“Decoding Real Madrid's Triumph: A Statistical Analysis of UEFA Champions League 2021-22”

Project Report

Affan Ul Ain (101443812)

School Of Electrical Engineering

Course Instructors: Pauliina Ilmonen

Department of Mathematics and Systems Analysis Aalto University, Finland

1. Introduction:

i. **Dataset Background:** The UEFA Champions League is an annual club football competition organized by the Union of European Football Associations and contested by top-division European clubs, deciding the competition winners through a round robin group stage to qualify for a double-legged knockout format, and a single leg final. The 2022 UEFA Champions League Final was the final match of the 2021–22 UEFA Champions League, the 67th season of Europe's premier club football tournament organized by UEFA, and the 30th season since it was renamed from the European Champion Clubs' Cup to the UEFA Champions League. It was played at the Stade de France in Saint-Denis, France, on 28 May 2022, between English club Liverpool and Spanish club Real Madrid. It was the third time the two sides have met in the European Cup final, after 1981 and 2018, the third final held here, after the 2000 and 2006 finals, and the first time the same two teams have met in three finals, with Real Madrid eventually thumping Liverpool to solidify their position as the club with most European title having won their 14th on that night.

ii. **Dataset Content:** This dataset contains all the key stats from the UCL 2021-22 for each player, having distributed different features in 08 different files, which are segregated on the basis of some key attributes:

- attacking.csv
- attempts.csv
- defending.csv
- disciplinary.csv
- distributon.csv
- goalkeeping.csv
- key_stats.csv

iii. **Variables Description:** To analyze this dataset containing a lot of different stats and variables, I first merged them into one table and got 42 different variables/features to analyze. So it was important to identify the most important ones which would help me to analyze better. So I decided the following 12 variables:

S. No.	Variable	Units	Description
1	Goals	Count	Number of goals scored by the player during the matches
2	Assists	Count	Number of assists provided by the player to teammates who score goals
3	Dribbles	Count	Number of times the player successfully maneuvers past opponents while in possession of the ball
4	Pass_completed	Count	Number of passes successfully completed by the player

5	Cross_accuracy	Percentage	Percentage of successful crosses made by the player out of total attempted crosses
6	Cross_attempted	Count	Total number of crosses attempted by the player
7	Pass_accuracy	Percentage	Percentage of successful passes made by the player out of total attempted passes
8	T_won	Count	Number of tackles successfully won by the player
9	Balls_recovered	Count	Number of times the player regains possession of the ball after it has been lost
10	Clearance_attempted	Count	Number of attempts by the player to clear the ball from a dangerous area
11	Saved	Count	Number of shots on goal saved by the goalkeeper during matches
12	Conceded	Count	Number of goals conceded by the goalkeeper during matches

iv. **Research Question: "***What were the key performance factors contributing to Real Madrid's success in winning the 2021-22 UEFA Champions League despite challenges, focusing on individual player contributions and team statistics?"*

The reason that motivated me to choose this research question is my strong love for Real Madrid, a club with the most decorated history in club football. Despite this illustrious history, the 2021-22 campaign was particularly unique. Real Madrid embarked on this journey as underdogs in every tie, comprising a squad with numerous young players facing a transitional phase at the club. Despite being outplayed in many matches, the team consistently capitalized on crucial moments to secure victories, even when playing below their usual standards. This led me to think whether Real Madrid's success was actually attributed to luck or if it could be backed by statistical analysis, considering both individual and team performances.

