QTM 350 - Data Science Computing

Final Project Instructions

1 Overview

This project tests your ability to combine the programming concepts covered in QTM 350 and produce your own analysis for real-world datasets. In groups of 4-5 students, you will conduct a comprehensive analysis of real-world datasets from the World Bank's World Development Indicators, focusing on one of the topics indicated below. You will use SQL for data cleaning and descriptive statistics, then switch to Python for data modelling and visualisation. The project will involve version control using GitHub and document the entire workflow in a Quarto report. The report should be fully reproducible and the repository should indicate the contributions of each group member with the number of commits and lines of code added.[1]

2 Dataset

You will use data from the World Bank's World Development Indicators (WDI) database. The dataset contains 1,600+ indicators for 200+ countries from 1960 to 2023, although not all indicators are available for all countries or years. The dataset is large, so you will need to select a subset of variables and countries to focus on for your analysis. I suggest that you work with only a few of them to keep the analysis manageable. You can download the dataset from the World Bank's website or use the excellent `wgapi` Python package to access the data directly from the World Bank's API. The package documentation is available on the package's GitHub repository. There are other packages available for accessing the World Bank's data, such as `wbdata`, which you can find here.

[1] For more information on how to use GitHub, see the GitHub Guides. GitHub provides these statistics in the Insights tab of your repository.

2.1 Topics

You can choose one of the following topics for your analysis. Each topic has three indicators that you can use to explore the topic further. You can use additional indicators if you find them relevant to your analysis. Feel free to work with particular countries (more than one) or regions (e.g., Sub-Saharan Africa, Latin America, etc.) to make your analysis more focused. You are also encouraged to assess the relationships between the indicators and how they have evolved over time, as well as to choose your own research questions.

Every indicator below has a unique code that you can use to access the data from the World Bank's API. The code is provided in the description of each indicator. Please visit the indicator's website and click on "Details" to read more about the indicator and its definition.

- **Economic Development**
- Datasets:
  - GDP per capita (constant 2015 US$) – Indicator: `NY.GDP.PCAP.KD`. Website.
  - Employment to population ratio, 15+, total (%) – Indicator: `SL.EMP.TOTL.SP.ZS`. Website.
  - GDP growth (annual %)– Indicator: `NY.GDP.MKTP.KD.ZG`. Website.
- **Population dynamics**
- Datasets:
  - Life expectancy at birth, total (years) – Indicator: `SP.DYN.LE00.IN`. Website.
  - Mortality rate, under-5 (per 1,000 live births) – Indicator: `SH.DYN.MORT`. Website.
  - Adolescent fertility rate (births per 1,000 women ages 15-19) – Indicator: `SP.ADO.TFRT`. Website.
- **Education**
- Datasets:
  - School enrollment, primary (% gross) – Indicator: `SE.PRM.ENRR`. Website.
  - School enrollment, secondary (% gross) – Indicator: `SE.SEC.ENRR`. Website.
  - School enrollment, tertiary (% gross) – Indicator: `SE.TER.ENRR`. Website.

3 Deliverables

The project will be due on December 8 at 11:59 PM. Please send the link to your repository on Canvas. You will submit the following items:

1. GitHub repository with the project files. The repository should include the following:
    - A README file with a brief description of the project and instructions on how to run the code.
    - A `data` folder with the dataset.
    - A `documentation` folder with the codebook and entity-relationship diagram.
    - A `figures` folder with the plots and tables generated in the analysis.
    - A `scripts` folder with the SQL and Python scripts used in the analysis.

    All group members should contribute to the repository. It can be helpful to assign roles to each member to ensure that everyone contributes equally to the project. Although the repository will be hosted on someone's GitHub account, all members should have access to it.

2. A Quarto report in PDF and HTML format. When you submit an HTML file, please make sure it is rendered on GitHub and all figures are displayed correctly. The report should include the following sections:
    - Title, names of project members and their Emory IDs.
    - Introduction.
    - Data description.
    - Data analysis.
    - Results and discussion.
    - Conclusion.

    The report should be concise, but not too brief. You may aim for 5 pages, excluding the code. You should focus on the main results and insights from the analysis. The report should be fully reproducible, meaning that anyone should be able to run the code and generate the same results as in the report.

If you have any questions, please let me know.

4 Grading Rubric

Please find below the grading rubric for the final project. The rubric is based on a total of 20 points, with each section having a different weight. The rubric is detailed to help you understand what is expected in each section of the project.

Component	Detailed Points	Total Points
Overall		2
Organisation and aesthetics	1	
Originality of approach and insights	1	
Introduction		2
Description of topic and research question	1	
Summary of main findings	1	
Data Description		5
Clear introduction of the dataset and context	1	
Data merging steps	1	
Data cleaning and pre-processing	2	
Summary of key variables (columns)	1	
Data Analysis		5
Interpretation of main results and trends	2	
Well-formatted tables and plots	3	
GitHub Management		3
Consistent version control with meaningful commit messages	1	
Effective use of branches and merges	1	
Clear README file outlining project structure and instructions	1	
Discussion		2
Clarity, conciseness, and relevance of discussion	1	
Reflection on findings and any limitations	1	
Technical Requirements		1
Report reproducibility using Quarto	1	
Total		**20**