EDAV Final Project - Data Exploration for Dota2 - The International 2018

Minghao Chen, Han Ding, Feihong Liu, Zilin Zhu

1 Introduction

Dota2 is one of the most famous MOBA (Multiplayer Online Battle Arena) game in the world with more than 400,000 active players per day. And its international tournament, TI (The International) has become an annual grand event and attracts millions of fans' attention. We have chosen three aspects combining with our understanding of this game to present your analysis of the data from 401 games of TI8 (from qualifiers to the main event). We first describe the collection and quality of data in Section 2. And then we introduce our main analysis of the data in three aspects in Section 3 including heroes data, players data and the interesting fact we found during the exploration respectively in Section 3.1, Section 3.2 and Section 3.3. And finally, we draw our conclusion in Section 4.

---

2 Description of Data

In this section, we described our data including the data source, data obtained, data quality and data clear processing. All our data is obtained from the open source project and all the collection and clear part are done by using SQL, Python, Jupyter Notebook.

2.1 Collection of Data

We collect our data of Dota 2 from *OpenDota* which is an open source Dota 2 data platform providing highly detailed data by parsing game replay files. By exploring this website and use the API provided, we can easily obtain the amount of high-quality data with a specific query. From *OpenDota/Exploration* we can use the SQL query to obtain specific Dota2 game data, such as data of each hero, player, and team within specific date and league. There are 4 important tables in the database that we mainly focus on: matches, player_matches, heroes, and notable_players. Description of each table is as follow:

- **matches**:information about this match including the match_id, winner, start_time and so on.
- **player_matches**: information about each player in each match, including the match_id, player_slot, hero_id, items, gold, experience, fight_log and so on. This table is the most important and detail table for our project.
- **heroes**:information of 115 heroes in Dota2, including hero_id, hero_localized_name, and hero attribution.
- **notable_players**:infomation of 200+ famous profession players in Dota2, including player_id, nick_name, team.

More detail description of the tables please refer to the *schema of the database*

For obtaining the above tables from OpenDota platform, we implement the SQL query and save the result as CSV file. To get the same data we used in the project, implement following SQL queries in *OpenDota/Exploration API.*

- **matches**:

```
SELECT
matches.*
FROM matches
JOIN leagues using(leagueid)
WHERE TRUE
AND leagues.name = 'The International 2018'
ORDER BY matches.match_id NULLS LAST
```

- **player_matches**:

```
SELECT
player_matches.*
FROM matches
JOIN leagues using(leagueid)
JOIN player_matches using(match_id)
WHERE TRUE
AND leagues.name='The International 2018'
ORDER BY matches.match_id NULLS LAST
```

- **heroes**:

```
SELECT * FROM heroes
```

- **notable_players**:

```
SELECT * FROM  notable_players
```

2.2 Quality of Data

Quality of the Data from OpenDota is high as the game data are all parsed from game replay file so there is little data missing and all kind of detail data is included in. However, there is some missing value in the TI 8 league and notable player.

In fact, there are 115 heroes in Dota2, and there are only 111 heroes used by players during all games in TI 8 League, other 4 heroes are too weak or useless to compete with others in the world tier one competition. Also, there is some missing value of in the notable_player table. Players change from team to team occasionally, there is also some player change their nickname from time to time. So there are some missing or outdated value of team or nickname of the players.

2.3 Clear of Data

To extract useful data from the raw data obtained to draw figures, we use python and Jupiter notebook to deal with the original CSV files. Win, pick, ban and the kill-assist-death ratio of each hero and player are count from the origin data. Because not all heroes are played during TI 8, there is some missing win, pick and ban data, however, there are only several (3 or 4) missing data compare with 110 non-missing data, and their heroes are not what we focus on, so we just ignore them. All our pre-process code could be found in our *github repo*

The object of the data clear and reformat is that we care about hero and player rather than each game. We want to know which hero is strong so that leads to a win and which player is good during the TI 8 league. Therefore, we extract all the relevant data from the matches, then we analyze and visualize these data.

