Kasra Samadi 993623030

HW2 DataMining

Tasks :

1- Read data set-2.csv

2- Obtaining information (#Rows, #Columns, Types of Columns, describe of dataFrame, Missing values, Inconsistent datas)

3- Correlation Matrix

4- Visualize the linear relationship between two numerical features to illustrate the concept of correlation

Task 1 And 2

Import libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Reading data set-2.csv

In [2]:
```
df = pd.read_csv("data set-2.csv")
```

Missing datas are marked with NaN

In [3]:
```
df
```

Out[3]:

	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
0	6	74	132	4	23.8	4
1	10	43	263	4	56.7	4
2	3	81	145	2	28.0	6
3	9	50	196	4	45.1	3
4	2	80	131	5	20.8	2
...	...	...	...	...	...	...
1213	7	56	264	5	58.2	5
1214	5	78	129	1	22.5	1
1215	5	77	138	3	26.8	1
1216	3	89	156	2	34.0	2
1217	7	59	273	1	61.4	5

1218 rows × 6 columns

This dataFrame has 1218 rows and 6 columns

In [4]:
```
df.shape
```

Out[4]:
```
(1218, 6)
```

Find columns names and their types

In [5]:
```
df.dtypes
```

Out[5]:
```
Insulation         int64
Temperature        int64
Heating_Oil        int64
Num_Occupants      int64
Avg_Age          float64
Home_Size          int64
dtype: object
```

Show The describe of this dataFrame

In [6]:
```
df.describe()
```

Out[6]:

	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
count	1218.000000	1218.000000	1218.000000	1218.000000	1218.000000	1218.000000
mean	6.214286	65.078818	197.394089	3.113300	42.706404	4.649425
std	2.768094	16.932425	56.248267	1.690605	15.051137	2.321226
min	2.000000	38.000000	114.000000	1.000000	15.100000	1.000000
25%	4.000000	49.000000	148.250000	2.000000	29.700000	3.000000
50%	6.000000	60.000000	185.000000	3.000000	42.900000	5.000000
75%	9.000000	81.000000	253.000000	4.000000	55.600000	7.000000
max	10.000000	90.000000	301.000000	10.000000	72.200000	8.000000

Find the number of missing values for each columns

There are no missing values AND Inconsistent datas in this dataFrame

In [7]:
```
df.isna().sum()
```

Out[7]:
```
Insulation       0
Temperature      0
Heating_Oil      0
Num_Occupants    0
Avg_Age          0
Home_Size        0
dtype: int64
```

In [8]:
```
print(df["Insulation"].value_counts())
print(df["Temperature"].value_counts())
print(df["Heating_Oil"].value_counts())
print(df["Num_Occupants"].value_counts())
print(df["Num_Occupants"].value_counts())
print(df["Home_Size"].value_counts())
```

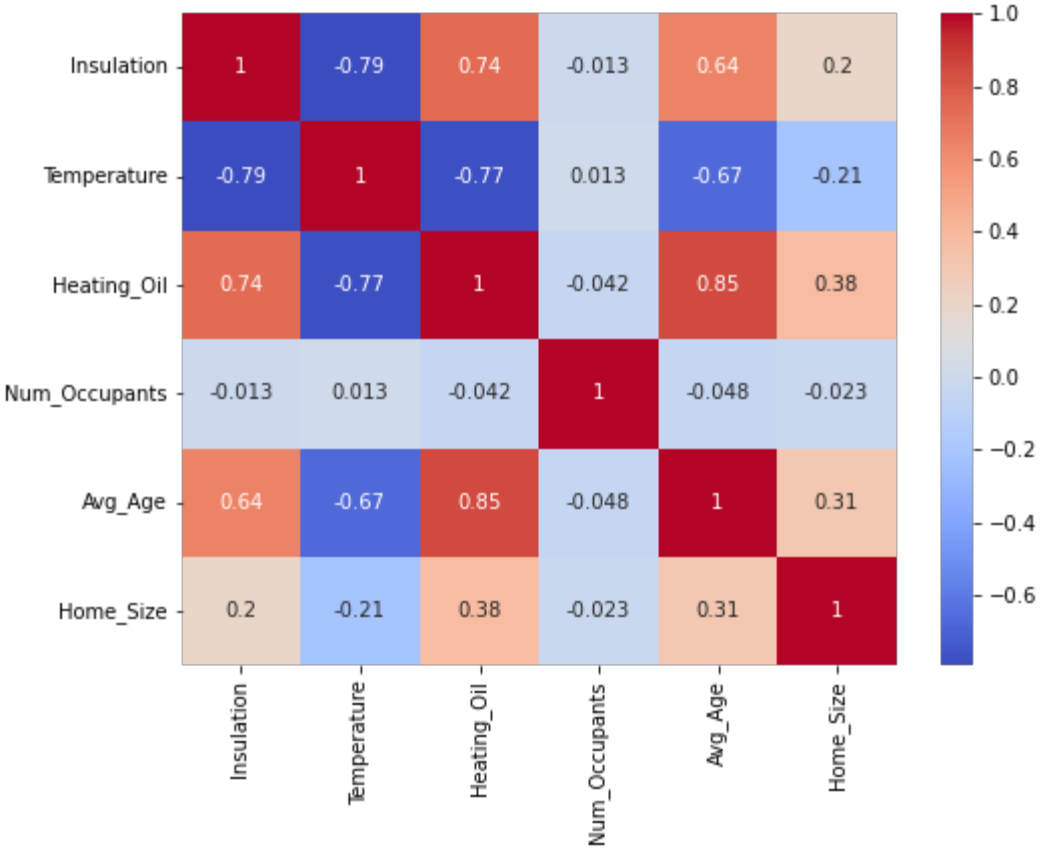
```
10    227
3     179
4     164
5     144
9     139
8     121
2      95
7      83
6      66
Name: Insulation, dtype: int64
76    57
89    49
77    43
42    43
55    42
48    41
75    41
88    40
40    38
83    37
56    37
58    36
43    35
81    33
57    33
73    33
86    32
80    32
59    31
84    30
74    30
78    29
82    27
49    27
87    26
47    25
39    25
54    24
45    24
90    23
52    22
50    22
53    20
51    18
41    17
46    16
79    15
72    15
```

```
44    14
85    14
60    12
38    10
Name: Temperature, dtype: int64
131    21
142    18
183    18
156    16
288    16
       ..
216     1
218     1
268     1
222     1
269     1
Name: Heating_Oil, Length: 178, dtype: int64
1     252
3     249
2     233
4     226
5     209
6      14
8      12
10      9
7       9
9       5
Name: Num_Occupants, dtype: int64
1     252
3     249
2     233
4     226
5     209
6      14
8      12
10      9
7       9
9       5
Name: Num_Occupants, dtype: int64
8    174
7    168
4    154
6    151
1    149
5    147
3    139
2    136
Name: Home_Size, dtype: int64
```

