



**Swinburne University of Technology Hawthorn Campus**  
**Dept. of Computing Technologies**

**COS10022 Data Science Principles**  
**Assignment 2 - Semester 1, 2025**

**Assessment Title:** Data Cleaning, Integration, and Analysis

**Assessment Weighting:** 20%

**Due Date:** Sunday, 25<sup>th</sup> May 2025 at 11.59 pm (AEDT)

**Assessable Item:**

- One (1) piece of a written report no more than 10-page long with the signed Assignment Cover Sheet.
- One (1) zip file containing your KNIME workflow, the input file, and the output file or any intermediate files produced in your workflow execution process.
- The submitted report must pass the Turnitin check on Canvas with no more than 30% similarity except the parts from the template or the short answers.

The submitted report should answer all questions listed in the assignment task section in sequence.

You must include a digitally signed Assignment Cover Sheet with your submission.

Submitting the zip file containing the input/output files and your fully functional KNIME workflow is essential for your submission to be marked. If the submitted zip file cannot execute properly or the execution result differs from what you have in the report, you will not get the mark even if you put in the correct answer.

---

There are 100 marks on this assignment. Your answers must address the following tasks.

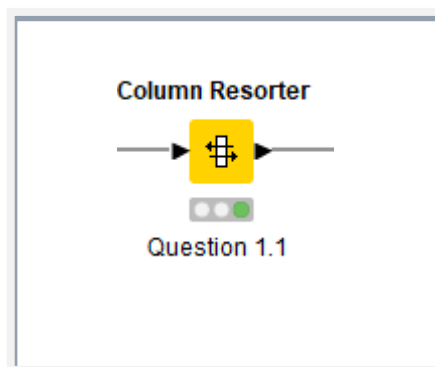
1. Answer the questions based on your findings:

**[26 marks]**

1.1. List the node you used to arrange the column order:

**(2 marks)**

**Ans:** The node that I used to arrange the column order is the Column Resorter node.



**Figure 1: Column Resorter node**

- 1.2. Take a screenshot of the output after arranging the columns in the specified order. The screenshot should contain the column names and the first five tuples. **(2 marks)**

**Ans:** **Figure 2** demonstrates arranged columns after using Column Resorter and Color Manager nodes.

Filtered table - 3:2 - Column Filter (Arrange the)

File Edit Hilite Navigation View

Table "default" - Rows: 1000 Spec - Columns: 7 Properties Flow Variables

Row ID	Resident ID	Resident	DoB	Current...	Education	Location	Income
Row0	2313	Ralen	1993/3/22	32_	college	Joondalup	271000
Row1	1622	Janan	2006-2-11	19	junior college	Karratha	284000
Row2	490	Dahlia	1978-10-4	-46	college	Carnarvon	112000
Row3	1395	Elus	1979-10-17	45	under HS	Warwick	338000
Row4	920	Galus	1978-11-3	46	college	Murray Bridge	151000

**Figure 2: Arranged columns****(2 marks)**

- 1.3. Observe the "Resident" data, list and describe the abnormal patterns you observed in this column.

**Ans:** From my perspective, I see that there are two kind of errors in the "Resident: column.

+ Leading Apostrophes: Comma above before the name started such as 'Corion, 'Damian, and 'Kelis.

+ Trailing Spaces (Special Characters): A slash and letter 's' at the end of each name 'Fenon\s', 'Ozias\s', and 'Felix\s'.

Consequently, I use the String Manipulation node with the below funtion to assign value "True" for error datas in this case. In contrast, the normal data will be False.

```
regexMatcher($Resident$, "(^[A-Za-z].*)/(.*\\s$)")
```

- 1.4. What percentage of data in the "Resident" column is noisy? Post the screenshot of where you find the answer, too. **(2 marks)**

**Ans:** For calculating the percentage of noisy data in the "Resident" column, in logical thinking and theory, we need to know the number of "True" values by using Value Counter node. As a result, it returns 20 values so the percentage will be calculated by the formula:  $20/1000 * 100\% = 2\%$ .

Group table - 3:58 - GroupBy

File Edit Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 2 Properties Flow V

Row ID	Noisy_data	Percent(Resident)
Row0	False	98
Row1	True	2

**Figure 3: Noisy Data Percentage**

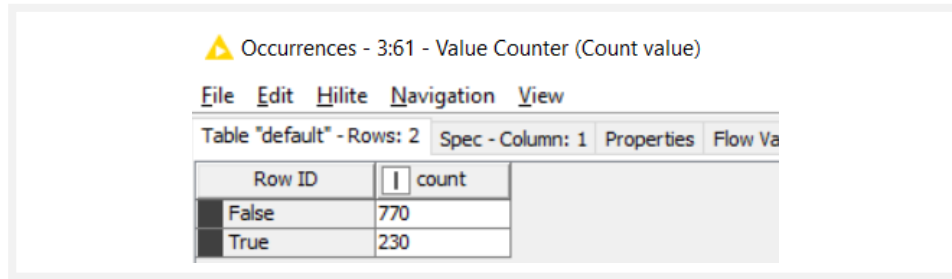
- 1.5. Observe the "DoB" data, list and describe the abnormal patterns you observed in this column. **(2 marks)**

**Ans:** In my opinion, there are two distinct data formats in the "DoB" column

which are yyyy/MM/dd and yyyy-MM-dd.

- 1.6. How many instances in the “DoB” column are noisy? Post the screenshot of where you find the answer, too. **(2 marks)**

**Ans:** To answer this question, I decide to assign value “True” with data format “yyyy/MM/dd” and False for “yyyy-MM-dd” by utilizing String Manipulation node. Then, I implement Row Filter nodes and figure out that there are 770 False values.



Occurrences - 3:61 - Value Counter (Count value)

File Edit Hilite Navigation View

Table "default" - Rows: 2 Spec - Column: 1 Properties Flow Va

Row ID	count
False	770
True	230

**Figure 4: Value Counter node**

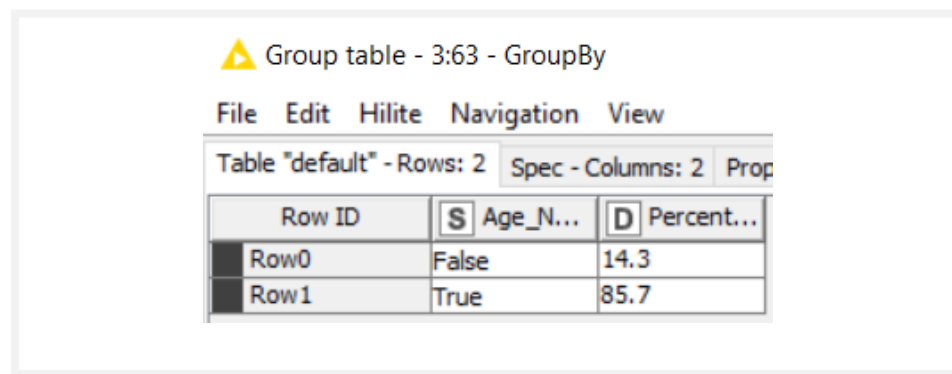
**(2 marks)**

- 1.7. Observe the “Age” data, list and describe the abnormal patterns you observed in this column.

**Ans:** From my perspective, I see various error data in the “Age” data including negative numbers (age cannot be negative), an underline after ages, a dot after ages, decimal ages, and negative ages have underline at the end. In summary, the integer numbers without special characters or symbols will be normal.

- 1.8. What percentage of data in the “Age” column is noisy? Post the screenshot of where you find the answer. **(2 marks)**

**Ans:** To answer this question, I assign “True” value for integer numbers without special characters or symbols by using String Manipulation node. Next, I use GroupBy node to calculate the percentage of noisy data in the “Age” column. According to the **Figure 5**, it demonstrates that there are 14.3% of noisy data.



Group table - 3:63 - GroupBy

File Edit Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 2 Prop

Row ID	Age_N...	Percent...
Row0	False	14.3
Row1	True	85.7

**Figure 5: Percentage of datas in the “Age” column**

- 1.9. How many different terms are included in the “Education” column? (For example, “PhD” and “master” are two different terms.) Post the screenshot of where you find the answer, too. **(2 marks)**

**Ans:** There are nine distinct terms included in the “Education” column including HS, PhD, college, high-school, junior college, master, postgrade by course, under HS, and under high-school.

Group table - 3:72 - GroupBy (Cout different)

File Edit Hilite Navigation View

Table "default" - Rows: 9 Spec - Column: 1 Properties

Row ID	S Education
Row0	HS
Row1	PhD
Row2	college
Row3	high-school
Row4	junior college
Row5	master
Row6	postgrade by course
Row7	under HS
Row8	under high-school

Figure 6: Educational Levels

- 1.10. How many instances in the "Education" column contain the term "junior college"? Post the screenshot of where you find the answer, too. (2 marks)

Ans: By using Value Counter node, there are about 250 instances in the "Education" column contain the term "junior college".

Occurrences - 4:28 - Value Counter

File Edit Hilite Navigation View

Table "default" - Rows: 9 Spec - Column: 1 Properties Flow Variables

Row ID	count
HS	328
PhD	20
college	250
high-school	22
junior college	250
master	15
postgrade by course	65
under HS	49
under high-school	1

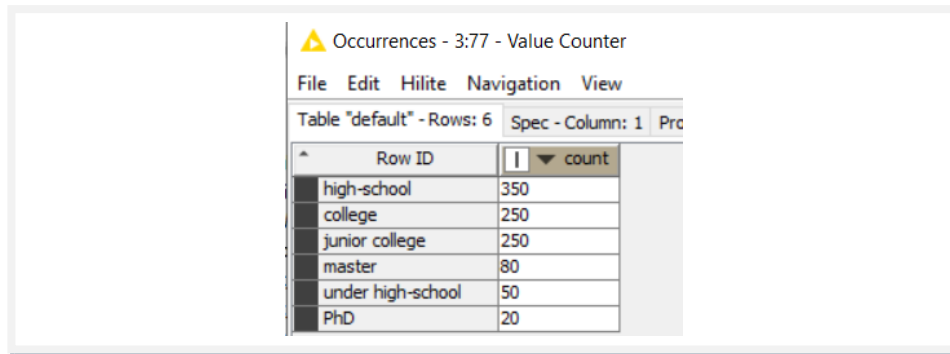
Figure 7: Number of people in each degree level

- 1.11. The data collectors used different terms to record the education levels. In our project, we only accept "under high-school", "high-school", "junior college", "college", "master", and "PhD" to be used in the data source. ***It looks like part of the "high-school" items are recorded as "HS", and some "master" terms are recorded as "postgraduate by course".*** How many instances are modified after converting these terms into the desired format?

Ans: There are about 393 instances are modified after converting into the desired format.

- 1.12. How many instances include residents with the education level of "master" after cleaning the data? Post the screenshot of where you find the answer, too. (2 marks)

**Ans:** After using Value Counter node, there are about 80 residents with the education level of “master”.



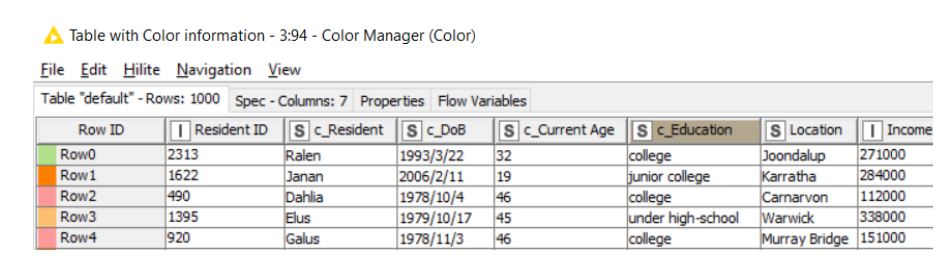
Row ID	count
high-school	350
college	250
junior college	250
master	80
under high-school	50
PhD	20

**Figure 8: Numbers of residents in distinct educational levels**

**(2 marks)**

- 1.13. Remove all columns containing data before cleaning and temporary columns used to generate data for comparison. Realign the remaining column in the sequence of “Resident ID,” “c\_Resident,” “c\_DoB,” “c\_Current Age,” “c\_Education,” “Location,” and “Income”, where “c\_” stands for the cleaned data. Take a screenshot of your realigned table with the first five instances with it.

