Lab Work Report

Data Preparation & Processing

SYEINRITA DEVI A/P ANBEALAGAN

MCS221022

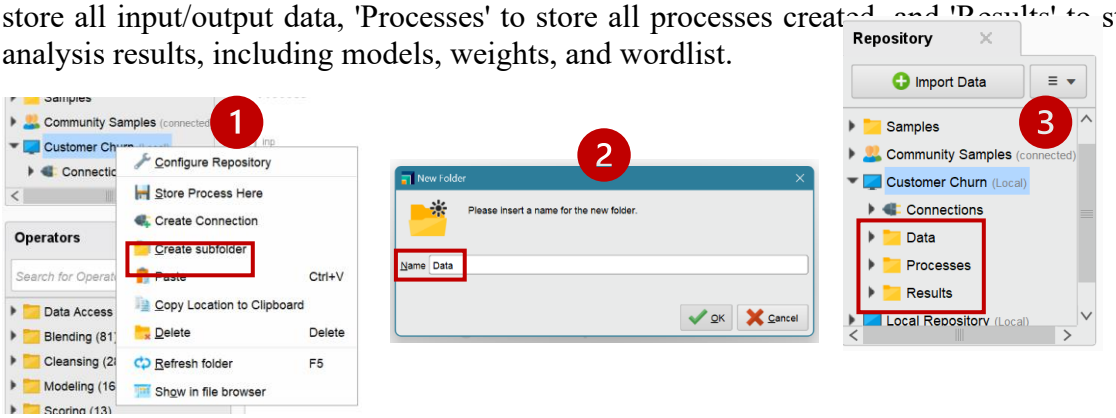
Table of Contents

INTRODUCTION

For this report, a data set titled "Customer Churn" is utilised to do preprocessing in order to get the data ready for analysis using the RapidMiner software. This report will outline the steps used throughout the data preparation.

EXERCISE: CREATE SUBFOLDER

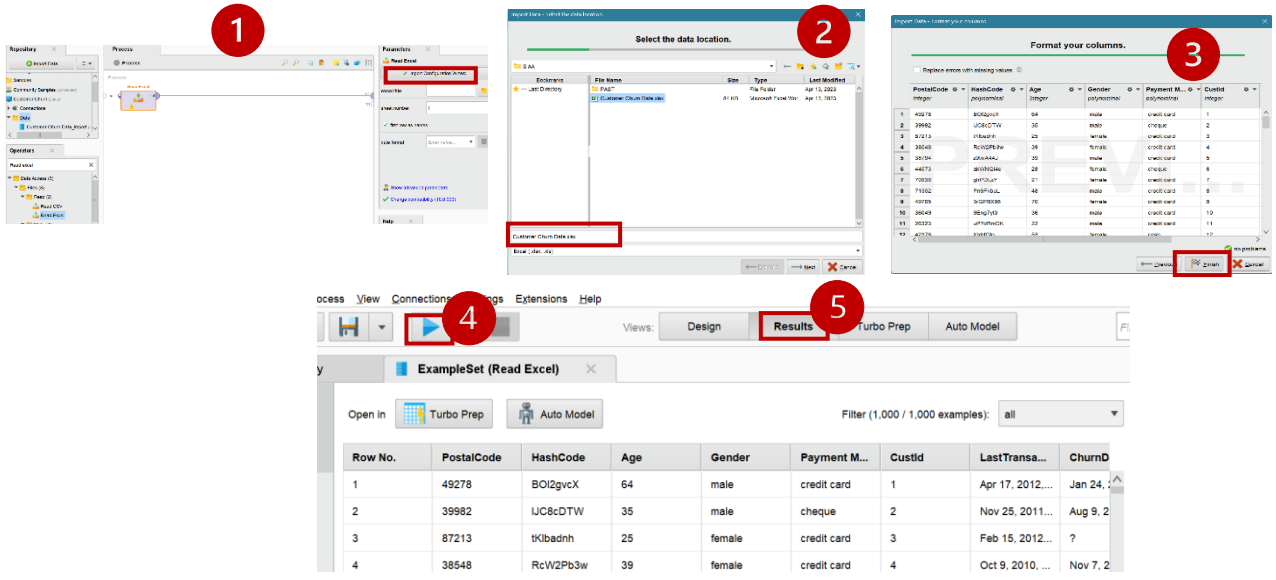
By right clicking the Customer Churn repository created, three subfolders are created: 'Data' to store all input/output data, 'Processes' to store all processes created, and 'Results' to store all analysis results, including models, weights, and wordlist.



