ISOM 675 - Data Visualization Final Project: United States Baby Name Trends in 1910-2017

Rush Bhardwaj, Yi Cai, Meng Wang, Nicholas Yang

INTRODUCTION

On searching "name trend" on Google, more than 1,210,000,000 result come up. The use case is for new parents, wondering what to name their new born; for social professionals wanting to interpret the meaning behind a name trend; for business owners, who print thank you cards and want to capture the trend and forecast of names in the next couple of years. It is important for any individual to have readily access to an easy tool to analyze such information. On the web, such tools are sparse. Balancing between a simple list of name recommendation and a profound analysis report, our dashboard should deliver a concise quantitative message - a combination of time series analysis and predictive analysis on the count of newborn babies' names. By using the data from the Social Security Administration (SSA) across 137 years with 93,889 unique names, we want to show the trends for different names across different years, states and gender.

We aim to let users be able to answer the question: what is the trend of newborn's name, across different years, states and gender, and what the trend will be in the upcoming year? Users can monitor the information at a glance through analyzing a map, word cloud, ranking chart, and bar chart. Knowing that our users are from a different demographic, we incorporate interactive actions, tooltips, a web url[5] for meaning of a name and search boxes in our dashboard. This way users can customize their experience through selecting on their point of interests and get the

detailed information needed on demand. Our visualizations help analyze the discrete data and extract meaning out of it.

While formulating our ideas on the drawing board, we tried to deliver our quantitative message efficiently and effectively. 2015 AP Top Weekly College Football Rankings[1] dashboard inspired us on how to visualize the ranking charts and put pen to our thoughts and ideas. A Word map (or a word cloud) was taken into consideration to show the popularity of few names in a region. The word cloud also facilitates the interaction between charts and provides more directive options for filtering. We used a bar chart/bar chart to show distribution of a name over every birth year and a rank chart to show the most popular name across the country in a birth year.

DATA AND DATA PREPERATION

Our dataset comes from the SSA baby name database[2](The dataset is intended for public access and use). It is national level data collected from 1910 to 2017 for males and females in all 50 states and the District of Columbia. While exploring the data we observed some names like "Rushang" and "Abhinav" did not appear in the dataset. As per the SSA website[3], if a name is where the year of birth, sex, state of birth is on record, and where the given name is at least 2 characters long will be a part of the dataset. Some names which are unique to each gender, for example: Alexis would be a part of both males and females attribute and instances would be counted accordingly.

To be able to provide our quantitative message we decided to prepare three data frames on R and then eventually export them for dashboarding on Tableau. The first (1) was a compilation of all the states datasets merged into one data frame for our word cloud and frequency

visualization. The second (2) was a cleaner and smaller data frame of (1) with rank of a name by sum of instances and filtered by birth year and gender. The second dataset was used for our map visualization and the tooltip. The third data frame (3) was a subset of (1) which had rank of a name by sum of instances by birth year. This was used for our rank visualization.

We decided not to join our data frames on Tableau as the computation time for making changes to a visualization got long (close to 30 seconds to a minute). While we were working on our map visualization and using the rank attribute to find the top names per state, we saw that some states had two to three top ranked names. This created some troubles for us as all the top ranked name showed up on our visualization per state. We manually edited this information on excel for each occurrence. For example, if both Rushang and Abhinav were ranked one in the state of Georgia for the year 2018, then we made Rushang rank 1 and Abhinav rank 1.3 to make Rushang's name appear on the visualization. For our rank visualization, we wanted to portray the top ten names per generation. This required online research of which period does a generation[4] belong. We incorporated that in our data with a rank of name per year. This was a fairly small data frame for a powerful visualization.

Methodology and Dashboard Design

As mentioned in the introduction, the quantitative message we tried to deliver is a combination of time series and predictive analysis of names for national and state level. User interactions, valuable insights, and predictive analysis are the three main focuses.

To begin, our visualization was focused on how the popularity of baby names has changed over time and what factors have led to this change. First and foremost, visualization we tried was

a map. We used a custom format and not the traditional maps making it a cartogram. The map provides users the chance to interact with the graph by year, gender and states. A user can click on different states and a “floating window” will show the information and a graph related to that specific state (tooltip). From this map, we were able to show audiences specific patterns of the popularity and highlight the contagious effect of a name amongst states. Besides, external factors like cultural influence or popular TV show can be another potential reason for the popularity changing and out of the scope of this project.

