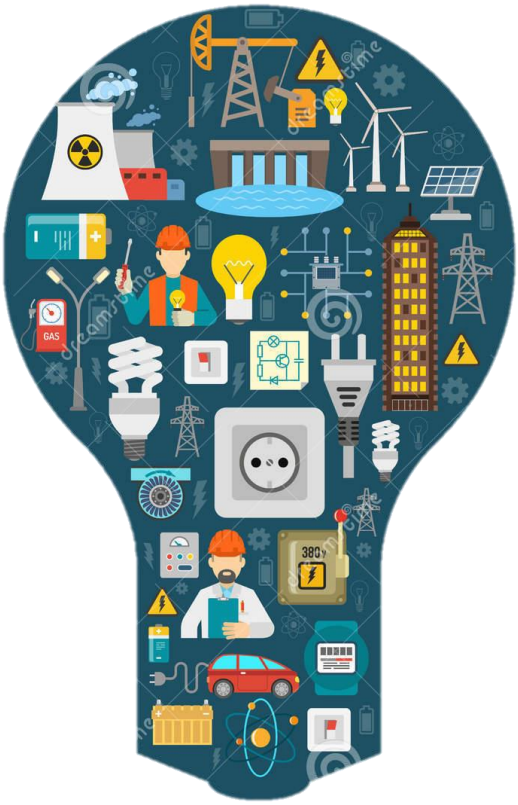
CS 536: Data Mining

Energy Consumption: Final Report



Submitted by: Group 10

Shehryar Sharif

Mariyam Tanveer

Mujtaba Ahmed Khwaja

Muhammad Hassan Tanveer

Overview

The data we were given reported the KWH of energy used by different groups of customers. The data was extensive, disseminating 93 files with a million rows each, and spread over 6 attributes, ‘LCLid’, ‘stdorToU’, ‘Datetime’, ‘KWH/hh (per half hour)’, ‘Acorn’ and ‘Acorn_grouped’. The stdorTou column consisted of two groups: standard or dynamic customers. The standard customers were charged at a fixed rate for the energy consumption, whereas the dynamic customers were charged based on the KWH consumption at different times of the day. Similarly, Acorn_grouped, also gave information about different sub-divisions of the customers.

Exploratory Data Analysis

Given the large amount of data, it took a lot of time to read and manipulate it, hence, the rows were filtered on the basis of days, with each entry corresponding to the sum of attributes on that particular day. The filtered data was aggregated in a new file, 'dataCombined', and contained 129177 rows only. Further, some columns were renamed for easy access, and the 'Acorn' column was completely dropped because it posed no significance in our analysis.

```
[ ] dataCombined.head(10)
```

	CustomerID	StdorToU	DateTime	Acorn_grouped	KWH
0	MAC000002	Std	2012-10-12 01:00:00	Affluent	7.098
1	MAC000002	Std	2012-10-13 01:00:00	Affluent	10.555
2	MAC000002	Std	2012-10-14 01:30:00	Affluent	12.569
3	MAC000002	Std	2012-10-15 02:00:00	Affluent	9.744
4	MAC000002	Std	2012-10-16 02:30:00	Affluent	8.886
5	MAC000002	Std	2012-10-17 03:00:00	Affluent	9.975
6	MAC000002	Std	2012-10-18 03:30:00	Affluent	9.224
7	MAC000002	Std	2012-10-19 04:00:00	Affluent	7.408
8	MAC000002	Std	2012-10-20 04:00:00	Affluent	16.004
9	MAC000002	Std	2012-10-21 05:00:00	Affluent	21.306

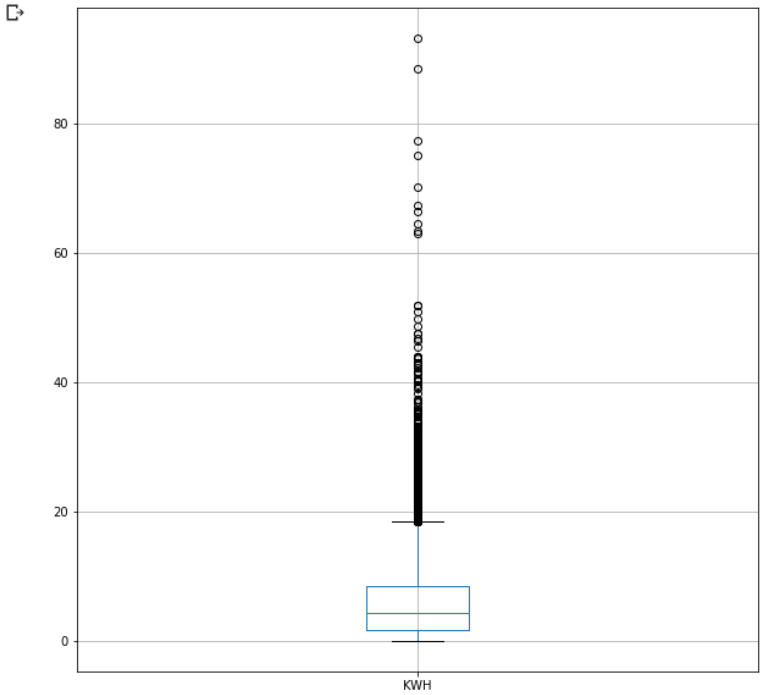
Data Cleaning

KWH was the only numerical attribute in the data. The null values in the column were replaced with 0 and its type was changed to float64. The type of DateTime column was also converted from object to datetime64[ns].

