# SIT107 – 2.3P – Preparing Data

## 212147139 – Matthew Gray

## Data Analysis

The humidity data provided for location X contains several inconsistencies, which have been identified by analysis.

Most faults were identified by graphing the data, which immediately highlighted any out-of-range values. This was followed by a visual scan of the data to further identify any inconsistencies.

* Row 342 Contains no data
* Between row 8480 and 8481, 5 minutes of missing data
* Row 318 – formatting error for the humidity reading
* Between 20256 and 20265 – no humidity reading
* 26606 and 26607 no humidity data
* 26951 – Formatting error for stamp
* 14911 – Date formatting error
* 24384 – No RowID

To fix the data the following are proposed

* Fix formatting errors
* Remove blank rows
* Remove rows with missing humidity data
* Leave missing rows as-is
* Enter data for missing stamp that would be appropriate (not critical data)

Q1. Github repository:

Q2. Summary of fixes are as follows:

|  |  |  |
| --- | --- | --- |
| Inconsistency Found | Approach Used for Fixing | Pros/Cons |
| 342 – no data | Removed row | Doesn’t replace lost data, hole in data can cause problems for graphical formatting |
| 8480 – 8481 5 minutes missing data | Leave as is | Will cause problems with graphical formatting, however data should not be manufactured. |
| 318 – Humidity reading formatting | Corrected | Nil Cons, data is consistent with range of data |
| 20256 – 20265 no Humidity reading | Remove rows | Will cause problems with graphical formatting, however data should not be manufactured. |
| 26606/26607 no humidity reading | Remove rows | Sensing data cannot be manufactured, thus removed. Data accuracy maintained, but |
| 26951 – Formatting for stamp | Correct formatting | Nil Cons |
| 14911 – Date Formatting Error | Corrected | Nil Cons |
| 24384 – No RowID | Row ID Created at mid-point between previous and next | No impact on sensor data, thus can be manufactured to allow improved readability. |