THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MODULE 8 UNIT 2
Activity submission

Learning outcome:

LO6: Practise fitting a clustering model on a data set in R.

Plagiarism declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. This assignment is my own work.

3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work.

4. I acknowledge that copying someone else's assignment (or part of it) is wrong and declare that my assignments are my own work.

Name: Vasileios Tsoumpris

1. Instructions and guidelines (Read carefully)

Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M8 U2 Activity Submission** – **e.g. Lilly Smith M8 U2 Activity Submission. NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, "Start writing here" under each question. Please type your answer there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

PLEASE NOTE: Plagiarism cases will be investigated in line with the Terms and Conditions for Students.

IMPORTANT NOTICE: Please ensure that you have checked your course calendar for the due date for this assignment.

Guidelines

1. There are seven pages and three questions in this assignment.

2. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully but concisely and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

3. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

4. At the end of your assignment, please provide feedback on areas where you require further assistance or would like the Assessor to expand on.

2. Mark allocation

Each question receives a mark allocation. However, you will only receive a final percentage mark and will not be given individual marks for each question. The mark allocation is there to show you the weighting and length of each question

Question 1	18
Question 2	18
Question 3	12
TOTAL	**48**

3. Questions

After completing the steps in the assessment IDE activity, you have conducted appropriate data exploration, and applied *k*-means and hierarchical clustering to the wine quality data set in R. With specific reference to what you’ve learnt in the assessment IDE activity, answer the questions that follow.

Question 1

Consider the data exploration done in Steps 1 to 8, and answer the following questions.

1.1 What inferences can be made about the variables that influence the wine quality?

1.2 What predictions can be made about the differences in variables between white and red wines?

(Max. 300 words)

Start writing here:

The reporting graphs through steps 1 to 8 depict some remarkable insights between wine quality and the other features. In particular, alcohol has one of the strongest positive correlations to quality (0.381). Furthermore, the pH is the most neutral variable in terms of correlation with the quality variable, while density (corr=-0,232), volatile acidity (corr=-0,229) and chlorides (-0.172) are the most negatively correlated variables with it. Other variables are somewhat neutral, as their values are around zero. From the above observations, there is a profound inference that the better the quality of the wine, the more the proportion of alcohol tends to be. Lastly, the density and the volatile acidity show a downtrend when the quality of the wine increases. To validate the above inferences, the scatterplots of variables against the quality also portray the same trends.

In addition, there are many differences between white and red wines. For example, the density graph demonstrates that red and white wines have different distributions of total sulphur dioxide, followed, for instance, by chlorides and volatile acidity. Also, in the density plot of citric acid, the white wines have a more significant frequency on the moderate citric acid values. On the contrary, red wines do not reach this frequency peak but have more observations in the low and high range of citric acid values than white wines. Furthermore, examining the scatterplot, it is easily noticed that most of the wines are around 5, 6, and 7 quality values.

Additionally, although red wines display lower values of total sulphur dioxide than white wines, their fixed and volatile acidity, chlorides, and sulphates are generally higher than white wines' values. Another notable insight is that red wines' volatile acidity shows a downtrend as the quality increases while the white wines' values remain at the same levels throughout all qualities.

Question 2

Refer to the *k*-means clustering execution in Steps 9 to 16. Assume that you are presented with a description of a new wine from a supplier that sets out the physicochemical properties of the wine. The goal is to determine the quality of the wine based on the given properties.

With due consideration of the scatterplot generated in Step 15 and the table generated in Step 16, what inferences regarding the quality of the newly launched wine can you make given the type of wine, the alcohol level (***alcohol***) of the wine, and the total sulphur dioxide (***total.sulfur.dioxide***) levels of that particular wine?

(Max. 300 words)

Start writing here:

Through steps nine to sixteen, a k-means model was created to categorise the wines into different clusters. Every category has different characteristics regarding the colour of wine, alcohol and total sulphur dioxide levels. From the three groups created, it is notable that most of the red wines are in the green (2nd) cluster, while the majority of white wines are in both the red (1st) and blue (3rd) clusters. To infer the group of the new wine, one of the parameters that would primarily affect it, is its colour. Of course, there are some exceptions, so it is necessary to investigate each cluster's properties further.

Exploring the table in step 16 and the scatterplot above, it is noticeable that one characteristic of the green cluster apart from the red colour is the low and medium levels of alcohol. Furthermore, the green group consists of wines with generally low levels of total sulphur dioxide. Therefore, the above variable is one of the most vital indicators for grouping the wines as its difference between the mean of clusters is significant (according to the table). Moreover, if the total sulphur of white wine is at medium levels while its alcohol is in a medium to high range, the wine probably belongs to the blue cluster. In other words, the most robust characteristics of the third cluster are the high alcohol values, the white colour and then a medium range of values in total sulphur dioxide. On the contrary, the first cluster consists of white wines with high total sulphur dioxide values and low alcohol levels.

Concluding, there is a blending area where wines from every cluster are observed. In this case, we should further consult the table about the most significant differences between clusters' mean in every variable (i.e. residual.sugar, chlorides etc.) to determine other strong characteristics.

Question 3

Consider the execution of hierarchical clustering performed in Steps 17 to 19. How do the clusters generated by hierarchical clustering in Step 19 compare to the three clusters found by *k*-means clustering in Step 15? What properties do the clusters share, and how do they differ?

