In-Depth Exploratory Data Analysis (EDA)

Analysis of residential energy consumption in London

Camila Malagón Suárez

Andrés Camilo Viloria García

Yesid Rivera

Oscar Nieto Garzón

Eduardo González Fierro

Didier Santander

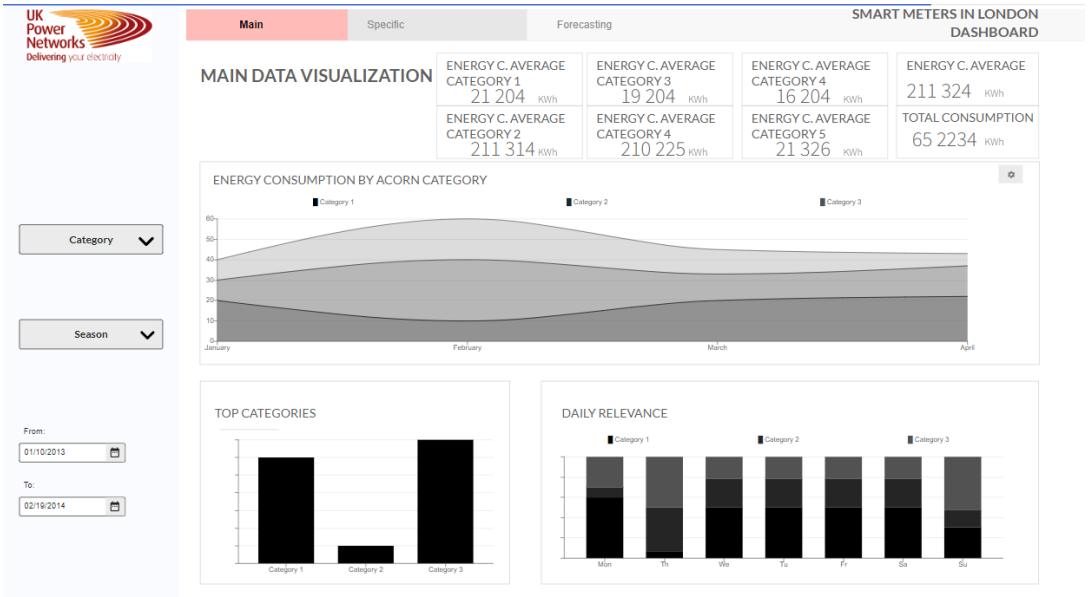
Hollman Baez

Front - End

Mockup

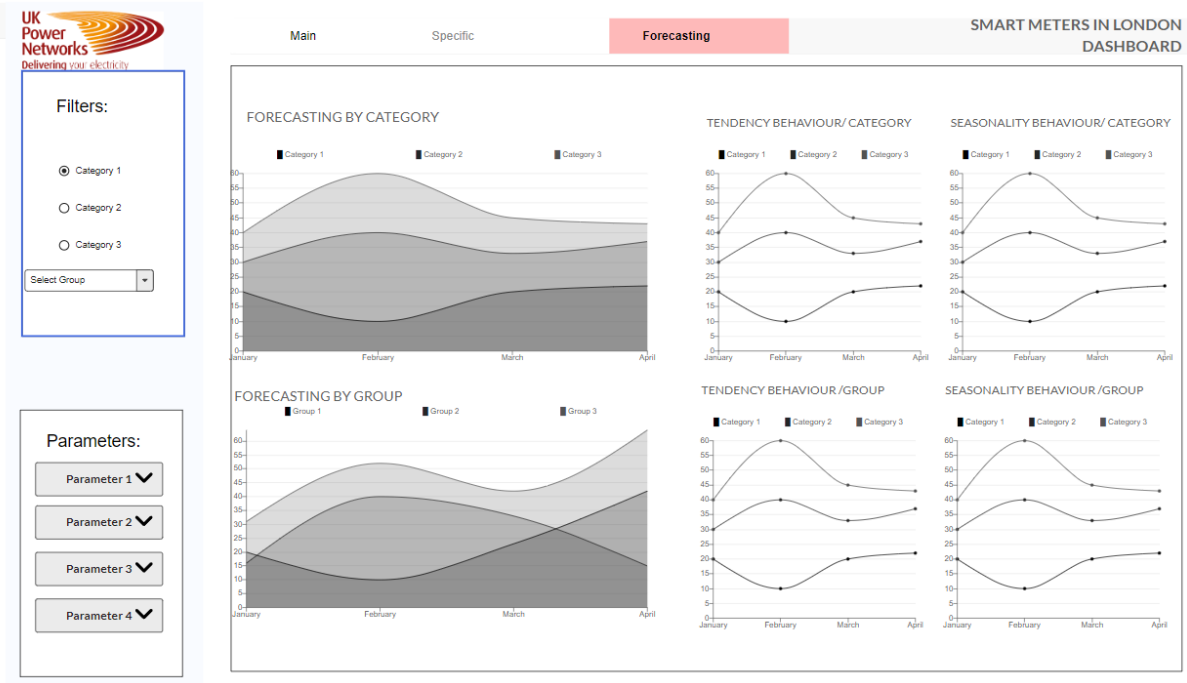
The visualization will be done in DASH with own and external components. The board will be composed of 3 parts:

1. General Visualization: General and comparative trends of the trend of energy consumption over time. This display will allow you to filter by categories and groups, as well as a specific date range. It will show a graph of the total consumption by category and also the proportion of consumption per day by category. Also the numerical average is shown by category together with the total and the sum of the energy consumption of the filters made on the board.



2. Specific Visualization: The detail of the energy consumption trend by group can be visualized and compared in the trend graph. In addition, a group can be selected in a unique way to review its trend over time, it is also possible to see the daily consumption by selected group in the filters section.

3. Forecast: The results of the model are shown in the prediction of the following months. In addition, by category and group, the behavior of the trend and seasonality over time is also shown.



Cleaning strategies

As a first step before starting the exploration and analysis of data, it is needed to perform a review for searching null or duplicated values, and in general issues that could cause a miss understanding of the reality, therefor a table with this overview is shown as follows:

Table	Size (rows)	Issue	Column	Quantity	Action
acorn_details	826	-	-	-	-
daily_dataset	3,510,433	Null values	energy_median	30	Data with less than 48 records by day were dismissed
		Null values	energy_mean	30	
		Null values	energy_max	30	
		Null values	energy_std	11,331	
		Null values	energy_sum	30	
		Null values	energy_min	30	
hhblock	3,469,352	Null values	hh_19	2	Imputed with the average for that household for that half-hour for the same week (weekdays)
		Null values	hh_25	21	
		Null values	hh_26	2	
		Null values	hh_30	5,460	
		Null values	hh_36	1	
households	5,566	Classification	acorn	49	49 rows with group U (value = ACORN-U) were assigned to group R, and 2 rows with group 'ACORN-' were dismissed
holidays	25	-	-	-	-
weather_daily	882	Null values	cloud_cover	1	Imputed with the value of the previous day
		Null values	uv_index	1	Imputed with the value of the previous day
		Null values	uv_index_time	1	The same day
weather_hourly	21,165	Null values	pressure	13	Imputed with the average of pressure for that day (all of the hours)

One of the biggest findings is relative to the daily dataset with more than 11 thousand null values. Going deep into this information we can see there are some days with troubles for the smart-meter recording as 2014-02-28, where there is just one record for the 48 half-hours for every household for all of the households in the study, that is the explanation for 4,987 records with null standard deviation, as well as other days have less than 48 counts (half-hours) and this is a problem because we can have for one day one household with 24 records, but it is not possible to know what of the 48 half-hours we are looking at (ie. morning hours vs night hours). The total number of days by households with less than 48 records is 41,081 (1.18% of the whole data set), then those data will be dismissed.

For 'hhblock' table the issues are focused on half-hour-30 for 5,460 records (week day) of the total of 3.5 millions, then, that null values will be completed using the average of the same half-hour-30 for that household using the consumption of the week of the missing day (avoiding a change of conditions like season).

For households there are 2 issues, one of them is households without group (acorn-) where there is no way to find the right group, then, those records will be dismissed too. And there are 49 rows with group U, those will be recategorized as 'acorn-R' for the new category 'Not Private Households'.

For the weather information there is a little group of null values which will be filled with data from the same day of the missing value or the day before (avoiding again a significant change of the climate conditions).

The review gives a 0 number of duplicate data for all the tables, then, this is not an issue to deal with.

Analysis per dataset

Daily dataset

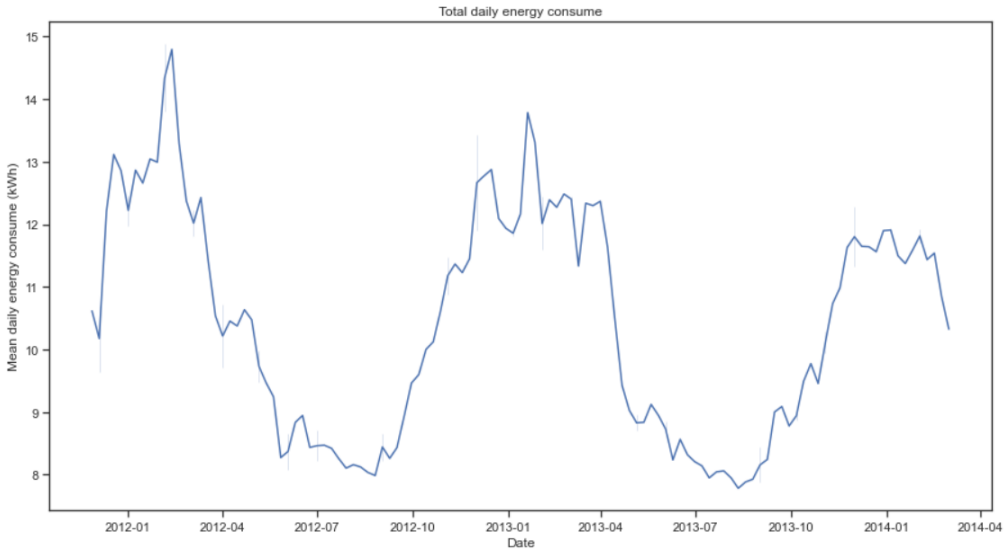
This dataset contains the records that were collected for the Smart meters. We have the daily consumption, which was obtained by adding up the 48 records each half hourly while the 24 hours of the day. Also, the dataset has date, and ID as categorical variables and median, mean, máximum, mínimum and standard deviation of the records.

3469352 were the rows that we're gonna work on after cleaning data. One row is data for each column, we have 17 columns.

Descriptive statistics for consumption of energy

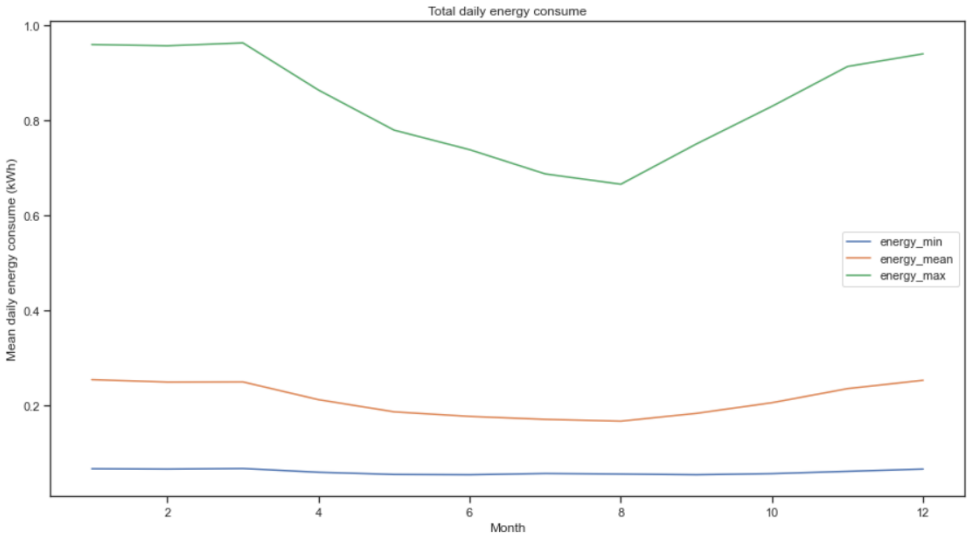
	energy_count	energy_min	energy_sum	energy_max
count	3469352.0	3469352.00	3469352.00	3469352.00
mean	48.0	0.06	10.16	0.84
std	0.0	0.08	9.13	0.67
min	48.0	0.00	0.00	0.00
25%	48.0	0.02	4.72	0.35
50%	48.0	0.04	7.84	0.69
75%	48.0	0.07	12.60	1.13
max	48.0	6.39	332.56	10.76

Consumption of energy by date



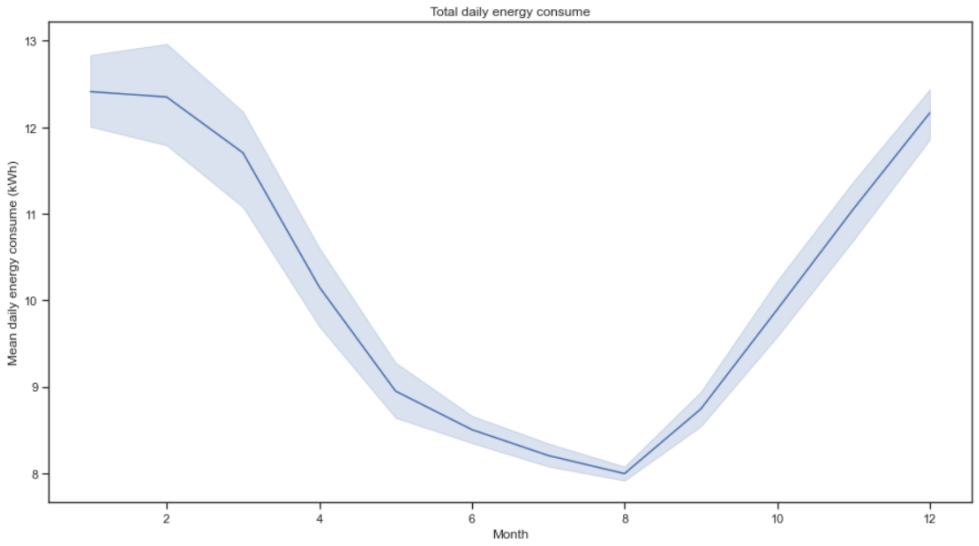
In this plot we can see that the demand of energy has a seasonal behaviour, but at this moment we can't affirm anything.

Maximum, minimum, and mean consumption of energy by month



This plot shows us the min, max and mean consumption of energy. We see that the minimum is 0, it's say every hour there are places where they don't use energy. The maximum consumption was close to 1 kw/h. and the mean is 0.3 kw/h approximately.

Mean consumption of energy by month during 2011, 2012, 2013.



There, we can see the months and this consumption. The behaviour shown in the plot is that between January and March the demand is high and between April and August the consumption decreases until the minimum. Then, between September and December the consumption increases.

Half-hourly dataset

The dataset that contains half-hourly data has more than 160 million entries, with 3 columns. The numeric column with the half-hourly energy consumption per household contains 5558 missing values in total, of which a major part corresponds to observations registered on a non-standard frequency of observation, i.e. 15:13:37 instead of 15:30:00. Also, this table contains 5566 unique households ids, matching the total number of households reported in other datasets.

The mean of the same half-hour of the corresponding week of the day with the missing value was computed to impute these values. This mean was used to fill in the missing values. Then, we aggregated the data into an hourly dataset, aggregating the recorded value of the same hour of the corresponding date and household. From this process, we obtain a new dataset with almost half the entries, around 80 million without missing values.

We can see that the number of observations was reduced almost to half and that the mean is almost double the half-hourly data set mean. Also, there is a great difference between the 75th percentile value and the maximum value in the two time scales.

Table 1. Basic statistics of half-hourly and hourly datasets

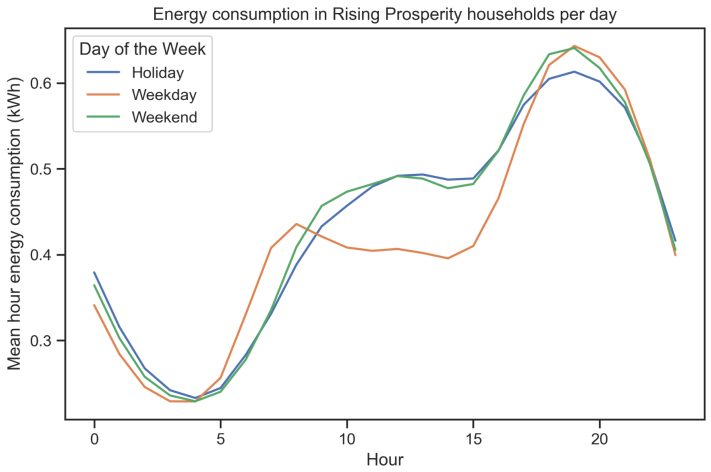
	energy (kWh)		energy (kWh)
count	167774203.000	**count**	83263824.000000
mean	0.212	**mean**	0.423531
std	0.297	**std**	0.561182
min	0.000	**min**	0.000000
25%	0.058	**25%**	0.122000
50%	0.117	**50%**	0.244000
75%	0.239	**75%**	0.491000
max	10.761	**max**	20.199000

This new hourly dataset was used to perform the Exploratory Data Analysis on different time scales.

Analysis for day of the week

First, the analysis per day of the week was done, considering weekdays, weekends, and holidays across the entire period of study. The dates of holidays were extracted from an additional dataset that listed the UK bank holidays, between 2011 and 2014.

Mean hourly energy consumption by day of the week



WD	Holiday	Weekday	Weekend
count	1921944.000	57630504.000	23711376.000
mean	0.434	0.418	0.437
std	0.585	0.555	0.574
min	0.000	0.000	0.000
25%	0.119	0.121	0.124
50%	0.242	0.241	0.250
75%	0.502	0.483	0.509
max	13.903	20.199	17.218

The hourly consumption pattern shows that there is a peak of energy consumption, between 18:00 and 20:00 hours, in the evening. Also, it seems that holidays and weekends have a similar consumption profile, with an increase in the values in the morning and mid-day. In contrast, on weekdays household energy consumption tends to rise earlier in the morning but maintains a stable value until late afternoon.

Also, the number of observations shows that a major part of the dataset corresponds to weekday values, with more than half of the total observations. The mean of each three groups of days is similar but on weekdays and weekends, the maximum value is larger than on holidays.

Additionally, to test that there is a meaningful difference in the hourly energy consumption between weekdays and weekends t-test is used, quantifying the difference between their means. In this case, the significance is defined as α = 0.05.

Test 1: the output of the t-test between the hourly energy consumption per household between weekdays and weekends is presented with the corresponding statistics.

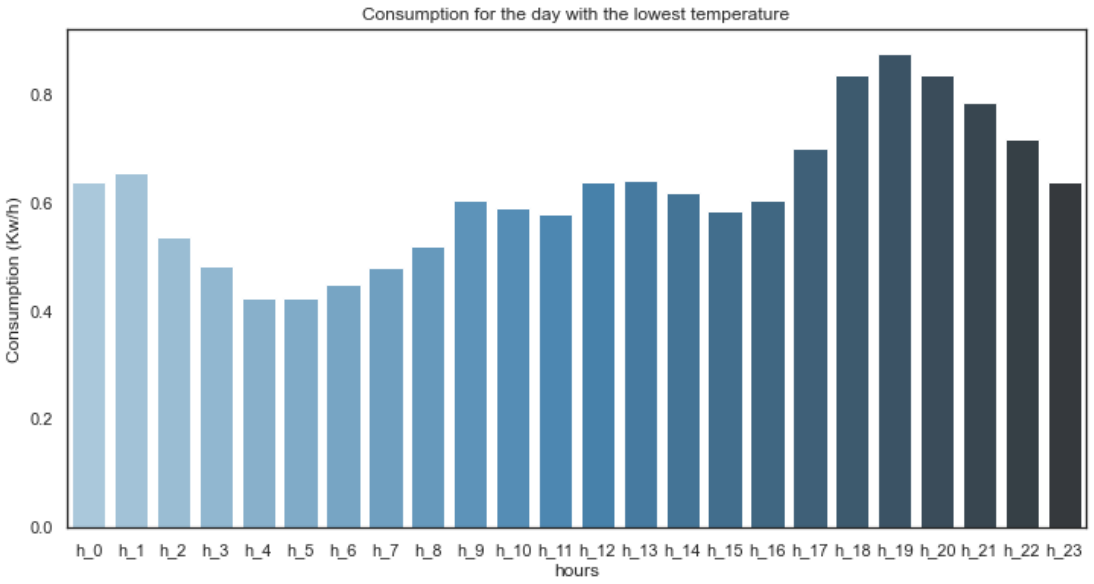
	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-140.84	4.282155e+07	two-sided	0.0	[-0.02, -0.02]	0.035	inf	1.0

We can see that the p-value (0.0) < α (0.05), so the null hypothesis can be rejected, concluding that there is a significant difference in the energy consumption per day of the week and their means in the studied households.

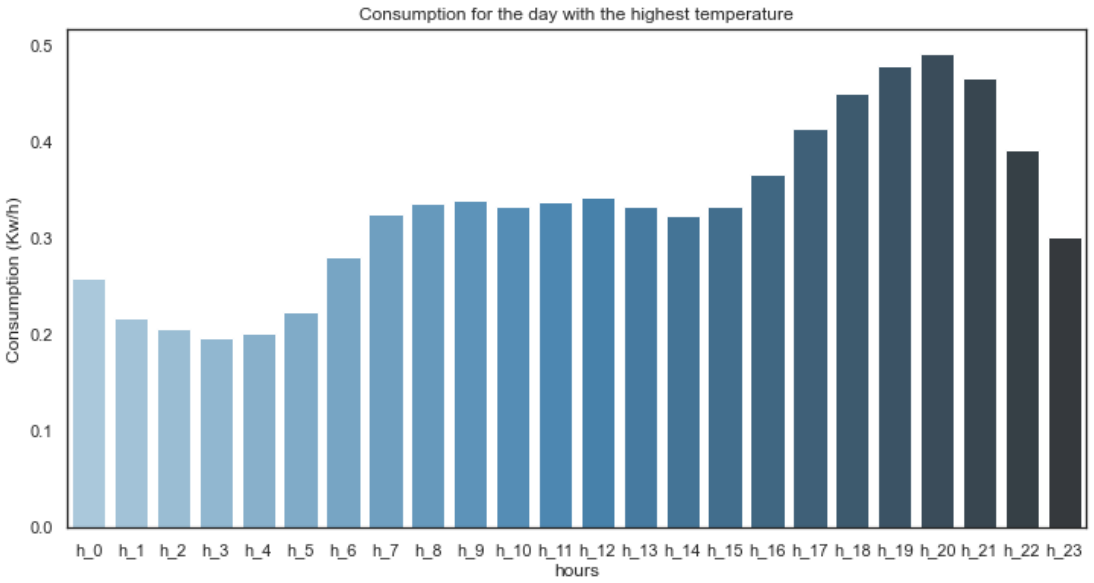
Analysis for hour of the day

In the line of describing the consumption per hour of the day, one relevant variable to include is the temperature, then, to see the possible pattern changes an analysis of two scenarios will be performed as follows:

Scenario 1: The coldest day for the period of the study has been selected on 2012-02-11, and no matter the category or group, the next chart shows the consumption for that day, broken down per hour:



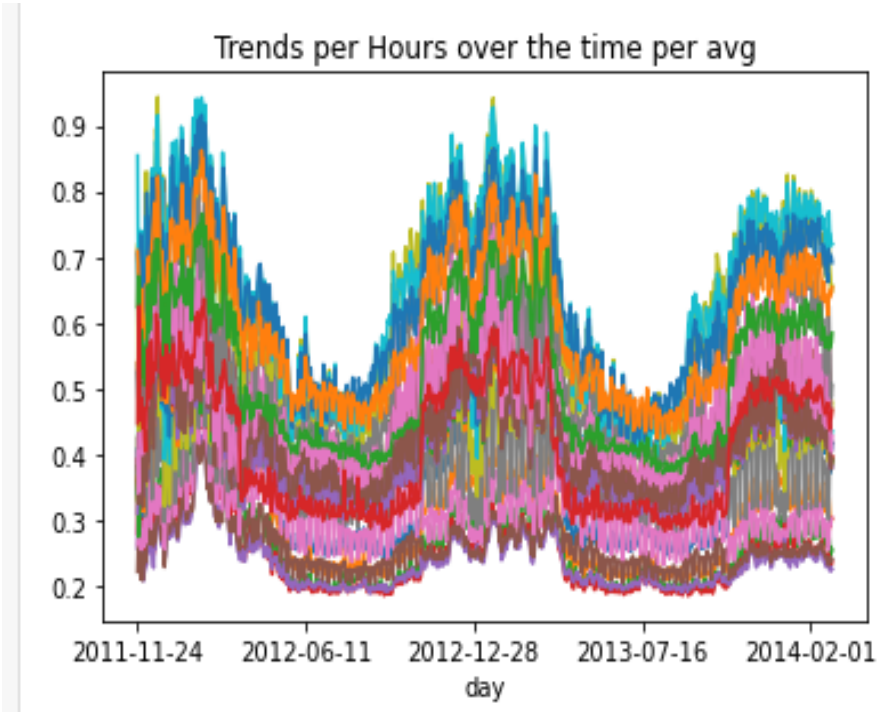
Scenario 2: Now the hottest day has been selected on 2013-07-31, and the same structure for information is shown below:



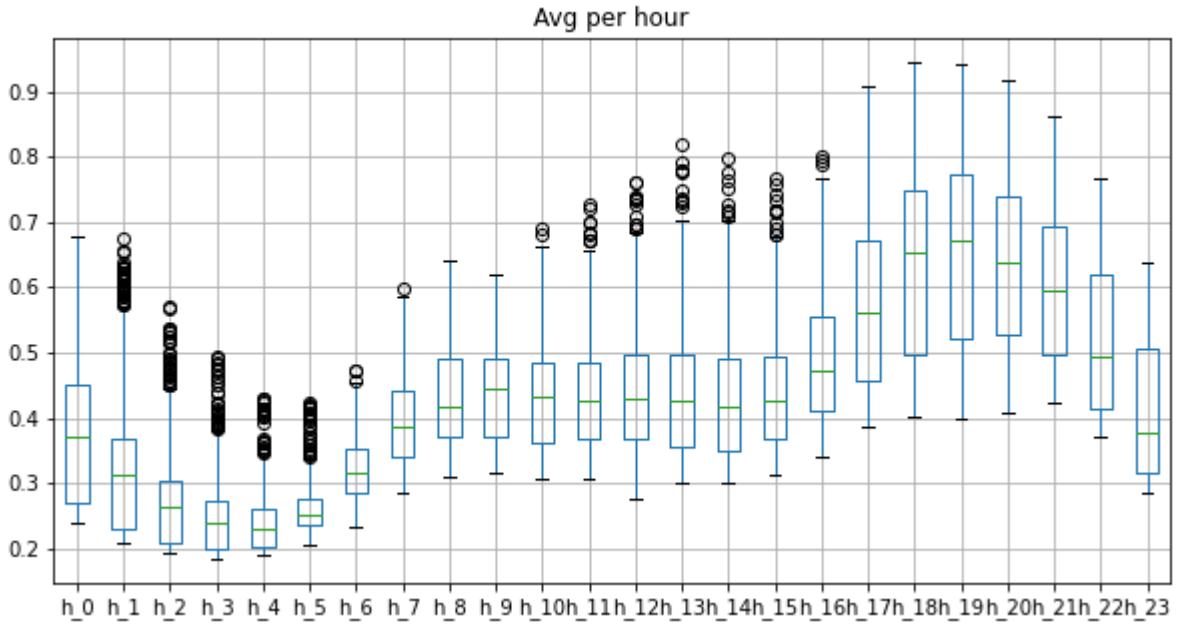
It is illustrative to see the general consumption with valley hours at the first hours of the morning (before 6am, light bars) and then a constant rising to get the highest values at the beginning of the night (around 8pm, dark bars); but for the coldest day the gap in lower and the minimum consumption is recorded around 4am to 5am with an average value of 0.4 Kw/h, as well as there is no huge reduction between 8pm and 1am; meanwhile the hottest day has a higher variation through the day, with the lowest point at 1am to 5am with a consumption a little greater than 0.2 Kw/h (2 times longer and 50% lower versus the coldest day). In addition, there is a flat range between 7am and 3pm (a significant part of the day) and the trend after 8pm is for a fast decrease to get the minimum again.

The maximum value for the coldest day is greater than 0.8 Kw/h while for the hottest day it is lower than 0.5 Kw/h, which tells many things about the patterns of consumption in function of temperature, over the hour patterns even.

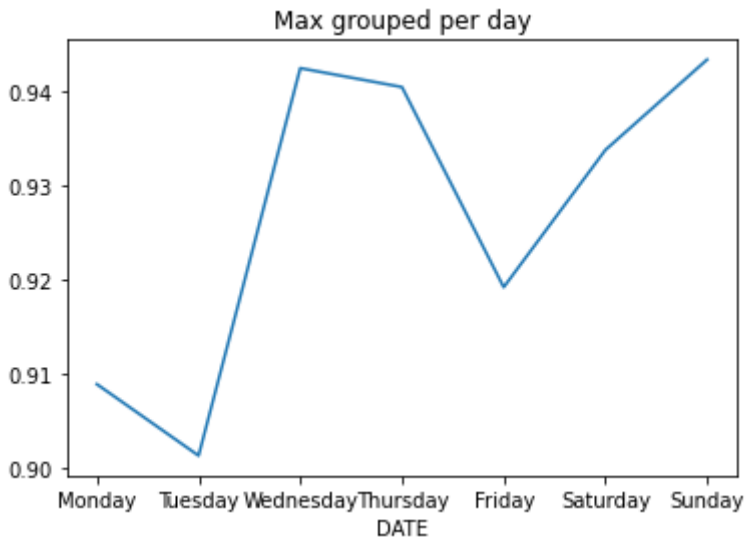
Checking the trends per hour and throughout the years, we realise a possible pattern for hours on the chart below:



There could be an explanation since this could be correlated to the stations and days of the week. Let's check the average consumption per hour.

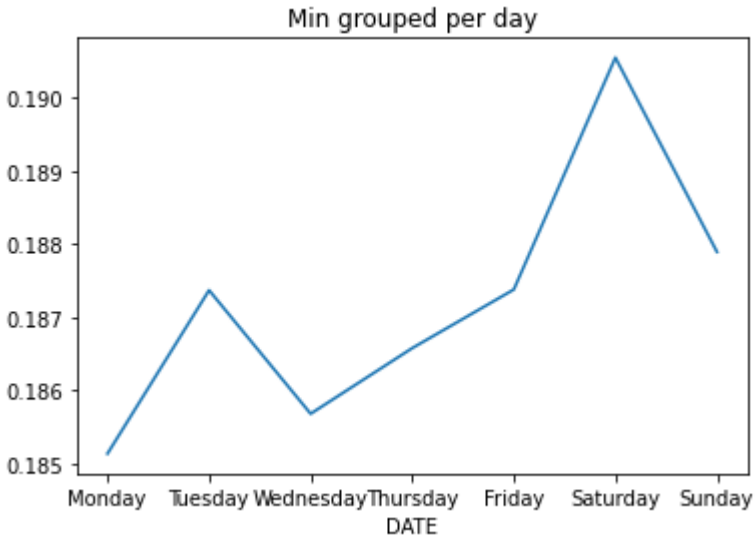


This graph shows the average consumption starts increasing from 17h to 22h, this will be the peak hour for us. When it comes to checking the day with the most consumption, it is explained for this graph, Wednesday, Thursday, Sunday.

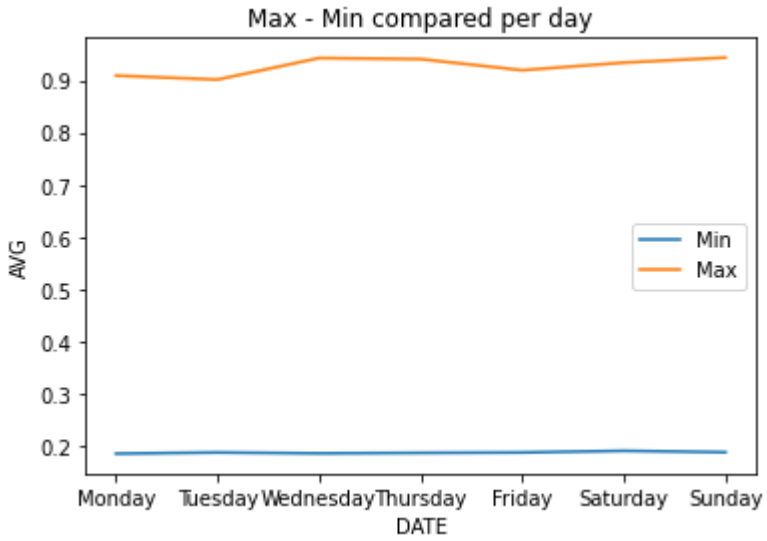


Now, there won't be a 0 demand for energy, so, checking the minimum consumption for the people, businesses and the company, it will look like this, stable and almost perfectly aligned.

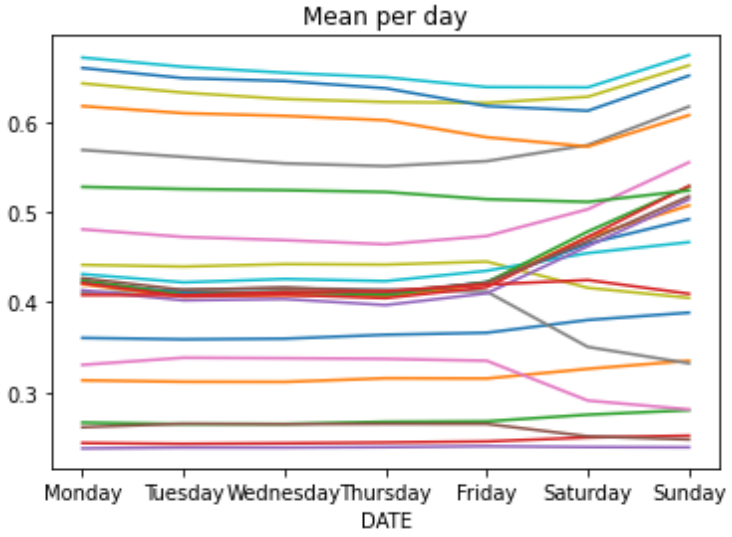
Let's check the minimum consumption:



This will be interpreted together to get a hint on how the demand will be covered when having peaks and valleys, ups and downs



To finish, we will plot the mean of the days to check if the weekends increases the consumption:

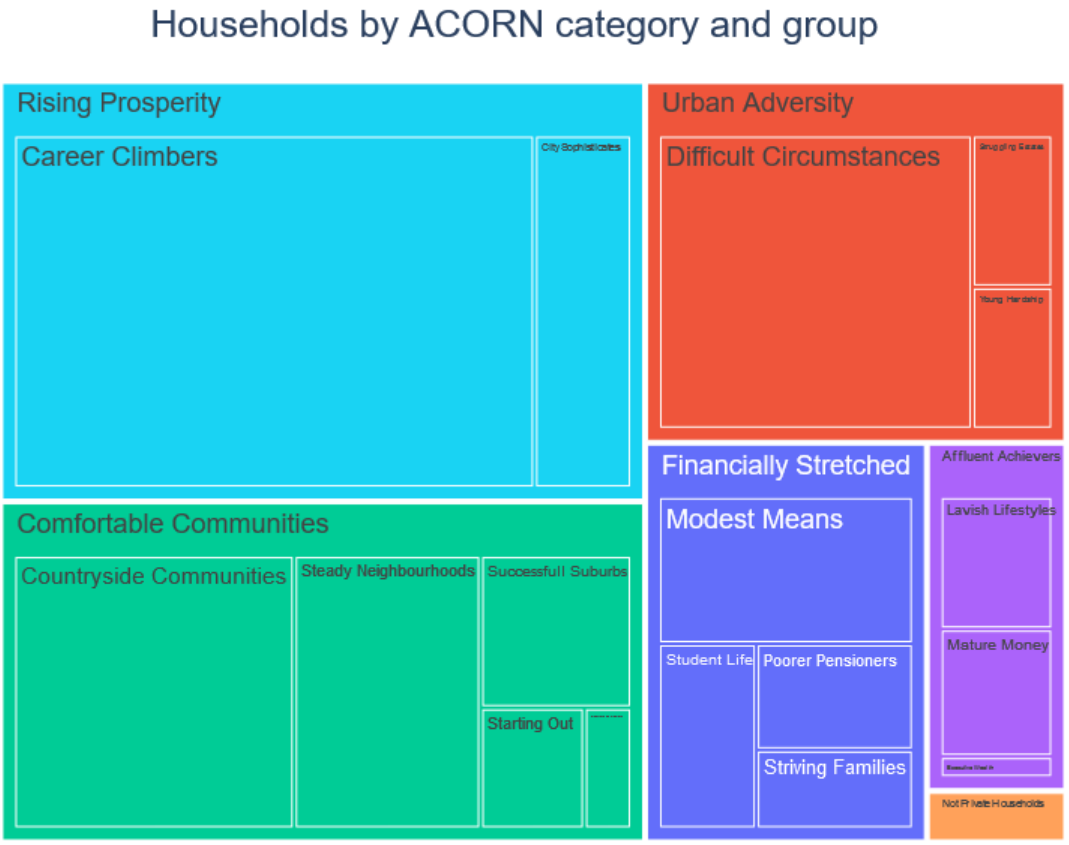


Takeaways so far, the consumption increases on weekends, and weekdays starting 16 to 22 hours. This will give us a hint on how to forecast the consumption and the days when we need to back up and guarantee the service.

Analysis by Group

The households table contains the ACORN category and group in which each household was categorized. ACORN is a consumer classification based on demographic data, social factors, population, and consumer behavior, that segments UK postcodes into 6 main categories and 17 groups. Each category is composed of one or several groups.

The dataset contains a total of 5556 different households, of which there are households that belong to the six ACORN categories and the 17 groups. On the following graph, the number of households by category and group is presented.



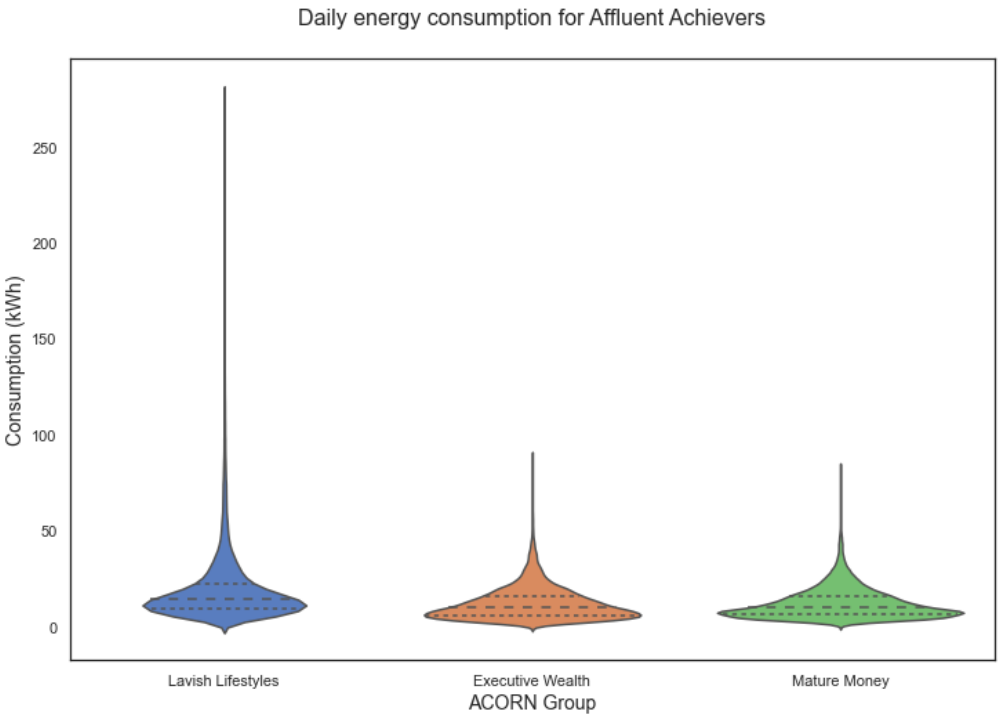
Analysis by Category

Category 1: Affluent Achievers

The descriptive statistics for this category let think about a low number of households, there are less than 200 thousand records (48 measures for each day for each household) as well as a high energy consumption because the maximum of the *energy_sum* near to 280 Kw/h for a single day for a single household, too far from the average (18 times higher) which suggests the presence of outliers but it has to be analysed in depth later to see if it is not caused for one of the groups within the category that could have a much higher mean but few households, so, few records.

	energy_count	energy_sum	energy_min	energy_max	energy_median	energy_mean	energy_std
count	194,747.00	194,747.00	194,747.00	194,747.00	194,747.00	194,747.00	194,747.00
mean	48.00	15.36	0.10	1.09	0.25	0.32	0.23
std	0.00	13.68	0.15	0.74	0.27	0.28	0.17
min	48.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	48.00	7.31	0.04	0.56	0.11	0.15	0.11
50%	48.00	11.82	0.07	0.97	0.17	0.25	0.19
75%	48.00	18.56	0.11	1.45	0.29	0.39	0.31
max	48.00	277.97	5.05	9.14	5.52	5.79	2.56

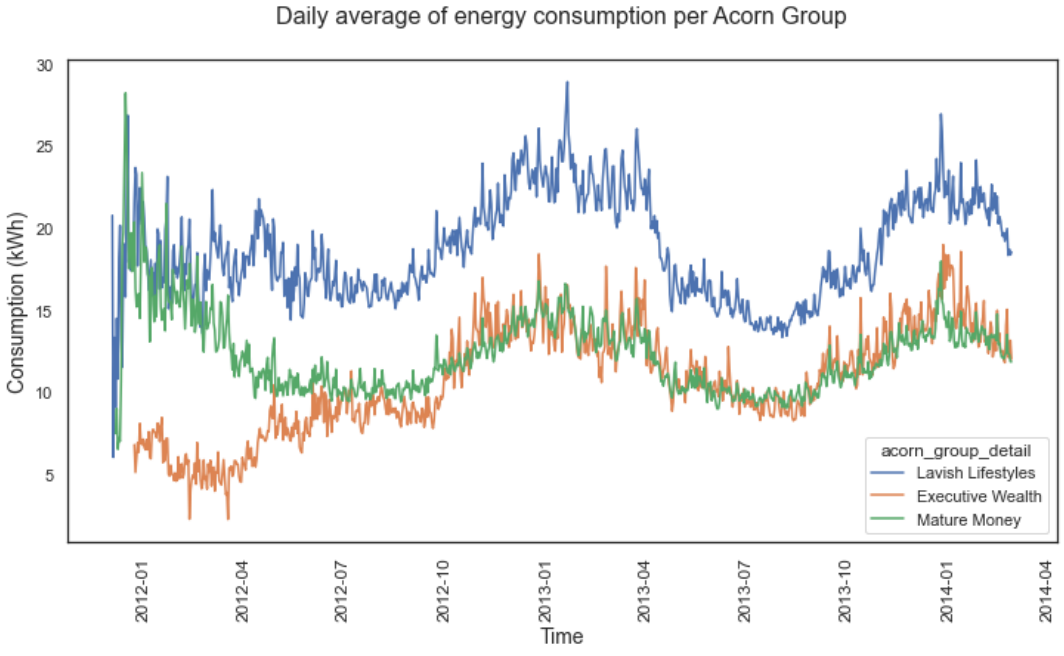
In order to determining the behaviour of each group, the next chart shows a violin plot:



The plot shows a similar consume pattern for the households of each group, it is possible to see the outlier introduced for the previous table and it belongs to the wealthier group, not only within the category but of the entire ACORN classification, then it is interesting to look for outliers like this one in other groups.

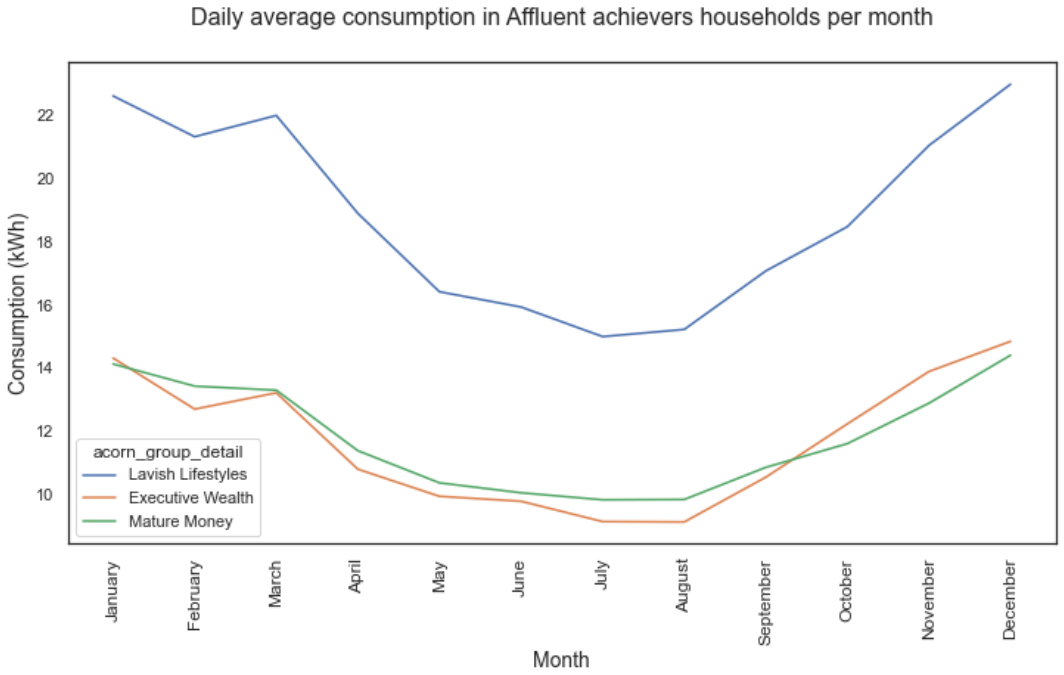
An interesting observation is that for the 3 groups most of the records are below the mean, making the distribution skewed to the left, one possible explanation are the records of houses for vacations periods and people out from home with 0 consumption for the whole day (travelling maybe, reasonable due to the profile of the population of this category), so here rises the need for a means contrast with a subset of the groups.

To understand the behaviour of the consumption across the time, a line plot is shown next:

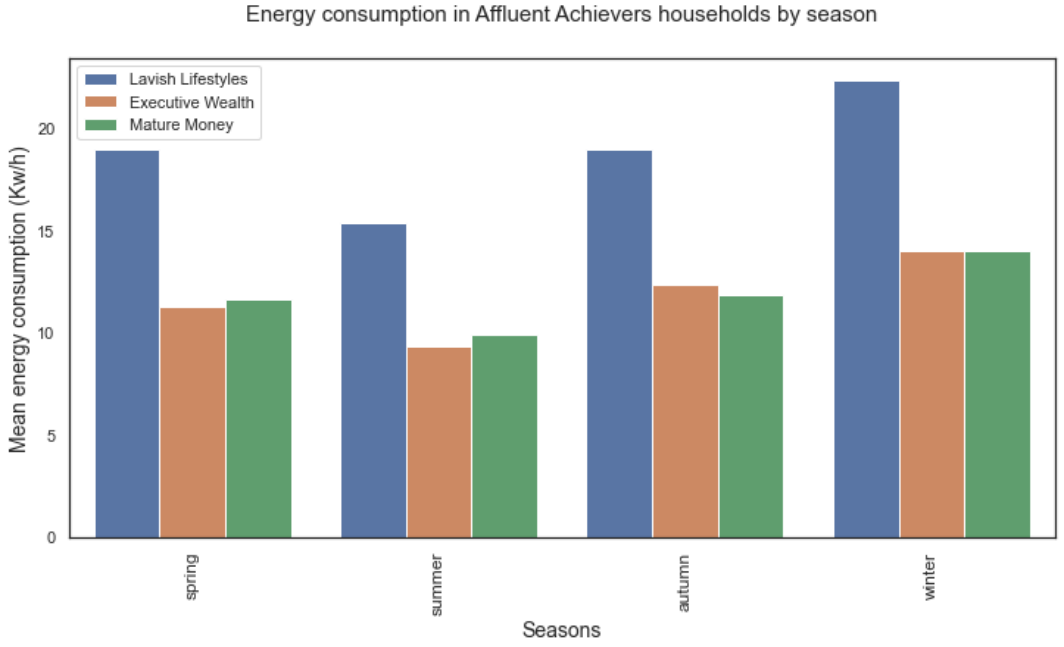


There is a clear pattern in general for all the groups, and it is a higher consumption for the end of each the year and the beginning of the next one, as well as a lower consumption for the months of the middle of the year, therefore that pattern must be analysed in according to the average per month and the seasons of the year. Also, there is some strange noise at the beginning of the period with a large amount of volatility that needs to be seen further later, maybe just the first e months of the available data.

To continue with the behaviour across the time, next there is the calculation of the average daily consumption for each one of the months of the year, separate by group to see possible different trends:



This chart confirms the rising trend of consumption from September to February, and the decreasing trend from March to August. Although the there is a clear difference between the groups, it is possible to see two of the groups: *Executive Wealth* and *Mature Money* with a similar consumption across the year, but the *Lavish Lifestyles* group shows a much higher consume for every month, about 2 times higher; then a mean contrast for consumption between groups is a good approach to consider.



The previous chart shows the consumption segmented for group and season, and there is a clear rising trend beginning in the summer, growing in autumn and reaching the highest point for the winter, no matter the group, the trend is clear, and once again there is a remarkable difference between *Lavish Lifestyles* and the other two groups.

In order to compare the means for the groups of the category, a t-test will be performance as follows:

Test 1: A subset of the data frame with the consumption per group is taken just with the *energy_sum* (total consumption per household per day) and the group, *Lavish Lifestyles* this case, and the same is performed with the group *Executive Wealth*, then the t-test is applied, and the result is the next:

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	78.35604	35418.602133	two-sided	0.0	[6.94, 7.3]	0.439063	inf	1.0

Interesting to see the p-value, which is 0.0, it means, with a 0.05 significance level that the null hypothesis must be rejected in favour of the alternative hypothesis, then it can be concluded that there is statistical difference between the average consumption of the *Lavish Lifestyles*group and the *Executive Wealth* group.

Test 2: The same procedure is conducted again but now to compare *Lavish Lifestyles* and Mature *Money* groups:

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	113.222307	134330.177636	two-sided	0.0	[6.96, 7.21]	0.524337	inf	1.0

Once again, the result is to reject the null hypothesis, it is notable to see the confidence interval of the two tests and how close they are. The conclusion is as expected, there is a statistical difference between the average consumption of the *Lavish Lifestyles* group and the *Mature Money* group.

Test 3: Finally, the same method is used to the *Executive Wealth* and *Mature Money* groups:

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-0.482	18825.470415	two-sided	0.629812	[-0.19, 0.11]	0.004492	0.011	0.07891

In this case, with the same 0.05 significance level, the p-value is greater by far than this, the confidence interval includes negative and positive values, therefore it is impossible to reject the null hypothesis and a statistical difference between the means of those groups cannot be affirmed.

