ANALYSING AND PREDICTING EXPENSIVE FOOTBALL TRANSFERS FROM 2000-2021

MINOR PROJECT REPORT

Submitted by

RACHIT NIGAM (RA1911027010035)

MUDIT KRISHNA (RA1911027010044)

ABHIMANYU BHADAURIA (RA1911027010045)

VIDUSHI GUPTA (RA1911027010046)

Under
Dr.A.Shanthini

of

B.Tech. BDA
Department of Data Science and Business Systems
School of computing

SRM Institute of Science and Technology

KATTANKULATHUR
October 2021

SRM UNIVERSITY

KATTANKULATHUR

BONAFIDE CERTIFICATE

Certified that this project report **"ANALYSING AND PREDICTING EXPENSIVE FOOTBALL TRANSFERS FROM 2000-2021"** is the bonafide work of "**RACHIT NIGAM, MUDIT KRISHNA, ABHIMANYU BHADAURIA AND VIDUSHI GUPTA"** who carried out the Mini project work under my supervision.

SIGNATURE

Dr. A. Shanthini

INTERNAL EXAMINER

SIGNATURE

HEAD OF THE DEPARTMENT
Data Science and Business and Systems

ACKNOWLEDGEMENT

The success and the final outcome of this project required guidance and assistance from different sources and we feel extremely fortunate to have gotten this all along the completion of our project. Whatever we have done is largely due to such guidance and assistance and we would not forget to thank them.

We express our sincere thanks to our faculty Dr. A. Shanthini, our HoD ma'am, our friends, and our family for providing support throughout this project.

We are thankful to and fortunate enough to get constant encouragement, support, and guidance from all the Teaching staff of the Department of Information Technology which helped us in successfully completing our minor project work. Also, we would like to extend our sincere regards to all the non-teaching staff of the department of Information Technology for their timely support.

Rachit Nigam (RA1911027010035)

Mudit Krishna (RA1911027010044)

Abhimanyu Bhadauria (RA1911027010045)

Vidushi Gupta (RA1911027010046)

TABLE OF CONTENTS

ABSTRACT

The purpose of this project was to analyze expensive football player transfers and create a machine learning-based model which could predict the transfer fee of football players based on some range of data. Nowadays, concepts of Data Science and AI are widely used in almost every sphere. Many football clubs use such techniques to maximize their profit.

The project demonstrates one way to develop such a technique. The first phase of the project included data collection, analysis, visualization, and interpretation. Largely the concept of Linear regression was applied for the project. After model evaluations, respective conclusions, visualizations, and interpretations were made.

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS

1. AI – Artificial intelligence
2. G colab - Google Collaboratory notebook
3. PSG – Paris Saint Germaine
4. EPL – English premier league
5. Vs – versus

INTRODUCTION

Nowadays, soccer teams buy and sell thousands of players during each transfer window, spending millions of dollars to buy them. The top players of the game are even worth a couple of hundred players during transfers, football teams aim to maximize the efficacy of each transfer. Some football teams use AI-based technologies to predict football players' transfer fees and market value. According to an article published on BBC, if a player transfers before their contract expires, the new club pays compensation to the old one. This is known as a transfer fee. While a player's market value is an estimate of the amount for which a team can sell the player's contract to another team. The market values attached to the players do not play a key role as in many cases a player is sold for a much higher price than his market value or much lower price. At the moment of his transfer from Barcelona to PSG Neymar's market value was 100 million euros, but PSG paid more than double the price to sign him. So, what are the main qualities of the player or other factors, that decide his transfer price? Also, when is the best time to sell the player? Which of the transfers paid off for the teams? And in general, are there any special connections among the teams or leagues of the transfer market? Those are some of the questions that we are going to provide answers to. One of our main goals is to identify whether the easily interpretable statistical measures of the players are solid predictors for his transfer fee and how well can these variables predict the player's price. As the importance of a statistical measure varies depending on the position of the player on the field, we will implement the predictions for each position separately. In the end when we find the best prediction model we can investigate the underrated and overrated players based on their transfer fee.

LITERATURE REVIEW

Various studies and projects conducted by individuals or organizations for different purposes, which tried to investigate the features that may impact the transfer fee of the football players and conduct predictive models and methods.

- First, there is a study conducted by CIES Football Observatory. They published a document called "Scientific assessment of football players' transfer value" in October 2018. The data that they have collected consists of over 2,400 transfers, involving top-5 league players between July 2011 and August 2018. The number of features is 36 including information about footballers' contract duration, year of transfer, book value, loan status, nationality, and economic level of the releasing club. They used multiple linear regression which included only significant features, leading to the overall model to be very significant (p-value < 0.00001) with an adjusted coefficient of determination evaluated as 0.86. The study concluded that the model is useful for defining the starting value, an initial salary of the players, as well as understanding the value of the club in the future based on the price of the players (olim, Ravenel, & Besson, 2018).

- Another study we examined was "Football player's performance and market value" published by Ricardo Cachuchino, Miao He and Arno Knobbe, published in 2015. The authors of the paper wanted to understand the relationship between the performance of the player and his market value. They also underlined the current problem of player's economic valuation and deviations in the market values and transfer fees. Their dataset included information about the player's performance, his ratings from WhoScored,

market information from TransferMarkt and some other performance assessment metrics gained by juries. The scope of the project only included LaLiga players for the half of the 2014-2015 season. The authors built their model of evaluating the market value using Lasso Regression. They emphasized the importance of choosing the right lambda in filtering the important features. Later they figured out the importance of evaluating the players based on their position and continued their work putting the main emphasis on evaluating the forwards. They came out with a simple linear model for evaluating the forward's performance based on his behavior on the field (fouls, yellow or red cards), shots from different areas, goals. (M. He, Cachucho, & Knobbe, 2015).

- We also found another similar study that tried to predict the market value of footballers with linear models. This study included only EPL data about footballers in the season 2017-2018. It also included the ratings of the player from Fantasy Premier League and the number of the player's Wikipedia page views. The overall formula of their model was based on the ability of the player and his performance. They regarded the number of page views as a proxy variable for ability. (Maurya, 2018).

- Another similar project was made by Yuan He. This project was the first case where the data was collected not only for one season but for 5 seasons. They divided the data into multiple parts, one for the player's personal information such as his nationality year of birth, race, height, position, and other similar attributes, and, they had a dataset aligned for the player's performance metrics such as the number of games played, the number of goals, the number of clean sheets for the goalkeepers and other metrics.

Also, this project included the national status of the players. Yuan Hees' project used mainly two models, OLS (ordinary least squares), KNN (k nearest neighbors) and Ridge, Principal Component Regressions. The authors used 10-fold cross-validation techniques and used RMSE as the main criterion for judging their models. The authors also added various performance ratios after the EDA, such as international caps to age and other interconnected attributes. The authors started to do predictions and check the accuracy of the model using multiple approaches. The approaches chosen were. Taking overall mean as a sole prediction. RMSE ≈ 54. Taking the responsive model for each player. RMSE ≈ 27.2. Using value from last year. RMSE ≈ 25.64 Training original data matrix without cross validation. RMSE ≈ 21.27 .The authors concluded that last year's value of the player made the highest contribution to the response. Goals, Assists, International caps, and other similar attributes also made a high contribution, while the personal information of the player and the ratios added by the authors did not have high significance (Y. He, 2012).

PROBLEM DEFINITION

Football players transfer from time to time from one club to another, it might be a profitable practice to analyze previous trends and make predictions based on that. To objective is to predict the transfer fee of football players based on some range of data by analyzing, interpreting, extrapolating and predicting results.

REQUIREMENT ANALYSIS

Requirement analysis involves defining, analyzing, validating, and aligning stakeholders' expectations for new projects while considering all possible conflicts. It is a process of **identifying, analyzing, and managing project requirements** to determine what the project should accomplish and eliminate any ambiguities or conflicting requirements in your project plan.

- **Functional Requirements**

They are product features or functions that the developers must implement to enable users to accomplish their tasks. Generally, functional requirements describe system behavior under specific conditions.

- **Graphical Representation of Dataset**

  Our project should interpret the dataset of transfer of football players graphically so that the relationship between features can be easily studied and effectively understood by others.

- **Conclusions**

  Main aim of our project is to conclude the findings and understanding of the dataset, so that we easily convey our interpretations to the user, without any confusions.

- **Predictions**

  Users should be able to conveniently obtain predictions like the cost of a player or the transfer fee based on information related to that specific player.

- **Non-Functional Requirements**

Nonfunctional Requirements (NFRs) define system attributes such as **security, reliability, performance, maintainability, scalability, and usability**. They serve as constraints or restrictions on the design of the system across different backlogs.

- **Accuracy**

  Numerics and graphs plotted in project should be accurate i.e., the graphs should depict and represent correct information as per the given dataset to create no discrepancies in conclusion.

- **Data Label**

  The data which is presented in the project should be properly labeled in order to avoid confusion from the user side and allow the user to explore the data himself.

- **Hardware and Software Requirements**

→ R Studio for data analysis and code implementation.

→ Google Co-Lab for providing a workspace to combine the code and collaborate with the team members.

→ Kaggle for getting all the data related information necessary for data visualization as well as getting the dataset for project.

→ Tableau for dynamic visualization used in project.

→ Data exploration libraries like ggplot, tidyverse, and others for graphical representation of data

→ A machine with a minimum of 4 GB RAM, 50 GB HDD or SSD, and a dedicated GPU for data processing and prediction making.

PROPOSED METHODOLOGY

The methods that were used or proposed for this project are quite simplified and easy to understand. We have mentioned all those methods stepwise from the start till the end.

- **Understand the dataset**

Before performing any operations on the data, we need to be familiar with the information that the datasets want to convey, more like a headline which defines the dataset well.

- **Break the problem statement into subproblems**

The original problem should be simplified further into subproblems so that it is easier to properly divide the work and then individually try to solve the subproblems one by one.

- **Scrim through the dataset**

It should be decided beforehand which columns are useful, and which all are of no use to us. So, it becomes important to scrim through the dataset to find such columns and divide our dataset accordingly.

- **Focusing on filtered data and finding relationships between features**

In order to solve a subproblem, we need to find the appropriate relationship between the columns or features and analyze those relationships one by one to conclude for a specific subproblem.

- **Graphical representation of data**

Often, Numerics do not help us understand the relationship between features. Hence, graphical representation of data is very important. Not only do they easily clarify our confusion, but also help us to find more conclusions or more subproblems, allowing us to better understand the data. Apart from that, for dynamic visualization, Tableau can also be used.

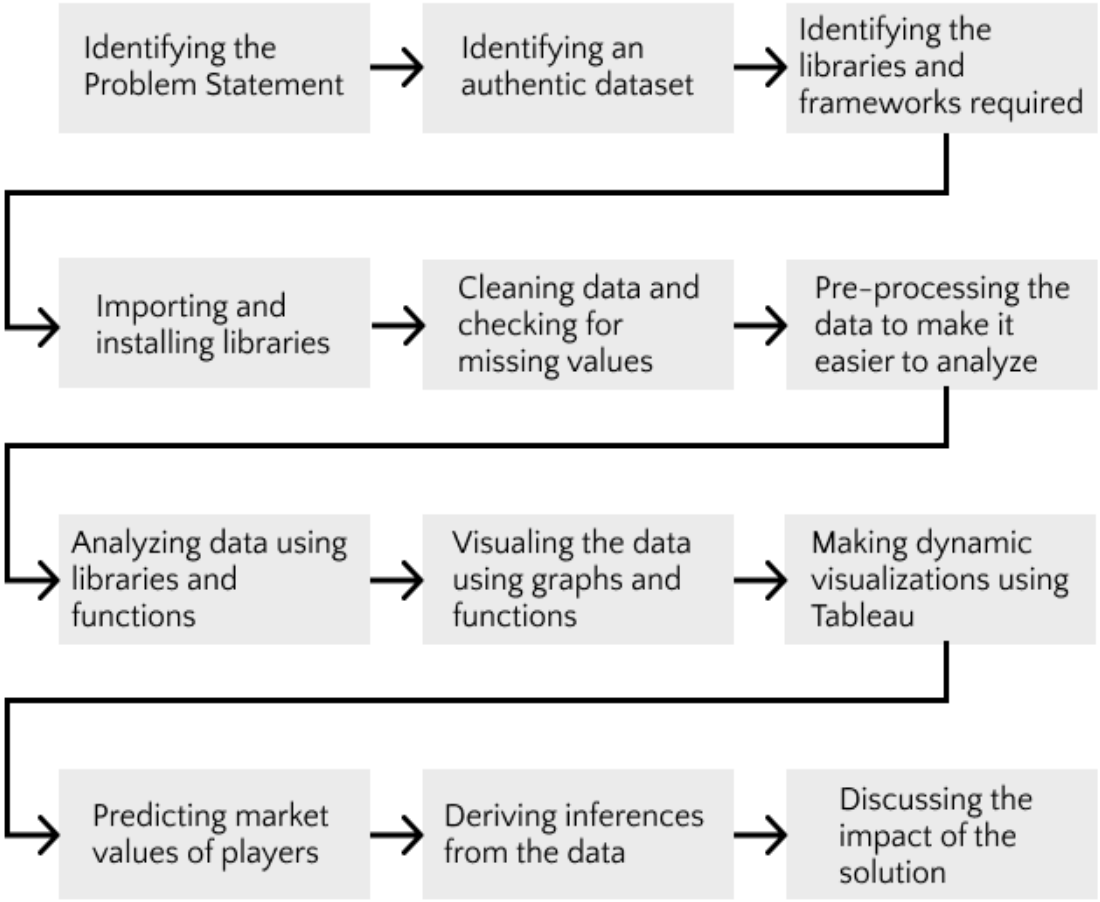
- **Conclusion**

Based on all the findings related to subproblems, we can now integrate them and create a conclusion for our original problem statement.

- **Predictions**

Now that we are familiar with the relationships between features, and have explored the data thoroughly, it is time to use those findings and information to create predications. Here, based on information of the player, like club, position, year, region, etc., we can predict the transfer fee of that player.

FLOW DIAGRAM



IMPLEMENTATION OF THE PROJECT

The project has been implemented step-by-step as per the flow diagram. The coding in R language has been done on Google Collaboratory (G Colab) notebook using the R extension. A G Colab notebook has been used in order to provide seamless collaboration and version control among all the team members. The dynamic visualizations have been performed on Tableau using the student login for efficient use. The implementation involved the following steps:

- Identifying a problem statement.
- Identifying a reliable and authentic dataset for analysis.
- Identifying the libraries and frameworks suitable for data analysis.
- Importing and installing libraries.
- Cleaning the data and checking for missing values (if any).
- Pre-processing the data to make it easier to analyze.
- Analyzing data using libraries and functions.
- Visualizing the analysis using graphs and charts by libraries.
- Making dynamic visualizations using Tableau.
- Predicting the value of football players using machine learning in R,
- Deriving conclusions and key takeaways.
- Discussing the impact of the solution.

DATASET IDENTIFICATION

The Dataset was acquired From Kaggle.

The chosen features are Rank, Origin, Player, Country, Club, Position, Fee, Year, Born. These features are ideal as they allow us to make accurate predictions of the transfer fee. We also noticed that this dataset is tailor-made for such a regression problem. Hence making it the ideal pick for our mini project.

PACKAGES AND LIBRARIES USED

The Packages that were used are as follows:

Tidyverse - The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures. See how the tidyverse makes data science faster, easier and more fun with “R for Data Science”.

Janitor - The janitor package is a R package that has simple functions for examining and cleaning dirty data. It was built with beginning and intermediate R users in mind and is optimized for user-friendliness.

GGthemes - Provides 'ggplot2' themes and scales that replicate the look of plots by Edward Tufte, Stephen Few, 'Fivethirtyeight', 'The Economist', 'Stata', 'Excel', and 'The Wall Street Journal', among others. Provides 'geoms' for Tufte's box plot and range frame.

Highcharter - Highcharter is a R wrapper for Highcharts javascript library and its modules. Highcharts is very flexible and customizable javascript charting library and it has a great and powerful API. The main features of highcharter are: Chart various R objects with one function: with hchart(x) you can chart data.

Hrbrthemes - Additional Themes, Theme Components and Utilities for 'ggplot2' A compilation of extra 'ggplot2' themes, scales and utilities, including a spell check function for plot label fields and an overall emphasis on typography.

DATA PRE-PROCESSING

There were a Few steps in Data Pre processing, they are as follows

- Acquire the Dataset, This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases.

- Import all Crucial libraries, The predefined Python libraries can perform specific data preprocessing jobs. Importing all the crucial libraries is the second step in data preprocessing in Data Science.

- Import the Dataset, In this step, you need to import the dataset/s that you have gathered for the project at hand. Importing the dataset is one of the important steps in data preprocessing.

- Initially we converted the csv file to a data frame to make it more readable.

- Then we removed the NULL values to avoid garbage values while making predictions, In data preprocessing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data.

- We also randomized the dataset to ensure we don’t have overfitting.

S.No	Club
1	Paris Saint-Germain
2	Barcelona
3	Atlético Madrid
4	Manchester City
5	Manchester United
6	Real Madrid
7	Juventus
8	Liverpool
9	Chelsea
10	Arsenal
11	Bayern Munich
12	Internazionale
13	Napoli
14	Shanghai SIPG
15	Tottenham Hotspur
16	Monaco

Table 1: Clubs with one or more expensive transfers

To Club	Total Spending in Euro(million)
Barcelona	657.
Real Madrid	634
Manchester City	590.
Paris Saint-Germain	590.
Manchester United	438.
Juventus	337
Chelsea	286.
Liverpool	207
Atlético Madrid	196
Arsenal	144.
Bayern Munich	80
Internazionale	80
Napoli	70
Shanghai SIPG	61
Monaco	60
Tottenham Hotspur	60

Table 2: Football clubs that spent the most in transfers

Rank	Player	From Club	To Club	Position	Fee(Million)
1	Neymar	Barcelona	Paris Saint-Germain	Forward	222
2	Kylian Mbappé	Monaco	Paris Saint-Germain	Forward	180
3	Philippe Coutinho	Liverpool	Barcelona	Midfielder	145
4	João Félix	Benfica	Atlético Madrid	Forward	126
5	Antoine Griezmann	Atlético Madrid	Barcelona	Forward	120
6	Jack Grealish	Aston Villa	Manchester City	Midfielder	117

Table 3: The most expensive players in transfers

Rank	Player	From Club	To Club	Position	Fee(Million)
3	Philippe Coutinho	Liverpool	Barcelona	Midfielder	145
5	Antoine Griezmann	Atlético Madrid	Barcelona	Forward	120
7	Ousmane Dembélé	Borussia Dortmund	Barcelona	Forward	105
18	Luis Suárez	Liverpool	Barcelona	Striker	82
27	Frenkie de Jong	Ajax	Barcelona	Midfielder	75
34	Sweden Zlatan Ibrahimović	Internazionale	Barcelona	Striker	69.5

Table 4: Expensive player transfers by Barcelona

Rank	Player	From Club	To Club	Position	Fee(million)
9	Gareth Bale	Tottenham Hotspur	Real Madrid	Forward	100
11	Eden Hazard	Chelsea	Real Madrid	Forward	100
12	Cristiano Ronaldo	Manchester United	Real Madrid	Forward	94
24	Zinedine Zidane	Juventus	Real Madrid	Midfielder	76
28	James Rodríguez	Monaco	Real Madrid	Midfielder	75
37	Kaká	Milan	Real Madrid	Midfielder	67

Table 5: Expensive player transfers by Real Madrid

Rank	Player	From.Club.	To.Club	Position	Fee(million)
6	Jack Grealish	Aston Villa	Manchester City	Midfielder	117.0
29	Kevin De Bruyne	VfL Wolfsburg	Manchester City	Midfielder	75.0
31	Rodri	Atlético Madrid	Manchester City	Midfielder	70.0
35	Rúben Dias	Benfica	Manchester City	Defender	68.0
36	Riyad Mahrez	Leicester City	Manchester City	Forward	67.8
39	Aymeric Laporte	Athletic Bilbao	Manchester City	Defender	65.2

Table 6: Expensive player transfers by Manchester City

Rank	Player	From Club	To Club	Position	Fee(million)
1	Neymar	Barcelona	Paris Saint-Germain	Forward	222
2	Kylian Mbappé	Monaco	Paris Saint-Germain	Forward	180
41	Edinson Cavani	Napoli	Paris Saint-Germain	Striker	64
42	Ángel Di María	Manchester United	Paris Saint-Germain	Midfielder	64
50	Achraf Hakimi	Internazionale	Paris Saint-Germain	Defender	60

Table 7: Expensive player transfers by PSG

Origin	Number of transfers	Total Fee(million)
France	9	861.2
Brazil	6	629.5
Portugal	6	515.0
Belgium	4	340.0
England	4	351.5
Argentina	3	229.6

Table 8: Transfers by country of origin

Year	Number of transfers	Total Fee(million)
2019	14	1142.0
2018	11	978.7
2017	5	538.5
2020	5	347.0
2009	3	230.5
2013	3	224.5

Table 9: Transfers year-wise

Position	Number of transfers	Total Fee(millions)
Midfielder	20	1551.6
Forward	14	1491.3
Striker	10	720.5
Defender	8	584.7
Goalkeeper	2	142.5

Table 10: Transfers by position of playing

DATA ANALYSIS AND VISUALIZATION

R is a language and environment for statistical computing and graphics. It is a GNU project which is like the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John

Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open-Source route to participation in that activity.

