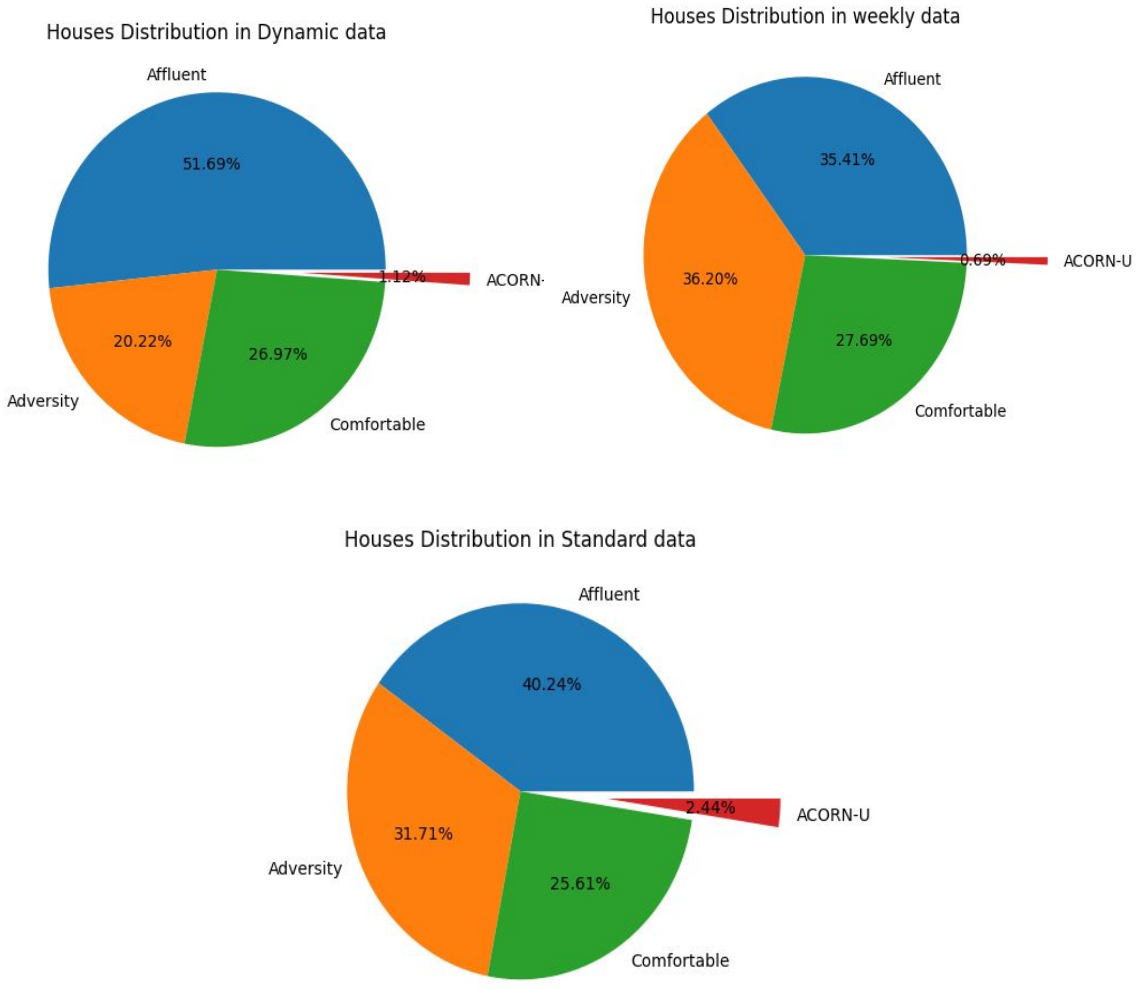
Group 3 Report

Chapter 1: Exploratory Data Analysis

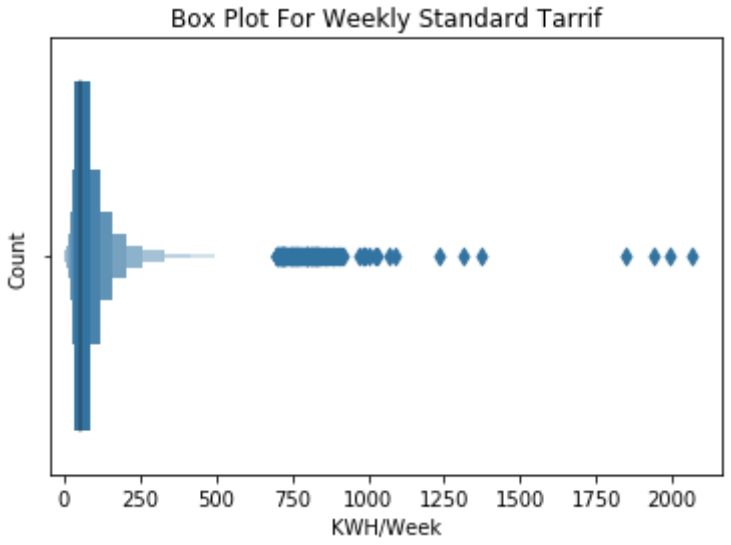
Summary Statistics



Starting off we will look at some numbers to describe what our data looks like. We broke up the data into three different dataframes. A standard dataframe that reformats the data into weekly time slots rather than per half hour. This is data taken from the first 90 files of the overall data and contains roughly 300000 observations from 3044 houses with all the various customer groups. The second

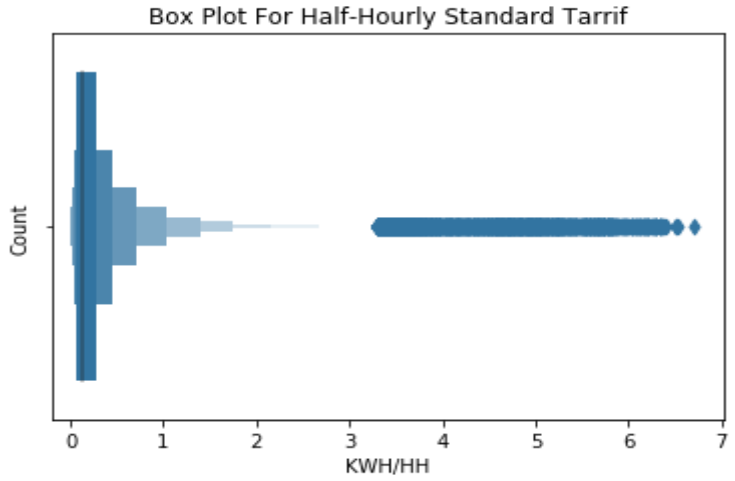
dataframe contains the raw data we received on a per half hour time basis and is from the first 3 files of the overall data and contains 3000000 observations from 82 houses. The last dataframe is for the dynamic customers and is from the last 3 files of the raw data. It contains per-half hour data and has roughly 3000000 observations from 89 houses.

Standard Weekly Data



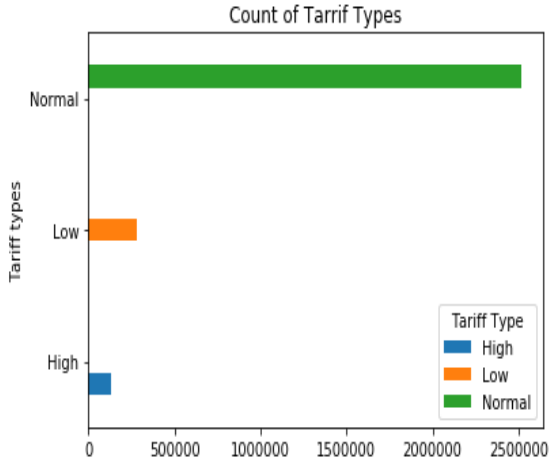
We used weekly rates to get a big picture of the data without getting side-tracked by the details, as a quick preliminary survey. For the 3044 houses in this category, we discovered that the majority of the data lay between 30 and 90 KWH/Week, specifically 50% of the data, from the 25th percentile to the 75th percentiles is within 34.6 and 88.4 KWH/Week. The median is 56.6 KWH/week whereas the mean is 71 KWH/Week, indicating a rightward skew. This can be seen in the many outliers.

