## **Demand Estimation Data Analysis Error Reporting and Findings**

Objective – To report the errors, limits and findings post cleaning and analysing the data of the daily gas consumption that is provided by the industry, as it is extremely important that only excellent quality data is passed for the modelling of data to predict behaviours and see patterns.

Errors in the data:

1.While there were no null values in the data there were duplicates in the Gasdemand file

See below:

Table

Description automatically generated

The duplicated were removed for further analysis as they can skew the results and predictions.

2. After merging the data sets together I came across the null value

See below:Table

Description automatically generated

These value stems from the gasdemand dataframe where the values for some dates are not available.The dataframe size is 18,240 while for 50 locations for 366 days it should be 18300

There could be several reasons for missing values and that require further investigation and more information to work with.

The Null values however were dropped as there were not major implications on other column values

### Limitation:

1. Even though not an error but there is a limitation that in the dataset. The usage value is annual while CWV data is a daily data which limits the scope to find any correlation to establish between the two. As the gas usage and CWV are linearly related to each other, and it would have been good to visualise but that could not be done.

### Findings:

1. The maximum and minimum values for Energy\_Value are following:

maximum CWV in the given period is 103936.0 for the site

9247 WM21

minimum CWV in the given period is -216.0 for the site

2446 SC22

There is a large variation in this data that has led to outliers that cannot be ruled out.

This need more information to be analysed further. The outliers were not dropped because

they are site dependent

See the boxplot and scatter plot showing for both Scotland and West Midlands

Scotland

Chart

Description automatically generatedChart

Description automatically generated

West Midlands**Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated**

As can be seen from both the scatter and boxplot of both the regions sites from Energy value is mostly concentrated on the lower side of the median with some values being on the higher side but as the data is dependent on

different site there can be a multitude of reasons for this behaviour and the outliers cannot be dropped.

1. There is a strong correlation between Energy\_Value and Annual\_Usage with the correlation value (Pearson)

being/The value shows that the Energy\_Value is related to the Annual\_Usage

Scatter plot below:

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

1. There is a strong negative correlation between Energy\_Value and CWV. With value of indicating that as the

value of CWV goes down the Energy\_Value goes up and vice-versa.

Scatter Plot below:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

## Conclusions

1. Data provided is over a period of a year and that helps to establish a relation between key factors like usage and energy value and CWV.
2. As mentioned in limitation there is not enough usage data available for analysis in respect to the CWV
3. On a glance a lot of values seem like outliers, but further analysis shows they cannot be dropped and need more thorough analysis and information also emphasizing the point that visualization is very important for a fair and correct analysis
4. More information can be helpful for further analysis like details on site references, more historical data, information on energy value and the units of all the numerical data in the data set.