Progetto fondamenti di scienza dei dati

Correlazione tra spesa in ricerca scientifica e premi Nobel

Aleksa Aleksic

Il progetto nasce dalla personale curiosità e dalla volonta di sfruttare alcune delle conoscenze apprese a lezione per provare a dare vita a delle considerazioni interessanti. La scelta del dataset sui premi Nobel deriva dalla volontà di scegliere un dataset interessante ma allo stesso tempo ben costruito.

I datset in se sono stati scaricati da due fonti:

- Kaggle;
- OECD (Organizzazione per la cooperazione e lo sviluppo economico);

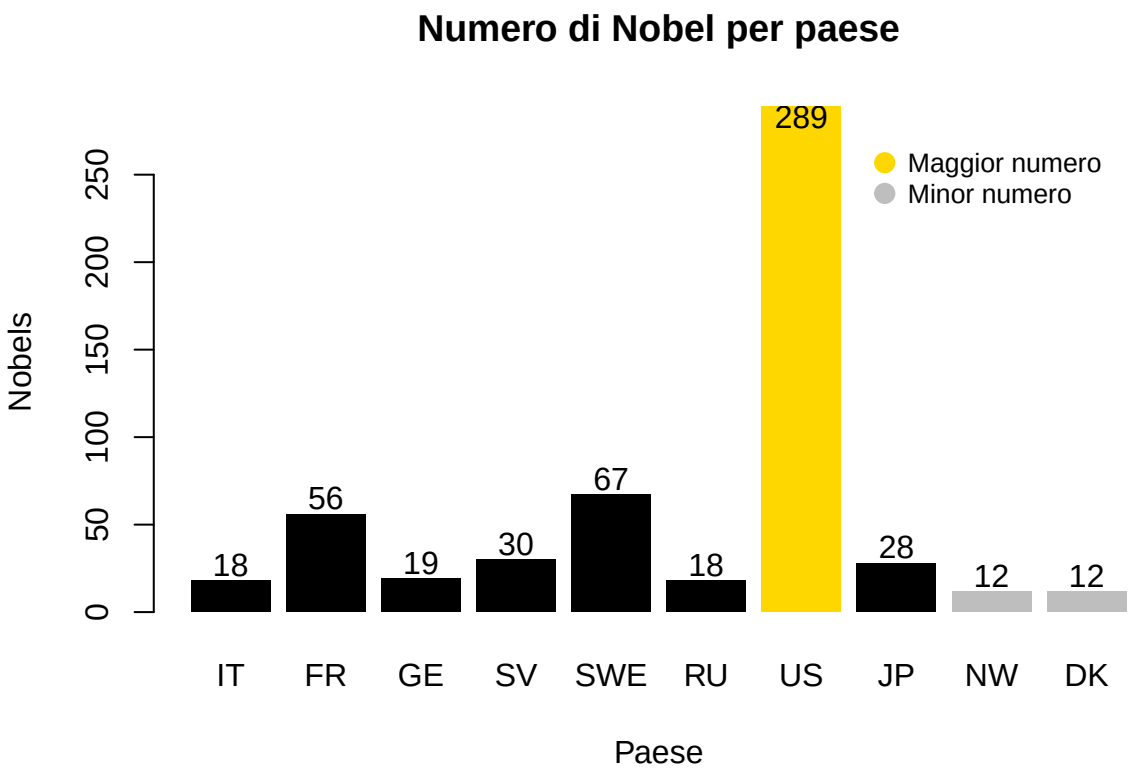
In se tutto il progetto ruota intorno a una serie di domande,ovvero:

- Quale paese ha vinto più premi nobel?
- Quanti ne ha vinti nel secolo corrente?
- Quanto ha speso tale paese in ricerca scientifica?
- Si può trovare una correlazione?
- Si può predire il prossimo numero di vincitori?

E' importante sottolineare che la prima parte del progetto considera i Nobel vinti dall'inizio della competizione (1901) fino all'anno 2021, esclude apposta il 2022 in quanto lo tiene buono per la parte finale del progetto nella quale sarà utilizzato per provare a prevedere il numero di vincitori e capire quanto vicino riesca ad andarci l'algoritmo. La parte in cui si guardano gli esempi da vicino considera invece il periodo 2000-2021, mentre la parte finale considera gli investimenti per tutto il periodo presente nel dataset quindi 1981-2021.

Quale paese ha vinto più premi Nobel?

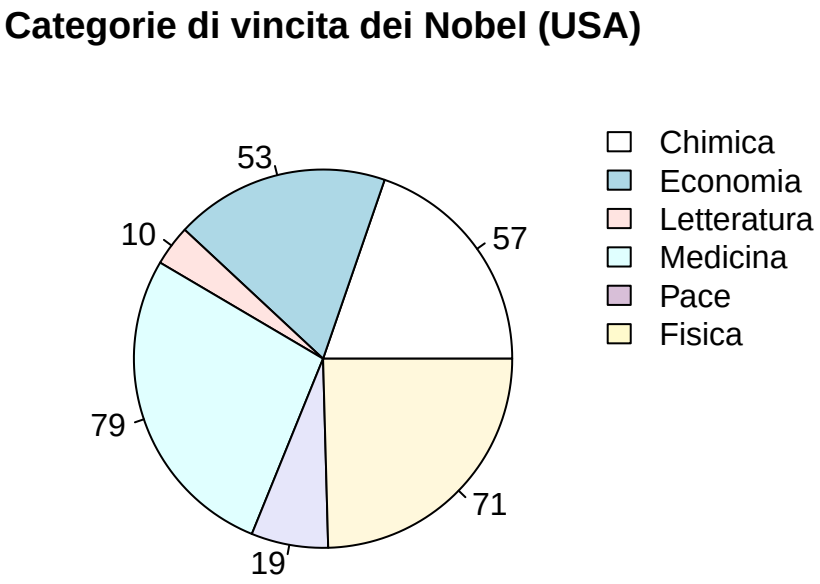
La risposta viene data elaborando un grafico a barre il quale prende in considerazione diesci paesi, ritenuti personalmente i più interessanti, e cerca di capire quanti nobel avesse vinto ognuno di essi.



L'attenzione si sposta quindi sul paese più titolato, ovvero gli Stati Uniti d'America.

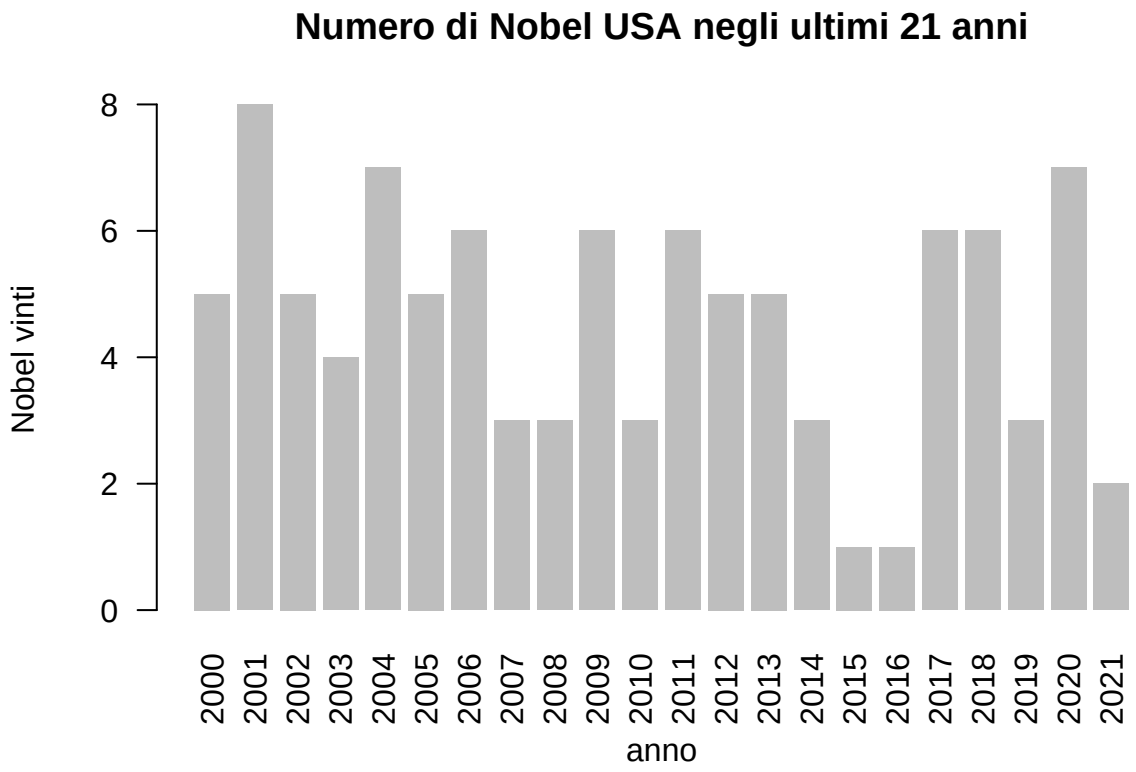
Quanti Nobel vinti per categoria?

