Data Wrangling Final Project

Introduction

First I need to make an introduction about the first data I will use.
This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2018. It includes data from the two current leagues (American and National), the four other “major” leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875.
This database was created by Sean Lahman, who pioneered the effort to make baseball statistics freely available to the general public. What started as a one man effort in 1994 has grown tremendously, and now a team of researchers have collected their efforts to make this the largest and most accurate source for baseball statistics available anywhere.
The package contains several main tables:
Master (people): Player names, dates of birth, death and other biographical info.
Batting: Player statistic of act of facing the opposing pitcher and trying to produce offense for one’s team
Pitching: A pitch is the act of throwing a baseball toward home plate to start a play.
Fielding: The performance of each player.
The reason that I choose this dataset because I am interested in baseball and its history. I want to explore the statistic aspect of baseball.

##facts about the People Data First we need to load the master data. I want to see the average weight and height of each year player. From the output we could tell that the average height does not vary a lot troughout years. However, the average weight does vary a lot. Even in consecutive years, for example, 1997 and 1998. The average weight has a difference about 20 pounds. And I also check the player statistic for 2019 MLB season, the average player weight is about 207 pounds. From the plot we also could see that the average height has consistently increase with a stable trend. However, the average weight jump around a lot throughout years.

birthYear <int>	avg_weight <dbl>
1994	199.9815
1995	202.9107
1996	197.9565
1997	204.5000
1998	185.0000
NA	*NA*

6 rows

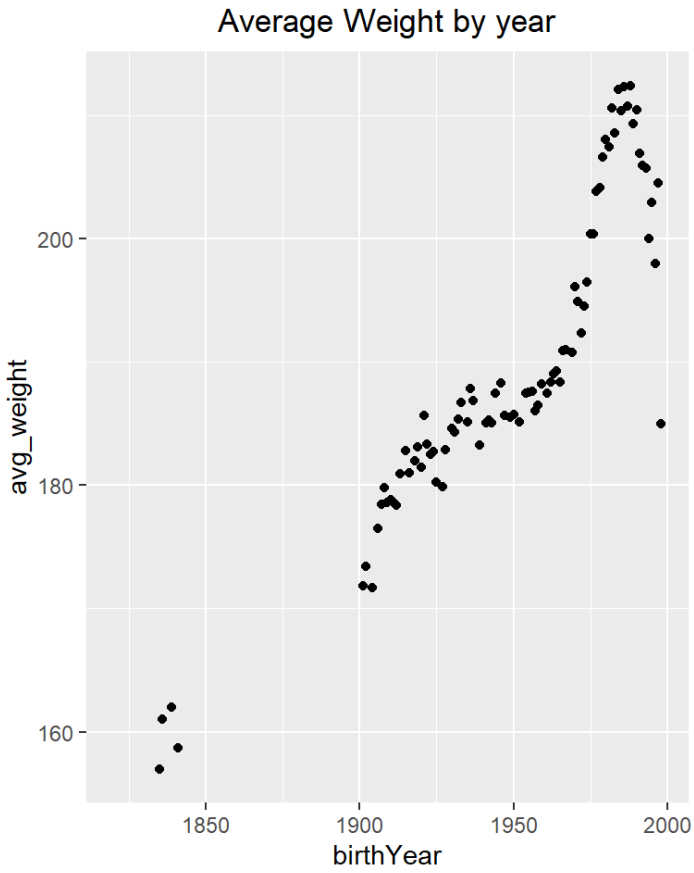
```
## [1] 189.4292
```

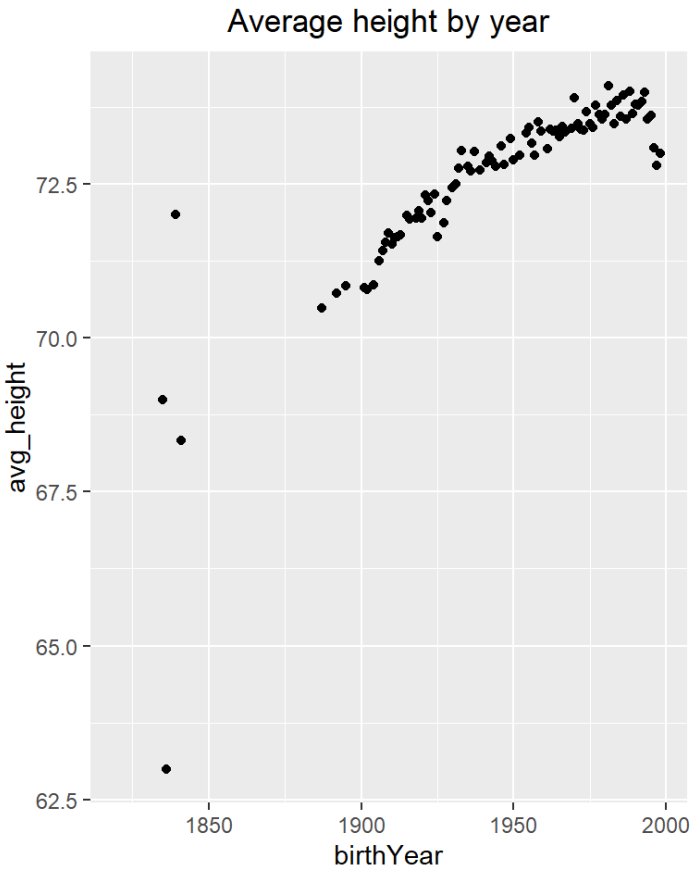
birthYear <int>	avg_height <dbl>
1994	73.55556

birthYear <int>	avg_height <dbl>
1995	73.60714
1996	73.08696
1997	72.80000
1998	73.00000
NA	*NA*

6 rows

```
## [1] 72.59008
```





State graph of Hall of Fame member The National Baseball Hall of Fame is a nonprofit committed to preserving the history of America's pastime and celebrating the legendary players, managers, umpires and executives who have made the game a fan favorite for more than a century. It is a pretty meaningful organization.
I want to explore the number of members in Hall of Fame in each state. I decide to use inner join on playerID on player and Hall of fame tables to get the players who are in the Hall. Then I find out that there is state information in college playing, using inner join on them could help me get the state information. Then I get the number of players for each state. However, while I plan to use the choropleth graph function. I realize that one important issue I have not fixed. The state name are in abbr rather than the actual name. I first try to rename each row and I know this is a stupid method. It turns out that it does not work.

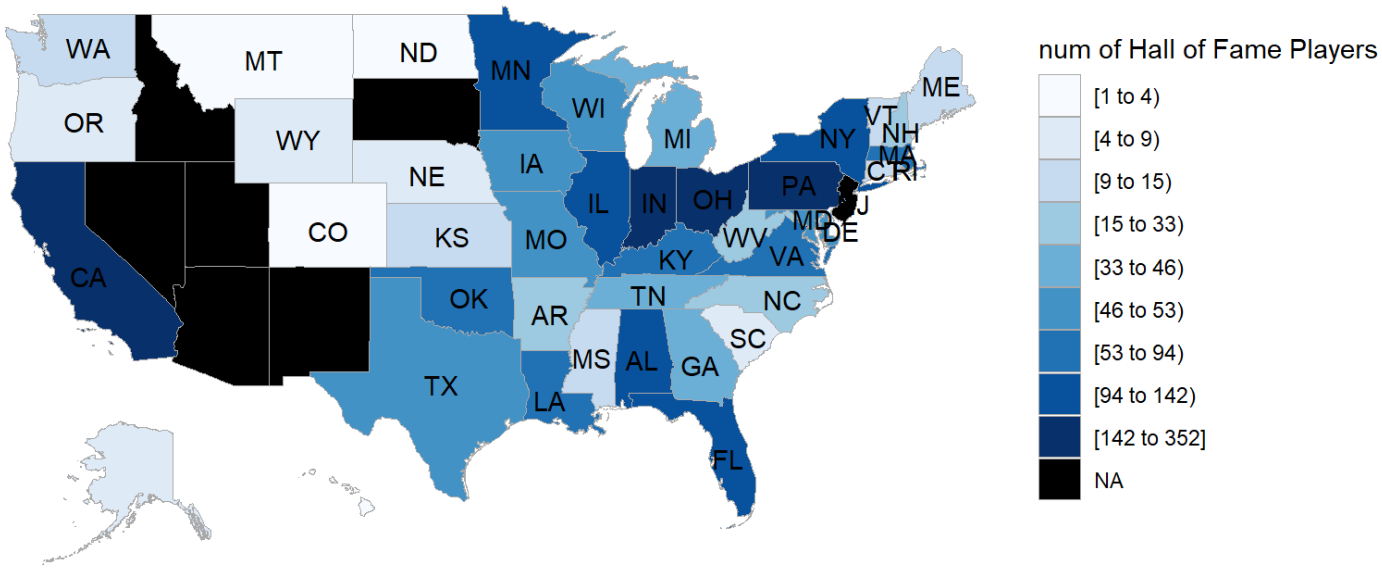
Therefore, I kind of cheat here. I first export the file to CSV, and rewrite the csv file, changing the abbreviation state name into the actual name. And then I could work on the graph.

