Summative Assessment

Statistical Computing and Empirical Methods, Teaching Block 1, 2024

Introduction

This document contains the specification for the summative assessment for the unit Statistical Computing and Empirical Methods, TB1 2024. Please read carefully the following instructions before you start answering the questions.

Deadline. Your report is due on 28 November 2024 at 13:00.

Rules: This is *an independent task*. For the summative assessment you should not share your answers with your colleagues. 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!

Support: Whilst this is an independent task, there is a lot of support available if you need it. If you are unclear about what is required for any part of the assessment then discuss this issue with the our teaching team in the computer lab or contact your unit director.

Plagiarism: Be very careful to avoid plagiarism. For more details, you should consult the "Academic Integrity" section under the Assessment tab within the central Blackboard page for the School of Engineering Mathematics & Technology.

The use of generative AI: The use of generative AI, such as *ChatGPT*, is prohibited. Any use of generative AI in this assessment will be considered as plagiarism.

Extenuating circumstances: For more details on the procedure for extenuating circumstances consult the "Assessment support options" section under the Assessment tab within the central Blackboard page for the School of Engineering Mathematics & Technology.

Late submission panalty: Coursework that is submitted after a deadline should be subject to a late submission penalty, unless there is an extension or a justified exceptional circumstance. The more details, you should consult the central Blackboard page for the School of Engineering Mathematics & Technology or contact the School office.

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 when necessary. Your answer should clearly demarcate which part of the question you are answering. Whenever possible, include pieces of well-written codes in your report to promote clarity.

Programming language: *For Section A* of this coursework you should use `Tidyverse` methods within the `R` programming language. For *Section B* and *Section C*, you can use either R or Python. 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.

Multiple submissions: Submitting the coursework multiple times before the deadline is allowed. However, only the last submission will be considered for marking. You can try to submit an temporary copy before your final submission if you like.

Submission structure: Please submit a single zip file that contains a folder named "`SCEM_???`" where "???" should be replaced by your unique UoB username (e.g., lf22553). The folder should contain three subfolders named "`A`", "`B`" and "`C`".

1 Subfolder "`A`" should include 1) a PDF file that contains your answers to Section A, and 2) a folder containing the code and data being used for Section A.

2 Subfolder "`B`" should include 1) a PDF file that contains your answers to Section B, and 2) a folder containing the code and data being used for Section B.

3 Subfolder "`C`" should include 1) a PDF file that contains your answers to Section C, and 2) a folder containing the code and data being used for Section C.

Time allocation: Section A & B and Section C both contain 50 marks, but we recommend that you allocate more time for the tasks in Section C, for example 40% on Section A & B and 60% on Section C.

Section A (20 marks)

General instruction: In this part of your assessment, you will perform a data wrangling task using *R programming*. 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* in R.

(Q1). First download the files entailed "debt_data.csv", "country_data.csv" and "indicator_data.csv" which are available within the Assessment section within Blackboard.

The file "debt_data.csv" contains debt data for different countries under different indicators, from 1960 to 2023. The indicators are represented by indicator codes (for example, NY.GNP.MKTP.CD). The file "indicator_data.csv" contains a list of the indicator names as well as their associated indicator codes. The file "country_data.csv" contains information about the country code, income levels, and regions for each country.

First, Load the file "debt_data.csv" into an R data frame called "debt_df", load the file "country_data.csv" into an R data frame called "country_df", and load the file "indicator_data.csv" into a data frame called "indicator_df".

Second, use R to check the number of columns and the number of rows that the data frame "debt_df" has. Display your results.

(Q2). Update "debt_df" by reordering its rows such that the values of the indicator "DT.NFL.BLAT.CD" is in descending order. Display a subset of the updated "debt_df" consisting of the first 4 rows and the columns "Country.Code", "Year", "NY.GNP.MKTP.CD", and "DT.NFL.BLAT.CD".

(Q3). In the data frame "debt_df", the indicators are represented by their associated indicator codes rather than by their names. The data frame "indicator_df" contains a list of indicator names and their corresponding indicator codes. Create a new data frame called "debt_df2" by combining the data from the two data frames "debt_df" and "indicator_df". The new data frame "debt_df2" should be equivalent to "debt_df" except that "debt_df2" now contains indicator names rather than indicator codes. The indicator names in "debt_df2" should match the indicator codes in "debt_df" according to their correspondence described in "indicator_df".

Display a subset of "debt_df2" consisting of the first 5 rows and the three columns "Country.Code", "Year", and "Net financial flows, others (NFL, current US$)".

