Lesson Assignment - EnergyLoads EDA

Author - Kaza Razat

Instructions

In this exercise you will use visualization and summary statistical methods to explore an energy efficiency data set.

This data set contains characteristics of over 750 buildings. The efficiency of the building is measured as either heating load or cooling load. Your goal is to find **three interesting relationships within the variables** which help you understand the energy efficiency of these buildings.

The exercise is deliberately open-ended. Whenever you approach a new data set some open-ended exploration is required. Expect this exploration to be an iterative process. You may need to try several ideas before you find truly interesting relationships.

Tip:

There are no categorical variables in this data set. There are two numeric variables, Orientation and Glazing Area Distribution, with only 4 and 2 discrete values. You can convert these integer values to factors and examine the effect on the distribution of Heating Load and Cooling Load.

In [1]:

```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
%matplotlib inline
```

In [3]:

```
# Cleaned up Dataset location
fileName = "https://library.startlearninglabs.uw.edu/DATASCI410/Datasets/EnergyEfficiencyData.csv"
```

In [4]:

```
energy_data = pd.read_csv(fileName)
energy_data.columns
```

Out[4]:

```
Index(['Relative Compactness', 'Surface Area', 'Wall Area', 'Roof Ar
ea',
       'Overall Height', 'Orientation', 'Glazing Area',
       'Glazing Area Distribution', 'Heating Load', 'Cooling Load'],
      dtype='object')
```

Summary statistics

Using pandas features like head, dtypes, describe and some initial histograms to get a better feel for the data.

In [5]:

```
energy_data.dtypes
```

Out[5]:

```
Relative Compactness         float64
Surface Area                 float64
Wall Area                    float64
Roof Area                    float64
Overall Height               float64
Orientation                    int64
Glazing Area                 float64
Glazing Area Distribution      int64
Heating Load                 float64
Cooling Load                 float64
dtype: object
```

In [7]:

```
energy_data.describe()
```

Out[7]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000

In [9]:

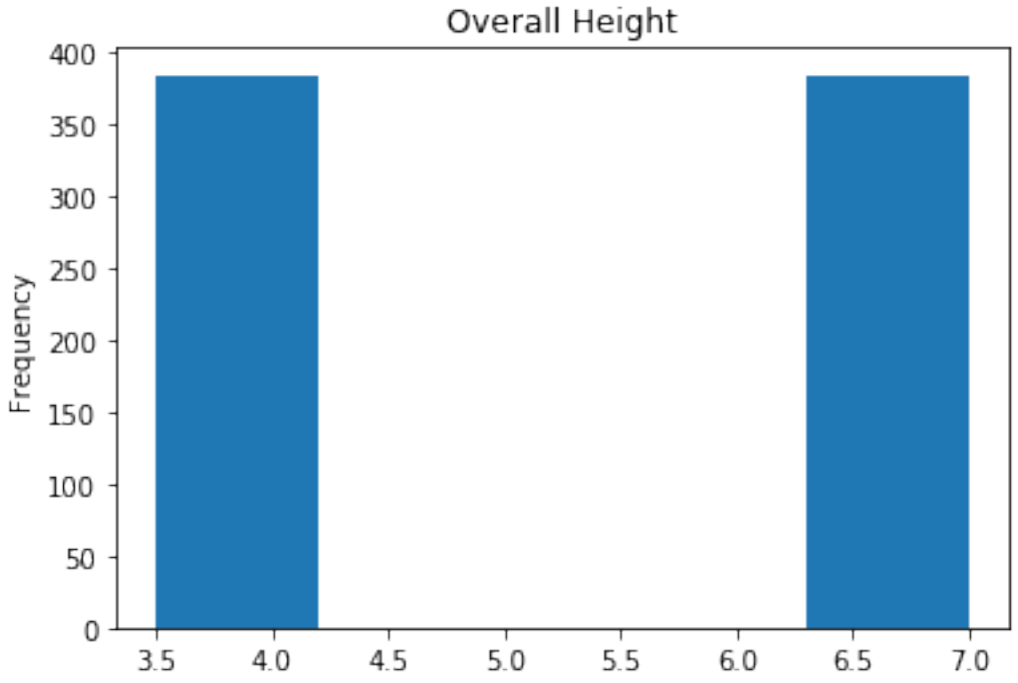
```
energy_data.tail(10)
```

Out[9]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heating Load
758	0.66	759.5	318.5	220.5	3.5	4	0.4	5	14.92
759	0.66	759.5	318.5	220.5	3.5	5	0.4	5	15.16
760	0.64	784.0	343.0	220.5	3.5	2	0.4	5	17.69
761	0.64	784.0	343.0	220.5	3.5	3	0.4	5	18.19
762	0.64	784.0	343.0	220.5	3.5	4	0.4	5	18.16
763	0.64	784.0	343.0	220.5	3.5	5	0.4	5	17.88
764	0.62	808.5	367.5	220.5	3.5	2	0.4	5	16.54
765	0.62	808.5	367.5	220.5	3.5	3	0.4	5	16.44
766	0.62	808.5	367.5	220.5	3.5	4	0.4	5	16.48
767	0.62	808.5	367.5	220.5	3.5	5	0.4	5	16.64

