COMP5048 Assignment 1

Yifan Wang
SID (510364904)

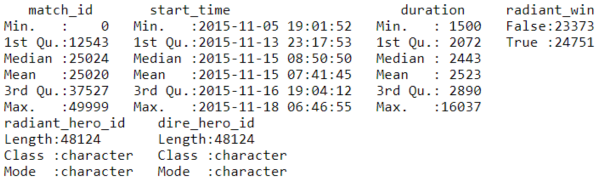
***Abstract*—Polman predicted in 2018 that the economic worth of the e-sports industry will reach 23.5 billion U.S. dollars in 2020 [1]. E-sports is a rising industry that has developed rapidly in recent years [2]. The most important and influential one is the professional competitions of various games, especially the annual DOTA2 The International(TI) competition. This article mainly discusses the data set related to the DOTA2 game. The purpose is to visualize the data and provide some help for professional players or coaches to analyze the data.**

I. Introduction

The rapid development of the electronics industry has spawned a series of high-value game competitions. Among them, the prize money of the TI event of DOTA2 breaks the record of the prize money of the e-sports event created by itself every year. At TI10 in 2021, the player prize pool has exceeded 40 million U.S. dollars, and in 2019 it was 34 million U.S. dollars (not held due to the COVID-19 in 2020) [3]. In order to win huge bonuses, in recent years, professional teams have trained their own data analysts or handed over to specialized data analysis teams to help them analyze the data in the game. Dota2 has complex gameplay, diverse lineup options, and changeable gameplay, all of which bring great potential market and value to data analysis and artificial intelligence. For example, OpenAI defeated high-end players [4], and various machine learning and deep learning models were used to simulate lineups to predict the win rate of the game [5]. Next, we will comprehensively analyze the visualization methods of DOTA2 competition data from eight perspectives, including data collection, data description, and visualization analysis, etc.

II. Nature of the data

The data set I used was previously used by COMP5310. At that time, I extracted the key data and used the deep learning models to predict the winning percentage of the lineup. This data set was first created by Joe Ramir in 2015 [6], Then Devin Anzelmo released an improved second version [7], and all we use are the second version. This data is collected by the Opendota website and contains 50,000 ranking ladder matches from the Dota 2 data dump [8]. This data set has 18 csv files and one json file (generated by Opendota when the data is obtained), and the total size is about 1.31GB. These data mainly describe all the information contained in a DOTA2 game, such as start time, match ID, selected heroes, which side wins, the duration of the victory, etc. Our data is extracted from player.csv and match.csv. Since there are 10 heroes in a match, the data contained in the player.csv file is 10 times that of match.csv. We extract ”match_id”, ”hero_id” from the player.csv file. We extract ”match_id”, ”radiant_win”, ”duration”, ”start_time” from the mathc.csv file. We divide the 10 heroes into Radiant hero id and Dire hero id according to their different camps, each of which contains the id of five heroes [9]. We have also removed matches that are less than 15 minutes long, because these matches are of little significance. Dota2 matches usually last between 30 and 40 minutes. All these data preprocessing was done in COMP5310 assignment, and finally our data set contains 48124 obs. and 6 variables. There is a summary shown in Figure 1, completed by R language

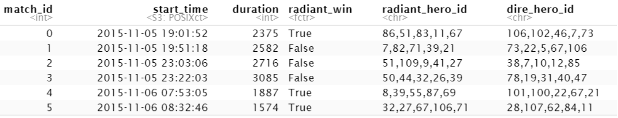


```
   match_id       start_time                      duration      radiant_win
 Min.   :    0   Min.   :2015-11-05 19:01:52   Min.   : 1500   False:23373
 1st Qu.:12543   1st Qu.:2015-11-13 23:17:53   1st Qu.: 2072   True :24751
 Median :25024   Median :2015-11-15 08:50:50   Median : 2443
 Mean   :25020   Mean   :2015-11-15 07:41:45   Mean   : 2523
 3rd Qu.:37527   3rd Qu.:2015-11-16 19:04:12   3rd Qu.: 2890
 Max.   :49999   Max.   :2015-11-18 06:46:55   Max.   :16037
 radiant_hero_id    dire_hero_id
 Length:48124       Length:48124
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Fig. 1. Summary of the data set.

Figure Labels: **match_id**: Represents the number id of each game. **start_time**: Represents the start date of each game. **duration**: Represents the duration of each game. **radiant_win**: Represents whether Radiant wins each game. **radiant_hero_id**: Represents the id of the Radiant heros who played in each game. **dire_hero_id**: Represents the id of the Dire heros who played in each game.

Figure 2 shows the specific information of the first 5 obs, in order to more clearly express what the specific content of each attribute is. The specific analysis and explanation of each attribute is placed in section III.



match_id	start_time	duration	radiant_win	radiant_hero_id	dire_hero_id
<int>	<S3: POSIXct>	<int>	<fctr>	<chr>	<chr>
0	2015-11-05 19:01:52	2375	True	86,51,83,11,67	106,102,46,7,73
1	2015-11-05 19:51:18	2582	False	7,82,71,39,21	73,22,5,67,106
2	2015-11-05 23:03:06	2716	False	51,109,9,41,27	38,7,10,12,85
3	2015-11-05 23:22:03	3085	False	50,44,32,26,39	78,19,31,40,47
4	2015-11-06 07:53:05	1887	True	8,39,55,87,69	101,100,22,67,21
5	2015-11-06 08:32:46	1574	True	32,27,67,106,71	28,107,62,84,11

Fig. 2. First 5 obs of the data set.

III. Data analysis

A. Consumers of the data

The main consumers of this data set are divided into three categories, one is data analysts or students who are interested in game data, one is data analysts and professional players who participate in professional teams in the e-sports industry and the other is potential users who do not use these data but are interested in the results of data analysis include high-ranking players and experienced players who have a deep understanding of the game. Data analysts and students who research data usually use some machine learning models or deep learning models to classify or predict data [5]. Analysts and professional players of professional teams usually care more about the results of the analysis rather than the process. Their main purpose is to use the results of the analysis to help them increasing the chance of winning the game. For the third category of consumers, it is their purpose to get some interesting tips or uncommon sense knowledge through data analysis, which can make them more enjoyable in the game.

B. Typical visualization

There are some examples of typically visualizing DOTA2 data set. Figure 3 shows win rate heat map when two heroes are in the same team [10]. Most of the points in the figure are in the 50% area, but there are still some outliers, indicating that some hero combinations have obvious high or low winning rates. I think one disadvantage of this figure is the color distribution. The red with high winning rate and the blue-black with low winning rate are difficult to distinguish at a glance against the yellow-green background.

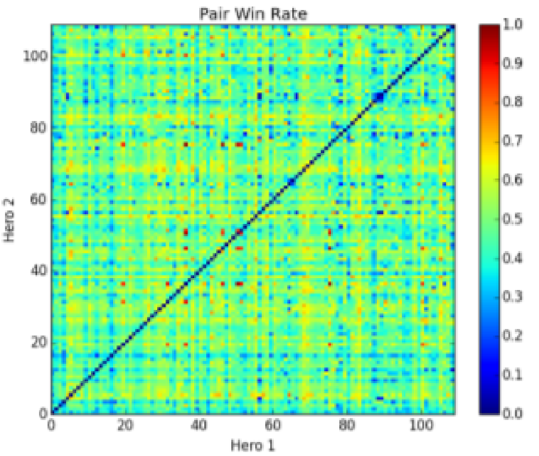


Figure 5: Pair Win Rate Heat Map

Fig. 3. Pair Win Rate Heat Map from other resources

The Figure 4 shows a histogram of the number of matches started in a period of time [11]. After about 1447300000, the number of matches suddenly increased. The author of the picture understood that this was due to OpenDota changing the algorithm of the collection game [11]. The disadvantage of this is that the time is displayed in Unix timestamp format, which does not look intuitive, and it would be better if it can be converted.

Figure 5 was made in my previous COMP5310 homework. It shows a histogram of the 10 heroes with the highest usage rate and the 10 heroes with the lowest usage rate. One disadvantage is that although the color changes according to the number of times the hero is selected, the order of the abscissa is the hero id, which makes the graph look a bit messy. If the abscissa is sorted in the order of the number of selections, the effect should be better.

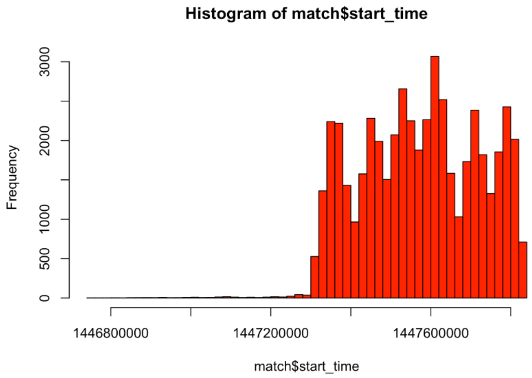


Fig. 4. Histogram of match$start_time from other resources

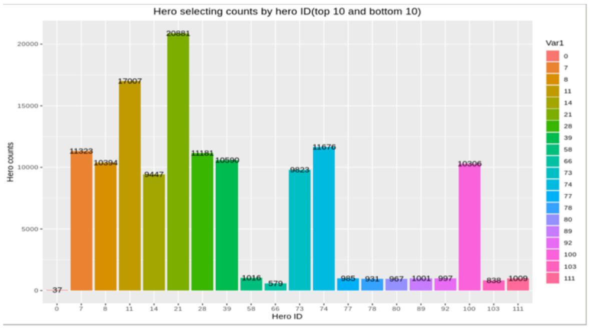


Figure 2. Hero selecting counts (top 10 and bottom 10)

Fig. 5. Hero selecting counts (top 10 and bottom 10)

Table 1 below analyzes the attributes of the 6 variables in the data in detail. Since hero id is norminal data, it does not have measurement significance, so when we visualize data, we can count the number of each id for comparison, as shown in Figure 5. Because the start time is interval data, the values are equidistant, so when visualizing, we can select an equidistant range to count the quantity, such as Figure 4, instead of hero id, only individual can be counted. Like radiant win is norminal data, we can combine hero id to the counted number of Radiant wins, and then convert it into winning percentage, as shown in Figure 3. For the match id, although it belongs to norminal data, it has no meaning in itself, and the role of id is only convenient for people to find. It is basically not used for data visualization. Duration is ratio data, which is meaningful and equidistant. It is of great value when doing some special analysis, such as finding a lineup that ends the battle quickly, or analyzing the change in winning percentage of a lineup under different lengths of time.

TABLE I
TABLE OF DATA TYPES

Attribute	Data type	Analysis
match_id	Norminal Data	The value is id, and has no quantitative meaning.
start_time	Interval Data	The values are dates, in order and the values are equally spaced.
duration	Ratio Data	The value is the length of time, equidistant, can be multiplied and divided.
radiant_win	Norminal Data	The value is boolean true or false, and has no quantitative meaning, the value is just a substitute for the name.
radiant_hero_id/ dire_hero_id	Norminal Data	The value is hero id, and has no quantitative meaning, the value is just a substitute for the name.

C. *Typical mistakes*

1) Survivorship bias: Sometimes data can be seen by people as belonging to survivors and may lead the wrong direction. For example, (Luna, Lycan) has a 90% win rate in Figure 3 [10], but in a real game, few people choose Luna and Lycan at the same time, because they both belong to the first position and require a lot of economy. If you choose two positions at the same time, it is more likely to lead to a loss of the game. In the heat map, the reason for their high winning rate may be that it is precisely because the sample of the number of games that choose both at the same time is small, which happens to lead to the extreme winning rate of the game that chooses both.

