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. The weightage per question is shown along with the task.

Requirements

This assignment has the following requirements:

Assignment Type	**Individual Task** Task A (50 marks) Task B (50 marks) Total is 100 marks which will be scaled to **10%** of your overall grade for this unit
Response Format	Two files: 1. **PDF file** containing your code, answers and explanations to questions and a 2. **Jupyter notebook file (.ipynb)** containing your Python code to all the questions respectively *Remark on PDF file:* ***Option 1:*** *use any text processing tool to prepare your PDF file.* ***Option 2:*** *convert your Jupyter notebook file as pdf as well as long as the answers, explanations and visualisations are included.*

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), 25 March 2024 (Monday of Week 5)
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: ● **The PDF file** should contain: ○ **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.) *[Remark] Please do not include screenshots of used code.* ○ 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. ● **The .ipynb file** should contain: ○ 1. **A copy of your work using python code** to answer all the questions. ● 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

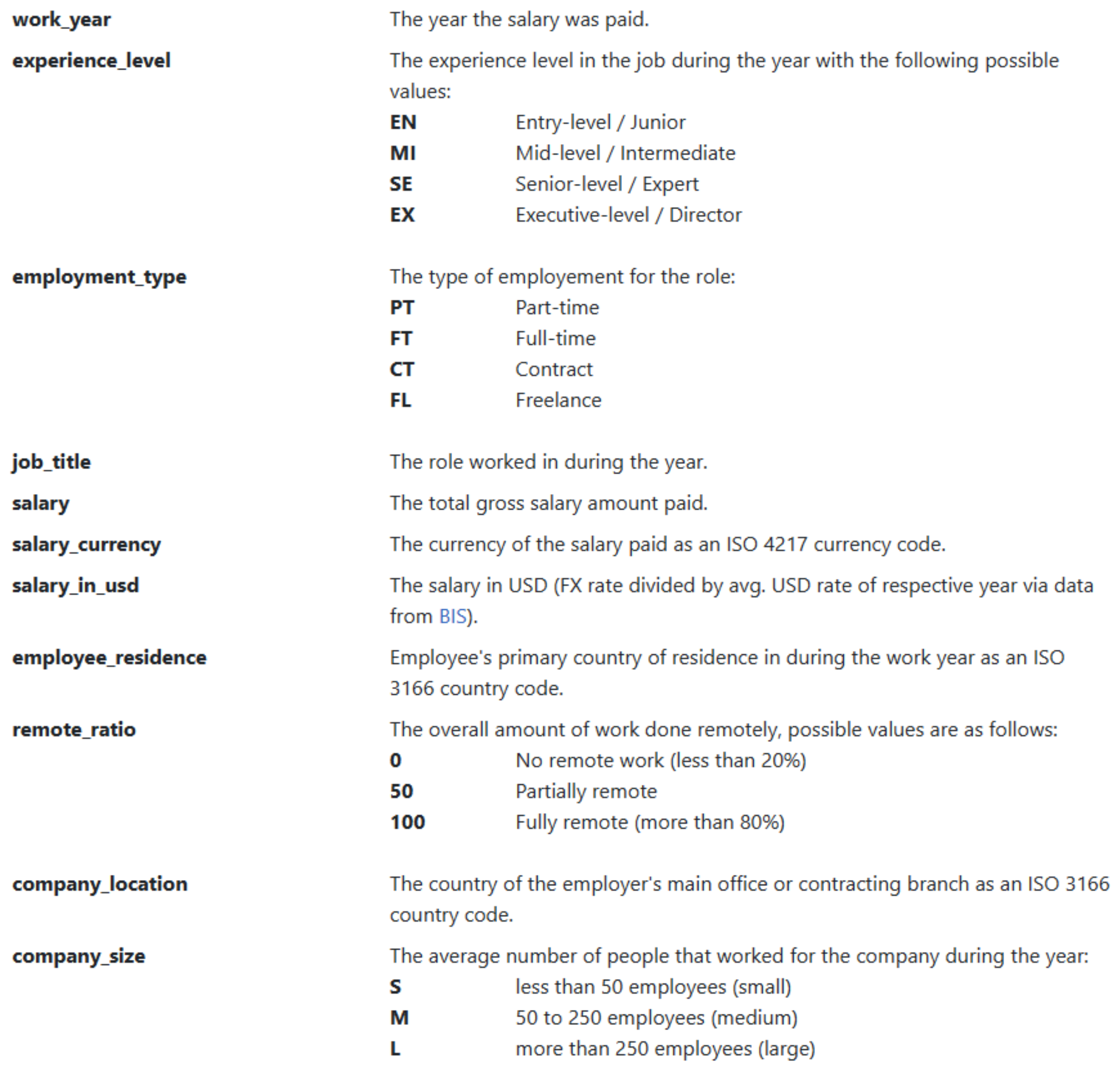
4. communicate your findings in your report.

