The Task:

The purpose of this analysis and case study is to predict the market value of football players and to help a football club strategize accordingly so that the overall performance can be improved. Accordingly, a player's market value can be defined as “an estimate of the amount of money a club would be willing to pay in order to make an athlete sign a contract, independent of an actual transaction”. So, to solve this problem, the nature of the relationship between of each variable with Value must be understood along with the individual characteristic of each variable. A statistical model is then adopted to further the analysis and arrive at the results and interpretation.

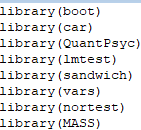
The Dataset:

The dataset contains the following variables: ****

The statistical model:

We have adopted the multiple linear regression model analysis in this case. A regression analysis is used to predict the value of one variable which is the dependent variable with the help of the other independent variables. So, to predict the market value of football players from the given data with the help of the other variables, the multiple linear regression model is best suited to the purpose. In the following pages of documentation, the approach steps have been clearly outlined.

Setting up the R model by loading the required libraries:



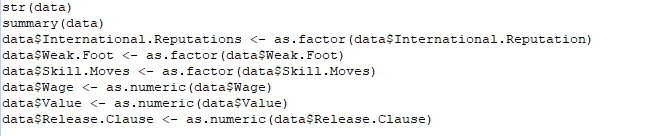
Data:

In the next step, the data is read into the R environment from the file.

C:\Users\ADMIN\Desktop\Fifa_19\data fitting.PNG

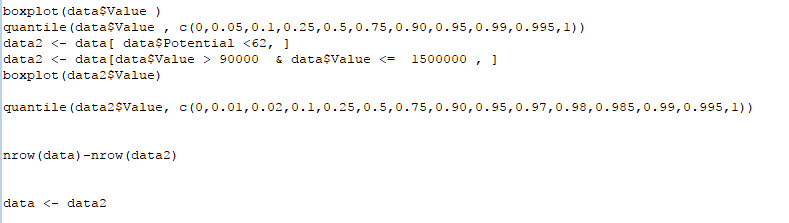
Data selection and data type modification:

In the given dataset, some of the columns which is named as “X”, “ID”, “Name”, “Age”, “Photo”, “Nationality”, ”Flag”, “Club”, “Club Logo”, “Special”, “Body Type”, ”Position”, “Real Face”, “Joined”, “Jersey Number”, “Loaned From”, “Contract Valid Until”, “Height”, “Weight”, “LS”, “RS”, “ST”, “LW”, “LF”, “RF”, “CF”, “RW”, “LAM”, “RAM”, “CAM”, “LM”, “CM”, “RM”, “RCM”, “LCM”, “LWB”, “RWB”, “LDM”, “RDM”, “CDM”, “LB”, “RB”, “CB”, “LCB”, and “RCB” are no relevance to our model and is thus excluded. Two of the variables, Weak.foot, International.Reputation are converted into factors and three of the variables, Wage, Value and Release.Clause are converted into numbers.

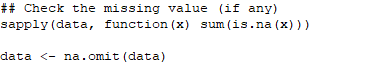


Data Cleaning:

In this step all the numeric variables are checked for the presence of outliers and using boxplots and quantile functions, they are removed. In each case, care is taken to keep at least 80% of the data. Only the data points that lie above or below the maximum and minimum

  
entries of the boxplot are removed from the dataset.

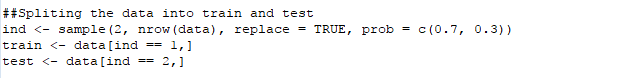
Checking for missing values:



After the data has been cleaned off all the outliers, it is then checked for any missing values in the following manner. No missing values were found.

