**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 **genre** column was split 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 for easy processing.

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.

Only certification with at least 500 occurrences were eligible to be studied. 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 were 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 supersets of each set return TRUE when matched. As an example, {Action, Crime} only returns TRUE for any sets which contains the two elements and FALSE otherwise.

**Data Visualization**

* **Plot**

The first tab allows us to visualize a line graph using 5 different settings which are **Plot Type**, **Year Range**, **First Genre**, **Second Genre** and **Third Genre**.

**Plot Type** allows us to choose the type of data to be plotted. The 4 choices available are **Count** which is the total number of movies made per year, **Rating** which is the average rating of movies per year, **Runtime** which is the average runtime of movies per year and **Votes** which is the average number of votes of the movies per year.

**Year Range** allows us to specify the range of years to be plotted on the line graph.

**First Genre**, **Second Genre** and **Third Genre** lets us filter the movies to be plotted. Only movies with supersets of the chosen genres are plotted. There are 17 choices of genres, 16 of which are the csv file names while the last one is a “-”. The “-” is treated as an empty set, {}. Duplicate choices are only counted once. As an example, {Action} and {Action} combine to be just {Action}.

* **Table**

The second tab allows us to present the data in a data table. There are 3 different settings which are **Table**, **Column** and **Type**.

**Table** allows us to choose the primary factorization of the **Movies** data frame. The choices given are **Certificate**, **Decade**, **Genre1**, **Genre2**, **Genre3**, **Rating**, **Runtime** and **Votes**.

**Column** allows us to choose the secondary factorization of the **Movies** data frame. The choices given are **rating**, **genres**, **runtime**, **votes**, **decade** and **certificate**. As mentioned above, a table divided by its own factor returns itself. So, only 5 out of the 6 choices given are meaningful.

**Type** allows us to choose the type of data to be visualized. There are 5 choices which are **total**, **original**, **row**, **column** and **order**.

**Total** displays the **total** number of movies for each primary factor and its **percentage** by column. **Original** displays the **total** number of movies for each secondary factor while **row** and **column** convert the **original** data to percentages by row and column respectively. **Order** orders the data by its percentage by row. The secondary factor is also given beside the percentage.

* **Chart**

The third tab allows us to visualize the data in a bar chart. There are 3 different settings which are **Table**, **Column** and **Position**.

**Table** and **Column** are similar to its instance in the second tab. They each allow us to choose the primary and secondary factorization of the **Movies** data frame respectively. The available choices of **Table** and **Column** are also similar to its instance in the second tab.

**Position** allows us to choose the method of visualization of the bar chart. The 3 choices available are **fill**, **stack** and **dodge**.

**Fill** allows us to visualize the percentage of each secondary factor in each primary factor. Normally, the bars add up to 1.00 or 100%. However, an empty bar signifies that there are no instances of the primary factor.

**Stack** allows us to visualize the actual number of each secondary factor in each primary factor. The bars are stacked on top of each another and we can clearly see the proportions of each primary factor.

**Dodge** also allows us to visualize the actual number of each secondary factor in each primary factor. However, the difference between **dodge** and **stack** is that the bars of the secondary factor are arranged next to each other instead of being in a stack.

**Questions**

1. **What is the general trend of the movies of the line graph?**

Generally, the trend of the movies tend to be dominated and decided by major genres such as **Action**, **Adventure** and **Romance**.

The number of movies made each year generally increases. This is expected due to technological advancements which make movie production easier. The increase in the number of humans and disposable income each year causes the potential market and the average gross for each movie to increase. These factors thus incentivize movie production.

There are 2 distinct periods which can be seen from the line graph which are movies made before and after the year 2000. The line of best fit for the number of movies made before year 2000 is flatter while the gradient of the line after year 2000 is much steeper. This can be attributed to the advancements in 3D technology and the explosion in the release of IMAX films.

For average **ratings**, movies made before 1920 tend to have more volatile ratings. After the year 1920, the average ratings have stabilized around the value 6.

For average **runtime**, movies made before 1920 are also much more volatile before stabilizing around 100 minutes. This is probably due to diminishing returns. A longer movie would need more money to produce while making less money in return.

For average number of **votes**, the value slowly increases before reaching a peak around the year 2010. After that, the average number of **votes** drops sharply. This might be due to the lost in trust in the website with the rise of fake reviews and review bombing. So, less people would use and vote on the website.

1. **Are there any outliers in the movies of the line graph?**

There is a sudden drop in the number of movies made around the year 2020. This is caused by the COVID-19 pandemic which imposed strict restrictions on travel and promotes social distancing. This causes the number of movies made during that period to drop due to the lower expected market for movies. Fortunately, after the lifting of COVID-19 restrictions, movie production has risen back to pre-pandemic levels.

There is also another drop is the year 2023. However, this is expected as the dataset only covers movies made until the first quarter of 2023 so the data is incomplete.

Movies with the **Film-Noir** genre are found only during the years 1920 – 1960 and have too small a sample for accurate analysis. So, the behavior exhibited may seem erratic.

For movies with the genre **War**, movie production typically increases after major conflicts such as the First World War (1910s) and the Second World War (1940s). After that, movie production drops before rising after the next major conflict.

1. **What are the most and least popular genre combinations?**

For movie genre analysis, the percentages are calculated based on the total number of movies which is 207919 and not the column sum. This is because a movie can have several genres so there are duplicate entries.

For single combination, **Romance** has the highest percentage of movies made at 22.83%. This is followed by **Action** and **Crime** at 21.96% and 16.36% respectively. The 7th entry onwards all have percentages less than 10% while the lowest three genre are **War** at 3.58%, **Sport** at 2.21% and **Film-Noir** at 0.88% which is the only genre to have less than 1% of the total percentage.

For double combination, {**Action**, **Crime**} has the highest percentage at 4.28% followed by {**Action**, **Adventure**} at 3.95% and {**Horror**, **Thriller**} at 2.89%. The 15th entry onwards all have less than 1% of the total percentage. Only 4 combinations never occur which are {**Animation**, **Film-Noir**}, {**Family**, **Film-Noir**}, {**Film-Noir**, **History**} and {**Film-Noir**, **War**}.

For triple combination, the top 3 entries are {**Horror**, **Mystery**, **Thriller**} at 0.54%, {**Action**, **Crime**, **Thriller**} at 0.53% and {**Action**, **Adventure**, **Animation**} at 0.50%. After this point, all entries have less than 1000 movies. The 303th entry onwards or the last 258 combinations never occur.

1. **What are the most popular movie genres of each rating level?**

For terrible movies rated at **0 – 2**, the category is dominated by the genre **Horror** followed by **Action** and **Adventure**. Similarly, bad movies rated at **2 – 4** are also dominated by the genre **Horror** and **Action** with third place taken by **Thriller**. This implies that horror movies are more likely to be bad which is possible since horror films are not known for their good writing and are characterized by its excessive use of loud noises and jump scares.

For good movies rated at **6 – 8**, the genre **Romance** takes the top spot followed by **Crime** and **Action**. For excellent movies rated at **8 – 10**, first place is also held by the genre **Romance** followed by **Action** and **Thriller**. This shows that romance movies are more likely to be rated higher by viewers. This is because humans are social creatures who seek companionship.

Finally, for average films rated at **4 – 6**, the top genre is **Action** followed by **Romance** and **Thriller**. For **unrated** films, the genre **Action** is also on top followed by **Romance** and **Crime**. The reason that the genre **Action** appears in the top 3 of each rating level is because action movie is an umbrella term which covers a wide variety of movies which may be good or bad and thus is represented in all rating levels.

1. **What are the most popular movie genres of each decade?**

For the 1900s, **Biography** was the most popular genre followed by **Adventure** and **History**. This changed in the 1910s where **Crime** took the top spot followed by **Romance** and **Adventure**. The 1920s saw **Crime** drop to 3rd place while **Romance** and **Adventure** rise to 1st and 2nd place.

For the 1930s, 1940s and 1950s, **Romance**, **Crime** and **Adventure** became the 3 most popular genres. This was followed by the 1960s where **Action** shot up to 2nd place, pushing **Crime** and **Adventure** to 3rd and 4th place. **Romance** still retains the top spot.

The 1970s, 1980s and 1990s also saw dominance by the same 3 genres which were **Action**, **Romance** and **Crime**. The streak was finally broken in the 2000s by the genres **Romance**, **Action** and **Thriller**.

