COURSEWORK SPECIFICATION
ECMM443 – Introduction to Data Science
Module Leader: Xiaoyang Wang
Academic Year: 2024/25

Title: Investigating Heart Rate Variability

Submission deadline: 20 November 2024 midday 12pm

This assessment contributes **20%** of the total module mark and assesses the following **intended learning outcomes**:

Module Specific Skills and Knowledge

1. Design a data science pipeline for a given problem in a chosen domain.
2. Investigate a dataset using mathematical and visualisation tools.

.

Discipline Specific Skills and Knowledge

3. Apply principles of a statistical pattern recognition to a given problem.
4. Articulate a decision problem to be solved with data science affecting business or society.

Personal and Key Transferable / Employment Skills and Knowledge

5. Critically read and report on research papers
6. Present the results of a piece of data science work in the form of a report

This is an individual assessment and you are reminded of the University's regulations on collaboration and plagiarism. You must avoid plagiarism, collusion, and any academic misconduct behaviours. Further details about academic honesty and plagiarism can be found at https://ele.exeter.ac.uk/course/view.php?id=1957.

GenAI in assessments:

This assessment is categorised as **AI-prohibited**. By submitting the course work, you declare that ***you have not used generative AI in preparing your work.***

GenAI statement

Use of GenAI tools in Coursework - Investigating Heart Rate Variability for ECMM443 Introduction to Data Science.

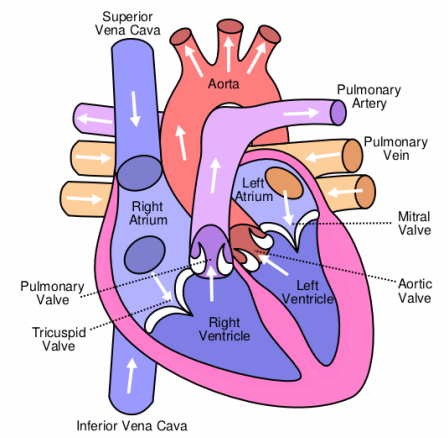
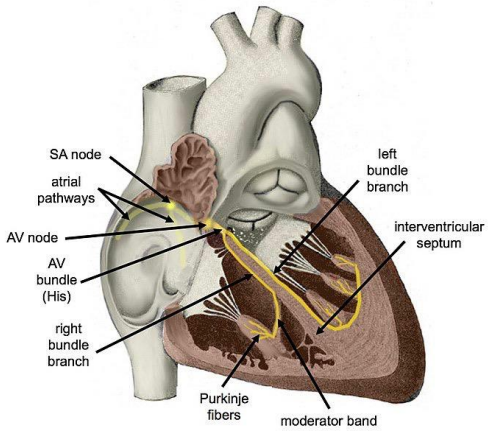
The University of Exeter is committed to the ethical and responsible use of Generative AI (GenAI) tools in teaching and learning, in line with our academic integrity policies where the direct copying of AI-generated content is included under plagiarism, misrepresentation and contract cheating under definitions and offences in TQA Manual Chapter 12.3. To support students in their use of GenAI tools as part of their assessments, we have developed a category tool that enables staff to identify where use of Gen AI is integrated, supported or prohibited in each assessment. This assessment falls under the category of ***AI-prohibited****. This is because the use of GenAI in assisting problem solving, generating code for data analysis and interpret results are prohibited in this assignment.*

You can find further guidance on using GenAI critically, and how to use GenAI to enhance your learning, on Study Zone digital.

ECMM443 cw: Investigating Heart Rate Variability

Background Information

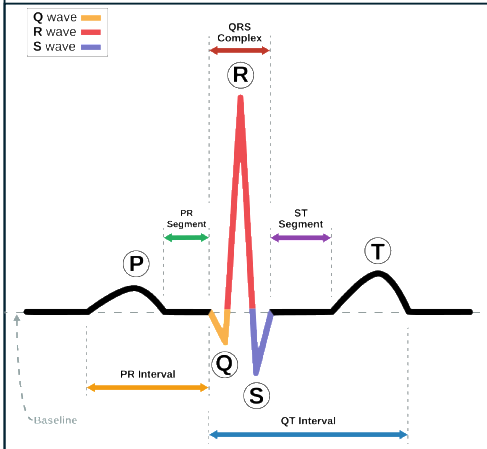
The human heart consists of a series of chambers surrounded by muscles that contract and relax rhythmically to pump blood around the human body.

