



## Assignment Sheet

<b>Unit Name</b>	Introduction to Data Science
<b>Unit Code</b>	FIT 1043
<b>Unit Teacher Name</b>	Ts. Dr. Sicily Ting
<b>Assignment Name</b>	Assignment 2 (20%)
<b>Aim of this assignment</b>	to conduct predictive analytics, by building predictive models on a dataset using Python in the Jupyter Notebook environment

## Learning Outcomes

This assignment assesses the following learning outcomes:

<b>Learning Number</b>	<b>Outcome</b>	<b>Learning Outcome Description</b>
5		Classify the kinds of data analysis and statistical methods available for a data science project;
6		Locate suitable resources, software and tools for a data science project.

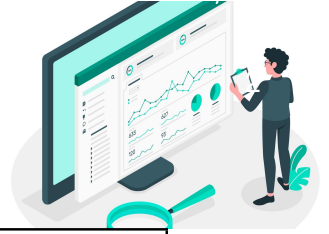
## Weighting

This assignment is worth **[20%]** of your overall grade for this unit.

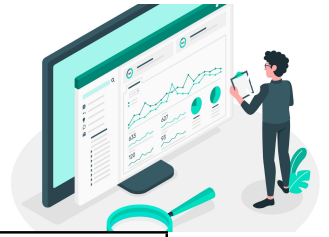
## Requirements

This assignment has the following requirements:

<b>Assignment Type</b>	<b>Individual Task (20%)</b>
<b>Response Format / Hand-in Requirements</b>	<p>Please hand in the following 4 files (including a <b>PDF file</b> containing your code, answers and explanations to questions, a <b>Jupyter notebook file (.ipynb)</b> containing your Python code to all the questions, <b>CSV file</b> for your prediction in task A4 and the <b>video file</b> respectively):</p> <ol style="list-style-type: none"> <li>1. <b>PDF file</b> should contain : <ol style="list-style-type: none"> <li>a. Answers to the questions. In order to justify your answers to all the questions, make sure to</li> </ol> </li> </ol>



	<ul style="list-style-type: none"> <li>i. Include <b>screenshots/images of the graphs or outputs</b> you generate (You will need to use screen-capture functionality to create appropriate images.)</li> <li>ii. <b>Copy/paste of your Python code (not screenshots of your code)</b>.</li> </ul> <p>2. <b>Jupyter notebook file (.ipynb)</b> containing your Python code to all the questions respectively</p> <ul style="list-style-type: none"> <li>a. <b>A copy of your working Python code</b> to answer the questions.</li> </ul> <p>3. The video file should contain:</p> <ul style="list-style-type: none"> <li>a. <b>A recording of yourself, explaining your answers to Part B.</b></li> <li>b. You can use Zoom to prepare your recording.</li> <li>c. Note each student is required to explain their approach in Part B only. Please see Part B for more details.</li> </ul> <p>4. A csv file of your predictions in task A4</p> <hr/> <p>[1] You can use Word or other word processing software to format your submission. Just save the final copy to a PDF before submitting or you can directly convert lpynb file into pdf, if the lpynb file includes everything required.</p>
<b>Response Specifications</b>	<p>1.) <b>Moodle Submission Link:</b>  <b>4 separate</b> files (i.e., .pdf file, .ipynb file, video file and csv file). Zip, rar or any other similar file compression format <b>is not acceptable</b> and <b>will have a penalty of 10%.</b></p>
<b>Due Date</b>	<b>11.55pm (MYT), Monday (2 October 2023)</b>
<b>Disclaimer</b>	<p><b><i>Generative AI tools cannot be used for any assessments in this unit.</i></b></p> <p><i>In this unit, you must <b>not</b> use generative artificial intelligence (AI) to generate any materials or content in relation to your assessment. (see <a href="#">Learn HQ</a>)</i></p>



<b>Notes:</b>	<p>The main submission must be done via the Moodle site's submission link.</p> <p>Kindly refer back to the late penalty on the Assessment tab of Moodle site.</p>
<b>Sanity Checks</b>	<ul style="list-style-type: none"> <li>• After you are done with the tasks, do sanity checks.               <ul style="list-style-type: none"> <li>◦ Run the code and make sure it can be run without errors.</li> <li>◦ You should never submit code that immediately generates an error (warnings are usually fine) when run!</li> </ul> </li> <li>• Make sure that your submission contains everything we've asked for.</li> </ul>

## Aim

The main objective of Assignment 2 is to conduct predictive analytics, by building predictive models on a dataset using Python in the Jupyter Notebook environment.