Data

The data we will use is salaries.csv file, it's AI/ML/Big Data salary data obtained anonymously through the survey at ai-jobs.net/salaries.

The primary goal of this dataset is to have data that can provide better guidance in regards to what's being paid globally. So newbies, experienced pros, hiring managers, recruiters and also startup founders or people wanting to make a career switch can make better informed decisions.

This file contains a single table with all salary information structured as follows:



work_year	The year the salary was paid.
experience_level	The experience level in the job during the year with the following possible values: **EN** Entry-level / Junior **MI** Mid-level / Intermediate **SE** Senior-level / Expert **EX** Executive-level / Director
employment_type	The type of employement for the role: **PT** Part-time **FT** Full-time **CT** Contract **FL** Freelance
job_title	The role worked in during the year.
salary	The total gross salary amount paid.
salary_currency	The currency of the salary paid as an ISO 4217 currency code.
salary_in_usd	The salary in USD (FX rate divided by avg. USD rate of respective year via data from BIS).
employee_residence	Employee's primary country of residence in during the work year as an ISO 3166 country code.
remote_ratio	The overall amount of work done remotely, possible values are as follows: **0** No remote work (less than 20%) **50** Partially remote **100** Fully remote (more than 80%)
company_location	The country of the employer's main office or contracting branch as an ISO 3166 country code.
company_size	The average number of people that worked for the company during the year: **S** less than 50 employees (small) **M** 50 to 250 employees (medium) **L** more than 250 employees (large)

- The file (salaries.csv) is available on the unit Moodle site under Assessments
- Acknowledgement: Dataset is sourced from ai-jobs.net

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 salaries.csv dataset. Have a look at the CSV file (salaries.csv) and then answer a series of questions about the data using Python.

A1. Dataset size [4 marks]

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

A2. Data auditing [6 marks]

Print out the first 8 rows, last 12 rows and random 6 rows of data.

A3. Data Types [4 marks]

What are the different data types for each column?

A4. Conversion [6 marks]

1. Convert the data in the ‘salary_in_usd’ column to **Malaysian ringgit (MYR)** based on the currency exchange rate **(1 USD = 4.47 MYR*).**
2. Create a new column and name it as ‘salary_in_myr’ and insert the converted salary data in MYR to your ‘salary_in_myr’ column

**Currency exchange rate is based on the date when we prepared this document. You may use it as it is, no need to change it to the latest rate.*

A5. Descriptive Statistics [6 marks]

1. Calculate summary statistics for the current dataframe.
2. What does it tell you? Discuss **at least two observations.**

A6. Exploring Job Titles [12 marks]

1. How many different (unique) job titles are recorded in the ‘job_title’ column?
2. What are those different job titles and how many instances are recorded for each job title?

3. What is the percentage of ‘Data Scientist’ records as one of the job titles in the ‘job_title’ column?

A7. Exploring location of Companies [12 marks]

1. What are the different locations for the companies and how many instances are observed for each location?

   Hint: Check the 'company_location' column.

2. [Use the dataframe from **A7.1**] What is the total number of **‘L’** size companies in the **US** ?

   Remark: May use the ISO 3166 Country Code to understand the full name of the countries.

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. It’s up to you to select the appropriate plots/graphs [line graph/scatter plot/bar graph/pie chart/histogram/Motion chart] and provide some basic insights to the following questions

B1. Investigating Employment Type [12 marks]

1. Which job gives the highest salary for Full-time (FT) employment type? [Answer it with a visualisation of your choice and provide your analysis]

2. Which job gives the highest salary for a Part-time (PT) employment type? [Answer it with a visualisation of your choice and provide your analysis]

3. **[Bonus]** Compare the highest salary for each employment type of the job (answer from B1.1) via visualisation and share your insight on it.

B2. Investigating Remote Ratio [16 marks]

1. What are the top three countries that have the highest recorded instances?

2. What is the distribution of remote ratio (0 - No remote work (less than 20%) ; 50 - Partially remote; 100 - Fully remote (more than 80%) for each of these three countries? [Answer it with a visualisation of your choice and provide your analysis]

   [Hint: choose the best plot/graph for this question and provide justification on your choice]

B3. Investigating Experience level [20 marks]

1. Is there any association between Experience Level and the Salary for all jobs? [Answer it with a visualisation of your choice and provide your analysis]

2. Which job has the highest association between Experience Level and the Salary? [Answer it with a visualisation of your choice and provide your analysis]

   [Hint: The job that have highest relevance that the more experience you have, the higher is the salary]

3. Explain any observations and comment on the distribution.

Good Luck!