Summative Assessment for EMATM0061

Statistical Computing and Empirical Methods
Teaching Block 1, 2022

Introduction

This document contains the specification for the summative assessment for the unit EMATM0061, TB1 2022. Please read carefully the following instructions before you start answering the questions.

Deadline. Your report is due on Tuesday 10$^{\text{th}}$ January 2022 at 13:00.

Rules: For the summative assessment you should not share your answers with your colleagues. This is an independent task. The experience of solving the problems in this project will prepare you for real problems in your career as a data scientist. If someone asks you for the answer, resist! Instead, you can demonstrate how you would solve a similar problem.

Support: Whilst this is an independent task, there is a lot of support available if you need it. Talk to your classmates and book office hours. If you are unclear about what is required for any part of the assessment then discuss this issue with the Unit Director in the computer lab or email `rihuan.ke@bristol.ac.uk`, including the unit code EMATM0061 in the subject of your email.

Plagiarism: Be very careful to avoid plagiarism. For more details, you should consult the "Plagiarism" section within the central Blackboard page for the Data Science MSc.

Extenuating circumstances: For more details on the procedure for extenuating circumstances consult the "Extenuating circumstances" section within the central Blackboard page for the Data Science MSc.

Clarity: Clarity is highly important. Be careful to make sure you clearly explain each step in your answer. You should also include comments within your code. Your answer should clearly demarcate which part of the question you are answering.

Programming language: For Section A of this coursework you should use `Tidyverse` methods within the `R` programming language. For Sections B and C, you can use either R, Python or Julia. Regardless of your choice of language, it is essential that your answers are clear and well-written.

Submission points: To submit your solutions, please visit the "Assessment, submission and feedback" tab on the course webpage at Blackboard. Make sure your submission follows the submission structure described below.

Submission structure: Your submission should include the following four parts (there will be 4 separate submission points at Blackboard for the four parts). You can submit each of the four parts separately but make sure you submit each part.

- Part 1 (Section A). A report that contains your answers to section A.
- Part 2 (Section B). A report that contains your answers to section B.
- Part 3 (Section C). A report that contains your answers to section C.
- Part 4 (Source code and data). In this part, you should submit a single folder. This folder should contain three sub-folders entitled "`username_EMATM0061_A`", "`username_EMATM0061_B`" and "`username_EMATM0061_C`" (corresponding to the three sections A, B, and C, respectively) where"`username`" is replaced with your unique University of Bristol username. Each sub-folder should contain any data and any supporting code (e.g., .Rmd files) used to create the report for the corresponding section.

For each of Part 1, Part 2 and Part 3, your report should be a single PDF file. So if you have your solutions in another file format (such HTML or images), please convert it to a PDF file before the submission. It is important that your approach to solving the questions is visible within these reports and you are encouraged to include pieces of clear and well-written code along with explanatory pros within the report itself.

Be careful to read each question in each section (A, B, and C) when writing your report.

Section A (20 marks)

In this part of your assessment, you will perform a data wrangling task with some finance data.

Note that clarity is highly important. Be careful to make sure you clearly explain each step in your answer. You should also include comments within your code when necessary. In addition, make the structure of your answer clear through the use of headings. You should also make sure your code is clean by making careful use of Tidyverse methods.

A.1

Begin by downloading the .csv file available within the Assessment section within Blackboard entitled "finance_data_2022". The file contains data about the cumulative commitments funds from the International Finance Corporation (IFC) as well as Loan & Guarantee participations, across different IFC regions and countries.

Next load the "finance_data_2022" csv file into a `R` data frame called "`data_original`".

How many rows and how many columns does this data frame have? Use `R` commands to display the numbers.

A.2

Using the data frame "`data_original`", generate a new data frame called "`finance_data`" with 5 columns:

The 1st column should be called `IFC` and correspond to the `IFC Region` column in the csv file.

The 2nd column should be called `IFC_CC` and correspond to the "`IFC Cumulative Commitments (US$ Thousands)`" column in the csv file.

The 3rd column should be called `Country` and correspond to the "`Country`" column in the csv file.

The 4th column should be called `Loan_Guarantee_CC` and correspond to the "`Loan & Guarantee participations Cumulative Commitments (US$ Thousands)`" column in the csv file.

The last column should be called `Date` and correspond to the "`As of Date`" column in the csv file.

A.3

Create a new data frame called `data_part1` by choosing a subset of the data frame `finance_data` that contains all rows satisfying that `IFC_CC` is no less than 300000 and `Loan_Guarantee_CC` is no more than 500000. The result should be stored in the new data frame `data_part1` and your data frame `finance_data` should not be changed in this task.

Reorder the rows of the data frame `data_part1` such that the values in the column `IFC_CC` are in descending order.

Display a subset of the `data_part1` data frame consisting of the first 4 rows and the three columns "`IFC`", "`IFC_CC`", and "`Loan_Guarantee_CC`".

A.4

Add a new column called IFC_ratio to the data frame finance_data. For each row of the data frame, the element of the IFC_ratio column is computed by $\alpha/(\alpha + \beta)$ where α denotes the element of the IFC_CC column, and β denotes the element of the Loan_Guarantee_CC column. Now, your data frame finance_data should have 6 columns.

Display a subset of the data frame finance_data consisting of the first 5 rows and the 4 columns "IFC", "IFC_CC", and "Loan_Guarantee_CC" and 'IFC_ratio".

A.5

Your data frame finance_data has a column called Date which corresponds to the "As of Date" column in the .csv file. The elements in the Date column are in the format of the day, month and year separated by the forward slash character "/", for example "06/30/2017". Divide the Date column into three columns called day, month, year. That is, for each row of the data frame finance_data, the day, month and year in the Date column are separated into three different columns called day, month, year, respectively. Make sure each of the day, month, year columns is of numeric type (rather than characters). Now, your data frame finance_data should have 8 columns.

Display a subset of the data frame finance_data consisting of the first 5 rows and the 4 columns "IFC_CC", "day", "month", and "year".

A.6

Next generate a summary data frame called "summary_data" from the "finance_data". The summary data frame "summary_data" should have the following properties:

Your summary data frame should have 7 rows corresponding to the 7 different IFC regions specified in the IFC column of "finance_data".

Your summary data frame should also have 7 columns:

(a) IFC - The IFC regions: "East Asia and the Pacific", "Europe and Central Asia", "Latin America and the Caribbean", "Middle East and North Africa", "South Asia", "Sub-Saharan Africa", "Worldwide"

(b) ifc_mn - the mean of "IFC Cumulative Commitments (US$ Thousands)" for the corresponding IFC region.

(c) ifc_21q - the 0.21-quantile of "IFC Cumulative Commitments (US$ Thousands)" for the corresponding IFC region.

(d) ifc_var - the variance of "IFC Cumulative Commitments (US$ Thousands)" for the corresponding IFC region.

(e) lg_mn - the mean of "Loan & Guarantee participations Cumulative Commitments (US$ Thousands)" for the corresponding IFC region.

(f) lg_21q - the 0.21-quantile of "Loan & Guarantee participations Cumulative Commitments (US$ Thousands)" for the corresponding IFC region.

(g) lg_var - the variance of "Loan & Guarantee participations Cumulative Commitments (US$ Thousands)" for the corresponding IFC region.

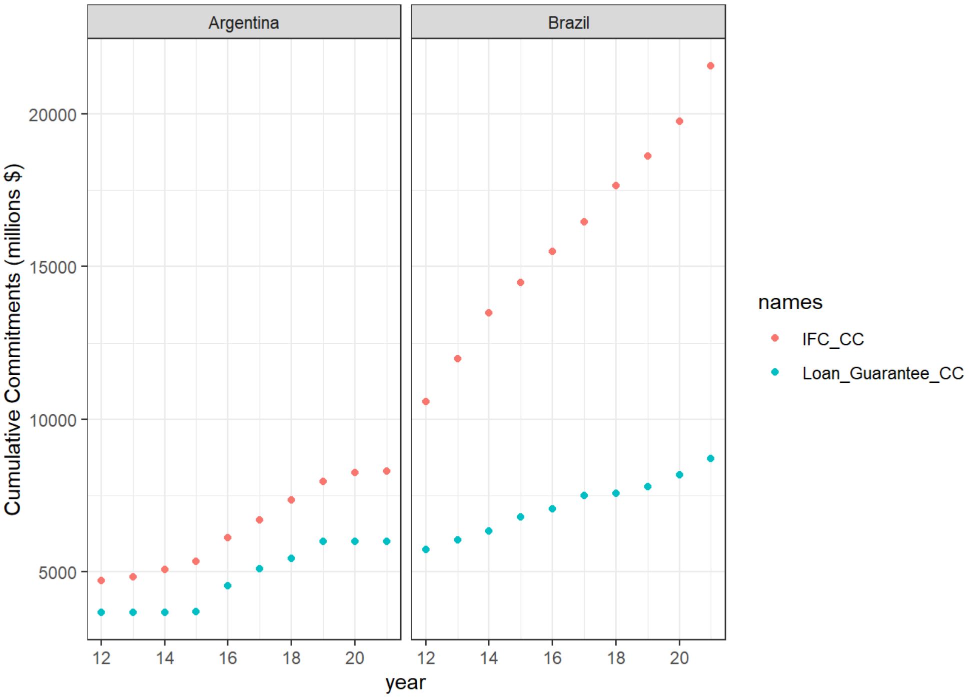
You should use Tidyverse methods to create your "summary_data" data frame. The missing values (NA) should not be taken into account when computing the summary data. Your data frame finance_data should not be changed in this task.

Display the "summary_data" data frame.

A.7

Using your data frame finance_data, create a plot to display the "IFC Cumulative Commitments" and "Loan & Guarantee participations Cumulative Commitments" as functions of the years, for two different countries Argentina and Brazil, respectively. Your plot should have two panels, each of which displays results corresponding to one of the two countries. In your plot, the years are represented by their last two digits (for example, 2022 is represented by 22). The Cumulative Commitments should be in the unit of million dollars (million $). Your data frame finance_data should not be changed in this task.

Your plot is expected to look as follows:



A.8

Create a function called "impute_by_quantile" which takes as input a vector numerical values, which may include some "NA"s, and replaces any missing values ("NA"s) with the 0.9-quantile of the vector.

Next, apply your function "impute_by_quantile" to each of the columns "IFC_CC", "Loan_Guarantee_CC", and "IFC_ratio" in your data frame finance_data. This aims to replace the missing values (NA) with the 0.9 quantile of the corresponding column, within the data frame finance_data.

Next, display a data frame of three columns ("IFC_CC", "Loan_Guarantee_CC", and "IFC_ratio") and 1 row. The "IFC_CC" column should contain a single number representing the mean of the "IFC_CC" column of your data frame finance_data. The "Loan_Guarantee_CC" column should contain a single number representing the mean of the "Loan_Guarantee_CC" column of your data frame finance_data. The "IFC_ratio" column should contain a single number representing the mean of the "IFC_ratio" column of your data frame finance_data.

Section B (30 marks)

B.1 (7 marks)

In this question, we consider the lifetime of products from a light bulb factory. The factory has two LED bulb production lines (Line A and Line B) that independently produce light bulbs. The light bulbs, however, are not always of high quality, and the products from Line A and Line B have different lifetimes. For a light bulb produced by Line A, the probability that its lifetime is equal to or bigger than 2 years is p_A (a number in $[0, 1]$). A light bulb from Line B generally has a shorter lifetime, and the probability of its lifetime being equal to or bigger than 2 years is p_B (which is less than p_A).

The light bulbs produced from Line A and Line B are considered to be the same products by the factory and they are sold randomly to the customers. These light bulbs are sold in the same packages such that, given a light bulb, a customer can not identify which of the two production lines the light bulb is made from. Suppose that if we buy a light bulb produced by the factory, then it must come from either Line A or Line B, and the probability that it comes from Line A is p. A customer bought a light bulb whose lifetime was unfortunately less than 2 years. The customer suspects that this light bulb was made in Line B.

(1). Let α denote the probability that this light bulb was made in Line B, given that its lifetime is less than 2 years. Derive a mathematical expression for α in terms of p_A, p_B and p.

(2). Suppose that $p_A = 0.99$, $p_B = 0.5$ and $p = 0.1$. What is the numerical value of α?

Next, fix $p_A = 0.99$, $p_B = 0.5$ and $p = 0.1$. Conduct a simulation study to estimate the conditional probability α with samples.

(3). Your simulation study should contain 100000 trials. In each of the trials, generate a sample of a light bulb. Each sample is represented by a pair of randomly generated numbers called (*Line*, *LessThen2Years*):

  (i). The random number *Line* represents the product line that the light bulb was made from (with *Line*=0 representing Line A and *Line*=1 representing Line B). The probability of *Line*=0 should be equal to p.

  (ii). The random number *LessThen2Years* should be either 0 or 1, where *LessThen2Years* =0 represents that the lifetime of this bulb is not less than 2 years, and *LessThen2Years* =1 represents that the lifetime is less than 2 years. The number *LessThen2Years* should be generated by taking into account the value of *Line*. If *Line*=0, then the probability of *LessThen2Years*=0 is equal to p_A. If *Line*=1, then the probability of *LessThen2Years* =0 is equal to p_B.