EXERCISE: LOAD DATA

Following the data's upload to the 'Data' subfolder, the following procedures are used to load the data:

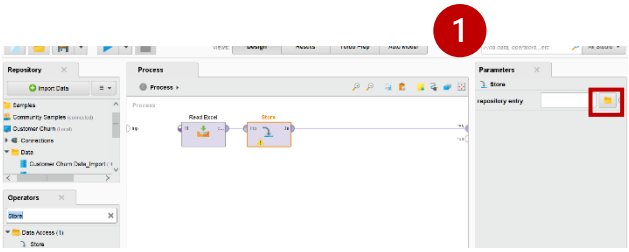
1) Type "read excel" into the Operators Panel.
2) Drag the Read Excel operator to the Process panel.
3) Join the res port and the out port.
4) Select the Import Configuration Wizard link in the Parameters panel.
5) Choose the data location.
6) Click 'Finish' after clicking 'Next' twice.
7) To view the results, click the Run button.

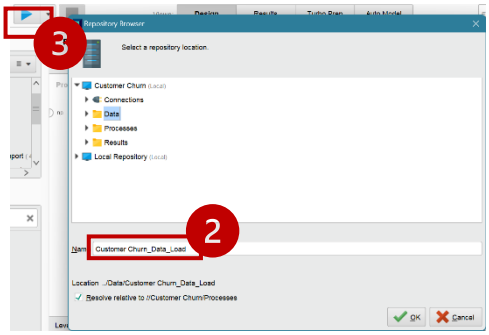


EXERCISE: STORE DATA

To store the data, the following procedures are followed:

1) Search store in the Operators panel.
2) Move the Store operator over to the Process panel.
3) Connect the Store operator next to Read Excel operator.
4) Click the button in the Parameters panel and select the location.
5) Add the name Customer Churn_Data_Load and press OK.
6) Run the process.

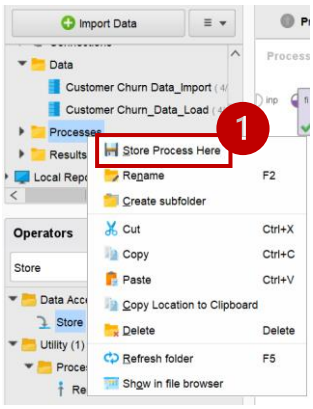


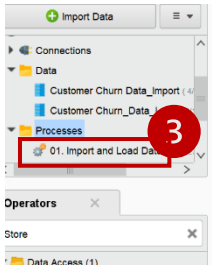


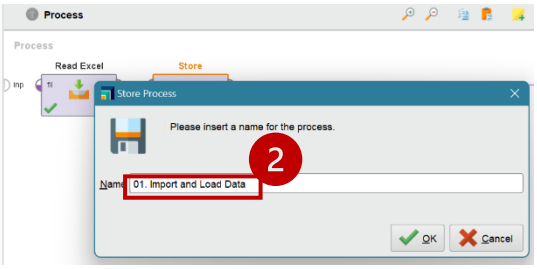
EXERCISE: STORE PROCESS

To store the process, following steps are followed:

1) Right-click the Process subfolder in the Repositories Panel and select Store Process Here.
2) At Store Process Windows, insert Name as Data Import and Loading. Input OK.







EXERCISE: DATA EXPLORATION

Potential Error	Example of Error	Recommended Solution
Some nominal classes with few observations (outliers),	372 persons aged 40-60, 4 person weighted 60-80.	Since it is an outlier in the data, that particular nominal class can be **eliminated**.
A nominal variable with more classes than it should have.	A third class of gender labelled as 2 other than 0 (male) or 1 (female).	Correct the third-class gender by **changing** "lady" to "female," "gent," and "männlich" to "male."
Data values that lie outside the expected or allowable range.	Age of -2 and 152 are found.	Outliers in the data should be **removed**.
Variables contain a high proportion of missing values.	60% of middle name of customers are missing.	To conduct an effective analysis, the entire column should be **removed** because more than 50% of the values are missing.

