**Lab Assignment 1**

**Were Vincent Ouma**

**Lab Assignment 1**

**a)Data Quality Check:**

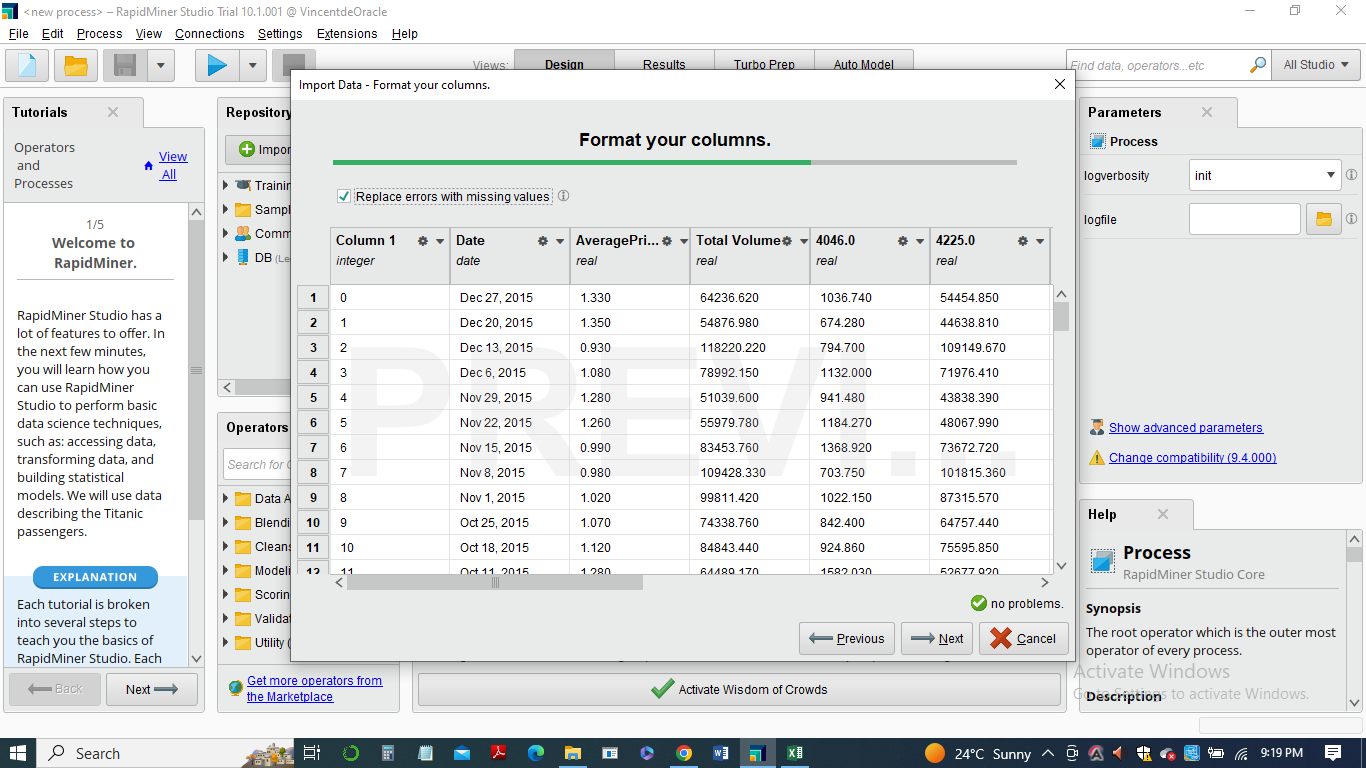
**a.1)** In RapidMiner Studio, the data quality is checked by using the “Data Quality” operator. This operator detect any bias in the dataset by calculating descriptive statistics of the dataset and visualizing the distribution of the data. It can also detect outliers and missing values. The “Data Quality” operator also provides information about the number of rows and columns in the dataset.

**a.2)** To handle the data quality issues such as duplications and missing values, the “Data Cleansing” operator is used. This operator allows the user to select the columns to be cleaned and specify the rules for cleaning the data. For example, the user can specify the rule to remove all duplicate values or to replace missing values with the median of the column feature in that specific “Region” variable.

