MVA 3011 Assignment

Submission date 3pm Thursday 14th December in the School of Computer Science and Statistics Reception Office in the O'Reilly Institute.

Late submissions are not accepted without supporting approval from your tutor.

All submissions MUST be stapled/self-contained (no loose sheets), and must contain all student names and ID numbers.

This assignment is worth 20% of your mark for ST3011.

This project is for students to work either on their own or in groups of 2 or 3. If working in groups you can delegate out the work required and discuss within your group ideas and conclusions. However, any difficulty in teamwork within the group will be considered as a reflection of all individuals within the group, and as such, only a single identical mark will be given to all group members whose names are listed on the final submission.

The file weather_diag.csv is available at:
https://www.scss.tcd.ie/Brett.Houlding/ST3011.html

This data file records 14 different weather related variables measured daily at Dublin Airport over the duration January 1st 2003 until May 31st 2014. A google search of the variable names should provide a link to the Met Eireann description or otherwise.

Also recorded is a daily indicator variable listing whether or not a patient in Ireland was diagnosed with Anti–glomerular basement membrane (Anti-GBM) disease, a rare autoimmune disease of which little is known concerning its cause. One hypothesis is that the cause of this disease may in someway be related to environmental or pollutant factors, as has been observed in the case of a related Kawasaki disease affecting individuals in Japan.

Whilst it is unlikely that any student will find a connection between the recorded environmental data and the diagnosis indicator, this data-set was compiled to begin such a possible investigation. However, it should be noted that, due to the delay between symptom onset and the patient presenting at a specialist medical practioner, if the environmental factors listed had any role then they would be because of effects they had in an unspecified period prior to the actual date of diagnosis.

In any case, this is clearly a multi-variate data set and one in which you can demonstrate your understanding of the techniques and skills we have been developing as part of ST3011. As such your task is to analyse this data in an appropriate manner and to report back the insights that you have obtained.

You should present a written report of your analysis using a suitable computer word processing package that is of length no greater than five sides of A4 at minimum 1.5 line spacing and font size no less than 10 pt.

Relevant graphs and R summary output may be added in addition to the written report in the form of suitable appendices (provided that every such output or graph is commented upon within the report itself). You can also attach any code that you felt is important in your investigation, but please be aware that interest is in the 5 page report and any such code will only be glossed over if checked at all.

In grading your report I will be looking for you to demonstrate as much of what you have learned as possible in an appropriate way for this data. In particular the following will be considered:

A: Introduction - 5%
Is there a problem statement and setting? outline of method? brief results? introduce document layout?

B: Middle - 70%
15% - Descriptive Statistics/Plots: What is the data telling us and how has this informed your question?

30% - Method. What method are you using to answer your questions? Why this method (based on notes and bonus for additional references)? Do you understand what is happening 'under the hood'? Implementation (BRIEF description of computational method.)

25% - Results. Detailed results based on method? What was the output? Do you understand it, and is it what you expected? How have you explained your results?

C: Conclusion - 10%
What are your conclusions based on your research question and your results? Is this what you expected? If not, why not? (Based on method chosen to achieve results not on nature of data)
What further analysis would you carry out and why? If your results were unexpected, what would you do differently? Any other comments?

D: References - 5%
Class notes/Tutorial sheets, data set, R, Rstudio and R packages used, any other additional resources.

E: Overall Structure - 5%
Did you follow instructions? Is the layout as above (A, B, C, D) or similar? What referencing system did you use? Does the introduction and conclusion tie the document together?

F: Bonus - 5%
Did you read outside of class? Did you look beyond just one research question/method?
General feel - is this just another assignment or is additional effort/interest detectable?