Due: _____/_____/_____

R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **Data Science**

DSCI-633-2215
Lab #2
Data Cleansing with Python

(Not a Team Assignment**)**

Overview

In this exercise, you will have an opportunity to investigate how you can use Python to cleanse data in preparation for machine learning.

After completing this exercise, you should be able to:

- First, show how you can use Python to check for problems in data.
- Then, discuss why data cleansing is a complicated but essential process.

For this exercise, you will need:

- Python/Jupyter Notebook & Python Libraries like Pandas by providing comments properly.
- A copy of the data file, patients.txt, is available under the Lab #2 folder on the RIT *my*Courses for this course. The delimiter of the patients' data file is not a comma.

Step #1: *Get the Data*

Download the patient data file onto your machine. Open and investigate the data. Note the number of rows.

The format of the data in the patients.txt[1] file is:

Variable Name	Description	Length	Data Type	Valid Values
patientNo	Patient ID	3	Character	Numbers only; If missing, duplicate or none alpha, assign a unique number;
gender	Patient gender	1	Character	'M' or 'F'
visit	Visit date	8	Character (MMDDYYYY)	Any valid date; If missing, 1/1/1900; If month>12, 12; If day>31, 31; If year>1999, 1999; If non-digit, 1/1/1900

[1] Source: "Cody's Data Cleaning Techniques, Using SAS Software," Ron Cody, SAS Institute Press, 1999.

HR	Heart rate	3	Numeric	≥ 40 and ≤ 100; If missing, 40; Otherwise, outliers?
SBP	Systolic blood pressure	3	Numeric	≥ 80 and ≤ 200; If missing, 80; Otherwise, outliers?
DBP	Diastolic blood pressure	3	Numeric	≥ 60 and ≤ 120; If missing, 60; Otherwise, outliers?
DX	Diagnosis code	3	Character	1- to 3-digit number; If missing, 999; If non-digit, 999
AE	Adverse event	1	Character (Boolean)	'0' or '1'; If missing or invalid, 0;

Step #2: *Data Cleansing*

You need to cleanse and validate the file from applying operations, including

- Missing values
- Duplicates
- Bad characters or NULL values
- Invalid values or data types

If you are new to Python, the following articles will guide you on how to cleanse a dirty dataset:

- **Cleansing a messy dataset using Python**
- **Data Cleaning and Preprocessing for Beginners**

Step #3: *Submit your py or ipynb file as Lab #2_LastName into the DropBox by 11:59 PM, Monday, February 14.*