NORTHEASTERN UNIVERSITY

A Project Report on

"WORLD FOOTBALL PLAYER ANALYSIS"

By

Akash Chitrey	**001888848**
Akshay Jain	**001873110**
Reuben Mathew Philip	**001888758**
Sankalp Vatsh	**001246262**

Under the guidance of

Professor. Rajesh Jugulum

TABLE OF CONTENTS

Introduction

Analytics has rapidly integrated into the sport industry to optimize scheduling, assist with resource allocation, and examine the legal environment within sports organizations. The goal of sports analytics is to analyze available data and identify all the correlations, trends and patterns that will affect the decision-making process, and, thus, improve the performance or achieve better results.

Many sports, like baseball and golf, have been successfully implementing it for some years. Technological breakthroughs in the last several decades have contributed to the development and popularity of sports analytics. With the introduction of the hawk-eye system in 2005, tennis has become another sport that makes an analysis based on the statistical data provided by hawk-eye tracking software. Today, most big sports teams have an analytics expert, which is a rapidly growing profession.

Even though sports analytics have been part of decision making in many sports for a long time, Roger Neilson, a former hockey coach and game innovator, was one of the first to use analytics to assess players and their on-court performance. In 2014, the National Hockey League (NHL) started to implement detailed analytical approach using advanced Hockey statistics.

Data analytics are playing an important role in sports because of the following:

- **Building a better and more successful team** – the objective of every hockey team is to achieve better results. Team or single player performance can be evaluated with the use of available quantitative and qualitative data. The playing strategy and evaluation of it is also largely dependent on the statistical data.
- **Competition** – statistical data of most hockey teams is usually available online. There are countless sites and forums dedicated to sports analytics. Teams use this data for the comparison, and to identify opponent's strengths and weaknesses. If interpreted right, sports analytics can become a team's competitive advantage.
- **Prevent mistakes** – analytics will provide you with a detailed analysis of a team's performance – what was done well and what went wrong. The goal is to minimize mistakes and achieve better cohesion for future games.

In our project, we used data from a popular video game FIFA to analyze the various attributes of a football player and tried to establish a correlation between the skills and success of a player. It also tries to correlate how a better player tends to have better wages. We tried to establish how better players have higher wages and clubs spending the maximum amount on these players tend to be more successful in the major competitions.

We have used R to analyze the data. We have then analyzed the results and generated various actionable insights, as well as inferences out of them.

The objectives of this project are:

- Determine what makes a football league successful
- Establish a relation between individual attribute and overall attribute and suggest ways to improve a player
- Establish a correlation in wage and the overall attribute of a player
- Study the distribution of players and their wages

P-Diagram

We have constructed a P-diagram to clearly communicate the working of our model.

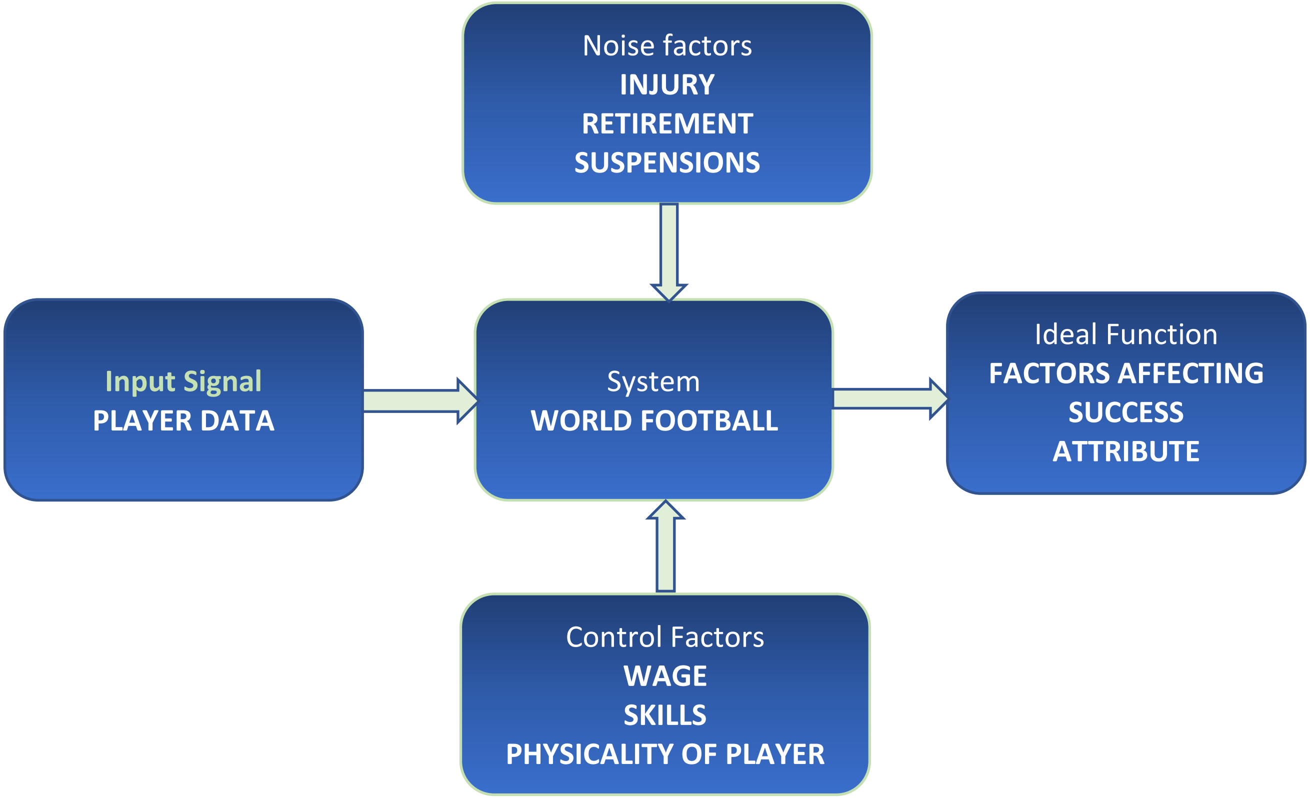


Figure 1: P-Diagram

Choosing the data

Data was collected for football players across all the major leagues in the world. To narrow down our analysis, we focused our analysis on the data for the leagues with the maximum wages, as we got a positive correlation between the overall attribute and wages. This allowed us to perform our analysis on the best players and the strongest leagues around the world. The correlation coefficient between the overall attribute and the wages is 0.6847895.

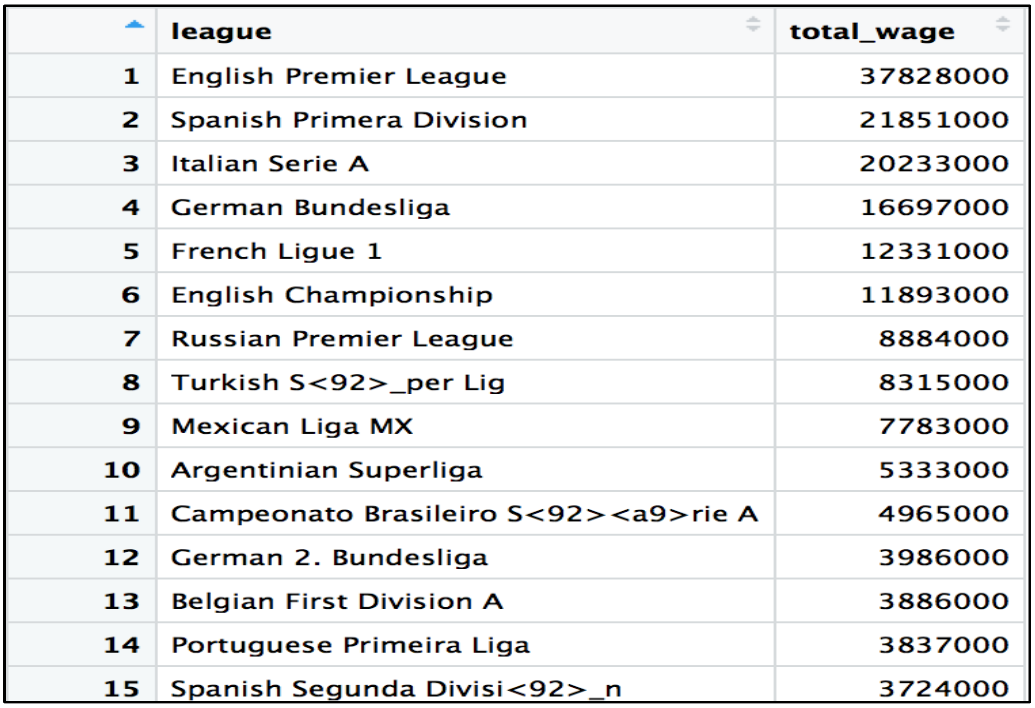
Command for the calculation:

```
> cor (Overall Attribute, wage)
```

Leagues spending the maximum amount on Wages

Command in R:

```
topleagues <- fifa.data %>% group_by(league) %>% summarise(total_wage = sum(eur_wage))
%>% arrange(desc(total_wage))
```



	league	total_wage
1	English Premier League	37828000
2	Spanish Primera Division	21851000
3	Italian Serie A	20233000
4	German Bundesliga	16697000
5	French Ligue 1	12331000
6	English Championship	11893000
7	Russian Premier League	8884000
8	Turkish S<92>_per Lig	8315000
9	Mexican Liga MX	7783000
10	Argentinian Superliga	5333000
11	Campeonato Brasileiro S<92><a9>rie A	4965000
12	German 2. Bundesliga	3986000
13	Belgian First Division A	3886000
14	Portuguese Primeira Liga	3837000
15	Spanish Segunda Divisi<92>_n	3724000

Figure 2: Top leagues with highest wages

We then plotted the overall attribute against the wage level, in order to observe this relationship. We can see that the wage level increases exponentially with increase in the overall attribute. While most plots follow this relationship, there are a few outliers as we can see in the graph.

Command in R:

```
ggplot(fifa5.data,aes(x=OverallAttribute,y=Wage,color=league))+geom_point()
```

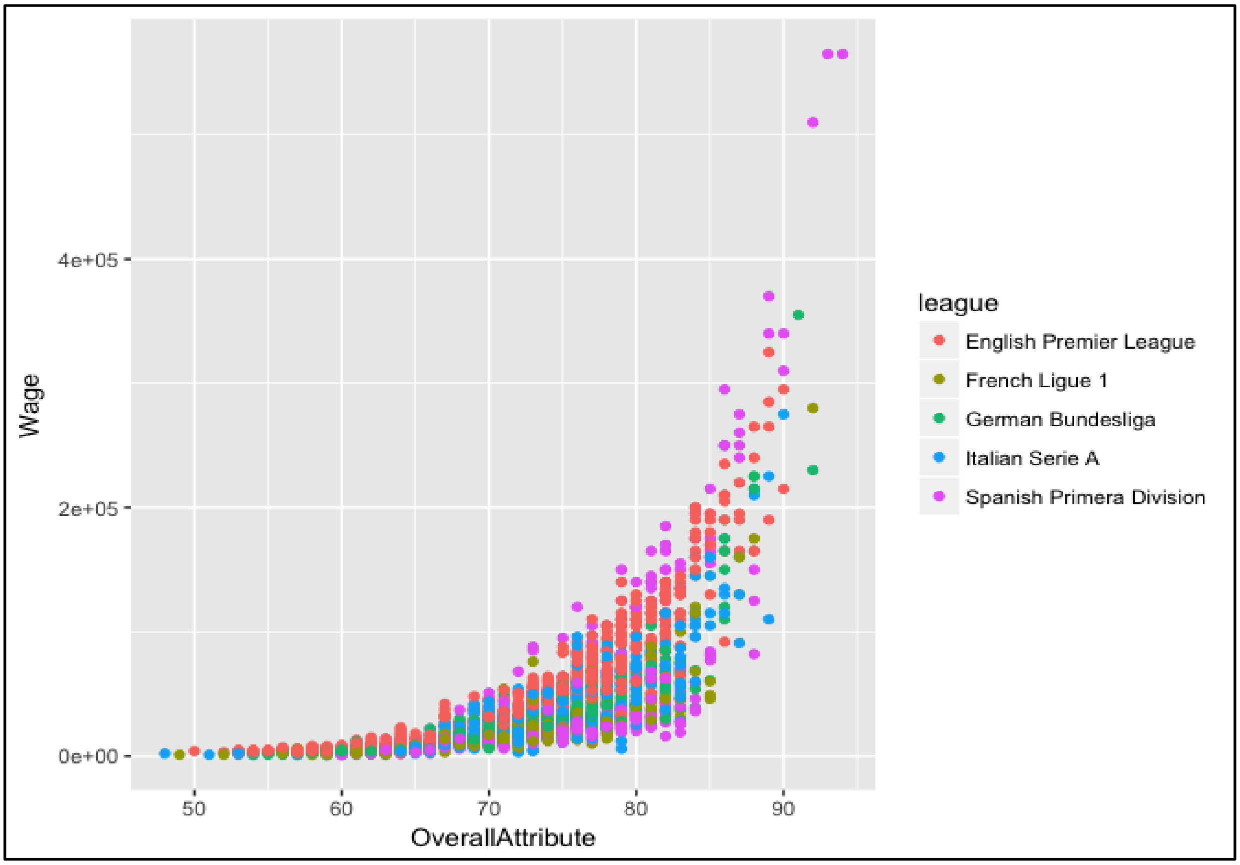


Figure 3: Correlation between overall attribute and wages

Predictive Analysis

We have tried to establish a relationship between individual attributes and the overall attribute. Here, we have narrowed down to evaluating 5 major attributes i.e. Pace, Shooting, Dribbling, Physical, and Defense. We observe that the sum of these attributes appears to have a linear relationship with the overall attribute, as you can see in the plot below.

The red line represents a prediction line, that predicts the increase in the overall attribute, as a function of the individual attributes are considered.

Command in R:

```
fifa5.data%>%add_predictions(fit) %>% ggplot(aes(x=Pace+Shoot+Dribbling+Physical+Defence))
+ geom_point(aes(y=OverallAttribute)) + geom_line(aes(y=pred,color='red'))
```

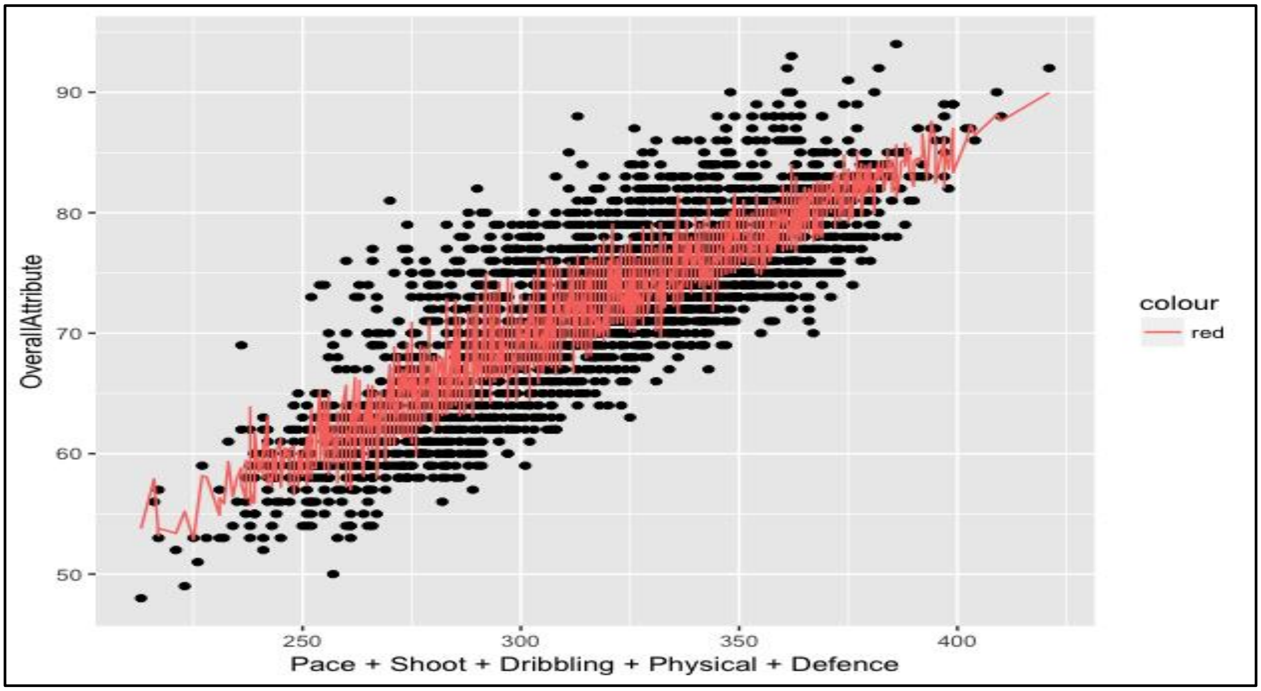


Figure 4: Predictive analysis for player attributes

The next analysis gives a clearer picture of the skills and the individual contribution of each to the overall attribute of a player. The programming language R has been used to get calculate these t values.

