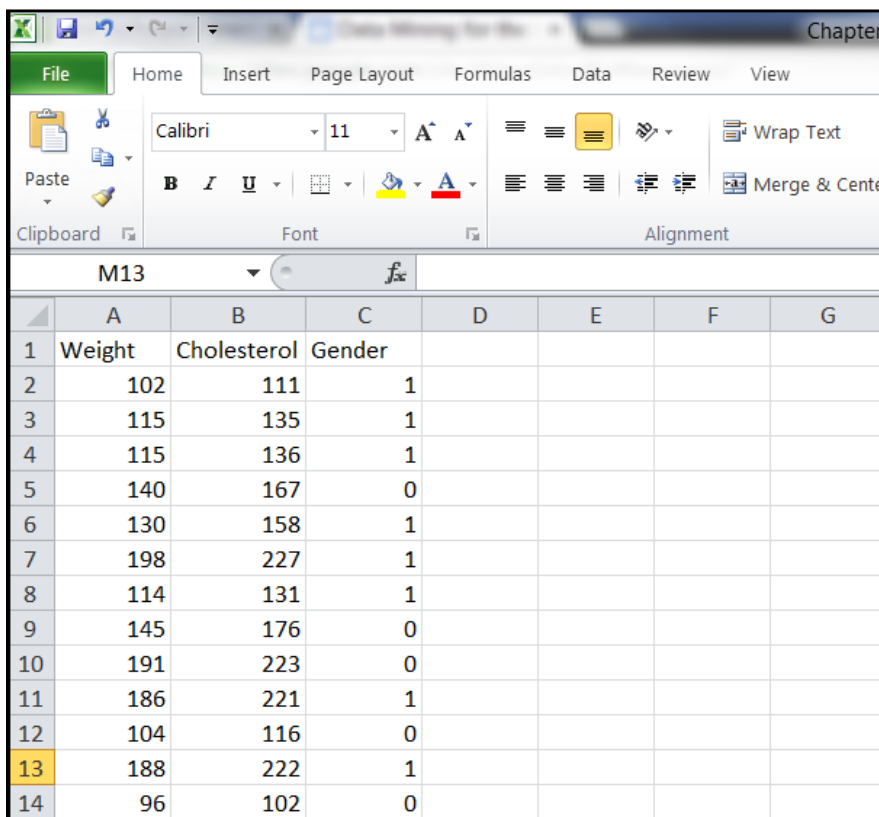


Special Note: Learning can be a difficult endeavor. At any point in this hands on demonstration, if you have any trouble, PLEASE reach out to me at 570-372-4465 or [robbinsr@susqu.edu](mailto:robbinsr@susqu.edu). I am very serious about your learning.

Imagine that you are a healthy lifestyle program director for a healthcare insurance company<sup>1</sup> known as the Prosperential Insurance Company of America. Your responsibilities at Prosperential include offering programs and incentives to your customers to help them transition to healthier lifestyles. If your programs are successful, some customers will need less health care services than would be the case otherwise. As part of this responsibility, you want to identify clusters of individuals that your company insures who are at risk of developing coronary heart disease as a result of being overweight and/or having high cholesterol and who may be benefit from participating in a "heart-healthy" program.

1. If you have not installed RapidMiner Studio (Starter Version), please do so prior to working through this demonstration. A free "Starter" version of RapidMiner is available at <http://rapidminer.com>.
2. Please download the file named **Chapter06DataSet.csv** from <http://bit.ly/1h0URh4> into your Downloads folder.
3. Open the **Chapter06DataSet.csv** file in Excel if you would like to review the data. The data has 547 records. Each record represents a person and has three values: a person's gender, a person's cholesterol level, and a person's weight, in pounds. A screenshot of the first 14 records is shown below so that you can become familiar with how the data is provided to us, prior to our use of it in RapidMiner. Note that in this data, a "1" represents a male and a "0" represents a female.

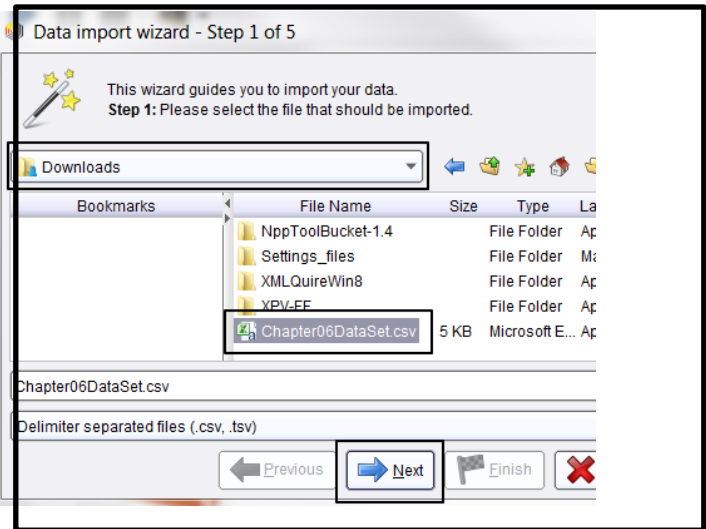
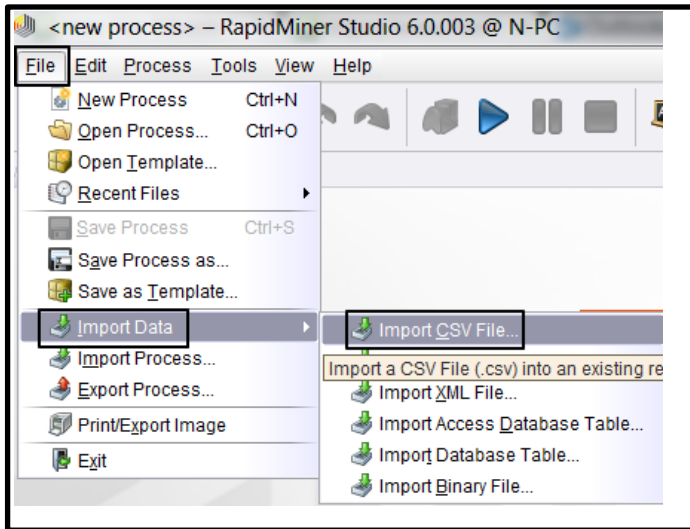


	A	B	C	D	E	F	G
1	Weight	Cholesterol	Gender				
2	102	111	1				
3	115	135	1				
4	115	136	1				
5	140	167	0				
6	130	158	1				
7	198	227	1				
8	114	131	1				
9	145	176	0				
10	191	223	0				
11	186	221	1				
12	104	116	0				
13	188	222	1				
14	96	102	0				

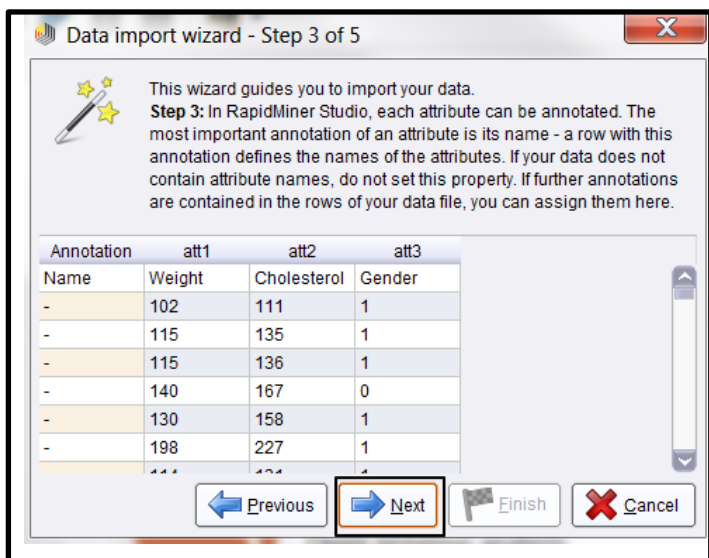
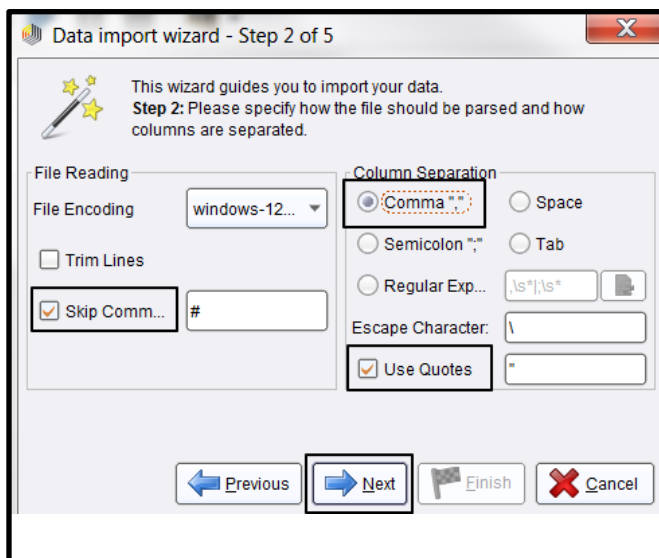
Please proceed to next page.






