Question 1: (2 points) Install the Tableau (free for students), and work through the tutorials specified in class. Download the data sets, Cereals.csv and Cereals.xlsx, use Tableau to plot the 13 histograms like the samples below.

**Answer**: The plot below matches the diagram in the sample shown for the homework with Protein on our y-axis and the Cereal name/ type in the x-axis.

A screenshot of a graph

Description automatically generated

Question 2: Download the dataset cars.xls for Cars in the shared folder. Use Tableau  
Desktop/Laptop to answer the following questions.

a. What are the types of each variable?

**Answer:** Variables: Model (Qualitative), MPG (Quantitative), Cylinders (Quantitative), Displacement (Quantitative), Horsepower (Quantitative), Weight (Quantitative), Acceleration (Quantitative), Year (Quantitative), Origin (Qualitative).   
b. Sketch the graph representing the regression relationship between Weight of Cars and  
MPG, with the color of dots to encode the third variable the Origin of Cars

**Answer:** y-axis represents Weight of vehicles while x-axis represents MPG of vehicles. We have a downward sloping regression line showing that as the weight of cars decrease the miles per gallon of vehicles increase.

A screen shot of a graph

Description automatically generated  
c. Use small multiples to encode the fourth variable Number of Cylinders of the Cars; print  
three similar figures to the samples below

A screen shot of a graph

Description automatically generated

**Answer:** The figure representing the Cylinders of each of the cars are represented below the first chart. It compares the cylinders on the y-axis to the MPG on the x-axis. The trend line or regression line also shows the direct comparison between car origins and the different axis that are represented in each model.

d. Select all four variables above and use the hint on the right pane from Tableau and use  
the suggested diagrams to represent the relationship among the variables and print the  
results.

**Answer:**  We went over this question at the end of class, and at first, I was not correctly grouping the columns by the Cylinders field. However, with the Cylinders field grouped (image below), you can see that the weight is being shown directly in each section split up per Cylinder input: 3,4,5,6,8.

A screen shot of a graph

Description automatically generated  
e. Set up a filter to hide blue and red colors but only show the green color variable in the  
second panel. Next, print the screenshot showing only the blue color variable but hiding  
the red and green colors

A screen shot of a graph

Description automatically generated

A graph with a line

Description automatically generated

Question 3: 2 points) Register Tableau Cloud and use the same dataset above to answer the same questions of problem 2 on the Tableau Cloud. Please summarize the difference in  
functionalities between Tableau Desktop and Tableau Cloud.

**Answer:** The biggest differences between the Cloud version and Desktop version of Tableau are some of the limitations of the features available via the cloud version vs. the desktop version. Specifically, the desktop version has more functional features for advanced users than the cloud version. Users are responsible for the updates regarding the Desktop version, whereas the cloud version updates live. Obviously, in terms of functionality, the Desktop version is limited to the downloaded installation on someone’s local machine vs. the web-based cloud version that be accessed via any network connection. It’s much easier for teams to use the web-based version for sharing and collaboration. This source provides an even further description as to what the fundamental functionality differences are between Tableau Desktop and Cloud.

1. Data source editing is limited in Tableau Server. You can only edit some information. For example, in Tableau release 2020 you still cannot edit joins or relationships within Tableau Server’s web authoring environment. As an alternative, you can use Tableau Prep (Prep is a Tableau ETL tool similar to Alteryx). Tableau Prep is available within the browser in versions 2020.4 and above.
2. Tableau Desktop provides some analytic capabilities not found in Tableau Server’s web authoring environment. What is available depends on the version of Tableau Server. For a complete list of capabilities by version see Tableau’s [Web Authoring and Tableau Desktop Feature Comparison](https://help.tableau.com/current/pro/desktop/en-us/server_desktop_web_edit_differences.htm).
3. As of Tableau version 2020.4, formatting options are more limited in Tableau Server. You cannot format at the item or worksheet level, only at the worksheet level. (Kumar, 2024).

Question 4: (4 points) This question is based on the following videos:  
The Data Detective by Tim Harford: <https://www.youtube.com/watch?v=rmqi09ldF_E>  
Ten Rules for Thinking Differently About Numbers: <https://youtu.be/FkvdtqtL1aM>  
How to Lie With Statistics by Darrell Huff: <https://www.youtube.com/watch?v=Vu7WKcWbg24>   
a. Use your words to describe what are the top three misconceptions about the role of  
Statistics.

