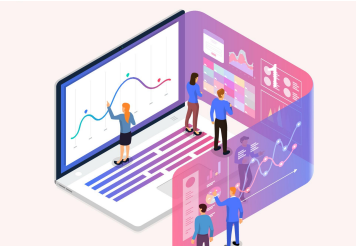
MONASH University
MALAYSIA



## Assignment Sheet

| | |
|---|---|
| **Unit Name** | Introduction to Data Science |
| **Unit Code** | FIT 1043 |
| **Unit Teacher Name** | Ts. Dr. Sicily Ting |
| **Assignment Name** | Assignment 1 (10%) |
| **Assignment Number/Reference** | Exploratory Data Analysis, data visualisation and Wrangling- Python |

## Learning Outcomes

This assignment assesses the following learning outcomes:

| Learning Outcome Number | Learning Outcome Description |
|---|---|
| 1 | Explain the role of data in different styles of business |
| 3 | Identify tasks for data curation and management in an organisation; |

## Weighting

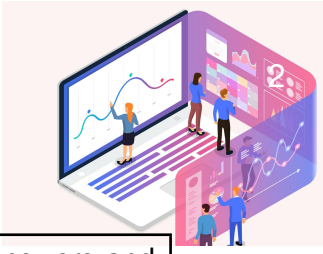
This assignment is worth [**10%**] of your overall grade for this unit. Please see the assignment rubric for the weightings of each assessment criterion.

## Requirements

This assignment has the following requirements:

| | |
|---|---|
| **Assignment Type** | **Individual Task**<br>Task A (44 marks)<br>Task B (56 marks)<br>Total is 100 marks which will be scaled to **10%** of your overall grade for this unit |
| **Response Format** | Two files:<br>1. **PDF file** containing your code, answers and explanations to questions and a<br>2. **Jupyter notebook file (.ipynb)** containing your Python code to all the questions respectively |
| **Response Specifications** | **two separate** files (i.e., .pdf file and .ipynb file). Zip, rar or any other similar file compression format **is not acceptable** and **will have a penalty of 10%.** |
| **Due Date** | 11.55pm (MYT), 29 August 2022 (Monday) |

MONASH University
MALAYSIA



| | |
|---|---|
| **Submission Process** | Please hand in a PDF file containing your code, answers and explanations to questions and a Jupyter notebook file (.ipynb) containing your Python code to all the questions respectively:<br>● **The PDF file** should contain:<br>○ **1. Answers and explanations to the questions.** Make sure to include screenshots/images of the graphs you generate and your Python code (copy and paste your code) to justify your answers for all the questions. (You may need to use screen-capture functionality to create appropriate images.)<br>*[Remark] Please do not include screenshots of used code.*<br>○ 2. You can use Microsoft Word or other word processing software to format your submission, and save the final copy to a PDF before submitting.<br>● **The .ipynb file** should contain:<br>○ 1. **A copy of your work using python code** to answer all the questions.<br>● You will need to submit two separate files (i.e., .pdf file and .ipynb file). Zip, rar or any other similar file compression format is not acceptable and will have a penalty of 10% |
| **Notes:** | The submission must be done via the Moodle site's submission link. |

## Aim

The aim of this assignment is to investigate and visualise data using Python as a data science tool. It will test your ability to:

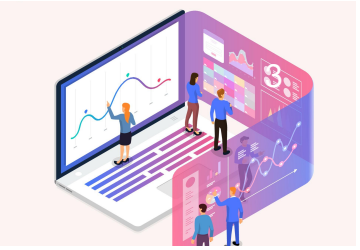
1. read a data file in Python and extract related data from it.

2. use various graphical and non-graphical tools for performing exploratory data analysis and visualisation.

3. use basic tools for managing and processing data and

4. communicate your findings in your report.

MONASH University
MALAYSIA



## Data

The data we will use contains the number of monthly smartcard replacements by reason and type in Queensland and comes from the Queensland government open data initiative.

- The monthly smartcard replacements dataset (monthly_smartcard_replacements.csv) contains all recorded smartcard replacements in Queensland for different smartcard types and reasons each month.
- The information is given under variables; Month (including year and month), Transaction, Smartcard.Type, Action.Reason and Number.of.transactions.
- The file (monthly_smartcard_replacements.csv) is available on the unit Moodle site under Assessments.

## Assignment Tasks:

There are **two main tasks (A and B)** that you need to complete for this assignment. Students that complete **only tasks A1-A7 and B1-B2** can only get a **maximum of Distinction**. Students that **attempt task B3** will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the **highest grade**.

Note: You need to use Python to complete all tasks.

## Task A: Data Exploration and Auditing:

In this task, you are required to explore the dataset and do some data auditing on the monthly smartcard replacements dataset. Have a look at the CSV file (monthly_smartcard_replacements.csv) and then answer a series of questions about the data using Python.

**A1. Dataset size**

How many data instances and variables exist in the given dataset as indicated by the rows and columns?

**A2. Missing values in the dataset**

Are there any null values in the dataset? Report the number of null values in each column.

**A3. Data Types**

What are the different data types for each column?

**A4. Convert Data Type**

Convert data type of column ‘Month’ to a datetime format.

*Hint: Use pandas.to_datetime function to convert the type of 'Month' column to a datetime format as shown in one of your tutorials.*

**A5. Descriptive Statistics**

Calculate summary statistics for the Number.of.Transactions column. What does it tell you? Discuss **at least two observations.**

**A6. Exploring Smartcard Types**

1. How many different (unique) smartcard types are recorded in the 'Smartcard.Type' column? What are those different smartcard types and how many instances are recorded for each type?
2. What is the percentage of Driver Licence Card records as one of the smartcard types in 'Smartcard.Type' column?

**A7. Exploring Reasons for Smartcard Replacement**

1. What are the different reasons for smartcard replacements in the given data and how many instances are observed for each reason?

   *Hint: Check the 'Action.Reason' column.*

2. What is the total number of months in which 100 or more smartcard replacements are reported due to being "Lost"?

## Task B: Group Level Analysis and Visualisation:

In this task, you are required to perform analysis based on data subsets or groups with visualisations where required.

**B1. Investigating Annual Smartcard Replacements**

1. Create a new column named 'Year' extracting the year from the 'Month' column.

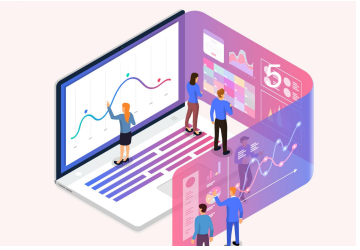
   Hint: you can extract year from column 'Month' using method .dt.year and create a new column for year as follows:

   >>> your_dataframe['Year']=your_dataframe['Month'].dt.year

2. Create a line plot showing the total number of annual smartcard replacements (number of transactions) against year.
3. Explain the trend as observed from the chart. Are there any years that are different from others and if so, what is the reason behind it?

**B2. Investigating Reasons for Smartcard Replacement**

1. Create a barchart showing the total number of transactions for each 'Action.Reason' using the available data.



2. What are the top three reasons for smartcard replacement?
3. Total number of transactions of which 'Action.Reason' is between 1000 and 2000?

**B3. Investigating Reasons over Annual Smartcard Replacement**

1. Find out the annual number of transactions for each 'Action.Reason' over different years that data is available.
2. For each action reason calculate the number of years that the number of annual transactions exceeds 10000.
3. Which action reasons have at least one year with the number of annual transactions exceeding 10000?
4. Create a histogram to analyze the distribution of the annual number of transactions per action reason as calculated in B3.1.
5. Explain any observations and comment on the distribution.

Good Luck!