SI 348
Individual Assignment
2025

Assignment Overview

You are required to complete a data analysis project as part of this module. The goal of this project is to apply the skills that you've learned in data transformation, cleaning, visualisation, and communication to a real-world dataset. While this assignment will focus on exploratory data analysis and clear reporting of your findings, all projects will involve substantial data processing and cleaning. Your attention is drawn to chapter 10 of R4DS (`https://r4ds.hadley.nz/eda`) and chapter 7 of Computational Analysis of Communication (`https://cssbook.net/content/chapter07.html`) for relevant material on EDA.

Project Selection

You must choose **2 out of the 6 project options**. Each project involves working with a dataset, conducting exploratory data analysis, and presenting your findings in a structured report.

Project Options

Project 1: Global Development Indicators

Conduct an exploratory analysis of global development indicators, examining relationships between, for example, variables like GDP, literacy rates, life expectancy, and internet access, among others. This project will use a dataset retrieved from the World Bank Development Indicators service.

Project 2: Customer Segmentation in Retail Data

Perform an exploratory analysis of customer purchasing behaviour, focusing on identifying key customer segments based on spending patterns using an ecommerce dataset for all sales between December 2011 and January 2020 at an online retail platform.

Project 3: Olympic Games Performance Analysis

Investigate trends in Olympic performance by country, sport, or gender. Analyse changes over time, visualise the data, and discuss factors influencing performance. The dataset includes data all athletes, medallists, hosts, and results from 1906 to 2022.

Project 4: Global Population Growth and Urbanisation
Analyse population growth trends across different regions, with a focus on urbanisation. Create visualisations to show changes in urban versus rural populations over time and discuss the implications for infrastructure and resource management. The dataset comes from the US World Population prospectus.

Project 5: Gender disparities in workforce participation and remuneration
Investigate gender disparities in workforce participation across different regions. Visualise trends in gender equality indicators and discuss the factors contributing to or hindering progress toward gender equality in the workplace based on the dataset provided by the World Bank Gender Indicators service.

Project 6: Student's choice
Choose your own dataset. Ensure that the dataset is of sufficient complexity to allow for meaningful exploratory data analysis. If you choose this option, email me with your proposed dataset and a brief description of your intended analysis before proceeding.

Project Requirements

For each selected project, you are required to produce a report on your exploratory data analysis. Specifically, you should complete the following steps for each of the two selected projects:

1. **Introduction**
   - Provide a brief introduction to the project. Outline the research question and objectives that you aim to explore in your analysis. Here, a degree of creativity is required.
2. **Data Cleaning and Preparation**
   - Load the provided dataset into R and perform an initial inspection.
   - Address any missing values, filter or categorise data as needed, and perform necessary tidying, transformations and variable calculations.
   - Document all steps taken during the data preparation process.
3. **Exploratory Data Analysis**
   - Provide basic summary statistics to describe the key variables in the dataset.
   - Create a variety of visualisations (e.g., scatter plots, bar charts, histograms, box plots, etc.) to explore patterns, trends, and relationships within the data.

- Conduct an in-depth exploratory analysis to uncover key insights related to the dataset.

4. **Reporting**[1]

    - Present your findings using clear and informative visualisations. Each visualisation should be accompanied by a concise explanation.
    - Write a narrative that ties your analysis together, focusing on key insights and what they mean in the context of the project.
    - Summarise the key takeaways from your analysis. Discuss any potential implications, limitations, or interesting avenues for further research.

Submission Requirements

For each of the two selected projects, you are required to submit the following files:

1. **Quarto File (.qmd)**

    - This file should contain all of your R code, data preparation steps, exploratory data analysis, and visualisations. Organise the file with appropriate headings and sections. Ensure that your code is well-commented to explain each step of your analysis.
    - Use the following format for the file name: `StudentNumber_ProjectTitle.qmd` (e.g., `12345678_CrimeTrends.qmd`).

2. **PDF Report**

    - This file should contain only the following elements:
        - Clearly labelled sections corresponding to the project requirements (Introduction, Data Cleaning and Preparation, EDA, Reporting).
        - Include key results, tables, and figures generated during your analysis.
        - Provide a clear narrative that explains your findings, supported by visualisations. Describe the steps you took in the analysis and discuss the implications of your findings.
    - Your code should not be included in the PDF output/version of your report.
    - The report should be well-organised, visually appealing, and formatted for clarity. Ensure that all visuals are properly labelled and referenced in the text.

[1] Note, while some degree of insight is required, it is not assumed that you are necessarily an expert in the domain of the data. Therefore, you will not be marked on the accuracy of your interpretations/insights, but rather on whether they match the data and are plausible.

- Use the following format for the file name: `StudentNumber_ProjectTitle.pdf` (e.g., `12345678_CrimeTrends_Report.pdf`).
- This PDF file should be generated using Quarto (see chapters 28 and 29 of R4DS).

How to Submit

- Ensure that you submit both the Quarto file (.qmd) and the corresponding PDF report (.pdf) for each of the two selected projects.
- Create a single zip file that contains:
  - Project 1: .qmd and .pdf files
  - Project 2: .qmd and .pdf files
- Upload your files to the link on SUNLearn by 23h59 on 10 October 2025.

Evaluation

Use the rubric on the next page to guide your work and understand how your project will be assessed.

Project Rubric

Category	**Criteria**	**Excellent (80–100%)**	**Good (65–79%)**	**Satisfactory (50–64%)**	**Needs Improvement (0–49%)**
Data Preparation (30%)	Data Cleaning & Transformation	Meticulously cleaned, transformed, and documented. Steps logical, clear, and appropriate.	Adequately cleaned/transformed. Most steps logical, minor issues.	Some steps incomplete or inappropriate. Weak documentation.	Poorly executed or missing.
	Handling Missing Data	Appropriately handled with clear justification.	Handled well, though some methods not fully justified.	Issues present or justification lacking.	Poorly handled or not addressed.
EDA (35%)	Summary Statistics	Comprehensive, accurately interpreted, key insights provided.	Adequate and accurate, most key insights included.	Incomplete or unclear interpretations.	Mostly missing or incorrect.
	Visualizations	Expertly crafted, relevant, well-labelled, clear and effective.	Clear, relevant, well-labelled, effective but less polished.	Basic or limited visuals; labelling/aesthetics could improve.	Unclear, irrelevant, or poorly executed.
Reporting (15%)	Narrative & Interpretation	Cohesive, insightful, well explained.	Clear, ties analysis together, mostly accurate.	Lacks cohesion or depth; weak/unclear.	Disjointed or missing narrative.
	Results Presentation	Logical, organised, highlights key findings, excellent storytelling.	Well-presented and organised; storytelling could improve.	Somewhat disorganised, unclear findings.	Poorly presented or disconnected.
	Conclusion & Implications	Insightful, discusses implications and limitations in depth.	Sound conclusions, some implications discussed.	Vague or weak, limited discussion.	Missing or unsupported.
Technical Execution (15%)	R Code Quality	Efficient, well-structured, reproducible, with clear comments.	Mostly efficient and reproducible, adequate comments.	Some inefficiencies, unclear comments.	Inefficient, poorly structured, lacking comments.
	Quarto Integration	Meticulously organised, professional PDF, error-free.	Well-organised, minor integration or formatting issues.	Somewhat unclear or disjointed.	Poorly organised, unprofessional report.
Critical Thinking (5%)	Insight & Originality	Deep insight, original thought, novel findings.	Good insight, some originality, meaningful findings.	Limited or superficial insight.	Minimal or trivial findings.