Journal of Mathematics and Applied Statistics
Vol. 2, No. 1, June 2024.
DOI: https://doi.org/10.35914/mathstat.v2i1.179

Multivariate Analysis of the Attributes of a Central Attacking Midfielder

Nwokike Innocent Chukwudozie[1*], Erumaka Ephraim Ngozie[1], Nwaigwe Chrysogonus[2], Obi Martin Chuks[1]

[1] Department of Mathematics, Federal University of Technology, Owerri, Nigeria
[2] Department of Statistics, Federal University of Technology, Owerri, Nigeria

**Correspondent Email: chukwudozienwokike@gmail.com*

ABSTRACT

Central attacking midfielders are one of the most vital squad members because of their decisive contributions to the team in the field of play. With teams now acquiring some talented players for as much as 103 million euros, it has become more important than ever to critically investigate the attributes of a central attacking midfielder. This study is designed to investigate the correlation between attributes (variables) such as crossing, finishing, short passing, volleys, dribbling, curve, long passing, ball control, acceleration, sprint speed, agility, stamina, vision, penalties, and composure, using multivariate analysis techniques. The study used canonical correlation analysis (CCA) to analyze these correlations. The study produced three statistically significant $(p < 0.05)$ *canonical functions* (CF) *with canonical correlations of* 0.951, 0.661, *and* 0.570, *respectively. The study also showed that canonical function 1* (CF_1) *showed more significance (Wilks's* $\lambda = 0.036$ *and* $F(27, 132.066) = 10.345$*). A principal component analysis (PCA) was conducted to determine the most important variables for a central attacking midfielder. The PCA generated four components. The first component showed itself to be optimal, with variables such as short passing, long passing, vision, composure, and ball control highlighted as the most relevant attributes of a central attacking midfielder. The study also developed a dynamical system that can be analyzed to understand the efficiency of the central attacking midfielder with respect to the given variables and parameters.*

Keyword: football, central attacking midfield, canonical correlation analysis, principal component analysis.

Article History:
Received, 27-06-2024; Revised, 15-07-2024; Accepted, 21-07-2024.

1. Introduction

Central attacking midfielders (CAM) in the sport of football, such as Bruno Fernandes (Manchester United F.C.), Kevin De Bruyne and Bernardo Silva (Manchester City F.C.), Jamal Musiala (FC Bayern Munich), Jude Bellingham and Luka Modric (Real Madrid CF), and co., are central attacking midfielders who are positioned between central midfield and the team's forwards, which is considered a more advanced midfield position. The central attacking midfielders are molded to be primarily offensive and can have attributes such as dribbling, ball control, and short passing. The study is interested discovering the most relevant attributes that describe a standard central attacking midfielder. Players in this position have very expensive price tags because of their immeasurable usefulness to the team. But finding prodigies with lower price tags can also be an option for teams that are interested in recruiting from other teams.

The club football transfer market is usually very intense, with Jude Bellingham costing Real Madrid CF 103 million euros to secure his signature from Borussia Dortmund (Mukherjee, 2024). It is important for club owners, management, and football coaches to exercise caution when recruiting central attacking midfielders with heavy price tags. Players can sometimes not replicate the performance for which the current club got attracted while they were in their former clubs (Ritchie, 2024). In light of this uncertainty, it is important for teams to be specific about the attributes they are looking for in their preferred central attacking midfielder. It is also important to note that these central attacking midfielders' attributes, such as crossing, finishing, short passing, volleys, dribbling, curve, long passing, ball control, acceleration, sprint speed, agility, stamina, vision, penalties, and composure, can sometimes be correlated with one another. Sometimes, recruiting agents might be focused on the ratings of CAM players on certain attributes and might be discouraged when these players don't rate highly in those attributes, which might be a loss to the interested clubs, especially if these players are available in the transfer market with considerable price tags.

The study has utilized canonical correlation analysis (CCA) and principal component analysis (PCA), which are important multivariate statistical methods that can determine the correlations between these attributes and therefore highlight substitute attributes to recruiters in case the players of interest are not well rated in their attributes of interest. The footballer's attribute rating data was collected from FIFA players ratings online sources. The image below shows the playing position of the CAM on the playing field.

Canonical correlation analysis is a measure of the interrelationships between sets of multiple dependent variables and multiple independent variables (Akbaş et al., 2005). The method is a generalization of multiple regression analysis with more than one set of dependent variables. It explores the relationship between two multivariant sets of variables (vectors) measured on the same individual (Iweka et al., 2018; Kuss et al., 2003; Filho et al., 2022). Until recent years, CCA was a relatively unknown statistical technique (Yang et al., 2008). The technique is efficient for data reduction and interpretation (Lu et al., 2014; Akbaş et al., 2005; Filho et al., 2022; Nayir et al., 2022).

The principal component analysis is a multivariate data analysis that also has its roots in linear algebra. PCA uses sophisticated underlying mathematical principles to transform a number of correlated variables in a collection into a smaller number of variables called principal components or factors (Richardson, 2009; Ramakrishna et al., 2024), The PCA is one of the most important results of applied linear algebra (Shlens, 2005). It can effectively be used to analyze large data sets. The vector space transform is deployed by PCA to reduce the dimensionality of large data sets. Since the number of components is reduced by using principal components, the complexity of the analysis itself is also reduced by avoiding analyzing a large number of output variables (Janićijević et al., 2022; Divya et al., 2024). The position of a CAM player on the field can be seen in the image below.

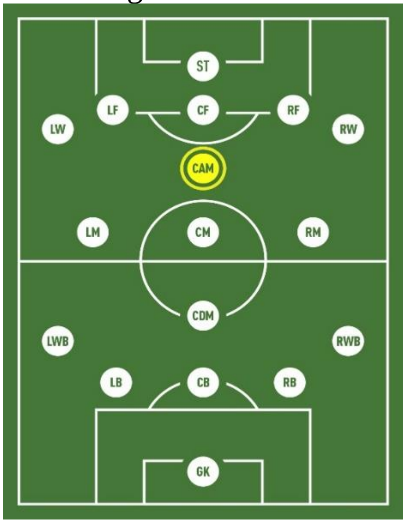

Figure 1.1. The position of a CAM player (Image credit: Google.com).

2. Material and Methods

The study is an analysis of attributes associated with a football central attacking midfield player using canonical correlation analysis, and principal component analysis, which are multivariate analysis techniques.

