Women in the Film Industry

Project Documentation

Prepared by Group-7-data1

Daisy Watson, Katie Roberts, Catherine Jones,
Maria Pilipenko, Sahr Anan Ahmad

INTRODUCTION

Is there a gender pay gap in the Film Industry?

Aims and objectives of the project

The aim of our project was to research whether there is gender inequality in the film industry through analysing how much female actors are being paid in comparison to male ones.

The main topic we came up with was 'Women in the Film Industry'

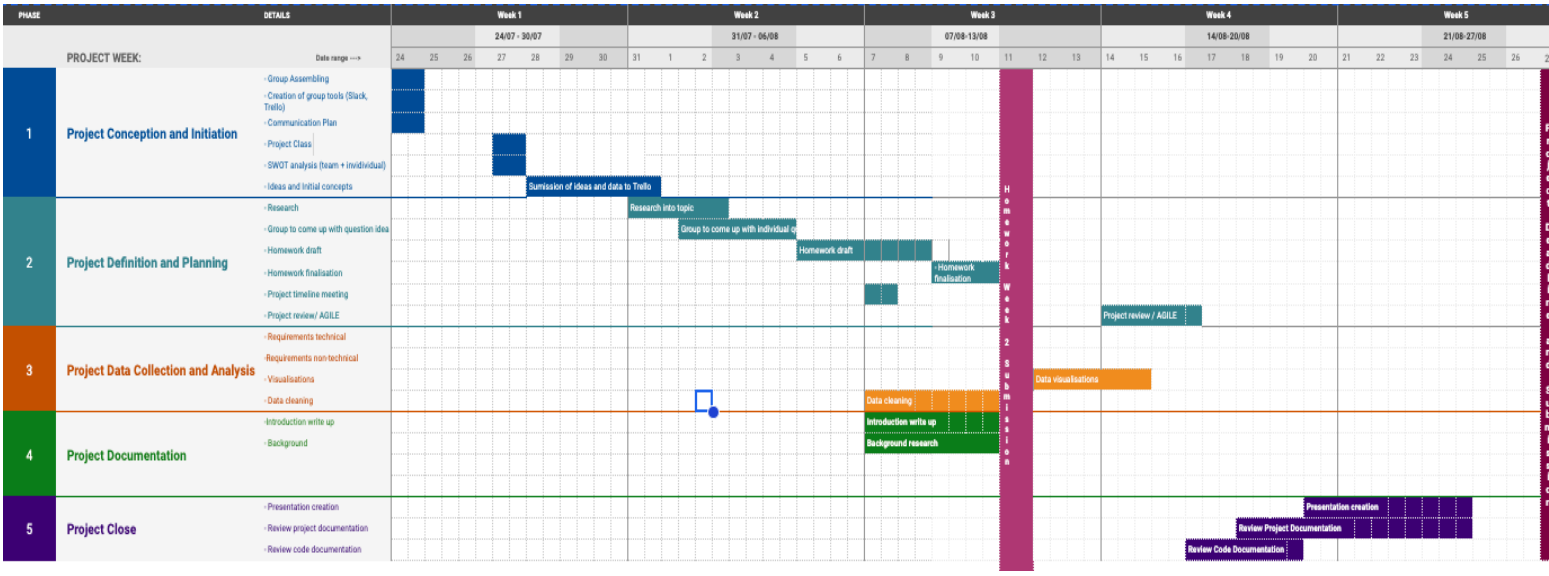
The questions we initially wanted to ask were:

- Are there any differences in earnings between female/male actors?
- Have earnings of female/male actors changed over the years?

We think earnings can be considered as the most established means of testing equality and the film industry would be a great instance of showing any potential gaps due to its popularity.

We decided to cover Hollywood actors in particular as the cinema of the United States is the oldest and one of the biggest industries in the modern world so analysing whether a pay gap exists in this entertainment capital could potentially be projected on the other regions.

