SI 618 Team Final Project

Football player and Analysis through EA Sports FIFA 24

Tzu-Yu Peng (typeng)
Shu-Ting Lin (shooting)
Chia-Yun Li (chiayun)

I. Motivation

In recent years, the growing popularity of the FIFA World Cup has sparked our interest in football. Coincidentally, a team member who has a solid understanding of the subject explained to us some of the regulations regarding the selection of soccer stars. Intrigued by this information, we aim to analyze whether a player's position, personal information, and other relevant characteristics can impact their salary.

Our team wishes to identify specific player traits, such as position, and their characteristics like weight, height, club, or league, and observe their correlation with player salaries. The ultimate goal is to develop a predictive model that can anticipate a player's salary based on these features. This model would serve as a valuable tool, enabling us to make salary predictions for new players by inputting their characteristics. We believe that understanding these relationships will provide valuable insights into the factors influencing a player's compensation in the world of football.

II. Data Sources

Our datasets for our project are from Kaggle. The primary dataset for this project encompasses football player data sourced from "EA Sports FC 24 Complete Player Dataset". It includes crucial features such as player attributes assigned by EA Sports (e.g., overall rating, dribbling, passing), personal details (age, nationality), and club-related information (wage, value, club name). The dataset we used is for male players, containing approximately 180,000 records, which is presented in CSV format.

The first secondary dataset is associated with football teams and encompasses features such as team name, league information, nationality, and statistics. This dataset, which comprises around 1,000 records, is also available in CSV format, and sourced from " EA Sports FC 24 Complete Player Dataset".

The second secondary dataset, "FIFA 23 Players Dataset", comprises features like wage value, full name, nationality, position, similar to the primary dataset. Similarly, it is presented in CSV format and contains approximately 1,000 records.

The final dataset we used for analysis and predictions underwent data preprocessing, resulting in approximately 13500 rows and 80 columns.

III. Data Manipulation Methods

Firstly, we conducted data cleaning to focus on the most recent updates in our dataset, spanning from September 17, 2014, to September 22, 2023. Specifically, we prioritized players whose data had been updated as of September 22, 2023. Additionally, to enhance the effectiveness of our analysis and predictions, we opted to concentrate on the league with the highest proportion, identified by the league_id equal to 1.

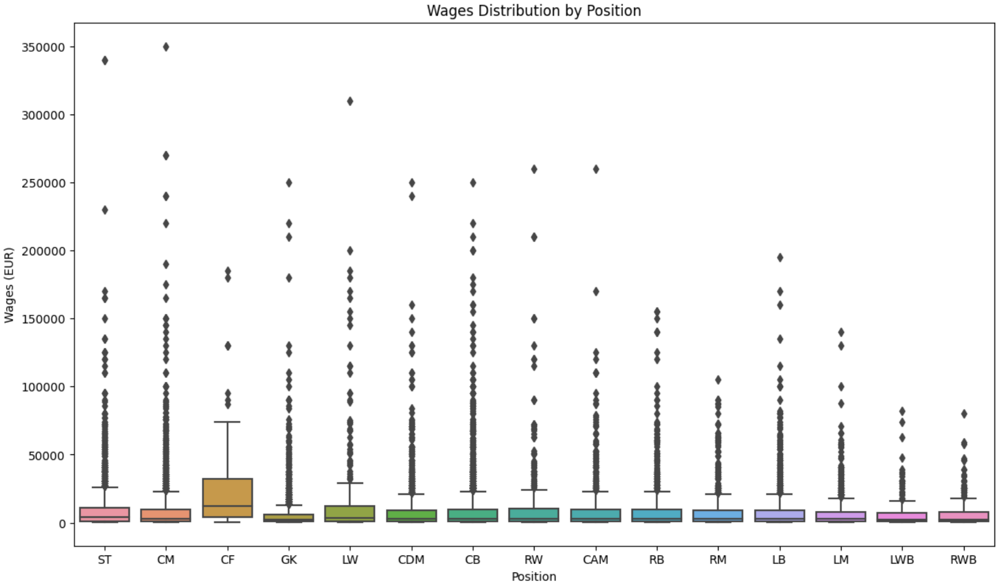
Given that the dataset comprises players from various leagues, we aimed for a more refined analysis by considering only the predominant league (league_id = 1). Furthermore, recognizing that a player may have multiple positions recorded in this dataset, we retained only the primary position. This primary position corresponds to the first occurrence in the player_position column.

