

STAT432 Final Project Proposal

Project Name

Investigations on the FIFA20 dataset

Team Member

Chen Si (chensi3)

Yuchen Cao (yuchenc5)

Dataset

The original dataset ([players_20.csv](#)¹) is obtained from Kaggle. This dataset contains FIFA20 player data. We pre-process the dataset with python to only obtain the data columns that we want to use in this final project and removed all data of Goalkeepers, which has a different record form than other players. The updated dataset [players_20_edited.csv](#)² contains 16242 non-goalkeeper soccer players and their physical data, scores on skills, and positions.

The full list of columns is:

Basic Info:

'short_name', 'long_name', 'age', 'height_cm', 'weight_kg',
'nationality', 'club', 'overall', 'potential', 'value_eur', 'wage_eur',
'player_positions', 'preferred_foot', 'international_reputation';

Featured Scores (a player's different professional abilities):

'weak_foot', 'pace', 'shooting', 'passing', 'dribbling', 'defending',
'physic', 'attacking_crossing', 'attacking_finishing',
'attacking_heading_accuracy', 'attacking_short_passing',
'attacking_volleys', 'skill_dribbling', 'skill_curve',
'skill_fk_accuracy', 'skill_long_passing', 'skill_ball_control',
'movement_acceleration', 'movement_sprint_speed', 'movement_agility',
'movement_reactions', 'movement_balance', 'power_shot_power',
'power_jumping', 'power_stamina', 'power_strength', 'power_long_shots',
'mentality_aggression', 'mentality_interceptions',
'mentality_positioning', 'mentality_vision', 'mentality_penalties',
'mentality_composure', 'defending_marking', 'defending_standing_tackle',
'defending_sliding_tackle';

Position Scores (a player's ability to play different positions):

'ls', 'st', 'rs', 'lw', 'lf', 'cf', 'rf',
'rw', 'lam', 'cam', 'ram', 'lm', 'lcm', 'cm', 'rcm', 'rm', 'lwb', 'ldm',
'cdm', 'rdm', 'rwb', 'lb', 'lcb', 'cb', 'rcb', 'rb', 'classification'

¹ The original dataset [players_20.csv](#) could also be found at https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?resource=download&select=players_20.csv if the link fails.

² The updated dataset [players_20_edited.csv](#) could also be found at <https://uofi.box.com/s/91m9qt18v7zhxh00d18edv7fx6x95fsw> if the link fails.

Goal of analysis

The goal of this project is to try to figure out the relationship between the player's overall score and his different ability scores. We would also like to discover differences between differences between players' positions based on their various abilities (passing, shooting, dashing, etc.).

The following table demonstrates some of the responsible variable – explanatory variable pairs that we would like to discover in the project.

