ENERGY CONSUMPTION REPORT

Group 4

Abstract

This document contains information extracted from an energy consumption report gathered by SmartMeter technology of UK Power Networks. The report will contain findings for the exploratory data analysis, cluster and outlier analysis, and recommendations and conclusions.

Group Members

Yahya Haque 21100133
Saad Bin Waqas 21100241
Iman Aleem 21100245
Hamza Bin Aqeel 21100126

Table of Contents

Chapter 1: Exploratory Data Analysis

1. Loading the dataset and Preprocessing the Sample

The separated csv files were amalgamated into one data frame and saved separately. In this section, the sample dataset provided was reduced from **half hourly** to **hourly** readings with the help of grouping and aggregating by hour, effectively reducing the dataset by half. The sample dataset and tariffs were merged along the DateTime column to make tariffs per hour available in the same data frame.

Null and zero values in the **KWH** column (formerly named **KWH/hh (per half hour)**) were accounted for by replacing them with the mean of the hours taken from the sample dataset. Compared to choosing mode, median, or interpolation, the mean gave greater security of the value not becoming an outlier.

Furthermore, a mapping function was used to form a new column: 'Amount Paid'. The formula it deployed was as follows:

Amount Paid = KWH * Standard/Dynamic Rate

The standard flat rate and the dynamic high/normal/low rates were used from the provided guiding document.

	LCLid	DateTime	Year	Month	Day	Hour	stdorToU	KWH	Acorn	Acorn_grouped	Amount_Paid	Tariff	Tariff
0	MAC000002	10-12 00	2012	10	12	0	Std	0.331851	ACORN-A	Affluent	4.721575	Normal	Normal
1	MAC000002	10-12 01	2012	10	12	1	Std	0.269969	ACORN-A	Affluent	3.841117	Normal	Normal
2	MAC000002	10-12 02	2012	10	12	2	Std	0.237100	ACORN-A	Affluent	3.373455	Normal	Normal
3	MAC000002	10-12 03	2012	10	12	3	Std	0.223529	ACORN-A	Affluent	3.180368	Normal	Normal
4	MAC000002	10-12 04	2012	10	12	4	Std	0.222612	ACORN-A	Affluent	3.167328	Normal	Normal
...	...	...	...	...	...	...	...	...	...	...	...	...	...
46444309	MAC005565	06-21 04	2012	6	21	4	ToU	0.062000	ACORN-C	Affluent	0.247380	Low	Low
46444310	MAC005565	06-21 05	2012	6	21	5	ToU	1.150000	ACORN-C	Affluent	13.524000	Normal	Normal
46444311	MAC005565	06-21 06	2012	6	21	6	ToU	0.261000	ACORN-C	Affluent	3.069360	Normal	Normal
46444312	MAC005565	06-21 07	2012	6	21	7	ToU	0.025000	ACORN-C	Affluent	0.294000	Normal	Normal
46444313	MAC005565	12-19 12	2012	12	19	12	ToU	0.441254	ACORN-C	Affluent	1.760602	Low	Low

46444314 rows × 12 columns

Figure 1: The Final Dataset

2. Processing the Dataset

The techniques for preprocessing applied on the sample dataset were then applied on the main dataset, with a greater frequency of storage on drive to avoid loss of data cleaning (since the dataset was large).

3. Exploratory Data Analysis

The data analysis section constitutes the following sub-sections:

3.1. Calculation and Visualization of Summary Statistics

Descriptive statistics for numerical and value counts for categorical data on the data frame were calculated. Pie charts, box plots, and distribution plots relating to data items were displayed. The following features were noted:

- Patterns can be observed but exact values cannot because the dataset is large;
- The distribution of **KWH**, and consequently **Amount Paid** have early peaks in their distribution plots, because there are many larger but infrequently occurring values in the dataset.

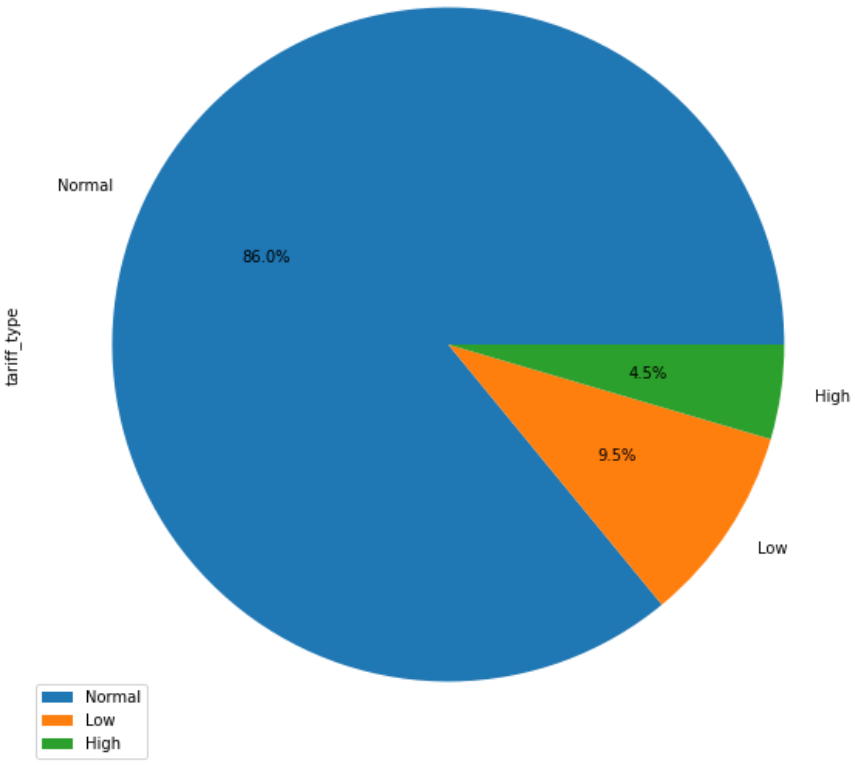

Figure 2: Weightage of Tariffs in the Dataset

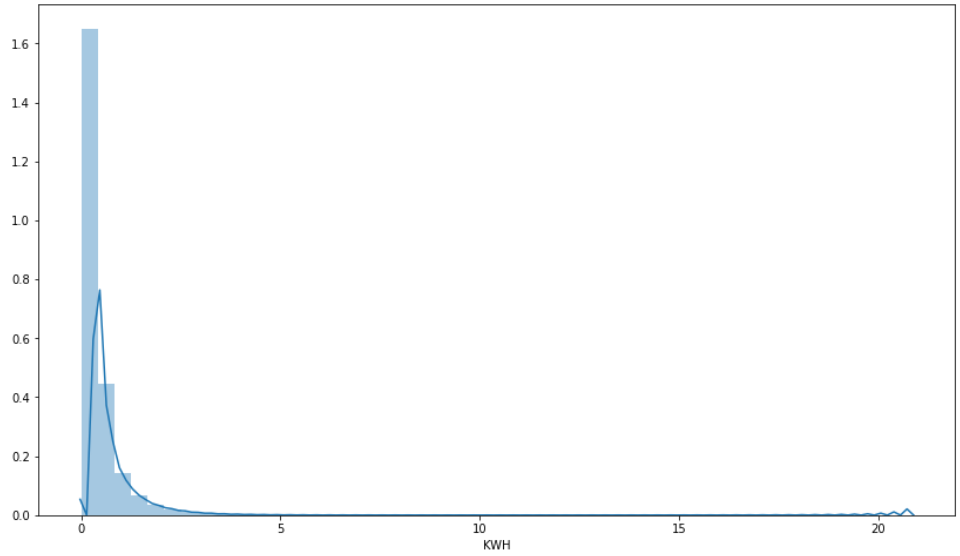

Figure 3: Distribution of KWH (Not Normalized)

3.2. Correlation and Dependence between Categorical and Numerical Values

Dictionaries were used to map tariff type, tariff rate, and group type with numerical numbers. The obtained correlation matrix and heat map help in concluding the following (irrelevant correlations such as between **KWH** and **Amount Paid** have been excluded):

- **KWH** and **Month** were slightly positively correlated, meaning that as the year progressed consumption increased;
- **Hour** and **Tariff** were slightly positively correlated, meaning that as the day progressed a higher tariff rate (where applicable) was imposed;
- **StdorToU** and **Acorn Group** were negatively correlated, meaning that as addresses went from Acorn- to Affluent, dynamic tariffs were imposed.

