Great Bay Community College
Course Content Outline

Department:	**Math**	**Date: September 28, 2014**	
Program:	**ALL**	**Prepared by:**	**Mary Rudis**
Course Number:	**DATA210**	**Course Title:**	**Elements of Data Science**
Theory Hours:	**2**	**Lab Hours: 2**	
Credits:	**3**		

Prerequisites: CIS110 (or higher) with a grade of C or better or by department approval. It is recommended to complete CIS112 prior to this course but not required.

Catalog Description:
This course is the foundation for introducing students to key topics in data science, including data acquisition/preparation and exploratory data analysis. Major topics include an introduction to the R programming language and RStudio integrated development environment, working with modern data formats (e.g. XML, CSV, JSON, XLS, XHTML), data import/export (e.g. files, APIs – application programming interfaces – , web sites, databases), finding data to augment analyses, and exploratory data analysis & visualization.

Desired Student Competencies
The student will be able to:

1. Use the core data structures of R including vectors, lists and data frames
2. Write and organize analysis scripts that utilize the functional programming nature of R and vectorization model unique to R
3. Work with all modern data formats, including XML, CSV, JSON, XLS (Excel), XHTML (web pages), and understand how to appropriately transform this data for use in structured analysis projects
4. Use web APIs, including Census & Bureau of Labor Statistics data sets and database APIs to import data for structured analysis projects
5. Explain the difference between discrete & continuous data and the basic operations that can be applied to each type
6. Understand the concept of "tidy" data and demonstrate how to transform data for use in structured analysis projects
7. Apply the "split-apply-combine" analysis pipeline paradigm to any data analysis problem
8. Search for authoritative data sets to help augment their analysis projects
9. Visualize data for use in exploratory data analysis as a pre-cursor to statistical analysis of data sets
10. Effectively communicate preliminary results of data analysis projects

Required Texts:

- Discovering Knowledge in Data, Wiley, 2014
- An Introduction to Data Science, Stanton, Syracuse University (free/online)
- The Split-Apply-Combine Strategy for Data Analysis, Journal of Statistical Software April 2011, Volume 40, Issue 1. (free/online)

Outline of Topics To Be Covered:

1. Programming in the R Programming Language and using the RStudio IDE
2. Organizing Data Analysis Projects
3. Understanding the data analysis pipeline
4. Cleaning data sets and working with "tidy" data
5. Working with Discrete & Continuous Data
6. Working with modern data formats (e.g. CSV, XML, etc)
7. Accessing data via modern APIs
8. Accessing data in databases
9. Performing exploratory data analysis
10. Visualizing data
11. Communicating analysis results

Other topics may be added as time allows

Objective 1
Description: Demonstrate basic proficiency in R
Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures: Students will write programs that demonstrate their knowledge of core R data structures, control-flow constructs and packages. Students will also demonstrate their ability to locate and install packages that can assist in their analysis projects.

Objective 2
Description: Demonstrate proficiency with diverse data sets
Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures*:* Students will distinguish between "tidy" and "messy" data sets and perform appropriate transformations of data sets to facilitate use in structured data analyses.

Objective 3
Description: Demonstrate proficiency in all modern data formats
Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures*:* Students will perform and report on exploratory data analysis and visualization using all modern data formats.

Objective 4
Description: Demonstrate proficiency with modern APIs & databases

Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures*:* Students will demonstrate proficiency in accessing core/common APIs for data including U.S. Census data, Bureau of Labor Statistics data, Health & Human Services and learn how to locate and use many web-accessible APIs.

Objective 5
Description: Using the "web" as a data source
Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures*:* Students will demonstrate proficiency in "scraping" and transforming content from web pages where no API is available.

Objective 5
Description: Demonstrate proficiency with exploratory data analysis (EDA)
Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures*:* Students will demonstrate proficiency in describing the structure of data sets and performing/communicating basic numerical and statistical analyses on data sets

Objective 6
Description: Demonstrate proficiency with data visualization
Assessment: Quizzes, tests, projects, and cumulative exam to 60% or better proficiency.
Outcome Measures*:* Students will demonstrate proficiency in using appropriate data visualization techniques to communicate results of EDA and understand the cognitive science behind visual communication.