# ASSIGNMENT 1 – REPORT

## INDEX

# DATA WRANGLING AND CLEANSING

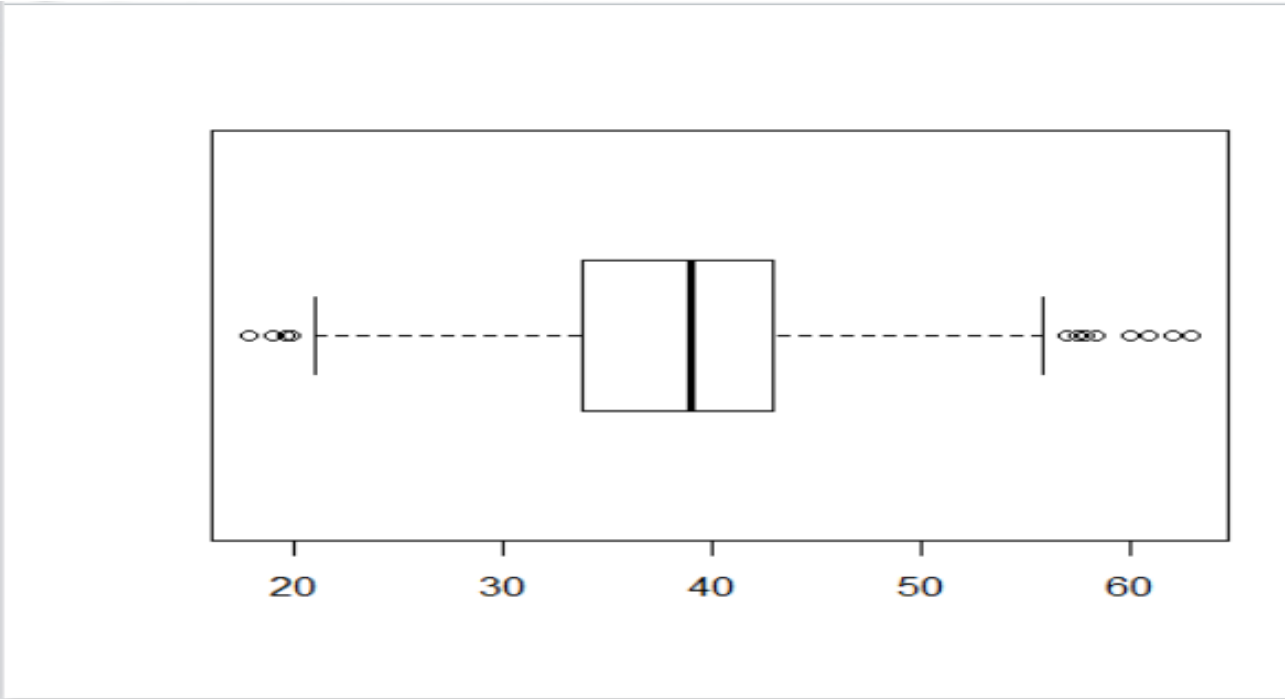
- Function to browse and select input file from the local system
- Performed functionalities on the “rawdata” Dataset
  1. Sorting the data by variable
  2. Rearranging the order of columns
  3. Renaming the columns
  4. Reshaping the data from wide to long format
  5. Visualizing the outliers using the BoxPlot
  6. Takes the start_date and end_date for getting the temperature data from the rawdata file without any manual interventions
  6. Function to detect and replace the outliers by “NA”
  7. Finding “NA” and replacing them with the mean of two consecutive observations for Temperature and kWh
  8. Implemented binning of data
  9. Manipulated the different datatypes of the columns
  10. Aggregation of data
  11. Merging of two different data frames (ie: rawdata and weatherdata) using LEFT OUTER JOIN
  12. Removed NA’s and Outliers for the merged data
  13. Rounding of the decimal points to ZERO in Temperature (given Format)

# REGRESSION MODEL AND PERFORMANCE METRICS

1. Datatype played an important factor for creating regression model
2. We observed taking columns factor data type is ideal for building regression model as it brings down the residual standard error
3. We split the data in 80-20% ratio for train and test data respectively
4. Implemented the Multi Linear regression model for training data set
5. The order of ignoring the predictors while building regression model played an important role in values of the coefficients as well as the residual standard error
6. Implemented the Performance evaluation on Test dataset
7. Calculated Predictive accuracy using the performance evolution on test data and Power consumption from the training data