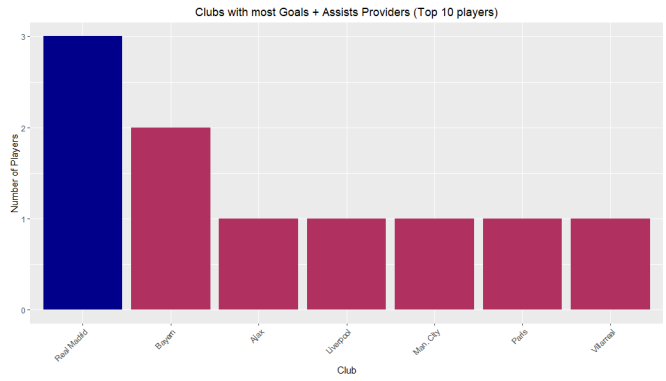
v. **Analysis Overview:** The analysis in this report will done on three different basis; Univariate, Bivariate and Multivariate Analysis. The variables defined above will be utilized in all the analysis.

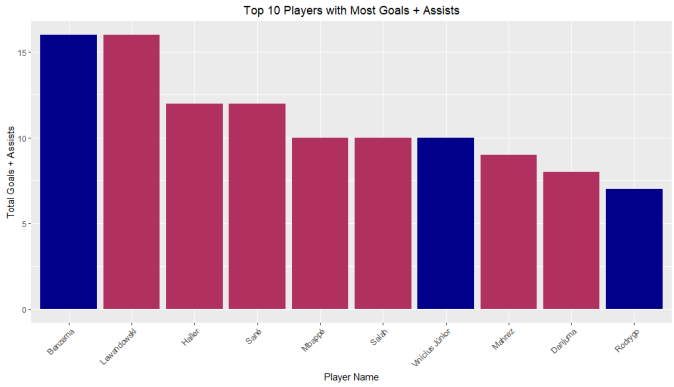
vi. **Dataset Resource:** The dataset is publicly available on Kaggle for analysis. It can be found using this link.

2. Univariate Analysis:

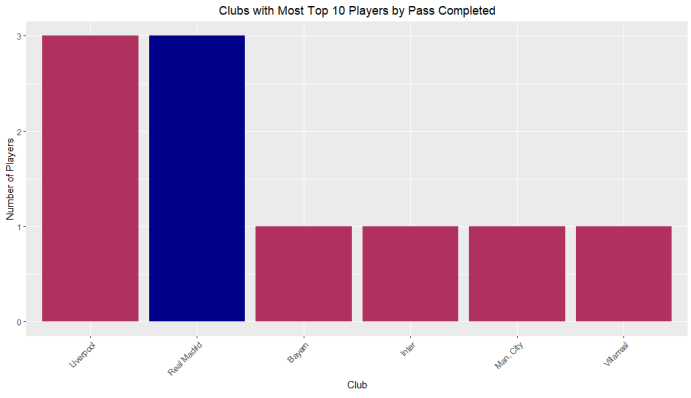
The univariate analysis is being conducted separately with respect to the different player positions: Attackers, Midfielders, Defenders, and Goalkeepers. The primary objective of these analyses is to understand whether Real Madrid as a club and which of its players dominate these key statistics. The blue bars in all the bar plots represent Real Madrid (players and club), whereas the maroon bars represent the others.

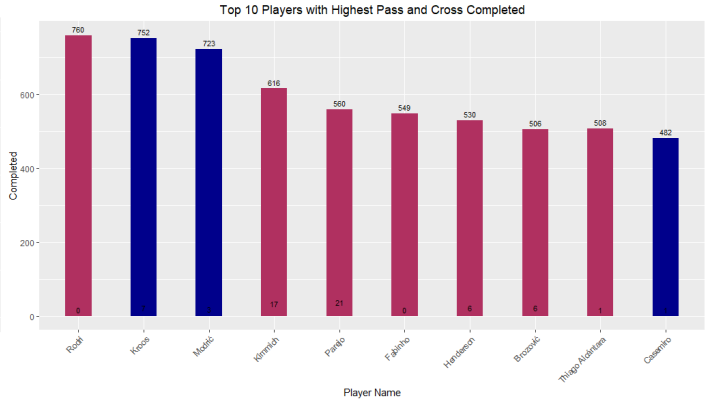
i. **Attackers:** We aim to identify the top 10 attacking players based on goals and assists, and determine which clubs have the highest number of these top 10 players. Our graph illustrates that Real Madrid contains the most, with 3 players out of the top 10 attackers: Benzema, Vinicius Jr., and Rodrygo.