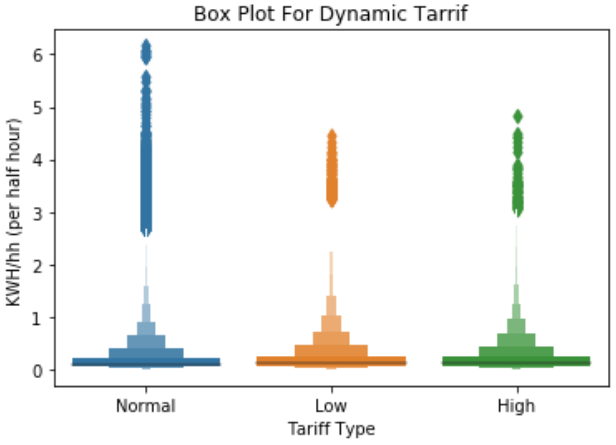
Standard Half Hourly Data



After the initial data survey we conducted a more in depth analysis to check if the data condensation had removed any patterns. We found the same right skew as before since the mean, 0.24 KWH/HH, is greater than 0.14 KWH/Week. The 25th and 75th percentiles are 0.066 and 0.27 KWH/HH respectively with the minimum value being 0 and the maximum value being 6.7 KWH/HH.

Dynamic Half Hourly Data





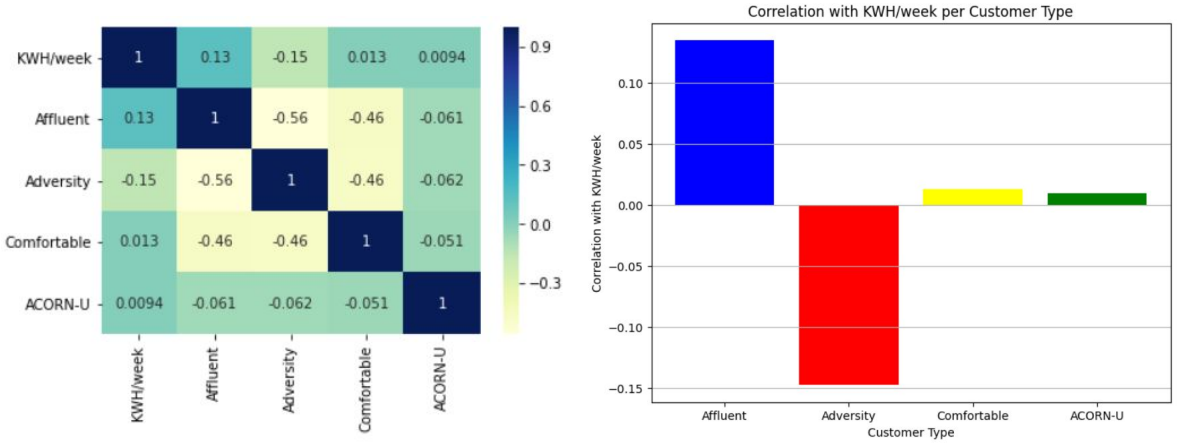
For the 89 houses that had dynamic tariffs we counted the number of low, normal and high tariffs and found a resounding majority to be normal tariffs followed by low then high. The distribution of high and low tariffs are very similar but there are much more outliers for normal tariffs. Overall, the

mean is 0.21 KWH/Week, while the percentiles, 25th, 50th and 75th are 0.053, 0.11 and 0.23 KWH/Week respectively.

Energy Usage Correlation with Categorical Attributes

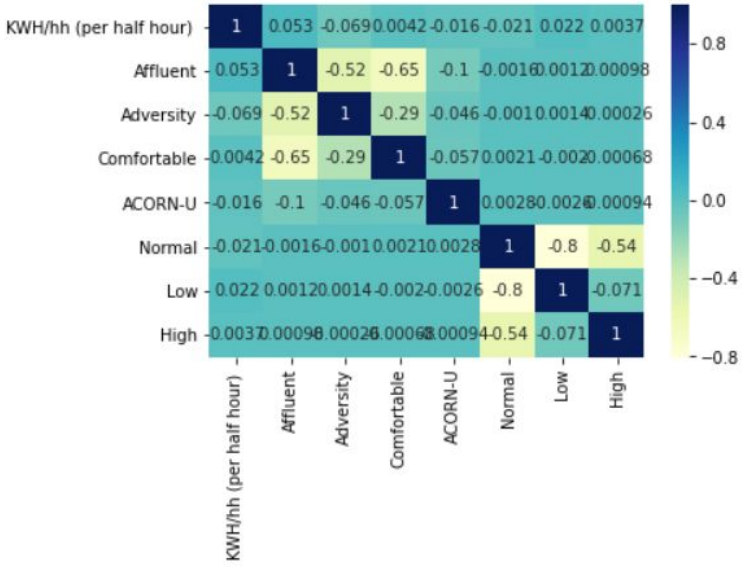
In this section we will define the correlation between the Energy Usage patterns we observed with 2 different categorical attributes. These include the customer group which separates the data on the basis of the income group that customer belongs to, and the tariff type for dynamic data which determines the rate being charged for the electricity in that time period

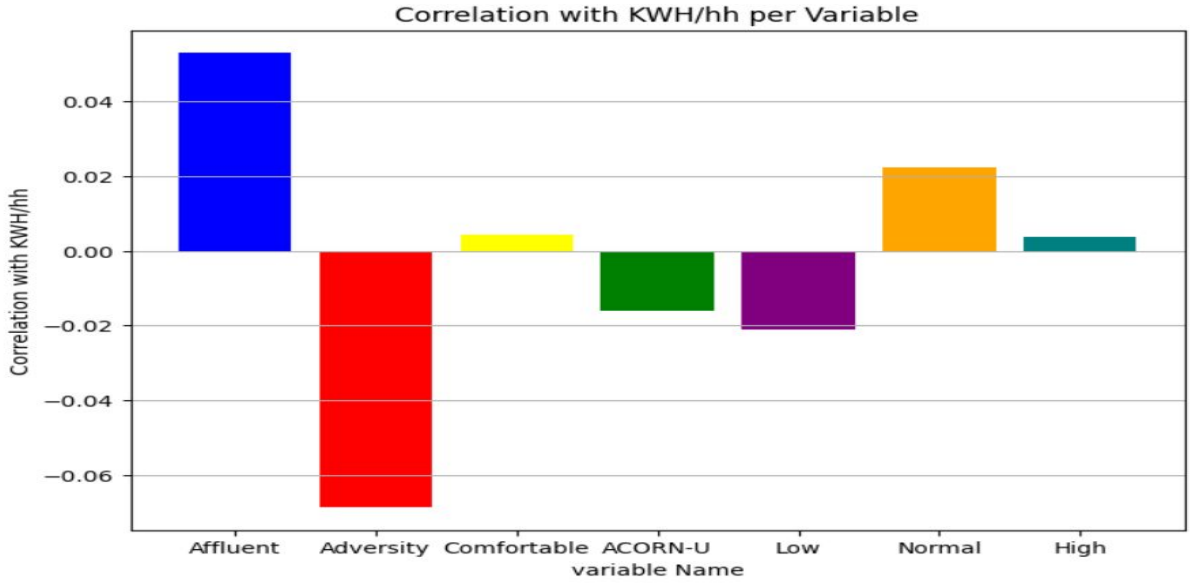
Customer Group



As we can see from the figures, Customer types “Affluent” and “Adversity” have a very weak relationship with KWH/week as the numbers range around 0.15. Whereas the “Comfortable” and “ACORN-U” customer type seems to have no linear relationship whatsoever with KWH/week, as their numbers are very close to zero. It makes sense that the Affluent type customer have a positive linear relationship with KWH/week, as they are likely to use more electricity, and it also makes sense that the Adversity Type has a negative relationship as they are likely to use less electricity because of their circumstances. The Comfortable and ACORN-U type of customers most probably are the people in the middle of society whose energy consumption lies around the average energy consumption.

