# Introduction

90% of the data in the world was generated in the last two years alone. With the emergence of Big Data, the need for processing and developing it has been unprecedented. Data mining allows us to discover patterns, insights, and relationships within mountains, tsunamis, and landslides of data. In this project, we experiment with various techniques of data mining to create a model for a sample dataset.

# Steps of Data Mining

* 1. Data Cleaning
  2. Data Integration
  3. Data Selection
  4. Data Transformation
  5. Data Mining (Methods, Modeling)
  6. Model Evaluation
  7. Knowledge Representation

# Understanding Data

* 1. There are two datasets in Chinese, one is in Excel and the other is in CSV format. The paper questionnaire and the Tencent questionnaire have different columns and formats, so we need to take note of that when cleaning the data later.
  2. Features
     1. Name (m\_name): The name of the investigator after anonymization
     2. Education (eduction): the educational information of the investigator.
     3. Major: The professional information of the investigator.
     4. Age: The age of the investigator.
     5. Working years (work\_seniority): The working years of the investigator.
     6. Job position (m\_work): The job position of the investigator.
     7. Child nickname (nickname): The baby's nickname. Since the real name of the child is also more sensitive, only the nickname is filled in here.
     8. Child\_age: The age of the baby.
     9. Child sex (child\_sex): The sex of the baby.
     10. Do you think your children are easy to take care of (care\_level): 1-4, 1 means good belt, 4 means particularly bad.
     11. Your current state (state\_now): A. Working outside the home full-time B. Working outside the home part time C. Working at home part-time D. Bringing the baby full-time E. Others
     12. If you raise your own baby, the reason (myself\_care\_reason): A. No one will bring them B. Don’t worry about others bringing the baby C. I especially want to bring my own D. Others
     13. As a woman, which do you think is more important, work or life (m\_import): A. Work B. Life C. All are important D. I can't tell which one is important
     14. If someone else is taking care of the baby, who is that person (care\_child\_people): A. Mom and Dad B. Parents-in-law C. Other relatives D. Child-rearing sister-in-law
     15. If parents or parents-in-law want to take their children to their hometown, you don’t have to worry about anything, whether you agree or not (go\_home\_or\_not): A. agree B. disagree
     16. Have you considered learning some professional knowledge to improve your competitiveness in your free time at work and with your baby (study\_or\_not): A. Yes B. No
     17. 17. Have you learned any professional parenting experience (parenting\_knowledge\_or\_not): A. studied B. not studied
     18. Do you consider yourself a qualified mother (qualified\_mothers\_or\_not): A. Competent B. Incompetent
     19. If there is a website that provides work from home, the work does not delay taking care of the baby, and the salary level is also good, would you consider registering an account and finding a suitable job on it (work\_home\_or\_not): A. consider B. not consider
     20. As a mother, do you sometimes feel that you are in a low mood, like you have depression, and you need emotional counseling from others (mothers\_mood):A. Yes B. No, I have always been in a good mood

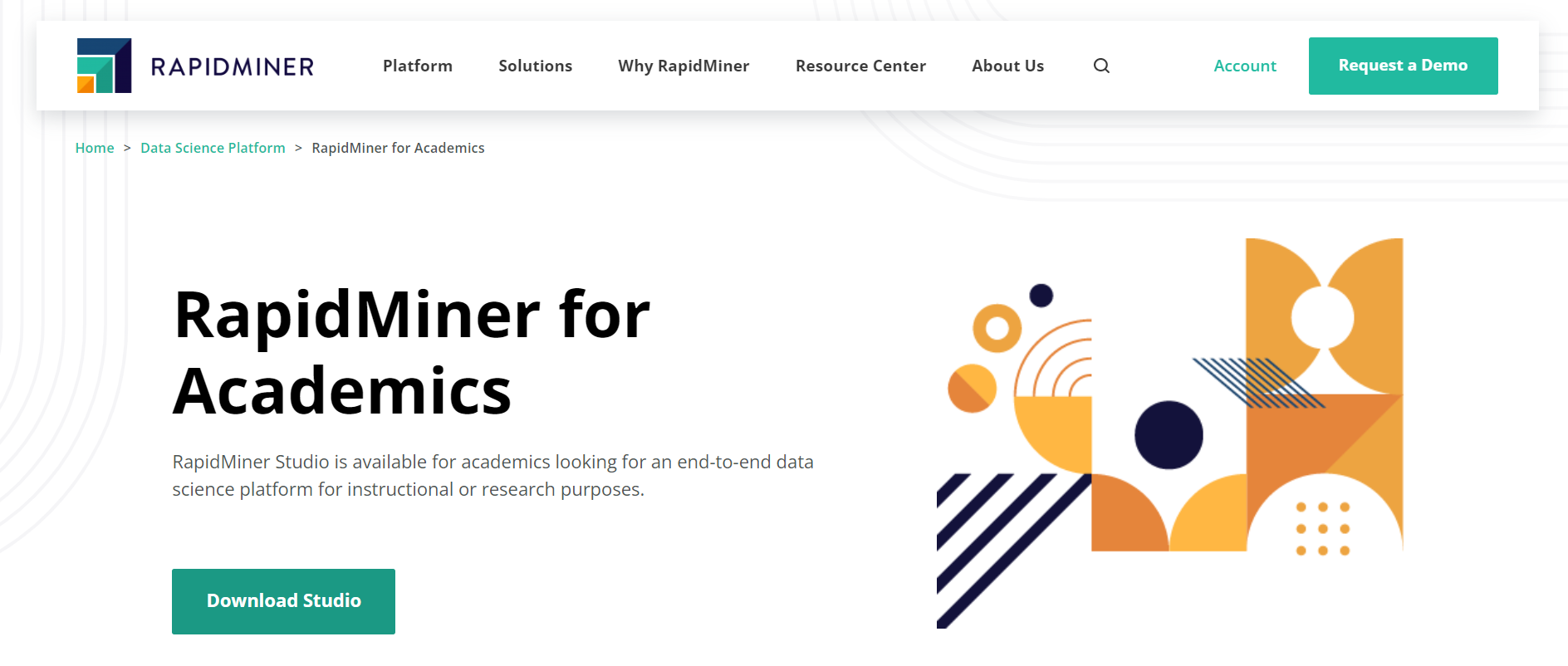
# RapidMiner

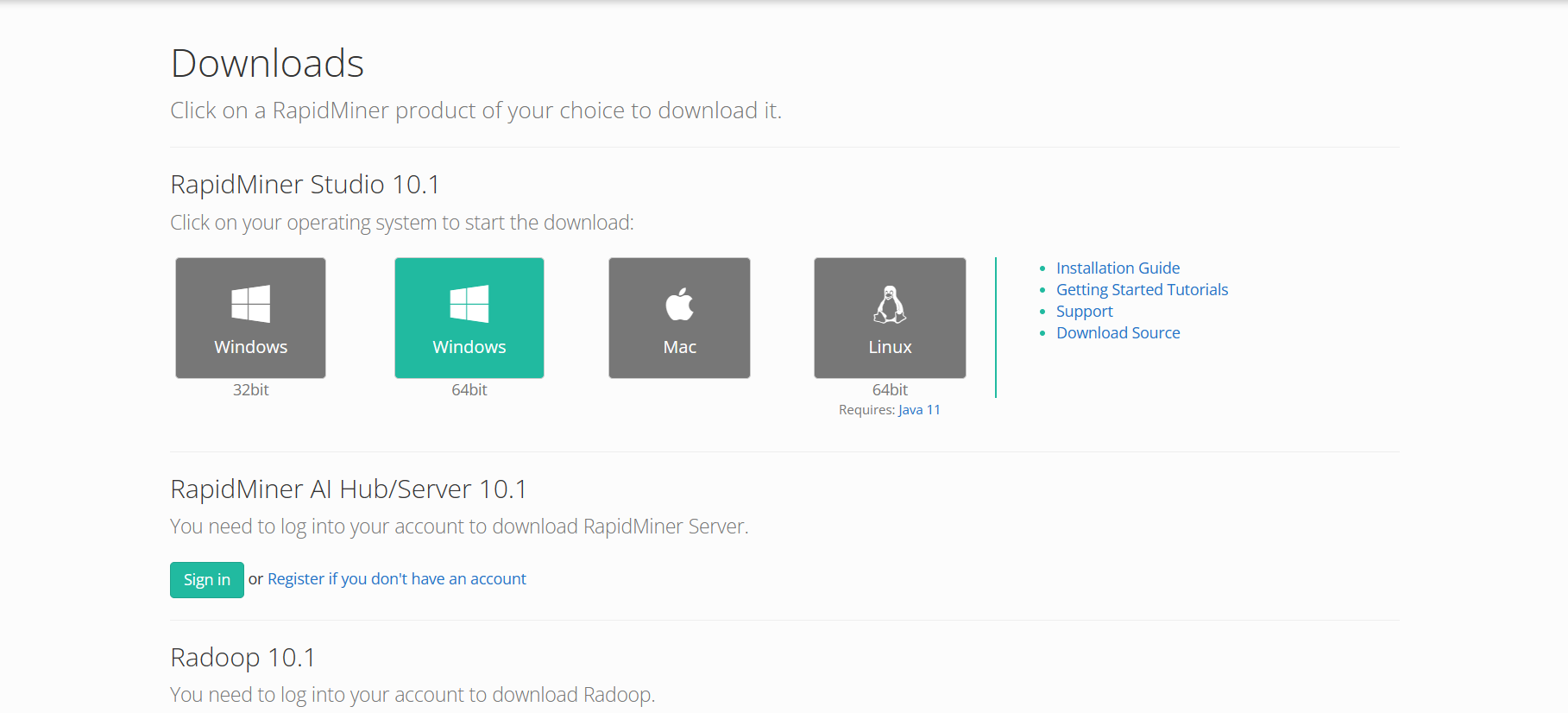
## Introduction to RapidMiner

* + 1. Rapidminer is a predictive analytics and data mining software. It is characterized by drag-and-drop operations, no programming required, fast computing speed, and open source and commercial versions. It has a rich collection of data mining analysis and algorithm functions that are often used to solve a variety of business critical problems. For example, typical business cases such as marketing response rate, customer segmentation, customer loyalty and lifetime value, asset maintenance, resource planning, predictive maintenance, quality management, social media monitoring and sentiment analysis. The solutions cover various fields, including automotive, banking, insurance, life sciences, manufacturing, oil and gas, retail and FMCG, communications, and utilities.

## RapidMiner Installation

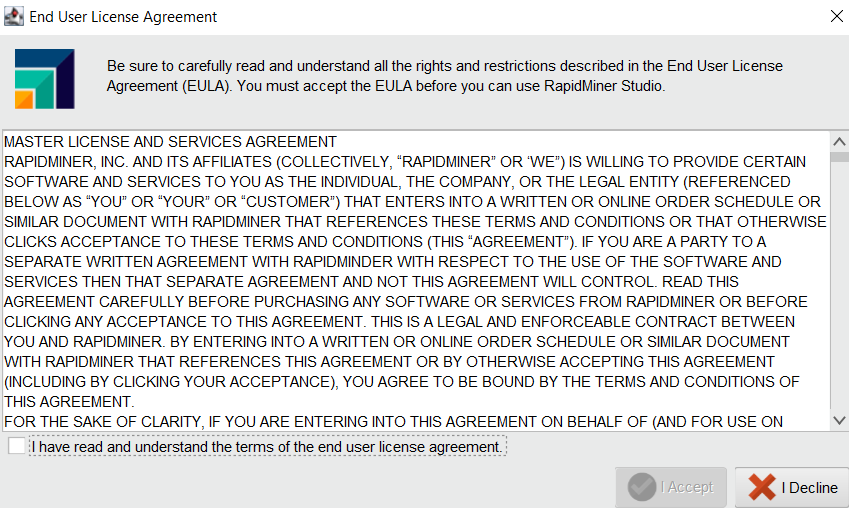
* + 1. Go to this link: <https://rapidminer.com/platform/educational/>
    2. Click “Download Studio, ” which opens a new tab.



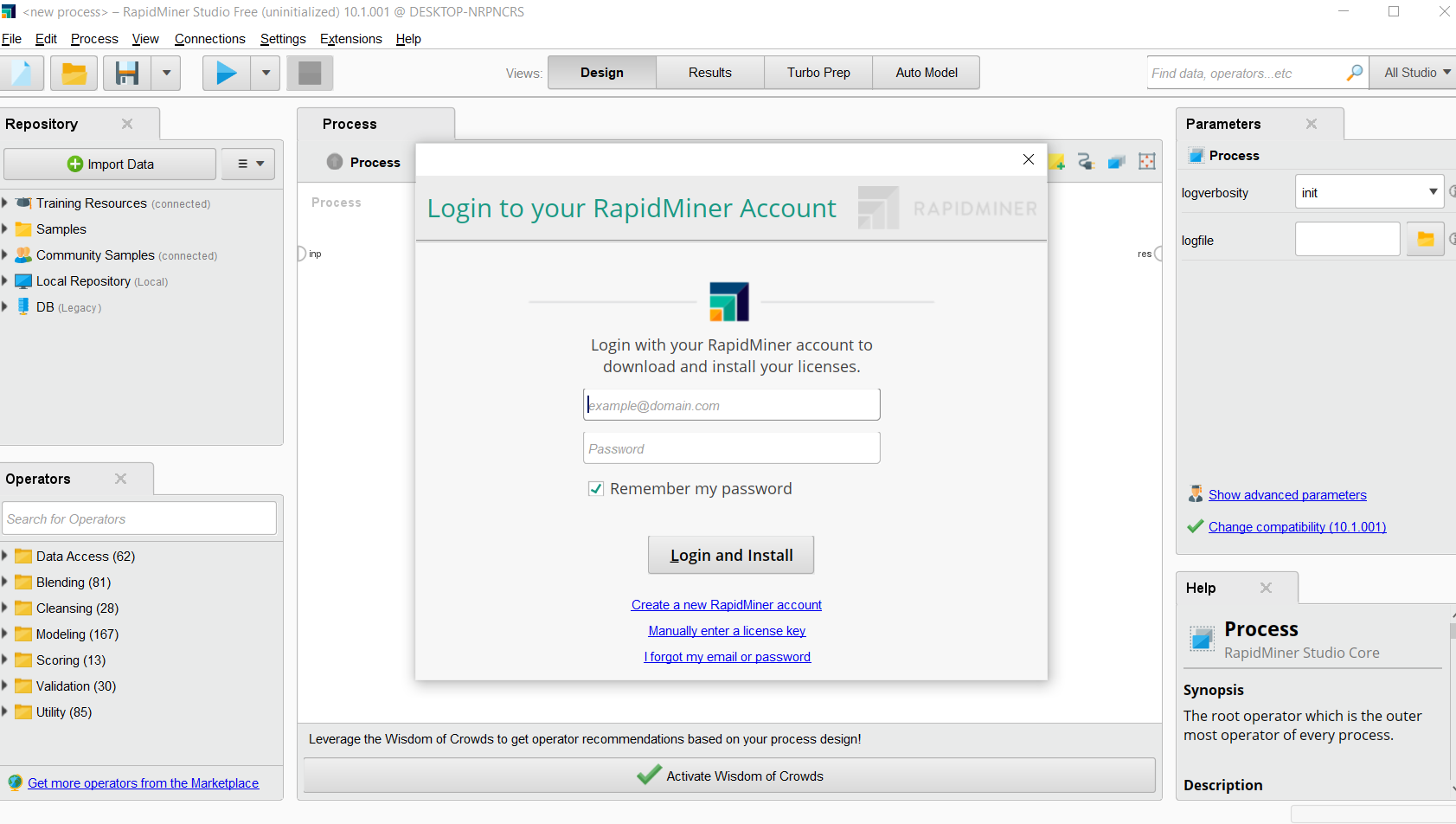
* + 1. Registration (skip to step 5 if you already have an account): Select *Educational Purposes* and fill in the corresponding information. When finished, click “Register.”
    2. A verification link will be sent to your email after signup, click verify.
    3. Log in to your RapidMiner account and go to this link: <https://my.rapidminer.com/nexus/account/index.html#downloads>. Select the download that corresponds with your operating system. 
    4. Click on the downloaded file, where a loading screen will show up as shown below.



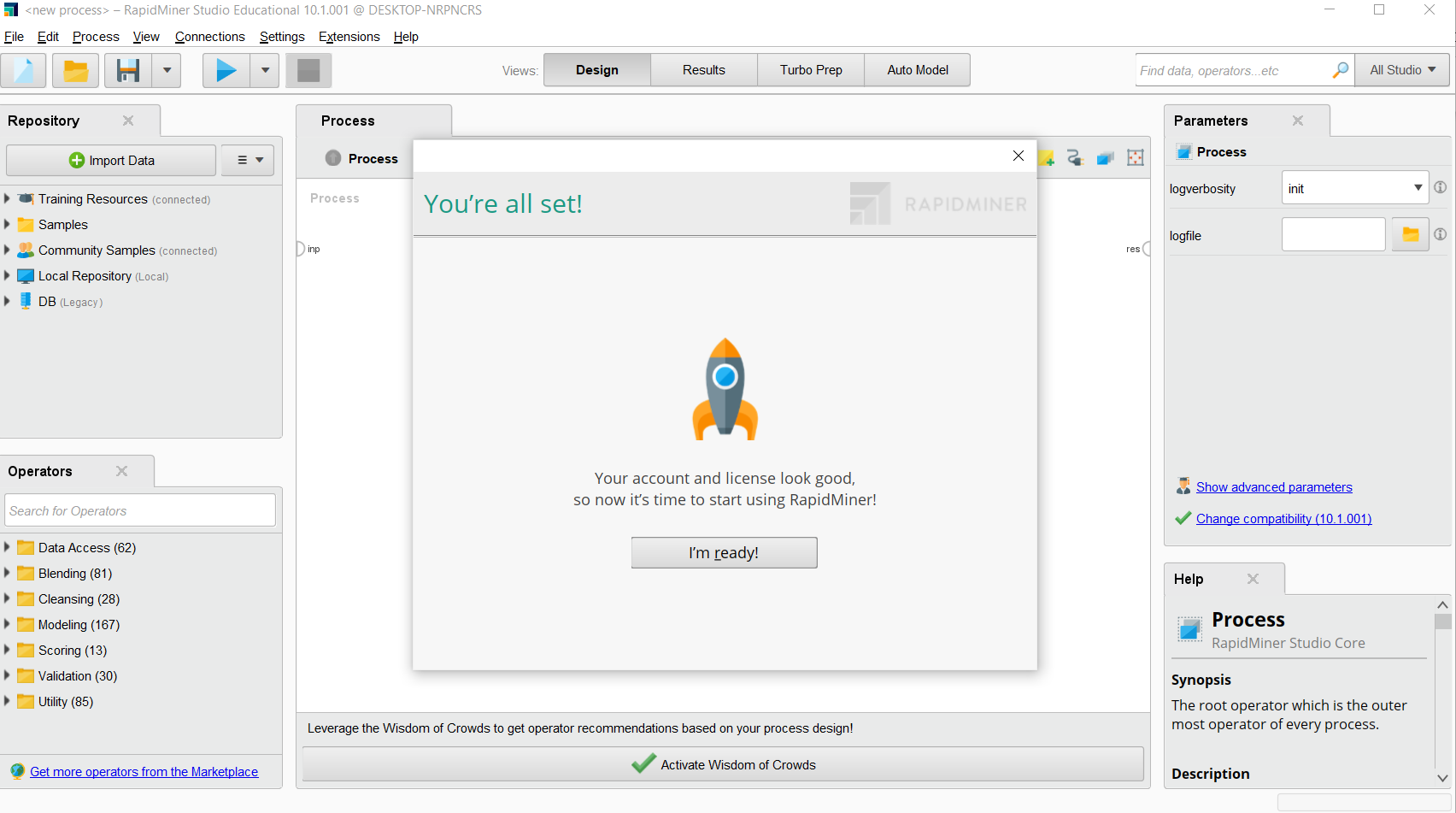
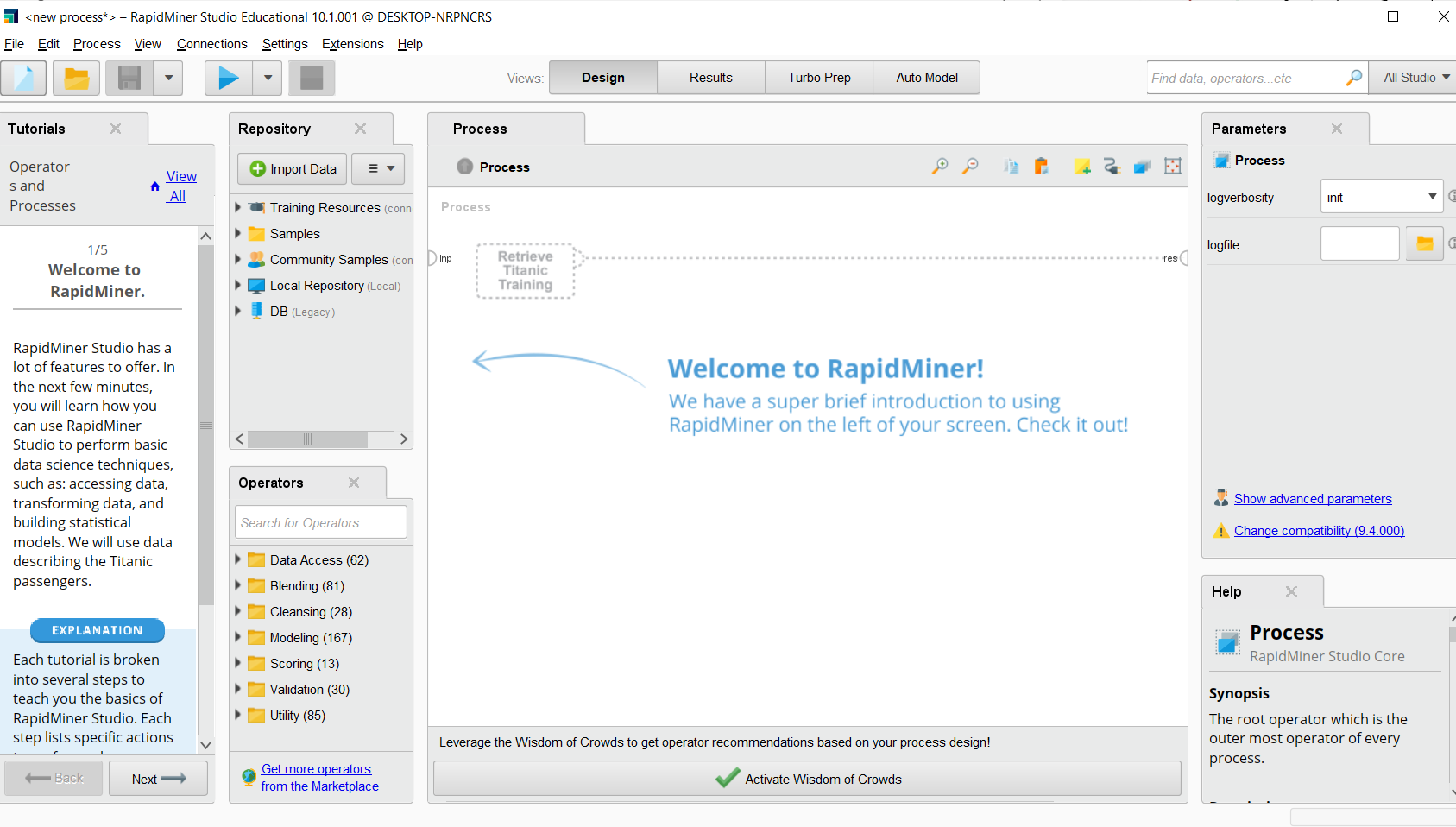
* + 1. Check “I have read and understand the terms of the end user license agreement”and click “I accept”



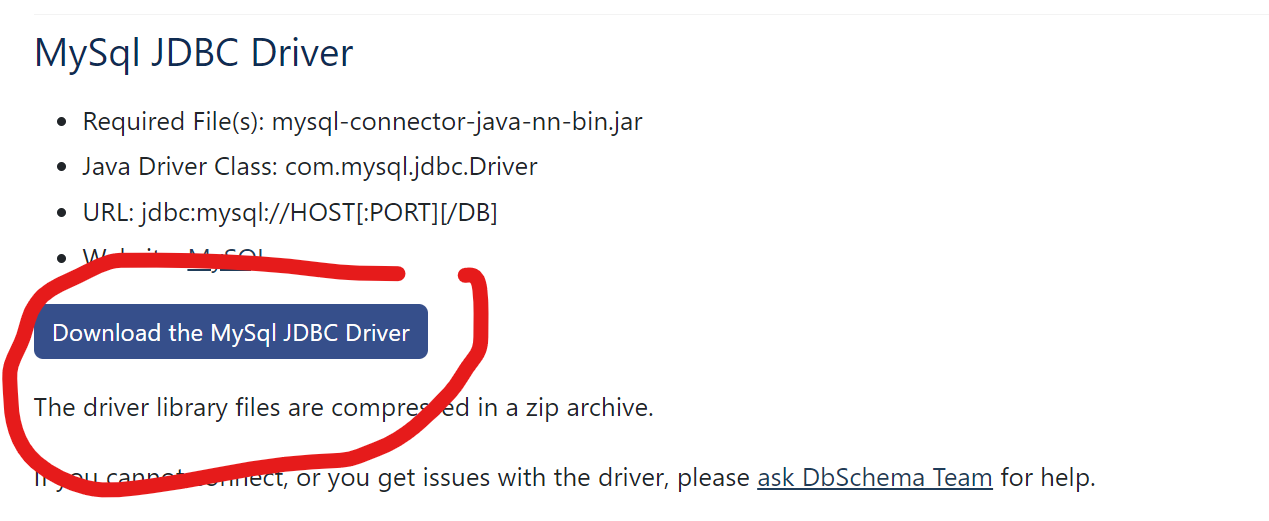
* + 1. Log in to your RapidMiner account.



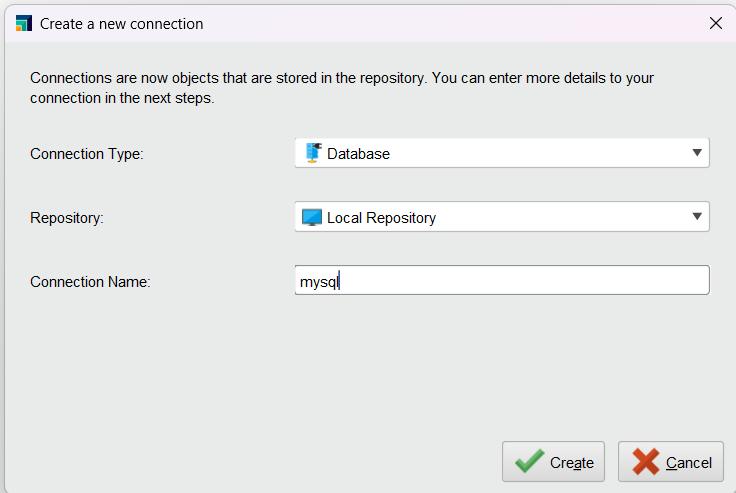
* + 1. If everything was successful, the following images should show up.