Furthermore, the map provided us with insights for how to further design our dashboard. Thus, we decided to use filter and highlighter functions to make interactions within different sheets to provide detailed information. With a such goal, we added a word cloud, a bump chart (rank chart) and a bar chart in the dashboard to provide users with specific information by clicking on a state with its most popular baby name.

The word map displays top 20 names of a state in a given year, with specified text color representing its gender and the text size representing the sum of instances a name has occurred in the database. We designed the bar chart which displays the national-level instances of a baby name through the six generations and the bump chart which aims to show the user the national-level ranks of baby names in a given generation; in such a way, it can provide users a sense how a specified name’s instances and rank changed through time.

We also noticed that a predictive analysis is crucial regards to both the purposes of user interaction and the trend. In such way, parents can use our dashboard as a reference resource for picking up baby names and also see a name’s five year forecast. Business owners can also refer to the prediction and plan their strategy. Thus, we added tableau’s built-in predictive analysis into the bar chart, showing the future trend of baby names for the next following five years.

Design Evolution & Functionality

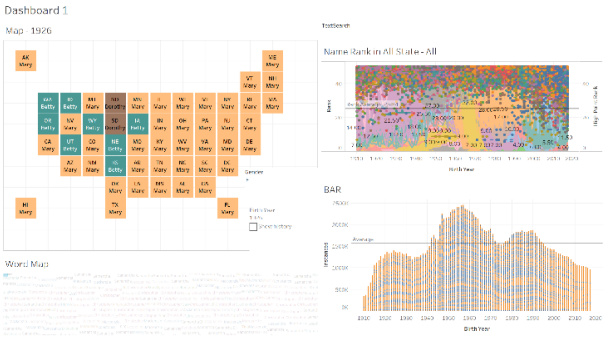

Figure 1

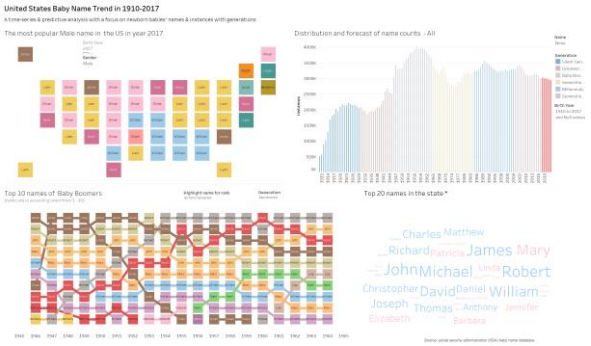

Figure 2

To understand our process of selecting the best visualization, figure 1 and 2 portrays where we started and our final dashboard design. We wanted to start with a map where we were able to highlight the names and the state, and we did that by adopting a cartogram layout from the very beginning. We added a row and a column section for every state in our data and were able to plot a box shape visualization. We took out the name of the state and put it in the tool tip if the user would actually be interested in knowing the state name. Our initial rough draft included a bar chart which plotted the frequency of top 50 instances of a name but find it hard to convey the information clearly. We decided to change the visualization with a periodical color scheme (using generation) and used our (1) dataset to show the sum of instances without any rankings involved. We wanted to show the top names on our dashboard. We started from a "name rank all state" visualization but the visualization was very unorganized. We switched to a simple line chart, but it could not depict the quantitative message we wanted to show. After doing our research, we decided to draw our inspiration from the visualization provided in module two (2015 AP Top Weekly College Football Rankings). We were able to provide our quantitative message of the top 10 names in every year.

We further added a filter of generation and plotted the visualization with apt technical precision. As we were doing many calculations on the word cloud, our computational time was very slow on Tableau. We eventually decided to use our original dataset and index over rank on Tableau. This improved our computational time and still provide the same visualization. The only change we did was to change the colors by name to gender.

