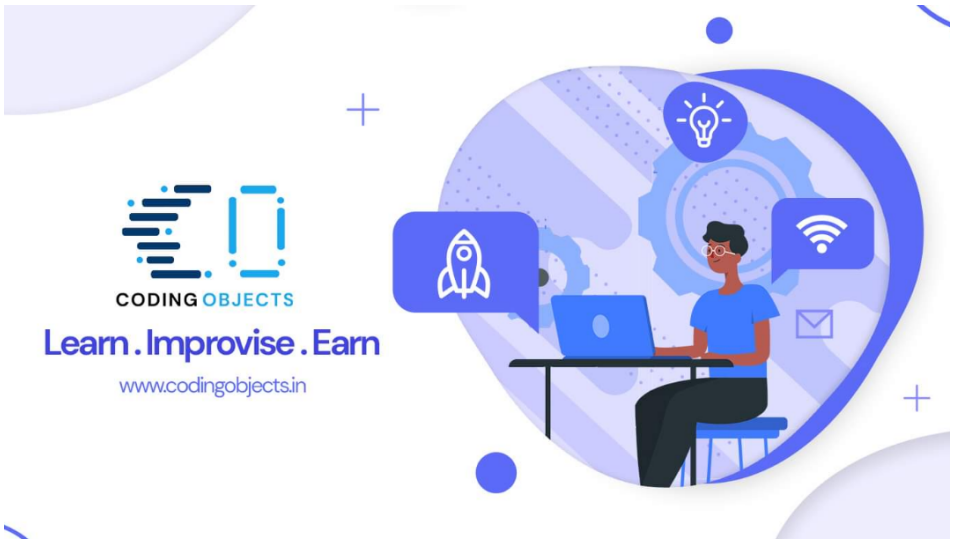
Contents

1 Heating Energy consumption and Correlates

In this Internship project, we conduct analysis on data collected from the many buildings on its construction date, yearly bill on Heating System and its geographical location.

1.1 Acknowledgement

I would like to thank softanbees for helping me learn and understand each basic aspect of data science and working with python.
I would like to extend my gratitude to Tamanna Ma'am for giving me the opportunity to work under her and for guiding me through the entire internship from the initial stage. I would also like to thank Tanvir sir for teaching the basic contents, and Surya sir for keeping us on track everyday on the course basics and assignments. Thank you for the support and encouragement throughout the course and internship.



1.2 Introduction

The project deals with Heating Energy Consumption and its correlates e.g. Climate change over the years, geographical position, building type etc. We collected the dataset from our mentor Tamanna Ma'am and converted it to our convenience. We included extensive data cleaning and curating because

our bottleneck here was not dimensionality of the dataset, rather its sheer size. So proper data cleaning and curating can give a good result later on.

1.3 Data Collection

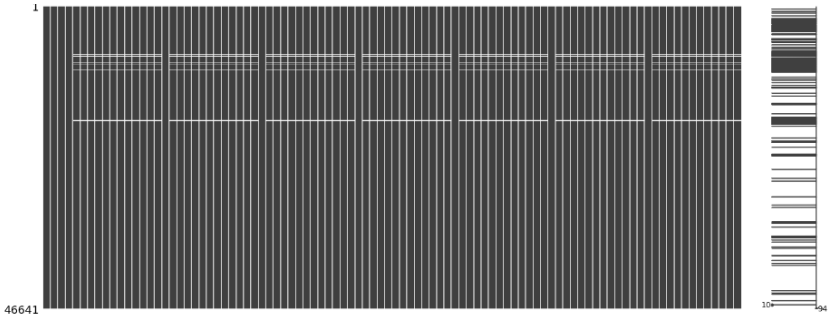
The data was collected by `Coding Objects` and provided to us by Tamanna Ma'am. The dataset has 45943 rows and 94 features. The data is in Excel format so we need to read the data appropriately.
One thing to notice that we have a lot of data so dropping some rows with repetitive empty cells is a good idea.

1.4 Data Cleaning

Looking at the data, we noticed some garbage values.

- vtj_{prt}
- gmlid

So we drop these features. Then we noticed many cells with `not_calculated` values. We replace this string with `NaN` so that we can deal with them using standard pandas functions. Then we performed a `missingno` visualization to see if the missing values are common in same rows.



And they indeed are! So we simply dropped the few 698 missing value rows.

1.5 EDA

Feature Overview to find our target feature.

- DHW = Domestic hot water for all monthes in 2020 (12+1 features)
- 2020, 2025, 2030, 2035, 2040, 2045, 2050 = 7*13 = 91
- longitutde, latitude, year = 3
- dropped unambiguous features = 2

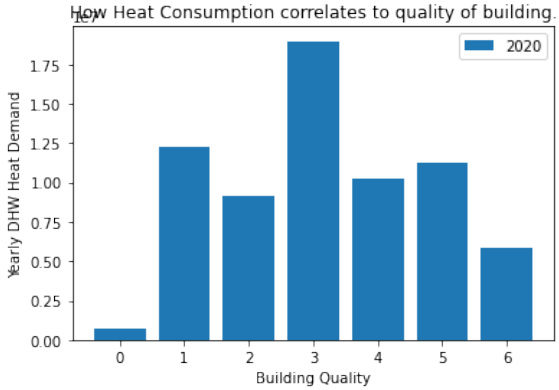
So total = 96

1.5.1 Finding Target Feature

1. Location So all data has different logitude, so each row correspondents to individual building.
   From google map, we saw that, buildings that are in same area, should have same three digits.
   23.450023, 61.23245
   23.423441, 61.24412
   They are in the same area.
   So we can assign area to each building and see which area has higher energy consumption over the years.

2. Change in Seasons and Global Warming We also notice years. We can use year 2020 and find out season where there were more energy consumptions. Then we can check if it remains consistent around all the years. A decrease in heat consumption indicated global warming. We can do a Chi-square test to see if the temperatures remain same in all seasons on these years.

3. Year of Construction The year a building was constructed has a nice correlation with its energy consumption.



1.6 Feature Engineering

In this stage, we fix our raw features before using them to build a model.

1.6.1 Defining Target Feature

So we have years from (1724 to 2020).
We will mark -

```
(1701-1800): Historical Monuments - 0
(1801-1900): Colonist Period Buildings - 1
(1901-1950): World War Era Buildings - 2
(1951-1990): Industrial Age Buildings - 3
(1991-2005): Modern Buildings - 4
(2006-2015): Recent Buildings - 5
(2016-2020): New Buildings - 6
```

With this, We build a new feature called `building_genre`. This is our target feature.

1.6.2 Type Conversion

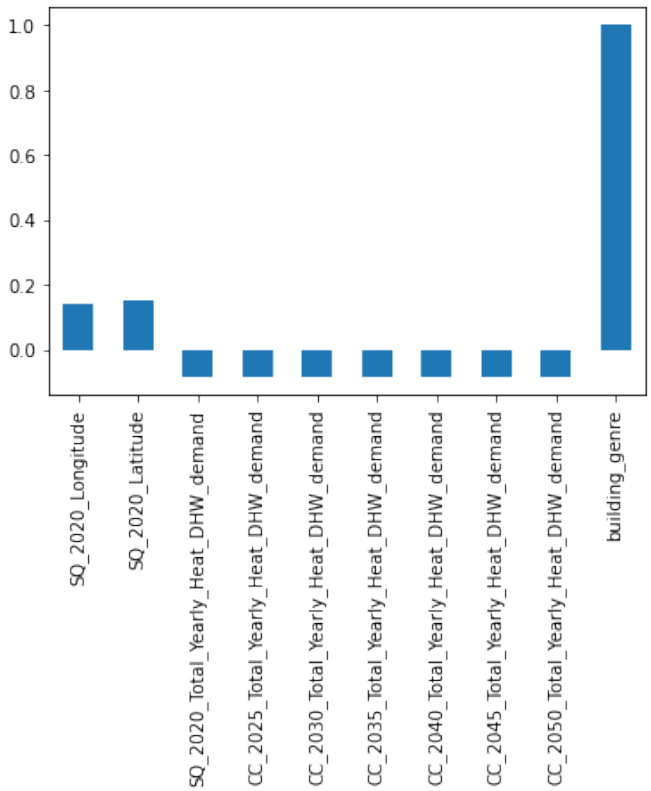
As we will be using the dataset later on model building, we wish to convert all datatypes into pandas datatypes. So we first see what types of features we have.

```
{dtype('int64'), dtype('float64'), dtype('O')}
```

We convert the `Object` datatype so something more usable - in our case, integer.

1.6.3 Correlation

We plot the correlation of target feature with other features.

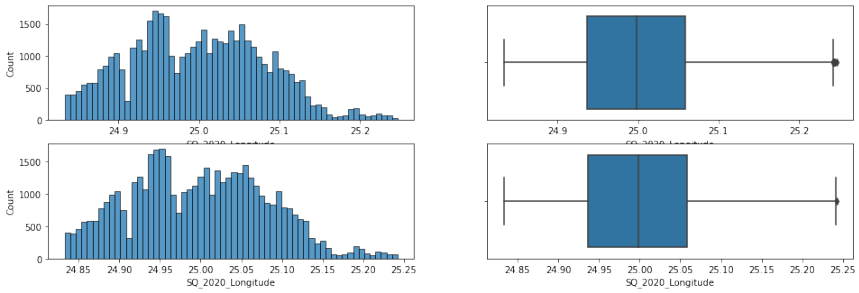


We notice that geographical location has comparatively better correlation with building genre.

