**BAN540 Lab 1 - Data Preparation in RapidMiner (30 points)**

*Adapted from “Data Mining for the Masses” Chapter 3*

Please follow the instructions carefully to finish lab assignment 1. In this assignment, you will be asked to make **8** **screenshots** and paste them to the “*BAN540 Lab1 Submission YourLastName.docx*” file. ***Hint***: you will make the first screenshot when you get to page 14 of this instruction file. Once you are done with all 8 screenshots, please submit the word file “BAN540 Lab1 Submission YourLastName.docx” (with your own last name 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.”

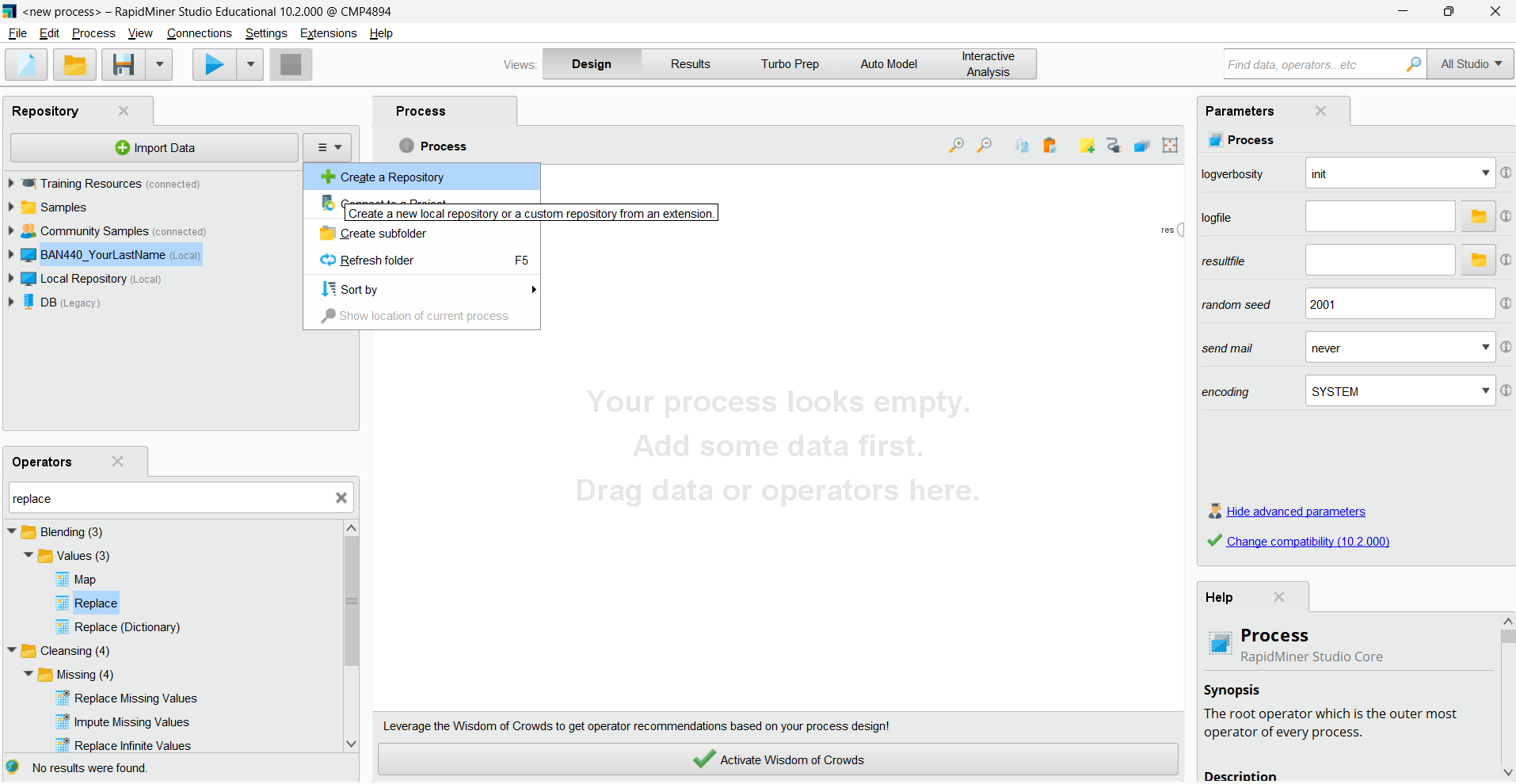
**A. CREATE YOUR OWN REPOSITORY**

1. Launch the RapidMiner application. This can be done by double clicking on your desktop icon named as “RapidMiner Studio” (as shown below), or by finding it in your application menu.



Within RapidMiner there are two main areas that hold useful tools: **Repositories** and **Operators**. The **Repositories** area is the place where you will connect to each data set. The **Operators** area is where all data mining tools are located. These are used to build models and otherwise manipulate data sets.

1. Follow the screenshot below to create your own new repository for BAN 540 class.

****

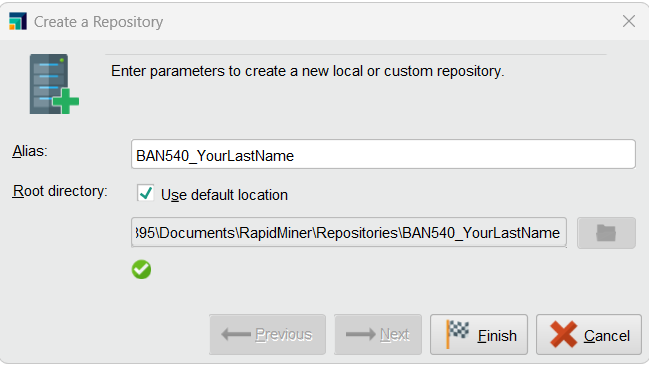
1. Click Next

Graphical user interface, text, application, email

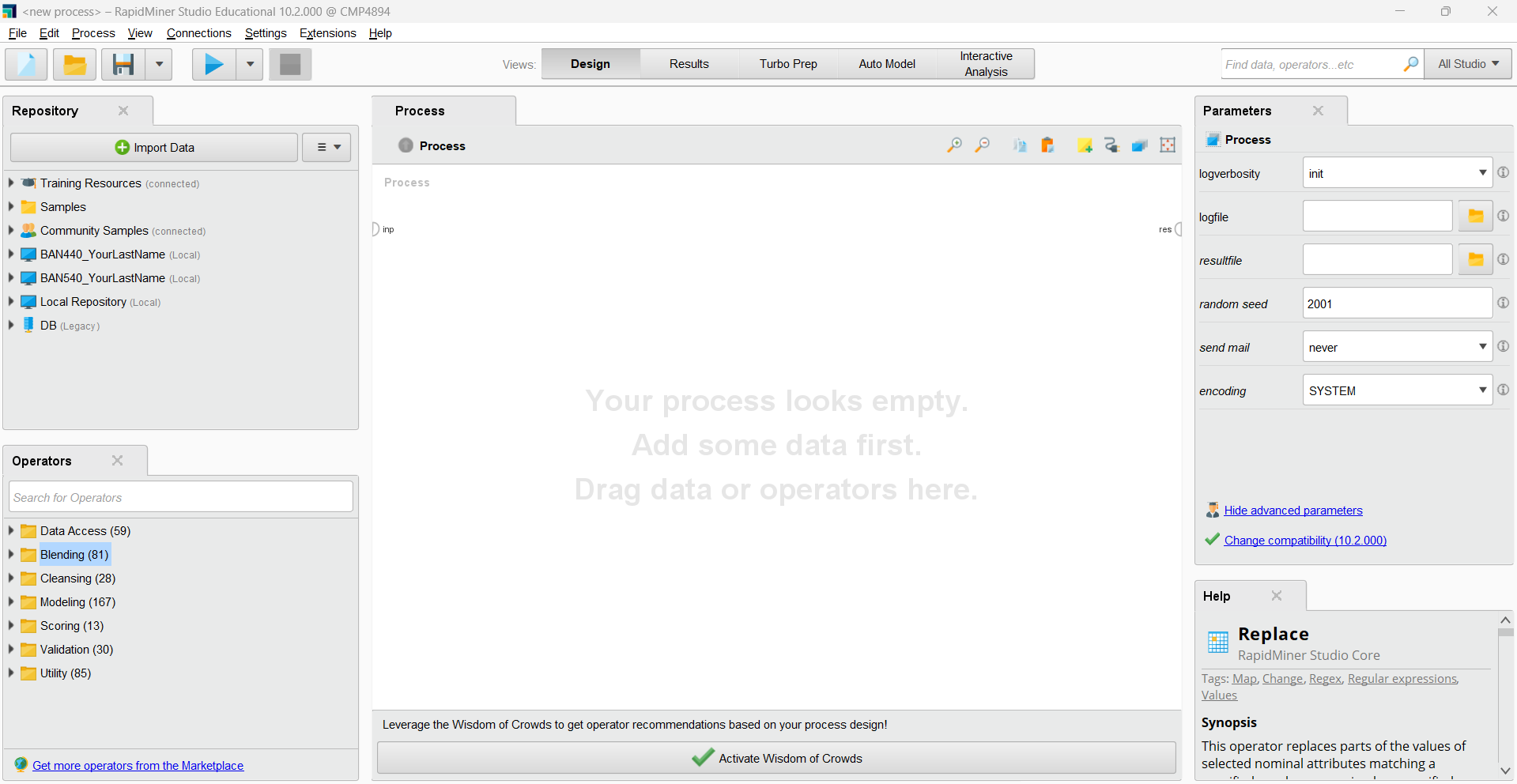
Description automatically generated

Please change the Alias name to **BAN540\_YourLastName**,and find a local folder where you want to put all the files related to this class. Then click Finish. ***(Important Note: you MUST name it as BAN540\_YourLastName to get credit for this step).***

*Hint: There is* ***no*** *specific requirement on where to save your repository for this lab.* *You may put it in your own FCB shared drive where you can access from any FCB computer.* *The repository you created here is computer specific, which means you may have to recreate it if you switch to a different computer.*

****

1. You will see a newly created repository named as BAN540\_YourLastName.

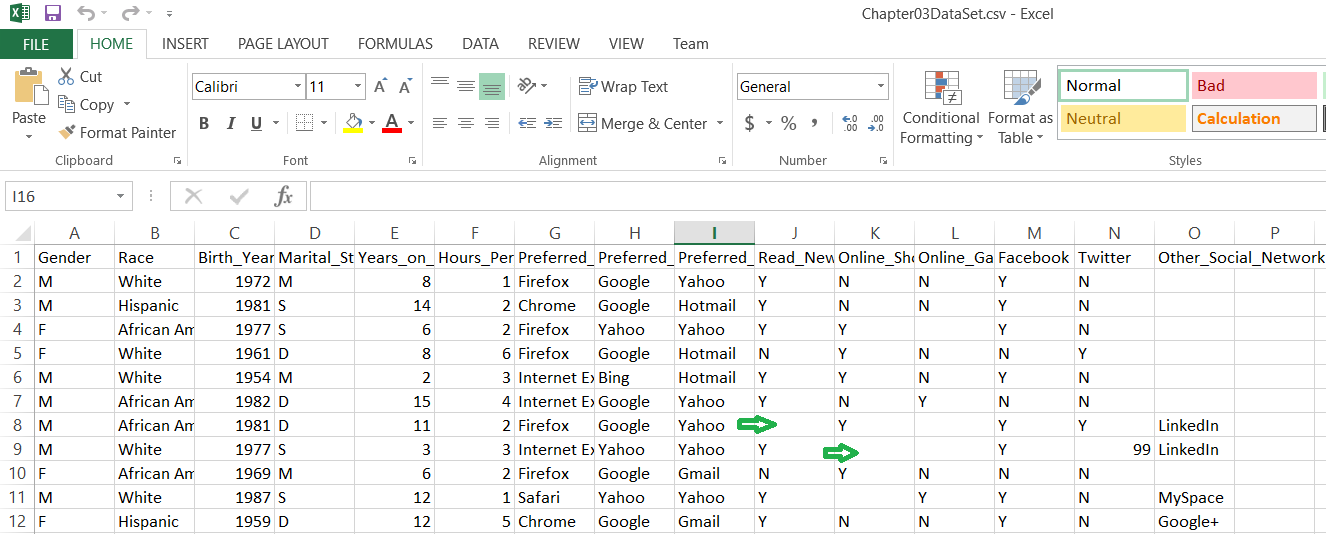
****

**B. DOWNLOAD AND IMPORT DATA**

1) Please download the dataset file “Chapter03DataSet.csv” from Canvas and save it to your local drive.

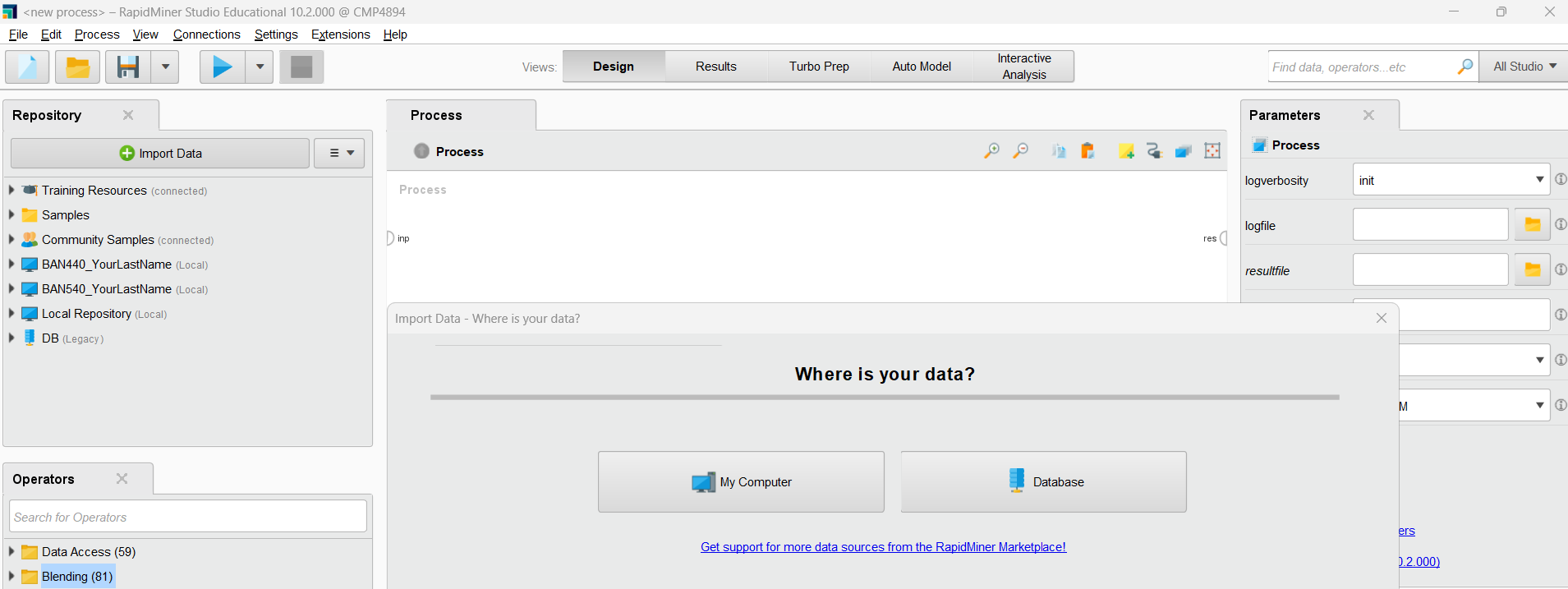
2) You can use Excel to view the downloaded file. This data set is very small, comprised of only 15 attributes and 11 observations. Our next step is to connect to this data set. When you browse this data set, you will notice there are some missing data as indicated by the green arrows (see below).

