ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

AQ049-3-M-MMDA

MULTIVARIATE METHODS FOR DATA ANALYSIS

APUMF1911DSBA(PR)

HAND OUT DATE: **25 MAY 2020**

HAND IN DATE: **22 JUNE 2020**

WEIGHTAGE: **40%**

INSTRUCTIONS TO CANDIDATES:

1. **This assignment should be submitted through outline facilities made available to the students.**
2. **Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing).**
3. **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.**
4. **Cases of plagiarism will be penalized.**
5. **You must obtain 50% overall to pass this module.**

STATISTICAL ANALYSIS IN FOOTBALL INDUSTRIES

AHMAD FUAD KHALIT

TP058497@mail.apu.edu.my

Asia Pacific University of Technology & Innovation

1.0 STEPWISE REGRESSION ON MLR

1.1 PROBLEM STATEMENT

In the past, using statistical method was not very popular among sport teams as they do not believe that mathematics can predicts or improved the performance of their teams. This has changes lately with the introduction of Sabermetrics methods in baseball games in United States popularize by Oakland Athletics' General Manager Billy Beane in his books “Moneyball” that has changed how baseball teams look at talent. From that, advance statistic has found its way in various sports including football as teams and their management realize the importance to integrate statistical model into their team management (Research and Markets. 2020). For this assignment, we will try to build a statistical model that can reveals some hidden pattern inside football match such as the most significant factor that affect scoring a goal in football match. By revealing the significant factor, it will help the team to look at the angle of improvement before and during match to improve their probability of winning a certain game.

1.2 INTRODUCTION

For this analysis, we used secondary data collected from Kaggle.com that consist data of European football match from 2014~2018 season (Lehkyi, 2019) and use the popular metrics in sport analytics called Advanced Metrics. The raw data contains 24 variable and 23,336 observation and the variable in the dataset include some non-metric (ordinal & binary) which we will remove later and metric (interval and ration) that is the requirement for this stepwise multiple linear regression analysis. Then we clean the data into 9 variable and 216 observation and pick random team from random league (Liverpool team from English Premier League). Before we start to analyze the data, first we will run a data exploration to understand what inside the dataset which we will conduct linearity, normality, absent of multicollinearity test and visual examination the homoscedasticity of error term. After that, we will analyze the

Advanced Metrics variable to measure the correlation of the independent variable to the dependent variable (scored) and analyze the hypothesis test result either accept or reject the null hypothesis and finally make conclusion and recommendation from our analysis on how to improve the team performance.

1.2.1 Understanding Football Advanced Metrics

Football in laymen's terms is a simple game. Try to gain possession of the ball and find a way to get it to over a line and into the opponent net. The most important statistics in this game is goals where it's the factor that differentiate between winning a game and losing a game. Advanced Metrics was the term used by sport analytics company such as Opta Sports. Its used basic principal in Sabermetrics methods in earlier days and as todays it has become more advanced that the researcher/statistician can capture every movement of player in games and derive significant values that give certain parameter with accurate metrics that able to measure athlete performance.

Here are the most widely used metrics(variable) that been used in Advanced Metrics with the explanation in order to understand how the metrics are derived from: -

Advanced Metrics Variable	Details Explanation
(xG) **Expected Goal**	The metrics derived from the quality of a shot player/athlete made based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance.
(xA) **Expected Assist**	The quality of a shot based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance.

(PPDA) **Passes Allowed Per Defensive Action**	Measure the passes allowed per defensive action on the opposition area when the team are in possession of the ball. Defensive Actions are possession-winning duels, tackles, interceptions, fouls.
(npxG) **Expected Goals without Penalties and Own Goals**	Expected goals without penalties and own goals. It provides a more accurate analysis. Since penalties have an (xG) of 0.76, they can significantly distort both a player's and team's expected goals.
(xGA) **Expected Goals Against**	Expected goals against means that the probabilities shot faced from opponent against the team.
(deep)	Passes completed within an estimated 20 yards of goal (crosses excluded)
(deep_allowed)	Opponent passes completed within an estimated 20 yards of goal (crosses excluded)
(scored)*	Numbers of successful goal against opponent
(missed)	Number of goals missed in games

***Selected dependend variable (DV)**

There are 2 ways of using the Advance Metrics where analyst can used for. One is for analyze player performance and another is to analyze team performance. For this analysis we will used it to analyze team performance instead.

1.3 RESEARCH OBJECTIVE

1.3.1 Aim

The aim for this analysis is to conduct statistical inclusive understanding on which metrics is significant that a team should focus on if the objective is to improved goal scored in match and identify other important correlation amongst variable that can contribute to the team success of scoring a goal by constructing Stepwise Multiple Linear Regression on the dataset.

1.3.2 Objective

- To analyze expected goal(xG) effect to the goal(scored) outcome
- To examine goal missed (missed) will have impact on goal(scored)
- To find if passes completed in 20 yard (deep) has significant impact on (scored)
- To develop useful MLR model

1.3.2 Hypothesis Test

Hypothesis 1

H_0 - Expected goal (xG) does not have effect on goal (scored)
H_1 - Expected goal (xG) have effect on goal (scored)

Hypothesis 2

H_0 - Passes completed in 20 yard (deep) has no influence on goal (scored)
H_1 - Passes completed in 20 yard (deep) has influence on goal (scored)

Hypothesis 3

H_0 - Goal missed (missed) will not have impact on goal(scored)
H_1 - Goal missed (missed) will have impact on goal(scored)

Hypothesis 4

H_0 - The model is not adequate
H_1 - The model is adequate

1.4 RESULT ANALYSIS & INTERPRETATIONS

Before we can jump into model summary of multiple linear regression, there are several model assumption that should be conducted first such as

- **Multivariate Normality**- Normality of the error term distribution
- **Linear relationship –** where it assumes there is linear relationship between independent variable and dependent variable
- **Homoscedasticity (equal variance) –** Constant variance of the error terms where it assume of error term are similar across the value of independent variable.
- **Minimum Multicollinearity –** Independence of the error term where it assumes that all independent variable are not highly correlated with each other (Hair et, el, 2014).

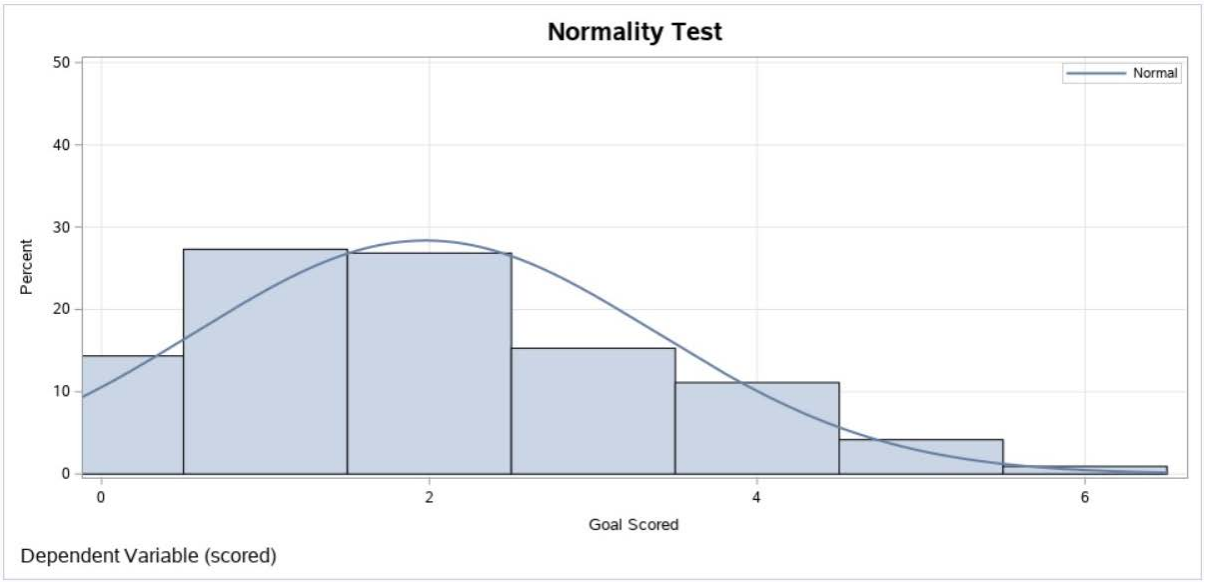
1.4.1 Data View

xG	xGA	npxG	npxGA	deep	deep_allowed	npxGD	ppda	missed	scored
1.33192000000000	1.55248000000000	1.33192000000000	1.55248000000000	6	9	-.220560000000000	11.333333333333300	1	2
.901889000000000	1.31375000000000	.901889000000000	1.31375000000000	7	11	-.411860999999999	18.222222222222200	3	1
2.74317000000000	.485920000000000	1.98200000000000	.485920000000000	4	6	1.496080000000000	12.120000000000000	0	3
.728097000000000	.701676000000000	.728097000000000	.701676000000000	5	1	.026421000000000	6.500000000000000	1	0
.378350000000000	1.26032000000000	.378350000000000	1.26032000000000	4	5	-.881970000000000	7.176470588235290	3	1
1.52860000000000	.475885000000000	1.52860000000000	.475885000000000	8	3	1.052715000000000	12.000000000000000	1	1
1.56163000000000	1.43840000000000	1.56163000000000	.677228000000000	10	5	.884402000000000	8.040000000000000	1	2
1.51813000000000	2.60852000000000	1.51813000000000	2.60852000000000	8	8	-1.090390000000000	6.300000000000000	2	3
.678924000000000	.222999000000000	.678924000000000	.222999000000000	11	2	.455924999999999	6.250000000000000	0	0
.258953000000000	1.56863000000000	.258953000000000	1.56863000000000	6	5	-1.309677000000000	9.375000000000000	1	0

1.4.2 Variable View

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
xG	Numeric	17	15		None	None	12	Right	Scale	Input
xGA	Numeric	17	15		None	None	12	Right	Scale	Input
npxG	Numeric	17	15		None	None	12	Right	Scale	Input
npxGA	Numeric	17	15		None	None	12	Right	Scale	Input
deep	Numeric	2	0		None	None	12	Right	Scale	Input
deep_allowed	Numeric	2	0		None	None	12	Right	Scale	Input
npxGD	Numeric	18	15		None	None	16	Right	Scale	Input
ppda	Numeric	17	15		None	None	15	Right	Scale	Input
missed	Numeric	1	0		None	None	12	Right	Scale	Input
scored	Numeric	1	0		None	None	12	Right	Scale	Input

1.4.3 Normality Test



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
scored	.179	216	.000	.922	216	.000

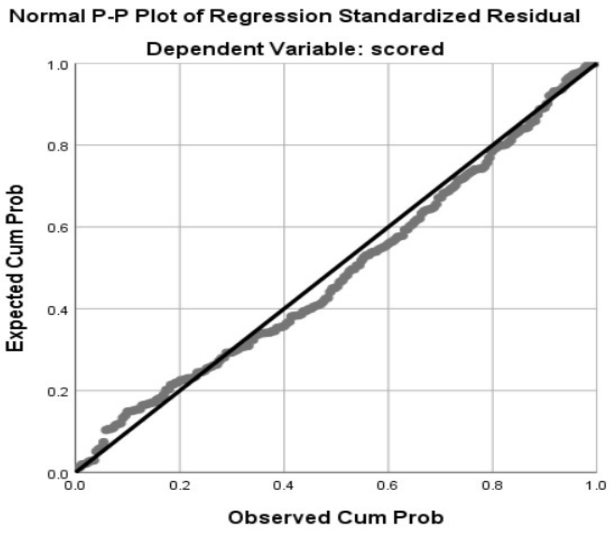
a. Lilliefors Significance Correction

Mean	Std Dev	Minimum	Maximum	N	Variance	Skewness	Kurtosis
1.9768519	1.4057759	0	6.0000000	216	1.9762059	0.5382379	-0.2976916

For the normality test of the dependent variable, based on visual representation, the histogram seems to right skew and are not normally distributed. This being confirm by the Shapiro-Wilk normality test that show the significant value is < 0.05 which mean that we rejected null hypothesis (data is normally distributed) and accept alternate hypothesis which mean the data is not normally distributed.

