Why Your Child's Name is Keanu

(and other "popular" names)

CSCI 403 - Databases - Project 9

Hanna Barringer, Dylan Kern, and Harry Krantz

May 4, 2017

1 About the Datasets

1.1 Baby Names

The Baby Names dataset is a collection of baby names reported to US Social Security from 1880 to 2014. It includes any name that was reported at least five times in a year. Since the minimum number of five is so low, we have a large number of names that we have never seen before, such as Zmarion, Brij, and Kla. This comes out to almost 2 million rows of data with many repeated names but for different years. This means that we are able to look at how popular names are through several years. The dataset was provided by data.gov, and then compiled into a single file by Kaggle. The data consists of an id, name, year, gender, and count of usages. A sample of the dataset is given by Table 1.

Table 1: Sample of Baby Names dataset

Id	Name	Year	Gender	Count
633420	Lisa	1965	F	60268
633421	Mary	1965	F	34270
633422	Karen	1965	F	32874
633423	Kimberly	1965	F	28833
633424	Susan	1965	F	26333
633425	Patricia	1965	F	23554
1100256	Keanu	1990	M	8

1.2 Movies

The Movie dataset came from "IMDB 5000 Movie Dataset" by chuansun76 on Kaggle, and was scraped from IMDB.com. It contains 5000 movies from 1916 to 2016 and has no usage restrictions. We removed non-English movies and non US movies in order to only look at the actors who would affect US baby names. This brought the data set down to about 3650 movies. The data has the movie title, director, release year, and the names of the lead 3 actors / actresses. It also has columns for the movie's budget, gross, and IMDB rating which allows us to gauge the popularity of the movie. A sample of the dataset is shown in Tables 2 and 3.

Table 2: Sample of Movies dataset

Movie Title	Year	Director	Actor 1	Actor 2
River's Edge	1986	Tim Hunter	Keanu Reeves	Daniel Roebuck
Bill & Ted's Excellent Adventure	1989	Stephen Herek	Keanu Reeves	George Carlin
Bill & Ted's Bogus Journey	1991	Peter Hewitt	Keanu Reeves	George Carlin
Bram Stoker's Dracula	1992	Francis Ford Coppola	Keanu Reeves	Anthony Hopkins
The Matrix	1999	Lana Wchowski	Keanu Reeves	Marcus Chong

Table 3: Sample of Movies dataset, Continued

Actor 3	Gross	Budget	IMDB Rating
Ione Skye	4600000	1900000	7.5
Al Leong	40485039	10000000	6.9
Alex Winter	38037513	20000000	6.2
Gary Oldman	82522790	40000000	7.5
Gloria Foster	1.71E+08	63000000	8.7

2 Processing

2.1 Cleaning the Data

The movies dataset was conditioned with Excel to remove unneeded columns and clean the data. There were several hundred instances of character encoding issues. There were also several duplicates in the dataset which caused primary key violations when trying to insert the data into the database. Furthermore, we removed movies that were not produced for the US film industry. The baby names dataset was perfect as found.

2.2 Loading the Data

We uploaded our data to flowers(under dkern account) and saved the datasets to two tables called Movies and BabyNames. To upload efficiently we used the \COPY command.

2.3 Compiling the Data

To determine which names had the biggest impact we made a new table called Rates. First we inserted into this table a row for every actor name, movie title, and year combination from the Movies table. Then the first name was saved in a separate column for each actor in the Rates table. Finally for each row in Rates, the count for the name is found in the BabyNames table for the year preceding the movie release, the year the movie was released and the four following years. The annual growth rate for each year is determined and saved in the Rates table. By ordering the table by the growth rate we found the names with the largest impact.

Table 4: This table shows the actors whose names had the greatest percent growth after a movie release. The annual growth rates are shown for the year the movie was released and the four following years.

Actor	Movie Title(s)	Year	+0 Years	+1 Years	+2 Years	+3 Years	+4 Years
Keanu Reeves	(Multiple)	1991	+975%	+36%	-1%	+80%	+86%
Farrah Fawcett	Logan's Run	1976	+921%	+80%	-78%	-39%	-11%
Denzel Washington	Mo' Better Blues	1990	+796%	+43%	+6%	+13%	-15%
Vivica A. Fox	(Multiple)	1997	+666%	+110%	-44%	+11%	-6%
Charlize Theron	The Astronaut's Wife	1999	+640%	+178%	+88%	-9%	+28%
Djimon Hounsou	Deep Rising	1998	+500%	-81%	-100%		+14%
Macaulay Culkin	My Girl	1991	+420%	+38%	-36%	+30%	-60%
Shailene Woodley	The Fault in Our Stars	2014	+383%	-100%			
Angell Conwell	(Multiple)	2001	+360%	+8%	-12%	+13%	-20%
Faizon Love	A Thin Line Between ...	1996	+340%	-50%	+100%	-22%	+35%

3 Results

Based on growth rate and overall count, we decided to look at 5 names from our list: Keanu, Farrah, Denzel, Charlize, and Macaulay. Other names were omitted because the growth rate did not reflect how popular the name actually was. For example, if a name only had 5 instances that grew to 15 in the next year, then this would be a 200% increase. An increase of 10 names could still be within normal fluctuation.

3.1 Keanu

The most popular Keanu is shown in Figure 1.

Figure 1: Left: Keanu Reeves in 1991 - pre-popularity, Right: Keanu Reeves in 2017 - post-popularity

A plot showing the popularity of Keanu Reeves' movies is given in Figure 2 and a plot of the name count vs year is given by Figure 3. The first 5 years of Reeves' movies did not have any noticeable affect on the name as the first data point is in 1990. The first distinct peak in the name "Keanu" came in 1995 for both genders, which followed Reeves' highly popular movie in 1994. Another, smaller, spike for both genders happened in 2002 following the release of the Matrix which was highly popular and released in 1999. Since 2002, the popularity of "Keanu" has decreased steadily.

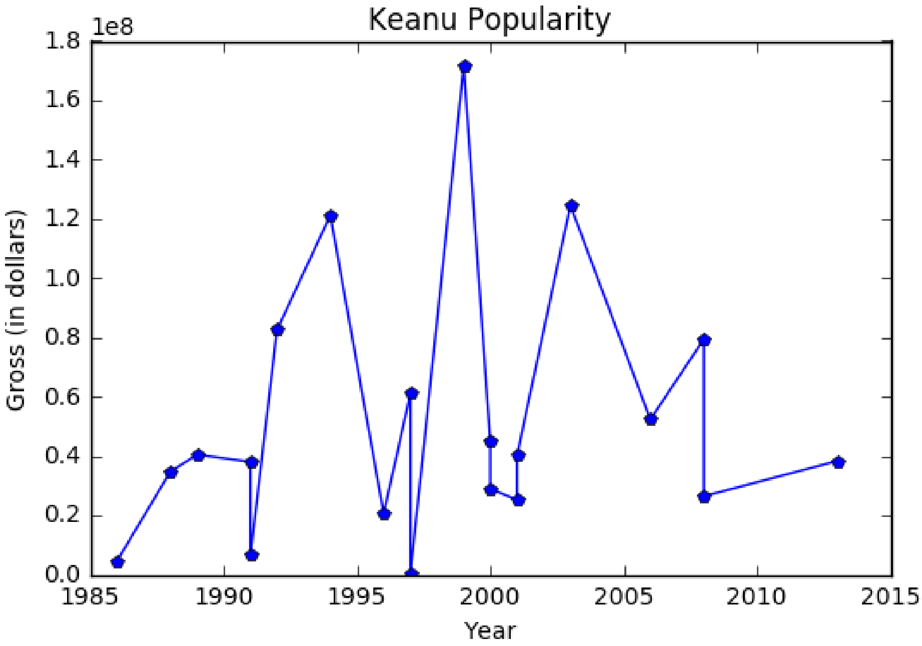

Figure 2: Plot of how popular Keanu's movies were, based on how much they grossed

