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 2 (20%)
Aim of this assignment	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:

Learning Outcome Number	**Learning Outcome Description**
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.

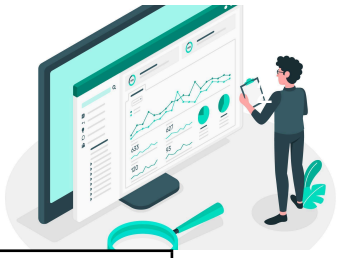
Weighting

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

Requirements

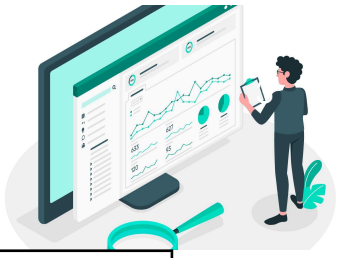
This assignment has the following requirements:

Assignment Type	**Individual Task (20%)**
Response Format / Hand-in Requirements	There are **2 submissions** for this, they are • Moodle submission • Kaggle submission (Competition's Link) **1.) Moodle Submission:**



	○ Submit the following 2 files (including a **Jupyter notebook file (.ipynb)** containing your Python code, answers and explanations(if required) to all the questions, and **CSV file** for your **prediction in task A4** respectively) 1. **Jupyter notebook file (.ipynb)** containing your Python code to all the questions respectively a. **A copy of your working Python code** to answer the questions. b. make use of markdown for any observation explanation/ justification. 2. A **csv file** of your predictions in task A4 **2. Kaggle Submission** The purpose of the Kaggle submission is to provide you with an introductory experience on how machine learning models are evaluated. Another file, called the "`FIT1043-MusicGenre-Submission.csv`" consists of data where there are no labels (no '`music_genre`' column). The whole purpose is to be able to predict those labels for this data set. You are to output the data to a CSV file that contains 6490 rows (6491 if include the headers) and 2 columns, the column "`instance_id`" and another column named "`music_genre`". A sample file without the '`music_genre`' entries is also available "`99999999-YourName-v1.csv`".
Response Specifications	1.) **Moodle Submission Link:** **2 separate** files (i.e., .ipynb file, and csv file). Zip, rar or any other similar file compression format **is not acceptable** and **will have a penalty of 10%.** **2.) Kaggle's Submission - the csv file with 2 columns (ref. "99999999-YourName-v1.csv"**
Due Date	**11.55pm (MYT), Tuesday (30 April 2024), Week 9**

MONASH University
MALAYSIA



Disclaimer	***Generative AI tools cannot be used for any assessments in this unit.*** *In this unit, you must **not** use generative artificial intelligence (AI) to generate any materials or content in relation to your assessment. (see Learn HQ)*
Notes:	The main submission must be done via the Moodle site's submission link. Kindly refer back to the late penalty on the Assessment tab of Moodle site/ the rubric file.
Sanity Checks	● After you are done with the tasks, do sanity checks. ○ Run the code and make sure it can be run without errors. ○ You should never submit code that immediately generates an error (warnings are usually fine) when run! ● Make sure that your submission contains everything we've asked for.

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.

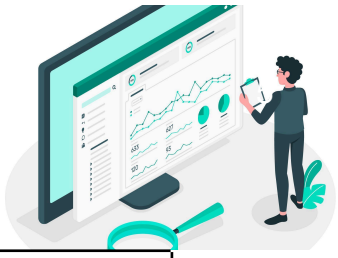
Data

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

1. FIT1043-MusicGenre-Dataset.csv
2. FIT1043-MusicGenre-Submission.csv

Format: each file is a single comma separated (CSV) file
Description: These two datasets were derived from a list containing features of the list of songs and their music genre.
Columns: There should be 15 columns consisting of the features of the song and the class/label of the song (Hint: the music_genre column)

Column Header	Description
instance_id	an unique ID assigned for each entry
artist_name	the name of the artist
track_name	the name/title of that song
popularity	The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.
acousticness	**A confidence measure from 0.0 to 1.0 of whether the track is acoustic**. 1.0 represents high confidence the track is acoustic
danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
duration_ms	The duration of the track in milliseconds.
energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
instrumentalness	Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
speechinesss	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio



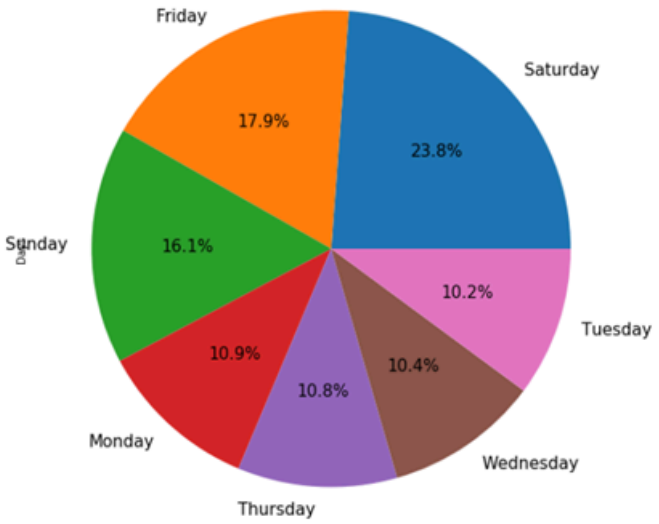
	book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
music_genre	Music genres are represented by the following code: 0 - Alternative 1 - Anime 2 - Blues 3 - Classical 4 - Country 5 - Electronic 6 - Hip-hop 7 - Jazz 8 - Rap 9 - Rock

