Exercise 1. Data cleaning and prep.

The first exercise consists in cleaning and merging different raw (or "semi" raw) data to create a unified dataset of players that can be used to run an analysis that relates a player's fantasy price to his performance metrics. The different datasets have different identifiers and different levels of aggregation. Remember to comment your code as you work on it.

You will find 4 datasets in the folder:

- **player_performance.rds**: includes player performance data, obtained via web-scraping
- **teams.csv**: includes team names and numeric IDs
- **players.csv**: includes player basic info (numeric ID, name, position, team, etc)
- **FPL Player Stats(2016-2024).csv:** includes various player statistics

For this task, you should:

1. Create a well-organised folder for the data and codes;
2. Read the different datasets in R;
3. From the dataset **FPL Player Stats(2016-2024).csv**: extract the player price (variable "value") for the first game week and the 2023/2024 season;
4. From the dataset **teams.csv:** extract team name and ID;
5. From the dataset **players.csv**: extract player web name, team and field position (goalkeeper-defender-midfielder-forward). Recode field position into a numeric variable;
6. Merge the different datasets into a unique dataset at player-level;
7. Label the variables;
8. Save the dataset in the appropriate folder.

Exercise 2. Scatterplots.

The dataset **player_performance.rds** contains various metrics of player performance. Consider fpl_points and fpl_points_fbref, which are two alternative measures of player performance, measured in fantasy football points. fpl_points is the "true" measure, while fpl_points_fbref is a measure built by the researcher by converting performance data into fpl_points, following fantasy football point assignments (e.g., 4 points per goal, 3 per assist, etc)

1. Generate summary statistics of these two variables. Do you see any reason why the researchers would like to use their own measure (fpl_points_fbref) instead of the true one?
2. Visualise, using a scatterplot, the relationship between the two variables. Highlight some players by adding their name next to the corresponding scatter plot's point.
3. Based on the Figure, is fpl_points_fbref a good approximation of fpl_points? What can you say about the precision of the researcher's measure compared to the true one?

Exercise 3. Visualisation of estimation results.

Now consider the csv dataset **spillover_results.csv**. This dataset contains the results of the estimation of the following regression, which estimates the effect of a school choice program called Meet The Parents (MTP) on the probability of enrolling in a secondary state vs private school:

$$y_i = \alpha\, MTP_i + \beta\, MTPI_i + \gamma\, MTP \cdot MTPI_i + \delta X_i + v_i$$

Where:

- α: Treatment effects of the MTP program;
- β: Spillover effects of the program;
- γ: Competition effects of the program.

The program consisted in meetings organised within local primary schools where parents have the opportunity to discuss various aspects of their secondary school. You estimate the regression for 4 groups of schools separately: all schools, only those promoted by MTP during their meetings, those promoted by MTP that were undersubscribed (they had less demand than places available) and those promoted by MTP that were oversubscribed (they had more demand than places available). In the dataset, you have the estimated coefficients and the corresponding standard errors, as well as the upper and lower confidence interval (though you are not sure at which level).

Prepare one bar graph showing in an effective way the different estimates for the different subgroups, making sure to include 95% confidence intervals.

Exercise 4: Mapping.

In the "geo" folder, you will find a shapefile of London boroughs. Shapefiles are geospatial data that can be used to plot geographic information. Set the coordinate reference system (crs) to 4326. In the "rege" folder you will find two csv files including, for Greater London:

- demolitions-coordinates-m: includes the list of public housing buildings redeveloped as mixed-income buildings – i.e., a mix of social and private market housing;
- demolitions-coordinates-nm includes the list of public housing buildings redeveloped as non-mixed-income buildings – i.e., fully social housing.

Create a map that plots the 33 London boroughs and the public housing buildings that were regenerated, making sure that the two samples can be clearly identified in the map. What can you infer about the distribution of the regenerations by looking at the map?