Data Visualization

Box plot - KWH

```
[ ]  boxplot = dataCombined.boxplot(column='KWH', figsize = (10,10))
```

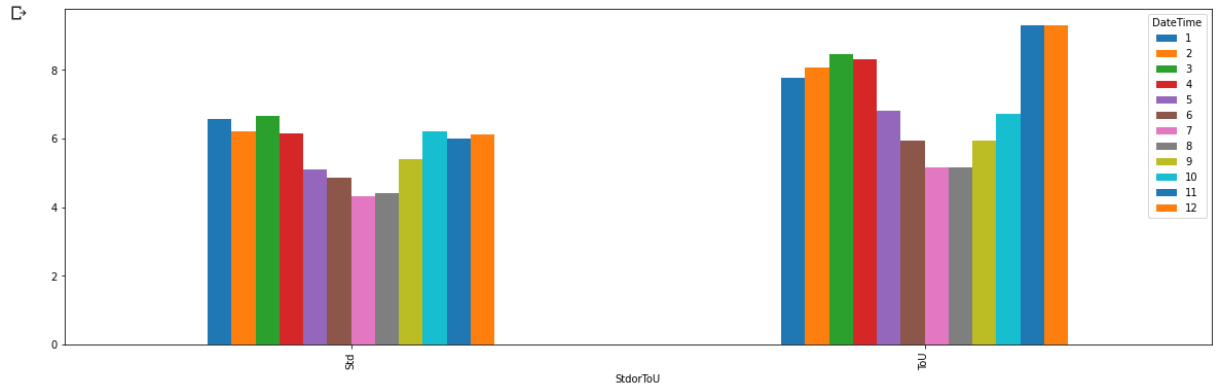


A box plot of the “KWH” entries was plotted to get a sense of the data’s shape. As the values are discrete and depend solely on the consumption of energy, removing the outliers might add a bias to the data.

The highest KWH value = 80 and the lowest is 0, with the interquartile range being around 7 KWH.

Energy usage vs customer group

```
[ ]  # Average energy usage in each month by each customer group
     temp = dataCombined['KWH'].groupby(by = [dataCombined['StdorToU'], dataCombined['DateTime'].dt.month]).mean().unstack().plot(kind='bar', figsize=(20,6))
     plt.xlabel('StdorToU')
     plt.ylabel('Average Daily Electricity Usage For Each Month /KWH')
     plt.title('Average Daily Electricity Usage For Each Month between November 2011 and February 2014')
```



The bar plots show the energy consumption by each group per month. The ToU group has an overall higher energy consumption as compared to the Std group. The overall pattern, however, is relatively same as the two have a similar shape. For both groups, the consumption decreases from March to

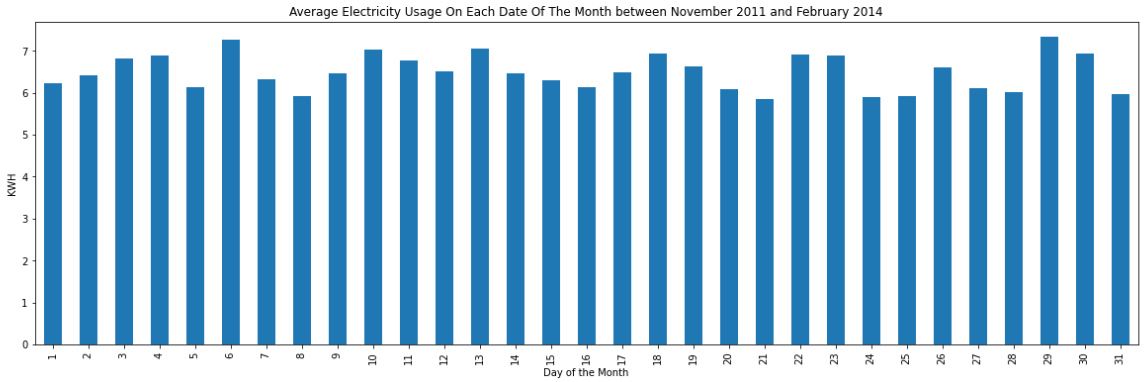
July, and then starts increasing again. For Std, there is a decline after October as well, as for ToU, there is a sharp increase from October to November.

The max consumption in Std group is in March and for the ToU group, it is in November-December. The variation might be a result of variable number of entries for each month. For instance, the August data covered 2 years – 2012 and 2013, whereas for some months, the data covered 3 years, 2012-2014.

