INVESTIGATION INTO WHETHER BABY NAMES IN THE UK ARE INFLUENCED BY MOVIE RELEASES

CODE FIRST GIRLS
SUE SCOTT
GEORGIA GRANT
FLAVIA FONSECA
KATIE PAPWORTH
LAURA FERNANDES

INTRODUCTION

The main objective of this investigation is to analyse data sources to see whether baby names are affected when movies are released with character names that appeal to new or expectant parents. The assumption is that any name may or may not be popular before the name of a character appears in a new movie, but this investigation will look at whether there was an increase in name usage (or a decrease) after the movie release date. It is important to note that this is not an investigation into a causal relationship instead the investigation aims to look at the possible influence of movies on name decisions made by new parents.
The research will use the highest rated movies (the most popular) by year and obtain the character name (or names if there are several main characters) for the hero and the villain (where there is one). Alongside this, baby name data will be selected for the research years.
This report shows the steps taken to achieve the objective and to investigate related points of interest. It will explain how the project team worked together to complete the research, it will examine the limitations of the project and will explain conclusions drawn.

BACKGROUND

The project began with an assumption that there would be a link between movie character names and baby names as members of the team had examples of people that had used names from characters in movies. For example, a friend of one team member called their baby "Riley" after the character in the Disney movie "Inside Out". Whilst it would not be possible to find out why a name had been chosen, the research could investigate the popularity of names. This research could be for people who are interested in the popularity of names over time. It may show that they called their child a name that was very popular in their year of birth and that name became popular after a certain movie was released. This project aims to show how pop culture, specifically movies, influence baby names.

STEPS SPECIFICATIONS

To begin the process of data analysis, the first task was to search for the data. This was carried out individually and then discussed as a team over a number of zoom calls. The most straightforward data to find was on baby names, taken from the Office of National Statistics (ONS) which provided baby names in England and Wales dataset in excel format. For an extensive overview of the data quality and birth statistics please refer to the ONS "user guide to birth statistics". The dataset covered the years 1996 to 2020 (latest data released on the 18th October 2021 for 2020); this provided plenty of time series data that could be used in the analysis.

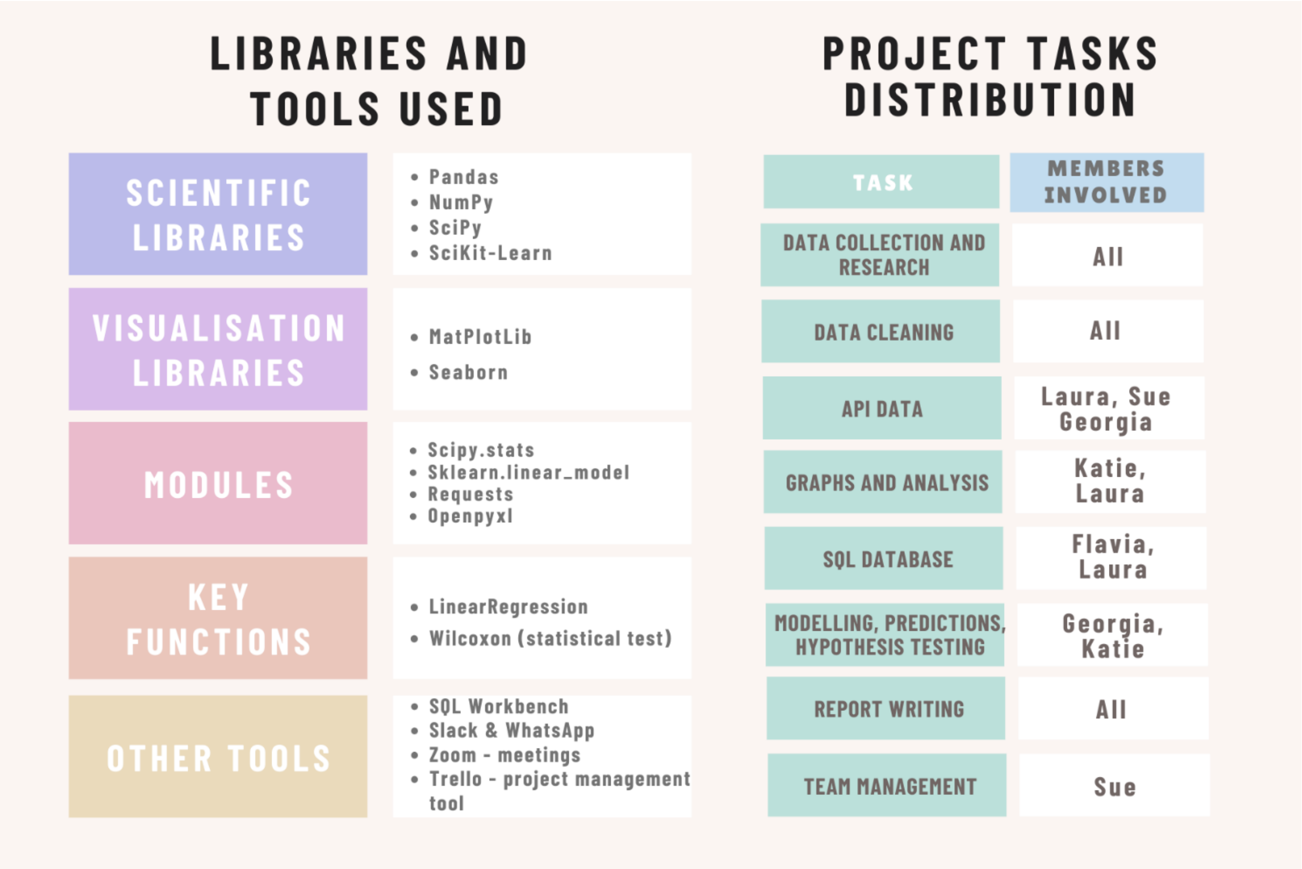
The second part of the data required was on movie releases by popularity. The movieDB website was identified as having an API with an API key which provided data on movies by year, popularity, and genre. This was a useful source of data however it had limitations in that it did not provide character names and it is a community led database which means the data can change. The team were unable to find a movie data source with character information so agreed to build a database in SQL using the top three most popular movies from 2009, 2011, 2013, 2015 and 2017 adding in the character names, gender and genre. Where applicable, both "good" characters (the heroic characters) and "bad" characters (the villains) were input into separate tables.