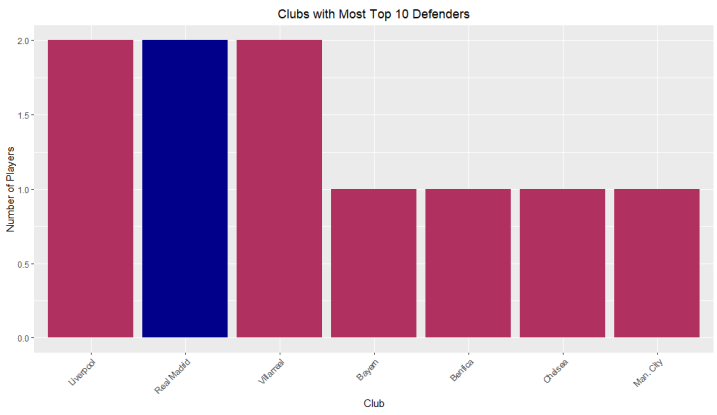


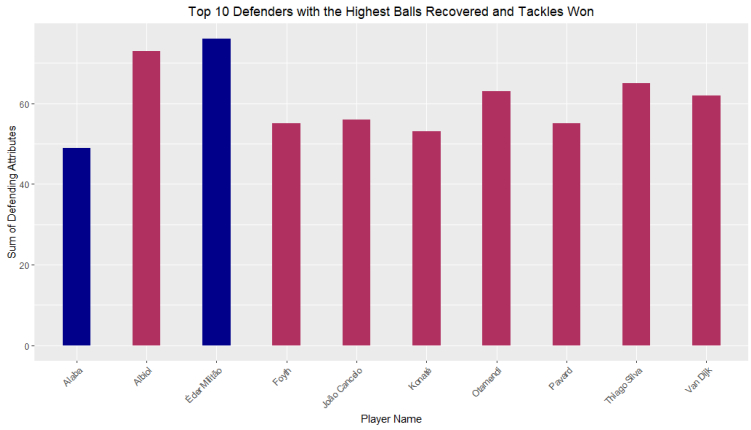
ii. **Midfielders:** We aim to identify the top 10 midfield players based on passes and crosses completed, and determine which clubs have the highest number of these top 10 players. Our graph illustrates that Real Madrid contains the most along with Liverpool, with 3 players out of the top 10 midfielders: Kroos, Modric, and Casemiro.



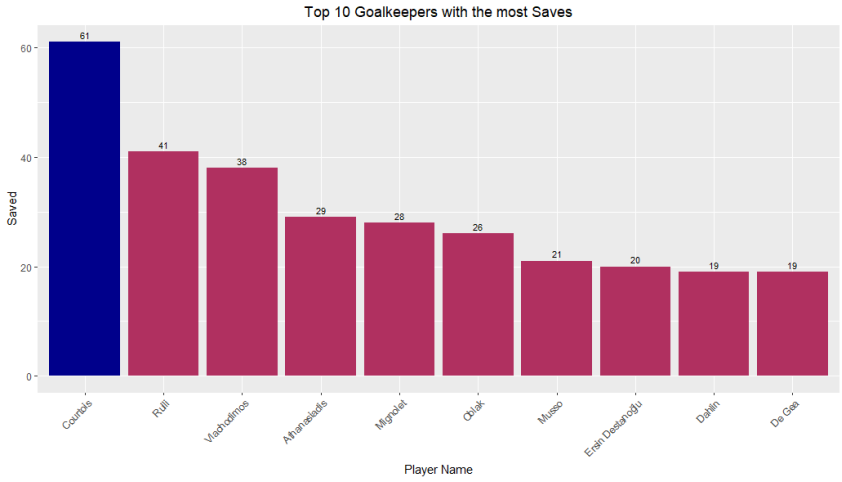


iii. **Defenders:** Following the same pattern, we aim to identify the top 10 defenders based on highest balls recovered and tackles won, and determine which clubs have the highest number of these top 10 players. Our graph illustrates that Real Madrid contains the most along with Liverpool and Villarreal, with 2 players out of the top 10 defenders: Alaba and Militao.





iv. **Goalkeepers:** Now, we aim to identify the goalkeeper with the most saves, and it's no surprise that it was Thibaut Courtois from Real Madrid with the highest number of saves, demonstrating his dominance throughout the tournament.



The univariate analysis for all player positions demonstrates Real Madrid's dominance, as they are the only club featured in all attributes with the highest number of players. We are on track to understand Real Madrid's dominance, not only through results backed by luck but also through statistics and numbers.

3. Bivariate Analysis - Pearson correlation:

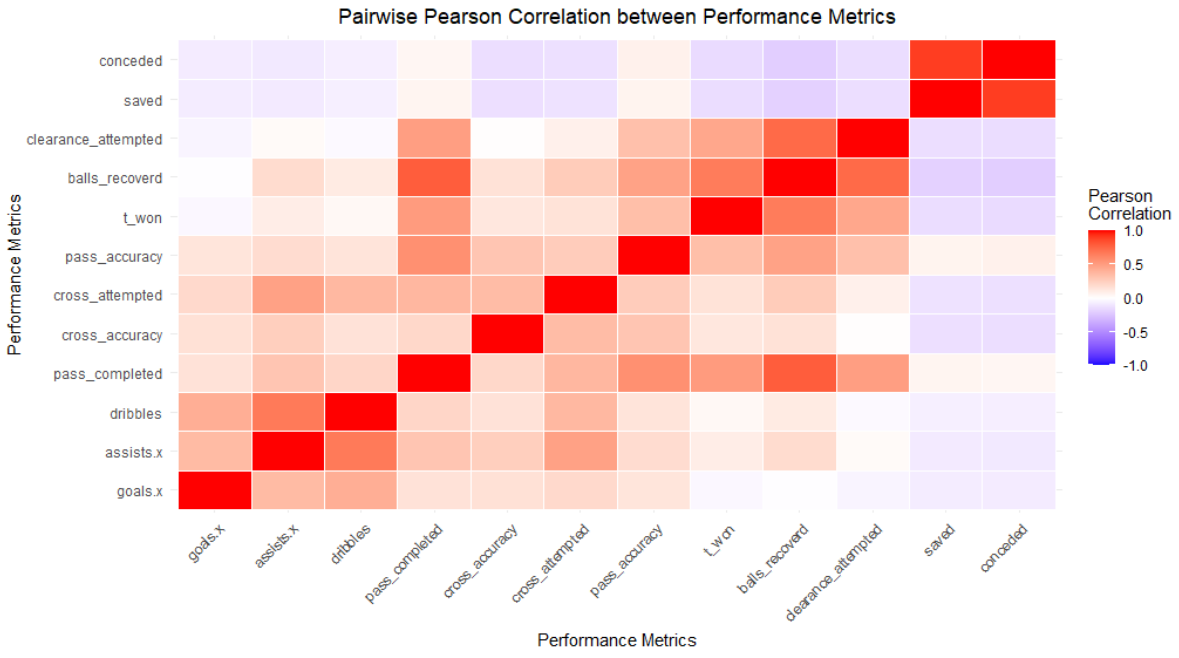
In the bivariate analysis, we aim to examine the pairwise relationships between the selected performance metrics of players using Pearson correlation. The correlation matrix generated provides insights into the strength and direction of the relationships between these performance metrics. We visualize the correlation matrix as a heatmap to facilitate interpretation, with warmer colors indicating stronger positive correlations, cooler colors indicating stronger negative correlations, and neutral colors indicating little to no correlation.