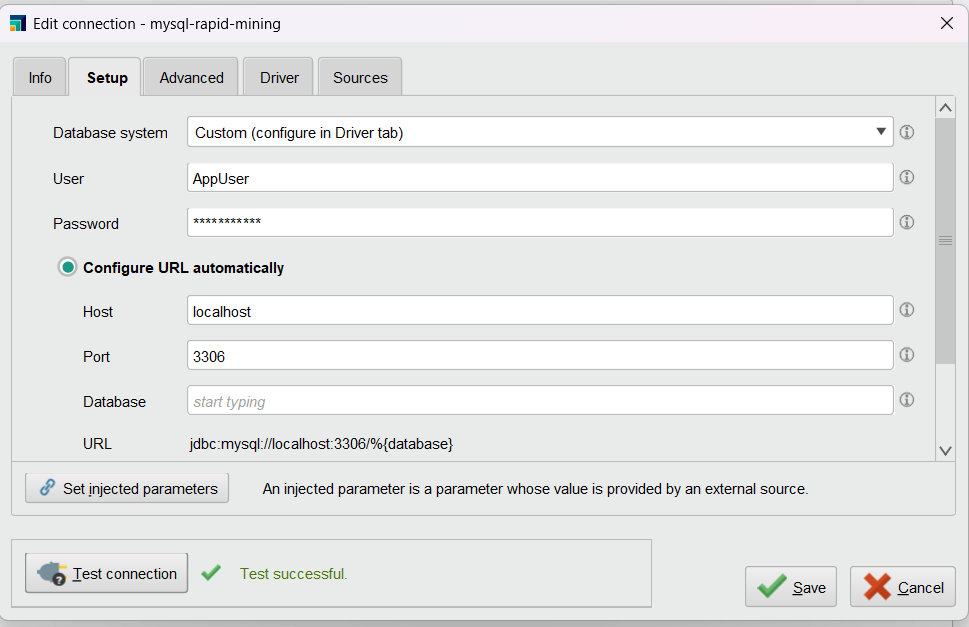
## RapidMiner Configuration

Install MySQL JDBC Driver if you do not have it yet. Here’s a link if you cannot find an appropriate one: <https://dbschema.com/jdbc-driver/MySql.html>

Click on connections >>> Create Connection

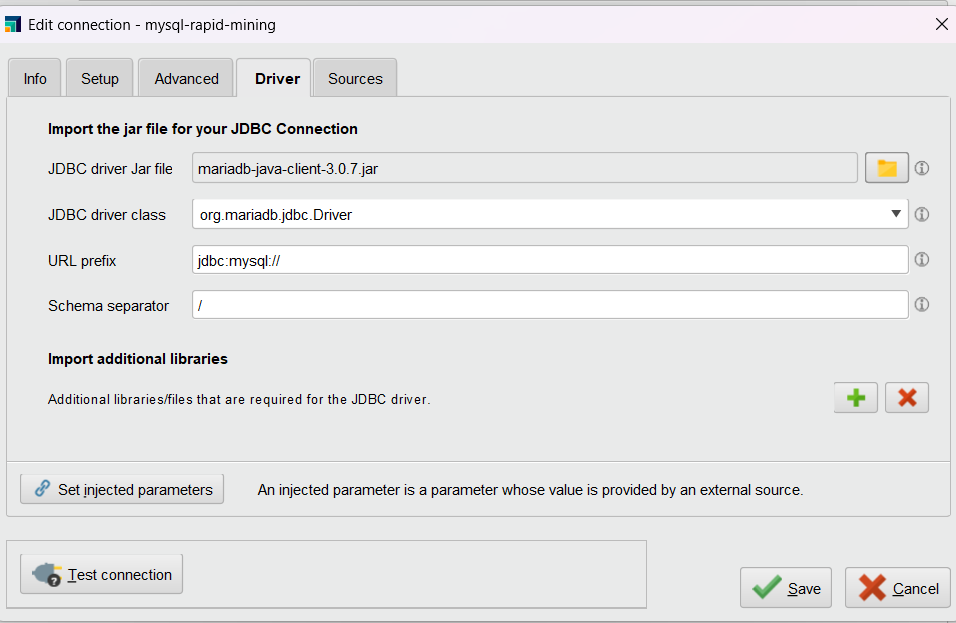


Give the connection a name then click Create



Fill in the User, Password, Host, Port and Database

Click on the driver tab and change the driver Jar file to the MySQL Driver we just installed



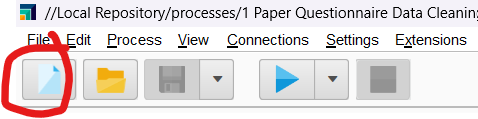
Test the connection then press save

# Data Preparation

## Data Collection

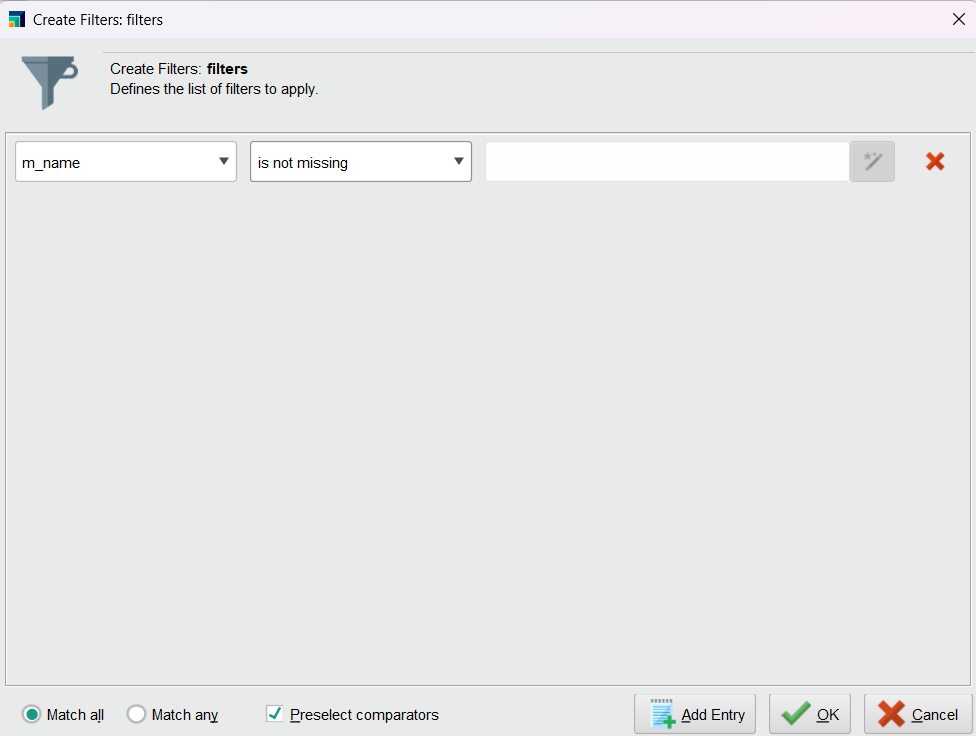
## Data Cleaning for Paper Questionnaire

* + 1. Create a new process

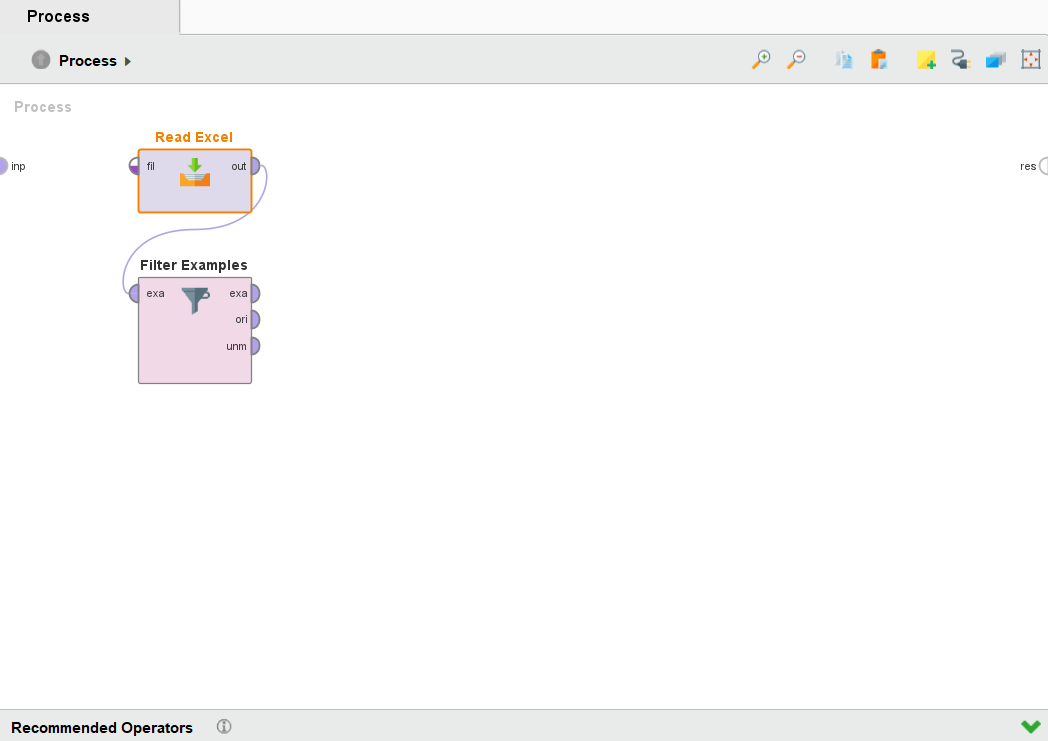


Then Right click and add a Read Excel operator

Click on the newly created operator and give it the directory of the Excel file

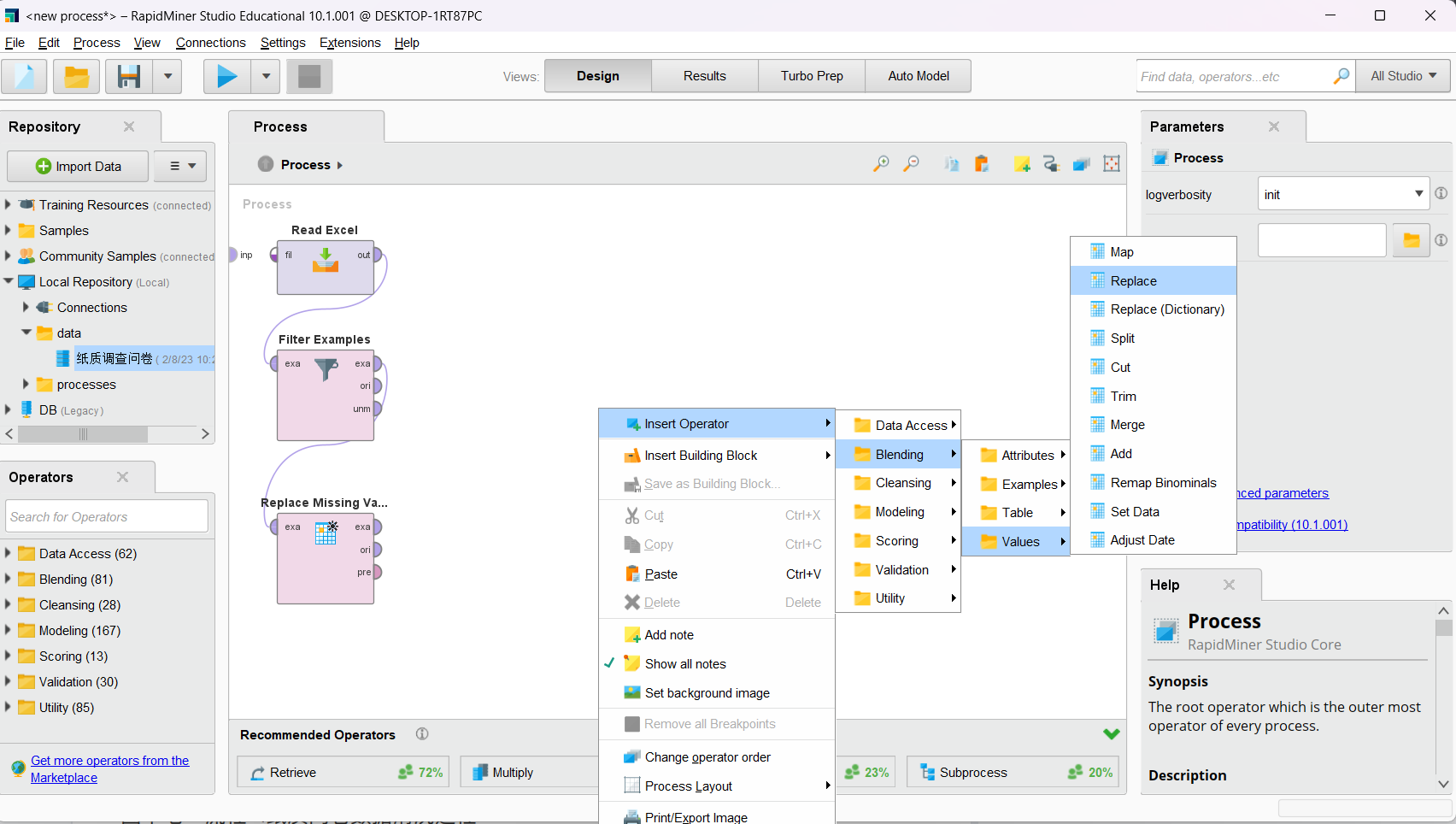
Add a filter operator then click on create new filter

Drag a line from Read Excel out to Filter Examples

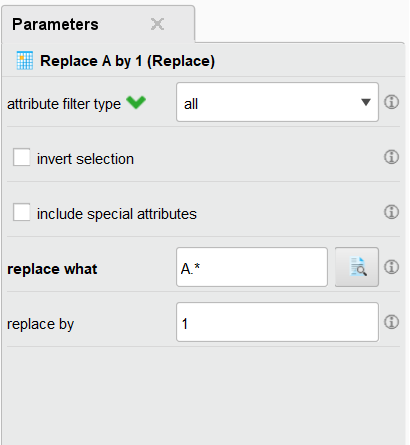


Add a Replace Missing Values operator then change the default to none

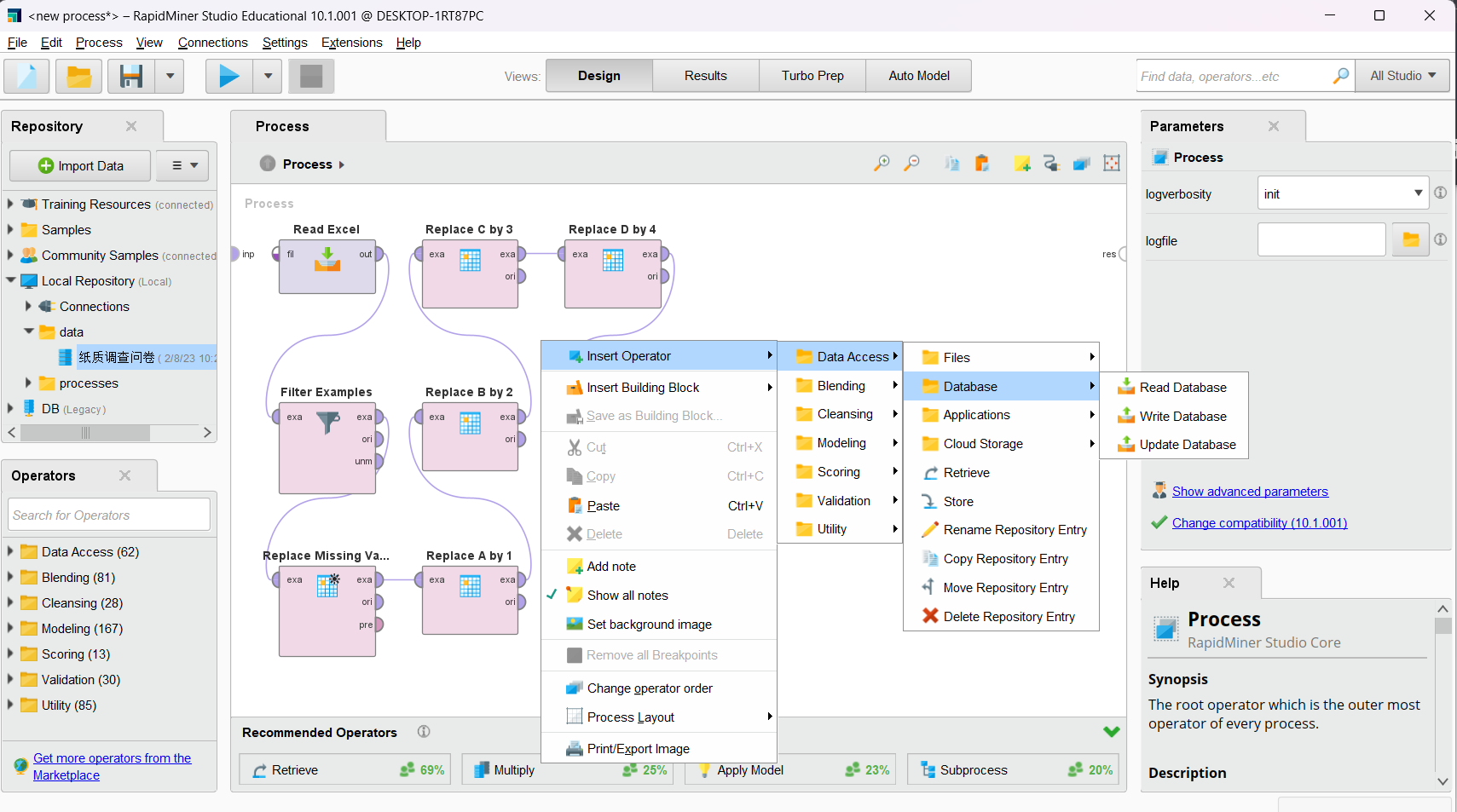
Drag a line to connect Filter Examples exa to Replace Missing Values exa



Add a Replace operator

We use the regular expression A.\* to replace any values with A as 1. “.\*” means arbitrary characters after A. 

Your process should look like this

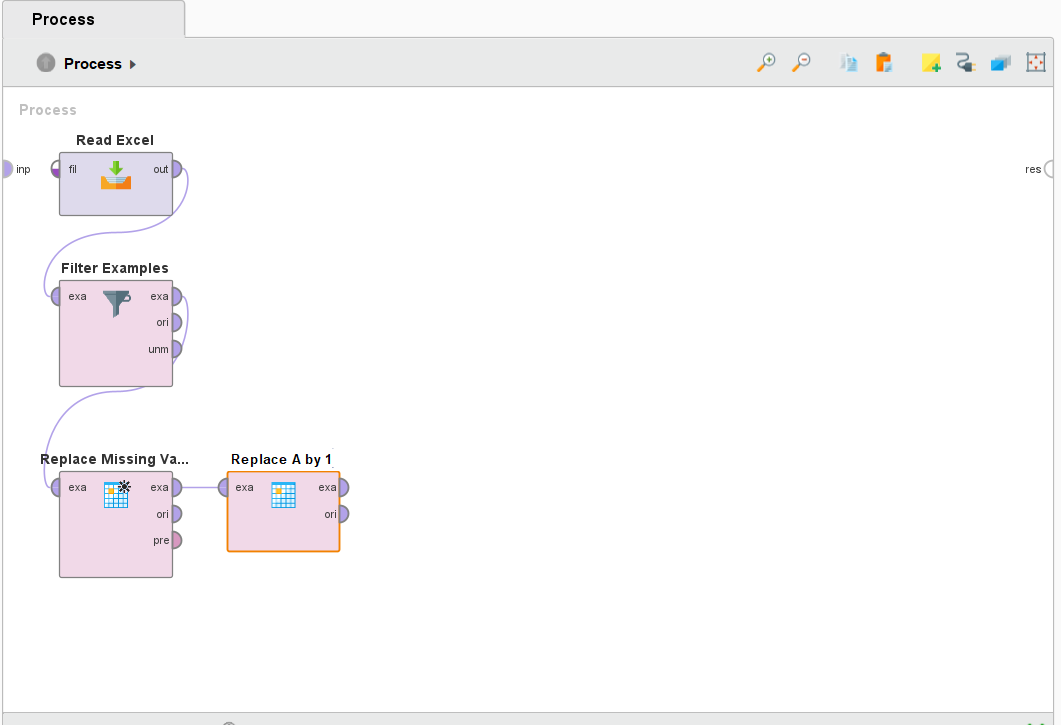


You can right click and copy operators

Copy the replace operator 4 more times

Replace A with B C D E

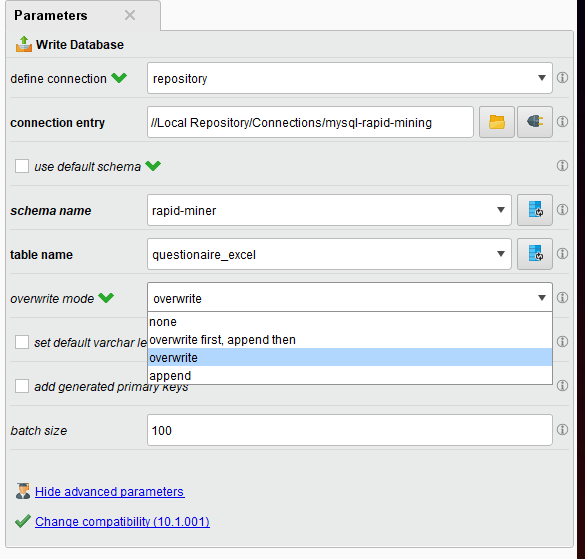
And 1 with 2 3 4 5

Chain them together, one after another

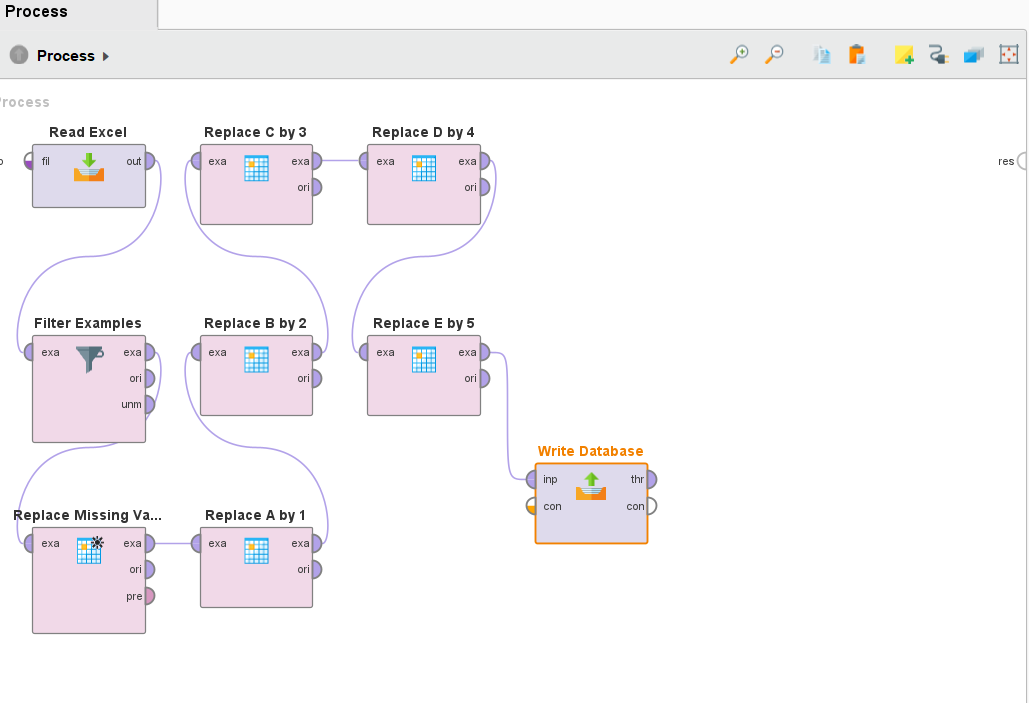
Now add a write database operator to write your data to.

Choose the table you would like to use. The dropdown will take some time to load initially. If you would like to create a new table, enter the new table name (we will name it “questionaire\_excel” for this dataset).

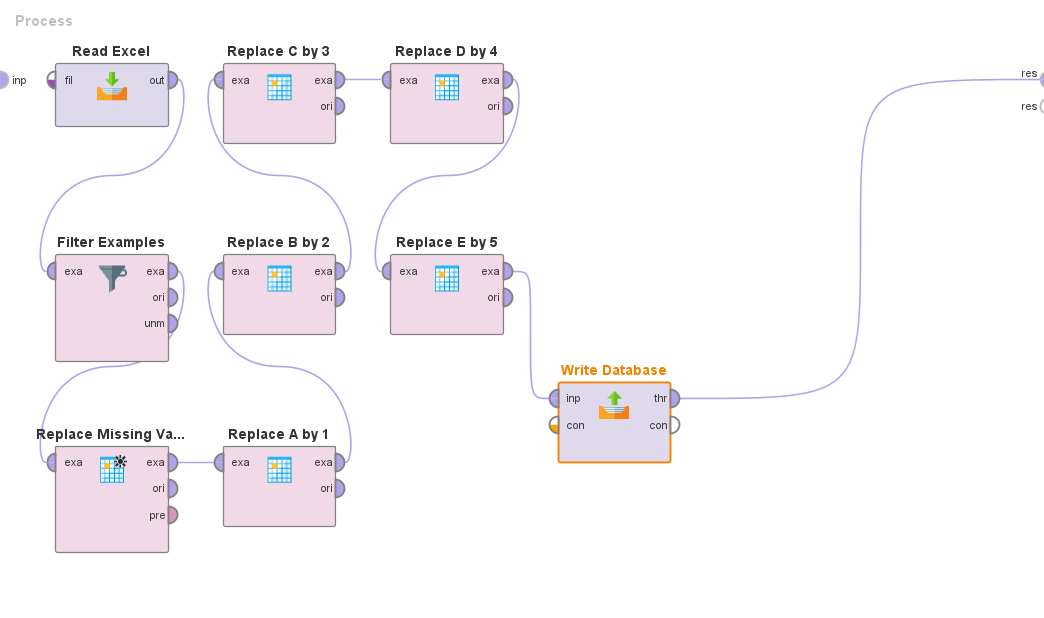
If you choose an existing table, then either the table structure will have to match the data format exactly, or you will encounter an error. Alternatively, you can enable overwrite mode and fill in a new name for the field **table name** so that Rapid Miner automatically creates a table for you. If you have an existing table and would like to overwrite that table, you can toggle on overwrite mode as shown below. (Note: all existing data in the original table will then be erased. Please proceed with caution.)



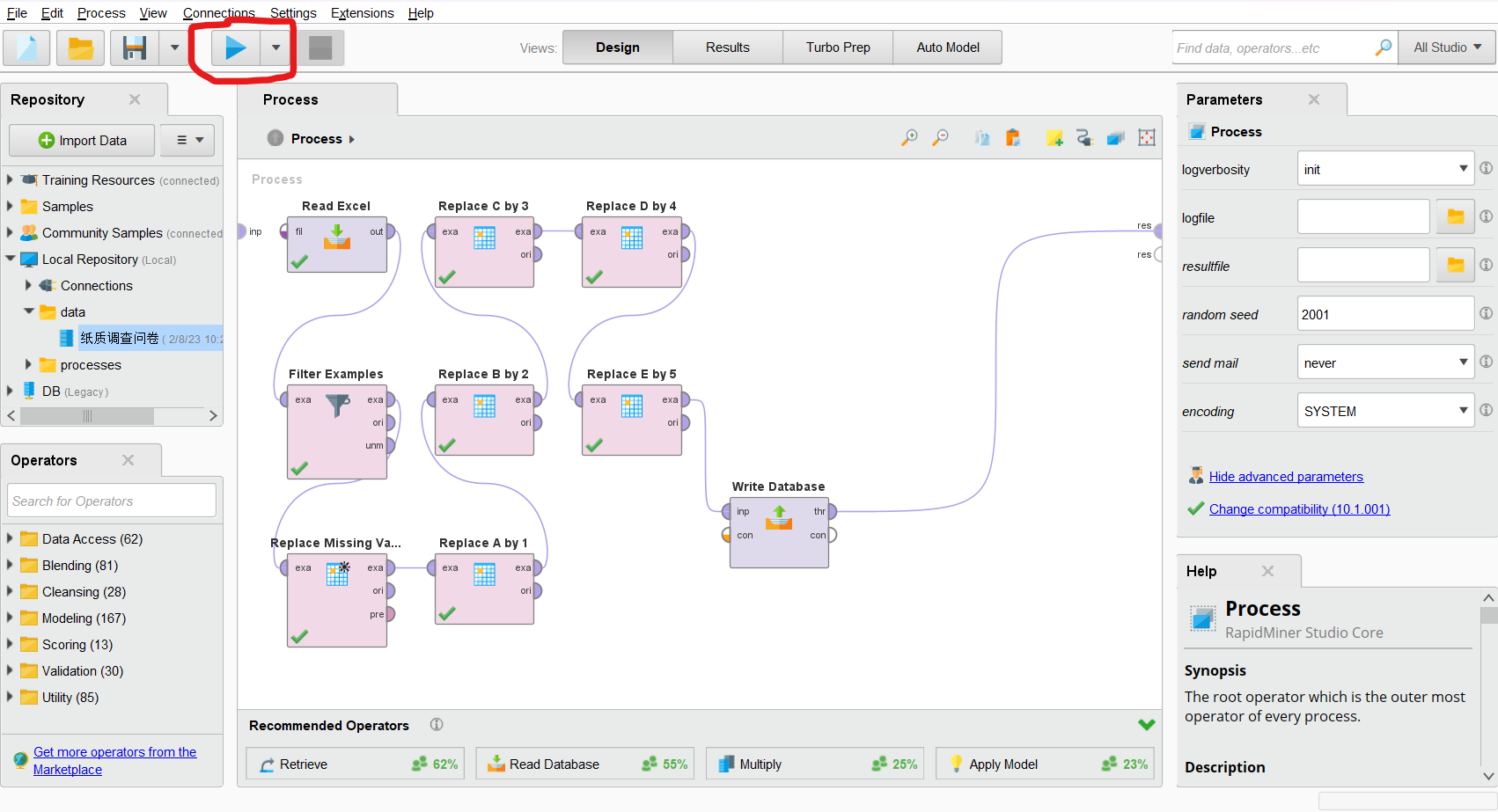
After you’re done configuring the Write Database component. Connect “Replace E by 5” to our new Write Database component.



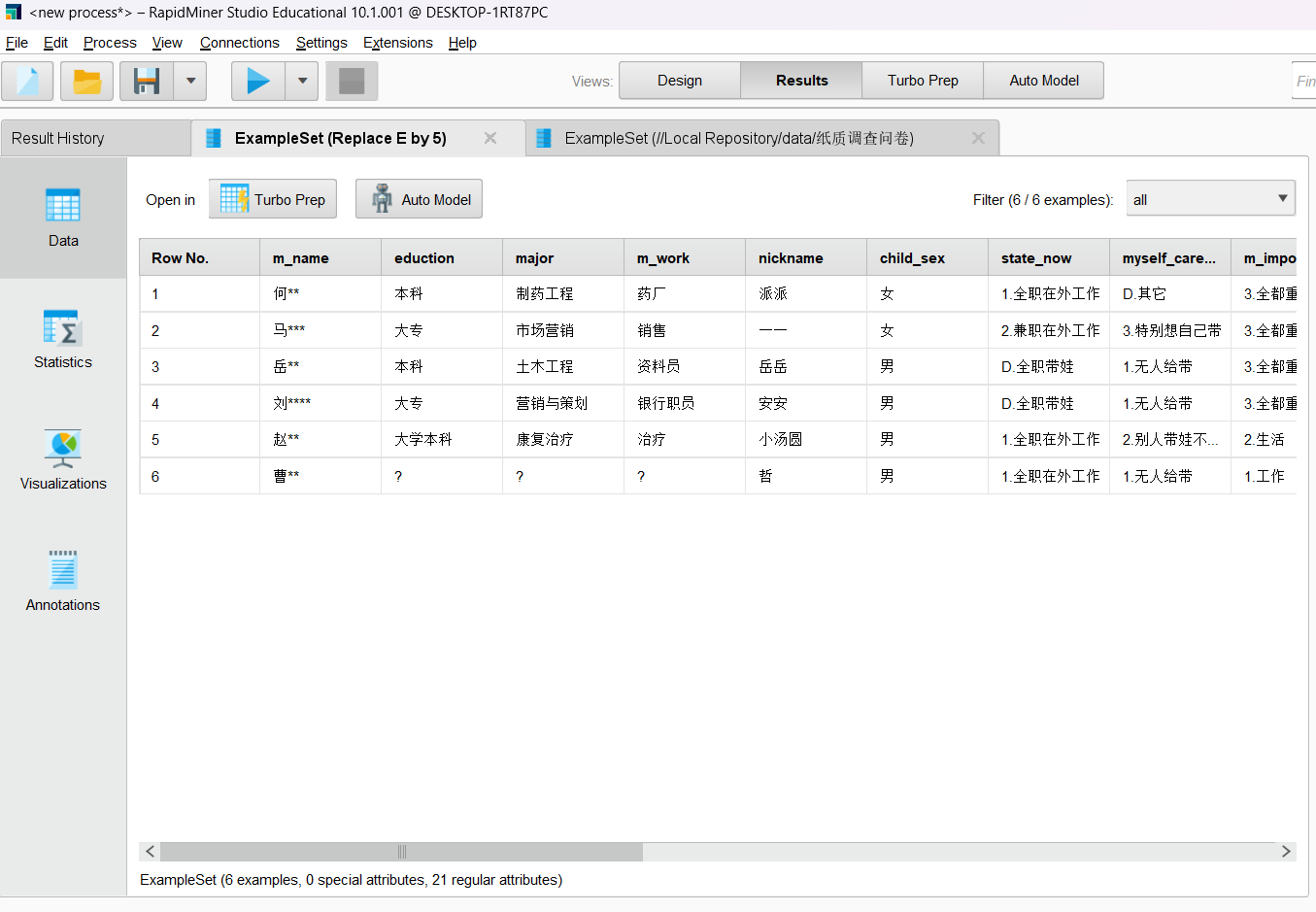
Finally, connect the “Write Database” component to the endpoint “res.”

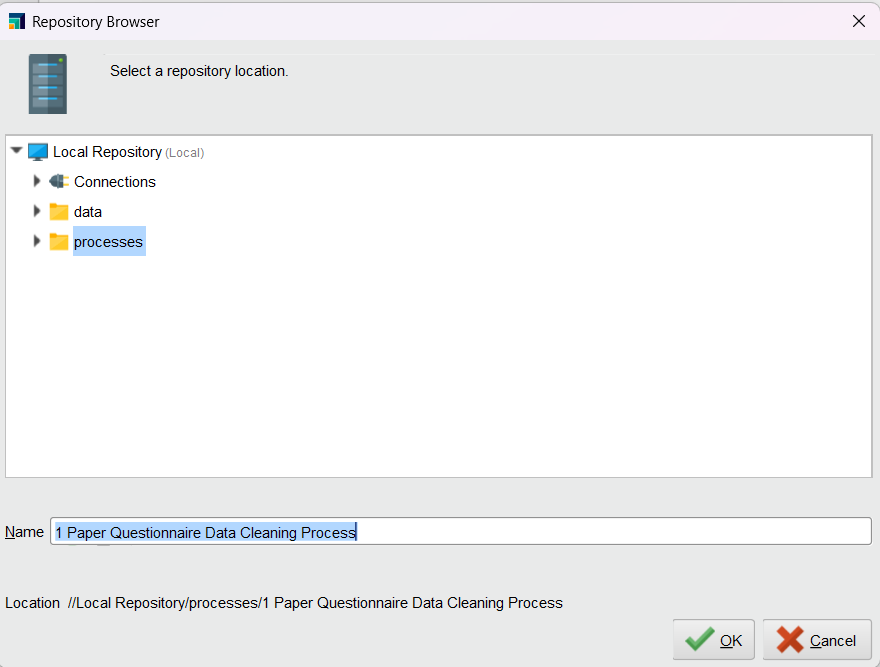
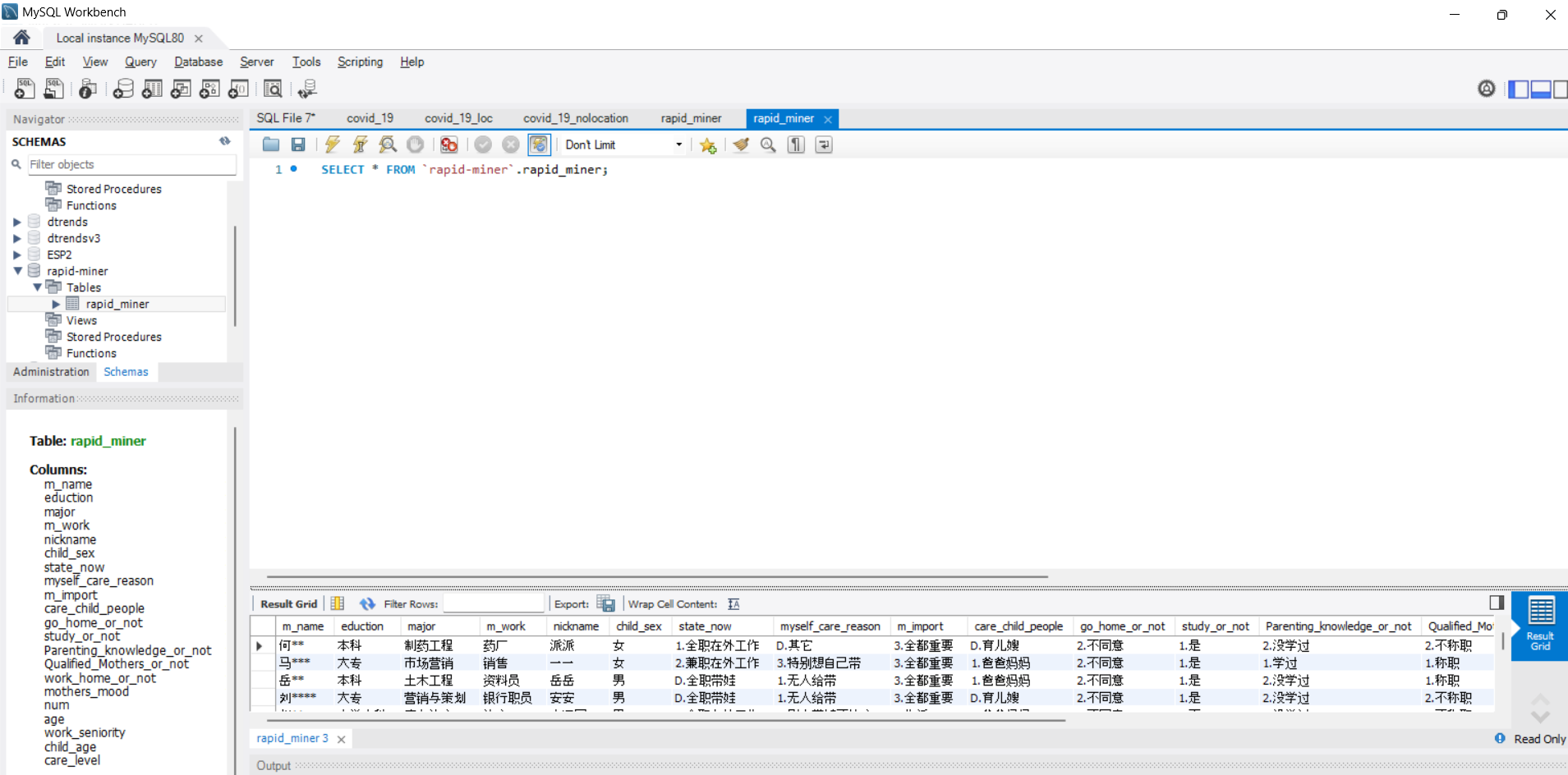


Now you can run the process by clicking this button.



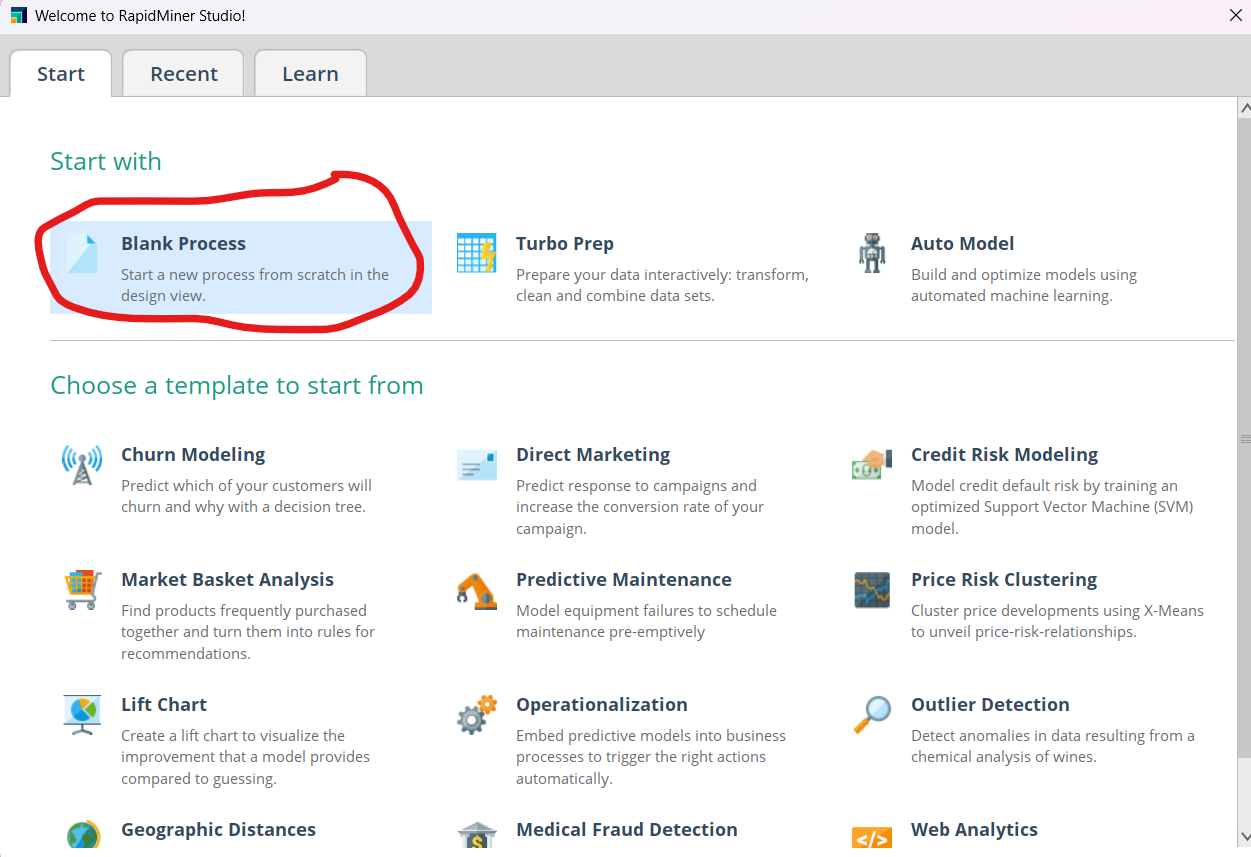
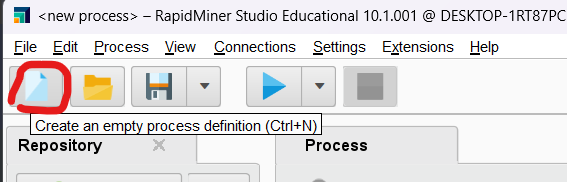
If everything ran successfully, the view should change to this. There should only be 6 rows because we’ve filtered out one of the columns with m\_name missing.

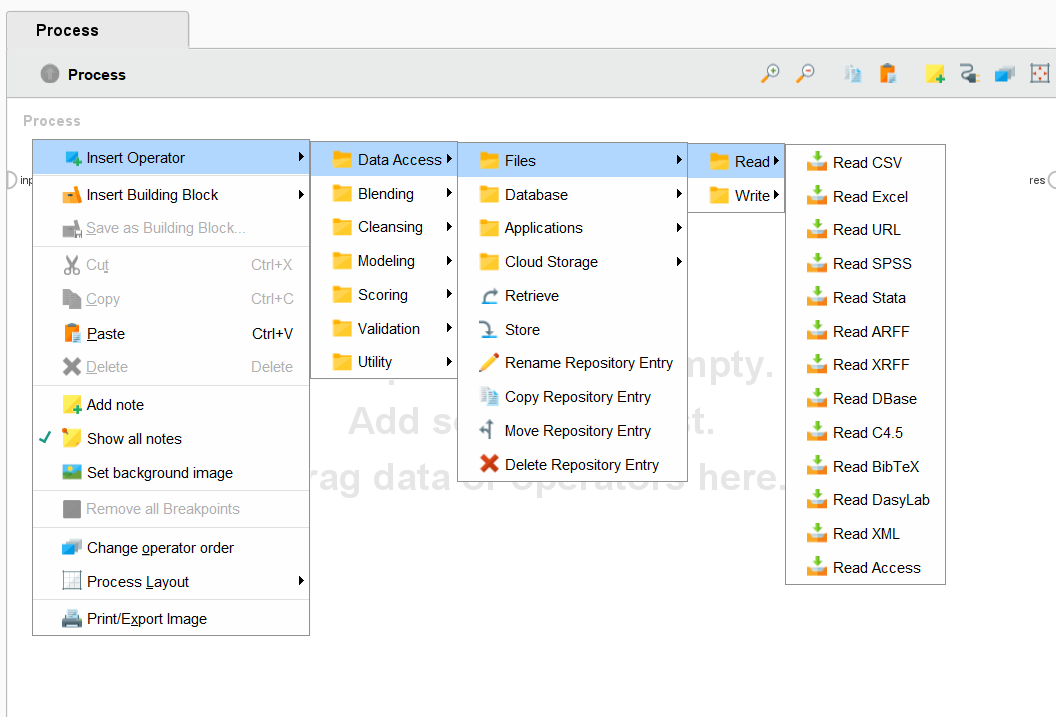
You can also verify this in MySQL Workbench.

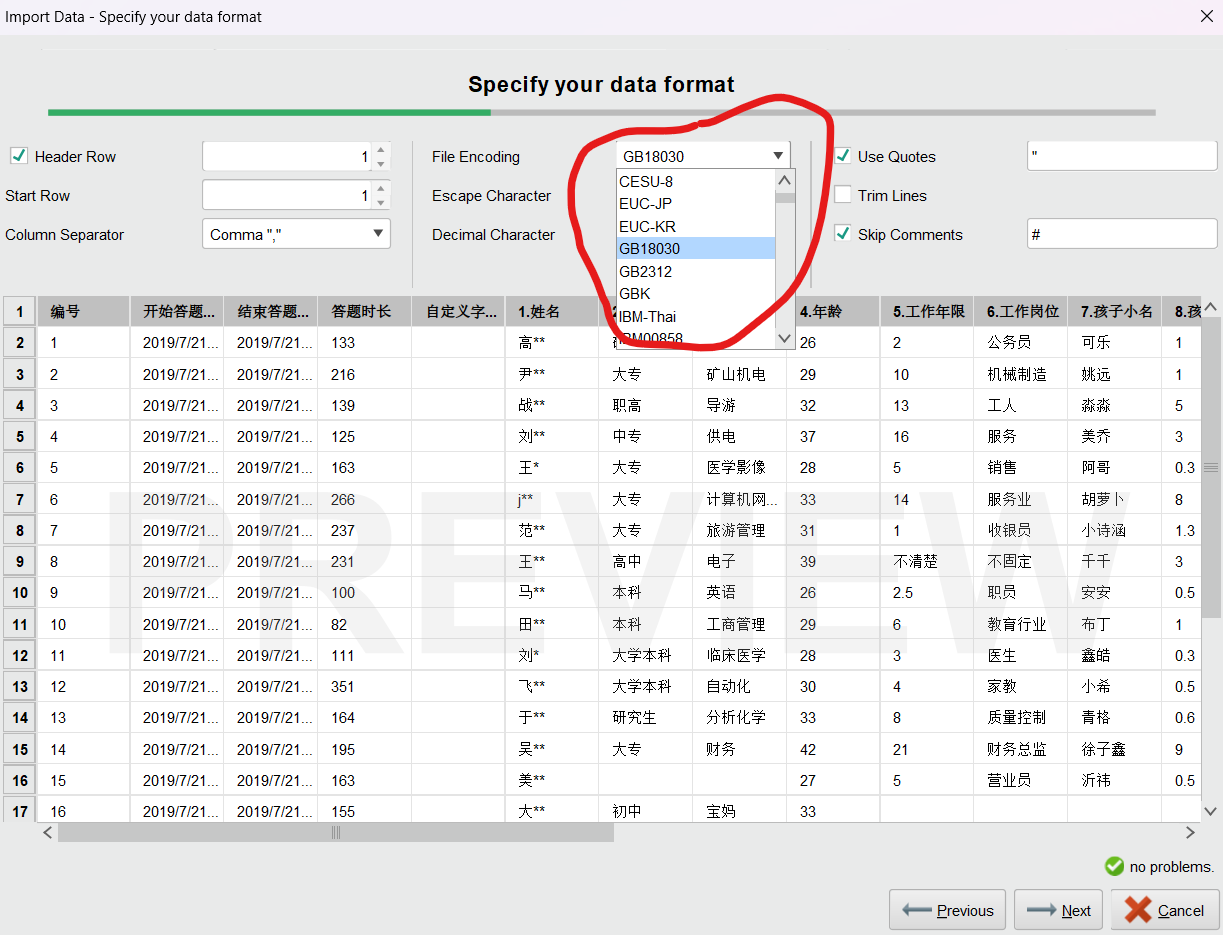
If the process is correct, save the process.

## Data Cleaning for Tencent Questionnaire

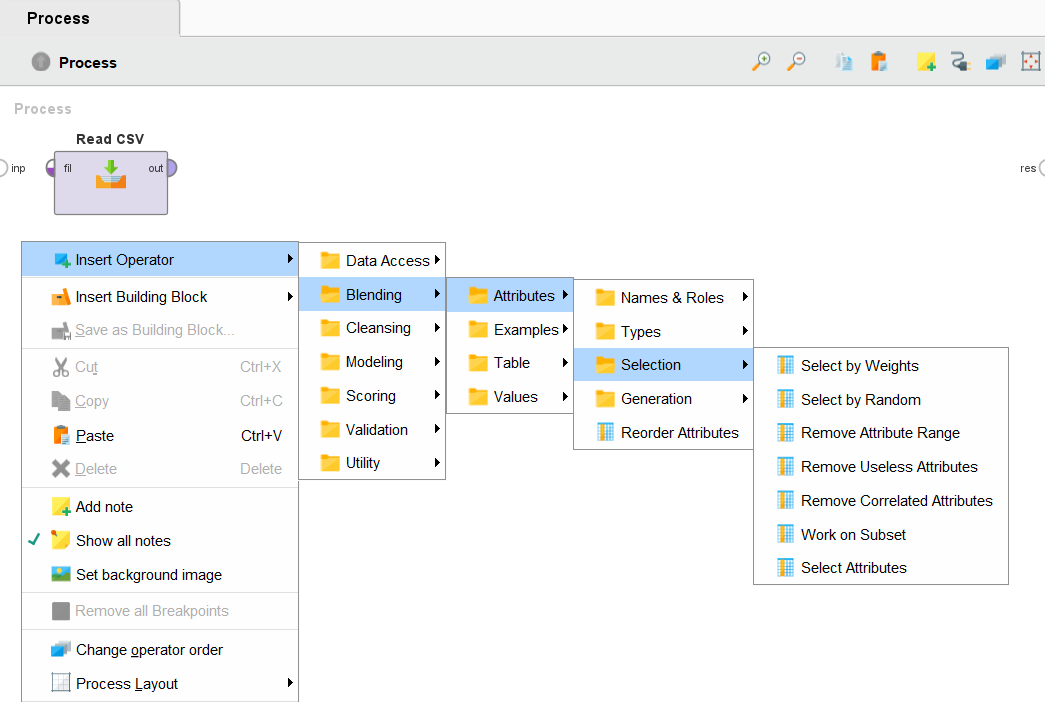
Now, let’s create a new process for the Tencent questionnaire data cleaning.

Click on the blank sheet icon in the upper left corner then select Blank Process

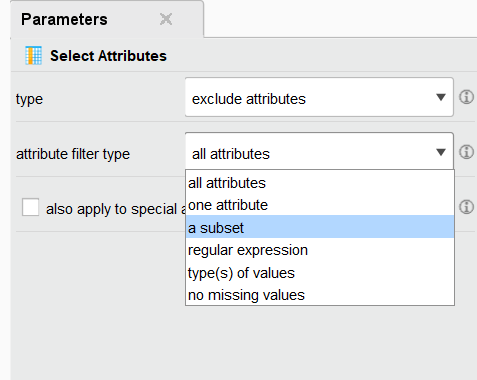
Create a read CSV module like before and select the directory for the csv file you wish to use.



Since our dataset uses Chinese characters we needed to select the right encoding method to properly display the characters

Click finish after ensuring your dataset is correct.

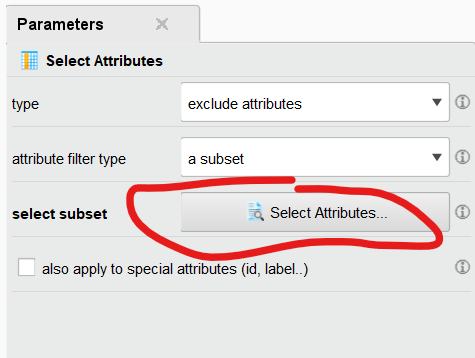
Add a Select Attribute operator



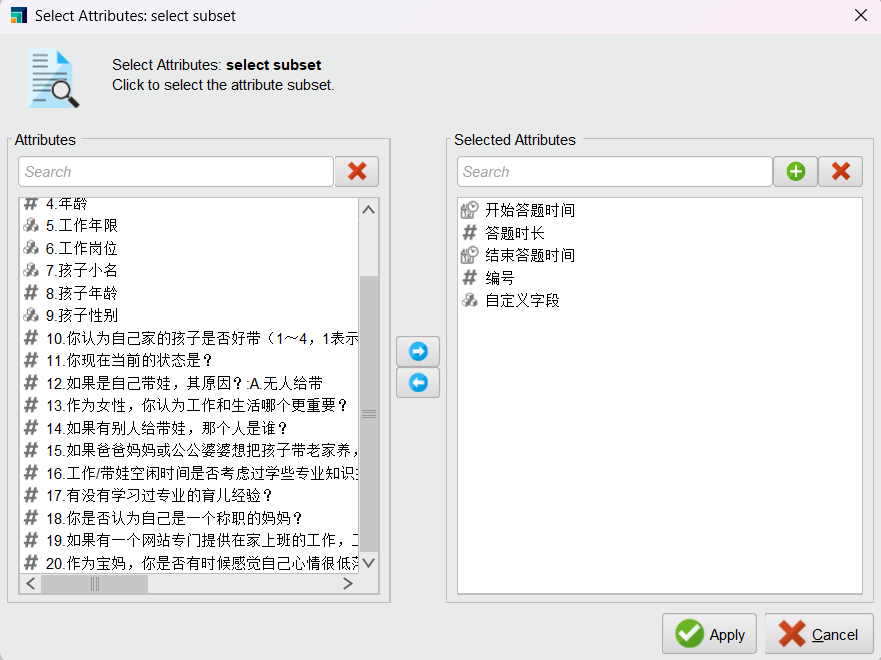
This interface will show up on the right.

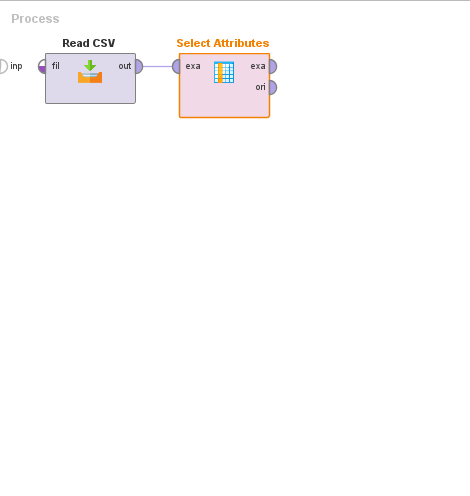
Set it to filter out a subset of attributes.

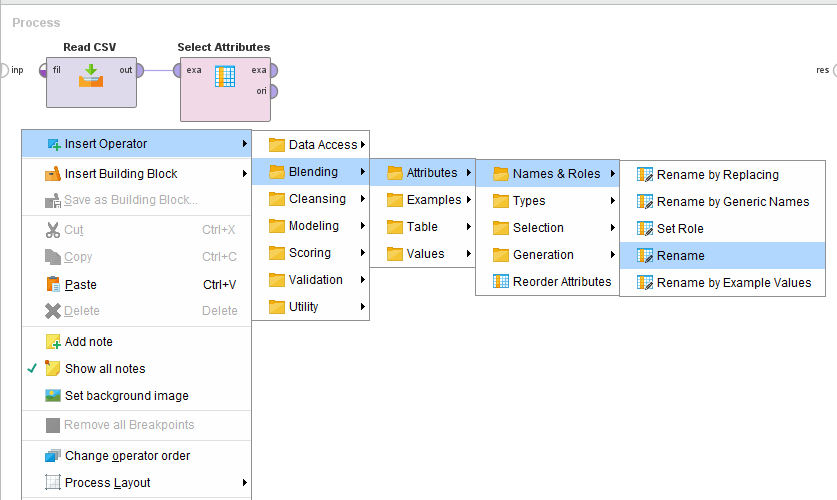
Then select the subset to be filtered out.



In the popup window, select the following attributes: 开始答题时间, 答题时长, 结束答题时间, 编号, 自定义字段 (as shown below). Then click Apply.

Drag a line from “Read CSV” component to exa of “Select Attributes” so that it looks like below.



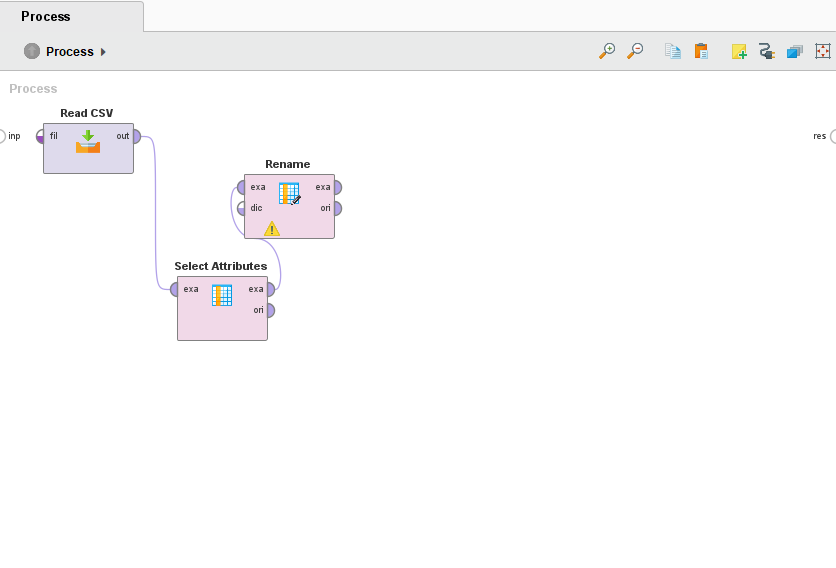
Add a new Operator to rename attributes

Once you click Edit List, the following window will show up. Use the Add Entry button on the bottom right corner to add 21 entries, and fill out the table as shown below.