The 2010s had the same 3 genres as the 2000s but in reverse which were **Thriller**, **Action** and **Romance**. Finally, the 2020s up until 2023 have **Thriller**, **Horror** and **Action** as their top 3 most popular genre.

1. **What are the most popular movie genres of each certification?**

To ensure better understanding, only the most popular movie certification will be analyzed which are **G**, **PG**, **PG-13** and **R**.

The other certifications such as **TV-14** and **TV-MA** are too obscure or cannot be properly separated such as **Unrated** and **Other** and thus are not analyzed here.

The **G** (General Audience) certification is for movies available to all ages. The top 3 genres are **Adventure**, **Family** and **Animation**. This is expected as most **G** movies are likely family-friendly cartoons.

The **PG** (Parental Guidance Suggested) certification also have the same top 3 genres as **G** which are **Adventure**, **Family** and **Animation**. **PG** movies are mostly the same as **G** movie but with the inclusion of cartoon violence and hidden adult humor.

The **PG-13** (Parental Guidance Strongly Cautioned) certification is dominated by the genres **Action**, **Romance** and **Adventure**. The movies have minimal swearing and mild depictions of violence.

The **R** (Restricted) certification is for movies with violence, gore and sexual themes. The genres most prevalent in **R** movies are **Thriller**, **Action** and **Crime**.

**Discussion and Reflection**

* **Visualization**

For data visualization, 3 methods were used which were line **plot**, **table** and bar **chart**.

Line **plot** was chosen to study the trend of the movie **count**, average **rating**, **runtime** and **votes** over the years. Outliers of the graph are also much easier to spot.

**Table** was chosen to present the actual values of the data and also their percentages. The ability of the table to sort by column was also important in analyzing the data.

Bar **chart** was chosen to visualize the data in the table. The setting **fill** was used to show the percentage of the factors in each bar. The setting **stack** and **dodge** was used to show the actual proportion on the data in a bar and also between bars.

* **Questions**

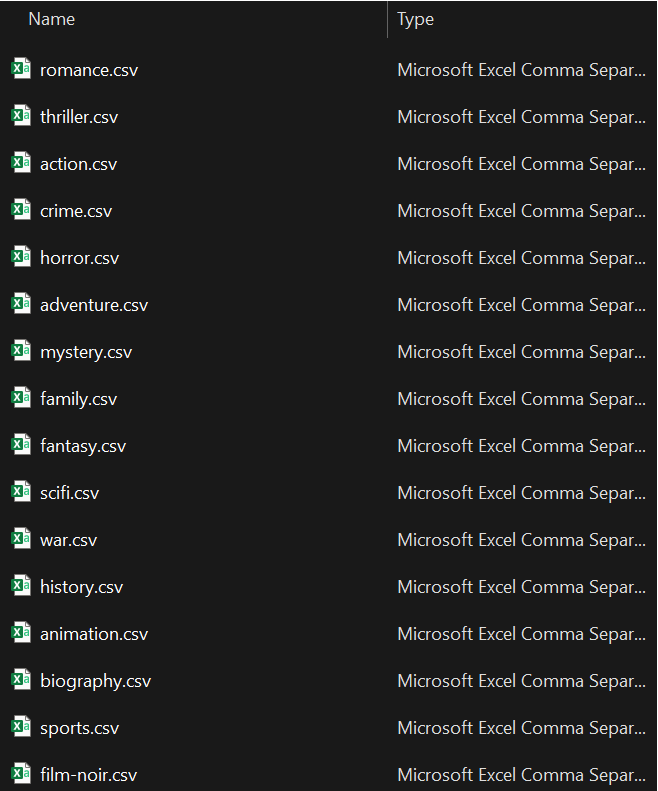
While each question was expected to have their own data cleaning and transformation steps, the data was instead cleaned and transformed only once during the beginning. This has streamlined the process of studying the data and forming questions. There are also no distinction between initial and further questions.

The code has also been transformed into a Shiny app. This is to encourage exploration and interactivity for the users so that they can form their own questions and find their own answers.