**Ans:** For remove unnessecary column, I tend to use Column Filter node. Then, I use Column Resorter node again to realign the “Resident ID,” “c\_Resident,” “c\_DoB,” “c\_Current Age,” “c\_Education,” “Location,” and “Income” columns which adapt the requirement.



Row ID	Resident ID	c_Resident	c_DoB	c_Current Age	c_Education	Location	Income
Row0	2313	Ralen	1993/3/22	32	college	Joondalup	271000
Row1	1622	Janan	2006/2/11	19	junior college	Karratha	284000
Row2	490	Dahlia	1978/10/4	46	college	Carnarvon	112000
Row3	1395	Elus	1979/10/17	45	under high-school	Warwick	338000
Row4	920	Galus	1978/11/3	46	college	Murray Bridge	151000

**Figure 9: Arrange columns**

2. Answer the questions based on your findings:

**[14 marks]**

- 2.1 After realigning the attributes, observe the content in the “Birthday.” List the abnormal patterns you observed in this column.

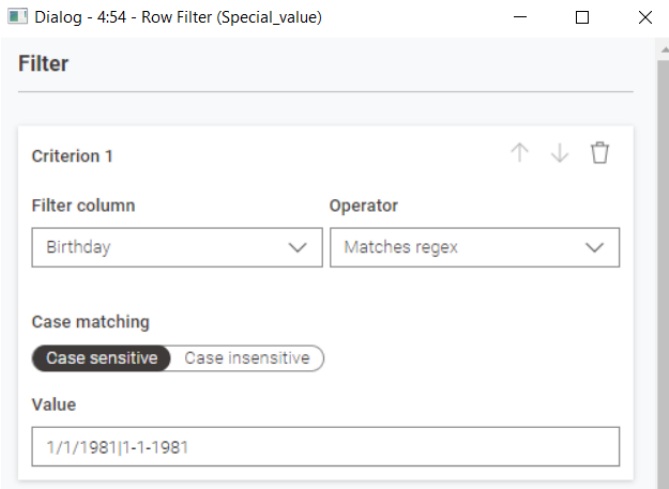
**(2 marks)**

**Ans:** From my observing, there are two formats of data in the “Birthday” column including dd-MM-yyyy and dd/MM/yyyy.

- 2.2 How many people in this dataset were born on the **first day of January in 1981**? Post the screenshot of where you find the answer, too.

**(2 marks)**

**Ans:** There are two people in the dataset who were born on the first day of January in 1981 by using Row Filter node.



**Filter**

Criterion 1

Filter column: Birthday Operator: Matches regex

Case matching: Case sensitive (selected), Case insensitive

Value: 1/1/1981|1-1-1981

**Included Rows - 4:54 - Row Filter (Special\_value)**

File Edit Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 31 Properties Flow V

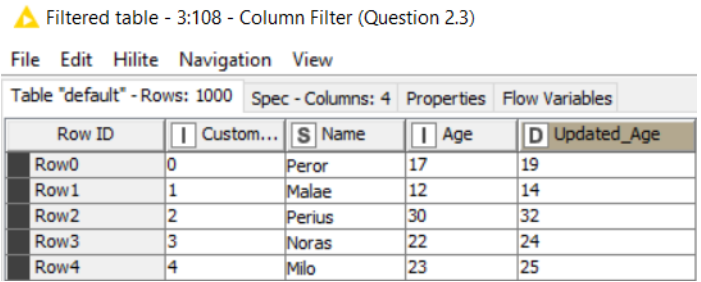
Row ID	Custom...	Name	Birthday
Row316	316	Parker	1-1-1981
Row559	559	Torvn	1-1-1981

(5 marks)

**Figure 10: Row Filter node**

- 2.3 We know that this dataset was collected two years before the other dataset used in this assignment. This means that the “Age” in this dataset is outdated. You should do data conditioning to make the “Age” values match the other dataset. Use a node in KNIME to complete the process and list the node name and screenshot the expression used in the node configuration. The processed data should be put in the column called “Updated\_Age.”

**Ans:** In this question, I decide to utilize the Math formula node to plus values in the column “Age” by two.



**Filtered table - 3:108 - Column Filter (Question 2.3)**

File Edit Hilite Navigation View

Table "default" - Rows: 1000 Spec - Columns: 4 Properties Flow Variables

Row ID	Custom...	Name	Age	Updated_Age
Row0	0	Peror	17	19
Row1	1	Malae	12	14
Row2	2	Perius	30	32
Row3	3	Noras	22	24
Row4	4	Milo	23	25

**Figure 11: Updated\_Age**

(5 marks)

- 2.4 Remove all temporary columns and the columns containing data before Semester 1, 2025

cleaning (conditioning). Realign the attribute in the sequence of "Customer ID," "Name," "Updated\_Birthday," "Updated\_Age," "Education Level," "City," "Purchase Date," "Shopping List," "Item List," and followed by "Item A" to "Item U." This will help you to observe the content in the dataset. You don't need to fill in anything for this question in the report. The tutors will examine your result in the submitted workflow.

3. Answer the questions based on your findings:

**[32 marks]**

- 3.1 Join the two datasets by matching the following five pairs to identify the tuples belonging to the same person: "name," "birthday," "age," "education," and "city." You should compare the values and types in the join columns. Include all tuples and all attributes from both datasets in the output. Creating new RowIDs will help you to identify tuples in the joined dataset. You should have exactly 1,000 tuples with 37 attributes in the joined dataset without missing values. If you don't get the result as mentioned. If your submitted workflow is not operable for checking the configurations and results, you will lose the mark. Explain how you discovered issues and what the corresponding solutions are that you adopted in your workflow.

**(30 marks)**

**Ans:** First and foremost, when I merge two datasets into one, the result returns that there are 2000 tuples and 37 attributes. At that time, I started to feel there's something wrong. Then, I emerged with each pairs respectively. Notably, I realized that two columns must be at the same data type so I utilize all powerful tools like AI Assistant and figured out what is the reason. Finally, I understood and solved the problem. To be honest, this question took me for two days but finally I did it.

- 3.2 Save the joined dataset and call it "2025\_A2\_integrated\_result.csv" List the name of the node you used to save the dataset. We also need to see the saved dataset being included in the zip file of your submission. You don't need to put in anything for this question in the report. Tutors will check your submitted workflow and your zip file for the combined dataset.

**(2 marks)**

**Ans:** In this question, I have to use Number to String and String to Number node to solve the complex issues. As a result, it works and export the table which adapt the requirement.

4. Answer the questions based on your findings:

**[12 marks]**

- 4.1 From the joined dataset, **drop the "c\_Current Age," "c\_Education," "Location," "Name," and "Item List" columns and realign the remaining columns in the sequence of "c\_Resident," "c\_DoB," "Updated\_Birthday," "Updated\_Age," "Education Level," "City," "Resident ID," "Customer ID," "Income," "Purchase Date," "Shopping List," and followed by "Item A" to "Item U."** Rename "c\_Resident" to "User Name," "Updated\_Birthday" to "Birthday," "Updated\_Age" to "Age," "Resident ID" to "RID," and "Customer ID" to "CID." You don't need to put anything in the report for this question. Tutors will check your workflow to find out whether your data fits the requirements.

**(5 marks)**

- 4.2 Take the proper part of the data from 4.1 and feed it into the Apriori algorithm to find corresponding association rules with the minimum support and minimum confidence to be 35% and 80%, respectively. List the count of item N and Item Q appear simultaneously. Post the screenshot of where you got the answer, too.

**(2 marks)**

**Ans:** Based on the **Figure 12**, the tem N and Item Q do not appear simultaneously in any itemset while setting the minimum support and minimum confidence to be 35% and 80% persepectively according to the requirement.

