CMU MSP 36602 Pipeline Example March 6, 2019

You are a data scientist for MallWart, a large retailer. The advertising department wants you to create a "one liner" that will compare rates of purchases of any particular for above vs. below a specified age during any specified date range. We need a mean difference with a 95% CI. Previous experience suggests that we need a random intercept for each zip code. The data is restricted to customers who are part of the company reward program.

When a customer joins the rewards program a record is created in the "customer.tab" file. The columns (in order) are customer id, birthday, zip code, and gender. Missing values are coded "NA".

When a customer makes a purchase, one record is created in the "sales.tab" file with columns sale id, sale date, customer id, and store id. In addition one record is added to the "item.tab" file for each item in the sale, with columns sale id and product id.

The three files are much larger than any one computer disk.

1) What overall strategy can you use to accomplish your task?
2) What is the input from the user? How could you verify it?
3) What specific strategy would you use for extracting the data and computing the two proportions for each zip code?
4) How will you pass the data to the next stage?
5) How will you analyze the extracted data?
6) How can you assure that the user is specifically informed about any problems encountered?

Using the sample data files (hget data sales.tab, etc.) and the provided template, complete steps 3 and 4.

Write code to complete step 5.

Write a simple version of step 1. Add in steps 2 and 6.