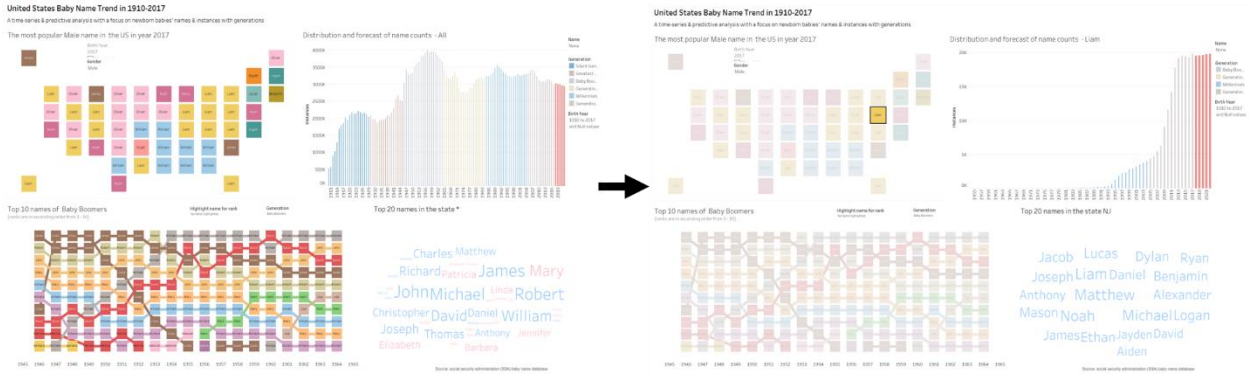

Figure 3

We used the map as a filter (figure 3) for the word cloud visualization to show the top 20 names in a state when selecting a state. We did not want to complicate the word cloud and the sole purpose of the word cloud is to show some important names and not any sort of detail oriented quantitative message. We also wanted to use the map to filter the bar chart visualization by name (figure 3). Users could select a common name in a state and see their five year name forecast. This filtering is different from the tooltips as the tooltip shows the distribution of name in the particular state and not the overall in the country and at the same time there is no forecast involved in the tool tip.

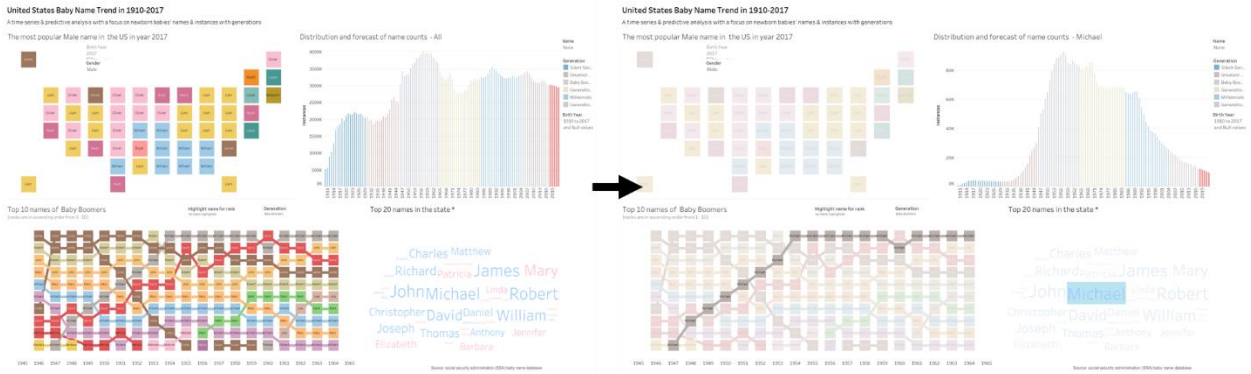

Figure 4

We used the word cloud as a filter for the same idea as the previous one (figure 4). To be able to see a popular names' distribution was at the core of our visualization. The main filters were coming through the map and the word cloud. We also used the two visualization as highlighters (figure 5). This time while using maps and word cloud we introduced a highlighter in the rank visualization by name. As the rank visualization is very rigid and cannot be filtered, we used a popular name highlight to see if the name is top ranked for a generation. We also added a highlighter to the word cloud where we chose the name and see the state's top name.

