**BAN 540 Lab 3 – Linear Regression (30 points)**

*Adapted from “Data Mining for the Masses” Chapter 8*

Please follow the instructions carefully to finish lab assignment 3. In this assignment, you will be asked to make **6** **screenshots** and paste them to “*BAN540 Lab3 Submission YourLastName.docx*” file. Once you are done with these 6 screenshots, please submit the word file “BAN540 Lab3 Submission YourLastName.docx” (with YourLastName in the file name) to Canvas via the submission link.

**Note:** “YourLastName” in this document refers to your own last name. Don’t literally type in “YourLastName” to name any of your repositories/processes/dataset files.

**CONTEXT AND PERSPECTIVE**

Sarah, the regional sales manager from the lab assignment 2 example (Chapter 4 of *“Data Mining for the Masses”* textbook), is back for more help. Business is booming, her sales team is signing up thousands of new clients, and she wants to be sure the company will be able to meet this new level of demand. She was so pleased with our assistance in finding **correlations** in her data (in lab assignment 2), she now is hoping we can help her make some **predictions** as well. She knows that there is some correlation between the attributes in her data set (like temperature, insulation, and occupant ages), and she’s now wondering if she can use the data set from lab assignment 2 to predict heating oil usage for new customers. You see, these new customers haven’t begun consuming heating oil yet, as there are a lot of them (42,650 to be exact), and she wants to know how much oil she needs to expect to keep in stock in order to meet the demand of these new customers. Can she use data mining to examine household attributes and known past consumption quantities to anticipate and meet her new customers’ needs?

**LEARNING OBJECTIVES**

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

* Explain what linear regression is, how it is used and the benefits of using it.
* Recognize the necessary format for data in order to perform predictive linear regression.
* Explain the basic algebraic formula for calculating linear regression.
* Develop a linear regression data mining model in RapidMiner using a training data set.
* Interpret the model’s coefficients and apply them to a scoring data set in order to deploy the model.

**ORGANIZATIONAL UNDERSTANDING**

Sarah’s new data mining objective is pretty clear: she wants to anticipate demand for a consumable product. We will use a **linear regression** model to help her with her desired predictions. She has data, 1,218 observations from lab assignment 2 data set that give an attribute profile for each home, along with those homes’ annual heating oil consumption. She wants to use this data set as training data to predict the usage that 42,650 new clients will bring to her company. She knows that these new clients’ homes are similar in nature to her existing client base, so the existing customers’ usage behavior should serve as a solid gauge for predicting future usage by new customers.

**DATA UNDERSTANDING**

As a review, our data set from lab assignment 2 (Chapter 4) contains the following attributes:

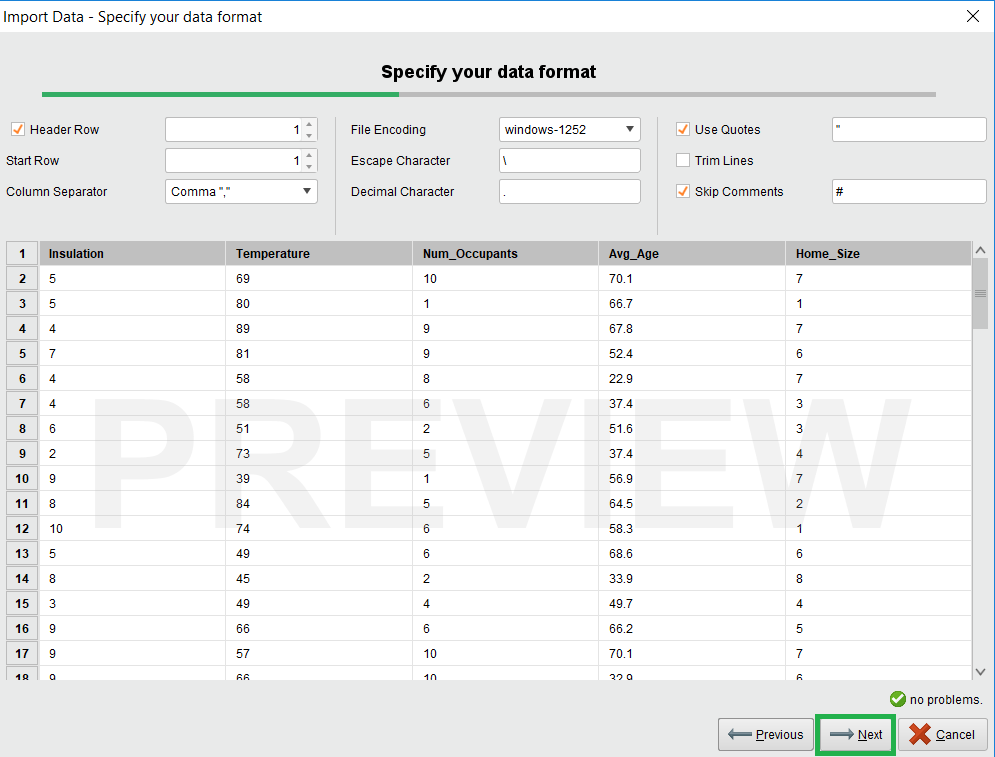
* **Insulation**: This is a density rating, ranging from one to ten, indicating the thickness of each home’s insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation.
* **Temperature**: This is the average outdoor ambient temperature at each home for the most recent year, measured in degrees 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.

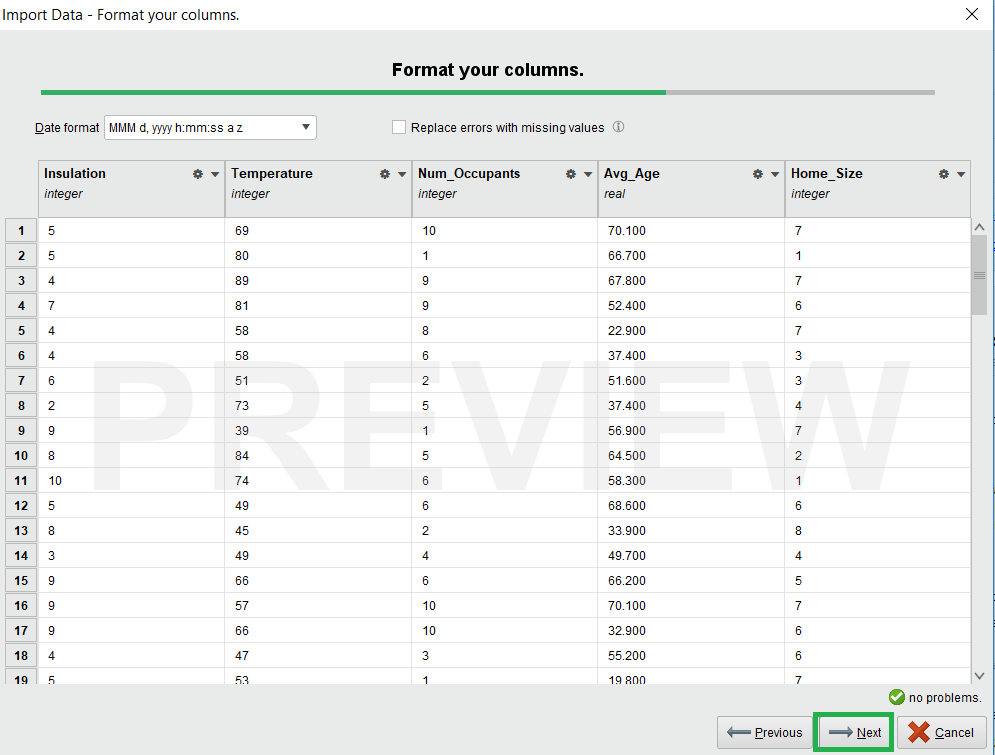
We will use the lab assignment 2 data set (Chapter 4) as our **training data set** in this assignment. Sarah has assembled a separate Comma Separated Values file containing all of these same attributes, **EXCEPT** of course for **Heating\_Oil**, for her 42,650 new clients. She has provided this data set to us to use as the scoring data set in our model.

**DATA PREPARATION**

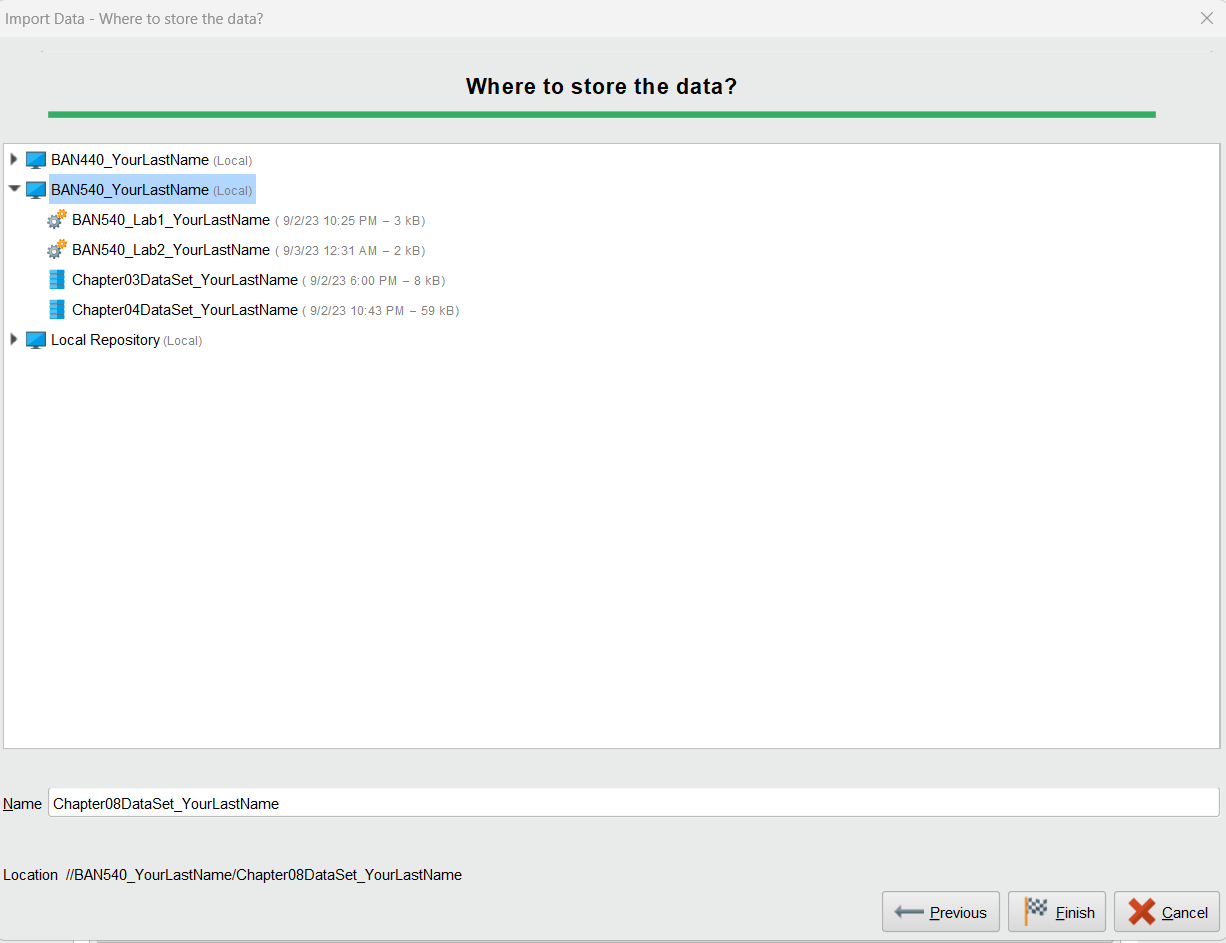
Please download “Chapter04Dataset.csv” and “Chapter08Dataset.csv” from Canvas and save them to your local drive.

