Stage 1: Data Description & Exploratory Data Analysis (individual assignment #1)

Assignment Submission Instructions:

- Please use **point-form full-sentences** in your report to facilitate grading.
- Clearly **indicate your project group number** at the top of your report.
- Use the **correct submission page based on your group number**, otherwise there will be a 10% penalty.
- This is an individual assignment! Every student needs to write and submit their own assignment. However, you are part of a group and at the end of the term you will submit a group report. We encourage you to talk and discuss your ideas with other group members. You don't need to agree or have the same answers. In fact, it will be insightful for the group if you explore the data from different angles.
- Each of you must submit two files in Canvas:
  - The source Jupyter Notebook (.ipynb file)
  - The rendered final document (.html file)

What Sections to Include?

Section 1: **Data Description**

1. Provide a descriptive summary (criteria given below).
   - Provide a brief description of the dataset assigned to your group. Note that the dataset will probably contain more variables than you need. In fact, exploring how the different variables in the dataset affect your model may be a crucial part of the project.
   - *Regardless of which variables you plan to use*, in this section, **provide a full descriptive summary of the dataset**. Please use a table or bullet points to describe the variables in the dataset.
   - At a minimum, you should include information on:
     - the number of observations
     - the number of variables
     - the name and type of variables
2. Source and information (2-3 sentences)
     - If known, indicate how the data has been collected. In all cases, include the data source and citation as requested by the owner(s)Tip: some datasets do not provide full information

about certain variables. Report what is available here and use NA for incomplete information

3. Pre-selection of variables (2-3 sentences)
   - Some variables will be discarded if they contain redundant information or if they won't be needed in future analysis. Include all variables in this section, regardless! Identify those that will be dropped and explain briefly why. Call this part "Pre-selection of variables".

Section 2: **Scientific Question**

1. Clearly state the question you want try to answer using the dataset. (2 sentences max)
   - Your question should involve one random variable of interest (the response) **and more than one explanatory variables** in the dataset.
   - Tip: your question should consider as many input variables as possible and you don't need to name all. For example: for the `Bikeshare` dataset presented in lectures, the question can be: "We want to examine the association between the number of bikes rented in a bikeshare program (response) and predictors related to climate conditions (e.g., temperature, wind speed, rain) and usage scenarios (e.g., commuting to work or biking around on the weekend)."
2. Name the response. (1 sentence)
   - You need to clearly state what the response is.
   - There may be more than one variable that can be used as a response, obviously it depends on the question and it will affect the model used (MLR, Logistic or Poisson). However, only some variables in the data can be explained by others, not all directions in the association are reasonable. For example, you won't explain how much it rained by the number of bikes rented -- the other direction makes more sense.
3. Explain whether your question is focused on prediction, inference, or both. (2 sentences max)
   - It is fine to have the same question as other group members but each member must have their own explanation. Again, you don't need to agree on a unique common question for now. In fact, usually many questions can be answered with the same dataset.
   - Tip: you may be interested in both but since some methods differ depending on the aim, state here the *primary* focus of your project. This can change later, in consultation with your TA.

Section 3: **Exploratory Data Analysis and Visualization** (no more than one plot per student)

1. Provide your **reproducible code** for

   - demonstrating that the dataset can be loaded into R,
   - cleaning and wrangling your data into a tidy format, and
   - performing visualization to explore the data.
2. Provide a **visualization** that you consider relevant to address your question or to explore the data.
   - Propose a high-quality, creative plot (you are allowed to use facets to explore a plot according to values of another variable).
      - Tip: Be ambitious with your plot! Try to explore at least 3 variables at once!! ***See here for some examples (https://canvas.ubc.ca/courses/171918/files/41928265?wrap=1)*** ***(https://canvas.ubc.ca/courses/171918/files/41928265/download?download_frd=1)*** that would be given full marks.
      - Tip: You are allowed to show multiple plots as long as they are combined into a single one.
      - Not all variables need to be included in the visualization, justify briefly your choices.
      - **IMPORTANT**: you are not allowed to use pairs() for this assignment. Push yourself to create a more informative visualization!
3. Provide the following **Interpretations** (max 2-3 sentences for *each* point below)
   - Explain why you consider this plot relevant to address your question or to explore the data.
   - Interpret briefly the results obtained.
   - What do you learn from your visualization?
      - Tip: this visualization does not have to illustrate the results of a methodology. Instead, you are exploring which variables are relevant, potential problems that you anticipate encountering, groups in the observations, etc.