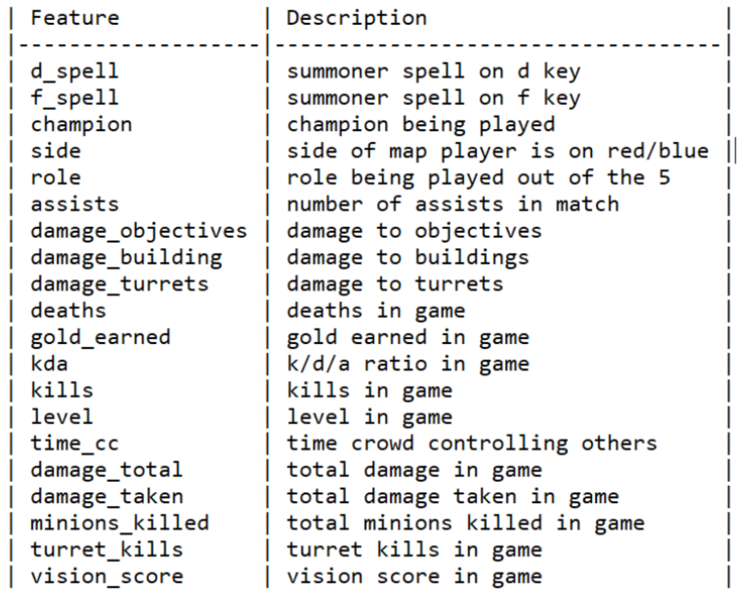
Table of Contents

Aim

The aim of this research project is to determine whether a relationship exists between a player’s performance and which team they are in. This was done by dividing players into a team and role category and comparing average performance. The results of this experiment may help game developers to redesign the game to allow for a more balance between teams. The results may also be useful by providing insight into whether players who play as ‘Jungler’ perform better on average than those who play using other roles. The project is therefore designed to help game makers determine whether team side or role gives a player an advantage and thus enable game makers to change the design of the game and/or redesign roles so that players are on an even playing field. This could allow more competitive games and the intrinsic skill level of a player to be reflected through their game statistics.

Datasets

Three datasets were utilized in this experiment, each drawn from the League of Legends Challenger matches from 3 individual servers, namely North America (NA), Europe West (EUW), Korea (KR). Each of the datasets were initially in excel format. All challenger players' match history was collected from January 2022 and each record comes from a randomly chosen individual in a single game. Each game has only one player drawn from it and an individual has only been collected once from all games for that region to ensure independence.



Feature	Description
d_spell	summoner spell on d key
f_spell	summoner spell on f key
champion	champion being played
side	side of map player is on red/blue
role	role being played out of the 5
assists	number of assists in match
damage_objectives	damage to objectives
damage_building	damage to buildings
damage_turrets	damage to turrets
deaths	deaths in game
gold_earned	gold earned in game
kda	k/d/a ratio in game
kills	kills in game
level	level in game
time_cc	time crowd controlling others
damage_total	total damage in game
damage_taken	total damage taken in game
minions_killed	total minions killed in game
turret_kills	turret kills in game
vision_score	vision score in game

Figure 1. Data Dictionary Game (LOL) Dataset

Pre-processing Method & Wrangling

The data was pre-processed and wrangled in such a way as to limit the use of supervised machine learning algorithms to avoid biased results in the clustering stage of the experiment.

Kills, deaths, assists and kda

Kills, assists, deaths and kda are related by the following equation:

$$kda = \frac{kills + assists}{deaths}$$

Therefore, if any of these attributes were missing they were calculated using this equation, provided the other three variables were available. In the case where kills, assists and kda are all equal to zero and hence deaths cannot be calculated using the formula, deaths were estimated using linear regression of damage taken. If the other three variables were not available the record was removed.

Damage objectives and damage turrets

Damage Objectives and Damage Turrets contain the same data and therefore any missing data from one variable could be extracted from the other variable. Where both attributes were missing from a record the record was removed. Following this Damage Turrets was removed to avoid duplicity.

Levels

As no game duration data was provided, the variable Levels was used as a way of measuring the duration of the game played by the respective player. This was used to normalize the numerical data in the dataset to produce more accurate and less inflated results. Missing values of Levels were not filled in because Levels was being used for standardization and this can affect the accuracy of results. Instead all null values were removed.

