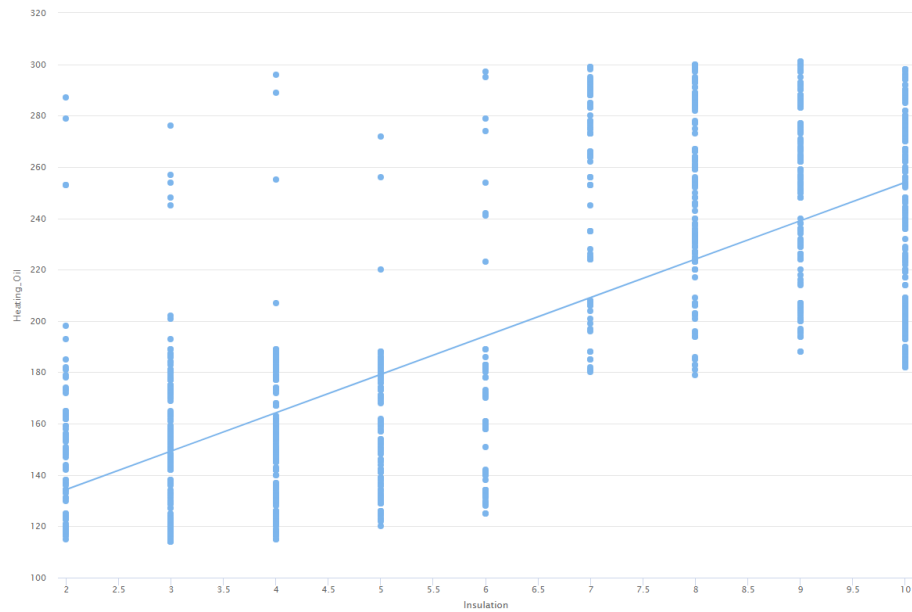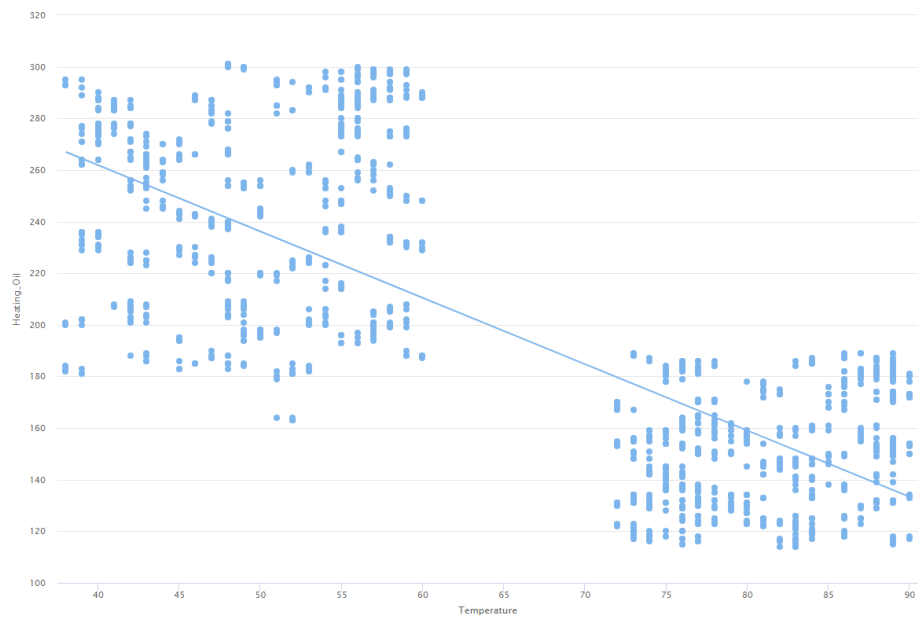# Homework 7