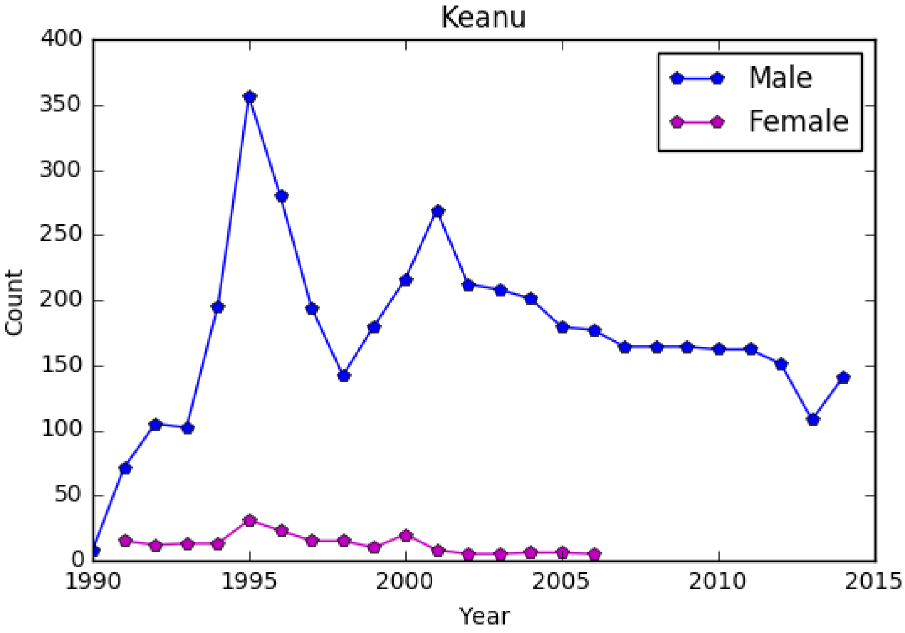

Figure 3: Plot of how many people were named Keanu separated by gender

3.2 Farrah

A plot showing the popularity of Farrah Fawcett's movies is given in Figure 4. She did do more movies, but the database did not include them. A plot of name count vs year is given by Figure 5. Farrah is most definitely a female name, as the male line has a single data point in 1980. The spike in the trend is in 1977, just one year after Fawcett's movie in 1976. It quickly dies off and then has a small spike in 2010.

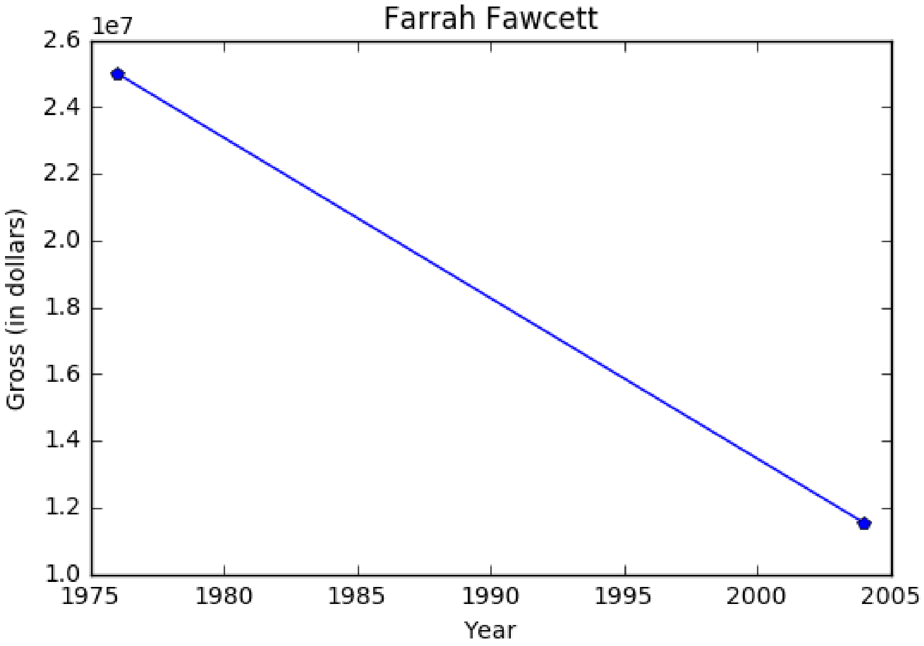

Figure 4: Plot of how popular Farrah's movies were, based on how much they grossed