Structure and chambers of the human heart	Nervous system (yellow) of the human heart
	

The behaviour and health of the heart can be studied using electro-cardiogram or ECG recordings, in which electrodes capture the electrical activity associated with the beating heart.

The electrical impulses to trigger heartbeats are generated by specialised pacemaker cells in the cardiac muscle. Normal or “sinus” rhythm heartbeats are triggered when pacemaker cells located in the sinoatrial node depolarise causing an impulse that propagates as a wave, triggering the pacemaker cells of the other regions in a coordinated sequence, such that a characteristic waveform:

ECG waveform sequence: Normal Sinus Beat



- **P-wave** generated by with atrial depolarization
- **QRS-wave-complex** generated by venticular depolarization
- **T-wave** generated by ventricular repolarization

However it is also possible for pacemaker cells of the atrial or ventricular regions to spontaneously depolarise without receiving a signal from the sinoatrial node. This triggers

what is known as a premature or ectopic heart beat. These are classified by where the premature depolarisation occurs:

- PAC Premature atrial contraction
- PVC Premature ventricular contraction

Such contractions may occur at a low rate in healthy individuals, but may be more frequently occurring or occur with characteristic repetitions in individuals with heart conditions.

Analysis of ECG recordings and Heart Rate Variability

To assess heart health a patient is asked to wear an ECG monitor that records their hearts electrical activity over an extended period. The recording is then annotated to:

- identify each beat by referencing the location of the R wave of the QRS complex.
- labels the type of each beat using N for normal sinus rhythm or another character to signify the type of non-normal type.

The resulting ECG can then analysed by an expert to look at, so they can try to diagnose the patient's heart health.

In addition to a manual inspection, several metrics can be automatically calculated from the annotated ECG recording. The most basic metric is the average interval beat interval (e.g. in **ms**) which may also be written as average beats-per-minute (BPM) value.

$$\text{mean BPM} = \frac{1000}{\text{mean beat interval in } \mathbf{ms}} \times 60$$

When calculated from the full set of annotated beats, this is known as the **average R-R interval**. However to better characterise the interval of the heart in its normal sinus rhythm, another reported metric might be the **average N-N interval**, which is found by filtering the recording so that only the intervals between consecutive N-type beats are included.

The rate at which hearts beat reflects the needs of the circulatory system to supply the body with oxygenated blood, such that the heart rate will be higher during physical exercise and lower during rest. However, in addition to this type of variability which reflects the changing needs of the body, the heart rate is also observed to vary beat-by-beat. Studies have shown that poor heart health is associated with decreased beat-to-beat variability (where beats occur in a regular frequency like a metronome).

Therefore there has been considerable effort to come up with ways to measure the variability of the heart and explore more fully how it varies with characteristics such as gender, age and heart health.

In this report we will study the following metrics which can be calculated from an annotated ECG recording. These include calculations based on the difference in beat-to-beat intervals, for example, if a beat of 800ms duration is followed by a beat of 870ms duration, then there is a beat-to-beat variation of +70ms.

HRV metric	Description
Mean NN	the mean duration of all N-N intervals in the recording
Mean BPM	the mean number of beats per minute calculated using the duration of all N-N intervals in the recording
SDNN (ms)	the standard deviation of the set of all N-N intervals
RMSSD (ms)	the square root of the mean squared value of the beat-to-beat difference for all sequential N-type beats in the recording.
pNN20 (%)	the proportion of beat-to-beat interval differences, where the absolute difference is greater than 20ms for all sequential N-type beats in the recording.
pNN50 (%)	the proportion of beat-to-beat interval differences, where the absolute difference is greater than 50ms for all sequential N-type beats in the recording.

Investigation Task

The aim of this investigation is to reproduce an analysis reported in an early paper that considered heart rate variability between groups of young and old individuals.

https://journals.physiology.org/doi/abs/10.1152/ajpregu.1996.271.4.R1078

It involves a set of 20 ECG recordings from healthy individuals (10 young, 10 old and split evenly between male and female participants). Each recording is approximately 120 minutes in length and was taken while the individual watched the Disney film Fantasia in a supine position (laying face up). The files we will work with store the ECG annotation information that records the timings and type of every identified heartbeat.

