CHAPTER FOUR:
CORRELATION

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. She feels a need to understand the 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? Sarah believes that data mining can help her begin to formulate an understanding of these factors and interactions.

LEARNING OBJECTIVES

After completing the reading and exercises in this chapter, you should be able to:

- Explain what correlation is, and what it isn't.
- Recognize the necessary format for data in order to perform correlation analysis.
- Develop a correlation model in RapidMiner.
- Interpret the coefficients in a correlation matrix and explain their significance, if any.

ORGANIZATIONAL 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

In order to investigate her question, Sarah has enlisted our help in creating a correlation matrix of six attributes. Working together, using Sarah's employer's data resources which are primarily drawn from the company's billing database, we create a data set comprised of 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

A CSV data set for this chapter's example is available for download at the book's companion web site (https://sites.google.com/site/dataminingforthemasses/). If you wish to follow along with the example, go ahead and download the Chapter04DataSet.csv file now and save it into your RapidMiner data folder. Then, complete the following steps to prepare the data set for correlation mining:

1) Import the Chapter 4 CSV data set into your RapidMiner data repository. Save it with the name Chapter4. If you need a refresher on how to bring this data set into your RapidMiner repository, refer to steps 7 through 14 of the Hands On Exercise in Chapter 3. The steps will be the same, with the exception of which file you select to import. Import all attributes, and accept the default data types. When you are finished, your repository should look similar to Figure 4-1.

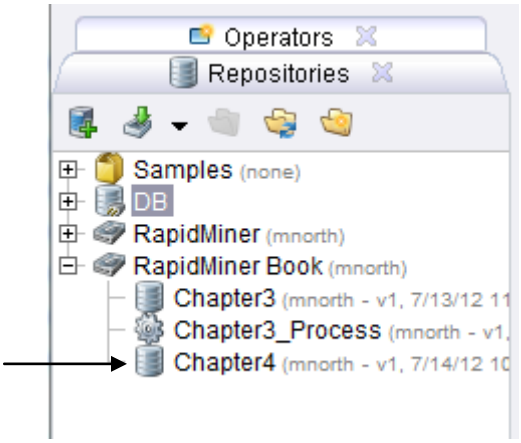
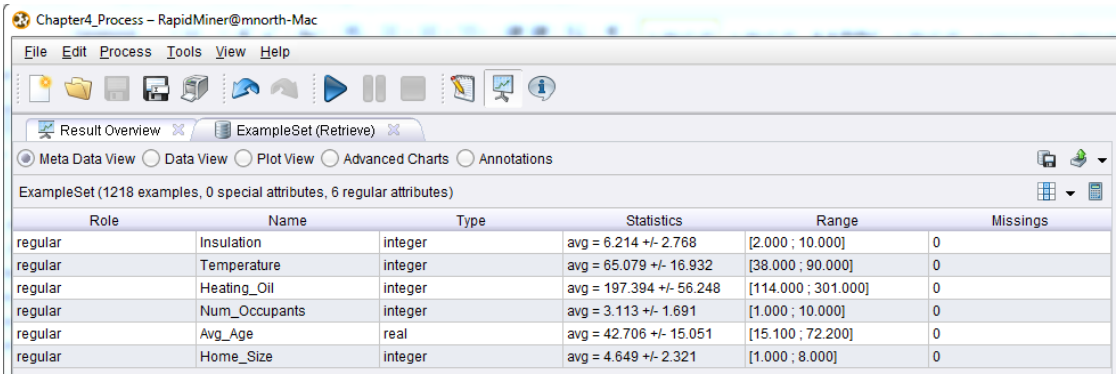


Figure 4-1. The chapter four data set added to the author's RapidMiner Book repository.

2) If your RapidMiner application is not open to a new, blank process window, click the new process icon, or click File > New to create a new process. Drag your Chapter4 data set into your main process window. Go ahead and click the run (play) button to examine the data set's meta data. If you are prompted, you may choose to save your new model. For this book's example, we'll save the model as Chapter4_Process.



