IT2381 DATA WRANGLING

Assignment (40%)

AY2025 Semester 1

Table of Contents

Learning Outcome

By the end of this assignment, you will be able to:

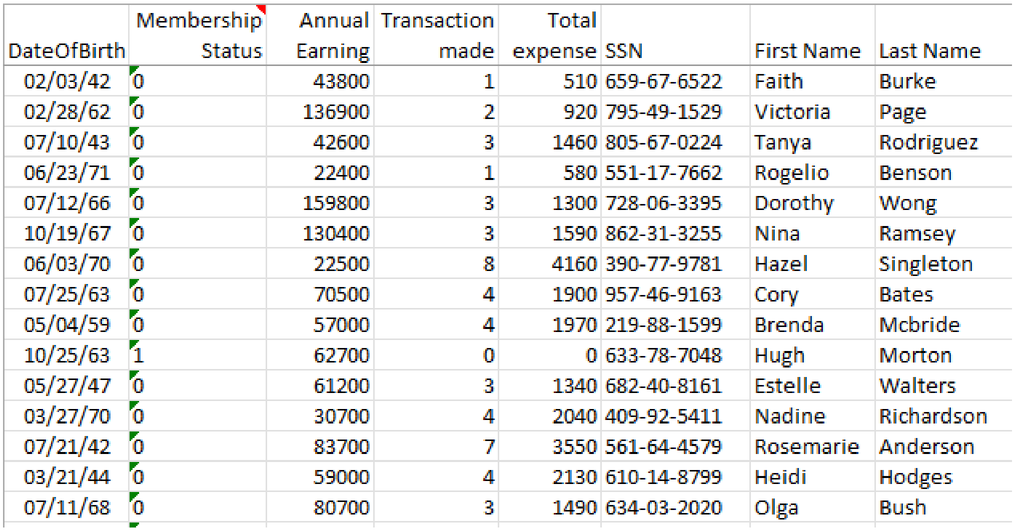
- Identify the common data problem
- Prepare and cleanse the data for analysis
- Derive business insights from the data

Business Scenario

You are a data analytist hired by a retail chain operating in 4 different malls to understand customer behavior and purchasing trends. Each datasets correspond to a specific mall, and the datasets have several issues. Your job is to merge and prepare the data for analyis, and then derive business insights from it.

Data Modeling is not required for this assignment.

Below is a portion of the four datasets.



	Membership	Annual	Transaction	Total			
DateOfBirth	Status	Earning	made	expense	SSN	First Name	Last Name
02/03/42	0	43800	1	510	659-67-6522	Faith	Burke
02/28/62	0	136900	2	920	795-49-1529	Victoria	Page
07/10/43	0	42600	3	1460	805-67-0224	Tanya	Rodriguez
06/23/71	0	22400	1	580	551-17-7662	Rogelio	Benson
07/12/66	0	159800	3	1300	728-06-3395	Dorothy	Wong
10/19/67	0	130400	3	1590	862-31-3255	Nina	Ramsey
06/03/70	0	22500	8	4160	390-77-9781	Hazel	Singleton
07/25/63	0	70500	4	1900	957-46-9163	Cory	Bates
05/04/59	0	57000	4	1970	219-88-1599	Brenda	Mcbride
10/25/63	1	62700	0	0	633-78-7048	Hugh	Morton
05/27/47	0	61200	3	1340	682-40-8161	Estelle	Walters
03/27/70	0	30700	4	2040	409-92-5411	Nadine	Richardson
07/21/42	0	83700	7	3550	561-64-4579	Rosemarie	Anderson
03/21/44	0	59000	4	2130	610-14-8799	Heidi	Hodges
07/11/68	0	80700	3	1490	634-03-2020	Olga	Bush

Here's the data dictionary of the data set:

Field Name	Description
DateOfBirth	Date of Birth of customer
Membership Status	Identify if customer is a member. 1 indicate yes and 0 indicate no
Annual Earnings	Annual income of customer
Transactions Made	Number of transactions customer has made
Total Expense	Total amount spent by customer

SSN	Social Security Number, unique identifier to identify customer
First Name	First Name of customer
Last Name	Last Name of customer

In this **individual** assignment, you are required to perform the tasks listed below to the given datasets.

