DS4A Final Project Report

Team 42

1. **Introduction and Problem Statement**

The British government is trying to upgrade energy supply and tackle climate change by better understanding the energy consumption by installing smart meters in every home in England, Wales, and Scotland. They recorded 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 2011 and February 2014.

In this project our objective was to explore the data from smart meters and discover energy consumption patterns and suggest strategies to upgrade the energy supply and tackle climate change.

We believe this would be impactful since, changes in climate can affect different industries and have harmful effects on public health. By taking advantage of smart meters in England we can analyze trends and identify different patterns over 2 years. We can forecast the energy consumption at household level, identify what factors are responsible for high energy consumptions in different households, identify different segments of the population, their behavior, based on their energy consumption and find correlations between different fields. By understanding and analyzing the data we can provide different recommendations to help save energy.

1. **Data**

The data contains the energy consumption readings for a sample of 5,567 London Households from November 2011 to February 2014, and is made up of the following files:

**hhblock\_dataset.zip** - a zip file containing 112 tables corresponding to 112 blocks in London. Each row represents a day for one household (as an array) with for example the hh\_0 column is the consumption between 00:00 and 00:30.

|  |  |  |  |
| --- | --- | --- | --- |
| **Columns** | LCLid | day | hh\_i |
| **Information** | Unique household ID | Date of measurement | Electricity usage measured every half hour |

**daily\_dataset.zip** - a zip file containing 112 tables corresponding to 112 blocks in London. Each table contains daily statistical information on the consumption of the households.

|  |  |  |  |
| --- | --- | --- | --- |
| **Columns** | LCLid | day | energy\_median/mean/max/min/std/sum |
| **Information** | Unique household ID | Date of measurement | Statistical information about electricity usage |

**weather\_hourly\_darksky.csv** - hourly weather information including visibility, temperature, precipitation, weather icons, and other weather-related data.

**informations\_households.csv** - household information.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Columns** | LCLid | stdorToU | Acorn | Acorn\_grouped | file |
| **Information** | Unique household ID | tariff type | Acorn group ID | Acorn group name | Where the LCLid is stored |

Our data gives information on Acorn classifications. Acorn is a segmentation tool which categorises the UK’s population into demographic types.

**acorn\_details.csv** - details on the acorn groups and the demographic profile of the people in the group. This dataset comes from CACI International Inc, which collects open source data and groups the UK population into 18 groups based on family structure, age, economic and geodemographic information, etc. The dataset contains the index for multiple parameters in comparison to the national that have an index of 100. The detailed information for the data collections and grouping method can be found at [this website](https://acorn.caci.co.uk/what-is-acorn), as well as in [CACI’s report](https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columns** | MAIN CATEGORIES | CATEGORIES | REFERENCE | ACORN-A/B...Q |
| **Information** | Characteristics categories | Detailed categories | Attributes in the category | Index of the attribute in the group |

**uk\_bank\_holidays.csv** - bank holidays in UK

|  |  |  |
| --- | --- | --- |
| **Columns** | Bank holidays | Type |
| **Information** | Date of the holidays | Name of holidays |

1. **Data Cleaning**

The team did the following work to clean the data:

1. Drop less than 2% of data in daily\_dataset.csv

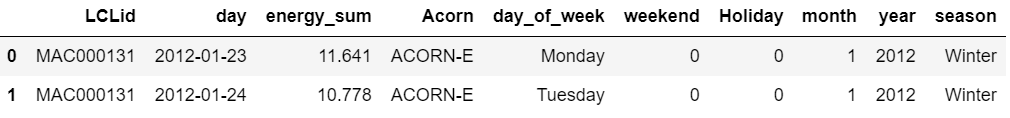
* Keep records which be recorded 48 times a day (energy\_count=48)
* Keep records after 01/23/2012
* Keep households which be recorded for more than 100 days

1. Drop outliers in daily\_dataset.csv
2. Keep only demographic, economical and environment-related features in Acorn\_cleaned\_boolean.csv, and convert index values to percentages
3. Merge daily\_dataset.csv, informations\_households.csv, and Acorn\_cleaned\_boolean.csv

The cleaned dataset has daily average energy consumption among 5492 households from January 2012 to February 2014.

The two final datasets that have been used for analysis and modeling are as follows:

* **Daily Dataset:**



**LCLid** represents the unique household id

**Day** represents the date of the energy consumption reading from the meter.

**Energy\_sum** represents the total energy consumption by the household on a particular day

**Acorn** represents the Acorn group to which the household belongs to.

**Day\_of\_week** represents the name of the week on that day

**Weekend** is a binary feature representing if the day is a weekend (1) or no (0)

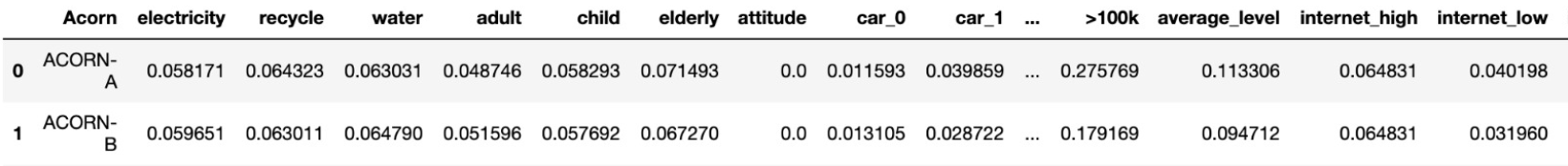
**Holiday** is a binary feature representing if the day is a bank holiday (1) or no (0)

