Data Cleaning

---

1. **Aim:**
   - Work with Git
     - Set-up GitHub repository
   - Clean the dataset
   - Functional programming
   - Follow Pandas best practices

2. **Things to Do:**
   - Check data types of features
   - Check validity of values for each feature
   - Check accuracy of values for each feature (extreme values)
   - Check for duplicates

3. **Things Not to Do:**
   - Learn any patterns in the dataset
   - Plots

4. **Flow:**
   - Set-up GitHub repository
   - Import libraries
   - Read the dataset
   - Preliminary Analysis:
     - Validate data types
     - Check for duplicates
   - Detailed Analysis:
     - Check the validity and accuracy of each feature individually
   - Document observations and changes to make
   - Implement Data Cleaning function
   - Split Cleaned Data into 3 subsets (X, y)
     - Train
     - Validation
     - Test
   - Export subsets
     - Implement a function for exporting data

- Take *X* and *y* as parameters
- Join the objects and export as a *.csv* file
- Display data for completeness