This assignment will test your ability to:

- Read and describe the data using basic statistics,
- Split the dataset into training and testing,
- Conduct multi-class classification using [Support Vector Machine](#) (SVM)\*\*,
- Evaluate and compare predictive models,
- Explore different datasets and select a particular dataset that meets certain criteria
- Deal with missing data,
- Conduct clustering using k-means

\*\* Not taught in this unit, you are to explore and elaborate these in your report submission.  
This will be a mild introduction to life-long learning to learn by yourself.

## Data

We will explore the following datasets in **Part A** (plus a dataset of your choice in **Part B**):

1. FIT1043-Essay-Features.csv
2. FIT1043-Essay-Features-Submission.csv

**Format:** each file is a single comma separated (CSV) file

**Description:** These two datasets were derived from a set of essays and are used to describe the essay features in numeric information.



**Columns:**

Column's header	Description
essayid	a unique id to identify the essay
chars	number of characters in the essay, including spaces
words	number of words in the essay
commas	number of commas in the essay
apostrophes	number of apostrophes in the essay
punctuations	number of punctuations (other than commas, apostrophes, period, questions marks in the essay
avg_word_length	the average length of the words in the essay
sentences	number of sentences in the essay, determined by the period (fullstops)
questions	number of questions in the essay, determined by the question marks
avg_word_sentence	the average number of words in a sentence in the essay
POS	total number of Part-of-Speech discovered
POS/total_words	fraction of the POS in the total number of words in the essay
prompt_words	words that are related to the essay topic
prompt_words/total_words	fraction of the prompt words in the total number of words in the essay
synonym_words	words that are synonymous
synonym_words/total_words	fraction of the synonymous words in the total number of words in the essay
unstemmed	number of words that were not stemmed in the essay
stemmed	number of words that were stemmed (cut to the based word) in the essay
score	the rating grade, ranging from 1 – 6

This data is pre-processed data on a set of essays that were provided on [Kaggle](#). You **DO NOT** have to download or process/wrangle the data from the original source.

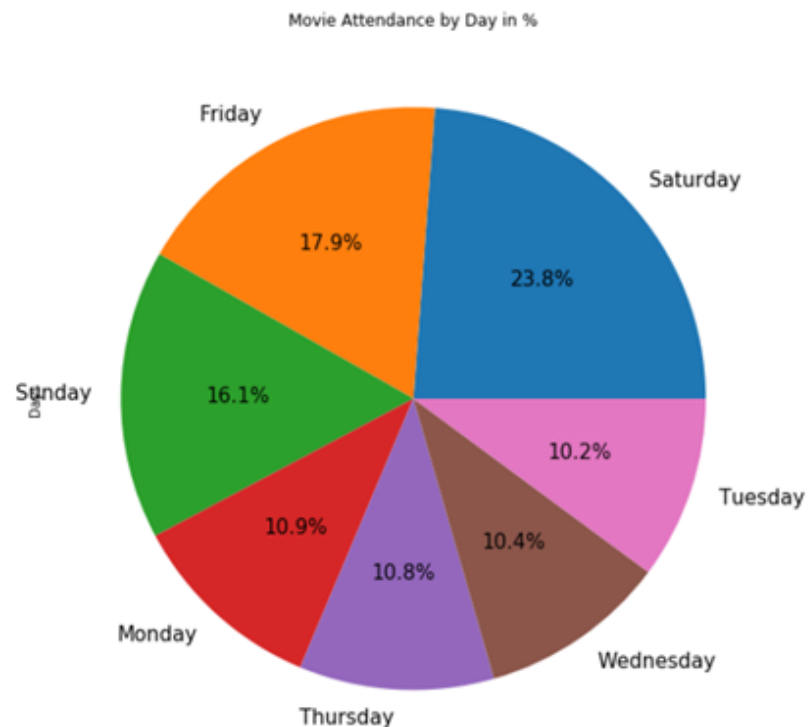


## Assignment Tasks:

This assignment is worth 20% 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. Below is an example from a past submission.

