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	CO2 Emissions by London Borough (2005-2014)/On tableau public sample data, from data.london.gov.uk
Data source 2	Titanic Passenger List/On tableau public sample data, no primary source listed. This data set is a popular one that I have also seen on Kaggle
Data source 3	Tuberculosis Burden by Country/ On tableau public sample data, Sourced from WHO

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 structured as a table of text from a primary source

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.)? *
elsx, stored in a well structured spreadsheet

Evaluate

7. Variables: list the data variables? What are the parameters? Give them names. What are the dependent variables and independent variables? *
1. independent vars: Code(string, identifier), Name(string, name of borough), Year(int), Type(String, category of emissions), Population(int). dependent vars: per capita emissions(float), CO2 emissions(float)

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

Assumptions, the data is accurate and specific. The categories of "Industry and Commercial", "Domestic", and "Transport" are well separated categories. The audience is the general public of the London area.

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. *

This data reflects the approximate average CO2 emissions from specific London boroughs. It also contains population data to compare emissions to. This can provide insights into the behavior of people within the boroughs.

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. *

A weakness of this data source is that it's organizational style is not "tidy". That is, each row is not "one observation" but several observations. This structure makes it difficult to parse with certain group by functions. Other than that the quality of the data is supurb.

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

The central focus of the data is comparing the different boroughs of London's emissions. This is clear from the stacked structure of the data.

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 provided is the raw observational data.

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. *

This data covers a time range for a specific area. The data is from 2005 to 2014 with observations for each year and each borough.