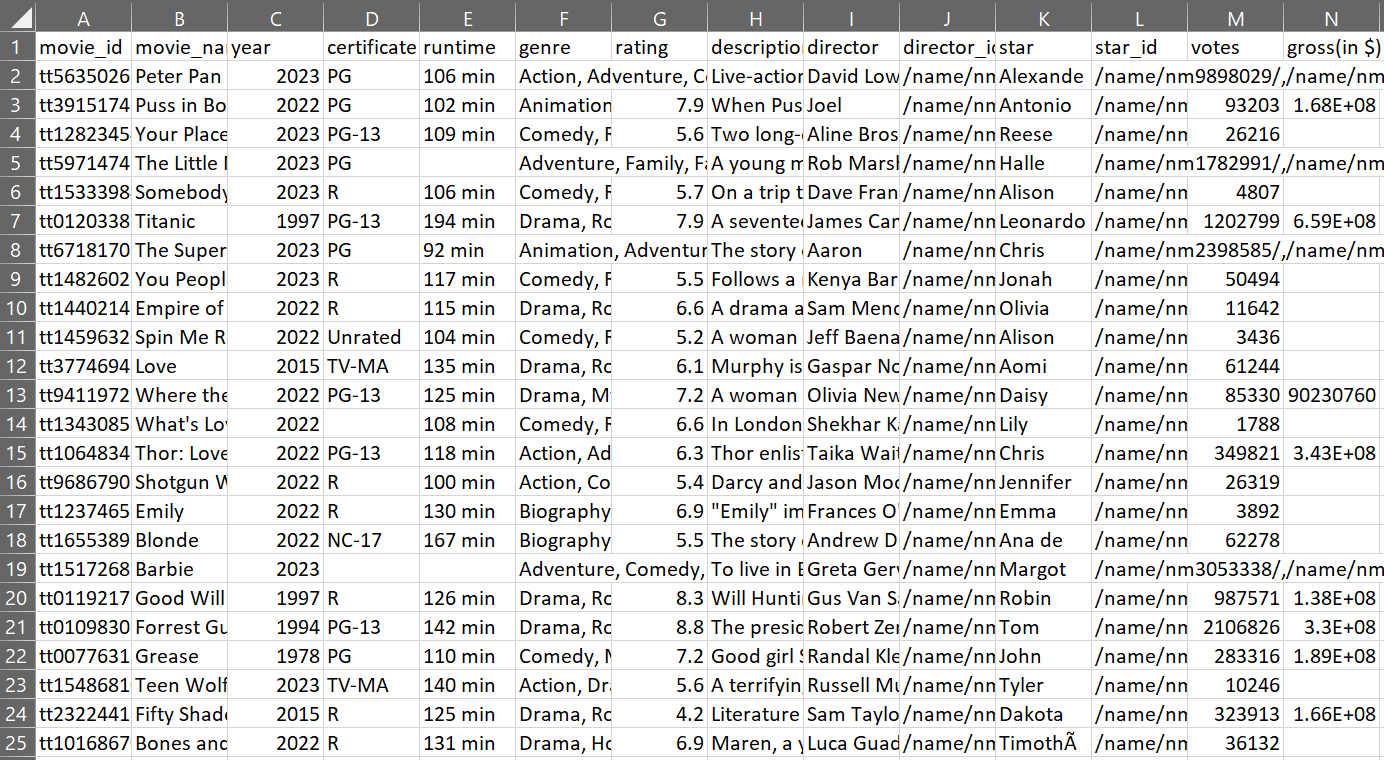
More questions were initially planned to be asked in the report. However, this was prevented by the maximum word count of 3000 of the report.

**Appendix**

* **Dataset**



*Files inside archive.zip*



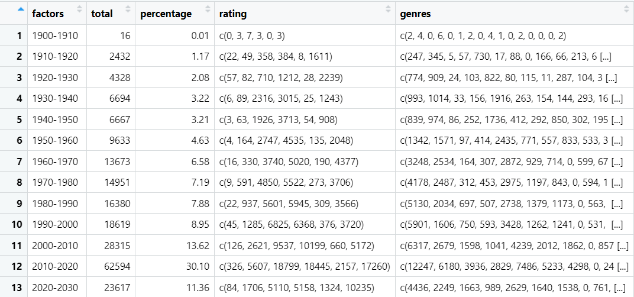
*Columns inside csv file*

* **Data Preprocessing**



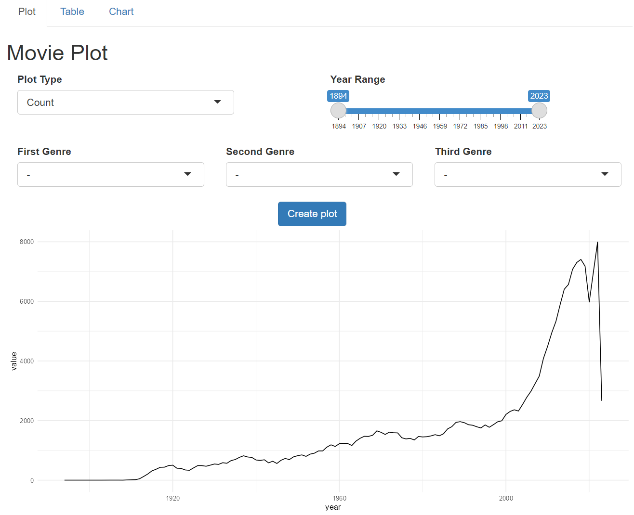
*Some columns inside Movies data frame*

* **Data Transformation**

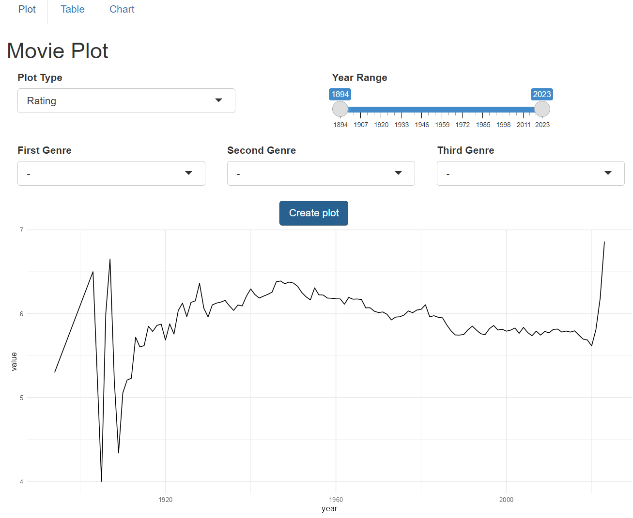


