MARCH 26, 2017

ENERGY & ME

A VISUAL EXPLORATION ON HOW VICTORIANS USE ENERGY

AARON (ARUNA) TILLEKERATNE

GDF SUEZ

Level 33, Rialto South Tower 525 Collins Street, Melbourne VIC 3000, Australia

BACKGROUND

In modern Victoria, we often pay little to no attention to our energy, other than the occasional instance where the bill is too high or getting energy connected at our new house. Yet it plays a crucial and irreplaceable role in our lives. We depend on the energy market for our basic necessities such as hot water, heating and light as well as luxuries of television and computers.

These two facts,

- We pay very little attention to the way we use energy (most of us)
- It is a basic need and we all use it

, gives Data Scientists, and data enthusiasts a potential gold mine of candid, unfiltered human behaviour to investigate.

CHALLENGE

The proposal for the energy usage is rather an exploratory data analysis piece as opposed to a problem solving one. That is, instead of a one central question, there are a number of minor, insights that are explored. We can summarise these insights we hope to gain as follows:

- How much on average are Victorians spending on their bill a year?
- How is this spread across the state?
- Are there distinct patterns in the daily usage?
- Are there distinct patterns in the yearly usage?
- Are there any correlations between the demographics in areas and their usage structure?

These are very fundamental and intuitive questions, which understandably, can give rise to even more questions, particularly, the question "why is there…?". Since the intent of this project is to visualise the '*kinematics'* of the system, we will not consider a deep dive into the causal implications of the data.

DATA SOURCES

For the above data exploration, we can define 3 conceptual data requirements.

- Consumption data
- Billing data
- Demographics data

CONSUMPTION DATA

The consumption data is the key to identifying the patterns and clusters in daily consumption usage. It was fortunate that an anonymised smart meter interval data has been made available to this project by the ENGIE group. It should be noted that this is not public data.

The layout of the data structure is in a star schema format with facts and dimensions clearly defined. There are also date keys and postcode keys assigned to observations.

BILLING DATA

The billing data is also facilitated by the ENGIE group and is available an anonymised, aggregated format. Also not available as public data, the typical structure will involve bill amounts, billing periods and consumption allocated to each bill.

DEMOGRAPHICS DATA

The demographic data is extracted via data packs in the Census data. The main data of interest will be the postal code level aggregated measures. These are available through the Census data website.

DATA PRIVACY

There will be absolutely no personally identifiable data made available in the output of this project. Post extraction, prior to any analysis algorithm or implementation of a cloud based version control system, all personally identifiable data (if unavoidable) will be hashed through a MD5 hashing algorithm.

In addition, non-public data will not be made explicitly available, if required, it may be requested from the author of the project. Author reserves the right to refuse the facilitation of any data, references relating to the data or any other entity, which may be non-publicly owned or owned by the ENGIE group, without the provision of any reason at any time.

DATA WRANGLING & STRUCTURES

Fortunately, most of the data available is mapped out in either snowflake schema or relational schema, which greatly reduces the data wrangling process.

There is however, still the need to clean and filter data that may be outliers or contain a large number of missing values.

Most of the unclean data can be addressed in the database layer by filtering out instances with too many missing values. An example may be, if 30-minute interval data isn't available for at least for 300 days of the year, that meter is not used in the analysis. Although it seems such a high filter, there is still approximately 200GB of data available in smart meter data in Victoria with that condition.

The Census data is already pre structured as a package in the Data Packs with postal-code used as the identifier. Identifier may change from the type of data-pack used, however this project will utilise the postal-code level aggregated data.

PRELIMINARY EDA

As somewhat of a motivator to the above questions, some preliminary exploratory data analysis was carried out on the consumption data. A database random set of 2000 meter readings were extracted across 2015 and 2016 for the preliminary exploratory data analysis.

The first topic of interest was the daily consumption. Are there clear clusters in consumption profiles? A use of a simple K-Means algorthm with K = 5 showed that there are indeed distinctly different groups of daily consumption profiles.

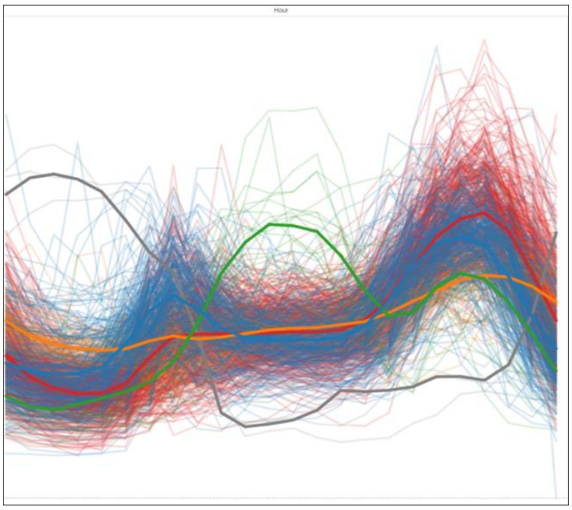

Figure 1: Daily consumption (2000 Database-Random Meters)

Secondary interest was the yearly consumption profiles. These were also subject to a simple K-Means treatment, where K = 7. The K-Means algorithm proved to be less effective in clustering in the yearly profiles, as its significance turned to focusing on overall usage rather than the shape of the curve.

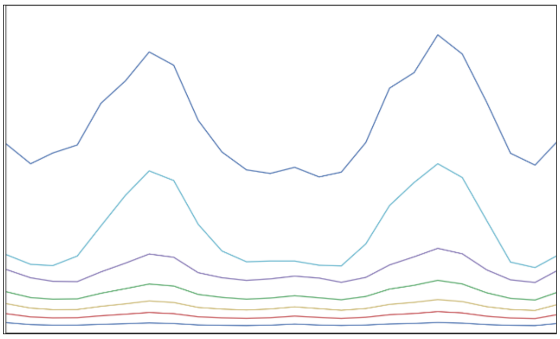

Figure 2: Yearly load profile averages.

DATA PRODUCTS & AUDIENCE

There are various products which may come as a result of the analysis, with target audiences ranging from retailers to home owners and energy users.

Typical load profiles and demographics can help customer targeting and product design for energy retailers as well as a basis for load forecasting for generators.

It can also serve as a basis for products for energy consumers to see how their consumption stacks up with the regional profiles and averages. Of course both of these will require more development from the initial proposal and visualisation of the data.

CURRENT STATUS

Currently, the project is migrating the 200GB of 30-minute interval data for all available meters in the ENGIE database. Next steps include the aggregation of bill data and going forward the visualisation of the Census data.

Once this has been completed, the integration of data sets and exploration of the data can begin.