**Month** represents the name of the month

**Year** represents the year

**Season** represents one of the 4 seasons(fall,summer,autumn,winter) to which the record belongs.

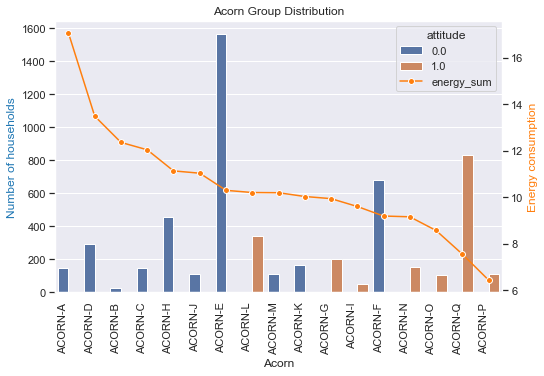
* **Acorn Dataset:**



Each row represents an Acorn group and the various values represent the proportion of the people in that particular group in that Acorn group

1. **Exploratory Data Analysis**
2. **Acorn group distribution**

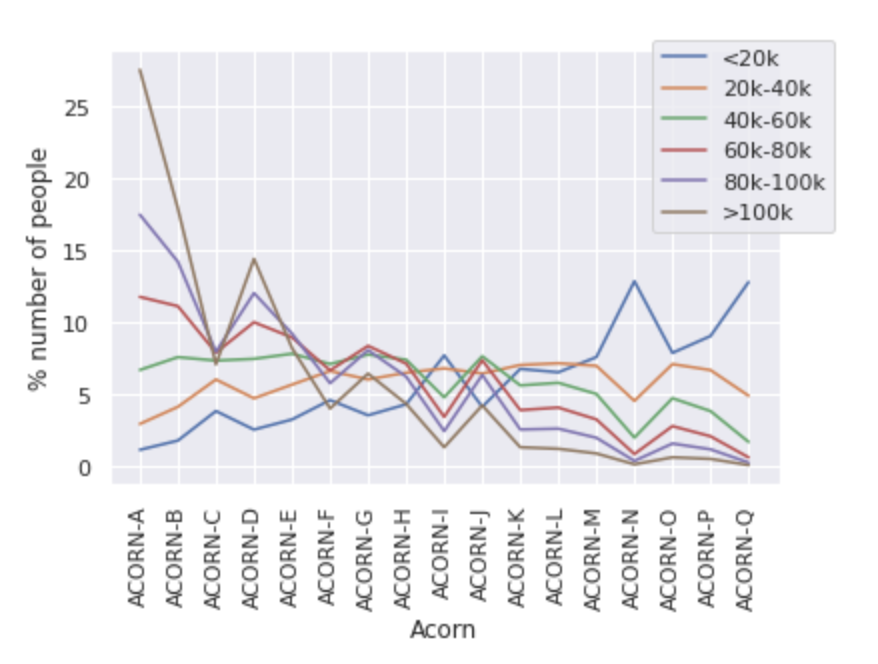
The following plot shows the population of each Acorn group and their average energy consumption of a household. The color of the bars shows the group’s attitude towards environment issues (attitude=0: don’t care, attitude=1: care). From the group we found:



**Figure-1**

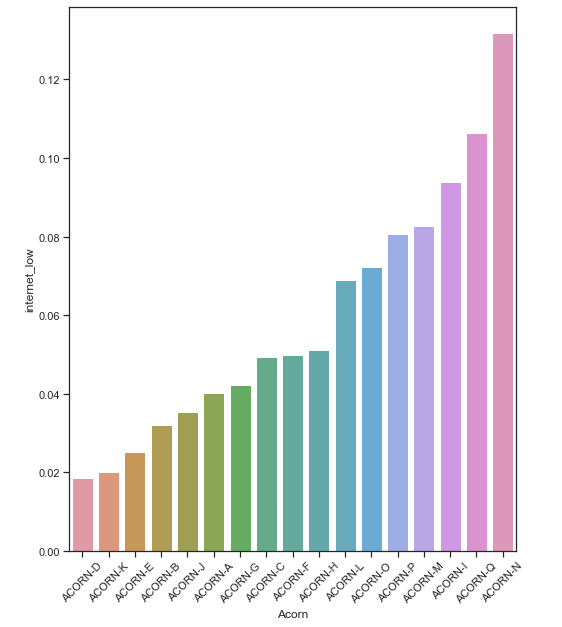
* The number of households of each Acorn group is unbalanced. However, because the energy consumption is measured by the average value in each group and almost each group has at least 100 samples, this problem should not cause serious biased results. Furthermore, we tried to see why Acorn-A and Acorn-B with such less number of households have high energy consumptions. We found out that this is because a significant number of households in these acorn groups belong to the wealthiest category as shown in the figure below. Acorn A and B have a higher ratio of households with income >100k.

**Plot showing the household distribution as per the income across acorn groups**



**Figure-2**

* From Figure 1, we can also see that the orange bar i.e groups with a positive environment attitude have lower energy consumption. There are some outliers which could be due to the the unbalanced dataset.

