
Department of Artificial Intelligence
Faculty of ICT
University of Malta

Dr Jean-Paul Ebejer
jean.p.ebejer@um.edu.mt
Mr Joseph Bonello
joseph.bonello@um.edu.mt

ICS5115: Statistics for Data Scientists

Study Unit Assignment

Version 1.01, 10th March 2019
(Last Updated: April 1, 2019)

Necessary Preamble: This document describes the assignment for study unit *ICS5115: Statistics for Data Scientists*. Please read it thoroughly. The marks allocated for this assignment are distributed as follows; **85%** are for the implementation and documentation of the specified task set out in this document, and **15%** are allocated for the presentation of your solution. You are expected to allocate approximately 70 hours to complete the assignment. The deadline for this assignment is **Friday, 21st of June, 2019 at noon**. Late submissions will **not** be accepted. Questions regarding the assignment should **only** be posted in the Assignment VLE forum, and should **not** be emailed directly to the lecturers of this study-unit. This is an **individual** assignment. Under **no** circumstances are you allowed to share the design and/or code of your task implementations (unless otherwise explicitly stated by the task specification). The Department of Artificial Intelligence takes a strict and serious view on plagiarism. For more details refer to plagiarism section of the Faculty of ICT website[1]. You are to upload all of your developed code and documentation on the VLE website. Additionally, you are required to submit a hard-copy of your documentation. You must also submit a signed copy of the plagiarism form to the secretary's office at the Department of Artificial Intelligence. Unless you do this, your assignment will **not** be graded. Note that **all** these deliverables have to be submitted by specified deadline. Uploading to VLE, and not submitting physical copies (or vice-versa) will result in your assignment **not** being graded.

[1] https://www.um.edu.mt/ict/Plagiarism

The main objective of this assignment is to demonstrate you understood the statistical concepts (in `R`) presented in class. Failure to provide evidence of this will have an adverse effect on your mark! You have to demonstrate an appropriate level of depth and understanding befitting of postgraduate study at an M.Sc. level.

1 Deliverables

The following deliverables are expected by the specified deadline. Failure to submit any of these artefacts in the required format will result in your assignment **not** being graded. Replace `NAME` and `SURNAME` with your name and surname respectively (doh!). Replace `IDCARD` with your national id card number (without any brackets), e.g. `121998G`. If you are a visiting student, and hence have no Maltese national identity card number, use your passport number instead. The **three** deliverables are:

- **`201819_ICS5115_SURNAME_NAME_IDCARD_assignment_code.zip` -** A `.zip` file containing your assignment code and data. This needs to be uploaded to VLE. It is your responsibility to make sure that this archive file has uploaded to VLE correctly (by downloading and testing it). Failure to open the archive file will result in your assignment **not** being graded. **Please include a `README.txt` file with additional library requirements and installation and running instructions in the top level directory of your archive file submission**. Please note that if your data (only) is too large to upload to VLE due to upload size limitation (presently 50MB), you should commit your data (no code) online (in github, google drive, or dropbox only) and specify the (accessible) URL in your documentation. Make sure this URL is not publicly available.
- **`201819_ICS5115_SURNAME_NAME_IDCARD_assignment_doc.pdf` -** The assignment documentation in `.pdf` format. The documentation has to be uploaded to VLE, together with your code. A hard-copy should be submitted to the secretary's office at the Department of Artificial Intelligence by the stipulated deadline.
- **Signed copy of the plagiarism form -** This should be submitted to the secretary's office (Ms Francelle Scicluna at Level 1, Block A, Room 4, ICT Building) at the Department of Artificial Intelligence.

1.1 A Note on Presentation

There are no marks assigned to the presentation of your assignment's documentation. However we expect your work to be of adequate postgraduate-study level all-around (including presentation). Your physical assignment

should be soft bound (no loose papers). Figures and tables should have captions and be numbered, should have axes labels (and units), and should be linked to the text and discussion. Mathematical formulas should be labelled and explained. References should be plentiful and in a correct (i.e. consistent) format. Sectioning should be logical and clearly labelled. Furthermore, we expect you to proof-read your assignment carefully. Please refer to the documentation section for further guidelines on the documentation. Consider each assignment as a training opportunity for when you are writing your M.Sc. dissertation.

2 Technical Specification

Any code you supply will be run on Linux (distribution Ubuntu 18.04.2 LTS). It is **your** responsibility to make sure your code runs on this OS platform. You are required to implement the assigned task using `R` (version 3.5.x). `R` libraries such as `ggplot2` will be installed on the assessment machine. Any other dependencies should be specified in a `README.txt` file as specified earlier. Use of an IDE, such as `RStudio`, is allowed but make sure that your code runs from the command line. At the beginning of each `R` script, you should briefly describe the main functionality of your script. Make sure to have plenty of comments. Note that you should **not** have any absolute paths hardcoded in any of your programs.

3 The Tasks

In this assignment you are required to apply statistics in practical Data Science tasks using `R`. You will be given three research questions. Choose **one** and investigate this thoroughly:

1. **What influence (if any) has Brexit had on financial markets in particular GBP/EUR exchange rates?** – Correlate major events (related to Brexit) happening in the UK to fluctuations in the currency. Determine if these events have effected the price of GBP/EUR trade in a significant manner. Have major cryptocurrencies been robust to these events? For this task you will need data from news agencies and foreign exchange markets.
2. **It is easier to make it to the top post in Hackernews using some topics instead of others** – Y-Combinator's Hackernews[2] is a news aggregation site for the technically-abled. People vote for the most interesting articles which make it to the top of the ranking. Some topics appear to be overrepresented in the top post. You are required to verify whether this is

[2] `https://news.ycombinator.com/`

true. Also, are there particular topics which make it more often to the front page?

