ST2195

ST2195 Programming for data science

The exam 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.

Please find questions on the following page.

ST2195: Zone A

Section A

This section contains six questions. Answer all six questions.

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

```
A = ['rain',2]
B = {
  "brand": "Ford",
  "model": "Focus",
  "year": 2016
  }
C = {'George','John','John'}
D = (2,3,A)
```

   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["year"])`
   (c) `print(C)`
   (d) `D[2]`

- *Marks*: 6

2. For each of the circumstances below do state, with a yes/no answer, whether side-by-side violin plots 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 study the empirical density of a single variable.
   (b) When we want to compare frequencies of one variable's categories across different categories of another variable.
   (c) When we want to explore the association between two continuous variables.
   (d) When we want to monitor changes in the distribution of a variable across different categories of another variable.

- *Marks*: 6

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) In classification we can assess the predictive performance with ROC curves.
   (b) In cross-validation some points may never be selected as test points.
   (c) Random forests can be used for classification.
   (d) Suppose that model A has lower training error than model B, but model A has higher test error than model B. We will choose model B.

- *Marks*: 6

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

   (a) The command `library("VGAM")` throws an error if the `"VGAM"` package has not been installed.
   (b) R is a programming language that works only on Windows and macOS.
   (c) There are only 4 source-code editors for R and 4 for Python.
   (d) Some source-code editors provide syntax highlighting and code auto-completion.
   (e) The help file of the `read.csv()` function in R can be accessed by typing `help read.csv`
   (f) There are source-code editors that can be used for both R and Python.

- *Marks*: 6

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

   (a) Jupyter notebooks can handle R and C++ code.
   (b) R Markdown is an authoring framework that allows adding marks in R code.
   (c) Jupyter notebooks were named after the languages Julia, Python and R.
   (d) The title of an R Markdown file can be specified in the YAML metadata.
   (e) Jupyter notebooks are special R Markdown notebooks.
   (f) R Markdown files can only be converted to PDF.

- *Marks*: 6

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.

```
for (i in 1:5)
    print(i)
```

   C2.

```
for i in [1,2,3,4,5]:
    print(i)
```

   C3. `a.shape`

   C4. `library(ggplot2)`

   C5. `[("a", "b", "c"), 2, (4, 3)]`

   C6. `df = pandas.read_csv('test.csv')`

   C7. `library("ggplot2")`

   C8. `v = (1, 11, 111, 11, 1)`

   C9. `list(33, c(1, 2), 1, 3)`

   C10. `head(df, 15)`

- *Marks*: 10

Section B

This section contains six questions. Answer all six questions.

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 data frame is also an array.
   (b) A matrix is also an array.
   (c) A list can contain matrices of different dimensions.
   (d) A vector is not a matrix.
   (e) A data frame is also a matrix.

- *Marks*: 10

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 connections to other tables.
   (b) SQLite does not require a server to operate.
   (c) The records in a table of an SQLite database can be deleted.
   (d) The SQL query adds the record with attributes "Buttons", "C", and 2 in the table Sales.

```
INSERT INTO Sales VALUES("Buttons", "C", 2)
```

   (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 and Category is not equal to 'A'.

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

- *Marks*: 10

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

   (a) numpy allows the addition of items in a tuple in Python.
   (b) A list in Python does not allow items with the same value unless they are one after the other.
   (c) `y | x`, in both R and Python, will return True only if both `y` and `x` are True.
   (d) A list in R is ordered and immutable.
   (e) A Git repository cannot be accessed without an internet connection.
   (f) `plot()` in R can only be used to produce scatterplot graphics.
   (g) The command `[[1,2,3],[3,2,1]]` in Python will create a matrix with 3 rows and 2 columns.
   (h) The command `git commit` is used to record a milestone in a project.
   (i) Structured data is all data that can be opened in Microsoft Excel.
   (j) RStudio is a special version of R.

- *Marks*: 10

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, 3, 2, 4)`

   C2. `[1, 3, 2, 4]`

   C3. `sapply(1:4, "*", 2)`

   C4. `[1, 2, 3, 4]*2`

   O1. `[1, 2, 3, 4]`

   O2. `1 3 2 4`

   O3. `2 4 6 8`

   O4. `[1, 2, 3, 4, 1, 2, 3, 4]`

- *Marks*: 10

5. Consider a data set consisting of several measurements on the quality of some wines produced over the last year in a country. The data contain the following variables:

   - `Price`: the price of the wine per litre
   - `Colour`: the colour of the wine (1: Red, 2: Rose, 3: White)
   - `Evaluation`: An evaluation from an independent committee (1: Not good enough, 2: so and so, 3: good, 4: excellent)
   - `Acidity`: a measurement of acidity (in pH) in 2 decimal places.

   (a) Describe what graphs you would produce to demonstrate how the colour of the wine, evaluation and acidity may affect the price of the wine.
   (b) How would your answer on part (a) change if `Evaluation` was considered as a continuous variable?

- *Marks*: 10

6. Consider the following numerical vectors:

```
Temp = [32,90,80,23]
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 script (could be either R or Python or just the necessary steps in plain English) that transforms the vector `Temp`, so that it only contains Fahrenheit degrees, and records this change into the vector `Type` as well. You can use the following formula for converting from Celsius to Fahrenheit:

$$\text{Fahrenheit} = \frac{9}{5}\text{Celsius} + 32$$

- *Marks*: 10