CSC 466 Lab 1 Report: Baby Name Trends in the United States

Rachel Roggenkemper, rroggenk@calpoly.edu
Kirina Sirohi, kasirohi@calpoly.edu
Dr. Alexander Dekhtyar
CSC 466: Knowledge Discovery from Data

Abstract.

We examined baby name trends in the United States using two datasets containing information about the total number of babies with a given name born each year from 1880 to 2014 in the US on a national and state level. Our analysis focused on three areas of investigation: variations of the name “John”, regional trends of Southern names, and impact of historic events. Through data manipulation and visualization, we found that we can view migration patterns of French, Nordic, and Hispanic people through observing variations of the name “John” over time. Additionally, we observed the regional differences of stereotypical Southern names and found that the male names were more common in the Southern region of the United States and observed that there was a big drop in counts in the 1970s, followed by a resurgence in the 1990s, which was probably due to social, historical, or cultural factors going on in the United States at the time. We found a steep decline in popularity of German names in the 1920s through the 1930s which continues to drop through the late 1900s, which corresponds with the timing of World War II. We don’t see a large difference in the popularity of Japanese after the years where Pearl Harbour and Hiroshima/Nagasaki occurred, but we did see a sharp decrease in popularity in the years prior to both Pearl Harbour and Hiroshima/Nagasaki. When we examined the trends of popular Middle Eastern names throughout time, on the national level and on the state level looking at just New York, we saw a steady increase from the 1960s to the 1990s which corresponds to a mass Arab immigration due to the Hart-Celler Immigration and Nationality Act. Additionally, on both the national and New York state level, right after 2001, we saw a very slight decrease, but then the popularity of Middle Eastern names started to increase in the years following.

Introduction.

Baby names are a glimpse into cultural and demographic trends into our society and we can learn so much from them. Parents put in a lot of effort when choosing their children’s names and that is largely defined by their culture, world events, and regional differences. In this report, we investigate different baby names to uncover patterns and cultural influences that allow us to understand the evolution of American society.

Dataset Description.

For this project, we are using a version of the Kaggle Baby Names dataset. There are two data files that we examined.

1. NationalNames.csv: This file contains information about the total number of babies with a given name born each year from 1880 to 2014 in the US. The file has five columns shown in the below table:

Column Name	**Meaning**
Id	Unique id of a row/line
Name	Baby name
Year	Year of birth
Gender	Gender of babies with given name (biological sex at birth)
Count	Number of babies of the given gender who were given the name in the given year

2. StateNames.csv: This file contains state-by-state information about the frequency of baby names in the years 1910 – 2014. The dataset contains six features (columns) which are shown in the below table.

Column Name	**Meaning**
Id	Unique id of a row/line
Name	Baby name
Year	Year of birth
Gender	Gender of babies with given name (biological sex at birth)
State	Two-letter state code
Count	Number of babies of the given gender born in a given state who were given the name in the given year

Research Questions.

The three research questions we focused on regarding baby name trends in the United States on a state and national level are:

1. **Variations of "John":** How do cultural variations and preferences influence the choice of different spellings (John, Jean, Jon, Juan) of the name 'John' over time?
2. **Southern Names:** What are the geographic variations in the usage of stereotypical Southern names, and which regions within the United States have the highest amount of these names?

3. **Historic Event Impact:** When a historic event occurs, how does the popularity of the typical cultural names associated with said event change?

Methods.

We performed our data manipulation and computations in Python, using the Pandas package. Our visualizations were created using the Matplotlib package.

1. **Variations of "John"**
   For this question, we were interested in seeing how different spellings of the name John would show migration patterns. We thought about four different variations: John, Jean, Jon, and Juan, and put those into a new dataframe to examine how they varied over time. We then plotted that to see the overall counts over time. Following that, we created a new variable that grouped each state into four regions and created a bar plot to see how the number of names varied around the country. For that we had to group by region and then summed the counts. This question used both the national and state datasets.

2. **Southern Names**
   We were interested in seeing the regional differences of stereotypical southern names. Similar to the previous question, we started out by creating a new dataframe that only contained stereotypical names. To generate the list of names, we searched online for stereotypical southern names as well as came up with some ourselves. From this we decided that the names should be: Beau, Sawyer, Blanche, Rufus, Atticus, Knox, Dawson, Charlotte, Maryann, Abigail, Dixie, Portia, Calpurnia, and Rhett. Using the state names dataset, along with the new region variable, we grouped by year and region to create a line graph representing the name count over time. We then separated the names into separate female and male dataframes and grouped by year and region and created the same plot as before.

