1. Project Preamble

This is an individual effort project. In this project, you are to demonstrate your understanding of how exploratory data analysis is carried out using R functions and visualisation tools.

2. Data

To keep the unit relevant to real-world data analytics while fulfilling the educational goals, we would like to provide the following options for sourcing data for the project.

- A specified dataset
- Data from public repositories

Specified data

If you do not have any datasets or domain of interests in mind, we suggest you to use the **Global YouTube Statistics 2023** dataset, which can be obtained from *kaggle.com*, collected by *Nidula Elgiriyewithana*.

You can create an account and log in to Kaggle to download the dataset (a .csv file) and access the data dictionary. A local copy is also provided in LMS under the *project* tab.

Data from public repositories

If you have specific domain of interests, for example, energy consumption, health sciences, transportation, sales, or sports, etc., you can browse some public dataset repositories to find a dataset that interests you. Note that the data needs to be in tabular form, i.e. not multi-media data, as we won't be able to deal with texts, images, videos, or audios. Also, the dataset should not be too simple. It should have continuous and discrete variables (columns) of various types (numerical, logical, character, etc). Your chosen dataset should have a similar level of complexity as the YouTube dataset.

A few well known public data repos are:

- **http://kaggle.com**
- **https://data.gov/**
- **https://data.gov.au/**
- **https://data.world/**
- **https://archive.ics.uci.edu/ml/datasets.html**

3. Exploratory Study of the Data

For this project, we present the data in two different ways to our stakeholders.

1. You are expected to produce an **HTML** file from your **R notebook** (a .Rmd file) that documents both the process and the R code that you used for your exploratory data analysis.
2. A **Shiny App** is also expected to coordinate the various plots with explanations to support interactive exploration of the data.

Using R functions to explore the data

Use R functions such as `str()`, `summary()` and `head()` to have a glance at the data. Document your interpretation of the data in the notebook.

Visualisation

Generate different types of plots and charts from the dataset for both single and multiple variable data exploration. Document any intuitions and observations you have in the notebook.

Data cleaning and transformation

Is there any missing values and data anomalies? Do you think it is important to do any data transformation? If so, document these in the notebook. Use visualisation to help justify the cleaning and transformation.

4. Marking Criteria

- **R functions (20%):** Correct use of **built-in R functions**, sensible interpretation of results, demonstrable proficiency in **writing one's own functions**.
- **Data cleaning and transformation (30%):** Meaningful cleaning and transformation are performed on the data. Operations performed are well described and justified. Good use of coding convention and data wrangling pipes.
- **Visualisation and Report Quality (20%):** At least 5 different types of plots are used correctly, with good justifications of the choice of *geoms* and annotations for meaningful readings of the data.
- **Visualisation and Shiny App Quality (20%):** Meaningful interactions are designed and implemented in the Shiny App to allow for smooth navigation of the various types of plots.
- **Process (10%):** Demonstrating a good understanding that EDA is an iterative process.

Note: the quality of a data science project is not about just meeting these requirements, which are often considered as the bare minimum. One often has to go out of the way, demonstrating professionalism, proficiency, effort and thoroughness to obtain marks in the HD range (80%-100%).
Please see the model project from 2018 for a submission in the high HD range.

5. Submission

1. Generate an html file with the name **project1.html** from your notebook and submit it to **cssubmit**. Ensure that your submitted file can be opened and viewed in a web browser (e.g., firefox, chrome).
2. Produce a short screencast to show how you can use your Shiny App, and upload to YouTube as an unlisted video. Submit the .R file (with the video link in the first line as a comment) to **cssubmit**

You should keep a record of your progressive work towards the final submission and also a copy of your latest work. You are encouraged to submit as often as you like to *cssubmit*. The latest submission will overwrite the previous version.

If you like version controlling your work, then you can keep your working copies on GitHub before the final submission to cssubmit; however, do make sure you keep your GitHub repo **private**. You can also try out Data Version Control at DVC. Note these are good practices that are encouraged but are not marked.

Submission Check List:

- Your name and student number are visible at the beginning of your notebook. To make it easier for us in the marking process,
    - please ensure that you **have your student number written correctly**.
    - please ensure that you **use exactly the same name as shown on LMS**.
    - please put your **surname in uppercase letters**, e.g., Michael CHEN, John SMITH, Xiaolian HUANG.

- Submit the generated **.html** file and the **.R** file used for Shiny App. Alternatively, you can insert the Shiny App code into the **.Rmd** file and submit the original .Rmd file as well as the **.html** file.
- Make sure the URL to your video is included in the **.html** file.
- If you are using data sourced from the web, make sure that you use the original URL in the data import function for your EDA.

How to get started with the project:

1. Create an R Markdown file by following the instructions.
2. Modify the code blocks with your own R code to read data into a data frame, extract useful columns to visualise using various plots.
3. Create a Shiny Web App, and open the App.R file, copy the code across into a code block in the R Markdown file. You will see you can use the Play button (the tiangle icon next to the code block) to launch the app. Then you can modify the App code inside the same R Markdown file.
4. Add descriptive text and explanations to document you EDA in a report style, as if you are written in a MS Word Document. Apply the markdown styling symbols when necessary, e.g. use a combination of typography such as headings, boldface, italics, tables, images, hyperlinks etc.
5. Make sure the sections are organised according to the marking criteria.
6. Please remember to **use R**for conducting ETL in Project 1. The use of Excel, Python, other programming languages, or GUI tools is not acceptable for this purpose.

***Requirements for the submited Video:

- You can prepare a 2-4 minutes video (marks maybe deducted if too short or too long) to explain and highlight anything outstanding with respect to:

    - Shiny APP design and implementation.
    - Navigation of the various types of plots
    - Meaningful interactions.
- You can take the video by various applications.