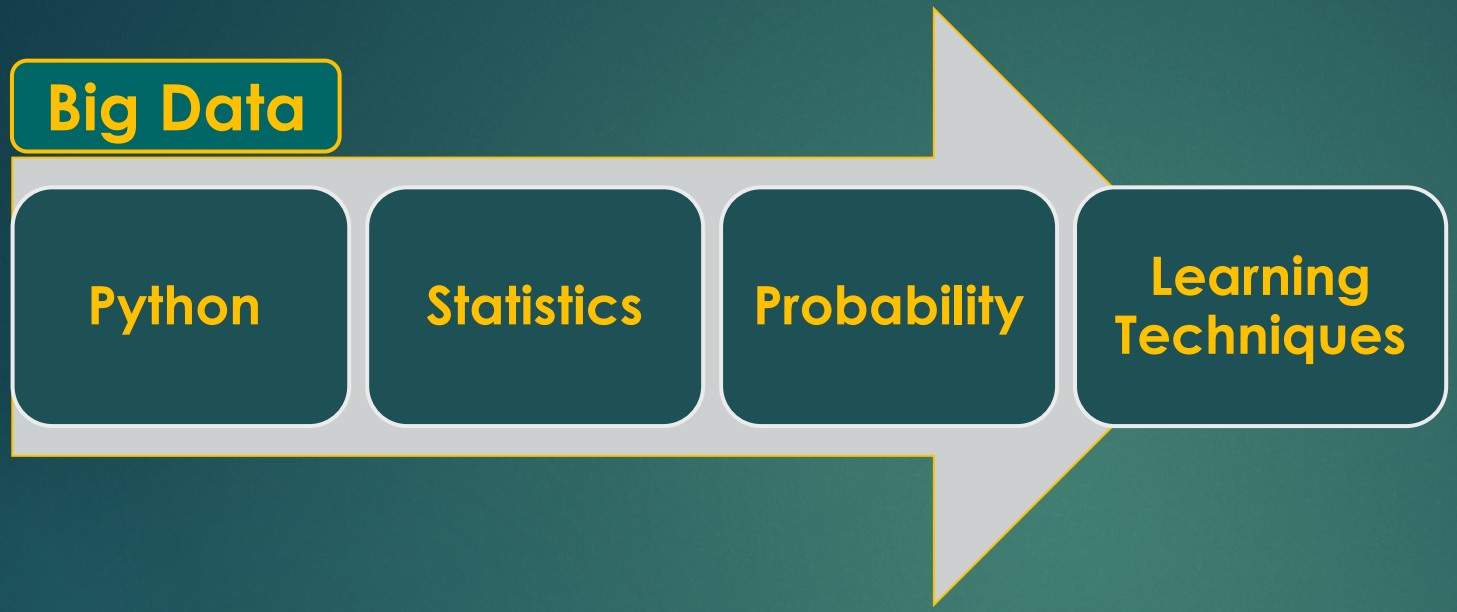
IS 669 Big Data and Information Systems
Fall 2023



Group Project Briefing

TASK SET 1

Professor HG Locklear
hlocklear@pace.edu

Phase 1

- In Phase 1 of your project, your team will explore, shape, and develop samples of your data.
- This phase serves several important purposes in the data analysis process:
  - Understanding the Data Structure
  - Identifying Patterns and Trends
  - Handling Missing or Outlier Values
  - Statistical Summary
  - Feature Engineering
  - Checking Assumptions
  - Informing Preprocessing Steps
  - Guiding Model Selection

Phase 1 Submission and Evaluation

- The Phase 1 portion of your project and the submission of your single Jupyter notebook is due on Friday Nov 17th by 1 PM.
- The submission will be done on BrightSpace by the Group Leader only…no other submissions will be accepted.
- The portion of the project is worth a total of 10 points toward your overall 40 points for the project.
- There is NO extension for this portion and late submissions receive 0 points.
- Your notebook will be evaluated on its correctness and organization…be sure to follow the specifications provided exactly as they are described.
- Code that does not execute, is in the incorrect cell, or provides incorrect output receives 0 points.

Tasks for Phase 1

- In Phase 1 of your project, your team should create a single highly-organized Jupyter notebook which accomplishes the data exploration, shaping, and sampling task listed.
- Organize your Jupyter notebook as shown below.

Cell	Content
1	Project # and all member of your group
2	All import statements
3	Data Exploration Task 1 **Narrative**
4	Data Exploration Task 2 **Imputation function**
5	Data Exploration Task 3 **code…output**
6	Data Exploration Task 4 **code…output**
7	Data Exploration Task 5 **code…output**

Cell	Content
8	Data Shaping Task 1 **Methodology narrative and code**
9	Data Shaping Task 2 **Diagram and Narrative**

Do not include any other files that may be generated by your code with your submission. I will run your notebook and generate the files

Cell	Content
10	Data Sampling Task 1 **Code**
11	Data Sampling Task 2/3 **Sample 1 Creation Code**
12	Data Sampling Task 2/3 **Sample 2 Creation Code**
13	Data Sampling Task 2/3 **Sample 3 Creation Code**
14	Data Sampling Task 2/3 **Sample 4 Creation Code**

Data Exploration Tasks

1. Provide a narrative explanation of your dataset and what type of problem it represents.
   - List each feature or feature group and explain its purpose in relation to all other features.
2. Identify any missing values and define an imputation method for replacing them.
3. Provide the descriptive statistics of the most relevant feature in your dataset.
   - Minimum and Maximum
   - Mean Median and Mode
   - Range
   - Variance and Standard Deviation
   - 1st, 2nd, and 3rd Quartiles
4. Provide a Frequency distribution of your dataset for the most relevant feature.
   - If the feature has continuous values just group, it into ranges.
5. Provide the mean of the Frequency distribution.

All output should just be text-based and NOT include any graphical visualization.

Data Shaping Tasks

1. Generate an additional 2000 instances of your data based on a methodology that will closely emulate the range of the features in the dataset.
   - Provide an explanation of this methodology.
   - Provide the additional instances as a dataframe, list of objects or csv file
   - ***If your code generates a csv file, do not include it with your submission…I will run your code and generate it myself.***
2. Define a schema for the creation of a database that could contain your dataset.
   - Provide the schema as an ER diagram within the cell of your notebook.
   - Your database should have as a minimum three tables.
   - Be sure to explain the constraints and relationships between tables as well as why you made such groupings.
   - Do not create the database in your notebook…***just show the schema as an image***

Data Sampling Tasks

1. Create a multiple-purpose function that has parameters that allow the specification of how many records are to be part of a sample and what the criteria for the sample should be.
   - Use those criteria, defined in your Lecture on Data Sampling, that are appropriate to your dataset.
2. Split your dataset into 4 separate samples based on the criteria you feel is most relevant.
   - Each sample creation should have its own cell…the samples themselves can be dataframes, list of objects or csv files.
   - ***If your code generates a csv file, do not include it with your submission…I will run your code and generate it myself.***
3. Provide descriptive statistics for each sample. (See Slide 5).
   - The code to do this should be included with each sample in its cell. (See Slide 4)
   - Each set of descriptive statistics should be in its own cell of your notebook and displayed in an informative manner that is easy to understand. ***This should just be text-based and NOT include any graphical visualization***.