You are required to visualise the data, calculate HRV metrics, and assess the evidence that there exists a difference in the HRV metrics between the young and old age groups).

Marks below indicate the relevant weighting of the tasks in the grade calculation.

There are 60 marks in total for all required tasks.

Task 1 (10 marks)

In file **hrv.py** define a Python function **calculate_HRV_metrics()** which takes:

- argument **file_in** storing a file location (which may include path).

Your code should: load in the data from that file; count the number **n** of NN intervals it stores; and calculate the following metrics as defined above:

average NN, average BPM, SDNN, RMSSD, pNN50, pNN20

It should return a dictionary of the calculated metrics in the following format:

```
{'filename': 'y01.csv',
 'n': 7293,
 'mean_nn': 746,
 'mean_bpm': 80.4,
 'sdnn': 20.2,
 'rmssd': 26.6,
 'pnn20': 50.0,
 'pnn50': 8.33}
```

Note:

- **mean_nn** should be rounded to the nearest integer
- **mean_bpm sdnn**, **rmssd**, **pnn20**, **pnn50** should be rounded to 1 decimal place.

To access the highest grades (see markscheme) your code should handle the following issues as described below:

missing data file: you should let the program raise an exception as usual if there is an issue opening the named file

insufficient data: when there are fewer than 500 valid data points for the calculated metrics, the dictionary entry should be filled with a **None** value.

Suggested approach:

Load in the file into a PANDAS dataframe and write code to add the following columns:

time_next	stores the time of the next beat (one row below in the **time** column)
type_next	stores the type of the next beat (one row below in the **type** column)
rr	interval between the current and next beat using **time** and **time_next** columns
rr_type	combines the **type** and **type_next** columns to form a two character string signifying the two types of beats
rr_next	stores the period of the next rr interval (one row below in the **rr** column)
diff	stores the difference between the current interval and the next one
diff_squared	stores the squared value of the **diff** column
rr_next_type	stores the type of the next interval (one row below in the **type** column)

You can then use these columns to calculate the metrics as defined above. Be careful to remember that to calculate the metrics **mean_nn**, **mean_bpm**, and **sdnn** we use only beat intervals of **NN** type.

Task 2 (5 marks)

In file **hrv.py** add a function **process_HRV_files()** that can takes two arguments:

- **file_list_in** a list of input filenames to process (may include paths)
- **file_out** the name (and path) specifying where the resulting analysis output should be stored.

It should load and calculate the HRV metrics for each input file, and save this in **csv** format in file **file_out** storing the results with columns:

filename, mean_nn, mean_bpm, sdnn, rmssd, pnn20, pnn50

To access the highest grades (see markscheme) your code should handle the following issues:

***missing data file*:** where a datafile does not exist, the code handle the exception should log/print an appropriate message about the missing file, and omit it from the results, but should continue to process the other valid files in **file_list_in**

Task 3 (5 marks)

In a code notebook **HRV_analysis.ipynb** write code to load and process the data in the provided dataset creating file **fantasia.csv** storing the results.

Add code that loads **fantasia.csv** so that the loaded data is in a dataframe named **hrv_df**

Add code to merges the information on age and gender of the records that can be found in **fantasia_individuals.csv** to the **hrv_df** dataframe.

Display the **hrv_df** dataframe in full in your notebook.

Task 4 (10 marks)

Poincare plots visualise a heartbeat record by plotting the duration of each heartbeat interval (x-axis) against the duration of the subsequent heartbeat (y-axis) as a scatter plot e.g. if a beat of 800ms duration is followed by a beat of 870ms duration, then it would be plotted as a point at coordinates (800, 870).

In your notebook **HRV_analysis.ipynb** write code to make a multiplot figure (four rows and five columns) multiplot figure showing the Poincare plots for the records, arranging the plot order and formatting so that any differences between the groupings can be compared.

Ensure that your figure has suitable labels, and that the axes range for all plots is matched to allow easy comparison.

Task 5 (10 marks)

We want to reproduce the analysis that considers the statistical evidence that heart rate (as indicated by mean bpm) and heart rate variability (as indicated by metric RMSSD) differ by age group.

In your notebook write code that applies a t-test to assess the statistical difference in these variables between the two groups.

Add a code section that creates figures to illustrate this analysis, along with a short paragraph setting out the analysis results.

