# 970G1: Data Science Research Methods

*Report (1000 Words) T1 Week 8*

Word Count: 987

# Contents

# Introduction

## Background

SussexBudgetProductions' recent £500k comedy-action-thriller grossed only £100k. The CEO suggests making a romance or horror next. I aim to recommend a genre, director, and lead actor that will achieve the highest IMDb score and thus profit.

## The Dataset

The *IMDb dataset* (IMDb, 2024), contains 5043 rows with 28 columns. Ostensibly, each row refers to a unique movie created between 1916-2016. **Table 1** outlines each column, along with descriptions, data levels and summary statistics or examples:

**Table 1:** Summary Table *(Before data cleaning)* for Features in the *IMDb dataset*

***Note:*** *This table was produced in the PDF's .tex file, not the .py script. However, statistics in the table are found in the .py script (lines 38-56)*

| **Feature** | **Description** | **Data Level** | **Summary Statistics / Examples** |
|---|---|---|---|
| `actor_1_facebook_likes` | The number of likes on the lead actor's Facebook page/profile | Ratio | Range: 0–640,000<br>$\bar{x}$: 6,560.05<br>$\sigma$: 15,020.76 |
| `actor_1_name` | The name of the movie's lead actor | Nominal | 2,098 unique values<br><br>e.g., "Kate Winslet", "Joe Mantegna", "Emma Stone" |
| `actor_2_facebook_likes` | The number of likes on the 2nd lead actor's Facebook page/profile | Ratio | Range: 0–137,000<br>$\bar{x}$: 1,651.75<br>$\sigma$: 4,042.44 |
| `actor_2_name` | The name of the movie's 2nd lead actor | Nominal | 3,033 unique values<br><br>e.g., "Stockard Channing", "William Hurt", "Christopher Lee" |
| `actor_3_facebook_likes` | The number of likes on the 3rd lead actor's Facebook page/profile | Ratio | Range: 0–23,000<br>$\bar{x}$: 645.01<br>$\sigma$: 1,665.04 |

| | | | |
|---|---|---|---|
| `actor_3_name` | The name of the movie's 3rd lead actor | Nominal | 3,522 unique values<br><br>e.g., "G.W. Bailey", "Nick Gomez", "Jon Lovitz" |
| `aspect_ratio` | The width-to-height ratio the movie was filmed in | Ratio | 23 unique values<br><br>e.g., 1.33, 1.78, 1.85 |
| `budget` | Movie's budget, expressed in movie's native currency (e.g., US movie = USD($), South Korean movie = KRW(₩)) | Ratio | Range: $10^2$–$10^{10}$<br>$\bar{x}$: 39,752,620<br>$\sigma$: 206,114,900 |
| `cast_total_facebook_likes` | The cumulative number of likes on the cast's Facebook pages/profiles | Ratio | Range: 0–656730<br>$\bar{x}$: 9699.06<br>$\sigma$: 18163.80 |
| `color` | Whether the movie is in colour or black & white | Nominal | 3 unique values (1 NaN)<br><br>e.g., "Color", "Black and White" |
| `content_rating` | The (US-based) content rating for the movie | Nominal | 19 unique values<br><br>e.g., "PG", "R", "PG-13" |
| `country` | The country the movie was made in | Nominal | 66 Unique values (2 ["New Line", "Official Site"] are countries)<br><br>e.g.,<br>"USA", "Canada", "UK" |
| `director_facebook_likes` | The number of likes the director's Facebook page/profile has | Ratio | Range: 0–23000<br>$\bar{x}$: 686.51<br>$\sigma$: 2813.33 |
| `director_name` | The name of the movie's director | Nominal | 2399 unique values<br><br>e.g.,<br>"Dario Argento", "Tarsem Singh", "James Foley" |
| `duration` | The duration of the movie, in minutes | Ratio | Range: 7–511<br>$\bar{x}$: 107.20<br>$\sigma$: 25.20 |
| `facenumber_in_poster` | The number of faces that appear in the movie's promotional poster | Ratio | Range: 0–43<br>$\bar{x}$: 1.37<br>$\sigma$: 2.01 |

| | | | |
|---|---|---|---|
| genres | The genre(s) of the movie | Nominal | 914 unique values<br><br>e.g.,<br>"Comedy—Family",<br>"Action—Adventure",<br>"Crime—Drama" |
| gross | The amount of revenue (in US Dollars) generated by the movie in the US & Canadian market | Ratio | Range: $\$10^2$–$\$10^8$<br>$\bar{x}$: \$48,468,410<br>$\sigma$: \$68,452,990 |
| imdb_score | The average IMDb score of the movie | Ordinal | Rating scale: 1–10<br>$\bar{x}$: 6.47<br>$\sigma$: 1.06<br>Mode: 6.7 |
| language | The language the movie is filmed in | Nominal | 47 unique values (1 NaN)<br><br>e.g.,<br>"English", "German",<br>"French" |
| movie_facebook_likes | The number of like the movie's Facebook page has | Ratio | Range: 0–349000<br>$\bar{x}$: 7525.96<br>$\sigma$: 19320.45 |
| movie_imdb_link | The URL for the movie's IMDb page | Nominal | 4919 unique values |
| movie_title | The title of the movie | Nominal | 4917 unique values<br><br>e.g.,<br>"The Matrix",<br>"Brooklyn", "The<br>Princess Diaries" |
| num_critic_for_reviews | The number of critical reviews given for the movie | Ratio | Range: 1–813<br>$\bar{x}$: 140.19<br>$\sigma$: 121.60 |
| num_user_for_reviews | The number of written reviews given for the movie | Ratio | Range: 1–5060<br>$\bar{x}$: 272.77<br>$\sigma$: 377.98 |
| num_voted_users | The number of IMDb reviews given for the movie | Ratio | Range: 1–$10^6$<br>$\bar{x}$: 83,668.16<br>$\sigma$: 138,485.30 |
| plot_keywords | A list of keywords to describe the movie | Nominal | 4761 unique values<br><br>e.g.,<br>"hitman—outlaw...",<br>"based on comic<br>book—dc comics...",<br>"moral<br>challenge—morality..." |
| title_year | The year the movie was released | Ordinal | Year range: 1916–2016<br>Mode: 2009 |

