CSC 466 Lab 1: A Study of Baby Names in the US

By Kaanan Kharwa and Laura McGann

Abstract

Based on data counting baby names in the US on a national- and state-level from 1880 to 2014, we considered a few questions, including analysis of gender distribution of gender-neutral names, comparative popularity throughout the past century of the name “Laura” vs “Lauren,” and how migration during the Dust Bowl was reflected in baby-naming patterns. After filtering the given datasets to acquire relevant subsets, we visualized the data in various ways to illustrate trends. We drew preliminary conclusions to answer our initial questions and also discovered some other interesting data features.

Introduction

The purpose of this lab is to gain introductory experience to posing research questions, collecting relevant data, and analyzing it to gain further insights. The datasets used contain counts of babies born in the US with different names. The two datasets count baby names on a national and state level, respectively. We collected subsets of data to answer our specific research questions by filtering out irrelevant observations. To make sense of the data after processing, we generated several visuals to identify trends and compare data. Line graphs track different names over time and bar graphs illustrate a more quantitative comparison between babies born with the same name but different gender.

Research Questions

We used the given baby-naming data to answer the following three questions:

1. How has the ratio of male to female babies with common gender-neutral names changed throughout the years 1880 to 2014?
    a. There are many gender-neutral names, so we took five of the most common names - Casey, Riley, Jessie, Jackie, Avery [2] - to evaluate if the gender distribution of a gender-neutral name is, in fact, equal, and if this distribution has changed over time, perhaps due to cultural shifts.
2. When did the name “Laura” start to be considered old-fashioned in favor of the now prevalent name “Lauren”?
    a. This question derives from curiosity from personal experience. The data answering this question can hopefully indicate a time period for the shift in popularity and shed light on the reasons behind the change, perhaps characteristics of the name itself or pop culture.
3. Did the migration from the Midwest (states such as Oklahoma and Arkansas) to California due to the Dust Bowl affect baby-naming habits in California?

a. During the years 1935 – 1950, many people from Arkansas and Oklahoma migrated to California due to the Dust Bowl. We want to see if the naming habits of these states carried over into California.

Dataset Description

We were given two datasets containing baby name counts in the US, one on a national scale and one on a state-by-state basis. The national dataset spanned the years 1880 to 2014 while the state-by-state dataset spanned only 1910 to 2014. Each dataset contains the following columns with the following data types:

Column Name	Data Type	Description	National Dataset	State Dataset
Id	Integer	Unique ID for observation	X	X
Name	String	Baby name	X	X
Year	Integer	4-digit year	X	X
Gender	Enumerator {“M”, “F”}	Baby’s gender assigned at birth (based on sex)	X	X
State	String	State abbreviation		X
Count	Integer	Total number of babies born with the given name in the given year (and state) of the given gender	X	X

For each of the three questions, we collected a subset of the data:

1. We viewed only the national count data - for the entire time span 1880-2014 - for five of the most common gender-neutral names ("Casey", "Riley", "Jessie", "Jackie", and "Avery”), splitting counts for each name by gender.
2. We viewed the national count of female babies born with the names “Laura” and “Lauren” from the years 1880 to 2014.
3. We used the state-by-state dataset to determine the top 5 most common names (ignoring gender) from both Arkansas and Oklahoma in the year 1910. We then collected the count data for those names in California from the years 1910 to 1960. We chose the time period 1910 to 1960 to provide some buffer around the years of the height of the Dust Bowl migration [1].

Methods

For each research question, we filtered and organized the data subsets a little differently.

Question 1

First, we selected only the data concerning the names "Casey", "Riley", "Jessie", "Jackie", and "Avery.” Second, for each name, we split the data into two pandas DataFrames based on gender. Next, we generated one plot per name, plotting the corresponding male and female counts for each year (1880-

2014) on the same plot to compare naming by gender over time. To gain a more cohesive insight into gendered naming practices, we also generated a single plot with two series representing the sum of the five names' male and female counts, respectively, in each year of the same time frame. Finally, to clarify the holistic comparison between use of gender-neutral names for male vs female babies, we generated a bar graph using the two summed male and female count series. Each bar represents the difference between female and male babies born with the five chosen gender-neutral names in each year. A positive (red) value reflects that many more female babies were born with the names than male babies. A negative (blue) value reflects that many more male babies were born with the names than female babies.

