

Universidade do Minho
Escola de Engenharia

Sistemas de Aprendizagem e Extração de Conhecimento

José Machado

Diana Ferreira

RAPIDMINER



O RapidMiner é uma ferramenta comercial para análise de dados que utiliza *machine learning* e pode ser considerada uma alternativa para a ferramenta Weka. Esta ferramenta desenvolvida pela empresa com o mesmo nome, tem como principal missão acelerar o processo de criação de análises preditivas e torná-las mais fáceis para serem aplicadas em cenários práticos de negócios.

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



CORRELAÇÃO COM O RAPIDMINER

CONTEXTO E PRESPECTIVA



A Sara é gerente regional de vendas de um fornecedor nacional de combustíveis fósseis para aquecimento doméstico.

A recente volatilidade nos preços de mercado do óleo para aquecimento específico, juntamente com uma grande variabilidade no tamanho de cada pedido de óleo para aquecimento doméstico, tem preocupado a Sara.

Ela sente a necessidade de saber os tipos de comportamento e outros fatores que podem influenciar a demanda por óleo para aquecimento no mercado doméstico.

Que fatores estão relacionados com uso de óleo para aquecimento e como se pode usar o conhecimento desses fatores para gerir melhor o inventário e antecipar a demanda?

O Data Mining pode ajudá-la a compreender esses fatores e interações

BUSINESS UNDERSTANDING



O objetivo da Sara é entender melhor como a sua empresa pode ter sucesso no mercado de óleo para aquecimento doméstico.

Ela reconhece que existem muitos fatores que influenciam o consumo de óleo para aquecimento e acredita que, ao investigar a relação entre esses vários fatores, poderá monitorizar e responder melhor à demanda de óleo para aquecimento. A Sara decidiu selecionar a correlação como uma forma de modelar o relacionamento entre os fatores que pretende investigar.

A **correlação** é uma medida estatística que mede o quão fortes são os relacionamentos entre atributos num *data set*.

DATA UNDERSTANDING



Usando os dados do empregador da Sara, extraídos principalmente da base de dados de cobrança da empresa, foi criado um *data set* composto pelos seguintes atributos:

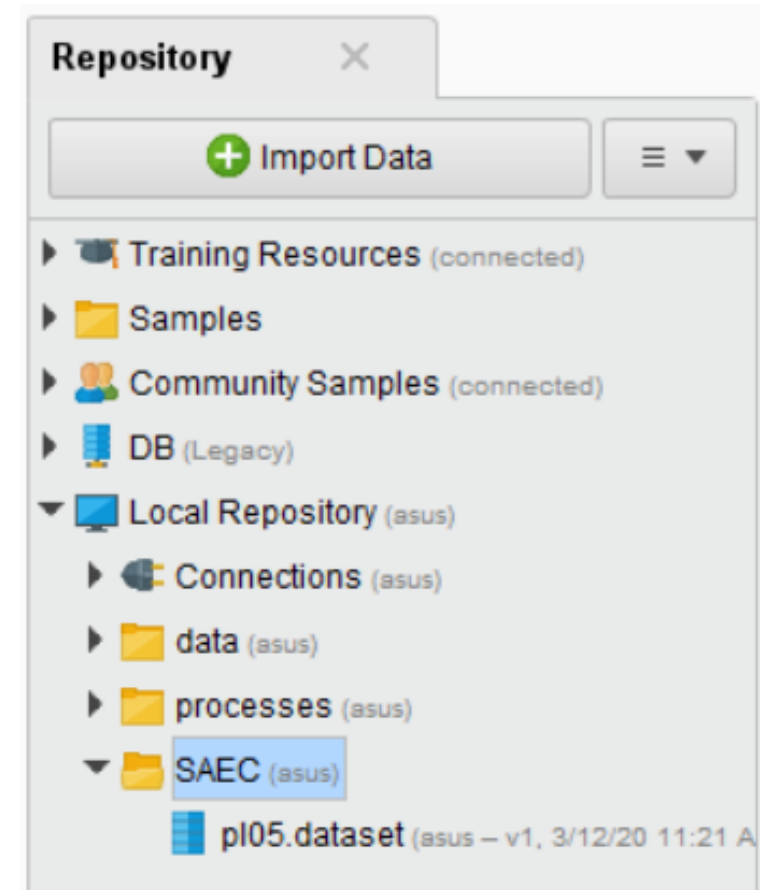
- **Insulation:** classificação de densidade que varia de 1 a 10 e indica a espessura do isolamento de cada casa. Uma casa com uma classificação de densidade de um é mal isolada, enquanto uma casa com uma densidade de dez possui um excelente isolamento.
- **Temperature:** temperatura ambiente média externa de cada casa no ano mais recente, medida em graus Fahrenheit.
- **Heating_Oil:** número total de unidades de óleo de aquecimento adquiridas pelo proprietário de cada casa no ano mais recente.
- **Num_Occupants:** número total de ocupantes que vivem em cada casa.
- **Avg_Age:** idade média dos ocupantes que vivem em cada casa.
- **Home_Size:** classificação, numa escala de 1 a 8, do tamanho geral da casa. Quanto maior o número, maior a casa.

