BaseballAnalysis.R

Code

Anu

Wed Feb 21 13:19:37 2018

Hide

```
library(ggplot2)
library(readr)
library(dplyr)
library(gridExtra)
options(warn=-1)
```

Hide

```
salary<-read.csv('salary.csv')
head(salary)
```

	year <int>	team_id <fctr>	league_id <fctr>	player_id <fctr>	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
6	1985	ATL	NL	chambch01	800000

6 rows

Hide

```
str(salary)
```

```
'data.frame':   25575 obs. of  5 variables:
 $ year     : int  1985 1985 1985 1985 1985 1985 1985 1985 1985 1985 ...
 $ team_id  : Factor w/ 35 levels "ANA","ARI","ATL",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ league_id: Factor w/ 2 levels "AL","NL": 2 2 2 2 2 2 2 2 2 2 ...
 $ player_id: Factor w/ 4963 levels "aardsda01","aasedo01",..: 229 288 324 689 795 805 1090 1447
1529 1854 ...
 $ salary   : int  870000 550000 545000 633333 625000 800000 150000 483333 772000 250000 ...
```

Hide

```
player<-read.csv('player.csv')
head(player)
```

	player_id <fctr>	birth_year <dbl>	birth_month <dbl>	birth_day <dbl>	birth_country <fctr>	birth_state <fctr>	birth_city <fctr>	dea
1	aardsda01	1981	12	27	USA	CO	Denver	
2	aaronha01	1934	2	5	USA	AL	Mobile	
3	aaronto01	1939	8	5	USA	AL	Mobile	
4	aasedo01	1954	9	8	USA	CA	Orange	
5	abadan01	1972	8	25	USA	FL	Palm Beach	
6	abadfe01	1985	12	17	D.R.	La Romana	La Romana	

6 rows | 1-9 of 24 columns

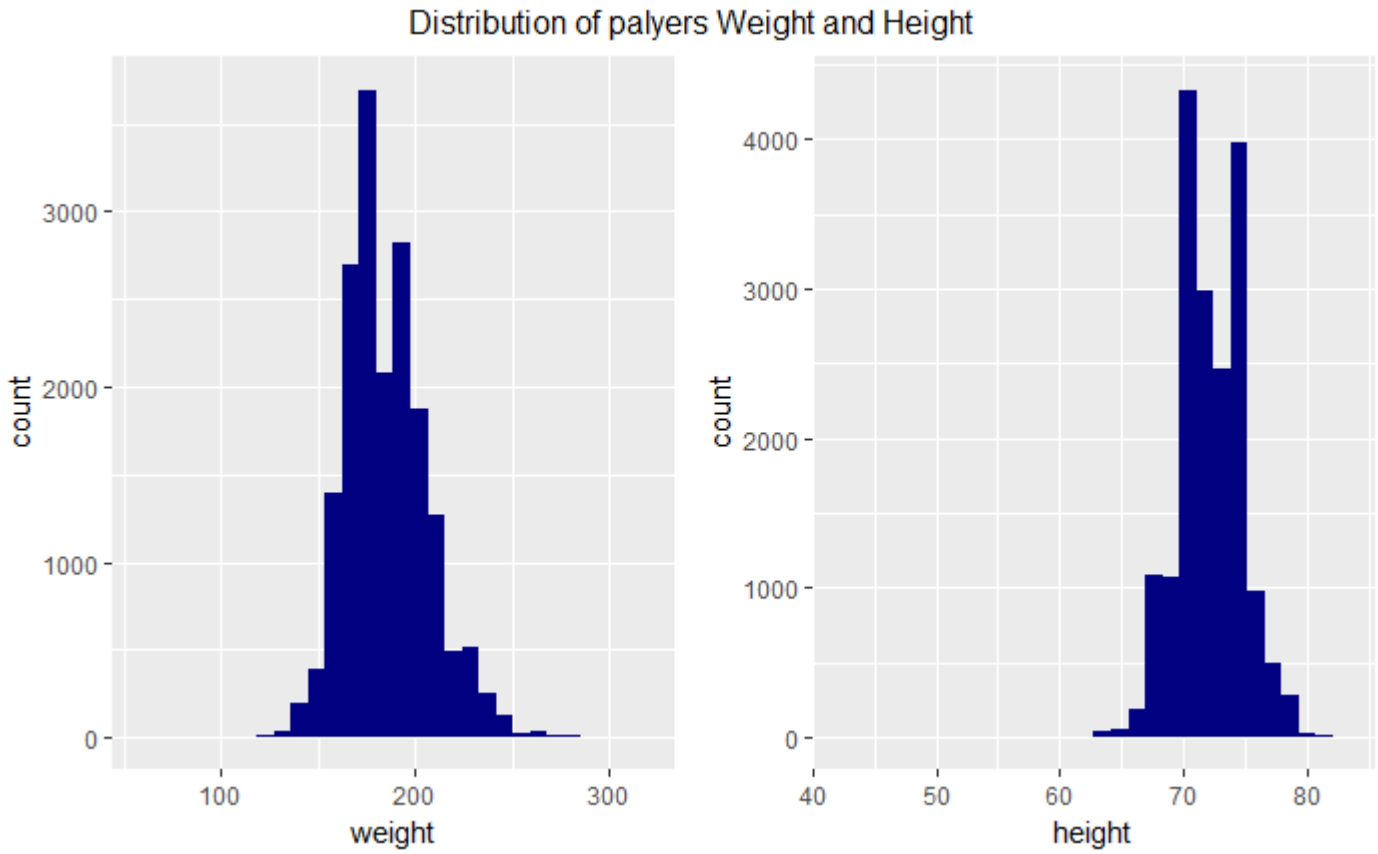
Hide

```
str(player)
```

```
'data.frame':   18846 obs. of  24 variables:
 $ player_id    : Factor w/ 18846 levels "aardsda01","aaronha01",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ birth_year   : num  1981 1934 1939 1954 1972 ...
 $ birth_month  : num  12 2 8 9 8 12 11 4 11 10 ...
 $ birth_day    : num  27 5 5 8 25 17 4 15 11 14 ...
 $ birth_country: Factor w/ 53 levels "","Afghanistan",..: 50 50 50 50 50 18 50 50 50 50 ...
 $ birth_state  : Factor w/ 246 levels "","AB","Aberdeen",..: 44 6 6 30 69 108 173 173 229 148
...
 $ birth_city   : Factor w/ 4714 levels "","Aberdeen",..: 1093 2718 2718 3092 3159 2212 3279 229
1 1337 1382 ...
 $ death_year   : num  NA NA 1984 NA NA ...
 $ death_month  : num  NA NA 8 NA NA NA 5 1 6 4 ...
 $ death_day    : num  NA NA 16 NA NA NA 17 6 11 27 ...
 $ death_country: Factor w/ 24 levels "","American Samoa",..: 1 1 22 1 1 1 22 22 22 22 ...
 $ death_state  : Factor w/ 93 levels "","AB","AK","AL",..: 1 1 26 1 1 1 57 25 88 12 ...