Roadmap of the report



{ The project timeline was prepared by Catherine at the beginning of the project.
The roadmap consists of 5 phases: Conception & Initiation, Definition & Planning, Data Collection & Analysis, Documentation, Project Close }

Background

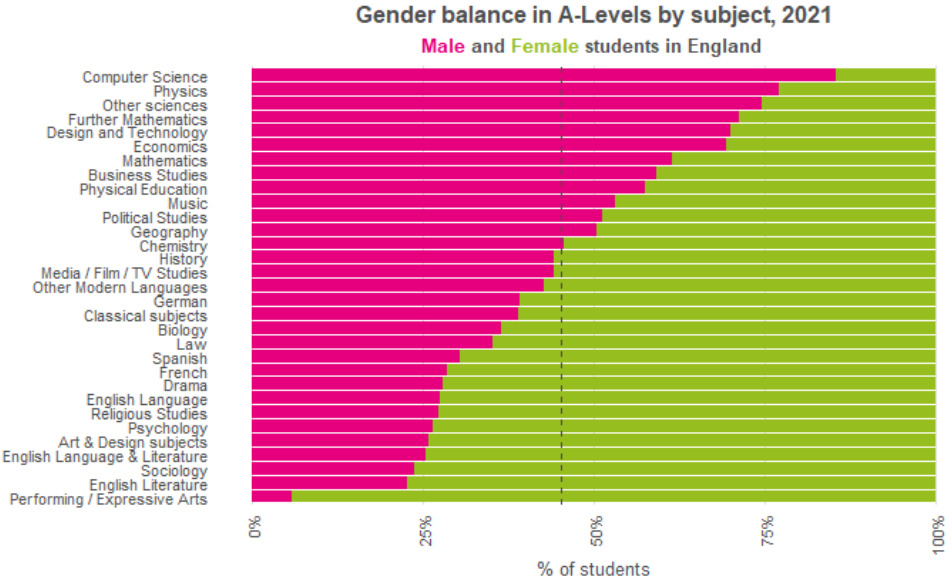
It was important to us as a group to concentrate on inequalities, particularly as we are completing the CFG Degree that promotes equal opportunities for women in the tech field.

Film industry analysis was the area of our interest as it is one of the biggest platforms across all industries to show the impact of the gender pay gap.

We determined as a group that our target audience would be **students completing Performing Arts/ Drama AS/A-Level college qualifications and their educators.**

Points we considered:

- The student population of drama courses are female dominated, which ties into our questions relating to potential differences in pay compared to gender for this profession
- We thought our target audience would first of all be familiar with current actors who are part of our data and would be interested if there is a difference
- As our target audience have chosen Performing Arts/Drama as one of their subjects to study, they may want to potentially pursue acting as a career and be interested in this topic for their chosen career path
- It is crucial for this age group to be familiar with potential change in earnings the change in earnings over time for women
- Our analysis would also be beneficial for educators, as they could use our findings/project ideas within their own teaching, to familiarise their own students with the topic we are discussing
- Educators could use our findings as a base for their own teaching, supporting their own students understanding of the profession



{ For more information, click [here](https://shorturl.at/fmAER) }

The above chart shows the gender balance in all A-Level subjects in the year 2021. Bear in mind that more female than male students sit A-Levels. In 2021, the overall cohort of A-Level students was 55% female. This is indicated on the chart below as a dotted line. The chart above shows there are less than 12.5% of males, taking Performing Arts as an A-level

