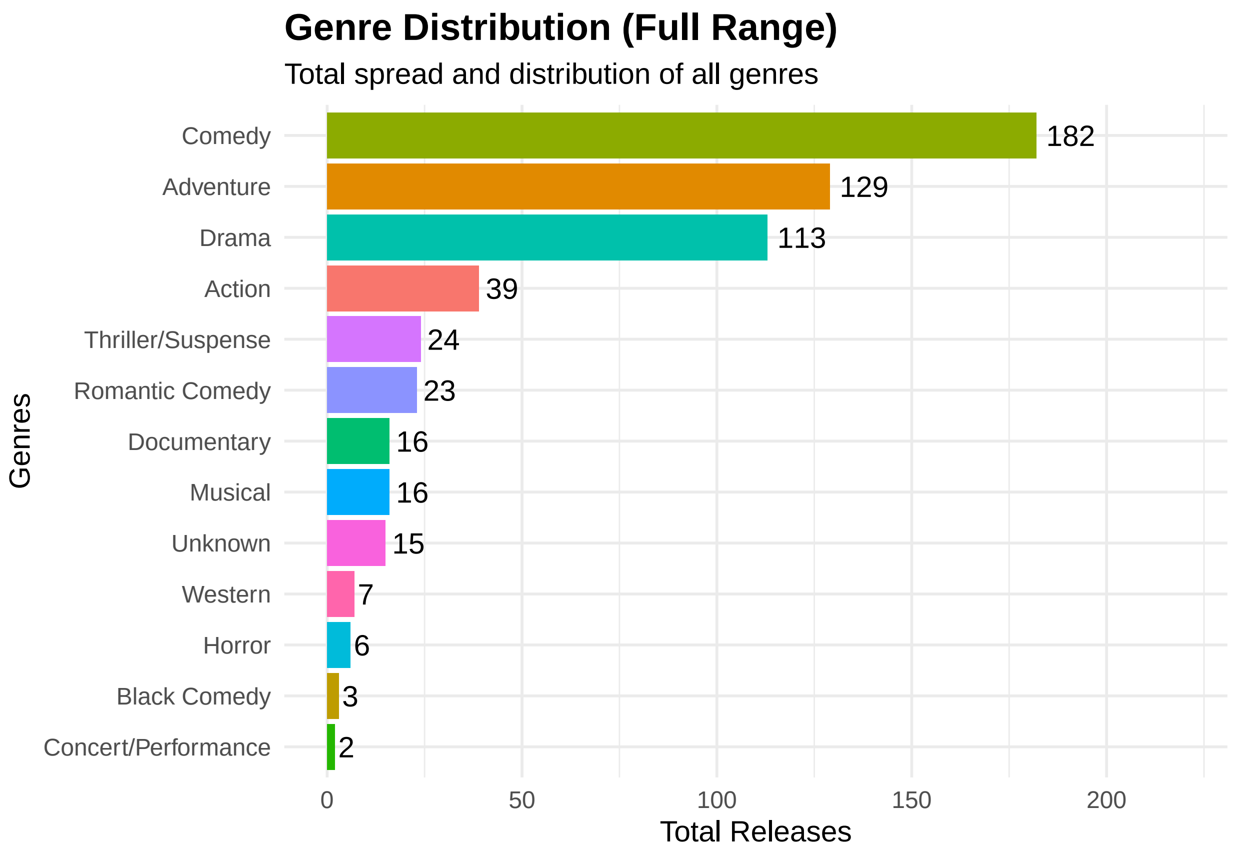
Overall, the graph illustrates a long-term escalation in real box office revenue, with a notable genre transition from **Comedy**-led dominance in the 1990s to **Adventure**-led dominance in the 2000s and beyond, highlighting a shift toward blockbuster oriented revenue models rather than simply increased release volume.

B. Short analysis of full time scale

In this section, we pivot from temporal trends to volume-based analysis by visualizing the total distribution of genres across **Disney's entire history**. By ranking the genres from most to least frequent, we can immediately identify the core pillars of Disney's production strategy. This visualization highlights whether the studio relies primarily on its classic "Musical" and "Adventure" roots or if high-volume categories like "Comedy" have historically dominated their release calendar.