2.1 Data Collection

The football player ratings data used in this study was extracted from players rating website (https://fifaratings.com, accessed on 18 February 2022). The study has used both Microsoft Word and Microsoft Excel as a data configuration and deposit bank before proceeding to use SPSS for the analysis.

2.2 Procedure and Data Analysis

The quantitative data used in this study was analyzed using SPSS 25 (IBM, 2017). The study tested the data for missing data, multivariate normality, linearity, multicollinearity, and singularity before conducting the multivariate analysis using canonical correlation analysis and principal component analysis techniques.

3. Multivariate Assumption Testing

In this section, the study shall subject the dataset with respect to the proposed variables for analyzing this playing position to multivariate assumption testing. The study shall ensure that the data satisfy multivariate analysis assumptions.

3.1 Tests of Normality

Table 3.1. Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Crossing	.975	57	.284
Finishing	.964	57	.084
Short Passing	.968	57	.129
Volleys	.962	57	.073
Dribbling	.965	57	.099
Curve	.960	57	.058
Long Passing	.974	57	.243
Ball Control	.975	57	.278
Acceleration	.976	57	.306
Sprint Speed	.984	57	.655
Agility	.982	57	.541
Stamina	.967	57	.116
Vision	.979	57	.418
Penalties	.981	57	.500
Composure	.985	57	.704

A Shapiro-Wilk test was conducted, and from Table 3.1, we can see the statistics for all the variables (crossing, finishing, short passing, volleys, etc.) in the first column, with their *p*-values given in the last column, and showing all of the variables having values ($p > 0.05$). This shows that the data did not show evidence of non-conformity or violation of the assumption of normality.

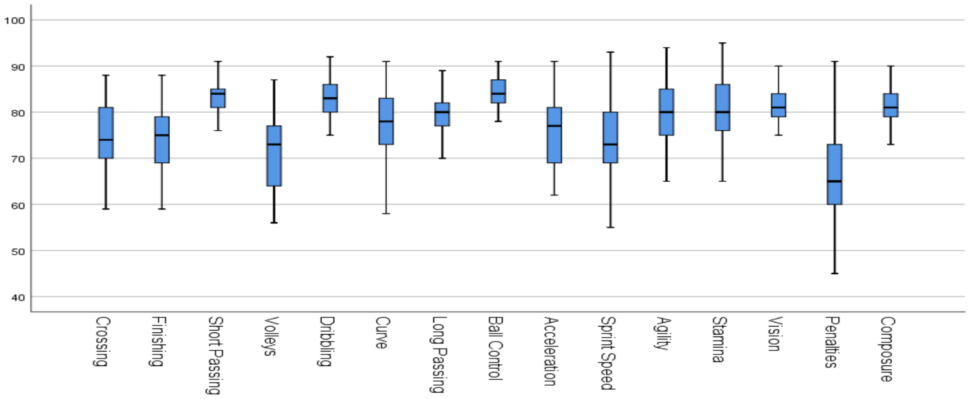

Figure 3.1. Boxplot for the data outlier check

The boxplot in Figure 3.1 shows that there are no more outliers in the data. The initial data consisted of extreme points (outliers), which, if not dealt with, may affect the outcome of this study. As recommended by the multivariate assumption on outliers, the study has ensured that only points within the interquartile range (IQR) were retained. Points below the $25th$ (Q_1) percentile and above the $75th$ (Q_3) percentile have been removed from the dataset. As we can see, variables such as crossing, finishing, short passing, vision, dribbling, etc. are within the IQR. Therefore, we infer that there are no outliers in the dataset, and the assumptions have not been violated (Grubbs, 1969; Ruan et al., 2005; Zimek et al., 2018; Hodge et al., 2004; Barnett et al., 1994; Ramaswamy et al., 2000; Schubert et al., *2012).*

3.2 Test for Non-Multicollinearity

The study has tested for multicollinearity, which, of course, occurs when the independent variables are highly correlated (Gujarati, 2009; Leamer, 1973; Giles, 2011). Hence, we have removed from the dataset one of the two variables that are correlated with values above 0.8 (Kalnins et al., 2023; Gujarati, 2009). The test shows that curve and crossing are both significantly highly correlated with a value of 0.822, which is above 0.80. Acceleration and sprint speed are also significantly highly correlated with a value of 0.802..This study has removed curve and sprint speed from the model.

3.3 Test for Independence of Observation

The study has tested for the independence of observation using the Durbin-Watson test statistics. The Durbin-Watson test statistics have values ranging between 0 and 4 (Durbin et al., 1950). A value of $0 \leq x \leq 2$ indicates a positive autocorrelation, and a value of $2 < x \leq 4$ indicates a negative autocorrelation (Durbin et al., 1951). Although Durbin-Watson test statistics value of $1.5 \leq x \leq 2.5$ can be accepted to show the independence of the observations (King, 1983; Dufour et al., 1985; Fahidy, 2006).

Table 3.2. Test for Independence of Observation

Model	Dependent Variables	Durbin-Watson
1	Finishing	**2.154**
2	Crossing	**2.068**
3	Short Passing	**2.001**
4	Volleys	**2.062**
5	Dribbling	**2.242**
6	Long Passing	**2.010**
7	Ball Control	**2.003**
8	Acceleration	**2.041**
9	Agility	**2.002**
10	Stamina	**2.372**
11	Vision	**2.013**
12	Penalties	**2.476**
13	Composure	**2.000**

The test for independence of observation using the Durbin-Watson test statistics shown in Table 3.2 indicates that there is a negative autocorrelation between the dependent variables and the predictor variables (finishing, crossing, stamina, acceleration, penalties, long passing, volleys, composure, ball control, vision, agility, short passing, and dribbling). Therefore, the study infers that the variables taken as the dependent variables, the assumption of independence of observations is not violated.

3.4 Test for Homoscedasticity

The study has checked whether the variance of the residuals is constant, by plotting the standardized residuals against the predicted values and visualize the plot using a histogram, line graph, and a scatterplot (White, 1980; Gujarati et al., 2009; Angrist et al., 2009; Long et al., 1993; Engle, 1982).

The study tested the data for homoscedasticity by plotting histograms of standardized residual. The study tested each of the variables independently to ensure that the assumption is not violated. The plots of the variables (finishing, crossing, short passing, volleys, dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure). A visual inspection of the histograms showed that the data contained approximately normally distributed errors, which is corroborated in the normal P-P plot of the standardized residuals of the respective variables.

The P-P plots showed that these variables (finishing, crossing, short passing, volleys, dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure) have an element of deviation from normality at some point along the line. This is because some points can be seen to not be exactly on the line but very close to it. We infer that these negligible deviations, which are expected, are also inconsequential, and we have chosen to neglect these slight deviations. We infer that the error terms of the residuals follow a normal distribution.

The scatterplots showed that the data of the variables (finishing, crossing, short passing, volleys, dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure) are evenly spread from left to right without any major clutter. This also affirms the claim that the data for the various variables under consideration are normally distributed. Considering results and discoveries from the histogram plots, P-P plots, and scatterplots of data for these variables, the study hence infer that these data do not violate the homoscedasticity condition.

4. Result

The study shall analyze the central attacking midfield data using both the canonical correlation analysis and the principal component analysis.

4.1 Canonical Correlation Analysis

Let the dependent variables be crossing, finishing, and short passing, while the independent variables be dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure.

Table 4.1. Canonical Correlations

	Correlation	Eigenvalue	Wilks Statistic	F	Num D.F	Denom D.F.	Sig.
1	.951	9.492	.036	10.345	27.000	132.066	.000
2	.661	.777	.380	3.578	16.000	92.000	.000
3	.570	.481	.675	3.232	7.000	47.000	.007

Table 4.1 shows the three canonical functions generated and their canonical correlations when the CCA was conducted using the variables crossing, finishing, and short passing as the dependent variables and dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure as the independent variables. The canonical functions (CF) with canonical correlations of 0.951, 0.661, and 0.570 can be seen in the table. The table also indicates that these canonical functions are statistically significant ($p < 0.05$). With Wilks's $\lambda = 0.036$ and $F(27, 132.066) = 10.345$, canonical function 1 (CF_1) showed more significance with correlation value of 0.951. The eigenvalues ($9.492, 0.777,$ and 0.481) are the shared variance between the three canonical variates of dependent and independent variables.

Table 4.2. Set 1 Canonical Loadings

Variable	1	2	3
Dribbling	-.480	.084	.523
Long Passing	-.484	.501	-.588
Ball Control	-.740	.081	.221
Acceleration	-.180	-.447	.215
Agility	-.375	-.023	.383
Stamina	-.180	-.344	-.758
Vision	-.921	-.021	.078
Penalties	-.435	-.107	.148
Composure	-.660	.020	-.076

The canonical loadings shown in Table 4.2 are the contributions or effects of the independent variables on the canonical functions. The table shows that three canonical functions were generated, and as seen from the second row, the variables have numerical values, which plays a significant role in the ranking of the canonical functions in terms of relevance. The values of -0.480, -0.484, and -0.740 for dribbling, long passing and ball control, respectively, implies that these variables have a negative effect on the first canonical function. The table also shows that the variable, dribbling, has a positive effect on both the second and third canonical functions with values 0.084, and 0.523, respectively.

Table 4.3. Set 2 Canonical Loadings

Variable	1	2	3
Crossing	-.772	.017	.636
Finishing	-.421	-.907	-.018
Short Passing	-.870	.346	-.352

Table 4.3 shows the canonical loadings of the three dependent variables on the canonical functions. It can be seen that the variables crossing, finishing, and short passing have values of -0.772, -0.421, and -0.870, respectively. This implies that they have negative effects on the first canonical function. In the second and third canonical functions, it can be seen that crossing has a positive effect of 0.017, while the other two both have negative effects.

Table 4.4. Proportion of Variance Explained

Canonical Variable	Set 1 by Self	Set 1 by Set 2	Set 2 by Self	Set 2 by Set 1
1	.299	.271	.510	.461
2	.066	.029	.314	.137
3	.163	.053	.176	.057

Table 4.4 shows the proportion of variance between the canonical loadings of the canonical functions in set 1 (Table 4.2) and set 2 (Table 4.3). The second and fourth columns

contain the variances explained by canonical loadings of each canonical function and themselves. The third and fifth columns contain the variances of each canonical function and other canonical functions.

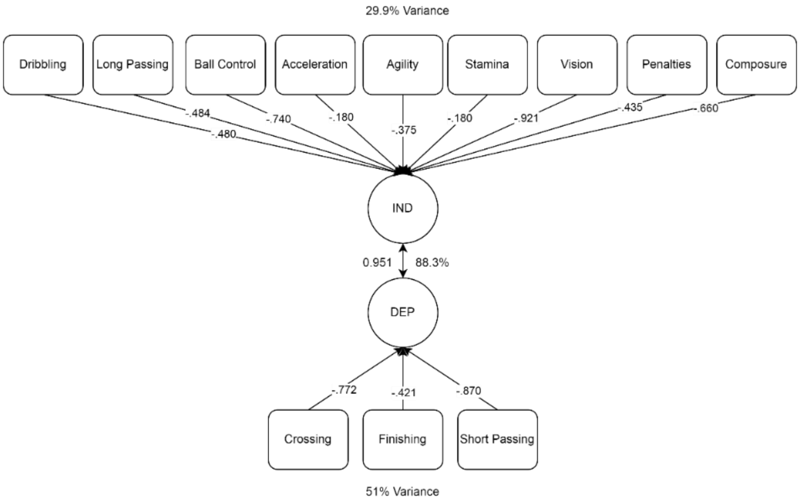

Figure 4.1. CC CF_1 diagram

The diagram in Fig. 4.1 is a pictorial representation of the canonical correlation of canonical function 1 (CF_1) in Table 4.1 with respect to the independent and dependent variables shown. It also shows the corresponding canonical loadings of the given canonical functions in Tables 4.2 and 4.3. The diagram also shows the variances of the independent and dependent variables, as seen in Table 4.4.

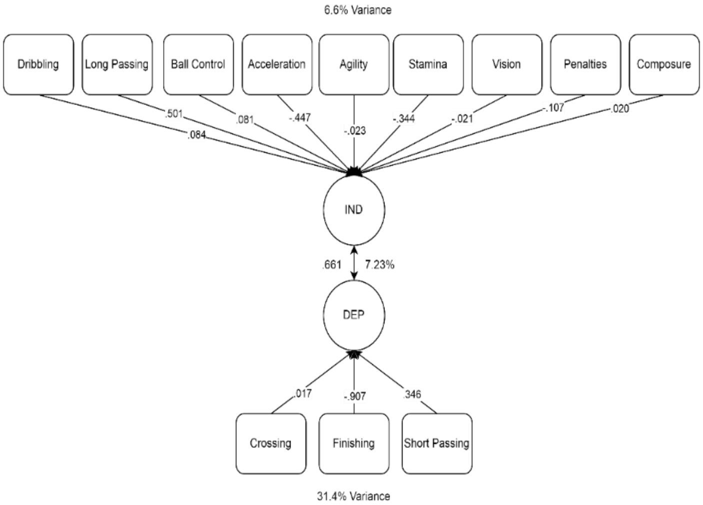

Figure 4.2. CC CF_2 diagram

The diagram in Fig. 4.2 is a pictorial representation of the canonical correlation of canonical function 2 (CF_2) in Table 4.1 with respect to the independent and dependent variables shown. It also shows the corresponding canonical loadings of the given canonical functions in Tables 4.2 and 4.3. The diagram also shows the variances of the independent and dependent variables, as seen in Table 4.4.

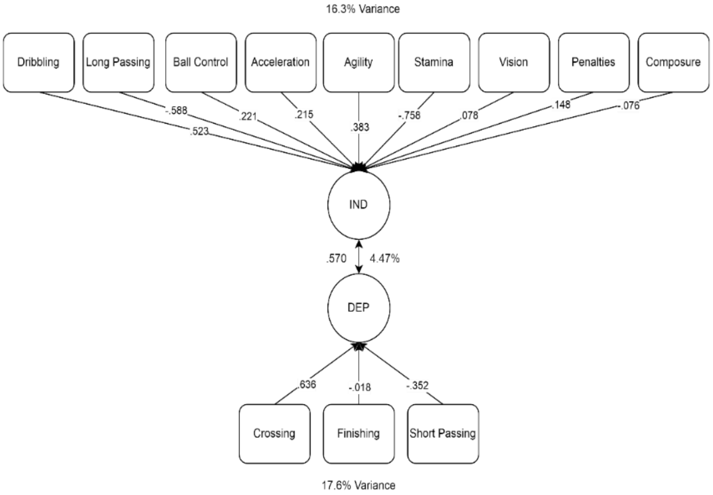

Figure 4.3. CC CF_3 diagram

The diagram in Fig. 4.3 is a pictorial representation of the canonical correlation of canonical function 3 (CF_3) in Table 4.1 with respect to the independent and dependent variables shown. It also shows the corresponding canonical loadings of the given canonical functions in Tables 4.2 and 4.3. The diagram also shows the variances of the independent and dependent variables, as seen in Table 4.4.

Considering the multiple canonical functions generated when using CCA, it is recommended that researchers only interpret those canonical functions that explain a reasonable amount of variance between the variable sets or risk interpreting an effect that may not be noteworthy (Sherry et al., 2005). In their example, they chose to interpret the first two canonical functions, as they explained 38.1% and 20.0% of the variance within their canonical functions, respectively. In this study, the study shall only interpret canonical functions as they have the highest variance between the variable sets, and in some cases, the first two canonical functions if the second canonical function is of considerable significance.

4.2 Principal Component Analysis

The study shall use the principal component analysis (PCA) to reduce the variables in the data for each of the playing positions we are considering in this study. The PCA can do this by grouping the variables into components called principal components. This mathematical technique generates a few linear combinations of the original variables that maximally explain the variance of all the variables.

The study shall carry out a principal component analysis (PCA) of the central attacking midfield data.

Table 4.5. Correlation matrix

		Crossing	Finishing	Short Passing	Volleys	Dribbling	Curve	Long Passing	Ball Control	Acceleration	Sprint Speed	Agility	Stamina	Vision	Penalties	Composure
Correlation	Crossing	1.000	.298	.453	.455	.543	.822	.148	.624	.205	.054	.414	-.146	.704	.372	.457
	Finishing	.298	1.000	.058	.508	.137	.232	-.101	.245	.338	.372	.160	.286	.380	.237	.253
	Short Passing	.453	.058	1.000	.282	.312	.484	.633	.586	.003	-.121	.228	.222	.741	.306	.565
	Volleys	.455	.508	.282	1.000	.298	.512	-.037	.376	.186	.199	.279	-.049	.411	.440	.354
	Dribbling	.543	.137	.312	.298	1.000	.597	-.196	.762	.606	.341	.730	-.248	.338	.060	.203
	Curve	.822	.232	.484	.512	.597	1.000	.095	.680	.220	.129	.435	-.172	.538	.444	.400
	Long Passing	.148	-.101	.633	-.037	-.196	.095	1.000	.139	-.273	-.299	-.154	.210	.486	.115	.314
	Ball Control	.624	.245	.586	.376	.762	.680	.139	1.000	.325	.101	.634	-.054	.521	.223	.483
	Acceleration	.205	.338	.003	.186	.606	.220	-.273	.325	1.000	.802	.732	.083	.163	.009	.039
	Sprint Speed	.054	.372	-.121	.199	.341	.129	-.299	.101	.802	1.000	.392	.122	-.005	.047	-.013
	Agility	.414	.160	.228	.279	.730	.435	-.154	.634	.732	.392	1.000	.005	.334	.165	.277
	Stamina	-.146	.286	.222	-.049	-.248	-.172	.210	-.054	.083	.122	.005	1.000	.098	.100	.246
	Vision	.704	.380	.741	.411	.338	.538	.486	.521	.163	-.005	.334	.098	1.000	.351	.532
	Penalties	.372	.237	.306	.440	.060	.444	.115	.223	.009	.047	.165	.100	.351	1.000	.329
	Composure	.457	.253	.565	.354	.203	.400	.314	.483	.039	-.013	.277	.246	.532	.329	1.000

a. Determinant = 5.974E-6

Table 4.5 shows the correction matrix of the central attacking midfield analysis, which indicates the level at which each variable correlates with other variables and itself. The table shows the correlations between these variables, which go from 0 to 1. The study shall consider values less than 0.3, meaning that the variables concerned are less correlated or weakly correlated. The study shall also consider values above 0.8 to be strongly correlated. The study shall neglect values near 0.8 in this study. Note that all variables will have correlation values equal to 1 between themselves. The determinant of the matrix can be seen at the bottom of the table, with a value of 5.974×10^{-6}. This shows that the correlation matrix is nonsingular.

Table 4.6. Sampling adequacy and sphericity test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.687
Bartlett's Test of Sphericity	Approx. Chi-Square	603.412
	df	105
	Sig.	.000

Table 4.6 shows the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. The KMO assesses the suitability of the data for principal component analysis by measuring the degree of coherence between variables. Values from the test go from 0 to 1, with values greater than 0.5 considered suitable and values within 0.9 considered excellent (IBM, 2024). The Bartlett's test measures the hypothesis of homogeneity of the correlation matrix. The significance value (*p*-value) of the Bartlett's test is appropriate when less than 0.05 (*Kaiser et al., 1974;* IBM, 2024; *Cureton et al., 2013; Dziuban et al., 1974)*. The table shows a KMO value of 0.687 and a significant value less than 0.0001 in the Bartlett's test, with both results indicating that central attacking midfield analysis data are suitable for the principal component analysis.

Table 4.7. Extracted component results

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	5.535	36.899	36.899	5.535	36.899	36.899	3.795
2	2.769	18.457	55.356	2.769	18.457	55.356	3.908
3	1.670	11.133	66.489	1.670	11.133	66.489	2.033
4	1.342	8.944	75.433	1.342	8.944	75.433	3.536
5	.734	4.893	80.326				
6	.672	4.477	84.803				
7	.498	3.323	88.126				
8	.455	3.036	91.162				
9	.396	2.643	93.804				
10	.315	2.101	95.905				
11	.256	1.705	97.610				
12	.129	.858	98.467				
13	.112	.748	99.216				
14	.068	.455	99.670				
15	.049	.330	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Table 4.7 shows the total variance explained by the variables in the central attacking midfield analysis. The initial eigenvalue compartment in the table contains three columns. This compartment accounts for the variances explained by the full set of initial factors. For optimality, the study choose the components whose eigenvalues are greater than one (IBM, 2024). The extraction sums of squared loadings compartment show that the PCA has extracted four components, which can account for 75.433% of the variances in the analysis. This implies that the PCA has reduced the factors to four. The last compartment in the table shows the rotation sums of squared loadings, which show the variances explained by the rotated factors.

Table 4.8. Component matrix

	Component			
	1	2	3	4
Ball Control	.832		-.253	.177
Curve	.819		-.264	-.240
Crossing	.816	.124	-.217	-.206
Vision	.768	.389		
Dribbling	.720	-.437	-.385	.180
Agility	.678	-.467	-.101	.297
Short Passing	.657	.557		.329
Volleys	.615		.192	-.537
Composure	.613	.373	.206	
Penalties	.463	.230	.220	-.459
Long Passing	.182	.755		.369
Acceleration	.474	-.743	.208	.297
Sprint Speed	.285	-.708	.400	.101
Stamina		.180	.776	.361
Finishing	.443	-.175	.617	-.324

Extraction Method: Principal Component Analysis.
a. 4 components extracted.

Table 4.8 shows the correlations between the variables and the four unrotated factors (components). We can see the four components generated by the PCA for the central attacking midfield analysis. The columns show the correlation between the variables and the components under consideration. The table shows that the correlation between the variable ball control and the first PCA component is 0.832.

Table 4.9. Pattern matrix

	Component			
	1	2	3	4
Short Passing	.911			
Long Passing	.850	.276	.185	.183
Vision	.690	-.124		-.318
Composure	.614		.119	-.284
Acceleration	-.125	-.955	.204	
Agility	.168	-.850	-.117	
Dribbling	.138	-.786	-.431	
Sprint Speed	-.307	-.726	.350	-.152
Ball Control	.494	-.519	-.323	
Stamina	.366		.848	
Curve	.283	-.232	-.476	-.457
Volleys			-.117	-.830
Finishing	-.106	-.164	.383	-.730
Penalties	.110	.202		-.720
Crossing	.353	-.195	-.419	-.448

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 22 iterations.

Table 4.9 shows the partial correlations between the variables and the four rotated factors (components) generated by the PCA for the central attacking midfield analysis after 22 iterations using the Oblimin with Kaiser normalization method. This study will consider values above 0.45 (IBM, 2024; *Larsen et al., 2010;* Kaiser et al., 1960; Velicer, 1976). The columns show that 14, 11, 12, and 9 variables have been grouped into the first, second, third, and fourth components, respectively.

Table 4.10. Component correlation matrix

Component	1	2	3	4
1	1.000	-.072	-.141	-.279
2	-.072	1.000	.094	.312
3	-.141	.094	1.000	.060
4	-.279	.312	.060	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Table 4.10 shows the correlations among each of the four components and themselves. The study will only consider correlation values that are less than or equal to 0.32 (Taherdoost et al., 2014; IBM, 2024). The table shows that the correlation between component one and the other three components is weak since the values are seen to be -0.072, -0.141, and -0.279, respectively. This result also implies that the PCA is adequate for the central attacking midfield data.

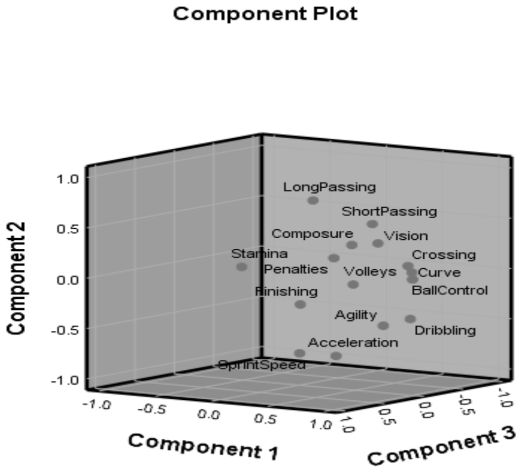

Figure 4.4. 3D component plot

Figure 4.4 shows the 3D component plot of the first three components extracted by the PCA for the central attacking midfield data. The axes are scaled from -1 to 1. We can see the variables represented on the plot. Coefficients close to -1 or 1 indicate that the variable strongly influences the component, either negatively or positively, (IBM, 2024; Barnett et al.,1987).

