# Data Science Project

## Executive Summary

The project I have created looks at player performance data of football players from 2015-2024. The players at show are the best performing players of that time. This Project answers the question ‘which players are consistently creating and converting chances and if so how is this affecting their rating?’ therefore this will be looked at by club scouts etc. to gather better insights on who would be a good potential signing and who will perform in the future. I have built an interactive dashboard which provides information in a user friendly way, whilst also having built an excel sheet which supports this dashboard. I have also created a basic linear regression model which predicts future goals of each player based on how they’ve performed at previous stages.

The methods used where all done in a way to allow for the best possible outcome, firstly the data was cleaned and standardised e.g. The formatting of columns to be the same throughout and making an outliers not stand out, next was ensuring I had any extra information from this data set taken out which may require calculations e.g. Goals per match. Once the excel data set was in a state I was happy with, I then used it to support me in creating a dashboard in power bi, this was created to allow for a more concise way to view the data, I then also used the data in excel to create a simple linear regression model which could predict the future amount of goals of players. This was done to support the question on who could be a good signing for clubs and who will perform in the future.

The key outcomes where clear. Where the goal scoring was led by the elite forwards and the top assist providers where the world class wingers, but the highest average rating of players was where the player could provide both and not just one of these abilities to the club. This was shown in the dashboard whilst also being obvious when doing a deeper dive into the excel sheet – and the predictive model allowed to highlight how much better or worse these players may get.

## Data Infrastructure and Tools

Throughout this project I used 3 main tools – Excel, Power Bi and Google Collab. Google collab was used as it’s a free and cloud based and doesn’t require any installation, which makes it ideal for working on any device etc. It allowed me to upload my Excel file which then I could explore the players football data further and create visuals etc. using phyton libraries. These visuals helped me identify trends in goals and other information which will support decision making of players strengths and weakness. I also built a simple predictive model to forecast future goals, as research has shown its ability to aid decision making (Dixon et al, 2024). Collab supports this process by providing computing resources and Gemini which allowed for corrections when struggling with my code. I also used excel where I cleaned and formatted my data – this then allowed for me to easily create my visuals in power bi as the data was standardised throughout and the key information was there – I then was also able to do a deeper dive into my data within excel where I could create models in excel. Lastly, Power Bi was used as it gave a user friendly way to see the information I had discovered whilst also allowing for a deeper dive in some of the stats that stakeholders may want to see – such as who was the top scorer of a certain season – that’s something that may take a while to find in excel whereas in power Bi that can be found quickly.

Data Engineering

Data engineering focused on transforming the raw dataset exported through kaggle into a structured, analysis ready format Ofosu, K. (2020). First column names were standardised throughout the dataset so it made more sense and was easier to read. (e.g. “Seasons Ratings” were called “Rating”, “Players” were changed to be called “Player”), and text fields were cleaned using the ‘=TRIM(A2)’ function which removes and unneeded extra spaces. Numeric fields such as Goals, Assists, Matches, and Rating were converted from text to numbers. The Season column contained mixed formats like “2023/2024” and “2023”, so I created a new column called, YearEnd, was created using the formula:

**=IF(ISNUMBER(FIND("/",A2)), RIGHT(A2,4), A2)**

This extracted the final year (e.g. 2019/2020 to 2020). For the outliers such as “2016/1017”, I just manually corrected them.

Next I created a Pivot Table which was used to aggregate the data by player, this made it a lot easier to understand and also analyse and create any visualisations Selwyn, N. (2020). Next, a Pivot Table was used to aggregate data by Player, calculating Sum of Matches, Sum of Goals, Sum of Assists, and Average of Rating. To compare all the players data I used fields which would do so, calculate fields which did this where:

* Goals per Match = Total Goals / Total Matches
* Assists per Match = Total Assists / Total Matches
* Goal Involvements per Match = Goals per Match + Assists per Match

The metrics I’ve created here account for the differences in all the players playing time and highlight the data we want to know rather than raw totals. To measure the consistency of players better I used standard deviation of rating which was calculated by adding rating to the pivot table and changing the value field setting to STDDEV, this gave an indication to me of how stable a players performance was across seasons. I also added in filters to reduce the bias of results such as if a player had less than 30 matches I filtered them out when ranking the avg. rating. Then missing values were handled because excel ignores blanks when doing calculations, so nothing was required. All of these changes can be seen below.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5. Original Dataset

A table of sports teams

AI-generated content may be incorrect.

Figure 6. Cleaned/ Standardised Dataset

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 7. Further information captured based on calculations etc. in excel

## Data Visualisation & Dashboards

Power Bi was used to create a clear and structured layout to communicate key findings and also dive deeper into some of the information. Charts and tables were selected to highlight trends in player performance over time and give the ability to look back. The design incorporated principles such as 5-50-500, which is the idea that the user will be able understand the KPIs in 5 seconds then further information in 50 seconds and then the full intricate deep dive in 500 seconds.

The visuals is connected throughout the project with e.g. The excel file and some of the findings there whilst also allowing for when having a deeper dive into the information trends and predictive insights to be reinforced and backed by the dashboard – used consistent colours and clear labels which ensures the visuals support all target audiences. Ameer, M., Rahul, S.P. and Manne, S. (2020)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 8. Dashboard created in power bi

Data AnalysisA graph with blue lines and red lines

AI-generated content may be incorrect.

Figure 9.example of forecasting model created

This example highlights a simple linear regression model which forecasts the next season goal tally of the player Yang, Y. (2006). In this example Cristiano Ronaldo’s historical goal trend illustrates a peak when he played for Real Madrid and Portugal where he was scoring 40+ goals where then there is a natural decline reflecting his aging and league transitions the model has forecasted his next tally to be approximately 20 goals. This estimation is derived from previous performances for club and country but it does not take into account any major changes. Therefore the prediction will be treated as a baseline and not a definite outcome. For example The league that Ronaldo has moved to now is much easier compared to the top 5 major leagues which the model has not taken into account so it cannot be fully trusted to provide a guaranteed prediction. Rather the model may produce better prediction for a player who is younger and isn’t planning on moving leagues. If I could change one thing about this model if I was to do this project again it would be to have the trend line be altered based on the age of the player.

A computer screen shot of a program

AI-generated content may be incorrect.

The Script I’ve created here predictions each players goal for the fore coming season using a season trend line. It looks at every player, and then skips anyone with few then **min\_seasons** years, totals their **Goals** per **YearEnd,** it ten fits a straight line (where it uses the year to predict goals). It projects one year ahead, it the stops unrealistic outcomes using **clip\_min** and **clip\_max**, rounds the results and returns a tidy table – which I have then created a visual off this based on the player I want to visualise.