3. **Markov Chain modelling – Stepping Stone Model** This question is inspired by studies in genetics[3]. It is called "The Stepping Stone Model". In this model we have an n-by-n array of squares, and each square is initially anyone of k different colours. For each step, a square is chosen at random.

   This square then chooses one of its eight neighbours at random and assumes the colour of that neighbour. It is assumed that the array is rolled out over a doughnut, such that if a square S is on the left-hand boundary, but not at a corner, then it is adjacent to the square T at the right-hand boundary in the same row. Note that under this arrangement, S would also be adjacent to the squares just above and below square T. The same assumption also holds for squares on the upper and lower boundaries.

   With these adjacencies, each square in the array is adjacent to exactly eight other squares. A simulation of this concept can be found online[4]. At any time, the probability that a particular colour will win out is equal to the proportion of the array of this colour.

   A state in this Markov Chain is a description of the colour of each square. For this Markov Chain, be aware that the number of states is k^{n^2}, which for even a small matrix is enormous. This is an example of a Markov chain that is easy to simulate but difficult to analyse in terms of its transition matrix.

   You are required to build a simulation of the Stepping Stone Model in R and then model the problem using a Markov Chain. Using the model, determine the position of the matrix at a particular time τ.

The aim of this assignment is that you run a descriptive and inferential statistical analysis on the collected data, and build enough evidence to support your conclusions. Note that your conclusions to the specified tasks may be for, against or inconclusive for a specific statement (this does not apply for Task 3).

Some required, critical steps (where applicable) in this assignment are:

a) Collect data for the task;

b) Augment this with data from another source – variety;

c) Clean noisy data;

d) Visualise your data (e.g. using `ggplot2` or other `R` graphical libraries). Visualisations should be appropriate, non-redundant and of different types and they should be used to support your statements;

[3] Sawyer, Stanley. (1976). Results for the Stepping Stone Model for Migration in Population Genetics. Ann. Probab. 4.

[4] `https://math.dartmouth.edu/archive/m20x11/public_html/test.html`

e) Apply meaningful descriptive statistics for the task;

f) Perform hypothesis testing to further support your conclusions. Build a null model, and compare the real data to the null model.

Note that in your documentation you should also discuss strategies you thought were interesting to try, but which were not successful. Keep in mind that this is an individual assignment and you should not discuss the strategy you pursue with anyone else. You can find many internet sources which implement solutions to the above questions, but you are required to show a deep understanding of your approach. Sources should be cited. Failure to do so will severely affect your grade.

4 Documentation

The documentation should be formatted as paper using the `IEEEtran` LaTeX class template[5]. Please make sure to include page numbers. The maximum page limit for this report is 6 pages, excluding figures, tables and references. A `twocolumn` layout should be used, and font size should be set to `11pt`, with default line spacing. You are not allowed to adjust margins.

A brief general introduction and concluding remarks (stating limitations of your approach and what you have learned in this assignment) should be included.

The document should contain following sub-sections:

1. Introduction – **What is the conclusion of your task study and what evidence (if any) do you have to support your conclusion.** The rest of the task section outlined below will discuss how you arrived to this conclusion.
2. Data Collection – Describe the data collection process in detail, e.g. the data source(s) used, sampling frequency, *etc.*.
3. Data Cleansing – Describe the data cleansing techniques used, e.g. duplicate removal, normalisation of data, which data you decided to use (filtering) and why, *etc.*
4. Data Storage – Describe the data storage and model, e.g. how the data was saved prior to analysis, *etc.*
5. Data Analysis – Evidence for the selected strategy (this includes statistical analysis, visualisations, descriptive statistics, modelling including assumptions taken, hypothesis testing, *etc.*)
6. Conclusion – Critical review of your solution. Also, a discussion of shortcomings and possible improvements is required. The *Future Work*

[5]More information is available from: `http://ctan.mirror.garr.it/mirrors/CTAN/macros/latex/contrib/IEEEtran/IEEEtran_HOWTO.pdf`

section should not be banal (e.g. improve performance), overly generic (e.g. apply ML or gather more data) and/or marginal (e.g. change parameter α to from 1.10 to 1.15).

Note that not all tasks lend themselves well to the above sectioning. Use your good judgement to supply a comprehensive document.

The documentation should **not** include any code listings. Also, it should **not** re-iterate the problem in the introduction (we know what problem we set for you, we *wrote* this specification document). This also applies for the presentation of your assignment.

5 Grading Criteria

The following criteria will be taken into consideration when grading your assignment:

- **Ability to answer the specified research task**.
- Expanding on the idea of this assignment to show other trends or interesting findings. Also, including other data sources which complement your analysis.
- Thoroughness, completeness, and correctness of solution.
- Ability to critically evaluate your work (shortcomings, assumptions, *etc.*). Note these should **not** be superficial or marginal improvements.
- Creativity and originality in solutions.
- Readability of code and adherence to coding standards (naming conventions, comments, consistency of style *etc.*).
- Quality of documentation (presentation, proper use of language, writing style, references, figures, tables, captions, *etc.*)

THE END