Subsequently, we engaged in data preprocessing, eliminating entries with missing values. For the remaining categorical data, we applied one hot encoding to transform them into an analyzable format. The merging of player and team data is performed using a column in the player data named "club_team_id" and a column in the team data named "team_id."

IV. Analysis

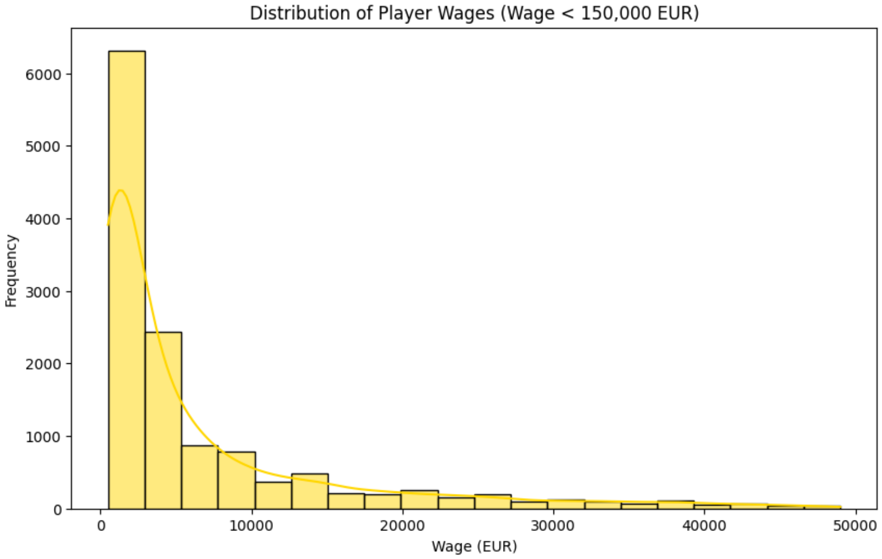
1. Visualization of Analysis

Observing the boxplot depicting wages and positions (plot a) reveals some intriguing findings. Firstly, the highest wage is associated with the position 'CM,' signifying 'Centre Midfielders,' while the highest average wage is found in the position 'CF,' denoting 'Centre Forward.' Notably, the highest quantile for all positions is also 'CF.' It's important to note that 'SUB' and 'RES' positions, being non-specific in the game, are excluded from our analysis.



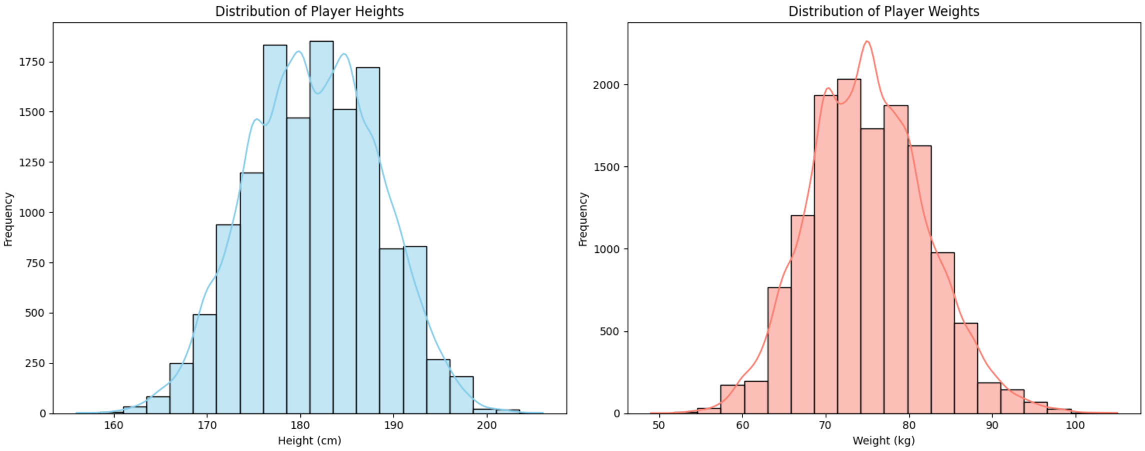
Plot a. Wages Distribution by Position

Second, moving on the distribution of players' wages (plot b), it can be observed that the distribution is right-skewed distribution, with players' wages concentrated primarily in the range of 0-500 euros.