3. **Historic Event Impact**
   i) To approach the question on whether the popularity of typical German names changed in the time period around World War II, we first had to determine what the typical German names were. To do this, we executed Wikipedia search, and thus determined that typical German male and female names were: Jakob, Hendrik, Marie, Arion, Luisa, Viki, Anton, Hans, Felix, Otto, Leon, Wolfgang, Jonas, Gerhard, Claus, Heinz, Hugo, Friedrich, Fritz, Werner, Johann, Emil, Andreas, Wilhelm, Ernst, and Lena. After determining these names, using the national names dataset, we then filtered the dataset keeping only the names listed above. We then grouped by year while summing the counts, and created our corresponding line plots to examine the trend of German name count over the years with this dataset, and whether the trends corresponded with any

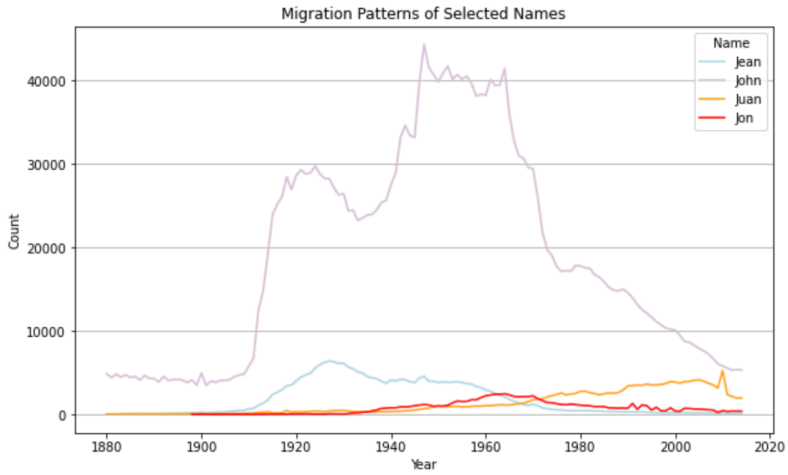
historical events.

ii) To approach the question on whether the popularity of typical Middle Eastern names changed in the time period around 9/11 (so the year 2001), we first had to determine what the typical Middle Eastern names were. To do this, we executed Wikipedia search, and thus determined that typical Middle Eastern male and female names were: Muhammad, Ali, Abdullah, Ibrahim, Amir, Ahmad, Omar, Aisha, Abbas, Hassan, Habib, Ahmed, Malik, Bilal, Hussain, Abdul, Mahdi, Amal, Adnan, Farah, Hussein, Aziz, Adil, and Maryam. After determining these names, using the national names dataset, we then filtered the dataset keeping only the names listed above. We then grouped by year while summing the counts, and created our corresponding line plots to examine the trend of Middle Eastern name count over the years with this dataset, and whether the trends corresponded with any historical events. We were also interested in seeing this trend in just the state of New York (where 9/11 took place), so using the state-level dataset and filtering to only include the state of New York, we then used an identical procedure to examine and visualize the trends.

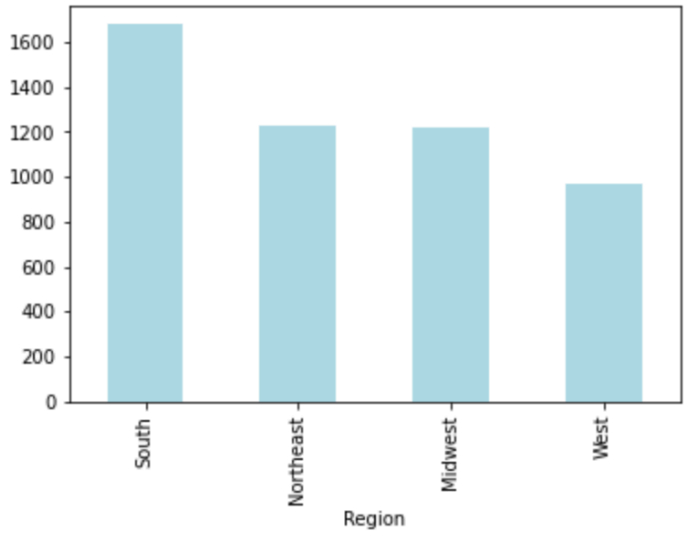
iii) To approach the question on whether the popularity of typical Japanese names changed in the time period around Pearl Harbour, Hiroshima, and Nagasaki, we first had to determine what the typical Japanese names were. To do this, we executed Wikipedia search, and thus determined that typical Japanese male and female names were: Akira, Aoi, Hinata, Haruki, Yuki, Haru, Hana, Sora, Akio, Ren, Sakura, Daiki, Aiko, Haruto, Akari, Emi, Isamu, Akemi, Asahi, Kei, Hikaru, Mei, Jiro, and Hiroshi. After determining these names, using the national names dataset, we then filtered the dataset keeping only the names listed above. We then grouped by year while summing the counts, and created our corresponding line plots to examine the trend of Japanese name count over the years with this dataset, and whether the trends corresponded with any historical events.