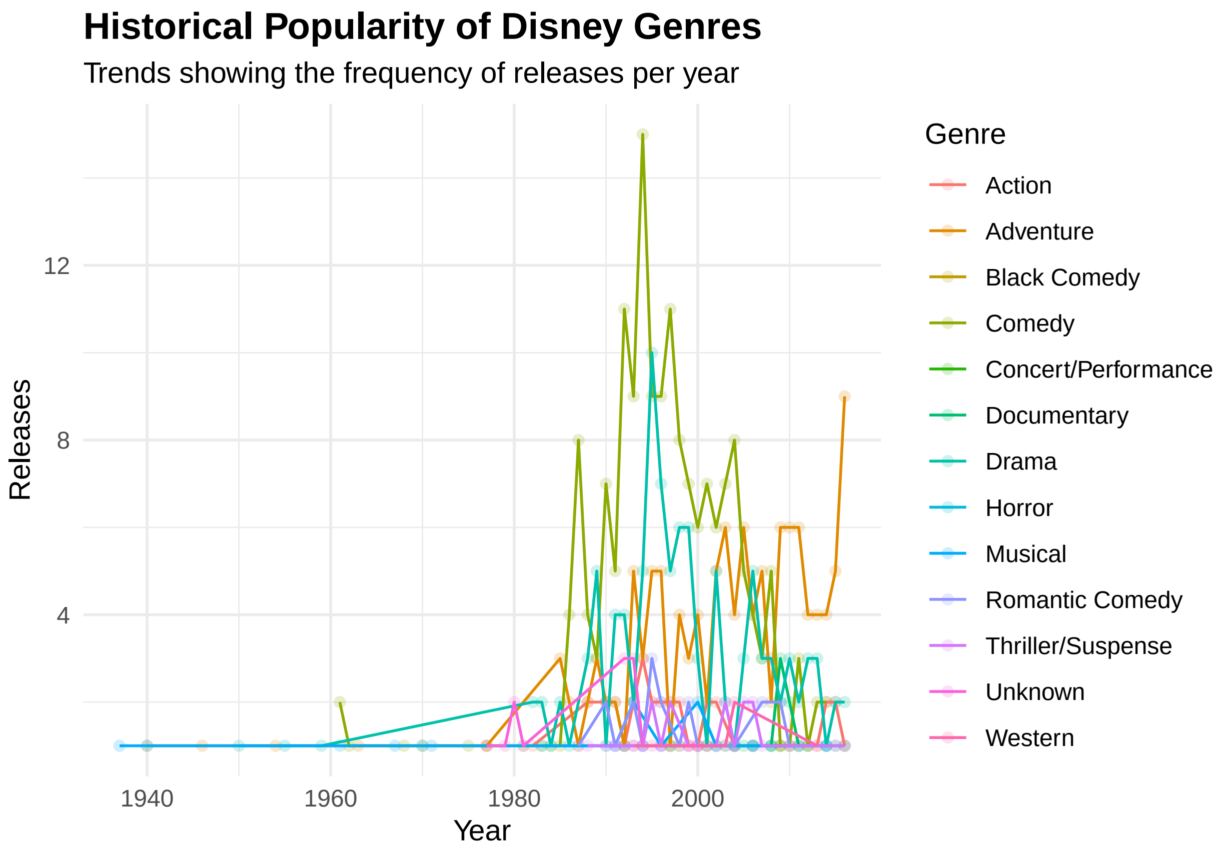
```r
ggplot(df_clean, aes(y = fct_rev(fct_infreq(genre)), fill = genre)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = ..count..), hjust = -0.2) +
  expand_limits(x = 220) + # Gives room for the text labels
  theme_minimal() +
  theme(legend.position = "none") +
  labs(
    title="Genre Distribution (Full Range)",
    subtitle = "Total spread and distribution of all genres",
    y="Genres",
    x="Total Releases"
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 14)
  )
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once per session.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



The distribution clearly shows that **Comedy** is Disney's most frequently released genre by a substantial margin, followed by **Adventure** and **Drama**. Together, these three categories account for the majority of all releases, indicating a strong emphasis on broadly accessible, repeatable formats rather than niche storytelling. Mid-tier genres such as Action, Thriller/Suspense, and Romantic Comedy appear with moderate frequency, reflecting periodic diversification without sustained focus. In contrast, genres like Western, Horror, Black Comedy, and Concert/Performance are rare, highlighting areas that Disney has historically deprioritized or only experimented with briefly. Overall, this distribution reinforces the idea that Disney's **long-term strategy favors salable, mass-appeal genres** while selectively deploying lower-frequency genres for experimentation or brand expansion, rather than as core production pillars.

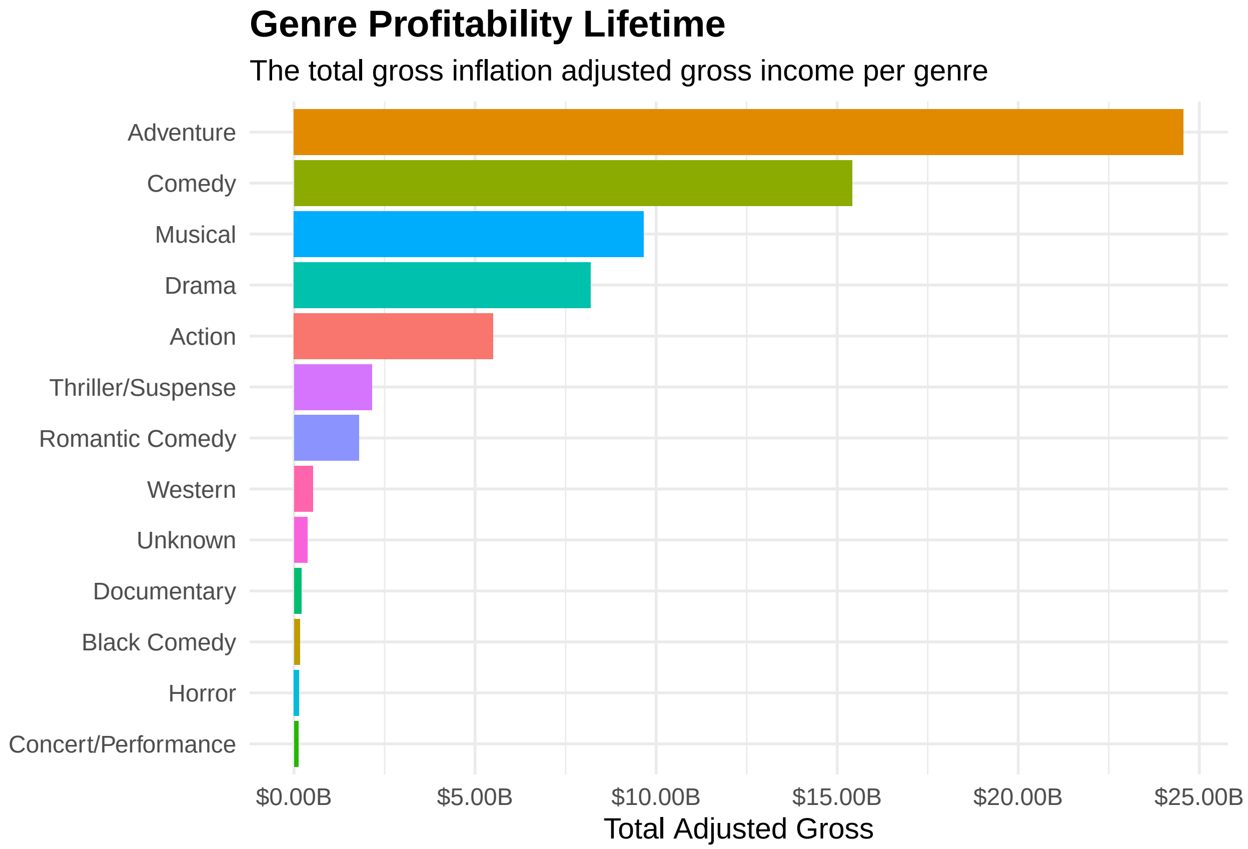
```r
df_yearly <- df_clean |>
  group_by(release_year, genre) |>
  summarise(n = n(), .groups="drop")

