# CS / INFO 3300 Project 2 Final Report

Cody Leung | cpL28
Carolyn Shi | cs925
Yingda Du | yd257

**Work Done by each Team Member**

Each team member contributed to the project equally corresponding to each member's expertise. We all used GitHub as a common source code base. We found that meeting up to work on the project was more effective than working separately. We ran into some trouble with GitHub and merge conflicts, so we decided to use one computer overall for the three of us as one person codes, and the other two give feedback and review.

Since last status update:

- Cody: Cody helped to style the project and make it look nice. He separated the code into different files for a cleaner structure.
- Carolyn: Carolyn helped to check over the code and did testing and quality assurance. She also wrote most of the final report.
- Yingda: Yingda implemented the skeleton code for the project and was able to set a good base for partners to implement corresponding features.

**Description of Data**

The data sets include various fields on baby names in the United States of America. When babies are born, they are given a name. Many babies are given the same name, and our project strives to display the frequency and trend of the most popular baby names. The data includes tuples including state, year, and other fields that we were able to display.

Sources:

1) Data source 1: https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data
2) Data source 2: https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-data-by-state-and-district-of-
3) Data source 3: https://www.kaggle.com/dvasyukova/d/kaggle/us-baby-names/interactive-names-explorer/code

Data filtration**:**

Overall, the data sets we got went all the way back to the 1800's! This was way too much data and we only wanted to focus on the years 1990 to present. Thus, for each data set we manually deleted the data that were from before 1990.

- The namesdata.json comes from the third link directly

- The baby names by state data is from the second link.
    - To process this data for the years from 1990 to 2014, we looped over the all baby names and for each state, we got the 20 most popular boy names and 20 most popular girl names PER year, and stored this in the M and F folders, respectively.
    - We chose to cut of all year before 1990 because the data file was too large to handle and felt that 1990 to 2014 was a long enough period of time to show a trend.
- The data files in data_year folder are from the first link.
    - We choose to get rid of all the files containing data before 2000 due to file size.
    - For the remaining files, we sorted each file by frequency and maintained the name information with baby names that had a frequency larger than 100.

Variables (~common to the data sets):

- Name: name that is given to the baby
- Year: year that baby was born
- State: state that baby was born in
- Gender: gender of baby
- Frequency: how many babies were given the name (in year)
- Popularity: popularity of baby name (=frequency/1000)

Shape files:

- US map

**Description of Mapping from Data to Visual Elements**

For each baby name, we choose to use d3 elements to represent them

- Treemap: rectangle elements
- US map: circles elements
- Bar graph: rectangle elements
- Word Cloud: text elements
- Line graph: path elements

Scales:

- Treemap:
    - Colors → sorted baby names by first letter
    - Rectangle size → frequency of baby name (with that letter)
- US Map:
    - Color → corresponds to first letter of baby name
    - Circle size → frequency of baby name
- Bar graph:
    - X-scale → linear scale of top 10 baby names
    - Y-scale → linear scale of popularity
- Word Cloud:
    - Text size → frequency of baby name (in year)

    - Color → just to distinguish names
- Line graph:
    - X-scale → linear scale of year
    - Y-scale → linear scale of frequency

**Story: What does It Tell Us?**

Our project researched and visualized how the frequency of baby names have fluctuated over time, with analysis starting in 1990. We grabbed the top baby names for each year, and filtered them by state or by year overall. Users can interactively choose 'By State' or 'By Year' to get a closer look at subtleties and are able to analyze and infer trends in baby names over time, pertaining to:

- Geographic regions in the US
- Year
- State
- Many more

Treemap: The treemap is separated by year and within that year, it displays the relative popularity of each baby name by gender. The top baby names per year, by state, are shown and the different colors separate the baby names by first letter. The larger the rectangle, the higher frequency of babies were named that corresponding name.

U.S. Map: The U.S. map in the top right shows the states that have the most number of babies named the selected name. This gives us a good visualization of where in the U.S. the baby name is most dense, and the viewer can possible correlate that to other reasons.

Bar graph: The bar graph in the bottom right shows the top 10 names of the selected year. This helps to group all of the top names together by year, and not by state. The viewer is able to see overall which baby names were most abundant for the selected year.

Word Cloud: The top baby names overall (all of U.S.)per year selected are shown in the graph to the left. A larger font size means a higher number of babies were named that name in the selected year.

Line Graph: The trend line graph of the baby names to the right shows the trend in baby names from 2000-2014 for those baby names in the year selected. This tells us that during certain years, baby names were most frequent while other years they were not. This allows us to see if there is a correlation with a world event during a certain year that caused the baby names to increase (or decrease).

**Surprises**

The bar graph is displaying the most popular names for that year and gender since we realized the treemap was not good about display the absolute popularity of names and you could only see the most popular name of each year most clearly.

**Future Extensions**

We tried to connect the main dataset of name, year, state to famous figures in history or popular culture but we ran into many problems. The main problem being that we could not find a dataset that covered the years we needed or had the qualitative variables that we could relate to the data. Even when suitable data was found, normalizing the data (i.e. dealing with famous people that only have a first name and no last name like “Sting” or “Kesha) was time-consuming and caused errors. The data file also seems to originally have accents included in the names so that in the CSV, it converted those letters to Chinese characters. When transforming the csv file to the json necessary for the relationship dependent visualization that we were pursuing, we could not establish the relationship