Swinburne University of Technology Hawthorn Campus
Dept. of Computer Science and Software Engineering

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

Assessment Title: Data Cleaning, Integration, and Analysis
Assessment Weighting: 20%
Due Date: Sunday, 25th 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.

Purpose of Assignment

This assignment aims to evaluate students' achievement of the following unit learning outcomes:

1. **Appreciate (and explain) the key concepts, techniques, and tools for handling the data and producing analytic outcomes.**
2. **Experiencing data cleaning and integration for a data science project.**

This is an individual assignment. Refer to the Unit Outline for the late submission penalty policy. You can ignore the high similarity that appears on the cover page, the template wording, and the short answers. You must make sure your submitted report has a similarity lower than 30% in total and less than 6% from a single source. Otherwise, your report will not be marked.

Key Lessons:

You are asked to use the specified dataset and then build models in KNIME analytic platform and explain your design concept. Two source datasets are provided, and you are expected to perform data cleaning and integration to create a combined dataset. Furthermore, you are expected to perform specified analysis with the produced dataset. KNIME must be used to find answers to all questions except the part asking you to observe the data.

Introduction

Assignment Goal

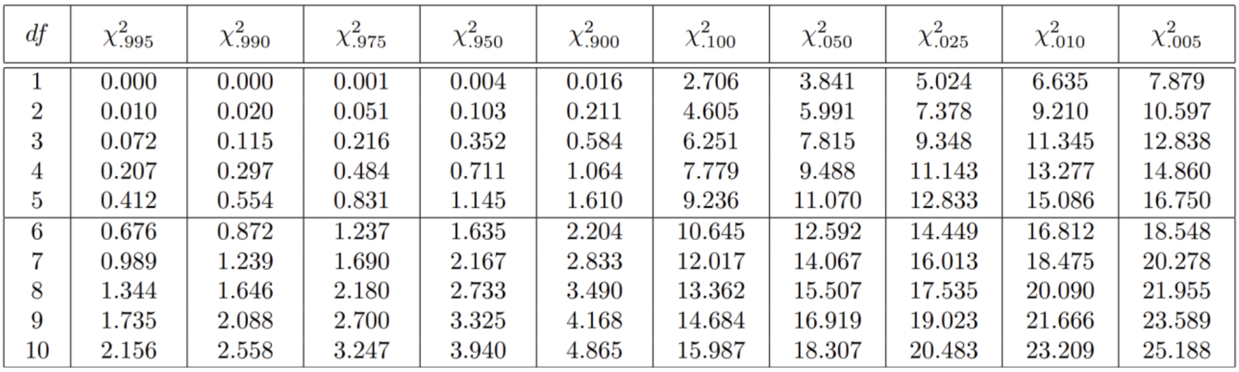
This assignment aims to build experiences for students to clean the data before integrating multiple data sources into a combined dataset and explaining the outputs. A small part of discovery and research component is included in the assignment for expending the skill set of the students.

Hints and Supplement Materials

- Here is the list of nodes that require you to research how to use them. You may use them at least once in the workflow to answer all questions in this assignment:

Column Comparator	*Joiner*	*Crosstab*
Column Combiner	*Math Formula*	-

- Here is a YouTube video regarding using the Crosstab node for the Chi-Square Test in KNIME: https://www.youtube.com/watch?v=YchT55Pywu4
- This assignment involves a heavy part of data cleaning. Observation is the key to completing the assignment correctly. Be careful about the data contained. You may need to do more processing than what was asked in the questions to get the result correct.
- The Chi-Square table is provided below:



df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Assignment Task

We have collected two data sets from different sources at different times. Your task is to clean the dataset according to the requirements and observations. The source files are specified in the tasks. The report should be prepared with the template. A table of Contents is not required.

The data source contains many details of the record. We aim to clean the data sets and then integrate them into one complete dataset before performing data analytics on it. Please follow the steps below to prepare your data after loading it into KNIME. When preparing the zip file, please put all input files under the same path as your KNIME workflow. This will make the marking easier for the tutors. You can keep growing the workflow across multiple steps or start an independent one, depending on your need to complete the tasks. Your submitted workflow must reveal all results corresponding to all questions in the task sheet:

1. Create a KNIME workflow to load the source file “2025_A2_src_d1.csv”. Align the column in the order of “Resident ID,” “Resident,” “DoB,” “Current Age,” “Education,” “Location,” and “Income.” Observe the content and perform data cleaning processes in KNIME. You may need to go back and forward in this process, as some abnormal cases will not be discovered until data integration.
2. Load the second source file entitled “2025_A2_src_d2.csv” in KNIME. Align the column in the order of “Customer ID,” “Name,” “Birthday,” “Age,” “Education Level,” “City,” “Purchase Date,” “Shopping List,” “Item List,” “Item A,” “Item B,” “Item C,” “Item D,” “Item E,” “Item F,” “Item G,” “Item H,” “Item I,” “Item J,” “Item K,” “Item L,” “Item M,” “Item N,” “Item O,” “Item P,” “Item Q,” “Item R,” “Item S,” “Item T,” “Item U.” Observe the content and perform data cleaning processes in KNIME. You may need to go back and forward in this process, as some abnormal cases will not be discovered until data integration.
3. For both data sets, drop the unused columns according to the instructions in the assignment sheet and then integrate them into a complete record.
4. Perform the association rule analysis on the part of the data that is suitable for use in the association

rule analysis.

5. Perform a Chi-Square Test to determine whether a specific group presents different purchasing behaviour on the specified product.

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: __________. Post the screenshot of the node inside your workflow, too. **(2 marks)**
   - 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)**
   - 1.3. Observe the "Resident" data, list and describe the abnormal patterns you observed in this column. **(2 marks)**
   - 1.4. What percentage of data in the "Resident" column is noisy? Post the screenshot of where you find the answer, too. **(2 marks)**
   - 1.5. Observe the "DoB" data, list and describe the abnormal patterns you observed in this column. **(2 marks)**
   - 1.6. How many instances in the "DoB" column are noisy? Post the screenshot of where you find the answer, too. **(2 marks)**
   - 1.7. Observe the "Age" data, list and describe the abnormal patterns you observed in this column. **(2 marks)**
   - 1.8. What percentage of data in the "Age" column is noisy? Post the screenshot of where you find the answer. **(2 marks)**
   - 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)**
   - 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)**
   - 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? Post the screenshot of where you find the answer, too. **(2 marks)**
   - 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)**
   - 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. **(2 marks)**
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)**
   - 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)**
   - 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 **(5 marks)**

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."

2.4 Remove all temporary columns and the columns containing data before 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. **(5 marks)**

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, or 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)**

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. Tutors will check your submitted workflow and your zip file for the combined dataset. **(2 marks)**

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)**

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. **(5 marks)**

5. Answer the questions based on your findings: **[16 marks]**

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)**

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)**

Submission Requirement

To fulfil the requirement of this assignment, the following items should be prepared in the MS Word or PDF format and submitted.

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.

Failure to adhere to the submission requirements will immediately result in no mark for this assignment.

Rubric:

Marks	**100 marks in total**
Description	Mark breakdown can be found in the above section.

-- End of Assignment --