Pre-processing:
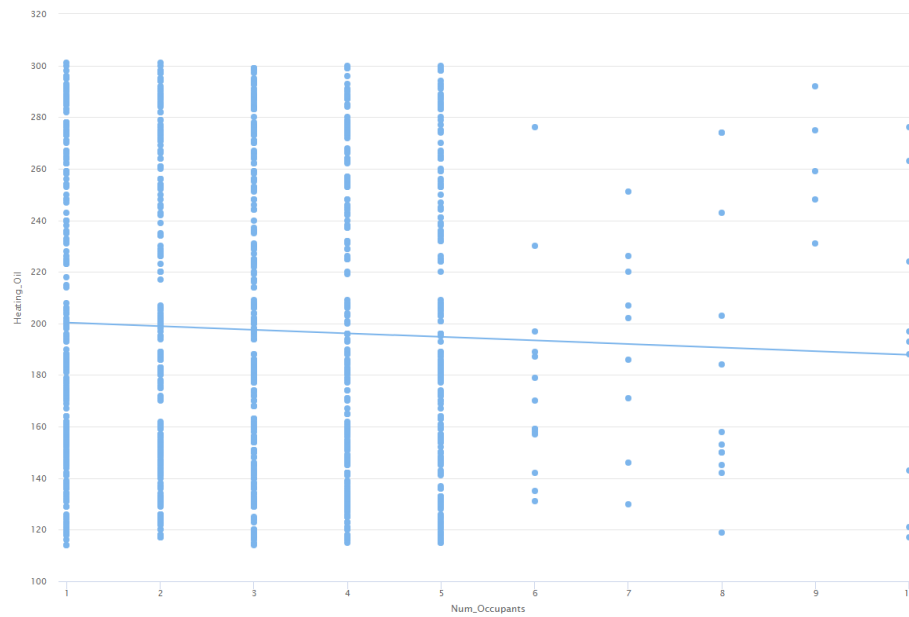
- Linearity check between independent and dependent variable.

  - Insulation – Heating_oil: Strong Positive
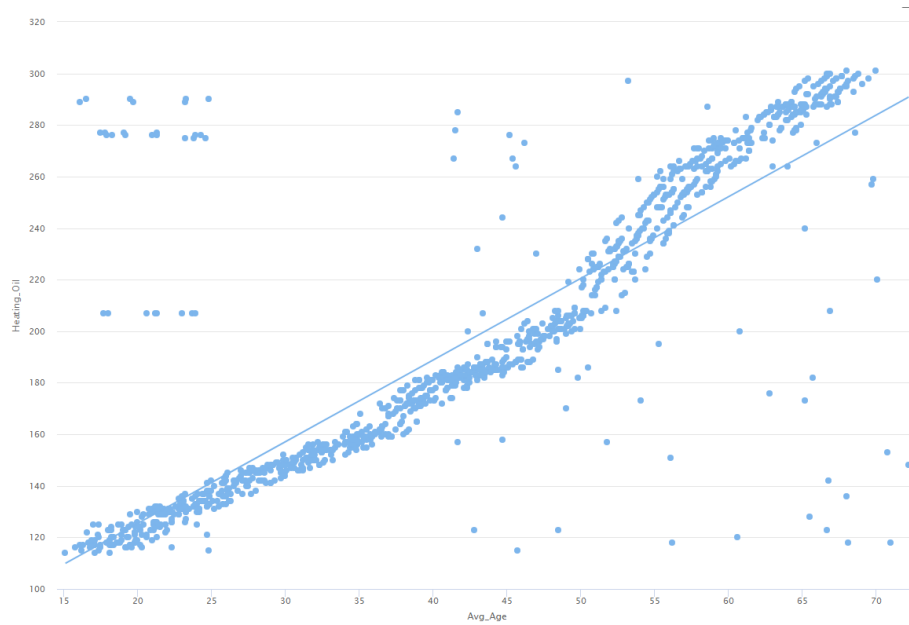

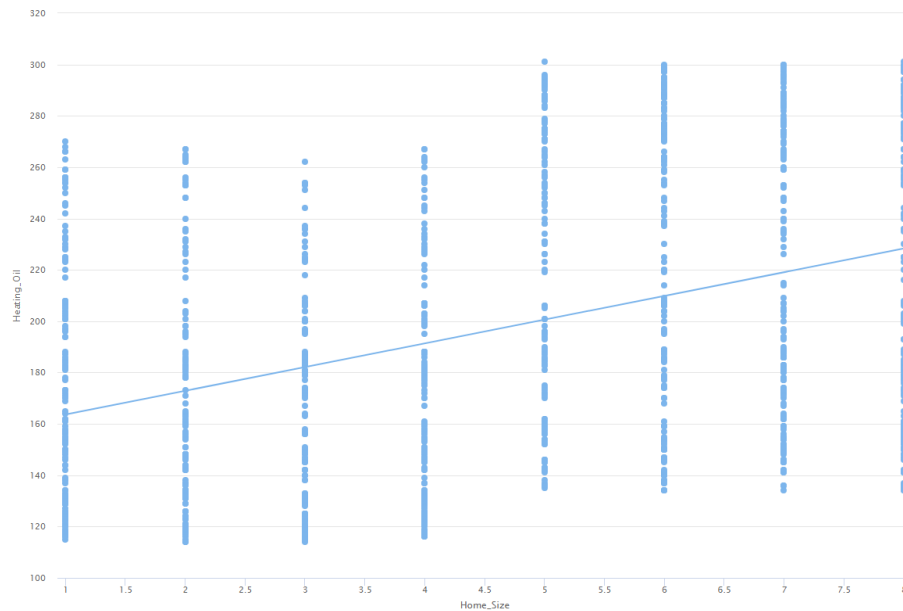 - Temperature – Heating_oil: Strong Negative