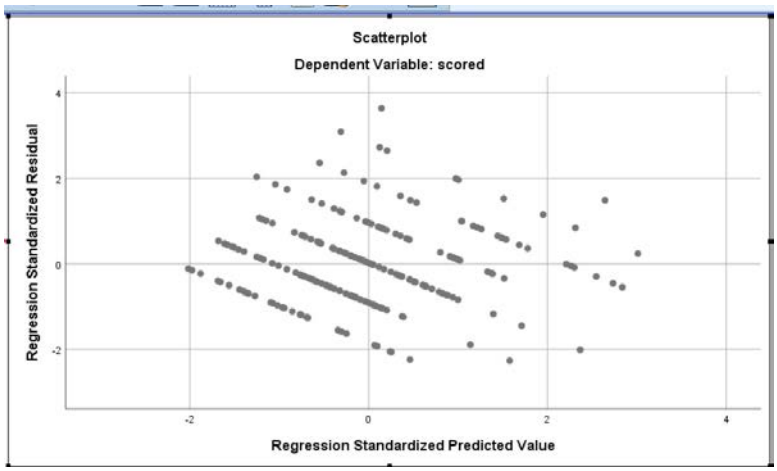
The value of skewness and kurtosis in the data also show that skewness is 0.538 which is >0 and its indicated the most of the distribution is at the left and its has longer right tail. Kurtosis in other hand, show negative value -0.298 and often called as platykurtic distribution that can be interpreted as extreme event probability is very low to happen in this distribution and very seldom the value will deviate far from its mean.

1.4.4 Linearity Test



Even though the dependent variable is not normally distributed, the independent variable have straight line relationship with dependent variable. this can be confirmed visually on above P-P plot where it show that all data point is near the regression line. Therefore, the linearity did not violate the normality assumption.

1.4.5 Homoscedasticity of Error term (Equal Variance)



On above picture, its show about homoscedasticity of the data based on predicted value (IV) and dependent variable (DV). This chart does not show any violation of equal variance assumption but its not very clear. Even though there is striking pattern of negative straight lines but this is because the dependent variable is on low value of 0 to 10 (as football is low scoring games) therefore its is believe that the model is adequate and unbiased in representing the populations.

1.4.6 Absent of Multicollinearity

Model	Collinearity Statistics	
	Tolerance	VIF
1		
	1.000	1.000
2		
	.800	1.249
	.800	1.249
3		
	.762	1.312
	.800	1.250
	.936	1.068

Above table show that the multicollinearity between the independent variable on final model. It shows that the tolerance level is acceptable and VIF (Variance Inflation Factor) are below 10. Therefore, it can be concluded that there is absent of multicollinearity between independent variable in the final model does not violate the no multicollinearity assumption (Tabachnick, Fidell, & Ullman, 2007).

1.4.7 Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.629[a]	.395	.392	1.096
2	.644[b]	.415	.409	1.080
3	.654[c]	.428	.420	1.071

a. Predictors: (Constant), xG

b. Predictors: (Constant), xG, deep

c. Predictors: (Constant), xG, deep, missed

Out of 9 variable we feed into SPSS, its has selected only 3 predictors into the equation. This is due to only these 3 predictors are making significant change on r^2 and adding the rest of not selected predictor only make the model redundant and unstable. The model r^2 is 0.428 which indicate from this dataset its able to explained almost 43% variance in (scored) and the rest (57%) are still hidden. The reason for this as this is a public dataset and not all important variable are reveal due to factor such as big analytics company are

not readily to give away some important variable to public without some remuneration of their well-design data collection.

1.4.8 Analysis of Variance (ANOVA)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	167.946	1	167.946	139.879	.000[b]
	Residual	256.939	214	1.201		
	Total	424.884	215			
2	Regression	176.267	2	88.133	75.507	.000[c]
	Residual	248.617	213	1.167		
	Total	424.884	215			
3	Regression	181.831	3	60.610	52.867	.000[d]
	Residual	243.053	212	1.146		
	Total	424.884	215			

a. Dependent Variable: scored

b. Predictors: (Constant), xG

c. Predictors: (Constant), xG, deep

d. Predictors: (Constant), xG, deep, missed

Hypothesis 4

H_0 - The model is not adequate
H_1 - The model is adequate

From the Analysis of Variance (ANOVA) table above, we achieve high f score (>50) for all the predictor where we can assume the group mean are spread out more and reflect differences at the population level. the final model p-value is less than 0.05 therefore its mean that there is less than 5% it's not due by chance of sampling error. On hypothesis 4, there is enough to warrant us rejecting the null hypothesis and assume that this model is adequate.

1.4.9 Coefficients of Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.210	.167		1.259	.209					
	xG	.989	.084	.629	11.827	.000	.629	.629	.629	1.000	1.000
2	(Constant)	.535	.205		2.613	.010					
	xG	1.099	.092	.699	11.925	.000	.629	.633	.625	.800	1.249
	deep	-.049	.018	-.156	-2.670	.008	.156	-.180	-.140	.800	1.249
3	(Constant)	.297	.230		1.290	.198					
	xG	1.144	.094	.727	12.223	.000	.629	.643	.635	.762	1.312
	deep	-.048	.018	-.154	-2.649	.009	.156	-.179	-.138	.800	1.250
	missed	.150	.068	.118	2.203	.029	-.045	.150	.114	.936	1.068

a. Dependent Variable: scored

(a)Estimated Multiple Linear Stepwise Regression Model

$\boldsymbol{Y}=0.210+0.989\boldsymbol{X_1}-0.049\boldsymbol{X_2}+0.150\boldsymbol{X_3}$

Where,

***Y**=Goal scored against opponent (scored)*

$\boldsymbol{X_1}$*=Expected Goal (xG)*

$\boldsymbol{X_2}$= Passes completed within an estimated 20 yards of goal (*deep)*

$\boldsymbol{X_3}$= Number of goals missed in games *(missed)*

(b)Explanation on Hypothesis 1

Hypothesis 1

H_0 - Expected goal (xG) does not have effect on goal (scored)
H_1 - Expected goal (xG) have effect on goal (scored)

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	xG	.989	.084	.629	11.827	.000	.824	1.154

Based on the table above, the p-value of expected goal (xG) is less than <0.05. Therefore, we can reject null hypothesis and accept alternate hypothesis that xG does have effect on goal (scored). This predicted variable influence 39.5% of final model and for every increment value in X_1 it will increase 0.989 in (scored) and it have largest effect on this final model based on stepwise method.

(c) Explanation on Hypothesis 2

Hypothesis 2

H_0 - Passes completed in 20 yard (deep) has no influence on goal (scored)
H_1 - Passes completed in 20 yard (deep) has influence on goal (scored)

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
2	deep	-.049	.018		-2.670	.008	-.086	-.013

On the other hand, the p-value of passes completed in 20 yard (deep) is less than <0.05. Therefore, we can reject null hypothesis and accept alternate hypothesis that deep does have effect on goal (scored). This predicted variable influence 2% of final model and for every increment value in X_2 it will decrease 0.049 in (scored) and it have second largest effect on this final model based on stepwise method.

(d) Explanation on Hypothesis 3

H_0 - Goal missed (missed) will not have impact on goal(scored)
H_1 - Goal missed (missed) will have impact on goal(scored)

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
3	missed	.150	.068	.118	2.203	.029	.016	.284

Finally, the p-value of goal missed (missed) is less than <0.05. Therefore, we can reject null hypothesis and accept alternate hypothesis that missed does have effect on goal (scored). This predicted variable influence 1.3% of final model and for every increment value in X_3 it will increase 0.15 in (scored) and it have third largest effect on this final model and the last variable that stepwise method included in the model.

1.5 CONCLUSION & REFERENCE

In order for the final model to be useful, the four-model assumption need to be met. In this analysis, 3 out of 4 assumption was met and did not violate the model assumption except for normality of dependent variable. This is normal since football is low scoring games where most of the score only single goal followed by 2 goal.

From this analysis, we can find some interesting finding such as how much expected goal (xG) influence the probability on scoring a goal where is influence almost 40% of our final model which is very big factor. Other interesting finding is how passes completed in 20 yard (deep) have negative effect on scoring goal. The general assumption among football fan and analyst is that the closer your team into the goal, the higher change are it can convert into goal. But it was proven otherwise by our model where it decreases the chance of scoring goal. The only way to explained this is that we can assumed the closer passes into opponent goal area, the more body and more aggressive the opponent in tackling so it will harder for a player to score from 20 yard.

Therefore, since out of 9 variable feed into this model can only explain 42.8% of the variance, there is need of improvement on this model such as finding important parameter thru video analysis technique in order to find the other hidden pattern. The more improved dataset with important parameter from the video analysis technique can improved this model thus able to get better performance outcome.

1.6 Reference

Hair, F. Joseph, W.Black, B.J. Babin & R.E.Anderson.(2014). Multivariate data analysis. 7th Ed Essex: Pearson Education Limited

Research and Markets. (2020) Sports Analytics Market - Growth, Trends and Forecast (2020 - 2025) [Online]. Available from https://www.researchandmarkets.com/reports/4703456/sports-analytics-market-growth-trends-and. [Accesses on 01 June 2020].

Sergi Lehkyi,2019, Football Data: Expected Goals and Other Metrics https://www.kaggle.com/slehkyi/extended-football-stats-for-european-leagues-xg

Tabachnick, B.G., Fidell, L.S. and Ullman, J.B., (2007). *Using multivariate statistics* (Vol. 5). Boston, MA: Pearson.

2.0 Factor Analysis

(a) The purpose of performing factor analysis

On the Question 1, we use Multiple Linear Regression in order for us to establish the causal relationship of dependent variable (scored) with independent variable. If the structure of variable to be analyze, then performing factor analysis on this dataset to study the interdependence between each of the independent variable are most suitable. Its also used to understand which variable are correlated with each other. Since in this dataset, some of the variable are derived from each other for example expected goal(xG) and non-penalty expected goal(npxG). The npxG are derived from xG value where npxG has excluded penalty goal from xG value to give analyst more accurate view on what happen in the field. As such, factoring analysis can be used to reduce the dimension of this dataset to avoid bias and redundancy by combining the similar variable into single component.

(b) Why the non-metric independent variables are not allowed in factor analysis.

Non-metric independent variables are not allowed in factor analysis due to non-metric does not distance between scale values. They do not possess a meter with which distance between scale values can be measured. This proved to be problematics when perform factor analysis where its produce factor loading based on Pearson correlation coefficient developed between metric variable.

(c) Perform factor analysis on SPSS

(i)Correlation Metrix

		xG	xGA	npxG	npxGA	deep	deep_allowed	npxGD	ppda	missed
Correlation	xG	1.000	-.276	.943	-.251	.447	-.115	.821	-.053	-.252
	xGA	-.276	1.000	-.279	.924	-.266	.326	-.677	.264	.614
	npxG	.943	-.279	1.000	-.266	.423	-.100	.870	-.037	-.252
	npxGA	-.251	.924	-.266	1.000	-.224	.348	-.706	.251	.545
	deep	.447	-.266	.423	-.224	1.000	-.288	.425	-.275	-.130
	deep_allowed	-.115	.326	-.100	.348	-.288	1.000	-.251	.327	.201
	npxGD	.821	-.677	.870	-.706	.425	-.251	1.000	-.155	-.464
	ppda	-.053	.264	-.037	.251	-.275	.327	-.155	1.000	.168
	missed	-.252	.614	-.252	.545	-.130	.201	-.464	.168	1.000

On this table, its show correlation on all 9-independent variable on this dataset.

(ii) Measure of Sampling Adequacy (MSA)

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.833
Bartlett's Test of Sphericity	Approx. Chi-Square	1606.172
	df	105
	Sig.	.000

The Kaiser-Meyer-Olking (KMO) statistic should be greater than 0.600 and the Bartlett's test should be significant (e.g. $p < .05$) which mean that the alternate hypothesis is accepted which mean there is correlation between independent variable.

(iii) Communalities

Communalities

	Initial	Extraction
xG	1.000	.935
xGA	1.000	.906
npxG	1.000	.955
npxGA	1.000	.885
deep	1.000	.632
deep_allowed	1.000	.572
npxGD	1.000	.964
ppda	1.000	.631
missed	1.000	.608

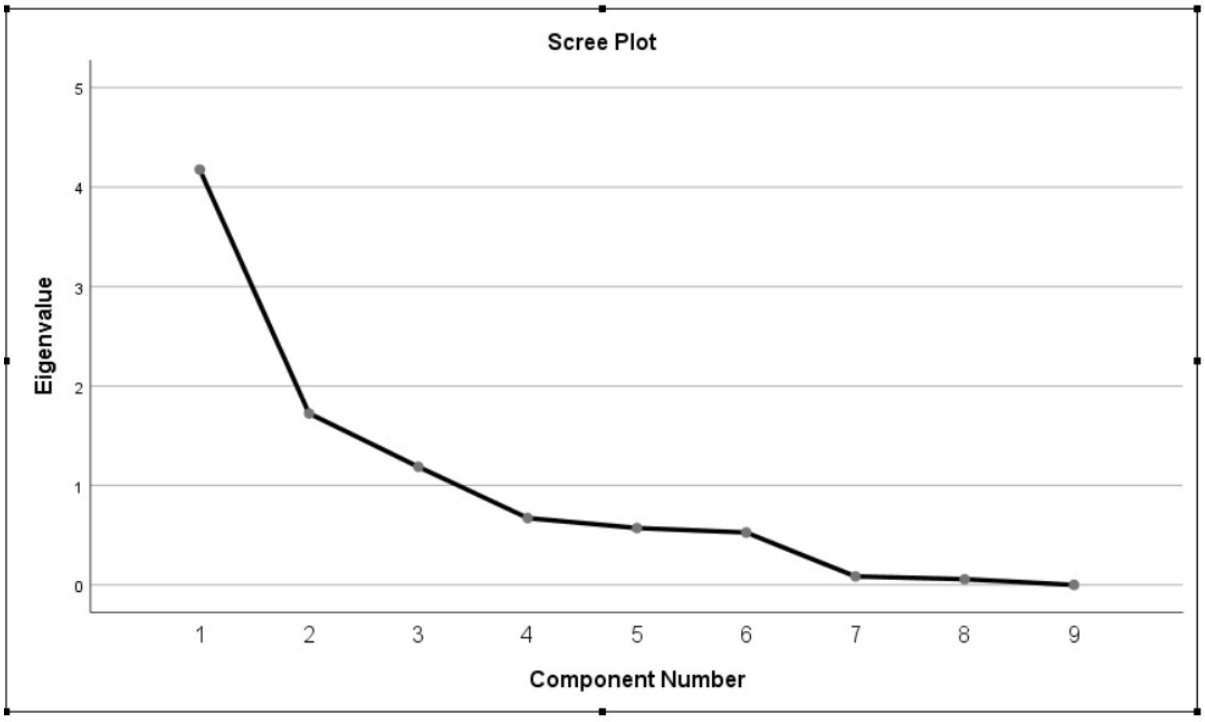
Extraction Method: Principal Component Analysis.

The above table is Communalities output using Principal component analysis method and all 9 variable was included and no variable was deleted due to all the value is above 0.5.

(iii) Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.176	46.400	46.400	4.176	46.400	46.400
2	1.724	19.155	65.555	1.724	19.155	65.555
3	1.187	13.194	78.749	1.187	13.194	78.749
4	.672	7.467	86.216			
5	.572	6.354	92.570			
6	.526	5.849	98.419			
7	.085	.946	99.365			
8	.057	.635	100.000			
9	-1.665E-16	-1.850E-15	100.000			

Extraction Method: Principal Component Analysis.



The SPSS has found 3 significant factor that contributed to 46.40% for factor 1, 19.15% for factor 2, 13.19% for factor 3 and 78.75% of cumulative total variance and can be visualize by the scree plot (eigenvalue above 1).

(iv) Component Matrix (Non-rotation & Rotation)

	Component		
	1	2	3
xG	.743	.618	.027
xGA	-.790	.471	.244
npxG	.752	.624	-.009
npxGA	-.775	.474	.245
deep	.542	.187	.550
deep_allowed	-.412	.382	-.507
npxGD	.948	.216	-.132
ppda	-.316	.395	-.613
missed	-.608	.350	.340

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Rotated Component Matrix[a]

	Component		
	1	2	3
xG	.957	-.136	-.027
xGA	-.167	.909	.228
npxG	.964	-.157	.004
npxGA	-.155	.901	.223
deep	.544	.018	-.579
deep_allowed	-.034	.234	.719
npxGD	.787	-.575	-.111
ppda	.032	.124	.784
missed	-.126	.768	.039

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

a. Rotation converged in 5 iterations.

The above table show component matrix before rotation and below table show component matrix after applied the varimax method of rotation.

(d) Group metric independent variables into factors.

Rotated Component Matrix[a]

	Component		
	1	2	3
npxG	.964	-.157	
xG	.957	-.136	
npxGD	.787	-.575	-.111
xGA	-.167	.909	.228
npxGA	-.155	.901	.223
missed	-.126	.768	
ppda		.124	.784
deep_allowed		.234	.719
deep	.544		-.579

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

a. Rotation converged in 5 iterations.

Based on table above, we can group the variable into 3 factors. The first factors member are as below:

Factor 1 *attacking moves*: npxG, xG, npxGD,

Factor 2 *defensive moves*: xGA, npxGA, missed

Factor 3 *buildup moves*: ppda, deep_allowed, deep

Then we can see that factor 1 are more on attacking metrics of the dataset which explained all the attacking value made into opponent area. Meanwhile, factor 2 are the opposite of factor 1 where we can see the defensive value are combine together in factor 2. Finally the factor 3 are the movement or buildup of the team toward opponent area here value such as ppda, deep and deep_allowed are put together as members.

(e) Factor analysis outputs

(i) Interpret the meaning of communality

Communality mean total amount of variance an original variable share with all other variable included in the analysis. Higher communality value indicates the larger amount of the variance in variable has been extracted by the factor solution.

(ii) Interpret the meaning of eigenvalue

Eigenvalue mean Column sum of squared loadings for factor and its also referred as the latent root. Its represents the amount of variance accounted for by factor.

(f) By assessing the relevant outputs obtained in part (c):

(i) How the factorability of your dataset can be improved prior to the task of factor analysis

The factorability of dataset can be improved by making sure the data matrix has sufficient correlation to justify application of factor analysis. Researcher can use method such as Bartlett's test of sphericity and measure of sampling adequacy (MSA). A ($sig<0.05$) indicates the sufficient correlations exist among the variables to proceed and MSA (>0.05) indicated the adequate sampling.

(ii) Explain what is meant by factor cross-loading

Component Matrix[a]

	Component		
	1	2	3
npxGD	.948	.216	-.132
xGA	-.790	.471	.244
npxGA	-.775	.474	.245
npxG	.752	.624	
xG	.743	.618	
missed	-.608	.350	.340
ppda	-.316	.395	-.613
deep	.542	.187	.550
deep_allowed	-.412	.382	-.507

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Cross-loading mean that a variable has two or more factor loadings exceeding the threshold value deemed necessary for inclusion in the factor interpretation process. Table above showed one of the example of cross-loading in this dataset.

(iii) How can the problem of cross-loading be reduced?

Researcher can reduce the cross-loading problem by employing the rotation method such as varimax. Varimax can improved the structure considerably by loading very high on single factor. However, if the problem persists, the researcher can consider possible deletion of the variable.