Objectives

Apply tools and methods of data analytics presented in this course. Select a suitable dataset for analysis; demonstrate analysis using R, Python, and SQL to produce statistical summaries and visualizations that support conclusions about the meaning and value of the dataset.

General Instructions

The Big Data Analytics Project is a course-long individual effort consisting of two separate deliverables. You must submit Deliverable 1 by the end of week 5, and Deliverable 2 by the end of week 7. You are strongly encouraged to work ahead on these deliverables. You may submit them early, if you would like, but they will not be assessed until after their respective due dates.

This project consists of two deliverables:

1) Dataset selection and description, and
2) Data analysis and interpretation

Specific Deliverables and Scoring

Read the sections below for the specific requirements of each deliverable. **Use these as templates for your reports.**

Scoring:

- Deliverable 1 – Dataset Selection & description: 20 points
- Deliverable 2 – Data Analysis & interpretation Report: 20 points

Project Analysis and Interpretation Templates

Deliverable 1 – Dataset Selection Template *(report is due end of week 5)*

Select a reasonably large and complex dataset from one of the following domains:

- Health
- Social Media
- Climate
- Politics
- Business
- Sports

Dataset description:

- Briefly *describe* the dataset: size (required storage), metadata (data items' meanings and types).

In your description, include the following:

1) **Who** (company, agency, organization) collected the data?
   a) Who they are, what do they do?
   b) What is their role/purpose?

2) Need
   a) *Why* did they collect this data?
   b) Why is this a big data problem?
   c) Describe any privacy, quality, ethical, or other issues with this dataset?

3) What potential *questions* could be answered by studying this data?
   a) List some *specific questions*, and *plan* to *answer them* in your analysis

4) What software and hardware resources will you need to study this data?

Deliverable 1 of the project is worth 10% of your overall project assignment grade.
*It is due by **Sunday, 11:59 pm ET**, at the end of week 5.*

Deliverable 2 – Data Analysis Template *(report is due end of week 7)*

Explore and *present* analysis of the dataset using relevant tools discussed in the course (R, SQL, Python, Tableau, etc.):

1) Prepare relevant *descriptive statistics* and *visualizations* for *selected* data items (i.e., you *don't need* to analyze *all* the items in the dataset if there is a very large number of them):
   a) *Prepare and include at least one of each*:
      i) Scatterplot, boxplot, correlation analysis, regression analysis, hypothesis test
      ii) Include an *SQL schema* for the data, and demonstrate several basic SQL-based queries of the dataset
   b) Graphics and tables must follow *good visualization practices* discussed in the course
      i) See Chapter 6 of the textbook for guidance

2) *Interpret* the results; what *conclusions* can be supported?
   a) This should reflect answers to the *specific questions* specified above
   b) Describe the *value* obtained from the study
   c) Include *explanations* of any *technical terms* relevant to the project domain

References

Provide *appropriate citations and references as described in the AIT-580 FAQ.*
Be sure to include a citation for the dataset; see http://infoguides.gmu.edu/citingdata.
Your report will be checked with *SafeAssign*

Project deliverable 2 is worth 25% of your overall project assignment grade.
*It is due by **Sunday, 11:59 pm ET**, at the end of week 7.*