Tariff Type



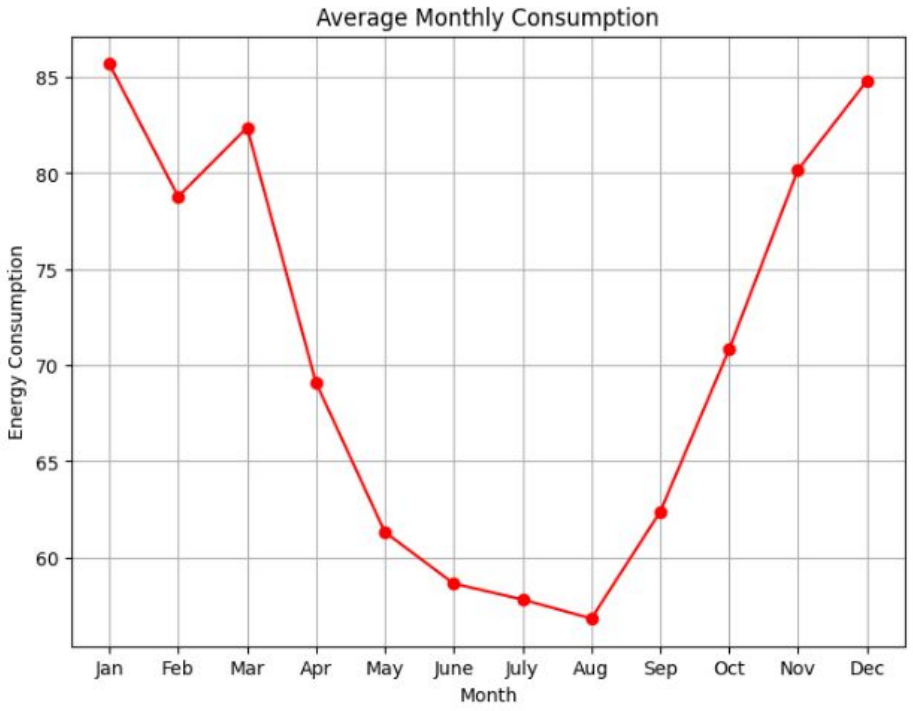


This data was a bit different; we can see that all the customer types have no correlation with KWH/hh, as their correlation values are very close to zero, maybe if we had accumulated the energy consumption to KWH/week we might have seen different results. Still Affluent and Adversity customer type does have a more significant correlation value than the other even though it is still very close to zero. The surprising thing to see here was that there was no significant linear relationship between the energy consumption with the type of use type. All the values are very close to zero, This means that the customers don't really care what time of usage is going on, their energy consumption remains the same. But it is interesting to see that the low type has a positive value and the other types have a negative one. This could show a very small variation in energy consumption.

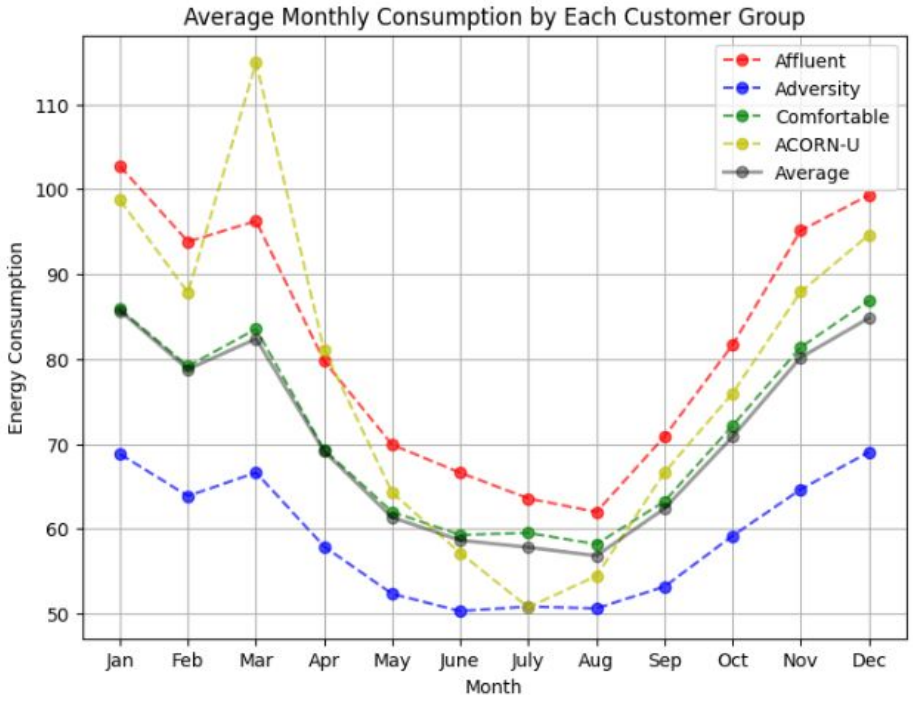
Energy Consumption Patterns Over Time

In this section we will be looking at how the energy consumption varies over periods of time and try to rationalize these variations and explain them.

Monthly Comparison of Energy Usage

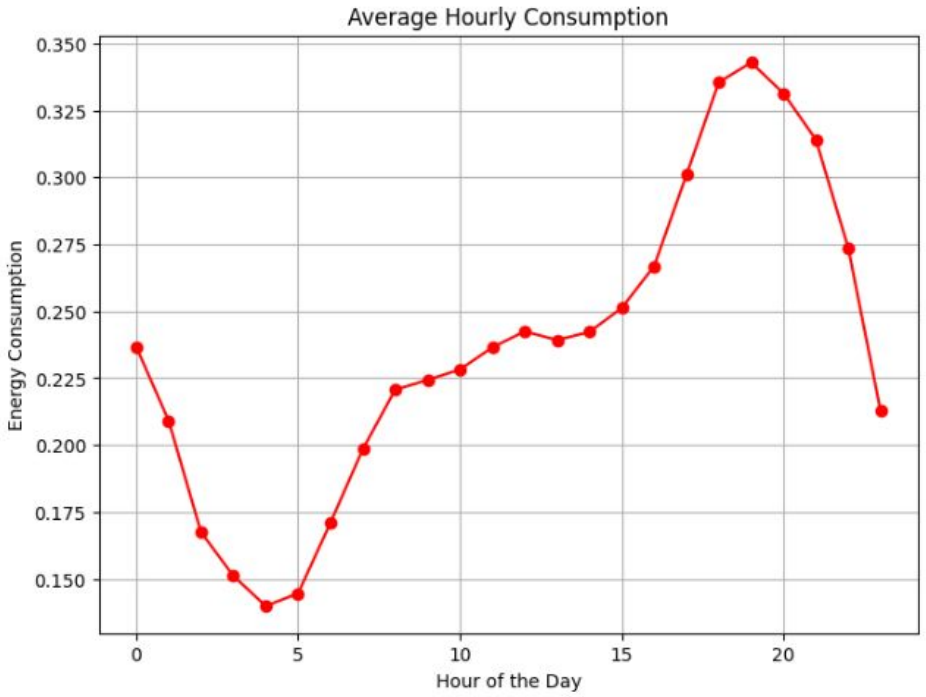


The experiment was conducted on the weekly standard dataframe in which we added up the energy consumption for every month and then divided by the number of observations. The above plot shows one clear pattern, winter season's call for higher energy consumption. The data was collected from the UK which sees harsh winters and mild summers hence,most of the energy in winters is consumed by heating appliances to keep houses warm. On the other hand, in summers, the UK doesn't require aggressive cooling methods such as air conditioners and 24/7 running fans hence the lower energy consumption. The energy consumption is almost symmetric as it falls when we approach the summer season and rises as we approach winter. The one point that obstructs this symmetry is that of March and we'll see this as a recurring pattern going forward as well.

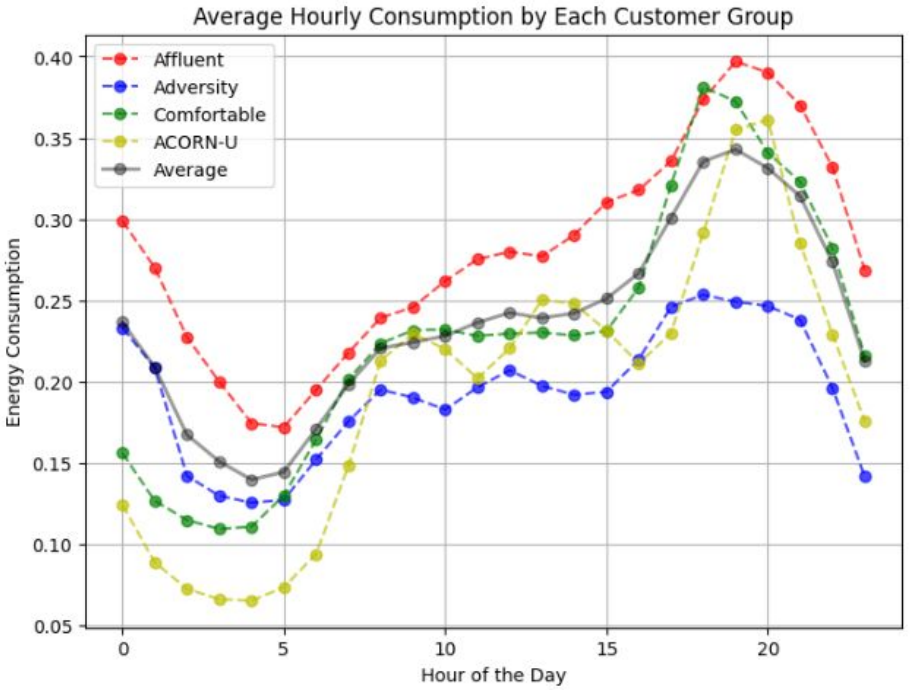


Over here we compared how this same metric varied for different customer groups. In the black non-dashed line, we see the average energy usage following the same trend we displayed earlier. Every groups' consumption can be evaluated based on how it compares to the average and we see that the affluent group is well above the average energy consumption level whilst the adversity group is consistently below average. This represents the economic divide which would lead to low income households being forced to consume less energy. The comfortable group sits close to but always above average. The ACORN-U group is a mystery but we can see that it's distribution sits comfortably above average in the winter months, however sharply dips in the summers, even going below average in the 3 hottest months of the summer.

Hourly Comparison of Energy Usage

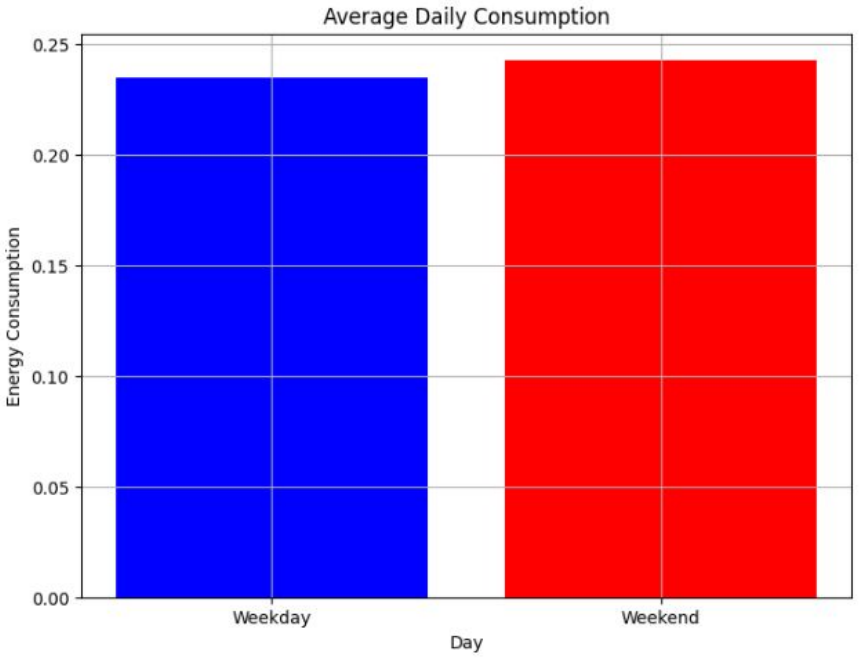


This analysis was conducted on the Hourly Dataframe which contains per half hourly data. The graph shows the expected distribution throughout the day as we see less energy consumption through the night as opposed to the day. During the hours of 12 am to 5 am, more and more people go to sleep and so we see the graph dip in these areas going below 0.150 KWH/hh. From 5 am to 8 am, people start to wake up and so the energy consumption rises rapidly and then slows down from 8 am to 12 pm as during this time people are only getting ready to start the day. From noon onwards we can see an exponential rise in energy consumption till 6 pm when a typical work day comes to an end. Finally from 6 pm to 8 pm, the families get ready for dinner before preparing to end the day and so we see a sharp dip from 9 pm to 11 pm as the majority of people with work starting early morning go to sleep.



We then follow the same exercise as before by evaluating how different customer groups stack up next to the average usage in a day. The one major thing to note here is the activity of the affluent group. We see that during night hours, all groups remain below average in energy consumption except the affluent group, which can probably be attributed to luxurious tendencies that would require the use of electricity. The adverse group remains below average throughout the day while the comfortable and ACORN-U group goes between below average and above average. The affluent group however stays well above average throughout the day.

Weekend vs Weekday Comparison

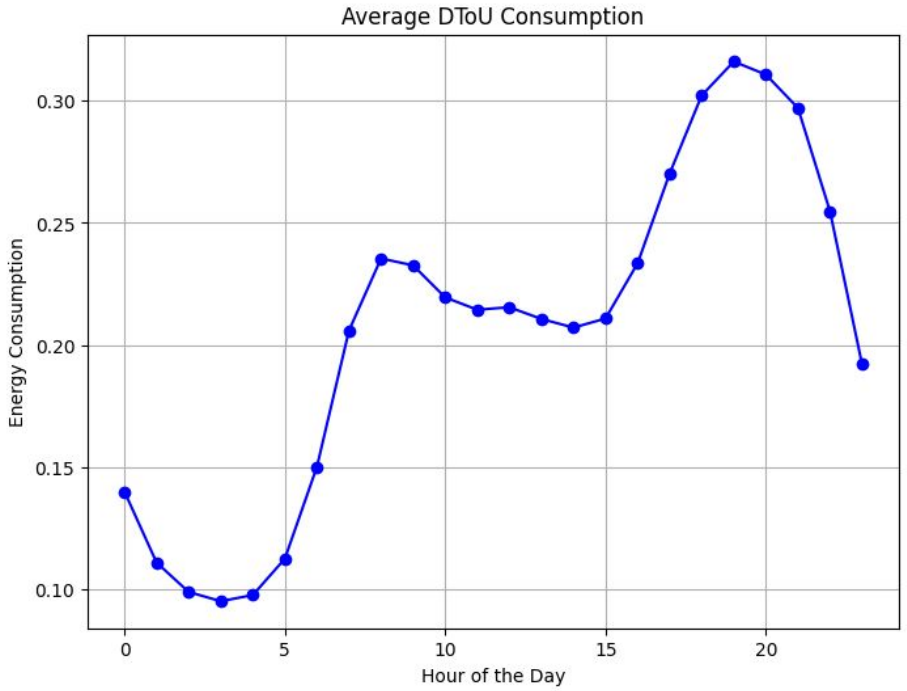


We tried to evaluate if there was any pattern in the difference of energy consumption on a weekday vs a weekend however we found little statistical difference in the results which indicate that weekends and weekdays see the same patterns in energy consumptions. Even when we proceeded on a day to day basis, we found that all days had roughly the same energy consumption in each day, with saturday and sunday having only a slightly higher energy consumption metric which lead to the marginal difference between the weekend and weekday bars.

Energy Consumption Patterns Over Time with Household Tariff Type (DToU)

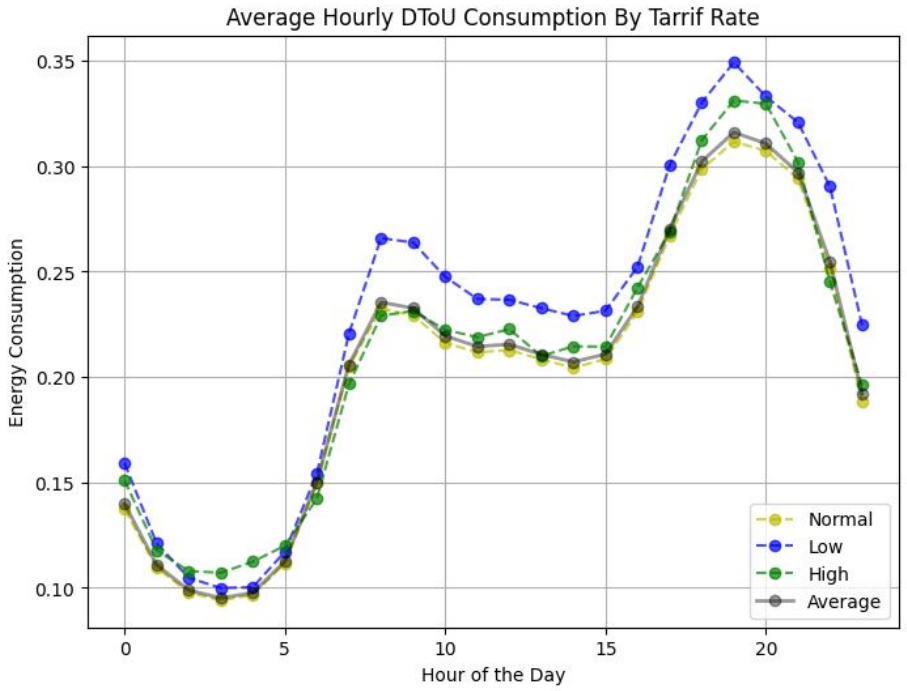
In this section we will be looking at Dynamic Households and their energy consumption patterns only. This is data from the dynamic dataframe which contains the data of 89 households. Since there was only one household that belonged to the ACORN-U customer group, it has been omitted in the below results due to lack of statistical representation.

Dynamic User Hourly Consumption



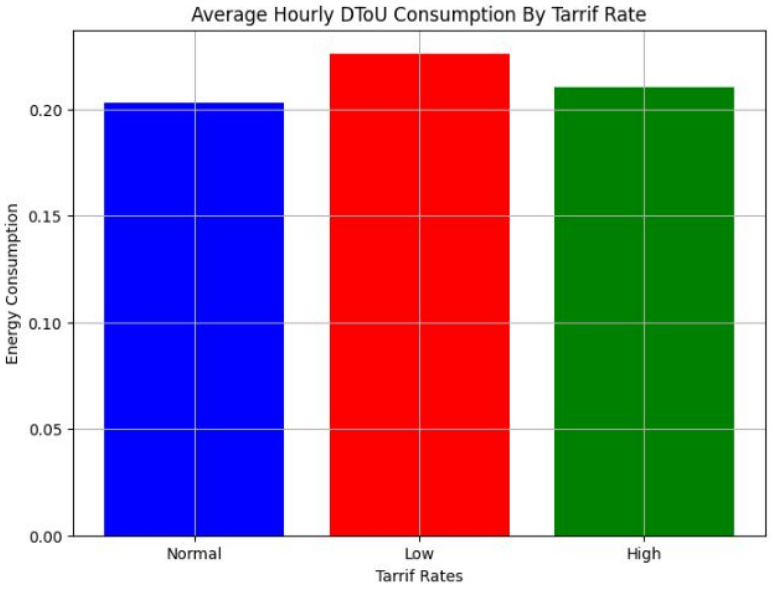
This analysis was conducted on the Tou Dataframe(Dynamic) which also contains per half hourly data. The analysis was achieved by amassing the total energy expenditure from the dynamic households recorded after every half hour. The total energy consumption for each half hour reading was then averaged by dividing it by the total number of readings for that half hour recording. The same is done for all readings until a pattern is created which helps to further elucidate upon the general hourly energy consumption trend across the dynamic households in the UK. The trend basically shows a deep minimum expenditure in the very early hours of the day (middle of the night) which is when the general populace is asleep. Mildly cool temperatures in the UK may prove that there is not as much need for air cooling or heating systems during this time of inactivity. There is a steep rise in early morning (between hours 5 - 8) indicating higher consumption. This is mainly due to more of the general populace waking up and using appliances and/ cooling systems. The consumption stays more or less the same till mid-day and then rises to a steep maximum indicating peak energy consumption. The need for heating systems due to lower temperatures at night as well as lighting may be the cause for the surge.

Dynamic User Hourly Consumption Comparison by Tariff

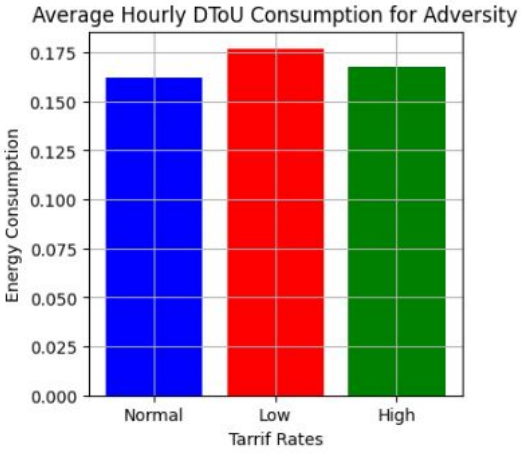


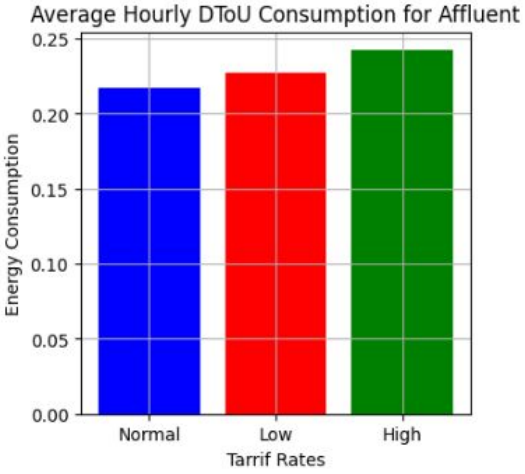
This analysis continues from the previous one done on dynamic households in the UK that are charged different tariff rates according to the time of recording. The three tariff rates are 'High', 'Normal' and 'Low' and are charged 67.20 p/kWh, 11.76 p/kWh and 3.99 p/kWh respectively. The normal and average graph are more or less the same indicating little change in usage when normal tariff is charged per kWh. The usage for low tariff charges is relatively the highest which would prove to be more economically rational owing to the colossal differences in the prices charged for lower tariff and the rest. The high tariff graph, however, shows that consumption is greater in some areas than the normal which could only mean that those numbers belong to the affluent customer households that generally consume more energy compared to other customer groups.

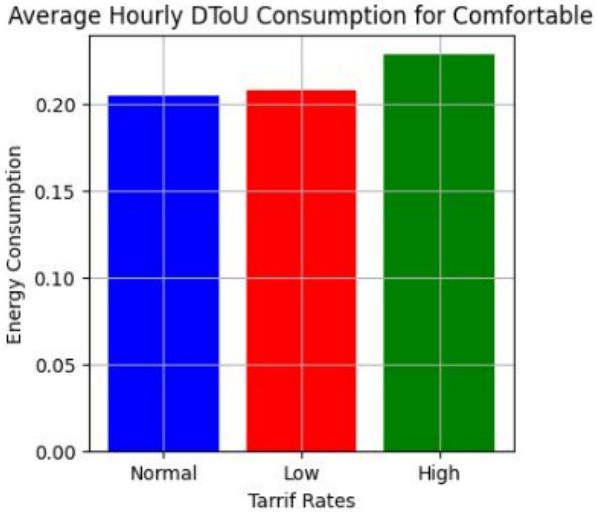
Average Energy Consumption



For this section we wanted to find out the average amount of energy that was being consumed by customers when they were being charged according to these different tarrif rates. Over the entire 3000000 observations, we see that there is very little difference in the amount of energy consumed on the different tariffs. Our theory is that since a household is being charged by the normal rate through roughly 90% of the year, most houses do not change their daily routines according to when the tariff has changed. Since the tariff so rarely changes, people likewise would not accommodate changes for such rare behavior hence the high similarity between the plots. However, we do see some small variations. This is because of the few people that do alter their behaviors with the tariff rate and we'll see these groups below







By inspecting the different graphs, we can extract two very interesting takeaways. Firstly, the adversity group has marginally higher usage during the Low tariff as compared to the normal or high tariff which would be expected as they would want to capitalize on cheaper energy however the problematic mix in the data is the higher consumption of the high tariff as well which goes against rational thought since the adverse group would want to minimize this consumption as it is more costly. The affluent and customer groups however follow a similar distribution with them consuming more energy on high tariff as opposed to low or normal. It is difficult to say whether these patterns have any statistical relevance or are just brought about by coincidence.

This concludes our EDA portion of the final report.