**Plot showing the different Acorn group internet usage** 

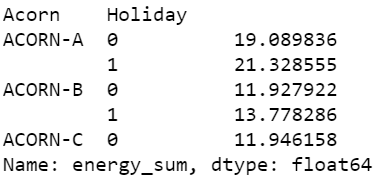
**Figure-3**

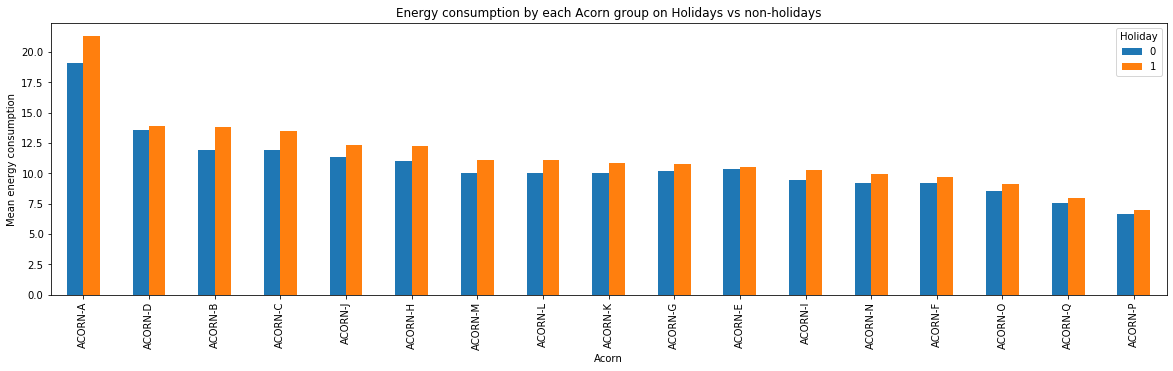
* In the above figure, the x-axis shows the different Acorn groups and the y-axis shows how many of these Acorn groups use low internet, a higher value indicates that the Acorn group hardly uses the internet and a low value indicates that the Acorn group uses the internet very often.

1. **Time related features**

We have used the daily\_dataset to perform the analysis. We then calculated the mean energy consumption for each Acorn group across that feature. The time related features that we have explored are:

* **Holiday**: This is a binary feature representing if the day is a bank holiday (1) or no (0). After transformation data looks like this

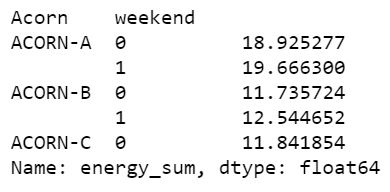


The plot for energy consumption by each Acorn group on Bank Holidays vs non- bank holidays 

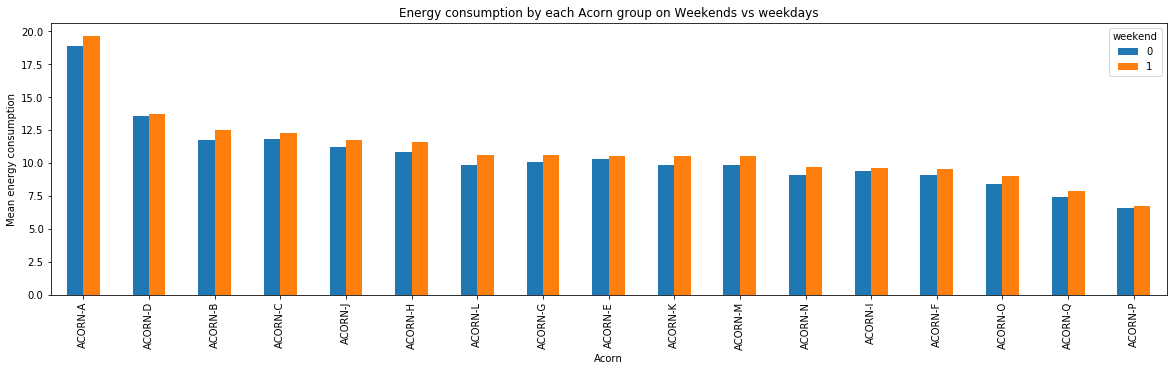
**Figure-4**

It is inferred that across each ACORN group energy consumption is more on the bank holidays than on non-bank holidays and top 3 energy consuming Acorn groups on bank holidays are ACORN A,D and B. This is in line with the fact that people usually prefer to stay at home on holidays and hence the energy consumption is more.

* **Weekend**: This is a binary feature representing if the day is a weekend(1) or no (0). After transformation data looks like this

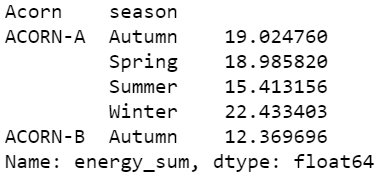


The plot for energy consumption by each Acorn group on weekends vs weekdays -

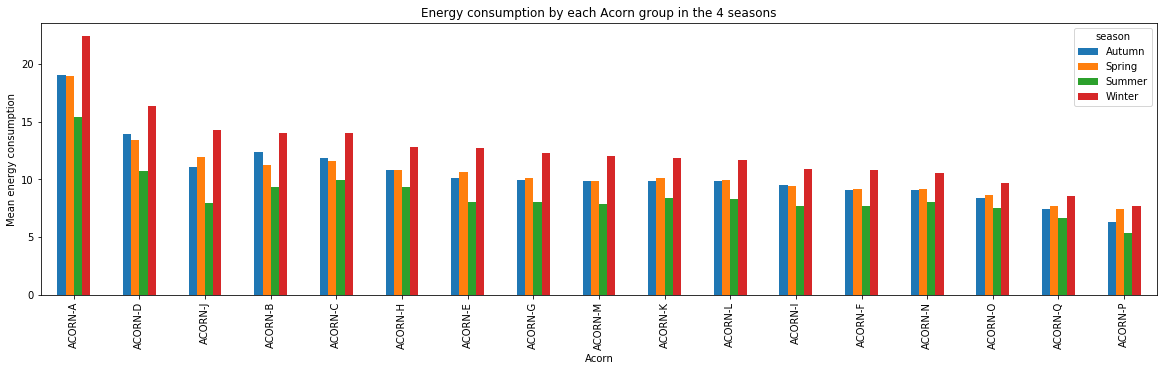
 **Figure-5**

It is inferred that across each ACORN group energy consumption is more on the weekends than on non-bank weekdays and top 3 energy consuming Acorn groups on holidays are ACORN A,D and B.