# Methods

## Hypotheses

$H_1$: "Significant difference in average IMDb score between romance and horror movies"

$H_2$: "$\geq 1$ ("better" genre) director with a significantly higher average IMDb score than the rest"

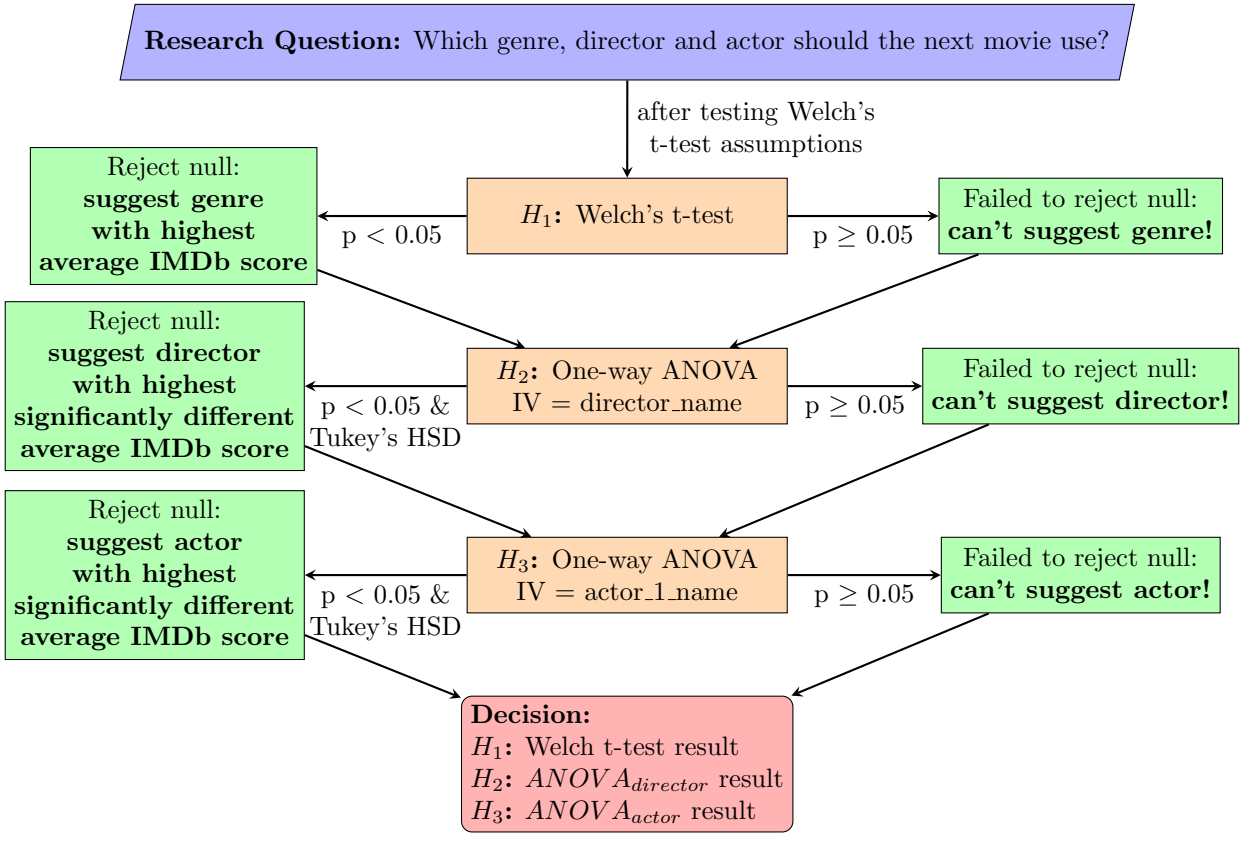
$H_3$: "$\geq 1$ ("better" genre) lead actor with a significantly higher average IMDb score than the rest"

## Analysis plan

Welch's t-test checks if two means differ without assuming equal population variances. One-way ANOVAs assess if any mean differs, and Tukey's HSD identifies the specific means that significantly differ. Significant differences allow us to select feature levels with the highest IMDb scores (i.e. genre, director & actor). If none are significant, solutions can be discussed. **Figure 1** visualises this analysis:

**Figure 1:** Analysis Plan

***Note:*** *This figure was produced in the PDF's .tex file, not the .py script.*



# Data Cleaning & EDA

### Duplicates

*movie_title*, *title_year*, and *director_name* were used to uniquely identify movies. Upon grouping by these features, 124 duplicate rows were removed.

