Student name: yizheng he

Student ID: 221411294

# SIT123: Data Capture Technologies

## Lab Report 2.2:

## Preparing data (30 marks)

In this task, you will investigate a given data set, look for inconsistencies in the data and propose methods to fix them.

## Due Date Friday 8:00pm, Week 3 – 29th July 2022

## Pre-requisites: You must do the following before this task

1. **Attend Class (Lecture)**
2. **Read this sheet from top to bottom**

## Task Objective

1. Use the provided dataset HumidityDataset.CSV on collection of Humidity values for location X. The file is available here: https://d2l.deakin.edu.au/d2l/le/content/1190434/viewContent/6267134/View
2. Investigate the data for inconsistencies. These inconsistencies could include:
   a. Missing data, rows and column values
   b. Mismatched data fields
   c. Mismatched date formats
3. Propose ways to fix consistencies, these fixes could include:
   a. Propose and use default values
   b. Remove missing rows
   c. Fix data format mismatches

4. Fix the data

## Task Submission Details

There are 2 questions in this task. Answer all of them in this document itself and submit to unit site.

### Q1: Once you have cleaned your data, submit the cleaned data file to unit site with this document.

(15 marks)

### Q2: Submit brief details on which inconsistencies you have found, what was your approach for fixing them and discuss the Pros. and Cons of your approach, using the given table below:

| Inconsistencies found | Approach for fixing | Pros. and Cons of your approach |
|---|---|---|
| Missing rows ID | Rewrites the missing data from the data before and after the missing line. | This is easier to do than deleting a whole row |
| Missing Stamp | Rewrites the missing data from the data before and after the missing line. | This is easier to do than deleting a whole row |
| Wrong dateline | Correct the wrong date below according to the date above | This is easier to do than deleting a whole row |

(You may add more rows to the table as required)
(15 marks)

Hum Null , “64.70”

| Inconsistencies found | Approach for fixing | Pros. and Cons of your approach |
|---|---|---|

| | | |
|---|---|---|
| Issue of Humidity | Change the incorrect temperature or null to the correct temperature or write null as 0 | This has the advantage of avoiding unnecessary deletions |
| Stamp duplication | Calculate the median data from the above and the following data | This has the advantage of avoiding unnecessary deletions |
| | | |