ggplot(df_yearly, aes(x = release_year, y = n, color = genre)) +
  geom_point(alpha = 0.2) +
  geom_line() +
  theme_minimal() +
  labs(
    title = "Historical Popularity of Disney Genres",
    subtitle = "Trends showing the frequency of releases per year",
    x = "Year",
    y = "Releases",
    color = "Genre"
  ) +
  theme(
    legend.position = "right",
    plot.title = element_text(face = "bold", size = 14)
  )
```



This figure provides a more detailed break down of each genre's popularity by release. Comedy (the light green line) is the clear dominant force during this era, hitting an all - time peak of approximately 15 releases in a single year around 1995. Drama (the teal line) also saw a significant spike during this period, briefly rivaling comedy with about 11 releases in one year. Several genres have remained consistently low - volume throughout Disney's history, Westerns, Horror, and Black Comedy: These genres rarely see more than 1 or 2 releases a year, remaining niche within the Disney portfolio.

C. Genre Popularity and Profitability

```
genre_gross = df_clean |>
  group_by(genre) |>
  summarise(genre_inflation_adjusted_gross = sum(inflation_adjusted_gross))

ggplot(genre_gross, aes(x = reorder(genre, genre_inflation_adjusted_gross), y = genre_inflation_adjuste
  geom_bar(stat="identity") +
  scale_y_continuous(labels = label_dollar(scale = 1e-9, suffix = "B")) +
  coord_flip() +
  theme_minimal() +
  labs(
    title = "Genre Profitability Lifetime",
    subtitle = "The total gross inflation adjusted gross income per genre",
    y = "Total Adjusted Gross",
    x = "Genre",
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    legend.position = "none",
    axis.title.y = element_blank(),
  )
```



The chart shows a significant “heavy-hitter” trend where the top four genres dominate the total gross income, then a noticeable drop-off after the top four with several genres represent a very small fraction of the total lifetime gross.

Adventure: The undisputed leader, generating nearly $25B. This likely reflects the high commercial success of blockbuster franchises.

Comedy: Ranks second, bringing in roughly $15B.

Musical & Drama: These round out the top tier, sitting between $8B and $10B.

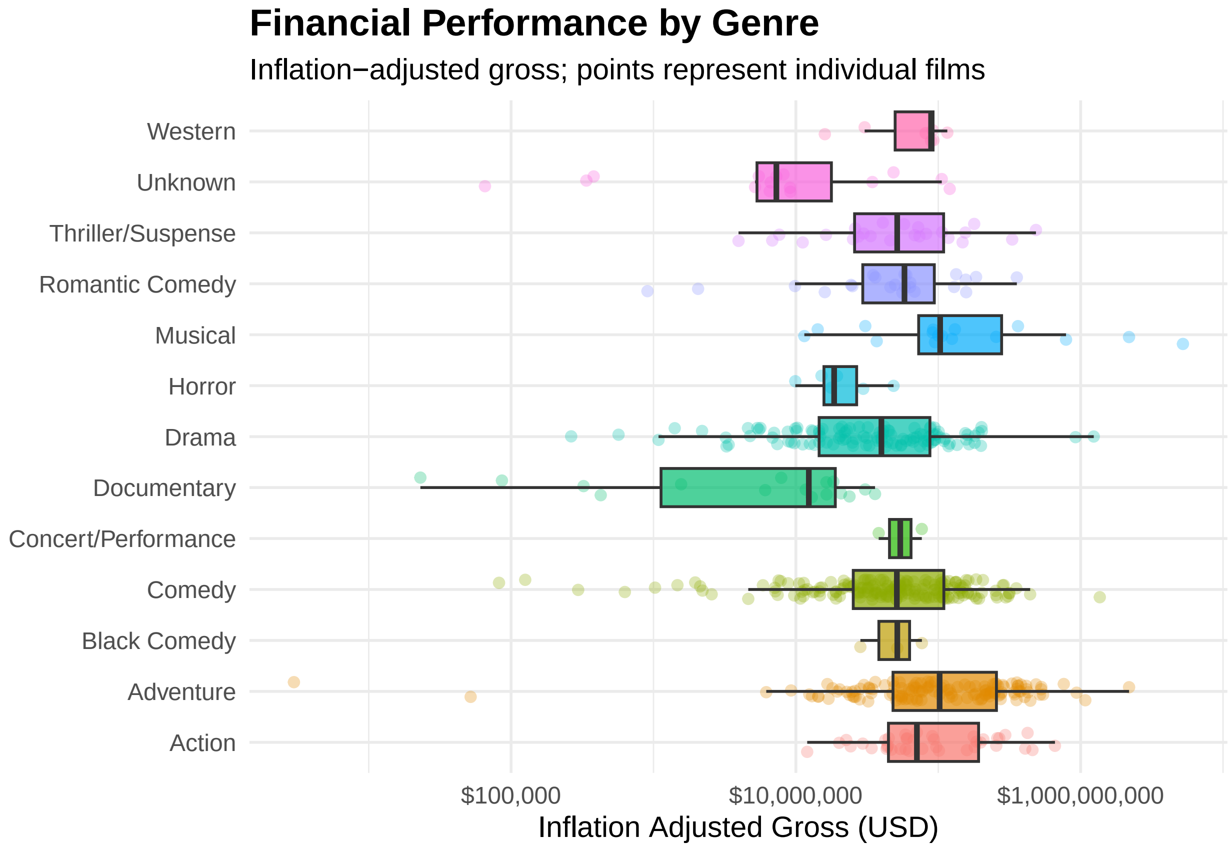
Action: Surprisingly sits in the middle of the pack (approx. $5,500M), though it’s worth noting that many “Action” films are often cross-categorized as “Adventure.”

Thriller/Suspense & Romantic Comedy: Both hover around the $2B - $2.5B mark.

Westerns and Documentaries: show minimal financial footprint compared to mainstream fiction.

Horror & Black Comedy: These genres appear at the very bottom. While often highly profitable relative to their low budgets (high ROI), their total gross volume is significantly lower than broad-appeal genres like Adventure.