SPECIFICATIONS AND DESIGN

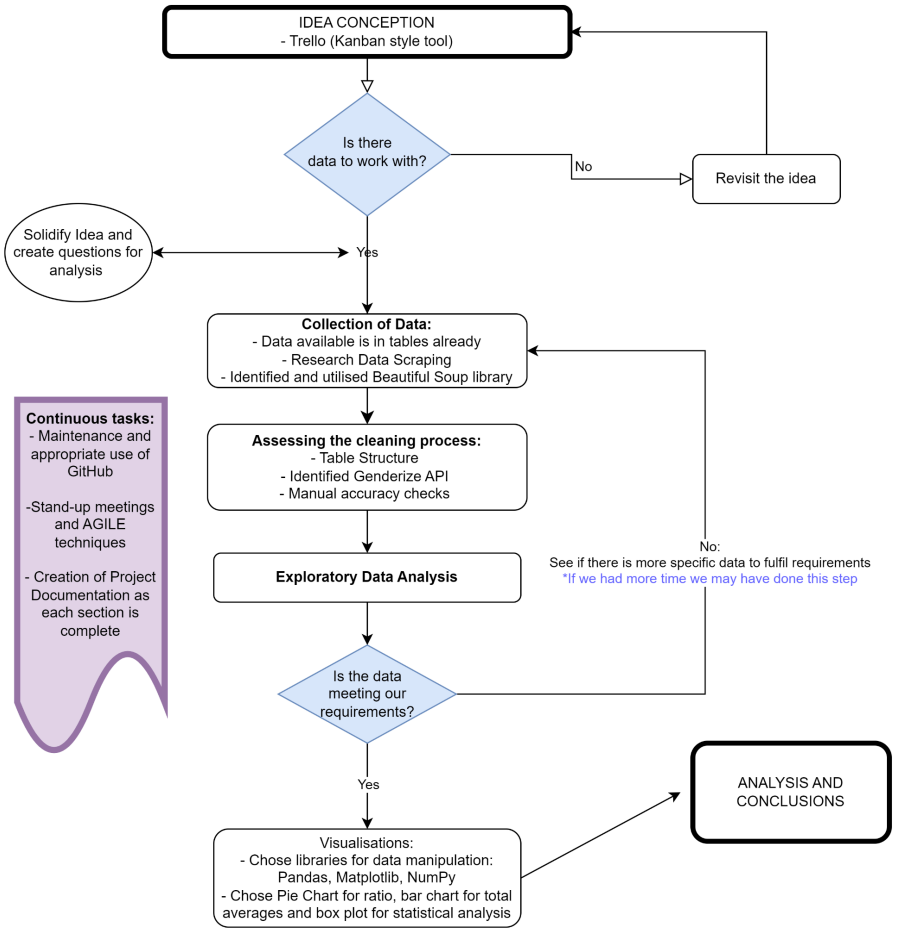
Requirements technical and non-technical

Technical requirements

- Readable and concise code is written in Python code, using the Google Colab Notebook
- Final code includes comments and clarifications
- Datasets are scraped from the reliable web sources using API
- Collected data has to be relevant
- API is applied for identifying gender of actors in the datasets
- Scraped datasets are cleaned with Pandas library
- Visualisations are done with Matplotlib library
- All the code scripts and required documentation are uploaded to Github library

Non-technical requirements

- Results of the analysis meet the set aim of the Project
- The visualisations are clear to understand by the target audience



IMPLEMENTATION AND EXECUTION

Development approach and team member roles

We used SWOT analysis to identify each member's weaknesses and strengths so everyone in the group is aware. To develop everyone's skills where possible, we have paired each person with their counterpart, so that there is a balance of knowledge (for example; someone who identifies themselves as weak using SQL paired with a confident SQL user). This allowed discussion between partners as first port of call and then questions can be posed to the wider group if a solution cannot be found. SWOT Analysis

Tools and libraries we used

1. Python to write the code
2. Data scraping using API (requests library)
3. Applying gender with the help of API (https://genderize.io/)
4. Data cleaning using Pandas, Numpy Library
5. Visualisations done using the Matplotlib library

Implementation Process. Our achievements, challenges

Scraping data

The data scraping process didn't go as smoothly as we initially thought. Although the data appeared to be complete and accurate when the .CSV file was created, it was discovered during the cleaning process that the accuracy of the data was brought into question. Due to the way the .CSV file was formatted, the tables were not clearly separated and so cleaning the data proved more challenging. To rectify this, the Top 10 dataset was re-scraped to streamline the preparation for cleaning.

Cleaning data

As multiple datasets were similar in shape and content, the more efficient way to clean it was offered by Catherine. She managed to loop through the files that required cleaning and automated the process for each dataset.

Applying gender

The datasets we fetched didn't have a gender column, so it wasn't convenient to analyse them. It was resolved to use an API to apply gender to all the datasets (the respective gender column was added to each table).

This API was very efficient and meant large quantities of data could be assigned at once and improved efficiency as a team member did not need to manually assign the gender. However, a weakness of this API is that it infers the gender based on the first name. The accuracy of this API did need to be checked manually and across the larger data sets. On average, it would miss interpret 1 in 50 names. However, this manual check also assumes prior knowledge of the actors or indeed that the team member assessing the data is also working on assumptions like the API.

