Machine Learning Report

# **Data Description and Research Questions**

------

# **Data Preparation & Cleaning**

First of all we load the data from our csv files and do data preparation and cleaning. The steps we have taken are following in sequence.

## **Removing Columns**

We removed all the unnecessary columns from our datasets. Weather dataset has 32 columns out of which only 11 are of use to us so we removed all the 21 columns which are not required.

## **Changing format of Date**

We change the format of the date column from String to Date Object type. We can use these Date methods on Data Objects to set and get month, day, year and many other functionality. We convert all out date columns to a standard Date Object which is needed for joining our datasets.

**Joining Datasets**

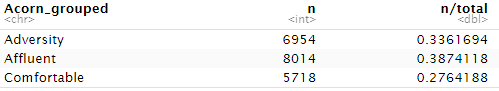
We joined the inform and household dataset on LCLid. Then join the resultant dataset with weather on date columns of respective datasets. Now we have joined all three dataset and can start analysing data more.

## **Creating New Features & Renaming Columns for Analysis**

We create new features and rename previous columns to give our dataset more intuition. We created new features weekday, month, year from the day column. We renamed the day column to date.

## **Stratified Sampling**

Our dataset is too large. To make it small we apply stratified sampling with Acorn\_grouped as the class. Stratified Sampling works in the way that it samples data with respect to class. It ensures that the sample reflects the diversity of the population. We have sampled our population to 2%.



## **Handling Not Applicable (NA) Columns and Re-ordering the columns**

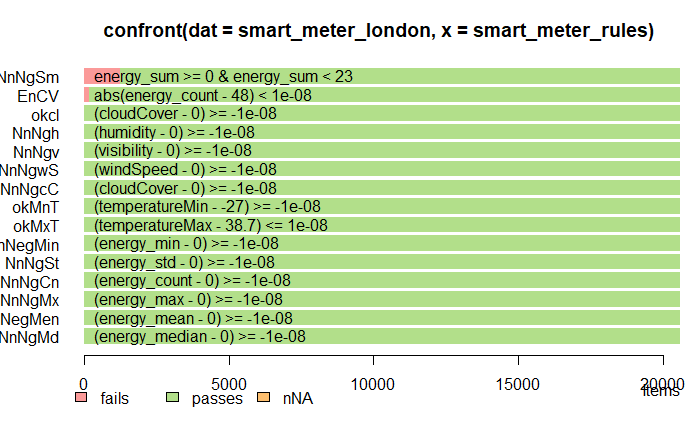
Now we check our columns for NA values and drop all the rows with NA values. NA values are the missing values that disturb our analysis.

## **Changing variable types**

We used Factor method to make categorical variables of categorical and numeric features of our dataset. It is memory efficient.

## **Validation the Data**

Validating the data means that we are checking whether the dataset meets the expectations we have about it? Therefore, we have defined some set of rules which check each record and each record will yield one answer whether it ‘Passes’ or ‘Fails’ that corresponding check.



The above results show that most of our data passes all the validation checks.

## **Cleaning Quality issue using Energy Count**

From the above validation check most of the records for ‘energy\_count’ are equal to 48 so we are only taking observations with ‘energy\_count’ 48.

## **Creating Categorical Variable from ‘energy\_sum’**

Now we have introduced a new feature ‘energy\_usage’ which yields output in the form of either ‘high’, ‘medium’ or ‘low’ depending on the value of the ‘energy\_sum’ variable.

# **Exploratory Data Analysis**

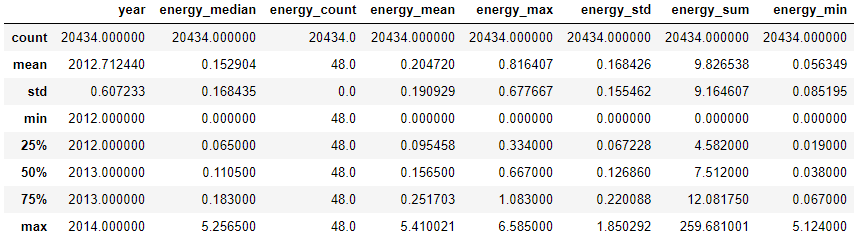
In Exploratory data analysis we would explore more about the data features, their correlation, distribution and relationship with output of the model.

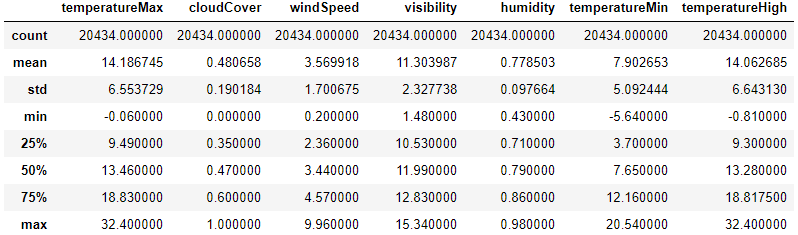
## **Exploring ‘summary’ column**

From the summary column we have generated ‘word cloud’ to know the different words used in summary.

