Project Report on FIFA 2022 Player Analysis

Introduction:

The FIFA world cup 2022 season in Qatar started in November and with that in mind, we chose the FIFA player dataset from 2022 for our project. One of the most esteemed sporting associations in the world is the Federation Internationale de Football Association (FIFA). It has 209 members and was founded in 1904 to encourage cooperation among national soccer associations. FIFA was established by seven national associations: Belgium, Denmark, France, Netherlands, Spain, Sweden, and Switzerland. They did this by supporting the scheduling of football games at all levels.

The objectives of FIFA were to "advance association football," to encourage cordial relationships between National Associations, Confederations, and their officials and players, and to exert control over all associations. We study the relationships between a number of player traits, including age, nationality, valuation, positions, endurance, agility, reaction time, clubs, and skills, and the player's overall performance on the field in FIFA matches. Additionally, we'll model and research how these characteristics affect overall performance.

Dataset Summary

The FIFA dataset 'players_22.csv' was taken from this link *(https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv)* which includes 2022 players data in FIFA. The dataset contains 19239 rows and 110 column attributes. The rows represent a record of a player and the column represents the player's various attributes. Some of the categorical variables of the player are nationality, name, preferred foot, etc. Some of the numeric variables are passing, dribbling, shooting, attacking and so on.

Hypothesis

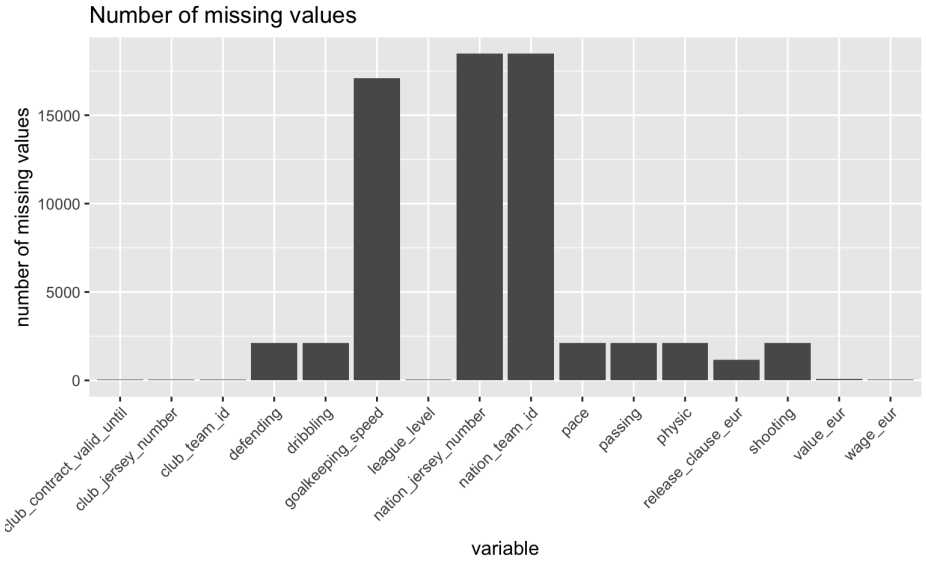
Player Analysis is a very important step to determining which player is taken off the team and which one can be switched out. After all, one of the key elements of FIFA is player ratings. This kind of analysis aids managers and coaches in selecting the best players for any given game. In this research, we use the player analysis to model each attribute and analyze how they affect overall performance.

The first question is whether a player's salary is influenced by on-field skills like shooting, passing, pace, and dribbling. We aim to see what skills contributes towards the salary of the player. Similarly, the validity of FIFA's use of overall ratings based on player traits is the subject of our second hypothesis. We aim to check if the fifa overall performance ratings are biased or they correspond to the various skills of player. To see if cluster analysis produces places for a representative playing team, we test the third hypothesis.

Data Cleaning

Out of 110 columns, we checked the amount of missing values in each of them. The bar chart below depicts the amount of missing values in the dataset. For those attributes that were not necessary for our project, we chose to drop them.

For the necessary variables like shooting, wage, value and others, we checked the skewness of the column values. Upon finding that it was heavily skewed, we combated that by replacing those missing values with median. Afterwards we handled the duplicates in the dataset. 20 duplicates were found and dropped, leaving us 19219 rows.