* **Season**: represents one of the 4 seasons(fall,summer,autumn,winter) to which the record belongs. After transformation data looks like this



The plot for energy consumption by each Acorn group in different seasons -



**Figure-6**

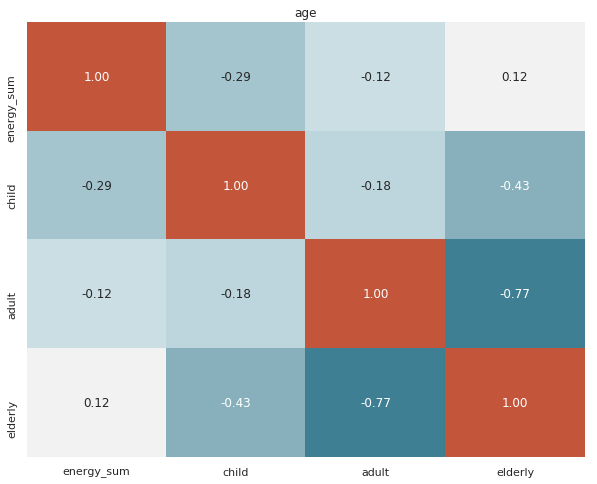
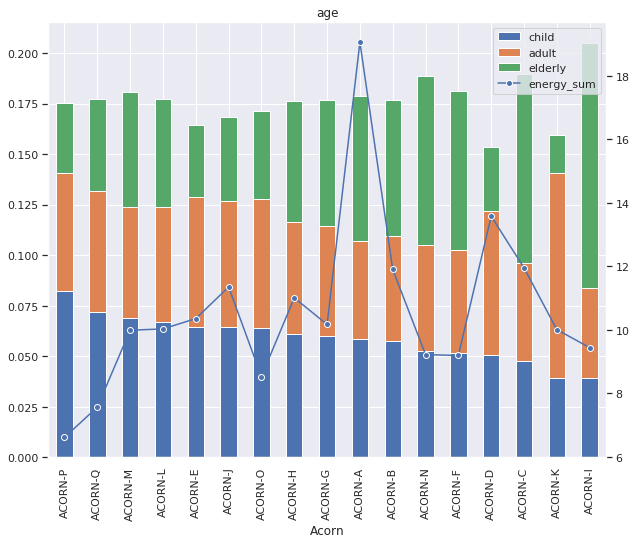
It is inferred that across each ACORN group energy consumption is more in the winter season. This is mostly due to the fact that a lot of high energy consumption is because of heaters, geysers and gas. Also, energy consumption is lowest in the summer season. This can be attributed to the fact that in the UK most of the households don’t have ACs and hence justifying low energy consumption in this season.

1. **Household Related Features**

Household related features include age, house-size and number of children in the house. The evaluation of correlation between these features and energy consumption is as follows:

1. Age

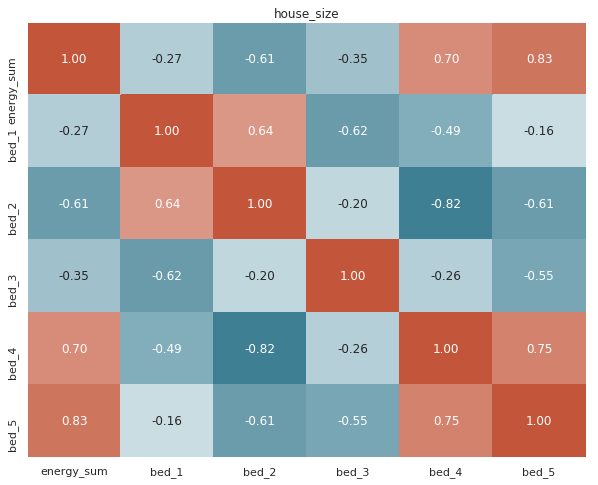
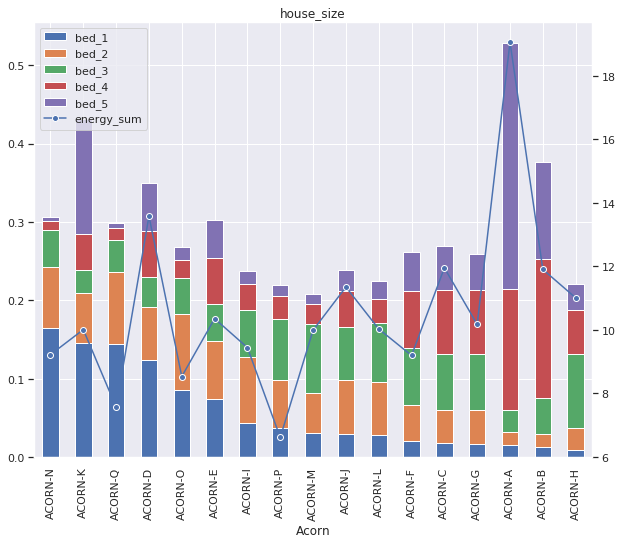
The correlation plot and the bar plot below shows that there is no significant correlation between age group and energy consumption.



