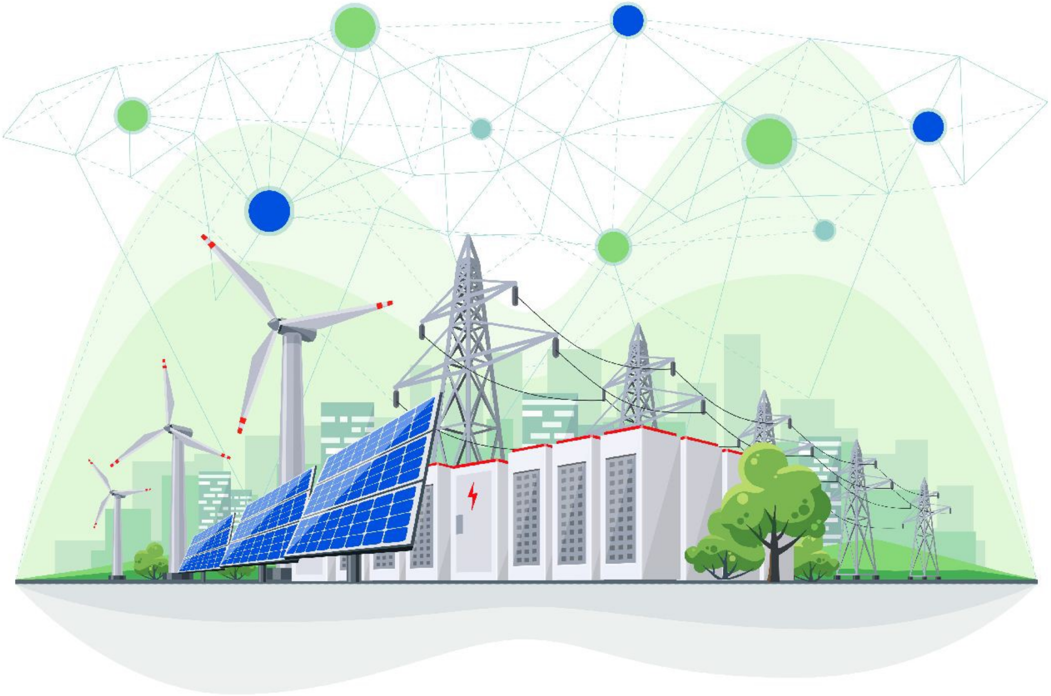


Energy Consumption Report

UK Power Networks and Smart Meter Technology

Group# 02
Mohammad Hamza - 20100135
Umer Farooq Zia - 20100148
Dilawer Ahmed - 20100177
Muhammad Hassan - 20100145

Chapter 1
Exploratory Data Analysis

Introduction

We were provided with the dataset of UK power networks gathered using SmartMeter Technology. The data set is gathered from 5,567 households between November 2011 and February 2014.

Dataset

The initially provided dataset contains 168 .csv files each file containing approximately 1 million rows, and the total dataset was around 10GB after unzipping. Later on due to computing issues, the numbers of files were reduced and 93 files were sampled for the analysis at this stage.

The dataset has five attributes, namely energy consumption in kWh (per half hour), unique household identifier, date and time, Acorn group, and tariff type. Meter readings were taken at half hourly intervals mostly, but there are a few exceptions in the dataset where erroneous readings were taken at irregular intervals. Such readings usually had NaN value and hence were replaced with ‘zero’ value in the processing stage. During the next phases of the project, we will see if these values need to be discarded or we can work to replace the null values with zeros.

Within the data set are two groups of customers. The first is a sub-group of customers who were subjected to Dynamic Time of Use (dToU) energy prices. Customers were issued following prices

1. High (67.20p/kWh)
2. Low (3.99p/kWh)
3. Normal (11.76p/kWh)

Prices were applied differently at different times of the day. The second sub-group of customers is Standard customers who were on a flat rate tariff of 14.228p/kWh.

Customers have been divided into different Acorn groups based on customer income. There are five different categories of Acorn groups, mentioned below:

- Adversity

- Affluent
- Comfortable
- ACORN-U
- ACORN-

Purpose of Document

This is an open ended phase of the project where we are required to perform Exploratory Analysis on the dataset. The goal is to find patterns, get acquainted with the data, and methods to look for meaningful inferences from the data.

In this chapter, we will discuss the results of some primary data exploration and analysis techniques and will try to draw some conclusions based on these results. In later stages of the project and in next chapters, we will come back to these hypotheses and their validity.

Basic Statistics and Correlation

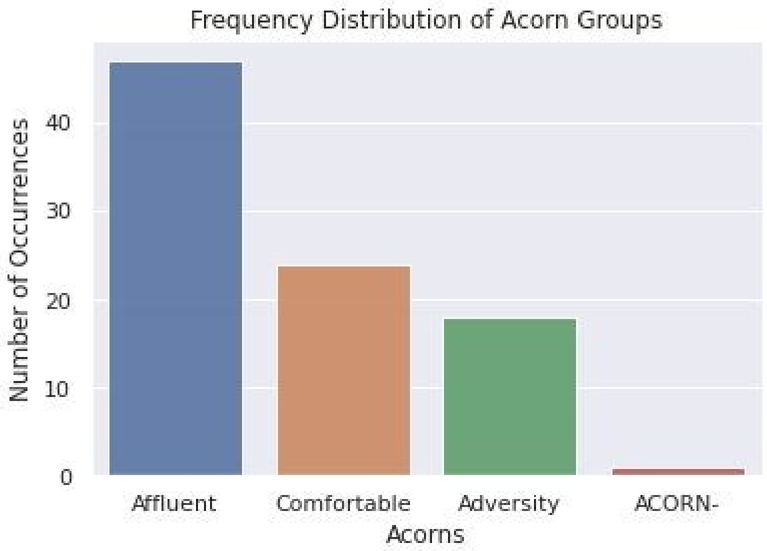
We divided the given dataset into two. One dataset of standard users and another dataset of dynamic users. Summary statistics of both types of users are provided in Table 1.

	Standard	**Dynamic**
Count	89,996,960	2,932,384
Mean	0.214	0.206
Standard Deviation	0.292	0.292
Minimum Value	00	00
25%	0.060	0.053
50%	0.120	0.107
75%	0.243	0.234
Maximum Value	10.761	6.162

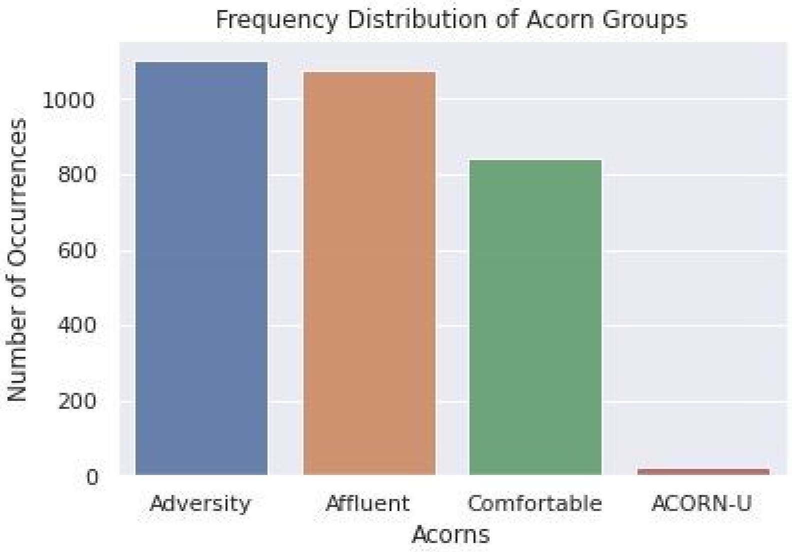
Table 1: Summary statistics of the different user tariff types

As discussed in the introduction of the dataset, households are divided into acorn groups of 5 kinds. On further analyzing the data, we find the following division of Acorn groups among customers.

