Scientific programming in Python

Sharon Cohen 208463463

1. Intro

The data set deals with the heating cots of buildings in a count where the primary source of heating power may be either oil or gas or electricity. The data set includes instances from past 10 years. Henceforth the subject matter used is Energy. There are 14 features in the data set. It includes features such as year, month when the measurement was taken. The typical sector of the building, size of the building, code of the district where the building resides, ground elevation at the location of the building, number of floors in the building, the year the building was built, minimum monthly temperature, maximum monthly temperature, mean monthly temperature, cost of heating for a given month and the target class which specifies the source of the heating power.

The data set includes 36000 instances. It includes both numeric and categorical variables. The size of the building, elevation of the building, number of floors in the building, etc comes under numeric type. On the other hand, variables like sector, month, district can be considered as categorical. The target class is a categorical variable.

1) Initial Data Analysis

The summary of the data set which consists of the mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, min value of the feature and maximum value of the feature is as shown in Figure 1:

	measure	year	month	size	district	elevation	floors	built	tmp_min	tmp_mean	tmp_max	cost
count	36000.000000	35471.000000	34583.000000	33374.000000	36000.000000	35424.000000	34512.000000	35431.000000	34645.000000	36000.000000	32496.000000	33349.00000
mean	17999.500000	2009.499704	6.500361	656.131030	4.486667	101.452942	7.910611	1966.782592	17.637988	24.058986	30.482232	4848.17067
std	10392.449182	2.871468	3.453408	561.181651	2.333090	100.214500	7.743878	26.525973	5.980053	5.488823	5.471901	6269.75903
min	0.000000	2005.000000	1.000000	122.000000	1.000000	-7.000000	1.000000	1920.000000	0.800000	12.800000	17.300000	7.75000
25%	8999.750000	2007.000000	3.000000	339.000000	2.000000	19.000000	3.000000	1944.000000	12.900000	19.500000	26.000000	1182.76000
50%	17999.500000	2009.000000	7.000000	497.000000	5.000000	55.000000	5.000000	1965.000000	18.700000	25.100000	31.700000	2625.02000
75%	26999.250000	2012.000000	10.000000	831.000000	6.000000	172.000000	9.000000	1992.000000	22.700000	29.200000	35.100000	5793.55000
max	35999.000000	2014.000000	12.000000	5947.000000	8.000000	337.000000	29.000000	2009.000000	77.700000	32.400000	43.900000	65493.79000

Figure 1: Summary of the data set

On examining the data set did consists of null values. The number of null values in each feature is as shown in Figure 2:

```
year          529
month        1417
sector       2618
size         2626
district        0
elevation     576
floors       1488
built         569
tmp_min      1355
tmp_mean        0
tmp_max      3504
cost         2651
source          0
dtype: int64
```

Figure 2: Number of null values of each feature

The next part is with respect to the handling of null values in the data set. The null values in the features year, month, sector and built year were remove from the data set since it does not make much sense to impute the missing year values. Next the null values in the other features were replaced with its median. This process is called as missing value imputation. Median was chosen as the choice since mean imputation is sensitive to outliers.

The next step in the process was to check for outliers. One of the popular methods to check for outliers is to use boxplots. It is vital to handle the outliers accordingly since many of the algorithms in data modelling step are sensitive to outliers. On examining the boxplots, there was some outliers present in the cost feature. The box plot of cost feature is as shown in Figure 3.