Question 2

To investigate this question, we again looked only at the national data, this time specifically for female babies (born in all available years) named "Laura" and "Lauren." We then plotted those two series on the same plot, count vs year, to show when "Lauren" started trending up in comparison to the decline of "Laura." In addition, we found the intersection of the two curves to determine in which year, exactly, "Lauren" surpassed "Laura" in popularity.

Question 3

To obtain the relevant data needed to answer this question, we created three subsets of data, each containing the names from one state (Oklahoma, Arkansas, or California) during the years 1910 to 1960. These years were chosen because the Dust Bowl began around 1935 and peaked around 1950 [1]. The reason the dataset begins in 1910 rather than 1935 is to see the initial naming patterns before the migrations. Likewise, the reason the dataset ends in 1960 is so we can see naming patterns after the peak of the migrations. After selecting out the data for only the relevant states and years, we found the 5 most common names in Oklahoma and Arkansas in the year 1910. There was a lot of overlap of the most common names between the two states, so we ended up combining the two lists resulting in one list containing the names "John", "Mary", "Ruby", "Ruth", "James", and "William." Lastly, we plotted the babies born with these names in California from the years 1910 to 1960 to see how the migration affected the naming patterns during the years 1935 to 1950.

Results

Question 1

Figure 1 below shows the results of the sum of our chosen five names' male and female counts in each year for which national data was available: 1880 to 2014. We can see that to begin with, in 1880 until about 1920, more females are given gender-neutral names than males. Starting in 1920, the count of female babies born with gender-neutral names begins to decline, dropping below the male count around 1928. Around 1950, the female count starts to rise again, surpassing the male count in approximately 1956 for a brief spike before dropping dramatically after 1960 and again falling below the male count around 1967. The female count remains below - although closely tracks - the male count until about 1993 when it spikes rapidly, quickly surpassing the male count in 2000 and continuing such a trajectory until the end of data in 2014. In contrast, the male count starts to decline rather rapidly around 2005. Overall, the male and female count roughly track each other from the beginning through 2000 when they rapidly diverge.

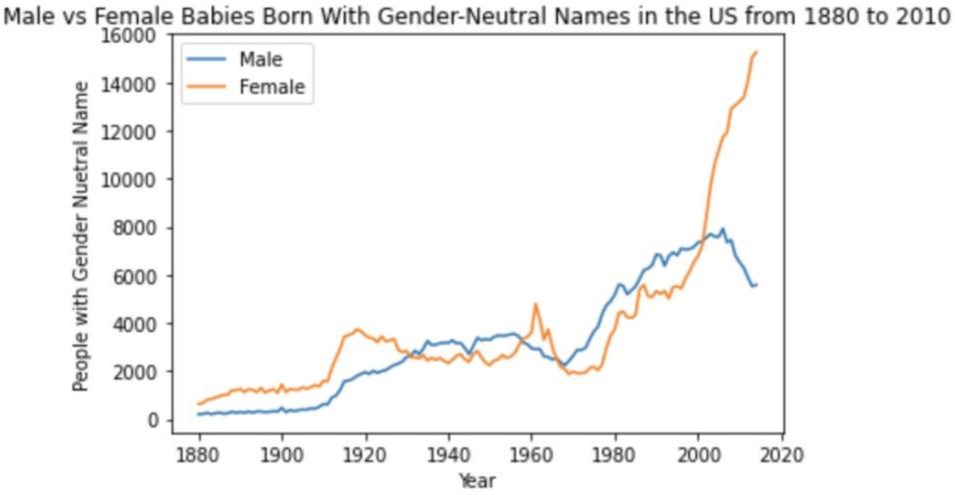

Figure 1. Line plot of male vs female babies born with any 1 of 5 gender-neutral names (Casey, Riley, Jessie, Jackie, Avery).

Figure 2 below yields further insights into how the distribution of male and female babies born with gender-neutral names changes over time. Coinciding with the information presented in the line plots of Figure 1 above, Figure 2 shows oscillations between which of the male or female count is greater. Again, reinforcing information from Figure 1, Figure 2 shows the male and female counts track each other relatively closely from 1880 to 2000 – never differing by more than 2,000 babies – until diverging rapidly after 2000 when the female count skyrockets to almost 10,000 more than the male count by 2014. Additionally, the bar chart tells us there were more years in which more female babies were given gender-neutral names than male babies.

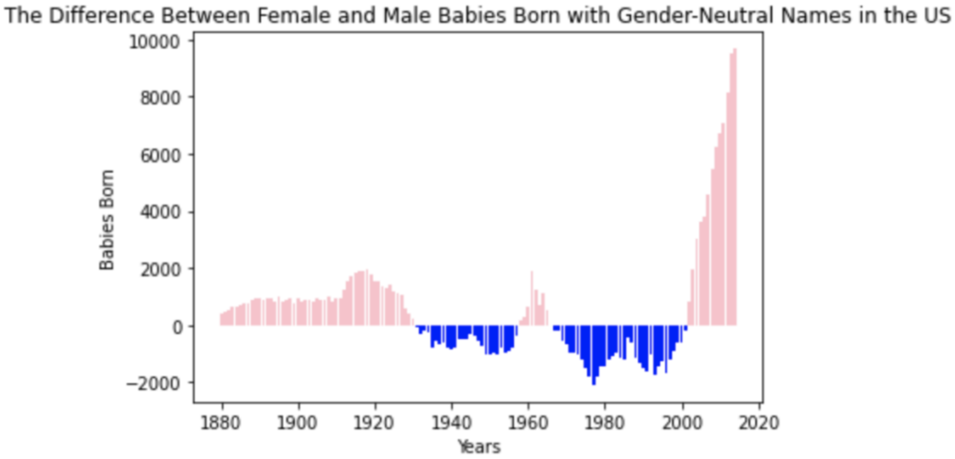

Figure 2. Bar graph showing the difference between female and male babies born with any 1 of 5 gender neutral names. Pink if more females have gender-neutral names in that year. Blue if more males have gender-neutral names in that year.

As shown above in Figure 3, the graphs for “Riley” and “Avery” follow a similar trend where the names initially start of as equal for both men and women but becomes a predominately female name as the date comes closer to the present. Both names start to become predominately female around the late 1990’s to early 2000’s. The graph for the name Jessie tells the opposite story where it starts out as a predominately female name and become more evenly distributed as time goes on. The name starts to get evenly distributed around the year 1980 and this trend continues into the present. The graph for the name Casey shows that the number of males and females named Casey is relatively the same throughout 1880 to 2014, but it is consistent that there are more males named Casey than females. The name Jackie starts off relatively even but becomes more male dominated from around 1925 to 1965. From 1965 to the late 1990’s it is a predominately female name. The after the 1990’s seems to be that the name is relatively evenly distributed amongst males and females. With the exception of the name Casey, whenever the names are at their most popular, it is always predominately female.

