Moe Pwint Phyu
Mount Holyoke College
Mentors: Emery Boose & Barbara Lerner
Group Project: Data Provenance in R

Accessible data provenance with debugging feature in R

Data analysis, a fundamental activity of nearly every scientific experiment, is critical to support hypotheses and report findings. In preparation for data analysis, raw data are transformed and manipulated to find useful information that will answer the scientist's questions. Therefore, recording the transformation of raw data (also known as data provenance) is important not only for the legitimacy of experiments but also for reproducibility by other scientists. To encourage scientists to record data provenance, this project previously produced RDataTracker, an R package that captures data provenance and DDG Explorer, a Java program that uses RDataTracker's provenance documentation, to visualize and query the data manipulation process. Before this summer, the two tools needed to be employed separately: after running RDataTracker, scientists needed to run the DDG Explorer program and select the correct data provenance file to see the visualization. To reduce the steps, I integrated DDG Explorer into the RDataTracker package, allowing scientists to see the visualization from within their R environment. In addition to viewing data provenance when the script is completed, I implemented a feature to incrementally draw as the script executes so that scientists can debug their R scripts better by visualizing what was done at each step. With these new features added to the software, we hope that data provenance will be a more accessible and useful resource for scientists.