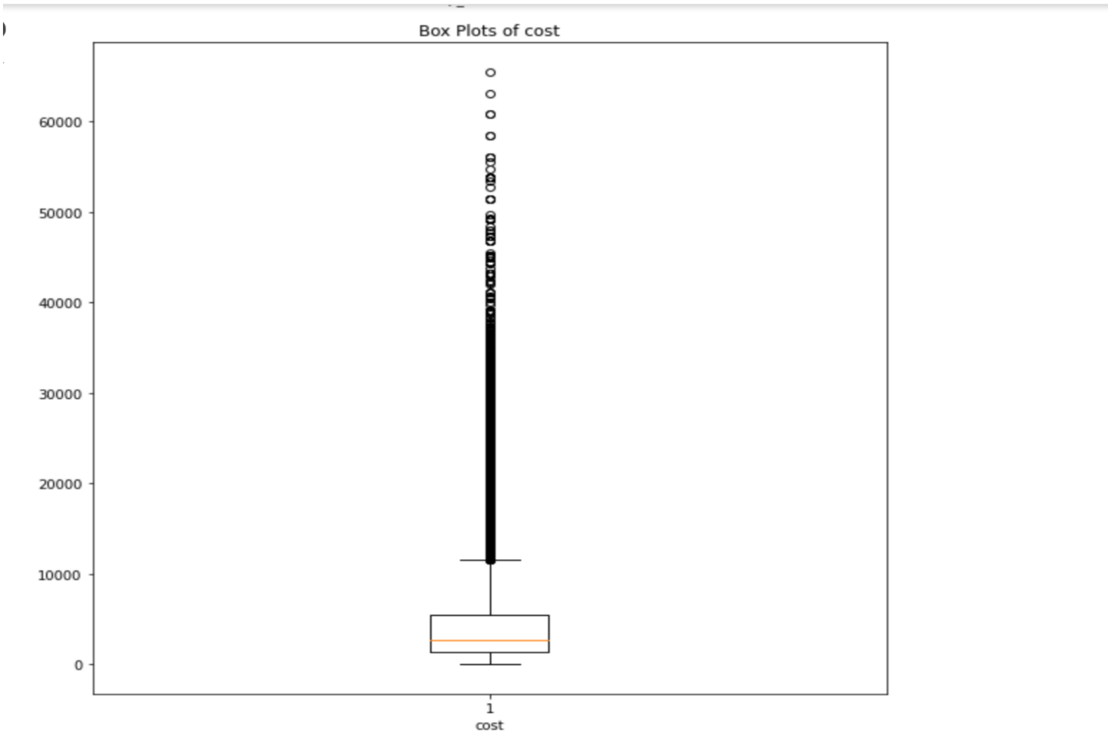

Figure 3: Box plot for cost feature

The black dots usually indicate the outliers. However, we have to make sure we won't be losing much of our data as well. So, we can choose a threshold as well. Here we remove the rows whose cost of heating is greater than 50000. Finally, the number of rows remaining after pre-processing is 31067.

2) Exploratory Data Analysis

Feature correlation is the process where you find the correlation between the existing features and the existing features to the target variable. Usually if there are multiple features which are highly correlated among themselves, the correlated features are removed except one. The features which have a high correlation to the target variable is chosen. The correlation matrix is as shown in the Figure 4:

	size	elevation	floors	tmp_min	tmp_max	cost
size	1.000000	0.077062	0.035680	0.009134	0.005102	0.393474
elevation	0.077062	1.000000	0.028844	0.003021	0.003300	0.050526
floors	0.035680	0.028844	1.000000	0.001409	0.001354	0.597677
tmp_min	0.009134	0.003021	0.001409	1.000000	0.845785	0.028894
tmp_max	0.005102	0.003300	0.001354	0.845785	1.000000	0.031496
cost	0.393474	0.050526	0.597677	0.028894	0.031496	1.000000

Figure 4: Corelation Matrix

We can see that there is a high correlation between maximum monthly temperature and minimum monthly temperature. So, we will go ahead and remove maximum monthly temperature from the feature's subset. Now coming to target classes, we will go ahead and check for the value counts of each target class. We will get to know whether we have a skewed data set or not. Bar graph in Figure 5 shows the same.

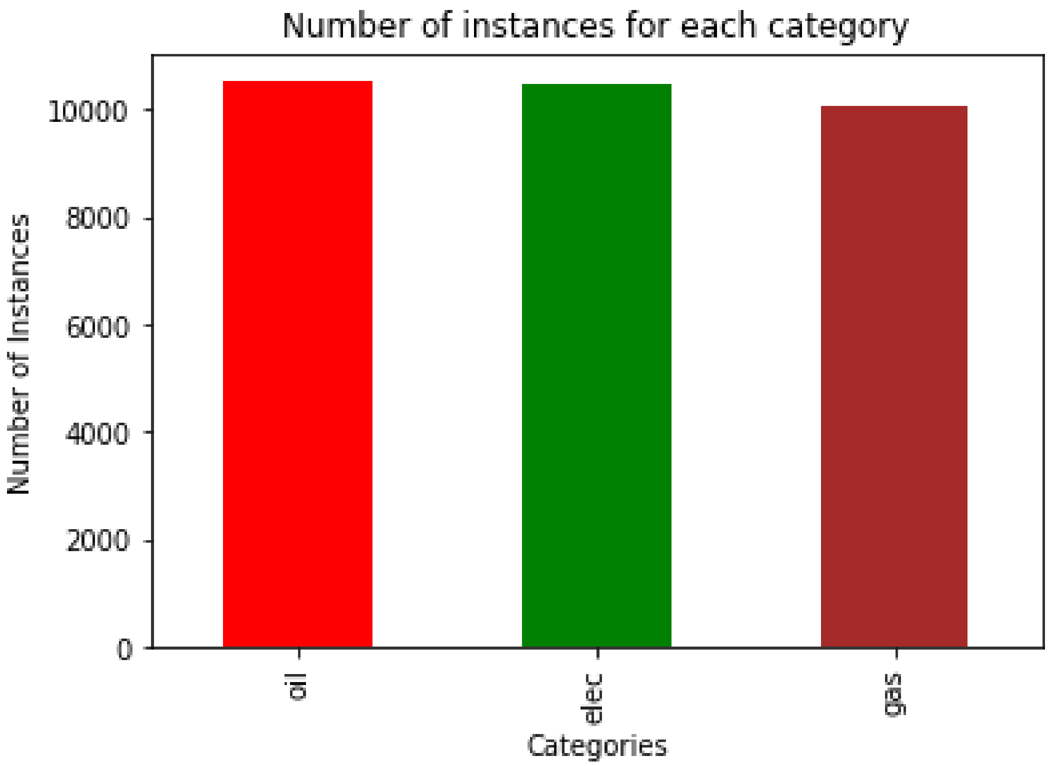

Figure 5: Number of instances of each class

Since the data set consists of time feature, we will go ahead and check for the mean cost of the heating per month across the years. We will use a time series graph for the same as shown in Figure 6.