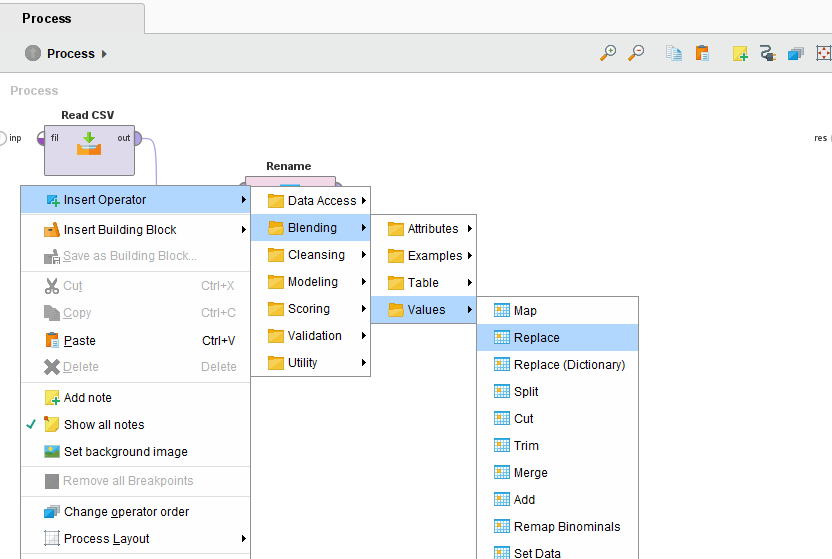
| **old name** | **new name** |
| --- | --- |
| 1.姓名 | m\_name |
| 2.学历 | eduction |
| 3.专业 | major |
| 4.年龄 | age |
| 5.工作年限 | work\_seniority |
| 6.工作岗位 | m\_work |
| 7.孩子小名 | nickname |
| 8.孩子年龄 | child\_age |
| 9.孩子性别 | child\_sex |
| 10.你认为自己家的孩子是否好带（1～4，1表示好带，4表示特别不好带） | care\_level |
| 11.你现在当前的状态是？ | state\_now |
| 12.如果是自己带娃，其原因？:A.无人给带 | myself\_care\_reason |
| 13.作为女性，你认为工作和生活哪个更重要？ | m\_import |
| 14.如果有别人给带娃，那个人是谁？ | care\_child\_people |
| 15.如果爸爸妈妈或公公婆婆想把孩子带老家养，什么都不用你们管，是否同意？ | go\_home\_or\_not |
| 16.工作/带娃空闲时间是否考虑过学些专业知识提高自己的竞争力？ | study\_or\_not |
| 17.有没有学习过专业的育儿经验？ | Parenting\_knowledge\_or\_not |
| 18.你是否认为自己是一个称职的妈妈？ | Qualified\_Mothers\_or\_not |
| 19.如果有一个网站专门提供在家上班的工作，工作不耽误带娃，薪资水平也不错，是否考虑注册一个账号，并在上面找一个合适的工作？ | work\_home\_or\_not |
| 20.作为宝妈，你是否有时候感觉自己心情很低落，像是得了抑郁症，很需要别人进行情感疏导？ | mothers\_mood |

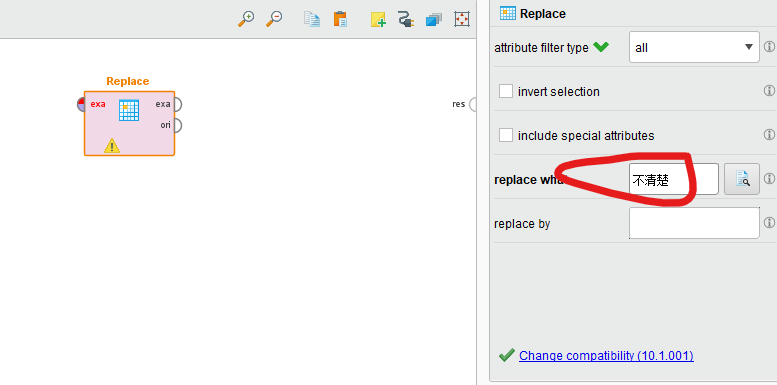
Click Apply when you’re done.

Connect the Rename operator so that the graph looks like this.

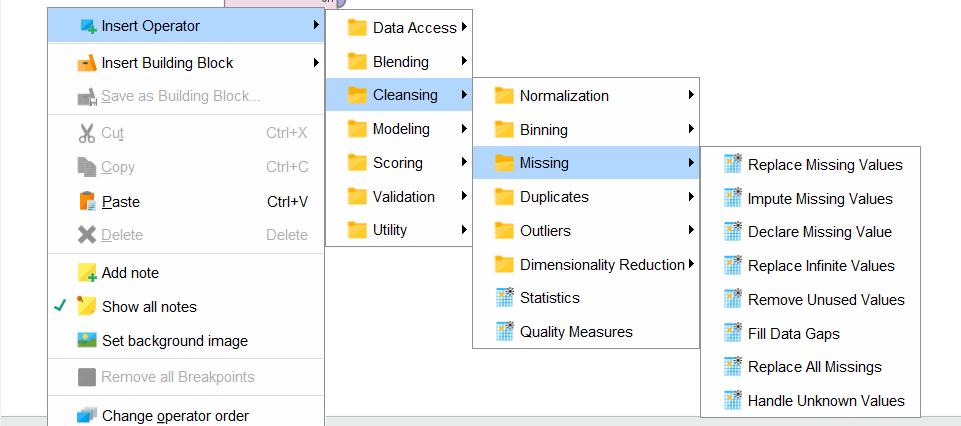


Next add a replace operator

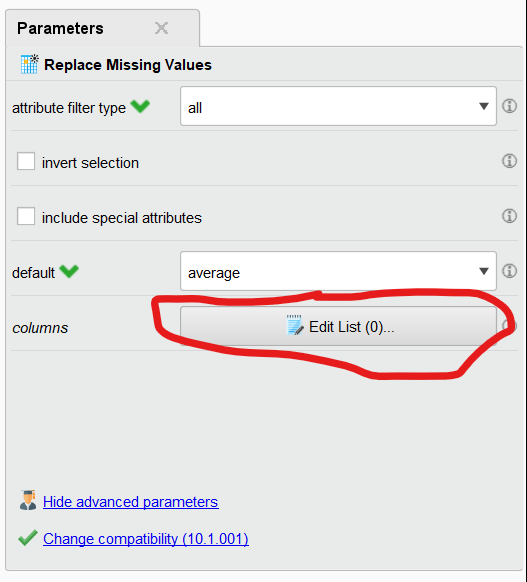
Edit the “replace what” field in the Replace window to be “不清楚” The purpose is to replace one of the values in our data that says “不清楚” for work seniority.



Drag a line from Rename to Replace



Create a replace missing values operator

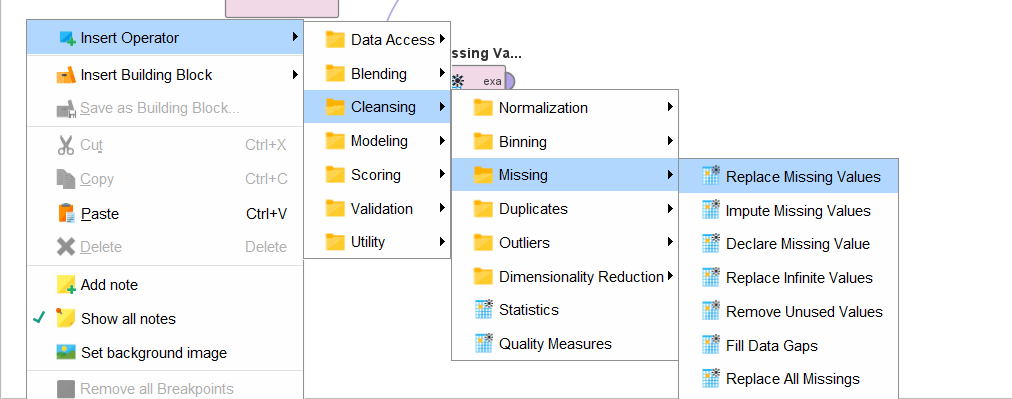
Add 7 entries using the Add Entry button below, and fill in the following values.

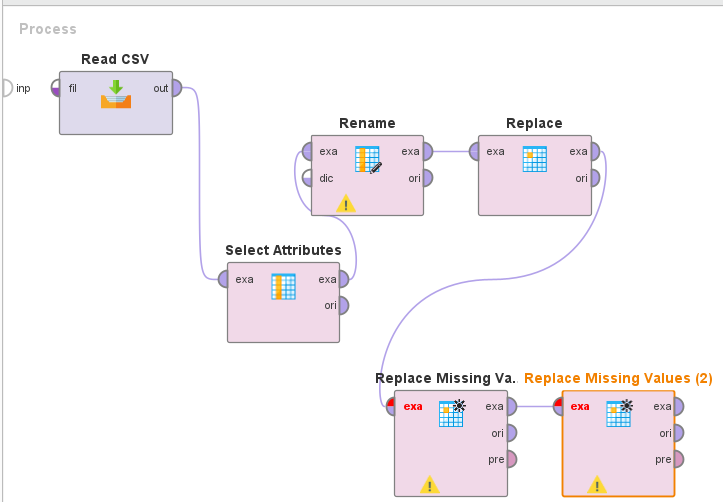
| **attribute** | **replace with** |
| --- | --- |
| 2.学历 | none |
| 3.专业 | none |
| 6.工作岗位 | none |
| 7.孩子小名 | none |
| 9.孩子性别 | none |
| 5.工作年限 | average |
| 8.孩子年龄 | average |

Click Apply when you’re done.

Connect a line from Replace to Replace Missing Values.

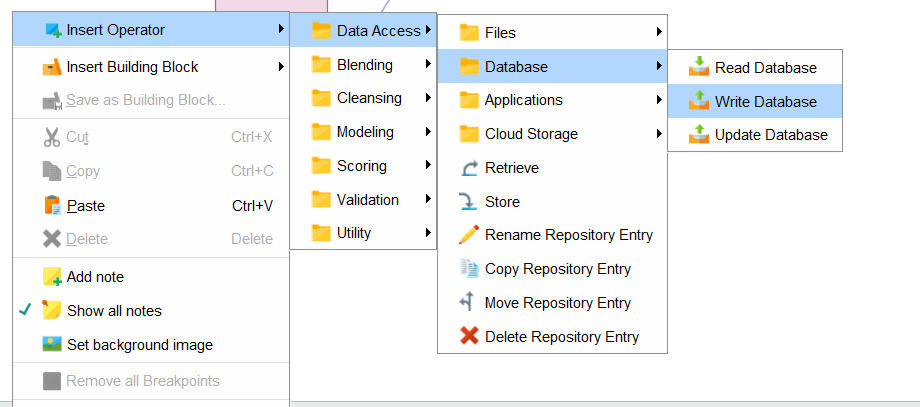
* + 1. F

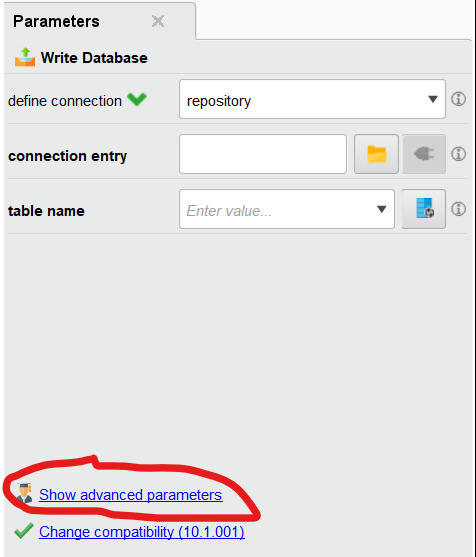


Use another replace missing values operator then add 1 entry that replaces missing values for the “12.如果是自己带娃，其原因？:A.无人给带” attribute and select replace with “value”

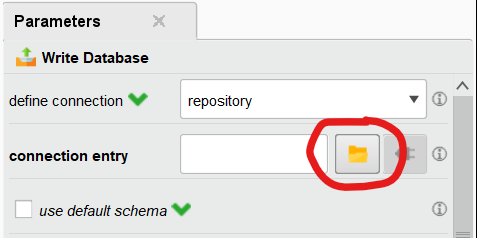
Then apply it and connect it to the our chain of operators

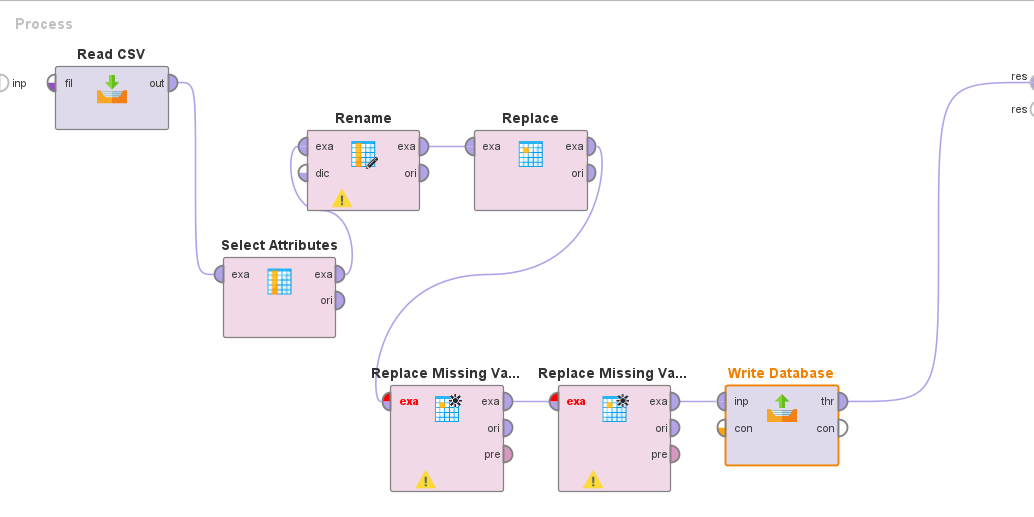
To save our changes to a database we will add a Write Database Operator





Like earlier, select the database that you want to write to then save the connection.

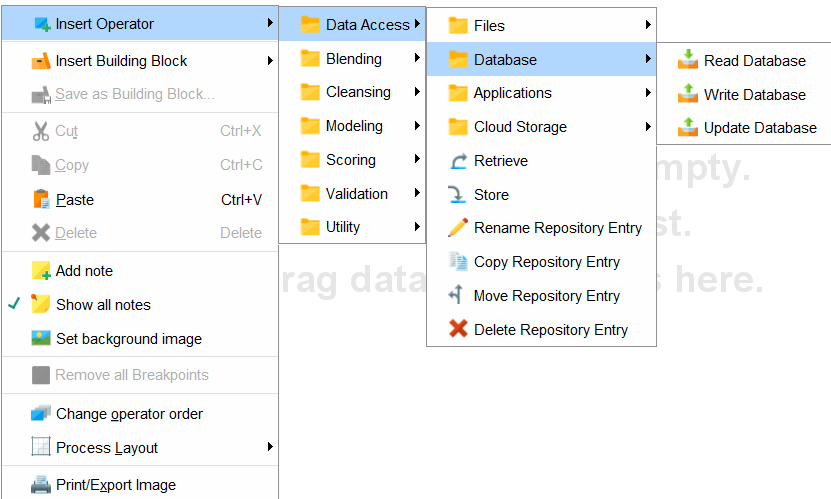
Fill in the schema name and table name (different table from before, our name here is called “questionaire\_csv”). Also, make sure overwrite mode is set to overwrite if you are overwriting a table or creating a new table.



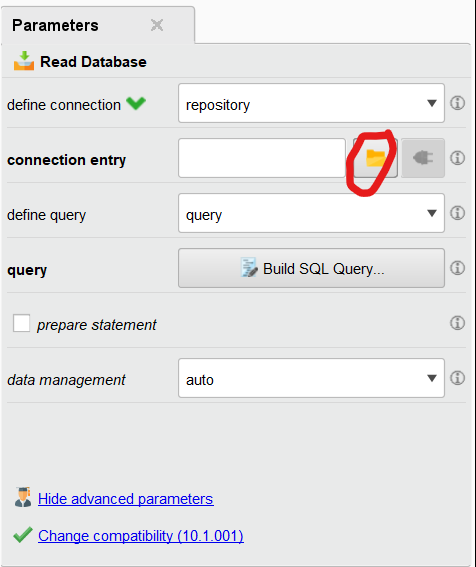
Write to Database to res.

Run the process to make sure that it works. The output should contain rows from both datasets. Then save the process

Read Database



Now that we have cleaned out data we will make a new process to “process” it.

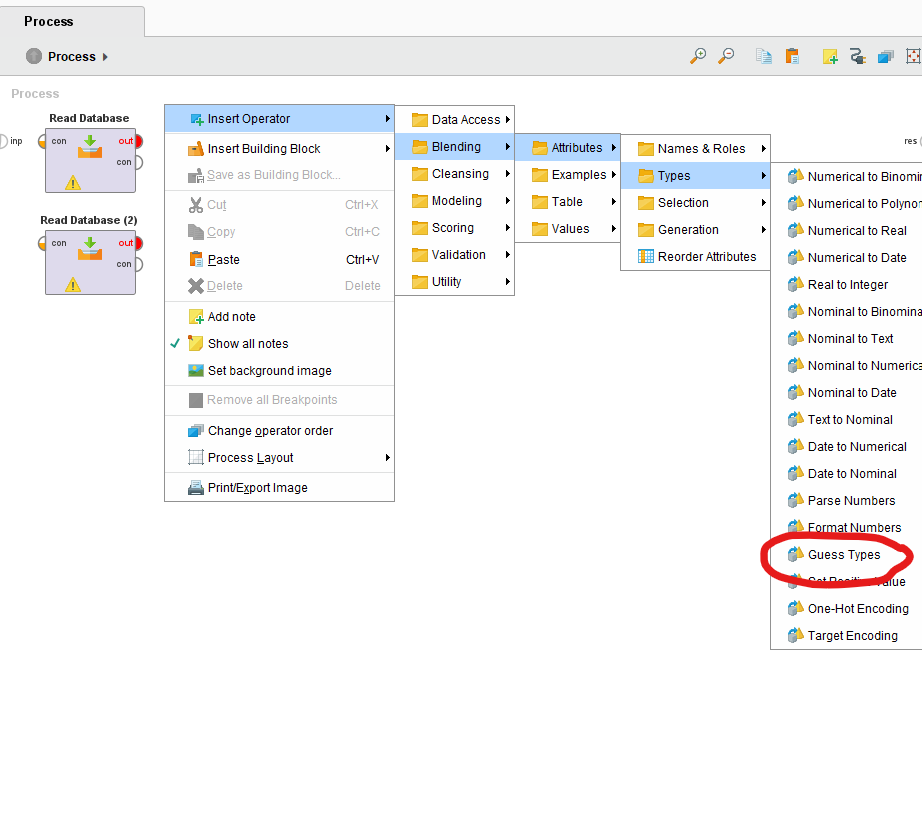
First we will need to have an operator read the data that we save to our database. Add a read database operator and select the database to read from

Since we wrote to a MySQL database we will have to retrieve our data by writing a query for it.

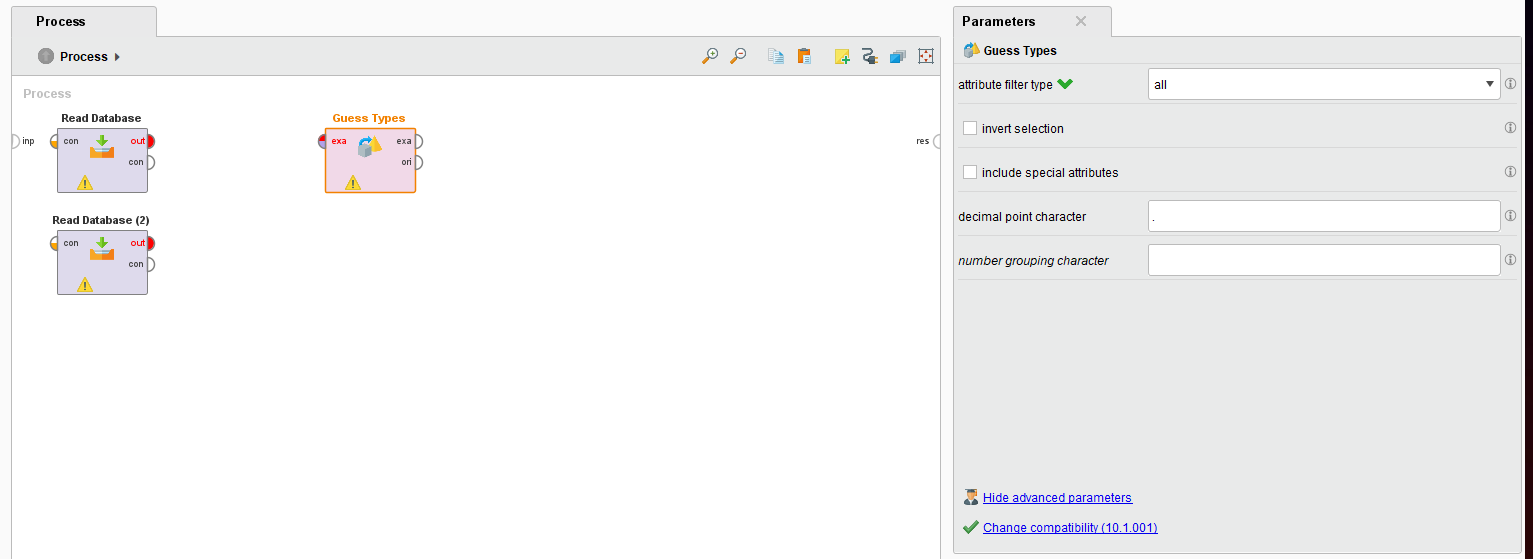
Se can simply select all data on the table

Make another operator to read from our second table and again select all from the table

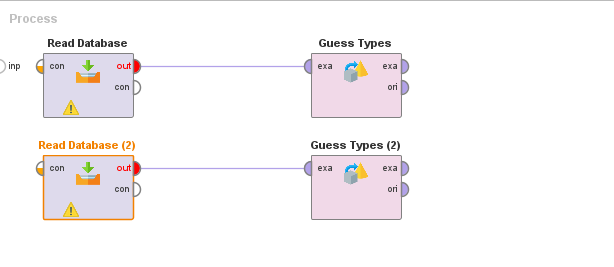
Now that we can read from both tables

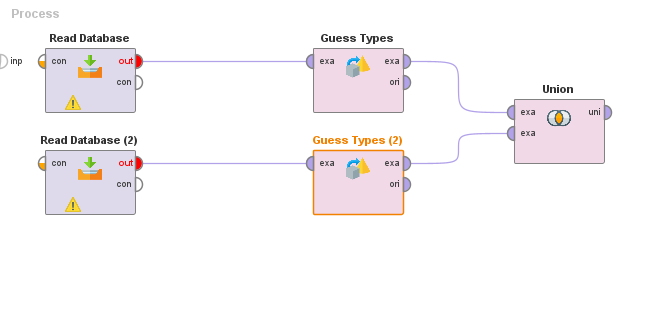


Your interface should now look like the following.

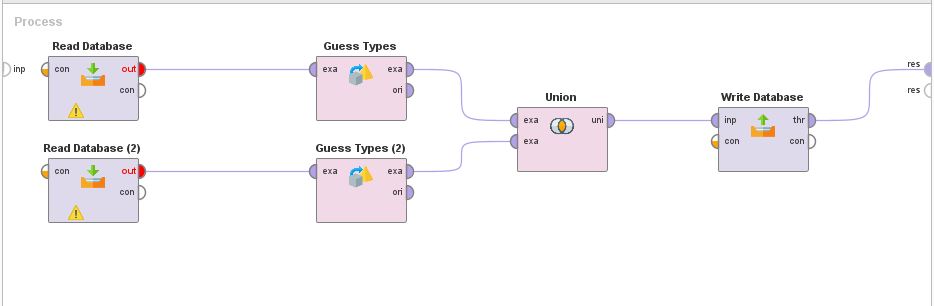


Create and connect a Guess Types operator with each read database operator



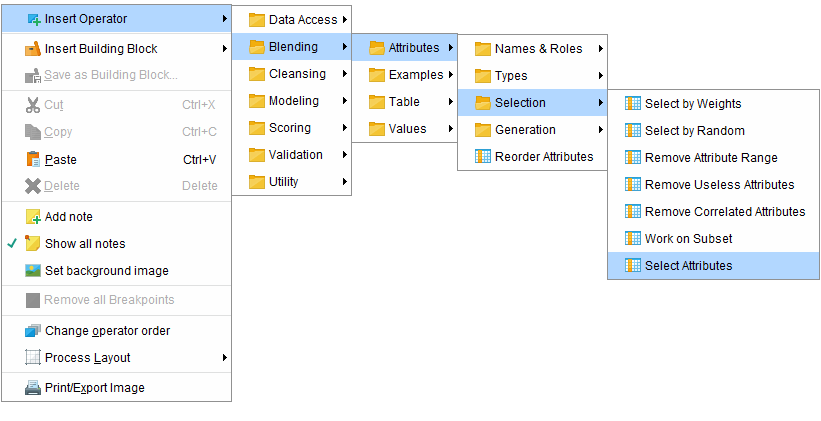
Now add a Union operator then connect both Guess Types to the Union

Now we will write the result to the database as before then connect the result to res



Run the process, and you should see the interface display the following table. We can also verify the new table exists in MySQL Workbench. Save the process with the name “2 Data Integration”

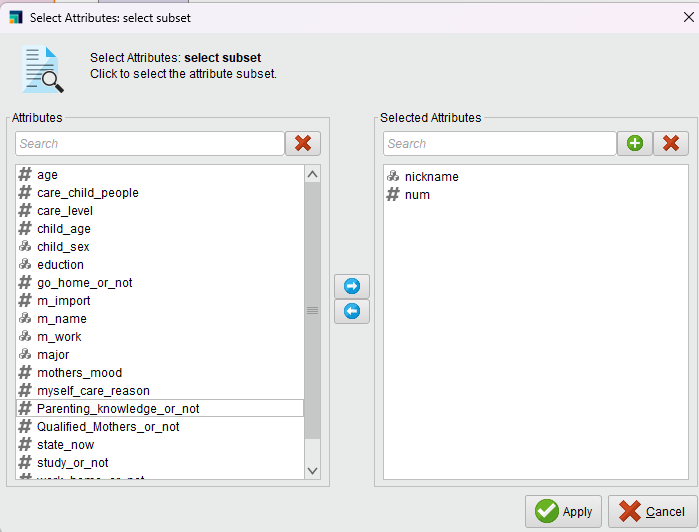
## Data Selection

Create a new process and add a read database process selecting all from the table we made earlier