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

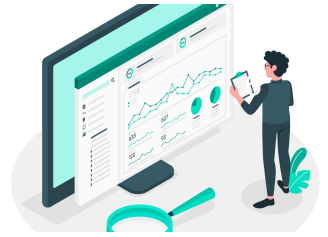
```
In [225]: # Display in pie chart as percentages
ticket_days.plot.pie(title= 'Movie Attendance by Day in %', figsize=(10,10), a
utopct='%1.1f%%', fontsize=15);
```



From our data we can see that Tuesday is the least popular day. The bar graph makes it a little harder to determine which day is the least popular because the bars for four columns are almost similar in height. Hence, a pie chart of percentages is displayed to show which day has the lowest percentage. According to the pie chart we can see that Tuesday has the lowest percentage, of 10.2%, and hence is the least popular day.


### Example 1

*This example has a code cell, the output, which is a rather nice pie chart (with some labels that aren't ideal) and a short explanation.*



## Good practice:

As good practice, 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**  


---

### 1. Introduction

The purpose of this report is to clean, wrangle, analyse, and present the data provided by a cinema. The dataset consists of data from only the month of April in the year 2017. With information such as ticket revenue, day, time, and others, exploratory analysis was performed to search for correlated data as well as to provide suggestions to improve the cinema's business.

The report's rough outline is as follows:

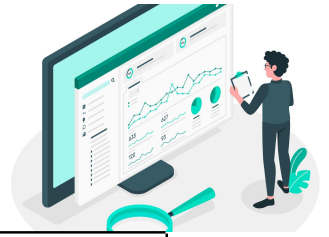
1. Introduction
2. Importing the necessary libraries
3. Simple edits
4. Data Auditing
5. Questions
6. Business Insights
7. Conclusion
8. References

A brief summary of the analysis questions is as follows:

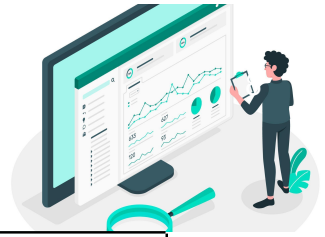
1. Film 5 is the movie that generated the highest revenue in the dataset.
2. The least popular day to watch a movie is Tuesday.
3. The most popular time of the day for movie goers is during the evening, which is from 16:00 to 19:59.
4. The user with the best averaged order time (turnaround) is User\_13.

### Example 2

*This is also a sample from past submissions..*

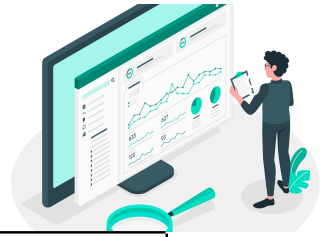


Assignment Task(s)	Description
<b>Part A : Classification</b>	
<b>A1. Supervised Learning</b>	1. Explain supervised machine learning, the notion of labelled data, and train and test datasets.
	2. Read the ' <b>FIT1043-Essay-Features.csv</b> ' file and separate the features and the label (Hint: the label, in this case, is the 'score')
	3. Use the <code>sklearn.model_selection.train_test_split</code> function to split your data for training and testing.
<b>A2. Classification (training)</b>	1. Explain the difference between binary and multi-class classification.
	2. In preparation for classification, your data should be normalised/scaled. <ol style="list-style-type: none"> <li>Describe what you understand from this need to normalise data (this is in your Week 7 applied session).</li> <li>Choose and use the appropriate normalisation functions available in <code>sklearn.preprocessing</code> and scale the data appropriately.</li> </ol>
	3. Use the Support Vector Machine algorithm to build the model. <ol style="list-style-type: none"> <li>Describe SVM. Again, this is not in your lecture content, you need to do some self-learning.</li> <li>In SVM, there is something called the kernel. Explain what you understand from it.</li> <li>Write the code to build a predictive SVM model using your training dataset. (Note: You are allowed to engineer or remove features as you deem appropriate)</li> </ol>
	4. Repeat <b>Task A2.3.c</b> by using another classification algorithm such as Decision Tree or Random Forest algorithms instead of SVM.