Champion

Champion could not be filled as each Champion was only allowed to appear once in each server to ensure independence hence any null value was removed from the data.

Role

Role could not be imputed as major conclusions at the end of the experimental analysis were being made based on this variable, hence any inaccuracy in this variable could lead to false conclusions about the experiment. Any ‘Jungler’ that gets played less than 0.5% of the total ‘Jungler’ pick was removed to prevent skewage. This is because the player may have role swapped and hence the data associated with this record cannot be affiliated with the role ‘Jungler’. The player may not be playing to their full potential and hence the role performance has not been maximized. Finally, the player may not have been suited to the role and hence the role performance may not have been maximized.

Other Numerical Data

All other numerical data was filled in using the mean value of the variable of interest calculated using only records with the same role as the role of the record being filled. This method is aimed to reduce biases in the final result.

Special Cases

Records for Champions “Ivern” and “Quinn” were removed as these champions only appeared once in the dataset, hence the missing numerical data could not be imputed.

Data loss during the cleaning process

It can be seen from the graph to the right that a substantial amount of the data remains after data processing and wrangling has occurred. Therefore, the conclusions that are drawn at the end of the experiment can still be considered valid.

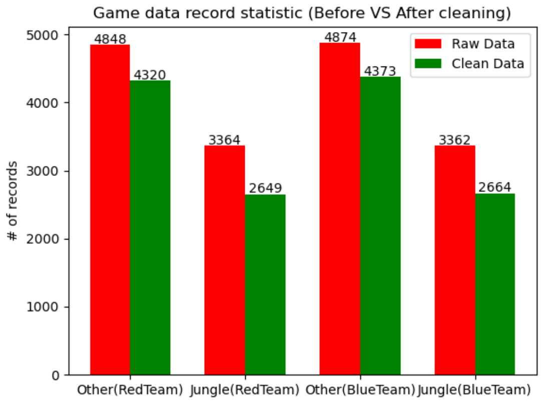


Figure 2. Game Data Statistic (Before VS After Data Preprocessing)

Analysis Method

Feature engineering

Firstly, as no game duration data was provided, the variable ‘Levels’ was used as a way of measuring the duration of the game played by the respective player and feature data normalized by this.

Normalizing data

Feature data was normalized based on ‘Levels' so that players could be compared despite playing games of varying lengths. Furthermore, feature data was normalized a second time to ensure that variables with larger values did not dominate variables with smaller values.

Binning Data

Feature data was binned by ‘Levels’ using equal frequency method so that the strength of all variable relationships, including those with non-linear relationships, could be measured by calculating MI scores.

Feature Selection

A heatmap of the relationship between different features based on their MI scores was developed to identify attributes that may have been dependent so that these could be eliminated from the feature set. This is because in PCA analysis the features used to determine clusters must be independent from each other. ‘Champions’, ‘d_spell’ and ‘f_spell’ are not included in this heatmap as these attributes do not explain variation in player performance. ‘Roles’ and ‘Side’ are not included because they are being analysed. This graph is depicted as below.