Role	Name	Type	Statistics	Range	Missings
regular	Insulation	integer	avg = 6.214 +/- 2.768	[2.000 ; 10.000]	0
regular	Temperature	integer	avg = 65.079 +/- 16.932	[38.000 ; 90.000]	0
regular	Heating_Oil	integer	avg = 197.394 +/- 56.248	[114.000 ; 301.000]	0
regular	Num_Occupants	integer	avg = 3.113 +/- 1.691	[1.000 ; 10.000]	0
regular	Avg_Age	real	avg = 42.706 +/- 15.051	[15.100 ; 72.200]	0
regular	Home_Size	integer	avg = 4.649 +/- 2.321	[1.000 ; 8.000]	0

Figure 4-2. Meta Data view of the chapter four data set.

We can see in Figure 4-2 that our six attributes are shown. There are a total of 1,218 homes represented in the data set. Our data set appears to be very clean, with no missing values in any of the six attributes, and no inconsistent data apparent in our ranges or other descriptive statistics. If you wish, you can take a minute to switch to Data View to familiarize yourself with the data. It feels like these data are in good shape, and are in no further need of data preparation operators, so we are ready to move on to…

MODELING

3) Switch back to design perspective. 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. You may be able to find it before you even finish typing the full search term. Once you've located it, drag it over into your process window and drop it into your stream. By default, the *exa* port will connect to the *res* port, but in this chapter's example we are interested in creating a matrix of correlation coefficients that we can analyze. Thus, is it important for you to connect the *mat* (matrix) port to a *res* port, as illustrated in Figure 4-3.

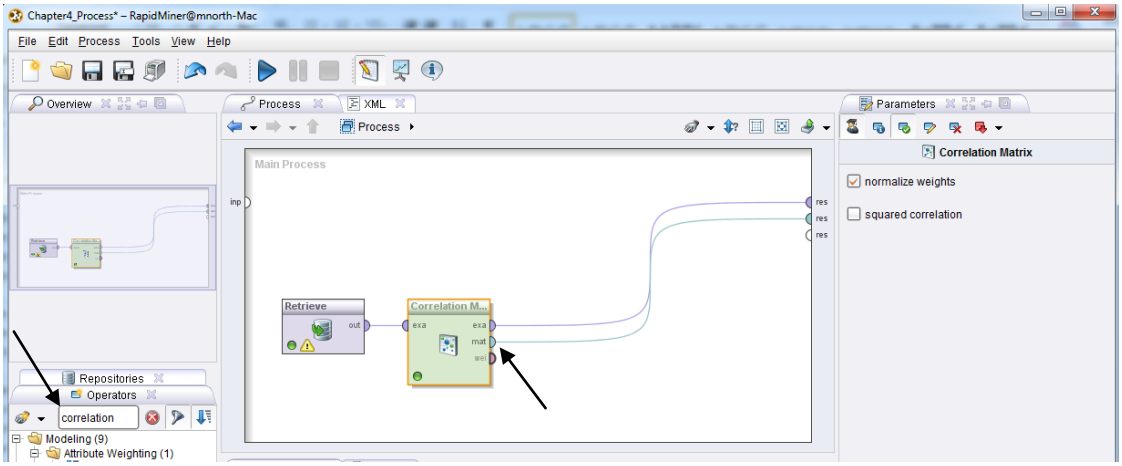
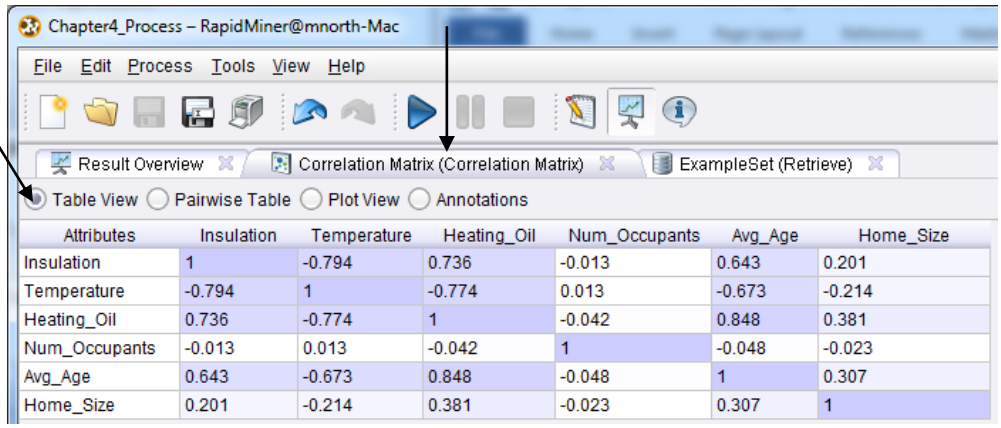


