Masters Programmes

Group Work Assignment Cover Sheet

Submitted by: 2016407, 2022502, 2022558, 2029468, 2084551, 2092650

Group Number: 22

Date: 22 March 2021

Module Title: Advanced Data Analysis

Module Code: IB98D0

Date/Year of Module: 2021

Submission Deadline: 22 March 2021

Question: 3

I declare that this work is being submitted on behalf of my group, in accordance with the University's Regulation 11 and the WBS guidelines on plagiarism and collusion. All external references and sources are clearly acknowledged and identified within the contents. No substantial part(s) of the work submitted here has also been submitted in other assessments for accredited courses of study and if this has been done it may result in us being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work.

Contents

Table of Tables

Table of Figures

Executive summary

This report provides an analysis of how players can improve their league index by improving certain metrics, what common features are exhibited by some groups of players and how specific metrics can be combined logically so that players can improve on some variables all at once. Methods of analysis include Multiple linear regression, Multiple discriminant analysis, Logistic regression, Cluster analysis and Factor analysis. The report finds that the speed of gameplay and technique-complexity are the most influential variables if players want to improve their level. There are six distinctive clusters of players which can further be divided into three groups, namely beginners, common gamers and experts in the dataset. Five hidden factors represent the association of the different variables: agility, complexity, offensiveness, defensiveness and experience. The report also examines the limitations of the analyses. Some of the limitations include: League Index is an ordinal variable and is not favourable in regression analysis, and hence some assumptions get violated; the league index 7 is underrepresented compared to other leagues.

Section 1 Regression Analysis

Which player attributes have the most significant impact on a players skill level? Multiple linear regression (MLR) was utilised to discover the underlying relationships, and the results were confirmed using multiple discriminate analysis (MDA) and logistic regression (LR). The results of the regression are given below.

1.1 Results

Based on the regression results, the most influential variables in determining a player's league index can be seen below in table 1, in descending order of importance. The standardized beta coefficients determine the weight of each variable's influence. So, the greater the absolute value of a variable's standardized beta coefficient (SBC), the more impact it will have in determining a player's league index.

Variable	Standardised Beta Coefficients (SBC)
APM	.259
AssignToHotKeys	.199
TotalHours	.163
NumberOfPACs	.162
ActionLatency	-.128
GapBetweenPACs	-.099

Table 1 Regression Variable Beta Coefficients

From Table 1 it can be seen that APM, or actions per minute, has the most significant influence on a player's league index. APM's large SBC is logical because the faster a player can collect resources, build defenses, build an army, and prepare for war, the more likely they will win. Speed of collecting resources is an important skillset required to succeed in this game.

All significant variables, except total hours, share one thing in common with APM, they all relate to a player's speed. AssignToHotKeys describes the number of shortcuts a player has assigned, allowing for faster actions. NumberOfPACs examines the number of attention shifts, followed by an action. NumberOfPACs examines a player's ability to simultaneously handle many different actions, which increase gameplay speed. The only variable that does not relate to a player's speed is TotalHours. Although it does not represent speed, its relationship with a player's speed is apparent: the more time players spend playing the game, the quicker their actions and reflexes will become.

The results from MDA and LR confirms this result. Table 6 shows the results from the MDA. Action latency has the most significant influence in discriminating between the two-player groups. The MDA's results are consistent with MLR. However, two new variables are introduced: WorkersMade and ComplexAbilitiesUsed. MDA reveals a critical aspect, the complexity of play which is equally essential for improving a player's expert level. Also, HoursPerWeek is introduced and has a negative relationship with the league index. This may indicate that players can play too much, which would create fatigue and impair judgement. Table 7 shows the results of the LR analysis. The results are congruent with MDA, indicating that speed and technique complexity determine a player's league index.

If players want to increase their league index, they need to work on their in-game speed and multitasking ability. A player can increase their speed in a couple of ways. First, players can spend more time playing and practicing. If players play more, actions will become more second nature, muscle memory will take over, and overall player speed will increase. Second, players can purchase gaming hardware that includes hotkeys, which will allow for faster actions.

1.2 Building the Models

The purpose of the MLR, MDA, and LR was to uncover variables that had a considerable influence on LeagueIndex. The process below outlines the modelling steps followed to ensure quality.

1.2.1 Research Design

The data used in the MLR, MDA, and LR was cleaned and prepared, as seen in *Data Preparation* in the appendix. Outliers and missing values were amended. The 684-observation sample size was also examined and deemed adequate for generalization. Representation in each league was not equal, but there were enough observations in each league to continue modelling. The ratio of observations to variables was 38:1, which exceeds the optimal ratio of 20:1.

1.2.2 Assumptions

Each model has different assumptions, so the assumptions Table 2 will be listed per model. A detailed evaluation of each model's assumptions is in *Modeling Assumptions* in the appendix.

Assumption Required	Assumption Broken: 'X'		
	MLR	MDA	LR
Linearity	X	X	
Homoscedasticity	X	X	
Multicollinearity	X	X	X
Normality of Residuals			
No Correlated Residuals			
Ill-Conditioned		X	X

Table 2 Assumption check

Assumptions were breached for all models but would not be rectified unless the model results also broke their assumptions.

1.3 Estimate Model & Fit

Because MDA and LR only confirmed the results of the MLR, the focus will be on MLR. The regression was built by performing a simple linear regression for each variable and then selecting

the most informative variables. Variables were selected while considering two things: The variables' r-squared and the total number of selected variables. Building a solid model with informative variables was necessary, but the principle of parsimony also needed to be considered. R-squared was chosen over p-value because they were synonymous in this instance, and r-squared was easier to visualize. Figure 1 *(left)* shows the r-squared for all the simple linear regressions. APM had the highest r squared. Figure 1 *(right)* shows the multiple linear regressions' r-squared changing as variables are added one at a time according to their r-squared value in Figure 1 *(left)*. In Figure 1 *(right)*, NumberOfPACs adjusted r-squared represents a model containing APM and NumberOfPACs, and ActionLatency's adjusted r-squared represents a model containing APM, NumberOfPACs, and ActionLatency. This pseudo-stepwise method was continued until a model containing all variables was built. Using the results seen in Figure 2, the model that optimized the adjusted r-squared while considering the principle of parsimony was at variable TotalHours. The adjusted r-squared at TotalHours was 1% less than the maximum r-squared and contained three fewer variables. So, the initial regression model built contained all variables up to TotalHours.

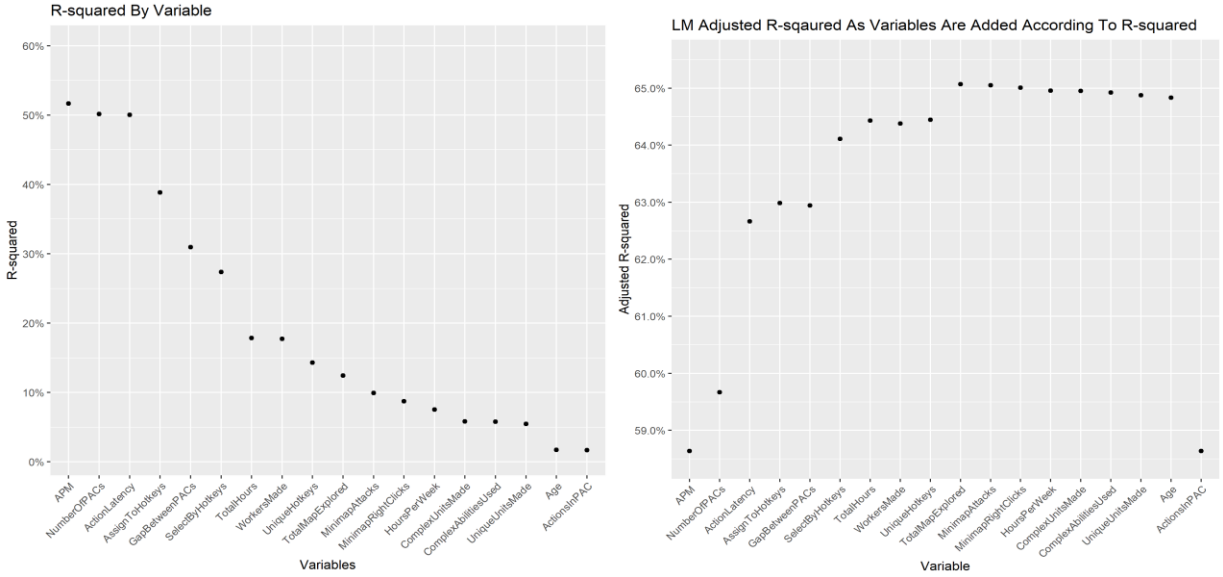

Figure 1 R-square By Variable (left) and MLR Adjusted R-squared as variables are added according to R-squared (right)

The initial model built contained the variables APM, NumberOfPACs, ActionLatency, AssugnToHotKeys, GapBetweenPACs, SelectByHotKeys and TotalHours. The 'Enter' method in SPSS was used to create the model because the user predefined the predictor variables. The initial model, Model 1, performed as was expected from Figure 1, showing an adjusted r-squared of 64.7%. Although, the variable SelectByHotKeys showed an insignificant p-value of .588, indicating that it should be removed. The .127 tolerance and 7.84 VIF for APM also created potential reasons to remove the variable for multicollinearity. The next model created, Model 2, was identical to Model 1 but did not include SelectByHotKeys. It is seen below in *Figure 2* and *3*.