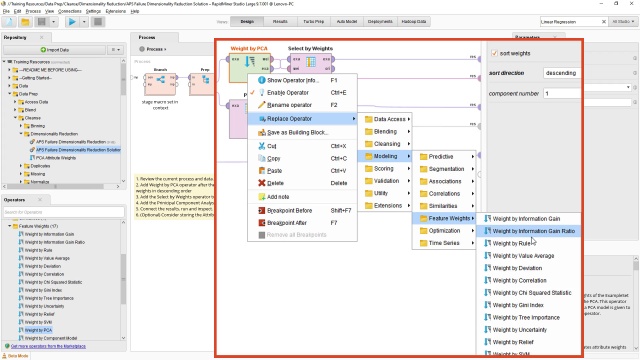
**a.3)** The number of rows and columns in the dataset is 18249 and 13, respectively.

**b)Data Cleaning and Preprocessing:**

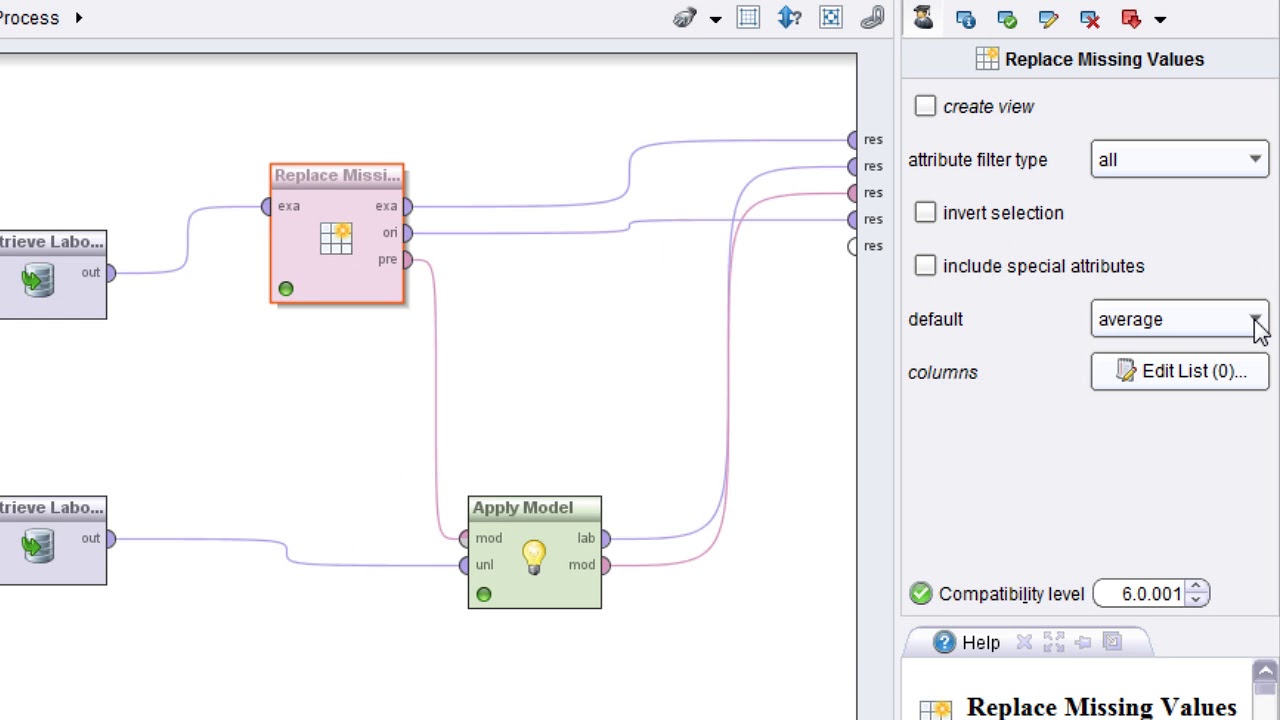
**b.1)**In RapidMiner Studio, the first column in the dataset is removed by using the “Select Attributes” operator. This operator allows the user to select or deselect the attributes that should be included in the dataset. The ‘year’ variable can be treated as nominal by using the “Nominal to Numeric” operator. This operator allows the user to convert the nominal values to numeric values.

