MATH 70096 - Exploratory Data Analysis & Visualisation

Spring 2022 - Assessed Coursework 5

Deadline: 12 April 2022, 23:59 (UK time)

You should submit a PDF report and the source file for an accompanying R Markdown notebook, via the Imperial College Blackboard VLE, by the deadline stated above. You should also submit an MP4 file containing your presentation; this should be uploaded to the Panopto folder linked on the Blackboard submission page. Further submission guidelines are included below; read these carefully.

Each submission will receive a mark out of 30: this will comprise a mark out of 25 for the report, and a mark out of 5 for the presentation. This coursework should involve no more than **12 hours** of effort.

This coursework counts for 60% **of your total mark for EDAV.**

Plagiarism: Your submission should be *your own work*. Note that software tools are used for plagiarism detection.

The data in the accompanying CSV file, `DevelopmentIndicators2017.csv`, contains the 2017 values of a selection of 'global development indicators' for 68 countries, published by the World Bank at https://data.worldbank.org/[1]. Each observation in this dataset corresponds to a single country, and most of the variables correspond to a development indicator; these are detailed in Table 1, below. Further information on each indicator is available at the above website.

Suppose you have applied for the position of Data Scientist at the World Bank, and that you have been offered an interview. Ahead of the interview, you are required to perform an exploratory analysis of this dataset, for discussion during the interview. You are advised that your analysis should focus, in particular, on:

- any issues of data quality, such as missing data or outliers;
- a brief exploration of the univariate and multivariate distribution of the data;
- any clustering behaviour indicated by the data, and whether dimension reduction can aid the discovery of any clusters.

You are required to produce a report for your interviewer (who is a Senior Data Scientist), in which your analyses and conclusions are detailed. Your analysis should also be communicated using well-designed data visualisation and any visualisation design decisions that are made to improve the communication of the data should be briefly explained.

In addition, you should record a two-minute presentation that summarises your findings. This presentation should be suitable for a non-technical audience; it should focus on the high-level conclusions of your analysis and any interpretations that you can offer.

You are also required to provide the (.Rmd) source file for an accompanying R Markdown notebook, in which the analyses are implemented, and visualisations constructed.

---

[1] These data were downloaded via the World Bank's API, using the 'WDI' package in R.

Variable Name	Description of Indicator	World Bank ID
country	Name of country	-
region	Geographical region	-
income	Income category, as specified by the World Bank	-
GDP.percap	GDP per capita (current USD)	NY.GDP.PCAP.CD
Market.Cap.pcntGDP	Market capitalization of domestic listed companies (% of GDP)	CM.MKT.LCAP.GD.ZS
Unemployment.female	Unemployment, female (% of female labour force)	SL.UEM.TOTL.FE.ZS
Unemployment.male	Unemployment, male (% of male labour force)	SL.UEM.TOTL.MA.ZS
Education.Expend	Government expenditure on education, total (% of GDP)	SE.XPD.TOTL.GD.ZS
Arable.Land.pcnt	Arable Land (% of land area)	AG.LND.ARBL.ZS
Life.Expect.female	Life expectancy at birth, female (years)	SP.DYN.LE00.FE.IN
Life.Expect.male	Life expectancy at birth, male (years)	SP.DYN.LE00.MA.IN
Mortality.u5	Mortality rate, under-5 (per 1,000 live births)	SH.DYN.MORT
CO2.emiss.mtpercap	CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC
Access2Elec.pcnt	Access to electricity (% of population)	EG.ELC.ACCS.ZS

Table 1: Description of variables included in `DevelopmentIndicators2017.csv`. Further information on each of the global development indicators can be accessed at https://data.worldbank.org/, using the corresponding World Bank ID.

Further Instructions / Guidelines

This is an individual assessment; you should complete the analysis, report and presentation on your own, and without communicating any details to your MLDS colleagues.

You should submit one PDF report, detailing the analyses, the motivation for these analyses and the corresponding results, and including a selection of visualisations to aid the communication of these analyses. You should also provide the source file for an accompanying R Markdown notebook. ***The report will be marked out of 25.***

- The (PDF) report should be well-formatted (with correct mathematical typesetting, e.g. using LaTeX or MS Word) and concise. The total report (*including* visualisations) should be no more than five A4 pages in length.
- Visualisations should be chosen judiciously – i.e. don't just plot everything you can think of! Wherever possible, visualisations should also be constructed using the `ggplot2` library in R.
- The (PDF) report should not simply be the knitted version of the .Rmd notebook - the report should be a more concise communication of the analyses performed, the motivation for these analyses, and their results.
- The R markdown document should reproduce the analyses and visualisations provided in the report; this notebook may contain supplementary material (e.g. further visualisations), but will not necessarily contribute to the final mark.
- The .Rmd file should list its dependencies at the top of the document, and 'knit' successfully when each of these are installed in the appropriate location; you do not need to submit the knitted PDF/HTML.

In addition, you should record a short video presentation, and submit this as an MP4 file. ***Your presentation will be marked out of 5.***

- You should record a 2-minute presentation, giving a high-level summary of your analyses and their conclusions. Note that marks may be deducted for presentations that are significantly longer than 2 minutes.
- Your presentation should be accompanied by a small number of well-chosen slides; the presentation can be delivered either with or without the camera on – the choice is entirely yours.
- One option is to make use of the recording features in MS Powerpoint to narrate each slide, and save the resulting narrated presentation as an MP4 file. Further guidance on how to do this is included on the Blackboard submission page.
- Each MP4 file should contain one presentation only.
- Your MP4 file should be uploaded to the required Panopto folder by the deadline; instructions for accessing and uploading to Panopto are included on Blackboard.
- I recommend that you prepare, record and upload your presentations well in advance of the deadline, in case you encounter technical difficulties and (in the worst case) need to re-record.

Your PDF report, R markdown document and MP4 file should all use the following naming convention:

`{your surname}_{your CID}_Assessment5.xxx`,

where `.xxx` is the corresponding file extension, e.g. `Martin_00123456_Assessment5.pdf`.