Exploring FIFA20 Data

Claire Yuan, George Ma, Nick Stoner, Brian Kang, Lyons Lu

Presented by team NO LUNCH

Which club and league is the most suitable for a young player below the age of 25 to play?

Top 3 Leagues to Join:

German Bundesliga

Spain Primera Division

Italian Serie A

Honorable Mention:

English Premier League

Top 3 Clubs to Join:

Bayern Munchen

Real Madrid

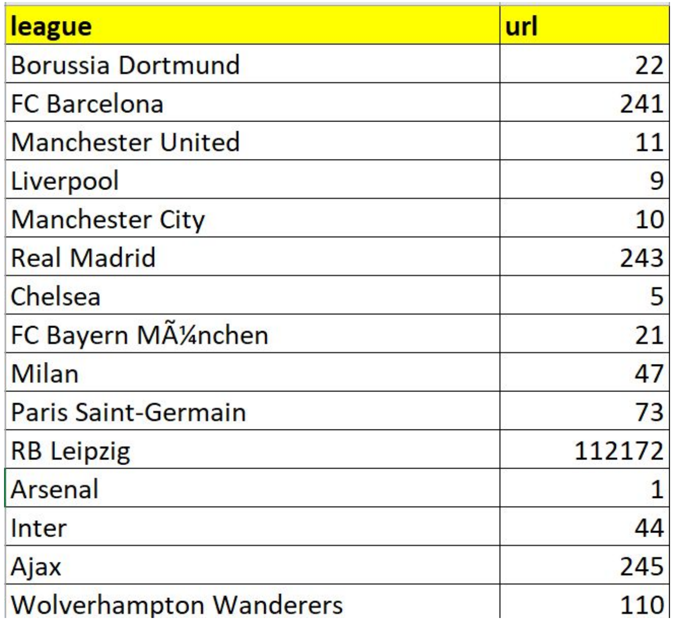
Juventus

Honorable Mention:

FC Barcelona

Web scraping & combining sets

Data source: sofifa.com



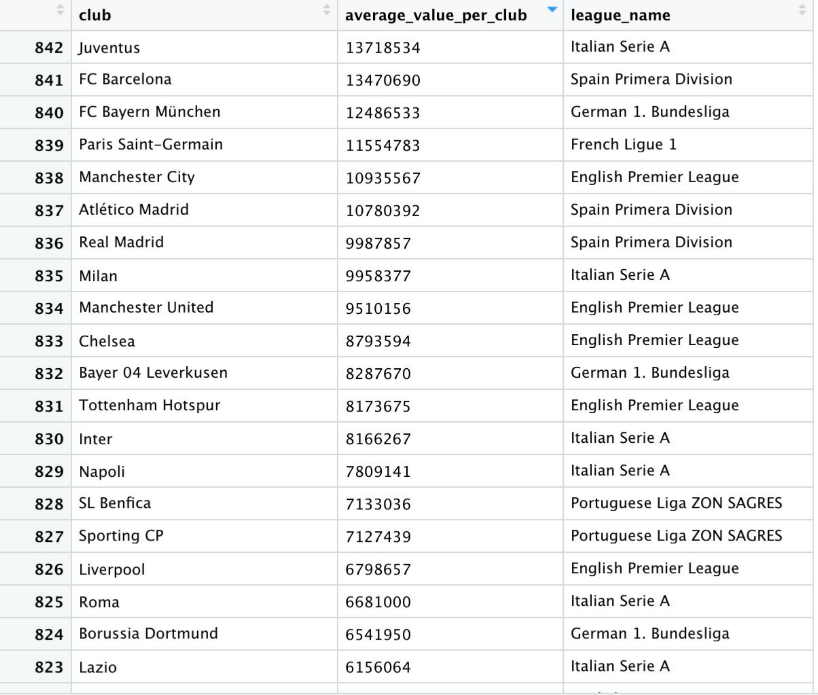
league	url
Borussia Dortmund	22
FC Barcelona	241
Manchester United	11
Liverpool	9
Manchester City	10
Real Madrid	243
Chelsea	5
FC Bayern MÃ¼nchen	21
Milan	47
Paris Saint-Germain	73
RB Leipzig	112172
Arsenal	1
Inter	44
Ajax	245
Wolverhampton Wanderers	110

cleaned columns:

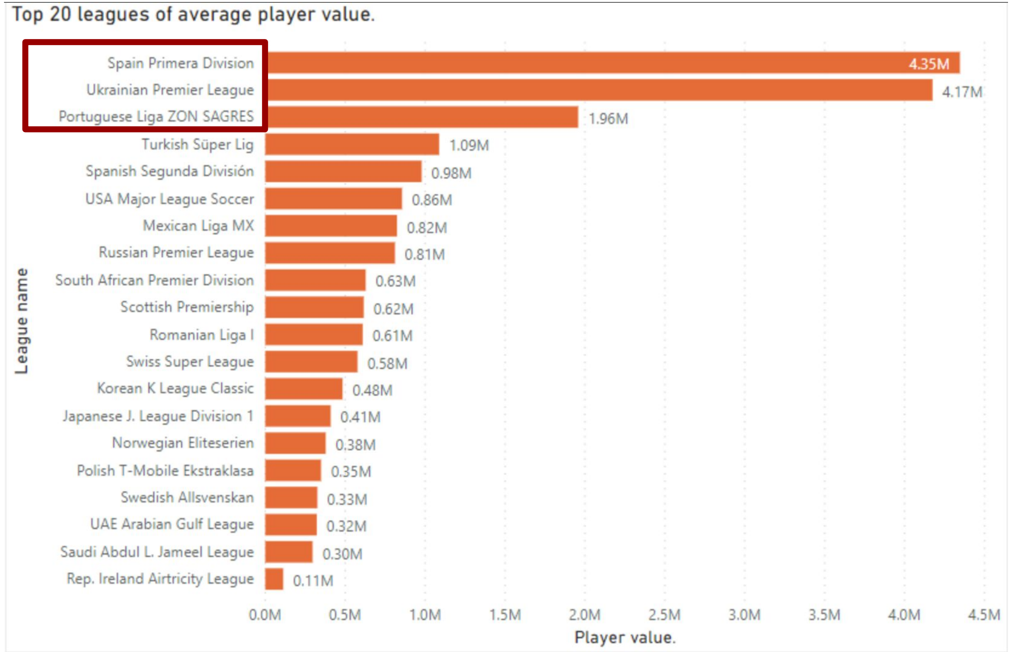
- Position columns (25 columns)
- Team jersey (2 columns)
- Player_url

Data analysis

Factor: Average player values



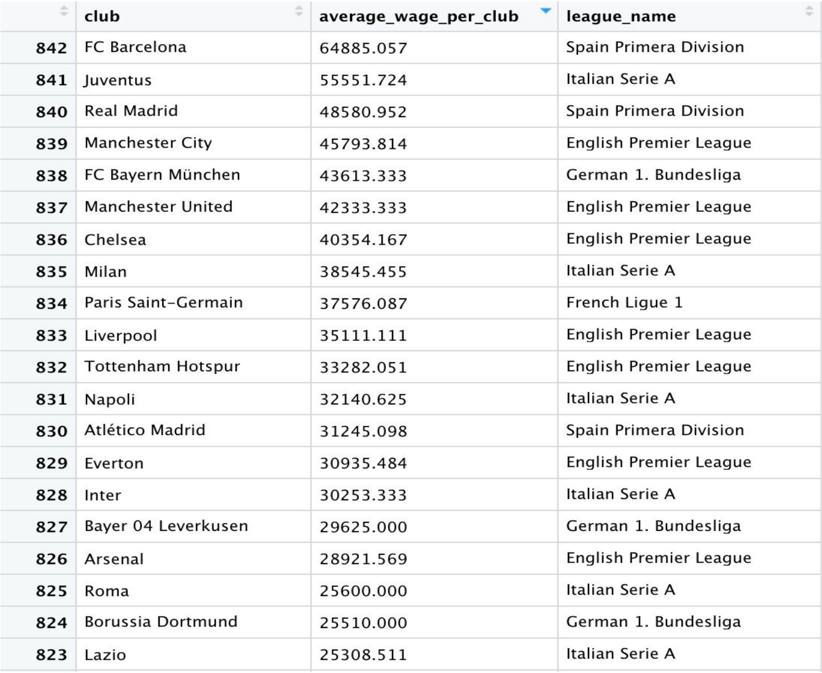
	club	average_value_per_club	league_name
842	Juventus	13718534	Italian Serie A
841	FC Barcelona	13470690	Spain Primera Division
840	FC Bayern München	12486533	German 1. Bundesliga
839	Paris Saint-Germain	11554783	French Ligue 1
838	Manchester City	10935567	English Premier League
837	Atlético Madrid	10780392	Spain Primera Division
836	Real Madrid	9987857	Spain Primera Division
835	Milan	9958377	Italian Serie A
834	Manchester United	9510156	English Premier League
833	Chelsea	8793594	English Premier League
832	Bayer 04 Leverkusen	8287670	German 1. Bundesliga
831	Tottenham Hotspur	8173675	English Premier League
830	Inter	8166267	Italian Serie A
829	Napoli	7809141	Italian Serie A
828	SL Benfica	7133036	Portuguese Liga ZON SAGRES
827	Sporting CP	7127439	Portuguese Liga ZON SAGRES
826	Liverpool	6798657	English Premier League
825	Roma	6681000	Italian Serie A
824	Borussia Dortmund	6541950	German 1. Bundesliga
823	Lazio	6156064	Italian Serie A



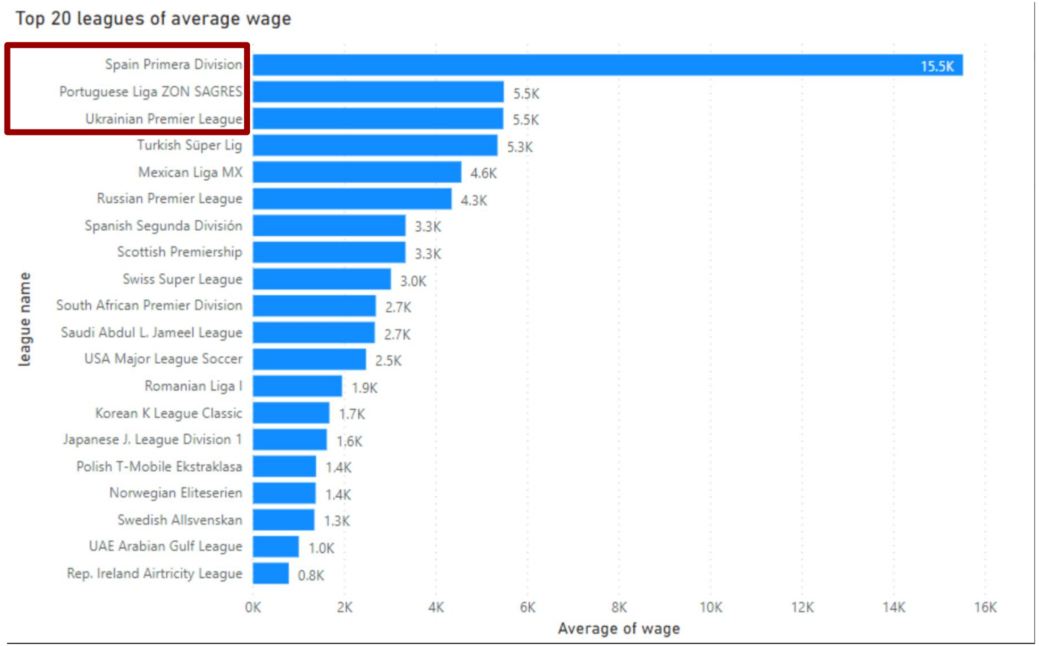
Top 20 leagues of average player value.

Data analysis

Factor: Average wage for players under age of 25



	club	average_wage_per_club	league_name
842	FC Barcelona	64885.057	Spain Primera Division
841	Juventus	55551.724	Italian Serie A
840	Real Madrid	48580.952	Spain Primera Division
839	Manchester City	45793.814	English Premier League
838	FC Bayern München	43613.333	German 1. Bundesliga
837	Manchester United	42333.333	English Premier League
836	Chelsea	40354.167	English Premier League
835	Milan	38545.455	Italian Serie A
834	Paris Saint-Germain	37576.087	French Ligue 1
833	Liverpool	35111.111	English Premier League
832	Tottenham Hotspur	33282.051	English Premier League
831	Napoli	32140.625	Italian Serie A
830	Atlético Madrid	31245.098	Spain Primera Division
829	Everton	30935.484	English Premier League
828	Inter	30253.333	Italian Serie A
827	Bayer 04 Leverkusen	29625.000	German 1. Bundesliga
826	Arsenal	28921.569	English Premier League
825	Roma	25600.000	Italian Serie A
824	Borussia Dortmund	25510.000	German 1. Bundesliga
823	Lazio	25308.511	Italian Serie A



Correlation Matrix (2020, Age < 25)

(Avgs)	Age	Value	Wage	Reputation	overall	potential	shooting	passing	dribbling
Age	1.000								
Value_euro	0.153	1.000							
Wage_euro	0.162	**0.861**	1.000						
Int’l Reputation	0.733	0.446	0.427	1.000					
overall	**0.887**	0.386	0.358	**0.776**	1.000				
potential	**0.817**	0.293	0.271	**0.779**	**0.952**	1.000			
shooting	0.560	0.324	0.292	0.506	0.668	0.624	1.000		
passing	0.706	0.382	0.352	0.635	**0.829**	**0.774**	**0.765**	1.000	
dribbling	0.715	0.346	0.322	0.644	**0.833**	**0.793**	**0.843**	**0.925**	1.000

Random Forest

Change in overall (2020 - 2015) =

age + position + potential + nationality + value_eur + experience year

```
#----------------------------------------------------------------------
# Model 1: OLS
lm.1 <- lm(formula, data = fifa20[train,])
summary(lm.1)
# prediction on test data to predict patient readmission or not
prob.lm.1 <- predict(lm.1, newdata = fifa20[-train,]) # team position throws errors
summary(prob.lm.1)
length(na.omit(prob.lm.1)) # count remaining observations
# test error
mse.1 <- mean((prob.lm.1-fifa20[-train,]$avg_overall)^2, na.rm=T)
#cat("\nMSE\n")
mse.1

#---------------------------------------------------------------------
# Model 2: OLS with feature selected by group
lm.2 <- lm(avg_overall ~ potential + value_eur + wage_eur + contract_valid_until
           + skill_moves + movement_reactions + mentality_penalties, data = fifa20[train,])
summary(lm.2)
# prediction on test data to predict patient readmission or not
prob.lm.2 <- predict(lm.2, newdata = fifa20[-train,]) # team position throws errors
summary(prob.lm.2)
length(na.omit(prob.lm.2)) # count remaining observations
# test error
mse.2 <- mean((prob.lm.2-fifa20[-train,]$avg_overall)^2, na.rm=T)
#cat("\nMSE\n")
mse.2

#
```

```r
#--------------------------------------------------------------------
# Model 3: LASSO then OLS
lasso.1 <- rlasso(formula , data = fifa20[train,], post = F)
#cat("Do LASSO on training set\n")
summary(lasso.1, all = F)

# get ceoffs that matter and make OLS formula
x <- which(coef(lasso.1)[-1]!=0)
#cat("\nCount and Kept Significant Variables by LASSO\nCount: ")
length(x)
#x
x <- paste(names(x), collapse = "+")
formula2 <- paste(c("avg_overall", x), collapse = " ~ ")

# name all extra variables created from doing OLS
fifa20$preferred_footRight <- fifa20$preferred_foot == "Right"
fifa20$nation_positionRB <- fifa20$nation_position == "RB"

# OLS regression on training set
olsLasso.1 <- lm(formula2, data = fifa20[train,])
summary(olsLasso.1)
#cat("\nDo OLS on training set using selected variables from LASSO\n")
summary(olsLasso.1)$coefficients[,1]
# prediction on test data to predict patient readmission or not
prob.lasso.1 <- predict(olsLasso.1, newdata = fifa20[-train,])
#cat("Predict on test set\n")
summary(prob.lasso.1)
#cat("\nCount remaining observations\n")
length(na.omit(prob.lasso.1)) # count remaining observations
# test error
mse.3 <- mean((prob.lasso.1-fifa20[-train,]$avg_overall)^2, na.rm=T)
#cat("\nMSE\n")
mse.3
```

Future focus