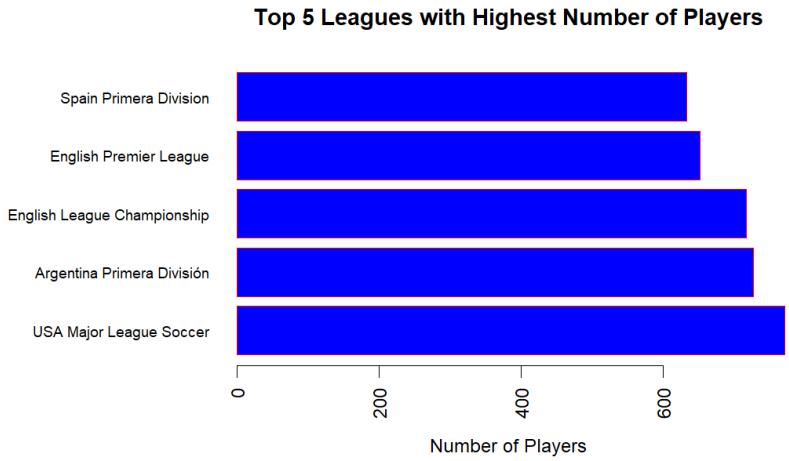


Data Analysis and Visualization

We wanted to see where the players(in the dataset) were coming from, from the age of the players to the countries and leagues of those players to get a sense of where the domination lies. We wanted to see the distribution of players in these categories.

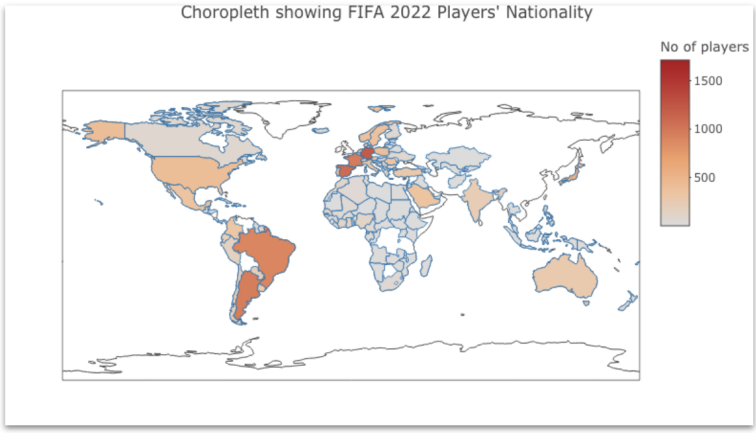
Leagues with highest players:

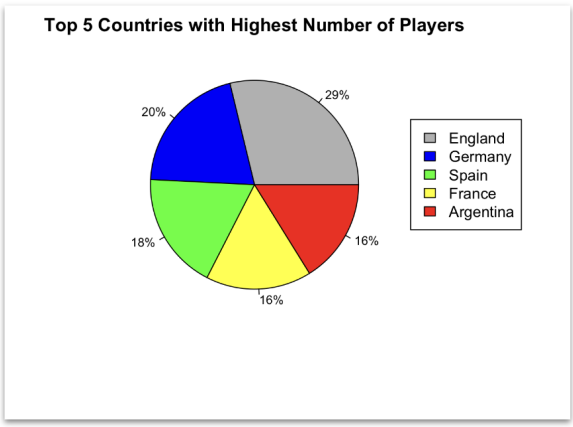
Surprisingly, the highest number of players in the league are mostly not from the elite leagues. The top 5 leagues with the highest number of players (from the most to the least) are USA Major League Soccer, Argentina Primera Division, English League Championship, English Premier League and Spain Primera Division. Assuming there's a limit to the number of players an elite league allows its team to register(Premier League allows 25 players per team), other leagues who don't have a ceiling in the number of players have got the maximum number of players.