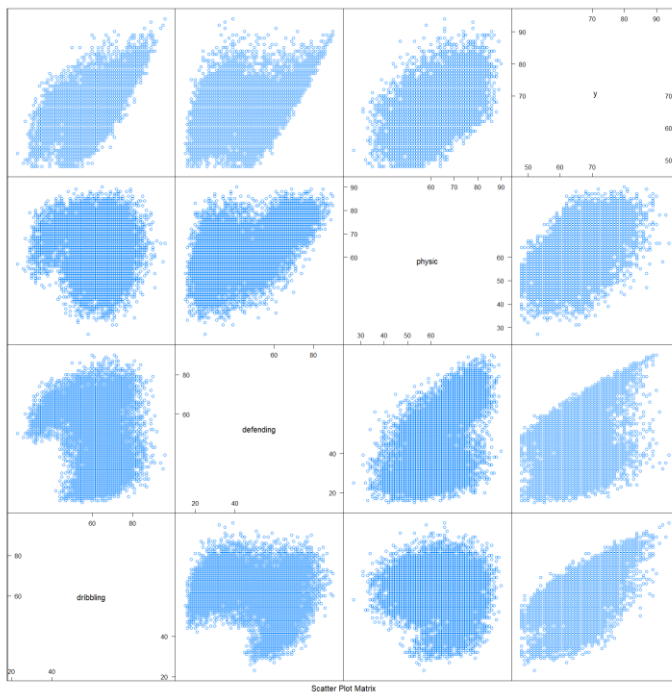
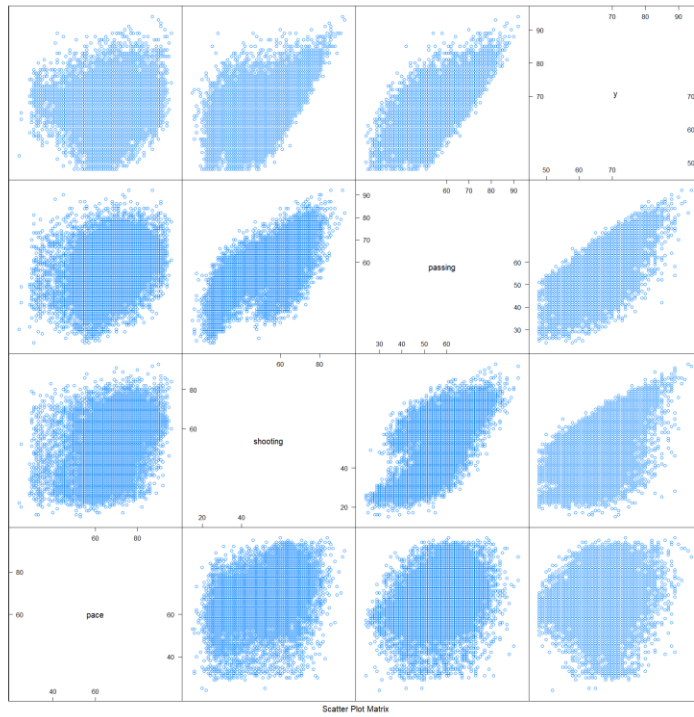
Response Variable	Explanatory Variable	Comments
Overall Score	Abilities (pace, shooting, passing, ...)	Explore how the game developer come up with a player's overall score given the player's different ability scores.
Potential Score	Nationality, Age, Abilities	Explore how the game developer come up with a player's potential score given the player's nationality, age and different ability scores.
Striker Score	Abilities	Explore how the game developer come up with a player's performance as a striker given the player's different ability scores.
L/RB, L/RWB	Abilities	Explore the difference between a player as a "back" or as a "Wing Back".
...

