Challenge_6: Basic Principles of Data-Driven Story-telling

AUTHOR
Mehak Nargotra

PUBLISHED
April 15, 2024

Make sure you change the author's name in the above YAML header.

Challenge Overview

In this challenge, we will mainly apply the principles we learned from Jane Miller's book in practice. You will review multiple examples of data description and presentation in text, table/number, and charts/figures. Please read the instructions for each part and complete your challenges.

For all the screenshots, images, or tables mentioned in the questions, please see the Challenge_6_Spring24.html file. You don't need to include any of these items in your rendered challenge file.

Part 1. Simple Applications

1. **Recall Jane Miller's Ws mentioned in Chapter #2. One of the W's (Who, What, When, and Where, Why) is missing from each of the following table descriptions. Rewrite each sentence to include that information. (10%)**

   a. "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."

   Ans: At the 2002 Winter Olympics, Germany did the best as it was ranked first in the final medal standings with a total of 35 medals (2 gold, 16 silver, 7 bronze) followed by United States who received a total of 34 medals (10 gold, 13 silver, 11 bronze) followed by Norway with a total of 24 medals (11 gold, 7 silver, 6 bronze) and then finally Canada with a total of 17 medals (6 gold, 3 silver, 8 bronze).

   b. "Gold, silver, and bronze medals each accounted for about one-third of the medal total."

   Ans: At the 2002 Winter Olympics, among the leading four countries — Germany, the United States, Norway, and Canada — the proportion of gold medals to the total number of medals, the proportion of silver medals to the total, and the proportion of bronze medals to the total all approximate one-third.

   c. "At the 2002 Winter Olympics, the United States won more medals than all other countries, followed by Canada, Germany, and Norway."

   Ans: At the 2002 Winter Olympics, the United States secured the highest count of bronze medals among all participating countries, totaling 11 bronze medals leding to a rise in their total medals count, closely followed by Canada with 8 bronze medals, Germany with 7 bronze medals, and Norway with 6 bronze medals.

2. **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. (10%)**

   a. Statistics on five types of air pollutants in the 10 largest US cities for a government report

Ans: I will use a table of numbers to represent this data. Each of the row (total 10) will represent a US city and the columns will represent the statistics of the air pollutants. This method allows a clear and organized display of information. Although we can use a bar graph, it will becomes less practical as if the number of cities and air pollutants increases, it will lead to cluttered information and difficult-to-read visualization. Therefore, a table provides a more efficient and effective means of presenting the data.

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

   Ans: A line graph will be the best for showing trends in the value of three stock market indices over a period of time (1 year) on a web page, as it visually represents changes over time and highlights trends effectively. We can use different colors for the line graphs to represent each of the three stock market indices.

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

   Ans: A prose of text is suitable for notifying employees in a corporation of a change in shipping fees which can notify the empoyees about the previous rate and updated rate. The previous rate and updated rate can be highlighted with different colors. We can also show it using a table with column titled as previous rate and update rate and values in these columns just to make it more presentable.

   d. 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

   Ans: A pie chart will provide an effective means of depicting the distribution of voter preferences regarding grade-level composition. Each segment/part of the pie will correspond to one of the grade-level options (5–8, 6–8, 6–9), with its size indicating the proportion of voters supporting that option.

   e. 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

   Ans: We can use a mix of prose of text, a table, and a bar graph to explain this. First, we'll discuss the topic of insurance coverage, the relevance of employment status and then the importance of understanding the national estimates can be highlighted. We can also present the key points/summary. Then, we'll summarize the main points. In the table, we can show the estimates, with one side for part-time workers and the other for full-time workers. Each row will show how many people don't have insurance. We can use a bar graph to show visually how many full-time and part-time workers don't have insurance to make the data more readable and understandable.

3. **Read the sentences below. What additional information would someone need in order to answer the associated question? (10%)**

   a. “Brand X costs twice as much as Brand Q. Can I afford Brand X?”

   Ans: What is the budget of the person and how much is the cost of brand Q or brand X.

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

   Ans: What are the dimensions of the car and is the length of the car and legroom large enough to accomodate the height of uncle and also the space for his legs.

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

   Ans: Information about the daily calorie intake of the person and the number of calories diet fizz juice has.

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

   Ans: It depends on the temperatures of the days in the previous months. Also is 25 in C or F. The units of temperature should be the same across various months.The average temperature of the month or the same month in previous years would be needed to confirm if it’s warmer than usual. We need to know the average temperature of the month and also the previous months in order to say if the month is warmer than the previous or not. In this case the measuring quantity needs to be same (i.e., temperagture in degree C or F).

