Get started Open in app

Jean-Michel D
150 Followers · About Follow

Smart meters in London (Part1) — Description and first insights — Jean-Michel D.

Jean-Michel D Jan 28, 2018 · 7 min read

Hello, the goal of this article is to offer a clear description of the dataset that I uploaded in November 2017 on Kaggle followed by some insights on the dataset.

Description of the dataset

To better follow the energy consumption, the government wants energy suppliers to install smart meters in every home in England, Wales and Scotland. There are more than 26 million homes for the energy suppliers to get to, with the goal of every home having a smart meter by 2020.

This roll out of meter is lead by the European Union who asked all member governments to look at smart meters as part of measures to upgrade our energy supply and tackle climate change. After an initial study, the British government decided to adopt smart meters as part of their plan to update our ageing energy system.

In this dataset, you will find a refacted version of the data from the London data store, that contains the energy consumption readings for a sample of 5,567 London Households that took part in the UK Power Networks led Low Carbon London project between November 2011 and February 2014. The data from the smart meters seems associated only to the electrical consumption.

To have an easier dataset to manipulate, different transformations have been applied on the dataset:

- Collection of all the data from a specific household in the same file (that was not the case in the original dataset)
- People from the same ACORN group are on the same file

The original and clean dataset can be find in the halfhourly_dataset zip file and one file looks like this snapshot.

LCLid	tstp	energy(kWh/hh)

	LCLid	tstp	energy(kWh/hh)
0	MAC000002	2012-10-12 00:30:00.0000000	0
1	MAC000002	2012-10-12 01:00:00.0000000	0
2	MAC000002	2012-10-12 01:30:00.0000000	0
3	MAC000002	2012-10-12 02:00:00.0000000	0
4	MAC000002	2012-10-12 02:30:00.0000000	0

As you can see the dataset is quite easy to manipulate with:

- **LCLid** that corresponds to the household id
- **tstp** the timestamp of the measure
- **energy(kWh/hh)** the energy consumes in the past 30 minutes in kWh

But to make the life easier for the user of my dataset, I created two others zip files that contains some pre-process data:

- the *daily_dataset* that contains daily informations on the consumption of the households

	LCLid	day	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
0	MAC000002	2012-10-12	0.1385	0.154304	0.886	46	0.196034	7.098	0.000
1	MAC000002	2012-10-13	0.1800	0.230979	0.933	48	0.192329	11.087	0.076
2	MAC000002	2012-10-14	0.1580	0.275479	1.085	48	0.274647	13.223	0.070
3	MAC000002	2012-10-15	0.1310	0.213688	1.164	48	0.224483	10.257	0.070
4	MAC000002	2012-10-16	0.1450	0.203521	0.991	48	0.184115	9.769	0.087

- the *hhblock_dataset* that contains the transpose data of a day for one household (as an array) with for example the hh_0 column is the consumption between 00:00 and 00:30

	LCLid	day	hh_0	hh_1	hh_2	hh_3	hh_4	hh_5	hh_6	hh_7	...	hh_38	hh_39	hh_40	hh_41	hh_42	hh_43	hh_44	hh_45	hh_46	hh_
0	MAC000002	2012-10-13	0.263	0.269	0.275	0.256	0.211	0.136	0.161	0.119	...	0.916	0.278	0.267	0.239	0.230	0.233	0.235	0.166	0.259	0.2
1	MAC000002	2012-10-14	0.262	0.166	0.226	0.088	0.126	0.082	0.123	0.083	...	1.075	0.956	0.821	0.745	0.712	0.511	0.231	0.210	0.278	0.1
2	MAC000002	2012-10-15	0.192	0.097	0.141	0.083	0.132	0.070	0.130	0.074	...	1.164	0.249	0.225	0.258	0.260	0.334	0.299	0.236	0.241	0.2
3	MAC000002	2012-10-16	0.237	0.237	0.193	0.118	0.098	0.107	0.094	0.109	...	0.966	0.172	0.192	0.226	0.203	0.211	0.168	0.213	0.157	0.2
4	MAC000002	2012-10-17	0.157	0.211	0.155	0.169	0.101	0.117	0.084	0.118	...	0.223	0.075	0.230	0.208	0.265	0.377	0.327	0.277	0.288	0.2

This is an overview of all the data from the smart meter, but to facilitate the exploration there is a table that stored all the households and their associated files (informations_households.csv).

	LCLid	stdorToU	Acorn	Acorn_grouped	file
0	MAC005492	ToU	ACORN-	ACORN-	block_0
2	MAC000002	Std	ACORN-A	Affluent	block_0
3	MAC003613	Std	ACORN-A	Affluent	block_0
4	MAC003597	Std	ACORN-A	Affluent	block_0
5	MAC003579	Std	ACORN-A	Affluent	block_0

In this table, there is:

- **LCLid** that correspond to the household id
- **stdorToU** the kind of tariff applied (*ToU* the dynamic tariff in function of the days or *Std* the classic fixed tariff)
- **Acorn** the ACORN group associated, that categorises the household
- **Acorn_grouped** this is another more global classification of the ACORN (fusion of different ACORN groups)
- **file** name of the file in the different zip files where you can find the data of the household