**Figure-7**

1. House Size

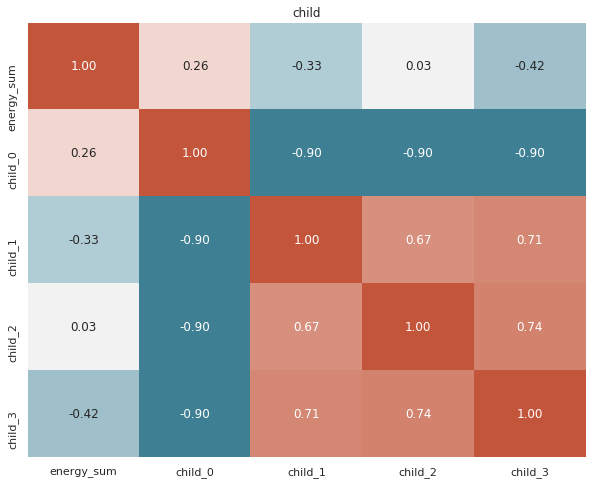
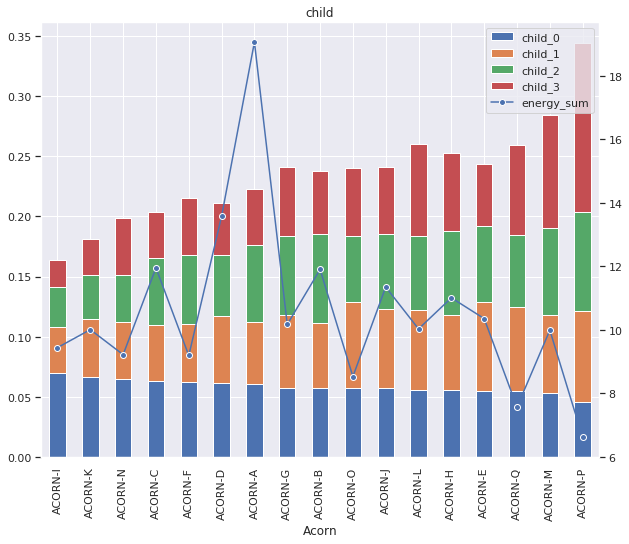
The correlation plot and the bar plot below clearly shows that more are the number of people with 4 or 5 beds, more is the energy consumption. This matches with our expectation.



**Figure-8**

1. Number of Children

The correlation plot and the bar plot below shows that there is no direct correlation between number of children and energy consumption.



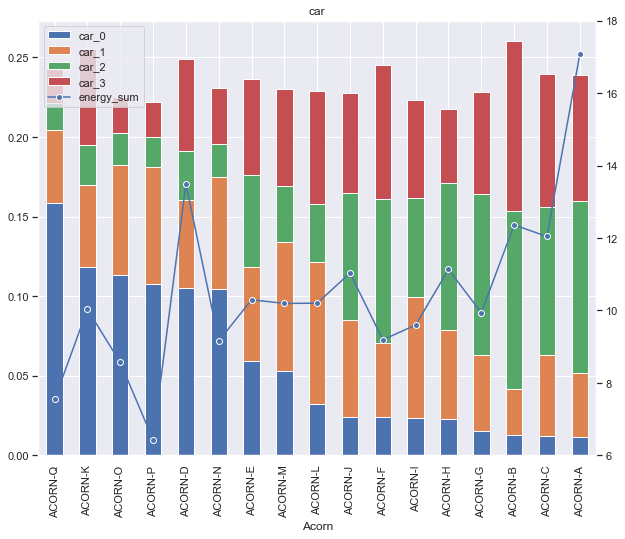
**Figure-9**

1. **Financial related features**

Financial related features include car ownership, income, and saving status. Based on the following bar plots, economical features (income, car ownership and financial situation) have a positive relationship with electricity consumption.

1. Car ownership

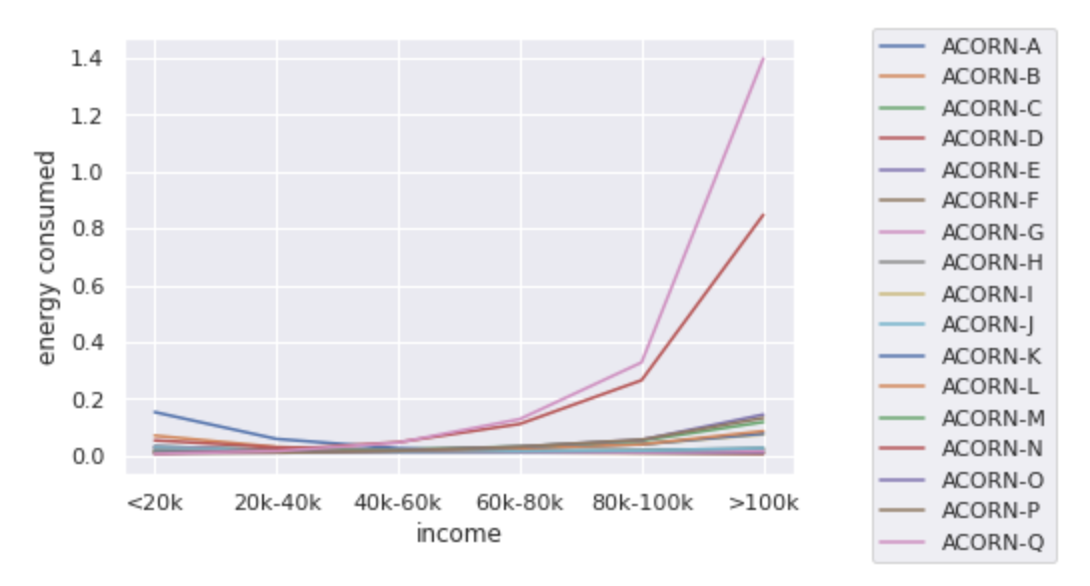
The higher ratio of households who own 3 cars, the more energy they tend to consume. This feature just indicates the wealth of the household. Thus, the more number of cars they own, the wealthier the households are and higher is their energy consumption.



