DSCI 325: Management of Structured Data

Instructor: Chris Malone
Office: Gildemeister 137
Email: cmalone@winona.edu
Website: www.statsclass.org

Text: There is no text required for this course. The following references may be useful.

- Modern Data Science with R by Baumer, B., Kaplan, D., Horton, N.
  Link: https://mdsr-book.github.io/mdsr2e/index.html
- Python Data Science Handbook by VanderPlas, J.
  Link: https://jakevdp.github.io/PythonDataScienceHandbook/

Course Description: This course will give students an overview of the issues related to the management of data. Topics to be covered in this course include: data warehousing, data integrity and quality, data cleansing, basic programming concepts, the construction of simple algorithms, and appropriate descriptive and graphical summaries of data. Commonly used software packages for the analysis and management of data will be emphasized.

Prerequisites: DSC 210 Data Science

Learning Outcomes: A student who has successfully completed this course will be able to:

- Demonstrate an understanding of the wide variety of issues related to the management of data in our data-centric world.
- Construct, manipulate, and manage data using software.
- Apply methods to summarize data for a wide variety of data structures.
- Apply basic programming concepts for the design and construction of algorithms necessary for data management.

Assessments:

Tasks (about 50% - 66% of grade)
You will be asked to complete several small tasks to demonstrate your mastery of concepts discussed during a particular class period. In addition, students will be required to complete more substantial tasks which will involve the use of several concepts / methods. Work submitted past the deadline will be assessed a 10% penalty for being late and will not be accepted solutions are posted.

Exams / Projects (About 33% - 50% of grade)
There will be a combination of exams and/or projects in this course. These exams / projects will be substantial in nature and will be worth between 33% and 50% of your grade. These assessments will test your ability to adapt your data management skills to new situations and/or extend your knowledge of methods presented in class. Exams will likely be a combination of in-class and out-of-class.

Grades:

Your grade will be determined by your performance on tasks and exams / projects. The tasks will be worth slightly more than the exams / projects. I do no weighting, so a point is worth a point in this class.

Your final grade will be determined using the following percentages.

Your Percentage	Grade
90% of greater	A
80% - 89.9%	B
70% – 79.9%	C
60% - 69.9%	D
Less than 60%	F

Extras:

- I encourage you to use a 3-ring binder for this class because class material will be a combination of note taking, handouts, and lots of computer output.
- Attendance in mandatory. If you miss class, it is your responsibility to get the material and get yourself caught up.
- If necessary, I reserve the right to make policy changes for this course as the semester progresses.

Topics Covered [Existing Course Outline]:

I. Introduction to Data Management (1.5 weeks)
   a. Basic Structure of Data
   b. Data Storage and Warehousing
   c. Integrity and Quality of Data
   d. Data Management Issues in Business, Healthcare, and Government
   e. Rules and Regulations for Data Collection and Management
      i. Institutional Review Board
      ii. HIPPA
      iii. Data Management Plans

II. Management of Data in Excel (1.5 weeks)
   Specific content should relate to Microsoft's Advanced Excel Certification Exam.
   a. Importing and Exporting Data
   b. Sorting and Filtering Data
   c. Using Functions to Manage and Manipulate Data
      i. Functions for Numerical Variables
      ii. Functions for String Variables
      iii. Other Functions
   d. PivotTables
      i. Creating and Managing PivotTables
      ii. Working with PivotTable options
      iii. Creating Visual Displays with PivotTables
   e. Creating Tables and Graphs for Report Writing
   f. Creating Macros for Repetitive Tasks
   g. Other potential topics: Conditional Formatting, Security Issues, Working with Auxiliary Data Sources

III. Management of Data in SAS (8 weeks)
   Specific content should relate to SAS Corporation's Certification Exams for Base, Advanced, and Clinical Trials Programmer.
   a. Data Warehousing and Data Structures
      i. Creating SAS Libraries
      ii. Creating temporary and permanent SAS datasets
      iii. Using PROC IMPORT to Retrieve Data from Other Sources
      iv. Using PROC EXPORT to Save Data to Other Sources
   b. Importing Raw Data
      i. Using the INFILE / INPUT Statements
      ii. Advanced features of INFILE / INPUT Statements
      iii. FORMAT and INFORMAT Statements
      iv. Using PROC CONTENTS
   c. Processing Data using the DATA STEP
      i. Creating New Variables
      ii. Modify Existing Variables
      iii. Using SAS Functions to Manipulate Numerical Variables
      iv. Using SAS Functions to Manipulate Character Variables
      v. Using the RETAIN Statement
      vi. Using ARRAYS in SAS
      vii. Using Basic Programming Concepts to Manipulate Data
         i. IF/THEN Statement

        - ii. IF/THEN/ELSE Statement
        - iii. DO Statement
    - viii. Data Cleansing Procedures
- d. Dataset Processing
    - i. Modify an Existing Dataset
    - ii. Obtaining Subsets of a Dataset
    - iii. Sorting a Dataset
    - iv. Merging Two or More Datasets
- e. Report Processing
    - i. Using PROC PRINT and PROC REPORT
    - ii. Generate a Custom Report within the DATA STEP
    - iii. Output Delivery System (ODS) in SAS
    - iv. Using SAS Procedures to obtain basic descriptive summaries
- f. SQL Procedure in SAS
    - i. Retrieve Data using SQL Procedure
    - ii. Generate Reports using SQL Procedure
    - iii. Compare and Contrast the SQL Procedure to programming with the DATA STEP
- g. Macros in SAS
    - i. Create User-defined Macros within the SAS Macro Language
    - ii. Using Macros to Enhance and Automate Programs
    - iii. Procedures for Debugging Macros
- h. Handling Errors in SAS
    - i. Procedures to Verify the Integrity and Quality of Data
    - ii. Recognize and Correct Syntax Errors in Programs
    - iii. Identify and Resolve Programming Logic Errors
- i. Other potential topics: PROC IML, Creating Graphs in SAS, Generating Random Variables, and Constructing Simulations Studies in SAS

IV. Management of Data in R (4 weeks)

- a. Introduction to R
- b. Working with Data in R
    - i. Manipulation of Vectors
    - ii. Manipulation of Arrays, Matrices, and Data Frames
    - iii. Importing and Exporting Data in R
    - iv. Using Basic Programming Concepts in R
        - i. IF Statement
        - ii. FOR Statement
        - iii. REPEAT and WHILE Statements
    - v. Using the apply() function
- c. Graphical Procedures in R
    - i. Overview of Available Procedures
    - ii. High-level Plotting Functions
        - i. Examples
        - ii. Optional Arguments
    - iii. Low-Level Plotting Commands
    - iv. Interacting with Graphs
    - v. Using the par() function
    - vi. Using the LATTICE Package

d. User-Defined Functions in R
    i. Creating Functions in R
    ii. Specifying Inputs for Functions
    iii. Specifying Outputs for Functions
    iv. Writing Efficient Functions
e. Other potential topics: Using Packages in R, Constructing Simulation Studies in R, Obtaining Descriptive Summaries in R, Creating Tables and Graphs for Report Writing