4.3 Mathematical Model Formulation: Attack-Oriented Play

This model is designed to analyze the central attacking midfield contribution to the game. The variables in consideration are both those derived from the optimal variables of the canonical correlation analysis and principal component analysis of attributes of the playing position and the in play attainable realities.

4.3.1 Model Assumption

The model was developed under these assumptions.

1. The contributing variables are attributes.
2. The parameters are rates at which the attributes contribute to the receiving (in play) variables.
3. The attributes are skills, and the in-play variables depend on the measure of the skills used.
4. Possession is retained throughout the period of analysis.
5. Initiating an attack is prioritized.
6. Increasing the use of these skills is prioritized.

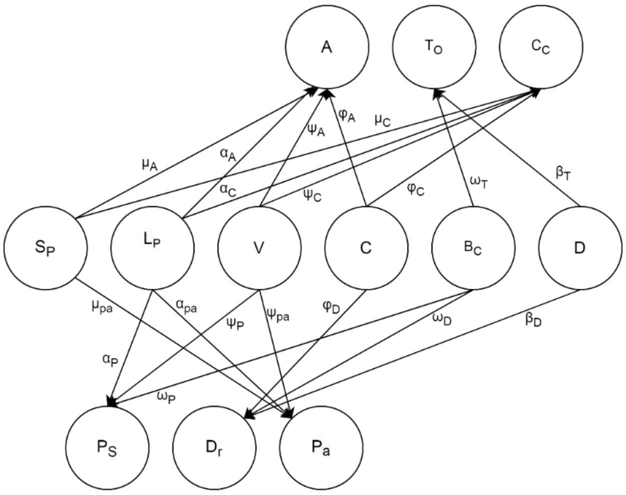

Figure 4.5. Flow chat of the model

Table 4.11. Description of variables and parameters

Variables	Description
$\boldsymbol{S_P}$	Short passing
$\boldsymbol{L_P}$	Long passing
$\boldsymbol{V}$	Vision
$\boldsymbol{C}$	Composure
$\boldsymbol{B_C}$	Ball control
$\boldsymbol{D}$	Dribbling
$\boldsymbol{A}$	Assists
$\boldsymbol{T_O}$	Take-on
$\boldsymbol{C_C}$	Goal scoring chances created
$\boldsymbol{P_S}$	Play switch
$\boldsymbol{D_r}$	Dribbles
$\boldsymbol{P_a}$	Pass accuracy

Parameters	Description	Value
$\boldsymbol{\mu_A}$	Rate at which short passes result to assists	--
$\boldsymbol{\mu_C}$	Rate at which short passes result to goal scoring chances	--
$\boldsymbol{\mu_{Pa}}$	Rate at which short passes are accurate passes	
$\boldsymbol{\alpha_A}$	Rate at which long passes result to assists	--
$\boldsymbol{\alpha_C}$	Rate at which long passes result to goal scoring chances	--
$\boldsymbol{\alpha_{Pa}}$	Rate at which long passes are accurate passes	--
$\boldsymbol{\alpha_P}$	Rate at which long passes result to play switch	--
$\boldsymbol{\psi_{Pa}}$	Rate at which vision contribute to accurate passes	--
$\boldsymbol{\psi_P}$	Rate at which vision contribute to play switch	--
$\boldsymbol{\psi_C}$	Rate at which vision contribute to take-on	--
$\boldsymbol{\psi_A}$	Rate at which vision contribute to assists	--
$\boldsymbol{\varphi_C}$	Rate at which composure contribute to the creation of goal scoring chances	--
$\boldsymbol{\varphi_D}$	Rate at which composure contribute to dribbles	--
$\boldsymbol{\varphi_A}$	Rate at which composure contribute to assists	--
$\boldsymbol{\omega_T}$	Rate at which ball control contribute to take-on	--
$\boldsymbol{\omega_D}$	Rate at which ball control contribute to dribbles	--
$\boldsymbol{\omega_P}$	Rate at which ball control contribute to play switch	--
$\boldsymbol{\beta_T}$	Rate at which dribbling contribute to take-on	--
$\boldsymbol{\beta_D}$	Rate at which dribbling contribute to dribbles	--

Following the model flow chart in Figure 4.5 and the assumptions in Table 4.11, we have the following partial differential equations;

$$\frac{\partial S_P}{\partial t} = -\mu_A S_P - \mu_C S_P - \mu_{Pa} S_P$$

$$\frac{\partial L_P}{\partial t} = -\alpha_A L_P - \alpha_C L_P - \alpha_{Pa} L_P - \alpha_P L_P$$

$$\frac{\partial V}{\partial t} = -\psi_{Pa}V - \psi_P V - \psi_C V - \psi_A V$$
$$\frac{\partial C}{\partial t} = -\varphi_C C - \varphi_D C - \varphi_A C$$
$$\frac{\partial B_C}{\partial t} = -\omega_T B_C - \omega_D B_C - \omega_P B_C$$
$$\frac{\partial D}{\partial t} = -\beta_T D - \beta_D D$$
$$\frac{\partial A}{\partial t} = \mu_A S_P + \alpha_A L_P + \psi_A V + \varphi_A C$$
$$\frac{\partial T_O}{\partial t} = \omega_T B_C + \beta_T D$$
$$\frac{\partial C_C}{\partial t} = \mu_C S_P + \alpha_C L_P + \psi_C V + \varphi_C C$$
$$\frac{\partial P_S}{\partial t} = \alpha_P L_P + \psi_P V + \omega_P B_C$$
$$\frac{\partial D_r}{\partial t} = \varphi_D C + \omega_D B_C + \beta_D D$$
$$\frac{\partial P_a}{\partial t} = \mu_{Pa} S_P + \alpha_{Pa} L_P + \psi_{Pa} V$$

Let the initial conditions be:

$S_P \geq (O), L_P \geq (O), V \geq (O), C \geq (O), B_C \geq (O), D \geq (O), A \geq (O), T_O \geq (O), C_C \geq (O), P_S \geq (O), D_r \geq (O), P_a \geq (O).$

5. Discussion

The canonical correlation analysis result for central attacking midfield Analysis was conducted using the variables crossing, finishing, and short passing as the dependent variables and dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure as the independent variables to evaluate the multivariate shared relationship between the two variable sets. The CCA yielded three canonical functions (CF) with canonical correlations of 0.951, 0.661, and 0.570 for each successive canonical function. It is important to state that all the canonical functions are statistically significant $(p < 0.05)$, but canonical function 1 (CF_1) showed more significance if we consider Wilks's $\lambda = 0.036$ and $F(27, 132.066) = 10.345$, as shown in Table 4.1. The variance between the dependent and independent canonical variates of CF_1 is 88.3% (Table 4.4). This was obtained from calculating the percentage of the eigenvalue of CF_1.

