

**Universidade do Minho**  
Escola de Engenharia

# K nowledge D iscovery

José Machado  
Cristiana Neto

# RAPIDMINER



RapidMiner is a commercial tool for data analysis that uses machine learning and can be considered an alternative to the Weka tool. This tool developed by the company with the same name, has as its main mission to accelerate the process of creating predictive analyzes and make them easier to be applied in practical business scenarios.

**Download RapidMiner Studio:** <https://rapidminer.com/get-started/>



# **CORRELATION WITH RAPIDMINER**

# CONTEXTO E PRESPECTIVA



Sarah is a regional sales manager for a nationwide supplier of fossil fuels for home heating.

Recent volatility in market prices for heating oil specifically, coupled with wide variability in the size of each order for home heating oil, has Sarah concerned.

She needs to find types of behaviors and other factors that may influence the demand for heating oil in the domestic market.

What factors are related to heating oil usage, and how might she use a knowledge of such factors to better manage her inventory, and anticipate demand?

**Data Mining can help her to understand these factors and interactions.**

# BUSINESS UNDERSTANDING



Sarah's goal is to better understand how her company can succeed in the home heating oil market.

She recognizes that there are many factors that influence heating oil consumption, and believes that by investigating the relationship between a number of those factors she will be able to better monitor and respond to heating oil demand. She has selected correlation as a way to model the relationship between the factors she wishes to investigate.

**Correlation** is a statistical measure of how strong the relationships are between attributes in a data set.

# DATA UNDERSTANDING



Using Sarah's employer data, extracted mainly from the company's billing database, a data set was created with the following attributes:

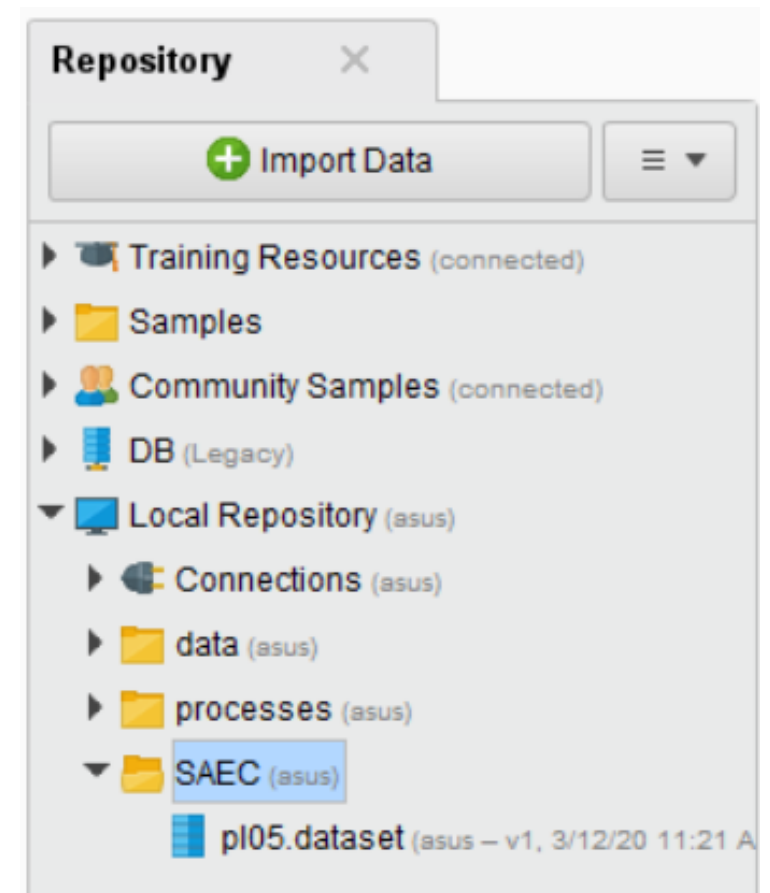
- **Insulation:** This is a density rating, ranging from one to ten, indicating the thickness of each home's insulation A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation
- **Temperature:** This is the average outdoor ambient temperature at each home for the most recent year, measure in degree Fahrenheit
- **Heating\_Oil:** This is the total number of units of heating oil purchased by the owner of each home in the most recent year
- **Num\_Occupants:** This is the total number of occupants living in each home
- **Avg\_Age:** This is the average age of those occupants
- **Home\_Size:** This is a rating, on a scale of one to eight, of the home's overall size The higher the number, the larger the home