2) Correlation and causation: Sometimes the data is highly correlated, but not necessarily causal. For example, in the second half of Figure 4, the number of games after 1447300000 has increased sharply than before. This does not mean that the number of DOTA2 games suddenly broke out after that point in time. For example, suddenly many more players came to play DOTA2. This was verified by the author of the picture, but the data collection website modified the algorithm for collecting data at that time point, which resulted in the seemingly correlation between the number of matches and time [11].

3) Misleading data visualization: Sometimes data visualization is not done well, which may mislead the audience very badly. For example, in Figure 5, the abscissa has shown the id of different heroes, and the color gradient is generally considered to reflect the value from high to low. Here I reflect the color gradient to the norminal data that has no practical meaning and order. On hero id, it may mislead the audience that the ten on the left have high winning percentages and the ten on the right have low winning percentages. In fact, the first one on the left contains a hero with the lowest winning percentage. Therefore, the correct way is to reflect the color gradient to a numerical value that can be compared, such as the number of wins on the ordinate, and the result of data visualization is more accurate.

IV. DATA VISUALIZATION

A. *Visually representation*

Because the game id itself doesn't make sense, we ignore it when visualizing the data and only deal with the remaining five features. First we visualize the game start time and duration. Duration We counted the sum of all the duration over the length of an hour. For start time we set the daily 7AM to 6PM to day time, 6PM to 12PM tonight, 0AM to 7AM to midnight, respectively, in yellow, blue, and gray. As shown in Figure 6, we can see that the total amount change is expected to be regular, with each peak in the night time period, indicating that the number of people online is highest during the 6Pm to 12PM period.

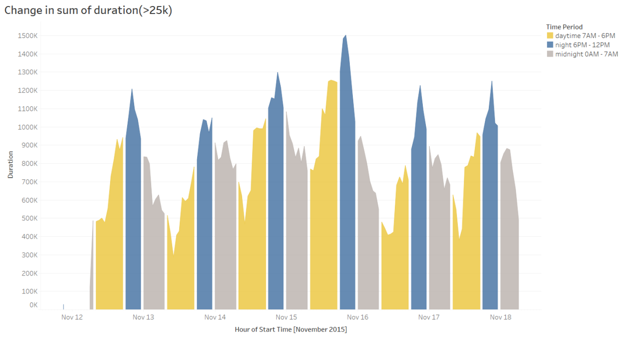


Fig. 6. Change in sum of duration

Then for Radiant Win, Radiant Hero Id, Dire Hero Id, these three attributes, direct visualization will be troublesome, because Radiant Hero Id and Dire Hero Id are made up of 5 id combinations, itself is just a string, and each combination is different, there is no correlation. So we used the R language and did some pre-processing of these three data. First of all, we count the number of appearances of all the heroes in the Radiant lineup, representing the total number of appearances of these heroes on the edge of Radiant. Similarly, we also count the number of times all heroes id for the night. These two data are reserved. Then we split the data set into two parts based on whether Radiant wins or not. In two parts, the number of heroic appearances of the winning side is counted separately. Divide this number of appearances by the number of just-spare appearances to get the winning percentage of each hero in the Radiant or Dire. In the end, as Figure 7 shows, the size of the block represents Dire's winning rate, the color from red to yellow to green represents Radiant's winning rate, and the number represents the hero's id. Perfectly blend the three properties of our dataset. And it's clear that Radiant and Dire don't differ much from heroes, such as the 57th and 42nd heroes with the highest winning percentages, who have a high winning percentage on either side. Some heroes can also be seen winning more than Dire in Radiant, such as Hero 48. But since there is no measure of winning, we can only judge by shape and color, so the modified version we discuss and present in Subsection C and D of Section IV.

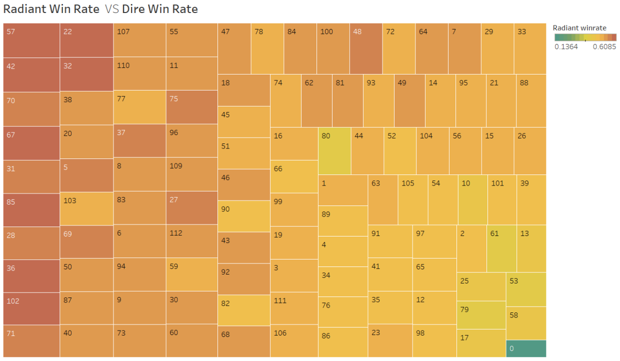


Fig. 7. Radiant Win Rate VS Dire Win Rate

B. Symbolic representations

The following two hand-drawn sketches Figure 8 and 9 represent symbolic expressions about Figure 6 and Figure 7. Comments on symbols are in Figure 10. Q is quantities, O is ordered component, C is color, S is the suqare size, Solid arrow is dimension of the plane utilized in HOMOGENEOUS, Dashed arrow is dimension of the plane utilized in HETEROGENEOUS.

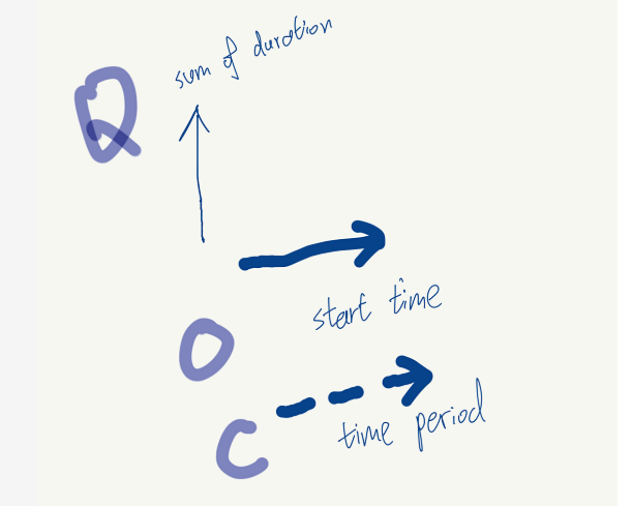


Fig. 8. Symbolic representations of Fig.6

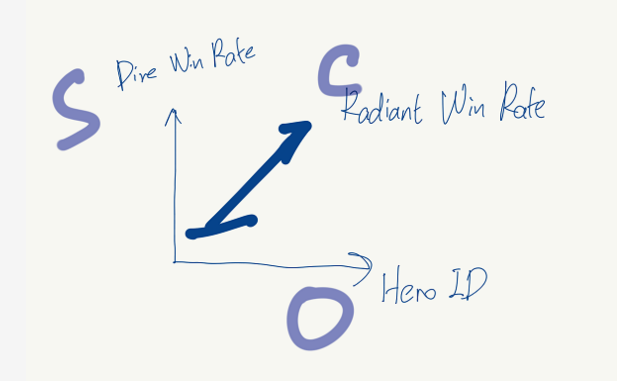


Fig. 9. Symbolic representations of Fig.7

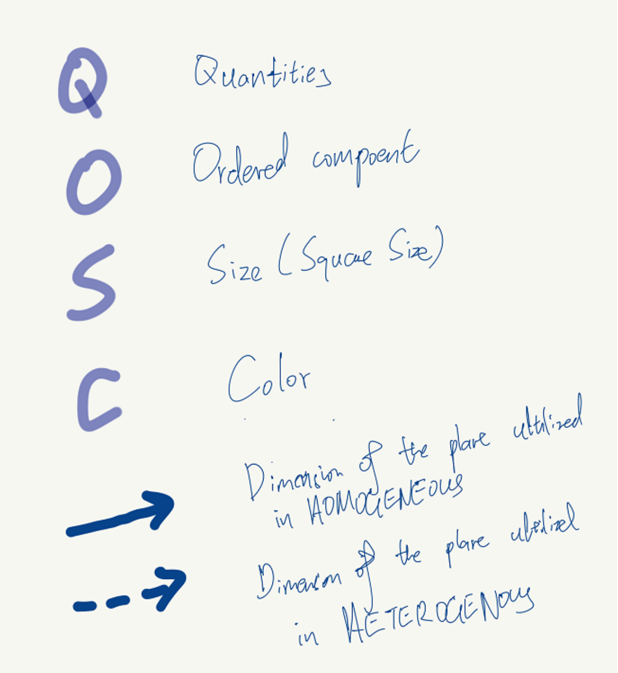

Q
Quantities
O
Ordered compoent
S
Size (Squae Size)
C
Color
Dimencion of the plane utilized in HOMOGENEOUS
Dimension of the plane utilized in HETEROGENOUS

Fig. 10. Symbolic meaning

C. Alternative symbolic representations

The following two hand-drawn sketches Figure 11 and 12 represent alternative symbolic expressions about Figure 6 and Figure 7. Comments on symbols are in Figure 10. For Figure 11, the change we made was to use Square Size to represent the size of sum of the duration, and then change the time of the start time from hourly to daily, and mark it on each block, with time period layered with Color to represent the same color sooner or later and before. For Figure 12, the change we made was to change the Dire Win Rate representation method to Square Size to Quantity, and then to change the Radiant Win Rate from Color to Quantity, so that we could visually see the winning number, and then put the hero id side as Color and add a numeric tag. The reason for this is that only numbers sometimes overlap and it's not easy to tell which hero is the one, and adding colors can be very distinguishable depending on the size of the numbers.