**Figure-10**

1. Income

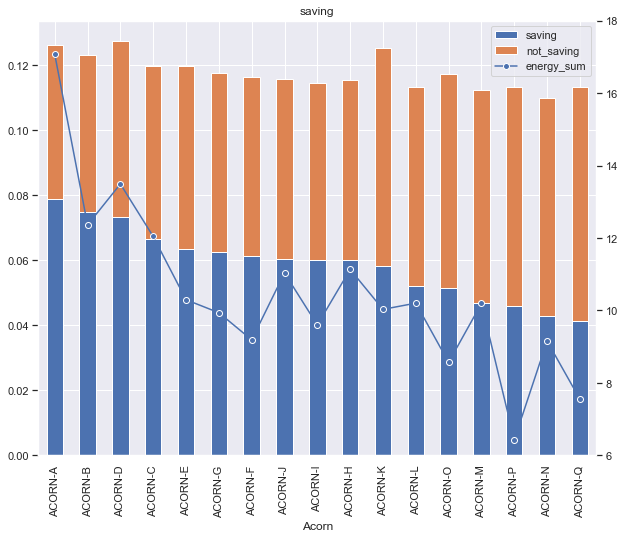
The plot below shows that more is the income, higher is the energy consumption. This behavior is expected as the households with higher income tend to be bigger with more electrical and luxury appliances, thereby, leading to higher consumption.



**Figure-11**

1. Saving status

The higher ratio of households who have savings, the more energy they tend to consume. This is opposite to our expectations. The reason for this could be that the ratio of households with saving attitudes is overpowered by the ratio of households with higher income leading to high consumption. Thus, even if the attitude considered alone as in the figure below shows negative correlation, we cannot conclusively believe that. The y-axis scale on the left showing the ratio of households with saving attitudes depicts that the ratio doesn’t vary much across acorn groups. Thus, the figure below does not illustrate a strong feature.



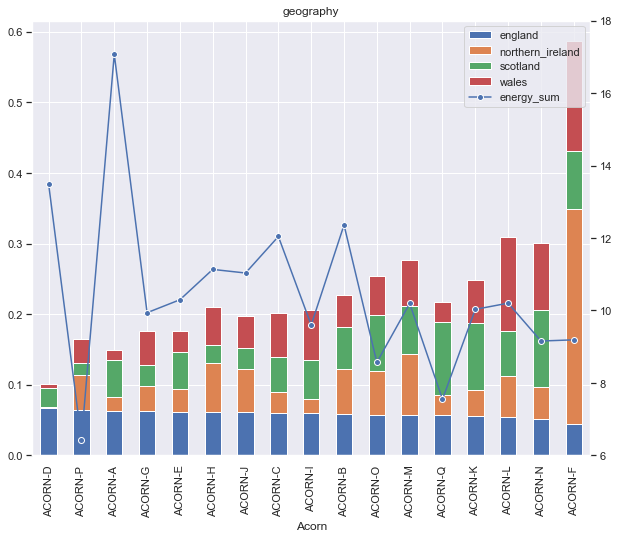
**Figure-12**

1. **Geography related features**

Geographical features in our analysis include their geo location and ethnicity.

1. Geo-location

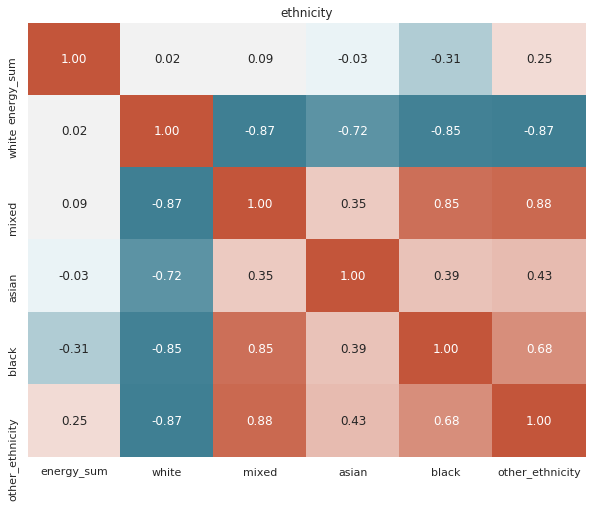
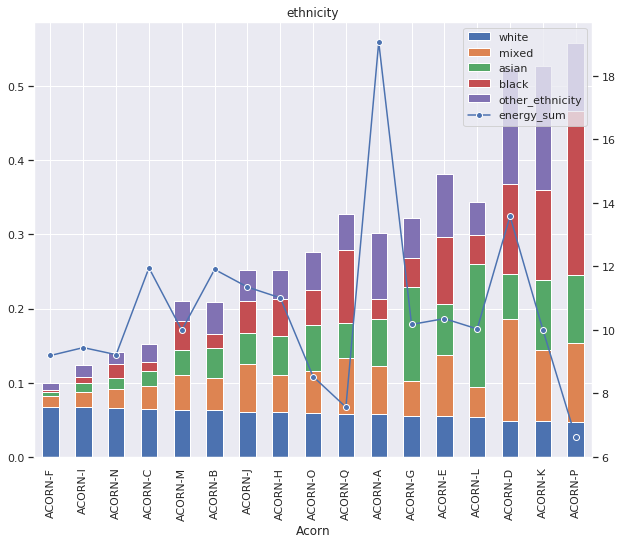
Based on correlation and the first barplot, geographical features do not have a strong relationship with electricity consumption.