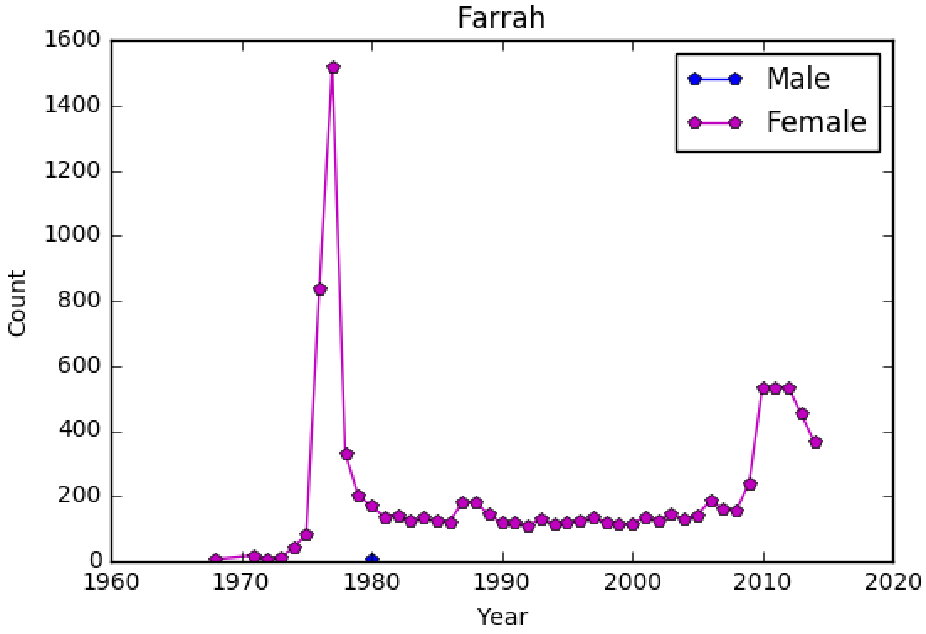

Figure 5: Plot of how many people were named Farrah separated by gender

3.3 Denzel

A plot of name count vs year is given by Figure 7, with Denzel's movie popularity shown in Figure 6. The name Denzel has sporadic but consistent occurrences since the early 20th century. The number of babies given the name Denzel grew when the actor Denzel Washington began appearing in movies in the late 1980s with a large spike in the early 1990's when his first big hits were released, starting with Mo' Better Blues. After this peak the number of new occurrences leveled off at a number much higher than previously.

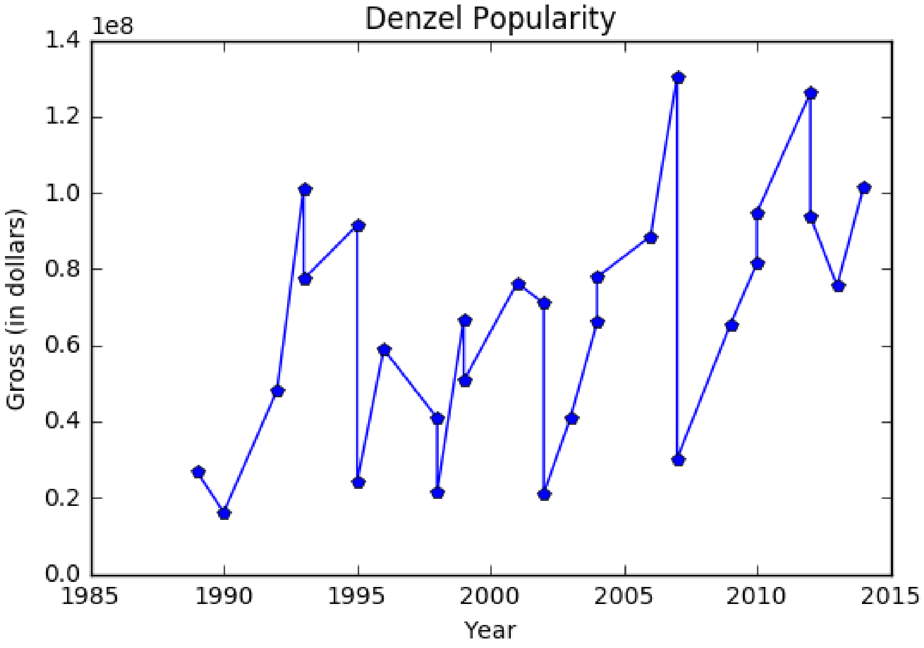

Figure 6: Plot of how popular Denzel's movies were, based on how much they grossed

