Response Summary:

Mine Worksheet

Goal: to identify patterns, extreme and subtle features about the data
Objectives: Students will identify basic descriptors for the data, and categorize the data according to the specifications from the Parse Worksheet
Outcomes: Three (3) specific questions to be answered using the data

1. Student Information *

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

2. Email Address *
cpascua@purdue.edu

3. Visualization Assignment *
- Lab Assignment

Analyze

4. Basic Descriptors: for each data component from the Parse Worksheet, identify basic descriptors (basic statistics). Explain *
Gender (String): mining procedure: mode
Year (Integer): mining procedure: median(?)
Name (String): mining procedure: mode, length
Count (Integer): mining procedure: maximum(8,259), minimum(5), average(80).

5. Categorize: consider what is similar and what is different? Categorize the data. Are the variables categorical (normal, ordinal, or rank). Are they quantitative (discrete or continuous)? Show categories. Explain. *
Textual: Name
Nominal: Gender
Interval: Year (discrete)
Ratio: Count (discrete)

6. Temporal: is the data streaming data? How is it stored (all at one time, over several years in years, days, minutes, seconds)? Explain. *

The data is temporal. It shows data from a set time range (1910-2020). It is stored over several years.

7. Range and Distribution: what is the distribution of the data? Few values, small size, evenly spread, sparse or dense? Explain. *

The dataset is very large, containing 394,179 rows of data. The data is not evenly distributed because values of count range from 5 to 8,259 and the average is 80. This means there are more small count values (under 100) than large count values (thousands range).

Evaluate

8. Questions and Assumptions: list at least 3 questions you plan to answer with the data or list the questions if they were provided. Must be complete sentences and end in a question mark. What assumptions are you making? *

Question 1	The data records all names and their occurrences in a single year. Is the minimum count for the name to be added to the list 5?
Question 2	Do more recent years record more names than past years (perhaps because of more names & their popularity)?
Question 3	Do high count names have anything in common (i.e. length, first initial, etc.)?
Assumptions	I assume names with high counts stay near the top of the list per year for a couple of years.