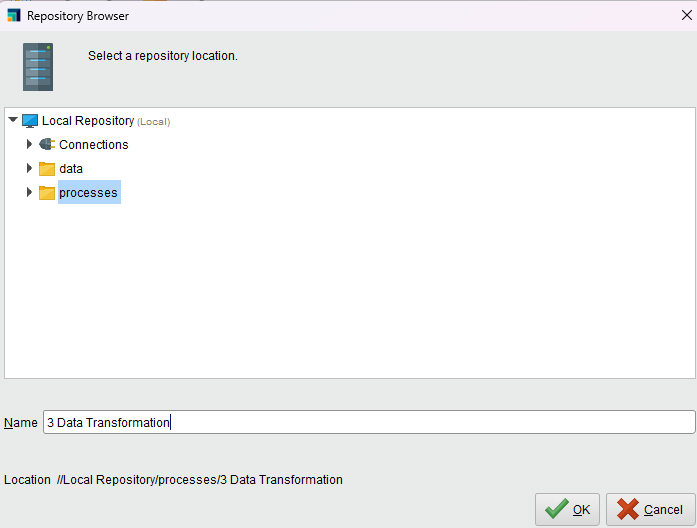
Now add a select attributes process

And set it to exclude a subset of attributes

Then connect the it to our database

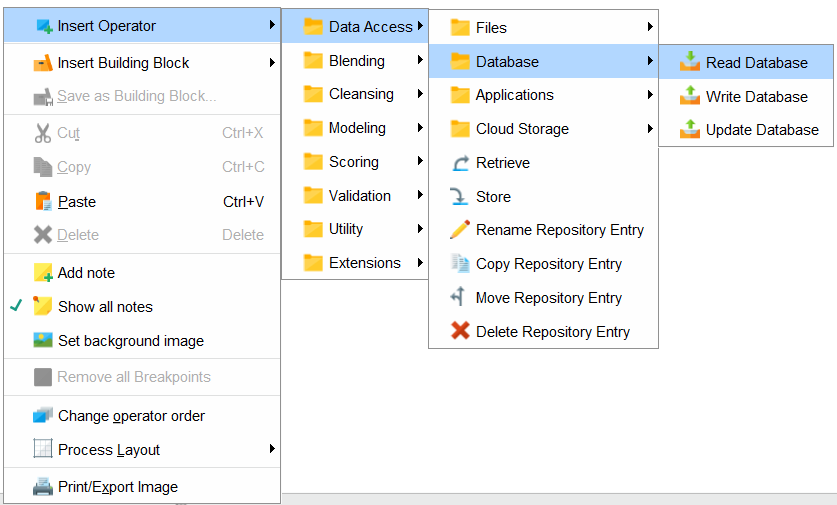


Add a write database operator and save like we have done before



## Data Transformation

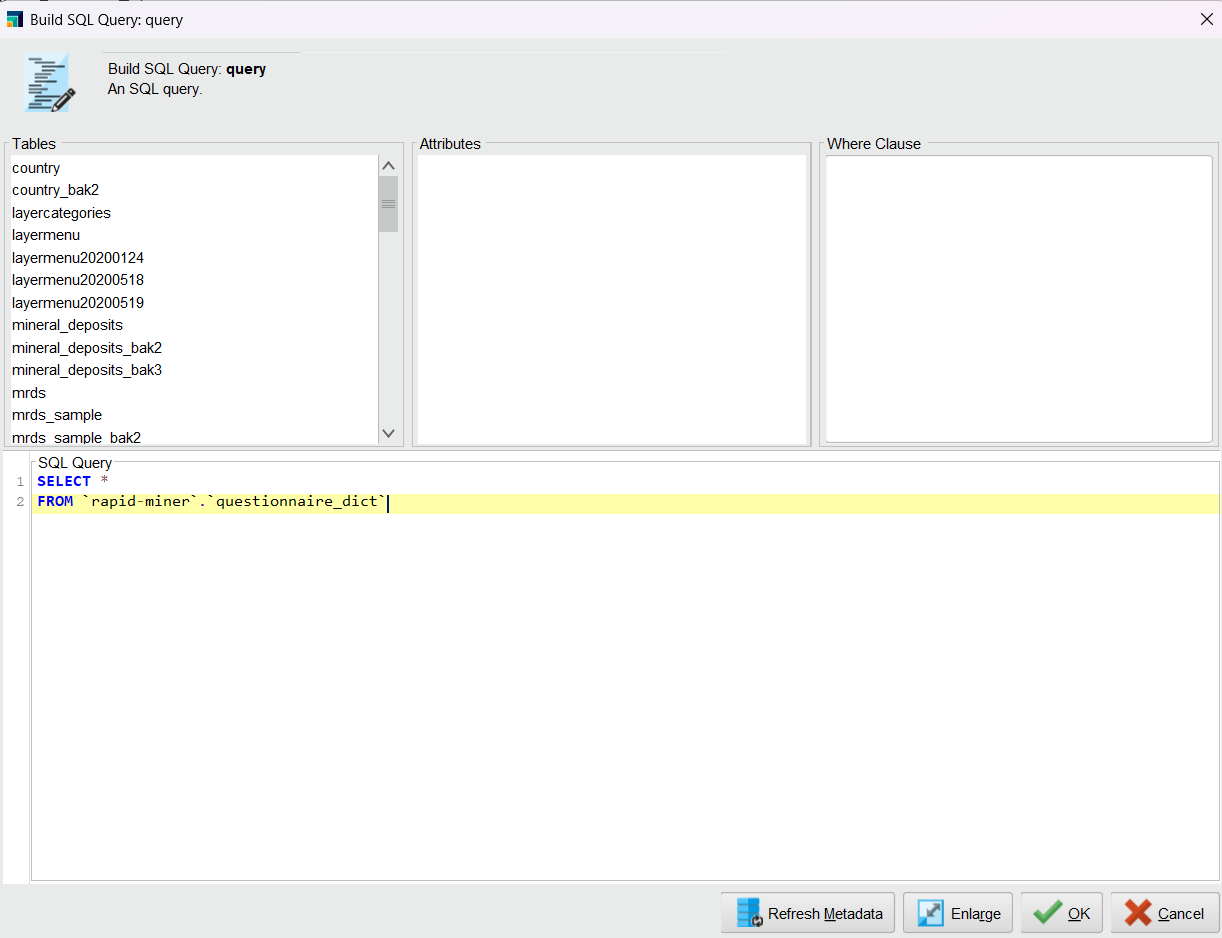
Create a new process and read the table we made in the prior process making sure to select all data



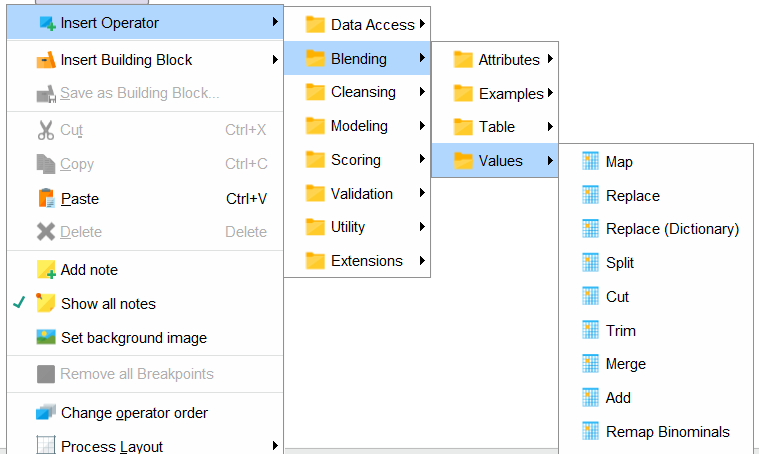
* + 1. Insert the following table into a new table in MySQL Workbench with the following values. We named the table questionnaire\_dict. [Sheet to download data](https://docs.google.com/spreadsheets/d/1atVSHCnXcL6gN-aXLh7xZnmfSPbfpp_g37eamqyVH2A/edit#gid=0)

| ID | attribute | chinese | value |
| --- | --- | --- | --- |
| 1 | eduction | 初中 | 1 |
| 2 | eduction | 高中 | 2 |
| 3 | eduction | 中专 | 3 |
| 4 | eduction | 大专 | 4 |
| 5 | eduction | 职高 | 5 |
| 6 | eduction | 大学本科 | 6 |
| 7 | eduction | 本科 | 6 |
| 8 | eduction | 硕士 | 7 |
| 9 | eduction | 研究生 | 7 |
| 10 | eduction |  | 0 |
| 12 | major | 宝妈 | 1 |
| 13 | major | 财务 | 2 |
| 14 | major | 会计电算化 | 2 |
| 15 | major | 行政管理 | 3 |
| 16 | major | 人力资源管理 | 3 |
| 17 | major | 计算机网络 | 4 |
| 18 | major | 计算机 | 4 |
| 19 | major | 生物工程 | 5 |
| 20 | major | 生物技术应用 | 5 |
| 21 | major | 市场营销 | 6 |
| 22 | major | 营销与策划 | 6 |
| 23 | major | 导游 | 7 |
| 24 | major | 旅游管理 | 7 |
| 25 | major | 财务总监 | 8 |
| 26 | major | 电子商务 | 8 |
| 27 | major | 电子 | 9 |
| 28 | major | 医学影像 | 9 |
| 29 | major | 制药工程 | 9 |
| 30 | major | 临床医学 | 10 |
| 31 | major | 英语 | 10 |
| 32 | major | 政治学 | 11 |
| 33 | major | 俄语 | 12 |
| 34 | major | 分析化学 | 13 |
| 35 | major | 工商管理 | 14 |
| 36 | major | 供电 | 15 |
| 37 | major | 精密机械 | 16 |
| 38 | major | 矿山机电 | 17 |
| 39 | major | 土木工程 | 18 |
| 40 | major | 自动化 | 19 |
| 41 | child\_sex | 男 | 1 |
| 42 | child\_sex | 女 | 2 |
| 43 | m\_work | HRM | 1 |
| 44 | m\_work | HR | 1 |
| 45 | m\_work | 服务业 | 2 |
| 46 | m\_work | 收银员 | 2 |
| 47 | m\_work | 营业员 | 2 |
| 48 | m\_work | 客服 | 2 |
| 49 | m\_work | 餐饮 | 2 |
| 50 | m\_work | 职工 | 2 |
| 51 | m\_work | 销售 | 3 |
| 52 | m\_work | 业务代表 | 3 |
| 53 | m\_work | 药厂 | 4 |
| 54 | m\_work | 治疗 | 4 |
| 55 | m\_work | 医生 | 4 |
| 56 | m\_work | 质量控制 | 5 |
| 57 | m\_work | 质量管理 | 5 |
| 58 | m\_work | 资料员 | 6 |
| 59 | m\_work | 银行职员 | 6 |
| 60 | m\_work | 公务员 | 6 |
| 61 | m\_work | 职员 | 6 |
| 62 | m\_work | 文员 | 6 |
| 63 | m\_work | 社区 | 6 |
| 64 | m\_work | 教育行业 | 7 |
| 65 | m\_work | 家教 | 7 |
| 66 | m\_work | 机械制造 | 9 |
| 67 | m\_work | 工人 | 10 |
| 68 | m\_work | 不固定 | 11 |
| 69 | m\_work | 服务 | 2 |

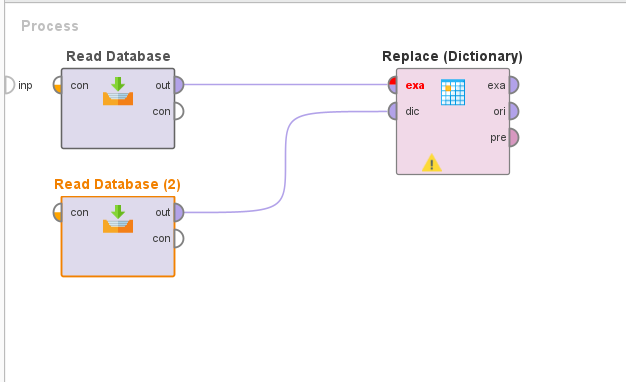
* + 1. For the Read Database(2) operator Parameters, configure the connection so it connects to the table we just created, as shown below.



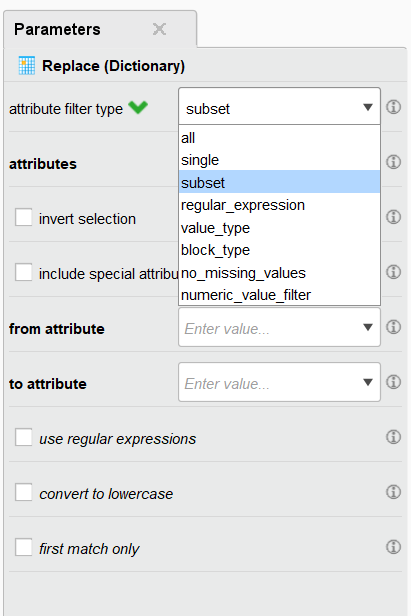
* + 1. Insert the Replace (Dictionary) operator

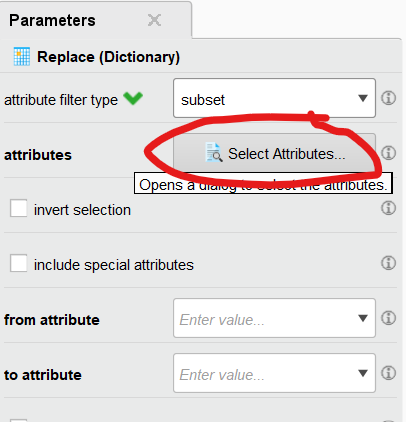


* + 1. Connect Read Database (2) to Replace Dictionary

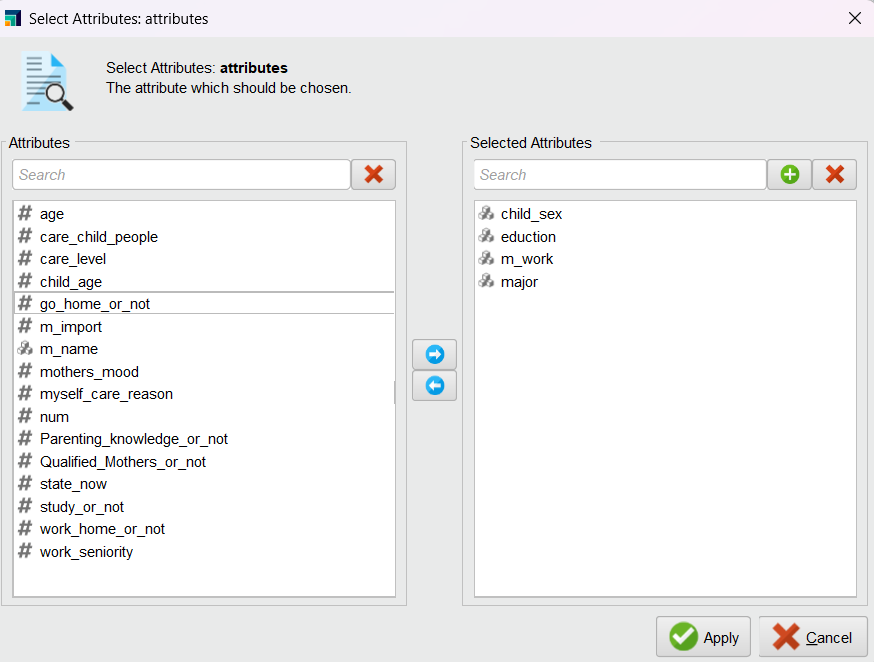


* + 1. Under Replace (Dictionary) Parameters, choose the subset for attribute filter type and click on the Select Attributes button.

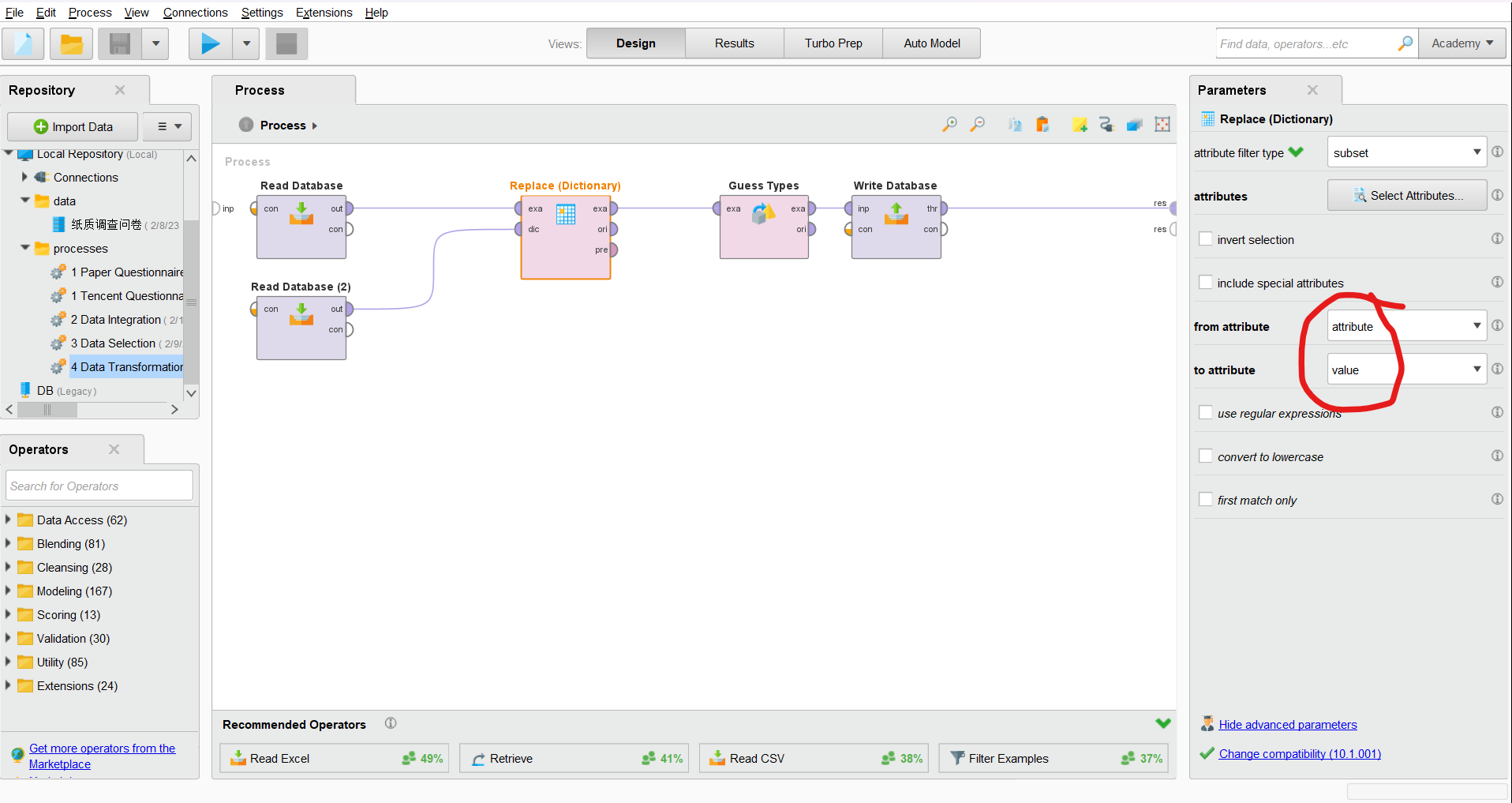




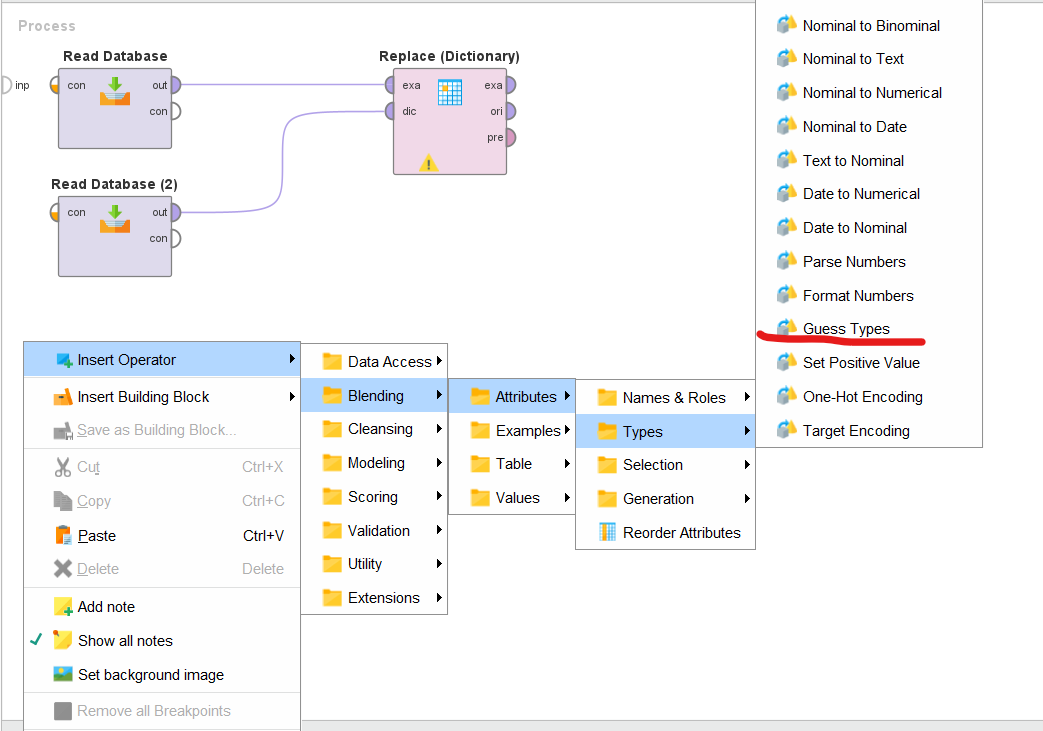
* + 1. Select child\_sex, eduction, m\_work, and major attributes, then click Apply.



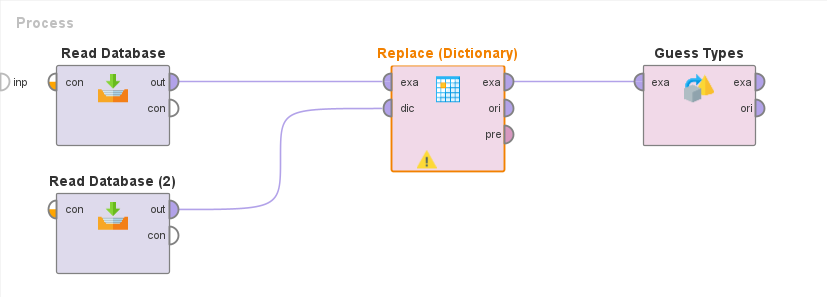
* + 1. Change from attribute to “attribute” and to attribute to “value”



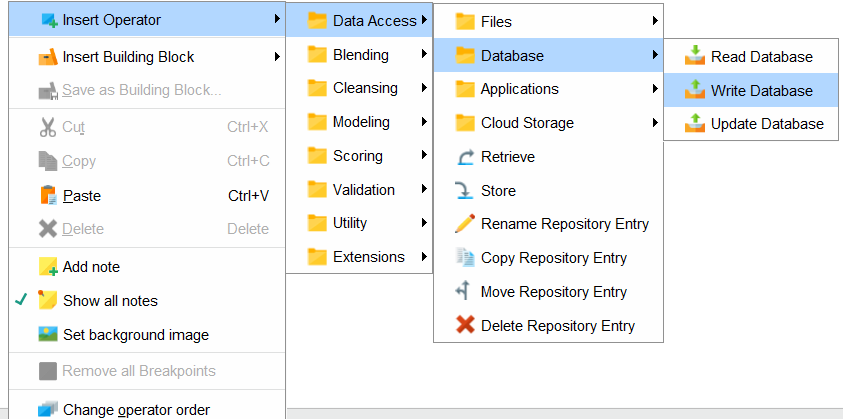
* + 1. Right click -> Insert Operator > Blending > Attributes > types > Guess Types



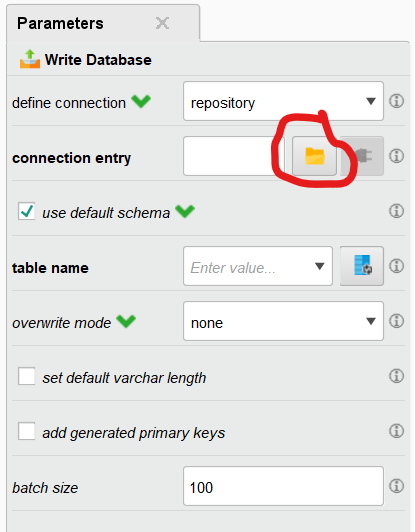
* + 1. Connect the Guess Types exa to the Replace Dictionary exa

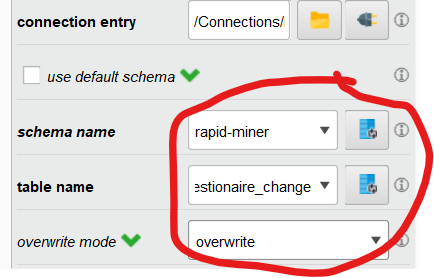
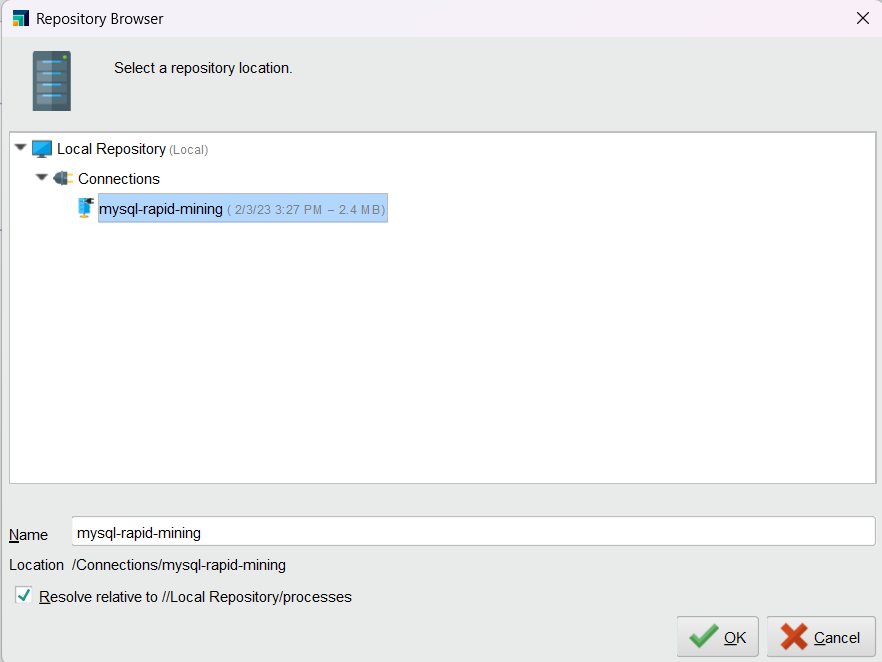


* + 1. Right click -> Insert Operator > Data Access > Database > Write Database

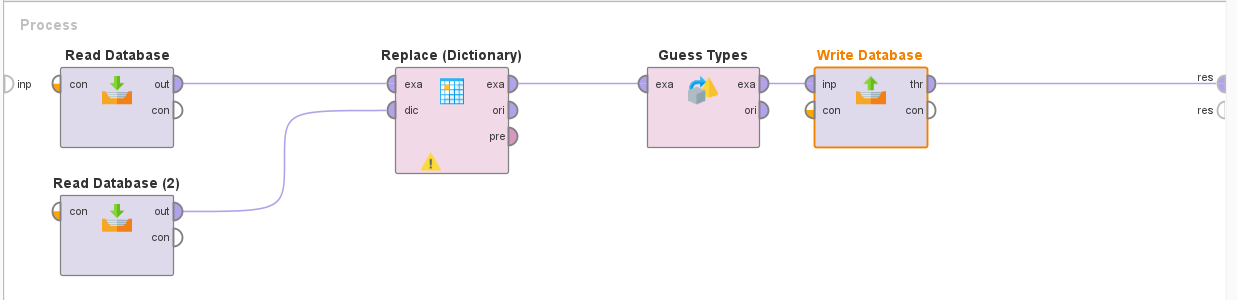


* + 1. Change the Write Database parameters so that the connection entry is properly configured. Make sure the table name is for “questionnaire\_change” and enable overwrite mode unless you already have a table with the correct columns.

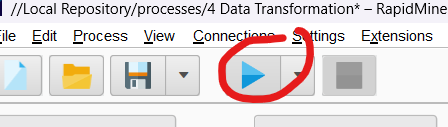




* + 1. Connect Write Database operator to Guess Types and the res endpoint.

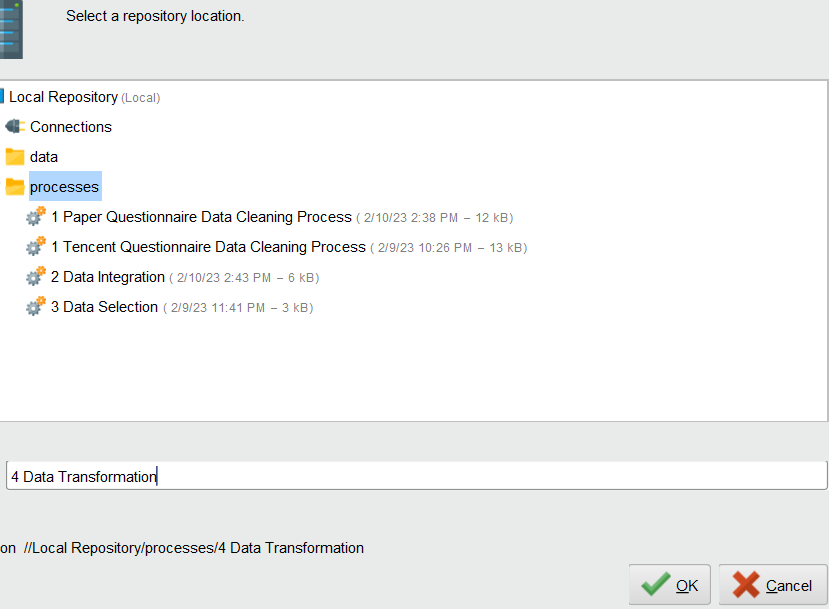


* + 1. Run the process.