## **Descriptive Statistics**

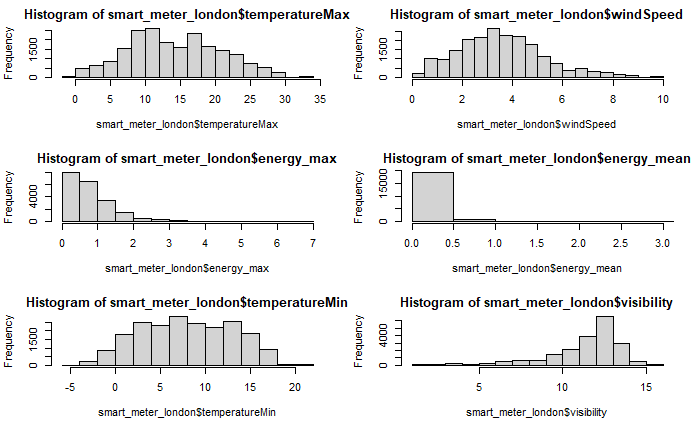
## Summary of the Dataset

Summary of the dataset gives us some insights of the data like the count, mean, median, quartiles, maximum and minimum. 

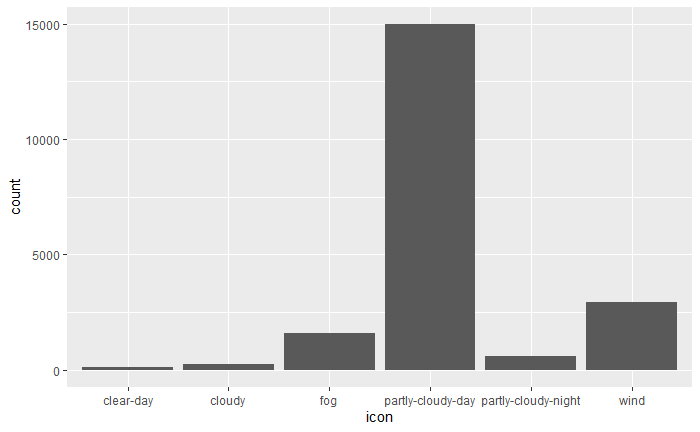


### 

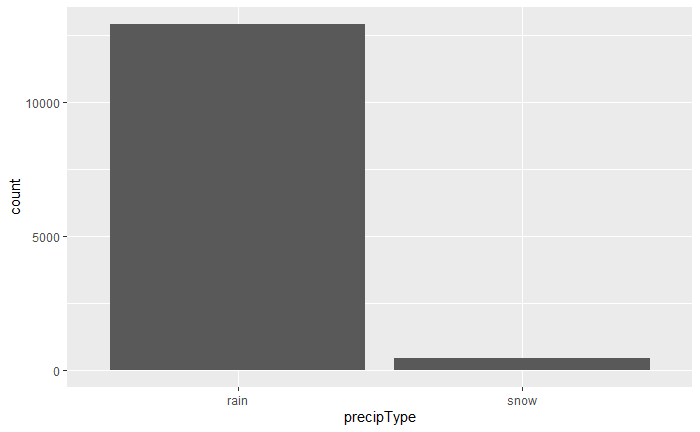
## Histograms of features

Histogram is used to show frequency distribution of the features.

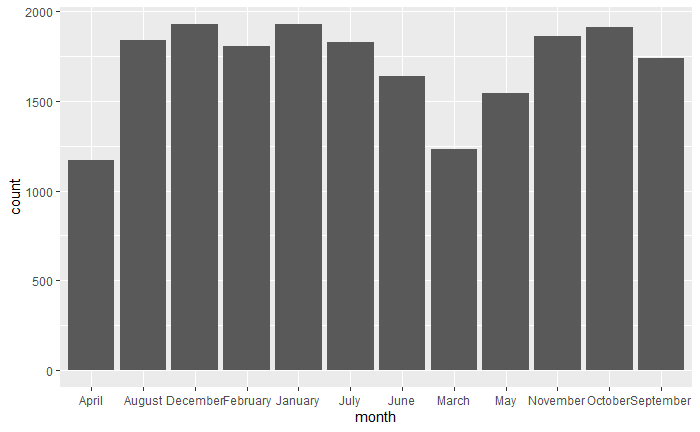
## Energy Consumption by different features



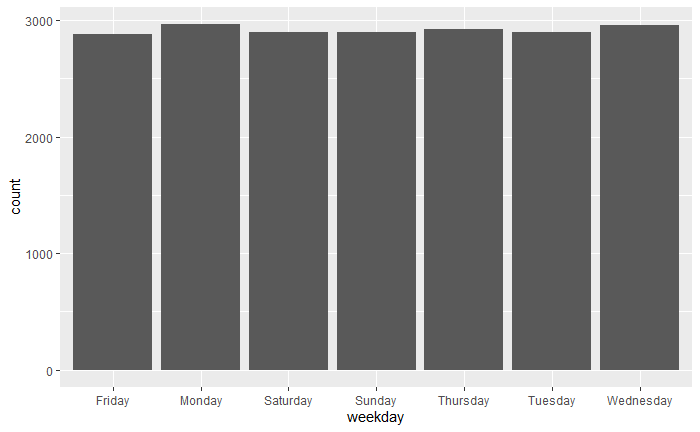
A higher value 15000 of energy sum on partly cloudy day indicates that the energy consumption on partly rainy day is highest.



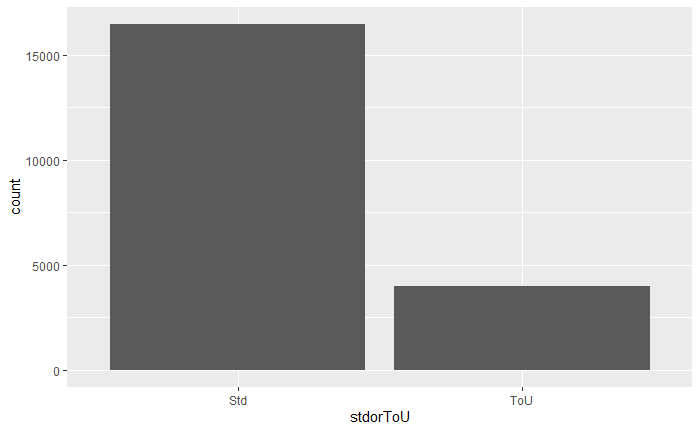
As the energy sum is high when rain. So energy consumption when rain, is higher than when it is snowing.



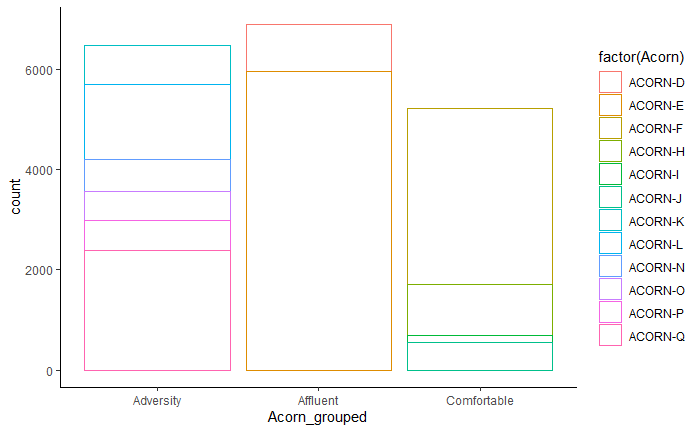
Energy Consumption in April and March are relatively less than the energy consumption in other months.



Energy consumption on every day of the week is approximately the same. It means that energy consumption is not dependent on the day of weeks but remain constant whole week.



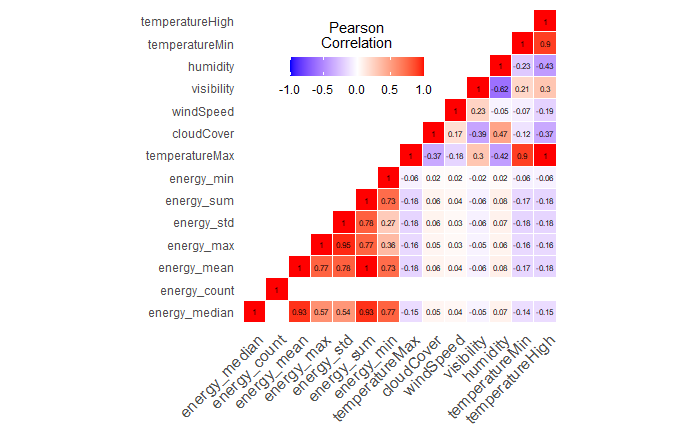
Energy consumption STD is very high compared to TOU.



For Acorn\_grouped Adversity class ACORN-Q has the highest energy consumption. For Affluent class ACORN-E has the highest consumption and for Comfortable class ACORN-F has the highest energy consumption. As a whole Affluent class has the highest energy consumption than other 2 classes.

## Correlation & Heat Map

Heat map is showing a strong relationship between energy\_min, energy\_max, energy\_std, energy\_sum, energy\_mean, energy\_count and energy median. These features have no correlation with any other features like temperature. Temperature\_min, temperature\_max and temperature\_high are strongly correlated. Cloud cover has correlation with humidity and visibility. Humidity is strongly correlated with visibility and temperature\_high.