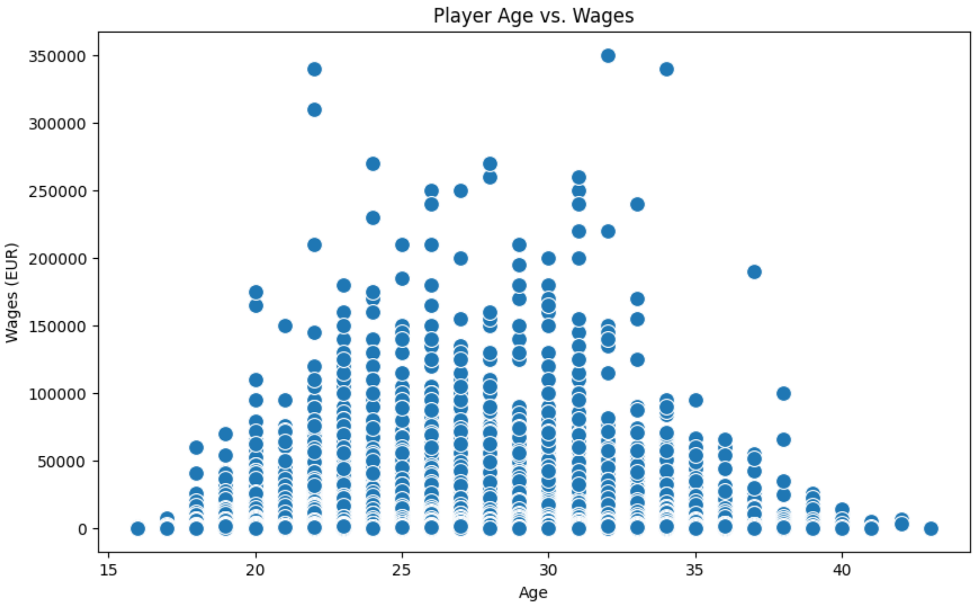
Plot b. Distribution of Players' Wages

Next, we want to observe the distribution of players' personal information (plot c), such as height and weight. From the histogram plot, we learn that these two characteristics both are normal distribution, which are aligned with our expectations.



Plot c. Distribution of Players' Heights and Weights

To know whether there are some relationships between players' age and wages, from the scatterplot of age and wages, out of our expectation, the higher wages didn't show at the middle of the plot, the range of 25 to 30, which I assume that age would be normal distribution.

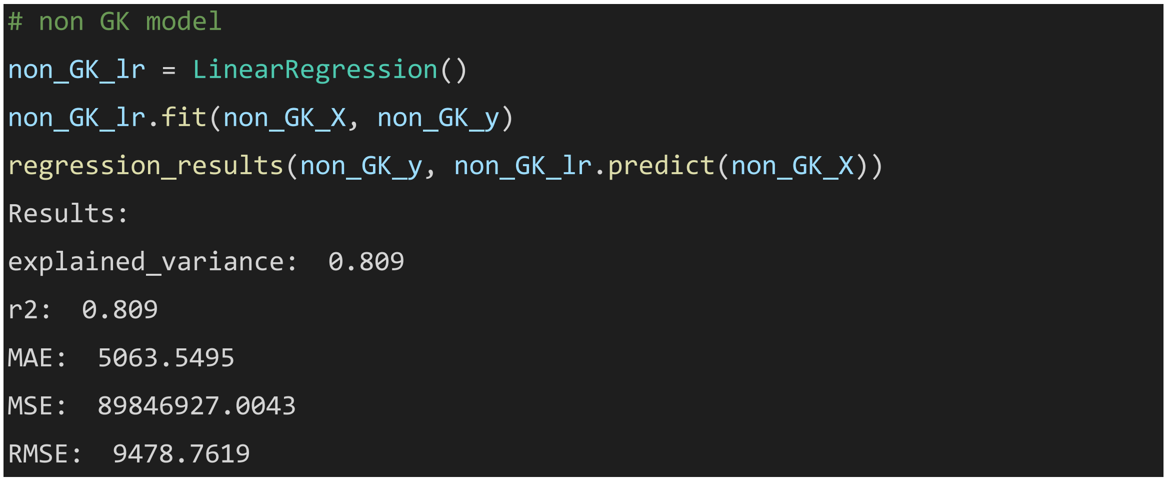


Plot d. Players' Age vs. Wages

2. Linear Regression Analysis and Prediction

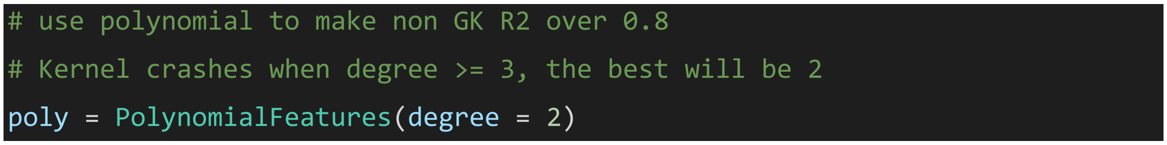
We aim to find a prediction model have well performance on predicting wages. First of all, we split the data two subsets, one is for position “GK”, which means goal keeper, and the other is non “GK”, for the reason why we split the data is that the skill sets required for goalkeepers differ significantly from those of attacking players such as forwards. Goalkeepers, for instance, do not emphasize abilities like shooting or dribbling as much. Therefore, the standout aspects of their abilities vary, influencing the resulting salary outcomes. The distinct nature of the required skills in these positions contributes to divergent factors affecting salary considerations.

We conduct linear regression analysis, using the subset of data with positions are not “GK”, for the basic linear regression model, its R-squared is 0.809, so this initial linear regression model demonstrates a reasonably good fit to the data. However, there is room for improvement in terms of prediction accuracy.

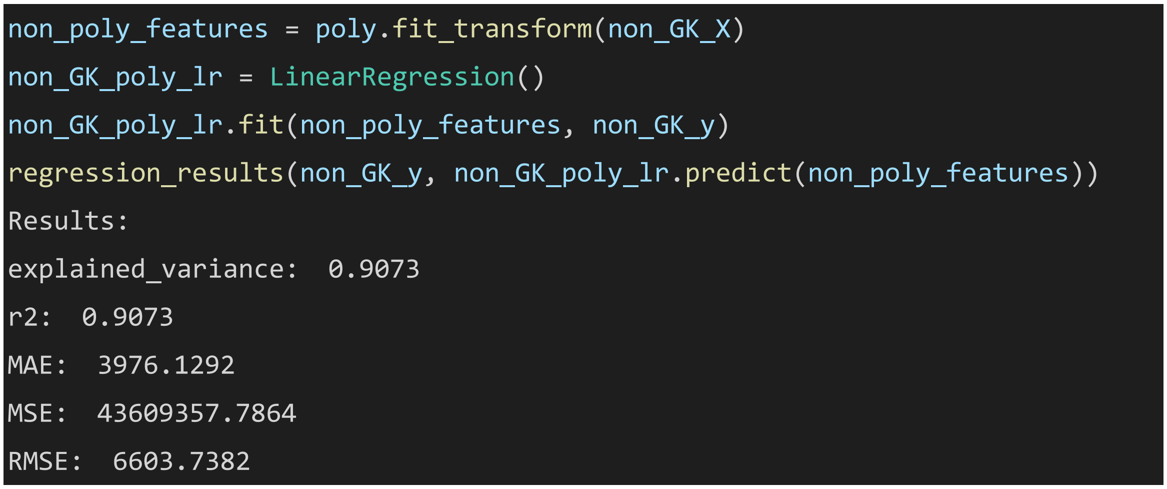


```
# non GK model
non_GK_lr = LinearRegression()
non_GK_lr.fit(non_GK_X, non_GK_y)
regression_results(non_GK_y, non_GK_lr.predict(non_GK_X))
Results:
explained_variance:  0.809
r2:  0.809
MAE:  5063.5495
MSE:  89846927.0043
RMSE:  9478.7619
```