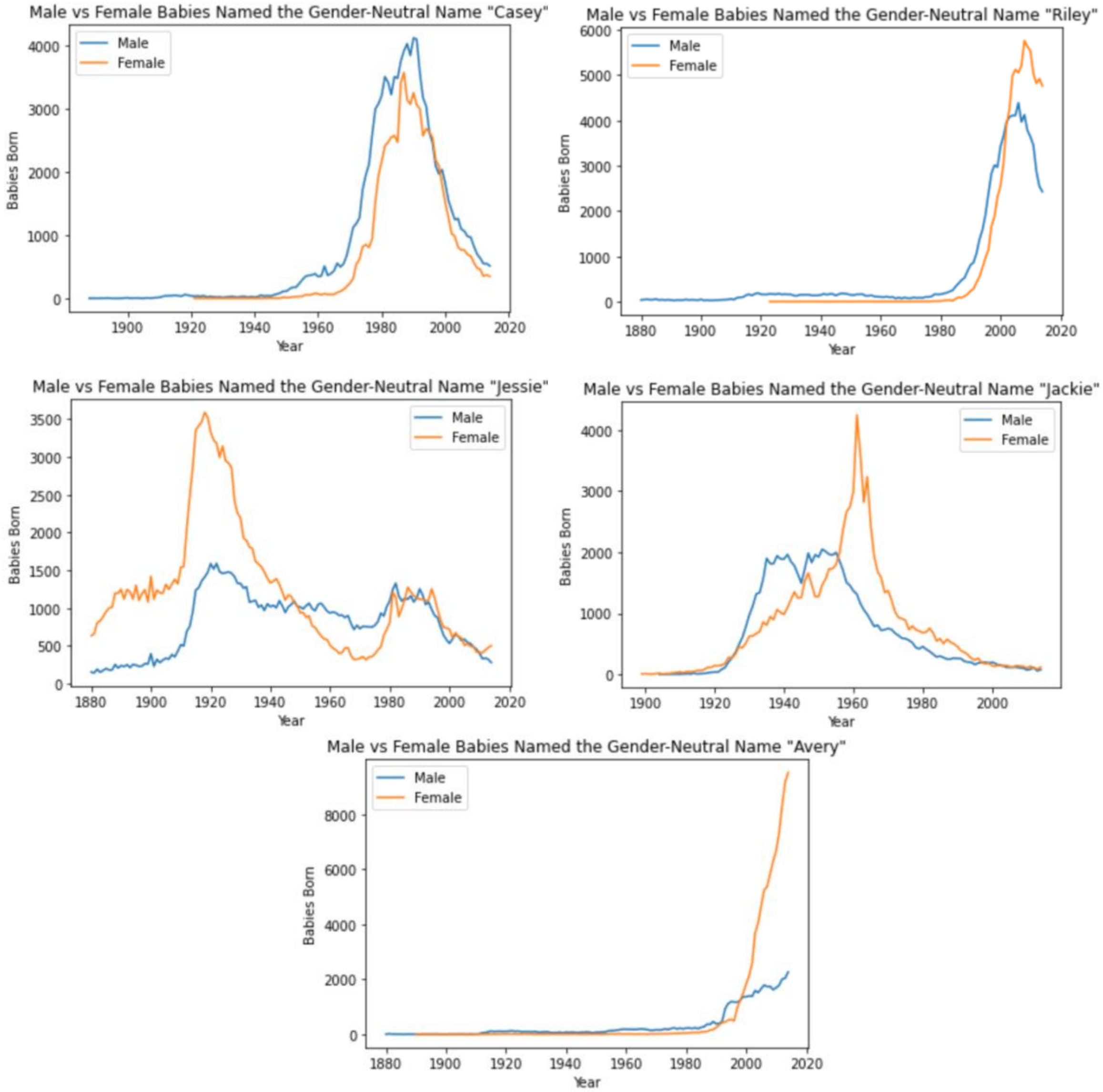

Figure 3. Line plots of male and female babies born with a specific gender-neutral name.

Question 2

Figure 4 shows the naming patterns of the names “Laura” and “Lauren” across the US from the years 1880 to 2014. The name Laura was not very common from 1880 to around 1945 as the number of babies born with those names never surpasses 5000. After 1945, the name Laura seems to dramatically increase in popularity where it peaks from around 1960 to 1970 where the babies born with that name is a little less than 20,000. There is another smaller peak around 1985 with around 16,000 babies with the name, but soon after the naming rate declines.

The name Lauren does not appear until close to 1920, where it remains an unpopular name until around 1975. After 1975, the rate at which babies are named Lauren increases until it peaks around 1990 with a little more than 20,000. After the peak, the naming rate declines until a smaller peak in the early 2000’s with around 15,000 babies named Lauren. However, shortly after, the name continues to decline.

The name Laura and Lauren coexist from around 1920 to 2014, which is the end of the data. Until around 1985, the name Laura is more popular than Lauren, but after, Lauren is the more popular name. Both names peaked at about the same amount, close to 20,000. However, Lauren had the higher peak. One interesting feature is that as the popularity of the name Lauren increases, so too does the popularity of the name Laura. The name Laura was initially in decline, but at around the time Lauren was gaining popularity, Laura’s popularity also increased. This resulted in the second peak of Laura. Another interesting feature is how both names had a smaller peak after their initial larger one signifying a resurgence of the name.