Some of the visualizations of dataset in R are as follows:

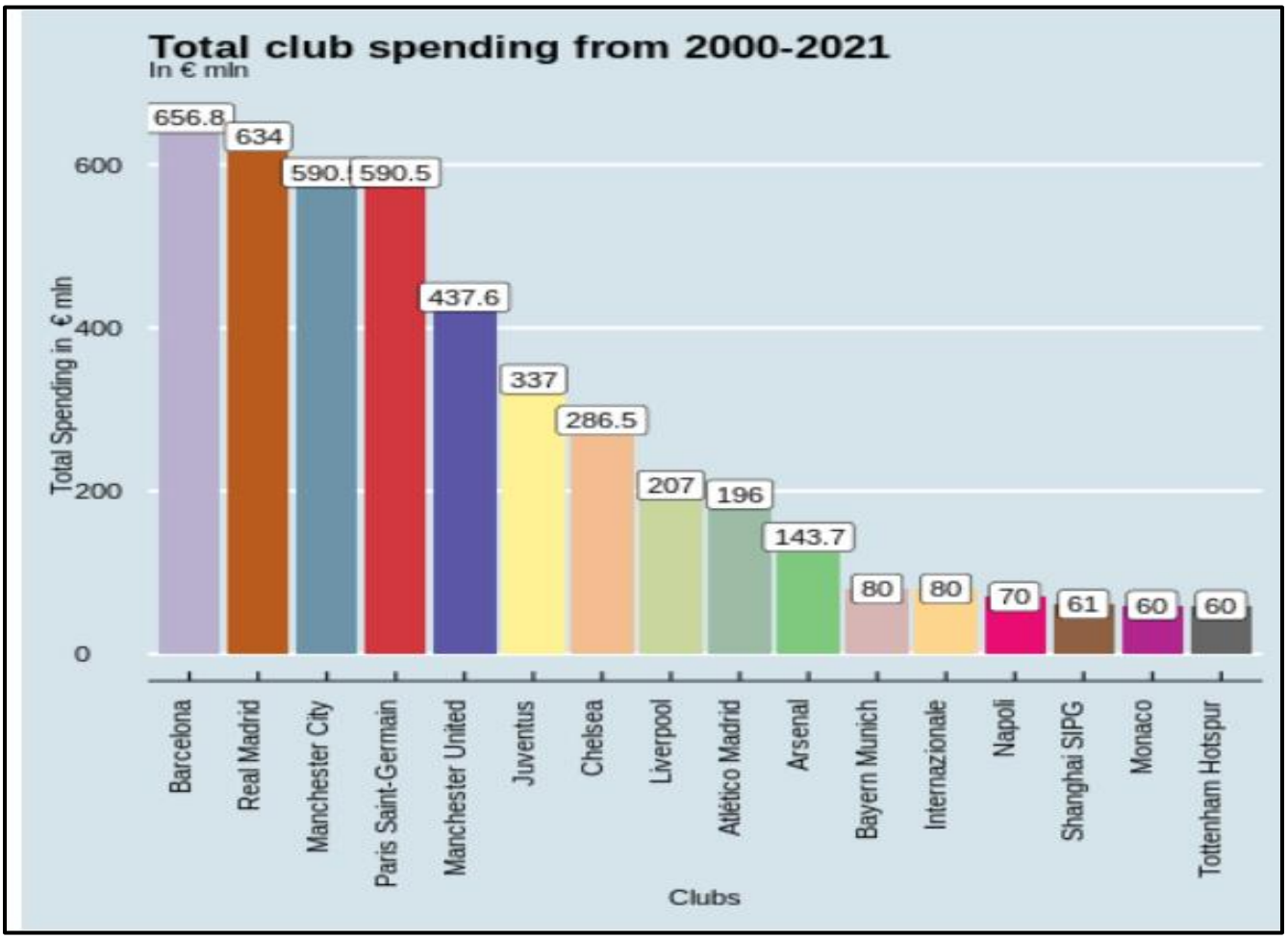


Figure 1: Total spending by football clubs from 2000-2021 on football transfers

The above graph shows that Barcelona is the club which has spent the most money in transfers, followed by Real Madrid and Manchester City.

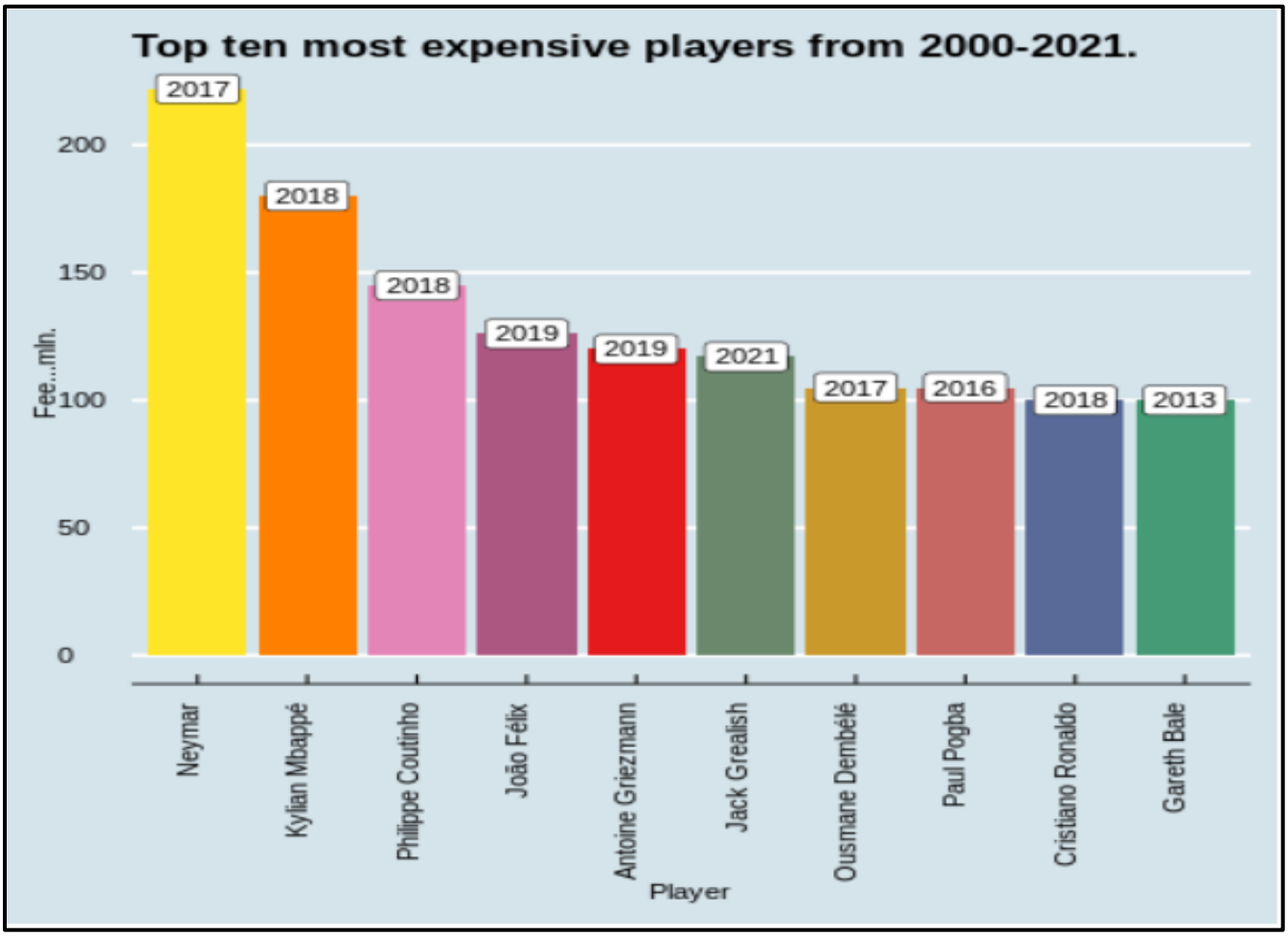


Figure 2: Top 10 expensive players

The graph tells that Neymar was the most expensive player in the transfer market in the year 2017, followed by Mbappe and Coutinho.

Now, we will see some of the expensive players from the top clubs.

Figure 3: Expensive players in Barcelona

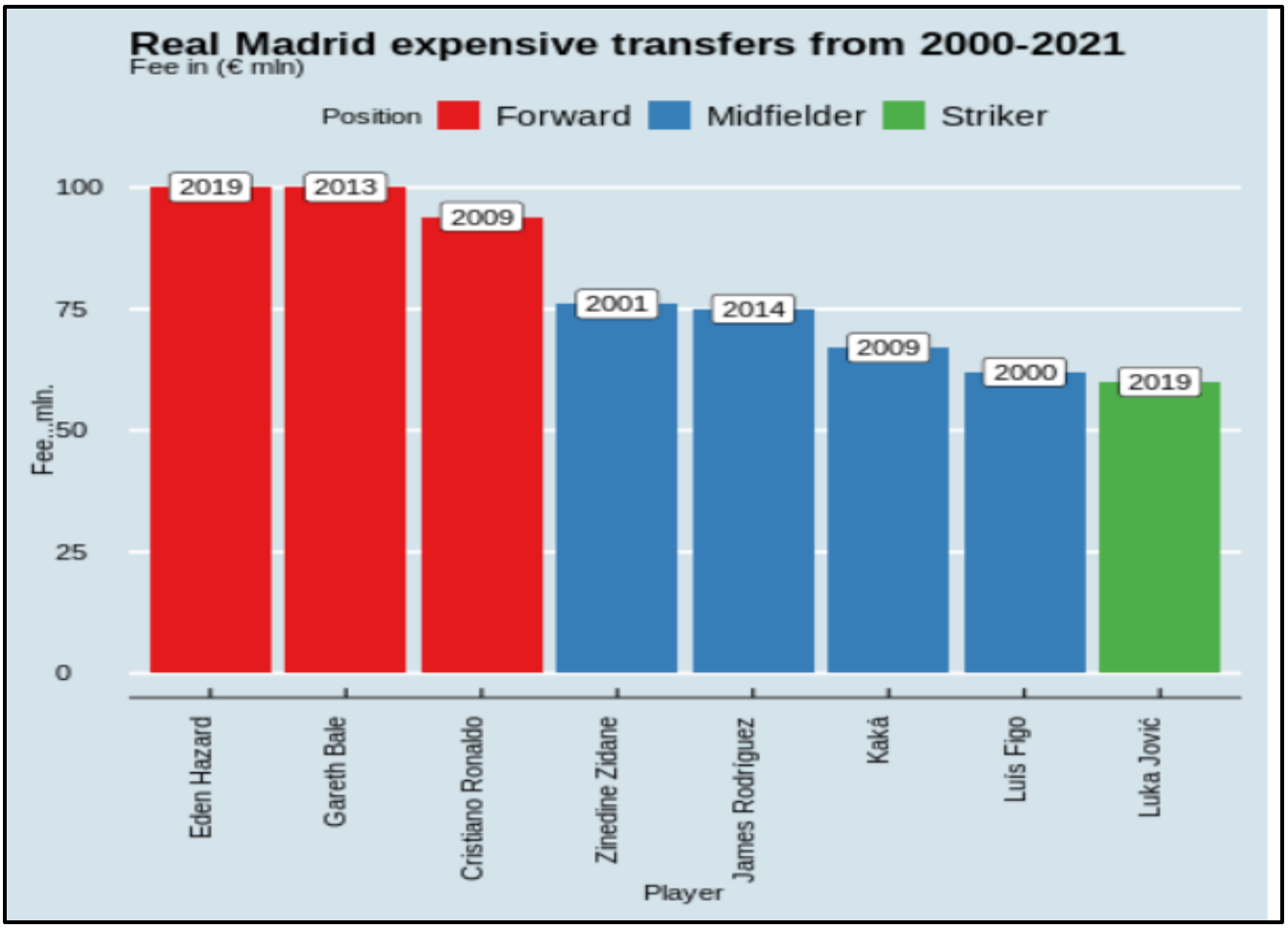


Figure 4: Expensive players in Real Madrid

Figure 5: Expensive players in Manchester City

Figure 6: Expensive players in PSG

These visualizations clearly depict the most expensive players from each club, the year in which the transfer took place, and the position to which that player belongs (midfield, attack, defense, etc.). Graphically, it is easier for us to interpret the data and come up with desired conclusions.

DATA ANALYSIS USING TABLEAU

Tableau is a data visualization tool founded with a guiding philosophy to make data understandable to ordinary people. It is considered to be the most popular visualization platform in the industry, well regarded within the business intelligence community for its ease of use and simple functionalities, which make it easy to create insightful dashboards in a few clicks. It is an end-to-end platform, designed with both data analysts and business users in mind.

Tableau has been considered in this project due to its wide scope in the industry and with the motive of learning a new technology throughout the course of this project.

The process involved connecting the server to a data source, which in our case was connecting to a Microsoft excel sheet with the data set. We can then manipulate the data using the homepage and filter out the relevant columns needed to visualize. After this, the dataset was connected to a worksheet where it was visualized using different kind of graphs. The output produced by Tableau is a dynamic chart.

These are the visualizations produced using Tableau:

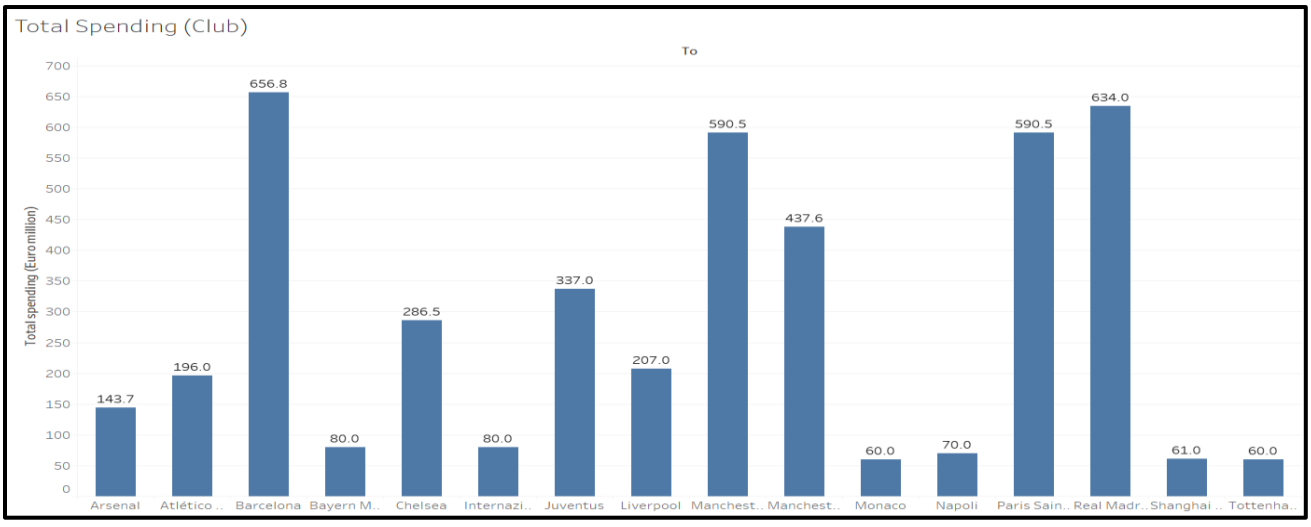


Figure 7: Total spending by football clubs from 2000-2021 on football transfers - Bar (Tableau)

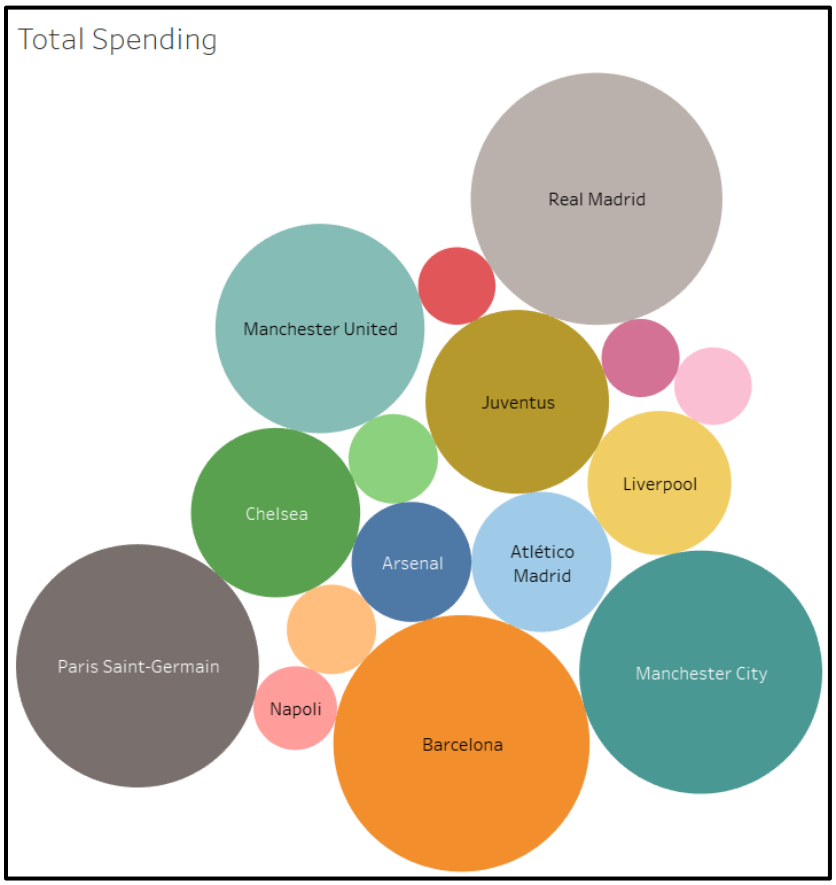


Figure 8: Total spending by football clubs from 2000-2021 on football transfers - Dot (Tableau)

The figures above show the total spending incurred by the football clubs in transfers from 2000 to 2021. As seen in the chart, Barcelona has the highest expenditure in football transfers, with Real Madrid, Manchester United, Manchester City and Paris Saint-Germain with significant expenditures. About 10 clubs have had very significant spending in the football transfers.

