OTH AMBERG-WEIDEN
Fakultät Elektrotechnik, Medien und Informatik
Prof. Dr.-Ing. Christian Bergler, Prof. Dr. Sandra Rebholz
Summer Semester 2024

Ostbayerische Technische Hochschule
Amberg-Weiden

Project Work Programming Starter (PRS)

Start: 14.06.2024 – 18:00:00 Uhr Deadline: 05.07.2024 – 17:59:59

Aim of the Project Work

The aim of this project is to process, analyze, and interpret the data material of the video game FIFA 15 – 23 and also the most recent EA Sports FC 24. The overall data pool provides historical data about all the players, teams, and coaches, differentiated between female and male. The period under review is the years 2015-2024. In particular, the aim is to gain deeper insights into the general development of the soccer market over time.

Modalities

- The project work is done in the programming language **Python**, with strong focus on the data analysis and visualization library, named `Pandas` and `Matplotlib`.
- As a **final result, a Jupyter-Notebook** named `lastname_firstname.ipynb` needs to be created and **submitted electronically via Moodle**, incldluing the entire programcode, together with all analysis results (embedded graphics, cell outputs, additional descriptive text). A separate written manuscript is not required. However, **the results of each of the individual tasks, next to all the derived findings, as well as the methodological procedure, must be documented in the form of a continuous text within the notebook**. In addition, a **set of slides**, named `lastname_firstname.pdf`, must be prepared for the **final presentation (10 minutes, 2-3 minutes Q & A)** and handed in as part of the overall project submission.
- The submitted **program code should be executable in the computer lab MBUT224** with the version of Python installed there. Indicate in the header of the document (Jupyter-Notebook) whether and, if so, which additional packages need to be installed.
- The project work may be completed in **groups of a maximum of two people**. In the case of a two-person submission, it is sufficient for one group member to submit the work electronically. All group members must be named in the document header of the submitted Jupyter-Notebook, as well as in the uploaded zip-archive.
- The documents must be submitted **latest by 5th of July 2024, 17:59:59,** via Moodle.
- A presentation of the final results in the form of an approx. **10-minute short talk** is expected to take place on **Wednesday, July 10, 2024** from **08:00 to 11:15 at MBUT224 & EMI108 (lecture slot)**. The attendance of each group is mandatory. There is no predefined presentation schedule for each group, since the presentation order is determined on site.
- The **written declaration** attached below (see Appendix) **must be signed by all group members, scanned and uploaded together with the submitted documents**.

Requirements and Evaluation Principles

- The code is executable and fulfills the requirements based on the given tasks.
- The code is clearly structured, easy to read, comprehensible and sufficiently commented.
- The code is elegant, efficient and, if available, uses existing Python functions to process the analysis tasks set.
- The submitted Jupyter-Notebook is appealing and clearly laid out. Use the structuring options offered by the Markdown language. The document should have an outline with links to the solutions of the individual tasks and vice versa.
- There should be a final summary of the analysis results and the knowledge gained.
- If you use external sources, list them in a bibliography.
- The insights gained in the individual subtasks are documented in detail visually and textually, while being put into the actual contex. All explanations are clearly formulated, comprehensible and can be substantiated by the data. The editorial quality of the document is included in the evaluation.
- The diagrams created are appealing and clearly laid out and convey a clear message. In particular, they are sufficiently labeled (e.g. title, axis labels, units, etc.).
- The steps carried out during data preparation and analysis are technically and methodically correctly carried out and sufficiently motivated and documented. Describe not only how you proceeded, but also why.
- The results are presented in a proper and convincing way. The quality of the slides, is part of the overall evaluation.

Data Corpus

EA Sports FC 24 Complete Player Dataset (EASD)
The provided datasets contain soccer data from the career mode of FIFA 15 – 23 and also the recent EA Sports FC 24. The overall datapool includes player-, coach-, as well as team-specific information for female and male athletes.
Among many possible statistics and other things, the data allows a broad historical spectrum of comparisons with respect to players, coaches, and teams across the last 10 versions of the video game. The original dataset is available at Kaggle. The data corpus includes the following contextual parts (see also EASD):

- Every player, coach, and team available in FIFA 15 – 23, and also EA Sports FC 24, split between the category female and male, stored in a total of 6 distinct CSV-files:
  - `female_coaches.csv, female_players.csv, female_teams.csv`
  - `male_coaches.csv, male_players.csv, male_teams.csv`
- 109 attributes for players, 8 attributes for coaches, and 54 attributes for teams
- Player positions, with the role in the club and in the national team
- Player attributes with statistics as pace, shooting, passing, dribbling, defending, and others
- Player personal data like nationality, club, date of birth, wage, salary, and others
- Team data regarding their coaches, their overall value, and tactics
- (...)

Project Tasks

Task 1 (Data Preprocessing)
The aim of this task is to prepare and preprocess the original EASD data corpus in order to build an appropriate and solid data foundation for any downstream data analyses & engineering approaches.

a) Read in the individual CSV-files `female_coaches.csv`, `female_players.csv`, `female_teams.csv`, `male_coaches.csv`, `male_players.csv` as well as `male_teams.csv`, and store all the information in stand-alone `Pandas-DataFrames` named `df_feco` (female coaches), `df_fepl` (female players), `df_fete` (female teams), `df_maco` (male coaches), `df_mapl` (male players), `df_mate` (male teams).

b) Verify the number of total columns for each `DataFrame` to check whether the numbers of overall attributes match with the numbers given in the official documentation. For the players (`df_fepl` & `df_mapl`) it should be 109, for the teams (`df_fete` & `df_mate`) it should be 54, and for the coaches (`df_feco` & `df_maco`) it should be 8 total attributes. How many data entries for players, teams, and coaches are provided by the EASD corpus?

c) Add an additional column with the name `Sex` to each of the 6 DataFrames and fill it accordingly with the String `male` or `female`. Afterwards, combine and merge the two sex-specific DataFrames for players (`df_mapl`, `df_fepl`), teams (`df_mate`, `df_fete`) and coaches (`df_maco`, `df_feco`), resulting in a total of 3 new DataFrames, named: `df_player`, `df_team`, and `df_coach`.

d) Create an updated and reduced version of `df_player`, `df_team`, and `df_coach` as the very final data foundation, by overwriting the respective DataFrames in such a way that only the subsequently mentioned attributes (columns) remain:

- Players (27 out of 109): `player_id, fifa_version, long_name, player_positions, value_eur, wage_eur, dob, height_cm, weight_kg, club_team_id, club_name, league_id, league_name, club_position, club_jersey_number, club_joined_date, club_contract_valid_until_year, nationality_id, nationality_name, nation_team_id, preferred_foot, pace, shooting, passing, dribbling, defending, physic`
- Teams (18 out of 54): `team_id, fifa_version, team_name, league_id, league_name, nationality_id, nationality_name, overall, attack, midfield, defence, coach_id, home_stadium, whole_team_average_age, captain, penalties, left_corner, right_corner`
- Coaches (4 out of 8): `coach_id, long_name, dob, nationality_name`

e) Check all the individual DataFrames (`df_player`, `df_team`, `df_coach`) for missing values and fill them with appropriate replacements, wherever it is reasonable, in order to guarantee a data collection as clean and error-free as possible. Be careful in which incidents replacing is not useful, where it is appropriate, and which replacement values make sense or not. Evaluate, verify, and comment on your decisions, next to reporting a feedback about the data quality, using various quality criteria.