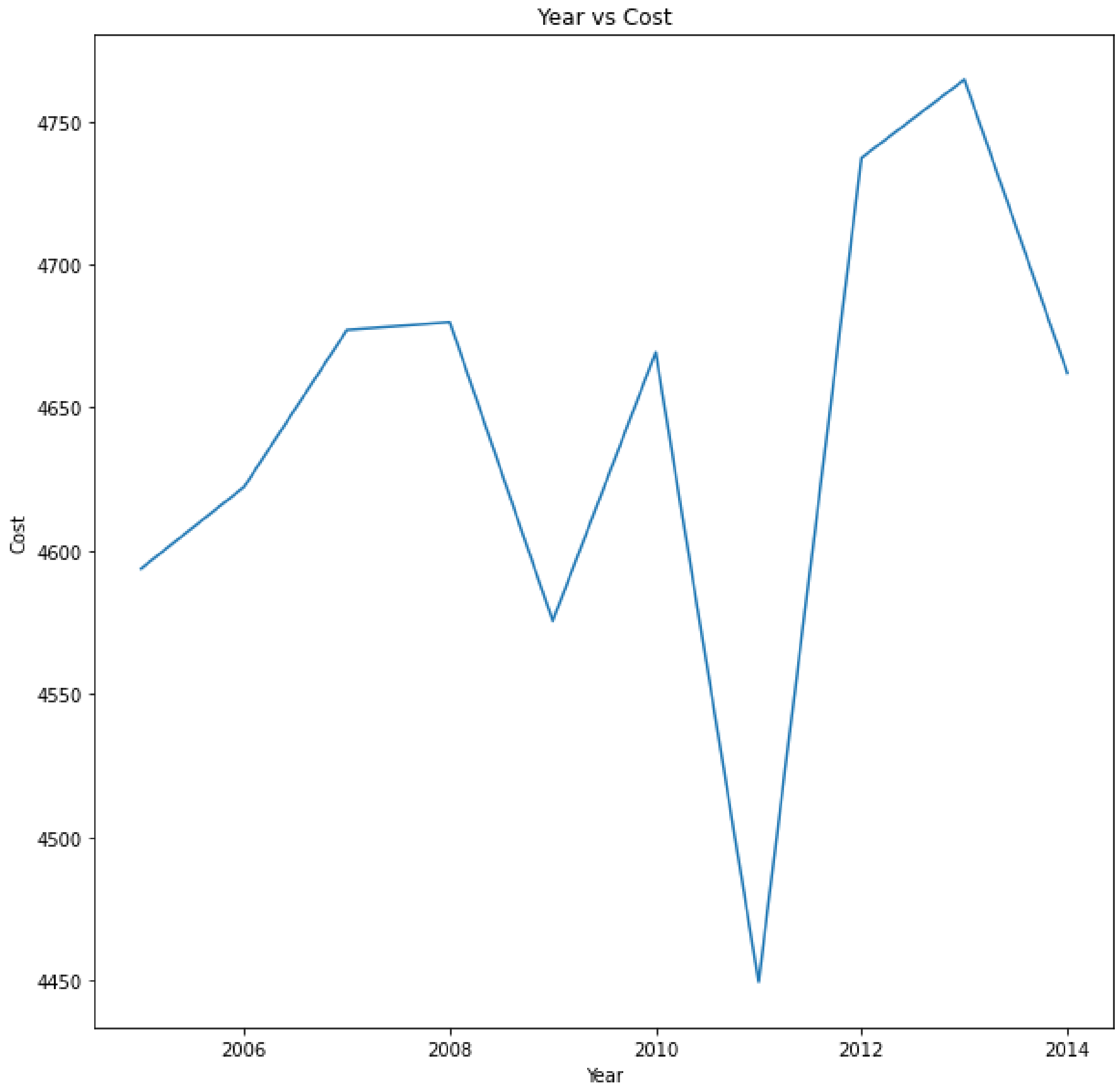

Figure 6: Mean cost of heating across years

The histograms of the continuous features are shown in Figure 7 and the pair plot is as shown in Figure 8.

