Element of Data Processing Ass2 report

Introduction @Liang

In this report we investigate How to get a higher Kda in League of Legends(LOL).By examining the relationship between some of the data categories and Kda, we aim to identify which aspect is important or can make a significant result in obtaining a higher Kda.

Target Audience @Liang

The main target audience we want to focus on are players of League of legends, especially the beginners, because they want to have a higher win rate by getting a higher kda.To achieve this, they want to investigate which area is helpful to increase their kda. Also, our research will attract the game designer of League of Legends, since by comparing how different areas affect the kda of player, game designer can make a reasonable balance.

Datasets used @Liang

We choose the data set of League of Legends Challenger matches from 3 servers: NA, EUW, KR from January 2022. Among these match records, we choose factors that we believe substantially impact the Kda. First of all,we choose the match records from all servers,since different game servers could have huge differences in average Kda due to players in each server could have uneven game level. Secondly, in the preparatory stage,we want to know whether Side, the lane we choose to play and champions we use would affect Kda. Finally, during the game phase,we predict that damage_objectives, damage_buildings, damage_turrents, damage_taken, turrents_kills, damage_total, gold_earned, minions_killed, vision score, time_cc could affect kda. However, there is some data missing in this dataset like what figure 1 shows.

Wrangling and Analysis methods @Liang & @ Jino

1. Data cleaning

Some values in our data sets are missing completely random,because we plan to combine all 3 servers' matches, and there are more than 15,000 matches, we decided to get rid of those MCAR instead of applying data imputation.

A	B	C	D	E	F	G	H	I	J	K
d_spell	f_spell	champion	side	assists	damage_ob	damage_bu	damage_tu	deaths	kda	kills
14	4	Leona	Side.blue	9	0	63	0	9		1
1	4	Ashe	Side.red		4885	9190	4885	5	2	3
4	12		Side.blue	13	1871		1871	7	3.571429	12

Figure 1: Example of data missing completely at random

2. Merge dataset

The dataset is provided in three different documents according to their servers (EU, KR, NA). For better analysis, we will Merge all three documents together, since there is no huge difference between these datasets.

3. Remove outlier

We decided to remove all suspected outliers and outliers since kda greater than 8 is not usual in the challenge players.

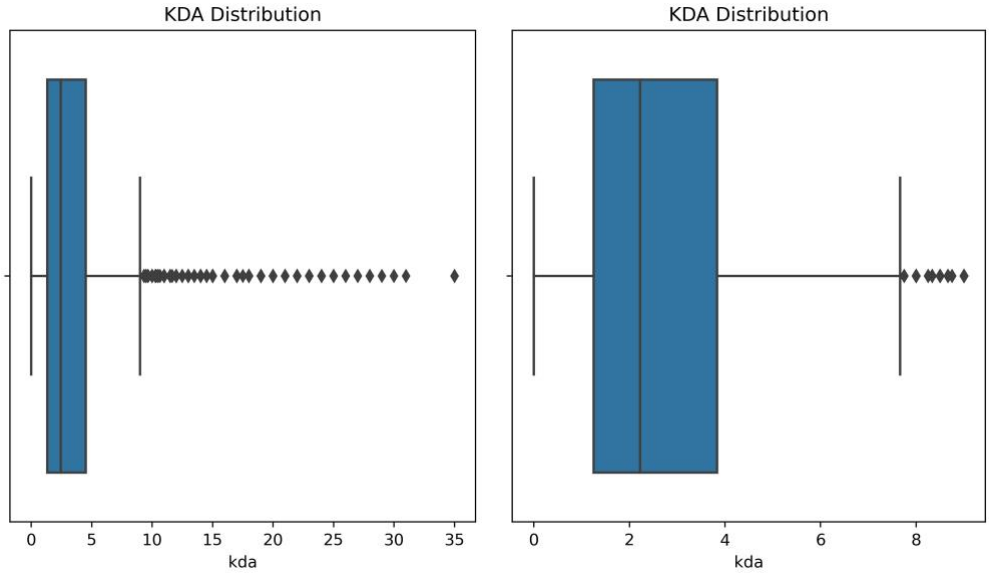

Figure 2: Boxplot of KDA Distribution before (left) and after (right) removing outliers

4. Data filtering

The time duration of each game in League of legends can vary largely. As the game length increase, most accompanied data will increase significantly.Therefore,In order to explore and analyze the relationship between each categorical variable better,we will drop those match records.However,there is no category in the dataset exhibit the time duration straight forward so we intend to filter the data by game level in each match,we will choose matches where level is between level 13 and 14 since the bar chart illustrates that majority of these matches end within this level range, the dataset will still be a good representation of the origin dataset.

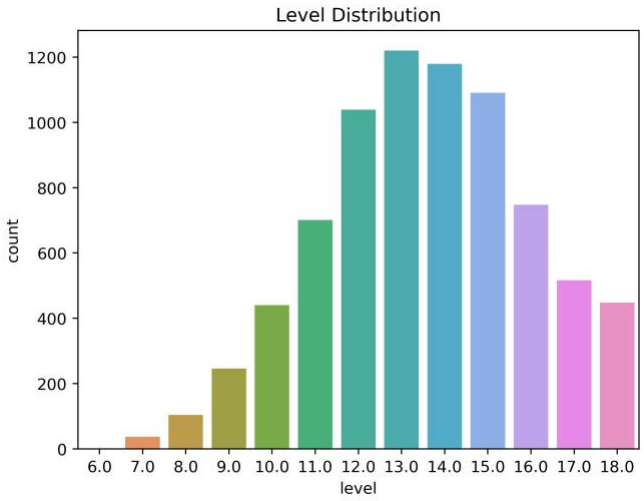

Figure 3: Level distribution plot

5. Visualization :
    - Box plot to identify the distribution of kda, helping remove outlier
    - Bar chart:
        - ❖ Find out the level with highest frequency
        - ❖ investigate the relationship between kda and champion pick, sides, whether top laner and jungler or not.
        - ❖ Investigate the champion with highest kda in three different servers.
    - Confusion matrix
    - Histogram to show the discreted kda distribution ● Regression plot and residual plot ● Pair plot:
        - ❖ Investigate the relationship between kda, kills, assists and deaths
        - ❖ Investigate the relationship between kda,damage building, turret kills, and gold earned

Data was then analyzed by searching for correlations between kda and our chosen features in order to see which aspects need to be focused to increase players' kda.We also wanted to see whether or not categories were correlated with each other to see if there were any relationship.To achieve this we made pairs plot between kda, kills, deaths, assists and minions killed to see if there were any correlations.We then calculated pairwise correlation coefficients and mutual information scores, and used different kinds of visualization tools on our data.

Results @Melody & @ William & @ Jino

To clearly see the relationship between kda and all other categories, a NMI plot and a correlation plot are sketched (Shown below). These two plots will be separated into two parts according to the role: TopLine_Jungle and Other.

TopLane_Jungle:

- Kills, assists, deaths:

    - ❖ Visualist:
      In both NMI plot and correlation plot, death is the most relevant category that affects the kda. kills and assists also strongly affect the kda, but it is to a lower extent than death. This can easily be observed from the NMI plot: the NMI of kills and assists is smaller than the NMI of deaths.

    - ❖ Mathematics:
      In League of Legends, the kda of a player in a game is simply calculated by the equation:

      $$Kda = (Kills + Assists) / Deaths$$

      It is clear to see that increasing on kills and assists will significantly increase the kda since these two categories are the numerator of the equation. And the kda is inversely proportional to the number of deaths which is the denominator of the equation.

      There is a simple rule in math: if both numerator and denominator increase by the same number, the fraction will get smaller. Therefore, reducing deaths is much more important than getting more kills or assists.

  These three categories will not be used in the modeling part due to their extremely close relationship with kda.

- Moreover, in the NMI plot, most categories, including damage_objectives, damage_buildings, damage_turrents, damage_taken, turrents_kills, damage_total, gold_earned, minions_killed, showcase some relations with kda. The correlation plot verifies these relations. Most categories are positively correlated with kda, except damage_taken which is inversely proportional to kda.