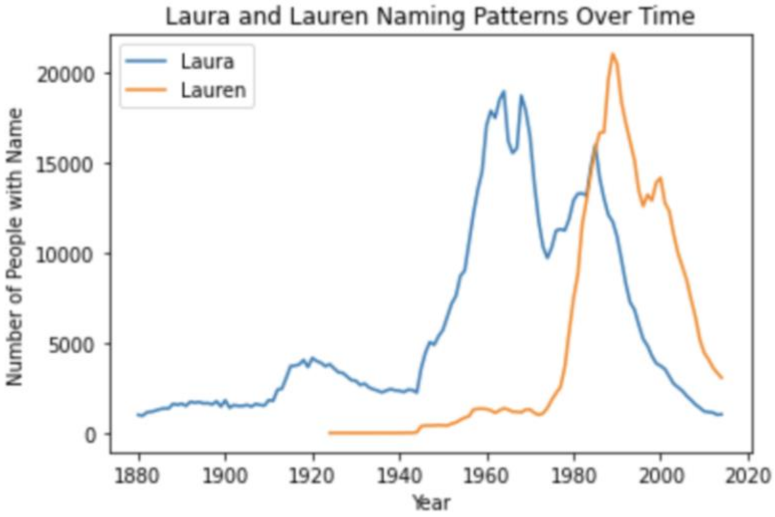

Figure 4: Line plot of babies named Laura and babies named Lauren from 1880 to 2014.

Question 3

Figure 5 below shows the naming patterns for the names “John,” “Mary,” “Ruby,” “Ruth,” “James,” and “William” in California alone from the years 1910 to 1960. The first vertical dashed line in 1935 marks when the Dust Bowl migration began, and the second vertical dashed line in 1950 marks when migration peaked. As we can see, before 1935, the naming rate of these names rose for about 15 years. The names then plateau, however, until the start of the Dust Bowl. Starting in 1935, the naming rates of most increase dramatically, plateauing again – although at a number more than double the pre-Dust-Bowl count – after 1950. An interesting exception to the clear trend are the names “Ruth” and “Ruby” which remain near-constant throughout the time period studied.

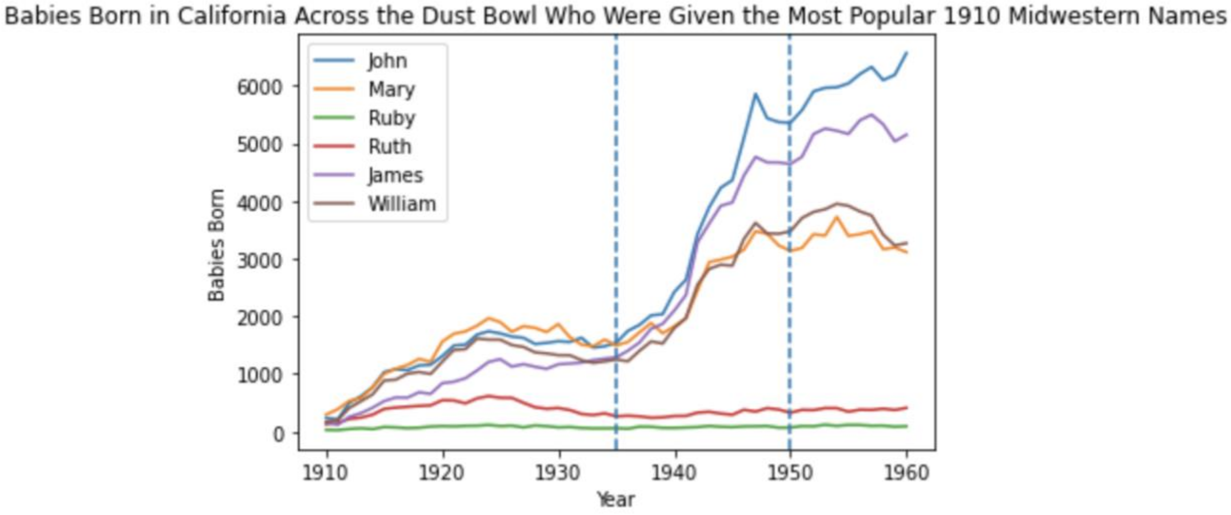

Figure 5: Line plot of most popular names in the Midwest in California from 1910 to 1960.