In [6]:

```
energy_data.loc[:,'Overall Height'].plot.hist(bins = 5, title="Overall Height");
```



Observation

The Overall Height values are evenly distributed and given a rational that building height could be a factor in its energy load I intend to create categorical data using these values to explore the possible relationship.

In [11]:

```
# Create a new column of categories based on the values of the Overall Height da
ta
height_categories = []
for i in energy_data.loc[:,"Overall Height"]:
    if i == 7.0:
        height_categories.append("taller")
    elif i == 3.5:
        height_categories.append("shorter")
    else:
        height_categories.append("unknown")

energy_data["Building Height"] = height_categories
```

Overall Height to Cooling Load Relationship

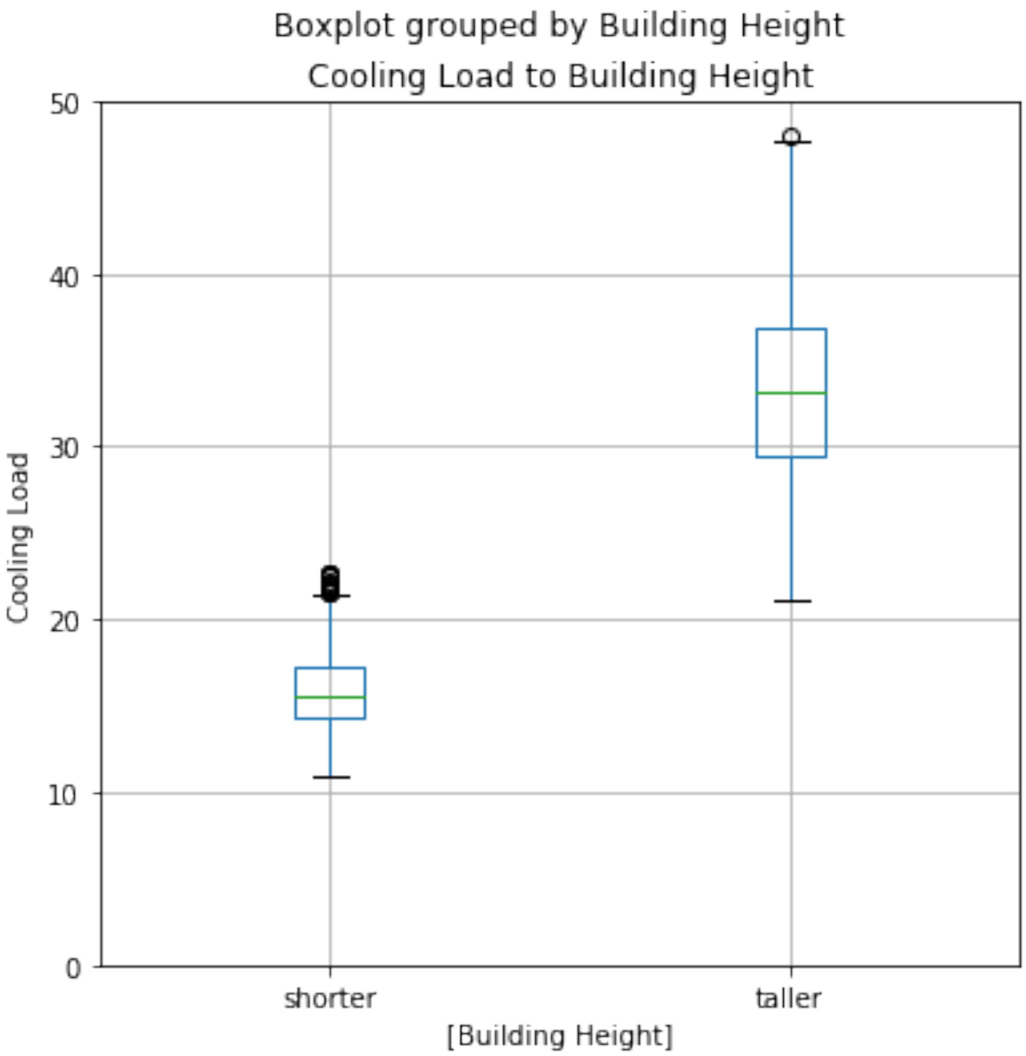
From this observation using a box plot it initially appears as if the taller buildings have the higher cooling loads