To enhance the model's performance, we employed a polynomial regression approach, introducing quadratic features with a degree of 2, and it has higher R-squared 0.9073, which is relatively high in most linear regression models. This indicates that the model with quadratic features better captures the underlying patterns in the data, resulting in reduced errors (lower MAE, MSE, and RMSE). The enhanced model demonstrates a more accurate representation of the salary prediction for non-goalkeeper players.



```
# use polynomial to make non GK R2 over 0.8
# Kernel crashes when degree >= 3, the best will be 2
poly = PolynomialFeatures(degree = 2)
```



```
non_poly_features = poly.fit_transform(non_GK_X)
non_GK_poly_lr = LinearRegression()
non_GK_poly_lr.fit(non_poly_features, non_GK_y)
regression_results(non_GK_y, non_GK_poly_lr.predict(non_poly_features))
Results:
explained_variance:  0.9073
r2:  0.9073
MAE:  3976.1292
MSE:  43609357.7864
RMSE:  6603.7382
```

We do same thing on the subset of data with position “GK”, there is also an improvement between the basic linear regression model with default parameters and the linear regression model with polynomial regression approach employed.

The result is:

Model	Linear Regression	Linear Regression with Polynomial Expansion
Explained Variance	0.7917	0.9457
R-squared	0.7917	0.9457
MAE	3931.0225	2551.4572
MSE	60290691.7806	15709857.7321
RMSE	7764.7081	3963.5663

Table a. Comparison of position “GK” about Linear Regression

For data pertaining to the “GK” (goalkeeper) positions, the basic linear regression model yields an R-squared value of 0.7917. Upon employing polynomial expansion, the R-squared significantly improves to 0.9457. This indicates that linear regression is highly suitable for constructing a model with this dataset.

In our pursuit of a well-performing model, we conducted model analysis. To ensure optimal performance, we aimed for a model with a Mean Squared Error (MSE) lower than 6000. Focusing on the subset data for positions excluding “GK,” we implemented various models. Notably, the Stacking Regressor, a hybrid of XGBoost, Gradient Boosting, and Random Forest Regressor, was among the models considered.

Model	**MSE**
Linear Regression	9658
Lasso	9654
Ridge	9657
Random Forest Regressor	5984
Adaboost	8375
Gradient Boosting	5527
XGBoost	**5238**
Stacking Regressor	5931

Table b. Comparison of Models for position "GK"

XGBoost has best performance, its MSE is 5238, which is the lowest in these eight models, and the last three hybrid model reach our goal of MSE lower than 6000.

Model Analysis of subset data with positions "GK", we try to use linear regression model first, and its MSE is 6657, which almost reach our goal, so we try the best performance model in "non GK" data, and its MSE becomes 4461, which is lower than 6000.

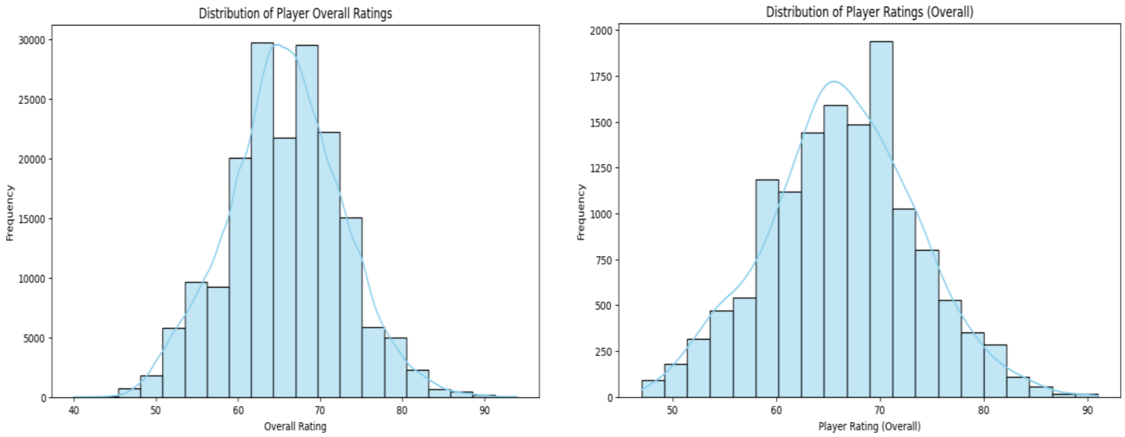
Model	**MSE**
Linear Regression	6657
XGBoost	**4461**