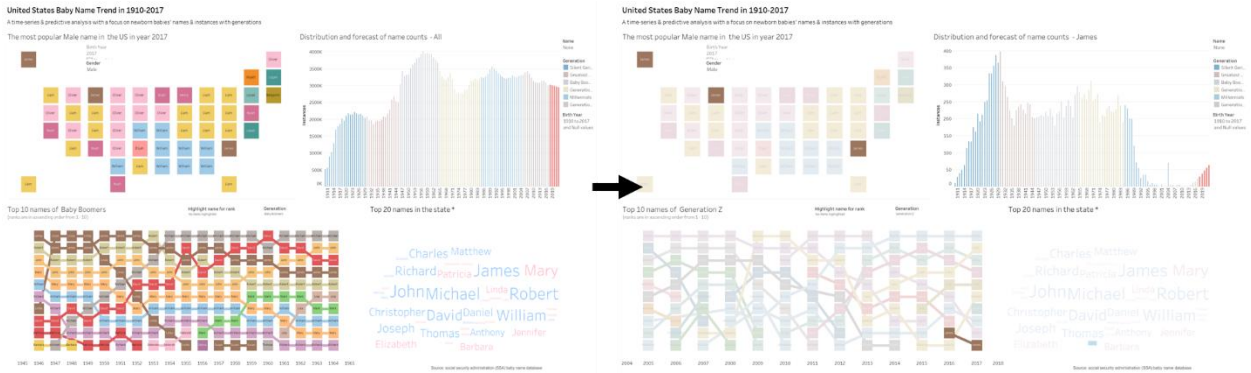

Figure 5

We had 6 visual cues in our visualization. We had two birth year cues, one had a playback slider (figure 6) and the other was a slider tool. As shown in figure 6, we used the playback slider (Birth Year) to depict the trend in change in name in the visualization. As it was synchronized with the map visualization, we decided to move it on top of the map to avoid any confusion for the user. For the same reasons, we moved the gender filter with the synchronized playback slider.

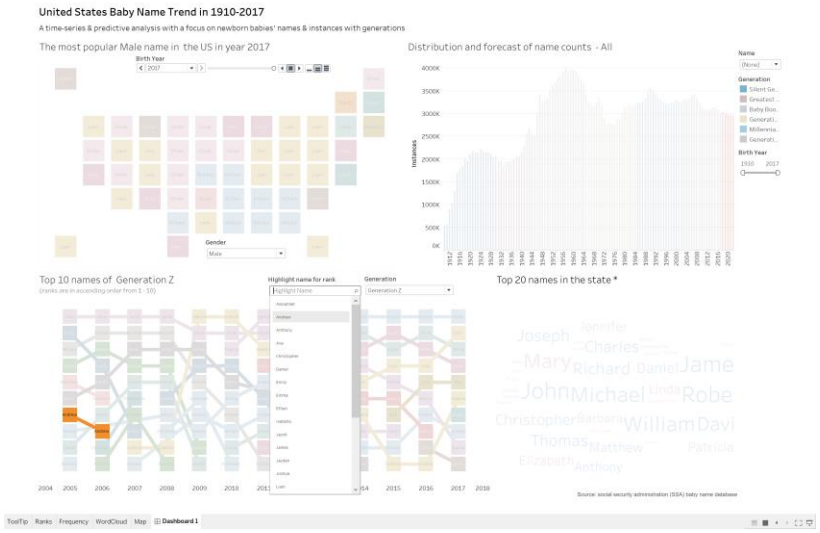

Figure 6

We wanted to highlight the trend of a top ranked name, so we introduced a drop down highlighter (figure 6) on the rank visualization to avoid confusion with the "name text search bar" in top right corner in figure 6. The highlighter dropdown was also connected to the map to highlight if the top name existed for a specific state for a generation. The name text search bar was provided right next to the frequency visualization to highlight trends in the distribution of a name and show a 5 year forecast. We also let the user be able to type few letters and be able to see a drop down list.