With SelectByHotKeys removed, the adjusted r-squared value stayed the same. The variable APM's tolerance and VIF also improved. All variables in Model 2 are significant and should be kept in the model. Model 2's Durbin-Watson is also close to 2, indicating that autocorrelation is not present. Model 2 showed statistical significance and, as seen at the beginning of this exert, practical significance.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.804[a]	.647	.643	1.053	.647	206.456	6	677	.000	1.966

a. Predictors: (Constant), APM, TotalHours, AssignToHotkeys, GapBetweenPACs, NumberOfPACs, ActionLatency

b. Dependent Variable: LeagueIndex

Figure 2 Model 2 Model summary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.966	.488		4.033	.000		
	NumberOfPACs	286.551	80.926	.162	3.541	.000	.250	3.995
	ActionLatency	-.010	.004	-.128	-2.546	.011	.206	4.857
	AssignToHotkeys	1692.849	251.765	.199	6.724	.000	.594	1.682
	GapBetweenPACs	-.009	.003	-.099	-3.050	.002	.497	2.011
	TotalHours	.001	.000	.163	6.381	.000	.799	1.251
	APM	.010	.002	.259	6.404	.000	.319	3.134

a. Dependent Variable: LeagueIndex

Figure 3 Model 2 Coefficients

Model 2 also did not show signs of a multicollinearity problem or autocorrelation. Model 2 was cross-checked in SPSS using a backwards stepwise method with all variables with 10% r-squared or more. The results confirmed the variables chosen and can be seen in the appendix in *Table AA and AB.* As seen below in Figure 4 *(left and middle),* the residuals show an almost-perfect normal distribution. Figure 4 *(right)* seems worrying, as there is a clear pattern, although, considering the almost categorical nature of LeagueIndex, the seven lines in the residuals makes sense. With this in mind, we can ignore the seven lines and interpret the results: Figure 4 *(right)* seems to fluctuate around a zero mean, is homoscedastic, and does not show a pattern.

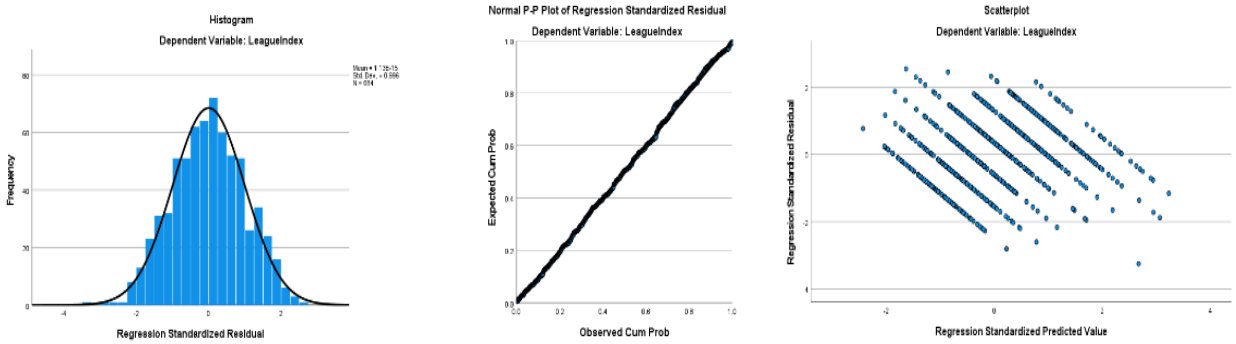

Figure 4 Histogram of Residuals (left), Normal P-P Plot of Residuals (Middle) and Scatter Plot of Residuals (right)

1.3.1 Interpretation

The results' interpretation was based on the Standardized Coefficients Beta seen in Figure 3 and is seen above in the *Results* section.

1.4 Results from MDA

The objective of performing a Multiple discriminant analysis (MDA) is to support the results found in MLR. The results from MLR indicate the variables that are most significant in describing the

Leagueindex. MDA works best when there are fewer groups (2-3). Multiple combinations of groups were used to build the best model. The final dataset that was used in the model has the following characteristics: The league index was reduced from 7 to 2 groups and all the independent variables except APM and Actions in Pac were included. The first group contain Leagues 1, 2 and 3 and the second group contain leagues 4, 5, 6 and 7. This grouping was a result of attempting multiple combinations of groups and variables and finalizing the model with the best classification accuracy and ease of interpretation. The results for other combinations of groups are added in the *appendix table A12:A16.* The summary of the best model is outlined below.

Stepwise estimation was used to perform the analysis as the objective is to identify the most discriminating variables and also a large number of potentially relevant variables are available for analysis. One significant discriminant function is identified within the two groups explaining 100% of the variance.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.539	418.934	8	.000

Table 3 Wilk's lambda

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.855[a]	100.0	100.0	.679

a. First 1 canonical discriminant functions were used in the analysis.

Table 4 Eigenvalues

Functions at Group Centroids

LeagueIndex	Function 1
1	-.956
4	.891

Unstandardized canonical discriminant functions evaluated at group means

Classification Results[a]

		LeagueIndex	Predicted Group Membership 1	4	Total
Original	Count	1	281	49	330
		4	71	283	354
	%	1	85.2	14.8	100.0
		4	20.1	79.9	100.0

a. 82.5% of original grouped cases correctly classified.

Table 5 Function at Centroid (left) and Classification results(right)

As shown in Table 5 above, the discriminant function successfully classifies the two groups separately with a substantial difference of 1.85 between the group centroids. The classification accuracy is 82.5%. The accuracy for group 1(LeagueIndex 1,2,3) is 85.2% and for group 4(LeagueIndex 4,5,6,7) is 79.9%. 120/684(17.5%) cases have been misclassified. This confirms that the two groups have significant differences between them, and MDA has been successful in capturing this difference.

As can be seen in Table 6, Action Latency is the most important variable. The variables that discriminate the two groups can be interpreted as the speed of play represented by SelectByHotkeys, AssignToHotkeys, MiniMapAttacks and ActionLatency, time spent represented by TotalHours and HoursPerWeek and complexity of play represented by ComplexAbilitiesUsed and WorkersMade. The time spent and the speed of play are related to each other in a way that if a player spends more time, the player gets quicker in reacting to changes in the gameplay. The HoursPerWeek has a negative coefficient, indicating that a player cannot expect to improve their gameplay by just spending more time constantly playing.

Standardized Canonical Discriminant Function Coefficients

	Function 1
HoursPerWeek	-.193
TotalHours	.330
SelectByHotkeys	.207
AssignToHotkeys	.294
MinimapAttacks	.135
ActionLatency	-.530
WorkersMade	.196
ComplexAbilitiesUsed	.117

Classification Function Coefficients

	LeagueIndex 1	LeagueIndex 4
HoursPerWeek	.129	.099
TotalHours	.001	.003
SelectByHotkeys	174.578	285.329
AssignToHotkeys	12551.288	15578.147
MinimapAttacks	4168.377	6200.771
ActionLatency	.395	.339
WorkersMade	5839.311	6602.926
ComplexAbilitiesUsed	4676.874	5615.581
(Constant)	-22.290	-20.852

Fisher's linear discriminant functions

Table 6 Discriminant function coefficients and classification function coefficients

After a point, there might be a fatigue factor that might affect the player's gameplay and worsen their abilities. This conclusion supports the results found in the MLR model. The speed of play and time spent is directly related and there is an extra factor of complexity that is revealed by MDA.

1.5 Results of Logistic Regression

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 7[9]	TotalHours	.002	.001	15.221	1	.000	1.002
	SelectByHotkeys	164.050	82.660	3.939	1	.047	1.762E+71
	AssignToHotkeys	2599.434	1144.967	5.154	1	.023	.
	MinimapAttacks	7838.390	2301.240	11.602	1	.001	.
	NumberOfPACs	1537.231	335.347	21.013	1	.000	.
	WorkersMade	1164.192	480.141	5.879	1	.015	.
	ComplexUnitsMade	3634.129	1821.909	3.979	1	.046	.
	Constant	-8.645	1.118	59.830	1	.000	.000

Table 7 Variables in the Logistic regression model

The variables identified as significant in the Logistic Regression model conducted on two groups (LeagueIndex 1-2-3, and LeagueIndex 4-5-6-7) are consistent with the results found in the MDA analysis, with variables explained in three different aspects: Speed, Timespent and Complexity.

Setting the cut-off value at 0.5, the LR model generates a satisfying classification outcome with 86.2% of the observation correctly classified in the training set and 81.7% in the validation set, which indicates a significant difference between these two groups explained by the most prominent variables and can be generalised to an independent dataset *(results in table A11 in appendix*).

1.6 Validation of MLR

The results were validated using a 50% random, stratified subset of the prepared data. Validation of Model 2 was performed by rerunning the model using the subset, and this model will be referred to as Model V2. The results are in tables *A7* and *A8* in *Appendix A.* The validation shows that Model 2 and Model V2 are very similar. The adjusted r-squared between the two models is only 1.9% different. The variables standardized coefficients beta are slightly different. APM is still the most important predictor, but now NumberOfPACs has a more significant influence on LeagueIndex than AssignToHotKeys and TotalHours. The order of importance for all other variables stayed the same. Model V2 also shows no signs of a multicollinearity problem, and its residuals are normally distributed and almost the same as Model 2 *(Appendix Figure A10, A11 and A12).*

While Model V2 performed almost the same as Model 2, there is a difference in variable significance. Model V2 sees GapBetweenPACs and ActionLatency as insignificant (p-value > 0.05). The insignificant p-values may have been because these variables need additional observations to have statistical power. However, because of this difference in model results, the regression results cannot be generalized to the population.

1.6.1 Validation of MDA & LR result

The dataset was sampled in a stratified manner composed of 50% training and 50% test data points. The results of MDA on this dataset are in appendix *Table A10*. The accuracy in the training set is 85.3% and in the test set is 80.5% which indicates that the MDA model is fairly accurate in classifying the different groups. This also means that the model can be generalized to an independent dataset.

Section 2 Factor Analysis

Continuing from our discussion of the variables that were the most important in explaining the league of a player, further analysis was performed to identify the various groups of variables that contributed to the expert levels. The endeavour is to explain the groups of variables in simple terms understandable to readers who do not have expert knowledge of the gaming world. The objective was to perform a factor analysis to extract different groups of variables.

There were multiple analyses performed to extract the best factors that had minimal cross-loading and that had the most logical meaning for the factors. A PCA analysis was performed to decide how many variables were to be extracted. The scree plot dropped steadily after 5 factors (*See Appendix Figure B5*) and hence 5 factors were extracted and rotated with Promax (kappa = 8). The figures for other combinations of rotations and extraction methods are added in *Appendix Figure B1:B4*.