Task 1: Perform Data Wrangling Tasks using Knime

You are required to perform the following tasks using Knime.

1. Merge the 4 datasets (Mall A, Mall B, Mall C and Mall D) into a single dataset for analysis.
2. Find and fix errors in the dataset. Under Personal Data Protection Act, you will need to conduct the necessary anonymization in your final cleansed dataset.
3. Perform data transformation to prepare the data for analysis.
4. Save the cleansed dataset as CSV or Excel.

Task 2: Prepare a Data Wrangling Report

Part A: Document the steps taken to mashup, clean and transform the data

Part B: List Down the insights you gained from the cleaned dataset

Name your file using your admin number and submit your report, cleansed dataset and Knime workflow in NYP LMS (Brightspace).

Submission Format and Mode

Below are the required deliverables for this assignment.

1. Knime Workflow
2. Cleaned Dataset in MS Excel format
3. Data Wrangling Report in MS Word format

Please be reminded to submit all the deliverables via NYP LMS (Brightspace) **by 15 June (Sunday) 2359hrs.**

The following late submission policy applies.

Days Late	Marks Deduction
If submission is <= 5 days	Cap at 50% of total marks e.g. submission is from 16 June – 20 June 2025 (inclusive) • If learner scored 30 marks, a score of 20 marks will be awarded. • If learner score 15 marks, a score of 15 marks will be awarded.
If submission > 6 days	0 mark will be awarded

Please refer to **Annex A** for detailed assessment rubrics of this assignment.

The base marks of this assignment are **40 marks** and it constitutes **40%** of your total ICA marks for this competency unit.

Copy work from other people or the internet is strictly prohibited. Taking and using the whole or any part of ideas, words or works of others, including contents generated by AI tools and passing it off as one's own work without acknowledgement of the original source will be considered a case of plagiarism and is subject to disciplinary actions.

Annex A: Assessment Rubrics

Tasks	Allocation of Marks
1. Data Mashup	4
2. Data Fixing	11
3. Data Transformation	12
4. Report Part A	5
5. Report Part B	8
Total	40

Rubrics for Data Wrangling Tasks

Criteria	Not Competent (F) 0% to <50%	Developing (D) 50% to <60%	Functional (C) 60% to <70%	Competent (B) 70% to <80%	Proficient (A) 80% to 100%
Data Mashup (4 marks)	Not able to merge the datasets into a single dataset	Able to merge some of the datasets into a single dataset manually	Able to merge some of the datasets into a single dataset	Able to merge all the datasets into a single dataset but with missing records	Able to merge all the datasets into a single dataset with no errors
Data Fixing (11 marks)	Not able to identify the data errors Inappropriate column data type	Able to identify some data errors Few of the columns are formatted to the correct data type	Able to identify the data errors but did not fix Some of the columns are formatted to the correct data type	Able to identify the data errors and fixed some Majority of the columns are formatted to the correct data type	Able to identify and fix all data errors All the columns are formatted to the correct data type
Data Transformation (12 marks)	Not able to identify data transformation needed	Able to identify 1 column for data transformation but did not perform transformation	Able to identify some columns for data transformation but did 1 data transformation	Able to identify all columns needed for data transformation and transformed some	Able to identify and perform all data transformation

Rubrics for Report

Criteria	Not Competent (F) 0% to <50%	Developing (D) 50% to <60%	Functional (C) 60% to <70%	Competent (B) 70% to <80%	Proficient (A) 80% to 100%
Part A: Document the steps (5 marks)	No documentation	Poor/inappropriate documentation of data preparation steps	Some relevant documentation of data preparation steps	Relevant and appropriate documentation of data preparation steps	Relevant and appropriate documentation with explanation on the data preparation steps taken
Part B: List down the insights (8 marks)	No meaningful insights, unclear, or lacks data evidence.	Poor insights, weak justification, limited connection to business impact.	Fair insights with partial justification, but needs improvement..	Good insights, backed by strong data evidence and business impact..	Excellent insights, backed by strong data evidence and business impact.