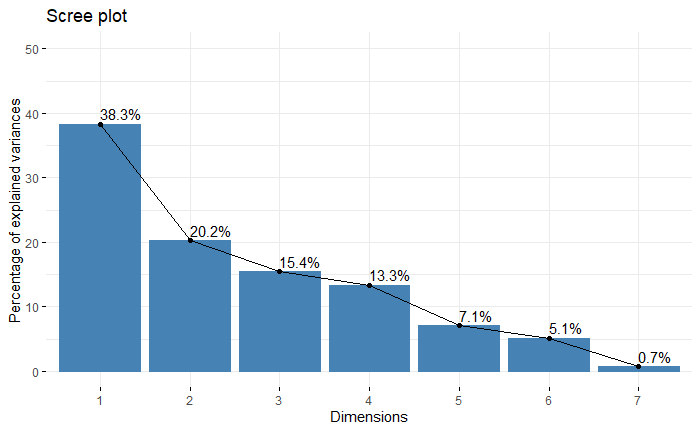


## Feature Selection

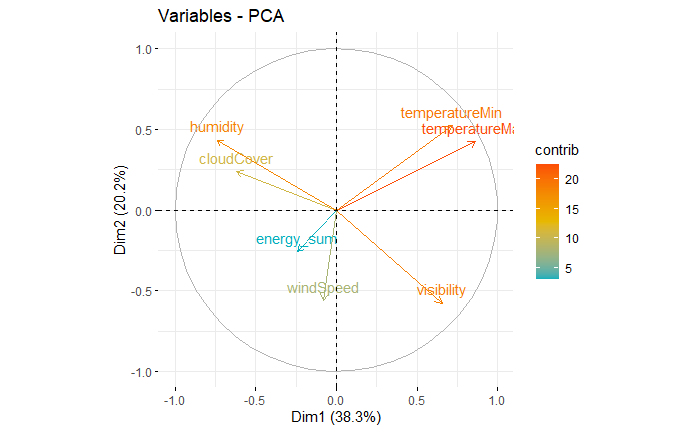
We selected energy\_sum, temperatureMax, cloudCover, windSpeed, visibility, humidity, temperatureMin as they are highly correlated.

## Principal Component Analysis of selected Features

We are applying PCA on selected features to reduce dimensionality.



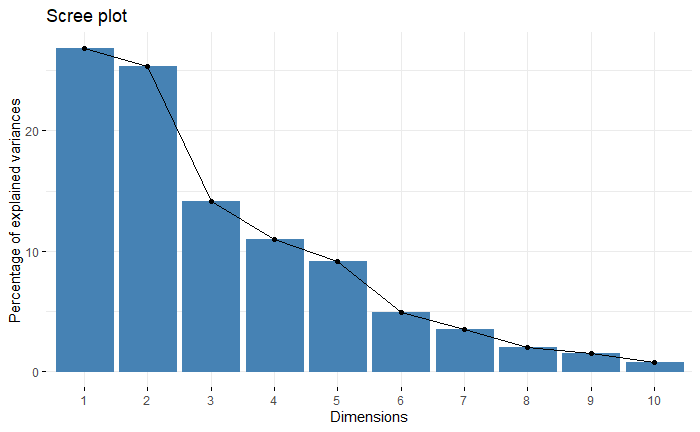
Above figure displays the proportion of variance explained by the 7 PCs. It is possible to notice that the first PC recovers around 38.3% of the variance of the original variables. This shows the great relevance of the component in summarizing the data, and, at a certain extent, this was expected, since many of the variables are correlated, therefore the component is able to explain the variance of these variables at the same time. The second component, instead, recovers around 20.2% of the variance. Although the percentage is not very high, we believe it could be interesting to analyse the second component, since it might recover the variance of a specific variable. We are not considering the remaining components relevant for the analysis. In conclusion, we will analyse the first 2 principal components, which recover a cumulative proportion of 60.5% of the variance.



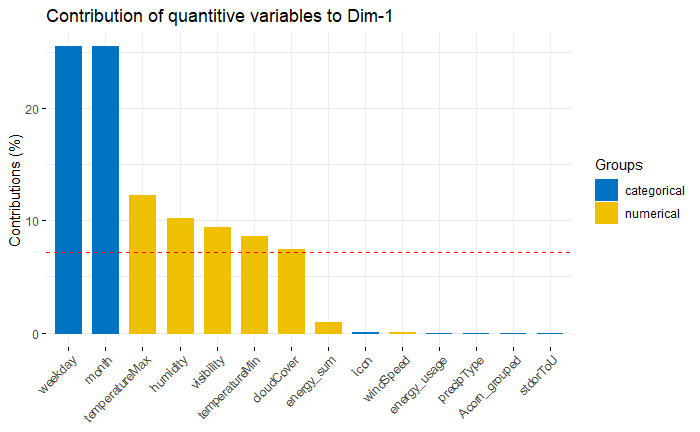
TemperatureMax, humidity, temperatureMin and visibility are important in the definition of the first principal component. On the other hand, Visibility, windSpeed and temperatureMin are more important in defining the second component. As a whole temperatureMax has the most contribution and temperatureMin, humidity, visibility also have large contributions as seen in the above figure.