Elaborando un grafico a torta cerchiamo di capire quali categorie hanno ottenuto più vincite e quali meno, diamo quindi un quadro un po' più specifico.



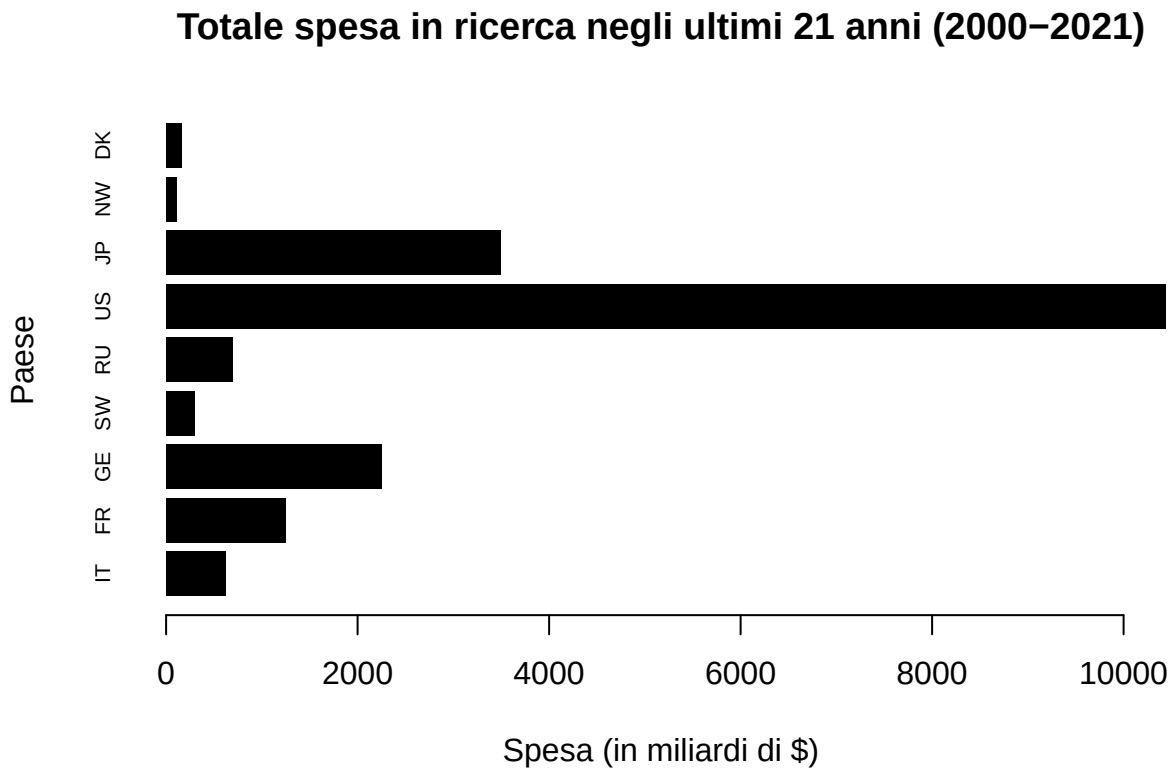
Quanti ne ha vinti negli ultimi anni?

Ristringiamo il campo di intersse, concentriamoci sull'ultimo secolo (2000-2021) per capire quanti Nobel fossero stati vinti per ogni anno nell'epoca moderna.