### "Non-movies"

Rows referring to "non-movies" (e.g., TV shows), aren't relevant to this analysis and should be removed. Since every movie has a director, inspecting the data shows known "non-movie" rows do not (e.g., index-459: Daredevil). Removing these drops 102 rows.

### "Very Low" Review Count

Too few reviews means sample IMDb scores are unreliable estimates; data inspection could help define "too few".

**Figure 2:** Distribution of Number of Unique Reviewers for Each Movie

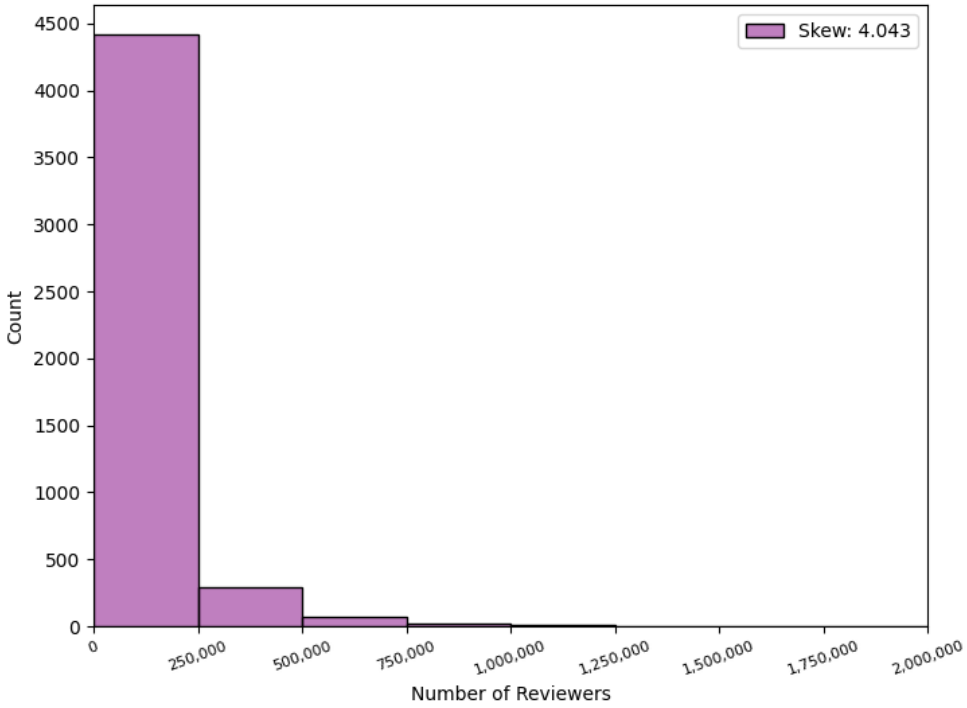


**Figure 3:** *Logarithmic* Distributions of Number of Unique Reviewers for Each Movie

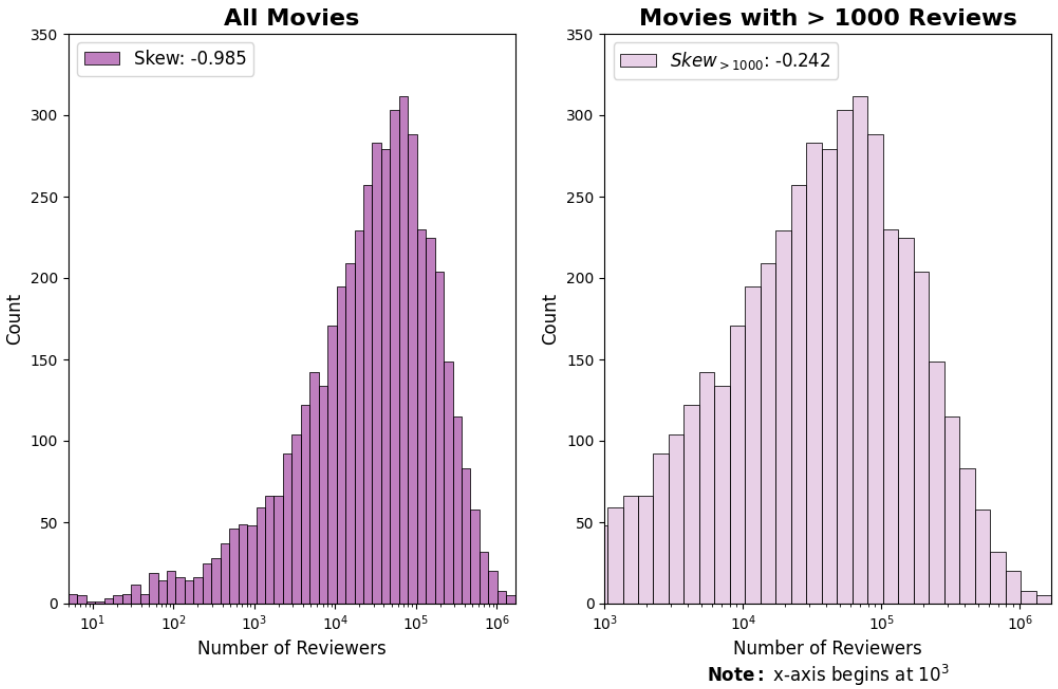



**Figure 2** shows an approximate log-normal distribution, thus removing values where $x < 3\sigma$ is inappropriate (*i.e.* $\bar{x} - \sigma = -55,658.79$). For *num_voted_users*, converting to a logarithmic x-axis shows the negative skew (**Figure 3**) is mainly due to reviews $\lessapprox$ 1000. Thus 364 rows with $\lessapprox$ 1000 reviews were removed.