Task 2 (Explorative Data Analysis)
The aim of this task is to gain deeper insights into the EASD data corpus using various statistics and data processing techniques, in order to derive interesting, analytic, and descriptive insights, all together presented and visualized using various diagrams. For all downstream tasks, please always use the pre-processed & updated DataFrames – `df_player`, `df_team`, `df_coach` – as the starting point, comprising all the (female/male) information. Please always try – if possible – to verify the plausibility of your results. Comparatively often the original data themselves contain implausible and/or incorrect information.

a) Rename and convert the column `value_eur` to `value_mio_eur` in the DataFrame `df_player` and change the numerical values accordingly, whereby an accuracy up to the 6th decimal place should be retained (e.g. 51000000 to 51.000000 or 325800 to 0.325800). In addition, the unit of the `height_cm` column should be changed. Rename the column to `height_m` and convert the numbers accordingly, while keeping an accuracy of 2 decimal places (e.g. 188 to 1.88). Furthermore, convert the date of birthday (column `dob`) from `YYYY-MM-DD` to the desired format `DD-MM-YYYY`.

b) Calculate the total number of (unique) players that have been involved in the video game over all the years (Note: a player can appear several times in the DataFrame, for example Kevin De Bruyne is part across all versions – FIFA 15 – 23 and EA Sports FC 24). Who is currently the youngest player in the game?

c) Which player is the smallest/tallest and thinnest/thickest across all FIFA versions? Is it possible to draw conclusions and associations with a player's skills from his physical constitution (e.g. slim players are fast but less robust, tall players are rather slow but more robust).

d) Which player has received the highest wage (`wage_eur` refers to the weekly player salary) since the start of the game and who is the highest paid player in the current EA Sports FC 24? Return the top-10 highest paid players in the current EA Sports FC 24 game!

e) What are the names of all the German stadiums that were involved in FIFA15 up to EA Sports FC 24?

f) Who is the oldest active coach nowadays?

g) Visualize all different national teams for male athletes, together with the current (2024) number of associated players per team (Note: not every player of a specific nation is part of the national team – see `nation_team_id`)?

h) Visualize the relation between body size, body weight, and pace of a player as part of a Scatter-Plot only considering the EA Sports FC 24 version?

i) Determine the average skill level of each player over the years in each version of the game, by computing the mean including the `pace, shooting, passing, dribbling, defending, & physic` attribute. Which female and male player in the current EA Sports FC 24 version has the best average skill? Which player has the best average skill ever measured across all versions and who are the top-10 ever?

j) How are the squad strengths distributed in the latest version of the game? Calculate the number of players belonging to each team (excluding national teams, only league teams), sort everything in descending order, and visualize the output accordingly!

k) How many teams (excluding national teams) are part of the game in each version (FIFA15 up to EA Sports FC 24)? Is there a recognizable trend considering the overall changes in the number of teams throughout the distinct versions?

l) Visualize the distribution of all the different nations and their associated teams (excluding national teams, only league teams) for each year (version of the game). In addition, identify for each year the nationality of every trainer from the distinct league teams, and visualize all the different nations plus the amount of associated coaches. Is there a temporal and deomographic trend between the number of coaches and teams per nation throughout all the years? Which country holds the most clubs and which country holds the most coaches across all the years?

m) Analyze and visualize the chronological evolution of the averaged skills, based on `pace, shooting, passing, dribbling, defending, & physic` – see also (i) – for the following players from the first game version to the latest game version: `Alexia Putellas (female), Erling Haaland (male), Kilian Mbappe (male), Lionel Messi (male), Christiano Ronaldo (male)`. Please also analyze, compare, and present their associated progression of wages across all the game versions! Is it possible to recognize an overall trend between skill and salary?

Appendix to the Project Work Programming Starter

Summer Semester 2024

Prof. Dr.-Ing. Christian Bergler, Prof. Dr. Sandra Rebholz

Please fill in the following declaration either together or for each group member and upload a scanned version to Moodle with your submission. A signature must be provided by each of the group members.

Last name, first name – Group member 1:

Matriculation number – Group member 1:

Last name, first name – Group member 2:

Matriculation number – Group member 2:

Declaration

I/We hereby declare that the submitted project work was created exclusively by the above-mentioned persons. All aids and sources used have been referenced in the work.

Place, Date

Signature(s)