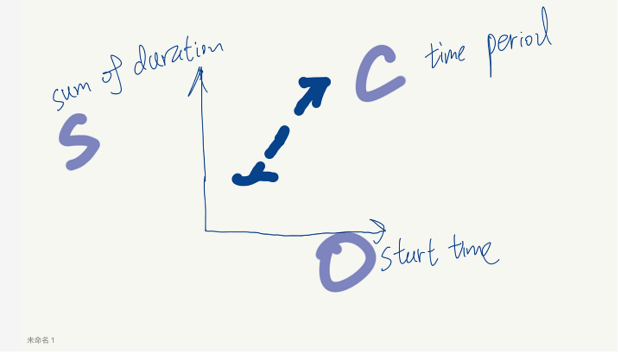


Fig. 11. Alternative symbolic representations of Fig.6

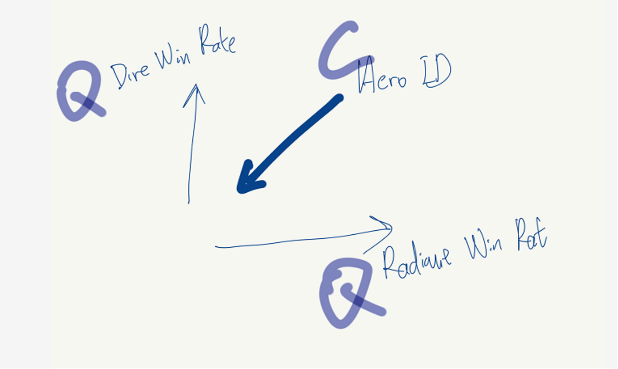


Fig. 12. Alternative symbolic representations of Fig.7

D. Example of alternative visualization

Figure 13 and Figure 14 are examples drawn with alternative representations, respectively. First look at Figure 13, we have a new discovery, the 15th day whether day, night or midnight game total length is greater than other days, indicating that the number of players online this day, estimated that the game has been updated or added to the new game caused. Overall, the day is slightly larger than the night block, as if it were different from our previous conclusion that the largest number of people are online at night, in fact, because we changed the timescale from hour to full day, and the day ratio is greater than the night so that the total length of the day exceeds the night.

For Figure 14, it is now clear that each influence in Radant and Dire's specific winning rate value, this figure we ignore Hero id is 0 hero, because this hero in Radiant or Dire have a terrible low winning rate, we look at the data to know that id 0 represents the player left at that time, so the winning rate will be so low. This chart clearly shows the strength of the hero, if you want to win the game, choose the hero in the upper right corner first, and you can adjust the hero selection based on being in Radiant or Dire.

