Mini Project 2

Ruilin Wu

June 10th 2024

1 Executive Summary

This project analyzes 11 datasets from season 0 to season 10. Each dataset contains 8 original variables, with an additional season variable added manually to facilitate combining all datasets into a single dataset.

- Categorical Variables: id, location, gender, sport, and season.
- Numeric Variables: player_value, training_hours(biweekly), age, and injuries
- Using Season 0 data as the baseline data, analyze the overall situation and the relationship between player value and other variables.
- Comparing seasons against player_value by randomly selects 10 individuals

2 Figures

Transform the input values (numbers) of `sport`, `gender`, and `location` in the original data into the provided actual values (specific characters) for further and more intuitive analysis and comparison.

2.1 Baseline Data

Using the `season0` dataset as the baseline data, it contains the following missing values in different variables: `training_hours`, `age`, and `gender` each have 3 missing values. `Injuries` has 4 missing values. `Location` and `sport` each have 1 missing value.

For season 0, players in boxing have the highest average player value and average training hours, while basketball players have the lowest average injuries. The average player values and training hours do not differ significantly between genders, but there are slight variations in average age and injuries.

sport <fctr>	avg_player_value <dbl>	avg_training_hours <dbl>	avg_age <dbl>	avg_injuries <dbl>
Baseball	98869320	45.33333	31.40000	11.166667
Basketball	44562945	46.50000	31.54545	3.000000
Boxing	133852045	48.00000	28.00000	10.000000
Football	56625840	46.80000	29.33333	6.555556
Soccer	67660736	45.52632	30.75000	9.166667
NA	64793556	46.00000	27.00000	8.000000

gender <fctr>	avg_player_value <dbl>	avg_training_hours <dbl>	avg_age <dbl>	avg_injuries <dbl>
Female	62328377	46.88889	30.22222	7.900000
Male	66059971	45.65714	30.80000	7.323529
NA	67228356	48.00000	29.33333	7.000000

Figure 1: Summary by Sport and Summary by Gender

2.2 Relationship Between Player Value and Other Variables at Baseline

The analysis of the relationships between player value and the variables of age, training hours, and injuries at baseline(season 0) reveals varied insights. All the plots are attached at 5.1.

- The scatter plot `Player Value vs.Training Hours`, the hypothesis that there is a slight positive relationship between player value and training hours at baseline. Players with more training hours tend to have slightly higher player values, but the relationship is weak, as indicated by the scattered distribution of points around the regression line. This suggests that while training hours might have some influence on player value, other factors are likely at play.
- The scatter plot `Player Value vs.Age` supports that there is a relationship between player value and age at baseline. Specifically, younger players generally have higher player values, and there is a slight negative correlation between age and player value. This relationship is important for understanding how player value might change as players age.
- The scatter plot `Player Value vs.Injuries` indicates that there is no significant relationship between player value and the number of injuries at baseline. The lack of a discernible trend and the horizontal regression line suggest that injuries do not have a major impact on player value at this stage. However, there is still a slight indication that players with fewer injuries might have higher values, though this relationship is not very strong.

2.3 Compare Seasons Against Player_Value

Using set.seed(455) ensures that the sample of individuals is reproducible then randomly selects 10 individual IDs from 1 to 50. The line plot illustrates the changes in player value across 10 seasons for 10 randomly selected individuals.

Most players exhibit a general decline in player value over the seasons, though some show occasional fluctuations. Notably, player ID 21 started with the highest value around 120 million and, despite some decreases, maintained a relatively high value above 90 million by season 10.

The plot highlights a visible gap between the highest and lowest player values across the seasons, indicating diverse performance trajectories among the players.

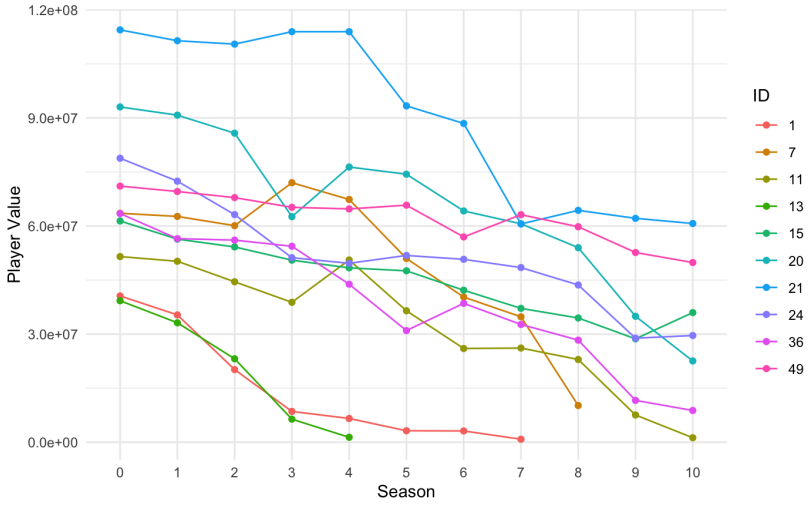


Figure 2: Player Value Across Seasons for 10 Random Individuals

2.4 Average Player_Value Across Seasons by Location, Sport, and Gender

The plots for average of each location, sport, and gender against player value during 10 seasons highlight the common trend of declining player value over time, with variations based on location, sport, and gender. The plots are attached at 5.2.

The geographical and sports-related differences suggest that external factors specific to these categories significantly influence player value trajectories. For instance, Saudi Arabia starts with the highest average player value, maintaining a leading position until season 9, after which it sees a sharp decline. Boxing consistently starts and remains the highest in average player value. The overall trends for both genders are quite similar, indicating uniform factors affecting player value irrespective of gender, but females begin slightly lower.