<sup>1</sup> This sample demonstration uses a data set and extends an example presented in the text Data Mining for the Masses, authored by Matt North and available for purchase at <http://amzn.to/1mKkibB>. The data set is available at <http://bit.ly/1h0URh4>.

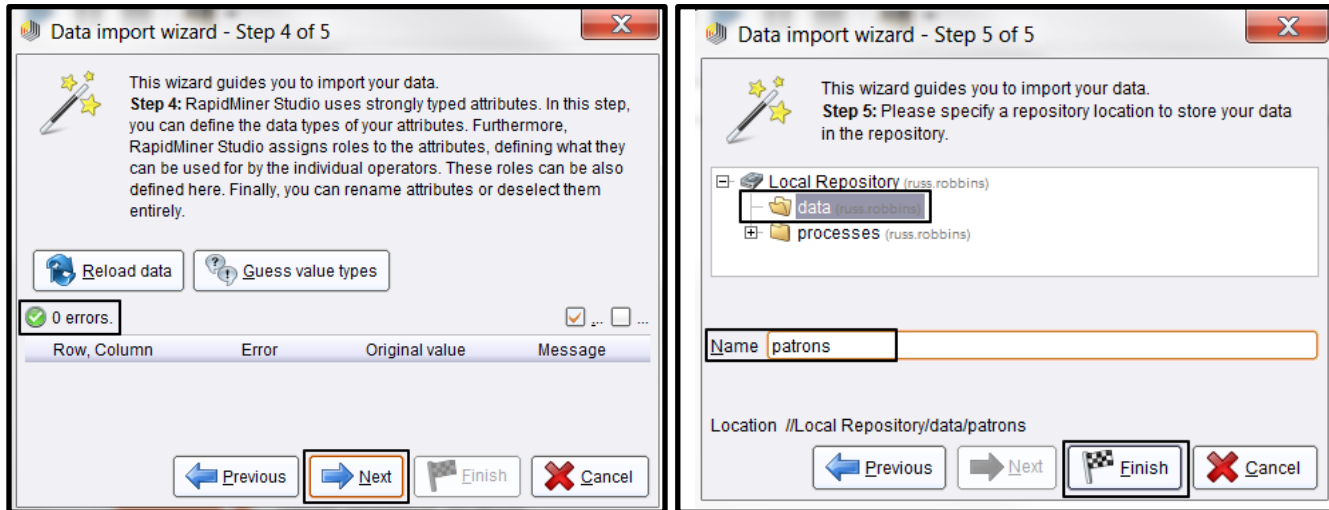
4. Open RapidMiner through your **Start**  button.
5. Import the **Chapter06DataSet.csv** file into RapidMiner by...
  - a. Selecting **File** (File)
  - b. Selecting **Import Data** (Import Data)
  - c. Selecting **Import CSV File...** (Import CSV File)
  - d. Selecting your **Downloads** (Downloads) folder.
  - e. Selecting the **Chapter06DataSet.csv** (Chapter06DataSet.csv) file.
  - f. Selecting **Next** (Next).




6. Continue importing the **Chapter06DataSet.csv** file into RapidMiner by...
  - a. Assuring that ☒ **Skip Comm...** (Skip Comments) is checked.
  - b. Assuring that the radio button for ☒ **Comma** (Comma) is selected.
  - c. Assuring that ☒ **Use Quotes** (Use Quotes) is checked.
  - d. Selecting **Next** (Next) on the Step 2 of 5 screen.
  - e. Selecting **Next** (Next) on the Step 3 of 5 screen. **Please proceed to next page.**

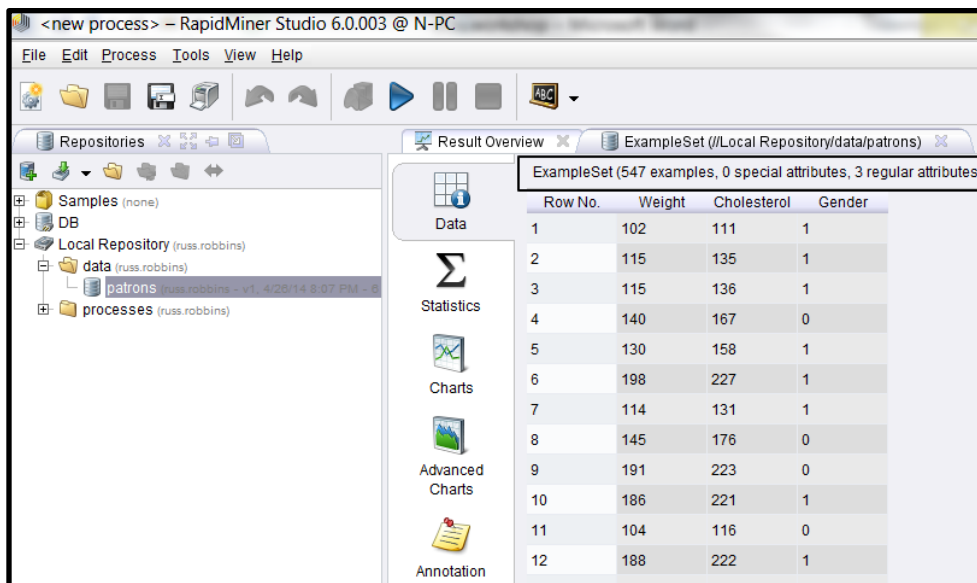


- f. Assuring that  0 errors. (0 errors) have occurred.
- g. Selecting  (Next) on the Step 3 of 5 screen.
- h. Selecting the  data (russ.robbins) (data) folder.
- i. Typing **patrons** in the  (name) field.
- j. Selecting  (Finish).



7. You should see a screen similar to the screen below. If you do not, please start at step 1 again. (No worries.)

- a. Assure you have imported  ExampleSet (547 examples, 0 special attributes, 3 regular attributes) (547 examples, 0 special attributes, 3 regular attributes).



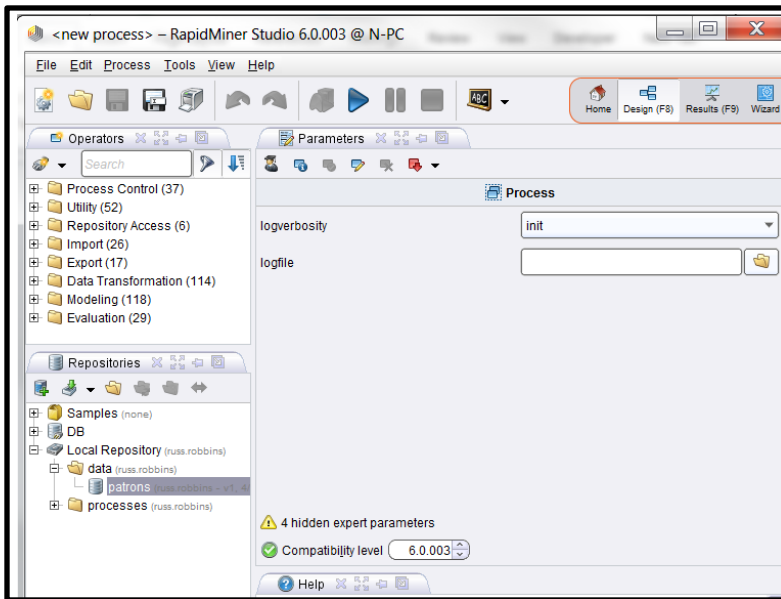
Please proceed to next page.

