FIT1043 Assignment 1 Specifications

Due date: Monday 20th April 2020 - 6:00 pm

Objective

The objective of this assignment is to investigate and visualise data using **Python** in the **Jupyter Notebook** environment. This assignment will test your ability to:

- Read data from files in Python,
- Manipulate the data,
- Describe the data using basic statistics,
- Produce non-graphical and graphical visualisation to explore the data,
- Communicate your findings as business insights, and
- Use a 3rd party tool to visualise (bonus).

Data

The data is provided in two comma separated (CSV) files. They are sanitized transactional data from an actual cinema (not completely sanitized but sufficiently anonymized for this exercise).

- The file "`FIT1043-ticket-trx.csv`" contains information about the ticket sales. There are many fields (columns) in the file. Most of the fields are self-explanatory but here's an incomplete list of them:
    - "`Transaction.Number`" relates to a particular ticket sale and "`Transaction.Sequence.Number`" relates to the activity ID that the particular ticket is involved in. *Together they form a unique identifier*.
    - "`Film.Code`" is just an ID for the film and "`Film.HO.Code`" is the long name (which has been sanitized).
    - "`User`" is the worker ID.
    - "`Workstation`" refers to the ID of the Point of Sales (POS) or the cashier station.
    - There are two types of taxes being applied on each ticket sale. One is known as Entertainment Tax and the other is Goods and Services Tax (GST).

- The file "`FIT1043-ticket-seating.csv`" contains information about the location of the seats. What you need to note for this is that for each "`Screen.Name`" (which is the cinema hall ID), the number of seats are different.

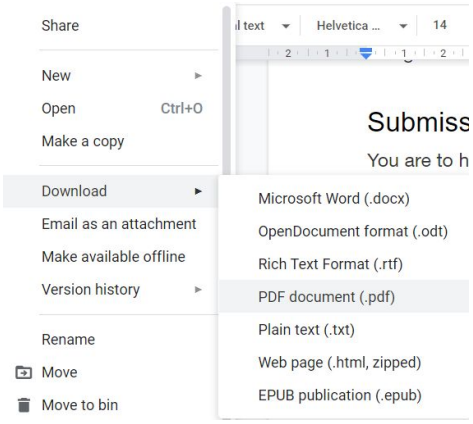
These two files are proprietary and are only for FIT1043 usage. You are NOT to share this file outside of this Unit. It is solely this Assignment's usage.

Submission

This assignment has to be done using the Jupyter Notebook only. Your Jupyter Notebook has to use the Markdown language for proper formatting of the report and answers, with inline Python code and graphs.

You are to hand in two files:

1. The **Jupyter Notebook file (.ipynb)** that contains a working copy of your report (using Markdown) and Python code that answers the questions.
2. A **PDF** file that is generated from your Jupyter Notebook. Execute your Python code and then download it as a PDF document. This can be done using the File menu as show below:



Clarifications

This assignment is not meant to provide step by step instructions and I expect to have lots of questions. The questions can be as simple as “What is a PDF file” to something relating to the possible meaning of a column in the CSV file. But, you are not to post answers directly.

As we are online for this semester, I would like you to post these questions on the Moodle Forum and I strongly encourage interactions between all of you in the forum. Some of the questions probably don’t have a single answer or a correct answer. For example, what does the `HO` in “`Film.HO.Code`” stands for?

Link to Moodle Forum (https://lms.monash.edu/mod/forum/view.php?id=6617339)

Assignment

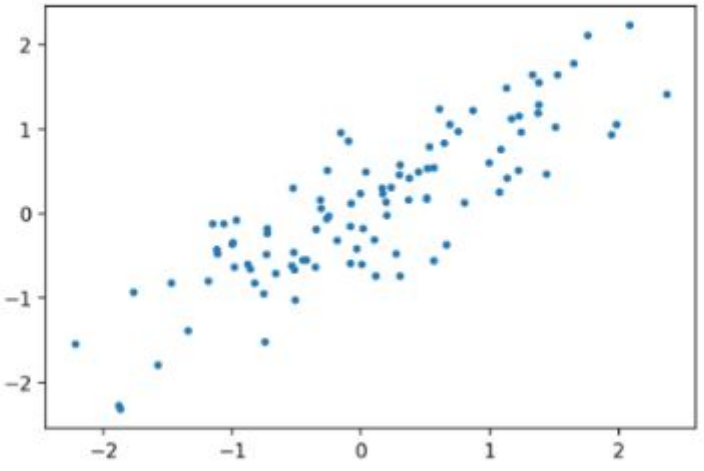
This assignment is worth 20 marks, which makes up for 10% of this Unit's assessment. This assignment has to be done using the **Python programming** language in the **Jupyter Notebook environment**. It should also be formatted properly using the Markdown language. As an example, your Jupyter file should produce something like below (image taken from https://solutions.rstudio.com/examples/python/reports/). For each section, you are to write about your approach, then your code and the output (can be non-graphical or graphical).

This is a .pmd markdown file

The *.pmd* file was written with Atom using Pweave.