1) For Chapter 8 data set, we do not have **Heating\_Oil** attribute as it is unknown for new customers. We are going to predict it using ***linear regression model*** in this assignment. As we already imported “Chapter04Dataset.csv” in the assignment 2, we will now import Chapter 8 data set (by clicking on the **Import Data** button in the ***Repository*** pane of RapidMiner) by following the steps shown in the screenshots below.

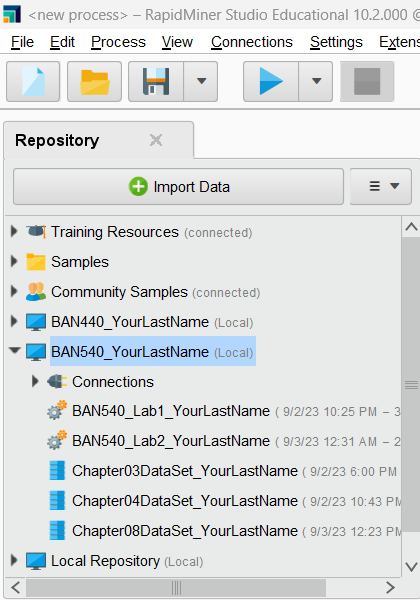




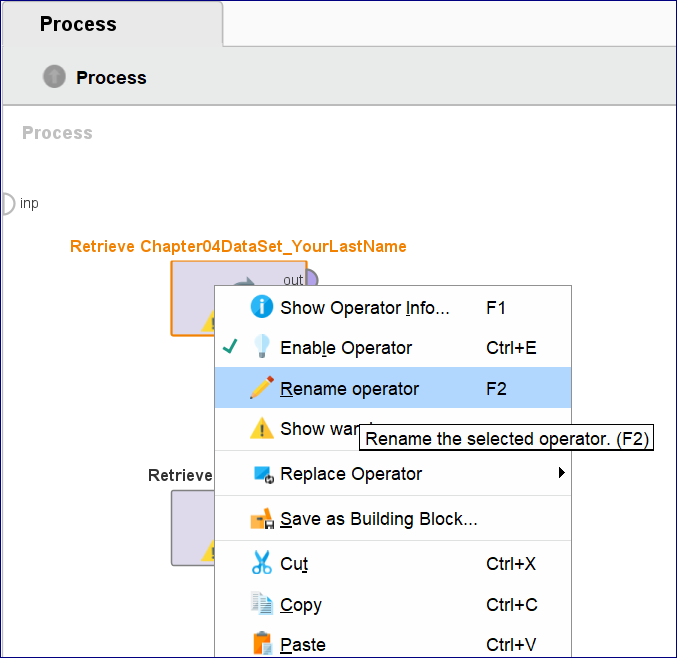
2) The final step is to choose a repository to store the data set in, and to give the data set a name within RapidMiner. As shown in the following screenshot, please store the data set in the repository that you already created for the previous assignments, which is ***BAN540\_YourLastName*(If you do not have this repository, you must recreate it now. Refer to assignment 1 instruction on how to do it)**, and name it as **Chapter08DataSet\_YourLastName**. Then click Finish. ***(Important Note: you MUST name it as Chapter08DataSet\_YourLastName to get credit for this step. Please also make sure to have Chapter04DataSet\_YourLastName which was created in lab 2; if not, please import this data set again).***

******

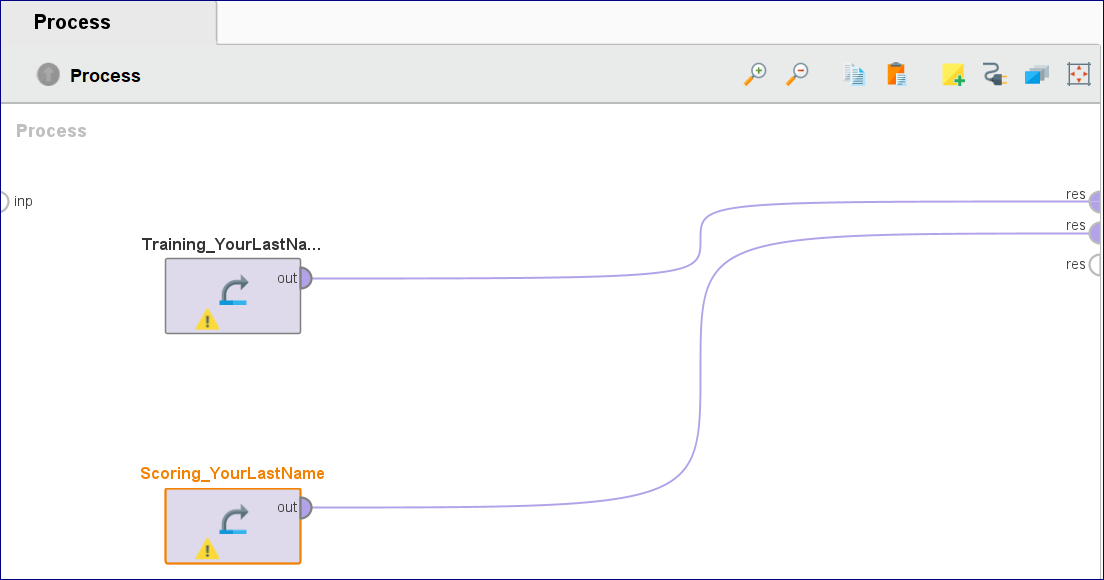
3) Once you have both Chapter 4 and Chapter 8 data sets imported into your RapidMiner data repository, your screen should look like the following screenshot. Please click on “**Design**” button to go to the **Design Perspective**, and make sure in **Repository** pane, you can see ***BAN540\_YourLastName*** with ***Chapter04DataSet\_YourLastName*** and ***Chapter08DataSet\_YourLastName*** in it. All other files are optional.



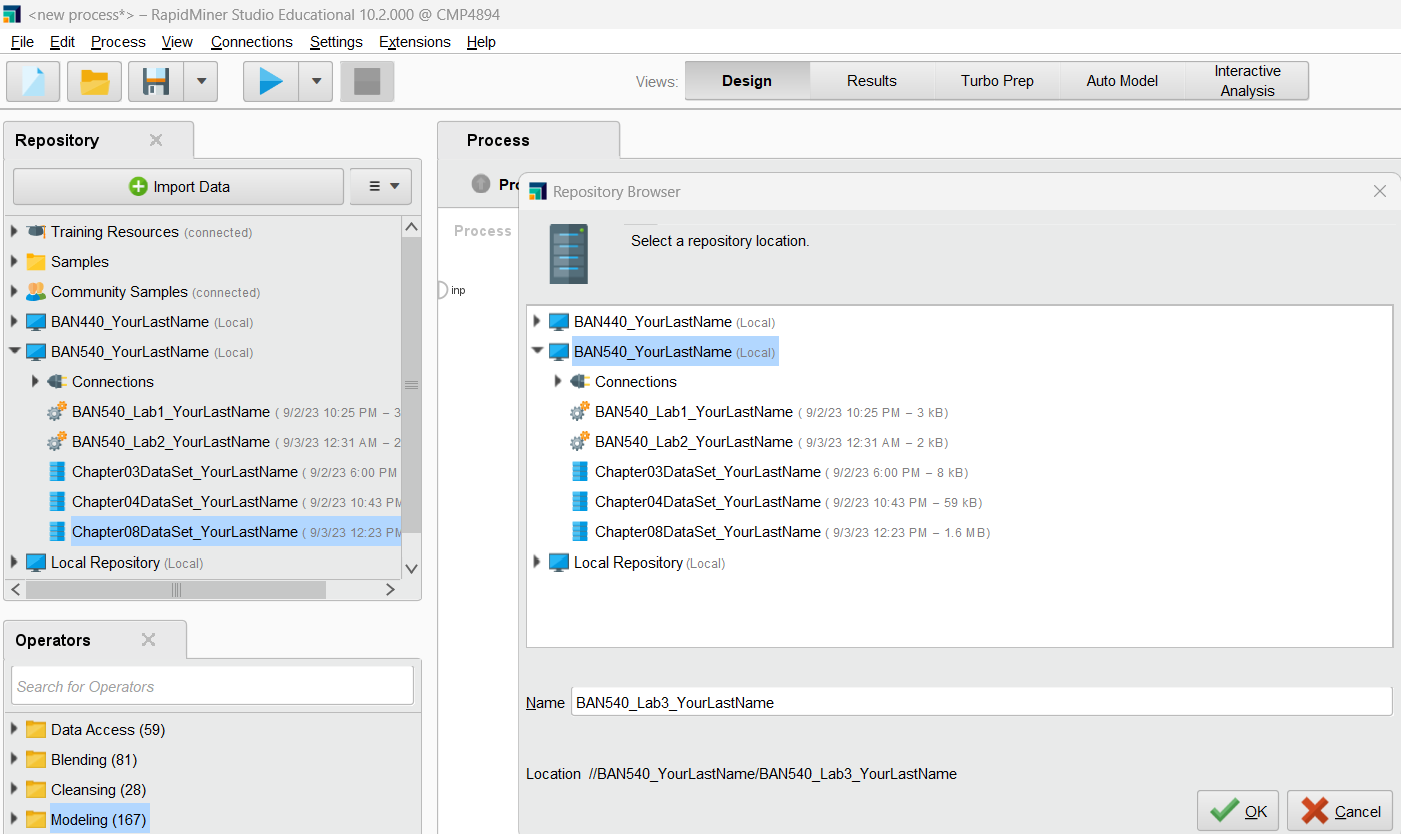
4) Please drag and drop both data sets onto the main **Process** window/pane in RapidMiner. Then, right click on both of the **Retrieve** operators to rename the Chapter 4 data set to ‘***Training\_YourLastName***’, and the Chapter 8 data set to ‘***Scoring\_YourLastName***’. ***(Important Note: you MUST name them correctly in order to get credit for this step.)***