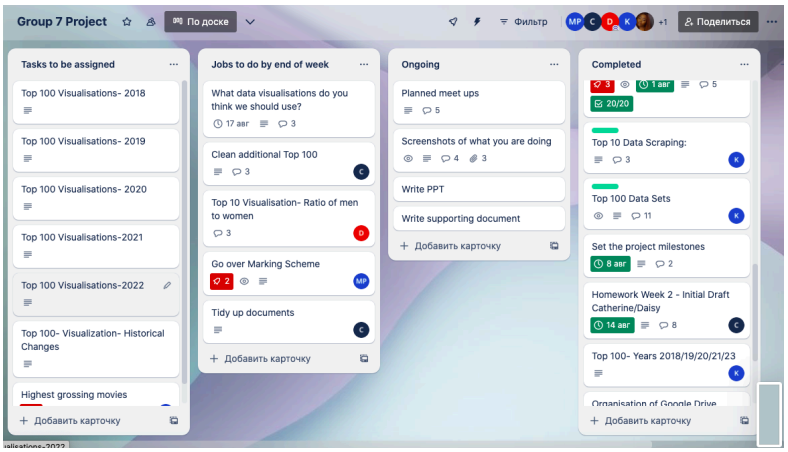
Filtering available data for analysis

Initially we came up with more questions on the topic and data sets that could be used. Some of these included the number of awards women / men have, evaluation of their popularity across ordinary users and critics and also the gender split in different roles associated with film. The scope and time constraints of the project meant that we needed to use AGILE techniques to reassess our approach. We streamlined our analysis and decided to concentrate on the pay gap analysis. Thus, some of the datasets we used, explored and cleaned at the beginning were set aside.

Agile development elements we implemented

Trello Cards – Throughout the project we were using Trello board (a web-based, kanban-style, list-making application) as a way to keep track of our activities and assign tasks. We found this an inclusive way to share ideas and comment on specific tasks, assign dates so tasks are completed by a certain date and keep track of what tasks still needed to be assigned. It was an inclusive tool as members of our group may have differing timetables than others, therefore everyone was able to have one source of organisation to refer to.

{ This is how the Trello board looked at the visualisation stage of the Project }



Stand-Up meetings – Each week we held several Stand-ups via Zoom for discussing current accomplishments, challenges and concerns. Each member of our group was able to lead the stand-up and made the notes to the Google Doc linked here [Daily Stand-Up]

Iterative approach – This was implied with data scraping. The scraping process was repeated several times in order to achieve better results and cleaner datasets. Each task also once completed, has been checked by another group member, so feedback, particularly in the daily stand ups, could be given to improve.

Refactoring – Initially the datasets were cleaned individually. All the datasets could be cleaned in an automated way. Also with data visualisation, the use of a for loop to improve data so it met with specifications needed to visualise all data.

Implementation challenges we faced

Visualisation

One problem initially occurred with the visualisation process, particularly with pie charts - the acceptance of 0 values. For example, Top 10 data from 2017 had zero female actors, however the pie chart visualisation would not initially accept this. Therefore, we came up with an IF statement that took this into consideration and plotted the chart, else take the count of gender in the data and plot this instead.

When it came to creating visualisations for our data, it did highlight some weaknesses in our data sets as we didn't have the depth of data that would have been really beneficial for the analysis. Knowing more about the importance of statistical analysis and methods has prepared us for future projects to consider more carefully the data required.

DATA COLLECTION

What information do we need?

After choosing the topic of women in the film industry, it was evident that we needed to look for the data having common variables that could be compared regardless of gender; numerical data sets showing key variables such as earnings would be of importance as a variable that is understandable, especially taking into account our target audience. It was also useful to find data sets over a longer period of time to identify any patterns and trends.

What information is available?

From the data gathered, there was information on a variety of subjects, for example academy awards won in a number of categories, actor accreditation in films, most successful films in the industry and their critics, budgets of films and salaries of actors in the industry. There was a range of selection with datasets and each could branch from the topic of women in the film industry industry well. However, as a group, we decided a financial approach to our topic would be more relevant. After conducting background research and the high amount of students undertaking acting as an A-level, we thought this would be a good topic of choice.

What is our data source?