All these informations are from the original dataset but to complete the informations available to make other study there is an addition of some new datasets:

- *acorn_details.csv* : that contains the index for multiple parameters in comparison of the national (that have an index of 100)

	MAIN CATEGORIES	CATEGORIES	REFERENCE	ACORN-A	ACORN-B	ACORN-C	ACORN-D	ACORN-E	ACORN-F	ACORN-G	ACORN-H	ACORN-I	A
0	POPULATION	Age	Age 0-4	77.0	83.0	72.0	100.0	120.0	77.0	97.0	97.0	63.0	
1	POPULATION	Age	Age 5-17	117.0	109.0	87.0	69.0	94.0	95.0	102.0	106.0	67.0	
2	POPULATION	Age	Age 18-24	64.0	73.0	67.0	107.0	100.0	71.0	83.0	89.0	62.0	
3	POPULATION	Age	Age 25-34	52.0	63.0	62.0	197.0	151.0	66.0	90.0	88.0	63.0	
4	POPULATION	Age	Age 35-49	102.0	105.0	91.0	124.0	118.0	93.0	102.0	103.0	76.0	

- *uk_bank_holidays.csv* : the bank holidays for the period of the study

	Bank holidays	Type
0	2012-12-26	Boxing Day
1	2012-12-25	Christmas Day
2	2012-08-27	Summer bank holiday
3	2012-05-06	Queen?s Diamond Jubilee (extra bank holiday)
4	2012-04-06	Spring bank holiday (substitute day)

- *weather_daily_darksky.csv* : the daily informations on the weather from darksky at London during the study

	temperatureMax	temperatureMaxTime	windBearing	icon	dewPoint	temperatureMinTime	cloudCover	windSpeed	pressure	apparentTemperatureMinTim
0	11.96	2011-11-11 23:00:00	123	fog	9.40	2011-11-11 07:00:00	0.79	3.88	1016.08	2011-11-11 07:00:0
1	8.59	2011-12-11 14:00:00	198	partly-cloudy-day	4.49	2011-12-11 01:00:00	0.56	3.94	1007.71	2011-12-11 02:00:0
2	10.33	2011-12-27 02:00:00	225	partly-cloudy-day	5.47	2011-12-27 23:00:00	0.85	3.54	1032.76	2011-12-27 22:00:0
3	8.07	2011-12-02 23:00:00	232	wind	3.69	2011-12-02 07:00:00	0.32	3.00	1012.12	2011-12-02 07:00:0

4	8.22	2011-12-24 23:00:00	252	partly-cloudy-night	2.79	2011-12-24 07:00:00	0.37	4.46	1028.17	2011-12-24 07:00:0

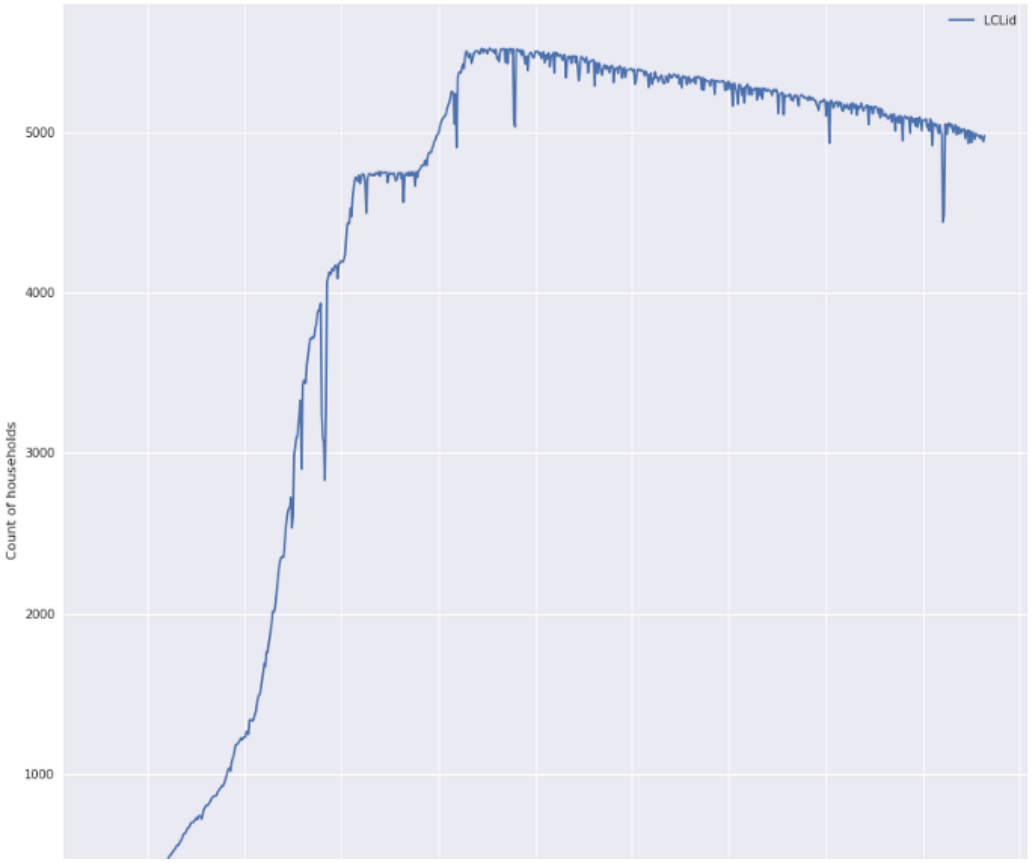
- *weather_hourly_darksky.csv* : the hourly informations on the weather from darksky at London during the study