Dynamic priced customers distribution into acorn groups is shown in the graph below:

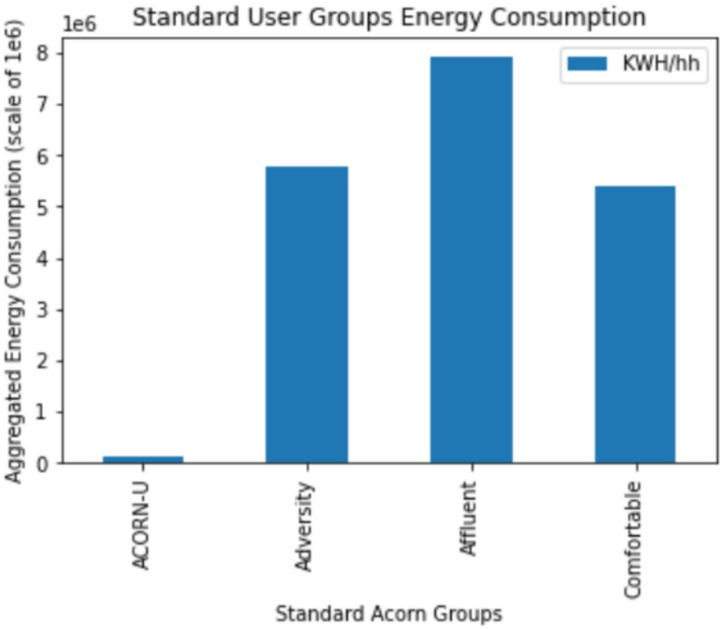


The standard priced customers had following distribution into Acorn groups as shown below.

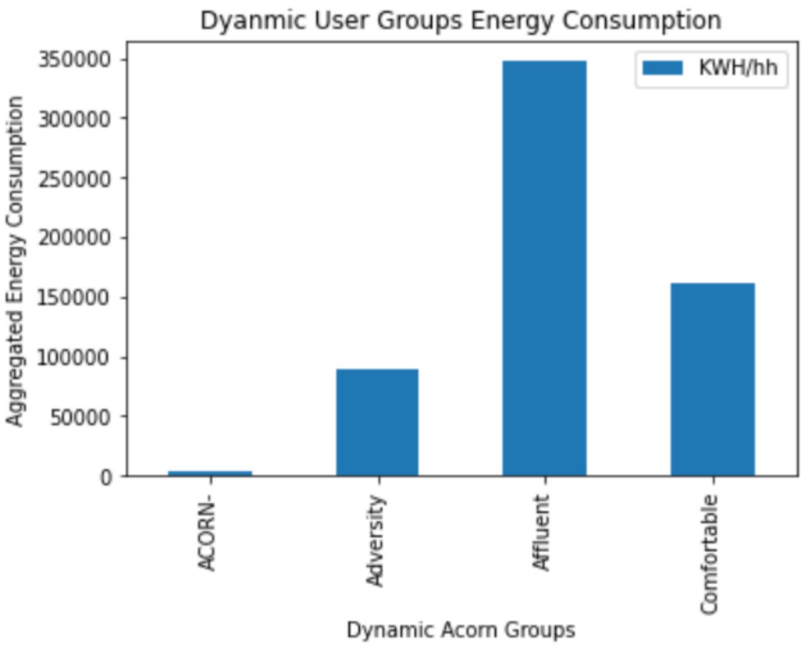


From the graphs, we can infer that the majority of customers who are asked to pay the dynamic prices belong to affluent and comfortable classes, while adversity and affluent classes are dominant in standard pricing. From a business point of view, this is an intelligent move considering that Affluent and Comfortable classes would be more willing to pay the dynamic prices, since the dynamic prices fluctuate to considerably very high value. While on the other hand, if we have a closer look at the distribution of Acorn groups into standard pricing, we find out that although ‘Adversity’ is dominant, almost ⅔ of the standard pricing customers are either ‘Affluent’ or ‘Comfortable’ households. Perhaps, in order to increase the revenue, the electricity providing company should try to shift some of these households to dynamic pricing. This can serve two purposes, reduce the load during the time when demand is high and also increase the revenue by dynamic pricing from a larger group.

If we analyze the total consumption of each pricing plan, we observe the following pattern for Standard Pricing customers. We observe that the Affluent are third most dominant in the number but their consumption is highest among all.



Below this paragraph is the total consumption graph for Dynamic Pricing plan. We realize that the total consumption of ‘Affluent’ households is greater than the combined consumption by the rest of the households in this category.



Analysis Across Energy Usage

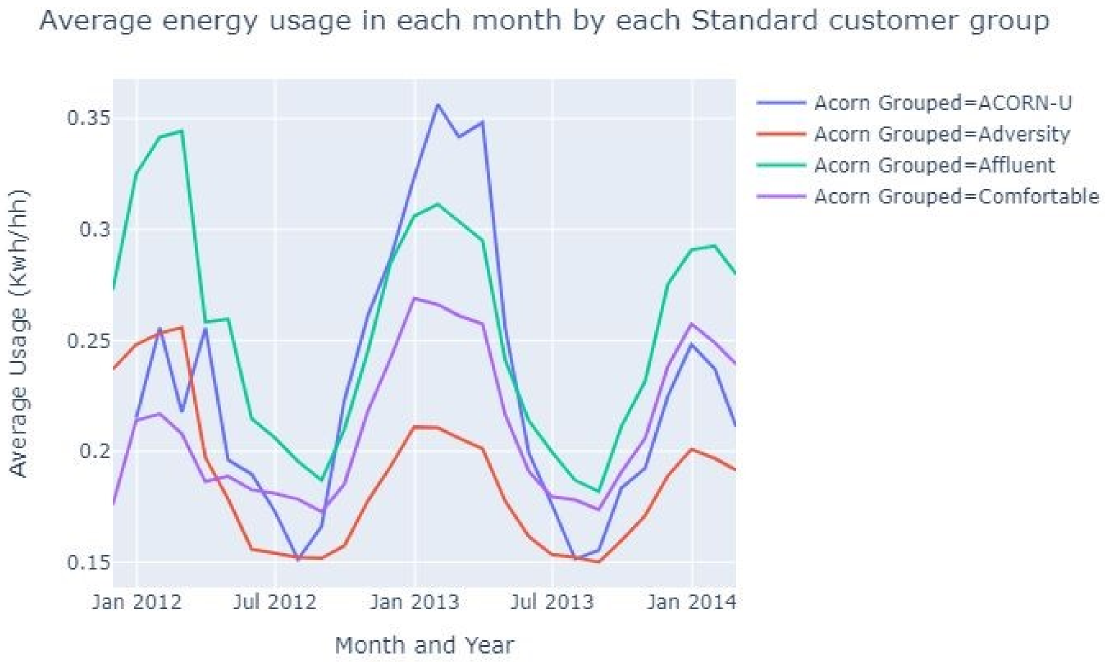
We have analyzed the dataset and observed the trends of usage of energy on many levels. We have done the following analysis on energy usage.