(4). Based on the samples generated in the simulation study, compute an estimate of the conditional probability α. First, select the subset of samples in which *LessThen2Years* =1. Second, within this subset of samples, compute the number of samples in which *Line*=1, and divide it by the total number of samples within this subset to get an estimate of the conditional probability α. Display your estimate to at least 5 decimal places.

B.2 (9 marks)

In this question, we will explore statistical estimation for parameters in continuous random variables.

Suppose a product is being sold in a supermarket. We are interested in knowing how quickly the product returns to the shelf again after it is sold out. Let X be a continuous random variable denoting the length of time between the time point at which it is sold out and the time point at which it is placed on the shelf again. So X should be a non-negative number, and $X = 0$ means that the product gets on the shelf immediately after it is sold out. Here, we assume that the probability density function of X is given by

$$p_\lambda(x) = \begin{cases} ae^{-\lambda(x-b)} & \text{if } x \geq b, \\ 0 & \text{if } x < b, \end{cases}$$

where $b > 0$ is a known constant, $\lambda > 0$ is a parameter of the distribution, and a is to be determined by λ and b.

(1). First, determine the value of a: derive a mathematical expression of a in terms of λ and b.

(2). Derive a formula for the population mean and stand deviation of the exponential random variable X with parameter λ.

(3). Derive a formula for the cumulative distribution function and the quantile function for the exponential random variable X with parameter λ.

(4). Suppose that $X_1, \cdots, X_n$ are independent copies of X with the unknown parameter $\lambda > 0$. What is the maximum likelihood estimate λ_{MLE} for λ?

Now download the .csv file entitled "`supermarket_data_EMATM0061`" from the Assessment section within Blackboard. The .csv file contains synthetic data on the length of time (in seconds) taken by a product to get on the shelf again after being sold out. So the sample is a sequence of time lengths. Let's model the sequence of time lengths in our sample as independent copies of X with parameter λ and known constant $b = 300$ (seconds). Answer the following questions (6) and (7).

(5). Compute and display the maximum likelihood estimate of λ_{MLE} of the parameter λ.

(6). Apply the method of Bootstrap confidence interval to obtain a confidence interval for λ with a confidence level of 95%. To compute the Bootstrap confidence interval, the number of resamples (i.e., subsamples that are generated to compute the bootstrap statistics) should be set to 10000.

Next, conduct a simulation study to explore the behaviour of the maximum likelihood estimator:

(7). Conduct a simulation study to explore the behaviour of the maximum likelihood estimator λ_{MLE} for λ on simulated data $X_1, \cdots, X_n$ (as independent copies of X with parameter λ) according to the following instructions. Take $b = 0.01$ and consider a setting in which $\lambda = 2$ and generate a plot of the mean squared error as a function of the sample size n. You should consider a sample size between 100 and 5000 in increments of 10, and consider 100 trials per sample size. For each trial of each sample size generate a random sample $X_1, \cdots, X_n$ (as independent copies of X with parameter $\lambda = 2$), then compute the maximum likelihood estimate λ_{MLE} for λ based upon the corresponding sample. Display a plot of the mean square error of λ_{MLE} as an estimator for λ as a function of the sample size n.

B.3 (14 marks)

Consider a bag of a red balls and b blue balls (so the bag has $a+b$ balls in total), where $a \geq 1$ and $b \geq 1$. We randomly draw two balls from the bag without replacement. That means, we draw the first ball from the bag and, WITHOUT returning the first ball to the bag, we draw the second one. Each ball has an equal chance of being drawn. Now we record the colour of the two balls drawn from the bag, and let X denote the number of red balls minus the number of blue balls. So X is a discrete random variable. For example, if we draw one red ball and one blue ball, then $X = 0$. Answer the following questions from (1) to (11).

(1). Give a formula for the probability mass function $p_X : \mathbb{R} \to [0,1]$ of X.

(2). Use the the probability mass function p_X to obtain an expression of the expectation $\mathbb{E}(X)$ of X (i.e., the population mean) in terms of a and b.

(3). Give an expression of the variance $\mathrm{Var}(X)$ of X in terms of a and b.

(4). Write a function called `compute_expectation_X` that takes a and b as input and outputs the expectation $\mathbb{E}(X)$. Write a function called `compute_variance_X` that takes a and b as input and outputs the variance $\mathrm{Var}(X)$.

Additionally, suppose that $X_1, X_2, \cdots, X_n$ are independent copies of X. So $X_1, X_2, \cdots, X_n$ are i.i.d. random variables having the same distribution as that of X. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean.

