MATH40005 Coursework Spring 2023

Overview of coursework

The goal of this coursework is to use the R programming language to perform an analysis on a dataset and create a report. A report must be submitted in PDF format on Blackboard. The report must contain your College CID number and must be named correctly (see the Submission and deadline section below).

All analysis must be done using the R programming language and all code used to create any figures/compute any statistics must be included in the report. If the code is not included, no marks will be awarded for that (part of the) question. For this reason, using the provided R Markdown template is highly recommended.

If you use the R Markdown template, make sure to modify the `author` field at the top of the document to contain your **College CID number**. Do not include your name anywhere on the coursework, only include your CID.

Submission and deadline

The report must be submitted as a PDF document named: `MATH40005_CID_coursework.pdf` where `CID` is replaced with your College CID number, e.g. `12345678`.

In the Assessments and Mark Scheme folder on Blackboard there is a **Coursework** Turnitin folder where the file can be submitted.

The deadline for submission is **Thursday 9 March 2023, 13.00 UK time**.

Note that this assessment will count **7%** towards the final grade for this module.

Files provided

On Blackboard the zip file `MATH40005_Coursework_Spring_2023.zip` contains:

- The dataset `salaries.txt`
- An R Markdown template `MATH40005_CID_coursework.Rmd`

Notice that the R Markdown template provides code chunks that may be helpful to you, but you also may not need to use. Please be sure to read the Frequently Asked Questions section on the last page of this document.

Presentation

Note that one mark will be awarded for presentation. For example:

- Your CID should be at the top of the script,
- The answer to each question should start on a new page,
- Brief commments to clarify the code can be included.

The R Markdown template includes code which starts each question on a new page.

Coursework description

The goal of this coursework is to use the R programming language to perform an analysis on a dataset described below and create a report.

Data

The dataset in the file `salaries.txt` contains the partial results of fictional surveys over the past ten years for data scientists working in different cities across the globe. The dataset has the following five columns:

- `id`: a five-digit number anonymising the respondent,
- `date`: the date the survey was completed,
- `age`: the age of the respondent,
- `salary`: the annual salary of the respondent (in £),
- `city`: the city in which the respondent lives.

The annual salaries for respondents in New York, London, Paris and Singapore are included (all values are in British pounds) along with their ages and the date on which they completed the survey. Each respondent also has a respondent ID which anonymises the respondent.

For example, the first respondent in the dataset is a data scientist in Singapore, who is 23 and completed the survey on 10 October 2013, and earned £41,234 that year.

Task

The goal is to analyse the dataset `salaries.txt` using the R programming language by completing the questions below, and provide the answers in the form of a report. All R code used to answer the questions must be included. It is strongly suggested that an R Markdown document is created.

Academic integrity

You are welcome to refer to any sources, so long as you cite sources that you have used. You are welcome to discuss this work with other students on the module, but you should write your own submission, including all code.

Part A (1 mark)

Read the dataset `salaries.txt` into a dataframe `df`.

Hint: it may be necessary to carefully inspect the first few lines of the file in order to determine how to read in the data correctly.

Part B (1 mark)

Print the **median** of the values in the `salary` column of the `df` dataframe, as follows:

```
The median salary is: 12345
```

if `12345` had been computed as the median salary.

Hint: the `cat` function is helpful for printing text and values to screen.

Part C (2 marks)

Using an appropriate type of plot, create a figure showing the number of respondents from each city.

Part D (1 mark)

Using an appropriate type of plot, create a figure showing the distribution of the ages of all the respondents.

Part E (2 marks)

Compute the mean salary, across all age groups, for respondents in each of the four cities. For each city, output the information to screen in the form:

```
Mean salary for CITY is 12345.67
```

where `CITY` is replaced with the city name, and the computed mean (e.g. `12345.67`) is **rounded to two decimal places**. Since there are four cities, there should be four such lines as output.

Part F (3 marks)

Save the respondents from London in the dataframe `df_L`. For the salaries of respondents in London, find any outliers in this dataset, explaining the criterion used to do this, and use an appropriate plot to display the distribution of salaries from respondents in London along with any outliers.