Results and Discussion.

1. Variations of “John”

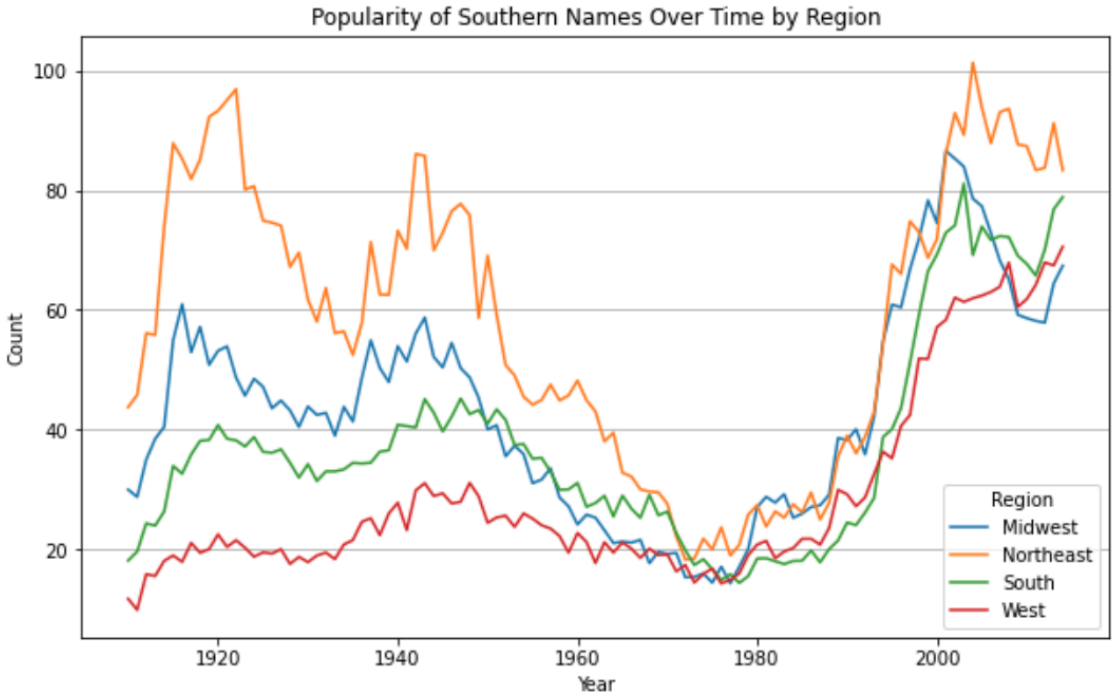


Our first research question involved looking at the popularity of the name “John” through the years to understand different migration patterns. We looked at four variations of the name, those being John, Jean, Jon, and Juan. We stuck to “John” being the U.S. default and that is evident in the plot with that name having the highest counts overall across all years. This roughly follows a normal distribution, slightly bimodal with peaks in the 1920s and 1960s. In general, it is not shocking that this name had the most occurrences because it is the traditional American name. What is interesting are the variations of the name that became popular at different times and how they stayed or left over time. First we will talk about the French variation, Jean. From the 1930s to 1960s, we can see that Jean was becoming more popular and that can be attributed to economic decline in Europe and labor rights movements. Additionally, the US had more job opportunities and better conditions of life so a lot of French people decided to migrate and went specifically to areas in the South as evident from the bar plot shown below.



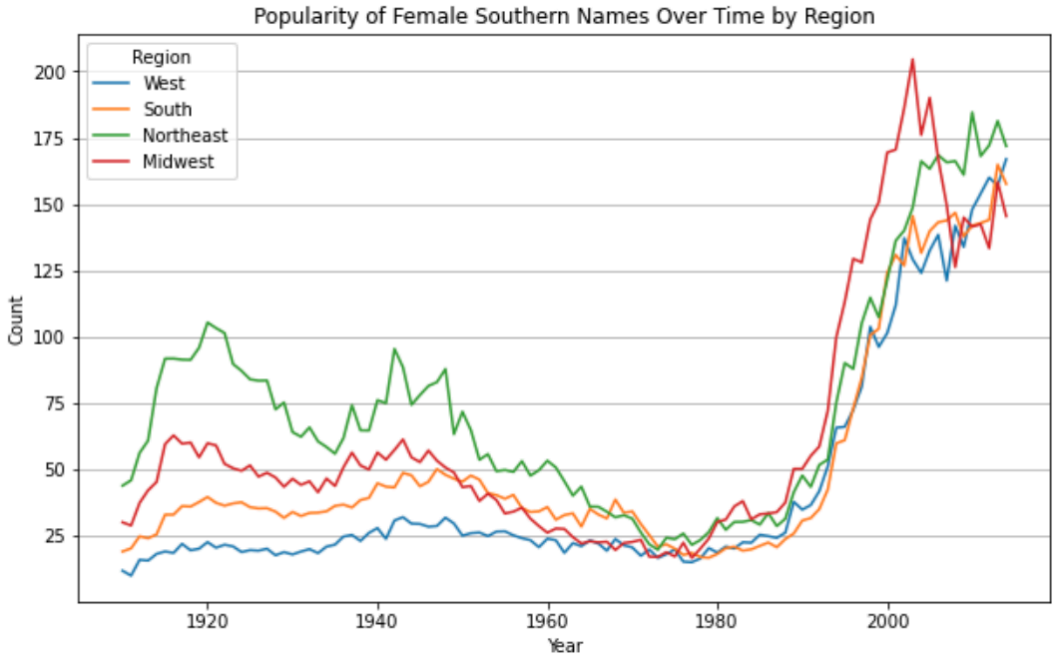
Next, we wanted to see how the name Jon was represented in the United States and we can see that they had a large jump in name popularity in the 1970s. This name originates from Nordic countries like Norway, Sweden, Finland, and Denmark. In Sweden around that time there were major power disputes and the king lost all power which might have led a lot of people to migrate. Finland also was not in NATO at the time and being so close to Russia, many of them might have wanted to flee to a safer country/region. Finally, Juan has been a name that has been steadily rising in the United States since the 1940s and that is most likely due to the United States having better living conditions than other hispanic countries that have been dealing with population growth, falling wages, and weak economic conditions. In sum, we can view migration trends by looking at name variations and see how they have changed over time.

