due Thursday, October 20 at 6:00pm **Homework 3** DATA 227
Fall 2022

Please complete each problem and submit a PDF with your solutions. If you use code to solve the problems, you should include it in your submission.

Problem 1

The Center for Near Earth Object Studies has collected a database of large, bright objects that have impacted the Earth and were detected by the flashes they made, day or night, in the atmosphere (Fireball and Bolide Data, Center for Near Earth Object Studies). As of Oct 12, 2022 the list has 932 events.

This question asks you to create effective visualizations of the distribution of variables related to these fireballs.

A. Begin by considering the variables included: Peak Brightness Date Time, Latitude, Longitude, Altitude, Velocity, Velocity Components vx, vy, and vz, Total Radiated Energy, and Calculated Total Impact Energy. Consider the nature of each variable–should we expect the variables to be numerical (discrete or continuous)? Categorical (nominal or ordinal)? Write down what you expect without viewing the storage types of the data. (1 pt)

B. Now, read in the data. Rename the variables to `Peak.Brightness.Date.Time`, `Latitude`, `Longitude`, `Altitude`, `Velocity`, `vx`, `vy`, `vz`, `Total.Radiated.Energy`, and `Calculated.Total.Impact.Energy` for convenience. View the storage types of each variable. Do they match what you expected in Part A? If not, convert the storage types. Note: For now, you can ignore converting the date time types. (3 pts)

C. For visualizing distributions, histograms or cumulative distribution plots are suitable. Plot the distribution of longitudes using a histogram, and describe the center, shape, and spread of the distribution. Indicate the presence of any outliers. Is this distribution consistent with what you might expect? (2 pts)

D. Plot the distribution of Total Impact Energies using a histogram, and describe the center, shape, and spread of the distribution. Indicate the presence of any outliers. Is this distribution consistent with what you might expect? (2 pts)

E. For right-skewed graphs, a log-scale is often appropriate for visualizing differences in the data. Plot the distribution of transformed Total Impact Energies using a histogram, and describe the center, shape, and spread of the distribution. Indicate the presence of any outliers. Do your findings change from Part D? (2 pts)

Problem 2

For this problem, you will create a static visualization from the CDC birth database.

The CDC, for fifty years, has been compiling data about each birth in the United States, making most of the data from these standard forms available. This data consists of about 50 fields, almost all of which are categorical data referring to geography, presence of absence of risk factors, method of delivery, duration of gestation, age of parents, date of birth, etc. This will require some amount of effort to wrangle.

There are 2-4 million births per year, and, if you desire to examine mortality, around 20,000 infant deaths. The CDC started obfuscating the exact day of birth at some point to maintain privacy. Examine the data, find a fact that is contained within the data, and design a visualization that communicates that fact.

Include a figure caption and just one paragraph discussing your findings and the graphical design. Attach any code you used to produce the visualization. The figure caption should describe the origin of the dataset.

You do not have to use any specific tools to produce the visualization (you could even draw it by hand) but you need to find something interesting and display it effectively.