Each movie was assigned a unique ID so that the data could be normalised. The database relationship is one to many. There is an Entity Relationship Model for the database in Appendix A. A connection between the SQL Movies database and Pandas was coded in order to use the data from the database in the analysis described later in this report.

IMPLEMENTATION AND EXECUTION

The team started the project keen to share all tasks from finding the data to drafting the conclusions, therefore following the waterfall methodology (working through each stage step by step). However, due to the limited time (three weeks after the team had agreed the subject matter and found the data sources) to conclude the project, it made sense to divide the remaining tasks between the team members so there were a number of meetings held with less of the team present according to task and availability.

After data cleaning was completed, the team followed Agile methodology, assigning tasks according to the strengths of each member. The Agile methodology, specifically scrum, allowed the group to adapt to any roadblocks. This methodology also allowed for multiple team members to work simultaneously on the same section of the project and for the work to be done in stages and reviewed after each one. Furthermore, a Kanban methodology was implemented in conjunction with Agile. Kanban allowed for physical representations of our workflow in the form of online applications e.g., Trello. The team are from diverse backgrounds: Flavia has a background in Geology (Natural Sciences) and data analysis, both Georgia and Katie are mathematicians, Laura has a master's in Astrophysics and Sue has a master's in Human Resources and is an HR Business Manager. Based on these backgrounds and our SWOT analysis (Appendix B), the tasks were assigned according to figure 1.



LIBRARIES AND TOOLS USED

SCIENTIFIC LIBRARIES	• Pandas • NumPy • SciPy • SciKit-Learn
VISUALISATION LIBRARIES	• MatPlotLib • Seaborn
MODULES	• Scipy.stats • Sklearn.linear_model • Requests • Openpyxl
KEY FUNCTIONS	• LinearRegression • Wilcoxon (statistical test)
OTHER TOOLS	• SQL Workbench • Slack & WhatsApp • Zoom - meetings • Trello - project management tool

PROJECT TASKS DISTRIBUTION

TASK	MEMBERS INVOLVED
DATA COLLECTION AND RESEARCH	All
DATA CLEANING	All
API DATA	Laura, Sue Georgia
GRAPHS AND ANALYSIS	Katie, Laura
SQL DATABASE	Flavia, Laura
MODELLING, PREDICTIONS, HYPOTHESIS TESTING	Georgia, Katie
REPORT WRITING	All
TEAM MANAGEMENT	Sue

Figure 1. Summary of project task distribution (left) and the libraries and tools the team used through the project (right).

Figure 1 also shows the libraries, modules, functions and tools that the team used throughout the project.