2. Southern Names

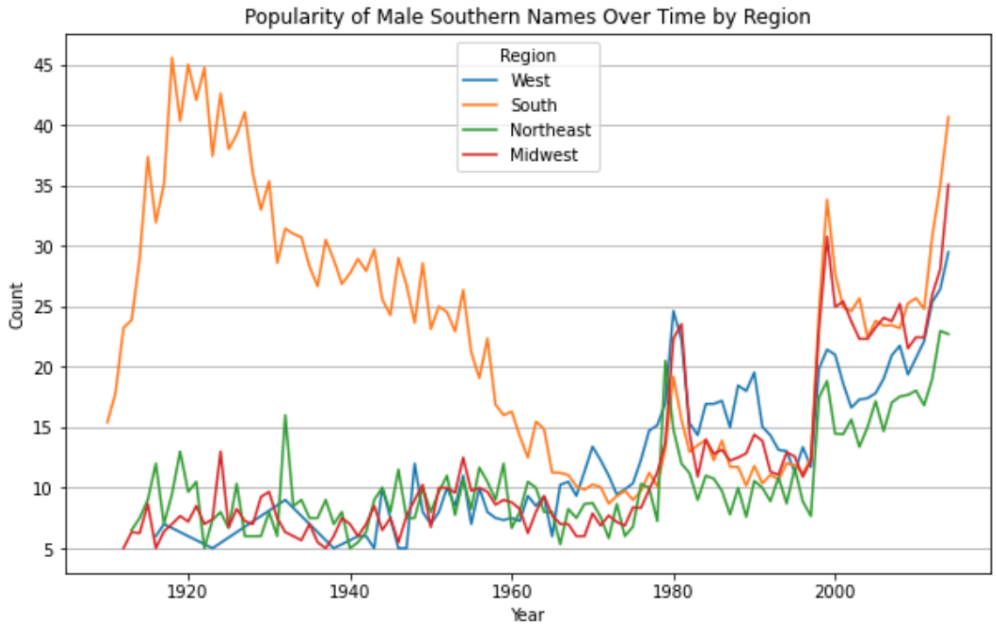


Our next research question involved seeing how stereotypical southern names were spread across the country, both as a whole as well as separated by males and females. We decided to look at a subset of the names Beau, Sawyer, Rufus, Atticus, Knox, Dawson, and Rhett for males and Blanche, Charlotte, Maryann, Abigail, Dixie, Portia, and Calpurnia for females. When looking at the combined male and female names graph above, we can see that on average, the Northeast region tends to have those names at a frequency higher than any other region. The South tends to have the second least frequency for those names which is very surprising. The northeastern region is known for very different cultural differences from the south, but that didn't seem to affect the names it seems.

A potential reason we came up with for the difference in counts was overall population of the regions, but we found that the south region has the highest population in comparison to the other four so that is probably not the reason. The northeast actually has the fewest number of people which is why it is shocking that they have the most southern names. Additionally, looking at the graph, we can see that there was a big dip in the 1970s, followed by a resurgence in the 1990s, which was probably due to social, historical, or cultural factors going on in the United States at the time.



Additionally, we looked at graphs of male and female southern names separately to see if there were any discrepancies between the individual vs. overall graph. As shown in the graph above for females, we can see that it generally follows the same pattern as the combined graph with the northeast leading in names, as well as a dip in the 1970s and a spike in the 1990s. The difference between regions isn't too substantial when looking at females which could indicate that those are just neutral names around the country.



However, when looking at the graph of male names above, we can see that it follows a very different pattern than the overall and female graphs. For the first time, we can see that the popularity of southern names is highest in the southern region of the United States, and it's by quite a bit up until the 1970s when it becomes more normal. Unlike the other graphs, this one does not seem to have a large dip and spike and the name count appears to be rising steadily.

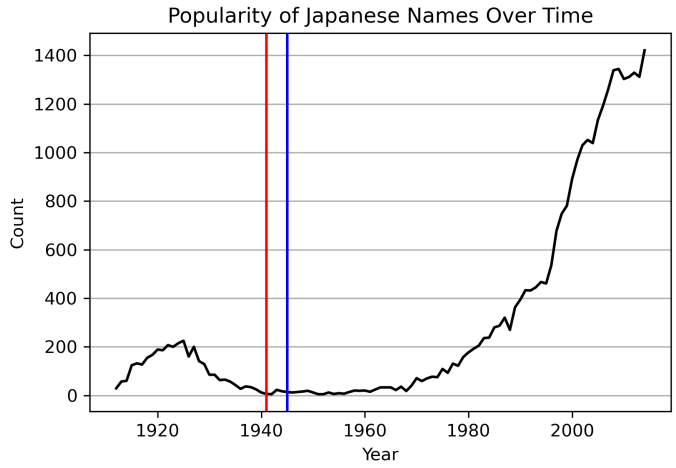
3. Historic Event Impact

i) We first examined how typical German names changed on a national level by year.

