Response Summary:

1. Student Information *

First Name	David
Last Name	Luo
Major	Data Science
Course *(e.g. CGT 270-001)*	CGT 270-003
Term *(e.g. F2019)*	S2022

2. Email Address *
(University Email Address is required.)
luo354@purdue.edu

3. Visualization Assignment *
- Lab Assignment

Generate

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

Data source 1	The Punxsutawney Phil data set was provided by lab instructor.
Data source 2	The Punxsutawney Phil data set can be found on Kaggle. It is the same data provided by the instructor. The data is managed by the Punxsutawney Groundhog Club and is in the public domain.
Data source 3	The york daily record has groundhog predictions from 1886 to 2022 but does not have temperature data. This dataset will not be used

5. Data format: what format is the data in? Structured vs instructed? All text, a combination, multiple sources? Is it primary or secondary data? *
The data is given in a csv format. The table has 10 features and 131 records. The data was generated by combining data from the Punxsutawney Groundhog Club about weather predictions and average monthly weather was provided by NOAA's National Climatic Data Center.

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 data is stored as a csv and can be accessed by excel or any other csv parser such as pandas. the delimiter of this csv is a single comma. the data includes data types ints, floats, and strings. The variable "Punxsutawney Phil" is stored in a string and is categorical with 3 values. The variable can either be nominal or ordinal depending on interpretation.

Evaluate

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

Year: integer value representing the year.
Punxsutawney Phil: String value representing if Phil's shadow
February Average Temperature: float representing average temperature
February Average Temperature (Northeast): float representing average temperature
February Average Temperature (Midwest): float representing average temperature
February Average Temperature (Pennsylvania): float representing average temperature
March Average Temperature March: float representing average temperature
Average Temperature (Northeast): float representing average temperature
March Average Temperature (Midwest): float representing average temperature
March Average Temperature (Pennsylvania): float representing average temperature
The predictors include Punxsutawney Phil, and February temperatures. The dependent variables are March Temperature. An additional variable may be added called "Correct Prediction" which would be a boolean value showing if the prediction was correct.

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

Assumptions:
Year is independent and does not affect the groundhog prediction.
Records are accurate.
Audience: Groups interested in groundhog day for any reason.

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 is about a groundhog's supposed ability to predict the coming of spring over the course of a century.

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. *

There exists missing data in the first few rows. The data is in the format needed but may need additional features to be useful.

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

The central focus of the data is weather or not the groundhog predicted spring.

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 raw data and is in tidy format. The data is not summarized. The data is presented as a table of observations.

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 scope of the data is February and March temperature in the Northeast and Midwestern united States, and Punxsutawney Phil predictions from 1895- 2016