Examiners' commentaries 2022

ST2195 Programming for data science

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2021–22. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2021). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

General remarks

Learning outcomes

At the end of the course and having completed the essential reading and activities you should be able to:

- convert raw data to relational databases such as SQL
- import data to Python and R, apply data manipulation and visualisation
- program in Python and R
- develop software using version control via Git.

Planning your time in the examination

You have two hours to complete this paper, which consists of twelve compulsory questions, split over two sections. Remember that each of these questions is likely to cover more than one topic. This means that it is really important that you make sure you have a reasonable idea of what topics are covered before you start work on the paper! Also, keep in mind that the questions in Section A require a simple answer without justification and, as such, may not take as long as Section B. We suggest you divide your time as follows during the examination:

- Spend the first 10 minutes annotating the paper. Note the topics covered in each question and its subquestions.
- Allow yourself no more than 40 minutes for Section A and no more than 60 minutes for Section B. Then try to spend around 6–7 minutes for each of the Section A questions, and 10 minutes for each of the Section B questions. Don't allow yourself to get stuck on any one

question, but don't just give up after two minutes! Moreover, note that in Section A negative marks apply for some questions so, only for these questions, do not feel forced to put an answer that you are not confident about.

- This leaves you with 10 minutes. Do not leave the examination hall at this point! Check over any questions you may not have completely finished. Make sure you have labelled and given a title to any tables or diagrams which were required.

What are the examiners looking for?

In Section A only specific answers are required such as '(a) and (b)' or 'True'. In Section B, unless stated otherwise, the examiners are looking for very brief justifications from you. They want to be sure that you:

- have covered the syllabus as described and explained in the subject guide
- know the basic concepts given and, more importantly, when and how to use them
- understand and answer the questions set.

You are **not expected to write long essays** with lengthy explanations. However, clear and accurate language, both mathematical and written, is expected and marked. The explanations below and in the specific commentary for the examination paper should make these requirements clear.

Key steps to improvement

The most important thing you can do is answer the question set! This may sound very simple, but these are some of the things that candidates often do not do, though asked! Remember:

- If you are asked to label a diagram (which is almost always the case!), please do so. What do the data describe? What are the units? What are the x and y axes?
- Do not waste time calculating things which are not required by the Examiners.
- When asked to provide the necessary steps in 'R or Python or plain English', note that there is no need to provide code which is syntactically correct; it suffices to demonstrate the logic and structure in your suggested program.

How should you use the specific comments on each question given in the Commentaries?

We hope that you find these useful. For each question and subquestion, they give:

- further guidance for each question on the points made in the last section
- the answers, or keys to the answers, which the examiners were looking for
- where appropriate, suggested activities from the study material which should help you to prepare, as well as similar questions.

Any further references you might need are given in the part of the subject guide to which you are referred for each answer.

Memorising from the Examiners' commentaries

It is generally noted in similar examination papers that a small number of candidates appears to be memorising answers from previous years' *Examiners' commentaries*, for example plots, and therefore produce the exact same image of them without looking at the examination questions at all! Note that this is very easy to spot. The *Examiners' commentaries* should be used as a guide to practise on sample examination questions and it is pointless to attempt to memorise them.

Examination revision strategy

Many candidates are disappointed to find that their examination performance is poorer than they expected. This may be due to a number of reasons, but one particular failing is '**question spotting**', that is, confining your examination preparation to a few questions and/or topics which have come up in past papers for the course. This can have serious consequences.

We recognise that candidates might not cover all topics in the syllabus in the same depth, but you need to be aware that examiners are free to set questions on **any aspect** of the syllabus. This means that you need to study enough of the syllabus to enable you to answer the required number of examination questions.

The syllabus can be found in the Course information sheet available on the VLE. You should read the syllabus carefully and ensure that you cover sufficient material in preparation for the examination. Examiners will vary the topics and questions from year to year and may well set questions that have not appeared in past papers. Examination papers may legitimately include questions on any topic in the syllabus. So, although past papers can be helpful during your revision, you cannot assume that topics or specific questions that have come up in past examinations will occur again.

If you rely on a question-spotting strategy, it is likely you will find yourself in difficulties when you sit the examination. We strongly advise you not to adopt this strategy.

Examiners' commentaries 2022

ST2195 Programming for data science

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2021–22. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2021). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions

The examination consists of Section A (40 marks) and Section B (60 marks). There are 6 questions in each section, all of which should be answered. The questions in Section A are multiple choice type or require very short answers. For Section A, there is no need to provide justification for your answers, but note that negative marks may apply; see the individual questions to check if this is the case. Section B contains questions that either require justification for your answers or some extra effort in terms of programming. Please pay attention to the number of marks for each question and allocate your time accordingly.

Section A

This section contains six questions. Answer all six questions.

Question 1

Consider the data structures that are created by running the following Python script:

```
A = [2,4,3]
B = {1,2,1}
C = ('John','George',A)
D = {
  "brand":  "Mercedes",
  "model":  "B Class",
  "year":  2014
  }
```

Provide the output of the following commands if they were to be run after the Python script above. If you think the code will result in an error, just type 'error'.

(a) `A[1]`

(b) `print(B)`

(c) `C[2]`

(d) `print(D["model"])`

(6 marks)

Reading for this question

This is a question on data structures in Python. Make sure to cover the relevant notes on data types and structures and apply the corresponding code in them in your computer.

Approaching the question

The answers are given below:

(a) `4`.

(b) {`1,2`}.

(c) `[2,4,3]`.
Note: 0.5 marks were awarded if the response was `A`.

(d) `B Class`.

Question 2

For each of the circumstances below do state, with yes/no answer, whether side-by-side boxplots provide an appropriate choice. All subquestions carry equal weight. One mark will be deducted for each incorrect answer. The minimum mark for this question is zero.

(a) When we want to monitor changes in the distribution of a variable across different categories of another variable.

(b) When we want to explore the association between two continuous variables.

(c) When we want to compare frequencies of one variable's categories across different categories of another variable.

(d) When we want to study the empirical density of a single variable.

(6 marks)

Reading for this question

This is a question on data visualisation and graphics. Make sure to cover the relevant notes and apply the corresponding activities on your own.

Approaching the question

Note that negative marks apply for this question. This is done to discourage random guessing with yes/no answers. Hence, try to figure out the answer and write what you think it is. But if you are really not sure about the answer, it is probably better to leave it blank.

The answers are given below.

(a) Yes.

(b) No.

(c) No.

(d) No.

Note that a single yes or no answer is required as shown above. However, to provide some explanations for the purposes of this commentary, in (b) scatter plots are more appropriate, (c) would be suitable if a continuous variable was monitored across the categories of another variable, and (d) erroneously refers to a single variable.

Question 3

For each of the statements below state whether it is correct or incorrect. All subquestions carry equal weight. One mark will be deducted for each incorrect answer. The minimum mark for this question is zero.

(a) Random forests cannot be used for classification.

(b) Suppose that model A has lower test error than model B, but model A has higher training error than model B. We will choose model A.

(c) In classification we can assess the predictive performance by looking at the sensivity.

(d) In cross-validation all points will be selected as test points.

(6 marks)

Reading for this question

This is a question on machine learning frameworks. Make sure to cover the relevant notes and apply the corresponding activities on your own.

Approaching the question

Note that negative marks apply for this question. This is done to discourage random guessing with incorrect/correct answers. Hence, try to figure out the answer and write what you think it is. But if you are really not sure about the answer, it is probably better to leave it blank.

The answers are given below.

(a) Incorrect.

(b) Correct.

(c) Correct.

(d) Correct.