## Missing data

Missing data in key features is problematic. Examining missing values in romance and horror movies (removing 17 romance-horror rows) and comparing them to the overall dataset avoids inadvertently removing these genres if they happened to contribute disproportionately to missing data.

**Table 2:** Top 5 Columns with the Most Nulls (Entire Dataset).

| Column Name | Nulls | Percentage of Total Rows |
|---|---|---|
| gross | 508 | 11.4 |
| budget | 306 | 6.9 |
| aspect_ratio | 115 | 2.6 |
| content_rating | 106 | 2.4 |
| plot_keywords | 32 | 0.7 |

**Table 3:** Top 5 Columns with the Most Nulls (Horror Subset, N = 474).

| Column Name | Nulls | Percentage of Total Rows |
|---|---|---|
| gross | 97 | 20.5 |
| budget | 24 | 5.1 |
| aspect_ratio | 14 | 3.0 |
| content_rating | 11 | 2.3 |
| plot_keywords | 2 | 0.4 |

**Table 4:** Top 5 Columns with the Most Nulls (Romance Subset, N = 1002).

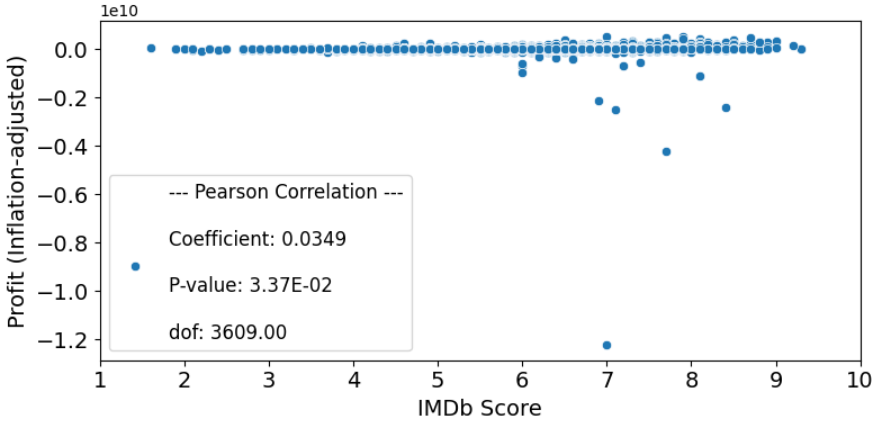
| Column Name | Nulls | Percentage of Total Rows |
|---|---|---|
| gross | 97 | 9.7 |
| budget | 71 | 7.1 |
| content_rating | 24 | 2.4 |
| aspect_ratio | 19 | 1.9 |
| plot_keywords | 5 | 0.5 |

**Table 3 & 4** show subsets aren't disproportionately representative of nulls. Removing them will not omit our target genres. *gross* and *budget* are our measure of profit when assessing the IMDb-profit assumption, so we'll remove null rows from these columns. Doing so drops 263.

### Assessing IMDb-profit Assumption

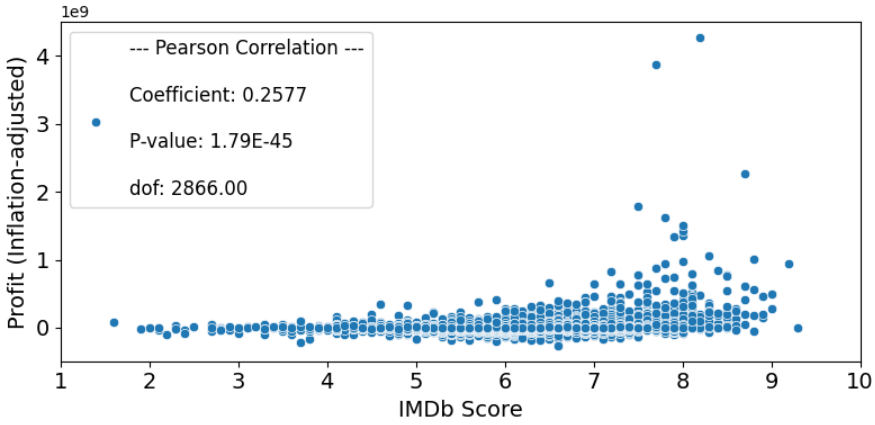
Creating a *profit* column (*gross* minus *budget*) and correlating it with *imdb_score*:

**Figure 4**



A near-zero correlation $r(3609) = .03,\ p < .05$. The lowest data point in Figure 4 shows the budget is in native currency ($budget = 12,215,500,000,\ gross = 2,201,412$), not USD. Furthermore, *gross* and *budget* aren't inflation-adjusted. After removing 778 non-US movies and adjusting values to 2023 USD using CPI data (Bureau of Labor Statistics, 2024), the correlation becomes stronger but still weak at 0.26, limiting conclusions.

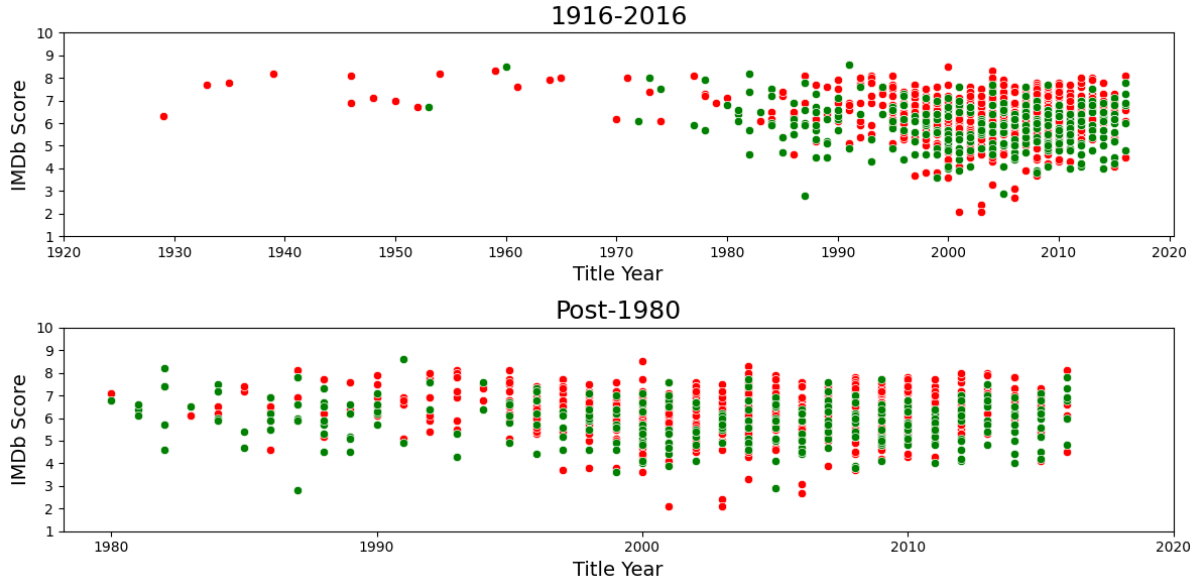
**Figure 5**




## Older Movies

I kept all release years since IMDb scores are consistent over time for both genres:

**Figure 6:** Scatter Plot Comparing IMDb Score Trends Over Time for Romance (red) and Horror (green) Movies



## Summary

42% of the data was removed, reducing from 5,043 to 2,918. This limits the analysis to the US market but greatly improves reliability.

# Results

## Top 10 Highest Rated Movies

**Table 5** shows the top 10 highest-rated movies, ranked first by IMDb score, then by review count.

**Table 5:** Top 10 Movies with the Highest IMDb Ratings.

| Rank | Movie Title | Year Released | IMDb Score | Number of User Reviews |
|---|---|---|---|---|
| 1st | The Shawshank Redemption | 1994 | 9.3 | 1,689,764 |
| 2nd | The Godfather | 1972 | 9.2 | 1,155,770 |
| 3rd | The Dark Knight | 2008 | 9.0 | 1,676,169 |
| 4th | The Godfather: Part II | 1974 | 9.0 | 790,926 |
| 5th | Pulp Fiction | 1994 | 8.9 | 1,324,680 |
| 6th | The Lord of the Rings: The Return of the King | 2003 | 8.9 | 1,215,718 |
| 7th | Schindler's List | 1993 | 8.9 | 865,020 |
| 8th | The Good, the Bad and the Ugly | 1966 | 8.9 | 503,509 |
| 9th | 12 Angry Men | 1957 | 8.9 | 447,785 |
| 10th | Inception | 2010 | 8.8 | 1,468,200 |

**Note:** Based on Movies with $N > 1000$ User Reviews, including non-US movies

## Top 5 Most Common Genres

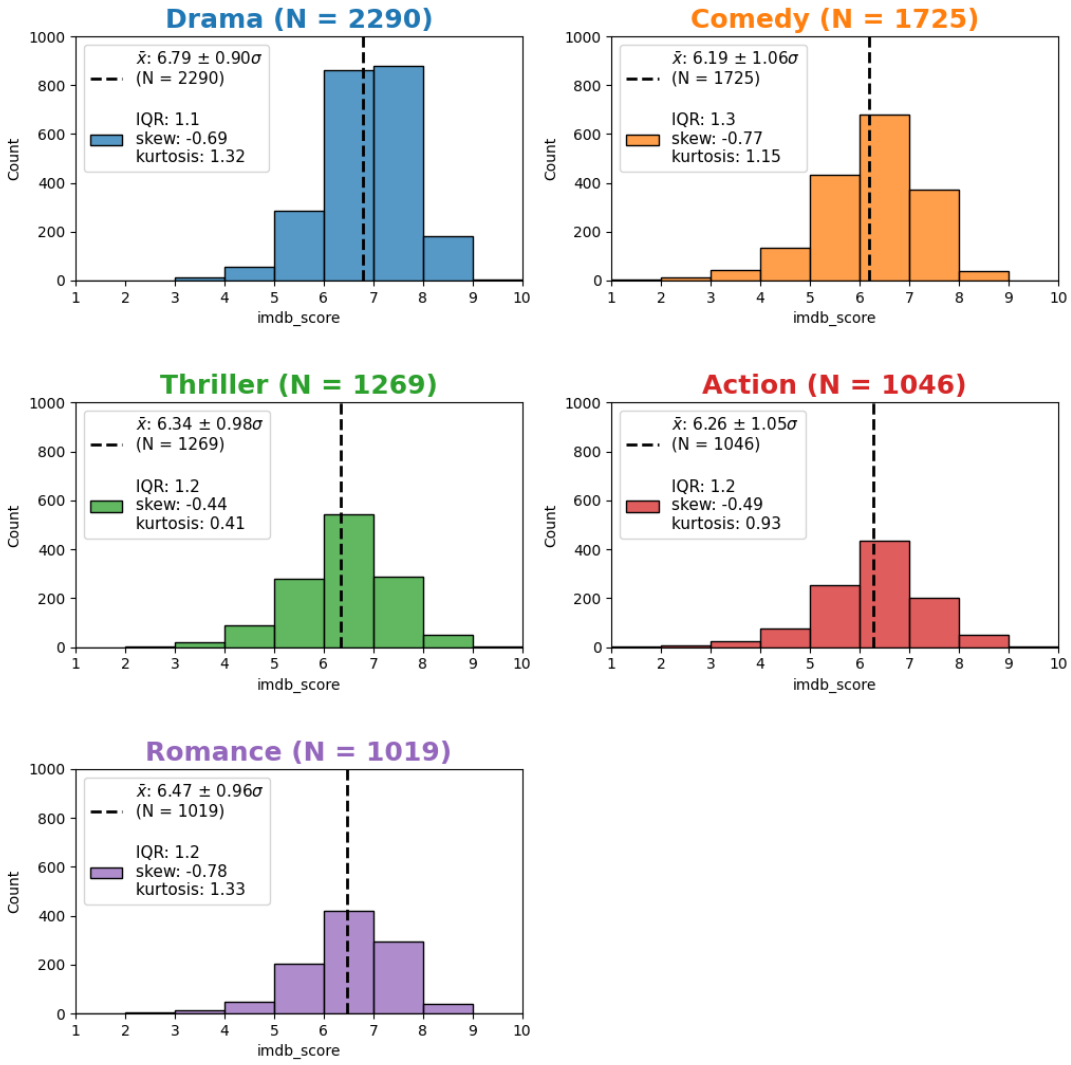
**Table 6** shows the top 5 genres by movie count, supported by **Figure 7** showing IMDb score distributions and summary statistics.

**Table 6:** Top 5 Individual Genres with the Most Number of Movies.

| Genre | Number of Movies |
|---|---|
| Drama | 2,290 |
| Comedy | 1,725 |
| Thriller | 1,269 |
| Action | 1,046 |
| Romance | 1,019 |

**Note:** Based on Movies with $N > 1000$ User Reviews, including non-US movies

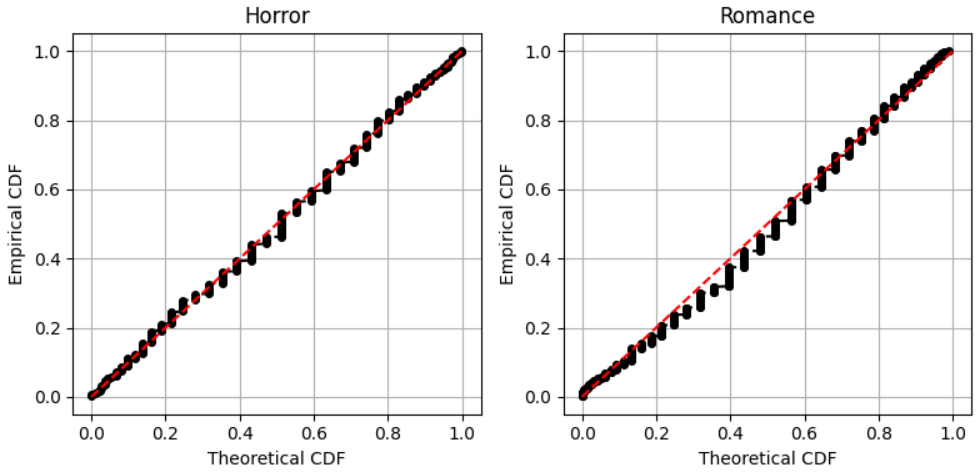
**Figure 7:** Histograms Displaying the Distribution of IMDb Scores for Each of the Top 5 Genres with the Most Movies, Along with Summary Statistics for Each Plot.