Linear Regression Model

In order to further analyze which attribute affect the overall attribute and to what degree, we plot the intercepts of the attributes and generate their correlation coefficients.

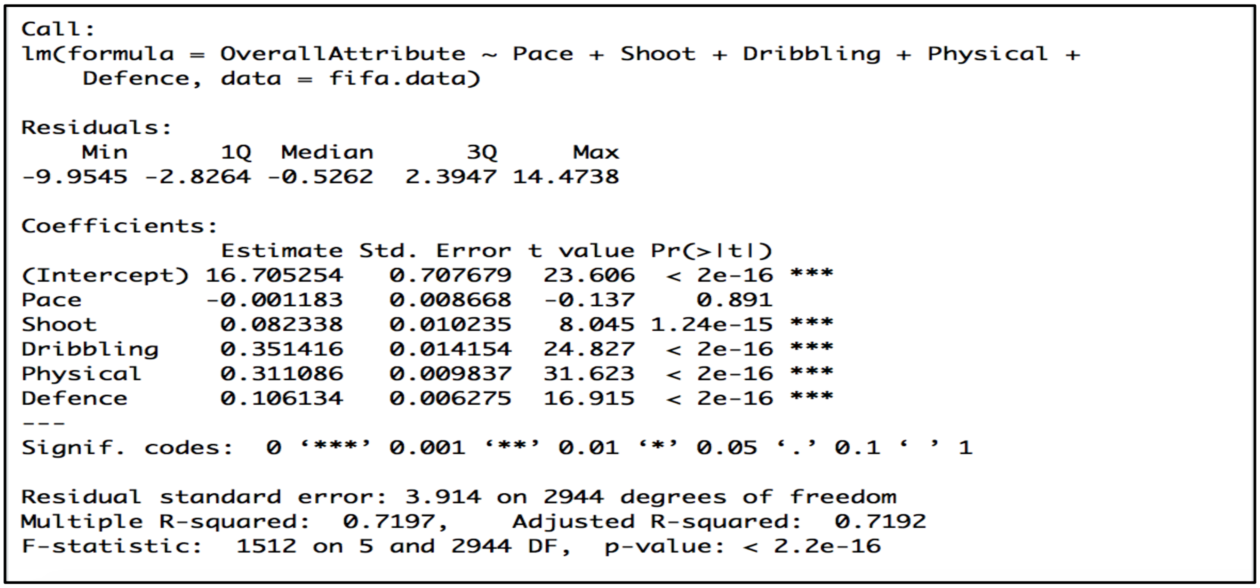
We observe that while Dribbling, Physical, Defence and Shooting are positively correlated to the overall attribute (in that order), Pace is not as correlated (slightly negatively correlated) to the overall attributed. Further, we see that the probability of generating the t value, given that there is no relation between Pace and the overall attribute is high. This gives us a strong reason to conclude that Pace is counterproductive towards the overall attribute when these 5 attributes are considered.

Command in R:

fit<-lm(OverallAttribute~Pace+Shoot+Dribbling+Physical+Defence+Pass, data=fifa5.data)

In the figure below:

T value- represents the test statistic value
Pr(>|t|)- represents probability of getting t value given that there is no relationship between the particular attribute and overall attribute
Residual Standard error – represents the RMSE
Multiple R-squared- percentage of variation accounted for in the output



```
Call:
lm(formula = OverallAttribute ~ Pace + Shoot + Dribbling + Physical +
    Defence, data = fifa.data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9545 -2.8264 -0.5262  2.3947 14.4738

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.705254   0.707679  23.606  < 2e-16 ***
Pace        -0.001183   0.008668  -0.137    0.891
Shoot        0.082338   0.010235   8.045 1.24e-15 ***
Dribbling    0.351416   0.014154  24.827  < 2e-16 ***
Physical     0.311086   0.009837  31.623  < 2e-16 ***
Defence      0.106134   0.006275  16.915  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.914 on 2944 degrees of freedom
Multiple R-squared:  0.7197,    Adjusted R-squared:  0.7192
F-statistic:  1512 on 5 and 2944 DF,  p-value: < 2.2e-16
```

Figure 5: Contribution of each attribute to the overall

This insight can be used by players trying to improve their overall attribute (as well as get paid more, we have established that the two are highly correlated). In order to improve their overall attribute, they should ideally work on their individual attributes in the descending order of correlation to the overall attribute. That is, beginning with Dribbling, and then Physical, Defence and then Shooting.