*Some columns inside the new data frame*

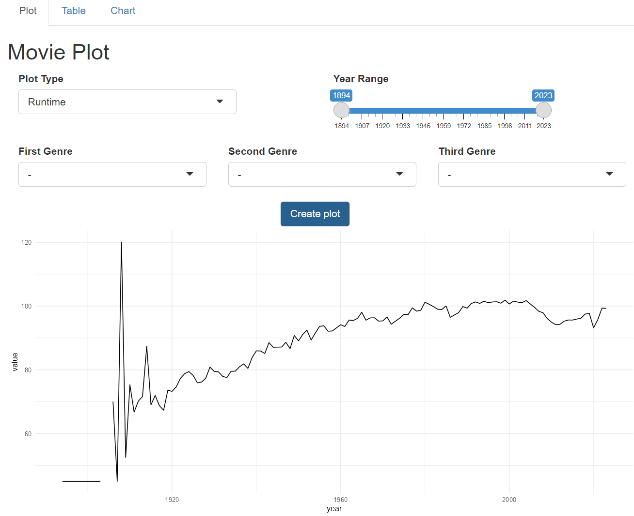
* **Question 1**



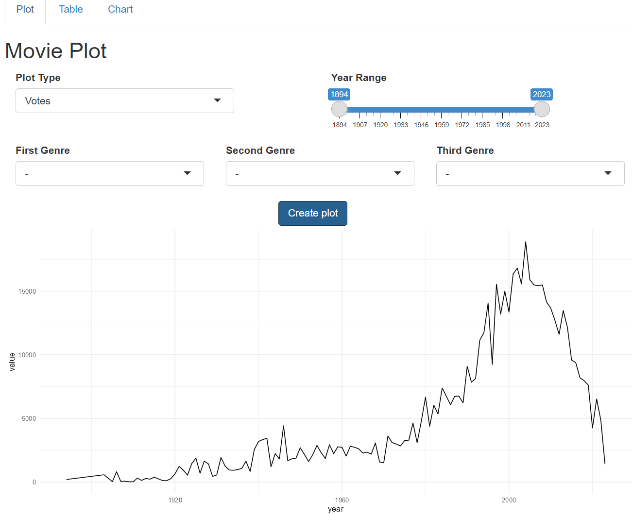
*Movie Count*



*Average Rating*

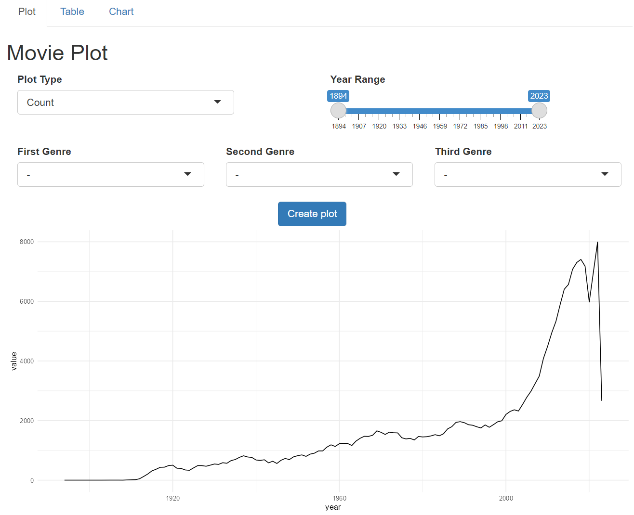


*Average Runtime*

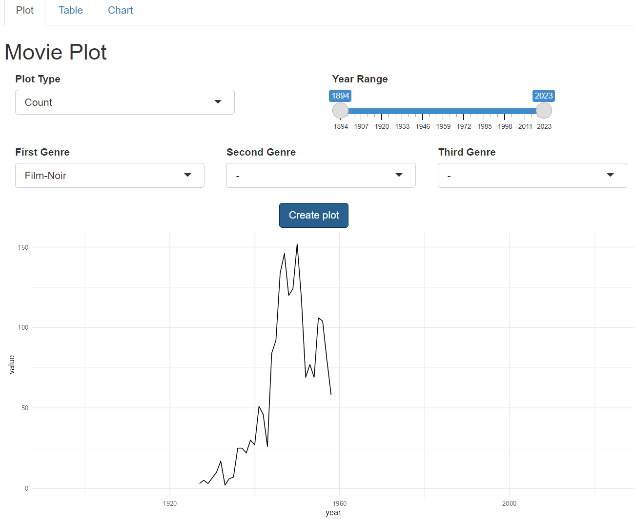


*Average Votes*