From the graph we could see the state with the most number of Hall of Fame players are California, pennsylvania, and Indiana.

region <chr>	value <int>
CA	352
PA	185
IN	165
OH	142
NY	123
IL	102

6 rows

Distribution of Hall of Fame Players in USA



Largest slugging percentage

In baseball statistics, slugging percentage (SLG) is a measure of the batting productivity of a hitter. It is calculated as total bases divided by at bats, through the following formula, where AB is the number of at bats for a given player, and 1B, 2B, 3B, and HR are the number of singles, doubles, triples, and home runs, respectively.
Unlike batting average, slugging percentage gives more weight to extra-base hits such as doubles and home runs, relative to singles. Plate appearances resulting in walks are specifically excluded from this calculation, as an appearance that ends in a walk is not counted as an at bat.
From my perspective, it is one of the best criteria to qualify a baseball player.

We could not obtain the SLG using the orignal equation below, because we do not have the single score data. However, we have the number of hits a player make. Therefore, we could transform the equation to another one.
$$

$$\begin{aligned} SLG &= ((1 * B) + (2 * 2B) + (3 * 3B) + 4 * (HR))/AB \\ &= (1 * (H - 2B - 3B - HR)) + (2 * 2B) + (3 * 3B) + 4 * (HR))/AB \\ &= (1 * H + 2B + 2 * 3B + 4 * HR)/AB \end{aligned}$$

$$ After figuring out the equation, we need to inner join on batting and player table to get the information we need. In order to prevent from too much information, I set the filter that only include players with more than 50 at-bats in the season.

	playerID <chr>	yearID <int>	stint <int>	teamID <fctr>	lgID <fctr>	G <int>	AB <int>	R <int>	H <int>
1	abercda01	1871	1	TRO	*NA*	1	4	0	0
2	addybo01	1871	1	RC1	*NA*	25	118	30	32
3	allisar01	1871	1	CL1	*NA*	29	137	28	40
4	allisdo01	1871	1	WS3	*NA*	27	133	28	44
5	ansonca01	1871	1	RC1	*NA*	25	120	29	39
6	armstbo01	1871	1	FW1	*NA*	12	49	9	11

6 rows | 1-10 of 23 columns

	playerID <chr>	birthYear <int>	nameFirst <chr>	nameLast <chr>	SLG <dbl>
1	spencsh01	1972	Shane	Spencer	0.9104478
2	willite01	1918	Ted	Williams	0.9010989
3	bondsba01	1964	Barry	Bonds	0.8634454
4	ruthba01	1895	Babe	Ruth	0.8490153
5	ruthba01	1895	Babe	Ruth	0.8462963
6	bakerje03	1981	Jeff	Baker	0.8245614

6 rows

```
## [1] 0.3543645
```

Above is the data in history and I want to explore the data of 2019 season.
Comparing with the data between the history and 2019 season, we could see that the SLG score in 2019 season is much lower than the SLG score before. I think there are two reasons behind this fact:
1. The competition of the legend has been increased from last century to 21st century. The disparity of skill among players has benn shrinked. Therefore, it is very hard to see the arise of a superstar in MLB today.
2. The history data has been collected for a long period of time, which means that the population is very huge

compared with our population in 2019 season. Therefore, it is not surprising to see some outstanding player statistic exists in a long run period of time. And we could see that the average is not that high, which proves my assumption.