  All of these categories mentioned above in both plot is correlated with kda, therefore, they will be mainly used in the modeling part.
- The rest categories including d_spell, f_spell, side, level, time_cc, vision_score, server(na_server = 0, kr_server = 1, eu_server = 2) do not have clear relation with kda in NMI plots, therefore, they are considered not correlated with kda.
-

Therefore, they will be excluded in the modeling part.

Other:

- Most categories for Other have the same pattern as TopLine_Jungle.
- Damage_building: In both NMI plot and correlation plot, TopLine_Jungle has higher correlation or NMI than Other, indicating that for TopLine_Jungle, damage_building is a highly relative category with kda, but for Other, it does not make such a huge difference.
- Damage_taken: In the correlation plot, we can clearly notice that the others have lower correlation than topLine_Jungle. It means that others have a higher negative relation compared with topLine_Jungle.
- Vision_score: In the NMI plot, Other has much higher NMI than TopLine_Jungle. So vision score is a relative category with kda for Other but not for TopLine/Jungle. For this discovery, we haven't found a proper way to explain it.
- Damage_total: Differing with TopLine/Jungle, the damage_total of Other shows a negative correlation with kda. This strange difference is caused by the existence of a support role in Other, who deals little damage but holds a high kda.

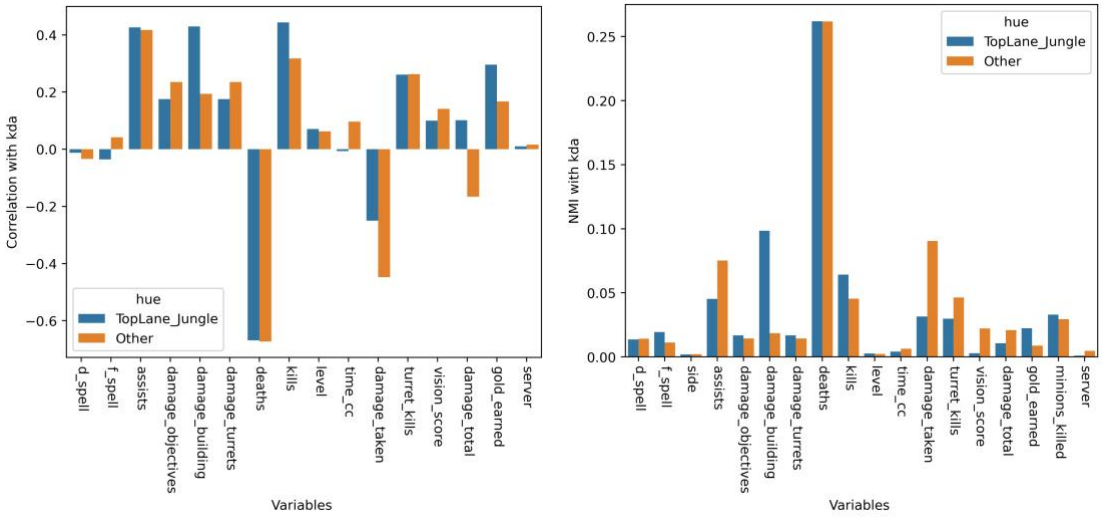

Figure 4: Correlation plot (left) and NMI plot (right)

The low correlation between kda and categories was largely to be expected, since kda does not simply depend on kills,deaths and assists,it is caused by a multitude of factors and can't be linearly mapped by any individual variable.kda has the largest mutual information score of 0.2619, while kills,assists,damage building,turret kills and gold earned have higher Pearson value,suggesting these are the features with high correlation to kda.
The pair plot between kda, kills, assists and deaths proves the conclusion above that kda is positively correlated with kills and assists and negatively correlated with deaths.
Besides, we can also find out that there is no significant relationship between kills, assists and deaths.