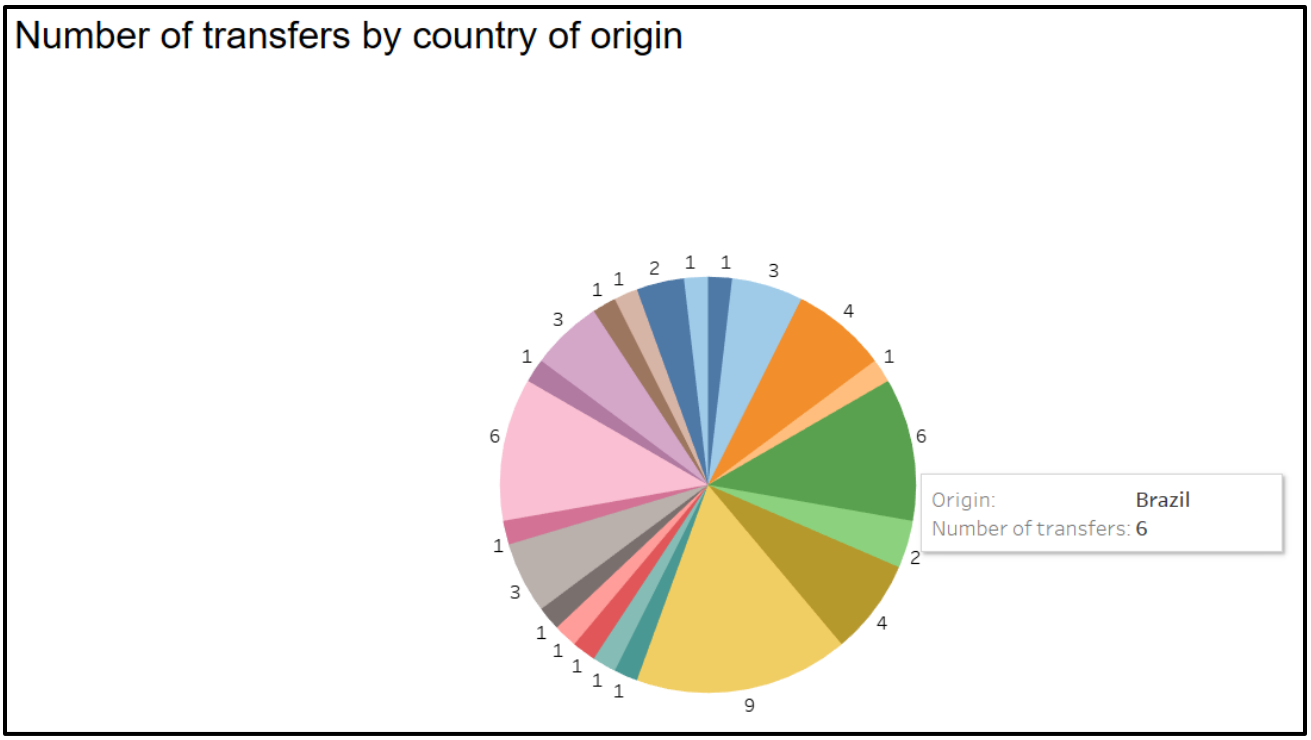


Figure 9: Number of transfers by country of origin from 2000-2021- Pie (Tableau)

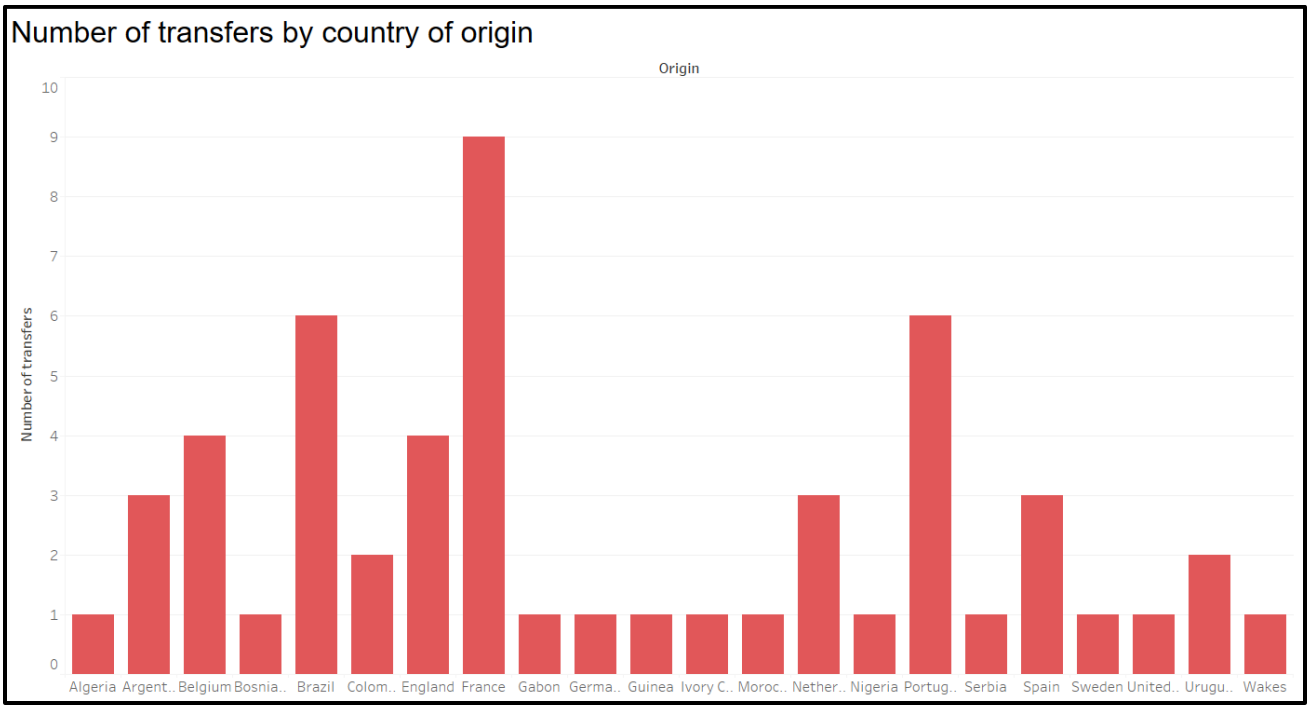


Figure 10: Number of transfers by country of origin from 2000-2021- Bar (Tableau)

The above charts mentions the total number of transfers by the country of origin. According to the pie chart, the highest number of transfers have been 9, which have originated by France. Other significant high number of transfers have been originated by Brazil, Portugal, Belgium, and England.

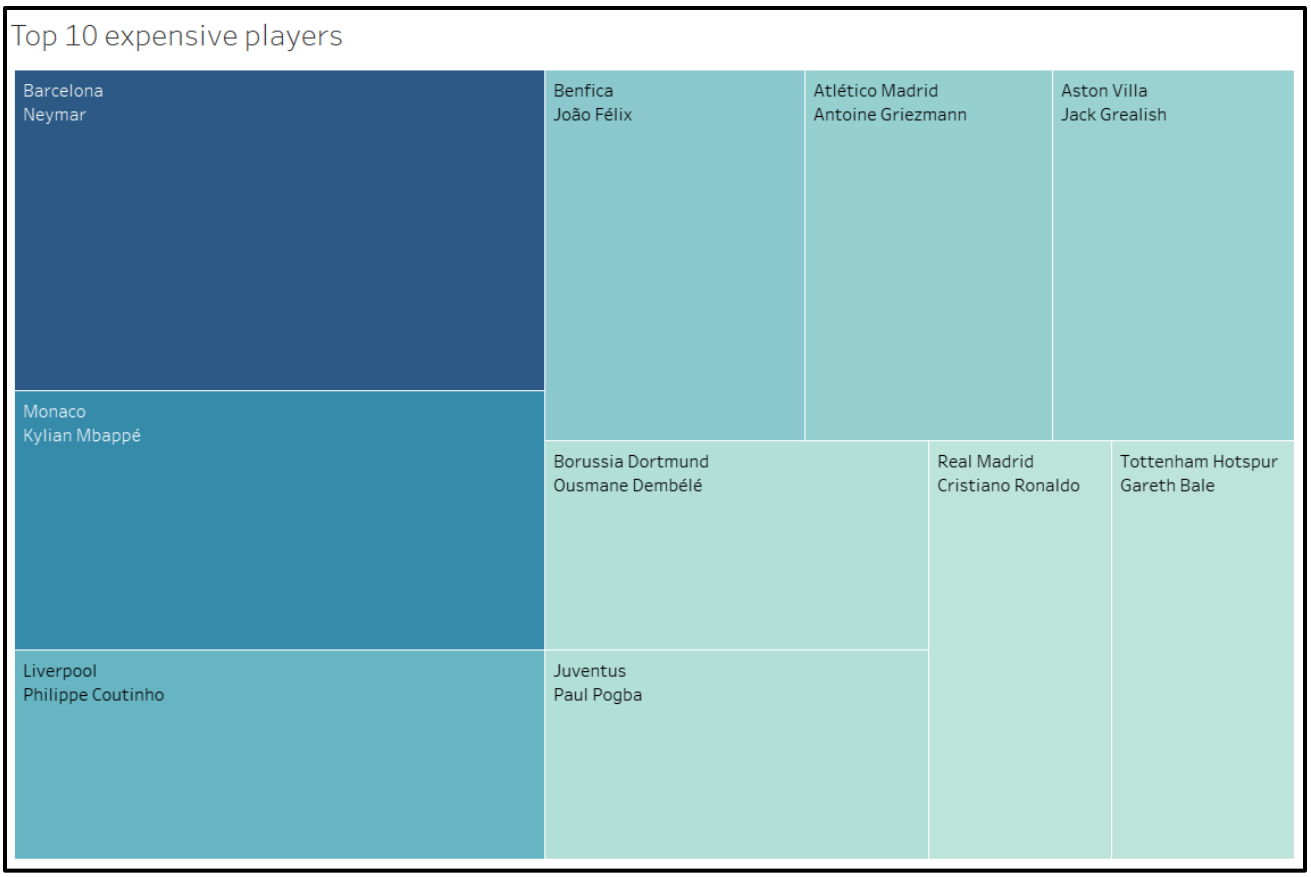


Figure 11: Top 10 expensive players – Square (Tableau)

The square plot shows the top expensive players. This means that the clubs had to incur greater expenditure in the buying of these players. The most expensive player, as clearly shown by the square plot, is Neymar from the Barcelona club with a fee of 222 euro million in 2017. Other significantly expensive players are Kylian, Phillipe, Jao and Antoine from Monaco, Liverpool, Benfica, and Atletico

Madrid, respectively. The range of the top 10 expensive players is about 122 euro million.