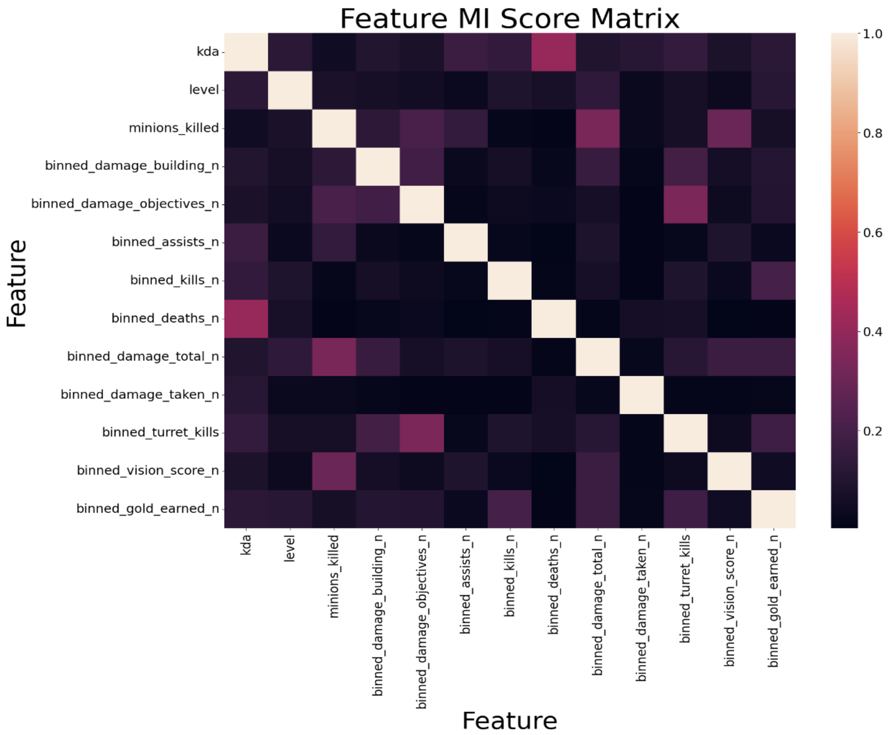


Figure 3. Feature MI Score Matrix

The distribution of the MI scores was represented on a line graph as seen below. From the graph, 4 features were determined to be dependent with each other due to the large gap between the cluster to the left of the red line marking an MI score of 0.25 and the 4 MI scores to the right of this line.

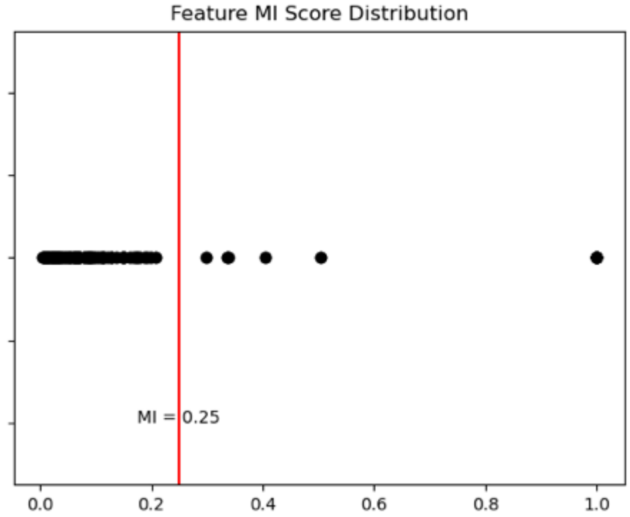


Figure 4: Feature MI Score Distribution

The original heatmap was modified to exclude attributes with MI scores greater than 0.25 by setting their MI score to equal 1. These features were therefore not used to analyze variation in the dataset.

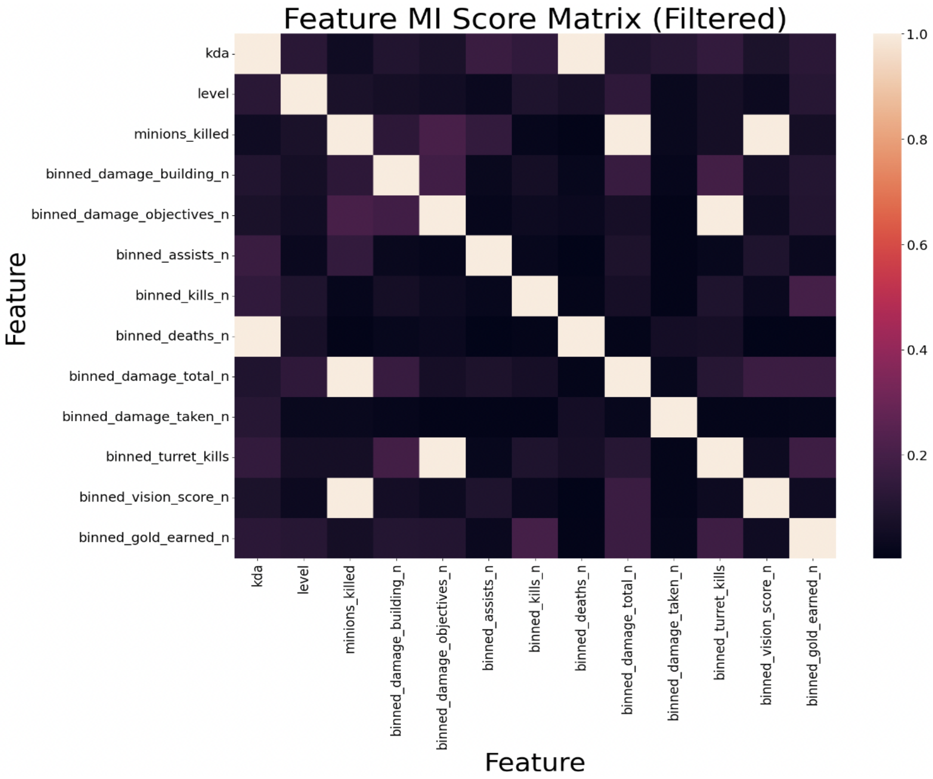


Figure 5: Feature MI Score Matrix (Filtered)

From this heatmap the following variables were determined to be the best variable to exclude from the list of variables that can explain variation in the dataset; “Deaths”, “Minions Killed” and “Turret kills”. “Deaths” was removed as it is dependent with ‘kda’ and intuitively ‘kda’ was predicted to explain more of the variation in the data. “Minions Killed” was removed because it is dependent with 3 different features which together were predicted intuitively to explain more of the variation in the data than “Minions Killed” alone. “Turret kills” was removed as it is dependent with “Damage Objectives” and intuitively this was predicted to explain more of the variation in the data because it had a greater range of values. Thus, after excluding a total of 3 variables from the set of 13 that made up the MI heatmap, 10 independent features remained that could be used to explain variation in the dataset.

This Weighted Feature Coefficient (PCA) Distributions was then graphed on a line graph to represent the distribution of these PCA score and the best features to select. The distribution of this graph showed that 5 features would best explain the variation in the dataset. These are the 5 different features to the right of the red line marking the PCA coefficient value of 0.125 as shown below i.e. the 5 features with PCA values greater than 0.125 were selected to form clusters out of the records in the data.

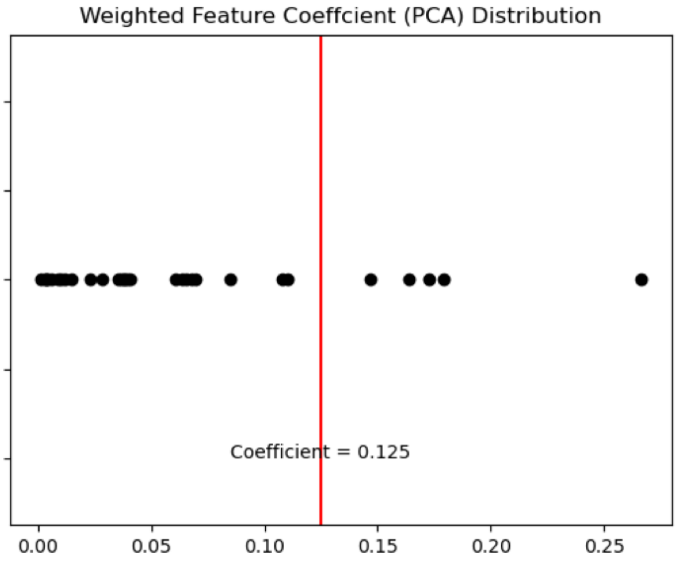
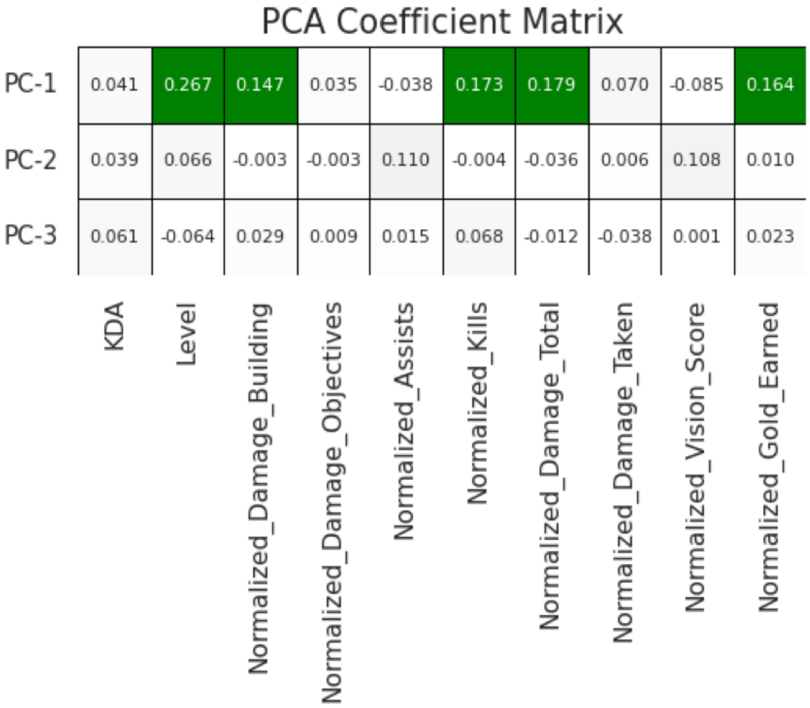