Communalities

	Initial	Extraction
HoursPerWeek	1.000	.771
TotalHours	1.000	.699
APM	1.000	.903
SelectByHotkeys	1.000	.641
AssignToHotkeys	1.000	.598
UniqueHotkeys	1.000	.480
MinimapAttacks	1.000	.545
MinimapRightClicks	1.000	.586
NumberOfPACs	1.000	.814
GapBetweenPACs	1.000	.588
ActionLatency	1.000	.740
ActionsInPAC	1.000	.756
TotalMapExplored	1.000	.701
WorkersMade	1.000	.445
UniqueUnitsMade	1.000	.650
ComplexUnitsMade	1.000	.761
ComplexAbilitiesUsed	1.000	.732

Extraction Method: Principal Component Analysis.

Pattern Matrix[a]

	Component				
	1	2	3	4	5
SelectByHotkeys	.927				
NumberOfPACs	.843				
AssignToHotkeys	.838				
APM	.817				
ActionLatency	-.718				
UniqueHotkeys	.677				
GapBetweenPACs	-.534				
ActionsInPAC		.931			
MinimapRightClicks		.744			
WorkersMade		.440			
ComplexAbilitiesUsed			.834		
ComplexUnitsMade			.832		
TotalMapExplored				.769	
UniqueUnitsMade				.707	
MinimapAttacks				.658	
HoursPerWeek					.906
TotalHours					.751

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 8 iterations.

Table 8 Communalities of PC extraction(left) and Pattern matrix(right)

Table 8 shows that after rotation and extraction, both communalities and factor loading were acceptable. There were no significant cross-loadings and all the communalities were above 0.5 except for WorkerMade (0.445) and UniqueHotkeys(0.480). Since these were close to 0.5, they were treated as borderline cases and included in the factor solution.

2.1 Factor Analysis – Factors Interpretation

Factor 1: SelectByHotkeys, NumberOfPACs, AssignToHotkeys, APM, ActionLatency, UniqueHotkeys, GapBetweenPACs.

This factor is related to a player's reaction abilities, i.e., the recognition cycle. Players with higher comprehensive recognition ability and cognitive flexibility tend to perform better in Real-time strategy (RTS) games like StarCraft[2]. A suitable name for this factor is '**Agility**'. More specifically, NumberOfPACs and GapBetweenPACs are both related to the screen movement, which is positively affected by the player's 'Agility'. Players with higher agility or speed tend to react faster to changes, are cognitively flexible and move the screen more frequently, therefore causing shorter gaps between PACs[2]. ActionLatency, the mean latency from the onset of a PACs to their first action in milliseconds, is also a consequential variable of 'agility'. High agility leads to smaller ActionLatency. APM, as in Actions Per Minute, is another strong indicator of agility. APM is a consequence of recognition ability and experience and can be improved by practice. As for hotkey related variables, they help players make actions in the game swiftly by pressing the hotkeys thereby reducing the time taken to perform complex actions through keystrokes.

Factor 2: ActionsInPAC, MinimapRightClicks, WorkersMade.

This factor is related to a player's strategic tendency to develop their resources i.e., the ability to develop the headquarter and collect resources. A suitable name for this factor is '**Defensiveness**'. Players with a defensive mindset do not necessarily perform better in terms of league index, as RTS games take both operation skills and attacking skills to win[1]. More specifically, ActionsInPAC reflects the frequency of the player's action in a single PAC. For players who focus on developing their headquarters, each of their PAC tends to contain more actions [4]. Similarly, to maintain more mine sites and generate more resources, additional workers must be produced, which would explain the inclusion of WorkersMade in this factor. As the players focus on operational strategy, players send their workers to operate the mine site causing an increase in the MinimapRightClicks and ActionsInPAC metrics.

Factor 3: ComplexAbilitiesUsed, ComplexUnitsMade.

This factor is related to a player's tendency towards complex gameplay. A suitable name for this factor is '**Complexity tendency**'. Players with a higher "complexity tendency" like to produce complex units and use complex abilities. This factor is not necessarily correlated with a player's performance on the league index because even though complex units are powerful and useful in the game, they take a large amount of time and resource to produce and are likely to slow down the player's action[5.]

Factor 4: TotalMapExplored, UniqueUnitsMade, MinimapAttacks.

This factor is related to a player's tendency for aggressive offensive actions in the game, i.e., the tendency to be aggressive towards the enemy and launching frequent attacks to distract the enemy from developing their resources. A suitable name for this factor is '**Offensive tendency**'. Players with a high offensive tendency do not necessarily perform better in terms of league index[3]. More specifically, players with a higher offensive tendency are more likely to send troops to explore the

map and detect enemy headquarters and therefore tend to have higher TotalMapExplored. A particular player can maintain only one unique unit at any time. The more often a player attacks their enemy, the more likely their unique unit is destroyed and rebuilt, therefore accumulating higher UniqueUnitsMade.

Factor 5: HoursPerWeek, TotalHours

This factor is related to a player's experience of the game. A suitable name for this factor is '**Experience**'. Players with a higher level of experience tend to perform better in terms of league index as a great way of improving APM as well as gaining strategical insights is through constant practice.

2.2 Factor Analysis - Validation

Two analyses were run separately on stratified data to check the stability of the FA analysis above.

Pattern Matrix[a,b]

	Component				
	1	2	3	4	5
SelectByHotkeys	.988				
AssignToHotkeys	.828				
APM	.821				
NumberOfPACs	.787				
UniqueHotkeys	.768				
ActionLatency	-.651				
WorkersMade	.478				
GapBetweenPACs	-.477	-.406			
ActionsInPAC		.911			
MinimapRightClicks		.796			
ComplexAbilitiesUsed			.905		
ComplexUnitsMade			.868		
MinimapAttacks				.781	
TotalMapExplored				.715	
UniqueUnitsMade				.666	
HoursPerWeek					.890
TotalHours					.746

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.[a,b]

a. Rotation converged in 7 iterations.

b. Only cases for which section = 1 are used in the analysis phase.

Pattern Matrix[a,b]

	Component				
	1	2	3	4	5
NumberOfPACs	.905				
SelectByHotkeys	.866				
APM	.828				
AssignToHotkeys	.804				
ActionLatency	-.790				
GapBetweenPACs	-.562				
UniqueHotkeys	.555				
ActionsInPAC		.973			
MinimapRightClicks		.631		.444	
WorkersMade		.509			
ComplexUnitsMade			.829		
ComplexAbilitiesUsed			.796		
TotalMapExplored				.775	
UniqueUnitsMade				.686	
MinimapAttacks				.642	
HoursPerWeek					.896
TotalHours					.713

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.[a,b]

a. Rotation converged in 10 iterations.

b. Only cases for which section = 0 are used in the analysis phase.

Table 9 Pattern matrix for validation dataset

From Table 9, only two variables, GapBetweenPACs and WorkersMade, were classified into different factors. Furthermore, both results were identical to the original FA.

Section 3 Cluster Analysis

3.1 Clustering

So far in the report, the relationships between the variables and the league index were explored. It was assumed that players in different league indexes had different capabilities and shortcomings. This raises the question if any groups of players share common features and span across league indexes. This idea was explored further through cluster analysis (CA).

Z-scores of all variables were used to do hierarchical and k-means CA to distinguish the players' identifiable sets. When performing hierarchical CA, the furthest-neighbour cluster method and squared Euclidean distance measure were used. The difference in cluster coefficients saw a

significant jump at stage 6. Thus, the cases would be grouped into 7 clusters (*Appendix: Figure C6*). Table 10 is the result of k-means. Two clusters contain less than 5 cases, indicating that there are outliers in the dataset. As CA is sensitive to outliers, it is necessary to use the dataset without outliers before carrying out cluster analysis.

Cluster Index	1	2	3	4	5	6	7
Number of Cases	144	107	187	64	189	4	1

Table 10 Number of cases in clusters(original dataset)

Multicollinearity may also influence the result of cluster analysis. From the VIF results, there are three variables with VIF values greater than 10: APM, SelectByHotkeys, and NumberOfPACs. Based on the dataset with outliers removed, a new dataset was created without these three variables. After removing outliers and selecting variables, two separate analyses were performed, one on the original dataset and another on the modified one (with 16 variables). Six different combinations of cluster methods (Within-groups linkage; Furthest Neighbor; Nearest Neighbor) and interval measures (Euclidean distance interval; Squared Euclidean distance interval) were tried in hierarchical cluster analysis. After comparing the results, the combination of the Furthest Neighbor cluster method and Squared Euclidean distance measure was chosen. After the removal of the three variables, the results were unchanged, and hence a more straightforward model was built with selected variables. According to hierarchical CA, the cases were clustered into six groups (*Appendix: Figure C3*). Table 11 shows the result of k-means.

Cluster Index	1	2	3	4	5	6
Number of Cases	152	73	192	36	65	166

Table 11 Number of Cases in Clusters (Prepared Dataset-Selected Variables)

Cluster	1	2	3	4	5	6
1		4.713	2.990	5.885	4.530	2.614
2	4.713		6.673	3.854	4.315	3.875
3	2.990	6.673		7.181	5.750	3.435
4	5.885	3.854	7.181		4.643	4.580
5	4.530	4.315	5.750	4.643		3.440
6	2.614	3.875	3.435	4.580	3.440	

Table 12 Distances between Final Cluster Centers (Prepared Dataset-Selected Variables)

From Table 12, the distance between cluster 4 and cluster 3 is the largest, and these two clusters have the most dissimilarities from other clusters. ANOVA revealed that all 19 variables contribute significantly to the cluster analysis, and the variable APM contributes the most.

Appendix figure C2 shows that there are a few outliers in the clusters. However, the result of our cluster analysis is valid and effective for dividing players into different groups. Figure 6 shows the z-scores and average values of all variables in the final clusters.

Some patterns are revealed in these clusters (*Figure C7 in appendix*). Cluster 2 and 4 have players from the higher expert levels, while cluster 3 has players from the lower expert levels. Cluster 1,5,6 share a similar trend with the average level. Based on this, players can be grouped into **experts, beginners and common gamers**. Figure 5 shows how the six clusters perform in each of the distinguishing features. Each cluster is colour coded (see below) and measured using z-scores. Distinct colour sets are used to represent the three big groups: purple: **experts (cluster 2&4)**, green: **common gamers (cluster 1,5,6)**, orange: **beginners (cluster 3)** and light blue: average level of all the players.