- Average usage in each month by each customer group
- Average usage by time of day
- Average usage by day of week
- Average usage by weather in DToU and STD group
- Average electricity around Christmas and New Year

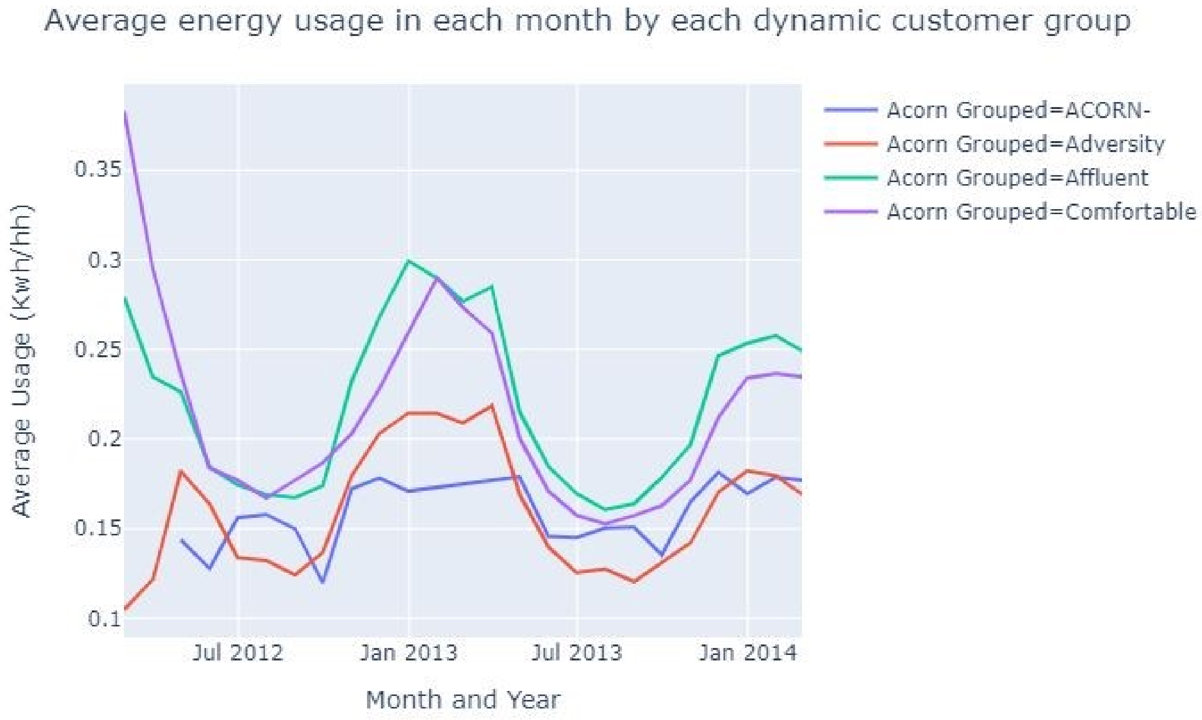
Average Usage In each month by each customer group

The following graph shows the average monthly usage of both groups. An interesting thing that can be observed is that the average usage is linked to how well off a group is. The most economically stable population i.e. Affluent have more per month usage compared to the Adversity group. This can be due to economically stable households being larger in size and have more electrical appliances than the other groups.

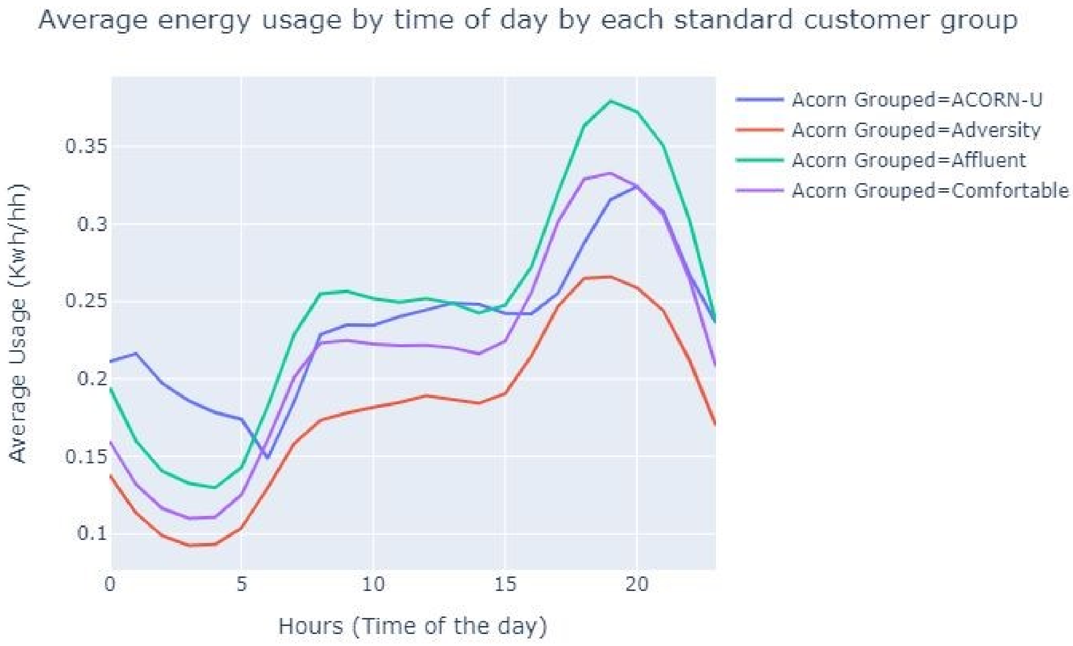
Standard Users:



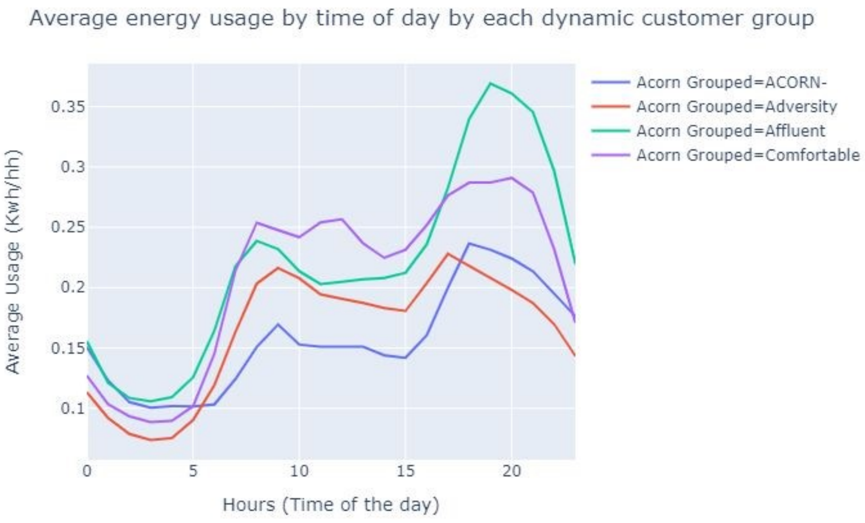
Dynamic Time of Use Users:



Average Usage By Time Of Day


Average energy usage by time of day by each standard customer group

The graph above shows the average usage per time of day for standard users. The trend follows the expected value for household use. Usage is low during night time from 11pm to 6am. Then it increases and stays constant from 7am to 3pm. At evening time it starts to increase, reaching peak at around 7-8pm. This graph can produce important insights for electricity production. Power plants can be run and their production can be adjusted accordingly. Electricity rates can be adjusted accordingly as well to flatten the curve. The same trend is followed for dynamic users.