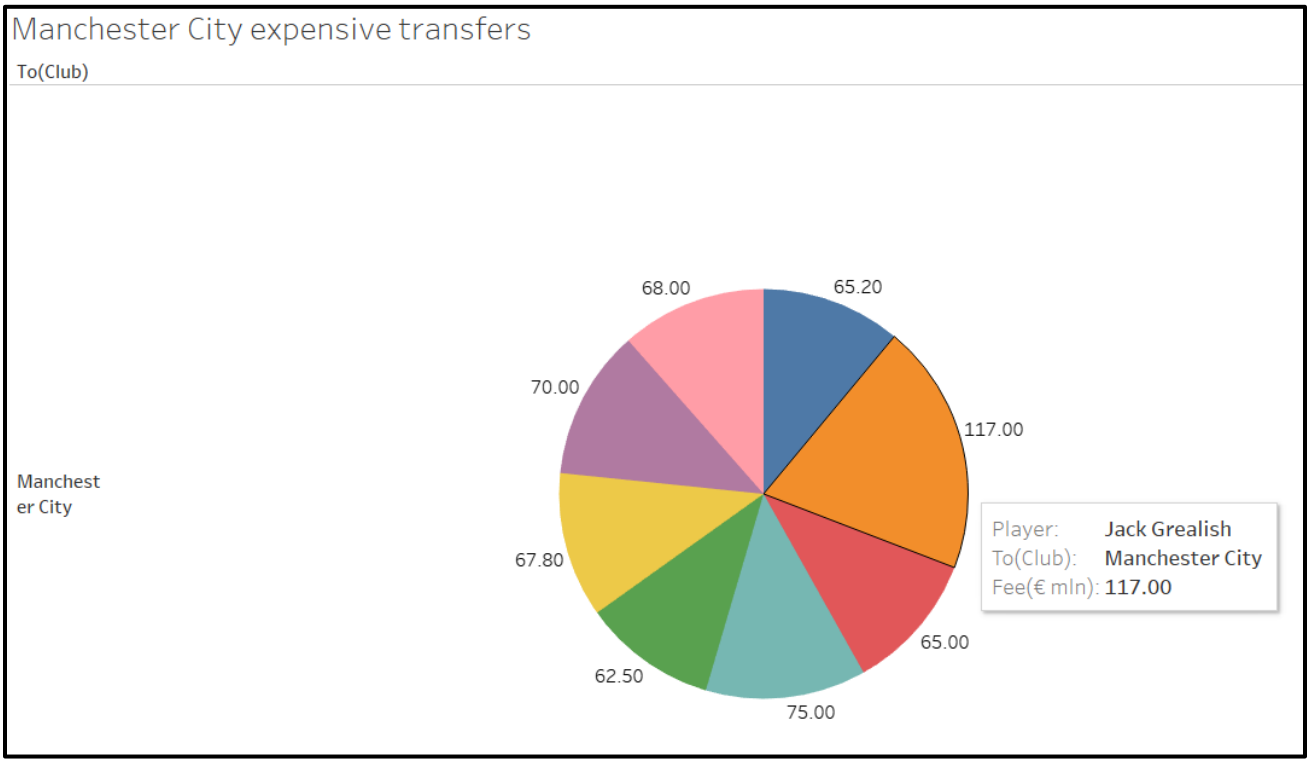


Figure 12: Manchester City expensive transfers – Pie (Tableau)

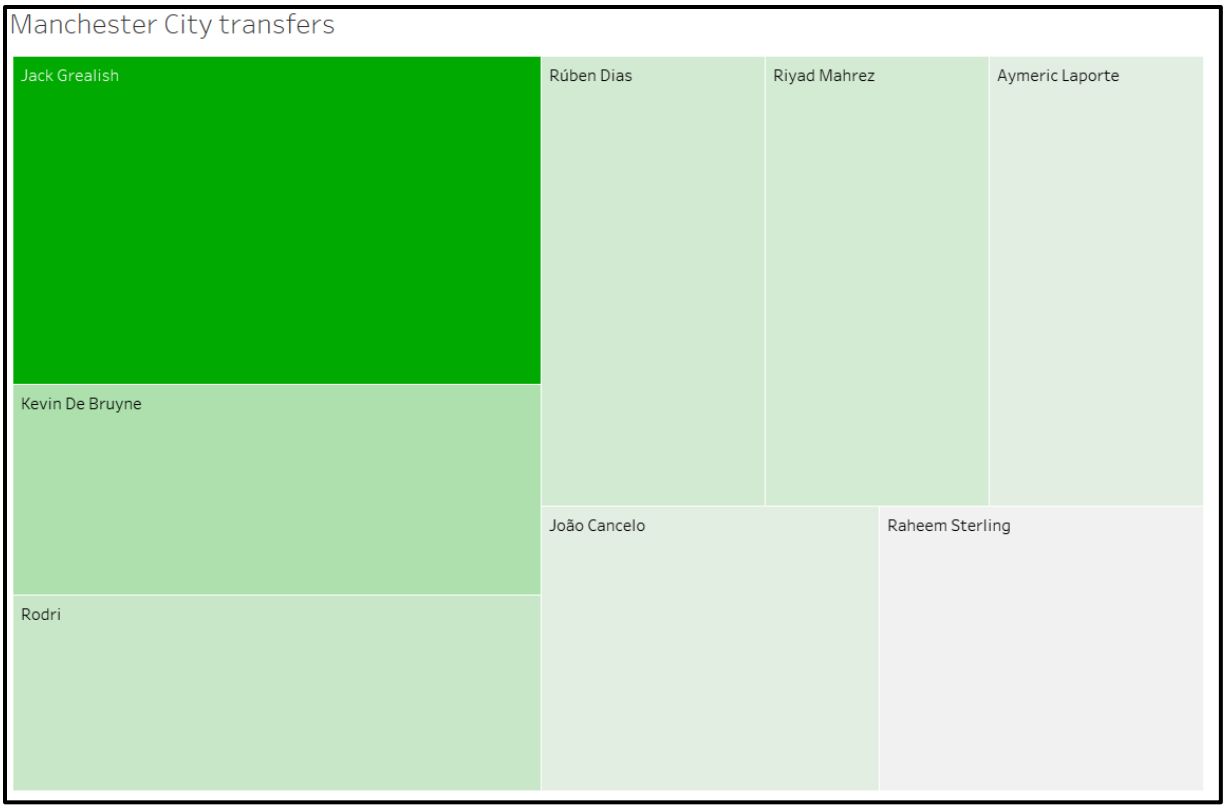


Figure 13: Manchester City expensive transfers – Square (Tableau)

The two charts above show the expensive transfers that the very famous club Manchester city incurred from 2000 to 2021. The most expensive transfer was for the midfielder Jack from the Aston Villa club for 117 euro million. Along with this, other significant expensive transfers include Kevin, Rodri, Ruben and Riyad for 75, 70, 68 and 67 euro million, respectively.

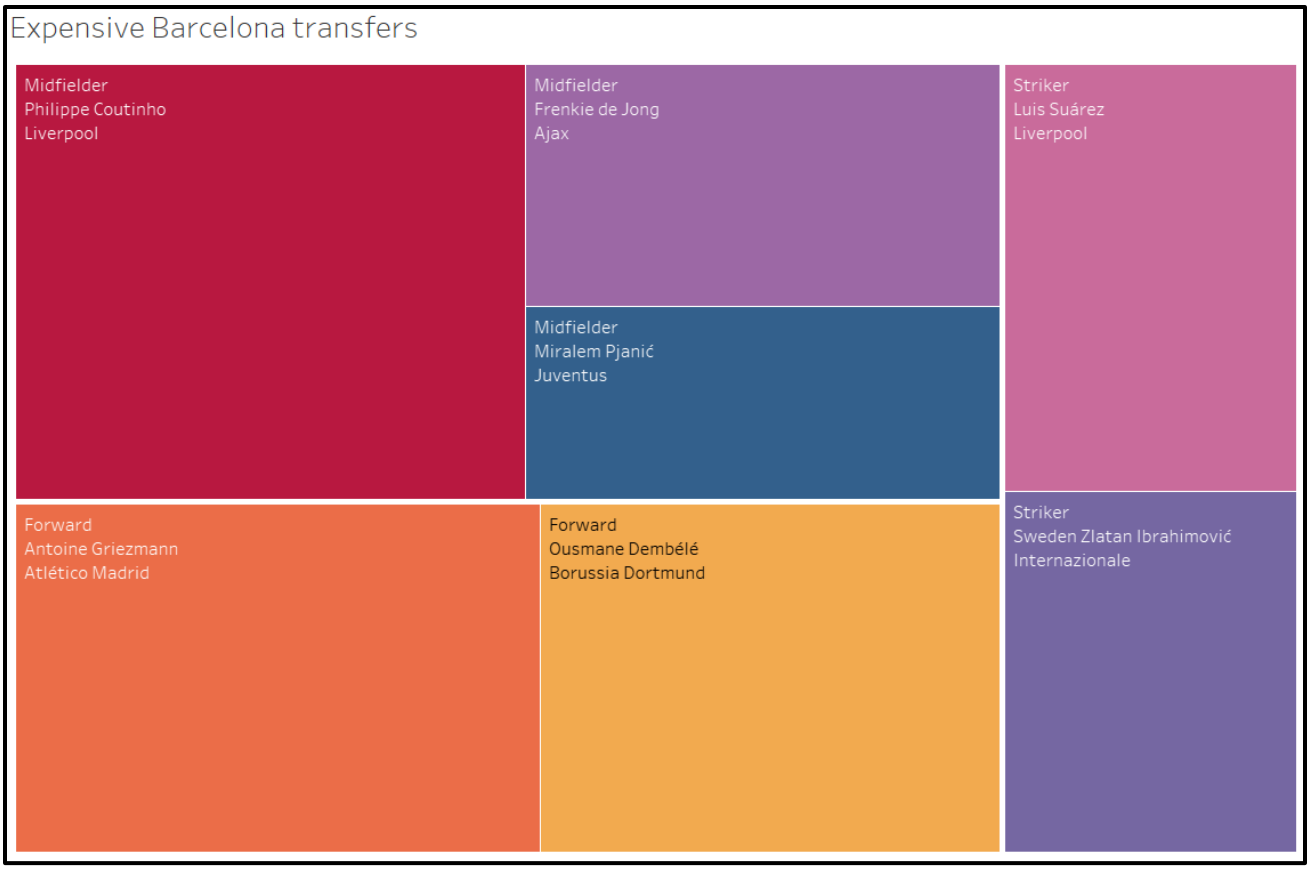


Figure 14: Barcelona expensive transfers – Square (Tableau)

The above visualization shows the most expensive transfers that Barcelona incurred. The most expensive transfer involved Phillipe from Liverpool for 145 euro million. With this, Antoine, Ousmane, Luis and Frenkie were other significantly expensive transfers for 120, 105, 82 and 75 euro million, respectively.