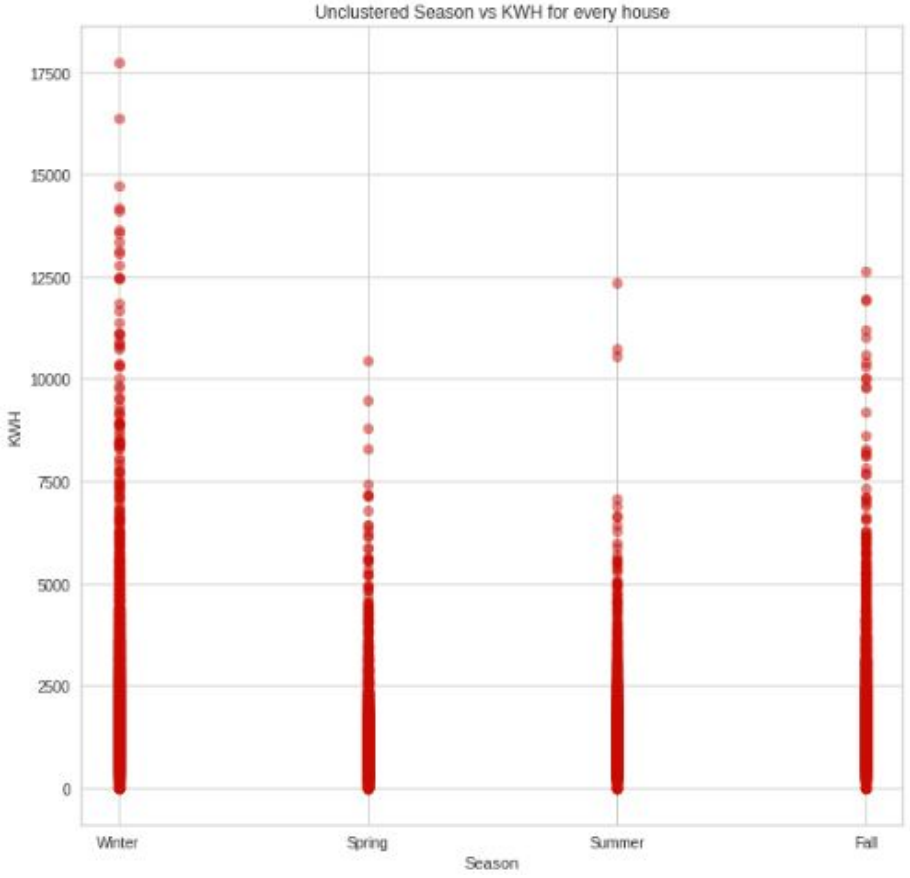
Chapter 2: Cluster Analysis

Part 1: Clustering

We conducted cluster analysis on the standard dataframe containing 90 files and entries from 3000+ std tariff type households. In the initial processing the energy usage for each household was summed up over each season giving us a dataframe in the following format:

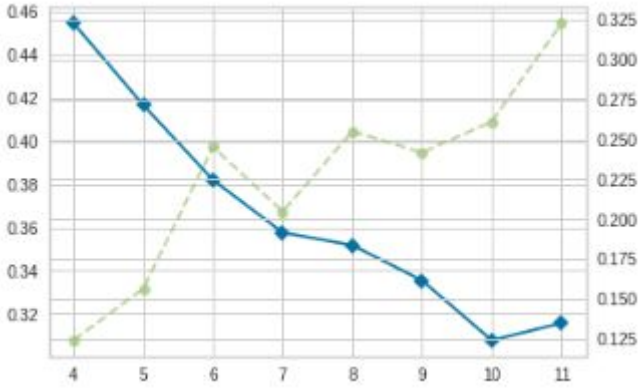
	Winter	Spring	Summer	Autumn
MAC000002	2428.027001	1082.450000	824.909000	1765.985000
MAC000003	5333.965000	3340.574001	2229.078000	3200.903002
MAC000004	331.533000	197.364000	277.058000	314.964000
MAC000006	645.342000	531.712000	445.952000	545.319000
MAC000007	1919.199000	928.044000	744.335999	1362.720000

In this dataframe, each row represents a unique household, and each column represents that household's energy usage for that respective season. We plotted this dataframe and it gives this resulting figure:

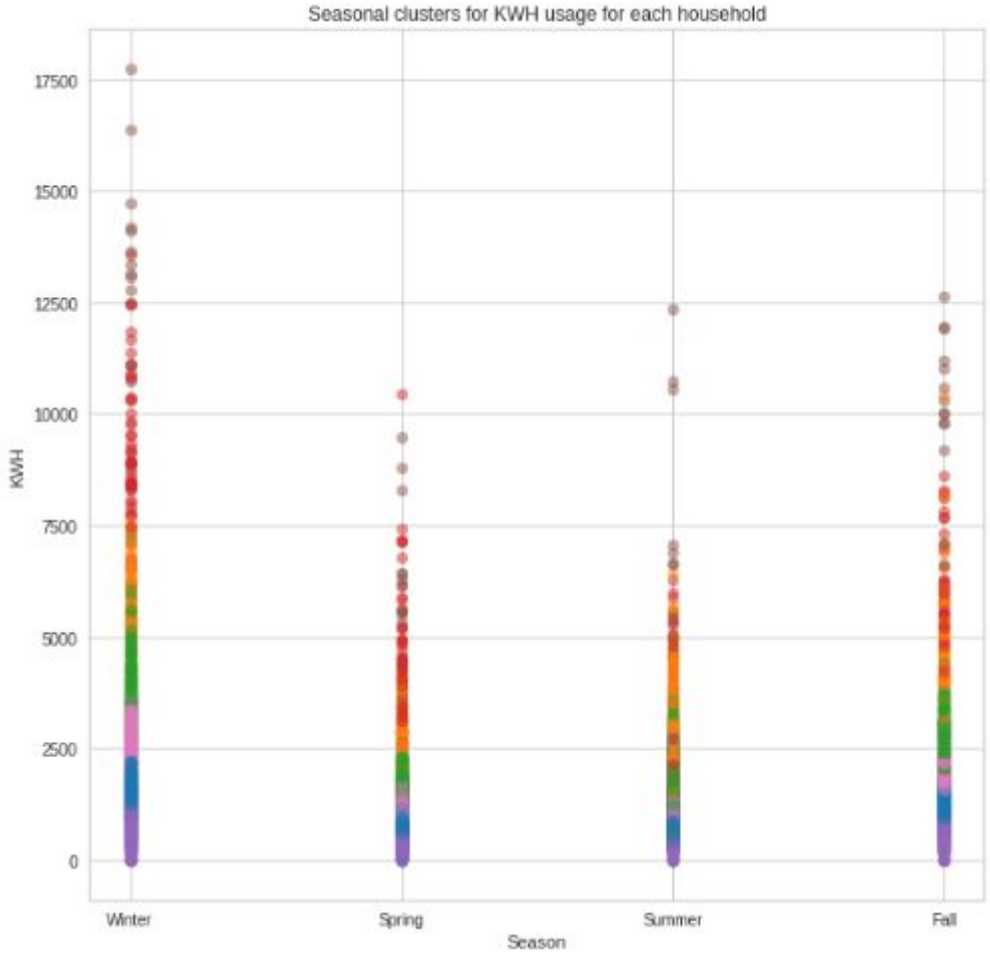


This figure shows us how the energy usage of these households is arranged, with the darker concentration of red at the bottom of the graph showing that most households consume energy at these levels.

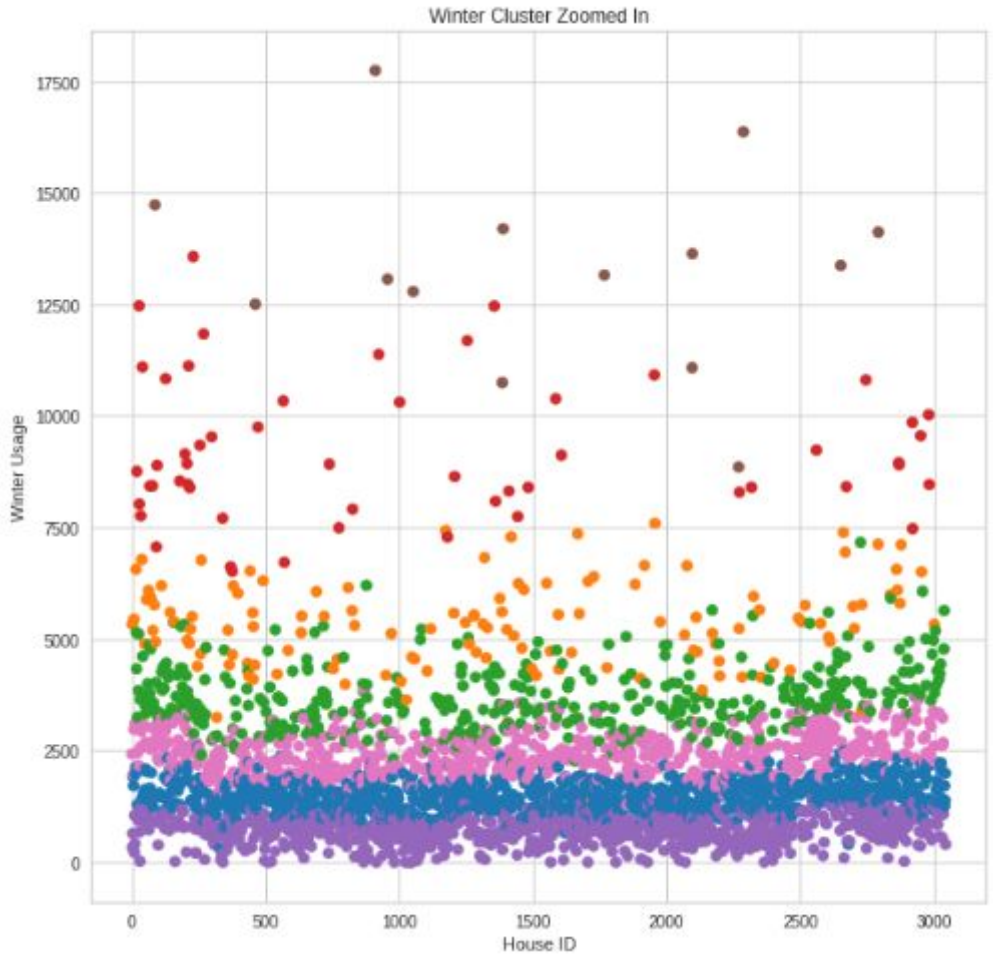
The next step was to figure out the number of clusters to use on the data as well as the clustering type to be used. We tried two different types of clustering and two approaches to tuning their parameters. Firstly we used KNN to try to find the eps value for DBSCAN. After tweaking the DBSCAN algorithm several times, we found that every configuration reported the majority of the points as outliers leading us to changing our approach. We then shifted to K-means clustering and to find the initial K value, we used the elbow method. The score that was computed to compare different K values (ranging from 4-12), was the silhouette score. After plotting the results, here's what we saw:



We tried other scoring metrics as well, but 7 seemed to be the right answer in most cases. Thus, we decided to use 7 clusters for our data. After fitting our data to the K-means algorithm, this is the resulting clustering that we saw:



The data is split into 7 consumer groups and before analyzing the clusters from this figure, let us zoom in on one season and see up closely what the clustering looks like:



From this figure of the winter cluster, it becomes much more clear what each cluster represents

1. Purple cluster: Lowest income and least energy consuming group, most like adversity customers.
2. Blue cluster: Below every energy consumption and income group, most likely upper class adversity group
3. Pink cluster: Average energy consumption, sitting right around 2500KWH for the season
4. Green cluster: The comfortable customer group sitting above average in most cases.
5. Orange cluster: A mixture of comfortable and affluent groups that sit well above average never going below 2500KWH energy consumption
6. Red cluster: Never going below 5000KWH for the season. Upper end of the affluent group of customers.
7. Grey cluster: Potential outliers, with only a few number of points, consuming roughly 5x the energy of an average customer.

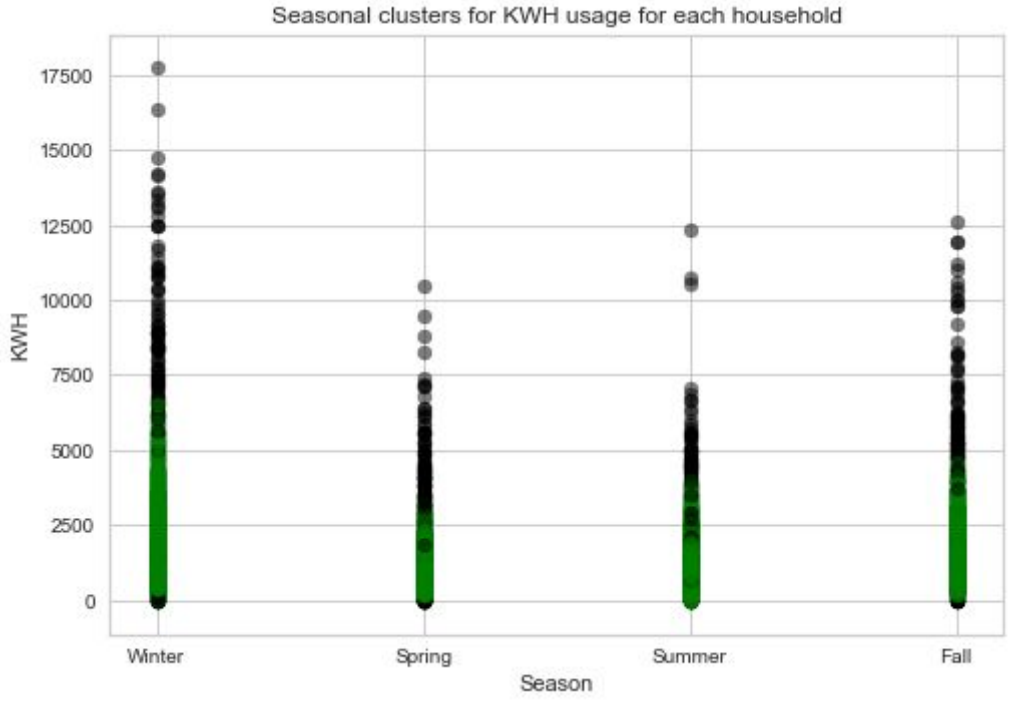
There could be some clusters we could merge in this case if we were using hierarchical clustering, for example, the purple and blue clusters could be merged to represent the below average and low income strata of the population. The pink and the green clusters could also be merged to represent the comfortable group of customers. Which would leave the orange and red clusters to represent

different tiers of affluent customers and the grey clusters as outliers.The graphs for the other seasons can be found in the notebook which follow a similar distribution.

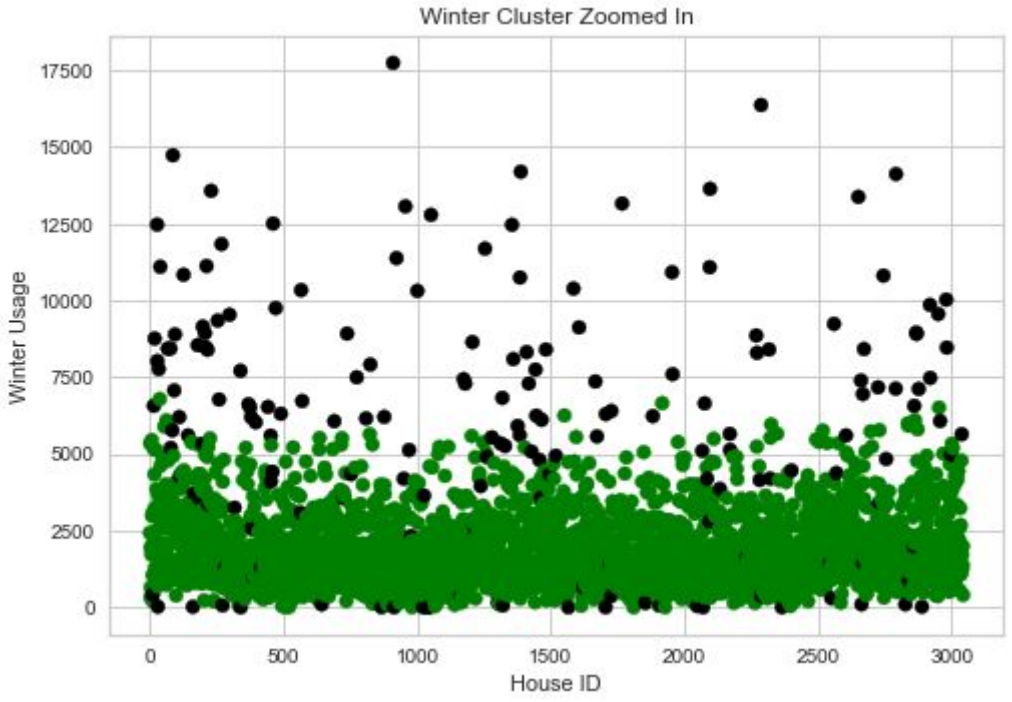
Part 2: Outlier Analysis

In order to conduct outlier analysis on the data we had multiple choices to choose from. Since the data was not labelled we had to employ unsupervised outlier detection and decided to go with the two major unsupervised approaches Isolation Forest and Local Outlier Factor. We used the same dataframe that we used for clustering earlier, for our anomaly detection.

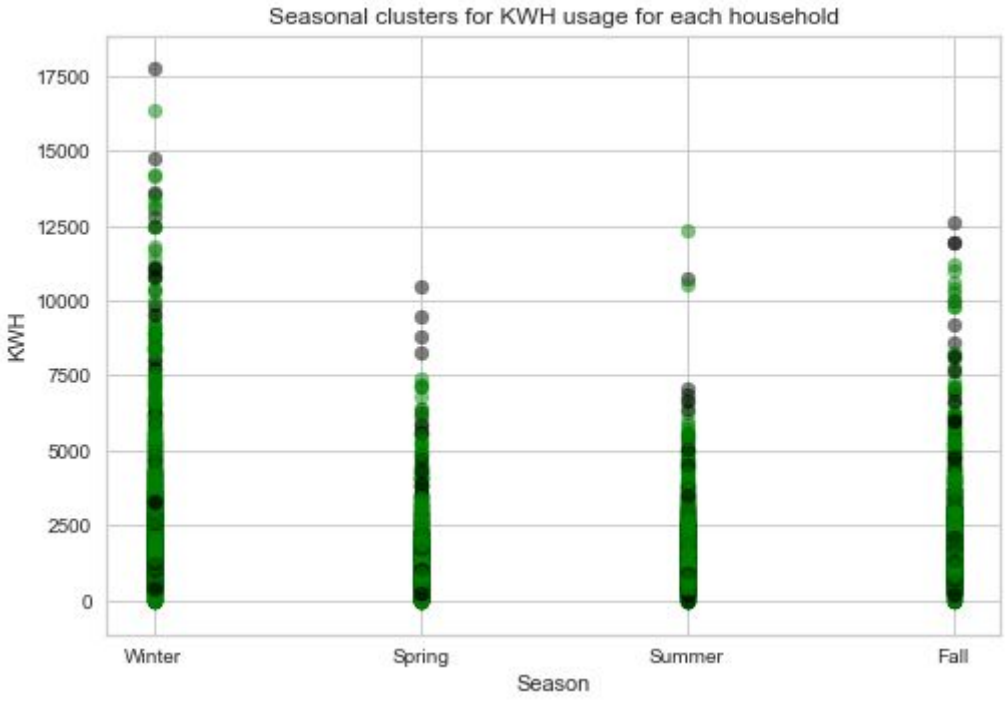
We ended up tuning a single parameter, max_samples, for Isolation Forest to reduce the number of false anomalies by repeating the number of experiment many times and eventually settling upon the number 1500. Performing this for our data frame we get:



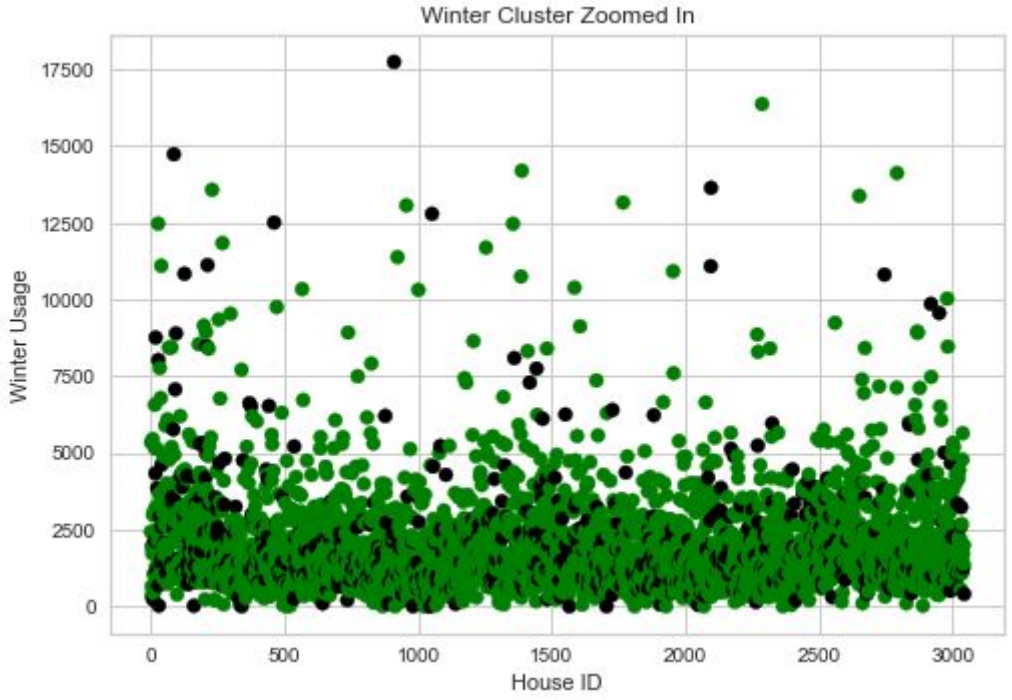
As can be seen the Isolation Forest does a fair job in isolating anomalies but gives a lot of false positives i.e it assigns a high anomaly score to many of the points. Looking at the figure we can see that the black dense 'anomaly' regions in the upper regions of each season. Once again in order to better understand the anomalies and the regions we zoom into the Winter season. This is shown below, where we can see that the Isolation forest classifies most of the members of the 5th cluster and all of cluster 6 and 7 as outliers. Due to the nature of these points, further away from each other less dense and fewer in number they are considered to be anaomilies. This same pattern of the above clusters being classified as anomalies can be seen across all the seasons.



In an attempt to compare this result and make sure that the points labelled anomalies were in fact so and not merely noise or wrongly labelled we further attempted a similar anomaly detection. For the second method we used a density based outlier detection technique, namely Local Outlier Factor (LOF). For the LOF method we attempted to first find the ideal number of neighbors. After multiple attempts of various values we finally used a single neighbor, so that we ended up with:



This zoomed version of winter usage shows how LOF outliers are spread. We can see that unlike Isolation Forest this technique seems to show outliers for each cluster. There is no single region where all the black anomalies cluster but rather regions of varying density so we can see black and



green intermixed with slightly more outliers along the upper and right regions. We can see that this technique is more well suited to identifying outliers than Isolation Forest which classified entire clusters as outliers. When looking at other season we see similar patterns of outliers interspersed with the normal point but some slightly denser regions of black, i.e outlier regions.

This concludes chapter 2.

Chapter 3: Conclusion and Recommendations

Conclusions

1. The average energy consumption in the standard tariff households is greater than the median, indicating a right skew meaning non affluent groups which are in majority are consuming less energy however the affluent groups are consuming enough energy to raise the average above the median.

2. The number of households in the dynamic data frame that were charged normal tariffs was overwhelmingly large. The second largest number was for the low tariffs and the least was obviously the high tariffs.

3. The Affluent customer type has a positive linear relationship with kWH/week and thus the highest consumption of electricity, while Adversity is conversely the least energy consuming customer type with a negative linear correlation with kWH/week. The ACORN-U and Comfortable types use an energy amount close to the average consumption.

4. In the UK, the summers are mild and cool while the winters are harsh and cold. The resultant patterns of monthly energy consumption over the course of a year show that the energy consumption in UK households is highest in the winters (January - March, Oct - December) when the climate is the coldest and the need for heating appliances and air conditioning is at the maximum.

5. The hourly energy consumption patterns show a minimum consumption in the early hours of the day, then steeply rising to a higher sub-constant level at afternoon/midday, going to the highest near dinner time at night, and steeply falling late at night. The energy consumption pattern explains the activity cycle of the average UK household.

6. People spend the same amount of electricity on weekends and weekdays with little to no difference.

7. Hourly energy consumption patterns for dynamic households is similar to households belonging to the standard dataframe. This is since households don't care for the tariff time, as can be seen from conclusion no 4.

8. Dynamic households for lower tariff rates show a slightly higher average energy consumption than other tariff rates, with high tariff average consumption having a minute edge over normal tariff average energy consumption.

9. The adversity group has a marginally higher energy usage during low tariff rates which makes rational sense, since low tariff rate saves a lot of money. However, high tariff energy usage is still higher for the same group than normal tariff which opposes the same

aforementioned rationality. The affluent and comfortable group however have a higher usage for high tariff rates so although it is highly likely for the results to be merely coincidence, it is hard to tell.

Recommendations

1. The power network should be producing 71kWH per household per week on average (the mean energy consumption over the week). It can also try to produce 0.24kWH per half an hour per household on average (the mean energy consumption over half an hour).

2. The power network should try to get more people aboard the dToU scheme as the number of customers using it is very low.

3. The power network can allocate energy in the grid according to neighborhoods having different customer types Affluent customers should be supplied with more power and adversity customers with less.

4. The customers don't really care what time of usage and the applicable tariff is going on, their energy consumption remains the same. The power network can extend timings for high tariffs rates to maximize profit.

5. The network can produce more and less energy according to the weather. They should produce more energy in the winters and less during the summers.

6. The network could produce energy according to the time of the day. With early in the morning being the minimum and near dinner time being the maximum to maximize efficiency.

7. The network does not need to adjust production patterns depending on if it is the weekend or not.

8. The network can update their tariff rates to maximize efficiency and profit because right now people are not that affected by them.

9. The network can potentially increase high tariff times for affluent and comfortable groups as they don't seem to care but not do so for the adversity group.

10. Using the clusters, the power network can allocate power according to each cluster and supply energy to the customer according to what cluster they lie in to maximize efficiency.

11. They can also find out which customers they can potentially charge more according to their clusters without it being a problem.

12. They can calculate a consumption threshold per cluster group to find out when after how much consumption they should start charging customers at a higher rate.