Figure 4-3. The addition of a Correlation Matrix to our stream, with the *mat* (matrix) port connected to a result set (*res*) port.

4) Correlation is a relatively simple statistical analysis tool, so there are few parameters to modify. We will accept the defaults, and run the model. The results will be similar to Figure 4-4.



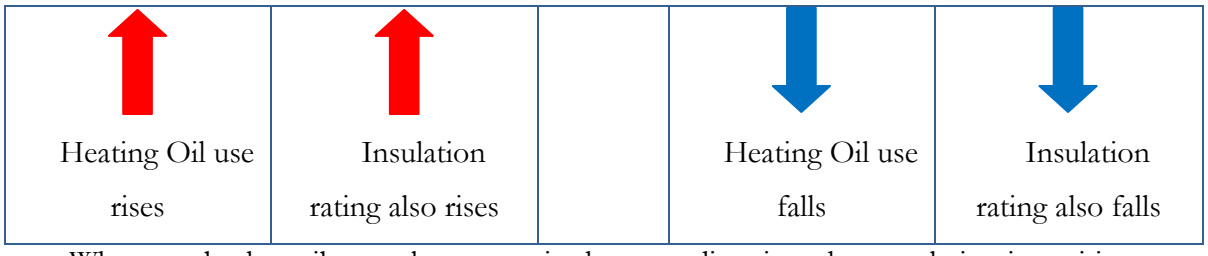
Attributes	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_Oil	0.736	-0.774	1	-0.042	0.848	0.381
Num_Occupants	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1

Figure 4-4. Results of a Correlation Matrix.

5) In Figure 4-4, we have our **correlation coefficients** in a matrix. Correlation coefficients are relatively easy to decipher. They are simply a measure of the strength of the relationship between each possible set of attributes in the data set. Because we have six attributes in this data set, our matrix is six columns wide by six rows tall. In the location where an attribute intersects with itself, the correlation coefficient is '1', because everything compared to itself has a perfectly matched relationship. All other pairs of attributes will have a correlation coefficient of less than one. To complicate matters a bit, correlation coefficients can actually be negative as well, so all correlation coefficients will fall somewhere between -1 and 1. We can see that this is the case in Figure 4-4, and so we can now move on to the CRISP-DM step of…

