CMPT 898 HOME EXERCISE 6

INITIAL EXPOSURES TO SPARK FILE DATA SOURCE AND BREEZE-VIZ

This exercise will provide an experience of manipulating file data in Spark, and in creating figures using this data in Breeze-Viz.

Please open the CSV file that we opened in class (“CCHS.csv”)

1. Please read the CSV into a dataset, per the schema distributed to you.
2. Please select from that dataset the columns for Age Category (DHHEGAGE), Sex (DHHE_SEX), Body Mass Index [BMI, a measure of weight given height] (HWTEGBMI) and daily estimated energy expenditure (PACEDEE).
3. Please create a case class for this dataset (judging types based on the above schema)
4. Create a dataset (with statically typed rows) that use the above case class.
5. Using a breeze-Viz histogram (created by the call *hist* and taking a DenseVector of values and a count of bins), please provide a histogram of both
   a) BMI (via HWTEGBMI).
   b) Estimated energy expenditure (via PACEDEE).
6. Please show a cross-tabulation (“crosstab”) table that gives, for each increment of 4 in BMI, and each incrmeent of the mean of 0.4 in PACEDEE, the count of respondents in the dataset who associated with those levels of BMI and PACEDEE (please note that what is being asked for is a completely standard crosstab functionality; what you will be setting is the appropriate “bin” sizes for BMI and PACEDEE. Please label each axis and give a title according.
7. Using a breeze-Viz scatterplot (similar to that created in class), please provide a scatterplot of datapoints, each representing a single individual. For a given point (person) on the scatterplot, the X location of the point should be given by the associated person’s PACEDEE, and the Y location of that point should be given the associated person’s HWTEGBMI. As above, please label each axis and give an appropriate title according.

Due in class November 1, 2016