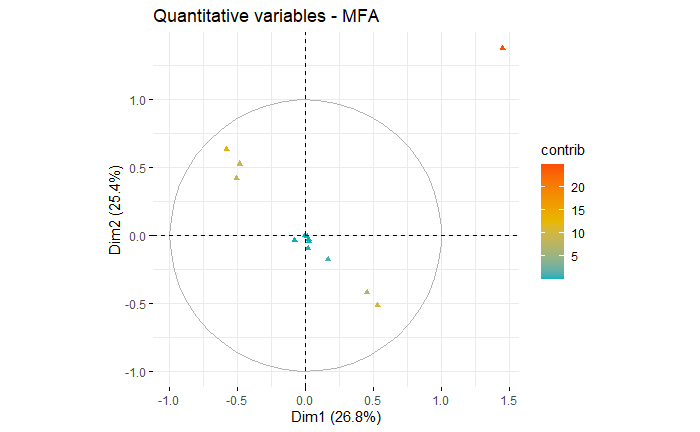
## **Multiple Factor Analysis (MFA)**

Multiple Factor Analysis (MFA) is a method for summarizing and visualizing complex data tables in which individuals are structured into groups. The number of variables in each group may differ and the nature of the variables can vary from one group to the other but the variables should be of the same nature in a given group .



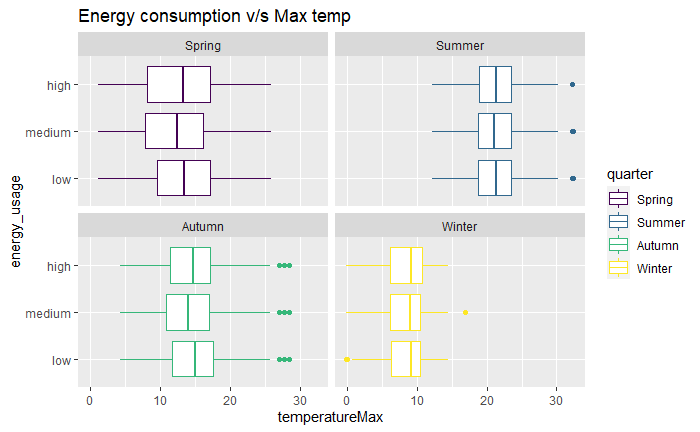
It is noticed that dimension 1 and dimension 2 have the most variance of original variables. Dimension 1 and dimension 2 are contributed by Categorical and Numerical Data almost equally. For our analysis we will take Dimension 1 and Dimension 2 only as they recover a high cumulative proportion.