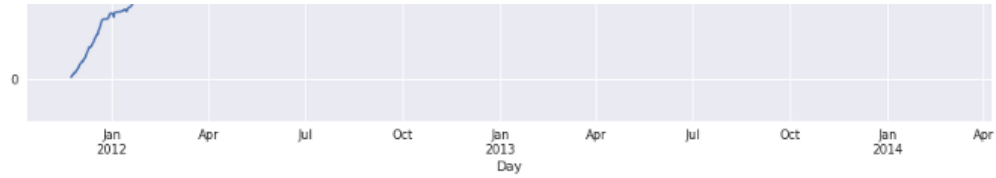
	visibility	windBearing	temperature	time	dewPoint	pressure	apparentTemperature	windSpeed	precipType	icon	humidity	su
0	5.97	104	10.24	2011-11-11 00:00:00	8.86	1016.76	10.24	2.77	rain	partly-cloudy-night	0.91	
1	4.88	99	9.76	2011-11-11 01:00:00	8.83	1016.63	8.24	2.95	rain	partly-cloudy-night	0.94	
2	3.70	98	9.46	2011-11-11 02:00:00	8.79	1016.36	7.76	3.17	rain	partly-cloudy-night	0.96	
3	3.12	99	9.23	2011-11-11 03:00:00	8.63	1016.28	7.44	3.25	rain	fog	0.96	
4	1.85	111	9.26	2011-11-11 04:00:00	9.21	1015.98	7.24	3.70	rain	fog	1.00	

This first part offers a general overview of the content of the dataset, it's time now to obtain a clearer vision on the data from the smart meter.

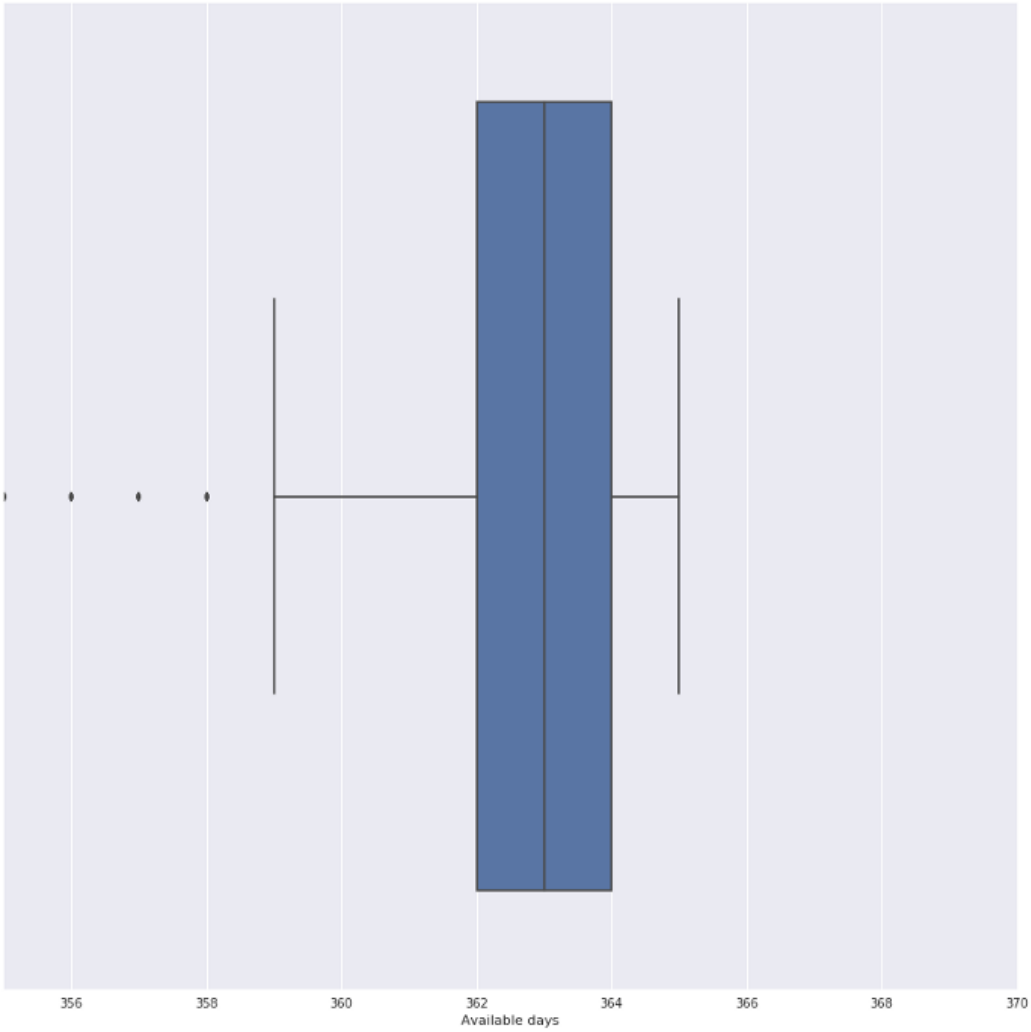
Exploration of the dataset

First step on this study is to find the best period to make the comparison. In my previous article on the electrical consumption in France there was a seasonal effect so a great period to study will be at least one year of data. In the next figure there is an illustration of the count of households with data (the 48 timestamps in the day) per day of the study.





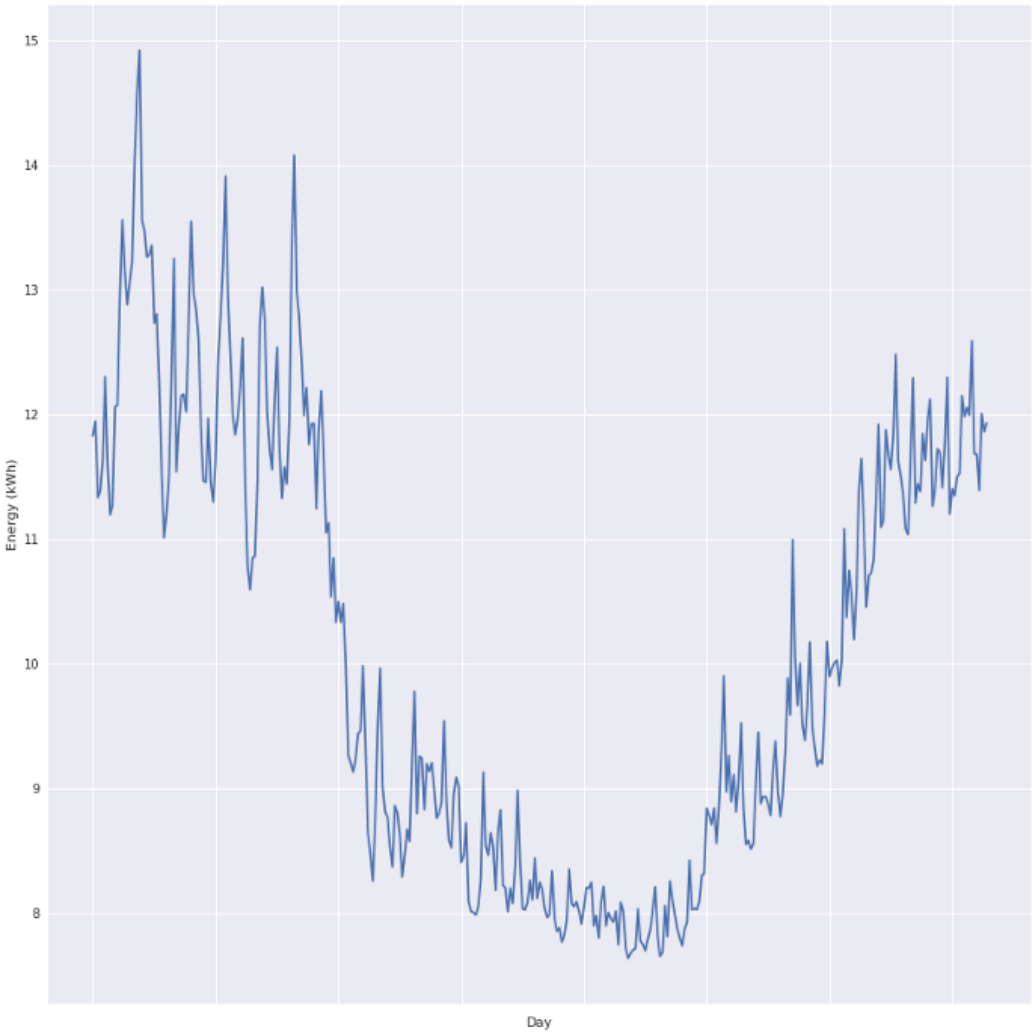
Notes: There is clearly an increase of the number of available households since the start of the study in end 2011, the peak is reach in 2013. A good period for our study could be 2013 (and i choose this one). But it's now important to know the distribution of the available days for this period in the households of the experiment, in the following figure there is a representation of this distribution in a boxplot.