(5). Give an expression of the expectation of the random variable $\overline{X}$ in terms of a, b.

(6). Give an expression of the variance of the random variable $\overline{X}$ in terms of a, b and n.

(7). Create a function called `sample_Xs` which takes as inputs a, b and n and outputs a sample $X_1, X_2, \cdots, X_n$ of independent copies of X.

(8). Let $a = 3$, $b = 5$ and $n = 100000$. Compute the numerical value of $\mathbb{E}(X)$ using the function `compute_expectation_X` and compute the numerical value of $\mathrm{Var}(X)$ using the function `compute_variance_X`. Then use the function `sample_Xs` to generate a sample $X_1, X_2, \cdots, X_n$ of independent copies of X. With the generated sample, compute the sample mean $\overline{X}$ and sample variance. How close is the sample mean $\overline{X}$ to $\mathbb{E}(X)$? How close is the sample variance to $\mathrm{Var}(X)$? Explain your observation.

Moreover, let $\mu := \mathbb{E}(X)$ and $\sigma := \sqrt{\mathrm{Var}(X)/n}$, and let $f_{\mu,\sigma} : \mathbb{R} \to [0,\infty)$ be the probability density function of a Gaussian random variable with distribution $\mathcal{N}(\mu, \sigma^2)$, i.e., the expectation is μ and the variance is σ^2. Next, conduct a simulation study to explore the behaviour of the sample mean $\overline{X}$.

(9). Let $a = 3$, $b = 5$ and $n = 900$. Conduct a simulation study with 20000 trials. In each trial, generate a sample $X_1, \cdots, X_n$ of independent copies of X. For each of the 20000 trials, compute the corresponding sample mean $\overline{X}$ based on $X_1, \cdots, X_n$.

(10). Create a scatter plot of the points $\{(x_i, f_{\mu,\sigma}(x_i))\}$ where $\{x_i\}$ are a sequence of numbers between $\mu - 3\sigma$ and $\mu + 3\sigma$ in increments of 0.1σ. Then append to the scatter plot a curve representing the kernel density of the sample mean $\overline{X}$ within your simulation study (with 20000 trials). Use different colours for the point $\{(x_i, f_{\mu,\sigma}(x_i))\}$ and the curve in the kernel density plot of the sample mean $\overline{X}$.

(11). Describe the relationship between the density of $\overline{X}$ and the function $f_{\mu,\sigma}$ displayed in your plot. Try to explain the reason for the observed relationship.

Section C (50 marks)

In this section, you are asked to complete a Data Science report which demonstrates your understanding of a statistical method. The goal here is to choose a topic that you find interesting and explore that topic in depth. You are free to choose a topic and data set that interests you.

There will be an opportunity to discuss and get advice on your chosen direction in the computer labs.

Below are two flexible example structures you can consider for this section of your report. If you are unsure what to do, choose one of the following. Note that you should not submit more than one of the example tasks below.

Example task 1

Investigate a particular hypothesis test e.g. a Binomial test, a paired Student's t test, an unpaired Student's t test, an F test for ANOVA, a Mann-Whitney U test, a Wilcoxon signed-rank test, a Kruskal Wallis test, or some other test you find interesting.

Note that clarity of presentation is highly important. In addition, you should aim to demonstrate a depth of understanding. For this hypothesis test you are asked to do the following:

1. Give a clear description of the hypothesis test including the details of the test statistic, the underlying assumptions, the null hypothesis and the alternative hypothesis. Give an intuitive explanation for why the test statistic is useful in distinguishing between the null and the alternative.

2. Perform a simulation study to investigate the probability of type I error under the null hypothesis for your hypothesis test. Your simulation study should involve randomly generated data which conforms to the null hypothesis. Compare the proportion of rounds where a Type I error is made with the significance level of the test.

3. Apply this hypothesis test to a suitable real-world data set of your choice (some places to find data sets are described below). Ensure that your chosen data set is appropriate for your chosen hypothesis test. For example, if your chosen hypothesis test is an unpaired t-test then your chosen data set must have at least one continuous variable and contain at least two groups. It is recommended that your data set for this task not be too large. You should explain the source and the structure of your data set within your report.

4. Carefully discuss the appropriateness of your statistical test in this setting and how your hypotheses correspond to different aspects of the data set. You may want to use plots to demonstrate the validity of your underlying assumptions. Draw a statistical conclusion and report the value of your test statistic, the p-value and a suitable measure of effect size.