DATA PREPARATION



Download do dataset: pl05-dataset.csv

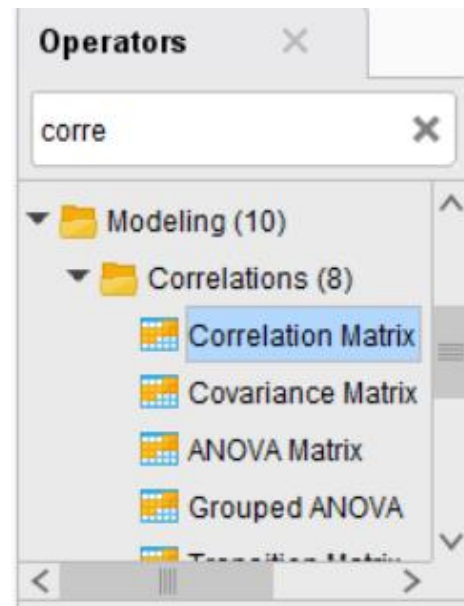
1. Importar o CSV para o repositório rapidminer (Import Data -> My Computer)
2. Verificar a *view* dos resultados e inspecionar os dados CSV importados (Data, Statistics)



MODELING



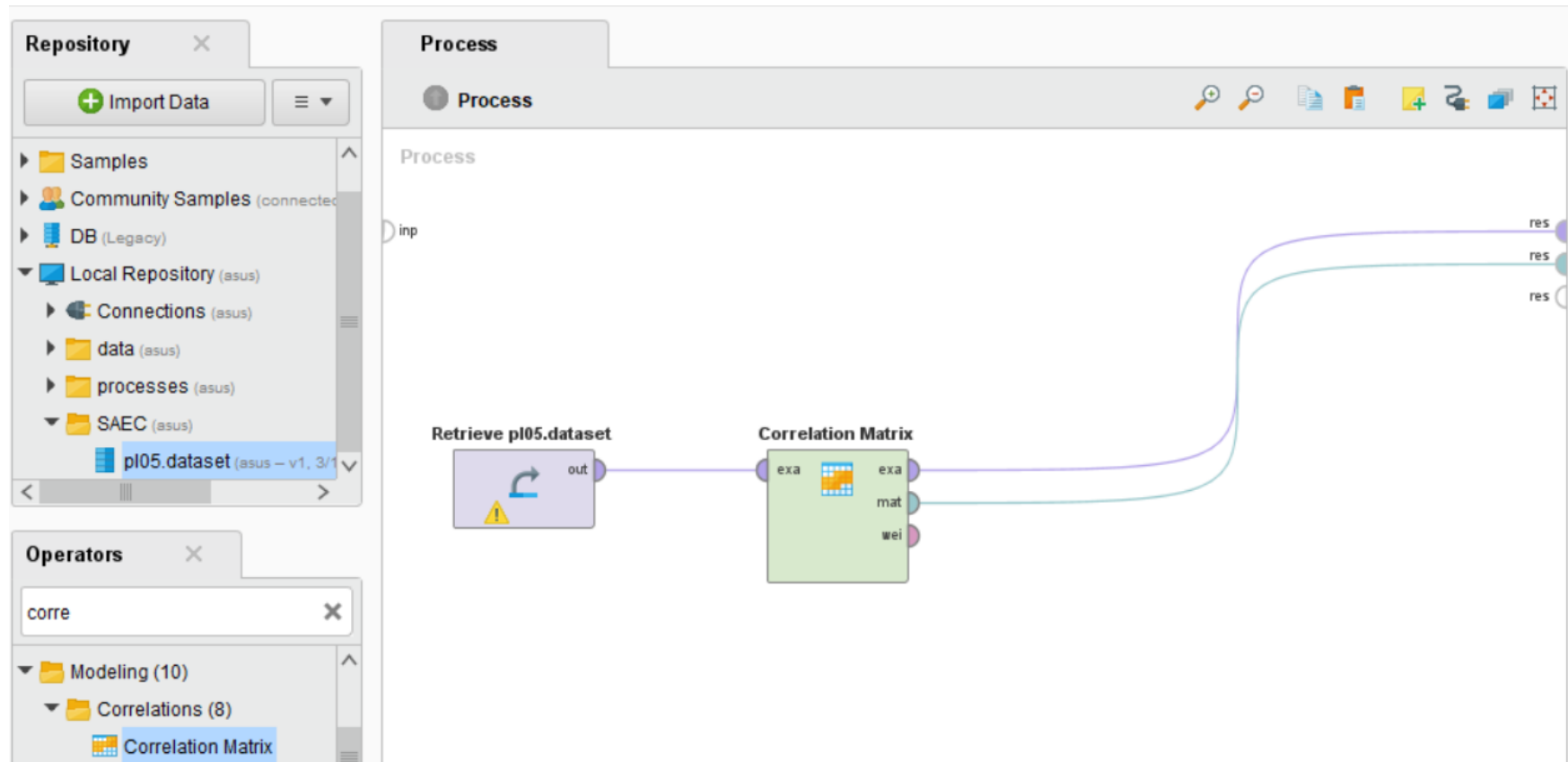
1. Mude para a perspectiva de design e arraste o *dataset* para a janela do processo.
2. No separador Operadores (secção das ferramentas de Data Mining), no canto inferior esquerdo, use a caixa de pesquisa e escreva a palavra 'correlation'. A ferramenta necessária chama-se 'Correlation Matrix'. Arraste-a para a janela do processo e solte-a.



