ITB431

Data Wrangling

|  |  |
| --- | --- |
| Submitted By: | **Chee Bo Yu (214322H)** |
| PEM Group: | BA2102 |
| Tutor: | Mr Heng Rui Jie |
| Deadline: | 13 June 2022 |

Assignment

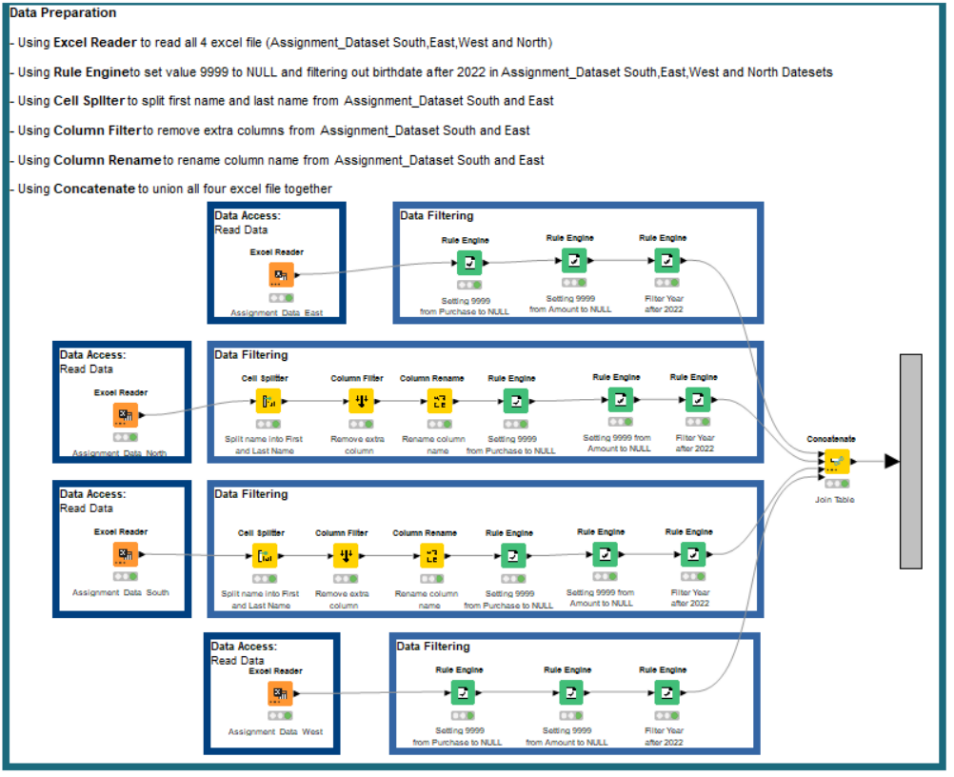
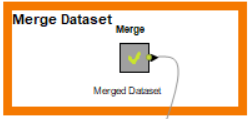
Report Title: Data Wrangling Report

**Table of Contents**

[Task 1: Steps taken to mashup, clean and transform the data 2](#Task1)-10

[Task 2: Challenges](#Task2) 10

**Task 1: Steps taken to mashup, clean and transform the data**



**Metanode**

Using **Metanode** in my assignment is a useful component to make my workflows neater for people to view.

Shape, arrow

Description automatically generated

By using **New Workflow Annotation,** I can describe what is the node about and the purpose of using the node for my workflows.

**New Workflow Annotation** Graphical user interface

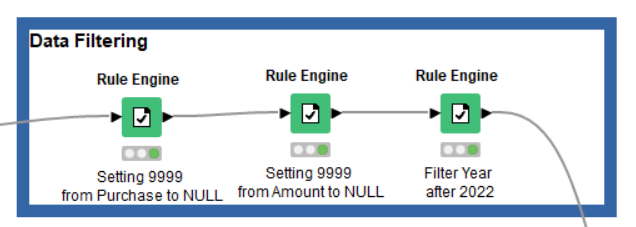
Description automatically generated

Table

Description automatically generated with medium confidence

**Excel Reader Transformation Tab**

By using **Excel Reader**, I can load the excel file to my workflow and use **Transformation Tab** to check the column would like to view.

**Graphical user interface, text, application

Description automatically generated Rule Engine**

Table

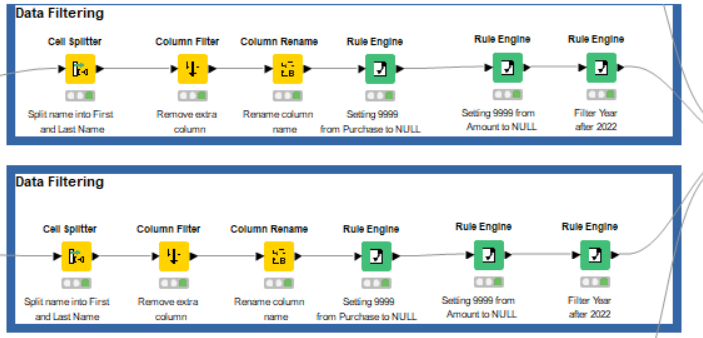
Description automatically generated Table

Description automatically generatedText

Description automatically generated

**Result Expression Command**

By using **Rule Engine**, I can set the column values with 9999 to NULL by using the command above and using the same method I can set the amount with 9999 to NULL, BirthDate after 2022 to NULL.



Graphical user interface, text, application, email

Description automatically generated

By using **Cell Splitter**, I can split First and Last Name into two columns from the original column name called Name.

**Cell Splitter**

Table

Description automatically generated with low confidence

**Before Splitting Columns**

Table

Description automatically generated with medium confidence

**After Splitting Columns**

Graphical user interface, application

Description automatically generated

By using **Column Filter**, I can remove the extra column called Name\_Arr[0] in the table.

**Column Filter**

A picture containing table

Description automatically generated

**Results**

Graphical user interface, text, application

Description automatically generated

By using **Column Rename**, I can change the column name from Name\_Arr[1] and [2] to First and Last Name respectively.

A picture containing table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

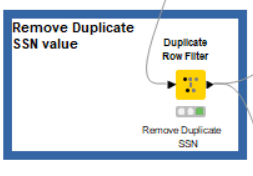
Table

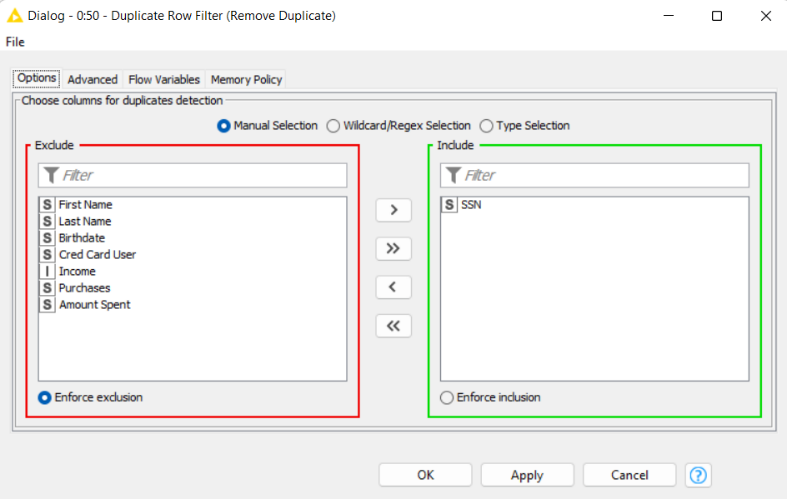
Description automatically generated with medium confidence

**Results**

**Concatenate**

By using **Concatenate**, I can put all 4 excel files (Assignment\_Dataset South, East, West and North) together to union.

