The submission deadline for this assignment is **12pm Monday 11th December**. You will need to submit via Turnitin, via the Ultra page. I **strongly recommend** submitting at least a few hours ahead of the deadline, in case of technical issues. Only one member of the group should submit the report.

There will also be a presentation component to this work, instructions for which have been released separately. The report itself is worth 20% of your mark for the module, with the presentation worth an additional 15%.

1 Introduction

People, by and large, want to be happy.

Every year, data is collected to allow the countries of the world to be placed on a "happiness ladder". The higher on the ladder a country is, the more happy the population of that country are on average. The ladder score is calculated by asking 1000 people in the country to rate their happiness between 0 and 10.

The data presented with this assignment represent an adaptation of the 2023 data. 137 countries (for perhaps a broad definition of the term "country") are represented. The value `ladder score` tells us how happy the population of a country are on average.

There are any number of variables that might affect the level of happiness within a country. In this data set, we concentrate on six. These six variables are given in Table 1, along with a description of the type of variable they are, and a brief summary of what they represent.

We assume in this assignment that national governments would rather their citizenry were happy, all else being equal.

2 Assignment

Imagine your group has been hired by a national government. This government is very keen to see the happiness level of its population rise over the next decade. They want to know where to focus their efforts in order to do just that.

You are to use the data in the file `Happy.csv` to write a report which gives accessible information to government officials regarding which inde-

Variable	Type	Description
LGPD	Continuous	Natural logarithmic value of Gross Domestic Product
Support	Continuous	Proportion of respondents answering “Yes” to question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”
HLE	Continuous	Healthy Life Expectancy - how long a person is expected to live absent accidents or chronic medical conditions
Freedom	Continuous	Proportion of respondents answering “Yes” to question “Are you satisfied with your freedom to choose what you do with your life?”
Corruption	Continuous	Mean of proportion of respondents answering yes to questions “Is corruption widespread throughout the government in this country?” and “Is corruption widespread within businesses in this country?”
Continent	Nominal	Continent the country is (usually) placed within

Table 1: Predictor Variables

pendent variables seem most closely related with happiness ladder score. Your work should draw upon both descriptive statistics, and a multivariate linear regression model (which we will cover in Week 7), using the variable `Ladder_score` as the response variable. During the writing of your report, you should consider

- Are there outliers/extreme values in the data, and if so, what should be done with them?
- What statistical values/diagrams could best support any points you wish to make?
- Is your multivariate linear regression model performing well? (**Note**: we will cover additional ways to check this in Week 8)
- Is your multivariate linear regression model plausibly meeting the assumptions needed for such a model?

(**Note:** the above list is not intended to include all considerations you should bear in mind).

3 Report

The report you write should be divided into four sections.

1. A one page executive summary, allowing a **non statistician** to understand the results you have presented. Any recommendations for future governmental policy should be given here, in bullet point form.

2. Findings (max 3 pages, including figures and tables).

   This section should contain descriptions of your main findings and recommendations for the future, as you would present them to senior government officials who will be tasked with carrying out the work you recommend. These officials will not be statisticians, so keep statistical jargon to a minimum, and use figures or tables to support your findings and recommendations.

   This section's aim is to provide the officials with an overview of what you found, to persuade them your recommendations are worth following. You should also summarise any limitations of your approach, and of the data used. Honest and constructive criticism will be valuable when these governmental officials write their own reports at a later date regarding the overall project of increasing national happiness.

3. Statistical methodology (max 6 pages, including figures and tables).

   This is a section written for other data scientists who may wish to check your work, or build on it themselves. You should describe here the methods you used in statistical language. This description should include any changes you made to the data, any decisions you made on variables to exclude or transform, how you assessed the performance of your model, and how you performed residual diagnostics on your model.

   A major goal of this section is to provide other data scientists with enough information that, should they wish to, they could completely reproduce your results.

4. Appendix (max 3 pages, including figures and tables).

   Here you should include any annotated `R` code and any additional figures or results supporting statements in Sections (1) and (2) but not included there. Do **not** put `R` code in Sections (1)-(2)

Note that the above page counts are strict - you will lose marks if you go over the page count in any section. The page counts are also deliberately quite low. This is to get you to think very carefully about what information does and does not truly **need** to be included in your report. We are imagining you are handing this report to people who will be extremely busy indeed!

You will also need to produce a presentation of your results, as detailed in the presentation assignment.

4 Mark Scheme

The report will be marked out of 20. The marking criteria for the report is given in Table 2, and is adapted from the university's marking criteria.

Mark	**Criteria**
18-20	The report is exemplary, providing clear evidence of a complete grasp of both the statistical methods employed, and their interpretation in the given context. The report is exceptionally well-designed, concisely offering relevant information and strong commentary at all points. The English used is faultless.
15-17	The report is excellent, providing strong evidence of a complete grasp of the statistical methods employed, along with a thorough understanding of how to interpret them in the given context. The report is very well-designed, frequently offering relevant information and strong commentary. The English used is very strong.
13-14	The report is good, providing evidence of a strong grasp of the statistical methods employed, along with some understanding of how to interpret them in the given context. The report is well-designed, offering relevant information and satisfactory commentary. The English used is good.
11-12	The report is acceptable, and provides evidence of a reasonable grasp of the statistical methods employed, although with limited evidence of an ability to interpret them in context. The report is acceptable in design, though some information is not relevant, or is inappropriately placed. There are some flaws in the English used.
8-10	The report contains insufficient evidence of a reasonable grasp of the statistical methods employed, although there is some evidence present. The context of the data is under-served or presented inaccurately. The report is flawed in design, with limited examples of relevant information. The English used is flawed, and sometimes poor.
6-7	The report is unacceptable, containing little evidence of a reasonable grasp of the statistical methods employed. The context of the data is under-served or presented inaccurately. The report is badly flawed in design, difficult to read and with limited examples of relevant information and numerous errors. The English used is often poor.
3-5	The report is unacceptable, containing little evidence of a reasonable grasp of the statistical methods employed. The context of the data is ignored or presented with major errors. The report is extremely badly flawed in design, extremely difficult to read and with almost no examples of relevant information and numerous serious errors. The English used is very poor.
0-2	The report is completely unacceptable, containing no evidence of any grasp of the statistical methods employed. The context of the data is ignored or presented without any accuracy. The report is exceptionally badly flawed in design, almost impossible to read and with no examples of relevant information, and with numerous serious errors. The English used is essentially unreadable.

Table 2: Marking criteria