1.6.4 Outliers

Outliers negatively impact a lot of models so we first trim and cap the outliers. We use extensive visualization to make sure the procedure is as

intended and we did not trim or cap any naturally skewed dataset.



1.6.5 Dealing with Highly Correlated Features

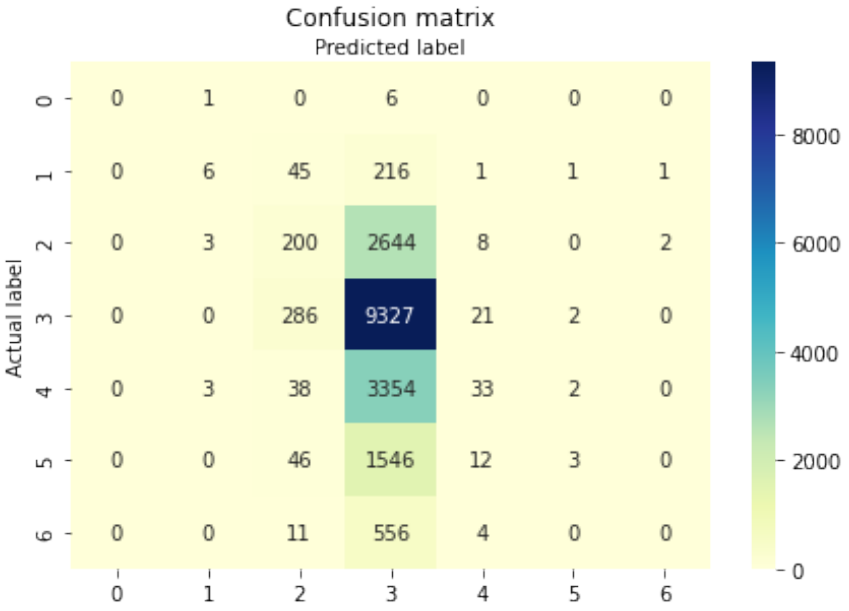
If a feature has high correlation with another feature, then just one of those features is enough to represent the data. We found these highly correlated features.

```
{'CC_2025_Total_Yearly_Heat_DHW_demand',
 'CC_2030_Total_Yearly_Heat_DHW_demand',
 'CC_2035_Total_Yearly_Heat_DHW_demand',
 'CC_2040_Total_Yearly_Heat_DHW_demand',
 'CC_2045_Total_Yearly_Heat_DHW_demand',
 'CC_2050_Total_Yearly_Heat_DHW_demand'}
```

And its intuitive - the sum of monthly power consumption does represent yearly power consumption. So we drop them.

1.7 Model Implementation

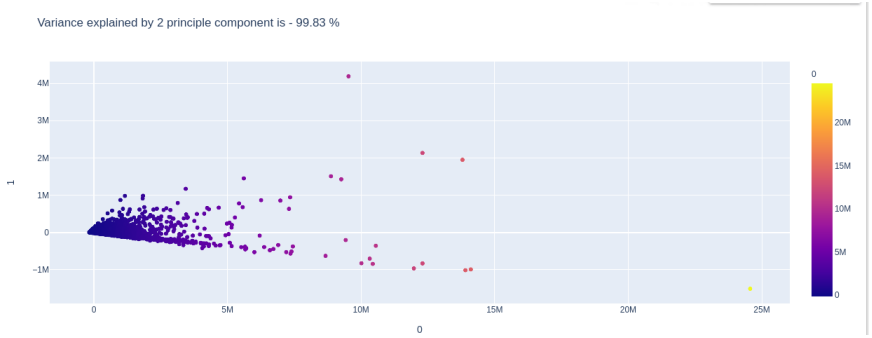
We split the dataset in test and train dataset. We first train the model, in our case, logistic regression model using the train dataset and validate it using test dataset. We do a 80:20 split of data as we have huge amount of data.



1.8 Conclusion

We notice that building genre has a great correlation with its geographical correlation and also, energy consumption. It drives us to the following suggestions.

- The old model buildings seem to have inefficient Heating systems that results in a lot of power-loss. So we suggest to update them.
- The building's geographic location does correlate with its state. So identifying regions with most frequent old buildings and renovating their heating system is a good approach.
- Highly correlated features made this data analysis pretty self-explanatory.



2 Future Outlooks

Due to limited computer resource, Extensive visualization and more complex models were not introduced. However, the two other possible target features leaves many interesting outlooks for future.

- Power Consumption and Geographical Location
- Change in Season on particular region and Climate Change