8. At this point we can use RapidMiner to do "k-means" clustering on our data. Please choose the



(Design View) from the top right of your screen so that we can design a process that will create clusters from this data.

a. You should see a screen similar to the one below.



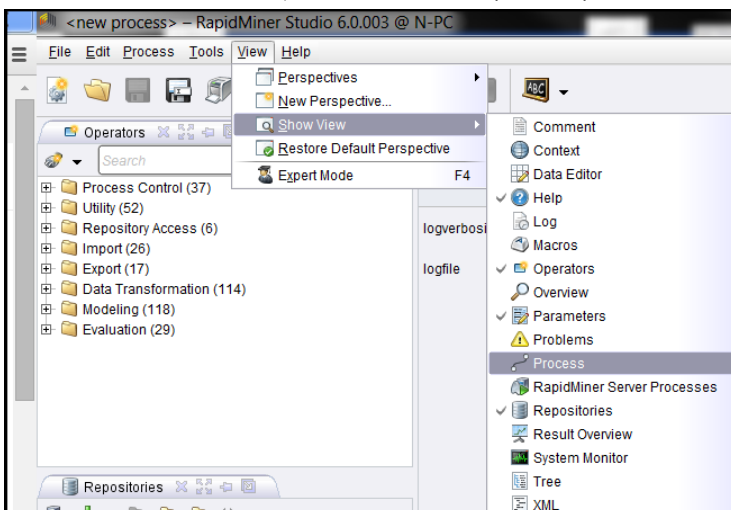
9. K-means clustering is simply a mathematical algorithm that compares records of data by comparing how "close" each pair of records are and then creates groups of records. In our case, you are trying to identify a subset of individuals you can focus on in terms of offering each person in the group incentives and the opportunity to participate in a "heart healthy" program. So for example, a person with a cholesterol level of 215 who is male and weighs 190 and is male may be placed in the same group (or cluster) as another person who has a cholesterol level of 210, weighs 198, and is male. If you are interested, a quick video that explains k-means clustering is here:

10. At this point you want to add a new "window" to your screen so that you can begin designing the process that will perform the k-means clustering algorithm.

a. Select **View** (View) from the top menu.

b. On the submenu, select **Show View** (Show View).

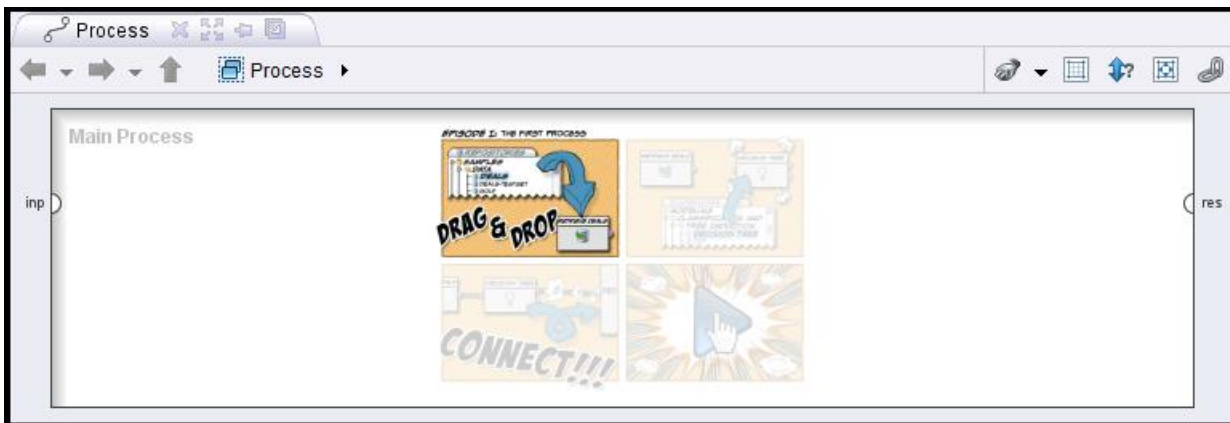
c. On the next submenu, select **Process** (Process).





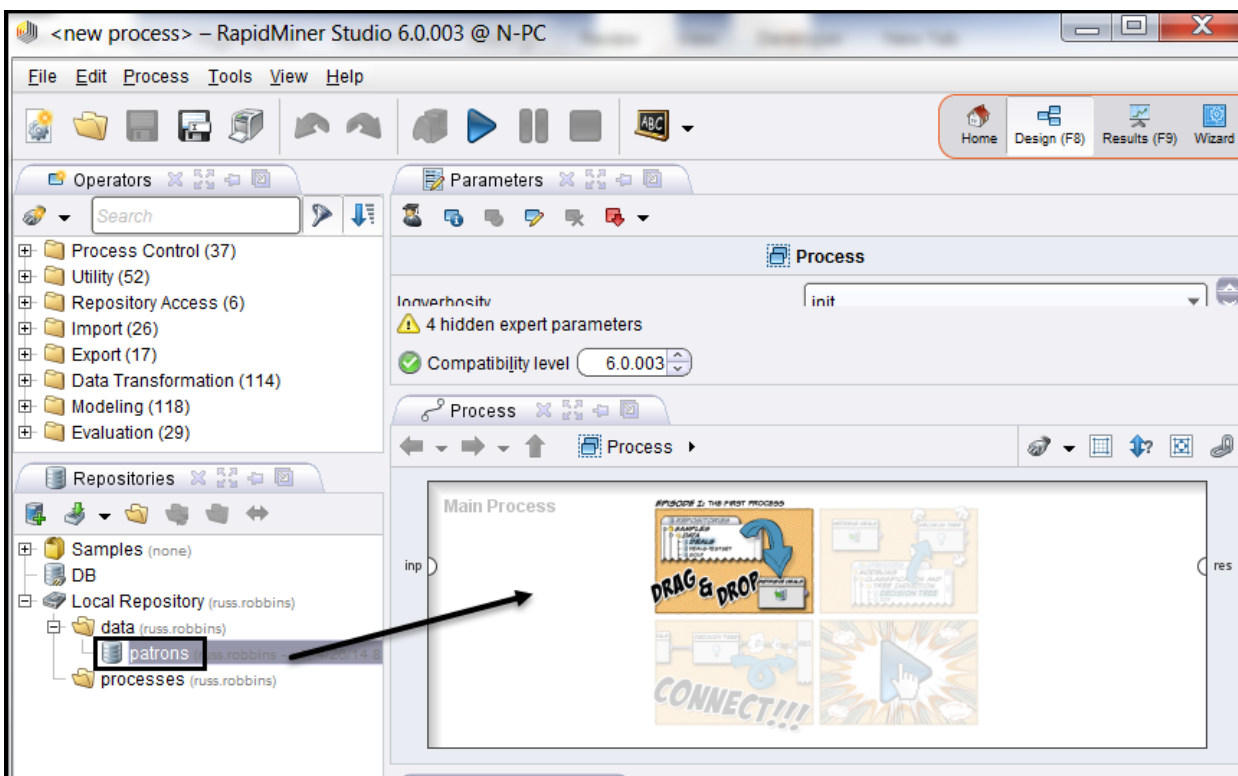
Please proceed to next page.

11. Your screen should now have the following Process window within it.

a. This window is known as the Process Perspective.



12. Select the data that you have imported  (patrons) and "pull" this data into the  (Process) perspective. By placing the imported data into the process you are indicating that the software should retrieve and use this data.

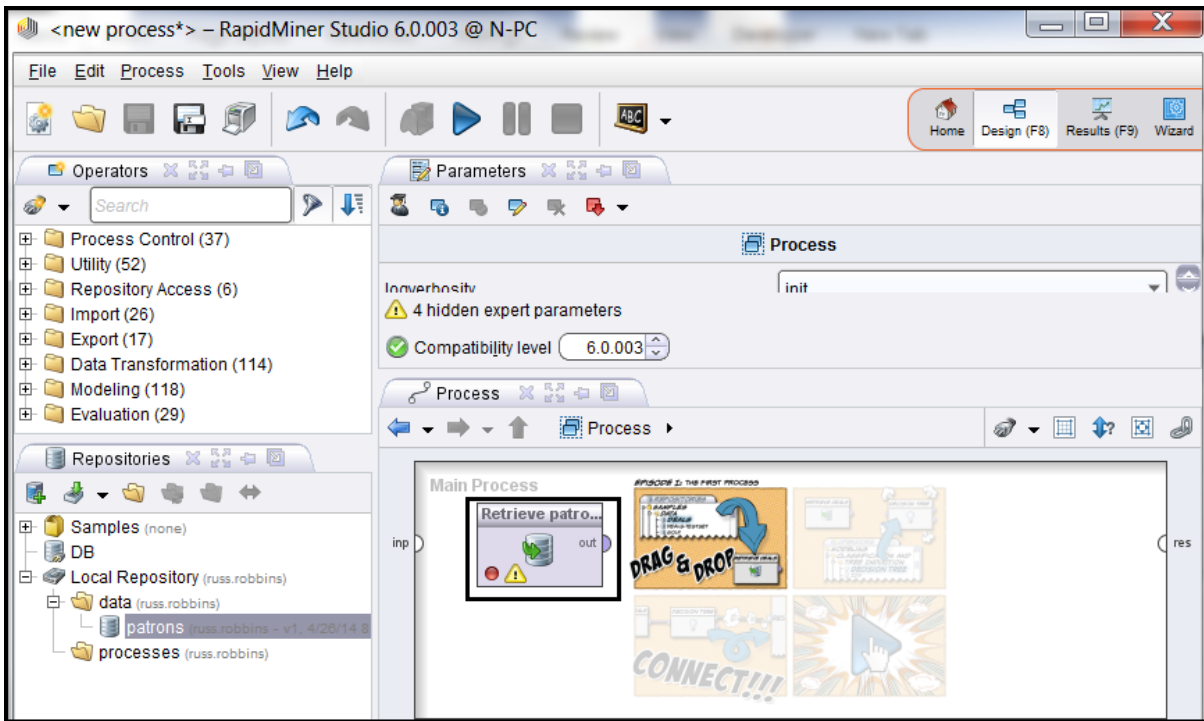


Please proceed to next page.

13. The result should be as shown below.

b. Note that the square box that now in the Process perspective has the name "Retrieve patrons".

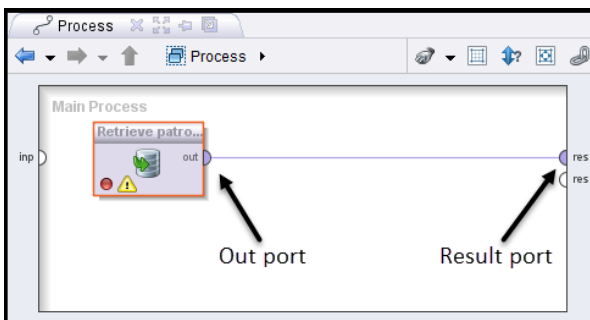
c. Note that this square box is referred to as an "operator."



14. On the right upper side of the "Retrieve patrons" operator, you will see a half-circle that is purple. This is the "out" "port" of the operator.

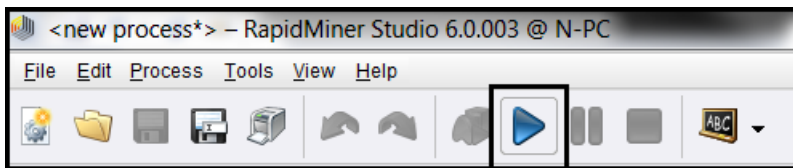
15. On the right side of the Process perspective, you will see another half circle that is purple, and which is named "res". This is a "result" port. (A "port" is just a place where data goes out or in, just like an ocean port.)

16. Select the out port and drag your mouse over to the res port and then release your mouse. The result should be what you see below.

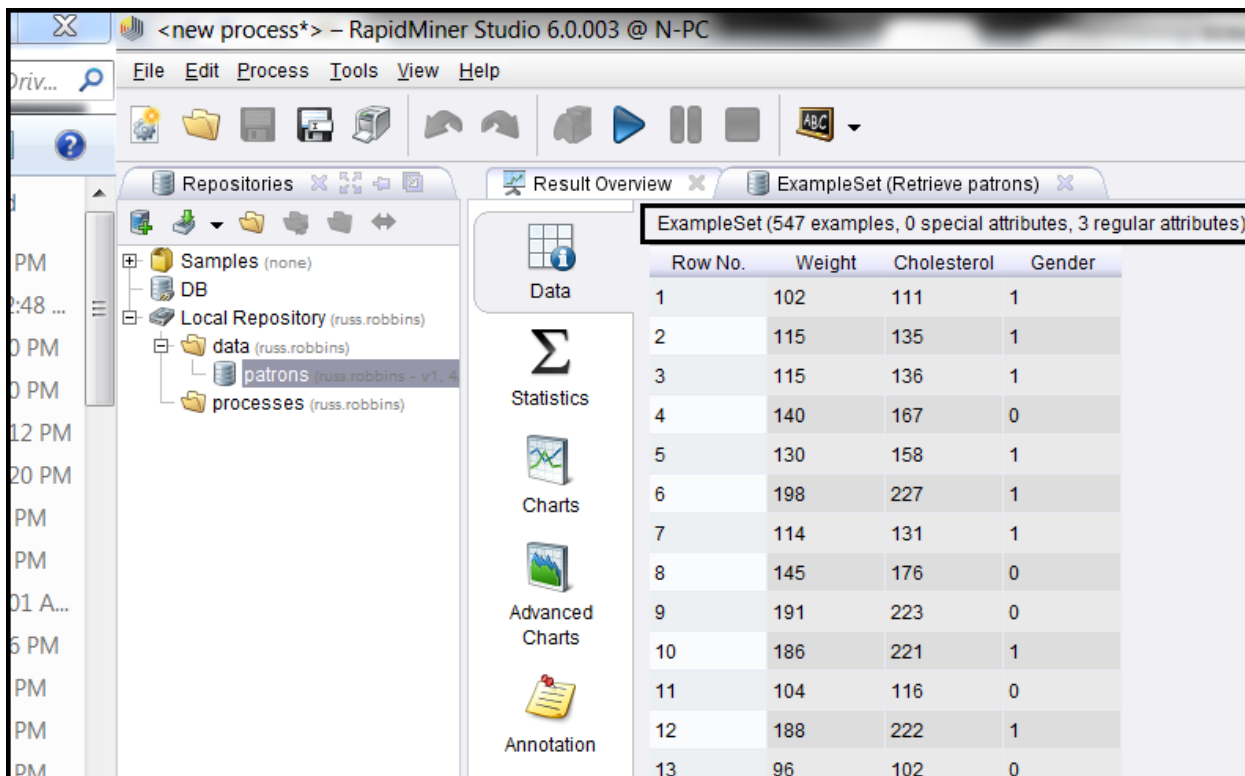


Please proceed to next page.

17. Click the large blue play button at the top of your screen to run the model.



18. Your results should look like the following screenshot. As before, confirm 547 examples, 0 special attributes, and 3 regular attributes.



19. Select Design View.

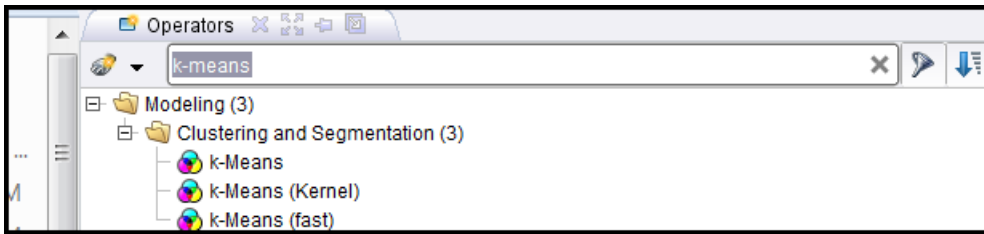


20. In the Operators Search Box in the top left of your screen, type **k-means** (k-means). (Be sure to include the hyphen.)