Our quantitative message required the names to be highlighted by color. For this we stuck with the Tableau 20 color pallet to highlight names. We could not use a gradient color scheme or triadic color scheme as there were many names. Also, on the map and a rank chart we want to highlight the trend of change in top names, this could not have been done without multiple colors. We used a binary color connotation for the word cloud. Men were blue, and women were given pink. As the distribution involved showing the generations, we added a dull color scheme. To not confuse the user by the colors chosen in map and the rank visualization, we made the color scheme for the bar chart visualization dull in contrast and hue. To focus on the 5 year forecast we decided to highlight it in red and increase the size of the bars. For our text we used Tableau default. In our visualization, we tried maximizing the data ink ratio rule set by Tufte. We kept a simple white background of all the visualization and dashboard, took unnecessary axis names away, deleted all the gridlines for the map and the rank plot, except for the bar chart. As we did not have long written lines in our dashboard, a sans serif family font was to be used and tableau default fulfill our needs.

Insights & Evaluation

By using our ideas, we can clearly see and understand the trend of newborn babies' names. Furthermore, such a trend has a high level of transitivity as popular names always started with a spreading central state. Then these central states followed a similar path as an epidemiological transition process to influence direct neighbor states until nationwide. Popular name trends commonly last for over 10 years. Those most popular/common names such as "Mary" lasted for over 20 years. However, we do find that the common names' active duration is decaying through these 100 years as the contemporary culture is more diverse and more distinct than before.

Although the overall culture environment shifted drastically over a century of time, population clustering/gathering are still largely based on geographical condition. Hence, newborn babies' names were distributed/segregated with clear locational partition evidence. With new names entering the pool, we also found that both East Coast and West Coast areas served as the entrance or initiator of new names. We believe this phenomenon might have a substantial correlation with internet access status, which requires us to conduct further investigation in the future. On the other hand, we revealed that besides interstate influence, name trend follows the step of popular TV shows/stars more frequently in the past 20 years. Such finding surprisingly justified our expectation of the connection between names and popular culture (internet/information access). For example, baby girls' name "Emma" started to take a trend in the year 2002, which we believe was inspired by the popular TV show Friends. Same patterns were found for both male and female babies' names. Finally, from the babies' name distribution bar chart, we can easily distinguish between generations of newborn babies even without the help of coloring. The newborn population generally follows a normal distribution curve with reflections of history events/eras. But it also reveals the seasonal nature of the current population structure as Millennials were born from the generation of Baby Boomers. Hence, we have more Millennials than Generation X's. Despite the fact of decreasing fertility rate, we should expect the population of the next generation to be larger than that of Generation Z.

With this visualization project, we can easily deliver the quantitative message we intend to deliver - a time-series predictive analysis with a focus on newborn babies' names & instances with generations. This project will enable audiences the ability to conduct further researches related to population shift, generation comparison, human resource allocation, and related-industry analysis. As a demo of our greater plan, we designed this project with an intention to attract parenting and

gift industries' business owners. Moreover, we were hoping this visualization can give newborn babies' parents a platform to be better prepared for finding a balance between common and unique names. For business owners, it is convenient for them to have a straightforward sense of name trend in a very specific state especially with a textbox search function. With the instance's prediction provided in the visualization, business owners are also able to benchmark production and marketing estimations, especially when their products and services are names or newborn quantity related, for example, holiday decorations with names pre-printed on.

Although this project is featured with multiple interactions for audiences to apply, a constraint is encountered while working with this visualization - running speed of both tableau software and audiences' device (internet & memory). Frequent selections & page transitioning may encounter unpleasure idle time for audiences. This problem cannot be only be solved with a sacrifice of data size and visualization technical complexity.

For future work and visualization revise, we can improve it in the following four ways:

1. Increase the scope of prediction on both rank and instance data,
2. Increase the range of searchable information and corresponding highlighting interactions,
3. Integrate the function of online search such as the origin of names,
4. Reconstruct the project with JavaScript & D3 (potentially mobile friendly).

APPENDIX

1. https://public.tableau.com/views/APTop252015/APTop252015?:embed=y&:loadOrderID=0&:display_count=yes&:showTabs=y&:showVizHome=no
2. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data
3. https://www.ssa.gov/oact/babynames/background.html
4. https://genhq.com/faq-info-about-generations/
5. https://www.behindthename.com/
6. https://www.youtube.com/watch?v=9XZTN5eNysg&feature=youtu.be
7. https://public.tableau.com/profile/mengwanganalytic#!/vizhome/ISOM675DataVisualizationFinalProject--Team9/Dashboard1?publish=yes