Average energy usage by time of day by each dynamic customer group

Average Usage By Day Of Week

The data was analyzed to observe the average usage with respect to weekdays. We want to see if usage is more on weekends or working days.

Here is the mapping of the days:

0 = Monday

1 = Tuesday

2 = Wednesday

3 = Thursday

4 = Friday

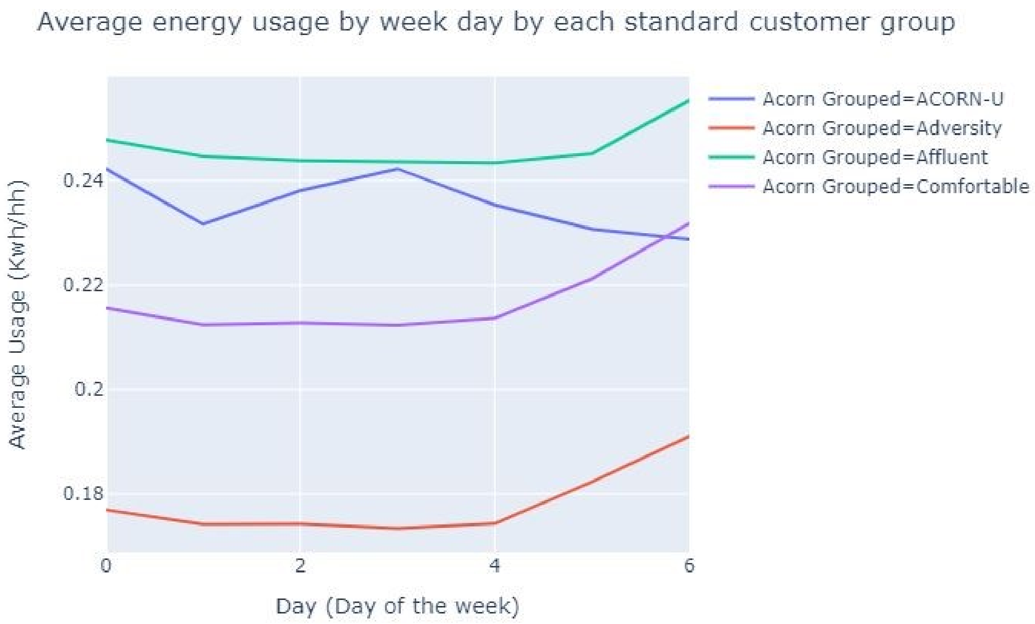
5 = Saturday

6 = Sunday

The graph for each customer group is given below. We found out that energy consumption was relatively constant throughout the day, but during the weekend there was a spike. Since this data is from SmartMeters in UK households, we can infer that the spike has been caused by people spending their weekends at home and not going to work places.

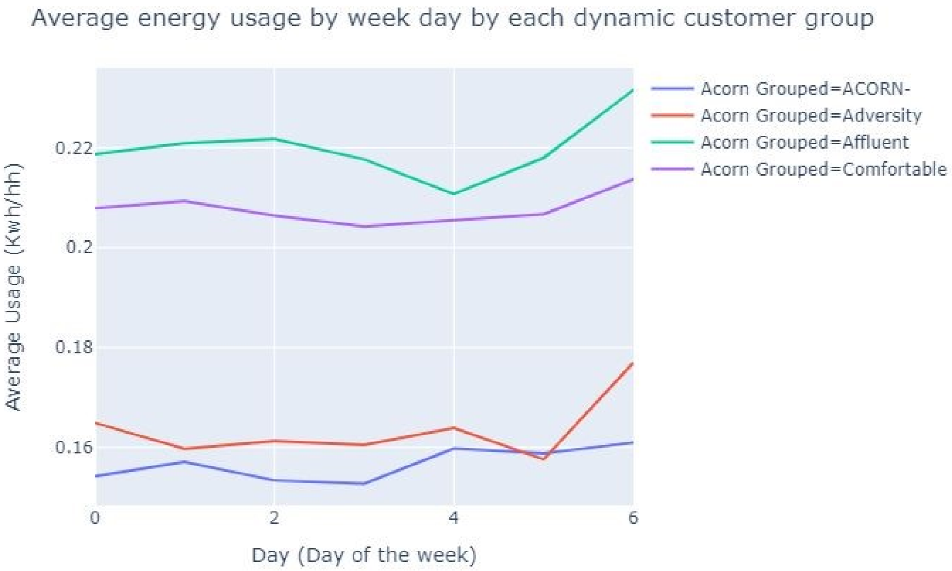
Further, we can infer that the graphs correlate with the economical question that Affluent and Comfortable households are spending more on electricity as compared to other classes.

Standard Users:



Dynamic Users:

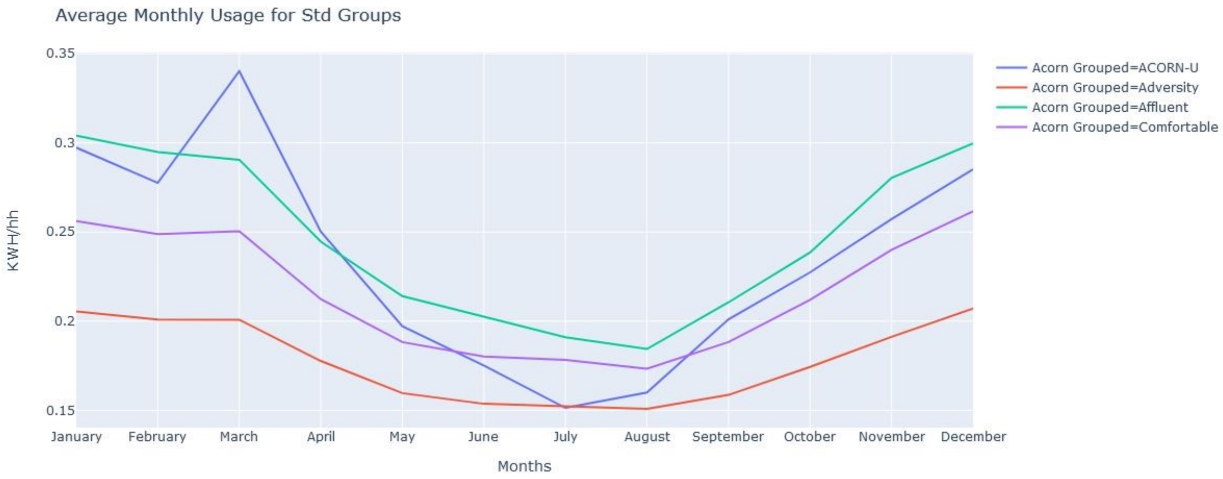
For the dynamic user data, We were able to see a large difference between usage of Affluent, Comfortable and Adversity and Acorn-. We observed the same trend of increased usage over the weekend.



Average usage by weather in DToU and STD group

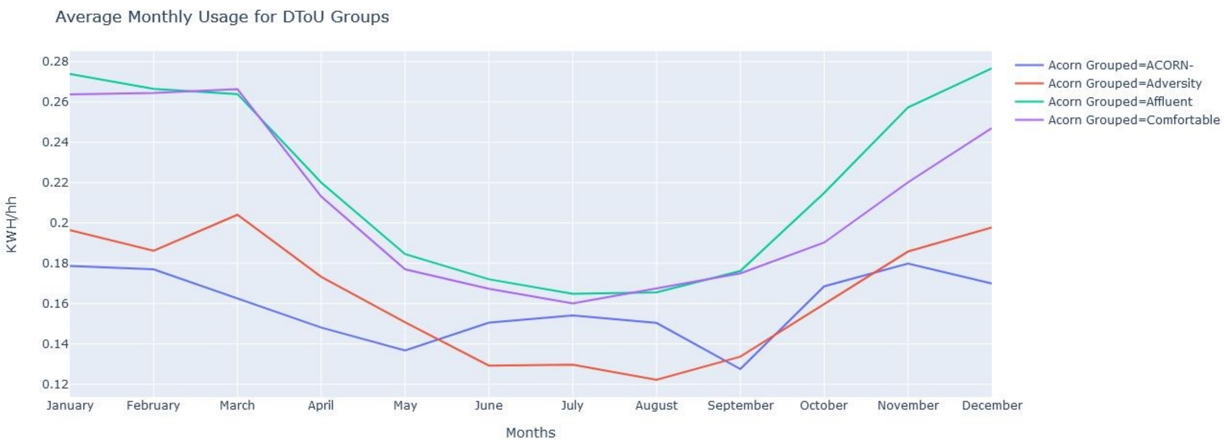
There are four seasons in UK (source):

- Spring: March - May
- Summer: June to August
- Autumn: September - November
- Winter: December to February



In the above figure, the graph is for average "KWH/hh" usage per "LCLid" for a month (STD Group). This group has four "Accron_grouped" in it, namely: "ACORN-U", "Adversity", "Affluent", "Comfortable". As expected, "KWH/hh" follows a trend of seasonality for all "Accron_grouped". During winters, the average usage zeniths and steadily decreases as moving towards summers. During summers, the average usage is the lowest. During spring and autumn, the average usage is similar. One conclusion that can be drawn from this is that electricity is used for heating purposes more during the winters. Correlating this with winter and summer conditions in the UK, where winters are very cold and harsh, and summers are very pleasant.

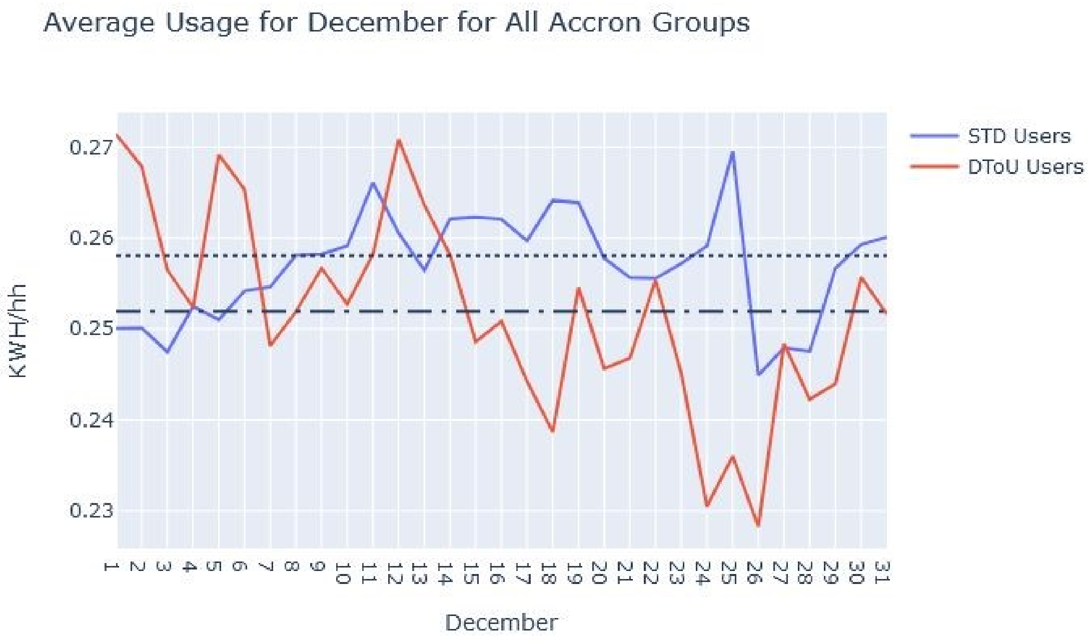
Up next is the graph for average "KWH/hh" usage per "LCLid" for a month (DToU Group). This group has four "Accron_grouped" in it, namely: "ACORN-", "Adversity", "Affluent", "Comfortable".



DToU also follows a similar usage pattern to STD i.e. the trend of seasonality in the electricity usage.

Average Electricity around Christmas and New Year

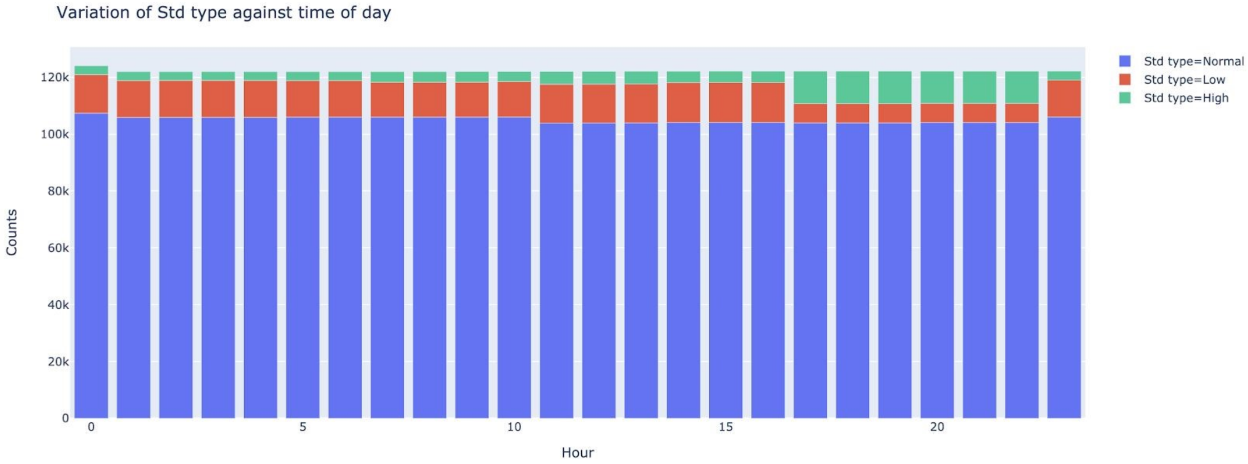
Following are the graphs for daily average and monthly average in December for DToU and STD groups. The **dotted line** represents the mean electricity usage for the month of STD groups (0.258 KWH/hh), whereas the **dashed line** shows the mean electricity usage (0.252 KWH/hh) by DToU groups in December.



Looking at this graph, it is difficult to say that the average daily electricity usage strictly follows some trend in the month of December. Although in the case of DToU, the usage was reduced in the second half of December, i.e. from 16th December to 31st December. This could be a result of people getting together (electricity usage falling for some houses) for Christmas and New Year events.