Energy usage on each day of the month

```
[ ] # Average Electricity Usage On Each Date Of The Month
    Daily_Usage = dataCombined['KWH'].groupby(by = dataCombined['DateTime'].dt.day).mean()
    plt.figure(figsize=(20,6))
    Daily_Usage.plot.bar()
    plt.xlabel('Day of the Month')
    plt.ylabel('KWH')
    plt.title('Average Electricity Usage On Each Date Of The Month between November 2011 and February 2014')
```

```
Text(0.5, 1.0, 'Average Electricity Usage On Each Date Of The Month between November 2011 and February 2014')
```

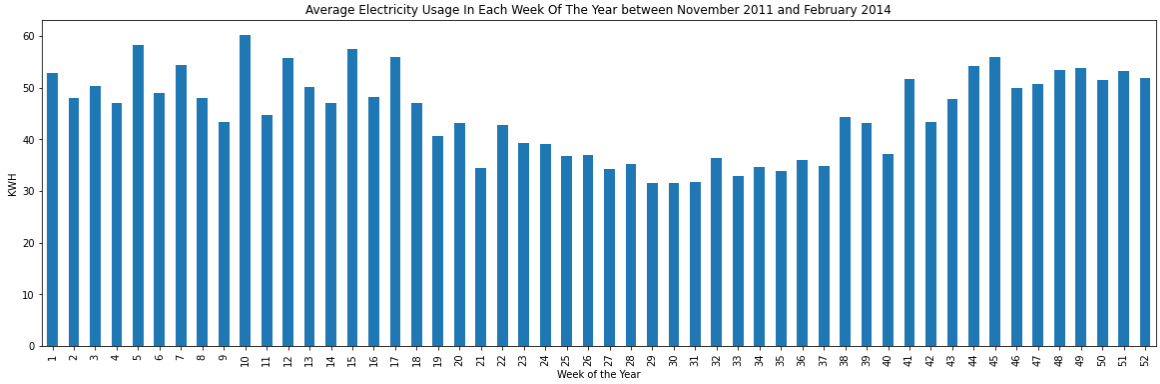


The bar graph represents the overall KWH of energy used on each day – over the given years. The overall consumption stays relatively similar, with slight fluctuations. The highest value is for 6th day of the month and is 7 KWH. On the other hand, the lowest value is around 5.5 KWH, observed mostly in the last week of the month.

Electricity usage per week

```
[ ] # Average Electricity Usage In Each Week Of The Year
    Weekly_Usage = dataCombined.groupby(by = dataCombined['DateTime'].dt.week).mean()
    Weekly_Usage = Weekly_Usage['KWH']*7
    plt.figure(figsize=(20,6))
    Weekly_Usage.plot.bar()
    plt.xlabel('Week of the Year')
    plt.ylabel('KWH')
    plt.title('Average Electricity Usage In Each Week Of The Year between November 2011 and February 2014')
```

```
Text(0.5, 1.0, 'Average Electricity Usage In Each Week Of The Year between November 2011 and February 2014')
```

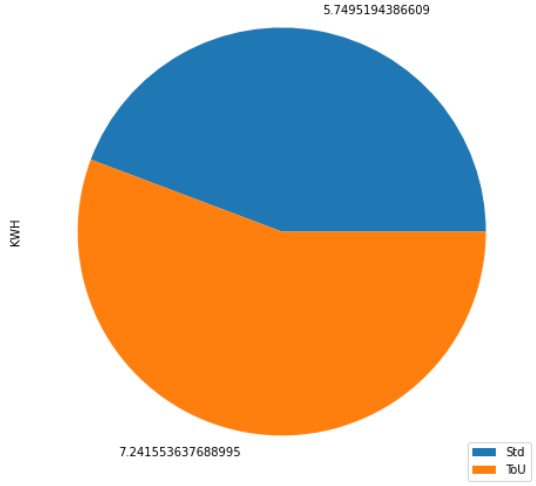


The average electricity usage in each week of the year keeps varying, with a small dip seen in the middle. There is no perfect trend as the values are fluctuating constantly. However, to overview, the values start decreasing at around week 19 and are lowest at week 30, after which they start increasing again. The lowest values are observed during the month of July, whereas the highest value is in week 10 – around mid-February.

Average Electricity Usage per Tariff Type

```
[ ]  # Average Electricity Usage Per Household Tarrif Type
     Usage_per_household_type = dataCombined.groupby('StdorToU').mean()
     vals=Usage_per_household_type['KWH']
     Usage_per_household_type.plot.pie(subplots=True, figsize=(8,8), labels = vals)
     plt.legend(['ACORN-U', 'ToU'], loc="lower right")
     plt.title('Average Electricity Usage Per Household Tarrif Type between November 2011 and February 2014')
```

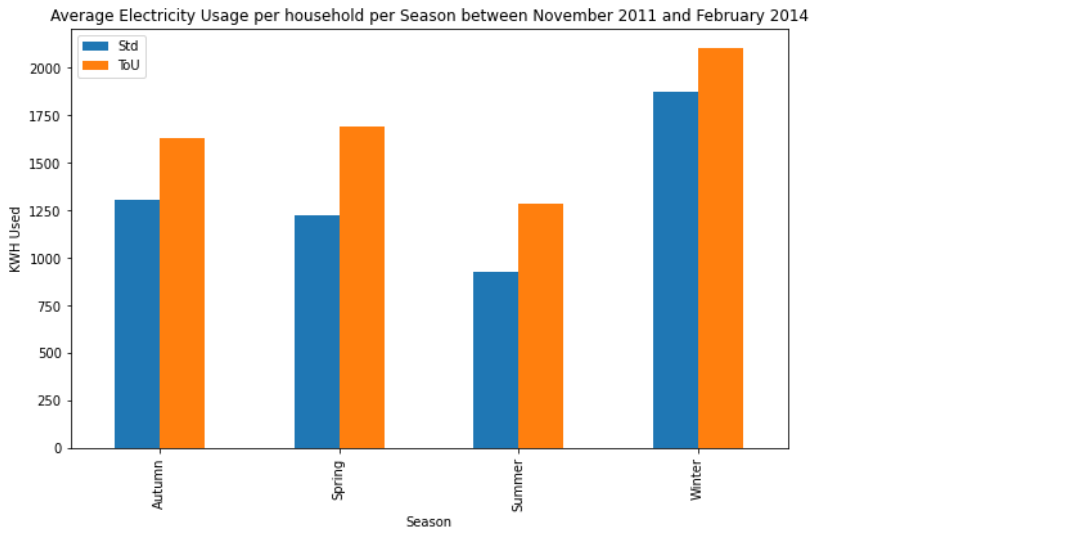
```
Text(0.5, 1.0, 'Average Electricity Usage Per Household Tarrif Type between November 2011 and February 2014')
```



The pie graph shows the overall consumption of energy by the two groups. It is higher for the group ToU with a value of 7.241, as compared to Std group, for which it's 5.749.

Average Electricity consumption per season

```
Text(0.5, 1.0, 'Average Electricity Usage per household per Season between November 2011 and February 2014')
```

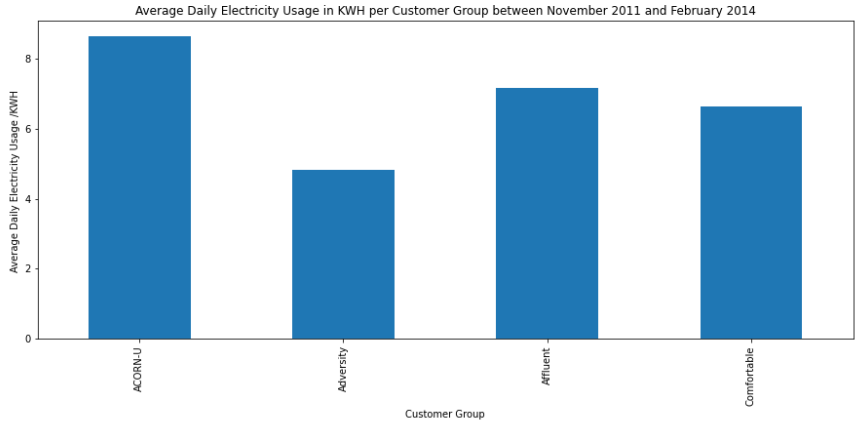


The months were divided into 4 seasons, Spring: March to May, Summer: June to August, Autumn: September to November and Winter: December to February. The most energy consumption for both groups is in Winter, and the least is in Summer. The increase in winter consumption might be due to an increased use of electric heaters etc. for warmth/heating.

Average Daily Electricty per customer group

```
[ ] # Average Daily Electricity Usage in KWH per Customer Group
    tempdf = dataCombined.copy()
    tempdf = tempdf.groupby('Acorn_grouped').mean()
    values = tempdf['KWH']
    new_values = [round(elem, 2) for elem in values]
    new_values
    tempdf.plot.bar(legend = False, figsize=(15,6))
    plt.title('Average Daily Electricity Usage in KWH per Customer Group between November 2011 and February 2014')
    plt.ylabel('Average Daily Electricity Usage /KWH ')
    plt.xlabel('Customer Group')
```

```
Text(0.5, 0, 'Customer Group')
```



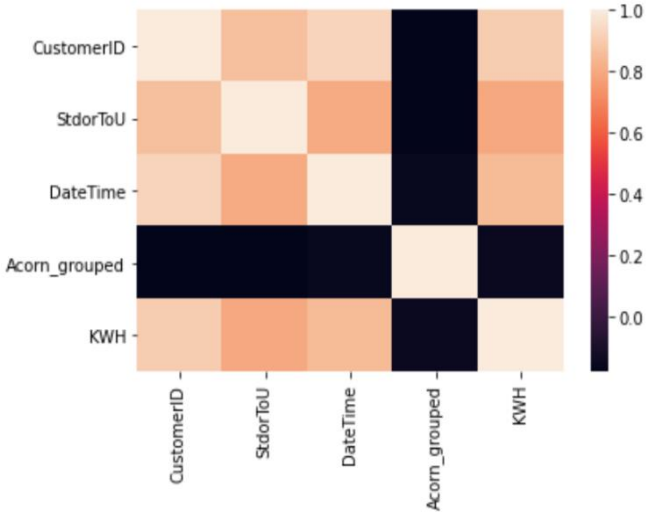
The Accorn-U has the most per day electricity consumption as compared to Adversity, for which it's the lowest. The max value is around 8.8 KWH and the lowest is 5 KWH.

Correlation

```
[ ] #Finding Out Correlation between Attributes

    sns.heatmap(dataCombined.apply(lambda x : pd.factorize(x)[0]).corr(method='pearson', min_periods=1))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb2a182e780>
```



The heatmap represents the correlation between the attributes. A correlation value of 0 shows no linear relationship and a value of 1 signifies a complete linear relationship. According to the heatmap, the Acorn_grouped has no dependency relation with any of the other attributes. StdorToU and KWH have a positive linear relationship with a value of 0.8. Further, the KWH and Datetime columns also have a positive relationship that is close to linear.

