FIT1043 Assignment 2 Specifications

Due date: Monday 18th May 2020 - 6:00 pm

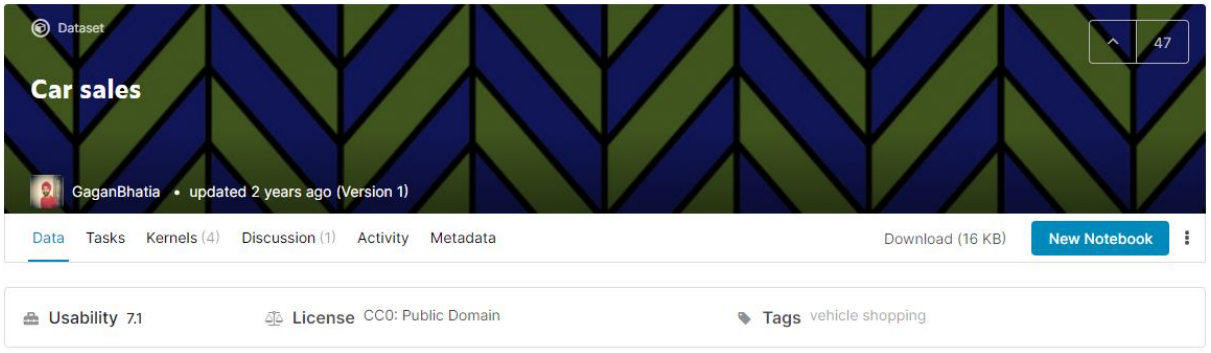
Objective

The objective of this assignment is to conduct machine learning on a small dataset using Python in the Jupyter Notebook environment. This assignment will test your ability to:

- Read and write data to a file in Python,
- Describe the data using basic statistics,
- Wrangle the data,
- Clustering and describe the output,
- Split your dataset into training and testing data for binary and multi-class classification,
- Conduct classification and communicate the output of your analysis, and
- Experience independent model evaluation.

Data

The data is provided in a single comma separated (CSV) file. It is an augmented (modified) version of the file that was originally obtained from a Kaggle Dataset called "Car sales". The data is under CCO: Public Domain and hence can be used freely.



For the assignment (note, there is another set of data for the competition submission), you are given a single file "`FIT1043-vehicle-classifier.csv`", which contains vehicles from around (year) 2000. The fields in the CSV file are self explanatory, but here are some description of the dataset:

- The measurements are US measurements, where the dimensions, such as height, are in inches and the weight is in pounds (lbs).

- Fuel capacity is in US gallons (different from the rest of the world) and fuel efficiency is in miles per gallon (mpg).
- There are 3 columns for vehicle classification;
    - “Vehicle_class” is the common classification that we know,
    - “Vehicle_alt_class” are additional classification that are used for more detailed classification, and
    - “US_vehicle_type” is a simple two class (binary) classification.
- “US_vehicle_type” classifies vehicles based on size where in general “passenger” are for people carriers and “car” is multi-utility.

Submissions

There are 2 submissions for this, they are

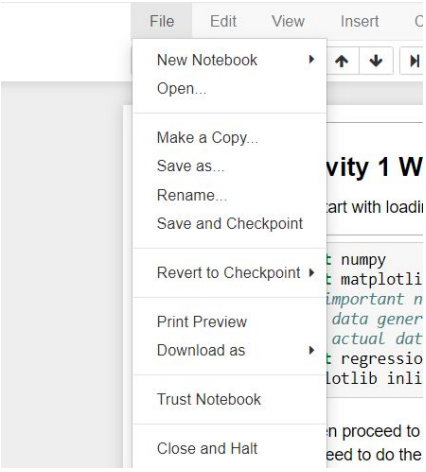
- Moodle submission
- Kaggle submission (https://www.kaggle.com/t/5665004d5a91460f8ed3cb34c188923c)

Moodle Submission

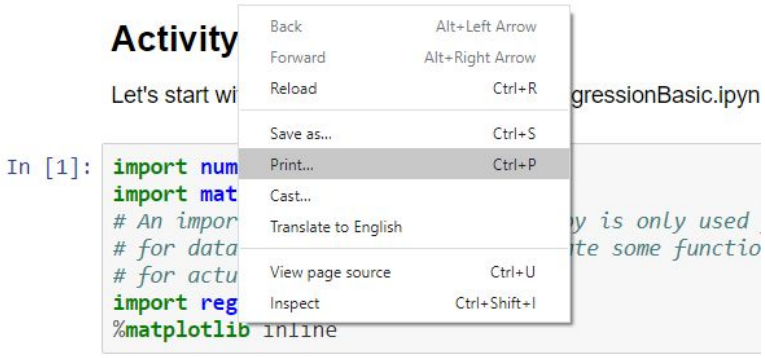
This assignment has to be done using the Jupyter Notebook only. Your Jupyter Notebook has to use the Markdown language for proper formatting of the report and answers, with inline Python code and graphs.

You are to hand in two files:

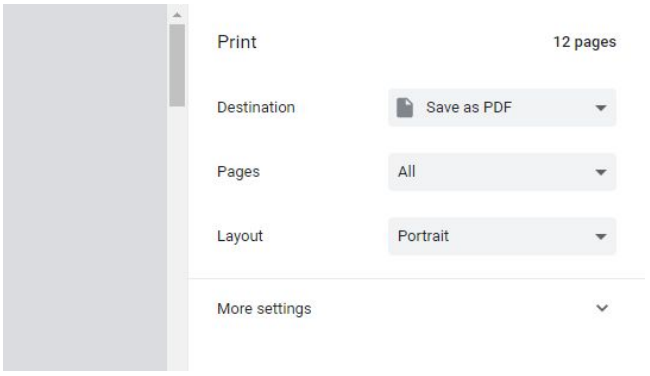
1. The **Jupyter Notebook file (.ipynb)** that contains a working copy of your report (using Markdown) and Python code for the data analytics.
2. A **PDF** file that is generated from your Jupyter Notebook. Execute your Python code, select "**Print Preview**"



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).



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



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).

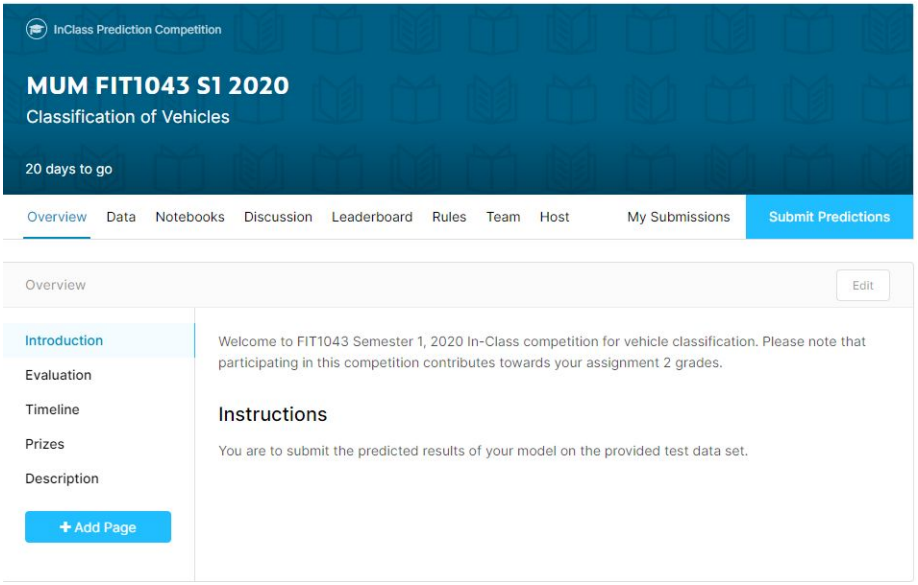
Kaggle Submission

The purpose of the Kaggle submission is to provide you with an introductory experience on how machine learning models are evaluated. Training data ("`FIT1043-kaggle-train-data.csv`") is provided on the Kaggle competition site (https://www.kaggle.com/t/5665004d5a91460f8ed3cb34c188923c) for building the model. This training data is the same as the data you used for the main part of your assignment. The difference is that the vehicle brand, model and 2 other columns have been removed to keep it neater and to ensure that you build a model that is appropriate for the evaluation. You are expected to split this data into your training and test dataset while you build your model. (Note that you can also build your model without testing, but you won't really know how good it is if you don't test).