(Q4). The data frame "country_df" contains information about Region, Income groups, and country name for each country. Create a new data frame called "debt_df3" by combining data from the two data frames "debt_df2" and "country_df". The new data frame "debt_df3" should contains a) all columns from "debt_df2" and b) 3 columns from "country_df" called "Region", "IncomeGroup", and "Country.Name". Make sure that in each row of "debt_df3", the "Region", "IncomeGroup", and "Country.Name" match "Country.Code" according to their correspondence described in "country_df".

Your data frames "debt_df3" and "debt_df2" should have the same numbers of rows, but "debt_df3" has three more columns.

Display a subset of "debt_df3" consisting of the first three rows and 4 columns called "Country.Name", "IncomeGroup", "Year", and "Total reserves in months of imports".

(Q5). Rename the following 5 columns from their original names to the new names specified below

Original column names	New column names
Total reserves in months of imports	Total_reserves
External debt stocks, total (DOD, current US$)	External_debt
Net financial flows, bilateral (NFL, current US$)	Financial_flow
Imports of goods, services and primary income (BoP, current US$)	Imports
IFC, private nonguaranteed (NFL, US$)	IFC

(Q6). Next generate a summary data frame called “debt_summary” from the data frame “debt_df3” with the following properties:

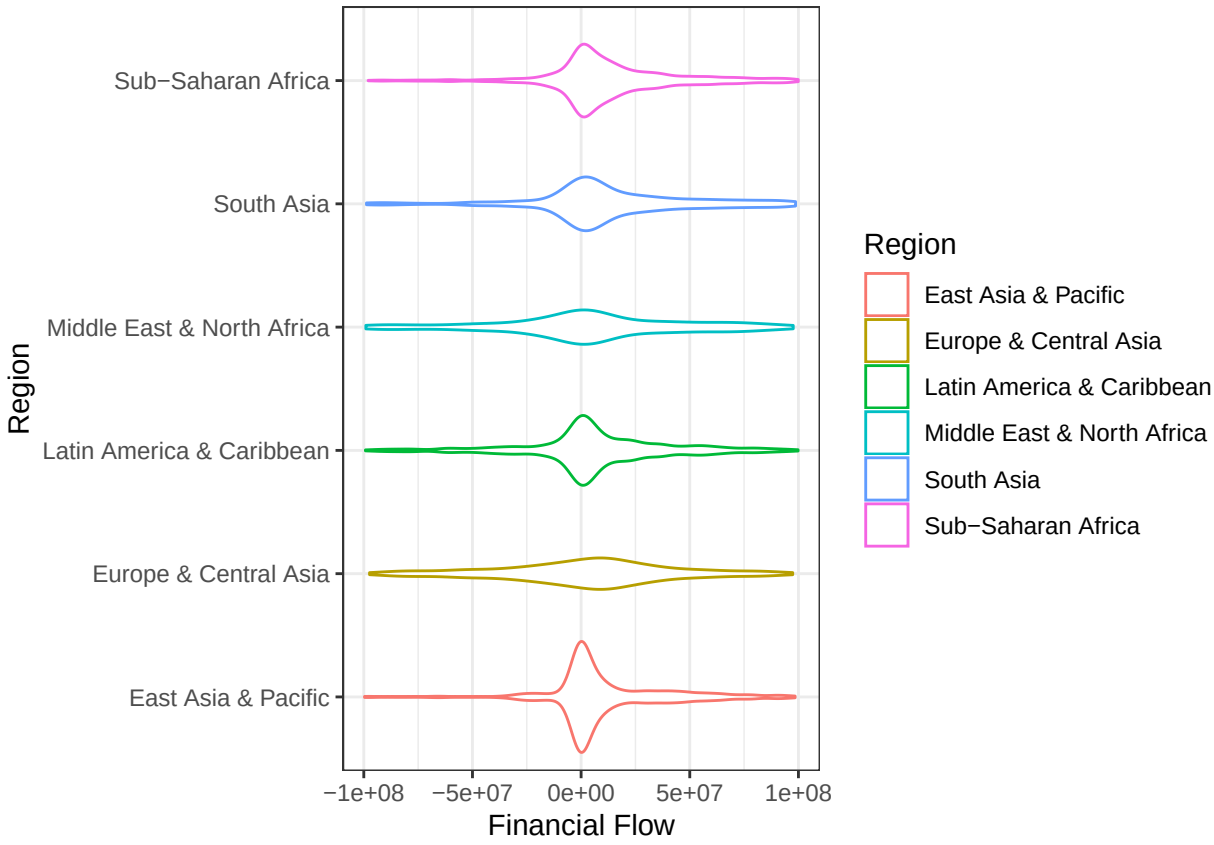
Your summary data frame “debt_summary” should contain 7 rows corresponding to the 7 different Regions, and it should also have 5 columns:

"Region" - the names of the 7 different regions including "East Asia & Pacific", "Europe & Central Asia" etc.
"TR_mn" - the average of "Total_reserves" in each region.
"ED_md" - the median of "External_debt" in each region.
"FF_quantile" - the 0.2 quantile of "Financial_flow" in each region.
"IFC_sd" - the standard deviation of "IFC" in each region.