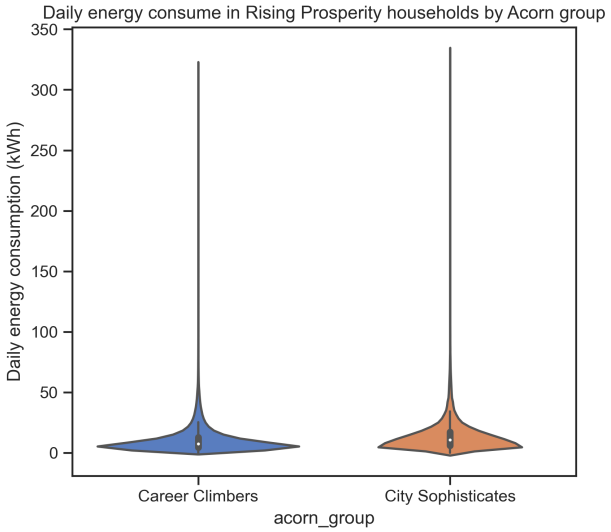
Category 2: Rising Prosperity

The descriptive statistics for the numerical data show that the columns on the daily dataset include many statistics generated from the half-hourly dataset. In this case, the data of the Rising Prosperity category have around 1 million entries, with 7 numerical columns.

	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
count	1198287.000	1198287.000	1198287.000	1198287.000	1198287.000	1198287.000	1198287.000
mean	0.167	0.227	0.882	48.000	0.187	10.896	0.063
std	0.198	0.223	0.737	0.000	0.177	10.704	0.099
min	0.000	0.000	0.000	48.000	0.000	0.000	0.000
25%	0.061	0.094	0.339	48.000	0.067	4.504	0.019
50%	0.109	0.162	0.716	48.000	0.140	7.770	0.038
75%	0.195	0.280	1.205	48.000	0.250	13.446	0.072
max	6.905	6.928	10.761	48.000	3.347	332.556	6.394

We can see that there are no negative values on the data, with the minimum value of energy consumption being 0. Also, within this category, the statistics have been computed only with days with 48 half hours, so there are no missing or incomplete values.

The maximum values of all the energy-related columns are far from the 75th percentiles, so surely there will be outliers. In this case, we see that there are some 0 values on the energy_sum and energy_min column that can be associated with periods of time where a specific household didn't consume energy across an entire day.

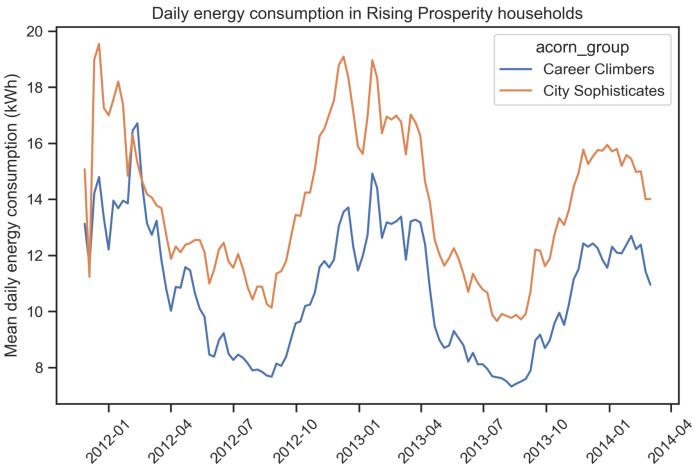


The violin plots of the main feature (energy_sum) by group confirm that in both cases we have outliers with very large values. However, the main portion of the values shows a concentration of the observations below the mean, which can be associated with a distribution skewed to the left. This type of distribution seems to be possible because a great number of observations describe lower values, near or almost 0.

We can visualise the trend of a household's energy consumption across different time scales.

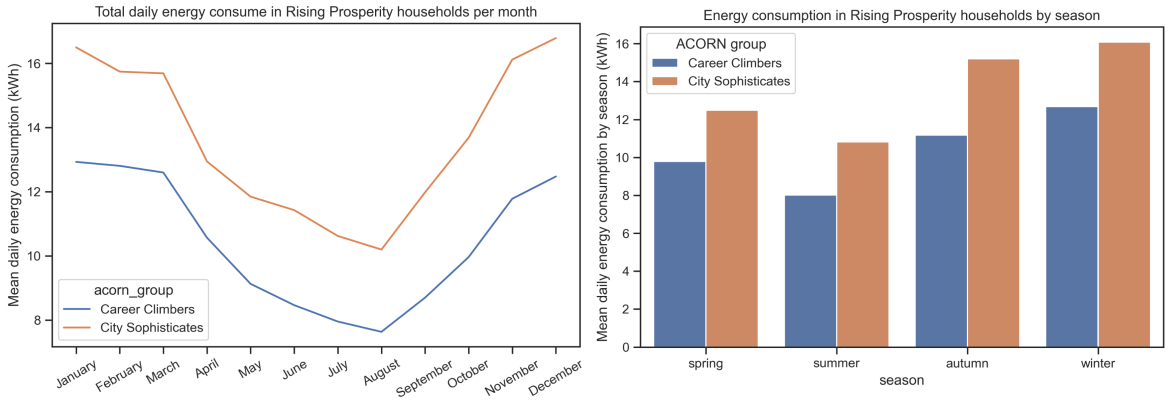
Mean daily energy consumption in Rising Prosperity households per day

First, the daily energy consumption by each group is calculated for the entire period.



Mean daily energy consumption in Rising Prosperity households per month

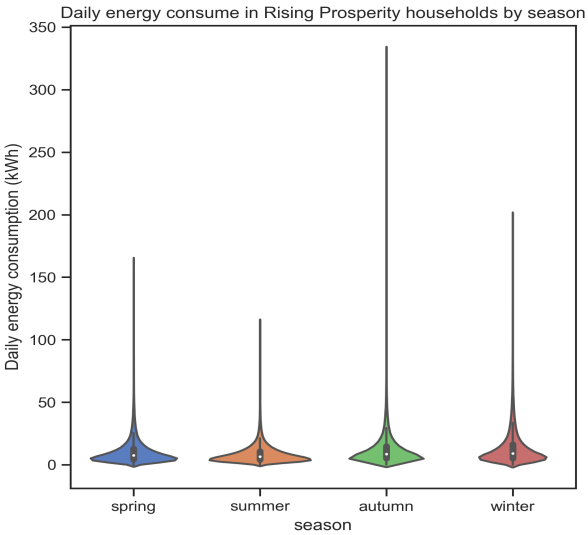
Then, the mean monthly and seasonal changes by group are calculated and plotted on the following graphs.



We can see that daily energy consumption varies with specific monthly changes within each year. In the first and last months of the year, the mean daily consumption of both groups is larger than in the rest of the months. It seems that the City Sophisticates group tends to have higher energy consumption than the Career Climbers.

The mean daily energy consumption by season shows that the described behaviour can be related to the changes within each year through the seasons.

Violin plot of daily energy consumption per season



acorn_group	season	count	mean	std	min	25%	50%	75%	max
Career Climbers	spring	216349.0	9.791778	9.216837	0.0	4.1820	7.086	12.019	164.000999
	summer	276350.0	8.006632	6.457934	0.0	3.8060	6.278	10.211	115.191999
	autumn	264944.0	11.177642	10.847992	0.0	4.6790	8.000	13.773	321.696998
	winter	253764.0	12.676896	12.858808	0.0	4.9040	8.583	15.492	199.818000
City Sophisticates	spring	39581.0	12.490218	10.498010	0.0	5.5210	9.894	16.013	153.242001
	summer	51791.0	10.810215	8.672987	0.0	5.0865	8.694	13.889	101.597000
	autumn	49111.0	15.199174	14.953183	0.0	6.5570	11.789	18.929	332.556001
	winter	46397.0	16.078485	14.205625	0.0	6.7510	12.395	20.401	176.024000

Daily mean energy consumption also tends to vary across the different seasons, with a higher mean magnitude in Autumn and Winter in both groups. The violin plots show that the maximum daily energy consumption was in Autumn but in Winter the overall percentile values are the highest.

Finally, to test that there is a meaningful difference in the daily energy consumption between the two groups it is used a statistical test to quantify the difference between their arithmetic means. In this case, the t-test allows us to perform this comparison, having a null hypothesis that their means are equal and an $\alpha = 0.05$.

Test 1: the output of the t-test between the subset of the total energy consumption per household per day for each group is presented with the corresponding statistics.

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-105.14587	235035.958842	two-sided	0.0	[-3.3, -3.18]	0.304194	inf	1.0

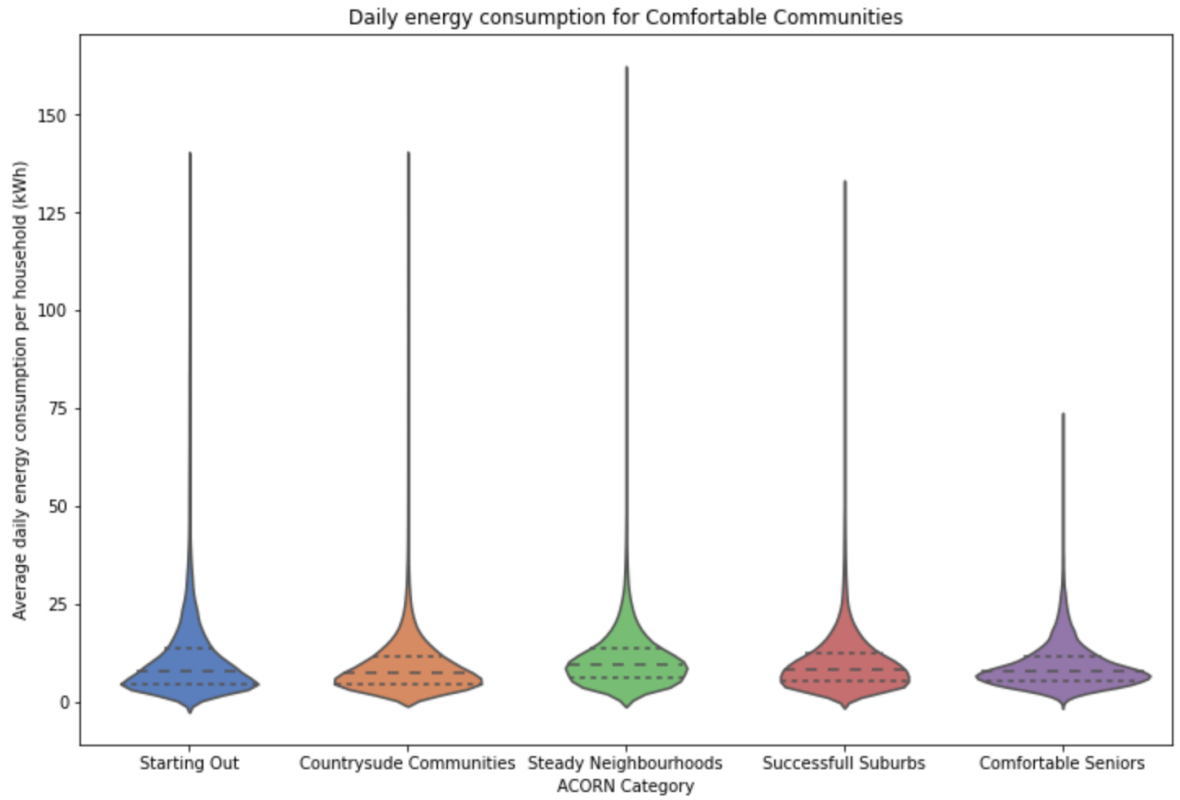
We can see that the p-value (0.0) < α (0.05), so the null hypothesis can be rejected, concluding that there is a significant difference between the group's total energy consumption per day and their means.

Category 3: Comfortable Communities Households

The descriptive statistics for the Comfortable Communities ACORN category show that this category has almost 920 thousand observations, with 7 numerical columns. This category is composed of 5 groups, and the average daily energy consumption is 10.04 kWh.

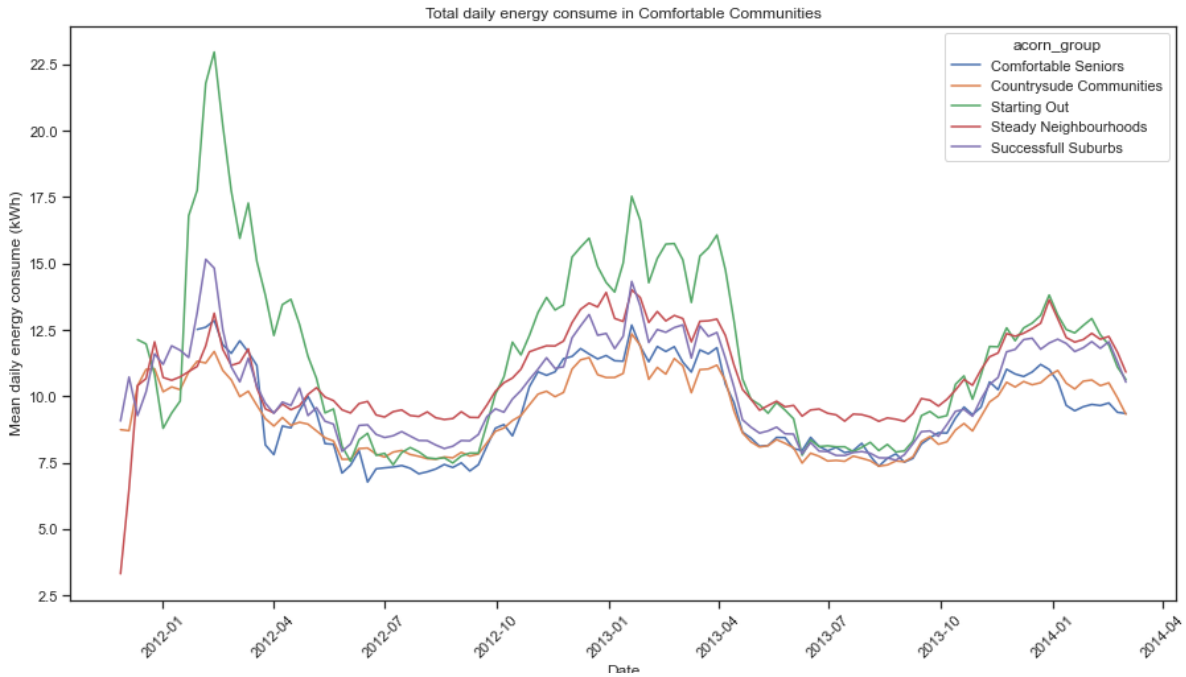
	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
count	926337.000	926337.000	926337.000	926337.000	926337.000	926337.000	926337.000
mean	0.158	0.209	0.827	48.000	0.168	10.042	0.059
std	0.148	0.164	0.616	0.000	0.135	7.850	0.074
min	0.000	0.000	0.000	48.000	0.000	0.000	0.000
25%	0.074	0.108	0.378	48.000	0.075	5.177	0.022
50%	0.123	0.174	0.709	48.000	0.137	8.339	0.042
75%	0.197	0.264	1.108	48.000	0.224	12.650	0.072
max	3.437	3.358	9.257	48.000	2.067	161.177	3.004

To understand and see possible trends and differences between the groups of this category, the next figure shows a violin plot with the consumption per group and also a tree map showing the categories under analysis:



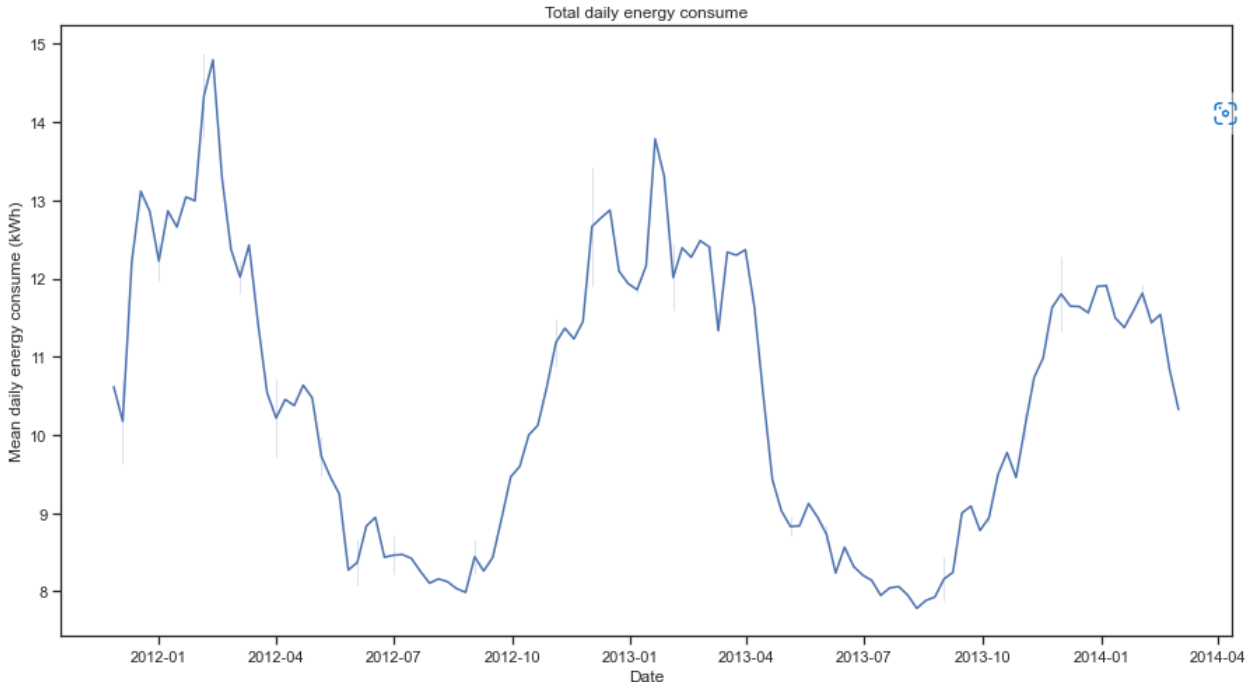
It is possible to see a common behaviour relative to the average consumption, in general not higher than 25 Kw/h, although there are outliers in all the groups. Also, there is something interesting to observe and it is the distribution of the groups Steady Neighbourhoods and Successful Suburbs, because in contrast with the other groups (even versus the groups of other categories), they show a lower skewness, the concentration around the median is not as intensive as the other groups. According to this analysis, a t-test to contrast the means of the groups does not seem a good approach since the similar consumption pattern within groups.