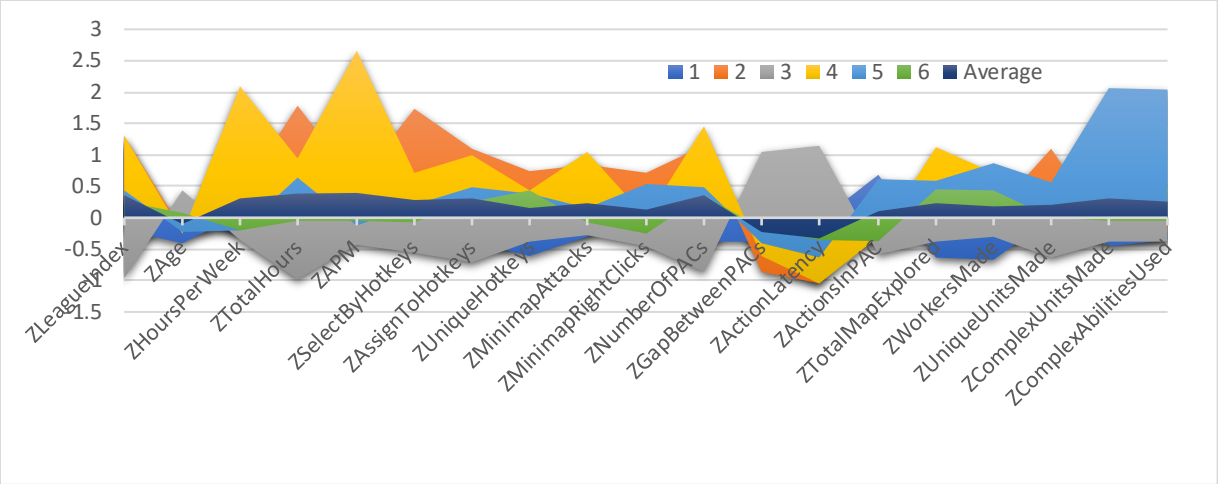

Figure 6 Area Chart of Average Z-score of Different Groups of Players on All Variables

Figure 5 Cluster membership interpreted using variables in factors from FA

3.2 Cluster Interpretation

The Beginners group cluster performs poorly for all variables as they are much below average. These players are relatively older and spend less time playing the game, leading to a lack of agility. These players have trouble progressing into higher league indexes because they are weak in developing resources, offence and cannot handle complex moves well.

Cluster 2 and 4 represent the expert group that have a highly developed style of play. Players in cluster 2 are loyal fans of the game and have devoted many hours to develop their strategies. They utilize hotkeys to improve their in-game speed and do not make complex units until the game's latter stages. However, players in cluster 4 prefer a much slower pace. Like players in cluster 2, they tend to be more agile, but they utilize complex units more often. Players in cluster 4 also have developed heavily offensive strategies. Additionally, they have a natural talent for this game as they do not play that much but still have high league indexes.

The Common Gamers group consist of Clusters 1, 5 and 6. Players in cluster 1 attempt to imitate the professional strategies of cluster 2 players. They create and sacrifice complex units many times. However, their poor agility cannot support this sacrificial strategy. What is worse, they fail to find a balance between defence and offence. If they practice more, they could also be experts. As for members in cluster 5, their entire focus is on making as many complex units as possible and ignoring almost all the other winning factors. It could be said that they would want to imitate the strategies of senior players in cluster 4 to win. However, they lack talent even though they have spent many hours playing the game. Cluster 6 gamers are probably new to this game. They perform poorly on the majority of indexes and need to explore this complex system.

3.3 Validation

A validation set was used to validate the methods and parameters of our cluster model. As mentioned above, the validation set contains 50% of the original dataset. Furthest neighbour cluster method and Squared Euclidean distance measure was applied to do hierarchical CA on the validation set, and the result showed that the cases should be clustered into six groups. Hence the cluster method and distance measure are appropriate.

Post k-means analysis was carried out, and the data points were allocated to their original clusters. GameID was used to check which data points belonged to the same cluster, and the accuracy measure was calculated to find the accuracy of cluster membership. The accuracy was estimated to be 78.23%, which indicates that the result is fairly stable (*Appendix: Figure C5*).

References

[1] Chen, P.Y., Qi, Z., Pan, Y. and Cheng, S.M., 2015, September. Multivariate and Categorical Analysis of Gaming Statistics. In 2015 18th International Conference on Network-Based Information Systems (pp. 286-293). IEEE.

[2] Glass BD, Maddox WT, Love BC. Real-time strategy game training: emergence of a cognitive flexibility trait. *PLoS One*. 2013;8(8):e70350. Published 2013 Aug 7. doi:10.1371/journal.pone.0070350

[3] McColeman, C., Thompson, J., Anvari, N., Azmand, S.J., Barnes, J., Barrett, R.C., Byliris, R., Chen, Y., Dolguikh, K., Fischler, K. and Harrison, S., 2020. Digit eyes: Learning-related changes in information access in a computer game parallel those of oculomotor attention in laboratory studies. Attention, Perception, & Psychophysics, 82, pp.2434-2447.

[4] Thompson, J., Blair, M., Henry, A. and Chen, B., 2013. StarCraft 2 Replay Analysis.

[5] Yong, H.R., Kim, D.J. and Hwang, H.S., 2015. A study of analysing real-time strategy game data using data mining. Journal of Korea Game Society, 15(4), pp.59-68.

Appendix

Record of Meetings

We had 6 meetings in total. We divided the 3 questions among us equally by forming 3 groups of 2 people each and solving one question each. All the group members discussed their findings with their partners and the larger group and decided on the final analysis and report. Each group worked on writing their part of the report and in the end all the questions were combined into one coherent report. Every member was sincere and did their best to answer the questions. There was a lot of interaction and cooperation amongst the group which made the whole process smooth.

Appendix A Regression, MDA, LR

Data Preparation:

Sample Size: The preferred ratio for regression is 20:1 (observations: variables). The SkillCraft data set has a ratio of 38:1. The proportions of each LeagueIndex are seen in *Table A1*. Each LeagueIndex is not represented equally, but there seem to be enough observations in each index to carry out modelling.

Table A1: LeagueIndex Proportions

League 1	League 2	League 3	League 4	League 5	League 6	League 7
14.3%	16.5%	17.1%	17.2%	15.9%	14%	5%

Missing Values: There was a missing value in TotalHours, in LeagueIndex 5. The error was remedied by replacing the missing value with the mean of TotalHours in LeagueIndex 5.

Outliers: Outliers were identified using Mahalanobis Distance. As seen from the graph below, 13 points had a Mahalanobis Distance larger than four. These points were removed from the data set.

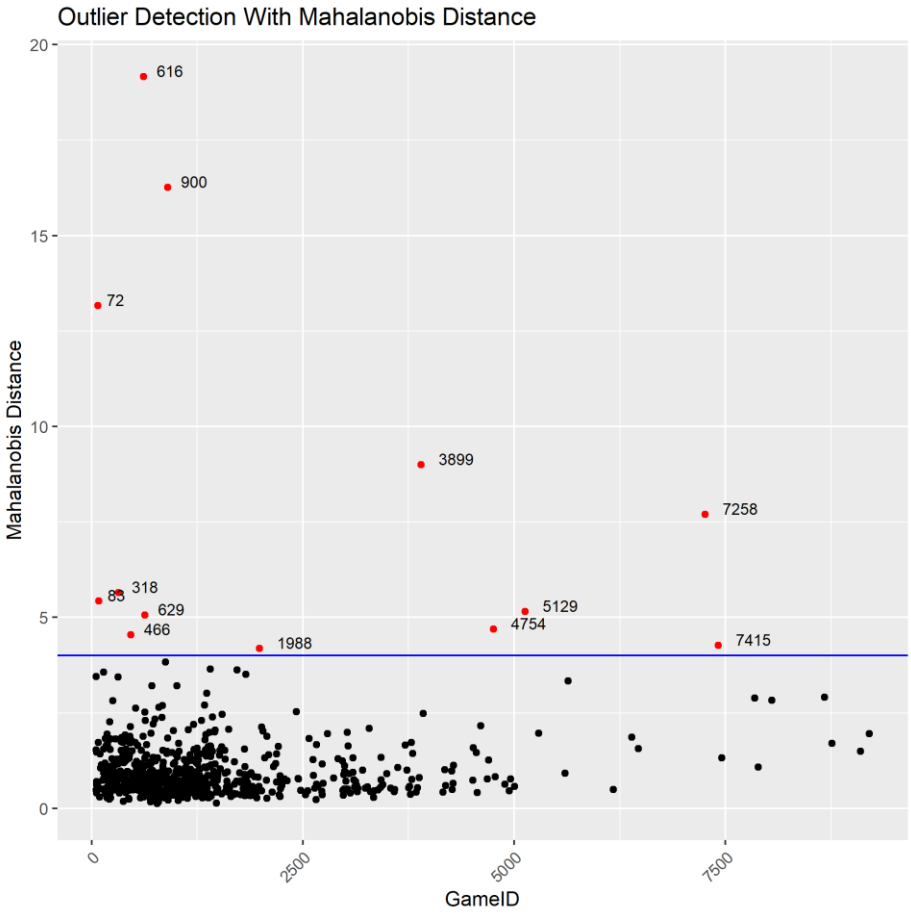

Figure A2: Outlier Detection With Mahalanobis Distance

Modelling Assumptions

Multicollinearity: Figure A3 shows the Pearson Correlation Coefficient between all variables in the data set. LeagueIndex has a high correlation with APM, SelectByHotKeys, AssignToHotKeys, NumberOfPACs, ActionLAtency, and WorkersMade, which is good.

There are two cases of multicollinearity: APM and SelectByHotKeys, and ActionLatency and NumberOfPACs. It makes sense why these variables would be multicollinear. Using hotkeys can significantly improve player speed. So, it only makes sense that APM would increase as SelectByHotKeys does. ActionLatency and NumberOfPACs also make sense as multicollinear variables. As ActionLatency decreases, the NumberOfPACs should increase. The faster a player can make an action, the more actions that player will make per timestamp. These variables will be removed if the modelling results are poor.

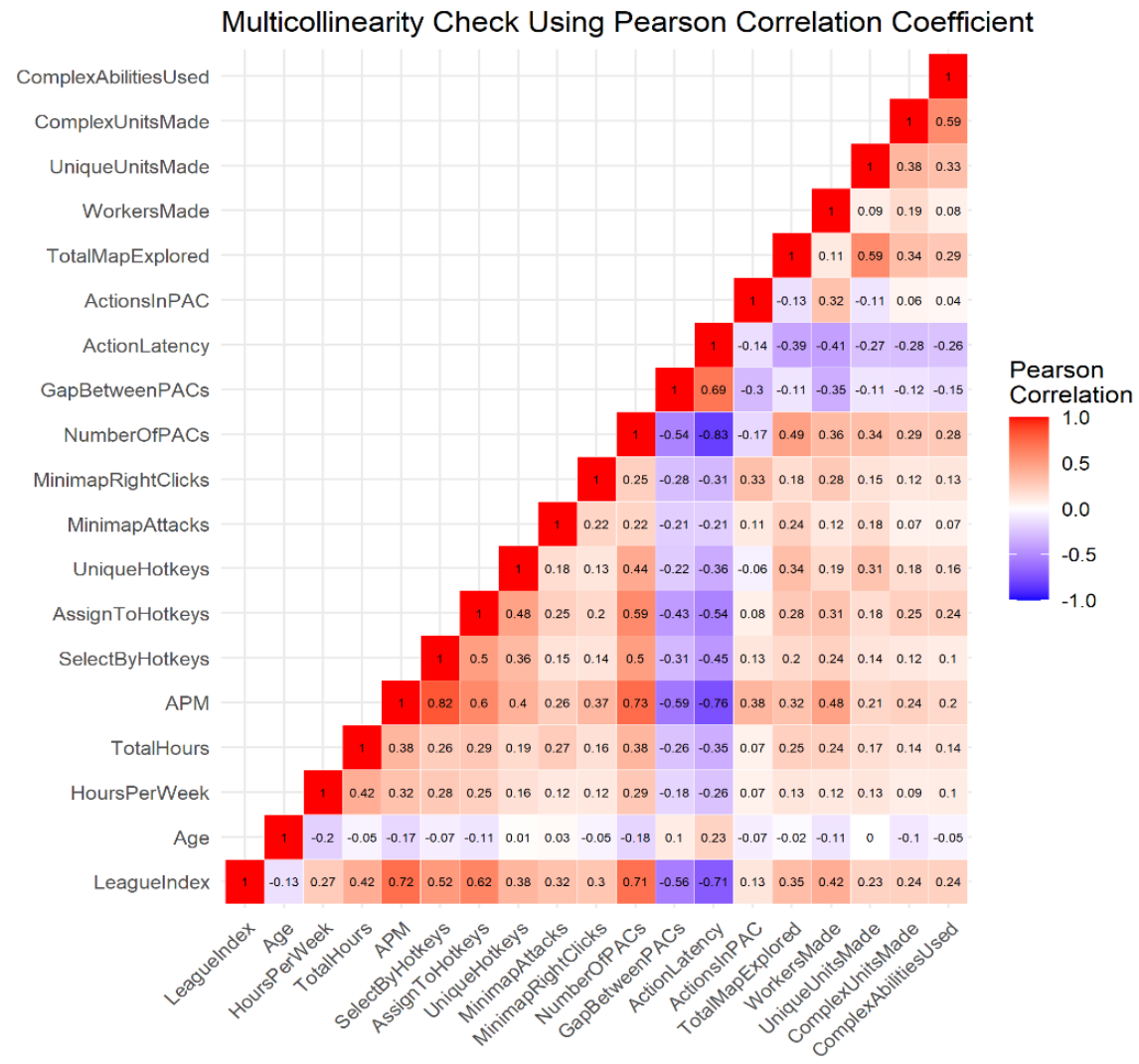

Figure A3: Multicollinearity Check Using Pearson Correlation Coefficient

<u>Linearity and Homoscedaticity:</u> The scatter plots below show only three variables from the data set, Age, APM, and ComplexAbilitiesUsed, respectively. Not every variable is needed to understand that there is an issue with homoscedasticity and linearity. Many other variables share these issues. If problems in the modelling results were present, log transformations would be considered.

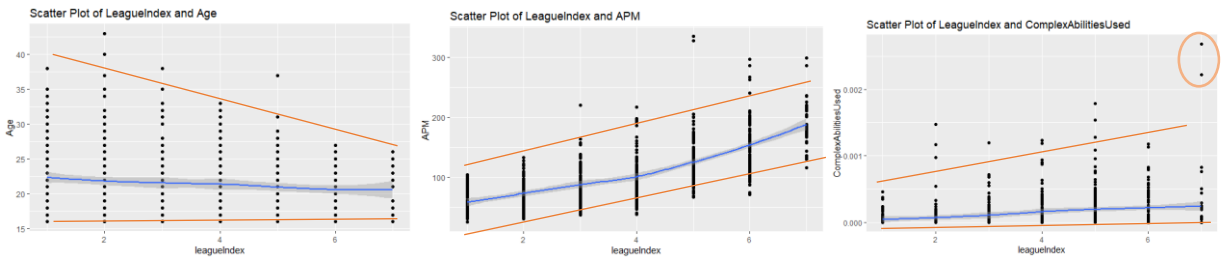

Figure A4: Linearity and Homoscedaticity Check Using Scatter Plots

Ill-Conditioned**:** An ill-conditioned test was performed by creating simple linear regressions with an independent variable as the dependent and using all remaining independent variables as predictors. If the independent variables could explain all (adjusted r-squared = 1) of the fake dependent's variance, then the variable was ill-conditioned. APM, NumberOfPACs, and SelectByHotKeys had more than 90% of their variance explained by other variables. APM had 97% of its variance explained by other variables.

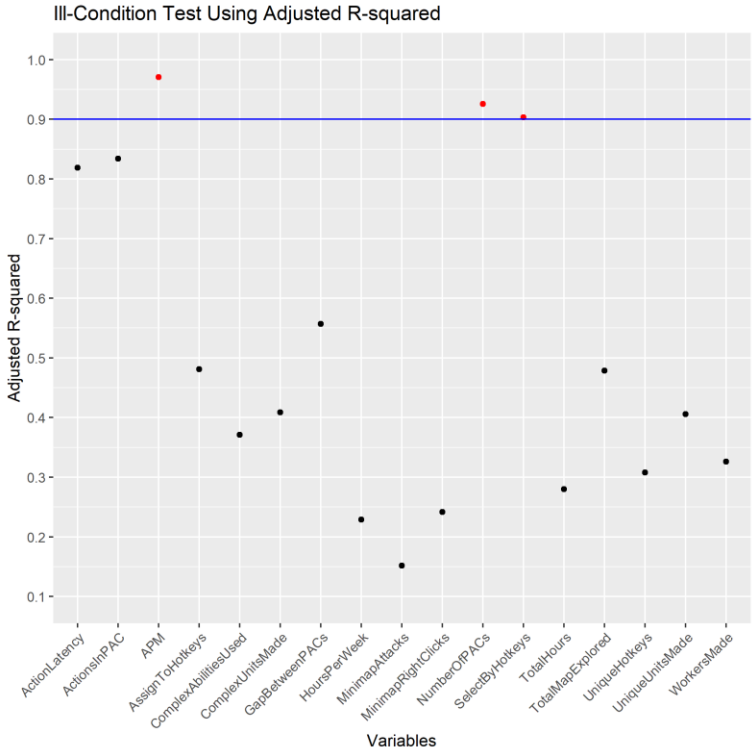

Figure A6: Ill-Condition Test Using Adjusted R-squared

Validation

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.796[a]	.633	.628	1.075	.633	129.307	9	674	.000	
2	.795[b]	.632	.628	1.076	-.001	1.655	1	674	.199	2.005

a. Predictors: (Constant), TotalMapExplored, GapBetweenPACs, TotalHours, UniqueHotkeys, WorkersMade, SelectByHotkeys, NumberOfPACs, ActionLatency, APM

b. Predictors: (Constant), TotalMapExplored, GapBetweenPACs, TotalHours, WorkersMade, SelectByHotkeys, NumberOfPACs, ActionLatency, APM

c. Dependent Variable: LeagueIndex

Table AA: Model Summary of Backwards Stepwise with Variables with R-squared of 10% and above

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2.203	.565		3.899	.000	1.094	3.312					
	APM	.005	.003	.142	1.959	.051	.000	.011	.726	.075	.046	.104	9.607
	NumberOfPACs	339.279	85.995	.191	3.945	.000	170.430	508.129	.701	.150	.092	.231	4.327
	ActionLatency	-.011	.004	-.139	-2.577	.010	-.020	-.003	-.701	-.099	-.060	.187	5.351
	UniqueHotkeys	.025	.020	.034	1.286	.199	-.013	.064	.359	.049	.030	.763	1.311
	GapBetweenPACs	-.013	.003	-.149	-4.301	.000	-.019	-.007	-.565	-.163	-.100	.452	2.212
	SelectByHotkeys	56.113	22.138	.122	2.535	.011	12.645	99.581	.549	.097	.059	.235	4.262
	TotalHours	.001	.000	.165	6.307	.000	.000	.001	.477	.236	.147	.798	1.254
	WorkersMade	317.360	99.294	.092	3.196	.001	122.397	512.322	.426	.122	.075	.659	1.517
	TotalMapExplored	.012	.007	.052	1.840	.066	-.001	.025	.334	.071	.043	.691	1.447
2	(Constant)	2.197	.565		3.887	.000	1.087	3.307					
	APM	.005	.003	.138	1.910	.057	.000	.011	.726	.073	.045	.104	9.593
	NumberOfPACs	357.737	84.830	.202	4.217	.000	191.174	524.300	.701	.160	.098	.238	4.207
	ActionLatency	-.011	.004	-.137	-2.540	.011	-.019	-.002	-.701	-.097	-.059	.187	5.347
	GapBetweenPACs	-.013	.003	-.150	-4.312	.000	-.019	-.007	-.565	-.164	-.101	.452	2.212
	SelectByHotkeys	60.275	21.911	.131	2.751	.006	17.253	103.297	.549	.105	.064	.240	4.171
	TotalHours	.001	.000	.165	6.295	.000	.000	.001	.477	.235	.147	.798	1.254
	WorkersMade	324.917	99.168	.094	3.276	.001	130.203	519.632	.426	.125	.076	.661	1.512
	TotalMapExplored	.014	.007	.058	2.086	.037	.001	.027	.334	.080	.049	.711	1.406

Table AB: Model Coefficients of Backwards Stepwise with Variables with R-squared of 10% and above

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.817[a]	.668	.662	1.023	.668	111.585	6	333	.000	1.264

a. Predictors: (Constant), TotalHours, GapBetweenPACs, AssignToHotkeys, NumberOfPACs, APM, ActionLatency

b. Dependent Variable: LeagueIndex

Table A7: Model V2 Model Summary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.755	.705		2.488	.013	.367	3.142					
	APM	.011	.002	.300	5.390	.000	.007	.015	.750	.283	.170	.322	3.104
	NumberOfPACs	320.919	111.013	.186	2.891	.004	102.543	539.296	.719	.156	.091	.241	4.153
	ActionLatency	-.010	.006	-.123	-1.641	.102	-.022	.002	-.730	-.090	-.052	.177	5.655
	AssignToHotkeys	1444.747	343.078	.174	4.211	.000	769.874	2119.620	.618	.225	.133	.582	1.719
	GapBetweenPACs	-.008	.004	-.087	-1.883	.061	-.016	.000	-.572	-.103	-.059	.472	2.119
	TotalHours	.000	.000	.131	3.668	.000	.000	.001	.474	.197	.116	.787	1.270

Table A8: Model V2 Coefficients

Variable	**Standardised Beta Coefficients**
APM	.3
NumberOfPACs	.186
AssignToHotKeys	.174
TotalHours	.131
ActionLatency	-.123
GapBetweenPACs	-.087

Table A9: Model V2 Standardized Beta Coefficients

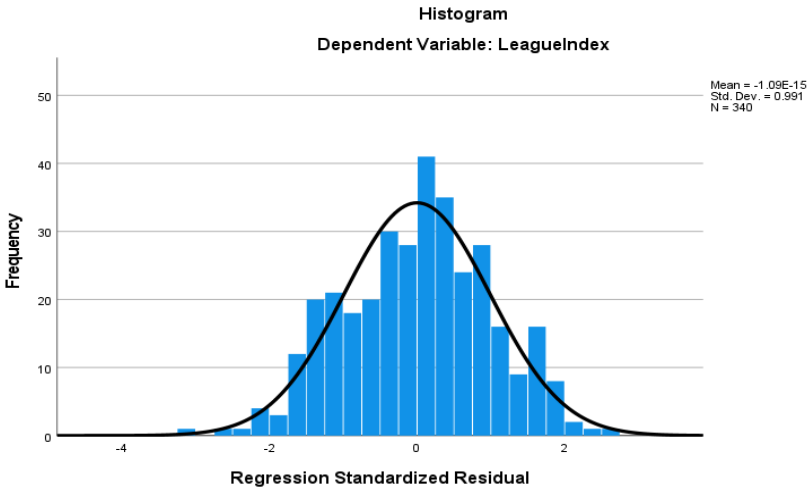

Figure A10: Model V2 Residual Histogram

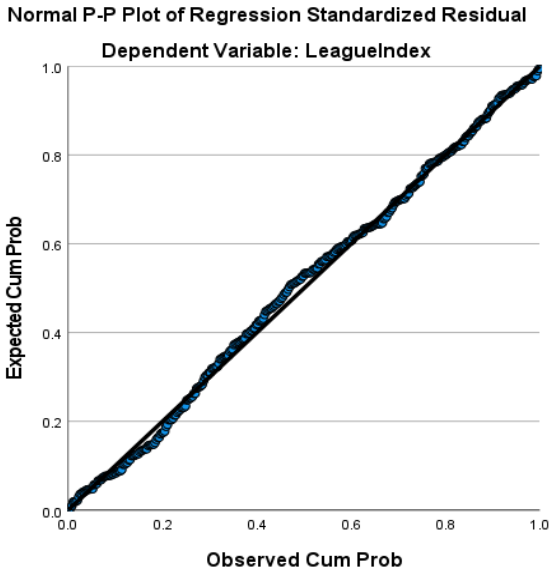

Figure A11: Model V2 Residual Normal P-P Plot

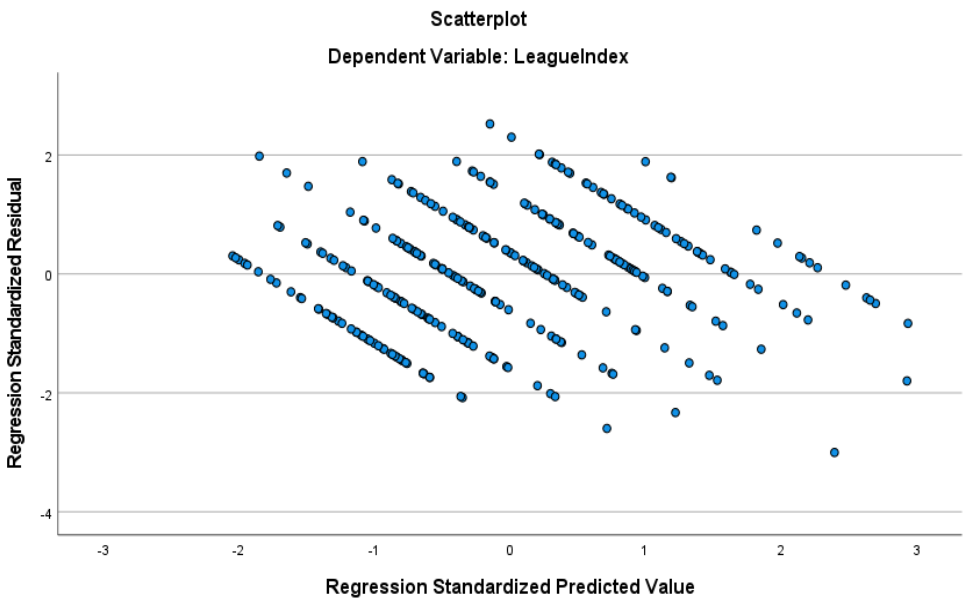

Figure A12: Model V2 Residual Scatterplot

Classification Results[a,b]

				Predicted Group Membership		
			LeagueIndex	1	4	Total
Cases Selected	Original	Count	1	145	19	164
			4	31	145	176
		%	1	88.4	11.6	100.0
			4	17.6	82.4	100.0
Cases Not Selected	Original	Count	1	141	25	166
			4	42	136	178
		%	1	84.9	15.1	100.0
			4	23.6	76.4	100.0

a. 85.3% of selected original grouped cases correctly classified.

b. 80.5% of unselected original grouped cases correctly classified.

Table A10: Classification results of MDA

Structure Matrix

	Function 1
ActionLatency	-.803
NumberOfPACs[a]	.712
AssignToHotkeys	.635
SelectByHotkeys	.526
GapBetweenPACs[a]	-.503
TotalHours	.470
WorkersMade	.430
UniqueHotkeys[a]	.351
MinimapAttacks	.310
TotalMapExplored[a]	.274
ComplexUnitsMade[a]	.255
ComplexAbilitiesUsed	.253
MinimapRightClicks[a]	.205
UniqueUnitsMade[a]	.178
HoursPerWeek	.163

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Table A10: Structure matrix results of MDA

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			LeagueIndex		Percentage Correct	LeagueIndex		Percentage Correct
	Observed		0	1		0	1	
Step 7	LeagueIndex	0	142	22	86.6	138	28	83.1
		1	25	151	85.8	35	143	80.3
	Overall Percentage				86.2			81.7

Table A11: Classification results of Logistic regression

Appendix for MDA:

The results for other combinations of groups:

Classification Results[a]

			Predicted Group Membership							
		LeagueIndex	1	2	3	4	5	6	7	Total
Original	Count	1	80	8	6	4	1	0	0	99
		2	57	15	27	8	6	0	0	113
		3	37	11	33	21	12	4	0	118
		4	16	10	26	37	23	7	1	120
		5	3	4	15	25	42	18	1	108
		6	0	0	5	12	21	53	6	97
		7	0	0	0	0	2	10	17	29
	%	1	80.8	8.1	6.1	4.0	1.0	.0	.0	100.0
		2	50.4	13.3	23.9	7.1	5.3	.0	.0	100.0
		3	31.4	9.3	28.0	17.8	10.2	3.4	.0	100.0
		4	13.3	8.3	21.7	30.8	19.2	5.8	.8	100.0
		5	2.8	3.7	13.9	23.1	38.9	16.7	.9	100.0
		6	.0	.0	5.2	12.4	21.6	54.6	6.2	100.0
		7	.0	.0	.0	.0	6.9	34.5	58.6	100.0

a. 40.5% of original grouped cases correctly classified.

Table A12: MDA classification result- 7 groups

Classification Results[a]

			Predicted Group Membership				
		LeagueIndex	1	3	5	7	Total
Original	Count	1	161	46	5	0	212
		3	67	139	31	1	238
		5	2	64	133	6	205
		7	0	0	13	16	29
	%	1	75.9	21.7	2.4	.0	100.0
		3	28.2	58.4	13.0	.4	100.0
		5	1.0	31.2	64.9	2.9	100.0
		7	.0	.0	44.8	55.2	100.0

a. 65.6% of original grouped cases correctly classified.

Table A13: MDA classification result- 4 groups with level 1-2 in group 1, level 3-4 in group 3, level 5-6 in group 5, level 7 in group 7

Classification Results[a]

			Predicted Group Membership			
		LeagueIndex	1	3	5	Total
Original	Count	1	156	52	4	212
		3	69	140	29	238
		5	3	62	169	234
	%	1	73.6	24.5	1.9	100.0
		3	29.0	58.8	12.2	100.0
		5	1.3	26.5	72.2	100.0

a. 68.0% of original grouped cases correctly classified.

