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	Cristina
Last Name	Pascua
Course ***(e.g. CGT 270-001)***	270-009
Term ***(e.g. F2019)***	F2021

2. Email Address *
cpascua@purdue.edu

3. Visualization Assignment *
- Assignment 1

Generate

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

Data source 1	TopBabyNamesbyState.csv from Tableau Public Resources.
Data source 2	Popular Baby Names by State from Social Security website; found through a Google search.
Data source 3	Most Popular Baby Names in California from a search on data.gov.

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

Data source 1 and 3 are in a structured format. Data source 2 is in an unstructured format. All data sets are primary data sources.

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.)? *

Data source 1 contains character, integer, and string data. It is stored in a csv file and structured in a spreadsheet.
Data source 2 contains string data and is stored as txt files by state. There is no structure to the data.
Data source 3 contains integer and string data. Similarly to the first data source, it is stored in a csv file and structured in a spreadsheet.

Evaluate

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

The dependent variable of the data is the top name that is the result of the count or occurrences of that name. The independent variables are the year, gender, and state. The parameter for the gender would be male or female. The parameters for state would be the abbreviations of each state.

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

The audience is parents. I assume that the data records every baby that was born and named in each state of each year.

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 data reflects population data on a timeline because it totals the occurrences of names by state.

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. *

Data source 2 is not in a structured format, so there many assumptions that can be made about what the data is presenting (variables & parameters).

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

The central focus of the data is the names of the babies. It is the dependent

The central focus of the data is the names of the babies. It is the dependent variable that changes based on the number of babies with that name on

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 data is summarized because it does not show every name of each baby in a year and location. It summarizes the data by giving the total occurrences of a name and the name of the highest total occurrences. Data sources 1 and 2 are categorized by state, but it can be categorized by year or gender as well.

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. *

The data is specifically for people in the United States because it only lists the top baby names by state. Data source 3 has an even smaller scope as it only records data from California. Data source 1 and 2 have a wider time range that goes from 1910 to 2012, but data source 3 only goes from 1960 to 2019.