In [12]:

```
fig = plt.figure(figsize=(6, 6)) # Define plot area
ax = fig.gca() # Define axis
energy_data.loc[:,['Cooling Load','Building Height']].boxplot(by = 'Building Hei
ght', ax = ax)
ax.set_title('Cooling Load to Building Height') # Give the plot a main title
ax.set_ylabel('Cooling Load')# Set text for y axis
ax.set_ylim(0.0, 50.0) # Set the limits of the y axis
```

Out[12]:

```
(0.0, 50.0)
```



Overall Height to Heating Load Relationship

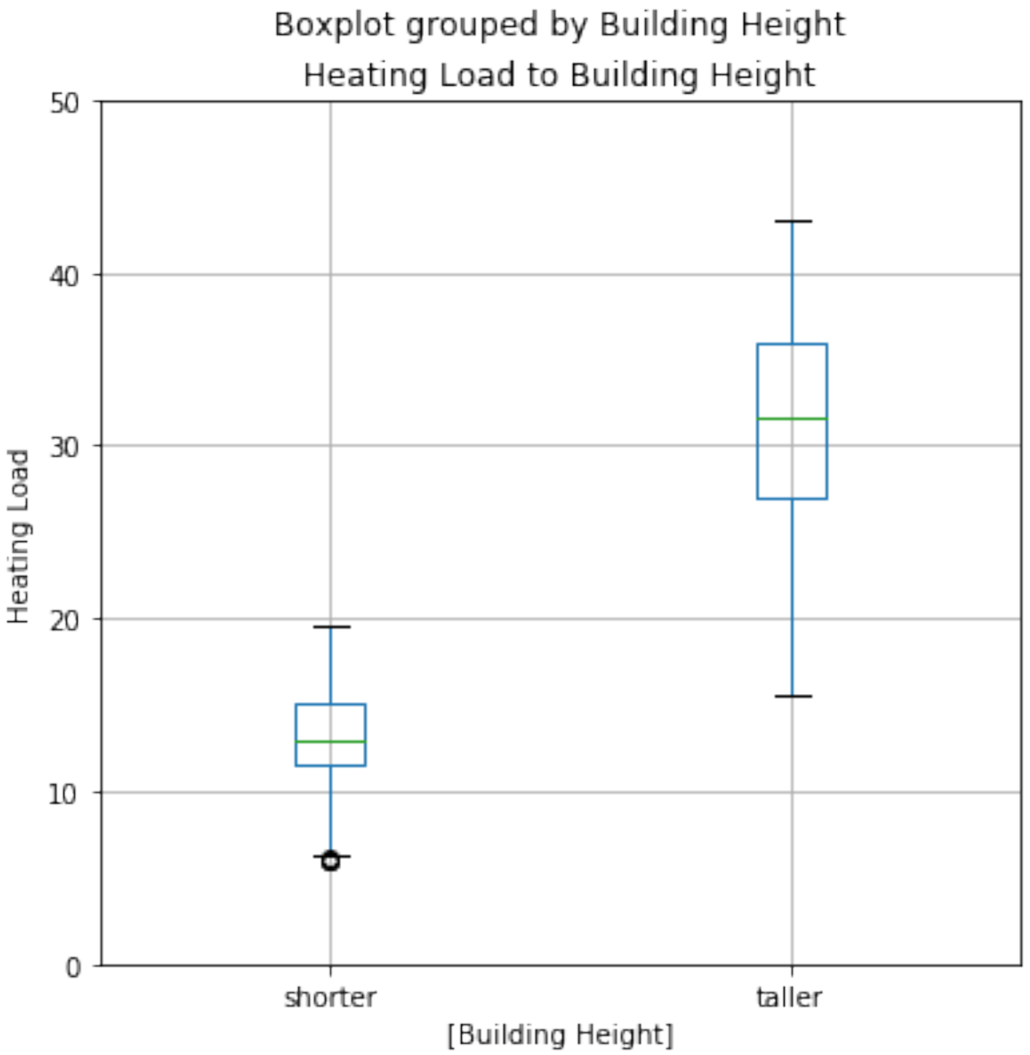
From this observation using a box plot it also appears that the taller buildings have the higher heating loads

In [13]:

```
fig = plt.figure(figsize=(6, 6)) # Define plot area
ax = fig.gca() # Define axis
energy_data.loc[:,['Heating Load','Building Height']].boxplot(by = 'Building Hei
ght', ax = ax)
ax.set_title('Heating Load to Building Height') # Give the plot a main title
ax.set_ylabel('Heating Load')# Set text for y axis
ax.set_ylim(0.0, 50.0) # Set the limits of the y axis
```