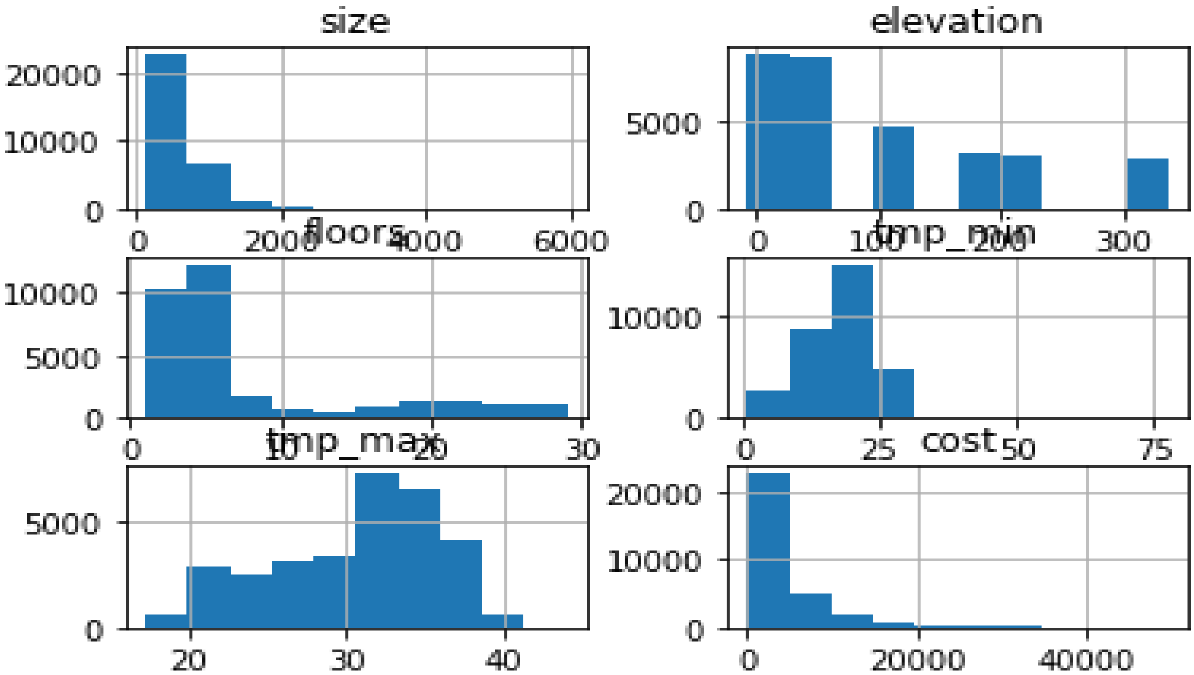

Figure 7: Histogram of Features

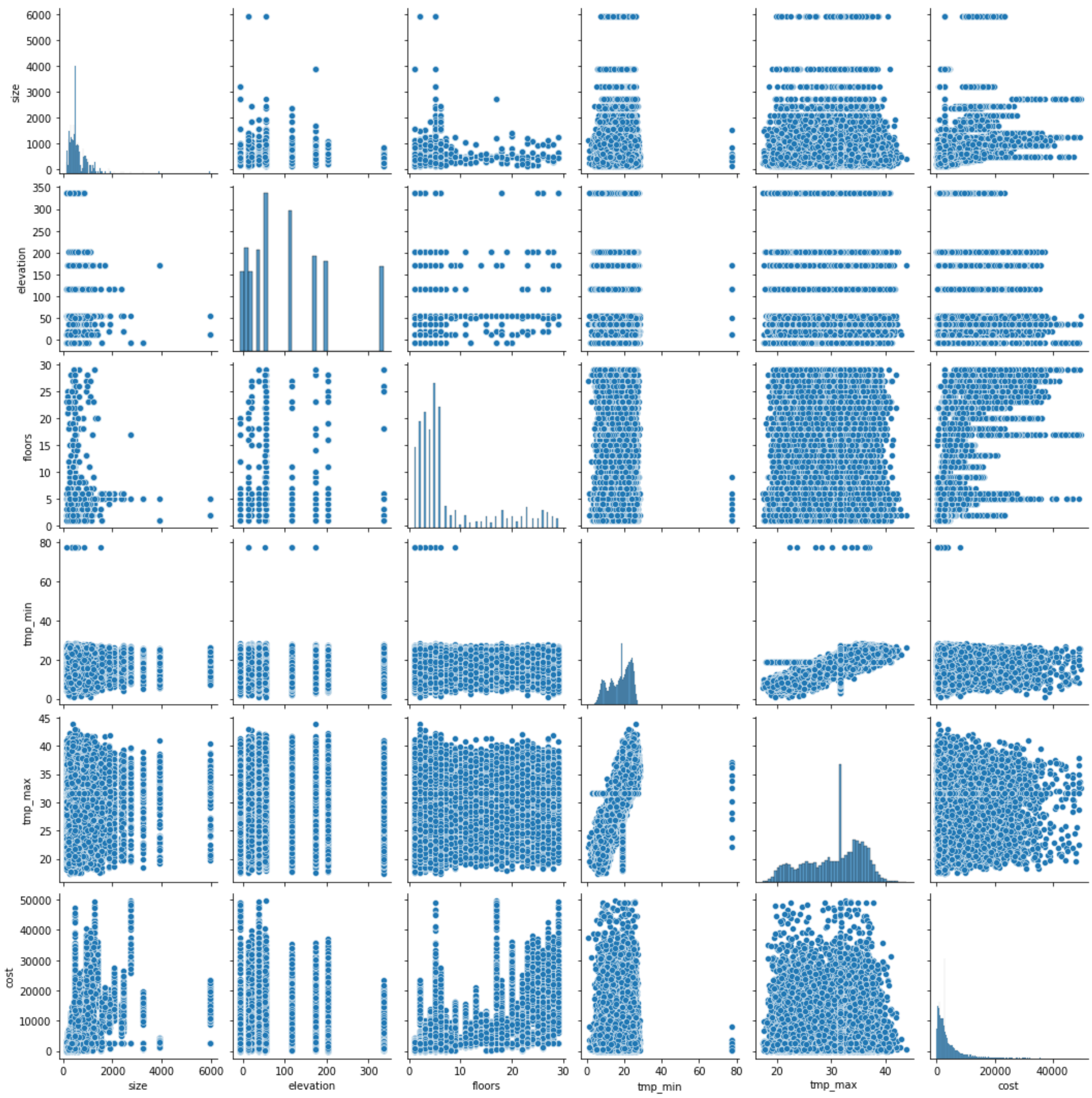

Figure 8: Pair Plot of features

Now the additional features include removal of features like year, month, built, it also included one hot encoding the sector feature. Later we standardise the features to a common scale. This is done so that model does not treat the higher values in a few features with high importance.