DATA COLLECTION AND CLEANING

ONS Data

The team downloaded the excel file from the Office of National Statistics and had a look at the data by reading the file. Based on the investigation objective, it was agreed that separating the boys and girls data (which was already on separate tabs on the excel spreadsheet) would need to be the first task. The data cleaning process was then applied to the boys and to the girls data separating the count (the number of times a name had been given) and the rank (the position of the name). The resulting dataframes were saved to csv files ('girls_rank_clean', 'girls_count_clean', 'boys_rank_clean' and 'boys_count_clean'). The notebook for the ONS data cleaning is 'Data_Cleaning_and_Exploration_ONS_data'.

Movie Data (API)

To use an API from The MovieDB, a TMDb account was created, and an API key was requested. Reading the required documentation enabled the appropriate retrieval of the API data. Research on the various genre options led to three API requests being made for the following genres: Animation, Family and Fantasy. The response of the API requests for these three genres were discussed and it was noted that some movies appeared in more than one genre.

It was decided that the family genre would be best suited for the analysis of this project. The data was then cleaned, leaving the team with only the information they were interested in, namely the movie title, year of release and rating. Conditioning on the years column and sorting by the rank we printed and saved the top 3 rated movies for the years 2009, 2011, 2013, 2015 and 2017.

Figure 2. was created using the python library Matplotlib (see Jupyter Notebook 'Analysis_Movies&BabyNames_Subplots'). It is a visual representation of the data collected from the moviedb API, specifically the cleaned data, where the top three ranking movies in 2009, 2011, 2013, 2015 and 2017 are shown. The notebook for pulling the data via the API and cleaning the data pulled, is 'Movie_API_ Genre_Family'.

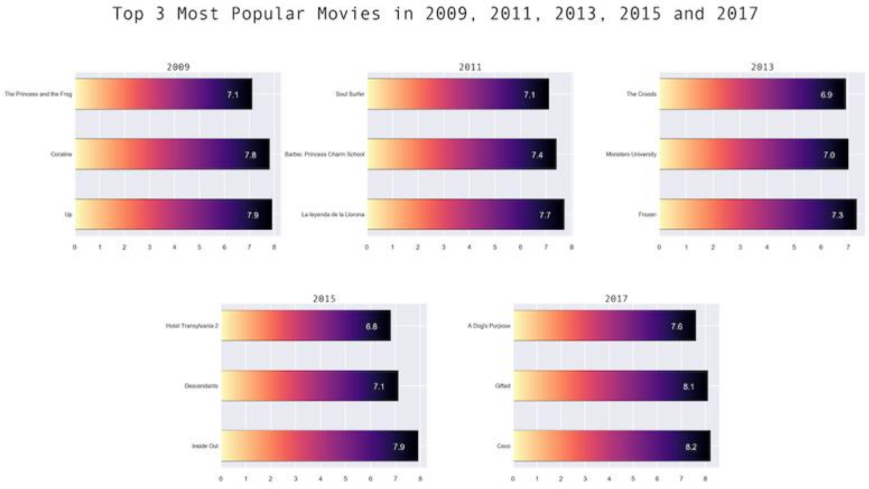
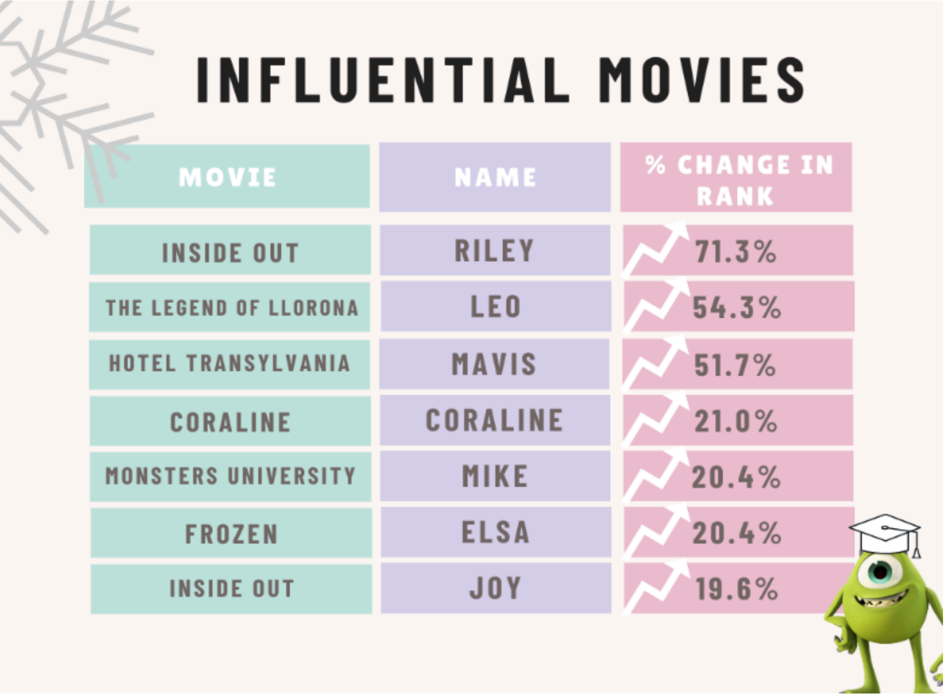


Figure 2. Rating vs Movie Title horizontal bar chart for the top 3 highest rated movies from top to Bottom row; 2009, 2011, 2013, 2015 and 2017. Each horizontal bar represents a movie, where the height of the bar determines its rating value. A sequential plot is used to colour-code the bars.

RESULT REPORTING

I. Which movies held more influence on baby names?

To understand which movies influenced baby names the most, in the 'Exploratory_Analysis' notebook, the team calculated the percentage change between the mean rank 3 years before and after the release of the movie, and then sorted the data according to this value. This was saved into an excel file (Key Statistics Heroes and Villains). The top results are displayed in figure 3. It is possible to conclude that among the movies analysed, the ones that held most influence over baby names were Inside Out, The Legend of Llorona and Hotel Transylvania as these are the movies where the rank of the character's name, climbed the most in the subsequent years after movie release. The name 'Riley' (female) went from an average rank of 965 all the way up to 277 (average) after Inside Out's release. Also, 'Mavis' climbed from an average rank of 4480 (years 2012-14) up to 2165 (year 2016-18) which is an enormous 51.7% increase.



INFLUENTIAL MOVIES

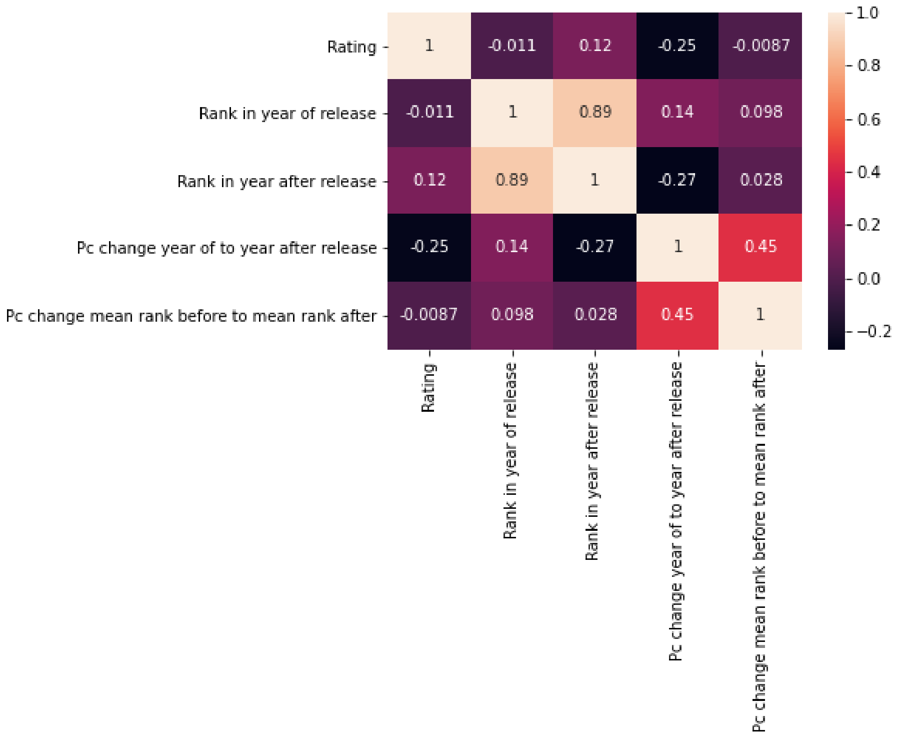
MOVIE	NAME	% CHANGE IN RANK
INSIDE OUT	RILEY	71.3%
THE LEGEND OF LLORONA	LEO	54.3%
HOTEL TRANSYLVANIA	MAVIS	51.7%
CORALINE	CORALINE	21.0%
MONSTERS UNIVERSITY	MIKE	20.4%
FROZEN	ELSA	20.4%
INSIDE OUT	JOY	19.6%

Figure 3. Summary of Influential Movies

II. Do movie ratings affect the usage of characters' names?

A heatmap plotting the correlation between 'Rating' and other key variables ('Rank in year after release') was produced using Seaborn (figure 4 below and the 'Exploratory_Analysis' notebook). This highlighted that, for the data we have, there is no correlation between rating of a movie and the subsequent ranks/popularity of baby names chosen due to all the correlation coefficients lying around the zero value. This implies that, baby names are not necessarily chosen based on a movies' merit.

Figure 4. Heatmap of correlations



III. If there is a trend between movies and baby names, how long does the trend last?

An overview of hero and villain movie characters and given names is depicted in figure 5. The movies shown are the top 3 highest-rated movies across the selected years. The hero character(s) within each movie is marked with an upward '^' pointer and is colour-coded. The villain or protagonist of the same movie has a down-facing pointer 'v' and is colour-coded in black. The smaller the value - the higher the rank e.g., 1 is considered the highest rank.

This section analyses some of the more noticeable trends and patterns between movie character names and names given to babies born in the selected analysis years. In the first row of plots, we have the highest-rated top 3 movies that were released in 2017, these were Coco, Gifted and A Dog's Purpose. Ernesto (villain character) appears to show no change in relation to the release year of Coco but here was a drop in the rank of Ernesto one year after the movie was released then an increase in the subsequent year. The hero in the same movie, Miguel, has had a consistent ranking over the past 20 years, which does increase in rank the year following the movie release.

In Gifted, the hero and heroine character names, Frank and Mary had dropped in their rank value the year the movie was released. However, in 2018, one year after the movie was released both names became more popular and improved in ranking; by 2019 the rank (popularity of the names) begins to fall again.
Within the same year the protagonist's name, Todd, in A Dog's Purpose did not show a considerable deterrence for using Todd after the movie was released. Whilst the rank for Todd decreases one year following the release of the movie, its' rank briefly increases in the subsequent years. Overall, the villain's name does not seem to have been affected by its' use in the movie and therefore there must have been an external influence for the change in rank.

In the second row of subplots for Inside Out, Descendants and Hotel Transylvania 2 which were released in the year 2015. In the first two movies there were no villain characters. The two heroine characters' names in Inside Out show an increase in their rank after the movie is released. The given baby names related to the hero/heroine characters in the Descendants movie have varied in rank between 2000 and 2020: the names 'Chad' and 'Audrey' became popular (improving in rank value) the year the movie was released.

Hotel Transylvania 2 shows an interesting correlation. There is an increase in ranking for the heroine Mavis and a decrease in ranking for the villain Bela (also called Samuel) in the years following the movie release. It is worth noting that the heroine Mavis' rank was increasing prior to the release of Hotel Transylvania 2, although this increase in ranking for the name Mavis could also be a consequence of the popularity in the first movie (Hotel Transylvania, 2012) which contains the same heroine's name.

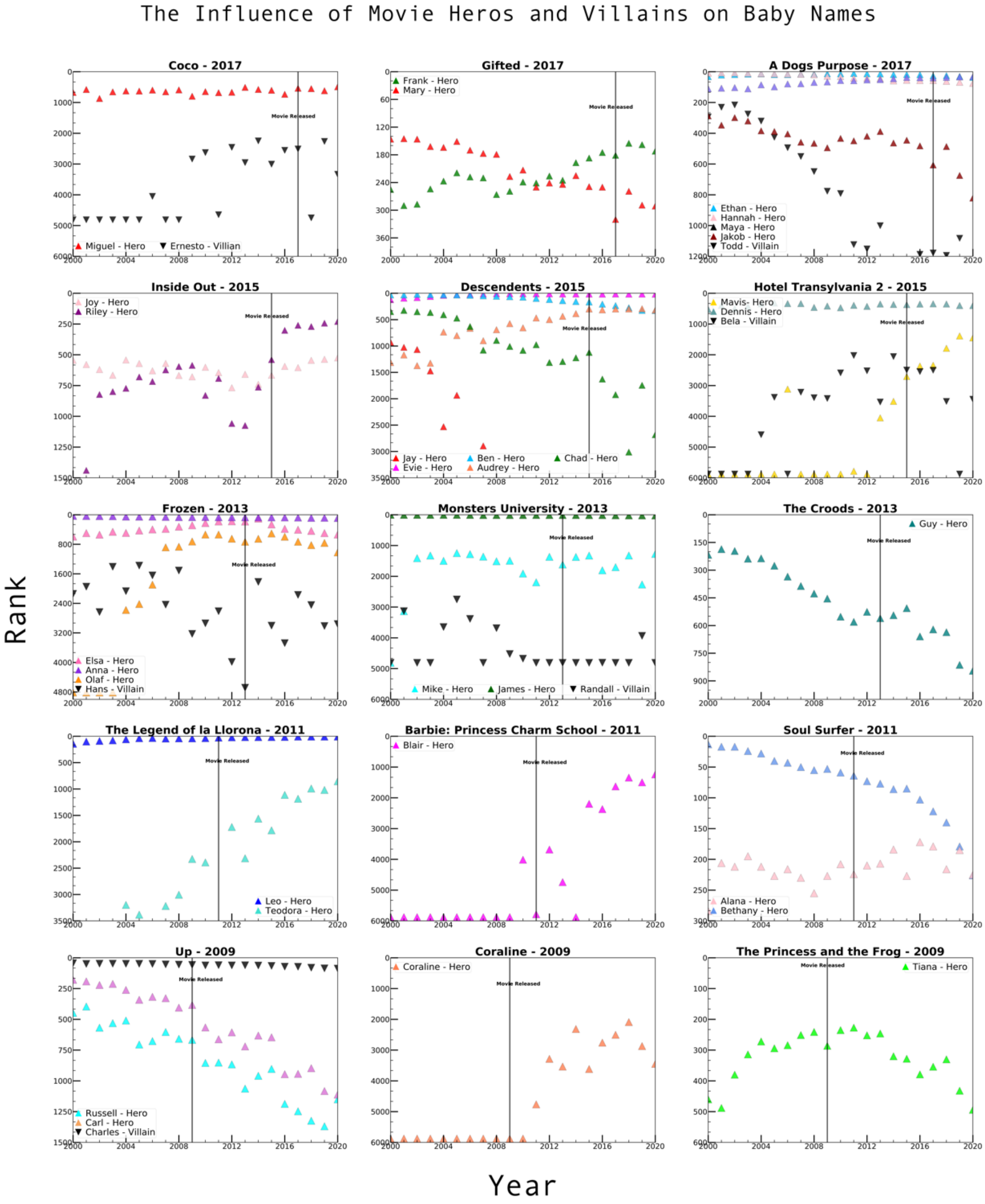


Figure 5. Year vs Rank scatter plot on the influence of movie hero and villain characters on baby names between the years of 2000 and 2020, where lower values are assigned a higher rank. The figure consists of 15 subplots. From top to bottom, each row contains the top 3 movies released in 2017, 2015, 2013, 2011 and 2009, respectively. Each panel is labelled with a movie title and the year of release. Within the panels the names of characters deemed heroes are marked with an up-arrow point (various colours used) and where possible a villain/protagonist in the same movie is marked with a down arrow point (black). The matplotlib python code is in the Jupyter notebook" Analysis_Movies&BabyNames_Subplots".

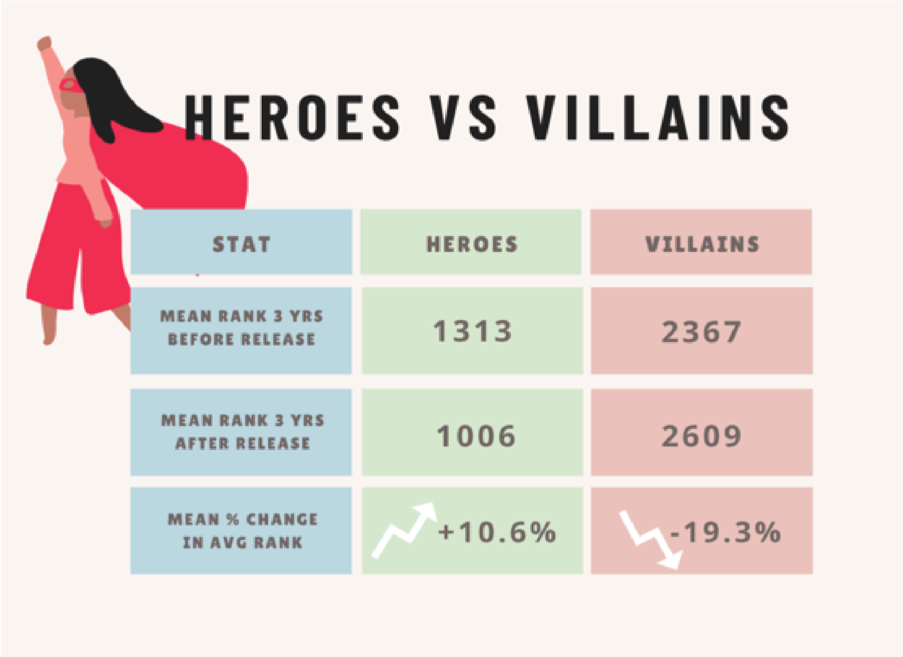
In the third row of figure 5, the current top-3 highest rated movies released in 2013 were Frozen, Monster University and The Croods. For all three movies, the hero/heroine character name increased in ranking at least one year after the movie was released; this suggests that the popularity of these movies as well as the character names had a positive influence on given baby names. The villain character name in Frozen shows no distinct pattern in ranking between 2000 and 2020. However, in the year following the movie release, the ranking for the boys' name "Hans" increased dramatically. This indicates that the villain movie character name became popular possibly due to the movie Frozen.

