**Capstone Project Proposal**

Jorge Moscat

* **An introduction to the problem (based on your earlier Capstone submissions).**

I want to understand how can energy consumption data from smart meters installed in homes can be use to model the routines and behavior of customers. My plan for this capstone project is to first do an initial data exploration and then implement the following use cases/analyses:

* Determine relation between weather and energy consumption
* Model customers' behavior: weekly energy usage
* Identify when customers are away from home based on usage analysis
* **A deeper dive into the data set:**
  + **What important fields and information does the data set have?**

The dataset I will use for my capstone is a pipe-delimited file, each record correspond to a customer and a day of between March 2013 and October 2015; for each record there are 24 columns that correspond to the electricity consumption readings for each hour of the day. Below I am including a quick description of the each of the fields in the dataset and a short explanation of which fields I am including in the scope of the project:

**DAY**: The date on which the energy usage occurred, in the following format: YYYY/MM/DD

**H1, H2,., H25**: The time at which the energy usage occurred. There are 25 entries due to the time change when the clocks are set back one hour on the last Sunday of October each year, which leads to a total of 25 hours on this day.

**ACTIVA\_H1, ACTIVA\_H2,., ACTIVA\_H25**: Active energy consumed per hour as measured in kWh (kilowatt-hour). This is the useful energy that the customers absorb from the grid and transform into work and/or heat at home.

**CITY**: The municipal district to which the customer belongs. Geographical reference.

**TARGET\_TENENCIA\_CUPS**: Probability that the municipal district in question is already equipped with a natural gas distribution network (which does not imply that the customer had contracted natural gas service).

* + Not part of the analysis: This column did not add any relevant information to the specific analyses I performed; therefore I decided to not include it.

**CLIENT\_ID**: The unique, customer reference number which allows for segmentation of usage per customer.

**CNAE**: (National Classification of Economic Activities) This value indicates whether the customer is domestic (T1) or not (T2).

* + Domestic customer: To reduce the scope and complexity of the analysis I decided to only include domestic clients. Trying to model the behavior of non-domestic clients (i.e. commercial clients) can be far more complex as there are many other factors that impact the electricity consumption of this kind of customers (type of business, size of the company, seasonality, etc.)

**PRODUCT**: Tariff / electrical product that the customer has contracted; there are up to 120 products.

**MARKET**: This value indicates whether the customer has a regulated tariff (M1) or a free market tariff (M2).Regulated Tariff: the price of the electricity is regulated periodically by the corresponding authority. Free market: the price of the electricity is freely agreed upon by the provider and the customer.

* + For all the analyses, I am only considering the M2 customers. For an utilities company, free market tariff customers are more profitable than regulated ones (free market vs. electricity price fixed by government), and therefore more relevant to analyze in detail.

* + **What are its limitations i.e. what are some questions that you cannot answer with this data set?**

Two main limitations:

* The dataset only contains a subset of customers per city, therefore for any conclusions from the analysis that we might reach to, we should keep in mind that it cannot be made extensible to the full population
* We do not have any kind of personal information about the client (e.g. age, gender, socio-economic level, etc.), therefore some of the analysis will be just limited to understanding how the customer behave based on his/her energy consumption but will not be able to do profiles of customers with greater detail.
  + **What kind of cleaning and wrangling did you need to do?**

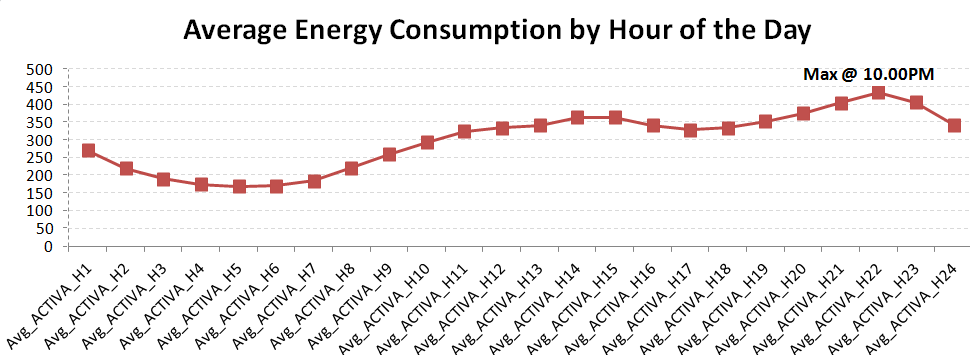
The first step that I will need to take would be to use command-line tools to select a subset of the full dataset. The dataset that I want to use for my capstone is >7.5Gb which make it impossible to load the full file into R to perform a holistic analysis.

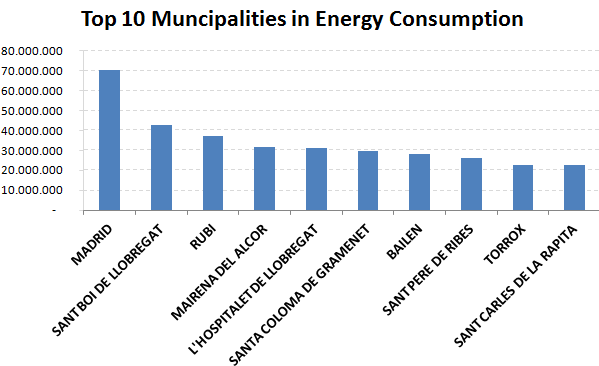
Once I have been able to parse the full dataset and select a subset of my interested I would need to remove some unnecessary columns, parse the date field to convert it R date data type and potentially derive some additional data.frame with a different configuration of fields (i.e. gather some column) to ease out plotting.

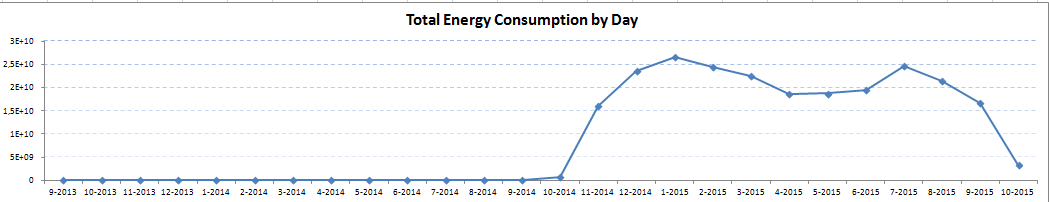
Once I start analyzing the data more in depth I would most certainly also need to identify and filter out usual outliers and empty fields

* **Any preliminary exploration you’ve performed and your initial findings.**

Before jumping into using command-line tools to extract a meaningful subset of data, I performed some basic and very high level analysis on the full dataset using a tool call Alteryx. With this tool I created the following plots:







* Up until Oct-2014, there is significant less data than for the 2014-2015 years which explain the difference in the energy consumption total. This might be due to the incremental installation of smart meters in home across the country.
* **Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?**

The approach will be very similar as the one described initially (see below) with one exception. I will need an extra initial step to parse my huge dataset and select a subset to be used for the analysis using command line tools:

0 – Use command-line tools to parse the initial dataset and extract a reduced subset that can be loaded into R

1 – Initial exploration of the dataset

2 – Data clean up and selection of a subset

3 – Online research of what are the most popular kind of analysis performed on smartmeter data

4 – Select 3-4 analysis to be perform

5 – Code and test hypothesis, re-evaluate goals if necessary

6 – Deliver the project