Countries with highest players:

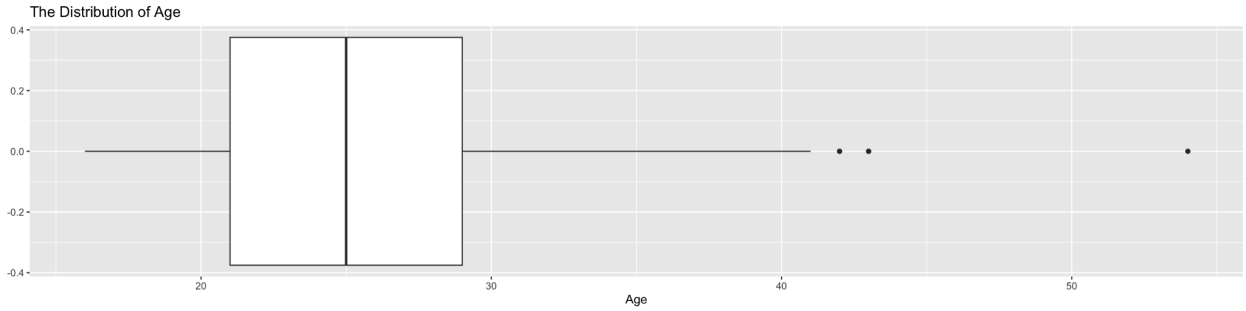
There are 209 countries listed in the FIFA 22 dataset. Out of those countries, the top countries with the highest number of players(from the most to the least) are England, Germany, Spain, France and Argentina. We can see Europe's dominance in soccer with the help of this stat: Only Argentina is from South America while the rest are from Europe.





Age groups of players:

Then, to represent the various age groups of FIFA players, we have a boxplot. Players begin to sign up for FIFA around the age of 16, and our boxplot shows the same. The data shows that there is a substantial group of players between the ages of 20 and 30. The histogram further shows some outliers above the age of 41.



The first table below shows some outstanding players who are still playing despite their old age. The second table shows that the top 5 youngest players are all 16 years old matching ot the recruitment age requirement of FIFA.

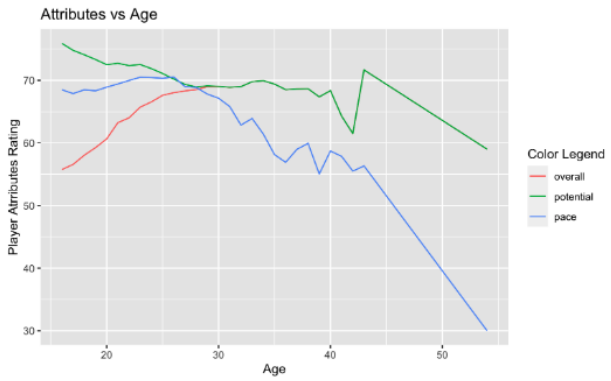
	short_name <chr>	nationality_name <chr>	age <int>
1	K. Miura	Japan	54
2	G. Buffon	Italy	43
3	C. Lucchetti	Argentina	43
4	S. Nakamura	Japan	43
5	D. Vaca	Bolivia	42

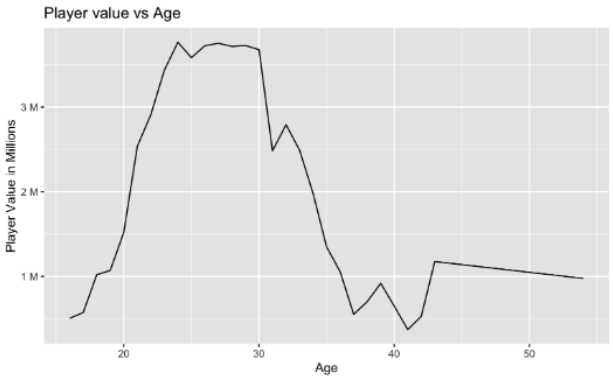
	short_name <chr>	nationality_name <chr>	age <int>
1	Gavi	Spain	16
2	V. Barco	Argentina	16
3	A. Kalogeropoulos	Greece	16
4	Yayo	Spain	16
5	R. van den Berg	Netherlands	16

Attributes over age:

We visualized the average pace, potential and overall performance of players with respect to age. For this we chose a line graph to compare those quantitative variables to visualize the changes over time. The results that we typically anticipate, such as increased stamina and acceleration with age—which peaks between the ages of 20 and 30—or adulthood—and then declines with growing age clearly visible. We saw that a player's performance reaches his peak potential at the age of 30.