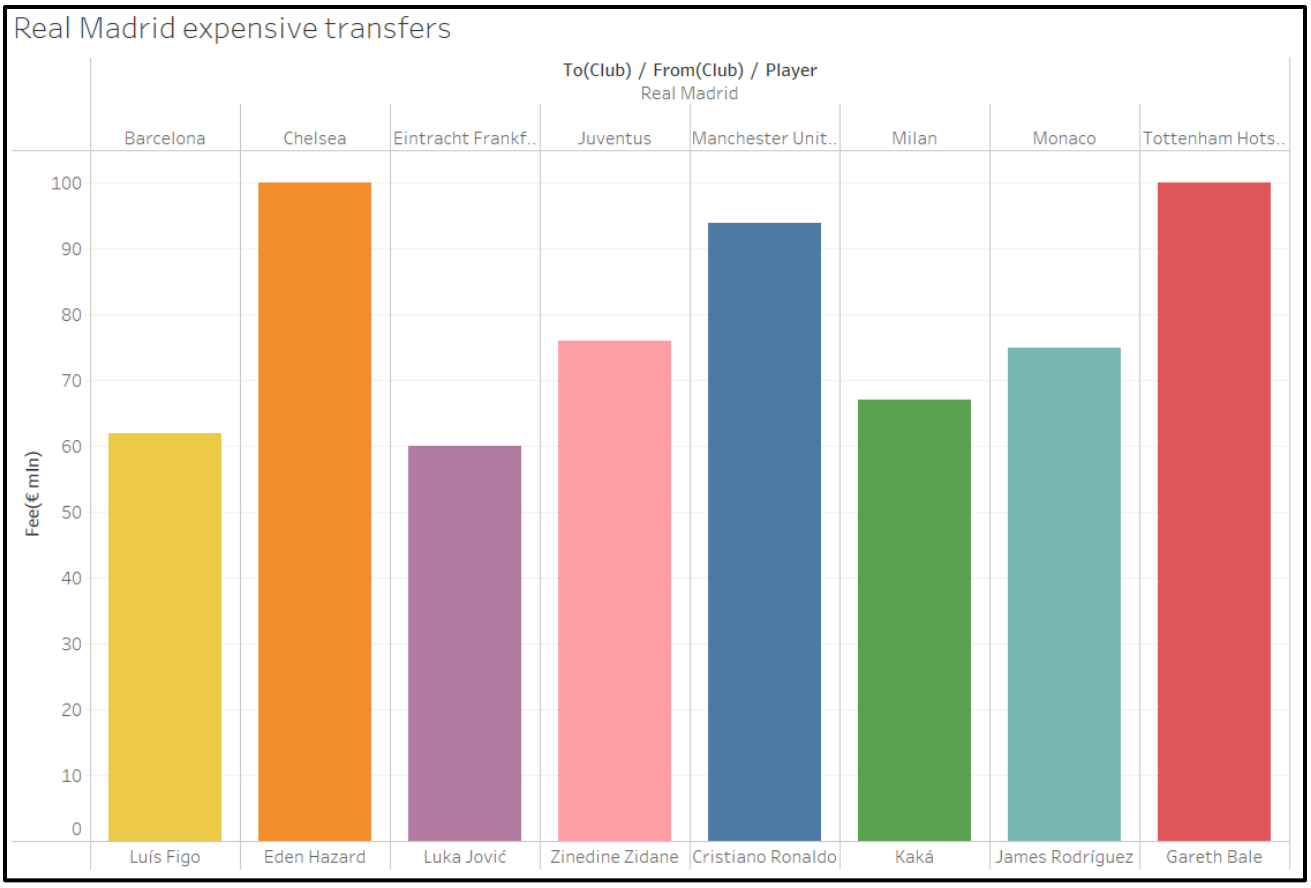


Figure 15: Real Madrid expensive transfers – Bars (Tableau)

The bar chart above shows the expensive transfers that Real Madrid incurred from 2000 to 2021. The most expensive transfer was for the forward position player Gareth from the Tottenham Hotspur club and Eden for 100 euro million each. Along with this, other significant expensive transfers include Christiano, Zinedine and James for 94, 76 and 75 euro million, respectively.

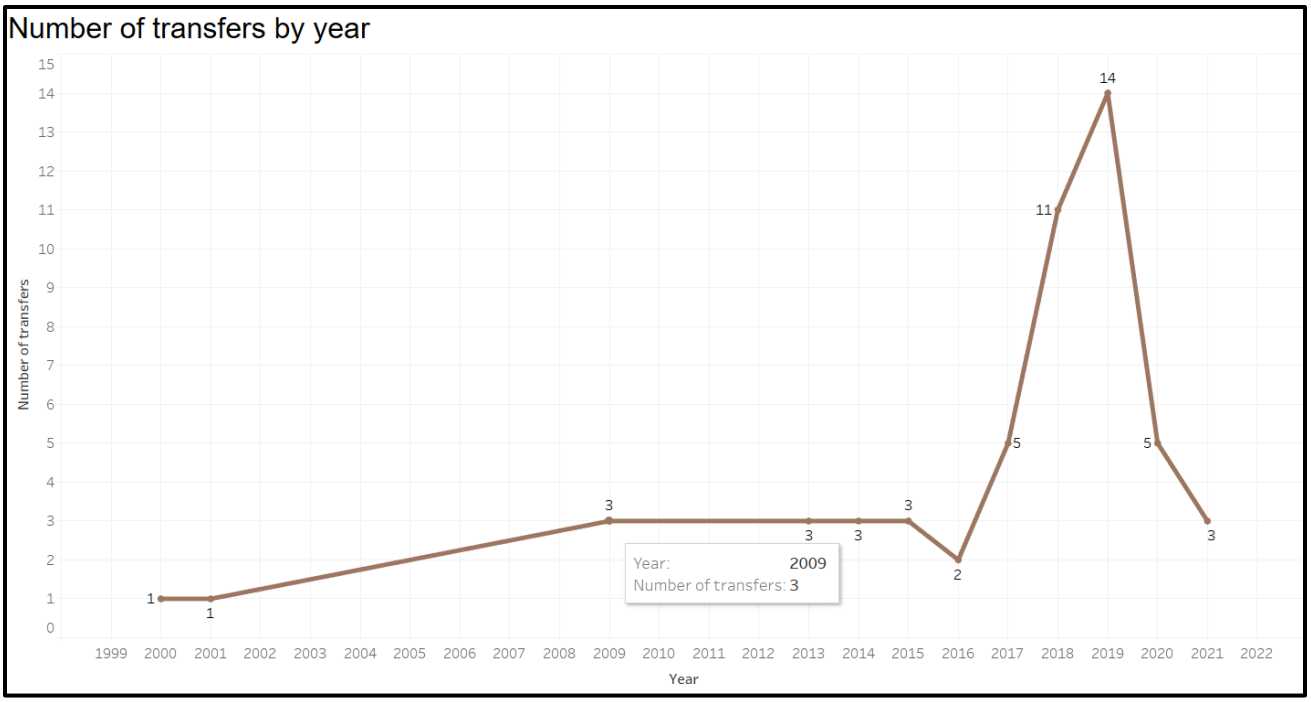


Figure 16: Total number of transfers year wise – Line (Tableau)

The line graph above shows a time-series analysis of the total number of transfers that took place from 2000 to 2021. The maximum number of transfers took place in 2019 which were 14. Other years which also have significant number of transfers are 2018, 2017 and 2020 with 11, 5 and 5 transfers, respectively.

PREDICTION USING LINEAR REGRESSION

Having the best players in your team always increases your chance of winning. But with more teams getting hands on more funds from the investments, transfer costs have started to increase exponentially every transfer window. Keeping the same in mind, it would be interesting to see if we can use Machine learning to predict the outstanding transfer fees teams spend on players.

We used a dataset from Kaggle featuring all the players from the Premier League in the season of 2017/18. The dataset contains the names of the players, their age, the position they play at, their value, nation, and age.

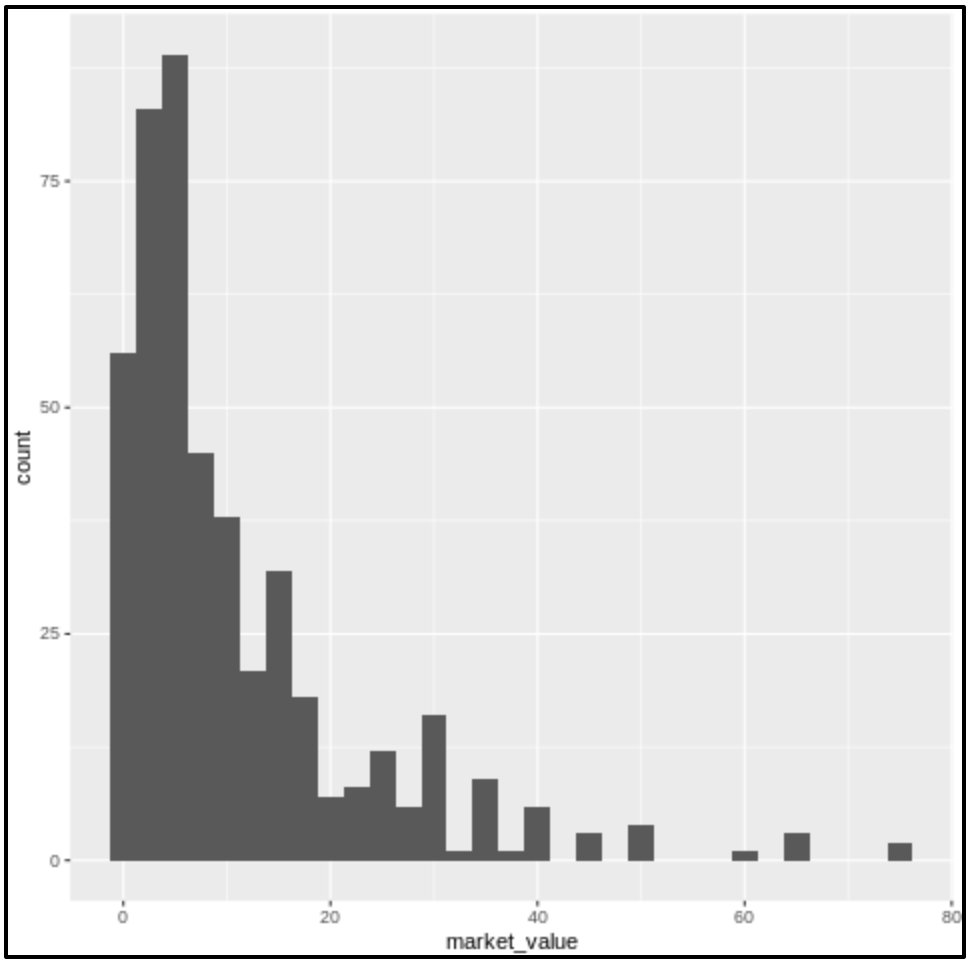


Figure 17: Distribution of market value

If we look at the distribution of market value, it can be clearly seen that most of the players have an average value of around 10 million with only a handful of players having a market value greater than 40 million. This is clearly not a normal distribution but was expected, as teams only have a few star players.

