**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](https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?resource=download&select=players_20.csv)[[1]](#footnote-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](https://uofi.box.com/s/91m9qt18v7zhxh00d18edv7fx6x95fsw)[[2]](#footnote-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'

**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.

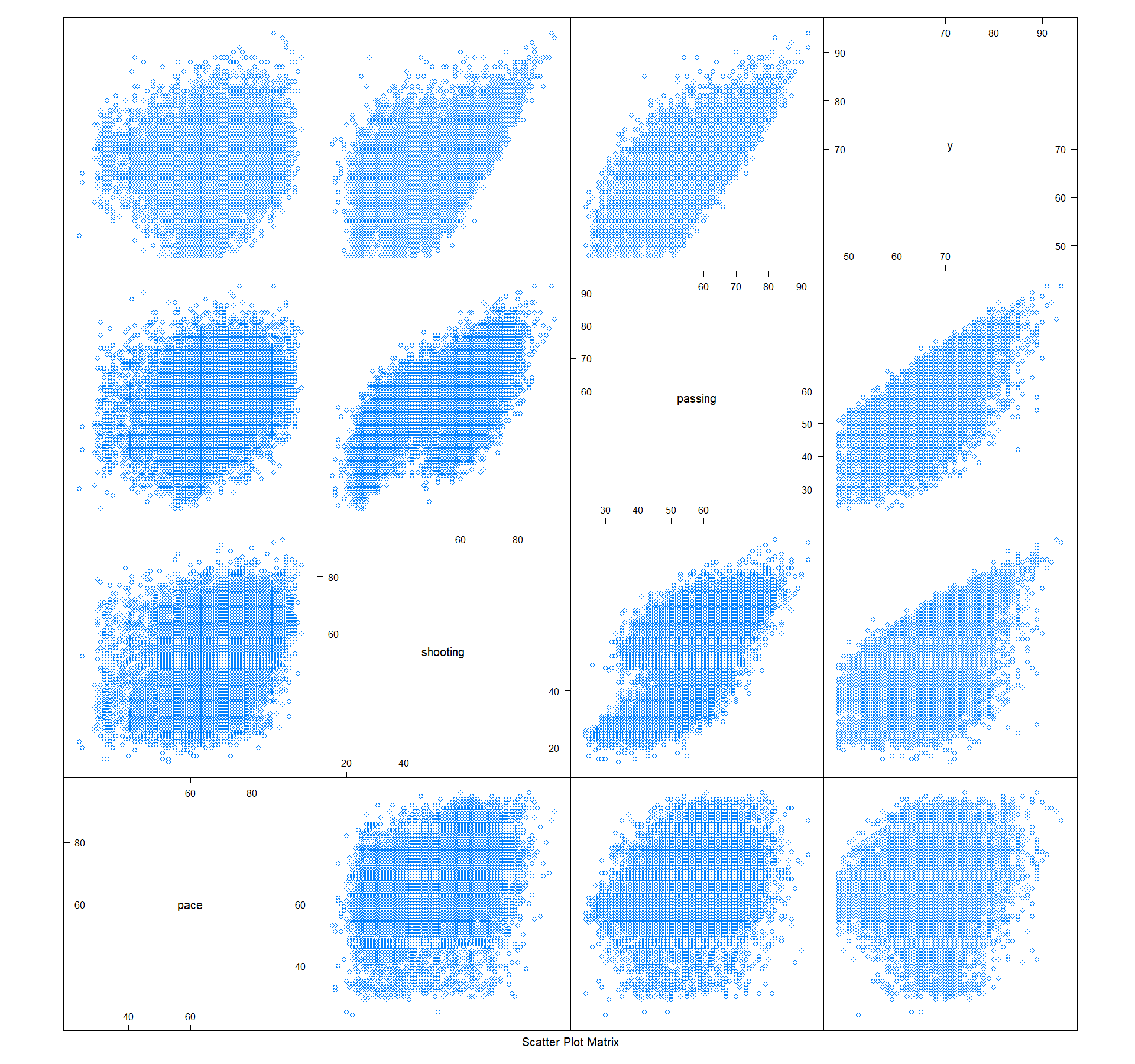
|  |  |  |  |
| --- | --- | --- | --- |
| Response Variable | Explanatory Variable | Comments | Type of Learning |
| Overall Score | Abilities  (pace, shooting, passing, …) | Explore how the game developer come up with a player’s overall score given the player’s different ability scores. | Regression |
| 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. | Regression |
| 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. | Regression |
| L/RB, L/RWB | Abilities | Explore the difference between a player as a “back” or as a “Wing Back”. | Classification |
| … | … | … |  |

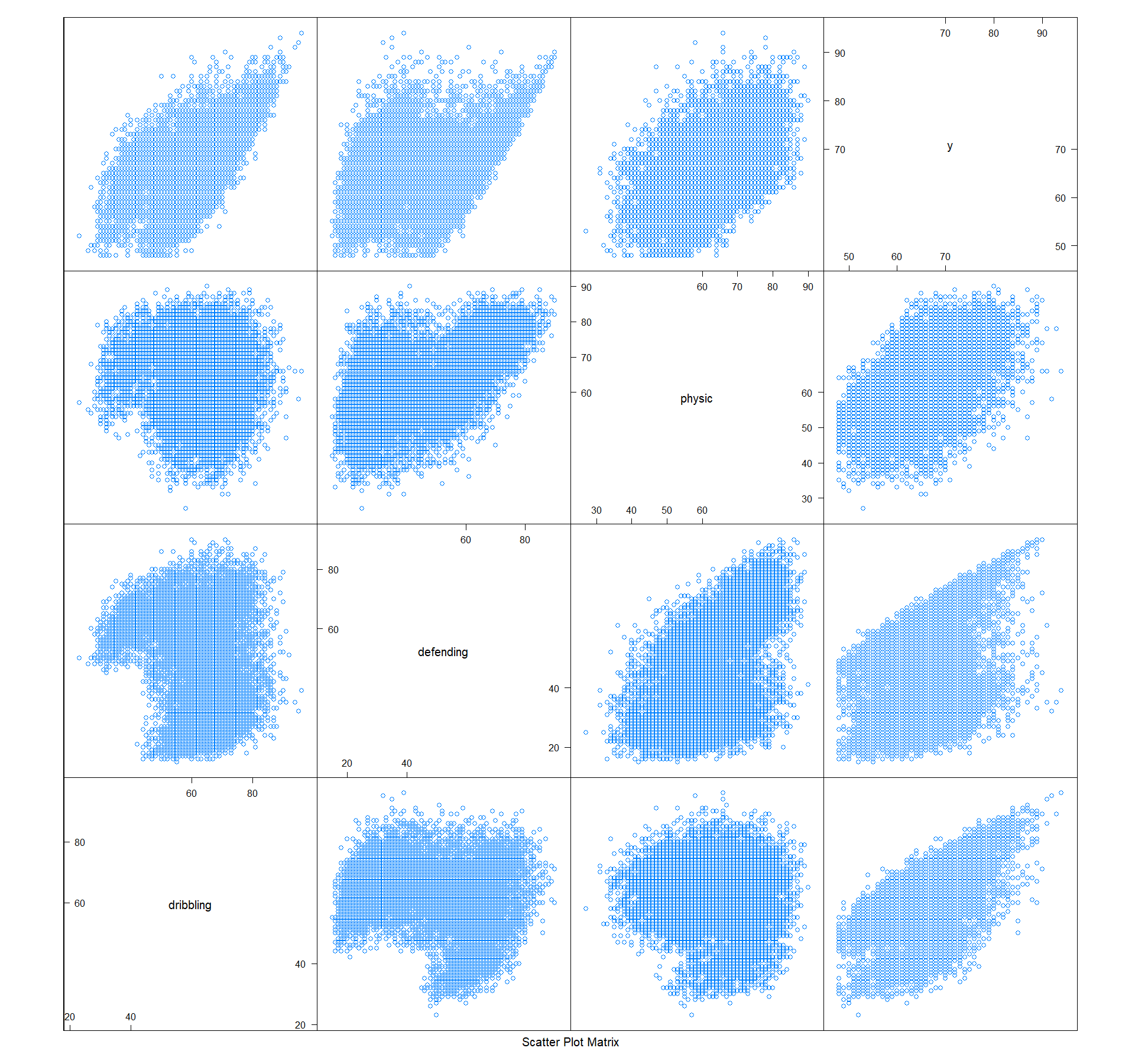
**Preliminary Visualization**

Some of the scatterplot matrices that we conducted as EDA are

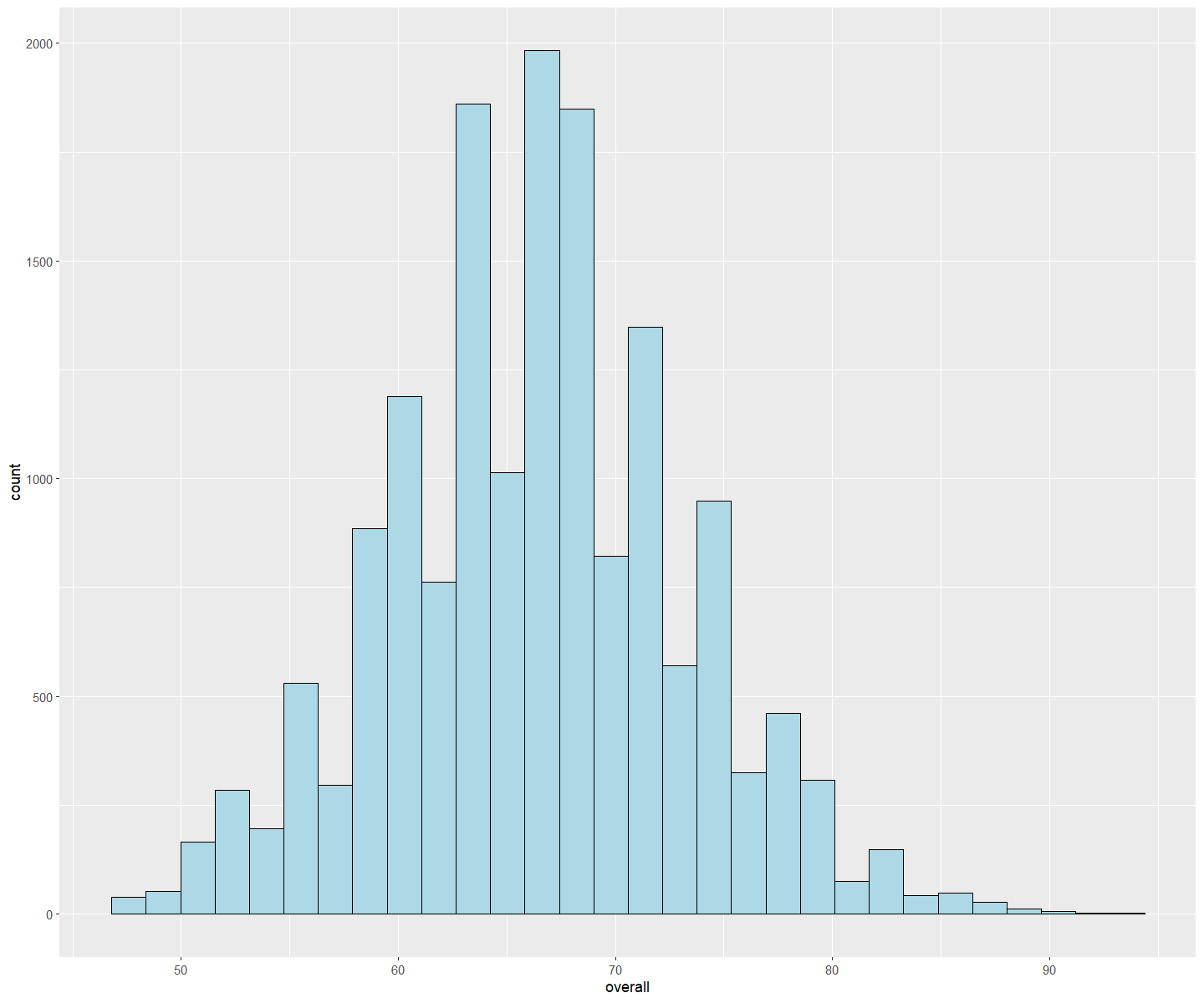
1. Overall ~ .

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

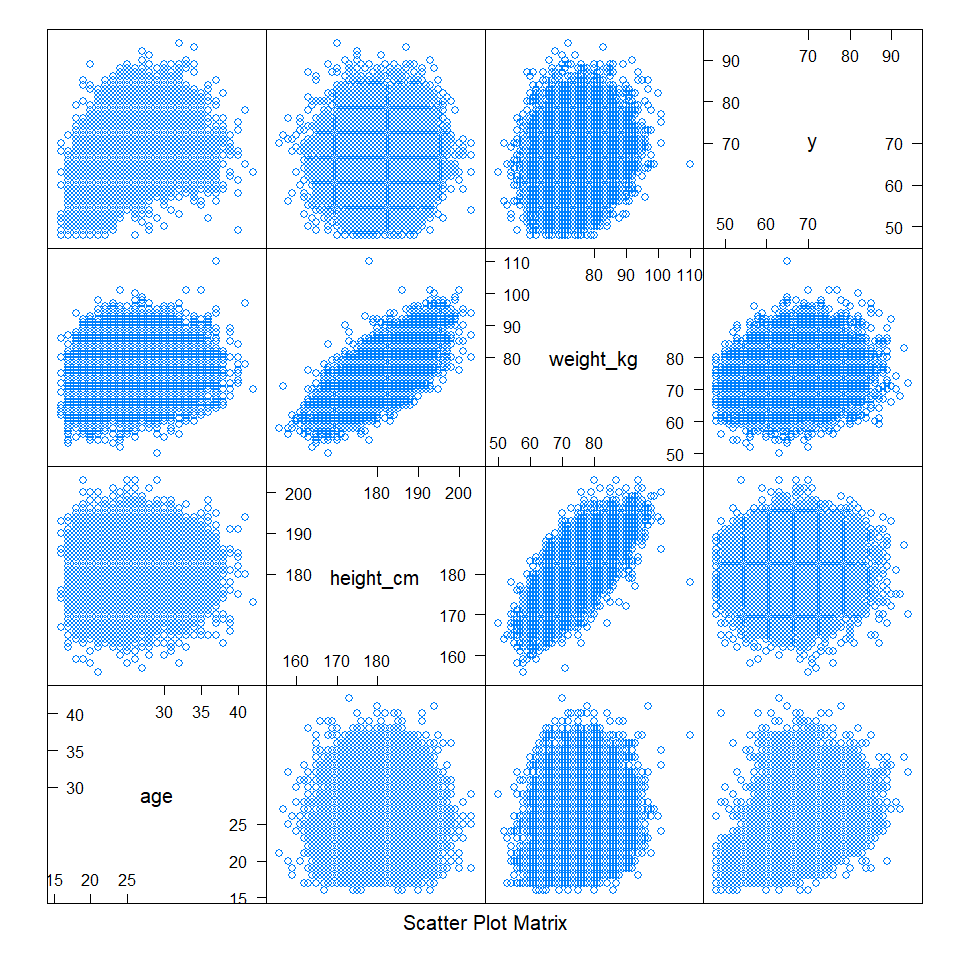




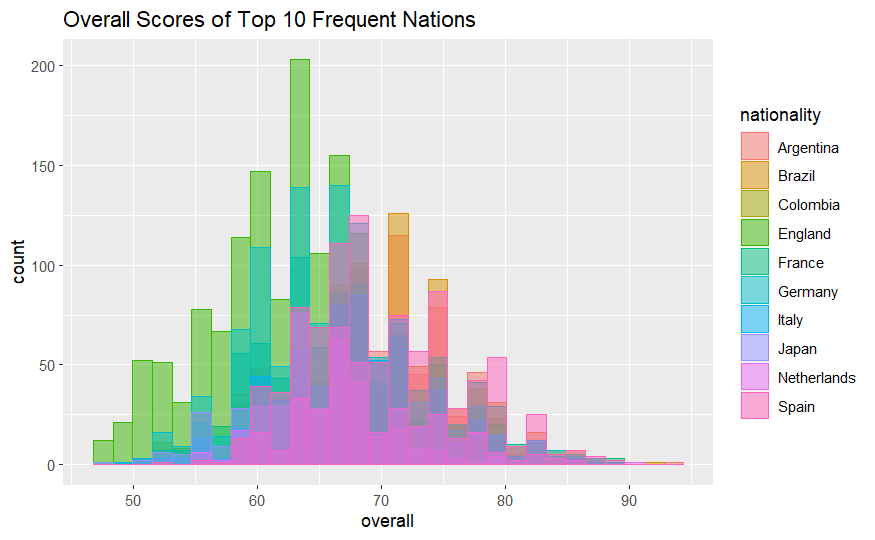
2. Histogram of Overall scores



3. Overall ~ age + weigth + height



4. Histogram of overall scores for Top 10 frequent nations



5. Word could count of all players’ nationalities



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. [↑](#footnote-ref-1)
2. The updated dataset players\_20\_edited.csv could also be found at https://uofi.box.com/s/91m9qt18v7zhxh00d18edv7fx6x95fsw if the link fails. [↑](#footnote-ref-2)