Given the previous analysis and in line to find out if there are significant differences between the groups, a line plot to see the consumption across the time is presented below:

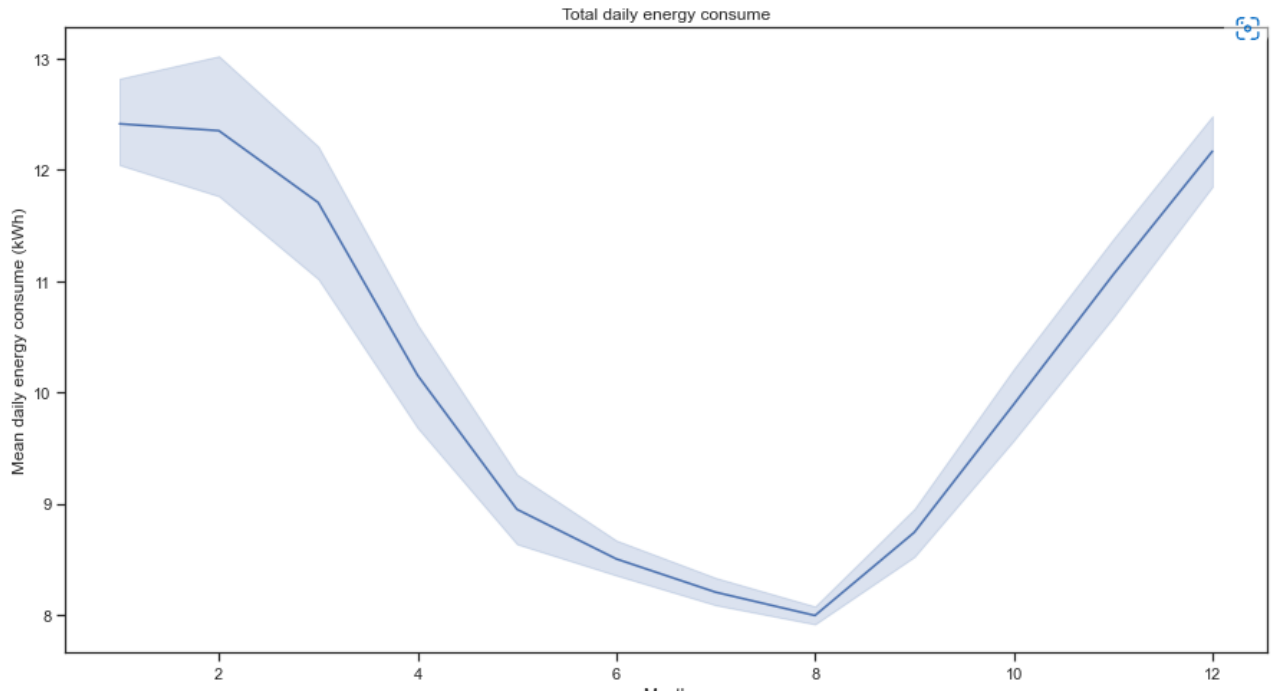


Once again, the consumption for all the groups seems to be really similar, except for the first months of the study (first quarter of 2012) when the chart shows a higher consumption for the green group (Starting Out). The trend is the same as usual, it is higher for the coldest months and lower for the hottest ones; but no matter the group, the amount of used energy has a common range from almost 7 Kw/h to 15 Kw/h.

Now checking the trends over the time, it really has the same behavior over the past 2 years, so the numbers match and it indicates we are going the right way with this stakeholders



But to get further details about this segment, this is the estimation over the months, the hypothesis summer is likely to get lower numbers since it is not that necessary the warm up power. This is estimated with an possible error, and this proves the Summer Theory



Category 4: Financially Stretched Households

Figure 1. Basic statistics of financially streched households of energy consume

	energy_count	energy_min	energy_sum	energy_max
count	463116.0	463116.00	463116.00	463116.00
mean	48.0	0.06	9.90	0.83
std	0.0	0.06	6.52	0.57
min	48.0	0.00	0.00	0.00
25%	48.0	0.02	5.54	0.40
50%	48.0	0.04	8.56	0.72
75%	48.0	0.07	12.48	1.11
max	48.0	1.43	90.10	6.39

The descriptive statistics for the numerical data show that the columns on the daily dataset include many statistics generated from the Daily dataset. In this case, the data of the Financially Streched category have around 460 thousands entries, with 9 numerical columns.

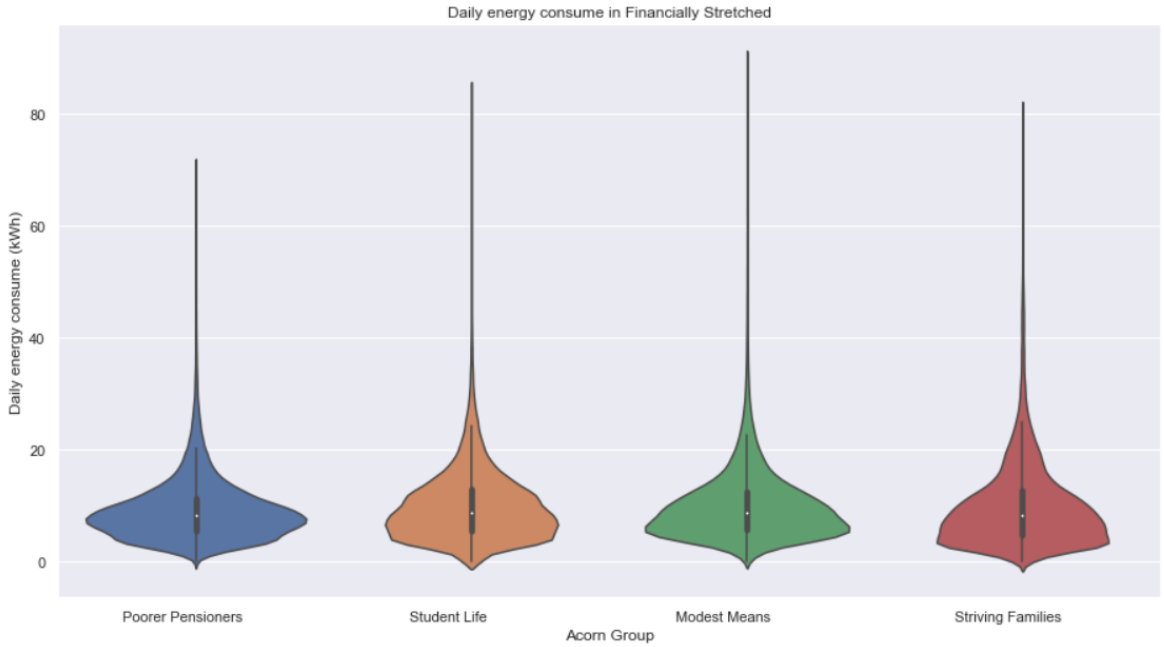
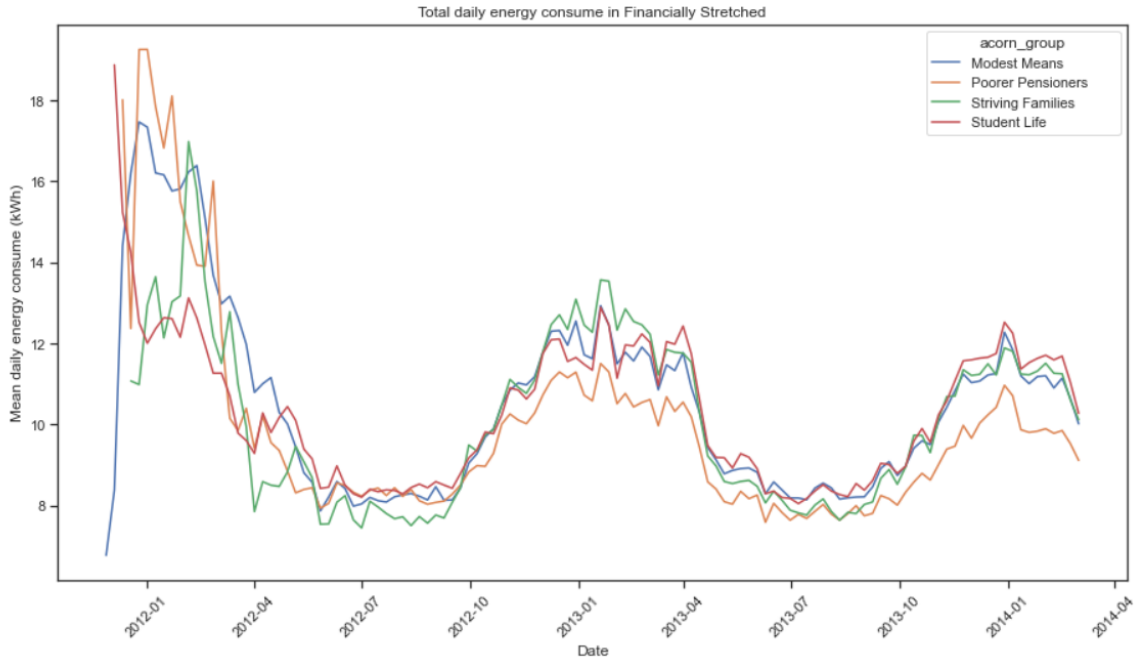


Figure 1. Historical behaviour of daily energy consumes in financially streched



Financially streched households, as well as all households considered have a increase in the energy consumption during the winter months and decrement in summer, registering the lowest daily energy consume in August. The increment in the electric consumption could be explained by the use of electric heaters in the months where the lowest temperatures are recorded.

Figure 1. Mean daily energy consume in financially streched households per month

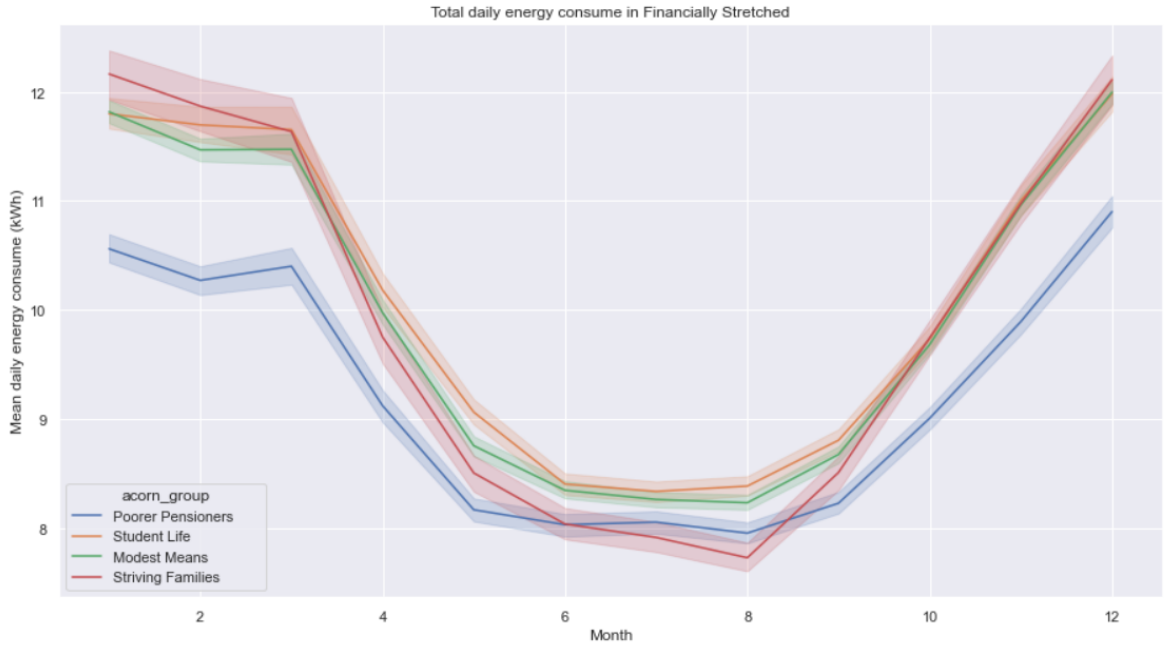
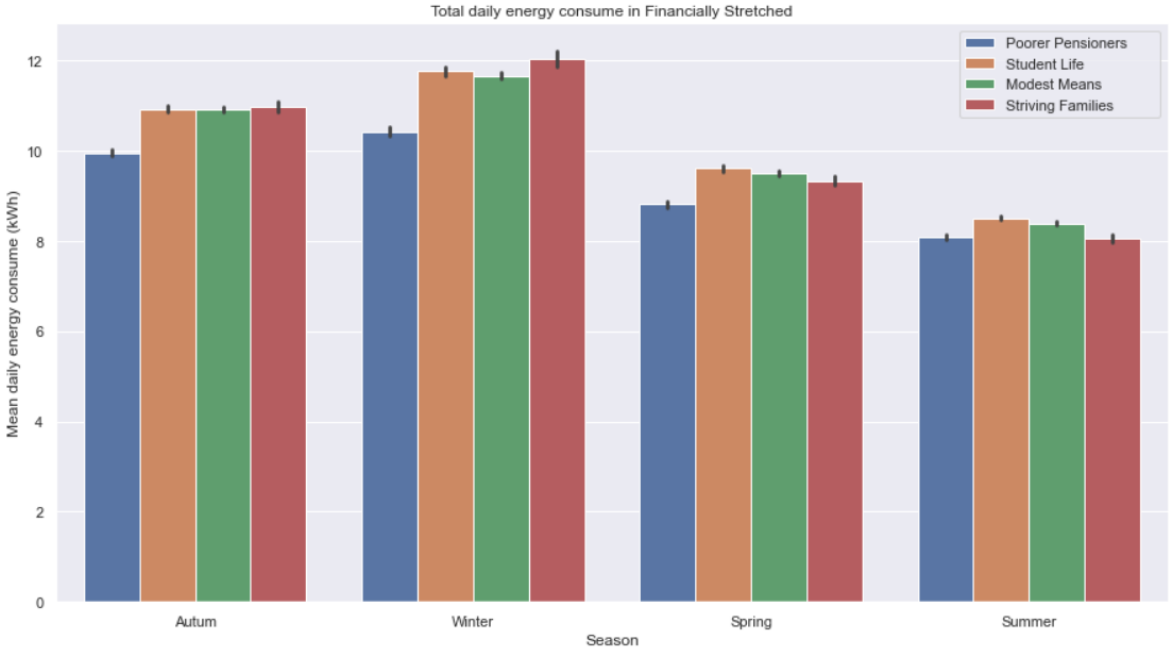
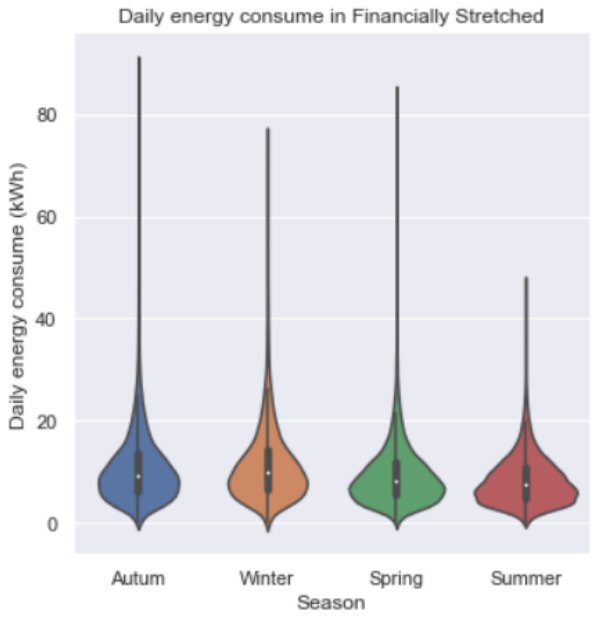


Figure 1.Mean daily energy consume in not private households per season



According with the violin graphic plotted below, there was explored the existence of a difference in the mean of the daily energy consume by financially streched households in seasons. In the autumn and winter, there 25% of daily energy records are greater than 10 kWh. While 25% of daily energy collected in summer and spring are greater than 8 kWh.

