Final Project: Analysis of All Time Top Earning DotA 2 players

Zixin Qian

Course: MSDS Data Wrangling and Husbandry

Instructure: Prof. Jason M. Klusowski

Introduction

DotA 2 is a multiplayer online battle arena (MOBA) vedio game developed from 2009 and offically released. In each game, ten players are evenly splitted into two teams battling each other. Recent years DotA 2 tournments are famous for its high prizes for the winners. Personally I played Dota and DotA 2 a few years ago and I want to analyse the top players' earnings. This project includes webscraping the data from a website named "Liquipedia", cleaning the data and some further analysis as well as interperutations.

Data

All the data are webscraped from Liquipedia. Multiple tables are read from different sections of the website. A total of 13 tables are read from the website: one includes personal information of all time DotA 2 professional players, two are the total earnings of those players (the table could not fit in one page so website divided it in two) and the rest ten tables are the tournment earning details of ten players with the highest earnings today.

Procedure

First is to download the data frames of all players and earnings

Player	list [1]	List of length 1
[[1]]	list [841 x 5] (S3: data.frame)	A data.frame with 841 rows and 5 columns
[[1]]	logical	NA NA NA NA NA ...
ID	character [841]	'.Ark' '13abyKnight' '1437' '290' '2hoi' '33' ...
Name	character [841]	'Egor Zhabotinskii' 'Jon Andersen' 'Sivatheeban Sivanathapillai' 'Zeng Chen' 'Ch ...
Team	character [841]	'' '' 'Tigers' 'Cola' 'XCN Gaming' 'Ninjas in Pyjamas' ...
Links	logical	NA NA NA NA NA ...

The earning table was divided in two due to the insufficient capacity of the webpage.

Player_earnings1	list [1]	List of length 1
[[1]]	list [500 x 8] (S3: data.frame)	A data.frame with 500 rows and 8 columns
[[1]]	integer [500]	1 2 3 4 5 6 ...
[[2]]	logical	NA NA NA NA NA NA ...
ID	character [500]	'KuroKy' 'N0tail' 'Miracle-' 'MinD_ContRoL' 'MATUMBAMAN' 'JerAx' ...
[[4]]	integer [500]	37 21 19 23 21 12 ...
[[5]]	integer [500]	14 16 7 10 14 10 ...
[[6]]	integer [500]	11 13 4 4 4 6 ...
Premier	integer [500]	21 9 8 6 6 4 ...
Earnings	character [500]	'$4,168,314' '$3,743,308' '$3,731,424' '$3,519,295' '$3,506,036' '$3,323,042' ...

Player_earnings2	list [1]	List of length 1
[[1]]	list [409 x 8] (S3: data.frame)	A data.frame with 409 rows and 8 columns
[[1]]	integer [409]	1 2 3 4 5 6 ...
[[2]]	logical	NA NA NA NA NA NA ...
ID	character [409]	'.Ark' 'ling' 'TnK' 'Ben' 'Nix' 'Poloson' ...
[[4]]	integer [409]	4 5 1 7 2 6 ...
[[5]]	integer [409]	3 5 3 3 3 6 ...
[[6]]	integer [409]	2 4 3 1 1 4 ...
Premier	integer [409]	0 0 0 0 0 0 ...
Earnings	character [409]	'$8,319' '$8,283' '$8,277' '$8,274' '$8,268' '$8,210' ...

Then I remove the columns that cannot be read in all three tables, combine the last two tables into a new one, replace the column names that could not be read (which were pictures but able to understand) with words and join the two tables into a new one. Since I want both the players' personal and earning information, I choose to use inner join.

ID <chr>	Name <chr>	Team <chr>	champion <int>	runnerup <int>	secondrunnerup <int>	Premier <int>	Earnings <chr>	numoftop3 <int>
.Ark	Egor Zhabotinskii		4	3	2	0	$8,319	9
.Ark	Egor Zhabotinskii		4	3	2	0	$8,319	9
13abyKnight	Jon Andersen		12	6	4	0	$101,159	22
1437	Sivatheeban Sivanathapillai	Tigers	7	10	4	0	$227,767	21
290	Zeng Chen	Cola	0	0	2	0	$2,869	2
2hoi	Chang Tu Hai	XCN Gaming	0	3	2	0	$1,827	5

However, Earnings are character variables that obviously can not be directly processed.

#remove "$" and "," and change data type into numerical values for Earnings

ID <chr>	Name <chr>	Team <chr>	champion <int>	runnerup <int>	secondrunnerup <int>	Premier <int>	Earnings <dbl>	numoftop3 <int>
.Ark	Egor Zhabotinskii		4	3	2	0	8319	9
.Ark	Egor Zhabotinskii		4	3	2	0	8319	9
13abyKnight	Jon Andersen		12	6	4	0	101159	22
1437	Sivatheeban Sivanathapillai	Tigers	7	10	4	0	227767	21
290	Zeng Chen	Cola	0	0	2	0	2869	2
2hoi	Chang Tu Hai	XCN Gaming	0	3	2	0	1827	5

Then I Analyzed about the names of all professional players and found five most commonly used names.

FirstName <chr>	n <int>
Zhang	12
Chen	10
Alexander	8
Liu	8
Zhou	8

It seems four of the five names are from Chinese players, somehow indicates the large number of Chinese professional DotA 2 players.

Then I want to see which players have the most earnings till now

ID <chr>	Name <chr>	Team <chr>	champion <int>	runnerup <int>	secondrunnerup <int>	Premier <int>	Earnings <dbl>
GH	Maroun Merhej	Team Liquid	15	7	2	5	3124576
JerAx	Jesse Vainikka	OG	12	10	6	4	3323042
KuroKy	Kuro Salehi Takhasomi	Team Liquid	37	14	11	21	4168314
MATUMBAMAN	Lasse Aukusti Urpalainen	Team Liquid	21	14	4	6	3506036
MinD_ContRoL	Ivan Borislavov Ivanov	Team Liquid	23	10	4	6	3519295
Miracle-	Amer Al-Barkawi	Team Liquid	19	7	4	8	3731424
N0tail	Johan Sundstein	OG	21	16	13	9	3743308
ppd	Peter Dager	Ninjas in Pyjamas	18	13	7	9	2891201
SumaiL	Syed Sumail Hassan	Evil Geniuses	11	10	10	7	3313043
UNiVeRsE	Saahil Arora		20	20	9	12	3042820

We can see that the players that have the top 10 earnings are: GH, JerAx, KuroKy, MATUMBAMAN, Mind_ContRoL, Miraclle-, N0tail, ppd, SumaiL, UniVeRsE

I also created a word cloud of them

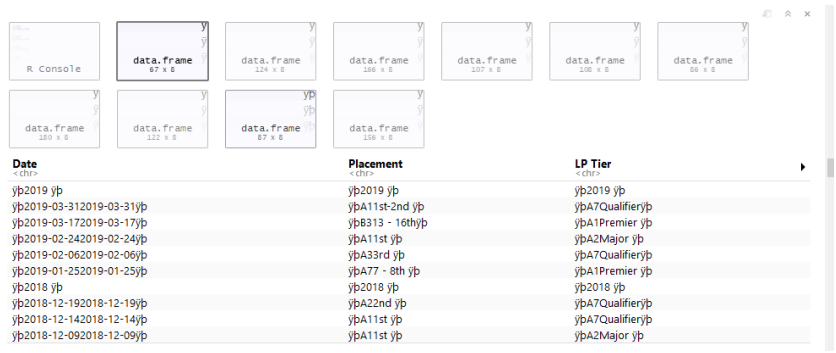
Note that sometimes one or two players' names could not fit on page, I tried to plot with a larger device but not working.

Then I run a linear regression of Team on Earnings. Coefficients of many teams are not significant but most teams has significant coefficients are good teams that tend to have higher earnings.

Note that the result only calculates the current team players' total earnings. i.e. It is does not indicate how much each team earns

Now I want to find the details of top earning players' tournment history. Webscrape them from the website. Since webscraping them one by one would be quite time consuming, I wrote a function called "ReadWeb" to webscrape those data and save it under the players' names.

```
ReadWeb <- function(web){
  a <- web %>%
    read_html() %>%
    html_nodes("table") %>%
    html_table(fill = TRUE)
  return(a)
}
TopearningPlayers <- list(GH,JerAx,KuroKy,MATUMBAMAN,Mind_ContRoL,Miraclle,N0tail,ppd,SumaiL,UNiVeRsE)
TopearningPlayers <- TopearningPlayers %>%
  map(ReadWeb)
```



Now we get a list of 10 data frames that has player earnings details, do some cleaning

Combine the 10 data frames into a new tibble

Each player's data on the website is categorized by year and the titles that indicates the year of each part were also read, remove those years.

Again, convert the variable type of earnings from character into numerics

remove the unreadable and useless variables

```
> Topearning
# A tibble: 1,130 x 7
   Date       Placement   `LP Tier`  Tournament                                Results..6 Prize Names
   <date>     <chr>       <chr>      <chr>                                     <chr>      <dbl> <chr>
 1 2019-03-17 B313 - 16th A1Premier  DreamLeague Season 11                     0 : 1      10000 UNiVeRsE
 2 2019-02-21 A99 - 10th  A2Major    ESL One Katowice 2019                     0/3/2       5000 UNiVeRsE
 3 2019-02-06 A33rd       A7Qualifier DreamLeague Season 11 North America Qualifier 2 : 0        0 UNiVeRsE
 4 2019-01-22 B313 - 16th A1Premier  The Chongqing Major                       0 : 1      10000 UNiVeRsE
 5 2019-01-04 A77 - 8th   A3Minor    LOOT.BET Winter Masters                   0 : 1          0 UNiVeRsE
 6 2019-01-03 A55 - 8th   A3Minor    WePlay! Dota 2 Winter Madness             1 : 2       2500 UNiVeRsE
 7 2018-12-08 A55 - 6th   A2Major    MegaFon Winter Clash                      1 : 2      12510 UNiVeRsE
 8 2018-11-30 A22nd       A7Qualifier The Chongqing Major North America Qualifier 2 : 0        0 UNiVeRsE
 9 2018-11-13 A99 - 12th  A1Premier  The Kuala Lumpur Major                    1 : 2      15000 UNiVeRsE
10 2018-10-26 A77 - 8th   A2Major    ESL One Hamburg 2018                      0 : 2       7500 UNiVeRsE
# ... with 1,120 more rows
```

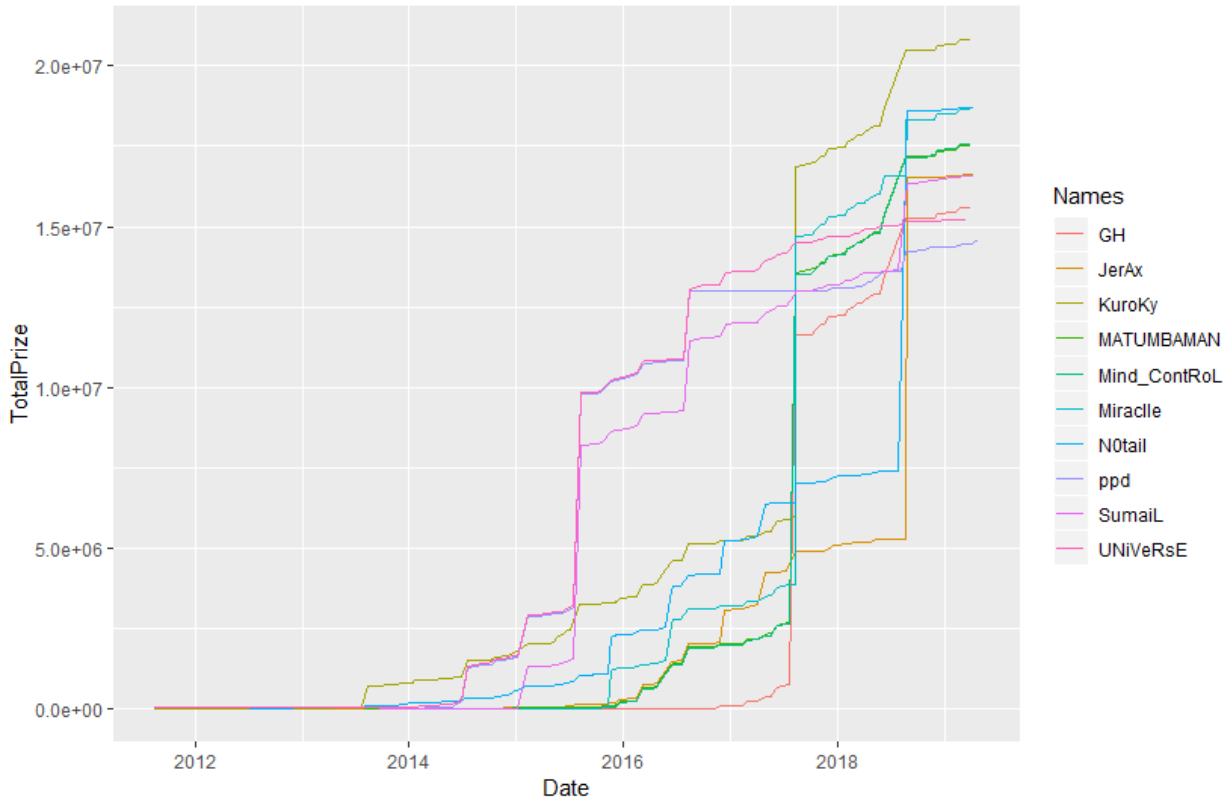
Get the new tibble.

Then add a new variable called TotalPrize that cumulates the total earnings of each player.

```
> Topearning.final
# A tibble: 1,130 x 8
   Date       Placement   `LP Tier`   Tournament                                      Results..6 Prize Names TotalPrize
   <date>     <chr>       <chr>       <chr>                                           <chr>      <dbl> <chr>       <dbl>
 1 2014-11-01 A55 - 8th   A2Major     MSI Beat IT 2014                                0/-/2          0 GH              0
 2 2015-06-28 A11st       A6Weekly    The Impress Dad Indy 500                        2 : 0        500 GH            500
 3 2015-07-20 A11st       A6Weekly    The Impress Dad Andy 500 #3                     2 : 1        500 GH           1000
 4 2015-09-23 A33 - 4th   A3Minor     PGL Dota 2 Pro-AM Qualifier #1                  0 : 2        500 GH           1500
 5 2015-11-11 A22nd       A5Monthly   paysafecard Go4Dota 2 October Finals            0 : 2        161 GH           1661
 6 2016-01-19 A33 - 4th   A7Qualifier ProDotA Cup EU #2 Open Qualifier                0 : 2          0 GH           1661
 7 2016-04-12 B313 - 16th A3Minor     ProDotA Cup Europe #5                           0 : 1          0 GH           1661
 8 2016-06-24 A22nd       A7Qualifier The International 2016: European Open Qualifier #2 0 : 2        0 GH           1661
 9 2016-07-11 A99 - 12th  A3Minor     ProDotA Cup Europe #8                           0 : 1          0 GH           1661
10 2016-07-12 A11st       A7Qualifier WellPlay Invitational #3: Closed Qualifier      2 : 0          0 GH           1661
# ... with 1,120 more rows
```

Save this table as a csv file as my final data set.

Get a plot of the players' total earnings.



From the plot we can see that the cumulative prize of players are mostly around 0 before 2014, that is because from The International (TI) 2013 (the most significant tournment) onward, its prize pool was allowed to be crowdfunded through a type of optional in-game battle pass called the "Compendium", which raises money from players buying them and connected lootboxes to

get exclusive in-game cosmetics and other bonuses offered through them. There are tournments through out every year and there are four major tournments each year while TI is the one with highest awards.

From the plost, we can see ppd and UNiVeRsE have almost identical trend before mid 2016, that is because they were in the same team: Team EG. Also, SumaiL join EG at late 2014 so that these three people share the same trend but Sumail is a little bit below. We can conclude that EG was the best team from 2015 to 2017.

On the other hand, Team Liquid became very competitive from 2016, winning TI 2017 gave the team members huge increase in earnings. Among these ten top earning players today, five of them are from Team Liquid: KuroKy, GH, MATUMBAMAN, Miracle and Mind_Contrl.

Conclusively, Winning one single TI from 2015 would award the players so much that it could take up more than half of the total earnings of some good players maintianing a high level of competitiveness in a long term.

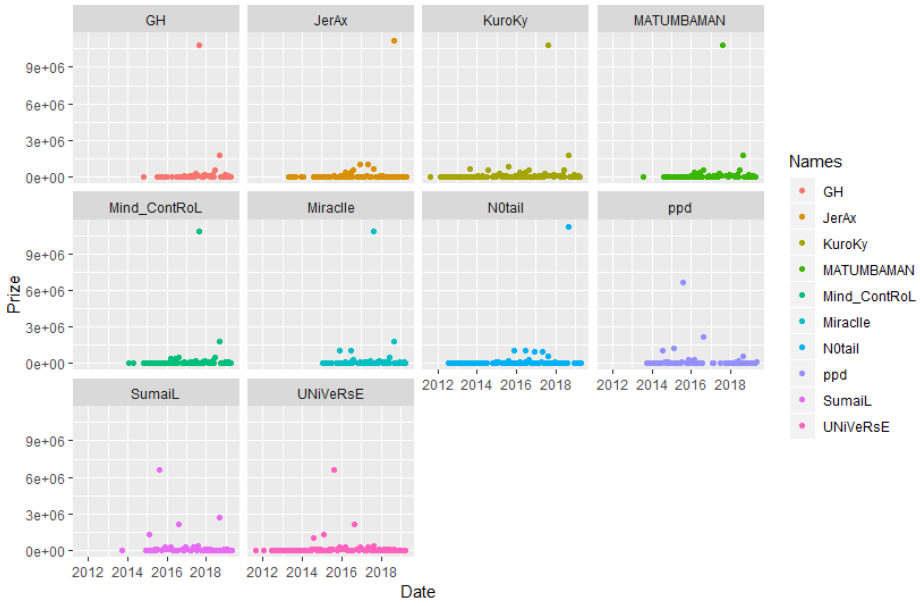
Player GH is a great example: his earning increased from 1 million to 12 millions of dollar just because he won TI 2017 as a team member of Team Liquid. Yet his total earning today is just abour 1560 millions of dollars.

Now let's see which tournments those players earned the most from

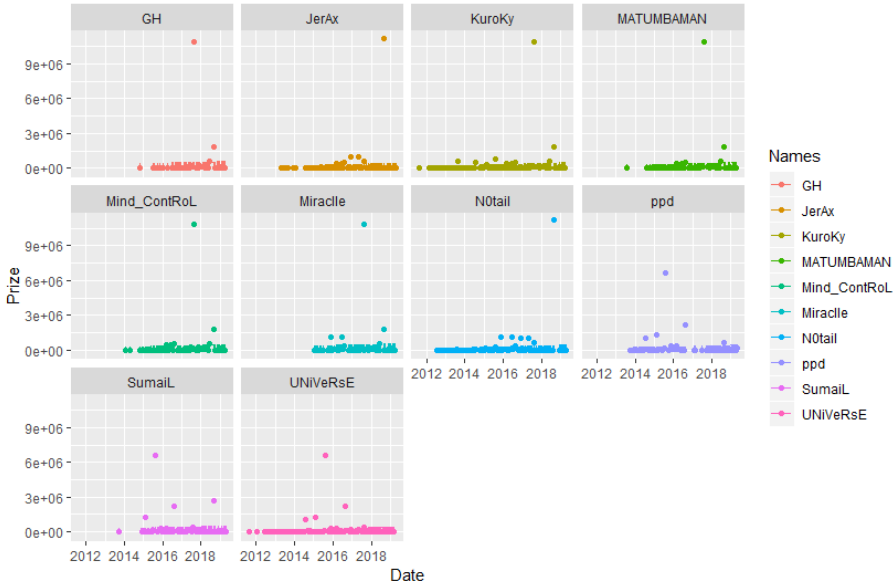
```
# A tibble: 22 x 7
   Date       Placement `LP Tier` Tournament                     Results..6    Prize Names
   <date>     <chr>     <chr>     <chr>                          <chr>         <dbl> <chr>
 1 2016-08-13 A33rd     A1Premier The International 2016         1 : 2       2180898 UNiVeRsE
 2 2015-08-08 A11st     A1Premier The International 2015         3 : 1       6634661 UNiVeRsE
 3 2015-02-09 A11st     A1Premier Dota 2 Asia Championships 2015 3 : 0       1284158 UNiVeRsE
 4 2018-08-25 A33rd     A1Premier The International 2018         0 : 2       2680879 SumaiL
 5 2016-08-13 A33rd     A1Premier The International 2016         1 : 2       2180898 SumaiL
 6 2015-08-08 A11st     A1Premier The International 2015         3 : 1       6634661 SumaiL
 7 2015-02-09 A11st     A1Premier Dota 2 Asia Championships 2015 3 : 0       1284158 SumaiL
 8 2016-08-13 A33rd     A1Premier The International 2016         1 : 2       2180898 ppd
 9 2015-08-08 A11st     A1Premier The International 2015         3 : 1       6634661 ppd
10 2015-02-09 A11st     A1Premier Dota 2 Asia Championships 2015 3 : 0       1284158 ppd
11 2018-08-25 A11st     A1Premier The International 2018         3 : 2      11234158 NOtail
12 2018-08-24 A44th     A1Premier The International 2018         0 : 2       1787252 Miraclle
13 2017-08-12 A11st     A1Premier The International 2017         3 : 0      10862683 Miraclle
14 2018-08-24 A44th     A1Premier The International 2018         0 : 2       1787252 Mind_ContRoL
15 2017-08-12 A11st     A1Premier The International 2017         3 : 0      10862683 Mind_ContRoL
16 2018-08-24 A44th     A1Premier The International 2018         0 : 2       1787252 MATUMBAMAN
17 2017-08-12 A11st     A1Premier The International 2017         3 : 0      10862683 MATUMBAMAN
18 2018-08-24 A44th     A1Premier The International 2018         0 : 2       1787252 KuroKy
19 2017-08-12 A11st     A1Premier The International 2017         3 : 0      10862683 KuroKy
20 2018-08-25 A11st     A1Premier The International 2018         3 : 2      11234158 JerAx
21 2018-08-24 A44th     A1Premier The International 2018         0 : 2       1787252 GH
22 2017-08-12 A11st     A1Premier The International 2017         3 : 0      10862683 GH
```

Other than "Dota 2 Asia Championships 2015", all the top earning tournments are The Internationals from 2015 to 2018. Again this show how much a player could earn by winning a TI chamipionship.

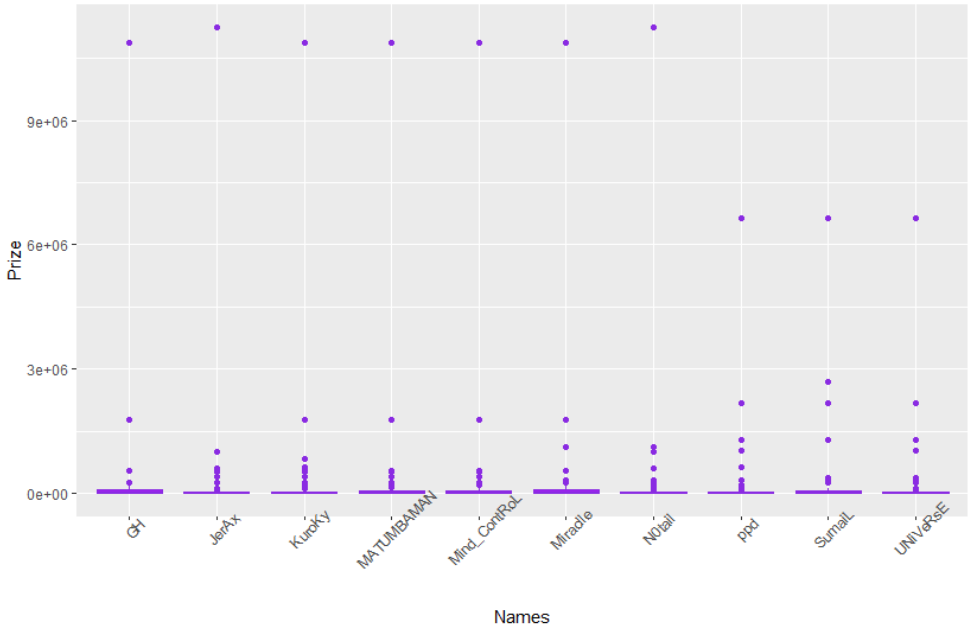
Then I get plots of each players' earnings plot



add the upper and lower confidence bounds

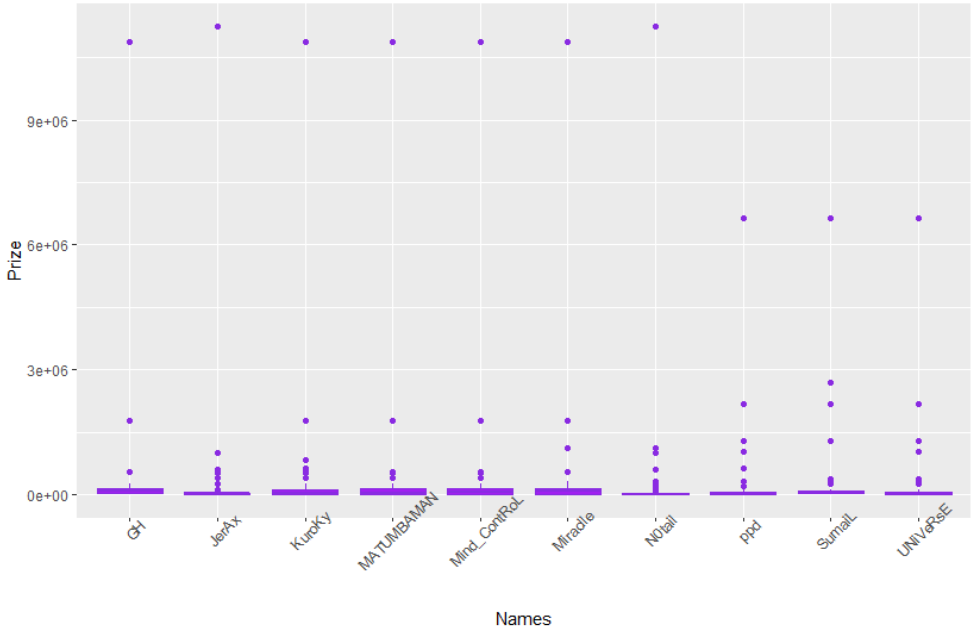


Then I made a boxplot



The boxplot seems extremely strange because most of the tournment earnings are quite low, even for these top players.

remove the tournment earnings below 10000



Results and Disscussion

Wining one single TI would often give a player around 10 millions of dollars earning which is crazy. As long as DotA 2 have sufficient players and consumers, TI would be the most desired championship for all professional players. Players do not get good rankings in tournments may not so happy because the salary of professional players playing DotA 2 is relatively low, especially in contrast with League of Legends (the largest competitor of DotA 2) professional players.

In conclusion, webscraping data from Liquipedia is mostly feasible. However, there are a few difficulties as well:

1. From the player information table, there was a variable indicates the nationality of each player but the nations are represented by a picture of nation flags (also a link) that could not be webscraped. Otherwise I would do some analysis about each nations.
2. The word cloud sometimes automatically removes one or two players since they could not fit on the picture. Trying to plot with a larger device did not work.
3. The last two plots are not very nice because many of the prize earnings are quite low while some major tournments contributes too much to the players' earnings.