Additional Data Pre Processing

1. map categorical features to string

2. One-Hot Encoding- Encode categorical features as a one-hot numeric array.

```
measure       int64
year        float64
month       float64
sector       object
size        float64
district      int64
elevation   float64
floors      float64
built       float64
tmp_min     float64
tmp_mean    float64
tmp_max     float64
cost        float64
source       object
dtype: object
```

```
year             float64
month             object
sector            object
size             float64
district          object
elevation        float64
floors           float64
built            float64
tmp_min          float64
tmp_mean         float64
cost             float64
source            object
elec               uint8
gas                uint8
oil                uint8
district_1         uint8
district_2         uint8
district_3         uint8
district_4         uint8
district_5         uint8
district_6         uint8
district_7         uint8
district_8         uint8
commercial         uint8
education          uint8
factory            uint8
office             uint8
residential        uint8
```

```
Apr              uint8
Aug              uint8
Dec              uint8
Feb              uint8
Jan              uint8
Jul              uint8
Jun              uint8
Mar              uint8
May              uint8
Nov              uint8
Oct              uint8
Sep              uint8
dtype: object
```

check the correlation between each features using heatmap after One-Hot Encoding

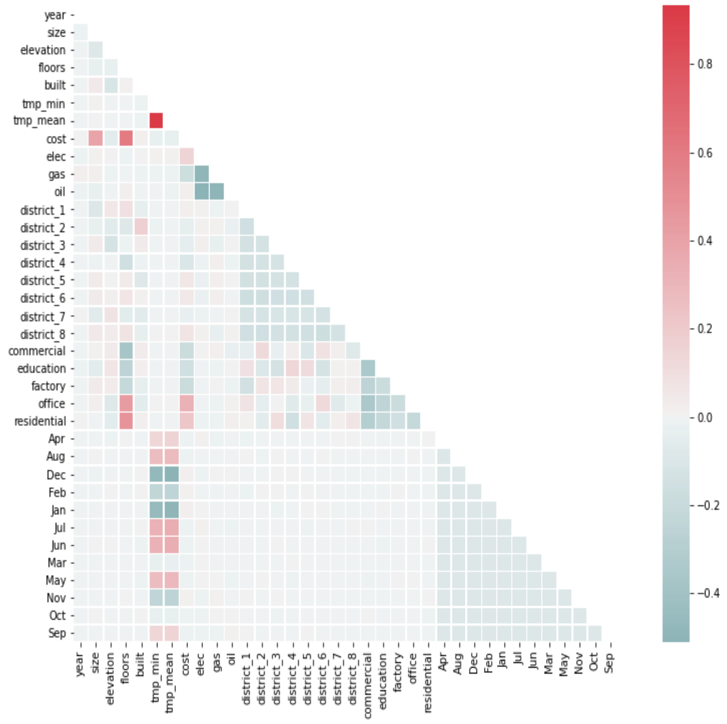

Figure 9 Heatmap

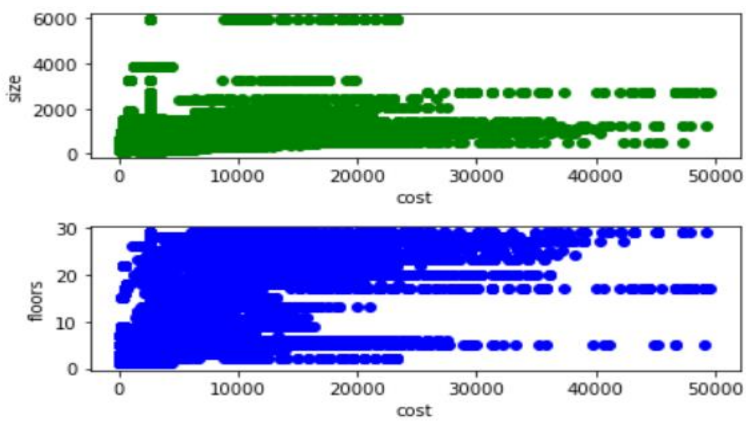

Figure 10 subplots Size +floors vs cost

According to Figure 10 the correlation between cost and size is vaguely seen. To continue researching we will use scatter plot 4 to see the 3rd feature we saw a high correlation with the cost feature

X-SIZE

Y-FLOORS

SIZE CYCLE – COST

COLOR- SECTOR(Before that we had to create order for the sector)

```
conv_dict={'commercial':1,'education':2,'factory':3,'office':4,'residential':5,'None':np.nan}
```

