DSA2101 Project Guideline

AY22/23 Semester 2

Aims and scope

For the group project, you will be using data from the TidyTuesday repository. You will apply the techniques acquired in this course to explore and visualize real-world data. You can **choose one** of the following topics:

1. Wealth and income (2021-02-09)
2. UN Votes (2021-03-23)
3. Deforestation (2021-04-06)
4. Billboard Top 100 (2021-09-14)

Please only use data sets from your chosen TidyTuesday project.

Your task is to come up with **two** distinct and meaningful questions, answer each of them with two visualizations using `ggplot2`, and write-up your method and findings. You are expected to give your best effort using tools acquired in class.

Workflow and timeline

1. **Weeks 8-9**: Read the project guideline carefully. Form a team of **up to 5 persons**. Discuss with your team members and determine the topic that interests you most.
2. **Week 9**: Create your own group on Canvas. You may want to refer to the Canvas wiki here.
    - A group can either by "open" (to allow any one to join) or "by invitation" (to include only a specific set of classmates).

- The group name should be the name of the chosen TidyTuesday project.
- Please complete this no later than **March 17 11:59pm (Friday of Week 9)**. This is to help you with workflow – changing topic/group member after this point of time would likely be counterproductive.

3. Think about **two distinct questions** that interest you the most and approach to answer them using two sets of variables that do not completely overlap each other. Answer each question with two data visualizations of different types and write-up your method and findings.

4. **Friday of Week 13**: Submission of project work, including an `Rmd` file and the knitted `HTML` file, by **April 14 11:59pm (Friday of Week 13)**.

Write-up

The write-up should contain **two** questions about the data; each question is answered with **two** visualizations of different types. The structure the write-up could be as follows.

1. **Introduction** (1 paragraph): Briefly introduce the data set. You can repeat some of the data description found on the TidyTudsday repository, paraphrasing on your own terms. Imagine that your reader has no prior knowledge about the data set.

2. **Descriptive statistics** (1-2 paragraphs and a few code blocks): Mention any transformation or cleaning you have applied to the data. Present key descriptive statistics to give a taste of what the data are like. Resist the temptation to report descriptive statistics in bulk and be more selective in reporting only the most interesting/relevant statistics.

3. **Question 1**: The title should relate to the question you are answering.

    - **Introduction** (1 paragraph): Introduction to the question and which parts of the data set are used to answer the question. Discuss why you are interested in this question.
    - **Methodology** (1-2 paragraphs): Describe the visualizations you make to answer your question. For each plot, explain what variables you are plotting and provide a clear explanation as to why it is the best for providing the information you are asking about. Two plots for the same question should be of different types.

- **Visualizations** (2-3 code blocks and 2 figures): In this section, provide the code that generates the plots. All plots must be made with `ggplot2`. Do not use base `R` or other plotting packages.
- **Discussions** (1-2 paragraphs): Describe the primary information you hope the reader will gain from your visualization. Identify any trends revealed (or not revealed) by the plots. Explore and explain why the data looks the way it does.

4. **Question 2**: Same structure outlined for Question 1, but for your second question. The visuals made in Question 2 can be of the same type as those in Question 1, but they should be of different types from each other.

5. **Reference**: List any references in this section. At a minimum, you should list your data source.

Be concise in your write-up. A paragraph should typically not be longer than five sentences. Remember, in a real-life scenario, your audience has very limited attention span.

You are not required to perform any statistical tests or modeling in this project, but you may do so if you find it helpful to answer your question.

Teamwork

You are to complete the assignment as a team. All team members should do an equal amount of work. Anyone judged to not have sufficiently contributed to the final product will have their grade penalized. While different team members may have different backgrounds and abilities, it is the responsibility of every team member to understand how and why all code and approaches in the project works.

Grading

The total marks allocation is **15 marks**. It will be graded based on

- The introduction provides a clear explanation of the questions of interest and the data to answer the questions, including the description of all relevant variables.
- The chosen visualizations are clearly explained and justified. They are appropriate, easy to read, and properly labelled.

  - Note that you do not need to visualize all the data. Instead, you want to focus on the quality of your visualizations – a single high quality plot will receive a much higher grade than a large number of poor quality charts.

- Discussion of results is clear, correct, and concise. It has some depth without being excessively long.
- Code is correct, easy to read, properly formatted, and properly documented.
- All required files are submitted. The `R` Markdown file can knit without any error and reproduce all outputs.

Submission requirements

The group project is due on **11:59pm on April 14, Friday of Week 13**. Your submission should include two files:

- An `R` Markdown file.
  - Note that the submitted code should include only the operations that generate the reported results in the sequence they are reported. You may have generated results that are not as interesting or relavent, please keep them out of the submission files.
- A knitted `HTML` file
  - Verify that your `R` Markdown file produces exactly the same results as reported in the `HTML` file.
- Only one submission is needed for each group.