---

3 Main Analysis

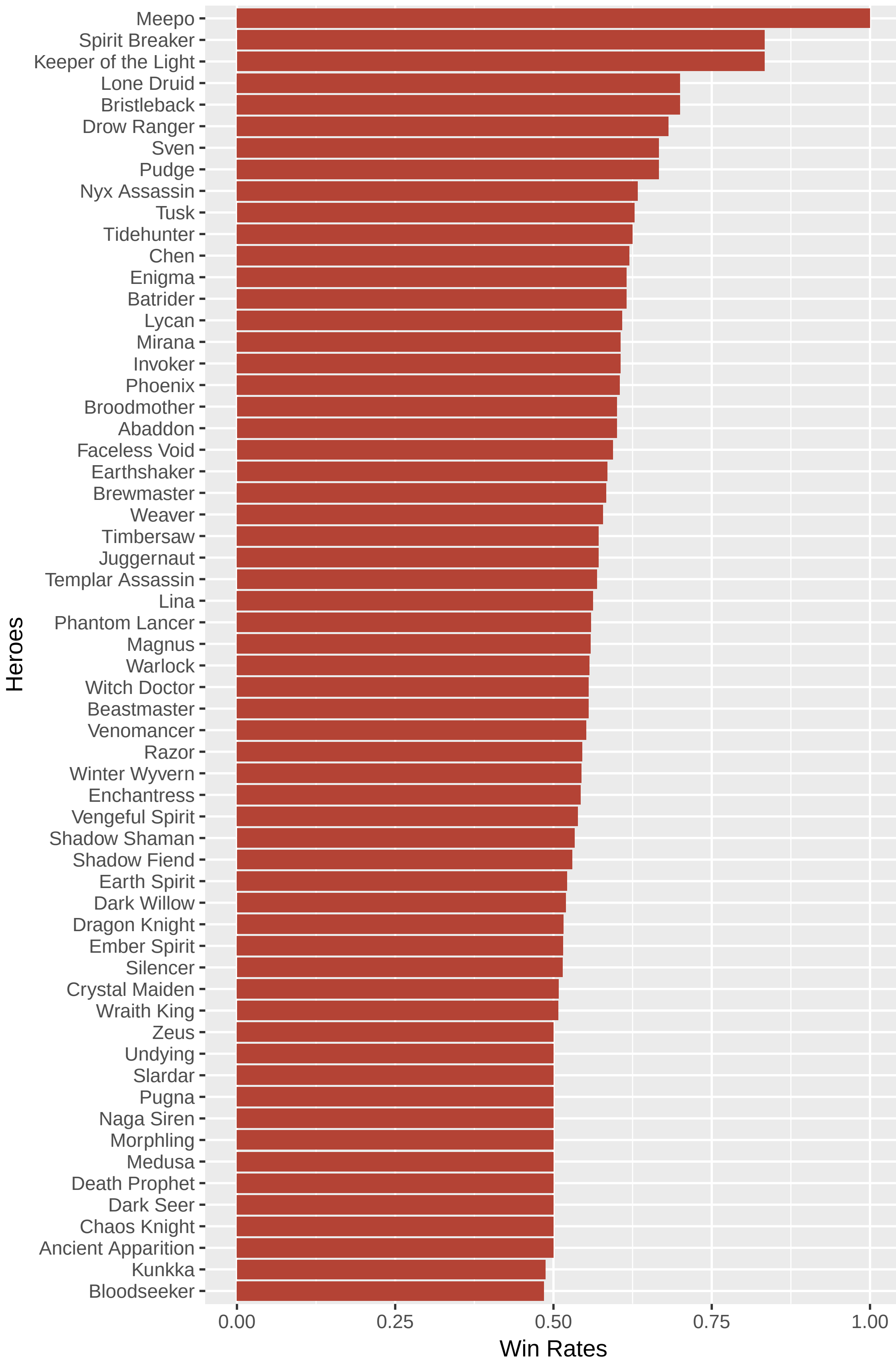
3.1 Hero Analysis

First of all, we would like to learn more about the heroes played in TI 8, we want to know which hero leads the team to a win, which hero is popular that everybody either picks or ban it. So, we will first draw the bar chart of win rate to show the "winnest" hero among them. Then we will draw a stacked bar chart of ban/pick/neither ban nor pick rate, and finally, we will draw a paired plot to show the correlation between the rates and kill-death-assist (kda) ratio.

3.1.1 Win Rate

Following bar chart shows the win rate of all the heroes.

```r
library(ggplot2)
df<-
  read.csv('../data/hero_part.csv')
order_<-
  order(-df[,'win_rate'])
selected_<-
  df[order_,][1:60,]
ggplot(selected_,
       aes(x=reorder(hero,win_rate),y=win_rate))+
  geom_bar(stat='identity',fill='#B44335')+
  coord_flip()+
  xlab('Heroes')+
  ylab('Win Rates')
```


Meepo
Spirit Breaker
Keeper of the Light
Lone Druid
Bristleback
Drow Ranger
Sven
Pudge
Nyx Assassin
Tusk
Tidehunter
Chen
Enigma
Batrider
Lycan
Mirana
Invoker
Phoenix
Broodmother
Abaddon
Faceless Void
Earthshaker
Brewmaster
Weaver
Timbersaw
Juggernaut
Templar Assassin
Lina
Phantom Lancer
Magnus
Warlock
Witch Doctor
Beastmaster
Venomancer
Razor
Winter Wyvern
Enchantress
Vengeful Spirit
Shadow Shaman
Shadow Fiend
Earth Spirit
Dark Willow
Dragon Knight
Ember Spirit
Silencer
Crystal Maiden
Wraith King
Zeus
Undying
Slardar
Pugna
Naga Siren
Morphling
Medusa
Death Prophet
Dark Seer
Chaos Knight
Ancient Apparition
Kunkka
Bloodseeker
Heroes
0.00
0.25
0.50
0.75
1.00
Win Rates

We can see that there are several heroes have dramatic win rate such as 100% or 80%. Let's have a look at them.

```
df<-read.csv('../data/hero_part.csv')
order_<-order(-df[,'win_rate'])
selected_<-df[order_,][1:3,]
selected_
```

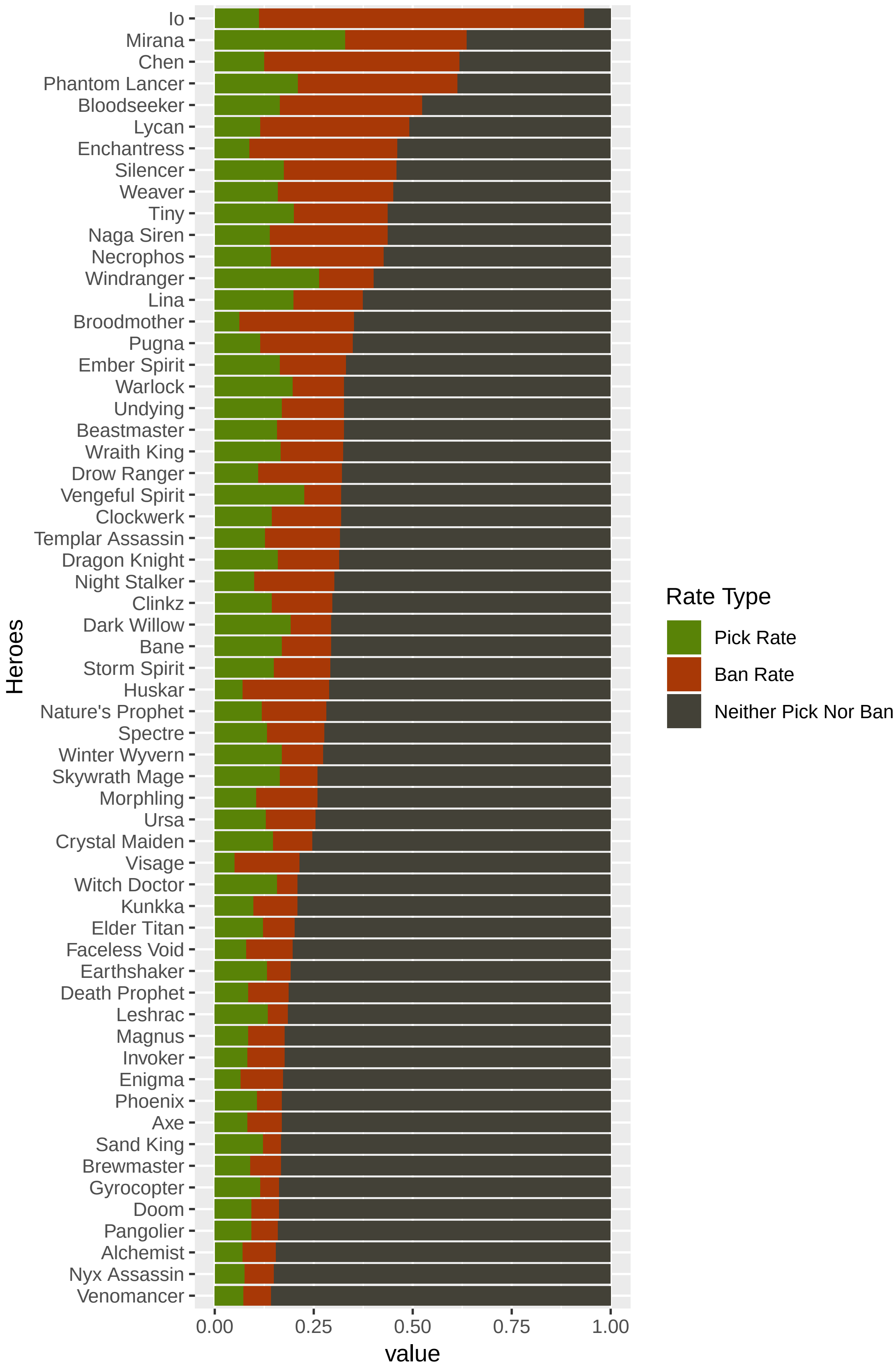
```
##                   hero win  win_rate      kda count pick ban other
## 79               Meepo   5 1.0000000 4.047619     5    5  22   374
## 77      Spirit Breaker   5 0.8333333 4.144823     6    6   4   391
## 80 Keeper of the Light   5 0.8333333 2.881944     6    6   3   392
##     pick_rate    ban_rate other_rate attr   type
## 79 0.01246883 0.054862843  0.9326683  agi  Melee
## 77 0.01496259 0.009975062  0.9750623  str  Melee
## 80 0.01496259 0.007481297  0.9775561  int Ranged
```