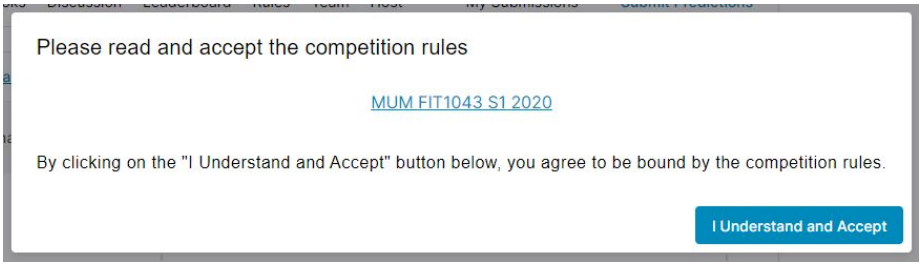
Another file called the "`FIT1043-kaggle-test-data.csv`" consists of data where there are no labels. The whole purpose is to be able to predict those labels for this data set. You are to output the data to a CSV file that contains 17 rows and 2 columns, the column "`Id`" and another column named "`Predicted`".

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.



Drag and drop your CSV file (please name it using your "`studentID-name-version.csv`" into the space allocated and give it a description, such as submission number 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 a very simple evaluation method, by simply computing the number of correctly predicted categories (% of correct predictions).

Clarifications

This assignment is not meant to provide step by step instructions and as per Assignment 1, do use the Moodle Forum (https://lms.monash.edu/mod/forum/view.php?id=6617339) 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.

Alternatively, the slack unofficial channel can also be used.

Assignment

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 are two examples from sample submissions of Assignment 1.

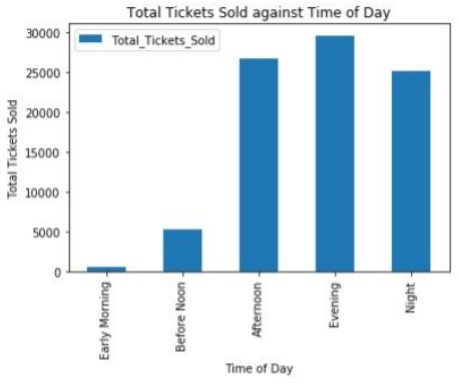
A bar chart was plotted to easily see the difference in the number of tickets sold based on the different times of the day.

In [1009]:

```
df3_mpt.plot.bar(x='Time_Of_Day',y='Total_Tickets_Sold')
plt.xlabel('Time of Day') # setting a label for x axis
plt.ylabel('Total Tickets Sold') # Setting a label for y axis
plt.title('Total Tickets Sold against Time of Day') # Setting the title of chart
```

Out[1009]:

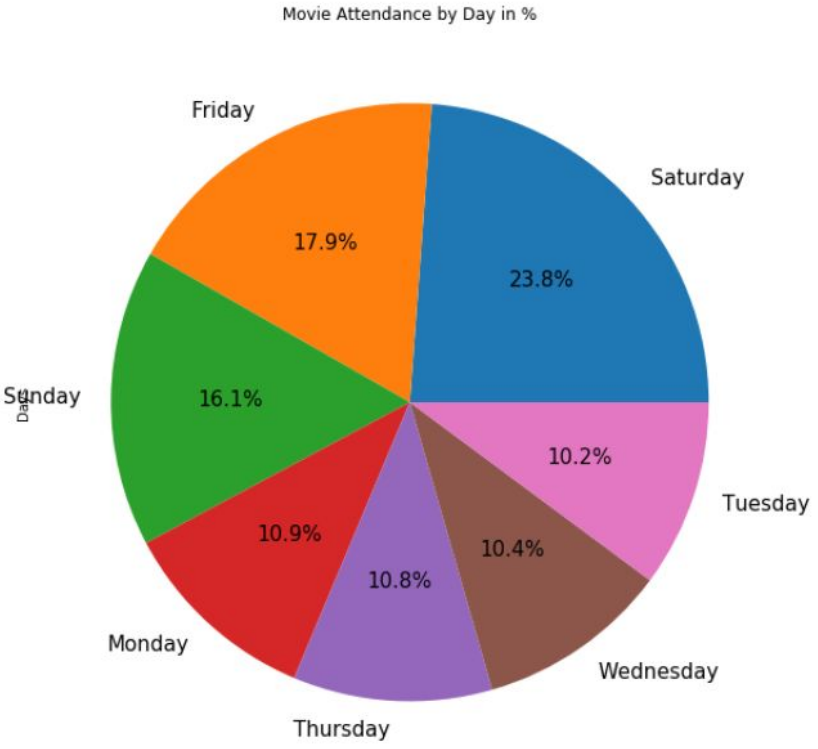
```
Text(0.5, 1.0, 'Total Tickets Sold against Time of Day')
```



However, the bar charts were pretty close in terms of height, and the figures of each column could not be seen clearly. Thus, a pie chart was made as well.

Example 1
This example has some markdown, code, output (graph) and further explanation on why he thinks that maybe the bar chart is not suitable.

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);
```



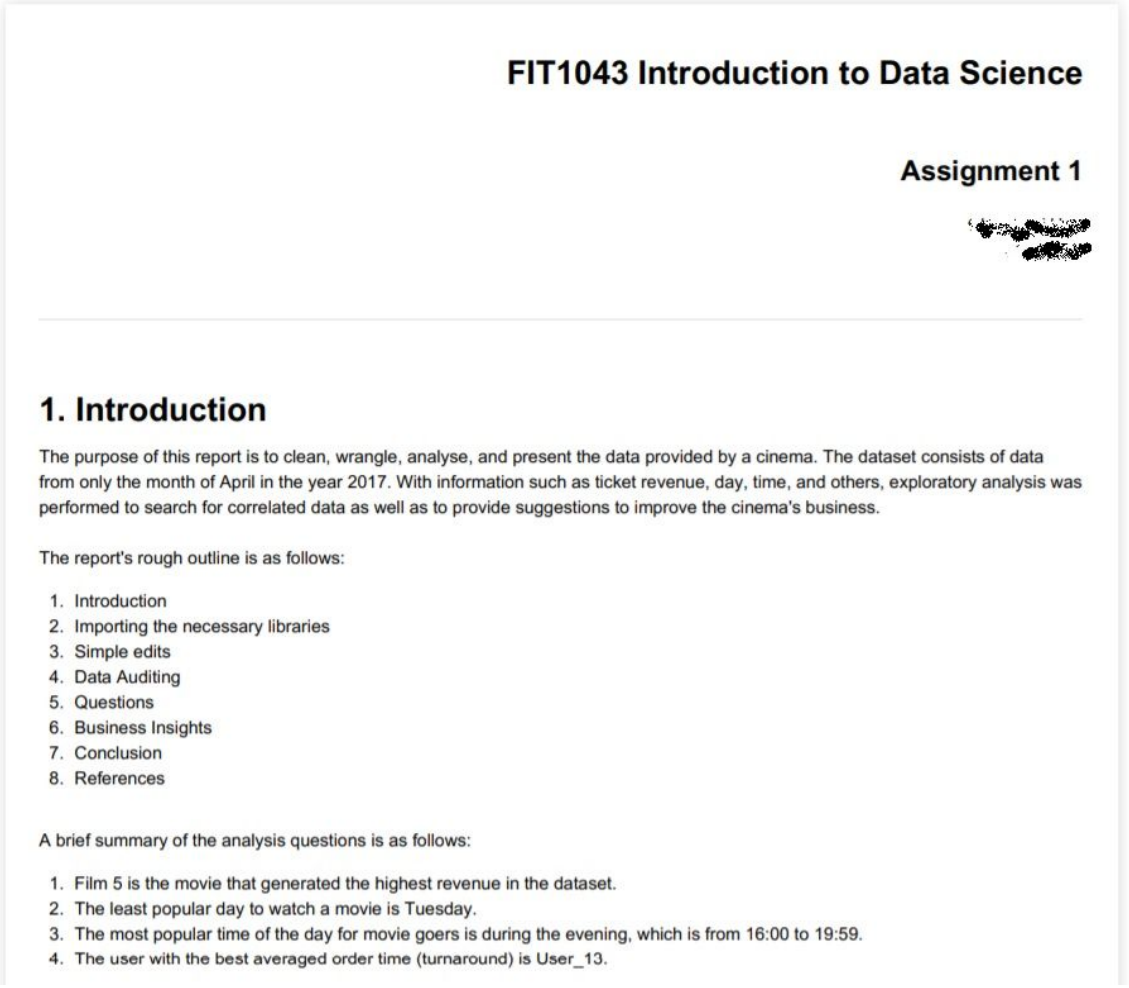
From our data we can see that Tuesday is the least popular day. The bar graph makes it a little harder to determine which day is the least popular because the bars for four columns are almost similar in height. Hence, a pie chart of percentages is displayed to show which day has the lowest percentage. 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 2

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.

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

[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 3
This is also a sample from Assignment 1 submission.

Thereafter, you are to complete the following tasks:

1. Introduction
    a. Start with an introduction to the assignment.
    b. Importing the necessary libraries, read the file, and provide some description of the data you have read.
    c. After data inspection, do the appropriate data wrangling that you deem necessary and describe the wrangled data.
2. Clustering
    a. For clustering purposes, remove the columns "`Vehicle_class`", "`Vehicle_alt_class`" and "`US_vehicle_type`". We will only use this for the classification part of the assignment.

b. Start with an explanation of un-supervised machine learning.
c. Before you start to cluster, explain the input that you have chosen. You do not need to use all the remaining fields. Justify your selection.
d. Conduct clustering using *k*-means, where you are to set *k* to be equal to 2, 5 and 7. (Three models). Note: You can create new features based on the existing data, for example if you want to have power output per litre of engine capacity, you can create a new column and use the "`Horsepower`" column divided by the "`Engine_size (litre)`" column. Whether you are doing a 2-dimension or *n*-dimension, do provide the visualisation for it.
e. Explain your clustering output. To explain the output, you can inspect the cluster and look at the data (observations) that have been assigned to the various clusters, do they have a certain pattern?
f. Where appropriate, you may re-do your clustering using an "improved" set of inputs. What were the results and why do you think it would be a better cluster?

3. Classification
   a. Explain the difference between binary and multi-class classification.
   b. Explain supervised machine learning, the notion of labelled data and the training and test datasets.
   c. Conduct a binary classification using the decision tree algorithm, where the labelled data is the "`US_vehicle_type`".