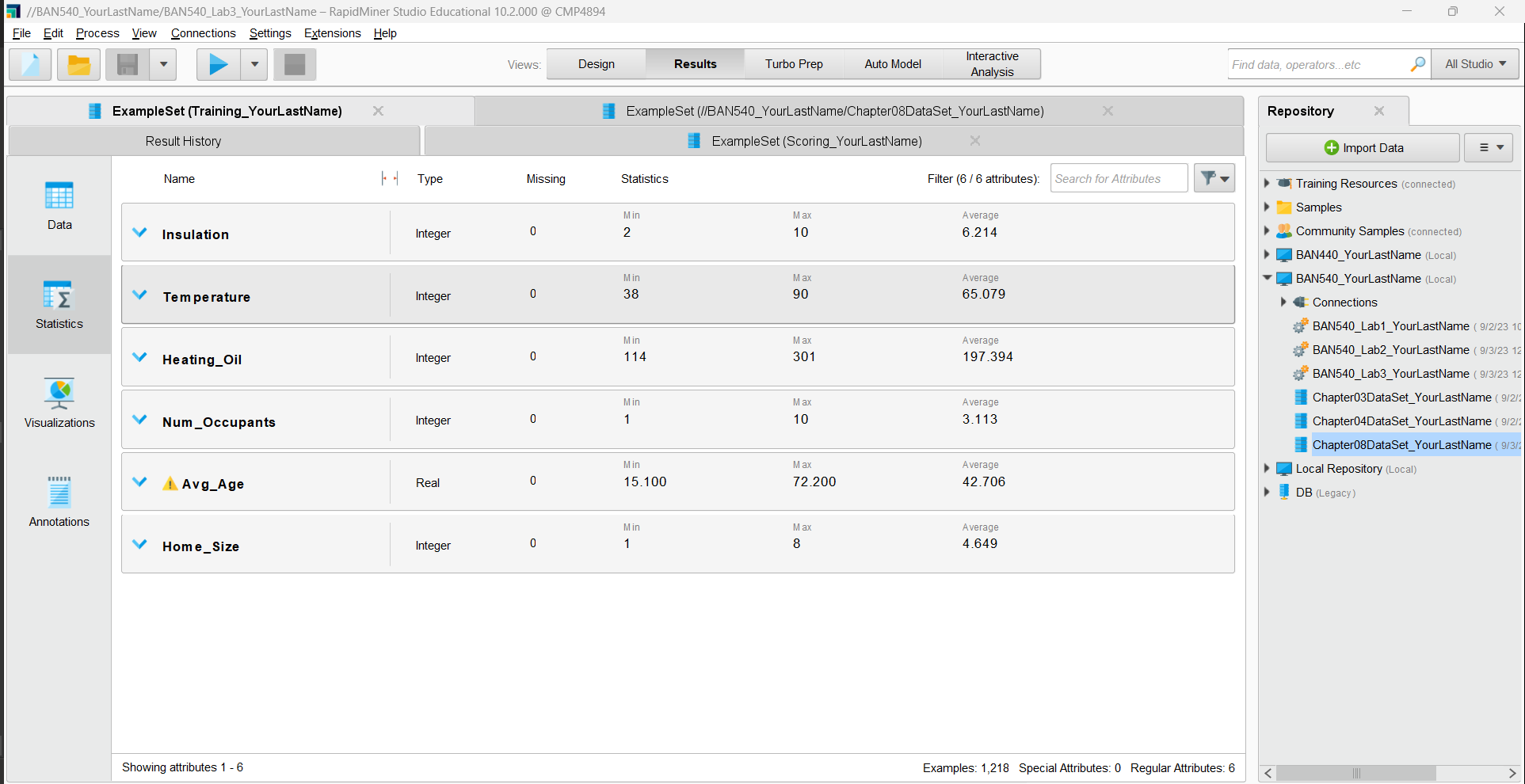
5) Connect both **out** ports to **res** ports as shown in the screenshot below.

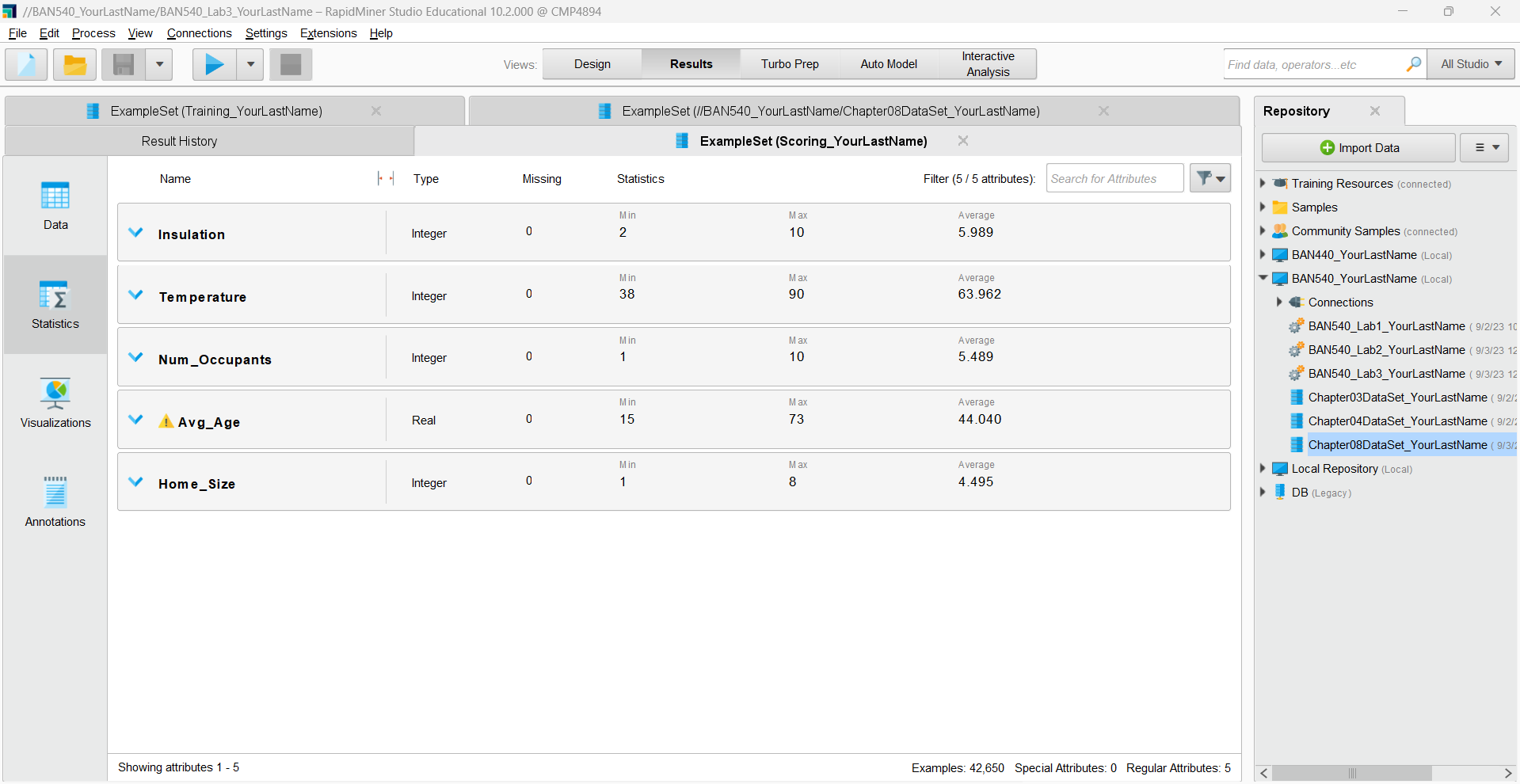


6) Please save the process (by clicking on the **Save** icon on the top left corner) into the repository “**BAN540\_YourLastName” (you should see Your own last name here).** And name your process as **BAN540\_Lab3\_YourLastName.** Then click OK. ***(Important Note: you MUST name it as BAN540\_Lab3\_YourLastName to get credit for this step).***