Further data analysis led us to the conclusion that this result was not solely caused by stamina and acceleration, as we had previously believed, but also by a number of other talents, such as pass and shot, which develop over time and 30 is on an average the skill that players reach their peak performance on those skills. Furthermore, the value of a player is greatly influenced by age factor. Player's value is also heavily dependent on how long a player can play for a team, how well he can perform and so on. With Messi of age 35 and Mbappe of age 23, although both play incredibly well for **Paris Saint-Germain Football Club,** Mbappe is regarded as more valuable than Messi. The line chart shows that the age where players are most valuable is around 22-32.





Data Modelling

Our first model was made to show the correlation between a player's individual skills and their salary, as well as to show how significant each skill is in affecting a player's salary. The model we used was able to predict the average salary of each player given their skills, however, the exact salaries of each player is also affected by many unknown variables outside of the dataset we used which introduces variability in the actual salaries which our model can not account for.

We initially divided the dataset into two parts with 70% train and 30% test dataset. We developed the models on training and applied the model on test data to get the predictions.
Despite this variability, we are able to find traits which will often get players higher pay. One example of this is the player's reaction time, which significantly increases a player's pay as it gets better. A trait that negatively affects a player's pay would be age as the higher a player's age the less they get paid.

For the multiple linear model for overall performance, the significant skills all have fairly similar effects on each player's performance. The model was much less variable than the salary model.The table below shows what skills had strong relationships with salary and overall. We would be able to reject the null hypothesis for attributes with p-value < 0.05.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.437e+00	2.140e-03	671.663	< 2e-16	***
pace	8.007e-05	5.697e-05	1.405	0.159924	
shooting	-2.228e-03	1.882e-04	-11.840	< 2e-16	***
passing	1.869e-04	1.093e-04	1.711	0.087173	.
dribbling	1.983e-03	1.562e-04	12.692	< 2e-16	***
defending	2.351e-03	1.562e-04	15.053	< 2e-16	***
physic	-2.428e-04	7.689e-05	-3.159	0.001589	**
age	7.849e-04	3.976e-05	19.742	< 2e-16	***
preferred_footRight	-2.582e-03	3.533e-04	-7.307	2.88e-13	***
attacking_crossing	2.026e-04	2.981e-05	6.795	1.13e-11	***
attacking_finishing	1.120e-03	8.725e-05	12.840	< 2e-16	***
attacking_heading_accuracy	1.487e-04	2.475e-05	6.009	1.91e-09	***
attacking_short_passing	5.811e-04	4.706e-05	12.348	< 2e-16	***
attacking_volleys	1.369e-05	2.424e-05	0.565	0.572090	
skill_dribbling	-6.984e-04	7.608e-05	-9.179	< 2e-16	***
skill_curve	-9.598e-05	2.257e-05	-4.252	2.13e-05	***
skill_fk_accuracy	-2.607e-05	2.008e-05	-1.299	0.194121	
skill_long_passing	-2.347e-04	3.135e-05	-7.486	7.52e-14	***
skill_ball_control	4.855e-04	5.558e-05	8.737	< 2e-16	***
movement_acceleration	2.942e-04	3.644e-05	8.073	7.45e-16	***
movement_sprint_speed	1.676e-04	3.890e-05	4.309	1.65e-05	***
movement_agility	-1.655e-04	2.574e-05	-6.430	1.32e-10	***
movement_reactions	2.153e-03	2.971e-05	72.450	< 2e-16	***
movement_balance	-2.470e-04	2.066e-05	-11.953	< 2e-16	***
power_shot_power	8.340e-04	4.436e-05	18.800	< 2e-16	***
power_jumping	5.397e-05	1.613e-05	3.345	0.000824	***
power_stamina	2.120e-04	2.617e-05	8.100	5.97e-16	***
power_strength	3.443e-04	3.682e-05	9.351	< 2e-16	***
power_long_shots	1.579e-04	4.470e-05	3.533	0.000412	***
mentality_aggression	-3.608e-05	2.295e-05	-1.572	0.115967	
mentality_interceptions	-6.039e-04	4.110e-05	-14.696	< 2e-16	***
mentality_positioning	-3.453e-04	2.706e-05	-12.761	< 2e-16	***
mentality_vision	-7.748e-05	2.741e-05	-2.826	0.004718	**
mentality_penalties	1.295e-04	2.286e-05	5.663	1.52e-08	***
mentality_composure	5.667e-04	2.495e-05	22.719	< 2e-16	***
defending_marking_awareness	-4.961e-04	5.006e-05	-9.911	< 2e-16	***
defending_standing_tackle	-5.847e-04	5.939e-05	-9.845	< 2e-16	***
defending_sliding_tackle	-2.556e-04	3.922e-05	-6.516	7.47e-11	***