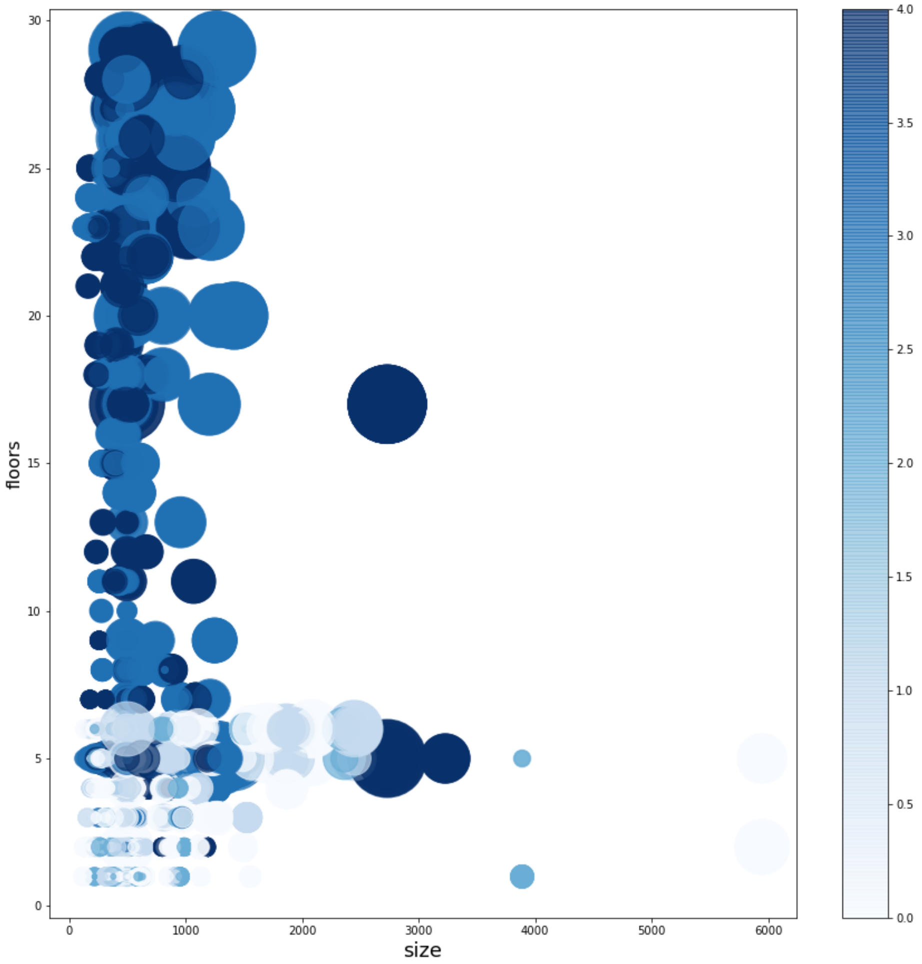

Figure 11 scatter plot 4D

3) Classification Model

4.1) Gaussian Naïve Bayes

How to find the 2 important features of GNB?