Figure 1. Violin graphic daily energy consume per season



									energy_sum
		count	mean	std	min	25%	50%	75%	max
acorn_group	season								
Modest Means	Autum	59963.0	10.91	6.88	0.00	6.34	9.52	13.55	90.10
	Spring	52690.0	9.50	6.15	0.00	5.49	8.17	11.76	76.82
	Summer	49491.0	8.39	4.71	0.00	5.03	7.40	10.75	46.36
	Winter	39579.0	11.65	7.61	0.00	6.67	9.99	14.30	75.83
Poorer Pensioners	Autum	26732.0	9.95	6.10	0.00	5.94	8.77	12.35	70.76
	Spring	23242.0	8.80	5.15	0.00	5.32	7.90	10.85	55.65
	Summer	23441.0	8.08	4.46	0.00	4.93	7.41	10.22	38.11
	Winter	17388.0	10.42	6.38	0.00	6.30	9.12	12.76	62.03
Striving Families	Autum	19614.0	10.96	8.06	0.34	5.29	9.14	14.14	80.43
	Spring	16866.0	9.33	7.08	0.00	4.55	7.67	11.84	75.02
	Summer	16146.0	8.05	5.30	0.00	4.10	6.83	10.72	41.28
	Winter	12835.0	12.03	9.66	0.00	5.43	9.52	15.33	70.25
Student Life	Autum	29050.0	10.92	7.00	0.00	5.91	9.61	14.20	76.27
	Spring	28459.0	9.60	6.10	0.00	5.34	8.39	12.48	84.38
	Summer	28558.0	8.51	4.81	0.00	4.74	7.81	11.47	47.25
	Winter	19062.0	11.75	7.53	0.00	6.39	10.19	15.10	68.65

In order to compare the means for the groups of the category, a t-test will be performance as follows:

Test 1: A subset of the data frame with the consumption per group is taken just with the *energy_sum* (total consumption per household per day) and the group, *Poorer Pensioner* this case, and the same is performed with the group *Student Life* , then the t-test is applied, and the result is the next:

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-29.21328	195919.688004	two-sided	3.329459e-187	[-0.85, -0.74]	0.131017	4.065e+182	1.0

Interesting to see the p-value, which is near to 0, it means, with a 0.05 significance level that the null hypothesis must be rejected in favour of the alternative hypothesis, then it can be concluded that there is not statistical difference between the average consumption of the *Poorest Pensioners* group and the *Student Life* group.

Test 2: The same procedure is conducted again but now to compare *Poorest Pensioners* and Modest Means groups:

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-34.078528	200705.843339	two-sided	8.204352e-254	[-0.85, -0.76]	0.128837	2.104e+249	1.0

Once again, the result is to reject the null hypothesis, it is notable to see the confidence interval of the two tests and how close they are. The conclusion is as expected, there is not a statistical difference between the average consumption of the *Poorest* Pensioners group and the *Modest Means* group.

Test 3: Finally, the same method is used to the *Poorest Pensioners* and *Strivings Families* groups:

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-29.21328	195919.688004	two-sided	3.329459e-187	[-0.85, -0.74]	0.131017	4.065e+182	1.0

In this case, with the same 0.05 significance level, the p-value is near to 0, the confidence interval includes negative and positive values, there is not a statistical difference between the average consumption of the *Poorest* Pensioners group and the *Strivings Families* group.

Category 5: Urban Adversity

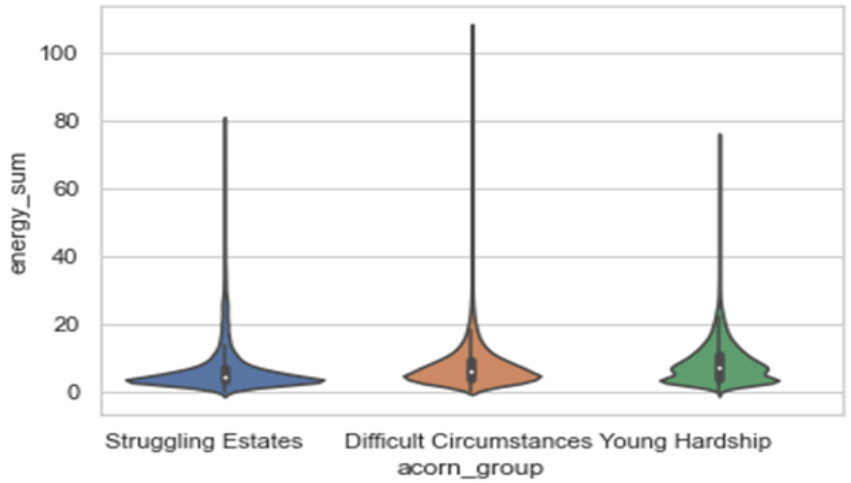
This category includes 3 subcategories, classified like: Acorn O= Young hardship, Acorn P= Struggling Estates, Acorn Q= Difficult Circumstances.

Those are the descriptive statistics by the category.

	energy_count	energy_min	energy_sum	energy_max
count	657698.0	657698.000000	657698.000000	657698.000000
mean	48.0	0.040515	7.584384	0.693320
std	0.0	0.046536	5.948498	0.599896
min	48.0	0.000000	0.000000	0.000000
25%	48.0	0.014000	3.769000	0.280000
50%	48.0	0.029000	6.090000	0.523000
75%	48.0	0.053000	9.540000	0.918000
max	48.0	1.548000	107.601999	8.285000

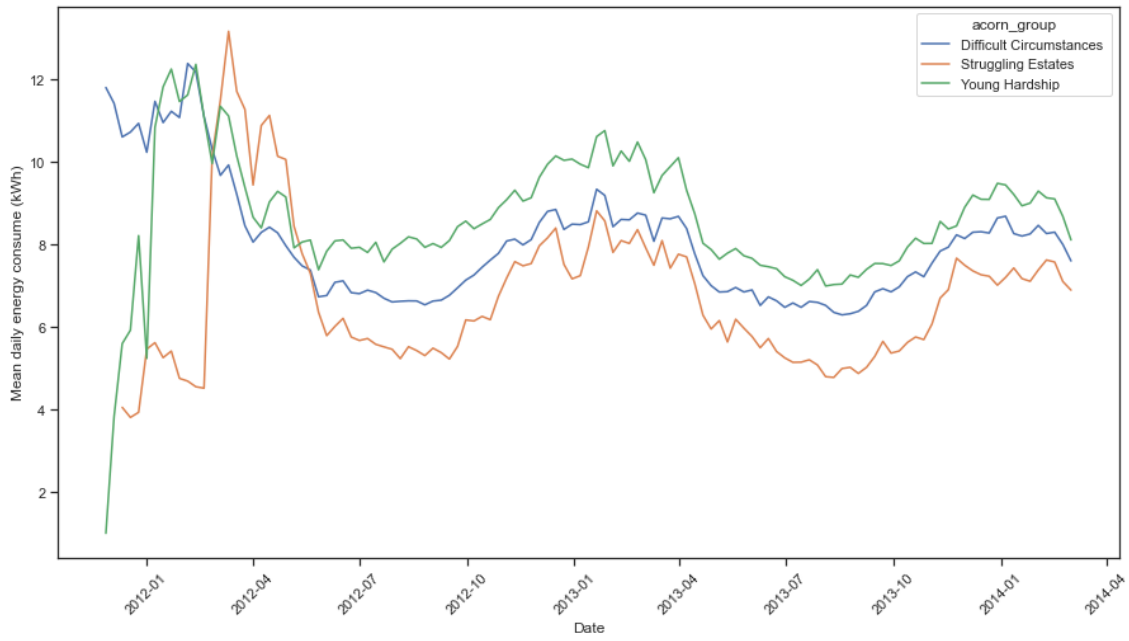
The descriptive statistics show us that we have 48 counts, a record each half hourly, while the 24 hours of the day. The mean consumption by day in this group is 7,58 kw/h and the estándar deviation show us that the data are scattered.

Now, we can see the information in a violín plot:

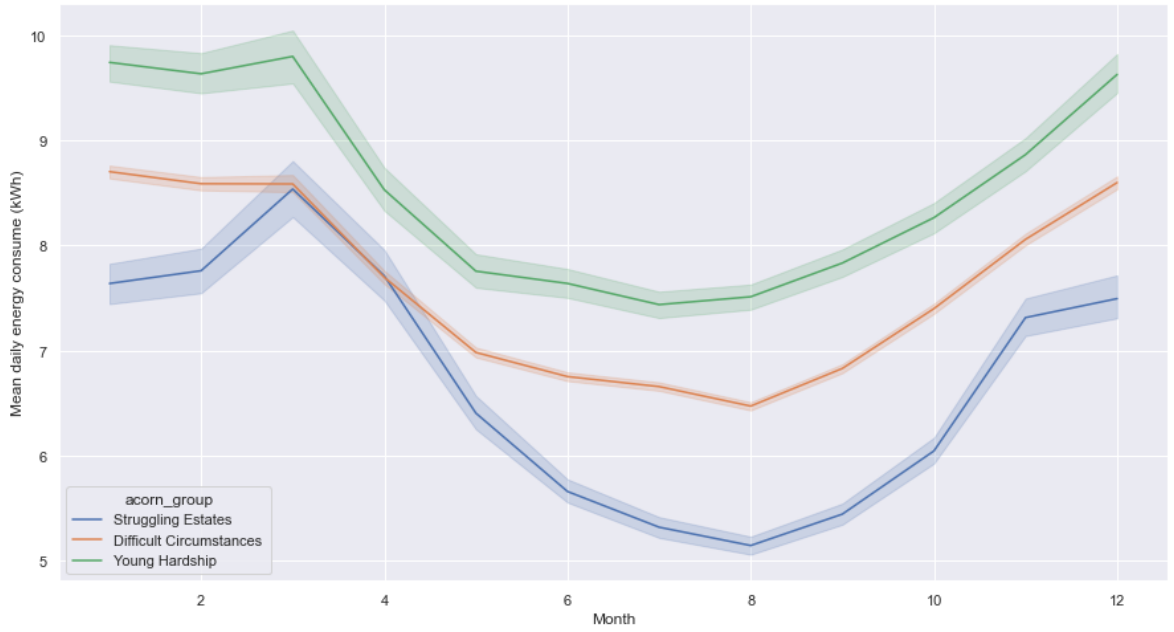


The plot shows the distribution in each one of the groups inside Urban Adversity. The median seems to be similar and the major difference is in the máximum values. Struggling Estates show more consumption than Difficult circumstances and the least more consumption than Young Hardship.

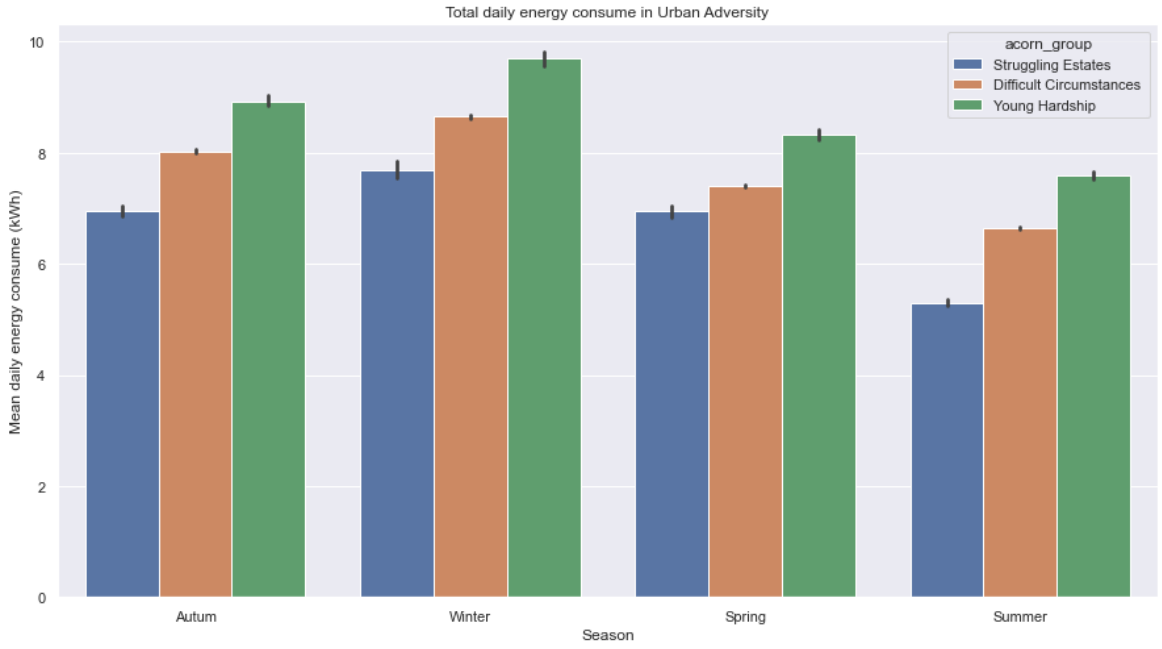
We're gonna see the behaviour of the daily consumption of energy through time.



The demand of energy shows a Deep since 2012 and their behaviour is seasonal.

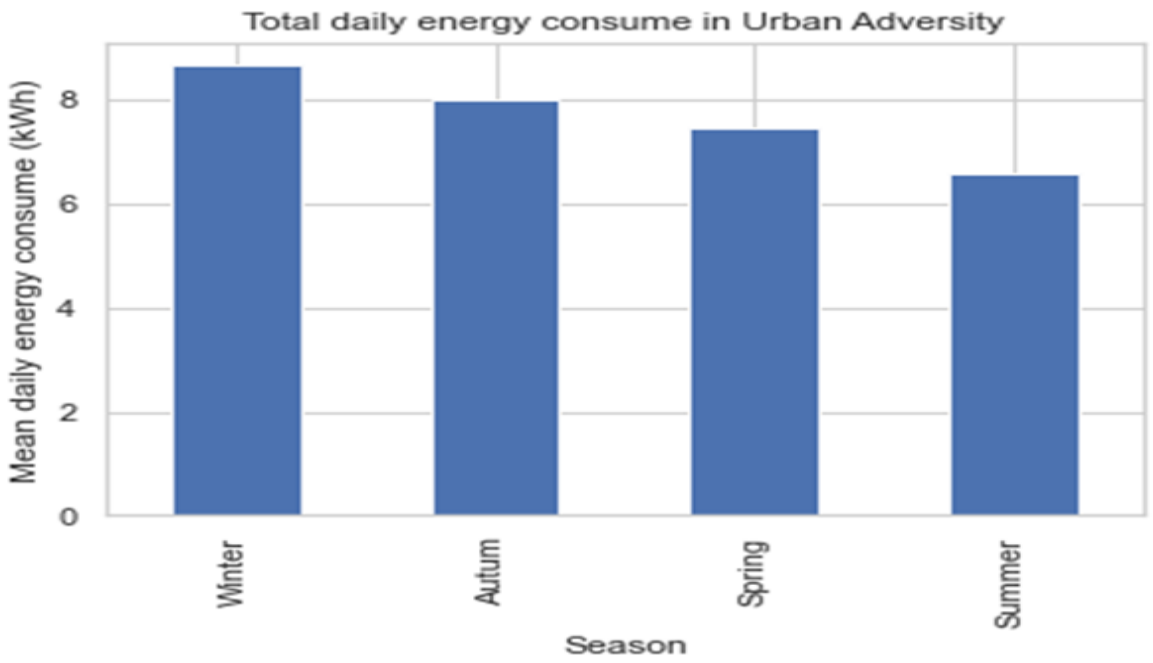


That is the behaviour of the monthly consumption of energy between 2011-2013.



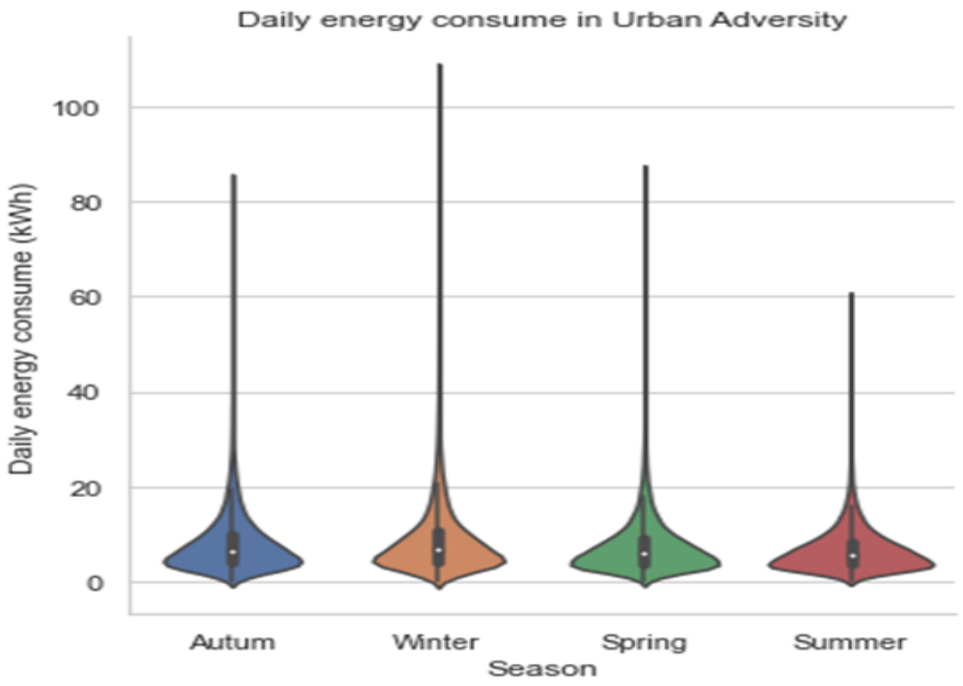
The demand for energy between January and March is high, then it falls until august and then grows until december.

Daily consumption of energy by season of the year



We can see that the Winter is the season when the demand of the energy is the highest, followed by autumn, spring and the summer show the lowest consumption.

¿How is the data distributed in the daily consumption season by year?



The violin plot shows the distribution of the consumption of energy by season. The median seems to be similar and the difference is in the maximum values. The seasons when the consumption is higher, the maximum values are higher.

Those are the descriptive statistics or the daily consumption by season:

season	count	mean	std	min	25%	50%	75%	max
Autum	183114.0	7.998615	6.303241	0.0	3.92700	6.396	10.12375	84.575000
Spring	176813.0	7.430022	5.767566	0.0	3.69700	5.978	9.33600	86.565000
Summer	177345.0	6.588925	4.429899	0.0	3.49800	5.569	8.49400	60.073000
Winter	120426.0	8.647124	7.225961	0.0	4.13625	6.705	10.70775	107.601999

Category 6: Not Private Households

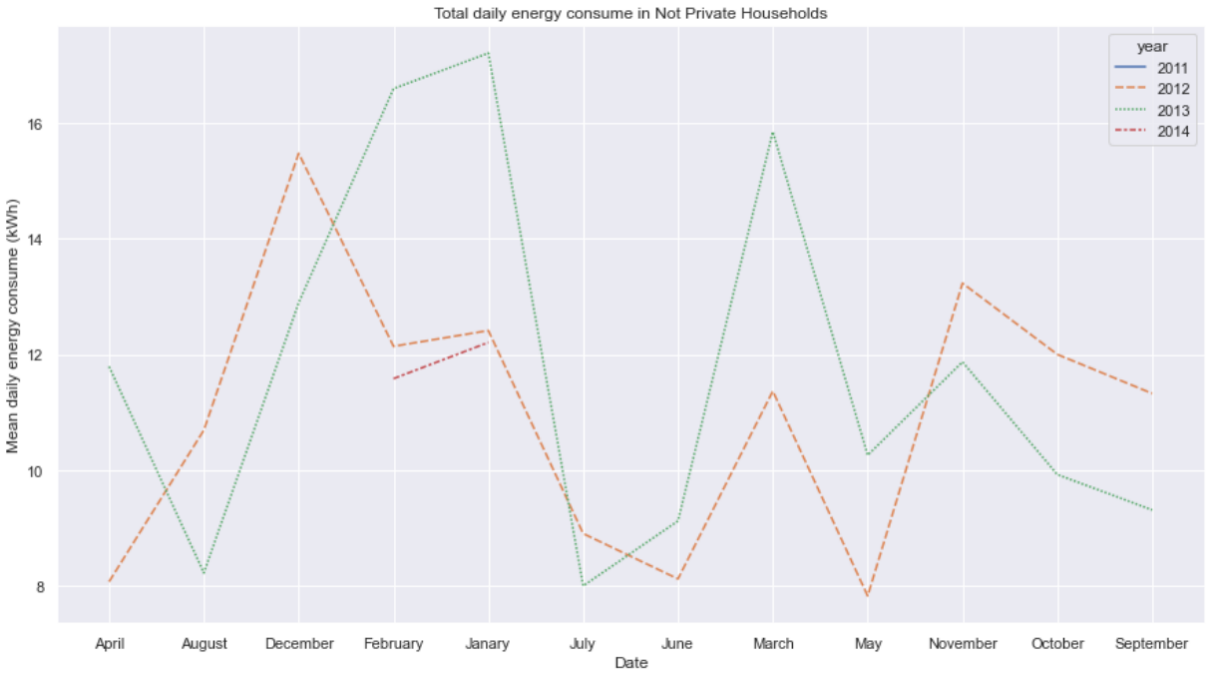
In the category “not private households”, all comunal, business and non-residential areas are grouped. During the development of the study, this group was considered to collect their daily and half-hourly energy consumption. Below, you will find a summary chart with the basic statistics of energy consumption.

Figure 1. Basic statistics of not private households of energy consume

	energy_count	energy_min	energy_sum	energy_max
count	29167.0	29167.00	29167.00	29167.00
mean	48.0	0.06	11.68	0.92
std	0.0	0.10	13.20	0.86
min	48.0	0.00	0.00	0.00
25%	48.0	0.01	4.03	0.30
50%	48.0	0.03	7.40	0.75
75%	48.0	0.06	14.69	1.29
max	48.0	2.16	150.36	8.75

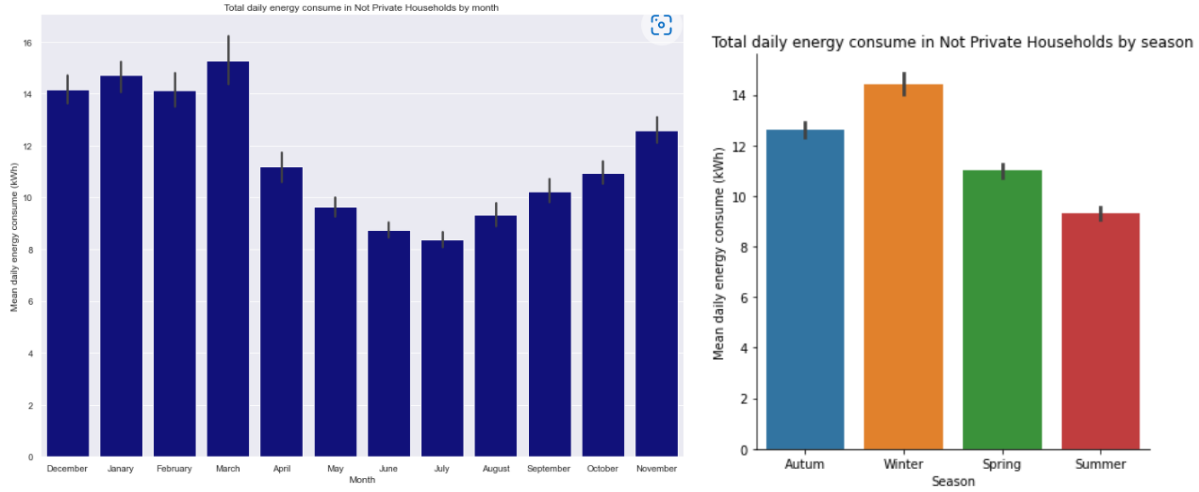
The following graphic shows the historical behaviour of daily energy consumption in the group “not private households” during the time the study was conducted.

Figure 1. Historical behaviour of daily energy consumes in not private households



Not private households, as well as all households considered have a increase in the energy consumption during the winter months and decrease in summer, registering the lowest daily energy consumption in July. The increment in the electric consumption could be explained by the use of electric heaters in the months where the lowest temperatures are recorded.

Figure 1. Daily energy consume in not private households per month and seasonally adjusted

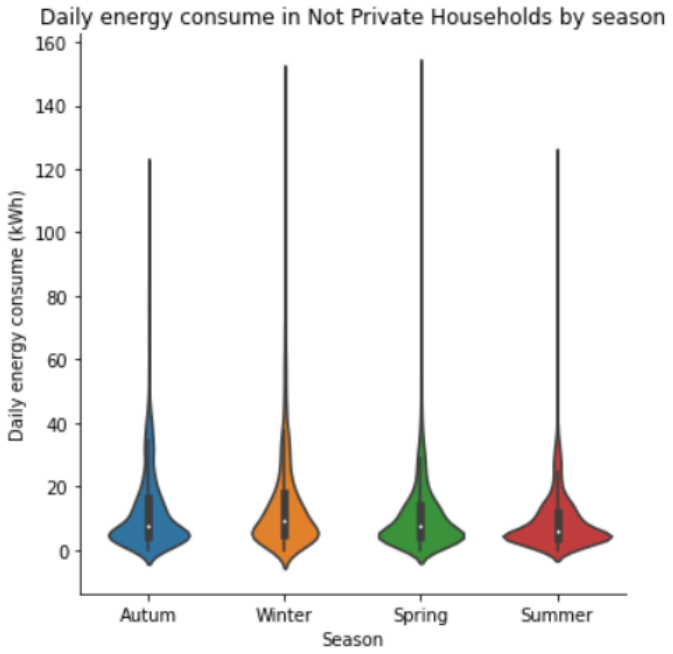


[a] [b]

[a] Mean daily energy consume in not private households per season. [b] Mean daily energy consume in not private households per season.

According with the violin graphic plotted below, there was explored the existence of a difference in the mean of the daily energy consume by not private households in seasons. In the autumn and winter, there 25% of daily energy records are greater than 17 kWh. While 25% of daily energy collected in summer and spring are greater than 14 kWh.

Figure 1. Violin graphic daily energy consume per season



season	count	mean	std	min	25%	50%	75%	max
Autum	8384.0	12.599724	13.569710	0.0	4.123	7.746	16.3485	118.811000
Spring	7477.0	11.011429	12.468938	0.0	3.990	7.291	13.9630	150.362001
Summer	7657.0	9.314072	10.289270	0.0	3.704	6.151	12.0940	122.918000
Winter	5649.0	14.431079	16.131951	0.0	4.646	9.423	17.8070	146.932999

To determine if there is a statistically significant difference between non-private households' daily energy consumption in the first and second half of the year, a T-test was conducted with a significant level of 5% obtaining a p-value of 4.57e-93. This result proves that there is a difference between the means of both periods.

Figure 1. **T-test mean difference between first and second half of the year**

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	20.544624	26504.198483	two-sided	4.570196e-93	[2.88, 3.49]	0.242989	1.185e+89	1.0

Exploring the difference in the seasons of the first half, spring and summer, an additional mean difference T-test was conducted with a significant level of 5 %. The obtained p-value is lower than the significant level, then the nule hypothesis of mean equality is rejected proving than the mean daily consumption of spring and summer are different.

Figure 1. **T-test mean difference between spring and summer**

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-9.122395	14474.334676	two-sided	8.301753e-20	[-2.06, -1.33]	0.148653	1.849e+16	1.0

Finally, the difference between the means of consumption of the autumn and winter seasons was verified with a significance level of 5%. Below are the results of the T-test developed.

Figure 1. **T-test mean difference between autumn and winter**

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-7.021315	10681.372216	two-sided	2.331462e-12	[-2.34, -1.32]	0.124964	9.178e+08	1.0

Daily weather dataset

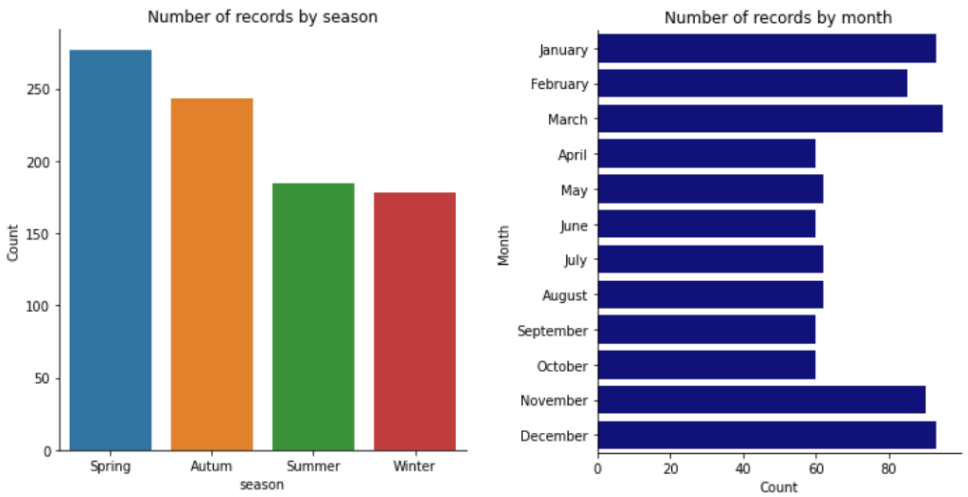
In the daily weather dataset, there is meteorological information about the daily climate in London, as minimum and maximum temperature, cloud cover, visibility, time in which the highest and lowest temperature were registered, and wind. The dataset has 882 records, one for each day in which the study was conducted, just one record has null data in the columns cloud cover and uvIndex. To complete this information, an imputation of the mean values of total records from cloud cover and uvIndex was made. Below is a table with the summary of the most important descriptive statistics of the dataset:

Table 1. Basic statistics of daily weather dataset

	temperatureMax	windBearing	cloudCover	windSpeed	pressure	visibility	uvIndex	temperatureMin	moonPhase
count	882.00	882.00	882.00	882.00	882.00	882.00	882.00	882.00	882.00
mean	13.66	195.70	0.48	3.58	1014.13	11.17	2.54	7.41	0.50
std	6.18	89.34	0.19	1.69	11.07	2.47	1.83	4.89	0.29
min	-0.06	0.00	0.00	0.20	979.25	1.48	0.00	-5.64	0.00
25%	9.50	120.50	0.35	2.37	1007.44	10.33	1.00	3.70	0.26
50%	12.62	219.00	0.47	3.44	1014.62	11.97	2.00	7.10	0.50
75%	17.92	255.00	0.60	4.58	1021.76	12.83	4.00	11.28	0.75
max	32.40	359.00	1.00	9.96	1040.92	15.34	7.00	20.54	0.99

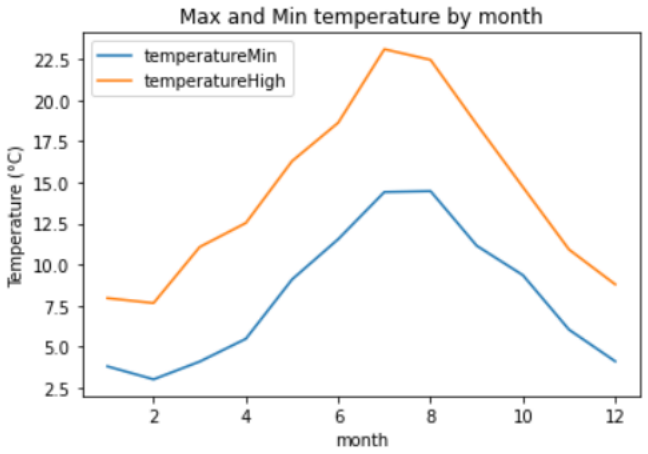
To carry out analyses by season of the year, an additional column was created from the date of registration, having the greatest number of records collected in the season of spring and the lowest in winter. This is aligned with the fact that the study began in November 2011 and ended in February 2014. For that reason, during the months of November and March, there are the greatest number of meteorological records.

Figure 1. Number of records by season and month



In this project, we are interested in discovering the relationship between energy consume per household and meteorological data. The first variable to consider is temperature due to the different seasons that London population live in. The following figure shows minimum and maximum temperatures per month, having the greatest temperature record between July and August and lowest in February. The lowest temperature register is 3°C in February.

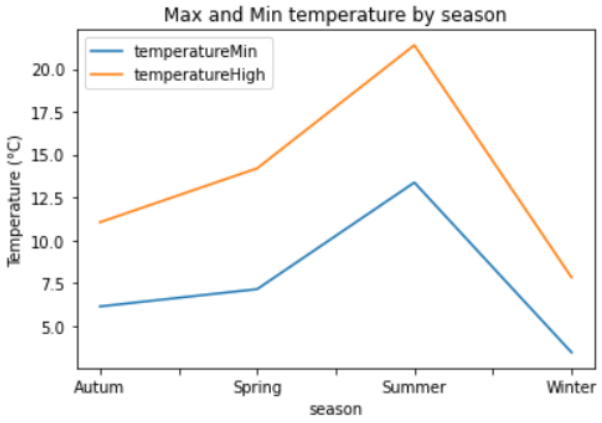
Figure 3. Mean maximum and minimum temperatures by month



month	temperatureMin	temperatureMax
1	3.82	8.29
2	3.03	7.76
3	4.11	11.12
4	5.50	12.56
5	9.10	16.35
6	11.55	18.63
7	14.41	23.10
8	14.47	22.46
9	11.15	18.58
10	9.36	14.81
11	6.05	11.13
12	4.12	9.09

Additionally, an analysis per season was conducted to identify the mean values of minimum and maximum temperature during each season.