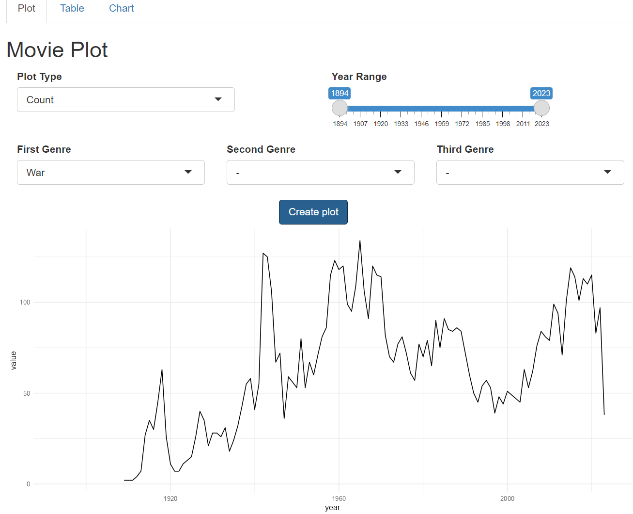
* **Question 2**



*Movie Outlier*

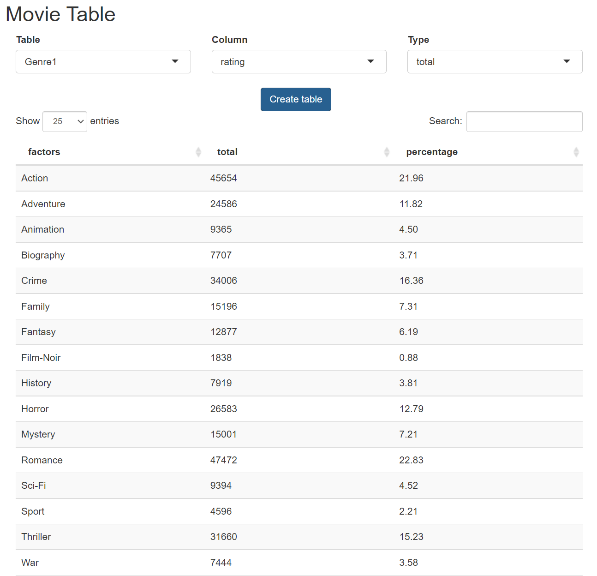


*Film-Noir Outlier*

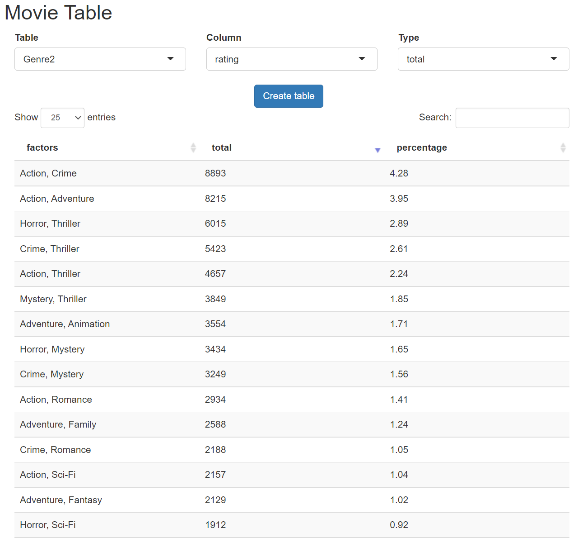


*War Outlier*

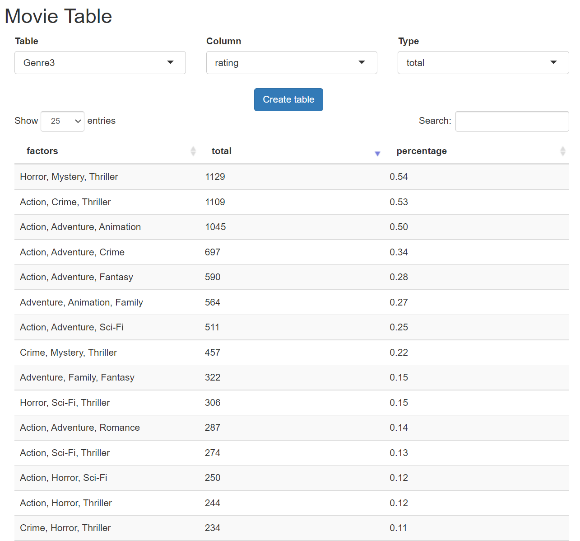
* **Question 3**



*Single Genre*

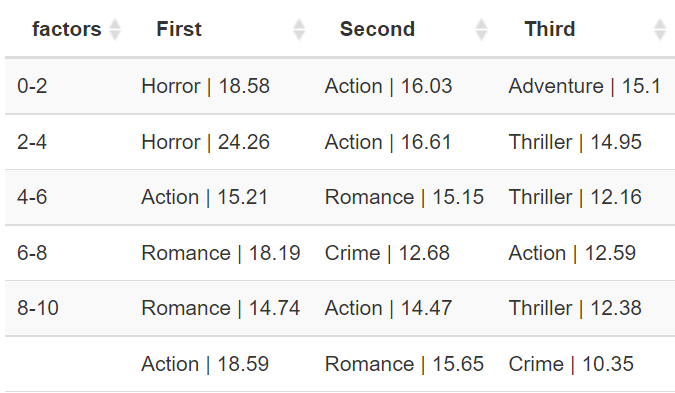


