Data Mining 1 Assignment 1

Apply EDA procedures to datasets met in class

Dr Bernard Butler, SETU

4th October, 2023

Issue Date: Wednesday 4th Ocober 2023

Submission Date: Thursday 26th October 2023 at 21:00

1 Submission Instructions

- Students should submit one (or more) Jupyter notebooks, together with any supporting resources needed to run the analysis, so that they can be ``run'' by the lecturer.
- If a student uses any additional python packages, please indicate where they can be found too.
- The submission should also include the outputs of running the notebook.
- All files in the submission should be added to a zip archive, with the following naming convention: *StudentID-JohnDoe*-CA1.zip, replacing *StudentId* and *JohnDoe* as needed.
- You should submit all your work electronically through moodle by Thursday 26th October 2023 at 21:00.
- 10 percent will be deducted for each day you are late submitting without good cause and prior agreement with your lecturer.
- Queries about this CA should be made via the Slack Channel.

2 Aim

The purpose of this assessment is to give students the opportunity to show what they have learned in the module relating to Exploratory Data Analysis (EDA). Each student is expected to apply the EDA procedure described in Weeks 4 and 5 to two of the datasets, chosen from Server Tips, Titanic passenger survival, and Algae Blooms. Students will also be able to draw upon what they have learnt in other weeks to aid their interpretation of the data.

3 Submission mode and Deadline

Assignment attempts should be uploaded via the relevant moodle submission page. The deadline for submission is **Thursday 27 October 2021 at 21:00**.

4 Specification

A detailed specification has been prepared in the format of a python notebook that should be used as a template for your answer. Essentially - you are asked to "fill the blanks".

For more details, please refer to the attached CA1spec.ipynb notebook.

5 Rubric

This assessment is worth 20% of your overall marks.

id	Criterion	Marks
1	First pass - load data set and initial clean	25
2	Second pass - individual features and target	25
3	Third pass - relationships between features and target	20
5	Derive Insights and show initiative	30

6 General Guidance

Students are reminded to use the markdown cells to explain their reasoning, discuss outputs and report on the progress of their data investigation. Students are also advised to use both tabular and plotted output to present their work. Each notebook should read like a technical report, with all findings supported by analysis of the data.

The key to success in this assessment is good data preparation and the ability to make, present and defend hypotheses based on what you see in the data.