**Answer:** I think the first misconception about statistics that is widely misunderstood is that just because you are seeing a graph doesn’t make it “correct” or accurate. Likely, there is inherent bias as to what you are being shown for someone’s job or perspective. Secondly, I think another big one is that correlation is equal to causation (Learn Statistics Easily, 2024). Most people might believe that if you can show them a trend line or regression analysis, in other words, that it will directly show the correlation and provide you with the cause of such trend between those two features. It’s not always accurate as to what is causing a certain metric to move in a particular direction. Typically, there are other features that might have a stronger more foundational impact on why a graph looks a certain way that isn’t being shown. With that in mind, there is a level of domain knowledge that can be extremely insightful when identifying truly causational & correlated features for a particular dataset. Thirdly, in my opinion, Statistics are misunderstood due to the implicit nature of them being mathematically involved. I think most people want to look at an image or metric and believe it. I don’t think they want to put in the legwork to understand how it’s analyzed or measured. Often in the first two situations described, the third can play a large factor in influencing the audience and why certain statistics are being accepted as factual evidence.

b. Give two examples of how you use statistics or computational thinking in your daily life  
to make decisions. Do you estimate your uncertainty in statistical terms?

**Answer:** I think one way that I use statistics in my daily life is thinking about the likelihood or probability of me getting a parking spot on campus when driving to this class for instance. I have experienced the lots being full in past semesters as well as I know the number of incoming students. The number of classes that are in the morning outweigh the number of classes that are in the afternoon/ evening as well. All those factors combined lead me to believe that it’s going to be very hard to acquire a parking spot on campus during the 10am time block. Therefore, I choose to spend the extra money to park at the parking garage as to mitigate my chances of being late to class. Another example, I use the weather forecast to plan my day sometimes because there are higher probabilities on certain days or times that forecast inclement weather. I would argue that I do estimate my uncertainty in statistical terms. Often, in conversation, I will use percentages as a measure of confidence.

c. Can you give two examples of how Statistics are used to mislead the public in media  
and political debates? It is OK to Google it, but you need to provide the link to the source  
information.

**Answer:** One example comes directly from the slides that were shown on the first day of class. It was a representation of a couple different pie charts that don’t add up to 100%. This false metric to statisticians is easy to see. However, to the public eye, it can be confusing or misleading due to them seeing the same portion size on each of the slices of the pie chart to be around the same size, and each portion measures up to be around the same. We know that it isn’t possible for the chart to represent an amount over 100%, so it sticks out as misinformation used to persuade one party vs. the other regarding political campaigns. A second example can be seen below in the screenshot I found when researching this topic. (Statistics How To, 2024) The graph shows how the chart has continued to grow, but due to the size of the chart in comparison to the actual measurements being used, it seems like a larger trend is happening than the reality of the situation. If it started out at 0 or even 10,000,000 instead of 94,000,000, then it would show a more accurate representation of the increase, which is marginal in relation to the size of the metrics we are looking at. In two years, it actually represents an increase of less than 10%.

A screenshot of a computer

Description automatically generated

**Answer:**  
d. Can you offer your opinion about the opinion from the claims of the source information  
above?

For example, you may find the original source is from a liberal-lean Journalist or  
more conservative point of view and which parts are facts and which are simply  
opinions.

**Answer:** This is a great question because inherently there will almost certainly be bias on either side. There is bias in this question because my answer may coincide with what the writer/ information source provided, or it may differ entirely based on beliefs that I have about the economy, politics, etc. To answer the actual question, I believe that the person writing this article is most likely a liberal since they are picking on Fox news quite a bit in this article. Traditionally, in my experience, Fox tends to run more conservative from the conversations that I’ve had about it. That said, I don’t think that these statements are wrong in the article, but it does have a target that’s being used repeatedly. They are using their power as an author to represent Fox in a negative light; however, it is not wrong with what they are providing. I think it would provide a more equitable argument if they provided some misleading statistics from other news sources as well that lean more towards the liberal side like CNN for example. It would provide a less-biased opinion on the topic of statistics being misleading overall.

e. You may suggest what type of data source (you do not need to actually collect them)  
may help to validate your opinion.

**Answer:** The data source that would help validate my opinion would be to find a more conservative author who is bashing on CNN for instance. It would at least provide some equity in the argument that all news broadcasting stations have some inherent bias and tailor make their charts/ visualizations a certain way to persuade their viewers towards their side of the argument. Frankly, I don’t have much of an opinion on what’s a better news station, but I do think that the truth matters. Truth being portrayed a certain way to sway opinions can very well be labeled as misinformation at a certain point of exaggeration. I think it would be best if we could provide an equal number of examples in one article to show that news stations are providing biased information overall.

Question 5:

(2 points) When you read files to your Google Colab, it is critical that you work in the right  
project directory or use the exact right path to open the data files. Before you run the codes of Molin’s book, insert a cell on top and run the following codes to mount your Google drive with Colab, then set your project work directory (pwd) as the folder you put your unzipped file. Make sure to change the ... below as the folder where you extract the zip files.  
Instead of using the data.csv, you will use the linked data sets Cereals.csv. (hint: you need to download the linked file and move it to the same project work directory above).

