COMP60711 - Part 2 Coursework 1

Course Unit:	COMP60711: Data Engineering
Responsible Staff Member:	Professor John Keane
Marks:	This course is worth **25%** of the overall marks for this unit.
Submissions:	This is the **3rd** of **4** assessed submissions.
Method of Submitting:	This notebook, after completion, should be saved as a HTML document and submitted using Blackboard
Deadline:	Thursday 28th October 9AM (UK time)
Late Submissions:	Extensions will only be granted as a result of formally processed Mitigating Circumstances. Marks for late submissions will be reduced in line with the University policy.

Please complete the questions in the spaces provided (under the "Answer" block for each question), then download the notebook in HTML format and submit to Blackboard.

Please also add your student ID and name below.

Student ID (7-8 digit number)	Full Name
12345678	Name Here

Reminders

- **Please make clear any assumptions and provide evidence to justify your answers**
- Jupyter notebooks use markdown. A brief summary of how to use markdown can be seen here. Otherwise, please refer to the brief guide on Blackboard.
- You **must** cite any sources used, from web pages to academic papers and textbooks.
- Please ensure your code has no errors, and that the output is shown in your submitted version.
- We have added some general notebooks on Blackboard to cover the basics of plotting in Python, Jupyter notebooks, and anaconda.
- Some questions require a mixture of code and text to answer the question. Marks are awarded based on the output of your code (i.e. graphs) and the explanation provided, not on the code itself.

Question 1: Clustering (16 marks)

The following sub-questions are about clustering. In general, the topics covered are as follows:

- Question 1.1 focuses on clustering algorithmic behaviour, and their sensitivity to data.
- Question 1.2 focuses on method for estimating the number of clusters.
- Question 1.3 uses a large real-world dataset to look at how clustering can be used for knowledge discovery.

The following reading is likely to be of use for the following questions:

1. Book chapter on clustering: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf
   - Note that certain sections may be useful, you are not expected to read it all!
2. Here is an example of running KMeans using scikit-learn: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

For this question, we will use multiple datasets, which can all be found on Blackboard. To load these datasets in, we can use the following code. Note that you may need to adjust the path to the dataset, depending on where they are located on your system.

In [ ]:
```
import pandas as pd

def load_dataset(file_name, **kwargs):
    try:
        data = pd.read_csv(file_name, **kwargs)
    except FileNotFoundError:
        raise FileNotFoundError(f"Cannot find {file_name}, please either plac
    labels = data.values[:, -1]
    data = data.values[:, :-1]
    return data, labels

# This loads the dataset used in Q1.1a
simple_data, simple_labels = load_dataset("simple.csv")
```

Q1.1 (6 marks)

Q1.1a (2 marks)

Using `simple.csv`, run the K-Means and single-linkage algorithms (available in `scikit-learn`) on the dataset, using the true number of clusters ($K = 5$). Produce a graph (e.g. bar chart) showing the performance (measured using the adjusted rand index) across 10 independent runs. Discuss the results obtained, using your knowledge of how the algorithms work to explain why the behaviour observed occurred.

Hints:

- This question is more difficult without the use of error bars!
- For single-linkage, use `AgglomerativeClustering(n_clusters=5, linkage="single")`
- For KMeans, use `KMeans(n_clusters=5, init="random", n_init=1)` as arguments.

Q1.1a Answer

In [ ]:
```
# Please enter any relevant code for Q1.1a below.
# You can use multiple cells if it helps break up the code execution
```

Q1.1b (2 marks)

Using the dataset named `q1-1b.csv`, perform the same experiment as in Q1.1a. Discuss the ARI results obtained, particularly how and why the results are different from Q1.1a.

Hint:

- Visualizing the data itself may help to support your discussion.

Q1.1b Answer

In [ ]:

Q1.1c (2 marks)

Using the dataset named `q1-1c.csv`, perform the same experiment as in Q1.1b. Discuss the results obtained, particularly how and why the results are different from Q1.1b.

Hints:

- Visualizing the data itself may help to support your discussion.

Q1.1c Answer

In [ ]:
```
# Please enter any relevant code for Q1.1c below.
# You can use multiple cells if it helps
```

Q1.2 (5 marks)

Q1.2a (3 marks)

Discuss **two** methods that can be used to estimate the *true* number of clusters. A discussion of the suitability and potential issues with each method is expected. No marks are given beyond the first two methods.

Q1.2a Answer

In [ ]:
```
# Please enter any relevant code for Q1.2a below
# You can use multiple cells if it helps
```

Q1.2b (2 marks)

Using the `simple.csv` dataset, apply one of the two methods discussed in Q1.2a to estimate the true number of clusters. Was the estimate correct? Discuss your result.

Q1.2b Answer

In [ ]:
```
# Please enter any relevant code for Q1.2b below
# You can use multiple cells if it helps
```

Q1.3 (5 marks)

For this question, we will use the `online_retail_full.csv` dataset, which is a real-world dataset of transactions for an online retail store. Full information about the dataset can

be found here. Here, we do not have true labels, and need to explore the data instead. This is a common scenario in practice, and will require you explore the data and use clustering (likely requiring multiple iterations and tweaks) to try to find patterns.

We're going to investigate whether there are groups of customers, how they are similar, and what they may represent. For simplicity, we will start by using KMeans as our model, and we'll remove some of the columns from our input data. Use a range of K values and whichever techniques in Q1.2 are useful to propose interesting K value(s). Comment on the clusters that are produced in terms of the context of the data.

Hints:

- As this dataset has no truth, there is a lot of scope in this question - remember to have some justification for why you have taken the steps you have.
- The quality of your final clusters is not important for marks, as long as you have taken reasonable steps.
- The overall aim is to try to find patterns in the data. KMeans is suggested as a starting point, but it is not always the best algorithm to use as we have seen in previous questions.
- You can create features from the existing ones. For example, the quantity and price can be multiplied to get a total amount (thus simplifying the data). Other features may require transformation before they can be used.

Q1.3 Answer

In [ ]:
```
# Load the data (the path may be different for you, adjust if needed)
retail = pd.read_csv("online_retail_full.csv", index_col=False)
```

Question 2: Itemset Rule Mining (4 marks)

For this question, we will be using a real-world dataset which gives the votes of 435 U.S. congressmen on 16 key issues gathered in the mid-1980s, and also includes their party affiliation as a binary attribute. This is a purely nominal dataset with some missing values (corresponding to abstentions). It is normally treated as a classification problem, the task being to predict party affiliation based on voting patterns. However, association-rule mining can also be applied to this data to seek interesting associations.

We will be using Weka, both for its utility for itemset rule mining, and to use a different approach for exploring data. You should have some experience using Weka from the first (non-assessed) week.

You may need to take screenshots of Weka and include them in your answer below, or copy & paste the relevant rules. Please ensure that your answer and rules are clearly legible.

Q2.1 (3 marks)

In Weka, run `Apriori` on this dataset using default settings. Comment on any patterns you see in the rules that are generated. Also discuss their support, confidence, and lift,

demonstrating that you understand how they are calculated, their role, and how to interpret these values.

Q2.1 Answer

Q2.2 (1 mark)

It is interesting to see that none of the rules in the default output involve `class = republican`. Why do you think that is?

Q2.2 Answer