# DATA PREPARATION



**Download dataset: pl05-dataset.csv**

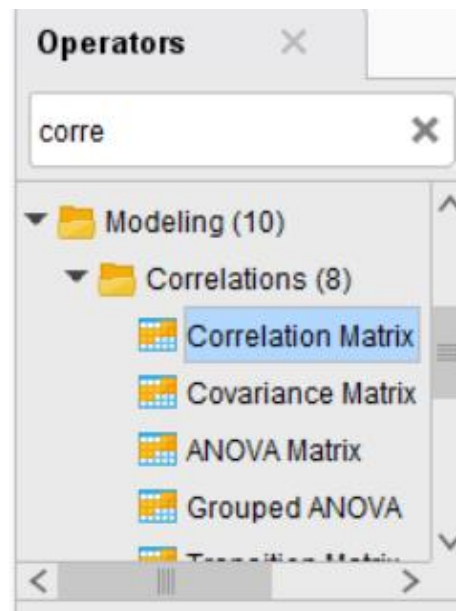
1. Import the CSV to a rapidminer repository (Import Data -> My Computer)
2. Verify the results *view* and inspect the data imported from the CSV (Data, Statistics)



# MODELING



1. Switch to the design perspective and drag the dataset to the process window.
2. On the Operators tab (Data Mining tools section), in the lower left corner, use the search box and type the word 'correlation'. The necessary tool is called 'Correlation Matrix'. Drag it to the process window and drop it.

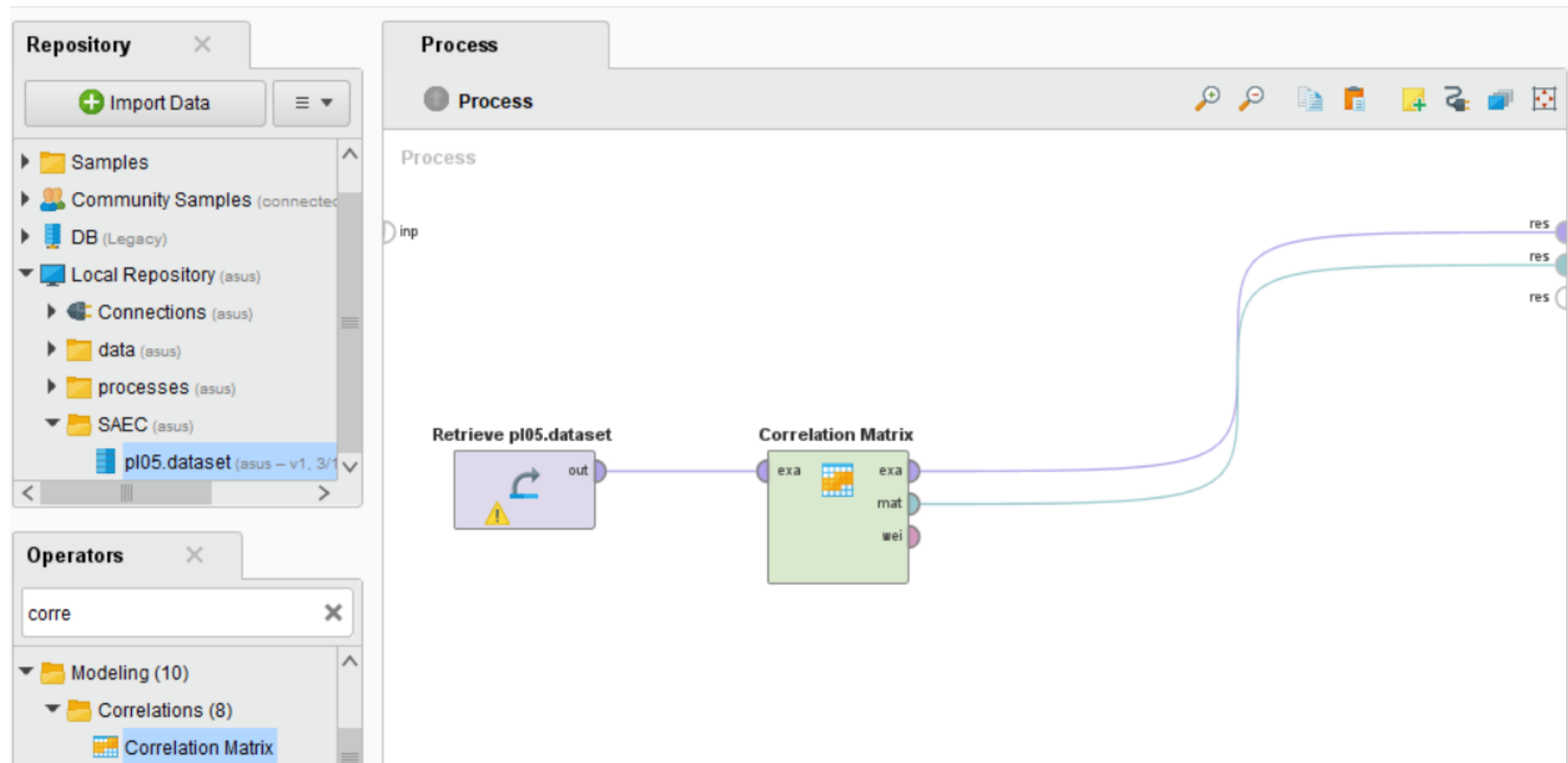




# MODELING



3. Make the connections as shown in the figure. Click on Run.



# MODELING



## Correlation Matrix

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1

# EVALUATION



Correlation Coefficients



between 0 and 1



between 0 and -1



Positive Correlations

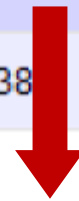


Negative Correlations

# EVALUATION



Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1



The attributes *Heating\_Oil consumption* and *Insulation rating level* have a positive correlation of 0.736.

What does this mean?

# EVALUATION







## What does this mean?

Correlations that are positive mean that as one attribute's value rises, the other attribute's value also rises But, a positive correlation also means that as one attribute's value falls, the other's also falls.

# EVALUATION



When attributes' values move in the same direction, the correlation is **positive**.

				
Heating Oil use rises	Insulation rating also rises		Heating Oil use falls	Insulation rating also falls

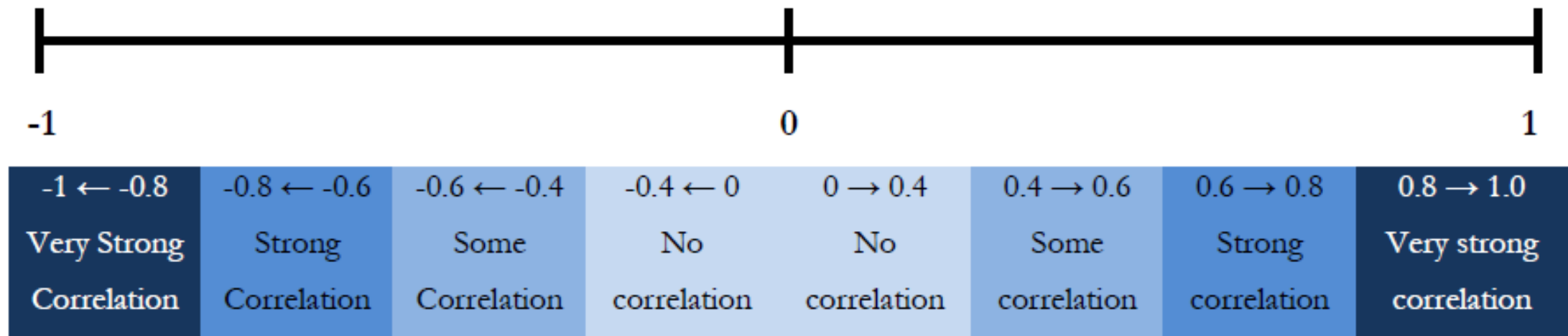
When attributes' values move in opposite directions, a correlation is **negative**.

				
Temperature rises	Insulation rating falls		Temperature falls	Insulation rating rises

# EVALUATION



The correlation coefficients not only allow us to determine the relationship between attributes, but they also tell us something about the **strength** of the correlation.