- Num_Occupants – Heating_oil: No correlation

- Avg_Age – Heating_oil: Strong Positive

- Home_Size – Heating_oil: Weak Positive

- Check of possible co-linear features.

Attributes	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Insulation	1	-0.794	-0.013	0.643	0.201
Temperature	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.643	-0.673	-0.048	1	0.307
Home_Size	0.201	-0.214	-0.023	0.307	1

จาก correlation matrix จะเห็นได้ว่า (Insulation, Temperature), (Insulation, Avg_Age) และ (Temperature, Avg_Age) มีค่า correlation coefficient ที่เข้าใกล้ 1, -1 ทำให้สรุปได้ว่า feature แต่ละคู่นั้นมีความเป็น collinearity กัน

Fit regression model:

- Fit the regression model to predict Heating_oil. P-value of 0.05 (also known in statistical significance level).

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Insulation	3.097	0.499	0.149	0.423	6.208	0.000	****
Temperature	-0.884	0.083	-0.264	0.399	-10.587	0	****
Num_Occupants	0.177	0.474	0.005	0.998	0.374	0.708	
Avg_Age	2.007	0.077	0.539	0.464	26.004	0	****
Home_Size	2.965	0.365	0.123	0.889	8.132	0.000	****
(Intercept)	135.317	9.029	?	?	14.987	0	****

* สร้าง regression model โดยแบ่ง training set : testing set เป็น 0.7 : 0.3

Feature ที่มีค่า p-Value สูงกว่า 0.05 คือ Num_Occupants ดังนั้นจึงไม่ควรนำ feature Num_Occupants มาใช้ในการสร้าง regression model

Use of model:

- Check the ranges of attributes between training attributes & testing attributes. Use operator Filter Examples operator with the parameters attribute_value_filter Avg_Age>= 15.1 | Avg_Age <=72.2, on the unseen dataset.

  - training attributes

			Min	Max	Average
Label Heating_Oil	Integer	0	114	301	195.379
Insulation	Integer	0	2	10	6.110
Temperature	Integer	0	38	90	65.645
Avg_Age	Real	0	15.100	72.200	42.370
Home_Size	Integer	0	1	8	4.580

  - testing attributes

			Min	Max	Average
Label Heating_Oil	Integer	0	114	300	202.104
Insulation	Integer	0	2	10	6.458
Temperature	Integer	0	38	90	63.756
Avg_Age	Real	0	16.100	70.100	43.492
Home_Size	Integer	0	1	8	4.811

* ไม่นำ Num_Occupants มาใช้เนื่องจากมีค่า p-Value สูงกว่า 0.05

- unseen dataset after filter

			Min	Max	Average
∨ **Insulation**	Integer	0	2	10	5.988
∨ **Temperature**	Integer	0	38	90	63.949
∨ **Num_Occupants**	Integer	0	1	10	5.489
∨ **Avg_Age**	Real	0	15.100	72.200	43.674
∨ **Home_Size**	Integer	0	1	8	4.497

- unseen dataset after filter (without Num_Occupants)

			Min	Max	Average
∨ **Insulation**	Integer	0	2	10	5.988
∨ **Temperature**	Integer	0	38	90	63.949
∨ **Avg_Age**	Real	0	15.100	72.200	43.674
∨ **Home_Size**	Integer	0	1	8	4.497

- Write the equation for predicting Heating_oil from the predictors in the model.

```
  3.098 * Insulation
- 0.884 * Temperature
+ 2.006 * Avg_Age
+ 2.965 * Home_Size
+ 135.869
```

   * สมการที่ได้มาจากการใช้ training set = 0.7 ในการสร้าง regression model

- What Heating_oil is predicted for the following attributes:

   o Insulation: 6

   o Temperature: 67

   o Avg_Age: 35.4

   o Home_Size: 5

Row No.	prediction(Heating_Oil)	Insulation	Temperature	Avg_Age	Home_Size
1	181.100	6	67	35.400	5