Hypothesis testing (Goodness of Fit)

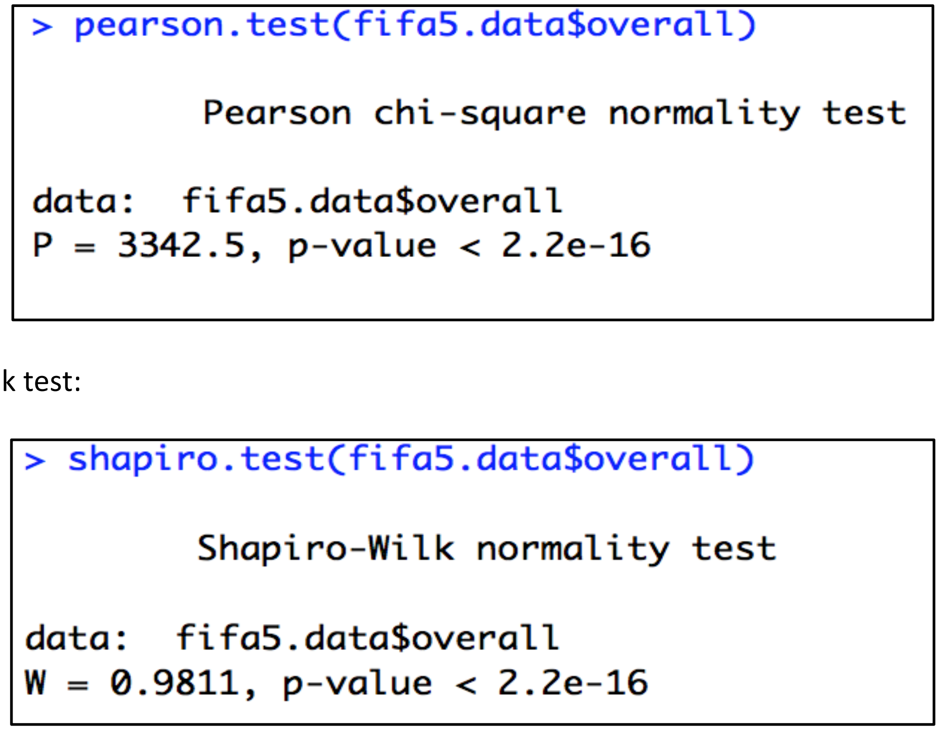
We also performed hypothesis testing on the data to check if the overall attributes are normally distributed or not. Two tests were performed to have a better understanding of the results, the Pearson chi-square normality test and Shapiro-Wilk normality test. We defined the null and alternate hypothesis as follows,

H_0: Overall Attribute is normally distributed

H_1: Overall Attribute is not normally distributed

The p-values we obtained from these tests were very small and thus based on these values we rejected the null hypothesis and accepted the alternate hypothesis.

Pearson test:



```
> pearson.test(fifa5.data$overall)

        Pearson chi-square normality test

data:  fifa5.data$overall
P = 3342.5, p-value < 2.2e-16
```

Shapiro-Wilk test:

```
> shapiro.test(fifa5.data$overall)

        Shapiro-Wilk normality test

data:  fifa5.data$overall
W = 0.9811, p-value < 2.2e-16
```

Figure 6: Hypothesis test results

When we generate a frequency plot of the overall attribute of the players, we see that the skill level of the players is skewed to the left. We performed two tests to test for normality, and both rejected the Null Hypothesis (that the distribution follows normality). We believe the skewness can be explained by the fact that to play major league football in Europe, especially to be hired by the major clubs which we have considered (view data set section), the overall attribute of the players would need to be high. Hence there would be a relatively low number of players with a low overall attribute.

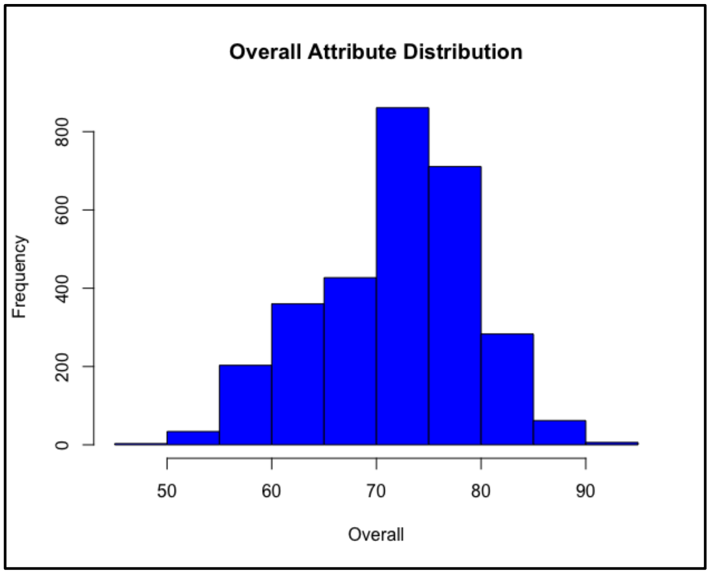
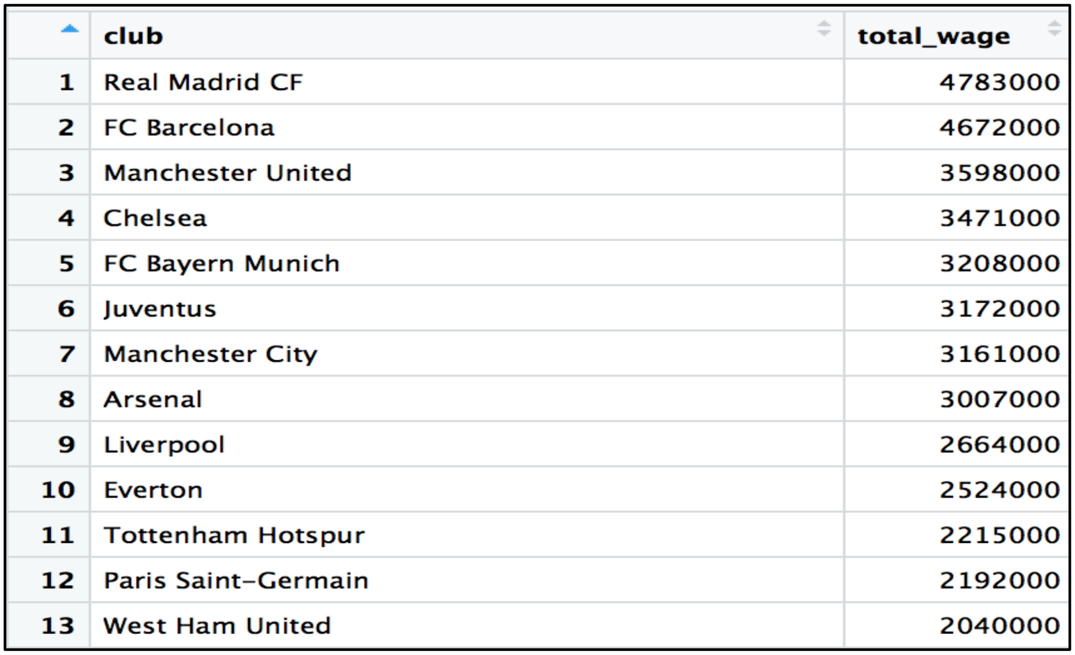
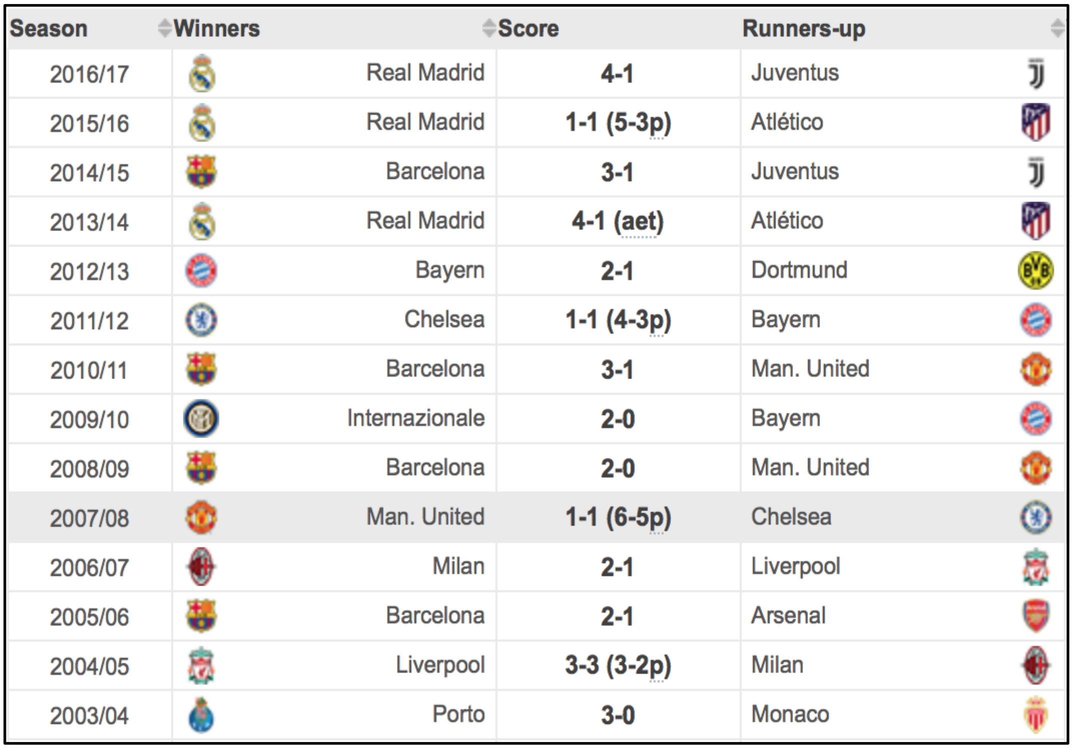


