CMPT 826 - Data and Process Modeling and Analytics

moodle.cs.usask.ca/mod/assign/view.php

In this assignment you will examine the processing of operationalizing a behaviour, and the impact of different types of operationalizations. This assignment uses the location and battery tables of the SHED10 dataset.

STEP 1: Preparation (20 Marks)

Using code from or similar to the previous assignment, filter out participants with less than 50% of possible battery data, and GPS points outside of Saskatoon, using the same bounding box as the previous assignment (52.058367, -106.7649138128), (52.214608, -106.52225318), and a reported accuracy of 100m or better. Aggregate location to the duty cycle level. Grid space at 100 m, and express all locations as positions on this grid.

STEP 2: Trip Definition (10 Marks)

Operationalize three different trips, using the N-times definition with N equal to 1, 3 and 5 duty cycles. For each participant determine the number of trips, the length (in grid cells) of the trips, and the time (in duty cycles) of each trip. Provide commented code as part of the hand-in. This can most easily be accomplished by creating and parsing visit strings, either explicitly, or on the fly.

STEP 3: Presentation (20 Marks)

Create the distributions for trip number (over participants), trip length (over participant-trips), and trip duration (over participant-trips) and plot these distributions for each N as multiple curves on the same plot (3 plots with 3 curves). Change the axis to show what you consider the best representation of these distributions (e.g. linear for Gaussian, log-linear for exponential, log-log for power law). Explain how you concluded which representation best fit the data. Plot a heat map over all participants which includes only trips, and another which includes only non-trips for each N, for a total of 6 maps.

STEP 4: Interpretation (30 Marks)
In a series of paragraphs, address the following questions.

4.1 What is the impact of changing N on the distribution of trip number, length and duration? What trips are being captured and which are being ignored? Given an example of a research question where the differences would be important, and an example of a question where they would be unimportant.

4.2 What distinguishing features did you see in the heatmaps? Where there points included in either map (trip, not trip) at any N that seemed out of place? How would you change the operationalization to eliminate these points?

Quality of writing and presentation (10 Marks)
Quality of code and code comments (10 Marks)