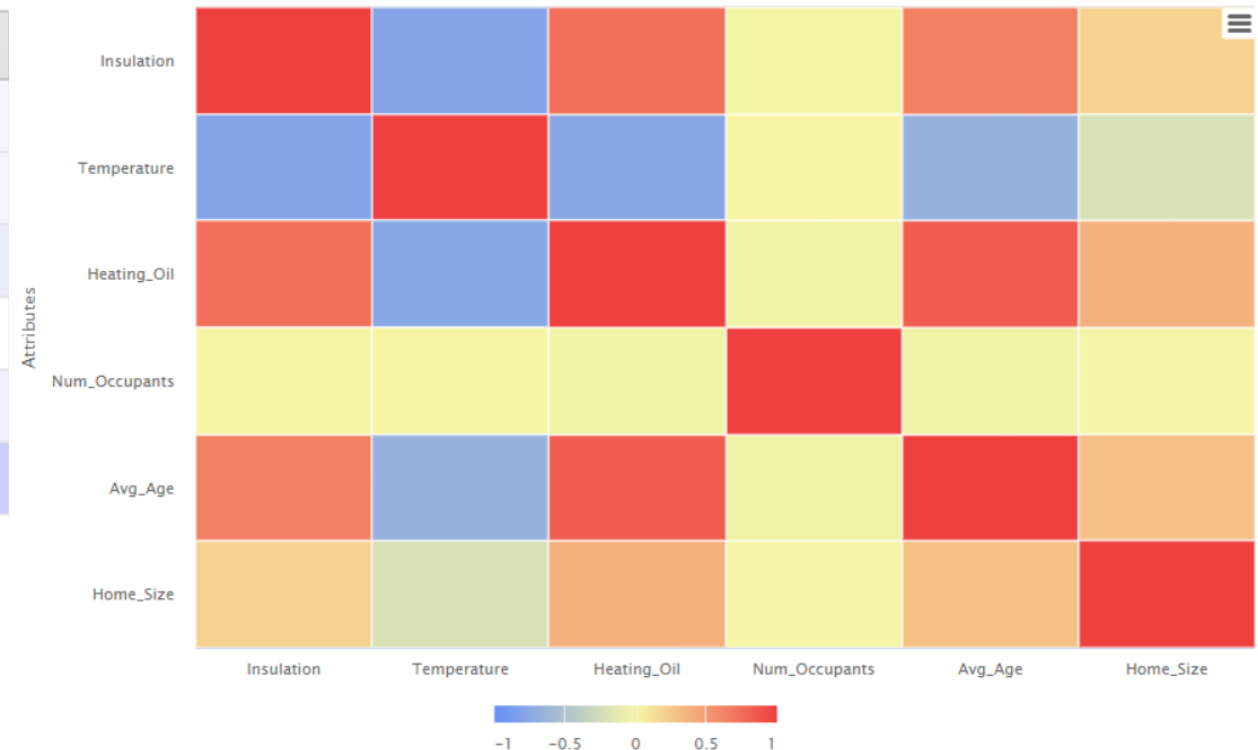
The closer a correlation coefficient is to 1 or -1, the stronger the correlation of the attributes.

# EVALUATION



RapidMiner helps to recognize strong correlations through color coding on both the *Data* tab and the *Matrix Visualization* tab.

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1





# DEPLOYMENT



With this study it was possible to notice that the two most strongly correlated attributes are *Heating\_Oil* and *Avg\_Age*, with a coefficient of 0.848.



As the average age of the occupants of a home increases, so does the use of heating oil in that home. Why? We don't know.



The assumption that correlation proves causation is dangerous and often false.

# DEPLOYMENT



The correlation coefficient between *Avg\_Age* and *Temperature* is -0.673 → strong negative correlation



“As the age of the inhabitants of a house increases, the outside temperature decreases; and as the temperature increases, the age of the residents decreases.”



Although there is a statistically correlation between these two attributes, there is no logical reason why the average age of the occupants of a home may have any effect on the external temperature of the home and vice versa.



The assumption that correlation proves causation is dangerous and often false.

# DEPLOYMENT



Another misinterpretation is that the correlation coefficients are percentages (%).



A correlation coefficient of 0.776  $\neq$  77.6% of variability between these attributes.



The mathematical formula underlying the calculation of the correlation coefficients measures only the strength, as indicated by the proximity of 1 or -1, of the interaction between the attributes.

# DEPLOYMENT



The concept of deployment in data mining means doing something with what you've learned from your model; taking some action based upon what your model tells you.



There are several things that Sarah can do to act based on the model obtained:

Removing the  
**Num\_Occupants**  
attribute

Investigating the role  
of home **insulation**.

Adding greater  
**granularity** in the  
data set.

Adding additional  
**attributes** to the  
*data set*

# DEPLOYMENT



Removing the  
**Num\_Occupants**  
attribute



The number of people living in a home may logically appear to be a variable that influences energy use, but it has not been significantly correlated with any other attributes.

Investigating the role  
of home **insulation**.



The Isolation attribute was highly correlated with a number of other attributes. This means that there may be an opportunity to partner with a company that specializes in adding insulation to existing homes or even creating your own company.

# DEPLOYMENT



Adding greater  
**granularity** in the  
data set.



This data set has attributes of low granularity such as the average annual temperature. Temperatures fluctuate throughout the year and, therefore, monthly or even weekly measurements would show more detailed results and closer to reality.

Adding additional  
**attributes** to the  
*data set*



For example, perhaps the number of instruments that consume heating oil in each home, such as ovens and/or boilers, would add something to Sarah's study.