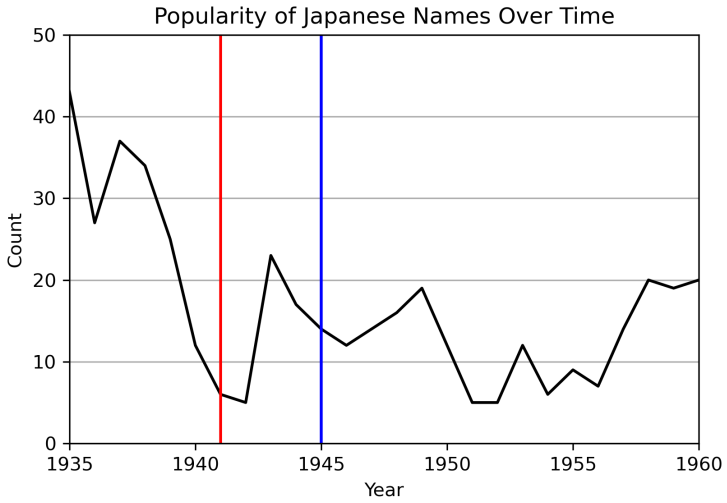
As we can see from the above figure, we see a fairly steady rise of babies with German names beginning in 1880 to the early 1900s, which correlates with when hundreds of thousands of Germans were immigrating to the United States. In 1912, we see a steep rise in popularity of German names, and then beginning in the 1920s through the 1930s, we see a steep decline in

popularity of German names. The popularity of German names continues to drop from the 1940s until around the 1980s, where it levels up and then slowly begins to rise again in the late 1990s. We highly suspect that the reasoning behind the sharp decline in popularity can be attributed to World War II which started in 1939, and the increasing hostility between the Allies (which the United States were apart of) and the Axis powers (which Germany was a part of).

ii) Next, we wanted to look into how typical Japanese names changed in the United States around the time of Pearl Harbour, Hiroshima, and Nagasaki.

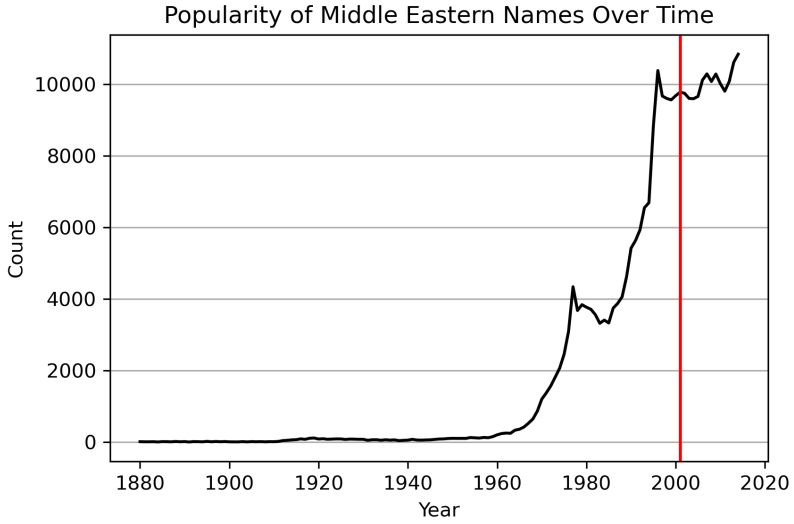


In the figure above, the red line signifies the year Pearl Harbour occurred: 1941. The blue line corresponds to the year Hiroshima and Nagasaki occurred: 1945. In the grand scheme of years throughout the 20th century, we can see that typical Japanese names had a small peak around 1925 because dipping and leveling off until 1970 where the popularity started to increase, and sharply increased around the 1990s. However, to closely examine the pattern around the years of Pearl Harbour, Hiroshima, and Nagasaki, we will now look at the years 1935 through 1960 specifically.

