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	Joy (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

Analyze

4. Basic Descriptors: for each data component from the Parse Worksheet, identify basic descriptors (basic statistics). Explain *
State: N/A because it is categorical
Gender: N/A because it is categorical
Year: Minimum is 1910, Maximum is 2012, but this field is also more categories instead of measurements
Top Name: N/A because it is categorical
Occurences: Minimum is 8, Maximum is 10023, Average is 1185.77, Median is 718, and Mode is 237, but I don't think these descriptors can be particularly useful with the data in its current form because these numbers are all separated by occurrences per state.

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. *

State: Nominal, because it is qualitative but without any order between the categories
Gender: Nominal, because it is qualitative but without any order between the categories
Year: Ordinal, because the year is being treated like a category to sort other data into
Top Name: Nominal, because it is qualitative but without any order between the categories
Occurrences: Ratio, discrete because they are numbers, but you can only count the occurrences with integers.

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

This data is temporal, and it only represents names up until 2012. It should be done once per year for the year.

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

The data has a lot of values, so it is large. The data appears to be skewed to the right, because the range is from 8 to 10023, but the average is 1185.77, which is much closer to 8 than 10023.

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	What is the most popular first letter for top names in each US State?
Question 2	Over the course of history, is there more consistency with top names of one gender over the other?
Question 3	Are there any names that only became the most popular in one state?
Assumptions	I am assuming that the data is complete and that the original analysis was correct, since the data is not particularly granular and I don't have access to the raw statistics of each name besides the most popular one in each state.