i. **Strong Positive Correlation:** The plot correctly illustrates attributes within player positions, with variables for the same positions showing strong positive correlations. For example, attacking attributes such as goals, assists, and dribbles all exhibit strong correlations, as do midfield, defending, and goalkeeping attributes.

ii. **Moderate Postive Correlation:** The plot indicates that all players also closely correlate with other positional attributes. For instance, midfield attributes like pass_completed, cross_accuracy, cross_attempted, and pass_accuracy demonstrate intermediate positive correlation with attacking and defensive attributes such as goals, assists, dribbles, tackles won, balls recovered, and clearance attempted. This clearly indicates that midfielders contribute significantly to both attacking and defensive roles.

iii. **Weak or Zero Correlation:** Since attackers and defenders typically operate at opposite ends of the formation, they usually don't share similar key attributes. This is evident in the weak or zero correlation between attacking attributes (goals, assists, and dribbles) and defensive attributes (tackles won, balls recovered, and clearance attempted).

iv. **Strong or Negative Correlation:** As all on-field players tend to contribute both defensively and offensively as per tactical and game requirements, except for the goalkeeper whose core role is to protect the goal post, the correlation matrix appropriately shows that goalkeeping attributes (saved and conceded) exhibit strong negative correlations with all other key performance positional attributes, except for passes_completed and pass_accuracy.



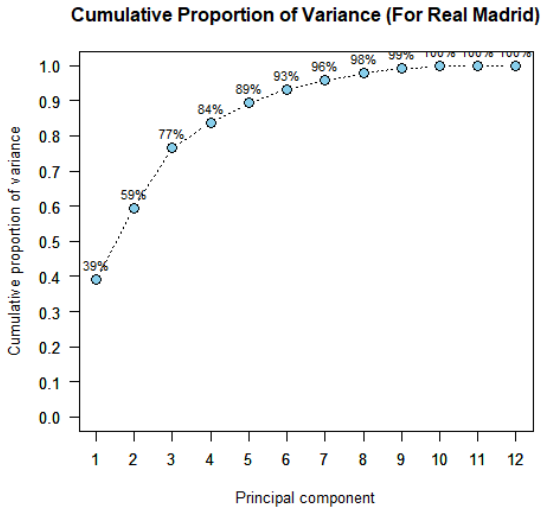
4. Multivariate Analysis – Principal Component Analysis:

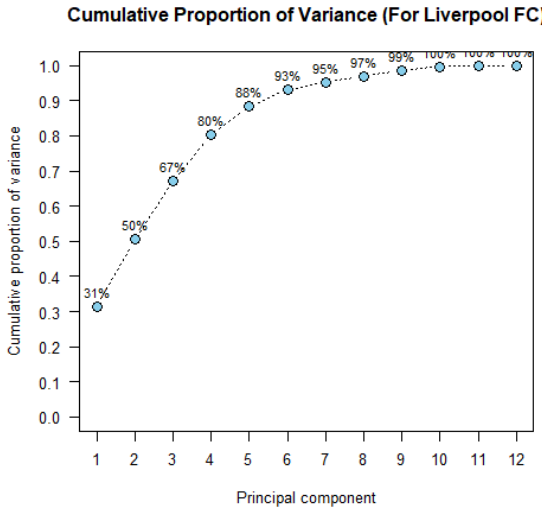
Method Selection: I have employed Principal Component Analysis (PCA) with the correlation matrix as my multivariate analysis method because I am analyzing player performance data with multiple variables related to different aspects of player performance. My dataset contains variables measured on different scales, making it essential to standardize them for fair comparison; hence, I opted for PCA with the correlation matrix. Moreover, since our interest lies in understanding the relative relationships between performance metrics for Real Madrid's success, PCA with the correlation matrix emphasizes these relationships, focusing on patterns of correlations rather than absolute magnitudes. Furthermore, it yields interpretable principal components, capturing underlying factors driving player performance. Additionally, PCA is typically used for continuous variables, ensuring its suitability for our analysis.