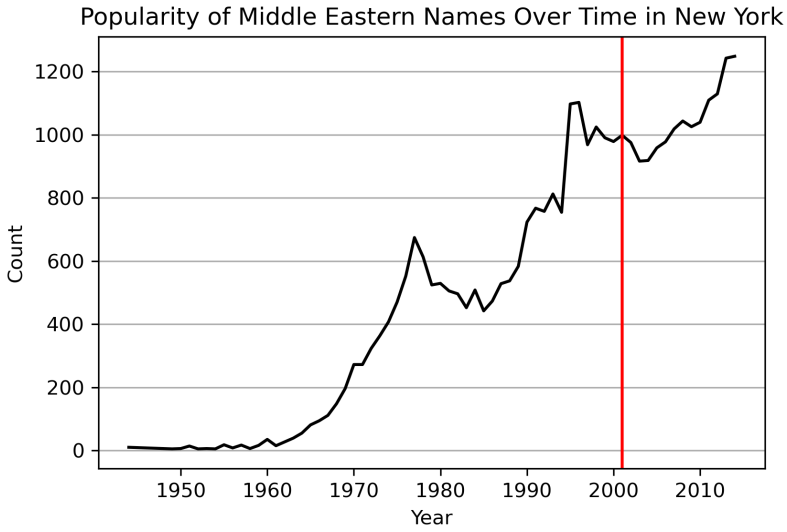


Now that we can see the pattern more closely, we can see that in the years leading up to Pearl Harbour, there is a sharp decrease in Japanese baby names nationally. This could be attributed to the growing tension between the United States and Japan. Right after Pearl Harbour and Hiroshima/Nagasaki, there seems to be a very slight decrease that accounts to maybe a loss in 1-2 baby names. Although we don't see a large difference in popularity of Japanese names after the two corresponding events, we do see a sharp decrease in popularity in the years prior to both Pearl Harbour and Hiroshima/Nagasaki.

iii) Lastly, we were interested in seeing if and how the trends of popular Middle Eastern names changed in the years surrounding 9/11.



In the figure above, we can see that there were practically no babies in the United States on a national level with typical Middle Eastern names until the 1960s, where there is a sharp increase in popularity which continues through the 1990s, which is when it starts to level off. This sharp increase lines up with the mass Arab immigration of about 400,000 people between 1966 and 1990, due to the Hart-Celler Immigration and Nationality Act which took place in 1965 and eliminated the discriminatory quota system from 1924, allowing more people from outside Europe to immigrate to the United States.
The red line in the figure represents the year 9/11 occurred: 2001. We do see a very slight drop after 2001, but it starts to increase in the following years.
We were also interested in how this trend would appear on the state-level, more specifically in New York, which is the year the 9/11 attacks took place.



Looking at the figure above, we see a similar trend in New York as we saw on the national level. There is a steady increase from the 1960s to the 1990s. Right after 2001, there is a very slight decrease, but then the popularity of Middle Eastern names starts to increase in the years following.

Conclusion.

Examining the John variation trends, we observed that the American version of the name had the highest counts, but the most important thing we saw was the migration patterns observed with the other name variations. With a spike in the 1930s, we observed that French people were migrating to the United States due to European turmoil. Additionally, with a spike in the 1970s, we saw that the Nordic name Jon became increasingly popular due to the looming threat of the Soviet Union. Finally, we saw that the Hispanic name Juan has been steadily increasing since the 1940s and will most likely continue due to various factors such as worse economic conditions.

Looking at the Southern names data, we observed that, surprisingly, the Northeast region tends to have southern names at a higher rate than other regions. We also observed that there was a big drop in counts in the 1970s, followed by a resurgence in the 1990s, which was probably due to social, historical, or cultural factors going on in the United States at the time. We also observed that female names generally followed the same trend, however male names were more variable and those names tended to be more stereotypical Southern, as the south region has the highest frequency.

When we examined the trends of popular German names throughout time, in the 1920s through the 1930s, we saw a steep decline in popularity of German names and continued to drop until the 1980s, which we highly suspect is associated with the timing of World War II which started in 1939, and the increasing hostility between the Allies (which the United States were a part of) and the Axis powers (which Germany was a part of). When we examined the trends of

popular Japanese names, although we don't see a large difference in popularity after in the years where Pearl Harbour and Hiroshima/Nagasaki occurred, we did see a sharp decrease in popularity in the years prior to both Pearl Harbour and Hiroshima/Nagasaki. When we examined the trends of popular Middle Eastern names throughout time, on the national level and on the state level looking at just New York, we saw a steady increase from the 1960s to the 1990s which corresponds to a mass Arab immigration due to the Hart-Celler Immigration and Nationality Act. Additionally, on both the national and New York state level, right after 2001, we saw a very slight decrease, but then the popularity of Middle Eastern names started to increase in the years following.

Bibliography.

[1] https://www.thebump.com/b/southern-baby-names
[2] https://www.iamexpat.de/expat-info/german-expat-news/german-names
[3] https://www.familyeducation.com/baby-names/first-name/origin/japanese
[4] https://momlovesbest.com/middle-eastern-names
[5] https://en.wikipedia.org/wiki/German_Americans#:~:text=The%20 largest%20flow%20 of%20German,the%20 largest%20group%20 of%20 immigrants.
[6] https://www.history.com/news/arab-american-immigration-timeline