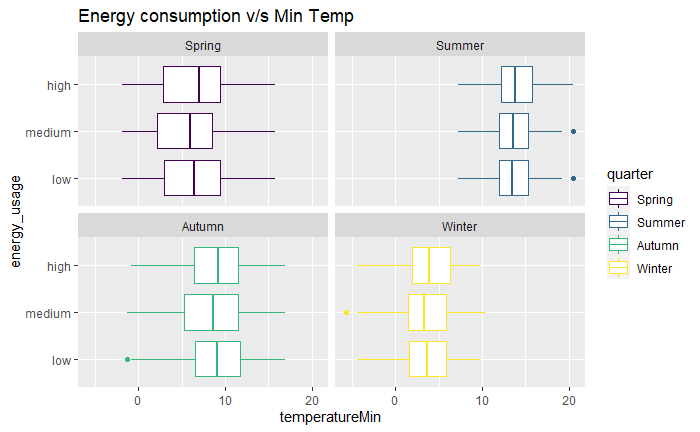


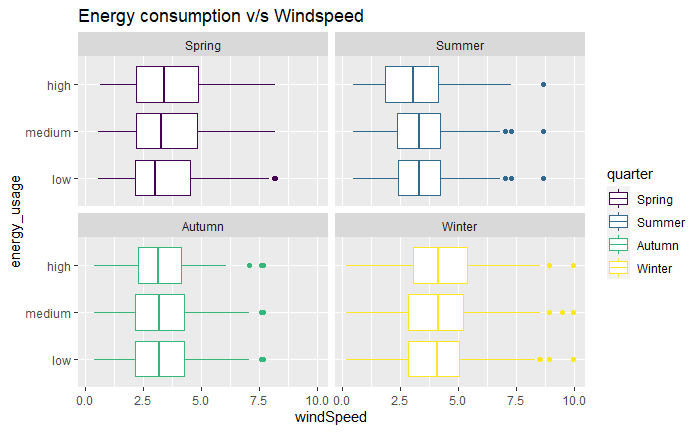


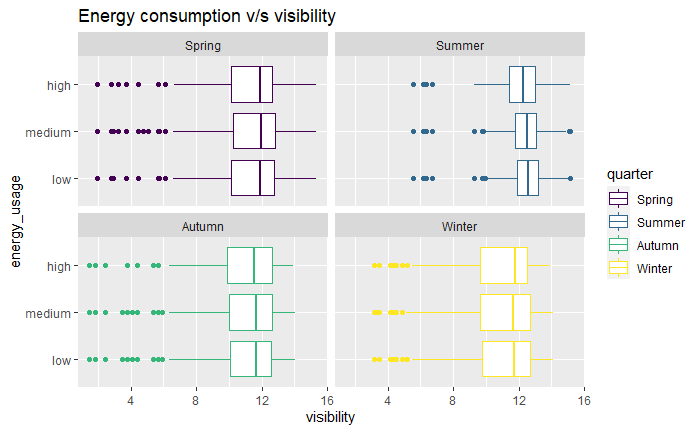
As we can see from the figures that the weekday and month categorical features have most contribution in dimension 1 and dimension 2 as they are far from the x and y axis. Some numerical features in yellow have high contribution in dimension 1 or dimension 2. Features in the center of the graphs have very less contribution in dimension 1 and 2.