Figure 6: Weighted Feature Coefficient (PCA) Distribution

The identities of these variables and their PCA scores were determined using a PCA coefficient matrix. This matrix is included below and highlights the values of the 5 greatest PCA values labeled with the variables that possessed these values.



PCA Coefficient Matrix

	KDA	Level	Normalized_Damage_Building	Normalized_Damage_Objectives	Normalized_Assists	Normalized_Kills	Normalized_Damage_Total	Normalized_Damage_Taken	Normalized_Vision_Score	Normalized_Gold_Earned
PC-1	0.041	0.267	0.147	0.035	-0.038	0.173	0.179	0.070	-0.085	0.164
PC-2	0.039	0.066	-0.003	-0.003	0.110	-0.004	-0.036	0.006	0.108	0.010
PC-3	0.061	-0.064	0.029	0.009	0.015	0.068	-0.012	-0.038	0.001	0.023

Figure 7: PCA Coefficient Matrix

Clustering

Having obtained the relevant features for clustering (The 5 features with PCA values highlighted above) the number of clusters was selected via the elbow method. The graph of “Distortion vs K” was plotted and can be seen below. This graph proved that the optimal K value was 3 i.e. the dataset is best clustered into 3 groups of records. This is because the difference in distortion when increasing K beyond 3 becomes marginal.

A scatter plot matrix containing an individual scatter plot for every combination of variables in the clustering feature set. This matrix shows the relationship between each pair of variables and categories each data point based on what cluster they belong to by colour. As such, patterns between pairs of variables within the same cluster can be seen and the overall performance of players within each cluster can be predicted intuitively.

Cluster 0, represented by the colour green, consistently has the highest statistics indicating that on average cluster 0 contains the best performing players in the dataset. Cluster 2, represented by the colour orange, consistently has mid-range statistics indicating that on average it contains the 2nd highest performing players in the dataset. Cluster 1, represented by the colour purple, consistently has the lowest statistics indicating that on average it contains the worst players in the dataset.

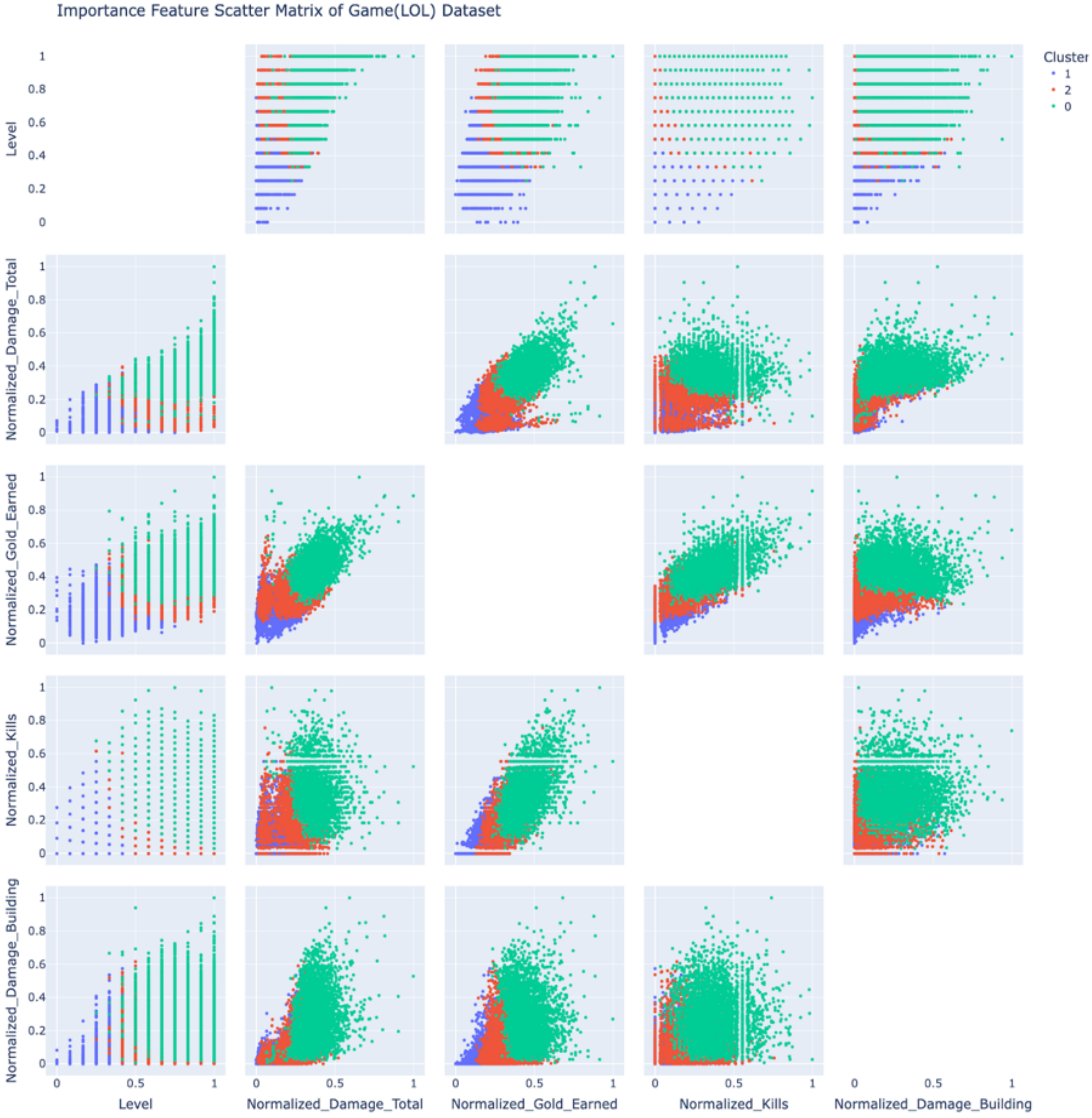


Figure 8: Importance Feature Scatter Matrix of Game (LOL) Dataset

Discussion

All players in the dataset can now be categorised into three categories; ‘Worst Performer’, ‘Moderate Performer’ and ‘Best Performer’. Further, these players can be separated into ‘Red Team’ and ‘Blue Team’. As such, all players can be categorised into 6 categories and the aim to determine whether a relationship exists between a player’s performance and which team they are in can be determined.

The graph below has been developed from the records of all players who played as ‘Jungler’. These players have been separated into the three categories; ‘Worst Performer’, ‘Moderate Performer’ and ‘Best Performer’ each represented by a bar in the graph. The bar in each category represents the percentage margin between players in Red Team and players in Blue Team. The percentage is calculated from the total number of records in the respective category with role Jungler. It can be seen that the number of ‘Worst Performers’ in each team is relatively similar. The number of ‘Moderate Performers’ in Red Team is 2.91% greater than those in Blue Team. The number of ‘Best Performers’ in Blue Team is 2.62% higher than the number in Red Team. Therefore, ‘Junglers’ who play as Blue Team are able to perform at a higher level than ‘Junglers’ who play as Red Team. So, playing in Blue Team may give a player an advantage if they are playing the role of ‘Jungler’. However, this advantage is only small.

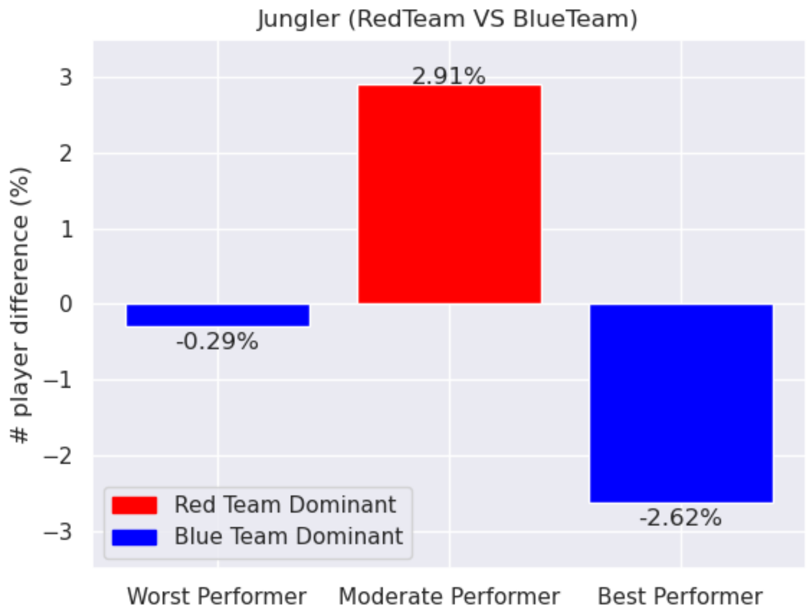


Figure 9. Jungler Performance Difference (RedTeam VS BlueTeam)

The graph below has been developed from the records of all players who play the role ‘Other’ with the same categorization. The bar in each category represents the percentage margin between players in Red Team and players in Blue Team. The percentage is calculated from the total number of records in the respective category with role ‘Other’. It can be seen that the number of ‘Worst Performers’ is 1.24% greater in Blue Team than in Red Team. The number of ‘Moderate Performers’ is relatively similar in both teams. Finally, the number of ‘Best Performers’ in Red Team is 1.46% higher than the number in Blue Team. Therefore, ‘Others’ who play as Red Team are able to perform at a higher level than ‘Others’ who play as Blue Team. So, playing for Red Team may give a player an advantage if they are playing as role ‘Other’.

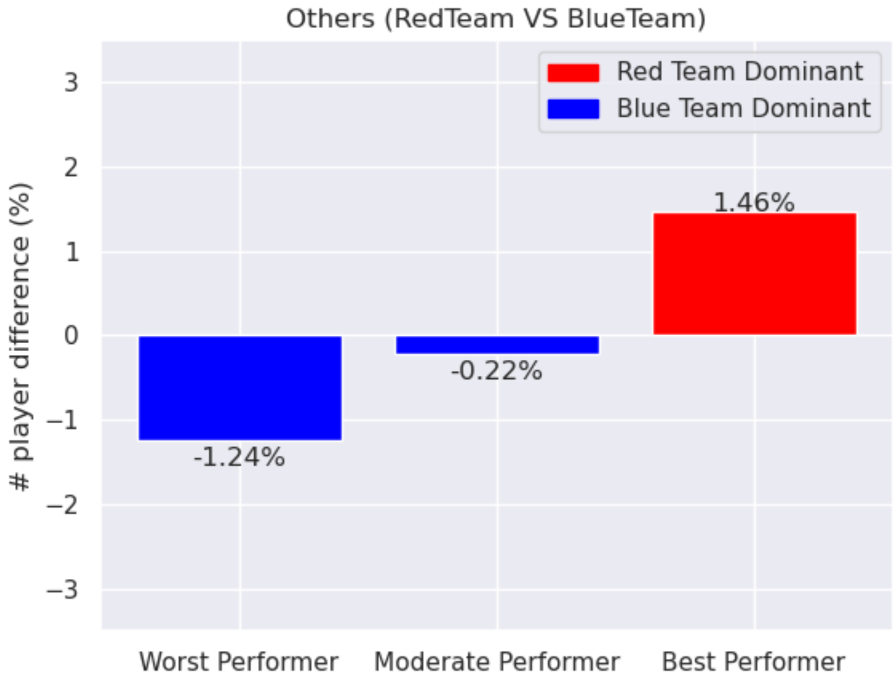


Figure 10. Other Performance Difference (RedTeam VS BlueTeam)

After considering the conclusions drawn from the above two graphs it can be seen that ‘Jungler’ is more suited to Blue Team and ‘Other’ is more suited to Red Team. Therefore, game designers may want to investigate ways of redesigning the game or roles so that neither ‘Jungler’ nor ‘Other’ players have an advantage when playing for either team.