Part G (3 marks)

Suppose that after completing their degree, one of your friends wants to go and work as a data scientist in either New York or Singapore, but cannot decide which city to go live in. Your friend thinks it would be better to go to the city which has the highest mean salary for 20-29 year old data scientists.

From the dataframe `df`, create a sub-dataframe `df_NY` which contains all the rows of `df` where the respondents are from New York **and** are aged between 20 and 29 (inclusive).

Similarly, from the dataframe `df`, create a sub-dataframe `df_S` which contains all the rows of `df` where the respondents are from Singapore **and** are aged between 20 and 29 (inclusive).

Compute and print the means of the salaries in each dataframe `df_NY` and `df_S` to screen to two decimal places; the output must be in the format:

```
Mean salary for respondents aged 20-29 in New York: 34567.89
Mean salary for respondents aged 20-29 in Singapore: 65432.10
```

where the numbers `34567.89` and `65432.10` are replaced with the computed sample means.

Part H (3 marks)

Your friend does not want to rely on only the sample means to make a decision, and wants to use Student's *t*-test to test for a significant difference. However, one of the assumptions of this test is that the data are observations of normal random variables.

Create an appropriate plot to help you decide if the salaries in `df_NY` follows a normal distribution. Interpret the plot and state whether you think the data follows a normal distribution or not, providing justification.

Similarly, create an appropriate plot to help you decide if the salaries in `df_S` follows a normal distribution. Interpret the plot and state whether you think the data follows a normal distribution or not, providing justification.

Part I (3 marks)

Write your own function called mytest to implement Student's two-sample *t* test. The function should take arguments x (first sample vector), y (second sample vector) and alpha (significance threshold).

Note that you may not use the built-in R function t.test or any other function which directly computes the *t*-test.

Regardless of the outcome of Part H, let's assume the salaries in df_NY and df_S follow the required assumptions for Student's *t*-distribution, and let's assume that both samples have equal (but unknown) variance, and further assume we use a significance threshold of $\alpha = 0.05$. Using your function mytest, compute the result for Student's two-sample *t*-test for the data in df_NY and df_S, where the null hypothesis is that the means of the two samples are equal.

Print the output to screen in the format:

```
t-statistic: 2.345
alpha: 0.95
threshold: 1.234
decision: reject
```

where the values above are replaced with the appropriate values (and if the decision is not to reject the null hypothesis, then the last line will say `decision: fail to reject`).

Presentation (1 mark)

1 mark will be awarded for presentation: correct layout and formatting, each question starts on a new page, CID at the top of the script, and if there are appropriate comments included.

Total: 20 marks

Frequently Asked Questions

Do I need to use R Markdown?

While you are not required to use the R Markdown template, it is **highly encouraged** because this makes it easy to include code, figures and text all in one document. Note that you may want to look at Problem Sheet 10 to become familiar with R Markdown, as well as the accompanying videos on Blackboard. If you do not want to use R Markdown, you can include the figures, code and text in a document using a word processor such as Microsoft Word; however, this may end up taking you more time.

Do I need to include my R code?

All R code used in the analysis must be included. The format for your report must allow for easy copy/pasting of your code into an R console. Saving your code as an image (picture) and including that in an MS Word document will not be acceptable, because it will not be possible to copy/paste and check your code. **If your code is not included, then that part of the question will not receive any marks.**

What type of file must be submitted?

The submission file must be in PDF format.

Can I do the analysis in Excel?

The analyses **must be done using the R programming language**, and the code for each question (or part of the question) must be included along with the results. If the R code is not included, no marks will be awarded to that part of the question.

How long will it take to do this coursework?

The coursework is designed to be completed in 2-3 hours if a student has been keeping up to date with the lectures and problem sheets. Although you have two weeks to do the coursework, you are encouraged to start the coursework within the first week to allow for enough time to complete it.

Can I work on this with another classmate/in a group?

This is an individual assessment and so **you must work on the coursework on your own**, and abide by the College's policy regarding collaboration on assessed work.

Can I get an extension?

For extensions, please contact the Senior Tutor.