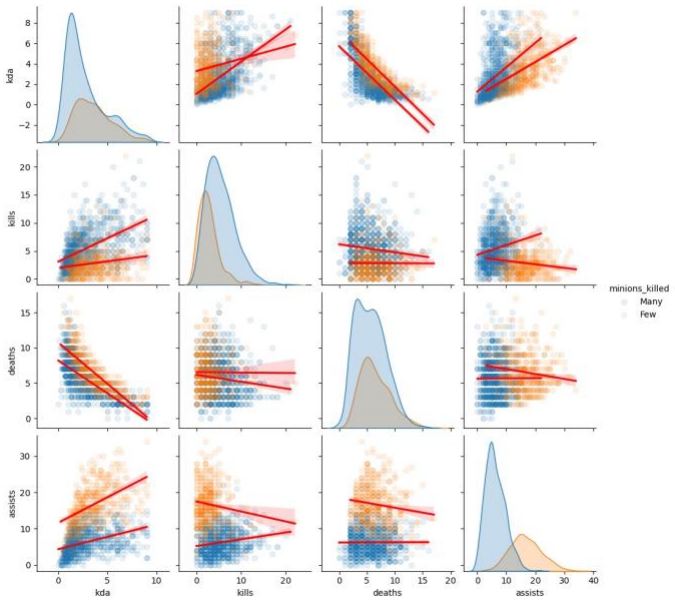

Figure 5: The pair plot between kda, kills, assists and deaths

As we can see, there is a positive relationship between kda,damage building, turret kills and gold earned.Moreover,correlation between kda and damage building is the highest.

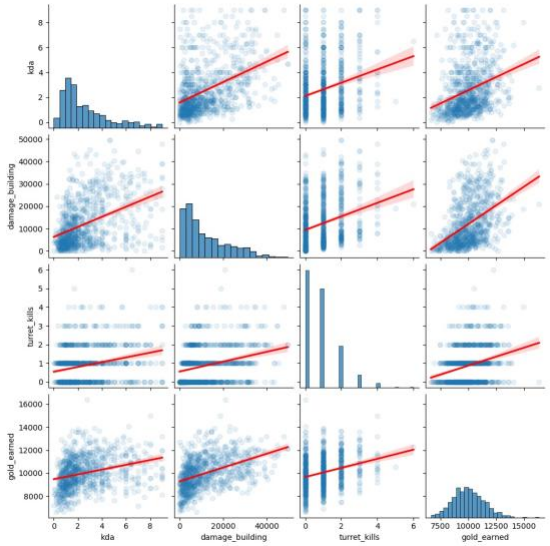

Figure 6: The pair plot between kda,damage building, turret kills, and gold earned

We made a confusion matrix by discreting kda values into 0,1,2 which are low, medium and high respectively.From the confusion matrix, the accuracy of Top and Jungle is 132/181=0.7293

which is relatively higher than others 0.6008,we also notice that for high and medium kda, the accuracy is significantly lower than low kda which is expected since this is based on an imbalanced dataset,the amount of data with low kda is larger than with medium and high kda.