Table c. Comparison of Models for position excluding "GK"

Besides, we also try to split the "GK" subset to different positions, but the result didn't become better, and we also tune the hyperparameters when using linear regression model, it took us lots of time and we tuned about 750 different hyperparameters combination, but the result was worse than the default linear regression model.

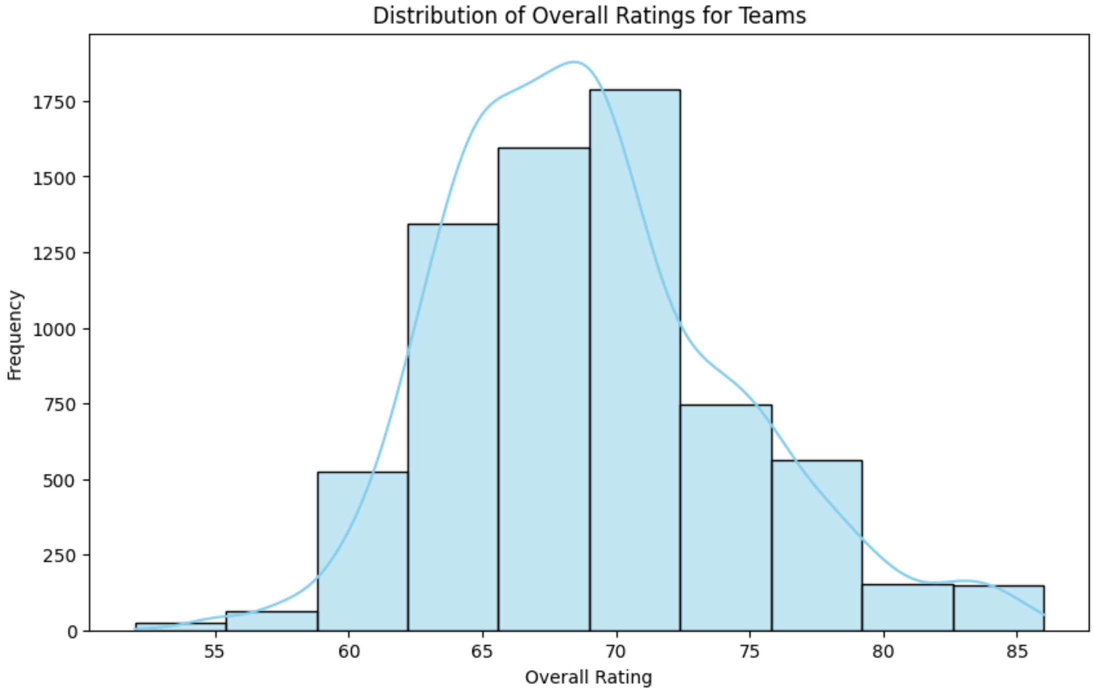
V. Visualization

First, we want to make sure of the ratings of the players, we think the ratings of players are highly related with the players' wages. Thus, we need to know the distribution of the ratings. Besides, we conduct data preprocessing, so we compare the result of original dataset and the modified dataset.