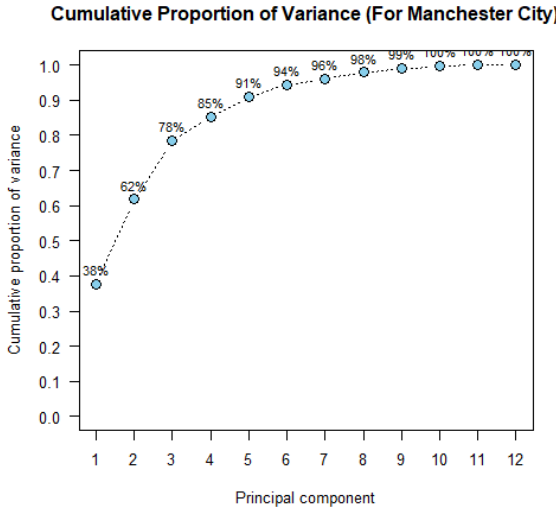
Analysis Approach: Since we have already reduced our performance metrics from 42 to 12 factors, we possess data for 747 players who participated in the tournament. Analyzing all of them simultaneously proved to be challenging. Therefore, we have narrowed our focus to the four semi-finalists of the tournament. We will conduct individual analyses of their performances to identify patterns in their gameplay. This multivariate analysis will focus on the tactical aspects of the gameplay of all the 04 teams.

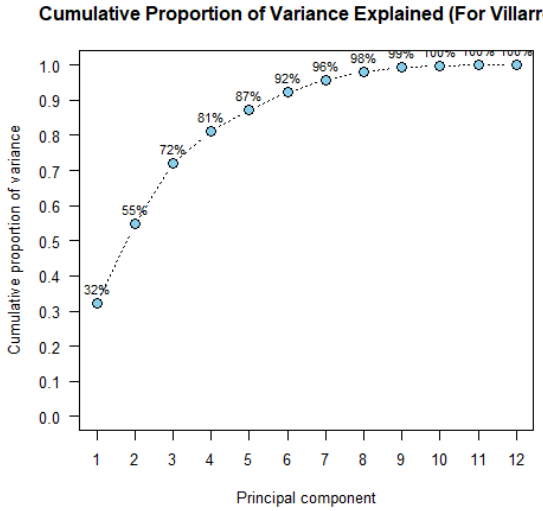
i. **Analysis on Variation of Gameplay:** We have plotted a scree plot for all four semi-finalists: Real Madrid, Liverpool, Manchester City, and Villarreal, to understand how many principal components define the majority of variations. Through this analysis, we can identify clubs with more homogeneous gameplay patterns (e.g., characterized by fewer dominant components) versus those with more diverse gameplay styles (e.g., requiring more components to explain their variance). For instance, if a club's scree plot shows a steep drop in the beginning, it means that a lot of the differences in their gameplay can be explained by just a few factors, such as scoring goals or making assists. This suggests that their gameplay is more straightforward or predictable.

As observed in the graph below, the differences between the scree plots of all the clubs aren't significant, as all the clubs exhibit similar steepness of the curves. Real Madrid and Manchester City start with more variation explained by the first two principal components compared to Liverpool and Villarreal, indicating that they were more consistent and predictable with their game plans throughout the tournament.