Discussion and Conclusion

Question 1

One interesting trend we noticed in the name Jackie is how famous people influenced the naming rates. Jackie Robinson started his career in the year 1946 [3]. As the first African American baseball player he was a well-known person, and we can see the increase in males named Jackie around the same time. However, this was also the time when World War II had just ended, so this increase in naming could simply just be from the fact that people were having more kids after the war. Furthermore, popularity of the name Jackie rose rapidly during the presidential campaign of John F. Kennedy in the years prior to the start of his term in 1961 [4]. JFK’s wife, Jackie Kennedy, who became First Lady and was a prominent public figure, may be responsible for the increase in babies named Jackie around the time. This spike was so significant that it caused the total female baby count – across all five chosen names – to surpass

Furthermore, one thing we noticed is that whenever a name is at its most popular, it is always the female graph that is at its peak rather than the male, with the name Casey being the exception. This may indicate that female names are more affected by fads. This may also show how people may be more conservative when deciding a male name for a baby as they are less likely to join in on the trend of a gender-neutral name. However, the parallel increase of male and female babies born with the name Riley in close to modern times may suggest a divergence from traditional naming conventions, but due to the beginnings of seeming decline right before 2014, these spikes may be another fad and cannot suggest anything about choice of gender-neutral names.

Overall, the naming distribution between male and female babies born with gender-neutral names oscillates and seems more dependent on celebrity figures than progressive or conservative attitudes towards gender-neutrality and naming conventions. However, gender-neutral names seem to be largely dominated by female babies. This may suggest the presence of more conservative gendered views applied more strongly to males, but further research needs to be done to be conclusive.

Question 2

One interesting trend we noticed was how the name Laura began to decline from around 1970 to 1975. After 1975, the name Lauren began increasing at a dramatic rate, but also the name Laura stopped declining and began increasing again as well. We think that since the names are similar sounding, the popularity of Lauren is carried over to the name Laura helping it increase its popularity. Alternatively, Laura may have rebounded on its own, in turn causing the creation of variants, such as Lauren, which rapidly gained popularity as a new name. Regardless, after Laura and Lauren reach the same popularity around the year 1980, Lauren continues to increase while Laura dramatically declines. Since Laura again drops off only after Lauren surpasses it, we can infer Lauren's popularity made Laura seem old-fashioned and outdated, a perception parents are noticeably affected by, ultimately causing Lauren to replace the name Laura. In conclusion, the name Laura began to seem old-fashioned around the year 1980, which was 20-30 years after its peak as a name.

Question 3

The trends in naming patterns from 1935 to 1950 show how naming patterns were brought from the Midwest to California. For the names "John," "James," "William," and "Mary," while there was an initial increase from 1880 to 1925, the plateau from 1925 to 1935 suggests Californians at the time were not especially interested in the Midwestern names under investigation. The plateau in name counts after the peak of the Dust Bowl migration similarly indicates native Californians were not responsible for the sudden change. Thus, the more-than-double increase in favorite Midwestern names during the years of the Dust Bowl strongly suggest the influx of Midwestern migrants was responsible for this change. The naming patterns of "Ruth" and "Ruby" suggest the opposite may be true as well, that Midwesterners were influenced by Californian naming habits. "Ruth" and "Ruby were part of the most popular names in the Midwest, but it seems as migrants came to California, they were reluctant to name their children these names. This may be because "Ruth" and "Ruby" seem more Midwestern and, therefore, parents did not want to give their kids these names to help them assimilate into Californian culture. This is but one possible explanation for the differing behavior observed in the names counts for "Ruth" and "Ruby," but further investigation is needed to be conclusive either way. Overall, the increase in common Midwestern names in California from the years 1935 to 1950 clearly attests that the migration during the Dust Bowl brought Midwestern baby-naming conventions to California.

Bibliography

[1] "Dust Bowl Migration - Rural Migration News: Migration Dialogue." *Rural Migration News*, UC Davis, 13 Oct. 2008, https://migration.ucdavis.edu/rmn/more.php?id=1355.

[2] Flowers, Andrew. "The Most Common Unisex Names In America: Is Yours One of Them?" *FiveThirtyEight*, FiveThirtyEight, 10 June 2015, https://fivethirtyeight.com/features/there-are-922-unisex-names-in-america-is-yours-one-of-them/.

[3] "Jackie Robinson Timeline 1919-1949." *MLB.com*, https://www.mlb.com/dodgers/history/jackie-robinson/timeline-1919.

[4] Freidel, Frank, and Hugh Sidey. "Life of John F. Kennedy." *Life of John F. Kennedy | JFK Library*, The White House, 2008, https://www.jfklibrary.org/learn/about-jfk/life-of-john-f-kennedy.