**Energy Project**

You are given data of UK Power Networks gathered using SmartMeter technology. This dataset is gathered from sample of 5,567 households between November 2011 and February 2014. Readings were taken every half hour.

Data Columns:

* Customer/Household Id (LCLid)
* Household tariff type (stdorToU):
  + Within the data set are two groups of customers. A sub-set of 1,100 customers (Dynamic Time of Use or dToU) were given specific times when their electricity tariff would be higher or lower price than normal – High (67.20p/kWh), Low (3.99p/kWh) or normal (11.76p/kWh). The rest of the sample (around 4,500) were on a flat rate of 14.228p/kWh.
* DateTime: Reading of every half hour.
* KWH/hh : KWH per half hour
* Accron : Ingore this field
* Accron\_grouped : Customer groups on income.

This project would have 3 deliverables. Keep on adding the deliverables to the same document. This way, final report would be a comprehensive report.

1. **Exploratory Data Analysis (EDA)**

For this deliverable, focus on initial data exploration and graphs. This could include but not limited to:

1. Finding missing values and replacing them by an appropriate strategy.
2. Calculating and visualization of summary statistics (consider boxplots)
3. Finding correlation between attributes.
4. Finding out dependence between categorical and numerical attributes.
5. Average energy usage in each month by each customer group
6. Average energy usage by time of day (If Any single day)
7. Average usage by weather in DToU and STD group
8. **Cluster Analysis**
9. Perform cluster analysis of customer and seasons (Use your own imagination for how many clusters should be made. It depends on the objective you are trying to achieve. So first think of what insight you can get and then go for actual implementation).
10. Perform outlier analysis and report your findings.
11. **Final Report**

For this deliverable, focus on following points. Try to apply all the concepts learned during the course. You would not be told which techniques to apply but rather what objectives to achieve. You should be able to identify the approaches and techniques. Your conclusions, recommendations, reasoning, and findings should be supported by the relevant graphs/visualizations considering the energy domain. You need to identify what meaningful information can be extracted from this data.

**Dataset:**

The full raw data is available to download from this drive folder.

<https://drive.google.com/open?id=12Be8raikn_cKS6qrD2k77lYceG6vgyik>

The data folder has 168 csv files, each file contains 1 million data rows. A sample file is also available in main folder.