3 Next Steps

1. Reports on data quality, including any known issues with missing values, data entry errors, or inconsistencies in the data. There is a lot of NA value which may affect the analysis the data of each season.

2. Access to historical data beyond the 11 seasons already analyzed. This could help in understanding long-term trends and patterns.

3. Information on any additional variables that might be available but not included in the current datasets. This might include financial data, player contracts, team performance, or health data beyond injuries, such as physical fitness scores, etc.

4 Data

Categorical Variables:

- The variables id, location, gender, sport, and season should be treated as categorical. Id is a unique identifier for each player. Location represents different geographical areas. Gender indicates the sex of the players (e.g., 'Male', 'Female'). Sport categorizes the type of sport each player is involved in. Season denotes different time periods within the dataset and functions better as a categorical variable to represent distinct cycles.
  - The whole dataset includes 161 entries for Spain, 79 entries for Germany, 66 entries for France, 63 entries for the USA, 52 entries for Saudi Arabia, and 86 entries for other locations. There are also 30 entries with missing location data.
  - The dataset contains 114 entries for female players, 397 entries for male players, and 26 entries where the gender is not specified.
  - The dataset includes 66 entries for baseball, 108 entries for basketball, 11 entries for boxing, 105 entries for football, and 222 entries for soccer. Additionally, there are 25 entries with unspecified sports.
    - There is no data on the sport of `Tennis`.
  - Each season from 0 to 10 has approximately 50 entries.

Numeric Variables:

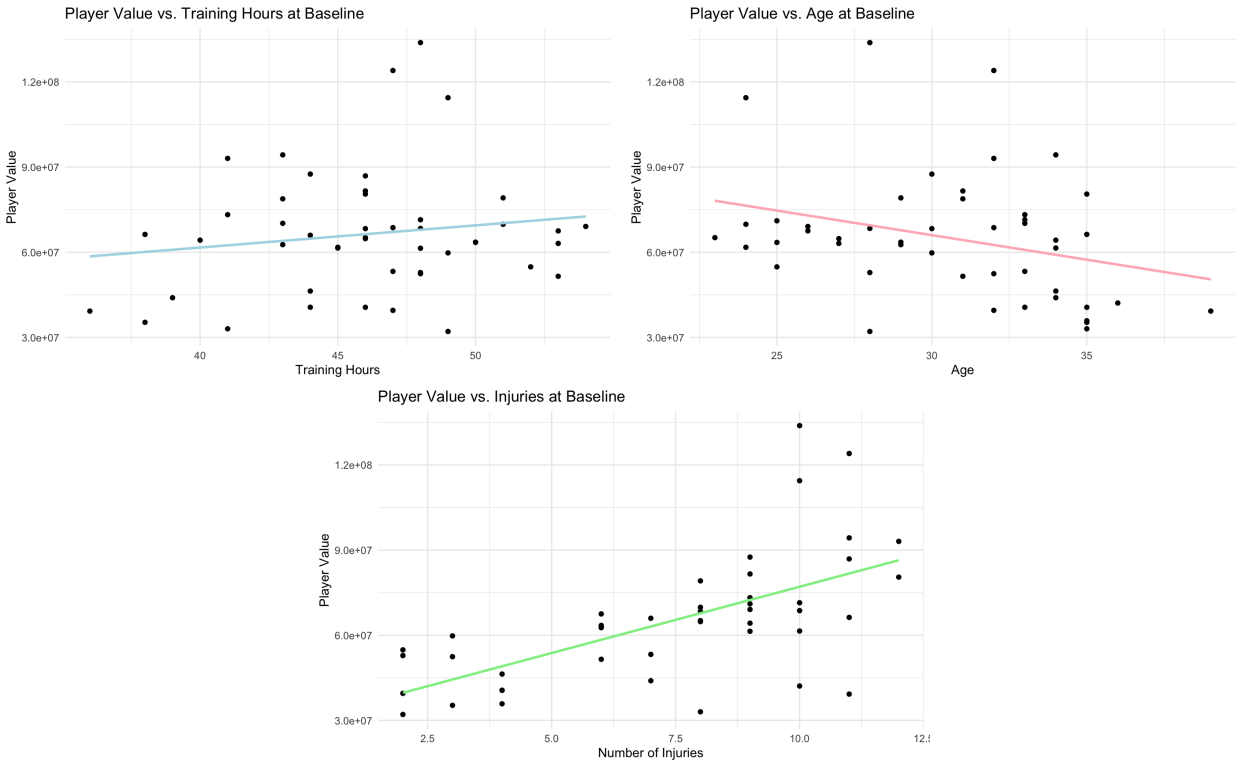
- The variables player_value, training_hours, age, and injuries should be treated as numeric. player_value represents the monetary value of a player. training_hours indicates the amount of time spent training, measured in hours. Age represents the player's age in years. Injuries count the number of injuries a player has sustained, quantifying the frequency of occurrences.
  - player_value range from a minimum of 524,820 to a maximum of 133,852,045.

- training_hours range from 20.00 to 94.00 hours.
- Ages range from 23 to 46 years.
- The number of injuries ranges from 2 to 30.

5 Appendix

5.1 Plots For 2.2

`Relationship Between Player_Value and Other Variables at Baseline`



5.2 Plots For 2.4

`Average Player_Value Across Seasons by Location, Sport, and Gender`