1. Use head to check the first 8 rows of the data frame.

A screenshot of a computer

Description automatically generated

b) Use loc or iloc to select half of the rows of the data, assign the new data frame as  
half\_df.  
A screenshot of a computer

Description automatically generated

c) Select only two columns of your interest (e.g., protein, fat, or sugar, etc.) and assign  
it as half\_df.

A screenshot of a computer

Description automatically generated

d) Use describe to compare both the original data frame and your chopped data frame.  
A screenshot of a computer

Description automatically generated

e) Sort the data according to the quantity of one selected

A screenshot of a computer

Description automatically generated

**Answer:** The above questions should be completed correctly in this [notebook](https://colab.research.google.com/drive/1ulebv86_mU7yKbR7RWnsJaYuCpr7zjLo?usp=sharing).

Question 6:

A screenshot of a computer program

Description automatically generated

**Answer:**

**A&B)**

**A screenshot of a computer

Description automatically generated**

**C)**

**A screenshot of a computer

Description automatically generated**

**D & E)**

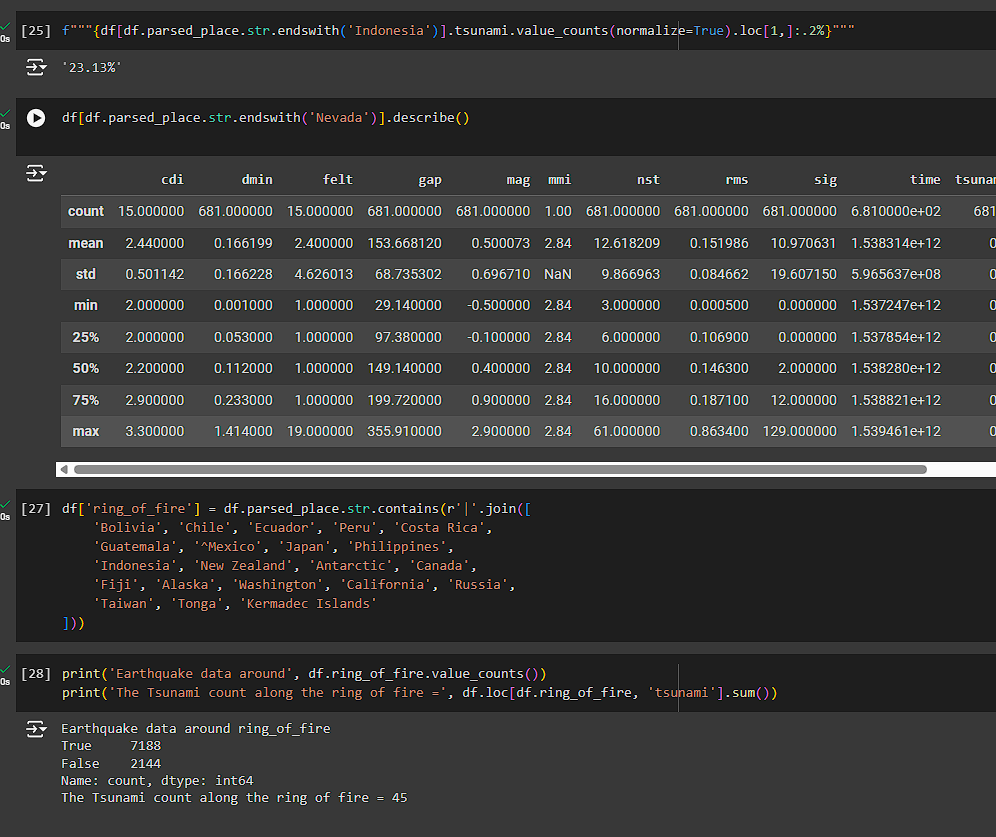
**A screenshot of a computer

Description automatically generated**

Question 7: A text on a page

Description automatically generated

**Answer:**



Question 8:

1. (3 points) Complete the following exercise using the skills you gained above. a. The two datasets Salaries.csv is stored in your shared folder. Download the file into your project work directory (pwd, e.g. chapter 2 of Molin’s subfolder). b. (Use Pandas to read\_csv to read the Salaries file into your working environment and assign to data frame as df (make sure the directory and file names are correctly spelled). c. Check the first 8 record of the Salaries file. d. Find the median and means of Assistant Professors, Associate Professors, and Full Professors. e. Find the median and means of the female and male professors.

**A&B) A screenshot of a computer

Description automatically generated**

**C&D&E)**

**A screenshot of a computer

Description automatically generated**

Question 9:

(2 points) Read the article by Gelman and Unwin, Infovis and Statistical Graphics: Different  
Goals, Different Looks.

Answer the following questions:  
Use a few pairs of sentences or a table to compare the different goals and different views  
between statisticians and information visualization designers. What type of attitudes does the author suggest that we should take in regard to the differences?