Out[13]:

```
(0.0, 50.0)
```



Cooling and Heating Loads Have Similar Distribution

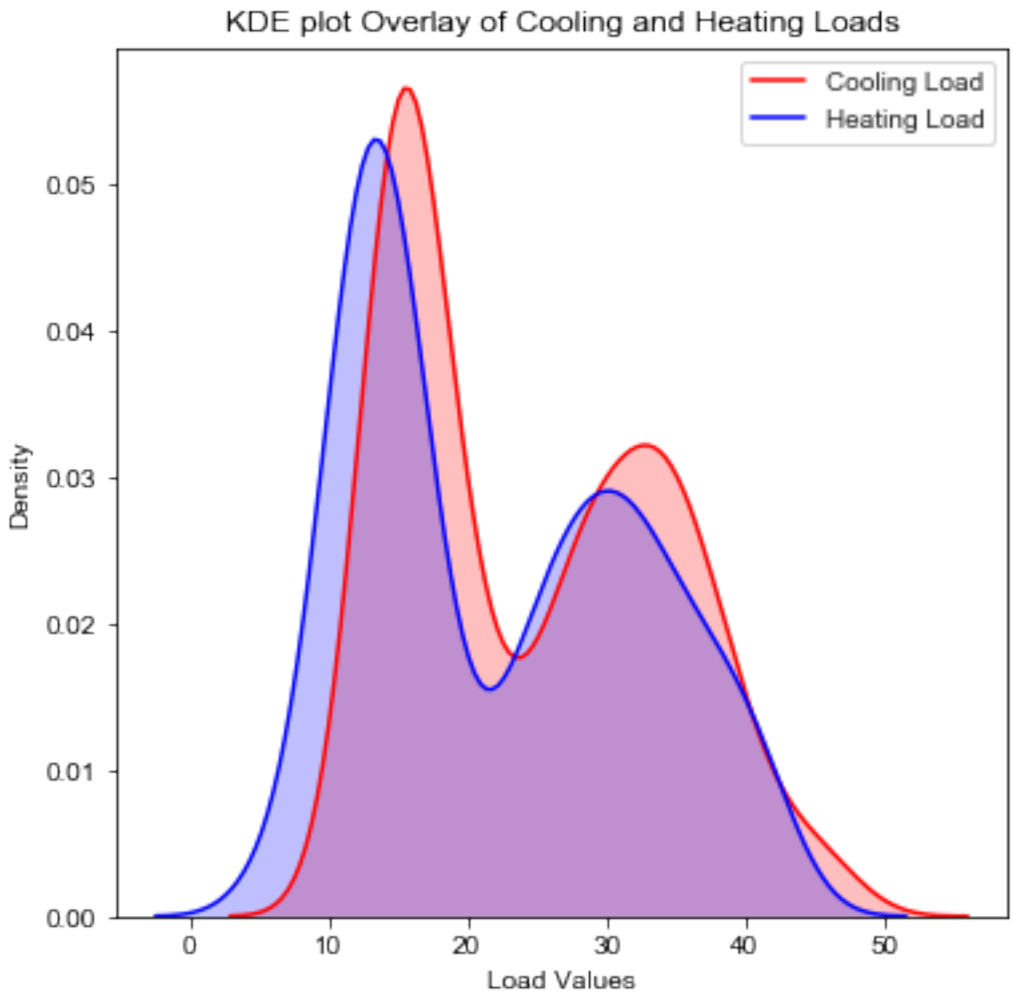
Using an overlay of kernal density plots we can observe that the distributions of cooling and heating loads are very similar across all the buildings.

In [14]:

```
fig = plt.figure(figsize=(6,6)) # Define plot area
ax = fig.gca() # Define axis
sns.set_style("whitegrid")
ax = sns.kdeplot(energy_data.loc[:, 'Cooling Load'], shade=True, color="r")
ax = sns.kdeplot(energy_data.loc[:, 'Heating Load'], shade=True, color="b")
ax.set_title('KDE plot Overlay of Cooling and Heating Loads') # Give the plot a
main title
ax.set_xlabel('Load Values') # Set text for the x axis
ax.set_ylabel('Density')# Set text for y axis
```

Out[14]:

```
Text(0, 0.5, 'Density')
```



EDA Conclusions

1.) There is a relationship between the Overall Height of a building and the Cooling Load.
2.) There is a relationship between the Overall Height of a building and the Heating Load.
3.) The distributions of Cooling Load and Heating Load are very similar indicating some correlation.