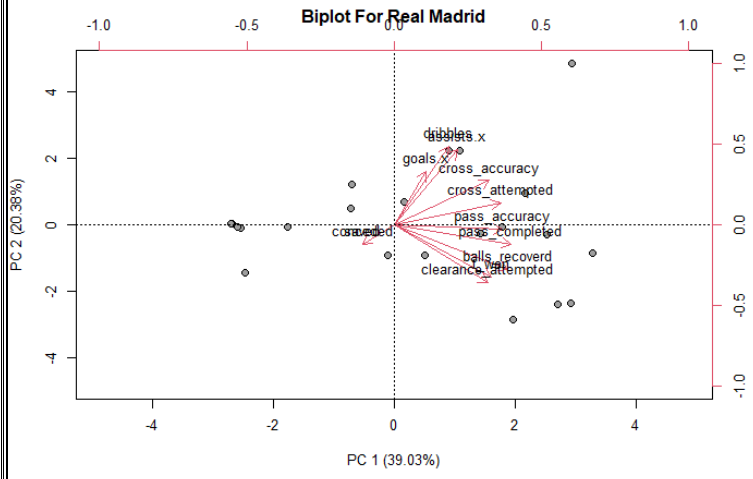


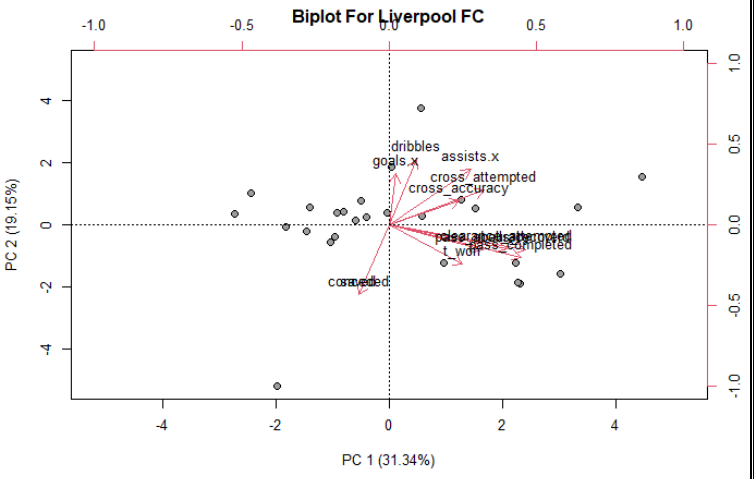


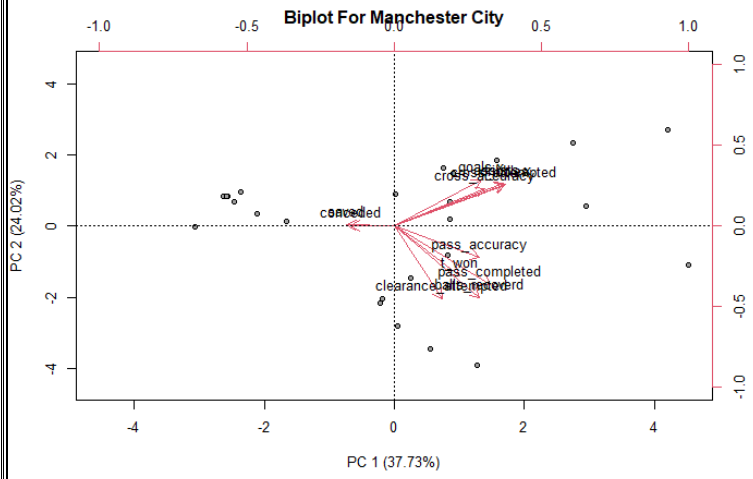
ii. Analysis on Style of Gameplay:

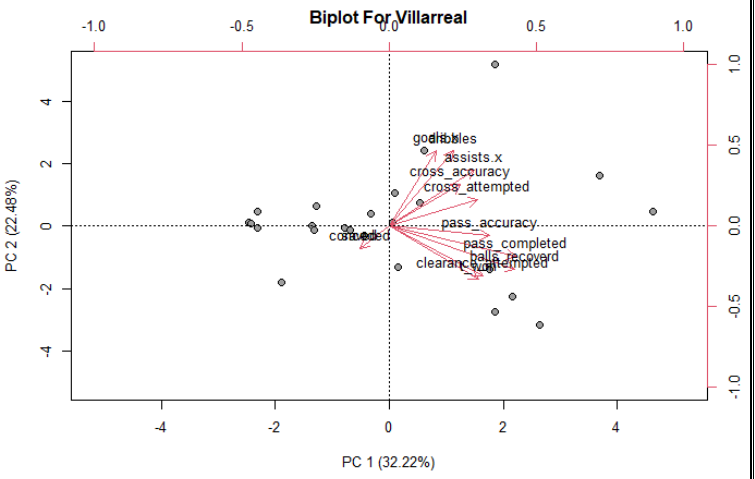
From the biplot, firstly we can analyze the relationships between the players' attributes and the principal components. Each point on the biplot represents a player, and the direction and distance of the point from the origin indicate how strongly and in what direction the player's attributes contribute to the principal components. So, this can help us analyze which players contribute the most for teams among these performance metrics, but we have already looked at it in the univariate analysis, so we won't be focusing on that anymore here.

