EPL Exploration

Jasin Chow

11/29/2021

# English Premier League Data Exploration, 2021

Hello :)

First things first, a little introduction.

My name is Jasin (pronounced as Jason with an ‘i’ instead of an ‘o’ :p). I have been an avid follower of football since I was 10. The rise of the use of Data Science in the sport is certainly intriguing, and being in the midst of pursuing a degree in Data Science, I hope to one day be able to work as a Data Scientist at a European football club.

This little exploratory project makes use of 2 datasets : 1. Individual player stats from the COVID-hit 2020/2021 season 2. All results since the Premier League’s Inception

# Setting up the environment

player\_stats\_20\_21 <- read.csv('EPL\_20\_21.csv')  
all\_results <- read.csv('results.csv')  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

# Player stats

## Top starters

Let’s see if any outfield player managed to play all 38 games last season, which would be quite the feat given that pre-season was extremely rushed with the delayed end of 2019/2020 due to COVID putting a pause on all matches midway through the 19/20 season

filter(player\_stats\_20\_21, Matches == 38, Position != 'GK') %>% select(Name)

## Name  
## 1 Andrew Robertson  
## 2 Georginio Wijnaldum  
## 3 Youri Tielemans  
## 4 Tomáš Souček  
## 5 Jarrod Bowen  
## 6 Pierre Højbjerg  
## 7 Stuart Dallas  
## 8 Luke Ayling  
## 9 Patrick Bamford  
## 10 Matt Targett  
## 11 James Ward-Prowse  
## 12 Ashley Westwood

Here we can see the TWELVE players who achieved this feat, which is commendable. In particular, Liverpool’s Robertson, West Ham’s Souček, Spurs’ Højbjerg, Leeds’ Dallas and Ayling, Villa’s Targett, Southampton’s Ward-Prowse and Burnley’s Westwood all STARTED the 38 matches as well!

## Top Scorers

I shall first do a simple arrangement of the players based on the number of goals scored.

arrange(player\_stats\_20\_21, desc(Goals)) %>%  
 select(Name, Goals) %>%  
 top\_n(10, Goals)

## Name Goals  
## 1 Harry Kane 23  
## 2 Mohamed Salah 22  
## 3 Bruno Fernandes 18  
## 4 Son Heung-min 17  
## 5 Patrick Bamford 17  
## 6 Dominic Calvert-Lewin 16  
## 7 Jamie Vardy 15  
## 8 Ollie Watkins 14  
## 9 İlkay Gündoğan 13  
## 10 Alexandre Lacazette 13

Harry Kane leads the charts, followed by Mohamed Salah and Bruno Fernandes. Here, the appearance of Alex Lacazette in the top 10 is surprising, given how much backlash the Frenchman is getting from Arsenal supporters for his poor performances. Gundogan’s goal scoring exploits last season were evident as well, as he, quite amazingly, became City’s top scorer with 13 goals from midfield, even though goal scoring is not a trait we commonly associate with the German playmaker.

## Top non-penalty scorers

Now, most fans are used to seeing top scorers in seasons gone by. Personally, I find them to be a little uninformative, especially since they account for penalties. I am acutely aware of fans on social media calling players ‘Penalty Merchants’ i.e. players with more goals than their performances deserve due to penalty taking duties. Let’s check the top 10 goalscorers excluding penalties

player\_stats\_20\_21 <- mutate(player\_stats\_20\_21, Non\_penalty\_Goals = (Goals - Penalty\_Goals), .after = Goals) #permanently add non-penalty goals to the data  
arrange(player\_stats\_20\_21, desc(Non\_penalty\_Goals)) %>%   
 select(Name, Non\_penalty\_Goals) %>%   
 top\_n(10, Non\_penalty\_Goals)

## Name Non\_penalty\_Goals  
## 1 Harry Kane 19  
## 2 Mohamed Salah 16  
## 3 Son Heung-min 16  
## 4 Dominic Calvert-Lewin 16  
## 5 Patrick Bamford 15  
## 6 Ollie Watkins 13  
## 7 İlkay Gündoğan 12  
## 8 Kelechi Iheanacho 12  
## 9 Marcus Rashford 11  
## 10 Sadio Mané 11  
## 11 Gareth Bale 11

So Kane and Salah still lead the way, with the most noticeable absentee from the previous table being Bruno Fernandes, who doesn’t even make the top 10 for non-penalty goals scored. Perhaps the name ‘Penandes’ is warranted….

## Top attackers

In recent years, the popularity of the metric ‘expected goals / xG’ has risen exponentially. Goals are no longer the sole useful metric to judge attackers, as they may be putting away low quality chances which require excellent finishing. This is not always sustainable across an extended period of time, requiring the attacker to consistently be in top condition to put away poor chances. This could be used to identify ‘one season wonders’, where a player performs exceptionally within a season, but completely nosedives in the subsequent seasons. Some of these, in recent seasons, include : Papiss Cisse (Newcastle), Pavel Pogrebnyak (Reading).

Expected goals take into account various factors when a shot has been placed : goalkeeper’s position, shooting position, shooting body part, defender’s position. These are then used to determine how likely, based on past data, a player should score given these factors when a shot has been taken.

### Best attackers (Expected goals)

Let’s see the 20 players with the top expected goals for the season (xG \* Matches)

mutate(player\_stats\_20\_21, expected\_goals = (xG \* Matches)) %>%  
 select(Name, Position, expected\_goals, Goals) %>%  
 arrange(desc(expected\_goals)) %>%  
 top\_n(20, expected\_goals)

## Name Position expected\_goals Goals  
## 1 Mohamed Salah FW 22.57 22  
## 2 Jamie Vardy FW 21.08 15  
## 3 Harry Kane FW 21.00 23  
## 4 Patrick Bamford FW 19.38 17  
## 5 Sadio Mané FW 17.15 11  
## 6 Bruno Fernandes MF 16.65 18  
## 7 Ollie Watkins FW 15.91 14  
## 8 Dominic Calvert-Lewin FW 15.84 16  
## 9 Neal Maupay FW,MF 15.18 8  
## 10 Chris Wood FW 14.52 12  
## 11 Roberto Firmino FW 14.40 9  
## 12 Timo Werner FW 14.35 6  
## 13 Alexandre Lacazette FW 14.26 13  
## 14 Michail Antonio FW 13.52 10  
## 15 Raheem Sterling FW 13.33 10  
## 16 Fábio Silva FW 12.80 4  
## 17 Callum Wilson FW 12.74 12  
## 18 Anwar El Ghazi FW 12.60 10  
## 19 Edinson Cavani FW 12.48 10  
## 20 Tammy Abraham FW 12.32 6

From the top 20, we can make a couple of observations : 1. As expected, the top 20 players in terms of xG are all forwards 2. While some players’ goalscoring numbers are similar to their xG numbers, such as Salah and Fernandes, there are also those with a relatively bigger disparity, such as Mane and Maupay. Naturally, I would want to see which players’ had the greatest differences, both positive and negative, in terms of goals and expected goals.

### Overperforming attackers

mutate(player\_stats\_20\_21, expected\_goals = (xG \* Matches), xG\_diff = Goals - expected\_goals) %>%  
 select(Name, Position, Goals,xG\_diff) %>%  
 arrange(desc(xG\_diff)) %>%  
 top\_n(20, xG\_diff)

## Name Position Goals xG\_diff  
## 1 Son Heung-min FW 17 5.90  
## 2 Jesse Lingard MF,FW 9 3.88  
## 3 Stuart Dallas DF,MF 8 3.44  
## 4 Joe Willock MF 8 3.38  
## 5 Wilfried Zaha FW 11 3.20  
## 6 Kurt Zouma DF 5 3.08  
## 7 Matheus Pereira MF,FW 11 2.75  
## 8 James Ward-Prowse MF 8 2.68  
## 9 Lewis Dunk DF 5 2.36  
## 10 Danny Ings FW 12 2.14  
## 11 James Maddison MF,FW 8 2.11  
## 12 James Rodríguez FW,MF 6 2.09  
## 13 Harry Kane FW 23 2.00  
## 14 Riyad Mahrez FW 9 1.98  
## 15 Ashley Westwood MF 3 1.86  
## 16 John Stones DF 4 1.80  
## 17 İlkay Gündoğan MF 13 1.80  
## 18 Harvey Barnes FW,MF 9 1.75  
## 19 Issa Diop DF 2 1.64  
## 20 Angelo Ogbonna DF 3 1.60  
## 21 Michael Keane DF 3 1.60

Here we can see Son overperforming his expected goals by a staggering 6 goals, with no other player on more than 4. This goes to show how good the Korean’s finishing is, but whether this overperformance is sustainable remains to be seen. There are a few defenders on this list, which is a testament to how big a threat these players are on set pieces. It is also of little surprise, personally, that Ward-Prowse features here, given that a majority of his goals are scored from free-kicks, all of which are taken from outside the penalty area and therefore have much lower expected goals. I also see that 3 West Ham defenders (Zouma, Diop, Ogbonna) feature, implying that set pieces played a major part in the Clarets’ push for top 4 that season. ### Underperforming attackers

mutate(player\_stats\_20\_21, expected\_goals = (xG \* Matches), xG\_diff = expected\_goals - Goals) %>%  
 select(Name, Position, Goals,xG\_diff) %>%  
 arrange(desc(xG\_diff)) %>%  
 top\_n(20, xG\_diff)

## Name Position Goals xG\_diff  
## 1 Fábio Silva FW 4 8.80  
## 2 Timo Werner FW 6 8.35  
## 3 Aleksandar Mitrović FW 3 8.34  
## 4 Matěj Vydra FW 3 8.20  
## 5 Neal Maupay FW,MF 8 7.18  
## 6 Eddie Nketiah FW 2 6.84  
## 7 Tammy Abraham FW 6 6.32  
## 8 Sadio Mané FW 11 6.15  
## 9 Jamie Vardy FW 15 6.08  
## 10 Kai Havertz MF,FW 4 5.99  
## 11 Alireza Jahanbakhsh FW,MF 0 5.88  
## 12 Olivier Giroud FW 4 5.86  
## 13 Dwight Gayle FW,MF 1 5.48  
## 14 Martinelli FW 2 5.42  
## 15 Roberto Firmino FW 9 5.40  
## 16 Joshua King MF,FW 0 5.28  
## 17 Richarlison FW 7 5.24  
## 18 Kevin De Bruyne MF 6 5.00  
## 19 Josh Maja FW,MF 3 4.95  
## 20 Anthony Martial FW 4 4.80  
## 21 Aaron Connolly FW 2 4.80

Wow. From this table we can see how much some players underperformed in terms of goals scored, with Fabio Silva of Wolves being the biggest culprit, having scored close to 9 goals less than he reasonably should have. Timo Werner also had his critics last season, and the data shows that it was justified. On that note, the top 10 already features 3 Chelsea forwards (Werner, Havertz and Abraham). This would probably go a long way to explaining why the London club had been actively searching for a top-class centre forward for the new season, and splashed 100m on Romelu Lukaku. Another notable inclusion is Neal Maupay of Brighton, who scored 7.18 less goals than expected. This is siginificant as Brighton underperformed the most in terms of points vs expected points (which are, of course, based on xG). We can see that a big part of this might be due to Maupay’s finishing.

## Top playmakers

I shall repeat the analyses above on assists and expected assists (xA).

select(player\_stats\_20\_21,Name, Position, Assists) %>%  
 arrange(desc(Assists)) %>%  
 top\_n(20, Assists)

## Name Position Assists  
## 1 Harry Kane FW 14  
## 2 Kevin De Bruyne MF 12  
## 3 Bruno Fernandes MF 12  
## 4 Son Heung-min FW 10  
## 5 Jack Grealish FW,MF 10  
## 6 Marcus Rashford FW 9  
## 7 Jamie Vardy FW 9  
## 8 Raphael Dias Belloli MF 9  
## 9 Timo Werner FW 8  
## 10 Aaron Cresswell DF 8  
## 11 Jack Harrison MF 8  
## 12 Pascal Groß MF,DF 8  
## 13 Raheem Sterling FW 7  
## 14 Andrew Robertson DF 7  
## 15 Trent Alexander-Arnold DF 7  
## 16 Roberto Firmino FW 7  
## 17 Sadio Mané FW 7  
## 18 Vladimír Coufal DF 7  
## 19 Patrick Bamford FW 7  
## 20 Lucas Digne DF 7  
## 21 James Ward-Prowse MF 7

Notice how Kane and Son feature again? This goes to show how crucial the duo were to Spurs last seasoon. On the other hand, the fact that they finished 7th last season also shows how poor the rest of the squad is, a far cry from the Pochettino team which finished 2nd in 2015/2016. The same could be said for Bruno Fernandes and Jamie Vardy.

mutate(player\_stats\_20\_21, expected\_assists = (xA \* Matches)) %>%  
 select(Name, Position, expected\_assists, Assists) %>%  
 arrange(desc(expected\_assists)) %>%  
 top\_n(20, expected\_assists)

## Name Position expected\_assists Assists  
## 1 Bruno Fernandes MF 11.84 12  
## 2 Kevin De Bruyne MF 11.50 12  
## 3 Alireza Jahanbakhsh FW,MF 11.34 1  
## 4 Son Heung-min FW 9.62 10  
## 5 Raphael Dias Belloli MF 9.30 9  
## 6 Jack Grealish FW,MF 9.10 10  
## 7 Mason Mount MF,FW 8.64 5  
## 8 Trent Alexander-Arnold DF 8.64 7  
## 9 Saïd Benrahma FW,MF 8.40 6  
## 10 Jack Harrison MF 7.92 8  
## 11 Harry Kane FW 7.70 14  
## 12 Timo Werner FW 7.35 8  
## 13 Riyad Mahrez FW 7.29 6  
## 14 Roberto Firmino FW 7.20 7  
## 15 James Maddison MF,FW 7.13 5  
## 16 Marc Albrighton DF,FW 7.13 5  
## 17 Xherdan Shaqiri MF,FW 6.86 2  
## 18 Andrew Robertson DF 6.84 7  
## 19 Gylfi Sigurðsson MF 6.84 5  
## 20 Che Adams FW 6.84 5  
## 21 Dwight McNeil MF 6.84 5

Now when we include expected assists, something immediately stands out. Jahanbakhsh has over 11 expected assists but only ONE assist. This corroborates with my previous statement of Brighton being the biggest underperforming team, and Neal Maupay is probably culpable for a few of these chances which should have been put away.

### Big gaps in assists?

Next I will see where the biggest differences in expected and actual assists come from.

mutate(player\_stats\_20\_21, expected\_assists = (xA \* Matches), xA\_diff = Assists - expected\_assists) %>%  
 select(Name, Position, Assists,xA\_diff) %>%  
 arrange(desc(xA\_diff)) %>%  
 top\_n(20, xA\_diff)

## Name Position Assists xA\_diff  
## 1 Harry Kane FW 14 6.30  
## 2 Marcus Rashford FW 9 4.93  
## 3 Jamie Vardy FW 9 3.56  
## 4 James Ward-Prowse MF 7 2.82  
## 5 Eberechi Eze MF 6 2.60  
## 6 Jesse Lingard MF,FW 4 2.56  
## 7 Patrick Bamford FW 7 2.44  
## 8 Wilfred Ndidi MF,DF 4 2.44  
## 9 Callum Wilson FW 5 2.40  
## 10 Harvey Barnes FW,MF 4 2.25  
## 11 Vladimír Coufal DF 7 2.24  
## 12 Michail Antonio FW 5 2.14  
## 13 Pierre Højbjerg MF 4 2.10  
## 14 John McGinn MF 5 2.04  
## 15 Stuart Armstrong MF 5 2.03  
## 16 Ben Chilwell DF 5 2.03  
## 17 Jordan Ayew FW,MF 3 2.01  
## 18 Aaron Cresswell DF 8 1.88  
## 19 Bernardo Silva MF,FW 6 1.84  
## 20 Raheem Sterling FW 7 1.73

We see that Kane has outperformed his expected assists by more than 6 goals. Previously we saw how Heung-min Son overperformed in the goals department, so this should come as no surprise. I also notice quite a few of the West Ham players featuring in this top 20, namely Lingard, Coufal, Antonio and Cresswell. It might be fair to say that the finishing of the entire team was above average, since only Lingard featured in the overperforming attackers chart. Cresswell’s stats might be attributed to the defenders’ outscoring their xG, as mentioned above, since he is their set piece taker.

Now for the biggest negative differences…

mutate(player\_stats\_20\_21, expected\_assists = (xA \* Matches), xA\_diff = expected\_assists - Assists) %>%  
 select(Name, Position, Assists,xA\_diff) %>%  
 arrange(desc(xA\_diff)) %>%  
 top\_n(20, xA\_diff)

## Name Position Assists xA\_diff  
## 1 Alireza Jahanbakhsh FW,MF 1 10.34  
## 2 Xherdan Shaqiri MF,FW 2 4.86  
## 3 Adama Traoré FW 2 4.66  
## 4 Ivan Cavaleiro FW,MF 0 4.32  
## 5 Ross Barkley MF 1 4.04  
## 6 Matt Phillips MF 1 3.95  
## 7 Adam Lallana MF,FW 1 3.80  
## 8 Mason Mount MF,FW 5 3.64  
## 9 Anwar El Ghazi FW 0 3.64  
## 10 İlkay Gündoğan MF 2 3.60  
## 11 Rhian Brewster FW 0 3.51  
## 12 Martinelli FW 1 3.48  
## 13 Hakim Ziyech FW,MF 3 3.44  
## 14 Cengiz Ünder MF,FW 2 3.13  
## 15 Fred MF 0 3.00  
## 16 Alexis Mac Allister MF,FW 1 2.99  
## 17 Callum Hudson-Odoi FW,DF 3 2.98  
## 18 Ian Carlo Poveda MF,DF 0 2.94  
## 19 Ben Osborn MF,DF 0 2.88  
## 20 Érik Lamela FW,MF 0 2.76

As expected, Jahanbakhsh comes out on top here. 4 Chelsea players feature here (Mount, Barkley, Ziyech, Hudson-Odoi) which is no surprise given how many of the team’s forwards we saw in the underperforming scorers’ chart. Traore and Cavaleiro both play for Wolves, for whom Fabio Silva plays. I would assume their combined 9 ‘missing’ assists can be largely attributed to the young Portugese forward.

### Decisive passers

One thing I thought would be interesting to explore is the number of passes made per xA for the players. I define this to be a measure how ‘pass decisiveness’. Of course, this isn’t the perfect metric since this stat might be affected by team style. For example, a possession dominant team would most likely take more passes to break down the opposition, while the converse is true for counter attacking teams who look to get the ball from one end of the pitch to the other as quickly as possible.

select(player\_stats\_20\_21, Name, Passes\_Attempted, Matches, xA, Assists) %>%  
 mutate(total\_xA = Matches \* xA, Passes\_Per\_xA = Passes\_Attempted/total\_xA, Passes\_Per\_A = Passes\_Attempted/Assists) %>%  
 select(Name, Assists, Passes\_Per\_xA, Passes\_Per\_A) %>%  
 arrange(Passes\_Per\_A) %>%  
 filter(Assists > 5)

## Name Assists Passes\_Per\_xA Passes\_Per\_A  
## 1 Jamie Vardy 9 83.08824 50.22222  
## 2 Harry Kane 14 121.68831 66.92857  
## 3 Patrick Bamford 7 110.96491 72.28571  
## 4 Saïd Benrahma 6 62.73810 87.83333  
## 5 Timo Werner 8 112.38095 103.25000  
## 6 Jack Grealish 10 120.87912 110.00000  
## 7 Kevin De Bruyne 12 122.26087 117.16667  
## 8 Raphael Dias Belloli 9 113.65591 117.44444  
## 9 Son Heung-min 10 124.63617 119.90000  
## 10 Bertrand Traoré 6 170.08547 132.66667  
## 11 Marcus Rashford 9 303.19410 137.11111  
## 12 Jack Harrison 8 146.71717 145.25000  
## 13 Sadio Mané 7 168.88889 152.00000  
## 14 Raheem Sterling 7 213.85199 161.00000  
## 15 Matheus Pereira 6 170.01595 177.66667  
## 16 Riyad Mahrez 6 148.97119 181.00000  
## 17 Roberto Firmino 7 181.66667 186.85714  
## 18 Bruno Fernandes 12 192.82095 190.25000  
## 19 Eberechi Eze 6 340.58824 193.00000  
## 20 Pedro Neto 6 177.71261 202.00000  
## 21 Pascal Groß 8 245.44118 208.62500  
## 22 Bernardo Silva 6 343.02885 237.83333  
## 23 Lucas Digne 7 312.96296 241.42857  
## 24 Vladimír Coufal 7 355.46218 241.71429  
## 25 Aaron Cresswell 8 336.60131 257.50000  
## 26 James Ward-Prowse 7 626.55502 374.14286  
## 27 Trent Alexander-Arnold 7 340.39352 420.14286  
## 28 Andrew Robertson 7 469.88304 459.14286

For this analysis, I have only included players with more than 5 assists. Unsurprisingly, the top 4 are all forwards. These players tend to be less involved in overall buildup play, and their touches, and hence passes, tend to be in and around the opposition penalty area, so these are more likely to end up in goals. The exception here would probably be Harry Kane, who often drops deep to build up play as well, as we can see from his Passes Per xA, which is much higher than his Passes Per Assist, implying his low PPA is due to great finishing, as discussed earlier. It is also clear that most of the players in the top 10 are playing for counter attacking teams such as Leicester City, West Ham United and Aston Villa. Leeds’ Bamford and Raphinha are also involved due to their high pressing style which leads to balls being won higher up the pitch and therefore decisive passes are easier to make.

## Entertaining teams

Let’s take a step back and look at data on a team level now. As followers of the sport, fans’ main form of entertainment is goals. I shall now see which teams were involved in the most number of goals.

player\_stats\_20\_21 %>%  
 group\_by(Club) %>%  
 summarize(goals\_scored = sum(Goals)) %>%  
 arrange(desc(goals\_scored))

## # A tibble: 20 x 2  
## Club goals\_scored  
## <chr> <int>  
## 1 Manchester City 82  
## 2 Manchester United 70  
## 3 Tottenham Hotspur 66  
## 4 Liverpool FC 65  
## 5 Leicester City 64  
## 6 Leeds United 60  
## 7 West Ham United 60  
## 8 Chelsea 56  
## 9 Arsenal 53  
## 10 Aston Villa 52  
## 11 Southampton 47  
## 12 Everton 45  
## 13 Newcastle United 44  
## 14 Brighton 39  
## 15 Crystal Palace 39  
## 16 Wolverhampton Wanderers 34  
## 17 West Bromwich Albion 33  
## 18 Burnley 32  
## 19 Fulham 26  
## 20 Sheffield United 19

Here we see Man City being, by far, the most prolific team in the league, averaging more than 2 goals scored a game. It is little wonder they won the title that year.