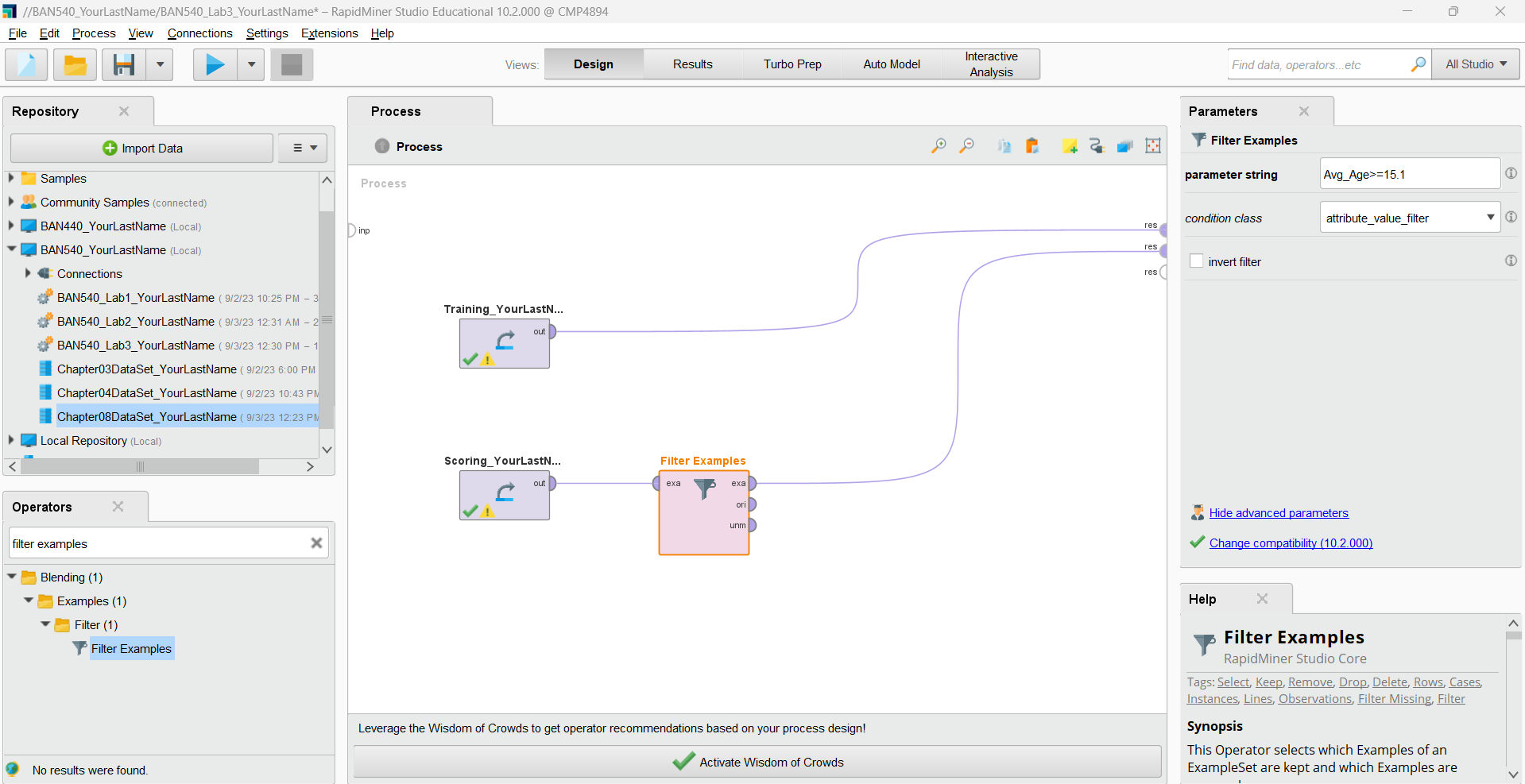


7) Now, please run your model. The following two screenshots show comparisons of the training and scoring data sets. When using linear regression as a predictive model, it is extremely important to remember that the ranges for all attributes in the scoring data **MUST** be within the ranges for the corresponding attributes in the training data. This is because a training data set cannot be relied upon to predict a target attribute for observations whose values fall outside the training data set’s values.

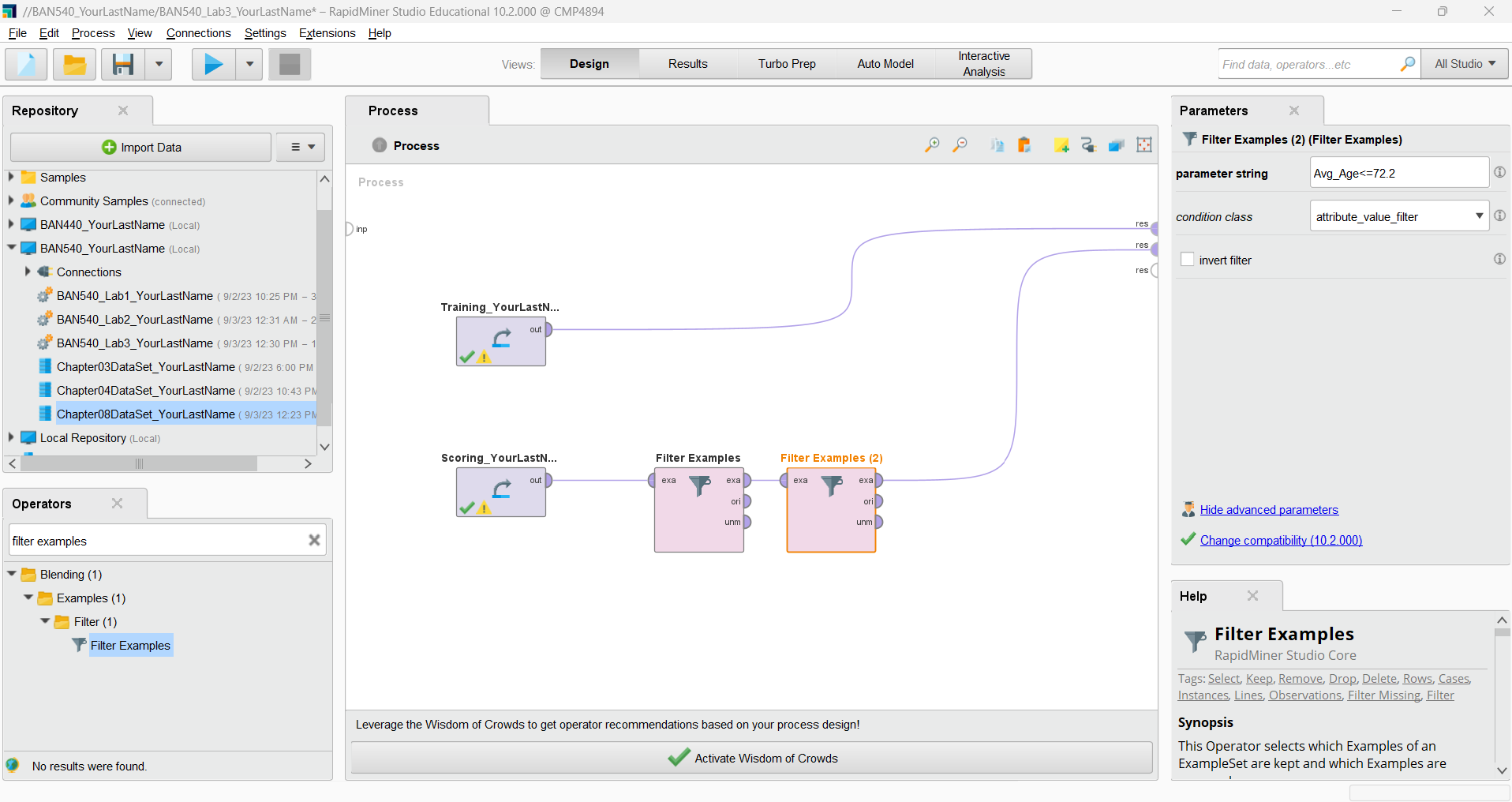




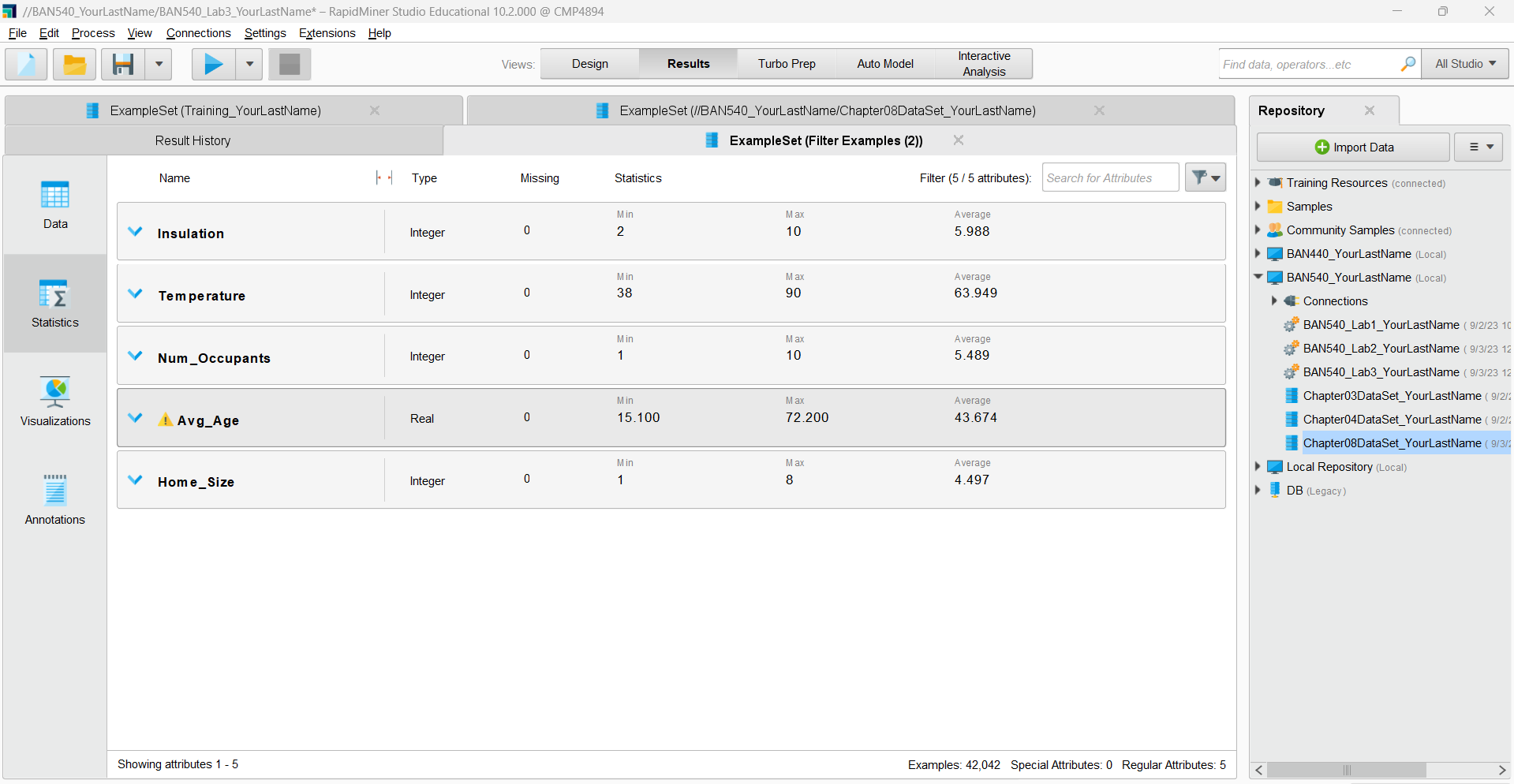
8) We can see the ranges are the same for all attributes except Avg\_Age. In the scoring data set, we have some observations where the Avg\_Age is slightly below the training data set’s lower bound of 15.1, and some observations where the scoring Avg\_Age is slightly above the training set’s upper bound of 72.2. You might think that these values are so close to the training data set’s values that it would not matter if we used our training data set to predict heating oil usage for the homes represented by these observations. While it is likely that such a slight deviation from the range on this attribute would not yield wildly inaccurate results, we **cannot** use linear regression prediction values as evidence to support such an assumption. Thus, we will need to **remove these observations** from our data set. To do it, please add a **Filter Examples** operator. Please search and find it from the Operators Pane. Please drag and put it after the “*Scoring\_YourLastName”* operator (making sure it is connected in the path). While the “Filter Examples” operator is being selected, please go to the **Parameters** Pane to choose ‘***attribute\_value\_filter****’* in the “condition class” drop-down list, and then specify ***Avg\_Age>=15.1*** in the “parameter string” textbox.



9) Please add another **Filter Examples** operator and put it right after the one you just created (see the screenshot below). Still, choose ***attribute\_value\_filter*** from the “condition class” drop-down list and specify ***Avg\_Age<=72.2*** in the “parameter string” textbox.

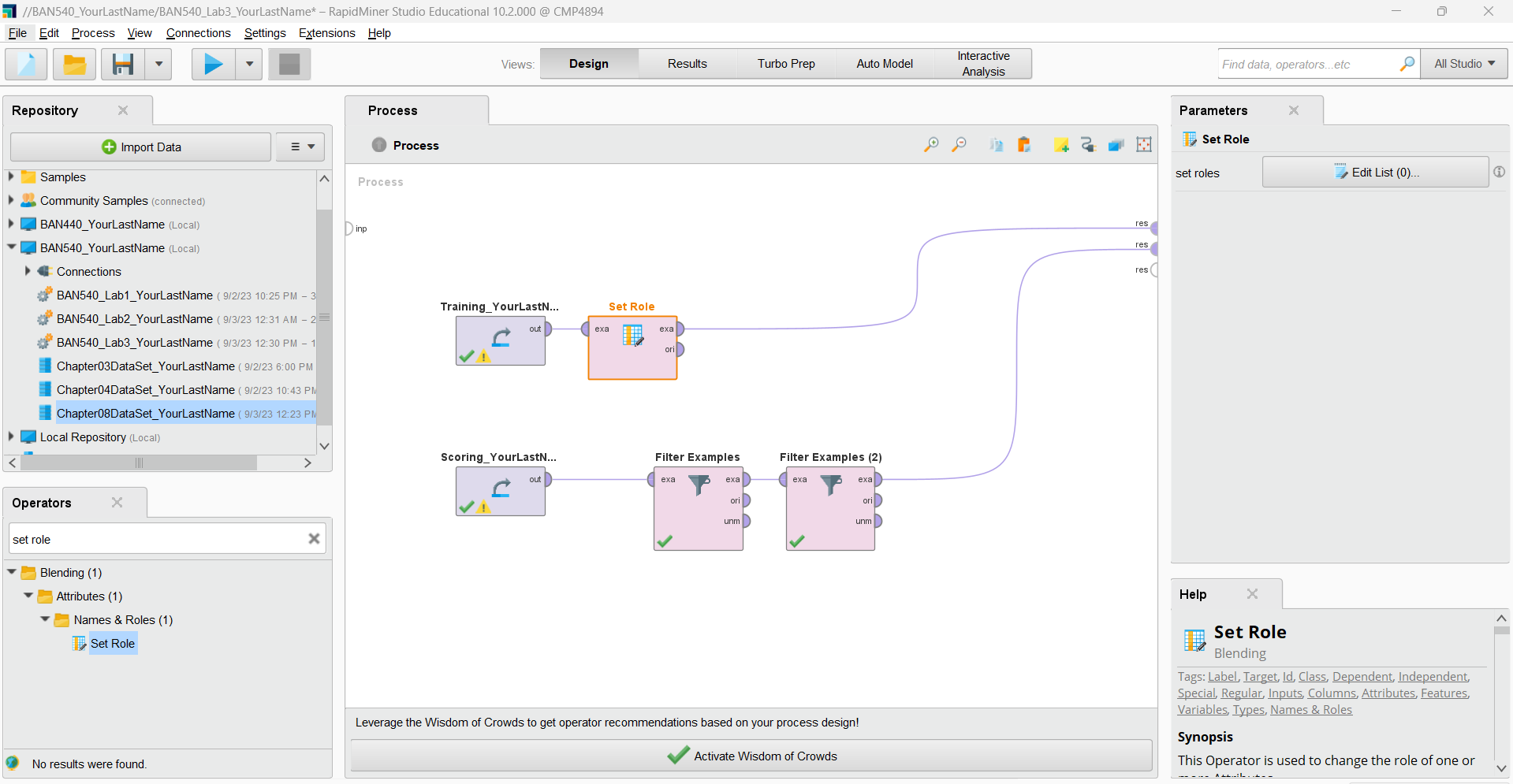


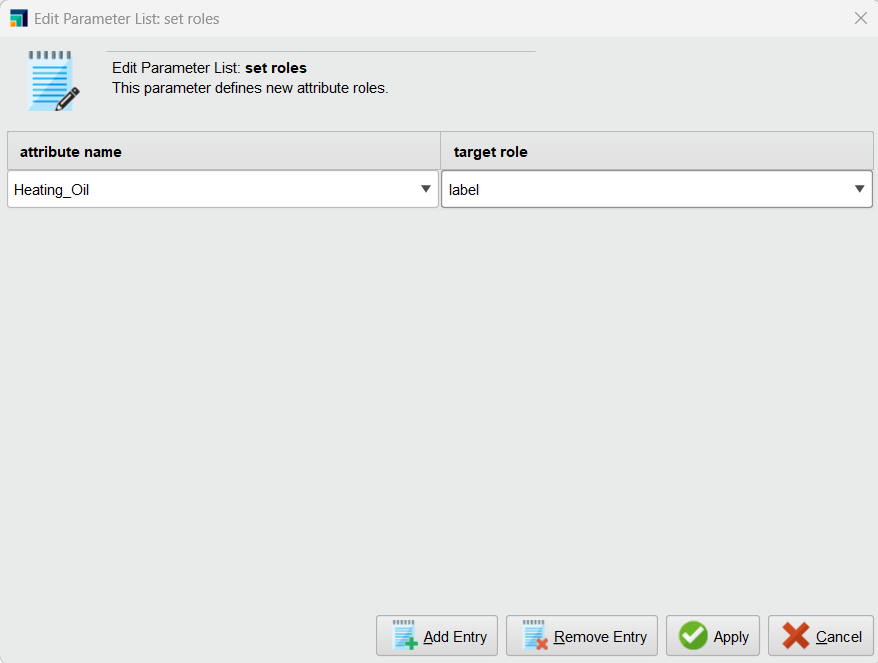
Now run your model, and you should see **42,042** observations remaining. Please check the ranges again to ensure that none of the scoring attributes now have ranges outside those of the training attributes. **Please make a screenshot now and replace my screenshot #1 with yours in the submission file (named as “BAN540 Lab 3 Submission YourLastName.docx”). Please make sure your screenshot shows your own last name in the related items we have added so far (see the red boxes in the Repository pane on the right-hand side of the Results view.)**



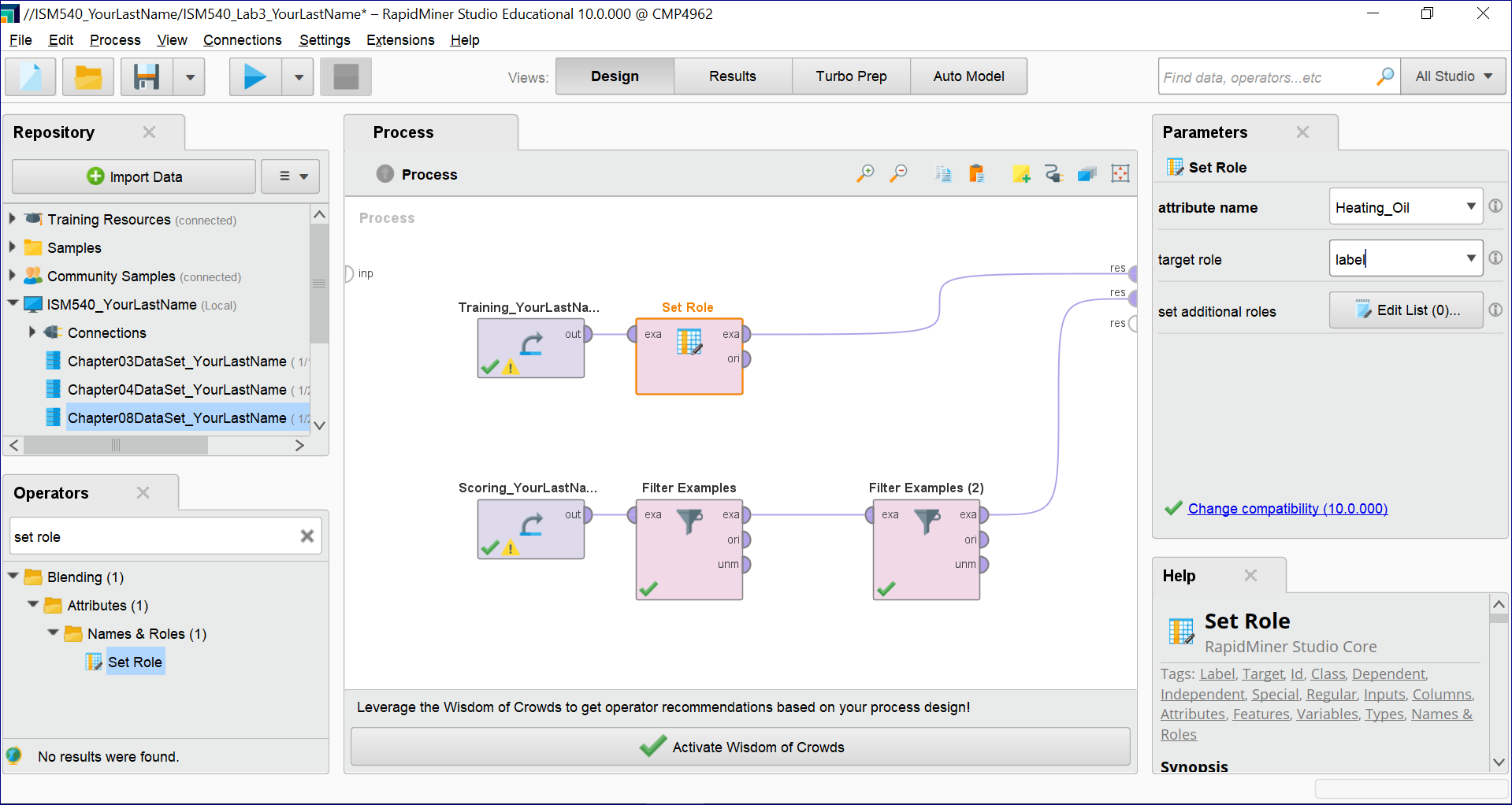
10) Now, return to **Design** perspective.

