Recitation 2

Data, Inference, and Applied Machine Learning

Friday 9 September 2022

Objectives of the assignment

- Extract and download data from different databanks, APIs and websites
- Compare different variables and study the relationships between them
- Describe and/or explain the relationship between variables
- Visualize datasets through different kinds of plots
- Synchronize timestamps in time series datasets
- Calculate summary statistics of datasets

Question 1: Malnutrition and GDP per capita

- Download the datasets for “GDP per capita (current US$)” and “Prevalence of underweight, weight for age (% of children under 5)” from World Bank DataBank (https://data.worldbank.org/indicator)
- “Malnutrition prevalence, weight for age (% of children under 5)” is also known as “Prevalence of underweight, weight for age (% of children under 5)”
- **Question:** What relationship do you think should exist between these two variables?
- Extract GDP and Malnutrition data values as matrices and produce the required scatter plot.
- **Question:** Now that you have the graph, what relationship exists between the two indicators?

Question 1 (continued)

- Ensure that the matrices of the metadata and those of the two main datasets are synchronized.
- Using the metadata, extract the countries for the six developing regions (excludes North America).
- Produce a scatter plot of the regions on the same figure, using a separate colour for each region.
- Label the graph properly with a meaningful title and an appropriate legend.
- **Question:** Any observations?
- Repeat the last four steps for the nine income levels and provide interpretations of your results.

Question 2: Time series plot commodity prices

- Visit https://github.com/quandl/Matlab or https://github.com/quandl/quandl-python to setup Quandl and get the unique API.
- Download the three datasets using the Quandl codes below and synchronize the time stamps.
  - ❖ Wheat Prices = ODA/PWHEAMT_USD
  - ❖ Crude Oil Prices = WGEC/WLD_CRUDE_WTI
  - ❖ Gold Prices = BUNDESBANK/BBK01_WT5511
- Produce a plot of the time series for the prices of Wheat, Crude Oil and Gold in $ on the same graph.
- Indicate the maximum and minimum prices of each of the three commodities on the graph using coloured dots and add a legend to explain each one.

Question 3: Summary statistics of datasets

- Download the required datasets (CO2 emissions (metric tons per capita) and School enrollment primary (% net) from World Bank DataBank.
- **Question**: How would you handle the NaN values?
- Compute the summary statistics required for each of the datasets and provide the results in tables.

Question 4: Fertility rate and GDP per capita

- Download the indicated datasets (Fertility rate, total (births per woman) and GDP per capita (current US$)) and produce the required scatter plot.
- Produce and **plot on the same axes**, cumulative distribution functions for the fertility rate variable using data of 1990 and 2010 respectively.
- Use vertical lines to indicate the mean and median of each of the distributions on the well-labelled graph.
- **Question:** What inference can you make with respect to change in fertility rate over the 20-year period?

Question 5: Happiness and Corruption

- Access the links provided and read about the various datasets and how they were obtained.
- Download the indicated datasets (Happy Planet Index for the year 2016, and Corruption Perceptions Index) from their respective links.
- Explore both files to understand how they are structured and pick out the sheets to use.
- Find the countries that are in both datasets and make a **fully annotated scatter plot** to demonstrate the relationship between the indices using the ranks of those countries in both cases.
- Use country abbreviations to annotated the points on the graph.
- **Question:** Can you identify any country or group of countries that appear unusual? Can you explain why you think they are unusual?

Submission instructions

Submission process:

1. Put the source code **file and data files** in a single folder
2. Name of the folder should be the same as your Andrew ID
3. **Zip this folder and attach the zipped file on the assignment submission page (CANVAS)**
4. After attaching the zipped file, click on “Add Another File” from the assignment submission page and **attach your report**
5. Submit your assignment

N.B. This new process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Submission instructions...

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID_DIAML_AssignmentNo. For example, mcsharry_DIAML_Assignment1, mcsharry_DIAML_Assignment2 and mcsharry_DIAML_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

Submission instructions...

The student is responsible for checking that their submission is complete. Students will lose 10% as for usual late submission even if the submission is repaired during the 24 hours after the deadline has passed and receive 0 for the assignment if it is not repaired.

The submission deadline is **Eastern Time (ET) on Monday 19, September 2022 17:59 / Rwandan Time (CAT) on Monday 19, September 2022 23:59**.

Q&A