MODELING



3. Estabeleça as ligações tal como representadas na figura. Clique em Run.



MODELING



Matriz de correlação

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



COEFICIENTES DE CORRELAÇÃO

entre 0 e 1

entre 0 e -1

Correlações positivas

Correlações negativas

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

Os atributos *Heating_Oil consumption* e *Insulation rating level* possuem uma correlação positiva de 0.736

O que é que isto significa?

EVALUATION



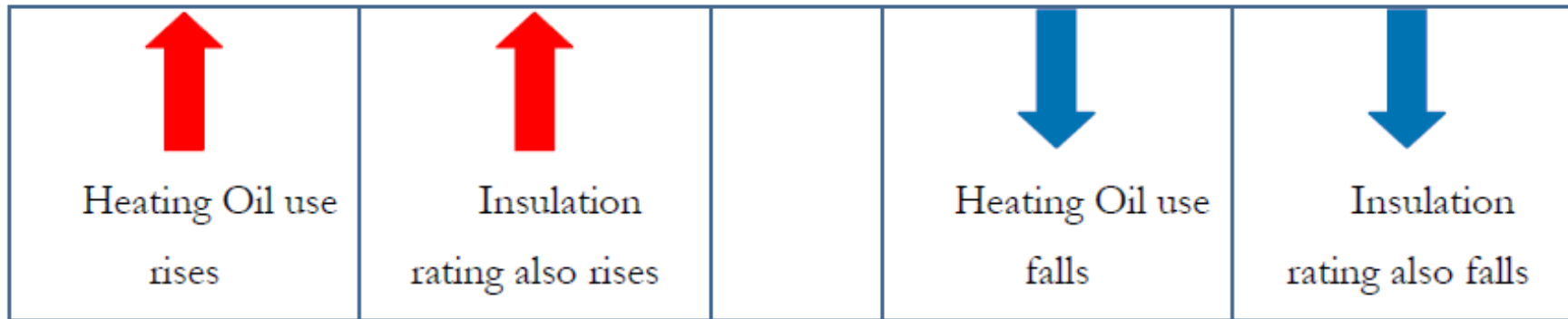
O que é que isto significa?