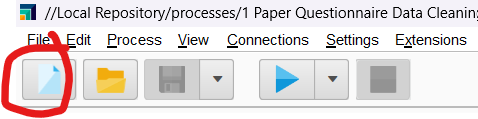
* + 1. If everything ran successfully, click the save button in the upper left corner. We recommend the name “4 Data Transformation”

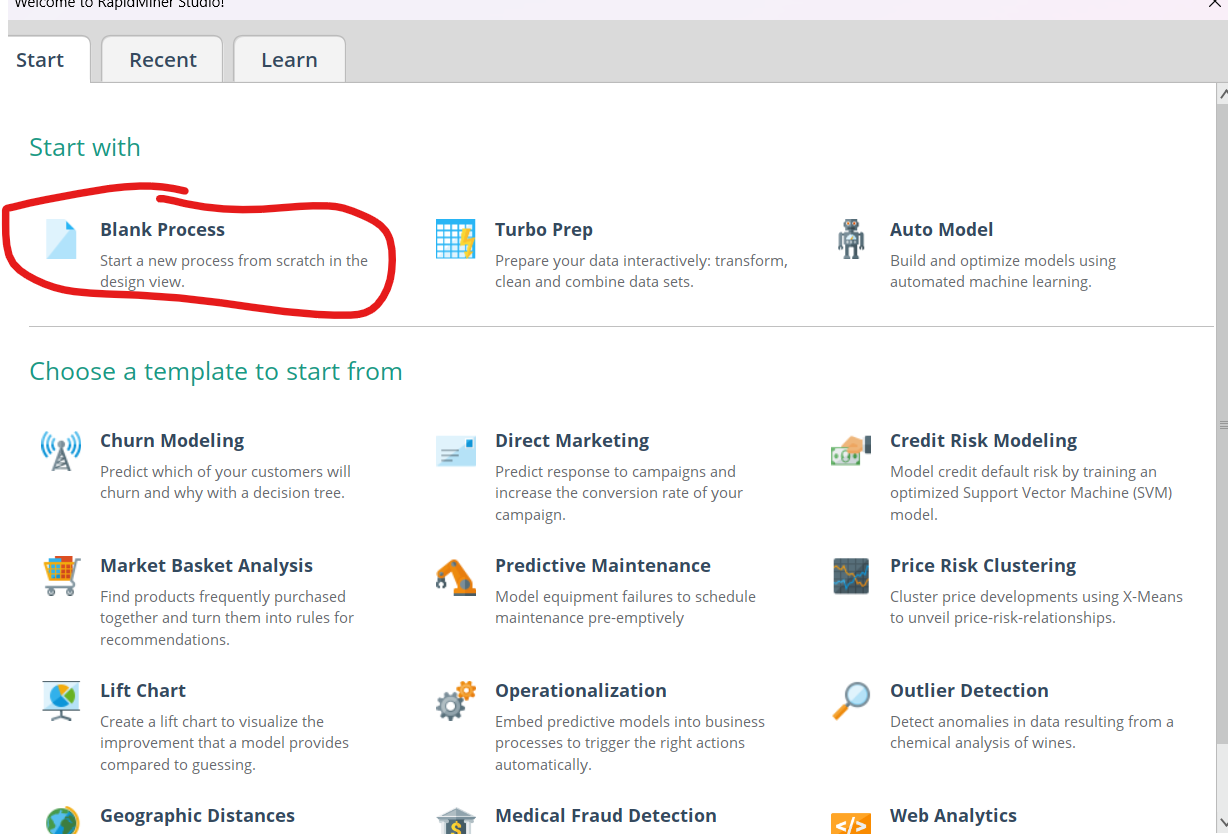




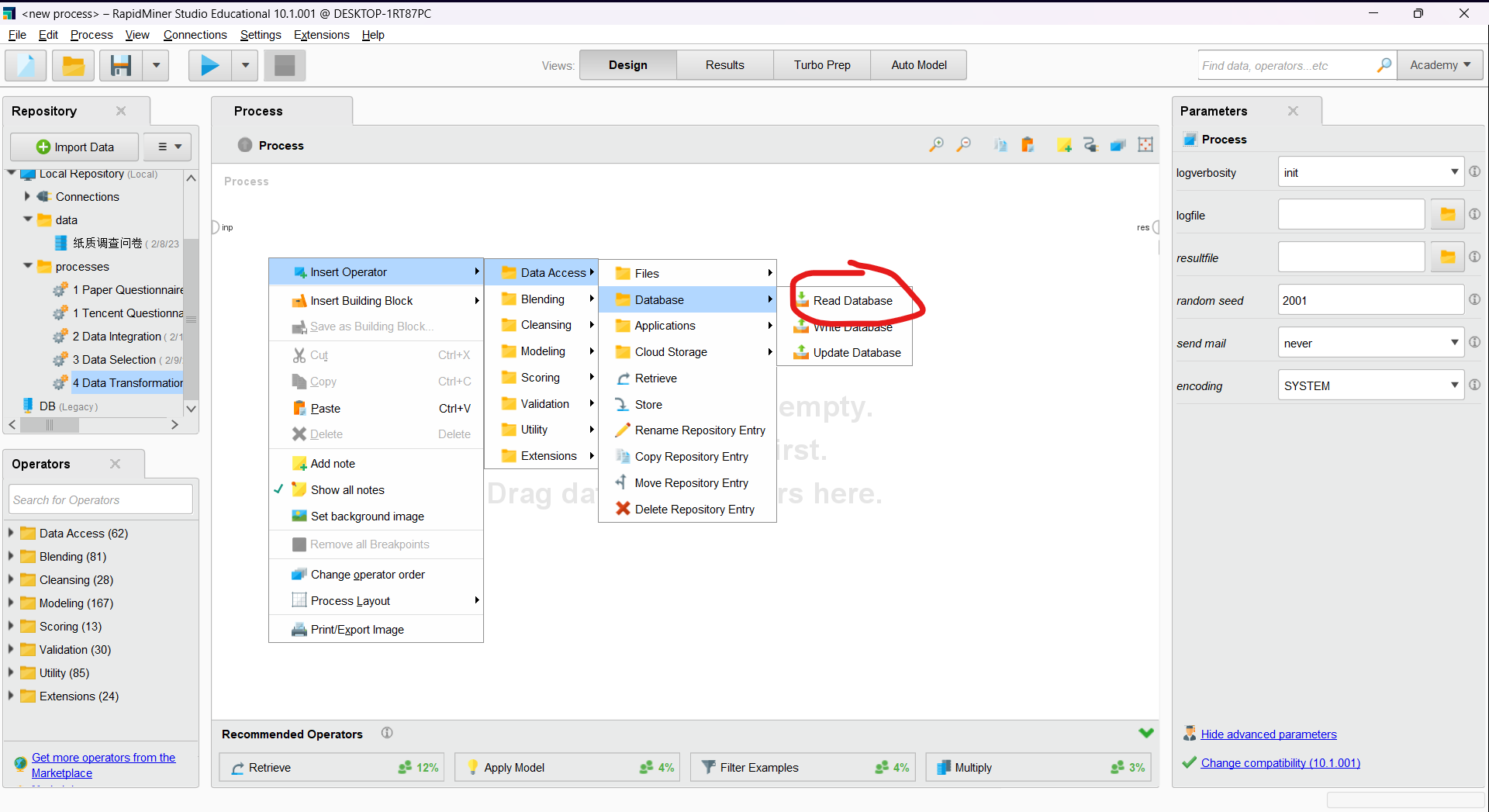
## Outlier Analysis

* + 1. Create a blank new process by clicking on the top left hand corner and selecting the blank process. The purpose of this process is to analyze the data for non-standard data. We may find that the analyzed outliers are really problematic or there might be new findings.

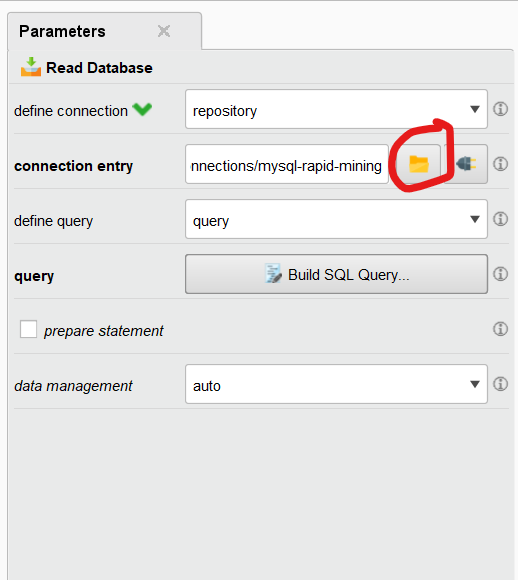




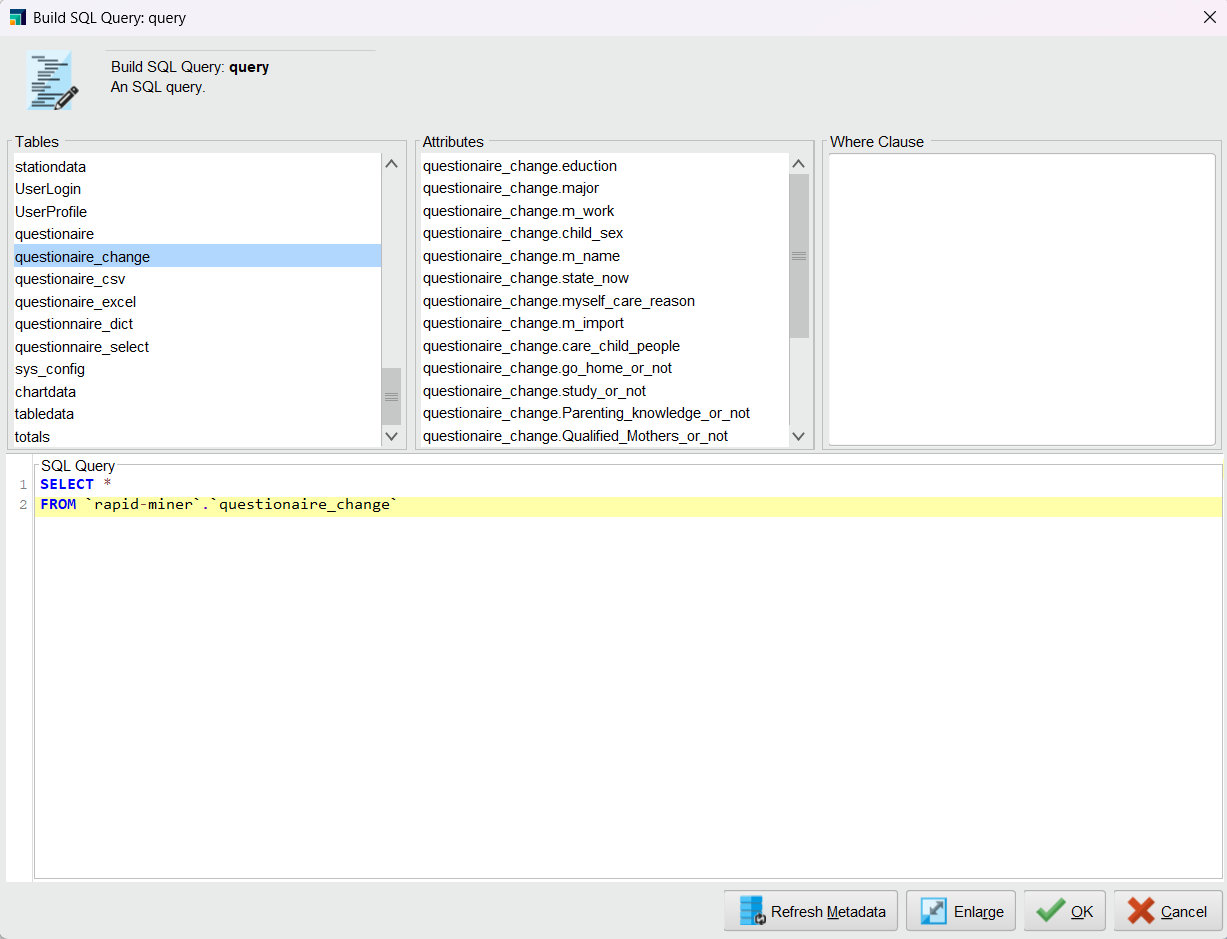
* + 1. Right click -> Insert Operator > Data Access > Database > Read Database



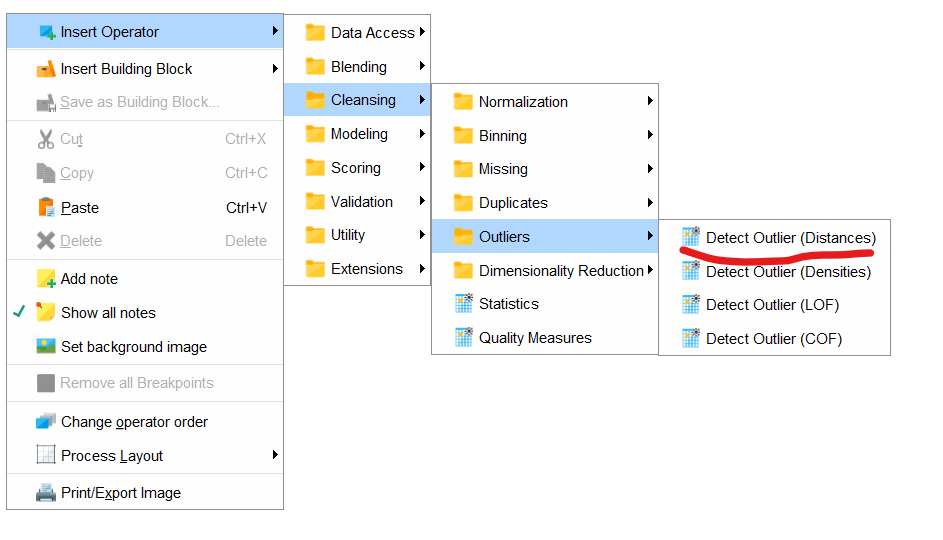
* + 1. Under parameters of Read Database, configure the SQL connection entry like before by clicking on the folder icon.



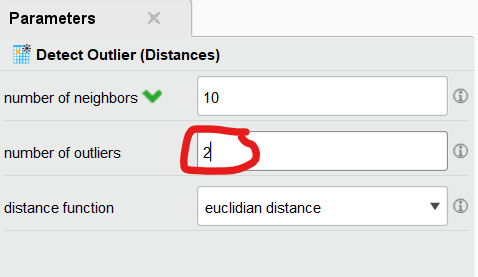
* + 1. Click on Build SQL Query. Look for the “questionnaire\_change” table under the table section. If you do not find the table, try clicking on Refresh Metadata and check again. If you have multiple schemas, you might want to manually alter the SQL statement to include the schema name before the table name, as shown in the picture below. Click OK when you’re done



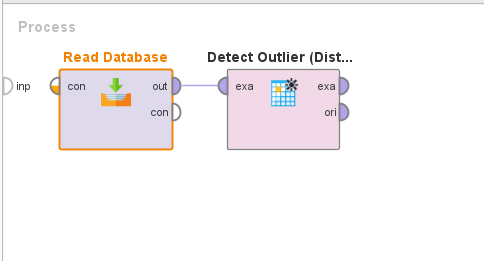
* + 1. Right click => Insert Operator > Cleansing > Outliers > Detect Outlier (Distances)



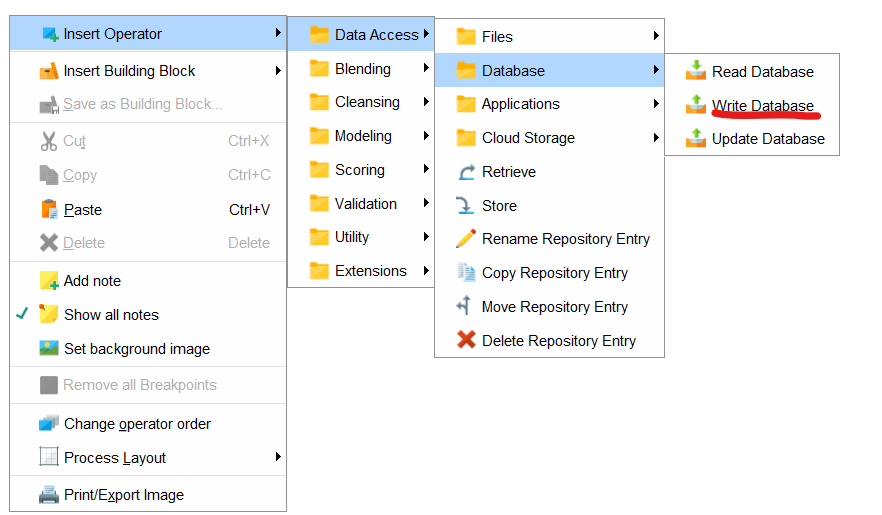
* + 1. Change the number of outliers to 2 for the Parameters of the Detect Outliers (Distances) operator. We do this so we don’t eliminate all of our data due to it being an “outlier”



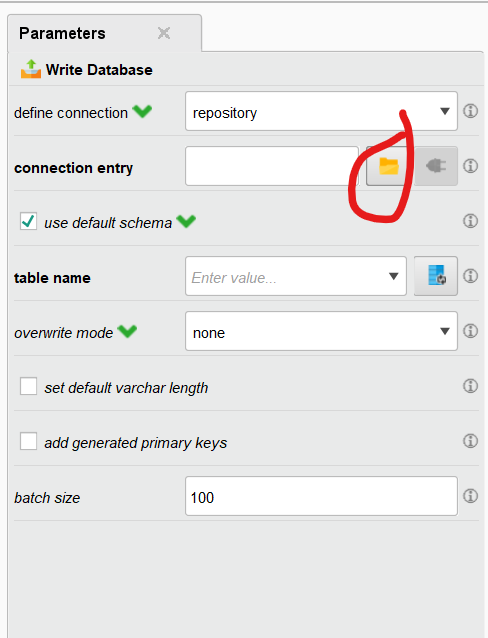
* + 1. There should now be 2 operators connected with each other, like below.

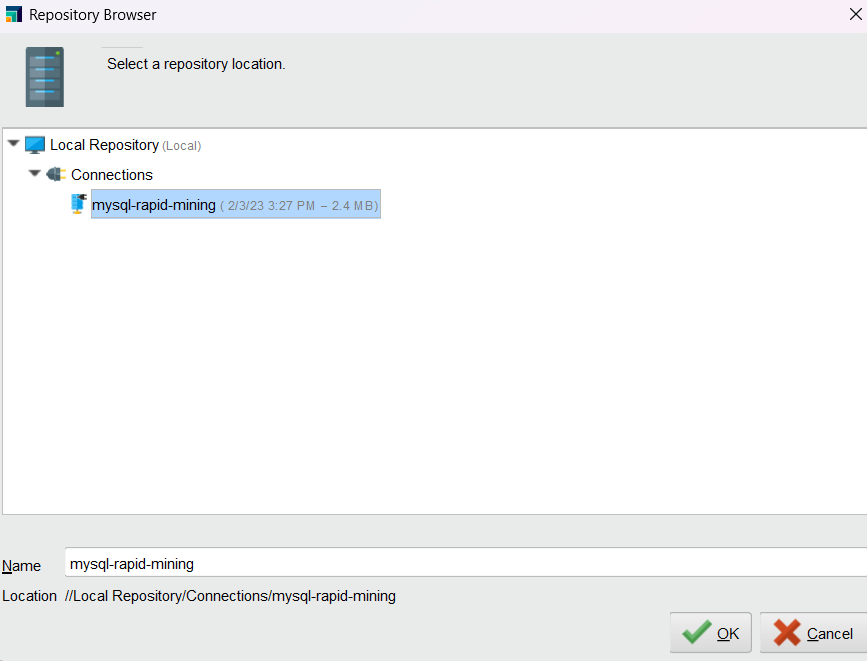


* + 1. Right click -> Insert Operator > Data Access > Database > Write Database

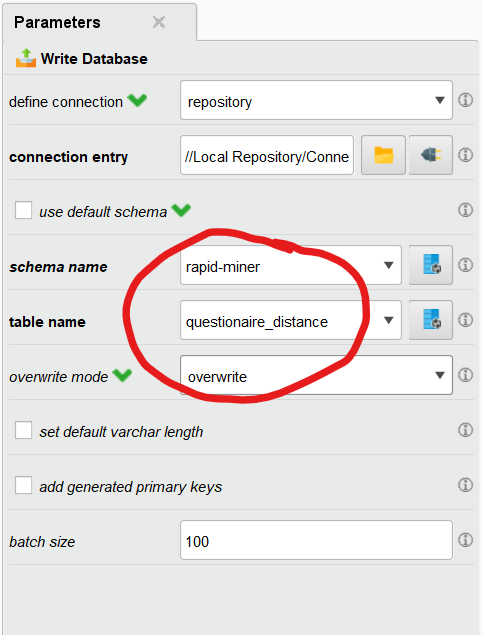


* + 1. Configure the MySQL connection like before by clicking on the folder icon for connection entry in the Parameters of Write Database.

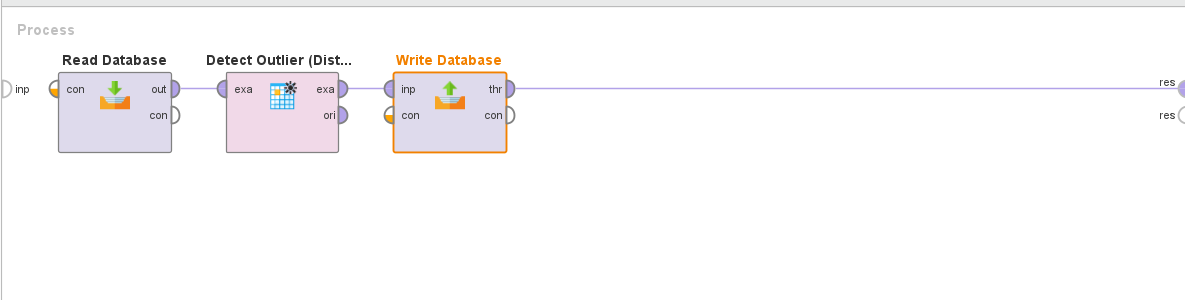




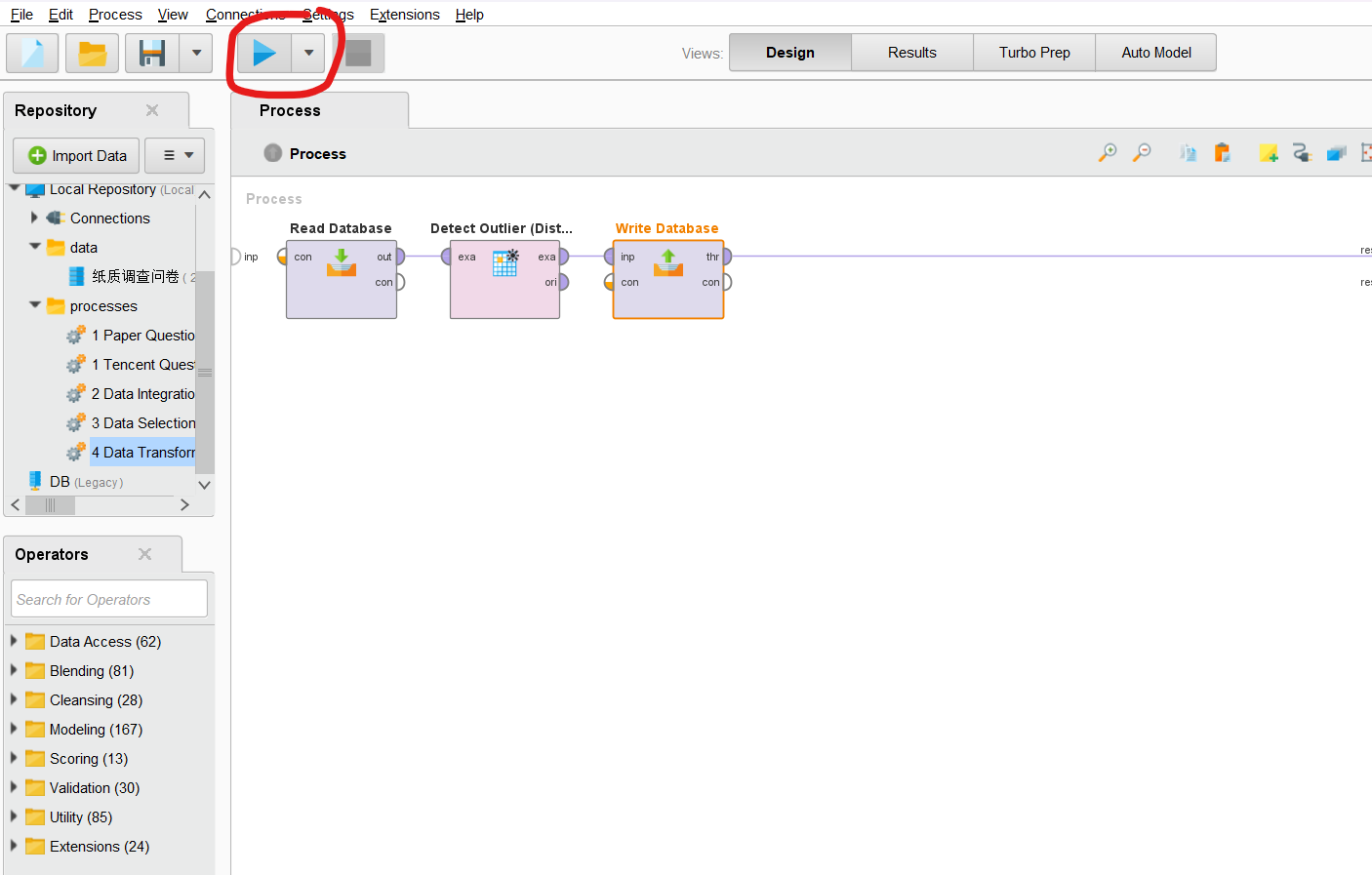
* + 1. Change the schema name and table name to the schema that contains the table we created in our previous challenge (“questionnaire\_distance”). Enable overwrite mode unless a table already exists with the correct format and column names.



* + 1. Connect the Write Database operator so the final operator graph looks like the below.