**Missing data** are data that do not exist in a data set. As you can see in the screenshot, missing data is not the same as zero or some other value. It is blank, and the value is **unknown**. Missing data are also sometimes known in the database world as **null**. Depending on your objective in data mining, you may choose to leave missing data as they are, or you may wish to replace missing data with some other value. We will deal with the missing data in later steps of this assignment.

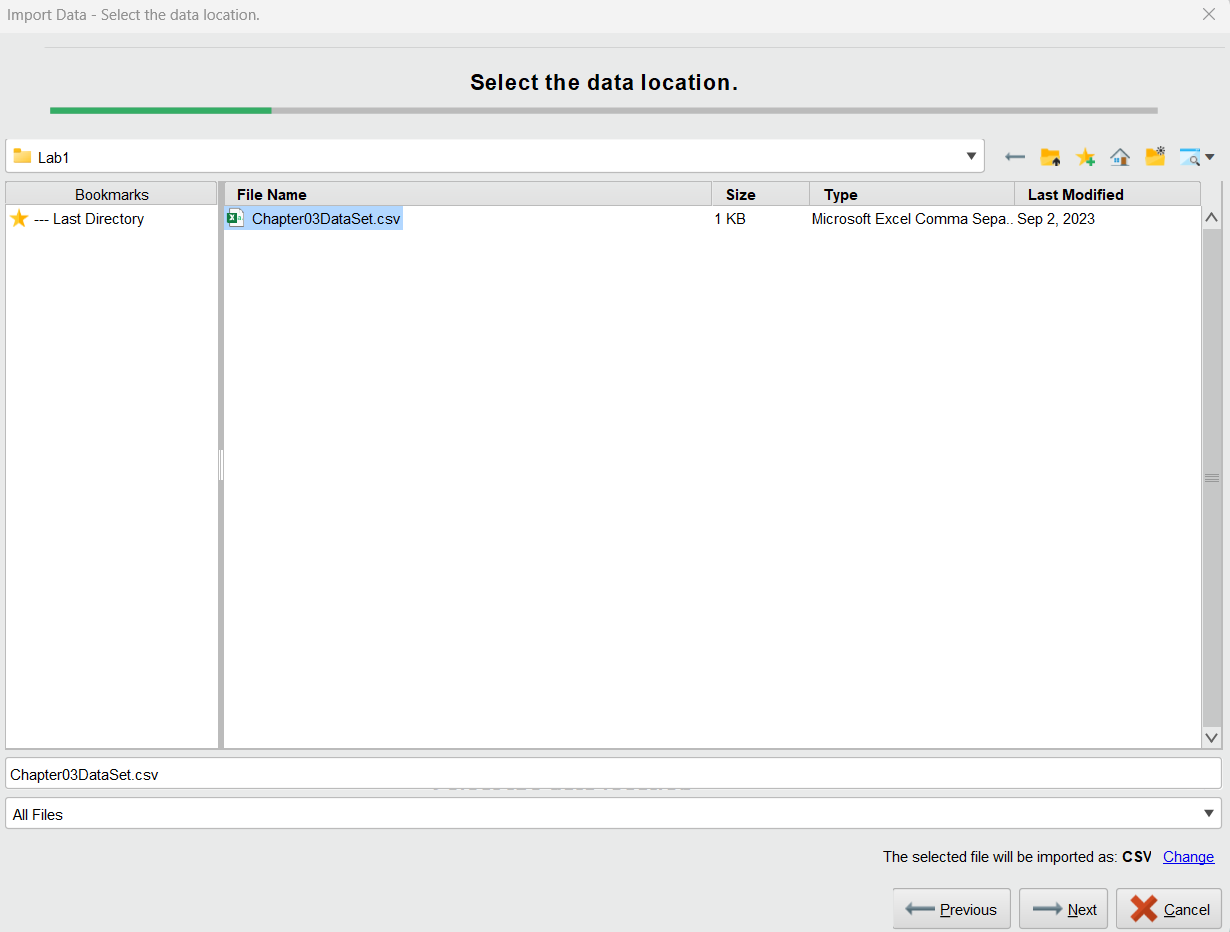


At this point, we could do a number of complicated and technical things, such as connecting to a remote enterprise database. This, however, would likely be overwhelming for now. For the purposes of this lab assignment, we will only need to connect to comma separate values (CSV) files. Please be aware that in the real world, most data mining projects incorporate extremely large data sets, encompassing dozens of attributes and thousands or even millions of observations. We will use smaller data sets in this assignment, but the foundational concepts illustrated here are the **same** as for larger ones.

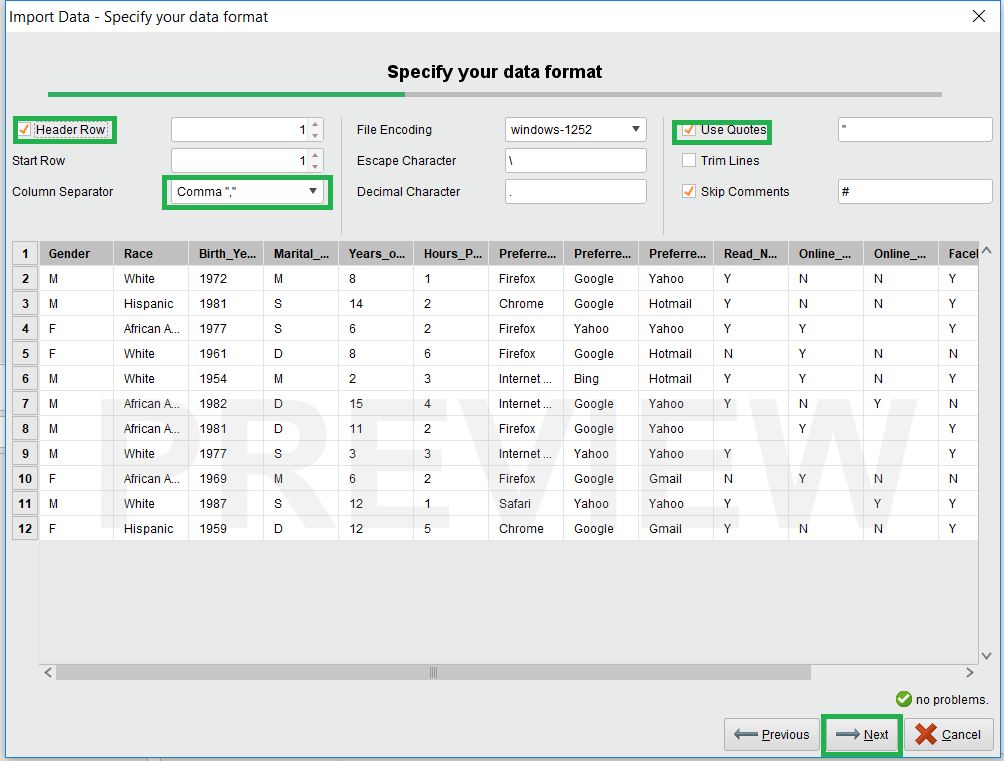
3) Click on the “**Import Data**” icon, as indicated in the red rectangle box on the picture below. Then click on “**My Computer**.” Note that by importing, you are bringing your data into a RapidMiner file, rather than working with data that are already stored elsewhere. If your data set is extremely large, it may take some time to import, and you should be mindful of disk space that is available to you.



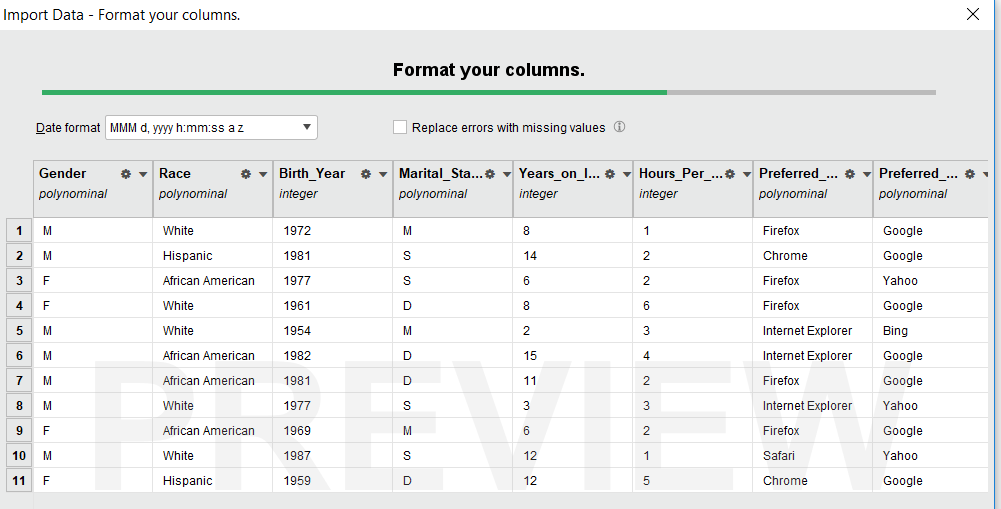
4) Locate the file (Chapter03DataSet.csv), and then click on Next.



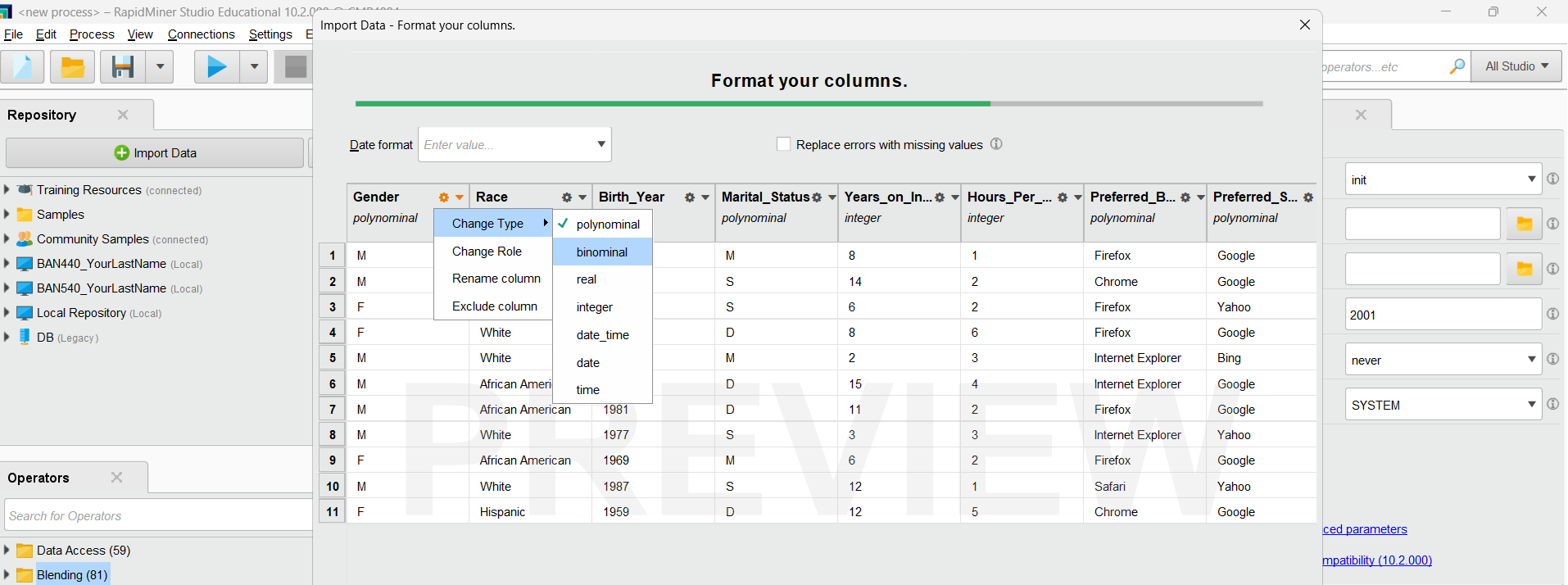
5) The column separation delimiter is Comma. Keep the default settings as shown in the screenshot below and click on Next.



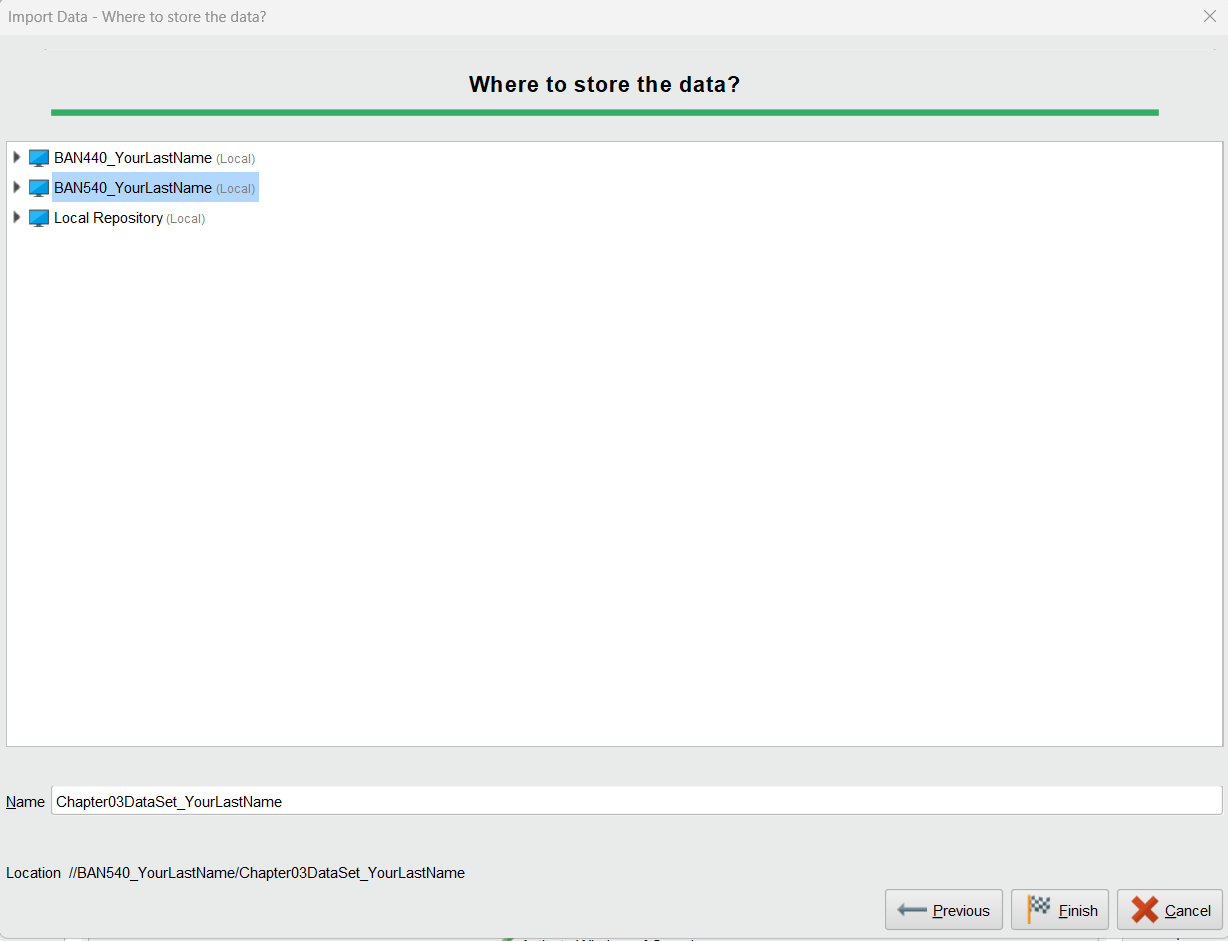
6) RapidMiner will take its best guess at a **data type** for each attribute. The data type is the kind of data an attribute holds, such as polynominal, integer, or text.



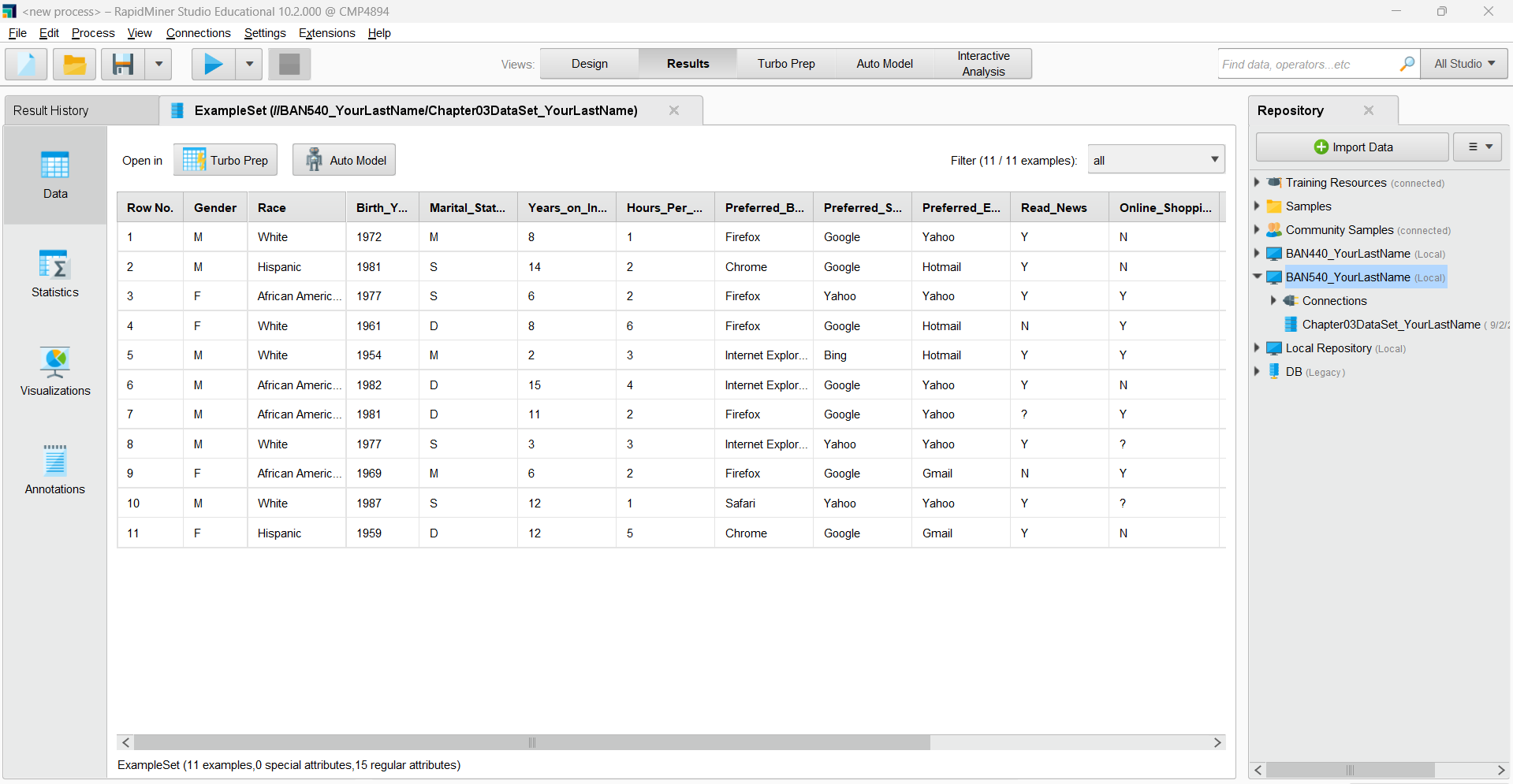
7) **Date types** can be changed by following the screenshot below. Please change **Gender** from “polynominal” to “**binominal**.” RapidMiner also indicates a **Role** for each attribute to play. By default, all columns are imported simply with the role of ‘attribute’, however we can change these by clicking on “**Change Role**” if we know a particular attribute is going to play a specific role in a data mining model that we will create. Since roles can be set within RapidMiner’s main process window when building data mining models, we will just accept the default ‘attribute’ whenever we import data sets for our class. Also, you may note that “**Exclude Column**” allows you to ***not***import some of the attributes if you don’t want to. Again, attributes can be excluded from models later in needed, so for this class, we will always include all attributes when importing data. Click on Next.



8) The final step for importing 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 you just created, which is **BAN540\_YourLastName**, and name it as **Chapter03DataSet\_YourLastName**. Then click Finish. ***(Important Note: You MUST name it as Chapter03DataSet\_YourLastName to get credit for this step).***

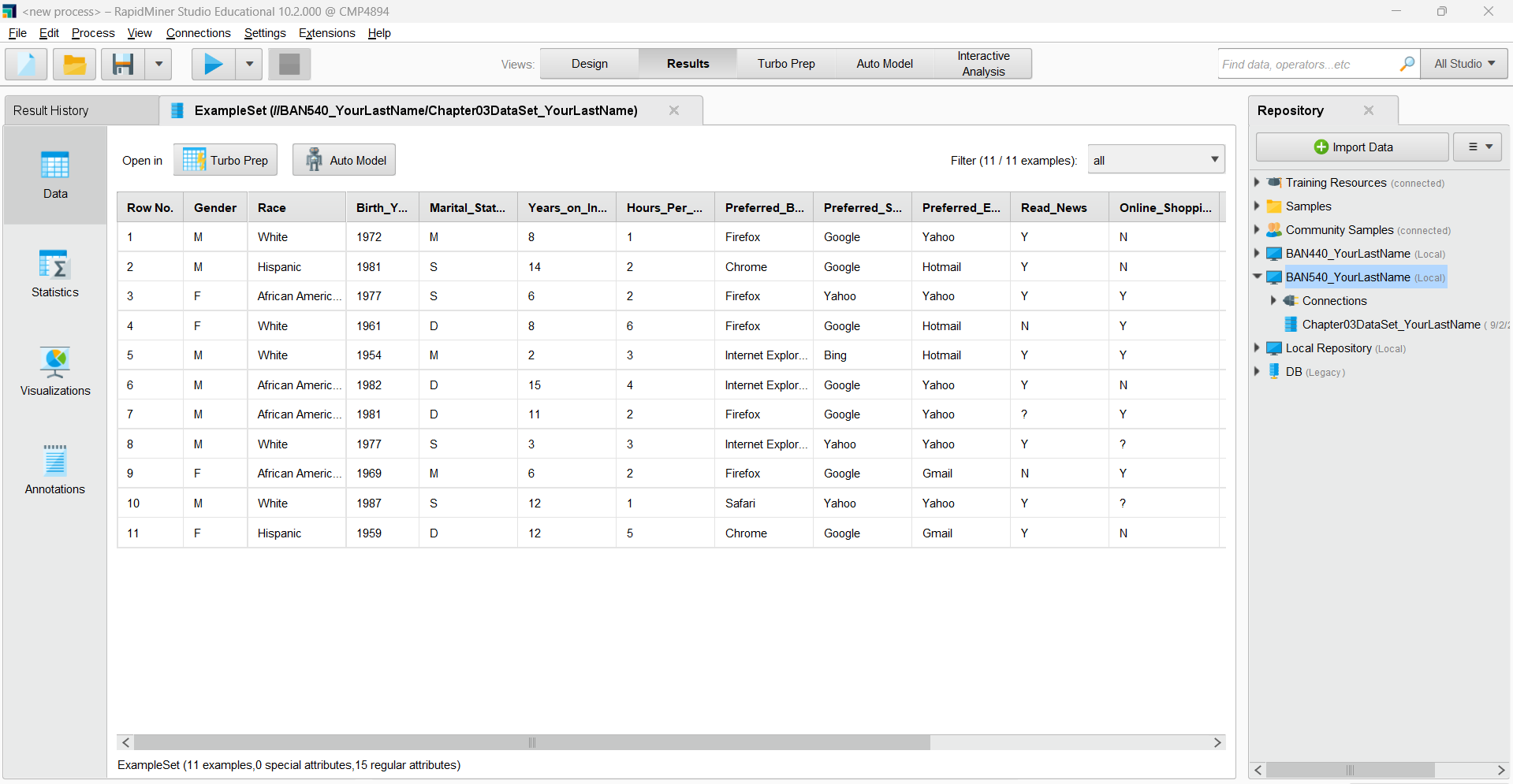


9) Once you click on Finish, this data set will become available to you for any type of data mining process you would like to build upon it. The following screen shows you the **Results Perspective**.

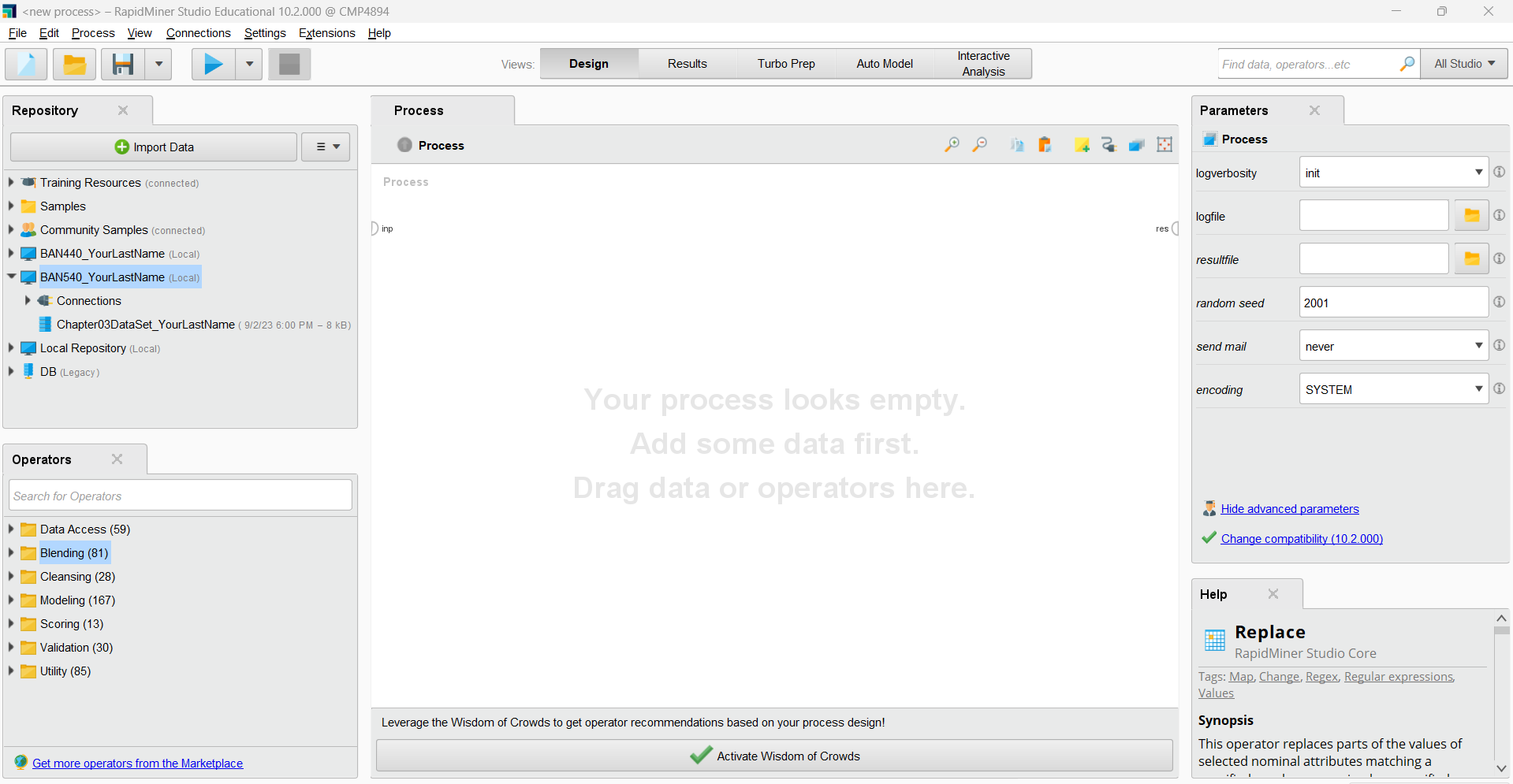


**C. RETRIEVE DATA OPERATOR**

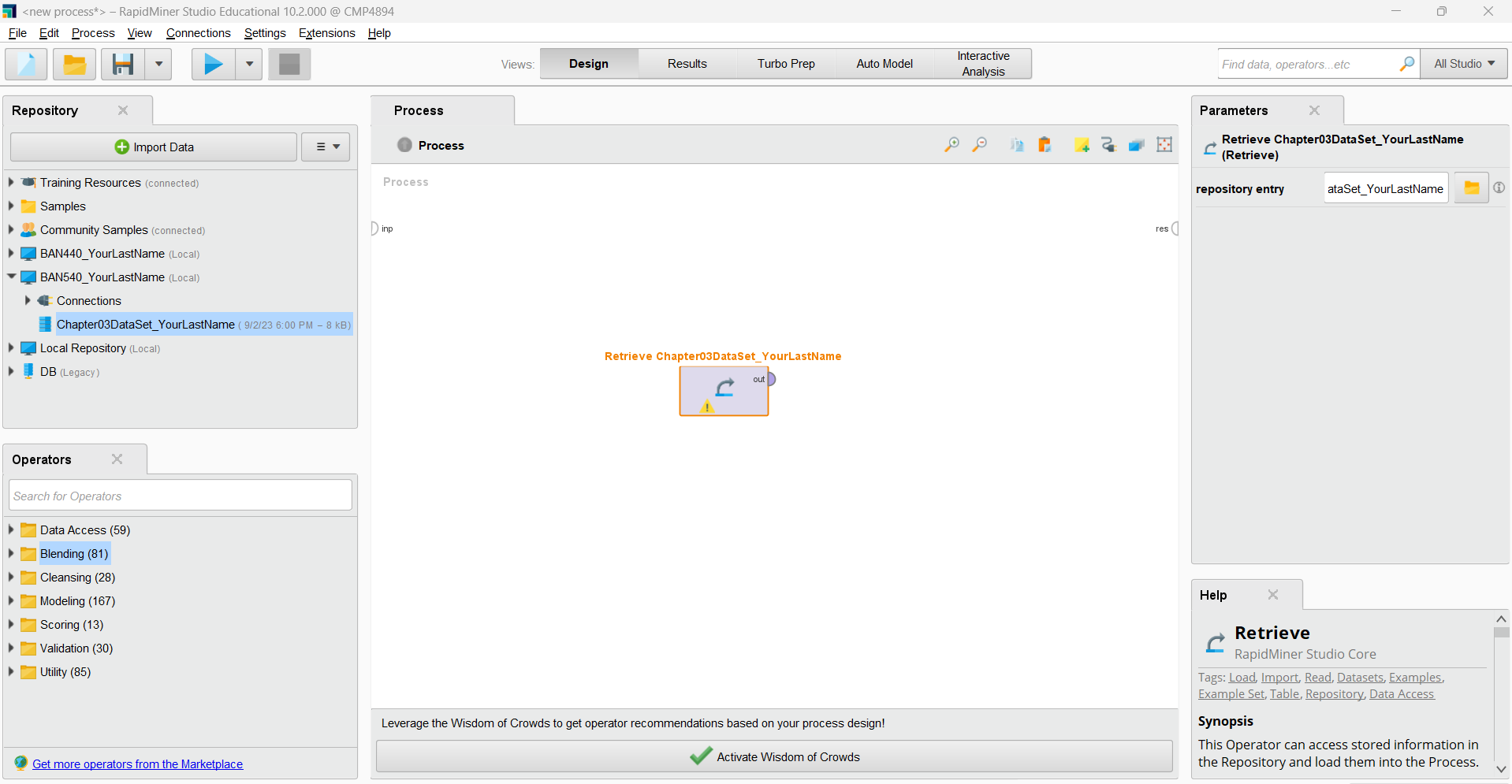
1) To continue, please click on “**Design**” tab on the top to switch back to **Design Perspective**.