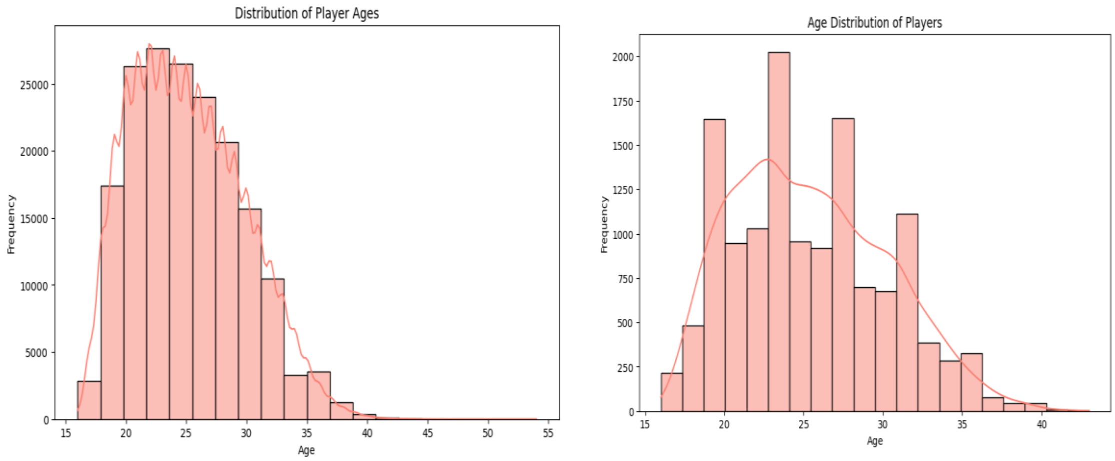
Plot e. Comparison of the Distribution of Players' Ratings

Compare the distribution between players, these two plots all show normal distribution, and we also compare the distribution of ratings for teams.



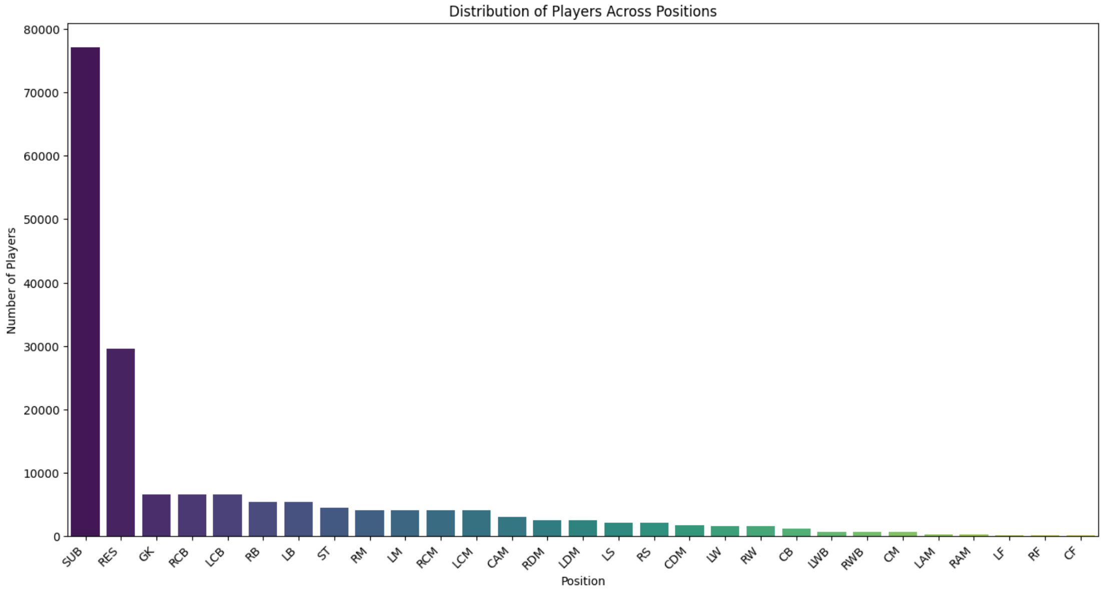
Plot f. Distribution of Teams Ratings

Second, we make a histogram about player's age, and it shows that most of the players are at 20-25 years old. Also, we compare this result with the distribution of players' ages came by the modified dataset. (Left is by the original dataset, and the right is by the cleaning dataset)



