Unit: FIT 1043 – Introduction to Data Science
Semester 1, Year 2022 Prepared by: Ts. Dr. Sicily Ting

FIT1043 Assignment 2 Specifications (20%)

8th April 2022, Version 1.0

Due date: Friday 29th April 2022 – 11.55 pm

Table of Content:

1. Objectives and learning outcomes

Assignment 1 covered the process of conducting *descriptive analytics*, whereas the objective of this assignment (Assignment 2) is to conduct *predictive analytics*, through machine learning 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 and exposure to some natural language processing (NLP)** terms,
- Split the dataset into training and testing (for this assignment, it may sometimes be referred to as validation dataset),
  - And opportunity (**not required** for this Assignment) to create multiple training and testing datasets, in order to conduct cross validation.
- Conduct multi-class classification using Support Vector Machine / Regression (SVM)**,
- Communicate the output of your analysis using the Quadratic Weighted Kappa (QWK) measure**,
- Experience independent model evaluation through submission for an in-class Kaggle competition.

** 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.

Contribute to the following learning outcomes:

LO2. Demonstrate the **size and scope of data storage** and data processing, and classify the basic technologies in use;

LO3. Identify **tasks for data curation** and management in an organisation;

LO6. Locate **suitable resources**, software and tools for a data science project;

2. Data

Format: a single comma separated (CSV) file

Description: This data was derived from a set of essays and is used to describe the essay features in numeric information.

Columns:

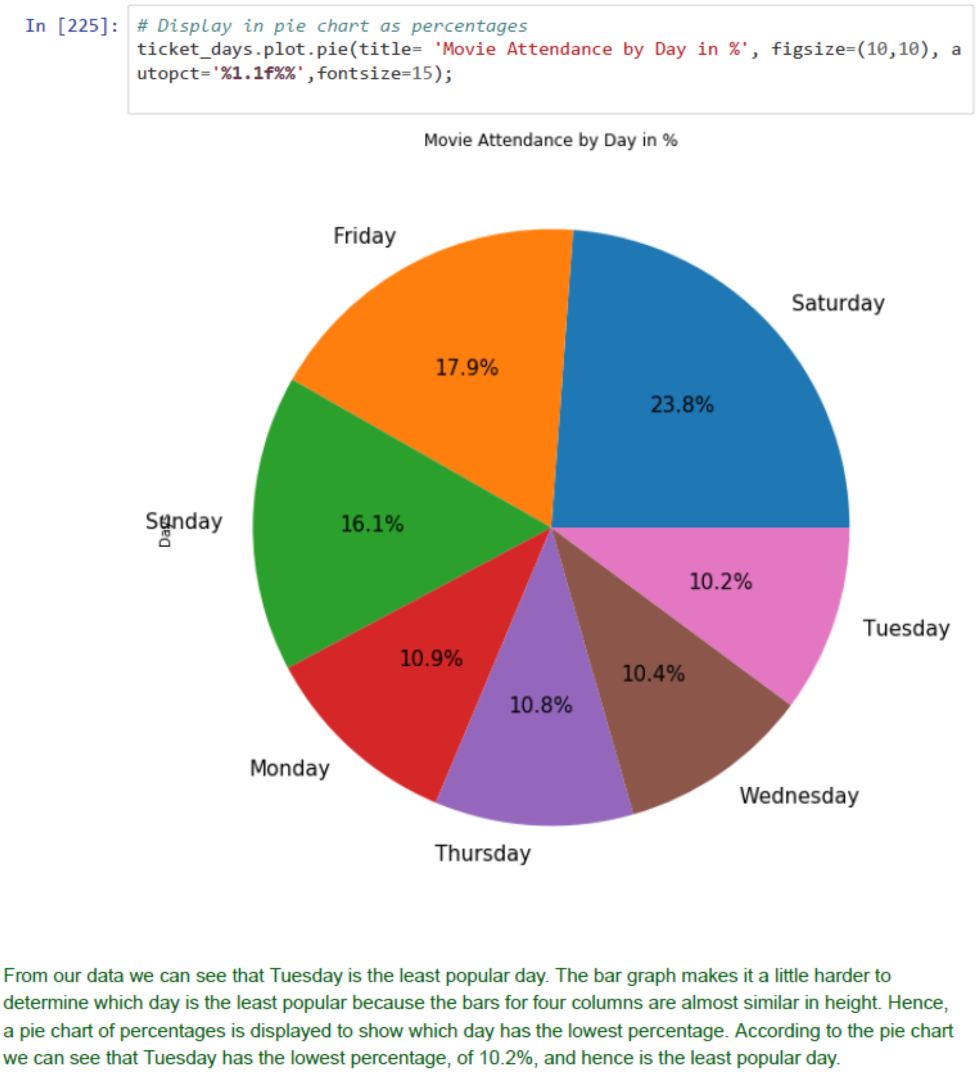
Column's header	Description
essayid	a unique id to identify the essay
chars	number of characters in the essay, including spaces
words	number of words in the essay
commas	number of commas in the essay
apostrophes	number of apostrophes in the essay
punctuations	number of punctuations (other than commas, 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 a 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.

