**1.   Assignment Information**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| * Due Date **Monday, April 6, 2015** 11:59 PM | | * Points Possible **10** | |  | |

* **Homework 1** **Due: Midnight April 6, 2015**   You are being provided with data in a CSV file that contains student records. Each row is a record about a student containing their name, demographic information, grades, test score, attendance, and whether they graduated or not.
  1. The first task is to load the file and generate summary statistics for each field as well as probability distributions or histograms. The summary statistics should include mean, median, mode, standard deviation, as well as the number of missing values for each field.
  2. You will notice that a lot of students are missing gender values . Your task is to infer the gender of the student based on their name. Please use the API at [www.genderize.io](http://www.genderize.io/) to infer the gender of each student and generate a new data file.
  3. You will also notice that some of the other attributes are missing. Your task is to fill in the missing values for Age, GPA, and Days\_missed using the following approaches:
  4. Fill in missing values with the mean of the values for that attribute
  5. Fill in missing values with a class-conditional mean (where the class is whether they graduated or not).
  6. Is there a better, more appropriate method for filling in the missing values? If yes, describe and implement it.

You should create 2 new files with the missing values filled, one for each approach A, B, and C and submit those along with your code.   Please submit the Python code for each of these tasks as well as the new data files for this assignment.  [mock\_student\_data.csv](https://chalk.uchicago.edu/bbcswebdav/pid-2808823-dt-content-rid-5579006_1/xid-5579006_1)