MIS2502: Exam 2 Study Guide (Spring 2024)

Instructor: Jeremy Shafer

The exam will be a combination of multiple-choice and short-answer questions. It is a closed-book, closed-notes exam. We will take it on paper in our regular classroom.

The exam has two parts. Part A will be a collection of multiple choice questions worth 4 points each. Answers to Part A will be recorded on a SCANTRON. Only answers recorded on the SCANTRON will count towards Part A. There will be roughly 20 questions in Part A.

Part B will be a collection of short-answer questions worht 4 points each. There will be roughly 5 questions on Part B.

In part B, students may be asked to write one or two short sentences. Or, write a short portion of Python code (such as a loop, if-statement, or Pandas statement used to retrieve data.) Student might be asked to evalauate a diagram, state a hypothesis, or interpret a t-test.

The following is a list of items you should review to prepare for the exam. Note that *not every item on this list will be on the exam, and there may be items on the exam that are not on this list.*

Python Basics

- The role of packages/libraries in Python; how to import a package in your script.
- Generate and explain the basic syntax for Python, for example:
    - Create a variable and assign a value to it.
    - Manipulate variables in calculations.
    - Use indexing to get a character or a substring from a string.
    - Identify functions versus variables.
- Control structures
    - Create conditional expressions.
    - Syntax of “if” and “else” statements.
- Data structures
    - Lists – how to define, how to add elements, how to work with individual elements.
    - Sets – how to define, how to check membership.
    - Dictionaries – how to define, how to get an element by the key.
    - Nested data structures – how to get access to an individual element.

Semi-Structured Data

- What is semi-structured data? Examples? What does it mean to have no formal data model?
- What is unstructured data? Examples?
- Compare CSV, XML, and JSON data formats and explain the advantages/disadvantages of each.
- Construct a CSV, XML, and JSON data file from raw data.

Python and Semi-Structures Data

- "for" loops for iterating over the data
- "json" package and the use of ".dump()" and ".load()" methods
- Iterate over JSON data to answer questions, e.g.,
    - Display values of a certain attribute.
    - Find the maximum value of a certain attribute.
- "pandas" package
    - Read CSV data.
    - Select a subset of columns.
    - Descriptive statistics using ".describe()" and ".value_counts()" methods.
    - Select a subset of rows.
    - Use the ".groupby()" method.

ETL

- What is it? Why is it important?
- Explain the purpose of each component (Extract, Transform, Load)
- How do inconsistencies in data get resolved?

Data Visualization

- Be able to assess a visualization by applying data three visualization principles.
    - Tell a story
    - Graphical integrity (lie factor)
    - Minimize graphical complexity (data-ink, chartjunk)
- Explain how visualization can be improved based on those principles.
- Understand basic chart types. Be able to choose an appropriate chart type given a scenario.

Hypothesis Testing

- Be able to read and interpret sample (descriptive) statistics.
- Be able to read and interpret results from simple hypothesis testing (e.g., t-test).