**Linear regression** is a predictive model, and thus will need an attribute to be designated as the **label**—this is the target, the thing we want to predict. Search for the **Set Role** operator in the Operators pane and drag it onto your training stream/path. While it is selected, go to the **Parameters** pane to choose **label** from the “target role” drop-down list, and specify **Heating\_Oil** in the “attribute name” textbox.

You will see below parameters pane if you are using the latest version of RapidMiner. Click on “Edit List”. 



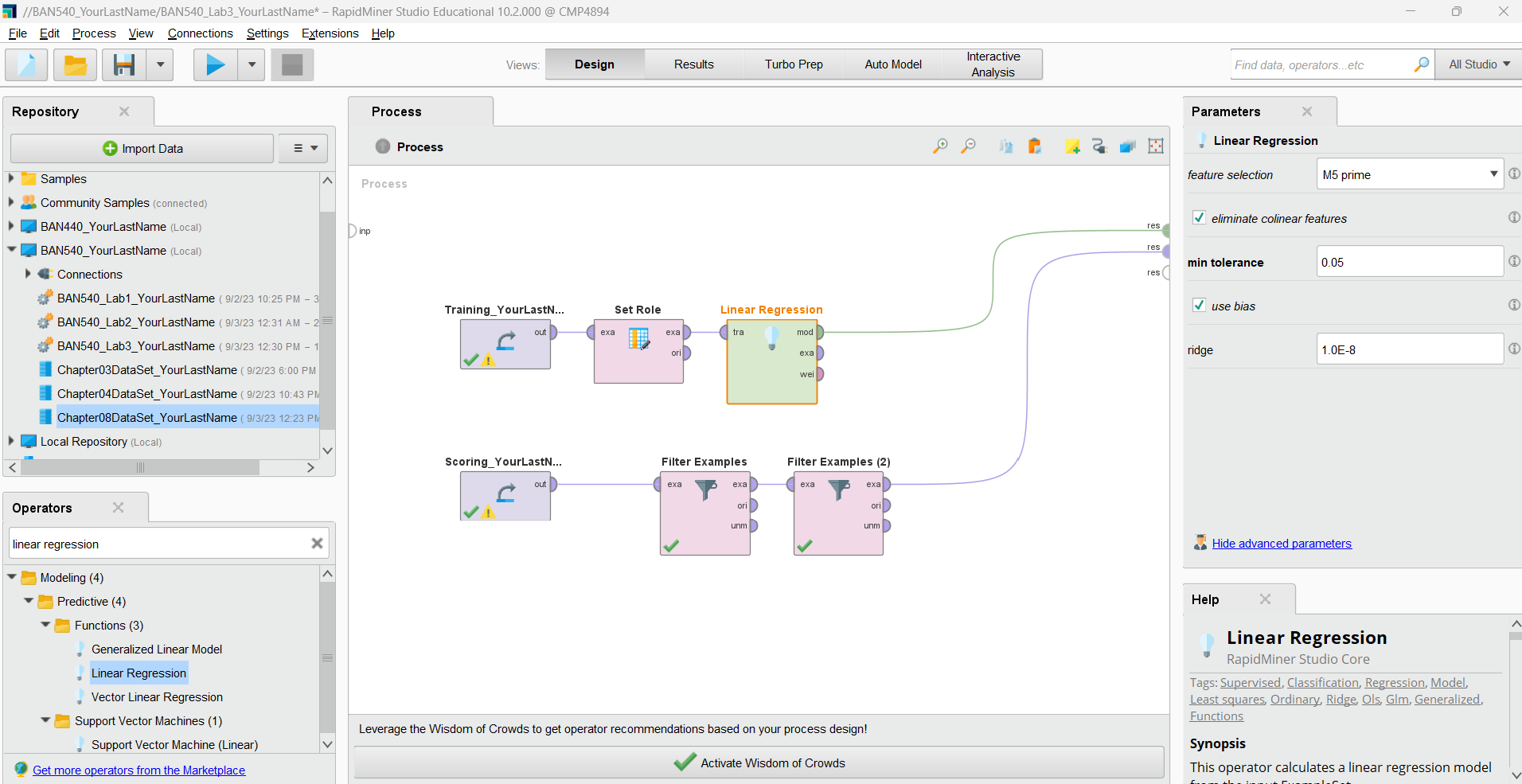
You will see below parameters pane if you are using an older version of RapidMiner:



With this step completed, our data sets are now prepared for…

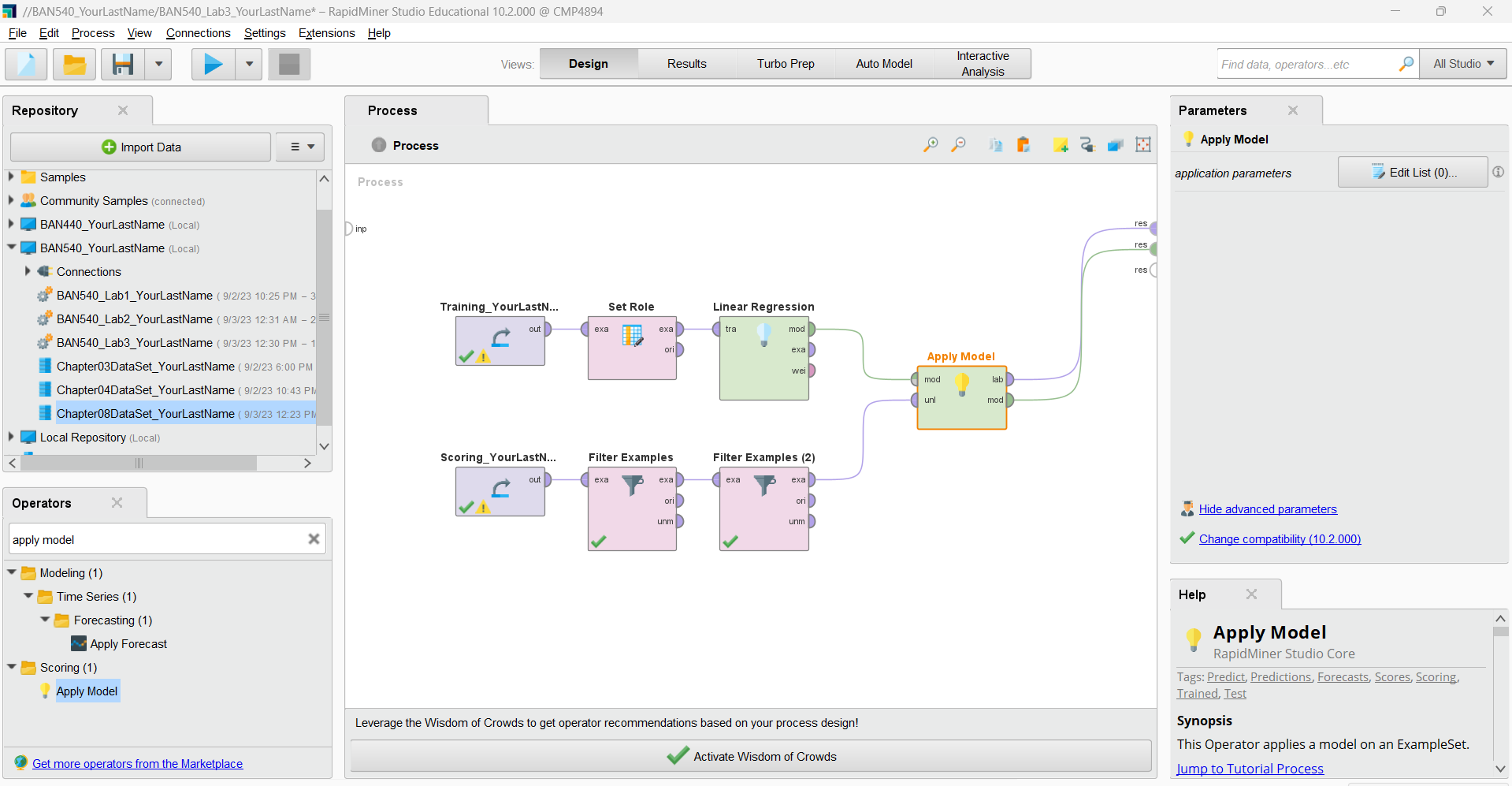
**MODELING**

1) Using the search field in the **Operators** pane again, locate the **Linear Regression** operator and drag and drop it onto your training data set’s stream. Note that the Linear Regression operator uses a default tolerance of .05 (also known in statistical language as the **confidence level** or **alpha level**). This value of .05 is very common in statistical analysis of this type, so we will accept this default.



2) The final step to complete our model is to use an **Apply Model** operator to connect our training stream to our scoring stream. Be sure to connect both the **lab** and **mod** ports coming from the **Apply Model** operator to **res** ports.

**Please make a screenshot now and replace my screenshot #2 with yours in the submission file (named as “BAN540 Lab 3 Submission YourLastName.docx”). Please make sure your screenshot shows YourLastName in the related items we have added so far (see the red boxes in the Repository pane on the left.)**



3) Now run the model. Having two splines coming from the **Apply Model** operator and connecting to res ports will result in two tabs in **Results** perspective. Let’s examine the **LinearRegression** tab first, as we begin our…