**Figure-13**

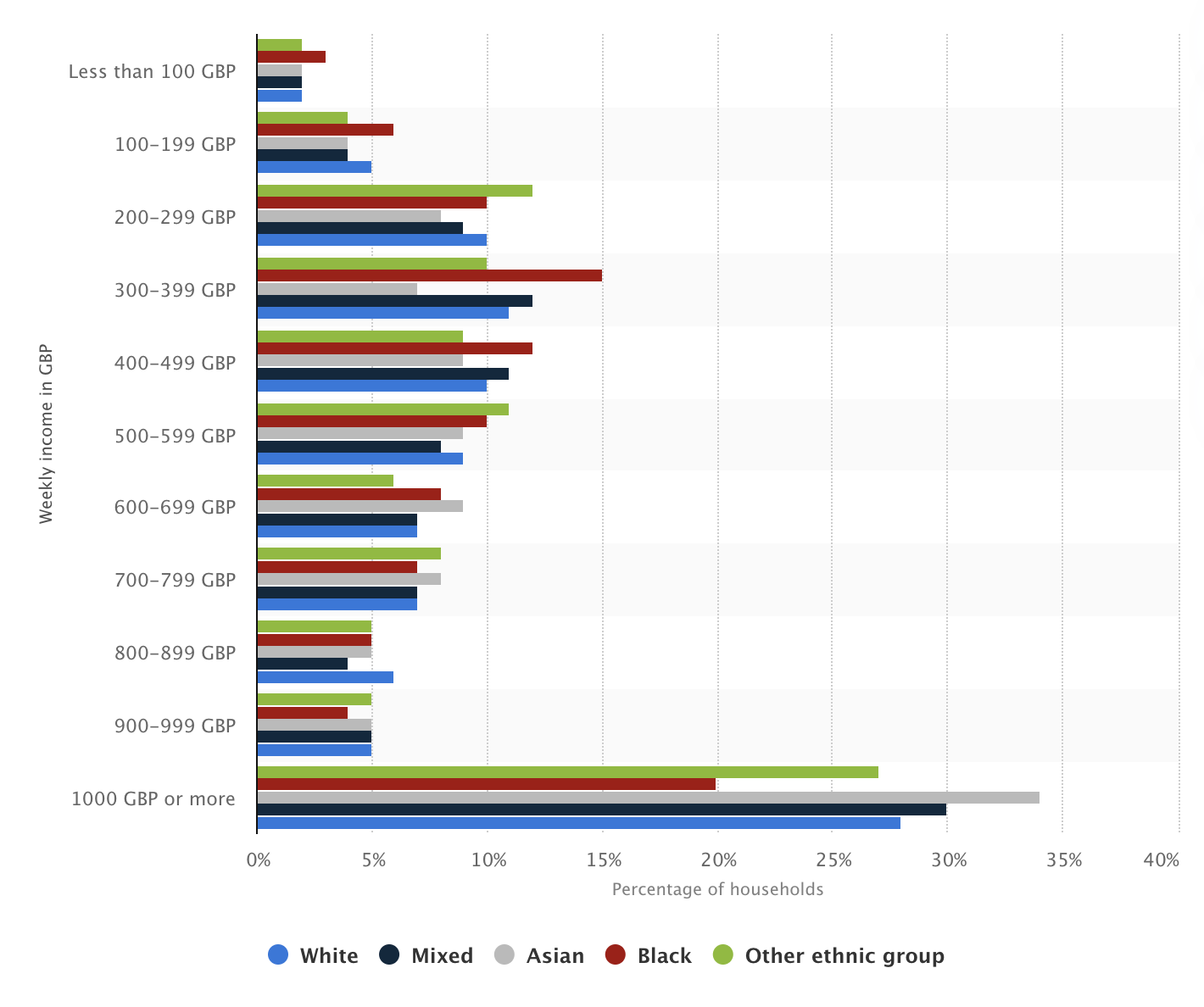
1. Ethnicity

The following plot shows the energy consumption vs the ratio of ethnic groups in each Acorn group. The correlation and the plot below show no significant correlation between ethnicity and energy consumption. However, there is a slight negative correlation between the ratio of blacks and energy consumption.



**Figure-14**

**Plot showing weekly income of UK from 2016/17 to 2018/19, by ethnic group of household head (**[**source**](https://www.statista.com/statistics/944052/household-income-by-ethnicity-in-the-uk-2017/) **)**



**Figure-15**

The above figure explains why blacks could have negative correlation with energy consumption. As seen in the plot, the number of blacks with high income is very less. Thus, we can say that most minority blacks have lower income and thus, lower energy consumption. This brings up the ethic pay gaps observed in UK and shows that ethnicity has no direct correlation to energy consumption.

**Takeaways**

After doing extensive EDA we realized that the most important features that influence the energy consumption are:

* Bank Holiday
* Season
* Weekends
* Income
* House size in terms of number of bedrooms
* Ethnicity

1. **Modeling**

There are 17 different Acorn groups after data cleaning. And daily energy consumption for each Acorn group is available in the data Starting from January 23, 2012 to February 28,2014. Time Series forecast model is fitted on each Acorn group to forecast average daily energy consumption in that Acorn group. 17 different Time Series forecast models, one for each Acorn group is fitted.

80% data is used for training the model and 20% data is used for testing. Data from January 23,2012 to October 31, 2013 is used as training data. Data from November 1, 2013 to February 28, 2014 is used for testing the model.