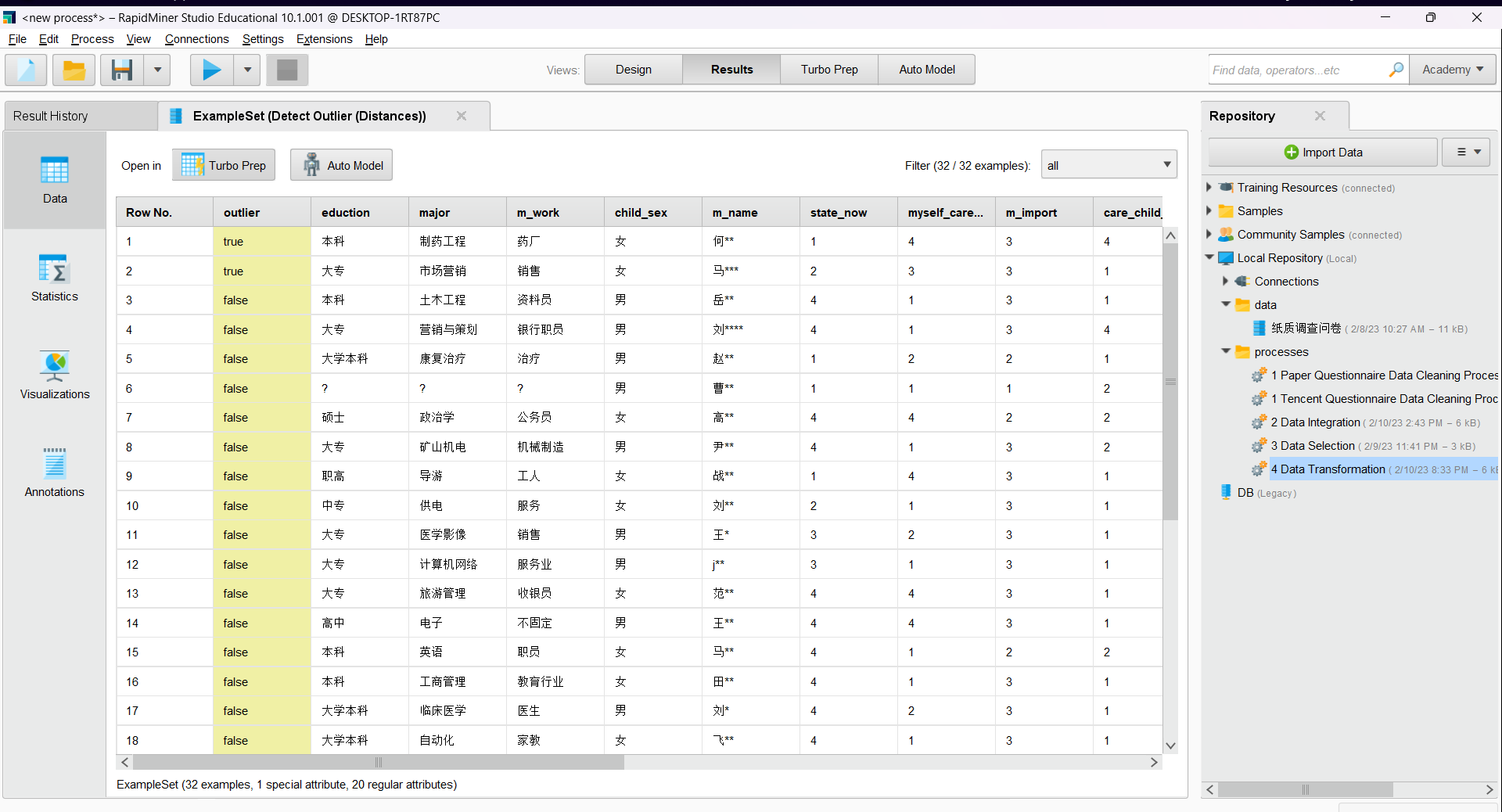


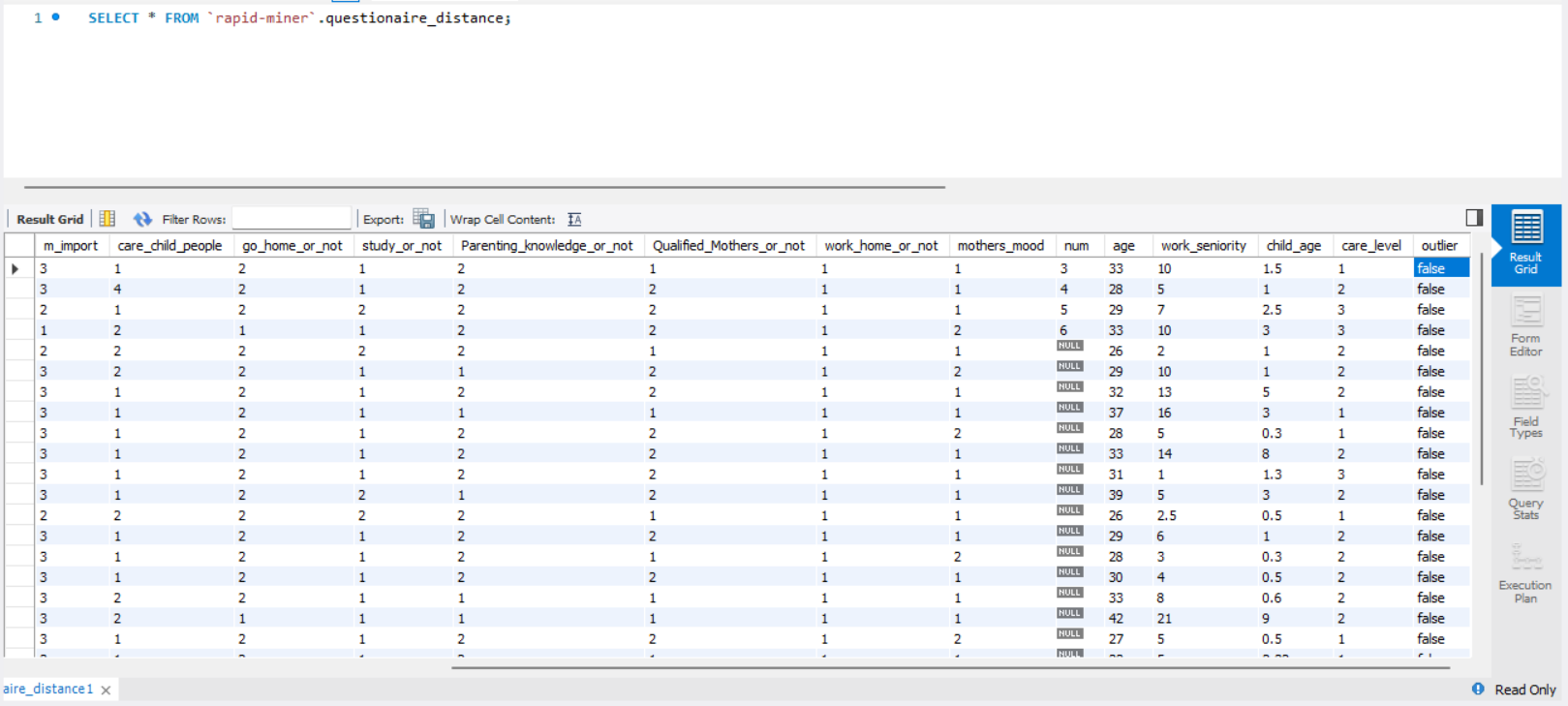
* + 1. Run the process by clicking on the top left hand corner.



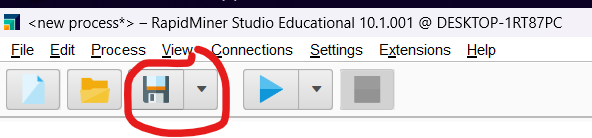
* + 1. After executing the process, 2 outliers were found". Among them, "吴\*\*" has a working life of 21 years which is higher than the working years of other people. But the person is also very old at the same time, so this outlier is reasonable.

Another "马\*" has an age of 7 years old. A 7-year-old Bao Ma, obviously there is a problem with the data. It’s very likely that the respondent wanted to write 27 instead of 7. We can remove this as well.

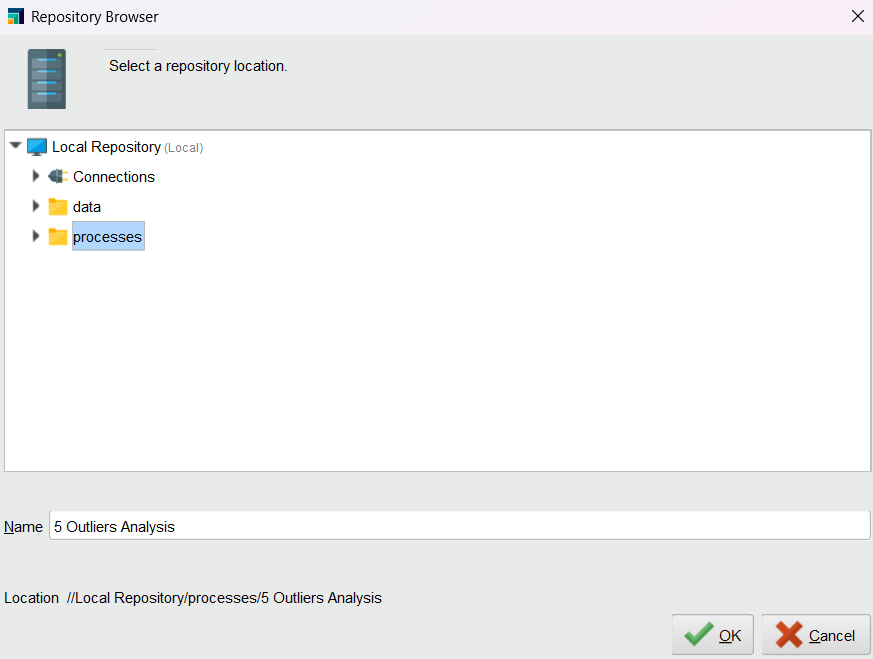




* + 1. Save the process by clicking on the button in the top left corner.

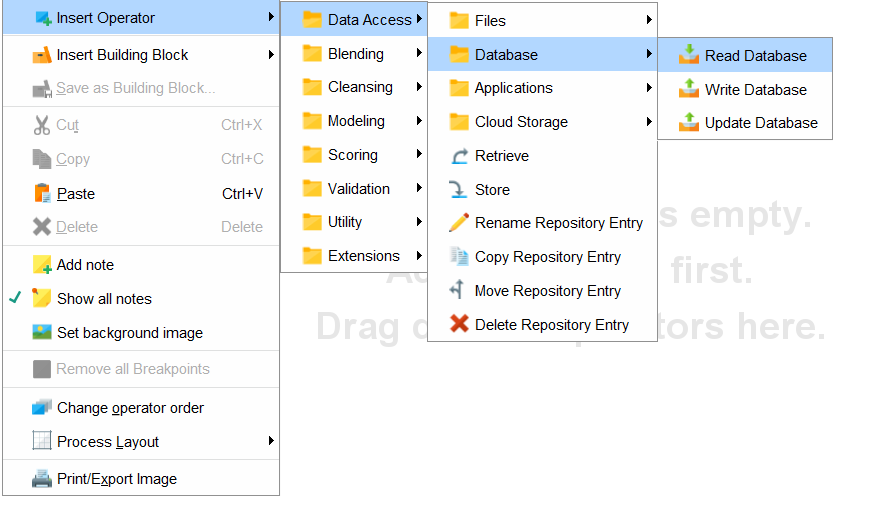


* + 1. It is recommended to name the process “5 Outlier Analysis”

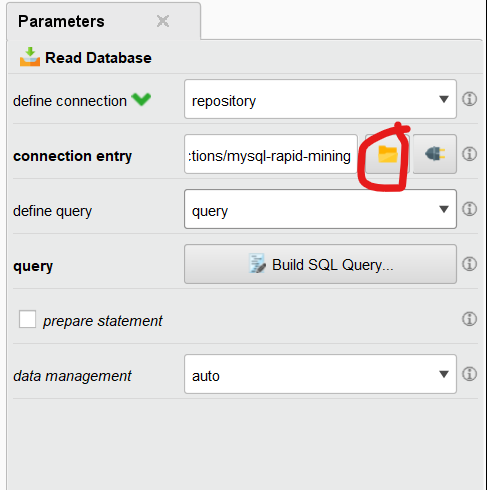


## Handling Outliers

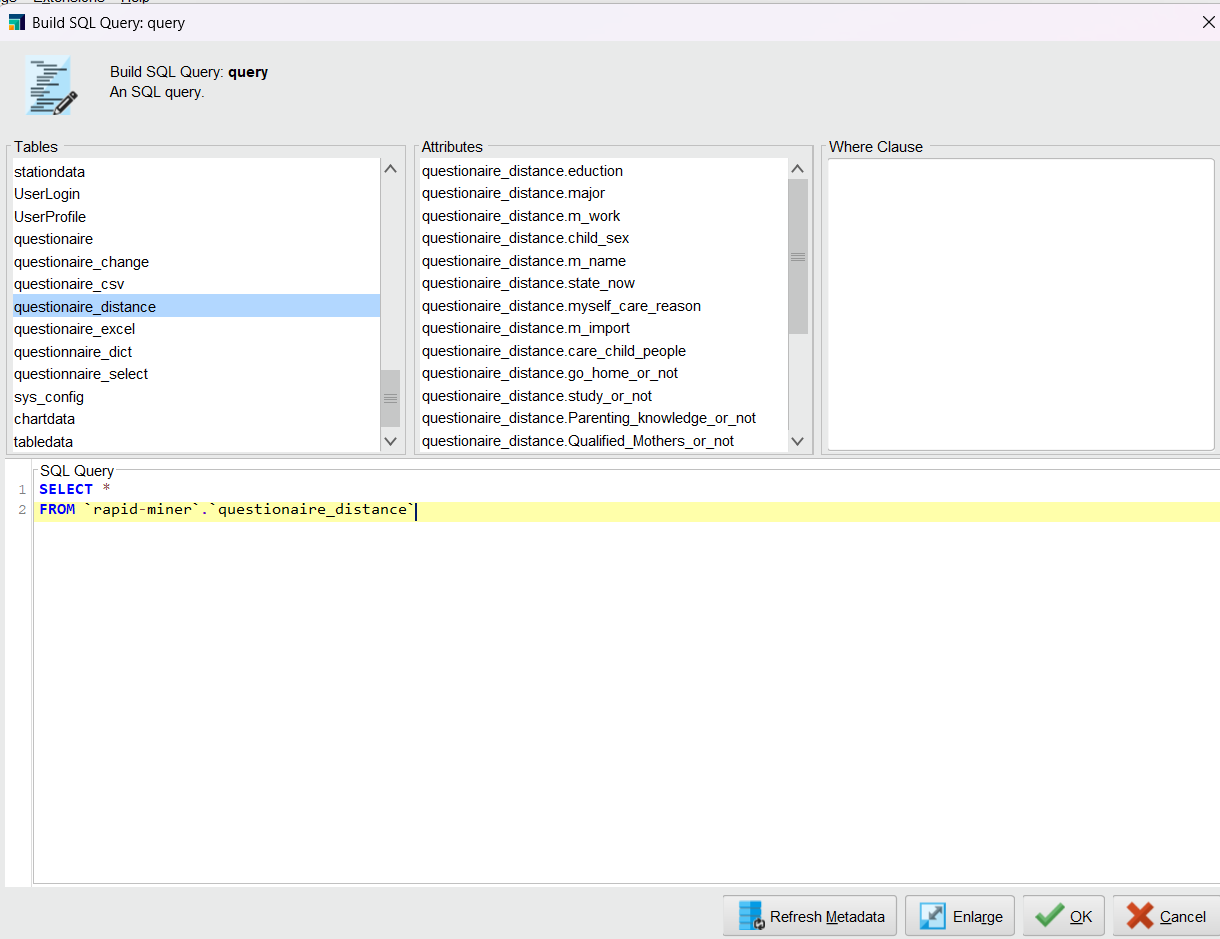
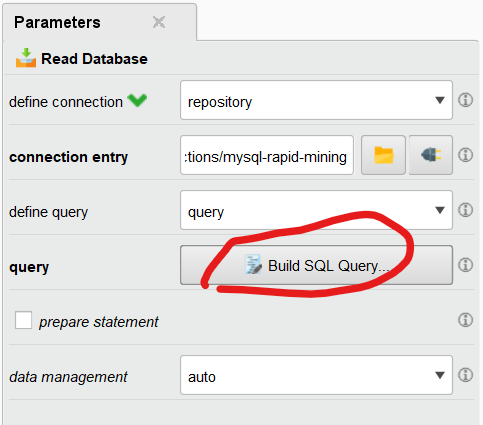
* + 1. Create a new process like before. The purpose of this process is to read the results from the 5th process, which is in the “questionaire\_distance” table and remove outliers so our model can perform better.
    2. Right click => Insert Operator > Data Access > Database > Read Database



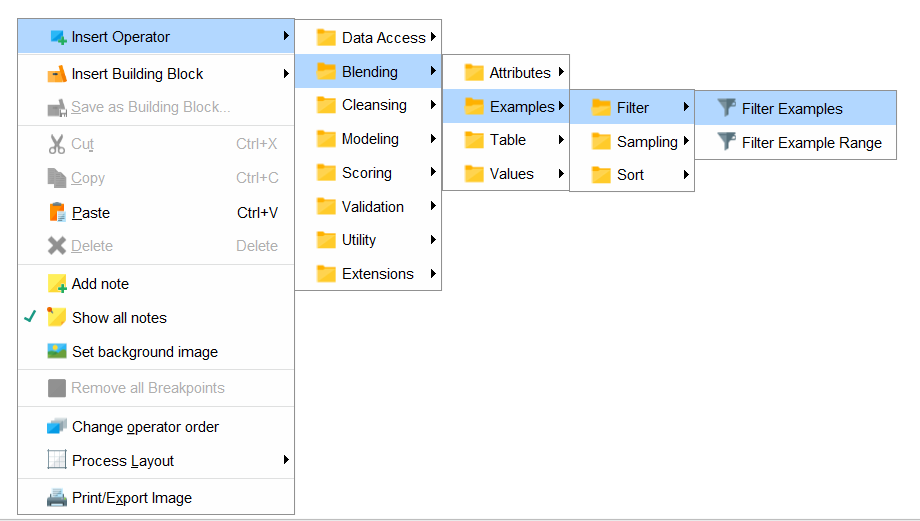
* + 1. Click on the Read Database operator and navigate to its Parameters pane. Click on the folder icon for connection entry and configure the SQL connection like before.



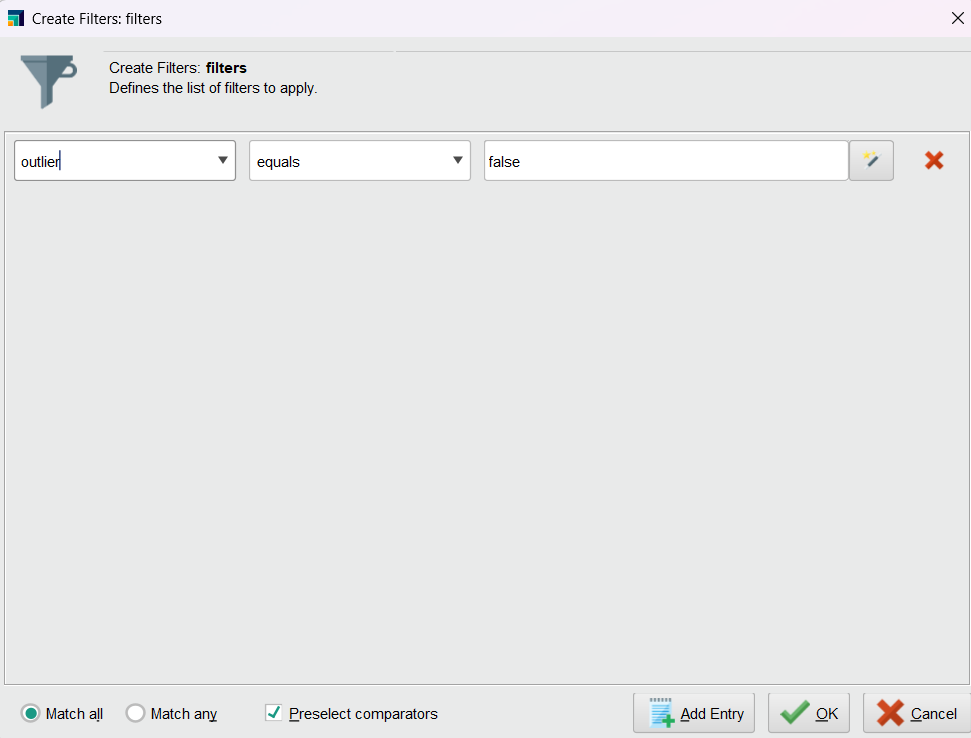
* + 1. Click on Build SQL Query. Look for the “questionnaire\_distance” table under the table section. If you do not find the table, try clicking on Refresh Metadata and check again. If you have multiple schemas, you might want to manually alter the SQL statement to include the schema name before the table name, as shown in the 2nd picture below. Click OK when you’re done.



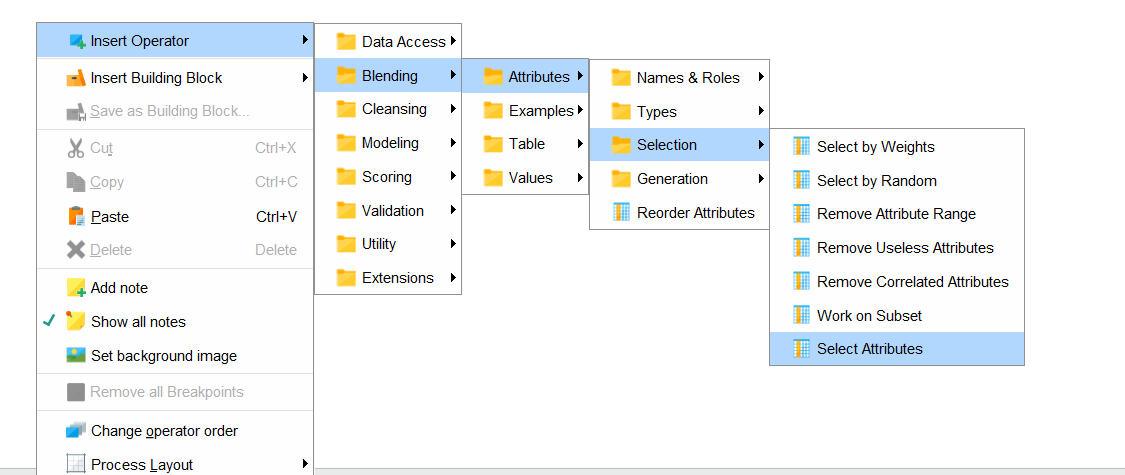
* + 1. Right click => Insert Operator > Blending > Examples > Filter > Filter Examples



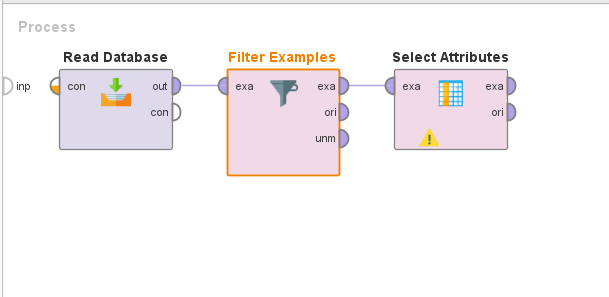
* + 1. Drag a line from Read Database to Filter Examples.
    2. Click on Add Filters. Set outlier equals to false. Here we’re going to remove the outliers from the dataset so that our model will perform better. Then click OK.



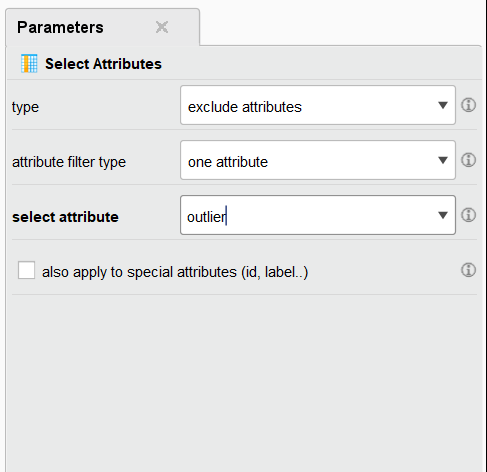
* + 1. Right click => Insert Operator > Blending > Attributes > Selection > Select Attributes



* + 1. Connect Filter Examples exa with Select Attributes exa.



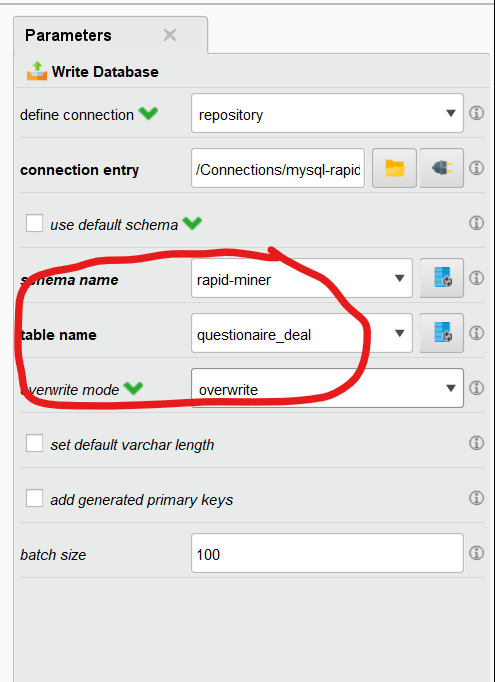
* + 1. Change Select Attributes Parameters to below. We’re doing this so we don’t include the outlier column as a feature in the model.



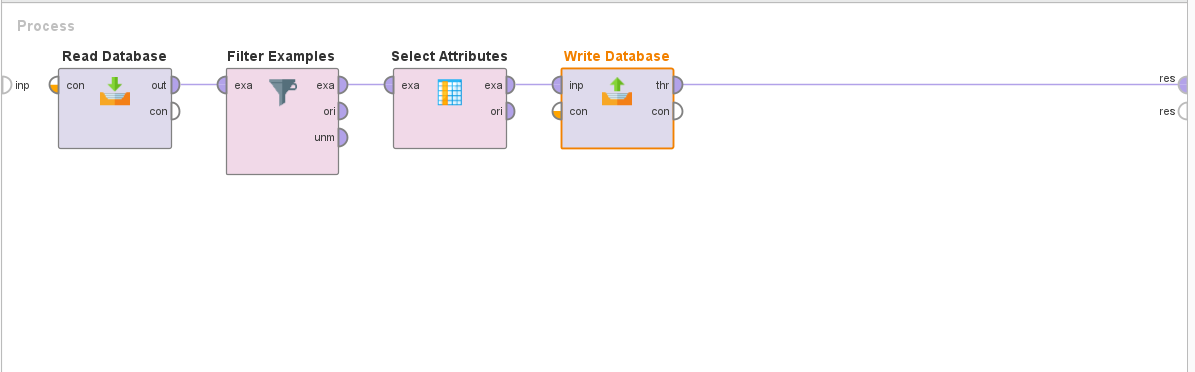
* + 1. Right click => Insert Operator > Data Access > Database > Write Database



* + 1. Change the parameters of the Write Database operator to below.



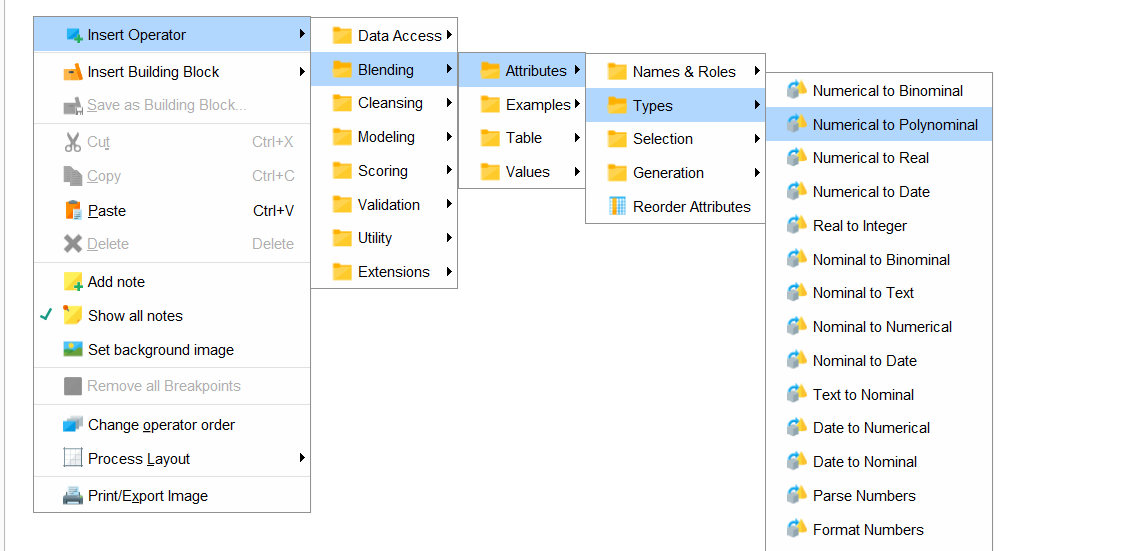
* + 1. Your final process should look like this. Try running it, and if it works, save it as “6 Handling Outliers”



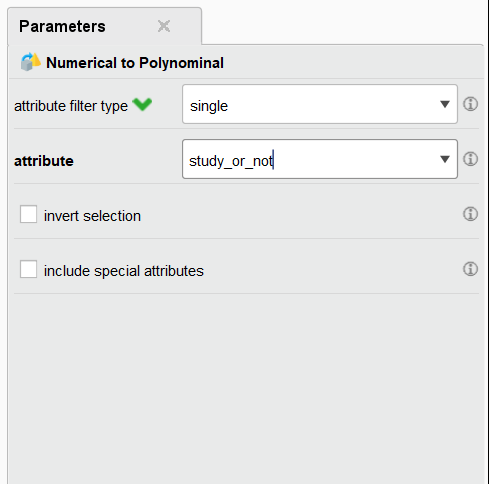
# Modeling

## Decision Tree Modeling

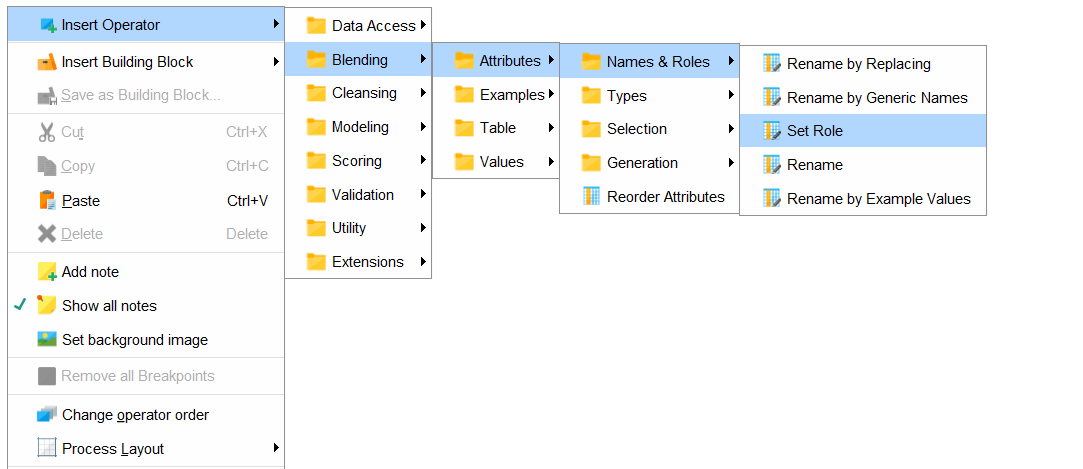
* + 1. Create a blank process.
    2. Insert a Read Database operator. Configure the connection so that it connects to the “questionnaire\_deal” table.
    3. Right click > Insert Operator > Blending > Attributes > Types > Numerical to Polynomial



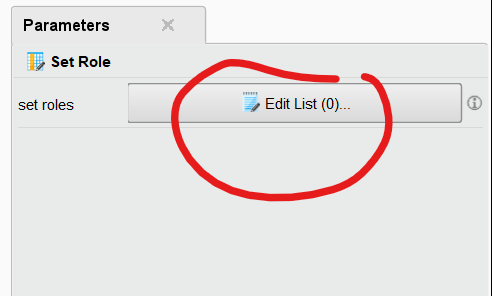
* + 1. Connect the Read Database operator to the Numerical to Polynomial operator.
    2. Set the parameters of Numerical to Polynomial operator to have the attribute filter type = “single” and attribute = “study\_or\_not”



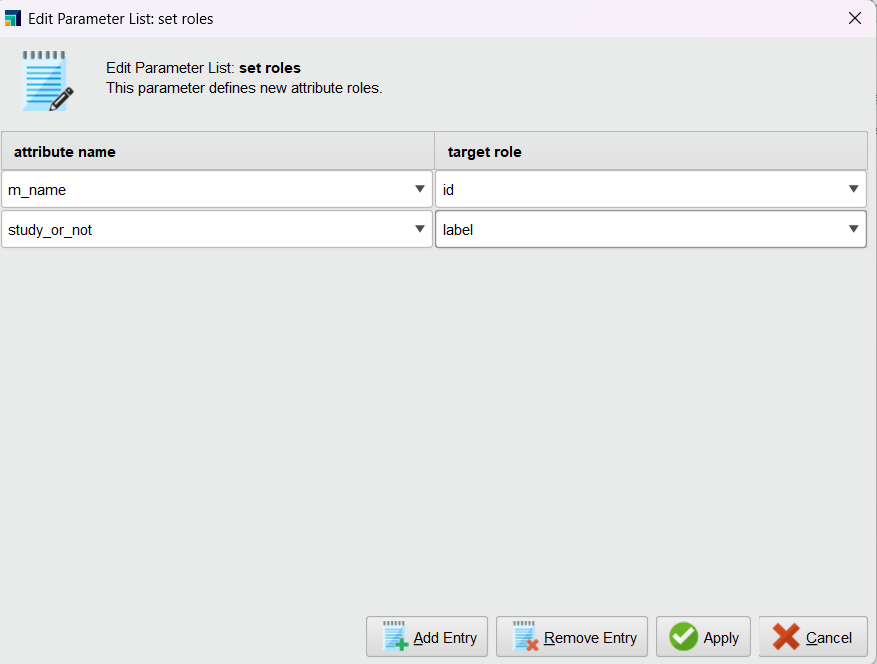
* + 1. Right click > Insert Operator > Blending > Names & Roles> Set Role



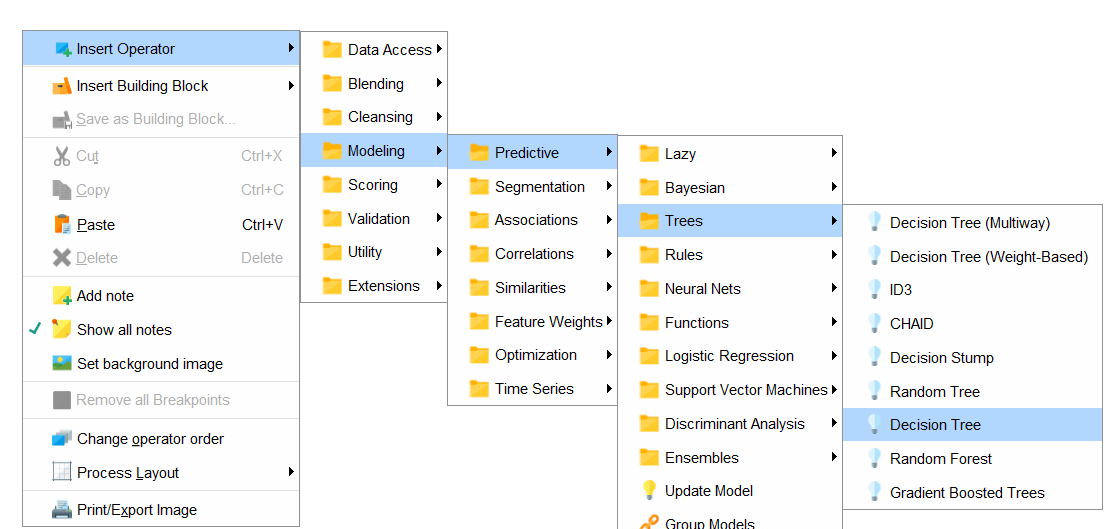
* + 1. Drag a line from Numerical to Polynomial to Set Role. For the parameters of Set Role, click on Edit List.



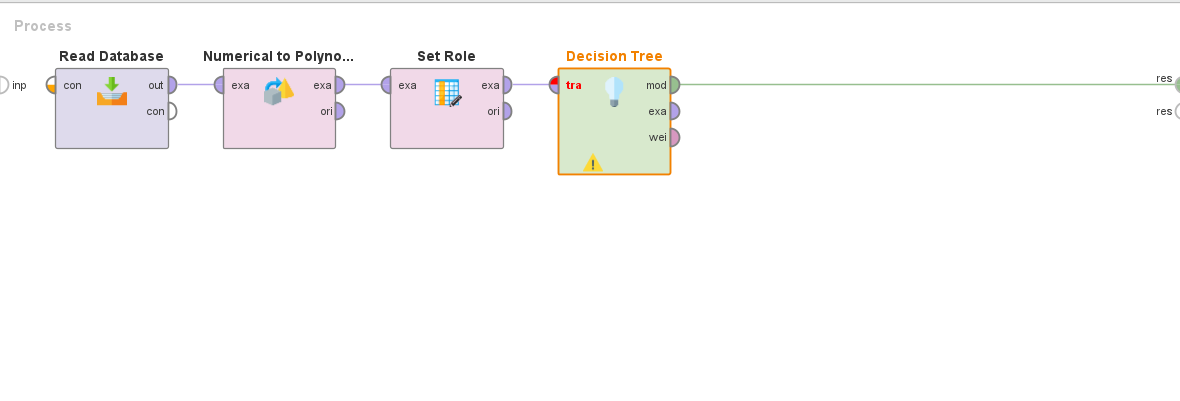
* + 1. Set the m\_name attribute to have the target role of id and study\_or\_not to the target role of label. Then click Apply.



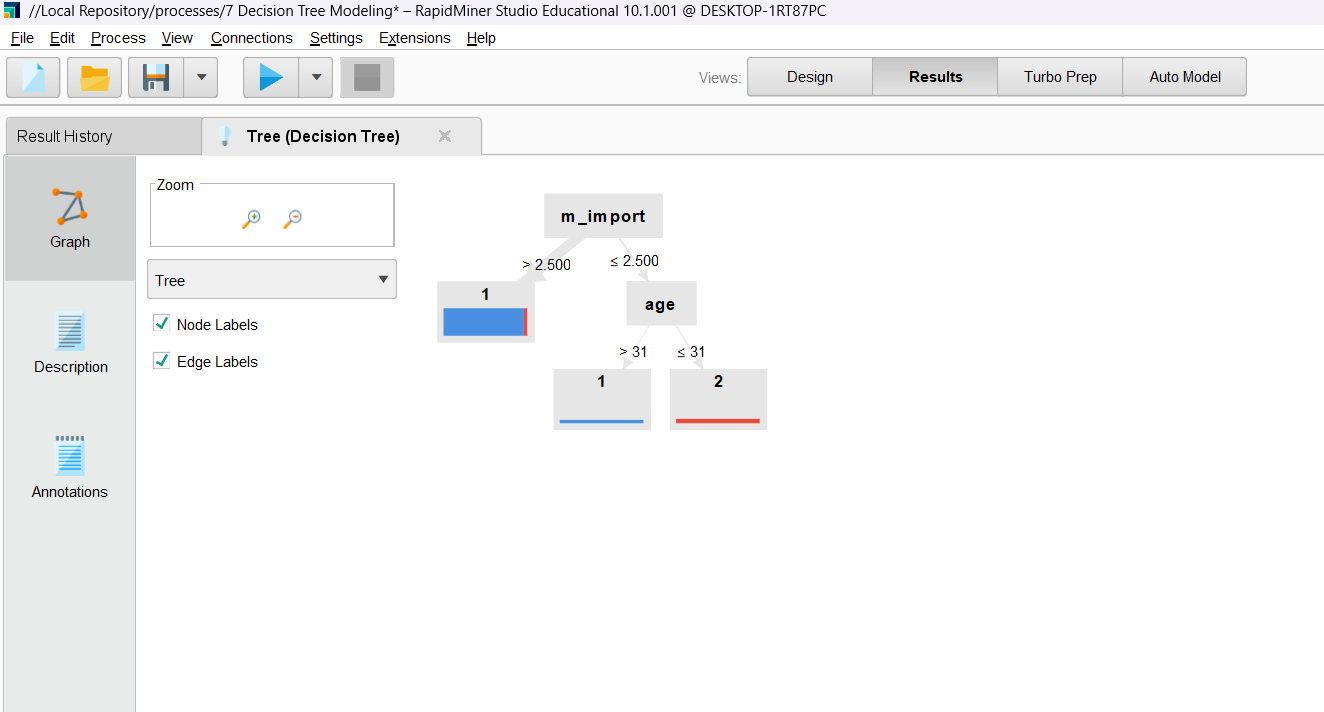
* + 1. Right click -> Insert Operator > Modeling > Predictive > Trees > Decision Tree.



* + 1. Connect the Set Role operator to the Decision Tree operator. Then connect the Decision Tree to the res endpoint so that the chart looks as follows:

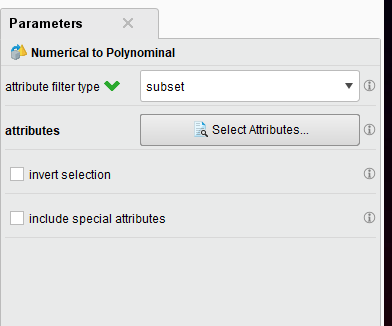


* + 1. Run the process. If it ran correctly, the Results should show a visual Decision Tree. Save it with the name “7 Decision Tree Modeling”



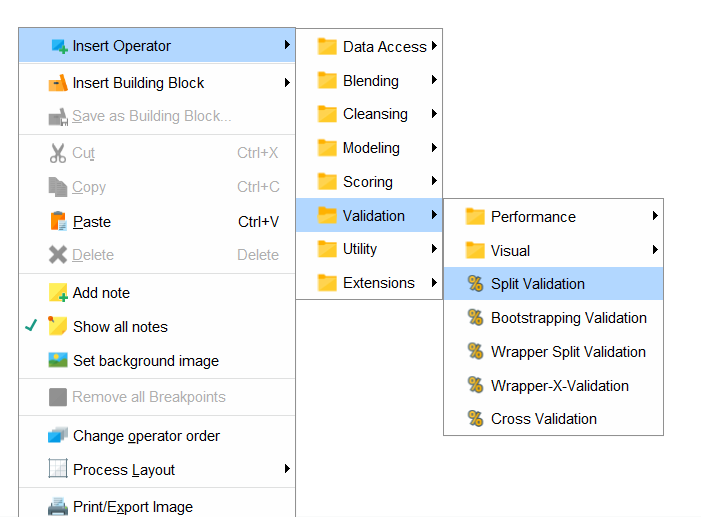
# Evaluation

* 1. Evaluate Decision Tree Model
     1. Create a new process in Rapidminer named "8 Decision Tree Evaluation". The main purpose of this process is to judge which model is most in line with reality. We need to evaluate the built model and select the one with the most accurate prediction by evaluating the results.
     2. Add a Read Database operator so we can read the questionnaire\_deal table from the database.
     3. Add a Numerical to Polynomial operator and connect it to the Read Database operator, just like you did in the last process.
     4. Change the parameters of Numerical to Polynomial operator so that attribute filter type is set to subset. Click Select Attributes and select all attributes except for child\_age, age, num, work\_seniority.

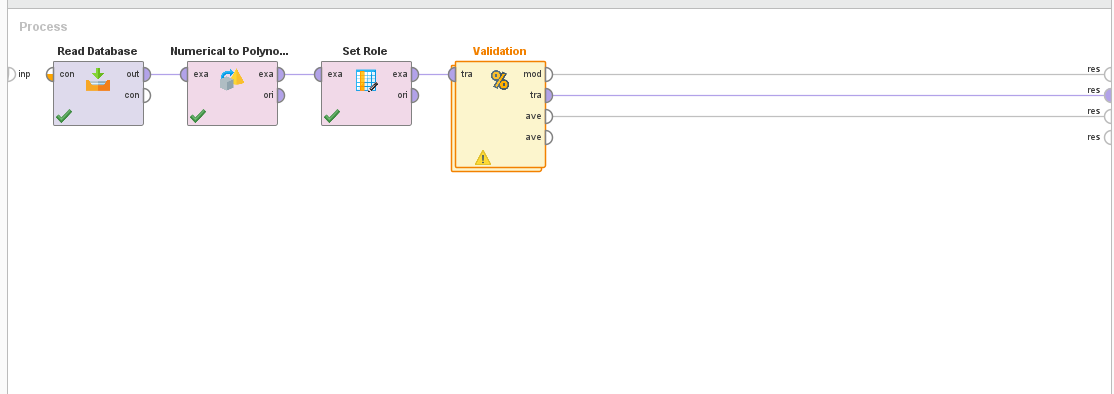




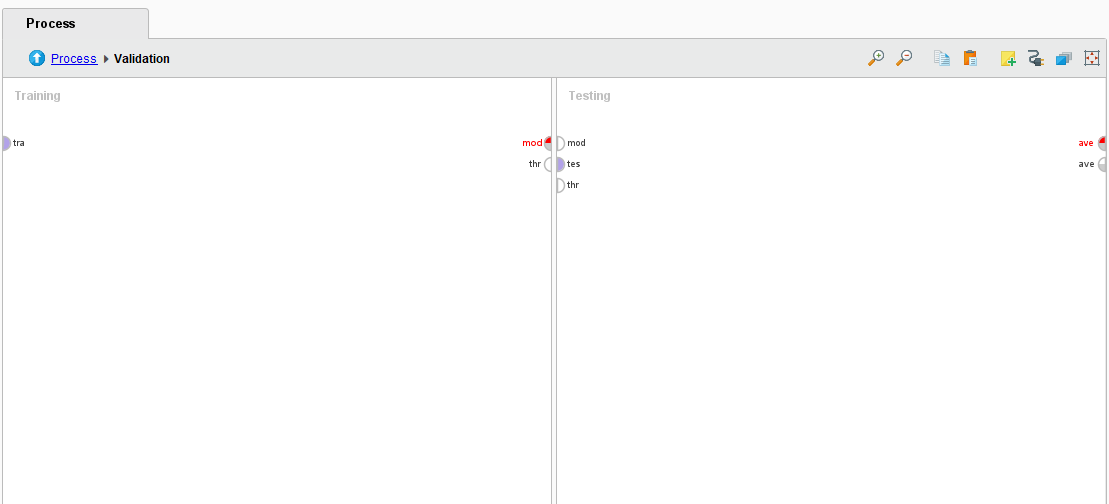
* + 1. Right click and add a Set Role operator like our last process, and set the role of m\_name to id and the role of study\_or\_not to label. Connect the Set Role operator to the Numerical to Polynomial operator. (These steps are the same as our 7th process)
    2. Right click => Insert Operator > Validation > Split Validation and connect the Split Validation operator to the Set Role operator.



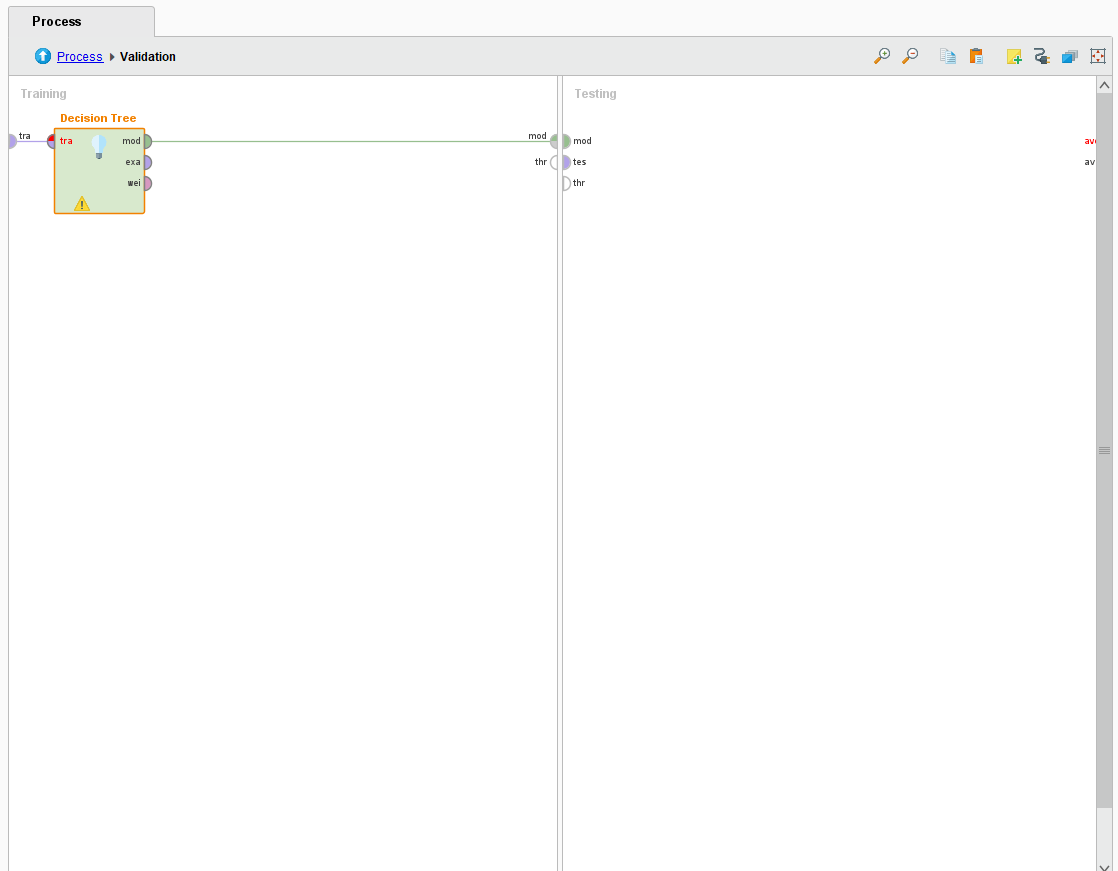
* + 1. Use the default parameters for the Split Validation operator. This sets the distribution ratio between training data and testing data to 0.7, which means that this control treats 70% of the data as training data and 30% as test data.
    2. Connect the Validation operator to 3 res endpoints as shown below.



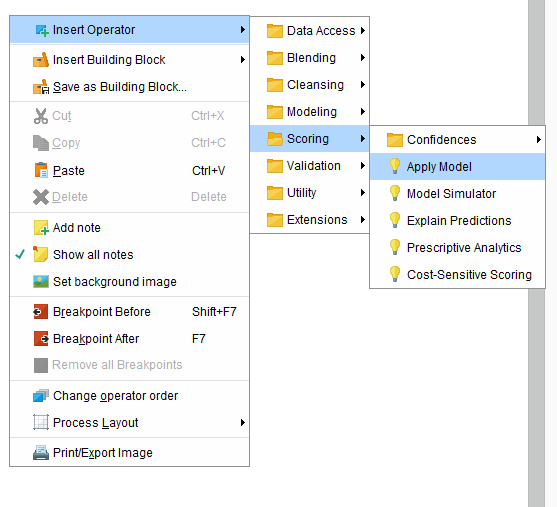
* + 1. Run the process, and the following interface should show up.



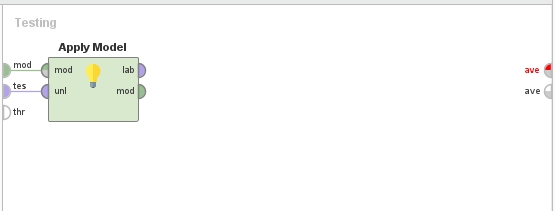
* + 1. Add a Decision Tree in the left Training panel. Then connect it to the tra starting point and the mod end point.



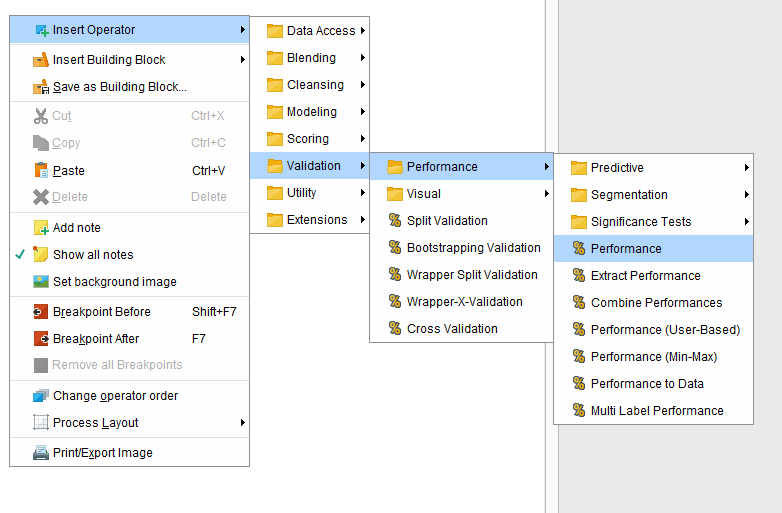
* + 1. Right click in the Testing panel => Insert Operator > Scoring > Apply Model



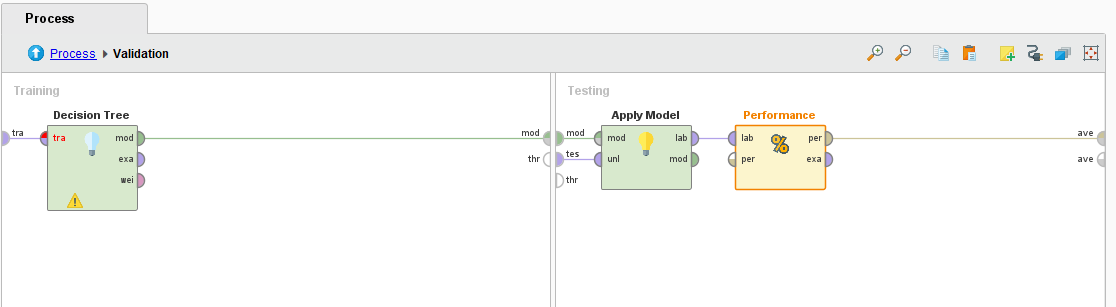
* + 1. Create two connections from mod to mod and tes to uni. It should now look like the following.



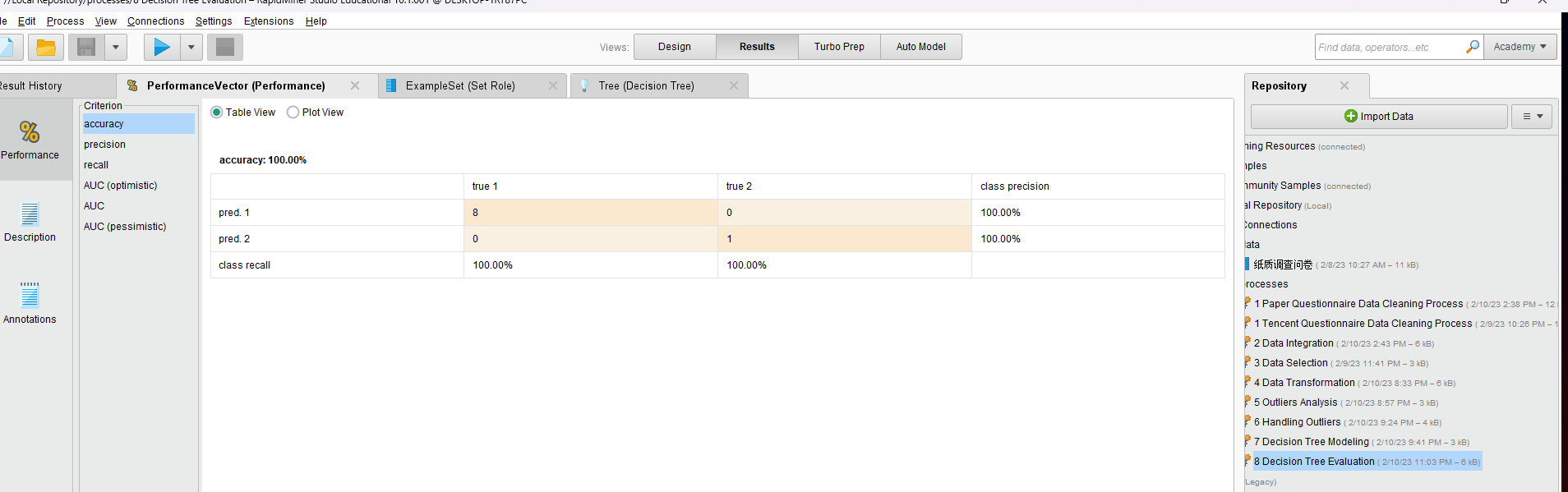
* + 1. Right click on Testing panel => Insert Operator > Validation > Performance > Performance



* + 1. Connect the Performance operator to the Apply Model operator and the ave endpoint. The entire interface should now look something like this.

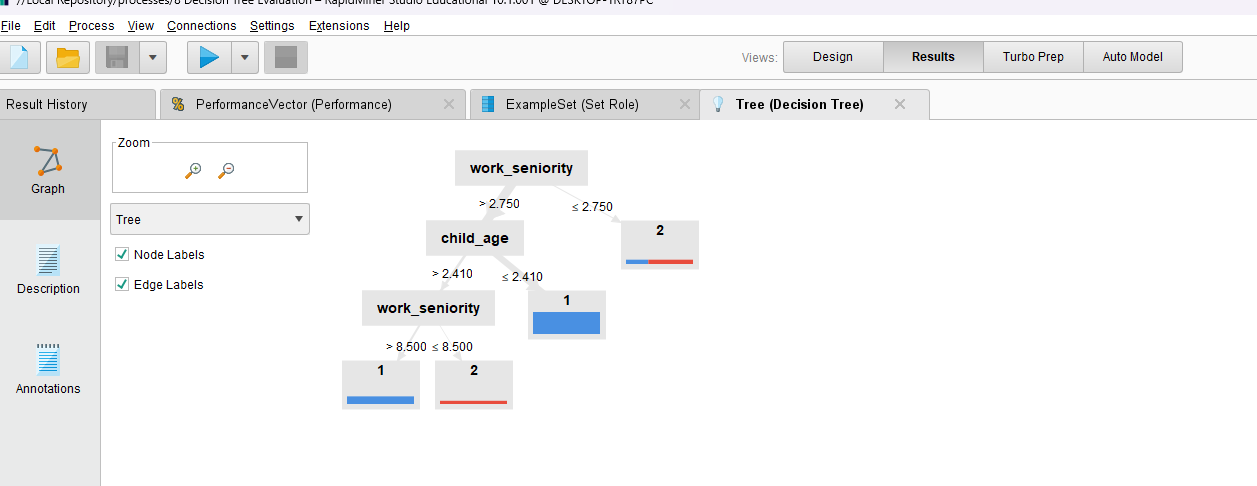


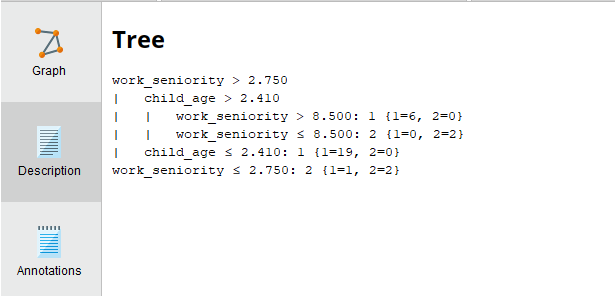
* + 1. Now, try running the process again. The results should show the following. Save this process as “8 Decision Tree Evaluation”



# Deployment

## Visual Display





# Summary

In conclusion, data mining is an important part of decision making in today's world. Through this project, we explored various techniques and algorithms to extract valuable insights. We have seen how data mining can help us make sense of sample amounts of data from user surveys. As the amount of data grows, the importance of data mining will only increase, making it a valuable skill to have in the years to come. Through this assignment we have better learned to use the tools and techniques utilized to clean, process and interpret data.