This data is pre-processed data that was extracted from Spotify and 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

[redacted]

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:

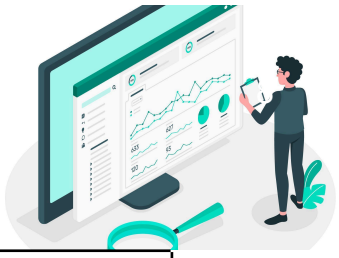
1. Introduction
2. Importing the necessary libraries
3. Simple edits
4. Data Auditing
5. Questions
6. Business Insights
7. Conclusion
8. References

A brief summary of the analysis questions is as follows:

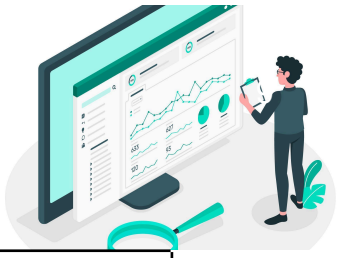
1. Film 5 is the movie that generated the highest revenue in the dataset.
2. The least popular day to watch a movie is Tuesday.
3. The most popular time of the day for movie goers is during the evening, which is from 16:00 to 19:59.
4. The user with the best averaged order time (turnaround) is User_13.

Example 2

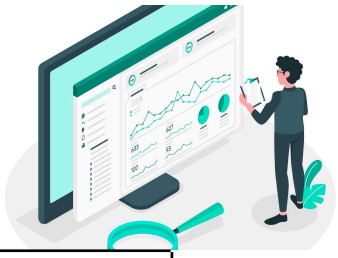
This is also a sample from past submissions..



Assignment Task(s)	Description
Part A : Classification	
A1. Supervised Learning	1. Explain supervised machine learning, the notion of labelled data, and train and test datasets.
	2. Read the '**FIT1043-MusicGenre-Dataset.csv**' file and separate the features and the label *(Hint: the label, in this case, is the 'music_genre')*
	3. Use the *sklearn.model_selection.train_test_split* function to split your data for training and testing.
A2. Classification (training)	1. Explain the difference(s) between binary and multi-class classification.
	2. In preparation for classification, your data should be normalised/scaled. a. Describe what you understand from this need to normalise data (this is in your Week 7 applied session). b. Choose and use the appropriate normalisation functions available in sklearn.preprocessing and scale the data appropriately.
	3. Use the Support Vector Machine algorithm to build the model. a. Describe SVM. Again, this is not in your lecture content, you need to do some self-learning. b. In SVM, there is something called the kernel. Explain what you understand from it. c. 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)
	4. Repeat **Task A2.3.c** by using another classification algorithm such as Decision Tree or Random Forest algorithms instead of SVM.



A3. Classification (prediction)	1. Using the testing dataset you created in **Task A1.3** above, conduct the prediction for the 'music_genre' (label) using the two models built by SVM and your other classification algorithm in **A2.4.**
	2. Display the confusion matrices for both models (it should look like a 10x10 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.
A4. Independent evaluation	1. Read the '**FIT1043-MusicGenre-Submission.csv**' file and use the best model you built earlier to predict the 'music_genre' for the songs in this file.
	2. Unlike the previous section in which you have a testing dataset where you know the 'music_genre' class and will be able to test for the accuracy, in this part, you don't have a 'music_genre' and you have to predict it and submit the predictions along with other required submission files. a. Output of your predictions should be submitted in a CSV file format. It should contain 2 columns: 'instance_id' and 'music_genre'. It should have a total of 6491 lines (1 header, and 6490 entries).
A5. Kaggle Competition	Submit to the Kaggle Submission site with the **2 columns** csv file (Obtained from A4.2.a)) with the naming as *"StudentID-YourName-VersionNumber.csv"* e.g.: 99999999-SicilyTing-v1.csv *Remark: A sample file has been provided "99999999-YourName-v1.csv"* *Bonus mark on students that are placed at Top 10% of the leaderboard placement.



Part B : Selection of Dataset and perform Clustering	
B1. Selection of a Dataset with missing data, Clustering	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: https://www.kaggle.com/datasets 1. Select a suitable dataset **that contains some missing data and at least two numerical features**. Please **note** you cannot use the same data set used in the applied sessions/lectures in this unit. Please include a link to your dataset in your report. You may wish to: ● provide the direct link to the public dataset from the internet, or ● place the data file in your Monash student - google drive and provide its link in the submission.
	2. 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 (k>=2)
	3. Visualise the data as well as the results of the k-means clustering, and describe your findings about the identified clusters.

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/15857/discussion/1841024) 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! ☺