Dependency Matrix

THE DEPENDANCY MATRIX

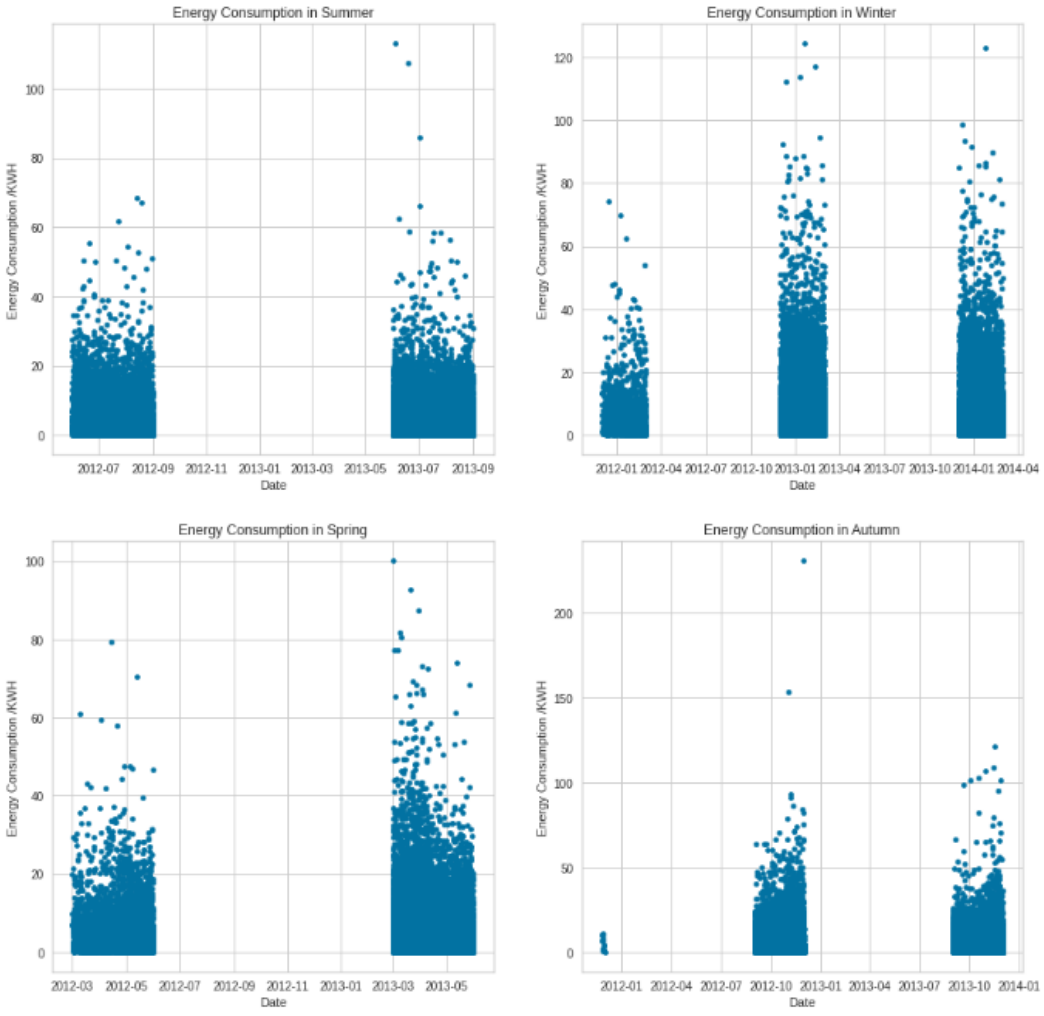
	StdorToU	Acorn_grouped	KWH
StdorToU	1.000000	0.181754	0.098727
Acorn_grouped	0.181754	1.000000	0.081725
KWH	0.098727	0.081725	1.000000

StdorToU and Acorn_grouped are categorical and KWH is the numerical attribute. The dependency value for KWH and StdorToU is 0.098, and between KWH and Acorn_grouped, it's 0.1817. Since the values are close to 0, there is almost no dependency between the attributes.

Cluster Analysis

The cluster analysis was done based on the Kilowatts energy consumed per hour. Since our dataset consisted of over 100,000 rows, partitional clustering approach was used since it has better complexity on larger data sets. Specifically, KMeans was applied to cluster the data.

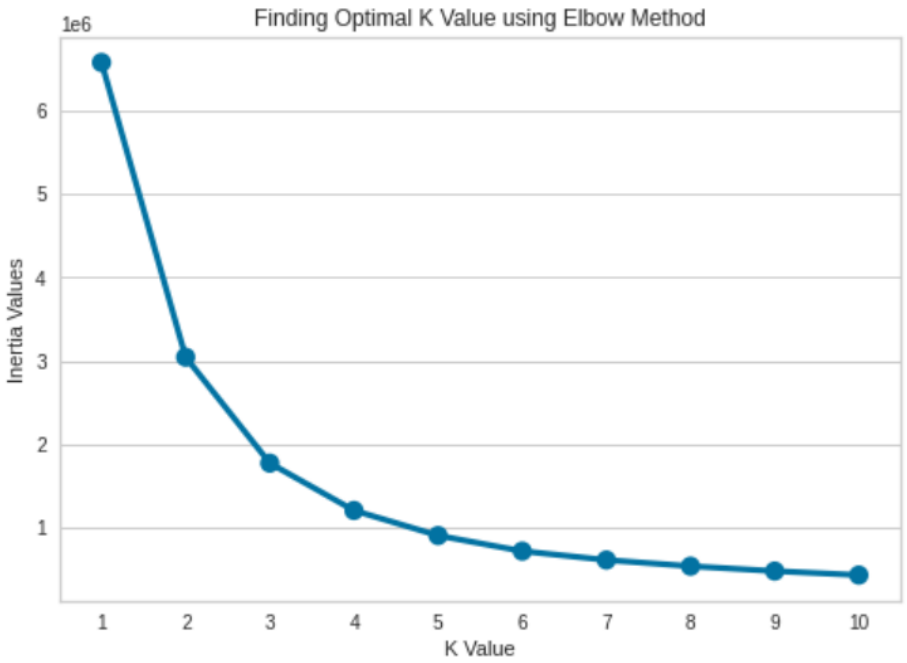
Initially, the data was divided into 4 seasons, Spring: March to May, Summer: June to August, Autumn: September to November and Winter: December to February. Scatter plots of KWH vs Months were made for all four seasons to get an approximate idea of how the data was spread in the particular seasons.