2) The following screenshot shows the **Design** view. We can see that the data set “**Chapter03DataSet\_YourLastName**” is now available for use in RapidMiner.

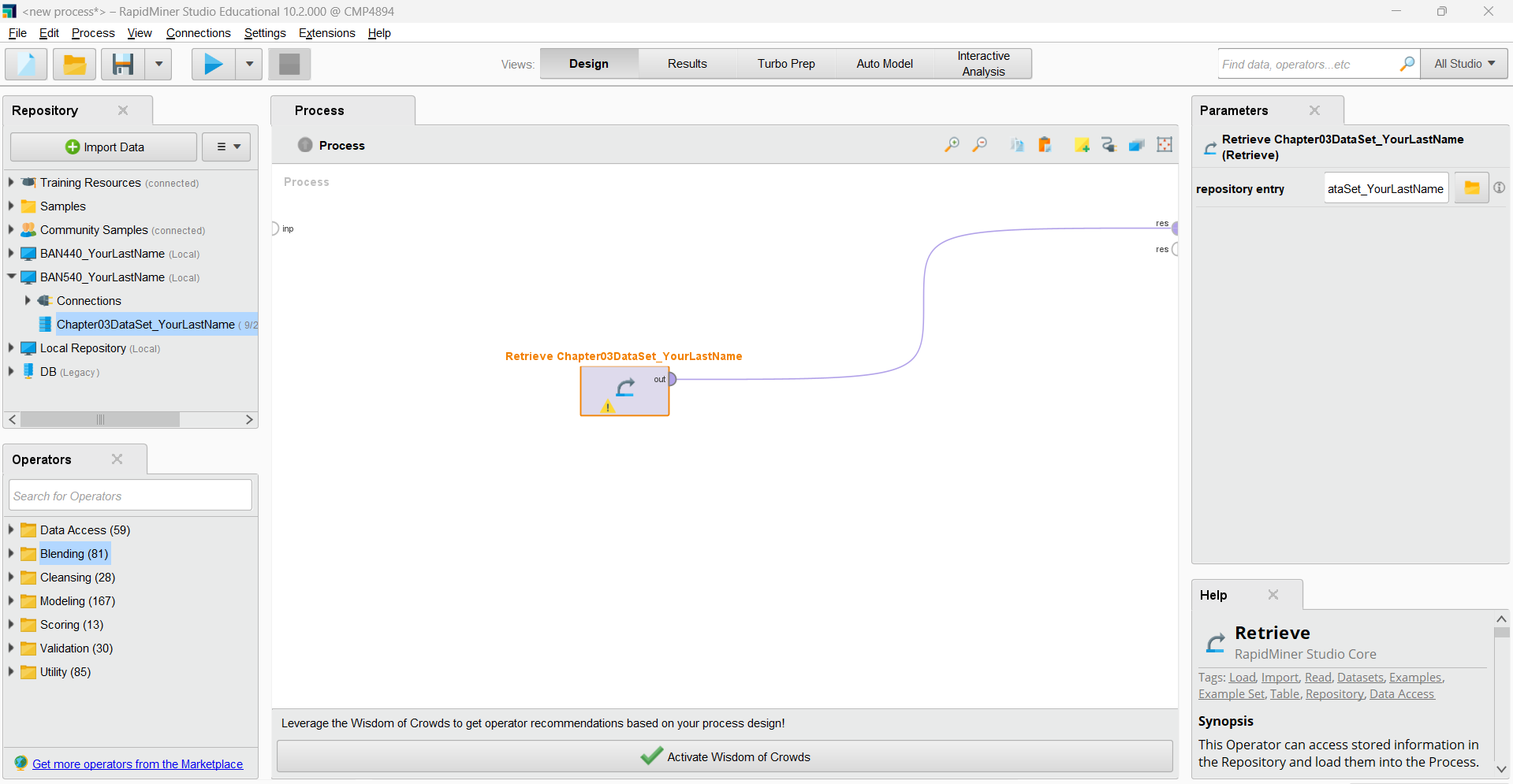


3) To begin using it in a RapidMiner data mining process, simply drag the data set and drop it onto the **Main Process** window.

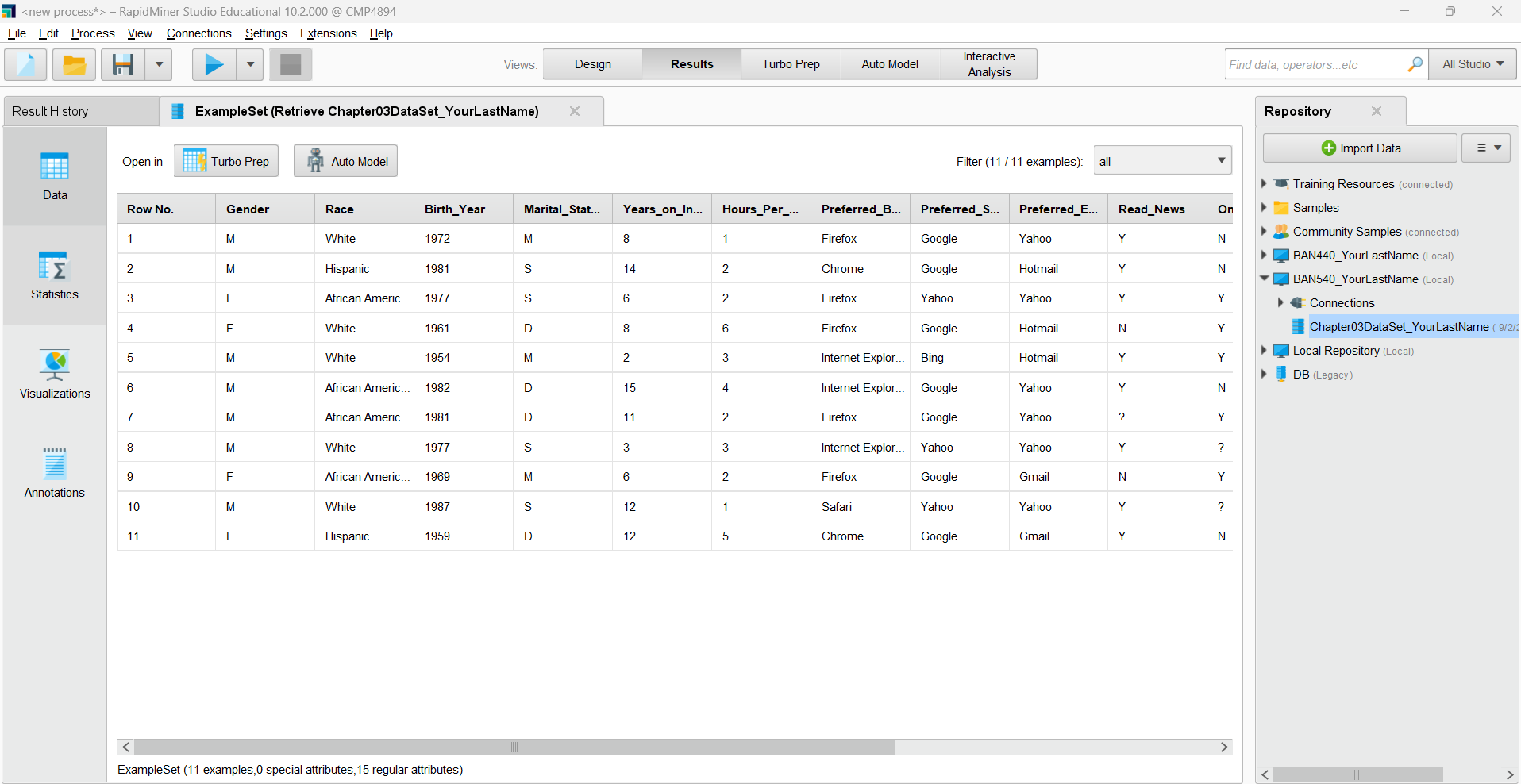


4) Each rectangle in a process in RapidMiner is called an **operator**. The Retrieve operator simply gets a data set and makes it available for use. The small half-circles on the sides of the operator, and of the Main Process window, are called **ports**. In the following screenshot, an output (*out*) port from our data set’s Retrieve operator is connected to a result set (*res*) port via a **spline**. To draw the spline, please put your mouse cursor to the *out* port and then move your mouse while holding it, to connect to the *res* port (on the very left side of the Process window).

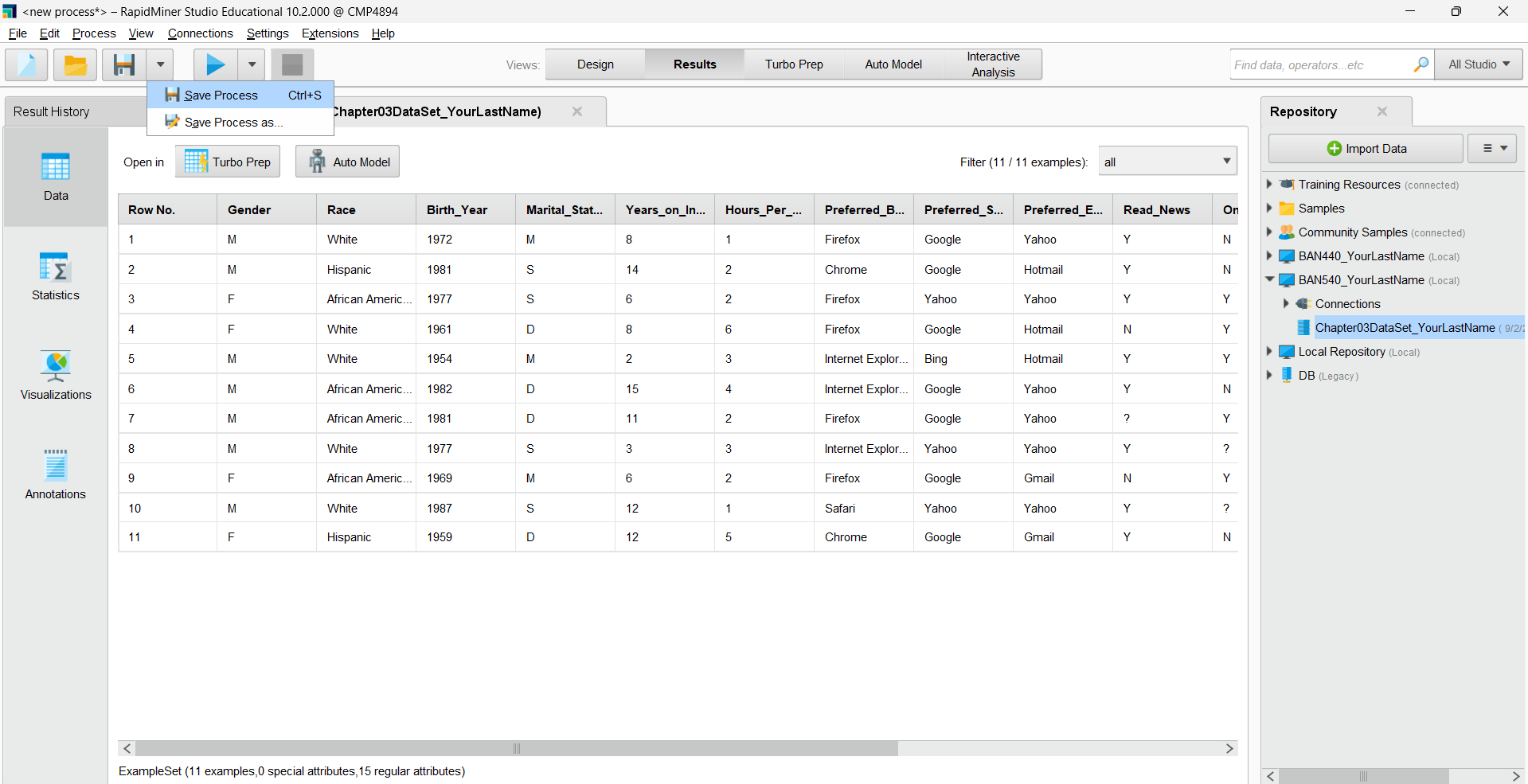
The splines, combined with the operators connected by them, constitute a **data mining stream**. To run a data mining stream and see the results, click on the blue, triangular **Play** button on the toolbar at the top of the RapidMiner window.



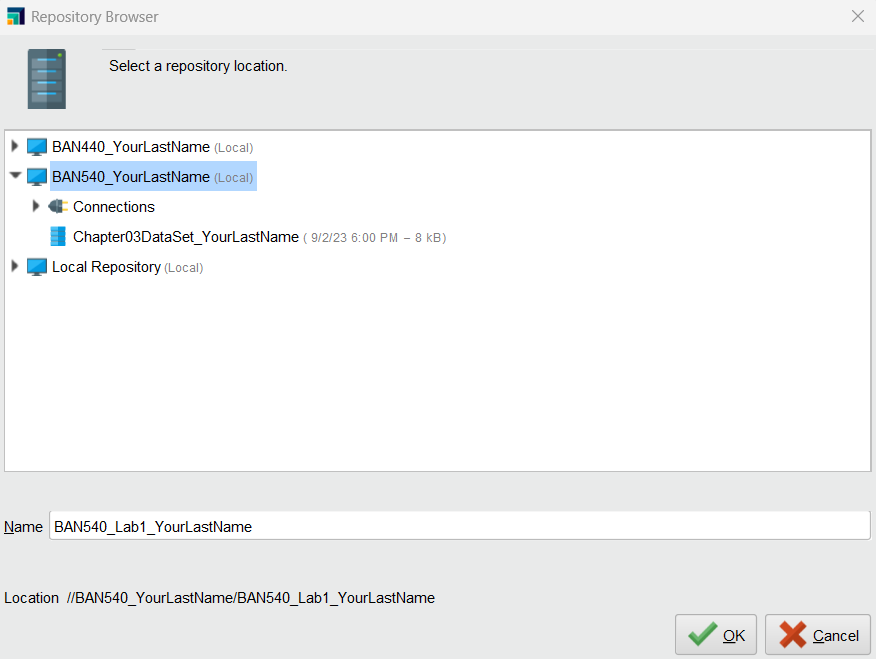
5) This will change your view from **Design Perspective**, which is the above screenshot where you can change your data mining stream, to **Results Perspective**, which shows your stream’s results, as pictured in the following screenshot.