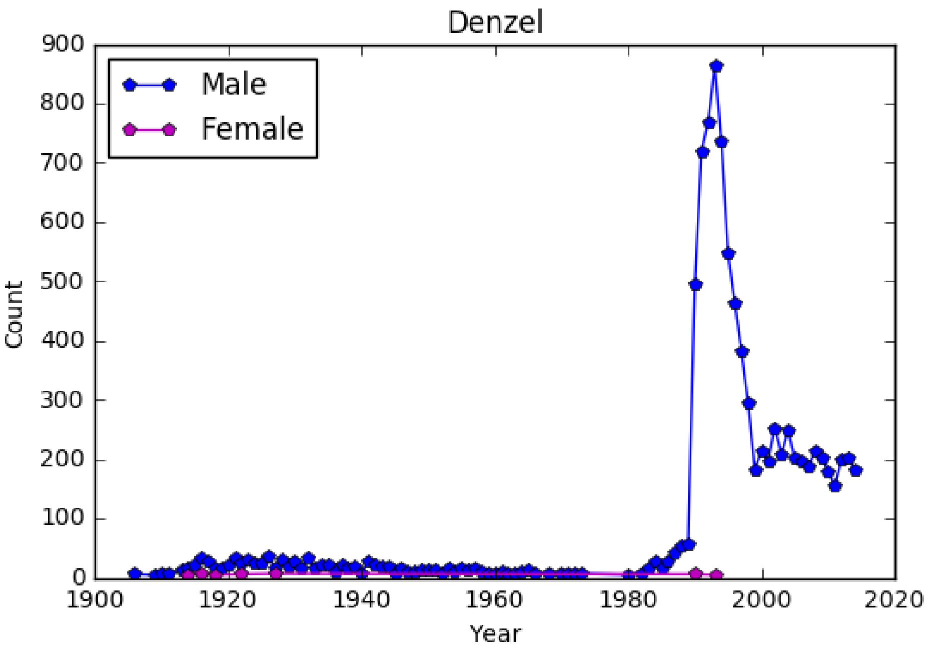

Figure 7: Plot of how many people were named Denzel separated by gender

3.4 Charlize

A plot of name count vs year is given by Figure 9, and her movie popularity is given by 8 . The name Charlize shows steady growth since it came onto the scene in 1998 from actress Charlize Theron. It is possible that advertising for the movie caused initial growth, and the movie popularity then pushed it into more common usages. Her second most popular movie happened in 2003, which directly preceded the peak of the name in 2004. There has never been a year where at least 5 male children have been named Charlize.