 $ death_city   : Factor w/ 2554 levels "","Aberdeen",..: 1 1 91 1 1 1 1735 758 462 1984 ...
 $ name_first   : Factor w/ 2313 levels "","A. J.","Aaron",..: 513 918 2087 599 73 775 1148 649
158 335 ...
 $ name_last    : Factor w/ 9713 levels "Aardsma","Aaron",..: 1 2 2 3 4 4 5 6 7 7 ...
 $ name_given   : Factor w/ 12437 levels "","A. Harry",..: 2491 5099 11411 2886 3732 3766 6605 3
201 954 1714 ...
 $ weight       : num  220 180 190 190 184 220 192 170 175 169 ...
 $ height       : num  75 72 75 75 73 73 72 71 71 68 ...
 $ bats         : Factor w/ 4 levels "","B","L","R": 4 4 4 4 3 3 4 4 4 3 ...
 $ throws       : Factor w/ 3 levels "","L","R": 3 3 3 3 2 2 3 3 3 2 ...
 $ debut        : Factor w/ 10037 levels "","1871-05-04",..: 8636 4698 5145 6222 8419 9383 106 1
132 875 934 ...
 $ final_game   : Factor w/ 9029 levels "","1871-05-05",..: 8991 5859 5560 6745 7984 9028 103 18
51 1078 1110 ...
 $ retro_id     : Factor w/ 18793 levels "","aardd001",..: 2 3 4 5 6 7 8 9 10 11 ...
 $ bbref_id     : Factor w/ 18846 levels "","aardsda01",..: 2 3 4 5 6 7 8 9 10 11 ...
```

Hide

```
options(repr.plot.width=5, repr.plot.height=3)
p1<-player %>%ggplot(aes(x=weight))+geom_histogram(fill="navyblue")
p2<-player %>%ggplot(aes(x=height))+geom_histogram(fill="navyblue")
grid.arrange(p1,p2,nrow=1,ncol=2,top="Distribution of palyers Weight and Height")
```



Hide

```
player$debut<-as.Date(player$debut,"%Y-%m-%d")
player$final_game<-as.Date(player$final_game,"%Y-%m-%d")
player$dob <- as.Date(with(player, paste(birth_year, birth_month, birth_day,sep="-")), "%Y-%m-%
d")
player$dage<-round(as.numeric(difftime(player$debut, player$dob,  unit="weeks"))/52.25,0)
head(player$dage)
```

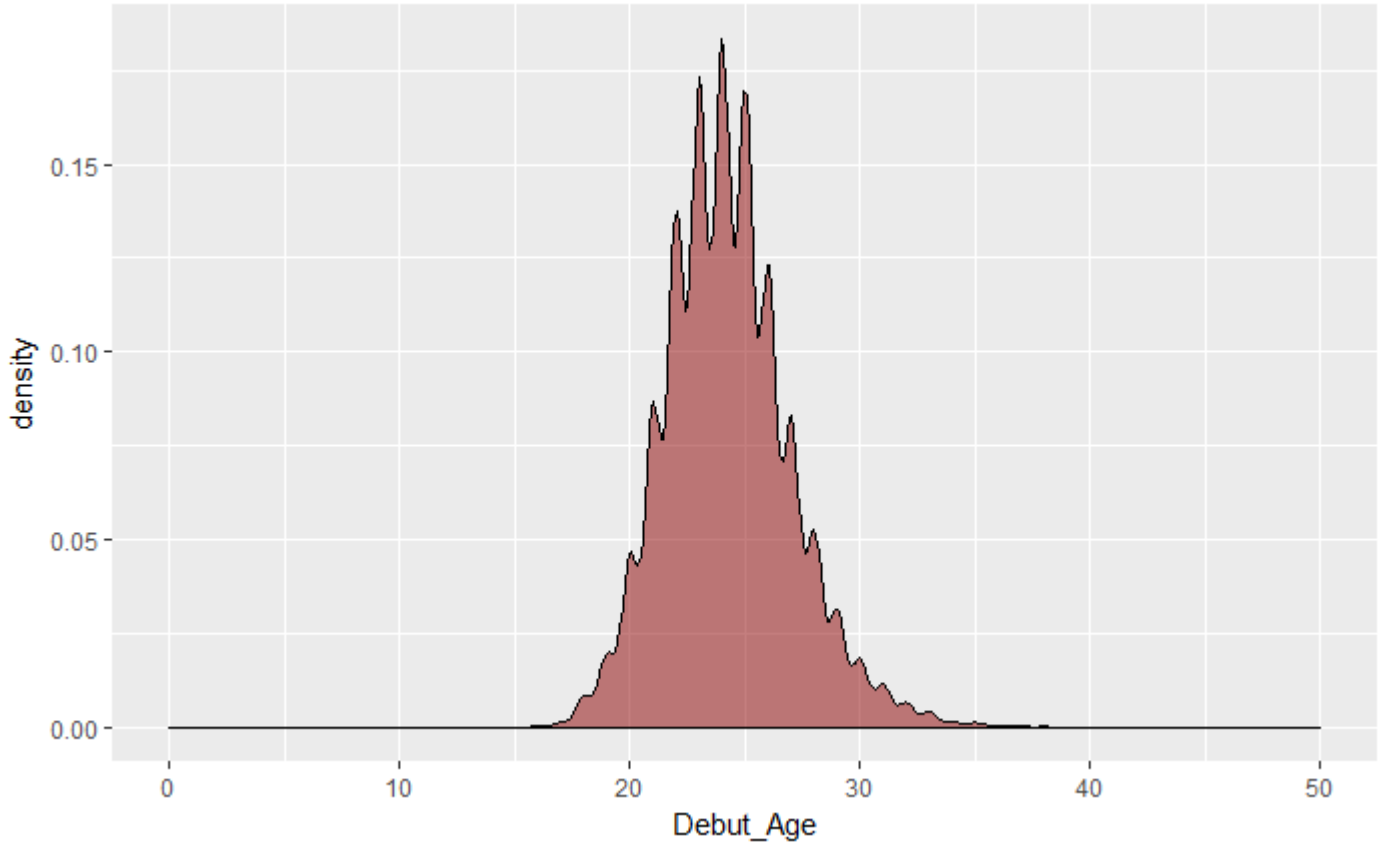
```
[1] 22 20 23 23 29 25
```

Weight of the players was measured in pounds and mostly weight ranges from 150 to 200 pounds. Height was measured in inches, most players has height 65 to 75.

Data Manipulation Most of the date columns are factors, so convert it into date format. Also do computation on birth date and debut.

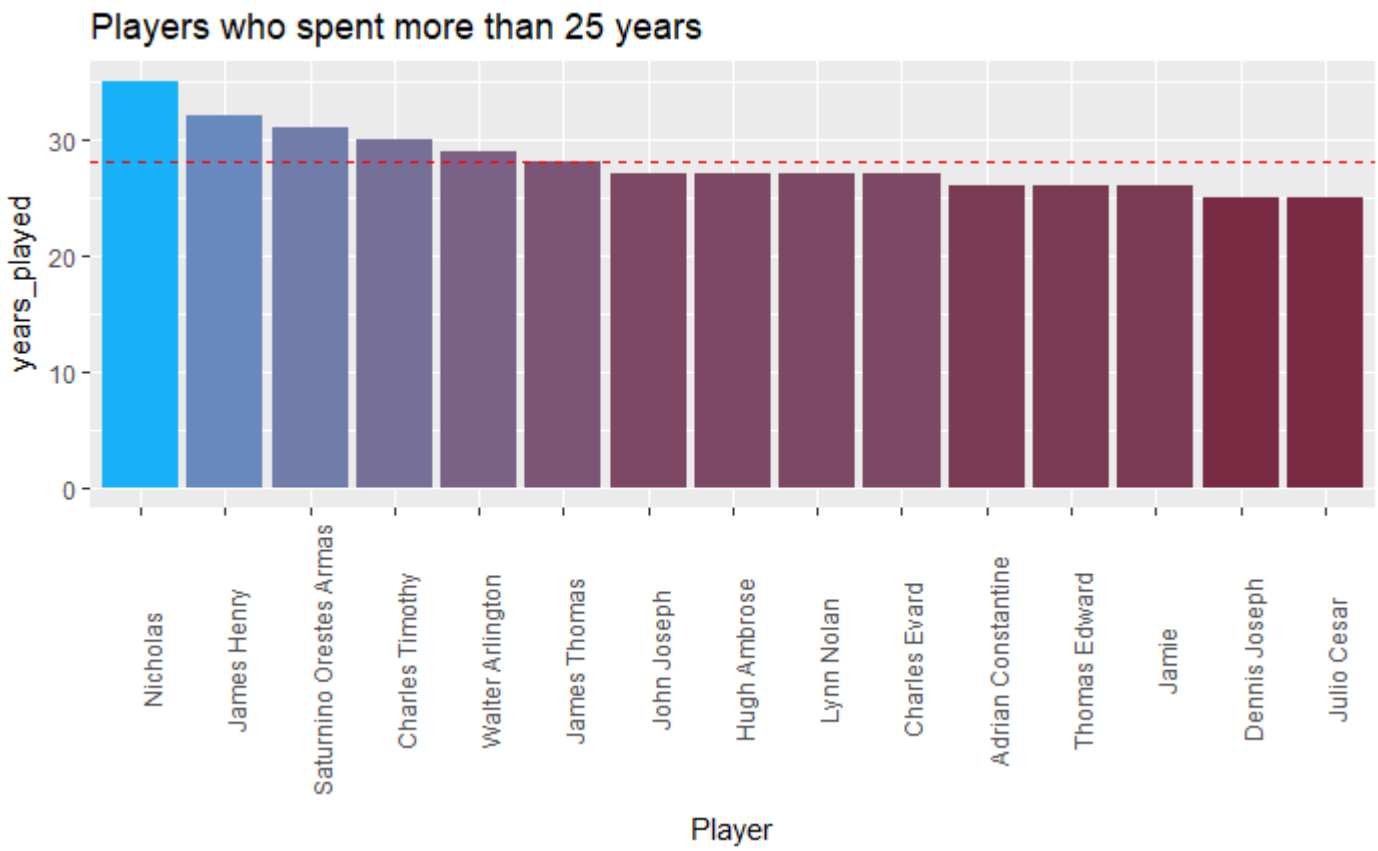
Hide

```
options(repr.plot.width=5, repr.plot.height=4)
player %>% select(name_given,dage)%>% arrange(desc(dage))%>%ggplot(aes(x=dage))+geom_density(fil
l="red4",alpha=0.5)+scale_x_continuous(limits=c(0,50))+labs(x="Debut_Age")
```



Hide

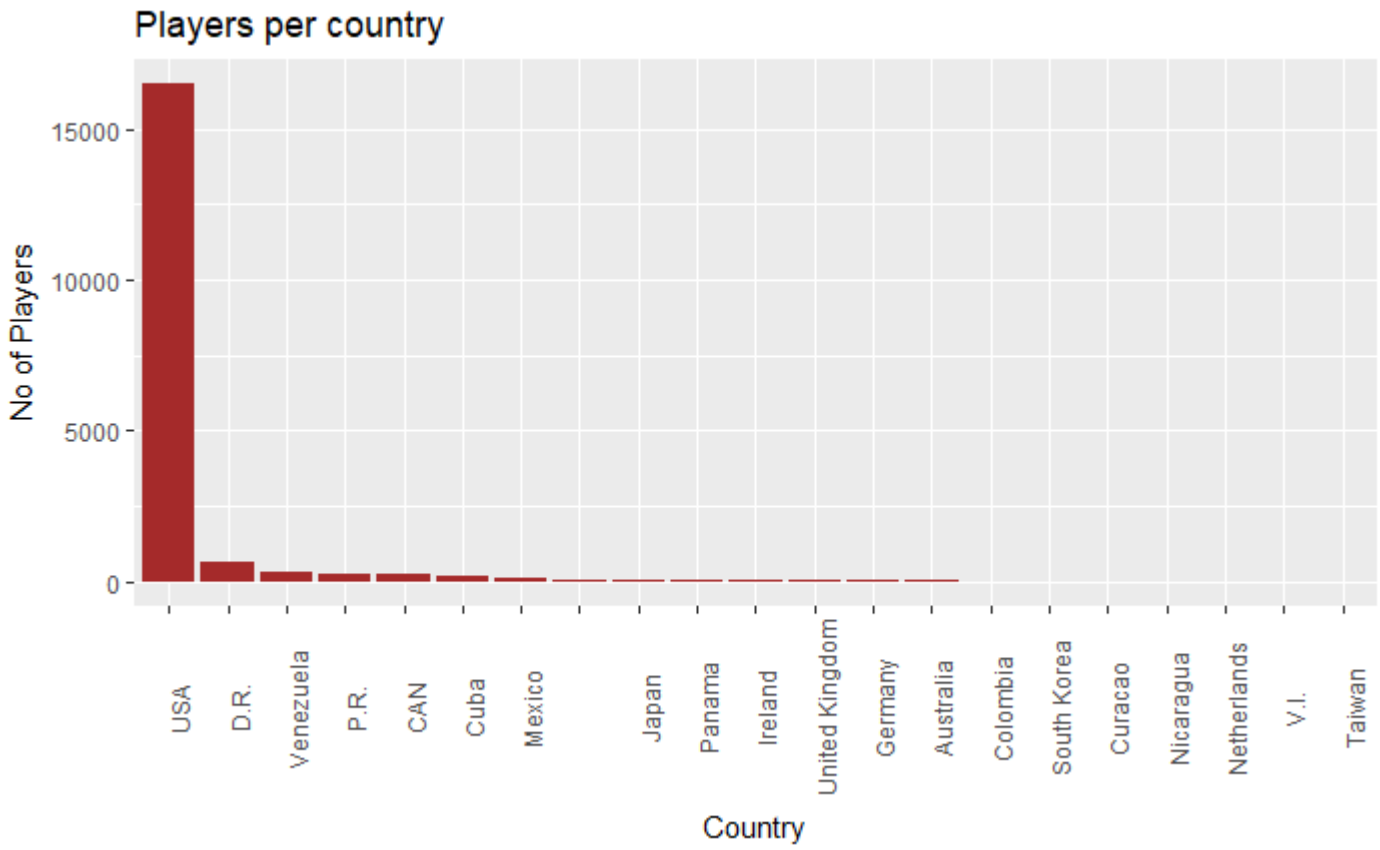
```
player$years_played<-round(as.numeric(difftime(player$final_game, player$debut,  unit="weeks"))/
52.25,0)
options(repr.plot.width=5, repr.plot.height=4)
player %>% select(name_given,years_played)%>% arrange(desc(years_played)) %>% filter(years_playe
d>=25)%>%ggplot(aes(x=factor(name_given,levels=name_given),y=years_played,fill=years_played))+
geom_bar(stat="identity")+theme(axis.text.x=element_text(angle=90),legend.position="none")+labs
(x="Player",title="Players who spent more than 25 years")+scale_fill_gradient(low = "#782B43", h
igh = "#16B1F7")+
geom_hline(aes(yintercept=mean(years_played)),col="red",linetype="dashed")
```



Most of the players started at the age of 22 to 25, Minimum age was at 16 and max age age was 36.

Hide

```
player %>%select(birth_country)%>%group_by(birth_country)%>%summarise(pcount=n())%>%arrange(desc(pcount))%>%filter(pcount>=10)%>%ggplot(aes(x=factor(birth_country,levels=birth_country),y=pcount))+geom_col(fill="brown")+theme(axis.text.x=element_text(angle=90))+labs(x="Country",y="No of Players",title="Players per country")
```



Hide

```
player %>%select(birth_country)%>%group_by(birth_country)%>%summarise(pcount=n())%>%arrange(desc(pcount))%>%filter(pcount>=10)%>%ggplot(aes(x=factor(birth_country,levels=birth_country),y=pcount))+geom_col(fill="brown")+theme(axis.text.x=element_text(angle=90))+labs(x="Country",y="No of Players",title="Players per country")
```

USA has got more number of baseball players , which was not at all comparable to other countries. ```