Task 3 : Correlation Matrix

In [9]:
```
plt.figure(figsize=(8, 6))
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm")
```

Out[9]: <AxesSubplot:>

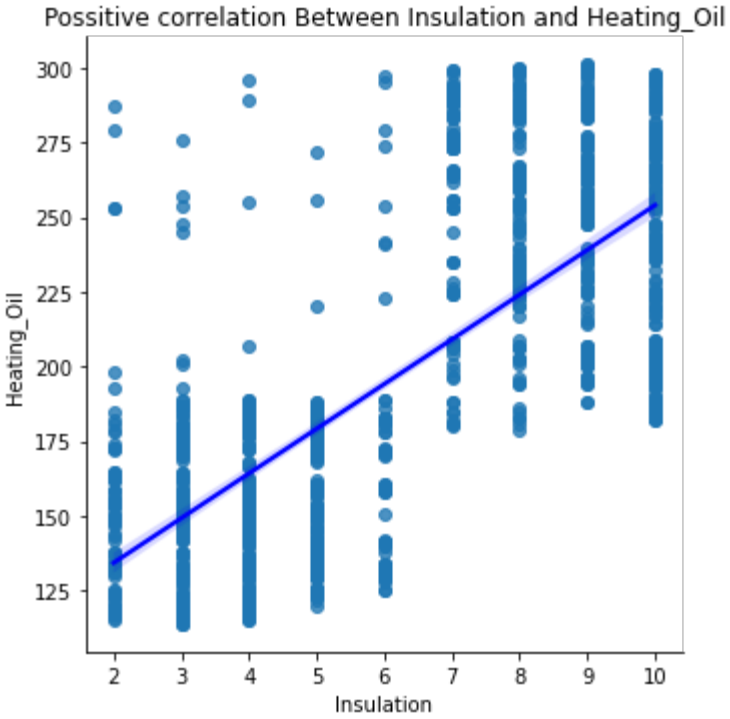


Task 4 : Visualize the linear relationship

Possitive correlation

In [10]:
```
sns.lmplot(x='Insulation', y='Heating_Oil',data=df, line_kws={'color': 'blue'}
```

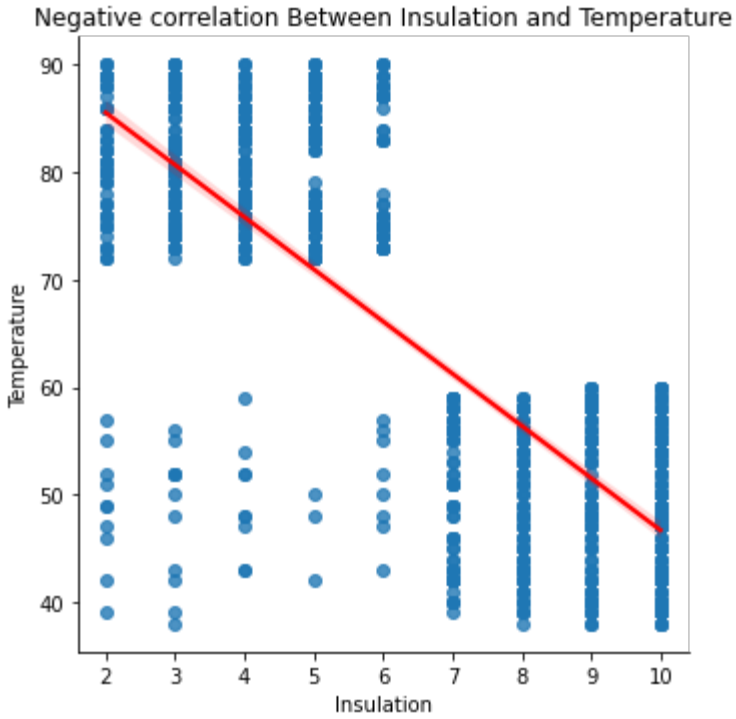
Out[10]:
```
<seaborn.axisgrid.FacetGrid at 0x127f6a14310>
```



Negative correlation

In [11]:
```
sns.lmplot(x='Insulation', y='Temperature',data=df, line_kws={'color': 'red'})
```

Out[11]:
```
<seaborn.axisgrid.FacetGrid at 0x127f6b3a280>
```



No correlation

In [12]:
```
sns.lmplot(x='Num_Occupants', y='Temperature',data=df, line_kws={'color': 'gre
```

Out[12]:
```
<seaborn.axisgrid.FacetGrid at 0x127f6bbbd30>
```