Table A14: MDA classification result-3 groups with level 1-2 in group 1, level 3-4 in group 3, level 5-6-7 in group 5

Classification Results[a]

		LeagueIndex	Predicted Group Membership 1	5	Total
Original	Count	1	420	30	450
		5	65	169	234
	%	1	93.3	6.7	100.0
		5	27.8	72.2	100.0

a. 86.1% of original grouped cases correctly classified.

Table A15: MDA classification result- 2 groups with 1-2-3-4 in group 1, level 5-6-7 in group 5

Classification Results[a]

		LeagueIndex	Predicted Group Membership 1	6	Total
Original	Count	1	536	22	558
		6	52	74	126
	%	1	96.1	3.9	100.0
		6	41.3	58.7	100.0

a. 89.2% of original grouped cases correctly classified.

Table A16: MDA classification result- 2 groups with 1-2-3-4-5 in group 1, level 6-7 in group 6

Notice in this division method, even though overall accuracy peaks at 89.2%, the performance for group 1 and group 6 is not balanced, hence, it is not considered as an un-biased model with good classification accuracy.

First 3 stages of MDA:

Stage 1: Define Problem

1. Understand differences between prior defined groups: Which features are important for different expert level? (Understand the group differences on different expert level group).
2. Determine which features discriminate between different expert level groups.

Stage 2: Research Design

Dependent Variable:

1. 7 expert levels form naturally 7 categories in the original dataset, while two groups of expert level are finalised as dependent variables, with group 1 containing expert levels 1, 2, 3 and group 2 containing expert levels 4, 5, 6, 7.
2. Exhaustive: Number of categories should include all options for dependent variable
3. Must be mutually exclusive: each object can only belong to one category

Independent Variables

17 variables other than Age and GameID are identified as independent variables.

After checking ill-condition problem (need Graham graph), 3 variables () have been identified as suspected ill-conditioned variables and 2 variables including () has been excluded from independent variables list, with 15 variables being used in the finalised model.

The detailed process and impacts will be analysed in Stage 5.

Sample size check:

Overall sample size has been checked more than enough, with sample size in each group over 20+ observations (minimum at 29 in level 7).

Stage 3: Assumptions check

Normality, multicollinearity, homoscedasticity, homogeneity of dispersion assumptions have been checked in the assumptions check in the appendix.

Appendix B Factor analysis

Other FA attempts

Pattern Matrix[a]

	Component				
	1	2	3	4	5
SelectByHotkeys	.804				
APM	.788				
NumberOfPACs	.785				
AssignToHotkeys	.760				
ActionLatency	-.704				
UniqueHotkeys	.622				
GapBetweenPACs	-.548		-.427		
ComplexUnitsMade		.866			
ComplexAbilitiesUsed		.861			
ActionsInPAC			.873		
MinimapRightClicks			.624		.448
WorkersMade			.463		
HoursPerWeek				.901	
TotalHours				.770	
MinimapAttacks					.668
TotalMapExplored					.618
UniqueUnitsMade		.469			.540

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

a. Rotation converged in 19 iterations.

Figure B1: PCA analysis with Direct Oblimin rotation

Pattern Matrix[a]

	Factor					
	1	2	3	4	5	6
ActionLatency	-.828					
NumberOfPACs	.758					-.319
GapBetweenPACs	-.695					
TotalMapExplored		.795				
UniqueUnitsMade		.648				
MinimapAttacks						
ComplexUnitsMade			.843			
ComplexAbilitiesUsed			.696			
TotalHours				1.051		
HoursPerWeek				.444		
SelectByHotkeys					.923	
APM	.317				.633	
AssignToHotkeys					.428	
UniqueHotkeys					.356	
ActionsInPAC						.992
MinimapRightClicks						.375
WorkersMade						

Extraction Method: Unweighted Least Squares.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 12 iterations.

Figure B2: Unweighted least square analysis with Direct Oblimin rotation

Pattern Matrix[a]

	Factor				
	1	2	3	4	5
SelectByHotkeys	1.082				
APM	.582	.386			
AssignToHotkeys	.367	.312			
HoursPerWeek					
NumberOfPACs		.890	-.308		
ActionLatency		-.780			
GapBetweenPACs		-.629			
WorkersMade		.325	.319		
TotalHours					
ActionsInPAC			.998		
MinimapRightClicks			.376		
ComplexAbilitiesUsed				.846	
ComplexUnitsMade				.681	
TotalMapExplored					.818
UniqueUnitsMade					.669
UniqueHotkeys					
MinimapAttacks					

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 16 iterations.

Figure B3: Maximum analysis with Direct Oblimin rotation

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
APM	.823	.406			
NumberOfPACs	.803				
SelectByHotkeys	.775				
AssignToHotkeys	.745				
ActionLatency	-.744				
GapBetweenPACs	-.577	-.492			
UniqueHotkeys	.577				
ActionsInPAC		.842			
MinimapRightClicks		.680			
WorkersMade	.426	.492			
ComplexUnitsMade			.845		
ComplexAbilitiesUsed			.838		
TotalMapExplored				.734	
UniqueUnitsMade			.444	.658	
MinimapAttacks				.594	
HoursPerWeek					.867
TotalHours					.760

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

a. Rotation converged in 8 iterations.

Figure B4: PCA analysis with varimax rotation

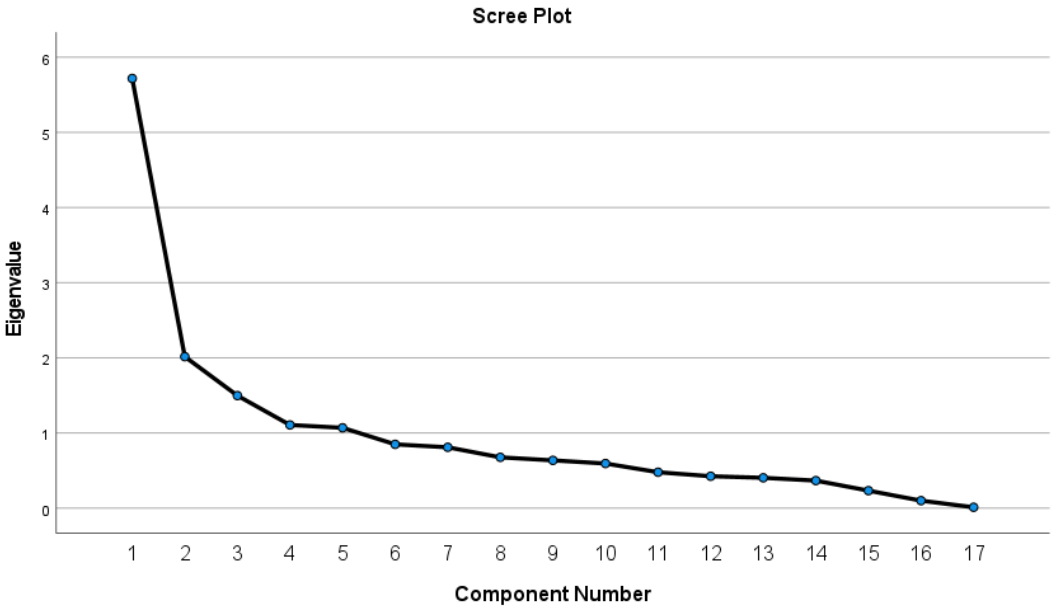

Figure B5: Scree plot from PCA analysis.

Appendix C Cluster analysis

Using all variables do cluster analysis on original data set

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Appears		Next Stage	Differences between coefficients	Cluster numbers
	Cluster 1	Cluster 2		Cluster 1	Cluster 2			
668	335	374	61.939	656	654	680	0.61619735210	Cluster 28
669	68	91	64.272	665	663	679	2.33280915498	Cluster 27
670	107	341	64.797	614	649	685	0.52516078193	Cluster 26
671	516	680	65.616	597	0	688	0.81904547810	Cluster 25
672	300	306	65.692	664	657	677	0.07587198366	Cluster 24
673	2	20	69.311	655	635	676	3.61918605293	Cluster 23
674	529	580	70.448	0	636	682	1.13712657059	Cluster 22
675	576	684	70.544	640	634	683	0.09593467603	Cluster 21
676	2	3	71.715	673	659	679	1.17084802654	Cluster 20
677	300	585	84.659	672	662	680	12.94430679514	Cluster 19
678	1	8	85.008	666	651	684	0.34935395585	Cluster 18
679	2	68	88.296	676	669	684	3.28756886738	Cluster 17
680	300	335	90.018	677	668	683	1.72170216540	Cluster 16
681	480	508	96.990	661	627	688	6.97197006609	Cluster 15
682	117	529	99.460	667	674	687	2.47038384047	Cluster 14
683	300	576	99.896	680	675	686	0.43607951506	Cluster 13
684	1	2	117.503	678	679	692	17.60669711742	Cluster 12
685	107	681	119.128	670	527	691	1.62537787047	Cluster 11
686	217	300	121.978	572	683	687	2.84972460957	Cluster 10
687	117	217	123.560	682	686	690	1.58257212321	Cluster 9
688	480	516	143.737	681	671	690	20.17652811197	Cluster 8
689	83	129	146.530	0	0	692	2.79333241812	Cluster 7
690	117	480	164.645	687	688	691	18.11423014538	Cluster 6
691	107	117	211.960	685	690	694	47.31587860273	Cluster 5
692	1	83	212.439	684	689	694	0.47819650631	Cluster 4
693	505	663	319.536	0	0	695	107.09736833345	Cluster 3
694	1	107	360.371	692	691	695	40.83501766252	Cluster 2
695	1	505	492.870	694	693	0	132.49911444035	Cluster 1

Figure C1: Hierarchical Cluster Analysis (Original Data Set-All Variables)