Perhaps it is worth clarifying in (b) that while a low training error is not necessarily a bad feature, minimising test error is the primary aim. Finally, as with all questions in this section, no justification is required, only whether each subsection is correct or incorrect.

Question 4

Which two of the following statements are correct? You may only give up to two answers.

(a) R is a programming language that works only on Windows.

(b) An active internet connection is necessary for the command `library(ggplot2)` to work in R.

(c) Some source-code editors provide code indentation and syntax highlighting.

(d) There are only 2 source-code editors for R and 5 for Python.

(e) There are source-code editors that can be used for both R and Python

(f) The help file of the `read.csv()` function in R can be accessed by typing `?read.csv`

(6 marks)

Reading for this question

This is a question covering mostly topics on code editors and their operation for Python and R. Candidates are advised to read the relevant notes.

Approaching the question

Note that the question asks to provide two correct answers. This does not mean that the correct answers in the list below are two, in fact there are three. To get the marks for this question candidates must provide any pair of correct answers out of these three. But if an incorrect answer was provided, no marks were given.

The answers are given below.

(a) Incorrect.

(b) Incorrect.

(c) Correct.

(d) Incorrect.

(e) Correct.

(f) Correct.

As with all questions in this section, no justification is required, only the answers '(c), (e) and (f)'. If more than two answers were provided, the first two were chosen.

Question 5

Which three of the following statements are correct? You may only give up to three answers.

(a) R notebooks is an authoring framework that combines Markdown with R

(b) Jupyter notebooks can handle Julia and Python code.

(c) R Markdown notebooks are special Jupyter notebooks.

(d) Jupyter notebooks were named after the first names of its creators, Julia and Peter.

(e) R Markdown files can only be converted to HTML.

(f) The author and date of an R Markdown file can be specified in the YAML metadata.

(6 marks)

Reading for this question

This is a question covering R notebooks, R Markdown and Jupyter notebooks. Candidates are advised to read the relevant notes.

Approaching the question

Note that the question asks to provide three correct answers. This does not mean that the correct answers in the list below will necessarily be three, although there were three on this

occasion. To get the marks for this question candidates must provide all three correct answers. If an incorrect answer was provided, no marks were given.

The answers are given below.

(a) Correct.
(b) Correct.
(c) Incorrect.
(d) Incorrect.
(e) Incorrect.
(f) Correct.

As with all questions in this section, no justification is required, only the answers '(a), (b) and (f)'. If more than three answers were provided, the first three were chosen.

Question 6

Note from which language (R or Python) each of the following code chunks is from. Give your answer in the form 'question number, language', e.g. 'C11, R'.

C1. `import pandas`
C2. `ggplot(df, aes(x = v1, y = v2))`
C3. `plot(log, 0, 10)`
C4. `typeof("a")`
C5. `type("a")`
C6. `blurp = "Coding rocks"; blurp.upper()`
C7. `read.table("df.csv")`
C8. `sapply(1:10, "-", 1)`
C9. `filter(df, Course == "ST2195")`
C10. `df[-c(1, 3, 4), ]`

(10 marks)

Reading for this question

This is a question asking to distinguish R from Python code. There is no specific place in the notes for it, this is just something that comes with practice on writing code in R and Python.

Approaching the question

The answers are given below.

C1. Python.
C2. R.
C3. R.
C4. R.
C5. Python.
C6. Python.
C7. R.
C8. R.
C9. R.
C10. R.

Section B

This section contains six questions. Answer all six questions.

Question 1

For each of the following statements about R, state if they are correct or not. Provide justification for your answer of no more than two sentences.

(a) A list contains vectors of different lengths.

(b) A data frame is also a matrix.

(c) A data frame is also an array.

(d) A vector is also a matrix.

(e) A matrix is also an array.

(10 marks)

Reading for this question

This is a question covering R data types and structures. It is essential to read the relevant notes and get a clear understanding of types such as lists, vectors, matrices, data frames and arrays.

Approaching the question

In all questions of this section justification is required, unless explicitly stated otherwise, and no marks are awarded if it is missing. Another thing to note is the rule of two sentences; answers with more than two sentences were penalised. The questions are designed so that they can be answered in a brief manner. In fact, in the answers given below, only one sentence was used.

(a) Correct. Lists can even contain elements of different types let alone different lengths.

(b) Incorrect. Data frames can have different data types unlike matrices.

(c) Incorrect. Data frames can have different data types unlike arrays.

(d) Correct. A vector of length n, can be viewed as an $n \times 1$ matrix.

(e) Correct. A matrix is also a 2-dimensional array.

Question 2

For each of the following statements about R, state if they are always correct or not. Provide justification for your answer of no more than two sentences.

(a) The rows of a table in a relational database represent attributes.

(b) The records in a table of an SQLite database cannot be altered.

(c) SQLite uses a separate server process to operate.

(d) The SQL query adds the record with attributes 'KitKat', 'A', and 1 in the table Products.

```
INSERT INTO Products VALUES("KitKat", "A", 1)
```

(e) The following R code chunk finds all records in the data frames Products and Sales that have matching values of Name, and returns only those records where Number is greater than 1 or Category is not equal to 'A'.

```
inner_join(Products, Sales, by = "Name") %>%
  filter(Number > 1, Category != "A")
```

(10 marks)

Reading for this question

This question covers databases. It is essential to read the relevant notes, conduct the corresponding activities and practise.

Approaching the question

As before, justification is required and no marks are awarded if it is missing. Another thing to note is the rule of two sentences; answers with more than two sentences were penalised. The questions are designed so that they can be answered in a brief manner. In fact, in the answers given below, only one sentence was used.

(a) Incorrect. The rows of a table in a relational database represent records.

(b) Incorrect. The records in a table of an SQLite database can be altered (for example, deleted, updated etc.).

(c) Incorrect. SQLite reads and writes to a single file on your disk.

(d) Correct. Expect some text here.

(e) Incorrect. '... is greater than 1 and Category is not equal to 'A'.'

Question 3

Explain in no more than 2 sentences, why the following statements are wrong.

(a) RStudio is the IDE that R comes with.

(b) pandas allows the removal of items in a tuple in Python.

(c) The only way to create a Git repository is in Github.

(d) Adding a double to a vector of doubles is not supported in R.

(e) Base Python does not allow the creation of a list of lists.

(f) A matrix in R is immutable.

(g) The command `git commit` is used to let your collaborators know that you will make an edit.

(h) The command `[[1,2,3,4],[3,2,1]]` in Python will create a matrix with 4 rows and 3 columns.

(i) `plot()` in R can only be used to produce lineplot graphics.

(j) Structured data is all data that can be opened in Python.

(10 marks)

Reading for this question

This is a question asking several things related to the course ranging from version control, data types and structures, code editors, IDEs and plots. There is no specific place in the notes for it, just review the relevant notes to the best of your ability.

Approaching the question

As before, justification is required and no marks are awarded if it is missing. Another thing to note is the rule of two sentences; answers with more than two sentences were penalised. The questions are designed so that they can be answered in a brief manner. In fact, in the answers given below, only one sentence was used.

(a) RStudio is an IDE for R that needs to be installed independently from R.

(b) A tuple in Python is immutable.

(c) A Git repository can be set up and accessed on a local computer.

(d) Adding a number to a vector of numbers in R will add the number to each element of the vector.

(e) A list of lists is a valid Python object.

(f) A matrix in R is mutable.

(g) The command `git commit` captures a snapshot of the project's currently staged changes.

(h) No, this command will create a list.

(i) `plot()` in R is a method that can construct different graphics depending on the class of its input.

(j) Structured data is data that is organised according to a predetermined set of rules; unstructured datasets are datasets for which it is difficult to have a predetermined set of rules for organising them.

Question 4

Match the commands C1–C4 with the output in O1–O4. There is no need to provide justification in this question.