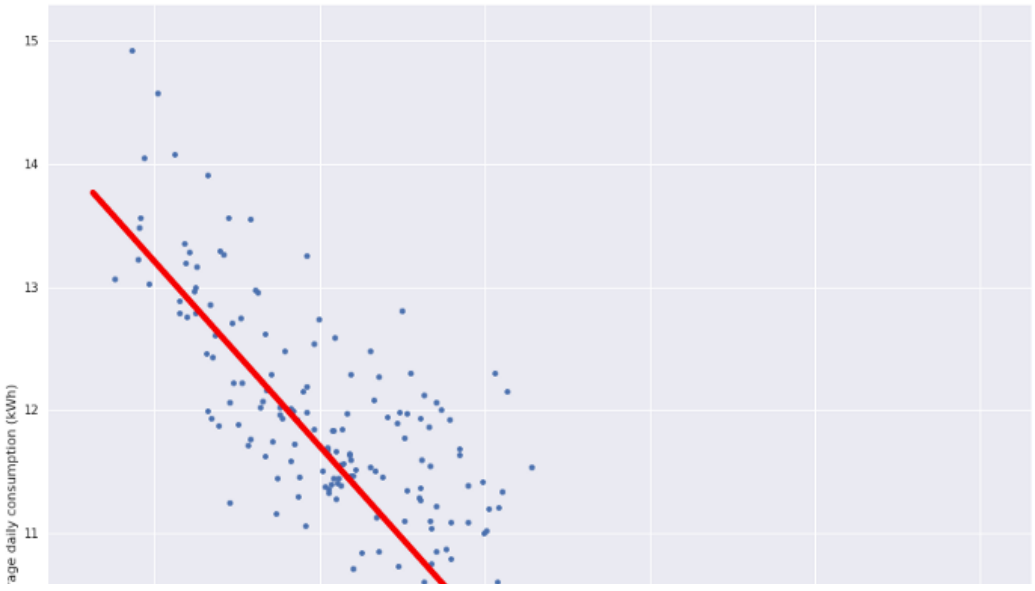
The decision has been done to use the Households that possess at least 357 days, so on the original dataset that represents a **lost of 38 households** on the 5566 available in the dataset that's totally acceptable (less than 1%).

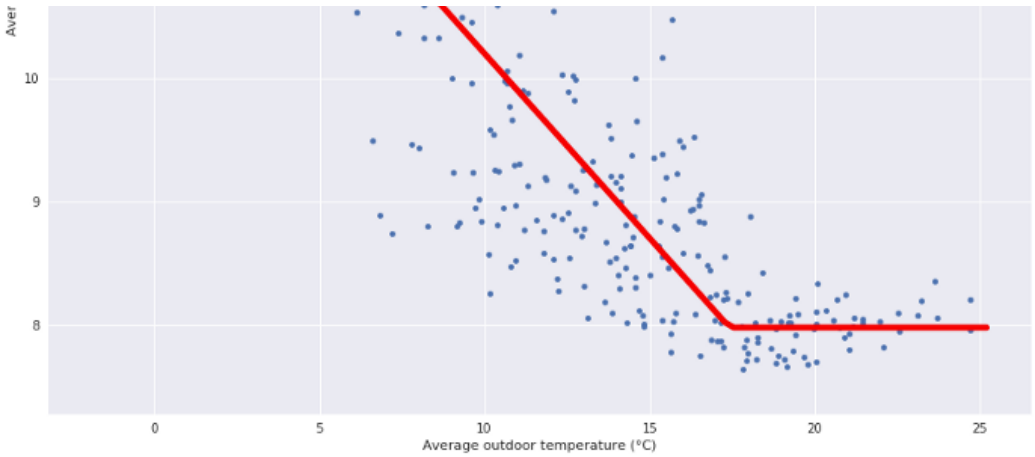
Overview of the panel

One of the first realisation is to display the average consumption per day of these households during the year of 2013, in the following figure there is the average global consumption of these households during the period.

