Data Cleaning Project (Project 2)

**Data:**

This data set covers hourly weather data from 122 weathers stations of southeast region of Brazil. The southeast include the states of Rio de Janeiro, São Paulo, Minas Gerais e Espirito Santo.

Dataset Source: INMET (National Meteorological Institute - Brazil).

Content:

This data has 8 climate parameters that are continuous values from these 122 stations. The complete data list is from 2010 to Sept of 2016.

Columns:

* wsid Weather station id
* wsnm Name station (usually city location or nickname)
* elvt Elevation
* lat Latitude
* lon Longitude
* inme Station number (INMET number) for the location
* city City
* prov State (Province)
* mdct Observation Datetime (complete date: date + time)
* date Date of observation
* yr The year (2000-2016)
* mo The month (0-12)
* da The day (0-31)
* hr The hour (0-23)
* prcp Amount of precipitation in millimetres (last hour)
* stp Air pressure for the hour in hPa to tenths (instant)
* temp Air temperature (instant) in celsius degrees
* dewp Dew point temperature (instant) in celsius degrees
* hmdy Relative humid in % (instant)
* wdsp Wind speed in metres per second
* wdct Wind direction in radius degrees (0-360)
* gust Wind gust in metres per second
* Highlighted Columns will not be touched

**Project parameters:**

You are to clean this data set in preparation to do analysis for things like predicting rainfall or temperatures forecasting. You will not be asked to do the predictions as this exercise is to use the libraries we have studied and cleaning techniques we have discussed to prepare the data for processing.

Things to consider but you need to find others as well:

1. Are the precipitation numbers in-line with what the standard deviations/normal range?
2. Are there any zero’s where there should be numbers?
3. Are there any blank, null or NaN entries in the data set?
4. Are the temperature numbers within a standard deviation like
   1. Is a temperature change a bad sensor reading or a real change (i.e. drop 20 degrees in an hour)
5. Is the longitude and latitude of the stations consistent for the same station across the data set?
6. Find 3 more anomalous themes and resolve across the data set.

**Deliverables:**

1. A Report outlining what you did, what methods you used to clean/sanitize/create/augment the data set. Why you chose those specific methods and what the outcome was
2. You Python code (where you used the libraries we have discussed) along with documentation on what libraries and versions
3. An output file of cleaned and pre-processed data.