

Data Analysis of Hollywood Movies

Evangeline Mapstone

A bowl of popcorn and a cup of soda

Description automatically generated

The assignment is to analyse the performance of Hollywood movies, to start we download the data from the link provided.

A screenshot of a computer

Description automatically generated

The data above displays the title, genre, lead studio, audience score %, profitability, rotten tomatoes % , worldwide gross and year of release for each Hollywood movie 2007-2012.

A screenshot of a computer

Description automatically generated

Import the data set into R Studio for data cleaning. I have set the data frame name to HDF (for Hollywood Data Frame), making the title shorter and easier for Data transformation in RStudio.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**View the Hollywood Movies dataframe (HDF).**

The above images display the original data now imported into R Studio ready to be analysed and cleaned.

A screenshot of a computer

Description automatically generated

Install the Tidyverse Library from Packages, to transform the data. The tidyverse library assists with data import, tidying, manipulation, and data visualisation.

**A screenshot of a computer

Description automatically generatedView the data frame and then check the data types.**

Columns included in the data set: title, genre, lead studio, audience score %, profitability, rotten tomatoes % , worldwide gross and year of release for each film 2007-2012.

**Check the data types**:

The rows and columns displaying the data are known as observations and variables, there are 74 observations and 8 variables in this data frame.

A screen shot of a computer

Description automatically generatedThe data types include in the data frame are Character, Numeric and Integer.

**Check for any missing values**

A close-up of a white background

Description automatically generatedA close up of a text

Description automatically generated

Missing values in the dataset: 3 for Profitability, 1 for Audience score and 1 for Rotten tomatoes score.

A close up of a text

Description automatically generated

A screen shot of a movie

Description automatically generatedRemove the missing values. Returns the dataset with the missing values removed.

**A white background with black text

Description automatically generatedCheck to make sure that the rows have been removed.**

A screenshot of a computer

Description automatically generated

The na.omit() function is used to drop the missing values from the dataset. To make sure that these are removed I created a new table; cleaned\_hdf = na.omit(HDF). This then removes all the rows which have missing values. Only four rows were removed as one of the rows had two missing values.

**Summary Statistics**

A close up of text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

The Summary returns the min, median, mean, max of the numerical data within the data frame. It also includes the 1st and 3rd quartile of each section of data. The quartiles show the data spread i.e variation, which along with the median are not affected by any outliers in the data set.

**Data visualisation in R Studio**

Create a scatterplot to show the range of rotten tomatoes scores for each studio

A black text on a white background

Description automatically generated

A graph of a chart

Description automatically generated with medium confidence

The scatterplot shows that there are different ranges of rotten tomatoes scores for each of the lead studios. For example, Independent has a wide range of rotten tomatoes scores for their films whereas The Weinstein Company has a smaller range of higher scores so it looks like a successful studio, however there are only three points for The Weinstein Company this is a very small amount of data so any conclusions taken from this scatterplot may be unreliable.

A black and white text

Description automatically generated

Create Bar chart to see the number of films released per year. This chart does not show any trend in the number of films released per year.

A graph of a graph showing the number of years

Description automatically generated with medium confidence

**Export cleaned data**

Now that the data has been cleaned and checked it is ready to be exported. The following command is used for this.

A black and red text

Description automatically generated

A screenshot of a computer

Description automatically generatedI need to make sure I know where the cleaned data is located as this information will be needed when I import it into PowerBI for visualisation.

Import the cleaned data into PowerBI as a Text/CSV.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

While creating my dashboard I realised that the data type for the Year column was number and would be used as such in Power BI so I had to convert this data into the correct type. I found three ways of doing this, the first was to convert the data type from integer to character in R studio. I would then have to re-export and import into Power BI.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe second way was to go into the Power Bi query editor and change the data type in the year column to Date

The simplest way was to go to the Model view of PowerBI, open the Properties panel and under Formatting set the data type to date and set the Date time format to Year.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Visualisation in PowerBI**

I used the visualisation tools in PowerBI to create my dashboard to show my charts used to analyse the performances of Hollywood movies.

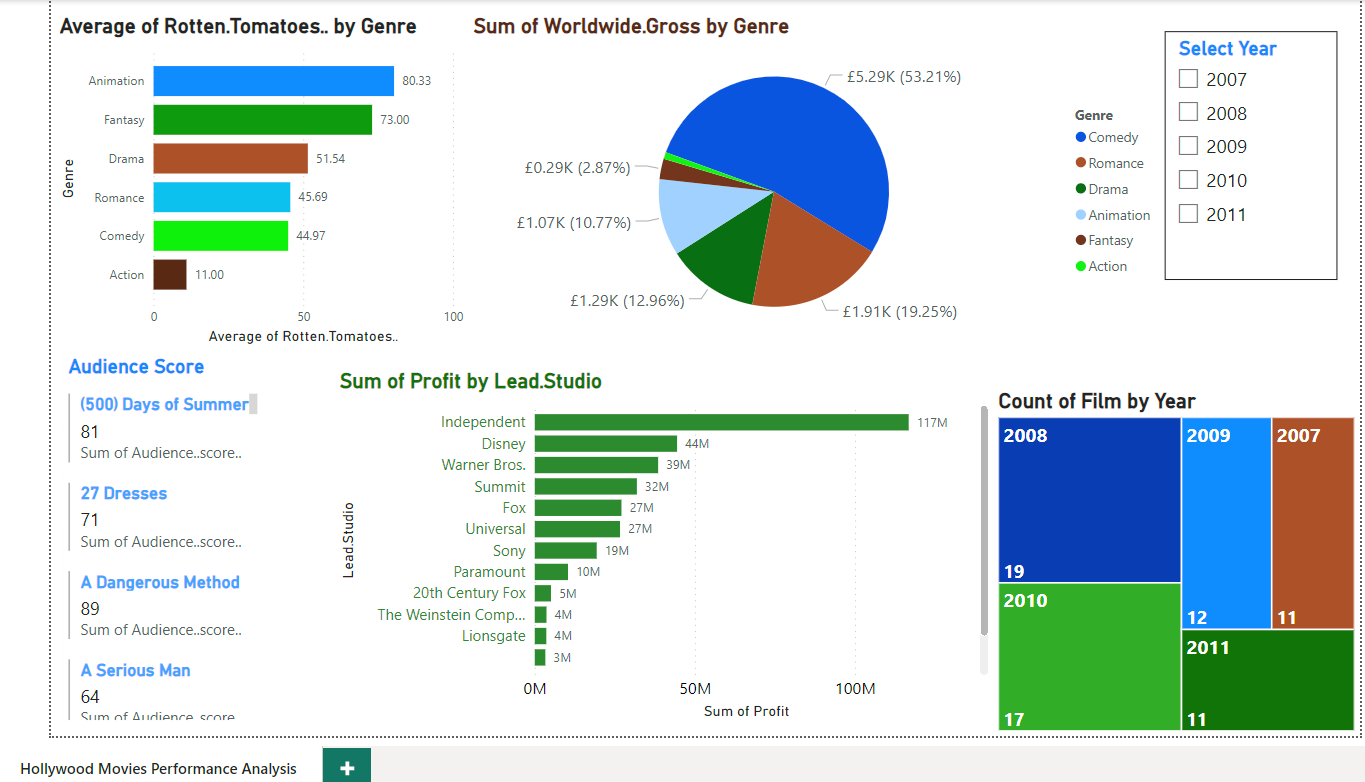
A screenshot of a computer

Description automatically generatedTo achieve this I need to select the right chart that could be used to effectively display data required by the client using the Build Visual panel. I then needed to format the visuals, think about the size of the charts, the values, the titles, making sure everything is clear and easy to read. In terms of colours the client’s request was green, blue and brown so I used different shades of each in my final dashboard.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**The Final Dashboard for the Hollywood Movies Performance Analysis**

The final dashboard displays the data that has been imported and cleaned in R Studio. The cleaned data has then been imported into Power BI where visualisation tools have been used to display the data to make it more appealing and simpler for the client to assess the performances of Hollywood movies.

For this assignment the client has requested the following data to be displayed:

1. The average Rotten Tomatoes ratings of each genre
2. The number of movies produced per year
3. The audience score for each film
4. The profitability per studio
5. The worldwide gross per genre

I have used different visuals for each of these, a stacked bar chart for the rotten tomatoes score of each genre, a treemap displaying the number of movies released each year from 2007-2011. For the audience score of each film I chose the multi-row card which is useful for a longer list of data. A clustered bar chart displaying the profit made by the lead studios and a pie chart for the worldwide gross made from each genre. On the top right hand corner there is also a slicer which can be used to analyse the data in each year.

An interactive exploration of this dashboard is accessible via the following link:

[Hollywood Movies Performance Analysis - Power BI](https://app.powerbi.com/groups/me/reports/5123719c-d18e-40e0-9176-d64d0778fe9e/ReportSection?experience=power-bi)