Correlações que são positivas significam que à medida que o valor de um atributo aumenta, o valor do outro atributo também aumenta. Mas, uma correlação positiva também significa que à medida que o valor de um atributo diminui, o valor do outro atributo também diminui.

EVALUATION



Quando os valores dos atributos se movem na mesma direção, a correlação é positiva.



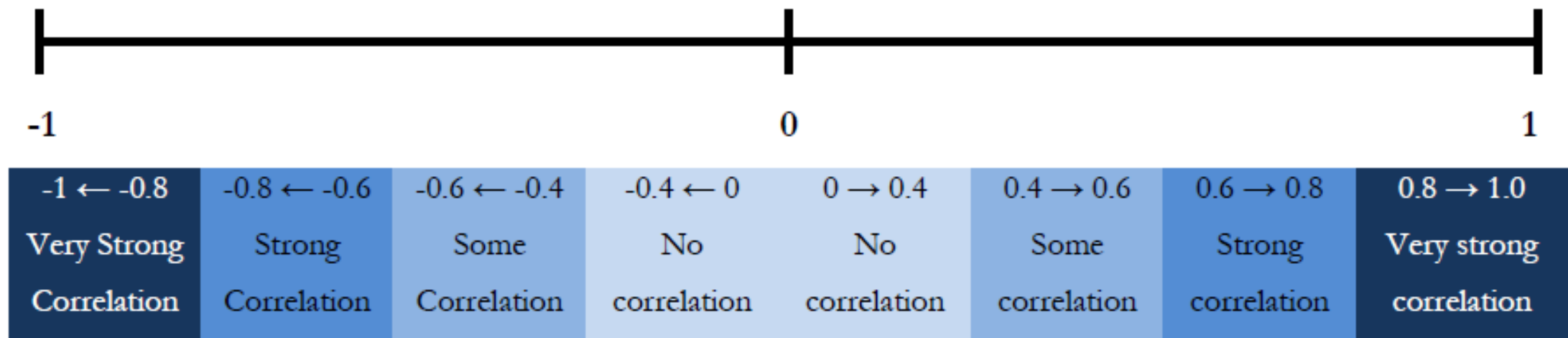
Quando os valores dos atributos se movem em direções opostas, a correlação é negativa.



EVALUATION



Os coeficientes de correlação não permitem apenas determinar a relação entre atributos, mas também nos dizem algo sobre a **força** da correlação.



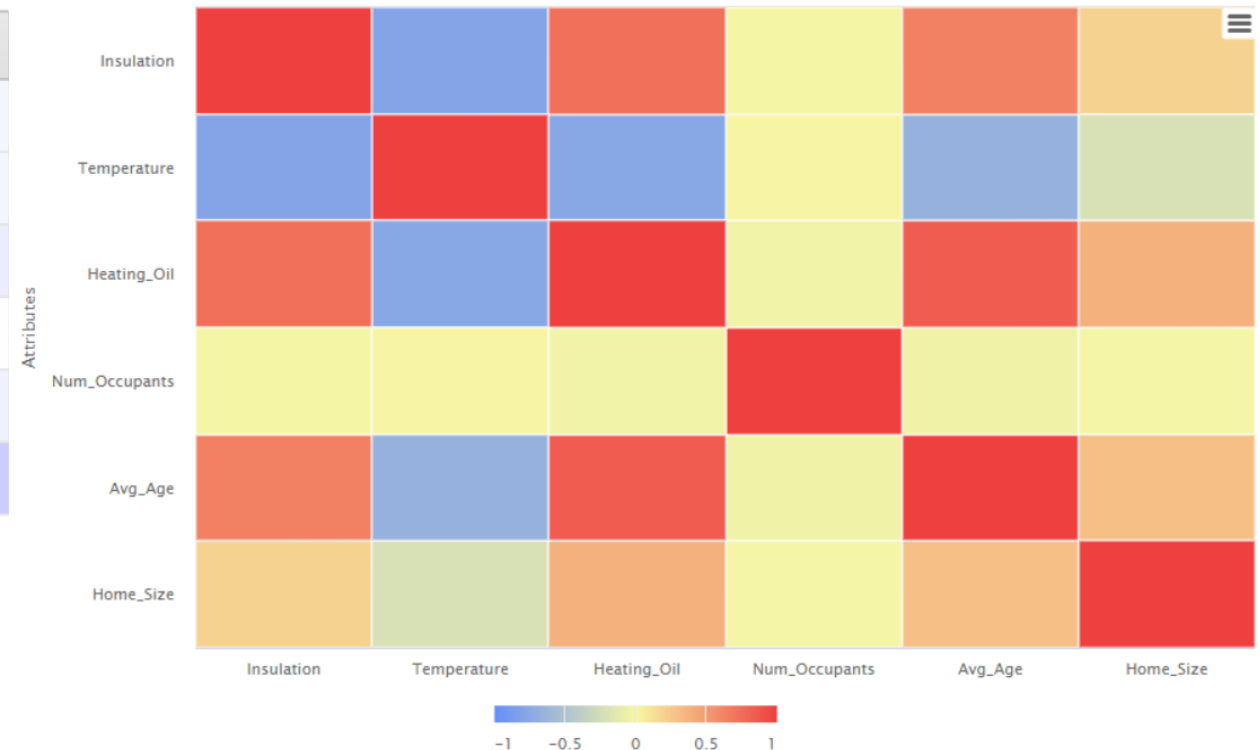
Quanto mais próximo um coeficiente de correlação estiver de 1 ou de -1, mais forte é a correlação dos atributos.