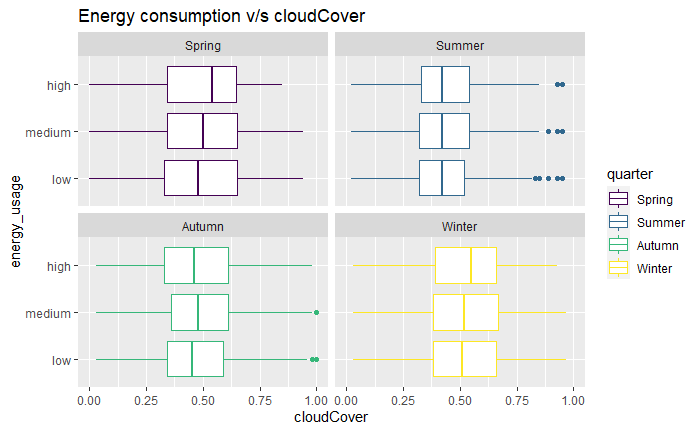
**Bi Variant Plots**

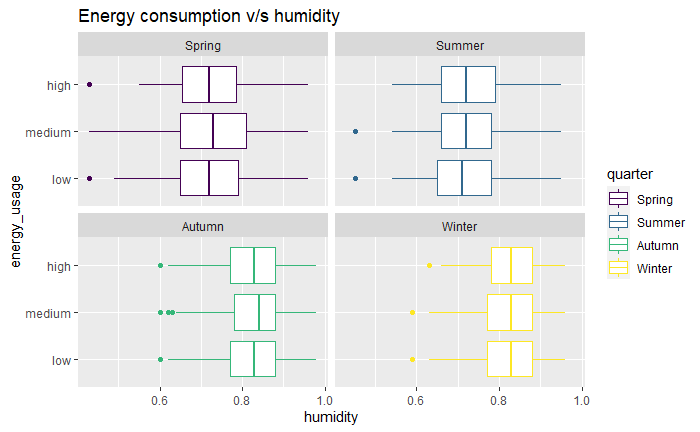
****

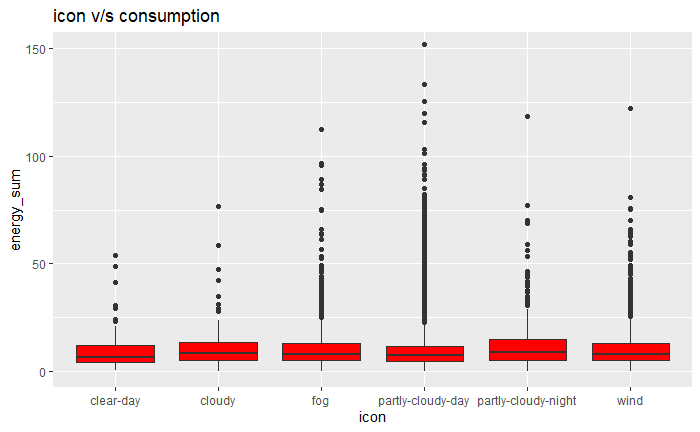
****

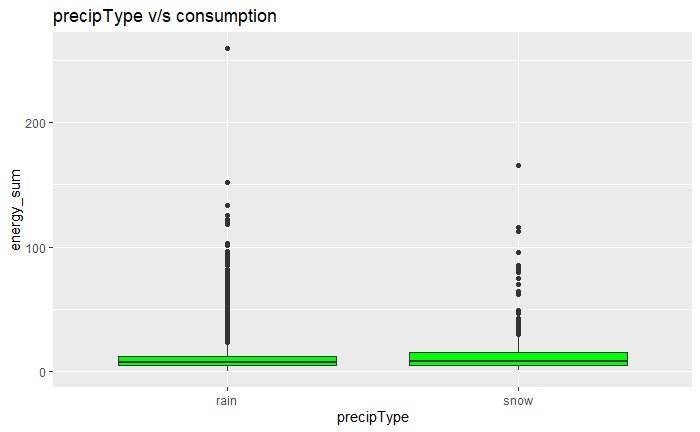
****

****

****

****

****

****

## **Removing Outliers**

An outlier is an observation that is at an abnormal distance from another observation of the random sample. We removed outliers.

# 

# **Machine Learning & Prediction**

**Applying Classification**

In classification, a class label is predicted for a given example of input data. We have to select input and output of the models.

## **Input**

After exploring the dataset, now we are selecting numeric variables and scaling them. We also have categorized non numeric variables into numeric values. We have also dropped features which are not relevant.

## **Output**

The output variable for our task in ‘energy\_usage’.

1. **Binary Target Variable**

The variable ‘energy\_usage’ contains two classes ‘medium’ or ‘high’ and with 0 as respective numeric value for ‘normal’ and 1 as respective numeric value for ‘high’.

1. **Three classes Target Variable**

The variable ‘energy\_usage’ contains three classes ‘low’, ‘medium’ or ‘high’ with 1 as respective numeric value for ‘low’, 2 as respective numeric value for ‘medium’ and 3 as respective value for ‘high’.



## **Model:**

We have split our data 20% and 80% for testing and training respectively. We applied Classification with Input Provided above and predicted ‘energy\_usage’.

We have used the following classification models:

* **Support Vector Machines (SVMs)**

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the image below.

* **K-Nearest Neighbors (KNN)**

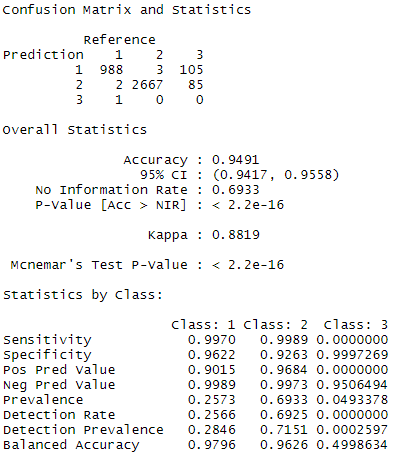
KNN is a simple algorithm which uses data and classifies new data points based on similarity measures (e.g. distance function).

* **Random Forest Classifier**

Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

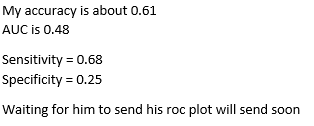
## **Support Vector Machines using target variables with 3 Classes.**

The first model which was implemented on the dataset was a support vector machine. The K-Cross Validation method was used while validating which involves splitting the dataset into k-subsets. The value of K was set to 10 in our case. For each subset is held out while the model is trained on all other subsets. This process is completed until accuracy is determined for each instance in the dataset, and an overall accuracy estimate is provided.



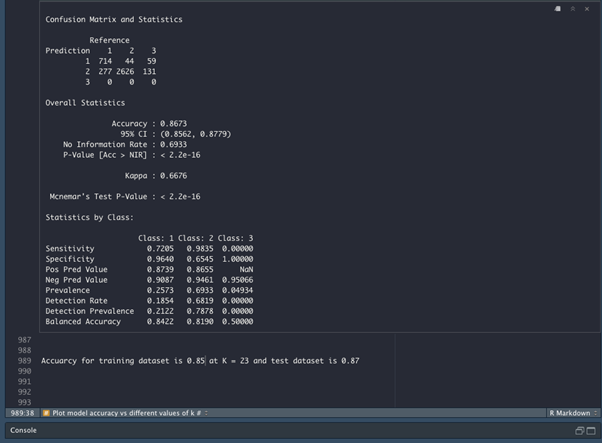
## **Support Vector Machine using Binary target variable**

The second model which was implemented on the dataset was SVM on a binary target variable. The accuracy of this model was lower as compared to SVM on 3 classes.



**K-Nearest Neighbors (KNN) using target variables with 3 Classes.**

The third model which was implemented on the dataset was K-Nearest Neighbors. The accuracy of this model was lower than the SVM model on target variable with 3 classes but greater than all other models.



