CSC 466 Lab 1 Report

An Analysis of Popularity, Region, and Gender Trends in Baby Names in the U.S.

Authors:
Sophia Chung, spchung@calpoly.edu
Anagha Sikha, arsikha@calpoly.edu

Abstract

In this report, we studied three different questions that looked at the impact of popularity, geographic differences, and gender differences of baby names. First, we looked at how the popularity of female musical artists in the U.S. impacted female baby names, where we found that successful album releases and deaths of artists were associated with an increase in female baby names after the artists. We also looked to see how the prevalence of common biblical names differs by coast, specifically looking at California versus Florida. We found that Florida has historically had higher proportions of babies with biblical names compared to California. Lastly, we looked to see the gender breakdown of the unisex name “Rory” and gender breakdowns of its spelling variations, “Rori” and “Rorie,” in the U.S. during the 20th century. We found that “Rory” had a higher proportion of being a male name than a female name, while “Rori” and “Rorie” had higher proportions of being a female name than a male name.

Introduction

Do you ever wonder if people actually name their kids after their favorite artists? What about if living on the West Coast versus the East Coast of a predominantly Christian country means you are surrounded by more people with biblical names? How about whether unisex names are actually unisex, and if over time they were more common among men than women or vice versa? Well luckily, you do not have to look any further. The answers to all your questions lie in this report below.

Dataset Description

We used the NationalNames and StateNames datasets. The national file contains information about the frequency of babies with a given name born from 1880 to 2014 in the U.S. The columns in this dataset include the Id, Name, Year, Gender, and Count. The state file contains information about the frequency of baby names in the years 1910 to 2014 for each state. The columns in this data set include the Id, Name, Year, Gender, State, and Count.

Research Questions

1. Is there an association between important events related to top female musical artists and naming babies after these artists?
    a. We have heard people state that they were named after their parents' favorite celebrity or musical artist, so we wanted to see if the popularity of certain female musical artists influenced baby names.
2. Is there a difference in the popularity of the most common biblical names on the West Coast (California) versus the East Coast (Florida)?
    a. Both of us grew up in California, and wanted to compare the diversity of a state on the East Coast to the diversity of our state. We chose Florida because it is the most populous state on the East Coast, and California is the most populous state on the West Coast.
3. Can we track the gender breakdown of the name "Rory" and its other spellings "Rori" and "Rorie" over the 20th century?
    a. There are multiple unisex names common in the U.S. but we noticed that people associate certain unisex names to a specific gender. We wanted to investigate the gender breakdown of the unisex name "Rory," and how the gender breakdown also differs depending on the different spellings.

Methods

Question 1

We decided to look at the top five female musical artists in the U.S. who were active in the 1980s to 2000s: Whitney Houston, Mariah Carey, Selena Quintanilla, Aaliyah, and Britney Spears. First, we filtered the NationalNames dataset to look at only the first names from our list of artists: ['Britney', 'Selena', 'Whitney', 'Aaliyah', 'Mariah']. We also filtered the dataset to look at names from 1970 to 2010. We chose this range as the artists we selected started their careers in the 80s, 90s, and early 2000s, so we looked from 10 years before to 10 years after their debut dates. We then grouped the filtered dataset by Year and Name and summed on the Count values. In order to plot a line from this dataframe, we unstacked the grouped data frame and filled in any NaN values with 0. Then, we plotted Figure 1, displaying the frequency of the female musical artist names over time. We found the peak for each artist by calculating the maximum value for the counts of baby names, and then found the corresponding year, as seen in Table 1. In order to compare the rise and decline of the artists' names, we calculated the percentage increase and decrease one year leading up to and one year after each respective artist's peak, as seen in Table 2.