EVALUATION



O RapidMiner ajuda a reconhecer as correlações fortes através de uma codificação por cores tanto no separador *Data* como no separador *Matrix Visualization*.

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



Com este estudo foi possível perceber que os dois atributos mais fortemente correlacionados são o *Heating_Oil* e o *Avg_Age*, com um coeficiente de 0,848.



À medida que a idade média dos ocupantes de uma casa aumenta, aumenta também o uso de óleo de aquecimento nessa casa. Porquê?



A suposição de que uma correlação prova causalidade é perigosa e muitas vezes falsa

DEPLOYMENT



O coeficiente de correlação entre *Avg_Age* e *Temperature* é de -0.673 → correlação negativa forte



“À medida que a idade dos moradores de uma casa aumenta, a temperatura externa diminui; e à medida que a temperatura aumenta, a idade dos moradores diminui.”



Embora estatisticamente exista uma correlação entre estes dois atributos, não há nenhuma razão lógica para que a idade média dos ocupantes de uma casa possa ter algum efeito sobre a temperatura externa da casa e vice-versa.



A suposição de que uma correlação prova causalidade é perigosa e muitas vezes falsa

DEPLOYMENT



Outra falsa interpretação é que os coeficientes de correlação são percentagens (%).



Um coeficiente de correlação de 0,776 \neq 77,6% de variabilidade entre esses atributos.



A fórmula matemática subjacente ao cálculo dos coeficientes de correlação mede apenas a força, como indicado pela proximidade de 1 ou -1, da interação entre os atributos.

DEPLOYMENT



O conceito de *deployment* (implementação) em *Data Mining* significa fazer algo com os resultados do modelo, ou seja, tomar algumas medidas com base no que o modelo aprendeu.



Existem várias coisas que a Sara pode fazer para agir com base no modelo obtido:

Remover o atributo
Num_Occupants

Investigar o papel do
isolamento da casa

Aumentar a
granularidade do
data set

Adicionar **atributos**
ao *data set*

DEPLOYMENT



Remover o atributo
Num_Occupants



O número de pessoas que vivem numa casa pode logicamente parecer uma variável que influencia o uso de energia, mas este não se correlacionou de forma significativa com mais nenhum atributo.

Investigar o papel do
isolamento da casa



O atributo de Isolamento foi bastante correlacionado com uma série de outros atributos. Isto significa que pode haver a oportunidade de fazer parceria com uma empresa especializada em adicionar isolamento às casas existentes ou até mesmo criar a sua própria empresa.

DEPLOYMENT



Aumentar a
granularidade do
dataset



Este *data set* tem atributos de baixa granularidade como a temperatura média anual. As temperaturas flutuam ao longo do ano e, portanto, medidas mensais, ou mesmo semanais, mostrariam resultados mais detalhados e próximos da realidade.

Adicionar atributos
ao *data set*



Por exemplo, talvez o número de instrumentos que consomem óleo de aquecimento em cada casa, como fornos e/ou caldeiras, acrescentasse algo ao estudo da Sara.