**Random Forest Classifier (RFC) using Binary Target Variable with different sampling techniques**

The fourth model which was implemented on the dataset was Random Forest Classifier. Different sampling techniques were used in this model (over sampling, undersampling, both and ROSE).

| **Random Forest Classifier** | **Accuracy** | **Sensitivity** | **Specificity** | **ROC** |
| --- | --- | --- | --- | --- |
| no sampling | 0.6457 | 0.1794 | 0.87076 | 0.5596 |
| oversampling | 0.5532 | 0.4906 | 0.5834 | 0.5589 |
| undersampling | 0.5376 | 0.5461 | 0.5335 | 0.5578 |
| both samping | 0.5421 | 0.5196 | 0.5530 | 0.5508 |
| ROSE | 0.5617 | 0.5458 | 0.5694 | 0.5836 |

# **High Performance Computation Implementation**

---------

# **Performance Evaluation & Comparison of Methods**

We used 4 different approaches for solving our problem which was to predict energy consumption.

| **Models** | **Accuracy** | **Sensitivity** | **Specificity** |
| --- | --- | --- | --- |
| ***SVM (3 classes*)** | **0.9491** | Class 1 = 0.9970  Class 2 = 0.9989  Class 3 = 0.0000 | Class 1 = 0.9622  Class 2 = 0.9263  Class 3 = 0.9997 |
| SVM (binary) | 0.61 | 0.68 | 0.25 |
| KNN (3 classes) | 0.8673 | Class 1 = 0.7205  Class 2 = 0.9835  Class 3 = 0.0000 | Class 1 = 0.9640  Class 2 = 0.6545  Class 3 = 1.0000 |
| RFC  (no sampling)  (binary) | 0.6457 | 0.1794 | 0.87076 |
| RFC (oversampling)  (binary) | 0.5532 | 0.4906 | 0.5834 |
| RFC (undersampling)  (binary) | 0.5376 | 0.5461 | 0.5335 |
| RFC  (both samping)  (binary) | 0.5421 | 0.5196 | 0.5530 |
| RFC  (ROSE)  (binary) | 0.5617 | 0.5458 | 0.5694 |

## **Discussion of Findings**

**Dataset and Analysis**

All in all, we can conclude our findings in the following points:

* Most observations have energy\_count of 48.
* The **energy consumption** is higher on a partly cloudy day compared to other days as the energy sum of partly cloudy days is much higher than other days. Higher energy sum is observed when it is raining compared to when it is snowing.
* In the months of March and April there is less energy consumption than the rest of the year. Energy consumption remains constant all over the week. It means that energy consumption is not dependent on the day of the week. Energy consumption for STD is much higher than TOU.
* A strong **correlation** exists between energy\_min, energy\_max, energy\_std, energy\_sum, energy\_mean, energy\_count and energy median. They all are highly correlated with each other but have no correlation with any other feature of the dataset.
* Temperature\_min, temperature\_max and temperature\_high are strongly correlated. Cloud cover has correlation with humidity and visibility. Humidity is strongly correlated with visibility and temperature\_high.
* **In PCA**, the 1st PC recovered around 38.3% of the variance of the original variables. This shows the great relevance of the component in summarizing the data, and, to a certain extent, this was expected, since many of the variables are correlated.
* The second component recovers around 20.2% of the variance. This might not be as high as the first component but it would be interesting to analyse as it might recover the variance of a specific variable.
* First 2 principal components recover a cumulative proportion of 60.5% of the variance of variables.
* TemperatureMax, humidity, temperatureMin and visibility are important in the definition of the first principal component. As we have discussed earlier, they are highly correlated. Due to their strong correlation the component is able to explain the variance of these variables at the same time. Visibility, windSpeed and temperatureMin are more important in defining the second component.
* By plotting features on Component 1 and Component 2 as axis we revealed the temperatureMax has the most contribution in the components and temperatureMin, humidity, visibility also have large contributions in the components.
* **In MFA**, we choose to do analysis on 2 dimensions which contain the highest proportion of variance. Both dimensions have almost equal contribution of Categorical and Numerical data.
* In Dimension 1, the top 2 contributors are weekday and month which are both categorical in nature. All other contributors with large contributions are numerical.
* The weekday and month categorical features have most contribution in dimension 1 and dimension 2 as well they are far from the x and y axis of plot with Dimension 1 and Dimension 2 as axis. Some numerical features in yellow have high contribution in dimension 1 or dimension 2. Features in the center of the graphs have very less contribution in dimension 1 and 2.

**Models**

Different models with different approaches were used to analyze the findings:

* **SVM** using 3 classes as target variable resulted in highest accuracy of 0.9491
* **KNN** using 3 classes as target variable resulted in the second highest accuracy of 0.8673
* **Random Forest Classifier** and SVM using **binary target** variables did not perform well as their accuracies were moderate.
* The results showed that using **3 classes** ‘low’, ‘medium’ and ‘high’ instead of 2 binary classes ‘medium’, and ‘high’ yields higher accuracy!

## **Data Management Planning and author Contribution statement**

----