6) When you hit the Play button, you may be prompted to save your process, and you are encouraged to do so. If not, please follow the screenshot below to “**Save Process**”



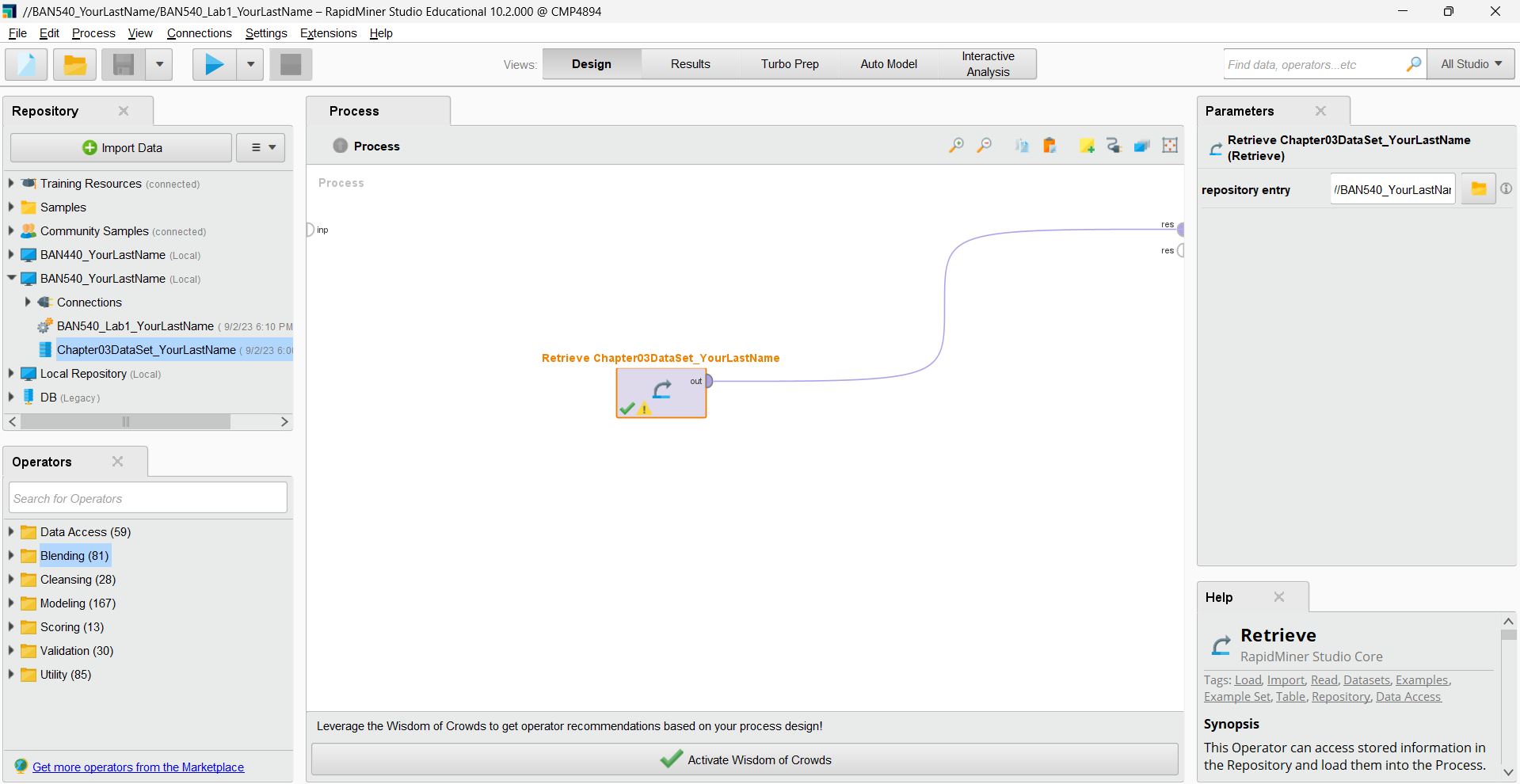
7) Please save the process into the repository you just created, which is **BAN540\_YourLastName.** Name your process as **BAN540\_Lab1\_YourLastName.** Then click OK. ***(Important Note: You MUST name it as BAN540\_Lab1\_YourLastName to get credit for this step).***



8) You will then see the following screenshot. In the **Result Perspective**, you can find the repository we created, which is “**BAN540\_YourLastName**” on the right side of the screen. You should also be able to see the dataset “**Chapter03DataSet\_YourLastName**” and the process “**BAN540\_Lab1\_YourLastName,**” both under the “BAN540\_YourLastName” Repository.



9) Please switch back to the **Design Perspective** by clicking on “Design” as shown below. You will find the repository “BAN540\_YourLastName,” the dataset “Chapter03DataSet\_YourLastName,” and the process “BAN540\_Lab1\_YourLastName” on the left side of the screen.

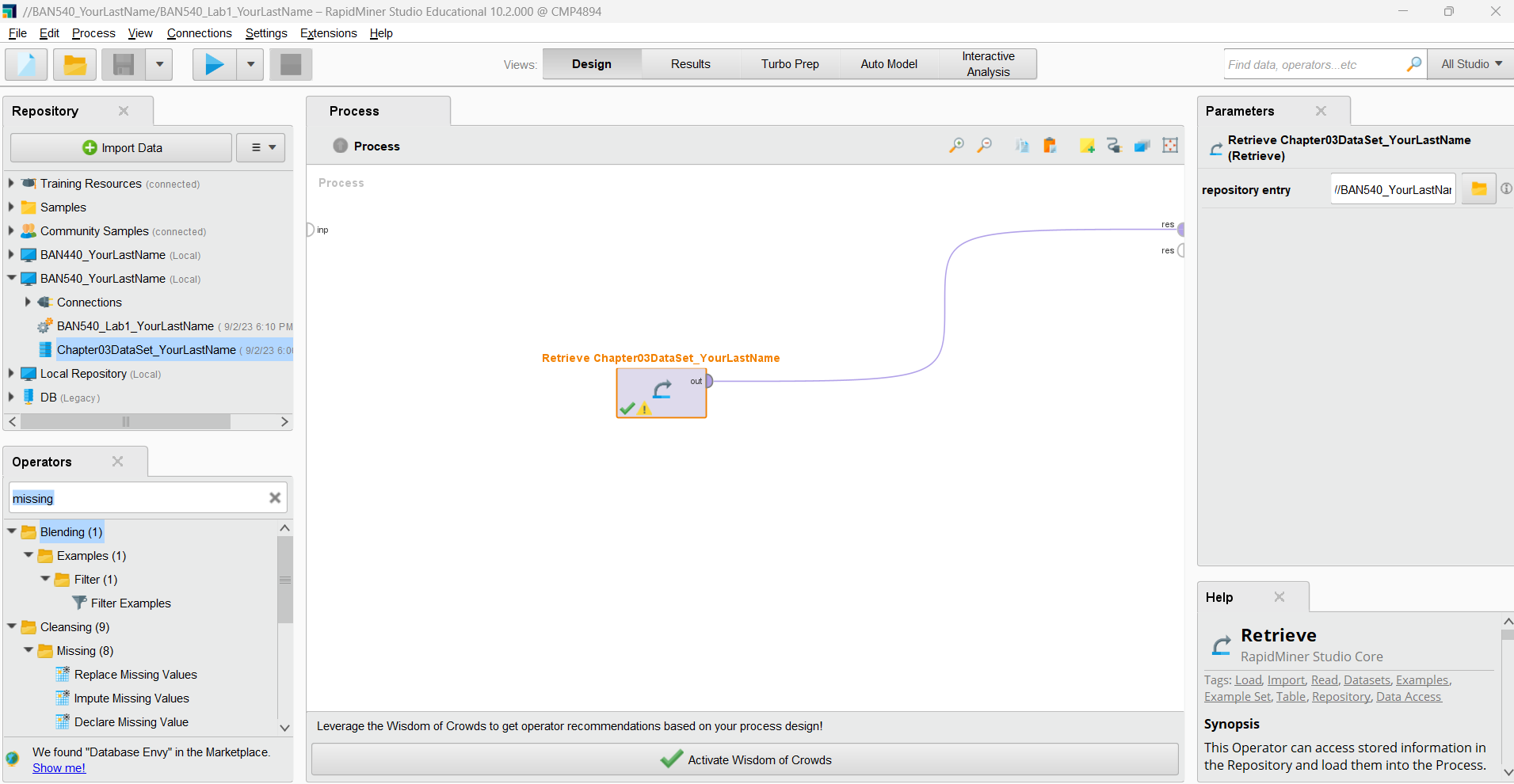


10) You can toggle between **design** and **results** perspectives by clicking on “**Design**” or “**Result**.”

**D. REPLACE MISSING VALUES**

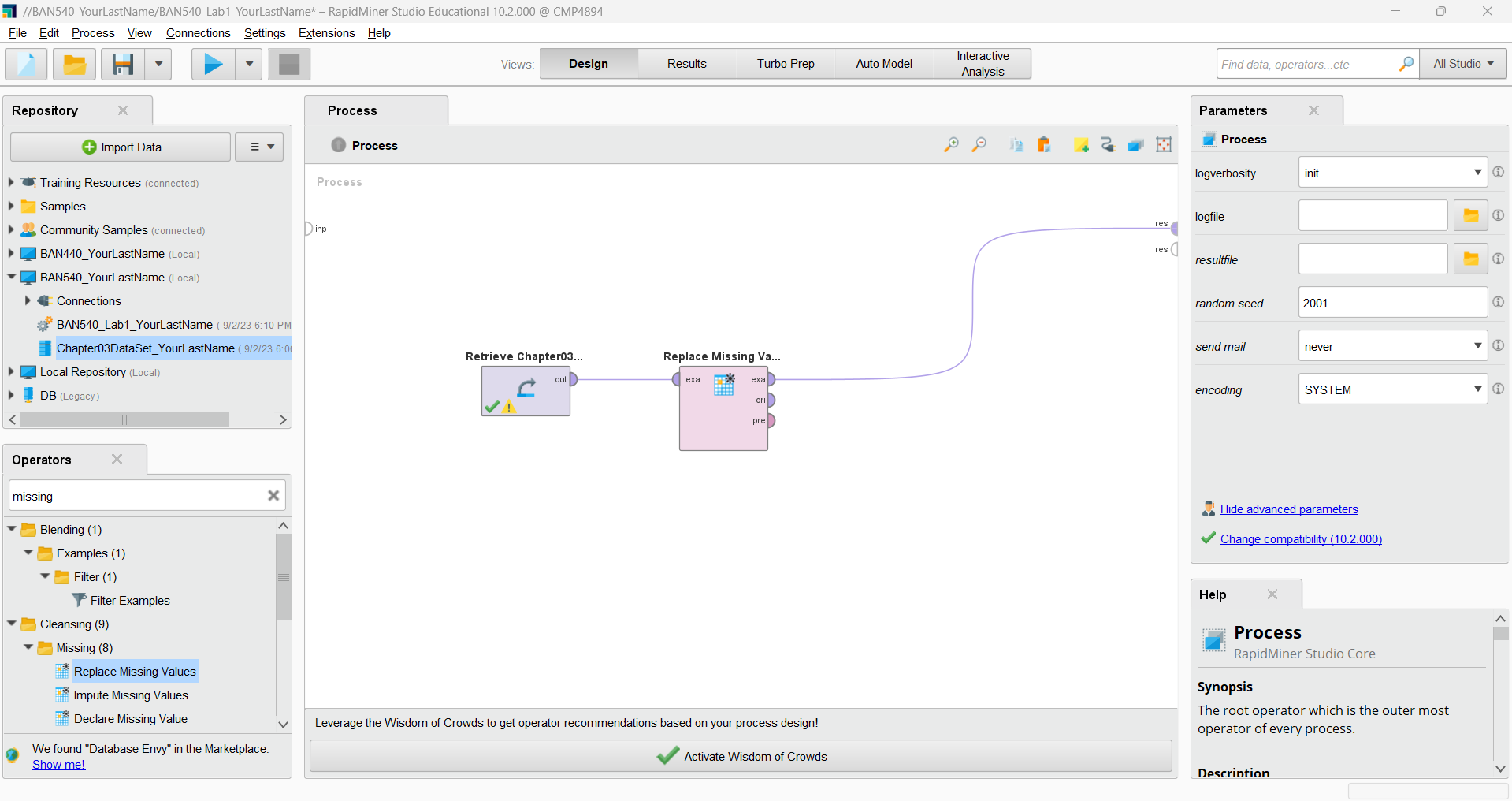
1) In order to find a tool (or an **operator**) in the **Operators** area, you can navigate through the folder tree in the lower left-hand corner of the screen. RapidMiner offers many tools/operators and sometimes, finding the one you want can be tricky. There is a handy search box, indicated by the red rectangle in the screenshot below that allows you to type in key words to find tools/operators that might do what you need.

Type in the word ‘**missing**’ into this search box, and you will see that RapidMiner automatically searches for tools/operators containing this word in their names. We want to replace missing values, and we can see that there is an operator called **Replace Missing Values**.



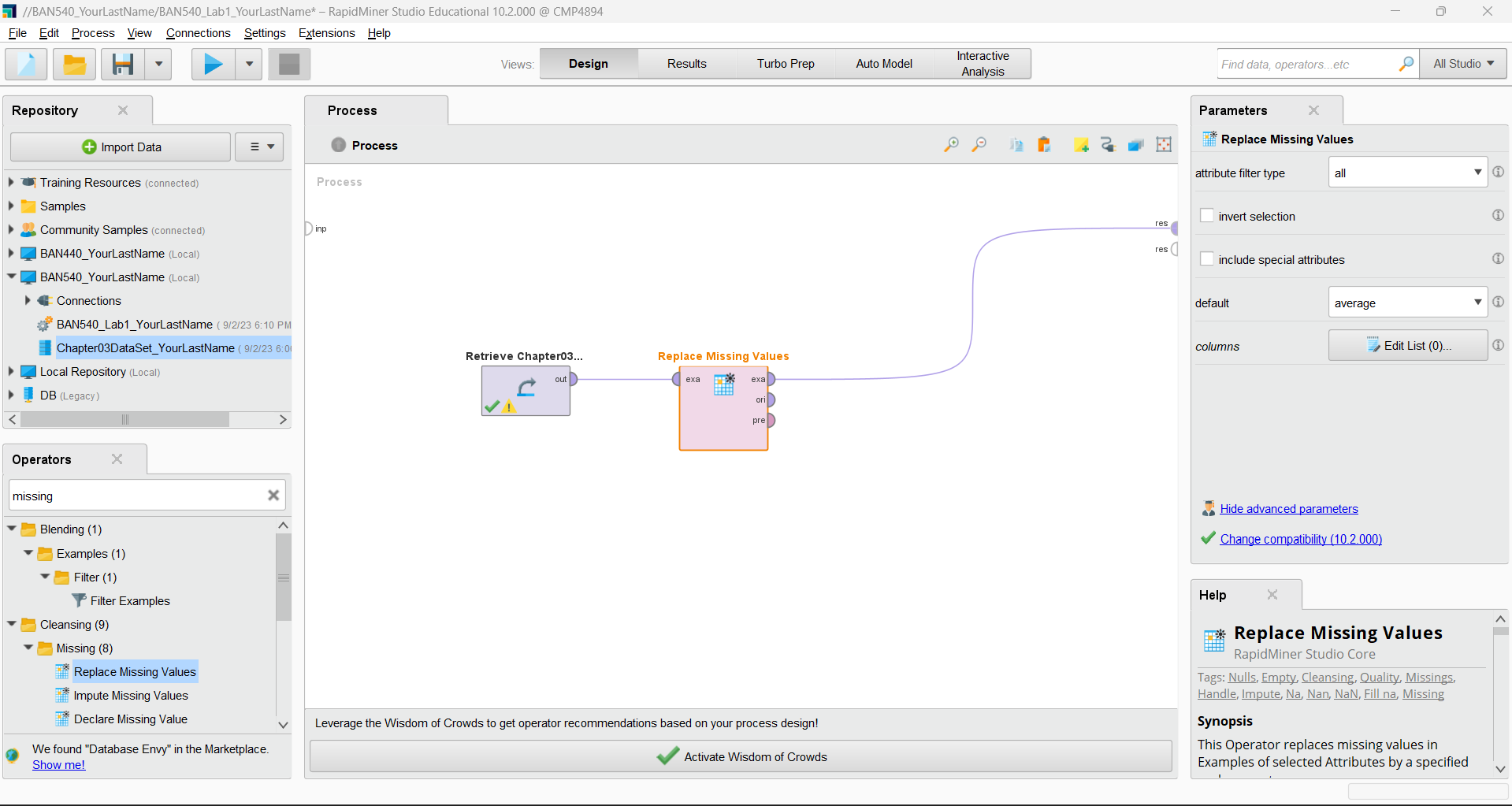
2) Now, let’s add this operator to our stream. Please click and hold on the operator name (from the left-hand side Operators pane), and drag it up to your spline. When you point your mouse cursor on the spline, the spline will turn slightly bold, indicating that when you let go of your mouse button, the operator will be connected into the stream.

