Clustering and Fitting (30%)

40 Points Possible

12/12/2024

Attempt 1

In Progress

NEXT UP: Submit assignment

Add comment

Unlimited Attempts Allowed

11/18/2024 to 12/17/2024

Details

Weighting %:	30	**Submission deadline (for students):**	12/12/24 at 12pm (Midday)
Authorship:	Individual	**Target date for returning marked coursework:**	10/1/25
Tutor setting the work:	Dr. William Cooper	**Number of hours you are expected to work on this assignment:**	20

This Assignment assesses the following module Learning Outcomes (from Definitive Module Document):

1. Be able to maintain and develop code using the git version control system.
2. Be able to apply different techniques for cleaning data and preparing it for analysis.
3. Be able to design and implement algorithms for clustering, classification and regression problems.
4. Be able to communicate their findings to others, including a critical assessment of performance.
5. Demonstrate knowledge and understanding of the concepts of version control for code development.
6. Demonstrate knowledge and understanding of key data manipulation techniques for data preparation.
7. Understand how to approach a range of different data science problems to obtain an efficient solution.

Assignment Tasks:

You will create a well-written report performing clustering and fitting within a dataset. You can download any dataset from Kaggle/Worldbank/etc. Be sure to include your name, student number and a link to your GitHub repository in the report. There will be at least four plots: a histogram/bar chart/pie chart; a line/scatter graph; a confusion matrix/heatmap/corner/box/violin plot; an elbow/silhouette plot. The code will contain evidence of the creation of any displayed graphs (one graph per function) and the creation of any shown clustering/fitting technique. The minimum expected techniques will be that of k-means clustering and line fitting.

This will build on the statistics and trends assignment into a full report as would be produced by a professional data scientist. However, do **not** use the same report/dataset as previously (this will be checked), as self-plagiarism is still academic misconduct.

Note: You may be asked to attend a viva if there are suspicions of derivative work, i.e., collusion (submitting near-identical work at any quantities); plagiarism (using someone else's work, including from online); or the use of generative AI (e.g. ChatGPT). If you do not attend this viva, the grade for the relevant work will be 0.

Submission Requirements:

A three page PDF report, including a functional link to your GitHub repository containing your python code (either notebooks or plain python). Check that your repository link is both clickable and links to a **public** repository. The plots **must** be in the report to be marked, not within the notebook. Do not waste space in the report with trivial content such as how you downloaded and read a file. The prose should flow as: short introduction to the topic; discussion on what the first plot is showing, using statistics and describing the distributions; then same for second plot then third plot.

Marks awarded for:

See rubric.

Type of Feedback to be given for this assignment:

Written feedback within the rubric.

Additional information:

- Regulations governing assessment offences including Plagiarism and Collusion are available from **https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct.pdf (https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct.pdf)** (UPR AS14) .
- Guidance on avoiding plagiarism can be found here: **https://herts.instructure.com/courses/61421 (https://herts.instructure.com/courses/61421)** (see the Referencing section)
- For postgraduate modules:
  - a score of 50% or above represents a pass mark.
  - late submission of any item of coursework for each day or part thereof (or for hard copy submission only, working day or part thereof) for up to five days after the published deadline, coursework relating to modules at Level 7 submitted late (including deferred coursework, but with the exception of referred coursework), will have the numeric grade reduced by 10 grade points until or unless the numeric grade reaches or is 50. Where the numeric grade awarded for the assessment is less than 50, no lateness penalty will be applied.

View Rubric

Assignment 2: Clustering and Fitting

Criteria	Ratings						Points
Relational Graph Quality view longer description	**2 pts** **Full marks** The graph will convey an xy relation. The axes labels will be fully readable without effort and the relation(s) will be clear.		**1 pts** **Fair quality** The graph will convey an xy relation. The axes labels may be too small to read comfortably. There may be an overcrowding of the figure.		**0 pts** **No marks** Missing graph from report or missing axes labels.		/ 2 pts
Categorical Graph Quality view longer description	**2 pts** **Full marks** The graph will compare multiple categories. The axes labels will be fully readable without effort and the appearance will be clear.		**1 pts** **Fair quality** The graph will compare multiple categories. The axes labels may be too small to read comfortably. There may be an overcrowding of the figure.		**0 pts** **No marks** Missing graph from report or missing axes labels.		/ 2 pts
Statistical Graph Quality view longer description	**2 pts** **Full marks** The graph will communicate a statistical relation. The axes labels will be fully readable without effort and the appearance will be clear.		**1 pts** **Fair quality** The graph will communicate a statistical relation. The axes labels may be too small to read comfortably. There may be an overcrowding of the figure.		**0 pts** **No marks** Missing graph from report or missing axes labels.		/ 2 pts
Quality of Analysis view longer description	**5 pts** **Full marks** The explanation is clear and coherent. Statistics are used to support statements. There is a connecting storyline.	**4 pts** **Very high marks** The explanation is clear and coherent. Statistics are used to support statements. There may be some connecting storyline.	**3 pts** **High marks** The explanation is mostly clear and coherent. There may be some statistics supporting some statements. There may be some storyline.	**2 pts** **Fair marks** The explanation is mostly coherent. There may be a majority of statements without statistical support. The report is more descriptive.	**1 pts** **Poor quality** The report is almost entirely descriptive without meaningful statistics.	**0 pts** **No marks** No description of any merit.	/ 5 pts
Spelling and Grammar view longer description	**1 pts** **Good** The spelling and grammar is acceptable enough to communicate complex ideas.		**0.5 pts** **Acceptable** The spelling and grammar use is acceptable enough to communicate basic ideas.		**0 pts** **No marks** Very poor English, making idea communication challenging.		/ 1 pts
Relational Graph Function view longer description	**1 pts** **Good** Function with docstring which only creates one plot.		**0.5 pts** **Acceptable** Function without docstring or function produces multiple plots.		**0 pts** **No marks** No/not useable GitHub link or no function used to create plot.		/ 1 pts

Assignment 2: Clustering and Fitting

Criteria	Ratings							Points
Categorical Graph Function view longer description	**1 pts** **Good** Function with docstring which only creates one plot.		**0.5 pts** **Acceptable** Function without docstring or function produces multiple plots.		**0 pts** **No marks** No/not useable GitHub link or no function used to create plot.			/ 1 pts
Statistical Graph Function view longer description	**1 pts** **Good** Function with docstring which only creates one plot.		**0.5 pts** **Acceptable** Function without docstring or function produces multiple plots.		**0 pts** **No marks** No/not useable GitHub link or no function used to create plot.			/ 1 pts
Statistical Depth view longer description	**3 pts** **Full marks** All major moments shown (mean/median, standard deviation, skewness, kurtosis). Correlation matrix and basic 'describe' used.	**2 pts** **High marks** First two major moments shown (mean/median, standard deviation). Correlation matrix and basic 'describe' used.	**1 pts** **Fair marks** Correlation matrix and basic 'describe' used.		**0 pts** **No marks** No/not useable GitHub link or no use of 'describe' and correlation matrix.			/ 3 pts
Code Quality view longer description	**2 pts** **Full marks** Code is easy to read and follows the major PEP-8 recommendations: import > functions > variables order; functions separated by exactly two lines (one if in a class) or sole occupier of notebook cell; spaces after commas and around assignment/mathematical operators.		**1 pts** **Fair marks** Code is mostly easy to read and may have a few slips from the major PEP-8 recommendations: import > functions > variables order; functions separated by exactly two lines (one if in a class) or sole occupier of notebook cell; spaces after commas and around assignment/mathematical operators.		**0 pts** **No marks** No/not useable GitHub link or code is difficult to read with many divergences from the major PEP-8 recommendations: import > functions > variables order; functions separated by exactly two lines (one if in a class) or sole occupier of notebook cell; spaces after commas and around assignment/mathematical operators.			/ 2 pts
Clustering Function view longer description	**1 pts** **Good** Function with docstring which does not create a plot.		**0.5 pts** **Acceptable** Function without docstring or function also creates plots.		**0 pts** **No marks** No/not useable GitHub link or no function used to perform clustering.			/ 1 pts
Fitting Function view longer description	**1 pts** **Good** Function with docstring which does not create a plot.		**0.5 pts** **Acceptable** Function without docstring or function also creates plots.		**0 pts** **No marks** No/not useable GitHub link or no function used to perform fitting.			/ 1 pts
Clustering Quality view longer description	**6 pts** **Full marks** The clusters will appear well grouped. The data will have been normalised and back scaled to present. Clear use of silhouette score/elbow method to select cluster amount. The graph will have coloured groups and labelled cluster centres in the legend. The	**5 pts** **Very high marks** The clusters will appear well grouped. The data will have been normalised. Clear use of silhouette score/elbow method to select cluster amount. The graph will have coloured groups and labelled cluster centres in the legend. The data will be	**4 pts** **High marks** The clusters will appear well grouped. The data may have been normalised. Use of silhouette score/elbow method to select cluster amount. The graph will have coloured groups and labelled cluster centres in the legend. The data will be appropriate for clustering.	**3 pts** **Fair marks** The clusters will appear well grouped. The data may have been normalised. The graph will have coloured groups and labelled cluster centres in the legend. The data will be appropriate for clustering.	**2 pts** **Poor quality** The clusters may appear well grouped. The data may have been normalised. The data will be appropriate for clustering.	**1 pts** **Very poor quality** The clusters are not well grouped. The data may be appropriate for clustering.	**0 pts** **No marks** The clusters are not well grouped. The data is not appropriate for clustering, or no graph in report.	/ 6 pts

Assignment 2: Clustering and Fitting

Criteria	Ratings						Points
	data will be appropriate for clustering.	appropriate for clustering.					
Fitting Quality view longer description	**5 pts** **Full marks** The data will be well fitted and suitable for fitting. The plot will include a good confidence interval and errorbars.	**4 pts** **High marks** The data will be well fitted and suitable for fitting. The plot will include a good confidence interval or errorbars.	**3 pts** **Fair marks** The data will be well fitted and suitable for fitting. The plot may include reasonable errorbars.	**2 pts** **Poor quality** The data will be fitted and suitable for fitting.	**1 pts** **Very poor quality** The data will be poorly fitted but suitable for fitting.	**0 pts** **No marks** The data is not suitable for fitting, or no graph in report.	/ 5 pts
Clustering Prediction view longer description	**4 pts** **Full marks** Several predicted points will be attached to appropriate groups, and clearly labelled and coloured.	**3 pts** **High marks** Several predicted points will be attached to appropriate groups.	**2 pts** **Fair marks** Predictions will be made for different group memberships.	**1 pts** **Poor quality** An attempt at predictions on unseen data will have been made for different group memberships.	**0 pts** **No marks** No predictions made or no/not useable GitHub link.		/ 4 pts
Fitting Prediction view longer description	**3 pts** **Full marks** Several predictions with good, associated uncertainties are shown.	**2 pts** **Fair marks** Several predictions are given.	**1 pts** **Poor quality** An attempt at predictions on unseen data are made.	**0 pts** **No marks** No predictions made or no/not useable GitHub link.			/ 3 pts
Submission Guidelines view longer description	**0 pts** **Expected** The report is at the required length with no overly small text or minimised margins.		**-4 pts** **Not expected** The report is not at the required length, either overrunning or too short (by at least a third of a page). Alternatively, there may be overly small text or minimised margins.				/ 0 pts
							Total points: 0

Choose a submission type

Upload

More

Choose a file to upload

File permitted: PDF

or

Canvas Files

I agree to the tool's **End-User License Agreement** (https://api.turnitinuk.com/api/lti/1p0/user/static_eula)
This assignment submission is my own, original work

Submit assignment