Summary of Model for Overall

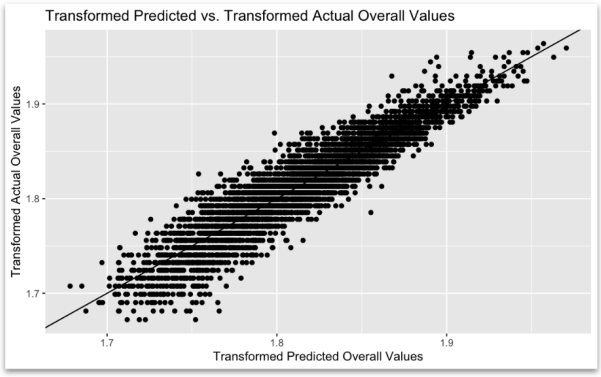
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.1012120	0.0495589	-2.042	0.041145	*
pace	-0.0010307	0.0013196	-0.781	0.434752	
shooting	-0.0173958	0.0043590	-3.991	6.62e-05	***
passing	-0.0018153	0.0025307	-0.717	0.473201	
dribbling	0.0176433	0.0036188	4.875	1.10e-06	***
defending	0.0266347	0.0036178	7.362	1.92e-13	***
physic	-0.0027722	0.0017809	-1.557	0.119593	
age	-0.0093548	0.0009210	-10.158	< 2e-16	***
preferred_footRight	-0.0123259	0.0081832	-1.506	0.132030	
attacking_crossing	0.0026669	0.0006906	3.862	0.000113	***
attacking_finishing	0.0077875	0.0020210	3.853	0.000117	***
attacking_heading_accuracy	0.0040710	0.0005732	7.102	1.29e-12	***
attacking_short_passing	0.0022091	0.0010900	2.027	0.042716	*
attacking_volleys	0.0023400	0.0005614	4.168	3.09e-05	***
skill_dribbling	-0.0028624	0.0017622	-1.624	0.104324	
skill_curve	0.0020949	0.0005228	4.007	6.19e-05	***
skill_fk_accuracy	-0.0013095	0.0004650	-2.816	0.004868	**
skill_long_passing	-0.0008253	0.0007262	-1.136	0.255807	
skill_ball_control	0.0015643	0.0012873	1.215	0.224301	
movement_acceleration	0.0016593	0.0008441	1.966	0.049339	*
movement_sprint_speed	0.0035286	0.0009010	3.916	9.04e-05	***
movement_agility	-0.0011727	0.0005963	-1.967	0.049233	*
movement_reactions	0.0233212	0.0006883	33.883	< 2e-16	***
movement_balance	-0.0016850	0.0004786	-3.521	0.000431	***
power_shot_power	0.0091222	0.0010275	8.878	< 2e-16	***
power_jumping	0.0010828	0.0003737	2.898	0.003765	**
power_stamina	-0.0011059	0.0006062	-1.824	0.068130	.
power_strength	0.0025534	0.0008529	2.994	0.002760	**
power_long_shots	-0.0000397	0.0010354	-0.038	0.969411	
mentality_aggression	0.0018055	0.0005316	3.396	0.000685	***
mentality_interceptions	-0.0073270	0.0009519	-7.697	1.49e-14	***
mentality_positioning	-0.0035675	0.0006268	-5.692	1.28e-08	***
mentality_vision	0.0004527	0.0006350	0.713	0.475894	
mentality_penalties	0.0012220	0.0005295	2.308	0.021023	*
mentality_composure	0.0086118	0.0005778	14.905	< 2e-16	***
defending_marking_awareness	-0.0055073	0.0011595	-4.750	2.06e-06	***
defending_standing_tackle	-0.0065354	0.0013756	-4.751	2.05e-06	***
defending_sliding_tackle	-0.0022125	0.0009084	-2.436	0.014879	*

