Soccer's Next Generation: Leveraging Data to Identify and Develop Talents

Author: Huu Dang Nguyen

Date: 9/13/2023

Class: ITSS 4v95 - Data Science Fundamentals

Table of Contents:

1. Overview
2. Data Overview and Pre-processing
3. Exploratory Data Analysis (EDA)

Overview

In recent years, soccer has become a favorite sport for many Americans. The MLS, Major League Soccer, has been around for 30 years. However, it has never experienced much exciting news until recent years; the MLS signed a 10-year deal with Apple TV; in 2022, for the first time in the MLS history, an MLS team, Seattle Sounders FC, won the CONCACAF Champions League; Most exciting news in the last summer, Lionel Messi, the 2022 World Cup champion and seven-time Ballon d'Or winner, the greatest soccer player in the history, joined Inter Miami FC. More importantly, the United States is the primary host of the World Cup 2026, the biggest World Cup ever with the appearance of 48 teams, played across 12 cities in America.

With the rapid growth of the MLS and soccer in the U.S., we also face many challenges, such as salary caps, stadium infrastructures, and talents recruitment. Attracting and recruiting young talents for long-term growth is always the biggest problem for MLS teams. To shape future success, talent recruiters need to focus on the right performance metrics and market analysis to provide a comprehensive assessment of potential recruits.

This project aims to provide valuable insights into identifying and pursuing young talents who can contribute to the long-term success of a team. The recommendations presented in this analysis will help the team make informed decisions in the player recruitment process.

Data Overview and Pre-processing

Data Overview

● The data collected by Transfermarkt,https://www.transfermarkt.us, uploaded on https://www.kaggle.com/datasets/davidcariboo/player-scores, the latest update used for this

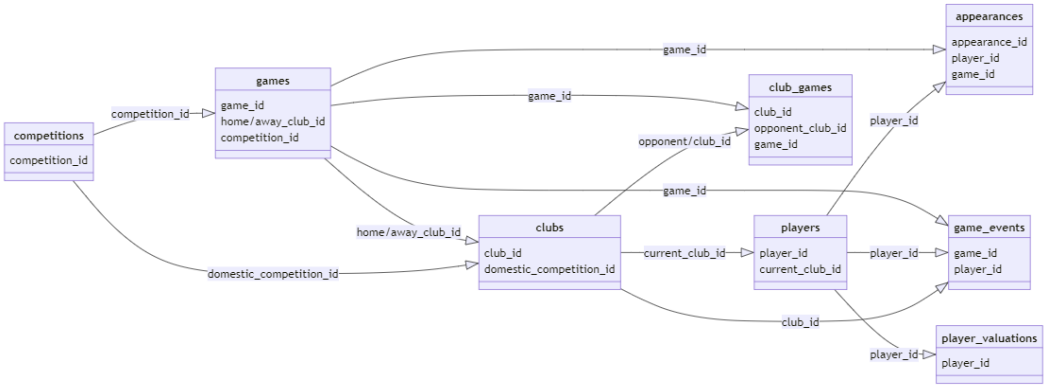
project is on 9/11/2023.

● The data contains information about more than 60,000+ games from many seasons on all major leagues, 400+ clubs, 30,000+ players, 400,000+ player market valuations historical records, and 1,200,000+ player appearance records from all games.

● The dataset has 8 tables: appearances, club_games, clubs, competitions, game_events, games, player_valuations, and players.

● 8 tables have total of 111 columns.

Schema:

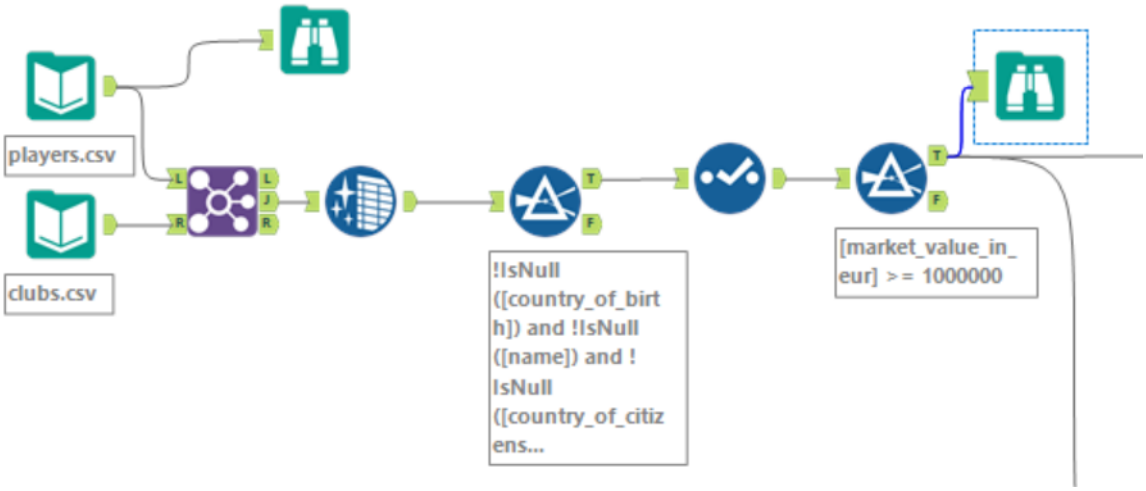


Data Pre-processing

I perform data pre-processing using Alteryx. First, I joined "players" table with "clubs" table on "club_id". Next, I performed data cleasing and dropped all records that have Null values in important fields as "name", "market_value_in_eur", "contract_expiration_date", "country_of_birth", and "country_of_citizenship". Then, I selected 11 fields that useful for exploratory data analysis. The fields are "name", "country_of_birth", "country_of_citizenship", "date_of_birth", "position", "height_in_cm", "market_value_in_eur", "highest_market_value_in_eur", "contract_expiration_date", "current_club_name", and "average_age". The last step is removing all players that have "market_value_in_eur" under 1,000,000.

From 30,266 data records and 28 columns, I now have 4,939 data records and 11 columns after data pre-processing.

Here is the my Alteryx workflow and the first 6 records:



Record	name	country_of_birth	country_of_citizenship	date_of_birth	position	height_in_cm	market_value_in_eur	highest_market_value_in_eur	contract_expiratio
1	Stefan Ortega	Germany	Germany	1992-11-06	Goalkeeper	185	9000000	9000000	2025-06-30
2	Claudio Gomes	France	France	2000-07-23	Midfield	180	1400000	1400000	2027-06-30
3	AdriÃ¡n BernabÃ©	Spain	Spain	2001-05-26	Midfield	170	6000000	6000000	2025-06-30
4	JuliÃ¡n ÃlIvarez	Argentina	Argentina	2000-01-31	Attack	170	60000000	60000000	2028-06-30
5	Matheus Nunes	Brazil	Portugal	1998-08-27	Midfield	183	45000000	45000000	2028-06-30
6	Kevin De Bruyne	Belgium	Belgium	1991-06-28	Midfield	181	70000000	150000000	2025-06-30

Exploratory Data Analysis

Through Exploratory Data Analysis (EDA), I aim to uncover hidden gems, and shed light on the crucial role of age in talent development. I will visualize player trajectories, and offer actionable insights to shape the future of soccer. Let's dig in exploring the synergy of data and sports.

Data Transformation:

I am interested in looking at range of age and its summary statistics, I started by calculate players' age from their dates of birth. But before explore how create the fomula, I had to figure out what type of data I want to look at. I selected "date_of_birth" and "market_value_in_eur", the reason that I chose those two because to find a potential young player, age and their market value are the most important besides of their performance on the pitch.

To calculate players' age in Alteryx, I used the Formula tool with this following fomula: Floor(DateTimeDiff(DateTimeToday(), [date_of_birth],"Months")/12). The entire expression calculates the age of players in years by finding the difference between the current date and their date of birth in months, then I converted it into years and rounding it down to the nearest whole number.

Here are the outputs of the first 5 records:

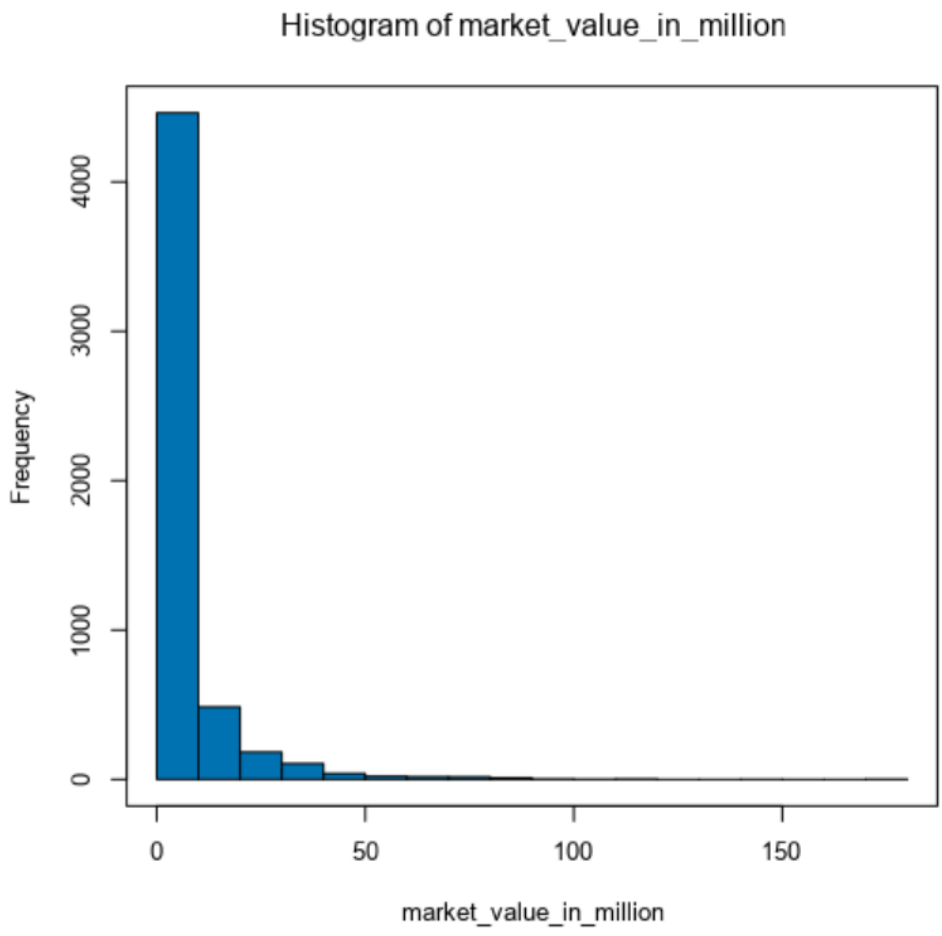
Record	date_of_birth	market_value_in_eur	Age
1	1994-02-07 00:00:00	2000000	29
2	1997-08-11 00:00:00	1000000	26
3	1996-01-19 00:00:00	1800000	27
4	1997-10-04 00:00:00	1000000	25
5	2002-01-22 00:00:00	1500000	21

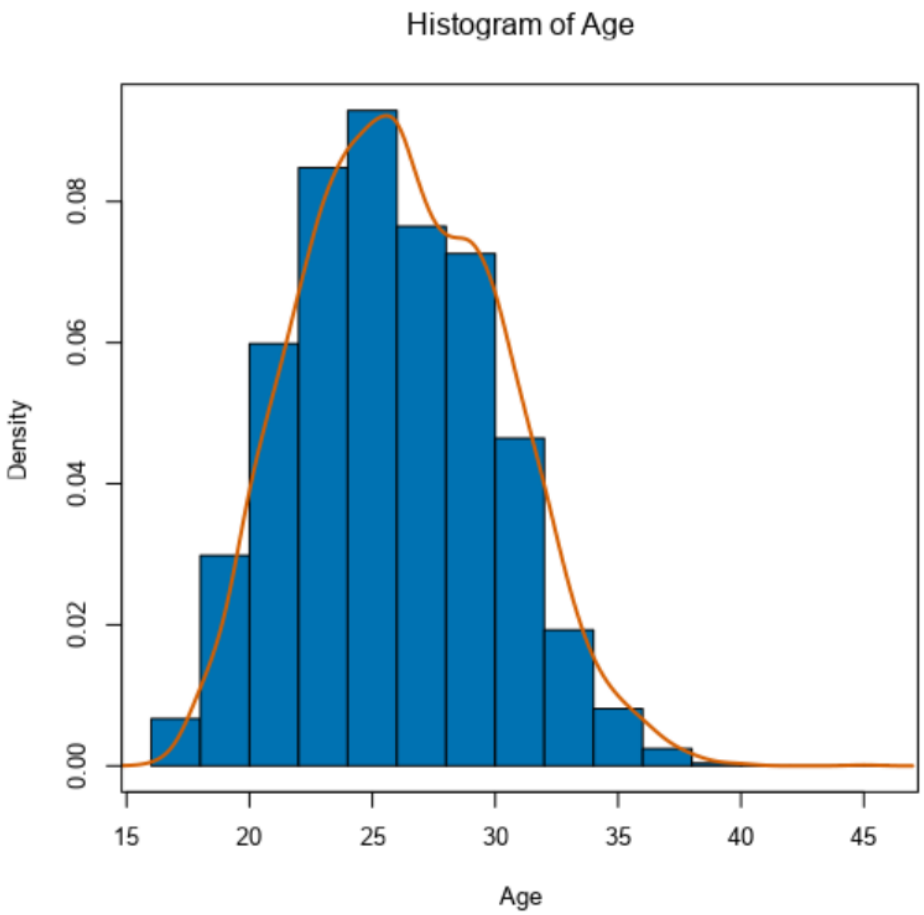
Descriptive Statistics:

Summary statistics: having the needed data ready, I performed summary statistics for "market_value_in_eur" and "Age" by using Field Summary tool. Here is my result:

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	market_value_in_eur	Numeric	1000000	180000000	3000000	12386848.967528	0	87	7342726.719416
2	Age	Numeric	16	45	26	4.016756	0	26	26.210793

Frequency distributions:





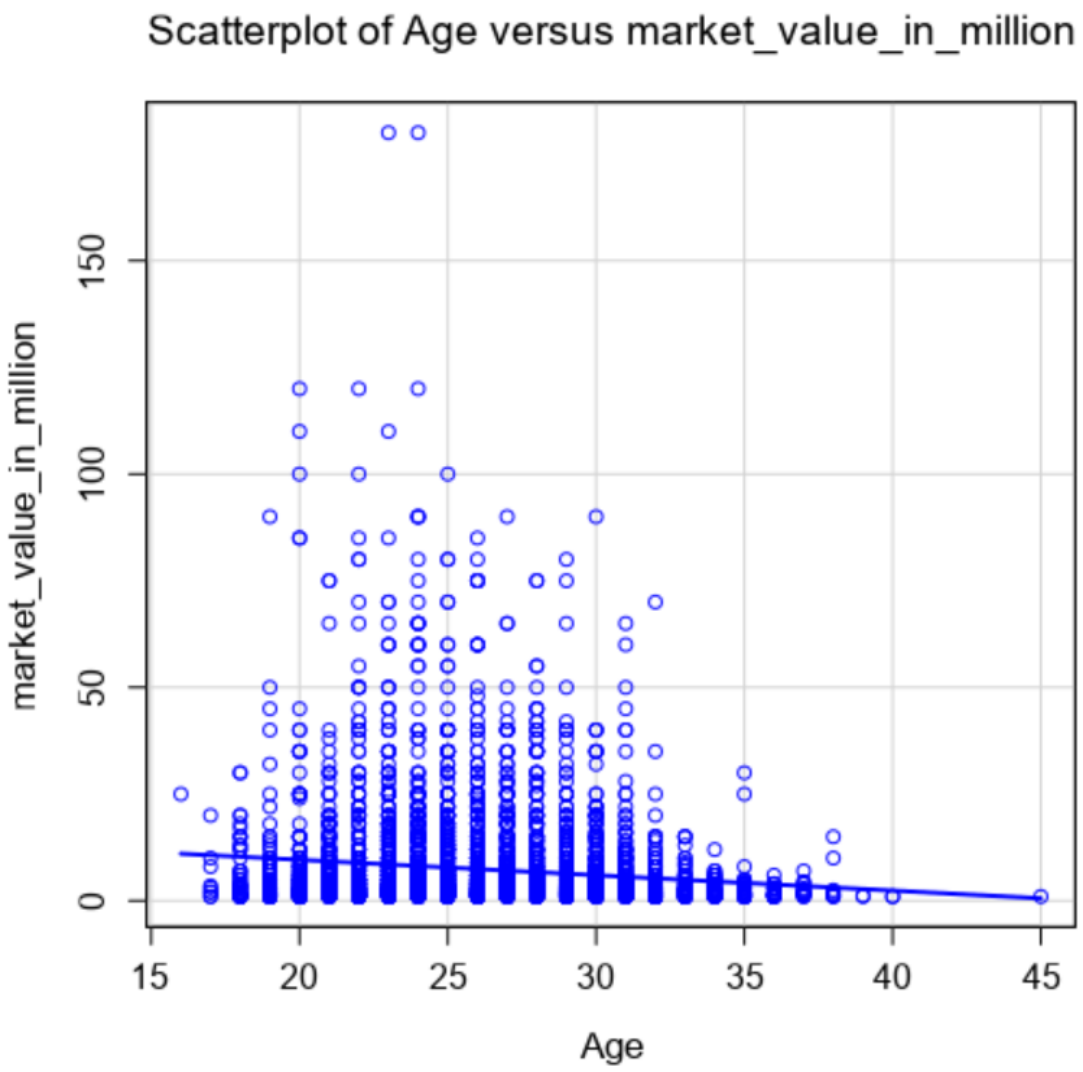
From the summary statistics and frequency distributions through histogram charts, here are the insights that I gain:

● Market value: The highest market value is 180M, the lowest as we filtered before is 1M, the average is at 3M, and the mean is about 73.5M. As the histogram chart, we could see the highest frequency is from ~1M to ~10M.

● Age: The highest age value is 45, the lowest is 16, the average is 26, the mean is 26.2, and the standard deviation is 4. As the histogram chart, we could see the central tendency is from 20 to 32.

Outlier Detection

I decide to use scatter plot to detect outlier because I have 2 variables which are age & market_value_in_eur, and I am interted in figuring out whether there is any correlation between those 2 variables. First, I divide market_value_in_eur/1,000,000 to market_value_in_million for visual purpose. Then I use Scatterplot tool to generate the visualization with x-axis is age and y-axis is market_value_in_million.



As we see from the scatter plot, we have several outliers:

● 2 outliers at the top represents 2 players valued at 180M

● 3 outliers at the right bottom corner represents a 38-year-old player, a 39-year-old player, and a 45-year-old player. And they are all valued at 1M.

● The correlation between age and market value is negative, which means the older a soccer player gets, the less valuable he is.

Additional Observations

Curiosity gets me again! Now, I even want to look deepper into the data, specifically in the frequency distribution of position and players' citizenship. Let's go!

● Country of citizenship:

I use Summarize tool to group_by and count the country_of_citizenship. Here is the first 10 results after I sort descending:

Record	Count	country_of_citizenship
1	449	Spain
2	373	France
3	334	Brazil
4	291	England
5	274	Germany
6	253	Italy
7	195	Portugal
8	183	Netherlands
9	144	Argentina
10	121	Russia

Where is our USA? Our population is 330M and we don't even make it to top ten? Here we are!!

21	59	United States

Top 21 with 59 talents, how many of them are young-potential players? We will find out later.

Now I want to see more information on the frequency distribution, and Frequency Table tool will help me with it. Here is first 10 records:

Record	Field_Name	Field_Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	country_of_citizenship	Spain	449	9.11	449	9.11
2	country_of_citizenship	France	373	7.57	822	16.68
3	country_of_citizenship	Brazil	334	6.78	1156	23.45
4	country_of_citizenship	England	291	5.90	1447	29.36
5	country_of_citizenship	Germany	274	5.56	1721	34.92
6	country_of_citizenship	Italy	253	5.13	1974	40.05
7	country_of_citizenship	Portugal	195	3.96	2169	44.00
8	country_of_citizenship	Netherlands	183	3.71	2352	47.72
9	country_of_citizenship	Argentina	144	2.92	2496	50.64
10	country_of_citizenship	Russia	121	2.45	2617	53.09

- Position:

I perform the same tasks as I do with country_of_citizenship. Here are the results:

Record	position	Count
1	Attack	1513
2	Defender	1606
3	Goalkeeper	315
4	Midfield	1495

Record	Field_Name	Field_Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	position	Defender	1606	32.58	1606	32.58
2	position	Attack	1513	30.70	3119	63.28
3	position	Midfield	1495	30.33	4614	93.61
4	position	Goalkeeper	315	6.39	4929	100.00

My exploratory data analysis has illuminated a path to uncovering soccer's future stars, age profiling, and market dynamics. What come next? I will combine performance metrics for further analysis of our journey in finding rising soccer stars. Before we continue, let's take a step back and look at the work flow of the EDA in Alteryx.