As we decided to scrape the data from the web, websites with reliable business information in the sphere of Hollywood filmmaking were our choice. The sources we used for our analysis:

- Forbes, American business magazine (the link to the data: https://shorturl.at/ijrsX)
- Wikipedia, six Forbes' lists were scraped from there (the link to the data: https://shorturl.at/yDF01)
- The Numbers database (the link to the data: https://shorturl.at/gsAGN)
- The Variety, American magazine (the link to the data: https://shorturl.at/bnI78)

Forbes publishes yearly lists of the highest paid actors and actresses based on a 12 months scoring frame (total earnings from 1st June the previous year to 1st June the current year). Rankings are based on screen and extracurricular earnings.

- Lists for years 2015-2020 on Top 10 actors were scraped from Wikipedia as they were already formatted as tables. We made sure that Wikipedia refers to Forbes and contains reliable information

- Earnings estimates in these datasets are based on data from Nielsen, ComScore, Box Office Mojo and IMDB, as well as interviews with industry insiders. All figures are pre tax and don't include fees for agents, managers and lawyers are not deducted.

- Figures for 2021 and 2022 from Variety magazine were based on confirmed salaries for single production role of the actors(accounting only highest paid role)

The Numbers website is a freely-available database of movie business information that gets all the data from the source. Top 100 lists were scraped from there.

- The lists from the Numbers show the highest grossing stars based on the domestic box office (which is defined by the North American movie territory) of the movies they had a leading role in in 2023 and the two preceding years (voice-only roles excluded).

How we collected our data:

We collected data through a variety of means to ensure we had a more reliable and extensive data set. This included web scraping and utilising APIs to later assign data to our data sets.

The Top 10 data sets came from Wikipedia so it is important to note that this form of data can be altered by people other than the original data creators (additionally, we made sure tables from Wikipedia were based on a reliable source of information - Forbes magazine). To strengthen our research we then conducted data scraping from another source of the Top 100 highest grossing actors. The issue found with this initially was that we had to inspect the website's HTML to ensure that we had permissions to scrape the data. After

finding that we did, we conducted the web scraping. This was a fairly straightforward process and we were able to use the same foundation coding for all of the years we used in the Top 100 data sets.

We utilised the Requests and Beautiful Soup libraries in Python to conduct the web scraping.

ANALYSIS

To answer the stated questions, we divided our analysis into two logical parts:

1. **Analysis of the Top 10 datasets for 2015 - 2022** (analysed by Maria and Daisy)
2. **Analysis of the Top 100 datasets for 2018 - 2023** (analysed by Catherine and Sahr)

The full version of the analysis can be found in the Notebook submission. But here is the brief summary of our findings:

Mini-conclusion drawn from Top 10 data set:

The average earnings of female actors throughout the selected period of time stay consistently below average earnings of male actors. Additionally, the representation of women should be taken into account to make the picture full. Even though at some point women may have an income comparable to men, the percentage of women represented in the datasets is very poor. It is evident that over 8 years the percentage of women in the Top 10 highest paid actors is significantly lower than the percentage of men.

Mini-conclusion drawn from the Top 100 data set:

The ratio of women to men stays consistent over time (65% men to 35% women); however it has been showing a decline in the last two years. This compounds the findings in the Top 10 data set, showing there isn't much movement towards gender equality even over a larger data set. To improve our conclusions, finding numerical data sets like the Top 10 would provide more solid conclusions as with the Top 100 we are inferring a lot from the rank variable without any financial data to solidify our conclusions.

CONCLUSION

- **Is there a gender pay gap in the Film Industry?** Yes, male actors dominate both in the Top 10 and Top 100 highest paid actors data sets
- **Are there any differences in earnings between female/male actors?** Yes, on average, at the very minimum women earn $10 million dollars less going up to extreme years like 2018 where they earned over $30 million less
- **Have earnings of female/male actors changed over the years?** Yes, the average of both men and women earnings have declined over time from the Top 10 data. We would like to find more numerical data over a larger data set and wider time frame to solidify these conclusions

To sum up, as a team, we have successfully answered the initial questions we set out at the beginning. We have used techniques that we have learnt throughout the duration of the CFG degree to retrieve, analyse and present our data. Not only that but through AGILE techniques and use of development tools we have been able to grow together as a team, to support each other and develop perceived weaknesses.

Given more time we would have enjoyed developing our analysis further to provide our target audience with more insights. With the help of our analysis they could make data-driven career choices, be aware of and challenge gender inequality in the Film Industry.