	Rd <chr>	Pick <int>	Player <chr>	Draft Tm <chr>	L <chr>	PA <int>	AB <int>	R <int>	H <int>
1	1	1	Adley Rutschman	Orioles	ALL (3)	155	130	19	33
2		*NA*				*NA*	*NA*	*NA*	*NA*
3	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
4	1	2	Bobby Witt Jr.	Royals	ROK	180	164	30	43
5		*NA*				*NA*	*NA*	*NA*	*NA*
6	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

6 rows | 1-10 of 29 columns

	Rd <chr>	Player <chr>	AVG <dbl>	SLG <dbl>	OPS <dbl>
1	17	Connor Cannon	0.324	0.669	1.063
2	1	CJ Abrams	0.393	0.647	1.083
3	24	Bryce Ball	0.329	0.628	1.023
4	27	Kyle MacDonald	0.354	0.604	1.052
5	30	Jimmy Govern	0.344	0.603	1.060
6	19	Kerry Carpenter	0.303	0.579	0.966

6 rows

Salary

The next thing I am interested in is about salaries.
I first want to check the MAX, MIN and average salary of different teams.
From the output we could see that LAA team(Los Angeles Angels) has the best average salaries among the team while NYA(New York Yankees) has the highest salary among other team.

	yearID <int>	teamID <fctr>	lgID <fctr>	playerID <chr>	salary <int>
1	1985	ATL	NL	barkele01	870000
2	1985	ATL	NL	bedrost01	550000
3	1985	ATL	NL	benedbr01	545000
4	1985	ATL	NL	campri01	633333
5	1985	ATL	NL	ceronri01	625000

	yearID <int>	teamID <fctr>	lgID <fctr>	playerID <chr>	salary <int>
6	1985	ATL	NL	chambch01	800000

6 rows

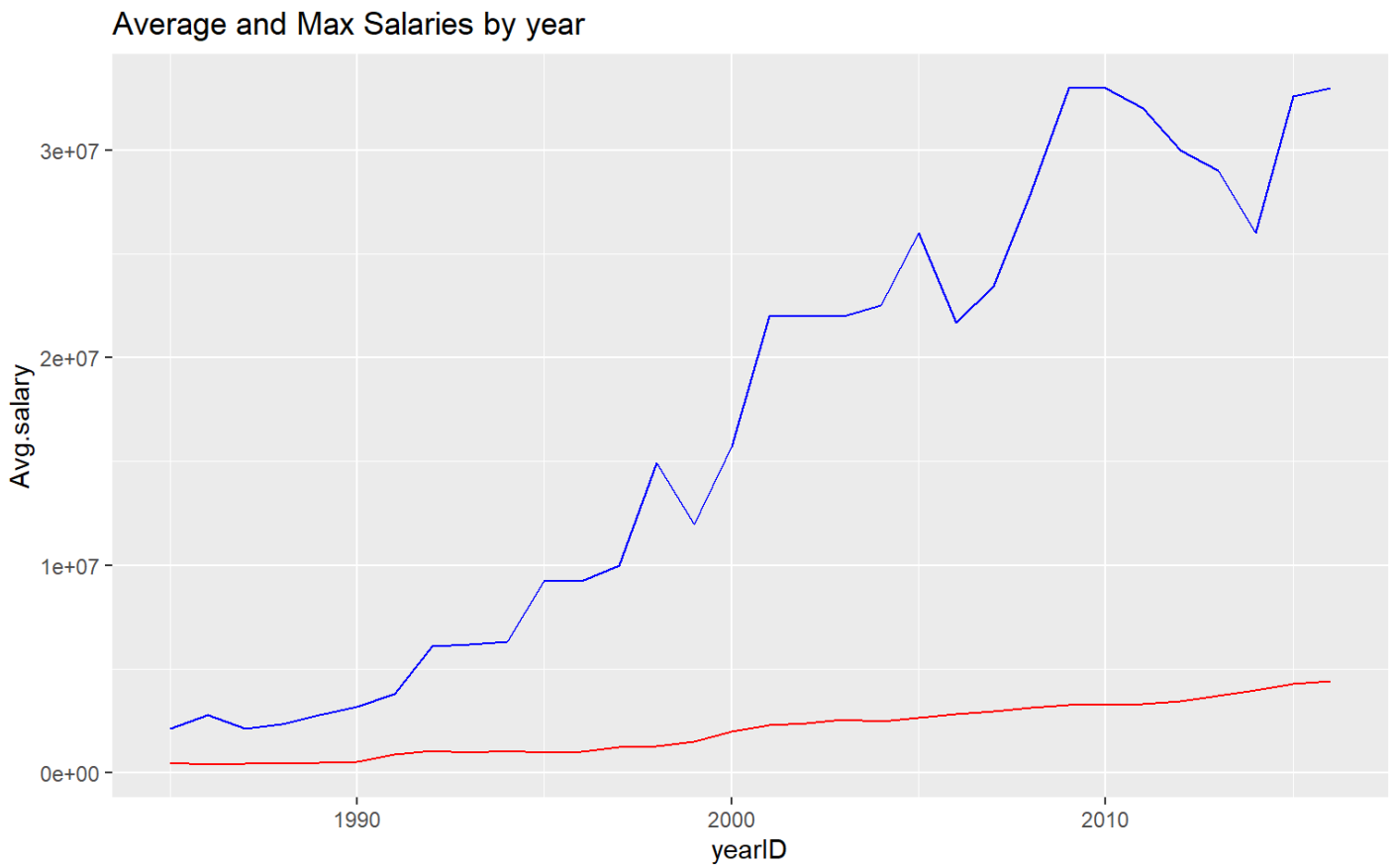
teamID <fctr>	AVG.salary <dbl>	MAX.salary <int>	MIN.salary <int>
LAA	4291454	26187500	316000
NYA	3968910	33000000	0
BOS	2968591	30000000	60000
WAS	2935073	22142857	316000
LAN	2795034	33000000	60000
MIA	2606827	19000000	480000

6 rows

I also show a plot of the average and max salaries thoughout years. From the plot we could see that average salary is increasing consistently because the econnomic development, while the max salary has some bump around. This might be due to some huge baseball star arises when the peak exists.

yearID <int>	Avg.salary <dbl>	MAX.salary <int>	MIN.salary <int>
1985	476299.4	2130300	60000
1986	417147.0	2800000	60000
1987	434729.5	2127333	62500
1988	453171.1	2340000	62500
1989	506323.1	2766667	62500
1990	511973.7	3200000	100000

6 rows



Popular name

I also want to determine the 5 most popular first names in baseball among players. After inital try, I find out that there are too many data statistic because this is a quite large database. Therefore, we need to set some constraints on the dataset we get. Since there are many players who just played serveral games in baseball league, I decide to filter out the players who just play several games in career.
First, I need to filter out the NA statistic in game record. Then I realized another issue that the table just records the game a player played in that single year. They do not have the games a player played in his career. Therefore, I need to first make a variable career total to sum up all the games a single player plays.

```
## Joining, by = "playerID"
```

playerID	career_total	nameFirst	nameLast	nameGiven	birthYear
<chr>	<int>	<chr>	<chr>	<chr>	<int>
rosepe01	3528	Pete	Rose	Peter Edward	1941
aaronha01	3020	Hank	Aaron	Henry Louis	1934
ripkeca01	2977	Cal	Ripken	Calvin Edwin	1960
cobbty01	2954	Ty	Cobb	Tyrus Raymond	1886
vizquom01	2940	Omar	Vizquel	Omar Enrique	1967
mayswi01	2929	Willie	Mays	Willie Howard	1931

6 rows

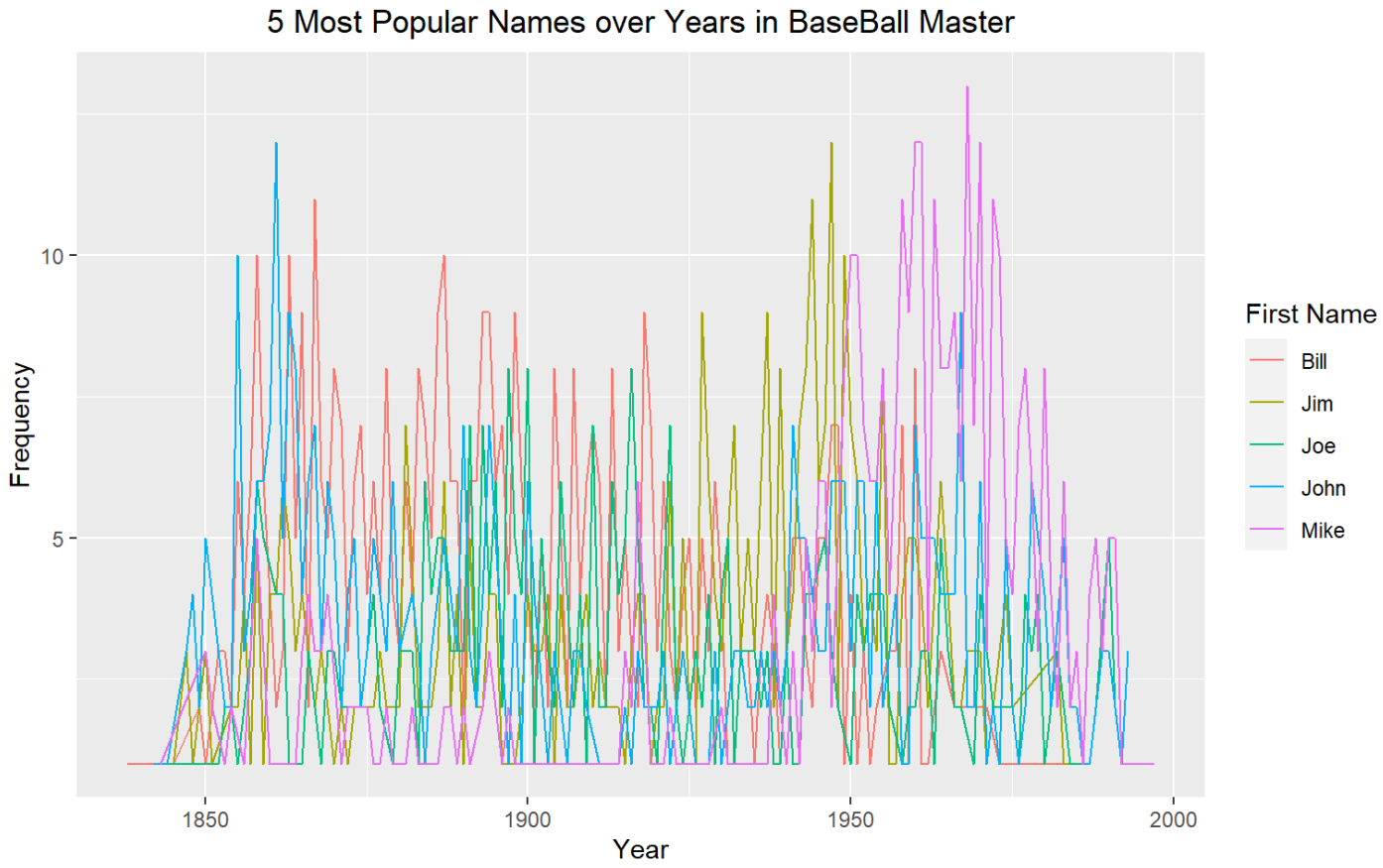
After collecting the career total games data, I am able to work out to determine the 5 most popular first names in baseball among players who played at least 500 games.

```
## Joining, by = c("playerID", "birthYear", "nameFirst", "nameLast", "nameGiven")
```

nameFirst <chr>	**count** <int>
Mike	80
Joe	60
John	56
Bill	55
Jim	52
Bob	48
George	45
Dave	43
Johnny	36
Jack	34

1-10 of 878 rows Previous **1** 2 3 4 5 6 … 88 Next

After get the name and count tables, we could see that the most popular names are Mike, Joe, John, Bill and Jim. Then I want to plot them on the same graphs.

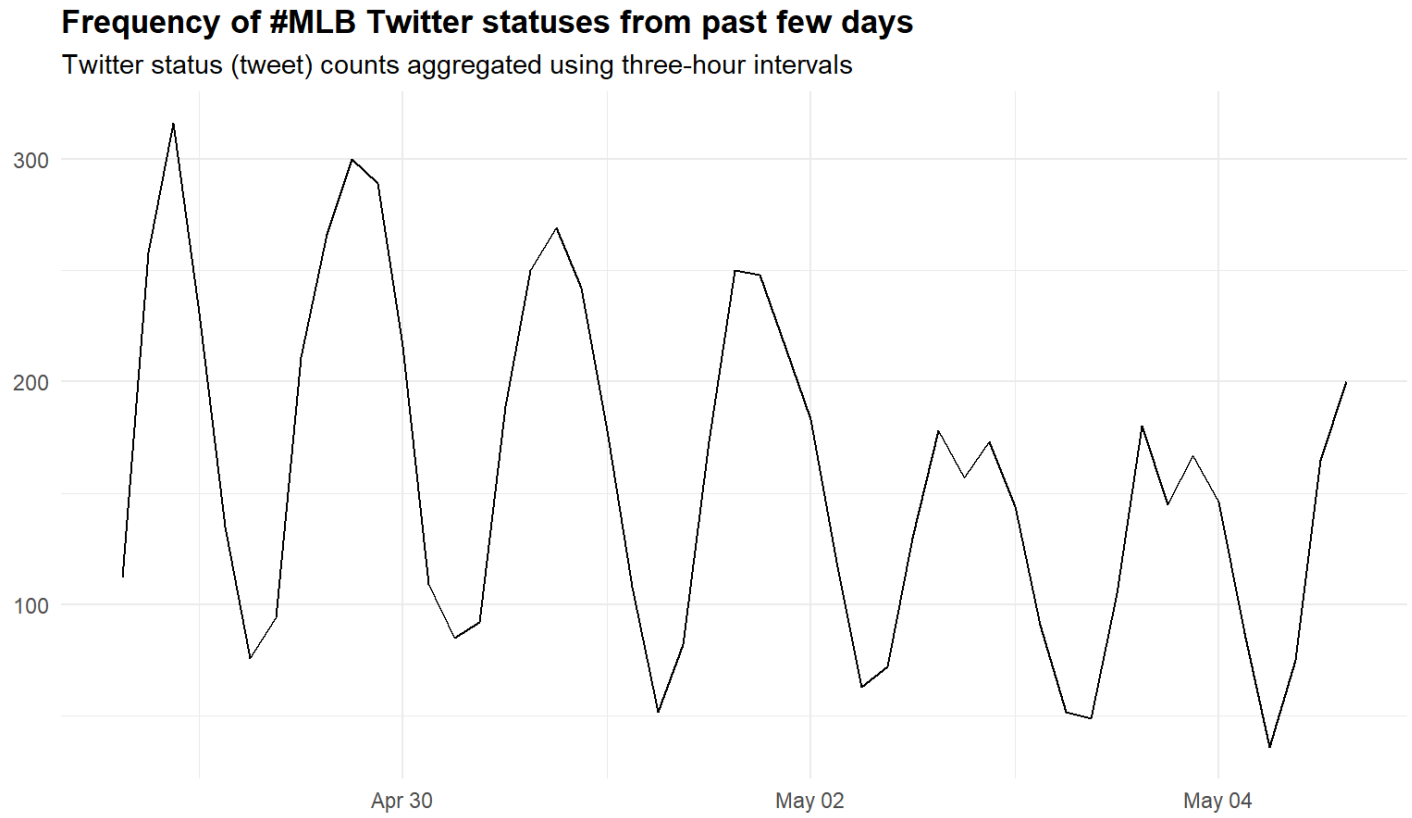


Connect with twitter

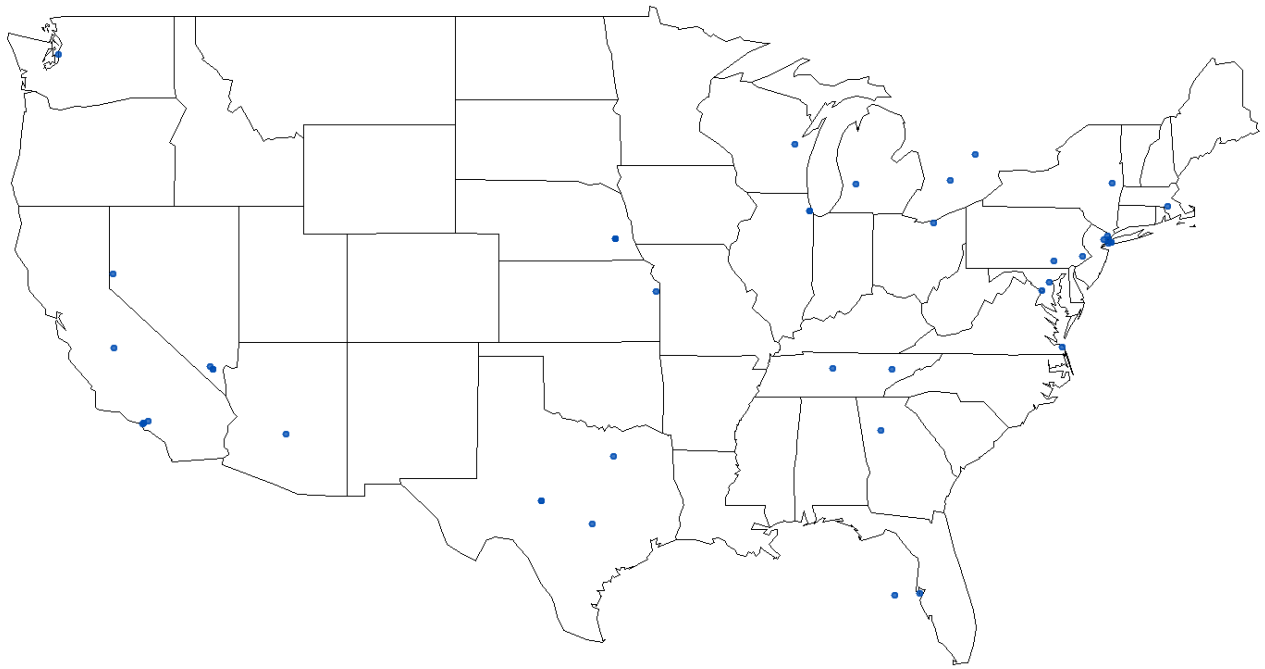
I want to see how people are concerning with MLB baseball league during such a special period of time. Using the package rtweet could help me gather the information from twitter, you will need to authorization using your own twitter account.
Here is the link for reference: “https://www.rdocumentation.org/packages/rtweet/versions/0.4.0 (https://www.rdocumentation.org/packages/rtweet/versions/0.4.0)”
And in order to retain the orginal information, I filter out the retweet data.

```
## <Token>
## <oauth_endpoint>
##  request:   https://api.twitter.com/oauth/request_token
##  authorize: https://api.twitter.com/oauth/authenticate
##  access:    https://api.twitter.com/oauth/access_token
## <oauth_app> MLB search
##   key:    tov89vja0cKNmma5vtNKLkWlH
##   secret: <hidden>
## <credentials> oauth_token, oauth_token_secret
## ---
```

From the plot we could see that there are less and less poeple are concerning about MLB. I think this is because at this special period, with MLB has already been lockout, the information of MLB are quite not popular among people.



And I also want to share a note on in which state people are concerned about the MLB.
However, I do not know if there is something wrong with my code or not. My plot contains little information. As you cansee, there are quite little information about the people who sent tweets related to MLB. And many of the states do not have any tweets related to MLB hashtag.



Conclusion

I think this project is quite meaningful for me. Because I do learn a lot from this course lecture and homework. However, I do not have a time to review them in a comprehensive aspect. This project provide a chance to go over what I have learnt in this course.

I think my project touch upon most of the stuffs covered in the class. One thing I think I do not include is regular expression part because my data does not require related method. However, I do go over the regular expression stuffs and I think I get a comprehensive understanding on it because my HW grade on it is good.

Overall, I think this is the best course in this semester because I do learn a lot useful techinique and methods in this class and I believe that they are helpful in my future career. Thank you for your lecture!