Analysis of Subgoal Data in Computer Science Principles - Data Cleaning

Hari Ramilison, Graduate Student, Management Information System

Faculty Mentor: Briana Morrison, Computer Science

A study to evaluate the effectiveness of students learning to solve programming problems with subgoal labels was conducted within the Code.org Computer Science Principles online course. Teachers opted into the study which meant that their students were presented with an alternative version of one unit within the curriculum. That data for these students and a comparison group were pulled from the Code.org database and provided to researchers for analysis. The first step in this process was to clean the data and combine it into a usable format, two-dimensional tables, for analysis. Three steps have been applied during the data cleaning procedure: data extraction, data transformation, and removal of incomplete data. This presentation reports on automated processing of all three steps along with the final results yielding approximately 2000 rows for students in the study and approximately 10000 rows for students in the comparison group.

The first step is (1) data extraction. It consists of retrieving questions that are designed for two groups of participants and their related answers. These data are stored in multiple tables. The second step is (2) data transformation. It converts the structure of the data into unified two-dimensional (or flat) tables. This process is performed by stored procedures. The flat tables are created based on the types of the questions asked to the participants. Questions are classified into either Multiple-Choice Assessments, Free Response Assessments or Subgoal Comments. The third step consists of (3) removing unnecessary data and inconsistencies. Unnecessary data include participants who did not respond to a certain number of questions. Various rules are created to perform the deletion. After completing these three main steps, a data validation concludes the data cleaning process in order to check the accuracy of the generated records.