The canonical loadings (Table 4.2 and 4.3), which are the values contributed to the canonical functions by each of the variables in the independent set (set 1) and dependent set (set 2), have values of -0.480, -0.484, -0.740, $-0.180, -0.375, -0.180, 0.921$, -0.435, and -0.660, respectively. The variance of these nine variables is 29.9%. This was obtained from Table 4.1, column 1 (proportion of variance explained). Looking at the dependent variables in Table 4.3 (set 2), we can see that the canonical loadings of the variables crossing, finishing, and short passing on CF_1 are $-0.772, -0.421$, and -0.870. Also, from Table 4.4, column 2, the variance of these three variables is 51%. We can see from above that the highest contributors to the canonical function CF_1 are vision, ball control, and composure, with values of $-0.921, -0.740$, and -0.660, respectively. While in the response set (set 2), we can see that short passing and crossing are most affected, with response values of -0.870 and -0.772, respectively, for the independent variables. This is not to say that the other predictor and response variables are inconsequential, as we can see that their values are considerably high as well, apart from acceleration and stamina, whose values can be seen to be less than -0.2.

A quick look at CF_2 shows that the independent variables (dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, and composure) have values of $0.084, 0.501, 0.081, -0.447$, -0.023, -0.344, -0.021, -0.107, and 0.020, respectively. The CF_2 dependent variables have coefficients of $0.017, -0.907$, and 0.346. The variance of these nine control variables is 6.6%, while the variance of these three response variables is 31.4%. We can see that the highest control variable coefficients of CF_2 are long passing and acceleration, with values of 0.50 and -0.447, respectively. The most affected variable in the response set is finishing (-0.907). Although we shall not consider working with CF_2 and CF_3 because the variance between the dependent and independent canonical variates is low (7.23% and 4.47%, respectively).

The principal component analysis (PCA) result for central attacking midfield analysis was conducted using the variables crossing, finishing, short passing, volleys, dribbling, long passing, ball control, acceleration, agility, stamina, vision, penalties, curve, sprint speed, and composure. The PCA reduced the factors to four, showing the results of the four retained components (factors). The component correlation matrix (Table 4.10) showed that the four extracted components have weak correlations between themselves. Table 4.9 shows the pattern in which the variables correlate with the four rotated factors (components) generated by the PCA for the central attacking midfield data. We shall be considering variables with correlation values above 0.45, which implies that we shall accept only the first and fourth components. The optimal variables in the first component are short passing, long passing, vision, composure, and ball control. The fourth component will contain only one variable, stamina. We infer that after 22 iterations, the PCA grouped the variables in the first and fourth components as the most relevant variables (attributes) for a central attacking midfield player.

6. Conclusion

The canonical correlation analysis showed that the first canonical function showed better correlation between the variates. The attributes dribbling, long passing, ball control, agility, vision, penalties, and composure showed huge influences on the attributes crossing, finishing, and short passing, which are more associated with players playing in such positions (central attacking midfield). This analysis suggests that central attacking midfielders should be scouted for possible recruitment and considered for possible selection in games, considering the first attributes listed above. These attributes have been considered by this analysis to be the major influencers of performance attributes (crossing, finishing, and short passing). The principal component analysis (PCA) result produced four principal factors. The results showed that the four extracted components have weak correlations between themselves. The first component can be considered optimal, with variables such as short passing, long passing, vision, composure, and ball control considered by the analysis as the most relevant qualities of a central attacking midfielder. The dynamical system in figure 4.5 was generated to show how these optimal attributes relate with each other to produce an efficient central attacking midfielder. The mathematical model can be analyzed to understand the efficiency of the CAM with respect to the given variables and parameters.

Conflicts of Interest: The authors declare no conflicts of interest.

7. References

Akbaş, Y., & Takma, Ç. (2005). Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers. *Czech Journal of Animal Science,* 50(4), 163–168. https://doi.org/10.17221/4010-cjas.

Angrist, J. D., & Pischke, J. (2009). Mostly Harmless Econometrics: An Empiricist's

Companion. *Princeton University Press*. Doi:10.1515/9781400829828. ISBN 978-1-4008-2982-8.

Barnett, T. P., & Preisendorfer R. (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. Monthly Weather Review. 115 (9): 1825. doi:10.1175/1520-0493(1987)115<1825:oaloma>2.0.co;2.1516440.

Barnett, V., & Lewis, T. (1994), *Outliers in Statistical Data (3 ed.), Wiley,* ISBN 978-0-471-93094-5.

Cureton, E. E., & d'Agostino, R. B. (2013). *Factor Analysis.* doi*:*10.4324/9781315799476.

Divya, T., & Praveen, L. (2024). A method for human behavior identification based on integrated sensor data using XGBoost classifier with PCA techniques. *Physica Scripta*. https://doi.org/10.1088/1402-4896/ad328c

Dufour, J.-M., & Dagenais, M. G. (1985). Durbin-Watson tests for serial correlation in regressions with missing observations, *Journal of Econometrics,* Volume 27, Issue 3, 1985, Pages 371-381, ISSN 0304-4076, https://doi.org/10.1016/0304-4076(85)90012-0.

Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression, I. *Biometrika*. 37 (3–4): 409–428. Doi:10.1093/biomet/37.3-4.409.

Durbin, J., & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression, II. *Biometrika*. 38 (1–2): 159–179. Doi:10.1093/biomet/38.1-2.159.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin. 81 (6): 358–361.* doi*:*10.1037/h0036316.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica.* 50 (4): 987–1007. Doi:10.2307/1912773. ISSN 0012-9682. Peter Kennedy, A Guide to Econometrics, 5th edition, p. 137.

Fahidy, T. Z. (2006). An application of Durbin–Watson statistics to electrochemical science, *Electrochimica Acta,* Volume 51, Issue 17, 2006, Pages 3516-3520, ISSN 0013-4686, https://doi.org/10.1016/j.electacta.2005.09.046.