Question 2

We researched the most popular biblical names, getting the top twelve for each gender (including spelling variations). Using the StateNames dataset, we created separate dataframes for California data and Florida data. For each state, we filtered based on top biblical names for each gender. We then concatenated these more specific data frames to create a dataframe containing all instances of the twenty-four biblical names in California and Florida. We added a binary categorical column to indicate whether a name is a top biblical name, and then we were able to drop duplicates to include names that are not a top biblical name. From here, we created a cross tabulation to calculate the proportion of names that are top biblical names for each Year and State. To visualize this numerical summary, we plotted a line graph, as seen in Figure 2, displaying the proportion of top biblical names over time. Given that we have biblical names for each gender, we also wanted to look at the gender breakdown of these top biblical names by State. Thus, we made a grouped bar chart, with State as the x-axis and sum of Count as the y-axis, all grouped by Gender.

Question 3

We filtered the NationalNames dataset to look only at the name “Rory” from 1901 to 1999. We used cross tabulation to get the total counts of each gender of “Rory” for each year. Then, we calculated the proportion of female babies named “Rory.” We used a line plot to visualize this proportion and found the local maximums to analyze when “Rory” was more often used to name female babies rather than male babies, as seen in Figure 4. We then wanted to see if different spelling variations of the name “Rory” resulted in different gender breakdowns and better understand how spelling affects gender association. Thus, we repeated the previous steps for the names “Rori” and “Rorie,” which we streamlined by writing our own functions. Then, we overlaid the proportion of female babies named “Rori” and “Rorie” on the initial graph, as seen in Figure 5.

Results

Question 1

Figure 1.

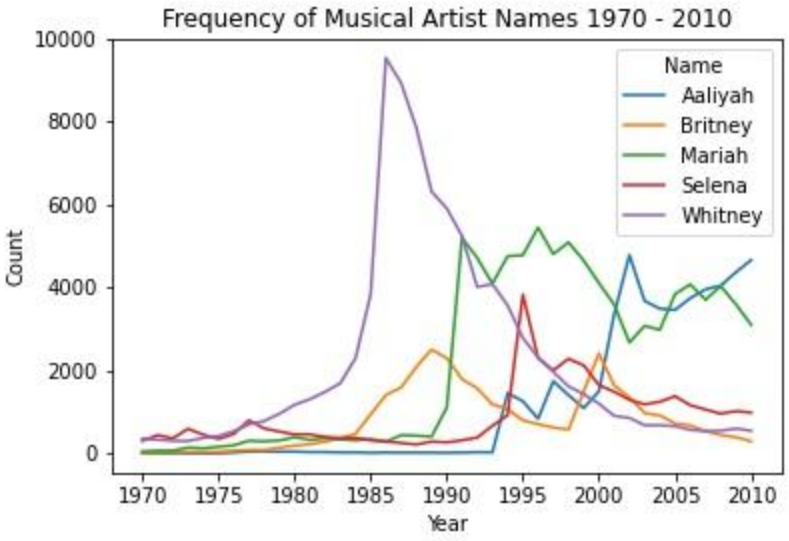


Table 1: Year of Most Frequent Naming of Babies After Female Musical Artists

Name	Year	Highest Peak Count
Aaliyah	2002	4778
Britney	1989	2496
Mariah	1996	5445
Selena	1995	3824
Whitney	1986	9532

Table 2: Percentage Increase and Decrease in Frequency of Baby Names

Name	% Increase in Popularity, One Year Leading up to Peak	% Decrease in Popularity, One Year After Peak
Aaliyah	42.542	-9.517
Britney	19.885	-10.134
Mariah	14.055	-0.587
Selena	317.012	-149.836
Whitney	148.682	-132.794

In Figure 1, we see that Whitney had the highest peak in 1986 with 9532 names. Through research, we found that Whitney Houston released her debut album on February 14, 1985, and it became the biggest selling album by a debut artist at the time, selling over 14 million copies in the U.S. We associate the success of her album release with the peak of the baby name “Whitney” in 1986.

