**Report 2**

**Word Count 2000**

**Due Date 9/09/2022**

Report 2 must relate to following course learning objectives 2, 3 and 5:

2. analyse and apply strategies processes and underlying technologies for effective management of data to make evidence-based decisions.

3. critically analyse organisational and societal problems using descriptive and predictive analysis and internal and external data sources to generate insight, create value and support evidence-based decision making;

5. communicate effectively in a clear and concise manner in written report style for both senior and middle management with correct and appropriate acknowledgment of the main ideas presented and discussed.

**Task 1 Data Quality (45 Marks)**

1.1 Define the concept of data quality and discuss the key dimensions of data quality **(30**

**marks 1000 words)**

1.2 Explain why data quality is so important for effective predictive analytics in an

organisation, drawing on a real world example **(15 marks 500 words)**

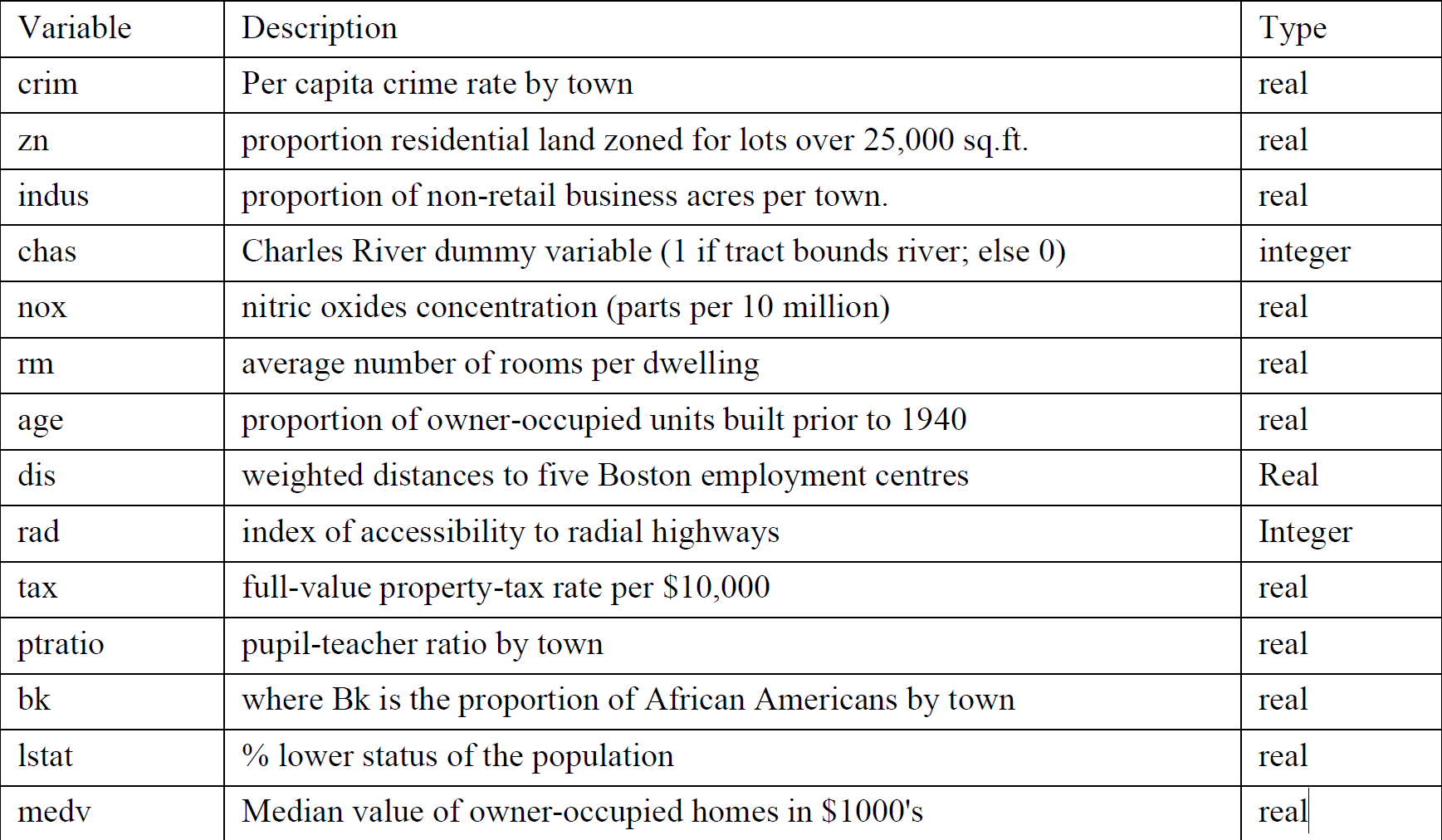
**Task 2 Exploratory Data Analysis and Linear Regression Analysis (45 Marks)**

Carefully study the Data Dictionary for Boston Housing Data Set (Table 1) and

accompanying description of each variable. It is important to understand this data set used

for Task 2. Each record in the *housing.csv* data set describes a Boston suburb or town.

Data was drawn from Boston Standard Metropolitan Statistical Area (SMSA) in 1970.



**Note:** You should conduct some desktop research to identify determinates/drivers of

Housing prices in order to fully understand and interpret the key findings of your

exploratory data analysis (EDA) and Linear Regression Model for the *housing.csv* data set

for Task 2.

**Task 2.1)** Conduct and report on exploratory data analysis (EDA) of the *housing.csv* data

set using RapidMiner Studio data mining tool**.** Note this will require use of a number of

RapidMiner operators

**Provide following for Task 2.1:**

**(i)** a screen capture of your final EDA process, briefly describe your EDA process

**(ii)** summarise key results of your exploratory data analysis in Table 2.1 Results of

Exploratory Data Analysis for *housing.csv.* **Table 2.1** should include key

characteristics of each variable in *housing.csv* set such as maximum, minimum

values, average, standard deviation, most frequent values (mode), missing values

and invalid values etc.

**(iii)** Discuss key results of exploratory data analysis presented in Table 2.1 and provide

a rationale for selecting top 5 variables for predicting median house value (medv),

in particular focusing on the relationships of independent variables with each other

and with dependent variable median house value (medv) drawing on results of

EDA analysis and relevant literature on determinates of house prices

**(25 marks 300 words)**

**Hint:** Statistics Tab and Chart Tab in RapidMiner Studio provide a lot of descriptive

statistical information and the ability to create useful charts like Barcharts, Scatterplots,

Boxplot charts etc for EDA analysis. You might also like to look at running correlations

and/or chi square tests as appropriate to determine which variables contribute most to

predicting median house value (medv).

**Task 2.2)** Build and report on your Linear Regression model for predicting medv using

RapidMiner data mining process and appropriate set of data mining operators and a reduced

set of variables from *housing.csv* data set as determined by your exploratory data analysis in

Task 2.1.

**Provide the following for Task 2.2:**

(i) A screen capture of Final Linear Regression Model process and briefly describe

your Final Linear Regression Model process

(ii) Table 2.2 named Results of Final Linear Regression Model for Task 2.2 for

*housing.csv* data set.

(iii) Discuss the results of Final Linear Regression Model for *housing.csv data set*

drawing on key outputs (coefficients, standardised coefficients, t-statistics values,

p-values and significance levels etc) for predicting median house value (medv)

and relevant supporting literature on interpretation of a Linear Regression Model.

**(20 marks 200 words)**

Include all appropriate outputs such as RapidMiner Processes, Graphs and Tables that

support key aspects of exploratory data analysis and linear regression model analysis of

the *housing.csv* data set in your Report 2.

**Note:** export Processes and Graphs from RapidMiner using File/Print/Export Image option,

include in Task 2 section or in Appendix 2 of Report 2.

**Report quality structure, presentation, writing style and referencing (Worth 10**

**marks)** Your Report 2 must be presented in report format, written in an appropriate style

and supported where required with appropriate in text references using Harvard

Referencing Style

**Your report 2 must be structured in report format as follows:**

Report 2 Cover/Title Page

Table of Contents

Body of report 2 – Task 1 main heading with appropriate sub headings Task 1.1, Task 1.2,

Task 2 main heading with appropriate sub headings Task 2.1 and Task 2.2…etc

List of References

List of Appendices

**You submit one file for Report 2: f**or Tasks 1 and 2 in Word document format with

extension *.docx*