The top-3 highest rated movies of 2011 show no clear influence on given names to babies born in the years following the release of these movies: the movie character names ranked high prior to the movie releases and fluctuates in more recent years.

In the bottom row subplots of figure 5, the 2009 top 3 highest rated movies are: Up, Coraline and The Princess and the Frog. Although a popular movie, the hero character names in Up decreased in rank in the year following the release. The given name to baby girls, Coraline had the lowest ranking until the release of the movie, where both the movie title and the heroine character share the same name. There is an increase in the rank of the name Coraline in the years following the movie release. The heroine Tiana in The Princess and the Frog also increases in rank a year after the movie is released. Although, the popularity of the name and influence of the movie was short-lived as the rank decreases in 2012 and further in 2014.

IV. Are more babies named after heroes/heroines than villains?

In the 'Exploratory_Analysis' notebook, the data was split into two data frames to separately analyse the ranks of baby names that corresponded to hero and villain characters. The key statistics gained from this analysis can be seen in figure 6. Overall, the 'Hero' baby names rank higher than villains, even before release of movies. This suggests movie writers base the names for villains on less popular names, and heroes on more popular names.
In addition, when a movie is released with a hero named in it, on average, this baby name's rank will increase in popularity by 10.6% (for the average rank 3 years after release). Conversely, a villain's name in a movie release, will result in that baby name rank declining, on average, by 19.3% in the subsequent 3 years.
To conclude, more babies are named after heroes than villains and this will be explored further in the next question.



HEROES VS VILLAINS

STAT	HEROES	VILLAINS
MEAN RANK 3 YRS BEFORE RELEASE	1313	2367
MEAN RANK 3 YRS AFTER RELEASE	1006	2609
MEAN % CHANGE IN AVG RANK	↗ +10.6%	↘ -19.3%

Figure 6. Summary of heroes vs villains mean rank three year before and after and mean % change in the average rank.

V. Do movies influence baby names? If so, how?

For the empirical analysis of the data, refer to the 'Hero_Villains_Hypothesis_Testing' notebook. The team decided to conduct a hypothesis test to determine whether the change in rank seen in baby names is significant. In statistics, a significant change is defined by one that has low probability of occurring. This change can be explained as occurring as a result of the external factor we are investigating (an external factor could be that a movie is released with said hero/villain name). After seeing the raw data and the graphs that were previously the team decided to conduct one and two-tail Wilcoxon signed rank tests.

Results:
For the hero dataset there was strong evidence to say that a baby name increases in popularity (this was equivalent to a decrease in rank) after a movie was released with a hero of said name.

For the villain dataset, there was weak evidence to show that a baby name decreases in popularity (this is equivalent to an increase in rank) after a movie was released with a villain of said name.

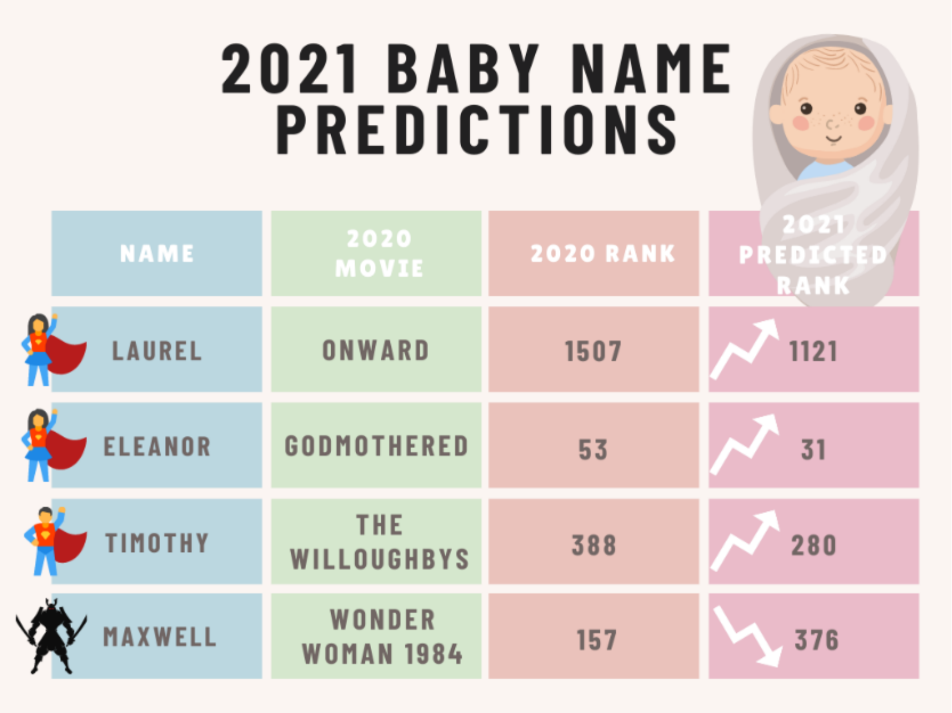
For characters in general, not separating heroes and villains, there was weak evidence that a baby name increases in popularity (this was equivalent to a decrease in rank) after a movie was released with a character of said name. Since the results were so different for the influence that heroes and villains have on baby names, the team reasoned that the overall trend seen for all characters was just a reflection on heroes (positively) influencing the baby names. For this reason, it was important to separate the cases for heroes and villains rather than looking at all character names.

CONCLUSION

The analysis conducted for this report shows that there is a relationship between hero and villain character names in movies and the rank of baby names in the UK. A hero character would, on average, increase the popularity of a baby name by 10.6%, whereas a villain character would result in a 19.3% decrease in popularity of a baby name. This is further evidenced through the statistical hypothesis testing where the team found strong evidence that hero names have a (positive) influence on baby names and weak evidence that villains have a (negative) influence on baby names. Out of the movies the team analysed, Inside Out, The Legend of Llorona and Hotel Transylvania held the most influence on baby names (see figure 3).

PREDICTIONS

As the team found evidence of a relationship between baby names and character names in some cases, they decided to look at whether it would be possible to predict the rank of character names from a selection of 2020 movies.



2021 BABY NAME PREDICTIONS

NAME	2020 MOVIE	2020 RANK	2021 PREDICTED RANK
LAUREL	ONWARD	1507	↗ 1121
ELEANOR	GODMOTHERED	53	↗ 31
TIMOTHY	THE WILLOUGHBYS	388	↗ 280
MAXWELL	WONDER WOMAN 1984	157	↘ 376

Figure 7. Baby name predictions for 2021

Linear Regression modelling was used to fit linear models to the data for both heroes and villains separately in the Exploratory Analysis notebook, with the intention to predict rankings for baby names in 2021 based on 2020 movie releases and the historical baby name data. There was more data for the heroes model, and as a result, the heroes linear model had an R^2 score of 0.92, meaning that the model fit 92% of the data. Whereas the villain linear model fit 66% of the data, so it is slightly less reliable. Four different character names (3 heroes and 1 villain) from 2020 movie releases were selected and inserted into the relevant linear models. The results from these predictions can be seen in figure 7, key findings are that the model(s) predict hero ranks to increase and villain ranks to decrease. This is as expected, and as you can see, 'Eleanor' from Godmothered is predicted to rise from rank 53 in 2020 to rank 31 in 2021. 'Maxwell' is the villain in the 2020 movie 'Wonder Woman 1984' and the model predicts the rank will decrease from 157 in 2020 all the way down to 376 in 2021.

LIMITATIONS AND FURTHER WORK

The limited time for this project meant that the team have only scraped the surface of what could be achieved with this research. Below are a few examples of other areas we would have liked to delve into:

1. Royal baby names – how does the names given to babies in the royal family affect the rank/popularity of given names from the office of national statistics dataset.
2. Did the pandemic and lockdown effect the relationship between movie character names and baby names? If so, could there then be a relationship between character names on Netflix (as well as other streaming services) and names given to new-borns in 2020/2021, compared to movies?
3. Do some movie genres perform better than others when it comes to the popularity of given baby names?
4. Research into a popular movie/tv character names and that name being used for a child but spelt slightly differently (e.g. Esme and Esmae)
5. How does education, cultural and social class backgrounds impact the types of movies people watch and how does this relate to the baby names they choose.
6. Demographics – what are the ages of people watching the highest rated/popular movies. Are they of child-bearing age? Are there more males or females watching movies? ONS reports that mothers aged 35 or older choose more traditional names. Therefore, are women under the age of 35 more likely to be influenced by movies, when choosing baby names?