SimilarGroup Data Quality Test

November 24, 2016

In this test you will need to analyze different sources of data to determine their quality. The attached file is a TSV formatted database containing the following fields.

1. Source ID
2. Site
3. Percent Unique
4. Number of Visits

There are 4 source IDs in the database numbered 0-3. Source 0 provides the actual traffic results for each site and will be referred to as the "Learning Set". Sources 1-3 are external sources of data based on different samples from populations of Internet users, and will be referred to as "External Sources".

The column headings for the TSV file are as follows. "Site" is the main domain of the sites we have in the Learning Set. "Percent Unique" is an internet traffic metric representing the percent of Internet users entering a given site. "Number of Visits" is an Internet traffic metric representing the number of visits to each site. Notice that a single user can visits a site more than once. Each of these metrics are calculated per month.

You can use any relevant Python package to analyze the relationship between the External Sources and the Learning Set in order to provide a quantitative recommendation for quality of the each of the External Sources. Given that the Learning Set represents actual measured Internet traffic, you will basing the quality of the External Sources on their agreement or lack thereof to the Learning Set. Try to think of several different metrics for determining source quality and choose the most relevant, comparing and contrasting your various approaches.

Please provide your Python script and Power Point presentation summarizing your approach to this data analysis exercise, your results, and your final conclusions.