C1. `c(1,2,3)`

C2. `[1,2,3]`

C3. `ifelse(-2, "A", "B")`

C4. `ifelse(is.character(1), "A", "B")`

O1. `1 2 3`

O2. `A`

O3. `B`

O4. `[1, 2, 3]`

(10 marks)

Reading for this question

This is a question on R and Python code. There is no specific place in the notes for it, just review the relevant notes to the best of your ability.

Approaching the question

Not that the question clearly indicates that there is no need for justification. Hence, despite being in Section B, there is no need to provide justification; only the matchings are required.

The answers are given below.

(a) C1–O1.

(b) C2–O4.

(c) C3–O2.

(d) C4–O3.

Question 5

Consider a data set consisting of several flat sales in a city over the past year. The data contain the following variables:

- `Price`: **the price of the flat sale**
- `Station`: **whether a bus or a train station is nearby (1: yes, 0: no)**
- `Floor`: **0: underground, 1: ground floor, 2: 1st floor, 3: 2nd floor or above**
- `Building_age`: **years since the building was first erected.**

(a) Describe what graphs you would produce to demonstrate how the presence of a bus or train station, floor and building age may affect the flat sale price.

(b) How would your answer on part (a) change if `Floor` was considered as a continuous variable?

(10 marks)

Reading for this question

This is a question on graphics and data visualisation. It requires going through all the notes in the relevant block. In addition to knowing what different plots depict, it is also essential to have an idea on when they are useful and how they can be used in the context of data analysis on specific domain-based questions.

Approaching the question

There is no unique answer for this question, hence an indicative one is provided below.

(a) A scatterplot of `Price` with `Building_age` could be used to get some idea of the kind of association between these two (positive/negative, linear/non-linear). Separate versions of this plot with the points labelled according to `Station` and `Floor` will also be useful to explore whether this association changes across these categories. Also, side-by-side boxplots, violin or ridgeline plots of the `Price` variable to monitor its distribution across the categories of `Station` and `Floor`.

(b) It could be useful to produce a matrix scatterplot based on the variables `Price`, `Floor` and `Building_age`, to explore the association between each of these pairs. The points labelled with different colours according to the `Station` variable to check if any of these association changes in flats with and without a station nearby. The side-by-side plot of the `Price` variable across the categories of `Station` will still be helpful.

Marks were awarded for a sensible choice of plots, mention of the appropriate variables, understanding of the scale of the variables and also understanding of what to look for in the plot.

Question 6

Consider the following numerical vectors:

```
Temp = [12,50,70,25]
Type = [1,0,0,1]
```

The first vector contains temperatures whereas the second vector indicates whether these degrees are in Celsius or Fahrenheit units (0: Farhenheit, 1: Celsius). Using a `for` loop and a conditional statement, write a program (could be either R or Python or just the necessary steps in plain English) that transforms the vector `Temp`, so that it only contains Celsius degrees, and records this change into the vector `Type` as well. You can use the following formula for converting from Fahrenheit to Celsius:

$$\textbf{Celsius} = \frac{\mathbf{5}}{\mathbf{9}}(\textbf{Fahrenheit} - \mathbf{32}).$$

(10 marks)

Reading for this question

This is a question on programming and in particular on structuring computer programs. All the relevant sections are useful with focus on iterations and conditional statements (also a little hint of data types).

Approaching the question

There is no need to actually write code that works in R or Python, in fact there is no need to write in any of these two languages, unless you find it convenient. What the examiners are looking for is the correct use of a loop, correct application of conditional statement and, overall, a script that is correct logically (not syntactically). Marks were also allocated for changing the vectors `Temp` and `Type`.

There is no unique answer for this question, hence an indicative one is provided below:

```
for i=1:4
    if Type[i] = 0
      Temp[i] = (5/9) *( Temp[i] - 32)
      Type[i] = 1
    end if
end for
```