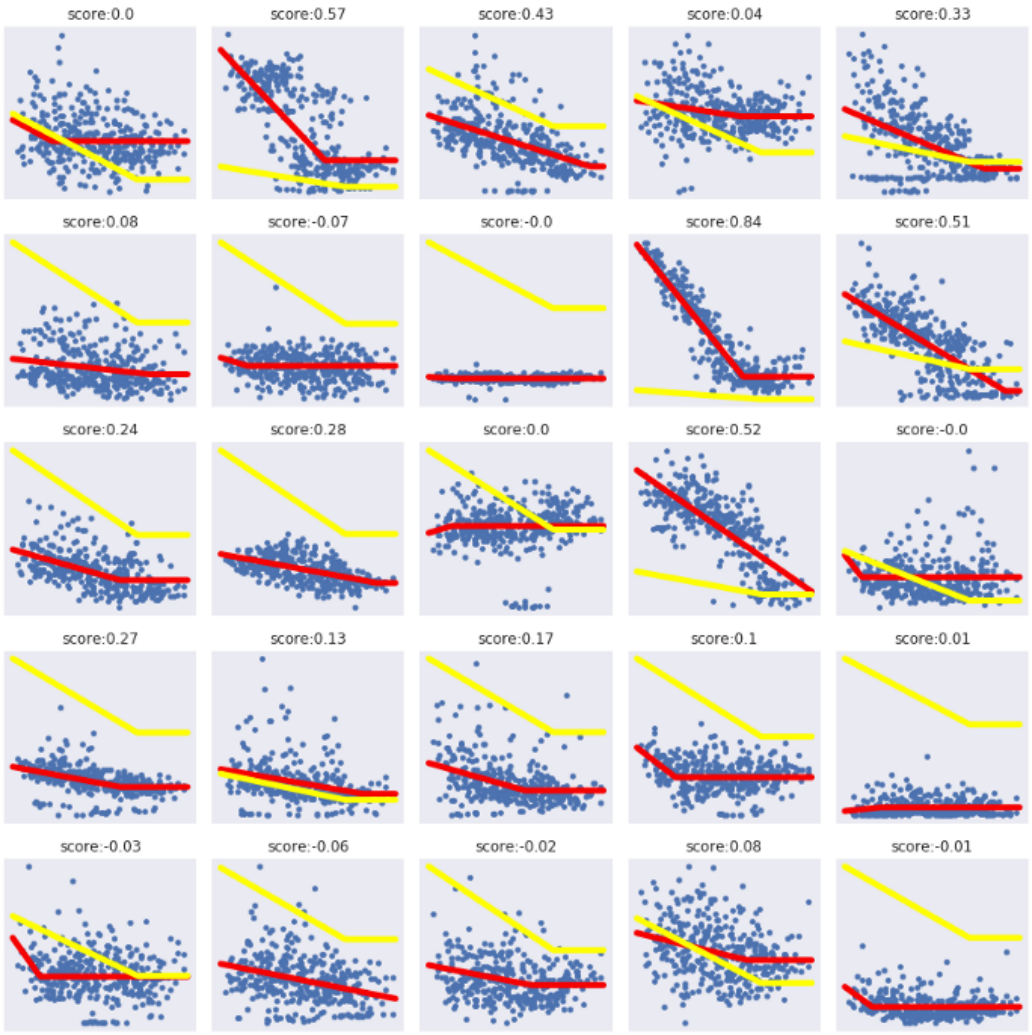


Notes: This is obvious that there is a link between the electrical consumption and the day of the year (same result than in my previous article). The seasonal effect is very clear so in this panel there is a lot of people that are using the electricity as an heating source. If the average daily outdoor temperature and the total daily consumption of the panel are crossed, the following figure display the relation:



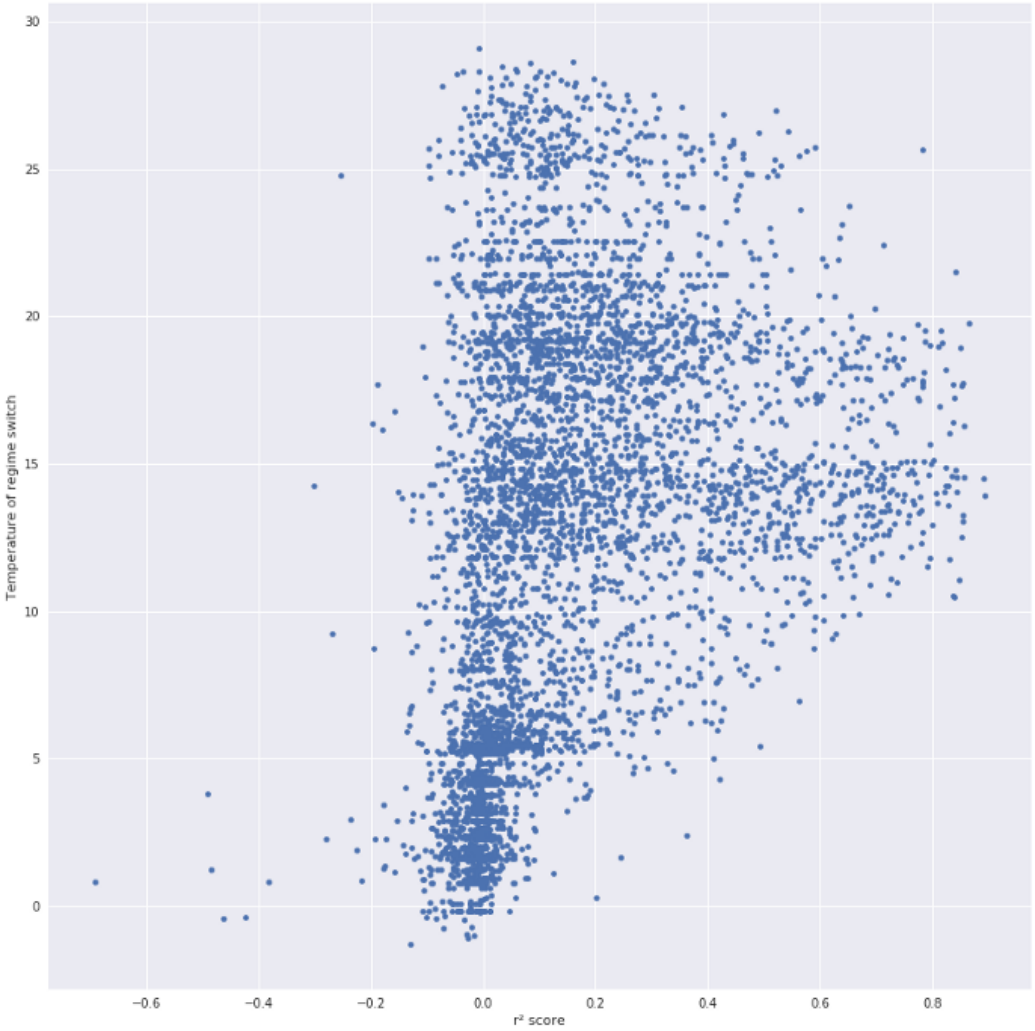


This general observation offers a clear vision that the PTG (The red plot) from the previous article can ba calculated for each household.In the following figure, there is a representation of the daily consumption and the ptg associated to this household (and theri r^2 score).