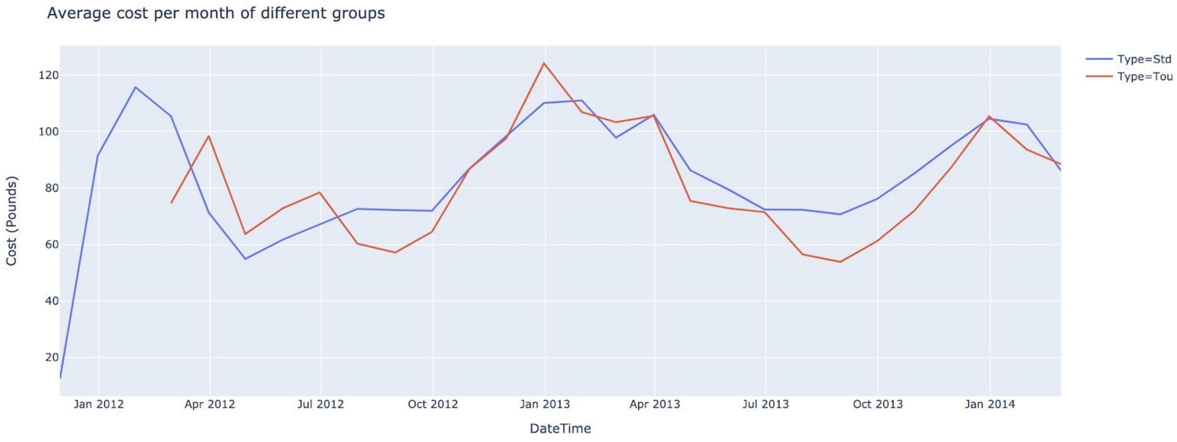
Tariff types of Dynamic Customer by time of Day

The stacked bar graph above shows the tariff types with respect to time of day. As can be seen mostly normal tariffs are in place. Around 6-10pm high tariffs are relatively more charged as compared to the rest of the day.



Comparison of Average Cost of Standard and Dynamic Customer

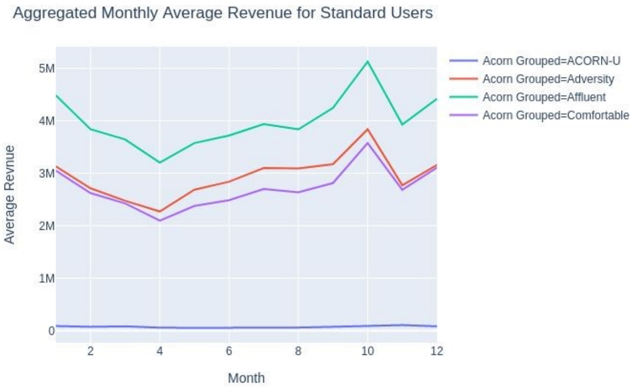
The graph does the average monthly cost per customer in UK Pounds for Standard tariff and Dynamic Time of Use Tariff. There is no clear indication of if one is better than the

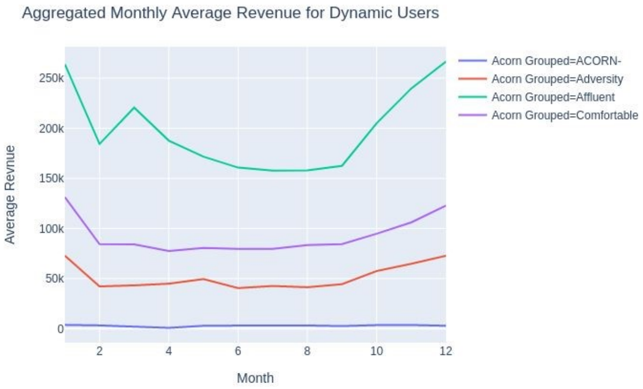


other but overall Dynamic tariffs are ever so slightly lower compared to standard. For each customer, the amount of electricity cost is calculated as following:

cost = (kwh_consumed_high * cost_of_high) + (kwh_consumed_low * cost_of_low) + (kwh_consumed_normal * cost_of_normal)

Analysis of Revenue Generation


Aggregated Monthly Average Revenue for Standard Users


Aggregated Monthly Average Revenue for Dynamic Users

In this section we look at the average revenue generated every year. On a close look at the standard user group, we find the following information.

- The revenue generated is lower in summer as compared to winters.
- The revenue is not constant across any point and keeps varying constantly for standard customers.
- The revenue is lowest in April and highest in October.
- The highest revenue is generated from affluent households, followed by adversity, comfortable and acorn-u households.
- The lowest revenue is generated from the ACORN-U households, owing to their low numbers.

Now if we look at the revenue analysis of the dynamic households, we infer the following information.

- Similar to standard customers, the revenue generation is high in winters as compared to winters.
- The revenue is constant across the summer season, i.e from April to August.
- The revenue generation is maximum at the end of December and at the start of January and lowest in summer months.
- The affluent households generate maximum revenue, followed by comfortable, adversity and acorn- households.

- The acorn- households generate almost constant revenue throughout the year.

Summary of Exploratory Data Analysis

- Affluent and Comfortable were spending more in terms of electricity cost as compared to others.
- Affluent were spending more in standard connections as well, the company can transfer them to Dynamic usage and try to maximize the profits. This is one of the hypotheses drawn from current inferences, we can only verify this in later stages where we will have better evidence for such business decisions.
- We tried to correlate categorical data through label encoding of categorical features, further we employed Pearson and Kendall methods in Python, but we could not find any strong correlation. Given this scenario, we referred to graphing our findings to correlate different attributes.
- The energy consumption was greater in winters as compared to summers, and is possible due to weather conditions in the UK.
- The energy consumption was greater on weekends as compared to weekdays
- The energy consumption was greater from 7 pm to 11 pm than the rest of the day.
- During holidays like New year and Christmas, the energy consumption has decreased. Maybe it's due to the fact that people go and visit their friends and relatives during this time period.
- Dynamic customers are being charged at Normal for most of the time but around 6pm to 10 pm high tariffs are relatively more charged as compared to the rest of the day.
- The average cost of Dynamic and Standard customer per month was relatively similar, we could not find any interesting points from it.

Chapter 2
Clustering and Outlier Analysis

Introduction

Traditionally, from the literature review on clustering analysis, we found out that the clustering was done mostly on the energy data points but lately the clustering techniques have started to incorporate other relevant factors of households as well. The need for such methods arise from the fact that there are more complex factors now involved in determining the possible energy demands of various households and customer types. This is more relevant since the energy demand varies very dramatically owing to the multiple factors including, but not limited to, season of the year, costing and revenue, pricing package and time of the day.

Smart metering is increasing in the energy sector since it provides better and more accurate insights; in a relatively short span of time. The data gathered from smart meters is analyzed in this section, since clustering customers based on attributes derived from the smart meter will create better understanding of the behavioral energy groups in the given dataset.

Clustering Analysis

One of the main objectives of the smart metering is to derive meaningful and useful results from the information which is presented. The aim of clustering applications on smart meter electricity dataset is to place data elements, i.e customers, into their related groups. This is done by partitioning the data into classes based on different attributes using a clustering algorithm.

K Means Clustering

In our analysis, we have applied the K-means clustering algorithm, which is an unsupervised machine learning algorithm. K-means is a centroid based distance algorithm which tries to minimize the distance between the points and center in a particular cluster. The numbers of cluster, K, is a predefined value.

Clustering on Electricity Usage by Hour

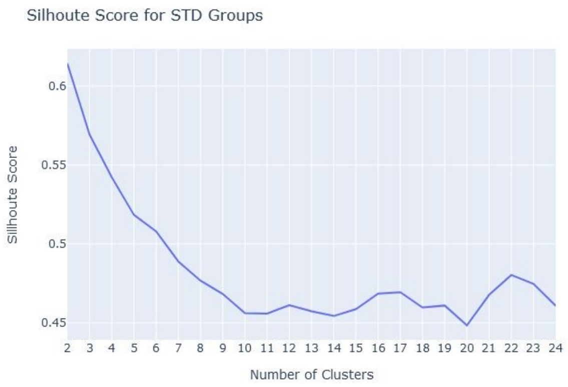
We clustered average electricity usage by the hour of the day. This clustering would allow us to visualize how we can optimize energy production to meet the demand.

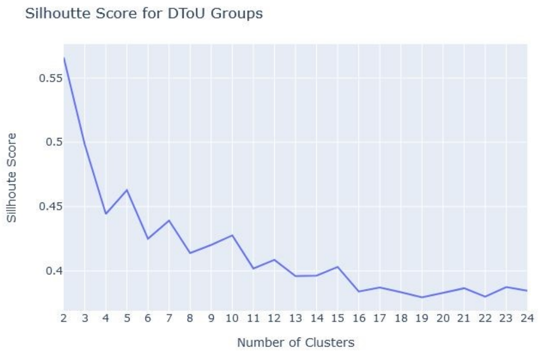
Dataframe and Attributes

We decided that the appropriate way of going about this would be to cluster the both dynamic and standard users separately. We prepared the data by keeping the only required columns, namely, 'DateTime' and 'KWH/hh'. We converted the date to hours by grouping the data and taking average for a given hour.

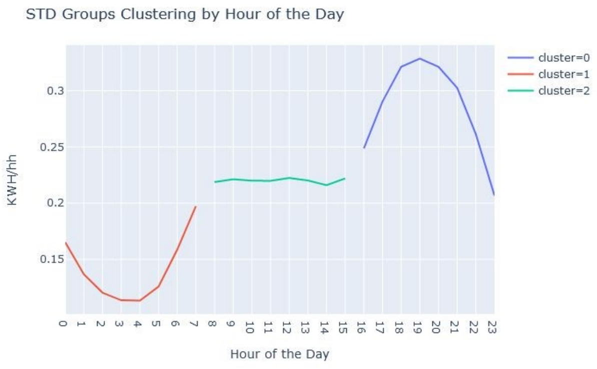
Clustering Results

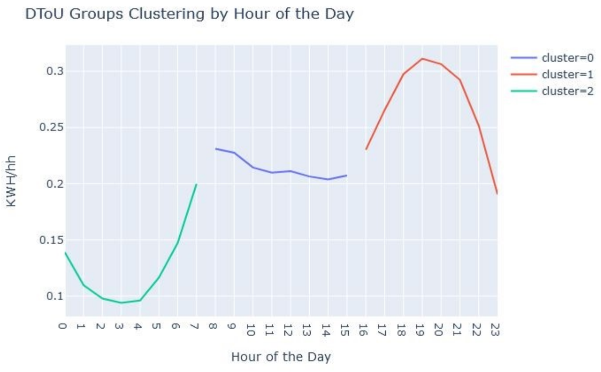
Moving forward, we calculated the Silhouette Score for the data. As you can see in the following images, the score is higher for the lower number of clusters.





We decided on using 3 for the number of clusters. As stated earlier, we used the K-means clustering algorithm. Once we had done the clustering, we averaged the electricity for each cluster for a given hour. The plots of the clustering are as follows:





As you can see that 3 clusters are quite well formed here. Both standard and dynamic electricity usage can be described as:

- Night time usage: This is from 12am to 7am. The usage first decreases and then increases. It is between the range of 0.1 KWH/hh to 0.2 KWH/hh. During this time, most people are asleep and the low usage is expected.
- Morning / afternoon usage: This is from 8am to 3pm. The usage is relatively steady during this period. It is around 0.22 KWH/hh. Most people are at offices and schools during this period, thus the usage more than night time but constant.
- Peak usage: This is from 4pm to 11pm. The usage is between a range of 0.32 KWH/hh to 0.19 KWH/hh. In this time, people return from schools and offices and perform different activities which lead to the highest electricity usage in the entire day.

Clustering on Season

We clustered customers on the basis of energy consumption in each season. Through this we want to see if there are some customers who change their energy consumption pattern from one season to another. Or if there are customers who have constant use of electricity throughout the year.

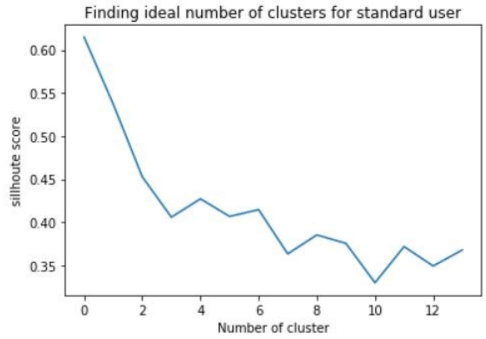
Dataframe and Attributes

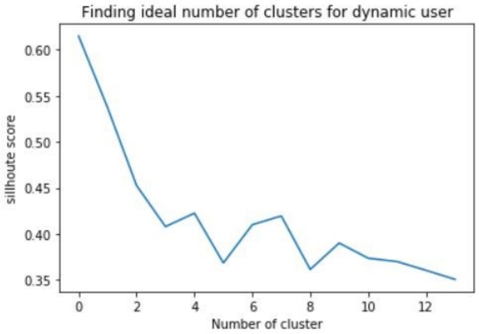
First, we prepared the data. We processed and extracted the data from the given dataset such that our new dataset contains one row for each customer and 5 columns. 4 of the 5 columns represent the season namely Summer, Autumn, Winter and Spring. Last column represents the group of the customer. We defined boundaries of seasons which are listed in Table 2

Season	**Months Included**
Winter	December, January, February
Spring	March, April, May
Summer	June, July, August
Autumn	September, October, November

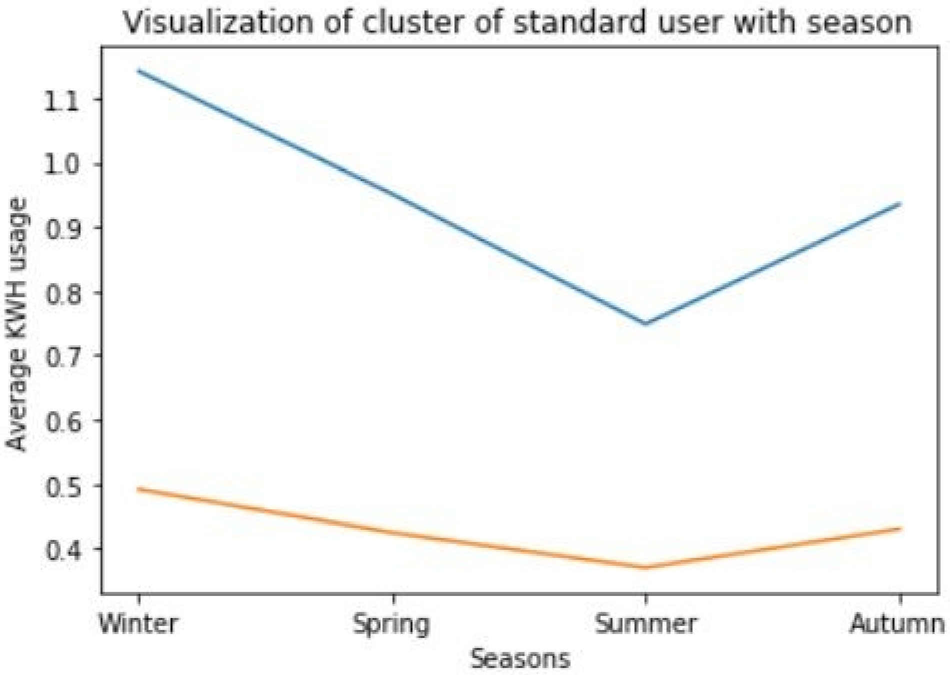
Table 2: Months included in each season during division of data for clustering