The graphs show two major clusters for Spring, Autumn and Summer, and three for Winter. The few extra points forming a smaller cluster in Autumn can be possible outliers.

K Value

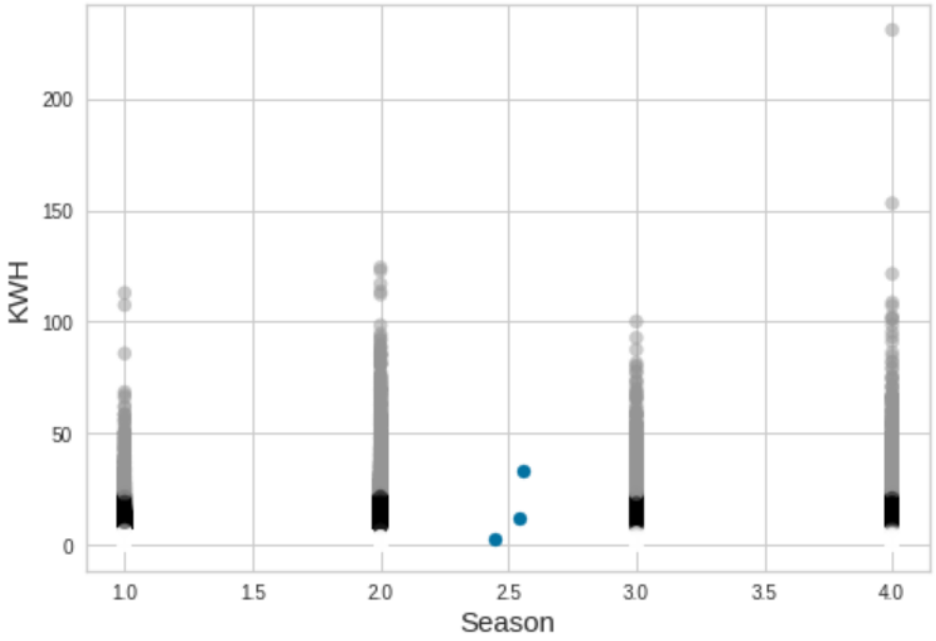
Since we had to apply K-means to the data, an optimal k value was required. The datetime and CustomerID columns were dropped as they did not pose any significance in our analysis. Also, K Means is suitable for numerical data only, hence, data reduction was carried out and the two columns were removed. The Acorn_grouped and StdorToU columns were mapped to numerical values to help in the evaluation.



The elbow method was used to fit the model with the values of k ranging from 1 to 11. The point of inflection of the curve is a reasonable indication that the data fits the model best at that particular value. In our case, the optimal k value was 3.

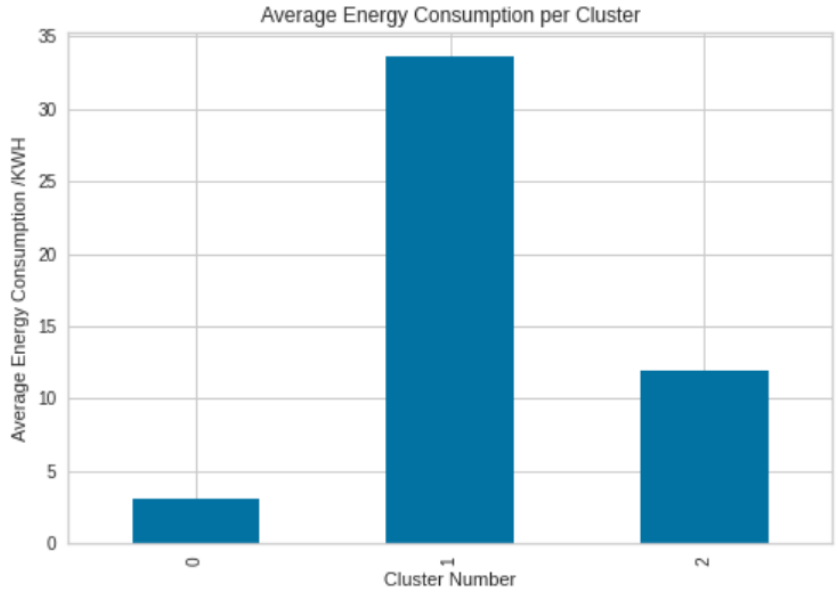
K Means

The built-in K Means algorithm with number of clusters = 3 was used to fit the dataset. A plot of KWH values vs Season was plotted, which showed four clustered regions – each showing the frequency of points in each season. The plot was solely for visualizing, to help in the further analysis.