EXERCISE: FILTER EXAMPLES

A filter example operator, filters data from a dataset by verifying the user-specified criteria.

The procedures below are used to filter the 'Age' variable in accordance with the company's policy, which states that clients must be between the ages of [17,99].

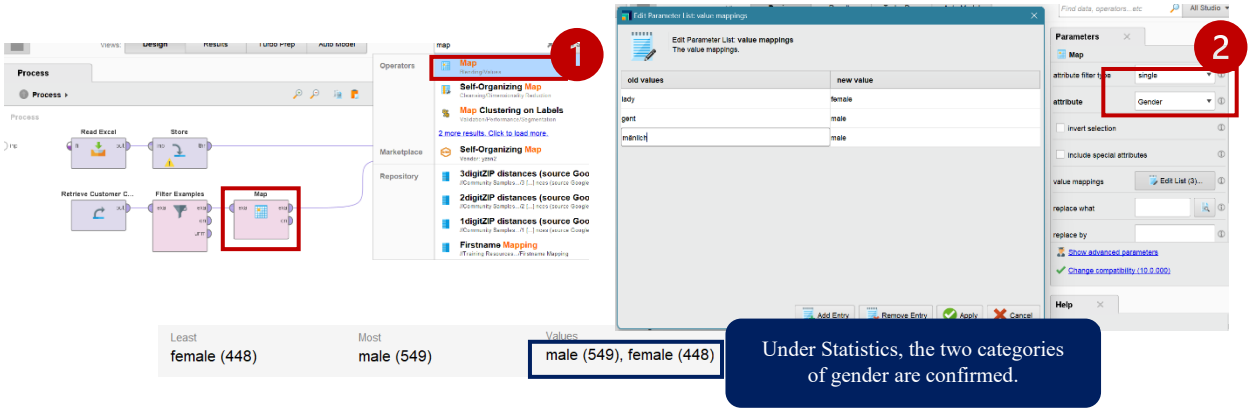
1) Drag Search Filter Example to the Process panel. Following that, link it to the Customer Churn Data_Load.
2) In the parameters panel, click add filter, and then enter the 'Age' condition that is requested. To join the filters, click "match any" and then "OK."
3) To eliminate the values, select "invert filter."
4) Run the process, to view results.

EXERCISE: MAP

The map operator is used to replace a given data value with a new value.

To replace wrong gender values, map is used by following the steps below:

1) Search Map and drag to Process panel. Connect it to the Filter Examples after that.
2) In the Parameters panel, select "single" and "Gender," then enter the previous values and the new ones to replace them.
3) To see results, run the process after clicking "apply."



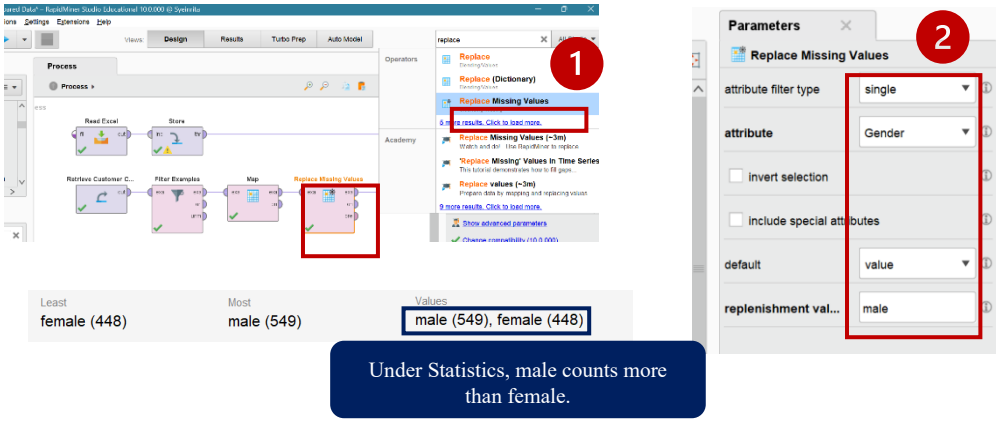
EXERCISE: REPLACE MISSING VALUES

The replace missing value operator is used to replace any values in data that are empty or null.