All missing values should be discarded when computing the summary data.

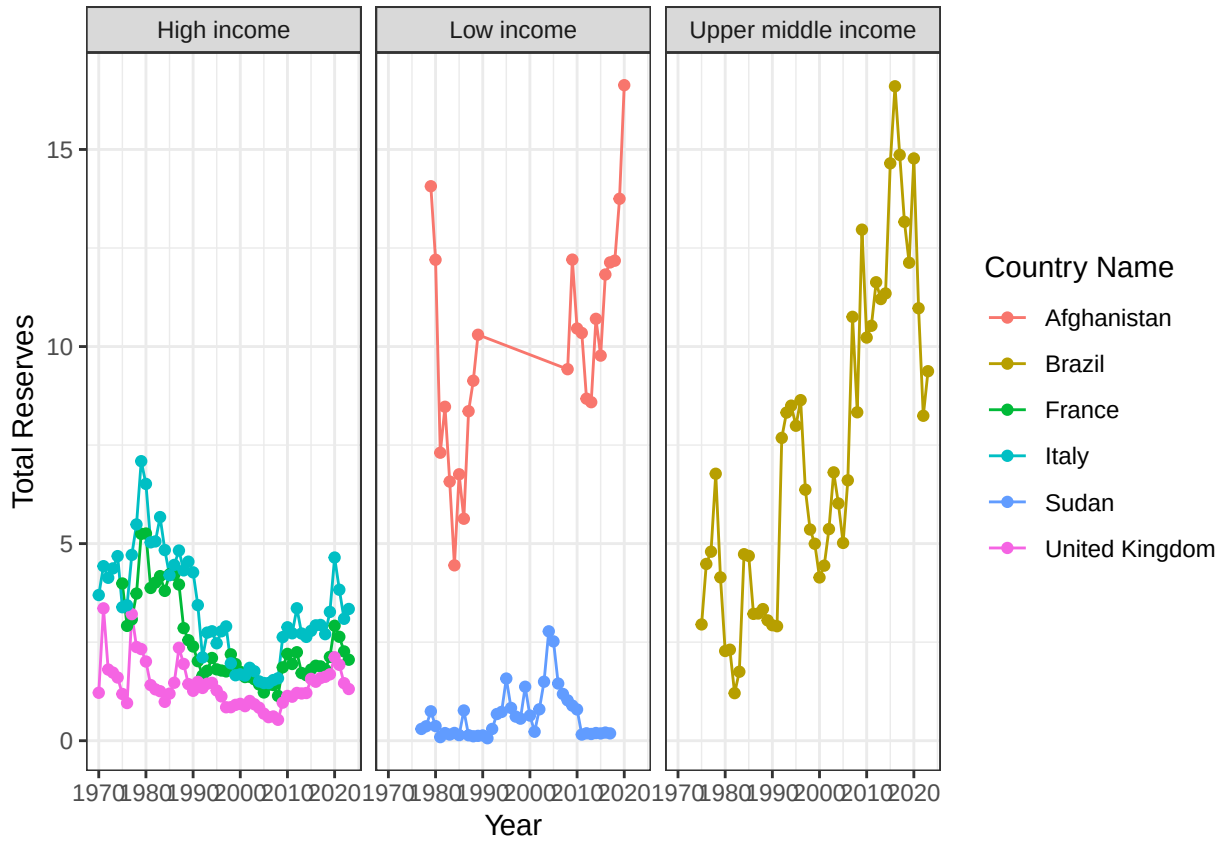
(Q7). Based on your data frame “debt_df3”, create a violin plot of "Financial_flow" for each of the regions. The violin plots should be displayed in the same figure and with different colors representing different regions. Ignore all missing values and all values that are smaller than -10^8 or bigger than 10^8.

Your plot is expected to look as follows.



(Q8). Based on the data frame "debt_df3", create a plot which displays the "Total_reserves" as a function of the years (from 1960 to 2023), for each of the following countries: Italy, France, United Kingdom, Sudan, Afghanistan, and Brazil. Additionally, the values of "Total_reserves" should be displayed in different panels according to the income groups of the countries. Use different colors to represent different countries.