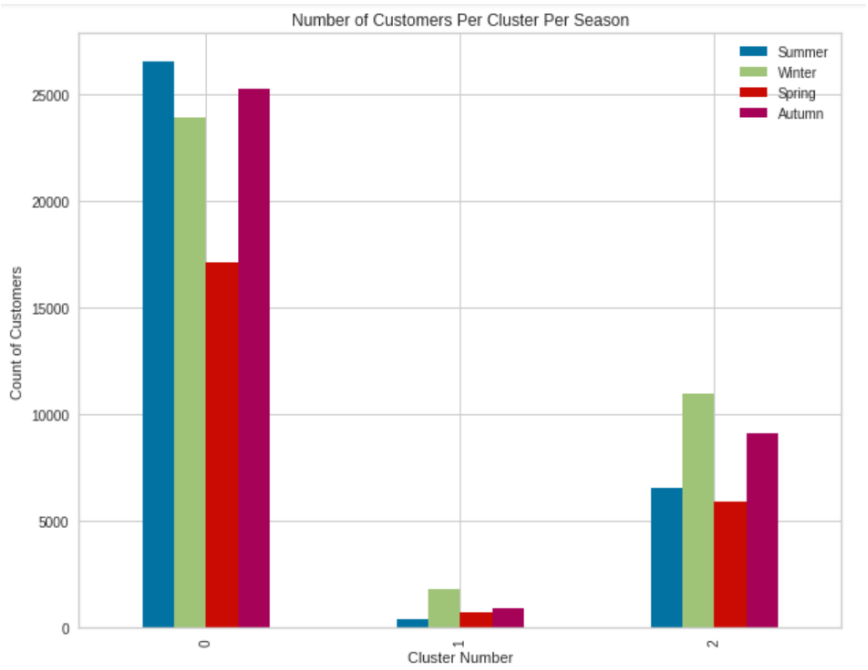
Bar plots were made to see the average energy consumption and the frequency of customers in each cluster.

Average energy consumption per cluster



Cluster 0 contained customers with the least average energy consumption, whereas cluster 1 had the ones who consumed the highest. For Cluster 0, the average value is around 3 KWH and for cluster 1, it is approximately 33 KWH. Cluster 2 lies between the two with an average KWH consumption of 12.

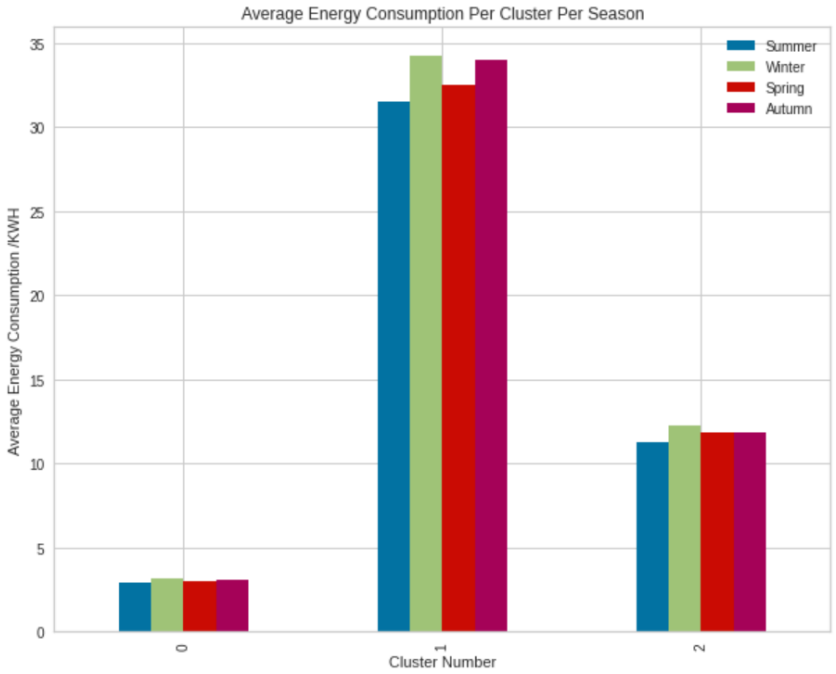
Customers per season per cluster



A bar graph indicating the frequency of customer in each cluster was plotted. The highest number of customers are in cluster 0, followed by cluster 2 and then cluster 1. For cluster 0, the highest frequency is in Summer and the lowest is in Spring. For cluster 2, the highest is in Winter and the

lowest in Spring. On the other hand, cluster 1 has the highest frequency of customers in Winter, while the lowest is seen in Summer.

