Practice Assignment 10

Create a GitHub repository called “st2195_assignment_10” and include a package in R and Python able to replicate:

1. the basic structure of the EDA in the assignment for modules on data visualisation [4 points]
2. the analysis in practice assignment 9 [4 points]

for any given dataset, pre-set of features and a categorical target. Add also two examples (one in R and one in Python) based on the titanic dataset and a small documentation to describe your packages and their basic functionalities. [2 points]

Note that GitHub provides different ways for documenting your repo. More details can be found at https://guides.github.com/features/wikis/.

Additional Notes:

- Objective -- Create both R and Python packages based on the analysis we have done for Practice Assignments 7 to 9 on titanic dataset
    - Relevant practice assignments:
        - R – Use Practice Assignments 7 and 9 as a starting point
        - Python – Use Practice Assignments 8 (Part 1) and 9 as a starting point
    - Convert the code in those practice assignments into separate functions within a package, so that it can take any dataset and do the following:
        1. Exploratory data analysis (EDA)
        2. Fit models
        3. Plot and compare models

- Hints/Suggestions:
    - Exploratory data analysis (EDA)
        - R and Python
            - For each variable (including target): (i) plot bar chart if variable is factor/category; (ii) plot histogram if variable is numeric
            - Additionally, for each non-target variable: (i) plot stacked bar charts by target if variable is factor/category; (ii) plot violin charts by target if variable is numeric
    - Fit models
        - R
            - Allow user to specify any dataframe as input data, the target variable, and the machine learning algorithms/models
            - Many of the algorithms follow a similar process for fitting a model that includes pre-processing and training/fitting
        - Python
            - Most of the code for fitting a model are the same across algorithms, with differences mainly in the pipeline steps.
            - You can store the specific pipeline for each algorithm in a dictionary and access the appropriate one to fit the model as needed.
            - Similarly, you can store the fitted models as a dictionary.
    - Plot and compare models
        - R
            - Once the models have been fitted, you can plot and compare the differences (e.g., using boxplots)
        - Python
            - Since you have stored the fitted models in a dictionary, you can just plot for each of the models in there (e.g., ROC curve).

- Putting your R package on GitHub (let's name it "analysis")
  - Useful reference -- https://kbroman.org/pkg_primer/pages/github.html
  - GitHub:
    - Create a new empty repository with a suitable name.
      - Navigate to Settings -> Developer Settings -> Personal Access Tokens to generate an authorization token
      - Make sure to select the required scope for it to work
      - Note down the authorization token
  - Command Prompt:
    - Change to the local package directory -- in our case, the directory that contains the "analysis" folder
    - Initialize the repository with "git init"
    - Add and commit everything with "git add ." and "git commit"
    - Create a new repository on GitHub (as described in section above)
    - Connect your local repository to the one on GitHub
      - git remote add origin https://token@github.com/username/repository.git
      - git branch -M main
      - git push -u origin main

- Installing your R package from GitHub (named "analysis")
  - RStudio Console:
    - library(devtools) *#load library containing install_github function*
    - install_github("username/repository/analysis", auth_token="token") *#if repository set as private*
    - install_github("username/repository/analysis") *#if repository set as public*
    - library(analysis)