Preliminary Visualization

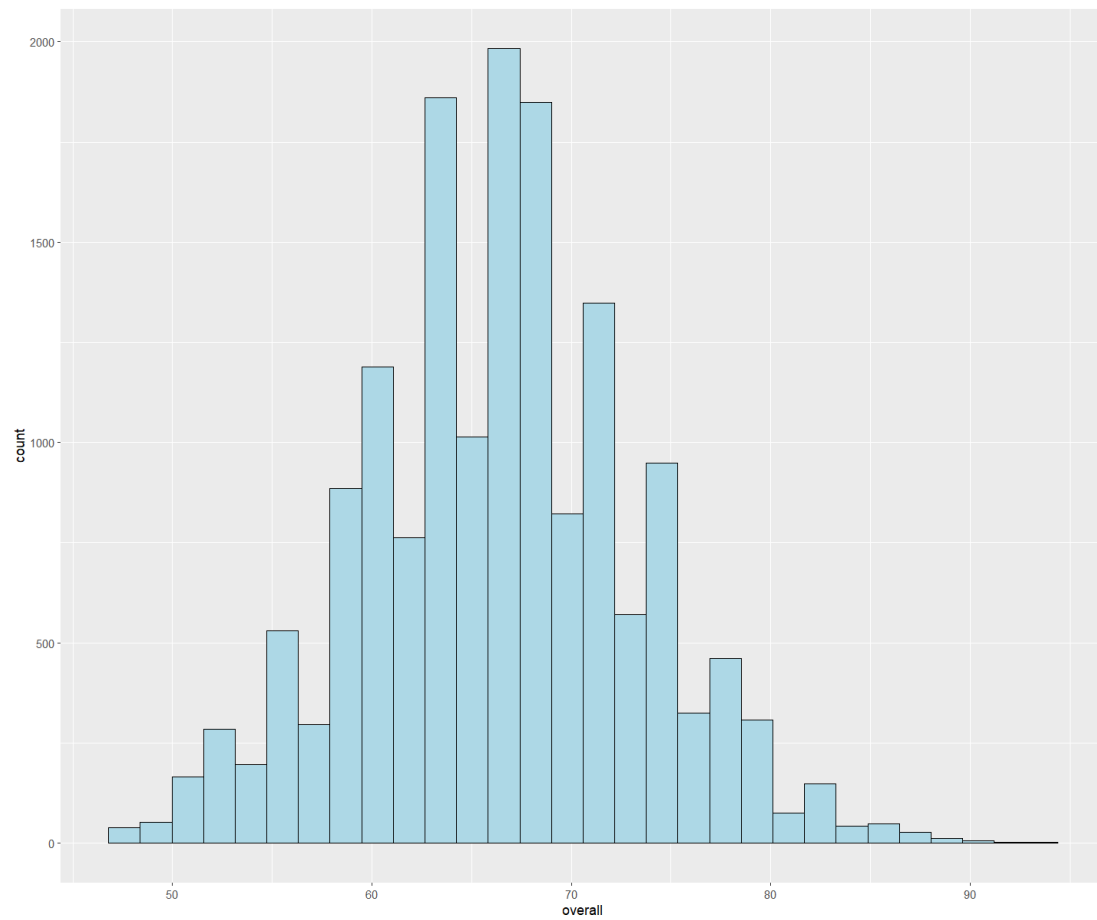
Some of the scatterplot matrices that we conducted as EDA are

1. Overall ~ .

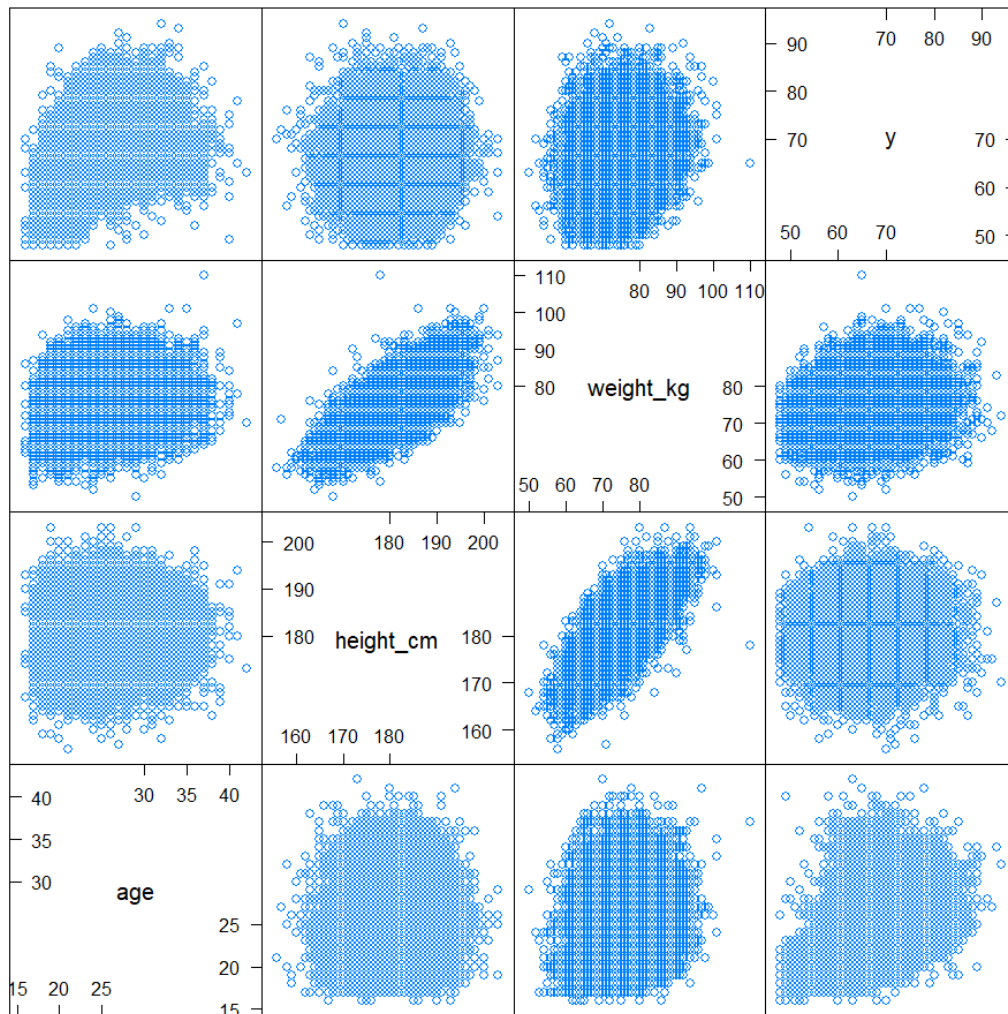
Distribution of overall vs passing, shooting, pace, dribbling, defending, physics scores.