- Gender

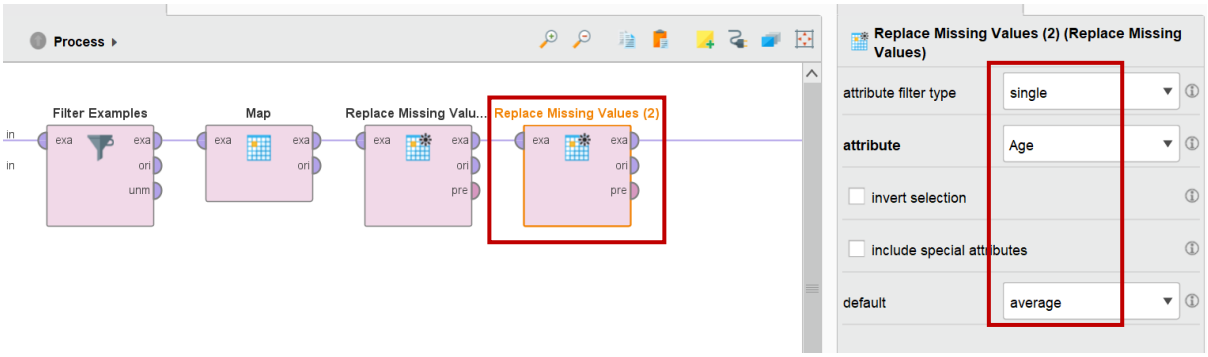
To replace missing value in gender with maximum count value, Replace Missing Values is used by foll0wing the steps below:

1) Drag to the Process panel after searching for Replace Missing Values. Connect it to the Map after that.
2) Select "single" and "Gender" in the Parameters panel, then select "value" and choose "male" to fill in the missing values.
3) Run the process to view the results.

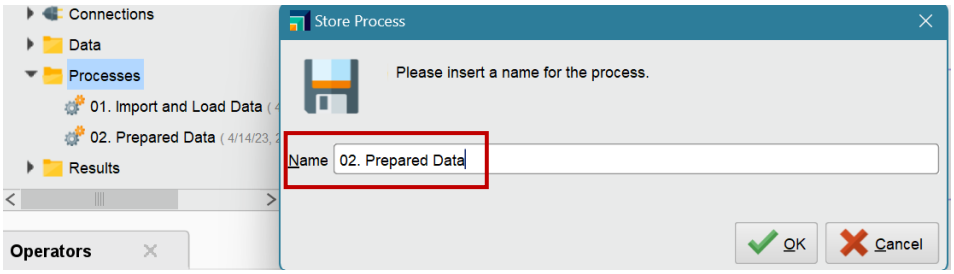


- Age

The steps are repeated to replace the missing value of Age with average of it.



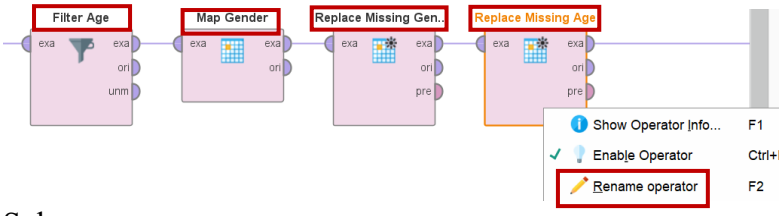
To store the process, the steps mentioned in Store Process section is repeated, and the process stored in the name of '02. Prepared Data' under Processes subfolder.



EXERCISE: UTILITIES FOR PROCESS PANEL

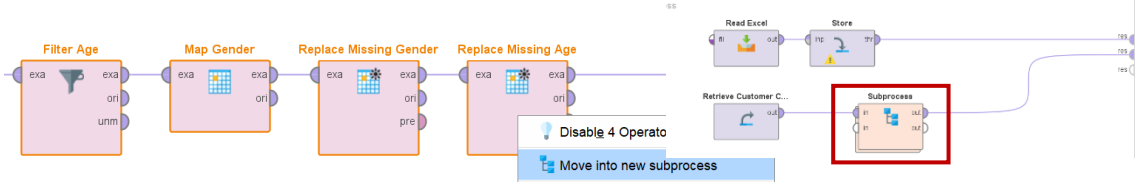
- Rename the operators.

The operators have been renamed accordingly by right clicking them, for better presentation.



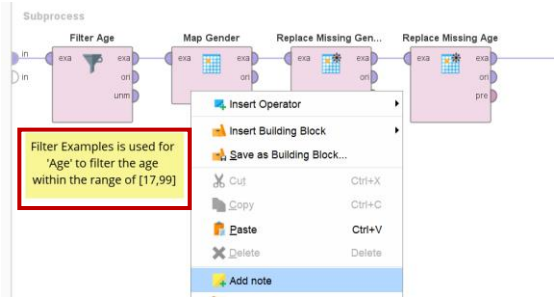
- Subprocess

Subprocess is to organise a group of operators which are working together into a folder. All the operators are selected and by right clicking, they are moved to new subprocess for better organisation.



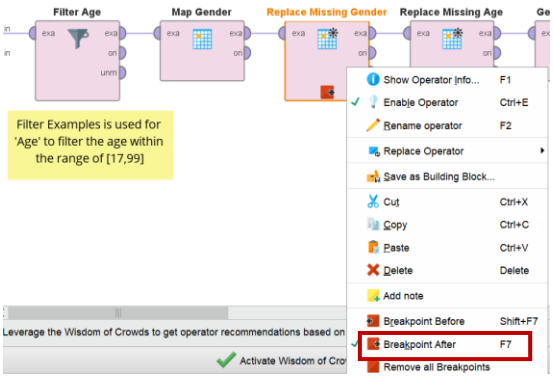
- Add notes

  Notes are used to record critical information or to provide detailed descriptions of actions taken while using an operator. Notes can be added by right clicking the operators.



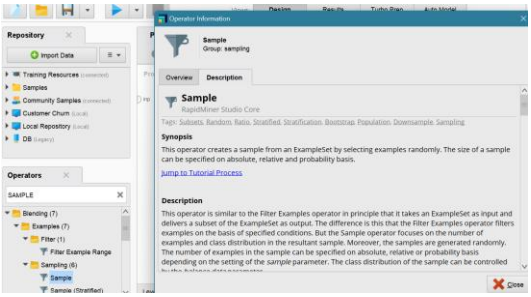
- Breakpoints

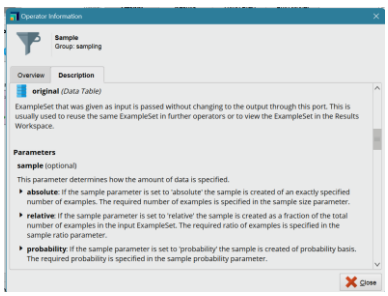
  Breakpoints enable process to be stopped so that outcomes from an operator's input or output can be examined. A breakpoint is inserted at Replace Missing Gender by right clicking, and when running the process, the operators till Replace Missing Gender will be executed and the results will be displayed. To continue can press play and the remaining operators will be executed or can press stop button to stop and restart.



- Sample process view

  From datasets, the sample operator will select a random sample. An absolute, relative, or probabilistic approach can be used to calculate the sample size. Here is a list of the operators' description and parameters.



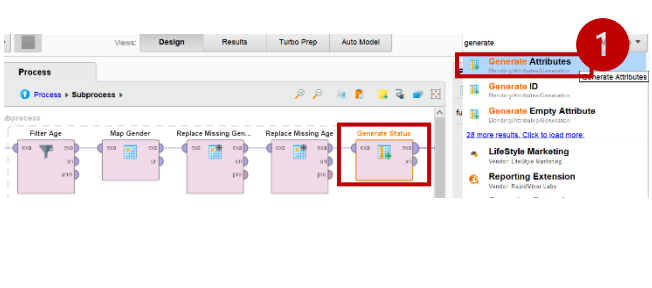


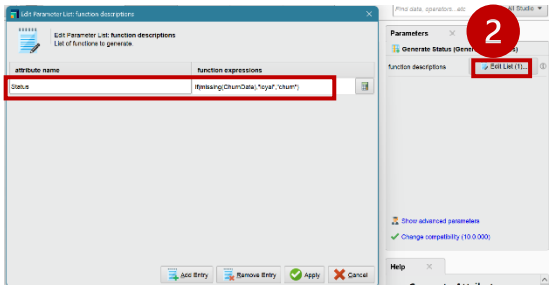
EXERCISE: GENERATE ATTRIBUTES

The Generate Attribute operator constructs new attribute from the attribute of the input data and arbitrary constants using mathematical expressions.

The steps listed below are used to apply Generate Attributes to a new attribute called "Status":

1) Search Generate Attributes and drag it to the Process panel. Connect it to Replace Missing Age after that.
2) In the Parameters panel, click Edit List. Enter the attribute name as "Status," and then create the function that displays "Loyal" if ChurnDate is missing and "Churn" otherwise. Rename it to Generate Status.
3) To see results, run the process after clicking "apply."



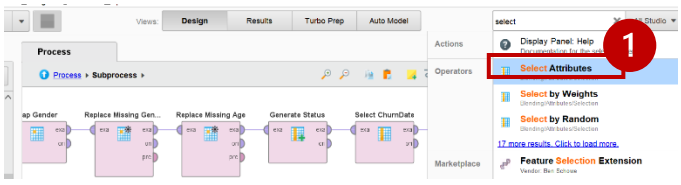


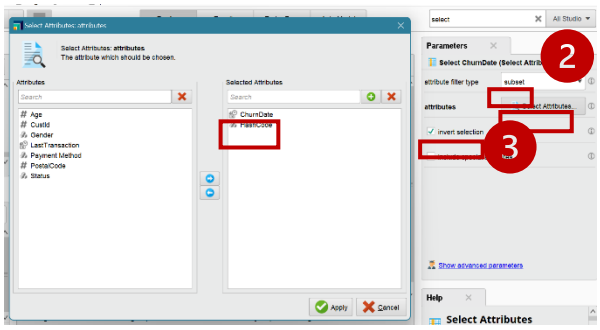
EXERCISE: SELECT ATTRIBUTES

Similar to the filter example, but much easier and simpler to apply, select attribute is used to extract a selection of attributes from a dataset while deleting all other attributes.

To remove ChurnDate and HashCode attributes, Select Attributes is used by following the steps below:

1) Search for and drag Select Attributes to the Process panel. Connect it to Generate Status after that.
2) In the Parameters tab, pick "subset" and then click "select attributes."
3) Click apply after selecting ChurnDate and HashCode. Rename it to Select ChurnDate.
4) Select "inverse selection" to exclude them, then click "run" to see the outcomes.





EXERCISE: DATA PREPARATION

To create Area attribute, Numerical to Polynominal, Generate Attributes and Select Attributes are used by following the steps below:

Numerical to Polynominal

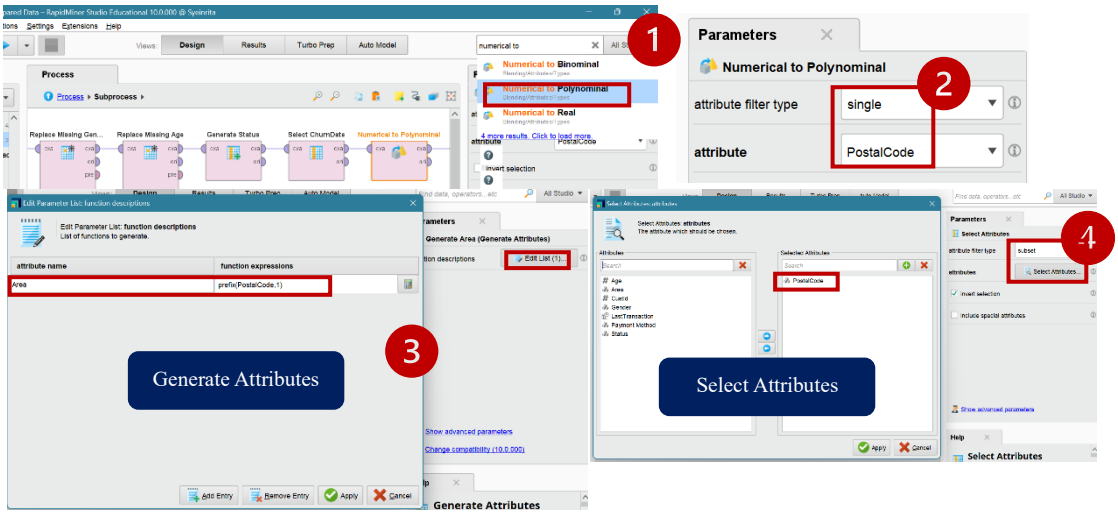
1) Drag the Process panel after searching Numerical to Polynominal. Link it to the Select ChurnDate after that.
2) To change from numerical to nominal, pick "single" and click select PostalCode attribute in the Parameters panel.