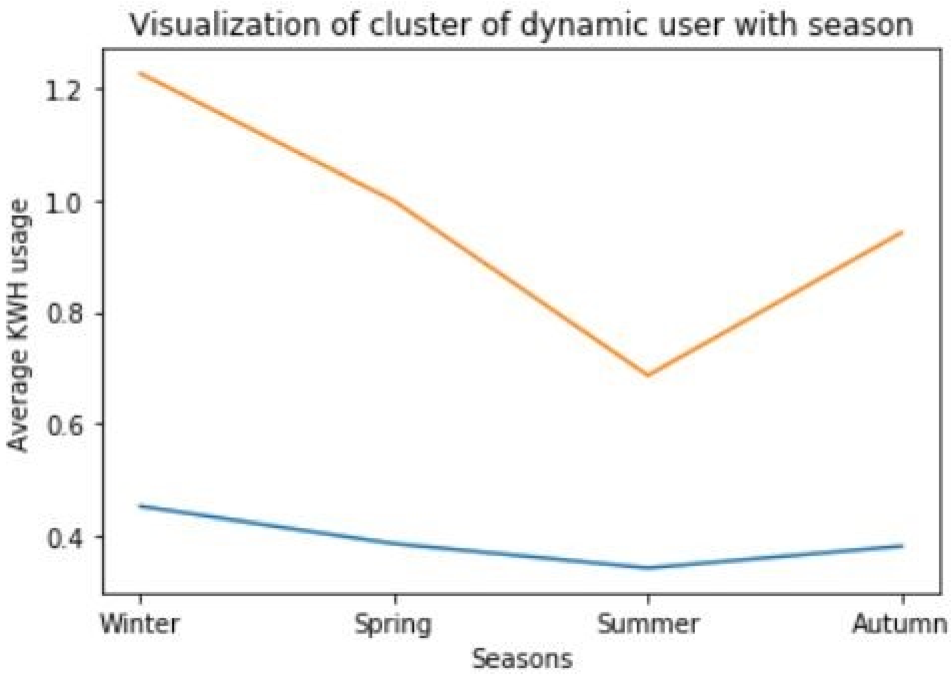
Clustering Results





We used silhouette score method for finding the optimal number of clusters and we found 2 to be the optimal number of clusters. Hence, we use the K-Means algorithm provided the K as 2. We did the clustering for the Dynamic tariff users and the standard tariff users seperately to ensure the low number of data of dynamic users isn’t in any way affected by the high number of users in the standard tariff and the results are comparable directly.





Outlier Analysis

Outlier analysis aims at detecting anomalous data points, i.e customer values in our datasets. These are customer values which stands out from the normal data points. The outliers customers for smart meters can provide very useful information about certain customers.

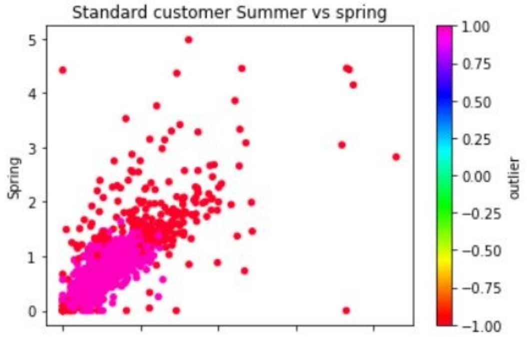
Isolation Forest Algorithm

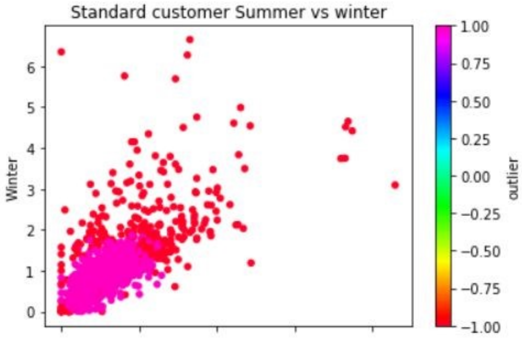
Isolation forest is an unsupervised learning algorithm for anomaly detection. Isolation forest algorithm works by randomly selecting a feature and then randomly splitting values of the selected feature, to calculate a threshold score for anomalous behaviour.

We ran the Isolation forest algorithm on the dataset of standard customer and dataset of dynamic customer separately, So that the number of customers in standard dataset do not overshadow the dataset of dynamic user.

Standard Customer:

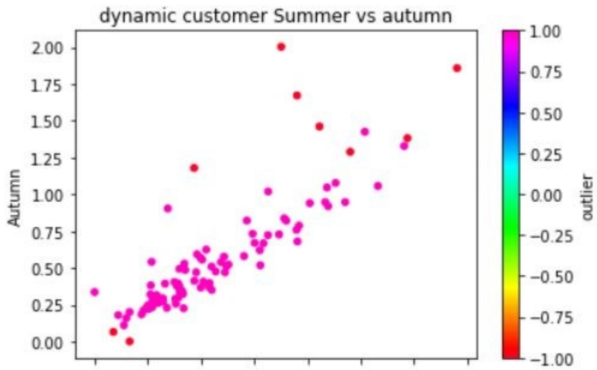
In this customer group, we had 3043 unique customers. The algorithm identified 305 (~10%) customers as outliers. The graph highlights that extreme large energy usage values and low energy usage values were highlighted as outliers.

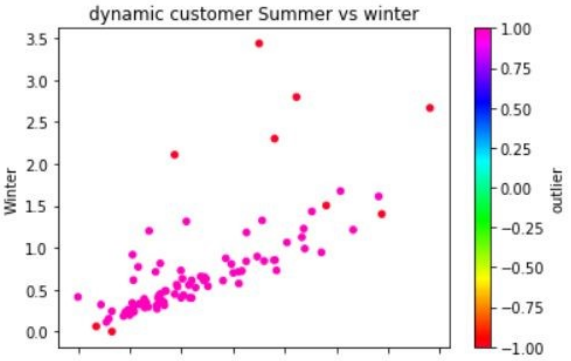




Dynamic Customer:

We had 90 unique customers for the dynamic group. Total 9 (10%) customers were identified by the algorithm as outliers. The graph highlights that extreme large energy usage values and low energy usage values were highlighted as outliers.


dynamic customer Summer vs autumn
Autumn
outlier


dynamic customer Summer vs winter
Winter
outlier