**Part 1. Simple Applications**

1. **Recall Jane Miller’s 5Ws. One of the W’s is missing from each of the following descriptions of the table. Rewrite each sentence to include that information.**
   1. “Germany did the best at the 2002 Winter Olympics, with 35 medals, compared to 34 for the United States, 24 for Norway, and 17 for Canada.”

In the 2002 Winter Olympics held in Salt Lake City Utah (USA) Germany won the most medals - 35 in total, followed by the USA with 34, Norway with 24 and 17 for Canada.

* 1. “Gold, silver, and bronze medals each accounted for about one third of the medal total.”

Gold, silver, and bronze medals each accounted for about one third of the medal totals at the 2002 Winter Olympics held in the USA, as the gold, silver, and bronze medals are awarded for each of the sporting competition.

* 1. “At the 2002 Winter Olympics, the United States won more medals than all other countries, followed by Canada, Germany, and Norway.”

At the 2002 Winter Olympics (held in the USA), the United states won more gold medals than all other countries, followed by Canada, German, and Norway.

1. **For each of the following situations, specify whether you would use prose of text, a table of numbers, or a particular type of chart/figure. Explain why you chose this way to present the data.**
   1. Statistics on five types of air pollutants in the 10 largest US cities for a government report

I would use a bar chart to explain these statistics, one for each of the five types of air pollutants, with the cities ranked from highest to lowest pollutants. I believe this would be a simple way to understand which cities had the best and worst records for each pollutant. It would help make the connection to what pollutants may be more prevalent based on their locations.

* 1. Trends in the value of three stock market indices over one year for a web page

I would use a like chart with three market indices overlayed lines across time (on X axis). This visual would make it easier to compare the three in one chart.

* 1. Notification to other employees in your corporation of a change in shipping fees

This may be as simple as written text. If there are many changes to be included text and a table of numbers may be most effective.

* 1. Distribution of voter preferences for grade-level composition of a new middle school (grades 5–8, grades 6–8, or grades 6–9) for a presentation at a local school board meeting

With three variables, this seems to be most understandable as a stacked bar chart with the three variables adding up to 100% of the voter preferences. A pie chart could also be a simple solution.

* 1. National estimates of the number of uninsured among part-time and full-time workers for an introductory section of an article analyzing effects of employment on insurance coverage in New York City

As this is specific to New York City’s insurance coverage, the national estimates could be as simple as two bar charts for the nation – counts along Y axis and variables part-time and full time along the X. It may be helpful to us percentages as well, as NYC (8 million) is being compared to the whole of the US (330 million).

1. **Read the sentences below. What additional information would someone need in order to answer the associated question?**
   1. “Brand X costs twice as much as Brand Q. Can I afford Brand X?”

The cost of brand Q would allow for a decision.

* 1. “My uncle is 6’6” tall? Will he fit in my new car?

You would need to know where your uncle is intending to sit, and measures of leg room and head room for the particular model.

* 1. “New Diet Limelite has 25% fewer calories than Diet Fizzjuice. How much faster will I lose weight on Diet Limelite?”

You would need additional information about the calorie count of Diet Fizzjuice, as well as the amount of each you intend on consuming over a period of time.

* 1. “It has been above 25 degrees every day. We’re really having a warm month, aren’t we?”

You would need to know the temperature scale of the 25 degrees. Celsius would indeed be warm, Fahrenheit or Kelvin would certainly not be.

1. **Indicate whether each of the following sentences correctly reflects table 4B. If not, rewrite the sentence so that it is correct. Check both the correctness and completeness of these sentences.**

Note: According to [Wikipedia](https://en.wikipedia.org/wiki/Voter_turnout), ” In political science, voter turnout is the participation rate (often defined as those who cast a ballot) of a given election. This is typically the percentage of registered, eligible, or all voting-age people.”

* 1. Between 1964 and 1996, there was a steady decline in voter participation.

Between 1964 and 1992 there was a noted decline in voter participation, with the exception of 1992 with an increase.

* 1. Voter turnout was better in 1996 (63.4%) than in 1964 (61.9%).

Voter turnout was better in 1964 (95.8%) than in 1996 (63.4%).

* 1. The majority of all registered voters participated in the 1964 US presidential election.

This statement is true.

* 1. The best year for voter turnout was 1992, with 104,600 people voting.

The best year for voter turnout was 1964 with 95.8% of registered voters and the best percentage of all voting age population.

* 1. A higher percentage of the voting-age population was registered to vote in 1996 than in 1964.

A higher percentage of the voting-age population was registered to vote in 1964 than in 1996.

1. **Identify terms that need to be defined or restated for a non-technical audience without much knowledge about the topic or statistical method. You don’t need to explain these terms. Just identify them.**
   1. “According to the latest study based on the [VDem Dataset](https://v-dem.net/data/the-v-dem-dataset/), the average Rule of Law score is statistically higher in democratic countries than non-democratic countries based on a t-test (p = 0.01).”

For a non-technical audience these terms would need to be explained: VDem\_Dataset, Rule of Law score, t-test (p=0.01)

b. ” According to the logistic regression results below, we can see a positive correlation between household income and the vote choice for G.W. Bush, with a positive coefficient (log-odd = 0.33). ”

For a non-technical audience these terms would need to be explained: logistic regressions, positive correlation, positive coefficient, log-odd=0.33

**Part 2. Practical Applications**

1. **Suppose you work as a data analyst in the music-producing industry. One day, you get a data report that studies the popularity of different genres of music. The following scatterplot is presented to you. There is no text description for either this table or the data.**
   1. What information can you describe or summarize based on the current graph?

Based on the information available, variable\_1 at approx. 2.6 had the most pop compared to all other variable\_1 measures, the second most was in variable\_1 at approx. 1.0

* 1. Thinking of the principles we learned in the week of visualization customization and Jane Miller’s principles, What additional information (Please describe at least three things (at least one thing that is NOT about graph customization, such as title, color, label, etc.) that you consider adding this table so that it can convey meaningful information.

This graph should have at title that explains what it is about. It also should have axes that support the data. A third piece of information should describe the population and basic information about the data.

(For your reference, this is the original source of the dataset: <https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year>)

1. **Two articles on scientific studies talk about the “risks.” Please read the titles and quotes from these two articles, and answer the following two questions.**
   1. Given the information provided, in which case is there a greater “risk”: the Pancreatic Cancer case or the Diabetes case?

These two pieces of information are lacking information that would allow this question to be answered. While the hazard ratio does imply an almost 2x risk, we do not know how prevalent pancreatic cancer is in the general population. Likewise, the information about soda and diabetes has information about risk in 10 years.

* 1. Think of Jane Miller’s principles. What additional information would you need to know to compare the “risk” in the two cases?

Both of the articles would need to have common information in order to know how the risks are comparable across two different diseases.