**Answer:** The different views that are discussed in this article for the statistician’s side of things really boil down to conveying data accurately and enabling comparisons to understand patterns. It is prioritizing precision and clarity, overall. While the Infovis side is mainly focused on telling a story and making the data interesting for the viewer. It’s mainly focused on leveraging different design principles to create visually appealing graphics. As for the type of attitude the author suggests the reader should take in regarding the differences is that both sides have value and their respective approaches can be complementary of one another; however, there are also some potential shortcomings. Such shortcomings could be described as unappealing to a wider audience for the statistician’s side, and the Infovis side might provide nice visually aesthetic designs that could potentially mislead the viewer. Overall, communication is key, and we want to provide accurate and clear visualizations. There certainly is an art to finding the balance between the two perspectives.

Question 10:

After reading the above article, answer the following questions:

a. What are your thoughts on the pros and cons of Wordle, which main purpose does the  
Wordle graph in Figure 3 serve?  
b. Comparing figure 5 FLORENCE NIGHTINGALE’S COXCOMB and line plot figure 6.  
• Which one do you prefer, and why?  
• Who are the potential readers? What are the purposes of the graphic designers of  
Figures 5 and 6?  
c. Compare Figure 9 THE BABY NAME WIZARD, and histogram Figure 10,  
• Which one do you prefer, and why?  
• Who are the potential readers? What are the purposes of the graphic designers of  
Figures 5 and 6?  
d. Why the author concluded that “the progress in InfoVis important—should be  
important—to statisticians.” Alternatively, what should statistician gain from the  
progress in InfoVis?

**Answer:**

1. The primary purpose of the Wordle graph in Figure 3 is to provide a quick and visually appealing overview of the most frequent words in the text. Like how a Word Cloud works.
2. I absolutely prefer the line plot in figure 6 over figure 5’s graph. The figure 5 graph is hard to read what’s happening, plus each of the two figures are different sizes. I can’t make out or distinguish what it’s wanting to show as easily. The purpose for figure 5 is to make it more visually appealing because it’s unlike most graphs I’ve ever seen; however, the figure 6 plot shows a real representation that is able to be read and interpreted easier than 5. The potential readers are those doing research on this time period and the mortality rates associated with those wars.
3. A very similar answer to part B quite honestly. The figure 9 visualization shows a very appealing graph with a color-coded graph that isn’t very easily readable for accuracy; however, it does provide eye catching information. With that being their goal, they do an exceptional job of it. I think the histograms shown in figure 10 are specific. They are showing growth in comparison to the last letter of boys’ names in two charts separated by 60 years’ time. It is very easily distinguishable what the data represents, and the accuracy of it is easy to read, for the most part. The only thing that I would critique is the y-axis are not the same. It does provide a little misleading information with the y-axis not being the same and the height of the bars. At first glance, someone who isn’t used to reviewing charts may not catch this difference. As such, it could be interpreted as a little misleading. I think the purpose of each of these graphic designers for these figures is simply to convey information in each of their respective perspectives. The potential readers are parents who are looking to name their children in the future most likely if they are planning on having children.
4. In summary, statisticians can gain from the progress in InfoVis by:

Improving their ability to communicate their findings effectively. Discovering new insights in their data. Enhancing their data exploration capabilities. Fostering collaboration with other fields. (Andrew Gelman & Antony Unwin, 2013)

The code in my .ipynb file can be found below. I have also attached a link [here](https://colab.research.google.com/drive/1ulebv86_mU7yKbR7RWnsJaYuCpr7zjLo?usp=sharing) to the Colab notebook for reference.



**References**

Andrew Gelman & Antony Unwin (2013) Infovis and Statistical Graphics: Different Goals, Different Looks, Journal of Computational and Graphical Statistics, 22:1, 2-28, DOI: 10.1080/10618600.2012.761137

Kumar, A. (2024, March 20). *Tableau Server vs. Tableau Desktop: Comparison*. Senturus. <https://senturus.com/blog/tableau-server-vs-desktop-comparison-2/>

Learn Statistics Easily. (2024, February 6). *7 myths about statistics you need to stop believing*. LEARN STATISTICS EASILY. <https://statisticseasily.com/7-myths-about-statistics/#:~:text=Highlights%201%20Correlation%20%E2%89%A0%20causation.%202%20Low%20p-value,matters.%207%20Not%20all%20stats%20are%20universally%20applicable>.

Agrawal, Rishabh. (2024, January 28). *Finding mean, median, mode in python without libraries*. GeeksforGeeks. <https://www.geeksforgeeks.org/finding-mean-median-mode-in-python-without-libraries/>

Statistics How To. (2024, July 20). *Misleading graphs: Real life examples*. <https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/>