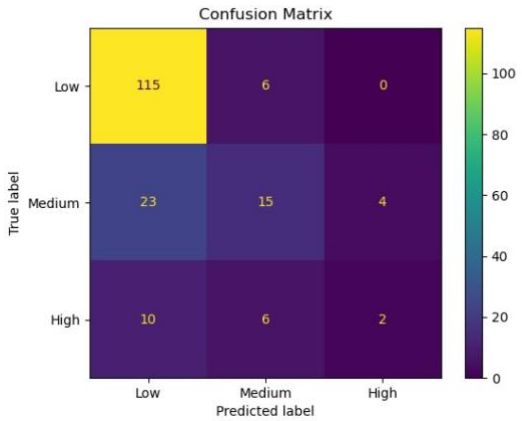

Figure 7:Confusion matrix of Top and Jungle

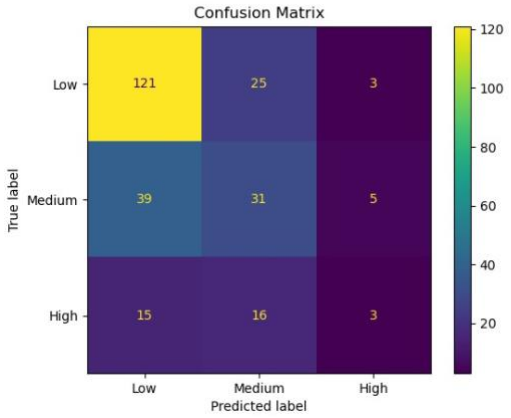

Figure 8: Confusion matrix of Other

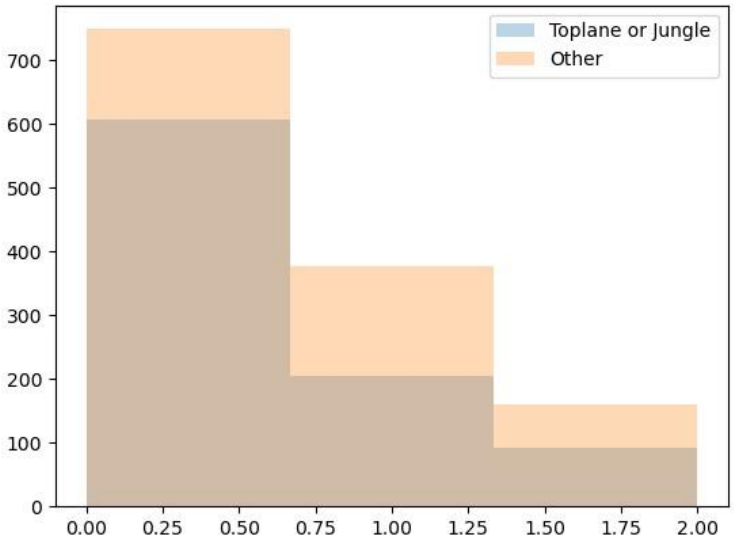

Figure 9: Discreted kda distribution

For the regression plot, it shows a relatively strong linear relationship between predicted value and actual value.

The residual is the difference between the observed and predicted value, as we can see the residual is linear,independent but does not have a constant variance

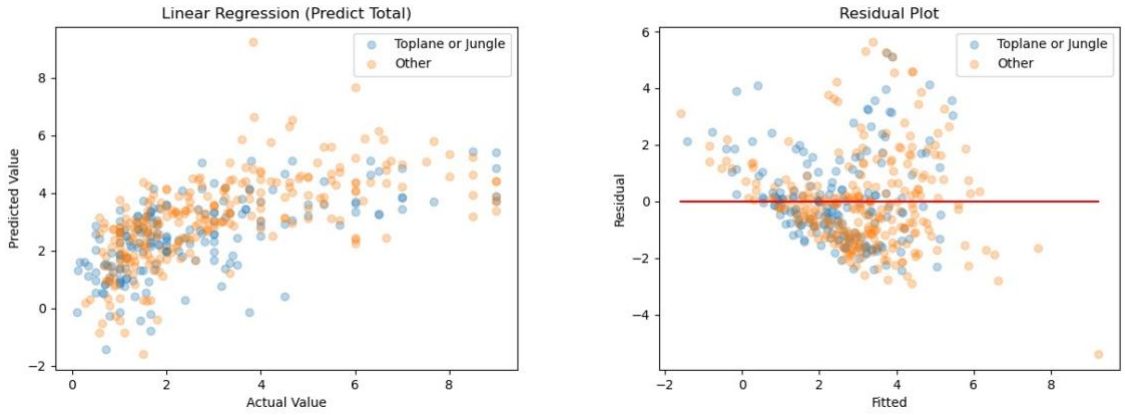

Figure 9: Regression and residual plot

Bar charts are used for analyzing the average kda for each champion in three different serves. In our graphs, we contain the top 10 highest kda champions (choose more than 5 matches) for recommending to get higher KDA.The MSE value for Top Lane and jungle is 2.4219 and 2.6889 for other.The r-squared value is 0.3856 for Top and jungle and 0.3853 for other, it is a relatively small value which means weak linear relationship.