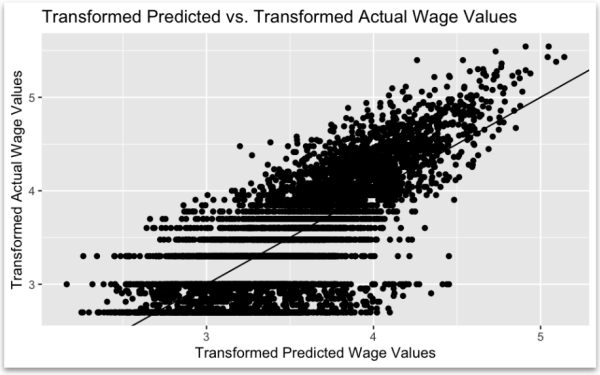
Summary of Model for Salary

One significant similarity between these two models is the effect of age on them. In both models age significantly decreases the overall performance and pay of a player. Other than age the two models share several significant skills such as the mental composure of a player increasing their pay as well as performance. Some skills, however, are valued highly in salary while not being reflected in performance and vice versa. One reason this may happen is due to scarcity or abundance of certain skills. When not as many players are capable of performing a certain skill, the value may increase for those who can, while having many good players for another skill may decrease the value of that skill. Neither of these scenarios would affect a player's performance but could affect their wage.

In order to check the accuracy of the model, we created a scatter plot with the actual values of the model vs the predicted values. The charts below show the x-y relations for the same. The Mean Square Error(MSE) for the models overall and wage were 0.0002738323 and 0.1543064 respectively.