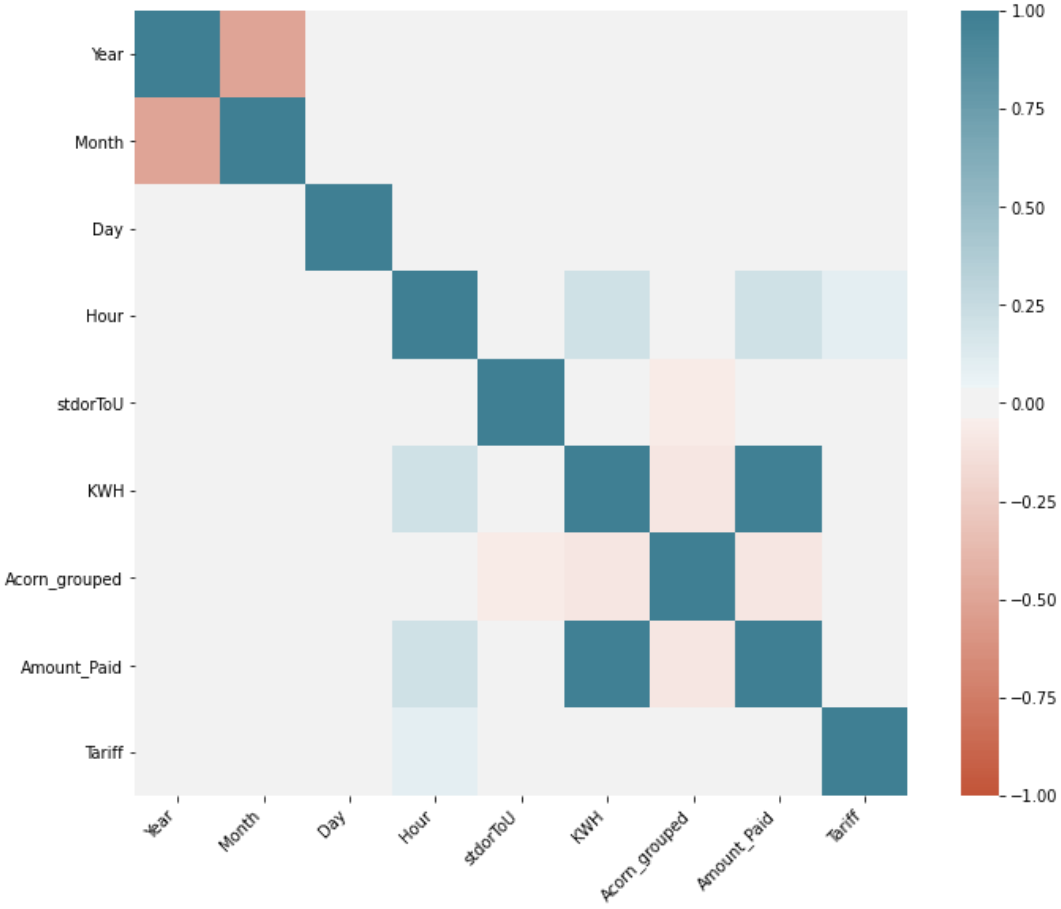

Figure 4: Correlation Heatmap

3.3. Average Energy Usage By Month

Note: **3.4** corresponds to the same analysis.

Data frame columns were dropped to accommodate only relevant columns for grouping. The grouping function was called on **Acorn Group, LCLid,** and **Month** respectively and the dataset was aggregated and summed to give consumption/amount paid per month of every household. The sums of all addresses were then summed again to produce the final data frame.

The graph plotted concluded the following:

- The monthly consumption between April and September is relatively stable for each group;
- September to December is a period of increase (with the exception of Accorn-);
- January to April is a period of decrease;

- **Accorn-** had a mean value of February's and April's consumption given to it for calculation because relevant data was missing from the dataset, which may impact further results as well.

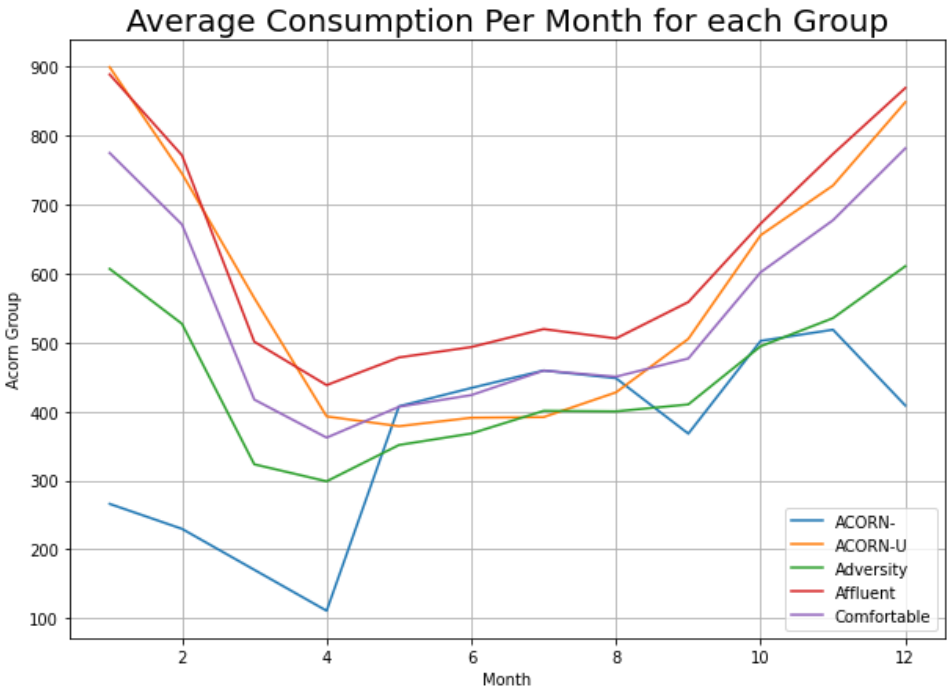

Figure 5: Average Consumption per month for each group

3.4. Covered Above

3.5. Average Energy Usage by Time of Day

Note: **3.6** corresponds to the same analysis.

A function was introduced to group the data frame by hour of the day, and the time setting was changed from 24 hour to 12 hour. Mapping was used to determine the average hourly consumption.

The consumption was plotted against the hour of the day and smoothened, giving the following conclusions:

- Around 9 am has the highest consumption time. This could be due to families leaving for work and school early in the morning and need to perform activities that consume energy, as well as other cleaning activities such as vacuuming the house;
- After 9 am the trend decreases. As all the residents leave for work and school.