Your plot is expected to look as follows.


High income
Low income
Upper middle income
Total Reserves
0
5
10
15
1970
1980
1990
2000
2010
2020
Year
Country Name
Afghanistan
Brazil
France
Italy
Sudan
United Kingdom

Section B (30 marks)

B.1

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/or b.

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

(3) Derive a formula for the cumulative distribution function and the quantile function for the 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_2024`" from the Assessment section within Blackboard. The .csv file contains 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 (X is the random variable mentioned above) with parameter λ and known constant $b = 300$ (seconds). Answer the following questions (5) and (6).

(5) Given the sample, compute and display the maximum likelihood estimate 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. Let $b = 0.01$ and the true parameter be $\lambda = 2$. Generate a plot of the mean squared error as a function of the sample size n. You should consider sample sizes from 100 to 5000 in increments of 10. For each sample size, consider 100 trials. In each trial, generate a random sample $X_1, \cdots, X_n$ (as independent copies of X with parameter $\lambda = 2$), and then compute the maximum likelihood estimate λ_{MLE} for λ based upon the 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.2

Consider a bag of a red balls and b blue balls (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 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/or 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 inputs 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 $\mathrm{Var}(X)$. Display your code.

In the following questions, we additionally assume 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 $n = 100000$. First, 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`. Second, 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 $\sigma := \sqrt{\mathrm{Var}(X)/n}$ (the random variable X is defined above), 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 $\overline{X}$ by answering questions (9)-(11).

(9) Let $a = 3$, $b = 5$ and $n = 100$. Conduct a simulation study with 50000 trials. In each trial, generate a sample $X_1, \cdots, X_n$ of independent copies of X. For each of the 50000 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 50000 trials). Use different colours for the point $\{(x_i, f_{\mu,\sigma}(x_i))\}$ and the density curve 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.

Section C (50 marks)

In this part of the assessment, 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 being considered, including the details of the test statistic and p-value, 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. What happens when a different significance level is used?
3. Choose a suitable real-world data set (for example, 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. You should also explain the related problem on which you want to perform the test.
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 you can 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?
6. Exploring further this hypothesis test on *one* topic/direction of your choice. This could be for example discussing a property of the test such as how the power of the chosen test changes with sample size, significance level, or effect size. As another example, how robust is the test when assumptions are violated and is there a robust alternative? How does the test compare to its non-parametric alternatives? How does the frequentist test compare with its Bayesian alternative? These are just a few examples. Make a clear statement on the question of interest and your conclusions. The details of your approach to support your findings should be visible within your report, and experiments or simulation studies can be included if needed.

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, including the underlying principles and any assumptions. 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 where this method can be applied. Perform a train, validation, and test split (for example, 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.)? You should also explain the associated problem that you will solve using your supervised learning method.
3. What is an appropriate metric for the performance of your model? Give a clear explanation of the metric. Explore how the performance of your model varies on *both* the training data and the validation data as you vary the amount of training data used. You should compare the performance of the models across different sizes of the training data.
4. Explore how the performance of your model varies on *both* the training data and the validation data 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?
6. Exploring further this supervised learning method on *one* topic/direction of your choice. This could be for example discussing how the bias-variance trade-off impacts the performance of the chosen method. As another example, is your model robust? How does the performance of the method change when applied to imbalanced datasets? Does your method work on small data and if not is there an suitable alternative? You could also investigate how different regularisation techniques affect the model's performance, or carefully compare the chosen method with other methods. These are just a few examples. Make a clear statement on the question of interest and your conclusions. The details of your approach to support your findings should be visible within your report, and experiments or simulation studies can be included if needed.

Further instruction for Section C.

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. Additional marks will be given for more advanced methods, provided that a very strong level of understanding is displayed. However, you should avoid choosing complex methods without properly

demonstrating your understanding. 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.
4. We expect that your approach should be visual and clear within the report itself. Therefore it is highly recommended to include pieces of clear and well-written code along with necessary comments and explanations within the report itself.
5. We expect that you interpret and make sense of the experiment results obtained, instead of displaying a list of the results without explanation or analysis. A high quality report should be able to use the experimental results to support its conclusions and findings in a consistent manner.
6. We do not have a page limit for the report. A *rough* guideline is that your report should ideally be no more than 10 pages, *if* all figures and large pieces of code were removed. However, this is not a strict constraint. Again, clarity is highly important, and you should include sufficient details to demonstrate your approach and the level of understanding of the key concepts.

Data sets

There are a vast number of freely available data sets across the internet. Below are 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.