We can see that though they have a high win rate, they have a low pick rate, that is 5 or 6 picks among more than 400 games. So we analyze the phenomenon as either this is a variance because of the little sample or this is because those heroes are unpopular but strong and play an important role in a specific scenario.

3.1.2 Pick and Ban

Then we focus on which hero is most popular during the league by plotting a stacked bar chart on the rate of pick, ban or neither of each hero.

```
library(ggplot2)
library(tidyr)
df<-read.csv('../data/hero_part.csv')
order_<-order(-(df[,'pick_rate']+df[,'ban_rate']))
selected_<-df[order_,][1:60,]
selected_['other_backup']=selected_['other_rate']
colnames(selected_)<-
  c('hero','win','win_rate','kda',
    'count','pick','ban','other',
    'a','b','c','attr','type')
tidy_df<-
  gather(selected_,key=type_of_value,value=value,'a','b','c')
ggplot()+
  geom_bar(aes(x=reorder(hero,-other),
               y=value,
               fill=type_of_value),
           data=tidy_df,
           stat="identity",
           position= position_stack(reverse = TRUE))+
  coord_flip()+
  scale_fill_manual(name='Rate Type',
                    values=c("#598307","#A83806","#434137"),
                    labels=c("Pick Rate", "Ban Rate", "Neither Pick Nor Ban"))+
  xlab('Heroes')
```

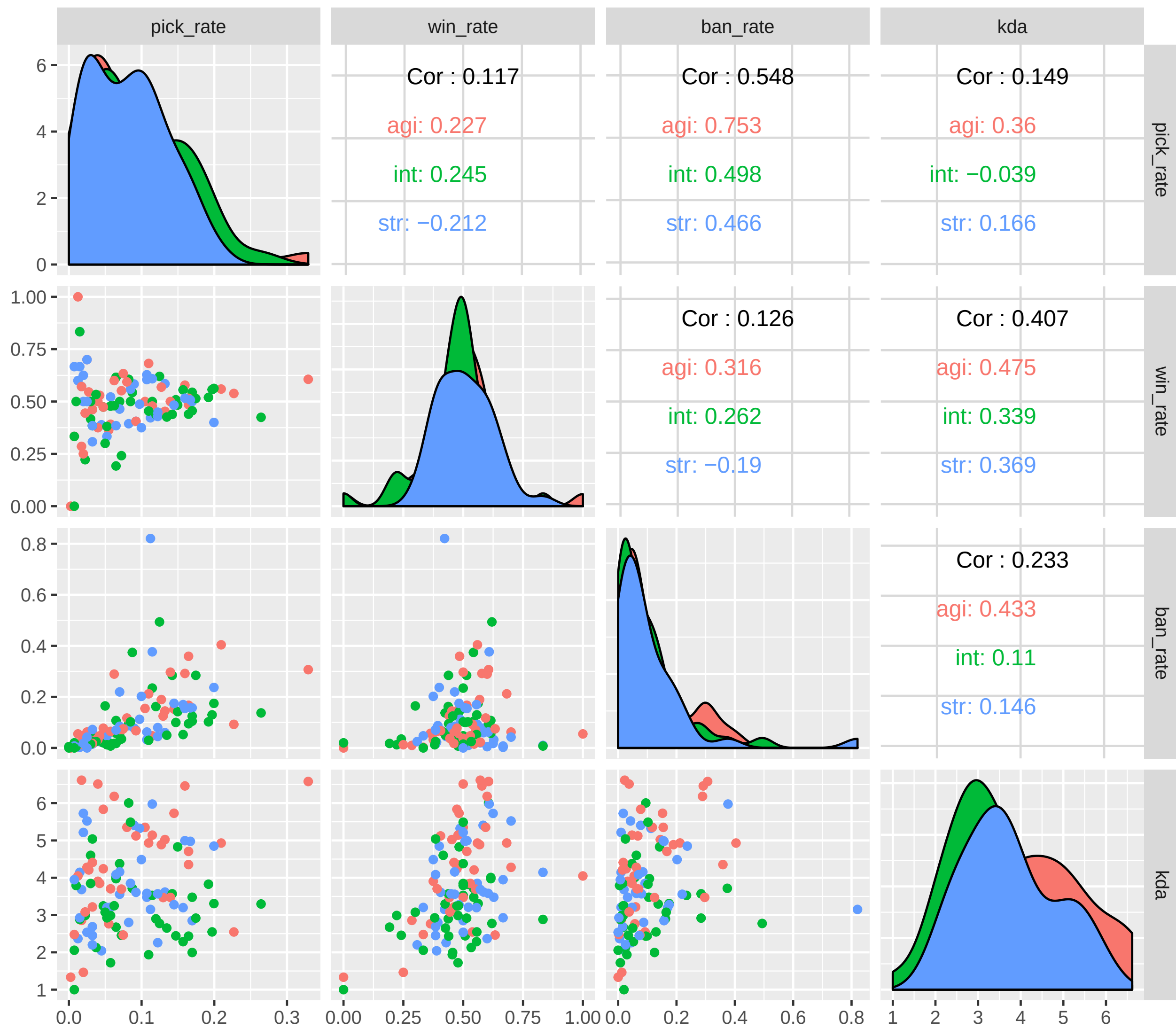

Io
Mirana
Chen
Phantom Lancer
Bloodseeker
Lycan
Enchantress
Silencer
Weaver
Tiny
Naga Siren
Necrophos
Windranger
Lina
Broodmother
Pugna
Ember Spirit
Warlock
Undying
Beastmaster
Wraith King
Drow Ranger
Vengeful Spirit
Clockwerk
Templar Assassin
Dragon Knight
Night Stalker
Clinkz
Dark Willow
Bane
Storm Spirit
Huskar
Nature's Prophet
Spectre
Winter Wyvern
Skywrath Mage
Morphling
Ursa
Crystal Maiden
Visage
Witch Doctor
Kunkka
Elder Titan
Faceless Void
Earthshaker
Death Prophet
Leshrac
Magnus
Invoker
Enigma
Phoenix
Axe
Sand King
Brewmaster
Gyrocopter
Doom
Pangolier
Alchemist
Nyx Assassin
Venomancer
Heroes
0.00
0.25
0.50
0.75
1.00
value
Rate Type
Pick Rate
Ban Rate
Neither Pick Nor Ban

We can see that the most popular heroes are IO, Mirana, Chen, and Phantom Lancer, these heroes either picked or banned by players because they are really popular among the games.

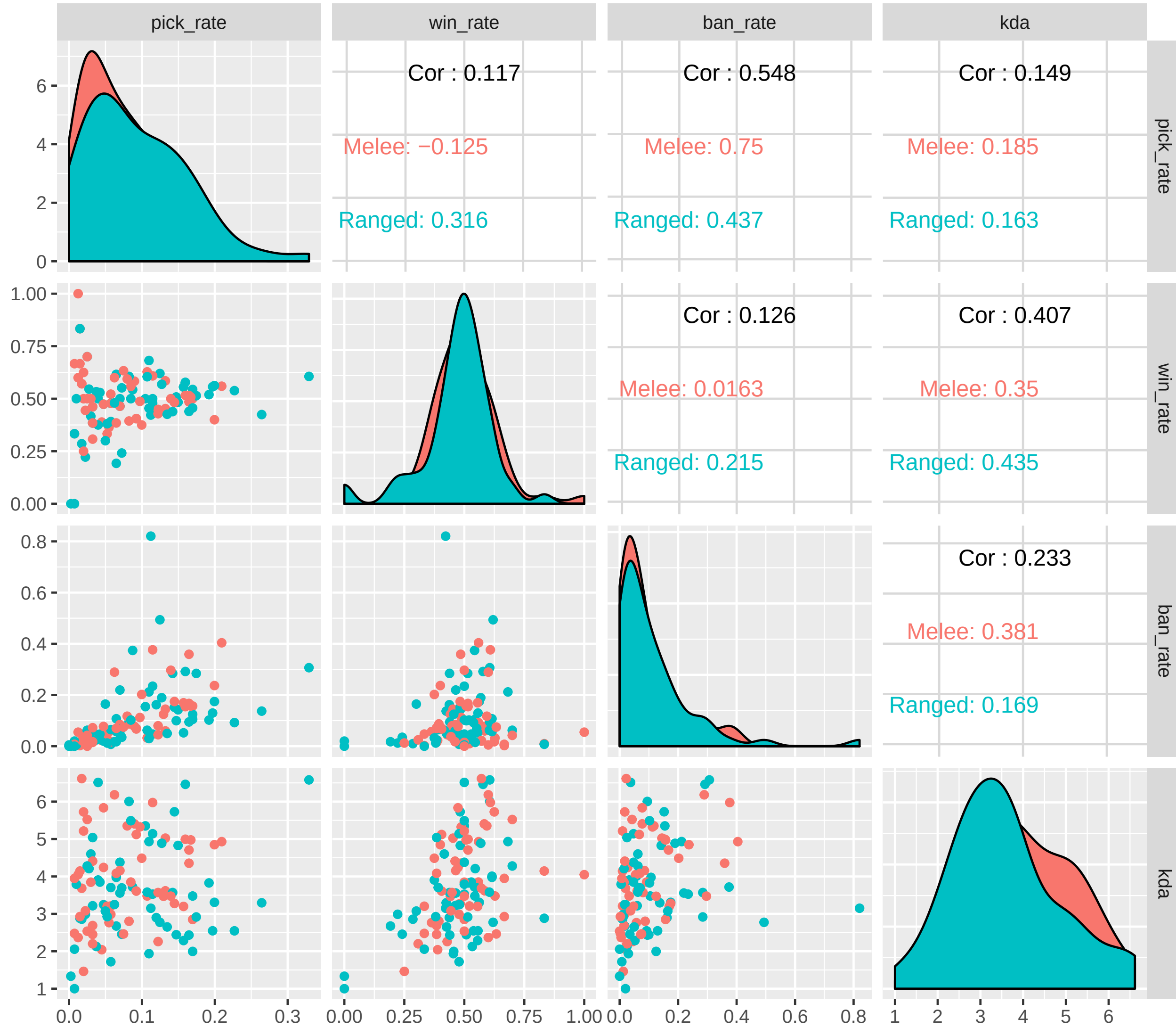
3.1.3 Correlations between Variables

We then analyze the correlation between the variables(win rate, pick rate, ban rate, and kill-assist-death ratio) by drawing pair plots. We separate heroes by their main attribution or attack type to see if the specific type of hero has specific correlation between the variables. The plot is as follow:

```r
library(GGally)
df<-read.csv('../data/hero_part.csv')
new_df<-df[c('pick_rate','win_rate','ban_rate','kda')]
ggpairs(new_df,mapping=ggplot2::aes(colour = df$attr))
```




```r
library(GGally)
df<-read.csv('../data/hero_part.csv')
new_df<-df[c('pick_rate','win_rate','ban_rate','kda')]
ggpairs(new_df,mapping=ggplot2::aes(colour = df$type))
```



It seems all kinds of heroes have same correlations that ban and pick have a strong correlation, kda and win have a strong correlation. If a hero is so popular that everyone wants to pick it, there is also a high probability that someone will ban this hero to counter their competitor. Meanwhile, a hero with high kda means it has a high probability killing other heroes and a small probability killed by others, which will lead to a higher chance of winning.

In general, hero Meepo wins always the games it plays, however, there are only 5 games Meepo is picked. Among more popular hero who picked more than 10 games, Draw Ranger is the one with the highest possibility of winning, followed by Nyx Assassin and Tusk.IO is the most popular hero who is picked or banned in more than 90% games while other heroes have at most 65% chance to be picked or banned, including Mirana and Chen.Meanwhile, lower the possibility to be picked, lower the possibility to be banned, but ban probability is always higher than the pick probability for the most popular heroes. Among the variables, ban rate and pick rate are most correlated variables suggesting that players tend to either pick or ban strong heroes. Also, KDA ratio and win rate are correlated somehow suggesting that more killing, more assisting and less death leads to winning.

3.2 Player

We also care about all the professional players in the TI 8 league. We would like to know who outperformed during the games by analyzing with a different dimension. The following subsections describe players from three dimensions: kill-death-assist ratio(KDA), player diversity and player stability.

3.2.1 Kill-Death-Assist Ratio

KDA refers to kills, deaths, assists.Its calculation formula is (kills + assists)/deaths. It's a direct indicator of players' performance in a game. People always care who is the best player, so we use the KDA that counts all the players who participate in TI 8 and use them as a measure of the player. At the same time, we also care about the stability of the players. We prefer stable players. Because we want players to be able to carry the team in most cases. Therefore, we performed a box plot analysis for each player to analyze the player's KDA distribution.

happydota

According to the graph of KDA, we find that 'Resolut1on' has the highest KDA mode and is 2 more than the second player, and he is a stable player. However, his team Forward Gaming just got 7th place at last. The second player is 'No[o]ne'. His KDA median is 5, and he has performed very well in how long. There is even a game his KDA has reached an astonishing 26. He is in Virtus.pro and only gets 5th place in TI 8. This indicates that a player can't master a game. DOTA2 is a team game.