5. Discuss what scientific conclusions can you draw from your hypothesis test. Discuss how these would have differed if the result of your statistical test had differed. Discuss key experimental design considerations necessary for drawing any such scientific conclusion. For example, perhaps an alternative experimental design would have allowed one to draw a conclusion about cause and effect?

Example task 2

Investigate a particular method for supervised learning. This could either be a method for regression or classification but should be a method with at least one tunable hyperparameter. You could choose one from ridge regression, k-nearest neighbour regression, a regression tree, regularized logistic regression, k-nearest neighbour classification, a decision tree, a random forest or another supervised learning technique you find interesting.

Note that clarity of presentation is highly important. In addition, you should aim to demonstrate a depth of understanding.

1. Give a clear description of the supervised learning technique you will use. Explain how the training algorithm works and how new predictions are made on test data. Discuss what type of problems this method is appropriate for.

2. Choose a suitable data set to apply this model to and perform a train, validation, and test split (some places to find data sets are described below). Be careful to ensure that your data set is appropriate for your chosen algorithm. For example, if you have chosen to investigate a classification algorithm then your chosen data set must contain at least one categorical variable. Your data set for this task does not need to be large to obtain good results. The size of your data set should not exceed 100MB and you should aim to use a data set well within this limit. Your report should carefully give the source for your data. In addition, describe your data set. How many features are there? How many examples? What type is each of the variables (e.g. Categorical, ordinal, continuous, binary etc.)?

3. What is an appropriate metric for the performance of your model? Explore how the performance of your model varies on both the training and the validation data change as you vary the amount of training data used.

4. Explore how the performance of your model varies on both the train and the validation data change as you vary a hyperparameter.

5. Choose a hyper-parameter and report your performance based on the test data. Can you get a better understanding by using cross-validation? Note that you will be graded on your understanding of the key concepts. It is far better to choose a simple hypothesis test and supervised learning algorithm, and apply sound statistical reasoning than to choose complex methods without properly demonstrating your understanding.

Alternative tasks

You could also choose an alternative task in which you explore a statistical method or methods which interest you.

A couple of elements to bear in mind here:

1. Demonstrate a solid level of understanding of the technique or techniques you consider.

2. Apply your chosen method or technique to a real data set. This data must be publicly available and should not exceed 100MB in size.

3. Where appropriate, use simulated data to explore and demonstrate the properties of your chosen method.

4. The subject of your report should be statistical methods or techniques and their performance and behaviour. Whilst you can consider techniques motivated by a particular application, the application itself should not become the focus of your report.

Note:

1. Do not complete and submit more than one of the above tasks. These are example tasks and you should only choose one. The goal here is to explore a topic in detail.

2. You will be graded on the level of understanding of the key concepts demonstrated within your report. With this in mind, it is far better to choose relatively simple methods, and apply sound statistical reasoning than to choose complex methods without properly demonstrating your understanding. You are welcome to include methods not covered within the lectures, provided that they are appropriate for the task at hand and that you are able to demonstrate a clear understanding of the methods used. A small number of additional marks will be given for more advanced methods, provided that a very strong level of understanding is displayed. However, the main focus here is a clear understanding and you should not sacrifice understanding for the sake of complexity. A clear understanding of the basic concepts is paramount.

3. You do not need to use large data sets. The dataset you choose should not be larger than 100MB. This is an upper bound. You should aim to use a data set well within this limit.

Data sets

There are a vast number of freely available data sets across the internet. Below is a few example sources. You are also welcome to use data sets from other sources. Any data you use should be freely available and accessible. The source of your data and the steps required to retrieve it should also be described within your main report.

You should also explain its structure e.g. the number of rows and the number of columns, and what the data in each column of interest represent for, $\cdots$. You are encouraged to use tabular data throughout.

https://www.kdnuggets.com/datasets/index.html
https://r-dir.com/reference/datasets.html
http://archive.ics.uci.edu/ml/datasets.php
http://lib.stat.cmu.edu/datasets/
http://inforumweb.umd.edu/econdata/econdata.html
https://lionbridge.ai/datasets/the-50-best-free-datasets-for-machine-learning/
https://www.kaggle.com/
https://www.ukdataservice.ac.uk/
https://data.worldbank.org/
https://www.imf.org/en/Data

Final remarks

Throughout your report you should emphasise:

- Reproducible analysis (be careful with randomised procedures).
- Clear and informative visualisations of your results.
- Demonstrate a depth of understanding.
- A clear writing style.