1 Introduction

This report introduces an interactive dashboard that visualises film data from 2007-2019, allowing users to explore trends, compare critic and audience scores, and analyse financial performance across genres. The goal is to make complex film data more accessible and insightful through engaging visualisations.

2 Tools/Technologies

I used Dash Python, a framework used to build data apps in Python, to create my dashboard of Film Data [1]. Preprocessing the data included converting any numerical columns to a numerical format that *pandas* recognises to allow for computations — such as calculating the mean of a column [2]. I converted everything to lowercase and removed any whitespace or irrelevant characters to allow for easy reading and computation of the data.

3 Dataset

The data set from which this dashboard is created is the Hollywood Insider data set, which contains film data from 2007 - 2019 [3]. This information includes ratings from various critic companies, financial performance, genre type, script type, whether the film won an Oscar, and associated Oscar details.

3.1 Dataset Overview

The following is a breakdown of the attributes and their potential data types:

Table 1: List of potential attributes and their data types

Attribute	Data Type	Description	Category
Film	String	The name of the movie	Nominal
Rotten Tomatoes critics	Integer	Critics' score on Rotten Tomatoes	Quantitative
Metacritic critics	Integer	Critics' score on Metacritic	Quantitative
Average critics	Float	Average of both critic scores	Quantitative
Rotten Tomatoes Audience	Integer	Audience score on Rotten Tomatoes	Quantitative
Metacritic Audience	Integer	Audience score on Metacritic	Quantitative
Rotten Tomatoes vs Metacritic deviance	Integer	Difference between critic scores	Quantitative
Average Audience	Float	Average of both audience scores	Quantitative
Audience vs Critics deviance	Integer	Difference between average audience and critic scores	Quantitative
Primary Genre	String	Primary genre of the film	Categorical
Genres	String	All genres associated with the film	Categorical
Script Type	String	Type of script, e.g., based on a true story	Categorical
Opening weekend ($million)	Float	Amount of money earned - opening weekend	Financial
Opening Weekend	Integer	Actual amount of money earned - opening weekend	Quantitative
Domestic gross ($million)	Float	Total earnings - domestic market (in millions)	Financial
Domestic Gross	Integer	Total earnings - domestic market (in dollars)	Quantitative
Foreign Gross ($million)	Float	Total earnings - international market (in millions)	Financial
Foreign Gross	Integer	Total earnings - international market (in dollars)	Quantitative
Worldwide Gross	Float	Total worldwide earnings (in millions)	Financial
Worldwide Gross ($million)	Integer	Total worldwide earnings (in dollars)	Quantitative
Budget ($million)	Float	Budget (in millions)	Financial
Budget recovered	Float	Percentage of budget recouped from earnings	Quantitative

Attribute	Data Type	Description	Category
Budget recovered opening weekend	Float	Percentage of budget recovered - opening weekend	Quantitative
Year	Integer	Release year	Temporal
Oscar Winners	String	Films that won an Oscar (if applicable)	Categorical
Oscar Detail	String	Details of the Oscars won (if applicable)	Categorical

3.2 Derived Attributes

A derived attribute, **Profit**, was calculated for each film by subtracting the **Budget** from the **Worldwide Gross**. This attribute was incorporated into the dashboard in several ways — including where the **Average Profit** for each genre was calculated annually — providing insights into genre-based financial performance. The **Profit** for individual films was visually represented in a custom graph, where the size of each film's data point could be enabled to be proportional to its profit.

3.3 Necessity of Visualisation

This dataset is complex due to its large **volume** (27 attributes and 1,694 rows, totalling 45,738 data points), which can overwhelm traditional analysis methods. As mentioned above, the **variety** of attribute types — ranging from financial figures to audience scores and film genres — creates a heterogeneous structure. The non-uniform distribution of values across these attributes, such as a few films having high earnings while the majority have lower earnings, further complicates the interpretation of the data. The **structure** of the dataset involves many attributes, meaning the multi-dimensional relationships require careful analysis, and its **visual complexity** risks over-plotting, making interpretation difficult.

4 Tasks

The following are some tasks that I support in my visualisation:

- **Compare**: The ability to analyse differences or similarities between films based on attributes like critic scores (eg. Rotten Tomatoes vs. Metacritic) or financial earnings across genres. The custom graph supports this by allowing users to compare multiple attributes, sort by genre, and visualise financial relationships, using size encoding for financial values.
- **Trend Analysis**: The ability to identify patterns over time, such as changes in critic scores, audience ratings, or profitability. Small multiples of scatter plots display the average profit per year by genre, helping users track profitability trends.
- **Correlation**: The dashboard helps identify relationships between different attributes, such as the correlation between critic scores and financial earnings, or between genre and profitability. The custom graph allows users to choose their attributes to compare and check if there is a correlation between attributes they find interesting.
- **Filter**: This allows viewers to choose to focus on a specific subset of data. The dashboard enables filtering data by genre, making it easier for them to isolate and identify trends or patterns.
- **Zoom**: Assists the user with examining data more closely. Zoom is especially useful for dense plots so viewers can examine the finer details more easily. Within the custom graph, the user can zoom in on a specific set of points and can hover over a point to see specific related attribute values and the name of the point's associated film.

5 Encoding Channels and Idioms

The main idiom used in my visualisation is a **dashboard layout**, which displays a variety of graph types to facilitate comparison and analysis. Below, I outline the encoding channels and idioms used, along with a justification for why each was chosen.

Small Multiples of Scatter Plots

A grid of scatter plots showing the average profit per year by genre, with each genre represented by a different colour. I encoded the, *year* and *average profit* attributes, by position within the scatter plot. Encoding the categorical attribute, *genre*, by colour allows for clear differentiation, allowing users to quickly compare trends over time while comparing genres. The use of small multiples allows for a **faceted view** of trends without overwhelming the user with a single dense graph.

Small multiples of Radar Charts

Each radar chart features four axes representing the Rotten Tomatoes critic and audience scores and the Metacritic critic and audience scores. These points are connected to form a polygon, highlighting a genre's

performance across these dimensions. **Colour encoding** is applied consistently across all graphs to represent genres, such as using red for the action genre. This uniformity helps users quickly identify and compare genres throughout the dashboard. Radar charts were chosen for their ability to allow users to see how a genre performs across all four critic dimensions at a glance, enabling comparisons within and across genres.

Cusotmisable Scatter Plot

A fully customisable scatter plot at the bottom of the dashboard allows users to select two attributes to plot on the x and y axes from dropdown menus. The user can also select a financial attribute from a dropdown menu to encode by size. Each data point is also encoded by **colour** based on its genre attribute. The custom scatter plot is flexible, supporting diverse comparison and correlation tasks. **Size** encoding allows financial metrics to stand out, while interactivity (e.g., dropdown menus, filtering, zooming, and hovering) provides a personalised exploration of the dataset.

Dynamic Table

For certain attribute combinations, such as when users choose "Film" and "Oscar Winner," a dynamic table is displayed instead of a scatter plot. This table lists the selected attributes and includes sorting functionality by genre. A table was chosen for its simplicity and clarity, making categorical data easier to interpret than visual encodings.

Interactivity

Interactivity is incorporated to enhance usability, preventing over-plotting and enabling the dashboard to be effective for broad overviews and detailed analysis. The main dropdown menu—choosing genres to display—works across all three visuals allowing for continuity and ease of understanding. Dropdown menus allow users to control the display, reducing visual clutter and enabling focused analysis. Zoom and hover functionalities address the issue of dense plots, giving users the ability to inspect individual data points or details without overwhelming the visualisation. The toggle button allows for further filtering and sorting by the genres chosen. Users can toggle specific genres on or off using the legend, focusing only on the data that interests them.

6 Novelty

This visualisation offers a more detailed and interactive approach compared to the *Information is Beautiful* visualisation [4]. While their version uses a single customisable graph, this is a dashboard that incorporates multiple visual elements, including small multiples of scatter plots and radar charts, allowing users to compare trends across genres, years, and attributes. The added interactivity—such as genre filtering, zooming, and the option to display data in a table—provides a more flexible and user-friendly experience, offering deeper insights into the dataset.

7 Critical Analysis

7.1 Strengths

1. **Interactivity and Flexibility**: One of the major strengths of this visualisation is its high interactivity. Users can filter data by genre, zoom in on specific points, and customize the display to focus on attributes that interest them making the tool adaptable to a wide range of user needs and preferences.
2. **Multiple View for Comparison**: The use of multiple visualisation types such as small multiples of scatter plots, radar charts, and a customisable scatter plot, allows users to analyse the dataset from different angles, offering flexibility in understanding relationships between attributes.
3. **Clear Visual Encoding**: The visual design —including colour coding for genres and size encoding for financial attributes — helps users quickly interpret the data. The use of the same colour for a genre across all visuals allows for continuity making it easier for users to understand the data and make quick comparisons.

7.2 Weaknesses

1. **Complexity for New Users**: While interactivity is a strength, it can also become a barrier for less experienced users. The variety of options and filters may overwhelm users, leading to a steeper learning curve.
2. **Potential for Overwhelming Data**: With multiple views displaying large amounts of data, there is a risk of overloading users. If they choose too many genres at once, it can result in cluttered plots. For this reason, there is a limit of five selected genres at a time.

Overall, the visualisation offers a powerful and interactive way to explore the dataset, but its complexity and potential for overwhelming users could be limiting factors.

8 Video and Repository Link

Please click here for the unlisted video uploaded to YouTube, or paste this link into your browser of choice: https://youtu.be/RaYIqv3fjxo

Please click here for the link to the GitHub repository that contains the source code for this project, or paste this link into your browser: https://github.com/daniellebuggle/BoxOfficeBreakdown

References

[1] Plotly Technologies Inc. Dash by plotly, 2024. URL https://dash.plotly.com/?_gl=1*1emj2k*_gcl_au*MTU3NTEzNDMyLjE3MzE2MTIzMjI.*_ga*MTEyNTEzNzk1NS4xNzMxNjEyMzIy*_ga_6G7EE0JNSC*MTczMzQ4NDczMS4xNS4xLjE3MzM0ODgzMjcuNjAuMC4w. Accessed: 2024-12-06.

