Download as PDF

URL to view Results	[Click Here]

Response Summary:

Acquire Worksheet

Goal: Identify appropriate data sources, analyze the data, identify data types, variables, list assumptions about the data
Objectives: Students will identify and acquire data from appropriate data sources
Outcomes: Data for the current visualization challenge

1. Student Information *

First Name	Chia-Hua
Last Name	Lin
Course *(e.g. CGT 270-001)*	CGT 270-LC4
Term *(e.g. F2019)*	F2021

2. Email Address *
lin1424@purdue.edu

3. Visualization Assignment *
- Training Data

Generate

4. Identify appropriate data sources: is the data publicly available? What search methods were used? *

Data source 1	public.tableau.com on Top Baby Names in the US, sorted by state. This data is publicly available, and I searched for it by scrolling down the list of categories it provided and picking one that is interesting to me
Data source 2	538, compiled from Social Security Association on The Most Common Unisex Names in the US. This data is publicly available, and I found it by finding an article on 538 that cited the data and described the processes they used to compile the data to be more specific to unisex names
Data source 3	538, compiled from Social Security Association on the most common full names in the US. This data is publicly available, I found it by looking for an article on 538 whose author had adjusted the data based on the databases for the most popular

	adjusted the data based on the databases for the most popular first names and the most popular last names.

5. Data format: what format is the data in? Structured vs instructed? All text, a combination, multiple sources? Is it primary or secondary data? *

All the datasets I found are structured and in a .csv file format. The data sources are all secondary data, with the primary data being provided by the Social Security Administration. Most of the data is numerical, with some qualitative categories labeled with text.

6. Data types: what types of data are in the data? How are they stored? What is the access to the data (API, JSON, txt, csv, etc.)? What structure holds the data (data base, spreadsheet, etc.)? *

The name that the row is for, the percentage that each gender has with the name, the gap between the percentage of the genders, the number of occurrences of each name per year, the state that the name shows up in

Evaluate

7. Variables: list the data variables? What are the parameters? Give them names. What are the dependent variables and independent variables? *

For the first source:
Data variables are State, Gender, Year, Top Name, Occurrences
The data is organized by year in ascending order, and then sorted state by state, and then separated by gender.
The independent variable is the year and the dependent variables are the top name and the number of occurrences.

8. Audience & Assumptions: list any assumptions you have about the data. Who is your audience? *

I am assuming that the original data had more or less full information on every person in every state in the US, and that the people doing the analyses did not make any mistakes in their methodology. My audience would mostly be people that are curious about names and their popularity, like a couple that are currently naming their baby and are curious to see what names are currently trending.

Generate

9. What real life behavior does the data reflect? Does it show patterns of activity, regularity of events, a timeline, population data, etc? Explain. *

The real life behavior that my data reflects would be the patterns of popularity and the trends when it comes to names throughout the years of history in the US.

11. What are the weaknesses of the data source? Is it likely that the source will be available in the future? Is the data complete? What is the quality of the data? Is it specific to your needs for. the current project? Is the data in the format you need? Are there missing data? Explain. *

The data that was compiled is not quite complete and may be hard to manipulate without the original files for the unadjusted data if I want to do anything myself. Some of the data is slightly outdated because the articles are from six or seven years ago.

12. What information is emphasized? What is the central focus of the data? Explain. *

The central focus of the data would be the popularity of names in the US. The first source focuses on the popularity of names in the US, sorted by gender and state throughout the years, while the second source focuses on specifically the unisex names that are popular in the US. The third data source focuses on the popularity of the combinations of both first names and last names, taking into account the people that have likely already passed away.

13. At what level of granularity is the data provided? Is the data summarized, or do you have access to the raw data? Is the data categorized or is the data in a format that allows you to create your own categories, etc. Explain. *

The sources are all summarized, but for the second source, we have access to the raw data in the repository provided on GitHub.

14. What is the scope of the data? What topics can be covered using the data? Is there a time range/frame? Is the data for a specific area/discipline/demographic etc.? Explain. *

For the first source: The data has the potential to cover topics such as average age of popular baby names per state and a comparison between which names are popular in which states in the US, and the time frame dates from 1910 to 2012.