Background

In this project, you will perform exploratory data analysis (EDA) on a subset of the Pediatric Clinically Important Traumatic Brain Injury (ciTBI) dataset. This dataset is from The Pediatric Emergency Care Applied Research Network (PECARN), a well-known multi-center research organization focused on the prevention and management of acute illnesses and injuries in children across the continuum of emergency medicine health care.

The dataset is available in the `data` folder of the repository as a CSV file named `citbi.csv`. Additionally, a data dictionary is provided in the same folder as a spreadsheet file named `citbi_data_dictionary.xlsx`. This data dictionary contains detailed descriptions of each variable in the dataset.

The dataset contains information on children who visited the emergency department with head trauma. Your task is three fold:

- Clean the dataset so that it can be analyzed effectively.
- Explore and explain interesting patterns in the dataset. This process is called exploratory data analysis (EDA), a critical step in any data science project.
- Report your findings to a hypothetical client, **Dr. Bayes**, a pediatric emergency physician and researcher who works with PECARN.

While you are completing each part of the project, you will have a few files to work in and submit via Github Classroom:

- **preparation.qmd**: The document containing your answers to each question.
- **report.qmd**: A document which will show selected data products to the client.
- **report.pdf** A rendered PDF of the `report.qmd` file.
  - Ensure your report has `format: typst` in the yaml header for it to render properly. Typst is a new markup based typesetting system similar to LaTeX. Although you won't be writing any typst code in your report, putting it in the yaml header will give Quarto the ability to render a nicely formatted PDF.

Part I: Data Cleaning

Hello, Stat 133 student! I'm Dr. Bayes, a pediatric emergency physician and researcher. I work with a large network of hospitals called PECARN studying children with traumatic brain injuries (TBI). I am currently dealing with a large and messy dataset and I need your help to make sense of it.

I have given you access to the dataset in the data folder of this repository.

The dataset contains information on children who visited the emergency department with head trauma. I want to understand the characteristics of these patients and their injuries. I also want to identify any patterns or trends that may help us better understand pediatric ciTBI.

To start, you will need to clone the github repository that contains the starter code and data. You can find this in the invite link in the project 3 assignment on **Ed**. To do so, you will need to run the following command in the terminal / git bash:

```
git clone <your-github-repo-link>
```

Your first step will be to clean the dataset in the `preparation.qmd` file. This includes (but is not limited to) the following tasks:

1. Address Missing Data:
   - How is missing data being represented? *Hint: Look at the data dictionary.*
   - Notice that some columns have a special 91 value that indicates a pre-verbal patients (children who haven't started speaking yet), some columns have a 92 that could be specific to that variable.

2. Change Variable Names:
   - Consider changing variable names. For example, names like "Clav" are unclear. A better name could be "clavicle_trauma".

3. Specify Variable Types:
   - Identify character, logical, and numeric vectors. Convert data types where appropriate.
   - Consider converting character vectors to factors.

Part II: Exploratory Data Analysis (EDA)

Great work cleaning the dataset! Next, I would like for you to perform exploratory data analysis (EDA). This is a critical step for most projects in data science. The goal of EDA is to better understand the data, identify patterns, and generate hypotheses. Additionally, EDA presents an opportunity to create data products that can be shared with others. (Still working withing the `preparation.qmd` file)

4. Create a Missing Data Summary:

- In Positron, navigate to the `session` tab in the top right where you can view a list of objects saved in your environment. Click the spreadsheet icon to the right of the data frame name you created in part I. You should be launched into a new tab with a split pane view. On one side, you will see mini-histograms of each variable along with the proportion of missing values. If you click onto the variables you can see some useful summary statistics. On the other pane, you can see the data frame itself.
- Now, **create a table** that summarizes the number and the proportion of missing values for each variable in the dataset. Does it match up with what you saw in the positron summary tab?

5. Create a histogram of the patient ages in months. Describe any interesting patterns you see.

6. Create a grouped summary table that shows the total count of patients for every combination of the loss of consciousness length and ciTBI outcome columns.

7. Create a bar chart to visualize the count for each category side-by-side in length of loss of consciousness by ciTBI outcome.

8. The variable (originally) named "GCSTotal" refers to the Glasgow Coma Scale (GCS) score, a neurological assessment used to evaluate the patient's level of consciousness. Create three visualizations for the relationship between total GCS score, Age in years, and ciTBI outcome.
   a. Create a stacked normalized bar chart to visualize how the proportion of patients with ciTBI varies across different ages.
   b. Explore whether the relationship between age and ciTBI differs across Glasgow Coma Scale scores. Create a stacked normalized bar chart that shows the proportion of patients with ciTBI across age, now faceted by total GCS score.
   c. Create a stacked bar chart that shows the number of patients with ciTBI across age faceted by total GCS score.

9. Create a scatter plot to visualize any two numeric variables in the dataset. Describe any interesting patterns you see.

10. Create a table that summarizes at least two statistics between two categorical variables. Describe any interesting patterns you see.

Part III: Explanatory Data Analysis

Now that you've cleaned and explored the data, your last task is to create a short report to communicate your findings to Dr. Bayes. You will need to highlight 3-5 insights from part II.

Your task

Select 3-5 of the most insightful data products from part II. For each data product, you will need to:

- Polish it so that it highlights a key message and is presentable.
- Provide a description below the data product that explains what it shows and why it is important.
- Consider creating new and related data products that can further highlight this insight. What do these new products contribute to the story?

Tips for an effective report

- Consider your audience. Dr. Bayes is a busy physician and medical researcher, not a statistician. He isn't expecting you to be an expert in the field, but he does want to understand the key findings.
- Be concise. Your report should be no more than 2 pages long.