# ITB431 DATA WRANGLING

## Assignment (40%)

AY2022 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

## Business Scenario

Nanyang Trading Company has customers spreading across four regions (North, South, East & West) in Singapore. The branch manager of each region maintained their own customers data. Nanyang Trading would like to analyse the amount spent by customers in each region. Analysis will include the demographics of the customers such as credit card user, income, age, number of purchases, etc. You have been tasked to prepare the data for analysis. Prior to analysis, the four datasets need to be merged. A portion of the four datasets appears below.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | SSN | First Name | Last Name | Birthdate | Cred Card User | Income | Purchases | Amount Spent |
| 2 | 659-67-6522 | Faith | Burke | 02-03-42 | 0 | 43800 | 1 | 510 |
| 3 | 795-49-1529 | Victoria | Page | 02-28-62 | 0 | 136900 | 2 | 920 |
| 4 | 805-67-0224 | Tanya | Rodriguez | 07-10-43 | 0 | 42600 | 3 | 1460 |
| 5 | 551-17-7662 | Rogelio | Benson | 06-23-71 | 0 | 22400 | 1 | 580 |
| 6 | 728-06-3395 | Dorothy | Wong | 07-12-66 | 0 | 159800 | 3 | 1300 |
| 7 | 862-31-3255 | Nina | Ramsey | 10-19-67 | 0 | 130400 | 3 | 1590 |
| 8 | 390-77-9781 | Hazel | Singleton | 06-03-70 | 0 | 22500 | 8 | 4160 |
| 9 | 957-46-9163 | Cory | Bates | 07-25-63 | 0 | 70500 | 4 | 1900 |
| 10 | 219-88-1599 | Brenda | Mcbride | 05-04-59 | 0 | 57000 | 4 | 1970 |
| 11 | 633-78-7048 | Hugh | Morton | 10-25-63 | 1 | 62700 | 0 | 0 |
| 12 | 682-40-8161 | Estelle | Walters | 05-27-47 | 0 | 61200 | 3 | 1340 |
| 13 | 409-92-5411 | Nadine | Richardson | 03-27-70 | 0 | 30700 | 4 | 2040 |
| 14 | 561-64-4579 | Rosemarie | Anderson | 07-21-42 | 0 | 83700 | 7 | 3550 |
| 15 | 610-14-8799 | Heidi | Hodges | 03-21-44 | 0 | 59000 | 4 | 2130 |
| 16 | 634-03-2020 | Olga | Bush | 07-11-68 | 0 | 80700 | 3 | 1490 |
| 17 | 032-41-3842 | Sherry | Stephens | 02-18-42 | 1 | 94500 | 2 | 1030 |
| 18 | 007-45-8600 | Norman | Osborne | 04-21-39 | 1 | 135600 | 4 | 2040 |
| 19 | 101-30-3651 | Rhonda | Fletcher | 08-03-72 | 1 | 26700 | 2 | 1030 |
| 20 | 997-55-2774 | Jill | Brewer | 01-24-75 | 0 | 40600 | 3 | 1560 |
| 21 | 702-86-2285 | Blanche | Jefferson | 03-29-46 | 0 | 48500 | 3 | 1560 |
| 22 | 105-09-3958 | Valerie | Wise | 07-12-47 | 0 | 109100 | 0 | 0 |
| 23 | 814-97-4282 | Myra | Mcdonald | 07-01-53 | 1 | 24400 | 1 | 550 |
| 24 | 874-40-5083 | Javier | Brady | 02-19-76 | 0 | 62800 | 2 | 970 |
| 25 | 038-72-8622 | Alfredo | Clayton | 04-19-65 | 1 | 117900 | 3 | 1460 |
| 26 | 366-03-5021 | Ignacio | Zimmerman | 09-23-34 | 0 | 121400 | 1 | 530 |
| 27 | 444-05-4079 | Harry | Patton | 01-01-32 | 1 | 23300 | 0 | 0 |

East

Here's the data dictionary of the data set:

| Field Name | Description |
|---|---|
| SSN | Social Security Number, unique identifier to identify customer |
| First Name | First Name of customer |
| Last Name | Last Name of customer |
| Birthdate | Date of Birth of customer |
| Cred Card User | Identify if customer use credit card. 1 indicate yes 0 indicate no. |
| Income | Annual income of customer |
| Purchases | Number of purchases customer has made |
| Amount Spent | Total amount spent by customer |

The data set has several problems. You are required to identify suspicious data values and prepare the data for analysis.

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 (North, South, East & West) into a single dataset for analysis.
2. Find and fix errors in the dataset. Under Personal Data Protection Act, you will need to anonymize the customer name 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

1. Document the steps taken to mashup, clean and transform the data
2. Challenges you have encountered and learnings you have gained while working on the assignment (not more than 500 words)

Name your file using your admin number and submit your report, cleansed dataset and workflow if any 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 13 June (Monday) 2359hrs.**

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. If found, it will be considered a case of plagiarism and is subject to disciplinary actions.**

# Annex A: Assessment Rubrics

| Tasks | Allocation of Marks |
|---|---|
| 1. Merge 4 datasets into a single dataset for analysis | 5 |
| 2. Find and fix the errors in the dataset | 10 |
| 3. Perform data transformation | 10 |
| 4. Final Cleansed Dataset | 5 |
| 5. Data Preparation Report | 5 |
| 6. Reflection Report | 5 |
| Total | 40 |

**Rubrics for Data Wrangling Tasks**

| Criteria | Unsatisfactory (0-4) | Approaching Standards (5-6) | Meeting Standards (7-8) | Exceeding Standards (9-10) |
|---|---|---|---|---|
| **Data Mashup** | Not able to merge the datasets into a single dataset | 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 Preparation** | Not able to identify the data errors<br><br>Inappropriate column data type | Able to identify the data errors but did not fix<br><br>Some of the columns are formatted to the correct data type | Able to identify the data errors and fixed some<br><br>Majority of the columns are formatted to the correct data type | Able to identify and fix all data errors<br><br>All the columns are formatted to the correct data type |
| **Data Transformation** | Not able to identify data transformation needed | Able to identify some columns for data transformation but did not perform transformation | Able to identify all columns needed for data transformation and transformed some | Able to identify and perform all data transformation |
| **Cleansed Dataset** | Cleansed dataset submitted with simple cleaning | Cleansed dataset submitted with not all columns required for data analysis is included | Cleansed dataset submitted with correct columns for data analysis but not formatted correctly | Cleansed dataset submitted with correct columns for data analysis |

## Rubrics for Report

| Criteria | Unsatisfactory (0-4) | Approaching Standards (5-6) | Meeting Standards (7-8) | Exceeding Standards (9-10) |
|---|---|---|---|---|
| **Report** | 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 with explanation on the data preparation stepstaken |
| **Reflection** | Lacks reflection and depth.<br><br>No justifications given and application of learning not described. | Shows little evidence of reasoned reflection.<br><br>Few justifications given and some applications of learning described. | Shows evidence of reasoned reflection but lack depth.<br><br>Some justifications given and application of learning described adequately. | Shows strong evidence of reasoned reflection and depth.<br><br>Adequate justifications given and application of learning clearly described. |