Northeastern University

**Course:** DA5020

**Assignment:** Term Project

**Total Points:** 100

**Date Due:** Posted on Blackboard

## Learning Objectives

In this assignment, you will learn how to:

- combine the technologies, concepts, and strategies learned in this course
- apply your new knowledge

## Project Goal

The term project is intended to allow you to revisit what you have studied this semester and deepen your knowledge of the material based on your interests. The project must be an R programming project that displays your competence in using the processes of collecting, storing and retrieving data to solve a specific problem. The project should be focused on these processes and NOT the analysis of the data set.

## Project Proposal

You must submit a proposal of what you plan to do. The proposal must be posted on the Blackboard Discussion Forum under the "Term Project" discussion thread. You must also provide feedback to **at least five** of the submitted proposals. The feedback must be taken into account when developing the final submission. The Blackboard post should be 1 - 3 paragraphs clearly outlining what you plan to do, how you plan to approach it, and what your final deliverable will be. As you are making progress, update your thread on Blackboard and keep contributing to the other threads.

## Project Ideas

Here are some ideas for a term project; you are not limited to these ideas and may propose anything else that would be about 30-60 hours of work.

- Collect visual data from medical diagnostic tools (MRI, CT Scans, ultrasound, etc.) store image files in a database, and retrieve the information in a manner useful for physician/patient.
- Retrieve data from social media web APIs, e.g. Twitter, Facebook, Instagram, etc., clean the data, and create an alternative data schema for a particular database management system.
- Collect publicly available economic data via web APIs offered by Yahoo Finance, US Bureau of Economic Analysis, and Quandl Resource Hub, etc., filter it for a particular industry/task, store in a database, and retrieve.
- Collect publicly available data produced by manufacturing and production systems with the aim of increasing efficiency for a certain manufacturing process. Try Amazon Public Datasets, DBpedia, or World Bank for dataset sources.
- Collect bioinformatics datasets from projects and databases such as the ENCODE consortium, NCBI, and DAVID that can be used to investigate a biological problem, parse the data, store it in an appropriate database, and demonstrate retrieval.

A well-designed project is one that takes a data set from some source(s), cleans it, stores it in a database, and retrieves the data. Your project must include an implementation of a database (SQL or NoSQL) .

## Deliverable

You must submit a detailed and well-formatted report in PDF format that is free of grammatical errors, contains proper references, and shows the work you did. The report should contain:

- A description of the problem that is driving the project (comprehensible by an educated lay person).
- A rationale for selecting the data collected used to investigate the problem.
- Justification for the processes implemented, and the technical choices you made for your project.
- A description of any issues you ran into and how you resolved them.
- Input and output results from an example session of your R code and insightful screenshots of the project running.
- Insights on what you learned and potential future work.

While this report does not need to be at publication level standards, it should be written as an academic report. Please refrain from using overly colloquial language.

A good report should be equivalent to 5 double-spaced pages of text, not including references. Include screen shots, graphs, and charts that enhance the report. Please include a title page with your name, course number, and the semester.