The baby name “Mariah” reached a peak in 1996 with 5445 names. When looking at the history of Mariah Carey, we found that she released her song “All I Want for Christmas Is You” in 1994. Also, her album "Daydream” was released in 1995, which was described as her best album to date. We associate these successes in 1994 and 1995 with the peak in the baby name “Mariah” in 1996.

The baby name “Selena” had the highest peak in 1995 with 3824 names. When looking at the history of Selena Quintanilla-Perez , we found that she released her album *Amor Prohibido* which topped multiple charts in the U.S., broke barriers in the Latin music world, and led her to have the title “Queen of Tejano Music.” We also found that she was murdered on March 31, 1995. We associate the peak in the baby name “Selena” with the unfortunate death of Selena.

We see a peak for the baby name “Britney” in 2000 with 2496 names. Through research, we found that Britney Spears released her debut album on January 12, 1999, which reach #1 on Billboard charts. We associate the success of her album in 1999 with the peak of the baby name “Britney” the following year.

The baby name “Aaliyah” reached its highest peak in 2002 with 4778 names. The artist Aaliyah debuted in 1994 but unfortunately died on August 24, 2001. In 2001, Aaliyah also released her third album, five years after her second album, and this album charted #2 in

Billboard 200. We associate her successful album and death in 2001 with the peak of the baby name “Aaliyah” in 2002.

When analyzing the rise of female musical artists’ names, as seen in Table 2, we see that “Selena” had the highest percentage increase at 317%. This means that the baby name “Selena” more than tripled in popularity in the year leading up to Selena’s peak. When analyzing the decline of female musical artists’ names, the baby name “Mariah” had the smallest percentage decrease at -0.58% showing that the baby name “Mariah” sustained popularity a year after Mariah Carey’s highest peak of fame.

Question 2

Figure 2.

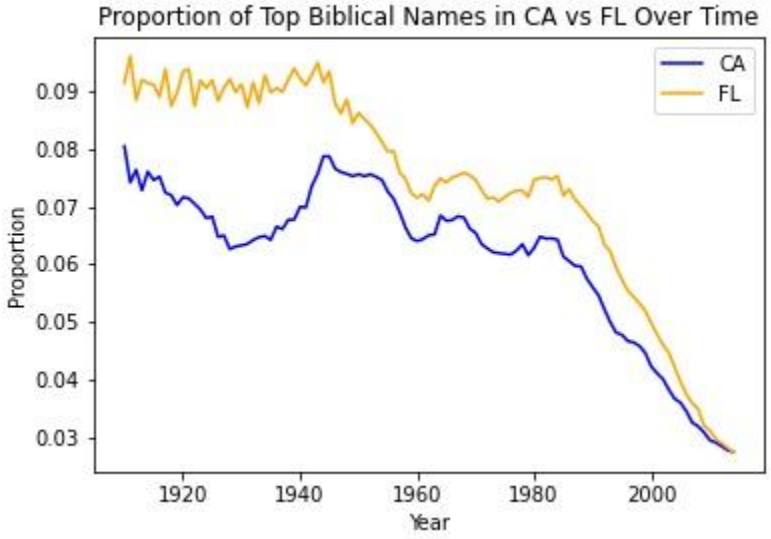


Figure 3.

