Final Report-Expected Data and Player's Value

Yuehan Duan

2019-5-5

There're dozens of leagues in the soccer world, there's no doubt that European football(soccer) is much stronger than other continents. And among European football leagues, England Premier Leagues (EPL) is the most-watched football league in the world, broadcast in 212 territories to 643 million homes and a potential TV audience of 4.7 billion people. What's more, EPL is also famous for big clubs like Manchester United, Liverpool, Arsenal, Chelsea, Manchester City, Tottenham. No other league has as many big clubs as the English Premier League does. They are among the world's very elite and awash with some of the most exciting talents.

Since I want to find out relationship between player's transfer value and their performance, I will focus on the EPL players and more specifically, Top 100 players in EPL ranked with their transfer values.

First, we must get the data we want. Let's start with the transfer markets website (https://www.transfermarkt.co.uk/) which is the most authoritative website in the field of soccer transfer.

The data of players were stored in 4 pages, I need to write a loop for them and find the nodes where the data I needed by viewing its CSS source through chrome and clear the data.

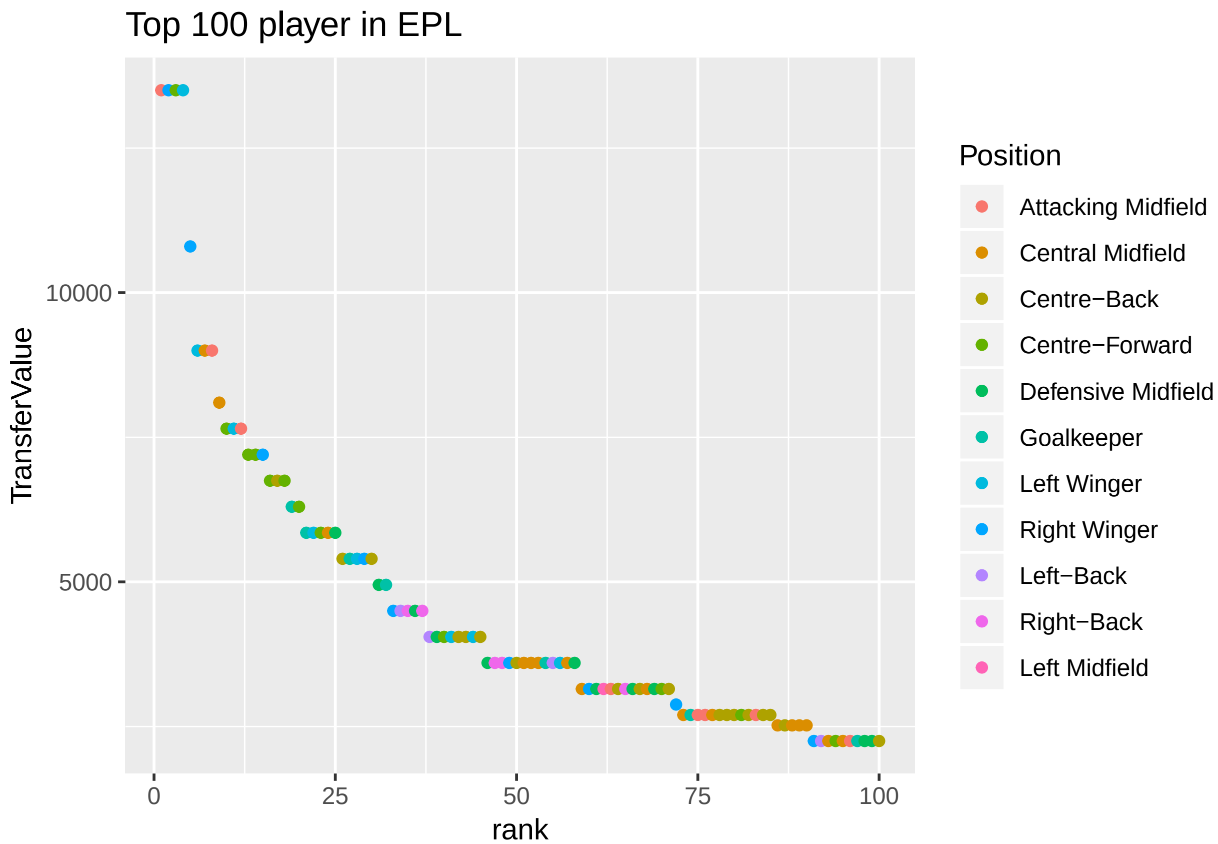
Then, get name, position, age and transfer value for each player, data samples of top10 player is shown below:

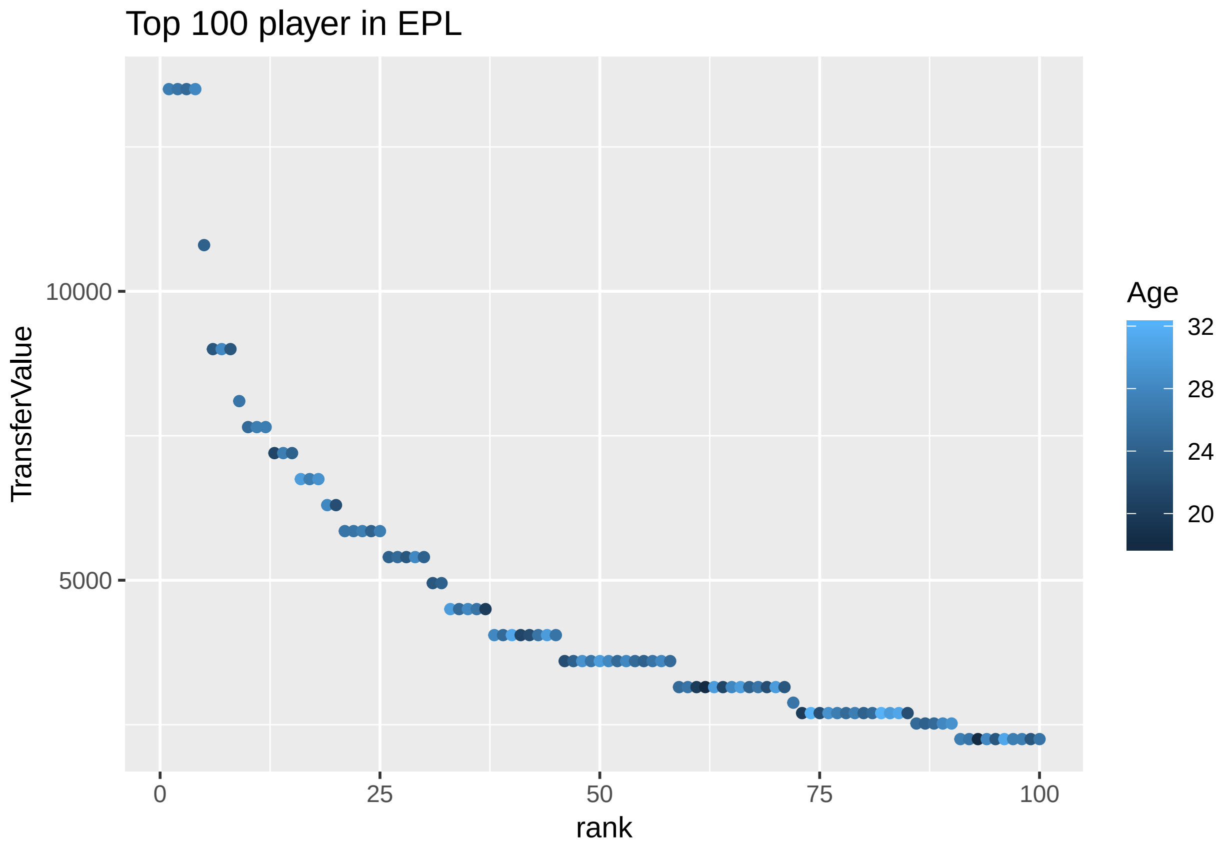
```
##         player_name TransferValue Age           Position
## 1  Kevin De Bruyne      135.00m  27 Attacking Midfield
## 2    Mohamed Salah      135.00m  26       Right Winger
## 3       Harry Kane      135.00m  25     Centre-Forward
## 4      Eden Hazard      135.00m  28        Left Winger
## 5  Raheem Sterling      108.00m  24       Right Winger
## 6       Leroy Sané       90.00m  23        Left Winger
## 7     N'Golo Kanté       90.00m  28   Central Midfield
## 8        Dele Alli       90.00m  23 Attacking Midfield
## 9       Paul Pogba       81.00m  26   Central Midfield
## 10   Romelu Lukaku       76.50m  25     Centre-Forward
```

The unit of transfer values in the data is million pounds with some Financial symbols, since all the Top 100 players' transfer values are the same unit, I think it's better to clean those symbols and turn them into numbers. I also add a feature called rank indicate the rank of player among Top 100 by transfer value. Data samples of top10 player is shown below, the unit of transfer values is 10000 pounds:

```
##         player_name TransferValue Age           Position
## 1  Kevin De Bruyne         13500  27 Attacking Midfield
## 2    Mohamed Salah         13500  26       Right Winger
## 3       Harry Kane         13500  25     Centre-Forward
## 4      Eden Hazard         13500  28        Left Winger
## 5  Raheem Sterling         10800  24       Right Winger
## 6       Leroy Sané          9000  23        Left Winger
## 7     N'Golo Kanté          9000  28   Central Midfield
## 8        Dele Alli          9000  23 Attacking Midfield
## 9       Paul Pogba          8100  26   Central Midfield
## 10   Romelu Lukaku          7650  25     Centre-Forward
```

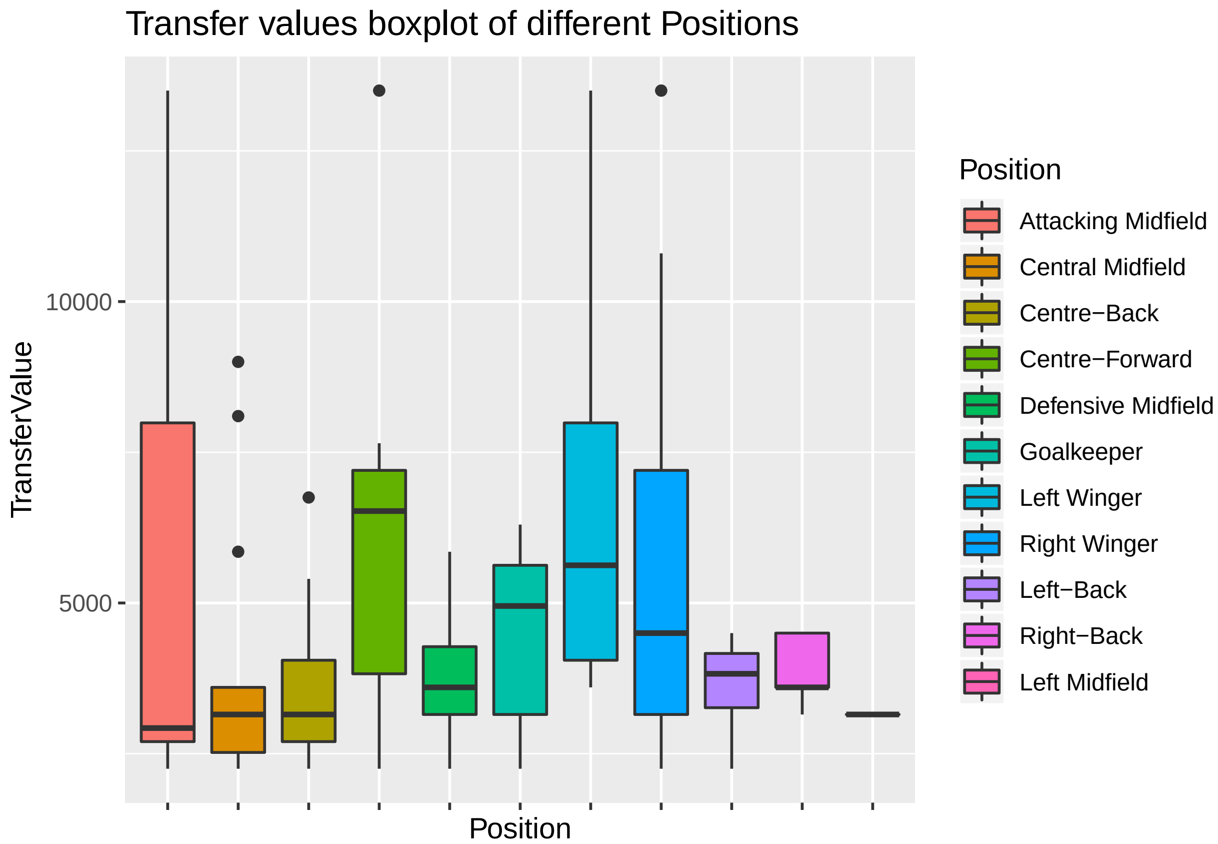
Let's go explore the data. Plot the transfer values again rank, and different color indicate different positions or ages.





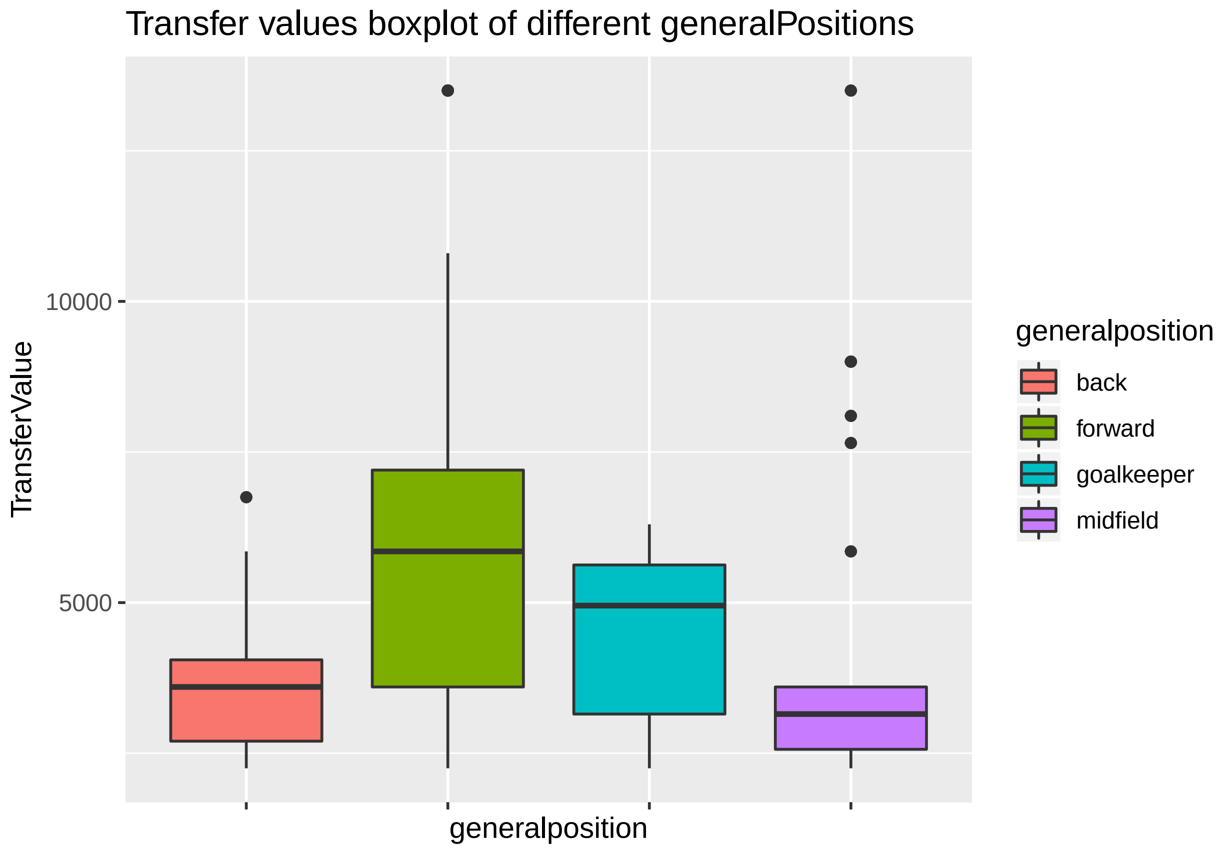
We can see that the distribution is not linear and transfer values for attacking players are seems higher than defensing players in average. And most players are under 28 and we can see that younger player are tend to have a higher transfer value. In fact, the average age of top 100 players is 25.79 which is lower than the average age of football player all over the league(27.08).

To explore more on the impact of positions, give a plot of average transfer values for each position.



We can see that center forward has the highest average transfer value, the values of wingers (left winger and right winger) are also high, it seems that forward's transfer value is higher than middle field and back's values.

Thus, we can mutate a new variable called "general position" which include forward, middle field, back and goalkeeper by the basic knowledge of soccer.




```
##          player_name TransferValue Age           Position rank
## 1  Kevin De Bruyne         13500  27 Attacking Midfield    1
## 2    Mohamed Salah         13500  26       Right Winger    2
## 3       Harry Kane         13500  25     Centre-Forward    3
## 4      Eden Hazard         13500  28        Left Winger    4
## 5  Raheem Sterling         10800  24       Right Winger    5
## 6       Leroy Sané          9000  23        Left Winger    6
## 7      N'Golo Kanté          9000  28   Central Midfield    7
## 8        Dele Alli          9000  23 Attacking Midfield    8
## 9       Paul Pogba          8100  26   Central Midfield    9
## 10   Romelu Lukaku          7650  25     Centre-Forward   10
##    generalposition
## 1         midfield
## 2          forward
## 3          forward
## 4          forward
## 5          forward
## 6          forward
## 7         midfield
## 8         midfield
## 9         midfield
## 10         forward
```

It's clear from the boxplot that forward has the highest average value, followed by goalkeeper, back and middle field.

```
## # A tibble: 11 x 2
##    Position               n
##    <fct>              <int>
##  1 Centre-Back           18
##  2 Central Midfield      17
##  3 Centre-Forward        12
##  4 Defensive Midfield    11
##  5 Right Winger           9
##  6 Attacking Midfield     8
##  7 Left Winger            8
##  8 Goalkeeper             7
##  9 Right-Back             5
## 10 Left-Back              4
## 11 Left Midfield          1
```

We can see most of players in Top 100 values are in center(center back, center midfield and center forward). It shows that center area is still the most important part for football.

To do deeper researches, more data is needed, and we need some data that can reflect player's performance in field. However, football(soccer) is a low scoring game that final match score does not provide a clear picture of performance. What's more, football players in different positions have different responsibility. You can't expect a goalkeeper to score a goal or a forward do a lot of defense. Those features of football indicate that basic data like goals and assists is not enough to evaluate player's performance and players in different positions needs different data to evaluate.

At first, I want to use data from whoscored (https://www.whoscored.com/), which is one of the most popular football data website and is famous for it's machine rating system for every players and every matches. I thought it would be a great data to evaluate player's performance.

However, the website has a system called Incapsula that can reject scraping. This system can identify whether you are using selenium, phantomJS, etc. Each time you enter the page, a cookie will be generated for the user's test results, and then the request will carry the test cookie and return other cookies to gain access to the site. But even if the access is authorized, too many requests will trigger the Incapsula system.

The data for each player stores in different pages that I need at least 100 request to get the data, however, only 5 or even less request will trigger this system. Thus I need find another website to get data.

For forward, the most important job of them is to score a goal. However, sometimes the chance is created by your teammate, but all the data says is just 1 goal. So, I use the statistical measure called expected goal (xG), which is measurement of the quality of chance player received range from 0 to 1 each time, to evaluate forward's performance of shooting.

The higher value of xG, the better the chance is; thus, we can also define expected assist (xA) which measure the quality of chances provided by player.

For this case, researchers trained neural network prediction algorithms with large dataset (>100000 shots, over 10 parameters for each), I scraping and cleaning this kind of data from understat(https://understat.com/) and the data was Json.

Join the 2 data i get from different websites, the cleaned data samples are shown below:

Since we need do some regression, clean the data into numbers and save in a csv file called "data".

```
##        player_name TransferValue Age           Position rank
## 1  Kevin De Bruyne         13500  27 Attacking Midfield    1
## 2    Mohamed Salah         13500  26       Right Winger    2
## 3       Harry Kane         13500  25     Centre-Forward    3
## 4      Eden Hazard         13500  28        Left Winger    4
```

```
## 5  Raheem Sterling         10800  24       Right Winger    5
## 6       Leroy Sané          9000  23        Left Winger    6
## 7     N'Golo Kanté          9000  28   Central Midfield    7
## 8        Dele Alli          9000  23 Attacking Midfield    8
## 9       Paul Pogba          8100  26   Central Midfield    9
## 10   Romelu Lukaku          7650  25     Centre-Forward   10
##    generalposition   id games time goals        xG assists        xA shots
## 1         midfield  447    18  954     2  1.429502       2  6.654021    30
## 2          forward 1250    37 3184    22 21.360759       8 10.468590   132
## 3          forward  647    28 2437    17 16.122394       4  4.562663   102
## 4          forward  701    36 2915    16 12.299006      15 11.548123    93
## 5          forward  618    33 2698    17 15.805114      10 10.650952    76
## 6          forward  337    31 1866    10  6.981944      10  8.101671    56
## 7         midfield <NA>    NA   NA    NA        NA      NA        NA    NA
## 8         midfield  645    24 1800     5  5.828909       3  3.293327    37
## 9         midfield 1740    34 2923    13 15.700942       9  5.142453   102
## 10         forward  594    32 2113    12 13.105178       0  2.320214    55
##    key_passes yellow_cards red_cards npg      npxG  xGChain xGBuildup
## 1          36            2         0   2  1.429502 12.07782  8.357447
## 2          68            1         0  19 19.077253 31.34062  7.809351
## 3          30            5         0  13 13.077756 18.83823  4.841164
## 4          97            2         0  12  9.254331 25.30644 11.546570
## 5          65            3         0  17 15.805114 32.32803 12.182243
## 6          40            1         0  10  6.981944 21.35401 10.558323
## 7          NA         <NA>      <NA>  NA        NA       NA        NA
## 8          27            4         0   5  5.828909 12.83371  5.540953
## 9          54            5         0   6  8.089253 20.69964 11.227801
## 10         21            4         0  12 13.105178 15.42570  5.426005
```

Where npg means none-penalty goal and npxG means none-penalty expected goal, xGChain means total xG of every possession the player is involved in and xGBulidup means total xG of every possession the player is involved in without key passes or shots. xGChain and xGBulidup can reflect how helpful this player is for the team during attacking.

Let's first analyze forward's data, the major job for them is attacking, so I choose data that related to attacking (goal, xG, assist, xA, shots, key passes, npg, npxG, xGChain, xGBulidup) and player's age to build a regression model for transfer values.

```
##
## Call:
## lm(formula = TransferValue ~ ., data = data100_forward)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68159 -0.16930 -0.03991  0.22069  0.79174
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.004106   0.092203   0.045  0.96506
## Age         -0.102238   0.124010  -0.824  0.42261
## time        -0.983816   0.254811  -3.861  0.00154 **
## goals        6.463452   2.689035   2.404  0.02961 *
## xG          -5.598029   3.124971  -1.791  0.09343 .
## assists     -0.589112   0.239951  -2.455  0.02677 *
```

```
## xA           1.397621   0.449778   3.107  0.00721 **
## shots        0.463812   0.288818   1.606  0.12914
## key_passes  -0.049434   0.279774  -0.177  0.86211
## npg         -5.054636   2.360871  -2.141  0.04910 *
## npxG         5.814058   2.937287   1.979  0.06643 .
## xGChain     -2.030005   0.890239  -2.280  0.03763 *
## xGBuildup    1.319783   0.462394   2.854  0.01206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4879 on 15 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8724, Adjusted R-squared:  0.7704
## F-statistic: 8.548 on 12 and 15 DF,  p-value: 0.0001101
```

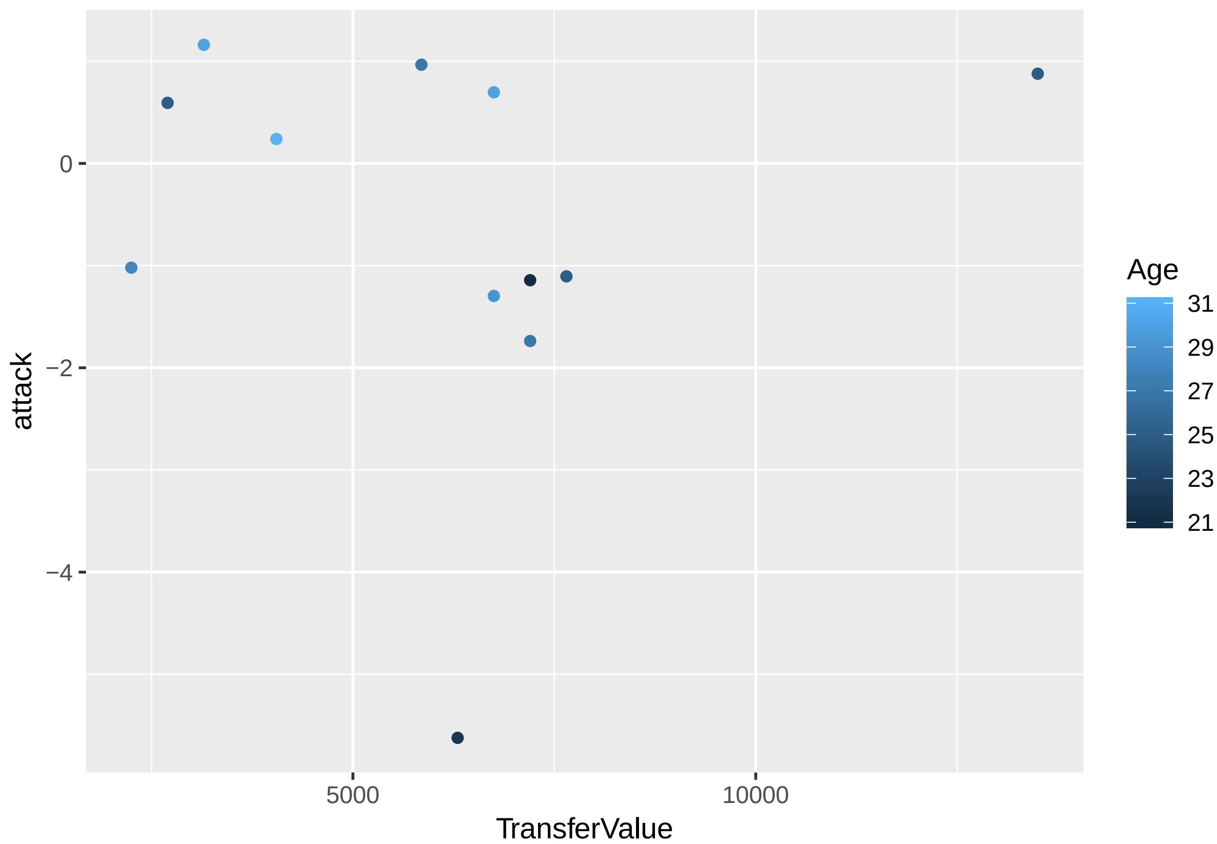
The R-square is high which means our model explain the data well, but when checking those coefs, we can see that it's strange that npg and xG are negative while goals and npxG are positive.

Since xG is the quality of chance received, maybe we should use the difference between xG and goals to evaluate the player's ability in attacking, and we should use per min data to show the player's efficiency.

Let's change the data and fit the model again.

```
##
## Call:
## lm(formula = TransferValue ~ ., data = data100_forward)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0598 -0.5385 -0.1959  0.3740  1.6634
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.004745   0.154567   0.031   0.9758
## attack            0.324266   0.179425   1.807   0.0858 .
## key_pass_per_min  0.053433   0.261654   0.204   0.8403
## shots_per_min     0.304598   0.231649   1.315   0.2034
## xGChain_per_min   0.393747   0.360207   1.093   0.2873
## xGBuildup_per_min -0.334526   0.326984  -1.023   0.3185
## xA_per_min        0.518761   0.302950   1.712   0.1023
## Age              -0.303893   0.170248  -1.785   0.0894 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8179 on 20 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.522,  Adjusted R-squared:  0.3547
## F-statistic:  3.12 on 7 and 20 DF,  p-value: 0.02144
```

Although the R square decrease a lot, the coefs of the model is more reasonable. Attack is xG minus goals, thus higher Attack means the player scored more goals than expected and that shows the ability of the player. By viewing the P-value we can conclude that the most important data for a forward's transfer value are attact and age. That fits our instinct.



We can see from the plot that player with high transfer values tend to have higher attack. What is interesting is that elder players are tend to have high attack but low transfer value.

