Homework-7

**Out Date:** 10/25/2019 (Friday)

**Due Date:** 11/07/2019 (Thursday) 11:59PM

Team#: \_\_\_

Team Member-1:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

Team Member-2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

**Submission**

1. Work on the Problme-1.
2. Prepare your Python file for Problem-1 (e.g., HW7\_P1\_Team#.py).
3. Upload the files to blackboard.

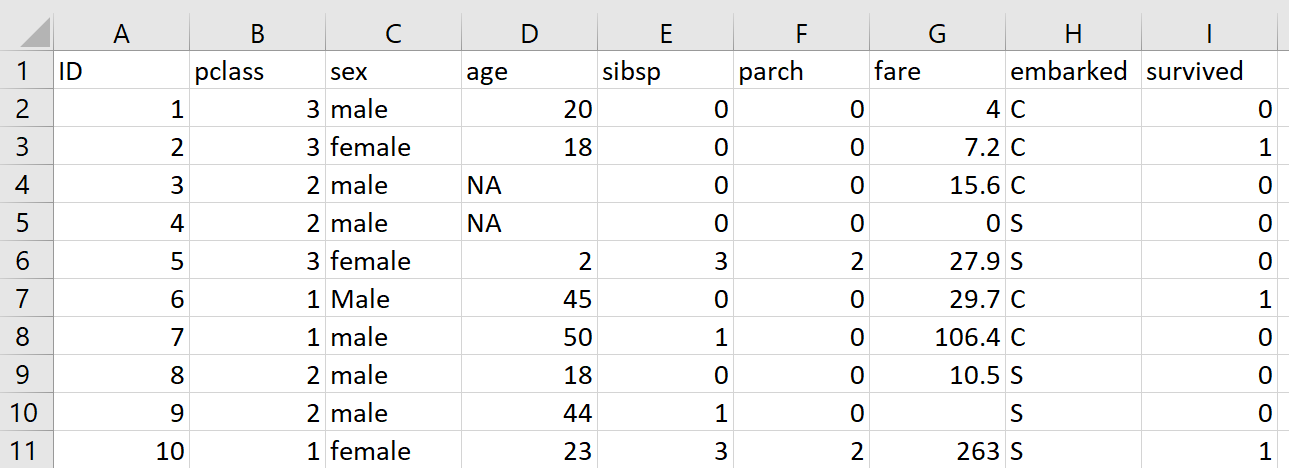
**Problem Statement:** Write a Python script to

* **Read-in** the *titanic* data (titanic\_traning.csv). The data is available in the homework folder. You may use read\_csv(). **[5 points]**
* **Identify the severity of the missing value problem and data inconsistency problem [20 points].** Specifically, generate summary of missing values, and inconsistent values for each of the features. Your script should generate a table similar to the one shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | Missing Values (MV) | % of MV  (MV/n) | Inconsistency Values (IV) | % of IV  (IV/n) |
| pclass |  |  |  |  |
| sex |  |  |  |  |
| age |  |  |  |  |
| sibsp |  |  |  |  |
| parch |  |  |  |  |
| fare |  |  |  |  |
| embarked |  |  |  |  |
| survived |  |  |  |  |

n is the total records.

An example of a missing value is NA in record 4 and the feature *Age*. Another example of a missing value is for record 9 and the feature *fare*.



An example of inconsistency is Male (with M upper case) in record 6 and the feature *sex*. Another example of inconsistency is Queenstown (in place of Q) in record 18 and the feature *embarked*.

* List/Display all the records with missing value(s) and/or inconsistent value(s). **[10 points]**
* **Handle the missing values [20 points].** Specifically, estimate missing values of quantitative features by computing the feature mean. Estimate missing values of nominal and ordinal features by computing the feature mode.
* **Correct the data inconsistency issue [20 points].** For instance, the feature values of *sex* should be lower case male and lower case female. Similarly the feature values of *embarked* should only have abbreviations.

* Store the clean data in a csv file. **[5 points]**

**Data Fields**

* **pclass:** Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
* **survived:** Survival (0 = No, 1 = Yes)
* **sex:** Sex (Male, Female)
* **age:** Age in years
* **sibsp:** # of siblings / spouses aboard the Titanic
* **parch:** # of parents / children aboard the Titanic
  + Parent = mother, father
  + Child = daughter, son, stepdaughter, stepson
  + Some children travelled only with a nanny, therefore parch=0 for them.
* **fare:** Passenger fare
* **embarked:** Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Please make sure your code follows the Python programing style guide available here: <https://www.python.org/dev/peps/pep-0008/> **[10 points]**.

Please make sure the code is well-commented **[10 points]**