▲ Frequent itemsets/Association rules - 3/77 - Association Rule Learner (Question 4.2)

File Edit Hilite Navigation View

Table "default" - Rows: 32 Spec - Columns: 6 Properties Flow Variables

Row ID	D Support	D Confidence	D Lift	S Consequent	S implies	[...] Items
rule0	0.35	1	1.855	Item S	<---	[Item T,Item G]
rule1	0.35	1	1.855	Item T	<---	[Item S,Item G]
rule2	0.351	1	1.855	Item S	<---	[Item T,Item K]
rule3	0.351	1	1.855	Item T	<---	[Item S,Item K]
rule4	0.353	1	1.855	Item S	<---	[Item T,Item H]
rule5	0.353	1	1.855	Item T	<---	[Item S,Item H]
rule6	0.355	1	1.855	Item S	<---	[Item N,Item T]
rule7	0.355	1	1.855	Item T	<---	[Item P,Item S]
rule8	0.356	1	1.855	Item S	<---	[Item P,Item T]
rule9	0.356	1	1.855	Item T	<---	[Item P,Item S]
rule10	0.357	1	1.855	Item S	<---	[Item T,Item I]
rule11	0.357	1	1.855	Item T	<---	[Item S,Item I]
rule12	0.358	1	1.855	Item S	<---	[Item T,Item F]
rule13	0.358	1	1.855	Item T	<---	[Item S,Item F]
rule14	0.358	1	1.855	Item S	<---	[Item T,Item U]
rule15	0.358	1	1.855	Item T	<---	[Item S,Item U]
rule16	0.361	1	1.855	Item S	<---	[Item T,Item A]
rule17	0.361	1	1.855	Item T	<---	[Item S,Item A]
rule18	0.362	1	1.855	Item S	<---	[Item T,Item L]
rule19	0.362	1	1.855	Item T	<---	[Item S,Item L]
rule20	0.363	1	1.855	Item S	<---	[Item T,Item R]
rule21	0.363	1	1.855	Item T	<---	[Item S,Item R]
rule22	0.364	1	1.855	Item S	<---	[Item M,Item T]
rule23	0.364	1	1.855	Item T	<---	[Item M,Item S]
rule24	0.367	1	1.855	Item S	<---	[Item T,Item E]
rule25	0.367	1	1.855	Item T	<---	[Item S,Item E]
rule26	0.372	1	1.855	Item S	<---	[Item T,Item B]
rule27	0.372	1	1.855	Item T	<---	[Item S,Item B]
rule28	0.385	1	1.855	Item S	<---	[Item T,Item J]
rule29	0.385	1	1.855	Item T	<---	[Item S,Item J]
rule30	0.539	1	1.855	Item S	<---	[Item T]
rule31	0.539	1	1.855	Item T	<---	[Item S]

**Figure 12: Association Rule Learner node**

(5 marks)

- 4.3** Following the result obtained in 4.2, what is the probability of item S appearing when items J and T appear? Post the screenshot of where you find the answer, too.

**Ans:** Based on the provided result of the association rule node, the probability of Item S appearing when Items J and T appear is 1, or 100% which means that whenever Item J and Item T are present together, Item S is always there as well. Therefore, this information can be found by looking at Rule ID 29 in your table, where the "Items" column shows "[Item J, Item T]" and the "Consequence" column is "Item S," with a "Confidence" of 1.



▲ Frequent itemsets/Association rules - 3:77 - Association Rule Learner (Question 4.2)