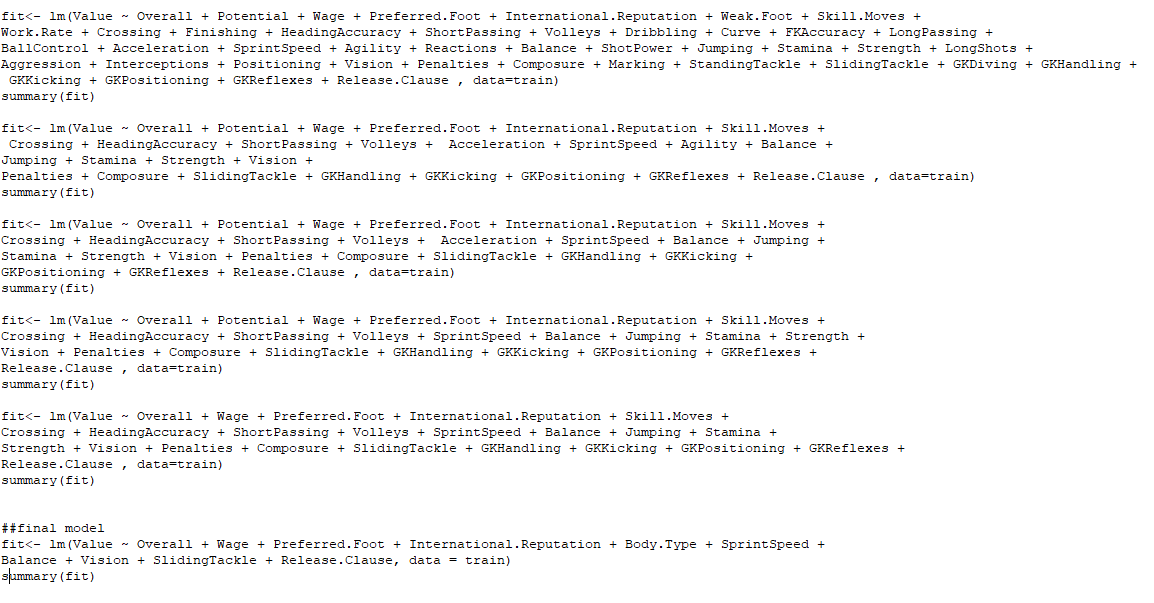
Splitting the data:

The data is then split into two parts. 70 percent of the data is split into ‘development’ and the remaining 30 percent is named ‘validation’. The development part is for training the linear regression model and the validation part is for testing the robustness of the model.



**Running the linear regression model:**

Once the data has been divided into two parts, a linear regression is performed with the value as the dependent variable. Once the model has been run, the anova value for each individual variable is checked. The variables with p value< 0.05 are removed from the model one by one such that only the statistically significant ones remain.



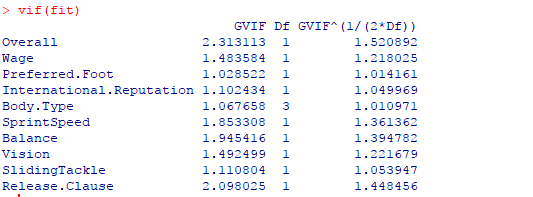
Goodness of fit of the model:

The coefficient of determination which is denoted by R2  is a key output of the regression analysis. It is interpreted as the proportion of variance in the dependent variable that can be explained by the independent variable. The value of R2  ranges from 0 to 1 and the closer the value is to 1, the better is the model. In our model, the value is 93% approximately which indicates a high amount of goodness of fit of the model.

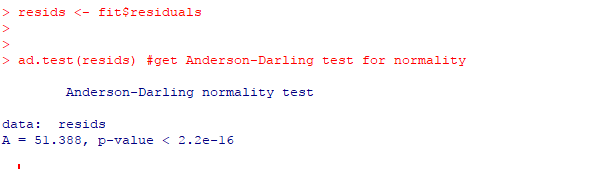
The adjusted R2 which provides a more honest value to estimate the R-squared for the population because with the normal R-squared, the value increases with the increase in the number of predictors. The value of the adjusted R-squared in our case also comes to around 93%.

**Checking the Assumptions for a linear model:**

1. ***Assumption of multicollinearity:*** - This is the most important of the assumptions of a linear model and it states that there should be no perfect linear relationship between two or more of the predictors or independent variables. This is tested with the vif function and any variable with a value of GVIF should be within 2. If it is greater than 10 then serious problem. In our case multicollinearity between independent variables was almost absent but nothing to worry about.