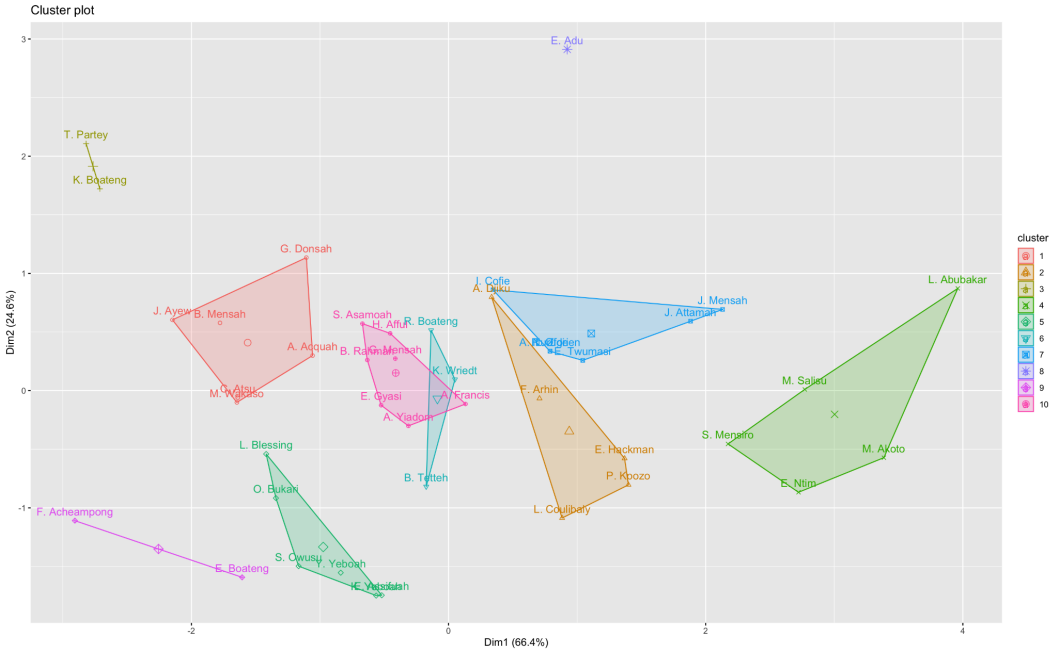
Overall scatterplot



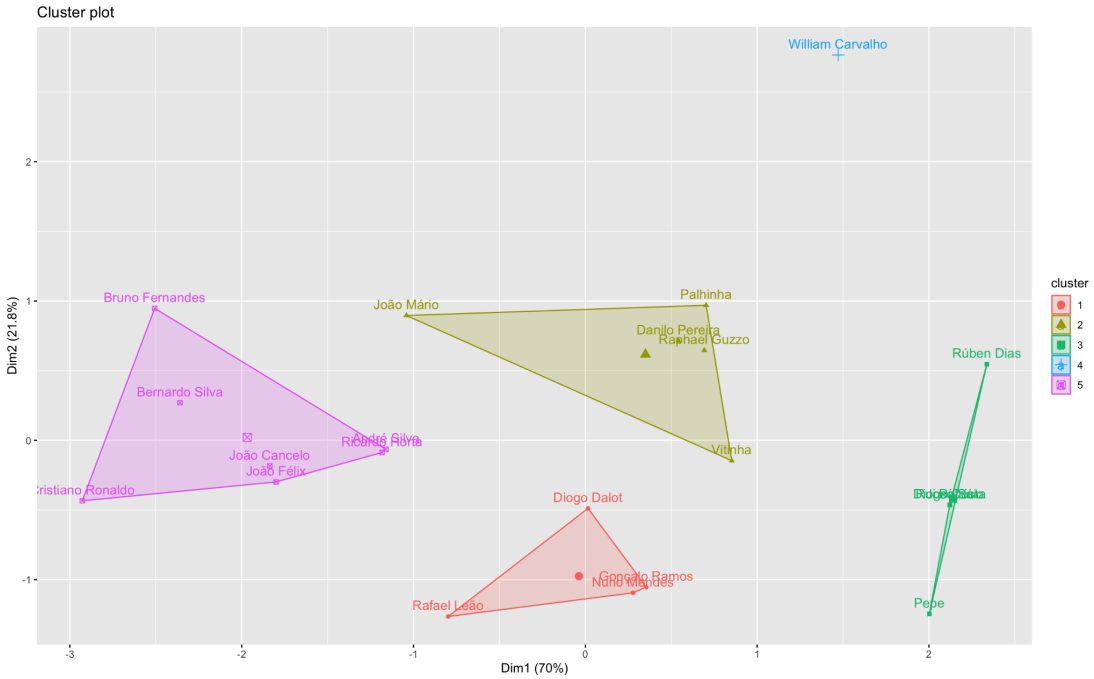
Wage scatterplot

Cluster Analysis in Players

In order to classify players based on their attributes, we filtered out the players with nationality Ghana and applied k means clustering. We chose k-means clustering because it can be used by coaches depending on the lineup he/she wants to use the team to play in the match. For all the players of Ghana, we obtained the following results with 10 numbers of clusters. We found out that the players within a cluster usually play in the same positions which matched our hypothesis that coaches use players with different sets of skills in different positions.



On the other hand, we also filtered out players that are playing in the FIFA 22 world cup in Qatar for team Portugal and clustered the players as per the same attributes. We saw that strikers like Christiano Ronaldo, Bruno Fernandez, Joao Felix and so on were in one cluster, whereas defenders like Pepe, Ruben Diaz and others were in the same cluster. This shows that our cluster analysis accurately represents how coaches use these players in various positions from the clusters based on their skills.



Conclusion

Football is very popular in Europe in comparison to other continents. The main reason might be because Europe's football culture and craze is unmatched. Although investments in football are increasing, as we can see the rise of USA Major League Soccer, the high-valued players still prefer playing in Europe at the peak of their career. After testing our first hypothesis using our models, we can conclude that salary is influenced by a player's attributes such as shooting, passing etc but also by factors such as the player's social media presence, negotiation styles, etc. Similarly, our overall rating model was pretty much accurate. The validity of FIFA's use of overall ratings based on player traits was just in our second hypothesis testing. The cluster analysis can classify a team into various clusters based on a player's attributes. That could be helpful to the coach/manager to pick their playing 11 and also find the best replacement for a player in cases of injury/bad performances. Different clusters represented different possible playing positions of players which was our hypothesis test three.

Authors : Anish Bhurtyal, Avanish Chaulagai, Nikita Gerzhgorin, Joseph Donohoe