Generate Attributes

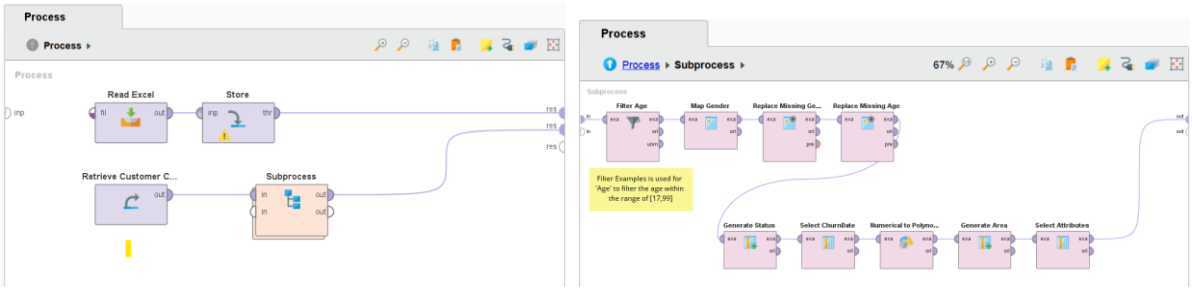
3) Search for Generate Attributes and drag it into the Process panel. Then join it to the Polynomial to Numerical.
4) Click Edit List in the Parameters panel. Type "Area" as the attribute name, and use prefix(PostalCode,1) to only utilise the first character of the PostalCode. Rename it to Generate Area.

Select Attributes

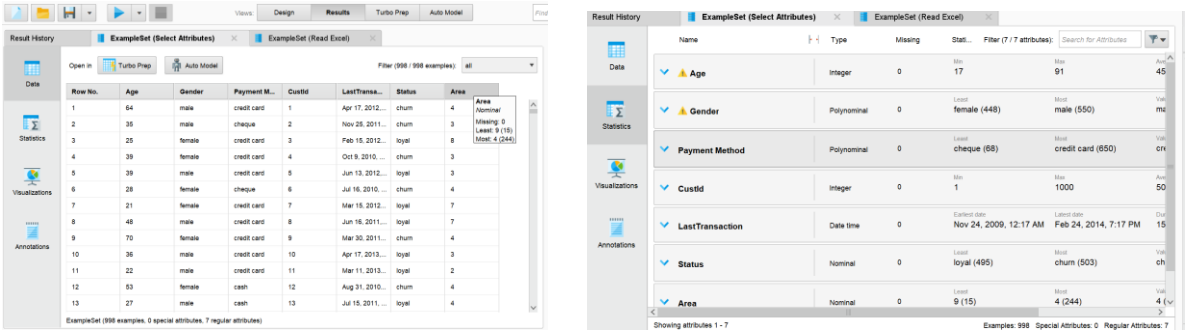
5) Search for Select Attributes and drag it to the Process panel. After that, connect it to the Generate Area.
6) Select "subset" from the Parameters tab, and then click "select attributes."
7) After choosing PostalCode, click apply.
8) Select "inverse selection" to exclude it, then click "run" to see the outcomes



FINAL RESULTS



Overview of Process Panel



Disaply of Final Data and its Statistics

CONCLUSION

To summarise the whole lab report, firstly the data named 'Customer Churn' is uploaded, loaded and stored in Rapidminer. The Age attribute is then filtered using Filter Examples to meet the company's criteria, and a map is used to correct incorrect Gender inputs by replacing them with the appropriate "male" or "female" label. After that, using Replace Missing Values to fill in the missing values for Gender and Age, the most prevalent "male" is chosen for Gender and the median age is used for Age.

To rename operators, organise, and other tasks, process panel utilities are utilised. In addition, General Attributes are used to construct the "Status" attribute and Select Attributes are used to remove the "ChurnDate" attribute because of its high missing value percentage. Then, using Numerical to Polynominal and General Attributes, the PostalCode attribute is changed from arbitrary to nominal. The 'Area' property is generated using the first number of the PostalCode, and the PostalCode attributes are removed using Select Attributes.

When we looked at the final result, we could see that the data had been prepared and that there were no longer any values that were missing, inaccurate data structure or outliers, which will be beneficial for future analyses.