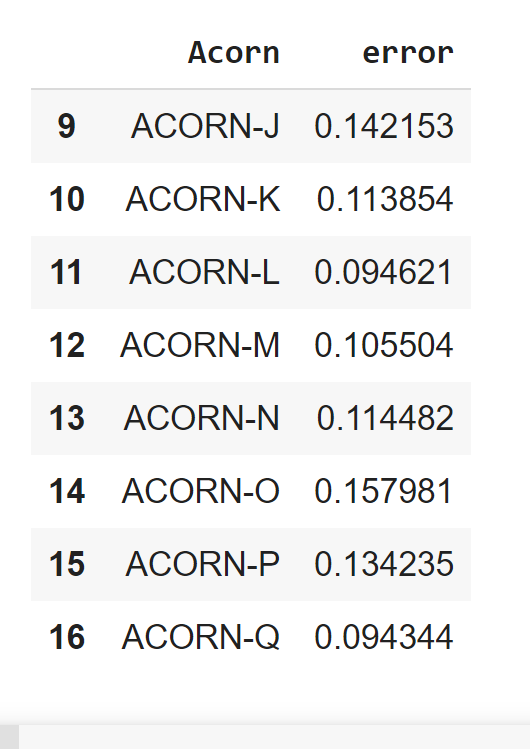
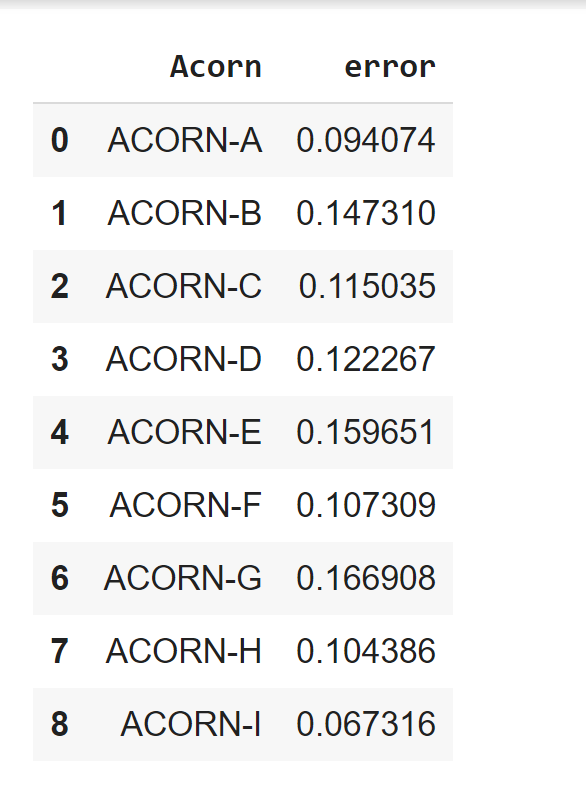
ARIMA models with exogenous variables are used for training. Exogenous variables used in this model are ‘Holiday’ and ‘Season’. Holiday is a binary variable indicating whether that date is holiday or not and Season will indicate which season that date belongs to.

ARIMA stands for auto-regressive integrated moving average. It’s a way of modelling time series data for forecasting (i.e., for predicting future points in the series), in such a way that:

* a pattern of growth/decline in the data is accounted for (hence the “auto-regressive” part)
* the rate of change of the growth/decline in the data is accounted for (hence the “integrated” part)
* noise between consecutive time points is accounted for (hence the “moving average” part)

Average percentage error is calculated for each Acorn group on testing data. In the future more data can be used to train the time series model which can improve the accuracy of the existing model. There is a bias in testing data, as most of the dates in testing data are from winter season, that’s why right now there is not a more robust way to measure the performance of the model.

Below are the results on the testing data:



1. **Inferences and Recommendation**

After performing exploratory data analysis we realized there are few important features that influence energy consumption. Based on these features we have thought of strategies that can be employed to reduce energy consumption and thus reduce carbon footprint and tackle climate change.

Suggestions

1. General suggestions
   1. Installation of solar heaters:The installment cost is high but from our analysis we inferred that high income groups consume more energy and thus they can afford to install solar heaters
   2. Hang laundry instead of using dryer
   3. Unplug unused electronics
2. We also observed through our study that people usually consume more energy on holidays as well as on weekends thus suggesting that they spend more time at home.
   1. Install programmable thermostat. Usually people have parties or gatherings at home on holidays. It is a fact that when more people visit your place due to the body heat the temperature of the room is higher than usual and hence the thermostat will sense that and control the temperature accordingly.
   2. Usually people have parties or gatherings at home on holidays. Use of paper plates can save on energy by not having to use dishwashers which consume a lot of energy.
   3. We can make sure that some campaign advertisements advocating smarter energy consumption are advertised more frequently on weekends and holidays.
   4. Use eco friendly decorations during holidays - have a timer for lights
3. For different Acorn groups: Based on our analysis we believe that Acorn groups A,D,B,C,H,J,E are the top Acorn groups that consume the most energy. We would like to target these Acorn groups and suggest ways by which they can reduce energy consumption. Based on certain behavioural and financial features, we infer the following:
   1. From Figure-1 we observe that, Acorn groups A, D, B,C, H, J and E don’t care too much about the environment, and thus consume too much energy. One solution to this would be to raise more awareness to these specific Acorn groups.
   2. From Figure-3 we can see that Acorn groups D, K, E, B, J and A use a lot of internet. These groups possibly use a lot of devices to connect to the internet or live stream/watch a lot of online shows or movies. This could cause high energy consumption. A recommendation to this would be to limit the use of streaming and using multiple devices at once to the internet.
4. Modeling: We can use our model to forecast daily energy consumption for each Acorn group. Hence, we can monitor the daily energy consumption of each household by comparing it with mean energy consumption of that Acorn group. This means that if the energy consumption of a household in an acorn group is more than the mean energy consumption of the acorn group to which it belongs then custom alert should be triggered and an SMS/email notification should be sent to that particular household. This will make sure that the people in the household are aware about how much energy they are consuming. It is often that we don’t realize how much energy we are consuming until we get the meter reading at the end of the month. This method will help spread more awareness and will be more effective as this will be targeted to an individual household.