**Dataset**

The dataset chosen in this coursework comes from Kaggle and is titled “**IMDb Movie Dataset: All Movies by Genre**”. It was submitted by Chidambara Raju G and **Version 3** of the dataset was used. It is stored as a zip file named **archive.zip**.

The dataset comes with 16 csv files which are named: **action.csv**, **adventure.csv**, **animation.csv**, **biography.csv**, **crime.csv**, **family.csv**, **fantasy.csv**, **film-noir.csv**, **history.csv**, **horror.csv**, **mystery.csv**, **romance.csv**, **scifi.csv**, **sports.csv**, **thriller.csv** and **war.csv**. Each csv file is named after a movie genre. Other movie genres such as drama and comedy exist but are not considered since they are not named.

Each csv file contains 14 columns which are named **movie\_id**, **movie\_name**, **year**, **certificate**, **runtime**, **genre**, **rating**, **description**, **director**, **director\_id**, **star**, **star\_id**, **votes** and **gross (in $)**.

Each movie can have 1 – 3 different genres associated with it and will appear in the corresponding csv files. As an example, a movie with the action and adventure genres will appear in action.csv and adventure.csv.

**Data Preprocessing**

* **Movie Extraction**

First, the list of csv files in **archive.zip** is obtained. We then loop through the list while extracting each csv file from the zip file into a data frame.

The columns **movie\_id**, **description**, **director\_id** and **star\_id** was removed from the data frame for being uninformative while the column **gross (in $)** was removed for having too many missing values.

Each processed data frame is then appended to the **Movies** data frame. Each csv file name was also saved to the **genres** list.

* **Runtime Processing**

The **runtime** column of the **Movies** data frame was originally stored as strings in the form “x min” which slowed comparisons.

The string “min” was removed from the column then the whole column was converted to integer for easy comparisons.

* **Year Processing**

The **Movies** data frame originally contains movies which have not been made. These movies can be easily identified by the values in their **year** column. Their years were either stored as roman numerals, had a number larger than 2023 or were NA.

The **year** column of the**Movies** data frame was converted to integer. This causes all roman numerals to be converted to NA. Then, any row with a number larger than 2023 or NA in their **year** column were removed to create a new, valid data frame.

* **Genre Processing**

The **genre** column of the **Movies** data frame was originally stored as strings separated by a comma between each individual genre. A movie can have 1 – 3 different genres.

The **strsplit** function was used to split the **genre** column into lists of genres. The number of genres in each list was uneven and can range from 1 – 3. Each list was then padded with NA to an equal length of 3.

The lists of genres were used to create 3 new columns in the **Movies** data frame named **genre1**, **genre2** and **genre3**. The original **genre** column was then removed.

* **Duplicate Removal**

The **Movies** data frame was created by appending 16 different csv files which contains duplicates. The data frame shrinks from 300k to 200k rows after their removal.

* **Certificate Processing**

The **certificate** column of the **Movies** data frame originally contains many different certifications from different agencies and time periods which need to be processed.

To facilitate easier analysis, only certification with at least 500 occurrences were considered. This leaves certification such as **G**, **R**, **PG**, **PG-13**, **TV-14**, **TV-MA**, **TV-PG**, **Passed** and **Approved**.

Certification which was empty, **Not Rated** and **Unrated** were all listed as **Unrated**. All other certification was listed as **Other**.

**Data Transformation**

The **Movies** data frame was then factored by **rating**, **runtime**, **votes**, **decade**, **genres** and **certificate** into different data frames.

Each data frame has 9 columns. The first 3 columns hold the **factors**, movie **total** by factor and movie **percentage** by column.

The last 6 columns hold lists produced when the factored movies are further divided by **rating**, **runtime**, **votes**, **decade**, **genres** and **certificate**. Lists were used instead of columns to reduce the number of columns needed to store the data.

Only 5 out of the 6 columns are useful as a table divided by its own factors returns itself. However, the extra column is saved for completeness during graph plotting.

* **Rating**

**Rating** was divided into factors of 2. By including NA, we get 6 factors which are **0 – 2**, **2 – 4**, **4 – 6**, **6 – 8**, **8 – 10** and **NA**.

* **Runtime**

**Runtime** was divided into factors of 30 minutes. The last factor includes numbers up to Inf. By including NA, we get 10 factors which are **0 – 30**, **30 – 60**, **60 – 90**, **90 – 120**, **120 – 150**, **150 – 180**, **180 – 210**, **210 – 240**, **240 – Inf** and **NA**.

* **Votes**

**Votes** was divided into factors of powers of 10. By including NA, we get 8 factors which are **1 – 10**, **10 – 100**, **100 – 1000**, **1000 – 10000**, **10000 – 100000**, **100000 – 1000000,** **1000000 – 10000000** and **NA**.

* **Decade**

**Decade** was divided into factors of 10 years. With no values of NA, we get 13 factors which are **1900 – 1910**, **1910 – 1920**, **1920 – 1930**, **1930 – 1940**, **1940 – 1950**, **1950 – 1960**, **1960 – 1970**, **1970 – 1980**, **1980 – 1990,** **1990 – 2000**, **2000 – 2010**, **2010 – 2020** and **2020 – 2030**.

* **Certificate**

**Certificate** is a categorical variable which is divided into 11 factors which are **G**, **R**, **PG**, **PG-13**, **TV-14**, **TV-MA**, **TV-PG**, **Passed**, **Approved**, **Unrated** and **Other**.

* **Genres**

**Genres** is also a categorical variable which is divided into 16 factors which are **Action**, **Adventure**, **Animation**, **Biography**, **Crime**, **Family**, **Fantasy**, **Film-Noir**, **History**, **Horror**, **Mystery**, **Romance**, **Sci-Fi**, **Sports**, **Thriller** and **War**.

For **Genres**, 3 different data frames were created which are **Genre1**, **Genre2** and **Genre3**. The number beside Genre denotes the number of genres which form a set to be matched with the **Movies** data frame.

**Genre1** has 16C1 = 16 sets, **Genre2** has 16C2 = 120 sets while **Genre3** has 16C3 = 560 sets.

Only subsets of each set return TRUE when matched. As an example, {Action, Crime} returns TRUE for {Action}, {Crime} and {Action, Crime} but returns FALSE for {Action, Adventure, Crime}.

**Data Visualization**

* **Plot**