V. Conclusion

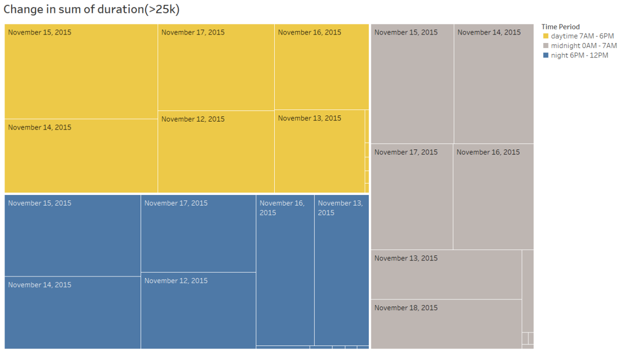


Fig. 13. Change in sum of duration per day

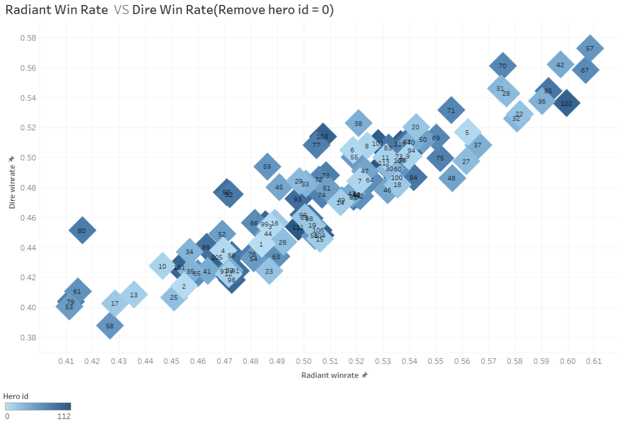


Fig. 14. Radiant Win Rate VS Dire Win Rate(Remove hero id = 0)

Through classroom learning, sometimes can not really understand how to do a good visualization, only through practice can we better train the use of tools, the ability to analyze data. At present, I am not familiar with the visualization tools, but I have just come into contact with the symbolic graphics of the content, but you can see the power of these methods, you can show the data is both beautiful and clear. These charts have room for further optimization as you learn more about the properties and symbolic graphical representations of visualization tools.

References

[1] Marelić, Marko, and Dino Vukušić. "E-sports: Definition and social implications." EQOL J 11.2 (2019): 47-54.

[2] Global eSports market revenue 2024 — Statista. (2021). From https://www.statista.com/statistics/490522/global-esports-market-revenue/

[3] DOTA2 Prize Pool Tracker. From https://dota2.prizetrac.kr/

[4] Raiman, Jonathan, Susan Zhang, and Filip Wolski. "Long-term planning and situational awareness in OpenAI five." arXiv preprint arXiv:1912.06721 (2019).

[5] Zhang, Lei, et al. "Improved Dota2 lineup recommendation model based on a bidirectional LSTM." Tsinghua Science and Technology 25.6 (2020): 712-720.

[6] Dota 2 Matches Dataset. From https://www.kaggle.com/jraramirez/dota-2-matches-dataset

[7] Dota 2 Matches Dataset. From https://www.kaggle.com/devinanzelmo/dota-2-matches

[8] OpenDota. From https://www.opendota.com/

[9] WebAPI/GetMatchDetails - Official TF2 Wiki — Official Team Fortress Wiki. From https://wiki.teamfortress.com/wiki/

[10] Kinkade, Nicholas, L. Jolla, and K. Lim. ”Dota 2 win prediction.” Univ Calif 1 (2015): 1-13.

[11] Exploring Dota2 Match Data. From https://rstudio-pubs-static.s3.amazonaws.com/246035_7d6a00ee44d745cd8428e8731fb1ab5a.html