3.2.2 Player Diversity

Hero pool of a player is the numbers of heroes he or she is familiar with and it is of vital importance for judging a player. Besides KDA, as an indicator of performance directly, diversity of heroes that player choose can reflect something else. For example, in Dota2 games, if we know the enemy mid has a small hero pool, sometimes we will ban his familiar heroes. Then he has to pick some unfamiliar heroes, which may lead to his bad performance in the game or their loss. Besides, if a player has a deep pool of heroes, the coach can arrange more tactics to fit more situation. Last but not least, there is a heroic restraint problem in DOTA2. If the player is able to master more heroes, he will be more likely to gain lane advantage and then help the team to win. For the above reasons, we hope to find out the players with the deepest hero pool in TI8. And clarify the relationship between the depth of the hero pool and the game wins and losses.

According to the graph, we notice that '33' has the deepest hero pool, he is a player from Optic Gaming. His team only get 8th place in TI 8. According to the analysis of KDA, '33' has high KDA among all the players. Therefore, we guess that the diversity of the hero pool may relate to the judgement of a player. The second player in this diversity rank is 'Ceb'. He is one member of OG, which won TI 8. And the fifth player 'JerAx' is also from OG. What's more, the third and fourth player 'Zai', 'Busen' are both from Optic Gaming. Although Optic Gaming didn't get a high place in TI 8, they gave the audiences a deep impression in mind. In conclusion, the hero pool of a player is strongly related to his performance.

3.2.3 Team Winning Rate

As players of Dota2, the most significant thing we concern about in TI 8 is who wins the TI 8 at last. Meanwhile, we would like to know which team perform well in the TI 8. As same as traditional sports, there's always some surprising things happen. For example, a team may have a high win rate in qualifier games but may be out early in the main event. Therefore, we would like to know whether there's a team that has a high win rate but gets out early in TI 8. First, we compute all teams' win rate and reorder them by decreasing.

Suprisely, the team has the highest win rate is not the champion of the TI 8, they are in third place in TI 8. OG, which has the second highest win rate, is the champion of TI 8. Meanwhile, the graph indicates that the team has a high win rate always got good places in TI 8.