2. Histogram of Overall scores

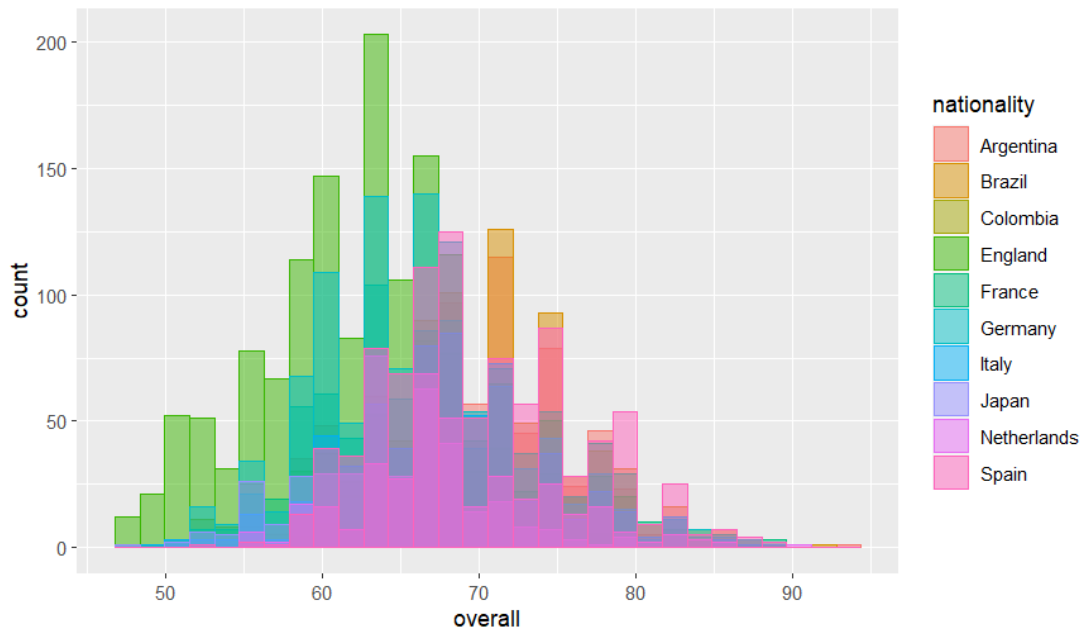


3. Overall \sim age + weight + height



Scatter Plot Matrix

4. Histogram of overall scores for Top 10 frequent nations



5. Word count of all players' nationalities