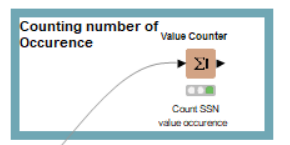


 **Duplicate Row Filter**

By using **Duplicate Row Filter**, I can remove the SSN Duplicate to create anonymize stage

Table

Description automatically generated

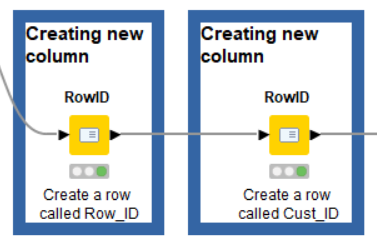


**Value Counter Results**

Graphical user interface, text, application

Description automatically generated

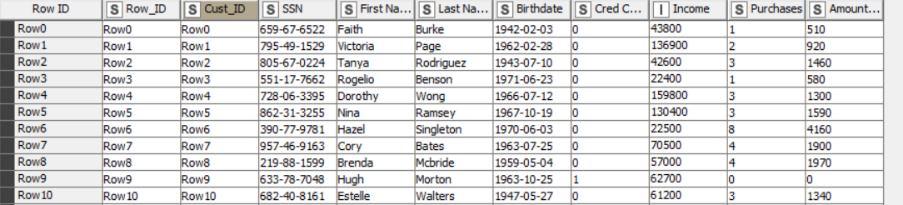
By using **Value Counter**, I can check whether there is any duplicate for the Row ID by looking at the count column.



**RowID**

Graphical user interface, text, application

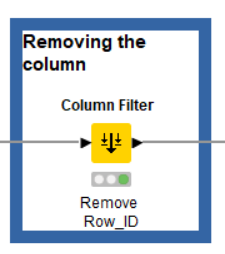
Description automatically generated



**Results**

By using **Row ID**, I can have a unique column on my dataset for anonymization, by creating two Row ID, this will remove \_dup in the Cust\_ID column after exporting to excel

Graphical user interface, application

Description automatically generated

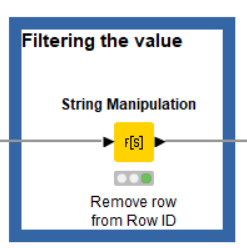
A picture containing table

Description automatically generated

**Results**

By using **Column Filter**, I can remove the column with the same name in my dataset.

A picture containing graphical user interface

Description automatically generated****

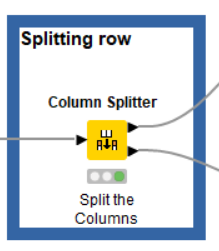
**Results**



By using **String Manipulation**, I can filter out the row from Cust\_ID which will only have the number of the ID.

Graphical user interface

Description automatically generated

 **Column Splitter**

Table

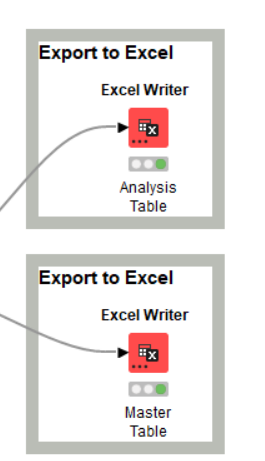
Description automatically generated with medium confidence

Table

Description automatically generated

**Results**

By using **Column Splitter**, I can drop the highest priority, which is the SSN, First and Last Name to complete anonymization.



By using **Excel Writer,** I can have two excel

file called Analysis Table and Master Table which

containing the lowest priority and highest priority.

**Task 2: Challenges**

The challenges that I face will be **Uncertain about each node in the Node Repository** and **whether my method of doing anonymising is correct or not.**

For my first challenge faced, with my determination and curiosity to find out what each node does by watching YouTube video, I can learn what each node works like Rule Engine, Column Filter, Column Rename, Cell Splitter and New Workflow Annotation.

For my second challenge faced, by checking with my lectures about my method of using anonymization, I can learn more about using Duplicate Row Filter, RowID, Value Counter, String Manipulation and Column Splitter for Anonymization rather than just using Anonymization Node to anonymize.