4. **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. (10%)**

   Note: According to Wikipedia, ” 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.”

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

   Ans: Sentence is reflected incorrectly. Corrected sentence: Between 1964 and 1996, the percentage of registered voters who actually voted showed a consistent decline until 1988, followed by an increase from 1988 to 1992, and then a decrease again from 1992 to 1996, while the percentage of people who voted out of the eligible voting population decreased steadily from 1964 to 1980, increased from 1980 to 1984, decreased in 1988, increased in 1992, and then finally dropped in 1996.

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

   Ans: Sentence is reflected incorrectly. Corrected sentence: In 1996, 63.4% of registered voters voted, which was higher than the 61.9% of the voting-age population who voted in 1964.

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

Ans: The sentence is correct The highest vote-to-registered-voter percentage was in 1964, reaching 95.8%.

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

Ans: Sentence is reflected incorrectly. Corrected sentence: In 1964, voter turnout was highest, with 95.8% of registered voters participating, and also had the highest turnout in terms of the voting age population, with 61.9%. Also, the year 1992 saw the highest number of votes cast, totaling 104.6 million.

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

Ans: Sentence is reflected incorrectly. Corrected sentence: In 1964, a larger proportion of the voting age population, at 35.38%, was registered to vote compared to 1996, where the proportion dropped to 25.5%.

5. **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 (you don't need to know any of the statistical methods mentioned). Just identify them. (10%)**

   a. "According to the latest study based on the VDem 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)."

   Ans: the average Rule of Law score, statistically higher, t-test (p = 0.01)

   b. " According to the logistic regression results in the screenshot 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). "

   Ans: logistic regression results, 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 scatter plot is presented to you. There is no text description for either this table or the data. (25%)**

   (1) What information can you describe or summarize based on the current graph? (10%)

   Ans: The graph shows many lines made of dots which are parallel to the y-axis. Each line represents a different group/category in the data. The difference in the dot density suggets variation in the number of data points across the different categories. The parallel lines show that for a group, the relationship between the variables stays the same as the y-axis variable changes. However, the dot density shows the difference in the size or significance of these populations.

(2) 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 to this table so that it can convey meaningful information. (15%)

Ans: Addition of some relevant labels would be helpful as the current labels do not provide any information. Some meaningful labels will help improve the understanding and information of the graphs. Currently the graph is difficult to understand as there are insufficient labels.

Secondly, adding some extra information next to the scatter plot, like labels for important or specific data points or special events. This helps to explain what's happening in the data and makes it easier to understand. This information will involve highlighting key observations, outliers, or significant events directly on the plot. This will help get some context for the data, matching with the visualizing customization principles and Jane Miller's principle of context establishment.

We can aslo add some text and interpretive insights to provide insights and context or comments to interpret the scatter plot so as to give a better understanding about it to those who are unfamiliar with the scatter plot or how to read it. It will help to interpret the scatter plot and also understand and get a summary of the graph.

We can also include the regression lines on the scatter plot to represent the pattern or association between the variables. This matches with both visualization best practices and Jane Miller's principle of summarizing the overall pattern.

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

2. **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. (25%)**

(1) Given the information provided, in which case is there a greater "risk": the Pancreatic Cancer case or the Diabetes case?

Ans: A hazard ratio of 1.87 means that people who drink two or more sodas a week are almost twice as likely to get pancreatic cancer compared to those who don't drink sodas. Similarly, those who have one sugary drink daily have an 18% higher chance of getting type 2 diabetes over ten years than those who don't drink sugary drinks. The hazard ratio measures an increased risk, where it shows an 18% rise in diabetes risk over time. For the initial risk for each condition in the general population, it is difficult to compare the magnitude precisely due to the lack of information. But even though pancreatic caner has a higher hazard ratio, understanding the greater risk would benefit if we get additional context and more information about the initial occurrence of each condition.

(2) Think of Jane Miller's principles. What additional information would you need to know to compare the "risk" in the two cases?

Ans:

1. Simple examples: We should use straightforward examples to explain how the studies were done, including their methods and any factors that might affect the results.

2. We should also mention any specific details about how drinking soda affects the risk of getting these diseases. Like any details about the association between drinking sodas and health problems for each case.

3. Proper context: Getting an understanding of how commony pancreatic cancer and type 2 diabetes occur in general population.

4. Use of simple language: Mentioning the details in simple language about all the risks about these health problems will be helpful.

5. We also need to look at different studies to understand what they are saying and what their statistics show. This will help eliminate any kind of bias of even consider the situtaions which might have been missed in the main article or any other reference articles.