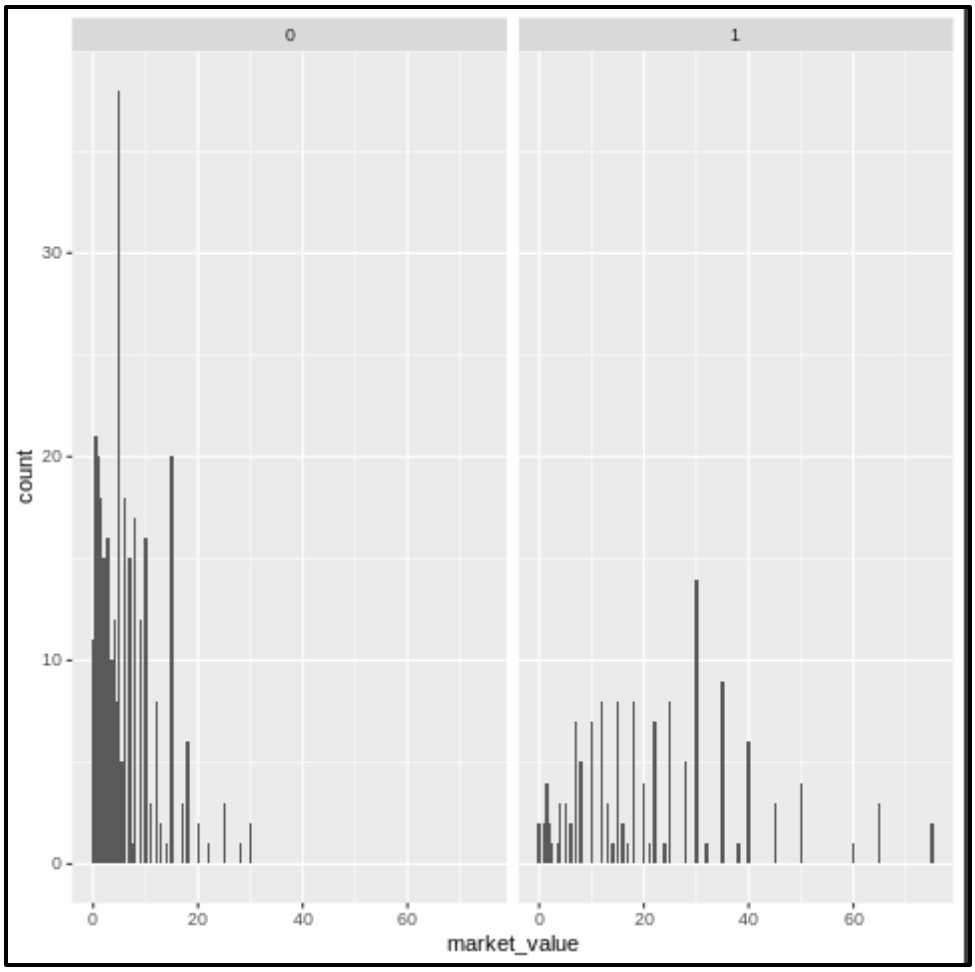


Figure 18: Distribution of market value top 6 vs the rest

The above figure compares the distribution of market in the top 6 teams of the league (Liverpool, Manchester City, Manchester United, Chelsea, Tottenham Hotspurs, and Arsenal) versus the rest of the league. Interestingly, the top 6 teams tend to have players with high and low market values, whereas the rest of the league have players with market values under 10 million, with the highest market value capping at around 35 million.

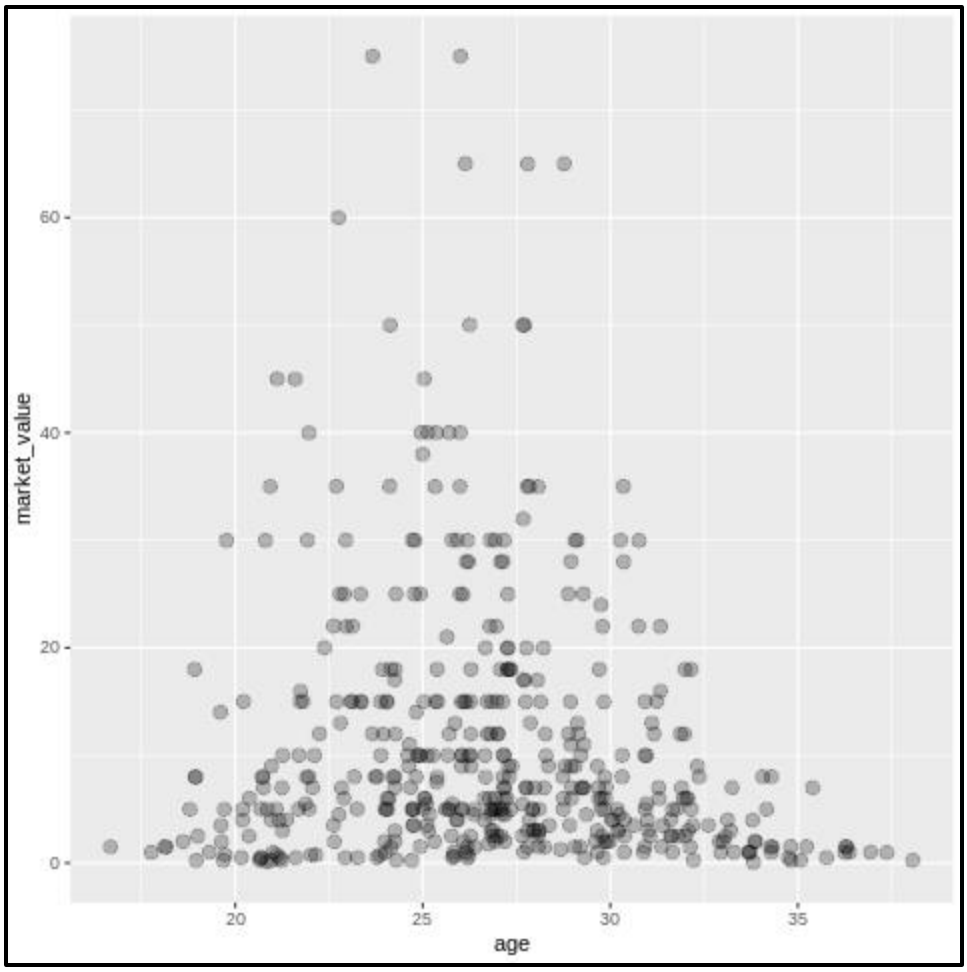


Figure 19: Distribution of market value with age

The above figure and table help us get a very important clue in finding the most important factor contributing to a player's market value, the player's age. A player's market value starts to increase from his early 20s and reaches a peak around his late 20s. As the player enters his 30s his market value starts to decrease.

Now for the prediction, we look at the number of times a player's page has been visited to find his popularity, the position he plays, his age and the stature of the club for which he plays. We used linear regression to predict the market values.

Sr. No.	Player Name	Market Value	Predicted Market Value
1	Sead Kolasinac	15	12.5
2	Alexandre Lacazette	40	21.9
3	Antonio Rudiger	25	10.4
4	Tiemoue Bakayoko	16	17.5
5	Davy Klaassen	18	9.4
6	Bernado Silva	40	21.2
7	Mo Salah	35	20

Table 11: Model predictions versus actual value

The above table shows the predicted market values compared to the actual values. The predicted values are less than the actual value because their respective all the mentioned players are not from England and subsequently would be less popular in the country.

CONCLUSION AND KEY TAKEAWAYS

Football clubs have been spending a lot of money in transfers to strengthen the team.

It is important for the football clubs to analyze their expenditure and inflow of funds to maintain funds for various other activities. These costs are covered in the revenue like stadium tickets, marketing, and broadcasting of football matches. With an understanding of the expenditure and profit, the clubs can plan out the scale of marketing and broadcasting.

A few findings from the analysis are as follows:

- **Barcelona** has incurred the highest expenditure in football transfers of 656.8 euro million over any other club from 2000 to 2021.
- The most expensive transfer has been **Neymar's** transfer from Barcelona to Paris Saint-Germain for 222 euro million in 2017.
- The **French** players have been most demanded in the market. About Nine players were purchased for a fee totaling of 861.2 euro million.
- With a total of **1142 million** euros involved in transfers, **2019** has been the year with the most transfers. There were 14 deals in 2019.
- Players from clubs that are of **a higher stature** are more likely to have a **higher market value.**
- The so called **"Big 6 clubs"** have players from low market to high market value.
- Rest of the clubs have most of the players **under 10 million** and only very few players having a market value more than 10 million.
- The market value of a player **peaks** when they are **around the age of 27**.

- Players from leagues other than the **EPL** are generally undervalued.

IMPACT OF THE SOLUTION

The analysis and predictions done in this project would help first and foremost the club. The clubs would be able to make smarter and more economical decisions regarding which player to buy and for how much they should try to get them for. It would also help the clubs manage their other expenses in a better way so that they know how much they can spend in their transfers. The visualization also helps the player realize where they stand in the current market.

APPENDIX 1

- **Project Review PPT**
  - PPT made for our project review.
  - Link

APPENDIX 2

- **Notebook with Data Analysis and Visualization**
  - Google Colab notebook with all the code required to recreate the visualizations.
  - Link

APPENDIX 3

- **Notebook with Prediction**
  - Google Colab notebook with all the code required to recreate the ML model to get the predictions.
  - Link

APPENDIX 4

- **Dataset 1**
  - Dataset used for all the visualizations.
  - Link

APPENDIX 5

- **Dataset 2**
    - Dataset used for the prediction.
    - Link

REFERENCES

Altexsoft 23 July 2021 'Requirement analysis'

CIES Football Observatory 2018 ‘Multiple Linear Regression’

Docs 2021 'R studio documentation'

Kaggle 2021 'Data related information'

Kirstie Sequitin 19 August 2021 'What Is Tableau and How Is It Used by Data Analysts?'

Lukas Barbuscak 2018 ‘Multiple Linear Regression’

R. Cachuchino A.Knobbe 2015 ‘Lasso Regression’

Shubham Maurya 2017 ‘Multiple Linear Regression’

Yuan He 2014 ‘OLS, KNN, Ridge Regression, Principal Component Regression’

Wikipedia 2021 'Exploratory data analysis'