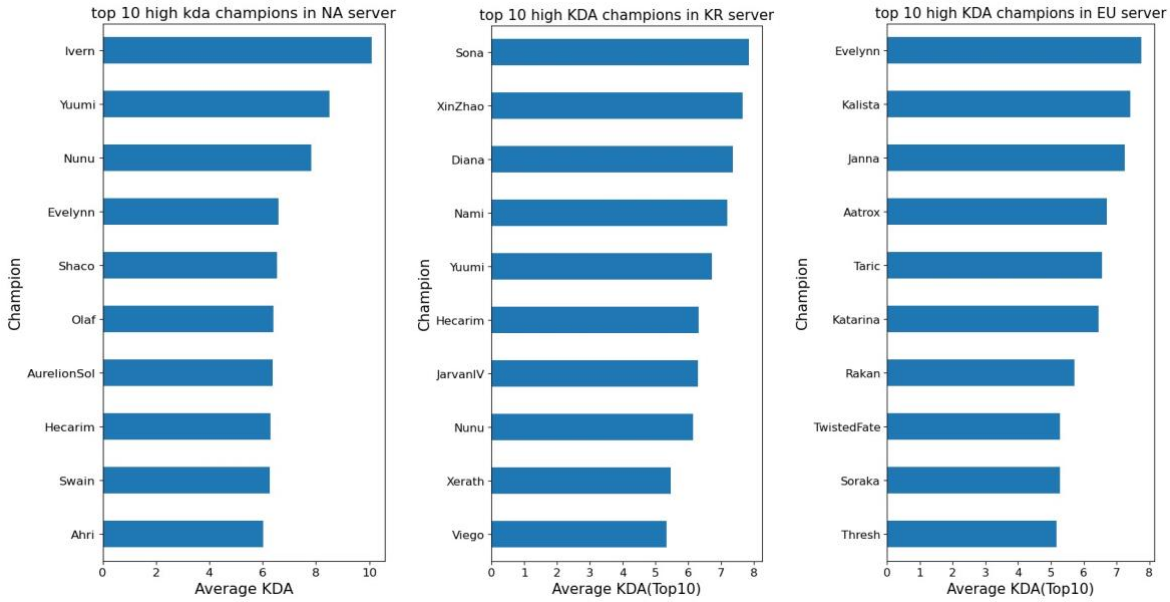

Figure 10: Top 10 highest kda champions in each serves

Limitation @William

1. The dataset misses some important data. Firstly, we do not know the result of matches. In LOL, the kda is usually higher in a winning game. Analyzing winning or not individually can make our analysis more reliable. Another important dataset missed is the time spent in one match. If a match time is much lower than average, the kda can be extremely high or low. Delete the extreme part will make the result more accurate.

2. The dataset only contains data from challenger matches. In LOL, the challenger only contains a very small part of all players.
3. There are some kinds of data which are not specific enough for analysis. Role is one of these kinds of data. In this dataset, it only has two kinds, top_jungle and others. Some other data could be influenced by a specific role. For example, a support is easy to have a higher vision score but they are not likely to have high damage_total which impacts our analysis of KDA. Another one is minions_killed, it only has two kinds: few and many. We do not even know the value divides two parts so we cannot easily use this data.

Future improvement @William

This dataset can contain more data like Master matches which will make the dataset have more enough data for analysis. Some missing data like the match result and time spent can be added to make our analysis more accurate. The data of role and minions killed should be more specific which make our analysis more accurate.

Conclusion @ Jino

According to the results we obtained and limitation of our design, we can conclude that in order to have a higher kda, first thing we can do is to choose a champion that has the higher average kda such as Ivern in NA server, Sona in KR server,Evelynn in EU server, it may vary largely due to different gaming environment but the idea is the same. On the other hand,in addition to kills, assists that are directly related to the calculation of kda, damage_buildings, damage_turrents, damage_taken, turrents_kills, damage_total, gold_earned, minions_killed should be focused, the higher these categories are, the higher kda will get.This result is similar to what we predicted, but we didn't find any correlation of vision_score, time_cc with kda. Finally, reducing the number of deaths is more important than getting more kills and assists.

Reference

https://www.kaggle.com/datasets/andrewasuter/lol-challenger-soloq-data-jan-krnaeuw