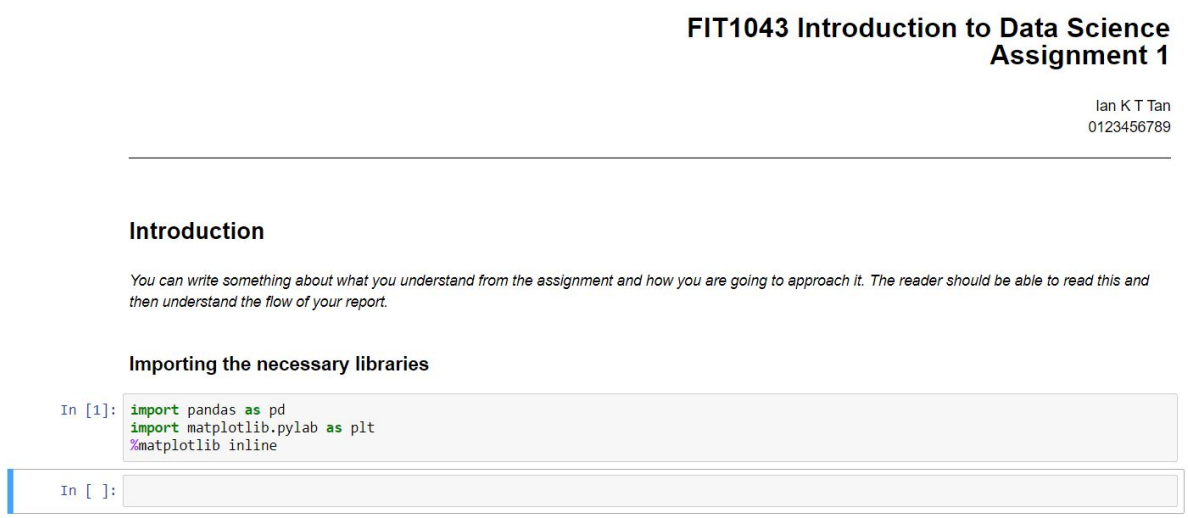
- evaluate single line/selection: `Shift-Enter`
- evaluate code cell using `Alt-Shift-Enter`
- create an empty Python code cell using `p-TAB` or `Alt-Cmd-i`

```
import numpy as np
import matplotlib.pyplot as plt

# Some incredibly difficult code
x = np.random.randn(100)
y = 0.7*x + 0.5*np.random.randn(100)

plt.plot(x,y,'.')
plt.show()
```



Tasks

You should start your assignment by providing the title of the assignment and unit code, your name and student ID, e.g.



FIT1043 Introduction to Data Science
Assignment 1

Ian K T Tan
0123456789

Introduction

You can write something about what you understand from the assignment and how you are going to approach it. The reader should be able to read this and then understand the flow of your report.

Importing the necessary libraries

```
In [1]: import pandas as pd
        import matplotlib.pylab as plt
        %matplotlib inline

In [ ]:
```

You are to start the coding portion by importing the necessary libraries, read the files, and merge the files correctly. You are to provide some description of the data that you have read, e.g. size and also minimally the 5 number statistics of the appropriate columns.

Then you are to answer the following question by showing and explaining your Python code and its output (graphical where you feel is relevant).

1. Which movie generated the **highest revenue** in the dataset?

2. Which day (as in Monday, Tuesday, etc and not date) is the **least popular day** to watch a movie? Hint: You need to convert the date to a day.

3. What is the **most popular time of the day** for movie goers. When you answer this question, we are expecting you to think from the intended audience perspective (the cinema management). Hint: Group the times into different categories (categorical data), e.g. before noon, early evening, etc.)

4. Using the column '`Order.Time..Secs`', determine (other than WEB user) which user has the **best averaged order time** (turnaround). This is basically how fast (efficient) the user is.

You are to end your report with some overall insights of the business. Note that you are NOT limited to providing insights solely based on the 4 questions above, you can do more if you like but your business insights has to be based on your exploratory data analysis.

Marking Rubrics

Administration	Timely submission of assignment	**1** mark
	Proper submission of the .ipynb and PDF files	**1** mark
Report	Appropriately formatted using Markdown (and HTML)	1 mark - Using at least 3 formatting types. **2** marks - Good and easy to read formatting.
	Content	1 mark - Introduction and Conclusion 2 marks - Overall well written report **3** marks - Business Insights
Code	Reading and describing the file content	1 mark 2 marks - State the size of files **3** marks - Basic statistics of the values in the files (only useful fields)
	Merging the files into one DataFrame	1 mark - Merging **2** marks - Neatly and no duplicated fields.
	Question 1	1 mark - Code (logical and executable) **2** marks - Appropriate answer
	Question 2	1 mark - Code (logical and executable) **2** marks - Appropriate answer and visualisation
	Question 3	1 mark - Code (logical and executable) **2** marks - Appropriate answer and visualisation
	Question 4	1 mark - Code (logical and executable) **2** marks - Appropriate answer and visualisation

I use the word “Appropriate answer” as you may make mistakes in the earlier parts but as long as your codes are correct, it will be marked appropriately.

Have Fun!

Upon completion of this assignment, you should have a high level experience of bits and pieces of Drew Conway’s Venn Diagram By completing this assignment, you would have shown that you have “hacking skills” (your Python code), you should have touched upon some basic statistics (although you have not used it effectively for understanding Machine Learning) and hopefully, you have managed to convince (anything logical will be good enough) us that you do know a little bit of the cinema business domain.