3. Tasks

3.1. Format

This assignment is worth 40 marks, which makes up for 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 past submission.



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.

Figure 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.

3.2. Tasks

Section	**Tasks**
1. Introduction	a. Start with an introduction to the assignment. b. Importing the necessary libraries, read the file (‘FIT1043-Essay-Features.csv‘), and provide some description of the data you have read (you do not need to repeat the description given in this file for each field).
2.Supervised Learning	a. Explain supervised machine learning, the notion of labelled data, and the training and test datasets. b. Separate the features and the label (Hint: the label, in this case, is the ‘score’) c. Use the `sklearn.model_selection.train_test_split` function to split your data for training and testing.
3.Classification	a. Explain the difference between binary and multi-class classification. b. In preparation for Support Vector Machine/Regression, your data should be normalised/scaled. i. Describe what you understand from this need to normalise data (this is in your Week 7 laboratory). ii. Choose and use the appropriate normalisation functions available in `sklearn.preprocessing` and scale the data appropriately. c. Use the Support Vector Machine/Regression algorithm to build the model. i. Describe SVM (in relation to Linear Regression). Again, this is not in your lecture content, you need to do some self-learning. ii. In SVM/SVR, there is something called the kernel. Explain what you understand from it. iii. Write the code to build the model using your training dataset. (Note: You are allowed to engineer or remove features as you deem appropriate) d. Predict i. Using the testing dataset you created in 2(c) above, conduct the prediction for the ‘score’ (label). ii. Display the confusion matrix (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.

	iii. Explain Quadratic Weighted Kappa (QWK). Again, this is not in your lectures. iv. Use the `sklearn.metrics` library to code and obtain the QWK score.
4.Kaggle submission's task	a. Read the 'FIT1043-Essay-Features-Submission.csv' file and use the model you built earlier to predict the 'score'. b. Unlike the previous section, you have a testing (also sometimes referred to as the validation) 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 it to the competition site. c. Output your prediction to a CSV file that contains 2 columns, 'essayid' and 'score'. It should have a total of 200 lines (1 header, and 199 entries).
5.Conclusion	Conclude your assignment

3.3 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.

4. Submission

There are 2 submissions for this assignment, they are

- **Moodle submission**
  - Files to hand in:
    - **Jupyter Notebook file (.ipynb)**
    - **PDF** file version of your **Jupyter Notebook**
- **Kaggle submission**
  Click this link to the kaggle competition page:
  https://www.kaggle.com/t/7d9f63612b794f9bbd94cf1aca80856c

4.1. Moodle 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 if applicable).

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 for the data analytics.
2. A **PDF** file that is generated from your Jupyter Notebook. Execute your Python code, select "**Print Preview**"

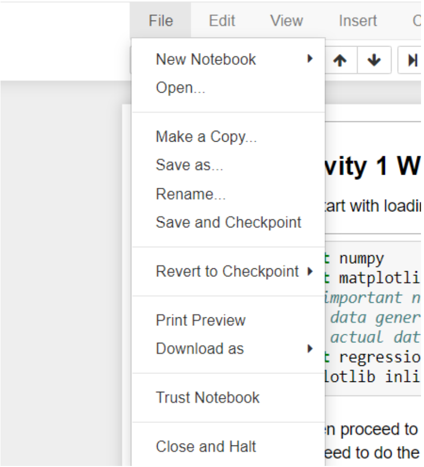


Figure 2. 'File' tab of Jupyter Notebook.

You will be presented with the output in your browser. If you are on Windows, you can then right click and select "**Print**" (similar function should be available on your Mac).

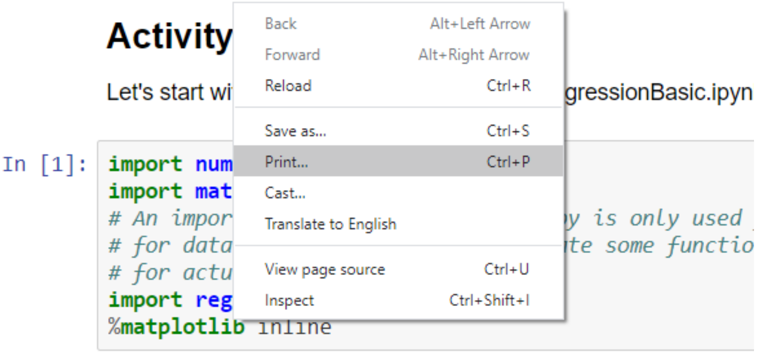
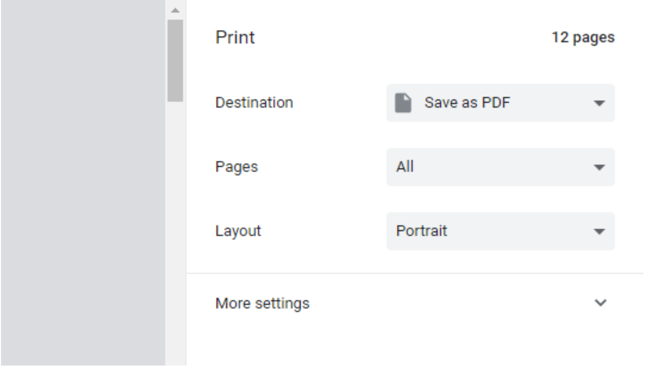


Figure 3. Pop-up menu after you right clicked on the file.

You should then be presented with a print dialog box, which should have a "**Save as PDF**" option instead of your printer.


Print 12 pages

Destination: Save as PDF

Pages: All

Layout: Portrait

Figure 4. Pop-up menu for 'Print'.

Save it as a PDF and submit this PDF file.

Note: that there were some problems with some browsers to be able to do this properly, so do try out other browsers (Chrome works).

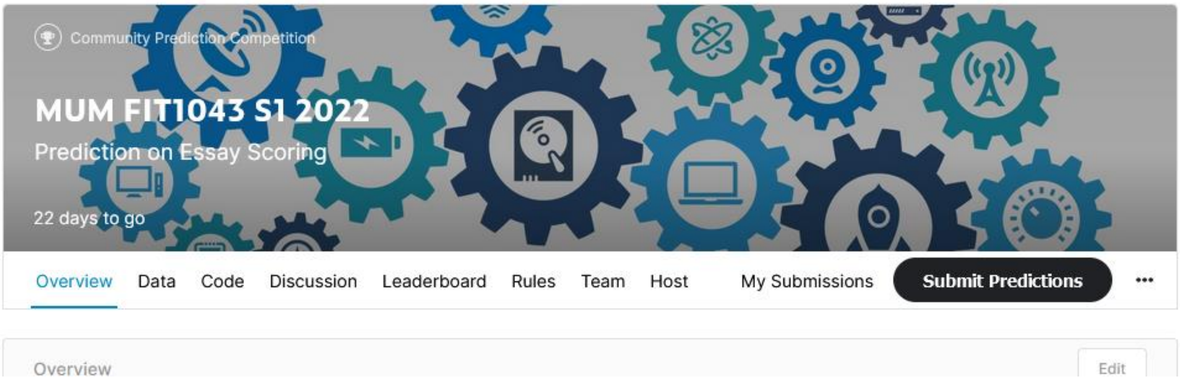
4.2. Kaggle submission

The purpose of the Kaggle submission is to provide you with an introductory experience on how machine learning models are evaluated. Data (“`FIT1043-Essay-Features.csv`”) is provided on the Kaggle (in-class) competition site for building the model (This is the same fie as the one on your Moodle Assessment posting).

Another file, called the “`FIT1043-Essay-Features-Submission.csv`” consists of data where there are no labels (no ‘`score`’ 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 199 rows (200 if include the headers) and 2 columns, the column “`essayid`” and another column named “`score`”. A sample file without the ‘score’ entries is also available “`YourID-YourName-v1.csv`”.

Competition Submission and Evaluation Method

When you logged on the competition page, submissions can be made on the rightmost tab.



You will be asked to read and understand the rules of the competition. When you have prepared your submission CSV file, drag and drop your CSV file (please name it using your “`YourID-YourName-version.csv`” into the space allocated and give it a description, so that you can track it. Please note that the submission process may take a **LONG** time, so please be patient. The system is evaluating your submission.

The competition uses the QWK (a function of error 😊) evaluation method, which will result in a score of between 0.0 and 1.0.

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 skipped the data wrangling and moved to focus on preparing your data for analytics, conducting machine learning using available libraries to build various models, output your results and got 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.

5. Marking Rubric

This assignment is worth 40 marks, which makes up for 20% of this Unit's assessment.

The marking rubrics is just a guideline, and it may vary slightly depending on your approach.

Report	Appropriately formatted using Markdown (and HTML) and content	1 mark – Good use of formatting codes (Markdown or HTML) **3 marks** - Good and easy to read submission, including introduction and conclusion.
Tasks	Reading and describing the file content	1 mark – Importing libraries and reading file(s) **2 marks** – Basic descriptive statistics of the values in the file(s)
	Supervised Learning	1 mark – explain supervised machine learning and the notion of labelled data 2 marks – explain the training and test datasets 3 marks – code the split of the features and labels 6 marks – code the split of the test and train dataset **7 marks** – explain binary and multi-class classification
	Data Normalisation	1 mark – explain the purpose of normalised / scaling of data **3 marks** – code the scaling using the library sklearn.preprocessing
	SVM Classification Model	1 mark – Explain SVM (must have some reference comparison to linear regression)

		2 marks – Explain the kernel in SVM **4 marks** – SVM Model building code
	Model Testing	2 marks – Code the prediction using testing data (note that the output should be integer values) 4 marks – Code, output and explain the confusion matrix 5 marks – Explain QWK **7 marks** – Code, output and explain the QWK score
	Kaggle Submission Preparation	3 marks – Read the competition data file, do the prediction. **6 marks** – Output to the right CSV format.
Kaggle submission		2 marks - submission and executable 5 marks - Below benchmark score but valid value (e.g. >0.0) **8 marks** - Passed the benchmark ** Extra 2 marks for submission placed at Top 10% leaderboard.

Have Fun!