# WORK-FLOW

## 1. Outliers- Boxplot




## 2. Data after merging rawdata and temperature data with NA values

```
> summary(sampleformat)
    Account               Date                 kwh             month
 Min.   :2.644e+10   Min.   :2014-01-01   Min.   : 60.55   Length:8760
 1st Qu.:2.644e+10   1st Qu.:2014-04-02   1st Qu.: 82.03   Class :character
 Median :2.644e+10   Median :2014-07-02   Median :122.59   Mode  :character
 Mean   :2.644e+10   Mean   :2014-07-02   Mean   :154.19
 3rd Qu.:2.644e+10   3rd Qu.:2014-10-01   3rd Qu.:231.33
 Max.   :2.644e+10   Max.   :2014-12-31   Max.   :405.98
                                          NA's   :1
     day                 year                hour          DayofWeek  Weekday
 Length:8760        Length:8760        Min.   : 0.00   Min.   :0   0:2496
 Class :character   Class :character   1st Qu.: 5.75   1st Qu.:1   1:6264
 Mode  :character   Mode  :character   Median :11.50   Median :3
                                       Mean   :11.50   Mean   :3
                                       3rd Qu.:17.25   3rd Qu.:5
                                       Max.   :23.00   Max.   :6

    Peakhour     Temperature
 Min.   :0.0   Min.   : 2.00
 1st Qu.:0.0   1st Qu.:37.00
 Median :0.5   Median :52.00
 Mean   :0.5   Mean   :50.95
 3rd Qu.:1.0   3rd Qu.:66.00
 Max.   :1.0   Max.   :92.00
               NA's   :73
```

There were 73 NAs in temperature and 1 NA in kWh after merging raw and weather data

## 2. Output after cleaning the rawdata

```
> summary(sampleformat)
    Account                Date                 kWh              month
 Min.   :2.644e+10   Min.   :2014-01-01   Min.   : 60.55   Min.   : 1.000
 1st Qu.:2.644e+10   1st Qu.:2014-04-02   1st Qu.: 82.04   1st Qu.: 4.000
 Median :2.644e+10   Median :2014-07-02   Median :122.59   Median : 7.000
 Mean   :2.644e+10   Mean   :2014-07-02   Mean   :154.18   Mean   : 6.526
 3rd Qu.:2.644e+10   3rd Qu.:2014-10-01   3rd Qu.:231.33   3rd Qu.:10.000
 Max.   :2.644e+10   Max.   :2014-12-31   Max.   :405.98   Max.   :12.000
      day              year           hour          DayofWeek     Weekday
 Min.   : 1.00   Min.   :2014   Min.   : 0.00   Min.   :0   Min.   :1.000
 1st Qu.: 8.00   1st Qu.:2014   1st Qu.: 5.75   1st Qu.:1   1st Qu.:1.000
 Median :16.00   Median :2014   Median :11.50   Median :3   Median :2.000
 Mean   :15.72   Mean   :2014   Mean   :11.50   Mean   :3   Mean   :1.715
 3rd Qu.:23.00   3rd Qu.:2014   3rd Qu.:17.25   3rd Qu.:5   3rd Qu.:2.000
 Max.   :31.00   Max.   :2014   Max.   :23.00   Max.   :6   Max.   :2.000
    Peakhour     Temperature
 Min.   :0.0   Min.   : 2.00
 1st Qu.:0.0   1st Qu.:37.00
 Median :0.5   Median :52.00
 Mean   :0.5   Mean   :51.01
 3rd Qu.:1.0   3rd Qu.:66.00
 Max.   :1.0   Max.   :92.00
>
```

## 3. Regression

```
> summary(lm.fit)

Call:
lm(formula = kWh ~ . - Account - Date - year, data = sampleformat)

Residuals:
     Min       1Q   Median       3Q      Max
-116.098  -42.321   -1.447   36.636  196.834

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.36588    3.11703  -9.421  < 2e-16 ***
month        -0.59160    0.16866  -3.508 0.000454 ***
day          -0.43060    0.06182  -6.966  3.5e-12 ***
hour          0.24802    0.07958   3.117 0.001835 **
DayofWeek     4.76223    0.27237  17.484  < 2e-16 ***
weekday      69.69959    1.20480  57.852  < 2e-16 ***
Peakhour    108.50743    1.10932  97.815  < 2e-16 ***
Temperature   0.06363    0.03247   1.960 0.050048 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.88 on 8752 degrees of freedom
Multiple R-squared:  0.6124,    Adjusted R-squared:  0.6121
F-statistic:  1976 on 7 and 8752 DF,  p-value: < 2.2e-16
```

# 4. Forecast

```
> summary(forecastData)
      Date                month              day                year               hour          DayofWeek      Weekday
 Min.   :2014-12-01   Length:744         Length:744         Length:744         Min.   : 0.00   Min.   :0.000   0:192
 1st Qu.:2014-12-08   Class :character   Class :character   Class :character   1st Qu.: 5.75   1st Qu.:1.000   1:552
 Median :2014-12-16   Mode  :character   Mode  :character   Mode  :character   Median :11.50   Median :3.000
 Mean   :2014-12-16                                                           Mean   :11.50   Mean   :2.903
 3rd Qu.:2014-12-24                                                           3rd Qu.:17.25   3rd Qu.:5.000
 Max.   :2014-12-31                                                           Max.   :23.00   Max.   :6.000
    Peakhour     Temperature
 Min.   :0.0   Min.   :17.96
 1st Qu.:0.0   1st Qu.:33.89
 Median :0.5   Median :39.02
 Mean   :0.5   Mean   :38.19
 3rd Qu.:1.0   3rd Qu.:42.98
 Max.   :1.0   Max.   :62.96
```

## INFERENCES AND CHALLENGES

- Handling date type format of date variable while pulling the data from wunderbur.com and converting it to date format
- To get the data formats to similar types before merging two data frames.