STAT40840 Data Programming with SAS (online) - Final Project

John O'Sullivan

- Submission date: see Brightspace.
- There is a maximum of 50 marks for this assignment. This assignment is worth 50% of your final grade.
- The marks available for each question are shown in brackets below.
- Late submissions will score 0, unless an extenuating circumstance form has been submitted and approved.
- You must submit two files only: **a single pdf** containing all required tables (please don't print full tables at any step) and **all answers**; and **a .sas programme file** containing all code used to answer the assignment.
- Code should run for me without errors in order to reproduce the analysis in your pdf. Code should be clearly commented throughout.
- Please do not email solutions to me. Any issues relating to upload to Brightspace must be identified clearly and well in advance of the deadline.
- You can re-submit until the deadline; only your most recent submission will be graded.
- Late submissions will be penalised by 5/50 up to 1 hour late, 10/50 up to 24 hours, 25/50 up to 48 hours and no grades will be awarded for work that is more than 48 hours late.
- Note that this is a project: differently from a final exam, further reading may become necessary to complete some parts.
- Please carefully read the school plagiarism guidelines accessed via the following link: `https://maths.ucd.ie/docserve?id=235`
- Please carefully read the SMS Honesty Code included on Brightspace under My Learning → Learning Materials → Module Information. Include a statement with your submission that you have read, understood, and agree to this code.

Plagiarism:

While you are encouraged to ask about the module material, this project should be completed individually. Any student who plagiarises will receive a 0 mark. Projects will be reviewed by the UCD plagiarism software. If you are unsure whether a question about the project would be considered as plagiarism, please email the question to the lecturer rather than posting on the discussion forums. The UCD Plagiarism Policy applies to all students. This can be consulted here:https://www.ucd.ie/secca/t4media/plagiarism_studentguide.pdf.

Final Project:

You should create all SAS tables in a library called s40840. This way, when I run your script, I should only need to change your first file path and no other lines or filenames.

Write steps in a single SAS programme file to answer all questions below, ensuring to comment your code clearly. Include answers in the pdf output as footnotes, where required.

Report [5 marks]

Complete your assignment using SAS OnDemand, and check that all of the output and comments are correctly and nicely shown in your final document. Once completed, submit the `.sas` file and the resulting `.pdf` document. Your pdf file must show answers to all questions, and function as a stand-alone document.

Data Analysis I [15 marks]

This task involves finding a dataset of interest to you, that contains a mix of categorical and numerical variables. As a guideline, the dataset would typically have a minimum of two categorical variables and three numerical variables; these miminum criteria are guidelines and not hard thresholds. Do not use an in-built SAS dataset or a dataset from kaggle. Do not discuss your choide of dataset with other students - datasets should be individually selected.

If you wish, you can make use of the following websites to find a dataset:

- The Irish government data repository: https://data.gov.ie/
- Google dataset search: https://datasetsearch.research.google.com/

The exercise then is to use the methods covered in this course to complete an analysis and write a report using SAS procedures on the data. The analysis of the data should involve the use of graphical summaries, tables and numerical summaries of the data. (This analysis only needs to include descriptive statistics - you do not need to perform hypothesis tests or any other inferential statistics tehcniques for this part of the project.)

This part of the project will be assessed in terms of:

- Using the functionality and settings of the appropriate procedures in SAS.
- Clearly annotating the code in the `.sas` file.
- Producing clear results for the data.
- Quality of the graphics included.
- Summarising the conclusions from the analysis appropriately.

Data Analysis II [20 marks]

The dataset `erasmus.csv` is in the `Final_Project` folder inside the shared course folder on the SAS OnDemand server. It contains information on individuals who undertook an Erasmus study programme between 2014 and 2020. Each row corresponds to a different individual.

1. Read the dataset `erasmus.csv` into SAS and call the resulting table `erasmus`, saving it in the `s40840` library. The file contains column names on the first row, with the first observation starting on the second row. You should ensure your code will overwrite any previous object of the same name.

   (a) Print the first 4 rows of the resulting `erasmus` table.

   (b) The duration variable is stored in months. Find the mean duration spent in the programme by students of Irish nationality ('IE'). How many students of Irish nationality are in the dataset?

   (c) One student is older than all other participants. What is the age of this student? In what city did this student study? In what academic year did they start?

   (d) Create a table of the nationality variable for students who are not from Ireland (that is, their nationality is not 'IE') and whose receiving city is Dublin. The table should be ordered from highest to lowest frequency. What is the most frequent nationality of non-Irish students who studied in Dublin?

   (e) In a single table, print the summary statistics for the age variable, divided into groups by both gender and academic year. Which cohort had the greatest mean age?

   (f) Produce a clustered bar chart of the 'academic year' variable, clustered by gender. Describe the resulting plot.

2. For this question, create a subset of the `erasmus` dataset which contains only those individuals whose receiving country is Ireland ('IE'). Call this subset `erasmus2` and use this subset for all of the follwing parts:

   (a) Conduct a univariate analysis of the `age` variable for those individuals in `erasmus2`. Write a short description of your findings, including key statistics and discussion of any plots produced.

   (b) Create boxplots of the `age` variable in `erasmus2`, grouped by `gender`. Ensure the plot is neat with an appropriate title etc. Comment on the resulting plot.

   (c) Conduct a hypothesis test to see if there is a statictially significant difference between the mean ages of female and male students, using as your sample data those students in `erasmus2`. State your hypotheses carefully, check all assumptions necessary, run your chosen test, comment on the resulting plots and state your conclusion clearly. Use a significance level of $\alpha = 0.01$.

`Tasks` demonstration [10 marks]

This exercise involves navigating to the following menu in the SAS OnDemand interface:
`Tasks and Utilities` → `Tasks`,
and then selecting one of the folders from the approximately 15 Tasks here. Note: you cannot select the `Data` or `Graph` folders, as we studied these in some detail in class. Also note that you could choose to explain and demonstrate the `IML` procedure with examples - this is found under `Snippets` → `Snippets`.

You must then explain the purpose of your chosen task and write a small report demonstrating its use. The report should demonstrate some of the key functionality of your chosen task, but doesnt need to demonstrate all of the functions (only the main ones).

This part of the project will be assessed in terms of:

- Clearly summarising the purpose of the task.
- Clearly demonstrating the functionality of some of the main functions in the task on appropriate data.
- Clearly commented code in your `.sas` script.
- Clearly showing the output for the demonstration examples in your report.