EVALUATION

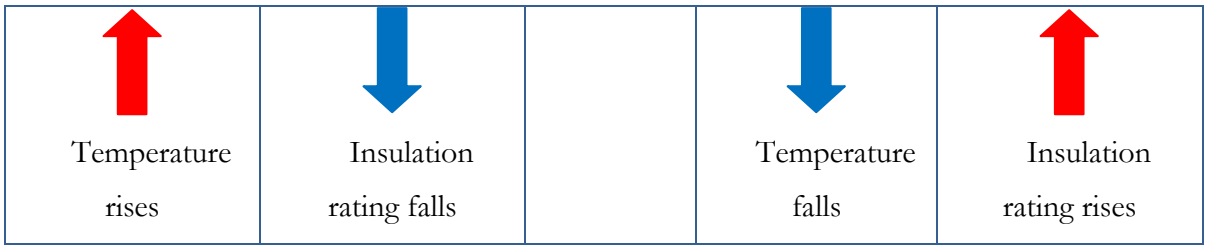
All correlation coefficients between 0 and 1 represent **positive correlations**, while all coefficients between 0 and -1 are **negative correlations**. While this may seem straightforward, there is an important distinction to be made when interpreting the matrix's values. This distinction has to do with the direction of movement between the two attributes being analyzed. Let's consider the relationship between the Heating_Oil consumption attribute, and the Insulation rating level attribute. The coefficient there, as seen in our matrix in Figure 4-4, 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. Data analysts sometimes make the mistake in thinking that a negative correlation exists if an attribute's values are decreasing, but if its corresponding attribute's values are also decreasing, the correlation is still a positive one. This is illustrated in Figure 4-5.



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

Figure 4-5. Illustration of positive correlations.

Next, consider the relationship between the Temperature attribute and the Insulation rating attribute. In our Figure 4-4 matrix, we see that the coefficient there is -0.794. In this example, the correlation is negative, as illustrated in Figure 4-6.



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

Figure 4-6. Illustration of negative correlations.

So correlation coefficients tell us something about the relationship between attributes, and this is helpful, but they also tell us something about the *strength* of the correlation. As previously mentioned, all correlations will fall between 0 and 1 or 0 and -1. The closer a correlation coefficient is to 1 or to -1, the stronger it is. Figure 4-7 illustrates the correlation strength along the continuum from -1 to 1.

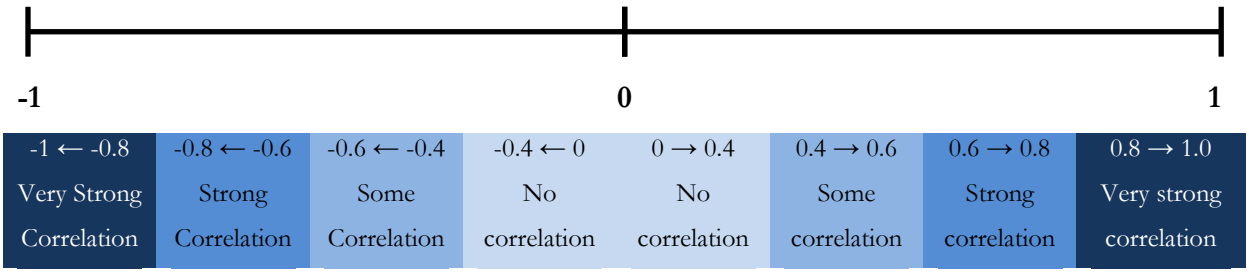


Figure 4-7. Correlation strengths between -1 and 1.

RapidMiner attempts to help us recognize correlation strengths through color coding. In the Figure 4-4 matrix, we can see that some of the cells are tinted with shades of purple in graduated colors, in order to more strongly highlight those with stronger correlations. It is important to recognize that these are only general guidelines and not hard-and-fast rules. A correlation coefficient around .2 does show some interaction between attributes, even if it is not statistically significant. This should be kept in mind as we proceed to…

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. In this chapter's example, we conducted some basic, exploratory analysis for our fictional figure, Sarah. There are several possible outcomes from this investigation.

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. What we *do not know* is why that occurs. Data analysts often make the mistake of interpreting correlation as causation. The assumption that correlation proves causation is dangerous and often false.

Consider for a moment the correlation coefficient between Avg_Age and Temperature: -0.673. Referring back to Figure 4-7, we see that this is considered to be a relatively 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. But 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.

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.

With these interpretation parameters explained, there may be several things that Sarah *can do* in order to take action based upon our model. A few options might include:

- Dropping the Num_Occupants attribute. While the number of people living in a home might logically seem like a variable that would influence energy usage, in our model it did not correlate in any significant way with anything else. Sometimes there are attributes that don't turn out to be very interesting.

- Investigating the role of home insulation. The Insulation rating attribute was fairly strongly correlated with a number of other attributes. There may be some opportunity there to partner with a company (or start one…?) that specializes in adding insulation to existing homes. If she is interested in contributing to conservation, working on a marketing promotion to show the benefits of adding insulation to a home might be a good course of action, however if she wishes to continue to sell as much heating oil as she can, she may feel conflicted about participating in such a campaign.

- Adding greater granularity in the data set. This data set has yielded some interesting results, but frankly, it's pretty general. We have used average yearly temperatures and total annual number of heating oil units in this model. But we also know that temperatures fluctuate throughout the year in most areas of the world, and thus monthly, or even weekly measures would not only be likely to show more detailed results of demand and usage over time, but the correlations between attributes would probably be more interesting. From our model, Sarah now knows how certain attributes interact with one another, but in the day-to-day business of doing her job, she'll probably want to know about usage over time periods shorter than one year.

- Adding additional attributes to the data set. It turned out that the number of occupants in the home didn't correlate much with other attributes, but that doesn't mean that other attributes would be equally uninteresting. For example, what if Sarah had access to the number of furnaces and/or boilers in each home? Home_size was slightly correlated with Heating_Oil usage, so perhaps the number of instruments that consume heating oil in each home would tell an interesting story, or at least add to her insight.

Sarah would also be wise to remember that the CRISP-DM approach is cyclical in nature. Each month as new orders come in and new bills go out, as new customers sign up for a heating oil account, there are additional data available to add into the model. As she learns more about how each attribute in her data set interacts with others, she can increase our correlation model by adding not only new attributes, but also, new observations.

CHAPTER SUMMARY

This chapter has introduced the concept of correlation as a data mining model. It has been chosen as the first model for this book because it is relatively simple to construct, run and interpret, thus serving as an easy starting point upon which to build. Future models will become more complex, but continuing to develop your skills in RapidMiner and getting comfortable with the tools will make the more complex models easier for you to achieve as we move forward.

Recall from Chapter 1 (Figure 1-2) that data mining has two somewhat interconnected sides: Classification, and Prediction. Correlation has been shown to be primarily on the side of Classification. We do not infer causation using correlation metrics, nor do we use correlation coefficients to predict one attribute's value based on another's. We can however quickly find general trends in data sets using correlations, and we can anticipate how strongly an observed movement in one attribute will occur in conjunction with movement in another.

Correlation can be a quick and easy way to see how elements of a given problem may be interacting with one another. Whenever you find yourself asking how certain factors in a problem you're trying to solve interact with one another, consider building a correlation matrix to find out. For example, does customer satisfaction change based on time of year? Does the amount of rainfall change the price of a crop? Does household income influence which restaurants a person patronizes? The answer to each of these questions is probably 'yes', but correlation can not only help us know if that's true, it can also help us learn how strongly the interactions are when, and if, they occur.

REVIEW QUESTIONS

1) What are some of the limitations of correlation models?

2) What is a correlation coefficient? How is it interpreted?

3) What is the difference between a positive and a negative correlation? If two attributes have values that decrease at essentially the same rate, is that a negative correlation? Why or why not?

4) How is correlation strength measured? What are the ranges for strengths of correlation?

5) The number of heating oil consuming devices was suggested as a possibly interesting attribute that could be added to the example data set for this chapter. Can you think of others? Why might they be interesting? To what other attributes in the data set do you think your suggested attributes might be correlated? What would be the value in knowing if they are?

EXERCISE

It is now your turn to develop a correlation model, generate a coefficient matrix, and analyze the results. To complete this chapter's exercise, follow the steps below.