Filho, A. C., & Toebe, M. (2022). Sample size for canonical correlation analysis in corn. *Bragantia*, 81(2017), 1–14. https://doi.org/10.1590/1678-4499.20210335

Giles, D. (2011). Econometrics Beat: Dave Giles' Blog: Micronumerosity. Econometrics Beat. Retrieved 3 September 2023.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics. 11 (1): 1–21. An outlying observation, or "outlier," is one that appears to deviate markedly from other members of the sample in which it occurs.* doi*:*10.1080/00401706.1969.10490657.

Gujarati, D. (2009). *Multicollinearity: what happens if the regressors are correlated*? Basic Econometrics (4th ed.). McGraw−Hill. Pp. 363. ISBN 9780073375779.

Gujarati, D. N.; Porter, D. C. (2009). *Basic Econometrics* (5th ed.). Boston: McGraw-Hill Irwin. ISBN 978-0-07-337577-9.

Hodge, V. J., & Austin, J. (2004), A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review, 22 (2): 85–126,* doi*:*10.1023/B:AIRE.0000045502.10941.a9.

http://www.brainmapping.org/NITP/PNA/Readings/pca.pdf

https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf

IBM. (2024). *KMO and Bartletts Test.* https://www.ibm.com/docs/en/spss-statistics/28.0.0?topic=detection-kmo-bartletts-test

Iweka, F., & Magnus-Arewa, A. (2018). Canonical Correlation Analysis, A Sin Quanon for Multivariant Analysis in Educational Research. *International Journal of Humanities, Social Sciences and Education,* 5(7), 116–126. https://doi.org/10.20431/2349-

0381.0507013

Janićijević, S., Mizdraković, V., & Kljajić, M. (2022). Principal component analysis in financial data science. *Advances in principal component analysis*. http://dx.doi.org/10.5772/intechopen.102928

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*. 20 (1): 141–151. doi:10.1177/001316446002000116.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement.* doi*:*10.1177/001316447403400115. .

Kalnins, A., & Praitis Hill, K. (2023). The VIF Score. What is it Good For? Absolutely Nothing. *Organizational Research Methods*. Doi:10.1177/10944281231216381. ISSN 1094-4281.

King, M. L. (1983). The Durbin-Watson test for serial correlation: Bounds for regressions using monthly data, *Journal of Econometrics*, Volume 21, Issue 3, 1983, Pages 357-366, ISSN 0304-4076, https://doi.org/10.1016/0304-4076(83)90050-7.

Kuss, M., & Graepel, T. (2003). The geometry of kernel canonical correlation analysis. *Biological Cybernetics*, http://www.kyb.mpg.de/publication.html?publ=2233

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods. 42 (3): 871–876.* doi*:*10.3758/BRM.42.3.871.

Leamer, E. E. (1973). Multicollinearity: A Bayesian Interpretation. *The Review of Economics and Statistics*. 55 (3): 371–380. Doi:10.2307/1927962. ISSN 0034-6535.

Long, J. S., & Trivedi, P. K. (1993). Some Specification Tests for the Linear Regression Model. In Bollen, Kenneth A.; Long, J. Scott (eds.). *Testing Structural Equation Models*. London: Sage. Pp. 66–110. ISBN 978-0-8039-4506-7.

Lu, Y., & Foster, D. P. (2014). Large scale canonical correlation analysis with iterative least squares. *Advances in Neural Information Processing Systems.*

Mukherjee, S. (2024, May 30). Dortmund chief says Jude Bellingham cost Real Madrid 'a lot more' than €103m as England star prepares to face former club in Champions League final. Goal. https://www.goal.com/en-ng/lists/dortmund-jude-bellingham-real-madrid-eur103m-champions-league-final/bltf003ee1080703668

Nayir, F., & Saridas, G. (2022). The relationship between culturally responsive teacher roles and innovative work behavior: Canonical Correlation Analysis. *Journal of Educational Research and Practice*, 12(1), 36–50. https://doi.org/10.5590/jerap.2022.12.1.03

Ramakrishna, K.K., Muddu, M., Fouzi, H., Anoop, V., & Ying, S. (2024). Robust Fault Detection in Monitoring Chemical Processes Using Multi-Scale PCA with KD Approach. *Chem Engineering*. https://doi.org/10.3390/chemengineering8030045

Ramaswamy, S.; Rastogi, R.; & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data – SIGMOD* '00. P. 427. Doi:10.1145/342009.335437. ISBN 1581132174.

Richardson, M. (2009). *Principal Component Analysis.*

Ritchie, C. (2024, March 24). 16 worst premier league signings in football history (ranked). Givemesport. https://www.givemesport.com/football-soccer-premier-league-worst-signings-ever/#:~:text=Some%20notable%20examples%20of%20bad,the%20club%2C%20or%20were%20overpriced

Ruan, D., Chen, G., & Kerre, E. (2005). Intelligent data mining: Techniques and applications. *studies in computational intelligence Vol. 5. Springer*. ISBN 978-3-540-26256-5.

Schubert, E., Zimek, A., & Kriegel, H. -P. (2012). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery. 28: 190–237.* doi*:*10.1007/s10618-012-0300-z.

Sherry, A., & Robin K. H. (2005). Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer, *Journal of Personality Assessment*. https://doi.org/10.1207/s15327752jpa8401

Shlens, J. (2005). *A Tutorial on Principal Component Analysis*. http://www.brainmapping.org/NITP/PNA/Readings/pca.pdf

Taherdoost, H., Sahibuddin, S., & Jalaliyoon, N. (2014). Exploratory Factor Analysis; Concepts and Theory. Advances in Applied and Pure Mathematics, 27, WSEAS, pp.375-382, 2014, *Mathematics and Computers in Science and Engineering Series*, 978-960-474-380-3.

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*. doi:10.1007/bf02293557.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 48 (4): 817–838. doi:10.2307/1912934.

Yang, H. J., Jing, Z. L., & Zhao, H. T. (2008). Tensor canonical correlation analysis. *Shanghai* *Jiaotong Daxue Xuebao/Journal of Shanghai Jiaotong University*, 42(7), 1124–1128.

Zimek, A., & Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. doi:10.1002/widm.1280. ISSN 1942-4787.