Notes: This is a good illustration that for some households the r^2 score is working great (this household should have a electric system) but for some households it doesn't work at all. The general model issued from the average daily consumption (the yellow curve)

illustrates that average daily consumption doesn't represent the general behaviour of the households.In the folowing figure there is the scatter plot of the pivot temperature in function of the r^2 score.



Another way to identify the households that have an electric heating system could be to compare the average consumption during the winter and the summer and make a simple ratio between these two consumptions.The data have been crossed with the informations of the households, and there is an extract of the new dataset.

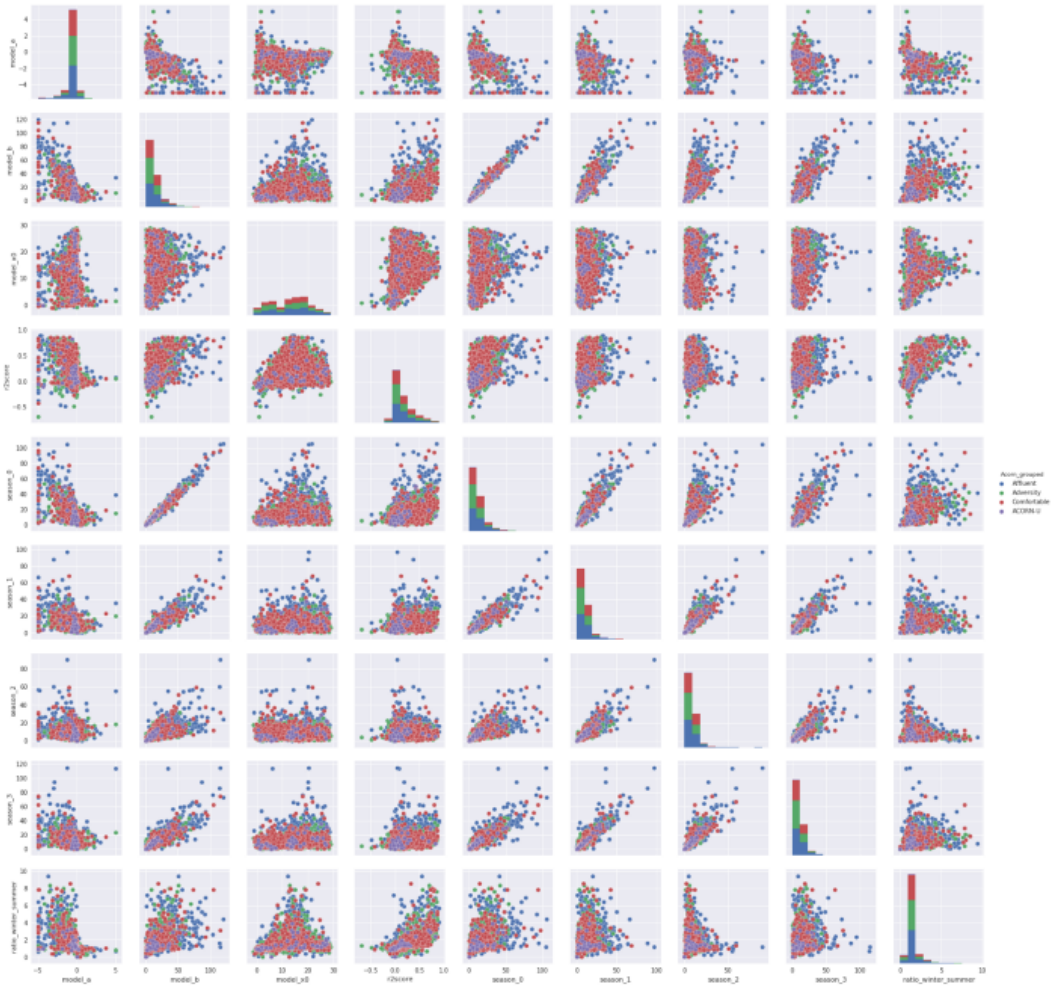
	model_a	model_b	model_x0	r2score	season_0	season_1	season_2	season_3	ratio_winter_summer	stdorToU	Acorn	Acorn_grouped
MAC000002	-0.249494	14.315588	19.976597	0.134495	13.677956	10.605000	9.082032	13.002841	1.051921	Std	ACORN-A	Affluent
MAC000003	-2.590097	40.789026	11.222564	0.315249	33.283000	17.459778	11.807021	14.316886	2.324737	Std	ACORN-P	Adversity
MAC000004	-0.018653	1.884939	20.000000	0.099579	1.886944	1.727923	1.475000	1.695057	1.113204	Std	ACORN-E	Affluent
MAC000005	-0.134637	5.904149	17.596062	0.243049	5.593722	3.852077	3.713745	5.002614	1.118160	ToU	ACORN-C	Affluent
MAC000006	-0.079921	3.722867	18.980000	0.094902	3.298278	2.868098	2.257532	3.158591	1.044224	Std	ACORN-Q	Adversity

In this dataframe, there is:

- **model_a** the slope of the ptg model (in the winter part)
- **model_b** the intersection of the ptg models

- **model_x0** the temperature of regime switch
- **r2score** the r^2 score of the ptg model on the household
- **season_0** the average consumption in winter
- **season_1** the average consumption in spring
- **season_2** the average consumption in summer
- **season_3** the average consumption in autumn
- **ratio_winter_summer** the ratio of the consumption winter/ration_winter_summer
- **stdorToU** the type of tariff
- **Acorn** the ACORN group
- **Acorn_grouped** tha aggregated ACORN groups

There is a serious amount of data to cross so in the following figure there is a pairplot that cross all this data andfilter thme in function of the Acorn_grouped.



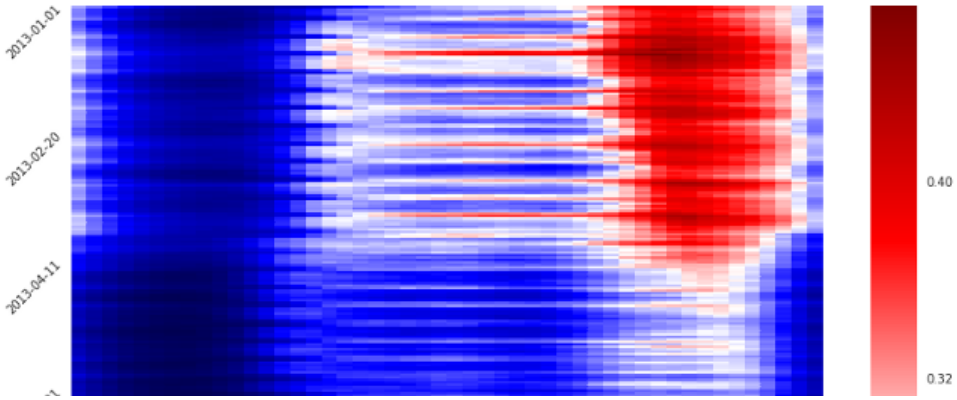
Notes: There is no obvious relation between all this index that defined the households except between the season_0 and the model_b but this two are winter related so that's

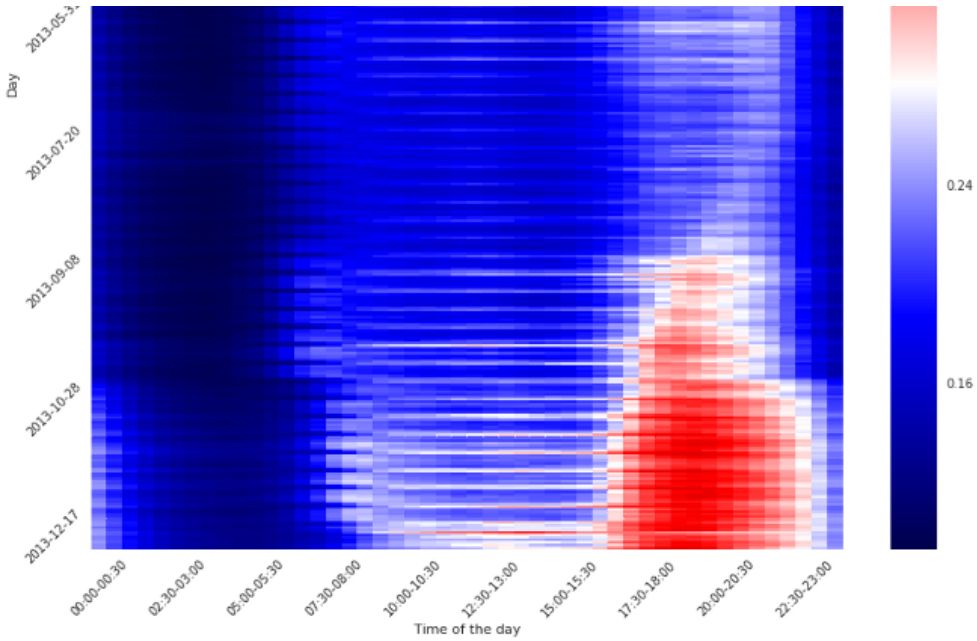
normal.But there is no link between these indexes and the Acorn_grouped, the result is similar with the Acorn, that's a little bit disappointing.

Next steps

As you can see this first exploration of the dataset has highlighted some characteristics of the electrical consumption in London like the influence of the weather in this consumption but there is a lot more things to do on this dataset. Some ideas for future analytics:

- Cross the ACORN data and the smart meter data
- Try to forecast the consumption of the different households
- Add new datasets like:
- EPC data from London
- extra data on London like some underground or train strike during the period
- Make some clusterings in the households data and the energy profiles, as you can see in the following heatmap there is a "pattern" in the total consumption of these households.





You can find all the code to make this article in this GitHub repo

Originally published at the-odd-dataguy.com on January 28, 2018.

Energy Python London Kaggle Data Science

About Help Legal

Get the Medium app

Download on the App Store GET IT ON Google Play