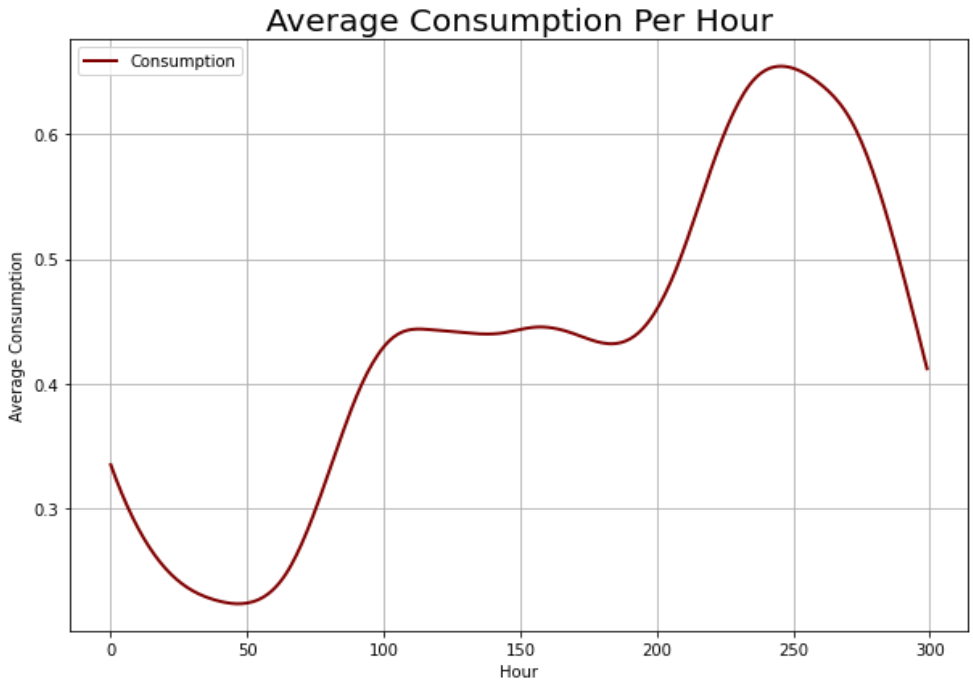

Figure 6: Average Consumption per hour

3.6. Covered Above

3.7. Average Usage By Weather:

A function was introduced to group the data frame by household tariff type. Then we grouped the months into seasons. With Winter lasting from December to February, Spring from March to May, Summer from June to August and Fall from September to November.

Then we checked the energy consumption for each group in these seasons. We found out that:

- The Std household tariff type is slightly higher than ToU tariff type;
- The highest consumption is during the season of winter. This may be because the dataset is from a region of colder climate hence the cost of electric heaters and heating up the house would be higher;
- The lowest consumption is during the season of summer. This might be because of the same reason as above. Due to colder climate summers aren't severe and it's easy to manage in those seasons without air conditioners.

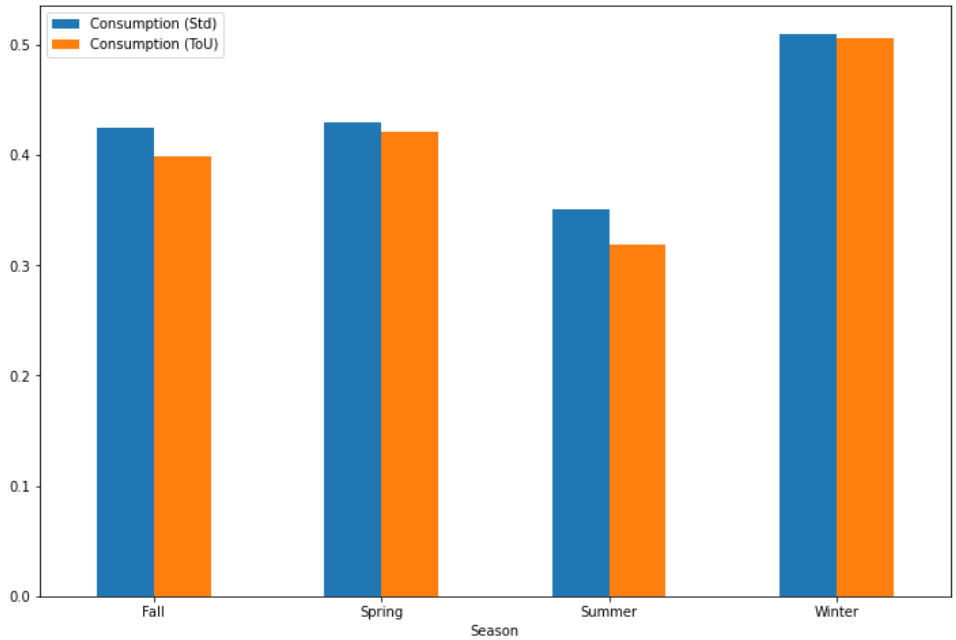

Figure 7: Average Consumption by Weather

Chapter 2: Cluster and Outlier Analysis

Clustering

We used K-Medoids method of clustering on our dataset since our data was in categorical form. Furthermore, data was clustered by weather and the following information was retrieved:

- Fall = 3129
- Winter = 3134
- Spring = 3112
- Summer = 3117

Each number depicts the number of households that were found for each weather.

The data was first aggregated according to household type and weather. This gave us a clean dataset to work on with all the customers and their consumption per weather. From that, we got different groups of data corresponding to each weather.