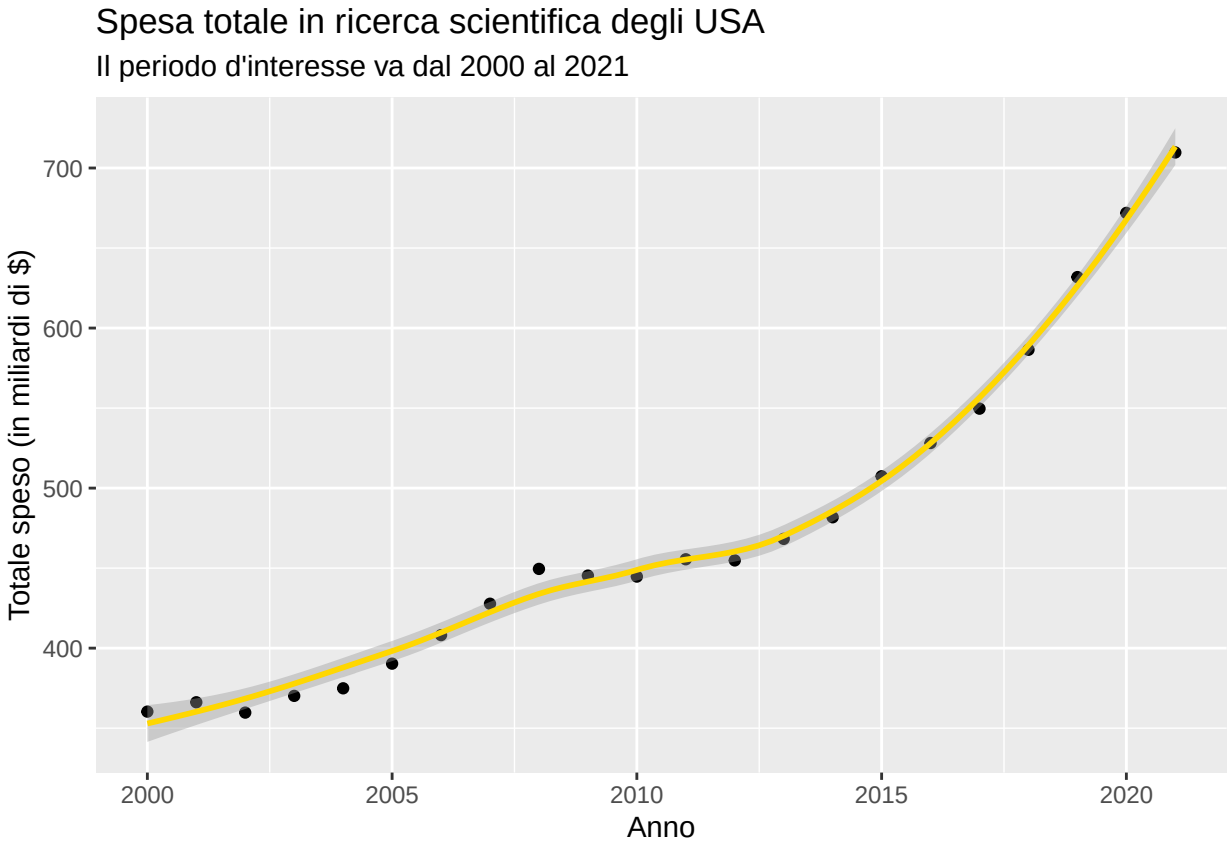
Quanto hanno speso in ricerca i paesi di interesse negli ultimi 21 anni?

Cerchiamo una correlazione tra le cose, consideriamo il dataset sugli investimenti in ricerca e sviluppo fornito dalla OECD e proviamo a capire quanto abbiano speso in ricerca e sviluppo i paesi che abbiamo considerato prima per vedere se spendere di più equivale a vincere di più. I dati sono relativi a tutti i paesi considerati prima a esclusione della Svizzera, per la quale nel dataset non sono presenti informazioni, e indicano la spesa in miliardi di $ per ogni paese.



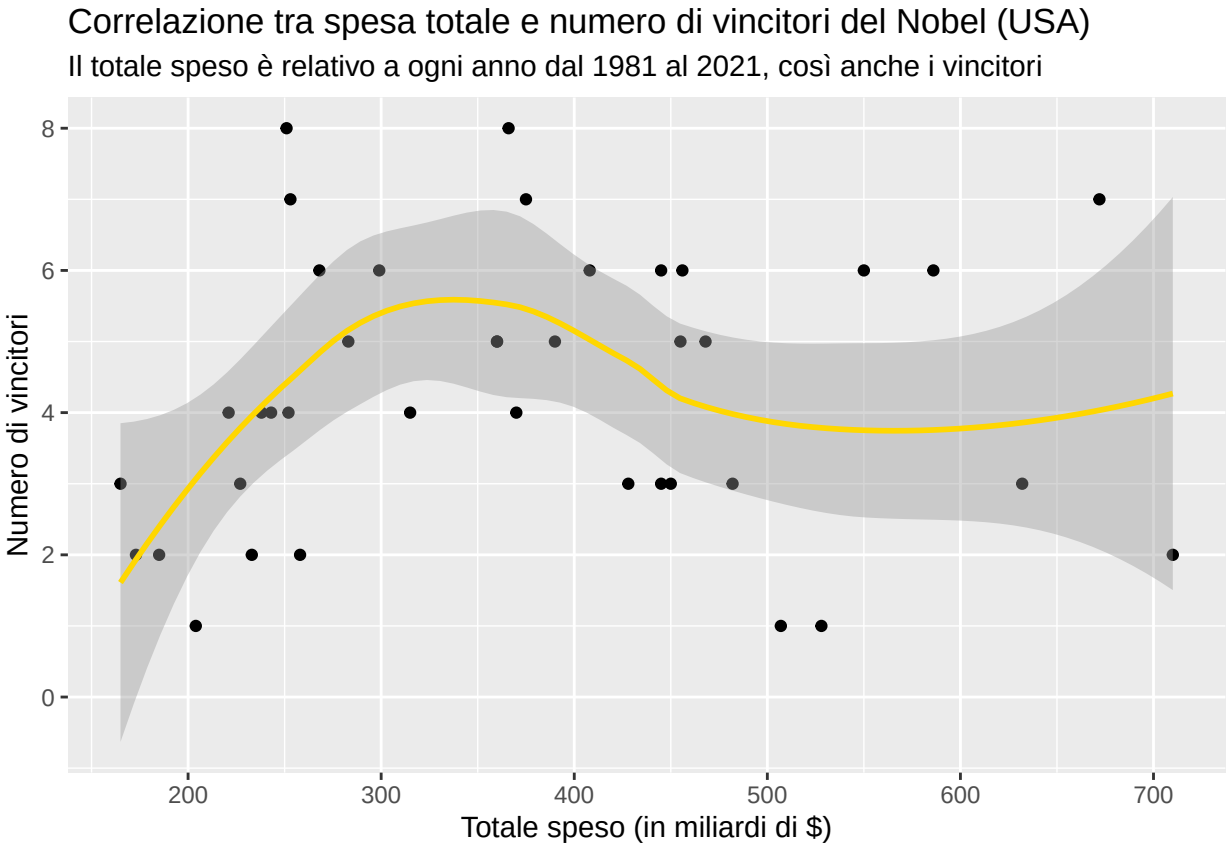
Come si è sviluppata la spesa degli USA negli anni?

Osserviamo quindi più da vicino la spesa effettuata dagli Stati Uniti e guardiamo quanti miliardi di $ hanno speso anno dopo anno dal 2000 al 2021, disegnamo quindi una linea di tendenza e notiamo che ogni anno tale spesa è aumentata sempre di più.



Si può trovare una correlazione?

Finiamo questa parte cercando una correlazione tra la spesa e i premi nobel vinti, in questo caso torniamo a considerare tutto il periodo di tempo che abbiamo a nostra disposizione, cioè dal 1981 al 2021, teniamo sempre fuori il 2022.



Si può anche provare a predire quanti ne avrebbe dovuti vincere nel 2022?

In conclusione, dopo aver tenuto fuori il 2022, cerchiamo di utilizzare le conoscenze acquisite per provare a predire il numero di vincitori del 2022 sapendo che la spesa fosse di 703 miliardi di $ e sapendo anche che gli effettivi vincitori fossero stati 6. La scelta del 2022 è stata fatta apposta, infatti scegliendo tale anno ci è possibile sfruttare tutto ciò che abbiamo calcolato prima e sopratutto ci è possibile confrontare ciò che otteniamo col il risultato effettivo.

Cerchiamo un indice di correlazione e gli indici in questo modo osservando tali valori possiamo vedere quanta correlazione c'è tra le due variabili.

```
## [1] 0.09001746
```

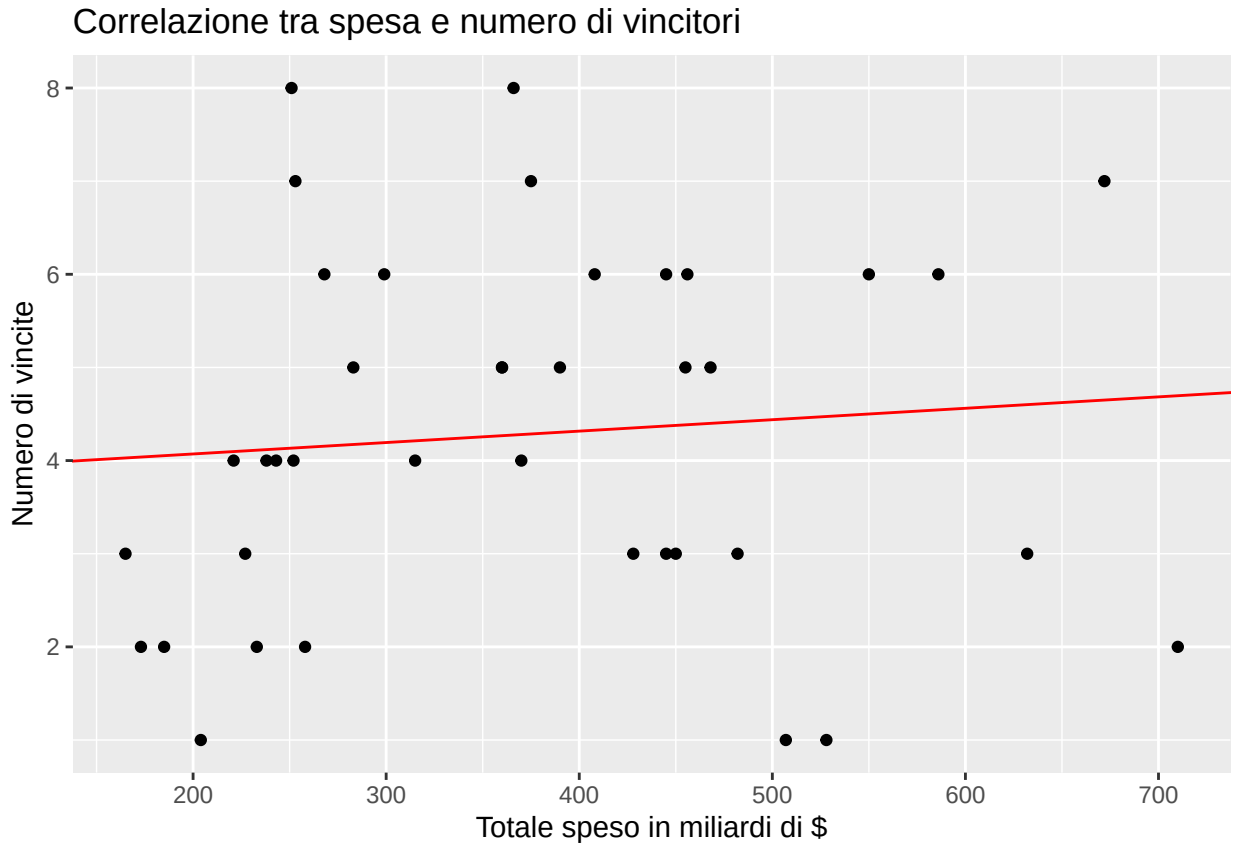
```
## (Intercept)       Value
## 3.825921409 0.001225902
```

```
## [1] 0.008103143
```

Il che vuol dire che possiamo riassumere la relazione come:

numero di vincitori = 0.001225902 *Totale speso + 3.825921409

Consideriamo il numero di vincitori come variabile dipendente (y) mentre il totale speso come variabile indipendente (x) e proviamo a tracciare un grafico della regressione lineare. Ovviamente, siccome la correlazione è minima e siccome la vincita di un premio Nobel non dipende solo dai soldi investiti in ricerca ma da una grande quantità di fattori, il rapporto tra le due variabili non è di totale linearià tuttavia notiamo che una minima linearità esiste, quindi in parte (anche se in piccola parte) possiamo dire che investire tanto in ricerca porta a vincere più Nobel.



Utilizziamo quindi la funzione predict() di R per provare a predire i vincitori per il 2022.

```
## [1] 4.69
```

Concludiamo il tutto estraendo il valore di tale predizione, ottenendo come risultato 4.69 (arrotondabile a 5), lo confrontiamo con il numero effettivo di vincitori,cioè 6, e considerando che i dati partivano dal 1981 e non dal 1901 (data della prima edizione del Nobel), e che andrebbero presi in considerazione molti altri fattori, possiamo ritenerci soddisfatti del risultato ottenuto.