```r
# considered removing outliers entirely.
# limit <- quantile(df_clean$inflation_adjusted_gross, 0.75) +
#   1.5 * IQR(df_clean$inflation_adjusted_gross)
#
# modified <- df_clean |>
#   filter(inflation_adjusted_gross <= limit)


ggplot(df_clean, aes(x = genre, y = inflation_adjusted_gross, fill = genre)) +
  geom_jitter(aes(color = genre), width = 0.2, alpha = 0.3) +
  geom_boxplot(outlier.shape = NA, alpha = 0.7, color = "grey20") +
  coord_flip() +
  # Use log scale because the gap between "Snow White" and others is too large to see detail
  scale_y_log10(labels = label_dollar()) +
  theme_minimal() +
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold", size = 14),
    axis.title.y = element_blank()
  ) +
  labs(
    title = "Financial Performance by Genre",
    subtitle = "Inflation-adjusted gross; points represent individual films",
    y = "Inflation Adjusted Gross (USD)",
  )
```



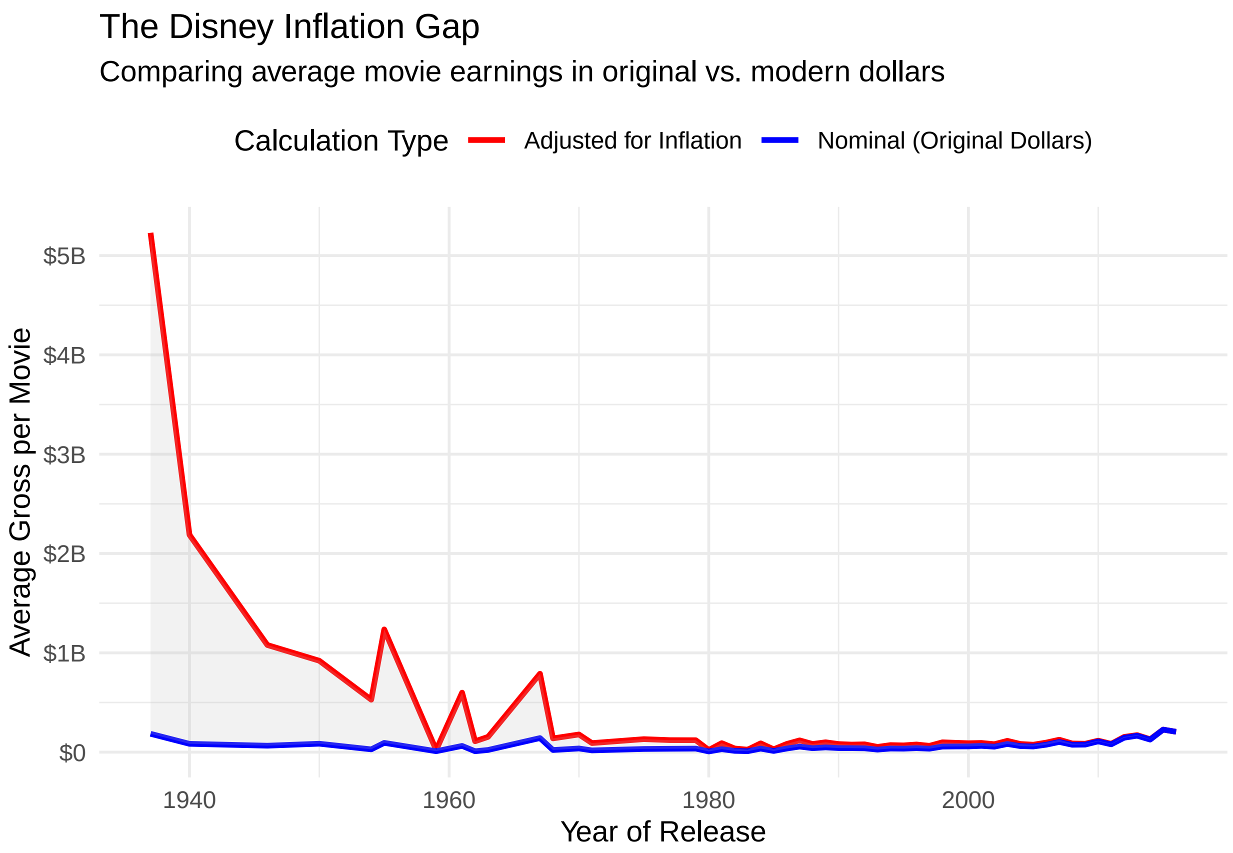
This box plot visualizes the financial performance of various film genres from the Disney Movies data set. It uses a logarithmic scale to show inflation-adjusted gross earnings, which allows us to see both the "blockbuster" hits and the smaller-scale productions in one view.

Top Performers and Stability: The **Adventure** & **Musical** genres appear to be the "heavy hitters." The **Adventure** genre has a high median gross and a significant number of films crossing the $1 billion mark (after inflation). **Musicals** show the highest median performance, with a relatively tight "interquartile range" (the box), suggesting consistent financial success.

Higher Risk areas include **Drama** & **Comedy:** Both genres show a massive number of individual data points (the dots). While their medians are healthy, the vast spread suggests these are high-volume genres with a high degree of unpredictability—some become massive hits, while others struggle to find an audience. Interestingly, the "Niche" genres like **Western**, **Black Comedy**, and **Concert/Performance** have very few data points. Their boxes are narrow, but this is likely due to a smaller sample size rather than guaranteed consistency.

D. The "Inflation Gap" - Stripping Away the Illusion of Inflation

