Studio dei successi in relazione allo stato economico nel Baseball

Matteo Pizzinato

31/5/2021

Il Baseball

> *"si presume abbia origine in Nord America verso la prima metà del XVIII secolo ma solo nel 1830 ci fu la prima codifica ufficiale di un regolamento."*
> *Wikipedia.org*

Uno sport antico quindi, che riunisce miliardi di tifosi sparsi in tutto il mondo, dagli Stati Uniti al Giappone passando per l'Europa, le squadre spendono milioni di dollari ogni anno per gli ingaggi dei giocatori e ne incassano altrettanti con il merchandise, le vincite dei vari campionati, divisioni e titoli.

Con questo progetto la mia intenzione è di scoprire non più di tanto perchè il baseball è molto seguito, quanto più di analizzare le possiblità economiche di una squadra in relazione ai suoi successi nei 30 anni che vanno dal 1985 al 2015:

- Quali sono le squadre di maggior successo nella storia del Baseball e quale invece quelle peggiori?
- E' vero che per avere successo servono molti soldi?
- Gli stipendi dei giocatori come sono evoluti nel tempo?
- Il baseball è uno sport che favorisce le squadre ricche?

Per rispondere a queste domande farò riferimento anche ad una particolare squadra, gli Oakland Athletics e alla rivoluzione del 2002 portata avanti dall'ex giocatore e all'epoca direttore generale degli Oak, Billy Beane, un uomo che seppe rivoluzionare il mondo del baseball come nessuno prima.

Fasi preliminari

Importo le librerie:

```
library(dplyr)
library(tidyverse)
library(ggplot2)
library(gridExtra)
library(maps)
library(ggrepel)
library(ggthemes)
library(xfun)
library(tinytex)
```

```
baseballTeams = read.csv(file= '../dataset/team.csv',na.strings=c("","NA"))
baseballPlayers = read.csv(file = "../dataset//player.csv",na.strings=c("","NA"))
baseballSalary = read.csv(file= '../dataset//salary.csv',na.strings=c("","NA"))
baseballCityTeams = read.csv(file = '../dataset//cityTeams.csv')
```

Modifico il dataset filtrando solo i dati che mi serviranno per il progetto:

```r
consYears = 1985:2015 # i 30 anni che prendo in considerazione# elimino delle colonne che non mi servon

baseballTeams = subset(baseballTeams, select = -c(team_id_retro, team_id_lahman45,team_id, attendance, l

names(baseballTeams)[names(baseballTeams) == "w"] = "wins"

names(baseballTeams)[names(baseballTeams) == "l"] = "losses"

names(baseballTeams)[names(baseballTeams) == "e"] = "errors"

names(baseballTeams)[names(baseballTeams) == "h"] = "hits"

names(baseballTeams)[names(baseballTeams) == "so"] = "strike_out"

names(baseballTeams)[names(baseballTeams) == "r"] = "runs"

names(baseballTeams)[names(baseballTeams) == "team_id_br"] = "team_id"
```

Introduzione

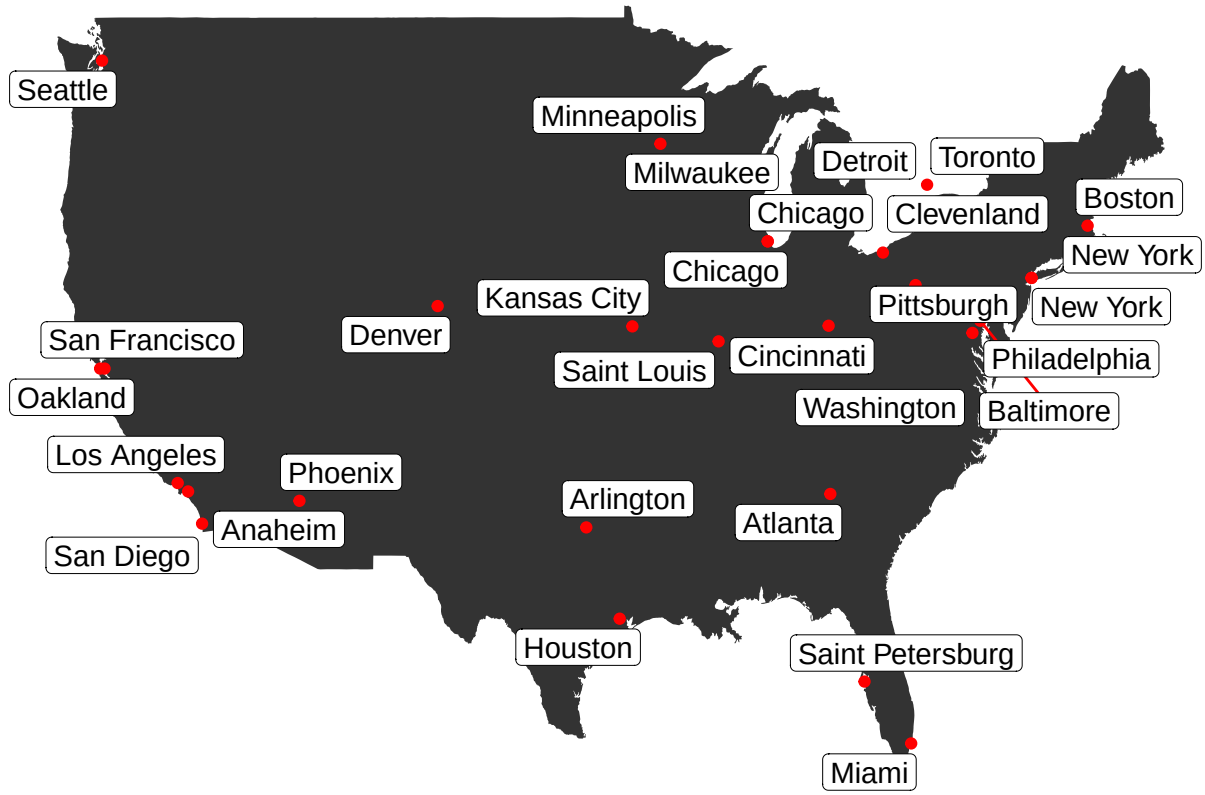
Per comprendere meglio alcuni passaggi fondamentali di questo progetto e per capire le operazioni eseguite sui dati stessi è meglio descrivere brevemente la struttura del campionato di Baseball:

In totale ci sono trenta squadre provenienti da tutti gli Stati Uniti che si sfidano in due campionati paralleli ovvero la American League e la National League, ogni campionato comprende la partecipazione di quindici squadre divise a loro volta in tre divisioni: East, Central e West. Ogni divisione comprende cinque squadre che si sfidano per il titolo divisione appunto. Le migliori di ogni divisione poi si sfideranno per il titolo campionato e le migliori parteciperanno poi al torneo per decretare la vincitrice delle World Series ovvero il titolo di maggior rilievo nel baseball.