Evaluation

There is a very limited set of ‘Roles’ included in the dataset which could be expanded by separating ‘Other’ roles into their respective roles.

No game time measuring variable was included in the dataset therefore ‘Levels’ had to be used with the assumption that ‘Levels’ is proportional to game time.

‘Damage Objectives’ and ‘Damage Turrets’ contained the same values however according to the website definition of these variables they should contain different data. Therefore, it seems highly unlikely that these variables should contain all the same values. Perhaps data was entered incorrectly into one of these variables’ columns or definitions were confused.

Appendix

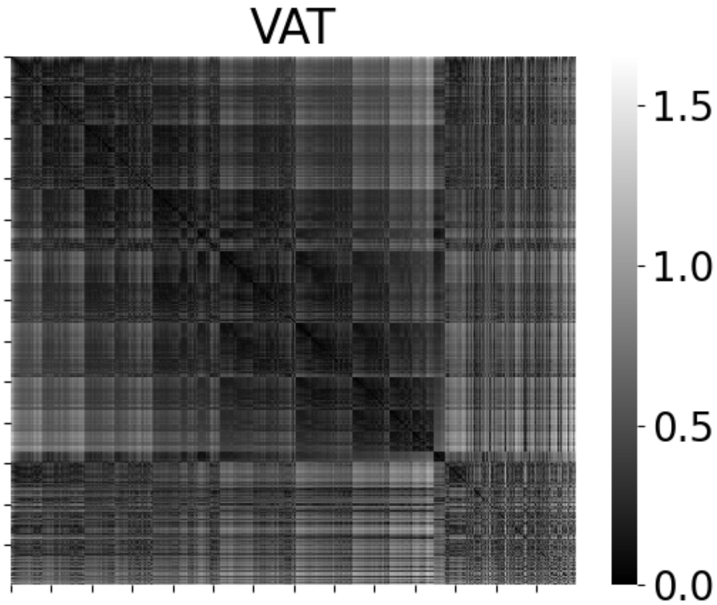


Figure 11. VAT (Importance Feature)

Level vs Norm_Damage_Total vs Norm_Gold_Earned

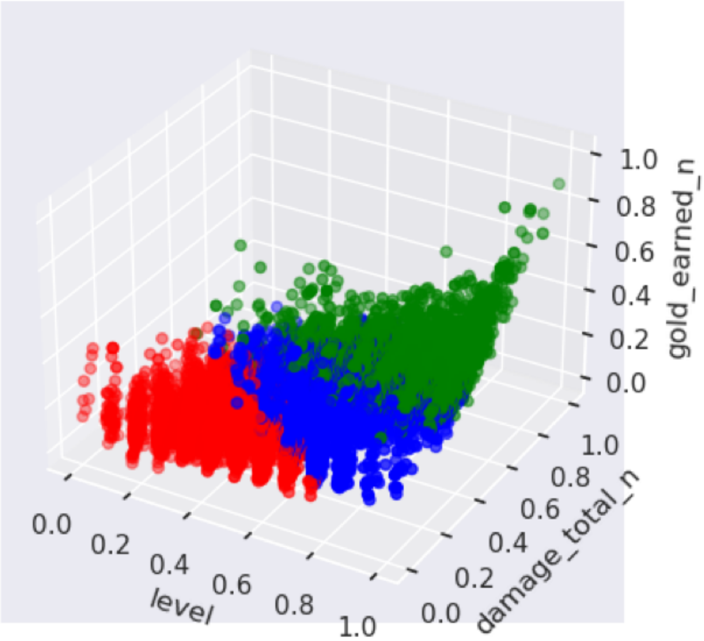


Figure 12. Sample Cluster Plot

Reference

Powell, V., & Lehe, L. (n.d). *Principal Component Analysis explained visually.* https://setosa.io/ev/principal-component-analysis/