Task 6

Add to your notebook **HRV_analysis.ipynb** answers to the following questions (use markdown boxes for written answers):

6a Write a short discussion that discusses the key observations that can be made from the Poincare plots, in terms of how they relate to the statistical analysis performed.

(5 marks)

6b In this investigation we have applied the t-test to detect statistical differences. Critically evaluate the applicability of the test and suggest what additional or alternative approaches would be suitable (you may demonstrate if your wish).

(5 marks)

6c What additional plots may be useful to view each recording before including it in the analysis. You may include an example plot/plots and explain why this would be useful.

(5 marks)

6d Explore the current research in this area and briefly set out the current understanding of heart rate and heart rate variability by age and gender.

(5 marks)

Submission Information

Please upload to ELE:

- the finished **hrv.py** code file
- the **HRV_analysis.ipynb** notebook
- your GenAI declaration form

via the link in the assessment section of the ELE ECMM443 module page.

Please ensure your notebook is clearly organised so that the relevant tasks and questions are clearly labelled.

A penalty of up to 20% may be applied at the discretion of the marker, if they have difficulty reviewing your submission e.g.: because the code file/notebooks are disorganised; lack sufficient commenting; missing labels; and/or are poorly formatted.

Marking criteria

For tasks involving writing code for data processing / analysis

Strong 1st class (80%+): functions / code work as defined in the brief; code deals with exceptions and issues (missing data, missing files) appropriately.

1st class (70-79%): functions / code work as defined in the brief with no-consequential issues in terms of functionality; however minor improvements/edits needed before work would be considered suitable for publishing (e.g. how it deals with exceptions and issues)

2:1 (60-69%): functions / code works as defined in the brief but have minor aspects in which improvements can be made (e.g. minor bug or issue with data processing/analysis; ways in which exceptions/issues are handled)

2:2 (50-59%): functions / code works as defined in the brief but have clear aspects in which improvements are possible (e.g. several minor bugs or issues with data processing/analysis and/or clear possible improvements to exception handling)

40-49%: functions / code demonstrates sufficient student ability, but have major issues required to be able to be used in practise, e.g. in terms of bugs / issues with data processing/analysis and/or clear problems with organisation / documentation / exception handling).

30-39%: Student demonstrates relevant work towards the requirements but insufficient to pass.

0-29%: Insufficient evidence of relevant work completed

For tasks involving data visualisation

Strong 1st class (80%+): figure is publication quality

1st class (70-79%): figure is high quality with only minor improvements/adjustments needed for publication (e.g. clarity of text, additional annotations, dpi)

2:1 (60-69%): figure is clear and fully labelled but some improvements/adjustments possible to improve clarity are present (optimisation of colour scheme, text size, axes)

2:2 (50-59%): figure is clear and fully labelled but there are clear/obvious improvements possible to improve (e.g. some aspect of formatting actively interferes with the ability of the reader to view the figure)

40-49%: result demonstrates sufficient student work in relation to expected output but contains major issues or many clear/obvious improvements to be made

30-39%: Student demonstrates relevant work towards the requirements but insufficient to pass.

0-29%: Insufficient evidence of relevant work completed

For tasks involving discussion and evaluation:

Strong 1st class (80%+): answer demonstrates deep and critical understanding of the topic going beyond the expected level of understanding, and with a standard consistent with publication quality scientific writing. Where appropriate referencing demonstrates full engagement with the most relevant research in the field discussed.

1st class (70-79%): answer is well written and demonstrates deep and critical understanding of the topic. Where appropriate referencing included that demonstrates good engagement with relevant related research.

2:1 (60-69%): answer is well written but contains minor issue/error/misconception/omission such that is not of 1st class standard. Where appropriate referencing included that demonstrates some engagement with relevant research context.

2:2 (50-59%): written work is of reasonable quality answers the question but contains several minor or single major issue/error/misconception/omissions(s). Where appropriate referencing included that demonstrates some engagement with research, but may be missing references in places.

40-49%: result demonstrates sufficient relevant knowledge or understanding to answer the question, but with several clear/major issue/error/misconception/omissions(s) or is written in such a way that it is below the expected standard (many typos or grammatical errors). May be missing many key references.

30-39%: Student demonstrates relevant work towards the requirements but insufficient to pass.

0-29%: Insufficient evidence of relevant work completed.