[2] The Pandas Development Team. Pandas, 2024. URL https://pandas.pydata.org/. Accessed: 2024-12-06.

[3] Google Sheets. Hollywood insider dataset - films, 2024. URL https://docs.google.com/spreadsheets/d/12bnGB7w5T03f7Bq1PG6gqFIdsLccdP6E-86xdVjbCOc/edit?gid=1382588702.

[4] David McCandless and Information is Beautiful. The hollywood insider, 2024. URL https://informationisbeautiful.net/visualizations/the-hollywood-insider/?_gl=1*rwbh2n*_ga*MTgyNjExNzQ1OC4xNzMwNzI0ODA5*_ga_73FXQD7Q23*MTczMzUwNjE3OC40LjAuMTczMzUwNjE3OC42MC4wLjA. Accessed: 2024-12-06.

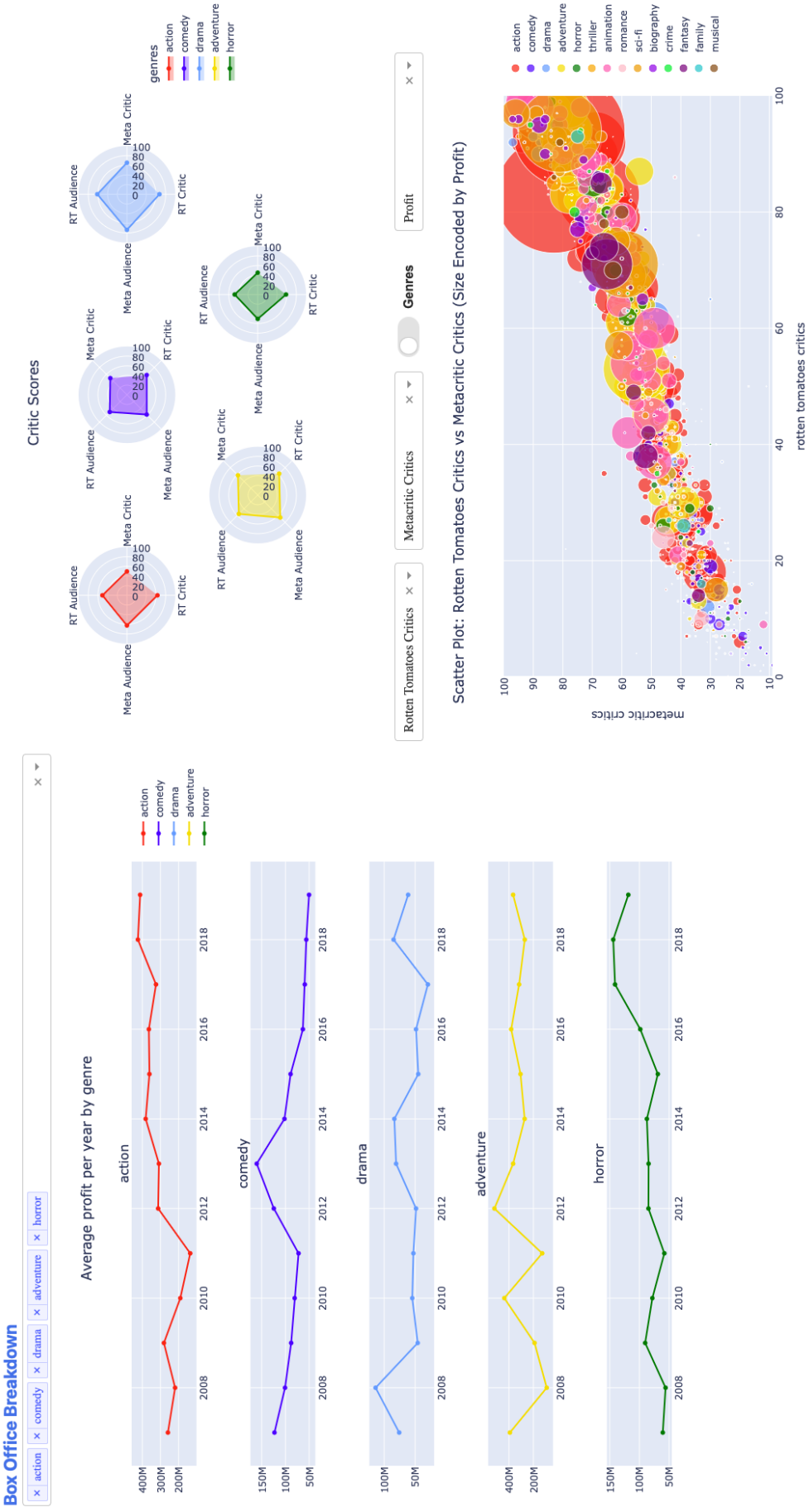

Figure 1: Box Office Breakdown