   d. Conduct a multi-class classification using the decision tree algorithm, where the labelled data is the "`Vehicle_class`".
   e. Conduct another multi-class classification with the "`Vehicle_alt_class`" as the label. You are to decide how you want to fill in the empty cells (if you have not already decided in Part 1 (c) above.
   f. For Parts 3 (c), (d) and (e), output your predicted results in a two column CSV file, the first column is a numbered sequence starting from 1 (to represent the first column of your test data) and the second column is the predicted output.
   g. Decide on the evaluation metrics and explain how you evaluated your output. You can be a little creative here.
4. Conclusion
   a. Re-do your Part 3 (d) using the Random Forest algorithm. Is the result better? (Keep the best model that you have for Part 3 (d) for the Kaggle in-class competition).
   b. Conclude your assignment.

Marking Rubrics

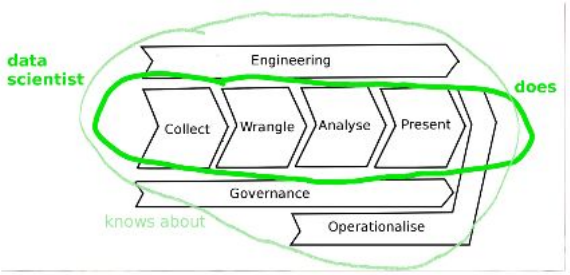
Administration	Timely submission of assignment	0 mark - More than 1 day late 1 mark - Late but less than 1 day **2** marks - timely submission
	Proper submission of the .ipynb and PDF files	1 mark - one of the file is correct format **2** marks - both files submitted
Report	Appropriately formatted using Markdown (and HTML)	1 mark - Using at least 3 formatting types. **3** marks - Good and easy to read formatting.
	Content	1 mark - Introduction **3** marks - Overall well written and organized report
Code	Reading and describing the file content	1 mark - Reading file **2** marks - Basic statistics of the values in the files
	Data Wrangling	1 mark - Identifying inconsistent data **2** marks - Appropriate treatment of data
	Clustering	1 mark - Unsupervised machine learning 2 marks - Feature (column) selection 4 marks - Code (logical and executable) for the 3 *k*-values. (1 + 0.5 + 0.5 marks for the code) **6** marks - Appropriate answers and visualisation
	Classification	1 mark - Binary and multi-class classification 2 marks - Supervised machine learning 3 marks - Correctly explaining and split training and testing data 6 marks - Code (logical and executable) **9** marks - Appropriate visualisation and evaluation.
	Conclusion	1 mark - Code (logical and executable) for Random Forest. 2 marks - Appropriate visualisation and evaluation **3** marks - Assignment conclusion.
	Kaggle submission	2 marks - Submission 4 marks - Submission and Executable 5 - **8** marks - Depending on placing in the ladder. (the placing doesn't carry much weight)

Have Fun!

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).

Data Scientist

Addresses the data science process to extract meaning / value from data



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 inspection 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.