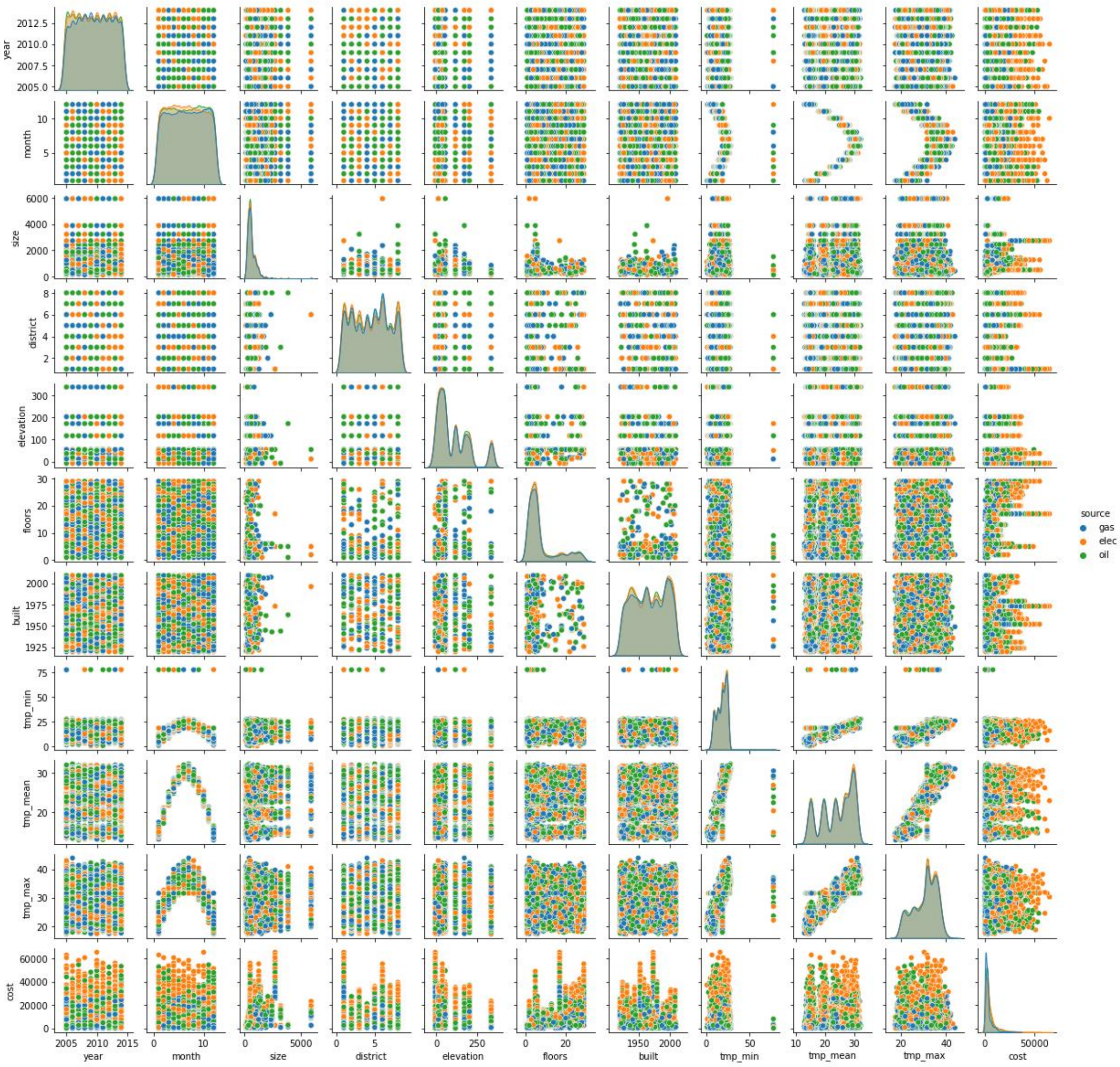

Figure 13 pairplot

- The diagonal plot which showcases the histogram. The histogram allows us to see the distribution of a single variable
- Upper triangle and lower triangle which shows us the scatter plot.
- The scatter plots show us the relationship between the features. These upper and lower triangles are the mirror image of each other.

The 2 important features of GNB classier were floors and cost (After trying 3 different possibilities)

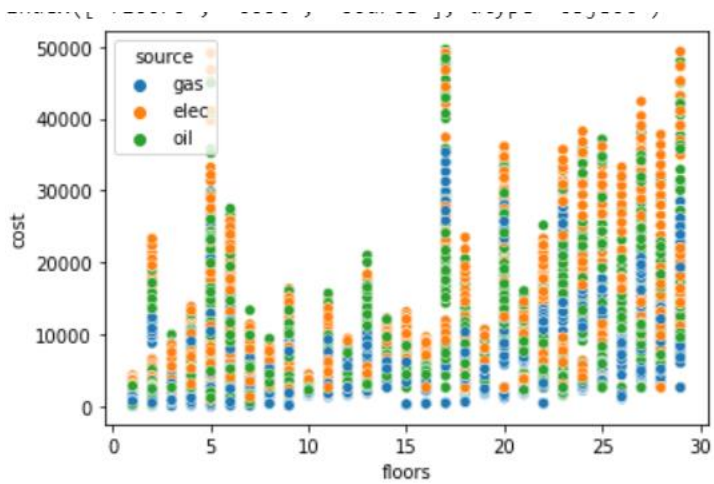

Figure 12 Scatter Plot1

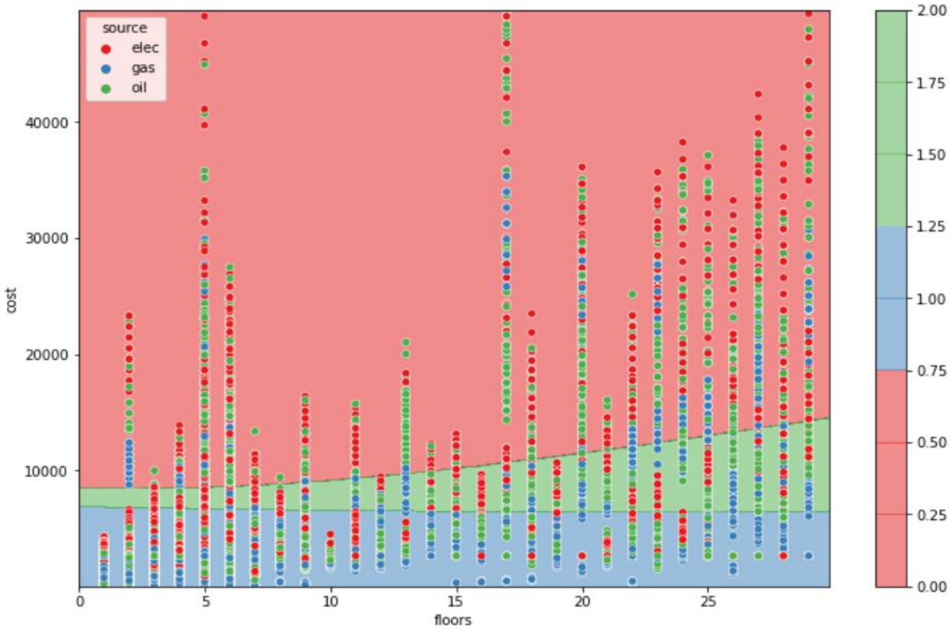

Figure 13 Scatter 2

4.2 Baseline Decision Tree Results

On fitting a decision tree to a baseline classification, the results were obtained as follows (shown in Figure 14)

	precision	recall	f1-score	support
elec	0.74	0.75	0.74	1783
gas	0.84	0.85	0.84	1820
oil	0.66	0.63	0.64	1851
accuracy			0.74	5454
macro avg	0.74	0.74	0.74	5454
weighted avg	0.74	0.74	0.74	5454