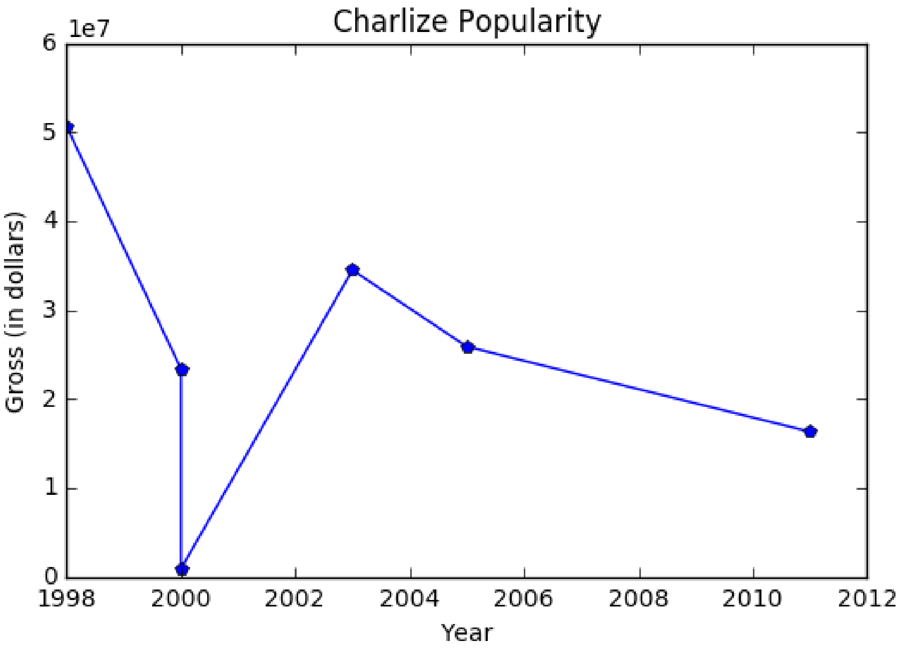

Figure 8: Plot of how popular Charlize's movies were, based on how much they grossed

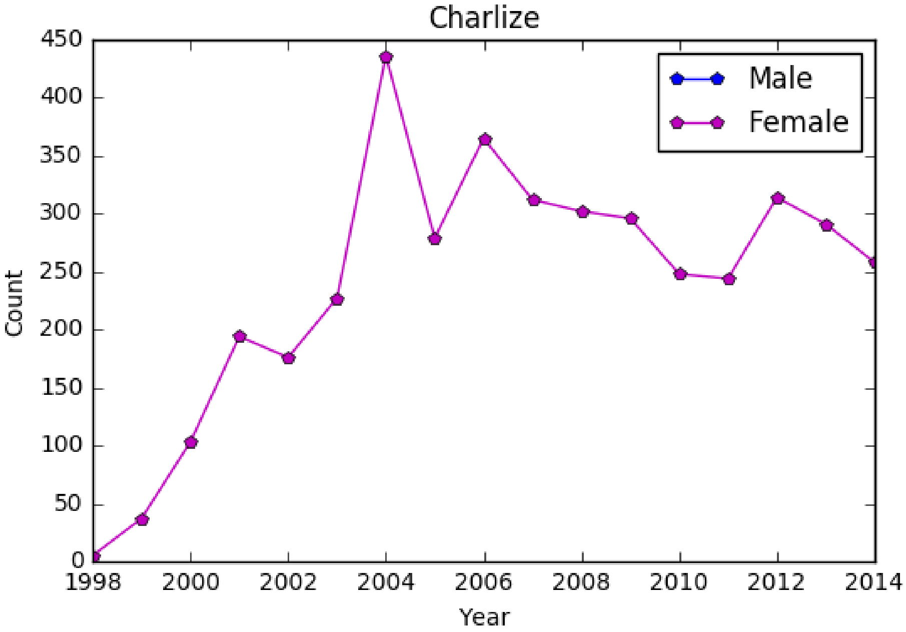

Figure 9: Plot of how many people were named Charlize separated by gender

3.5 Macaulay

A plot of name count vs year is given by Figure 11 and his movie popularity by Figure 10. The spike in popularity follows the Home Alone movie series starring Macaulay Culkin. Starting in 1990, the name grew in popularity to 35, but quickly dropped off and sits around 5. The minor increase in 1994 may be due to Macaulay's second most popular movie in 1992. The female name had a short life with a high of 10, then 5, then dropping below, where we do not have data.