## Romance vs Horror: Two-Tailed Welch's t-test

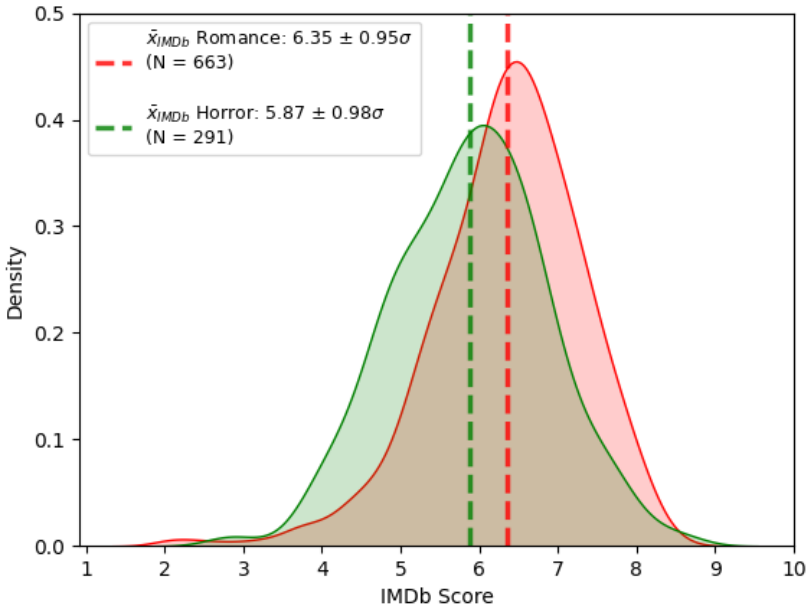
Given the IMDb distributions are approximately Gaussian:

**Figure 8:** P-P Plot Comparing the Empirical Cumulative Distribution Functions of Romance (red) and Horror (green) Movie IMDb Scores to a Theoretical Gaussian CDF




Welch's t-test shows a significant difference between IMDb scores for *romance* ($\bar{x} = 6.35,\ \sigma = 0.95,\ N = 663$) and *horror* ($\bar{x} = 5.87,\ \sigma = 0.98,\ N = 291$) movies, $t(537.40) = 7.03,\ p < .05,\ 95\%\ CI\ [0.346,\ 0.614]$. Rejecting the null hypothesis, *romance* movies have a significantly higher average IMDb.

**Figure 9:** KDE Plot Comparing IMDb Score Distributions of Romance (red) and Horror (green) Movies



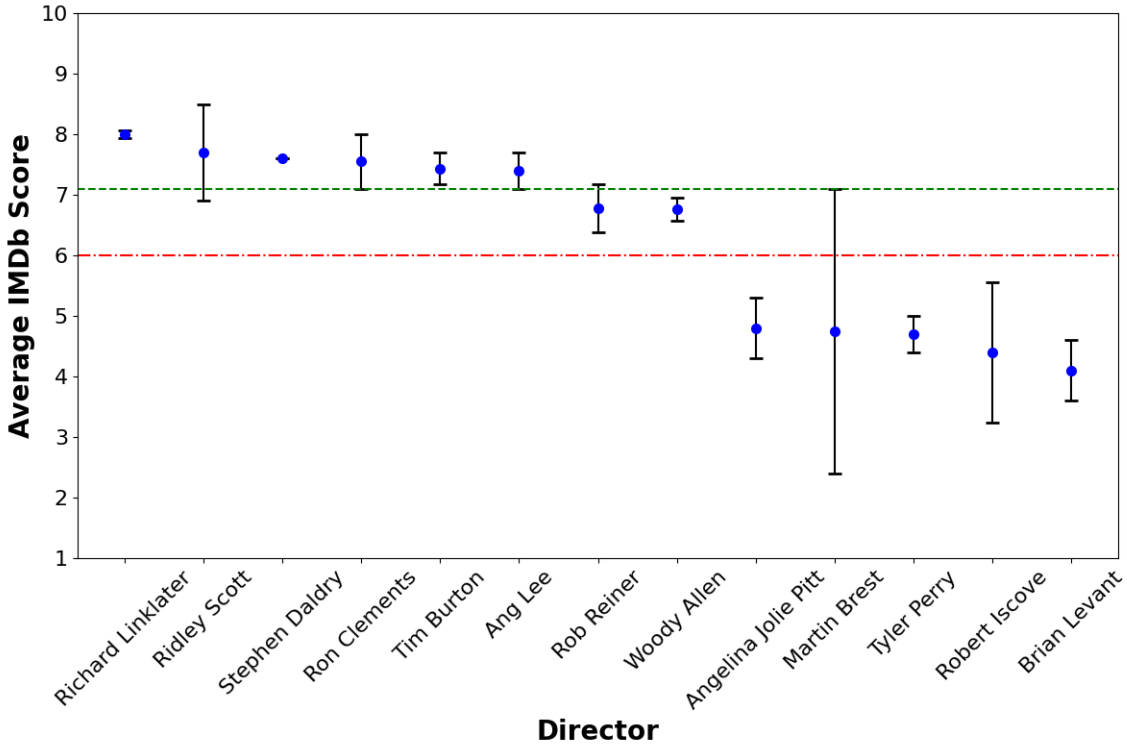
# Choosing a Director

### Two-tailed One-way ANOVA

ANOVA found at least 1 mean was significantly different $F(125, 199) = 1.89,\ p < 0.05$, thus the null hypothesis is rejected.