Figure 14: Classification Report for baseline Classification

4.3 Decision Tree based on Manipulated Data Set

On the manipulated data set, the classification report is as follows (as shown in Figure 15)

	precision	recall	f1-score	support
elec	0.70	0.70	0.70	2639
gas	0.78	0.78	0.78	2500
oil	0.63	0.63	0.63	2628
accuracy			0.70	7767
macro avg	0.70	0.70	0.70	7767
weighted avg	0.70	0.70	0.70	7767

Figure 15: Classification Report for Manipulated Data Set

The feature importance graph is as shown in Figure 12.

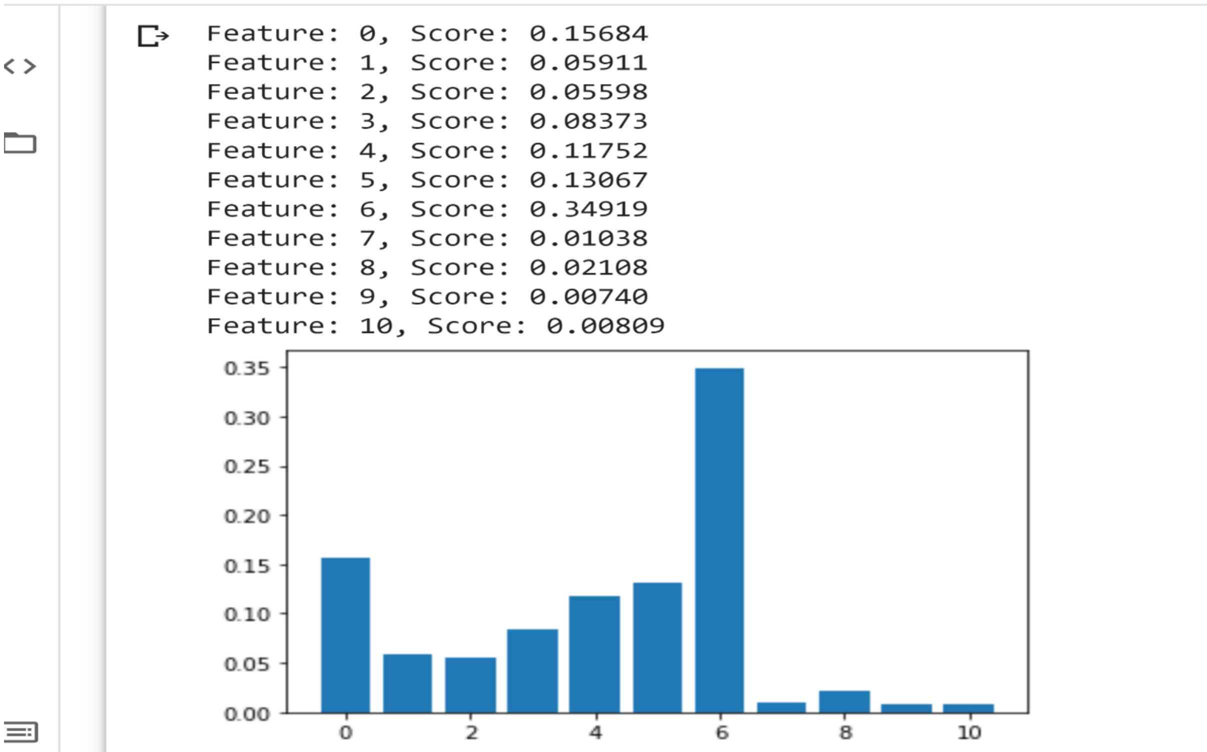

Figure 12: Feature Importance graph

The models performance could not be improved as compared to the baseline model. There might be various reasons accountable to this. Different imputation techniques might have to be tried out, may be the date and month formed important features to the target variables and should be further analysed. Hyperparameter techniques could be applied to the model. In these few ways, the performance of the model on the modified data set could be improved significantly.

Summary

Gaussian Naïve Bayes algorithm gives a very less accuracy. On the other hand, the baseline model and the model on manipulated data gives decent result with baseline model providing around 75 percent accuracy. However, this can be further improved tremendously. Accuracy can be considered as a suitable metric in evaluating the performance since it is a balanced data set. Some of the issues encountered include handling of missing values, categorical variables. Dealing with time series is difficult most of the times. Making meaning of time features becomes complicated sometimes. In our case we removed the time feature but manipulations could have been done to make some meaning out of it. Some of the insights form the analysis is The mean cost of heating is highest in 2013, the average size of the houses is 656 square feet. There is correlation between the monthly highest temperature and monthly minimum temperature. The average number of floors in the building is around 7. So this means mostly the medium-smaller apartments are the ones where measurement was taken