Predicting energy consumption at a hourly frequency for a region in US

# Problem

Electrical utilities need to diligently plan ahead of time the allocation of generating units in their power plants to match their regional energy demand (MW), because if the demand is higher than the generation it can cause several blackouts resulting in a huge loss to the economy; on the other hand if the generation is higher than the demand the extra electricity will be wasted and it can also create an unnecessary load on the transmission lines.

So, it is very important for the utilities to have a forecast of the energy consumption to be able to allocate appropriate resources to meet their demand. A year, month or day ahead forecast can help the utilities plan for a larger time scale but for smoother daily operations an hourly (or even better) forecast can prove very useful. For example, if the plant operators get a high energy forecast for the next hour, they can start the gas or coal fired power plants, which can take upto an hour to start, and thereby avoid a potential shortage.

The project will involve analyzing past 4 to 10 years’ of hourly energy consumption data of a region to find trends in energy consumption around hour of the day, day of the week, season of the year, outside temperature, etc. That is, a model can be built to predict the energy consumption given parameters like day of the week, time of the day, summer or winter, holiday or not, local weather, renewable capacity added in the local market, etc.

# Client

The developed prediction model can be utilized by the electrical utilities to effectively plan their energy generation operations and balance the demand with appropriate supply. An efficient forecast can prove very useful for the utilities in planning their day to day operations, meeting their customers’ energy demand, and avoiding excess generation of energy.

# Data

There are two potential datasets I am looking at to use for this project:

* Kaggle PJM electrical utility data
* The first dataset is available on kaggle ([here](https://www.kaggle.com/robikscube/hourly-energy-consumption)). This dataset consists of hourly consumption values in MWh for different sub-regions within the PJM electrical utility region (east coast). Some of the sub-regions have data ranging from 1998 to 2018 but that is not consistent across all the sub-regions; some sub-regions have data starting only from 2013.
* Also, along the years some of the sub-regions got divided into smaller sub-regions, for example, the PJM sub-region was divided into PJMW and PJME around 2002. So, to keep the integrity of the data but at the same time to be consistent across all regions of the dataset, I am planning to truncate the data such that all the columns will have equal number of non-null observations.
* I will also try to introduce the weather data parameters from NOAA’s climate.gov site ([here](https://www.climate.gov/maps-data/dataset/past-weather-zip-code-data-table)) into a separate column for each sub-region, this will give the model a chance to evaluate the effect of weather on energy consumption.
* For matching the weather files with the kaggle data and also to extract any kind of regional information from it, I will require some kind of regional coordinates for all the sub-regions included in the main data which at present I am not sure of.
* US Energy Information Administration (EIA)
* This dataset available on US EIA’s website ([here](https://www.eia.gov/opendata/qb.php?category=2122628)) includes hourly energy consumption, quarterly residential/industrial renewable energy capacity, annual population, quarterly heating and cooling degree days, etc. for larger regions (in size) as compared to the above kaggle data. For example, in EIA’s data, California has only 3 sub-regions and entire Arizona is represented as a single region.
* Since the regions are larger here, I am a bit hesitant to use local weather data because the weather can vary considerably across a larger region like California as compared to a smaller sub-region. But, if weather data is to be used, I am considering using the same NOAA data as mentioned [above](https://www.climate.gov/maps-data/dataset/past-weather-zip-code-data-table).
* The datasets available on this site range from Jul, 2015 to present day (~4 years).

# Approach

* First step will be to finalize one dataset out of the above two based on some initial analysis on both of them.
* After picking a dataset, the parameters to be included will need to be finalized. The default variables to be used will be hour of the day, day, month, weekday/weekend, holiday/working-day and summer/winter which can be extracted from the datetime index.
* Other parameters like the weather variables- temperature, rain, snow, etc. will need to be matched appropriately with the region and added as separate columns.
* After cleaning and preparing the above dataset, the next step will be doing EDA. Basic trends can be observed during the EDA and which parameters affect the energy consumption the most can also be determined.
* Initially I am planning to treat this as a prediction problem using the following models: regression, decision trees, gradient boosting, etc.
* After the initial analysis, I will model the problem as a time forecasting problem. I need to do more research on the time series forecasting models.

# Deliverables

* Jupyter notebook including the code.
* Detailed report explaining the entire process from data acquisition to discussing model solutions.
* Slide deck for presentation purpose.