File Edit Hilite Navigation View

Table "default" - Rows: 32 Spec - Columns: 6 Properties Flow Variables

Row ID	[D] Support	[D] Confidence	[D] Lift	[S] Consequent	[S] implies	[...] Items
rule0	0.35	1	1.855	Item S	<---	[Item T,Item G]
rule1	0.35	1	1.855	Item T	<---	[Item S,Item G]
rule2	0.351	1	1.855	Item S	<---	[Item T,Item K]
rule3	0.351	1	1.855	Item T	<---	[Item S,Item K]
rule4	0.353	1	1.855	Item S	<---	[Item T,Item H]
rule5	0.353	1	1.855	Item T	<---	[Item S,Item H]
rule6	0.355	1	1.855	Item S	<---	[Item N,Item T]
rule7	0.355	1	1.855	Item T	<---	[Item N,Item S]
rule8	0.356	1	1.855	Item S	<---	[Item P,Item T]
rule9	0.356	1	1.855	Item T	<---	[Item P,Item S]
rule10	0.357	1	1.855	Item S	<---	[Item T,Item I]
rule11	0.357	1	1.855	Item T	<---	[Item S,Item I]
rule12	0.358	1	1.855	Item S	<---	[Item T,Item F]
rule13	0.358	1	1.855	Item T	<---	[Item S,Item F]
rule14	0.358	1	1.855	Item S	<---	[Item T,Item U]
rule15	0.358	1	1.855	Item T	<---	[Item S,Item U]
rule16	0.361	1	1.855	Item S	<---	[Item T,Item A]
rule17	0.361	1	1.855	Item T	<---	[Item S,Item A]
rule18	0.362	1	1.855	Item S	<---	[Item T,Item L]
rule19	0.362	1	1.855	Item T	<---	[Item S,Item L]
rule20	0.363	1	1.855	Item S	<---	[Item T,Item R]
rule21	0.363	1	1.855	Item T	<---	[Item S,Item R]
rule22	0.364	1	1.855	Item S	<---	[Item M,Item T]
rule23	0.364	1	1.855	Item T	<---	[Item M,Item S]
rule24	0.367	1	1.855	Item S	<---	[Item T,Item E]
rule25	0.367	1	1.855	Item T	<---	[Item S,Item E]
rule26	0.372	1	1.855	Item S	<---	[Item T,Item B]
rule27	0.372	1	1.855	Item T	<---	[Item S,Item B]
rule28	0.385	1	1.855	Item S	<---	[Item T,Item J]
rule29	0.385	1	1.855	Item T	<---	[Item S,Item J]
rule30	0.539	1	1.855	Item S	<---	[Item T]
rule31	0.539	1	1.855	Item T	<---	[Item S]

Figure 13: Association Rule Learner node

[16 marks]

## 5. Answer the questions based on your findings:

- 5.1 Taking the output in 4.1 as the data source. Perform a Chi-Square test to determine whether having an income between 200,000 and 300,000 impacts the purchase behaviour of Item D. Explain which nodes you used to complete this task, why and how those nodes are used.

(8 marks)

**Ans:** For answering this question, I decide to use three distinct nodes named Column Filter, Rule Engine, and Crosstab respectively. Firstly, the Column Filter node is used to select "Item D" and "Income" columns. Next, the Rule Engine node is employed to transform the numerical income data into a categorical variable, specifically classifying incomes between \$200,000 and \$300,000 into one group and all others into another. Finally, a Crosstab node is used to create a contingency table, which summarizes the counts of individuals based on their categorized income and whether they purchased Item D, setting the stage for the Chi-Square test to assess independence between these two categorical variables.

- 5.2 Explain how you find the Chi-Square test related information step by step to question 5.1 with the critical value of 0.05, and draw a conclusion. Post the screenshot of where you find the answer, too.

(8 marks)

**Ans:** Based on the Chi-Square test performed, the relevant information is found directly in the statistics output of the Crosstab node. Here, the **p-value** for the Chi-Square test is **0.593**. Comparing this to the critical value (significance level) of 0.05, we observe that  $0.593 \geq 0.05$ . Therefore, we **fail to reject the null hypothesis**, concluding that there is no statistically significant evidence to suggest that having an income between \$200,000 and \$300,000 impacts the purchase behavior of Item D.

Statistics Table - 3:80 - Crosstab (Question 5.1 - 5.2)

File Edit Hilite Navigation View

Table "default" - Rows: 1 Spec - Columns: 4 Properties Flow Variables

Row ID	Chi-Square	Chi-Square (DF)	Chi-Square (Prop)	Fisher's Excact Test (2-Tail) (Prop)
Row0	0.286	1	0.593	0.6219832281185544

**Figure 14: Crosstab node**