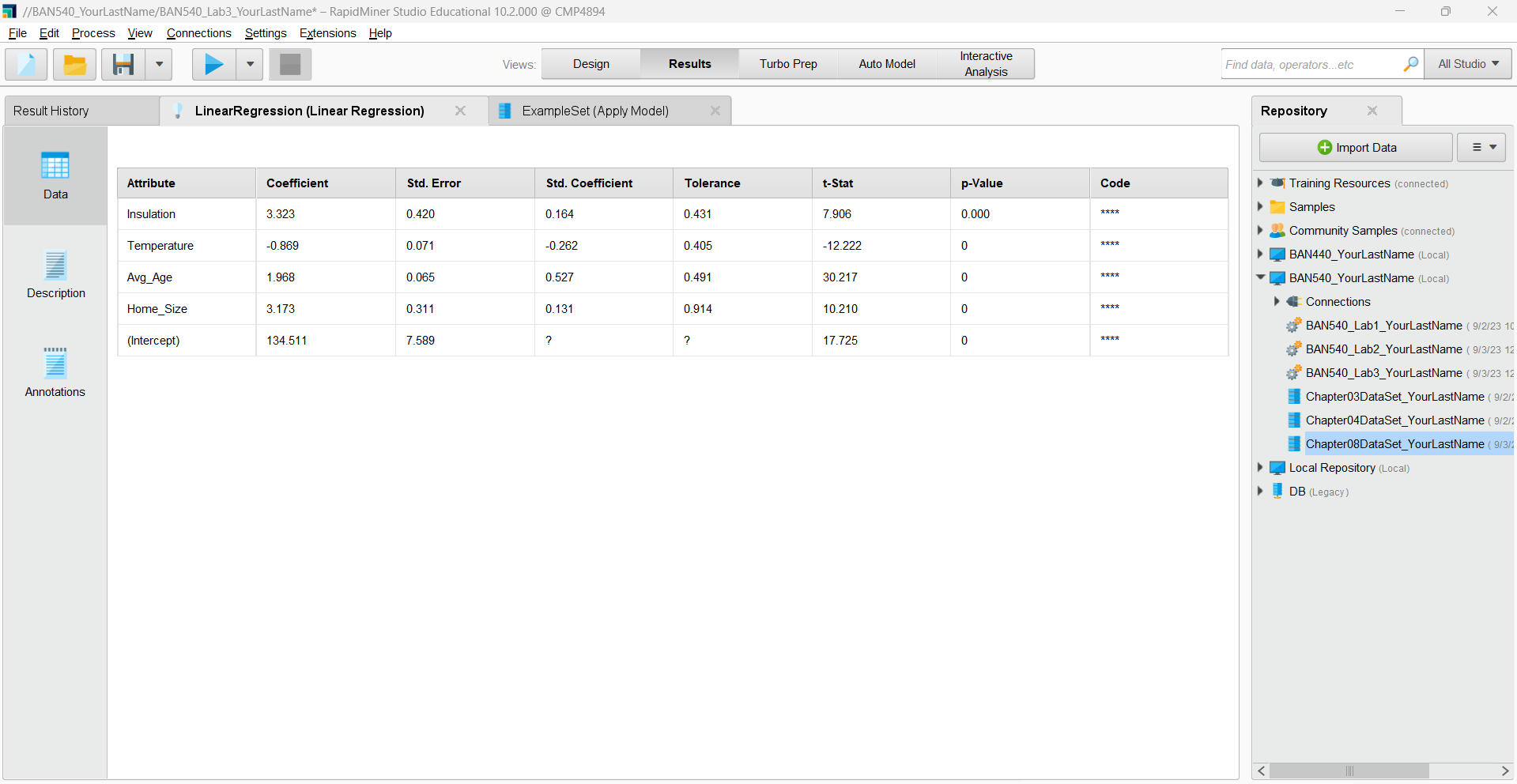
**EVALUATION**

1) The result we get in the following screenshot is a **multiple linear regression (MLR)** model, which is an extension of simple linear regression. The MLR model is:

***Predicted heating oil usage***

*= (****Insulation****\*****3.323****)+(****Temperature****\*-****.869****)+(****Avg\_Age****\*****1.968****)+(****Home\_Size****\*****3.173****)+****134.511***

**Please make a screenshot now and replace my screenshot #3 with yours in the submission file (named as “BAN540 Lab 3 Submission YourLastName.docx”). Please make sure your screenshot shows YourLastName in the related items we have added so far (see the red boxes in the Repository pane on the right.)**

****

What happened to **Num\_Occupants**? Why is it not in our MLR model? The answer is that Num\_Occupants was **not** a statistically significant predictor of heating oil usage in this data set, and therefore, RapidMiner removed it as a predictor. In other words, when RapidMiner evaluated the amount of influence each attribute in the data set had on heating oil usage for each home represented in the training data set, the number of occupants was so non-influential that its weight in the formula was set to zero. An example of why this might occur could be that two older people living in a house may use the same amount of heating oil as a young family of five in the house. The older couple might take longer showers, and prefer to keep their house much warmer in the winter time than would the young family. The variability in the number of occupants in the house doesn’t help explain each home’s heating oil usage very well, and so it was removed as a predictor in our model.

2) With the MLR model we just built, suppose we wanted to predict heating oil usage, using our model, for a home with the following attributes:

* Insulation: 6
* Temperature: 67
* Avg\_Age: 35.4
* Home\_Size: 5

Our formula for this home would be:

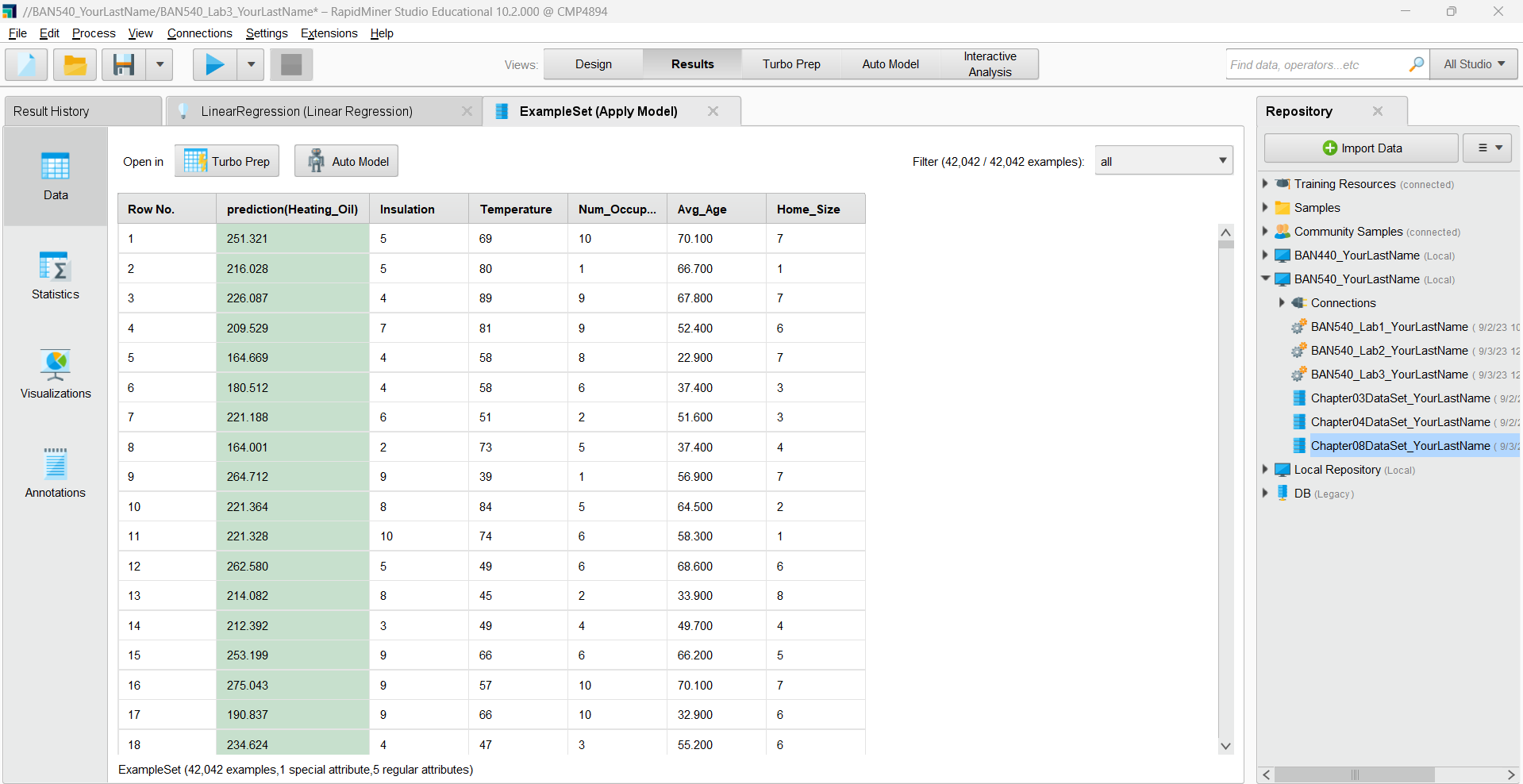
***Predicted heating oil usage*=(6\*3.323)+(67\*-.869)+(35.4\*1.968)+(5\*3.173)+134.511**

Our prediction for this home’s annual number of heating oil units ordered (y) is 181.758, or basically 182 units. Let’s check our model’s predictions as we discuss possibilities for…

**DEPLOYMENT**

1) While still in the **Results** perspective, please switch to the **ExampleSet(Apply Model)** tab, and select the **Data** tab. We can see in this view that RapidMiner has quickly and efficiently predicted the number of units of heating oil each of Sarah’s company’s new customers will likely use in their first year. This is seen in the **prediction(Heating\_Oil)** attribute/column.

**Please make a screenshot now and replace my screenshot #4 with yours in the submission file (named as “BAN540 Lab 3 Submission YourLastName.docx”). Please make sure your screenshot shows YourLastName in the related items we have added so far (see the red boxes in the Repository pane on the right.)**

