Uploaded 20/02/24, 17:16

Group Registration

This assignment can be performed in groups of two to three. Please register the groups until March 21st!

Team Partner Wanted!

We strongly recommend to perform this exercise in groups of two. You can use this forum to find a team partner.

Zoom

Zoom links to open labs and submission talks will be available here.

Exercise 1: Python Notebook (Groupwork)

The goal of this exercise is to get into some first contact with Python, Pandas data frames, and the many opportunities how to quickly generate static charts from data frames.

For the first exercise, load the provided CSV file, do some data cleaning, and check if your cleaning operation was successful using visualization. For inspiration, browse through some Kaggle notebooks. The original data was downloaded from Kaggle.

The main purpose of the notebook is to preprocess the CSV file for further visualization. The following steps have to be performed:

- Read the three CSV using Pandas. See the pandas.read_csv documentation to check how to parse the CSV correctly! Merge the two datasets player_data_per_36_min.csv and players.csv into one table. Add a team_name column to your new data frame based on the team_id column retrieving the name from the teams.csv dataset. You should now have a table that includes player info from **players.csv**, and players performance statistics from **player_data_per_36_min.csv**, and the team's name from **teams.csv**. (**3 points**)
- Take care of missing values. Sci-kit learn provides different data imputation methods. Remove unusable rows or columns, if necessary. If the player has no current team, replace the empty value with "Retired".
  (Hint: You may need to impute before you completely filter your dataframe. You need to make an educated judgment). (**3 points**)
- Create two new tables: one that groups rows based on **player_id**, another that groups rows based on **team_id**. (**2 points**)
- Visualize the data (twice). Every submitted notebook should contain **at least two visualizations** using at least **two different Python visualization libraries**. One visualization for each of the new tables (aggregated players or aggregated teams). A list of the most wide-spread Python visualization libraries can be found in this article. You must concisely describe and explain each visualization and your decisions in a Markdown field. You will not receive the points for this task if you did not add a description/explanation. (max. **5 points** per visualization)
- Save the resulting tables (the cleaned player_data_per_36_min and the two new aggregated tables) as CSV. To be sure that the data is correctly saved, you can load it again. You will have to work with these tables for the second exercise. (**2 points**)

Possible visualizations include, but are not limited to:

- Scatterplots and scatterplot matrices
- Parallel coordinates
- Radar charts
- Bar charts
- Box plots and histograms
- Choropleth maps
- ...

If you do not know these visualizations or you do not know which visualizations to use for which type of data, we recommend that you also attend the InfoVis VO lecture!

Useful links:

- Jupyter Notebook Documentation
- Pandas Data Analysis and Manipulation Tool: Documentation
- Matplotlib
- Pandas visualizations based on Matplotlib
- Seaborn
- Bokeh
- Plotly

Submit your notebook through TU Wien's JupyterHub: