Analytics Shootout Question and Answer Session #1 – February 25th, 2015

The following are questions submitted and discussed during the first Q&A session for the 2015 Analytics Shootout:

The variable Hour in SolarArray_Weather.sas7bdat ranges from 0 to 23, but Hour in SolarArray_Production.sas7bdat and SolarArray_SolarAngle.sas7bdat ranges from 1 to 24. Can we assume hour 0 in SolarArray_Weather.sas7bdat is hour 1 in SolarArray_Production.sas7bdat and SolarArray_SolarAngle.sas7bdat? **Yes. For a 1-24 clock, 24 is (23:00:00 - 23:59:59) and 1 is (00:00:00 – 00:59:59). For a 0-23 clock, 23 is (23:00:00 - 23:59:59) and 0 is (00:00:00 – 00:59:59).**

Two data sets, Powercity_population and Solararray_weather, cannot be open by SAS directly. Is it a part of this competition? **For some of the datasets, you must change certain variable names to read them into Base SAS. Another option would be to read the files into a different SAS program, such as JMP or Enterprise Guide.**

Question 4: In "Solararray_weather.sas7bdat", variables of "Precipitation" and "Pressure" have missing values after 2012. Is it a task of the competition or a data source problem? **As with any real data, sometimes there are instances of missing of incomplete records. In this instance, the sensor at the Solar Array did not record full precipitation and pressure over the course of the 5-year period. As a team, you must decide how to handle this in your modeling efforts.**

	Year	COUNT_of_Precipitation	COUNT_of_Pressure
1	2010	8729	8670
2	2011	8732	8642
3	2012	8730	8668
4	2013	1584	0
5	2014	957	0

In dataset powercity_weather_scenario.sas7bdat, there is a gap that has no record of Wind_Speed between 0-1.5m/s. Can you explain? **The sensor does not record wind speeds under 1.5 m/s.**

Data dictionary indicates that the year in dataset SolarArray_SolarAngle.sas7bdat is "Consumption", but in the dataset the year is real year from 2010 through 2014. Is there any explanation for that? **The data is from 2010 to 2014. This table can be used to build the models that will be used to score the "Scenario" year data.**

Data dictionary says there are 18704 rows for WindFarm_Production.sas7bdat, but there are actually 15381 rows in this dataset. Are there supposed to be more data supplemented? **15381 is the correct number of records. This is a typo in the Data Dictionary. An updated version of the Data Dictionary is being provided to you.**

What's the relationship/correlation between variable Precipitation (in dataset powercity_weather_scenario) and variable Precipitable_Water (in dataset powercity_weather_consumption)? **They are not the same variable. From the data dictionary:**

- **Precipitable_Water: The total precipitable water contained in a column of unit cross section extending from the earth's surface to the top of the atmosphere in millimeters**
- **Precipitation: Amount of precipitation during the hour in milimeters**

Are the value of all variables in SolarArray_SolarAngle.sas7bdat and SolarArray_Weather.sas7bdat are measured at the beginning of the hour? or at the end of the hour? or is the average of that hour? **This depends on the variable. Most variables are the average over the course of the hour. Precipitation, on the other hand, is the total amount during the hour.**

The wind_speed we are given for the scenario year appears to be from powercity, rather than the windfarm location. Are we to assume that powercity has installed a local windfarm and use their wind_speed information for power generation? **Yes, assume that the wind speeds in Power City are the same as the wind speed at the Wind Farm.**

This applies to solar weather information as well, as weather conditions at powercity are obviously different from the solar array site. **Yes. For the scenario year, assume that the Wind Farm and Solar Array is located in Power City. As such, use the Power City weather variables to score the Wind and Solar production for the scenario year.**

Is using an R code node to run models we can't fit using EM going to be frowned upon? **As long as you also use SAS software, additional software use is allowed.**

Are we to assume that population and sector use matrix will remain constant from the consumption to scenario years? **Yes, assume the population does not change in the scenario year.**