```
df_inflation <- df_clean |>
  group_by(release_year) |>
  summarise(
    avg_nominal = mean(total_gross),
    avg_adjusted = mean(inflation_adjusted_gross)
  ) |>
  pivot_longer(
    cols = c(avg_nominal, avg_adjusted),
    names_to = "gross_type",
    values_to = "amount")

wide = df_inflation |>
  pivot_wider(names_from = gross_type, values_from = amount)

ggplot(df_inflation, aes(x = release_year, y = amount, color = gross_type)) +
  geom_line(linewidth=1) +
  # Fill the area between lines to emphasize the "gap"
  geom_ribbon(wide,
              mapping=aes(x = release_year,
                  ymin = avg_nominal,
                  ymax = avg_adjusted),
              inherit.aes = FALSE,
              fill = "grey",
              alpha = 0.2
              ) +
  scale_y_continuous(labels = label_dollar(scale_cut = cut_short_scale())) +
  scale_color_manual(
    values = c("avg_adjusted" = "red", "avg_nominal" = "blue"),
    labels = c("Adjusted for Inflation", "Nominal (Original Dollars)")
  ) +
  theme_minimal() +
  labs(
    title = "The Disney Inflation Gap",
    subtitle = "Comparing average movie earnings in original vs. modern dollars",
    x = "Year of Release",
    y = "Average Gross per Movie",
    color = "Calculation Type"
  ) +
  theme(legend.position = "top")
```



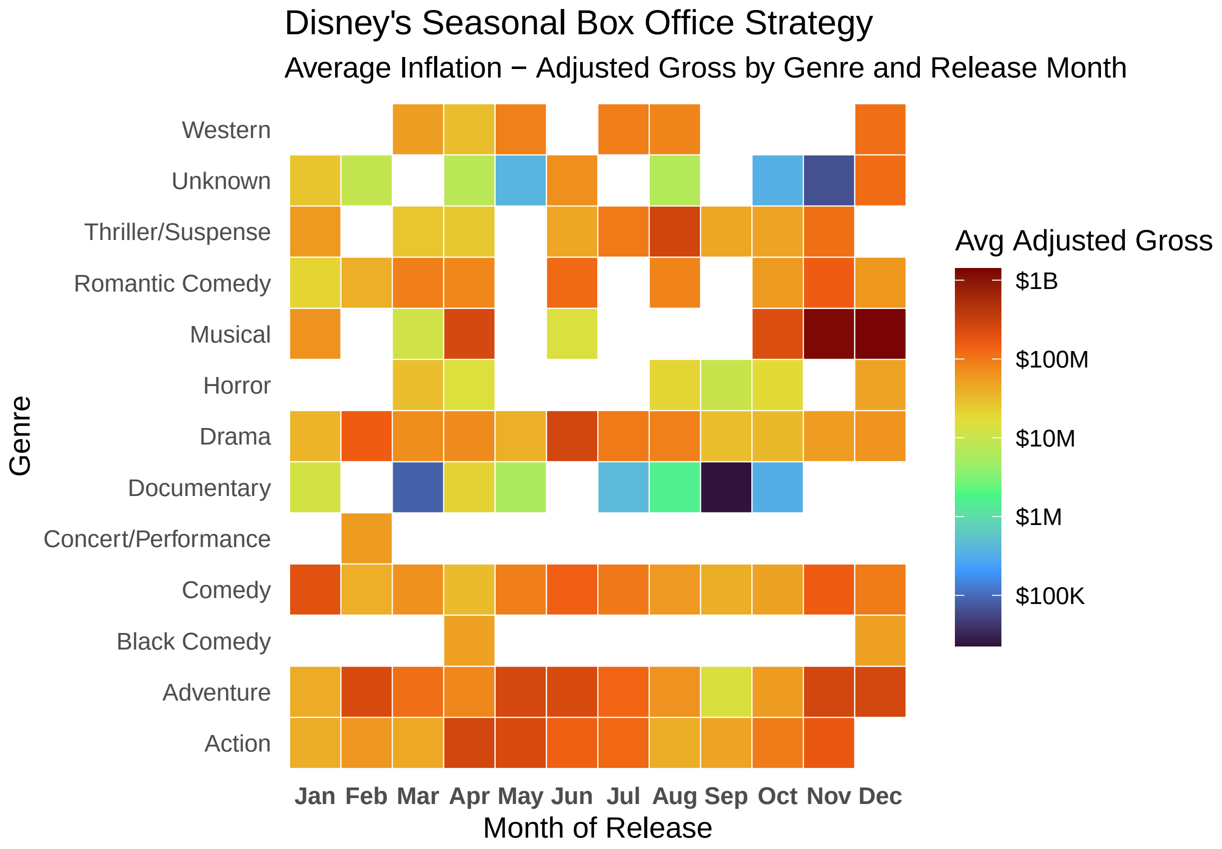
The most dramatic feature is the **red line** (Adjusted for Inflation) starting at over **$5B** around 1937. This represents the release of *Snow White and the Seven Dwarfs*. While the nominal earnings (**blue line**) at the time look like a flat line near zero, the inflation-adjusted value is astronomical. This highlights how dominant early Disney classics were in a market with far less competition. As we move toward the present day (the right side of the chart), the red and blue lines eventually **merge.** This happens because "modern dollars" are the benchmark. In very recent years, there is no "inflation gap" yet to calculate, so the nominal gross and the adjusted gross are essentially the same.

There are a few other key observations to make from this chart. You can see significant volatility in the red line during the mid-century. The spikes likely represent "Event" movies (like *Mary Poppins* in 1964 or *The Jungle Book* in 1967) that performed significantly better than the average Disney output of that era.

Recently we see Nominal Growth, the **blue line** (Nominal Dollars) shows a slow, steady climb starting around the year 2000. This reflects the modern era of blockbusters (Marvel, Star Wars, and Pixar) where movies consistently cross the $1B mark in actual, non-adjusted dollars.

E. Seasonal Success - Which Months Actually Drive Genre Performance?