(Max. 200 words)

Start writing here:

The way the hierarchical clustering performed the grouping of wines differs in some points from the k-means model. Firstly, hierarchical clustering is more rigid when placing observations into clusters. On the other hand, in k-means clustering, the observation placement can be moved to the most appropriate group. As a result, in hierarchical clustering, the boundaries of every cluster are more profound than in k-means.

In hierarchical clustering, the primary variable that groupings are based on is the total sulphur dioxide (and colour), followed by alcohol. For example, red wines with low total sulphur dioxide levels and medium to high alcohol levels are contained in one cluster. Another cluster is characterised by medium levels of total sulphur dioxide and low to high alcohol levels. Finally, the third group of the hierarchical clustering model is circumscribed by white wines with high values of total sulphur dioxide and low levels of alcohol.

On the other hand, in k-means clustering, all groups share a common characteristic, the medium level values of total sulphur dioxide. Apart from that, the way the groups are clustered in terms of alcohol (and wine colour) is kind of similar to the hierarchical clustering model.

4. Rubric

The following rubric will be used to grade your submission for this activity submission.

	Unsatisfactory	Limited	Accomplished	Exceptional
Question 1: Adherence to the brief *The answer provides a brief analysis of the output generated in the IDE notebook.* *The answer is substantiated based on the output of the IDE notebook, with specific reference to the data exploration section.*	No submission. OR The answer fails to adhere to any of the elements contained in the brief. (0)	The answer adheres to some elements contained in the brief, but some key elements are missing. The answer does not fall within the prescribed word count (50 words over the word count). (2)	The answer adheres to most, but not all, elements of the brief, and falls within the prescribed word count. Almost all information is provided and relevant. (4)	The answer adheres to all the elements of the brief and falls within the prescribed word count. All information provided is comprehensive and relevant. (6)
Question 1: Understanding and accurate use of module content *The answer demonstrates that the student engaged with the content.* *The answer demonstrates that the student has an informed grasp of how to interpret the output generated from the data exploration steps.*	The answer demonstrates that the student did not engage with the content. OR The answer fails to demonstrate a basic understanding of the content. (0)	The answer demonstrates that the student attempted to engage with the content. The answer demonstrates an inadequate understanding of the content. (2)	The answer demonstrates that the student engaged with the content and understands most of it. (4)	The answer demonstrates that the student engaged with the content and has an excellent understanding of the module's content. (6)

Question 1: Coherence and clarity *The answer is clearly structured and written in a way that is comprehensible.*	Answer is incoherent or lacks clarity. Answer is not logically structured, or it is incomprehensible. (0)	Answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. (2)	Answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. (4)	Answer is extremely well structured and written with exceptional clarity and coherence. (6)
Question 2: Adherence to the brief *The answer provides a brief analysis of the output generated in the IDE notebook.* *The answer is substantiated based on the output of the IDE notebook, with specific reference to the k-means clustering section.*	No submission. OR The answer fails to adhere to any of the elements contained in the brief. (0)	The answer adheres to some elements contained in the brief, but some key elements are missing. The answer does not fall within the prescribed word count (50 words over the word count). (2)	The answer adheres to most, but not all, elements of the brief, and falls within the prescribed word count. Almost all information is provided and relevant. (4)	The answer adheres to all the elements of the brief and falls within the prescribed word count. All information provided is comprehensive and relevant. (6)
Question 2: Understanding and accurate use of module content *The answer demonstrates that the student engaged with the content.* *The answer demonstrates that the student has an informed grasp of how to interpret the output generated from k-means clustering, and has the ability to use the output to interpret the clusters.*	The answer demonstrates that the student did not engage with the content. OR The answer fails to demonstrate a basic understanding of the content. (0)	The answer demonstrates that the student attempted to engage with the content. The answer demonstrates an inadequate understanding of the content. (2)	The answer demonstrates that the student engaged with the content and understands most of it. (4)	The answer demonstrates that the student engaged with the content and has an excellent understanding of the module's content. (6)

Question 2: Coherence and clarity *The answer is clearly structured and written in a way that is comprehensible.*	The answer is incoherent or lacks clarity. Answer is not logically structured, or it is incomprehensible. (0)	The answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. (2)	The answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. (4)	The answer is extremely well structured and written with exceptional clarity and coherence. (6)
Question 3: Adherence to the brief, coherence, and clarity *The answer provides a brief analysis of the output generated in the IDE notebook.* *The answer is substantiated based on the output of the IDE notebook, with specific reference to the hierarchical clustering section.*	No submission. OR The answer fails to adhere to any of the elements contained in the brief. (0)	The answer adheres to some elements contained in the brief, but some key elements are missing. The answer does not fall within the prescribed word count (50 words over the word count). (2)	The answer adheres to most, but not all, elements of the brief, and falls within the prescribed word count. Almost all information is provided and relevant. (4)	The answer adheres to all the elements of the brief and falls within the prescribed word count. All information provided is comprehensive and relevant. (6)
Question 3: Coherence and clarity *The answer is clearly structured and written in a way that is comprehensible.*	The answer is incoherent or lacks clarity. Answer is not logically structured, or it is incomprehensible. (0)	The answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. (2)	The answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. (4)	The answer is extremely well structured and written with exceptional clarity and coherence. (6)

Total: 48 marks

Feedback:

Start writing here: