Football Player Market Value Analysis

Samarth Gwalani

Dec 10, 2021

Table Of Contents

Introduction

With the increase in popularity of Soccer globally there is an in pour of money chasing the next best players. Many a time players from a certain nationality, age and position command inflated market prices due to players with a similar profile having great success in the past. With the help of data clubs can avoid overpaying for such players and understand correlations of these factors in order to get a better understanding of the transfer market. Data Analysis would help quantify the problem and back up decision making with concrete evidence.

I am trying to see if there is a correlation between soccer athletes salary and market value to factors such as playing position, nationality and age. Researchers in the past too have aimed to gain a deeper understanding by comparing skill and market value. While this is a great metric evaluating such a relation can be questionable as agreeing on a players skill as a number could be subjective. I believe hype and expectations for players of a certain nationality and position impact the valuation of players since past performances of players from a certain region inflate salaries. The study would provide insight into a valuation benchmark model that can help identify undervalued players in the market for possible transfers and also give clubs an understanding of fair player wage compensation.

Exploring the Data

Our dataset consists of the following variables:

Dependent Variables:

- Market Value

Independent Variables:

- Player's Age
- Overall Skill Level
- Nationality
- Position
- Preferred Foot

Snippet of dataset being used for Analysis:

ID	Name	FullName	Age	Height	Weight	PhotoUrl	Nationality	Overall	Potential	Growth	TotalStats	BaseStats	Positions	BestPosition	Club	ValueEUR	WageEUR	ReleaseClause	ClubPosition
158023	L. Messi	Lionel Messi	34	170	72	https://cdn.sofifa.com/players/158/023/22_60.png	Argentina	93	93	0	2219	462	RW,ST,CF	RW	Paris Saint-Germain	78000000	320000	144300000	RW
188545	R. Lewandowski	Robert Lewandowski	32	185	81	https://cdn.sofifa.com/players/188/545/22_60.png	Poland	92	92	0	2212	460	ST	ST	FC Bayern München	119500000	270000	197200000	ST
20801	Cristiano Ronaldo	C. Ronaldo dos Santos Aveiro	36	187	83	https://cdn.sofifa.com/players/020/801/22_60.png	Portugal	91	91	0	2208	457	ST,LW	ST	Manchester United	45000000	270000	83300000	ST
231747	K. Mbappé	Kylian Mbappé	22	182	73	https://cdn.sofifa.com/players/231/747/22_60.png	France	91	95	4	2175	470	ST,LW	ST	Paris Saint-Germain	194000000	230000	373500000	ST
200389	J. Oblak	Jan Oblak	28	188	87	https://cdn.sofifa.com/players/200/389/22_60.png	Slovenia	91	93	2	1413	489	GK	GK	Atlético de Madrid	112000000	130000	238000000	GK
192985	K. De Bruyne	Kevin De Bruyne	30	181	70	https://cdn.sofifa.com/players/192/985/22_60.png	Belgium	91	91	0	2304	485	CM,CAM	CM	Manchester City	125500000	350000	232200000	CM
190871	Neymar Jr	Neymar da Silva Santos Jr.	29	175	68	https://cdn.sofifa.com/players/190/871/22_60.png	Brazil	91	91	0	2183	454	LW,CAM	LW	Paris Saint-Germain	129000000	270000	238700000	LW
215914	N. Kanté	N'Golo Kanté	30	168	70	https://cdn.sofifa.com/players/215/914/22_60.png	France	90	90	0	2179	470	CDM,CM	CDM	Chelsea	100000000	230000	185000000	CM
202126	H. Kane	Harry Kane	27	188	89	https://cdn.sofifa.com/players/202/126/22_60.png	England	90	90	0	2205	456	ST	ST	Tottenham Hotspur	129500000	240000	246100000	ST
192448	M. ter Stegen	Marc-André ter Stegen	29	187	85	https://cdn.sofifa.com/players/192/448/22_60.png	Germany	90	92	2	1444	484	GK	GK	FC Barcelona	99000000	250000	210400000	GK
167495	M. Neuer	Manuel Neuer	35	193	93	https://cdn.sofifa.com/players/167/495/22_60.png	Germany	90	90	0	1534	501	GK	GK	FC Bayern München	13500000	86000	22300000	GK
200104	H. Son	Heung Min Son	28	183	78	https://cdn.sofifa.com/players/200/104/22_60.png	Korea Republic	89	89	0	2142	455	LM,CF,LW	LM	Tottenham Hotspur	104000000	220000	197600000	LW
200145	Casemiro	Carlos Henrique Venancio Casimiro	29	185	84	https://cdn.sofifa.com/players/200/145/22_60.png	Brazil	89	89	0	2219	462	CDM	CDM	Real Madrid CF	88000000	310000	180400000	CDM
165153	K. Benzema	Karim Benzema	33	185	81	https://cdn.sofifa.com/players/165/153/22_60.png	France	89	89	0	2116	446	CF,ST	CF	Real Madrid CF	66000000	350000	135300000	CF
192119	T. Courtois	Thibaut Courtois	29	199	96	https://cdn.sofifa.com/players/192/119/22_60.png	Belgium	89	91	2	1327	469	GK	GK	Real Madrid CF	85500000	250000	181700000	GK
208722	S. Mané	Sadio Mané	29	175	69	https://cdn.sofifa.com/players/208/722/22_60.png	Senegal	89	89	0	2192	465	LW	LW	Liverpool	101000000	270000	186900000	LW
203376	V. van Dijk	Virgil van Dijk	29	193	92	https://cdn.sofifa.com/players/203/376/22_60.png	Netherlands	89	89	0	2104	455	CB	CB	Liverpool	86000000	230000	159100000	CB
230621	G. Donnarumma	Gianluigi Donnarumma	22	196	90	https://cdn.sofifa.com/players/230/621/22_60.png	Italy	89	93	4	1377	481	GK	GK	Paris Saint-Germain	119500000	110000	230000000	GK
209331	M. Salah	Mohamed Salah	29	175	71	https://cdn.sofifa.com/players/209/331/22_60.png	Egypt	89	89	0	2211	468	RW	RW	Liverpool	101000000	270000	186900000	RW
210257	Ederson	Ederson Santana de Moraes	27	188	86	https://cdn.sofifa.com/players/210/257/22_60.png	Brazil	89	91	2	1583	501	GK	GK	Manchester City	94000000	200000	181000000	GK
212622	J. Kimmich	Joshua Kimmich	26	177	75	https://cdn.sofifa.com/players/212/622/22_60.png	Germany	89	90	1	2283	475	CDM,RB	CDM	FC Bayern München	108000000	160000	186300000	CDM
212831	Alisson	Alisson Ramses Becker	28	191	91	https://cdn.sofifa.com/players/212/831/22_60.png	Brazil	89	90	1	1393	486	GK	GK	Liverpool	82000000	190000	157900000	GK
192505	R. Lukaku	Romelu Lukaku	28	191	94	https://cdn.sofifa.com/players/192/505/22_60.png	Belgium	88	88	0	2064	445	ST	ST	Chelsea	93500000	260000	173000000	ST
155862	Sergio Ramos	Sergio Ramos García	35	184	82	https://cdn.sofifa.com/players/155/862/22_60.png	Spain	88	88	0	2251	461	CB	CB	Paris Saint-Germain	24000000	115000	44400000	CB
176580	L. Suárez	Luis Suárez	34	182	83	https://cdn.sofifa.com/players/176/580/22_60.png	Uruguay	88	88	0	2307	457	ST	ST	Atlético de Madrid	44500000	135000	91200000	ST
182521	T. Kroos	Toni Kroos	31	183	76	https://cdn.sofifa.com/players/182/521/22_60.png	Germany	88	88	0	2148	445	CM	CM	Real Madrid CF	75000000	310000	153800000	CM
239085	E. Haaland	Erling Haaland	20	194	94	https://cdn.sofifa.com/players/239/085/22_60.png	Norway	88	93	5	2102	458	ST	ST	Borussia Dortmund	137500000	110000	244100000	ST
193041	K. Navas	Keylor Navas	34	185	80	https://cdn.sofifa.com/players/193/041/22_60.png	Costa Rica	88	88	0	1428	477	GK	GK	Paris Saint-Germain	15500000	130000	28700000	SUB
212198	Bruno Fernandes	Bruno Miguel Borges Fernandes	26	179	69	https://cdn.sofifa.com/players/212/198/22_60.png	Portugal	88	89	1	2341	481	CAM	CAM	Manchester United	107500000	250000	206900000	CAM
202652	R. Sterling	Raheem Sterling	26	170	69	https://cdn.sofifa.com/players/202/652/22_60.png	England	88	89	1	2113	451	LW,RW	LW	Manchester City	107500000	290000	206900000	SUB
199556	M. Verratti	Marco Verratti	28	165	60	https://cdn.sofifa.com/players/199/556/22_60.png	Italy	87	87	0	2202	449	CM,CAM	CM	Paris Saint-Germain	79500000	155000	147100000	CM
177003	L. Modrić	Luka Modrić	35	172	66	https://cdn.sofifa.com/players/177/003/22_60.png	Croatia	87	87	0	2253	464	CM	CM	Real Madrid CF	32000000	190000	65599999	CM
183898	A. Di María	Ángel Di María	33	180	69	https://cdn.sofifa.com/players/183/898/22_60.png	Argentina	87	87	0	2177	455	RW,LW	RW	Paris Saint-Germain	49500000	160000	91600000	SUB
186153	W. Szczęsny	Wojciech Szczęsny	31	195	90	https://cdn.sofifa.com/players/186/153/22_60.png	Poland	87	87	0	1317	465	GK	GK	Juventus	42000000	105000	69300000	GK
189596	T. Müller	Thomas Müller	31	185	76	https://cdn.sofifa.com/players/189/596/22_60.png	Germany	87	87	0	2136	440	CAM,RM,RW	CAM	FC Bayern München	66000000	140000	108900000	CAM
192387	C. Immobile	Ciro Immobile	31	185	85	https://cdn.sofifa.com/players/192/387/22_60.png	Italy	87	87	0	2065	437	ST	ST	Lazio	67500000	125000	114800000	ST
195864	P. Pogba	Paul Pogba	28	191	84	https://cdn.sofifa.com/players/195/864/22_60.png	France	87	87	0	2222	472	CM,LM	CM	Manchester United	79500000	220000	147100000	CDM
209658	L. Goretzka	Leon Goretzka	26	189	82	https://cdn.sofifa.com/players/209/658/22_60.png	Germany	87	88	1	2314	496	CM,CDM	CM	FC Bayern München	93000000	140000	160400000	CDM
207865	Marquinhos	Marcos Aoás Corrêa	27	183	75	https://cdn.sofifa.com/players/207/865/22_60.png	Brazil	87	90	3	2074	452	CB,CDM	CB	Paris Saint-Germain	90500000	135000	174200000	CB
211110	P. Dybala	Paulo Dybala	27	177	75	https://cdn.sofifa.com/players/211/110/22_60.png	Argentina	87	88	1	2134	447	CF,CAM	CAM	Juventus	93000000	160000	160400000	CAM
216267	A. Robertson	Andrew Robertson	27	178	64	https://cdn.sofifa.com/players/216/267/22_60.png	Scotland	87	88	1	2163	465	LB	LB	Liverpool	83500000	175000	160700000	LB
228702	F. de Jong	Frenkie de Jong	24	180	74	https://cdn.sofifa.com/players/228/702/22_60.png	Netherlands	87	92	5	2229	478	CM,CDM,CB	CM	FC Barcelona	119500000	210000	253900000	CM
231281	T. Alexander-Arnold	Trent Alexander-Arnold	22	180	69	https://cdn.sofifa.com/players/231/281/22_60.png	England	87	92	5	2227	467	RB	RB	Liverpool	114000000	150000	219500000	RB
239818	Rúben Dias	Rúben Santos Gato Alves Dias	24	187	82	https://cdn.sofifa.com/players/239/818/22_60.png	Portugal	87	91	4	1886	407	CB	CB	Manchester City	102500000	170000	197300000	CB
153079	S. Agüero	Sergio Agüero	33	173	70	https://cdn.sofifa.com/players/153/079/22_60.png	Argentina	87	87	0	2068	424	ST	ST	FC Barcelona	51000000	260000	104600000	ST
167948	H. Lloris	Hugo Lloris	34	188	82	https://cdn.sofifa.com/players/167/948/22_60.png	France	87	87	0	1372	472	GK	GK	Tottenham Hotspur	13500000	125000	25700000	GK
233049	J. Sancho	Jadon Sancho	21	180	76	https://cdn.sofifa.com/players/233/049/22_60.png	England	87	91	4	2007	431	RM,CF,LM	CAM	Manchester United	116500000	150000	224300000	LM
201024	K. Koulibaly	Kalidou Koulibaly	30	187	89	https://cdn.sofifa.com/players/201/024/22_60.png	Senegal	86	86	0	1705	397	CB	CB	Napoli	55500000	105000	94400000	CB
232363	M. Škriniar	Milan Škriniar	26	188	80	https://cdn.sofifa.com/players/232/363/22_60.png	Slovakia	86	88	2	1829	413	CB	CB	Inter	74000000	150000	131400000	CB

Figure 1.1 - Snippet of dataset

Hypothesis

Our research surrounded analysing the best predictors for the market value in order to find the right model fit by running a linear regression analysis.

Null Hypothesis (H0)

All the independent variables have the same amount of influence on the market value.

Alternate Hypothesis (Ha)

All the independent variables have a different amount of influence on the market value.

Data Screening

Mahalanobis

We checked for any outliers with estimated mahalanobis distance. Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point and a distribution and therefore is an important step in our data screening process. We used the chi square distribution to calculate the cutoff score for our data and to then get rid of any outliers that might affect the data and regression analysis. We set our cut off at 99.9% The cutoff score was 45.3. After reviewing the mahalanobis score with the cutoff score, we came back with the output that we had no outliers.

Linear Regression Analysis

```
Call:
lm(formula = ValueEUR ~ Overall, data = dataset)

Residuals:
      Min        1Q    Median        3Q       Max 
-13554518  -2744895  -1072419   1180341 175465671 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -38111026     441350  -86.35   <2e-16 ***
Overall        622476       6670   93.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6311000 on 19258 degrees of freedom
Multiple R-squared:  0.3114,	Adjusted R-squared:  0.3114 
F-statistic:  8709 on 1 and 19258 DF,  p-value: < 2.2e-16
```

Figure 1.2 - Output of Linear Regression Model for ValueEUR and Overall

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable[1]. For the first model our explanatory variable is the Overall and our dependent variable was the ValueEUR. For this model, our t-statistic was 93.32, meaning that the ValueEUR was 93.32 standard errors from zero. The p-value is calculated using the t-statistic from the T distribution. The p-value, in association with the t-statistic, helps us to understand how significant our coefficient is to the model. In practice, any p-value below 0.05 is usually deemed as significant. When we say our model is significant, it means that we are confident that the coefficient is not zero, meaning the coefficient does in fact add value to the

[1] Thieme, C. (2021, June 16). *Understanding linear regression output in R*. Medium. Retrieved December 5, 2021, from https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3.

model by helping to explain the variance within our dependent variable[2]. In our first model of Overall and ValueEUR our p value < 0.00000000000000022.

The Multiple R-squared value tells us what percentage of the variation within our dependent variable that the independent variable is explaining. In other words, it's another method to determine how well our model is fitting the data[3]. In our first model our multiple R-squared value was 0.3114 which means Overall explains 31% of the variation with ValueEUR (our dependent variable).

The null hypothesis is that there is no relationship between the dependent variable and the independent variables and the alternative hypothesis is that there is a relationship. The alternative hypothesis is that at least one of them is not zero. The F-statistic and overall p-value help us determine the result of this test. $F(1, 19258) = 8709$, $p < 0.00000002$. Our F-statistic is 8709 which for a smaller model like ours signifies that the null hypothesis should be rejected. Additionally, our p value being way lower than 0.05 also indicates that our coefficient in our model is significant. Therefore, we can say that Overall is a good indicator for ValueEUR.

---

[2] Thieme, C. (2021, June 16). *Understanding linear regression output in R*. Medium. Retrieved December 5, 2021, from https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3.

[3] Thieme, C. (2021, June 16). *Understanding linear regression output in R*. Medium. Retrieved December 5, 2021, from https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3.

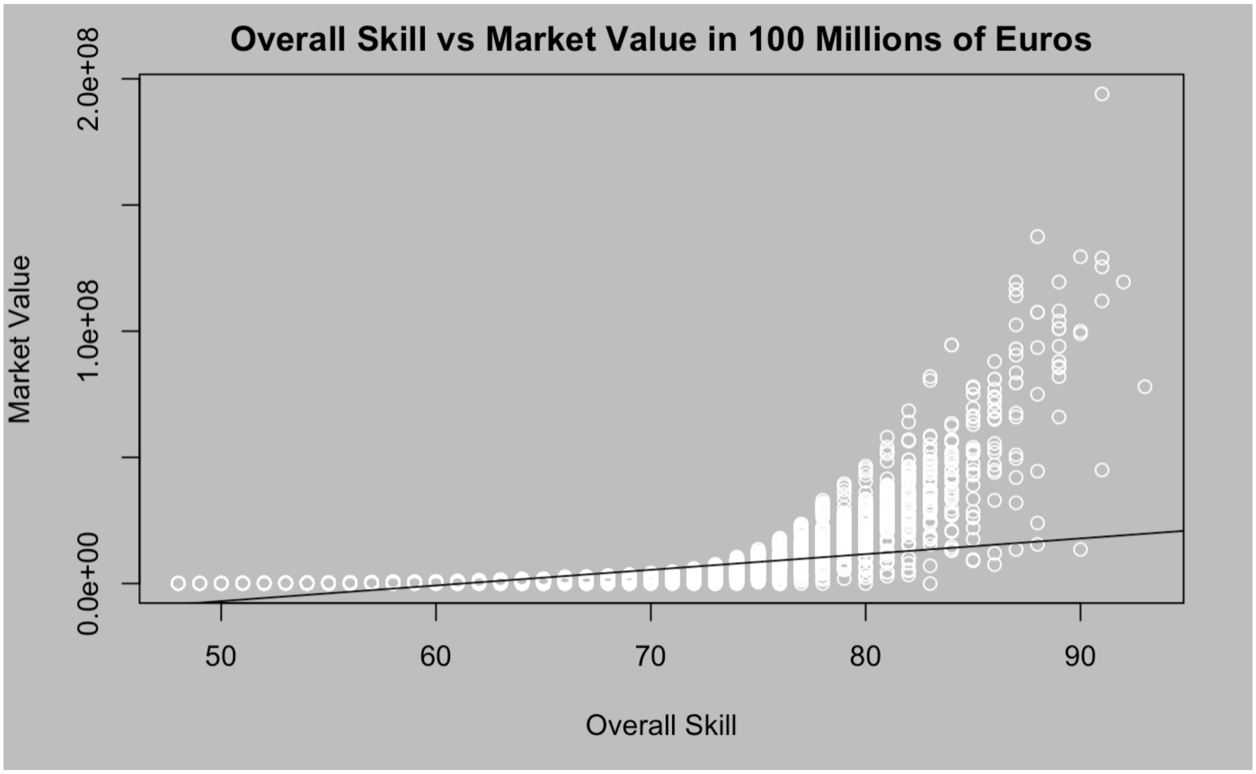


Figure 1.3 - Scatter plot of ValueEUR and Overall

Interpreting R-squared Results

The Linear regression best fit model is obtained by using the adjusted R squared method. The *lm function* is used to run a linear model. We ran the same analysis for all the independent variables to see how significantly they impacted the ValueEUR.

The highest value obtained was 68% for WageEUR. While the lowest value obtained was 0% for Height. The result is shown below.

```{r}
#Checking the highest influence for ValueEUR
summary(lm(ValueEUR ~ Overall, data = dataset)) #31%
summary(lm(ValueEUR ~ Height, data = dataset)) #0%
summary(lm(ValueEUR ~ Age, data = dataset)) #0.1%
summary(lm(ValueEUR ~ WageEUR, data = dataset)) #68%
summary(lm(ValueEUR ~ ContractUntil, data = dataset)) #4%
summary(lm(ValueEUR ~ WeakFoot, data = dataset)) #2%
summary(lm(ValueEUR ~ PreferredFoot, data = dataset)) #0%


```

Figure 1.6 - code for linear regression models for all independent variables

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. WageEUR and Overall having a higher R-squared value indicating that the better the regression model fits our observations.

Checking which independent variables cause most variance on ValueEUR(dependent variable):

Overall- (Multiple R-squared: 0.3114, F-statistic: 8709 on 1, p-value: < 2.2e-16)

Height- (Multiple R-squared: 9.199e-05, F-statistic: 1.772 on 1, p-value:0.1832)

Age- (Multiple R-squared:0.001379, F-statistic: 26.59 on 1, p-value: 2.54e-07)

WageEUR- (Multiple R-squared: 0.6806, F-statistic:4.103e+04 on 1, p-value:< 2.2e-16)

ContractUntil- (Multiple R-squared: 0.04744, F-statistic: 955.2 on 1, p-value: < 2.2e-16)

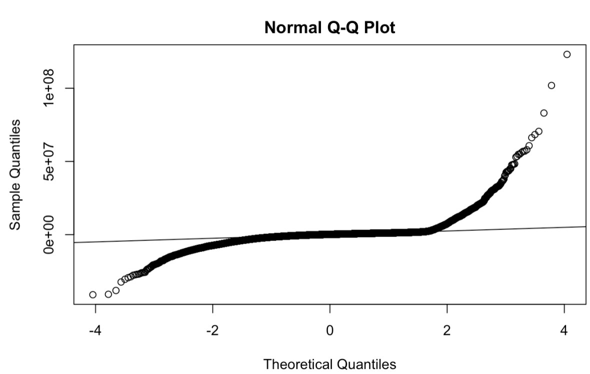
WeakFoot- (Multiple R-squared: 0.02221, F-statistic: 437.4 on 1, p-value: < 2.2e-16)

PreferredFoot- (Multiple R-squared: 0.0003646, F-statistic: 7.024 on 1, p-value: 0.008049)

Best Fit Model:

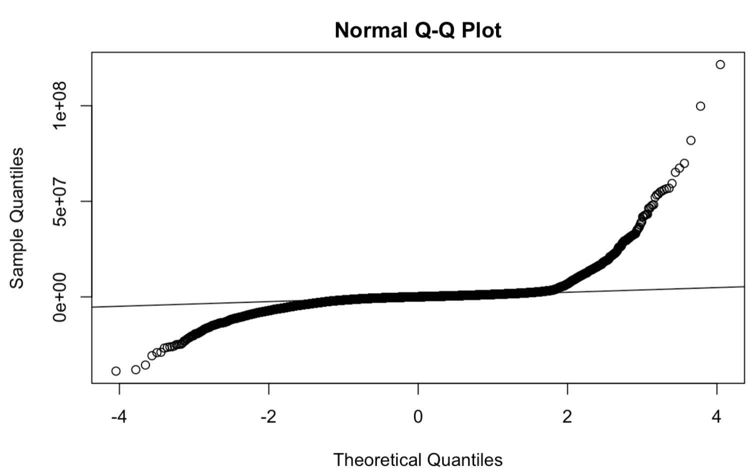
1st Iteration: In the 1st iteration, we ran linear models with the three highest R-squared values, those included- WageEUR, Overall, ContractUntil. After which we checked which model provides the highest adjusted R squared value. The highest value obtained was 0.6897 and the Multiple R-Squared for this model was 0.6897, signifying that the independent variables accounted for 68.9% of the variation in the ValueEUR.

```
Call:
lm(formula = ValueEUR ~ Overall + WageEUR + ContractUntil, data = dataset)

Residuals:
      Min        1Q    Median        3Q       Max
-41046331   -818161    189530    858979 123127369

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.578e+08  5.195e+07  -14.59   <2e-16 ***
Overall        9.978e+04  5.665e+03   17.61   <2e-16 ***
WageEUR        2.977e+02  1.986e+00  149.90   <2e-16 ***
ContractUntil  3.715e+05  2.570e+04   14.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4244000 on 19180 degrees of freedom
  (76 observations deleted due to missingness)
Multiple R-squared:  0.6897,    Adjusted R-squared:  0.6896
F-statistic: 1.421e+04 on 3 and 19180 DF,  p-value: < 2.2e-16
```

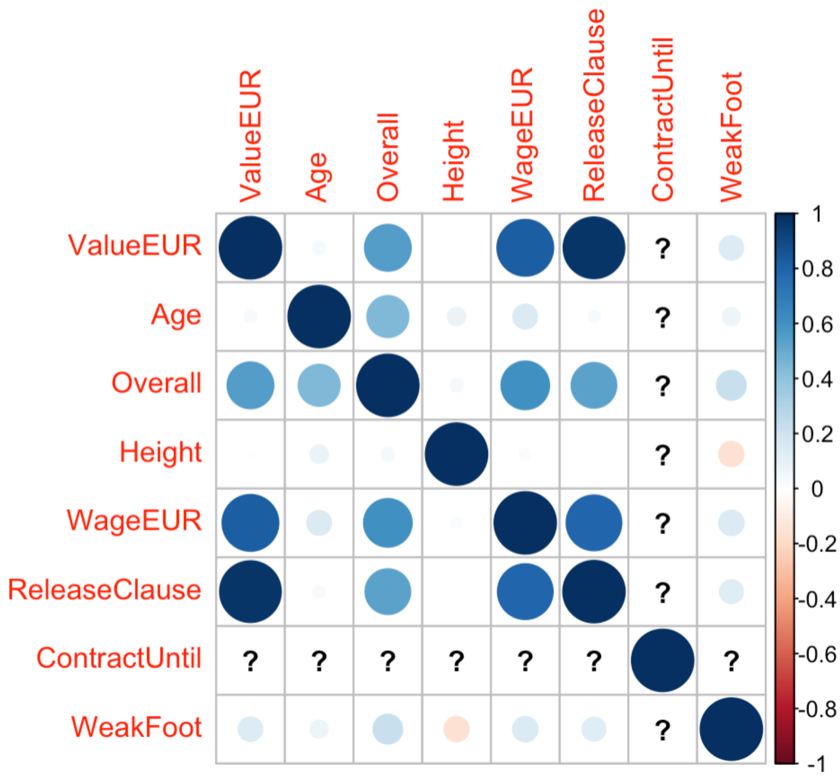


2nd Iteration: In the 2nd iteration, we ran linear models with all independent variables included. After which we checked which model provides the highest adjusted R squared value. The highest value obtained was 0.70.

```
Call:
lm(formula = ValueEUR ~ Overall + WageEUR + ContractUntil + Height +
    Age + WeakFoot + PreferredFoot, data = dataset)

Residuals:
      Min        1Q    Median        3Q       Max
-38818070   -820649     67800    836174 121549343

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.811e+08  5.370e+07  -3.371 0.000749 ***
Overall             2.002e+05  6.427e+03  31.160  < 2e-16 ***
WageEUR             2.887e+02  1.955e+00 147.640  < 2e-16 ***
ContractUntil       8.668e+04  2.658e+04   3.261 0.001113 **
Height             -6.404e+03  4.447e+03  -1.440 0.149873
Age                -2.437e+05  7.607e+03 -32.033  < 2e-16 ***
WeakFoot            2.556e+04  4.664e+04   0.548 0.583690
PreferredFootRight  4.366e+04  7.076e+04   0.617 0.537266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4133000 on 19176 degrees of freedom
  (76 observations deleted due to missingness)
Multiple R-squared:  0.7057,    Adjusted R-squared:  0.7056
F-statistic:  6569 on 7 and 19176 DF,  p-value: < 2.2e-16
```



In order to confirm our output from our linear regression model, we checked the correlation between all of the variables and then created a correlation plot to graphically visualise our results.



Conclusion

Our Project aimed at understanding the effects multiple variables had on the ValueEUR Score.

Through running different linear regression models and determining the best model fit, we came to the conclusion that while WageEUR with a Multiple R-Squared of 68%, seemed to be the most influential predictor on the ValueEUR, our multiple linear regression model proved that the independent variables that were the best predictors for the ValueEUR were Overall, WageEUR and ContractLeft.

We therefore reject the null hypothesis as our multiple linear regression models were statistically significant and accept the alternate hypothesis.