*Double Genre*

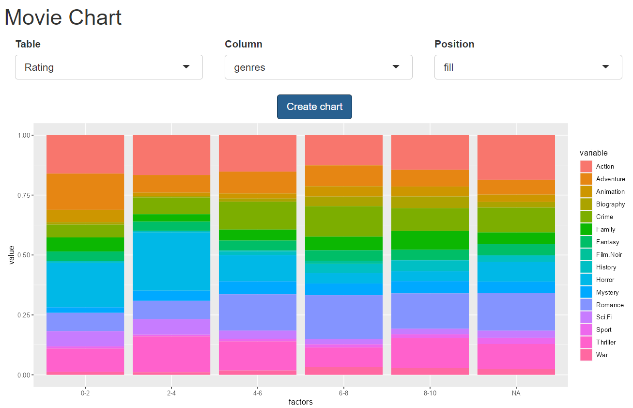


*Triple Genre*

* **Question 4**



*Table*

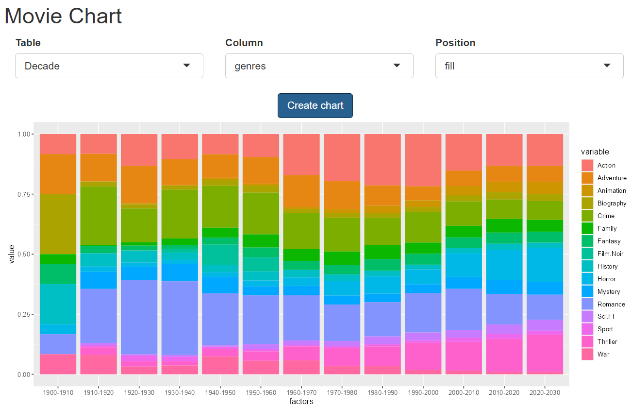


*Chart*

* **Question 5**

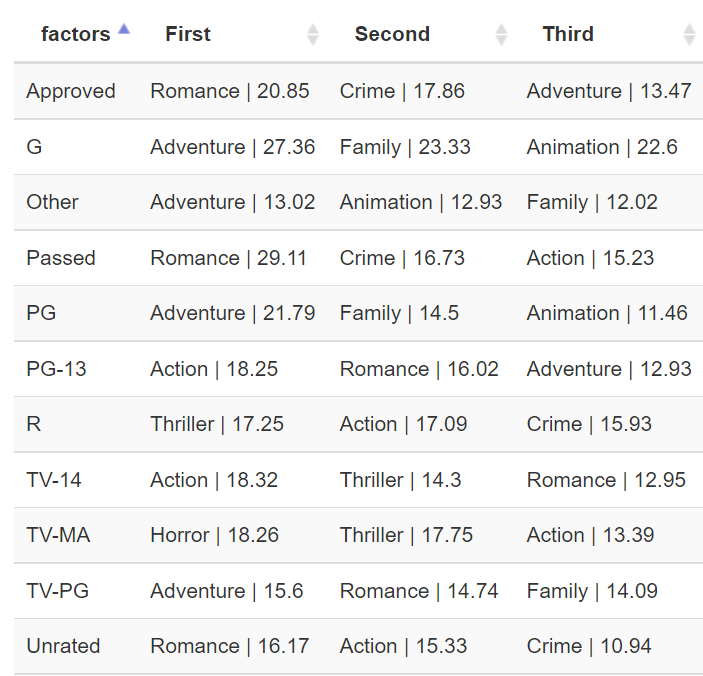


*Table*

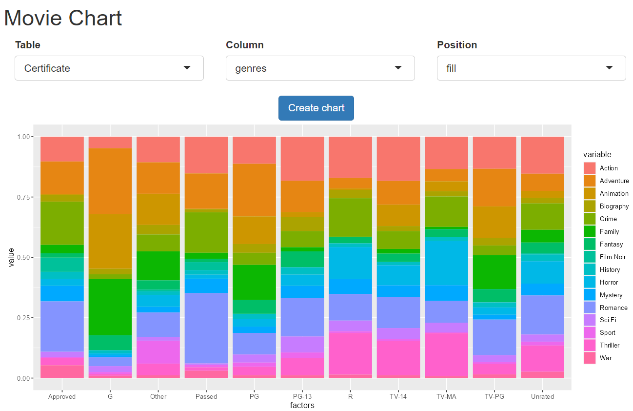


*Chart*

* **Question 6**



*Table*



*Chart*