```r
df_heatmap <- df_clean |>
  mutate(
    release_month = month(release_date, label=TRUE, abbr=TRUE),
  ) |>
  group_by(genre, release_month) |>
  summarise(avg_adj_gross = mean(inflation_adjusted_gross), .groups="drop")

ggplot(df_heatmap, aes(x = release_month, y = genre, fill = avg_adj_gross)) +
  geom_tile(color = "white") + # White borders make tiles pop
  # Using a viridis color scale (standard for heatmaps, colorblind friendly)
  scale_fill_viridis_c(
    option = "turbo",
    trans = "log10",
    labels = label_dollar(scale_cut = cut_short_scale()), # Formats as $100M
    name = "Avg Adjusted Gross"
  ) +
  theme_minimal() +
  labs(
    title = "Disney's Seasonal Box Office Strategy",
    subtitle = "Average Inflation - Adjusted Gross by Genre and Release Month",
    x = "Month of Release",
    y = "Genre"
  ) +
  theme(
    axis.text.x = element_text(face = "bold"),
    panel.grid = element_blank(),
    legend.position = "right",
    legend.key.height = unit(1, "cm")
  )
```



The chart reveals a very deliberate "calendar" for certain types of films. **"Dump Months",** notice the "greener" or "cooler" spots in September. Historically, this is a slower time for cinema as kids return to school. Disney's Adventure films, for instance, see a significant dip in performance here. During June and July, it show high performance for **Drama** and **Adventure**, as the studio capitalizes on family vacations.

Interestingly, **Horror** shows a slight uptick in October, leaning into the Halloween season, though it remains a much smaller earner for Disney compared to other genres. The “cool” colors (blue, purple, and green) highlight areas where Disney either struggles to find a massive audience or focuses on smaller, prestige projects. Documentaries are the lowest earners on the map, with several months (like September) hitting the dark purple range, signifying earnings around **$100K**. The “Unknown” genre row shows high volatility. It hits a peak in December but has significant “cold” spots in October and November. This could represent experimental titles or films that didn’t fit into a traditional marketing bucket. However we must consider our methodology, this might simply be a side affect of how we prepared the data, and assigned these labels to missing data points. It may simply be coincidental.

F. Statistical Reflection - Proof of Disney’s Content Formulas

```r
df_anova <- df_clean |>
  filter(mpaa_rating %in% c("G", "PG", "PG-13", "R")) # Remove Unrated

rating_anova <- aov(inflation_adjusted_gross ~ mpaa_rating, data = df_anova)

summary(rating_anova)
```

I. *Variance Analysis - Testing the Financial Impact of MPAA Ratings*

```
##              Df    Sum Sq   Mean Sq F value  Pr(>F)    
## mpaa_rating   3 3.046e+18 1.015e+18   12.13 1.1e-07 ***
## Residuals   516 4.318e+19 8.368e+16                    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This table shows the results of a **One-Way ANOVA (Analysis of Variance)**. This test is used to determine if there are statistically significant differences between the means of three or more independent group, in this case, how different `mpaa_rating` categories (like G, PG, PG-13, R) affect a numerical outcome.
Provided is a breakdown of these values and their meaning:

F-value 12.13: This is the ratio of variance between groups to the variance within groups. A value significantly greater than 1 suggests the groups are quite different from each other.

p-value (Pr(>F)) 1.1x10^-7: This is extremely small (0.00000011). Since it is well below the standard threshold of **0.05**, the result is highly statistically significant. In other words the impact of the mpaa rating on the adjusted gross mean is ***not*** coincidental and could use further analysis with more in depth metrics. The three stars (`***`) indicate a very high level of confidence. We can reject the null hypothesis that all MPAA ratings have the same mean outcome. In plain English: The MPAA rating of a movie has a significant impact on the result we are measuring.

Df (Degrees of Freedom) 3: This indicates there are 4 rating categories being compared, as Df = n - 1, where n is the number of categories.

However An ANOVA tells us *that* a difference exists, but it doesn’t tell us *where* it is. For example, we don’t know if PG movies outperform R movies, or if G movies are the ones dragging the average down. We can make guesses based on our own intuition, but that is out of scope of this analysis.

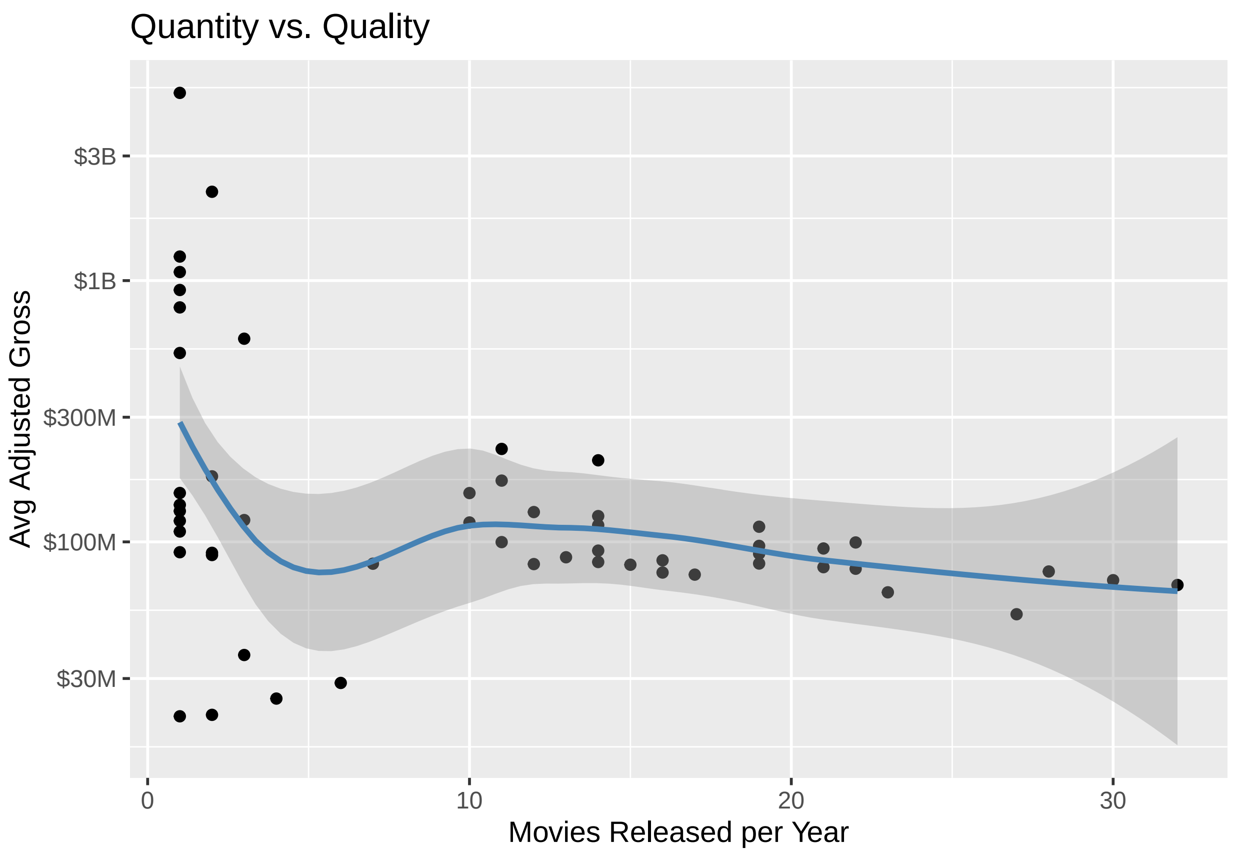
```
df_cor <- df_clean |>
  group_by(release_year) |>
  summarise(
    count = n(),
    avg_adj_gross = mean(inflation_adjusted_gross)
  )

cor_result <- cor.test(df_cor$count, df_cor$avg_adj_gross)

cor_result
```

II. Volume vs. Value - Measuring the Relationship between Output and Average Gross

```
##
##  Pearson's product-moment correlation
##
## data:  df_cor$count and df_cor$avg_adj_gross
## t = -2.3023, df = 52, p-value = 0.02535
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.52884044 -0.03961357
## sample estimates:
##       cor
## -0.304148
```

Our Pearson correlation analysis revealed a significant negative relationship between the annual volume of movie releases and the average inflation-adjusted gross per film ($r = -0.30$, $p < 0.05$). This suggests a **'Dilution Effect'**. In years where Disney's production slate was most crowded, individual film performance tended to decline. This could be due to internal competition (cannibalizing their own audience) or a spread of creative resources too thin across too many projects. For stakeholders, this suggests that a **'Boutique'** strategy focusing on fewer, higher-impact releases may be more financially efficient than a high-volume approach.

```
ggplot(df_cor, aes(x = count, y = avg_adj_gross)) +
  geom_point() +
  geom_smooth(color="steelblue", method = "loess", formula=y~x) +
  scale_y_log10(labels = label_dollar(scale_cut = cut_short_scale())) +
  labs(title = "Quantity vs. Quality", x = "Movies Released per Year", y = "Avg Adjusted Gross")
```



The scatter plot highlights a clear trade-off between production volume and financial density. By utilizing a logarithmic scale, we can see that the decline in average gross is consistent across orders of magnitude. While the 'Renaissance' years (lower-left) maintained high average yields through scarcity, the modern 'High-Output' era (right-side) shows a clear regression toward a lower mean, validating the -0.30 correlation and suggesting that market saturation may be a limiting factor for per-film performance. This is the visual representation of our -0.30 correlation. As the dots move further to the right (more movies released), they generally drop lower on the vertical axis (lower average gross). This suggests that Disney's "Sweet Spot" might be on the left side of the graph. High-volume years often fail to produce the same per-movie "punch" as leaner, more focused years. This is interesting when we return up to section 2a, where the graph of Disney's output suggest this ***is*** happening.

```
model <- lm(inflation_adjusted_gross ~ genre + mpaa_rating, data = df_anova)
summary(model)
```

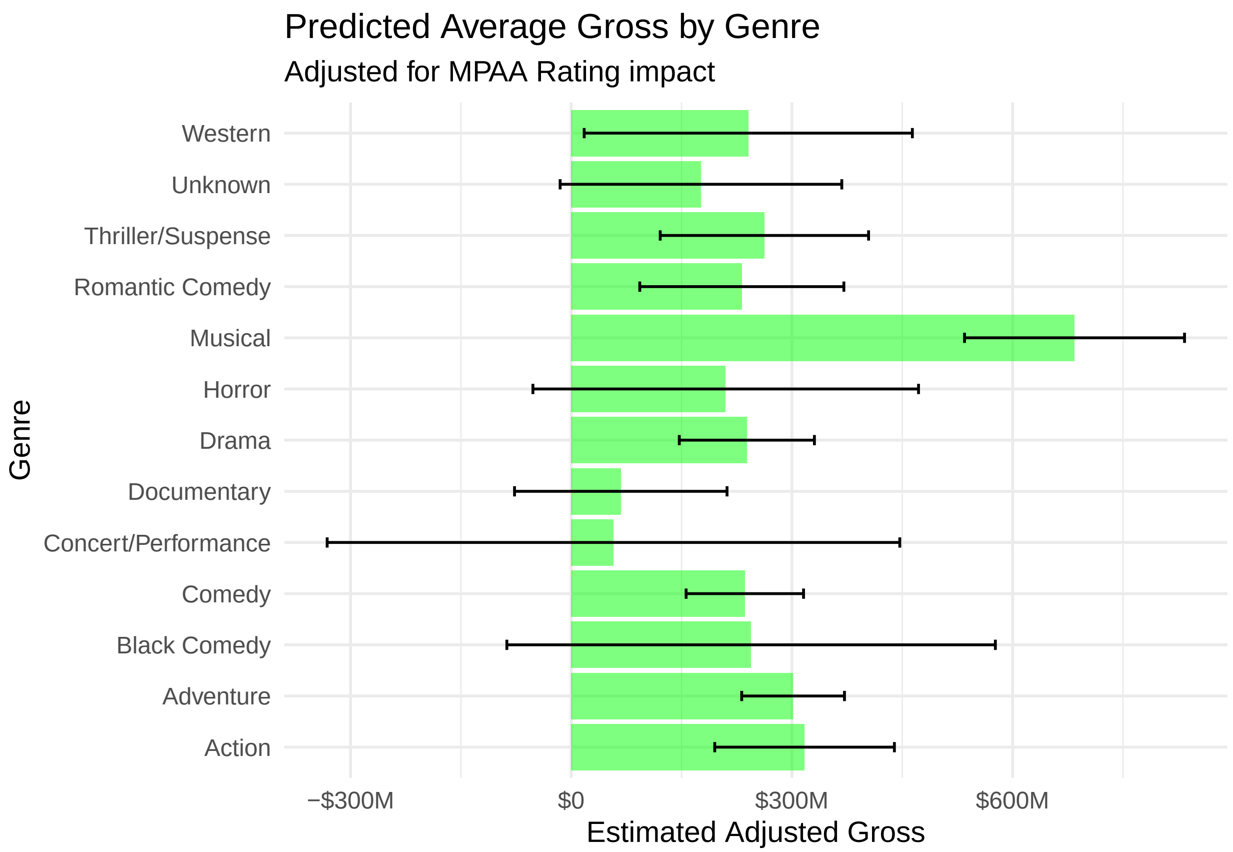
III. Predictive Modeling - Determining the Revenue Drivers of Disney's Catalog

```
##
## Call:
## lm(formula = inflation_adjusted_gross ~ genre + mpaa_rating,
##     data = df_anova)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -669901756  -66564655  -29706248   31748132 4544813351
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               317184154   62153379   5.103 4.74e-07 ***
## genreAdventure            -15608096   57641460  -0.271 0.786672
## genreBlack Comedy         -72659002  169943754  -0.428 0.669163
## genreComedy               -81231437   53268380  -1.525 0.127900
## genreConcert/Performance -259773315  207618691  -1.251 0.211441
## genreDocumentary         -249828380   91749218  -2.723 0.006695 **
## genreDrama                -78476615   54540015  -1.439 0.150805
## genreHorror              -107191347  134654640  -0.796 0.426380
## genreMusical              366955746   93165083   3.939 9.34e-05 ***
## genreRomantic Comedy      -85305427   76324082  -1.118 0.264239
## genreThriller/Suspense    -54618828   74843466  -0.730 0.465867
## genreUnknown             -140854056  100215915  -1.406 0.160487
## genreWestern              -76513806  115737764  -0.661 0.508853
## mpaa_ratingPG            -161454886   39318221  -4.106 4.69e-05 ***
## mpaa_ratingPG-13         -155491070   44807598  -3.470 0.000565 ***
## mpaa_ratingR             -192281660   49020140  -3.923 9.98e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 280200000 on 504 degrees of freedom
## Multiple R-squared:  0.1443, Adjusted R-squared:  0.1188
## F-statistic: 5.664 on 15 and 504 DF,  p-value: 8.587e-11
```

Our linear regression model ($p < 0.001$) reveals that Disney’s financial ‘North Star’ remains the **G-rated Musical**. While the industry has shifted toward PG-13 blockbusters, the data shows that a Musical adds an estimated $366.9 Million in inflation-adjusted value over Action films, while moving from a ‘G’ to a ‘PG-13’ rating actually correlates with an average $155 Million decrease in adjusted yield. This highlights that Disney’s core competitive advantage is not just in ‘Action,’ but in high-quality, family-accessible musical content. This is clearly evident in market trends, in the current day the most popular movie among young children K-Pop Demon Hunters, plays directly into these findings, as well as Disney’s own movies, Frozen, Wish, Encanto, all having significant musical elements.

While these films aren’t in the data set, it is notable that they follow the trends found from this data set.

```r
df_predict <- ggpredict(model, terms = "genre")

# Plot the predictions
ggplot(df_predict, aes(x = x, y = predicted)) +
  geom_bar(stat = "identity", fill = "green", alpha=0.5) +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.2) +
  coord_flip() +
  scale_y_continuous(labels = label_dollar(scale_cut = cut_short_scale())) +
  labs(
    title = "Predicted Average Gross by Genre",
    subtitle = "Adjusted for MPAA Rating impact",
    x = "Genre",
    y = "Estimated Adjusted Gross"
  ) +
  theme_minimal()
```



This bar chart illustrates the **Predicted Average Gross by Genre**, with values adjusted for the impact of MPAA ratings. It uses green bars to represent the point estimate and horizontal "error bars" (whiskers) to show the range of uncertainty or variability for each genre.
The musical genre is the undisputed leader in predicted gross, with an estimated average significantly higher than any other category (approaching **$700M**). Even its lower-bound error bar remains higher than the predicted average of most other genres.
Documentary and Concert/Performance have the lowest predicted average gross, and the highest uncertainty. Their error bars suggest high unpredictability in these genres.
Several genres cluster in the **$200M** to **$300M** range. Action and Adventure show strong, consistent predicted performance with relatively tight error bars, suggesting more predictable returns, with Comedy, Drama, and Western following closely behind.

The black horizontal lines are crucial for interpreting the "risk" or "volatility" of these predictions. Using all these details and predictions we can create a somewhat accurate chart of how each genre performs.

Performance Tier	**Genres**
High Performer	Musical
Consistent/Mid-Tier	Action, Adventure, Comedy, Drama, Western
Low Performer	Documentary, Concert/Performance, Horror
High Risk/Variance	Concert/Performance, Black Comedy

3. Conclusion & Business Recommendations

This analysis confirms that Disney's longevity is rooted in a highly specific financial formula. While the studio has expanded into diverse genres, its "North Star" remains high quality, family accessible content with strong musical elements.

Prioritize Quality over Quantity: The identified **Dilution Effect** suggests that increasing the annual volume of releases correlates with a decline in average per-film gross. Disney should maintain a "boutique" strategy, focusing resources on fewer, high-impact titles rather than saturating the market.

Double Down on Musical G-Rating Advantage: Statistical modeling shows that **G-rated Musicals** provide the highest predicted average gross, adding an estimated **$366.9 Million** in value over standard Action films. This remains Disney's most significant competitive advantage.

Optimize the Release Calendar: The seasonal heat map reveals a clear **"Dump Month"** effect in September. Adventure and Drama titles should continue to be anchored in the June-July and December windows to maximize the "family vacation" revenue spike.

Leverage Adventure for Stability: While Musicals offer the highest peaks, the **Adventure** genre provides the most consistent high-volume returns in the modern era. This genre should remain the backbone of the "Blockbuster" model, particularly for PG and PG-13 audiences where Musicals may have less reach.