



Exploring RapidMiner
Correlation process

PL05



Material

<http://hpeixoto.github.io/dc>



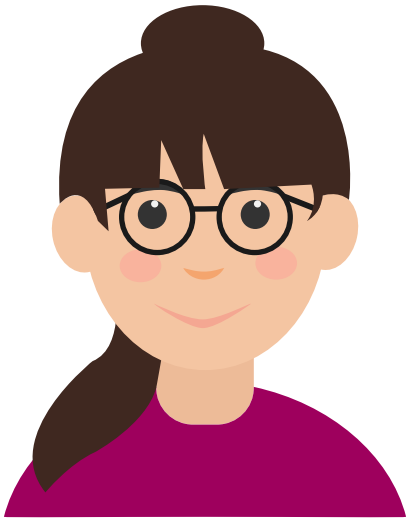
CONTEXT AND PERSPECTIVE

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**.

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.





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

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.

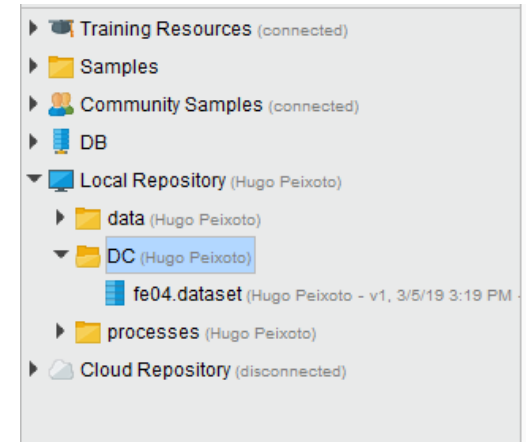


DATA PREPARATION

Download csv: <http://hpeixoto.github.io/dc/pl05/pl05.dataset.csv>

Import csv to rapidminer repository.

Check results tab and inspect metadata view of the imported csv.

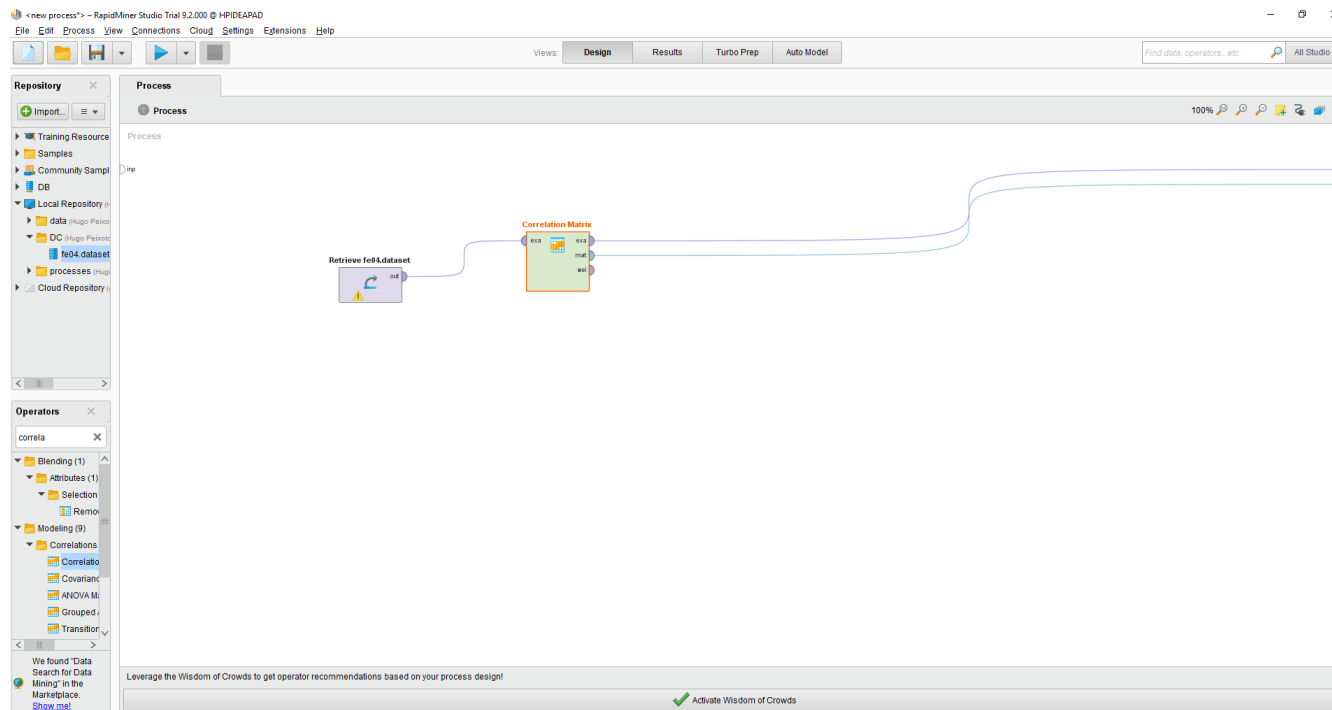




MODELING

On the Operators tab in the lower left hand corner, use the search box and begin typing in the word *correlation*.

The tool we are looking for is called *Correlation Matrix*. Drag and drop and click *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

All correlation coefficients between 0 and 1 represent **positive correlations**, while all coefficients between 0 and -1 are **negative correlations**.

Keep in mind the direction of movement between the two attributes

consider the relationship between the **Heating_Oil** consumption attribute, and the **Insulation** rating level attribute.





The coefficient there, as seen in our matrix, is 0.736. This is a positive number, and therefore, a positive correlation.

But what does that 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

Positive correlations

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

Whenever both attribute values move in the same direction, the correlation is positive.

Negative correlations

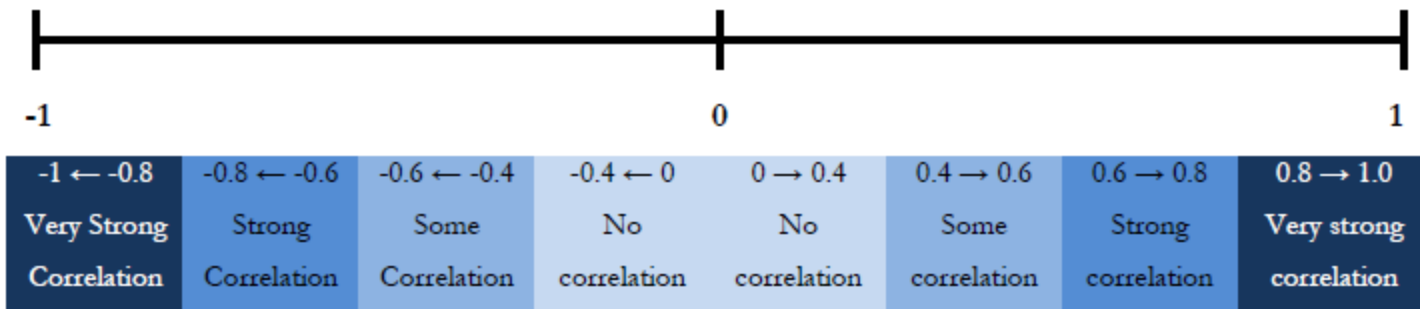
				
Temperature rises	Insulation rating falls		Temperature falls	Insulation rating rises

Whenever attribute values move in opposite directions, the correlation is negative.



EVALUATION

Correlations strengths





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:

We learned through our investigation, that the two most strongly correlated attributes in our data set are **Heating_Oil** and **Avg_Age**, with a coefficient of **0.848**

Thus, we know that in this data set, as the average age of the occupants in a home increases, so too does the heating oil usage in that home.



DRAWBACKS

Consider the correlation coefficient between *Avg_Age* and *Temperature*: -0.673 (strong negative correlation)

As the age of a home's residents increases, the average temperature outside decreases; and as the temperature rises, the age of the folks inside goes down.

Could the average age of a home's occupants have any effect on that home's average yearly outdoor temperature? **Certainly not.** If it did, we could control the temperature by simply moving people of different ages in and out of homes. This of course is silly.

While statistically, there is a correlation between these two attributes in our data set, there is no logical reason that movement in one *causes* movement in the other. The relationship is probably coincidental, but if not, there must be some other explanation that our model cannot offer.

Such limitations must be recognized and accepted in all data mining deployment decisions.



DRAWBACKS

Another false interpretation about correlations is that the coefficients are percentages, as if to say that a correlation coefficient of 0.776 between two attributes is an indication that there is 77.6% shared variability between those two attributes. **This is not correct.**

While the coefficients do tell a story about the shared variability between attributes, the underlying mathematical formula used to calculate correlation coefficients solely measures strength, as indicated by proximity to 1 or -1, of the interaction between attributes. **No percentage is calculated or intended.**



FE04



Exploring RapidMiner
Correlation process

PL05