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 is Reliable because they were reviewed by their editors.

Data Collection: This data is collected through a network of over 9000 reviewers known as EA data reviewers they're made up of coaches, professional scouts, and a lot of season ticket holders. They watch players, review their abilities and assigned them various ratings (i.e. survey data, it is collected manually). These 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 stats (scoured from other agencies) to determine the ratings.

Data Contents: This data contains information about EA Sports FIFA players ratings for 2021 with variables such as their bio information, clubs represented, wages, market value, and their ratings of different fields.

Resources: This link contain article where EA Sports explains how FIFA player ratings are calculated: https://www.vg247.com/how-ea-calculates-fifa-17-player-ratings , The second link is that of 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 i Chose this dataset:

I chose this dataset because Soccer is a sport i am passionate about and it is an area where data analytics is used immensely, example is the case of a soccer player of Manchester city FC who instead of the use of a traditional soccer agent hired the 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, i will like to familiarize myself with analytics in the sport space, because it contribution in this field is understated and it is growing.

Data Profile

Data cleaning and consistency checks

Dropped Columns: I dropped all the columns except 20 columns, for the following reasons;

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

Joined: too many missing data

Not needed : all other columns

Final counts of columns 20

Renaming columns: There are lot of column names i will like to change due to the vast volume of it,it is not worth it.

So i changed 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, namely;

league_name, team_position, league_rank : there are 225 values missing, for these 5 columns ,to have the same number of missing values this tells me they on the same rows, so i chose to delete them.

joined: there are 983 missing value, thats is more than 5 percent of the column, so i chose to drop the entire column because if i replace this with the mean value of the entire column, it might skew my result

Duplicate:

There is no duplicate

Basic descriptive statistics:

Rows:18719,

Columns: 20

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

Column 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, base on the method of collection, some reviewers can give some players higher reviews or ratings because he is a member of a club he/she support or if the players is from his/her country, even though the final review is done by team of editors, there is a potential measurement bias if one of them does not have adequate knowledge of training, it can skew the final ratings result.

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 players ratings?

Player ratings

Which nationality have the highest average player ratings?

Does player with higher ratings receive more wages ?

Wages/Overall value Eur

Which club have the highest wage bill?

Does player with higher market value receive more wages or vice versa

Age

Which club name have the youngest players

Does the age of players affect the player ratings