**ASSIGNMENT 1 – REPORT**

**INDEX**

**1. DATA WRANGLING AND CLEANSING**

**2. REGRESSION MODEL AND PERFORMANCE METRICS**

**3. PREDICTION**

**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. Visualizing the outliers using the BoxPlot

5. Function to detect and replace the outliers by “NA”

6. Finding “NA” and replacing them with the mean of two consecutive observations for Temperature and kWh

7. Implemented binning of data

8. Manipulated the different datatypes of the columns

9. Aggregation of data

10. Merging of two different data frames (ie: rawdata and weatherdata) using LEFT OUTER JOIN

11. Removed NA’s and Outliers for the merged data

12. 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