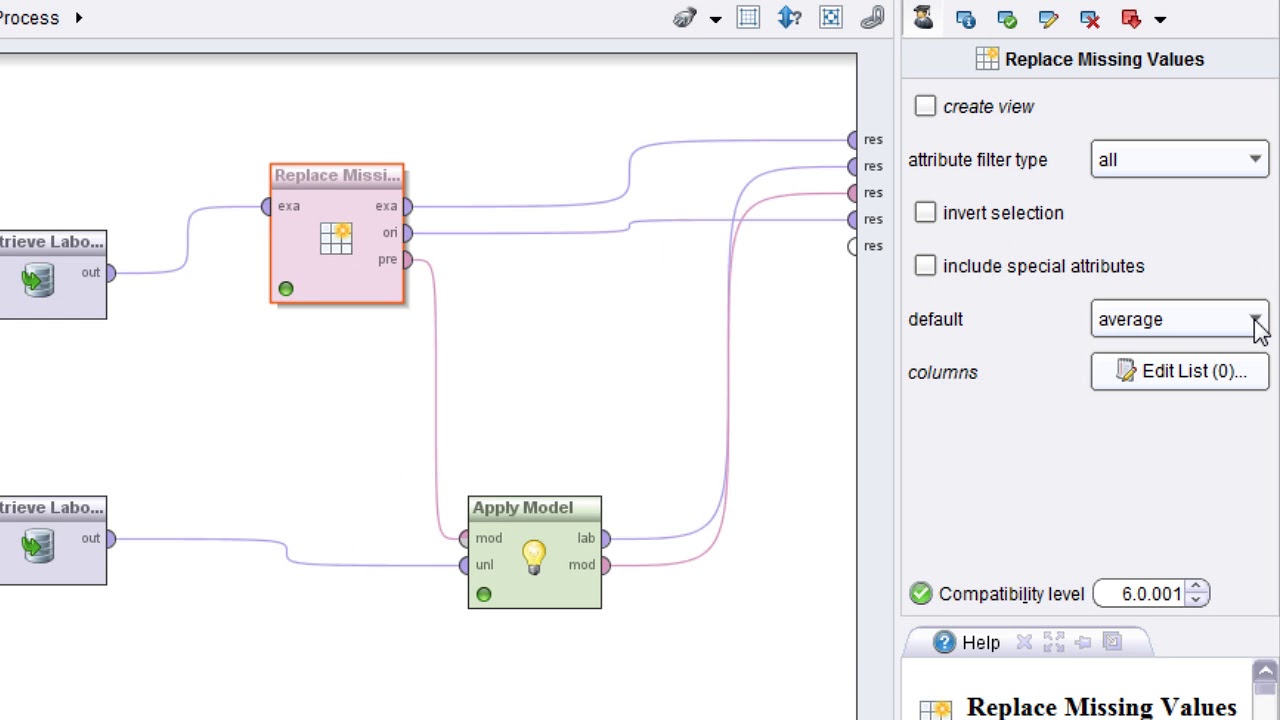


**b.2)**To check for duplicate values and remove them, the “Data Cleansing” operator is used. This operator allows the user to select the columns to be cleaned and specify the rules for cleaning the data. For example, the user can specify the rule to remove all duplicate values.



**b.3)**To check for missing values, the “Data Cleansing” operator is also used. This operator allows the user to select the columns to be cleaned and specify the rules for cleaning the data. For example, the user can specify the rule to replace missing values with the median of the column feature in that specific “Region” variable. If most column values in a data record are missing, the data record can be removed. The correlation between the variables is found by using the “Correlation Matrix” operator. This operator calculates the correlation coefficient between all the variables in the dataset and visualizes the correlation matrix. The result of the correlation matrix is shown in the screenshot below.





**b.4)**The correlation between the variables can affect the model accuracy. If two variables are highly correlated, they might provide redundant information, which can lead to overfitting of the model. On the other hand, if two variables are not correlated, they can provide complementary information which can help improve the accuracy of the model.

