Introduction

The aim of this project is to see if there's a relationship between a player's popularity and his market value, given the difficult nature of using summary statistics for this task.

Some Preliminary Analysis

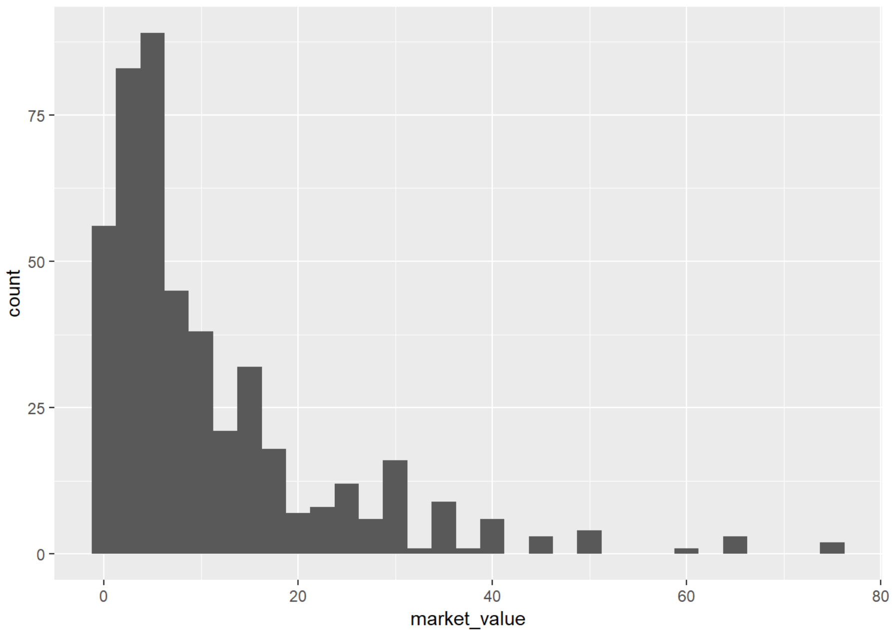
Who are the most valuable players in the EPL?

name	club	age	position	position_cat	market_value	page_views	fpl_value	fpl_sel	fpl_points	region	nationality	new_forei
Eden Hazard	Chelsea	26	LW	1	75	4220	10.5	2.30%	224	2	Belgium	
Paul Pogba	Manchester+United	24	CM	2	75	7435	8.0	19.50%	115	2	France	
Alexis Sanchez	Arsenal	28	LW	1	65	4329	12.0	17.10%	264	3	Chile	
Kevin De Bruyne	Manchester+City	26	AM	1	65	2252	10.0	17.50%	199	2	Belgium	
Sergio Aguero	Manchester+City	29	CF	1	65	4046	11.5	9.70%	175	3	Argentina	
Harry Kane	Tottenham	23	CF	1	60	4161	12.5	35.10%	224	1	England	

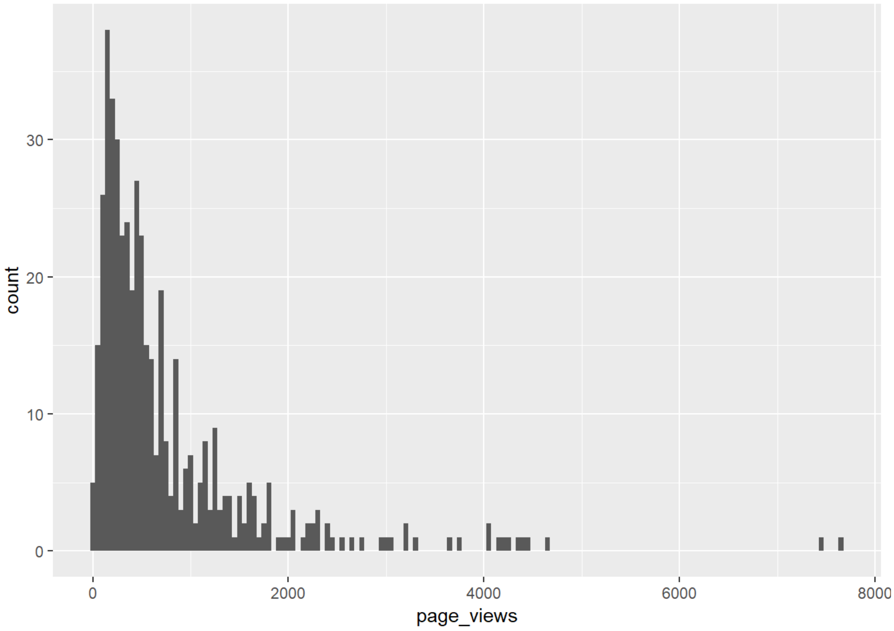
Who are the most popular players?

name	club	age	position	position_cat	market_value	page_views	fpl_value	fpl_sel	fpl_points	region	nationality	new_forei
Wayne Rooney	Everton	31	SS	1	15	7664	7.5	20.90%	76	1	England	
Paul Pogba	Manchester+United	24	CM	2	75	7435	8.0	19.50%	115	2	France	
Dele Alli	Tottenham	21	CM	2	45	4626	9.5	38.60%	225	1	England	
Diego Costa	Chelsea	28	CF	1	50	4454	10.0	3.00%	196	2	Spain	
Mesut Ozil	Arsenal	28	AM	1	50	4395	9.5	5.60%	167	2	Germany	
Alexis Sanchez	Arsenal	28	LW	1	65	4329	12.0	17.10%	264	3	Chile	

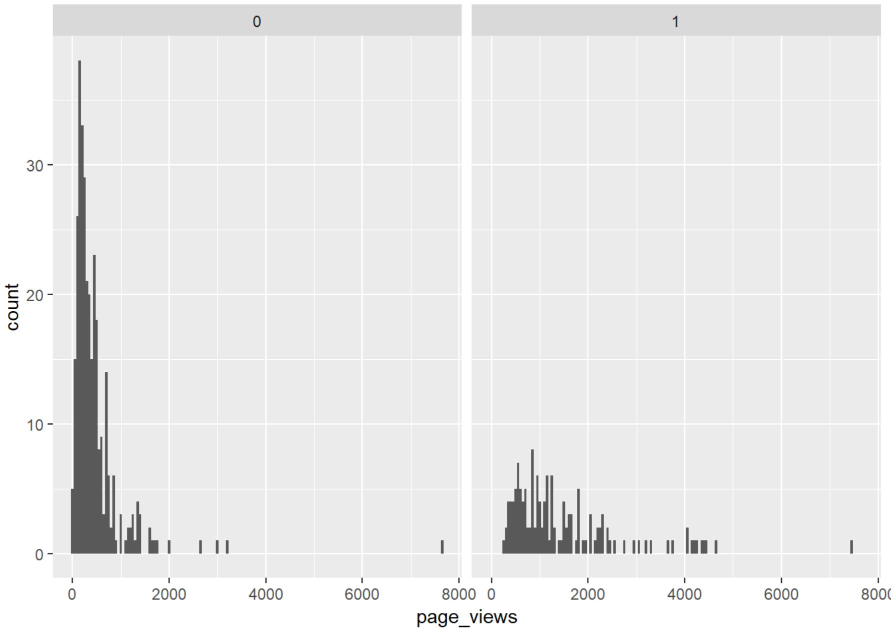
Distribution of Market Value



Clearly not a normal distribution, but this was expected. Teams tend to have few elite players, and a large number of low + mid value players in their *squads*. ### Distribution of popularity



Top 6 vs the rest

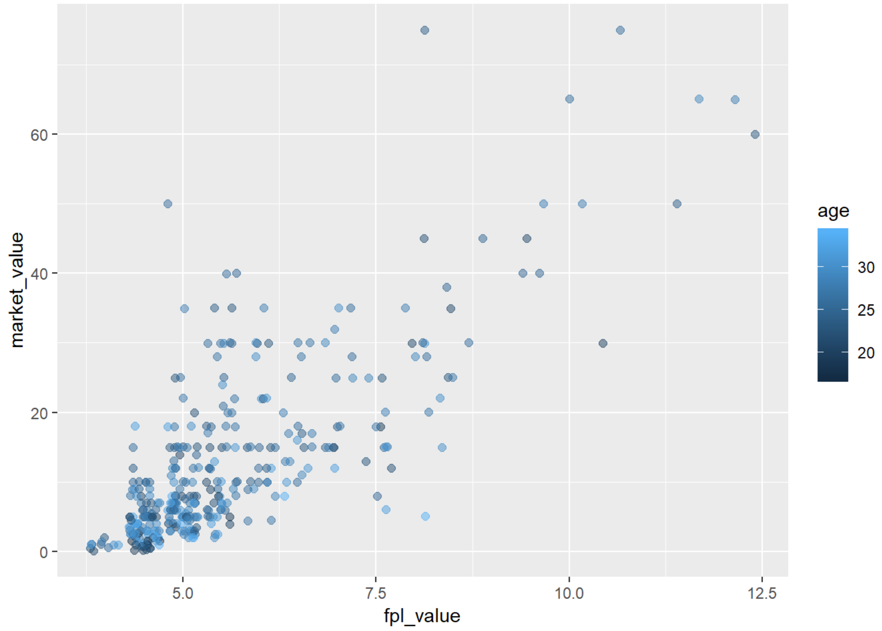


The top 6 clubs seem to have a spread of players popularity. Also, Wayne Rooney is at Everton now.

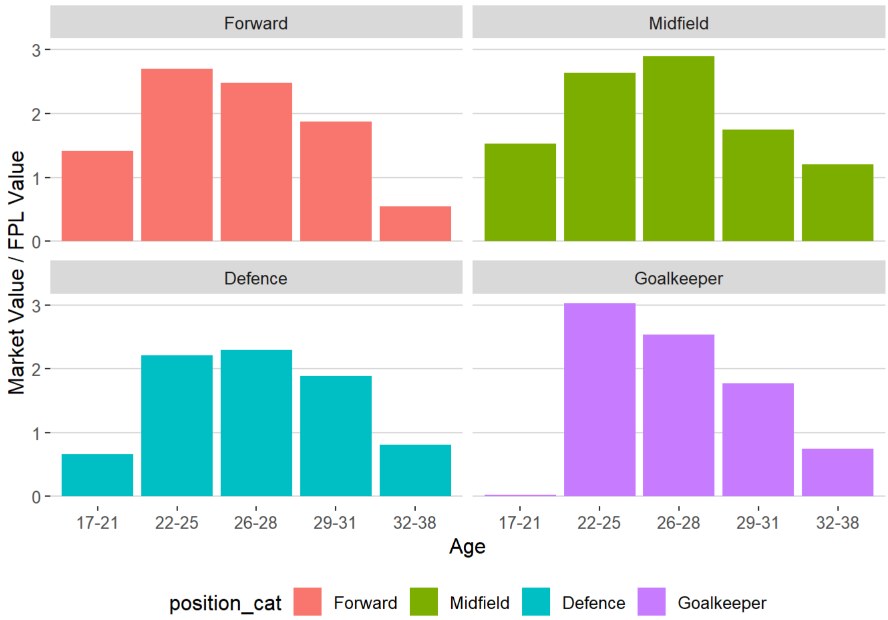
Detailed Analysis

There seems to be evidence of a player's market value being correlated with how popular he is. This is interesting because *ability* and *performance* are notoriously difficult to quantify in football. It varies with the position, the manager's tactics, the opposition, the league, the ability of your own teammates, and so on. Consequently, valuing a player is very hard to do, though it has to be done anyway.
Websites like WhoScored have a score for each player for each match, and Fantasy Premier League places a value on each player's head. It would be interesting to see if *popularity* can be used as a basic proxy for *ability*

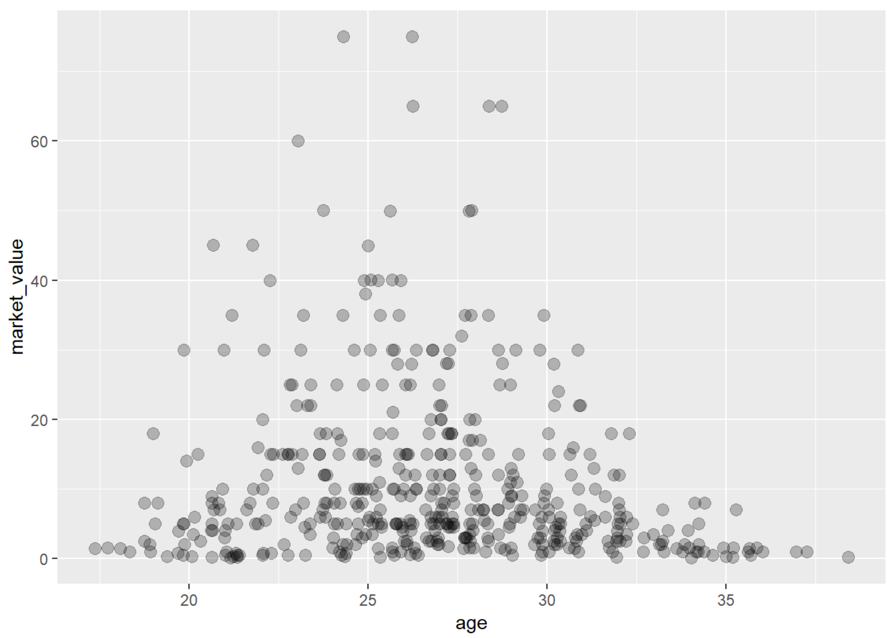
FPL Valuation



There seems to be nice agreement between the FPL value and transfermrkt value, despite the fact that FPL valuation is decidedly shorter term, so age would be less of a factor

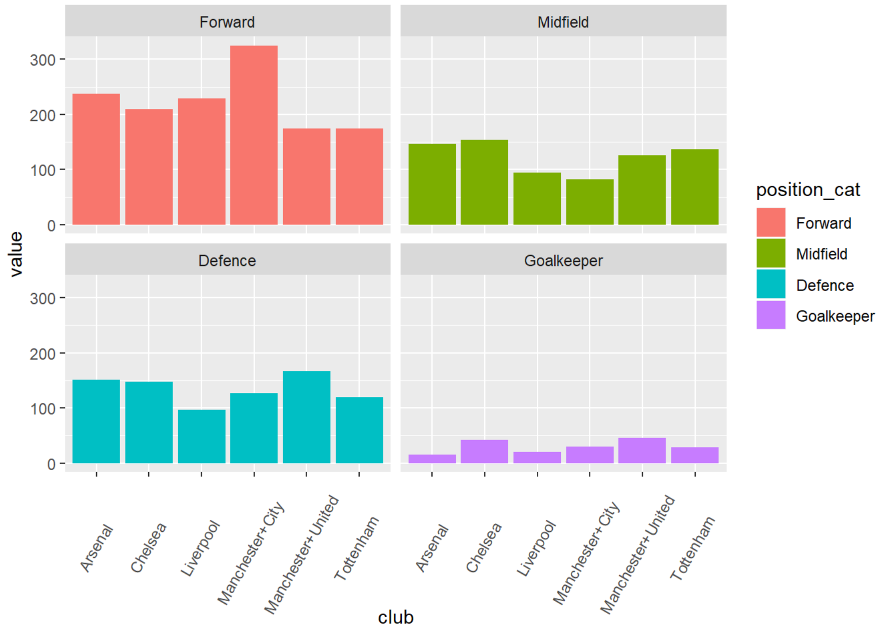


Market Value with Age



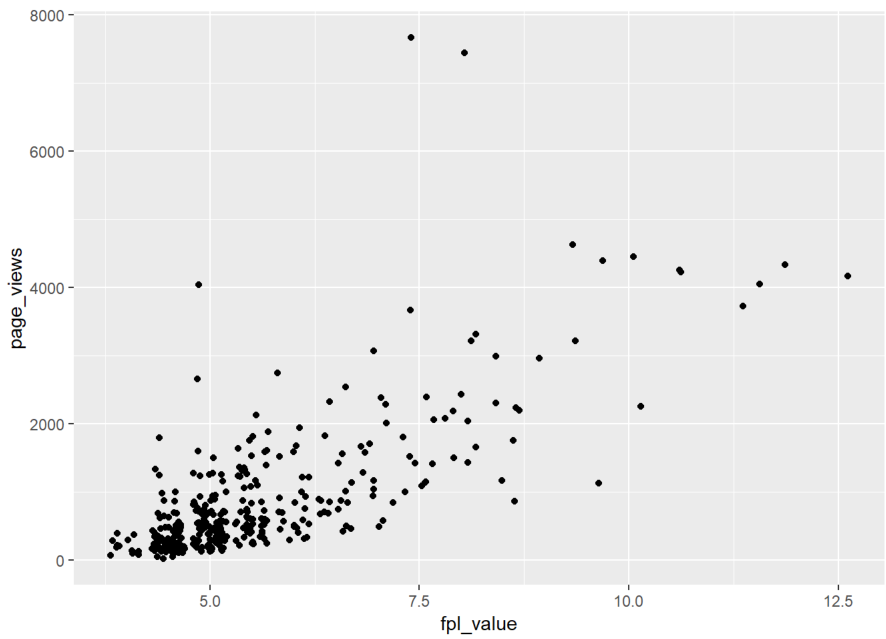
The high value players are clustered around the age of 24-32, peaking at about 27. It’s important to note that this is in no way a linear relationship, which is why I use age categories in the regression model that follows.

Who’s stocking up at which position?



Popularity as a proxy for Ability

Ability is difficult to measure and compare through performance indicators. Assuming **FPL valuation** is a fair measure of ability. While this may not be perfect, we should still be able to se a relationship between ability and popularity.



There seems to be a nice, linear relationship between FPL valuation and popularity, with a few notable exceptions. ## Regression Model

The main aim is to see whether market value can be determined using popularity as a proxy for ability. A player’s market value can intuitively be represented as -

```
market value ~ ability + position + age
```

In the model, I control for 1-4, but not for 5 and 6. Both 5 and 6 would require extensive work identifying breakouts and long-term injuries, which might be useful future additions to the model.

For factors 1 - 4:

1. Retrieved the nationality of each player, and put them into 4 buckets:

- 1 for England
- 2 for EU (Brexit made this a natural classification)
- 3 for Americas
- 4 for Rest of World

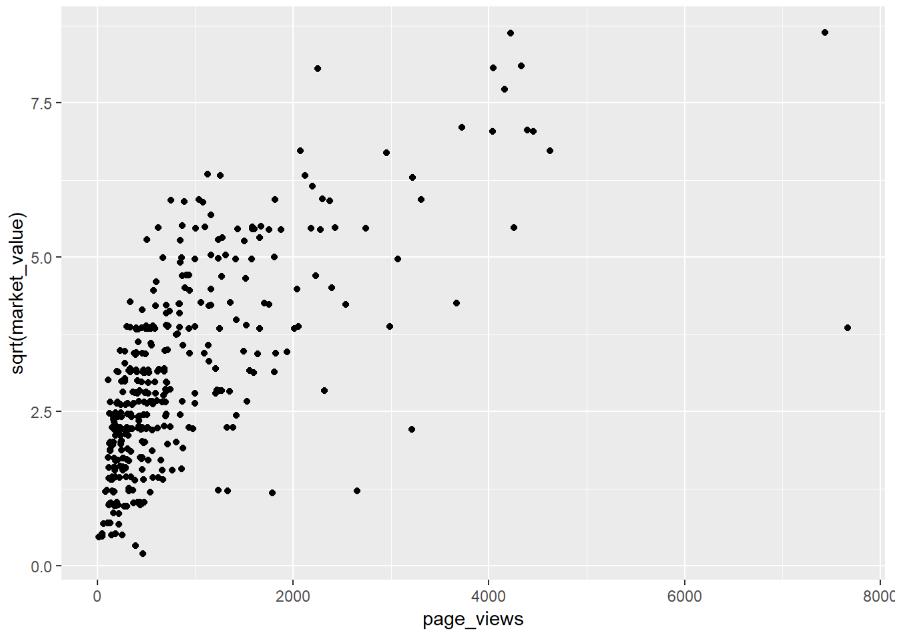
A new column called `region` was made, as a factor with 4 levels.

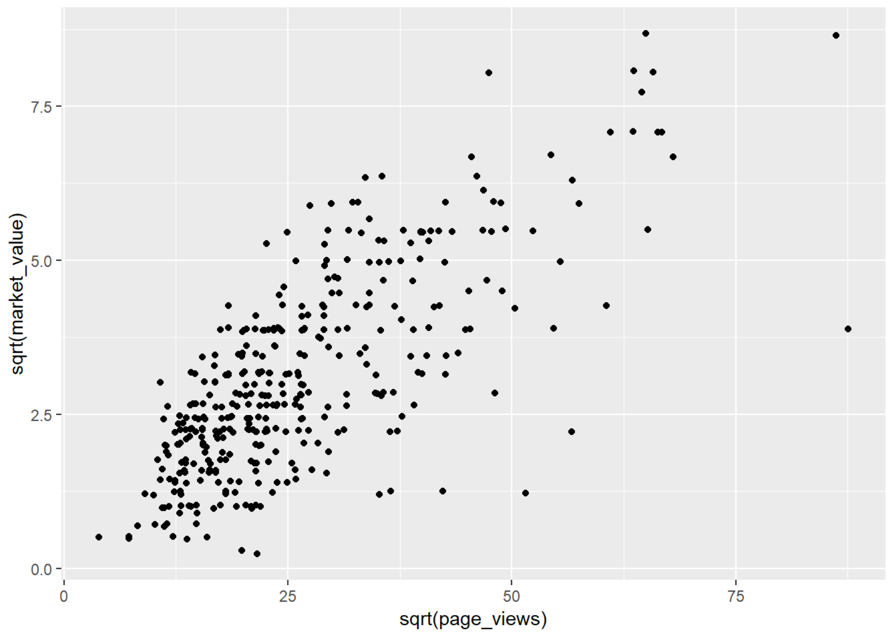
2. Included an interaction term for page views and position category.
3. Marked the new signings of 2016/17, and interacted that with page views.
4. A column `big_club` was created comprising of United, City, Chelsea, Arsenal, Liverpool and Tottenham. This was interacted with page views as well.

Apart from these interactions, age is also included as a categorical variable (due to its non-linear relationship with market value).

Dataset Modifications

1. sqrt values of `market_value` are taken, because `market_value` is right-tail heavy, which could lead to heteroscedasticity.
2. However, this leads to the relationship between `sqrt(market_value)` and `page_views` looking like this -





This looks roughly linear.

Now applying a multiple linear regression model yields the following R^2 value -

Call: lm(formula = sqrt(market_value) ~ page_views + age_category:position_cat + page_views:region + page_views:big_club + new_signing:page_views, data = df1)

Residuals: Min 1Q Median 3Q Max -2.34847 -0.55618 -0.01892 0.58945 2.24128

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.130222 0.267112 0.488 0.626199
page_views 0.054551 0.006235 8.750 < 2e-16 ***age_category17-21:position_cat1 0.256266 0.318218 0.805 0.421187***
age_category22-25:position_cat1 1.576169 0.298532 5.280 2.28e-07 age_category26-28:position_cat1 1.542115 0.288898 5.338 1.70e-07
age_category29-31:position_cat1 1.028739 0.309253 3.327 0.000973 age_category32-38:position_cat1 -0.845027 0.481267 -1.756 0.079996 .
age_category17-21:position_cat2 0.644210 0.366759 1.756 0.079884 .
age_category22-25:position_cat2 1.270794 0.315609 4.026 6.95e-05 ***age_category26-28:position_cat2 1.560059 0.294349 5.300 2.06e-07***
age_category29-31:position_cat2 0.886903 0.337567 2.627 0.008986 ** age_category32-38:position_cat2 0.231237 0.343605 0.673 0.501411

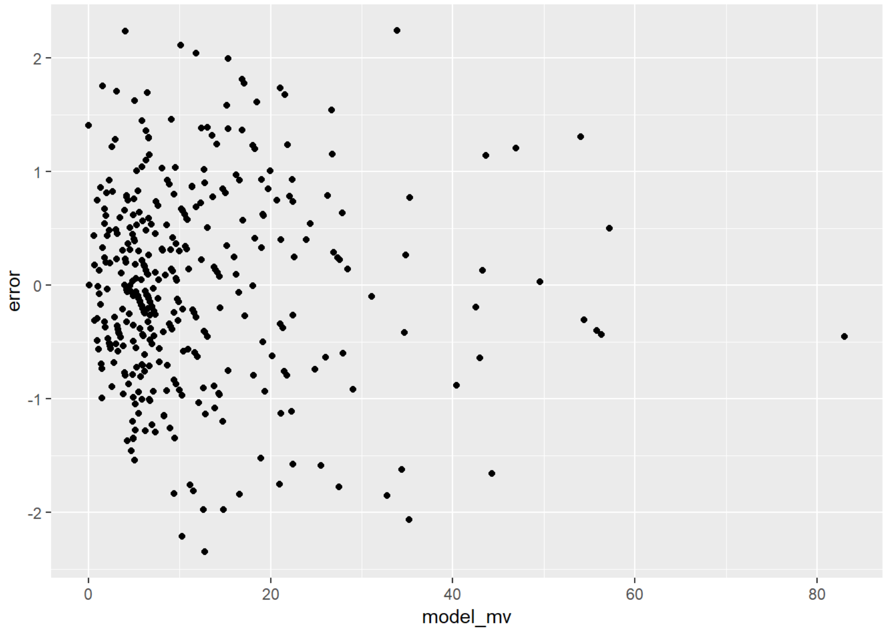
age_category17-21:position_cat3 0.282958 0.366732 0.772 0.440895
age_category22-25:position_cat3 1.229583 0.299980 4.099 5.17e-05 ***age_category26-28:position_cat3 1.571121 0.284149 5.529 6.32e-08***
age_category29-31:position_cat3 1.071693 0.303386 3.532 0.000467 ***age_category32-38:position_cat3 0.416735 0.323600 1.288 0.198668***
age_category17-21:position_cat4 -1.551477 0.921304 -1.684 0.093078 .
age_category22-25:position_cat4 1.352890 0.467968 2.891 0.004081 *age_category26-28:position_cat4 1.010454 0.508002 1.989 0.047476*
age_category29-31:position_cat4 0.842614 0.356848 2.361 0.018763 *
age_category32-38:position_cat4 NA NA NA NA
page_views:region2 0.011488 0.003962 2.899 0.003978 ** page_views:region3 0.013590 0.005710 2.380 0.017852 *
page_views:region4 0.006894 0.005604 1.230 0.219403
page_views:big_club 0.021383 0.004151 5.152 4.33e-07 *** page_views:new_signing 0.002002 0.004169 0.480 0.631360
— Signif. codes: 0 ‘’ ***0.001 ’’*** *0.01* ’ 0.05 ‘.’ 0.1 ‘’ 1

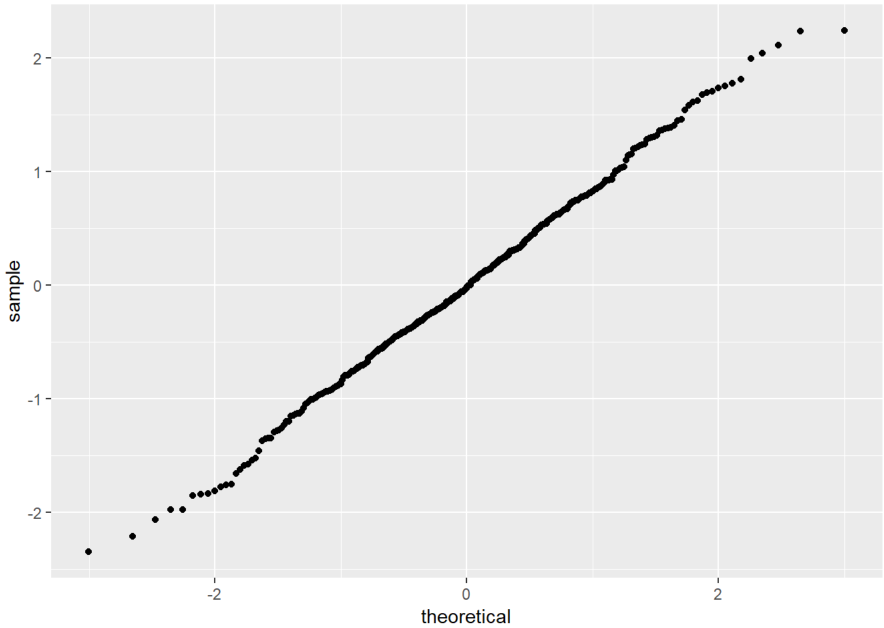
Residual standard error: 0.8852 on 348 degrees of freedom Multiple R-squared: 0.7218, Adjusted R-squared: 0.7018 F-statistic: 36.12 on 25 and 348 DF, p-value: < 2.2e-16

R^2 of over 70% ! Further, the coefficient of `page_views` is extremely significant. Clearly, there is a linear relationship between `sqrt(market_value)` and `sqrt(page_views)` .

What can residual plots tell us?

The residual plots should be able to tell us whether we have a heteroscedasticity problem in our data.





The residual plot seems to have randomly distributed errors, and the qq plot confirms that they are normally distributed.

EPL Popularity

An interesting by-product is to see how popular the Premier League is, compared to other leagues. Due to the small number of inward-transfers from foreign leagues, this remains a rough method. However, the differences are large enough to be greater than just noise.

name	market_value	predicted_mv
Sead Kolasinac	15.0	12.5

name	market_value	predicted_mv
Alexandre Lacazette	40.0	21.9
Antonio Rudiger	25.0	10.4
Tiemoue Bakayoko	16.0	17.5
Davy Klaassen	18.0	9.4
Sandro Ramirez	10.0	8.9
Vicente Iborra	9.0	4.3
Mohamed Salah	35.0	20.0
Ederson Moraes	22.0	9.2
Bernardo Silva	40.0	21.2
Victor Lindelof	22.0	16.4
Jan Bednarek	0.5	0.4
Roque Mesa	12.0	5.7
Kiko Femenia	4.0	5.2
Will Hughes	8.0	5.2
Ahmed Hegazy	1.0	5.1

The model works because it has *generally undervalued* players from other leagues. The reasoning is thus - a 20 million player in the EPL gets more hits than a 20 million player in Ligue 1. Because of this, the *value* of **each** page view is far lower in the EPL. But since the model is built using EPL data, the coefficient of page views is derived from EPL. Consequently, foreign players from less popular leagues are undervalued.