The clustering obtained based on the seasons is as follows:

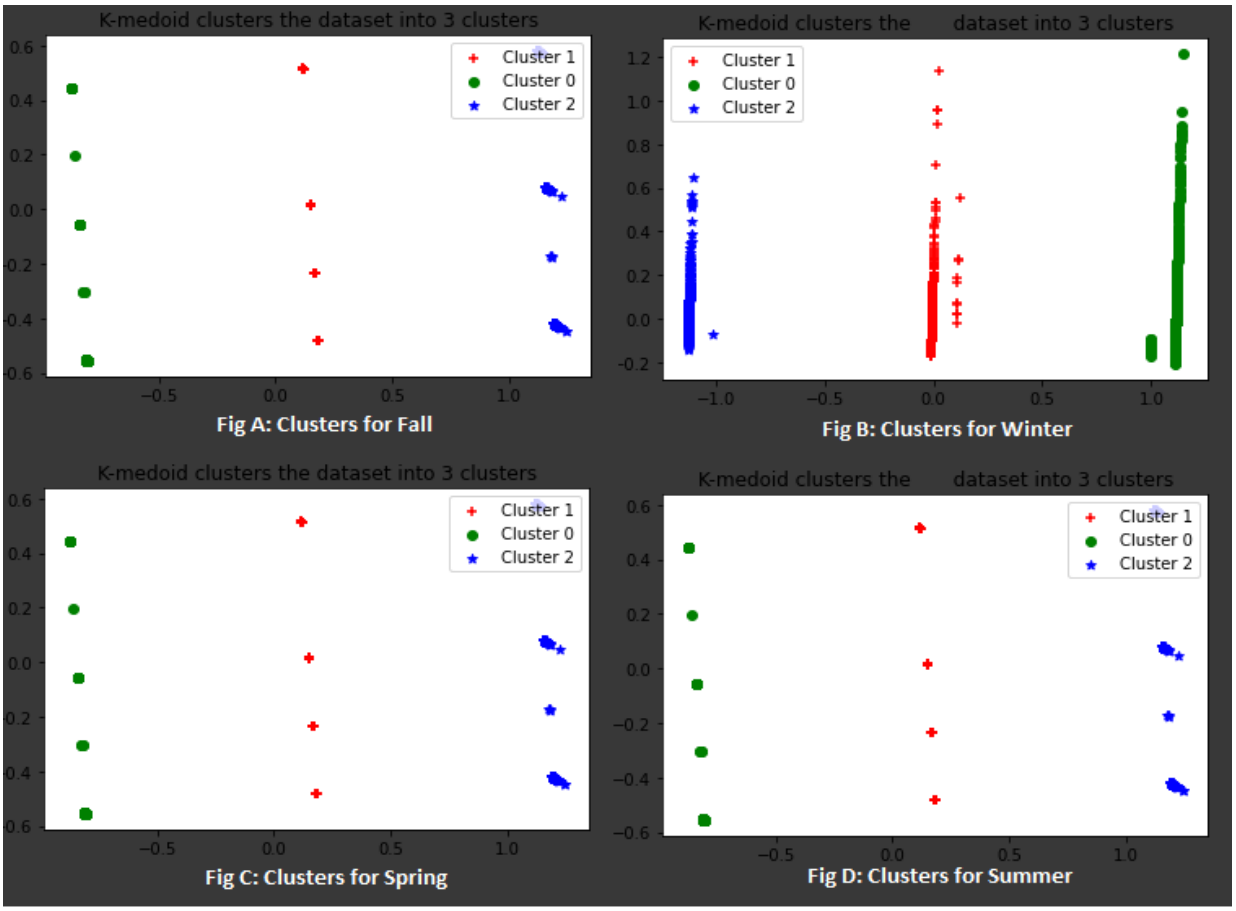

Figure 8: Clustering based on weather

The following factors explain the clusters which were obtained:

1. Principle Component Analysis was used for ease of visualization by reducing the dimensionality.
2. KWH was the only attribute which had a continuous distribution, and the remaining attributes were either dependent on KWH (Amount Paid) or were discrete, therefore naturally the number of clusters obtained through K-medoids was **3.**
3. Houses of types ACCORN- and ACCORN-U were very less in number (1 and 21 respectively) compared to the other three types, which is why they were clustered with the other three types and attempts to visualize them separately were not successful.