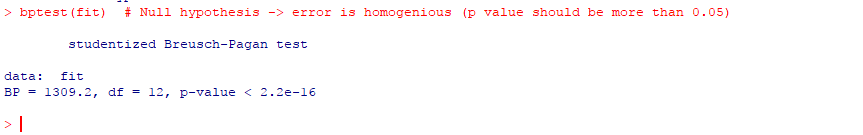


***2> Assumption of normality: -*** This assumption states that the errors or the residuals of the model should be normally distributed. This is checked using the Anderson-Darling normality test. The p value should be more than 0.05 for the assumption to hold but in this case the assumption doesn’t hold.

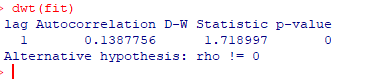


***3> Assumption of homoscedasticity:*** This assumption states that at each level of the predictor variables, the variance of the residual terms should be constant. This assumption is checked with the help of the Breusch-Pagan Test. Like before, for the null hypothesis to hold, a p value greater than 0.05 should be returned. This assumption however doesn’t hold in this case.

**Output:**



***4>Assumption of serial correlation: -*** This assumption states that for any two observations, the residual terms should be uncorrelated or independent. This assumption is checked with the help of the Durbin–Watson test. The value of the statistic ranges from 0-4 and a value close to 2 indicates no correlation. In this case, the assumption holds with a D-W statistic value of approximately 2.



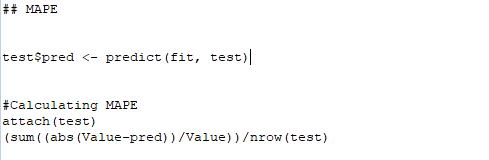
**Validation of the model:**

The predictions for the value variable are done using the validation part of our data. The predictions are then saved in a csv file for reference.

test$pred <- predict(fit, test)

write.csv(data2, "data2result.csv")

***MAPE:*** The MAPE or Mean Absolute Percentage Error is then checked using the validation data. The MAPE shows how much the predicted values differ from the actual values. The value for MAPE ranges from 0 to 1 and the lesser the MAPE, the better is the model. The calculation is :

 C:\Users\ADMIN\Desktop\Fifa_19\Mape2.PNG

We can assume the model to be good because of above result.

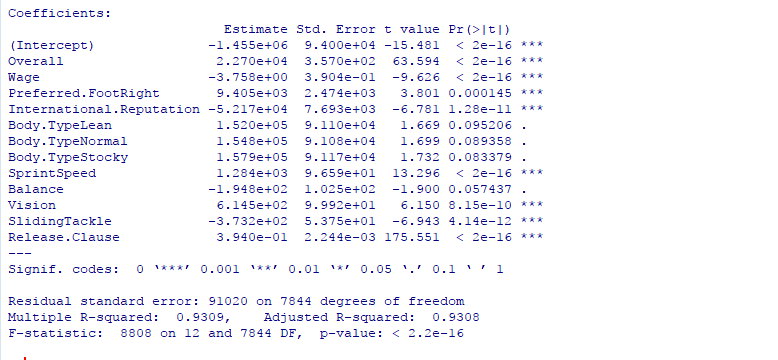
The significant variables and their significance:

The following image shows the variables that are significant to our model and the image below shows us the relationship of these variables with the dependent variable. The following are the significant variables with a positive relationship with customer lifetime value:

* Overall
* Players with right foot preference
* Body Type
* Sprint Speed
* Vision
* Release Clause

However, the following are the significant variables with a negative relationship with the dependent variable:

* International Reputation
* Balance
* Sliding Tackle



The Business Interpretation

For the Club to prosper they should focus on the following targets:

* The club should focus on players who have a high overall. High overall players are potential goldmines when it comes to value as they have positive relationship as inferred by the model.
* Players with right foot preference should also be targeted. They also have a strong positive relationship with value.
* Managers shouldn’t judge players by their Body Types. Even stocky or lean type of body has positive relationship.
* Players with higher sprint speed are also high value players and should be kept within the fold.
* Managers should pursue players who have better vision as it has a very positive relationship with value .
* The club must look into the Release Clause of the players as it has a significantly positive relationship with value.
* With increase in International Reputation the value may decrease as it have much negative relationship.
* Those with higher balance must not be pursued as it has significantly negative relationship with value.
* However, players those who make more sliding tackles have much lesser value and have significantly negative relationship with value.