<b>A3. Classification (prediction)</b>	1. Using the testing dataset you created in <b>Task A1.3</b> above, conduct the prediction for the 'score' (label) using the two models built by SVM and your other classification algorithm in <b>A2.4</b> .
	2. Display the confusion matrices for both models (it should look like a 6x6 matrix). Unlike the lectures, where it is just a 2x2, you are now introduced to a multi-class classification problem setting.
	3. Compare the performance of SVM and your other classifier and provide your justification of which one performed better.
<b>A4. Independent evaluation (Competition)</b>	1. Read the ' <b>FIT1043-Essay-Features-Submission.csv</b> ' file and use the best model you built earlier to predict the 'score' for the essays in this file.
	2. Unlike the previous section in which you have a testing dataset where you know the 'score' and will be able to test for the accuracy, in this part, you don't have a 'score' and you have to predict it and submit the predictions along with other required submission files. <ul style="list-style-type: none"> <li>a. Output of your predictions should be submitted in a CSV file format. It should contain 2 columns: 'essayid' and 'score'. It should have a total of 200 lines (1 header, and 199 entries).</li> </ul>
<b>Part B : Selection of Dataset, Clustering and Video Preparation</b>	
<b>B1. Selection of a Dataset with missing data, Clustering</b>	<p>We have demonstrated a k-means clustering algorithm in week 7. Your task in this part is to find an interesting dataset and apply k-means clustering on it using Python. For instance, Kaggle is a private company which runs data science competitions and provides a list of their publicly available datasets: <a href="https://www.kaggle.com/datasets">https://www.kaggle.com/datasets</a></p> <ol style="list-style-type: none"> <li>1. Select a suitable dataset <b>that contains some missing data and at least two numerical features</b>. Please <b>note</b> you cannot use the same data set used in the applied sessions/lectures in this</li> </ol>





	<p>unit. Please include a link to your dataset in your report. You may wish to:</p> <ul style="list-style-type: none"> <li>• provide the direct link to the public dataset from the internet, or</li> <li>• place the data file in your Monash student - google drive and provide its link in the submission.</li> </ul>
	<p>2. Perform wrangling on the dataset to handle the missing data and explain your procedure</p>
	<p>3. Perform k-means clustering, choosing two numerical features in your dataset, and apply k-means clustering to your data to create k clusters in Python (<math>k \geq 2</math>)</p>
	<p>4. Visualise the data as well as the results of the k-means clustering, and describe your findings about the identified clusters.</p>
<p><b>B2. Video Preparation</b></p>	<p>Presentation is one of the important steps in a data science process. In this task you will need to prepare a short video of yourself (you can share your code on screen) and describe your approach on the above task (<b>Task B1</b>).</p> <ul style="list-style-type: none"> <li>• Please make sure to keep your camera on (show yourself) during recording. You may want to share your screen with your code while you talk.)</li> </ul>

## Clarifications

This assignment is not meant to provide step by step instructions and as per Assignment 1, do use the Moodle Forum (<https://edstem.org/au/courses/12193/discussion/>) so that other students can participate and contribute. For postings on the forum, do use it as though you are asking others (instead of your lecturer or tutors only) for their opinions or interpretation. Just note that you are not to post answers directly.



**Upon completion** of this assignment, you should have some experience with the *Collect*, *Wrangle*, *Analyse* and *Present* process that is core to the role of a Data Scientist (See Lecture 1, Data Science Process).

## Congratulations!

By completing Assignment 1, you would have experienced looking, understanding, and auditing data. You would also have provided exploratory analytics using descriptive statistics and visualisation. In doing so, you would have had to spend some time sieving through the data to understand it. That was the intention to get you to experience it.

For Assignment 2, we moved to focus on preparing your data for analytics, conducting machine learning using available libraries to build various models, output your results and get the results to be independently evaluated.

You should now be ready to start to build a machine learning portfolio by entering proper Kaggle competitions. This should give you an introduction to the role of a data scientist.

Good Luck! 😊