Are the wind times or hours adjusted to reflect that some months are during standard time and some are during daylight time, or do we need to make the adjustment? **All datasets are in Local time and should follow the US rules for Daylight Savings Time. You can further examine the data to see if there is any additional consideration to make for the time.**

The Problem Statement states that the Solar Elevation data set contains data for Power City, the Wind Farm and the Solar Array, however the only location on the data set is Solar Array. Is this a mistake in

the data set or in the Problem Statement? **There are separate solar elevation/angle datasets for each location (Power City, Solar Array, Wind Farm).**

There are times when the wind speed is greater than 0 but the electrical energy produced by the wind farm is greater than 0. Can you comment on that? **When there is enough wind, the turbine will produce energy. However, there may be times when the turbine is either not running or the sensor is temporarily not working. You will need to make a decision about how to handle these occurrences.**

The data dictionary lists the hour range for the solar production data as 1-24. We cannot verify this since the solar production data only contains hours in the range 7-22. Is the data dictionary correct? **Yes, the Solar Production data uses a 1-24 hour clock**

There is a lot of variation in the hour range of the production data from year to year and even day to day. Should we assume that a lack of data for hours near the top and bottom of the hour range implies that no energy was produced during those hours? **Solar production is only possible during the daylight hours.**

Is the solar energy being produced by a Concentrated Solar Power (CSP) plant or a Photovoltaic (PV) Power plant? **Photovoltaic (PV)**

Is there a reason behind certain dates not having any data? Is this due to missingness at random or was no power generated or recorded? **You are provided with real data and the problems that often occur using real data sources. Therefore, you may find that some data is missing or incomplete.**

All of the electricity values are multiples of four. Is there a reason behind this? Should the values in our model only be multiples of four for this variable?
The electricity values of the windfarm production are a multiples of 4. This is how the sensor measured the data. You do not need to score your data in multiples of 4.

Will holidays and weekends indicate that sectors other than school will also be on holiday? Or do we use logic? **Correct. The type of day applies to all sectors of the city.**

Scenario year, month-02, day-29 is a Friday but Consumption year month-02, day-29 is Sunday. Hence scenario year and consumption year is different. What do these two years: scenario and consumption refer? **The Consumption Year and Scenario Year are indeed different, but you do not know what specific years they are.**

In solararray_production data, number of hours for power generation varies from 7 to 14 hours a day. It does not have any systematic order and it varies from year to year. Is it intentional or is there any discrepancy? **The solar array will only produce electrical energy when adequate solar conditions occur. Therefore, there are times when the conditions are such when no power is being generated. You can gain more insight into this once you build and examine your models.**

File: WindFarm_Production.sas7bdat; File is referring to wind farm production, but the description in the Data Dictionary says that the Electricity _KW_HR is from "solar energy." Should this be "wind farm energy" in the Data Dictionary? **Correct, this is a typo in the data dictionary and has since been updated.**

File: PowerCity_Population.sas7bdat; Tract_ID or any reference therein, appears to be in only one data file. Is this intentional? **The population file is structured similar to what is provided from the US Census Bureau. The population from the city was collected at the Census Tract level. However, you are only using the total city population and do not need to do any modeling at the Census Tract level.**

File: PowerCity_Population.sas7bdat; some of the values for the age groups appear to be in decimal format (i.e., maximum value for group <5 is 507.998); How are the values for the group calculated? Is it an average or some other summary statistic? **The population file was put together using only the portions of the Census Tracts that fell within the city boundary. By using an Area-Weighted distribution, segments of the tracts that did not fall within the city boundary were removed and the tract population was adjusted based on the area that was inside the city, thus resulting in the decimals.**

For the Calendar_Days dataset in the consumption year, some of the holidays do not occur on the correct date (Independence Day is shown on July 5th and Christmas is being shown as December 24th). Is there a reason for this? **The dataset designates the "US Federal" holiday for Independence Day and Christmas because they both fall on weekend dates. The true holiday dates are July 4th and December 25th, respectively. You may examine how or if there is any difference between the true holiday date and the federally recognized date.**