# Instructions for candidates

You should use the file (filename) provided on blackboard and you should load this file from your data folder / directory.

Save the Rmarkdown file provided as your studentID.Rmd; you will upload this file as your submission.

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The author of this document should be set to **your student ID**. Do not change the authorship to your name.

The subsubsections labelled **Answer** in the RMarkdown document indicate where you should put in your written answers. The template also provides blank code chunks for you to complete your answers; you may choose to add additional chunks if required.

# Data description

*A description of the variables will be here, alongside a description of what subset of observations you should use for the analysis.*

Once you have completed the instructions above for selecting which subgroup of the data to analyse, use your studentID as a seed and select (but retain the order) 1000 observations to use for the analysis.

*aside: in the exam, to expect a distinction, you should have code that nicely presents your results - including tables; selecting appropriate statistics to present, not just an "output dump".*

The general structure of the exam questions is as follows.

# Question

You will be expected to carry out some data preparation to enable subsequent analysis, including selecting the appropriate data. Marks allocated for your code and commenting.

# Question

# Exploratory Data Analysis

## Descriptive Statistics

1. What descriptive statistics would be appropriate for this dataset?
2. Implement those descriptive statistics in the code chunk below:
3. What have those descriptive statistics told you?

## Exploratory Graphs

1. What exploratory graphs would be appropriate for this dataset?
2. Implement those exploratory graphs in the code chunk below
3. What have those exploratory graphs told you?

## Correlations

1. What linear correlations are present within this data?

# Question

# Bivariate relationship

1. Which of the potential explanatory variables has the strongest linear relationship with the dependent variable?
2. Create a linear model to model this relationship.
3. Explain and interpret the model:
4. Comment on the performance of this model.

## Bootstrap

1. Use bootstrapping on this model to obtain a confidence interval of the estimate of the slope parameter.

# Question

# 2 explanatory variables

1. What do you think should be the next term to add to this model? Provide supporting evidence to justify your answer.
2. Explain and interpret the model:
3. Comment on the performance of this model.

# Question

# Multivariable relationship

Consider a model with [information redacted] explanatory variables included:

1. Explain and interpret the model:
2. Comment on the performance of this model.
3. What general concerns do you have regarding this model?

# Question

# Model simplification

Try to simplify the model made for the previous Question.

1. What grounds for model simplification are you considering?
2. Explain and interpret the selected "best" model. What model simplification (if any) has been achieved?
3. Comment on the performance of this model.

# Question

Write a short report for a client based in Bristol outlining your analysis and your findings, illustrating what they could learn about patterns in [information redacted] and any data collection recommendations you have for them so that they can optimise the analysis for their situation.

Highlight what may or may not be directly transferable from the scenario analysed here.