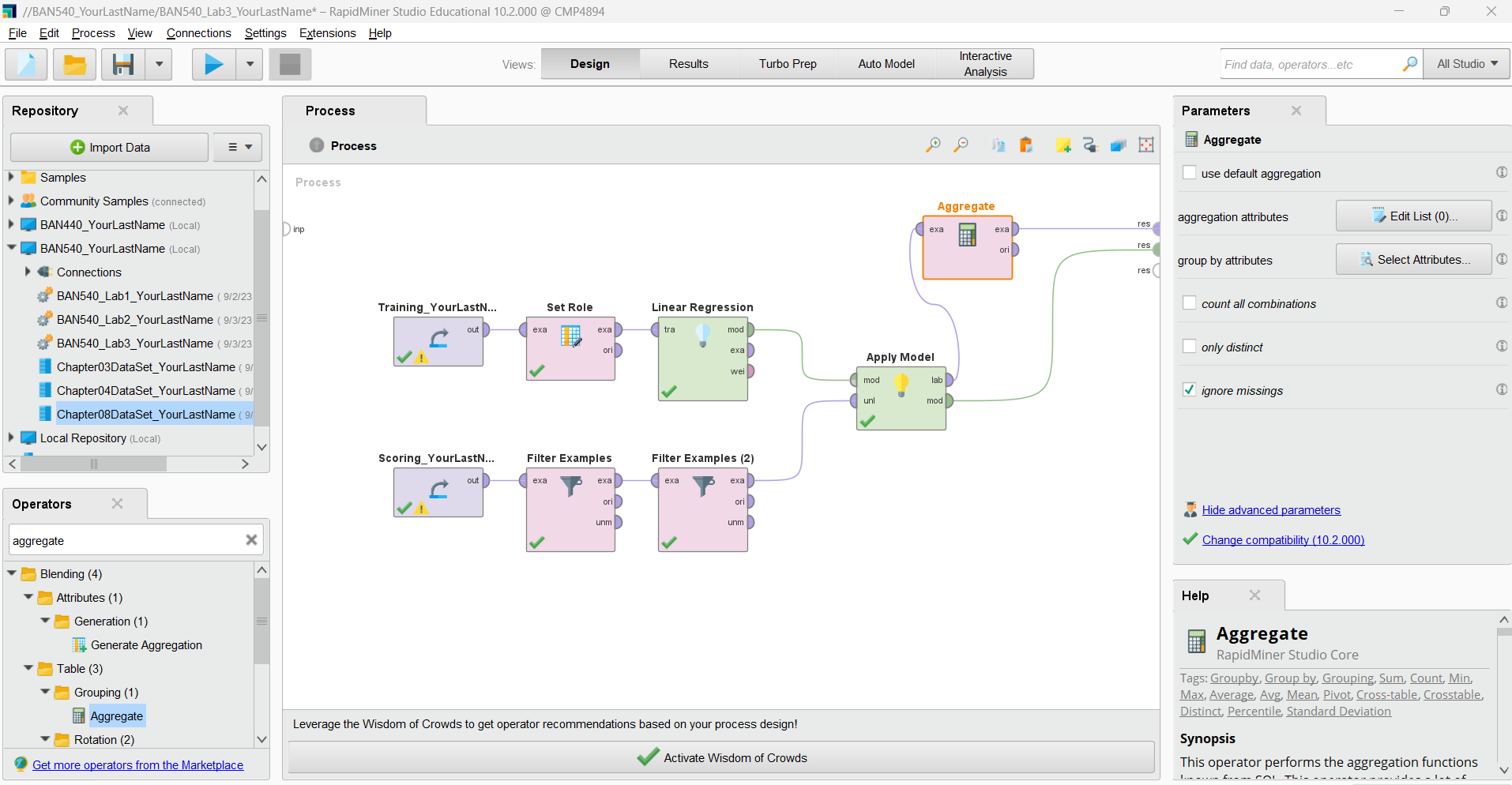
****

2) Let’s check the first of our 42,042 households by running the linear regression formula for row 1: *(5\*3.323)+(69\*-0.869)+(70.1\*1.968)+(7\*3.173)+134.511 =* ***251.321***

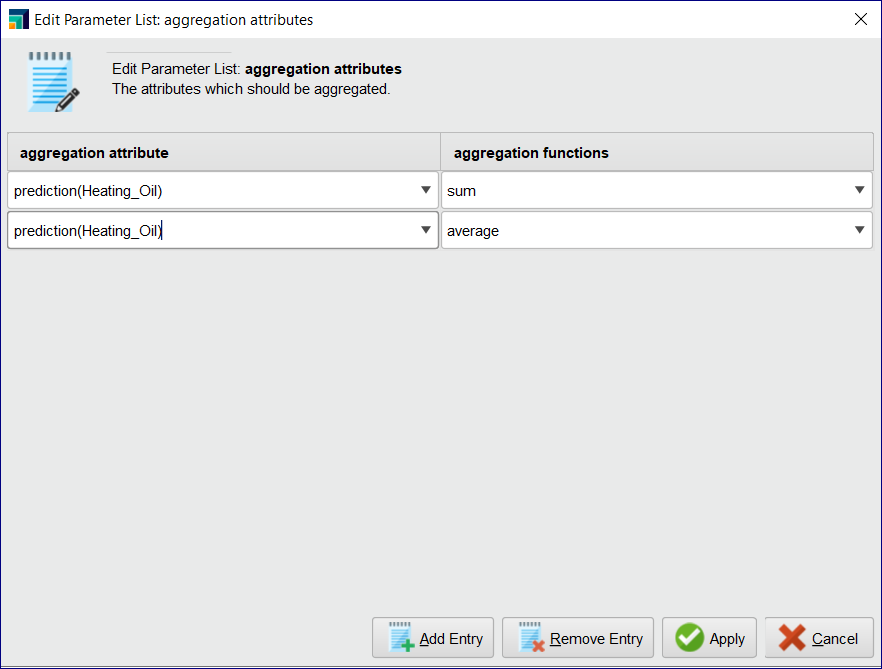
Note that in this formula we skipped the Num\_Occupants attribute because it is not predictive. The formula’s result does indeed match RapidMiner’s prediction for this home. Sarah now has a prediction for each of the new clients’ homes, with the exception of those that had Avg\_Age values that were out of range. How might Sarah use this data? She could start by summing the prediction attribute. This will tell her the total new units of heating oil her company is going to need to be able to provide in the coming year. This can be accomplished by exporting her data to a spreadsheet and summing the column, or it can even be done within RapidMiner using an **Aggregate** operator. We will demonstrate this briefly.

3) Please switch back to **Design** perspective. Search for the **Aggregate** operator in the **Operators** pane and add it between the **lab** (means labelled data) and **res** ports, as shown below.

**Please make a screenshot now and replace my screenshot #5 with yours in the submission file (named as “BAN540 Lab 3 Submission YourLastName.docx”). Please make sure your screenshot shows YourLastName in the related items we have added so far (see the red boxes in the Repository pane on the left.)**



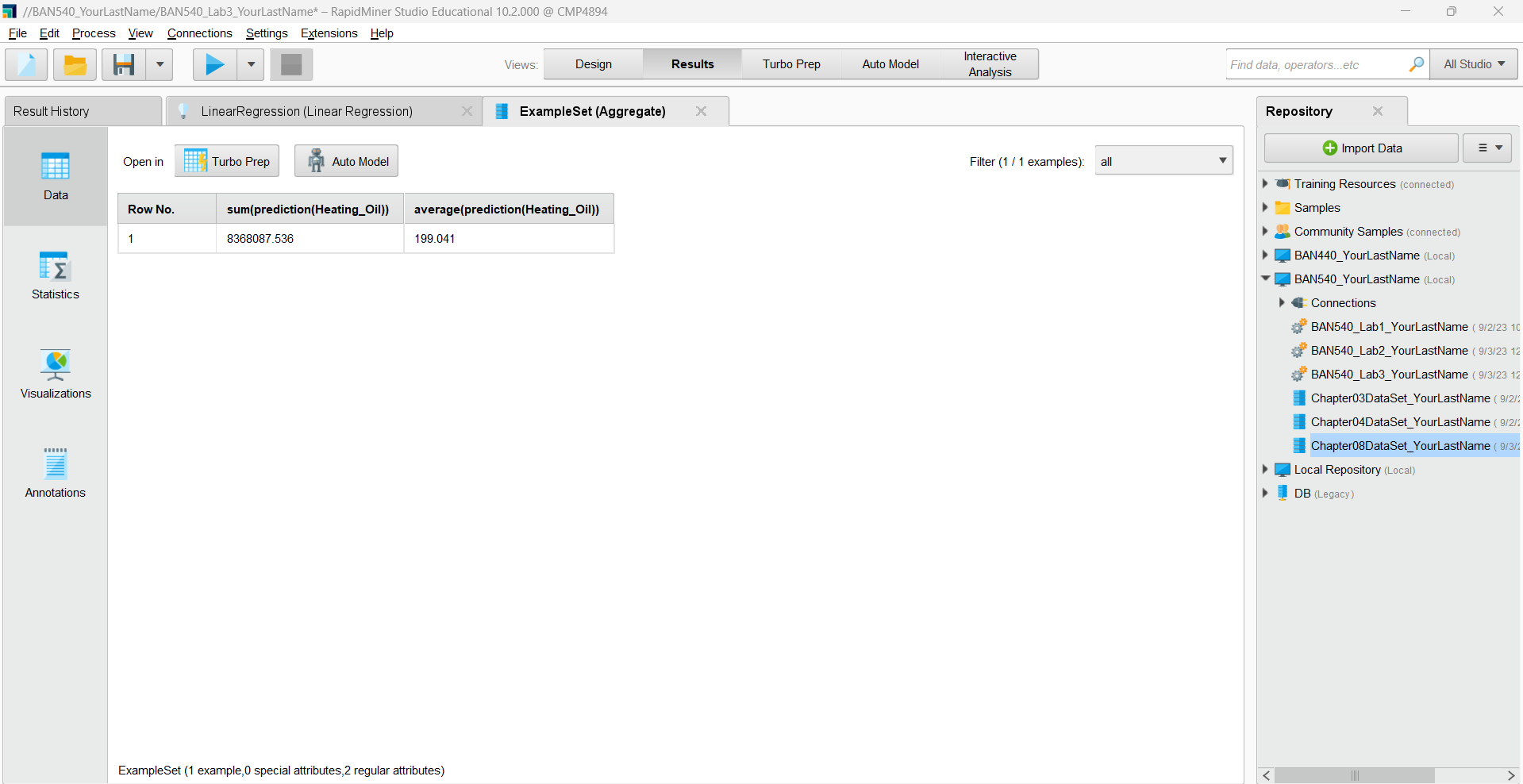
4) While the **Aggregate** operator is selected, please go to the **Parameters** pane. Click on the **Edit List** button in the ‘aggregation attributes’ field. Then, a window similar to the following screenshot will appear. Choose **prediction(Heating\_Oil)** from the ‘aggregation attribute’ dropdown list, and **sum** fromthe ‘aggregation functions’ dropdown list. Then click on **Add Entry** to add an **average** for **prediction(Heating\_Oil)** as well.



5) When you are done with the above settings, click **Apply** to return to your main process window.

6) **Run** the model. In the **Results** perspective, select the **ExampleSet(Aggregate)** tab, and then select the **Data** tab. The sum and average for the prediction attribute will be shown as in the following screenshot.

**Please make a screenshot now and replace my screenshot #6 with yours in the submission file (named as “BAN540 Lab 3 Submission YourLastName.docx”). Please make sure your screenshot shows YourLastName in the related items we have added so far (see the red boxes in the Repository pane on the right.)**



From this image, we can see that Sarah’s company is likely to sell 8,368,088 units of heating oil to these new customers. The company can expect that on average, their new customers will order about 200 units each. These figures are for all 42,042 clients together, but Sarah is probably going to be more interested in regional trends. In order to deploy this model to help her more specifically address her new customers’ needs, she should probably extract the predictions, match them back to their source records which might contain the new clients’ addresses, enabling her to break the predictions down by city, county, or region of the country. Sarah could then work with her colleagues in Operations and Order Fulfillment to ensure that regional heating oil distribution centers around the country have appropriate amounts of stock on hand to meet anticipated need. If Sarah wanted to get even more granular in her analysis of these data, she could break her training and scoring data set down into months using a month attribute, and then run the predictions again to reveal fluctuations in usage throughout the course of the year.