Outlier Analysis

We used Local Outlier Factor (LOF) for our outlier analysis because the dataset given does not fluctuate in density very frequently, which means that a local strategy is important for detecting outliers among potential neighbors.

Optimal number of outliers were selected based on a “Hit and Trial” method where graphs were visualized with a range of different neighbors to consider. The neighbors which gave the lowest values with appropriate and reasonable visualization were selected.

Based on the criteria above, the following optimal neighbors were obtained for each season:

- Fall: 10
- Winter: 40
- Spring: 50
- Summer: 40

1. Outliers for Fall

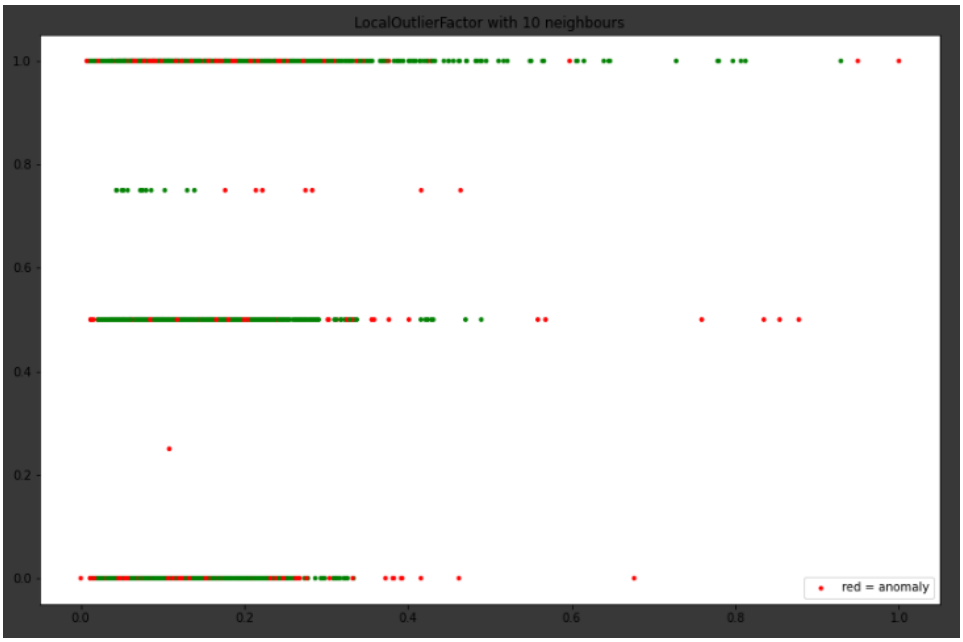

Fig 9: Fall LOF analysis

% of outliers in Fall = **10.45%**

2. Outliers for Winter

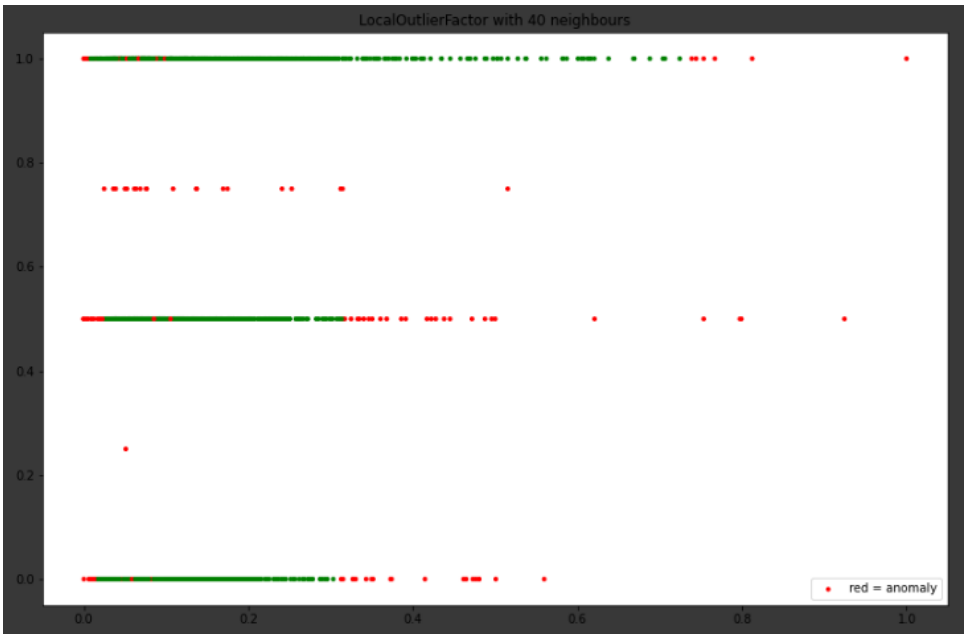

Fig 10: Winter LOF analysis

% of outliers in Winter = **5.74%**

3. Outliers for Spring

Fig 11: Spring LOF analysis

% of outliers in Spring = **9.96%**

4. Outliers in Summer

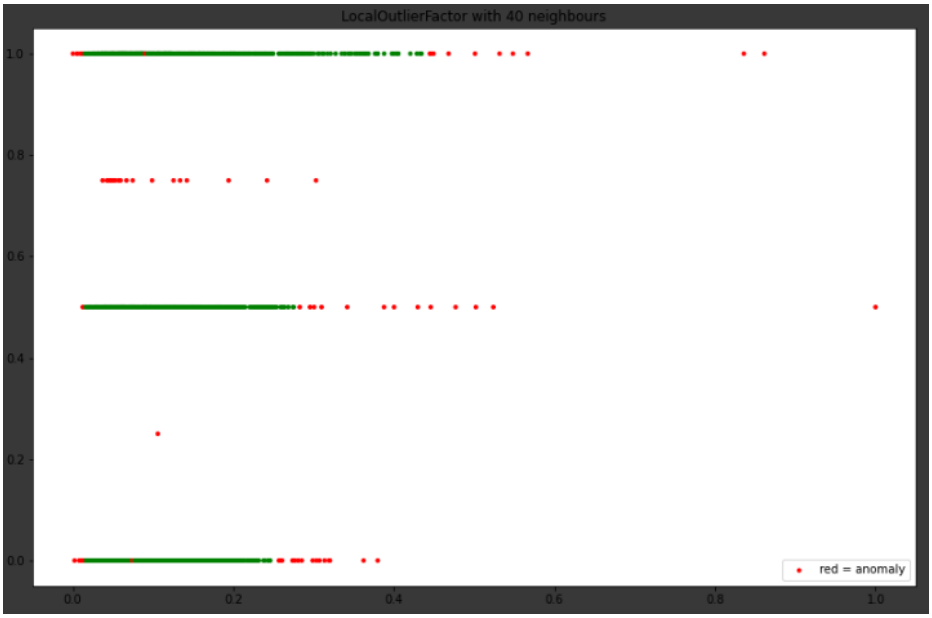

Fig 12: Summer LOF analysis

% of outliers in Summer = **5.42%**

Chapter 3: Recommendations and Conclusions

Conclusions:

1. Insufficient data lead to 3 natural clusters being visualized, since Acorn- and Acorn-U did not have sufficient values for separate clusters after Principle Component Analysis.
2. Most of the outliers shown in the visualizations belong to relatively small clusters of Acorn- and Acorn-U, and hence are not exactly outliers but values which can be clustered to higher/lower KWH consuming clusters.
3. Average consumption for standard tariff households was greater than dynamic tariff households, which is a claim supported by figure 7. However, it is to be noted that households with dynamic tariff type were much less in number.
4. Cold climate in UK makes greater consumption of energy in the winters necessary, which is supported by figure 5.
5. From figure 4 it can be seen that the highest consumption of energy is from 15:00 – 20:00 every day, and our data supports that this is the time where dynamic tariff type households have their tariffs changed to a high rate.

Recommendation

Based on the conclusions made above, a major recommendation for dynamic tariff households can be drawn. Since their average consumption per hour is maximum between 15:00 – 20:00, and this consumption is highest in Winters, these households should invest in low energy consuming heating devices to avoid being taxed highly in peak consumption times.