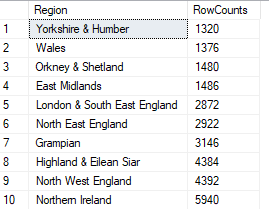
Weather Notes and Assumptions:

### Summary:

### Initially I browsed the data to understand the contents of the file.

My understanding from this was:

1. There are duplicates in the file.
2. Most of the columns were in string format.
   1. Change the necessary columns to correct data type to answer the questions from the use case. I.e. change ScreenTemperature from string to decimal, ObservationDate from datetime to date.
3. It seems that the ScreenTemperature ‘-99.0’ represents the cases when a temperature value is not available for an observation (i.e default value). This is the most frequent value.
4. It seems that for each given ForecastSiteCode, the temperature is recorded at every hour. This you will see in column ObservationTime where the recording starts at 0 hour and finishes at 23 hours on each day / date (i.e.ObservationDate)
5. There are 10 distinct regions with blank Country names in them. 
6. File and the Python script breakdown:

|  |  |
| --- | --- |
| **Lists** | **Comments** |
| CSV files (weather.20160201.csv, weather.20160301.csv) | These files were saved in the local folder to read. |
| Parquet file (output) | The convert file from .csv to parquet was saved in the local folder. |
| weather\_task.py | Python script where we read the source files and convert them to the parquet format while answering the use case questions.  The script is written in Pyspark and Spark SQL is used to query the data. |
| Output from the python file |  |