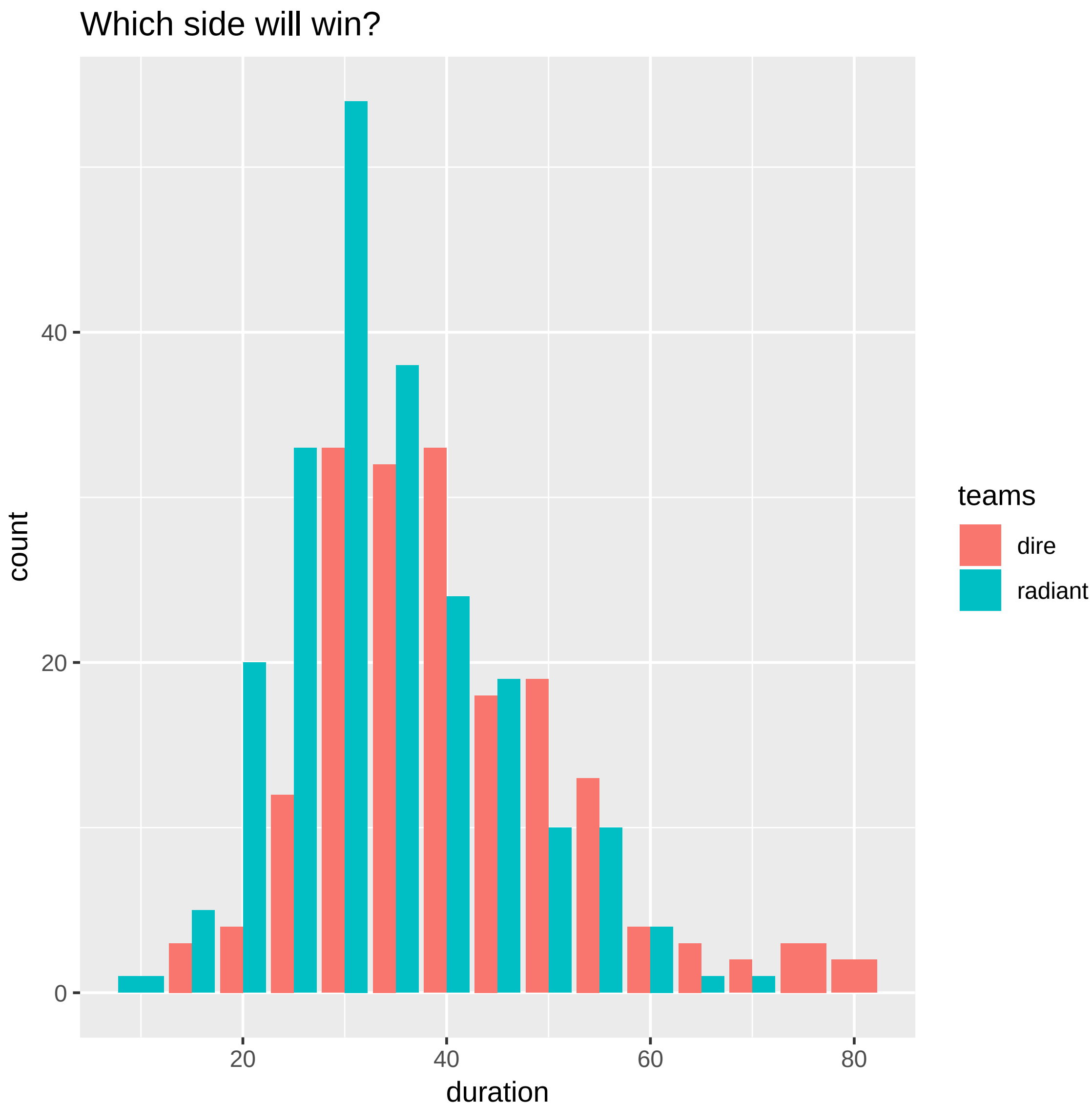
3.3 Interesting Discovery

Except for the discussion above, we want to predict which team will win during the game and know the factors affecting the result. As an experienced player, when I am playing Dota2, I notice that the map of Dota2 is asymmetry and the jungles are different for two teams. Meanwhile, the difference between position will influence the performance of players, since some players will feel more comfortable to move the mouse from down to up, like me. Therefore, we are curious about whether the different side of the game will affect the result of the game.

3.3.1 Winning Rate of Radiant and Dire

We find that dire win 181 games over all 401 games. Does it mean that radiant appears to be more winnable from the start of the game? If it's true, will the win rate appear to be different with different duration of the game? Will dire to be re winnable when the games become a late game? To answer these question, we divide the duration of the game into different periods and count all TI 8 games within each period separately.

```
library(ggplot2)
df = read.csv("../data/time_num_df.csv")
g <- ggplot(df, aes(x=duration,group = factor(teams), fill = teams))+
  geom_bar(position=position_dodge())+
  ggtitle('Which side will win?')
g
```



According to the graph, we could see that radiant appears to win more games when the game duration is less than 40 minutes. However, when games become late, dire appears to become more winnable and catch up with radiant and exceed radiant then. This is quite interesting. That means radiant are will win more in early game or mid game but radiant will win more late games.

What's more, we compute a series fractions of dire win rate in TI 8. And draw a dynamic line chart to display the relationship between dire win rate and game duration in our website(*happydota*). (Since the radiant's win rate line is a symmetry line about 0.5 with the dire's win rate line, we didn't draw it)

We could see an obviously increasing trend for dire's win rate. This incredible result may be caused by the ban and pick a direction or the difference of the map. But one thing can be sure that if the game become late, dire has more advantages.

3.3.2 Winning Rate with Gold

After discussion then relationship between win rate and sides, there's one more thing that every player care about - gold. Gold is one of the most important indexes of Dota2. Intuitively, a team with more gold could get better items and then do more damage and at last more likely to win the game. Therefore, we want to find whether it's true or not. To get further, we want to know the precise win rate when we get some extent gold advantage. The discussion is done in the following steps.

First, we define what is a gold advantage. Intuitively, one thousand gold advance at 10 minutes will be clearly different with one thousand gold advance at 40 minutes. So, we define the gold advantage by using the gold of first-team divided by the second team. Besides, we don't consider the influence of time in this question and divide the game into minutes to get more data(the 63-min game will have 63 gold advantage for one team, 126 for two). Then we calculate the win rate in every gold advantage for both team and draw the line chart. At first, we find that there's some point that appears to be impossible, as a team with a huge gold advantage but low win rate. After repeat checking, we find that it is because that there're few games at that gold advantage, so the result appears to be randomly distributed. To solve the problem, we use top coding, let the gold advantage less than 0.7 to be one category and gold advantage more than 1.4 to be another category. The result is displayed in our website(($happydota$)).

In conclusion, we find that dire appears to be more winnable in most case but when the gold advantages are some extreme values.