If you let go and the **Replace Missing Values** operator fails to connect into your stream, you can reconfigure your splines manually. Simply click on the ***out***port in your **Retrieve** operator, and then click on the ***exa***port on the **Replace Missing Values** operator. *Exa* stands for example set, and ‘examples’ is the word RapidMiner uses for observations in a data set. Be sure the *exa* port from the Replace Missing Values operator is connected to your result set (*res*) port so that when you run your process, you will have output. Your model should now look similar to the screenshot below.



**Please make a screenshot now and replace my screenshot #1 with yours in the submission file (named as “BAN540 Lab 1 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 box in the above screenshot.)**

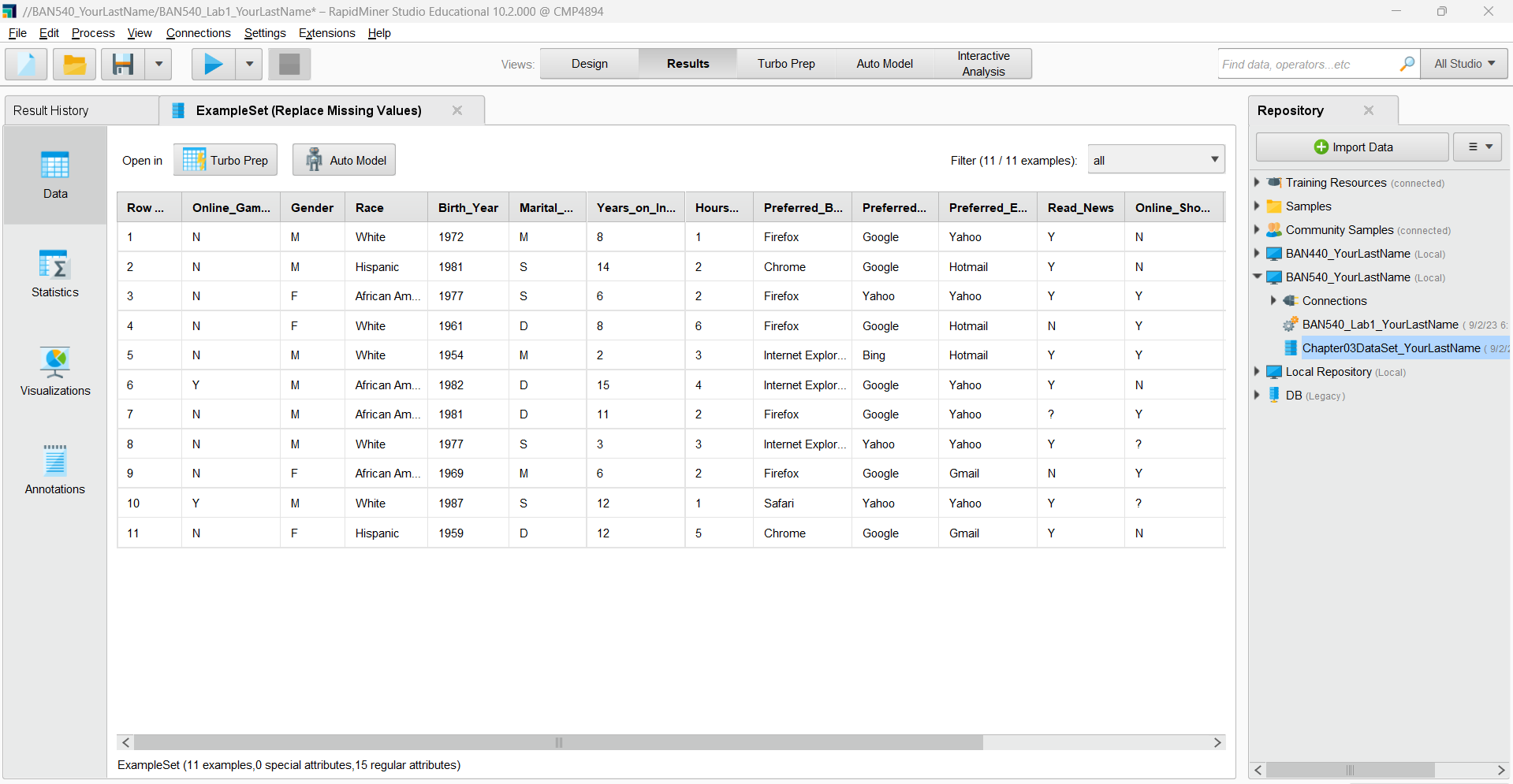
3) When an operator is selected in RapidMiner, it has an **orange rectangle around it**. This will also enable you to modify that operator’s **parameters**, or properties. The Parameters pane is located on the right side of the RapidMiner window (see below).



4) For this assignment, we have decided to change all missing values in the **Online\_Gaming** attribute to ‘**N**’, since this is the most common response in that attribute. To do this, please make sure the **Replace Missing Values** operator is selected (with the orange border), and then change the ‘attribute filter type’ to ‘**single**.’ Then, and you will see a dropdown box appears under it (for ‘attribute’), allowing you to choose the **Online\_Gaming** attribute as the target for modification. Next, expand the ‘default’ dropdown box, and select ‘**value’**, which will cause a ‘replenishment value’ box to appear. Type the replacement value ‘**N**’ in this box. Note that you may need to expand your RapidMiner window, or use the vertical scroll bar on the left of the Parameters pane in order to see all options, as the options change based on what you have selected. When you are done, your parameters should look like below.



5) Please note that there are many other options available to you in the parameters pane. We will not explore all of them here, but feel free to experiment with them. For example, instead of changing a single attribute at a time, you could change a subset of attributes in your data set. You will learn much about the flexibility and power of RapidMiner by trying out different tools and features. When you have your parameter set, click the **Play** button. This will run your process and switch you to **Results** perspective once again. Your results should look below.



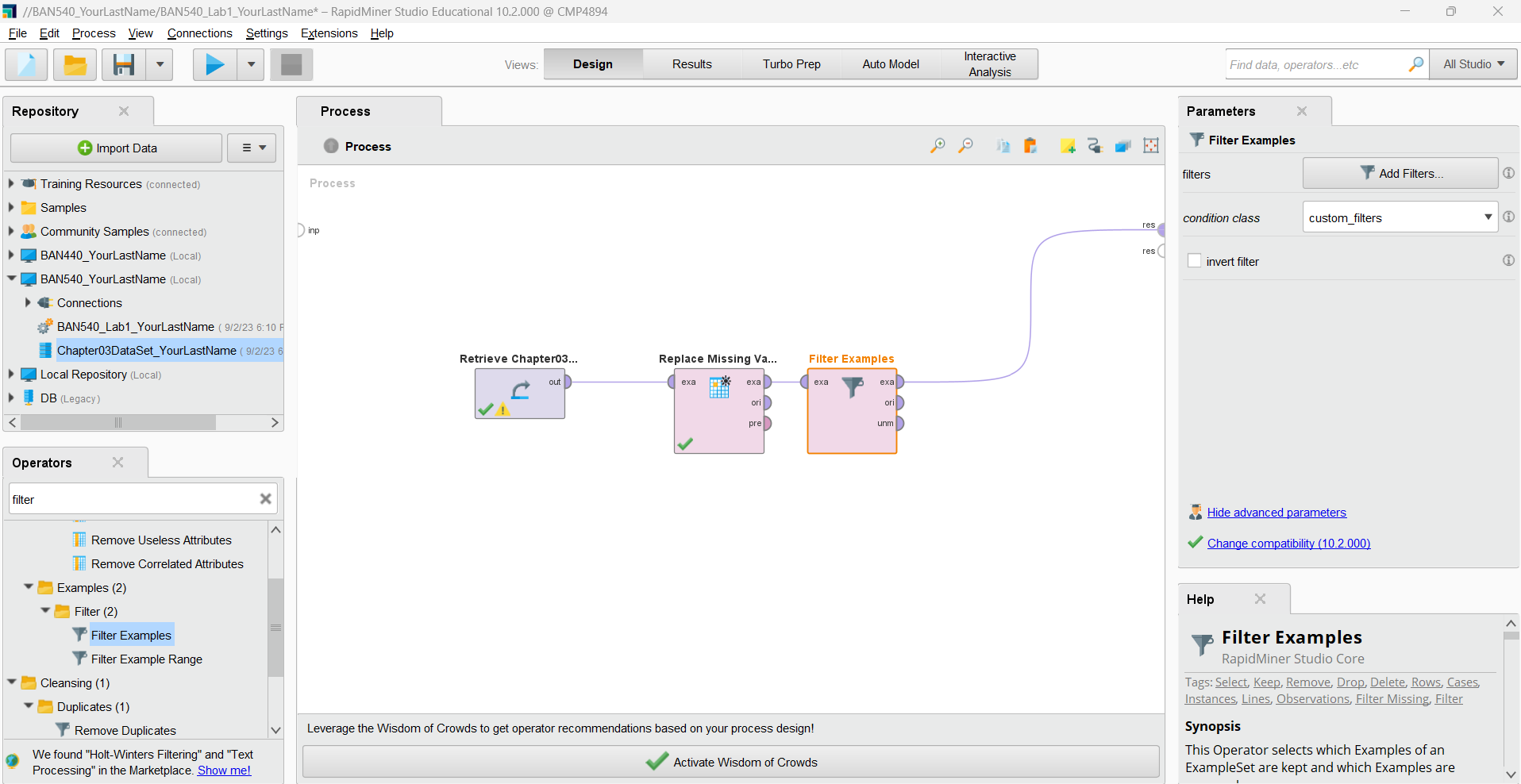
**Please make a screenshot now and replace screenshot #2 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”). Please make sure to show the Online\_Gaming attribute in your screenshot (see the red box), by scrolling to the very left.**

As you can see, now the **Online\_Gaming** attribute has been moved to the very left side of the attributes, and there are no missing values. All missing values for Online\_Gaming have been replaced by “N.” Now, let’s look at the **Online\_Shopping** attribute. A question mark (**?**) denotes a missing value in an observation. For this variable, suppose we do not wish to replace the null values with the mode, but rather, we wish to remove those observations from our data set prior to mining it. This can be accomplished through data reduction.

**E. DATA REDUCTION**

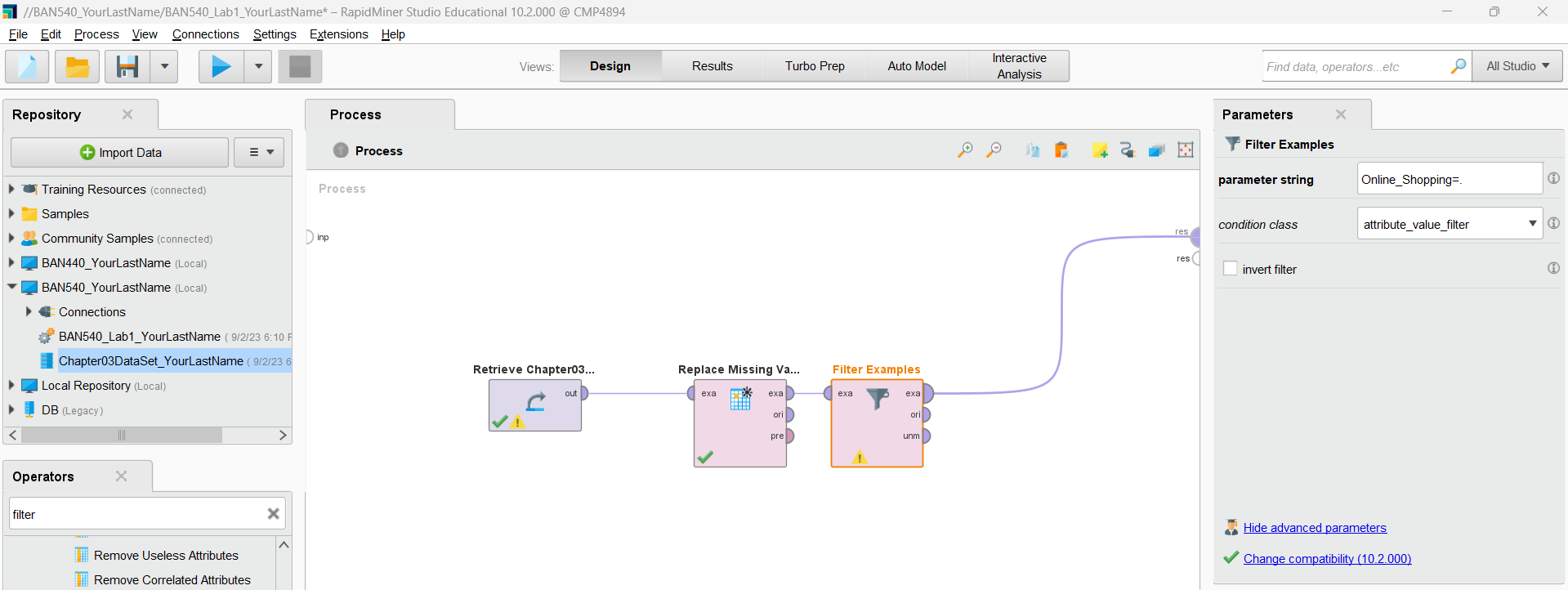
Go ahead and switch back to **Design** perspective. The next set of steps will teach you to reduce the number of observations in your data set through the process of **filtering**.

1) In the search box of the **Operators** pane, type in the word ‘**filter**’. This will help you locate the ‘**Filter Examples**’ operator, which is what we need to use. Drag the **Filter Examples** operator over and connect it into your stream, right after the **Replace Missing Values** operator. Your window should look like the following screenshot.

****

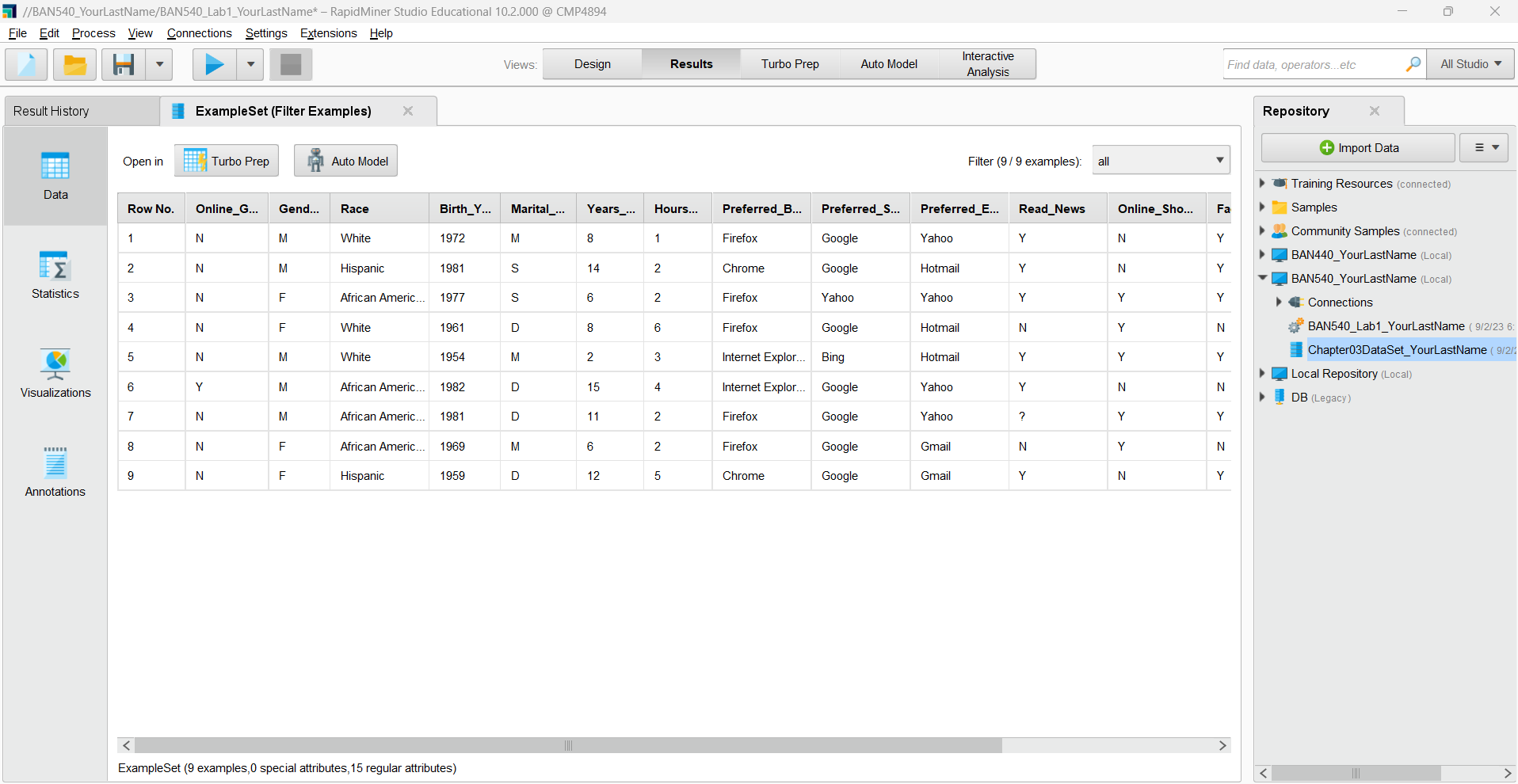
**Please make a screenshot now and replace screenshot #3 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”). Please make sure your screenshot shows your own last name in the related items in the Repository pane.**

2) In the ‘condition class’ dropdown box of the **Parameters** pane, choose ‘**attribute\_value\_filter**’, and in the ‘parameter\_string’ text box, type in the following: **Online\_Shopping=.** Be sure to include the period. This parameter string refers to our attribute, Online\_Shopping, and it tells RapidMiner to filter out all observations where the value in that attribute is missing. This is a bit confusing, because in Data View in Results perspective, missings are denoted by a question mark (?), but when entering the parameter string, missings are denoted by a period (.). Once you’ve typed these parameter values in, your screen should look like below.



