Queen Mary
University of London

MTH781P: Data Analytics **Autumn 2021**
Assessed Coursework **A. Baule, J. Griffin**

The subject of this coursework is the investigation of a historical dataset containing data on housing in the Boston area[1]. A summary of the variables of interest is given in the following table:

CRIM	Per capita crime rate by town
CR01	Crime rate dummy variable (=1 if above median; 0 otherwise)
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
LSTAT	Percentage of lower status of the population
MEDV	Median value of owner-occupied homes in $1000s

Table 1: Variables (columns) in the housing dataset.

The dataset is available on QMplus as *housing.csv*. Use R to analyze the dataset and address the following tasks:

1. Explore the data. Plot and produce summary statistics to identify the key characteristics of the data and produce a report of your findings. We would expect between 5 and 10 tables or figures accompanied by a description of your main findings. Interpret your statistical observations in the business context of the dataset. (40 marks)

2. Develop a regression model to predict **MEDV** from one or more of the other variables. Discuss your methodology including, for example, variable selection, goodness of fit, performance. Consider both linear and nonlinear models. Produce a report of your findings supported by plots and statistical analysis. (35 marks)

3. Develop a classification model to predict whether a neighbourhood has high (**CR01**=1) or low (**CR01**=0) per capita crime rate. Explore various subsets of predictors and discuss the performance of your model. (15 marks)

Additional marks will be given for the overall presentation of the coursework, the quality of figures and writing. (10 marks)

[1]The original dataset is available as part of the scikit-learn package and has been slightly modified for this assessment.

Guidelines for the coursework submission:

- The submission deadline is **23:59 on Monday 20st December 2021**.
- Please submit the report by the indicated deadline. For **late submissions**, the QMUL late-submission policy applies (see Section 3.48 of the QMUL Academic Regulations):

  "For every period of 24 hours, or part thereof, that an assignment is overdue there will be a deduction of five per cent of the total marks available (eg five marks for an assessment marked out of 100). After seven calendar days (168 hours or more late) the mark will be reduced to zero and recorded as 0FL (zero, fail, late)."
- The submission should be a single document in .pdf format containing any R-code used for the analysis. The **page limit** for the document is **12 pages including R-code and any appendices** (if applicable). Minimum fontsize is 11.
- Every student is assigned a group as allocated in a separate document. Students have been allocated randomly. The group should work together and produce a **single report**. Only one report can be submitted by each group on QMplus.
- If one or more students within a group are unable to complete the work and need to submit **extenuating circumstances** (EC) claims, the standard procedure for EC submissions applies. Any extensions to the submission deadline due to EC claims apply to the whole group.
- If a group is unable to work together, please get in touch with the module organizers to discuss this as soon as possible.
- Every group report will be checked for **plagiarism** using an automated system. Please refer to the QMUL Academic Regulations for more information about the definition of plagiarism and the related penalties.
- Plagiarism penalties will be applied to every member in a group.
- Every submission should contain the **declaration** below, signed by all group members. Please upload a scanned copy.

Declaration

We hereby declare that this project is our own work, written by ourselves in our own words. If quotations are included from published or unpublished sources, these are clearly indicated and acknowledged as such. We also declare that the distribution of workload, as indicated below, has been agreed among all members of the group.

Group number:

---

1. Student name:

Student number:

Responsible for the following content:

Signature:

---

2. Student name:

Student number:

Responsible for the following content:

Signature:

---

3. Student name:

Student number:

Responsible for the following content:

Signature: