DOTA 2 PROJECT REPORT
Nov.29, 2019

Tao ChengV00819010

Sijia Chen V00856044

Table of Contents

1. Introduction

The topic we chose is Dota 2 game. Dota 2 is a multiplayer online battle arena (MOBA) video game developed and published by Valve. The game is a sequel to Defense of the Ancients (DotA), which was a community-created mod for Blizzard Entertainment's Warcraft III: Reign of Chaos and its expansion pack, The Frozen Throne. Dota 2 is played in matches between two teams of five players, with each team occupying and defending their own separate base on the map. Each of the ten players independently controls a powerful character, known as a "hero", who all have unique abilities and differing styles of play. During a match, players collect experience points and items for their heroes to successfully defeat the opposing team's heroes in player versus player combat. A team wins by being the first to destroy the other team's "Ancient", a large structure located within their base.[1] It is important to remember that no two Dota games are the same, and teams’ strategies between games can vary wildly in regards to things such as lane composition (trilanes, dual lanes, jungling, etc.), focus (pushing, ganking, farming), and known opponent strengths and weaknesses. While terms regarding the lane positioning (off-laner, solo mid, trilane support, etc.) and hero roles (hard carry, semi-carry, roaming ganker, etc.) are frequently used when discussing individual games or laning setups, these descriptions can be lacking when used more generally.

There are two factions called “Radiant” and “Dire”. Our mission is to discover the win rate for these two different factions. Rach faction contains 5 different positions.

The most common system of naming positions in competitive Dota consists of a numerical scale from 1 through 5, each number representing a player’s farm priority, with the 1-player having the highest farm priority and the 5-player having the lowest. Below is a detailed description of each position, with example heroes, notable professional players for that position, and the ideal player characteristics needed for that position. By the way, the positions are same for “Radiant” and “Dire”.

POSITION 1 (AKA THE HARD CARRY)

As mentioned above, position 1 receives the highest farm priority on a team. This position is most often found in a tri-lane in competitive games, with two allied heroes nearby to support and ensure the 1 position player's early game farm. Due to the frequency of tri-lanes, heroes in this position tend to be more gold dependent than level dependent. Position 1 is almost exclusively filled by hard carry heroes, with the rare exception being very aggressive pushing strategies. This player's position is to survive the early game while securing as much farm as possible, and then to make smart decisions in the late game to survive and dominate engagements, as their team's success frequently depends on the 1 staying alive and controlling the flow of the game.

Characteristics: Ability to eke out early farm in any situation, a cool head and good game sense to not die unnecessarily, refrains from joining in team fights early unless absolutely needed.[2]

POSITION 2 (AKA THE SOLO, GANKER, OR SEMI-CARRY)

Position 2 is one of the more versatile positions in terms of role played for the team, but the laning setup is the most static: solo mid. Heroes in this position tend to be equally gold and level dependent, and are chosen for their mobility, their ability to excel in 1v1 situations, and their ability to strongly impact other lanes through ganking or quick scaling into the mid-game. This player's primary role is to outlane the opposing team's solo mid, giving their team an advantage in the mid game through mobile ganking and rune control, or through early solo farm to produce a strong mid-game hero. Heroes at this position tend to be strong and relevant throughout the entire duration of the game.

Characteristics: Ability to dominate a 1v1 laning situation, aggressiveness to gank sidelanes successfully, strong map awareness to avoid incoming enemy ganks from the sidelanes.[2]

POSITION 3 (AKA THE OFFLANER OR SUICIDE SOLO)

Position 3 is played in two primary ways. The first, commonly referred to as the suicide solo, is to go solo in the hard lane against an enemy's defensive trilane. As the name indicates, going up 1 versus 3 in the hard lane is a tough proposition, so the primary focus in this situation is simply to survive and absorb experience and farm when possible. Not dying is one of the primary roles of this position in the early game, and heroes in this role tend to be more level dependent than gold dependent (generally only remaining in the lane until their ultimate or a certain level is acquired, at which point they transition into more of a roaming ganker or initiator). Heroes in this role almost without excpetion have a strong set of escape or tank abilities, allowing them to survive the tough environment in which they're placed (though some exceptions such as Lone Druid and Nature's Prophet can instead use their summoned creeps and jungling capabilities to farm well despite being in a tough situation).

The second, commonly referred to as a farming offlane, is to go solo in the safe lane against the opposing team's hard lane solo while an aggressive tri-lane is employed. In this situation, the heroes played tend to fit the role of semi-carry and benefit from good farm, and the position 3 role becomes much more similar to the position 2 role (in fact, when aggressive tri-lanes are used, it's not uncommon for a team's position 2 player to take over the offlane farming position while the position 3 player takes over the solo mid role).
Characteristics: Ability to survive and soak up experience regardless of the laning situation, great game sense to know when opposing teams are going for a kill, aggression to capitalize on opportunities presented in lane and elsewhere on the map.[2]

POSITION 4 (AKA THE ROAMING SUPPORT OR JUNGLER)

Position 4 is generally played either directly in the tri-lane, or nearby in the jungle, poking their head into lane occasionally to provide ganks. This role is also generally in charge of pulling jungle creeps when possible to help maintain lane control in the safe lane for the carry. These heroes tend to gain a bit more farm than their position 5 support counterparts, and generally will build towards mid-game support items (Drums, Mekansm, Pipe) while the position 5 support tends to spend their little bit of gold on consumables such as wards and pooled regeneration items. Players in position 4 generally have more of a hand in early game kills (either in their own lane or mid) than any other role, and frequently are the key heroes involved in putting early pressure on enemy towers.
Characteristics: Ability to get by with less farm than other heroes, strong understanding of the jungle, good sense of the opposing team's vision and positioning, (frequently) good micro skills.[2]

POSITION 5 (AKA THE HARD SUPPORT OR BABYSITTER)

Position 5, as indicated by the number, is the least farm-dependent role on the

team. Heroes in this role are largely item and level independent, and it’s not uncommon to see a position 5 hero with only boots and a few branches 3-4 levels behind other heroes well into the game. Not to be underestimated, the position 5 player is often the backbone of their team, pooling their resources by buying consumables for the carries to use and maintaining map control through warding and counterwarding. In addition, the position 4 and 5 heroes have the strongest impact out of all roles in the early phase of the game. On top of spending their gold on team-wide consumables, position 5 heroes are also invaluable for protecting the position 1 player in the early phases of the game when the 1 position is vulnerable, and are frequently relied on for key disables to begin ganks or save teammates.
Characteristics: Strong decision-making in the early game, ability to get by with next to no items, ability to produce enough gold to buy wards and other consumables without taking away from other player’s farm, strong map awareness and knowledge to properly ward and de-ward.[2]

It is important to stress that these positions and roles are not set in stone, as the Dota metagame is very fluid. Many players regularly play multiple roles on their team and rotate based on what the situation requires. In addition, these heroes are merely suggestions, and many heroes are capable of filling many different roles, and can even change roles as a game is in progress. However, these general guidelines are how professional teams tend to draft and fill their rosters. Knowing these positions can help you not only to better understand professional Dota, but also to understand how teams and heroes in public games are supposed to work together as a cohesive unit, as opposed to five individual players all trying to obtain their own farm.

2. Project Goal

Our goal is to explore the win rate of “Radiant” and “Dire”, which faction has a higher win rate. Do different factions affect the balance of one game?

2.1 Project Domain

The datasets are huge because the game is very complex and unique. Thus, we use the data from previous matches found on kaggle to analyze if the players from one of the two teams all join the team fright, his team will have a greater probability of winning the entire game or not.

2.2 Tools and datasets

1. Dataset

   The dataset we use contains players’ statistics for over 23086 matches. It provides different players’ personal gameplay data in different games, which includes the farming, killing, assisting, and deaths.

2. Tools

   Jupyter notebook python 3

3. Classifier

   Naïve Bayes Classifier
   Logistic Regression
   Linear Regression

2.3 Data Information

The dataset we used is from Devin Anzelmo’s post, namely Dota 2 Matches on Kaggle, it is used about Dota 2 matches data for 23 different csv fies.

2.4 Type of attributes:

-duration
-tower_status_radiant
-tower_status_dire
-barracks_status_radiant
-barracks_status_dire

-postitive_votes
-negative_votes
-radiant_score
-dire_score

2.5 Problems we faced

The first problem we occurred is that the dataset is very huge and it made us hard to find the useful data we needed. In addition, the datasets are different from each other because they have different format for players' ids or other attributes. There are some missing data for some matches and it could provide a false conclusion and make us harder to predict. We sifted through a lot datasets before finding the data we could use.

3. Data Processing

Data Processing is the most important step in our project. Data processing is to find out the data that are useful for our research. And we will follow the steps:

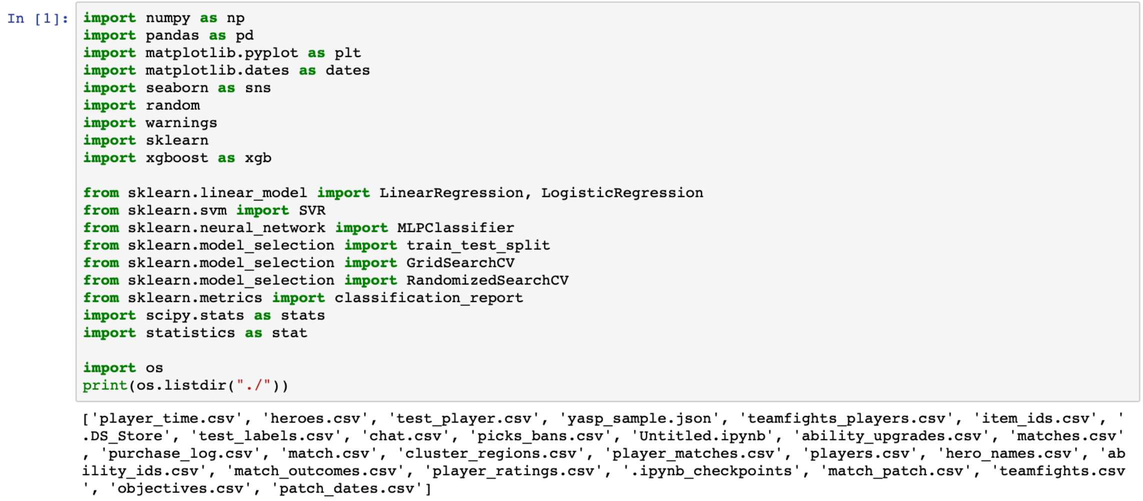
1. Comparing classifiers
2. Analyzing raw data
3. Data reduction and selection

3.1 Comparing classifiers

According to the study from the lectures and labs, we learned linear regression and logistic regression. In the process of data mining, we used these classifiers to analyze the dataset and then compare these results to see the advantage. For this project, Jupyter Notebook and Python3 are used to process the data.

3.2 Analyzing raw data

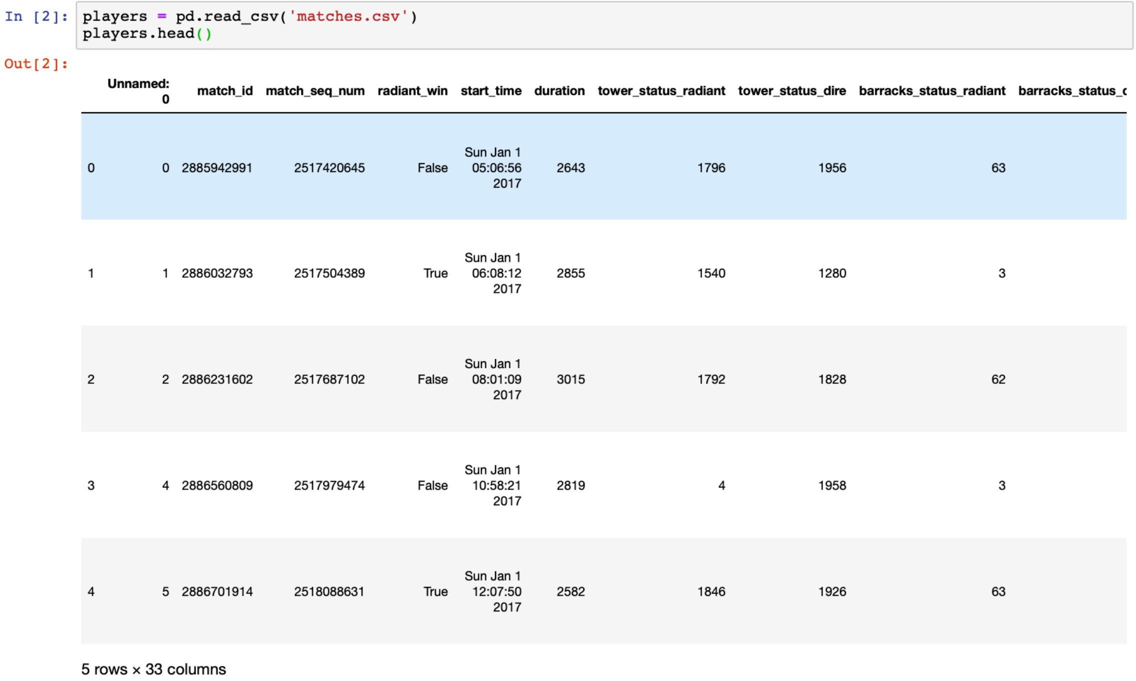
The original dataset comes from the statistics for approximately 23000 of the dota matches that provided by the Kaggle websites.



```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import matplotlib.dates as dates
        import seaborn as sns
        import random
        import warnings
        import sklearn
        import xgboost as xgb

        from sklearn.linear_model import LinearRegression, LogisticRegression
        from sklearn.svm import SVR
        from sklearn.neural_network import MLPClassifier
        from sklearn.model_selection import train_test_split
        from sklearn.model_selection import GridSearchCV
        from sklearn.model_selection import RandomizedSearchCV
        from sklearn.metrics import classification_report
        import scipy.stats as stats
        import statistics as stat

        import os
        print(os.listdir("./"))

        ['player_time.csv', 'heroes.csv', 'test_player.csv', 'yasp_sample.json', 'teamfights_players.csv', 'item_ids.csv', '
        .DS_Store', 'test_labels.csv', 'chat.csv', 'picks_bans.csv', 'Untitled.ipynb', 'ability_upgrades.csv', 'matches.csv'
        , 'purchase_log.csv', 'match.csv', 'cluster_regions.csv', 'player_matches.csv', 'players.csv', 'hero_names.csv', 'ab
        ility_ids.csv', 'match_outcomes.csv', 'player_ratings.csv', '.ipynb_checkpoints', 'match_patch.csv', 'teamfights.csv
        ', 'objectives.csv', 'patch_dates.csv']
```

3.3 Data Reduction

As we found, although there are tons of data we have, we think only five datasets can influence the result directly. Thus, we try to make them become the top five important datasets for mining in the next stage.

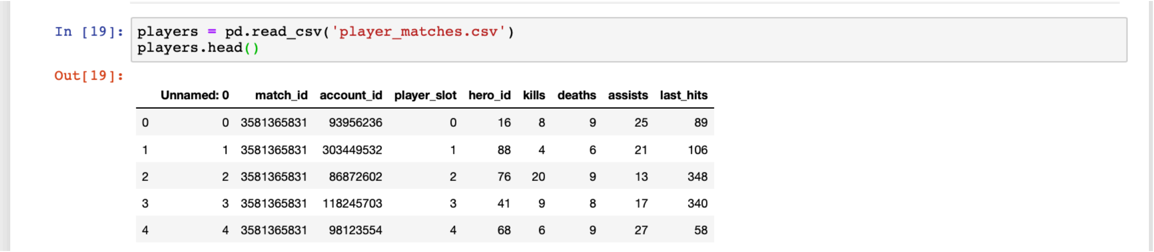


```
In [2]: players = pd.read_csv('matches.csv')
        players.head()
```

Out[2]:

	Unnamed: 0	match_id	match_seq_num	radiant_win	start_time	duration	tower_status_radiant	tower_status_dire	barracks_status_radiant	barracks_status_c
0	0	2885942991	2517420645	False	Sun Jan 1 05:06:56 2017	2643	1796	1956	63	
1	1	2886032793	2517504389	True	Sun Jan 1 06:08:12 2017	2855	1540	1280	3	
2	2	2886231602	2517687102	False	Sun Jan 1 08:01:09 2017	3015	1792	1828	62	
3	4	2886560809	2517979474	False	Sun Jan 1 10:58:21 2017	2819	4	1958	3	
4	5	2886701914	2518088631	True	Sun Jan 1 12:07:50 2017	2582	1846	1926	63	

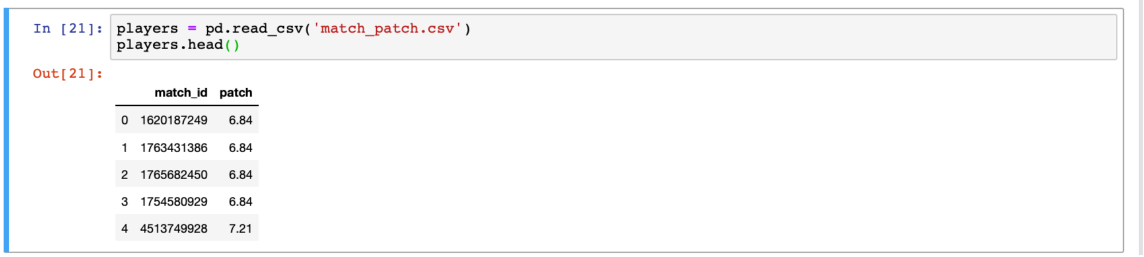
5 rows × 33 columns



```
In [19]: players = pd.read_csv('player_matches.csv')
         players.head()
```

Out[19]:

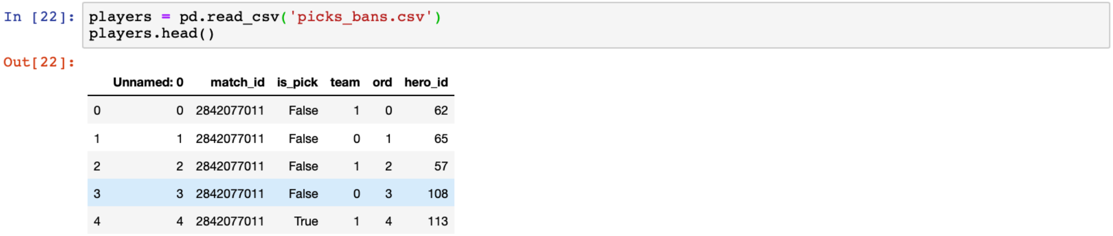
	Unnamed: 0	match_id	account_id	player_slot	hero_id	kills	deaths	assists	last_hits
0	0	3581365831	93956236	0	16	8	9	25	89
1	1	3581365831	303449532	1	88	4	6	21	106
2	2	3581365831	86872602	2	76	20	9	13	348
3	3	3581365831	118245703	3	41	9	8	17	340
4	4	3581365831	98123554	4	68	6	9	27	58



```
In [21]: players = pd.read_csv('match_patch.csv')
         players.head()
```

Out[21]:

	match_id	patch
0	1620187249	6.84
1	1763431386	6.84
2	1765682450	6.84
3	1754580929	6.84
4	4513749928	7.21



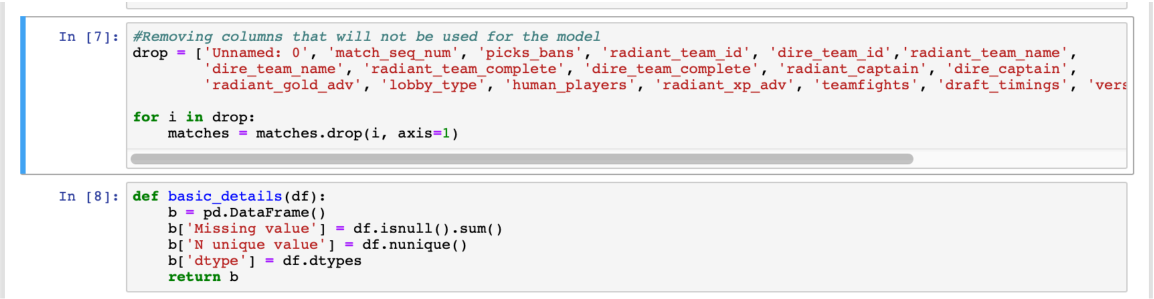
```
In [22]: players = pd.read_csv('picks_bans.csv')
         players.head()
```

Out[22]:

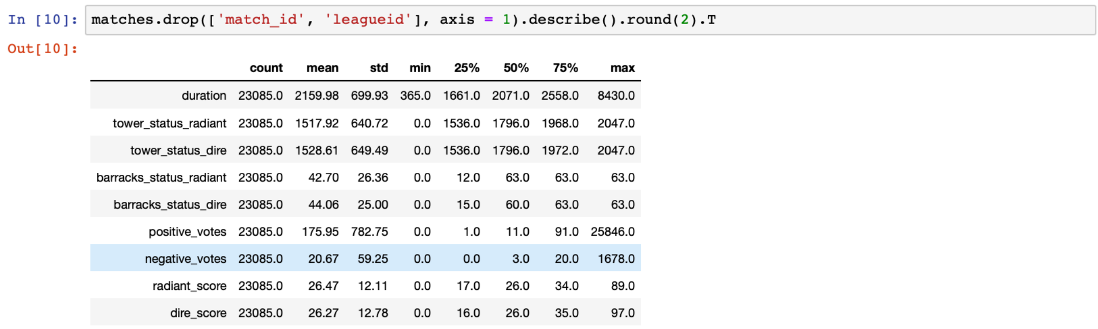
	Unnamed: 0	match_id	is_pick	team	ord	hero_id
0	0	2842077011	False	1	0	62
1	1	2842077011	False	0	1	65
2	2	2842077011	False	1	2	57
3	3	2842077011	False	0	3	108
4	4	2842077011	True	1	4	113

4. Data Mining

As what we planed from Data Processing section, we removed columns that will not be used for the model. And we found the useful attributes namely: duration, twoer_status_radiant, tower_status_dire, barracks_status_radiant, barracks_status_dire, positive_votes, negative_votes, radiant_score, dire_score.



```
In [7]: #Removing columns that will not be used for the model
        drop = ['Unnamed: 0', 'match_seq_num', 'picks_bans', 'radiant_team_id', 'dire_team_id','radiant_team_name',
                'dire_team_name', 'radiant_team_complete', 'dire_team_complete', 'radiant_captain', 'dire_captain',
                'radiant_gold_adv', 'lobby_type', 'human_players', 'radiant_xp_adv', 'teamfights', 'draft_timings', 'vers

        for i in drop:
            matches = matches.drop(i, axis=1)

In [8]: def basic_details(df):
            b = pd.DataFrame()
            b['Missing value'] = df.isnull().sum()
            b['N unique value'] = df.nunique()
            b['dtype'] = df.dtypes
            return b
```

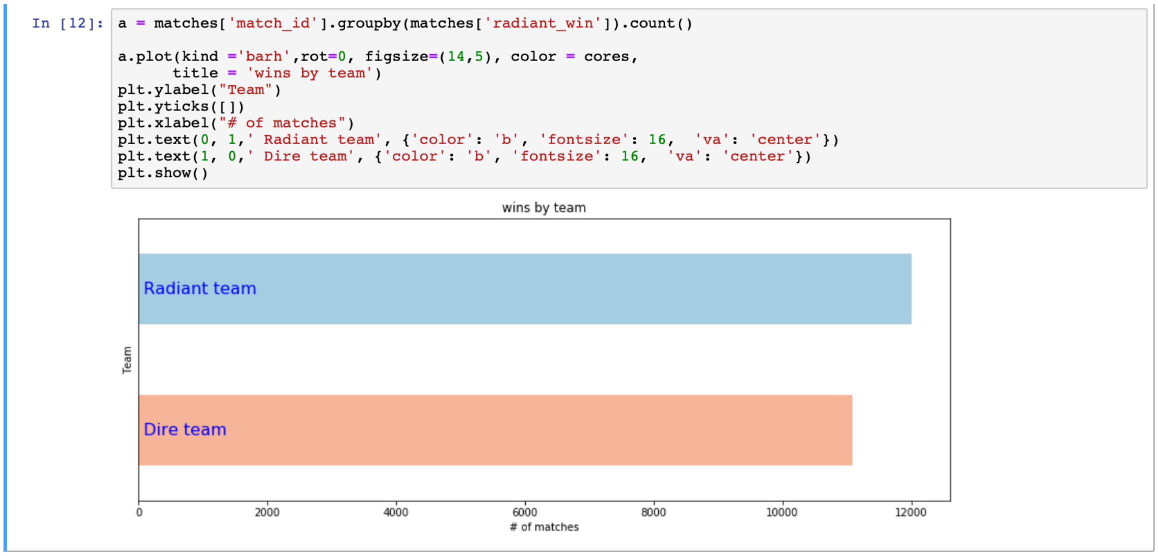


```
In [10]: matches.drop(['match_id', 'leagueid'], axis = 1).describe().round(2).T
Out[10]:
```

	count	mean	std	min	25%	50%	75%	max
duration	23085.0	2159.98	699.93	365.0	1661.0	2071.0	2558.0	8430.0
tower_status_radiant	23085.0	1517.92	640.72	0.0	1536.0	1796.0	1968.0	2047.0
tower_status_dire	23085.0	1528.61	649.49	0.0	1536.0	1796.0	1972.0	2047.0
barracks_status_radiant	23085.0	42.70	26.36	0.0	12.0	63.0	63.0	63.0
barracks_status_dire	23085.0	44.06	25.00	0.0	15.0	60.0	63.0	63.0
positive_votes	23085.0	175.95	782.75	0.0	1.0	11.0	91.0	25846.0
negative_votes	23085.0	20.67	59.25	0.0	0.0	3.0	20.0	1678.0
radiant_score	23085.0	26.47	12.11	0.0	17.0	26.0	34.0	89.0
dire_score	23085.0	26.27	12.78	0.0	16.0	26.0	35.0	97.0

4.1 Bar chart of wins by team

Just like we mentioned in the Data Processing section, the whole datasets that contained matches information is very huge and we are only interested in the top five important datasets. So we extracted the dataset and then print out them in the form of the bar chart. Clearly, team radiant has a high win-rate than team dire.

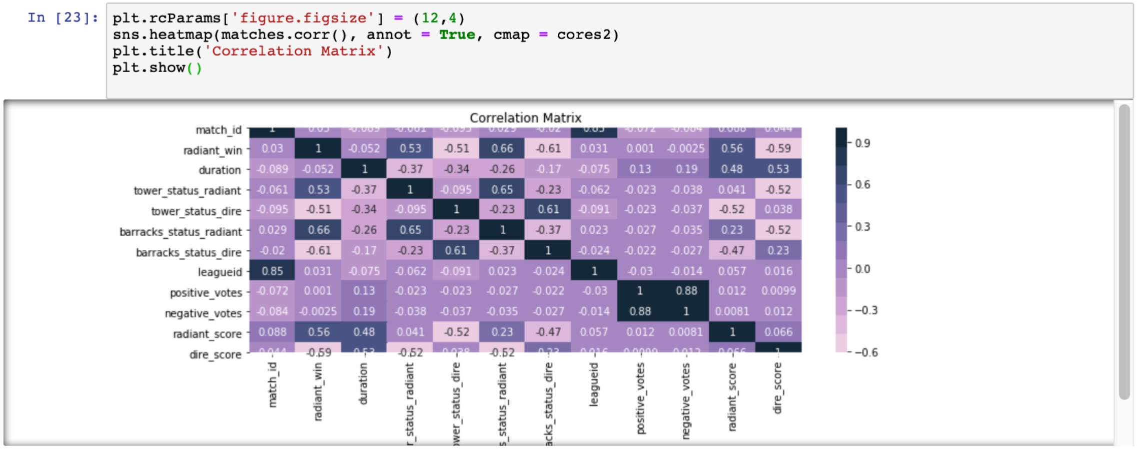


```
In [12]: a = matches['match_id'].groupby(matches['radiant_win']).count()

         a.plot(kind ='barh',rot=0, figsize=(14,5), color = cores,
               title = 'wins by team')
         plt.ylabel("Team")
         plt.yticks([])
         plt.xlabel("# of matches")
         plt.text(0, 1,' Radiant team', {'color': 'b', 'fontsize': 16,  'va': 'center'})
         plt.text(1, 0,' Dire team', {'color': 'b', 'fontsize': 16,  'va': 'center'})
         plt.show()
```

Figure 1

4.2 Correlation matrix

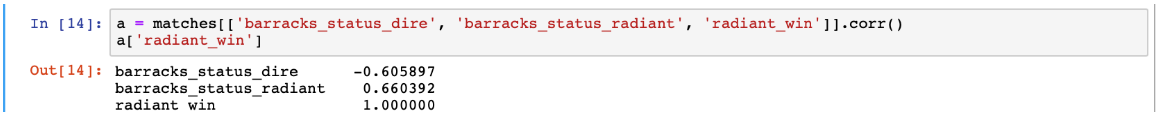
This is a good way to show the relationship between different attributes to the match result.



```
In [23]: plt.rcParams['figure.figsize'] = (12,4)
         sns.heatmap(matches.corr(), annot = True, cmap = cores2)
         plt.title('Correlation Matrix')
         plt.show()
```

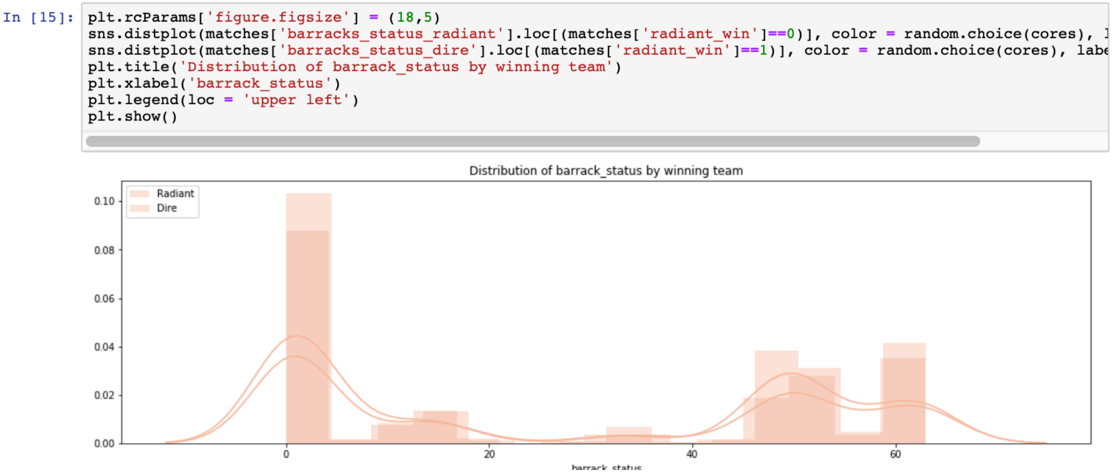
Figure 2

And the result just show that ratio of radiant wins is more than the ratio of dire wins



```
In [14]: a = matches[['barracks_status_dire', 'barracks_status_radiant', 'radiant_win']].corr()
         a['radiant_win']

Out[14]: barracks_status_dire       -0.605897
         barracks_status_radiant     0.660392
         radiant_win                 1.000000
```

4.3 Bar chart and Line chart of barrack_status



```
In [15]: plt.rcParams['figure.figsize'] = (18,5)
         sns.distplot(matches['barracks_status_radiant'].loc[(matches['radiant_win']==0)], color = random.choice(cores), l
         sns.distplot(matches['barracks_status_dire'].loc[(matches['radiant_win']==1)], color = random.choice(cores), labe
         plt.title('Distribution of barrack_status by winning team')
         plt.xlabel('barrack_status')
         plt.legend(loc = 'upper left')
         plt.show()
```

Figure 3

4.4 Scatter diagram of score

```
In [16]: sns.pairplot(x_vars='radiant_score', y_vars='dire_score', hue='radiant_win', data=matches, size=7, palette= cores
```

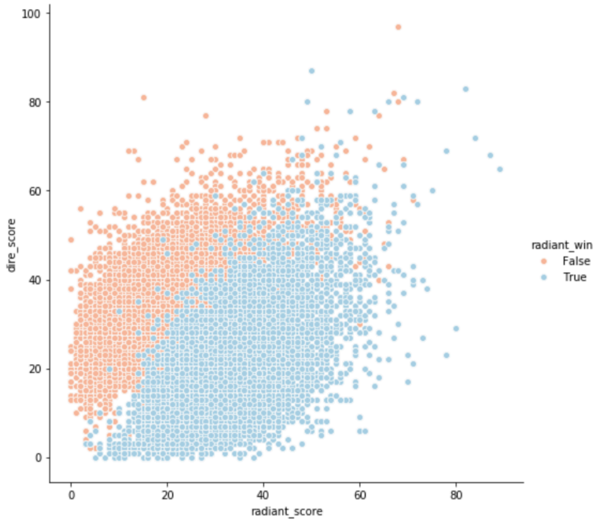
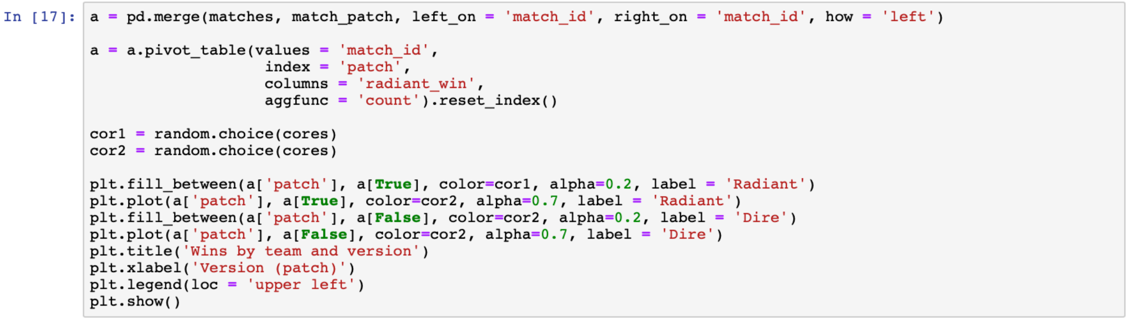


Figure 4

From the diagram, we still found that the range of radiant win is a little bit larger than dire win.

4.5 Area graph of version



```
In [17]: a = pd.merge(matches, match_patch, left_on = 'match_id', right_on = 'match_id', how = 'left')

         a = a.pivot_table(values = 'match_id',
                           index = 'patch',
                           columns = 'radiant_win',
                           aggfunc = 'count').reset_index()

         cor1 = random.choice(cores)
         cor2 = random.choice(cores)

         plt.fill_between(a['patch'], a[True], color=cor1, alpha=0.2, label = 'Radiant')
         plt.plot(a['patch'], a[True], color=cor2, alpha=0.7, label = 'Radiant')
         plt.fill_between(a['patch'], a[False], color=cor2, alpha=0.2, label = 'Dire')
         plt.plot(a['patch'], a[False], color=cor2, alpha=0.7, label = 'Dire')
         plt.title('Wins by team and version')
         plt.xlabel('Version (patch)')
         plt.legend(loc = 'upper left')
         plt.show()
```

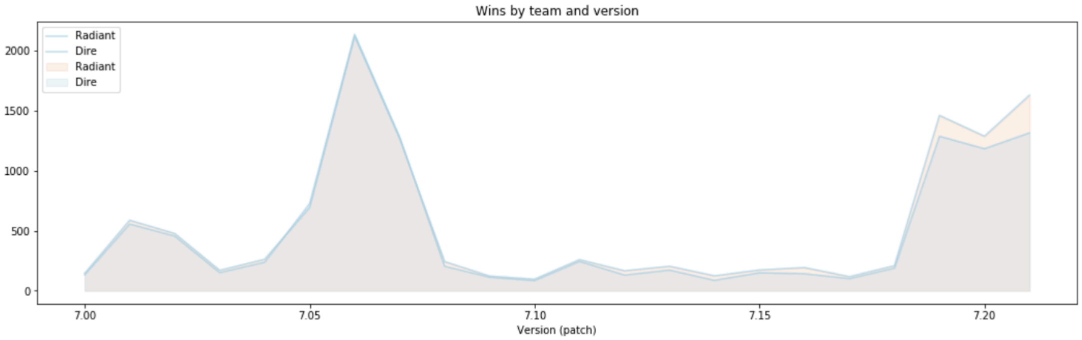


Figure 5

As we can see the area graph, the latest versions show that the ratio of radiant-win is a little bit larger.

5. Conclusion

In conclusion, by using the functions in python and jupyter book, we found that the two different factions “Radiant” and “Dire” indeed affect the outcome of the whole game. The heroes in Dota 2 are related to each other. In fact, the quality of heroes that belongs to “Radiant” or “Dire” can almost reach a flat state. However, “Radiant” breaks this balance a little bit, that is, there are more heroes in “Radiant” faction in the past days, the terrain of the two factions is different, and the boss “Roshan” is close to the “Radiant” faction which gives “Dire” team a better chance to get the buff. Also, the barracks’ quality influence the match result, which is should not happen in such a “fair” game. Overall, by analyzing the datasets, we found that the win rates for two different factions are different.

6. Reference

[1] Dota 2 from Wikipedia.
https://en.wikipedia.org/wiki/Dota_2

[2] DOTA 2: A GUIDE TO COMPETITIVE POSITIONS, Submission, 2013-05-17,
https://imperium.news/dota-2-guide-competitive-positions/

[3] Dota 2 Matches datasets. Link:
https://www.kaggle.com/devinanzelmo/dota-2-matches