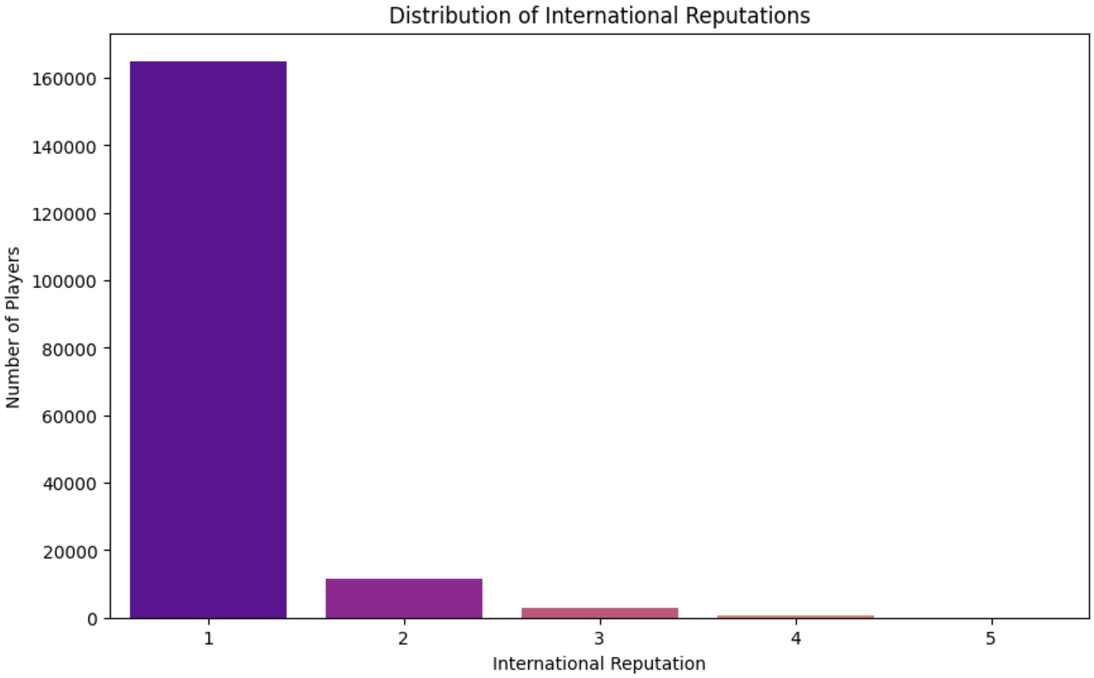
Plot g. Comparison of the Distribution of Players' Ages

Third, look at the distribution of players' position, the main position is "SUB", which means substitute players, it is normal in this dataset. The second one is "RES", which means reserve players. For other positions, their proportion are similar.



Plot h. Distribution of Players' Position

Next, we plot the distribution of international reputations, we learn that this is not a balanced feature; in this dataset, the majority of players have an international reputation score of 1. Therefore, in terms of this feature, there is not a significant variation among players.



Due to space constraints, we have selected a few key charts for analysis and explanation. Our code contains more in-depth exploratory data analysis (EDA) insights.

VI. Statement of Work

Member	Work
Tzu-Yu Peng	Data Preprocessing, Model Analysis
Shu-Ting Lin	Data Preprocessing, Visualization
Chia-Yun Li	Data Preprocessing, Report writing

VII. Reference

1. Bell, A., & Brooks, C. (2017). "The footballers' wage curve: New evidence from the financial crisis." Applied Economics Letters, 24(5), 305-308.
2. Pena, J. M., & Touchette, H. (2017). "A network theory analysis of football strategies." Advances in Complex Systems, 20(02), 1650013.
3. Lago-Penas, C., Lago-Ballesteros, J., Dellal, A., & Gomez, M. (2018). "Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league." Journal of Sports Science & Medicine, 17(2), 289.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning." Springer.

5. Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
6. Tukey, J. W. (1977). "Exploratory Data Analysis." Addison-Wesley.