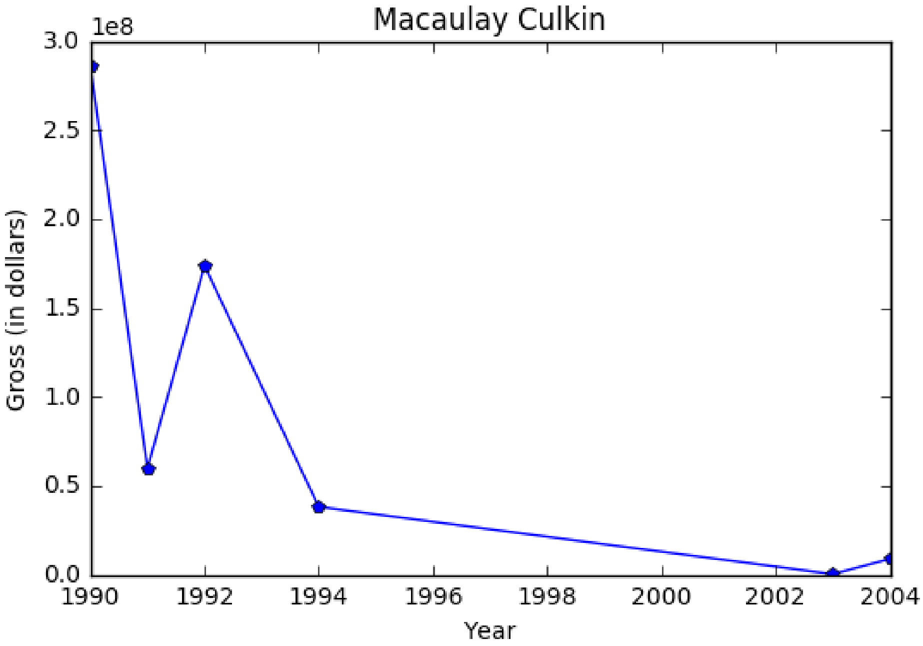

Figure 10: Plot of how popular Macaulay's movies were, based on how much they grossed

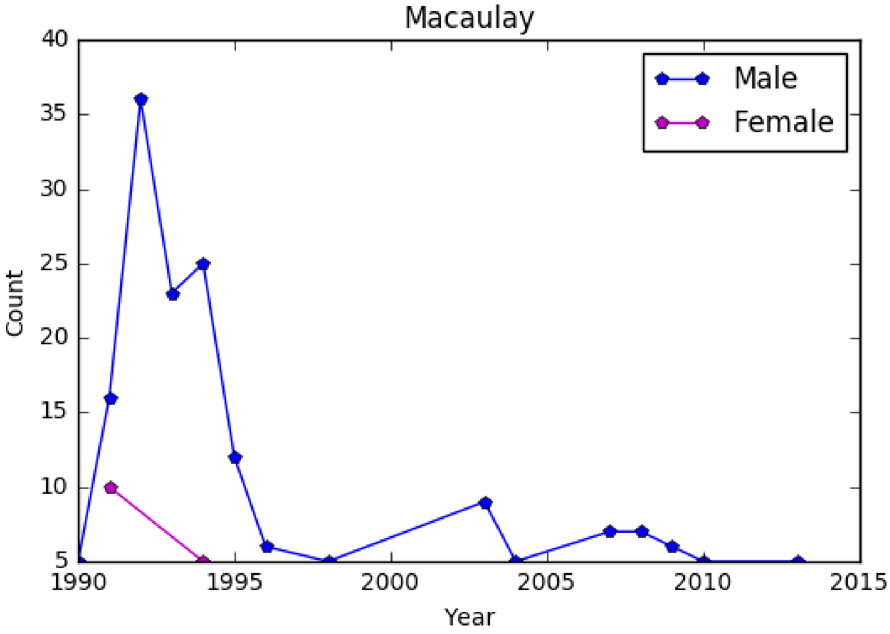

Figure 11: Plot of how many people were named Macaulay separated by gender

4 Technical Challenges

Calculating the growth rates was a challenge. The annual growth rate for name for a specific year was calculated by determining the difference in count between the given year and the previous year and dividing it by the previous year's count. The BabyNames table treats male and female counts separately. Both counts were added together to determine the overall growth rate.

$$rate = \frac{count_{year} - count_{year-1}}{count_{year-1}} \tag{1}$$

The equation was implemented in an SQL statement used to update the Rates table. The SQL statement had to use quite a few COALESCE functions to filter NULL values and replace them with zeros. Furthermore, a CASE statement was used to check if the previous year's count was zero, in which case the entire equation would return NULL to avoid a division by zero error. Although this successfully calculated rates for most rows in the table, it essentially ignored those with counts of zero. This means there is a possibility that we are missing potentially significant growth events in name occurrences. This would happen if the number of counts for a name grew from zero to a large number.

5 Conclusion

We can confidently conclude that there are multiple instances of strong correlation between an actor appearing in a popular movie and the number of babies with the same name as the actor in the years following. From our analysis we cannot conclude if this is a widespread practice for most names or only for less common names. The specific names we looked at are relatively uncommon with a maximum count of a couple thousand versus extremely common names with counts in the tens of thousands. This difference in general popularity makes the impact of a movie more apparent when looking at a single year. It is harder to detect this impact for very common names which see less violent fluctuation.