1) Select a professional sporting organization that you enjoy, or of which you are aware. Locate that organization's web site and search it for statistics, facts and figures about the athletes in that organization.

2) Open OpenOffice Calc, and starting in Cell A across Row 1 of the spreadsheet, define some attributes (at least three or four) to hold data about each athlete. Some possible attributes you may wish to consider could be annual_salary, points_per_game, years_as_pro, height, weight, age, etc. The list is potentially unlimited, will vary based on the type of sport you choose, and will depend on the data available to you on the web site you've selected. Measurements of the athletes' salaries and performance in competition are

likely to be the most interesting. You may include the athletes' names, however keep in mind that correlations can only be conducted on numeric data, so the name attribute would need to be reduced out of your data set before creating your correlation matrix. (Remember the Select Attributes operator!)

3) Look up the statistics for each of your selected attributes and enter them as observations into your spreadsheet. Try to find as many as you can—at least thirty is a good rule of thumb in order to achieve at least a basic level of statistical validity. More is better.

4) Once you've created your data set, use the menu to save it as a CSV file. Click File, then Save As. Enter a file name, and change 'Save as type:' to be Text CSV (.csv). Be sure to save the file in your data mining data folder.

5) Open RapidMiner and import your data set into your RapidMiner repository. Name it Chapter4Exercise, or something descriptive so that you will remember what data are contained in the data set when you look in your repository.

6) Add the data set to a new process in RapidMiner. Ensure that the *out* port is connected to a *res* port and run your model. Save your process with a descriptive name if you wish. Examine your data in results perspective and ensure there are no missing, inconsistent, or other potentially problematic data that might need to be handled as part of your Data Preparation phase. Return to design perspective and handle any data preparation tasks that may be necessary.

7) Add a Correlation Matrix operator to your stream and ensure that the *mat* port is connected to a *res* port. Run your model again. Interpret your correlation coefficients as displayed on the matrix tab.

8) Document your findings. What correlations exist? How strong are they? Are they surprising to you and if so, why? What other attributes would you like to add? Are there any you'd eliminate now that you've mined your data?

Challenge step!

9) While still in results perspective, click on the ExampleSet tab (which exists assuming you left the *exa* port connected to a *res* port when you were in design perspective). Click on the Plot View radio button. Examine correlations that you found in your model visually by creating a scatter plot of your data. Choose one attribute for your x-Axis and a correlated one for your y-Axis. Experiment with the **Jitter** slide bar. What is it doing? (Hint: Try an Internet search on the term 'jittering statistics'.) For an additional visual experience, try a Scatter 3D or Scatter 3D Color plot. Consider Figures 4-8 and 4-9 as examples. Note that with 3D plots in RapidMiner, you can click and hold to rotate your plot in order to better see the interactions between the data.