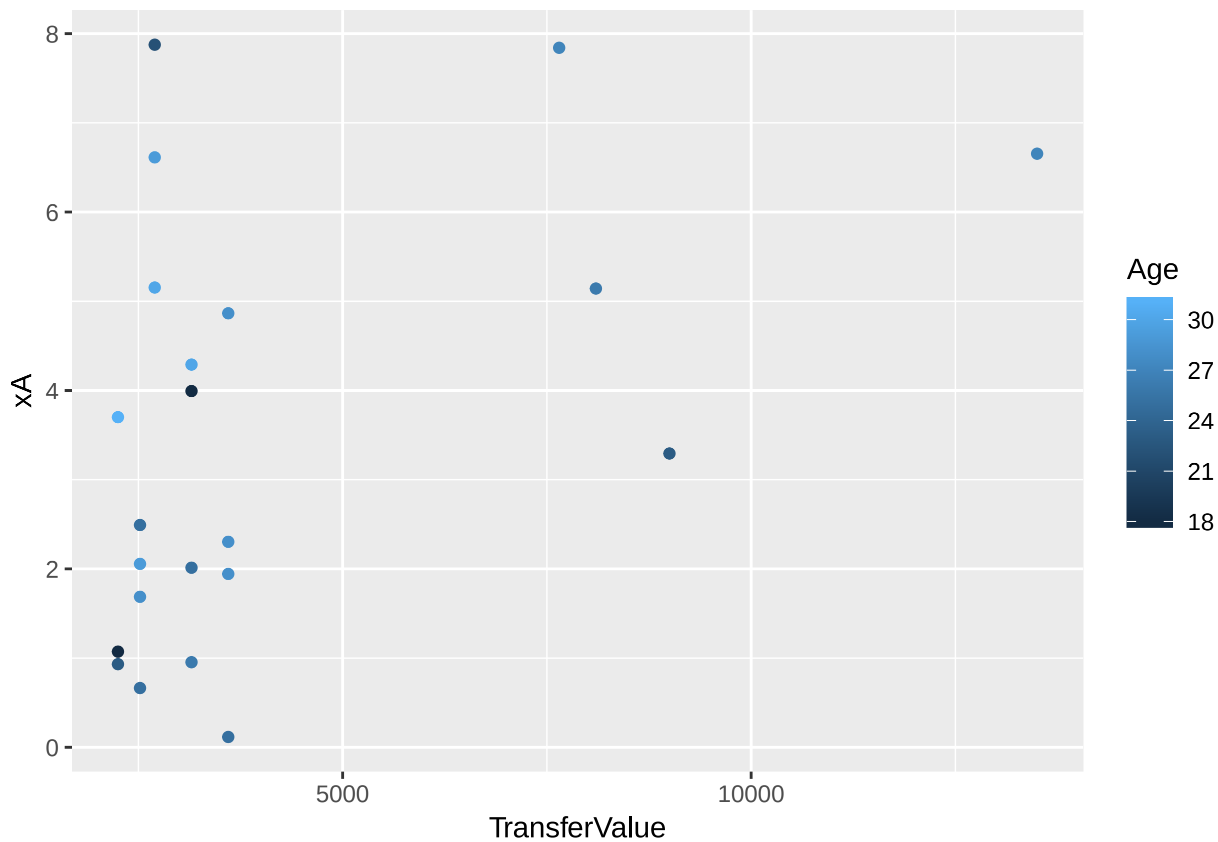
In fact, the average attack of Top100 forward player is 0.6375095 which is positive. That means "expensive" forward have the ablity to score more goal than expected.

Let's analyze middle field:

```
##
## Call:
## lm(formula = TransferValue ~ ., data = data100_midfield)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8422 -0.4968 -0.1580  0.4243  2.0412
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.05016    0.18253  -0.275   0.7875
## attack            -0.22275    0.23226  -0.959   0.3538
## key_pass_per_min  -0.50792    0.55031  -0.923   0.3717
## shots_per_min      0.45789    0.44086   1.039   0.3166
## xGChain_per_min   -5.87345    3.54187  -1.658   0.1195
## xGBuildup_per_min  5.02293    3.31164   1.517   0.1516
## xA_per_min         1.35583    0.47317   2.865   0.0125 *
## Age               -0.03689    0.24810  -0.149   0.8839
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8539 on 14 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5193, Adjusted R-squared:  0.279
## F-statistic: 2.161 on 7 and 14 DF,  p-value: 0.1042
```

We can see the coefs are changed a lot, the importanance of xA and xGBulidup stand out as the key factor. it's reasonable because midfields need do more with passes and assist, sometimes defense.



We can see from the plot that player with high transfer values tend to have higher xA, which fits the results of regression.

In conclusion, we can see that xG and xA can reflect the performance of a player better compared with basic data like goals and assists. However, age is also a very important part when discuss a player's transfer values.

Obviously, there are many other variables that may affect player's transfer values like nationality, club, height, commercial value and so on. But I believe the usage of expected data is a big step for football analyze.

The url of github is: https://github.com/redLeo-D/project_stat597