Figure 7: Overall attribute distribution

Establishing relationship between the money and success



	club	total_wage
1	Real Madrid CF	4783000
2	FC Barcelona	4672000
3	Manchester United	3598000
4	Chelsea	3471000
5	FC Bayern Munich	3208000
6	Juventus	3172000
7	Manchester City	3161000
8	Arsenal	3007000
9	Liverpool	2664000
10	Everton	2524000
11	Tottenham Hotspur	2215000
12	Paris Saint–Germain	2192000
13	West Ham United	2040000



Season	Winners	Score	Runners-up
2016/17	Real Madrid	4-1	Juventus
2015/16	Real Madrid	1-1 (5-3p)	Atlético
2014/15	Barcelona	3-1	Juventus
2013/14	Real Madrid	4-1 (aet)	Atlético
2012/13	Bayern	2-1	Dortmund
2011/12	Chelsea	1-1 (4-3p)	Bayern
2010/11	Barcelona	3-1	Man. United
2009/10	Internazionale	2-0	Bayern
2008/09	Barcelona	2-0	Man. United
2007/08	Man. United	1-1 (6-5p)	Chelsea
2006/07	Milan	2-1	Liverpool
2005/06	Barcelona	2-1	Arsenal
2004/05	Liverpool	3-3 (3-2p)	Milan
2003/04	Porto	3-0	Monaco

Figure 8: Top clubs with wages and Champions league results of last 10 years

The top 5 clubs from the first table are the ones who have been winning the most reputed trophy in Europe. As we can see from the two tables above the teams that spend the maximum amount on wages tend to be more successful. The 2nd table shows data from the UEFA Champions League website which is the most coveted club competition in the world. These tables show the direct correlation between the money spent and success. This brings us to the conclusion that players with higher skill sets tend to earn better wages and the clubs able to afford these wages get the best players and form a better team.

Another interesting observation is that while the overall attribute (skill) is skewed to the left, the wages are skewed to the right. This reveals that while players are paid according to their skill level, a majority of them are paid comparable amounts, while very few are paid really high amounts. We believe that this can be explained by various factors like fame and exceptional skill that give some players the ability to demand extraordinarily high paychecks.

Conclusions

- The wage and overall attribute is strongly correlated
- The players overall attribute is dependent on Shooting, Dribbling, Physicality and Defense and Pace doesn't play a major role in the overall attribute.
- Teams with higher wage structure tend to be more successful in the competitions
- Skill is skewed in major league football towards the left, while the wages are skewed toward the right.

Learning Outcomes

- We have learned that statistical analysis is indeed a useful tool to help us find actionable insights and inferences from various kinds of data. (Historical performance information, in our project)
- Results of Data Analysis only gives us information. We need to use these results and fit them into the process we are studying in order to turn them into knowledge and actionable insights.
- Noise is a given for any large data set. We need to ensure that our analysis takes this into account while producing results.
- We have learned that working in a group helps us see things from different perspectives, have constructive deliberations, critically analyzing each other’s propositions, and hence resulting in a better outcome.