### Tukey's HSD

**Figure 10:** Mean Plot Comparing Average IMDb Scores of Directors Found to be Significantly Different by Tukey's HSD




No director "dominates," but there's a bifurcation (red line) into "higher" and "lower" IMDb averages. The error bars show *Richard Linklater*, *Stephen Daldry* and *Tim Burton* have significantly higher averages than the "lower" group (green line). Given *Linklater* has the highest, they are the recommended director.

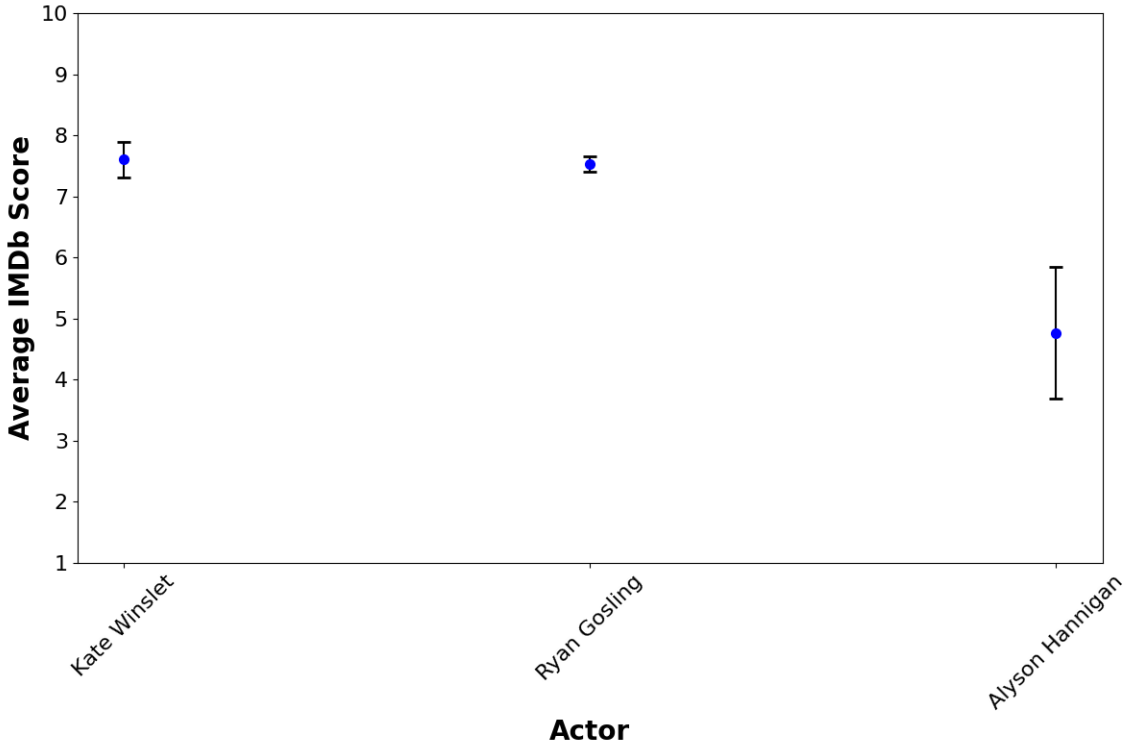
# Choosing a Lead Actor

### Two-tailed One-way ANOVA

ANOVA found at least 1 mean was significantly different $F(116, 246) = 1.81,\ p < 0.05$, thus the null hypothesis is rejected.

### Tukey's HSD

**Figure 11:** Mean Plot Comparing Average IMDb Scores of Actors Found to be Significantly Different by Tukey's HSD




Whilst *Kate Winslet* and *Ryan Gosling* have the highest average IMDbs, there is little to separate them. Given *Winslet's* average is slightly higher ($\bar{x} = 7.600$) than *Gosling's* ($\bar{x} = 7.525$), *Winslet* is the recommended actor.

# Conclusion

**Table 7:** Recommendations from the Analysis

| | |
|---|---|
| Genre: | Romance |
| Director: | Richard Linklater |
| Lead Actor: | Kate Winslet |

***Note:*** *This table was produced in the PDF's .tex file, not the .py script*

To conclude, with a cleaned sample of 2,918 US movies, a two-tailed Welch's t-test revealed *romance* movies - the recommended genre - have a significantly higher average IMDb score compared to *horror* movies. Furthermore, two-tailed one-way ANOVAs, followed by post-hoc Tukey HSDs provided *Richard Linklater* as recommended director, with *Kate Winslet* as recommended lead.

Such a triplet of suggestions is a best estimate for maximising profit from the next movie, with the following limitations kept in mind:

1. Conclusions only apply to the US market.
2. IMDb score shares a weak positive correlation with profit, and an undetermined causal relation.
3. Director/actor selection post Tukey's HSD arguably lacks statistical justification.
4. Only directors and actors who have previously appeared in romance movies were considered (since they have relevant data that can be analysed) - non-romance directors/actors may still outperform those recommended.

# References

Bureau of Labor Statistics Data. (2024). Bureau of Labor Statistics. https://data.bls.gov/timeseries/CUUR0000SA0

IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows. (2024). IMDb. https://www.imdb.com/