Secondly, we can comment on the overall gameplay of teams. The direction of the arrows indicates the relationship between the performance attributes and the principal components. Arrows pointing in the same direction as the principal component axes suggest that the corresponding attributes strongly contribute to those components. The length of the arrows represents the strength of this contribution; longer arrows indicate stronger contributions.









A. **For Real Madrid:** The longest vectors in the direction of PC1 are for defensive and midfield attributes, whereas the attacking attributes mostly change in the direction of PC2 and have comparatively shorter vectors. This shows that their midfield and defense have been their strengths throughout the tournament, and they have employed possession-based controlled football. The shortest conceded vector demonstrates the strength of their goalkeeper, who has been the best among all teams.
B. **For Liverpool:** The longest vectors in the direction of PC1 are for defensive and midfield attributes, whereas the attacking attributes mostly change in the direction of PC2 and have comparatively shorter vectors. This indicates that their midfield and defense have been their strengths throughout the tournament, and they have employed possession-based controlled footballing tactics. Also, their conceded vector is the longest among all the teams, which shows that they have conceded the most goals.
C. **For Manchester City:** The longest vectors in the direction of PC1 are for attacking and midfield attributes, whereas the defensive attributes mostly change in the direction of PC2 and have comparatively shorter vectors. This indicates that their midfield and attack have been their strengths throughout the tournament. The distance between the group of vectors is greater, suggesting they have employed one-dimensional attacking footballing tactics. Also, their conceded vector varies only in the direction of PC1 and has a comparatively shorter length than Liverpool, indicating that their goalkeeper has performed better than the Liverpool keeper.
D. **For Villarreal:** The longest vectors in the direction of PC1 are for defensive and midfield attributes, whereas the attacking attributes mostly change in the direction of PC2 and have comparatively shorter vectors. This shows that their midfield and defense have been their strengths throughout the tournament. Also, their conceded vector is comparatively shorter and has a shorter length than Liverpool, indicating that their goalkeeper has performed better than the Liverpool keeper.

From a tactical perspective, I believe Real Madrid was very flexible with their gameplay in different situations, as their performance vectors are evenly distributed and have almost equal distances between them. This contrasts with other clubs, which have them clustered. I believe it's the flexibility and the entire team putting defensive and attacking shifts as per the game requirements that have made them superior to their counterparts in various situations.

5. Critical evaluations:

While this report has attempted to analyze the performance metrics of team performances, both based on individual performances and overall team tactics, I feel there are a few aspects that I observed and analyzed as a viewer but couldn't address here. One of the most significant factors for Real Madrid was the timing and impact of substitute players, as they made comebacks in the Round of 16 against Paris Saint-Germain, the Quarter Final against Chelsea, and the Semi Final against Manchester City. Due to the limitation of missing data for this metric, I believe it could have further solidified our analysis and supported our research question. Additionally, another remarkable observation I made as a spectator was the average squad age. As I mentioned earlier, this campaign was unique because the club was in a transition phase, with the team consisting mostly of youngsters who outperformed the senior competitors in rival teams. Moreover, some aspects of the analysis, such as the interpretation of gameplay tactics, are subjective and can vary from other people's perspectives.