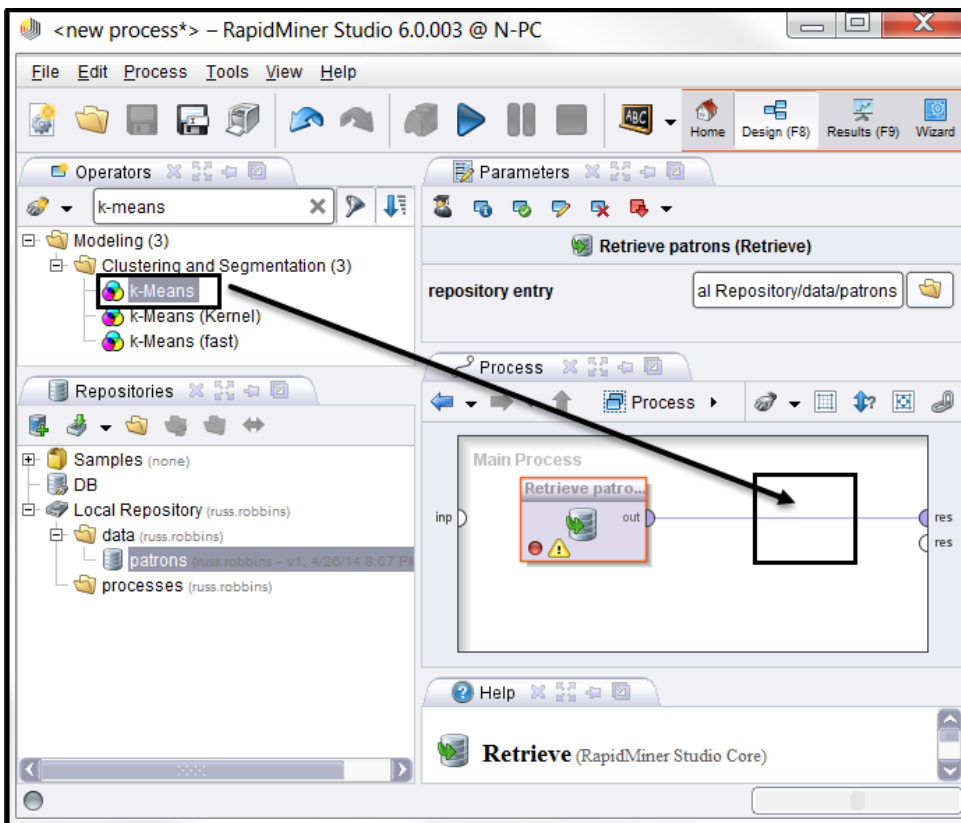


21. After typing "k-means" you should see the following. **Please proceed to next page.**

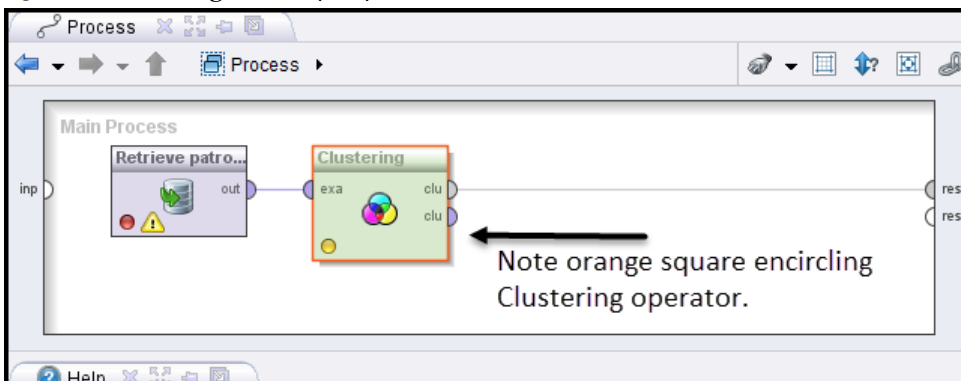




22. Select the **k-Means** (k-means) algorithm operator and "pull" it onto the purple line connecting the Retrieve patrons out operator and the result operator on the right side of the Process perspective.
- Do not release your mouse button until the purple line between the out port and the res port becomes **bold purple**.
  - If your k-means operator is surrounded with an orange-lined square, then you know that you have placed the operator correctly.



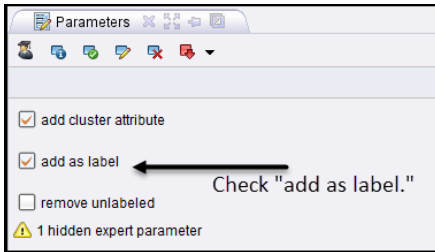
23. Your resulting Process perspective should look like the screenshot below.



Please proceed to next page.

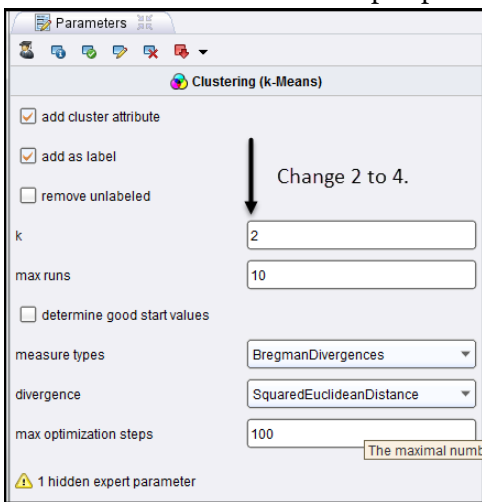


24. If "add as label" is not checked, please do this.

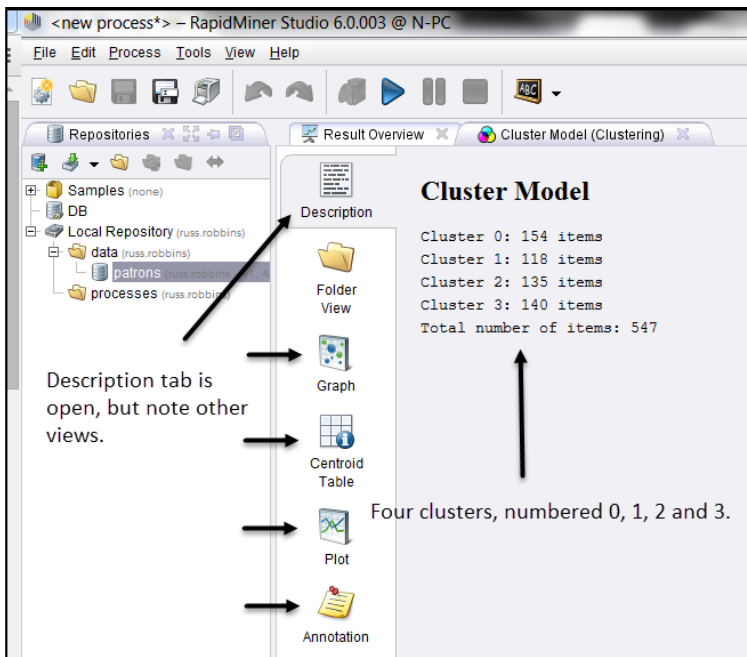


25. At this point you could run the "k-means" algorithm that you have "programmed" into RapidMiner. However, as the program director, you have decided you want four "clusters" or groupings of patrons, instead of what RapidMiner does as a default, which is two. So, if necessary, maximize your

Parameters perspective, by clicking the Parameters location noted that is next to the X. You should see a perspective (a window) that looks like what is below.



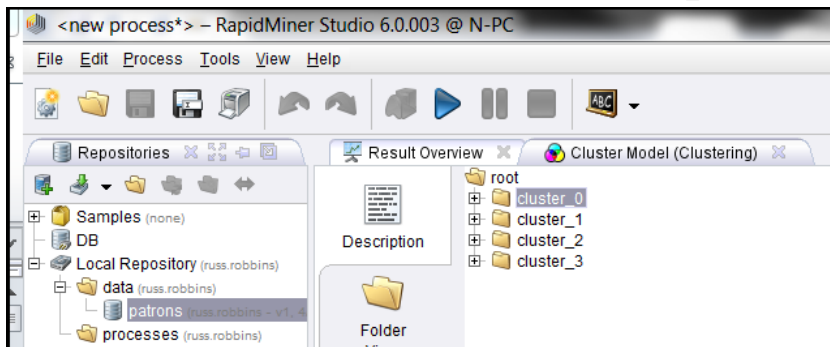
26. Run the k-means clustering algorithm on the data by now pressing the (Play) button. Note that four clusters have been created. Note that there are different "views" of the results. **Please proceed to next page.**



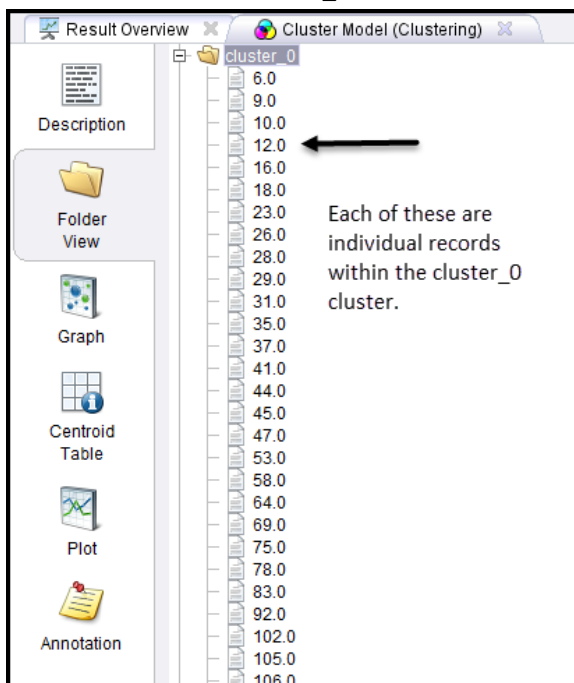
27. Select the folder view.

a. Note there are four clusters (cluster\_0, cluster\_1, cluster\_2, and cluster\_3).

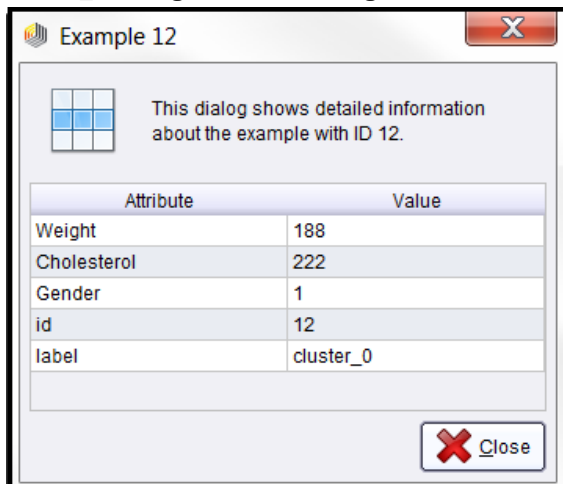
b. Note the names of the clusters include the underscore ( \_ ) character.



28. Double -click on cluster\_0. You should see a view similar to below.



29. Now use your mouse to click on record 12. This screen represents one record that has been classified into cluster\_0, using the k-means algorithm. Note that 12 is the "id" of the record and not data from the CSV file.



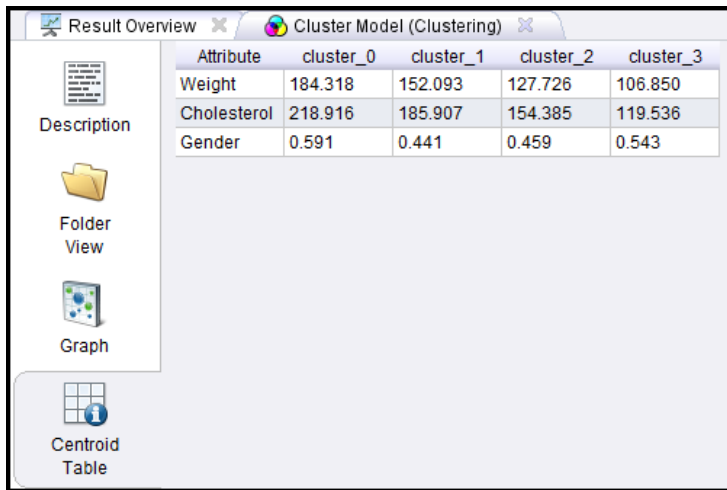
Please proceed to next page.

30. Select the Centroid View.

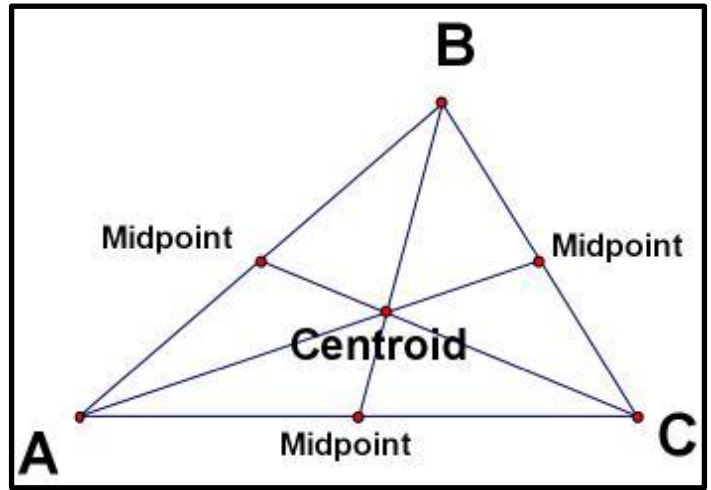
a. Note the centroid weights, cholesterol, and gender.

b. A centroid is not identical to the average but in this example, they are identical.

c. I have placed a graphic of a centroid to the right of the centroid table to help you see that a centroid is about geographical space while an average is a computation that can be made without any consideration of space.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	184.318	152.093	127.726	106.850
Cholesterol	218.916	185.907	154.385	119.536
Gender	0.591	0.441	0.459	0.543




31. Which cluster do you think you will want to target with your "heart healthy" programs? Why?

32. Which cluster would you target second?

33. What does the gender centroid of .591 for cluster\_0 mean?

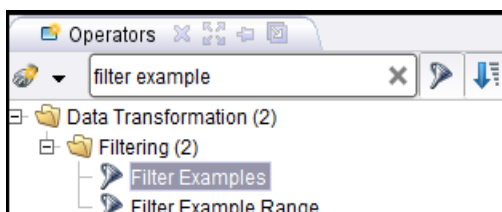
34. After answering these questions, feel free to review the Graph and Plot views.

35. To get more statistical information about the clusters, we can now add an operator to our process. The operator we are going to add is "Filter Examples."

36. Select Design View.  Then type "Filter Examples" into the Operators search field.

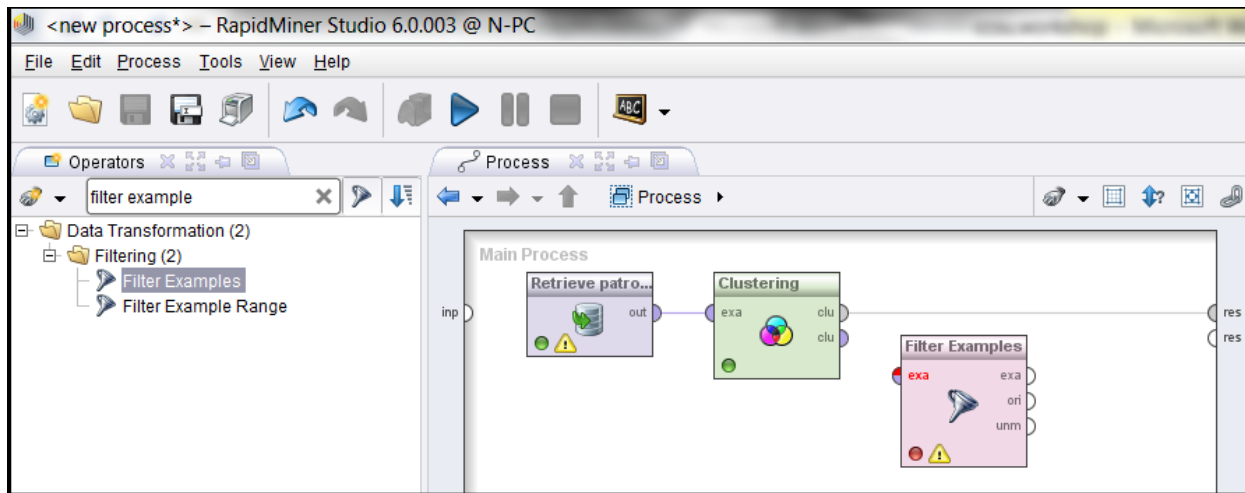


37. The result should be as shown below.

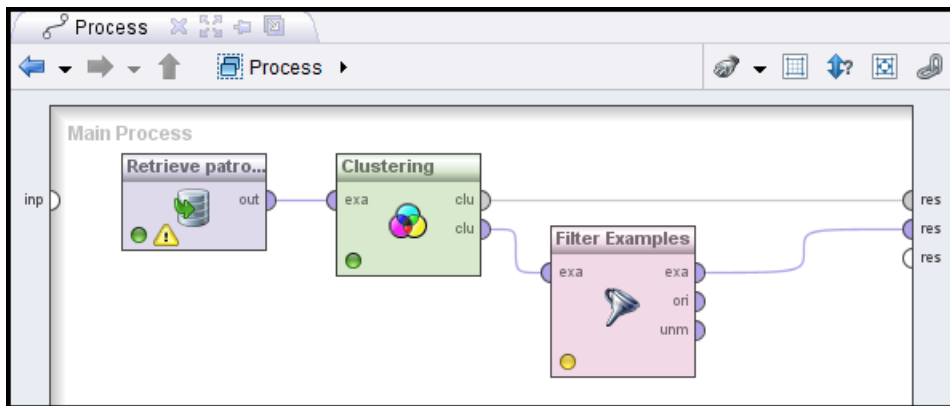


Please proceed to next page.

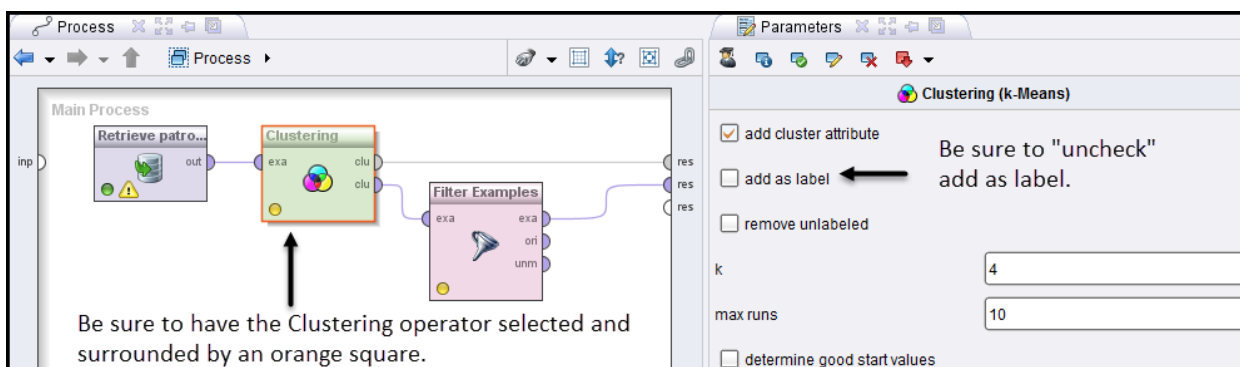
38. Then "pull" the Filter Examples operator into the Process perspective, so that your Filter Examples operator is as below.



39. Now connect the clu (clustering) port on the right side of the Clustering operator with the exa (examples) port on the left of the Filter Examples operator and the exa (examples) port on the right of the Filter Examples port to the res (results) port on the right of the Process perspective. The result should be as shown below.

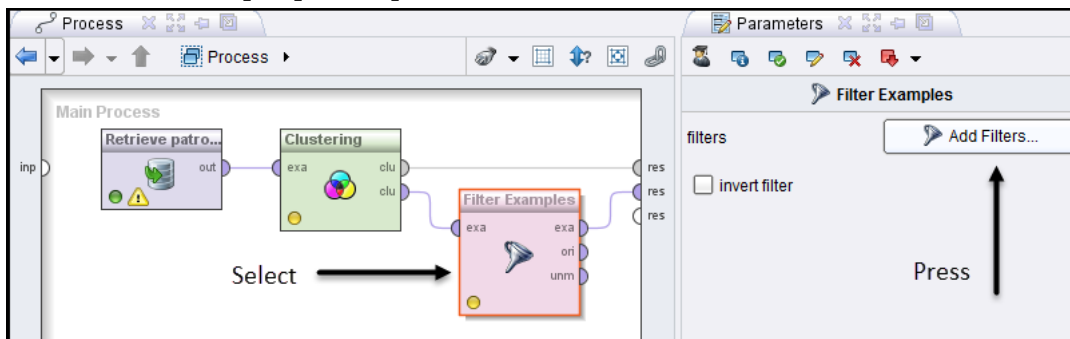


40. Before we use the new process, we have to set parameters on the Clustering and the Filter Examples operators.
- Select the Clustering operator in the Process perspective.
  - In the Parameters perspective, be sure to uncheck the "add as label" checkbox. **Please proceed to next page.**




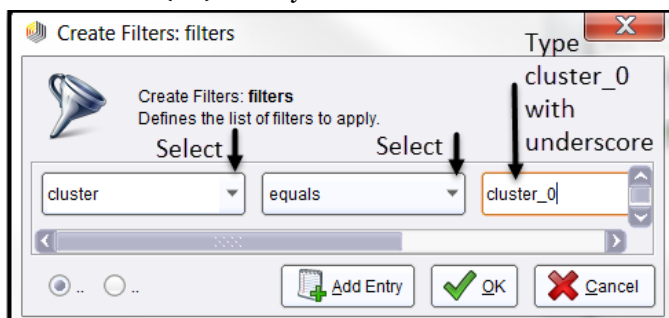
41. Now let's set parameters on Filter Examples operator.

  - a. Select the Filter Examples operator in the Process perspective.
  - b. In the Parameters perspective, press the "Add Filters..." button.

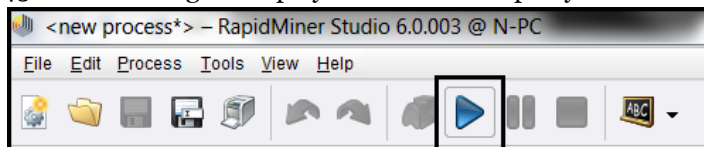



42. Fill in the Create Filters dialog box as below. Be sure to type "cluster\_0" exactly as below.

Press  (OK) when you are finished.



43. Click the large blue play button at the top of your screen to run the model.



44. Switch to the Statistics view.  Note this view which summarizes weight, cholesterol, and gender for cluster\_0. Note that most of the individuals in cluster\_0 are men, since the "average" gender is 0.591. The average is 0.591 because there are more records with 1 in the gender field than there are 0. If you remember, 1 refers to males.

Name	Type	Miss.	Statistics			
id	Integer	0	Min	Max	Average	Deviation
id			6	543	271.727	157.396
cluster	Nominal	0	Least	Most	Values	
cluster			cluster_3 (0)	cluster_0 (154)	cluster_0	
Weight	Integer	0	Min	Max	Average	Deviation
Weight			167	203	184.318	9.809
Cholesterol	Integer	0	Min	Max	Average	Deviation
Cholesterol			204	235	218.916	8.191
Gender	Integer	0	Min	Max	Average	Deviation
Gender			0	1	0.591	0.493

Please proceed to next page.

