STATS/CSE 780 - Homework Assignment 1

Pratheepa Jeganathan

2023-09-12

Instruction

- **Due before 10:00 PM on Tuesday, September 26, 2023.**
- **Submit a copy of the PDF with your report (2-3 pages) and technical supplemental material (less than 10 pages) to Avenue to Learn using the link that was emailed to you.**
  - Technical supplemental material can only include R or Python codes for the results reported.
- **Late penalty for assignments: 15% will be deducted from assignments each day after the due date (rounding up).**
- **Assignments won't be accepted after 48 hours after the due date.**

Assignment Standards

Your assignment must conform to the Assignment Standards listed below.

- Quarto or RMarkdown or LaTeXis strongly recommended to write the report.
- Quarto or, RMarkdown, or Jupyter Notebook must be used to write the supplemental material.
- Report is about the results by applying data science methods and how you interpret or discuss the results. Don't show in the report how you do the analysis using R/Python.

- Technical supplemental material is how you produce the report results using R/Python. Don't print chunk messages, warnings, or extended data frames in the PDF.
- Write your name and student number on the title page. We will not grade assignments without the title page.
- Eleven-point font (times or similar) must be used with 1.5 line spacing and margins of at least 1~inch all around (This document used these formats).
- Your report may not exceed **three pages**, inclusive of tables and figures. It would help if you chose the tables and figures accordingly for the report. You can also keep other tables and figures in the supplementary material (**less than 10 pages**) and refer to the report. In addition, you may use one page for the bibliography and one page for the title page.
- You may discuss homework problems with other students, but you have to prepare the written assignments yourself.
- No screenshots are accepted for any reason.
- The writing and referencing should be appropriate to the graduate level.
- Various tools, including publicly available internet tools, may be used by the instructor to check the originality of submitted work.
- Assignment policy on the use of generative AI:
  - Students are not permitted to use generative AI in this assignment. In alignment with McMaster academic integrity policy, it "shall be an offence knowingly to … submit academic work for assessment that was purchased or acquired from another source". This includes work created by generative AI tools. Also state in the policy is the following, "Contract Cheating is the act of"outsourcing of student work to third parties" (Lancaster & Clarke, 2016, p. 639) with or without payment." Using Generative AI tools is a form of contract cheating. Charges of academic dishonesty will be brought forward to the Office of Academic Integrity.

Question

Students can use either R (ggplot2 and R packages for data transformation), covered in class or Python (matplotlib and Python modules for data transformation).

Sustainable development is important for various reasons, as it addresses the critical interplay between economic growth, social progress, and environmental stewardship. In this assignment, students will use Open Government Portal data to address one of the Sustainable Development Goals by United Nations.

Open Government provides open data and info, enhancing accountability and transparency.

We can navigate to "Search for open data and information," called Open Government Portal. Once you land on this page, on the left, there are filter options (by organization, resources type.)

Now, students will use the filter option to choose a dataset with the following requirements.

1. Update Frequency - Annually or Monthly or Quarterly (the dataset must contain at least 10 time points)
2. Resources Type - `Dataset`.
3. Format - `CSV`.
4. Organization - For example, I choose Environment and Climate Change Canada.
5. Keywords - For example, I choose `wastewater` to address SDG 3 (Ensure healthy lives and promote well-being for all at all ages)
6. Dataset contains at least one quantitative and categorical variable across provinces and territories.

If there are too many variables and samples, you can choose a subset of data after downloading to make the following plots.

(i) Briefly describe your chosen dataset and clearly explain where it was sourced.

(ii) Clearly explain data transformation and the preprocessing methods you used to tidy the data.

(iii) Choose one (quantitative) variable for the following analysis. Then, use an appropriate visualization method to describe the trend of the variable in the selected time-frequency across provinces. Finally, clearly describe any statistical transformation used for visualization and interpret the results.

(iv) Aggregate (aggregate over the provinces) the selected variable (from iii) for Canada and inspect the trend over the selected time-frequency using an appropriate visualization method. Interpret the results.

(v) You can use either Shiny for R or Shiny for Python for the following analysis.

- In addition to provinces, choose a categorical variable with more than two categories—product type or health status, income status, etc.
- Use an appropriate plot to show the change of quantitative variable (from iii or any other quantitative variable) in the selected time-frequency across provinces when the user chooses the category.
- The supplementary material must include the R or Python codes you used to create the app.
- You must submit a link to your Shiny App. We (Instructor or TA) must have access to the app when we grade it; otherwise, no points for the app (only for the code if provided) are given. Describe your app in the report.

For all the questions, write a clear and concise interpretation of the plots and clearly state what conclusions can be drawn from the plots or graphs — these conclusions should be cast in the context of the chosen dataset and one of the sustainability development goals (SDG).

- Plots must be readable.
- Choose an appropriate font size for plots.
- Label all aesthetics and axes in the plot.
- Use appropriate statistical transformation for plots.

Grading scheme

(i)	Data description	Describe the chosen dataset (background of the dataset and the variables) [3]
(ii)	Data transformation	Did you choose all the downloaded variables and observations or a subset? Describe the reasons for using all the data or the subset. [2]
	Pre-processing	How did you identify missing values? How did you represent the missing values in tidy data? How did you identify outliers? If there were any outliers, how did you handle them? [4]
(iii)	Plot	Appropriate plot, the plot is readable, appropriate font size, label all aesthetics and axes, use appropriate statistical transformation, interpretation (how to read the plot), conclusion (any interesting patterns related to the chosen SDG) [4]
(iv)	Plot	Appropriate plot, the plot is readable, appropriate font size, label all aesthetics and axes, use appropriate statistical transformation, interpretation (how to read the plot), conclusion (any interesting patterns related to the chosen SDG) [4]
(v)	Shiny app	Link to shiny app works, the output is an appropriate plot, shiny app reacts to the user inputs (categories), description of the app is written in the report or the app [4]
	Plot	Plot is readable, appropriate font size, label all aesthetics and axes, use appropriate statistical transformation, interpretation (how to read one of the plots), conclusion (any interesting pattern in one of the plots related to the chosen SDG) [4]
References		Reference list starts on a new page, references are appropriate and list out in the **report** [2]
Supplementary material		Supplementary material starts on a new page, code readability, all codes are within the margins, the R or Python codes and the outputs for the questions are presented [3]
	Shiny app	Shiny app codes (don't execute the codes when you create the PDF) [2]

The maximum points for this assignment is 32. We will convert this to 100%.