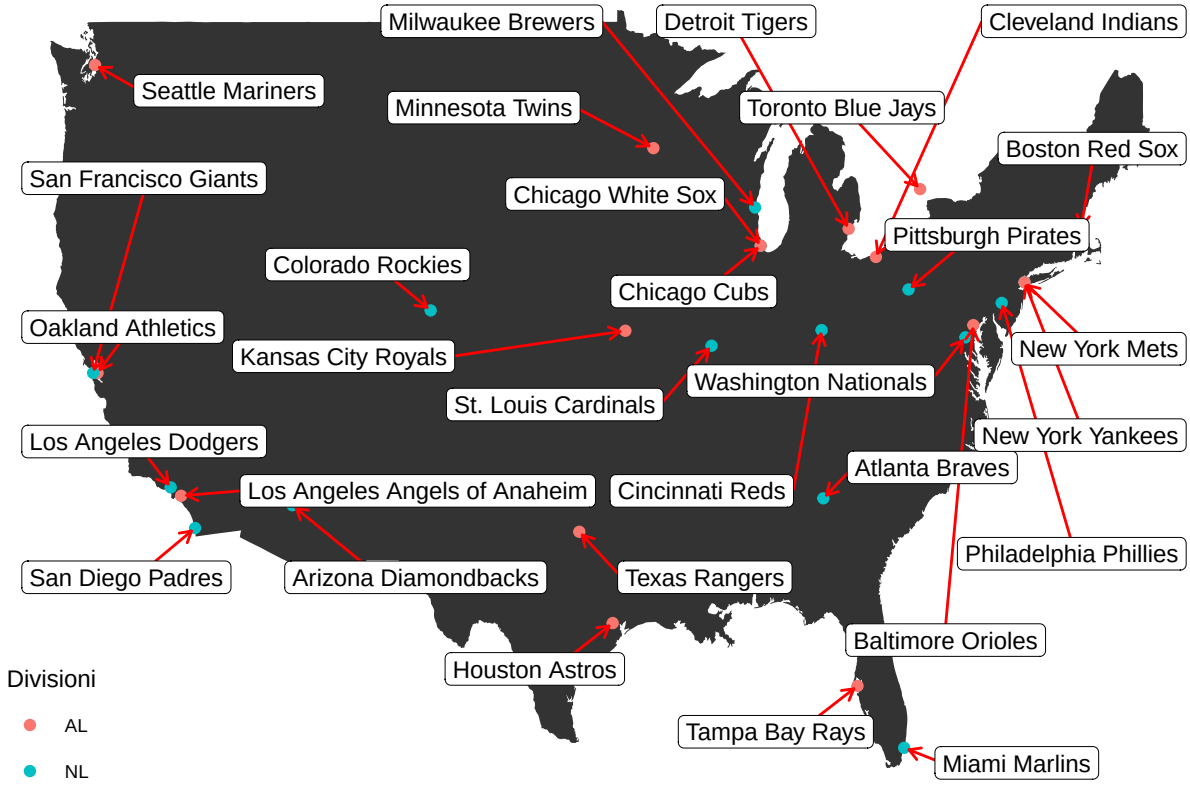
Di seguito vi è una mappa che mostra la collocazine delle città che ospitano le varie squadre e successivamente le squadre divise per campionato.

```r
x = baseballTeams %>%
  select(team_id, name, league_id) %>%
  distinct()

mapTeamCity = merge(x, baseballCityTeams, by="team_id")

# rimuovo i doppioni che is sono creati con i trasferimenti da una campionato all'altro da parte degli
mapTeamCity = mapTeamCity[!(mapTeamCity$team_id == "HOU" & mapTeamCity$league_id == "NL"),]
mapTeamCity = mapTeamCity[!(mapTeamCity$team_id == "MIL" & mapTeamCity$league_id == "AL"),]

# Mappa delle località delle squadre
options(ggrepel.max.overlaps = Inf)

usa = map_data("usa")
plotTeamCity = ggplot() +
  geom_polygon(data = usa, aes(x = long, y = lat, group = group)) +
  geom_point(data = mapTeamCity, aes(x = long, y = lat), color = "red") +
  geom_label_repel(aes(x = mapTeamCity$long, y = mapTeamCity$lat, label = mapTeamCity$city),
                   label.size = 0,
                   box.padding = unit(0.2, "line"),
                   label.padding = 0.2,
                   point.padding = 0.1,
```

```
                  min.segment.length = 2,
                  segment.color = 'red') +
  theme_map()
plotTeamCity
```




```
plotLeagueTeams = ggplot() +
  geom_polygon(data = usa, aes(x = long, y = lat, group = group)) +
  geom_point(data = mapTeamCity, aes(x = long, y = lat, color = league_id)) +
  guides(col = guide_legend("Divisioni")) +
  geom_label_repel(aes(x = mapTeamCity$long, y = mapTeamCity$lat, label = mapTeamCity$name),
                  label.size = 0,
                  box.padding = unit(0.7, "line"),
                  label.padding = 0.2,
                  point.padding = 0.1,
                  min.segment.length = 0.5,
                  segment.color = 'red',
                  verbose = TRUE,
                  seed = 123,
                  max.time = 5,
                  max.iter = Inf,
                  size = 3,
                  arrow = arrow(length = unit(0.015, "npc"))) +
  theme_map()
plotLeagueTeams
```




```
## 5.00s elapsed for 154760 iterations, 45 overlaps. Consider increasing 'max.time'.
```

Primi risultati e analisi generale

In questi 30 anni voglio estrarre i nomi delle squadre che hanno vinto più National League.

```
mostNLWins = baseballTeams %>%
filter(league_id == "NL" & lg_win == "Y") %>%
select(year, name, league_id, team_id) %>%
group_by(name) %>%
count(team_id, sort = TRUE)
```

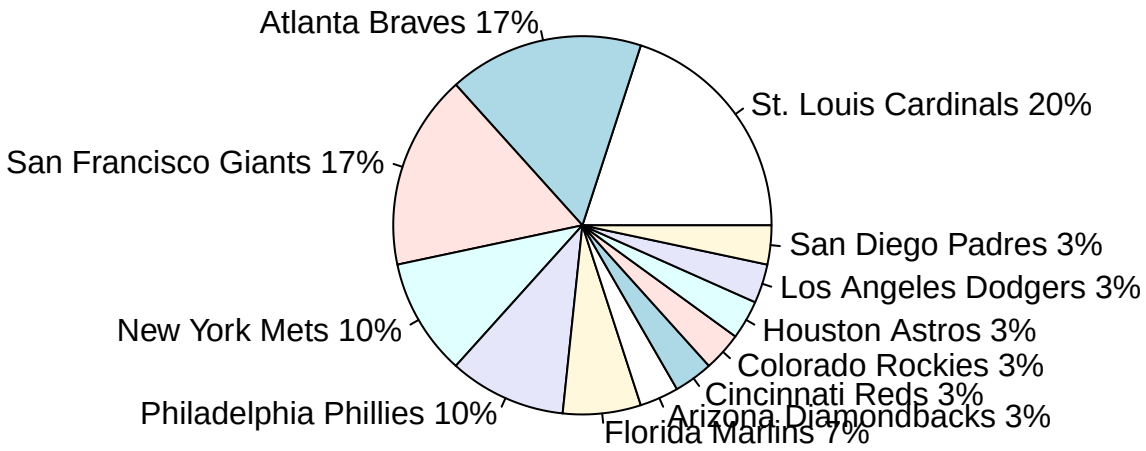
Ora considero le squadre che hanno vinto più American League.

```
mostALWins = baseballTeams %>%
filter(league_id == "AL" & lg_win == "Y") %>%
select(year, name, league_id, team_id) %>%
group_by(name) %>%
count(team_id, sort = TRUE)
```

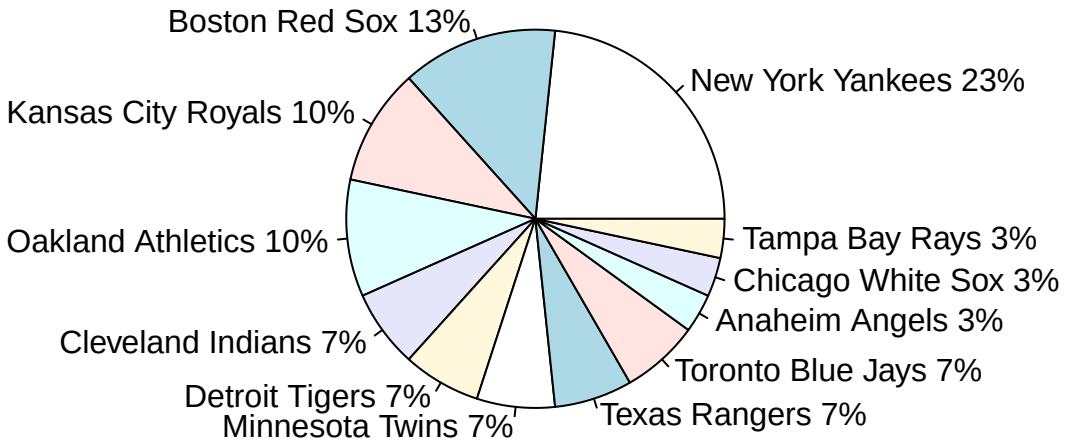
Stampo i dati in un diagramma a torta per vedere in quali proporzioni le squadre hanno ottenuto i loro successi nelle rispettive divisioni.

```
slicesNL = mostNLWins$n
lblsNL = mostNLWins$name
pctNL = round(mostNLWins$n/sum(slicesNL)*100)
lblsNL = paste(lblsNL, pctNL)
lblsNL = paste(lblsNL,"%",sep="")
```

```
plotNL = pie(slicesNL, labels = lblsNL, main = "Successi National League")
```




```
slicesAL = mostALWins$n
lblsAL = mostALWins$name
pctAL = round(mostALWins$n/sum(slicesAL)*100)
lblsAL = paste(lblsAL, pctAL)
lblsAL = paste(lblsAL,"%",sep="")
plotAL = pie(slicesAL, labels = lblsAL, main = "Successi American League")
```

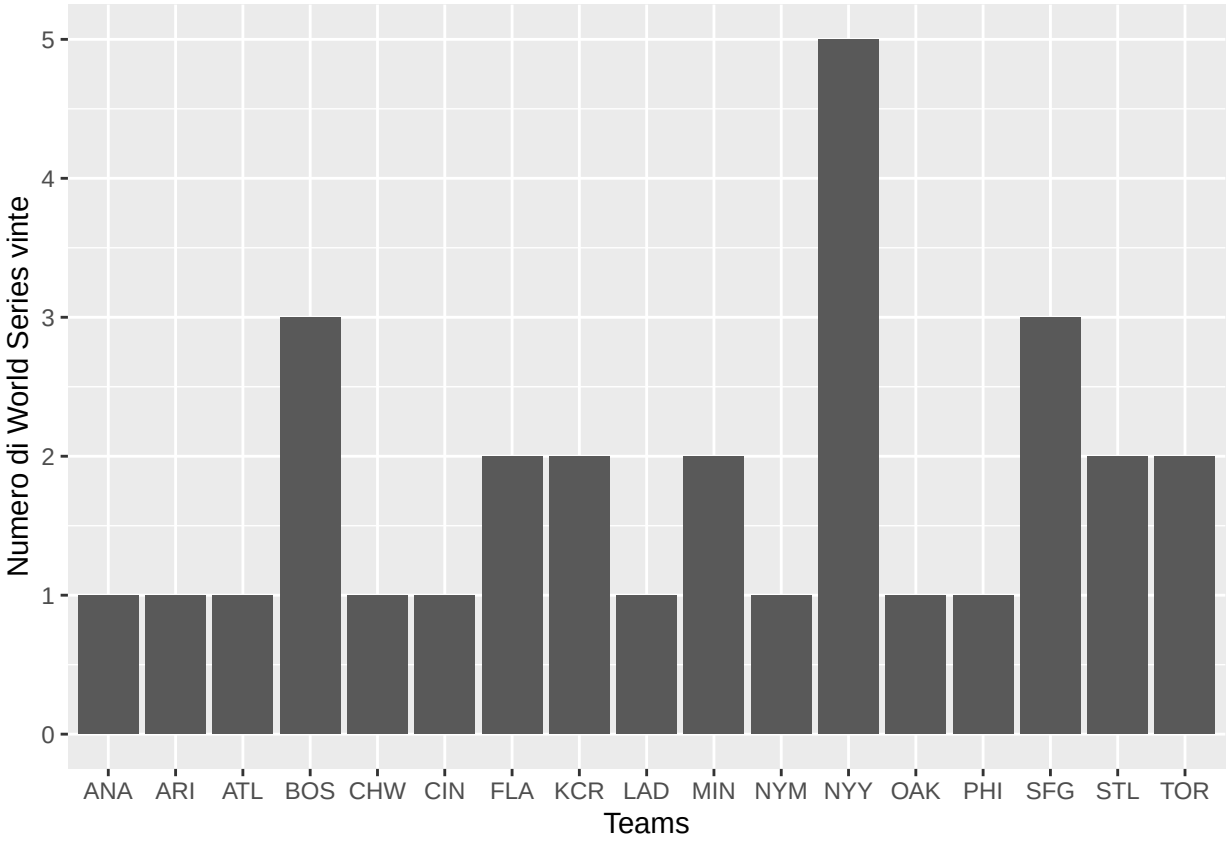


La World Series, nota anche come Fall Classic, costituisce la fase finale del campionato professionistico americano di baseball, per decretare la squadra campione della Major League Baseball (MLB). Wikipedia.org.

Quali sono le squadre che hanno vinto più World Series?

```
mostMWWS = baseballTeams %>%
  filter(ws_win == "Y") %>%
  select(year, name, league_id, team_id) %>%
  group_by(name) %>%
  count(team_id, sort = TRUE)
```

```
# Stampo i dati

plotWorldSeries = ggplot(data=mostMWWS, aes(x=team_id, y=n)) +
  geom_bar(stat="identity") + xlab("Teams") + ylab("Numero di World Series vinte")
plotWorldSeries
```



Vediamo che vi è un predominio dei New York Yankees con ben 5 world series vinte negli ultimi 30 anni.

Quale squadra ha vinto di meno?
Per ottenere questo dato devo fare delle considerazioni:

1. Cercare le squadre che hanno meno titoli divisione.
2. Dalle squadre ottenute al passaggio 1 conto quali hanno vinto almeno un titolo lega.
3. Infine elimino le squadre ottenute al passaggio 2 da quelle ottenute al punto 1.

I rimanenti team saranno quelli che hanno vinto di meno.

```
loseTeam = baseballTeams %>%
  filter(div_win == "N") %>%
  select(name, ws_win, lg_win, div_win, team_id) %>%
  count(team_id, name)
```

```
lTWLeague = baseballTeams %>%
  filter(team_id %in% loseTeam$team_id, lg_win == "Y") %>%
  select(name, team_id) %>%
  count(team_id)
```

```
loseTeam = loseTeam[!loseTeam$team_id %in% lTWLeague$team_id,]
```

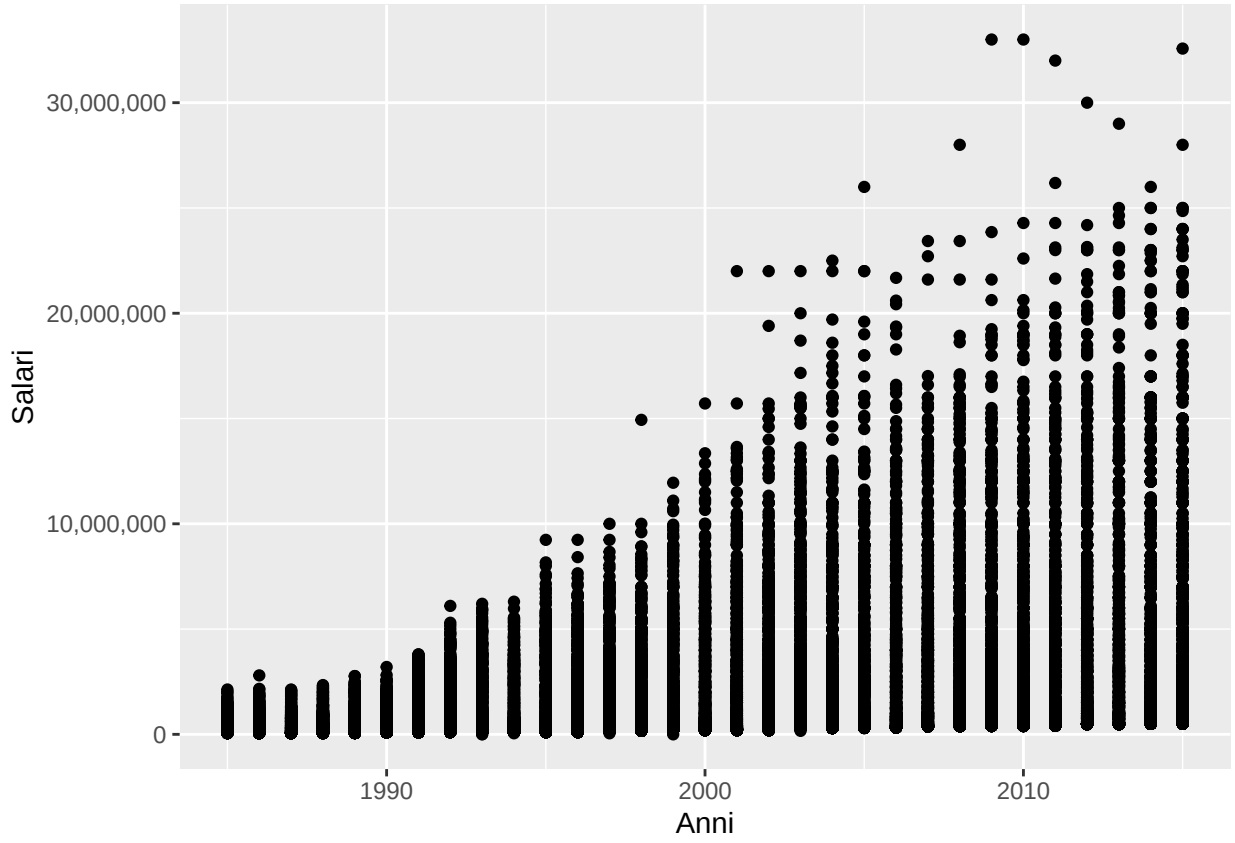
Infine riordinando per il numero di volte in cui un team ha perso la divisione, possiamo ottenere la squadra che ha vinto meno.

```
loseTeam = loseTeam[order(-loseTeam$n),]
loseTeam
```

```
##    team_id                          name  n
## 19     MIL             Milwaukee Brewers 29
## 4      BAL             Baltimore Orioles 28
## 26     PIT            Pittsburgh Pirates 27
## 28     SEA              Seattle Mariners 27
## 7      CHC                  Chicago Cubs 26
## 21     MON               Montreal Expos 19
## 6      CAL             California Angels 10
## 31     TBD          Tampa Bay Devil Rays 10
## 35     WSN          Washington Nationals  9
## 16     LAA Los Angeles Angels of Anaheim  6
## 18     MIA                 Miami Marlins  4
```

I salari dei giocatori sono aumentati nel corso degli anni?

```
plotSalary = ggplot(baseballSalary, aes(x=year, y=salary)) + geom_point() + scale_y_continuous(labels =
  ylab("Salari") +
  xlab("Anni")

# stampo il grafico

plotSalary
```



Possiamo notare che i salari sono aumentati in maniera graduale questo ci porta a pensare che nel corso degli anni ci sia stato un aumento dei tifosi e di conseguenza più merchandise venduto, quindi le squadre possono disporre di più soldi per pagare i giocatori.

E' vero che le squadre che vincono di più hanno i giocatori più pagati?

```
# per comodità unisco le due tabelle baseballPlayers e baseball Salary

pSD = inner_join(baseballPlayers, baseballSalary, by = NULL, copy = FALSE) # player salary and descript
```

Trovo quali sono i nomi dei giocatori e quali sono i loro salari.

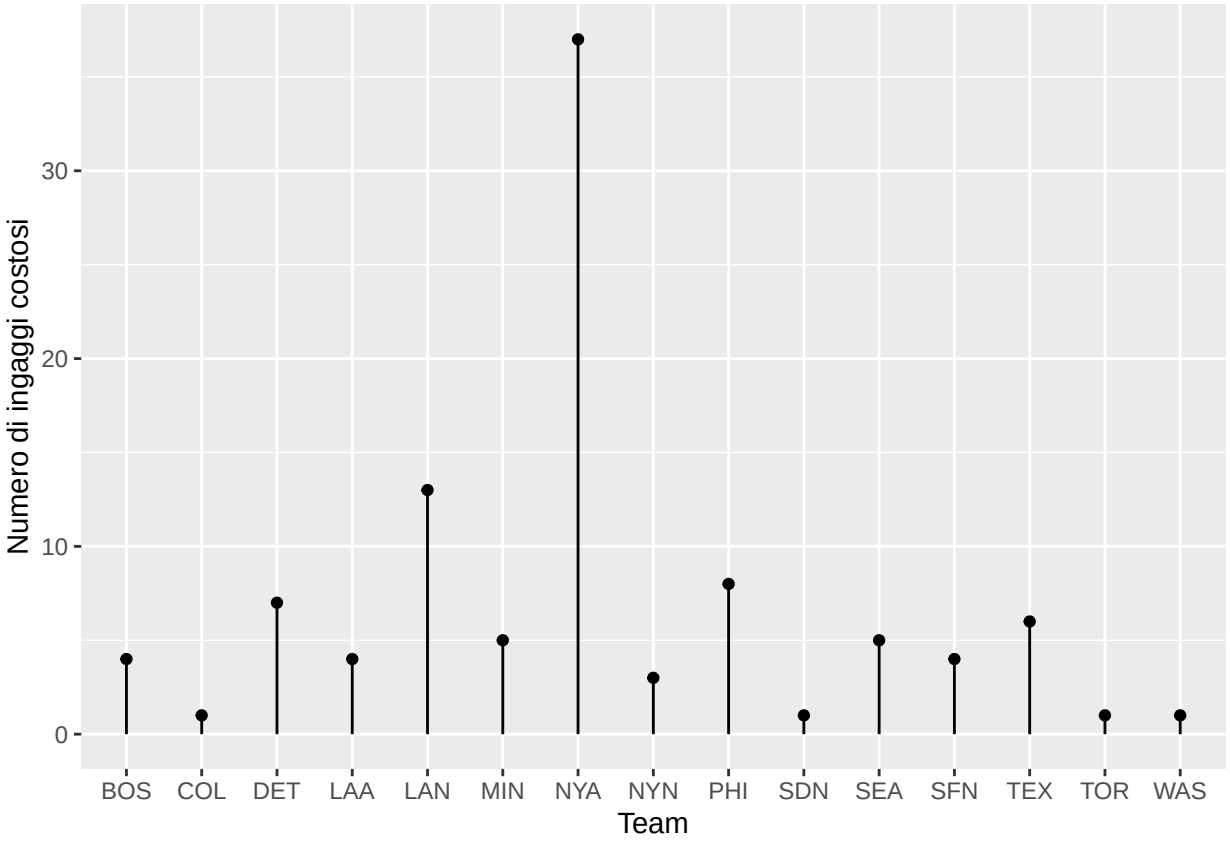
```
mPP = pSD %>% # most payed players

  filter(pSD$year %in% consYears) %>%
  select(year, team_id, name_given, salary)

# ordino i salari dal più alto al più basso

mPP = mPP[order(-mPP$salary),]
```

Considero solo i primi 100 giocatori con gli stipendi più alti negli anni considerati e conto il numero di squadre in cui hanno giocato.

```
topHundredSalary = head(mPP, 100) %>%
count(team_id, sort = T)
topHundredSalary
```

```
##    team_id  n
## 1      NYA 37
## 2      LAN 13
## 3      PHI  8
## 4      DET  7
## 5      TEX  6
## 6      MIN  5
## 7      SEA  5
## 8      BOS  4
## 9      LAA  4
## 10     SFN  4
## 11     NYN  3
## 12     COL  1
## 13     SDN  1
## 14     TOR  1
## 15     WAS  1
```

```
# plotto il grafico che mostra la quantità di giocatori che ogni squadra ha avuto negli anni e che rien

plotPPT = ggplot(topHundredSalary, aes(x=team_id, y=n)) +
  geom_point() +
  geom_segment(aes(xend=team_id, y=0, yend=n)) + xlab("Team") + ylab("Numero di ingaggi costosi")

plotPPT
```



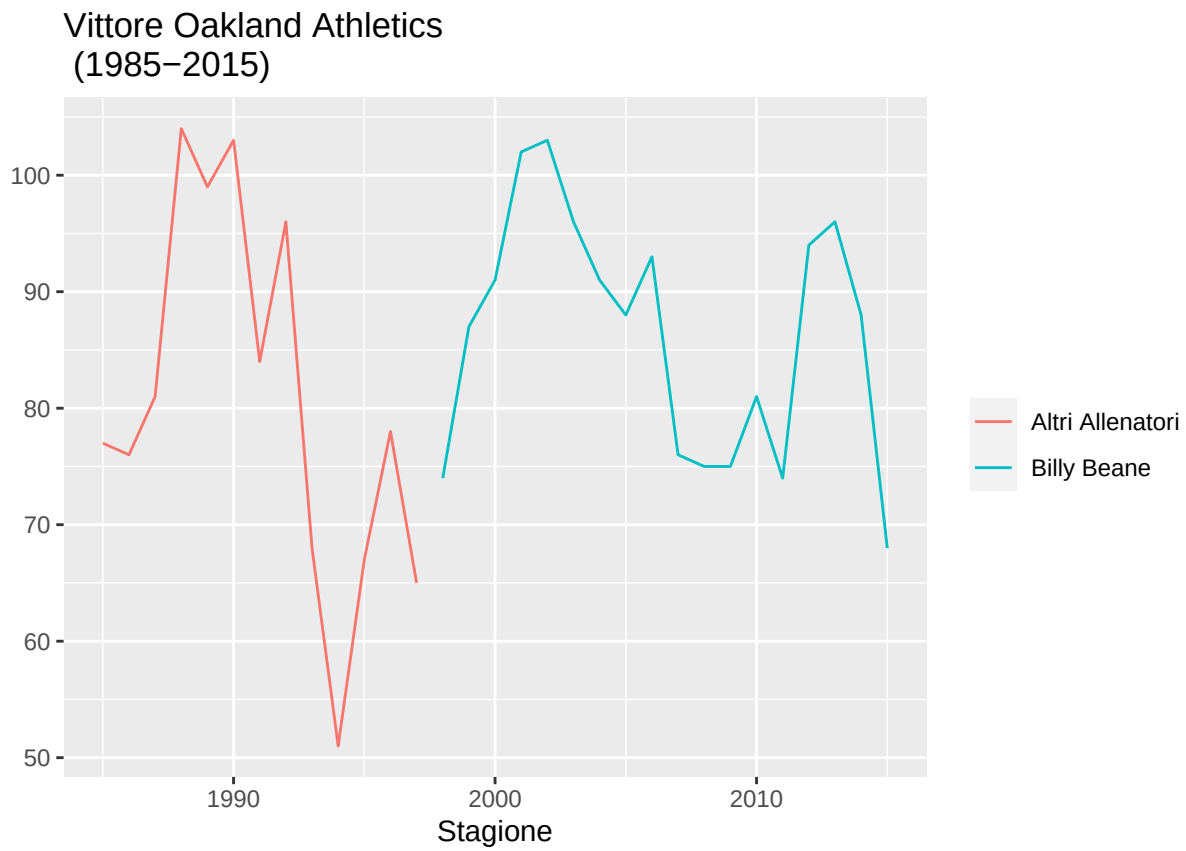
Notiamo che su tutti i New York Yankees hanno avuto per ben 37 volte dei giocatori con dei salari esorbitanti, questo indica che gli Yankees hanno un budget cospicuo e possono permettersi ingaggi elevati per vincere il campionato e i vari titoli, infatti come riportato in precedenza hanno vinto ben 5 world series negli anni considerati.

Ci sorge spontaneo chiederci se nel baseball per vincere serva davvero disporre di elevate quantità di denaro.

La rivoluzione di Billy Beane

Billy Beane è un dirigente sportivo ed ex giocatore di baseball, attualmente si occupa di amministrare e gestire le finanze della squadra degli Oakland Athletics ed è proprio da questa squadra che nel 2002 partì una rivoluzione che scosse tutto il mondo del baseball, insegnò alla Major League che non serve disporre di cifre spropositate di denaro per essere competitivi.

```
# trovo l'andamento generale degli Oakland Athletics sotto la guida di Billy e la paragono agli altri m

oaklandAth = baseballTeams %>%
        filter(name == "Oakland Athletics") %>%
        mutate(BB = ifelse(year >= 1998, "Billy Beane", "Altri Allenatori"))

plot = ggplot(oaklandAth, aes(x = year, y = wins, colour=factor(BB)))
plotStatsBB = plot + geom_line() + ggtitle("Vittore Oakland Athletics \n (1985-2015)")+ xlab("Stagione")
  theme(legend.title = element_blank())

current_div = oaklandAth$div_id[oaklandAth$year == 2015] %>%
              as.character
plotStatsBB
```



Dal grafico non sembra che Beane abbia riportato gli Oakland Athletics in vetta alle classifiche ma in realtà ha dimostrato che si può essere competitivi pur avendo una squadra che dispone di un budget limitato.

Ma facciamo un passo indietro e parliamo della stagione 2001, gli Oakland Athletics dopo la sconfitta contro i New York Yankees e l'esclusione dalle World Series dovettero affrontare un problema ancora più grande, la partenza di tre dei loro migliori giocatori, Johnny Damon, Jason Giambi, e Jason Isringhausen.

> "Ci sono squadre ricche e ci sono squadre povere. Poi ci sono venti metri di m***a. Poi ci siamo noi."
> Cit: Billy Beane (Brad Pitt), Moneyball, l'arte di vincere.

Sembrava quasi impossibile per Billy risolvere quel problema e assemblare in poco tempo una squadra competitiva con un budget limitato ma grazie all'aiuto di Paul DePodesta, suo assistente laureato in economia alla Harvard University sviluppò un algoritmo per valutare un giocatore in base a delle statistiche ben precise e non attraverso dei pregiudizi o delle stagioni di successo passate come accadeva fino a quel momento.

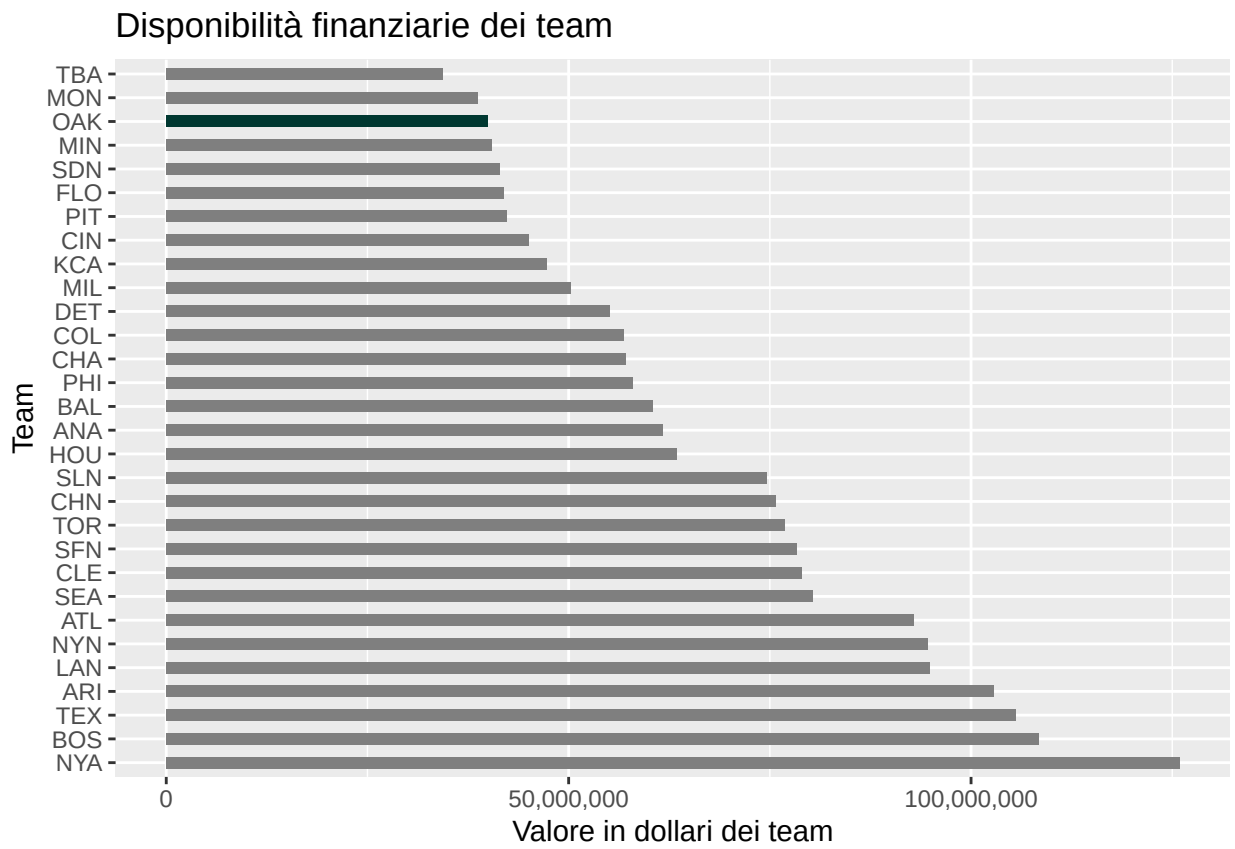
> "Tra i 20.000 giocatori che vale la pena di valutare credo ci sia una squadra da titolo di 25 giocatori che ci possiamo permettere... una specie di isola dei giocattoli difettosi"
> Cit: Peter Brand aka Paul DePodesta (Jonah Hill), Moneyball, l'arte di vincere.

Consideriamo quindi la squadra che mise assieme Billy nel 2002 e confrontiamo i budget che gli Oak hanno speso per gli ingaggi dei loro giocatori rispetto a quelli delle altre squadre.

Partiamo quindi dal presupposto che se una squadra spende molto per mantenere i propri giocatori disponga di cifre di denaro molto più significative rispetto a chi invece non può spendere o spende meno e che quindi il valore della squadra sia pari alla somma spesa per gli ingaggi dei giocatori.

```
# prendo i salari dei giocatori e li sommo ottenendo così il totale che ogni squadra deve spendere per

teamCost = pSD %>%
```

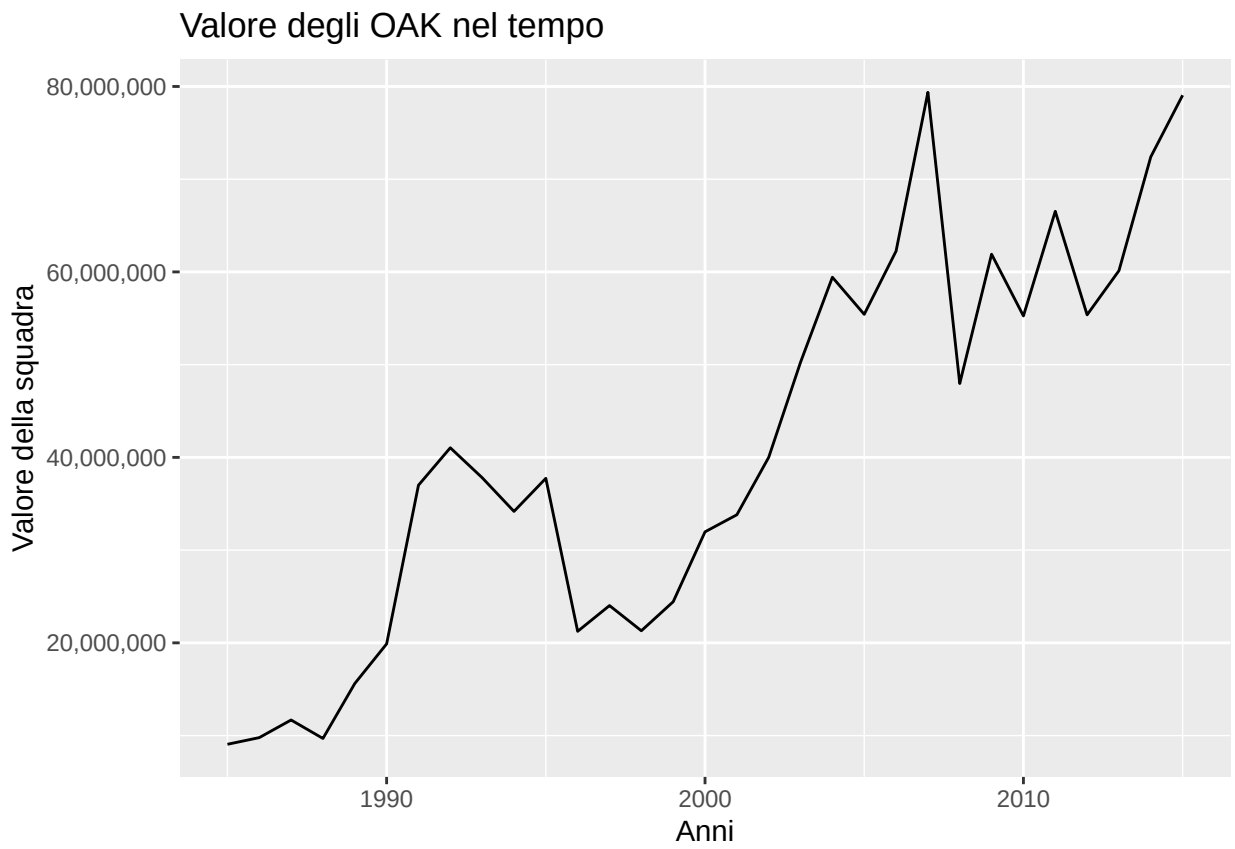
```
  group_by(year, team_id) %>%
  summarise(cost_salary = sum(salary)) %>%
  filter(year == 2002)

# prendo i team del 2002

baseballTeam2002 = baseballTeams %>%
             filter(year == 2002)

# stampo il grafico che ci fa capire meglio la spesa degli Oakland Athletics per i suoi giocatori rispe

area.color = c(baseballTeam2002$team_id)
area.color[!grepl('OAK', area.color)] = "OT"
cols = c("OT" = "grey50", "OAK" = "#003831")

plot = ggplot(teamCost, aes(x = reorder(team_id, -cost_salary), y = cost_salary, fill = area.color)) +
  scale_fill_manual(values = cols)

plotBudgetTeams = plot + geom_bar(stat = "identity", width = 0.5) +
  coord_flip() + xlab("Team") + ylab("Valore in dollari dei team") + scale_y_continuous(labels = scales
  ggtitle("Disponibilità finanziarie dei team") +
  theme(legend.position='none')

# stampo il grafico

plotBudgetTeams
```



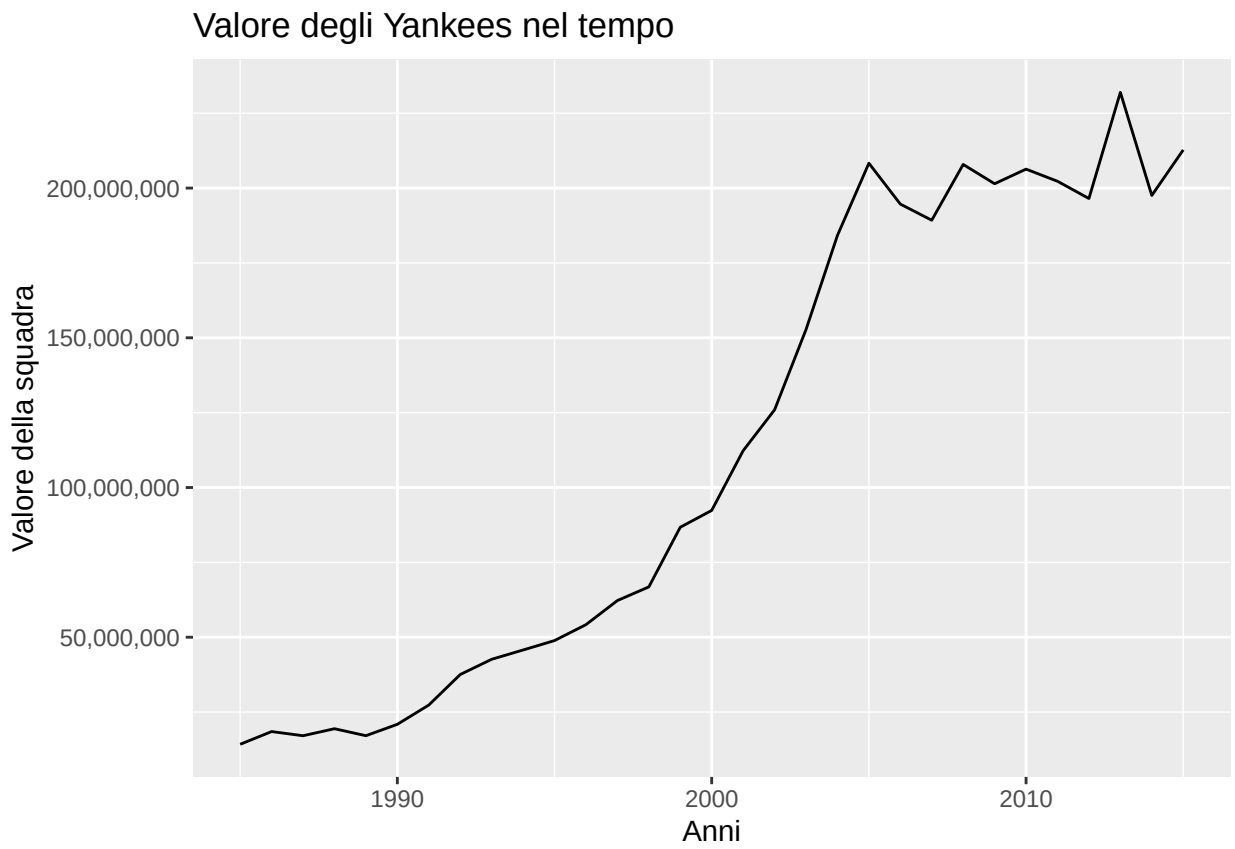
Ricordando quali sono le squadre che hanno vinto di meno nella storia possiamo capire quanto in realtà il denaro sia importante fino ad un certo punto in questo sport. I Seattle Mariners ad esempio sono tra le squadre meno competitive ma dispongono di ingaggi considerevoli.

Vediamo il valore della formazione degli Oakland Athletics e dei New York Yankees negli anni considerati.

```
oakCost = pSD %>%
  group_by(year, team_id) %>%
  summarise(cost_salary = sum(salary)) %>%
  filter(year %in% consYears, team_id == "OAK")

# creo e stampo il grafico di andamento nel tempo del valore degli Oak

plotOakValue = ggplot(oakCost, aes(x = year, y = cost_salary)) +
  geom_line() + scale_y_continuous(labels = scales::comma) +
  xlab("Anni") + ylab("Valore della squadra") +
  ggtitle("Valore degli OAK nel tempo")

plotOakValue
```




```
nYYCost = pSD %>%
  group_by(year, team_id) %>%
  summarise(cost_salary = sum(salary)) %>%
  filter(year %in% consYears, team_id == "NYA")

# creo e stampo il grafico di andamento nel tempo del valore degli Oak

plotnYYValue = ggplot(nYYCost, aes(x = year, y = cost_salary)) +
  geom_line() + scale_y_continuous(labels = scales::comma) +
  xlab("Anni") + ylab("Valore della squadra") +
  ggtitle("Valore degli Yankees nel tempo")

plotnYYValue
```



Quindi la squadra di Billy era in netto svantaggio economico rispetto alle altre squadre della Major League ma con appena un valore di

```
oakCost2002 = oakCost %>%
  filter(year == 2002)
oakCost2002
```

```
## # A tibble: 1 x 3
## # Groups:   year [1]
##    year team_id cost_salary
##   <int> <chr>         <int>
## 1  2002 OAK        40004167
```

Si mise in gioco affrontando squadre molto più blasonate che valevano e spendevano più del triplo delle possibilità degli Oakland Aths e i risultati furono sorprendenti.

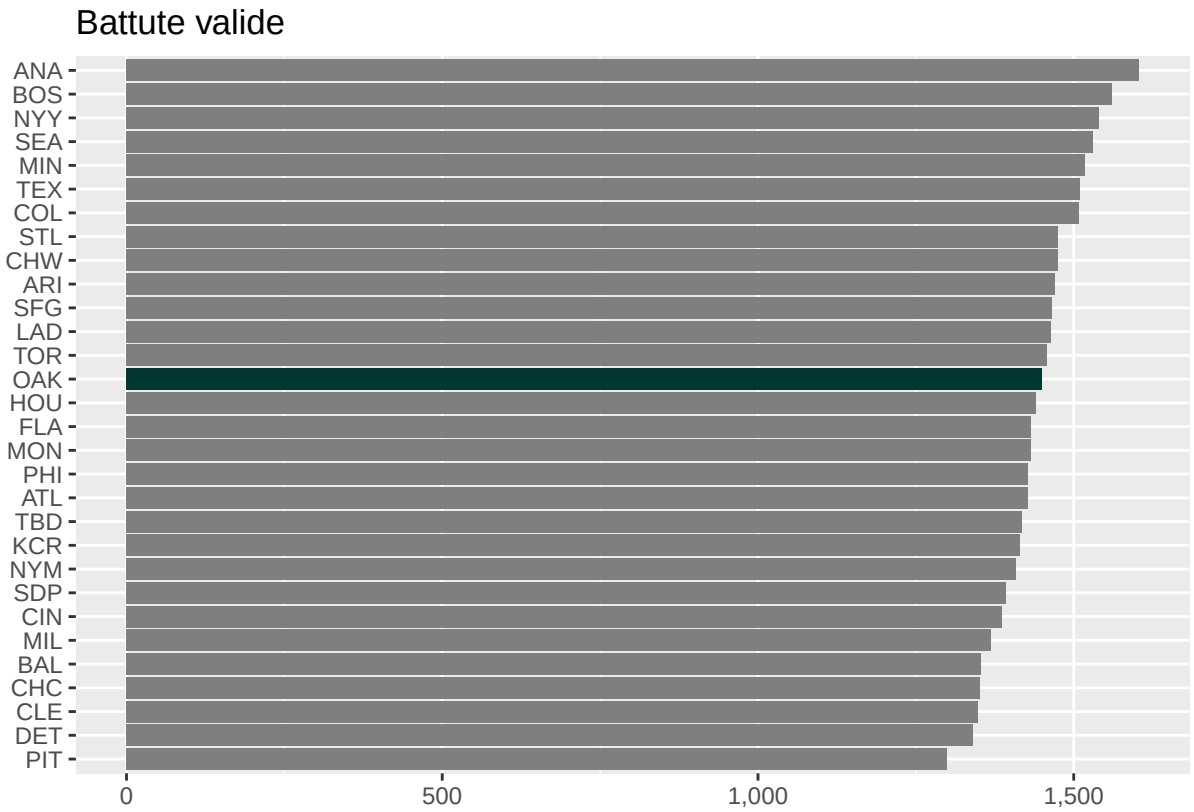
> "Lo scopo non deve essere comprare giocatori: lo scopo deve essere comprare vittorie."
> Cit: Billy Beane (Brad Pitt), Moneyball, l'arte di vincere.