3) Go ahead and run your model by clicking the **Play** button. In Results perspective, you will now see that your data set has been reduced from eleven observations (or examples) to nine. This is because the two observations where the Online\_Shopping attribute had a missing value have been removed. You’ll be able to see that they are gone by selecting the Data View. They have not been deleted from the original source data, they are simply removed from the data set at the point in the stream where the filter operator is located and will no longer be considered in any downstream data mining operations.

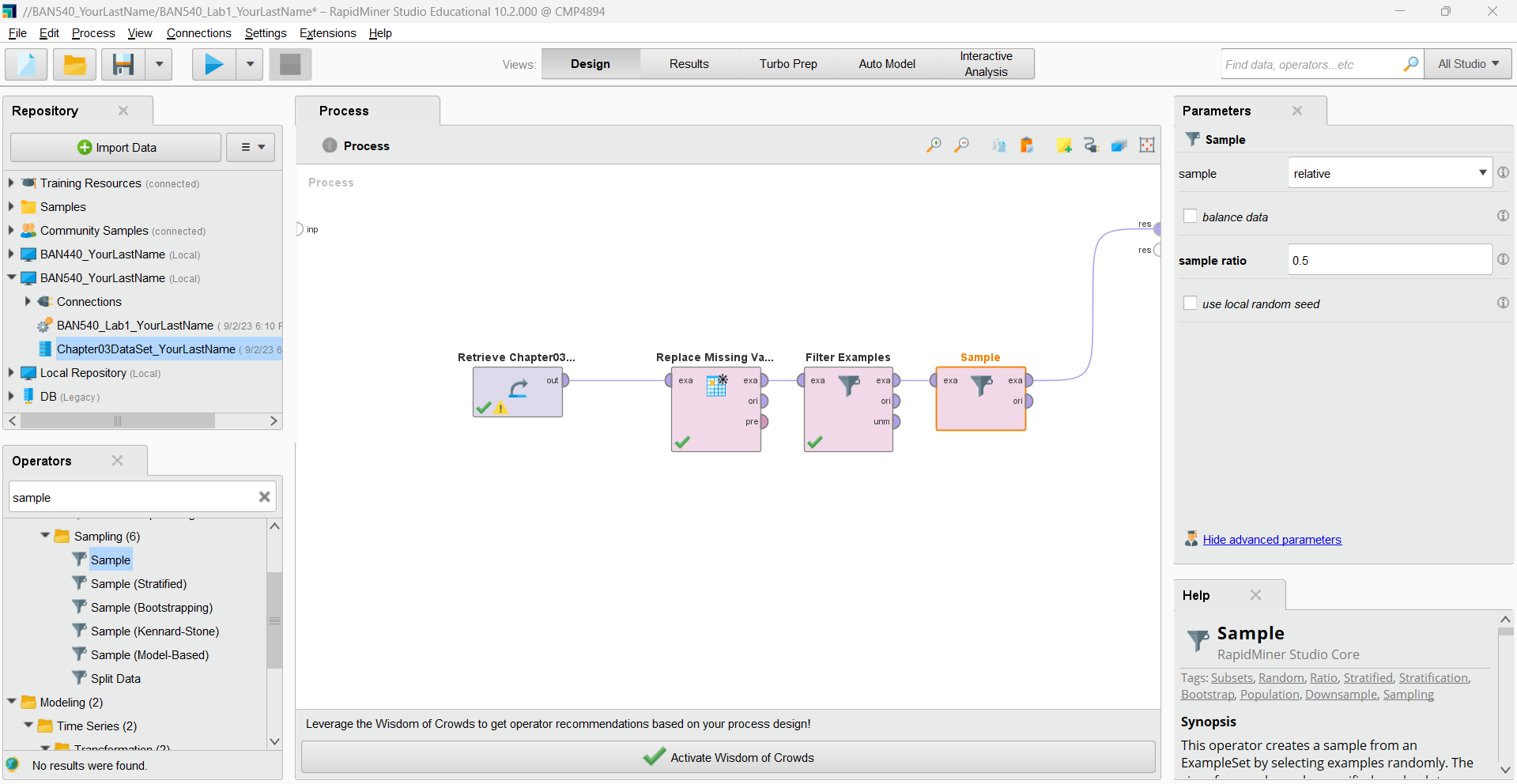
In instances where the missing value cannot be safely assumed or computed, removal of the entire observation is often the best course of action. When attributes are numeric in nature, such as with ages or number of visits to a certain place, an arithmetic measure of central tendency, such as **mean, median** or **mode** might be an acceptable replacement for missing values, but in more subjective attributes, such as whether one is an online shopper or not, you may be better off simply filtering out observations where the datum is missing.



**Please make a screenshot now and replace screenshot #4 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”).**

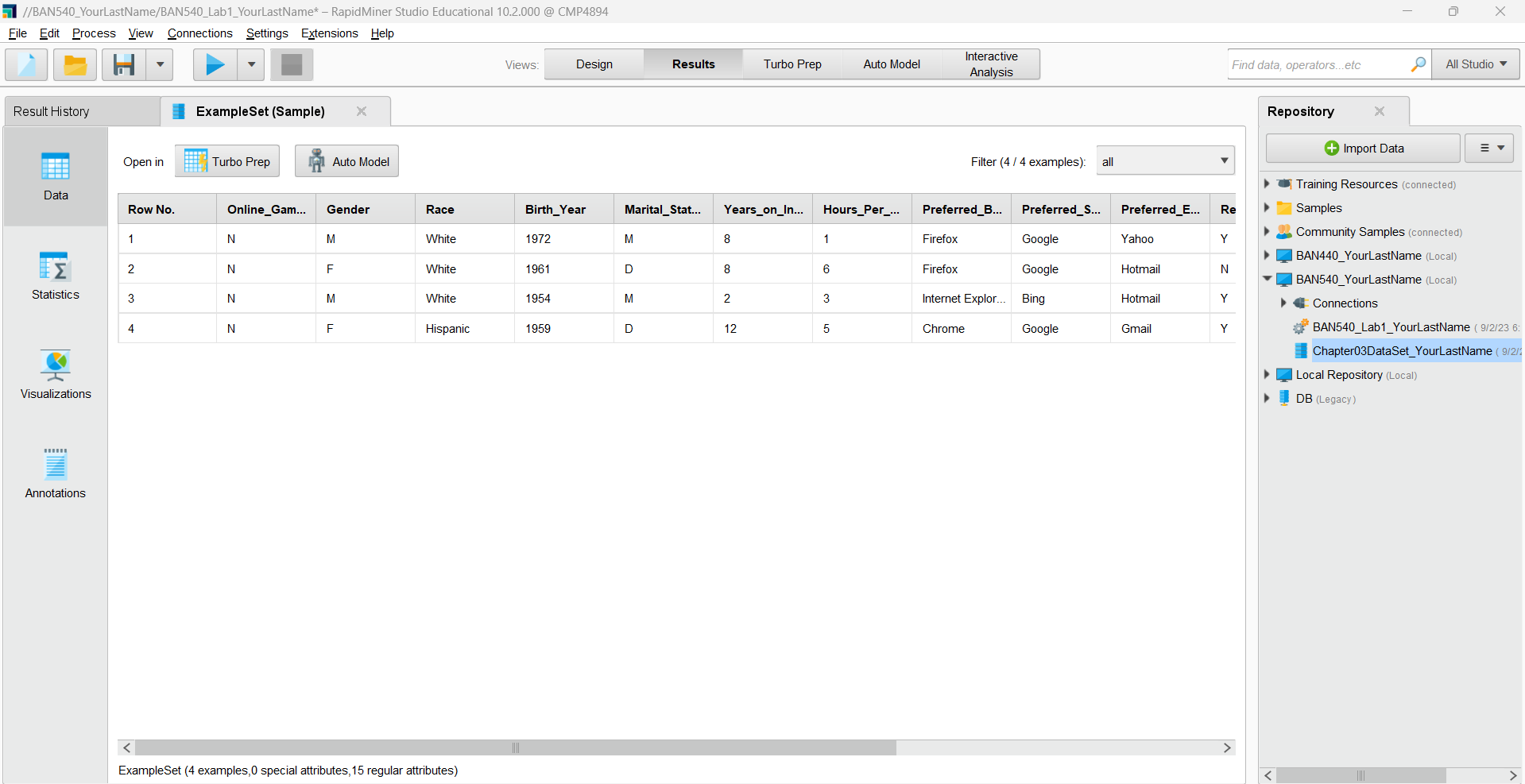
Data mining can be confusing and overwhelming, especially when data sets get large. It doesn’t have to be though, if we manage our data well. The previous example has shown how to filter out observations containing undesired data (or missing data) in an attribute, but we can also reduce data to test out a data mining model on a smaller subset of our data. This can greatly reduce processing time while testing a model to see if it will work to answer our questions. Follow the steps below to take a **sample** of our data set in RapidMiner.

4) Leveraging the search techniques previously demonstrated, use the **Operators** search feature to find an operator called ‘**Sample**’ and add this to your stream. In the **Parameters** pane, set the sample to be a ‘**relative**’ sample, and then indicate you want to retain 50% of your observations in the resulting data set by typing **.5** into the ‘sample ratio’ field. Your window should look like below.



**Please make a screenshot now and replace screenshot #5 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”). Please make sure your screenshot shows your own last name in the related items in the Repository pane.**

5) When you run your model, you will find that your results only contain four or five observations, randomly selected from the nine that were remaining after our filter operator removed records that had missing Online\_Shopping values.

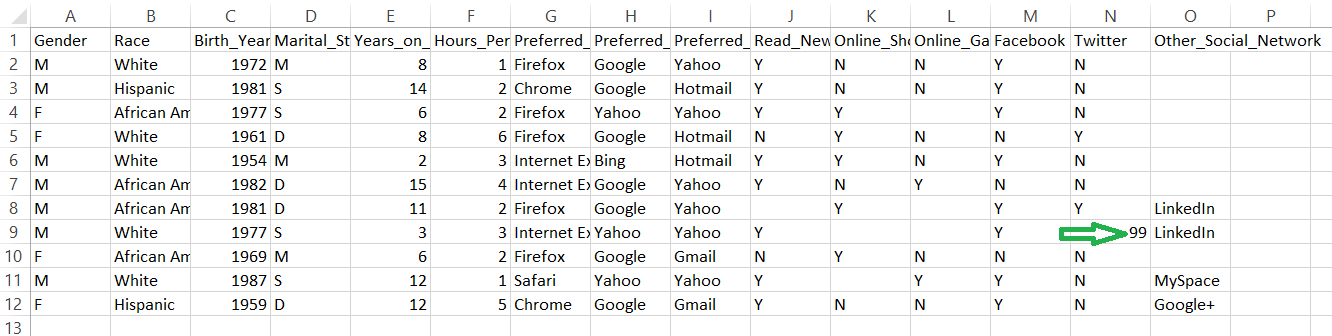


**Please make a screenshot now and replace screenshot #6 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”).**

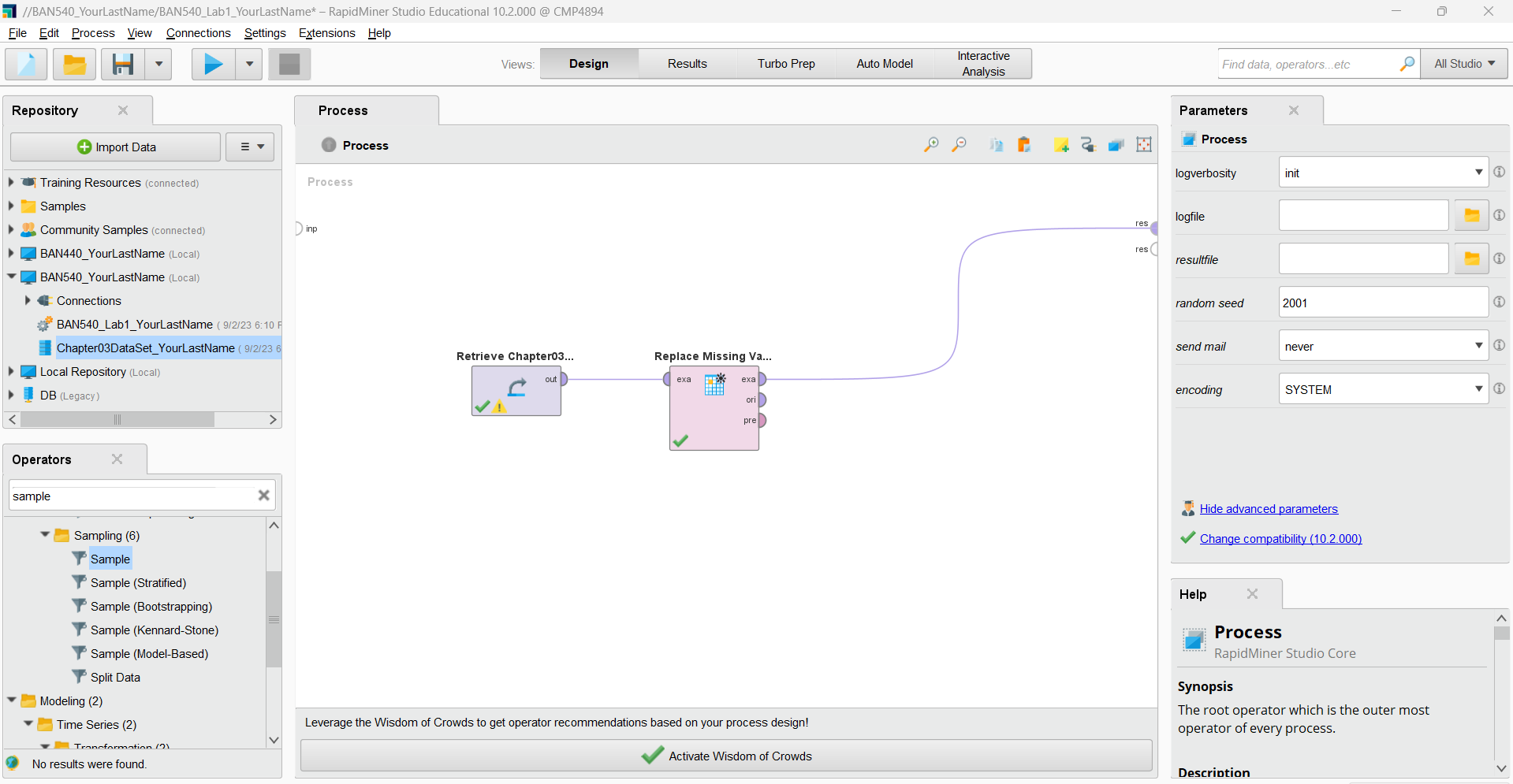
There are many ways, and various reasons to reduce data by decreasing the number of observations in your data set. We’ll now move on to handling inconsistent data, but before doing so, it is going to be important to reset our data back to its original form. While filtering, we removed an observation that we will need in order to illustrate what inconsistent data is, and to demonstrate how to handle it in RapidMiner. This is a good time to learn how to remove operators from your stream. To do it, switch back to **Design** perspective and click on your **Sampling** operator. Next, right click on it and choose **Delete**, or simply press the Delete key on your keyboard. Delete the **Filter** **Examples** operator at this time as well. Note that your spline that was connected to the *res* port may also be deleted. This is not a problem, you can reconnect the *exa* port from the **Replace Missing Values** operator to the *res* port, or you will find that the spline will reappear when you complete the steps under Handling Inconsistent Data.