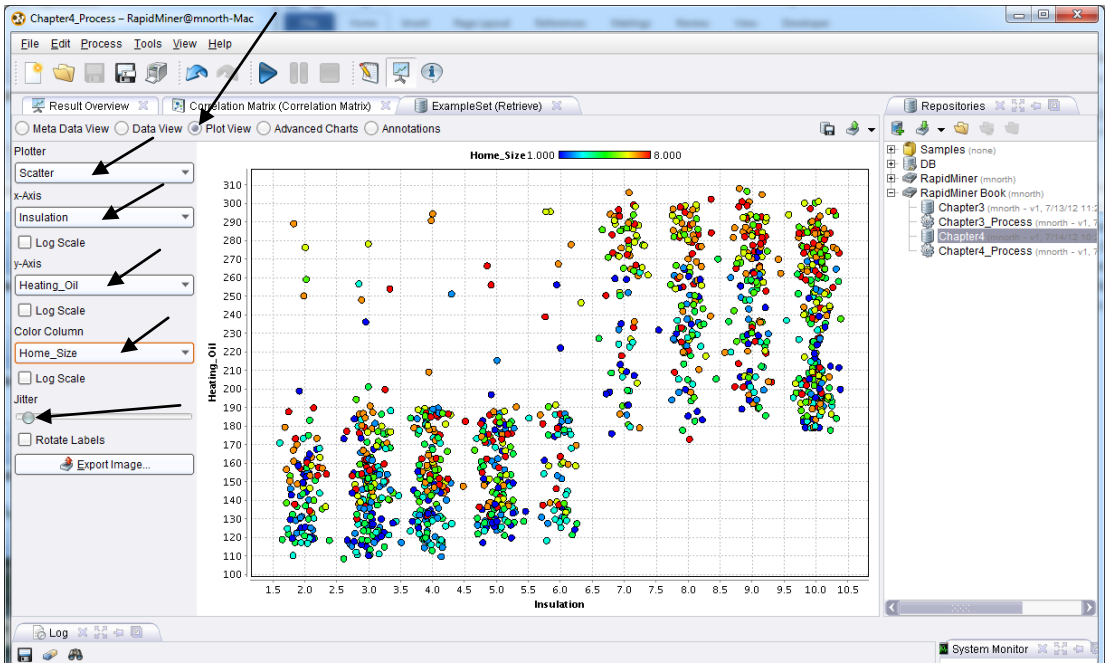


Figure 4-8. A two-dimensional scatterplot with a colored third dimension and a slight jitter.

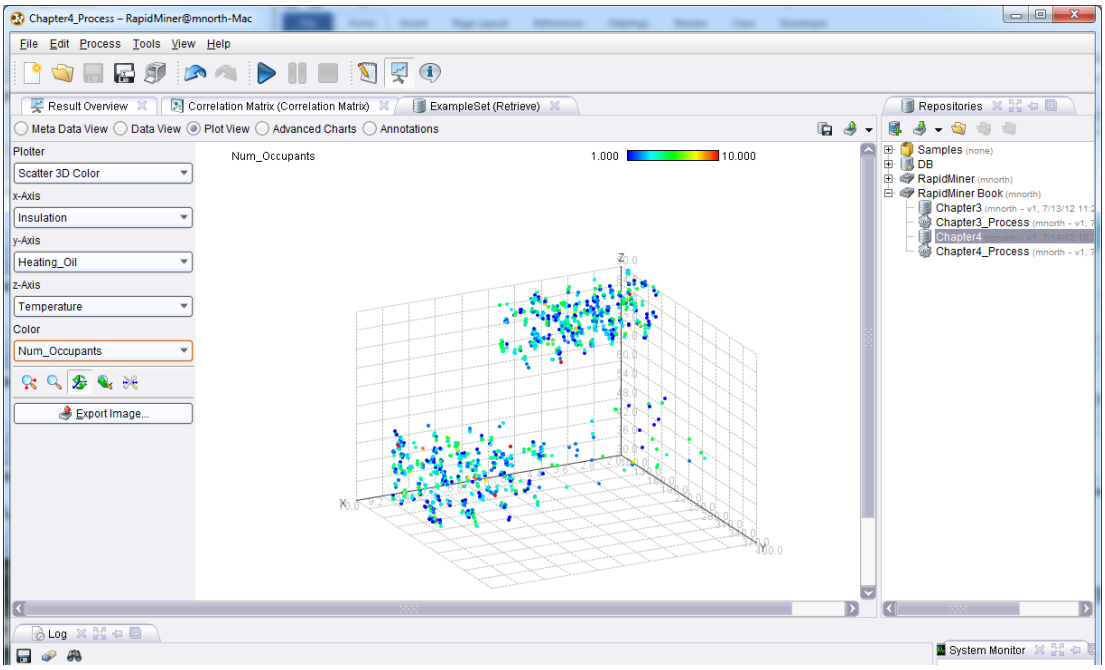


Figure 4-9. A three-dimensional scatterplot with a colored fourth dimension.