Cluster cetrals: Using all variables do kmeans analysis on original data set

CLUSTER_	ZLeagueIn	ZAge	ZHoursPer	ZTotalHou	ZAPM	ZSelectByI	ZAssignTo	ZUniqueH	ZMinimap	ZMinimap	ZNumberC	ZGapBetw	ZActionLa	ZActionsIn	ZTotalMa	ZWorkersI	ZUniqueU	ZComplex	ZComplexAbilitiesUsed
1	0.326514	0.037042	-0.19176	-0.07174	0.04956	-0.15533	0.211175	0.237126	-0.04216	0.026409	0.331348	-0.20196	-0.41517	-0.04585	0.656779	0.103674	0.834565	0.81556	0.606458
2	1.024274	-0.08892	0.125842	0.244314	1.171033	0.984796	0.646246	0.507311	0.18552	0.418303	0.709176	-0.6329	-0.81002	0.470546	-0.19565	1.067658	-0.27536	-0.31978	-0.35156
3	-0.33289	-0.35096	0.011789	-0.18653	-0.23137	-0.35291	-0.34363	-0.48788	-0.23204	-0.02559	-0.37889	-0.26058	0.009713	0.400835	-0.56221	-0.16157	-0.61974	-0.47023	-0.30918
4	1.371798	-0.34078	1.208213	0.990355	1.629616	1.483022	1.491372	1.114878	0.772679	0.681944	1.671891	-0.83406	-1.19421	-0.00288	1.180524	0.450859	0.879204	1.072023	1.000628
5	-0.95026	0.500377	-0.35295	-0.39492	-0.98155	-0.52851	-0.68505	-0.34827	-0.21989	-0.46631	-0.84725	1.049276	1.17533	-0.61077	-0.31224	-0.6477	-0.24754	-0.40547	-0.36006
6	0.74976	-0.54297	0.877183	8.399125	0.843175	0.094693	0.02706	-0.00478	1.95923	1.052578	0.745855	-0.61172	-0.92235	0.456145	0.948865	1.173075	0.28662	-0.12659	0.221635
7	0.74976	-0.31594	0.6346	-0.0012	0.45095	-0.15978	0.94118	0.82845	16.81791	1.03196	-0.0756	-0.634	0.08844	1.1172	0.81649	0.45356	2.42326	0.35429	0.70703

Figure C2: Calculated Cluster Centrals (Original Data Set-All Variables)

Using selected variables do cluster analysis on prepared data set

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Appears		Next Stage	Differences between coefficients	Cluster numbers
	Cluster 1	Cluster 2		Cluster 1	Cluster 2			
655	118	239	56.890	648	626	663	0.0413652043537	29
656	463	505	57.498	635	619	659	0.6079585422997	28
657	297	579	58.003	639	636	675	0.5053371216924	27
658	570	673	58.839	631	0	681	0.8352702196260	26
659	463	510	59.484	656	596	674	0.6457405188716	25
660	105	372	61.459	646	597	668	1.9744243542534	24
661	3	22	63.057	644	620	672	1.5984937729260	23
662	214	523	63.154	652	0	670	0.0969462129911	22
663	118	281	64.605	655	637	668	1.4505829178922	21
664	4	31	65.992	650	624	667	1.3868932172640	20
665	5	106	66.587	625	640	676	0.5956088987125	19
666	560	602	70.672	0	600	674	4.0841920248348	18
667	2	4	77.154	653	664	672	6.4820482602674	17
668	105	118	78.463	660	663	676	1.3088869946302	16
669	289	452	80.518	649	455	678	2.0554522606065	15
670	214	534	81.595	662	628	679	1.0766459711861	14
671	371	552	83.701	654	520	677	2.1064919844221	13
672	2	3	90.342	667	661	673	6.6413577475859	12
673	1	2	94.695	647	672	682	4.3526801215845	11
674	463	560	95.151	659	666	680	0.4559169161210	10
675	297	476	97.226	657	651	678	2.0749057338926	9
676	5	105	105.575	665	668	677	8.3494192726251	8
677	5	371	120.988	676	671	679	15.4121216690716	7
678	289	297	128.221	669	675	680	7.2337451302881	6
679	5	214	141.821	677	670	682	13.6000601047809	5
680	289	463	145.086	678	674	681	3.2646725252936	4
681	289	570	170.668	680	658	683	25.5822273485857	3
682	1	5	196.397	673	679	683	25.7283801625599	2
683	1	289	248.998	682	681	0	52.6015178156466	1

Figure C3: Hierarchical Cluster Analysis (Prepared Data Set-Selected Variables)

Cluster cetrals: Using selected variables do kmeans analysis on prepared data set

	ZLeagueIn	ZAge	ZHoursPer	ZTotalHou	ZAPM	ZSelectByI	ZAssignTo	ZUniqueH	ZMinimap	ZMinimap	ZNumberC	ZGapBetw	ZActionLa	ZActionsIn	ZTotalMap	ZWorkersI	ZUniqueU	ZComplex	ZComplexAbilitiesUsed
1	0.184633	-0.25871	0.080348	0.037465	0.273839	-0.06271	-0.12497	-0.46509	-0.1667	0.604418	-0.22196	-0.50098	-0.25198	0.958484	-0.45178	0.537784	-0.54474	-0.40913	-0.39794
2	1.360123	-0.27704	0.577262	0.643197	1.9941	2.141246	1.372945	0.891732	0.974108	0.700868	1.345541	-0.90091	-1.10714	0.386433	0.409769	0.841309	0.258386	0.039456	-0.03553
3	-0.89908	0.288509	-0.24644	-0.43352	-0.8889	-0.55991	-0.71512	-0.46353	-0.2904	-0.42291	-0.81209	0.815022	0.96725	-0.41277	-0.46497	-0.58384	-0.36244	-0.3909	-0.37402
4	1.241446	-0.22954	2.092995	3.020029	0.728318	0.357559	0.72273	0.442248	0.925286	-0.05468	1.322439	-0.56503	-0.99101	-0.31632	0.900689	0.194531	0.591143	0.889916	0.461387
5	0.469081	-0.23314	-0.20291	-0.12859	0.576237	0.071011	0.498341	0.392903	0.282028	0.614777	0.500792	-0.40119	-0.65161	0.592408	0.668555	0.547064	0.950173	2.065237	2.110952
6	0.251708	-0.00351	-0.19626	-0.10714	-0.0173	-0.06539	0.273763	0.417522	-0.09848	-0.29257	0.38748	-0.22648	-0.37305	-0.37827	0.420231	-0.09139	0.347535	-0.06619	-0.01558

Figure C4: Calculated Cluster Centrals (Prepared Data Set-Selected Variables)

Original Dataset		Validation Dataset						
GameID	Cluster Re	GameID	Cluster-va	Cluster Or	Cluster Original Inde		Accuracy S	Accuracy
61	3	61	1	3	3	val1-ori3	1	0.782353
171	1	532	1	3	3	val6-ori1	1	
363	3	699	1	3	3	val3-ori6	1	
532	3	731	1	3	3	val5-ori5	1	
590	3	779	1	3	3	val2-ori4	1	
600	1	1013	1	3	3	val4-ori2	1	
699	3	1016	1	3	3		1	
731	3	1141	1	3	3		1	
766	3	1228	6	3	1		0	
771	3	1338	1	3	3		1	
779	3	1447	6	1	1		1	
1013	3	1469	1	3	3		1	
1016	3	1679	1	3	3		1	
1074	3	1830	6	1	1		1	
1128	3	2178	1	3	3		1	
1141	3	2231	1	3	3		1	
1173	3	2533	1	3	3		1	
1228	3	2677	1	3	3		1	
1247	3	2988	1	3	3		1	
1298	1	3044	1	3	3		1	
1338	3	3135	1	3	3		1	
1347	3	3146	1	3	3		1	
1366	1	3265	1	3	3		1	
1447	1	3287	1	3	3		1	
1469	3	3339	1	3	3		1	
1580	1	3344	1	3	3		1	
1603	3	3389	6	3	1		0	
1604	3	3423	1	3	3		1	
1634	3	3428	1	3	3		1	
1648	3	3449	1	3	3		1	
1679	3	3494	1	3	3		1	
1719	3	3621	1	3	3		1	
1825	3	3734	1	3	3		1	
1826	3	3773	1	3	3		1	
1830	1	3782	1	3	3		1	
1852	3	3801	1	3	3		1	

Figure C5: Accuracy Score

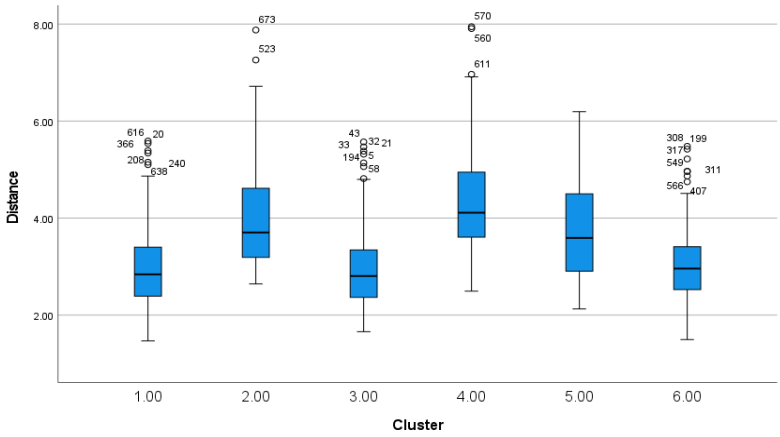

Figure C6: Boxplot of Cluster Memberships' Distances to Cluster Centers

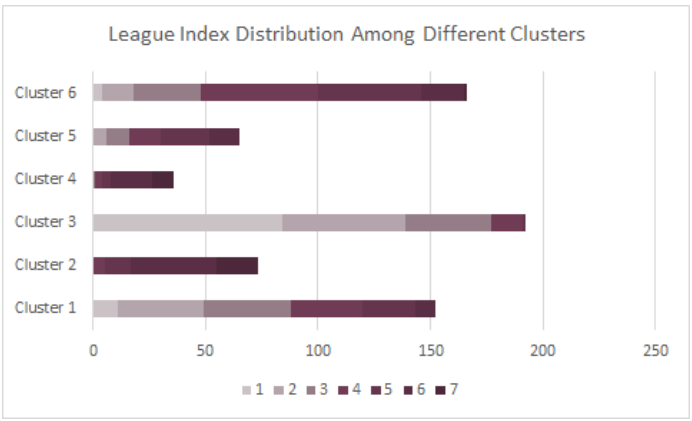

Figure C7: League Index distribution in different clusters