EA Sports FIFA 2021 complete player ratings datasets

Data Source

Summary

Data Source: This is a mix of external and internal data source. It is provided by EA sports, a developer and publisher of sport video games. The data provided are reliable because they were reviewed by their editors.

Data Collection: This data is collected through a network of over 9000 reviewers known as they are also called by the EA "data reviewers". Members are made up of coaches, professional scouts, and a lot of season ticket holders. For the data to be more comprehensive and be able to meet the business goal, the reviewers watch players, review their abilities and assign them various ratings (i.e. survey data, it is collected manually by observing different players). In addition, the reviewed data is then handled by 300 editors, which arrange it into 300 fields and 35 attribute categories. EA uses this subjective feedback in conjunction with its own statistics (scoured from other agencies) to determine the ratings.

Data Contents: This data contains information about EA Sports FIFA player's ratings for 2021 with variables such as their biography, age, geographical location, clubs represented, wages, market value, and their ratings of different fields.

Resources: This link contains articles where EA Sports explains how FIFA player ratings are calculated: https://www.vg247.com/how-ea-calculates-fifa-17-player-ratings , The second link reports the dataset: https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset,

To compile all the data into an excel file, it was scraped from the publicly available website https://sofifa.com.

Why did I choose this dataset:

I go for this dataset because Soccer is a sport I am passionate about and it is an area where data analytics is used immensely focusing on different areas such as coaches making use of big data set to create winning strategies, predicting the opponent strength and so on. There is also a case study of Manchester city soccer player who instead of the use of a traditional means to negotiate his contract, he rather hired an agent service of a data analytics firm, for the negotiation of his contract and it worked out great, he got a better deal out of the

negotiation. In addition to that, I will like to familiarize myself with analytics in the sport space as well because data can explain a lot about the soccer mechanisms.

Moreover, I would also love to use another data in the food industry since my focus is also more on that but due the time frame, I won't be able to meet up. In the future, I will still delve into the data in the food industry sector.

Data Profile

Data cleaning and consistency checks:

The original data set contains 18944 rows and 106 columns, so in order to narrow down my focus, I only imported the columns that will be needed for the analysis and this will also reduce the total number of columns and rows.

Dropped Columns: I dropped all the columns except 21 columns, for the following reasons:

PII/ Sensitive data: first name, last name, DOB

- Columns with too many missing data:
  - Joined
  - player_tags
  - loaned_from
- Columns with too many missing data rows were deleted.

Renaming columns

- There are lot of columns I would like to have renamed for them to give more meaning but due to the vast volume, it is not worth it to rename over 85 columns
- I renamed the column 'overall' to *player_rating*

Mix data types:

- There is no column of mix data types

Missing values: There are missing values in 5 columns:

league_name, *team_position*, *league_rank* : there are 225 values missing for these 5 columns. This gives me the insight that for having the missing values definitely they have to be on the same rows, so I deleted them.

Columns with large missing values:

Joined, *player_tags* and *loaned_from* have values that are more than 5 % of the column data set, whereas others are more that 30%. Therefore I decided to drop the entire column because if I replace this with the mean value of the entire column, it might skew my result since some variables are imputed in a discrete format.

Duplicate:

There is no duplicate

Basic descriptive statistics:

Rows:18719,

Columns: 21

Records:374,380

Some Continuous variables

Column	Min	Max	Mean	Frequency
Age	16	53	25	
height_cm	155	206	181	
weight_kg	50	110	75	
overall value_eur	0	105,500,008	225,155,5.06	
wage_eur	500	560,000	8,780	

Some Categorical Variables

Columns counts for all: 18719

Columns	Mode
Nationality	England
player_rating	65
preferred_foot	Right
team_jersey_number	7
contract_valid_until	2021

Limitations and ethical considerations

The data was collected manually, which makes it prone to human errors and it is collected from multiple sources which can make some data to be incorrect. Based on the method of collection, some reviewers can give some players higher reviews or ratings because the player is a member of the club they supported or from the same country as well as personal interest bias. The final review is done by team of editors, though there is a potential

measurement bias if one of them does not have adequate knowledge of training which can also affect the result generated.

The ethical concerns of the dataset is that, it contains some PII and sensitive information, this can cause data breach.

Define questions

Preferred foot

Which foot is the most preferred among players?

Does preferred foot determines the player's ratings?

Player ratings

Which nationality has the highest average player ratings?

Do players with higher ratings receive more wages?

Wages/Overall value Eur

Which club has the highest wage bill?

Do players with higher market value receive more wages or vice versa?

Age

Which club name has the youngest players?
Does the age of players affect the player ratings