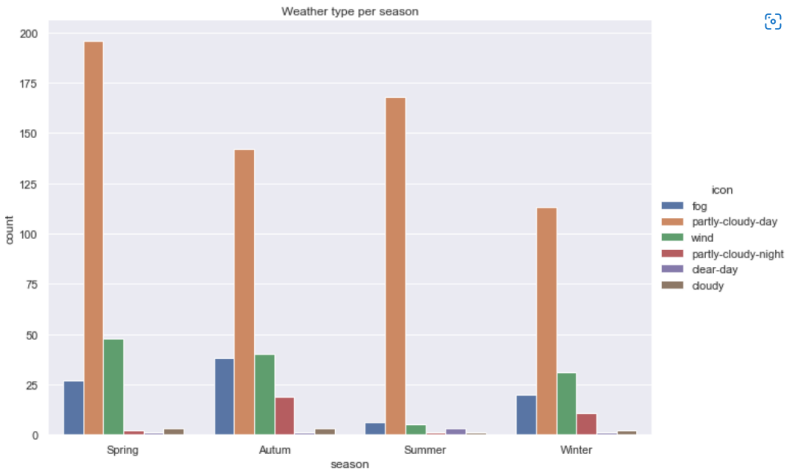
Figure 4. Mean maximum and minimum temperatures by season



season	temperatureMin	temperatureMax
Autum	6.13	11.26
Spring	7.14	14.23
Summer	13.37	21.41
Winter	3.44	8.04

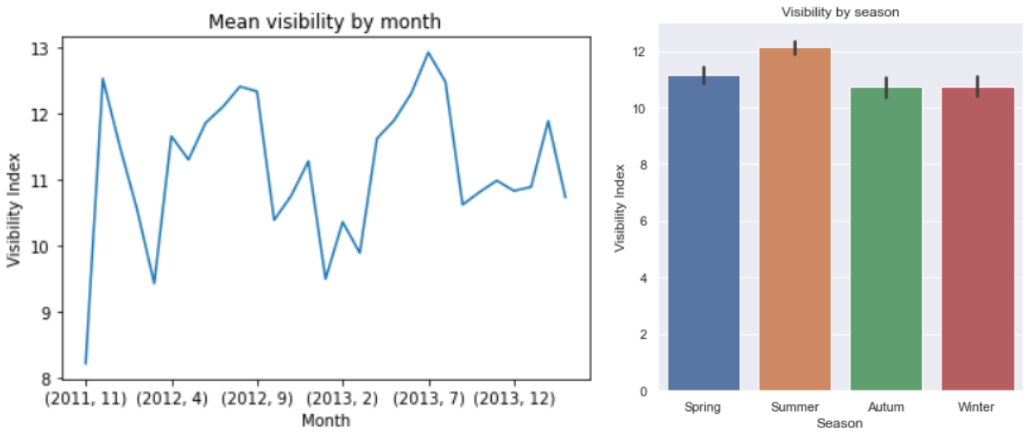
The second meteorological aspect to consider is icon used to describe weather in each day in London. According with the available data,during the months in which the study was developed, the 70% of the time the weather was partly-cloudly in London.

Figure 5. Weather type by season



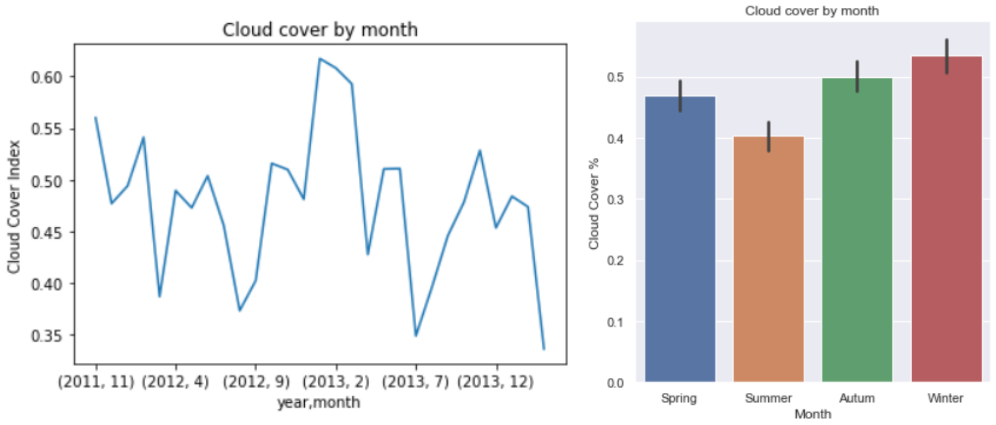
The third metereological variable is visibility, in order to find if there is any correlation between a lower visibility and the need to consume more energy for artificial lights. According with the data, in summer (July) is the greatest visibility index.

Figure 6. Mean visibility index per month and season



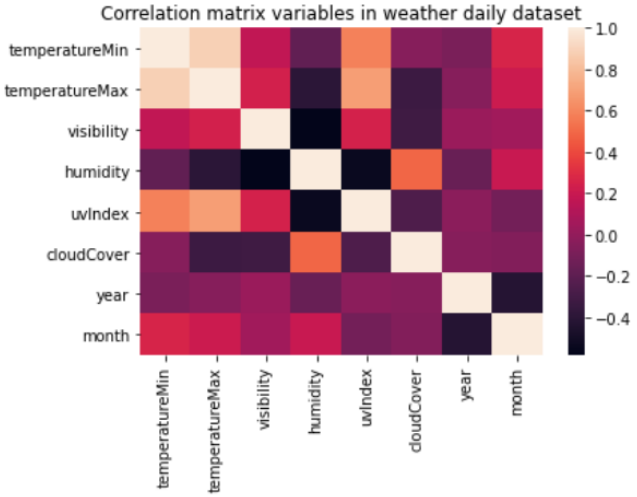
The last variable explored is the percentage of cloud cover, in summer used to be a cloud cover of almost 40%, while in winter, this measure increases to 54%. At the beginning of 2013, there were the highest cloud cover achieving more than 60%.

Figure 7. Mean cloud cover per month and season

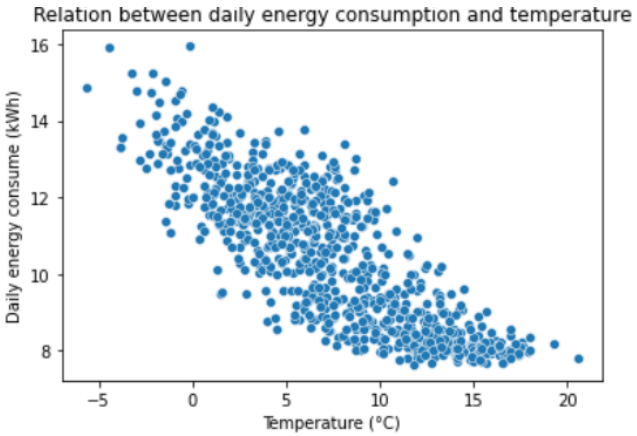


Finally, a correlation matrix is created to detect relations between variables in the dataset. It is shown that there is a positive correlation between the min, max and uvIndex. On the other hand, there is a negative relationship between temperature, humity and cloud cover percentage.

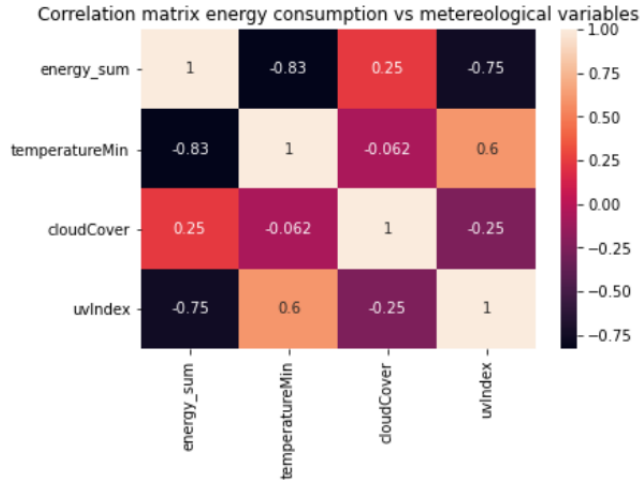
Figure 8. Heatmap of correlation matrix between the variable in daily weather dataset



Additionally, it was explored the relationship between daily energy consumption in London households and the registered temperature monthly. In the next figure is shown graphically that if there is an increase in temperature, the energy consumption in household will decrease.



To verify the correlation between the energy consumption and the meteorological variables explored in the weather daily report. In this graph, it is shown that the temperature has the greater negative correlation (-0.83) with the monthly energy consumption. The other variable with the greater negative influence to the energy consumption is the uvIndex (-0.75).



Hourly weather dataset

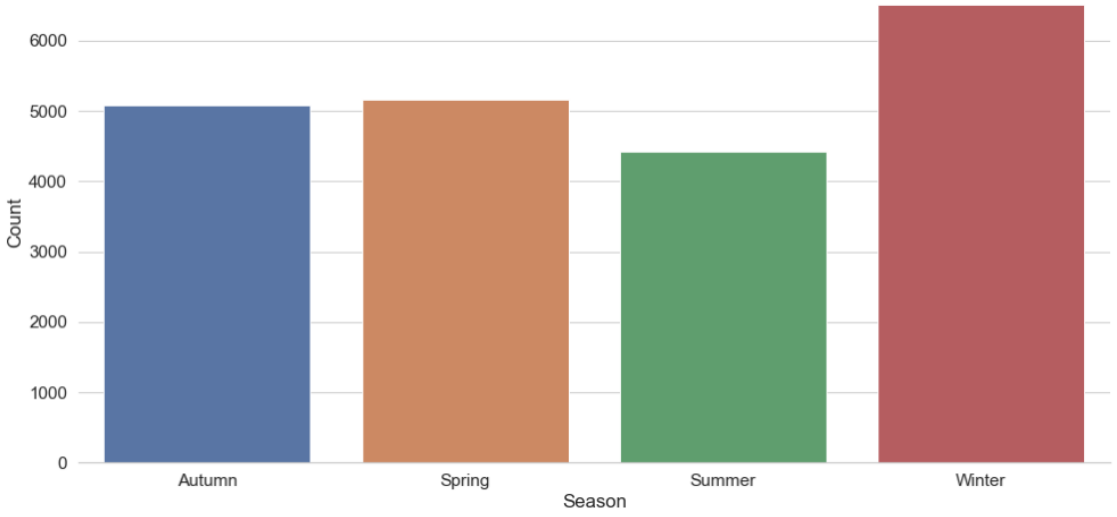
In the hourly weather dataset, there is meteorological information about the daily climate in London, as minimum and maximum temperature, cloud cover, visibility, time in which the highest and lowest temperature were registered, and wind. The dataset has 21.165 records, one for each hour in which the study was conducted, only 13 records have null data in the pressure column. To complete this information, an imputation of the mean values of total records from pressure was made. Below is a table with the summary of the most important descriptive statistics of the dataset:

Table 1. Basic statistics of hourly weather dataset

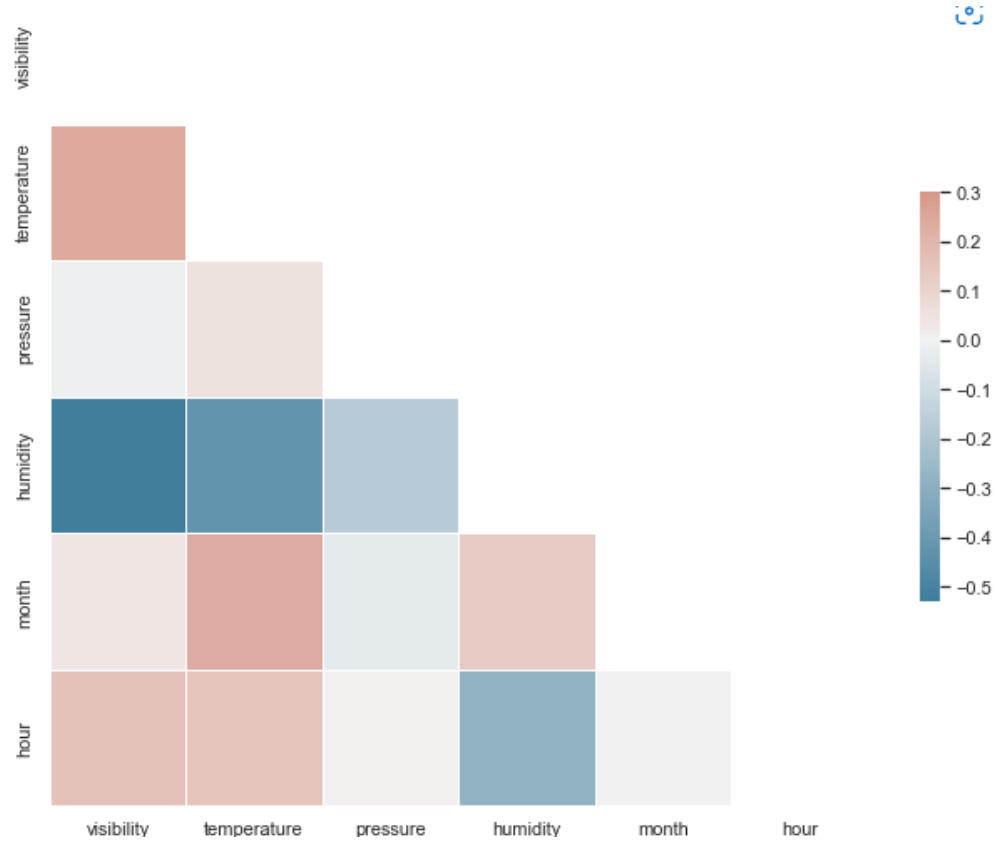
	visibility	windbearing	temperature	dewpoint	pressure	apparenttemperature	windspeed	humidity
count	21165.00	21165.00	21165.00	21165.00	21165.00	21165.00	21165.00	21165.00
mean	11.17	195.69	10.47	6.53	1014.13	9.23	3.91	0.78
std	3.10	90.63	5.78	5.04	11.38	6.94	2.03	0.14
min	0.18	0.00	-5.64	-9.98	975.74	-8.88	0.04	0.23
25%	10.12	121.00	6.47	2.82	1007.44	3.90	2.42	0.70
50%	12.26	217.00	9.93	6.57	1014.77	9.36	3.68	0.81
75%	13.08	256.00	14.31	10.33	1022.05	14.32	5.07	0.89
max	16.09	359.00	32.40	19.88	1043.32	32.42	14.80	1.00

In this project, we are interested in discovering the relationship between energy consume per household and meteorological data. The first variable to consider is temperature due to the different seasons that London population live in.

The following graph shows the total count of hours recorded for each season, indeed it was expected that the count would be higher in the winter because this season is longer than the others.



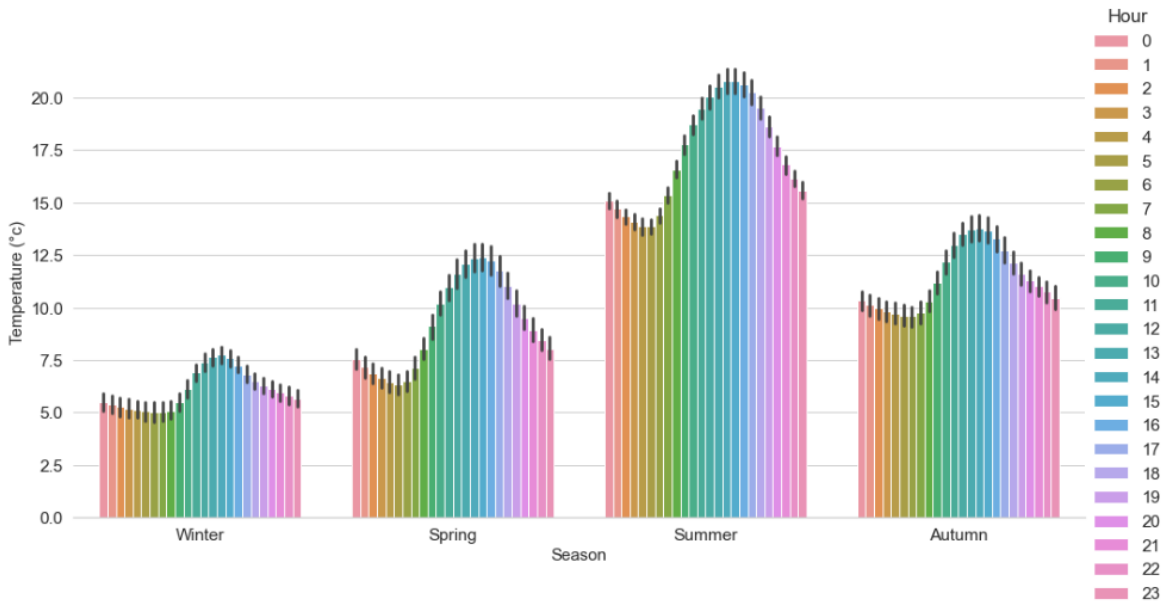
Regarding the most important variables that could have a potential in the incidence of energy consumption in households in London, this correlation is shown below.



It is without a doubt that energy consumption is influenced by temperature, and in turn this variable has a high dependence on other variables such as humidity, since the higher the temperature, the lower the humidity, therefore the correlation can be seen in red, for On the other hand, visibility also shows a positive correlation with temperature. Therefore, as it is already known how the temperature influences the humidity, in subsequent analyzes the temperature is compared with the visibility.

The following graph shows the temperature differences in each of the hours of the day in each season. It is evident in chronological order that as you go through the temperature increases and the same behavior in each season is very similar, forming a behavior like in S where from the first hour of the day until around 8 am there is a drop in temperature, but then from this time until about 2 pm, the trend decreases again. This occurs in each season but on a different scale, showing a much more aggressive change in summer than in other seasons.

From January to August is that when the summer ends, the temperature increases after this date then it begins to decrease.



The effect of temperature on visibility is confirmed, on occasions where the temperature decreases, visibility also decreases, effectively on those sunny days that are more common in summer are the days where there is more visibility throughout the year.

It can be evidenced that despite the fact that the temperature is not a determining factor in visibility, since on those days when the temperature is lower in the year, less visibility is expected and it is not like that on all occasions, therefore there are other factors that they influence. The end of the year has been the period where there is less visibility and this basically comes down to the snowfall.