Data 227 - Autumn 2023 HW 3

Data Visualization and Communication - Trimble

Due: Friday October 20, 2023 11:59pm

Fireball data

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 13, 2023 the list has 964 events.

   `https://cneos.jpl.nasa.gov/fireballs/`

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

   a For all of the columns in the database, 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. If any of the storage types need to be converted, convert them. Convert the date and time to a time data type. (1 point)

   c Is there anything unusual about the distribution of date and times of fireballs in the dataset? (1 pt)

   d For visualizing distributions, histograms or cumulative distribution plots are suitable. Visualize 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? (1 pts)

   e Visualize the distribution of the latitudes. Is this distribution consistent with what you might expect? (1 pts)

   f Plot a histogram of Total Impact Energies, and describe the center, shape, and spread of the distribution. Indicate the presence of any outliers. Do some of the outliers have their own wikipedia pages? Is this distribution consistent with what you might expect? (1 pts)

   g 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 f? (2 pts)

   h Count the number of fireballs in each degree of longitude. Make a plot that compares the histogram-of-the-histogram (number of degrees of longitude with 0, 1, 2, etc. events) with the Poisson distribution with the same number of average events. This will be a Poisson distribution with a mean of 964 / 180. Comment on their agreement / disagreement. (2pts)

Birth data

2. For this problem, you will create a static visualization from the CDC birth database. Data for births and infant deaths is available from the CDC Vital Statistics Online Portal

   `https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm` .

   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 myriad relationships in this dataset, and hundreds of papers have been written leveraging the birth census for making inferences about maternal and infant health.

There are 2-4 million births and about 20,000 infant deaths per year. Different years have different sets of fields. Examine the data, find a fact that is contained within the data, and design a visualization that communicates that fact.(6pts)

Include a figure caption and just one paragraph discussing your findings and the graphical design. (1pt) Attach any code you used to produce the visualization. The figure caption should describe the origin of the dataset.(1pt)

Please put your question 2 visualization in a PDF of its own, with its caption, separate from the code. (2 points)