Average energy consumption per cluster per customer

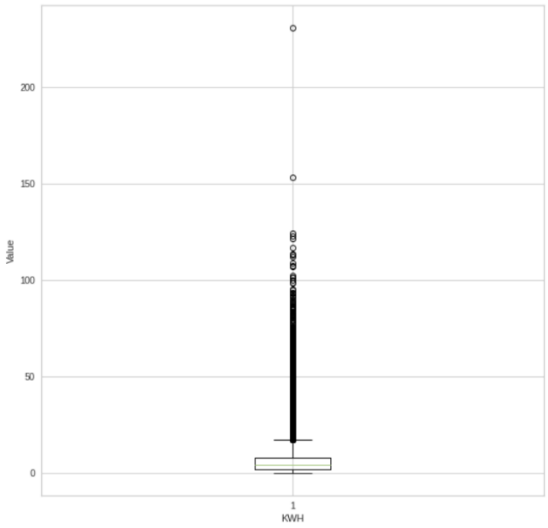


The average energy consumption per season in each cluster is approximately the same. Minor fluctuations are seen in Cluster 1. The energy consumption is highest in Winter and the lowest is in Summer for all three clusters. The consumption in Spring and Autumn is the same in cluster 2.

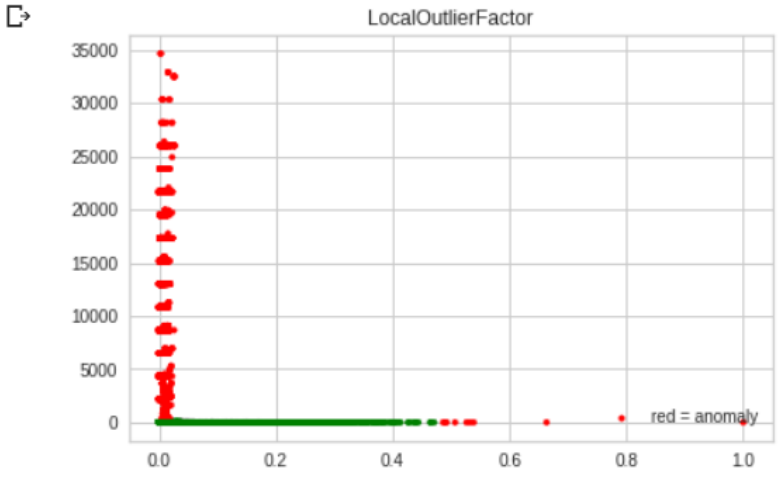
Outlier Analysis

Outlier analysis is done on the numerical data only, hence all the categorical attributes were dropped and only 'KWH' was retained. A box plot was made to get an idea of the outliers. Most of the outliers were clustered in the 30 – 100 value range, with two being above 150.

Box plot

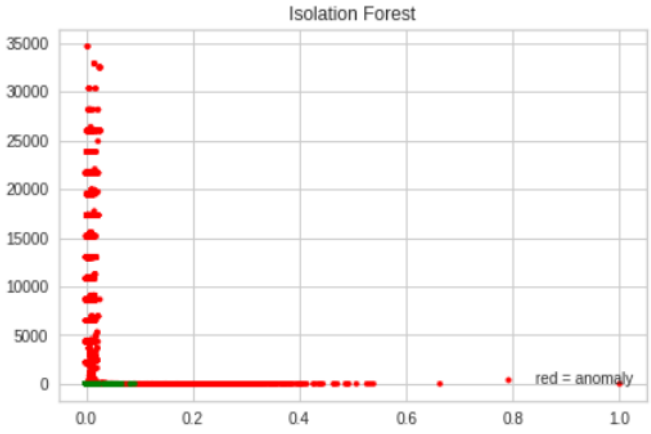


Local Outlier Factor



The values are normalized using MinMaxScaler function and the Local outlier factor method is applied. Two new columns were added – Error and Radius, and a scatter plot is made. The red points in the graph indicate anomalies. There is a range of anomalies, with different values.

Isolation Forest



Isolation forest algorithm was also applied on the data and a scatter plot was made. The number of outliers in this were more as compared to LOF, perhaps because it is more sensitive.

Conclusion

The project was a great learning opportunity for our group. By using different algorithms and filtering out data that was not necessary for our analysis, computations were made easier and time efficient. The analysis helped us understand the energy usage trends in different customer groups and how this data could be divided into clusters, which can then be used to describe the trends much more easily, since similar entries are grouped together. Outlier detection helped us eliminate the values that did not fit the model and were possible sources of error in our calculations.

Recommendations

Based on our cluster analysis, we can see that three groups of customers can be formed based on their energy consumption, namely Group 0 with average energy consumption around 3 KWH, Group 1 an average KWH consumption of 33 KWH, and Group 2 with average energy consumption of approximately 12 KWH. We can offer different rates based on these energy consumptions.

Also, we observed that different groups can be formed based on energy consumption in different seasons. The average energy consumptions of Group 0 and Group 2 remain more or less the same throughout different seasons and Group 1's energy consumption fluctuates a fair amount. Also, the information regarding the number of customers per clusters across different seasons is as follows:

0. Group 0 has a fairly large number of clients i.e. approximately 15,000 customers.
1. Group 1 has the lowest number of customers i.e. approximately 1,000 customers.
2. Group 2 has a fairly average number of customers i.e. approximately 7,500 customers.

Using the information presented above, the following recommendations can be made:

- The group with the highest customers, Group 0, can be offered relatively cheaper rates of electricity so as to retain their customer base.
- The group with the lowest number of customers i.e. Group 1 has the highest energy consumption so it should be offered higher rates of electricity.
- The remaining group, Group 2, which has a fairly average number of customers and a normal energy consumption, can be offered cheaper or subsidized electricity or some other incentives in order to benefit them. This could possibly attract more customers and increase the company's customers' base.