E le vittorie nel baseball si ottengono conquistando il maggior numero di basi, completando più home run possibili, eliminando il maggior numero di battitori e facendo il minor numero di errori.

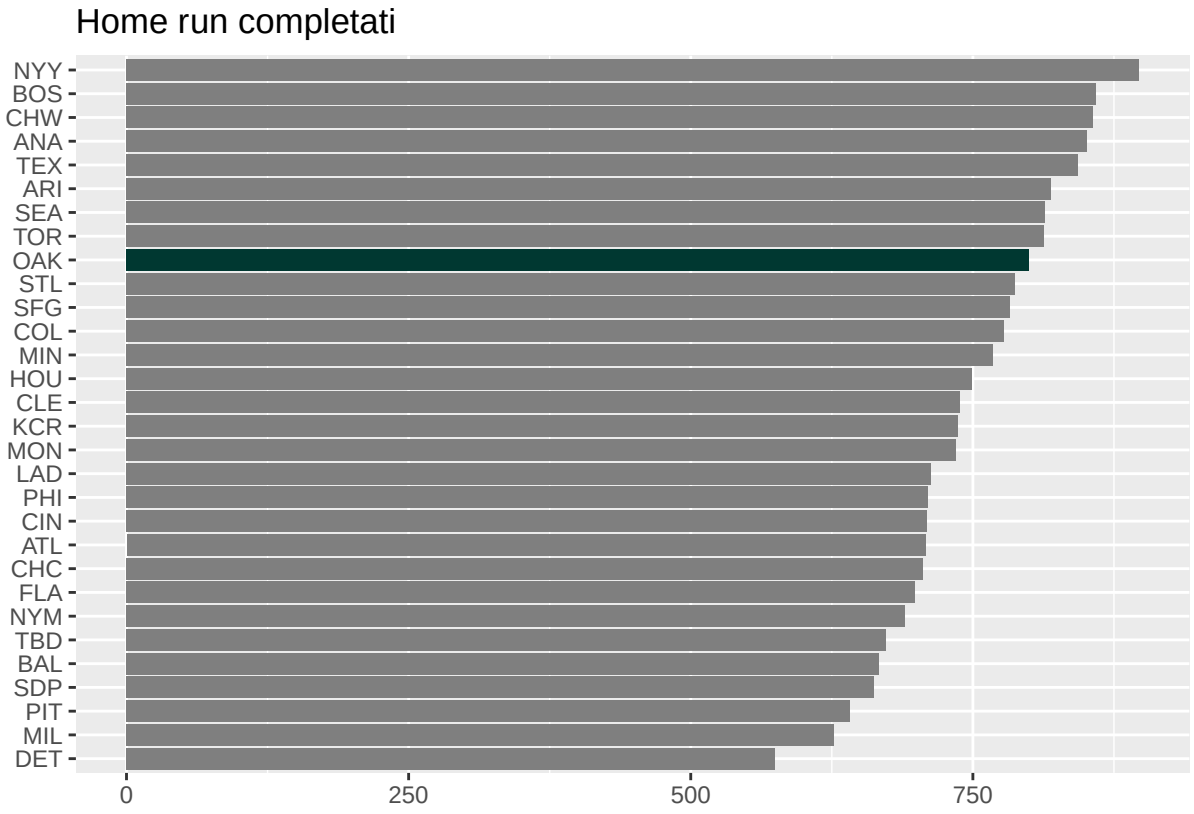
Prendiamo quindi i team della stagione 2002 e vediamo le statistiche nei campi descritti sopra.

```
# grafico delle battute valide, ovvero quando il battitore colpisce la palla e riesce a raggiungere la

plotGoodHit = ggplot(baseballTeam2002, aes(y = hits, x = reorder(team_id, hits), fill = area.color)) +
  scale_fill_manual(values = cols) +
  coord_flip() + xlab("") + ylab("") +
  scale_y_continuous(labels = scales::comma) + geom_bar(stat = "identity") + ggtitle("Battute valide") +
```

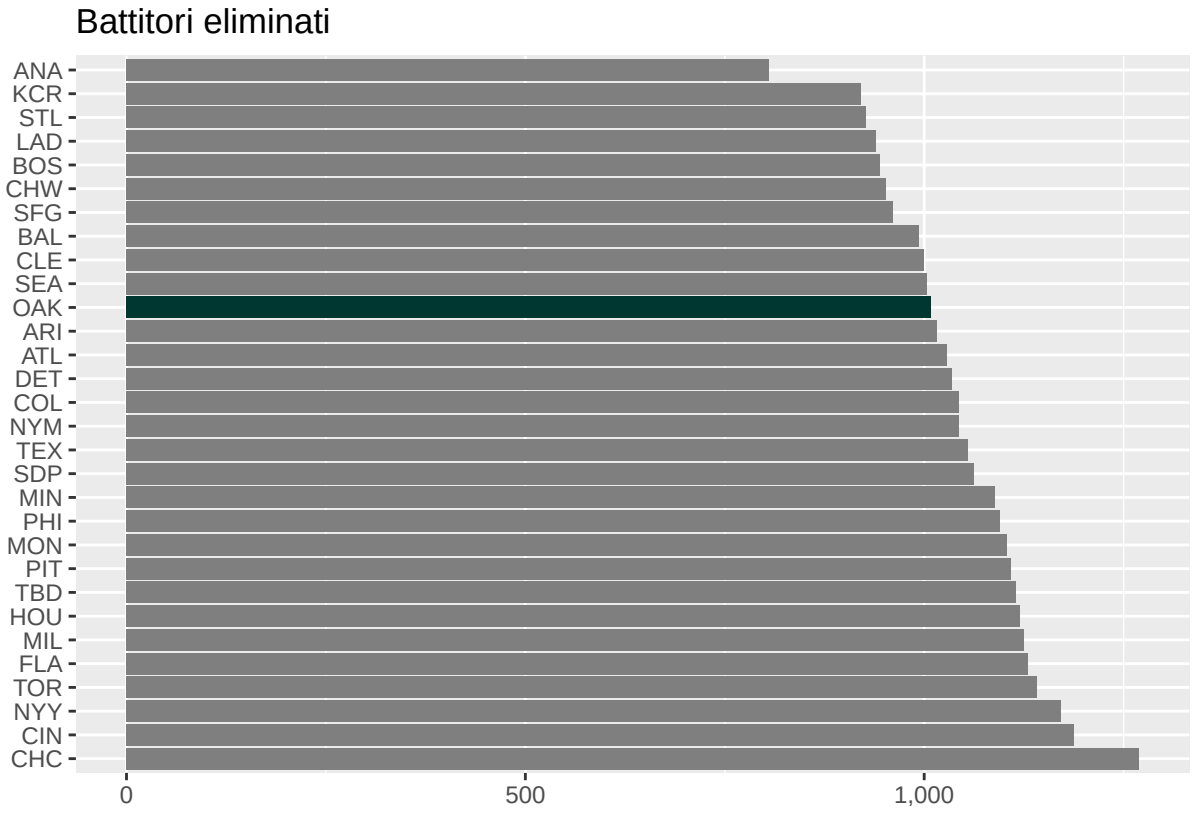
```
  theme(legend.position='none')
plotGoodHit
```



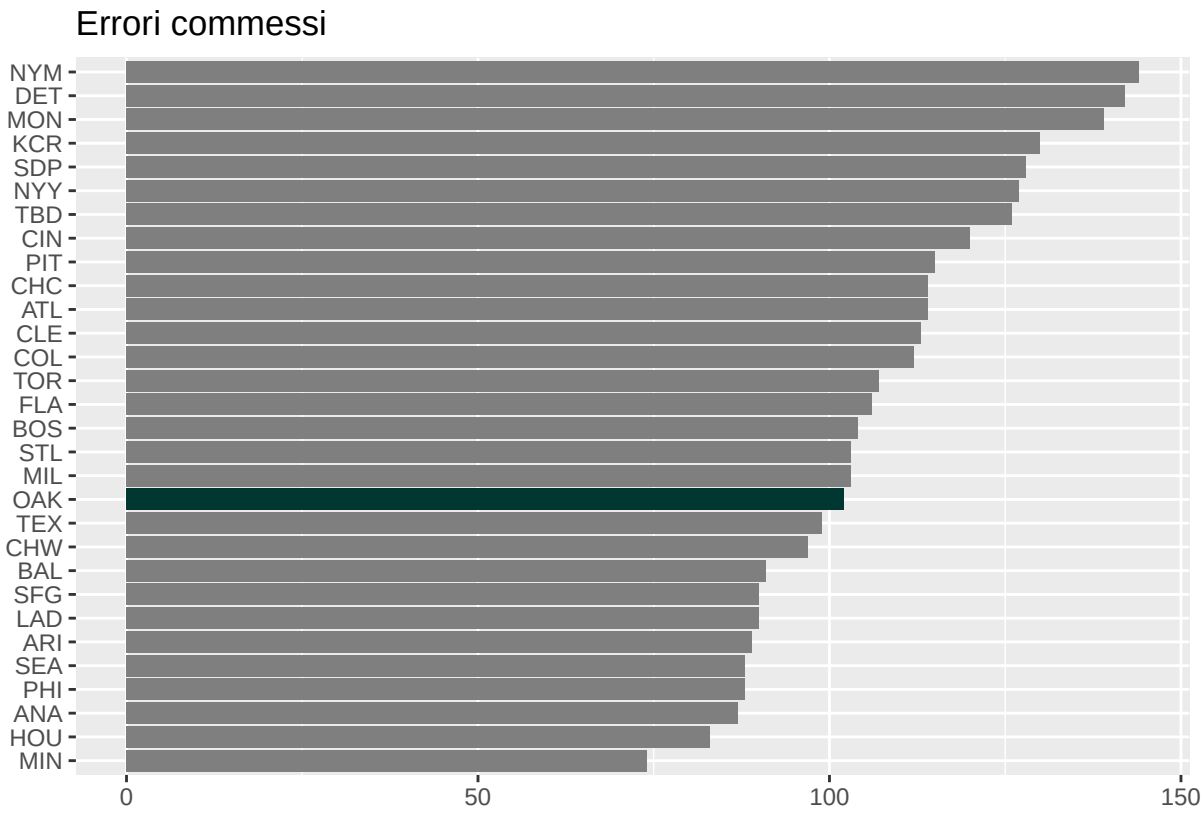

```
# grafico degli home run effettuati, più hr si effettuano maggiori saranno i punti e i battitori salvi

plotHomeRun = ggplot(baseballTeam2002, aes(y = runs, x = reorder(team_id, runs), fill = area.color)) +
  scale_fill_manual(values = cols) +
  coord_flip() + xlab("") + ylab("") +
  scale_y_continuous(labels = scales::comma) + geom_bar(stat = "identity") + ggtitle("Home run completa
  theme(legend.position='none')
plotHomeRun
```




```
# grafico degli strike che ogni squadra ha subito, in questo caso il grafico è da leggere "al contrario

plotStrike = ggplot(baseballTeam2002, aes(y = strike_out, x = reorder(team_id, -strike_out), fill = area
  scale_fill_manual(values = cols) +
  coord_flip() + xlab("") + ylab("") +
  scale_y_continuous(labels = scales::comma) + geom_bar(stat = "identity") + ggtitle("Battitori elimina
  theme(legend.position='none')
plotStrike
```

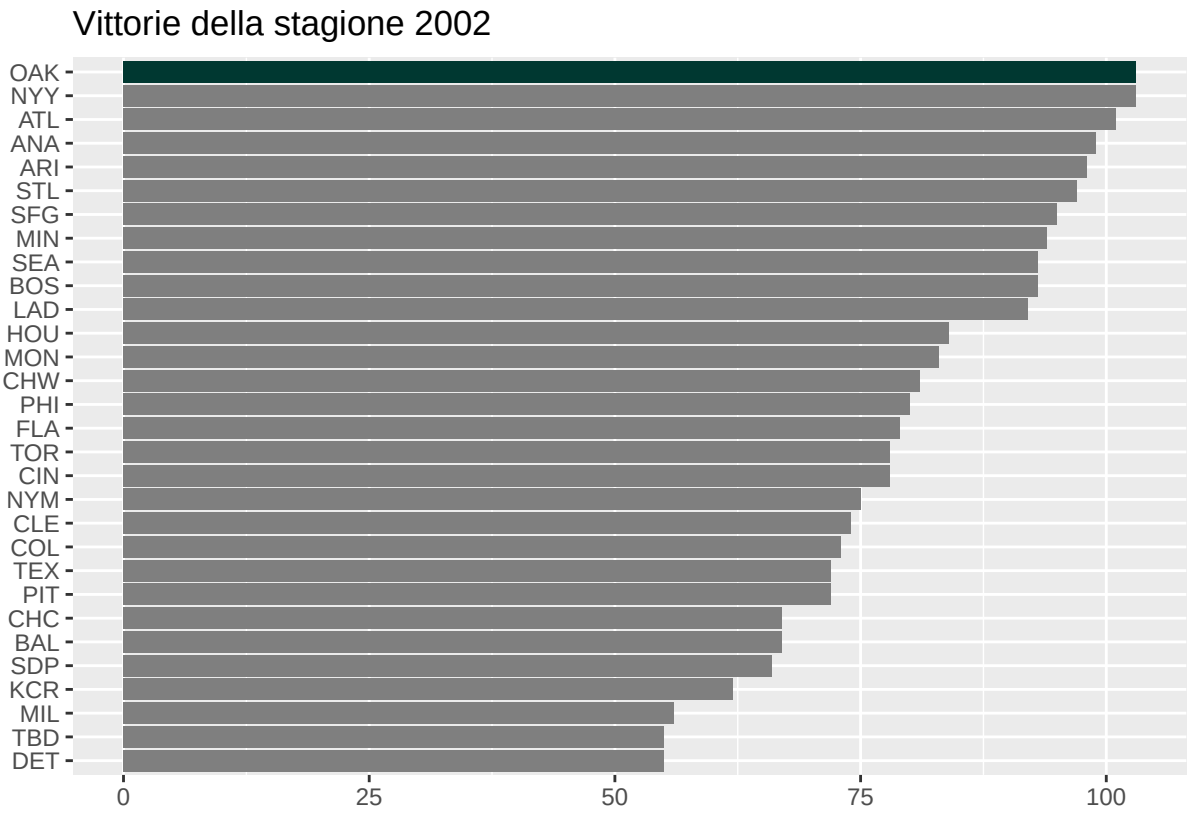



```
# grafico degli errori commessi

plotError = ggplot(baseballTeam2002, aes(y = errors, x = reorder(team_id, errors), fill = area.color)) +
  scale_fill_manual(values = cols) +
  coord_flip() + xlab("") + ylab("") +
  scale_y_continuous(labels = scales::comma) + geom_bar(stat = "identity") + ggtitle("Errori commessi") +
  theme(legend.position='none')

plotError
```



Notiamo come questi quattro fattori influiscano notevolmente sul numero di vittorie conseguite.

```
winPlot = ggplot(baseballTeam2002, aes(y = wins, x = reorder(team_id, wins), fill = area.color)) +
  scale_fill_manual(values = cols) +
  coord_flip() + xlab("") + ylab("") +
  scale_y_continuous(labels = scales::comma) + geom_bar(stat = "identity") + ggtitle("Vittorie della sta
  theme(legend.position='none')
winPlot
```



Gli Oakland Athletics dal 2000 al 2003 raggiunsero i playoff per quattro volte consecutive e divennero la prima squadra in oltre 100 anni di American League a vincere 20 partite di seguito.

Billy Bean ce l'aveva fatta, aveva dimostrato come si possono ottenere vittorie con un budget limitato e questo come già scritto in precedenza rivoluzionò il baseball.

Notiamo anche come nei grafici precedenti gli Anaheim Angels (ANA) si siano sempre posizionati meglio rispetto alle altre squadre, infatti nella stagione 2002 vinsero le World Series battendo in finale i San Francisco Giants (SFN).

Una squadra da:

```
ANA = teamCost %>%
  filter(year == 2002, team_id == "ANA")
paste(ANA$cost_salary, "$")
```

```
## [1] "61721667 $"
```

Riuscì a batterne una da:

```
SFN = teamCost %>%
  filter(year == 2002, team_id == "SFN")
paste(SFN$cost_salary, "$")
```

```
## [1] "78299835 $"
```

E gli Oakland Athletics, il cui valore era di:

```
oakCost = teamCost %>%
  filter(year == 2002, team_id == "OAK")
```

```
paste(oakCost$cost_salary, "$")
```

```
## [1] "40004167 $"
```

Ottenerono lo stesso numero di vittorie dei New York Yankees:

```
NYA = teamCost %>%
  filter(year == 2002, team_id == "NYA")
paste(NYA$cost_salary, "$")
```

```
## [1] "125928583 $"
```

Considerazioni finali Nel baseball serve davvero molto denaro per avere successo?

No, a differenza di altri sport il baseball può essere molto più equilibrato, certo, le squadre con un elevato quantitativo di capitale possono permettersi ingaggi più considerevoli ma Billy Bean ha dimostrato che non serve disporre di ingaggi esorbitanti per vincere.