**CS-521 Final Project - Program Description**

Jose Uzcategui (joseuz@bu.edu)

Saturday, May 4, 2019

**Overview**

The programs allows a user to create an object of class type ParseCSV and access to various methods.

The current version of this class allows the user to invoke methods that (a) verify the file’s column headers, (b) returns descriptive stats for selected columns and (c) uses Python random module to creates a simple random sample of the data set. This last method returns a sample dataset as well as the remaining dataset equivalent to the original file minus the generated random sample.

**Why is it useful?**

Training machine learning algorithms often require splitting dataset into, at the very least, two different sets: training and test. At the moment the program uses a few methods to describe data and the Python random module to generate a training and test set using a simple random method. That said, this class can be used as a foundation to further describe the dataset, as well as using other sampling methods to better fit specific industries or situations. For example, stratified sampling for web analytics or clustered sample for sentiment analysis.

**Methods of Interest in Main File**

The following methods are included in the **main.py** file.

* *headerCols()* - Prints the available column headers along with a numbered index for easy manipulation.
* *describeCols()* - Shows descriptive stats for columns input by the user. The current version differentiates between “string” and “numeric” type columns.
* *getSample()* - Returns a tuple with two elements based on the read file.
  + Index[0] returns a random sample
  + Index[1] returns the remaining table
  + Passing 'p' as an argument (e.g. *getSample(‘p’))* will create CSV files named “sample\_table.csv” and “remaining\_table.csv”.

**Methods of Interest in Helper File**

The **input\_validator.py** file includes various helper methods that are used to validate user input or to improve readability in the main.py file.

* *fileChecker*() - Check is a file exists and provides a message to the user if it doesn’t.
* *colInputVal*() - Prompts the user to enter valid column numbers to run descriptive stats.
* *sampleSizeValidate*() - Prompt the user to enter a valid percent value used to create the sample data set.