**F. HANDLING INCONSISTENT DATA**

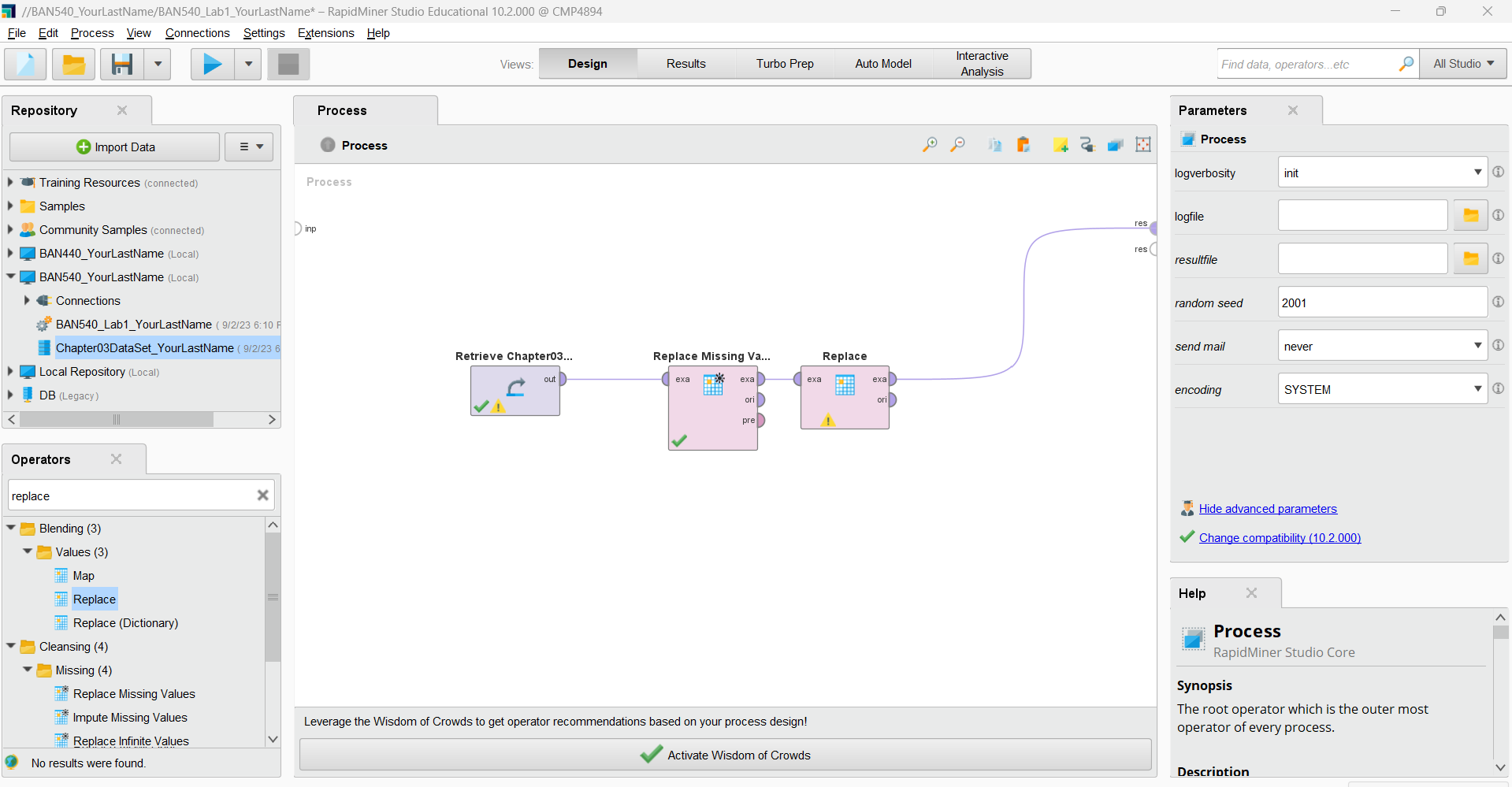
**Inconsistent data** is different from missing data. Inconsistent data occurs when a value does exist, however that value is not valid or meaningful. Please look at the following screenshot of our original dataset, what is that 99 doing there? It seems that the only two valid values for the Twitter attribute should be ‘Y’ and ‘N’. This is a value that is inconsistent and is therefore meaningless. As data miners, we can decide if we want to filter this observation out, as we did with the missing Online\_Shopping records, or, we could use an operator designed to allow us to replace certain values with others.



1) Return to **Design** perspective if you are not already there. Ensure that you have deleted your sampling and filter operators from your stream, so that your window looks like the following screenshot.

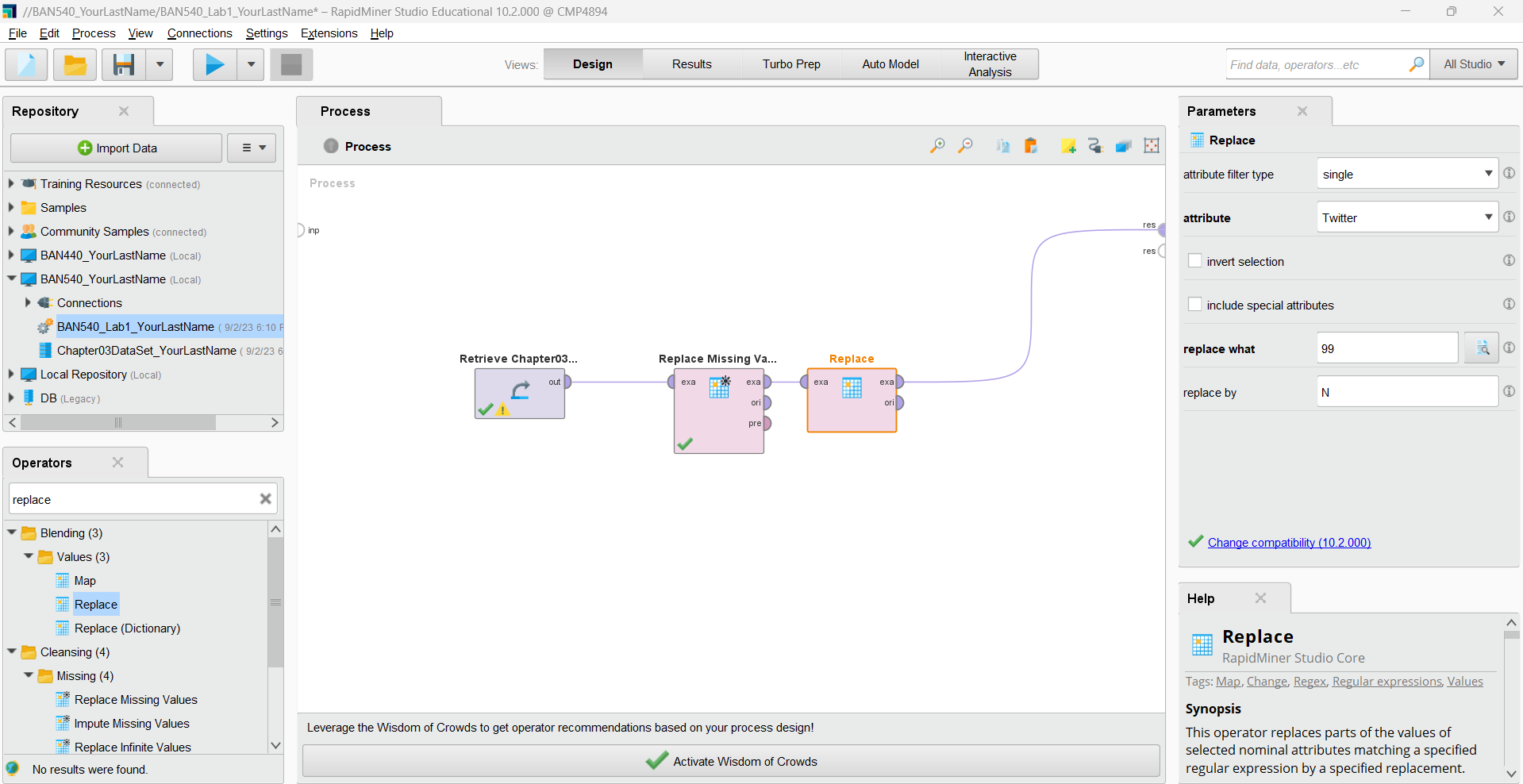


2) Note that we don’t need to remove the **Replace Missing Values** operator, because it is not removing any observations in our data set. It only changes the values in the Online\_Gaming attribute, which won’t affect our next operator. Use the **search** feature in the **Operators** tab to find an operator called **Replace**. Drag this operator into your stream.

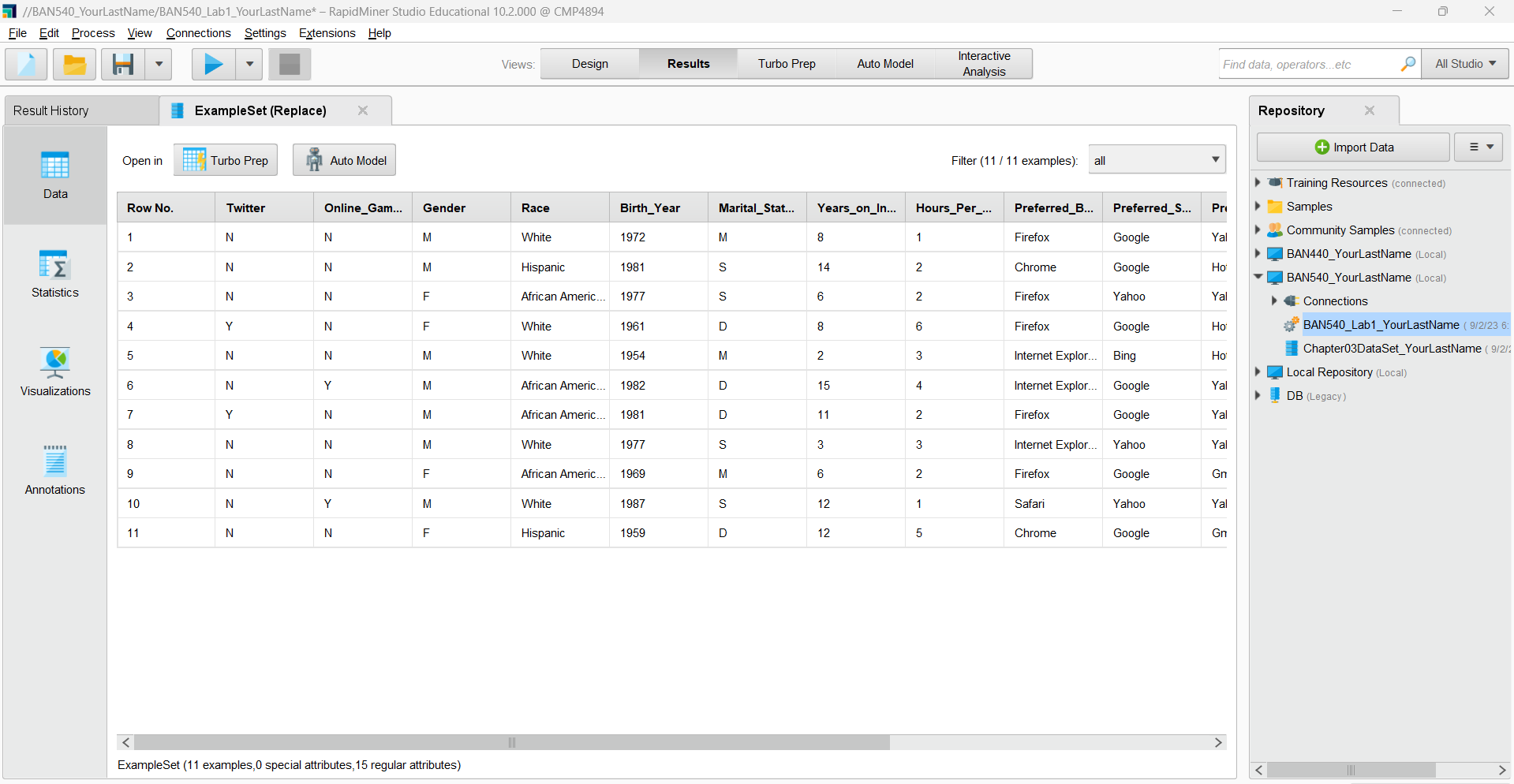


**Please make a screenshot now and replace screenshot #7 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”). Please make sure your screenshot shows your own last name in the related items in the Repository pane.**

3) In the **Parameters** pane of the Replace operator (making sure to select it first), change the ‘attribute filter type’ to **single**, then indicate **Twitter** as the ‘attribute’ to be modified. In truth, in this data set there is only one instance of the value 99 across all attributes and observations, so this change to a single attribute is not actually necessary in this example, but it is good to be thoughtful and intentional with every step in a data mining process. Most data sets will be far larger and more complex than the data set we are currently working with. In the ‘replace what’ field, type the value **99**, since this is the value we’re looking to replace. Finally, in the ‘replace by’ field, we must decide what we want to have in the place of the 99. If we leave this field blank, then the observation will have a missing (?) when we run the model and switch to Data View in results perspective. We could also choose the mode of ‘N’, and given that 80% of the survey respondents indicated that they did not use Twitter, this would seem a safe course of action. You may choose the value you would like to use. For now, we will enter ‘**N.**’



Now, please run it. You will then see that there are nine values of ‘N’, and two of ‘Y’ now for the Twitter attribute.



**Please make a screenshot now and replace screenshot #8 with yours in the submission file (named as “BAN540 Lab 1 Submission YourLastName.docx”).**

Keep in mind that not all inconsistent data are going to be as easy to handle as replacing a single value. It would be entirely possible that in addition to the inconsistent value of 99, values of 87, 96, 101, or others could be present in a data set. If this were the case, it might take multiple replacements and/or missing data operators to prepare the data set for mining. In numeric data we might also come across data which are accurate, but which are also statistical outliers. These might also be considered to be inconsistent data. Sometimes data scrubbing can become tedious, but it will ultimately affect the usefulness of data mining results, so these types of activities are important, and attention to detail is critical.