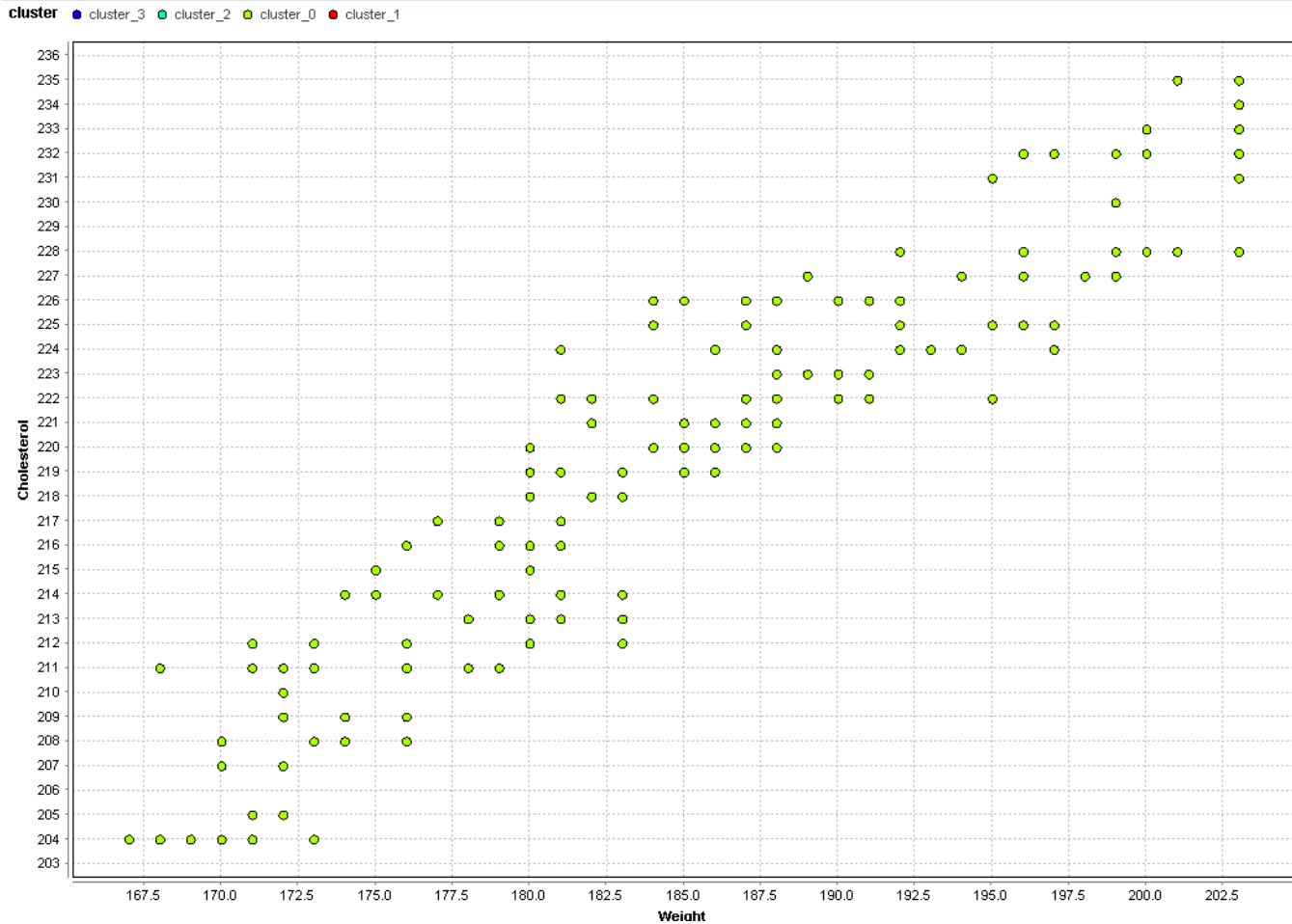


45. Switch to the (Charts) view of the cluster\_0 data.

a. Note the only data on the chart is for records summarizing individuals in cluster\_0.

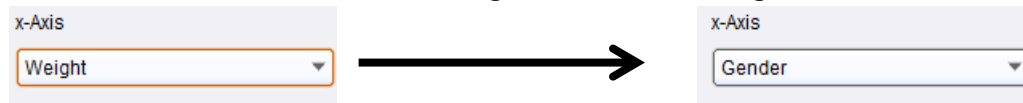
b. Note the linear relationship between weight and cholesterol. (As weight goes up, so does cholesterol.)

(You have an upcoming meeting with clinicians to begin developing the "heart healthy" programs. Perhaps you could ask them about this possible relationship between weight and cholesterol, and how this may figure into the development of the heart healthy programs.)



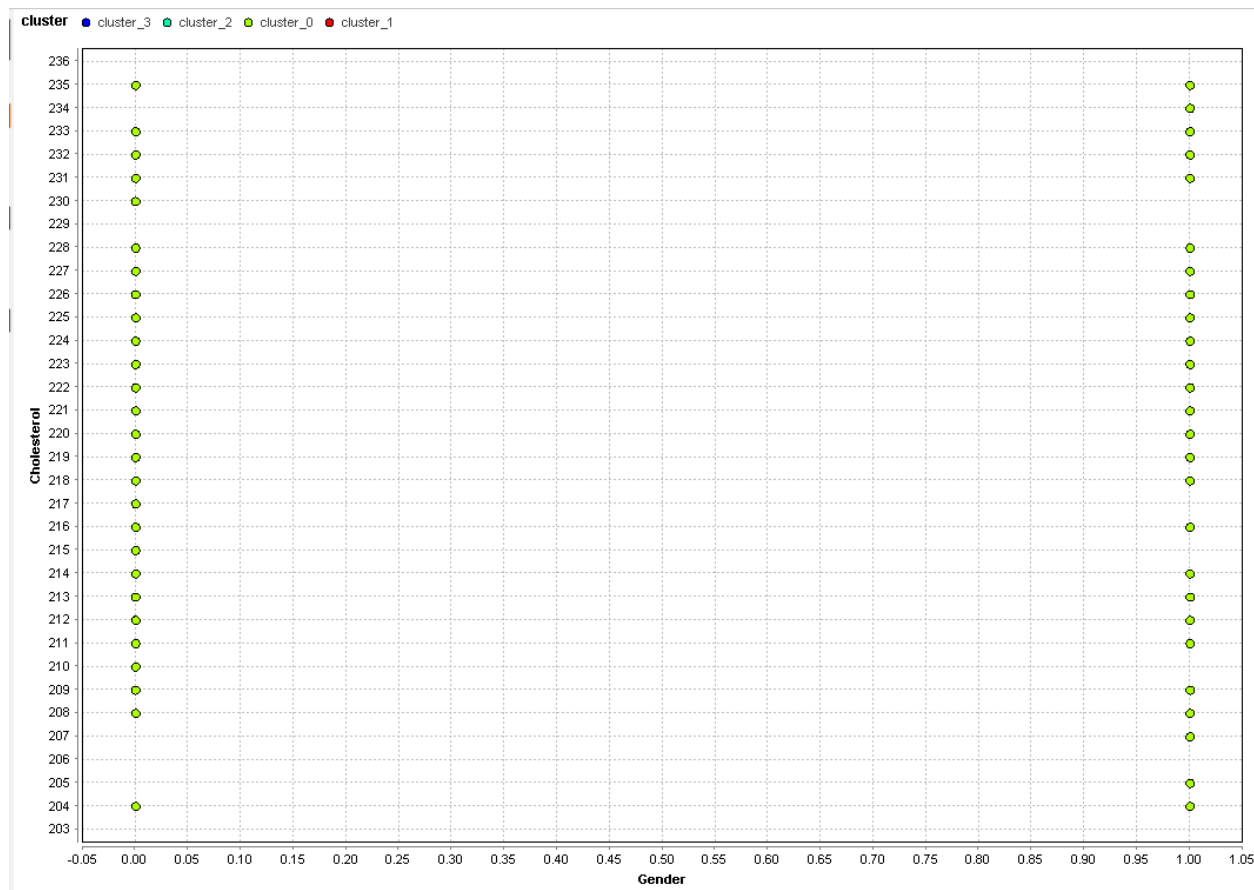
Please proceed to next page.

46. Remain in the Charts View and change the x-axis from Weight to Gender.



47. The chart you should now see should be as below.

a. Note that it appears that both genders (male, 1) and (female, 0) in the cluster\_0 seem to have cholesterol in the above 200 range, and, at least by looking at this chart, it doesn't seem like a person's gender can "explain" why a person might have high (over 200) cholesterol.



## Conclusion:

By using the k-means clustering algorithm, and knowing that those with the highest levels of cholesterol and weight are the most susceptible to heart disease, you have identified some patrons with the highest risk. With this information, you will be able to return to your company's database and identify the names and contact information for these individuals insured by your company. Using their contact information, you can begin to reach out to these individuals so that they can consider whether they want to be involved in your upcoming "heart healthy" program.

Congratulations! on completing this "hands on" demonstration of the use of the k-Means clustering algorithm.