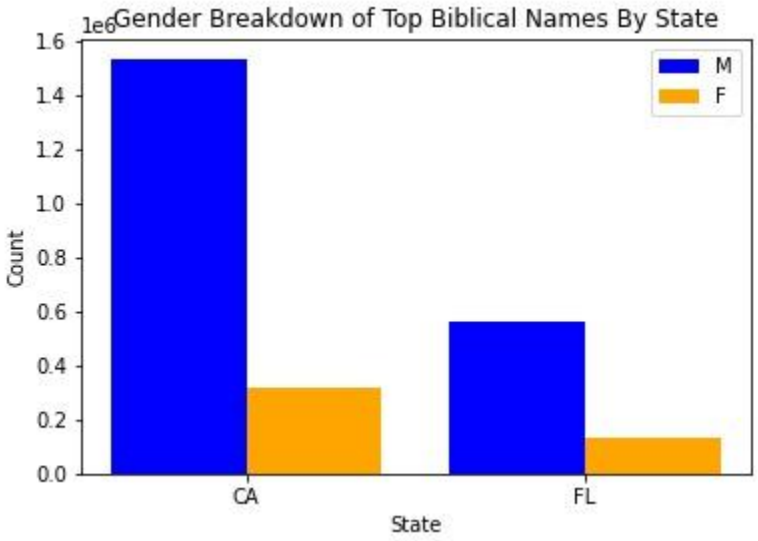


Figure 2 reveals that regardless of coast, these top biblical names are decreasing in prevalence. The last year of the dataset, 2014, marks the all-time lows of these top biblical names in both states. Around 2.75% of babies born in 2014 were named one of the top biblical names. This could be due to the diversification of the American population. We also see that California has consistently smaller proportions of babies with top biblical names compared to Florida. Using cross tabulation, we were able to find that in 1911 Florida, 9.6% of the babies had a top biblical name. This is the largest percentage of babies named after top biblical names for any Year and State combination.

The two lines seem to follow similar patterns except around the time of the Great Depression from 1929 to 1939. The proportion of top biblical names in Florida stays relatively constant, whereas the proportion of top biblical names in California takes a dip. Supposedly, California Church membership increased greatly during the Great Depression, so it seems contradictory that relatively less babies had top biblical names during this time in California.

Because we analyzed our data based on top biblical names for each gender, we wanted to look at the gender breakdown of top biblical names by state as well. In Figure 3, we see that overall in both California and Florida, there were more occurrences of male babies having a top biblical name than female babies having a top biblical name. Although Florida consistently had higher proportions as shown in Figure 2, California has higher counts as shown in Figure 3. This means that overall, California had more babies named with a top biblical name, but Florida had higher rates.

Question 3

Figure 4.

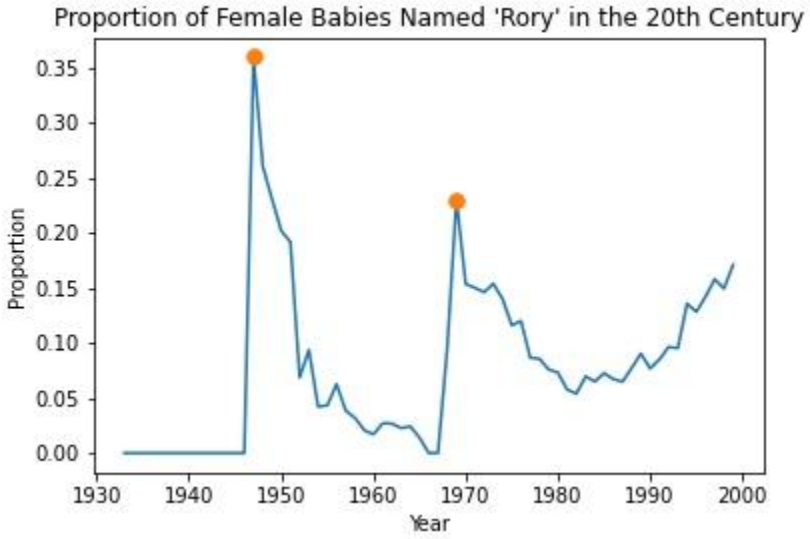
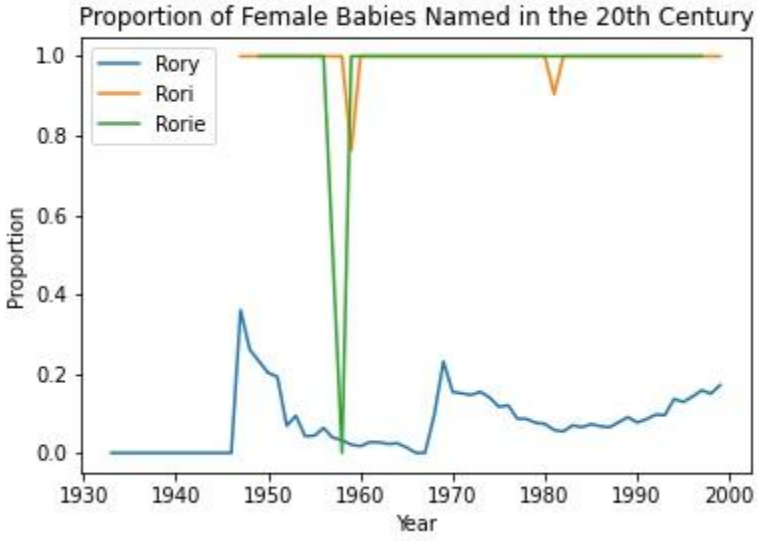


Figure 5.



The first year that “Rory” appears as a name in our dataset is 1933. From 1933 to 1946, “Rory” was only used to name male babies. After 1946, we see that “Rory” was used to name female babies as well. Looking at Figure 4, our line graph has two distinct peaks, one in 1947 with a female proportion of 0.36 and the other in 1969 with a female proportion of 0.23. We found that popular Australian-American actor Errol Flyn named his daughter “Rory” in 1947, and that politician Robert F. Kennedy named his daughter “Rory” in 1968. We associate the rise in the popularity of naming female babies “Rory” in 1947 and 1969 to these notable American figures naming their daughters with this unisex name.

Looking at Figure 5, we see that the names “Rori” and “Rorie” have much higher proportions of female baby names compared to female babies named “Rory.” The orange and green lines corresponding to “Rori” and “Rorie” in Figure 5 stay mostly constant at a proportion of 1, meaning that for most years in the 20th century, “Rori” and “Rorie” were solely used to name female babies. It appears that using the letter endings “i” or “ie” for names is more associated with female babies compared to using the letter ending “y” for names, which is more associated with male babies.

Discussion and Conclusion

Question 1

Overall we see an increase in babies named after popular female musical artists around impactful moments in their careers and lives such as releases of successful songs and albums or deaths. We also see that the percentage increases and decreases of these baby names within a year of the artists’ peaks vastly differed for each artist. A challenge we ran into was trying to find the location of the second highest peaks on the line graph to analyze other moments where an artist’s career could be associated with an increase in the baby names. We found packages to mark all peaks on the line graph, but had issues trying to get those y-values.

Question 2

Looking at the proportion of top biblical names in California and Florida over time, we see that Florida has higher rates of top biblical names than California does. However, both states seem to follow a similar pattern in the rise and decline of top biblical names. Filtering the data set based on gender, state, and whether the name was biblical with matching gender was tedious, and we eventually wrote functions to remove redundancy.

Question 3

From our results on the gender breakdown of the baby name “Rory”, we see that “Rory” was mostly used as a male baby name except for the years 1947 and 1968, where we see a noticeable rise in using “Rory” as a female baby name. Meanwhile when looking at the gender breakdown for the spelling variations of “Rory”, the names “Rori” and “Rorie” have much higher proportions of female baby names compared to female babies named “Rory.” We faced a similar challenge as before in finding the values of the second highest peaks in our line graph. To overcome this challenge, we filtered the dataset to the years 1960 to 1970, after eyeballing the years on our line graph, and looked at our table of proportions. From this table, we were able to find the year of the second highest proportion of female babies named “Rory.”

Bibliography

https://en.wikipedia.org/wiki/Whitney_Houston

https://en.wikipedia.org/wiki/Mariah_Carey

https://en.wikipedia.org/wiki/Selena

https://en.wikipedia.org/wiki/Britney_Spears

https://en.wikipedia.org/wiki/Aaliyah

https://wisevoter.com/state-rankings/states-by-population/

https://rsc.byu.edu/california-saints/depression-years-1929-1939

https://ohbabynames.com/all-baby-names/rory/